Bachelor Thesis

Development of Trip Generation Equations for the Gold Coast and Sunshine Coast Regions

Veitch Lister Consulting

Brisbane, Australia

08 February 2008

Robert van Beilen
s0043753

Trip Generation

Review and development of the VLC Trip Generation Equations for the Gold Coast and Sunshine coast region

Veitch Lister Consulting P/L,
Suite 3, Ground Floor,
80 Jephson Street, Toowong, QLD 4066, Australia.
http://www.veitchlister.com.au

Daily supervisor: Tim Veitch

Twente University
Faculty of Engineering Technology
Civil Engineering (& Management)
Drienerlolaan 5
7522 NB Enschede
http://www.cit.utwente.nl

Supervisors: prof. dr.ir. E. (Eric) C. van Berkum
dr. T.(Tom) Thomas

Author: R. (Robert) J. van Beilen
Waalstraat 219
7523 RG Enschede
r.j.vanbeilen@student.utwente.nl

Executive Summary

The current method of modelling trip production in Veitch Lister Consulting's Trip Generation Model is to use simple linear regression with stratified household attributes included as dummy 0-1 variables. While this is effective for calculating *average* trip rates for various household attributes, the nature of the data is such that the assumptions of linear regression are seriously violated. This means that the usefulness of these averages applied at various levels is questionable and measures of model accuracy are not to be trusted entirely. A technique commonly cited in the literature for accommodating the assumptions of linear regression with count data is to take a log transform of those counts. However, in this case, the data is also heavy in zeros and so a log transform will actually worsen the performance. In fact there is no transformation that can 'spread out' a large number of zeros to make data suitable for simple linear regression. Instead, Zero-Inflated Poisson and Zero-Inflated Negative Binomial models are proposed as an improvement and assume to be more effective in modelling count data heavy in zeros. These zero-inflated models however, are non-linear and thus it is not possible to simply apply them at the equations zonal (aggregate) level as can be done with linear models. This report details the model assumptions, fit, diagnostics and performance for each of the six trip production purposes under the linear regression technique. The zero-inflated models are finally discussed as possible future improvements but this will need some further investigation and research.

Table of Contents

Table of Figures

1 Introduction

The purpose of this report is to analyse the effectiveness of the current method of modelling trip generation used in Veitch Lister Consulting's (VLC) transport model and to propose alternatives. It provides an independent review of the trip generation equations as described in the technical report, *Calibration of an Integrated Travel Forecasting model* [VLCTN2, 1997]. The data used in the existing VLC model was collected from 1992/1995, incorporating data from 5565 households from Melbourne. The model was then tested against the 1992/'94 Household Travel Survey Data for the Brisbane Statistical Division and is currently implemented in the South East Queensland region. To facilitate an analysis the model structure will be recalibrated using available 2004/'05 South East Queensland Household Travel Survey (SEQHTS) data to obtain updated parameters and model fit estimates.

The current method of predicting trip productions is to use ordinary linear regression with dummy variables, see Ortuzar, J. de D. and Willumsen (1994) for an introduction. Briefly trips are split by purpose and for each purpose a separate regression equation is calculated relating the number of trips (for each purpose) to household attributes (such as the number of workers, dependants or vehicles). It is a relatively simple method of allowing zonal totals to be used as inputs for the model equations. However, some of these models have a very low explanatory power and violate some basic statistical assumptions of linear regression. These assumptions are that the range of dependant variables is unrestricted, the dependant and independent variables are linearly related, and that errors are normally distributed, uncorrelated with constant variance (that is, homoskedastic). The dummy formulation is used to avoid issues of non-linear relationships, but does not address the problems created by the restricted range of the dependent variable (i.e. the number of trips con only be positive) nor the non-normal, heteroskedastic errors. The use of simple regression is further complicated by the presence of a large number of zero trips for all trip purposes.

Thus the main objection to the current models is that the distribution of the variables is not appropriately considered. Often trip generation models use a log transform to account for the restricted range of the dependent variable and the heteroskedastic errors. However, the literature says count data with values close to zero may be more effectively modelled in a generalised linear model (GLM) framework with a Poisson distribution. The traditional alternative to linear regression is category analysis or cross-classification. This tends to have higher data demands than regression techniques, and does not provide measures of model fit or accuracy.

As a result, the following components are successively described in this report:

- First of al an outline of my personal objectives –what I want to accomplish– are presented;
- This is followed by an outline of the research plan which actually form the guiding principles how to tackle the problem;
- Chapter 4 gives an overview of the data collection and necessary preparations executed for this research;
- The literature on both regression and category analysis are reviewed in the following part;
- Chapter 6 outlines the specific methodology that has been used for updating the model;
- In chapter 7 the results of each model are given and compared in chapter 8;
- This is followed by a discussion and notion of possible future improvements;
- Finally conclusions and recommendations are given in the last chapter.

2 Personal Objectives

Before starting with the description of the research a short outline of my personal objectives (i.e. what I personally want to learn/accomplish with this project) is presented. These objectives will mainly concern knowledge and experience that I would like to gain in the field of traffic and transport engineering.

1) To get a better understanding of the actual field of work I would first of all like to learn and get more familiar with the use of transport related theories (e.g. trip generation, trip distribution, modal split and assignment etc.) in real-life projects. Questions I am asking myself regarding this case are: What part of the theory is really coming back in practice? What information/data is necessary to write a good and reliable traffic model?

2) I want tot gain knowledge about the use of OmniTRANS and the Zenith (VLC's transport modelling software) in practice as a tool for travel demand forecasting, transportation- / land use planning and traffic engineering.

3) Third thing I want to gain experience in is the consultancy work. The models are used to predict and simulate traffic, but what is actually done with the outcome? What does a typical advice look like? To what extend is the consultancy involved in the progress of a project?

Hopefully this internship will answers my questions and gives more insight the practical work of a consultancy in the field of transportation planning and traffic forecasting.

3 Research Plan

3.1 Company profile

Veitch Lister Consulting Pty Ltd (VLC) is one of the largest specialist transport and traffic planning consultancies in Australia. Established in 1986 by Mike Veitch, VLC specializes in providing a full range of transport planning services including:

- Travel demand forecasting (including model development and associated software);
- Transportation planning and policy development;
- Toll road revenue and demand forecasting;
- Public transport policy, route strategy and service planning;
- Traffic micro-simulation;
- Traffic engineering, parking analysis and local area traffic management;
- Socio-economic analysis;
- Economic evaluation;
- Studies of land use/transport interaction;
- Survey design, data collection, statistical and market analysis;
- Bicycle path and trail planning;
- Special event planning; and
- Transport planning training.

The company combines the skills of some of Australia's best transport planners and traffic engineers to offer a wide range of services to the public and private sectors. VLC has built on the extensive expertise of its staff to successfully complete assignments throughout Australia and in South-East Asia.

The company offers particular expertise in the areas of travel demand forecasting, toll modelling and public transport modelling and planning. VLC has revolutionized transport modelling in Australia, with the development of highly accurate travel forecasting models, which support planners in developing the most pragmatic and efficient solutions to a wide range of transport problems. VLC has invested 20 years, and millions of dollars in developing its models, which are by far the most accurate, sophisticated and well-researched in Australia. In particular their public transport models are at the cutting edge of world research, and are designed specifically for Australian conditions. The modelling techniques developed by VLC have become industry standard within Australia.

With Australia's most powerful modelling tools at their disposal, VLC's transport planning staff are able to add significant value in all areas of transport planning. Importantly, many of VLC commissions make use of their own integrated transport modelling software, ZENITH. This comprehensive modelling software allows detailed time-period modelling of both private and public transport. Integrated models can be used to estimate car and truck flows as well as a wide range of public transport data including trip patterns, boardings and alightings and load profiles.

The ZENITH model starts from a detailed description of the land uses within the modelled area to develop transit and vehicle trip patterns from 'first principles'. The assignment of these trip tables to the modelled infrastructure network is also state-of-the-art. Presentation of results is first class through integration with a graphical information system (GIS).

VLC's experience and technical expertise are highly regarded within the industry. VLC has been involved in the planning and modelling of several hundred major projects over an extended period of time, with a focus on close client collaboration, and providing the client with pragmatic, innovative and useful solutions [VLC Company Profile, 2007].

3.2 Project & Research Description

3.2.1 Background

Veitch Lister Consulting is the developer (and owner) of a comprehensive four-step model[1] of the South-East Queensland (SEQ) region. SEQ is a rapidly growing region of the state of Queensland and actually is the fastest growing region in Australia, fuelled principally by migration from the southern states. The SEQ region contains approximately two-thirds of the state population and is estimated to be 2.8 million which is heavily urbanised and concentrated along the coast. The three largest population areas are concentrated in Brisbane, Gold Coast and the Sunshine Coast which subsequently account for 90 per cent of the region's population. Future development plans of South East Queensland focus on slowing down coastal development, in order to prevent creating a 200 km city, and instead aim for growth in the west [WIKIQLD, 2007], [SEQIPP, 2007].

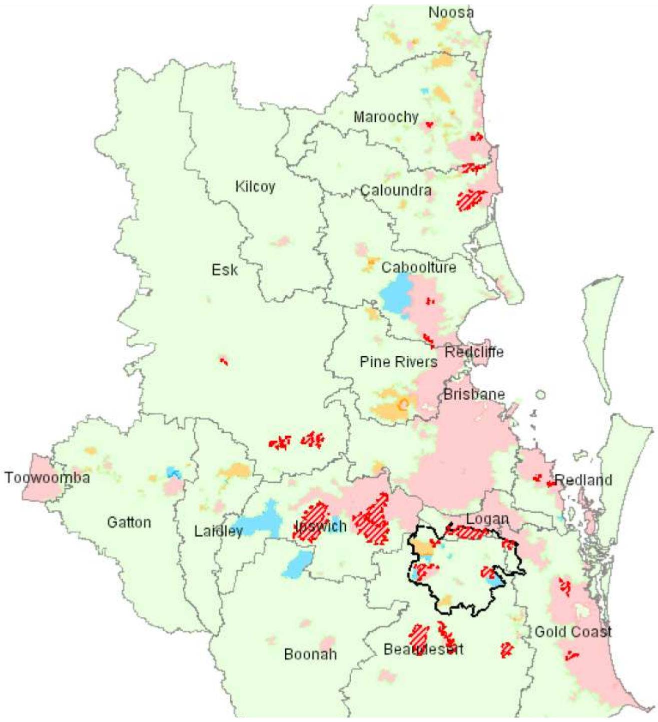


figure 1: The South-East Queensland Region

The modelled area spans nearly 400 km north-to-south, and nearly 200 km east-to-west, and is frequently utilized on a wide variety of transport planning and modelling projects. The Transport authorities and Planning agencies are required to make decisions on transport infrastructure and services worth billions of dollars. The decision making process for transport planning needs to be informed, accountable and founded on comprehensive, current and reliable data. One of the most important areas of information needed is an accurate description of travel behaviour of the people living in this area. The most effective way for transport planners to gather this information is by the conduct of a household travel survey. The last time a survey of this nature was conducted in South-East Queensland was in 1992/'94. This data is now dated and hence it was decided to conduct a new travel survey. The purpose of this research project is to re-calibrate the trip generation component of the model in light of newly available household travel survey data.

3.2.2 Problem analysis

The existing trip generation equations of the SEQ model are not up to date and based on 1992/'95 *Melbourne* survey data. The model was then tested against the 1992/'94 Household Travel Survey Data for the Brisbane Statistical Division. There are demographical differences between the South East Queensland region and the cities of Brisbane and Melbourne. The SEQ region consists of retired people and tourists highly. Furthermore it is clear that the SEQ region is undergoing fast changes in terms of growth. Currently the model under predicts traffic and the existing equations are manipulated manually to compensate for these drawbacks. So if VLC wants to say something credible about future developments in the region it would be highly implausible if their arguments and predictions are based on the 1992/'95 Melbourne and Brisbane data they currently use for their models. But what drives people to move to this region and what is the purpose of their trips? The newly available household

[1] The four step model is formed by:

1) Trip Generation – how many trips, of what purpose
2) Trip Distribution – where people travel to? By using the gravity model
3) Modal Split – the different modes of transport
4) Assignment

travel survey data should give a better prediction of the factors that influence trip generation in the region.

3.2.3 Objectives

As mentioned before, the overall objectives of this project for VLC are to have an updated version of the existing model in light of the newly available household travel survey data. The importance of this research lies in the fact that VLC wants to be able to give founded and justified recommendations, with the updated model, concerning future developments. With this advanced investment, VLC is able to be ahead of competition when new transport infrastructure projects are being planned. Having this model up to date gives the organisation the opportunity to state plausible travel demand forecasts and get involved in future development plans in early stages.

The objectives of my research are:

1. To gain a full understanding of the scope, structure and reliability of the new household travel survey, for the purposes of developing a four-step travel model
2. To gain a full understanding of VLC's current trip generation techniques, and to gain "hands-on" experience in using these techniques on a real-life project
3. To evaluate the soundness of VLC's trip generation techniques in comparison to other techniques; both theoretically, and empirically.
4. To implement a re-calibrated trip generation model for use in future VLC projects.

Summarizing; the objective of my thesis is to develop a new trip generation model for the South East Queensland region by giving insight in the newly available household travel survey data and by analysing the existing trip generation model in the light of other trip generation techniques which will give VLC the opportunity to be ahead of competition and state more plausible travel demand forecasts for the region.

3.3 Methodology

The execution of the project can be broken down into a number of stages that are described in the following sub-sections.

Familiarisation

First of all I will start with the 'familiarisation' stage. This will involve:

Investigation of the household travel survey – an extensive report including all the household travel survey data and an explanation/procedure manual of the survey is available. A working paper listing relevant questions of interest has been prepared (see research questions section paragraph 3.4). Following investigation of these questions, a short summary document should be produced describing and summarising the key features of the survey.

Introductory reading relating to trip generation – Chapter 4 of the text "Transport Modelling" by Ortuzar et al. should be read as introductory reading to trip generation.

Introduction to VLC's current trip generation techniques – a technical note describing the original calibration of VLC's trip generation model will be provided, and should be read. Following this, a VLC staff member will provide a hands-on demonstration of the currently implemented trip generation model. Finally, I will prepare a short report summarising VLC's currently implemented techniques.

Development of a New Trip Generation Model

When the familiarisation stage is finished the imposed development of the new trip generation model can start. This stage will involve:

Selection of a statistics package – a statistics technique (e.g. regression analysis or variance analysis) will need to be chosen to sort the data and remove inaccuracies. This can be done with R-project, SPSS, SAS or another statistics package.

Data Preparation - The household travel survey data will need to be cleaned, and manipulated into a form that is suitable for input into the statistics package.

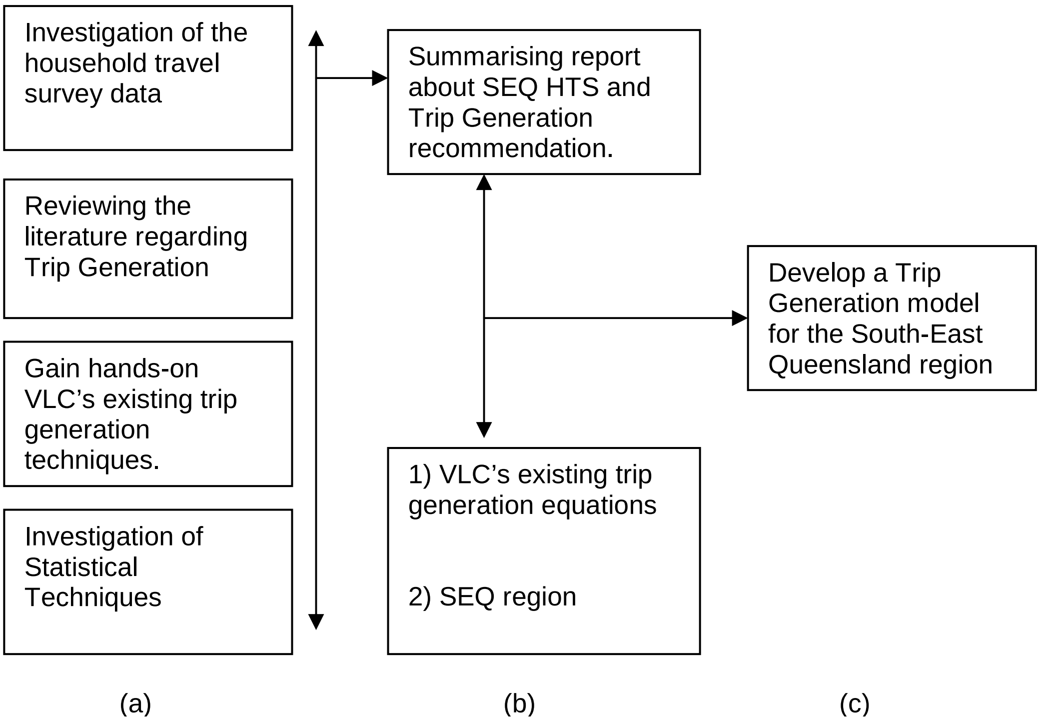
Re-calibration of VLC's existing trip generation equations – As a starting point, it will be useful to re-calibrate the trip generation equations that VLC currently uses. The validation of this new model should then be compared with the validation of the "old" equations using the survey data. Finally, the new equations should be implemented in VLC's model, to investigate how they affect overall model performance against traffic counts.

Calibration of other model forms – following a review of the literature, some attention should be devoted to other forms of trip generation models. Their performance should be compared to the re-calibrated versions of VLC's trip generation equations, with a judgement made about which should be adopted.

Implementation and Presentation
The initial plan was to implement the proposed new trip generation model within VLC's standard model, and consider its implications for overall model calibration. Due to time constraints, as well personal as within the organisation this did not occur.

3.3.1 The research model

To make sure the objectives of my research are going to be achieved it is useful to make a clear overview of the different stages that have to be accomplished to reach the final goal. The research model below is a schematic reproduction of the project and the global stages that are going to be executed.



(a) Investigation of the household travel survey, reviewing the available literature relating to trip generation modelling and getting hands on the existing trip generation techniques should lead to a report containing new insights and recommendations for improvement of the existing model (b). In this report, the factors/parameters that influence trip generation in the SEQ region are presented by which the performance of VLC's existing trip generation techniques can be evaluated. (c) This will finally lead to the development of a new trip generation model for the South-East Queensland region and implementation within VLC's standard model.

3.4 Research Questions

It is clear that my research aims on a practical application with the objective of re-calibrating VLC's existing trip generation equations. The objective asks for prescriptive knowledge that should give a proposal on how the model should be updated. To be able to do this, more knowledge concerning the background of the existing model and the purpose of the project is necessary. This will mainly be explanatory information that comes across by making inquiries. For that reason, the research is split up in central- and sub questions to get better insight.

3.4.1 Central questions

The answers to the central questions should tend to be sufficient to accomplish the overall objective of my research and are formulated as follows:

1. Why is a new trip generation model for the South East Queensland region necessary?
2. What criteria does the survey data need to satisfy to be reliable and useful for trip generation modelling?
3. What factors influence trip generation in the South East Queensland region?
4. To what extent does the existing VLC model meet the required outcomes?
5. To what extent do other trip generation models meet the required outcomes?

3.4.2 Sub questions

The sub questions are mainly meant to support and answer the central questions. The following sub questions are formulated:

1.1. What is VLC's purpose with the new trip generation model?
1.2. On what data is the existing trip generation model of the SEQ region based?

2.1. What purposes of travel are reported? How many trips of each type are there?
2.2. How is the time of the trip recorded? How precise? Is it the start or end of the journey, or are both recorded?
2.3. How are the origin and destination of the trip recorded? District? Suburb? Address? Or perhaps coordinates?
2.4. What household attributes are recorded? Number of people? Number of cars? Etc.
2.5. What individual attributes are recorded? Occupation (Employed/Student/Unemployed) – what categories? What categories of employment?
2.6. What dates are the surveys conducted? Beginning/End, during school holidays?

3.1. What factors affecting trip generation appear from the Household Travel Survey Data?
3.2. What other factors, that did not appear from the survey, should be taken into account?

4.1. What factors affecting trip generation are currently taken into account for the existing model?
4.2. To what extent an update of the existing model is necessary?

5.1. What factors affecting trip generation are taken into account in other trip generation models following a review of the literature?
5.2. How do other trip generation models perform regarding to trip predictions?

4 Dataset Description

This chapter starts with a short description how the survey data has been collected and which trip purposes are accounted for in this research. This is the actual basis and first part of my research and describes the household travel survey data and its overall characteristics. Paragraph 4.1 outlines the necessary data preparation that has been made for the development of the trip generation equations. For a full description of the household travel survey data I would like to refer to Appendix 1.

The Household Travel Survey collected data in the form of travel diaries from approximately 7000 households. Each household filled in the diary for a representative working day, during October – November 2003 or February – March 2004. Holidays have been avoided as good as possible but some private school and university holidays were unavoidably included. The previous 1992 Household Travel Survey had drawn a random sample from residential Energex (Energy Company) billing addresses and used a mail out – mail back survey. This sampling methodology was changed slightly for the 2003/'04 SEQTS to improve a dropping response rate observed in other cities. It was decided to change to a hand delivery – hand collection system, with follow-up telephone calls to encourage the completion of questionnaires. To facilitate this more labour intensive method a two-step process was used when sampling. First, census collection districts (CCD's) were sampled (320 out of 4185), and then households were randomly sampled within each district (at the rate of approximately 6%). Using these techniques the target response rate of 60% was maintained for the Brisbane region (60%) and the Sunshine Coast (62%). The response rate for the Gold Coast reached a reasonable (55%).

The variables pertinent to this research are: the number of trips (by purpose); household structure (blue collar, white collar and undefined workers, dependants by age groups A[0-17], B[18-64], C[65+] and household vehicles. (See following pages for full variable descriptions and histograms of responses). The nested relationship of the trip purpose is shown in figure 2.

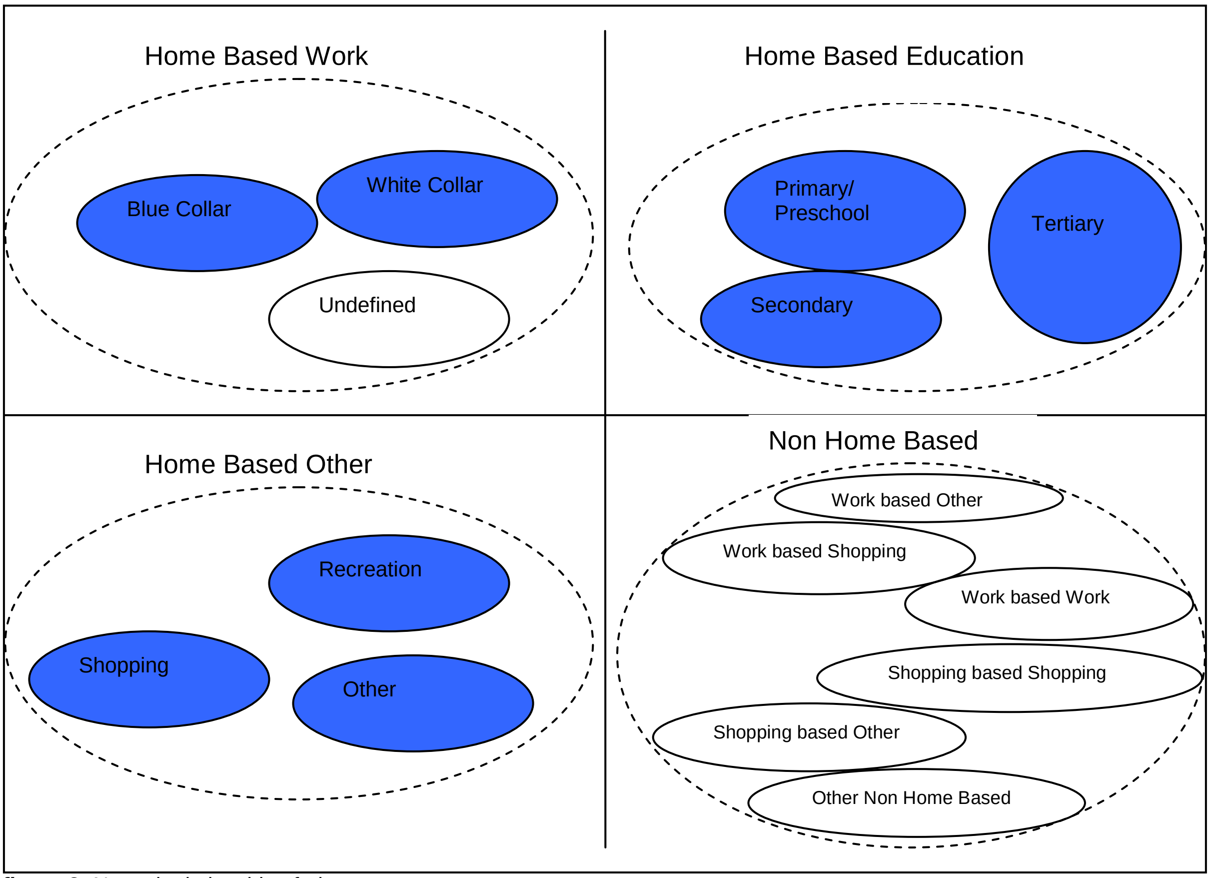


figure 2: Nested relationship of trip purposes

The blue shaded groups are taken into account while the non-home based travel is not considered in this analysis.

4.1 Data Preparation

The data as collected during the survey period and provided in the database can not be used directly for the development of the trip generation equations. The data namely is incomplete or missing in some cases, collected during school holidays, inconsistent or incorrectly classified. Furthermore data is also collected from visitors who normally aren't residents of the area. All these data needs to be filtered and excluded from the survey since they affect the trip generation calculations in a negative way which of course needs to be avoided.

4.1.1 Definition of Population

The entire survey contains a total population of 18.194 people. Visitors are included in the *TotalPersons* per household variable of the data, though they are not counted as household members and so not included in the *HHsize* variable. So *HHsize* is the variable we are interested in for household size (i.e. not including visitors). *PersonIDs* from people who are not residents (i.e. RESIDENT = 2) are filtered out, which leaves 17.783 *PersonIDs* remaining.

As mentioned in table 22, Appendix 1, total number of responding households is 6.978. While preparing the data it became clear that some records were inconsistent in terms that the count of PersonIDs in the household (where RESIDENT = 1) did not match the HHsize. In other words, where the number of people recorded did not match the household size. This situation (after removing the visitors) covers 10 household cases. These households have been removed from the data since it is impossible to discover what the actual values should be. In total these 10 households matched 34 people, which have been removed and left 6.968 *HHIDs* and 17.749 *PersonIDs* remaining.

Further investigation brought another inconsistency to the light. Some people seemed to note *Home_Based_Work(HBW) trips* without specifying their occupation. Effect on the trip generation in this case is that it is impossible to specify the trip as a *Blue* or *White* collar worker trip. Households that recorded unspecified HBW trips in this case where also entirely ignored and excluded from the data. In total 668 trips in 216 households were of this type of unspecified HBW trips, 216 *HHIDs* and all matching *PersonIDs* (i.e. 274) have been removed, remaining 6.752 *HHID* and 17.475 *PersonIDs.*

Households who were interviewed during school holidays / tertiary holidays actually need to be filtered out. From the 2003 and 2004 school calendar it became clear that the survey never was executed during public primary and secondary school holidays. But private school holidays and tertiary holidays tend to differ in some way from the public school holidays. This has actually not been taking into account since this would only be a small portion and it is hardly impossible to find out all private school holidays. So the final sample size contains 6752 households.

4.1.2 Definition of a Trip

The Access database of the survey contains a "SEQTS_TRIPS" table in which each individual person trip is recorded. If a trip involves interchanging then each "leg" will be recorded in the "SEQTS_STOPS" table, but the trip will only be recorded once in the trips table. So, the definition of a trip used for trip generation is a row in the trips table.
For each trip a "Start Purpose", "End Purpose" and "Overall Purpose" are recorded. After the first regression though, it looked like the classification as specified in the survey was not done properly or at least not as expected by VLC. The main problem noticed after regression in comparison with the existing VLC model, arose in a big difference between observed and estimated "Home Based Education"(HBE), "Home Based Other"(HBO) and "Home Based Recreation" (HBR)-trips. From this we could read that the total number of observed HBE_PrePrim(Pre- & Primary school)-trips were 2,65 times the model estimated. For HBE_Sec(Secondary) the observed trips were twice as much as estimated. And for the HBR applies they were underpredicted by a factor 1,6. On the other hand HBO-trips though were highly over predicted (almost 4 times as many as observed). This seeming shift from trips to different purposes was a reason to have a closer look at the different classifications of the Home Based-trips whereas especially *accompanying* and *serving passenger* trips where differently specified than expected. table 1 shows these trips and their specified *overall purpose* as given in the survey. The last column in the table shows the desired VLC purpose, which in each case is HBO for these specific trips. Reason for the proposed changes lies in the fact that even though someone's trip might have the specified destination (e.g. education, shopping, social or work) doesn't mean their

individual trip purpose is of the specified classification. All incorrectly specified trips are changed and recoded into the desired VLC overall purpose.

Start Purpose	End Purpose	Overall Purpose	VLC Overall Purpose
Home	Accompany EducationPrim	HBE - Pre Primary	Home Based Other
Home	Serve Pass. EducationPrim	HBE - Pre Primary	Home Based Other
Accompany EducationPrim	Home	HBE - Pre Primary	Home Based Other
Serve Pass. EducationPrim	Home	HBE - Pre Primary	Home Based Other
Home	Accompany EducationSec	HBE - Secondary	Home Based Other
Home	Serve Pass. EducationSec	HBE - Secondary	Home Based Other
Accompany EducationSec	Home	HBE - Secondary	Home Based Other
Serve Pass. EducationSec	Home	HBE - Secondary	Home Based Other
Home	Accompany EducationTert	HBE - Tertiary	Home Based Other
Home	Serve Pass. EducationTert	HBE - Tertiary	Home Based Other
Accompany EducationTert	Home	HBE - Tertiary	Home Based Other
Serve Pass. EducationTert	Home	HBE - Tertiary	Home Based Other
Home	Serve Pass. Pers. Bus/Welfare	Home Based Shopping	Home Based Other
Home	Serve Pass. Shopping	Home Based Shopping	Home Based Other
Serve Pass. Pers. Bus/Welfare	Home	Home Based Shopping	Home Based Other
Serve Pass. Shopping	Home	Home Based Shopping	Home Based Other
Home	Serve Pass. Social/Ent/Recreation	Home Bases Social	Home Based Other
Serve Pass. Social/Ent/Recreation	Home	Home Bases Social	Home Based Other
Home	Serve Pass. Work Related Bus.	Home Based Work	Home Based Other
Serve Pass. Work Related Bus.	Home	Home Based Work	Home Based Other

table 1: Reclassification of trip purposes

4.1.3 Table production

The structure of VLC's trip production model is a series of additive linear relationships (one for each trip purpose) that describe the number of person trips made by households of varying characteristics on a typical weekday where there are seven stratified attribute groups (SAG's) that in itself are a linear function as well. How these so called "dummy variables" exactly work is explained Appendix 2. First step in gaining the desired data is producing a table from the Access database that describes the household structure of each of the remaining 6.752 households in terms of the specified attribute groups. table 2 is an example of the produced table.

HHID	HHSIZE	CARS	BLUE	WHITE	Dep_0_17	Dep_18_64	Dep_65_Plus
Y03H010101	2	1	1	1	0	0	0
Y03H010102	4	1	0	0	0	4	0
Y03H010103	4	1	0	2	2	0	0
Y03H010104	1	0	0	1	0	0	0

table 2: Example HH_structure

This table tells us something about the household size (HHSIZE), #cars (CARS), #blue collar workers (BLUE), #white collar workers, dependants aged 0-17 (Dep0_17), dependants aged 18-64 (Dep18_64) and dependants aged 65+ (Dep_65_plus), per household.
Next step is to get information about the observed trip counts per households, per purpose. table 3 tells something about the total number of home based trips that have been produced by each household. Home based trips can be split up in eight different purposes – i.e. Education Pre/Primary, Education Secondary, Education Tertiary, Other, Shopping, Recreational, Blue Collar and White Collar trips.

HHID	HBE_PP	HBE_SEC	HBE_TER	HBO	HBS	HBR	HBW_BL	HBW_WH
Y03H010101	0	0	0	0	8	2	2	0
Y03H010102	0	0	4	0	5	1	0	0
Y03H010103	0	3	0	0	4	1	0	2
Y03H010104	0	0	0	0	0	0	0	2

table 3: Example Home Based Trip count per HHID per Purpose Category

From the tables above, a table of all households with a field for each stratified dummy variable and the number of trips, looking like table 4 below, is produced for each trip purpose. This set of 14 excel sheets containing 6752 rows (one for each household) and 32 columns (HHID, one for each dummy variable and trip count) are created by use of programming language "'*Ruby".*

HHID	HHSIZE_1	HHSIZE_2	HHSIZE_3	HHSIZE_4	HHSIZE_5	HHSIZE_6+	**HBE_PP_Trips**
Y03H010101	0	1	0	0	0	0	**0**
	BLUE_0	BLUE_1	BLUE_2	BLUE_3+			
	0	1	0	0			
	WHITE_0	WHITE_1	WHITE_2	WHITE_3+			
	0	1	0	0			
	Dep0_17_0	Dep0_17_1	Dep0_17_2	Dep0_17_3	Dep0_17_4+		
	1	0	0	0	0		
	Dep18_64_0	Dep18_64_1	Dep18_64_2	Dep18_64_3+			
	1	0	0	0			
	Dep_65+_0	Dep_65+_1	Dep_65+_2+				
	1	0	0				
	CARS_0	CARS_1	CARS_2	CARS_3+			
	0	1	0	0			

table 4: Example stratified dummy variable table

From the examples above we can read that household Y03H010101 has a household size of two, containing one blue collar worker, one white collar worker, zero dependants, one car and doesn't make any HBE_PP_Trips. This sounds quite logical since this household does not have any dependant household members in the age group of 0 – 17 who will basically make trips to pre- and primary school. Now we have the household structure and the observed number of trips per purpose of each household in the survey.

4.2 Characteristics of the home interview sample – household structure

The broad structure of the South East Queensland household sample is presented in *table* 5 with respect to the worker and the dependants attribute of households. Workers are considered as the sum of white collar and blue collar workers (see paragraph 5.3.1 for an explanation of different worker types). Dependants are presented as the sum of the co-occupying household members in the age groups of 0-17, 18-64 and 65+. For example, there are 319 households with 1 worker and 2 dependant co-occupiers in the entire survey. A more comprehensive breakdown of the household structure of the sample is presented in *table* 6, where the data construct adopted for *table* 5 is layered bye household car ownership.

Broad structure of SEQ_HHTS						
Workers	**Dependants**					**Total**
	0	**1**	**2**	**3**	**4 +**	
0	0	823	985	119	87	2014
1	641	625	319	283	172	2040
2	1020	445	544	199	63	2271
3+	223	127	54	18	5	427
Total	1884	2020	1902	619	327	6752

table 5: Broad structure of SEQ_HHTS

Furthermore the following histograms in figure 3 provide a better understanding of the household structure in the SEQ travel survey data.

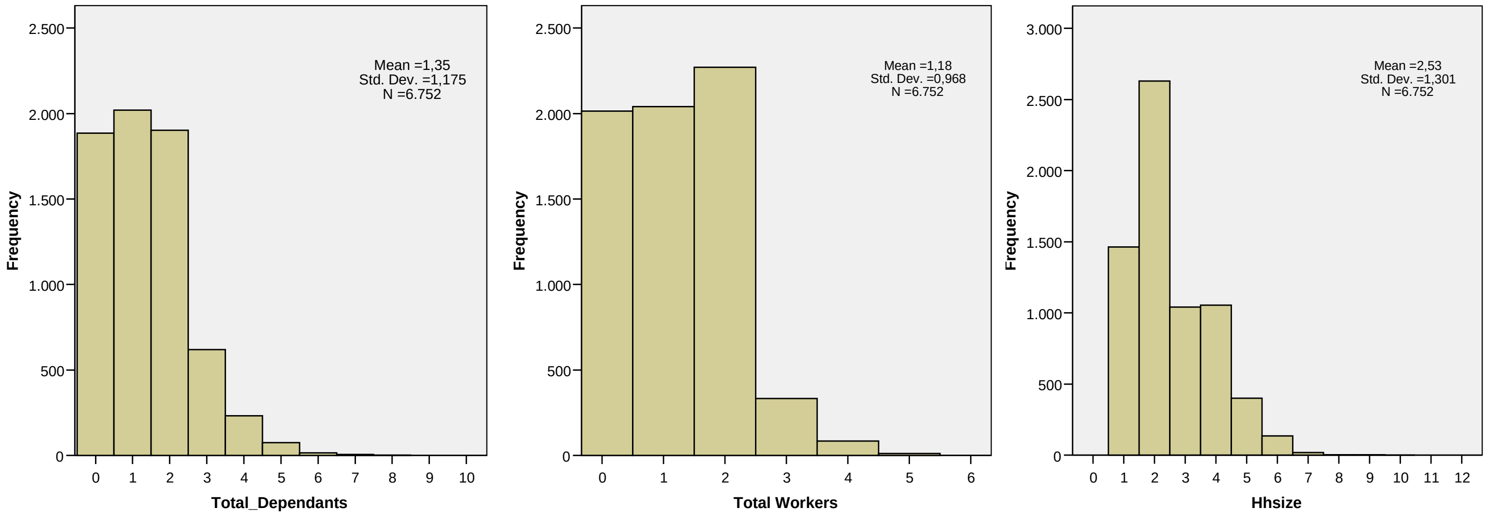


figure 3: Broad structure of SEQ Households

An appreciation of the structure of the household sample is important when considering the adequacy of the model in replicating the behaviour of various market segments. Interesting points to note from table 5, table 6 and the corresponding figures are:

Referring to *table* 5 and *figure 3* it can be seen that the average household size in the survey approximately is 2,53 persons per household, which sounds realistic and doesn't impose a closer look. The most common household contains two workers and no dependants followed closely by households containing no workers and two dependants instead (15,1 over 14,6 percent). Furthermore it is interesting to note there are about an equal number of zero worker-, one worker- and two worker households (28,9 percent, 30,2 percent and 33,6 percent respectively). The fact there are that much zero worker households sounds somewhat conspicuous so a closer look to this group tells there are:

- 280 zero worker households with dependants in the category 65+, zero cars
- 842 zero worker households with dependants in the category 65+, one car
- 213 zero worker households with dependants in the category 65+, two cars
- 23 zero worker households with dependants in the category 65+, three+ cars

This means that of all 2014 zero worker households, a total of 1358 are actually households with merely retirees which is 67,4 percent respectively.

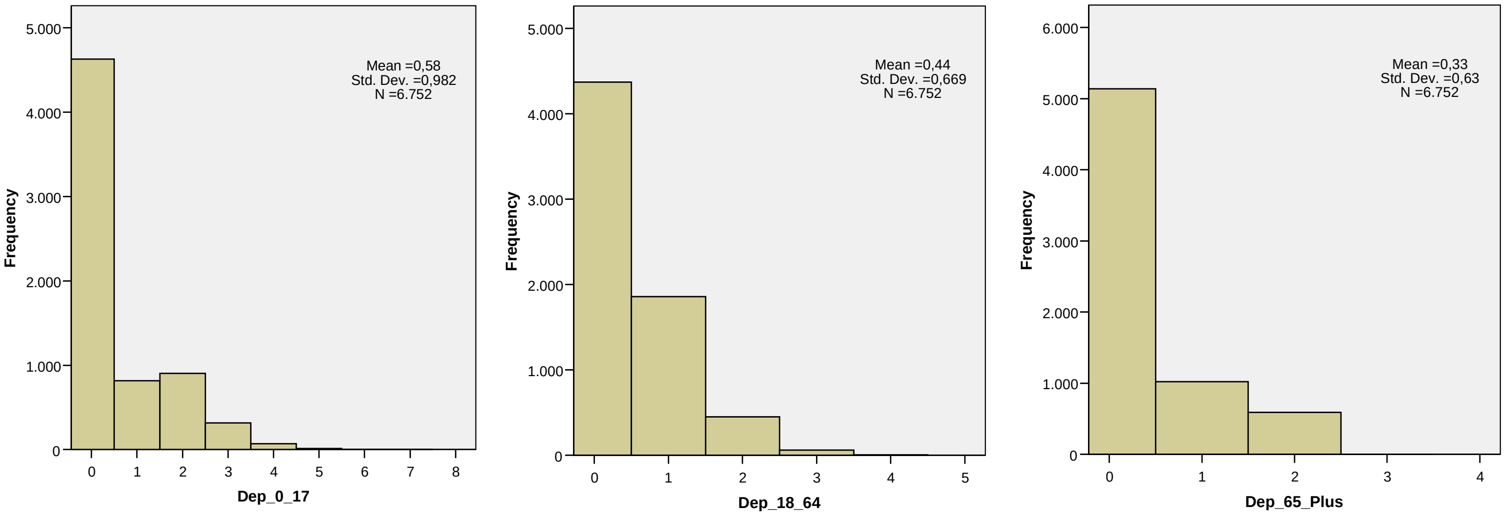
The more comprehensive breakdown of the household structure, layered by car ownership, is presented in the following table 6:

Structure of SEQ_HHTS – Workers – Dependants - Cars						
car ownership = 0						
Workers	Dependants					Total
	0	1	2	3	4 +	
0	0	278	85	15	6	384
1	66	25	11	2	3	107
2	10	4	2	1	0	17
3+	2	0	0	0	0	2
Total	78	307	98	18	9	510
car ownership = 1						
Workers	Dependants					Total
	0	1	2	3	4 +	
0	0	506	624	63	49	1242
1	499	284	136	70	36	1025
2	235	69	72	18	7	401
3+	16	7	2	0	0	25
Total	750	866	834	151	92	2693
car ownership = 2						
Workers	Dependants					Total
	0	1	2	3	4 +	
0	0	31	254	36	26	347
1	66	277	136	178	108	765
2	686	251	364	133	33	1467
3+	54	32	15	3	0	104
Total	806	591	769	350	167	2683
car ownership = 3+						
Workers	Dependants					
	0	1	2	3	4 +	Total
0	0	8	22	5	6	41
1	10	39	36	33	25	143
2	89	121	106	47	23	386
3+	151	88	37	15	5	296
Total	250	256	201	100	59	866

table 6: Structure of SEQ_HHTS – Workers – Dependants – Cars

From the table above it is interesting to note that only 510 out of 6752 households (7,6 percent) do not have access to a motor vehicle and the number of households with access to three motor vehicles or more even exceeds the number of non-car owning households. Furthermore the number of two car owning households can almost be said to be equal to the number of one car owning households.

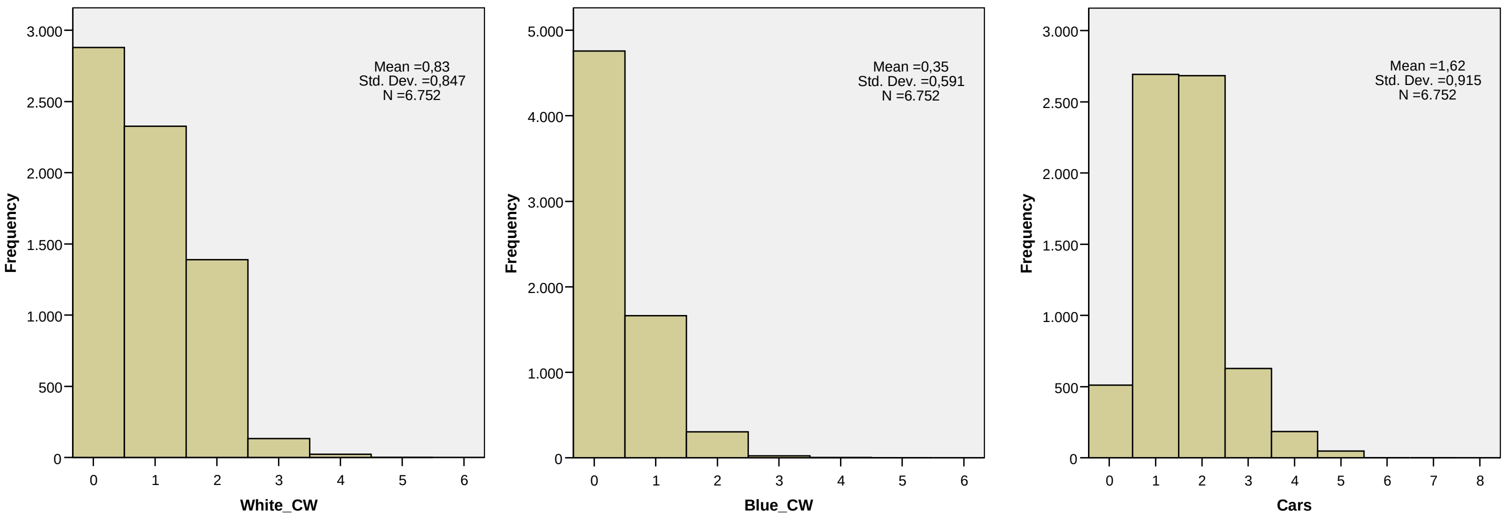
Further information about the household structure can be found in figure 4 where histograms and frequency tables of the number of dependants, blue- and white collar workers and number of cars are presented. Information about total trips per purpose is presented in the histograms of figure 5.



Dep_0_17	Frequency (HH)	Percent
0	4631	68,6
1	818	12,1
2	903	13,4
3	315	4,7
4	70	1,0
5	13	,2
6	1	,0
7	1	,0
Total	6752	100,0

Dep_18_64	Frequency (HH)	Percent
0	4372	64,8
1	1860	27,5
2	451	6,7
3	62	,9
4	7	,1
Total	6752	100,0

Dep_65_Plus	Frequency (HH)	Percent
0	5138	76,1
1	1022	15,1
2	589	8,7
3	3	,0
Total	6752	100,0



White_CW	Frequency (HH)	Percent
0	2878	42,6
1	2327	34,5
2	1389	20,6
3	133	2,0
4	23	,3
5	2	,0
Total	6752	100,0

Blue_CW	Frequency (HH)	Percent
0	4756	70,4
1	1664	24,6
2	303	4,5
3	24	,4
4	4	,1
5	1	,0
Total	6752	100,0

Cars	Frequency (HH)	Percent
0	510	7,6
1	2693	39,9
2	2683	39,7
3	628	9,3
4	186	2,8
5	48	,7
6	2	,0
7	2	,0
Total	6752	100,0

figure 4: Histograms of the number of households by person categories and cars

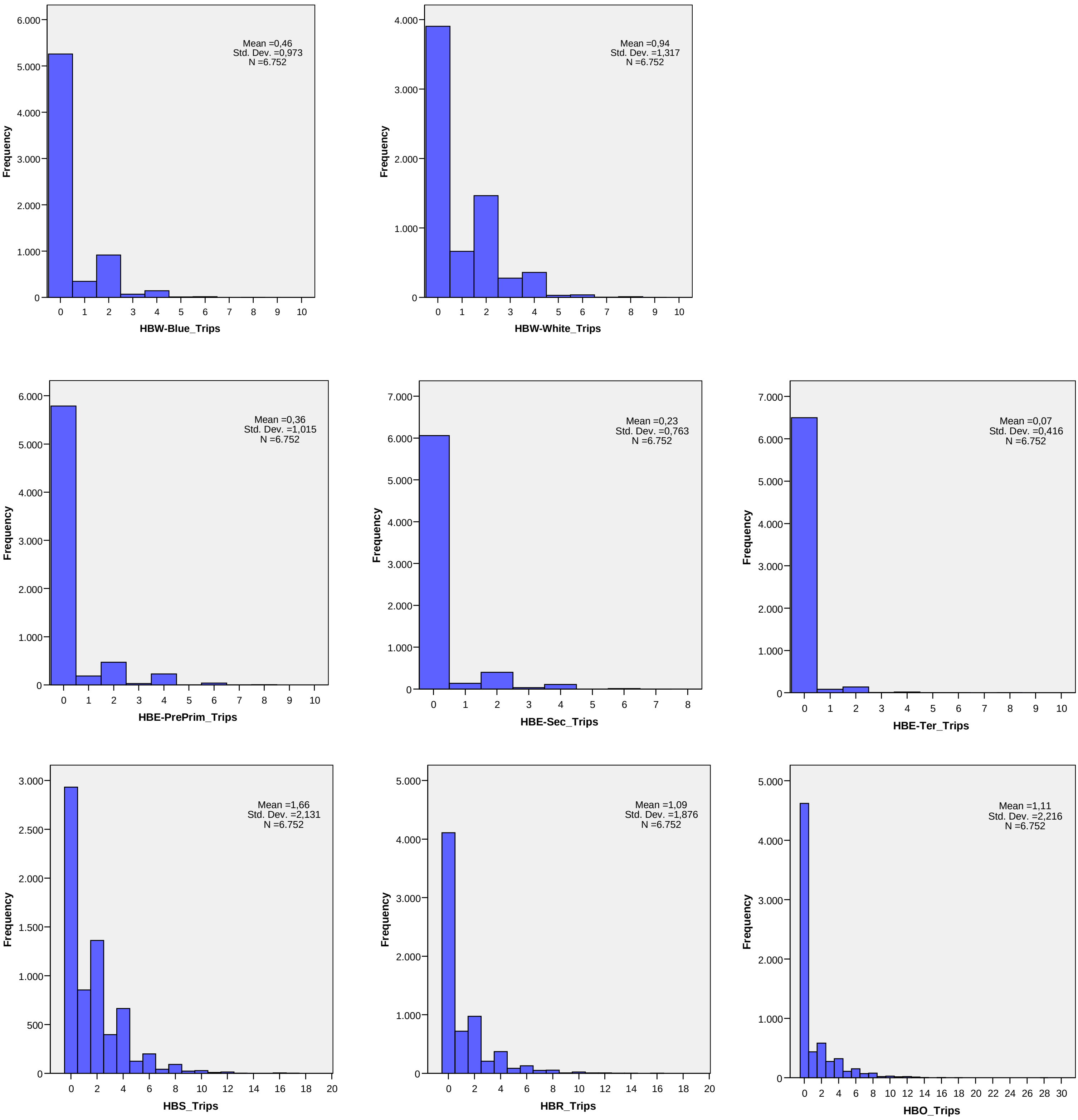


figure 5: Histograms of number of trips of all trip purposes

This research is concerned with models of event counts. An event count is the realization of a nonnegative integer-valued random variable and refers to the number of times an event occurs. In this case, the count data concerns the number of observed trips per purpose. From the histograms above we can clearly recognize there are a large number of zero trips for each trip purpose. These excess zero’s may cause trouble when we are trying to predict trips with a simple linear model. Since the data is heavy in zero's a linear model will be forced to predict more values near to zero probably better than it would predict non-zero's. A closer look linear models is discussed later on in this research.

4.3 Summary

Summarizing, we now have all the required data in its desired structure to actually start with the development of the new trip generation equations. It became clear that it takes some time to gain hands on all available data and filter out all the inconsistencies and things you do not want to use. Considering what is pointed out in this chapter we are now able to answer what VLC's purpose with the new trip generation model is. Furthermore it is clear that the existing trip generation model of the SEQ region is based on dated 1992 Melbourne data. Furthermore an overview of the different trip purposes and attribute groups is given.

5 Trip Generation Literature Review

This chapter is the next part of the research and is basically considered reviewing the literature regarding trip generation. The chapter starts with an introduction to the first phase of the traditional four step model and attention is paid to the manner VLC classifies their model. Paragraph 5.5 and 5.6 describe the most common used approaches in trip generation: category analysis and regression analysis.

5.1 Introduction

Trip generation modelling corresponds to the first stage of the conventional four step transport modelling theory, where the other stages are trip distribution, mode choice and traffic assignment to the transportation network. The purpose of the trip generation phase (e.g. production and attraction phase) in the 'classical transport model' is to predict the total person trips produced by and attracted to each zone in the study area based on social-economic information. The so called 'classic transport model' is a result of years of experimentation and development. As mentioned, the model consists of four stages and is depicted in figure 6 below:

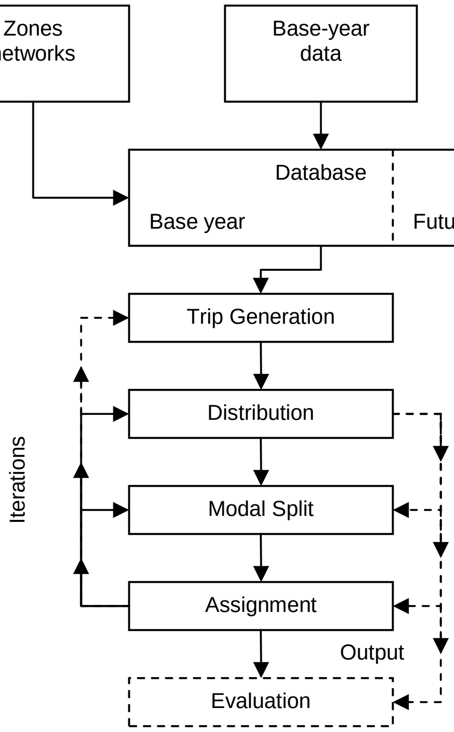


figure 6: Traditional 4-step Model

The approach starts by considering a zoning and network system, and the collection and coding of planning, calibration and validation data. These data would include base-year levels for population of different types in each zone of the study area as well as levels of economic activity including employment, shopping space, educational and recreational facilities. These data are then used to estimate a model of the total number of trips generated and attracted by each zone of the study area (trip generation) [Ortúzar and Willumsen., 2004].

This research project is focused on the development of trip generation equations with which a prediction of the total number of trips generated by origin (O_i) and attracted by destination (D_j) can be made. The focus of this research lies on the production stage of trip generation though. These models trip generation models can be further divided as home based, non-home based or work based trips. For the development of a good production-/attraction model it appears to be useful to make a classification of the trips by purpose, time of the day and person type. This will be discussed in

paragraph 5.3. Subsequently, factors affecting trip generation and attraction are discussed and attention is paid to modeling approaches for production and attraction of trips. Most commonly used approaches in trip generation are regression-analysis and category-analysis which will be discussed in paragraph 5.5 and 5.6.
Trip generation models can operate at different scales: zonal, household or person based. As well as the spatial scale, specific predictor variables also vary across regions and countries. Common predictors for production models are given in table 7.

Variable	Type	Sub-split
Income	Continuous or ordinal	
Dwelling type	Nominal	
Number of cars	Ordinal	
Household size	Continuous or ordinal	
Number of workers	Ordinal	White / blue collar
Number of dependents	Ordinal	Split by age
Racial/ethnic background	Nominal	

table 7: Common predictors in trip production models

5.2 Zones and networks

5.2.1 Background

A transport model is related to a certain study area. Displacements in this study area can start and finish on each address and travellers can make use of all kinds of transport. It is however impossible to collect and analyse data on the basis of individual data. A schematisation of reality is necessary. This schematisation includes the following components:

- Area classification; the study area is classified into a number of zones. We study the displacements of and to each of these zones. All displacements are considered to start and finish in an imaginary point within this area, the centroid.
- Networks; the transport system exists from a number of networks which represent the available transport modalities.

Significant parameters are the number of applied zones and their dimension. Within the model the centroids are basically connected by *connectors* that represent the roads and their characteristics. Zones with a population from 1000 to 3000 people tend to satisfy well. Regarding the number of zones within a typical regional transport model, 300 to 500 zones is said to be reasonable. More than 1000 zones can be classified as extensive. All zones should be approximately equal in terms of dimension concerning traffic production. Furthermore one should aim for homogeneity towards land use within each zone [Immers & Stada, 1998]. The transport system is represented by a network of junctions and links. Characteristics of the network are ascribed to the links, for instance; length, speed and capacity. The junctions are not specially qualified. Networks tend to be more and more represented as multimodal network models where different ways of transport are connected mutually instead of separate networks for different modalities. A typical regional transport model consists of 1000 to 5000 junctions; over 10.000 junctions can be classified as extensive [Immers & Stada, 1998].

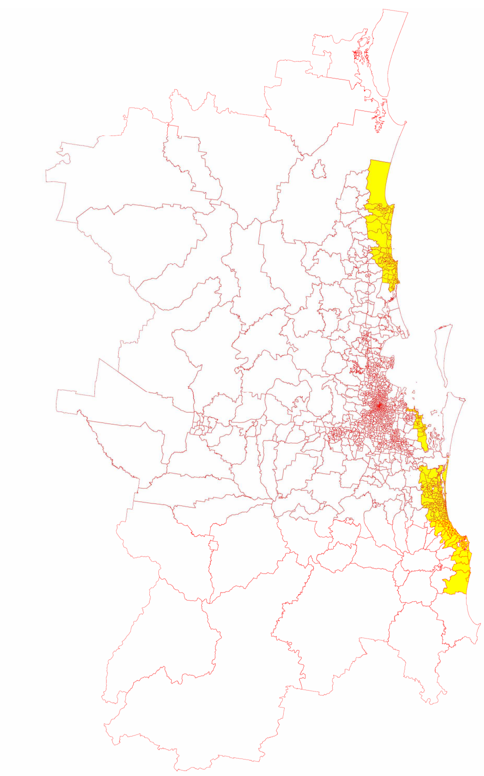


figure 7: VLC's SEQ zoning

5.2.2 VLC's SEQ Transport model

The household survey of the SEQ region consists of approximately 4200 Census Collection Districts (CCD) with approximately 225 households (which equals +/- 560 person) per CCD. It is not usually desirable to model the area with such an extensive number of zones since it will influence the complexity and dimension of the model. Based on geographical and demographical data the existing modelled area is brought to 1529 zones with a total number of 1.101.164 households and a total population of 2.787.092 people. So each zone approximately contains 720 households and 1823 people. figure 7 depicts the zoning of the South East Queensland region as currently modeled by VLC which is pretty extensive. VLC's zoning system also includes a significant "influence area", also known as a buffer region. This area is included to better estimate travel patterns within the core study area, and minimize the "boundary effects". For detailed modeling, VLC also utilizes a 5448 zone model, where each travel zone represents a CCD or smaller. In terms of transport infrastructure, VLC's model of SEQ consists of approximately 20,000 junctions, and over 50,000 one-way link sections.

5.3 Classification

Classification of trips according to Ortúzar and Willumsun is done by: trip purpose, time of day and by person type.

5.3.1 Trip purpose

Ortúzar and Willumsun identify different trip purposes and model them separately to obtain better trip generation, trip distribution and mode choice models. In the case of home based trips, five categories are distinguished:

- Trips to work
- Trips to school or college (education trips);
- Shopping trips;
- Social and recreational trips;
- Other trips;

The first two are usually called compulsory trips and all the others are often optional trips. In the past, non-home based trips were often not separated because they only amounted to a small percentage (15% - 20%) of all trips [Ortúzar, 2004]. However, VLC research showed that the scale of non-home based travel is rapidly increasing, and as such, they will be considered as part of their trip generation model.

The "*QT_Overall_Purpose_DMR"* column in the SEQTS_TRIPS table of the SEQTS Access database is a combination of the categories "*QT End"* and "*QT Start"* purposes to create a single purpose. Result is the following table 8, defining the possible overall trip purposes per person in the Household Travel Survery:

QT_OVERALL_PURPOSE_DMR	
Home Based Work	Shopping Based Shopping
Home Based Education - Pre Primary	Shopping Based Other
Home Based Education - Secondary	Non Home/Work/Shopping Based
Home Based Education - Tertiary	Work Based Other
Home Based Other	Work Based Shopping
Home Based Shopping	Work Based Work
Home Based Social	

table 8: Overall trip purposes as in the SEQ household survey data

Veitch Lister Consulting distinguishes home based trip purpose categories and non-home based trip purpose categories as in table 9:

VLC's Trip Purpose Categories	
Home-Based Work: Blue Collar (HBW-Blue);	Shopping-Based Personal Business and Shopping (SBS);
Home-Based Work: White Collar (HBW-White);	Shopping-Based Other (SBO);
Home-Based Education: Pre and Primary School (HBE-PPrim);	Other Non-Home-Based Trips (ONHB);
Home-Based Education: Secondary School (HBE-Sec);	Work-Based Other (WBO);
Home-Based Education: Tertiary Institution (HBE-Ter);	Work-Based Shopping/Personal Business (WBS);
Home-Based Shopping/Personal Business (HBS);	Work-Based Work (WBW);
Home-Based Social/Recreational (HBR);	
Home-Based Other (HBO);	

table 9: VLC's Trip purpose categories

Else than Ortúzar states, VLC does take into account the non-home based trips separately, since it is presumed that the amount these trips is growing and currently attributes more than 15% – 20% of all trips. Furthermore we can read from the tables above that the available data as collected in the Household Travel Survey almost entirely corresponds with the VLC's standard trip purpose categories. The main difference can be found in the home-based work purpose that is divided in blue collar and white collar work based trips, by VLC. Blue and white collar workers have been defined in terms of the following occupation classifications:

Blue Collar Occupations

- Tradespersons;
- Plant and Machinery Operators;
- Labourers and Administration.

White Collar Occupations

- Managerial and Administration;
- Professional;
- Para-Professional;
- Clerical;
- Sales and Personal Services.

White Collar workers tend to be situated at an office mainly and exhibit different travel behavior to blue collar workers. Using this distinction means the available data has to be translated to VLC standards. This is done by reviewing the occupation of each household member within the travel survey data and adding an extra value(definition) to each member. The following occupations in table 10 appear from the survey data and are translated to the VLC definitions:

CODE_DESCRIPTION	VLC_DEFINITION
N/A	-
Missing	-
Managers and Administrators	White
Professionals	White
Associate professionals	White
Tradespersons	Blue
Advanced Clerical and Service Workers	White
Intermediate Clerical, Sales and Service Workers	White
Intermediate Production and Transport Workers	Blue
Elementary Clerical, Sales and Service Workers	White
Labourers	Blue

table 10: SEQ Occupation descriptions specified to VLC definitions

If occupation data of a household member is N/A (not available) or missing these households were excluded entirely. As mentioned in paragraph 4.1.1 this was the case for 216 households.

5.3.2 By Time of Day

Trips are often classified into peak and off-peak period trips; the proportion of journeys by different purposes usually varies greatly within time of day [Ortzúzar, 2004]. For this research we won't classify trip generation by different times of the day and just take the daily data. This will actually be divided after the distribution phase in the four-step model.

5.3.3 By Person Type

This is another important classification, as individual travel behaviour is heavily dependant on socioeconomic attributes. The following categories are usually employed:

- *Income level*
- *Car ownership (typically three strata: 0, 1 and 2 or more cars);*
- *Household size and structure (e.g. six strata in most British studies);*

It is important to note that the total number strata can increase very rapidly and this may have strong implications in terms of data requirements, model calibration and use, particularly when using category analysis [Ortuzar, 2004]. In terms of data preparation for trip production, VLC does not specifically classifies trips by person type.

5.4 Factors affecting Trip Generation

Trip generation is on the one hand affected by production factors and on the other hand by attraction factors. Production factors are often based on household demographics and other socio economic factors whereas attraction factors tell us something about land use.

5.4.1 Factors affecting production

The following factors have been proposed for consideration in many practical studies:

- *Income;*
- *Car ownership;*
- *Household structure;*
- *Family size;*
- *Value of land;*
- *Residential density;*
- *Accessibility;*

The first four have been considered in several trip generation studies, while value of land and residential density are typical of zonal studies. Accessibility has rarely been used although most studies have attempted to include it [Ortuzar, 2004]. Ignoring the accessibility factor in the model though, means production and attraction of a zone are insensible for changes in the transport system. This is a basic shortcoming of the existing model. Since accessibility definitely is an important factor for determining productions and attractions several models have been developed but there isn't much accordance regarding the accuracy. Reason lays in the difficulty of quantifying the concept of accessibility [Immers & Stada, 1998]. In case of VLC's trip production model there are seven "stratified household attribute groups" (see Appendix 2 for further detail) considered as production factors, defined as follows:

- household size (1,2,3,4,5,6+)
- car ownership (0,1,2,3+)
- resident blue collar workers (0,1,2,3+)
- resident white collar workers (0,1,2,3+)
- resident (0-17); dependants (0,1,2,3,4+)
- resident (18-64); dependants (0,1,2,3+)
- resident (65+); dependants (0,1,2+)

VLC classifies the "dependants" in aforementioned attribute groups since it is assumed that the different dependant groups cause different affects on trip generation. For example, the dependants aged 0-17 will mostly generate home based - education trips. As can be seen, income level does not play a direct roll within VLC's current trip production model. This is due to the fact that income level and car ownership tend to be highly correlated, and VLC research has previously shown car ownership to be the more important variable.

5.4.2 Factors affecting attraction

The following factors have been proposed to influence the attraction of a zone:

- *Employment*
- *Land use*
    - *Industrial*
    - *Education*
    - *Commercial*
    - *Service industry (e.g. hospitals, banks, public institutions...)*
    - *Recreation (e.g. sport accommodations, sights and attractions...)*
    - *Storing and transfer (harbours, airports…)*
- *Accessibility*

This research is basically focused on factors affecting production rather than the factors affecting attraction. So no further attention is paid to these factors at this stage and deserves further research in future modelling phases.

5.5 Category analysis

Category analysis (also called cross-classification) requires each variable to be split into a number of categories, crossed with the categories of other variables, and a mean trip value calculated based on the number of observations that fall into each cell. That is, the number of trips divided by the number of households. In two dimensions this is easy to visualise as in table 11 for example where a cross classification by workers and dependants is presented.

Trips for designated purpose by workers and dependants					
Workers	Dependants				
	0	1	2	3	4+
0	trips/households	…	…	…	…
1	…				
2					
3+					

table 11: Cross-classification by workers and dependants

With three or more variables, repeated tables are needed to display classifications by the third variable (e.g. a table for each number of cars for instance).

Category analysis is basically called non-parametric (or distribution free), as it does not assume any distributional form for each of the variables. Additionally it is independent of the zone system of a region, interactions between variables are accounted for automatically and it is not possible to extrapolate beyond the original data [Stopher and McDonald, 1983]. Whether these are advantages or not depends on the context and demands of the model. Stopher and McDonald also give some clear disadvantages to category analysis:

1. There is no goodness of fit measure
2. Cell values vary in reliability (depending on number of observations in each category), and conversely, least reliable cells are likely to fall in the extremes of each variable and may also be most critical.
3. There are no systematic procedures to determine best categorisations/groupings within each variable. Or for determining the best predictors or trip rates.
4. There is no information on variances within each cell.

Some methods have been proposed to improve classification analysis in its raw form. One such method is the Multiple Classification Analysis (MCA) [Stopher and McDonald, 1983]. MCA is based on analysis of variance methods (ANOVA) and so provides a framework for selecting variables, classes and interactions using formal goodness of fit measures.

More recently, a probabilistic approach has been proposed [Rengaraju and Satyakumar, 1995]. This technique does not treat cells as independent, and so does not simply count up the observations in each category. Rather, the probability for any one cell is considered to be dependent on those around it using concepts of conditional probability. As it is not always possible to calculate conditional probabilities, multiple regression analyses are used to derive some values. In this approach even cells which have no data can have a probability assigned based on area wide effects. This means that higher order category analyses are then feasible as zero-count cells do not have to be avoided.

5.6 Regression analysis

Ordinary least squares regression is the most commonly used method for calculating productions and attractions, which is easy to visualise as finding the line of best fit through a scatter plot, see figure 8. With linear regression we try to predict the variable *Y* as a linear function of one or more explanatory variables X_i.

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + ...$$

Y is the variable that has to be predicted and is a so called dependant variable. The explanatory variables X_i are the independent variables. Coefficient *a* is a constant factor and the coefficient b_i are the regression coefficients. By using regression in the development of production and attraction models, the X_i variables normally form the social-economic factors as mentioned in paragraph 5.4.1 such as car ownership, household structure, income etc. The dependant variable *Y* represents the number of produced or attracted trips commonly classified by purpose. The constant factor *a* and the regression coefficients b_l are estimated with the help of social-economic data of a base year [Immers & Stada, 1998]. figure 8 below displays HBE Pre/Primary trips (dependant variable) as a function of total dependants in the household (the independent variable).

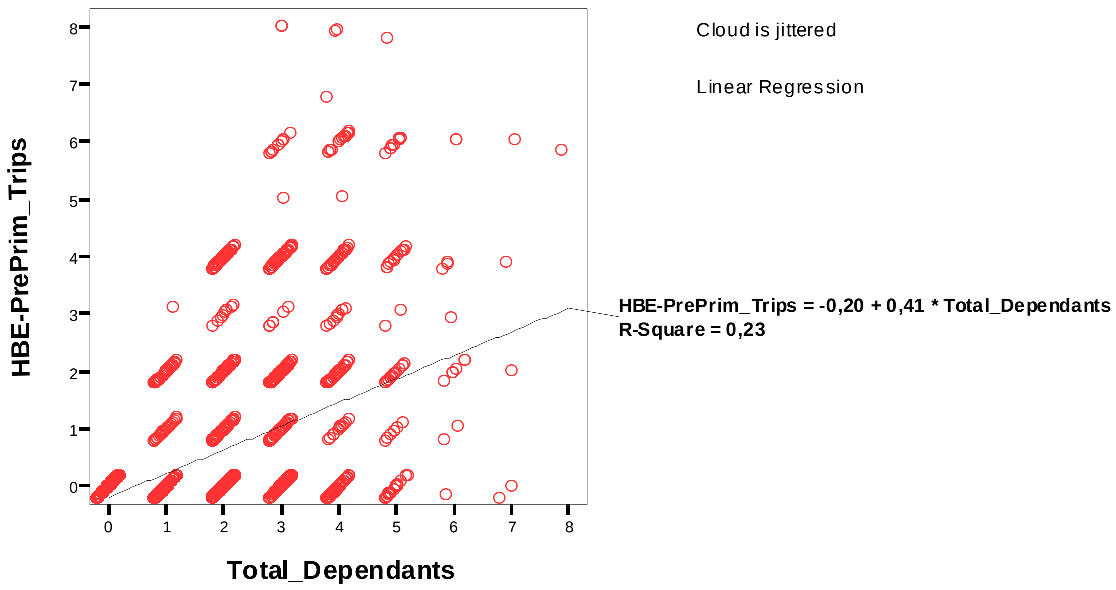


figure 8: Ordinary least squares regression

There are generally two types of errors in linear regression analyses, both of which cause problems in modelling trip generation. Firstly, there is a lack of theory to guide choices in which variables are important predictors of trip rates and models may be incorrectly specified [Washington, 2000]. Typically data dredging (practice in which large volumes of data are analyzed seeking any possible relationships between data) procedures, such as a step-wise regression, are employed to choose between competing models. However it may still be difficult to determine the best stratification of variables, and which interactions are important. Washington (2000) describes a method combining tree based regression with linear regression for a more effective means of specifying the best model. Washington proposes to fit an ordinary least-square regression model using relationships that are known to be important, and then create a tree-based regression using the residuals of the model as the dependent variable to identify relationships not already captured in the model. Dummy variables can be created to represent these relationships and then regression is repeated, and so on.

In the current context there is the larger problem of not meeting the linear modelling assumptions. When these key assumptions are not met, not only may it render a simple regression analysis invalid, but the analysis may be simply unable to detect relationships between variables even when they exist. These assumptions are:

1. that the range of dependent variables is unrestricted;
2. the dependant and independent variables are linearly related;
3. and that errors are normally distributed, uncorrelated with constant variance (that is, homoskeadastic).

Clearly in trip generation models the first condition is never strictly met, as the number of trips cannot be negative. However, if the distribution of trips is not concentrated near zero this may not cause too much of a problem, although a negative number of trips could still be predicted. In the regression plot above for example, the intercept is actually negative so there is a negative number of trips predicted for households without dependants. Simple inspection of scatterplots is a common non-statistical method of determining if nonlinearity exists in a relationship. In the scatterplot above no clear linear relationship can be found. Similarly, the remaining assumptions about the errors may not be satisfied. Again, this can be seen in the above scatter plot where the variance in the number of trips appears to be increasing as the number of dependents increases. Homoskedasticity assumes that variance is fixed throughout a distribution. In many cases a log-linear form may better describe the relationship between the dependent and independent variable, and reduce the effects of increasing variance. To test whether the errors are normally distributed a histogram of standardized residuals should show a roughly normal curve. A normal probability plot, also called a P-P Plot, is an alternative method, plotting observed cumulative probabilities of occurrence of the standardized residuals on the Y axis and of expected normal probabilities of occurrence on the X axis, such that a 45-degree line will appear when the observed conforms to the normally expected and the assumption of normally distributed error is met. To test whether the assumptions hold, these plots are produced for the SEQ data in Appendix 6. The problem of multicollinearity is shortley discussed in Appendix 3 and also displays a correlation matrix. Briefly, it occurs when there is a linear relation between the explanatory variables (i.e. multicollinearity in regression occurs when predictor variables (independent variables) in the regression model are more highly correlated with other predictor variables than with the dependent variable).

These problems are often avoided by using a stratified dummy variable regression approach, where all variables are categorised. The VLC model uses a stratified dummy variable regression technique as well, which essentially calculates average trip rates for various categories of households. As mentioned before, the home based trip purposes used in this model are:

- Home based Work: Blue Collar (HBW_BLUE)
- Home based Work: White Collar (HBW_WHITE)
- Home based Education: Pre and Primary School (HBE_PrePrim)
- Home based Education: Secondary School (HBE_SEC)
- Home based Education: Tertiary School (HBE_TER)
- Home based Shopping/Personal Business (HBS)
- Home based Other (HBO)

Each trip purpose is related to one or several household attributes (i.e. factors affecting production), these are:

- Household size (1,2,3,4,5,6+)
- Car ownership (0,1,2,3+)
- Resident blue collar workers (0,1,2,3+)
- Resident white collar workers (0,1,2,3+)
- Resident (0-17) dependants (0,1,2,3,4+)
- Resident (18-64) dependants (0,1,2,3+)
- Resident (65+) dependants (0,1,2+)

Each attribute is included in the regression equation as a dummy 0-1 variable (see Appendix 2). This means that rather than finding a linear relationship between the number of trips per household and the household attribute, the regression is simply calculating an average trip rate for each category of the household attribute. The addition of further variables can be interpreted as making adjustments to those averages for other household attributes. In this way, the interpretation of the regression model is in fact almost identical to that of category analysis. The difference is that category analysis includes all interactions by default up to the number of variables in the model, and dummy regression only includes the main effects as default although additional variables are usually included using a stepwise regression method.

The objection to category analysis is that it does not give measures of fit or accuracy, nor does it include methods for determining the importance of variables, furthermore it does not easily lend itself to aggregate predictions using zonal totals, which is relatively simple with a linear model. The problem with linear regression though is, that it relies on certain statistical assumptions that have been found to be seriously violated by the 2004/'05 SEQ data, which can be seen in the plots in Appendix 6 and is further discussed in chapter 9. Briefly, the trip data consists of counts close to zero (and heavy in zeros), and does not satisfy the assumptions of normality or homoskedastic erros, furthermore some independent variables seem to be highly correlated. This means that simple linear regression analysis is not truly appropriate, and should be interpreted with caution.

Regression does still have the advantage over traditional category analysis, as it is possible to calculate many data fit and performance measures. However, the linear regression could be improved without stratifying continuous variables by using a generalised linear model (GLM). GLM's relax many of the assumptions of ordinary linear regression, in particular the assumption of linearity between dependent and independent variables. Count data is often well described by a Poisson distribution, and a GLM with a log link is commonly used in this case. For example see texts in [Huet et al., 2004] and [McCullagh,1989]. To visually check if the log link is warranted, graphs can be compared of the log transformed and untransformed values of the dependent variable against the independent variable. Where there are large numbers of zeros in the data, a two-stage approach may be justified. First predicting whether or not a person/household makes trips of a certain type; then predicting how many trips are made. This has been explored for one dataset by Monzon et al. (1989), who found that the added complexity did not add greatly to model fit. A slightly different technique to the two-stage model, is the two-part / zero-inflated model, in which the parameters are estimated simultaneously. These models have been well tested in other applications [Cheung, 2002]; [Hall, 2000]; [Lambert, 1992] and [Lewsey and Thomson, 2004], but have not been applied to trip generation data.

6 Methodology

This chapter describes the methodology that has been applied to design and develop the new trip generation equations. For the full analysis, the data was split randomly into a calibration and validation datasets to facilitate comparison between models. A 50:50 split was generated which brings the calibration as well as the validation dataset to a size of 3376 household samples. This reduced calibration sample size should be adequate to estimate the parameters for each model.
Although it is said some basic assumptions of linear regression are seriously violated by the data, I have chosen to perform the stratified dummy variable regression approach as a starting point since VLC uses this technique and problems are said to be avoided by using this method, where all variables are categorised. Later on this method will be evaluated and discussed whether improvements on the model should be made.

6.1 Calibration

For each trip purpose the calibration dataset was created and then imported into SPSS to execute a linear regression analysis as described in paragraph 5.6 For each trip purpose, the observed number of trips per household is selected as the 'dependent' variable. Subsequently two different models with different 'independent' variables are run for calibration:

1) Firstly, a model containing the same independent variables as from the 1992 Melbourne survey will be used as a starting point. See table 12 in chapter 7 for the coefficients of this model.
2) Secondly, I have chosen to perform a stepwise regression method in which the choice of independent variables is carried out by an automatic procedure. See table 14 for the coefficients of this model.

The first model contains the same variables as the 'old model' (from the 1992 survey data), but with updated coefficients (based on the new 2004/'05 SEQ data). Some variables seemed not to be significant anymore (at the 5% significance level) after running the regression on the data. These variables are coloured *red and italicised* in the trip generation coefficients table 12. This doesn't mean these variables were not significant for the old Melbourne data though. It is then also possible to test each case with or without an intercept/constant, but whether this is significant needs to be questioned.
To understand how this works, consider figure 9 below. The predictor variable (total number of persons) is stratified (total persons = 1, 2, 3, 4, 5 or 6+) and for each stratification the number of produced trips is presented in the scatter plot.

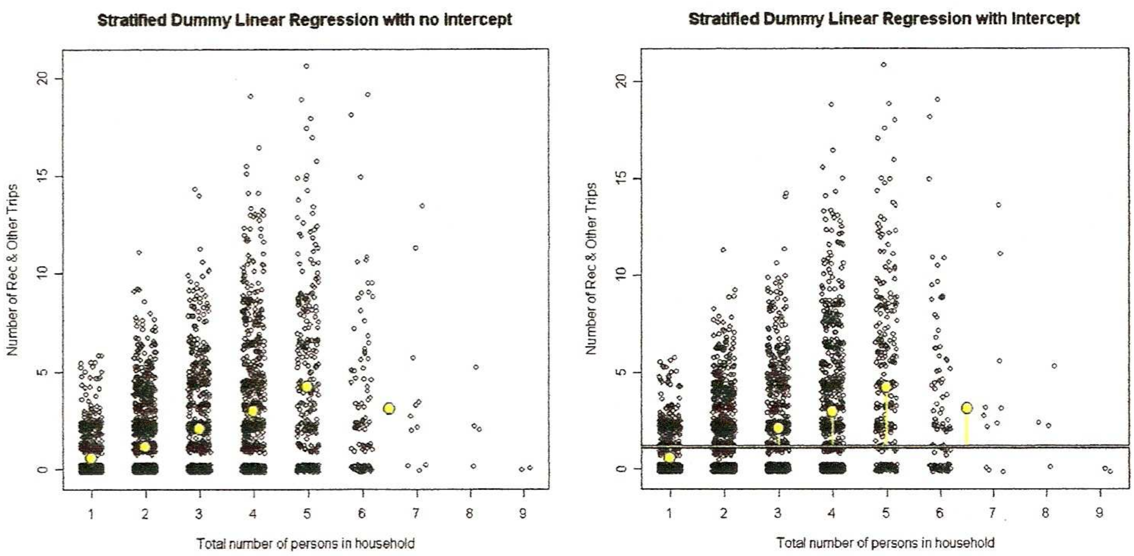


figure 9: Illustration of stratified dummy linear regression with (right diagram) and without (left diagram) intercept. The points on the graph have been 'jittered' so that they are not hidden underneath each other.

When there is no intercept (left diagram) the regression coefficient is equivalent to the average number of trips made by people in each category (yellow dots). When the intercept is present (right diagram), one of the dummy variables is excluded as it is considered the base case, in this case when total persons = 2. Each coefficient then represents the difference between the base case and the

average trip rate, and hence can be negative. Thus, for simple stratified dummy linear regression it is intuitively more appealing to exclude the intercept. Therefore, both new models were eventually forced without an intercept; notice that the old model contains a constant for several trip purposes though and should be interpreted with caution. Furthermore, trip purposes that are not considered satisfactory in the first model are those that include many variables (such as for shopping, recreation and other trips) and yet still had low explanatory power (low R-squared values). For that reason the insignificant variables (at the 5% level) were eventually excluded in the second model. And where two variables display high cross-correlations (see Appendix 3) both were tested separately (with no correlated variables still in the model) and the better performing variable was selected.
Taking these observations into account, the second model could be created based on simple but established relationships such as the number of blue collar work trips being related to the number of blue collar workers in the household. These assumptions become rather difficult though when considering HBS-, HBR- and HBO trips because it is difficult to choose which variables are important predictors for these purposes. That is why I have chosen to perform a stepwise regression to examine alternative model formulations instead, in which the choice of predictive variables is carried out by an automatic (forward and backward regression) procedure. Briefly, this method starts with the choice of the most significant predictor variable, namely the one for which R^2 is highest. Subsequently this step is repeated and variables are added to the model one by one. The variable, for which R^2 increases most, is chosen each time. The method stops when there are no variables anymore that allow R^2 to rise. In other words, this step is repeated until all significant variables (within the 5% level) are added and insignificant variables are excluded. In some cases, significant variables are added but in the end did not contribute much to the explanatory power of the model. That's why I have chosen to select that model run that explains more or less 98% of the end R-squared value. In other words, the R-squared value tends to approach an end value that doesn't change much anymore, even if more significant variables are added to the model. In some cases, I have chosen to exclude even more variables; this is when there are still correlated variables in the model.
The final model is selected on the basis of the F-statistic value (as discussed later on), model simplicity (number of predictor variables) and maximizing the R-squared (without compromising variable significance). That is, if two models are within the certain 2 percent of each other in terms of R-squared values, the simpler model is selected. Note that some models will have a very low explanatory power. Having a look at the case of shopping, recreation and other trips, a simpler model using just the total number of persons in the household had approximately similar explanatory power to models containing all of the other variables combined. In such case, the simpler model is preferred as there is little theory to support the relationships in the more complicated versions. The calibration results are discussed in chapter 7.

6.2 Validation

For the development of a simulation model, validation is an important process that takes place after calibration to check whether the model formulation is appropriate. On the basis of 50% of the data, that has not been used for the development of the model, the model outcomes are reviewed by testing whether the observed number of trips for each purpose and all attribute groups are sufficient in line with the estimated model outcomes. When model and reality have sufficient agreement one can speak of a valid and legitimate model. For each purpose, the following three trip production models are compared:

1) Firstly, the old VLC model containing the variables derived from the 1992 Melbourne data is tested;
2) Secondly, the model containing the same variables with updated coefficients derived from the 2004/'05 SEQTS survey is tested;
3) Thirdly, the model derived from the stepwise linear regression on the basis of simplicity and maximising R-squared is tested.

table 13 and table 15 in chapter 7 compare the reported number of home-based trips made by the sample households with estimates obtained using the equations contained in table 12 and table 14. To further ground the validity of the models a more detailed examination has been made of the performance of the trip production equations by comparing model forecasts with reported trip-making for a cross-classification of the broad households attributes used in the model derivation. The comparison was made by comparing trip rates for 80 categories of household incorporated in the model – i.e. workers (4) x dependants (5) x cars (4). This is an exhaustive test of the model's ability to predict trips at the disaggregate household level, and verifies the use of stratified multiple linear regression as a replacement for the more conventionally adopted "category analysis" technique. The validation results are discussed in chapter 7 as well.

7 Comparison of models

This chapter forms the actual development of the new trip generation equations for the South East Queensland region. At first, both new models are presented and their performance (compared to the old model) is tested in paragraph 7.1 and 7.2. Paragraph 7.3 presents the cross classification results.

7.1 1st Regression model

The first linear model is identical to the existing VLC model, except that it was created on the calibration dataset. This is so the performance of the original model can be determined and compared with the other models.

Stratified Variable	Trip Purpose							
	HBW-Blue	**HBW-White**	**HBE-Pre/Prim**	**HBE-Secondary**	**HBE-Tertiary**	**HBS**	**HBR**	**HBO**
1 Person								-1,261
2 Persons								-1,081
3 Persons								-0,808
4 Persons			***-0,048***	0,147				***0,039***
5 Persons			***-0,154***	0,417				***0,177***
6+ Persons			-0,235	0,502				
0 Cars					-0,037	***-0,316***	-0,635	***0,066***
1 Cars					-0,059	***0,017***	-0,288	0,431
2 Cars	***0,023***				***-0,073***	***0,028***	***-0,094***	0,401
3+ Cars	0,081	0,298						
0 Blue-Collar								0,147
1 Blue-Collar	1,228				***0,010***	0,349	***-0,058***	
2 Blue-Collar	2,653				***0,018***	0,460	***-0,032***	
3+ Blue-Collar	4,078				***-0,029***	***0,992***	***-0,121***	
0 White-Collar								
1 White-Collar		1,070			***0,014***	0,517	0,209	
2 White-Collar		2,075			0,104	0,942	0,418	
3+ White-Collar		3,217			***0,065***	1,435	0,698	
0 Dependants (0-17)		0,123			0,044			-3,251
1 Dependants (0-17)			0,361	0,506		0,514	0,269	-2,161
2 Dependants (0-17)			1,319	0,533		0,680	0,818	-1,475
3 Dependants (0-17)			2,304	0,589		0,806	1,356	***-0,406***
4+ Dependants (0-17)			3,354	0,962		1,439	1,194	
0 Dependants (18-64)								-0,215
1 Dependants (18-64)					0,144	1,041	0,320	
2 Dependants (18-64)					0,441	1,650	0,430	
3+ Dependants (18-64)					1,418	2,256	1,306	
0 Dependants (65+)								
1 Dependants (65+)						0,963	0,320	
2+ Dependants (65+)					***0,054***	1,935	0,714	
Constant	0,0	0,0	***0,004***	0,0	0,0	***0,140***	0,622	4,328
F-Statistic	1426	1440	370	189	55	21	17	146
R-squared (Adjusted)	0,679	0,681	0,434	0,281	0,182	0,097	0,077	0,376
Standard error of the Estimate	0,599	0,976	0,784	0,702	0,444	2,026	1,783	1,768

table 12: trip production model coefficients, calibration results home-based person trips per household, model 1

7.1.1 Formulation and fit

The information as presented in table 12 above originates from the output generated by SPSS after running the linear regression analysis. A comprehensive explanation of the model output is presented in Appendix 5.1. As mentioned in paragraph 6.1 some variables used in the 'old model' seemed not to be significant anymore (at the 5% significance level) after running the regression on the new SEQ data. These variables are coloured *red and italicised* in the trip generation coefficients table above.

Having a closer look to the results tells us the home based work trips (blue and white) perform quite well. The large adjusted R-squared values indicate the variables capture the data for a substantial amount. The regression and residual sums of squares (presented in the ANOVA table of the model output) are approximately 2:1, which indicates that about 68% of the variation in trips is explained by the equations in the model. Furthermore the significance value of the F statistic is less than 0.05, which means that the variation explained by the model is not due to chance, indicating that using the model is better than guessing the mean. Interesting to note are the higher rates of home based work trip making for blue collar workers relative to white collar workers. This was also the case in the 'old model coefficients' and could reflect the higher tendency for white collar workers to serve a passenger (i.e. drop their children at school for instance) or go shopping on the way home from work. This seems to be confirmed by the fact that not having children (0 Dep_0_17) has a positive effect on white collar work trips. The above assertion seems to be confirmed by the higher calibration coefficients obtained for white collar shopping trips as well. This assumption should be interpreted with caution though, due to the great number of variables in shopping trips.

Considering education trips, some problems arise. The coefficients table (see model output for more detail) shows that there are too many predictors in the model. First of all there are several non-significant coefficients (especially in tertiary trips this is remarkable), indicating that these variables do not contribute much to the model. Furthermore the problem of multicollinearity arises. The correlation matrix in Appendix 3 shows 'Household size' (HHsize) and 'Dependants 0-17' (Dep_0_17) are highly correlated. The collinearity statistics (see coefficients table in the model output) confirm there are problems with multicollinearity. The tolerance is the percentage of the variance in a given predictor that cannot be explained by the other predictors. Thus, the tolerances show that 40-50% of the variance in a given predictor can be explained by the other predictors. When the tolerances are close to 0, there is high multicollinearity and the standard error of the regression coefficients will be inflated. A variance inflation factor (VIF) greater than 2 or 3 is usually considered problematic, there are no clear rules though.

The behaviour of the remaining trip purposes (shopping, recreation and other) should be interpreted with caution. For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This cannot be compared to R Square for models, which include an intercept. The mentioned trip purposes, are such models with an intercept and hence the acceptance or rejection of any of the models should not be based on the R-squared only. Furthermore these trip purposes contain a lot of variables which make the interpretation of the effects of one single variable difficult. Whether the independent variables are a true cause of the changes in the dependent variable should be questioned.

7.1.2 Performance

The equations (using the coefficients in table 12) created on the calibration dataset were used to predict the number of trips per household for the validation dataset. These predictions were then compared with the observed number of trips per household and with the predictions generated by the 'old model'. The ability of the trip production equations to replicate observed trips by broad attribute classifications is presented in table 13. In other words, table 13 represents the total number of observed & estimated trips and its distribution over the different attributes. As an example consider trip purpose 'HBW-Blue' of which in total 1619 trips are observed in the validation dataset. Now if we, for instance, take a look at the HH description variable "Workers (Blue)", one can read from this table that households containing 1 blue collar worker made 1138 HBW-Blue trips. The 'new model' estimates this particular attribute makes 1092 HBW-Blue trips whether the 'old model' predicts 1060 HBW-Blue trips. So for this particular attribute we can conclude that the new model is a better predictor than the old model. Furthermore the new model gives a better prediction (for the HBW-Blue collar trip purpose) of the total production of HBW-Blue trips, this is 1619 observed against 1609 estimated and 1507 estimated by the old model. Whether this holds for all trip purposes and all attributes is discussed later on in chapter 8 "results".

HH description variable			HBW-Blue			HBW-White			HBE-PrePrim			HBE-Sec			HBE-Ter			HBS			HBR			HBO		
	Level	Sample	Obs.	Est.	Old	Obs.	Est.	Old	Obs.	Est.	Old	Obs.	Est.	Old	Obs.	Est.	Old	Obs.	Est.	Old	Obs.	Est.	Old	Obs.	Est.	Old
Persons	1	735	107	115	112	208	297	290	0	3	1	0	0	0	9	14	24	699	640	935	433	384	342	52	84	126
	2	1394	518	530	502	1107	1124	1084	36	35	25	42	41	33	36	127	125	2536	2176	2659	1472	1240	1045	420	565	707
	3	506	372	358	332	530	604	574	218	176	134	147	149	110	30	63	73	936	926	977	586	572	475	632	636	835
	4	499	443	413	382	651	673	642	509	467	384	291	295	246	65	60	70	1013	1036	1080	799	774	600	1360	1383	1535
	5	175	140	147	135	199	238	227	255	282	245	171	167	127	25	25	27	448	406	432	372	356	242	720	665	835
	6+	67	39	47	43	77	94	88	156	161	135	87	85	77	6	16	15	196	206	195	143	150	127	364	275	365
Totals		3376	1619	1609	1507	2772	3030	2904	1174	1124	924	738	737	592	171	306	334	5828	5390	6279	3805	3475	2830	3548	3608	4402
Cars	0	222	14	20	19	23	53	51	28	23	18	11	14	10	4	14	19	252	220	327	93	91	115	52	28	52
	1	1351	273	297	285	586	714	701	326	280	224	185	174	130	33	97	89	2233	2017	2259	1345	1146	950	952	1053	1279
	2	1366	863	820	781	1474	1483	1456	678	664	547	374	402	313	61	104	111	2421	2296	2726	1710	1653	1287	1933	1999	2378
	3+	437	469	472	421	689	779	696	142	157	135	168	147	139	73	92	115	922	856	966	657	585	478	611	528	695
Totals		3376	1619	1609	1507	2772	3030	2904	1174	1124	924	738	737	592	171	306	334	5828	5390	6279	3805	3475	2830	3548	3608	4402
Workers (Blue)	0	2336	27	34	26	2079	2156	2076	694	683	557	411	424	330	121	243	219	4125	3802	4534	2665	2455	1994	2180	2288	2796
	1	869	**1138**	**1092**	**1060**	606	753	722	416	395	325	276	263	213	41	52	89	1366	1350	1441	948	866	697	1175	1116	1372
	2	156	395	421	368	82	111	99	58	45	40	48	44	41	5	10	23	301	210	272	176	141	122	174	182	212
	3+	15	59	62	53	5	10	8	6	2	2	3	7	8	4	0	3	36	28	32	16	12	18	19	22	22
Totals		3376	**1619**	**1609**	**1507**	2772	3030	2904	1174	1124	924	738	737	592	171	306	334	5828	5390	6279	3805	3475	2830	3548	3608	4402
Workers (White)	0	1601	724	743	693	57	184	154	370	329	267	185	203	156	50	175	141	2922	2570	3017	1635	1418	1254	1092	1185	1439
	1	1133	774	719	682	1224	1349	1334	488	480	392	348	315	250	66	61	105	1805	1750	1959	1298	1201	948	1407	1439	1767
	2	585	105	128	116	1289	1295	1226	311	307	255	189	200	165	45	63	68	991	954	1144	799	773	546	983	911	1116
	3+	57	16	19	16	202	202	191	5	9	10	16	19	21	10	6	20	110	115	159	73	84	82	66	73	81
Totals		3376	1619	1609	1507	2772	3030	2904	1174	1124	924	738	737	592	171	306	334	5828	5390	6279	3805	3475	2830	3548	3608	4402
Dependants (0-17)	0	2356	974	979	915	1785	1906	1819	0	3	5	6	17	26	111	226	241	3843	3347	4217	2247	1918	1644	550	736	913
	1	402	271	272	253	382	438	421	178	139	114	230	225	195	33	36	43	735	766	729	498	445	383	704	730	819
	2	439	295	264	249	431	488	473	597	559	445	310	299	230	21	25	32	816	852	890	675	715	533	1326	1342	1622
	3	140	64	72	68	131	152	148	287	301	251	138	139	88	6	14	13	328	313	338	308	315	192	714	619	799
	4+	39	15	22	21	43	46	44	112	122	109	54	57	54	0	4	4	106	113	105	77	82	78	254	181	249
Totals		3376	1619	1609	1507	2772	3030	2904	1174	1124	924	738	737	592	171	306	334	5828	5390	6279	3805	3475	2830	3548	3608	4402
Dependants (8-65)	0	2171	1089	1123	1051	1995	2275	2192	680	626	512	428	395	311	13	40	98	3304	2916	3733	2178	2030	1633	1869	1850	2391
	1	940	464	429	404	661	634	603	404	426	352	240	277	222	96	124	127	1742	1870	1883	1232	1120	924	1385	1400	1626
	2	237	57	48	44	98	103	94	76	60	50	56	52	46	45	102	84	683	519	582	336	263	242	226	303	328
	3+	28	9	9	8	18	17	15	14	12	10	14	12	12	17	40	24	99	84	81	59	62	32	68	56	58
Totals		3376	1619	1609	1507	2772	3030	2904	1174	1124	924	738	737	592	171	306	334	5828	5390	6279	3805	3475	2830	3548	3608	4402
Dependants (65+)	0	2466	1539	1524	1427	2656	2797	2689	1139	1091	899	704	711	568	166	268	310	4023	3817	4382	2834	2652	2052	3260	3284	3996
	1	566	70	72	68	94	169	157	31	28	21	18	19	17	3	22	23	957	832	996	508	442	446	182	184	234
	2	344	10	13	12	22	63	57	4	5	4	16	7	8	2	15	1	848	741	900	463	381	332	106	140	172
Totals		3376	1619	1609	1507	2772	3030	2904	1174	1124	924	738	737	592	171	306	334	5828	5390	6279	3805	3475	2830	3548	3608	4402

table 13: Comparison of estimated (old and 1st new model) against observed home based person trip productions

7.2 2nd Regression model

The second linear model is generated, using a stepwise regression method on the calibration dataset in which the choice of independent variables is carried out by an automatic procedure. This means all variables are selected on the basis of maximum t-value and the 5% significance level (see model output for details).

Stratified Variable	Trip Purpose							
	HBW-Blue	HBW-White	HBE-Pre/Prim	HBE-Secondary	HBE-Tertiary	HBS	HBR	HBO
1 Person						-1,288		
2 Persons						-0,667	0,363	
3 Persons						-0,380	0,784	
4 Persons							1,121	
5 Persons							1,720	
6+ Persons						1,026	1,489	
0 Cars						-0,313	-0,333	
1 Cars								
2 Cars								
3+ Cars	0,075	0,239						
0 Blue-Collar							0,586	
1 Blue-Collar	1,241						0,272	
2 Blue-Collar	2,665					-0,389		
3+ Blue-Collar	4,089							
0 White-Collar								
1 White-Collar		1,170						
2 White-Collar		2,241						
3+ White-Collar		3,375						
0 Dependants (0-17)		0,246						
1 Dependants (0-17)			0,341	0,572				1,517
2 Dependants (0-17)			1,261	0,707				2,793
3 Dependants (0-17)			2,160	0,977				4,028
4+ Dependants (0-17)			3,130	1,457				4,354
0 Dependants (18-64)		-0,266				-0,637		
1 Dependants (18-64)					0,137			
2 Dependants (18-64)					0,444			
3+ Dependants (18-64)					1,415			
0 Dependants (65+)						2,464		0,364
1 Dependants (65+)						2,995		
2+ Dependants (65+)						3,412		
Constant	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
F-Statistic	1782	1230	834	317	237	247	183	659
R-squared (Adjusted)	0,678	0,686	0,497	0,272	0,173	0,421	0,302	0,494
Standard error of the Estimate	0,600	0,968	0,784	0,707	0,447	2,025	1,783	1,797

table 14: trip production model coefficients, calibration results home-based person trips per household, model 2

7.2.1 Formulation and fit

The information as presented in table 14 above originates from the output generated by SPSS after running the stepwise linear regression analysis. The comprehensive model output is presented in Appendix 5.2. The preferred model and selected variables are coloured *blue* in the various model output tables.

Having a closer look to the results tells us the home based work trips (blue and white) perform quite well again. The stepwise regression analysis removed insignificant- and added significant variables if necessary. The results of the stepwise regression on HBW-White trips display more variables than I finally selected in the model. This is due to the fact that the R-squared isn't improving and the std. error of the estimate is not reducing anymore after adding more variables. Furthermore the problem of multicollinearity seems to occur if more variables are added.
HBE-Pre/Prim improved in terms of R-squared and simplicity. The HBE-secondary model is simpler but apparently did not perform better compared to the previous model. Whether this really is the case in terms of predicting the number of trips can be seen (and is discussed) in paragraph 7.2.2. Regarding HBE-tertiary the regression analysis proposes to improve the model by adding more variables again. Even though multicollinearity does not arise and the R-squared can be improved there is not much theory to prove the more complicated model should be better. After validation of both the simple and the more complicated model it became clear there is not much difference so the simpler model is preferred.
The following trip purposes (shopping and recreation) are more complicated since it is difficult to determine and to prove (even though stepwise regression automatically selects most significant variables) which factors really drive these specific trips. Considering HBS first, the stepwise regression came up with three high coefficients for Dep_65+. The use of dummy variables requires the imposition of additional constraints on the parameters of regression equations. The possible constraints are (a) to set the constant term of the equation to zero, or (b) to omit one of the dummy variables from the equation to avoid perfect multicollinearity. Apparently everybody makes a daily shopping trip and there is no situation possible that one would not. For that reason one could argue to include a constant for this specific trip purpose instead of using the dependant_65+ variable. Including a constant would only influence the dep_65+ variable (see model output HBS stepwise regression also) where a clear shift is recognizable from the dep_65+ parameter values to the constant. I have chosen to keep the constant term zero though, since this makes comparison with- and understanding of the different trip purposes easier. Either way, the results in trip production estimates will be the same.
In terms of recreational trips it is interesting to note the effect of not having blue collar workers in the household. Apparently households without blue collar workers make more recreational trips than households containing a blue collar worker. Although it sounds logical it is also interesting to note the suppressing effect of not owning a car has on home based shopping & recreational trips.
The final trip purpose (HBO) basically concerns accompanying or serving passengers. This often implies dropping children at school or picking them up. Taking a closer look to the model output tells that having children is the major factor that drives this trip purpose. Adding more variables would improve the model slightly but especially household size is correlated with the attribute dependants(0-17) which should be avoided. For that reason the best predictor variable is chosen, which was dep_0_17.

7.2.2 Performance

The equations (using the coefficients in table 14) created on the calibration dataset were used to predict the number of trips per household for the validation dataset. These predictions were compared again with the observed number of trips per household and with the predictions generated by the old model. The ability of the trip production equations to replicate observed trips by broad attribute classifications is presented in table 15.

HH description variable	Level	Sample	HBW-Blue			HBW-White			HBE-PrePrim			HBE-Sec			HBE-Ter			HBS			HBR			HBO		
			Obs.	Est.	Old	Obs.	Est.	Old	Obs.	Est.	Old	Obs.	Est.	Old	Obs.	Est.	Old	Obs.	Est.	Old	Obs.	Est.	Old	Obs.	Est.	Old
Persons	1	735	107	115	112	208	244	290	0	0	1	0	0	0	9	17	24	699	601	935	433	350	342	52	149	126
	2	1394	518	518	502	1107	1119	1084	36	28	25	42	46	33	36	102	125	2536	2285	2659	1472	1181	1045	420	449	707
	3	506	372	355	332	530	596	574	218	165	134	147	176	110	30	57	73	936	899	977	586	607	475	632	700	835
	4	499	443	409	382	651	653	642	509	466	384	291	289	246	65	57	70	1013	1048	1080	799	757	600	1360	1260	1535
	5	175	140	146	135	199	231	227	255	290	245	171	144	127	25	24	27	448	374	432	372	372	242	720	632	835
	6+	67	39	46	43	77	93	88	156	166	135	87	79	77	6	15	15	196	225	195	143	128	127	364	278	365
Totals		3376	1619	1588	1507	2772	2936	2904	1174	1115	924	738	734	592	171	274	334	5828	5432	6279	3805	3394	2830	3548	3467	4402
Cars	0	222	14	20	19	23	36	51	28	22	18	11	16	10	4	11	19	252	222	327	93	86	115	52	85	52
	1	1351	273	300	285	586	659	701	326	269	224	185	192	130	33	99	89	2233	2028	2259	1345	1174	950	952	946	1279
	2	1366	863	796	781	1474	1469	1456	678	657	547	374	407	313	61	108	111	2421	2341	2726	1710	1565	1287	1933	1892	2378
	3+	437	469	472	421	689	772	696	142	167	135	168	119	139	73	56	115	922	841	966	657	570	478	611	545	695
Totals		3376	1619	1588	1507	2772	2936	2904	1174	1115	924	738	734	592	171	274	334	5828	5432	6279	3805	3394	2830	3548	3467	4402
Workers (Blue)	0	2336	27	14	26	2079	2127	2076	694	671	557	411	436	330	121	213	219	4125	3773	4534	2665	2391	1994	2180	2074	2796
	1	869	1138	1091	1060	606	706	722	416	394	325	276	259	213	41	53	89	1366	1422	1441	948	845	697	1175	1204	1372
	2	156	395	421	368	82	94	99	58	47	40	48	35	41	5	7	23	301	206	272	176	138	122	174	174	212
	3+	15	59	62	53	5	8	8	6	3	2	3	4	8	4	1	3	36	31	32	16	20	18	19	15	22
Totals		3376	1619	1588	1507	2772	2936	2904	1174	1115	924	738	734	592	171	274	334	5828	5432	6279	3805	3394	2830	3548	3467	4402
Workers (White)	0	1601	724	739	693	57	110	154	370	320	267	185	212	156	50	183	141	2922	2632	3017	1635	1306	1254	1092	1024	1439
	1	1133	774	712	682	1224	1336	1334	488	475	392	348	319	250	66	71	105	1805	1749	1959	1298	1196	948	1407	1497	1767
	2	585	105	119	116	1289	1289	1226	311	309	255	189	191	165	45	19	68	991	946	1144	799	797	546	983	892	1116
	3+	57	16	18	16	202	200	191	5	12	10	16	12	21	10	1	20	110	105	159	73	95	82	66	55	81
Totals		3376	1619	1588	1507	2772	2936	2904	1174	1115	924	738	734	592	171	274	334	5828	5432	6279	3805	3394	2830	3548	3467	4402
Dependants (0-17)	0	2356	974	967	915	1785	1883	1819	0	0	5	6	0	26	111	179	241	3843	3388	4217	2247	1845	1644	550	536	913
	1	402	271	270	253	382	411	421	178	137	114	230	230	195	33	42	43	735	715	729	498	499	383	704	752	819
	2	439	295	260	249	431	453	473	597	554	445	310	310	230	21	32	32	816	886	890	675	679	533	1326	1383	1622
	3	140	64	70	68	131	145	148	287	302	251	138	137	88	6	17	13	328	319	338	308	295	192	714	612	799
	4+	39	15	22	21	43	44	44	112	122	109	54	57	54	0	4	4	106	124	105	77	76	78	254	184	249
Totals		3376	1619	1588	1507	2772	2936	2904	1174	1115	924	738	734	592	171	274	334	5828	5432	6279	3805	3394	2830	3548	3467	4402
Dependants (8-65)	0	2171	1089	1111	1051	1995	2048	2192	680	616	512	428	404	311	13	0	98	3304	2933	3733	2178	1984	1633	1869	1943	2391
	1	940	464	423	404	661	739	603	404	424	352	240	276	222	96	129	127	1742	1932	1883	1232	1066	924	1385	1249	1626
	2	237	57	46	44	98	128	94	76	62	50	56	45	46	45	105	84	683	495	582	336	296	242	226	237	328
	3+	28	9	9	8	18	20	15	14	13	10	14	8	12	17	40	24	99	72	81	59	49	32	68	38	58
Totals		3376	1619	1588	1507	2772	2936	2904	1174	1115	924	738	734	592	171	274	334	5828	5432	6279	3805	3394	2830	3548	3467	4402
Dependants (65+)	0	2466	1539	1507	1427	2656	2787	2689	1139	1083	899	704	714	568	166	247	310	4023	3849	4382	2834	2659	2052	3260	3396	3996
	1	566	70	70	68	94	132	157	31	28	21	18	15	17	3	25	23	957	843	996	508	399	446	182	57	234
	2	344	10	11	12	22	17	57	4	5	4	16	5	8	2	2	1	848	740	900	463	336	332	106	15	172
Totals		3376	1619	1588	1507	2772	2936	2904	1174	1115	924	738	734	592	171	274	334	5828	5432	6279	3805	3394	2830	3548	3467	4402

table 15: Comparison of estimated (old and 2nd new model) against observed home based person trip productions

7.3 Cross Classification

As mentioned in chapter 6.2, a more detailed examination has been made of the performance of the trip production equations by comparing model forecasts with reported trip-making for a cross-classification of the broad household attributes used in the model derivation. This tests the model's ability to predict trips at the disaggregate household level and verifies the use of a stratified dummy variable regression. The results presented in table 16 and table 17 compare the estimated- (of the three models) and observed trip production rates per household. The household sample size in each category is given to indicate the significance of the comparison. An example is given on the next page.

Observed, Estimated and Old Model Home-Based Household Trip Rates
summary of trip rates by household structure - combined purposes

car ownership = 0

Workers	Status	Dependants 0 Trip Rate	0 Sample	1 Trip Rate	1 Sample	2 Trip Rate	2 Sample	3 Trip Rate	3 Sample	4 + Trip Rate	4 + Sample
0	Observ	0	0	1,430	135	2,381	42	7,600	5	6,667	3
	Model 1	0		1,120		2,985		6,891		9,878	
	Model 2	0		1,145		3,077		8,070		10,095	
	Old model	0		1,872		3,535		7,274		10,064	
1	Observ	1,650	20	2,571	7	8,000	3	6,000	1	12,000	1
	Model 1	1,419		3,283		5,944		9,219		13,566	
	Model 2	1,908		3,827		6,728		9,714		14,062	
	Old model	2,182		3,673		6,590		8,918		13,270	
2	Observ	4,000	1	6,000	2	16,00	1	0	0	0	0
	Model 1	3,478		5,977		9,656		0		0	
	Model 2	4,048		6,796		9,988		0		0	
	Old model	3,983		6,198		9,674		0		0	
3+	Observ	1,000	1	0	0	0	0	0	0	0	0
	Model 1	5,133		0		0		0		0	
	Model 2	5,683		0		0		0		0	
	Old model	5,393		0		0		0		0	

car ownership = 1

Workers	Status	Dependants 0 Trip Rate	0 Sample	1 Trip Rate	1 Sample	2 Trip Rate	2 Sample	3 Trip Rate	3 Sample	4 + Trip Rate	4 + Sample
0	Observ	0	0	2,082	293	4,351	373	6,375	40	12,000	25
	Model 1	0		2,100		3,905		7,048		11,180	
	Model 2	0		1,898		3,463		7,360		11,368	
	Old model	0		2,446		3,989		6,927		11,069	
1	Observ	2,387	225	4,576	132	6,425	73	9,500	26	12,600	10
	Model 1	2,357		4,575		7,375		10,466		12,573	
	Model 2	2,564		4,795		7,792		10,175		14,351	
	Old model	2,777		4,503		7,272		9,673		14,130	
2	Observ	4,341	82	7,960	25	12,609	23	14,500	8	20,000	4
	Model 1	4,426		7,154		10,560		13,760		13,079	
	Model 2	4,597		7,136		10,198		13,151		15,886	
	Old model	4,398		6,588		9,897		13,075		15,970	
3+	Observ	8,429	7	10,500	4	20,000	1	0	0	0	0
	Model 1	7,192		9,906		12,758		0		0	
	Model 2	6,565		8,951		11,852		0		0	
	Old model	6,967		8,603		12,089		0		0	

table 16: summary of trip rates by household structure(1)

To understand how this cross classification works, consider for example the car ownership = 2 table below: the validation dataset contains 361 households with two cars, two workers and zero dependants. A particular household with this structure produced 4,59 trips on average per day. The different models predict production rates of 4,482 – 4,580 and 4,855 respectively, which all should be classified as good.

Observed, Estimated and Old Model Home-Based Household Trip Rates
summary of trip rates by household structure - combined purposes

car ownership = 2

Workers	Status	Dependants 0 Trip Rate	0 Sample	1 Trip Rate	1 Sample	2 Trip Rate	2 Sample	3 Trip Rate	3 Sample	4 + Trip Rate	4 + Sample
0	Observ	0		1,950		4,425		7,900		12,455	
	Model 1	0	0	2,207	20	3,893	160	6,314	20	11,508	11
	Model 2	0		2,073		3,462		5,762		11,043	
	Old model	0		2,908		4,355		6,053		11,309	
1	Observ	2,200		4,439		6,129		10,269		15,479	
	Model 1	2,339	35	4,357	155	6,911	70	10,466	93	13,579	48
	Model 2	2,535		4,382		6,818		10,119		13,707	
	Old model	3,167		4,608		6,771		10,007		13,861	
2	Observ	4,590		7,009		10,320		13,966		14,933	
	Model 1	4,482	361	7,236	115	10,704	172	13,870	59	13,046	15
	Model 2	4,580		7,160		10,408		13,374		15,978	
	Old model	4,855		7,090		10,423		13,679		16,604	
3+	Observ	7,500		9,308		13,600		0		0	
	Model 1	7,467	14	10,171	13	12,690	5	0	0	0	0
	Model 2	6,710		8,672		11,309		0		0	
	Old model	7,450		9,385		12,055		0		0	

car ownership ≥ 3

Workers	Status	Dependants 0 Trip Rate	0 Sample	1 Trip Rate	1 Sample	2 Trip Rate	2 Sample	3 Trip Rate	3 Sample	4 + Trip Rate	4 + Sample
0	Observ	0		3,250		4,200		9,333		17,000	
	Model 1	0	0	2,671	4	4,013	10	6,354	3	14,857	2
	Model 2	0		2,605		3,705		5,929		13,307	
	Old model	0		3,105		4,552		6,118		13,020	
1	Observ	2,000		4,963		7,238		9,833		14,000	
	Model 1	2,727	3	4,498	27	6,617	21	10,147	18	14,699	11
	Model 2	2,872		4,708		6,711		9,653		14,674	
	Old model	3,410		4,757		6,833		9,448		14,811	
2	Observ	5,295		6,741		10,688		11,611		12,222	
	Model 1	4,547	44	6,764	58	10,001	64	12,183	18	14,708	9
	Model 2	4,877		6,897		9,650		11,312		14,623	
	Old model	5,021		6,817		9,655		11,691		14,464	
3+	Observ	7,667		9,638		10,769		18,000		22,000	
	Model 1	7,182	75	9,864	47	12,292	13	14,126	6	18,271	4
	Model 2	7,103		9,223		11,645		13,886		17,619	
	Old model	7,743		9,672		12,474		13,182		18,868	

table 17: summary of trip rates by household structure(2)

8 Results

In order to compare how effectively each regression can model the household data we can analyse the calibration results and the validation results. The calibration results summarize the overall ability of the data to be modelled by a linear function. The validation results, in which the estimated (old and new models) against observed home based person trip productions per attribute group, and the home based household trip rates are presented, analyse the models ability to predict the observed trip counts. Even though the presented tables give an overview of the broad performance of each model it is difficult to indicate which model is best due to the large number of values. In order to give a better overview of the model performances this chapter summarizes the results and criteria for choosing the model. The final model isn't selected on the calibration results as maximizing R-squared, F-statistics, standard error of the estimate and model simplicity only, but also on minimizing the mean error between observed and estimated (predicted) trips.

8.1 Calibration

Starting with the calibration results, the model coefficients were recalculated using the SEQ HTS '03/04 data using both the variables in the original model and a simpler set of variables. table 18 below summarizes the results. The R-squared values and Std. error of estimate of the old model were not available and hence where produced on another dataset. The available data of this old model (F-statistic and # variables) was presented though as a comparison.

	F-statistic			R-squared (adjusted			Std.error of estimate			# variables in model		
	Model 1	Model 2	Old Model	Model 1	Model 2	Old Model	Model 1	Model 2	Old Model	Model 1	Model 2	Old Model
HBW-Blue	1426	1782	1528	0,679	0,678	-	0,599	0,600	-	5	4	5
HBW-White	1440	1230	1149	0,681	0,686	-	0,976	0,968	-	5	6	5
HBE-PrePrim	370	834	452	0,434	0,497	-	0,784	0,784	-	8	4	8
HBE-Sec	189	317	162	0,281	0,272	-	0,702	0,707	-	7	4	7
HBE-Ter	55	237	43	0,182	0,173	-	0,444	0,447	-	14	3	14
HBS	21	247	26	0,097	0,421	-	2,026	2,025	-	19	10	19
HBR	17	183	19	0,077	0,302	-	1,783	1,783	-	19	8	19
HBO	146	659	250	0,376	0,494	-	1,768	1,797	-	15	5	15
Average	458	686	454	0,351	0,440		1,135	1,139		11,5	5,5	11,5

table 18: Summary of calibration results

The F-statistics suggest that there is substantially more error in the non-work trips than in the work trips. Model 2 has been able to reduce this error a bit, since the F value is the ratio of the mean regression sum of squares divided by the mean error (residual) sum of squares, indicating that the model is able to explain more of the variation in the dependent variable. The significance value of the F-statistic is less than 0.05 for both models and all trip purposes, which means that the variation explained by the model is not due to chance and so it is better to use the model. Based on the F-statistic, model 2 is preferred.

R squared is the proportion of variation in the dependent variable explained by the regression model. It ranges in value from 0 to 1. Small values indicate that the data is not well captured by a linear model and does not fit the data well. Note that this value cannot be compared to R Squared for models which include an intercept (for model 1, this concern HBS, HBR and HBO trips). Overall, both sets of models seem to give similar results.

The standard error of the estimate is a measure of the accuracy of predictions made with the regression line. The increasing standard error of the estimate across trip purpose is a cause for concern. Again, both sets of models give approximately similar results.

Finally, the model simplicity is expressed by the number of variables in the model. Overall I must conclude that the calibration results show that model 2 is preferred due to the fact that it is able to produce the same statistical test results with fewer variables in the model. Whether model 2 is a better predictor of trips as well, is discussed in the following validation results.

8.2 Validation

8.2.1 Home based trip productions

The ability of the trip production equations to replicate observed trips by broad attribute classifications (see table 13 and table 15) appear to be good at first sight. Comparing total observed trips in each attribute with the predicted values of the new models and with predicted values of the old model, tend to confirm the new models are better estimators then the old model. To verify whether the one model has an overall better performance than the other, I computed two different "*mean errors"*:

- the mean absolute trip production error;
- the mean weighted trip production error.

For each attribute level (30 in total) of the seven stratified attribute groups in each trip purpose, the absolute error (difference between the observed and estimated number of trips per level) is calculated as a starting point. For instance the absolute trip production error for HBW-Blue trips of households containing 1 persons is |107 – 115| = 8. This is repeated for each level and each trip purpose. Eventually, the mean absolute trip production error is calculated as follows:

- mean absolute error $\mu_{abs} = \frac{\sum_{i=1}^{N} \mu_i}{N}$ (in which $\sum_{i=1}^{N} \mu_i$ is the sum of all trip production errors (difference) per level divided by the number of variables/levels *N* = *30*).

This value represents the average trip production error of each level per attribute group per trip purpose. In other words, it represents the average deviation in each trip production cell. For instance, the mean absolute error for home based blue collar trips of model 2 (table 15) is calculated as: $\mu_{abs} = \frac{|107-115|+|518-518|+...+|10-11|}{30} = 17{,}33$. Indicating that each cell (except from the total) in the estimate column of HBW-blue has an error of 17,33 on average. Note that each level in a specific attribute group has equal influence on this value no matter how many households have that particular structure. In other words, the 3376 households in this dataset are distributed over the different levels of the household description variable, and are not accounted for. The mean absolute error calculation does not consider this effect. So, it is more plausible to give extra weight to the variables that contain more households. For that reason, the "*mean **weighted** trip production **error***" is calculated:

- mean weighted error $\mu_{wgt} = \frac{\sum_{i=1}^{N} freq_i \cdot \mu_i}{\sum_{i=1}^{N} freq_i}$ (in which $\sum_{i=1}^{N} freq_i \cdot \mu_i$ is the sum of the household sample size per level times trip production error per level, divided by the total sample size in the attribute group $\sum_{i=1}^{N} freq_i$).

This value represents the average weighted trip production error of each level per attribute group per trip purpose. In other words, it represents the weighted average deviation in each trip production cell. For instance, the mean weighted error for home based blue collar trips of model 2 (table 15) is calculated as: $\mu_{abs} = \frac{(|107-115|*735)+(|518-518|*1394)+...+(|10-11|*344)}{3376} = 19{,}87$.

So, the mean weighted error also evaluates the frequency (sample size) in order to compute the mean error per cell. Eventually the mean absolute- and mean weighted trip production errors (in exact values as a percentage of the total trips) for all trip purposes are presented in table 19 below:

	mean absolute error			percentage			mean weighted error			percentage		
	model 1	model 2	old model	model 1	model 2	old model	model 1	model 2	old model	model 1	model 2	old model
HBW-Blue	16,22	17,33	29,50	0,010	0,011	0,018	19,87	25,46	38,55	0,012	0,016	0,024
HBW-White	61,97	38,97	43,70	0,022	0,014	0,016	102,06	60,84	62,24	0,037	0,022	0,022
HBE-PrePrim	20,00	21,90	59,07	0,017	0,019	0,050	27,29	29,38	96,55	0,023	0,025	0,082
HBE-Sec	9,83	12,96	36,40	0,013	0,018	0,049	14,59	16,52	57,76	0,020	0,022	0,078
HBE-Ter	32,57	29,10	38,10	0,190	0,170	0,223	61,74	50,58	72,34	0,361	0,296	0,423
HBS	122,90	118,95	116,87	0,021	0,020	0,020	235,35	212,27	222,78	0,040	0,036	0,038
HBR	105,33	99,34	229,67	0,028	0,026	0,060	200,42	187,24	424,63	0,053	0,049	0,112
HBO	85,20	63,57	200,43	0,024	0,018	0,056	103,90	71,04	368,32	0,029	0,020	0,104
average	56,75	50,27	94,22	0,041	0,037	0,062	95,653	81,665	167,896	0,072	0,061	0,110

table 19: mean absolute- and weighted trip production error per trip purpose

Consider for instance HBW-Blue model 1, indicating that each cell has an average error of 16,22 trips, which is as much as 1% of the total observed trips and should be considered as good. HBE-tertiary is a matter of concern though, since the three models produce an error between 17% en 23%. Taking frequency of the levels into account this even becomes an error between 30% en 42%. Overall we can conclude model 2 is the better predictor of the observed trips and its ability to distribute the expected number of trips over the different attribute levels, by looking at the average values. This verifies the reconsideration of the old model and proves fewer variables are able to calculate an expected average trip rate for a household with particular attributes as well (given enough households).

8.2.2 Cross classification trip rates

At first sight, I must say that all three models do an overall pretty good job in calculating trip production rates in the cross classification tables. This confirms that the use of stratified dummy variables (as a replacement for the more conventionally adopted “category analysis”) is a suitable manner for trip production at the disaggregate household level. On the other hand, no clear patterns arise of one model constantly performing better than the other, which makes selection of the best fitting model more difficult. In table 16 and table 17 as presented in chapter 7, the best trip rate estimators are collard *blue*. To get a better insight in the performance and comparison between the different models the mean trip rate error for the different categories is calculated. For that reason all category trip rate errors (this is the difference between the observed and predicted trip rates for each individual household category/type) are calculated. I computed two different mean errors, namely:

- mean absolute error $\mu_{abs} = \frac{\sum \mu_i}{N}$ (in which $\sum \mu_i$ is the sum of all trip rate errors per household category divided by the number of errors *N*).

This value represents the average trip rate error per type of household. For instance (see table 17), the trip rate error for model 1 in the category: 2 cars, 1 worker, 0 dependants = | 2,200 – 2,339 | = 0,139. So, the mean absolute error of model 1 for the cross classification table with car ownership = 2, is calculated as $\mu_{abs} = \frac{|2{,}200 - 2{,}339| + ... + |14{,}933 - 13{,}046|}{17} = 0{,}64$. This value indicates how accurate the model predicts the trip rates per type of household on average. Note that each type of household has an equal influence on this value, since in this case sample size of a particular category is not accounted for. It is more plausible to give more weight to the categories that occur more frequently. To account for this drawback, the following is presented:

- mean weighted error $\mu_{wgt} = \frac{\sum_{i=1}^{N} freq_i \cdot \mu_i}{\sum_{i=1}^{N} freq_i}$ (in which $\sum_{i=1}^{N} freq_i \cdot \mu_i$ is the sum of the household sample size per category times the trip rate error per household category, divided by the total sample size in the broad category $\sum_{i=1}^{N} freq_i$).

This value represents the average trip rate error per household. For instance, the weighted error ($freq_i \bullet \mu_i$) for model 1 in the category: 2 cars, 1 worker, 0 dependants = | 2,200 – 2,339 | * 35 = 4,865. So, the mean weighted trip rate error of the broad category classification car ownership = 2 is:

- mean weighted error $\mu_{wgt} = \frac{(|2{,}200 \;\; 2{,}339| \cdot 35) + ... + (|14{,}933 - 13{,}046| \cdot 15)}{1366} = 0{,}363$.

So, the mean weighted error also evaluates the frequency (sample size) of the different household types in order to compute the mean error per household.

The absolute- and weighted trip rate errors are calculated for the 80 household categories incorporated in the model – i.e. workers (4) x dependants (5) x cars (4). Eventually the mean absolute- and mean weighted trip rate errors of the four cross classification tables, are presented below:

	mean absolute trip rate error			mean weighted trip rate error		
classification	model 1	model 2	old model	model 1	model 2	old model
car ownership = 0	1,82	1,92	1,81	0,504	0,523	0,711
car ownership = 1	1,39	1,61	1,53	0,331	0,589	0,459
car ownership = 2	0,64	0,78	0,69	0,363	0,366	0,338
car ownership = 3+	1,21	1,34	1,28	0,628	0,670	0,528
	1,266	1,412	1,329	0,456	0,537	0,509

table 20: mean absolute- and weighted trip rate error per classification category

The mean weighted trip rate error is a more valid result to compare than the mean absolute trip rate error since sample size is an important factor in this case to indicate the significance. Considering these errors, they indicate the trip rates average deviation per household. Apparently the old model (and model 1 with the same variables) containing more variables in the equation do a better job in calculating trip rates of cross-classified household types. This may be explained by the fact that more variables are able to influence the trip rates of the combinatorial household types. Overall it seems that model 1 is the best predictor of the observed trip production rates per household. Surprisingly the old model, computed from the 1992 Melbourne data, still performs pretty well in comparison. Apparently households structures have not changed that much in 12 years.

9 Discussion

According to the literature it is said that the available count data with excess zeros violates some basic statistical assumptions of linear regression (as discussed before). All trip purposes have their stochastic error terms, not to be normally distributed as depicted by the histogram charts, residual plot sand normal p-p plots in Appendix 6. This indicates the normality assumption is not satisfied. In the residuals plot there appears to be increasing variance for most of the trip purposes. There are clear patterns in the data, the linear model is not accounting for. A plot of standardized residuals against standardized predicted values of the dependent variable should show a random pattern when linearity is present. The horizontal line represents zero error and ideally the errors should be randomly scattered around this line. Furthermore, the normal scores plots are clearly showing non-normality for all trip purposes. Therefore, I cannot assess their statistical reliability by the classical tests of significance (F and R^2) because the latter are based on normal distributions.

9.1 Log-linear model

A common technique in the transport modelling literature is to transform the dependent variable (number of trips) onto the log scale. This has the effect of 'normalizing' the distribution and tempering the effects of increasing variance. The histograms below show the trip rates transformed by first adding 1 (as it is not possible to take the log of zero) and then by taking the 10log. In this case, it makes the spike at zero even more prominent, and so it is unlikely it will improve the model at all. Other transformations are possible, such as taking square roots, but there is no transformation that can spread out the zeros, and so non would significantly improve the model overall.

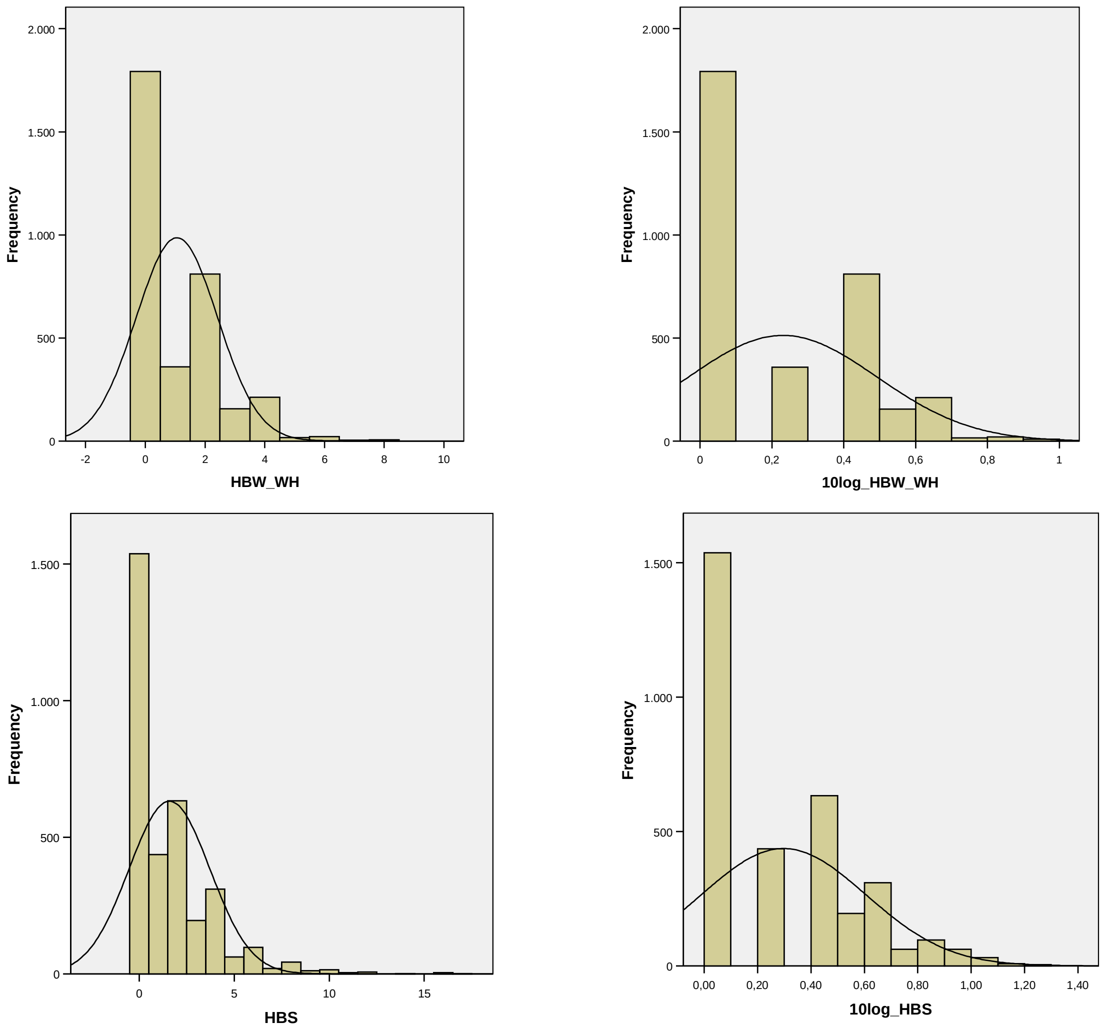


figure 10: Log-linear histograms of transformed trip rates of HBW_White and HBS

One of the difficulties in comparing different model formulations lays in the measures of fit. For example, while normal linear models can be compared using adjusted R-squared values (as well as for accounting for the significance of individual predictors) it is not sensible to compare R-squared values across models such as linear versus log-linear, nor is it even possible with different formulations of generalised linear models. *The term R-squared refers to the fraction of variance explained by a model, but what is the relevant variance that demands explanation? There are many transformations that may be applied to a variable before it is used as a dependent variable in a regression model: deflation, logging, seasonal adjustment, differencing. All of these transformations will change the variance and may also change the units in which variance is measured. Therefore, if the dependent variable in the regression model has already been transformed in some way, it is possible that much of the variance has already been "explained" merely by the choice of an appropriate transformation* [DUKE, 2007]. Another measure of fit however, Akaike's Information Criterion (AIC), might be a useful alternative. This value measures the distance from the true model (that is the observed data) and the approximating model, the larger the distance the worse the model [Burnham and Anderson, 2002]. AIC values can be compared across non-nested models as long as they are calculated from the same data. AIC is not an absolute measure of goodness of a model, rather it should be seen as a comparative measure between models. So it is the difference between AIC values that is important, not their absolute values.

9.2 Zero-inflated models

Count data with large number of zeros have been analysed in a number of fields. The zero inflated Poisson model described by Lambert (1992) is of a factory line which fluctuates between perfect and imperfect states. The perfect state producing no faults (that, all zeros) and the imperfect state producing some faults and this includes some zeros. The overall percentage of zeros in this example is 81% which is said to be clearly higher than can be explained by the Poisson distribution alone. The formula for calculating probabilities using the Poisson distribution is given by:

$$f(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

λ: mean number of successes in a given period
x: number of successes we are interested in
e: base of the natural logarithmic function ln

For example, a Poisson distribution with a mean of just 1 predicts only 37% zeros, mean of 2 predict 14% zeros and a mean of 3 predicts 5% zeros. For the trip production data the percentage of zeros by trip purpose is as follows:

Trip Purpose	% zero
Blue	78
White	56
Education Primary	86
Education Secondary	90
Education Tertiary	96
Shopping	43
Recreation	61
Other	68

table 21: Percentage of zeros for each trip purpose in the whole dataset

As with the study of defects in a factory line, there are also two types of zeros in trip generation. Zeros can result when a trip purpose is not applicable to any member of the household, and also when no member of the household chooses to make a trip for a particular purpose. For example, a house with no blue collar workers is most likely not to make any blue collar work trips. However, when a household does have blue collar workers there may still be zero trips produced on that day due to illness, personal holiday or for some other reason. In the case of shopping and recreation there is no category which excludes these trips, so zeros are solely do to choice. Thus a possible model has two components, the probability that no trips are made which may include the 'not applicable' groups and the 'choice' groups, and the distribution of trips modelled by a Poisson distribution.

For a zero-inflated Poisson regression Lambert (2002) and Jang (2005) describe the model as follows: the responses **Y** = $(Y_1,....Y_n)$ are assumed independent and

$y_i = 0$ with probability φ_i
$y_i \sim Poission(\lambda_i)$ with probability 1 - φ_i

So that,

$$P(Y_i = 0) = \varphi_i + (1-\varphi_i)\exp(-\lambda_i)$$

$$P(Y_i = k) = (1-\varphi_i)\exp(-\lambda_i)\lambda_i^{\ k} / k!$$

And the expected value is given by $E(Y_i) = (1-\varphi_i)\lambda_i$, as the mean of the Poisson distribution is λ itself. In words, the expected number of trips is the probability that a trip is made times the mean trip rate as defined by the Poisson distribution. Further research should make clear whether this methodology works for trip generation.

10 Conclusion and Recommendations

This analysis has investigated the use of stratified dummy variable linear regression for predicting trips for various purposes by household characteristics, and finally proposed the zero-inflated regression technique as a possible alternative method in the future.

Although it became clear that stratified dummy linear regression can be used to calculate averages and the overall total trip predictions came close to observed values, the measures of fit and performance given are not completely reliable. Whether the model can be interpreted as a description of the relationship between trip making and household characteristics should be questioned as well. The experience gained with the investigation of the South East Queensland data would suggest the following concerns regarding the old VLC model:

1) Possible high cross-correlations between some variables
    a) Total persons and number of dependants 0-17
    b) Number of cars and number of workers
    c) Number of cars and total persons (household size)
2) Possible inclusion of insignificant and false relationships
    a) Almost all variables are used in the shopping recreation and other trips, experience suggests that a model using only the total persons in the household gives equivalent results
3) No measures of standard error or model fit were given of the existing model (although these would not be entirely trustworthy in any case)
    a) The given F-statistic suggests that there is substantially more error in the non-work trips than in the work trips, which is expected, but large amounts of error render the model useless and should have been reported. The errors eventually became clear while testing the old model on the newly available data.

Of particular concern is the statement in section 6.0 of VLC's technical note 2, saying that: *The ability of the trip production equations to replicate observed trips by broad attribute classifications appears to be excellent and no significant biases are evident in the model.* The fact that the model can replicate the same totals as the data it was created on is to be expected, it does not indicate that the model is useful for predictive purposes though. Consider that the models calculate an (expected) average trip rate for a household with particular attributes. Given enough households, we would always expect that the total number of trips will almost be equal to the average times the number of households in the sample. While the average may perform acceptably over a large enough sample, the smaller the sample (or zone) the greater the chance of deviation from the average since the model does not capture the underlying structure of the data as shown in the diagnostics.

Since problems are said to be avoided by using the dummy variable regression approach, the model coefficients used in VLC's current trip production model were recalculated using the SEQHTS 2003/'04 data using both the variables in the original model and a simpler set of variables. It became clear that the old model seriously needs an update since it is not able to calculate reliable trip productions anymore and severe errors occur when distributing estimated trips over the different attribute groups. Both new sets of models came up with better results compared to the old model and eventually the second model, using the stepwise regression analysis, is preferred due to its model simplicity and overall performance.

For now, I would suggest to implement the updated coefficients of model 2 in the existing trip generation model. Making use of this updated model though, is only suggested if one decides to keep working on the same way. In the end I would suggest to investigate the use of other model formulations such as a zero-inflated Poisson model. The ease of the current model formulation though, is that it uses aggregated census data in the linear regression equations to predict total trips for each zone. For the zone we can sum up the attributes of all the households to get a zonal total in each category. Applying at the zonal level then gives the same answer as doing each household individually and summing them. This is very simple and effective, but unfortunately not possible with the non-linear model I just suggested. It is not possible to simply input the totals of households into non-linear equations so some form of simulation is necessary to replicate household data and calculate all predictions at the disaggregate level.

Another reason why a move to a more comprehensive model would be justified is the fact that the linear model provides mean estimates that are based on a large random heterogeneous sample of South East Queensland. The averages from small homogeneous zones will possibly deviate from this overall mean. The smaller and more homogeneous the zones are, the greater deviations from the large sample means can be expected.

11 References

[Burnham and Anderson, 2002]	Kenneth P. Burnham; David R. Anderson (2002). “Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach”, Second edition, Springer Verlag, New York.
[Cheung, 2002]	Cheung, Y.B. (2002). “Zero-inflated models for regression analysis of count data: a study of growth and development.” *Statistics in Medicine,* 1461-1469
[Cohen et al, 2003]	Cohen. J. (2003). “Applied multiple regression/correlation analysis for the behavioral sciences.” Mahwah, NJ Lawrence Erlbaum Associates
[Hall, 2000]	Hall, D.B. (2000). “Zero-inflated Poisson and Binomial regression with random effects: a case study.” *Biometrics,* 56 1030-1039
[Hardy, 1995]	Hardy, Melissa A. (1995) “Regression with dummy variables.” *Newbury Park, CA Sage Publications*
[Hellerstein, 1991]	Hellerstein, D.M. (1991). “Using count data models in travel cost analysis with aggregate data.” *American Journal of Agricultural Economics,* 73(3), 860-866
[Huet et al., 2004]	Huet, S., Bouvier, A., Poursat, M., and Joliet, E. (2004). *Statistical tools for nonlinear regression: A practical guide with S-Plus and R examples.* Springer, New York
[Immers & Stada, 1998]	L.H. Immers, J.E. Stada (1998); *Verkeersmodellen*; Dictaat Katholieke Universiteit Leuven;
[Jang, 2005]	Tae Youn Jang (2005); “Count Data Models for Trip Generation” *Journal of transportation engineering vol. 131 (2005), issue 6,* page 444-450 (7)
[Kynn, 2005]	Kynn M. (2005); “Analysis of the trip production equations in the BSTM”
[Lambert, 1992]	Lambert, D. (1992). “Zero-inflated Poisson regression, with an application to defects in manufacturing.” *Technometrics,* 34(1), 1-14.
[Lewsey and Thomson, 2004]	Lewsey, J.D. and Thomson, W.M. (2004). “The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross sectional and longitudinal DMF data examining the effect of socio-economic status. “*Community Dentistry and Oral Epidemiology,* 32 183-189
[Monzon et al, 1989]	Monzon, Jose, Konstadinos G. Goulias, Ryuichi Kitamura (1989) Trip Generation Models for Infrequent Trips. *Transportation Research Record* (1220), 40 - 46
[McCullagh,1989]	McCullagh, P. and Nelder, J.A. (1989). “Generalized Linear Models”, CRC Press
[Ortúzar and Willumsen, 2004]	Ortuzar, J. de D., and Willumsen, L.G. (2002). *Modelling Transport (Third Edition)*. John Wiley and Sons, Inc, Chicester, England

[OVO, 2000] Verschuren, P., Doorewaard, H. (2000), *Het ontwerpen van een onderzoek.* Utrecht: Uitgeverij LEMMA

[Rengaraju and Satyakumar, 1995] Rengaraju, V.R., and Satyakumar, M. (1995). “Three-dimensional category analysis using probabilistic approach.” *Journal of Transportation Engineering,* 121(6), 538-543

[SCOTT, 1997] Scott, J. Long (1997) “Regression models for categorical and limited dependent variables.” Thousand Oaks, California

[SEQIPP, 2007] The South East Queensland Infrastructure Plan and Program 2005 – 2026, (2007).

[Stopher and McDonald, 1983] Stopher, P.R., and McDonald, K.G. (1983) “Trip generation by cross-classification: An alternative methodology.” *Transportation Research Record,* 944 84-91.

[Washington, 2000] Washington, S. (2000). “Iteratively specified tree-based regression: theory and trip generation example.” *Journal of Transportation Engineering,* 126(6), 482-491

[VLCTN1, 1997] Veitch Lister Consulting Pty. Ltd (1997) *Calibration of an Integrated Travel Forecasting Model: Technical note 1 The Household Segmentation Model*

[VLCTN2, 1997] Veitch Lister Consulting Pty. Ltd (1997) *Calibration of an Integrated Travel Forecasting Model: Technical note 2 The Household Trip Production Model*

[VLCTN3, 1997] Veitch Lister Consulting Pty. Ltd (1997) *Calibration of an Integrated Travel Forecasting Model: Technical note 3 The Travel Market Segmentation Model and Vehicle Occupancy Model*

[VLCTN4, 1997] Veitch Lister Consulting Pty. Ltd (1997) *Calibration of an Integrated Travel Forecasting Model: Technical note 4 The Trip Attraction Model and Non-Home Based Trip Production Model*

Websites:

[WIKIQLD, 2007] South East Queensland description
http://en.wikipedia.org/wiki/South_East_Queensland
Obtained: May 2007

[VLC Company Profile, 2007] Veitch Lister Consulting information;
http://www.veitchlister.com.au/
Obtained: July 2007

[DUKE] Information regarding R-squared:
http://www.duke.edu/~rnau/rsquared.htm
Obtained: January 2008

[NCSU1] Linearity assumptions;
http://www2.chass.ncsu.edu/garson/PA765/assumpt.htm#linearity
Obtained: January 2008

[NCSU2] Dummy variable regression information
http://www2.chass.ncsu.edu/garson/PA765/regress.htm#dummy

Obtained: January 2008

[NCSU3] Multiple regression analysis information
http://www2.chass.ncsu.edu/garson/PA765/regress.htm
Obtained: July 2007

[NLREG] Understanding the results of an analysis
http://www.nlreg.com/results.htm
Obtained: December 2007

[COS] Poisson distribution:
http://infinity.cos.edu/faculty/woodbury/Stats/Tutorial/Pois_Form.htm
Obtained: January 2008

[GRAPHPAD] Article regarding multicollineartiy
http://www.graphpad.com/articles/Multicollinearity.htm
Obtained: December 2007

[COVENTRY] Information regarding Linear regression and assumptions
http://www.coventry.ac.uk/ec/~nhunt/regress/ass1.html
Obtained: November 2007

Appendix 1 Household Travel Survey Summary

This summary is a brief description of the procedures involved in the design, conduct and analysis of the South East Queensland Travel Survey (SEQTS).

1.1 Why was the Survey Needed?

The Transport Portfolio and other transport planning agencies are required to make decisions on transport infrastructure and services worth billions of dollars. The decision making process for transport planning needs to be informed, accountable and founded on comprehensive, current and reliable data. One of the most important areas of information needed is an accurate description of travel behaviour of the people living in an area. The most effective way for transport planners to gather this information is by the conduct of a household travel survey. The last time a survey of this nature was conducted in South-East Queensland was 1992/'94. This data is now dated and hence it was decided to conduct a new SEQTS.

1.2 What are the Survey Objectives?

Two key objectives for the SEQTS data have been identified:

- As a primary source of information for the development of transport analysis tools and models for personal travel. This allows the estimation/assessment of the impacts and transport outcomes of changes to transport infrastructure, systems and services; and
- To understand and quantify travel behaviour. This allows monitoring of the use of the transport system and assists in assessing the success of transport infrastructure, systems and services. These found key planning, policy development and decision-making processes relating to the provision of transport infrastructure and services.

1.3 What was the scope of the Survey?

The scope of the survey covers:

- All occupied private residential households within Brisbane, the Gold Coast and the Sunshine Coast areas;
- All persons (including visitors) staying at these households on the night preceding the household's Travel Day;
- Travel made by persons aged 5 and above on weekdays during the survey period.

The survey and collation of results was carried out by The Urban Transport Institute (TUTI), as the primary survey contractor.

1.4 Definition of Terms

Study Area	Brisbane, the Gold Coast and the Sunshine Coast
Occupied Private Dwelling	This term refers to a private residence which was occupied at the time of the survey
Trip Stage	A one-way travel movement from an origin to a destination for a single purpose (including change of mode) and by a single mode
Stop	A place where an activity (including change of mode) is undertaken.
Trip	A one-way travel movement from an origin to a destination for a single purpose (including picking up and delivering passengers), but perhaps by multiple modes.
Journey to work	The first occasion during the day when a person leaves their home place and arrives at their workplace (perhaps including other stops during the journey).
Journey from work	The last occasion during the day when a person leaves their workplace and arrives at their home place (perhaps including other stops during the journey).

Trip Chain	A sequence of trips which starts at one place and eventually returns to the same place.
Home-based Trip Chain	A sequence of trips which starts at home and returns to home.
Work-based Trip Chain	A sequence of trips which starts at work and returns to work.
Gross Sample	The complete list of household addresses drawn from the sample frame.
Sample Loss	Those addresses in the Gross Sample from which a response could not reasonably be expected (examples include vacant blocks, commercial premises, demolished houses, vacant houses, and houses that were unoccupied during the period of the survey).
Net Sample	The Gross Sample minus the Sample Loss.
Acceptable Household Return	An acceptable household return is a household return that when processed (including consistency checks) and delivered to the Principal has at least 95% of all data items completed and all key data items completed.
Response Rate	The number of Acceptable Household Returns divided by the Net Sample.
Key Data Items	The following are the Key Data Items for the survey:

Household Data
- Household ID number
- Date of Travel Day
- Household Address (geocodes randomised to prevent identification)
- Dwelling Type
- Number of persons in household
- Number of household vehicles (cars, motorcycles, vans/trucks)

Vehicle Data
- Vehicle ID number
- Year of manufacture
- Body type
- Fuel type
- Ownership

Person Data
- Person ID number
- Year of birth
- Student status

Travel Data (based on recording travel by trip stages)
- Start-of-day Location
- If no travel undertaken, reason for no travel
- Trip Stage ID number
- Trip Stage Starting Time
- Trip Stage Destination Arrival Time
- Trip Stage Destination Place-Type
- Trip Stage Destination Activity
- Trip Stage Destination Location
- Mode Used
- Departure Time from Destination

1.5 Time period and Methodology

Initially, the intention was to conduct the Brisbane survey in one period in the second half of 2003. However, given the need to conduct a Pilot Survey prior to the Main Survey, it was agreed that the survey period would need to be split into two periods; the first eight weeks in Oct-Nov 2003 and the second eight weeks in Feb-Mar 2004. The Coastal surveys were conducted in one eight week block in Oct-Dec 2004. The SEQTS methodology is based on a self-completion questionnaire, which is hand-delivered to, and hand-collected from the survey households. This process is also supplemented by telephone motivational calls, telephone and postal reminders, and telephone clarification calls. The 1992 SEQHTS survey used a mailout-mailback self-completion questionnaire, and achieved response rates in the mid-60% range. Over the past ten years, however, it has been observed in many different cities that response rates have been declining. Since the Brief specified a response rate of 60%, it was considered that this would be difficult to obtain, considering the type of data required, using the same mailout-mailback methodology.

1.6 Target population

The target population for SEQTS was all residents and visitors in occupied private dwellings in Brisbane, the Gold Coast and the Sunshine Coast on the night before the specified Travel Day. See figure 11*,* for an overview of the region. There are six aspects to the above definition of the target population, which involve specification of:

- Brisbane;
- The Gold Coast
- The Sunshine Coast;
- The specified Travel Day;
- Occupied Private Dwellings; and
- Residents and visitors.

Brisbane was defined as the Brisbane Statistical Division, as specified by the Australian Bureau of Statistics, and is an area around Brisbane bounded by Caboolture in the north, Ipswich in the west and Beenleigh in the south.

The Gold Coast was defined as the Statistical Sub-Division of Gold Coast City Part B. In the actual Coastal survey conducted in Oct-Dec 2004, areas within Gold Coast City Part A in the southern part of the Brisbane Statistical Division were also surveyed, but in the final data set these households are included within the Brisbane area.

The Sunshine Coast was defined as the Statistical Sub-Division of Sunshine Coast, plus the Statistical Local Areas of Noosa Balance, Maroochy Balance, Caloundra – Hinterland and Caloundra – Rail Corridor.

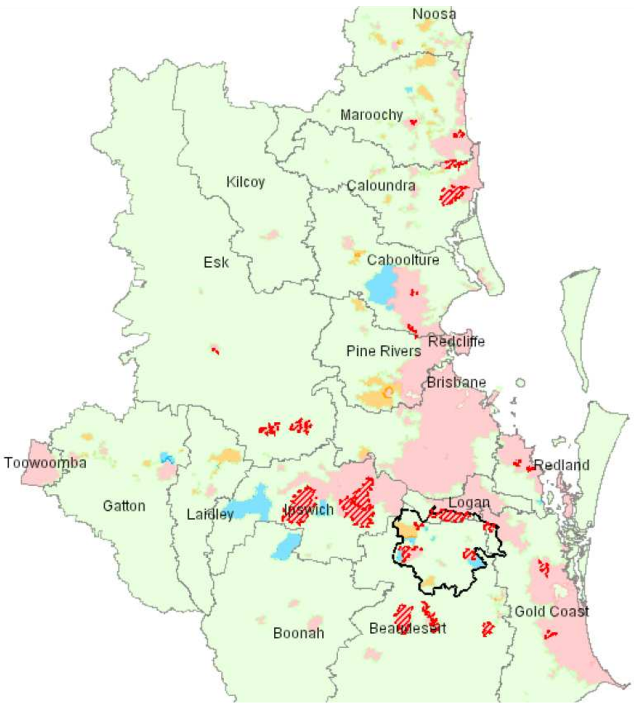


figure 11: Research Area

The specified Travel Day for each household was initially obtained by uniformly spreading the sample of Brisbane households over the sixteen weeks of weekdays (80 days in total) for the Brisbane survey. Following the completion of the first eight weeks of the survey, however, the Principal decided to increase the sample size for the second eight weeks from 1750 responding households to 2250 responding households for the second eight weeks. After the sample for the second eight weeks of the survey was increased, a uniform number of households were added to each of the 40 days in the second eight weeks of the Brisbane survey. The specified Travel Day for each household in the

Coastal surveys was obtained by uniformly spreading the sample of Coastal households over the eight weeks of weekdays (40 days in total) for the Coastal survey.

Occupied Private Dwellings consisted of those private residential addresses in the Study Area which were occupied on the night before the Travel Day.

Finally, *residents* were defined as those people who normally lived at the residential address (i.e. including those who were temporarily away such as people interstate for a few days, but excluding those who where usually away such as children at boarding school). *Visitors* were defined as those who slept overnight at the address on the night before the Travel Day.

For each area (Census Collection District or CCD) specified by TUTI, Energex (the electricity authority) provided a specified number of residential addresses (i.e. electricity connections on a residential rate) with the following details:

- CCD number
- Property name
- Lot number
- Flat number
- Street number
- Street name
- Locality (suburb)
- Postcode

The Census Collection District (CDD) is the smallest geographic unit of collection. Generally defined as an area that one Collector can comfortably cover delivering and collecting Census forms, there are on average around 225 dwellings per CCD. The total number of CCD's in the research area is 4184. It is impossible and unwanted to cover each CCD in the survey. Therefore a fraction of representative sample CCD's has been taken. In total the survey was conducted within 320 CCD's.

1.7 Sample size and composition

The sample size for the Brisbane survey was specified by Queensland Transport as 3500 acceptable household responses in the Brisbane Statistical Division. This was later increased to 4000 acceptable household responses. The sample sizes for the Coastal surveys were specified as 1400 acceptable household responses on the Gold Coast and 1110 acceptable household responses on the Sunshine Coast.

1.8 Interpretation of the results

There are a number of limitations that should be borne in mind when using and interpreting the data in SEQTS, as follows:

- The travel data is only for weekdays; no travel diaries have been completed on weekends.
- The travel data is for a limited period of the year; diaries were only completed in October, November, early-December, February and March. Thus, school holiday periods were largely avoided, and only limited periods when Universities were not in term.
- Travel diaries for children under five have been re-constructed from the diaries of other household members, and only record travel undertaken with at least one other member of the household; travel undertaken with no other members of the household is not recorded for children under 5.
- Multi-purpose stops within regional shopping centres (e.g. Westfield Indooroopilly) have been simplified to a single trip to the shopping centre, irrespective of the number of different activities undertaken while at the shopping centre.
- Only personal travel, including travel to and from work, of professional drivers was collected.
- As always, the SEQTS data is from a limited sample of about 4000 households in Brisbane, 1500 on the Gold Coast and 1300 on the Sunshine Coast; care should be taken when reporting analyses undertaken with significant segmentation of the data, since the sample sizes within some strata may be relatively low.

1.9 Summary

A summary of the entire survey data appears in table 22 below:

Survey Name:	South-East Queensland Travel Survey 2003-2004 (SEQTS)
Description:	A survey of day-to-day travel behaviour of persons living in a sample of private dwellings in South-East Queensland (Brisbane, Gold Coast and Sunshine Coast). Also includes some household characteristics.
Coverage:	Brisbane Statistical Division, Gold Coast City Part B Statistical Sub-Division, Sunshine Coast Statistical Sub-Division, Statistical Local Areas of Noosa (S) Balance, Maroochy (S) Balance, Caloundra (C) – Hinterland, and Caloundra (C) – Rail Corridor
Project Duration:	July 2003 – September 2005
Pilot Survey:	Dress-rehearsal of all survey stages for sample of 125 responding households in Brisbane in August 2003; Dress-rehearsal of all survey stages for sample of 120 responding households in Gold Coast and Sunshine Coast in July 2004
Main Survey Duration:	Brisbane: 16 weeks (Oct-Dec 2003, Feb-Mar 2004) Coasts: 8 weeks (Oct-Dec, 2004)
Target Sample Size:	Brisbane:4000 responding households Gold Coast: 1400 responding households Sunshine Coast: 1110 responding households
Actual Sample Size:	Brisbane (including those in Gold Coast City Part A surveyed in Coastal Surveys): 4115 responding households, 11091 persons, 6859 vehicles, 41610 stops (trip stages), 36426 trips, 3551 journeys-to-work, 3470 journeys-from-work Gold Coast (in Gold Coast City Part B) 1473 responding households, 3763 persons, 2498 vehicles, 13024 stops (trip stages), 12210 trips, 1121 journeys-to-work, 1093 journeys-from-work Sunshine Coast 1390 responding households, 3340 persons, 2219 vehicles, 12087 stops (trip stages), 11434 trips, 794 journeys-to-work, 788 journeys-from-work
Response Rate:	Brisbane: 60% Gold Coast: 55% Sunshine Coast: 62%
Sampling Method:	Multi-stage, variable-proportion, clustered sampling of household addresses within Census Collection Districts (CCD) within 11 regions of Brisbane, 10 regions of Gold Coast and 8 regions of Sunshine Coast
Survey Methodology:	- Self-completion questionnaires with stage-based one-day travel diaries; - Pre-contact Letter of Introduction from Client; - Personal Delivery of Questionnaires; - Motivational Phone call on evening before travel day; - Personal Collection of Questionnaires, with option of reply-paid mailback for those not contacted; - Non-respondent questions for refusals; - Reminder Phone Call after one week; - Reminder Letter for those not contactable by phone
Data Processing:	- Field Office visual check of returns; - Electronic scanning of questionnaires; - Geocoding of all destination locations; - Clarification Calls to households to clarify information.

table 22: SEQTS summary

Appendix 2 Dummy variable regression

In regression analysis we sometimes need to modify the form of non-numeric variables, for example sex, or marital status, to allow their effects to be included in the regression model. This can be done through the creation of dummy variables whose role it is to identify each level of the original variables separately. This is also the case in VLC's trip generation model which is discussed below.

2.1 VLC's stratified dummy variable regression

The development of a regression model for production can be based on aggregated data per zone or on disaggregated data per household. With models based on zonal data one tries to predict the total number of productions by a zone with social-economic data that characterise the entire zone. This method is only recommended in case one has only got aggregated zonal data. In case more data is available, for instance on household level or person level, aggregation will cause loss of information. A model based on household data is preferred since this makes the modal independent of the zonal classification. Furthermore the loss of information that occurs by aggregation of the data is reduced. The household is normally taken as unit instead of the person individual. Reason is that it is the household structure that determines the production rather than the individual person characteristics [Immers & Stada, 1998].

VLC uses a "*stratified dummy variable regression technique"* for the prediction of zonal person trip productions. The trip production models are disaggregate in the sense that they are calibrated, and could be applied at the household level. The model considers household attributes such as household size, workers within the household, dependants (i.e. non full-time working inhabitants) within the household, and car ownership. Each of these variables is segmented into several discrete levels which are represented by dummy variables. Each of the variables is considered independently, so that no cross-classifications are included. Advantages of *not* including cross-classifications include that the model is simpler to implement, data requirements are much less demanding, and less data is required for calibration of the parameters. The obvious disadvantage is that cross-classification effects are ignored. However, when deriving trip generation parameters in Melbourne, Australia, using an extensive household travel database, VLC found that cross-classification effects were negligible, supporting the case for the "stratified dummy variable" technique.
A major benefit of the "stratified dummy variable" technique is that by also implementing a relatively simple household segmentation model, it is possible to supply only "aggregate" zonal data to the model – variables like average zonal household size, average workers per household, average car ownership etc. This means that the technique has the rigour of a disaggregate model, but the data requirements of an aggregate model [VLCTN2, 1997].

A series of additive linear relationships (one for each trip purpose) have been derived that describe the number of person trips made by households of varying characteristics on a typical weekday. The generalised form of the relationship is as follows:

$$HPT_{purp} = \alpha_0 + SAG_1 + SAG_2 + ...SAG_n \qquad (1)$$

Where:

$HPT_{purp} =$	weekday person trips of a household for a particular purpose;
$\alpha_0 =$	a calibration coefficient;
$SAG_1 \ to \ SAG_n =$	stratified household attribute groups.

As mentioned in paragraph 5.6 there are seven stratified household attribute groups (SAG) in case of the trip production model. Each *SAG* in itself is an additive linear function. For example, the car ownership attribute group is of the following form:

$$SAG_{car} = c_0 CAR_0 + c_1 CAR_1 + c_2 CAR_2 + c_3 CAR_3 \qquad (2)$$

Where:

$SAG_{car} =$ the group of attribute variables describing the car ownership level of a household;

$c_n =$ calibration coefficients;

$CAR_n =$ dummy variables representing levels of car ownership.

The other stratified attribute groups are similarly structured, but the number of stratifications may vary – for example, the household size attribution is stratified into six variables. See paragraph 5.6 for the number of stratifications per attribute group. Each attribute variable within the SAG's is treated as a "dummy" variable - i.e. the only value it can have is zero or one (0 or 1). For a particular household a dummy attribute variable will equal unity if that variable describes the household - for example, if the household has two cars the CAR_2 variable will equal "1" - and will be set to zero if it falsely describes the household. Consequently only one attribute variable will equal unity within each attribute group, and all others will have the value zero. Therefore only seven of the 30 attribute variables (6 person, 4 car, 4 blue-collar worker, 4 white-collar worker, 5 (0-17) dependants, 4 (18-64) dependants and 3 (65+) dependants) will be non-zero - one in each SAG .

For example, a household that contains two employed residents (one blue-collar worker and one white-collar worker), four dependants (one child under 17, two young adults between 18 and 64 and a retiree over 65) and owns three cars will be estimated to make the following person trips on a typical weekday by combining equation (1) en (2).

$$Trips_{purp} = a_0 + h_6 HOUSEHOLD_6 + bc_1 BLUE\_COLLAR_1 + wc_1 WHITE\_COLLAR_1 +$$
$$dc_1 DEPEND\_CHILD_1 + da_2 DEPEND\_ADULT_2 + dr_1 DEPEND\_RETIREE_1 + c_3 CAR_3$$

As the above attribute variables have values of unity, the above equation therefore simplifies to:-

$$Trips_{prup} = a_0 + h_6 + bc_1 + wc_1 + dc_1 + da_2 + dr_1 + c_3$$

All other calibration coefficients are excluded, as their attribute dummy variables are equal to zero. The unit of travel demand forecasting is therefore the household. So the model has the sensitivity of a disaggregate model. However, because of its linear additive form it can be simply applied at the zonal level, without the need to forecast the joint distribution of household attributes. This has major benefits as simpler, and more robust, market segmentation models can be adopted.

The thirty household attribute coefficients that make the trip production model are calibrated using a step-wise multiple linear regression technique. The unit of observation is the household, the dependent variables are the number of reported person trips made by the household for each trip purpose and the independent variables are the mentioned dummy variables, describing the household characteristics (attributes) [VLCTN2, 1997].

Appendix 3 Assumptions and descriptions

3.1 Multicollinearity

In some cases a model fits the data well, even though none of the X variables has a statistically significant impact on predicting Y. How is this possible? When two X variables are highly correlated, they both convey essentially the same information. In this case, neither may contribute significantly to the model after the other one is included. But together they contribute a lot. If you removed both variables from the model, the fit would be much worse. So the overall model fits the data well, but neither X variable makes a significant contribution when it is added to your model last. When this happens, the X variables are collinear and the results show multicollinearity. To help assess multicollinearity, SPSS tells you how well each independent (X) variable is predicted from the other X variables. The results are shown as an individual a Variance Inflation Factor (VIF). When the VIF values are high for any of the X variables, the fit is affected by multicollinearity.

What to do about multicollinearity? The best solution is to understand the cause of multicollinearity and remove it. Multicollinearity occurs because two (or more) variables are related – they measure essentially the same thing. If one of the variables doesn't seem logically essential to your model, removing it may reduce or eliminate multicollinearity. Or perhaps you can find a way to combine the variables. For example, if height and weight are collinear independent variables, perhaps it would make scientific sense to remove height and weight from the model, and use surface area (calculated from height and weight) instead [GRAPHPAD].

3.2 Linear regression

In linear regression analysis it is assumed that the relationship between y and x can be expressed in the form:

$y = (a + b x) + e$

where e represents the unpredictable element in y due to random variation or measurement error. This random element explains why different values of y are obtained for the same value of x. Without this random element there would be no need for regression analysis since the value of y would be perfectly predictable from the value of x. Because of the random element you can only estimate the values of a and b by the intercept a and slope b of the line of best fit. Hence any predictions you make based on the line are also only estimates. To be able to quantify the reliability of these estimates you must make the following assumptions [COVENTRY]*:*

- *the assumption of linearity*
- *the assumption of independence*
- *the assumption of constant variance*
- *the assumption of normality*

3.3 Correlation matrix

	Hhsize	Cars	Blue_CW	White_CW	Dep_0_17	Dep_18_64	Dep_65+	HBW-Blue	HBW-White	HBE-PrePrim	HBE-Sec	HBE-Ter	HBS	HBR	HBO
Hhsize	1	0,496	0,26	0,386	0,804	0,272	-0,241	0,207	0,238	0,502	0,408	0,119	0,219	0,234	0,559
Cars	0,496	1	0,337	0,43	0,167	0,119	-0,255	0,289	0,339	0,072	0,124	0,093	0,12	0,138	0,148
Blue_CW	0,26	0,337	1	-0,129	0,077	-0,085	-0,257	0,775	-0,093	0,054	0,063	-0,016	-0,008	-0,015	0,054
White_CW	0,386	0,43	-0,129	1	0,165	-0,233	-0,436	-0,096	0,698	0,072	0,108	0,034	0,001	0,077	0,15
Dep_0_17	0,804	0,167	0,077	0,165	1	0,078	-0,275	0,051	0,043	0,649	0,453	-0,014	0,118	0,195	0,621
Dep_18_64	0,272	0,119	-0,085	-0,233	0,078	1	-0,228	-0,042	-0,099	0,038	0,057	0,283	0,182	0,091	0,112
Dep_65_Plus	-0,241	-0,255	-0,257	-0,436	-0,275	-0,228	1	-0,205	-0,32	-0,164	-0,128	-0,063	0,082	-0,007	-0,184
HBW-Blue_Trips	0,207	0,289	0,775	-0,096	0,051	-0,042	-0,205	1	-0,058	0,046	0,061	0,002	-0,045	-0,039	0,021
HBW-White_Trips	0,238	0,339	-0,093	0,698	0,043	-0,099	-0,32	-0,058	1	0,02	0,067	0,05	-0,092	0,007	0,038
HBE-PrePrim_Trips	0,502	0,072	0,054	0,072	0,649	0,038	-0,164	0,046	0,02	1	0,11	-0,017	0,009	0,085	0,399
HBE-Sec_Trips	0,408	0,124	0,063	0,108	0,453	0,057	-0,128	0,061	0,067	0,11	1	0,035	0,029	0,071	0,239
HBE-Ter_Trips	0,119	0,093	-0,016	0,034	-0,014	0,283	-0,063	0,002	0,05	-0,017	0,035	1	0,004	0,004	0,013
HBS_Trips	0,219	0,12	-0,008	0,001	0,118	0,182	0,082	-0,045	-0,092	0,009	0,029	0,004	1	0,11	0,046
HBR_Trips	0,234	0,138	-0,015	0,077	0,195	0,091	-0,007	-0,039	0,007	0,085	0,071	0,004	0,11	1	0,142
HBO_Trips	0,559	0,148	0,054	0,15	0,621	0,112	-0,184	0,021	0,038	0,399	0,239	0,013	0,046	0,142	1

table 23: Correlation matrix

Appendix 4 VLC model coefficients

Stratified Variable	Trip Purpose							
	HBW-Blue	HBW-White	HBE-Pre/Prim	HBE-Secondary	HBE-Tertiary	HBS	HBR	HBO
1 Person								-0,5965
2 Persons								-0,3906
3 Persons								-0,1091
4 Persons			0,0236	0,2302				0,1518
5 Persons			0,0125	0,5865				0,4049
6+ Persons			-0,3357	0,6961				
0 Cars					-0,0154	-0,4386	-0,3298	-0,2599
1 Cars					-0,0706	-0,3281	-0,1881	0,1149
2 Cars	0,0228				-0,0967	-0,0056	-0,034	0,0978
3+ Cars	0,042	0,1335						
0 Blue-Collar								0,1595
1 Blue-Collar	1,1988				0,0496	0,0108	0,0294	
2 Blue-Collar	2,3291				0,1001	0,2512	0,1307	
3+ Blue-Collar	3,4993				0,1621	0,6656	0,586	
0 White-Collar								
1 White-Collar		1,0901			0,0555	0,3122	0,2086	
2 White-Collar		2,0029			0,1254	0,6102	0,335	
3+ White-Collar		3,1713			0,295	1,5276	0,9218	
0 Dependants (0-17)		0,1095			0,0508			-5,5553
1 Dependants (0-17)			0,2791	0,4037		0,0861	0,2746	-4,2265
2 Dependants (0-17)			0,9994	0,2989		0,3933	0,5841	-2,7756
3 Dependants (0-17)			1,8182	0,0544		0,6437	0,6731	-1,0137
4+ Dependants (0-17)			3,1167	0,6844		0,9397	1,3044	
0 Dependants (18-64)								-0,1202
1 Dependants (18-64)					0,1239	0,6075	0,3369	
2 Dependants (18-64)					0,3603	1,3087	0,5833	
3+ Dependants (18-64)					0,8341	1,5262	0,5238	
0 Dependants (65+)								
1 Dependants (65+)						0,6861	0,4778	
2+ Dependants (65+)					0,0157	1,7200	0,7469	
Constant	0	0	0,0016	0	0	1,1056	0,3423	6,2497
F-Statistic	1528,1	1149,1	451,7	162	43,3	26,4	19,2	249,5

table 24: VLC Trip Generation Coefficients

Appendix 5 Model output

5.1 1st Regression model run

Home Based Work – Blue Collar

Model Summary

Model	R(a)	R Square(b)	Adjusted R Square (c)	Std. Error of the Estimate(e)
1	,824	,679	,679	,599

a) R is the correlation coefficient between the observed and predicted values of the dependent variable. It ranges in value from 0 to 1. A small value indicates that there is little or no linear relationship between the dependent variable and the independent variables.
b) For regression through the origin (the no-intercept model), R squared is the proportion of variation in the dependent variable explained by the regression model. It ranges in value from 0 to 1. Small values indicate that the model does not fit the data well. This CANNOT be compared to R Square for models which include an intercept.
c) The "adjusted R square" is an R^2 statistic adjusted for the number of parameters in the equation and the number of data observations. It is a more conservative estimate of the percent of variance explained.
e) The standard error of the estimate is a measure of the accuracy of predictions made with the regression line.

ANOVA

Model		Sum of Squares(a)	df	Mean Square	F(b)	Sig.
1	Regression	2562,527	5	512,505	1426,079	,000
	Residual	1211,473	3371	,359		
	Total	3774,000	3376			

a) A model with a large regression sum of squares in comparison to the residual sum of squares indicates that the model accounts for most of variation in the dependent variable. Very high residual sum of squares indicate that the model fails to explain a lot of the variation in the dependent variable, and you may want to look for additional factors that help account for a higher proportion of the variation in the dependent variable.
b) The F value statistics, test the overall significance of the regression model. Specifically, they test the null hypothesis that all of the regression coefficients are equal to zero. This tests the full model against a model with no variables and with the estimate of the dependent variable being the mean of the values of the dependent variable. The F value is the ratio of the mean regression sum of squares divided by the mean error (residual) sum of squares. Its value will range from zero to an arbitrarily large number. The significance value of the F statistic is the probability that the null hypothesis for the full model is true (i.e., that all of the regression coefficients are zero). For example, if Sig. has a value of 0.01 then there is 1 chance in 100 that all of the regression parameters are zero. If the significance value of the F statistic is small (smaller than say 0.05) then the independent variables do a good job explaining the variation in the dependent variable. The significance value of the F statistic is less than 0.05, which means that the variation explained by the model is not due to chance.

Coefficients

Model		Unstandardized Coefficients(a)		Standardized Coefficients(b)	t	Sig.(c)	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance(d)	VIF(e)
1	BLUE_1	1,228	,024	,564	50,426	,000	,762	1,313
	BLUE_2	2,653	,052	,524	51,170	,000	,910	1,099
	BLUE_3+	4,078	,162	,248	25,174	,000	,978	1,022
	CARS_2	,023	,018	,013	1,220	,222	,801	1,248
	CARS_3+	,081	,032	,027	2,583	,010	,843	1,187

a) Displays the regression coefficients with their standard errors.
b) To determine the relative importance of the significant predictors, look at the standardized coefficients. This is not important in dummy variable regression though since all variables are measured on the same scale and it is impossible to say something about relative importance.
c) The t statistic and its significance value are used to test the null hypothesis that the regression coefficient is zero (i.e. that there is no linear relationship between the dependent and independent variable). The t-statistic is computed by dividing the estimated value of the parameter by its standard error. This statistic is a measure of the likelihood that the actual value of the parameter is not zero. The larger the absolute value of t, the less likely that the actual value of the parameter could be zero. Sig. (often called p-value) is the probability of obtaining the estimated value of the parameter if the actual parameter value is zero.

The smaller the value of Sig., the more significant the parameter and the less likely that the actual parameter value is zero. For example, assume the estimated value of a parameter is 1.0 and its standard error is 0.7. Then the t value would be 1.43 (1.0/0.7). If the computed Sig. value was 0.05 then this indicates that there is only a 0.05 (5%) chance that the actual value of the parameter could be zero. When the significance level is small (less than 0,05), the coefficient is considered significant.
d) A statistic used to determine how much the independent variables are linearly related to one another (multicollinear). The proportion of a variable's variance not accounted for by other independent variables in the equation. A variable with very low tolerance contributes little information to a model, and can cause computational problems. When the tolerances are close to 0, there is high multicollinearity and the standard error of the regression coefficients will be inflated.
e) The reciprocal of the tolerance. As the variance inflation factor increases, so does the variance of the regression coefficient, making it an unstable estimate. Large VIF values are an indicator of multicollinearity. A variance inflation factor greater than 2 is usually considered problematic.

Residuals Statistics

	Minimum	Maximum	Mean	Std. Deviation(e)	N
Predicted Value (a)	,00	4,16	,44	,752	3376
Residual (b)	-4,160	5,347	-,002	,599	3376
Std. Predicted Value (c)	-,587	4,948	,000	1,000	3376
Std. Residual (d)	-6,939	8,920	-,004	,999	3376

a) This table displays statistics about the residuals and predicted values. The predicted value is the value predicted by the regression model.
b) A residual is the difference between the observed value of the dependent variable and the value predicted by the model. Residuals are estimates of the true errors in the model. If the model is appropriate for the data, the residuals should follow a normal distribution. A histogram or P-P plot of the residuals will help to check the assumption of normality of the error term.
c) Standardized predicted values are predicted values standardized to have mean 0 and standard deviation 1.
d) Similarly, standardized residuals are ordinary residuals divided by the sample standard deviation of the residuals and have mean 0 and standard deviation 1.
e) A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.
Residual statistics are not presented in the further output. The model diagnostics should be interpreted as a replacement.

Home Based Work – White Collar

Model Summary

Model	R	R Square(a)	Adjusted R Square	Std. Error of the Estimate
1	,825(b)	,681	,681	,976

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6857,098	5	1371,420	1440,248	,000(a)
	Residual	3209,902	3371	,952		
	Total	10067,000	3376			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	WHITE_1	1,070	,033	,369	32,849	,000	,751	1,332
	WHITE_2	2,075	,038	,586	54,195	,000	,808	1,238
	WHITE_3+	3,217	,104	,322	30,796	,000	,864	1,157
	Dep_0_17_0	,123	,025	,059	4,890	,000	,661	1,513
	CARS_3+	,298	,053	,062	5,615	,000	,787	1,270

Home Based Education – Pre- Primary School

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,659(a)	,435	,434	,784

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1593,779	7	227,683	370,163	,000(a)
	Residual	2070,998	3367	,615		
	Total	3664,777	3374			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	,004	,017		,259	,796		
	Dep_0_17_1	,361	,043	,114	8,317	,000	,894	1,119
	Dep_0_17_2	1,319	,058	,435	22,721	,000	,457	2,187
	Dep_0_17_3	2,304	,091	,490	25,381	,000	,450	2,222
	Dep_0_17_4+	3,354	,158	,373	21,227	,000	,543	1,842
	HHSIZE_4	-,048	,052	-,017	-,914	,361	,483	2,069
	HHSIZE_5	-,154	,080	-,037	-1,927	,054	,455	2,199
	HHSIZE_6+	-,235	,115	-,037	-2,040	,041	,503	1,988

Home Based Education – Secondary School

Model Summary

Model	R	R Square(a)	Adjusted R Square	Std. Error of the Estimate
1	,532(b)	,283	,281	,702

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	654,559	7	93,508	189,498	,000(a)
	Residual	1662,441	3369	,493		
	Total	2317,000(b)	3376			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Dep_0_17_1	,506	,037	,214	13,773	,000	,880	1,136
	Dep_0_17_2	,533	,051	,238	10,392	,000	,404	2,473
	Dep_0_17_3	,589	,081	,162	7,286	,000	,431	2,321
	Dep_0_17_4+	,962	,141	,136	6,807	,000	,537	1,861
	HHSIZE_4	,147	,046	,072	3,178	,001	,413	2,420
	HHSIZE_5	,417	,071	,130	5,839	,000	,428	2,335
	HHSIZE_6+	,502	,103	,102	4,882	,000	,491	2,036

Home Based Education – Tertiary School

Model Summary

Model	R	R Square(a)	Adjusted R Square	Std. Error of the Estimate
1	,430(b)	,185	,182	,444

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	150,666	14	10,762	54,545	,000(a)
	Residual	663,334	3362	,197		
	Total	814,000(b)	3376			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	BLUE_1	,010	,018	,010	,566	,572	,731	1,367
	BLUE_2	,018	,038	,008	,475	,635	,946	1,057
	BLUE_3+	-,029	,119	-,004	-,246	,806	,988	1,012
	WHITE_1	,014	,018	,016	,767	,443	,526	1,900
	WHITE_2	,104	,020	,103	5,155	,000	,606	1,650
	WHITE_3+	,065	,045	,023	1,429	,153	,944	1,059
	Dep_0_17_0	,044	,016	,073	2,695	,007	,329	3,036
	Dep_18_64_1	,144	,017	,153	8,424	,000	,733	1,364
	Dep_18_64_2	,441	,032	,226	13,972	,000	,925	1,081

	Dep_18_64_3+	1,418	,070	,318	20,310	,000	,987	1,013
	Dep_65_Plus_2+	,054	,032	,030	1,675	,094	,776	1,288
	CARS_0	-,037	,030	-,022	-1,230	,219	,760	1,316
	CARS_1	-,059	,019	-,075	-3,078	,002	,406	2,461
	CARS_2	-,073	,020	-,093	-3,703	,000	,385	2,600

Home Based Shopping

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,319(a)	,102	,097	2,026

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1564,883	18	86,938	21,171	,000(a)
	Residual	13785,726	3357	4,107		
	Total	15350,609	3375			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	,140	,209		,668	,504		
	BLUE_1	,349	,101	,069	3,442	,001	,657	1,522
	BLUE_2	,460	,200	,044	2,299	,022	,728	1,373
	BLUE_3+	,992	,560	,030	1,771	,077	,939	1,065
	WHITE_1	,517	,108	,116	4,809	,000	,460	2,175
	WHITE_2	,942	,146	,188	6,451	,000	,314	3,184
	WHITE_3+	1,435	,252	,115	5,695	,000	,660	1,515
	Dep_0_17_1	,514	,112	,079	4,598	,000	,902	1,109
	Dep_0_17_2	,680	,108	,110	6,278	,000	,876	1,142
	Dep_0_17_3	,806	,163	,084	4,946	,000	,933	1,072
	Dep_0_17_4+	1,439	,305	,078	4,724	,000	,975	1,025
	Dep_18_64_1	1,041	,094	,217	11,049	,000	,691	1,447
	Dep_18_64_2	1,650	,166	,189	9,948	,000	,744	1,343
	Dep_18_64_3+	2,256	,333	,116	6,773	,000	,914	1,094
	Dep_65_Plus_1	,963	,132	,154	7,288	,000	,597	1,675
	Dep_65_Plus_2+	1,935	,175	,237	11,038	,000	,582	1,719
	CARS_0	-,316	,189	-,041	-1,672	,095	,437	2,286
	CARS_1	,017	,142	,004	,123	,902	,252	3,961
	CARS_2	,028	,125	,006	,221	,825	,325	3,078

Home Based Recreation

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,287(a)	,082	,077	1,783

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	955,751	18	53,097	16,701	,000(a)
	Residual	10672,652	3357	3,179		
	Total	11628,403	3375			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	,622	,184		3,378	,001		
	BLUE_1	-,058	,089	-,013	-,654	,513	,657	1,522
	BLUE_2	-,032	,176	-,004	-,183	,855	,728	1,373
	BLUE_3+	-,121	,493	-,004	-,246	,806	,939	1,065
	WHITE_1	,209	,095	,054	2,211	,027	,460	2,175
	WHITE_2	,418	,129	,096	3,248	,001	,314	3,184
	WHITE_3+	,698	,222	,064	3,149	,002	,660	1,515
	Dep_0_17_1	,269	,098	,048	2,737	,006	,902	1,109
	Dep_0_17_2	,818	,095	,152	8,593	,000	,876	1,142
	Dep_0_17_3	1,356	,143	,162	9,463	,000	,933	1,072
	Dep_0_17_4+	1,194	,268	,075	4,456	,000	,975	1,025
	Dep_18_64_1	,320	,083	,077	3,856	,000	,691	1,447
	Dep_18_64_2	,430	,146	,056	2,944	,003	,744	1,343
	Dep_18_64_3+	1,306	,293	,077	4,455	,000	,914	1,094
	Dep_65_Plus_1	,320	,116	,059	2,751	,006	,597	1,675
	Dep_65_Plus_2+	,714	,154	,100	4,631	,000	,582	1,719
	CARS_0	-,635	,166	-,096	-3,826	,000	,437	2,286
	CARS_1	-,288	,125	-,076	-2,307	,021	,252	3,961
	CARS_2	-,094	,110	-,025	-,848	,396	,325	3,078

Home Based Other

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,615(a)	,379	,376	1,768

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6402,290	14	457,306	146,278	,000(a)
	Residual	10507,472	3361	3,126		
	Total	16909,762	3375			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	4,328	,283		15,299	,000		
	BLUE_0	,147	,072	,030	2,048	,041	,889	1,125
	Dep_0_17_0	-3,251	,376	-,681	-8,648	,000	,030	33,538
	Dep_0_17_1	-2,161	,367	-,317	-5,881	,000	,063	15,753
	Dep_0_17_2	-1,475	,353	-,227	-4,174	,000	,063	15,993
	Dep_0_17_3	-,406	,358	-,040	-1,135	,256	,147	6,790
	Dep_18_64_0	-,215	,069	-,046	-3,127	,002	,867	1,154
	HHSIZE_1	-1,261	,313	-,232	-4,031	,000	,056	17,888
	HHSIZE_2	-1,081	,297	-,233	-3,638	,000	,045	22,111
	HHSIZE_3	-,808	,278	-,132	-2,901	,004	,090	11,150
	HHSIZE_4	,039	,261	,006	,150	,881	,099	10,129
	HHSIZE_5	,177	,264	,020	,668	,504	,212	4,714
	CARS_0	,066	,160	,008	,415	,678	,464	2,156
	CARS_1	,431	,121	,094	3,548	,000	,262	3,816
	CARS_2	,401	,110	,087	3,636	,000	,320	3,129

5.2 2nd Regression model run (stepwise)

Stepwise regression Home based Work Blue Collar

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,576	,332	,332	,864
2	,784	,615	,614	,657
3	,824	,678	,678	,600
4	,824	,679	,678	,600

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1252,835	1	1252,835	1677,129	,000(a)
	Residual	2521,165	3375	,747		
	Total	3774,000	3376			
2	Regression	2319,611	2	1159,805	2690,602	,000(c)
	Residual	1454,389	3374	,431		
	Total	3774,000	3376			
3	Regression	2559,896	3	853,299	2370,619	,000(d)
	Residual	1214,104	3373	,360		
	Total	3774,000	3376			
4	Regression	2561,992	4	640,498	1781,968	,000(e)
	Residual	1212,008	3372	,359		
	Total	3774,000	3376			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	BLUE_1	1,255	,031	,576	40,953	,000	1,000	1,000
2	BLUE_1	1,255	,023	,576	53,911	,000	1,000	1,000
	BLUE_2	2,694	,054	,532	49,747	,000	1,000	1,000
3	BLUE_1	1,255	,021	,576	58,997	,000	1,000	1,000
	BLUE_2	2,694	,049	,532	54,440	,000	1,000	1,000
	BLUE_3+	4,143	,160	,252	25,837	,000	1,000	1,000
4	BLUE_1	1,241	,022	,570	56,206	,000	,927	1,078
	BLUE_2	2,665	,051	,526	52,416	,000	,946	1,057
	BLUE_3+	4,089	,162	,249	25,279	,000	,981	1,019
	CARS_3+	,075	,031	,025	2,415	,016	,866	1,155

Stepwise regression Home based Work White Collar

Model Summary

Model	R	R Square(a)	Adjusted R Square	Std. Error of the Estimate
1	,662	,439	,439	1,294
2	,734	,538	,538	1,174
3	,783	,614	,613	1,074
4	,822	,676	,675	,984
5	,824	,679	,679	,979
6	,824	,679	,679	,979
7	,825	,681	,681	,976
8	,829	,687	,686	,968
9	,829	,687	,686	,967
10	,829	,688	,687	,966
11	,829	,688	,687	,966

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4418,449	1	4418,449	2640,016	,000
	Residual	5648,551	3375	1,674		
	Total	10067,000	3376			
2	Regression	5417,771	2	2708,885	1965,870	,000
	Residual	4649,229	3374	1,378		
	Total	10067,000	3376			
3	Regression	6177,797	3	2059,266	1785,945	,000
	Residual	3889,203	3373	1,153		
	Total	10067,000	3376			
4	Regression	6803,427	4	1700,857	1757,365	,000
	Residual	3263,573	3372	,968		
	Total	10067,000	3376			
5	Regression	6836,181	5	1367,236	1426,559	,000
	Residual	3230,819	3371	,958		
	Total	10067,000	3376			
6	Regression	6834,330	4	1708,582	1782,223	,000
	Residual	3232,670	3372	,959		
	Total	10067,000	3376			
7	Regression	6857,098	5	1371,420	1440,248	,000
	Residual	3209,902	3371	,952		
	Total	10067,000	3376			
8	Regression	6911,686	6	1151,948	1230,325	,000
	Residual	3155,314	3370	,936		
	Total	10067,000	3376			
9	Regression	6917,253	7	988,179	1056,966	,000
	Residual	3149,747	3369	,935		
	Total	10067,000	3376			
10	Regression	6922,128	8	865,266	926,656	,000
	Residual	3144,872	3368	,934		
	Total	10067,000	3376			
11	Regression	6926,348	9	769,594	825,059	,000

	Residual	3140,652	3367	,933		
	Total	10067,000	3376			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Dep_65_Plus_0	1,286	,025	,662	51,381	,000	1,000	1,000
2	Dep_65_Plus_0	,895	,027	,461	33,225	,000	,710	1,408
	WHITE_2	1,323	,049	,374	26,930	,000	,710	1,408
3	Dep_65_Plus_0	,758	,025	,390	30,039	,000	,678	1,475
	WHITE_2	1,458	,045	,412	32,224	,000	,701	1,427
	WHITE_3+	2,807	,109	,281	25,674	,000	,955	1,047
4	Dep_65_Plus_0	,066	,036	,034	1,852	,064	,284	3,518
	WHITE_2	2,137	,049	,604	43,337	,000	,495	2,019
	WHITE_3+	3,444	,103	,345	33,353	,000	,899	1,113
	WHITE_1	1,118	,044	,385	25,425	,000	,419	2,386
5	Dep_65_Plus_0	,050	,036	,026	1,390	,165	,282	3,540
	WHITE_2	2,093	,050	,591	42,168	,000	,484	2,066
	WHITE_3+	3,253	,108	,326	30,160	,000	,816	1,226
	WHITE_1	1,099	,044	,378	25,034	,000	,417	2,399
	CARS_3+	,312	,053	,064	5,846	,000	,787	1,271
6	WHITE_2	2,140	,036	,605	59,423	,000	,919	1,088
	WHITE_3+	3,294	,104	,330	31,797	,000	,884	1,131
	WHITE_1	1,144	,029	,394	39,523	,000	,958	1,044
	CARS_3+	,317	,053	,066	5,975	,000	,792	1,263
7	WHITE_2	2,075	,038	,586	54,195	,000	,808	1,238
	WHITE_3+	3,217	,104	,322	30,796	,000	,864	1,157
	WHITE_1	1,070	,033	,369	32,849	,000	,751	1,332
	CARS_3+	,298	,053	,062	5,615	,000	,787	1,270
	Dep_0_17_0	,123	,025	,059	4,890	,000	,661	1,513
8	WHITE_2	2,241	,044	,633	51,237	,000	,609	1,643
	WHITE_3+	3,375	,106	,338	31,951	,000	,831	1,203
	WHITE_1	1,170	,035	,403	33,577	,000	,646	1,548
	CARS_3+	,239	,053	,049	4,487	,000	,771	1,298
	Dep_0_17_0	,246	,030	,117	8,282	,000	,467	2,142
	Dep_18_64_0	-,266	,035	-,125	-7,636	,000	,349	2,863
9	WHITE_2	2,292	,048	,648	47,326	,000	,496	2,017
	WHITE_3+	3,424	,107	,343	31,867	,000	,802	1,247
	WHITE_1	1,197	,036	,412	32,797	,000	,588	1,700
	CARS_3+	,217	,054	,045	4,016	,000	,749	1,336
	Dep_0_17_0	,288	,034	,137	8,396	,000	,350	2,858
	Dep_18_64_0	-,262	,035	-,123	-7,515	,000	,349	2,869
	BLUE_0	-,083	,034	-,041	-2,440	,015	,334	2,993
10	WHITE_2	2,310	,049	,653	47,111	,000	,483	2,070
	WHITE_3+	3,452	,108	,346	31,937	,000	,791	1,264
	WHITE_1	1,215	,037	,419	32,520	,000	,560	1,786
	CARS_3+	,230	,054	,048	4,245	,000	,740	1,352
	Dep_0_17_0	,286	,034	,136	8,354	,000	,350	2,860
	Dep_18_64_0	-,265	,035	-,124	-7,590	,000	,348	2,872
	BLUE_0	-,079	,034	-,039	-2,329	,020	,333	3,000
	HHSIZE_3	-,105	,046	-,024	-2,285	,022	,824	1,214
11	WHITE_2	2,313	,049	,654	47,180	,000	,483	2,072
	WHITE_3+	3,453	,108	,346	31,959	,000	,791	1,264
	WHITE_1	1,207	,038	,416	32,145	,000	,554	1,805
	CARS_3+	,216	,055	,045	3,962	,000	,729	1,371
	Dep_0_17_0	,320	,038	,152	8,479	,000	,288	3,471
	Dep_18_64_0	-,281	,036	-,131	-7,870	,000	,333	3,000
	BLUE_0	-,098	,035	-,048	-2,784	,005	,313	3,195
	HHSIZE_3	-,153	,051	-,035	-2,985	,003	,667	1,499
	Dep_0_17_1	,130	,061	,026	2,127	,033	,601	1,665

Stepwise regression Home Based Education – Pre- Primary School

Model Summary

Model	R	R Square(a)	Adjusted R Square	Std. Error of the Estimate
1	,445	,198	,198	,991
2	,613	,376	,376	,874
3	,697	,486	,485	,793
4	,705	,497	,497	,784

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	816,480	1	816,480	832,136	,000(a)
	Residual	3310,520	3374	,981		
	Total	4127,000	3375			
2	Regression	1553,102	2	776,551	1017,642	,000(c)
	Residual	2573,898	3373	,763		
	Total	4127,000	3375			
3	Regression	2003,885	3	667,962	1060,878	,000(d)
	Residual	2123,115	3372	,630		
	Total	4127,000	3375			
4	Regression	2052,356	4	513,089	833,696	,000(e)
	Residual	2074,644	3371	,615		
	Total	4127,000	3375			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Dep_0_17_3	2,160	,075	,445	28,847	,000	1,000	1,000
2	Dep_0_17_3	2,160	,066	,445	32,710	,000	1,000	1,000
	Dep_0_17_2	1,261	,041	,422	31,070	,000	1,000	1,000
3	Dep_0_17_3	2,160	,060	,445	36,011	,000	1,000	1,000
	Dep_0_17_2	1,261	,037	,422	34,204	,000	1,000	1,000
	Dep_0_17_4+	3,130	,117	,330	26,757	,000	1,000	1,000
4	Dep_0_17_3	2,160	,059	,445	36,423	,000	1,000	1,000
	Dep_0_17_2	1,261	,036	,422	34,596	,000	1,000	1,000
	Dep_0_17_4+	3,130	,116	,330	27,064	,000	1,000	1,000
	Dep_0_17_1	,341	,038	,108	8,875	,000	1,000	1,000

Stepwise regression Home Based Education – Secondary School

Model Summary

Model	R	R Square(a)	Adjusted R Square	Std. Error of the Estimate
1	,316	,100	,100	,786
2	,415	,172	,172	,754
3	,481	,231	,230	,727
4	,523	,273	,272	,707
5	,526	,277	,276	,705
6	,529	,280	,278	,704
7	,531	,282	,281	,703
8	,533	,284	,282	,702
9	,533	,285	,283	,702

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	231,862	1	231,862	375,291	,000(a)
	Residual	2085,138	3375	,618		
	Total	2317,000	3376			
2	Regression	398,953	2	199,477	350,896	,000(c)
	Residual	1918,047	3374	,568		
	Total	2317,000	3376			
3	Regression	535,117	3	178,372	337,648	,000(d)
	Residual	1781,883	3373	,528		
	Total	2317,000	3376			
4	Regression	632,704	4	158,176	316,672	,000(e)
	Residual	1684,296	3372	,499		
	Total	2317,000	3376			
5	Regression	641,091	5	128,218	257,904	,000(f)
	Residual	1675,909	3371	,497		
	Total	2317,000	3376			
6	Regression	647,612	6	107,935	217,890	,000(g)
	Residual	1669,388	3370	,495		
	Total	2317,000	3376			
7	Regression	653,785	7	93,398	189,186	,000(h)
	Residual	1663,215	3369	,494		
	Total	2317,000	3376			
8	Regression	657,381	8	82,173	166,760	,000(i)
	Residual	1659,619	3368	,493		
	Total	2317,000	3376			
9	Regression	659,307	9	73,256	148,794	,000(j)
	Residual	1657,693	3367	,492		
	Total	2317,000	3376			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Dep_0_17_2	,707	,036	,316	19,372	,000	1,000	1,000
2	Dep_0_17_2	,707	,035	,316	20,196	,000	1,000	1,000

	Dep_0_17_3	,977	,057	,269	17,144	,000	1,000	1,000
3	Dep_0_17_2	,707	,034	,316	20,950	,000	1,000	1,000
	Dep_0_17_3	,977	,055	,269	17,785	,000	1,000	1,000
	Dep_0_17_1	,572	,036	,242	16,055	,000	1,000	1,000
4	Dep_0_17_2	,707	,033	,316	21,545	,000	1,000	1,000
	Dep_0_17_3	,977	,053	,269	18,290	,000	1,000	1,000
	Dep_0_17_1	,572	,035	,242	16,511	,000	1,000	1,000
	Dep_0_17_4+	1,457	,104	,205	13,978	,000	1,000	1,000
5	Dep_0_17_2	,700	,033	,313	21,359	,000	,997	1,003
	Dep_0_17_3	,975	,053	,268	18,283	,000	1,000	1,000
	Dep_0_17_1	,562	,035	,238	16,227	,000	,995	1,005
	Dep_0_17_4+	1,447	,104	,204	13,912	,000	,999	1,001
	Dep_18_64_3+	,454	,111	,060	4,107	,000	,992	1,008
6	Dep_0_17_2	,677	,033	,303	20,318	,000	,961	1,040
	Dep_0_17_3	,800	,072	,220	11,152	,000	,550	1,818
	Dep_0_17_1	,548	,035	,232	15,746	,000	,983	1,018
	Dep_0_17_4+	1,427	,104	,201	13,733	,000	,997	1,003
	Dep_18_64_3+	,409	,111	,054	3,682	,000	,980	1,021
	HHSIZE_5	,234	,064	,073	3,628	,000	,529	1,890
7	Dep_0_17_2	,656	,034	,294	19,422	,000	,932	1,073
	Dep_0_17_3	,713	,076	,196	9,405	,000	,491	2,035
	Dep_0_17_1	,539	,035	,228	15,456	,000	,977	1,024
	Dep_0_17_4+	1,097	,140	,155	7,849	,000	,550	1,818
	Dep_18_64_3+	,330	,113	,044	2,920	,004	,942	1,062
	HHSIZE_5	,299	,067	,093	4,472	,000	,488	2,047
	HHSIZE_6+	,358	,101	,072	3,536	,000	,507	1,971
8	Dep_0_17_2	,550	,052	,246	10,632	,000	,396	2,522
	Dep_0_17_3	,626	,082	,172	7,607	,000	,416	2,402
	Dep_0_17_1	,507	,037	,215	13,827	,000	,880	1,137
	Dep_0_17_4+	1,014	,143	,143	7,098	,000	,525	1,906
	Dep_18_64_3+	,275	,115	,037	2,393	,017	,912	1,097
	HHSIZE_5	,379	,073	,118	5,187	,000	,408	2,450
	HHSIZE_6+	,442	,106	,089	4,177	,000	,463	2,158
	HHSIZE_4	,127	,047	,062	2,701	,007	,400	2,499
9	Dep_0_17_2	,571	,053	,256	10,817	,000	,380	2,629
	Dep_0_17_3	,648	,083	,178	7,810	,000	,408	2,448
	Dep_0_17_1	,526	,038	,223	13,885	,000	,824	1,214
	Dep_0_17_4+	1,043	,144	,147	7,267	,000	,519	1,926
	Dep_18_64_3+	,282	,115	,038	2,458	,014	,911	1,098
	HHSIZE_5	,384	,073	,120	5,251	,000	,408	2,453
	HHSIZE_6+	,438	,106	,089	4,141	,000	,463	2,159
	HHSIZE_4	,135	,047	,066	2,856	,004	,397	2,516
	CARS_2	-,045	,023	-,034	-1,978	,048	,732	1,365

Stepwise regression Home Based Education – Tertiary School

Model Summary

Model	R	R Square(a)	Adjusted R Square	Std. Error of the Estimate
1	,317	,101	,101	,466
2	,391	,153	,152	,452
3	,417	,174	,173	,447
4	,424	,179	,178	,445
5	,425	,181	,179	,445
6	,426	,182	,180	,445
7	,427	,183	,181	,444

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	82,049	1	82,049	378,324	,000(a)
	Residual	731,951	3375	,217		
	Total	814,000	3376			
2	Regression	124,222	2	62,111	303,811	,000(c)
	Residual	689,778	3374	,204		
	Total	814,000	3376			
3	Regression	141,478	3	47,159	236,526	,000(d)
	Residual	672,522	3373	,199		
	Total	814,000	3376			
4	Regression	146,080	4	36,520	184,372	,000(e)
	Residual	667,920	3372	,198		
	Total	814,000	3376			
5	Regression	147,048	5	29,410	148,645	,000(f)
	Residual	666,952	3371	,198		
	Total	814,000	3376			
6	Regression	147,895	6	24,649	124,706	,000(g)
	Residual	666,105	3370	,198		
	Total	814,000	3376			
7	Regression	148,734	7	21,248	107,602	,000(h)
	Residual	665,266	3369	,197		
	Total	814,000	3376			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Dep_18_64_3+	1,415	,073	,317	19,451	,000	1,000	1,000
2	Dep_18_64_3+	1,415	,071	,317	20,033	,000	1,000	1,000
	Dep_18_64_2	,444	,031	,228	14,363	,000	1,000	1,000
3	Dep_18_64_3+	1,415	,070	,317	20,286	,000	1,000	1,000
	Dep_18_64_2	,444	,031	,228	14,544	,000	1,000	1,000
	Dep_18_64_1	,137	,015	,146	9,303	,000	1,000	1,000
4	Dep_18_64_3+	1,407	,070	,316	20,240	,000	1,000	1,000
	Dep_18_64_2	,435	,030	,223	14,272	,000	,996	1,004
	Dep_18_64_1	,128	,015	,137	8,695	,000	,986	1,014
	WHITE_2	,076	,016	,076	4,820	,000	,982	1,019
5	Dep_18_64_3+	1,415	,070	,317	20,336	,000	,997	1,003
	Dep_18_64_2	,440	,031	,226	14,407	,000	,991	1,010

	Dep_18_64_1	,136	,015	,145	8,974	,000	,936	1,069
	WHITE_2	,085	,016	,085	5,217	,000	,923	1,083
	Dep_0_17_2	-,049	,022	-,037	-2,211	,027	,874	1,144
6	Dep_18_64_3+	1,426	,070	,320	20,443	,000	,992	1,009
	Dep_18_64_2	,447	,031	,229	14,553	,000	,979	1,021
	Dep_18_64_1	,144	,016	,153	9,212	,000	,879	1,138
	WHITE_2	,092	,017	,092	5,531	,000	,885	1,130
	Dep_0_17_2	-,055	,022	-,041	-2,455	,014	,860	1,163
	Dep_0_17_1	-,048	,023	-,034	-2,070	,039	,884	1,132
7	Dep_18_64_3+	1,412	,070	,317	20,162	,000	,982	1,018
	Dep_18_64_2	,436	,031	,224	14,007	,000	,951	1,051
	Dep_18_64_1	,138	,016	,146	8,648	,000	,847	1,181
	WHITE_2	,085	,017	,084	4,952	,000	,843	1,186
	Dep_0_17_2	-,054	,022	-,041	-2,422	,016	,860	1,163
	Dep_0_17_1	-,052	,023	-,037	-2,224	,026	,878	1,138
	CARS_3+	,048	,023	,035	2,062	,039	,848	1,179

Stepwise regression Home Based Shopping

Model Summary

Model	R	R Square(a)	Adjusted R Square	Std. Error of the Estimate
1	,521	,272	,272	2,272
2	,566	,320	,320	2,195
3	,602	,362	,361	2,127
4	,624	,390	,389	2,080
5	,637	,406	,405	2,053
6	,644	,414	,413	2,039
7	,648	,419	,418	2,031
8	,649	,421	,420	2,027
9	,650	,422	,421	2,026
10	,650	,423	,421	2,025

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6504,701	1	6504,701	1260,507	,000(a)
	Residual	17416,299	3375	5,160		
	Total	23921,000	3376			
2	Regression	7663,153	2	3831,576	795,169	,000(c)
	Residual	16257,847	3374	4,819		
	Total	23921,000	3376			
3	Regression	8659,372	3	2886,457	637,941	,000(d)
	Residual	15261,628	3373	4,525		
	Total	23921,000	3376			
4	Regression	9326,769	4	2331,692	538,738	,000(e)
	Residual	14594,231	3372	4,328		
	Total	23921,000	3376			
5	Regression	9707,866	5	1941,573	460,493	,000(f)
	Residual	14213,134	3371	4,216		
	Total	23921,000	3376			
6	Regression	9907,340	6	1651,223	397,086	,000(g)
	Residual	14013,660	3370	4,158		
	Total	23921,000	3376			
7	Regression	10029,071	7	1432,724	347,457	,000(h)

Model		Sum of Squares	df	Mean Square	F	Sig.
	Residual	13891,929	3369	4,123		
	Total	23921,000	3376			
8	Regression	10076,378	8	1259,547	306,412	,000(i)
	Residual	13844,622	3368	4,111		
	Total	23921,000	3376			
9	Regression	10098,389	9	1122,043	273,314	,000(j)
	Residual	13822,611	3367	4,105		
	Total	23921,000	3376			
10	Regression	10118,913	10	1011,891	246,776	,000(k)
	Residual	13802,087	3366	4,100		
	Total	23921,000	3376			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Dep_65_Plus_0	1,560	,044	,521	35,504	,000	1,000	1,000
2	Dep_65_Plus_0	1,560	,042	,521	36,741	,000	1,000	1,000
	Dep_65_Plus_2+	2,161	,139	,220	15,505	,000	1,000	1,000
3	Dep_65_Plus_0	1,560	,041	,521	37,916	,000	1,000	1,000
	Dep_65_Plus_2+	2,161	,135	,220	16,001	,000	1,000	1,000
	Dep_65_Plus_1	1,478	,100	,204	14,838	,000	1,000	1,000
4	Dep_65_Plus_0	1,761	,043	,589	40,600	,000	,861	1,162
	Dep_65_Plus_2+	2,161	,132	,220	16,360	,000	1,000	1,000
	Dep_65_Plus_1	2,159	,112	,298	19,313	,000	,759	1,317
	HHSIZE_1	-1,164	,094	-,203	-12,418	,000	,676	1,479
5	Dep_65_Plus_0	2,173	,061	,726	35,688	,000	,426	2,350
	Dep_65_Plus_2+	2,871	,150	,292	19,109	,000	,753	1,328
	Dep_65_Plus_1	2,611	,120	,361	21,732	,000	,640	1,561
	HHSIZE_1	-,953	,095	-,166	-10,020	,000	,639	1,564
	Dep_18_64_0	-,743	,078	-,225	-9,507	,000	,314	3,188
6	Dep_65_Plus_0	2,376	,067	,794	35,357	,000	,345	2,902
	Dep_65_Plus_2+	3,345	,164	,341	20,377	,000	,622	1,607
	Dep_65_Plus_1	2,888	,126	,399	22,950	,000	,576	1,737
	HHSIZE_1	-1,238	,103	-,216	-12,016	,000	,537	1,861
	Dep_18_64_0	-,679	,078	-,206	-8,685	,000	,309	3,234
	HHSIZE_2	-,582	,084	-,132	-6,926	,000	,476	2,101
7	Dep_65_Plus_0	2,279	,069	,762	32,919	,000	,322	3,107
	Dep_65_Plus_2+	3,212	,165	,327	19,427	,000	,608	1,644
	Dep_65_Plus_1	2,762	,127	,381	21,670	,000	,557	1,796
	HHSIZE_1	-1,163	,104	-,203	-11,233	,000	,528	1,895
	Dep_18_64_0	-,641	,078	-,194	-8,205	,000	,307	3,260
	HHSIZE_2	-,502	,085	-,114	-5,906	,000	,462	2,166
	HHSIZE_6+	1,179	,217	,074	5,433	,000	,923	1,084
8	Dep_65_Plus_0	2,437	,083	,815	29,237	,000	,221	4,518
	Dep_65_Plus_2+	3,383	,173	,344	19,599	,000	,556	1,797
	Dep_65_Plus_1	2,928	,136	,404	21,478	,000	,485	2,061
	HHSIZE_1	-1,321	,113	-,231	-11,651	,000	,439	2,280
	Dep_18_64_0	-,645	,078	-,195	-8,260	,000	,307	3,260
	HHSIZE_2	-,661	,097	-,150	-6,818	,000	,354	2,826
	HHSIZE_6+	1,020	,222	,064	4,603	,000	,881	1,134
	HHSIZE_3	-,386	,114	-,058	-3,392	,001	,594	1,685
9	Dep_65_Plus_0	2,445	,083	,817	29,322	,000	,221	4,524
	Dep_65_Plus_2+	3,417	,173	,348	19,738	,000	,552	1,810
	Dep_65_Plus_1	2,980	,138	,411	21,580	,000	,472	2,118
	HHSIZE_1	-1,254	,117	-,219	-10,725	,000	,412	2,428
	Dep_18_64_0	-,656	,078	-,199	-8,394	,000	,306	3,273
	HHSIZE_2	-,652	,097	-,148	-6,726	,000	,353	2,830
	HHSIZE_6+	1,012	,221	,064	4,570	,000	,881	1,135
	HHSIZE_3	-,378	,114	-,057	-3,323	,001	,593	1,686
	CARS_0	-,312	,135	-,034	-2,316	,021	,783	1,277

10	Dep_65_Plus_0	2,464	,084	,824	29,413	,000	,219	4,574
	Dep_65_Plus_2+	3,412	,173	,347	19,719	,000	,552	1,811
	Dep_65_Plus_1	2,995	,138	,414	21,676	,000	,471	2,123
	HHSIZE_1	-1,288	,118	-,225	-10,931	,000	,405	2,469
	Dep_18_64_0	-,637	,079	-,193	-8,107	,000	,302	3,312
	HHSIZE_2	-,667	,097	-,152	-6,865	,000	,352	2,843
	HHSIZE_6+	1,026	,221	,065	4,632	,000	,881	1,136
	HHSIZE_3	-,380	,114	-,057	-3,339	,001	,593	1,686
	CARS_0	-,313	,135	-,034	-2,320	,020	,783	1,277
	BLUE_2	-,389	,174	-,030	-2,237	,025	,923	1,084

Stepwise regression Home Based Shopping with intercept

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,209	,044	,043	2,086
2	,253	,064	,063	2,064
3	,275	,076	,075	2,051
4	,291	,085	,084	2,041
5	,305	,093	,092	2,032
6	,310	,096	,095	2,029
7	,313	,098	,096	2,027
8	,315	,100	,097	2,026
9	,318	,101	,098	2,025

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
7	Regression	1505,986	7	215,141	52,338	,000(g)
	Residual	13844,622	3368	4,111		
	Total	15350,609	3375			
8	Regression	1527,997	8	191,000	46,525	,000(h)
	Residual	13822,611	3367	4,105		
	Total	15350,609	3375			
9	Regression	1548,521	9	172,058	41,961	,000(i)
	Residual	13802,087	3366	4,100		
	Total	15350,609	3375			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
7	(Constant)	3,383	,173		19,599	,000		
	HHSIZE_1	-1,321	,113	-,255	-11,651	,000	,559	1,787
	Dep_18_64_0	-,645	,078	-,144	-8,260	,000	,881	1,135
	HHSIZE_6+	1,020	,222	,079	4,603	,000	,907	1,103
	Dep_65_Plus_0	-,946	,145	-,180	-6,541	,000	,353	2,833
	HHSIZE_2	-,661	,097	-,149	-6,818	,000	,558	1,791
	HHSIZE_3	-,386	,114	-,066	-3,392	,001	,705	1,418
	Dep_65_Plus_1	-,455	,173	-,073	-2,630	,009	,348	2,876
8	(Constant)	3,417	,173		19,738	,000		
	HHSIZE_1	-1,254	,117	-,242	-10,725	,000	,525	1,904
	Dep_18_64_0	-,656	,078	-,147	-8,394	,000	,878	1,139
	HHSIZE_6+	1,012	,221	,078	4,570	,000	,907	1,103

	Dep_65_Plus_0	-,972	,145	-,185	-6,709	,000	,351	2,851
	HHSIZE_2	-,652	,097	-,147	-6,726	,000	,557	1,794
	HHSIZE_3	-,378	,114	-,065	-3,323	,001	,704	1,420
	Dep_65_Plus_1	-,437	,173	-,070	-2,522	,012	,347	2,882
	CARS_0	-,312	,135	-,041	-2,316	,021	,856	1,168
9	(Constant)	3,412	,173		19,719	,000		
	HHSIZE_1	-1,288	,118	-,249	-10,931	,000	,517	1,936
	Dep_18_64_0	-,637	,079	-,142	-8,107	,000	,868	1,153
	HHSIZE_6+	1,026	,221	,080	4,632	,000	,906	1,104
	Dep_65_Plus_0	-,948	,145	-,181	-6,525	,000	,349	2,867
	HHSIZE_2	-,667	,097	-,151	-6,865	,000	,555	1,802
	HHSIZE_3	-,380	,114	-,065	-3,339	,001	,704	1,420
	Dep_65_Plus_1	-,417	,173	-,067	-2,404	,016	,346	2,890
	CARS_0	-,313	,135	-,041	-2,320	,020	,856	1,168
	BLUE_2	-,389	,174	-,037	-2,237	,025	,965	1,037

Model 10 (no intercept)		Model 9 (with intercept)	
(Constant)	0,000	(Constant)	3,412
Dep_65_Plus_0	2,464	Dep_65_Plus_0	-,948
Dep_65_Plus_1	2,995	Dep_65_Plus_1	-,417
Dep_65_Plus_2+	3,412	Dep_65_Plus_2+	0,000
Dep_18_64_0	-,637	Dep_18_64_0	-,637
HHSIZE_1	-1,288	HHSIZE_1	-1,288
HHSIZE_2	-,667	HHSIZE_2	-,667
HHSIZE_3	-,380	HHSIZE_3	-,380
HHSIZE_6+	1,026	HHSIZE_6+	1,026
CARS_0	-,313	CARS_0	-,313
BLUE_2	-,389	BLUE_2	-,389

Stepwise regression Home Based Recreation

Model Summary

Model	R	R Square(a)	Adjusted R Square	Std. Error of the Estimate
1	,456	,208	,208	1,899
2	,481	,232	,231	1,871
3	,502	,252	,252	1,846
4	,519	,269	,268	1,825
5	,528	,279	,278	1,814
6	,538	,290	,289	1,800
7	,548	,300	,299	1,787
8	,548	,300	,299	1,786
9	,549	,302	,300	1,785
10	,551	,303	,302	1,783
11	,552	,304	,302	1,782

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3199,766	1	3199,766	887,055	,000(a)
	Residual	12174,234	3375	3,607		
	Total	15374,000	3376			
2	Regression	3562,014	2	1781,007	508,730	,000(c)
	Residual	11811,986	3374	3,501		
	Total	15374,000	3376			
3	Regression	3879,544	3	1293,181	379,479	,000(d)

	Residual	11494,456	3373	3,408		
	Total	15374,000	3376			
4	Regression	4139,143	4	1034,786	310,578	,000(e)
	Residual	11234,857	3372	3,332		
	Total	15374,000	3376			
5	Regression	4284,884	5	856,977	260,514	,000(f)
	Residual	11089,116	3371	3,290		
	Total	15374,000	3376			
6	Regression	4457,013	6	742,836	229,308	,000(g)
	Residual	10916,987	3370	3,239		
	Total	15374,000	3376			
7	Regression	4619,187	7	659,884	206,712	,000(h)
	Residual	10754,813	3369	3,192		
	Total	15374,000	3376			
8	Regression	4619,180	6	769,863	241,235	,000(i)
	Residual	10754,820	3370	3,191		
	Total	15374,000	3376			
9	Regression	4638,152	7	662,593	207,927	,000(j)
	Residual	10735,848	3369	3,187		
	Total	15374,000	3376			
10	Regression	4664,825	8	583,103	183,384	,000(k)
	Residual	10709,175	3368	3,180		
	Total	15374,000	3376			
11	Regression	4677,521	9	519,725	163,597	,000(l)
	Residual	10696,479	3367	3,177		
	Total	15374,000	3376			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Dep_65_Plus_0	1,094	,037	,456	29,783	,000	1,000	1,000
2	Dep_65_Plus_0	,740	,050	,309	14,734	,000	,519	1,926
	BLUE_0	,537	,053	,213	10,172	,000	,519	1,926
3	Dep_65_Plus_0	,638	,051	,266	12,587	,000	,497	2,013
	BLUE_0	,544	,052	,216	10,439	,000	,519	1,926
	HHSIZE_5	1,232	,128	,149	9,653	,000	,925	1,081
4	Dep_65_Plus_0	,459	,054	,191	8,485	,000	,427	2,344
	BLUE_0	,559	,052	,222	10,849	,000	,519	1,928
	HHSIZE_5	1,391	,127	,169	10,912	,000	,907	1,103
	HHSIZE_4	,771	,087	,147	8,827	,000	,785	1,273
5	Dep_65_Plus_0	,407	,054	,170	7,501	,000	,418	2,393
	BLUE_0	,561	,051	,222	10,951	,000	,519	1,928
	HHSIZE_5	1,439	,127	,174	11,338	,000	,904	1,106
	HHSIZE_4	,820	,087	,156	9,418	,000	,780	1,282
	HHSIZE_6+	1,260	,189	,099	6,656	,000	,967	1,034
6	Dep_65_Plus_0	,233	,059	,097	3,954	,000	,349	2,863
	BLUE_0	,564	,051	,224	11,101	,000	,519	1,928
	HHSIZE_5	1,600	,128	,194	12,515	,000	,877	1,140
	HHSIZE_4	,988	,089	,188	11,046	,000	,728	1,373
	HHSIZE_6+	1,401	,189	,110	7,420	,000	,957	1,045
	HHSIZE_3	,653	,090	,122	7,289	,000	,756	1,323
7	Dep_65_Plus_0	,003	,067	,001	,045	,964	,268	3,735
	BLUE_0	,428	,054	,170	7,938	,000	,454	2,204
	HHSIZE_5	1,896	,134	,230	14,198	,000	,792	1,262
	HHSIZE_4	1,294	,099	,246	13,120	,000	,590	1,694
	HHSIZE_6+	1,663	,191	,131	8,706	,000	,921	1,085
	HHSIZE_3	,950	,098	,177	9,672	,000	,620	1,613

	HHSIZE_2	,519	,073	,147	7,128	,000	,487	2,055
8	BLUE_0	,429	,051	,170	8,363	,000	,502	1,993
	HHSIZE_5	1,898	,123	,230	15,482	,000	,939	1,064
	HHSIZE_4	1,296	,082	,247	15,852	,000	,858	1,165
	HHSIZE_6+	1,665	,185	,131	8,981	,000	,978	1,023
	HHSIZE_3	,952	,083	,177	11,405	,000	,858	1,166
	HHSIZE_2	,521	,064	,148	8,166	,000	,635	1,575
9	BLUE_0	,511	,061	,203	8,338	,000	,351	2,848
	HHSIZE_5	1,779	,132	,216	13,495	,000	,811	1,233
	HHSIZE_4	1,182	,094	,225	12,534	,000	,645	1,551
	HHSIZE_6+	1,547	,191	,122	8,080	,000	,916	1,092
	HHSIZE_3	,835	,096	,156	8,681	,000	,645	1,551
	HHSIZE_2	,414	,077	,117	5,343	,000	,430	2,324
	BLUE_1	,213	,087	,048	2,440	,015	,526	1,902
10	BLUE_0	,586	,066	,232	8,815	,000	,298	3,361
	HHSIZE_5	1,720	,133	,209	12,904	,000	,792	1,263
	HHSIZE_4	1,121	,096	,213	11,617	,000	,614	1,628
	HHSIZE_6+	1,489	,192	,117	7,748	,000	,906	1,104
	HHSIZE_3	,784	,098	,146	8,020	,000	,624	1,604
	HHSIZE_2	,363	,079	,103	4,585	,000	,410	2,441
	BLUE_1	,272	,090	,062	3,034	,002	,499	2,005
	CARS_0	-,333	,115	-,046	-2,896	,004	,833	1,201
11	BLUE_0	,586	,066	,233	8,828	,000	,298	3,361
	HHSIZE_5	1,744	,134	,211	13,037	,000	,786	1,273
	HHSIZE_4	1,151	,098	,219	11,791	,000	,599	1,668
	HHSIZE_6+	1,506	,192	,118	7,830	,000	,904	1,106
	HHSIZE_3	,873	,107	,163	8,126	,000	,515	1,941
	HHSIZE_2	,373	,079	,106	4,703	,000	,408	2,450
	BLUE_1	,274	,090	,062	3,060	,002	,499	2,005
	CARS_0	-,335	,115	-,046	-2,913	,004	,833	1,201
	Dep_0_17_1	-,208	,104	-,034	-1,999	,046	,708	1,413

Stepwise regression Home Based Other

Model Summary

Model	R	R Square(a)	Adjusted R Square	Std. Error of the Estimate
1	,490	,240	,240	2,201
2	,569	,324	,323	2,077
3	,651	,424	,423	1,917
4	,677	,458	,458	1,860
5	,703	,494	,494	1,797
6	,706	,499	,498	1,789
7	,709	,502	,501	1,784
8	,710	,504	,503	1,781
9	,711	,505	,504	1,778
10	,712	,507	,505	1,776
11	,713	,508	,506	1,774
12	,714	,509	,508	1,772
13	,714	,510	,508	1,770
14	,715	,511	,509	1,769

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5170,692	1	5170,692	1067,259	,000(a)
	Residual	16351,308	3375	4,845		
	Total	21522,000	3376			
2	Regression	6966,964	2	3483,482	807,505	,000(c)
	Residual	14555,036	3374	4,314		
	Total	21522,000	3376			
3	Regression	9121,898	3	3040,633	827,094	,000(d)
	Residual	12400,102	3373	3,676		
	Total	21522,000	3376			
4	Regression	9861,560	4	2465,390	712,949	,000(e)
	Residual	11660,440	3372	3,458		
	Total	21522,000	3376			
5	Regression	10637,779	5	2127,556	658,935	,000(f)
	Residual	10884,221	3371	3,229		
	Total	21522,000	3376			
6	Regression	10736,992	6	1789,499	559,166	,000(g)
	Residual	10785,008	3370	3,200		
	Total	21522,000	3376			
7	Regression	10803,991	7	1543,427	485,147	,000(h)
	Residual	10718,009	3369	3,181		
	Total	21522,000	3376			
8	Regression	10840,363	8	1355,045	427,256	,000(i)
	Residual	10681,637	3368	3,172		
	Total	21522,000	3376			
9	Regression	10876,647	9	1208,516	382,240	,000(j)
	Residual	10645,353	3367	3,162		
	Total	21522,000	3376			
10	Regression	10908,015	10	1090,801	345,925	,000(k)
	Residual	10613,985	3366	3,153		
	Total	21522,000	3376			
11	Regression	10935,006	11	994,091	315,965	,000(l)
	Residual	10586,994	3365	3,146		
	Total	21522,000	3376			
12	Regression	10964,210	12	913,684	291,125	,000(m)
	Residual	10557,790	3364	3,138		
	Total	21522,000	3376			
13	Regression	10983,451	13	844,881	269,613	,000(n)
	Residual	10538,549	3363	3,134		
	Total	21522,000	3376			
14	Regression	11001,958	14	785,854	251,144	,000(o)
	Residual	10520,042	3362	3,129		
	Total	21522,000	3376			

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF

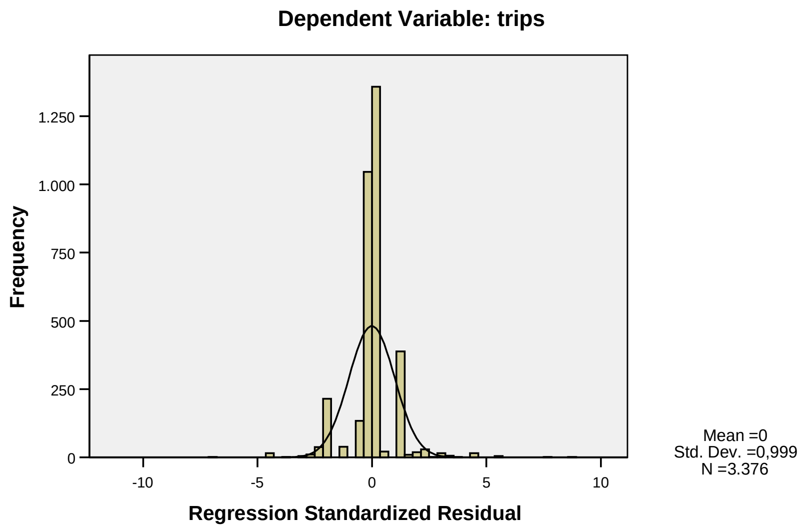
1	Dep_65_Plus_0	1,391	,043	,490	32,669	,000	1,000	1,000
2	Dep_65_Plus_0	1,031	,044	,363	23,485	,000	,838	1,193
	Dep_0_17_2	2,149	,105	,316	20,406	,000	,838	1,193
3	Dep_65_Plus_0	,752	,042	,265	17,859	,000	,776	1,289
	Dep_0_17_2	2,418	,098	,355	24,710	,000	,827	1,209
	Dep_0_17_3	3,648	,151	,329	24,211	,000	,925	1,081
4	Dep_65_Plus_0	,662	,041	,233	16,030	,000	,758	1,319
	Dep_0_17_2	2,505	,095	,368	26,342	,000	,824	1,213
	Dep_0_17_3	3,736	,146	,337	25,544	,000	,924	1,082
	Dep_0_17_4+	4,055	,277	,187	14,625	,000	,978	1,023
5	Dep_65_Plus_0	,364	,044	,128	8,203	,000	,615	1,626
	Dep_0_17_2	2,793	,094	,410	29,795	,000	,792	1,263
	Dep_0_17_3	4,028	,143	,363	28,250	,000	,908	1,102
	Dep_0_17_4+	4,354	,269	,201	16,208	,000	,973	1,028
	Dep_0_17_1	1,517	,098	,211	15,505	,000	,811	1,233
6	Dep_65_Plus_0	,326	,045	,115	7,301	,000	,601	1,664
	Dep_0_17_2	2,376	,120	,349	19,849	,000	,481	2,078
	Dep_0_17_3	3,978	,142	,359	27,969	,000	,904	1,106
	Dep_0_17_4+	4,391	,268	,203	16,416	,000	,972	1,029
	Dep_0_17_1	1,430	,099	,199	14,504	,000	,791	1,264
	HHSIZE_4	,607	,109	,097	5,568	,000	,485	2,062
7	Dep_65_Plus_0	,311	,045	,110	6,966	,000	,598	1,673
	Dep_0_17_2	2,180	,127	,320	17,193	,000	,427	2,344
	Dep_0_17_3	3,377	,193	,305	17,501	,000	,488	2,048
	Dep_0_17_4+	4,338	,267	,201	16,248	,000	,970	1,031
	Dep_0_17_1	1,358	,100	,189	13,641	,000	,771	1,296
	HHSIZE_4	,782	,115	,126	6,790	,000	,432	2,316
	HHSIZE_5	,789	,172	,081	4,589	,000	,477	2,098
8	Dep_65_Plus_0	,198	,056	,070	3,544	,000	,382	2,617
	Dep_0_17_2	2,157	,127	,317	17,012	,000	,425	2,351
	Dep_0_17_3	3,350	,193	,302	17,374	,000	,487	2,052
	Dep_0_17_4+	4,338	,267	,201	16,273	,000	,970	1,031
	Dep_0_17_1	1,354	,099	,188	13,619	,000	,771	1,297
	HHSIZE_4	,806	,115	,130	6,999	,000	,430	2,325
	HHSIZE_5	,815	,172	,084	4,748	,000	,476	2,102
	BLUE_0	,171	,050	,057	3,387	,001	,517	1,935
9	Dep_65_Plus_0	,227	,056	,080	4,029	,000	,373	2,681
	Dep_0_17_2	2,099	,128	,308	16,435	,000	,418	2,394
	Dep_0_17_3	3,284	,194	,296	16,972	,000	,482	2,073
	Dep_0_17_4+	4,262	,267	,197	15,957	,000	,963	1,038
	Dep_0_17_1	1,286	,101	,179	12,699	,000	,741	1,350
	HHSIZE_4	,783	,115	,126	6,797	,000	,429	2,334
	HHSIZE_5	,805	,172	,082	4,693	,000	,476	2,103
	BLUE_0	,243	,055	,081	4,447	,000	,438	2,285
	HHSIZE_1	-,268	,079	-,049	-3,388	,001	,691	1,446
10	Dep_65_Plus_0	,222	,056	,078	3,940	,000	,373	2,683
	Dep_0_17_2	1,973	,134	,290	14,756	,000	,380	2,630
	Dep_0_17_3	3,043	,208	,274	14,637	,000	,417	2,398
	Dep_0_17_4+	3,499	,360	,162	9,715	,000	,529	1,892
	Dep_0_17_1	1,240	,102	,172	12,137	,000	,726	1,377
	HHSIZE_4	,882	,119	,142	7,396	,000	,399	2,507
	HHSIZE_5	,992	,181	,102	5,473	,000	,424	2,356
	BLUE_0	,248	,055	,083	4,538	,000	,437	2,287
	HHSIZE_1	-,269	,079	-,050	-3,403	,001	,691	1,446
	HHSIZE_6+	,820	,260	,055	3,154	,002	,491	2,038
11	Dep_65_Plus_0	,262	,058	,092	4,526	,000	,352	2,842
	Dep_0_17_2	1,842	,141	,271	13,089	,000	,342	2,922
	Dep_0_17_3	2,855	,217	,257	13,134	,000	,381	2,628
	Dep_0_17_4+	3,227	,372	,149	8,686	,000	,496	2,018
	Dep_0_17_1	1,217	,102	,169	11,883	,000	,721	1,386
	HHSIZE_4	1,003	,126	,161	7,956	,000	,356	2,811
	HHSIZE_5	1,183	,192	,121	6,147	,000	,376	2,659
	BLUE_0	,249	,055	,083	4,565	,000	,437	2,287
	HHSIZE_1	-,291	,079	-,054	-3,669	,000	,685	1,459
	HHSIZE_6+	1,091	,276	,072	3,956	,000	,436	2,296
	CARS_3+	-,300	,102	-,042	-2,929	,003	,699	1,431
12	Dep_65_Plus_0	,212	,060	,075	3,529	,000	,326	3,070
	Dep_0_17_2	1,674	,151	,246	11,078	,000	,296	3,374

	Dep_0_17_3	2,687	,224	,242	11,996	,000	,358	2,797
	Dep_0_17_4+	3,043	,376	,141	8,095	,000	,483	2,071
	Dep_0_17_1	1,039	,118	,144	8,834	,000	,545	1,834
	HHSIZE_4	1,220	,145	,196	8,436	,000	,270	3,707
	HHSIZE_5	1,421	,207	,146	6,850	,000	,323	3,099
	BLUE_0	,241	,055	,081	4,417	,000	,436	2,293
	HHSIZE_1	-,251	,080	-,046	-3,120	,002	,667	1,500
	HHSIZE_6+	1,342	,287	,089	4,669	,000	,400	2,502
	CARS_3+	-,405	,108	-,057	-3,752	,000	,628	1,592
	HHSIZE_3	,333	,109	,052	3,050	,002	,494	2,025
13	Dep_65_Plus_0	,199	,060	,070	3,300	,001	,323	3,094
	Dep_0_17_2	1,681	,151	,247	11,133	,000	,296	3,375
	Dep_0_17_3	2,697	,224	,243	12,048	,000	,357	2,798
	Dep_0_17_4+	3,043	,376	,141	8,101	,000	,483	2,071
	Dep_0_17_1	1,043	,118	,145	8,875	,000	,545	1,835
	HHSIZE_4	1,218	,145	,196	8,422	,000	,270	3,708
	HHSIZE_5	1,418	,207	,145	6,839	,000	,323	3,099
	BLUE_0	,262	,055	,088	4,756	,000	,425	2,352
	HHSIZE_1	-,187	,084	-,034	-2,214	,027	,604	1,655
	HHSIZE_6+	1,342	,287	,089	4,673	,000	,400	2,502
	CARS_3+	-,409	,108	-,058	-3,791	,000	,628	1,593
	HHSIZE_3	,341	,109	,054	3,130	,002	,493	2,027
	CARS_0	-,288	,116	-,033	-2,478	,013	,804	1,244
14	Dep_65_Plus_0	,234	,062	,083	3,783	,000	,305	3,276
	Dep_0_17_2	1,644	,152	,241	10,844	,000	,293	3,409
	Dep_0_17_3	2,636	,225	,238	11,713	,000	,353	2,833
	Dep_0_17_4+	2,971	,377	,137	7,890	,000	,480	2,084
	Dep_0_17_1	1,015	,118	,141	8,601	,000	,540	1,852
	HHSIZE_4	1,257	,145	,202	8,648	,000	,266	3,755
	HHSIZE_5	1,483	,209	,152	7,102	,000	,317	3,152
	BLUE_0	,314	,059	,105	5,315	,000	,371	2,696
	HHSIZE_1	-,251	,088	-,046	-2,841	,005	,550	1,817
	HHSIZE_6+	1,392	,288	,092	4,836	,000	,398	2,514
	CARS_3+	-,407	,108	-,057	-3,775	,000	,628	1,593
	HHSIZE_3	,352	,109	,055	3,230	,001	,493	2,030
	CARS_0	-,303	,116	-,035	-2,602	,009	,802	1,247
	WHITE_2	-,200	,082	-,039	-2,432	,015	,575	1,740

Appendix 6 Diagnostics

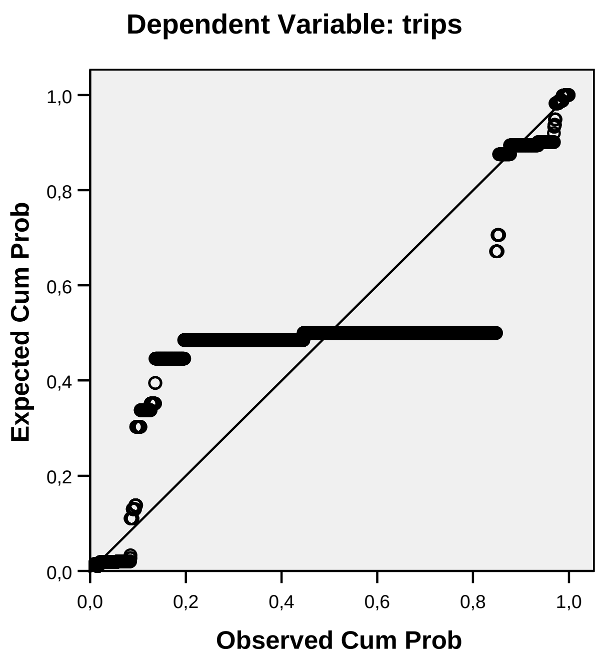
Home Based Work – Blue Collar

A residual is the difference between the observed and model-predicted values of the dependent variable. The residual for a given product is the observed value of the error term for that product. A histogram or P-P plot of the residuals will help you to check the assumption of normality of the error term.

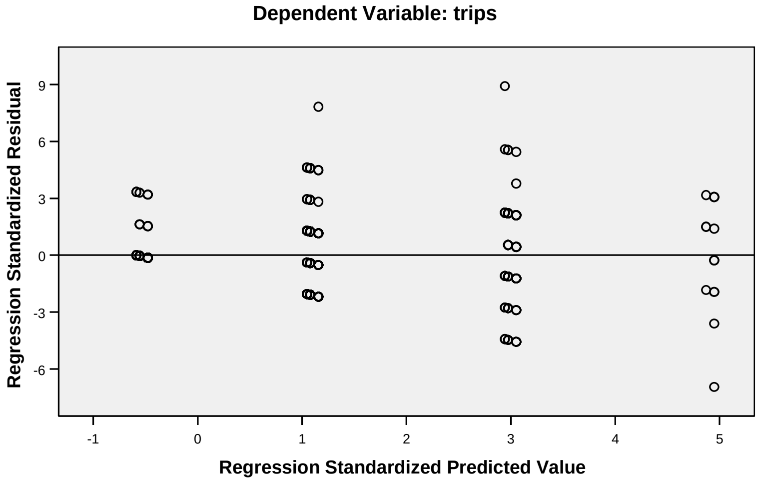


The shape of the histogram should approximately follow the shape of the normal curve. This histogram is not acceptably close to the normal curve.

Normal P-P Plot of Regression Standardized Residual



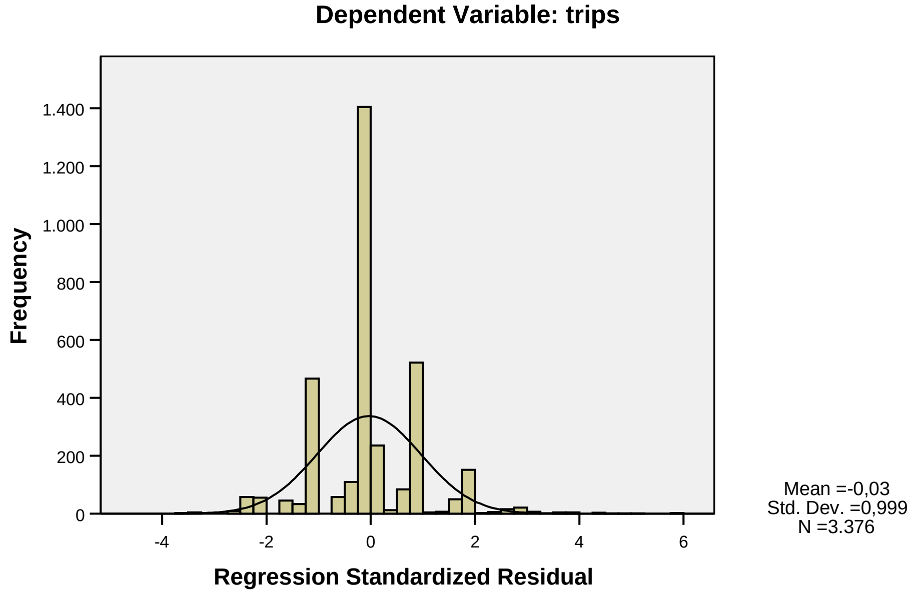
The P-P plotted residuals should follow the 45-degree line. As well as the histogram as the P-P plot indicate that the normality assumption are violated.

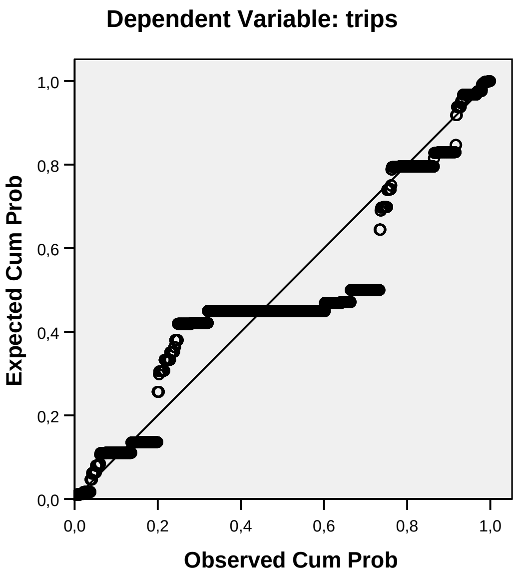


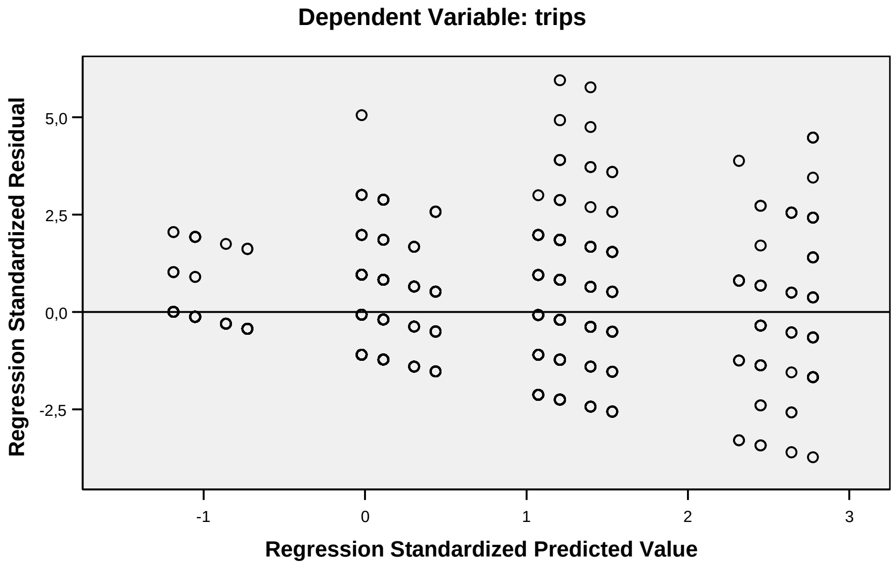
The plot of residuals by the predicted values shows that the variance of the errors increases with increasing predicted number of trips.

The standardized residual (ZRESID and ZPRED) is the residual divided by its standard error. Standardizing is a method for transforming data so that its mean is zero and standard deviation is one. If the distribution of the residuals is approximately normal, then 95% of the standardized residuals should fall between -2 and +2. If many of the residuals fall outside of + or – 2, then they could be considered unusual. However, about 5% of the residuals could fall outside of this region due to chance.

Home Based Work – White Collar

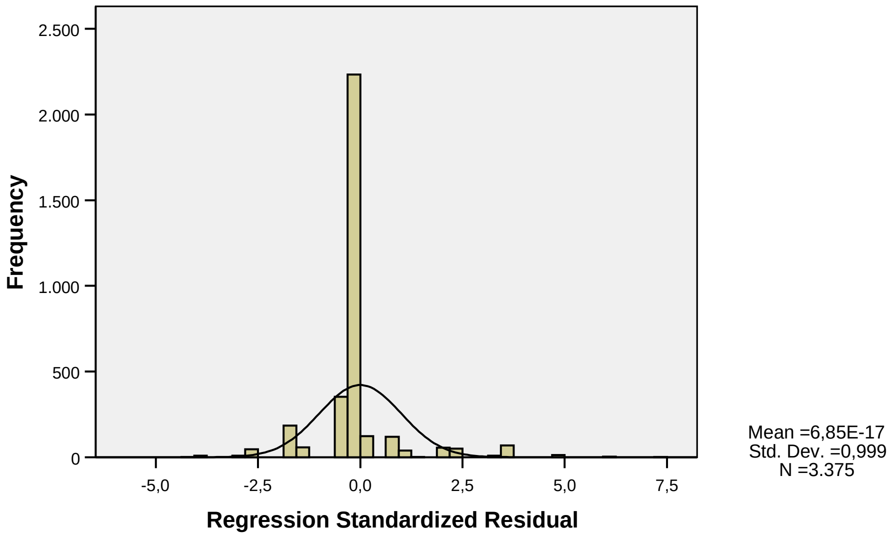




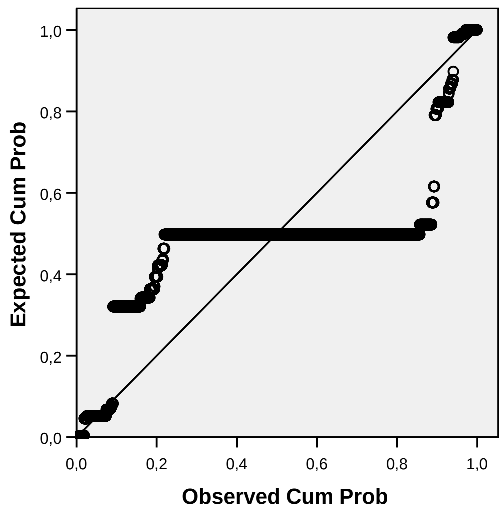


Home Based Education – PrePrimary

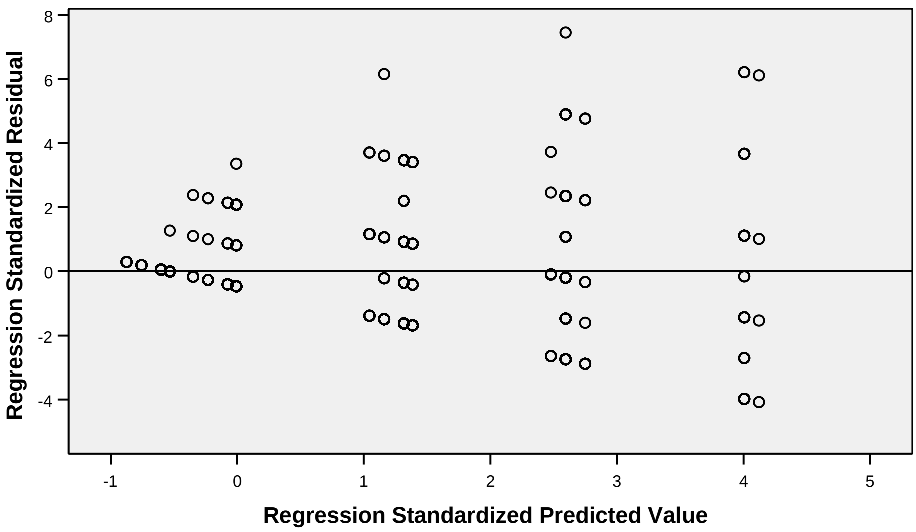
Dependent Variable: trips



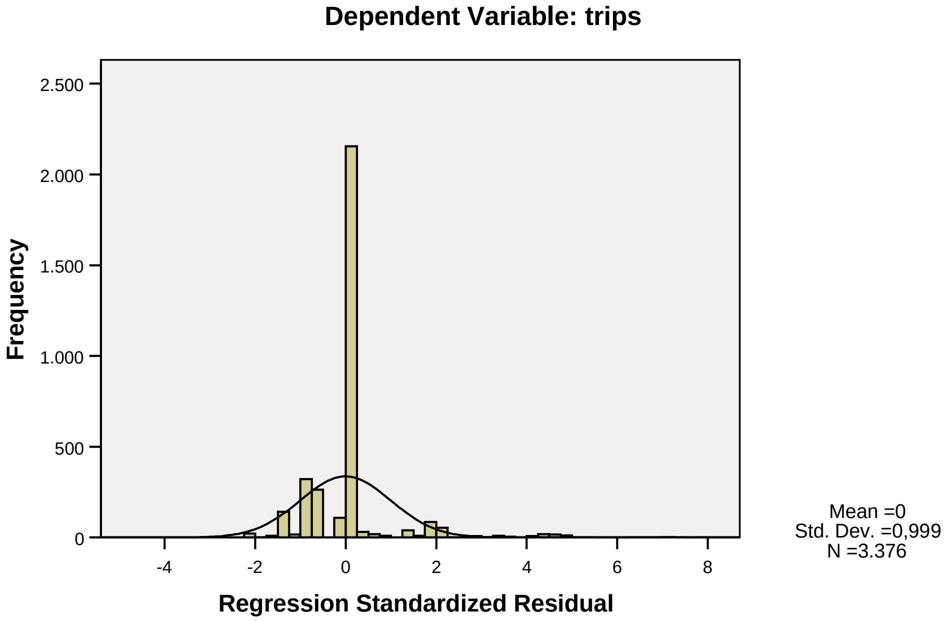
Dependent Variable: trips

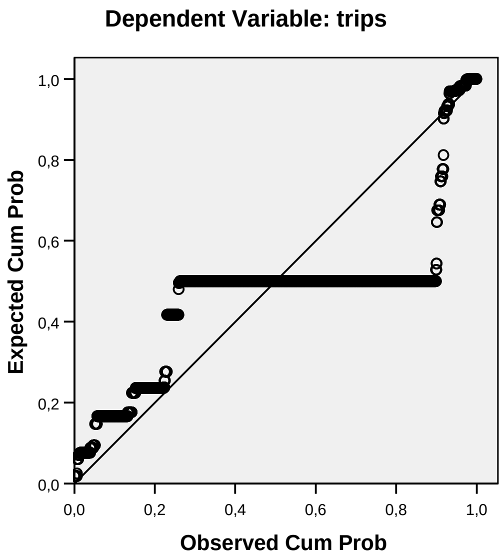


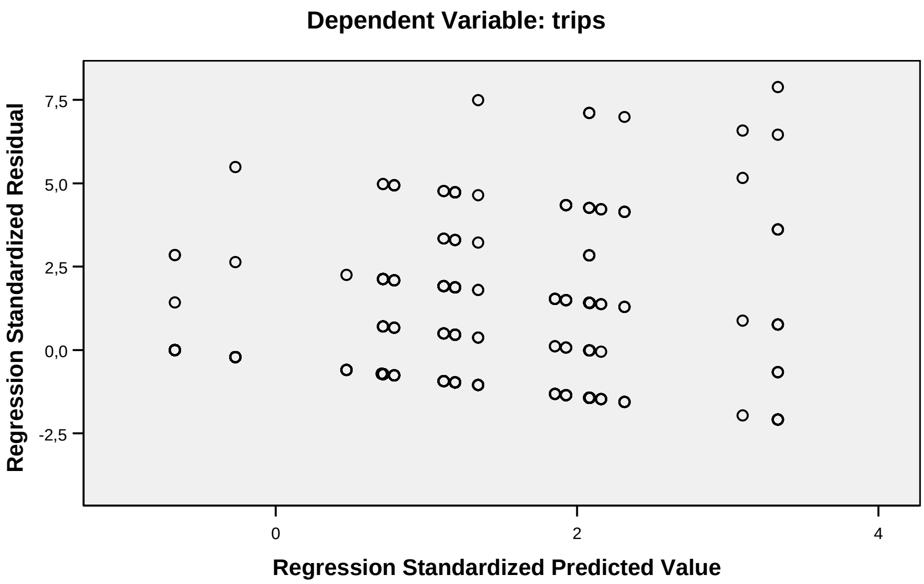
Dependent Variable: trips



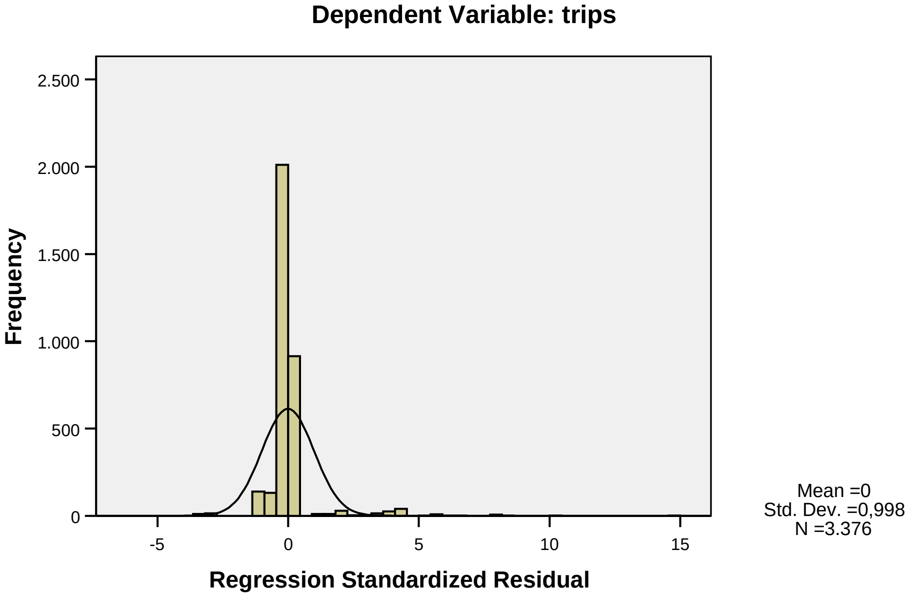
Home Based Education – Secondary

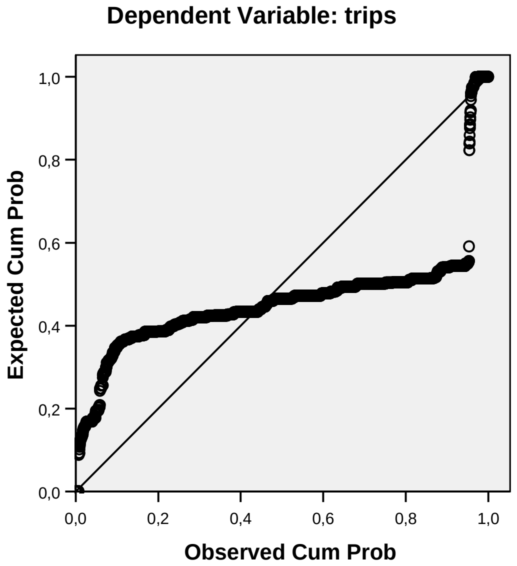


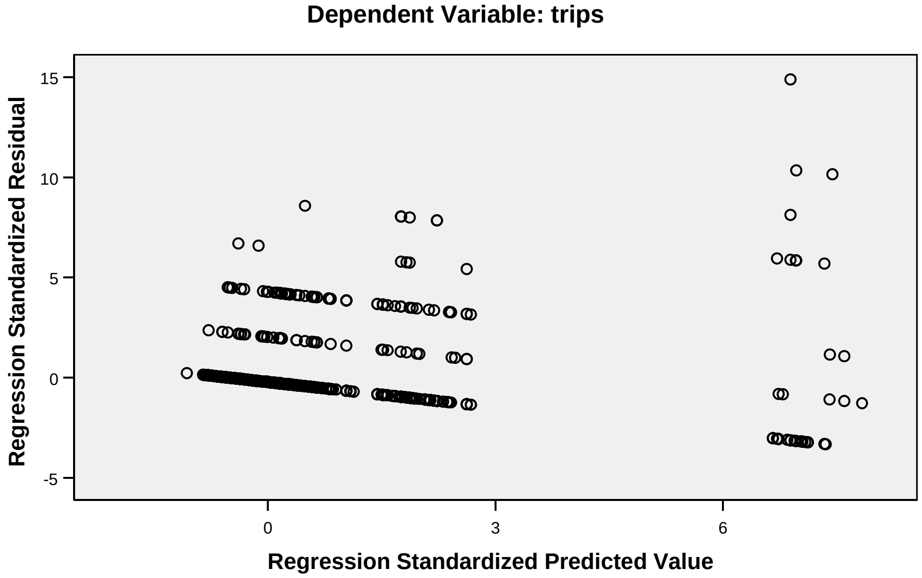




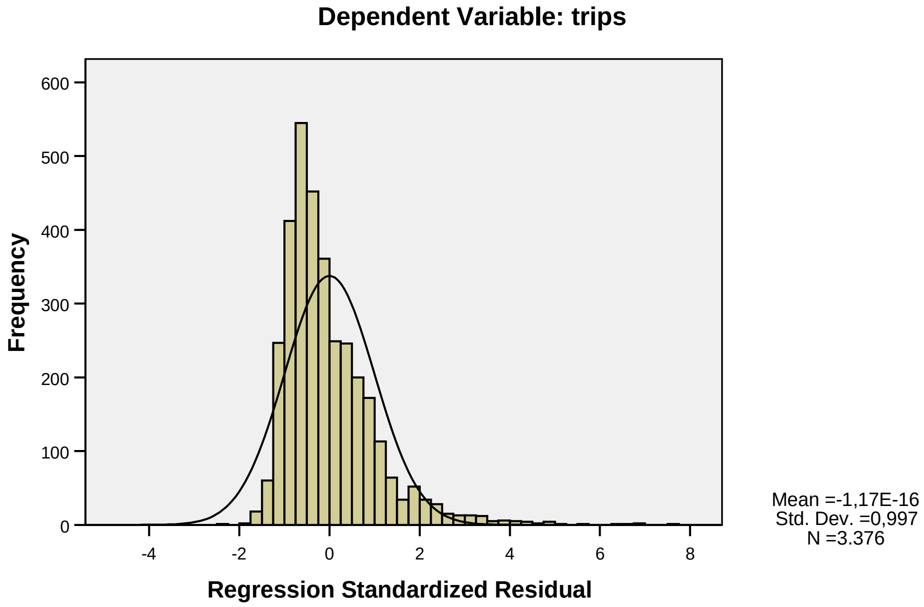
Home Based Education – Tertiary

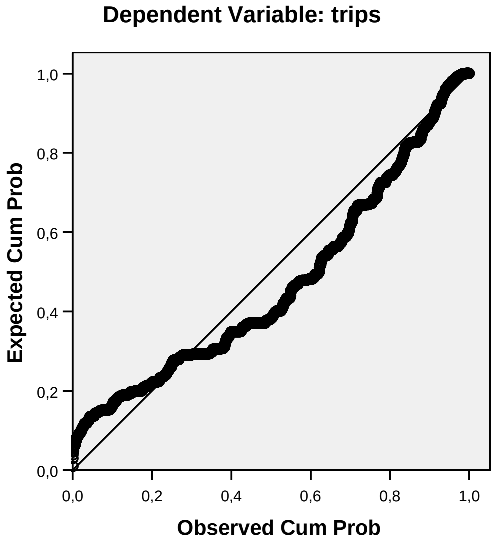


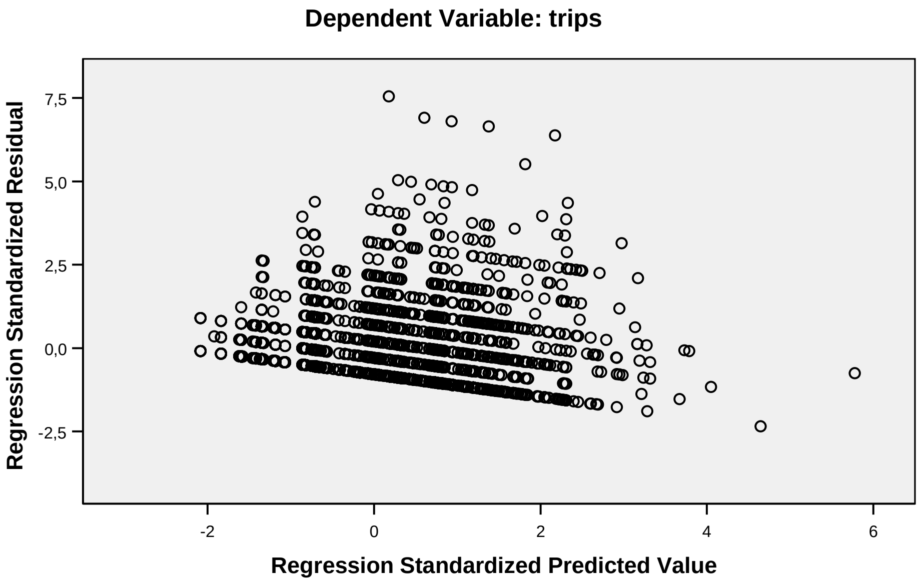




Home Based Shopping

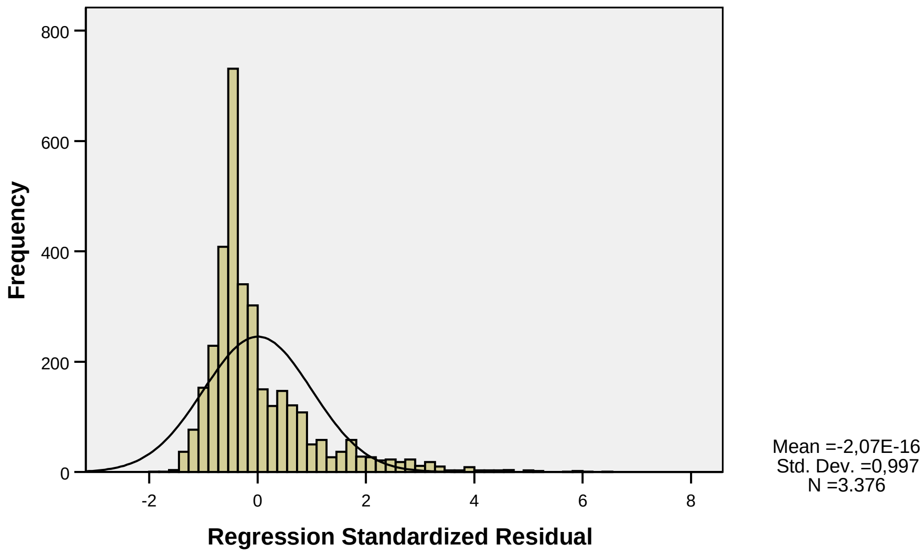




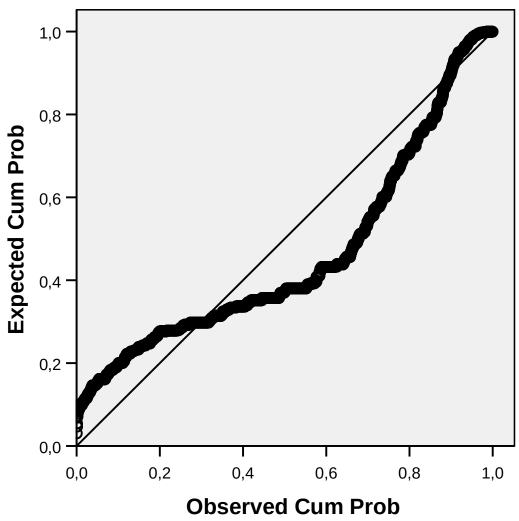


Home Based Recreation

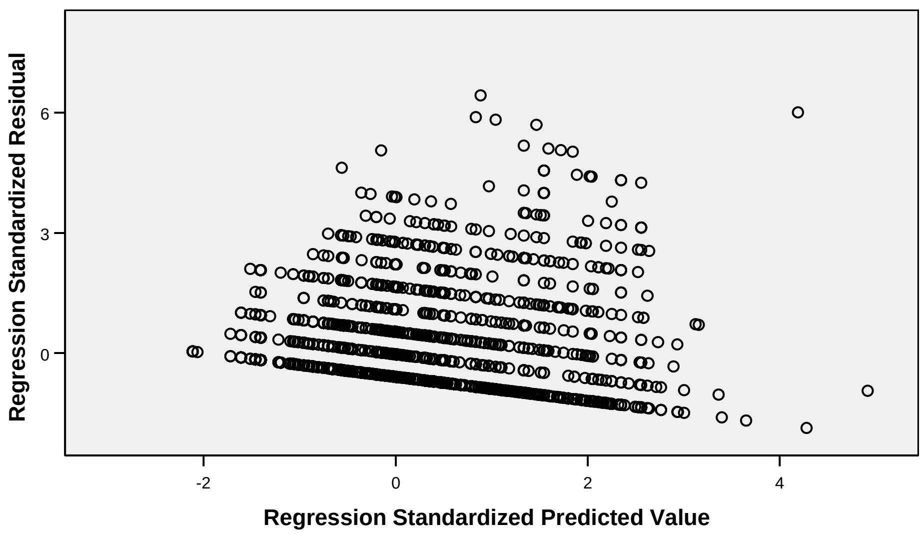
Dependent Variable: trips



Dependent Variable: trips



Dependent Variable: trips



Home Based Other