BRITTANY WHEATON

Brittwheaton13@gmail.com

Data Analysis Portfolio

Contents

Professional Resume 2

Time Series Analysis: U.S. Live Birth Data 3

Introduction 4

Methods 5

Trend 5

Seasonality 7

Residuals 8

Results 10

Model Fitting 10

Model Diagnosis 10

Predictions 12

Discussion 15

Summary 16

Appendix 16

Works cited 20

Software 21

Blog Post: Planning our Future Using Birth Data 22

Veterans Administration Lung Cancer Trial Survival Analysis 25

Introduction 25

Data 25

Methods and Results 26

Analysis of Car and Motorcycle Crashes in New Zealand in 2009 58

Introduction 58

Data 58

Analysis and Results 59

Conclusions/Discussion 64

Appendix 64

Professional Resume

Brittany Wheaton

Brittwheaton@gmail.com ● 210.857.5913 ● San Antonio, TX

SUMMARY

Data analyst with strong communication acumen skilled in applying quantitative techniques and translating results into clear, actionable insights for a variety of audiences.

EXPERIENCE

Teacher | 2014-2018
Northeast Independent School District

- Designed and implemented data- and technology-driven lesson plans as 6th grade reading team lead
- Documented student and teacher progress and taught strategies for student self-organization
- Participated in discussions with teams to collaborate toward teaching objectives

Girls' Athletic Coach | 2014-2016
Northeast Independent School District

- Built and enacted practice and game plans for volleyball, basketball, and track teams
- Leveraged statistical knowledge to inform future coaching strategy
- Facilitated relationship-building and encouraged collaboration among team members

Senior Writer | 2010-2017
City of Schertz, Public Affairs

- Developed communication material such as new and feature articles for monthly publication
- Gathered, organized, and assessed data and translated results to leaders during 2010 bond project
- Conducted interviews and transcribed information to transform into deliverable content

EDUCATION

Master of Science, Data Analytics | 2018-present (expected June 2020)
Oregon State University
GPA: 4.0

Relevant Coursework:

- Time Series Analytics
- Multivariate Analytics
- Applied Machine Learning
- Data Visualization
- Introduction to Econometrics
- Applied Survival Analysis

Master of Arts, Teaching | 2013-2014
Trinity University
Bachelor of Arts, Communication | 2008-2012
Trinity University

SKILLS

- Programming in R, Python, SAS, SQL, and STATA
- Computing with Google Cloud Platform
- Data wrangling with Apache Spark
- Data management and analysis in Excel
- Data Visualization in R
- Time series modeling and forecasting
- Multivariate techniques including principle component analysis, factor analysis, MANOVA, multivariate regression
- Technical/report writing
- Media writing
- Copy editing

Time Series Analysis: U.S. Live Birth Data

A Summary of Analysis of Live Births in the United States from 1978-2006

Introduction: Demography, or the study of population, plays a major role in the planning of a society. Being able to predict birth rates and accurately estimate the size of future populations is necessary for crucial tasks such as community design and resource allotment. Understanding trends is essential for ensuring long-term needs are met; for example, enough schools are built over time to effectively educate a young population. Furthermore, assessing seasonality can be critical in allocating appropriate amounts of resources throughout the year such as medical staff and space during peak birth seasons. This purpose of this report is two-fold: 1) to understand key trends and seasonality in the birth data and 2) to use that knowledge to predict future births.

Data Description: The dataset contains the monthly number of live births for the United States of America from 1969 to 2015. The data is available from the Demographics Statistics Database from the United Nations Statistics Division. The data have been collected since 1948 through a set of questionnaires dispatched annually to over 230 national statistical offices and have been published in the *Demgraphic Yearbook* collection, which contains statistics on population size and composition, births, deaths, marriage, and divorce as well as other topics. The dataset used for analysis includes U.S. data from 1978-2006. Prior years were removed from the analysis due to inconsistent or missing data. Years beyond 2006 were removed as a hold-out set for testing the predictive models.

Statistical Modeling: Two types of predictive models were used and compared in this report. First, an autoregressive integrated moving average, or ARIMA, model was fit to the time series. In the process of this model-fitting, the trend and seasonality were analyzed using smoothing methods, and the series was made stationary for fitting through differencing. The resulting model was a seasonal ARIMA model, or SARIMA(5,1,5)x$(1,2,1)_{12}$. Holt-Winter's forecasting was used as an alternative method for prediction. Based on a non-varying seasonality, an additive form was used.

Results: The time series was found to have an overall increasing trend with some cyclical variation in the form of a drop near the middle of each decade (1985, 1995, etc.). A clear seasonality emerged with births increasing steadily throughout the year peaking in August and declining in the fall and winter months. Predictions were performed for one additional cycle (year 2007) using the ARIMA and additive Holt-Winter's forecasting methods. These test group forecasts were then compared to the actual U.N. data for births in 2007, and the resulting error was recorded for each group. Error was generally lower for the Holt-Winter's method, though this was not consistent for each month. Overall, the actual number of births in 2007 was 4,306,242. Holt-Winter's predicted this value to be 4,364,731 (a difference of 58,489 births), while the chosen ARIMA model estimated the number of 2007 births to be 4,418,137 (an error of 111,895 – nearly double that of Holt-Winter's)

Conclusion: Both the ARIMA models and Holt-Winter's forecasting methods show promise for predicting future numbers of births in the United States. The relatively consistent trend and seasonality in the given time series suggest that these methods can be relatively useful planning purposes; however, it should be noted that the further forecasts are made into the future, estimates become far more uncertain.

Analysis of Live Births in the United States from 1978-2006

Brittany Wheaton

Introduction

Using live births by month of birth data downloaded from the “UN Data: A World of Information” for the United States of America, the monthly number of live births from 1978 to the end of 2006 is plotted in the time plot below. (http://data.un.org/Data.aspx?d=POP&f=tableCode%3A55)

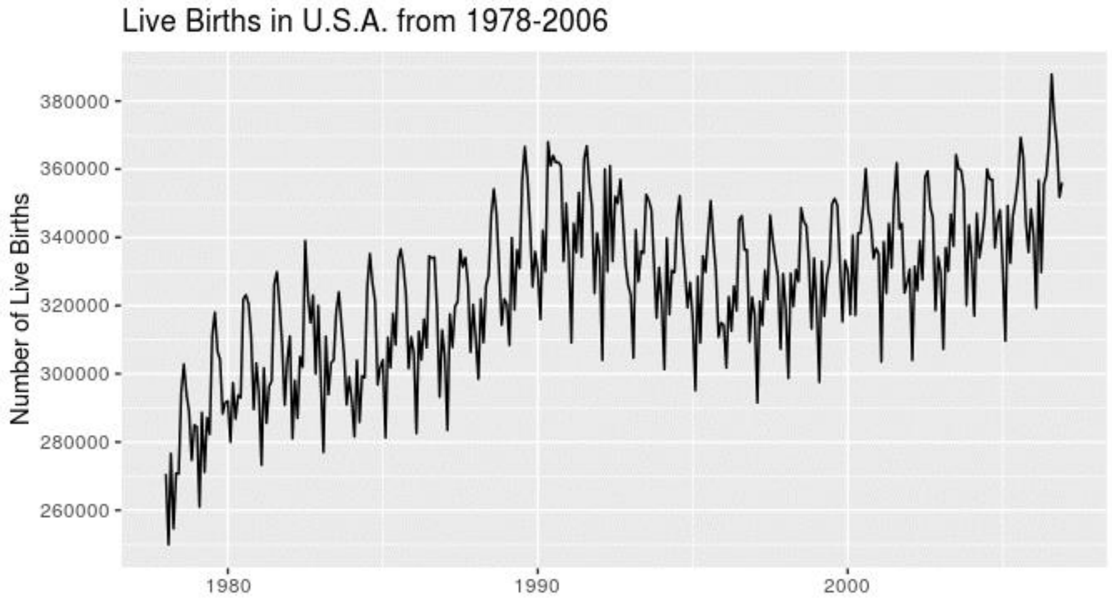


Figure 1. Number of live births in the United States

Darrow $et al.$, found that seasonal patterns differed among racial and ethnic groups, maternal education levels, and marital status (2009). A large national population laced with cultural, biological and environmental factors contribute to a complex seasonal variation. A goal of this analysis is to understand the trend and seasonal variation in the years 1978-2006 in order to develop an optimal ARIMA model for forecasting, which will be tested against a hold-out validation set of data from 2007. This model will compared with Holt-Winter’s forecasts to assess its performance.

Methods

To get a preliminary assessment of the primary components of the time series, the data was decomposed into the following elements: observed series, trend, seasonal, and random components. The trend component indicates a general incline with some possible cyclic variation. A clear seasonality emerges as well, though it will be necessary to examine this more closely in order to better understand when the high and low points occur throughout the year.

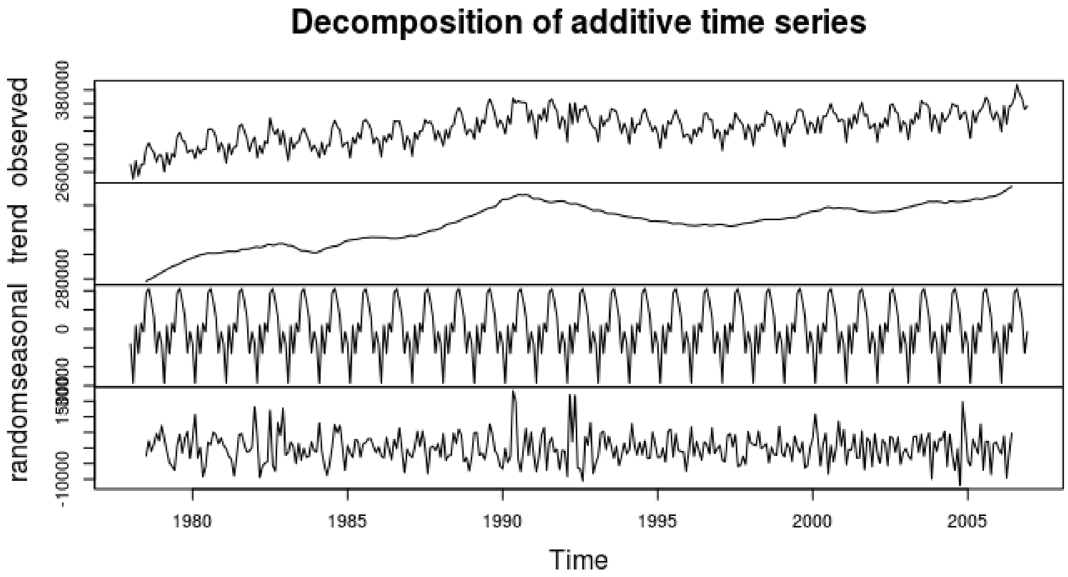


Figure 2. Decomposition of Births time series

Trend

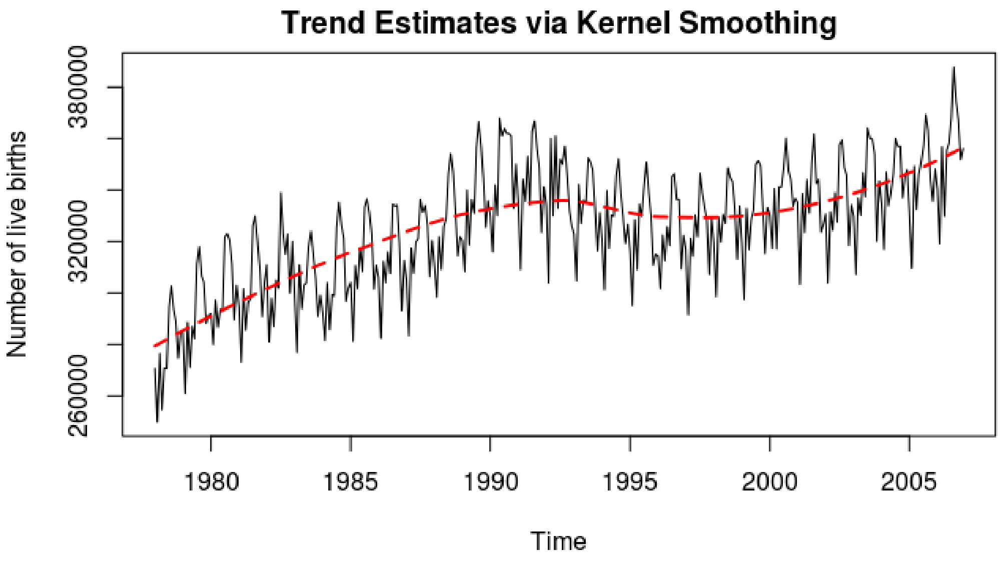


Figure 3. Trend line estimated via kernel smoothing method.

From the mid 1970's to the mid 1990's, there is an increasing trend in the number of babies born each month, though there appears to be some cyclical variation over the decades with small dips occurring about once every 10-15 years, most prominently in the mid-1980s and mid-1990s. The dashed red line above employs the kernel smoothing method to estimate the trend line.

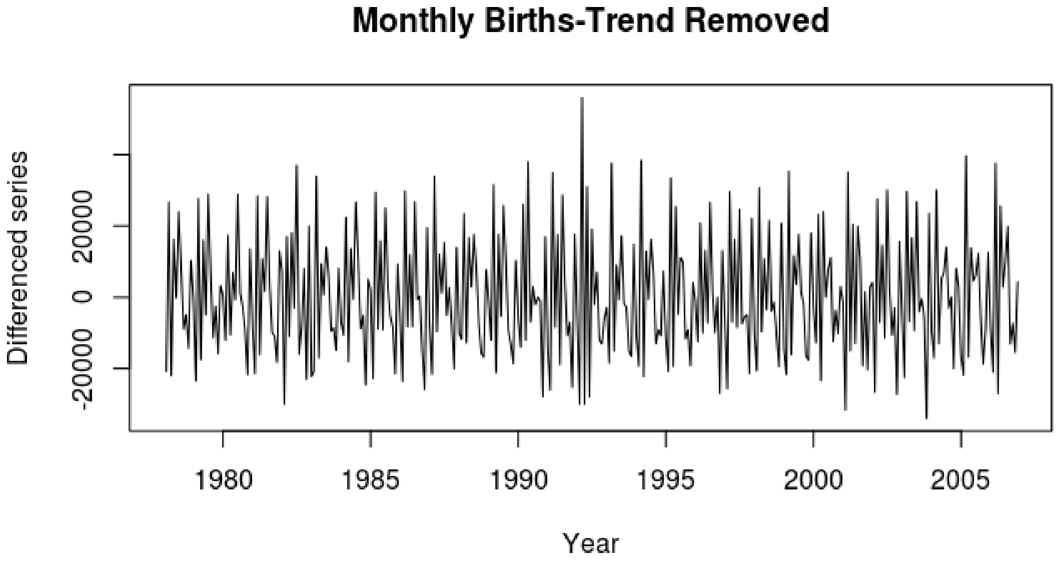


Figure 4. Time series after trend removal through first-order differencing.

In order to fit an ARIMA model for prediction, the time series must be differenced to achieve stationarity. A first-order difference was applied to remove the trend seen above. The remaining series still shows evidence of seasonality, which will be addressed in the next section.

Seasonality

From an overall view of the time series, a fairly regular seasonality emerges over the span of 28 years with the peak number of births occuring in August, declining into the fall and winter months. The low point in the year occurs at the beginning in January/February. This seasonality can be seen in the plot below featuring a smoothed seasonal trend across all of the years in blue. The magnitude of incline from April to August is comparable to the slope of the decline from August to December. Birth levels remain relatively flat from January to March.

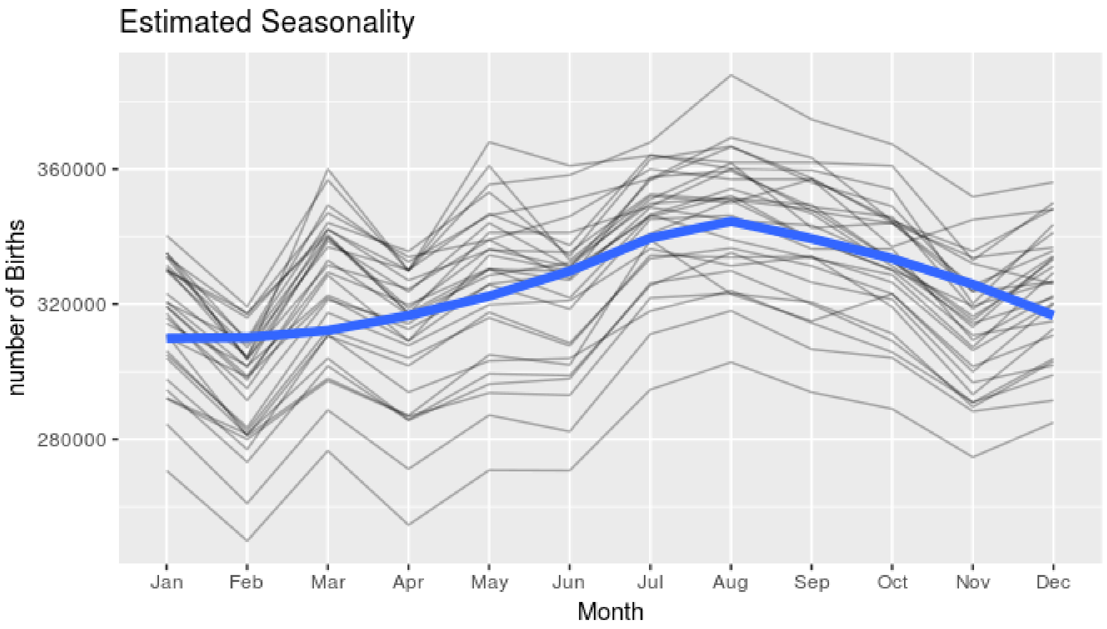


Figure 5. Estimated seasonality based on smoothing of all observed years.

Seasonality appears to be a dominant influence in the time series; in order to account for this and to finish rendering the data stationary, the time series can be differenced with a lag of 12 to make the time series stationary.

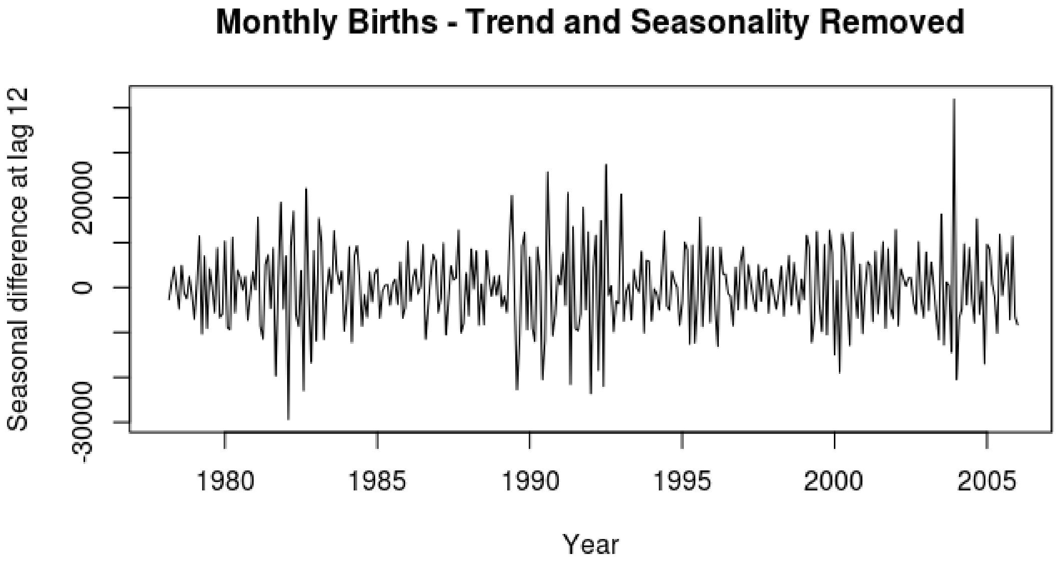


Figure 6. Time series after first-order and seasonal differencing.

After taking a first difference and applying a seasonal difference, the plot below resembles the targeted white noise; however, there remains a large spike around 2004. Additionally, some segments of time have much greater variation than others, which may have been due to some underlying cyclic variation. Examining residuals and the autocorrelation functions will determine whether or not the series is ready for the model-fitting process.

Residuals

After the seasonality and the trend has been removed from the time series, some autocorrelation remains. The ACF still has pronounced seasonal lags and the PACF slowly decreases to zero and also still features seasonal lags. Because of the remaining seasonality, a second seasonal difference was applied to help improve the stationarity of the series before fitting a model.

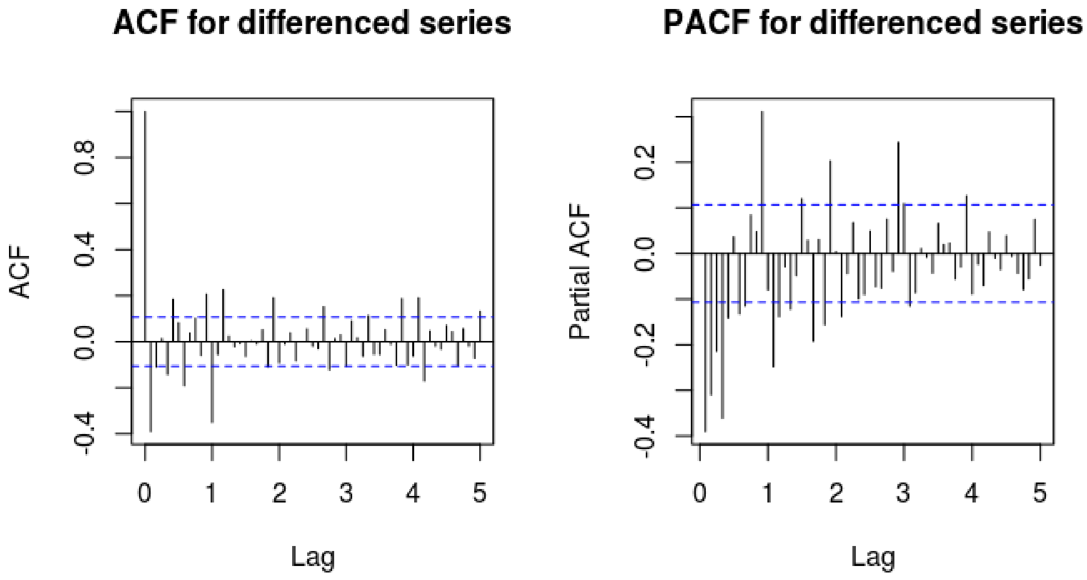


Figure 7. Autocorrelation and Partial Autocorrelation Functions for differenced series

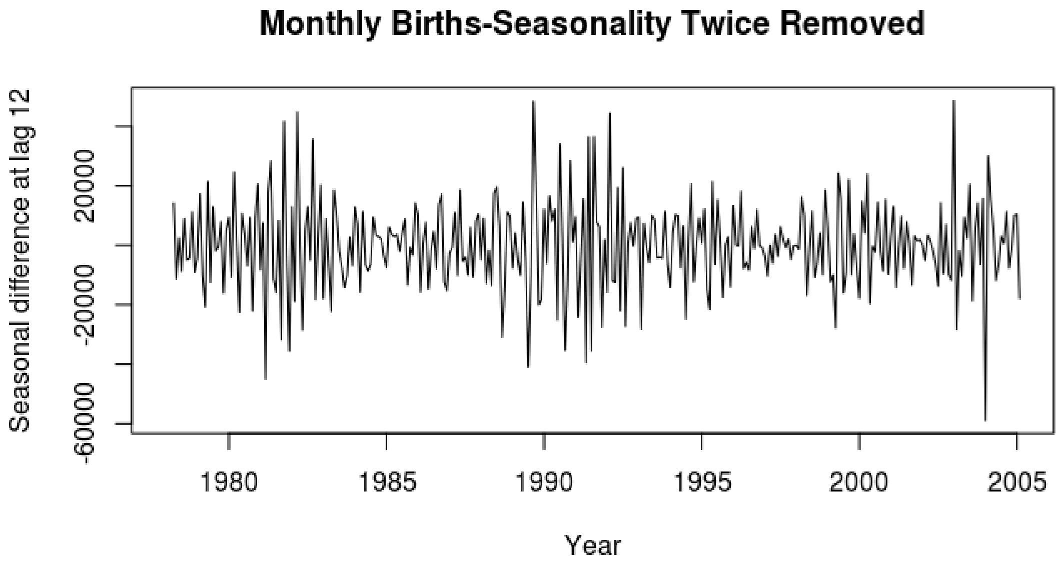


Figure 8. Time series after first-order differencing and two seasonal differences.

This second seasonal difference as seen below helps to minimize autocorrelation at later lags in the ACF, ensuring that the time series has better stationarity prior to fitting a model. There are still spikes in the differenced series around 2003-2004; however, the remainder of the series is more white noise-like. We see a stable mean around zero and variance generally unaffected by time; thus, we can conclude that the series is weakly stationary and ready for the model-fitting process.

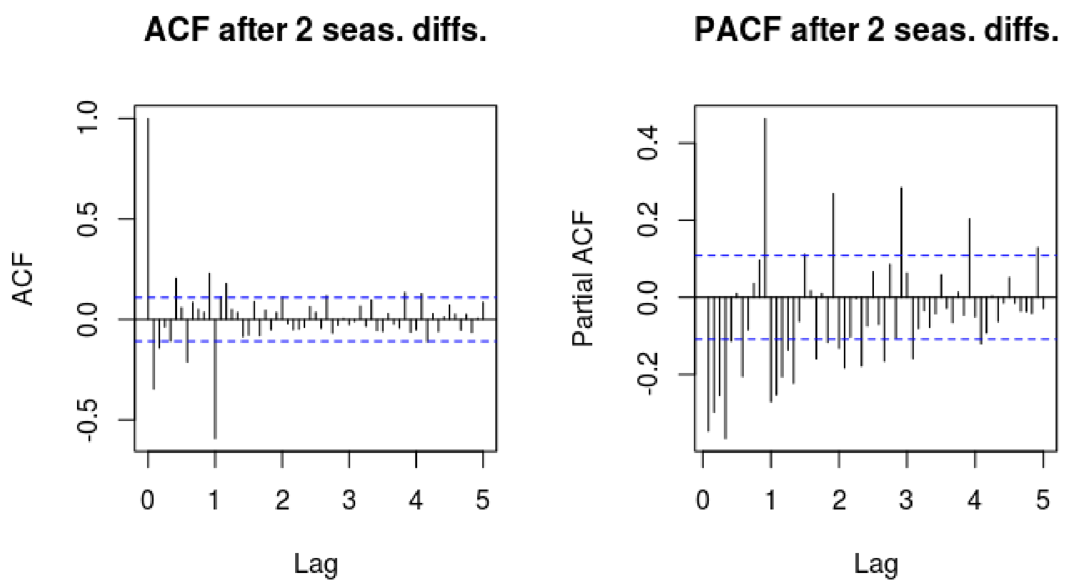


Figure 9. Autocorrelation and partial autocorrelation functions for series after applying multiple differences

Results

Model Fitting

In order to begin the model-fitting process, the R function auto.arima was used to estimate a best-fit model based on the births data. Based on this function, an ARIMA(0,1,1)(0,1,1)[12] with drift model was proposed. However, when this model was adapted to an $SARIMA(0,1,1)x(0,2,1)_{12}$ model to reflect the second seasonal differencing, the Ljung-Box statistic indicated remaining autocorrelation and poor model fit.

Instead, a model was fit based on the ACF and PACF plots above with orders chosen reflecting significant lags and accompanying cutoff points. The ACF showed significance at the first seasonal lag = 12, suggestive of a seasonal MA(1) element (Q=1). The PACF had non-zero values at all lags that were multiple of 12, indicative of a seasonal AR(1) element (P=1). To determine ideal non-seasonal orders, the lower level lags were examined. There were varying signficant and non-significant values in the ACF plot up to lag=6. Values in the PACF tailed off and cut off around lag=5. A model of $SARIMA(5,1,6)x(0,2,1)_{12}$ was then fit as well as neighboring $SARIMA(6,1,5)x(0,2,1)_{12}$, $SARIMA(5,1,5)x(1,2,1)_{12}$ and $SARIMA(5,1,5)x(0,2,1)_{12}$

After a process of trial and error, a $SARIMA(5,1,5)$ x $(1,2,1)_{12}$ model was selected as the model with the lowest Akaike Information Criterion of 6643.39.

Model Diagnosis

Diagnostics of the $SARIMA(5,1,5)$ x $(1,2,1)_{12}$ model indicate that it should be an adequate model for prediction. In the residuals above, we see few lags outside the bounds of significance except for seasonal lags. The Ljung-Box statistic plot at the bottom of the figure below confirms that the residual p-values for the fitted model are all greater than 0.5. Thus, there is no auto-correlation, and the model provides a good fit.

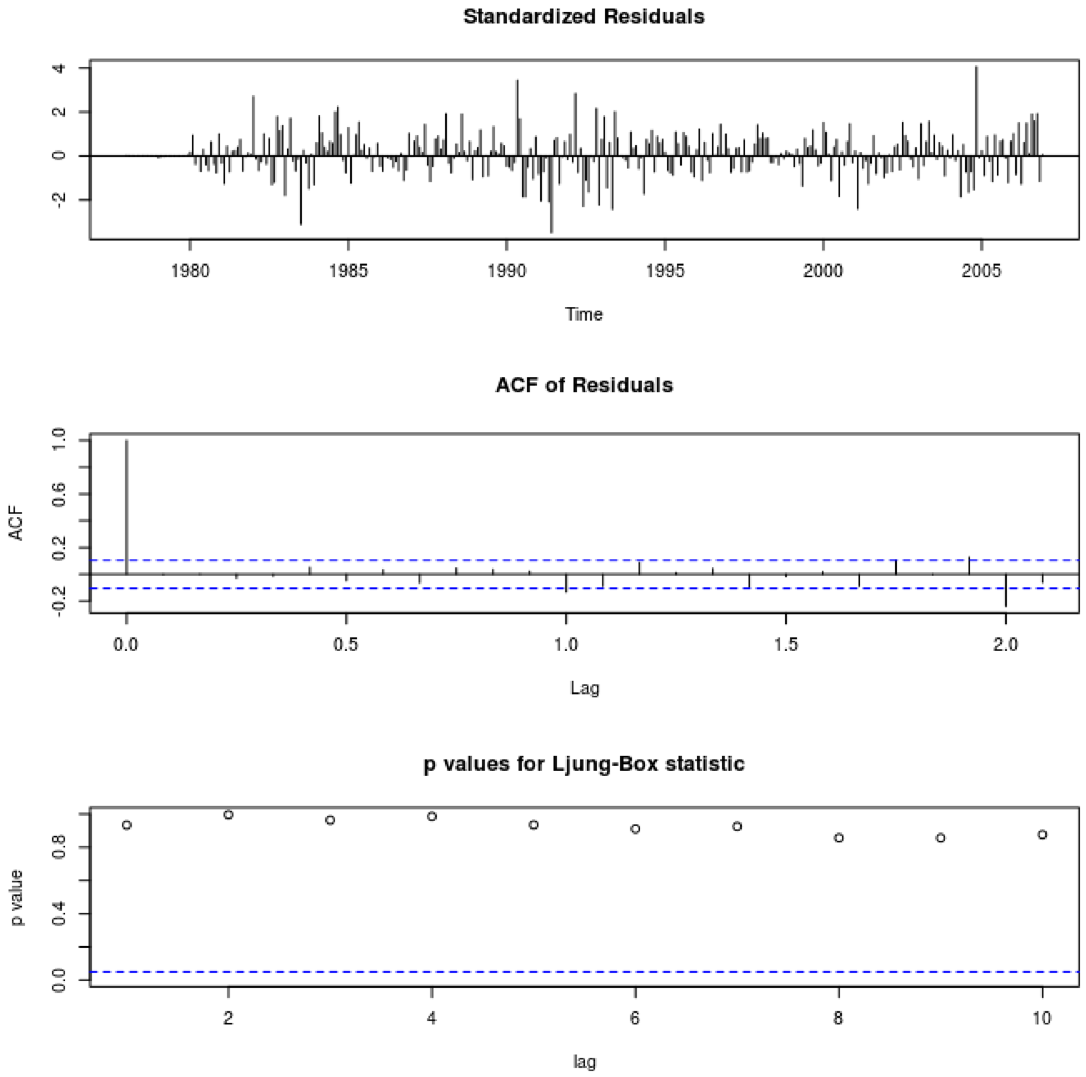


Figure 10. Diagnostics for *SARIMA* (5,1,5)x(1,2,1)$_{12}$ model.

Additionally, the Q-Q plot also shows that the assumption of normality is reasonable as most of the values fall along the given line with some exceptions toward the extremes. After assessing these diagnostic plots, we may proceed with the proposed model for prediction.

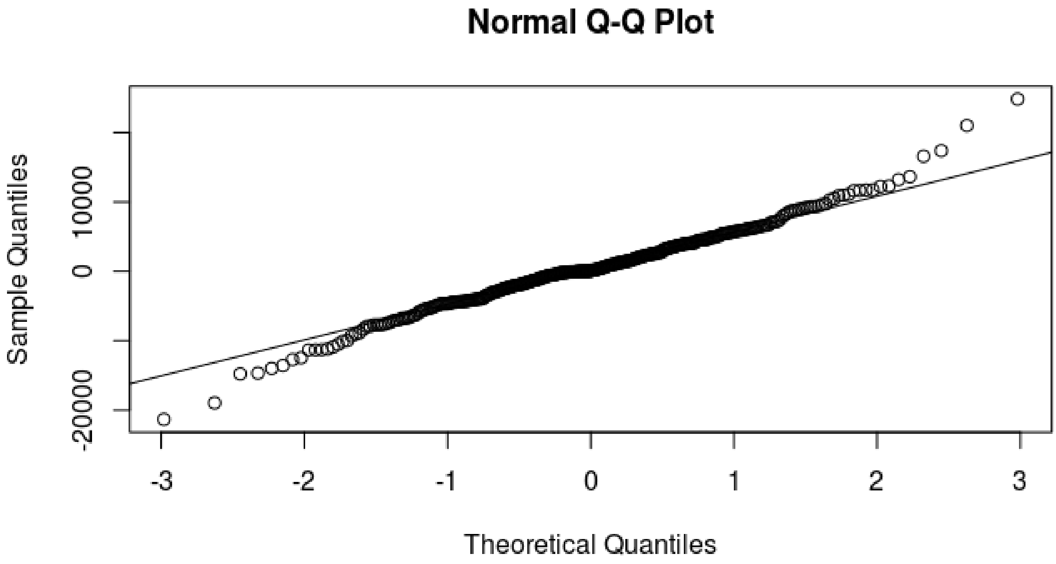


Figure 11. Q-Q plot for assessment of normality of *SARIMA* (5,1,5)x(1,2,1)$_{12}$ model.

Predictions

The proposed $SARIMA(5,1,5)$ x $(1,2,1)_{12}$ model was used to predict the number of births for each of the twelve months of 2007. These values can be compared with forecasts from the Holt-Winter's method with additive seasonality based on the relatively constant variance in the time series. Additonally, since the data for number of U.S. births for 2007 was available as part of the original dataset, it was excluded from model-fitting as a sample test set. We will examine the performance of both the Holt-Winter's and ARIMA methods in predicting the accurate number of births for each month.

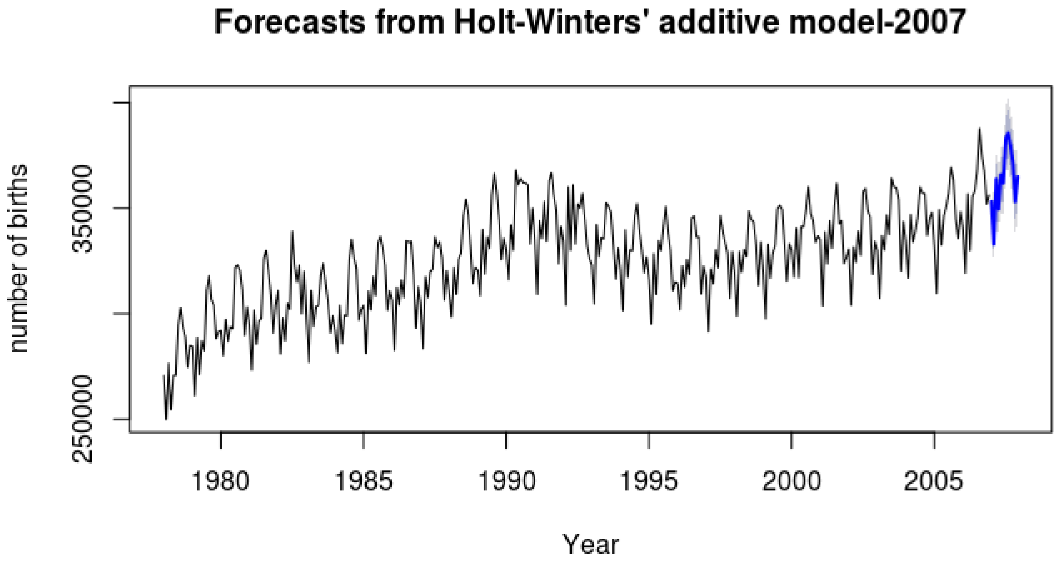


Figure 12. Year 2007 birth forecasts from Holt-Winter's additive model

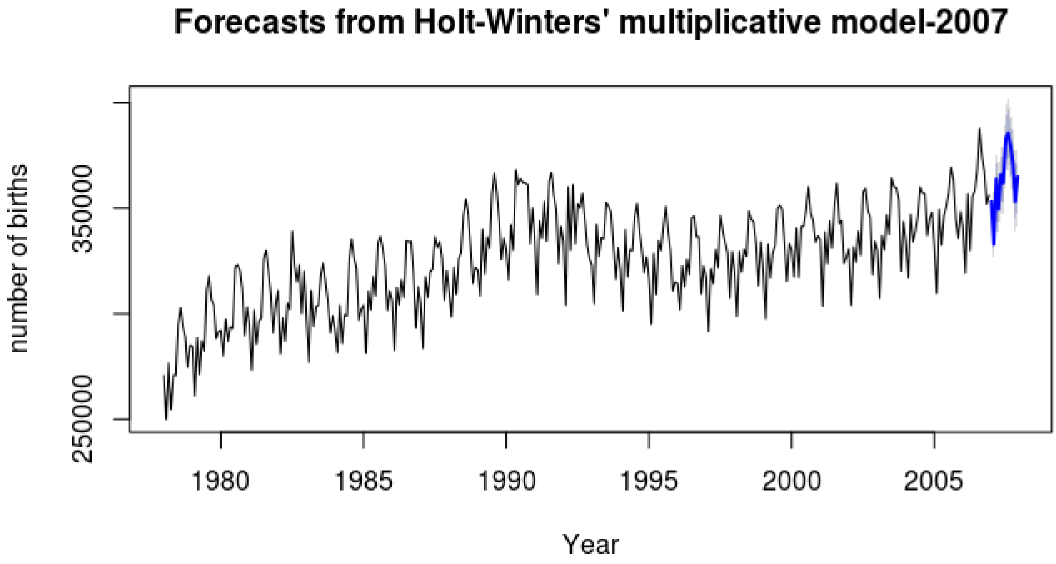


Figure 13. Year 2007 birth forecasts from Holt-Winter's multiplicative model

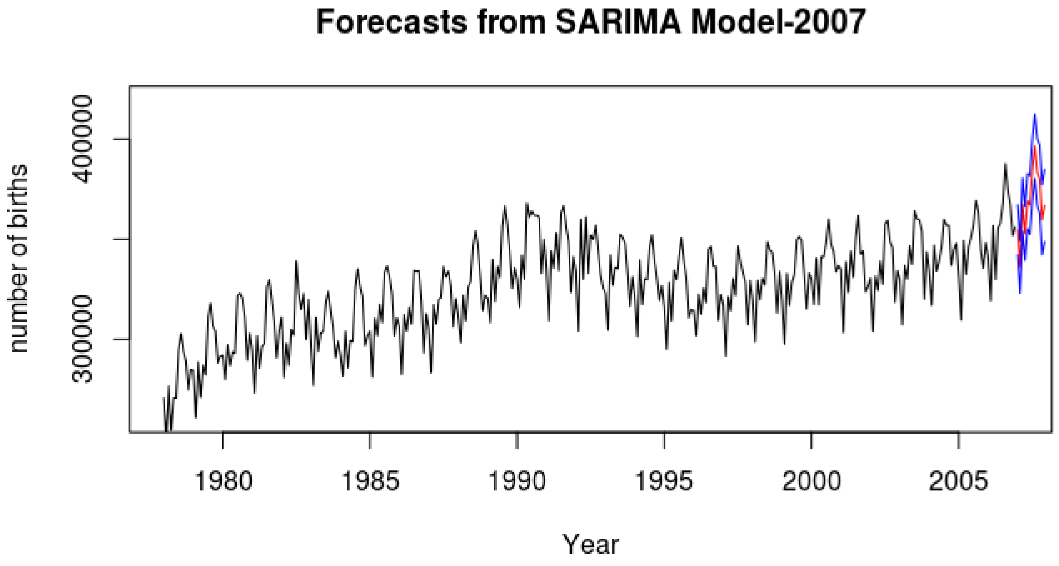


Figure 14. Year 2007 birth forecasts from SARIMA model with 95% confidence bands

Comparison of Predictions Methods

Month	2007 births	H-W forecasts	HW-M forecasts	ARIMA predictions	H-W error	HW-M error	ARIMA error
Jan	354943	353080	351707	354753	-1863	-3236	-190
Feb	326891	332871	329203	336157	5980	2312	9266
Mar	360828	363850	363126	367323	3022	2298	6495
Apr	328224	349449	346178	352878	21225	17954	24654
May	362319	365917	365228	369082	3598	2909	6763
Jun	358606	361906	360087	367065	3300	1481	8459
Jul	379616	382843	384136	384512	3227	4520	4896
Aug	390387	385597	388427	396454	-4790	-1960	6067
Sep	366904	379553	381210	383403	12649	14306	16499
Oct	369324	371667	373199	380076	2343	3875	10752
Nov	353660	353115	353171	359664	-545	-489	6004
Dec	354540	364881	365848	366769	10341	11308	12229

Table 1. Actual year 2007 births, predictions and error for 3 forecasting methods

The table above features the actual number of births in 2007 compared with forecasts from both of the Holt-Winter's smoothing method as well as ARIMA modeling (rounded to the nearest whole number). This chart includes forecasts from both additive and multiplicative Holt-Winter's forecasting methods, though they perform similarly.The final three columns consider the difference between each method's predictions and the actual values in order to calculate the error for each method given the test set available.

When judging performance of the two different methods of forecasting, Holt-Winter's and ARIMA modeling, the absolute values of these columns can be compared in order to determine which is more accurate. Based on these values, the predictions from Holt-Winter's perform better on average. When compared with the number of actual births in 2007, the error is less for the Holt-Winter's predictions in every month except for January.

Discussion

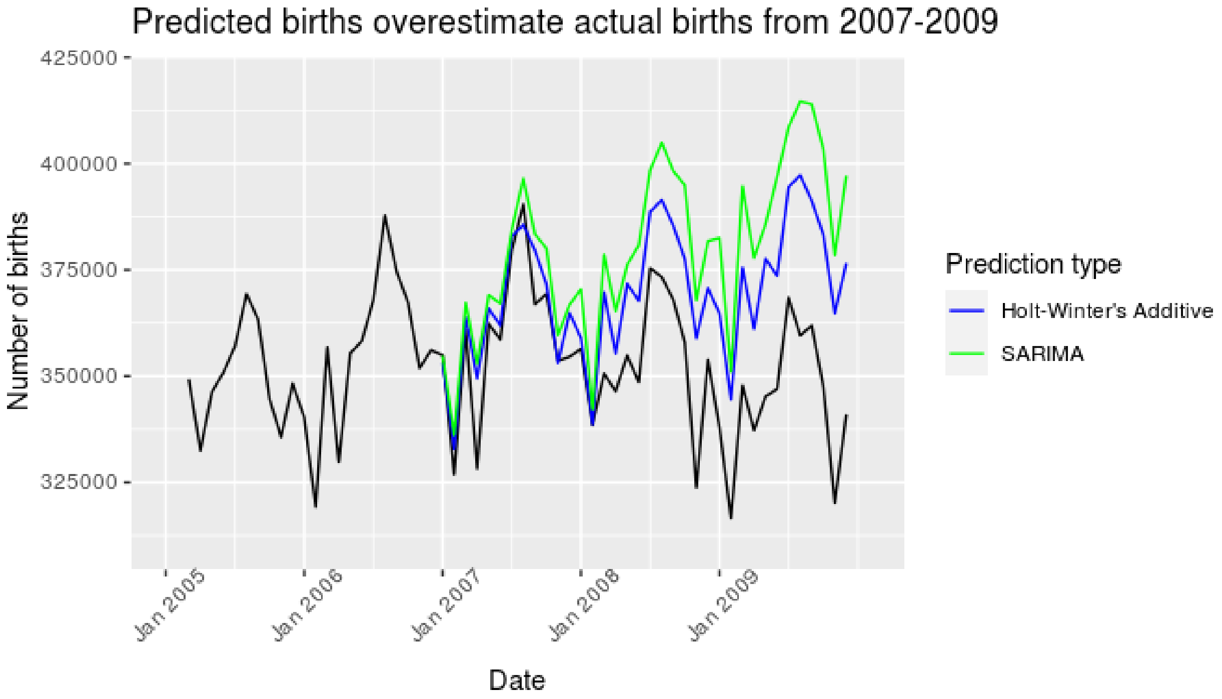


Figure 15. Predictions and actual values for 3 cycles following observed time series.

Examining the predictions compared to actual values for subsequent years can provide an additional layer of insight into the performance of the forecasting methods and into also the data itself. We see that the two different types of predictions mirror each other quite closely for most of 2007-2009 in terms of shape; however, SARIMA predictions generally are larger. It appears that the SARIMA model exaggerates the increasing trend observed in the series from 1978-2006, while the Holt-Winter's method is slightly more conservative. This seems to be the superior approach in this case as the actual data behave differently than we might rationally have expected; instead of continuing to increase, the trend appears to be decreasing during these 3 years. Without more information, it is difficult to know whether or not this is a true reversal of the trend or simply a cyclical drop as observed in prior decades.

In order to investigate more current trends and make more relevant and timely forecasts, it is essential to consider more up-to-date data in SARIMA modeling. Data is available from the United Nations Statistics Division up to 2019; however, all 12 months of 2010 are missing from the archives. Data imputation would need to be performed in order to build a more current model; a recommended procedure for this would be to assign values based on predictions from the Holt-Winter's method with data up to 2009.

Summary

In summary, there is an overall increasing trend in births in the United States since the late 1970s with both seasonal and cyclical variation. This seasonality indicates that more births occur in mid- to late summer, peaking in August and declining in the winter months. Variance in number of births is generally stable over time. While both Holt-Winter's additive smoothing and ARIMA modeling provide insight into this time series, Holt-Winter's holds the edge in terms of predictive power in this particular instance.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(fig.width=7, fig.height=4)
library(forecast)
library(ggplot2)
library(ggfortify)
#Avril
#births<-read.csv('/Users/avrilalfred/ST566_Time/ST566_Module_10/birth
s78_06.csv', header=T)

#Brittany
births <- read.csv('births78_06.csv')
#str(births)
#head(births)
# Number of babies born each month
babies<-births$Value

# Making the time series
births<-ts(data=babies, start=c(1978, 1), deltat=1/12)
str(births)
# Producing the time plot
plot(births,ylab="number of live births", xlab="Year", main="Live Birt
hs by Month-USA (1978-2006)")
births.dec <- decompose(births)
```

```r
plot(births.dec)
# create a time veriable
birth.time <- time(births)

#Approach 3: Kernel Smoothing
# fit a nonparametric trend
birth.loess <- loess(births ~ birth.time)

# get the trend
birth.loess.pred <- predict(birth.loess)

# change it into a time series object
birth.loess.trend <- ts(birth.loess.pred, start = c(1978, 1), frequenc
y = 12)

# overlay the trend on the time plot
plot(births, ylab = "Number of live births", main = "Trend Estimates v
ia Kernel Smoothing")
lines(birth.loess.trend, col = "red", lty = 2, lwd = 2)
diff.trend <- c(NA, diff(births))
diff.trend <- ts(diff.trend, start = c(1978,1), deltat = 1/12)
plot(diff.trend, xlab = "Year", ylab = "Differenced series", main="Mon
thly Births-Trend Removed")

# Data processing
month<- rep(1:12,len=length(births))
year<-rep(1978:2006, each = 12)
births_by_month<-as.data.frame(cbind(month,year,babies))

# Plotting a smoothed seasonal trend
qplot(month,babies, main = "Estimated Seasonality", geom="line", group
=year, data=births_by_month, alpha=I(0.3))+
  geom_smooth(aes(group=1), method="loess", se=FALSE, size=2)+
  ylab("number of Births")+xlab("Month")+
  scale_x_discrete(labels=c("Jan", "Feb", "Mar","Apr","May","Jun","Jul
","Aug","Sep","Oct","Nov", "Dec"),limits=c(1:12))
diff12 <- c(NA, diff(diff.trend, lag = 12))
diff12 <- ts(diff12, start = c(1978,1), deltat = 1/12)
plot(diff12, xlab = "Year", ylab = "Seasonal difference at lag 12", ma
in="Monthly Births - Trend and Seasonality Removed")
par(mfrow=c(1,2))
acf(diff12, lag.max = 60, na.action = na.pass,main = "ACF for differen
ced series", cex=0.5)
pacf(diff12, lag.max = 60, na.action = na.pass,main = "PACF for differ
enced series", cex=0.5)
#second seasonal difference
```

```
diff212 <- c(NA, diff(diff12, lag = 12))
diff212 <- ts(diff212, start = c(1978,1), deltat = 1/12)
plot(diff212, xlab = "Year", ylab = "Seasonal difference at lag 12", m
ain="Monthly Births-Seasonality Twice Removed")
par(mfrow=c(1,2))
acf(diff212, lag.max = 60, na.action = na.pass,main = "ACF after 2 sea
s. diffs.", cex=0.5)
pacf(diff212, lag.max = 60, na.action = na.pass,main = "PACF after 2 s
eas. diffs.",cex=0.5)
##
auto.model<-auto.arima(births)
summary(auto.model)

#adapted auto-suggested model
arima011021 <- arima(births, order = c(0,1,1),seasonal = list(order =
c(0,2,1), period = 12))
summary(arima011021) #aic = 6741.01, failed diag
tsdiag(arima011021)
#fits

arima615121 <- arima(births, order = c(6,1,5),seasonal = list(order =
c(1,2,1), period = 12))
summary(arima615121) #aic = 6644.64

arima516121 <- arima(births, order = c(5,1,6),seasonal = list(order =
c(1,2,1), period = 12))
summary(arima516121) #aic = 6668.69

arima515121 <- arima(births, order = c(5,1,5),seasonal = list(order =
c(1,2,1), period = 12))
summary(arima515121) ###orig mod, aic 6643.39 - winner, passes diag

arima515021 <- arima(births, order = c(5,1,5),seasonal = list(order =
c(0,2,1), period = 12))
summary(arima515021) ### aic = 6693.99


# For a table
a <- arima011021$aic
b <- arima615121$aic
c <- arima516121$aic
d <- arima515121$aic
e <- arima515021$aic

abcde <- rbind(a, b, c, d, e)
min(abcde)
```

```r
par(mfrow=c(1,2))
res <- arima515121$residuals
acf(res, lag.max = 36)
pacf(res, lag.max = 36)
tsdiag(arima515121)
par(mfrow=c(1,1))
qqnorm(res)
qqline(res)
#Using Holt-Winters Method
#additive
fore.birth <-hw(births, seasonal = "additive", h = 12)
round(fore.birth$mean)

#multiplicative
fore.birth.m <-hw(births, seasonal = "multiplicative", h = 12)
round(fore.birth.m$mean)
plot(fore.birth, main="Forecasts from Holt-Winters' additive model-2007", ylab="number of births", xlab="Year")
plot(fore.birth, main="Forecasts from Holt-Winters' multiplicative model-2007", ylab="number of births", xlab="Year")
#Model based predictions
pred <- predict(arima515121, n.ahead = 12)
round(pred$pred)
plot(births, xlim = c(1978, 2007),
ylab = "number of births", ylim=c(260000,420000), xlab="Year", main="Forecasts from SARIMA Model-2007")
####forecasted values
lines(pred$pred, col = "red")
####95% forecasting limits
lines(pred$pred-2*pred$se,col='blue')
lines(pred$pred+2*pred$se,col='blue')
####zoomed in predictions
plot(births, xlim = c(2005, 2008), ylim = c(260000, 420000), ylab = "number of births", xlab="Year", main="Zoomed-in Forecasts from SARIMA Model-2007")
lines(pred$pred, col = "red")
####95% forecasting limits
lines(pred$pred-2*pred$se,col='blue')
lines(pred$pred+2*pred$se,col='blue')
#difference between prediction methods
round(fore.birth$mean - pred$pred)
#get actual 2007 values
births2007 <- read.csv('births2007.csv')
births2007
#compare hw predictions to actual values
hwforevreal <- round(fore.birth$mean - births2007$Value)
```

```
#compare hw predictions to actual values
hwforemvreal <- round(fore.birth.m$mean - births2007$Value)
#compare arima predictions to actual values
arimavreal <- round(pred$pred - births2007$Value)
months <- c("Jan", "Feb", "Mar","Apr","May","Jun","Jul","Aug","Sep","O
ct","Nov", "Dec")
forecast_table <- cbind(months, births2007$Value, round(fore.birth$mea
n), round(fore.birth.m$mean), round(pred$pred), hwforevreal, hwforemvr
eal, arimavreal)
colnames(forecast_table) <- c("Month", "2007 births", "H-W forecasts",
"HW-M forecasts", "ARIMA predictions", "H-W error", "HW-M error", "ARI
MA error")
births_2007_total <- sum(births2007$Value)
births_2007_total

hwfore_total <- round(sum(fore.birth$mean))
hwfore_total

arima_total <- round(sum(pred$pred))
arima_total
library(knitr)
kable(forecast_table, align="c", caption = "Comparison of Predictions
Methods")
## Long-term performance - remove!
births08_15 <- read.csv("births_08_15.csv")
births08_15
#Using Holt-Winters Method
#additive
fore.birth.2 <-hw(births, seasonal = "additive", h = 96)
round(fore.birth.2$mean)

#multiplicative
fore.birth.2m <-hw(births, seasonal = "multiplicative", h = 96)
round(fore.birth.2m$mean)
plot(fore.birth.2, main="Forecasts from Holt-Winters' additive model-2
008-15", ylab="number of births", xlab="Year")
citation("forecast")
citation("ggplot2")
```

Works cited

Darrow, L. A., Strickland, M. J., Klein, M., Waller, L. A., Flanders, W. D., Correa, A., Marcus, M., & Tolbert, P. E. (2009). Seasonality of birth and implications for temporal studies of preterm birth. Epidemiology (Cambridge, Mass.), 20(5), 699–706. https://doi.org/10.1097/EDE.0b013e3181a66e96

UN data: A world of Information, United Nations Statistics Division, Demographic Statistics Database, Live Births by Month of Birth, United States of America. http://data.un.org/Data.aspx?d=POP&f=tableCode%3A55

Software

Allaire J, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R (2020). rmarkdown: Dynamic Documents for R. R package version 2.1, https://github.com/rstudio/rmarkdown.

Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2020). *forecast: Forecasting functions for time series and linear models*. R package version 8.11, <URL: http://pkg.robjhyndman.com/forecast>.

Hyndman RJ, Khandakar Y (2008). "Automatic time series forecasting: the forecast package for R." *Journal of Statistical Software*, *26*(3), 1-22. <URL: http://www.jstatsoft.org/article/view/v027i03>.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Xie Y, Allaire J, Grolemund G (2018). R Markdown: The Definitive Guide. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 9781138359338, https://bookdown.org/yihui/rmarkdown.

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.28.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

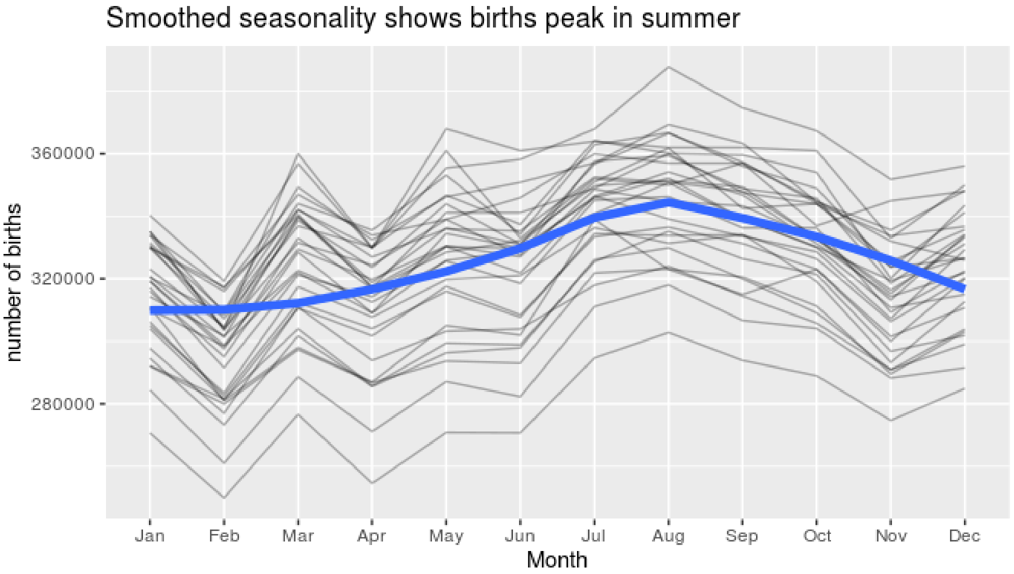
Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Blog Post: Planning our Future Using Birth Data

Take a moment to consider what life might look like in your world 10 years from now. What does you neighborhood look like? Are there more homes and residents? How is your commute? Are there freshly-paved roads to keep up with growth? What is the price of gas? What else differs from the world that surround you today? While there are many factors that might contribute to an altered distant reality, a driving force behind those is population growth. Politicians, economists, civil engineers, the medical industry and more depend on having a reasonable understanding of the structure of society in order to make decisions to best serve their communities.

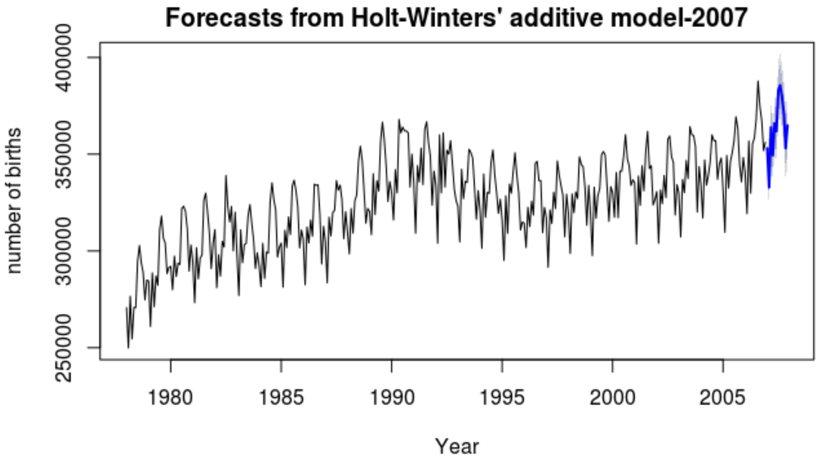
"The more you know about the past, the better prepared you are for the future". These words were spoken by one of our nation's most prominent leaders, Theodore Roosevelt and still ring true today. When we want to make predictions about what is to come, we must look at what is behind us in order to find trends that may give us insight into future patterns. This is precisely what statisticians do in time series analysis.

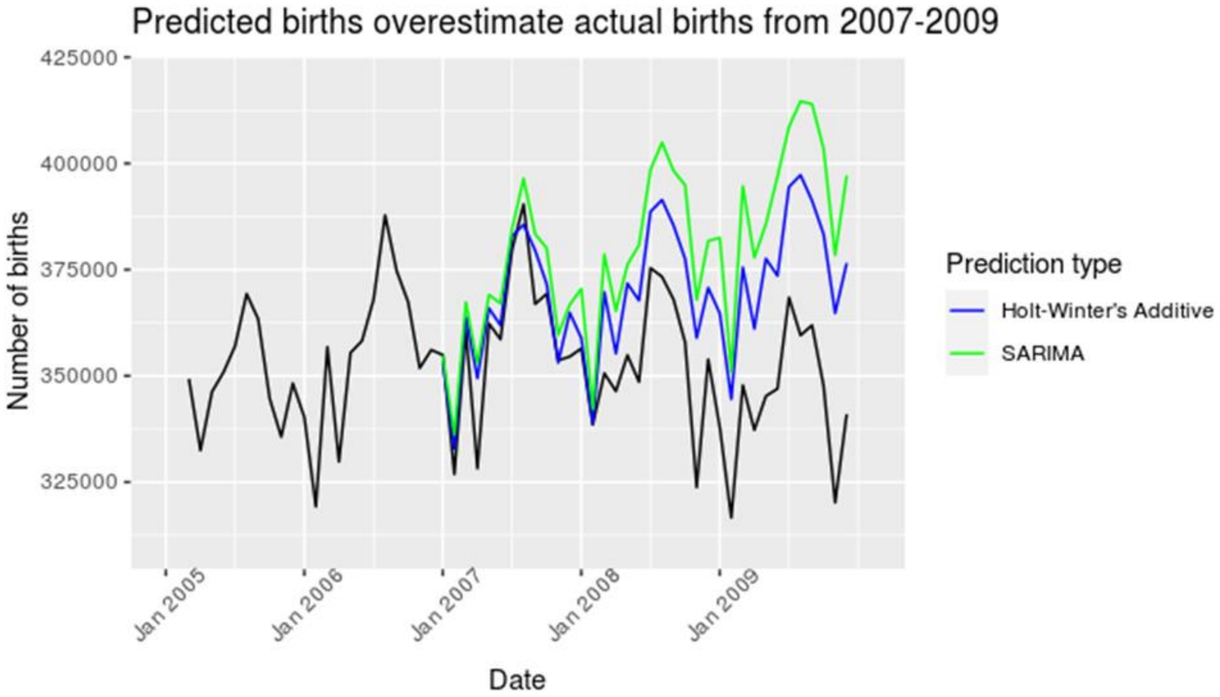
My research examines live birth data from the United States from 1978-2006 in order to better understand the evolution of our population in the last quarter of the 20th century. The goals of this analysis were twofold. The first objective includes finding trends and seasonality within the 28 years of data observed. Understanding these parts of the data allow us to understand when births rise and fall throughout the year to help medical communities plan for these events. Knowledge of trends also lead into the second goal of analysis, which is to forecast future births. This also allows for larger-scale community planning, though it should be noted births are only one aspect of calculating population growth.



I personally found this graphic of interest. It gives a smoothed average of all of the years' data for each month to see when the most births tend to occur. As an August baby myself, this made me realize I'm not quite as unique as I might have hoped. It also dispelled the myth I had always heard about all of the "Valentine's babies" born in November as numbers clearly begin to drop off at the end of the calendar year and don't pick back up again until Spring.

Another goal was to determine how many births we can expect in future years. In this analysis, two forecasting methods were utilized to make predictions: the Holt-Winter's additive model and a SARIMA (seasonal autoregressive integrated moving average) model. One of the reasons the data used only goes to 2006 is so the performance of these models could be assessed using actual birth rates from future years. In this particular case, Holt-Winter's consistently outperformed the chosen SARIMA model, and the predictions from it are shown in blue to the left. One can see that despite a few cyclical drops in births overall, the general trend of the data above is upward; that is, we can expect the number of births in the United States to continue to grow.





However, upon closer inspection, it appears that actual number of births (black line) beyond 2007 are dropping. What is going on here? Is the trend reversing and birth rates are dropping, or is this simply cyclical variation like we have seen in previous decades? We need more recent data in order to interpret this change. While there is always uncertainty in the future, the more historical information we have, the better equipped we will be for whatever lies ahead.

Summary of a Survival Analysis of Lung Cancer Patients

Introduction: According to the U.S. Centers for Disease Control and Prevention, since 2003, over 200,000 Americans each year are diagnosed with Lung and Bronchus cancer annually. This makes it the second most common type of cancer diagnosis in the United States, though it is the leading cause of cancer death (Ranchod, 2019). As a major health problem in the United States, different treatments are often being tested as an alternative to traditional chemotherapy. The Veteran's Administration study compares the survival outcomes of men who received a standard therapy or a test chemotherapy.

Data Description: The dataset comes from the U.S. Veteran's Administration and was first published in 1980 in *The Statistical Analysis of Failure Time Data* (Kalbfleisch and RL Prentice, 2002). The study followed 137 male patients with advanced inoperable lung cancer. There are nine censored observations as these patients left the study prior to death. The primary goal of the study was to assess the performance of the test chemotherapy and determine whether or not it was beneficial to patients in extending survival times. Secondary goals included analysis of covariates as prognostic variables. These covariates included the following: treatment type, cell type (histological type of the tumor), Karnofsky score, time between diagnosis and start of the study, age, and an indicator for whether or not the patient received prior therapy. Outcome of survival time was recorded in days.

Statistical Modeling: Kaplan-Meier estimates were used to visually assess the differences in survival between treatement groups. A weighted log-rank test (Wilcoxon) was used to formally test whether or not there was a significant difference in survival time based on treatment. A log-logistic accelerated failure time (AFT) model was fit to assess the effects of covariates.

Results: Based on the results of the Wilcoxon test, we find that there is no significant difference between the two treatment groups (Wilcoxon test statistic = 1.1184, p-value=.2903). To confirm this effect and those of other covariates in the dataset, a naïve Weibull model was first fit to find significant covariates; among these only cell type and Karnofsky score. Estimates for these effects were found using a log-logistic model with only cell type and Karnofsky score included.

Conclusion: To return our questions of interest, it was found that treatment did not have a significant effect on survival time between the two groups. Cell type and Karnosfky score, however, did. These effects can be quantified as follows: small cell and adeno cell types both have worse expected survival outcomes than a patient with large cell type (p-values = .0055 and .0043, respectively). The odds of a patient with small cell type surviving are approximately 49% lower than the odds of a patient with cell type 4, or large cell type holding the other covariates constant. The odds of a patient with cell type 3, adeno cell type, surviving are approximately 53% lower than the odds of a patient with cell type 4, large cell when the other covariates are held constant. Finally, Karnofsky score is another important variable in determining a patient's expected survival time (p-value <.0001) after adjusting for treatment and cell type. For every one unit increase in Karnofsky score, the odds of survival for a patient are 3.7% higher. A 10-unit increase in Karnofsky score would increase the relative odds of survival by 43%, holding all of the other covariates constant.

Veterans Administration Lung Cancer Trial Survival Analysis

Introduction

This analysis examines the effect of two different chemotherapy treatments on patients with advanced inoperable lung cancer. Specifically, two questions of interest will be addressed: 1) *Is there a significant difference in survival times between treatment groups?* 2) *Which covariates have significant effects on survival outcomes and what are these effects?*

Data

The data set 'VeteranLungCancer.csv' contains data from the Veteran's Administration Lung Cancer Trial (Kalbfleisch and Prentice). The 137 included patients were men with advanced, inoperable lung cancer and were treated with either standard chemotherapy (standard chemotherapy, n=69 patients), or a test chemotherapy (test chemotherapy, n=68). The purpose of the trial was to assess whether the test chemotherapy was beneficial; a secondary goal included the analysis of covariates as prognostic variables. Patients were followed from the beginning of treatment until death. A total of 128 patients died, while the other 9 dropped out before the conclusion of the study. One of the censored individuals was completely removed due to a suspected error in data entry, leaving a dataset of 136 patients, with 67 patients in the test chemotherapy group and 8 censored individuals. In the study, survival difference for individuals with four different cell types of lung cancer was measured: squamous cell, small cell, adeno cell and large cell. Karnofsky score, a measure of wellness, time since diagnosis, age and an indicator for prior therapy were also included as covariates. A full list of covariates can be summarized below:

- treatment: treatment type (categorical: 1 = standard chemotherapy, 2 = test chemotherapy)
- cell_type: histological type of the tumor (categorical: 1 =squamous, 2 = small cell, 3 = adeno, 4 = large)
- surv_time: survival time in days (continuous)
- c_status: censoring indicator (categorical: 0 = censored, 1= death)
- kscore: Karnofsky performance score that describes the overall patients status at the beginning of the study (continuous)
- diag_months: Time between diagnosis and start of the study in months (continuous)
- age: age of the patient in years (continuous)
- therapy: indicates if the patient has received another therapy before the current one (categorical: 0 = no, 10 = yes)

Methods and Results

Univariate analysis:

SAS software was used to perform a preliminary examination of the dataset. One observation caught our attention with a reported Karnofsky score of '99' in the type 3 cancer type. The Karnosky score of 99 may be a data entry error as the scores are typically multiples of 10. The observation was in a censored patient and was omitted from analysis (n=136). As shown in Table 1, the patients were evenly split across the treatments (n=69 and n=67) and the percentage of censored observations was very similar (5/69-7.2%, 3/67-4.5%) for both treatment groups. The censoring rate was consistent across cell types. In this study, the 9 censored observations were right censored due to patient withdrawal from the study. The minimal survival time for all patients was 1 day and the maximum was 999 days, with an overall mean survival time of 121.9 days. The minimum age was 34, the maximum age was 81, with a mean age for all patients of 58.3 years. The mean time since diagnosis was 8.82 months for all patients with treatment specific mean times of 8.65 and 8.82 for treatment 1 and 2 respectively. The minimum Karnofsky score among all of the patients was 10, and the maximum was 90, with a mean score of almost 60 (Patient requires occasional assistance, but is able to care for most of their personal needs.) In the first treatment group, 30.4% had previous therapy, where 28.4% had previous therapy in the second treatment. Individuals with squamous or large cell lung cancer had previous therapy more often than the other two kinds of lung cancer, but this was evenly split between both treatment groups.

Lung Cancer Type	Treatment	Number of Patients	Number censored
Type 1: Squamous n=35	Treatment 1	15	2
	Treatment 2	20	2
Type 2: Small cell n=48	Treatment 1	30	2
	Treatment 2	18	1
Type 3: Adeno n=26	Treatment 1	9	0
	Treatment 2	17	0
Type 4: Large cell n=27	Treatment 1	15	1
	Treatment 2	12	0

Table 1: Treatment and Censorship Numbers by Lung Cancer Cell Type

Bivariate analysis:

Bivariate analysis in Figure 1 shows the distribution of the survival times, Karnofsky scores, and time since diagnosis in months. The variables of Karnofsky score and months to diagnosis are negatively correlated (Pearson Correlation Coefficient -0.18, p-value= .03) and Karnofsky score and survival time are positively correlated (PCC 0.39, p-value= <0.0001).

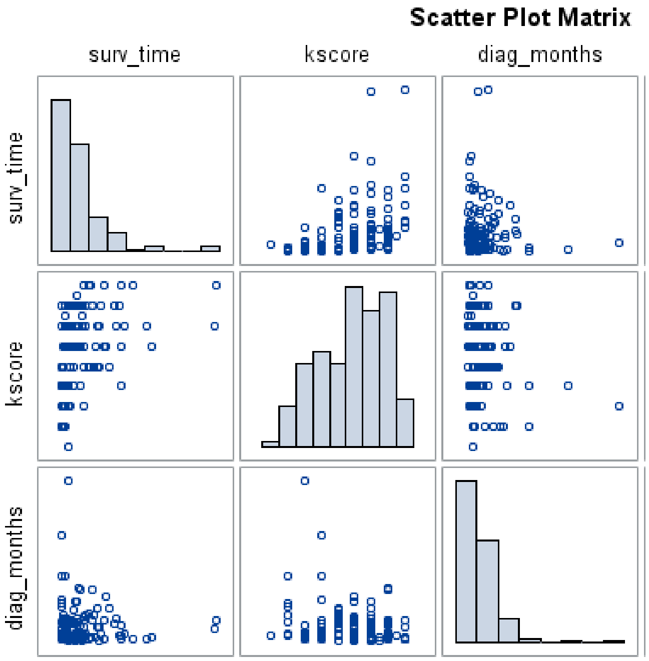


Figure 1. Correlation of covariates survival time, Karnofsky score, and months from diagnosis.

Analysis of Treatment Effect on Survival Outcomes

Question 1: Is there a significant difference in survival times between treatment groups?

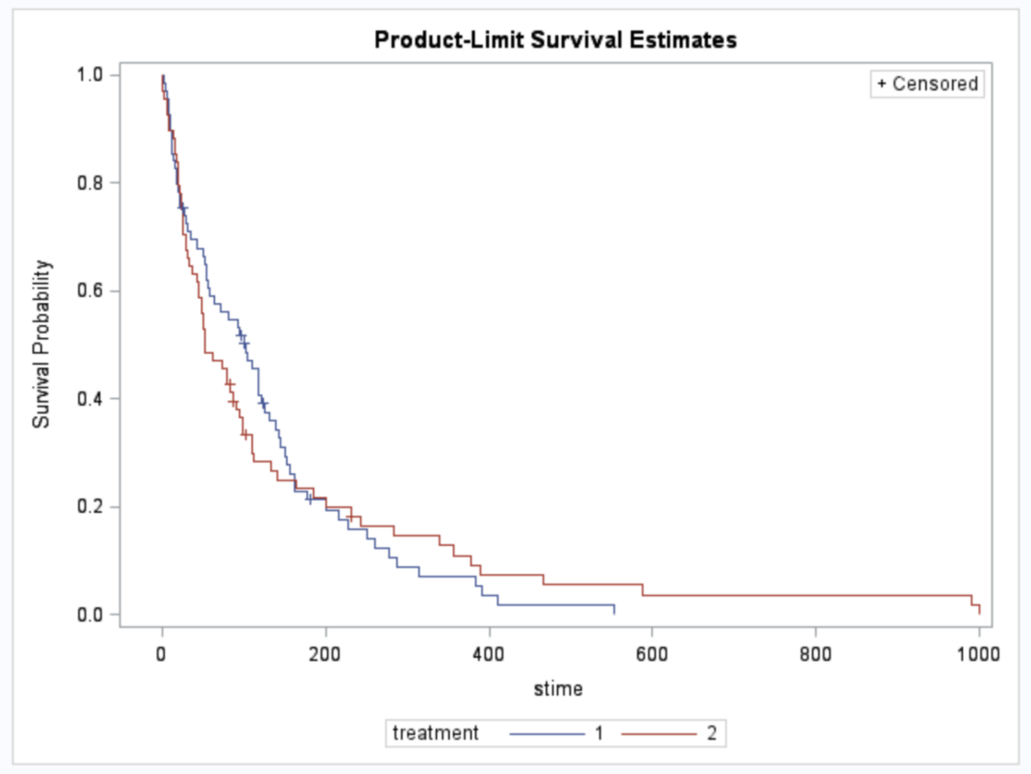


Figure 2. Kaplan Meier Survival Estimates Stratified by Treatment

To begin the process of testing whether or not the test chemotherapy drug had a significant effect on patient survival, a Kaplan-Meier estimate for each treatment group was produced (Figure 2). The survival curves for the two treatment groups are fairly close to each other for early survival times. Visually, the curves seem very similar, indicating an initial assessment that there may not be a significant difference in survival outcomes between the two groups.

A weighted log-rank test was chosen to answer this question for several reasons. First, the generic log-rank test statistic typically used to test the null hypothesis that there is no difference in survival times between groups is inappropriate due to violation of one its key assumptions. The assumption of non-informative censoring is likely reasonable; while there is no definitive explanation for the censoring provided, there are only eight censored observations in the dataset. The log-rank test also assumes proportional hazards for different covariates. Based on Figure 2, we might expect there to be a violation of the proportional hazard assumption because the survival curves cross each other in multiple places. Violation of this assumption in our dataset is confirmed and can be summarized in the following table:

Proportional Hazard assumption based on:	Kolmogorov-Type Supremum Test:	Schoenfeld residuals:
Treatment	Yes p-value = 0.3987	Yes
Cell Type 1-Squamous	Yes p-value = 0.1280	Reasonable
Cell Type 2-Small Cell	Yes p-value = 0.0559	No
Cell Type 3-Adeno	Borderline p-value = 0.0486	No
Karnofsky score	No p-value = <0.0001	No
Months since diagnosis	Yes p-value = 0.6135	Yes
Age	Yes p-value = 0.2127	Reasonable
Age (quadratic form)	Yes p-value = 0.1596	Yes
Previous therapy	Yes p-value = 0.0705	Reasonable

Table 2: Summary of proportional hazard assumption tests

As is shown in Table 2, the proportional hazard assumption was assessed in two different ways: the Komogorov-Type Surpemum Test and Schoenfeld residuals. The Komogorov Test simply provides a p-value for the null hypothesis that the proportional hazard assumption is met. The Schoenfeld residuals were assessed visually, and those that did not produce flat lines through a random pattern violated the proportional hazard assumption. Based on these two criteria, the covariates cell type and Karnofsky score were in clear violation of this assumption, making our power of the log-rank test weak, and the chance of successfully fitting a Cox Proportional Hazard model low.

Aside from being able to accommodate a non-constant hazard, a weighted form of the log-rank test may also be more appropriate than the aforementioned choices in order to weight times that are more meaningful in the context of our study. The Wilcoxon test gives greater weight to earlier failure times, which may be more appropriate for several reasons. First, lung cancer is a form of cancer that spreads extremely quickly; therefore, more attention should be given to early times as it is unlikely that the majority of patients will survive into later times. Additionally, in the context of drug testing, it can be advantageous to weight earlier failure times more heavily to pick up on any potential drug toxicity. Based on the results of the Wilcoxon test, we find that there is no significant difference between the two treatment groups (Wilcoxon test statistic = 1.1184, p-value=.2903).

Question 2: Which covariates have significant effects on survival outcomes and what are these effects?

To assess the effects of covariates other than treatment on survival outcomes, several alternative tests were conducted by stratifying by covariates with non-proportional hazards. A Wilcoxon weighted log-rank test was further utilized to investigate the survival differences between the four kinds of lung cancer and found that the cell types were significantly different in their survival without considering the effects of treatment, or other covariates. (Wilcoxon test statistic = 20.1816 df=3 p-value=.0002). This supports the intuitive notion that certain cancer types have the potential to be more lethal than others.

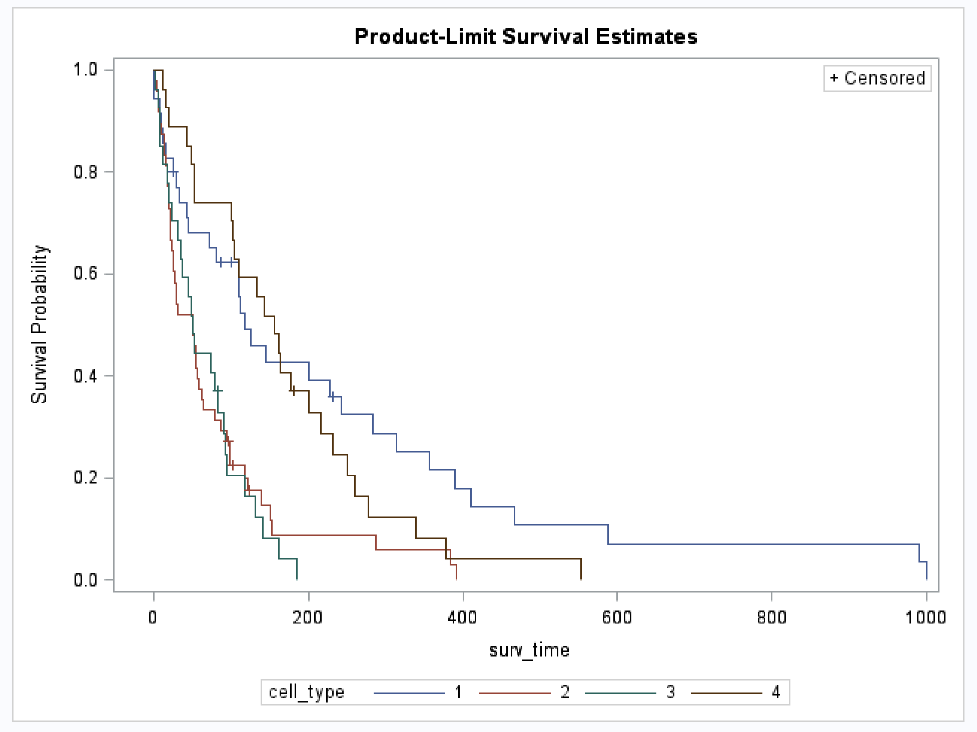


Figure 3. Kaplan-Meier Estimated Survival Curves Based on Cell Type.

Similarly, a Log-rank test to investigate the survival differences between different Karnofsky measurements also suggests that Karnofsky score will have a significant effect on survival as well. (Wilcoxon test statistic = 78.5532 df=10 p-value=<.0001)

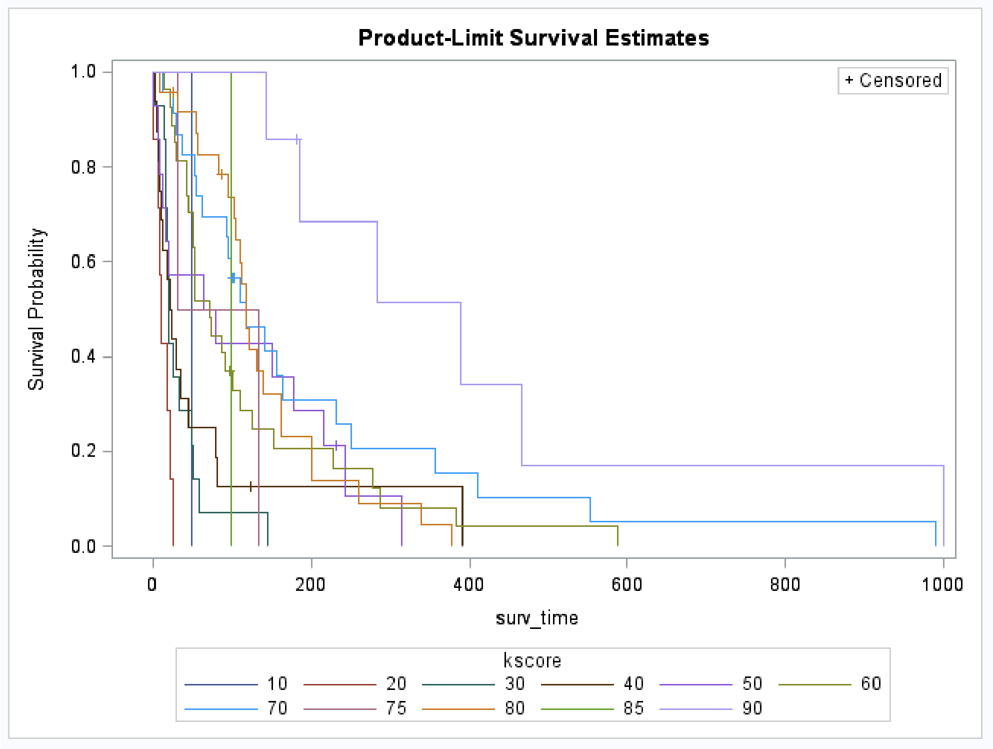


Figure 4. Kaplan Meier Estimated Survival Curves Based on Karnofsky Score.

The survival probability curves for the different treatment groups and cell types cross multiple times during the length of the study. Random crossing can occur when two curves are nearly identical or arising from small samples. Note the different scales of the horizontal axis. A Kaplan-Meier estimate of survival curves stratified by cell type was produced in order to visually assess the survival function of each cell type for each treatment in Figure 5 and 6.

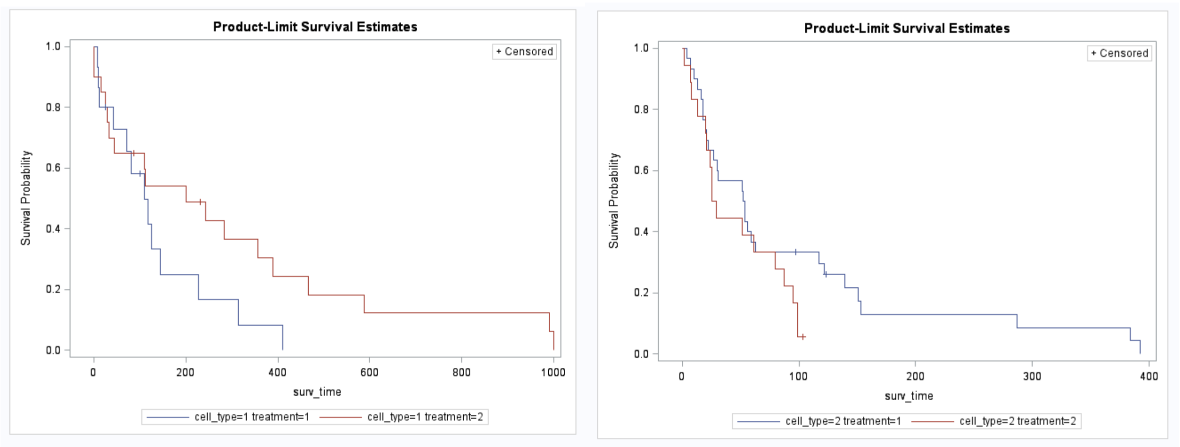


Figure 5. Kaplan Meier Estimated Survival Curves Stratified by Cell Type (1 and 2) and Treatment.

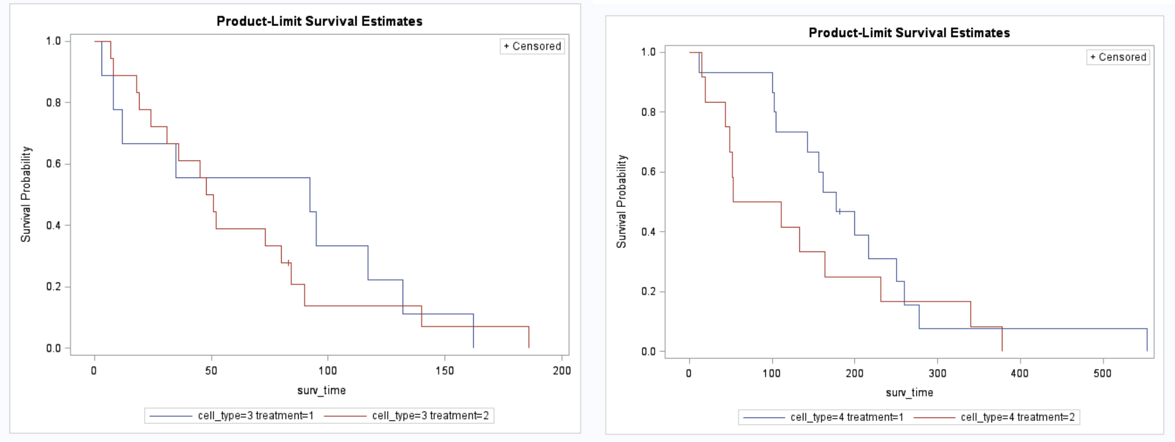


Figure 6. Kaplan Meier Estimated Survival Curves Stratified by Cell Type (3 and 4) and Treatment.

Comparing AFT models

In an attempt to fit a parametric distribution to our data in order to create a regression model to predict time to failure, several accelerated failure time (AFT) models were compared.

First, a full Weibull model with all covariates was considered in order to find significant parameters. The variables cell_type and kscore were found to have significant effects on survival time (Wald Chi-Square test, p-values <.0001 for both variables). A second Weibull model with only cell_type, kscore, and treatment as covariates was then fit. Treatment was included despite its lack of significance because it is the primary variable of interest. A likelihood ratio test was conducted to determine whether the full Weibull model with all available covariates or the reduced model with only significant covariates and treatment was a better fit to the data. Based on the likelihood ratio test comparing the full and reduced Weibull models, the reduced model was found to be sufficient (Likelihood ratio test, p-value = .178).

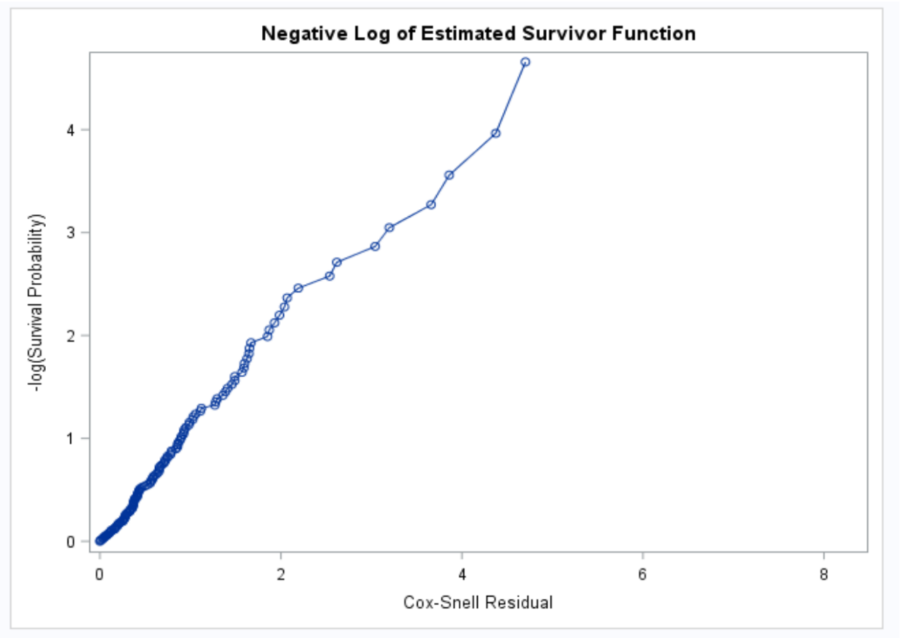


Figure 7. Cox-Snell Residuals for Reduced Weibull Model (includes covariates treatment, cell_type, and kscore).

To further assess fit of the chosen model, Cox-Snell residuals of this model were examined in Figure 7. The reduced Weibull model proves to be a reasonably good fit as the residuals produce an approximate line with a slope of 1; however, an examination of the hazard function for this data suggests a different distribution since a Weibull model would be monotonic, which is not the case in Figure 8. Instead, a unimodal distribution would be more sensible, leading to the fitting of log-normal and log-logistic models instead.

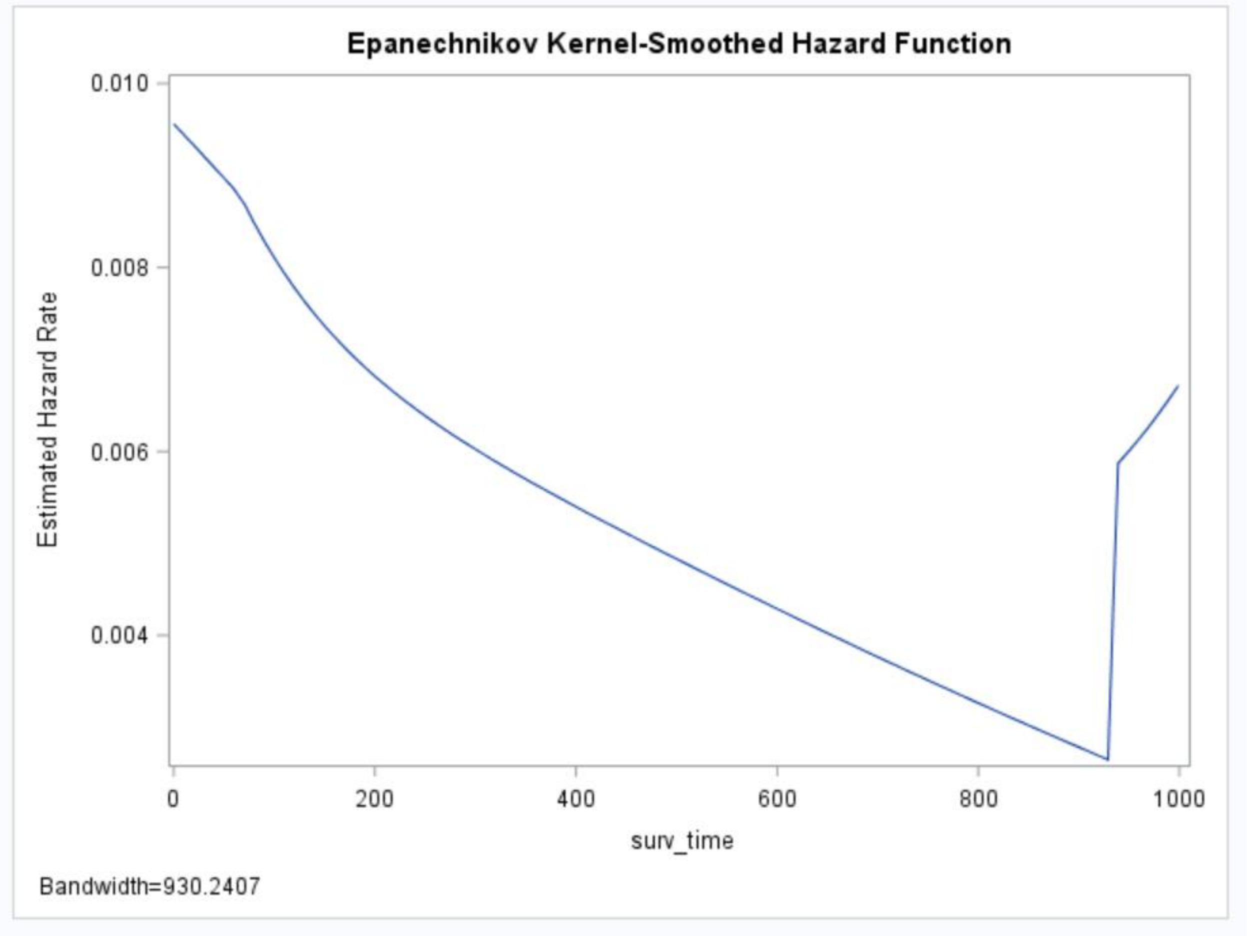


Figure 8. Estimated Hazard Rate for patients in Veteran’s Administration Lung Cancer Trial.

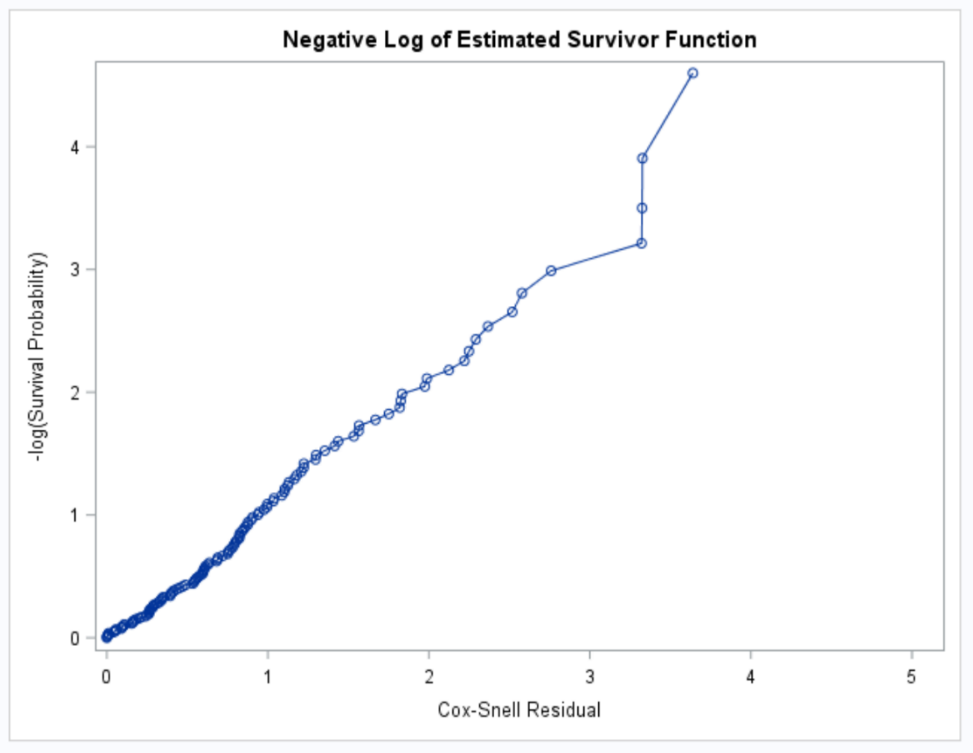


Figure 9. Cox-Snell Residuals for AFT Model with Log-Normal distribution and Covariates Treatment, Cell Type, and Karnofsky Score.

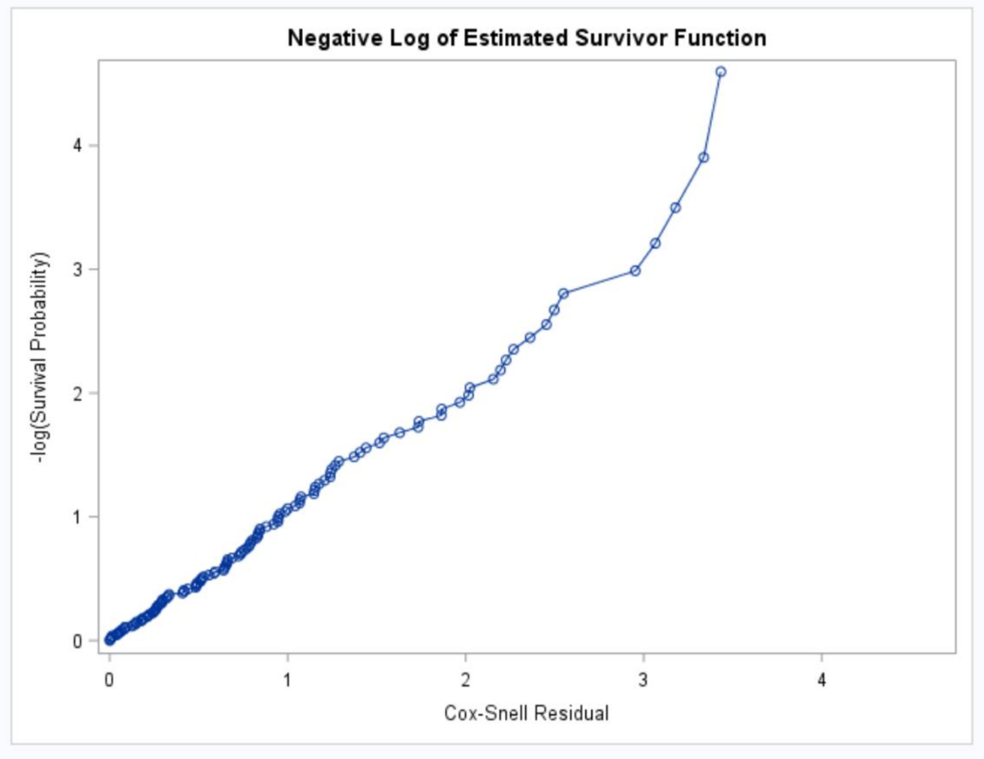


Figure 10. Cox-Snell Residuals for AFT Model with Log-Logistic distribution and Covariates Treatment, Cell Type, and Karnofsky Score.

Model Selection

Based on Figures 9 and 10, it appears the log-logistic model is preferred over the log-normal model as its residuals provide a smoother line with a more approximate slope of 1. The log-logistic model also has a lower AIC of 399.832, while the log-normal model AIC is 406. To confirm that the log-logistic distribution fits our data better than the previously fit Weibull model, a likelihood ratio test comparing the Weibull model with covariates treatment, cell_type, and kscore with the log-logistic model with the same covariates provides evidence of a better fit of the log-logistic model (likelihood ratio test, p=.01). An AIC comparison provides additional support for the log-logistic model in favor of the Weibull model (log-logistic AIC = 399.832, Weibull AIC = 406.010).

After fitting these models, however, it was noted that treatment was not a significant covariate (loglogistic model, p=.7416). Because of this, the models were changed to only contain the covariates cell type and kscore. This change resulted in a reduction of all AIC values with the lowest still being the log-logistic model (AIC = 397.941). The updated Cox-Snell residuals below confirm a satisfactory fit.

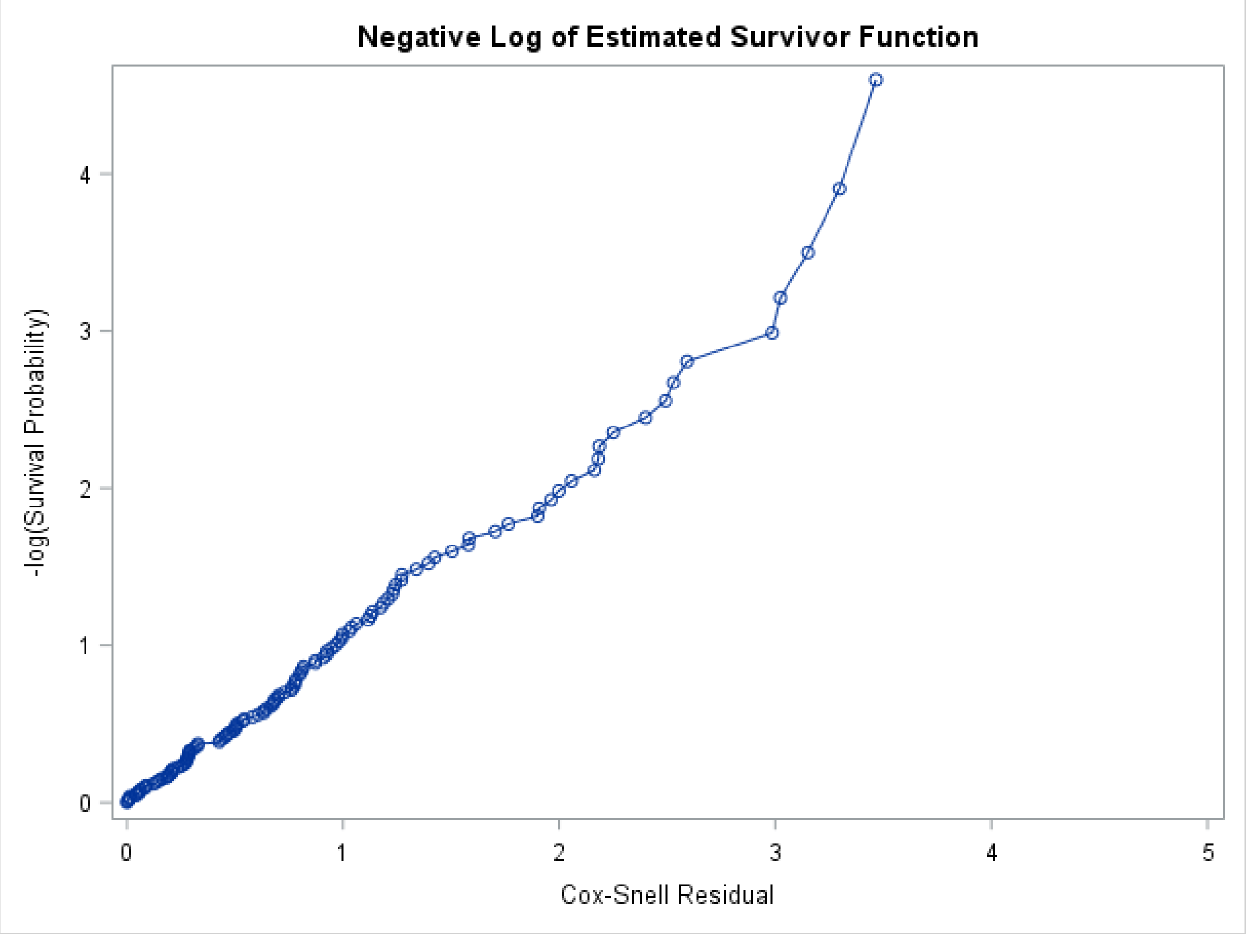


Figure 11. Cox-Snell Residuals for AFT Model with Log-Logistic distribution and Covariates Cell Type and Karnofsky Score.

Assumption checking

Before proceeding with inference based on the log-logistic AFT model, assumptions regarding non-informative censoring, sample size, and lack of multicollinearity must be verified. It is assumed that the eight censored patients in the dataset are representative of uninformative censoring and that all of the observations are independent measures. The log-logistic regression model also assumes that there is little or no multicollinearity of the covariates. Most of the covariates included in that data set have little to no correlation with the exception of some correlation between survival times and Karnofsky scores. The log-logistic regression model also requires adequate sample sizes for comparison of different groups. We are assuming that this assumption is met. Parametric survival models rely on the comparison of the data set to an underlying specified distribution. The plot in Figure 10 of the logit (log[(1-S(t))/S(t)]) against the log of survival time, produces a straight line providing evidence that log-logistic regression model is appropriate.

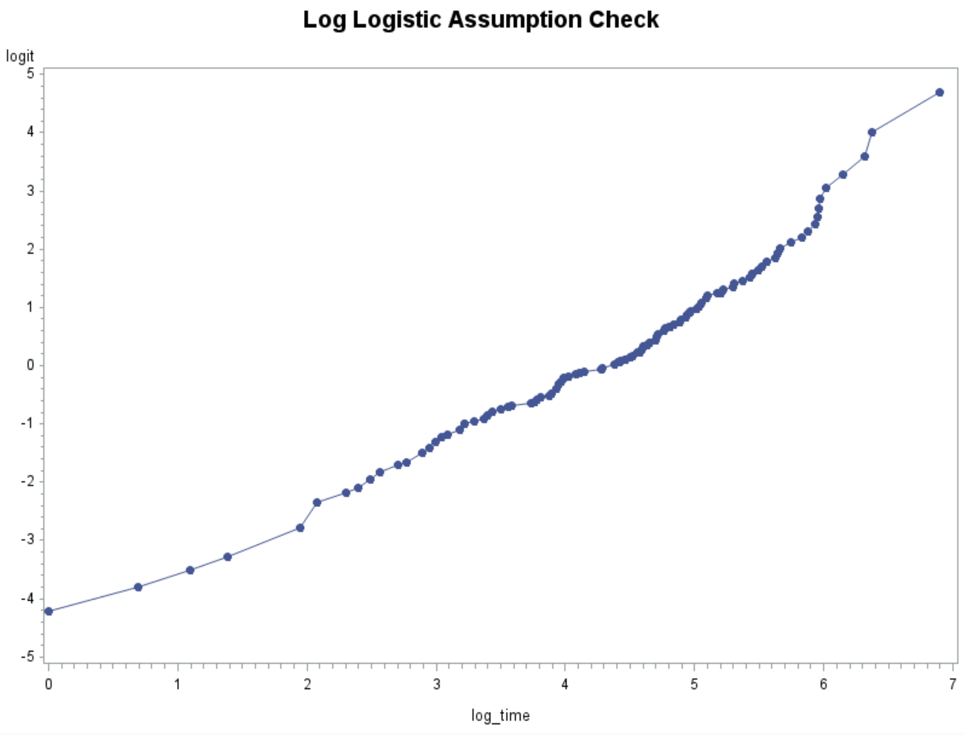


Figure 12. Linear Relationship Test for Log Time and the Logit of Survival

Interpretation

To answer the first question based on the log-logistic AFT model with covariates treatment, cell_type, and kscore, no significant effect for treatment was found, supporting the conclusion from our log-rank and Wilcoxon tests that there is no difference in survival between the two treatment groups, standard or test chemotherapy (p-value =.7416).

Other covariates have greater impact on survival, and these were assessed using an updated log-logistic model with treatment removed. Cell type affects survival outcomes after adjusting for treatment and Karnofsky score. Cell types 2 and 3, small cell and adeno, specifically show the greatest difference in survival times compared to cell type 4, large (p-values = .0055 and .0043, respectively). Both have worse survival times than patients with large cell type. The odds of a patient with cell type 2, or small cell type, surviving are approximately 49% lower than the odds of a patient with cell type 4, or large cell type holding the other covariates constant. The odds of a patient with cell type 3, adeno cell type, surviving are approximately 53% lower than the odds of a patient with cell type 4, large cell when the other covariates are held constant. There is no significant difference in the survival odds between patients with squamous and large cell types.

Finally, Karnofsky score is another important variable in determining a patient's expected survival time (p-value <.0001) after adjusting for treatment and cell type. For every one unit increase in Karnofsky score, the odds of survival for a patient are 3.7% higher. A 10-unit increase in Karnofsky score would increase the relative odds of survival by 43%, holding all of the other covariates constant.

Discussion

Based on the results of the log-rank test and Wilcoxon test, no significant differences in survival time exist between the two treatments of standard chemotherapy and the test chemotherapy drug. The log-logistic accelerated failure time model we chose with covariates treatment, cell type, and Karnofsky score also indicated that treatment had no significant effect on the expected survival time of patients. Cell type and Karnofsky score did, on the other hand, contribute to differences in survival times. Generally speaking, a higher Karnofsky score leads to better survival outcomes, as does having large cell or squamous cell types vs. small cell or adeno types.

A drawback of this analysis was the inability to successfully implement a Cox Proportional Hazard model or modified Cox model. While other valid methods of answering the questions of interest, the flexibility of the semiparametric Cox model would have been a useful tool in analyzing this data. Several covariates in the dataset violated the proportional hazard model, as seen in Table 2. This would have led to potentially stratifying the data based on the non-proportional covariates, but this presented two challenges: first, there were several non-proportional covariates, including the continuous covariate Karnofsky score. Based on this system of scoring, stratification by ranges of scores could have been considered; however, stratifying by this variable or cell type would have yielded strata with sample sizes too small to make meaningful inference. For example, cell type group 4 contains just 26 patients, while group 3 has only 27 patients. Perhaps if the original sample for this study had been larger, more options would have existed for a stratified Cox model.

Another issue with this study is the scope of inference. All patients in this study were strictly male, which means any insight we gain on survival estimates can only be used to predict survival for males. Additionally, this study was conducted by the United States Veteran's Administration, so we may want to consider whether or not we would extend inference for individuals who are not residents of the United States, and possibly even further - those who are not U.S. military veterans.

Finally, it would be helpful to know under what circumstances the censored individuals left the study. All nine of the censored observations were right-censored due to dropping out of the study rather than surviving beyond the study period. In order to make valid inference in most of the models and tests used, it is assumed that any censoring is non-informative - that is, the censoring occurred for reasons unrelated to survival outcome. This is not guaranteed based on the background information given, though since there were only 9 censored observations (prior to the removal of one observation), censoring status is unlikely to have a major impact on the results of this analysis.

Appendix

Works cited:

Cox, D. R. “Regression Models and Life-Tables.” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, 1972, pp. 187–220. *JSTOR*, www.jstor.org/stable/2985181.

https://communities.sas.com/t5/Statistical-Procedures/Proc-PhReg-Hazard-Ratio-and-Strata-statement/td-p/239948

Kaplan E.L, Meier P. “Nonparametric estimation from incomplete observations.” *Journal American Statistical Association,* vol 53, no. 282 1958, pp 457–481. *JSTOR*, https://www.jstor.org/stable/2281868.

https://www.lexjansen.com/phuse/2009/sp/SP02.pdf

Li H, Han D, Hou Y, Chen H, Chen Z (2015) Statistical Inference Methods for Two Crossing Survival Curves: A Comparison of Methods. PLoS ONE 10(1): e0116774. https://doi.org/10.1371/journal.pone.0116774

Ranchod, Yamini Phd (2019) “The 13 Most Common Cancer Types”. *Healthline.* https://www.healthline.com/health/most-common-cancers#breast-cancer

The data analysis for this paper was generated using SASsoftware, Version 9.4 of the SAS System for Windows. Copyright © [2015] SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Data:

Kalbfleisch, J.D and Prentice, R. L.(2002) *The Statistical Analysis of Failure Time Data,* Second Edition. John Wiley & Sons, Inc. pp 378-379. Retrieved from https://onlinelibrary-wiley-com.ezproxy.proxy.library.oregonstate.edu/doi/book/10.1002/9781118032985

SAS code:

/* Reading data into SAS */

```
data lung;

infile 'VeteranLungCancer.csv'  dlm=',' ;

input treatment cell_type surv_time c_status kscore diag_months age therapy;

run;
```

```
proc print data=lung (obs=10);
run;

/* Initial exploration of data set*/

proc means data=lung mean min q1 median q3 max std;
var surv_time kscore diag_months age;
class treatment cell_type therapy c_status;
run;

/*Check stratum size for cell_type*/
proc sort data=lung;
by cell_type;
run;

proc sql;
select count(*) as N_obs
from lung
where cell_type = 1 OR cell_type=2 OR cell_type=3 OR cell_type=4
group  cell_type;
quit;

proc sql;
select count(*) as N_obs
from lung
where cell_type = 1;
quit;
```

```
proc sql;
select count(*) as N_obs
from lung
where cell_type = 2;
quit;

proc sql;
select count(*) as N_obs
from lung
where cell_type = 3;
quit;

proc sql;
select count(*) as N_obs
from lung
where cell_type = 4;
quit;

/*Check sample size for each treatment*/
proc sort data=lung;
by treatment;
run;

proc print data=lung;
run;

/*Check number of deaths*/
proc sort data=lung;
by c_status;
```

```
run;

proc print data=lung;
run;

/* Histograms and correlation by treatment*/
proc corr data=lung plots(maxpoints=none)=matrix(histogram);
var treatment surv_time kscore diag_months age cell_type therapy;
run;

proc corr data=lung plots(maxpoints=none)=matrix(histogram);
var therapy surv_time kscore diag_months age cell_type;
run;

/*Obtain KM estimates and Log Rank Test*/
proc lifetest data=lung alpha=0.05 method=KM outsurv=survest
plots=(s, ls, lls) graphics;
time surv_time*c_status(0);
strata treatment;
run;

/* Overall survival time by cell type*/
proc gplot data=lung;
plot surv_time*cell_type;
run;
```

Bivariate analysis:

```
/* KM-estimates by strata of cell_type*/
```

```
proc lifetest data=lung alpha=0.05 method=KM outsurv=survest

plots=(s, ls, lls) graphics;

time surv_time*c_status(0);

strata cell_type;

run;


/* KM-estimates by strata of cell_type and treatment*/

proc lifetest data=lung alpha=0.05 method=KM outsurv=survest

plots=(s, ls, lls) graphics;

time surv_time*c_status(0);

strata cell_type treatment;

run;


/* KM-estimates by strata of kscore-all and Log Rank Test*/

proc lifetest data=lung alpha=0.05 method=KM outsurv=survest

plots=(s, ls, lls) graphics;

time surv_time*c_status(0);

strata kscore;

run;


/* KM-estimates by strata of cell_type..plot of single cell_type*/

proc lifetest data=lung(where=(cell_type=4)) alpha=0.05 method=KM outsurv=survest

plots=(s, ls, lls) graphics;

time surv_time*c_status(0);

strata cell_type treatment;

run;


/* KM-estimates by strata of previous therapy and Log Rank Test*/

proc lifetest data=lung(where=(therapy=10)) alpha=0.05 method=KM outsurv=survest
```

```
plots=(s, ls, lls) graphics;
time surv_time*c_status(0);
strata cell_type treatment;  /* or omit cell_type*/
run;
```

Model fitting

```
/* Try looking at the effect of treatment on each stratum (CoxPH)*/
proc phreg data=lung plots=survival;
class treatment cell_type  therapy kscore;
model surv_time*c_status(0)= diag_months age;
strata cell_type treatment;
run;

/*Weibull regression, intercept only*/
proc lifereg data=lung;
model surv_time*c_status(0) =  / dist=weibull;
run;

/*Weibull model full*/
/*class vars = treatment, cell_type, therapy
cont vars = kscore, diag_months, age
LR test says full is better*/
proc lifereg data=lung;
class treatment cell_type c_status therapy;
model surv_time*c_status(0) = treatment cell_type kscore diag_months age therapy / dist=weibull;
run;

/*Try Weib model with only significant covars cell type and kscore*/
```

```
proc lifereg data=lung;
class treatment cell_type c_status therapy;
model surv_time*c_status(0) = cell_type kscore / dist=weibull;
run;

/*Weib model with significant covars cell type and kscore plus variable of interest treatment*/
proc lifereg data=lung;
class treatment cell_type c_status therapy;
model surv_time*c_status(0) = treatment cell_type kscore / dist=weibull;
output out=expout Cresidual=csr sresidual=sr;
run;

/*Weib model with significant covars cell type and kscore - NO TREATMENT*/
proc lifereg data=lung;
class treatment cell_type c_status therapy;
model surv_time*c_status(0) = cell_type kscore / dist=weibull;
output out=expout Cresidual=csr sresidual=sr;
run;

/*Residual check for Weibull model*/
proc lifetest data=expout plots=(ls) notable graphics;
time csr*c_status(0);
run;

/*Lognormal model with significant covars cell type and kscore plus variable of interest treatment*/
proc lifereg data=lung;
class treatment cell_type c_status therapy;
model surv_time*c_status(0) = treatment cell_type kscore / dist=lognormal;
output out=expout Cresidual=csr sresidual=sr;
```

```
run;

/*Lognormal model with significant covars cell type and kscore - NO TREATMENT*/
proc lifereg data=lung;
class treatment cell_type c_status therapy;
model surv_time*c_status(0) = cell_type kscore / dist=lognormal;
output out=expout Cresidual=csr sresidual=sr;
run;

/*Residual check for lognormal model*/
proc lifetest data=expout plots=(ls) notable graphics;
time csr*c_status(0);
run;

/*Loglogistic model with significant covars cell type and kscore plus
variable of interest treatment better than log normal*/
proc lifereg data=lung;
class treatment cell_type c_status therapy;
model surv_time*c_status(0) = treatment cell_type kscore / dist=loglogistic;
output out=expout Cresidual=csr sresidual=sr;
run;

/*Loglogistic model with significant covars cell type and kscore plus
variable of interest treatment better than log normal - NO TREATMENT*/
proc lifereg data=lung;
class treatment cell_type c_status therapy;
model surv_time*c_status(0) = cell_type kscore / dist=loglogistic;
output out=expout Cresidual=csr sresidual=sr;
run;
```

```
/*Residual check for loglogistic model*/
proc lifetest data=expout plots=(ls) notable graphics;
time csr*c_status(0);
run;

/*Log logistic fit logit log time plot*/
proc lifetest data=lung outsurv=lifeout; /*(compute Kaplan-Meier estimator)*/
time surv_time*c_status(0);
run;

/*Transformations of t and S(t) for probability plots*/
data forplot;
set lifeout;
log_time=log(surv_time);                     /*(to get log(t) )*/
logit=log((1-Survival)/Survival);   /*(for log-logistic)*/
run;

proc gplot data=forplot;
title ‘Log Logistic Assumption Check’;
axis1=’Log((1-S(t))/S(t))’
symbol1 value=dot interpol=JOIN;
plot logit*log_time;
run;

/*Exp model with significant covars cell type and kscore plus variable of interest treatment*/
proc lifereg data=lung;
class treatment cell_type c_status therapy;
model surv_time*c_status(0) = treatment cell_type kscore / dist=exponential;
```

```
output out=expout Cresidual=csr sresidual=sr;

run;


/*Residual check for exp model - same as weib?*/

proc lifetest data=expout plots=(ls) notable graphics;

time csr*c_status(0);

run;


/*Log Rank Test - Treatment*/

proc lifetest data=lung plots=(s);

time surv_time*c_status(0);

strata treatment;

run;


/*Log Rank Test - Cell Type*/

proc lifetest data=lung plots=survival(strata=overlay);

time surv_time*c_status(0);

strata cell_type/test=logrank adjust=sidak;

run;


/*Log Rank Test - Therapy*/

proc lifetest data=lung plots=(s);

time surv_time*c_status(0);

strata therapy;

run;


/*Log Rank Test - Kscore*/

proc lifetest data=lung alpha=0.05 method=KM outsurv=survest

plots=(s, ls, lls) graphics;
```

```
time surv_time*c_status(0);
strata kscore;
run;

/*Log Rank Test - Diag_months*/
proc lifetest data=lung alpha=0.05 method=KM outsurv=survest
plots=(s, ls, lls) graphics;
time surv_time*c_status(0);
strata diag_months(10 20 30)/ test=logrank adjust=sidak;
Run;

/*Log Rank Test - Age*/
proc lifetest data=lung alpha=0.05 method=KM outsurv=survest
plots=(s, ls, lls) graphics;
time surv_time*c_status(0);
strata age(30 40 50 60 70 80)/ test=logrank adjust=sidak;
run;

/*Cumulative hazard  function estimate */
ods output ProductLimitEstimates = ple;
proc lifetest data=lung(where=(c_status=1))  nelson outs=outlung;
time surv_time*c_status(0);
run;

proc sgplot data = ple;
series x = surv_time y = CumHaz;
run;

/*regular hazard*/
```

```
proc lifetest data=lung(where=(c_status=1)) plots=hazard;
time surv_time*c_status(0);
run;

/*Alternative hazard function code*/
/* KM-estimates by strata of kscore*/
proc lifetest data=lung atrisk plots=hazard(bw=200) outs=outlung_kscore;
strata kscore(25 35 45 55 65 75 85);
time surv_time*c_status(0);
run;

/* KM-estimates by strata of treatment*/
proc lifetest data=lung atrisk plots=hazard(bw=200) outs=outlung_treatment;
strata treatment;
time surv_time*c_status(0);
run;

/* KM-estimates by strata of cell_type*/
proc lifetest data=lung atrisk plots=hazard(bw=200) outs=outlung_cell;
strata cell_type;
time surv_time*c_status(0);
run;
```

CoxPH models and proportional hazard check

```
/*Martingale and deviance residuals for full CoxPH model */
proc phreg data=lung;
class treatment cell_type therapy;
model surv_time*c_status(0) = treatment cell_type kscore diag_months age therapy;
```

```
output out=Outp xbeta=Xb resmart=Mart resdev=Dev;
run;

/*Overview Martingale/deviance-looking for outliers for full CoxPH model*/
proc gplot data=Outp;
plot Mart*Xb=c_status/vaxis=-3 to 3 by 0.5 vref=0;
run;
proc gplot data=Outp;
plot Dev*Xb=c_status/vaxis=-3 to 3 by 0.5 vref=0;
run;

/*Kolmogorov-Type Supremum Test*/
proc phreg data=lung;
class treatment cell_type therapy;
model surv_time*c_status(0) = treatment cell_type kscore diag_months age|age therapy;
assess ph/ resample=1234;
run;

/*Use Martingale residuals to assess functional forms of continuous covariates for full CoxPH model*/
/*age*/
proc phreg data=lung;
class treatment cell_type therapy;
model surv_time*c_status(0) = ;
output out= Outp resmart=Mart;
run;
proc loess data = Outp plots=ResidualsBySmooth(smooth);
model Mart = age/smooth=0.2 0.4 0.6 0.8;
run;
```

/*kscore*/

```
proc phreg data=lung;
class treatment cell_type therapy;
model surv_time*c_status(0) = ;
output out= Outp resmart=Mart;
run;

proc loess data = Outp plots=ResidualsBySmooth(smooth);
model Mart = kscore/smooth=0.2 0.4 0.6 0.8;
run;
```

/*diag_months*/

```
proc phreg data=lung;
class treatment cell_type therapy;
model surv_time*c_status(0) = ;
output out= Outp resmart=Mart;
run;
proc loess data = Outp plots=ResidualsBySmooth(smooth);
model Mart = diag_months/smooth=0.2 0.4 0.6 0.8;
run;
```

/*Refit CPH model with quadratic form for age*/

```
proc phreg data=lung;
class treatment cell_type therapy;
model surv_time*c_status(0) = age age*age;
output out=Outp resmart=Mart resdev=Dev;
run;
```

/*Re-check residuals against age and no more patterns left */

```
title "Martingale residuals by age";
proc loess data = Outp plots=ResidualsBySmooth(smooth);
model Mart = age/smooth=0.2 0.4 0.6 0.8;
run;

/*Schoenfeld residuals */
proc phreg data=lung;
class treatment cell_type therapy;
model surv_time*c_status(0) = treatment cell_type kscore diag_months age|age therapy;
output out=OutS ressch=schtreatment schcell_type1 schcell_type2 schcell_type3  schkscore schdiag_months schage schage2 schtherapy;
run;

proc loess data = OutS;
model schtreatment=surv_time / smooth=(0.2 0.4 0.6 0.8);
run;

proc loess data = OutS;
model schcell_type1=surv_time / smooth=(0.2 0.4 0.6 0.8);
run;

proc loess data = OutS;
model schcell_type2=surv_time / smooth=(0.2 0.4 0.6 0.8);
run;

proc loess data = OutS;
model schcell_type3=surv_time / smooth=(0.2 0.4 0.6 0.8);
run;
```

```
proc loess data = OutS;

model schkscore=surv_time / smooth=(0.2 0.4 0.6 0.8);

run;


proc loess data = OutS;

model schdiag_months=surv_time / smooth=(0.2 0.4 0.6 0.8);

run;


proc loess data = OutS;

model schage=surv_time / smooth=(0.2 0.4 0.6 0.8);

run;


proc loess data = OutS;

model schage2=surv_time / smooth=(0.2 0.4 0.6 0.8);

run;


proc loess data = OutS;

model schtherapy=surv_time / smooth=(0.2 0.4 0.6 0.8);

run;
```

/*Stratified on Cell_type Full CoxPH Model*/   (kscore and therapy violate PH assumption)

```
proc phreg data=lung;

class treatment cell_type kscore therapy;

model surv_time*c_status(0) = treatment age|age kscore diag_months therapy ;

strata cell_type;

output out= Out_full_CoxPH resmart=Mart;

run;


/*treatment*/
```

```
proc loess data = Out_full_CoxPH plots=ResidualsBySmooth(smooth);
model Mart = treatment/smooth=0.2 0.4 0.6 0.8;
run;

/*age*/
proc loess data = Out_full_CoxPH plots=ResidualsBySmooth(smooth);
model Mart = age/smooth=0.2 0.4 0.6 0.8;
run;

/*kscore*/
proc loess data = Out_full_CoxPH plots=ResidualsBySmooth(smooth);
model Mart = kscore/smooth=0.2 0.4 0.6 0.8;
run;

/*diag_months*/
proc loess data = Out_full_CoxPH plots=ResidualsBySmooth(smooth);
model Mart = diag_months/smooth=0.2 0.4 0.6 0.8;
run;

/*therapy*/
proc loess data = Out_full_CoxPH plots=ResidualsBySmooth(smooth);
model Mart = therapy/smooth=0.2 0.4 0.6 0.8;
run;

/*Stratified on Cell_type Larger CoxPH Model treatment age diag_months*/
proc phreg data=lung;
class treatment cell_type kscore therapy;
model surv_time*c_status(0) =  treatment age|age diag_months;
strata cell_type;
```

```
output out= Out_reduced_tadm resmart=Mart;
run;

/*treatment*/
proc loess data = Out_reduced_tadm plots=ResidualsBySmooth(smooth);
model Mart = treatment/smooth=0.2 0.4 0.6 0.8;
run;

/*age*/
proc loess data = Out_reduced_tadm plots=ResidualsBySmooth(smooth);
model Mart = age/smooth=0.2 0.4 0.6 0.8;
run;

/*diag_months*/
proc loess data = Out_reduced_tadm plots=ResidualsBySmooth(smooth);
model Mart = diag_months/smooth=0.2 0.4 0.6 0.8;
run;

/*Stratified on Cell_type Reduced CoxPH Model just age*/
proc phreg data=lung;
class treatment cell_type kscore therapy;
model surv_time*c_status(0) =  age|age;
strata cell_type;
output out= Out_reduced_age resmart=Mart;
run;

/*age*/
proc loess data = Out_reduced_age  plots=ResidualsBySmooth(smooth);
model Mart = age/smooth=0.2 0.4 0.6 0.8;
```

```
run;

/*Stratified on Cell_type Reduced CoxPH Model age and kscore*/
proc phreg data=lung;
class treatment cell_type kscore therapy;
model surv_time*c_status(0) =  age|age kscore;
strata cell_type;
hazardratio 'Effect of cell type' cell_type / at(treatment=ALL);
output out= Out_reduced_ak resmart=Mart;
run;

/*age*/
proc loess data = Out_reduced_ak  plots=ResidualsBySmooth(smooth);
model Mart = age/smooth=(0.2 0.4 0.6 0.8);
run;

/*kscore*/
proc loess data = Out_reduced_ak  plots=ResidualsBySmooth(smooth);
model Mart = kscore/smooth=(0.2 0.4 0.6 0.8);
run;
```

Summary of Analysis of New Zealand Road Crashes, 2009

Introduction: With any activity, comes risk. Drivers of cars, motorcycles, and other vehicles are made well aware of this risk by automobile insurance agencies, but they may wonder what is the extent of this risk? What factors might make a crash, injury, or fatality more likely? In this analysis, answers to these questions are explored with a dataset comprised of primarily time-related variables. With logistic regression and graphical analysis, the following questions will be answers:

1. How does peak crash time vary depending on type of vehicle?
2. What is the predicted number of car crashes that result in an injury or fatality based on the time of day and day of week?
3. Are certain days of the week and times of day associated with alcohol offenses?

Data Description: The dataset comes the New Zealand Ministry of Transport. The reports are derived from Traffic Crash Reports completed by police offers who attend the crashes. The outcome variable includes counts of crashes, deaths, and fatalities as well as counts of alcohol offenses committed at these scenes. While the original dataset includes a variety of vehicle types, this report focuses on cars and motorcycles only.

Statistical Modeling: A negative binomial model was chosen for both predictive and inferential purposes due to overdispersion present in a Poisson model. In the predictive model, car crashes (a combination of counts of crashes resulting injuries or fatalities) were predicted based on the hour of day (0-23) and day of the Week (Mon-Sun.). In the inferential model, the Hour and Day remained predictors with an outcome of alcohol offenses in place of crashes.

Results: Based on the results of predictive modeling, it was found that highest individual times predicted to have crashes were Saturday at 3 and 5 p.m. with a predicted count of crashes for the year equal to 163. Generally speaking, this window of time was predicted to have the most crashes leading injury or fatality for any day of the week. Another hour with a relatively high number of crashes was 8 a.m. Counts also tended to increase as the week progressed, peaking on Friday and Saturday. Weekends (Friday, Saturday, Sunday) were also associated with higher counts of alcohol offenses. This number tended to increase from 2 pm into early morning.

Conclusion: The first question of interest was addressed graphically and indicated similar patterns of high frequency of crashes (8 a.m. and around 4 p.m) for both cars and motorcycles; however, motorcycles had another peak not observed in car crashes, which was around noon on weekends. The second question is summarized in a table of predictions by hour of day and day of week, but with numbers reflective of the patterns indicated previously. Finally, it was found that Friday, Saturday, and Sunday were all statistically significant in explaining the likelihood of committing an alcohol offense, particularly in later hours of the day.

Analysis of Car and Motorcycle Crashes in New Zealand in 2009

Introduction

This report provides an analysis of road crashes in New Zealand over the year 2009, specifically limited to car and motorcycle crashes. Methods of analysis include logistic regression and graphical analysis used in order to answer the following questions:

1. How does peak crash time vary depending on type of vehicle?
2. What is the predicted number of car crashes that result in an injury or fatality based on the time of day and day of week?
3. Are certain days of the week and times of day associated with alcohol offenses?

The report finds the predicted number of car crashes that result in either injury or fatality are not practically different given day of week or time of day. Similarly, while car crashes are far more common than motorcycle crashes by count, analogous patterns emerge in what hours of the day crashes are most likely to occur for either type of vehicle. Finally, the report finds that an association exists between alcohol offenses and the day of the week on which they occurred.

Limitations of this report include a lack of comparison between all types of vehicles when examining crash time, though this could be added later depending on the vehicles of interest (for the purposes of this report motorcycles and cars are the vehicles of interest). Additionally, it is unclear in the alcohol offense data if all of the offenders were car drivers or if some may have been in another type of vehicle such as truck or motorcycle.

Data

All data used in this analysis came from the New Zealand Ministry of Transport. The reports are derived from Traffic Crash Reports that are completed by police officer who attend fatal and injury crashes. The datasets used for the analysis are aggregate numbers of crashes reported at each hour-day combination over the 2009 calendar year. The specific datasets used in this report include the aggregate number of injuries by car, number of fatalities by car, number of crashes involving motorcyclists, and the number of alcohol offenders from breath screening drivers.

These counts of the above information are a compilation of 4 data frames. Each of these data frames are 24 x 7 tables displayed in a count format – 24 for each hour of the day and 7 for each day of the week. At each intersection of a time of day and day of week, the count of injuries, fatalities, or offenses is recorded. For example, the first number recorded at the 0 hour (midnight) for Monday is 16, meaning that 16 car crash-related injuries occurred on Monday nights at midnight during 2009. Since this is well-recorded frequency data, there are no missing values in these datasets. There are zero counts in places, though these make sense as there were not fatalities at every day/hour combination over the year 2009 in New Zealand.

Analysis and Results

Question 1: How does peak crash time vary depending on vehicle – specifically, do motorcycles and cars have similar peak crash times?

In order to answer this question, we'll first explore data for each type of vehicle crash separately. Below is a plot of the number of car crash injuries in cars. The data was collected over the year 2009 and organized by hour of day (0 = midnight to 23 = 11 pm) and day of the week. Each day of the week is represented by a different colored line, and the count is recorded at each hour.

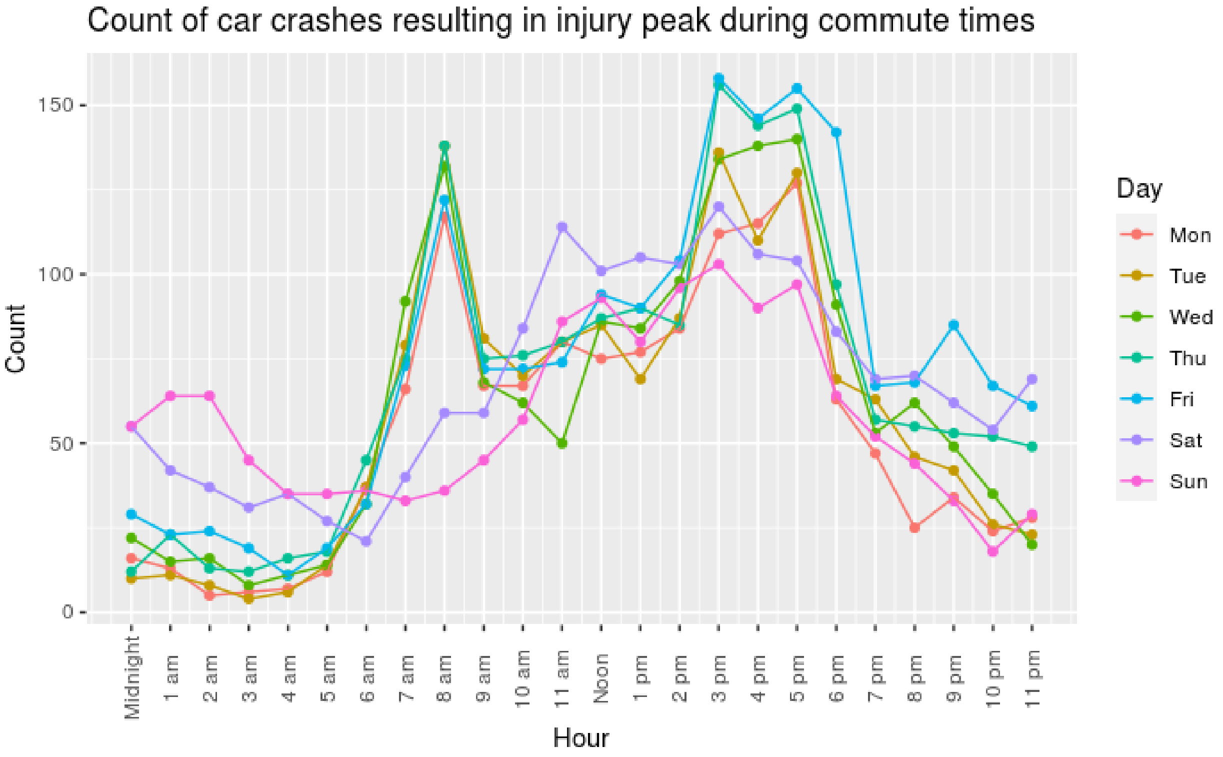


Figure 1. Count of car crashes resulting in injury for the year 2009.

The plot shows very similar trends for all of the weekdays (Monday-Friday) with some variation on the weekend (Saturday and Sunday). Broadly speaking, car crash injuries seem to rise and fall at similar hours throughout the day, regardless of the day of the week. These hours tend to align with the typical workday – around 8 a.m. when people are heading to work and between 3 and 5 p.m. when workers are returning home.

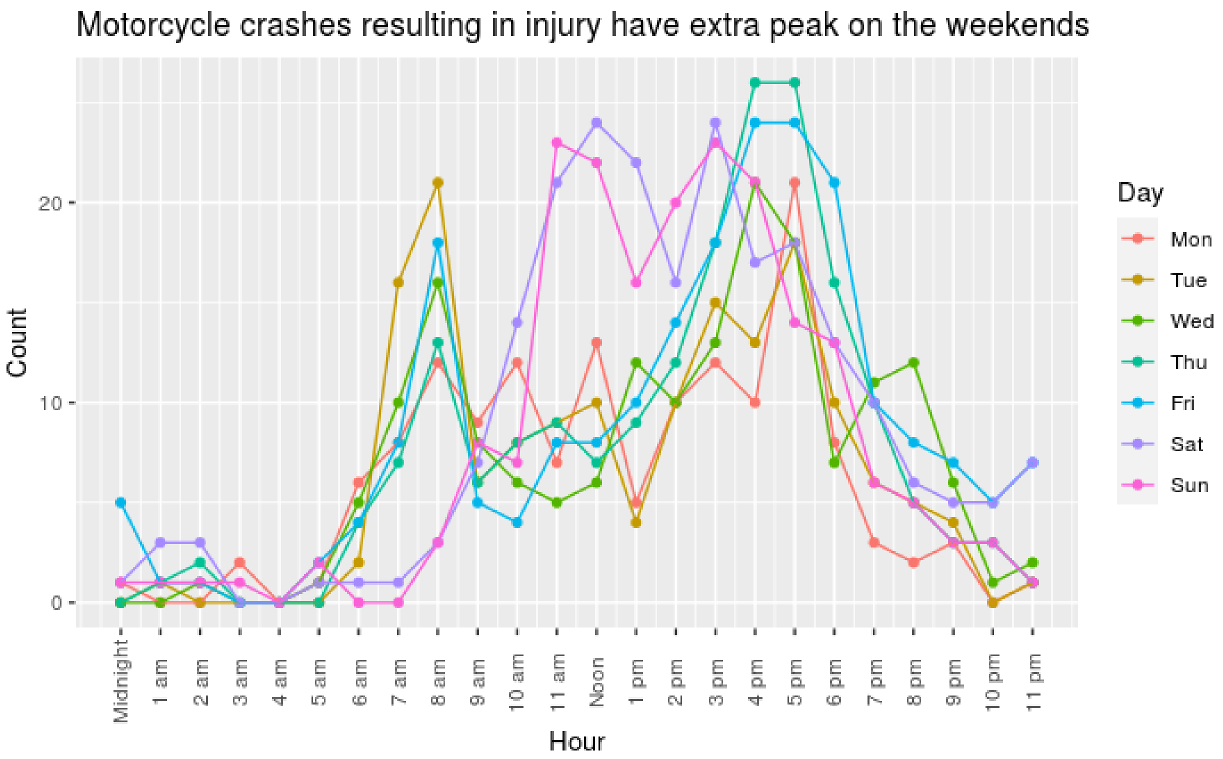


Figure 2. Count of motorcycle crashes resulting in injury over the year 2009.

Motorcycle crashes at first glance seem to have similar peak times, with more variation, particularly during midday between 10 a.m. and 3 p.m. This makes intuitive sense as the spikes during these times are on Saturday and Sunday when people are more likely riding their motorcycles leisurely. Like the car crash data, there seems to be more similarity between Monday-Friday with slightly different (but similar to each other) trends for Saturday and Sunday. Regardless of the day of the week, however, it seems as though the peak time for motorcycle injuries is between 3-5 p.m., which is the same peak time for car crash injuries.

Another way to look at this data is to aggregate it and look at the count of injuries at each hour averaged over all of the days of the week as seen in Figure 3.

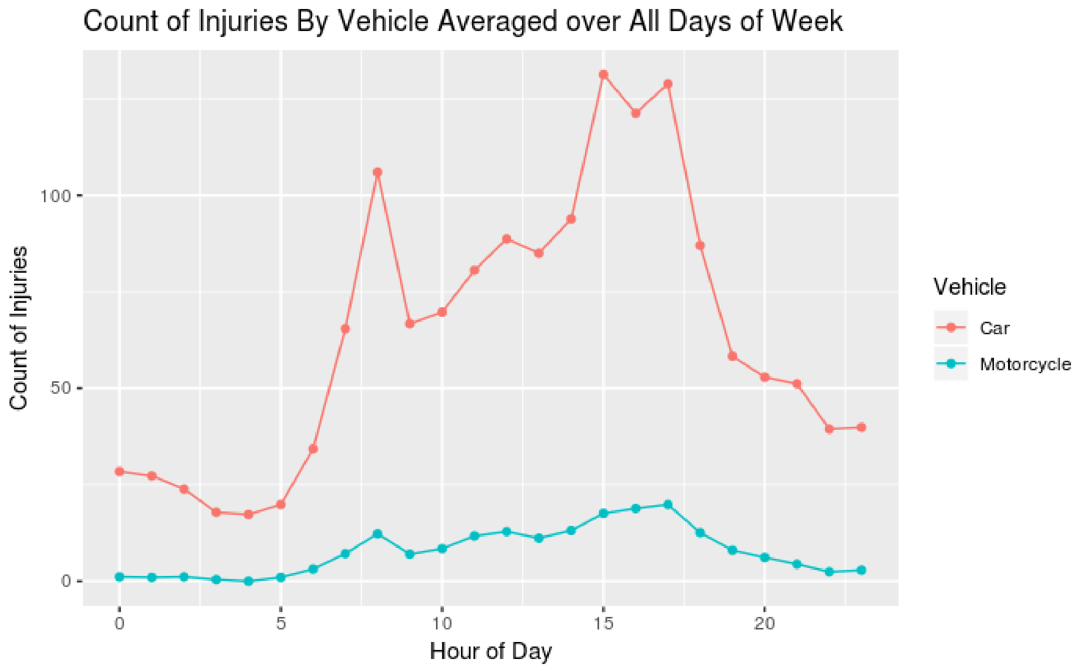


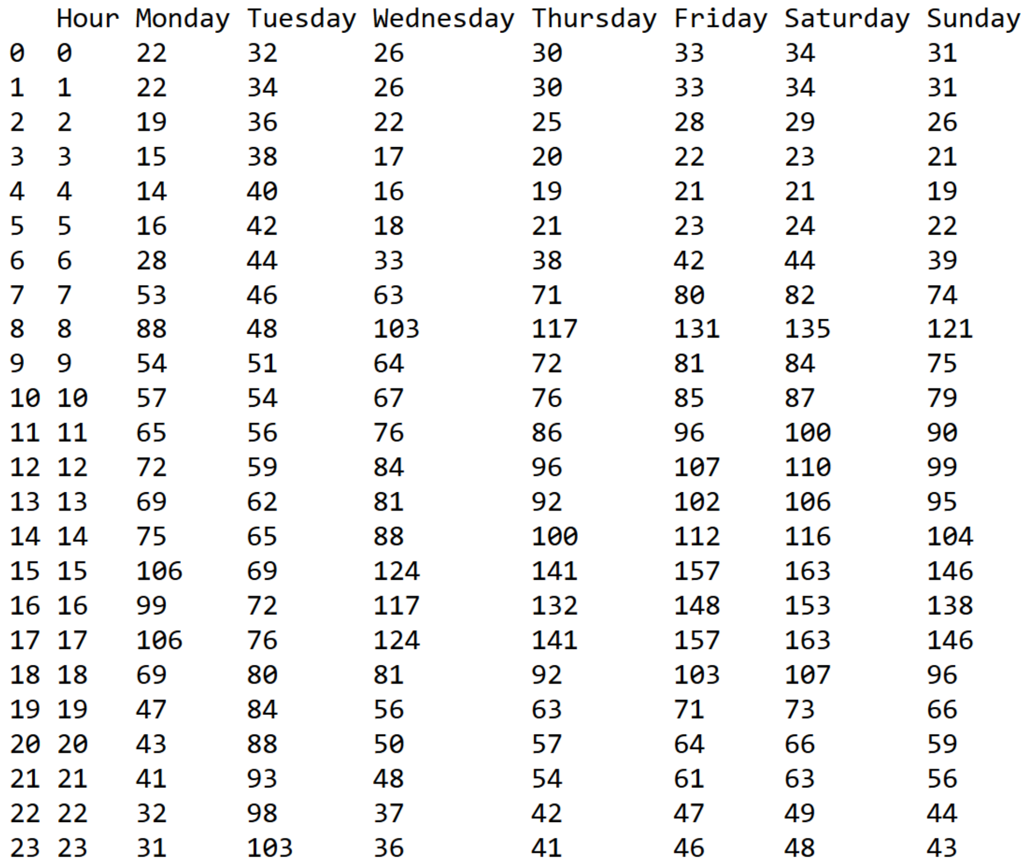
Figure 3. Aggregated counts of injuries by vehicle over the year 2009.

While it's clear the injury counts are generally much higher for cars, we can use the simplicity of this plot to see trends. If we look at the high points of each line, the maximum points are in the same places – around 8 a.m., and again between 3-5 p.m. Again, this makes intuitive sense as more people are on the road during these times going to and leaving work (this specifically applies to weekdays, but 5 of the 7 days are weekdays, so this trend will emerge in this view of the data).

Limitations of these descriptive plots include concerns about aggregation – we lose information and possibly gain misconceptions in this way. When comparing the last plot to the first two, however, the trends seem fairly consistent across the board. Perhaps a better way to answer this question would be to focus in on specific times or segments of time and examine more types of vehicles for comparison. In any case, it does seem apparent that there are similarities in peak accident time when comparing cars and motorcycles.

Question 2: What is the predicted number of car crashes that result in an injury or fatality based on the time of day and day of week?

In order to answer this question, a predictive model is needed. The first model fit was a Poisson regression model with day and hour as the explanatory variables (variables coded as is in the dataset – day as factor with 7 levels). Counts of injuries and fatalities were combined to create a "crashes" variable. This model resulted in strong evidence of overdispersion with an estimated dispersion parameter of 25.44. In order to address the overdispersion, a negative binomial model was fit. In this model, the overdispersion problem seems to be addressed with a new dispersion parameter of .4514 and a lower AIC of 1507.365 vs. 1846.772 for the Poisson model. The results of predicted number of crashes based on time of day and hour of day are summarized in the table below.



```
    Hour Monday Tuesday Wednesday Thursday Friday Saturday Sunday
0   0    22     32      26        30       33     34       31
1   1    22     34      26        30       33     34       31
2   2    19     36      22        25       28     29       26
3   3    15     38      17        20       22     23       21
4   4    14     40      16        19       21     21       19
5   5    16     42      18        21       23     24       22
6   6    28     44      33        38       42     44       39
7   7    53     46      63        71       80     82       74
8   8    88     48      103       117      131    135      121
9   9    54     51      64        72       81     84       75
10  10   57     54      67        76       85     87       79
11  11   65     56      76        86       96     100      90
12  12   72     59      84        96       107    110      99
13  13   69     62      81        92       102    106      95
14  14   75     65      88        100      112    116      104
15  15   106    69      124       141      157    163      146
16  16   99     72      117       132      148    153      138
17  17   106    76      124       141      157    163      146
18  18   69     80      81        92       103    107      96
19  19   47     84      56        63       71     73       66
20  20   43     88      50        57       64     66       59
21  21   41     93      48        54       61     63       56
22  22   32     98      37        42       47     49       44
23  23   31     103     36        41       46     48       43
```

Figure 4. Predicted number of car crashes resulting in injury or fatality based on hour of day and day of week (Hour 0 = Midnight).

Based on these findings, we see predictions that mirror the rise and fall of the lines in figures 1 and 2 with peaks during typical commute times (8 am and 3-6 pm). This trend continues for Saturday and Sunday as well despite these not being traditional work days.

Question 3: Are certain days of the week and times of day associated with alcohol offenses?

This question seeks to answer whether or not the time of day or day of week can explain the variation in the counts of alcohol offenses recorded. According to the New Zealand Police, alcohol and drug offense data includes all offenses relating to driving under the influence of alcohol and drugs. In 2014, this definition was expanded to include offenses committed by drivers aged under 20 who breach the zero breath and blood alcohol limit and those under the new lowered adult alcohol impairment limit; however, since this data is from 2009, it does not follow the most up-to-date New Zealand laws. Figure 5 gives an overall sense of when most offenses occurred during the day and during the week.

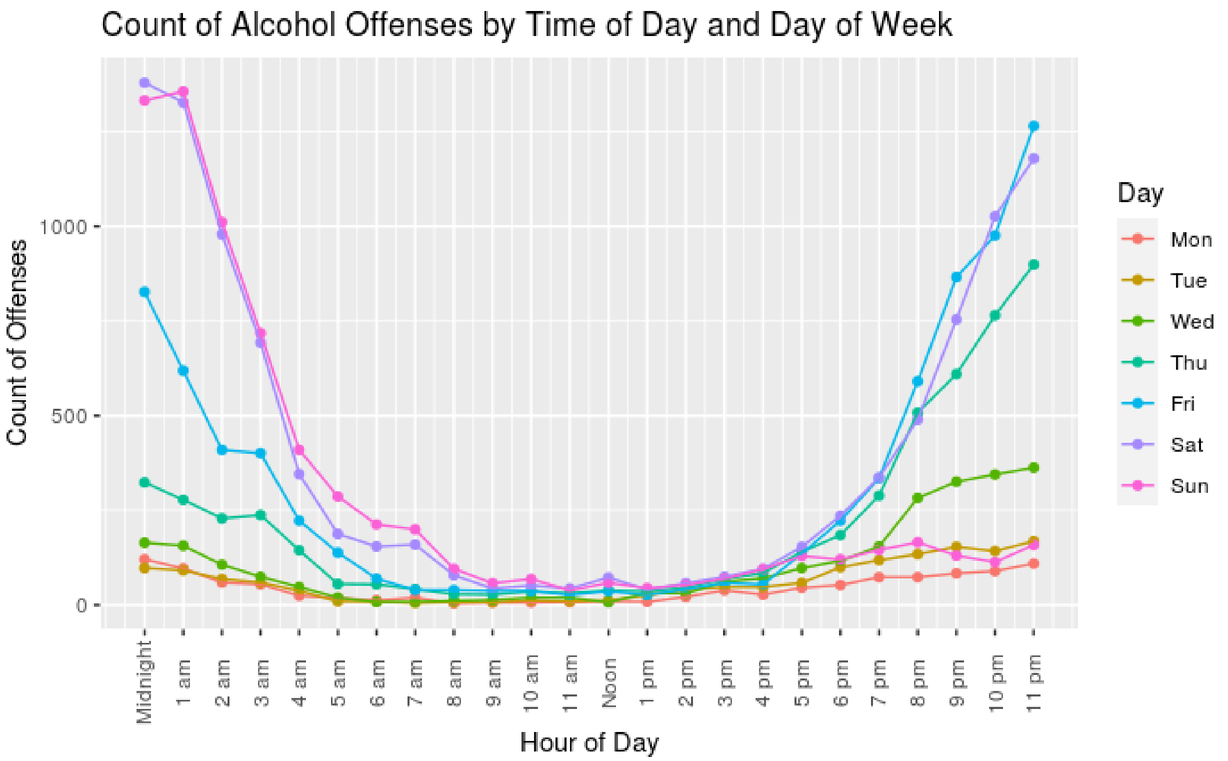


Figure 5. Count of alcohol offenses on the road over the year 2009.

The plot seems to indicate a “dead zone” during the work week from about 8-9 a.m. to about 2-3 p.m where there are several zero or near-zero counts. The weekends follow this same pattern during those times of day. These counts increase dramatically toward the edges of the graph – early morning and late evening, which makes sense given the time of day people generally drink. This effect is even more pronounced going into the weekend – counts are highest Thursday, Friday and Saturday late evening, and early Friday, Saturday, and Sunday.

To get a clearer idea of what levels of the day and time variables may be important, a negative binomial model with categorical variables for time and day were fit. A negative binomial model was chosen to model this count data over a Poisson model because of the significant overdispersion present (the estimated dispersion parameter for the Poisson model was 5.2978). An Akaike information criterion comparison confirms this choice (Poisson AIC = 8105.2, Negative Binomial AIC = 1784.7). In this model, Friday, Saturday, Sunday are all significant (p-value = 2e-16), which indicates strong evidence that more alcohol offense occur on the weekends as the coefficients for these three variables are all positive (1.307, 1.700, and 1.625, respectively). On the other hand, Monday has fewer alcohol offenses than the reference level Tuesday (p=.02), and Wednesday and Thursday have slightly more alcohol offenses with coefficient estimates of .393 and 1.06, respectively (p = .004 and 1.53e-15, respectively).

Most hours of the day were significant in this model as well. The reference was changed from midnight to noon to get a clearer idea of direction of coefficient signs. 1 pm was not significantly different from noon in terms of alcohol offenses (p=.58); however, hours from 2 pm – 7 am were statistically significant in that they all had more alcohol offenses; this peaked around 11 pm with a coefficient estimate of 2.9 (p-value = 2e-16).

Conclusions/Discussion

Through graphical analysis and logistic regression, the following questions were answered in this report:

1. How does peak crash time vary depending on type of vehicle?
2. What is the predicted number of car crashes that result in an injury or fatality based on the time of day and day of week?
3. Are certain days of the week and times of day associated with alcohol offenses?

Generally speaking, it was found that motorcycles and cars follow similar patterns in when during the day crashes occur, with an exception for motorcycles having an extra peak midday on the weekends. Further analysis could fold in data for different types of vehicles such as bicycles and trucks to see if they follow similar patterns.

A negative binomial model with Day and Hour as explanatory variables was found to be a good fit model for predicting the number of car crashes resulting in injuries or fatalities. These predictions mirrored the trends in the 2009 data, meaning that it is expected that weekday counts of crashes during 5 o'clock traffic may peak around 150. These predictions can help New Zealand emergency responders to staff appropriately.

It does appear that certain days of the week and times of day were statistically significant in explaining the number of alcohol offenses committed. As one might expect, weekend days (Friday, Saturday, and Sunday) were all associated with higher counts of alcohol offenses. This number also tended to increase from 2 pm into early morning.

A final caution regarding the prediction question is that this data came from the New Zealand Ministry of Transport in 2009. The predictions made from the model can only reasonably be applied to New Zealand roads. Furthermore, the data off which these predictions are based is over 10 years old. It is very likely there have been structural changes such as road updates or new driving laws that would affect the accuracy of our predictions.

Appendix

Data source:

"Motor Vehicle Crashes in New Zealand 2009." *Motor Vehicle Crashes in New Zealand 2009 | Ministry of Transport,* 9 Sept. 2013, www.transport.govt.nz/mot-resources/road-safety-resources/roadcrashstatistics/motorvehiclecrashesinnewzealand/2009/.

Code:

Start by loading packages

Load crash data

Combine and re-organize data to make it long rather than wide: #Organize crashi

```
##   Count Day Hour   Crash_type
## 1    16 Mon    0 Car Injuries
## 2    13 Mon    1 Car Injuries
## 3     5 Mon    2 Car Injuries
## 4     6 Mon    3 Car Injuries
## 5     7 Mon    4 Car Injuries
## 6    12 Mon    5 Car Injuries
```

#Organize crashf

#Organize crashmc

#Organize alcoff

```
hour <- rownames(alcoff) ## grab the hours
alcoff2 <- stack(alcoff) ## combine 7 columns of crashes into 1
names(alcoff2) <- c("Count","Day")
alcoff2$Day <- factor(alcoff2$Day,levels(alcoff2$Day)[c(2,6,7,5,1,3,4)
])  # make sure the days are ordered correctly
alcoff2$Hour <- as.numeric(rep(hour, ncol(alcoff)))
## add a column with hour and make it numeric (not categorical)
#add column for dataset origin
alcoff2 %<>% mutate(.,Crash_type = "Alcohol Offenses")
#head(alcoff2)
```

#Combine datasets

```
crash_new <- rbind(crashi2, crashf2, crashmc2, alcoff2)
#View(crash_new)
```

Questions of interest: 1) Is there a difference in peak crash time depending on type of vehicle? 2) Do car crash fatalities follow the same trends as car crash injuries? Is there a different peak time? 3) Is there a connection between alcohol offenses and car crash fatalities?

Q of I take 2: 1) How does peak crash time vary depending on type of vehicle? 2) What is the probability of a car crash-related injury or fatality based on time of day and day of week (predictive) 3) What variables are associated with alcohol offenses? (explanatory)

Explorations: #Plot count of crashes (Y) vs. Hour of day (X) by Day of Week (color) for Car Injuries #Plot count of crashes (Y) vs. Hour of day (X) by Day of Week (color) for Motorcycle Injuries

#Plot count of crashes (Y) vs. Hour of day (X) by Day of Week (color) for Alcohol Offenses

#Question 1 # Plot average car crash injuries vs. motorcycle injuries First, subset and take average for each hour over all days of the week - car injuries

Subset and take average for each hour over all days of the week - motorcycle injuries

Make new mini-data frame with motorcycle and car injuries

Plot curves of averages against each other - peaks appear to be in similar places - around 8 am and then again between 3-5 pm, which logically makes sense given typical traffic hours during the workweek. This plot includes all 7 days, but 5 of those are work days.

#Question 2 - predictive model for car crashes (injuries and fatalities)

```
## # A tibble: 168 x 4
##    Day    Hour `Car Injuries` `Car Fatalities`
##    <fct> <dbl>          <int>            <int>
##  1 Mon       0             16                1
##  2 Mon       1             13                0
##  3 Mon       2              5                1
##  4 Mon       3              6                1
##  5 Mon       4              7                1
##  6 Mon       5             12                0
##  7 Mon       6             37                1
##  8 Mon       7             66                1
##  9 Mon       8            117                2
## 10 Mon       9             67                1
## # … with 158 more rows

## [1] TRUE

##
## Call:
## glm(formula = crashes ~ Day + Hour, family = "poisson", data = car_
combo_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.0762  -1.4189  -0.1115   1.1490   7.3228
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.224e+00  7.392e-02  43.619  < 2e-16 ***
## DayTue       8.720e-02  3.783e-02   2.305 0.021175 *
## DayWed       1.463e-01  3.731e-02   3.922 8.78e-05 ***
## DayThu       2.469e-01  3.647e-02   6.770 1.29e-11 ***
## DayFri       3.303e-01  3.583e-02   9.218  < 2e-16 ***
## DaySat       2.463e-01  3.648e-02   6.753 1.45e-11 ***
## DaySun       7.272e-02  3.796e-02   1.915 0.055429 .
## Hour1       -3.114e-15  9.806e-02   0.000 1.000000
## Hour2       -1.671e-01  1.024e-01  -1.631 0.102870
## Hour3       -4.031e-01  1.096e-01  -3.679 0.000234 ***
## Hour4       -4.777e-01  1.121e-01  -4.263 2.02e-05 ***
## Hour5       -3.747e-01  1.086e-01  -3.449 0.000562 ***
## Hour6        1.759e-01  9.402e-02   1.871 0.061378 .
## Hour7        8.002e-01  8.347e-02   9.586  < 2e-16 ***
## Hour8        1.289e+00  7.831e-02  16.463  < 2e-16 ***
## Hour9        8.258e-01  8.314e-02   9.932  < 2e-16 ***
## Hour10       8.850e-01  8.241e-02  10.739  < 2e-16 ***
## Hour11       1.024e+00  8.084e-02  12.664  < 2e-16 ***
## Hour12       1.124e+00  7.981e-02  14.082  < 2e-16 ***
## Hour13       1.082e+00  8.023e-02  13.492  < 2e-16 ***
## Hour14       1.170e+00  7.937e-02  14.737  < 2e-16 ***
## Hour15       1.504e+00  7.666e-02  19.621  < 2e-16 ***
## Hour16       1.440e+00  7.711e-02  18.675  < 2e-16 ***
## Hour17       1.497e+00  7.671e-02  19.510  < 2e-16 ***
## Hour18       1.102e+00  8.003e-02  13.767  < 2e-16 ***
## Hour19       7.146e-01  8.462e-02   8.444  < 2e-16 ***
## Hour20       6.260e-01  8.590e-02   7.288 3.14e-13 ***
## Hour21       5.733e-01  8.670e-02   6.612 3.80e-11 ***
## Hour22       3.323e-01  9.086e-02   3.658 0.000255 ***
## Hour23       3.219e-01  9.106e-02   3.535 0.000407 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4248.75  on 167  degrees of freedom
## Residual deviance:  815.18  on 138  degrees of freedom
## AIC: 1846.8
##
## Number of Fisher Scoring iterations: 5
```

Overdispersion is present (overdispersion parameter = 49.027) - try negative binomial instead

```
##
## Call:
## glm.nb(formula = crashes ~ Day + Hour, data = car_combo_df, init.th
eta = 11.93907896,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1925  -0.6752  -0.0695   0.5210   3.2705
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.108180   0.144841  21.459  < 2e-16 ***
## DayTue       0.078591   0.095060   0.827 0.408374
## DayWed       0.160479   0.094677   1.695 0.090071 .
## DayThu       0.287555   0.094134   3.055 0.002252 **
## DayFri       0.398243   0.093707   4.250 2.14e-05 ***
## DaySat       0.432291   0.093584   4.619 3.85e-06 ***
## DaySun       0.325271   0.093984   3.461 0.000538 ***
## Hour1       -0.002267   0.184467  -0.012 0.990196    
## Hour2       -0.173532   0.186995  -0.928 0.353406    
## Hour3       -0.409299   0.191183  -2.141 0.032285 *  
## Hour4       -0.474515   0.192507  -2.465 0.013704 *  
## Hour5       -0.353641   0.190113  -1.860 0.062863 .  
## Hour6        0.234734   0.181565   1.293 0.196069    
## Hour7        0.871094   0.176329   4.940 7.81e-07 ***
## Hour8        1.366412   0.173973   7.854 4.02e-15 ***
## Hour9        0.884718   0.176248   5.020 5.17e-07 ***
## Hour10       0.930726   0.175982   5.289 1.23e-07 ***
## Hour11       1.062991   0.175280   6.065 1.32e-09 ***
## Hour12       1.164431   0.174800   6.662 2.71e-11 ***
## Hour13       1.121808   0.174996   6.410 1.45e-10 ***
## Hour14       1.212635   0.174587   6.946 3.77e-12 ***
## Hour15       1.551390   0.173346   8.950  < 2e-16 ***
## Hour16       1.490546   0.173540   8.589  < 2e-16 ***
## Hour17       1.551740   0.173345   8.952  < 2e-16 ***
## Hour18       1.130785   0.174954   6.463 1.02e-10 ***
## Hour19       0.752380   0.177081   4.249 2.15e-05 ***
## Hour20       0.649025   0.177810   3.650 0.000262 ***
## Hour21       0.599342   0.178187   3.364 0.000769 ***
## Hour22       0.345192   0.180417   1.913 0.055709 .  
## Hour23       0.326589   0.180602   1.808 0.070555 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Negative Binomial(11.9391) family taken t
o be 1)
## 
##     Null deviance: 772.69  on 167  degrees of freedom
## Residual deviance: 182.38  on 138  degrees of freedom
## AIC: 1507.4
## 
## Number of Fisher Scoring iterations: 1
## 
## 
##               Theta:  11.94 
##           Std. Err.:  1.72 
## 
##  2 x log-likelihood:  -1445.365
```

Negative binomial seems to be a better fit - check AIC

```
##           df      AIC
## mod_if_1  30 1846.772
## mod_if_nb 31 1507.365
```

—Proceed with mod_if_nb for prediction—

```
##   Day Hour
## 1 Mon   23

## $fit
##        1
## 31.02422
##
## $se.fit
##        1
## 4.338919
##
## $residual.scale
## [1] 1
```

```
pred_table <- cbind(hours, mon_preds, tue_preds, wed_preds, thu_preds,
fri_preds, sat_preds, sun_preds)
colnames(pred_table) <- c("Hour", "Monday", "Tuesday", "Wednesday", "T
hursday", "Friday", "Saturday", "Sunday")
pred_table

##    Hour Monday Tuesday Wednesday Thursday Friday Saturday Sunday
## 0  0    22     24      26        30       33     34       31
## 1  1    22     24      26        30       33     34       31
## 2  2    19     20      22        25       28     29       26
## 3  3    15     16      17        20       22     23       21
## 4  4    14     15      16        19       21     21       19
## 5  5    16     17      18        21       23     24       22
## 6  6    28     31      33        38       42     44       39
## 7  7    53     58      63        71       80     82       74
## 8  8    88     95      103       117      131    135      121
## 9  9    54     59      64        72       81     84       75
## 10 10   57     61      67        76       85     87       79
## 11 11   65     70      76        86       96     100      90
## 12 12   72     78      84        96       107    110      99
## 13 13   69     74      81        92       102    106      95
## 14 14   75     81      88        100      112    116      104
## 15 15   106    114     124       141      157    163      146
## 16 16   99     107     117       132      148    153      138
## 17 17   106    114     124       141      157    163      146
## 18 18   69     75      81         92        103    107       96
## 19 19   47     51      56         63        71     73        66
## 20 20   43     46      50         57        64     66        59
## 21 21   41     44      48         54        61     63        56
## 22 22   32     34      37         42        47     49        44
## 23 23   31     34      36         41        46     48        43
```

—data with day and time categories—

Try different variables - create copy of data and change variables = crash_new2

```
##   Count Day Hour   Crash_type Day.Cat   Time.Cat
## 1    16 Mon    0 Car Injuries Weekday Early.Morn
## 2    13 Mon    1 Car Injuries Weekday Early.Morn
## 3     5 Mon    2 Car Injuries Weekday Early.Morn
## 4     6 Mon    3 Car Injuries Weekday Early.Morn
## 5     7 Mon    4 Car Injuries Weekday Early.Morn
## 6    12 Mon    5 Car Injuries Weekday Early.Morn
```

Fit a negative binomial model with new categories as explanatory variables - AIC is slightly lower than previous NB model at 2834.9. These variables also might be more useful in a predictive capacity because if you're trying to figure out the probability of getting in a crash at a certain time, you're likely interested in the time of day (i.e. in the morning on the way to work) rather than a specific hour.

```
## 
## Call:
## glm.nb(formula = Count ~ Day.Cat + Time.Cat, data = subset(crash_ne
w2, 
##     Crash_type == "Car Injuries" | Crash_type == "Car Fatalities"), 
##     init.theta = 0.4906589171, link = log)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.1238  -1.3193  -0.4732   0.4507   1.5108  
## 
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         2.56989    0.24964  10.294  < 2e-16 ***
## Day.CatSat          0.05727    0.29394   0.195 0.845516    
## Day.CatSun         -0.01917    0.29403  -0.065 0.948028    
## Day.CatWeekday     -0.27408    0.23256  -1.179 0.238584    
## Time.CatMorn        1.18855    0.22343   5.320 1.04e-07 ***
## Time.CatAfternoon   1.57841    0.21522   7.334 2.23e-13 ***
## Time.CatEvening     0.79209    0.23458   3.377 0.000734 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for Negative Binomial(0.4907) family taken to
be 1)
##
##     Null deviance: 448.76  on 335  degrees of freedom
## Residual deviance: 398.76  on 329  degrees of freedom
## AIC: 2834.9
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.4907
##           Std. Err.:  0.0353
##
##  2 x log-likelihood:  -2818.9150
```

We'll also look at interaction terms to see if we can improve fit.

```
##
## Call:
## glm.nb(formula = Count ~ Day.Cat * Time.Cat, data = subset(crash_ne
w2,
##     Crash_type == "Car Injuries" | Crash_type == "Car Fatalities"),
##     init.theta = 0.5075152516, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1762  -1.3976  -0.2932   0.5105   1.1446
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z
|)
## (Intercept)                       2.43507    0.41412   5.880  4.1e-
09 ***
## Day.CatSat                        0.56066    0.58298   0.962  0.336
19
## Day.CatSun                        0.81981    0.58216   1.408  0.159
07
## Day.CatWeekday                   -0.55370    0.46446  -1.192  0.233
20
## Time.CatMorn                      1.19812    0.58129   2.061  0.039
29 *
## Time.CatAfternoon                 1.74041    0.55977   3.109  0.001
88 **
## Time.CatEvening                   1.15122    0.60935   1.889  0.058
86 .
```

```
## Day.CatSat:Time.CatMorn          -0.73333    0.82042  -0.894  0.371
40
## Day.CatSun:Time.CatMorn          -1.23080    0.82026  -1.500  0.133
49
## Day.CatWeekday:Time.CatMorn       0.59205    0.65092   0.910  0.363
06
## Day.CatSat:Time.CatAfternoon     -0.75780    0.78980  -0.959  0.337
32
## Day.CatSun:Time.CatAfternoon     -1.16821    0.78933  -1.480  0.138
87
## Day.CatWeekday:Time.CatAfternoon  0.35623    0.62696   0.568  0.569
91
## Day.CatSat:Time.CatEvening       -0.61181    0.86001  -0.711  0.476
84
## Day.CatSun:Time.CatEvening       -1.48833    0.86091  -1.729  0.083
85 .
## Day.CatWeekday:Time.CatEvening    0.04817    0.68260   0.071  0.943
74
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.5075) family taken to
be 1)
##
##     Null deviance: 462.61  on 335  degrees of freedom
## Residual deviance: 397.32  on 320  degrees of freedom
## AIC: 2839.5
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.5075
##           Std. Err.:  0.0367
##
##  2 x log-likelihood:  -2805.5280
```

In this case, it seems the model without interaction terms fits slightly better based on AIC. Let's check with a Drop in Deviance test.

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: Count
##                 Model     theta Resid. df    2 x log-lik.   Test
df
## 1 Day.Cat + Time.Cat 0.4906589       329       -2818.915
## 2 Day.Cat * Time.Cat 0.5075153       320       -2805.528 1 vs 2
```

```
9
##   LR stat.   Pr(Chi)
## 1
## 2 13.38669 0.1458754
```

This also confirms that we should proceed with the reduced model. Let’s use this model to make a specific prediction - say we want to know how many car crashes that result in an injury or fatality happen on weekday mornings.

```
## $fit
##        1
## 32.60134
##
## $se.fit
##        1
## 5.575661
##
## $residual.scale
## [1] 1
```

We want to know if this is more or less than the number of crashes resulting in an injury or fatality Saturday night.

```
## $fit
##        1
## 30.54667
##
## $se.fit
##        1
## 7.888181
##
## $residual.scale
## [1] 1
```

#Question 3 - what explanatory variables are associated with alcohol offenses? Look at alcohol offense subset

Fit logisitic regression model to see if we can understand this relative to the explanatory information:

```
##
## Call:
## glm(formula = Count ~ Day + Hour, family = "poisson", data = alc_su
b)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -28.9425   -3.0310   -0.2297    3.4271   17.9126
```

```
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  4.87762    0.03371 144.708  < 2e-16 ***
## DayTue       0.31609    0.04009   7.885 3.14e-15 ***
## DayWed       0.86167    0.03636  23.699  < 2e-16 ***
## DayThu       1.56865    0.03351  46.810  < 2e-16 ***
## DayFri       1.93952    0.03260  59.488  < 2e-16 ***
## DaySat       2.22583    0.03209  69.364  < 2e-16 ***
## DaySun       1.88374    0.03272  57.570  < 2e-16 ***
## Hour1       -0.07836    0.02214  -3.539 0.000402 ***
## Hour2       -0.39341    0.02418 -16.272  < 2e-16 ***
## Hour3       -0.63950    0.02611 -24.489  < 2e-16 ***
## Hour4       -1.23491    0.03233 -38.197  < 2e-16 ***
## Hour5       -1.77587    0.04033 -44.036  < 2e-16 ***
## Hour6       -2.09224    0.04630 -45.186  < 2e-16 ***
## Hour7       -2.18622    0.04829 -45.272  < 2e-16 ***
## Hour8       -2.75902    0.06287 -43.883  < 2e-16 ***
## Hour9       -3.07053    0.07288 -42.131  < 2e-16 ***
## Hour10      -2.89415    0.06701 -43.187  < 2e-16 ***
## Hour11      -3.16078    0.07610 -41.535  < 2e-16 ***
## Hour12      -2.87727    0.06648 -43.280  < 2e-16 ***
## Hour13      -2.97384    0.06960 -42.729  < 2e-16 ***
## Hour14      -2.67356    0.06040 -44.262  < 2e-16 ***
## Hour15      -2.30873    0.05104 -45.233  < 2e-16 ***
## Hour16      -2.18412    0.04824 -45.272  < 2e-16 ***
## Hour17      -1.71516    0.03930 -43.644  < 2e-16 ***
## Hour18      -1.41158    0.03467 -40.720  < 2e-16 ***
## Hour19      -1.07097    0.03038 -35.255  < 2e-16 ***
## Hour20      -0.63638    0.02609 -24.394  < 2e-16 ***
## Hour21      -0.37268    0.02403 -15.510  < 2e-16 ***
## Hour22      -0.20500    0.02290  -8.950  < 2e-16 ***
## Hour23      -0.02432    0.02184  -1.114 0.265474    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 52476.1  on 167  degrees of freedom
## Residual deviance:  6985.5  on 138  degrees of freedom
## AIC: 8105.2
## 
## Number of Fisher Scoring iterations: 5
```

It seems like day of the week is a significant variable here but not hour. I'm wondering if hour might be better off as a category like we modeled it in lab (early morning, morning, afternoon, evening) because based on the graph it seems like time of day might be important. There also seems to be evidence of overdipsersion, so this may not be the best model fit. I'll try a negative binomial model with categories and a negative binomial model without categories.

```
##
## Call:
## glm.nb(formula = Count ~ Day + Hour, data = alc_sub, init.theta = 5
.297796654,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5287  -0.8081  -0.0640   0.5647   1.8124
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.121443   0.205667  10.315  < 2e-16 ***
## DayTue       0.320955   0.137904   2.327 0.019945 *
## DayWed       0.713759   0.136378   5.234 1.66e-07 ***
## DayThu       1.381179   0.134723  10.252  < 2e-16 ***
## DayFri       1.627506   0.134324  12.116  < 2e-16 ***
## DaySat       2.020022   0.133851  15.092  < 2e-16 ***
## DaySun       1.945598   0.133927  14.527  < 2e-16 ***
## Hour13       0.141274   0.256974   0.550 0.582485
## Hour14       0.511161   0.253749   2.014 0.043964 *
## Hour15       0.928300   0.251165   3.696 0.000219 ***
## Hour16       0.960551   0.251002   3.827 0.000130 ***
## Hour17       1.392277   0.249230   5.586 2.32e-08 ***
## Hour18       1.697594   0.248344   6.836 8.16e-12 ***
## Hour19       2.014564   0.247658   8.134 4.14e-16 ***
## Hour20       2.397881   0.247064   9.705  < 2e-16 ***
## Hour21       2.614320   0.246815  10.592  < 2e-16 ***
## Hour22       2.734989   0.246697  11.086  < 2e-16 ***
## Hour23       2.903313   0.246554  11.776  < 2e-16 ***
## Hour0        2.726557   0.246705  11.052  < 2e-16 ***
## Hour1        2.623048   0.246806  10.628  < 2e-16 ***
## Hour2        2.292340   0.247207   9.273  < 2e-16 ***
## Hour3        2.087506   0.247527   8.433  < 2e-16 ***
## Hour4        1.506773   0.248868   6.055 1.41e-09 ***
## Hour5        0.918874   0.251213   3.658 0.000254 ***
## Hour6        0.591930   0.253172   2.338 0.019384 *
## Hour7        0.523793   0.253656   2.065 0.038925 *
## Hour8       -0.002241   0.258523  -0.009 0.993084
## Hour9       -0.184343   0.260774  -0.707 0.479624
## Hour10       0.019210   0.258280   0.074 0.940712
## Hour11      -0.180160   0.260719  -0.691 0.489557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(5.2978) family taken to
be 1)
##
##     Null deviance: 1458.60  on 167  degrees of freedom
## Residual deviance:  168.67  on 138  degrees of freedom
## AIC: 1784.7
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  5.298
##           Std. Err.:  0.604
##
##  2 x log-likelihood:  -1722.712
```

Negative binomial model with unaltered variables does not converge. Time still not significant.

```
##
## Call:
## glm.nb(formula = Count ~ Day.Cat + Time.Cat, data = subset(crash_ne
w2,
##     Crash_type == "Alcohol Offenses"), init.theta = 1.898504918,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5579  -0.9719  -0.3907   0.4566   2.5564
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         5.9384     0.1783  33.311  < 2e-16 ***
## Day.CatSat          0.4200     0.2109   1.991  0.04644 *
## Day.CatSun          0.3232     0.2110   1.532  0.12547
## Day.CatWeekday     -0.8999     0.1673  -5.378 7.55e-08 ***
## Time.CatMorn       -2.0088     0.1612 -12.458  < 2e-16 ***
## Time.CatAfternoon  -1.2555     0.1540  -8.150 3.63e-16 ***
## Time.CatEvening     0.4601     0.1667   2.761  0.00577 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.8985) family taken to
be 1)
##
```

```
##     Null deviance: 541.62  on 167  degrees of freedom
## Residual deviance: 180.52  on 161  degrees of freedom
## AIC: 1920.3
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.899
##           Std. Err.:  0.197
##
##  2 x log-likelihood:  -1904.316
```

Fit a model with interaction terms

```
##
## Call:
## glm.nb(formula = Count ~ Day.Cat * Time.Cat, data = subset(crash_ne
w2,
##     Crash_type == "Alcohol Offenses"), init.theta = 2.512782625,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6294  -0.9489  -0.2603   0.4941   2.3306
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z
|)
## (Intercept)                        6.0788     0.2583  23.536  < 2e-
16 ***
## Day.CatSat                         0.6289     0.3650   1.723   0.08
49 .
## Day.CatSun                         0.6692     0.3650   1.833   0.06
68 .
## Day.CatWeekday                    -1.3990     0.2893  -4.836 1.32e-
06 ***
## Time.CatMorn                      -2.3372     0.3701  -6.315 2.71e-
10 ***
## Time.CatAfternoon                 -1.6634     0.3540  -4.699 2.61e-
06 ***
## Time.CatEvening                    0.6140     0.3828   1.604   0.10
87
## Day.CatSat:Time.CatMorn            0.1181     0.5213   0.227   0.82
07
## Day.CatSun:Time.CatMorn            0.3121     0.5209   0.599   0.54
90
```

```
## Day.CatWeekday:Time.CatMorn        0.5195     0.4158   1.249   0.21
15
## Day.CatSat:Time.CatAfternoon      -0.3930     0.5000  -0.786   0.43
18
## Day.CatSun:Time.CatAfternoon      -0.6796     0.5004  -1.358   0.17
44
## Day.CatWeekday:Time.CatAfternoon   1.0057     0.3964   2.537   0.01
12 *
## Day.CatSat:Time.CatEvening        -0.6921     0.5412  -1.279   0.20
10
## Day.CatSun:Time.CatEvening        -2.3978     0.5423  -4.422 9.79e-
06 ***
## Day.CatWeekday:Time.CatEvening     0.3578     0.4285   0.835   0.40
37
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.5128) family taken to
be 1)
##
##     Null deviance: 712.04  on 167  degrees of freedom
## Residual deviance: 177.75  on 152  degrees of freedom
## AIC: 1887.2
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.513
##           Std. Err.:  0.270
##
##  2 x log-likelihood:  -1853.248
```