Samantha Oliszewski

Professional Portfolio

TABLE OF CONTENTS

SAMANTHA OLISZEWSKI

OBJECTIVE

To provide an analytical approach to decision-making and help a company grow through optimization of resources and informed innovation.

EXPERIENCE

2016 - Present NativeFAX, LLC Warm Springs, OR

Business Analyst

- Developed reports for board meetings that presented company profitability based on real-time service usage
- Performed service analysis to optimize resources and reduce costs based on estimated project requirements
- Recommended and implemented an efficient organizational structure for company data including relational databases and secure web interface for managing data
- Managed and implemented user portal for fax service that included account management and billing, and usage reporting in both graphical and verbose forms
- Drafted company policies and procedures which implemented and enforced PCI DSS and HIPAA compliance

2015 - Present Brookforest Farm Sherwood, OR

Operations Manager

- Determined the optimal number of trees to plant per orchard based on land dimensions and tree spacing
- Estimated project costs and assisted in purchasing bulk order of trees and irrigation equipment
- Scheduled optimized crop removal and planting processes based on the growing schedule
- Assisted in heavy machinery part stocking and repairs to ensure business continuity

2011 - Present FaxBack, Inc. Portland, OR

Software Engineer / Project Manager

- Created web applications using HTML5, CSS3, JavaScript/jQuery, TypeScript and SQL with Microsoft Visual Studio guided both from technical specifications written by management and myself
- Coordinated and maintained company-wide to-do list
- Recommended and implemented technological service enhancements that addressed company concerns
- Performed quality assurance and testing of products

EDUCATION

2017 - March 2019 Oregon State University Corvallis, OR

- Master of Science in Data Analytics (3.96 cumulative G.P.A.)
- Studied statistical theory, univariate and multivariate data analysis, and programming in R, Python and SQL with tools such as Apache Spark, Hadoop, Google Cloud Platform. Created formal data analysis reports, demonstrating written and oral communication skills

2012 - 2017 Portland State University Portland, OR

- Bachelor of Science in Economics (3.97 cumulative G.P.A.)
- Awarded Latin honor *summa cum laude*
- Awarded Harold Vatter Award for academic excellence by Department Chair (former Oregon State Economist Dr. Tom Potiowsky)

SKILLS

Univariate/multivariate data analysis, R, Python, SQL, Apache Spark, Hadoop, Google Cloud Platform, JavaScript/jQuery, TypeScript, HTML5, CSS3, Microsoft Visual Studio, Tableau, PCI DSS, HIPAA, Microsoft Office

SAM.OLISZEWSKI@GMAIL.COM
15720 SW OBERST LANE, SHERWOOD, OR 97140 • (503) 545-1844
RELOCATING TO 2769 MARION STREET, BELLMORE, NY 11710

PROJECT OVERVIEWS

Three data analysis reports are presented in this portfolio. A summary of each report is included below.

Analysis of the Behavior of Electricity Price in Oregon Households

This paper discusses an analysis of the electricity expenses from the American Community Survey for households in Oregon (ACS) during the year 2015. The analysis has both an explanatory and predictive question of interest it aims to answer. First, the explanatory question "Is there a difference in electricity expenses for people living in houses versus apartments?" is investigated. Next, the predictive question "Can a model be created to predict electricity costs for a household in Oregon?" is explored.

Analysis of the Crash Data for New Zealand Drivers

This analysis studies car crash data from New Zealand that contains information about the number of recorded crashes involving trucks, bicycles, and motorcycles grouped by the hour of the day and day of the week of the incident during the year 2009. Three different datasets were examined for this study and restructured to present the number of crashes by vehicle type during different time-of-day intervals for each day of the week. This analysis used the variable Time as a categorical description of the time of day, with the following possible values: early morning, morning, afternoon, and evening. The variable Crashes was used as the response variable for each of these datasets. Three questions were addressed in this study: 1. What day of the week should we expect to observe the greatest number of crashes? 2. How does the number of crashes change according to vehicle type? 3. How many bicycle crashes should we expect to observe on a Wednesday afternoon?

Analysis of the Behavior of Selected Sample Statistics on the Youth Risk Behavior Surveillance System Population

This paper discusses the analysis of some sample statistics using 2003 and 2013 data from a large survey of high-school students in America, called the Youth Risk Behavior Surveillance System (YRBSS). This analysis includes the inference on the population of high-school students based on the sample of YRBSS students. First, there is a determination as to whether high-school students have increased their BMI over time. Second, there is a determination as to whether male high-school students are more likely to smoke than female high-school students. Finally, there is an estimate as to how much TV the average high-schooler watches per day.

Analysis of Veteran's Administration Lung Cancer

This analysis studies data from the US Veteran's Administration involving male patients with advanced inoperable lung cancer that were randomly assigned to two treatments of either a standard therapy or a test chemotherapy. Time to death was recorded for each of the patients. Various covariates were also documented for each patient. The primary goal of the study was to assess if the test chemotherapy is beneficial. Secondary goals included the analysis of the additional covariates as prognostic variables.

Analysis of the Behavior of Electricity Price in Oregon Households

Sam Oliszewski

Oregon State University

ABSTRACT

This paper discusses an analysis of the electricity expenses from the American Community Survey for households in Oregon (ACS) during the year 2015. The analysis has both an explanatory and predictive question of interest it aims to answer. First, the explanatory question "Is there a difference in electricity expenses for people living in houses versus apartments?" is investigated. Next, the predictive question "Can a model be created to predict electricity costs for a household in Oregon?" is explored. To begin, an exploratory analysis was performed to provide insight on the data being studied. Next, the approaches for model determination in the explanatory and predictive settings are described, as well as limitations of the selected models. Finally, conclusions of the analysis are reported, and caveats of the analysis are presented to aid in the overall interpretation of the study results. It was determined that there is no significant difference in electricity expense for people living in houses versus apartments among Oregon households. Additionally, a model was presented that could be used for predicting the electricity expense of an Oregon household, but the predictive power is less than ideal. It is recommended that future research be conducted to identify additional characteristics of households that may influence the electricity expense and help make predictions more accurate.

Keywords: multiple linear regression, Oregon household electricity, predictive modeling

ANALYSIS OF THE BEHAVIOR OF ELECTRICITY PRICE IN OREGON HOUSEHOLDS

The American Community Survey (ACS) was conducted for households in Oregon during the year 2015. Two questions of interest are raised about this data— one explanatory and one predictive in nature. This analysis will first address the explanatory question "Is there a difference in electricity expenses for people living in houses versus apartments?" Next, the predictive question "Can a model be created to predict electricity costs for a household in Oregon?" is explored. The paper will begin by discussing the exploratory analysis that was performed to provide insight on the data being studied. Next, the approaches for model determination in the explanatory and predictive settings are described, in addition to discussion about the limitations of the selected models. Finally, conclusions drawn from the analysis are reported, and caveats of the analysis are offered to aid in the overall interpretation of the study results.

Exploratory Analysis

This section discusses the exploratory analysis performed to investigate the ACS dataset.

Assessing the Available Variables

Original Variables. The raw dataset contained fifteen predictor variables. The original variables included in the ACS dataset used for this analysis are: serial number (SERIALNO), type of unit (TYPE), number of people in the household (NP), lot size in acres (ACR), bedrooms in household (BDSP), units in structure (BLD), fuel cost (FULP), gas cost (GASP), house heating fuel type (HFL), number of rooms in household (RMSP), tenure (TEN), property value (VALP), year structure was built (YBL), presence of under age 18 persons (R18), and presence of over age 60 persons (R60). The response variable in this study is the price of electricity per household (ELEP). A sample of the data, as well as descriptions of possible values for each variable, can be referenced in the Appendix.

Removing Unnecessary Predictors. The dataset originally contained fifteen predictor variables and the modified set of predictors used in this analysis contained thirteen. The variables deemed unnecessary for the study were SERIALNO and TYPE. The SERIALNO variable was excluded from the data set because the value of this predictor is an identifier of the observation and has no relationship with the response variable ELEP. The TYPE variable was excluded from the data set because the value is identical for all the observations in the data set and therefore would not contribute to the change in the response variable ELEP. The predictor variables included in the reduced data set used for this analysis are: NP, ACR, BDSP, BLD, FULP, GASP, HFL, RMSP, TEN, VALP, YBL, R18, and R60. The response variable in this study is ELEP.

Addressing Missing Values

Explanation of Problem. Since there are several missing values in the data set, the number of observations able to be studied is reduced when left in its raw state. Leaving the data untouched would limit the ability to draw meaningful conclusions from the data.

Solution to the Missing Value Problem. To address the issue of missing values, imputation was performed to increase the number of observations able to be studied. There was a mix of integer and factor variables in the dataset and two different methods of imputation were performed as a result— median and mode imputation. For integer variables, median imputation was used because the median was a good measure of the center of the data that preserves the integer value in the result. Since integers are ordered, a median is a reasonable statistic to use for imputation. For the factor values, mode imputation was used. This method was preferred due to many of the levels being unordered.

Modifying Variable Structure for Different Goals

BLD in the Explanatory Problem Setting. Since the goal in the explanatory problem is to assess the difference in electricity expenses for people living in houses versus apartments, a variable needed to be chosen to classify a household as a house or an apartment. Two possible variables were available— TEN and BLD. Ultimately, the BLD variable was determined to be better for this classification because the responses for this variable were more explicit about the house versus apartment classification. However, the structure of the variable needed to be altered to fit the needs of the question. Therefore, the BLD variable was restructured from the original ten factors into two factor levels: "House" and "Apartment". The level House was taken from the original levels "One-family house detached" and "One-family house attached". The level Apartment was taken from the original levels "2 Apartments", "3-4 Apartments", "5-9 Apartments", "10-19 Apartments", "20-49 Apartments" and "50 or more apartments". After the variable was restructured, the observations in the dataset containing BLD values "Mobile home or trailer" and "Boat, RV, van, etc." were removed because they were not relevant to the study for the context of this question of interest.

BLD in the Prediction Problem Setting. In the prediction problem for this analysis, there is no need to use the restructured BLD variable as discussed previously. It is preferred to consider the original factor levels and their associated observations because the question does not specify that there should be a grouping of BLD type. Further, the specificity of BLD type could prove influential on the prediction and should thus be considered for this problem. Therefore, the BLD variable in the prediction problem will include the original ten factor levels "Mobile home or trailer", "One-family house detached", "One-family house attached", "2 Apartments", "3-4 Apartments", "5-9 Apartments", "10-19 Apartments", "20-49 Apartments", "50 or more apartments", and "Boat, RV, van, etc.". All the original observations are included in this problem.

Visualizing Relationships in Data

Figures of Predictors Against Response. The initial exploration of the data included plotting each predictor value against the response to determine any apparent relationships within the dataset. These relationships are depicted after imputation occurred. The resulting plots are shown below:

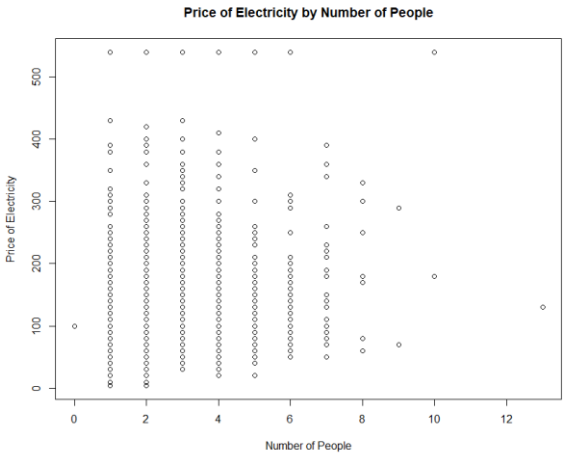


Figure 1. ELEP by NP

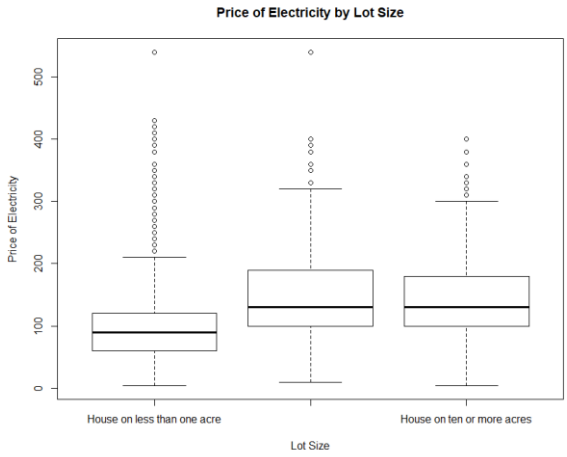


Figure 2. ELEP by ACR

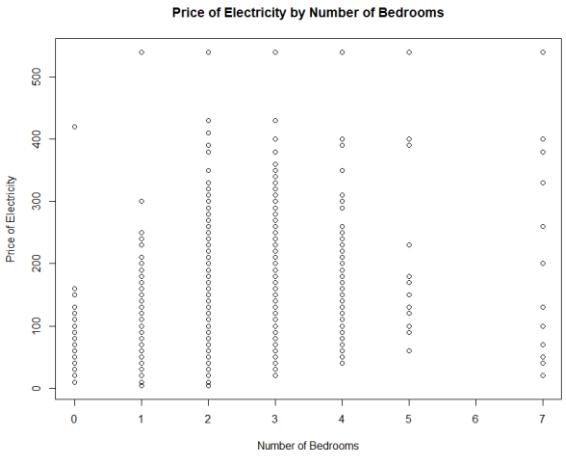


Figure 3. ELEP by BDSP

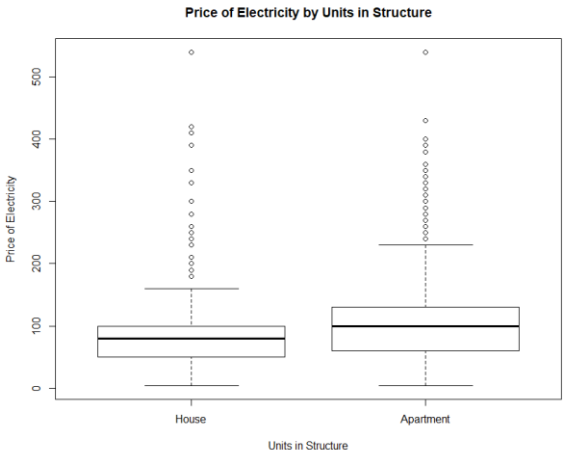


Figure 4. ELEP by BLD

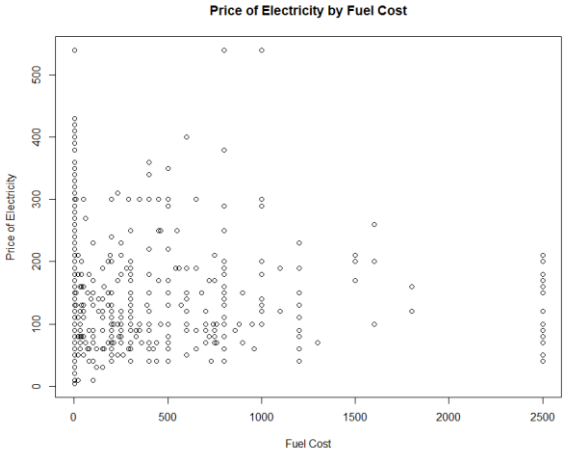


Figure 5. ELEP by FULP

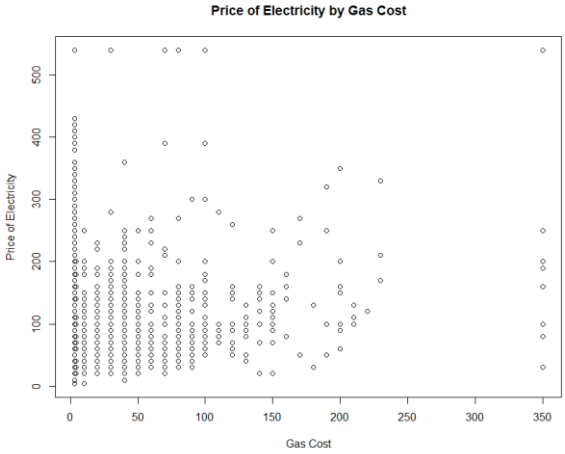


Figure 6. ELEP by GASP

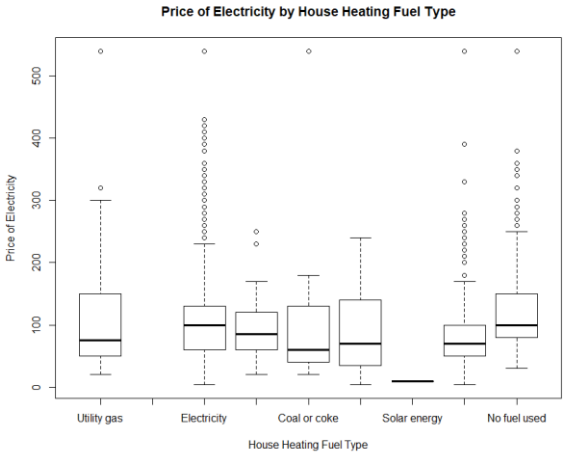


Figure 7. ELEP by HFL

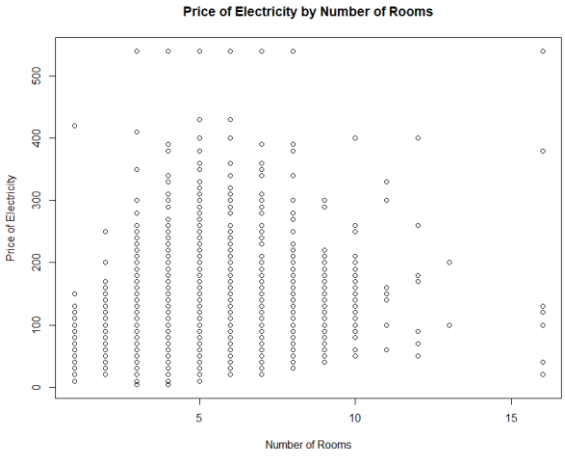


Figure 8. ELEP by RMSP

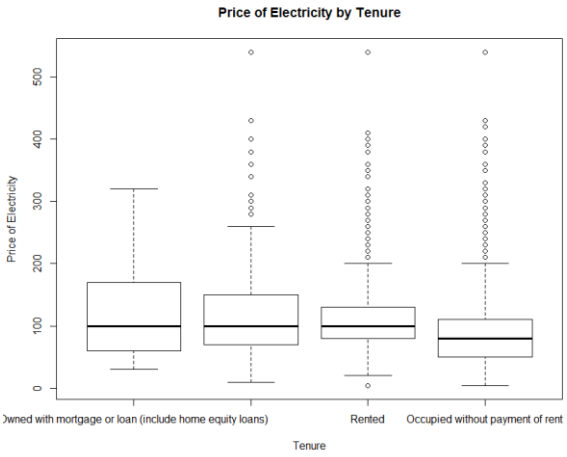


Figure 9. ELEP by TEN

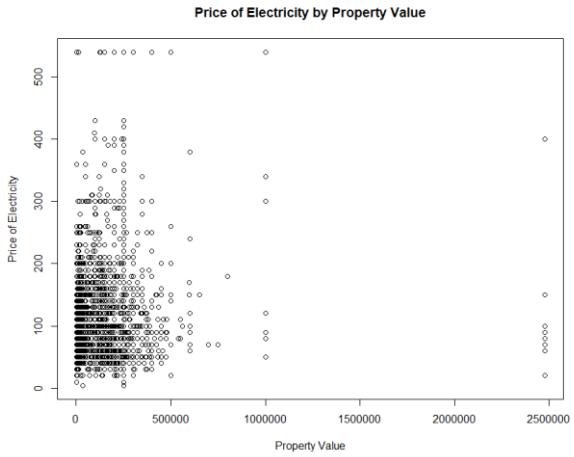


Figure 10. ELEP by VALP

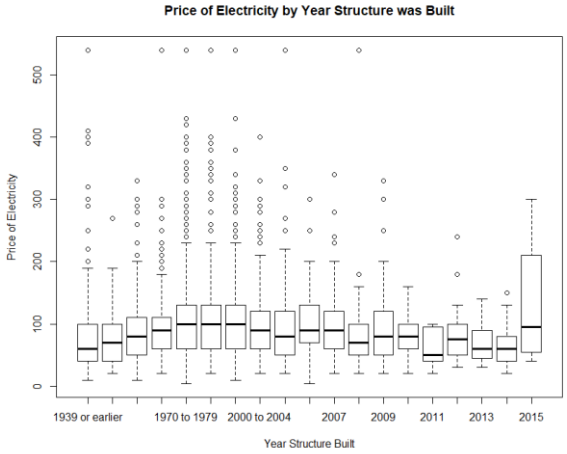


Figure 11. ELEP by YBL

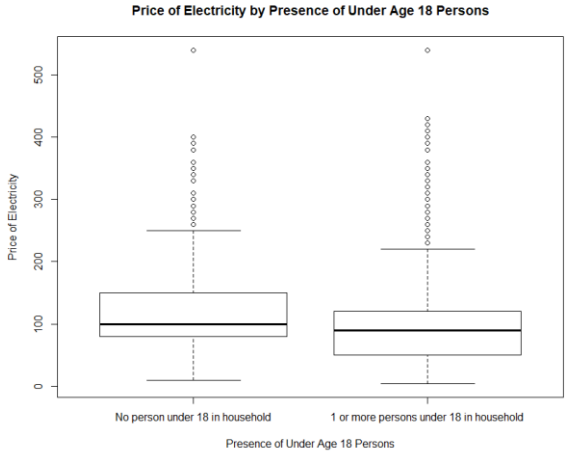


Figure 12. ELEP by R18

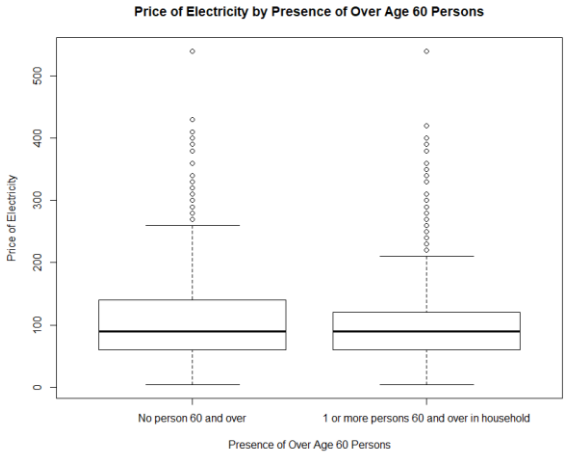


Figure 13. ELEP by R60

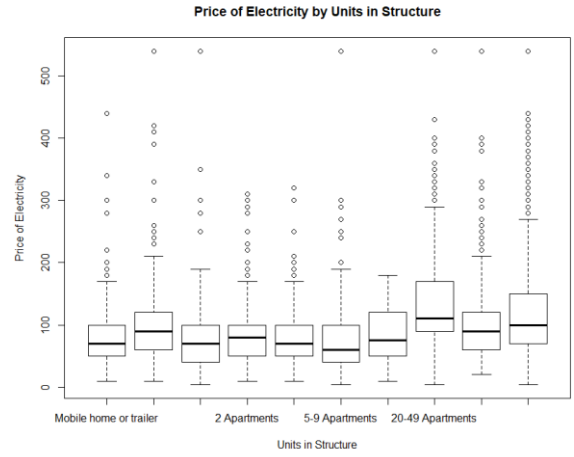


Figure 14. ELEP by Original BLD Factors

Conclusions. After examining these plots, there is no strong evidence of impact on the response variable for different values of the predictor for the predictor variables R60, TEN or VALP (most of the observations are clustered). It is seen that there is some evidence of impact on the response variable for

different values of the predictor for the predictor variables R18 (slightly higher for no 18 year old persons), YBL (clear outliers for 2008 and earlier years and 2015 is much higher), RMSP (consistency for number of rooms 4 and greater and lower prices for 3 and fewer rooms), HFL (fairly consistent medians across fuel types, but more high values for electricity), GASP (more high values when gas price is almost zero), FULP (more high values when gas price is almost zero), BLD (in the original factoring the 20-49 apartments level has a median higher than most other structure types Q3 value, and in the restructured BLD the apartment level is higher than house), BDSP (consistency with 2 or more bedrooms and lower electricity price with 0 and 1 bedrooms), ACR (lower electricity price with house on less than one acre, but more outliers), and NP (consistent electricity prices with 7 and fewer people, but lower values with 8 or more people per household).

Explanatory Problem

Overview

The goal in this section is to determine whether people living in apartments pay less on electricity than those living in houses and by how much. To make this determination, first a multiple linear regression (MLR) model must be constructed to reflect the dataset. This process will involve fitting both a model without interactions and a model with interactions. Both of the fitted models will include all the predictors in the dataset which means that the regression line will have its slope altered based on the values of each term because each predictor has an associated coefficient (beta) that represents the estimated amount by which the mean value of the response variable (ELEP) changes for a unit change in the predictor when all other predictors are held fixed. Once two models are fit, the preferred model will be determined, and a summary of the model will be reported that will identify the difference in electricity expenses between people living in apartments and people living in houses.

Methods

Fitting a Model without Interactions. An MLR model including all thirteen predictors of the following form was fit:

$$ELEP = \beta_0 + \beta_1 NP + \beta_2 ACR + \beta_3 BDSP + \beta_4 BLD + \beta_5 FULP + \beta_6 GASP + \beta_7 HFL + \beta_8 RMSP + \beta_9 TEN + \beta_{10} VALP + \beta_{11} YBL + \beta_{12} R18 + \beta_{13} R60$$

To assess the fit of this model, first a residuals versus fitted plot was generated. This plot is shown below:

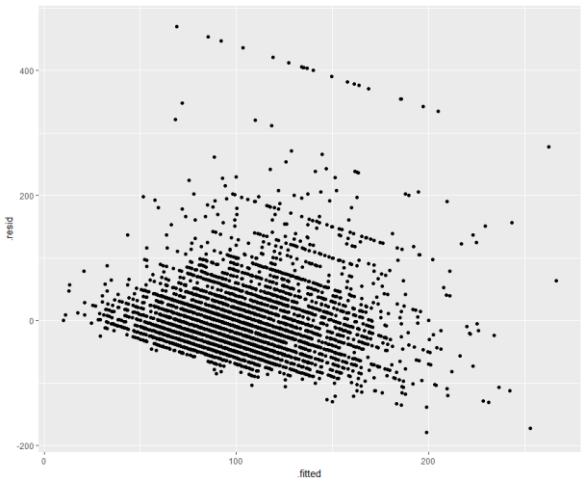


Figure 15. Residual versus Fitted – Model without Interactions

Of note in Figure 15 is the apparent parallel lines created by the plotted residuals. This is likely a reflection of the discrete nature of the response variable ELEP, forcing each value to take on integer values rather than continuous ones. This is also a possible cause of the suspected heteroscedastic behavior causing the residuals to take a shape. With this understanding of the response variable, it is reasonable to assume that the model satisfies the constant variance assumption for MLR. Further, the assumption of linearity is also satisfied because the plotted points would generally be random if the response variable were not discrete and thus MLR is still reasonable to pursue. The sample size is sufficiently large to assume that the assumption of normality for each Y around its mean is satisfied and the assumption of independence is also met because the households studied were selected at random and there is no reason to believe that one household's electricity price should be related to another's.

To further assess whether MLR is appropriate to perform, residuals versus explanatory variable plots for each of the thirteen explanatory variables were created. These plots are shown below:

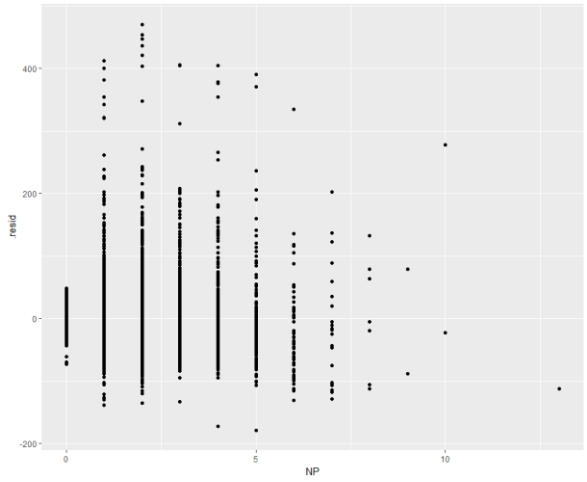


Figure 16. Residuals versus NP

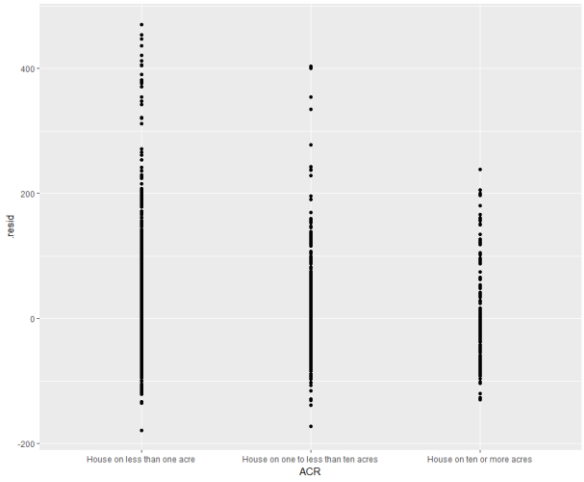


Figure 17. Residuals versus ACR

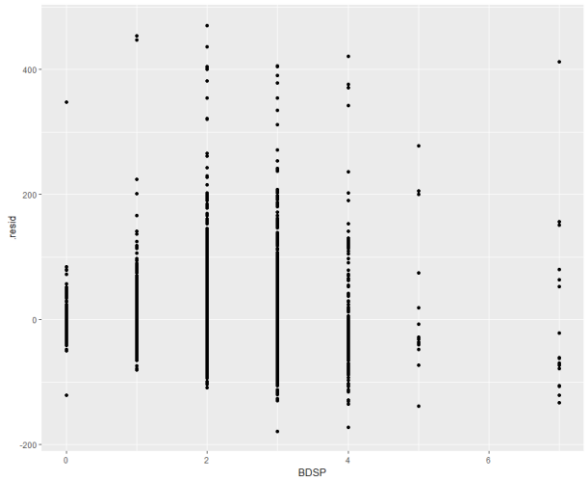


Figure 18. Residuals versus BDSP

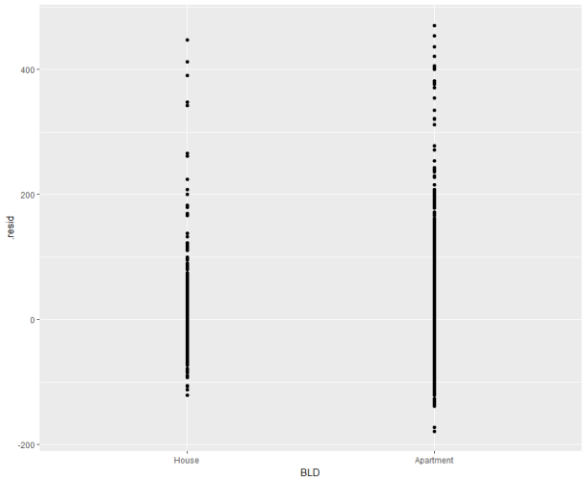


Figure 19. Residuals versus BLD

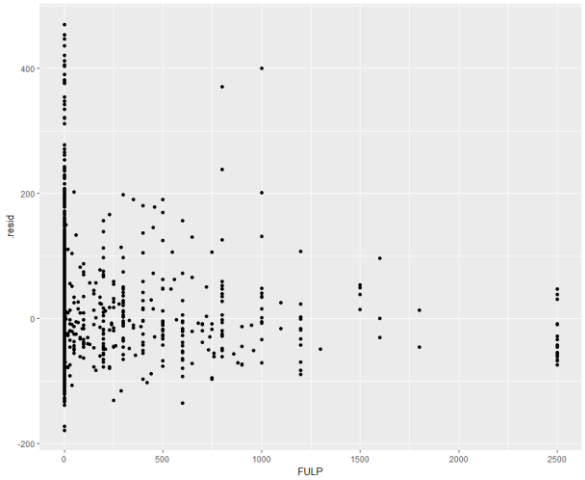


Figure 20. Residuals versus FULP

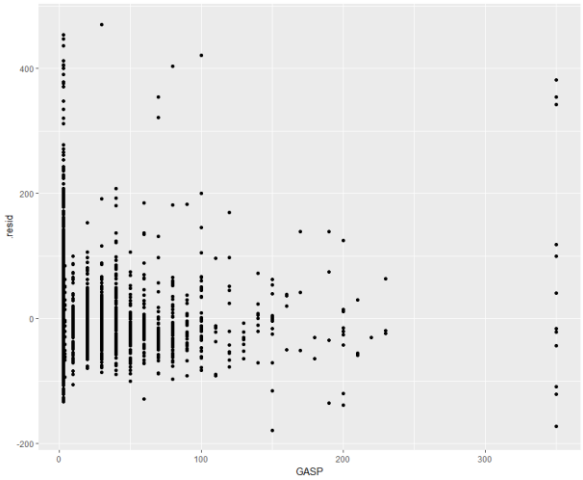


Figure 21. Residuals versus GASP

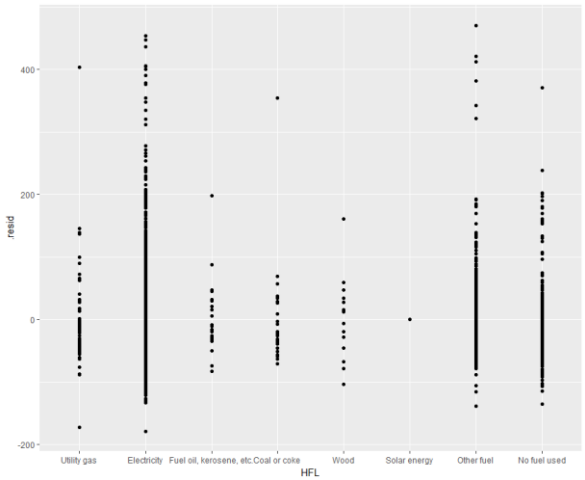


Figure 22. Residuals versus HFL

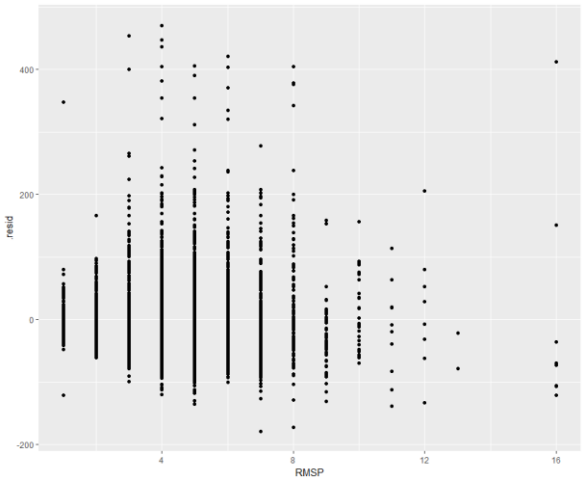


Figure 23. Residuals versus RMSP

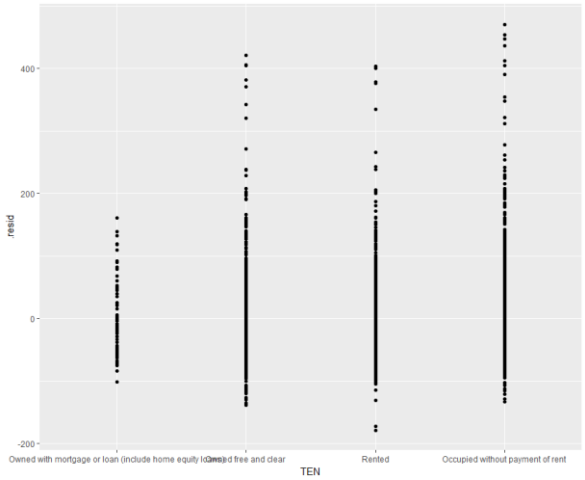


Figure 24. Residuals versus TEN

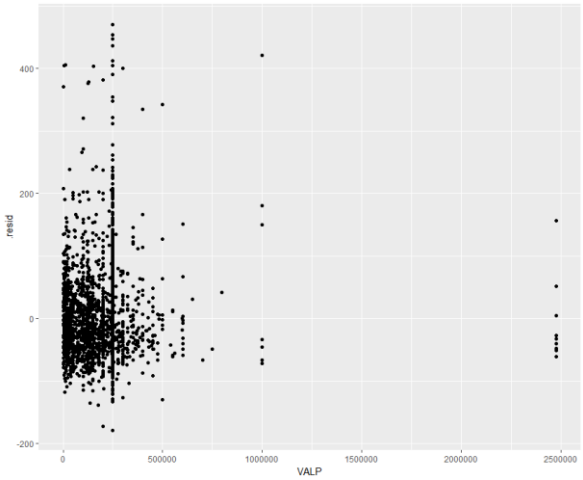


Figure 25. Residuals versus VALP

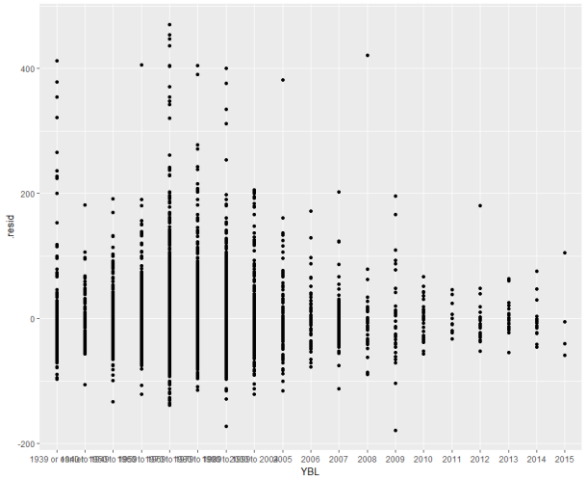


Figure 26. Residuals versus YBL

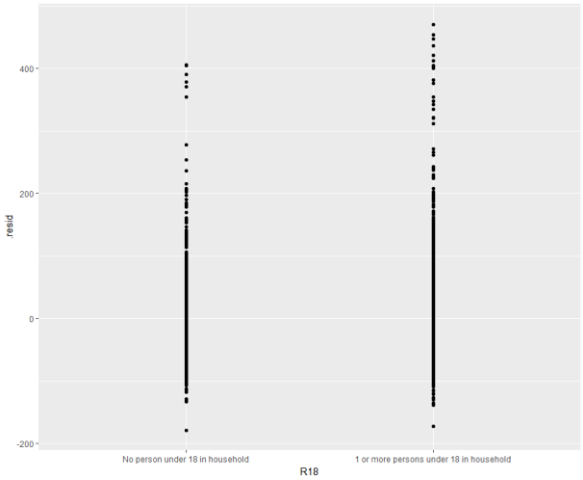


Figure 27. Residuals versus R18

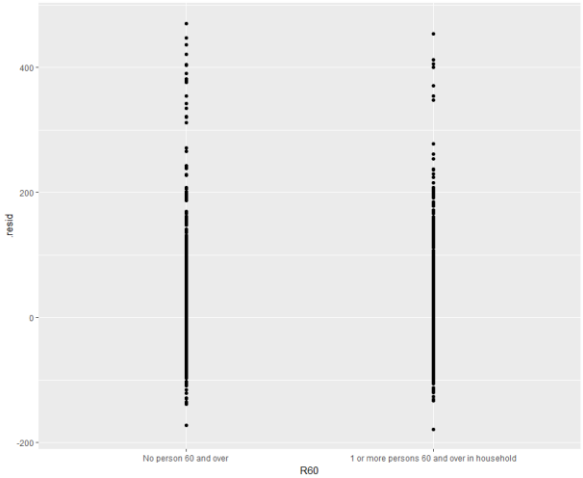


Figure 28. Residuals versus R60

While each of the thirteen residuals versus explanatory variable plots showed some values of electricity being higher than most of the data points, there are too many points to be considered outliers and do not present any evidence that the assumptions for MLR are not met. This is consistent with the findings from earlier that suggest MLR is appropriate. Further, while the residuals versus predictor plots seem to suggest non-constant variance, the variance is not due only to single predictor and therefore the residuals versus fitted plot is a more appropriate measure of model fit and does not indicate any MLR assumption violations the would halt inference from proceeding.

Fitting a Model with Interactions. A model including all thirteen of the predictors and their interactions of the following form was fit:

$$\begin{aligned} ELEP = \beta_0 + (&\beta_1 NP + \beta_2 ACR + \beta_3 BDSP + \beta_4 BLD + \beta_5 FULP + \beta_6 GASP + \beta_7 HFL + \beta_8 RMSP \\ &+ \beta_9 TEN + \beta_{10} VALP + \beta_{11} YBL + \beta_{12} R18 + \beta_{13} R60) * (\beta_1 NP + \beta_2 ACR \\ &+ \beta_3 BDSP + \beta_4 BLD + \beta_5 FULP + \beta_6 GASP + \beta_7 HFL + \beta_8 RMSP + \beta_9 TEN \\ &+ \beta_{10} VALP + \beta_{11} YBL + \beta_{12} R18 + \beta_{13} R60) \end{aligned}$$

To assess the fit of this model, a residuals versus fitted plot was generated. This plot is shown below:

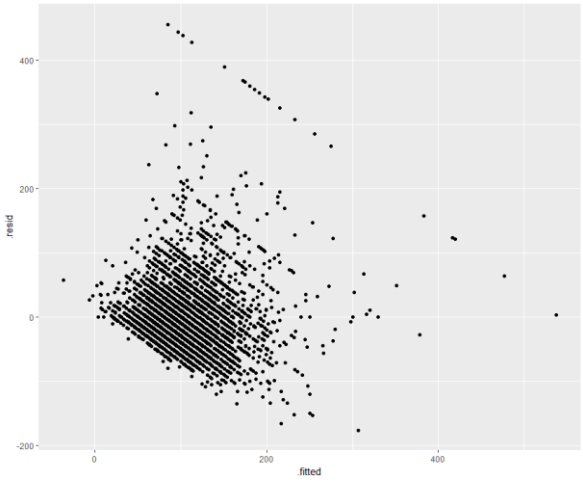


Figure 29. Residuals versus Fitted – Model with Interactions

This plot is very similar to the residuals versus fitted plot for the model without interactions and therefore does not present any evidence that the assumptions for MLR are not met, as were discussed previously.

Choosing the Best Model. After the two models were fit and the assumptions for MLR were proven to be satisfied, a decision needed to be made as to which of the models best fit the data. To make this determination, an ANOVA test was performed comparing the two models. The results of this test found that the model with interactions was preferred (Extra Sum of Squares comparing model with interaction to model without interaction, p-value = 2.2e-16). This analysis will thus proceed with this model.

Summarizing the Preferred Model. With the best model for the data selected, a summary was generated to determine whether there is a difference in the electricity expenses between Oregonians living in apartments and houses and how much that difference is. This showed that the estimated difference in the BLD term for people living in apartments versus houses is $84.23, with a standard error of $51.66, a t-statistic of 1.63, and a p-value of 0.10.

The following table summarizes all the relevant coefficients for the question of interest from the fitted model:

Term	**Point Estimate**	**Standard Error**	**t-value**	**p-value**
BLDApartment	84.23	51.66	1.63	0.10
BLDApartment:GASP	-38.11	12.49	-3.05	0.00
BLDApartment:TENOwned free and clear	-68.84	29.68	-2.32	0.02
BLDApartment:TENOccupied without payment of rent	-57.77	26.27	-2.12	0.03

Conclusion

Overall, with 95% confidence, there is no evidence that the mean price of electricity for people living in apartments is less than the mean price of electricity for people living in houses. It is estimated that the mean price of electricity for people living in houses is between $17.05 below and $185.51 above the mean price of electricity for people living in houses, with a point estimate of $84.23 (t-test, df= 4008, t =1.63, p-value = 0.10). However, it is important to note that the preferred model included interactions, which means that the influence of BLD on the electricity expense is more complicated than simply assessing the value of this one coefficient. There are three significant interaction terms including the BLD variable in the model, which also

can influence the electricity expense. These interactions appear to lower the price of electricity as well, which is an interesting observation. The interaction of an apartment household with gas price has a mean electricity price between $50.60 and $25.62 below the mean price of electricity for people without this interaction (t-test, df=4008, t=-3.05, p-value=0.00). The interaction of an apartment household with tenure that is owned free and clear has a mean electricity price between $98.52 and $39.16 below the mean price of electricity for people without this interaction (t-test, df=4008, t=-2.32, p-value=0.02). The interaction of an apartment household with tenure that is occupied without payment of rent has a mean electricity price between $84.04 and $31.50 below the mean price of electricity for people without this interaction (t-test, df=4008, t=-2.32, p-value=0.03).

Limitations. Of note on the conclusions from this study is that the values calculated may be somewhat inaccurate because the data was imputed instead of resulting solely from actual observations. This may underestimate the standard error associated with the generated values. Also noteworthy is that the missing data in the dataset requires investigation to see if it was missing completely at random as this is the only type of missing data that would not affect the inferential conclusions.

Prediction Problem

Overview

The goal in this section is to create a model that could be used to predict electricity costs for a household in Oregon. To do this, best subset selection will be performed along with k-fold cross validation to select the best model. Afterward, a summary of the best model will be reported.

Methods

K-Fold Cross Validation Approach. To determine the best model, thirteen models were fitted that included from one to thirteen predictors. Best subset selection was performed to determine which predictors to include in the model at each model size. Then, to test each model, the data was divided into k number of groups of even size. In this study, the value of k was equal to 10. Then, 10 experiments were performed where one of the 10 groups is used as a test set, and the remaining k-1 (10 - 1 = 9) groups acted as the training set. The MSE was computed for each model and the lowest test MSE among each calculated for the different model sizes was selected as the preferred model. This indicated the optimal number of predictors to include in the model. The k-fold cross validation approach resulted in a model that included all thirteen of the original predictors and had a MSE of 3846.57. The R^2 statistic for this model is 0.24.

Conclusion

The best model as chosen by the k-fold cross validation includes the following terms:

Term	Coefficient	Standard Error	t-statistic	p-value
Intercept	-5.46e00	7.04e00	-0.78	0.44
NP	1.36e01	5.23e-01	26.07	**2e-16**
BLDOne-family house detached	2.16e01	4.40e00	4.92	**8.96e-07**
BLDOne-family house attached	2.67e00	4.53e00	0.59	0.56
BLD2 Apartments	4.35e00	4.02e00	1.08	0.28
BLD3-4 Apartments	-1.84e00	4.04e00	-0.46	0.65
BLD5-9 Apartments	-9.21e-01	4.29e00	-0.22	0.83
BLD10-19 Apartments	2.96e01	1.29e01	2.29	**0.02**
BLD20-40 Apartments	3.95e01	3.67e00	10.76	**2e-16**
BLD 50 or more apartments	2.32e01	3.98e00	5.84	**5.31e-09**

BLDBoat, Rv, van, etc.	3.15e01	3.39e00	9.28	**2e-16**
ACRHouse on one to less than ten acres	2.75e01	1.64e00	16.83	**2e-16**
ACRHouse on ten or more acres	2.62e01	2.44e00	10.73	**2e-16**
BDSP	6.39e00	7.79e-01	8.20	**2.58e-16**
FULP	8.58e-03	2.12e-03	4.04	**5.35e-05**
GASP	1.48e-01	1.24e-02	11.87	**2e-16**
HFLBottled, tank, or LP gas	-2.03e01	4.50e01	-0.45	0.65
HFLElectricity	4.08e01	3.78e00	10.79	**2e-16**
HFLFuel oil, kerosene, etc.	2.00e00	5.07e00	0.39	0.69
HFLCoal or coke	2.56e01	1.01e01	2.54	**0.01**
HFLWood	2.85e01	7.15e00	3.98	**6.81e-05**
HFLSolar energy	-1.47e01	2.27e01	-0.65	0.52
HFLOther fuel	-9.19e00	3.69e00	-2.49	**0.01**
HFLNo fuel used	9.46e00	4.08e00	2.32	**0.02**
RMSP	1.69e00	3.34e-01	5.06	**4.30e-07**
TENOwned free and clear	-8.97e00	4.30e00	-2.08	**0.04**
TENRented	-2.39e00	4.24e00	-0.56	0.57
TENOccupied without payment of rent	-3.54e00	4.34e00	-0.81	0.42
VALP	2.36e-05	2.42e-06	9.76	**2e-16**
YBL1940 to 1949	8.22e00	2.54e00	3.24	**0.00**
YBL1950 to 1959	5.42e00	2.20e00	2.47	**0.01**
YBL1960 to 1969	3.34e00	2.16e00	1.55	0.12
YBL1970 to 1979	8.73e00	1.86e00	4.70	**2.65e-06**
YBL1980 to 1989	4.73e00	2.14e00	2.21	**0.03**
YBL1990 to 1999	-7.32e-01	1.93e00	-0.38	0.70
YBL2000 to 2004	-2.64e00	2.35e00	-1.12	0.26
YBL2005	-4.44e00	3.92e00	-1.13	0.26
YBL2006	-6.25e00	4.37e00	-1.43	0.15
YBL2007	-4.92e00	4.28e00	-1.15	0.25
YBL2008	-7.02e00	5.74e00	-1.22	0.22
YBL2009	2.56e00	6.10e00	0.42	0.67
YBL2010	-7.63e00	6.75e00	-1.13	0.26
YBL2011	9.72e-01	8.69e00	0.11	0.91
YBL2012	-3.89e00	7.84e00	-0.50	0.62
YBL2013	-1.20e01	7.13e00	-1.68	0.09
YBL2014	-1.15e01	7.83e00	-1.46	0.14
YBL2015	-2.27e01	1.42e01	-1.60	0.11
R181 or more persons under 18 in household	4.47e00	1.75e00	2.55	**0.01**
R601 or more persons 60 and over in household	-3.70e00	1.20e00	-3.09	**0.00**

Of note is that while this model is found to be the preferred model when compared to others, in terms of predictive power, it may not be particularly strong. The R^2 statistic is relatively low and the MSE is a bit higher than desired. This means that the predictions resulting from this model may not be as accurate as one would hope. The coefficients for each term presented in the table above indicate the influence on electricity expense that each term has based on the value of the term. One of the reasons the predictive power may be reduced is that earlier it was determined that there were interactions between some of the terms and therefore this may not be addressed fully in this predictive model.

Limitations. While this model was selected as the best fit for the data, there are some limitations to consider. One such limitation to the model is that imputation was performed on the data meaning that several observations are affected by the imputed values, rather than actual observations. As a result, the standard error is likely underestimated. Also of note is that the missing data in the dataset requires investigation to see if it was missing completely at random as this is the only type of missing data that would not affect the inferential conclusions. Another limitation to this predictive model is that these households are only representative of Oregon households and should not be considered representative of households in other states. Also affecting the performance of this predictive model is how the sample size is large and we may run into issues such as small effects being considered statistically significant, raising the question of practical versus statistical significance.

Conclusion

This analysis aimed to answer two questions of interest. The respective questions were "Is there a difference in electricity expenses for people living in houses versus apartments?" and "Can a model be created to predict electricity costs for a household in Oregon?" Ultimately, it was determined that there is no significant difference in electricity expense for people living in houses versus apartments among Oregon households. Additionally, a model was presented that could be used for predicting the electricity expense of an Oregon household, but the predictive power is less than ideal. Based on the results of this study, further investigation is recommended to identify additional characteristics of households that may influence the electricity expense and help make predictions more accurate.

APPENDIX

Sample Data:

SERIALNO	NP	TYPE	ACR	BDSP	BLD	ELEP	FULP	GASP	HFL	RMSP	TEN	VALP	YBL	R18	R60
70	4	1	House on less than one acre	2	One-family house detached	70	2	3	Wood	4	Rented	NA	1939 or earlier	1 or more	none
106	0	1	NA	2	2 Apartments	NA	NA	NA	NA	3	NA	NA	1970 to 1979	NA	NA
163	2	1	House on less than one acre	2	One-family house detached	100	600	3	Fuel oil, kerosene, etc.	7	Owned with mortgage or loan	225000	1939 or earlier	none	none
178	1	1	House on less than one acre	3	One-family house detached	60	2	110	Utility gas	8	Owned free and clear	315000	1939 or earlier	none	1 or more
243	2	1	House on less than one acre	4	One-family house detached	80	2	20	Utility gas	8	Owned free and clear	200000	1950 to 1959	none	1 or more

Description of Variables:

SERIALNO: Housing unit/GQ person serial number
***Possible Values*:** Any Discrete Integer

NP: Number of person records following this housing record
Possible Values: Integer between 0 and 20

TYPE: Type of unit
Possible Values: Categorical Integer

- (1) Housing unit
- (2) Institutional group quarters
- (3) Noninstitutional group quarters

ACR: Lot size
Possible Values: Categorical Text

- N/A
- House on less than one acre
- House on one to less than ten acres
- House on ten or more acres

BDSP: Number of bedrooms
Possible Values: Integer between 0 and 99

BLD: Units in structure
Possible Values: Categorical Text

- N/A (GQ)
- Mobile home or trailer
- One-family house detached
- One-family house attached
- 2 Apartments
- 3-4 Apartments
- 5-9 Apartments
- 10-19 Apartments
- 20-49 Apartments
- 50 or more apartments
- Boat, RV, van, etc.

APPENDIX CONTINUED

ELEP: Electricity (monthly cost)
Possible Values: Discrete Integer

- (bbb) N/A (GQ/vacant)
- (001) Included in rent or in condo fee
- (002) No charge or electricity not used
- (003 … 999) $3 to $999 (Rounded and top-coded)

FULP: Fuel cost (yearly cost for fuels other than gas and electricity)
Possible Values: Categorical Text

- (bbbb) N/A (GQ/vacant)
- (0001) Included in rent or in condo fee
- (0002) No charge or these fuels not used
- (0003 ... 9999) $3 to $9999 (Rounded and top-coded)

GASP: Gas (monthly cost)
Possible Values: Categorical Text

- (bbb) N/A (GQ/vacant)
- (001) Included in rent or in condo fee
- (002) Included in electricity payment
- (003) No charge or gas not used
- (004 ... 999) $4 to $999 (Rounded and top-coded)

HFL: House heating fuel
Possible Values: Categorical Text

- N/A (GQ/vacant)
- Utility gas
- Bottled, tank, or LP gas
- Electricity
- Fuel oil, kerosene, etc.
- Coal or coke
- Wood
- Solar energy
- Other fuel
- No fuel used

RMSP: Number of Rooms
Possible Values: Discrete Integer between 0 and 99

TEN: Tenure
Possible Values: Categorical Text

- N/A (GQ/vacant)
- Owned with mortgage or loan (include home equity loans)
- Owned free and clear
- Rented
- Occupied without payment of rent

VALP: Property value
Possible Values: Integer between 1 and 9,999,999

APPENDIX CONTINUED

YBL: When structure first built
Possible Values: Categorical Text

- N/A (GQ)
- 1939 or earlier
- 1940 to 1949
- 1950 to 1959
- 1960 to 1969
- 1970 to 1979
- 1980 to 1989
- 1990 to 1999
- 2000 to 2004
- 2005

R18: Presence of persons under 18 years in household (unweighted)
Possible Values: Categorical Text

- N/A (GQ/vacant)
- No person under 18 in household
- 1 or more persons under 18 in household

R60: Presence of persons 60 years and over in household (unweighted)
Possible Values: Categorical Text

- N/A (GQ/vacant)
- No person 60 and over
- 1 person 60 and over
- 2 or more persons 60 and over

Analysis of the Crash Data for New Zealand Drivers

Sam Oliszewski

Oregon State University

ABSTRACT

This analysis studies car crash data from New Zealand that contains information about the number of recorded crashes involving trucks, bicycles, and motorcycles grouped by the hour of the day and day of the week of the incident during the year 2009. Three different datasets were examined for this study and restructured to present the number of crashes by vehicle type during different time-of-day intervals for each day of the week. This analysis used the variable Time as a categorical description of the time of day, with the following possible values: early morning, morning, afternoon, and evening. The variable Crashes was used as the response variable for each of these datasets. Three questions were addressed in this study: 1. What day of the week should we expect to observe the greatest number of crashes? 2. How does the number of crashes change according to vehicle type? 3. How many bicycle crashes should we expect to observe on a Wednesday afternoon? The results from the data analysis showed: 1. The day of the week that one would expect to observe the greatest number of crashes varies by vehicle type and was determined to be Thursday, Friday, and Saturday for bicycles, trucks, and motorcycles, respectively. 2. Since the time of the day that most crashes are expected to occur is the same for all three of the vehicle types, the number of crashes changes according to vehicle type in that each vehicle has different days where one would expect to observe the greatest number of crashes. 3. It was determined that between 60 and 72.52 bicycle crashes are expected to be observed on a Wednesday afternoon.

Keywords: log-linear modeling, negative binomial modeling, New Zealand car crashes

ANALYSIS OF THE CRASH DATA FOR NEW ZEALAND DRIVERS

Three sets of car crash data from New Zealand drivers were examined in this data analysis. Each dataset contained information about the number of crashes involving a given vehicle type (trucks, bicycles, or motorcycles) grouped by the hour of the day and day of the week that the incident occurred during the year 2009. The original datasets were restructured to present the number of crashes for a given vehicle type during different time-of-day intervals for each day of the week. Two explanatory questions and one predictive question were addressed in this study: 1. What day of the week should we expect to observe the greatest number of crashes? 2. How does the number of crashes change according to vehicle type? 3. How many bicycle crashes should we expect to observe on a Wednesday afternoon? This paper will begin by discussing the exploratory analysis that was performed to provide insight on the data being studied. Next, the approach for model determination is described, in addition to discussion about the limitations of the selected models. Finally, conclusions drawn from the analysis are reported, and caveats of the analysis are offered to aid in the overall interpretation of the study results.

Exploratory Analysis

This section discusses the exploratory analysis performed to investigate the datasets.

Examining the Data

Original Variables. The raw data for this analysis came from three separate datasets in the R library VGAM: crashtr, crashmc, and crashbc. Each dataset was of the same structure and contained variables for each of the seven days of the week and row numbers corresponding to the time of day that crashes were observed. A sample of the data, as well as descriptions of possible values for each variable, can be referenced in Appendices A, B, and C.

Restructuring the Data. To perform this analysis, each of the original datasets was restructured. Each of the restructured datasets contained the following variables: Day, Time, and Crashes. Also of note is that the Time variable was converted into four general time-of-day categories, rather than representing the exact hour of the day. These categories were as follows: Early Morning, Morning, Afternoon, and Evening. The associated time frames for each category are: 12:00 AM-4:59 AM (Early Morning), 5:00 AM-11:59 AM (Morning), 12:00 PM-5:59 PM (Afternoon), and 6:00 PM-11:59 PM (Evening). A sample of the data, as well as descriptions of possible values for each variable, can be referenced in Appendices D, E, and F.

Visualizing Relationships in the Data

Figures of Predictors Against Response. To begin this analysis, it is beneficial to examine plots for each vehicle type depicting how the day of the week and time of the day are related to the number of crashes. These plots are shown below:

` Figure 3. Bicycle Crashes by Day and Time

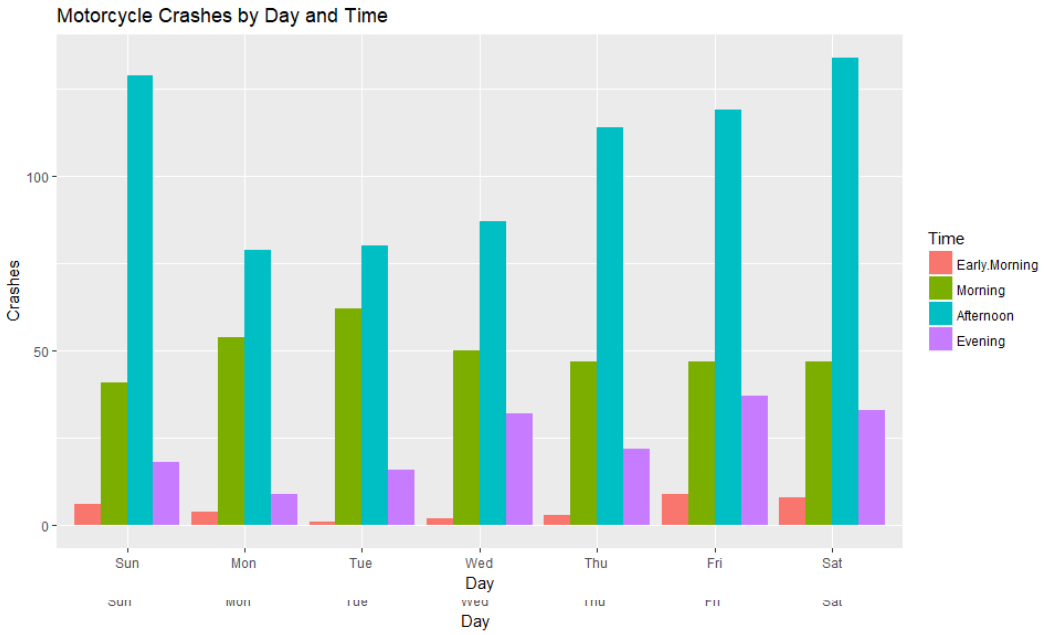


Figure 4. Motorcycle Crashes by Day and Time

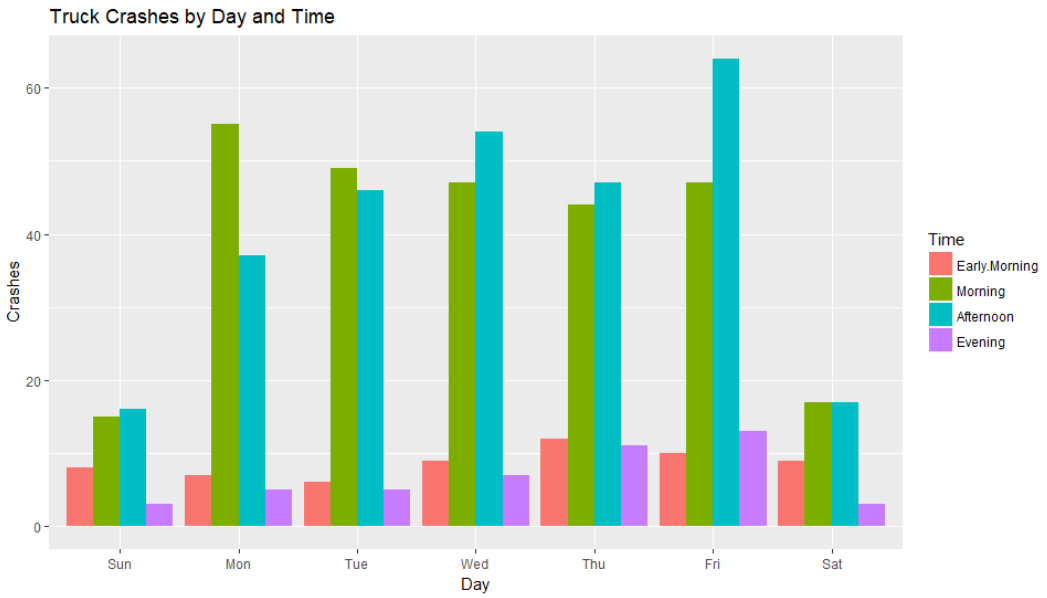


Figure 5. Truck Crashes by Day and Time

Conclusions. After examining these plots, there is some evidence of influence on the number of crashes according to the day of the week and time of the day for each of the vehicle types. For all vehicle types, the most crashes appear to occur in the morning and afternoons. For trucks and bicycles, the most

crashes appear to occur during weekdays, whereas for motorcycles the most crashes appear to occur on weekends. This is likely related to the popular types of vehicles used for transportation to and from work as well as vehicles used for commercial use. Additionally, there is no evidence of zero inflation in the data because there is only one occurrence of a zero count; therefore, there is no need to consider zero inflated models when determining the appropriate model to fit the data.

Explanatory Problems

Overview

There are two goals in this section: 1. To determine which day of the week we expect to observe the greatest number of crashes. 2. To determine how the number of crashes changes according to vehicle type. To accomplish these goals, first a log-linear regression model must be constructed for each of the three datasets. These models will help determine the dispersion parameter for the data. If over-dispersion is observed in any of the datasets, a negative binomial model will then be fit to the data. Once the best model has been identified for each of the datasets, a summary of the model will be reported that will identify the difference in the number crashes with respect to the day of the week. This will indicate which day of the week we would expect to observe the greatest number of crashes. The models will also indicate during which time of the day the greatest number of crashes are expected to occur as well. To determine how the number of crashes changes according to vehicle type, a comparison of the day of the week and time of the day where the greatest number of crashes are expected for each vehicle type will be performed.

Methods

Fitting a Log-Linear Model. A log-linear regression model for each vehicle dataset including the predictors Day and Time of the following form was fit:

$$log(Crashes_i) = \beta_0 + \beta_1 DayMon_i + \beta_2 DayTue_i + \beta_3 DayWed_i + \beta_4 DayThu_i +$$
$$\beta_5 DayFri_i + \beta_6 DaySat_i + \beta_7 TimeMorning_i +$$
$$\beta_8 TimeAfternoon_i + \beta_9 TimeEvening_i$$

First, the calculation of the dispersion parameter for each model must be calculated. The dispersion parameter for the truck, motorcycle, and bicycle datasets are 1.33, 2.95, and 1.12, respectively. This suggests there is evidence of over dispersion in the motorcycle dataset and a negative binomial model should be fit to this data. For the truck and bicycle datasets, there is no strong evidence of over dispersion according to the dispersion parameter alone. For these datasets, evaluating the residuals for the fitted models would give more insight as to whether the log-linear model is appropriate to fit.

These residual plots are shown below:

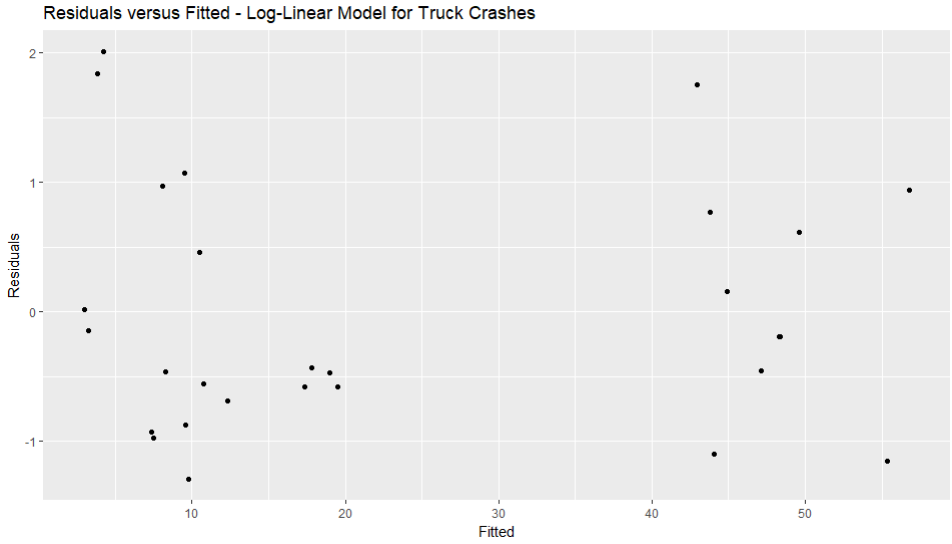


Figure 6. Residuals versus Fitted Truck Data

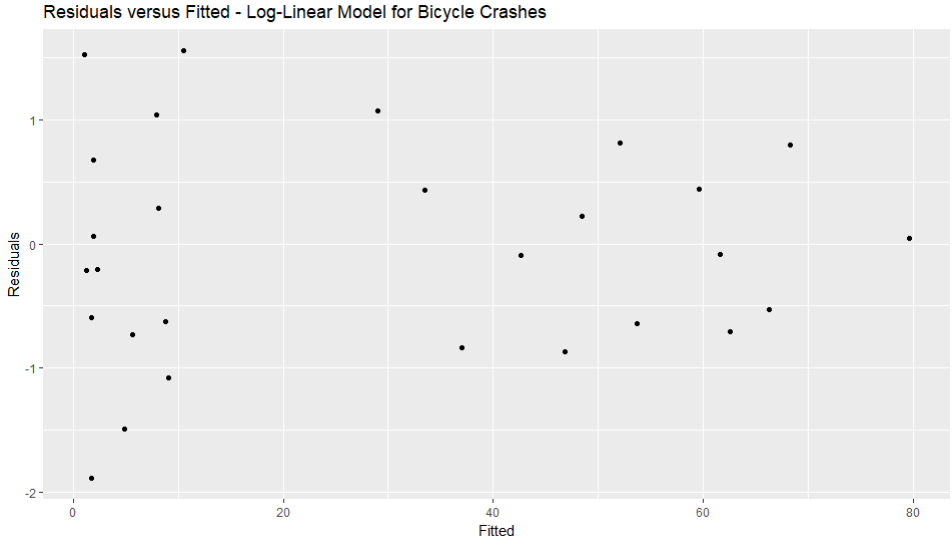


Figure 5. Residuals versus Fitted Bicycle Data

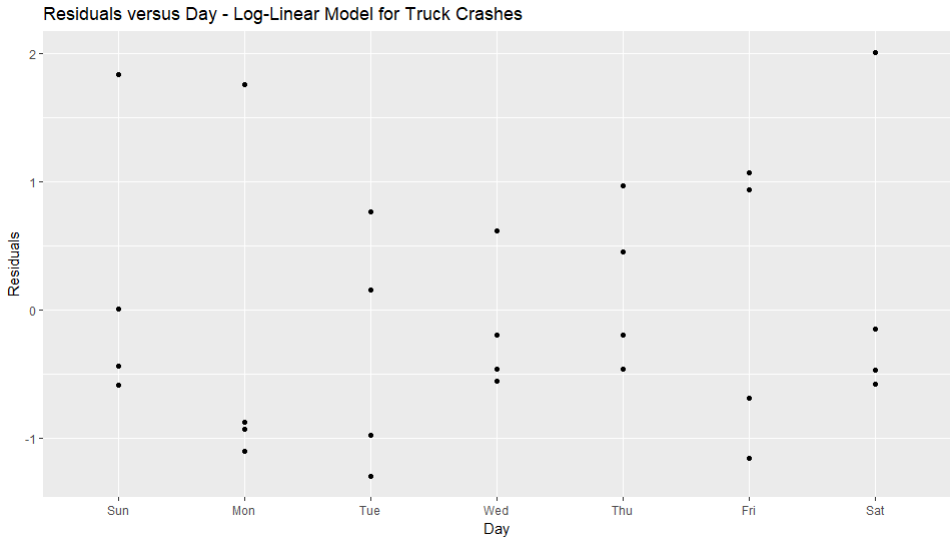


Figure 6. Residuals versus Day Truck Data

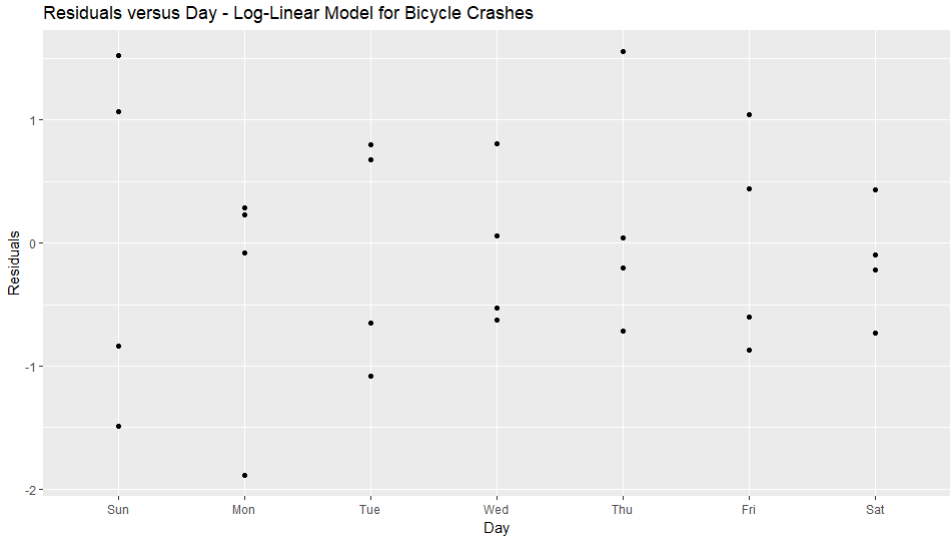


Figure 7. Residuals versus Day Bicycle Data

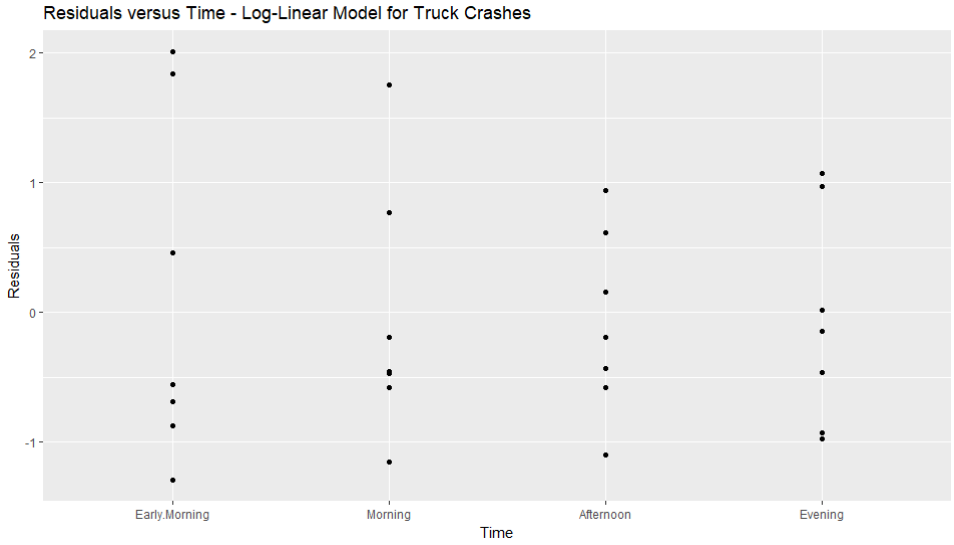


Figure 8. Residuals versus Time Truck Data

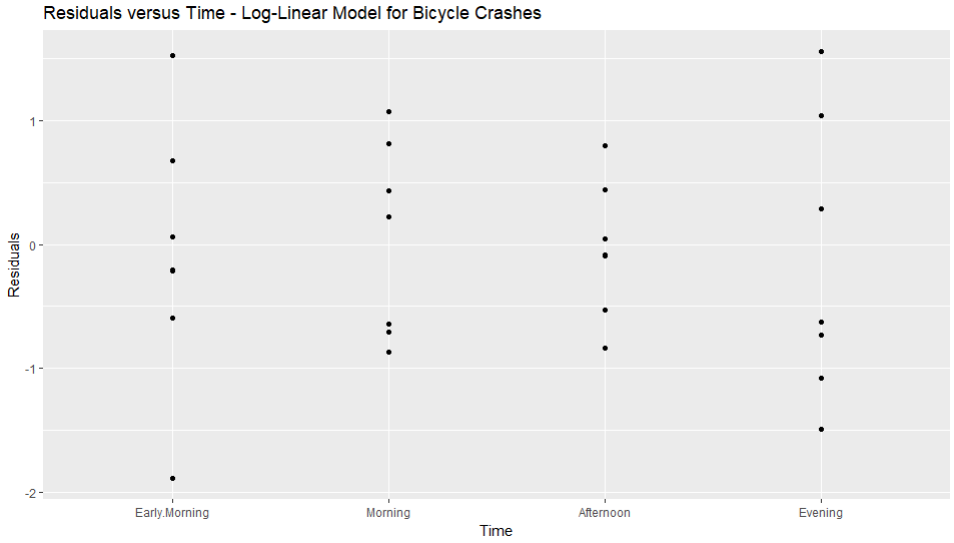


Figure 9. Residuals versus Time Bicycle Data

Based on these residual plots, there is no evidence that the log-linear model is inappropriate for the truck and bicycle datasets. The residuals are all relatively close to zero, ranging mostly between -2 and 2, and do not show any patterns that would suggest they are not normally distributed. Therefore, log-linear models are appropriate to apply to the truck and bicycle datasets.

Fitting a Negative Binomial Model. A negative binomial model for the motorcycle dataset including the predictors Day and Time of the following form was fit:

$$log(Crashes_i) = \beta_0 + \beta_1 DayMon_i + \beta_2 DayTue_i + \beta_3 DayWed_i + \beta_4 DayThu_i +$$

$$\beta_5 DayFri_i + \beta_6 DaySat_i + \beta_7 TimeMorning_i +$$

$$\beta_8 TimeAfternoon_i + \beta_9 TimeEvening_i$$

To assess the fit of this model, first a residuals versus fitted plot was generated. This plot is shown below:

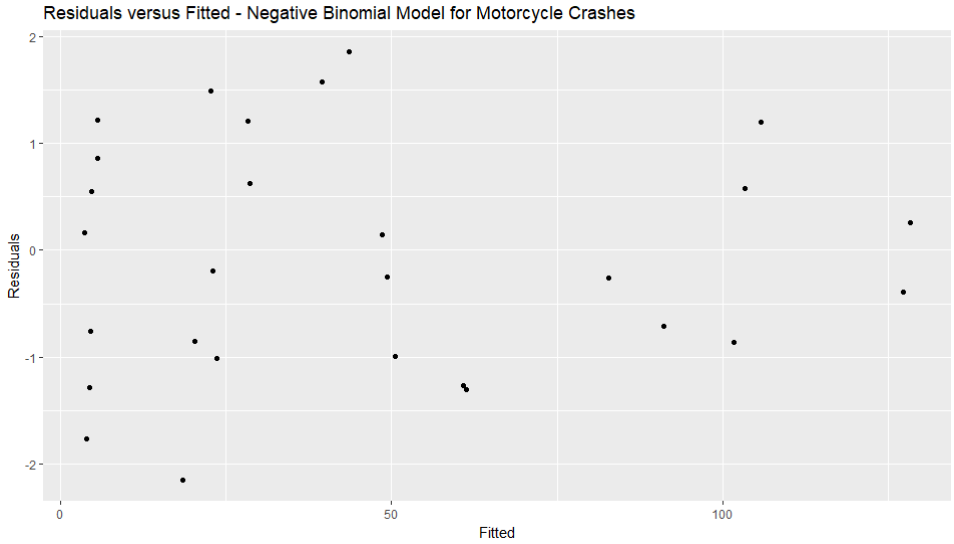


Figure 10. Residuals versus Fitted Motorcycle Data

To further assess whether the negative binomial model is appropriate to use, residuals versus explanatory variable plots for both of the explanatory variables were created. These plots are shown below:

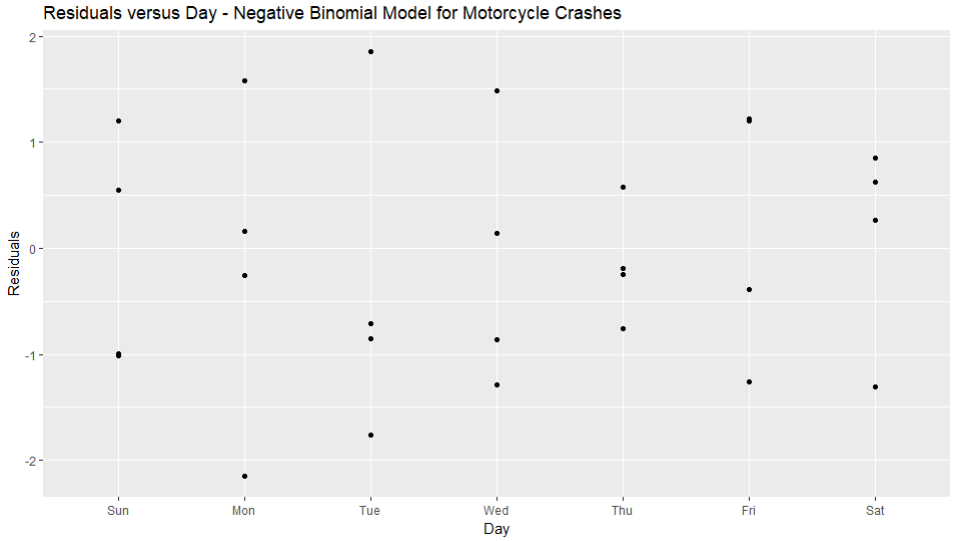


Figure 11. Residuals versus Day Motorcycle Data

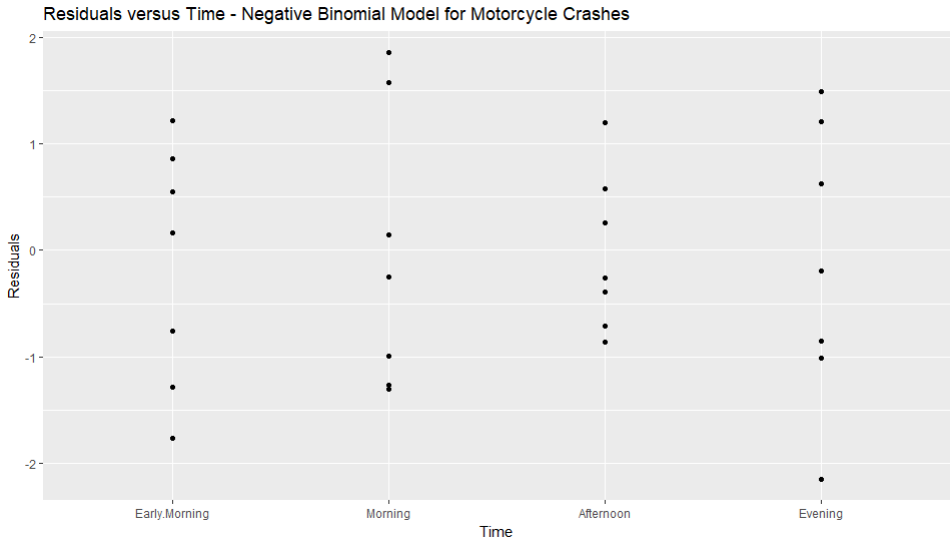


Figure 12. Residuals versus Time Motorcycle Data

Based on these residual plots, there is no evidence the negative binomial model is inappropriate for this data. The residuals are all relatively close to zero, mostly ranging between -2 and 2, and do not exhibit any pattern that would suggest they are not normally distributed. Therefore, the negative binomial model is appropriate to apply to the motorcycle data.

Summarizing the Preferred Models. With the best model for each dataset selected, a summary was generated to determine the following: 1. Which day of the week we expect to observe the greatest number of crashes. 2. How the number of crashes changes according to vehicle type.

The following table summarizes the fitted log-linear model for the truck data:

Term	**Point Estimate**	**Standard Error**	**z-value**	**p-value**
Intercept	1.35	0.20	6.87	**6.33e-12**
DayMon	0.91	0.18	4.96	**7.07e-07**
DayTue	0.93	0.18	5.08	**3.82e-07**
DayWed	1.02	0.18	5.70	**1.23e-08**
DayThu	1.00	0.18	5.53	**3.17e-08**
DayFri	1.16	0.18	6.56	**5.36e-11**
DaySat	0.09	0.21	0.43	0.67
TimeMorning	1.50	0.14	10.61	**2e-16**
TimeAfternoon	1.53	0.14	10.81	**2e-16**
TimeEvening	-0.26	0.19	-1.34	0.18

The following table summarizes the fitted log-linear model for the bicycle data:

Term	**Point Estimate**	**Standard Error**	**z-value**	**p-value**
Intercept	0.07	0.31	0.22	0.83
DayMon	0.51	0.15	3.43	**0.00**
DayTue	0.61	0.15	4.19	**2.74e-05**
DayWed	0.58	0.15	3.96	**7.37e-05**
DayThu	0.77	0.14	5.38	**7.61e-08**
DayFri	0.48	0.15	3.18	**0.00**
DaySat	0.14	0.16	0.88	0.38

TimeMorning	3.30	0.29	11.23	**2e-16**
TimeAfternoon	3.54	0.29	12.10	**2e-16**
TimeEvening	1.52	0.32	4.78	**1.77e-06**

The following table summarizes the fitted negative binomial model for the motorcycle data:

Term	**Point Estimate**	**Standard Error**	**z-value**	**p-value**
Intercept	1.55	0.21	7.35	**1.97e-13**
DayMon	-0.25	0.16	-1.49	0.14
DayTue	-0.15	0.16	-0.92	0.36
DayWed	-0.04	0.16	-0.25	0.80
DayThu	-0.02	0.16	-0.15	0.88
DayFri	0.18	0.16	1.18	0.24
DaySat	0.19	0.16	1.24	0.22
TimeMorning	2.38	0.20	12.01	**2e-16**
TimeAfternoon	3.11	0.19	16.04	**2e-16**
TimeEvening	1.62	0.21	7.85	**4.21e-15**

Using these models, the day of the week that we would expect to observe the greatest number of crashes can be determined by identifying the largest positive coefficient among the day of week terms. Sunday is incorporated into the intercept term and therefore any positive coefficient for other days of the week would indicate that it has a greater impact on the number of crashes than Sunday. For trucks, the day of the week we expect to observe the greatest number of crashes on is Friday, with a point estimate of 1.16, a standard error of 0.18, a z-statistic of 6.56, and a p-value of 5.36e-11. For bicycles, the day of the week we expect to observe the greatest number of crashes on is Thursday, with a point estimate of 0.77, a standard error of 0.14, a z-statistic of 5.38, and a p-value of 7.61e-08. For motorcycles, the day of the week we expect to observe the greatest number of crashes on is Saturday, with a point estimate of 0.19, a standard error of 0.16, a z-statistic of 1.24, and a p-value of 0.22. For all three vehicle types, the time of day when we expect to observe the greatest number of crashes is the afternoon. For trucks, the afternoon time indicator term had a point estimate of 1.53, a standard error of 0.14, a z-statistic of 10.81, and a p-value of 2e-16. For bicycles, the afternoon time indicator term had a point estimate of 3.54, a standard error of 0.29, a z-statistic of 12.10, and a p-value of 2e-16. For motorcycles, the afternoon time indicator term had a point estimate of 3.11, a standard error of 0.19, a z-statistic of 16.04, and a p-value of 2e-16. Based on these results, one can compare how the number of crashes changes according to vehicle type by comparing how day and time influence the number of crashes for each vehicle type.

Conclusion

Based on the raw output discussed above, several conclusions can be made and the two explanatory questions this analysis studied can be answered.

Identifying the Day When the Most Crashes Are Expected to Occur. Based on the summary of the fitted models for the truck, bicycle, and motorcycle data, the day that we expect the greatest number of crashes to occur differs for each vehicle type. For trucks, Friday is when we expect to observe the greatest number of crashes and with 95% confidence this day of the week is expected to increase the number of crashes by between 2.66 and 3.82 crashes. For bicycles, Thursday is when we expect to observe the greatest number of crashes and with 95% confidence this day of the week is expected to increase the number of crashes by between 1.88 and 2.48 crashes. For motorcycles, Saturday is when we expect to observe the

greatest number of crashes and with 95% confidence this day of the week is expected to increase the number of crashes by between 1.03 and 1.42 crashes.

Identifying How the Number of Crashes Changes by Vehicle Type. Based on the summary of the fitted models for the truck, bicycle, and motorcycle data, in addition to the discussion above regarding the influence on the number of crashes according to the day of the week, the crash behavior for each vehicle type can be compared. As noted previously, the day of the week where the greatest number of crashes is expected to occur varies for each vehicle type. Therefore, this is the first notable observation of how the number of crashes changes by vehicle type. Secondly, for all three vehicle types, the time of day when the greatest number of crashes is expected to occur is the same— the afternoon. This indicates that crash behavior for each vehicle type is similar based on the time of day, but each vehicle type has a different day of the week where more crashes are expected and may indicate that the odds of a crash are higher.

Limitations. Of note on the conclusions from this study is that there is no information in the original data that indicates the number of vehicles on the road for each vehicle type. Therefore, it is difficult to interpret the results of this analysis to determine whether the odds of a crash are more likely at a certain time of day or day of the week. All that can be determined is how the number of crashes is influenced by these predictors. Without context for these values, the results of this study have questionable practical significance.

Prediction Problem

Overview

The goal in this section is to use a model to predict the number of crashes that we expect to observe on a given day of the week and time of day. The objective is to use the predictive model to determine how many bicycle crashes would occur on a Wednesday afternoon. A summary of the prediction results will be reported.

Methods

Fitting a Log-Linear Model. In the previous section, a log-linear regression model for the bicycle dataset including the predictors Day and Time of the following form was fit:

$$log(Crashes_i) = \beta_0 + \beta_1 DayMon_i + \beta_2 DayTue_i + \beta_3 DayWed_i + \beta_4 DayThu_i +$$

$$\beta_5 DayFri_i + \beta_6 DaySat_i + \beta_7 TimeMorning_i +$$

$$\beta_8 TimeAfternoon_i + \beta_9 TimeEvening_i$$

Recall, that the following table summarizes the fitted log-linear model for the bicycle data:

Term	Point Estimate	Standard Error	z-value	p-value
Intercept	0.07	0.31	0.22	0.83
DayMon	0.51	0.15	3.43	**0.00**
DayTue	0.61	0.15	4.19	**2.74e-05**
DayWed	0.58	0.15	3.96	**7.37e-05**
DayThu	0.77	0.14	5.38	**7.61e-08**
DayFri	0.48	0.15	3.18	**0.00**
DaySat	0.14	0.16	0.88	0.38
TimeMorning	3.30	0.29	11.23	**2e-16**
TimeAfternoon	3.54	0.29	12.10	**2e-16**
TimeEvening	1.52	0.32	4.78	**1.77e-06**

Using this log-linear model, we can predict the number of crashes we would expect to occur on a given day of the week and time of day. For the prediction interval, we simply increase the level of error in the confidence interval. Using this model, the number of crashes on a Wednesday afternoon has a point estimate of 66.26, with a standard error of 6.26.

Conclusion

With 95% confidence, the number of bicycle crashes we expect to observe on a Wednesday afternoon is between 60 and 72.52. Referring to the initial plot of the number of crashes according to day and time from the exploratory analysis, this value is reasonable for the given day and time for bicycles.

Limitations. This model was able to determine the number of bicycle crashes one would expect to observe on a given day and time, but there are caveats to this model's predictive power. For example, all of the predictors in this model are indicator terms. Therefore, there are only a finite number of combinations possible and thus this model isn't particularly applicable to other settings. If other predictors were introduced into the model, there would be more room for honest prediction to occur, but this model simply can only evaluate what the expected number of crashes would be for a given day and time for bicycles in New Zealand in 2009. This does not give a whole lot of insight into what future values may be and thus has limited practical significance.

Conclusion

This analysis aimed to answer three questions of interest: 1. What day of the week should we expect to observe the greatest number of crashes? 2. How does the number of crashes change according to vehicle type? 3. How many bicycle crashes should we expect to observe on a Wednesday afternoon? Ultimately, it was determined that the day of the week that one would expect to observe the greatest number of crashes varies by vehicle type and was determined to be Thursday, Friday, and Saturday for bicycles, trucks, and motorcycles, respectively. Additionally, it was determined that the time of day that most crashes are expected to occur is the same for all three of the vehicle types— the afternoon. Therefore, the number of crashes changes according to vehicle type in that each vehicle has different days where one would expect to observe the greatest number of crashes. Finally, it was determined that between 60 and 72.52 bicycle crashes are expected to be observed on a Wednesday afternoon. To improve the value of this analysis, further research should be performed to determine the number of vehicles on the road according to the time of day and day of the week so that the number of crashes can be put into proper context. Additionally, data should be collected over multiple years to determine if there is trend, cycles, or seasonality in the data. Lastly, it would be beneficial to study how the number of crashes by day and time changes at different points of the year. This could be studied by recording the number of crashes by time for each day of the year and indicating what day of the week the date fell on. This additional research would provide a lot of meaningful insight into the crash behavior in New Zealand.

APPENDIX A

Dataset: crashtr (VGAM)
Description: This dataset from the VGAM library in R contains the number of crashes by day of the week and time of the day involving trucks in New Zealand during 2009.

Sample Data:

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
0	0	1	1	1	0	2	1
1	1	1	1	0	2	0	2
2	0	3	2	1	1	3	2
3	0	0	2	1	1	0	3
4	3	0	3	4	2	3	0

Description of Variables:

Variable	Description	Possible Values
{Row Num}	Corresponds to the time of day the crash occurred (0 = 12:00 AM).	Integer from 0-23
Mon	The number of crashes involving trucks on Monday.	Any integer
Tue	The number of crashes involving trucks on Tuesday.	Any integer
Wed	The number of crashes involving trucks on Wednesday.	Any integer
Thu	The number of crashes involving trucks on Thursday.	Any integer
Fri	The number of crashes involving trucks on Friday.	Any integer
Sat	The number of crashes involving trucks on Saturday.	Any integer
Sun	The number of crashes involving trucks on Sunday.	Any integer

APPENDIX B

Dataset: crashmc (VGAM)
Description: This dataset from the VGAM library in R contains the number of crashes by day of the week and time of the day involving motorcycles in New Zealand during 2009.

Sample Data:

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
0	1	0	0	0	5	1	1
1	0	1	0	1	1	3	1
2	0	0	1	2	1	3	1
3	2	0	0	0	0	0	1
4	0	0	0	0	0	0	0

Description of Variables:

Variable	Description	Possible Values
{Row Num}	Corresponds to the time of day the crash occurred (0 = 12:00 AM).	Integer from 0-23
Mon	The number of crashes involving motorcycles on Monday.	Any integer
Tue	The number of crashes involving motorcycles on Tuesday.	Any integer
Wed	The number of crashes involving motorcycles on Wednesday.	Any integer
Thu	The number of crashes involving motorcycles on Thursday.	Any integer
Fri	The number of crashes involving motorcycles on Friday.	Any integer
Sat	The number of crashes involving motorcycles on Saturday.	Any integer
Sun	The number of crashes involving motorcycles on Sunday.	Any integer

APPENDIX C

Dataset: crashbc (VGAM)

Description: This dataset from the VGAM library in R contains the number of crashes by day of the week and time of the day involving bicycles in New Zealand during 2009.

Sample Data:

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	2
3	0	0	0	0	0	0	1
4	0	0	0	1	0	0	0

Description of Variables:

Variable	Description	Possible Values
{Row Num}	Corresponds to the time of day the crash occurred (0 = 12:00 AM).	Integer from 0-23
Mon	The number of crashes involving bicycles on Monday.	Any integer
Tue	The number of crashes involving bicycles on Tuesday.	Any integer
Wed	The number of crashes involving bicycles on Wednesday.	Any integer
Thu	The number of crashes involving bicycles on Thursday.	Any integer
Fri	The number of crashes involving bicycles on Friday.	Any integer
Sat	The number of crashes involving bicycles on Saturday.	Any integer
Sun	The number of crashes involving bicycles on Sunday.	Any integer

APPENDIX D

Description: This dataset restructures the crashtr dataset from the VGAM library in R and contains the number of crashes by day of the week and time of the day involving trucks in New Zealand during 2009.

Sample Data:

Day	Time	Crashes
Fri	Early.Morning	10
Fri	Morning	47
Fri	Afternoon	64
Fri	Evening	13
Mon	Early.Morning	7

Description of Variables:

Day: Corresponds to the day of the week the crashes occurred.
Possible Values: Text

- Mon → Monday
- Tue → Tuesday
- Wed → Wednesday
- Thu → Thursday
- Fri → Friday
- Sat → Saturday
- Sun → Sunday

Time: Corresponds to the time of day the crashes occurred.
Possible Values: Text

- Early.Morning— Time between 12:00 AM and 4:59 AM
- Morning— Time between 5:00 AM and 11:59 AM
- Afternoon— Time between 12:00 PM and 5:59 PM
- Evening— Time between 6:00 PM and 11:59 PM

Crashes: The number of crashes that occurred.
Possible Values: Any integer greater than or equal to 0.

APPENDIX E

Description: This dataset restructures the crashmc dataset from the VGAM library in R and contains the number of crashes by day of the week and time of the day involving motorcycles in New Zealand during 2009.

Sample Data:

Day	Time	Crashes
Fri	Early.Morning	9
Fri	Morning	47
Fri	Afternoon	119
Fri	Evening	37
Mon	Early.Morning	4

Description of Variables:

Day: Corresponds to the day of the week the crashes occurred.
Possible Values: Text

- Mon → Monday
- Tue → Tuesday
- Wed → Wednesday
- Thu → Thursday
- Fri → Friday
- Sat → Saturday
- Sun → Sunday

Time: Corresponds to the time of day the crashes occurred.
Possible Values: Text

- Early.Morning— Time between 12:00 AM and 4:59 AM
- Morning— Time between 5:00 AM and 11:59 AM
- Afternoon— Time between 12:00 PM and 5:59 PM
- Evening— Time between 6:00 PM and 11:59 PM

Crashes: The number of crashes that occurred.
Possible Values: Any integer greater than or equal to 0.

APPENDIX F

Description: This dataset restructures the crashtr dataset from the VGAM library in R and contains the number of crashes by day of the week and time of the day involving bicycles in New Zealand during 2009.

Sample Data:

Day	Time	Crashes
Fri	Early.Morning	1
Fri	Morning	41
Fri	Afternoon	63
Fri	Evening	11
Mon	Early.Morning	0

Description of Variables:

Day: Corresponds to the day of the week the crashes occurred.
Possible Values: Text

- Mon → Monday
- Tue → Tuesday
- Wed → Wednesday
- Thu → Thursday
- Fri → Friday
- Sat → Saturday
- Sun → Sunday

Time: Corresponds to the time of day the crashes occurred.
Possible Values: Text

- Early.Morning— Time between 12:00 AM and 4:59 AM
- Morning— Time between 5:00 AM and 11:59 AM
- Afternoon— Time between 12:00 PM and 5:59 PM
- Evening— Time between 6:00 PM and 11:59 PM

Crashes: The number of crashes that occurred.
Possible Values: Any integer greater than or equal to 0.

Analysis of the Behavior of Selected Sample Statistics on the Youth Risk Behavior Surveillance System Population

Sam Oliszewski

Oregon State University

ABSTRACT

This paper discusses the analysis of some sample statistics using 2003 and 2013 data from a large survey of high-school students in America, called the Youth Risk Behavior Surveillance System (YRBSS). This analysis includes the inference on the population of high-school students based on the sample of YRBSS students. First, there is a determination as to whether high-school students have increased their BMI over time. Second, there is a determination as to whether male high-school students are more likely to smoke than female high-school students. Finally, there is an estimate as to how much TV the average high-schooler watches per day. Ultimately, the information gathered from the YRBSS data can help make useful inferences about the population of high-school students. The research defined in this paper provide the context for such inferences. Ultimately, it was observed that students have increased their BMI over time and that male students are more likely to smoke than females. It is also known that students on average are watching approximately two hours of TV per day. Understanding this data allows law-makers and educators to guide student behavior in a way that could improve student health. For future research, it would be useful to model the responses from the YRBSS data to quantify the risk each student has based on their responses. This would require further questioning about the overall health and well-being of the students to be done. Building a predictive model based on risk factors would help reveal the risk that certain behaviors have in the context of real health outcomes.

Keywords: Youth Risk Behavior Surveillance System (YRBSS), two-sample t-test, two-sample proportion test

ANALYSIS OF THE BEHAVIOR OF SELECTED SAMPLE STATISTICS ON THE YOUTH RISK BEHAVIOR SURVEILLANCE SYSTEM POPULATION

This paper provides an analysis of some sample statistics using 2003 and 2013 data from a large survey of high-school students in America, called the Youth Risk Behavior Surveillance System (YRBSS), and includes the inference on the population of high-school students based on the sample of YRBSS students. There are three explanatory questions of interest addressed in this analysis. First, there is a determination as to whether high-school students are observing an increase in their BMI over time. Second, there is a determination as to whether male high-school students are more likely to smoke than female high-school students. Finally, there is an estimate as to how much TV the average high-schooler watches per day. Ultimately, the information gathered from the YRBSS data can help make useful inferences about the population of high-school students. The research defined in this paper provide the context for such inferences. The paper will begin by discussing the exploratory analysis that was performed to provide insight on the data being studied. Next, the approaches for statistical test determination are described. Finally, conclusions drawn from the analysis are reported, and caveats of the analysis are offered to aid in the overall interpretation of the study results.

Exploratory Analysis

Original Variables. The raw data for this analysis came from two separate datasets containing the YRBSS responses for the years 2003 and 2013, respectively. Each dataset was of the same structure. A sample of the data, as well as descriptions of possible values for each variable, can be referenced in Appendices A and B.

Reducing the Number of Variables. For this analysis, not all of the variables in the original data are relevant to the questions of interest being addressed. Therefore, the data was reduced to only contain the variables required for the analysis. This included: year, BMI, sex, q33, and q81. A sample of the data, as well as descriptions of possible values for each variable, can be referenced in Appendices C and D.

Visualizing the Data

Figures for Each Variable by Year. For this analysis, it is beneficial to examine plots for each variable in the data by each year the survey was performed. These plots are shown below:

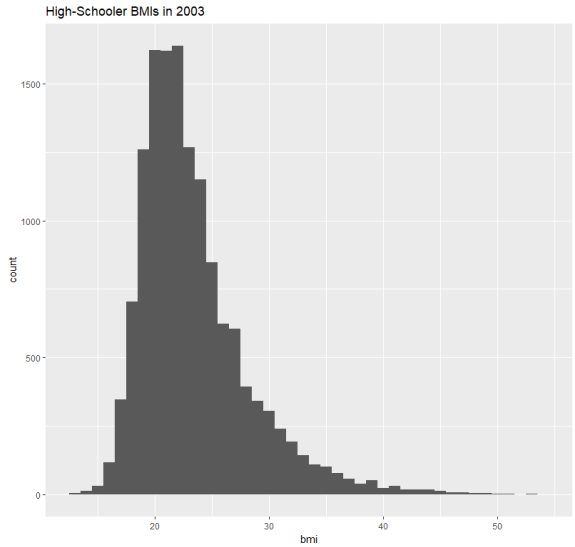


Figure 1. High School BMI in 2003

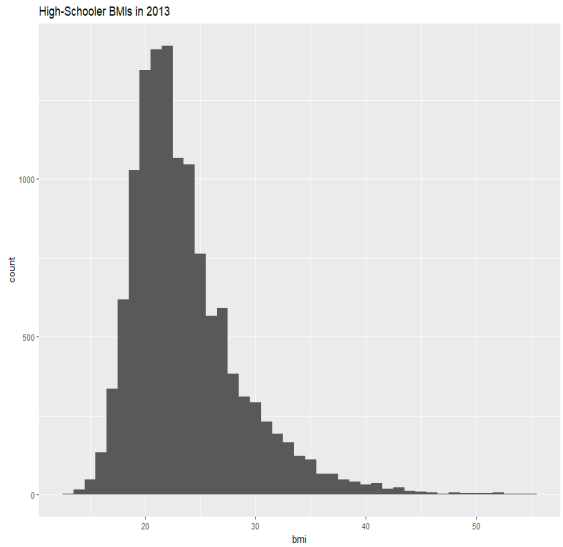


Figure 2. High School BMI in 2013

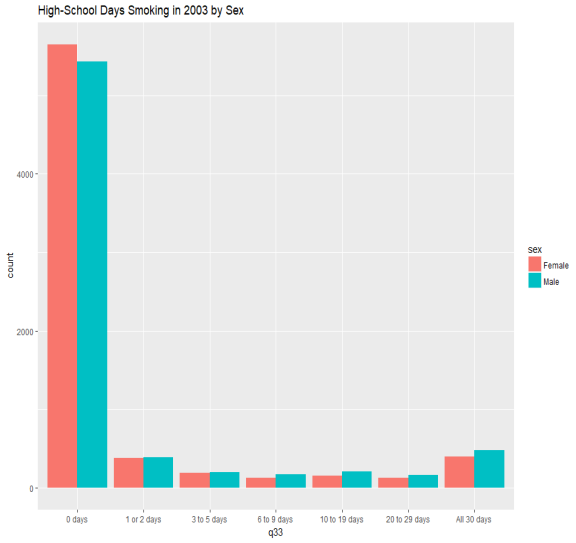


Figure 3. High School Days Smoking by Sex in 2003

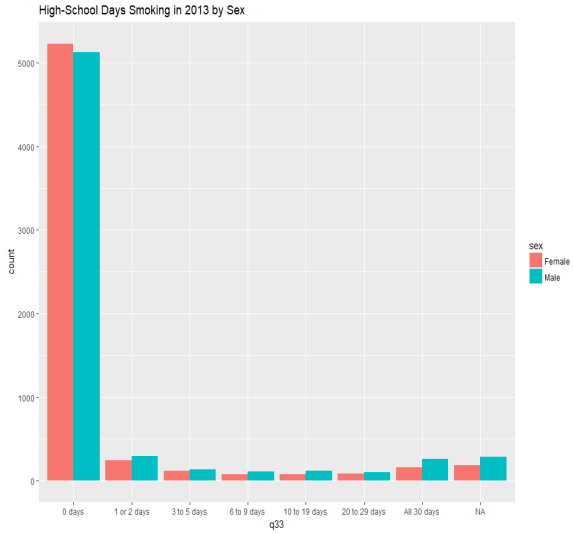


Figure 4. High School Days Smoking by Sex in 2013

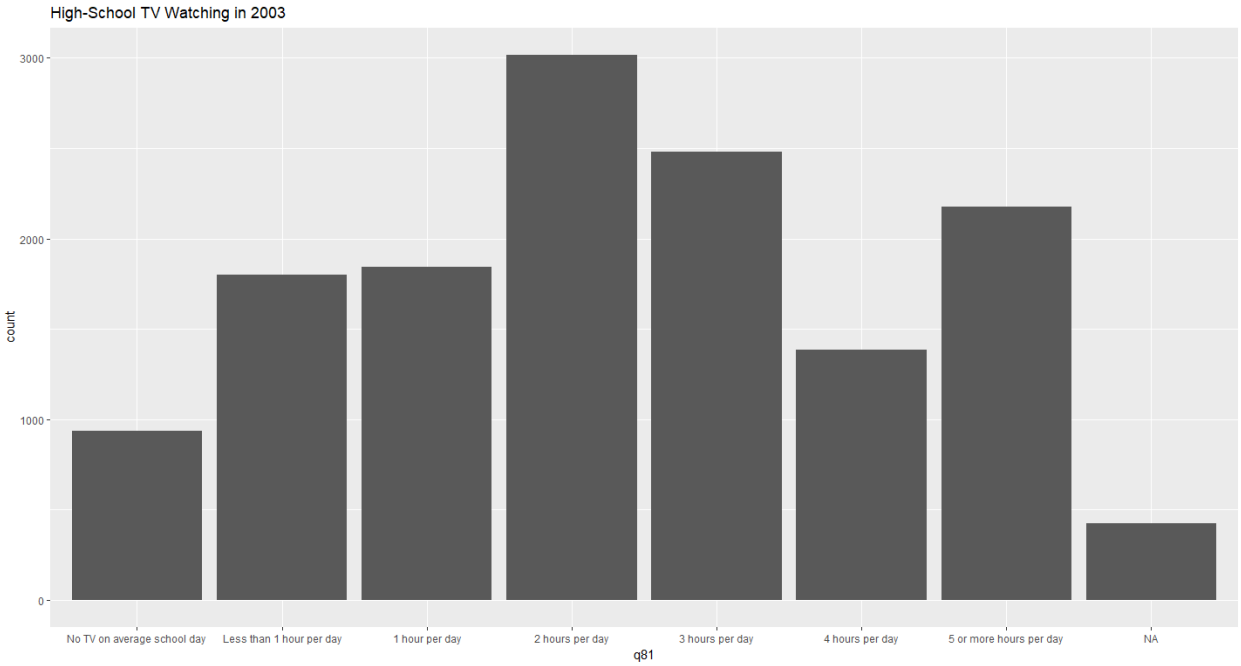


Figure 5. High School TV Watching in 2003

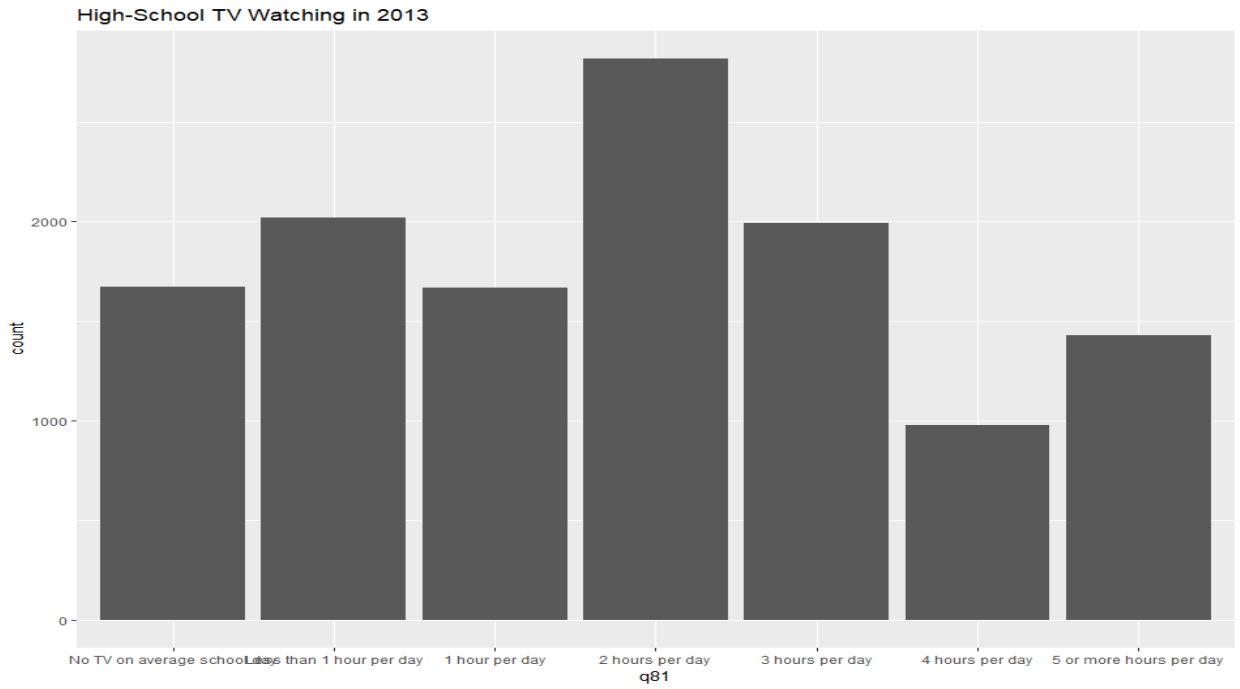


Figure 6. High School TV Watching in 2013

Based on these plots, there are small differences between 2003 and 2013 in terms of BMI; there appears to be less smoking overall between 2003 and 2013 and it is more common for male high schoolers than female; and the number of hours of TV watched per day appears to be slightly less on average, while the most common amount of time is 2 hours per day.

Handling Missing Values. Since there are a handful of missing values in the data, and the number of missing values is relatively small (<1%), median imputation will be performed on the data. This will provide a few more observations to be available in the data. Median imputation will be performed on the variables q33 and q81 as they are the only variables with observed missing values in the reduced data.

Explanatory Problems

How Has the BMI of High-Schoolers Changed Between 2003 and 2013?

Methods. To study the BMI of high-school students between 2003 and 2013, a Welch's two-sample t-test was run with a null hypothesis that the mean BMI of students in 2003 is equal to the mean BMI of students in 2013 and an alternative hypothesis that the mean BMI of students in 2013 is greater than the mean BMI of students in 2003. The variances are assumed to be unequal between the populations since there is no clear evidence that they should be assumed equal. The data from 2003 and 2013 are also assumed to be independent because they are not tracking the same students. A two-sample t-test is an appropriate choice for this study because the parameter in question is the mean. This test can be applied because the sample size is large validating the use of the approximately normal test statistic. Further, a t-distribution is a reasonable comparison to use, since the sample is approximately normal.

Results. The two-sample t-test yielded a p-value of 8.76e-05, suggesting that there is significant evidence to reject the null hypothesis that the mean BMI of students in 2003 is equal to the mean BMI of students in 2013. It is estimated that the mean BMI of students in 2013 is 23.64 and the mean BMI of students in 2003 is 23.41. With 95% confidence, the mean BMI of students in 2013 is at least 0.13 units larger

than the mean BMI of students in 2003. Additionally, the median BMIs in 2003 and 2013 are observed to be 22.29 and 22.49.

Figures. The following figure depicts the increase in high-value outliers in the 2013 BMI data that caused the increase in mean BMI from 2003 to 2013:

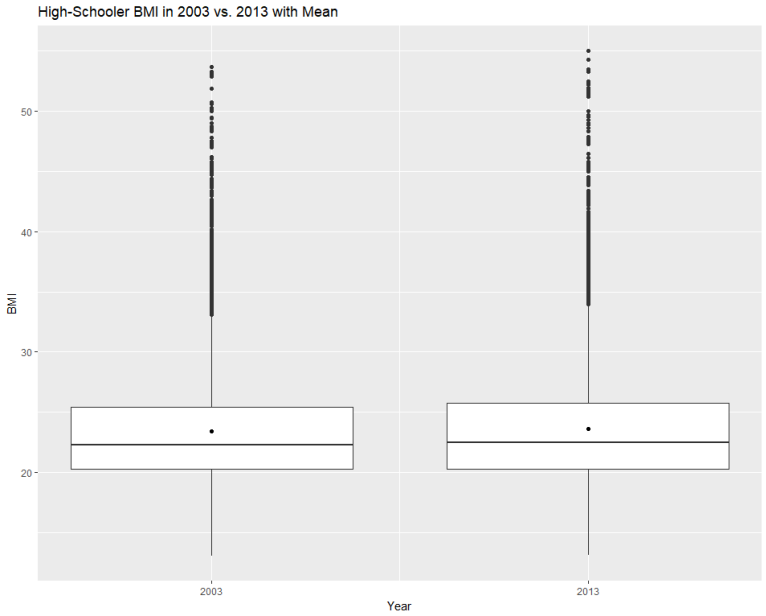


Figure 7. Boxplots of BMI Data From 2003 and 2013 with Mean Labeled.

Conclusion. The result of this test showed that high-school students in 2013 have higher BMIs than the students from 2003 (Welch's two sample t-test, t = 3.75, df = 25988, p-value = 8.76e-05). This suggests that the BMI of high-schoolers is increasing over time. Further, the median BMIs of 22.29 and 22.49 for 2003 and 2013, respectively, indicate that the BMI of high schoolers is increasing over time and that this result is not simply the result of influential outliers.

Are Male High-Schoolers More Likely to Smoke than Female High-Schoolers?

Methods. To study the likelihood that male high-schoolers are more likely to smoke than female high-schoolers, a two-sample proportion test was run with a null hypothesis that the proportion of male high-school smokers was equal to the proportion of female high-school smokers and an alternative hypothesis that the proportion of male high-school smokers is greater than the proportion of female high-school smokers. A two-sample proportion test is appropriate for this study because proportions are an appropriate tool for comparing binary data. In the case of determining whether students are smokers, the answer would be "yes" or "no" for the purpose of this question. Therefore, a proportion of "yes" responses would indicate how many students are smokers. Then, the comparison of proportions between the two sexes would indicate whether one is more likely to smoke than the other. The two-sample proportion test can be used because the following conditions are met: 1. The samples of female and male students are sufficiently large, and 2. The samples of female and male students are independent. To define a "yes" response for whether the student is a smoker, the student must have indicated for q33 that they have smoked greater than 0 days in the last 30 days, otherwise it will be assumed as a "no" response.

Results. The two-sample proportion test yielded a p-value of 1.02e-07, suggesting there is significant evidence to reject the null hypothesis that the proportion of male high-school smokers is equal to the proportion of female high-school smokers. It is estimated that the proportion of male high-school

smokers is 10.99% and the proportion of female high-school smokers is 8.25%. With 95% confidence, the proportion of male high-school smokers is between 2% and 100% larger than the proportion of female high-school smokers.

Figures. The following figure depicts that male high schoolers generally smoke more than females:

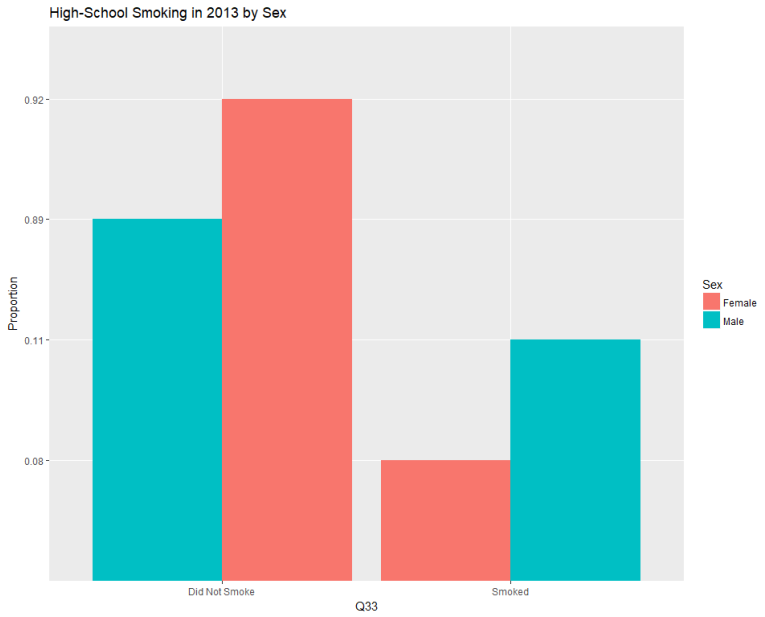


Figure 8. High School Smoking by Sex in 2013

Data. The following table displays the number of smokers versus total students and the calculated proportion of smokers by sex:

Sex	Number of Smokers	Number of Students	Proportion of Student Smokers
Male	705	6414	**10.99%**
Female	509	6166	8.25%

Table 1. High-School Student Smoker Proportions.

Summary. The result of this test showed that male high-schoolers are more likely to smoke than female high-schoolers (two sample proportion test, x-squared = 27.00, df =1, p-value = 1.02e-07).

How Much TV Do High-Schoolers Watch?

Methods. To study how much TV high-schoolers watch, the median TV time value reported by high-schoolers in 2013 was calculated. This value is assumed to reflect the population of high-schoolers in the present because it is the most recent data available. Since the responses on the YRBSS survey are categorical, the answer to this analysis will also be reported as categorical. The median is an appropriate measure of the average amount of time high-schoolers watch TV because the data recorded was in ordinal categories and therefore the median is an appropriate measure of the center of the data because other statistics, such as the mean, wouldn't apply to this type of data.

Results. The median value reported by students in 2013 was "2 hours per day".

Figures. The following figure depicts the responses by students in 2013 about how much TV they watch on average.

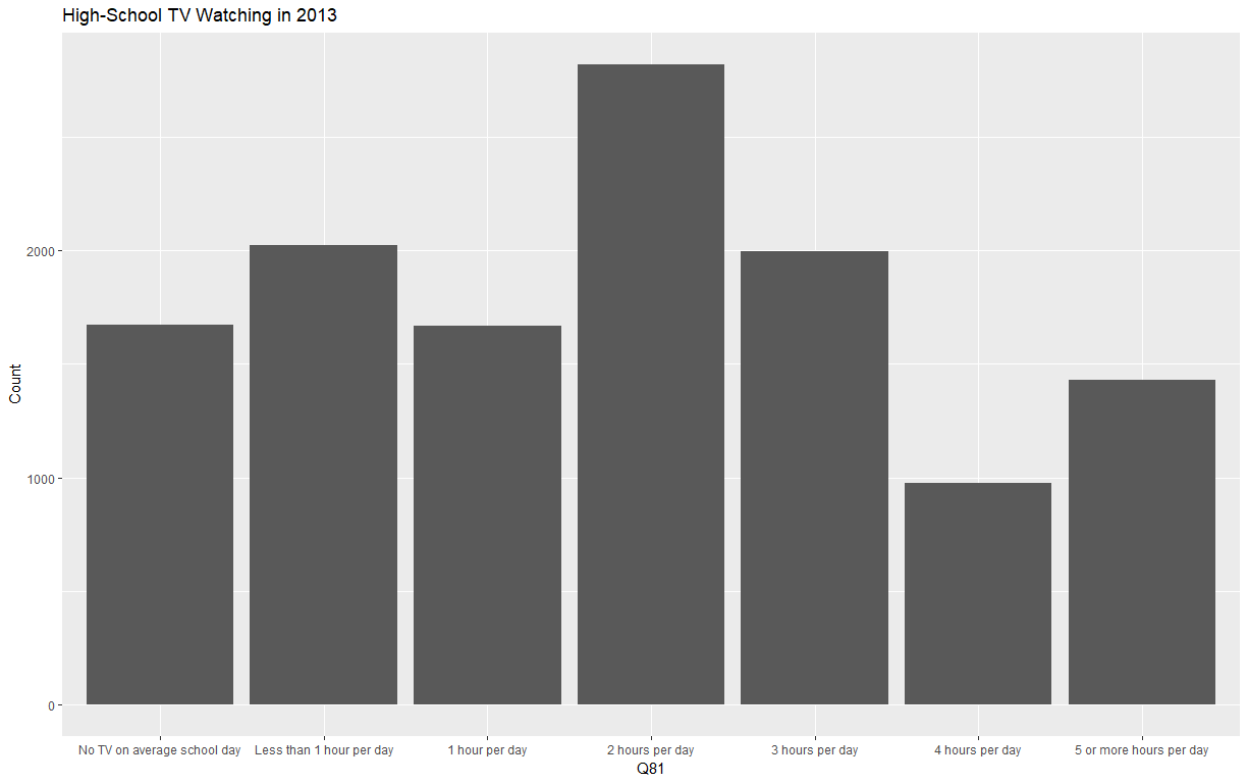


Figure 9. High School TV Watching in 2013

Data. This table displays the student responses to the amount of TV they watch per day and the frequency of each response:

TV Watching Time	Frequency
No TV on average school day	1,671
Less than 1 hour per day	2,021
1 hour per day	1,667
2 hours per day	**2,548**
3 hours per day	1,995
4 hours per day	977
5 or more hours per day	1,430

Table 2. Frequency of TV Watching Responses.

Summary. The amount of TV that high-schoolers watch is estimated to be two hours per day.

Conclusion

Based on the analysis of the YRBSS data, inferences about the population of high-school students can be drawn. It was observed that students have increased their BMI over time and that male students are more likely to smoke than females. It is also known that students on average are watching approximately two hours of TV per day. Understanding this data allows law-makers and educators to guide student behavior in a

way that could improve student health. For future research, it would be useful to model the responses from the YRBSS data to quantify the risk each student has based on their responses. This would require further questioning about the overall health and well-being of the students to be done. Building a predictive model based on risk factors would help reveal the risk that certain behaviors have in the context of real health outcomes.

APPENDIX A

Sample Data: YRBSS 2003

Year	BMI	Age	Sex	Grade	Race	Q9	Q33	Q77	Q80	Q81
2003	21.8	12 years old or younger	Male	11th	All other races	Rarely	0 days	NA	NA	5 or more hours per day
2003	21.5	12 years old or younger	Male	9th	All other races	Never	All 30 days	NA	NA	5 or more hours per day
2003	21.4	13 years old	Male	10th	Black or African American	Always	0 days	NA	NA	5 or more hours per day
2003	18.9	13 years old	Male	9th	Hispanic/Latino	Always	0 days	NA	NA	2 hours per day
2003	18.0	13 years old	Male	9th	White	Rarely	0 days	NA	NA	5 or more hours per day
2003	18.1	13 years old	Male	9th	White	Always	0 days	NA	NA	1 hour per day

Description of Variables:

Year: Year of the YRBSS survey response

Possible Values: Integer of the value 2003 or 2013

BMI: Student body mass index (BMI)

Possible Values: Number greater than 0

Age: Age of the student

Possible Values: Text

- 12 years old or younger
- 13 years old
- 14 years old
- 15 years old
- 16 years old
- 17 years old
- 18 years old or older

Sex: Sex of the student

Possible Values: Text

- Female
- Male

APPENDIX A CONTINUED

Grade: School grade of the student

Possible Values: Text

- 9th grade
- 10th grade
- 11th grade
- 12th grade
- Ungraded or other grade

Race: Race of the student

Possible Values: Text

- White
- Black or African American
- Hispanic/Latino
- All Other Races

Q9: Response to the question "How often do you wear a seat belt when riding in a car driven by someone else?"

Possible Values: Text

- Never
- Rarely
- Sometimes
- Most of the Time
- Always

Q33: Response to the question "During the past 30 days, on how many days did you smoke cigarettes?"

Possible Values: Text

- 0 days
- 1 or 2 days
- 3 to 5 days
- 6 to 9 days
- 10 to 19 days
- 20 to 29 days
- All 30 days

Q77: Response to the question "During the past 7 days, how many times did you drink a can, bottle, or glass of soda or pop, such as Coke, Pepsi, or Sprite? (Do not count diet soda or diet pop.)

Possible Values: Text

- I did not drink soda or pop during the past 7 days
- 1 to 3 times during the past 7 days
- 4 to 6 times during the past 7 days
- 1 time per day
- 2 times per day
- 3 times per day
- 4 or more times per day

APPENDIX A CONTINUED

Q80: Response to the question "During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day? (Add up all the time you spent in any kind of physical activity that increased your heart rate and made you breathe hard some of the time.)

Possible Values: Text

- 0 days
- 1 day
- 2 days
- 3 days
- 4 days
- 5 days
- 6 days
- 7 days
-

Q81: Response to the question "How many hours of TV do you watch on an average school day?"

Possible Values: Text

- I do not watch TV on an average school day
- Less than 1 hour per day
- 1 hour per day
- 2 hours per day
- 3 hours per day
- 4 hours per day
- 5 or more hours per day

APPENDIX B

Sample Data: YRBSS 2013

Year	BMI	Age	Sex	Grade	Race	Q9	Q33	Q77	Q80	Q81
2013	22.0	12 years old or younger	Male	9th	Black or African American	Never	NA	4 or more times per day	7 days	No TV on an average school day
2013	21.5	12 years old or younger	Male	10th	Black or African American	Sometimes	0 days	Did not drink soda or pop	7 days	3 hours per day
2013	19.0	12 years old or younger	Male	12th	All Other Races	Always	NA	4 or more times per day	2 days	5 or more hours per day
2013	21.9	12 years old or younger	Male	12th	Black or African American	Always	0 days	Did not drink soda or pop	7 days	2 hours per day
2013	17.6	13 years old	Male	9th	White	Always	0 days	1 to 3 times	7 days	No TV on an average school day
2013	28.9	13 years old	Male	9th	Black or African American	Always	0 days	1 to 3 times	6 days	3 hours per day

Description of Variables:

Year: Year of the YRBSS survey response

Possible Values: Integer of the value 2003 or 2013

BMI: Student body mass index (BMI)

Possible Values: Number greater than 0

Age: Age of the student

Possible Values: Text

- 12 years old or younger
- 13 years old
- 14 years old
- 15 years old
- 16 years old
- 17 years old
- 18 years old or older

APPENDIX B CONTINUED

Sex: Sex of the student

Possible Values: Text

- Female
- Male

Grade: School grade of the student

Possible Values: Text

- 9th grade
- 10th grade
- 11th grade
- 12th grade
- Ungraded or other grade

Race: Race of the student

Possible Values: Text

- White
- Black or African American
- Hispanic/Latino
- All Other Races

Q9: Response to the question "How often do you wear a seat belt when riding in a car driven by someone else?"

Possible Values: Text

- Never
- Rarely
- Sometimes
- Most of the Time
- Always

Q33: Response to the question "During the past 30 days, on how many days did you smoke cigarettes?"

Possible Values: Text

- 0 days
- 1 or 2 days
- 3 to 5 days
- 6 to 9 days
- 10 to 19 days
- 20 to 29 days
- All 30 days

APPENDIX B CONTINUED

Q77: Response to the question "During the past 7 days, how many times did you drink a can, bottle, or glass of soda or pop, such as Coke, Pepsi, or Sprite? (Do not count diet soda or diet pop.)

Possible Values: Text

- I did not drink soda or pop during the past 7 days
- 1 to 3 times during the past 7 days
- 4 to 6 times during the past 7 days
- 1 time per day
- 2 times per day
- 3 times per day
- 4 or more times per day

Q80: Response to the question "During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day? (Add up all the time you spent in any kind of physical activity that increased your heart rate and made you breathe hard some of the time.)

Possible Values: Text

- 0 days
- 1 day
- 2 days
- 3 days
- 4 days
- 5 days
- 6 days
- 7 days

Q81: Response to the question "How many hours of TV do you watch on an average school day?"

Possible Values: Text

- I do not watch TV on an average school day
- Less than 1 hour per day
- 1 hour per day
- 2 hours per day
- 3 hours per day
- 4 hours per day
- 5 or more hours per day

APPENDIX C

Sample Data: YRBSS 2003

Year	BMI	Sex	Q33	Q81
2003	21.8	Male	0 days	5 or more hours per day
2003	21.5	Male	All 30 days	5 or more hours per day
2003	21.4	Male	0 days	5 or more hours per day
2003	18.9	Male	0 days	2 hours per day
2003	18.0	Male	0 days	5 or more hours per day
2003	18.1	Male	0 days	1 hour per day

Description of Variables:

Year: Year of the YRBSS survey response

Possible Values: Integer of the value 2003 or 2013

BMI: Student body mass index (BMI)

Possible Values: Number greater than 0

Sex: Sex of the student

Possible Values: Text

- Female
- Male

Q33: Response to the question "During the past 30 days, on how many days did you smoke cigarettes?"

Possible Values: Text

- 0 days
- 1 or 2 days
- 3 to 5 days
- 6 to 9 days
- 10 to 19 days
- 20 to 29 days
- All 30 days

Q81: Response to the question "How many hours of TV do you watch on an average school day?"

Possible Values: Text

- I do not watch TV on an average school day
- Less than 1 hour per day
- 1 hour per day
- 2 hours per day
- 3 hours per day
- 4 hours per day
- 5 or more hours per day

APPENDIX D

Sample Data: YRBSS 2013

Year	BMI	Sex	Q33	Q81
2013	22.0	Male	NA	No TV on an average school day
2013	21.5	Male	0 days	3 hours per day
2013	19.0	Male	NA	5 or more hours per day
2013	21.9	Male	0 days	2 hours per day
2013	17.6	Male	0 days	No TV on an average school day
2013	28.9	Male	0 days	3 hours per day

Description of Variables:

Year: Year of the YRBSS survey response

Possible Values: Integer of the value 2003 or 2013

BMI: Student body mass index (BMI)

Possible Values: Number greater than 0

Sex: Sex of the student

Possible Values: Text

- Female
- Male

Q33: Response to the question "During the past 30 days, on how many days did you smoke cigarettes?"

Possible Values: Text

- 0 days
- 1 or 2 days
- 3 to 5 days
- 6 to 9 days
- 10 to 19 days
- 20 to 29 days
- All 30 days

Q81: Response to the question "How many hours of TV do you watch on an average school day?"

Possible Values: Text

- I do not watch TV on an average school day
- Less than 1 hour per day
- 1 hour per day
- 2 hours per day
- 3 hours per day
- 4 hours per day
- 5 or more hours per day

Analysis of Veteran's Administration Lung Cancer

Sam Oliszewski

Oregon State University

Introduction

In a study conducted by the US Veteran's Administration, male patients with advanced inoperable lung cancer were randomly assigned to two treatments of either a standard therapy or a test chemotherapy. Time to death was recorded for 137 patients, while 9 left the study before death. Various covariates were also documented for each patient including: tumor cell type, Karnofsky performance score, time between diagnosis and start of study (in months), age of patient, and an indicator of prior therapy.
This data set has been published in D. Kalbfleisch and R.L. Prentice (1980), The Statistical Analysis of Failure Time Data. Wiley, New York.

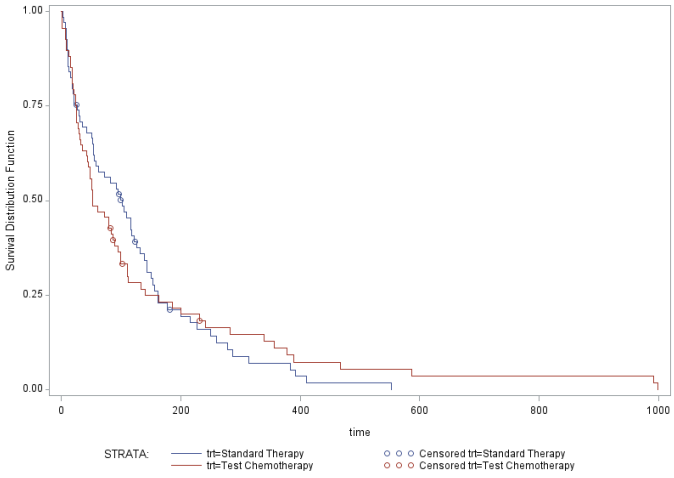
Methods and Results

The primary goal of the study was to assess if the test chemotherapy is beneficial. Secondary goals included the analysis of the additional covariates as prognostic variables.

The dataset of interest contains the following variables:

- trt: treatment type; 1=Standard Therapy, 2=Test Chemotherapy
- celltype: histological type of the tumor; 1=Squamous, 2=Smallcell, 3=Adeno, 4=Large
- time: survival time
- status: censoring status; 0=start of the observation period or censoring, 1= death
- karno: Karnofsky performance score that describes the overall patients' status at the beginning of the study (discretely scored)
- diagtime: time between diagnosis and start of the study (in months)
- age: age of the patient (in years)
- prior: indicates if the patient has received another therapy before the current one; 0=No Prior Therapy, 10=Prior Therapy

In a preliminary effort to determine if there appears to be a treatment effect on time to death for the V.A. lung cancer patients, a plot of the Kaplan Meier survival curves for each treatment group was generated. This plot is shown below:

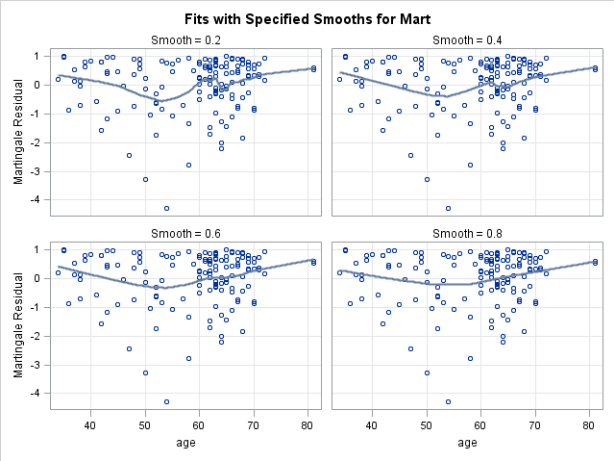


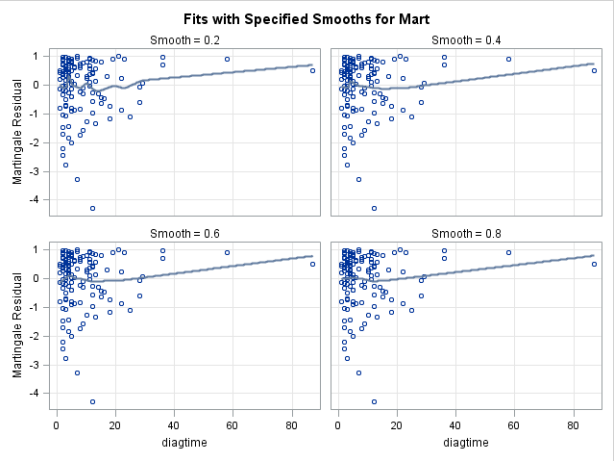
Based on these curves, there is not an obvious difference in the survival functions for the two treatment groups. Further analysis can confirm this intuition.

To determine the effect of treatment on time to death for the V.A. lung cancer patients, a Cox Proportional Hazard model was fit to the data. To verify that the Cox Proportional Hazard model is appropriate for this data, assessment of residual plots was required. Specifically, evaluation of the Martingale and Schoenfeld

residual plots was performed. These residual plots would indicate whether the covariates in the data had linear contributions to the model and whether proportional hazards were observed in the data. These are two important assumptions to satisfy in order to consider inference from the Cox Proportional Hazard model reliable.

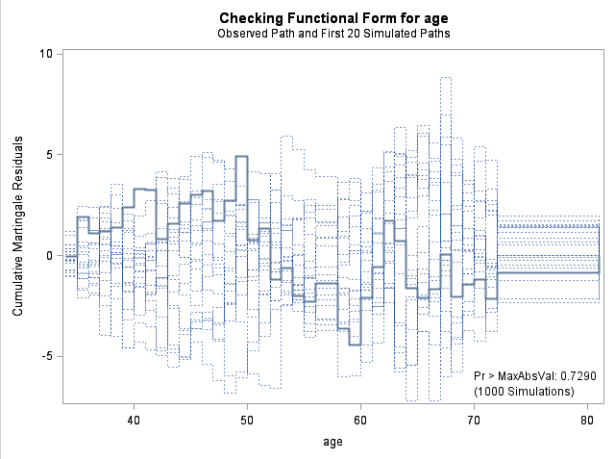
To begin, the Martingale residuals for each continuous covariate in the dataset were assessed. This included plots for the variables *age* and *diagtime*. These plots are shown below:

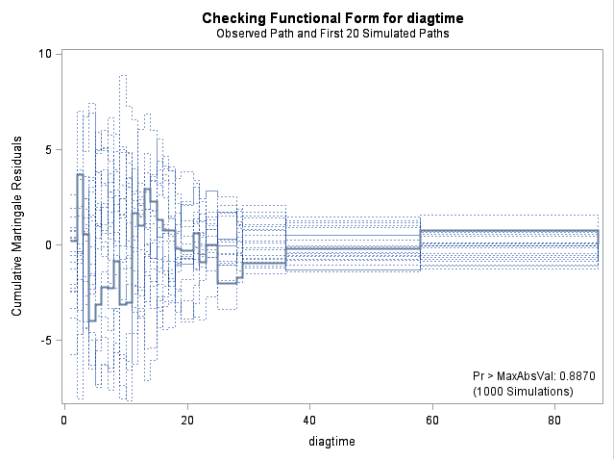




To verify the linear functional form of these covariates, we aim to see a flat line represented on the figure. We observe that as the smoothness increases for the fitted line, the curve flattens for both plots; however, neither is definitively flat nor strikingly curvy. Therefore, an additional test for the functional form of these covariates can help verify the linearity and satisfy this model assumption.

One assessment of the functional form of the covariates involves an evaluation of the cumulative Martingale residuals against each continuous covariate. These plots are shown below:



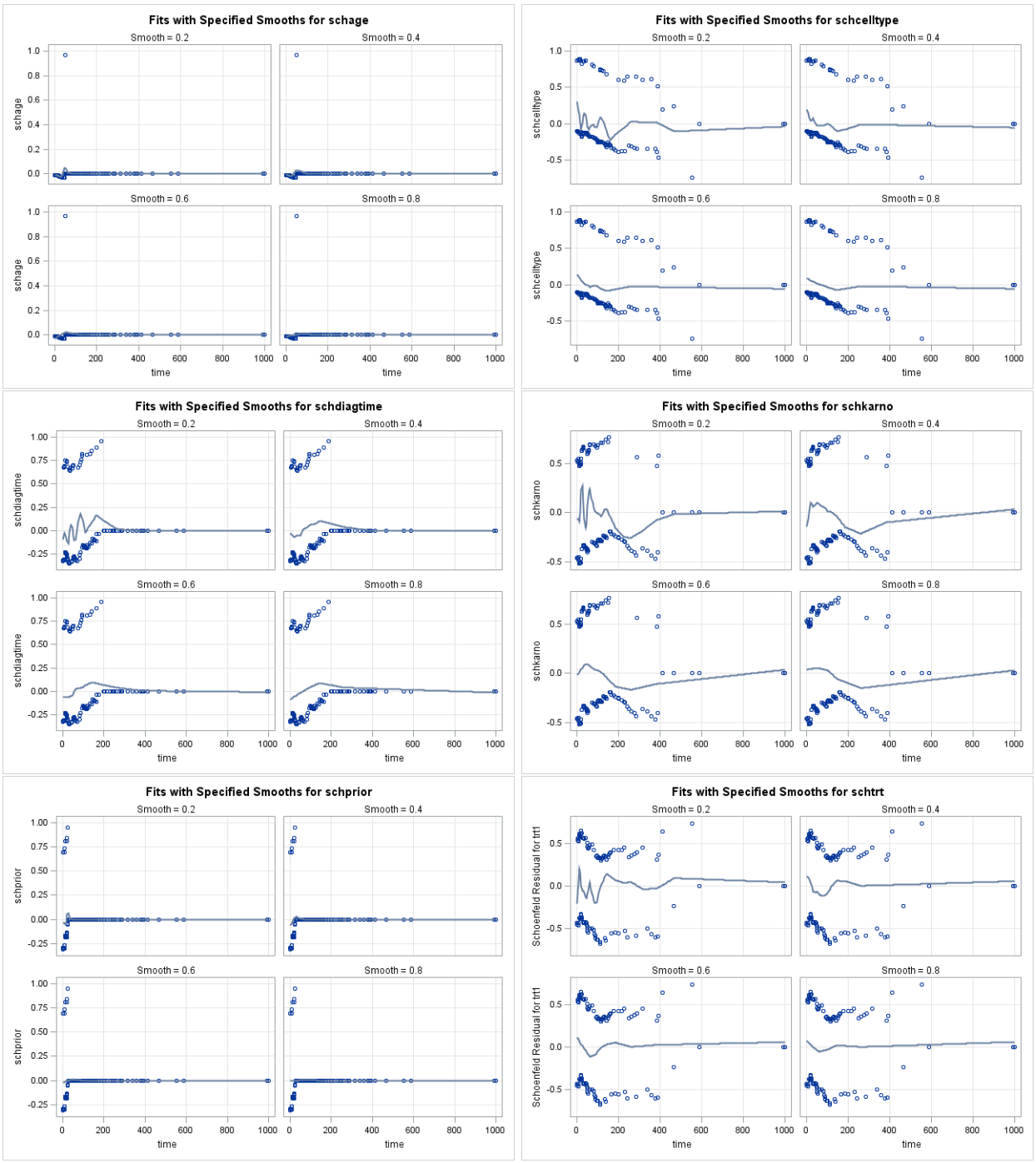


Confirmation of a linear functional form is depicted by the bolded line falling within the range of the dotted lines in the figure. Since both bolded lines in the two plots satisfy this, we have visual confirmation of a linear functional form of these continuous covariates.

A Kolmogorov-type supremum test for the functional form of each covariate can provide a quantitative measure for determining if there is evidence of a non-linear functional form. Therefore, computing such a test, in combination with information provided from the above plots, will allow for confirmation of the functional form of the continuous covariates in this dataset. The resulting p-values from the Kolmogorov-

type supremum test for the covariates *age* and *diagtime* were 0.73 and 0.89, respectively. These high p-values (p-value > 0.05) indicate that there is no evidence that these covariates do not have a linear functional form. Therefore, this assumption of the Cox Proportional Hazard model is satisfied.

The other important assumption for the Cox Proportional Hazard model is that there are proportional hazards in the data. To check this assumption, the plots for the Schoenfeld residuals must be considered. These plots are shown below:



This assumption is satisfied if the Schoenfeld residual plots for each covariate depict a flat smooth fitted line. The fitted lines for *age* and *prior* are very clearly flat lines and the fitted lines for *celltype, diagtime,* and *trt* are

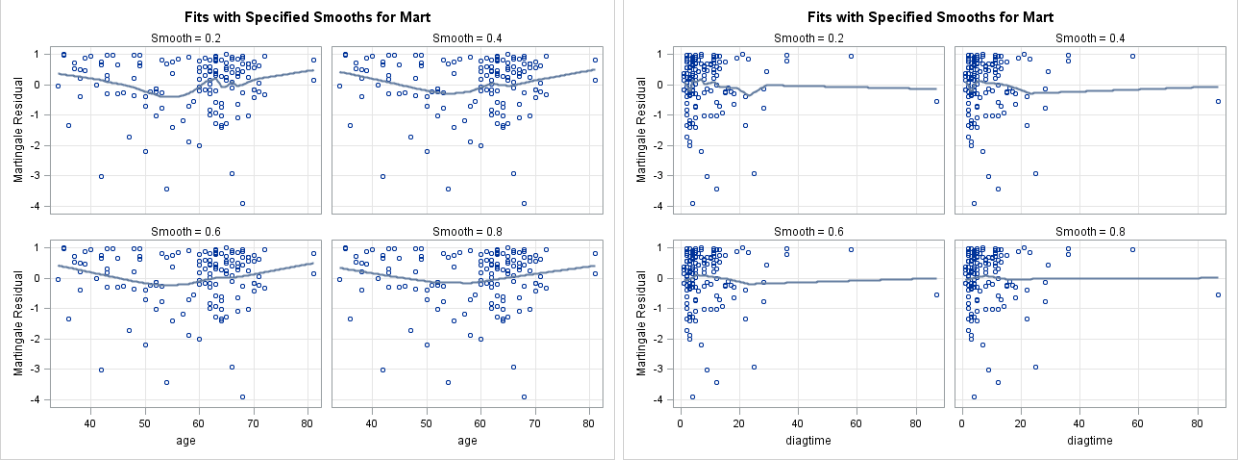
mostly very flat and thus give no indication that the proportional hazard assumption is violated. However, the fitted line for *karno* shows some evidence of curving and thus suggests that the proportional hazard assumption may be violated for this covariate.

Since the Schoenfeld residual plot for *karno* is inconclusive about violation of the proportional hazard assumption additional testing can be useful to determine whether the model assumption is satisfied. A Kolmogorov-type supremum test can be used to determine whether the proportional hazard assumption has been violated. Since the Karnofsky score variable is treated as categorical in the model, p-value are calculated for each increment of the metric. The associated p-values for each score are fairly large (p-value > 0.5), except for the scores of 40 and 80, which resulted in p-values of 0.02 and 0.00, respectively. The high p-values indicate there is no violation of the proportional hazard assumption, but since some of the Karnofsky scores indicated a violation of the proportional hazard assumption, either this covariate needs to be excluded from the model or the Cox Proportional Hazard model is inappropriate for the data. This analysis will therefore proceed with the exclusion of the Karnofsky score in the Cox Proportional Hazard model.

After excluding the *karno* covariate from the Cox Proportional Hazard model, we can run the Kolmogorov-type supremum test again to verify the model assumption again. This test indicated, with p-values for each covariate all larger than 0.05, that the proportional hazard assumption has been satisfied with this model.

With the exclusion of the *karno* covariate, it is wise to recheck the Martingale residuals to make sure the model is still appropriate.

The Martingale residuals for the model excluding *karno* are shown below:



We note that there are no significant changes to the Martingale residual plots.

Now that the Cox Proportional Hazard model has been justified to use for this data, a univariate analysis of the covariates in the model is useful to understand the influence of each covariate on the response. The model fit yielded the following regression coefficient output:

Parameter	DF	Estimate	Standard Error	Chi-Square	p-value	Hazard Ratio
trt=Standard	1	-0.16	0.20	0.66	0.41	0.85
celltype=Squamous	1	-0.29	0.29	1.03	0.31	0.75
celltype=Smallcell	1	0.75	0.26	8.20	0.00	2.11
celltype=Adeno	1	0.89	0.30	9.08	0.00	2.44
diagtime	1	0.01	0.01	1.16	0.28	1.01
age	1	0.00	0.01	0.26	0.61	1.01
prior=No Prior	1	0.07	0.23	0.08	0.77	1.07

The chi-square values reported above reflect the Wald test statistic and the p-value results from comparing the Wald statistic to a chi-square distribution with one degree of freedom (and thus reflect the results from a Wald test).

The full model thus takes the form:

$$H(t) = h_0(t)\exp(-0.16trtStandard - 0.29celltypeSquamous + 0.75celltypeSmallcell + 0.89celltypeAdeno + 0.01diagtime + 0.00age + 0.07priorNoPrior)$$

This output can be interpreted as follows:

1. The risk of death for patients given the standard lung cancer therapy is 0.85 times the risk of death for patients given the test chemotherapy.
2. The risk of death for patients with squamous tumor cells is 0.75 times the risk of death for the reference group (patients with large tumor cells).
3. The risk of death for patients with smallcell tumor cells is 2.11 times the risk of death for the reference group (patients with large tumor cells).
4. The risk of death for patients with adeno tumor cells is 2.44 times the risk of death for the reference group (patients with large tumor cells).
5. A one-unit increase in the value of *diagtime* results an increase in the risk of death by 1% (1.01-1.01=1%).
6. A one-unit increase in the value of *age* results in an increase in the risk of death by 1% (1.01-1.00=1%).
7. The risk of death for patients with no prior therapy is 1.07 times the risk of death for patients with prior therapy.

This output indicates that only one term is clearly significant in the model— *celltype*. Specifically, smallcell and adeno tumor cell types are significant in the model.

We can perform backwards selection to then create a reduced model with only relevant covariates. Each iteration, we will remove the covariate with the highest p-value, which is greater than 0.15, until all covariates in the model are significant (and including *trt*). This process resulting in selecting to include the covariates *trt* and *celltype* in the reduced model, as these were the only significant terms.

We determined that the log likelihood of the full model is 984.975. Fitting a reduced model with only the significant covariates *celltype* and *trt* (*trt* is kept due to treatment distinguishing), we find the log likelihood to be 986.217. We can then perform a likelihood ratio test to determine whether the full or reduced model is preferred. The likelihood ratio test statistic is 1.24 and when this statistic is compared to a chi-square distribution with 3 degrees of freedom (p-k → 7-4=3), we obtain a p-value of 0.74. This indicates that there is insufficient evidence that the coefficients of the extra covariates existing only in the full model are not equal to 0. Therefore, the reduced model is preferred.

Fitting the reduced model yields the following regression output:

Parameter	***DF***	***Estimate***	***Standard Error***	***Chi-Square***	***p-value***	***Hazard Ratio***
trt=Standard	1	-0.19	0.20	0.98	0.32	0.82
celltype=Squamous	1	-0.30	0.29	1.07	0.30	0.74
celltype=Smallcell	1	0.79	0.25	9.72	0.00	2.21
celltype=Adeno	1	0.87	0.29	8.83	0.00	2.38

The reduced model takes the form:

$$H(t) = h_0(t)\exp(-0.19trtStandard - 0.30celltypeSquamous + 0.79celltypeSmallcell + 0.87celltypeAdeno)$$

The following inferences can be made from the reduced ("best") model:

1. The risk of death for patients given the standard lung cancer therapy is 0.82 times the risk of death for patients given the test chemotherapy.
2. The risk of death for patients with squamous tumor cells is 0.74 times the risk of death for the reference group (patients with large tumor cells).
3. The risk of death for patients with smallcell tumor cells is 2.21 times the risk of death for the reference group (patients with large tumor cells).
4. The risk of death for patients with adeno tumor cells is 2.38 times the risk of death for the reference group (patients with large tumor cells).

Discussion

This analysis aimed to determine the treatment effect on time to death for V.A. lung cancer patients, as well as to determine if any other covariates could be prognostic tools. Treatment type was not determined to be a significant covariate in the model of time to death for this data, but it was determined that the risk of death is slightly lower for patients given the standard therapy. Further, it was determined that tumor cell type could be used to determine the risk of death for lung cancer patients and that squamous tumor cells were attributed to lower risk of death, versus the higher risk associated with the presence of smallcell and adeno tumor cells.

The inferences drawn from this analysis have some caveats. The first is that Karnofsky score is not considered in this model and could be relevant to the risk of death. Further analysis the reintroduces this covariate is recommended to understand its influence on the time to death. Since the primary goal of this study was to assess the treatment effect on time to death, the exclusion of this covariate from the model should not impact the conclusion drawn about treatment effect, however.

This analysis involved through examination of model assumptions to ensure the validity of the inferences made. The scope of the study was refined to only examine variables that could accurately be interpreted using the Cox Proportional Hazard model. Overall, the goals for this study were achieved using the methods discussed in this paper.