Modeling and Forecasting US Airline Flight Delays
2020 TAMIDS Data Science Competition

Team Big Data Energy
Johnathan Lo & Isaac Ke
Advisor: Dr. Huiyan Sang

April 8, 2020

Contents

Chapter 1

Introduction

Reliable transportation supports a strong economy by facilitating the rapid and timely exchange of goods and services and bolstering tourism revenue. In 2018, the United States transportation industry accounted for $648 billion, which was 3.16% of the GDP [Duffin (2019)]. Worldwide, the aviation industry contributes $2.7 trillion (3.6%) of the world's GDP. In fact, it is projected that by 2036, global air transportation will support $5.7 trillion of the global economy [Borders]. A key metric for evaluating the efficiency of airline industry production is flight delay time. In 2018, flight delays led to an economic loss of 31.2 billion dollars [America]. For individual companies, delays can influence consumer choice, and for the industry itself, unmitigated delays can impel consumers to switch to substitute services such as automotive or rail-based transport.

Therefore, a major goal of this project was to analyze flight delays and diagnose areas for improvement. We intended to create models using the provided datasets as well as other publicly available data that can accurately predict future delays. In doing so, our hope was to uncover significant and controllable factors that can help guide airline companies to reduce flight delays.

Chapter 2

Executive Summary

2.1 Problem and Approach

Ever since the first commercial airline flight was flown in 1914, the air transportation industry has played an integral part in both boosting the global economy as well as connecting people from all over the world. On occasion, flights can be delayed from their scheduled departure or arrival times, and this results in lost revenue and irritated customers. The goal of our analysis was to not only model flight delays but also create predictive models to forecast future late arrivals, specifically for the third quarter of 2019. We dove into this project by first gathering and tidying up our data, performing exploratory data analysis, then fitting and assessing unique predictive models. By gathering substantive knowledge about airlines, we then interpreted and applied our results to creating a model. Interpretability and thoroughness were the driving forces in our analysis.

2.2 Data Preprocessing

In addition to the datasets provided to us, we worked to gather historical weather information for all 11 million flights at both the origin and destination airport by fetching pertinent data from the NCDC (National Climate Data Center) API. After all data was gathered, we ensured our data was "tidy" by combining all flight, weather, airfare, route, airport, and geographical information into one data frame with each observation as a row and covariate as a column. Moreover, we changed appropriate independent variables to categorical variables, replaced all missing values with reasonable entries, and experimented with transformations of variables. For subsequent models, data were further reorganized to fit the nature of the problem. For example, for one of our models we formatted the data to resemble a time-series.

2.3 Exploratory Analysis

During the exploratory phase, we produced various plots and summary statistics to learn about the distribution and nature of flight delays. By using numerous packages in R along with connection to the Google maps API, we were able to produce many revealing plots. Among these, we were able to assess the geographic, temporal, carrier, and weather-based patterns of flight delays, to name a few. Through the construction of conditional density estimates and other tests, the main takeaway from our exploration was that flight delays are not a purely random event and that they behave very predictably across a myriad of factors. This motivated us to pursue the models that we did.

2.4 Model Creation and Assessment

Using the empirical distributions found during our exploratory analysis, we created a parametric distribution for the marginal values of the response. A parametric distribution was useful here because it provided interpretability and allowed regression and Bayesian methods to be used on the parameters. A key difficulty here was the skewness of the distribution, which we could not correct through transformation methods such as Box-Cox. Attempts to fit other well-known distributions were underwhelming. Ultimately, we created out own mixture distribution, $Y = UV + (1 - U)T$ where $U \sim Ber(p)$, $V \sim exp(\lambda)$, and $T \sim N(\mu, \sigma^2)$. Our parametric distribution has great descriptive utility in interpreting results. By using QQ plots and conditional density estimation, we were able to validate this parametric distribution for the majority of factor levels. Importantly, we discovered conditional densities have the same approximate distribution across factors. This distribution is henceforth referred to as the Lo-Ke distribution, named after the team members who created it.

Having the form of the conditional distributions allowed us to construct linear models using OLS (ordinary least squares). Observed non-normality of the errors prohibited conventional inference procedures; however, the assumptions for Gauss-Markov were not violated. To allow for prediction, we described the errors with the Lo-Ke distribution. Since the distribution is not location-scale parameterized, we instead approximated predictions by keeping other parameters constant while varying p.

Next, as we found that modeling the binary variable of whether or not a delay would occur was beneficial, we trained a logistic regression model. By splitting our data into training and testing sets, we regressed on the now-binary arrival delay response on a variety of other pre-departure factors such as weather conditions. Along with critically assessing model output, we generated a Receiver Operating Characteristics curve as well as a confusion matrix to further optimize our model and maximize its accuracy.

Lastly, we re-visited our linear model. We hypothesized that the observed pattern in the residuals might be the result of autocorrelation, or some type of time-dependency. To deal with these effects, we fitted a dynamic linear model by creating a bootstrap time series and fitting a multi-season ARIMA model (auto-regressive integrated moving average). We then fit a linear model using OLS to the residuals to account for the "left-over" variation that was unaccounted for in the time-series. Our final model was an additive model with one component being the output from the ARIMA time-based model and the other from the linear regression on the residuals of this ARIMA fit (to ensure minimal variation is left unexplained by our data).

2.5 Applications and Conclusions

The models we chose showed significant effects on arrival time for several different predictors. The most significant and predictable of these were seasonal effects by quarter and time of day. Flights during the 2nd and 3rd quarters were both significantly and substantially later than flights during the 1st and 4th. Time of day showed a clear effect with lower delays in the early morning, and delays peaking for flights departing around 5 PM. Day-of-week effects were also observed, with Thursday and Friday having the greatest delays. However, these differences (though significant) were not substantial. Different carriers also showed significant differences in mean arrival time. While most large airlines performed around the middle of the pack, several (notably Delta and Alaska Airlines), performed much better than expected, given their routes and revenue. Among quantitative variables, latitude, longitude, and precipitation showed usefulness as predictors. All in all, predicting flight delays involves a seasonal component as well as understanding the underlying distribution to then make predictions using the various models we developed.

Overall, model construction was challenging with this dataset- not only because of its size, but also because of the inherent difficulty in capturing all the possible sources of variation in flight delays. Every flight is made possible by the coalescence of hundreds, if not thousands of disparate factors; conversely, delays can result from a single anomaly among those. Most of the factors that we uncovered are out of the control of airlines themselves.

Chapter 3

Motivation, Data Collection, and Software & Code

3.1 Motivation

As stated in the introduction, flight delays can have a wide-ranging effect on the economy. Most airline companies have already done everything in their power to mitigate and reduce delays. We are interested in finding whether delays can be further alleviated, and whether those variables can be controlled by airline companies. To the extent that some delays are unavoidable or difficult to predict, we are also interested in devising methods to minimize the impact of those delays, whether by reducing the number of passengers affected, offering alternate routes to affected passengers, or discounting tickets. Overall, for the benefit of airline companies, consumers, and society-at-large, we should minimize flight delays or the impact thereof.

3.2 Data Collection

Our data was provided as .csv files by the competition organizers. The primary dataset was composed of roughly 11 million observations of 50 variables. Each observation was a distinct flight that occurred between 1/1/2018 and 6/30/2019. The 50 covariates included origin, destination, quarter, arrival delay, departure delay, distance, and many more variables pertaining to each flight. Auxiliary datasets included information on flight routes, airports, and market share.

In addition to these data, we also sought out additional information to enhance our analysis. We obtained geographic coordinates for each airport from *openflights.org* and historical weather data from the NOAA databases through the NCDC API. The geographic coordinates were given in decimal format, and our weather data described meteorological events near the origin and destination of each flight. Also, a key was obtained to connect to the Google Maps API through their Google Cloud Platform. Various plots were created using this connection. A more in-depth discussion can be found in the next chapter, under *Data Wrangling*. A full list of covariates along with a brief descriptions can be found in Fig 9.1 of the appendix. See Fig 9.2 for additional information on the stages and segments of air travel. *From here on out, figures starting with the number 9 can be found in the attached appendix, appearing in the order they are mentioned.*

3.3 Software and Code

All analyses were performed in R v3.6.3 using the RStudio IDE. Packages used include, but are not limited to, *ggplot2*, *ggmap*, *dplyr*, *caret*, *rnoaa*, and *tseries*. Individual datasets were loaded as *data.frame* objects and combined using *merge* along with various *dplyr* commands. In addition, Microsoft Power BI was utilized in order to tidy the data and perform computationally-intensive data rearranging. The final dataset and code for this project can be found on our GitHub repository: *github.com/johnathanlo/TAMIDS20*

Chapter 4

Exploratory Data Analysis

4.1 Data Wrangling

Our dataset was drawn from four different main sources - flight delays and airfare data, geographic coordinates from *openflights.org*, and weather data from NOAA. Flight delays and geographic coordinates were combined by merging on both common origin and destination names. The resulting data frame was then combined with the market data by common route, year, and quarter. Adding weather data was more challenging in that the observations related information collected by weather stations, and not the airports themselves. Thus, weather station coordinates were cross referenced with airport coordinates to find the closest active weather station to each airport. Due to this constraint, 10 airports corresponding to 99,980 observations (a negligible amount given the overall size of our data), were dropped due to the lack of NOAA weather stations nearby. Weather data was then merged with the rest of the data on common dates and airports, with separate variables for weather at the origin and destination. The final dataset containing all four sets of information is what will be referenced in this paper hereafter. Small tweaks to these data were made on a model-to-model basis, depending on the type of unique problem selected-whether it be classification or regression, for example.

The data contained a number of variables with numerical values that could be interpreted either as quantitative or categorical variables. Using substantive knowledge, a number of numeric variables were converted to factors, including day of week, month, quarter, and route number. Additionally, our data contained many missing values. Where applicable, missing categorical variable values were replaced by adding an additional factor level *Unk*. For quantitative variables, missing values were either replaced with 0 or observations were removed entirely based on substantive knowledge of the variable characteristics. Finally, we discovered that the flight delay dataset somewhat bewilderingly assigned canceled flights a delay time of 0. Canceled flights were removed from the dataset and analyzed separately. Throughout this report, we further to flight delays as specifically arrival delays as opposed to departure delays. In all, the final dataset consisted of 10,614,150 observations of 100 variables.

4.2 Distribution of Flight Delays

The dataset provided information on flights from 17 different carriers along 6,684 unique routes being flown between a combination of 362 airports in the United States (including Alaska and Hawaii and a couple airports in US territories such as Guam). 3,814,366 out of the 10,915,495 flights (35%) had some form of arrival delay. A histogram of all arrival delays is shown in Fig 9.3. We observe that most "early" flights are not incredibly early compared to the frequency at which late flights can become "significantly" delayed, say more than 30 minutes.

Clearly, the data is strongly right-skewed. To correct the skewness (to meet assumptions for our analysis), a cube-root transformation was performed. Subsequent Shapiro-Wilk tests provided strong evidence against normality for this transformation, so we turned to other methods, as will be discussed in chapter 5. To heuristically assess dependence between covariates and late arrivals, we examined various conditional distributions, as follows.

4.2.1 Geographic Distribution of Flight Delays

In Fig 9.4, the routes that have various average intervals of delay time are shown. Note straight lines are drawn for simplicity; the actual flight most likely flew a non-linear path toward the destination. Also, early arrivals are given a delay time of 0 in the computation of the mean. We observe many things. Almost all routes have an average delay that is positive. This spotlights the nature of this project: to focus on remedying delays to improve customer satisfaction and increase airline revenue. The conditional densities given certain latitudes and longitudes are show in Fig 9.5-9.6. Notice the similarity across distributions.

For the routes with a mean delay between 30 and 45 minutes, most of the delays are clustered on the eastern US with most of the routes either beginning or ending in the San Francisco, New York, or southern Florida regions. For more severe delays of 45+ minutes, these routes encompass more cross-country flights. Furthermore, it can be seen that an airport in the northeast, most likely JFK in New York, is involved in a lot of severe flight delays. In Fig 9.7, the average delays at certain airports is depicted. We see further evidence that the more problematic delays are centered in bigger cities, particularly those on the east and west coast. Fig 9.8 illustrates the relationship between the popularity of an airport and the amount of delays. As expected, airports that crank out more flights have a higher average delay time. This raises the question: Do more flights simply give airports a higher probability of having delays (by random chance), or does an increase in flights also bring in other factors that *cause* an increase in flight delays? In more broad terms, what factors correlate with delays and which ones can be controlled by the airlines?

4.2.2 Temporal Distribution of Flight Delays

Fig 9.9 shows the distribution of flight delays by quarter. Note the similarity of the marginal distributions. Distributions by month, day of week, and time of day are shown in Fig 9.10-Fig 9.12. Note that because our data contained a year and a half's worth of observations, the frequency of delays for quarter 1 and 2 are higher than other time periods that were not recorded twice. The histograms appear roughly symmetric with a slight right-skew. Specifically, early arrivals appear to follow a normal distribution while it transitions to an exponential distribution once delays become positive. We will explore this specific observation in the first section of our next chapter. Regardless, the general form of the distribution across these time factors does not change significantly.

To further investigate the role of temporal factors in flight delays, we produced a number of coplots, which show several interesting seasonal patterns, as depicted in Fig 9.13-9.15. Namely, the most problematic times for delays are quarter 2 and 3, Thursdays and Fridays, and around 5PM-6PM in the evening.

4.2.3 Weather-based Distribution of Flight Delays

Intuitively, the most obvious factor that most likely affects flight delays is the weather. Therefore, we analyzed the relationship between flight delays and average temperature and precipitation. These are described in Fig 9.16-Fig 9.17, respectively. In the scatterplot depicting rain, we see a lack of evidence that more extreme temperatures correlate with increased delays. This would be supported if the scatterplot showed a parabolic pattern, which is not seen. Surprisingly, we see the same idea in the scatterplot of precipitation. Contrary to intuition, as the amount of precipitation increases, there is no visual subsequent increase in the delay time. In Fig 9.18, we see slight changes in the distribution of arrival delays across different precipitation levels. We will return to this observation in chapter 5. Lastly, Fig 9.19 demonstrates that some dangerous weather events have a slightly bigger effect on delay times than others. For example, the distributions of "ice, sleet and hail" and "blowing or drifting snow" have more area in their right-tails. This is indicative of more occurrences where these events caused more significant delays. Besides these difference in tail-density, the overall shape remains very similar.

4.2.4 Carrier-based Distribution of Flight Delays

Next, looking at delays by carrier would provide us with insight as to whether some carriers are better at mitigating delays than others. Fig 9.20 shows us that the distribution across carriers stays roughly the same. Apart from the differences in frequencies (with some airlines being more popular or providing more flight routes), the shape of the distribution is basically homogeneous, especially when compared to the other conditional distributions of other factors. This means that delays across carriers *behave* in the same manner. To analyze if the average delay time (as opposed to the distribution) differs across airlines, we conducted ANOVA (analysis of variance). Specifically, we used Tukey's HSD to make multiple comparisons across each combination of carriers to see what carriers differed from one another. A plot of the resultant 95% confidence intervals for the mean difference between airport delays is show in Fig 9.21. We see that the majority of the intervals constructed do not fall within 0. Thus for those that didn't, we conclude that they do indeed have a difference in mean arrival delays. All in all, we observe that the *distribution* across carriers is similar, but the *quantitative amount of delays* across carriers is different.

4.2.5 Airport-based Distribution of Flight Delays

Fig 9.22 displays the histograms of delays for some of the most popular airports. As has been the trend thus far, the distribution across the airports does not change much save for the changes in frequency. Thus, we can conclude the behavior and process of flight delays is pretty universal and can be modeled with an explicitly defined distribution, as we will dive into in the next chapter.

4.2.6 Making Up Lost Time En Route

One last thing we wanted to explore was the ability for pilots to make up lost time as a result of a departure delay. We presumed that a longer flight, distance-wise, would allow for more opportunity for a flight delay to be alleviated. In the air, harnessing favorable air currents or taking shortcuts can remedy the lost time they left the ground with. Surely, in Fig 9.23, we see that flights over a longer distance have, on average, less arrival delays. Similarly, the shorter

the flight, the more severe the arrival delay is. These observations support our hypothesis stated above. Of course, correlation does not imply causation, so more substantive knowledge on how pilots navigate the flight route would provide more clarification on this.

Chapter 5

Model Formulation and Assessment

5.1 Constructing a Parametric Distribution for Delays

Fresh off our exploratory analysis, we were primarily interested in deducing the marginal distribution of arrival times from the data. Although it would have been relatively facile to estimate a valid empirical distribution, we decided that a parametric distribution would be more useful and intuitive. By establishing a set of parameters, further work could be directed towards estimating parameters under certain combinations of covariate values. With mostly categorical data, it then became feasible to estimate parameters for certain combinations of interesting variables. Parameterization also allows the density functions of the distribution to be expressed analytically.

As seen in Fig 9.3, the marginal distribution is strongly right skewed, and thus our first plan of action was to attempt a transformation to correct the skewness. A number of transformations were considered. With negative values in the data, log or square root transformations could only be applied by first shifting the data to be strictly positive. One way to do this was to simply add the smallest (most negative) number to each of our delay times; however, it was decided that this approach rendered our results somewhat uninterpretable. Consider, for example, a new observation where the flight arrived earlier (with a negative flight delay) than any other flight in the dataset. Such an observation would not be supported in a distribution of log-transformed values. We also briefly considered cube root transformation, but as seen in Fig 9.24, it did not result in normality or resemblance to any familiar parametric distribution. In lieu of transforming the data, we considered several well-known skewed distributions, but none of them fit well or appeared sensible.

Thus, we instead looked to construct a mixture distribution. The primary issue that we had been confronted with thus far was finding a distribution that appeared to have sensible parameterizations. Our search was rooted in the premise that not all delays are created equal. We suspected that the majority of delays are "run-of-the-mill" events that do not result from any extraordinary circumstances in particular, while a minority of delays have true substantial causes. This is similar to the rationale for a zero-inflated Poisson distribution used in manufacturing processes. This famous distribution assumes most machines are in good working order and do not produce any products with defects, but some machines with defects will produce problematic products by a Poisson distribution.

In this vein, we decided upon a mixture of the form:

$$Y = UV + (1 - U)T$$

with $U \sim Ber(p)$, $V \sim exp(\lambda)$, and $T \sim N(\mu, \sigma^2)$. In this model, U describes whether or not a delay with "true, substantial causes" occurs, T describes the distribution of arrival delays when no extenuating circumstances occur, and V describes arrival delays under defined circumstances that result in lateness. We calculated the cumulative distribution function (CDF) of this distribution is given by

$$F_X(x) = \sum \alpha_i F_i(x) = p(1 - e^{-\lambda x}) + (1 - p)\left(\frac{1}{2}(1 + erf(\frac{x - \mu}{\sigma\sqrt{2}}))\right)$$

e.g. a simple mixture of three distributions. The complementary probability density function (PDF) was found to be:

$$f_X(x) = F'_X(x) = p(\lambda e^{-\lambda x}) + (1 - p)\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\right)$$

To estimate our parameters, we constructed functions in R to perform maximum likelihood estimation for our distribution. Using *optim()* in R, we were able to generate estimates of the parameters for the marginal data and compare them to the empirical distribution, as shown in Fig 9.25.

We decided to name this distribution the Lo-Ke disitrbution after the members of this team who formulated it. To validate this distribution, a QQ plot was made of the observed arrival delays distribution against our Lo-Ke distribution under the parameter estimates from the MLE (Fig. 9.26). From this plot, it can be observed that our mixture distribution is elegantly able to describe the observed values of arrival delays quite accurately. The points hugging the diagonal line mean that the quantiles on our custom distribution match up with the quantiles of the data. Note that the arrival times are given as discrete values (in minutes), hence the "jumps" from percentile to percentile. It is important to note here that our distribution is not as strong in describing data points in the extreme upper range of observations, $p < .001$. Some delays were quite extreme, with values in excess of 1,000 minutes. To prevent such outliers from having undesired effect on our downstream analyses, we further pared the dataset here to include only data below the 99th percentile.

Further, we see that conditional density estimation agrees with our theoretical parametric distribution. In Fig 9.27, we show the marginal distribution for a subset of flights out of the ABE airport, and compare it to the theoretical distribution constructed with MLE. Additional examples of the empirical conditional density vs our MLE estimates are shown in Figs 9.28-9.29. Overall, we suggest that our theoretical Lo-Ke distribution can be used to accurately describe arrival delays across all combinations of factor levels.

5.2 Linear Model Using OLS

Ideally, when constructing a linear model using OLS (ordinary least squares) estimates, we want to satisfy four principal assumptions. Namely, these are: linear relationship between

predictors and responses, independence of errors, homoscedasticity, and normality of the errors. However, having seen that the conditional distributions of the response are strongly *non-normal* and that there exists no simple transformation to restore normality, the fourth assumption appears to be violated. It is important to remember then, that the Gauss-Markov theorem does *not* require normality of the errors for the OLS estimates to be the best linear unbiased estimator and that normally distributed errors are required only for conducting inference on the model via the t-distribution. For merely constructing a model equation and generating predictions that minimize the mean squared error, normality of the errors is not strictly required. Further, inference on the model parameters can still be conducted through a variety of alternative methods. With heavy-tailed distributions of residuals, usually the more general M-estimation is used in place of OLS estimation to provide robustness against non-normality when conducting hypothesis tests on model parameters. Alternatively, bootstrap can be used to generate the empirical error distribution. However, the size of our dataset obviates the need for either of these procedures since we can reasonably assume by the Central Limit Theorem that our estimates of the mean response, and therefore the coefficients of the model, are normally distributed. Therefore, the only inference that cannot be performed with the usual set of equations is prediction. A scatterplot matrix for the OLS regression is shown in Fig 9.30.

To conduct prediction and construct prediction intervals, we use our theoretical Lo-Ke distribution to describe the errors. The predictions thus have the form:

$$Y_{new}|X \sim X\beta + e$$

with variance

$$Var(Y_{new}|X) = X^T X Var(\beta) + Var(e)$$

and

$$E(Y_{new}|X) = E(X\beta)$$

The question then becomes how to select parameters for the theoretical distribution of our conditional responses. Particularly, how do the parameters relate to the mean and variance given above? For errors that are normally distributed, or distributed via some member of the location-scale family of distributions, the parameters are precisely those given by the variance and expectation above. However, for our Lo-Ke distribution, solving for the correct parameter values becomes more challenging. Analytically, this requires either solving for the method of moments estimators or the maximum likelihood estimators. Both analytical solutions become complex rather quickly. For example, we calculated the quite-complex moments of our theoretical distribution as follows:

$$M_1 = \frac{p}{\lambda} + (1-p)\mu$$

$$M_2 = p\left(\frac{2}{\lambda^2} - \frac{2\mu+1}{\lambda} - \mu^2 - \sigma^2\right) + \mu^2 + \sigma^2$$

$$M_3 = \frac{6p}{\lambda^3} - (p+1)(\mu^3 + 3\mu\sigma^2)$$

where we use M_i to denote moments rather than the usual μ_i, to avoid confusion. To actually solve for the four parameters, we would also need a fourth moment. Likewise, for an analytical solution using MLE, we would need to solve a system of four partial derivatives. While such a solution exists and is not as time consuming as it might appear, we can approximate a solution by conducting numerical MLE on a subset of data close to the given values of the predictors. We can also observe that the parameter with the greatest influence on the mean of the distribution is the parameter p. Therefore, as the conditional mean response varies in our linear model, we can vary p in the error distribution to allow for accurate inference on predictions. Using our above-defined moments, we can also calculate the variance of our distribution:

$$Var(x) = p\left(\frac{2}{\lambda^2} - \frac{2\mu+1}{\lambda} - \mu^2 - \sigma^2\right) + \mu^2 + \sigma^2 - \frac{p^2}{\lambda^2} - (1-p)^2\mu^2 - \frac{2p(1-P)\mu}{\lambda}$$

A full model, along with functions for prediction, is on our GitHub repository.

5.3 Logistic Regression

The majority of the models in data science can be put into three bins: regression, classification, and clustering. As we came to realize the deficiencies in our regression model, we turned to a classification method. In particular, we chose binary logistic regression to predict whether or not a flight would be delayed or not given certain pre-departure covariates. Our motivation for pursuing logistic regression was two-fold. For one, we wanted to improve on the accuracy of our predictive model, so by moving away from regressing on a continuous variable, we shifted our focus to predicting a binary success or failure. Secondly, our conditional density estimates illustrated that given certain covariate levels, the *shape* of our distribution changes more than the *mean* (location) of our distribution. Thus, regressing on the parameter p from our Bernoulli random variable from our mixed distribution would help us explain these observed changes in our conditional density estimates. The marginal distribution of U (our bernoulli random variable) in our model changes how thick-tailed our distributions become on either side of 0.

Using the *glm()* function in R, we fit a generalized linear model using the logit function as the link function and the binomial distribution family as the probability distribution. A 5% subset of our data frame was taken to shorten computation time.Next, 80% of our data was allotted for training and 20% for testing. The summary of our fit model is show in Fig 9.31. Our covariates included various factors on the time as well as weather phenomena in the departure and arrival airport. We interpret our logistic regression output in two separate ways for numerical and categorical data. For numerical data, the specific factor level's coefficient can be interpreted as the change in the log-odds of the "success" occurring – in our case, a delay. Exponentiating these values gives you a more interpretable estimate. For example, from our model output, we can say that for every additional millimeter of snow, the odds of a delay occurring (versus not occurring) increases by a factor of $e^{0.1283} = 1.013$. For categorical data, the interpretation is a little different. We illustrate by example. In our case, flying in April as opposed to January lowers one's chances of having a flight delayed by a factor of $e^{-0.2646} = 0.767$. Furthermore, we see that the majority of our coefficients are statistically significant as indicated by the asterisks next to each line. These significant predictors are thus being utilized in our model, as desired.

Since our logistic regression model outputted a *probability* that a delay will occur, we needed to determine the optimal probability threshold to determine whether or not a delay will take place. We thus improved our model by plotting an ROC curve, or a "Receiver Operating Characteristics" curve, as shown in Fig 9.32. By maximizing the area under the curve,

we are able to increase our accuracy. The y-axis gives the true positive rate while the x-axis gives 1 minus the true negative rate. Thus, by altering the probability threshold, we could then predict with better accuracy. Using this ROC curve and various functions, we found that the optimal probability threshold was 0.39. This means that any predictions that output the probability of a delay being over 0.39, we say that a delay will occur.

Finally, after validating our model on the testing data, we generated a confusion matrix and other measures of performance as shown in Fig 9.33. Our accuracy was 67.62% where out of our testing set of 92,667 observations, it correctly predicted 62,187 of them. Observe from the confusion matrix that we have significantly more false positives than false negatives. Thus, we interpret this as our model is on the pessimistic side and is more likely to forecast a delay when in reality there is not one.

5.4 Time-Based Model: Dynamic Regression

From our exploratory data analysis, we noted that several different categorical variables related to time showed significant effects on the mean arrival time. In particular, the scheduled time of a flight had a noticeable effect on arrival delays over the course of a day. While our original linear model accounted for time-related effects by including aspects of the date and time as categorical variables, the assumptions of linear regression do not allow for autocorrelation which we believed could play a role in the true model of arrival times. For example, one could imagine that a flight that randomly arrives 10 minutes late could cause a subsequent flight to also arrive $f(10)$ minutes late, where f is some autocorrelation function. We also noted that the residuals did not appear to be random; i.e. not independent or identical. To adjust our model to reflect this reality, we constructed a new model:

$$Y_t = X_t\beta + \eta_t$$

where η_t represents a time series following an auto-regressive integrated moving average (ARIMA) model, and $X\beta$ represents a linear combination of the other (non-time-related) variables fitted against the *residuals* from the time series.

The difficulty with constructing this model was that the dataset was not originally in a time series format. We wanted our time series model to take into account daily, weekly, and quarterly seasonality. We briefly considered constructing a time series on average measurements per time period, potentially split by some categorical grouping (e.g. flight routes, or carriers), but we considered the loss of information (through taking averages) too heavy of a price to pay. We wanted to construct time series data that consisted of one observation at each time interval: with 19 time intervals per day, 7 days per week, and 91.3125 days per quarter. To do this, we used a bootstrap method wherein we sampled with replacement from the original data: 6,935 observations per year, corresponding to 6,935 time periods across 100 years. This resulted in a bootstrap time series of 700,000 observations. It is important to note that the construction of this time series precludes any long-term trends, i.e. the series is stationary from year to year.

Using this bootstrap time series dataset, we fitted a multi-seasonal ARIMA (auto-regressive integrated moving average) model. The decomposition of this time series into constituent seasonalities and trends can be seen in Fig 9.34. Then, taking the residuals from our time series and appending it back to the original data, we fitted a linear model to predict these residuals

using our covariates that did not involve time (since we already accounted for this in our time-based model). Our goal in fitting another linear model to the residuals was to try and explain the variation in the time-series that could not be picked up by the seasonalities. The resulting model produced residuals that were approximately random (Fig 9.35), or at least more random than our previous models. Our R^2 was 0.369; 36.9% of the variation in our ARIMA residuals could be explained by the model we fit, with output shown in Fig 9.36. After analyzing the plots of our covariates, we saw that there was a slight non-linear relationship with precipitation (the variance was decreasing as the precipitation was increasing) (Fig 9.37), so we decided to correct for this by weighting observations by the square root of precipitation. After this, our R^2 increased to 0.48 (Fig 9.38), a huge improvement. We had now constructed an additive model that took into account both seasonality and other covariates describing weather, geographical, and market data.

Chapter 6

Forecasting Flight Delays for 2019 Q3

6.1 Challenges

Predicting flight delays is exceptionally important to facilitating economic efficiency, and for precisely that reason, it is vital to make predictions. As this is such an important topic, it is surely well-trodden territory with contributions made from many of the world's most preeminent statisticians. Additionally, as previously mentioned, each flight is affected by thousands of variables, each of which can potentially cause delays, either by themselves or in combination with others. Keeping track of all these variables and making predictions from them can be incredibly computationally intensive. Our exploration of the data confirms for us the inherent unpredictability of flight delays. It is clear that black swan type events are important in the genesis of flight delays, and the best safeguard against this is not an exhaustive predictive model, but robustness.

If one wants to predict solely whether or not a delay will occur or not, our logistic regression model can deal with this. If one wishes to predict specific numerical delays, they could utilize our final additive model. Our model was fit by averaging the coefficients generated by OLS on five separate samples of 50,000 observations.

6.2 Using our Forecast Model

To use our final forecast model, we have implemented an R function on our GitHub that will receive an observation and generate a prediction. As previously stated, our model is of the form

$$Y_t = X_t\beta + \eta_t$$

In order to forecast future delays, time covariates are fed into our multi-seasonal ARIMA time-series model, and the other covariates (weather data, market data, etc) are inputted into our linear model. The output from both of these constituent models, namely the ARIMA forecast and the fitted value of our regression are added together to obtain our predicted arrival delay. The predicted arrival delay will be distributed with a mean given by our forecast model and quantiles given by our created Lo-Ke distribution.

Chapter 7

Business Recommendations

7.1 Differences Between Carriers

As we saw in our analysis of variance in our data exploration, the average delay times across carriers do differ. More popular carriers such as American Airlines do incur more delays than say, a lesser known airline such as Pinnacle Airlines. For large carriers, Alaska Airlines and Delta do a very solid job of alleviating delays. It is also evident the more flights an airport hosts, the more chance there is at a delay occurring. From a customer's point of view, one could simply avoid flying with the most popular airlines if the cost and quality level of another airline is adequate. From this same perspective, flying during less delay-prone times, by leaving late at night, flying during the months of January and February, or leaving on a Saturday would help families and the like minimize their chances of experiencing delays.

7.2 Important Covariates

As a result from playing around so much with our covariates, we gained a solid idea of what does and what doesn't go into predicting a flight delay. Surprisingly, temperature has a minimal effect on delays, and although precipitation and weather events do contribute significantly to delays, they are rare events in the grand scheme of things, and are thus not the best way to predict future delays. On the other hand, airlines should schedule flights as to avoid propagation of delays throughout the day. Delay propagation occurs when a delayed flight early on in the day creates a ripple effect of increasingly delayed flights as the day progresses. In a sense, a delay becomes contagious as it quite literally permeates time and spreads across the country and to different airports. Thus, it is advantageous to commit resources to ensuring early flights do not get delayed. As the number of flights that take off toward the evening and into the night decreases, the airlines can catch-up during these phases to start fresh the next day. Survey data conducted on a random sample of customers could provide airlines with information regarding passengers' willingness to fly early in the morning or late in the evening. Thus, data on clusters of passengers could be analyzed to move certain flights to different time blocks while maximizing the amount of people who would be willing to purchase tickets for these.

Delays occurring simply comes down to the popularity of the airport. The more balls that are thrown into the air, the more likely one will slip the hand and fall to the ground. In the end, delays will be inevitable. If airlines can predict when they will occur, then they can be more prepared to respond by compensating customers with special packages or alerting commuters further in advance about the change in logistics.

Chapter 8

Closing Thoughts

8.1 Retrospect

There are many possible approaches to this type of statistical problem, so there are a plethora of algorithms and models that could end up providing the best descriptive or predictive power. Due to time and computing-power constraints, we simply employed the ones that we felt were the most efficient and promising given our circumstances. In the end, many of our critical choices were subjective. We chose our response variable, the arrival delay time, based on our own preferences and experiences with air travel. We felt that the arrival delay would be the variable of interest (as opposed to the departure delay, for example) for the majority of consumers and the most economically impactful. Choosing alternative responses such as departure delays or cancellations have their own justification. Additionally, the response could have been transformed in a number of different ways that could have made it more informative. For example, one might perhaps be interested in how long a delay will prolong only *if* it occurs. Or, one could desire to know the delay time as a proportion of the original flight time. Even more, some delay metric adjusted for the distance between origin and destination could be revealing. Regardless, our choice of predictor variables was informed by our own background and objectives. On one other hand, for researchers with backgrounds in econometrics, it may have been more interesting to study the dataset in the context of the economic environment of 2018-2019.

Technology is the other underlying component of successful statistical analysis. Being limited to our home computers for the majority of this project, we were unable to run algorithms requiring high computational loads and large memory. In such circumstances, it may have been wiser to familiarize ourselves with cloud computing services like Google Cloud's virtual machines before subjecting our own laptops to such heavy workloads. A solid amount of our models took hours and sometimes days to run. When finally finished computing, we would often need to make adjustments and re-run it again, which would require more waiting. Despite all this, overall we believe that we leveraged the available technological resources to our satisfactory.

8.2 Future Work

There are many potential analyses that can still be done on this dataset. With more computing assets, more complex machine and deep learning algorithms such as neural networks and high-dimensional clustering could be applied. We believe that one of the strengths of our analysis was compiling a dataset that incorporated both the data given to us by the organizers and large amounts of outside data. In fact, we found copious amounts of downloadable outside

data on airport characteristics such as runway lengths and airspace class that could have been incorporated were it not for the physical restrictions imposed by our limited RAM.

The question of how best to ameliorate arrival delays remains an issue of core importance to the economy. Generations of data scientists have tackled this problem, and future generations will continue to. It is obvious that the more delays are reduced, the more utility and convenience can be added to various modes of transportation. It is unlikely delays can be removed entirely, so the job of data scientists is to predict them as accurately as possible. With vast and expansive amounts of data out there, the trajectories to tackle this challenge are endless. That is the beauty of big data in this new era we are in. Without a doubt, airline industries will be harnessing the power of data science, both tomorrow and forever.

Chapter 9

Appendix

9.1 Additional Figures, Tables, and Data

[Covariate]	[Type]	[Description]
DEST	Factor w/ 362 levels	IATA code for destination airport
ORIGIN	Factor w/ 362 levels	IATA code for origin airport
YEAR	Factor w/ 2 levels	2018 or 2019
QUARTER	Factor w/ 4 levels	Q1, Q2, Q3, or Q4 of a year
FL_DATE	Factor w/ 546 levels	A unique day between 01/01/2018 and 06/30/2019
MONTH	Factor w/ 12 levels	Month of year
DAY_OF_MONTH	Factor w/ 31 levels	Day of the month
DAY_OF_WEEK	Factor w/ 7 levels	Day of the week
CARRIER	Factor w/ 17 levels	IATA code for airline carrier for that flight
FL_NUM	Integer	Flight number
Route	Factor w/ 6,684 levels	Unique route number for that flight path from origin to destination
DEST_CITY	Factor w/ 355 levels	Destination city
DEST_STATE	Factor w/ 52 levels	Destination state
CRS_DEP_TIME	Integer	CRS scheduled departure time
DEP_TIME	Integer	Actual departure time
DEP_DELAY	Integer	Difference in minutes between scheduled and actual departure time
DEP_DELAY_NEW	Integer	Same as DEP_DELAY except early arrivals are set to zero
DEP_DEL15	Binary	0=less than 15 minute delay, 1=more than 15 minute delay
DEP_DELAY_GROUP	Factor w/ 13 levels	Delay Intervals: every 15 minutes from -15 to 180 minutes
DEP_TIME_BLK	Factor w/ 19 levels	CRS scheduled departure time block (hourly)
TAXI_OUT	Integer	Taxi out time (minutes)
WHEELS_OFF	Integer	Wheels off time
WHEELS_ON	Integer	Wheels on time
TAXI_IN	Integer	Taxi in time (minutes)
CRS_ARR_TIME	Integer	CRS scheduled/expected arrival time
ARR_TIME	Integer	Actual arrival time
ARR_DELAY	Numeric	Difference in minutes between scheduled and actual arrival time
ARR_DELAY_NEW	Integer	Same as ARR_DELAY except early arrivals are set to zero
ARR_DEL15	Binary	0=less than 15 minute delay, 1=more than 15 minute delay
ARR_DELAY_GROUP	Factor w/ 13 levels	Delay Intervals: every 15 minutes from -15 to 180 minutes
ARR_TIME_BLK	Factor w/ 19 levels	CRS scheduled arrival time block (hourly)
CANCELED	Binary	0=not cancelled, 1=flight was cancelled
CANCELLATION_CODE	Factor w/ 5 levels	A, B, C, or D
DIVERTED	Binary	0=not diverted, 1=was diverted
CRS_ELAPSED_TIME	Integer	CRS allotted elapsed time of flight (in minutes)
ACTUAL_ELAPSED_TIME	Integer	Actual elapsed time of flight in minutes
AIR_TIME	Integer	Actual elapsed time of flight in minutes
DISTANCE	Integer	Distance between airports
CARRIER_DELAY	Integer	Carrier delay in minutes
WEATHER_DELAY	Integer	Weather delay in minutes
NAS_DELAY	Integer	National Air System delay in minutes
SECURITY_DELAY	Integer	Security delay in minutes
LATE_AIRCRAFT_DELAY	Integer	Late aircraft delay in minutes
PASSENGERS	Integer	Total number of passengers for this year, month, route and carrier
EMPFULL	Integer	Full-time employees for this year, month, route and carrier
EMPPART	Integer	Part-time employees for this year, month, route and carrier
EMPTOTAL	Integer	Total employees for this year, month, route and carrier
EMPFTE	Integer	Full-time equivalent employees for this year, month, route and carrier
NET_INCOME	Numeric	Net income for this year, month, route and carrier
OP_REVENUES	Numeric	Operating revenue for this year, month, route and carrier
ID	Factor w/ 350 levels	Unique ID of origin airport from NCDC database
AIRPORT_NAME	Factor w/ 350 levels	Full origin airport name
Latitude	Coordinate	Latitude coordinate of origin airport
Longitude	Coordinate	Longitude coordinate of origin airport
DIST	Numeric	Distance from origin airport to nearest active weather station (km)
Prcp	Numeric	Precipitation at origin (tenths of mm)
Snow	Numeric	Snowfall at origin (mm)
Tavg	Numeric	Average temperature at origin (tenths of degrees C)
Tmax	Numeric	Maximum temperature at origin (tenths of degrees C)
Tmin	Numeric	Minimum temperature at origin (tenths of degrees C)
Wt01	Binary	Fog, ice fog, or freezing fog (may include heavy fog)
Wt02	Binary	Heavy fog or heaving freezing fog (not always distinguished from fog)
Wt03	Binary	Thunder
Wt04	Binary	Ice pellets, sleet, snow pellets, or small hail
Wt05	Binary	Hail (may include small hail)
Wt06	Binary	Glaze or rime
Wt07	Binary	Dust, volcanic ash, blowing dust, blowing sand, or blowing obstruction
Wt08	Binary	Smoke or haze
Wt09	Binary	Blowing or drifting snow
Wt10	Binary	Tornado, waterspout, or funnel cloud
Wt11	Binary	High or damaging winds
ID.DEST	Factor w/ 350 levels	Unique ID of destination airport from NCDC database
AIRPORT_NAME.DEST	Factor w/ 350 levels	Full destination airport name
Latitude.DEST	Coordinate	Latitude coordinate of destination airport
Longitude.DEST	Coordinate	Longitude coordinate of destination airport
DIST.DEST	Numeric	Distance from destination airport to nearest active weather station (km)
Prcp.DEST	Numeric	Precipitation at destination (tenths of mm)
Snow.DEST	Numeric	Snowfall at destination (mm)
Tavg.DEST	Numeric	Average temperature at destination (tenths of degrees C)
Tmax.DEST	Numeric	Maximum temperature at destination (tenths of degrees C)
Tmin.DEST	Numeric	Minimum temperature at destination (tenths of degrees C)
Wt01.DEST	Binary	Fog, ice fog, or freezing fog (may include heavy fog)
Wt02.DEST	Binary	Heavy fog or heaving freezing fog (not always distinguished from fog)
Wt03.DEST	Binary	Thunder
Wt04.DEST	Binary	Ice pellets, sleet, snow pellets, or small hail
Wt05.DEST	Binary	Hail (may include small hail)
Wt06DEST	Binary	Glaze or rime
Wt07.DEST	Binary	Dust, volcanic ash, blowing dust, blowing sand, or blowing obstruction
Wt08.DEST	Binary	Smoke or haze
Wt09.DEST	Binary	Blowing or drifting snow
Wt10.DEST	Binary	Tornado, waterspout, or funnel cloud
Wt11.DEST	Binary	High or damaging winds
Nsmiles	Integer	Non-stop miles between these cities
Fare	Numeric	Average fare between these cities
Carrier_lg	Factor w/ 10 levels	Carrier with the largest market share between these cities
Large_ms	Numeric	Proportion of market covered by carrier with largest market share
Fare_lg	Numeric	Average Fare for Largest Carrier between these cities
Carrier_low	Factor w/ 12 levels	Carrier with smallest market share between these cities
Lf_ms	numeric	Proportion of market covered by carrier with smallest market share
Fare_low	Numeric	Average fare for carrier with smallest market share

Figure 9.1: List of the final set of covariates along with their type and a brief description.

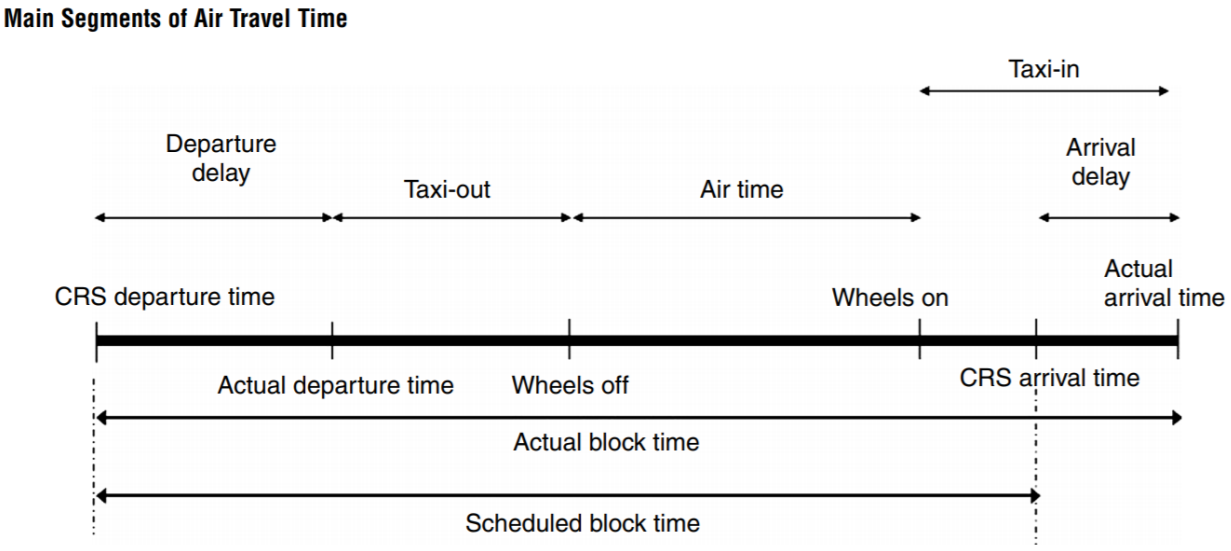


Figure 9.2: Main Segments of Air Travel Time

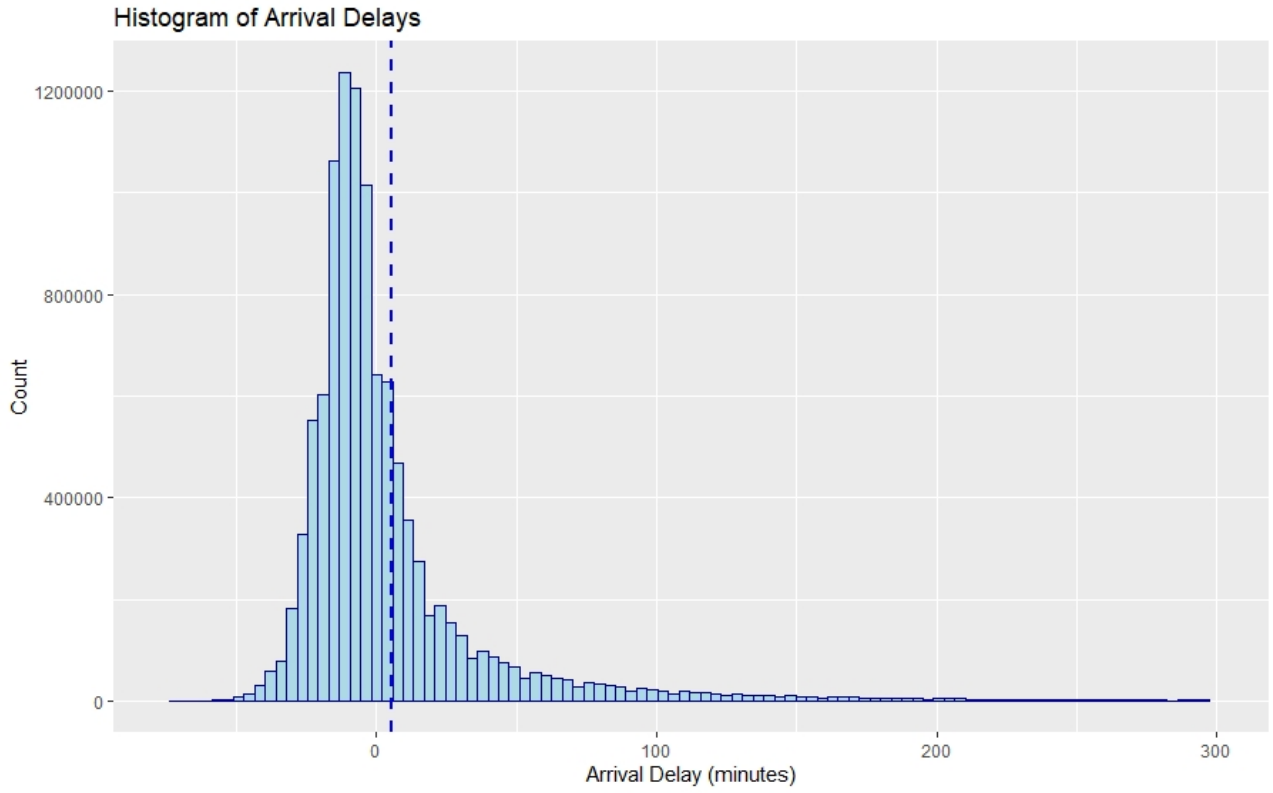


Figure 9.3: Histogram of arrival delays for all observations in the dataset. The dashed line depicts the mean.

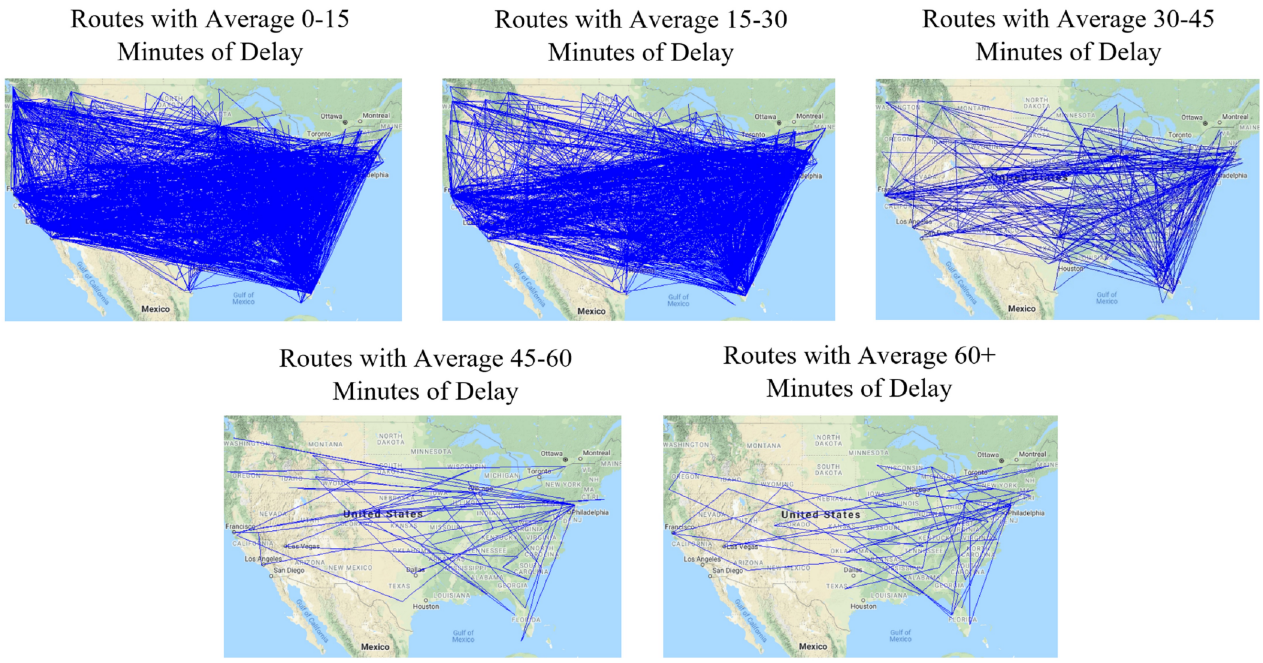


Figure 9.4: Routes with different averages of delay time

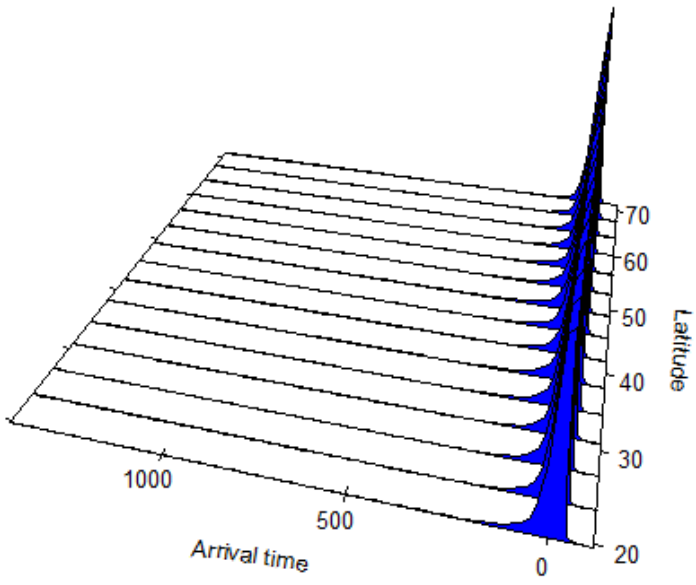


Figure 9.5: Conditional densities of arrival delays across different latitudes

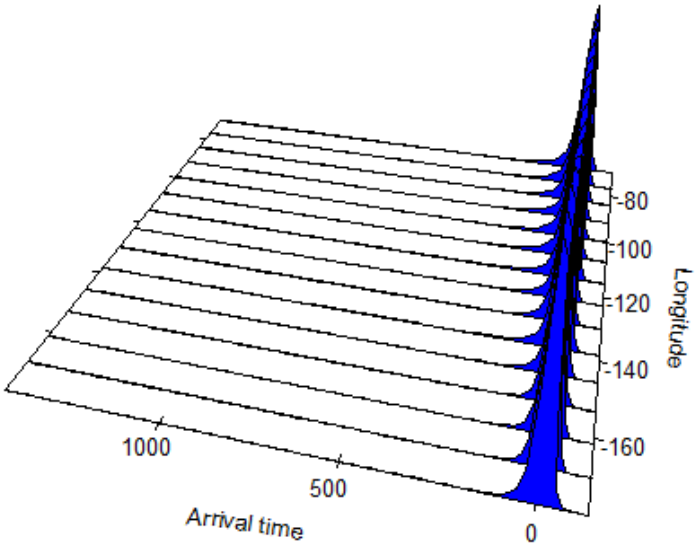


Figure 9.6: Conditional densities of arrival delays across different longitudes

Figure 9.7: Average delay at destination airports

Figure 9.8: Average delay and total number of flights at destination airports

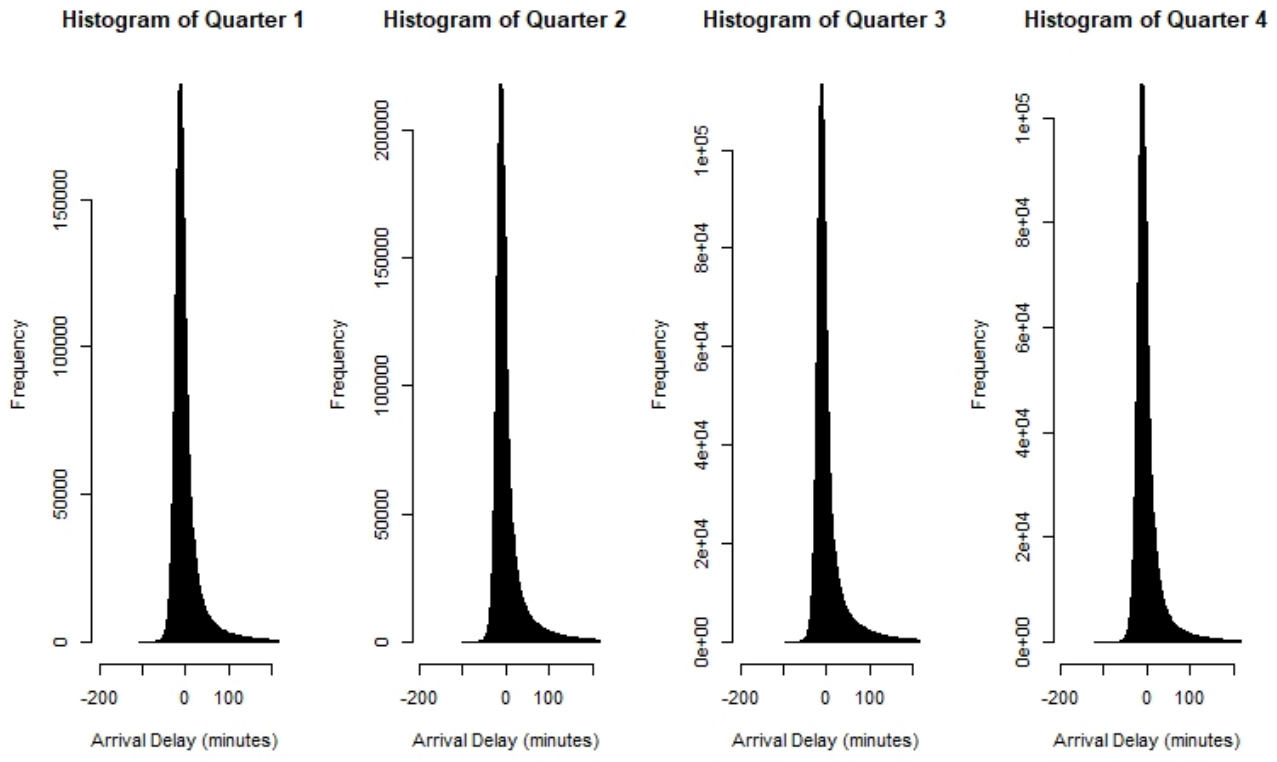


Figure 9.9: Delays by quarter

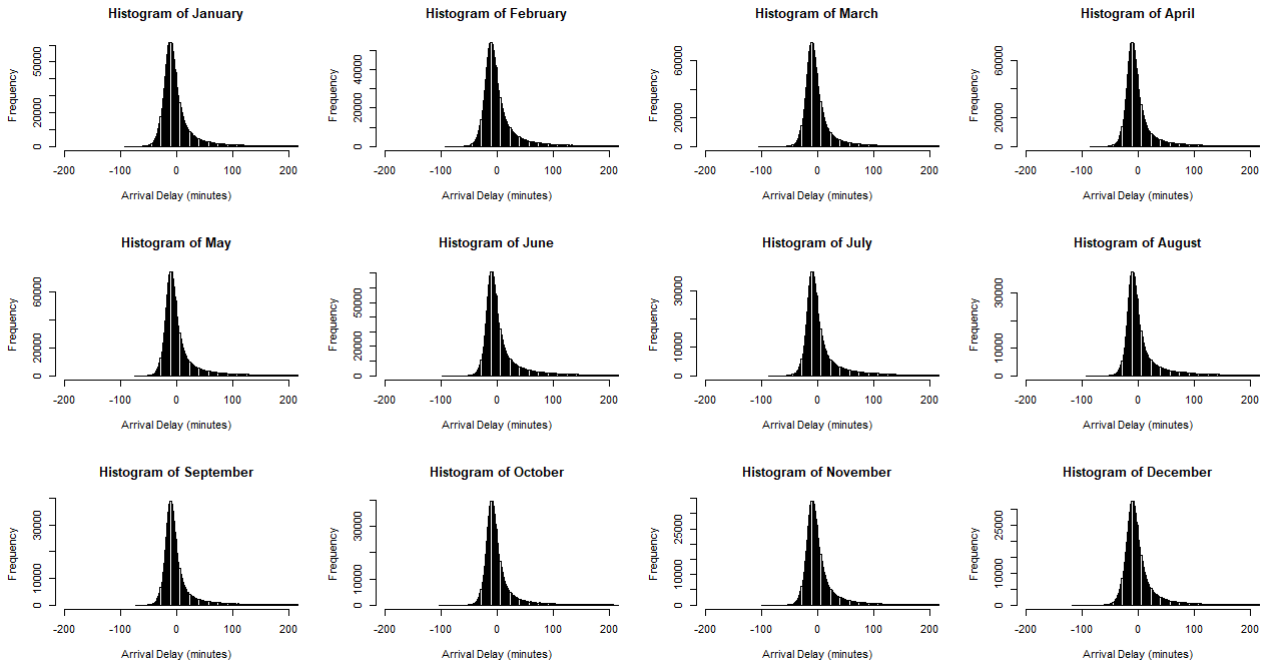


Figure 9.10: Delays by month

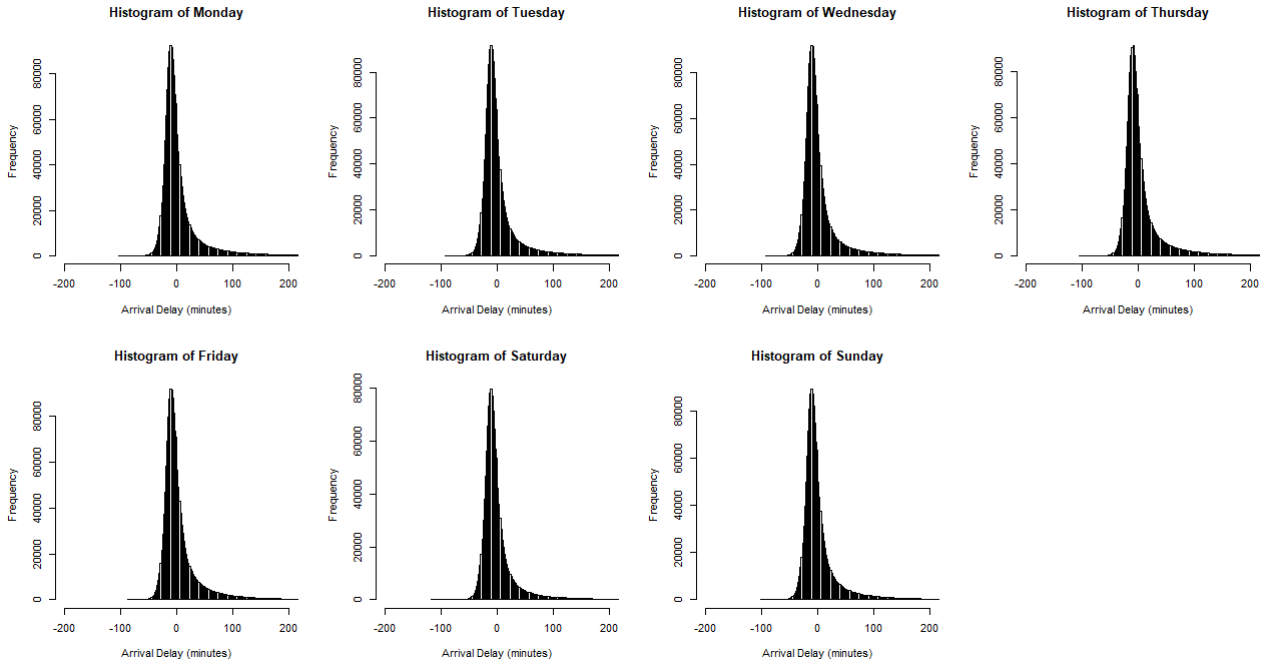


Figure 9.11: Delays by day of the week

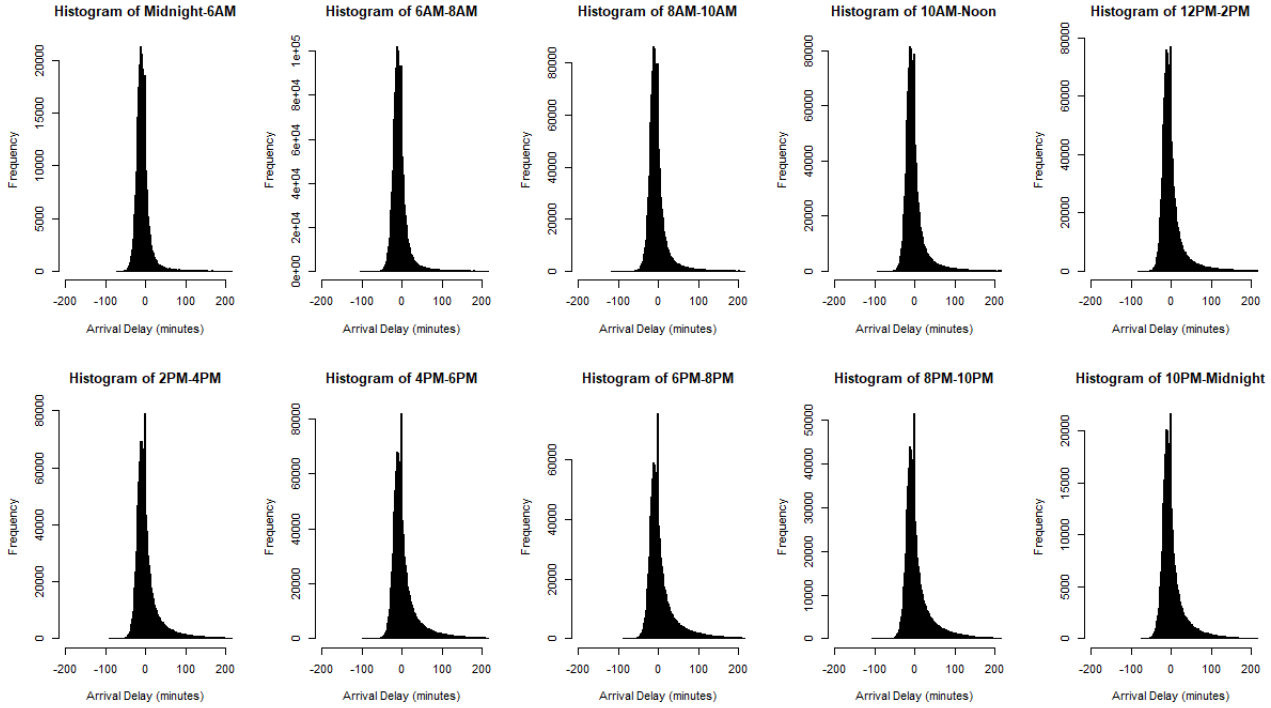


Figure 9.12: Delays by time of day

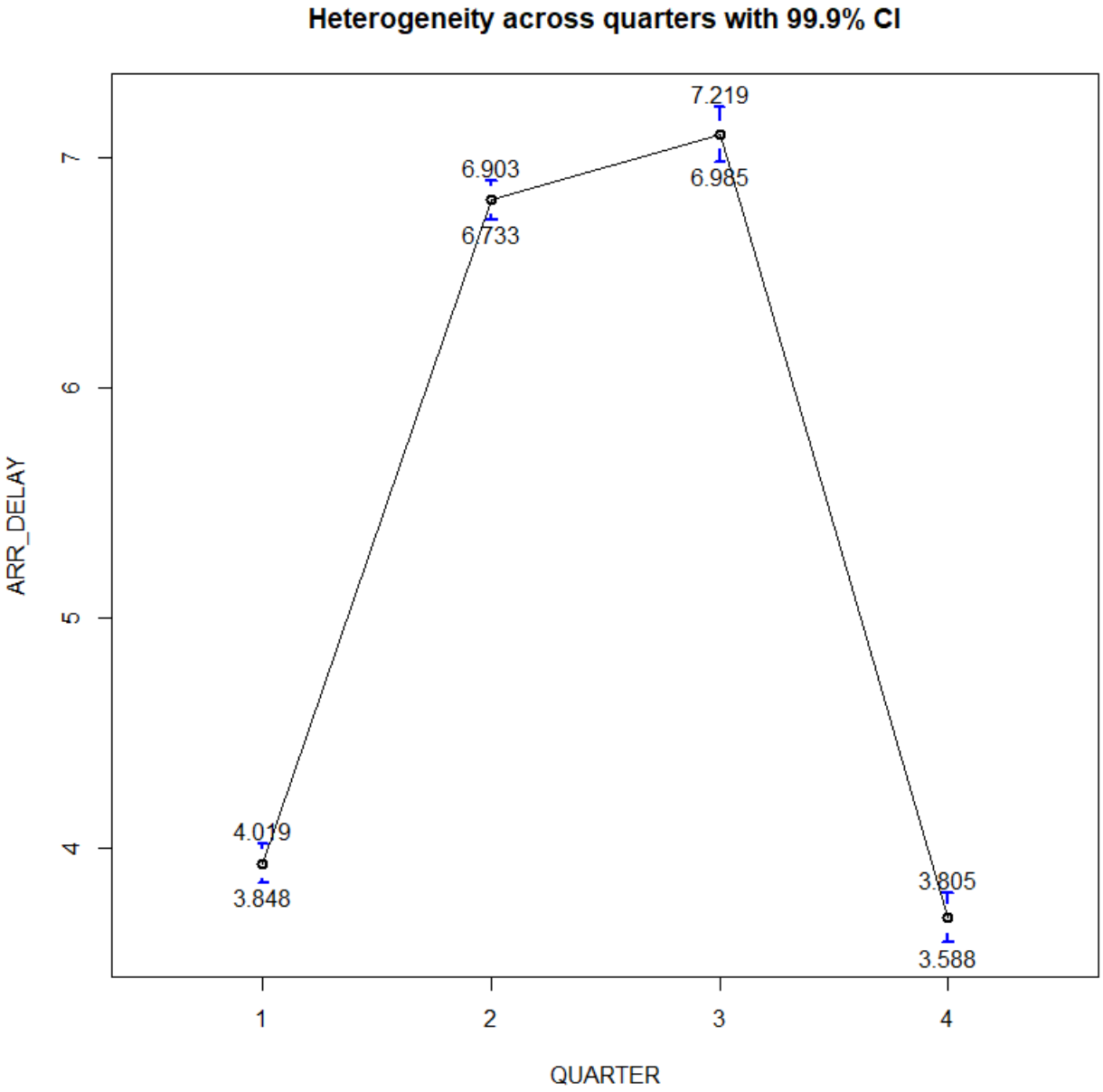


Figure 9.13: Heterogeneity by quarter

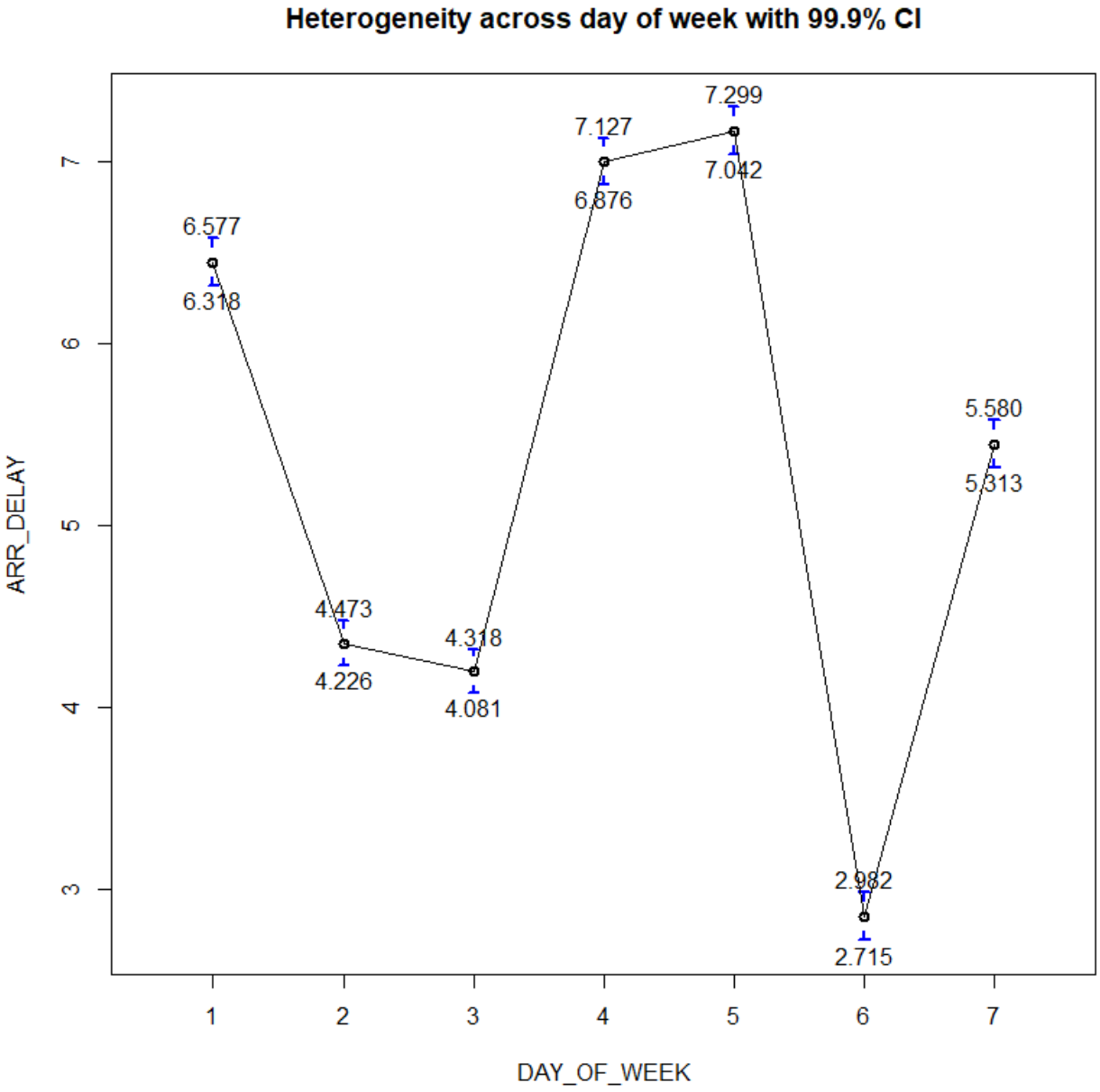


Figure 9.14: Heterogeneity by day of week

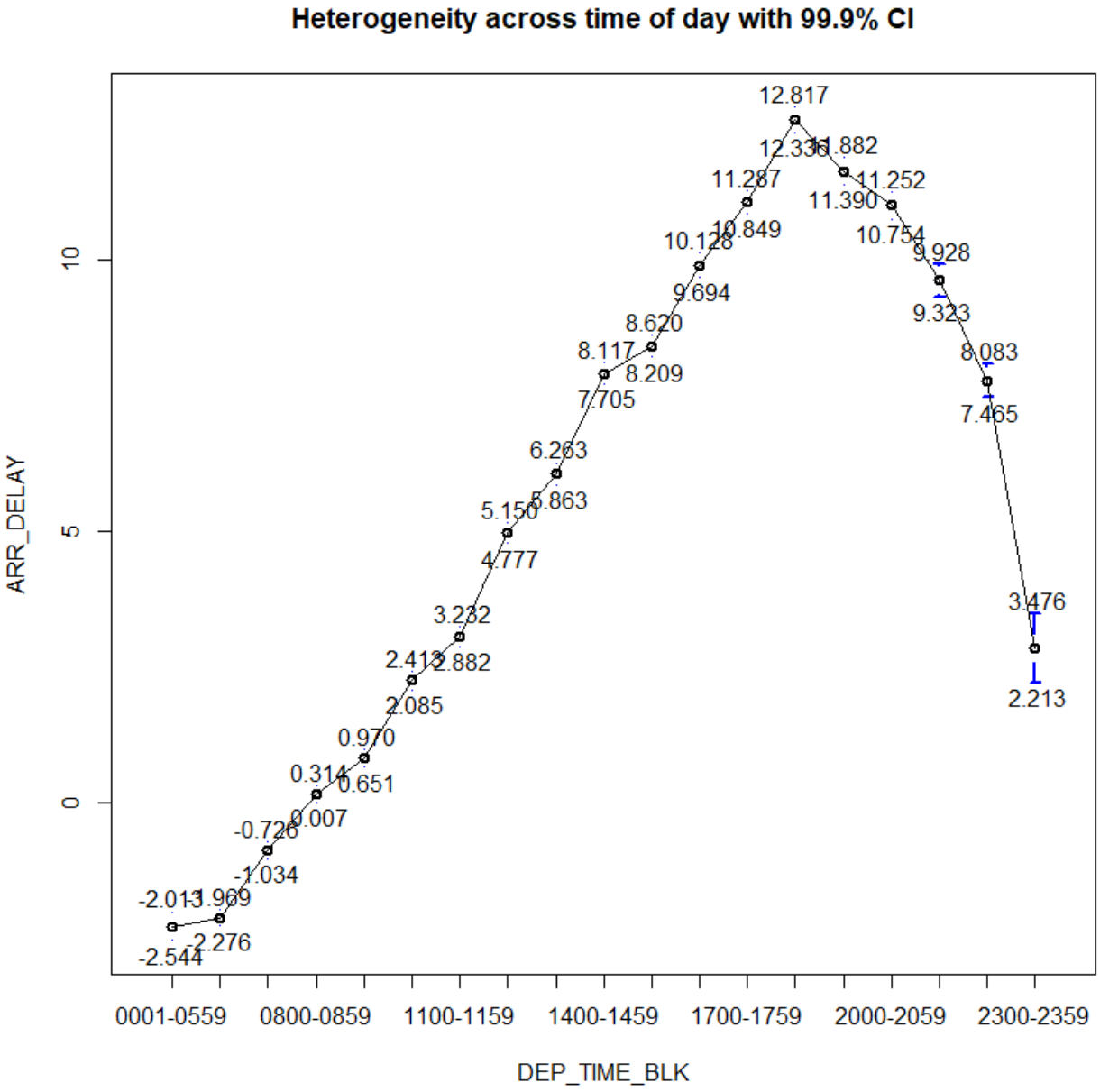


Figure 9.15: Heterogeneity by time of day

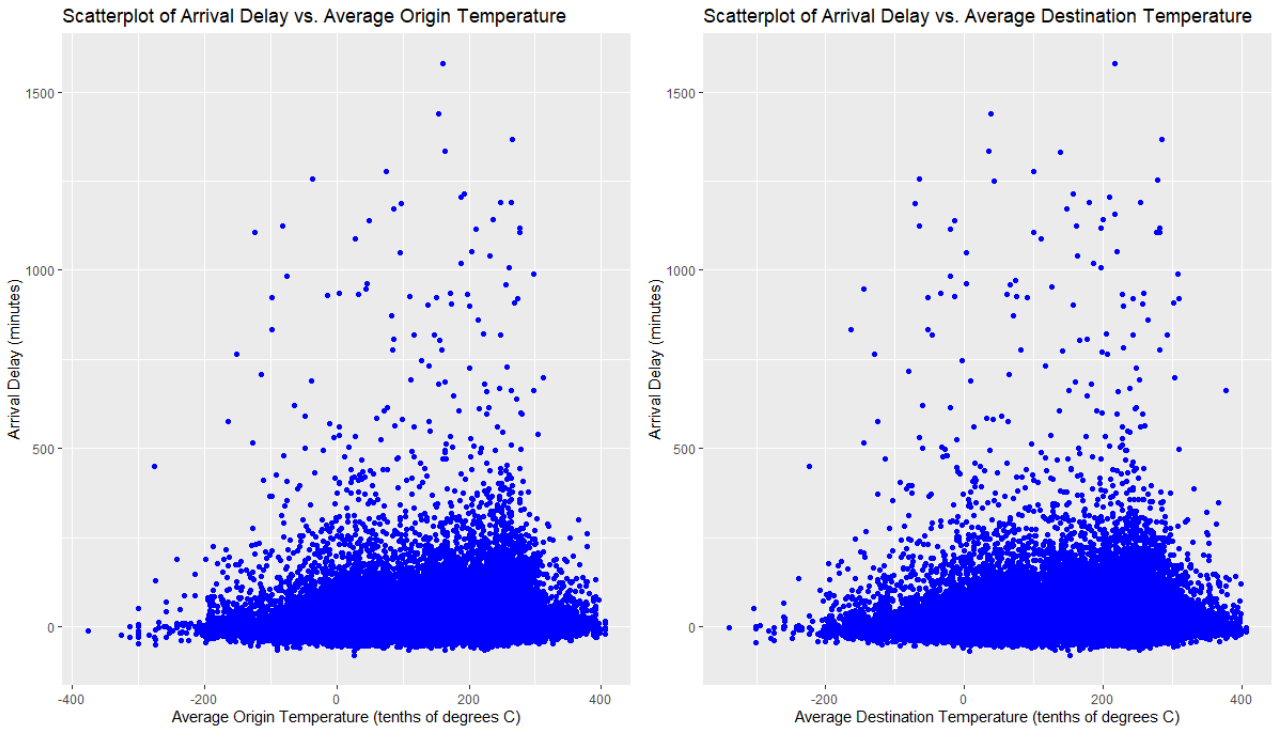


Figure 9.16: Delays vs. Temperature

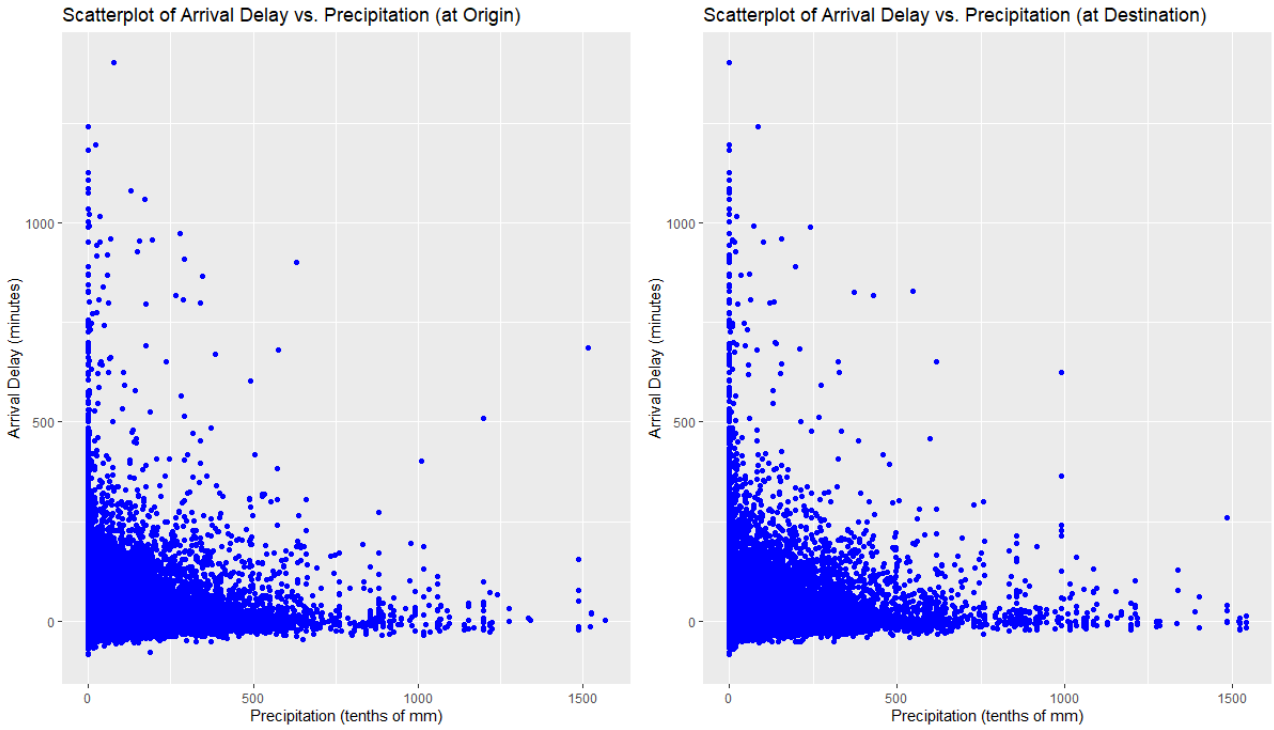


Figure 9.17: Delays vs Precipitation

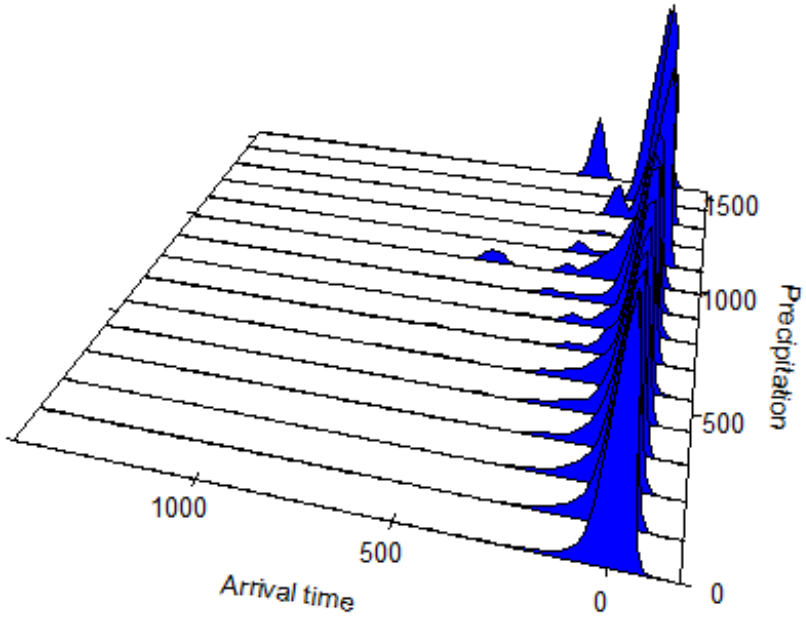


Figure 9.18: Conditional densities across levels of precipitation

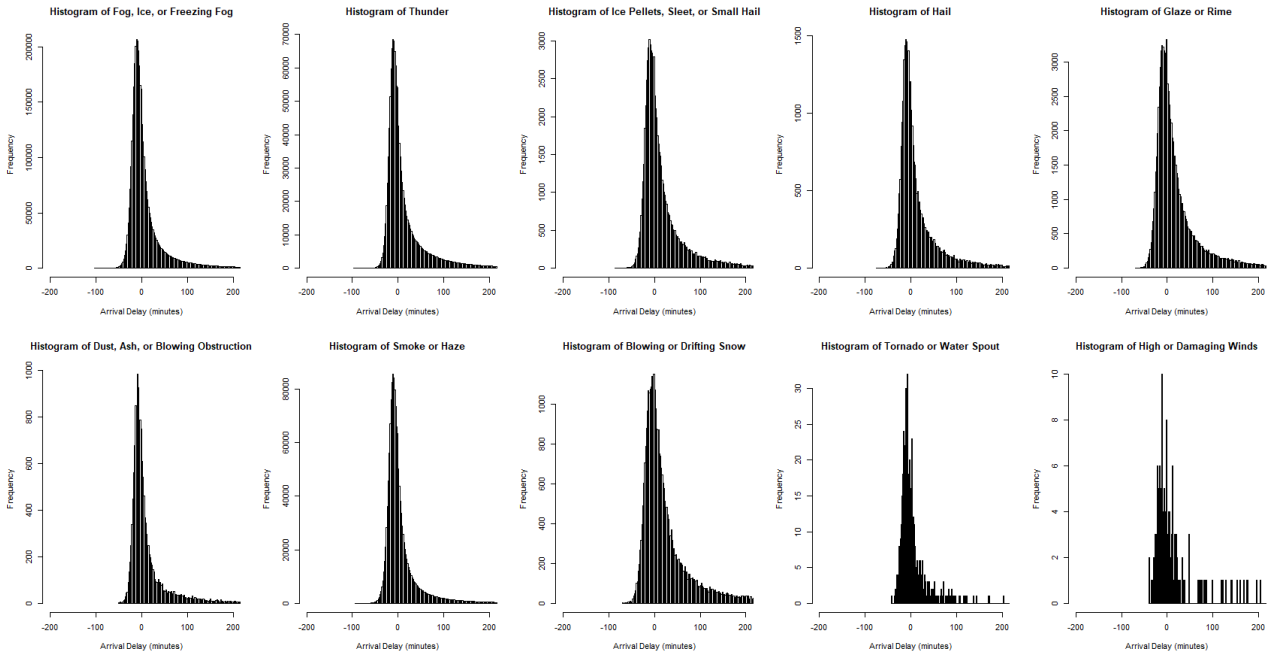


Figure 9.19: Delays by unique weather event

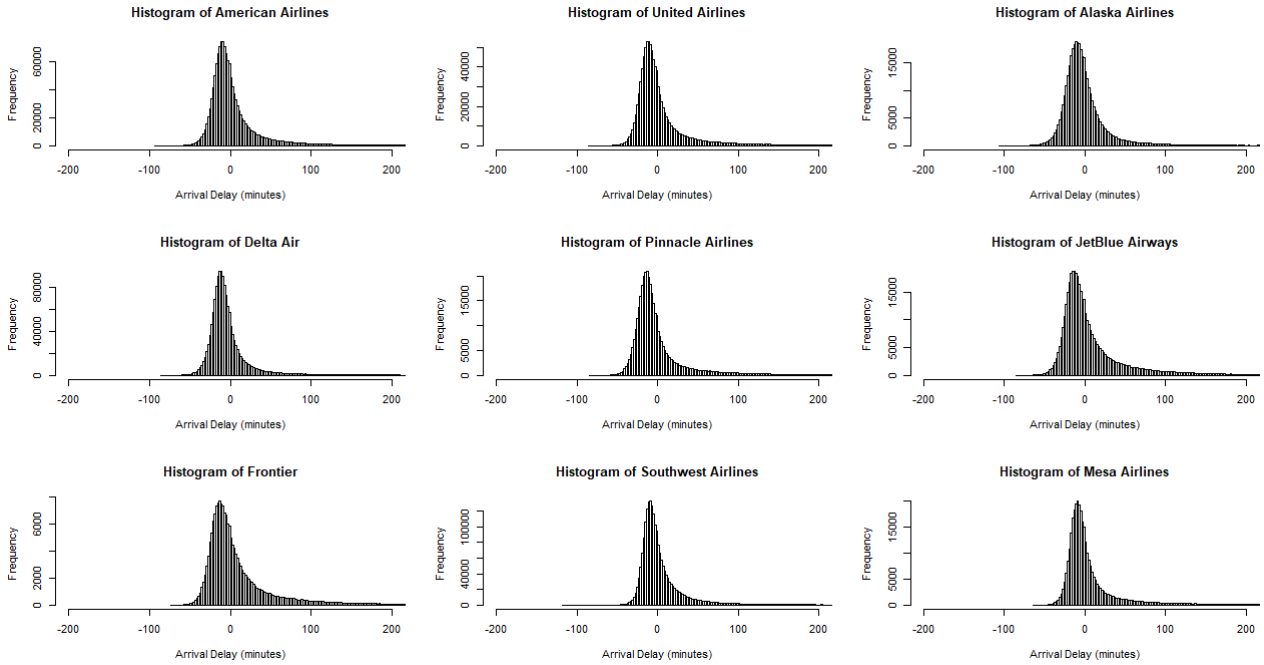


Figure 9.20: Delays by carrier

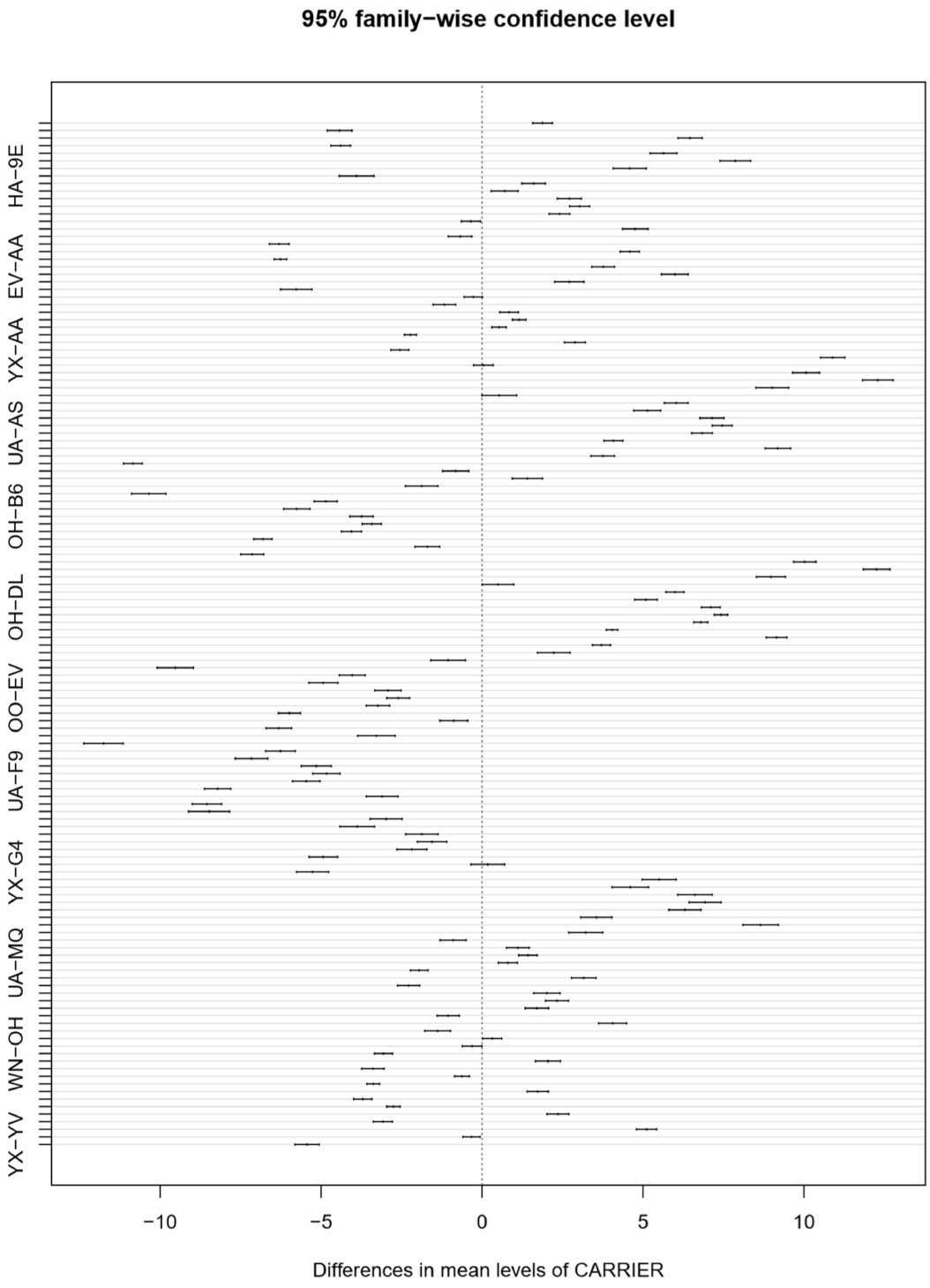


Figure 9.21: Tukey's HSD: 95% Confidence Intervals for the difference in mean delays across carriers

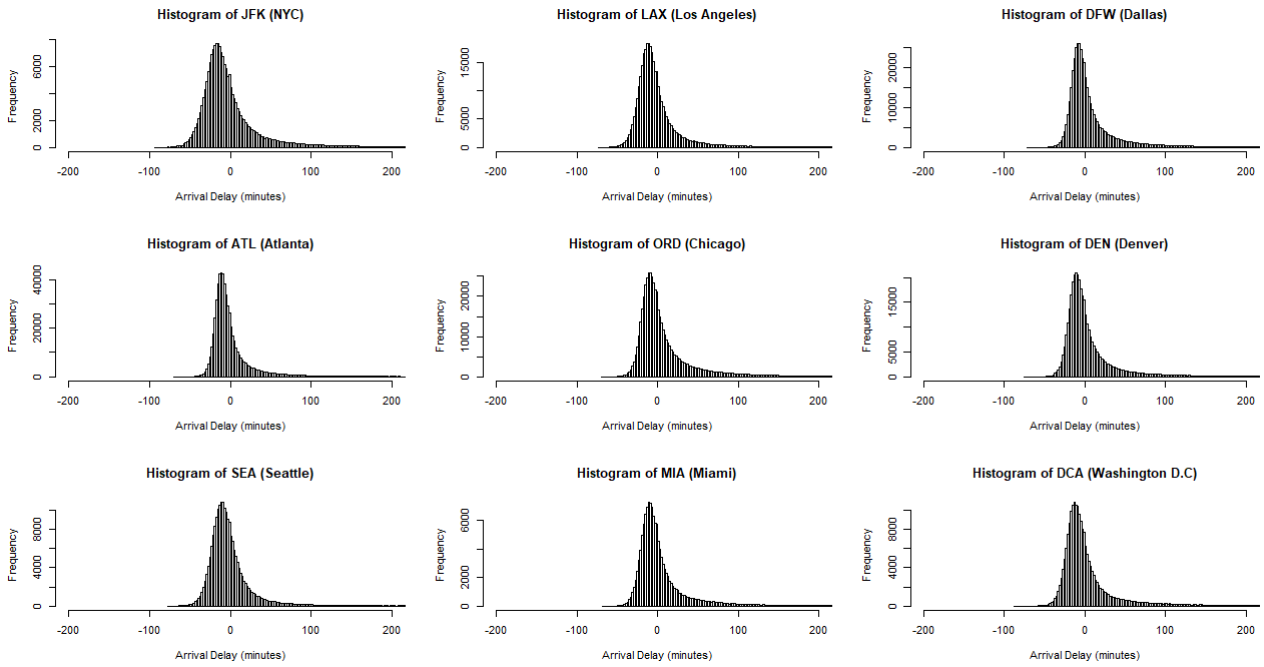


Figure 9.22: Delays by airport

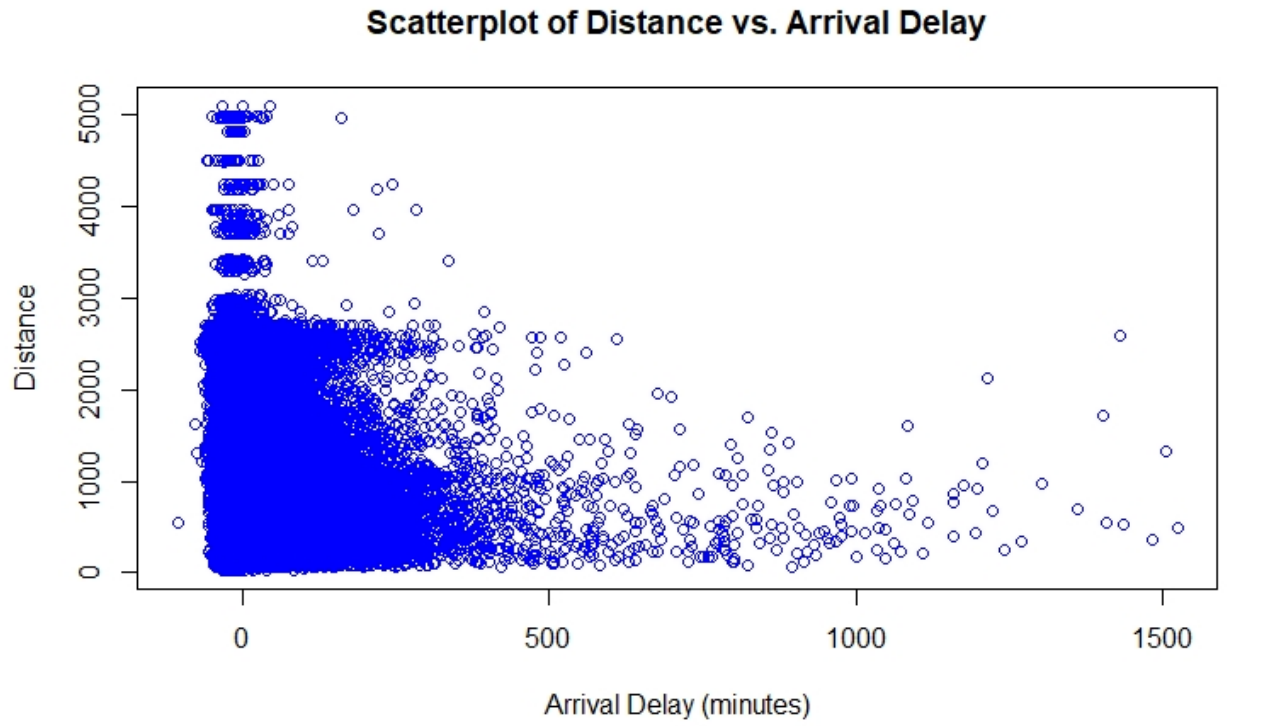


Figure 9.23: Do longer flights allow for lost departure time to be made up?

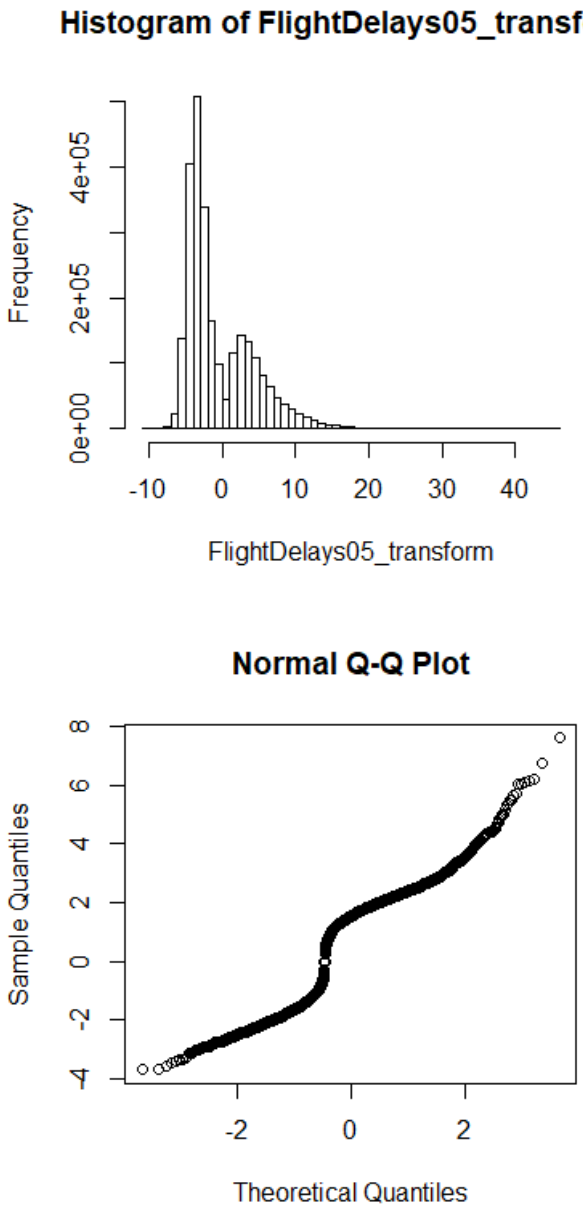


Figure 9.24: Top: Density plot of square root transformation vs normal distribution with same mean and variance. Bottom: QQ Plot of transformation against theoretical normal quantiles. We see that the transformation did not make the data normal.

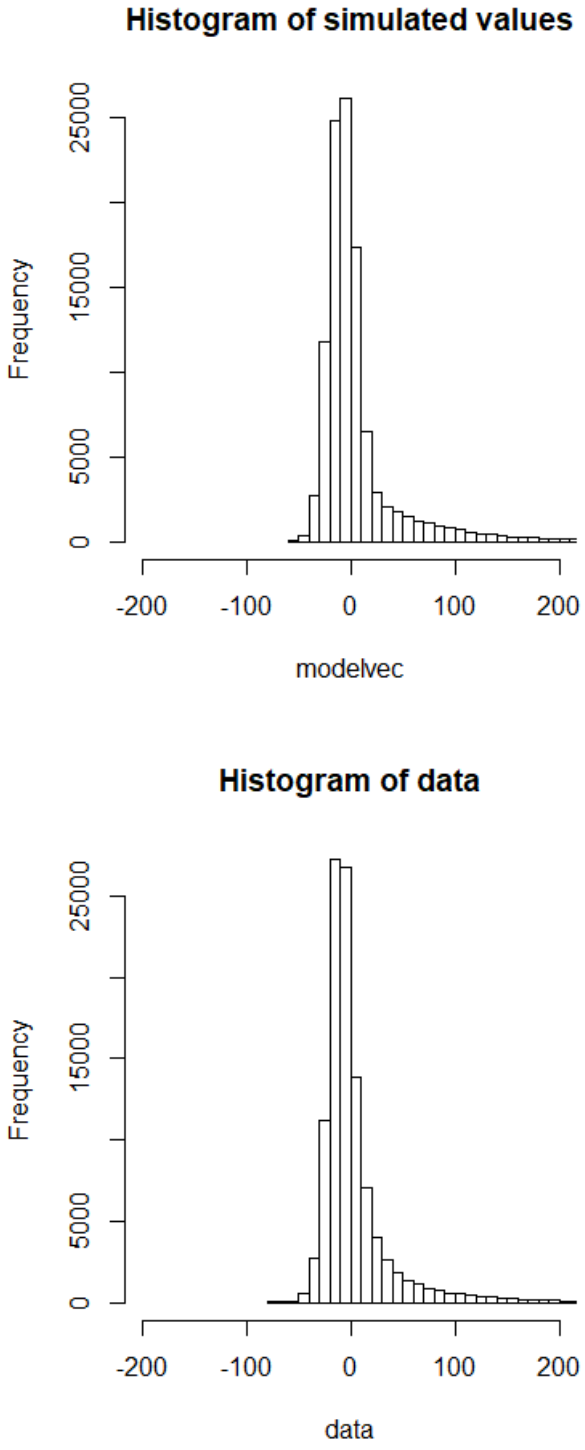


Figure 9.25: Top: Our estimated Lo-Ke distribution fit of marginal density. Bottom: Distribution of the actual marginal data. We see that our theoretical Lo-Ke distribution with estimated parameters very well captures the empirical distribution.

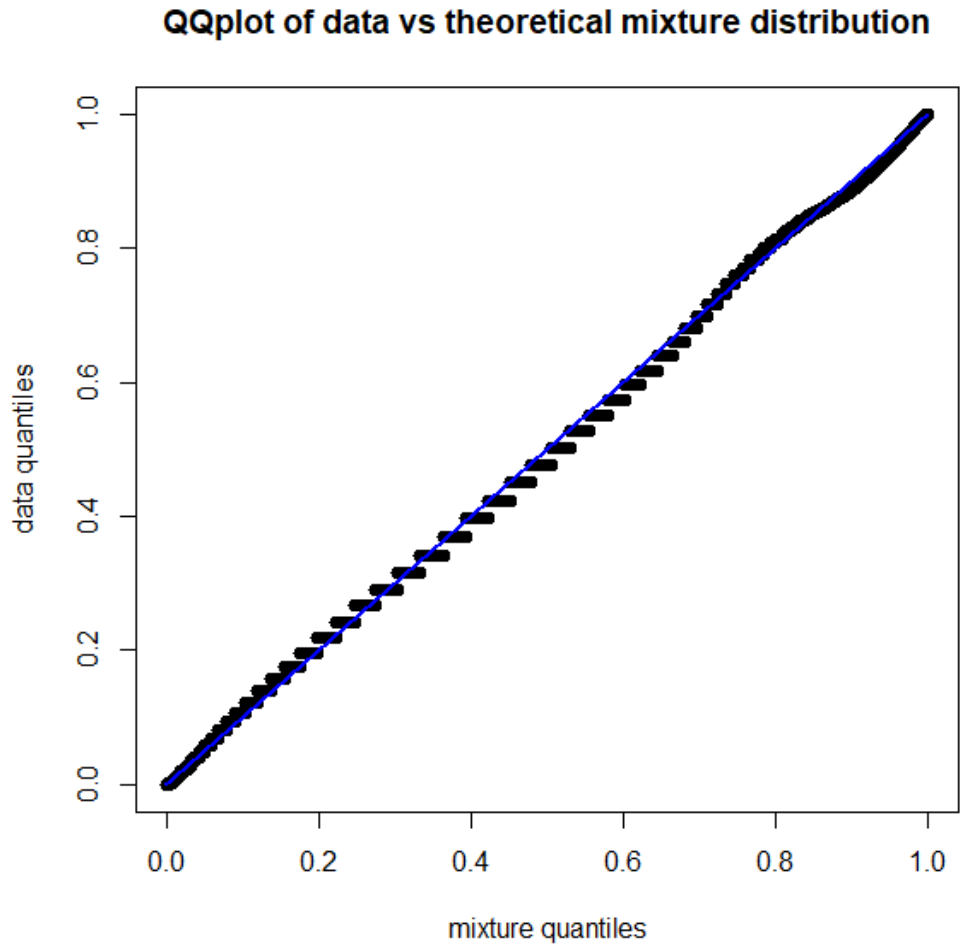
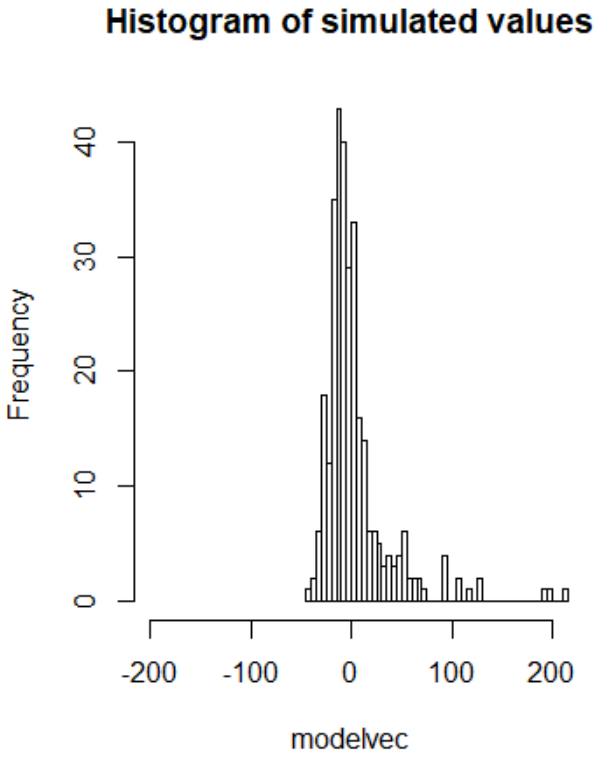


Figure 9.26: QQ plot of data vs theoretical Lo-Ke distribution. As the points follow the line y=x, we see that our created mixture distribution almost exactly matches the distribution of the data.



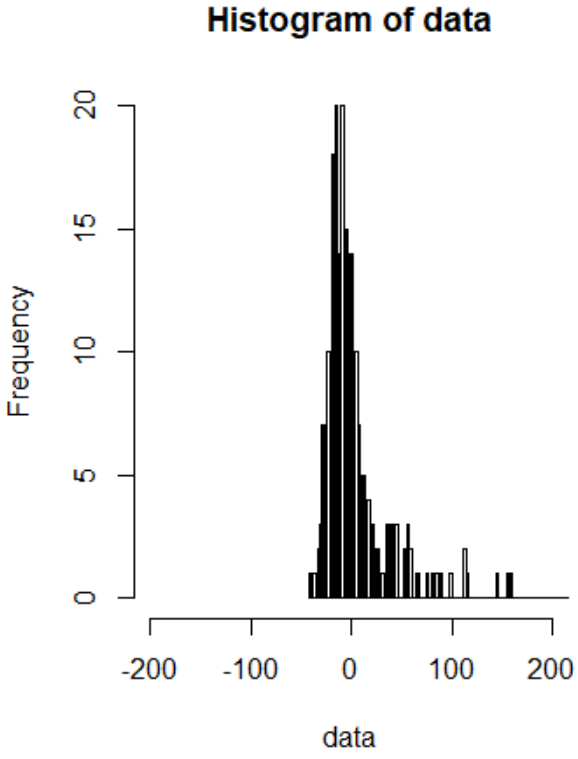


Figure 9.27: Top: Our Lo-Ke distribution estimate of the marginal distribution for ABE airport. Bottom: The actual distribution of arrival delays for ABE airport in the dataset.

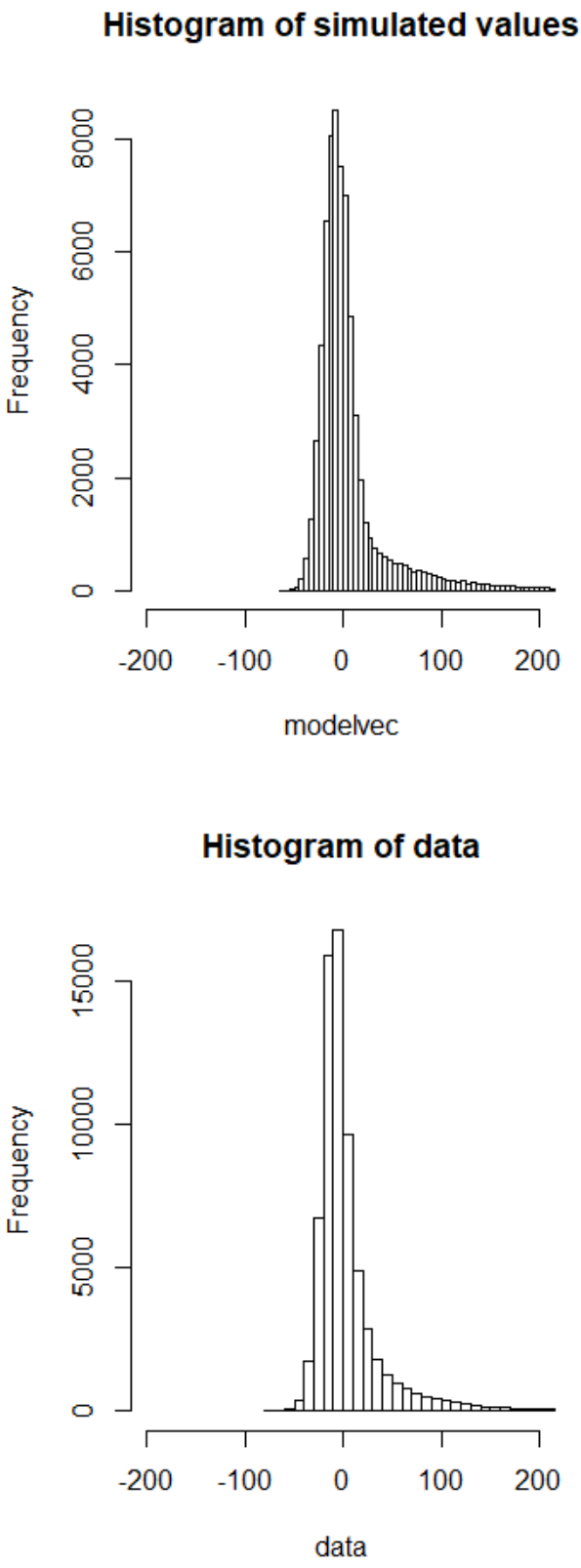


Figure 9.28: Top: Our Lo-Ke distribution estimate of the marginal distribution for American Airlines. Bottom: The actual distribution of arrival delays for American Airlines in the dataset.

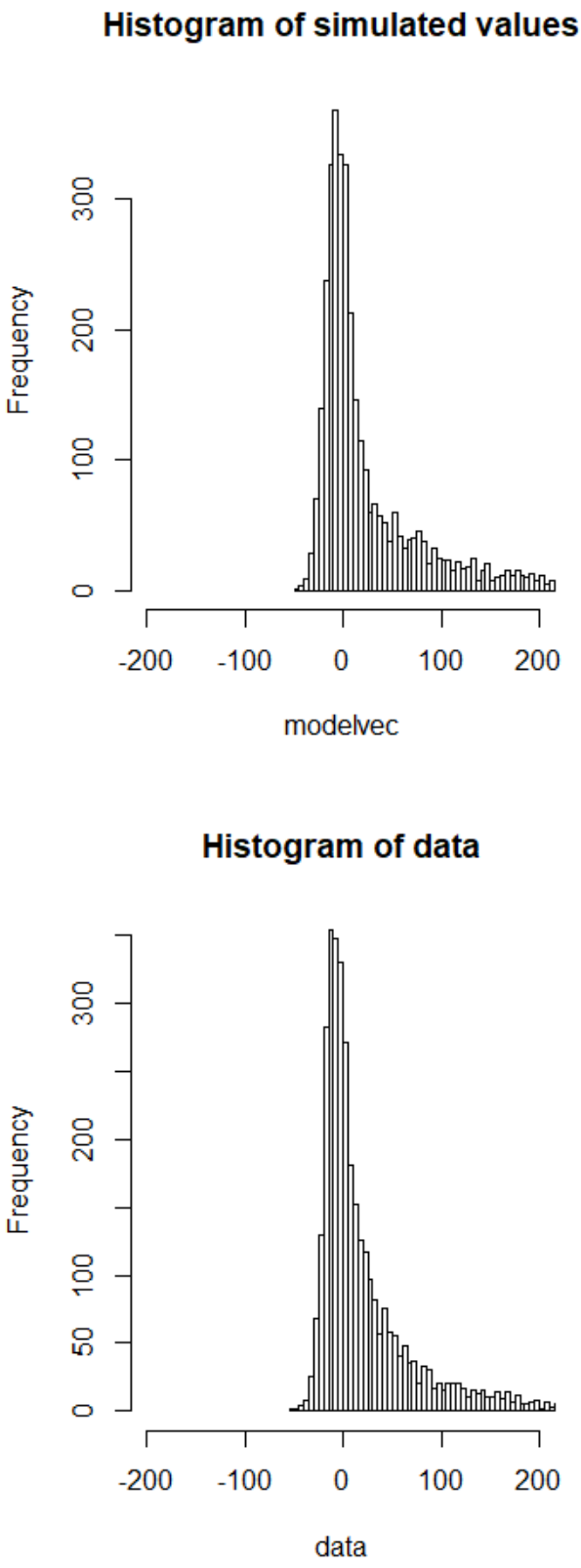


Figure 9.29: Top: Our Lo-Ke distribution estimate of the marginal arrival delay distribution for precipitation. Bottom: The actual distribution of arrival delays for precipitation in the dataset.

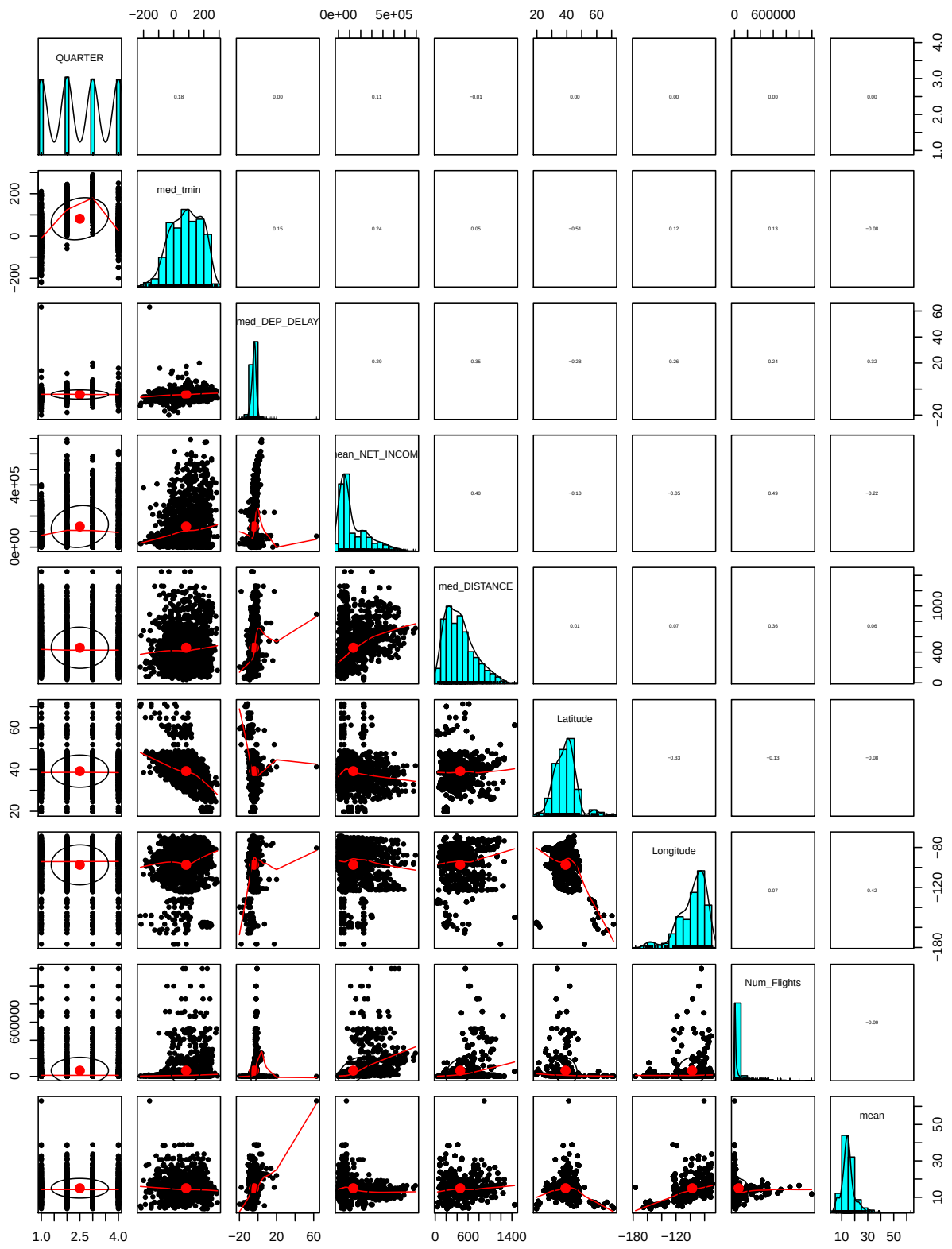


Figure 9.30: Scatterplot matrix of some of the OLS model covariates.

```
Call:
glm(formula = ARR_DELAY_NEW ~ YEAR + QUARTER + MONTH + DISTANCE +
    CRS_DEP_TIME + CRS_ARR_TIME + prcp + snow + tavg + tmax +
    tmin + wt01 + wt02 + wt03 + wt04 + wt05 + wt06 + wt07 + wt08 +
    wt09 + wt10 + wt01.DEST + wt02.DEST + wt03.DEST + wt04.DEST +
    wt05.DEST + wt06.DEST + wt07.DEST + wt08.DEST + wt09.DEST +
    wt10.DEST + wt11.DEST, family = "binomial", data = weather_training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9860  -0.9417  -0.7824   1.2917   1.9849

Coefficients: (3 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.789e+00  2.061e-02 -86.785  < 2e-16 ***
YEAR2019     1.771e-02  8.724e-03   2.030 0.042388 *
QUARTER2     3.384e-01  1.808e-02  18.715  < 2e-16 ***
QUARTER3    -5.200e-02  2.166e-02  -2.401 0.016351 *
QUARTER4     8.784e-02  1.915e-02   4.587 4.50e-06 ***
MONTH2       1.080e-01  1.597e-02   6.766 1.32e-11 ***
MONTH3       4.350e-02  1.556e-02   2.795 0.005182 **
MONTH4      -2.646e-01  1.561e-02 -16.943  < 2e-16 ***
MONTH5      -2.122e-01  1.459e-02 -14.544  < 2e-16 ***
MONTH6              NA         NA      NA       NA
MONTH7       3.672e-01  2.101e-02  17.478  < 2e-16 ***
MONTH8       3.174e-01  2.089e-02  15.197  < 2e-16 ***
MONTH9              NA         NA      NA       NA
MONTH10      1.509e-02  2.123e-02   0.711 0.477282
MONTH11      9.208e-02  2.094e-02   4.397 1.10e-05 ***
MONTH12             NA         NA      NA       NA
DISTANCE     9.370e-05  5.847e-06  16.025  < 2e-16 ***
CRS_DEP_TIME 4.218e-04  9.835e-06  42.892  < 2e-16 ***
CRS_ARR_TIME 1.833e-04  9.194e-06  19.940  < 2e-16 ***
prcp         1.136e-03  4.482e-05  25.338  < 2e-16 ***
snow         1.283e-02  4.741e-04  27.052  < 2e-16 ***
tavg         3.171e-03  2.993e-04  10.593  < 2e-16 ***
tmax        -2.564e-03  1.695e-04 -15.121  < 2e-16 ***
tmin        -1.032e-03  1.822e-04  -5.665 1.47e-08 ***
wt011        1.268e-01  8.750e-03  14.493  < 2e-16 ***
wt021        1.139e-01  1.764e-02   6.455 1.08e-10 ***
wt031        3.112e-01  1.168e-02  26.637  < 2e-16 ***
wt041        1.809e-01  4.224e-02   4.283 1.84e-05 ***
wt051        8.103e-02  5.948e-02   1.362 0.173120
wt061        5.035e-01  3.817e-02  13.190  < 2e-16 ***
wt071        3.611e-01  8.475e-02   4.262 2.03e-05 ***
wt081       -1.288e-02  1.043e-02  -1.235 0.216786
wt091        4.368e-01  6.170e-02   7.079 1.45e-12 ***
wt101        3.186e-01  4.673e-01   0.682 0.495330
wt01.DEST1   2.457e-01  8.186e-03  30.018  < 2e-16 ***
wt02.DEST1   1.052e-01  1.773e-02   5.935 2.94e-09 ***
wt03.DEST1   3.217e-01  1.130e-02  28.457  < 2e-16 ***
wt04.DEST1   2.225e-01  4.259e-02   5.225 1.74e-07 ***
wt05.DEST1   2.090e-01  6.328e-02   3.302 0.000959 ***
wt06.DEST1   1.171e-01  3.968e-02   2.951 0.003165 **
wt07.DEST1   2.508e-01  9.195e-02   2.728 0.006376 **
wt08.DEST1  -3.975e-02  1.053e-02  -3.775 0.000160 ***
wt09.DEST1   6.063e-01  5.838e-02  10.385  < 2e-16 ***
wt10.DEST1  -5.519e-01  4.478e-01  -1.232 0.217839
wt11.DEST1   5.033e-01  8.339e-01   0.604 0.546124
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 484413  on 371030  degrees of freedom
Residual deviance: 466603  on 370989  degrees of freedom
  (61647 observations deleted due to missingness)
AIC: 466687

Number of Fisher Scoring iterations: 4
```

Figure 9.31: R output for logistic regression

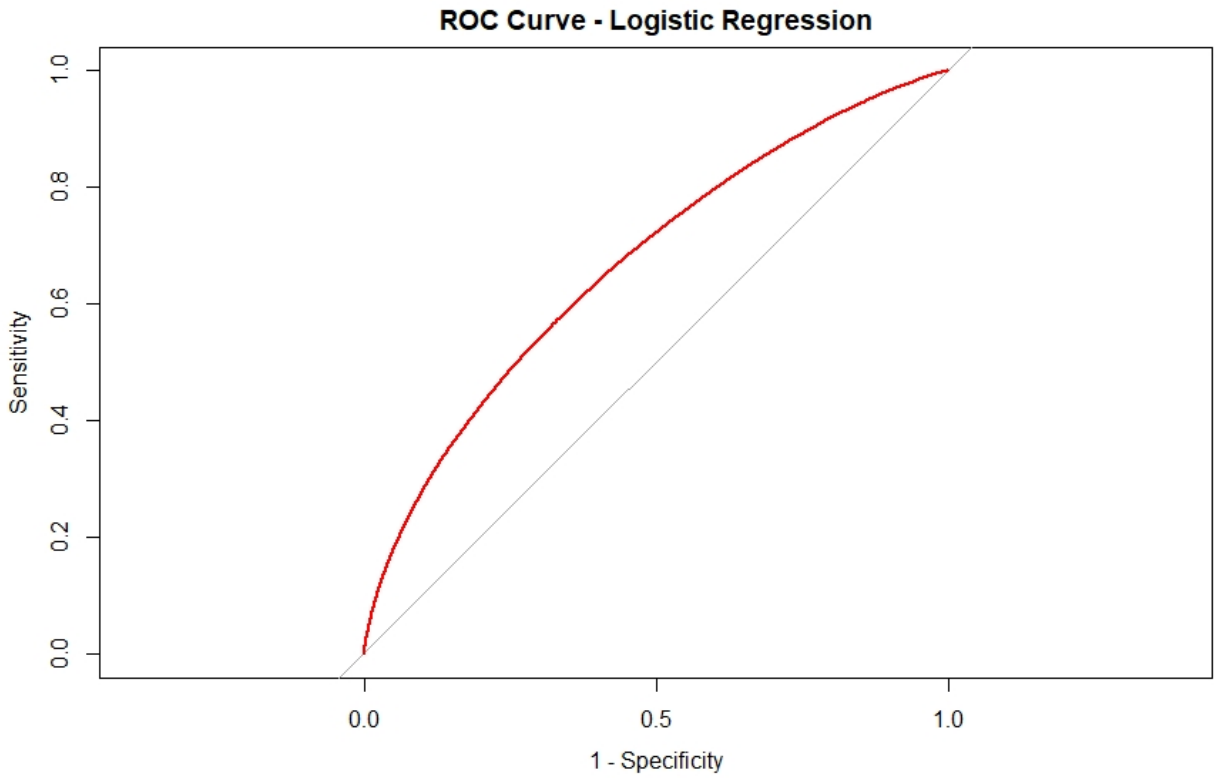


Figure 9.32: Receiver Operating Characteristic curve to determine probability cutoff threshold for logistic regression

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 53256  8931
         1 21076  9404

               Accuracy : 0.6762
                 95% CI : (0.6732, 0.6792)
    No Information Rate : 0.8021
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1836

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7165
            Specificity : 0.5129
         Pos Pred Value : 0.8564
         Neg Pred Value : 0.3085
             Prevalence : 0.8021
         Detection Rate : 0.5747
   Detection Prevalence : 0.6711
      Balanced Accuracy : 0.6147

       'Positive' Class : 0
```

Figure 9.33: Confusion matrix and other performance measures for logistic regression

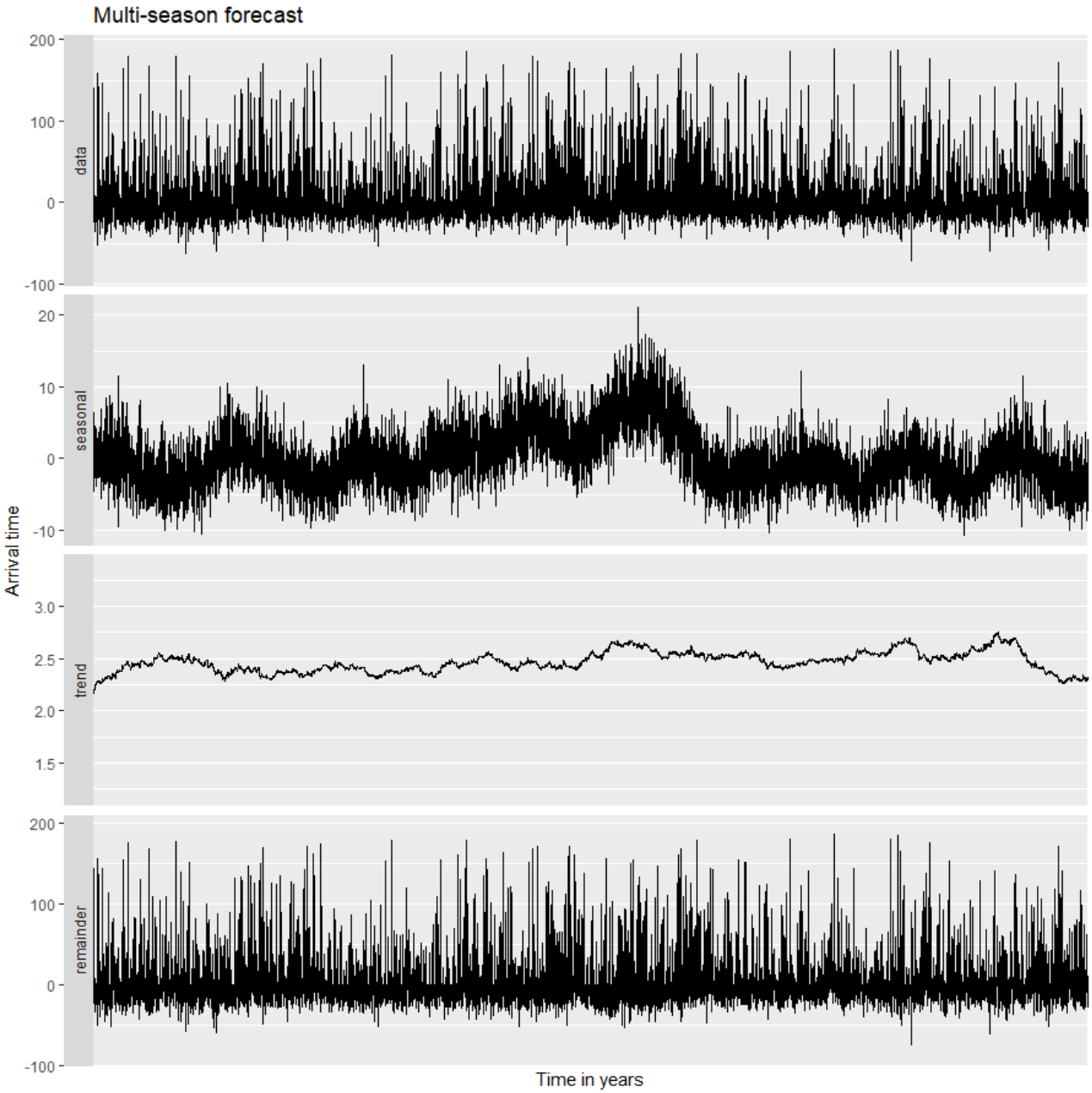


Figure 9.34: Multiseasonal time series decomposition for arrival delay using bootstrap method

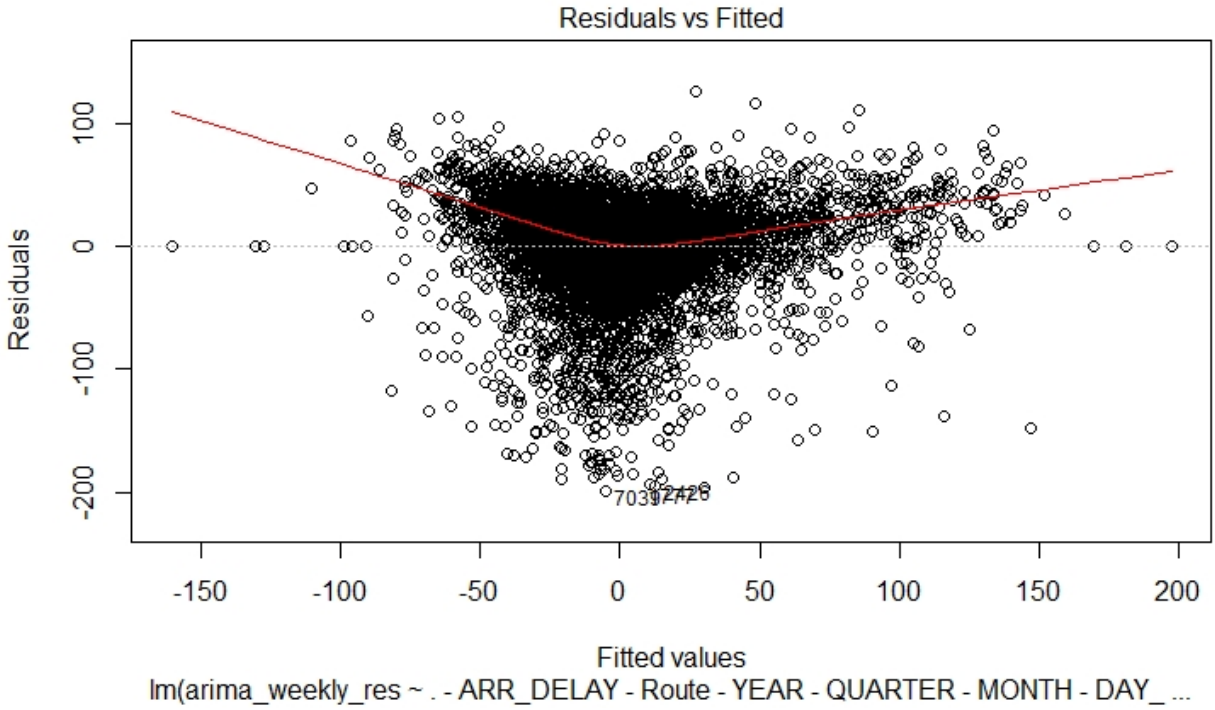


Figure 9.35: Plot of residuals vs. fitted values for our linear model of the residuals of our ARIMA model vs. the non-temporal covariates

```
Call:
lm(formula = arima_weekly_res ~ . - ARR_DELAY - Route - YEAR -
    QUARTER - MONTH - DAY_OF_MONTH - DAY_OF_WEEK - DEP_TIME_BLK -
    ARR_TIME_BLK, data = subset_residuals)

Residuals:
     Min       1Q   Median       3Q      Max
-205.035  -10.612    3.979   18.008  126.899


Residual standard error: 33.85 on 49234 degrees of freedom
Multiple R-squared:  0.3787,    Adjusted R-squared:  0.369
F-statistic: 39.23 on 765 and 49234 DF,  p-value: < 2.2e-16
```

Figure 9.36: R output for our linear model of the residuals of our ARIMA model vs. the non-temporal covariates

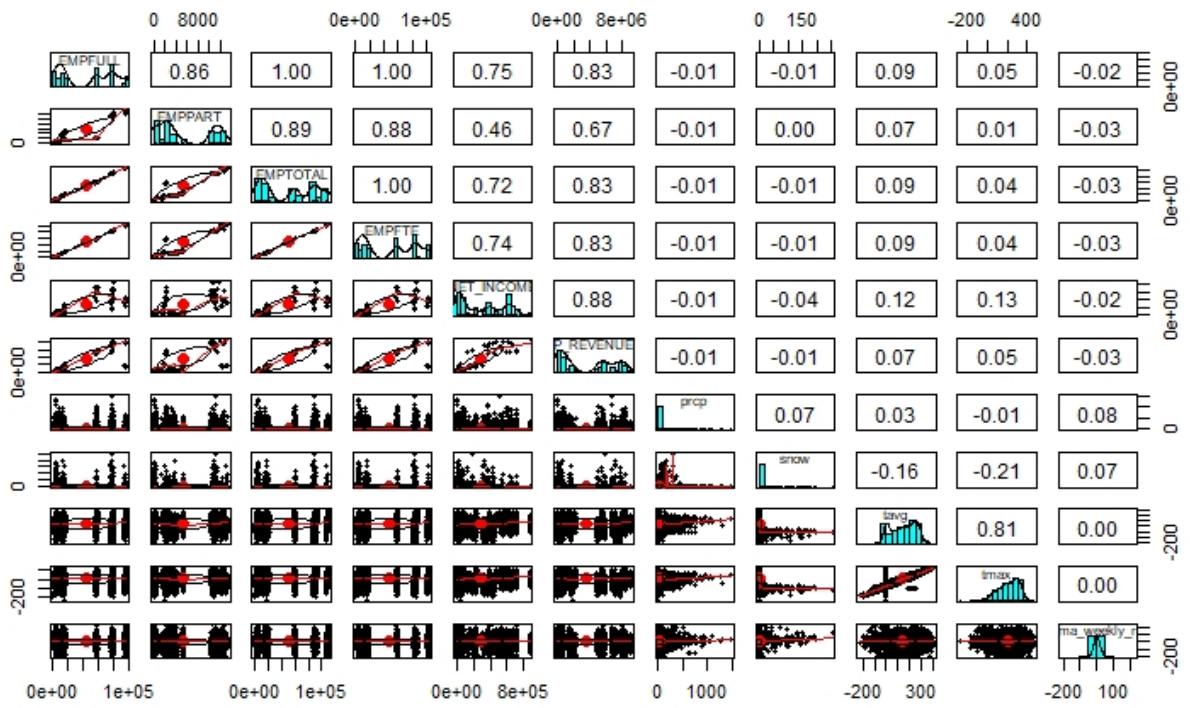


Figure 9.37: Scatterplot matrix for our dynamic linear model where we gained intuition that we should transform the precipitation covariate or provide weights for our model that involve this predictor

```
Call:
lm(formula = arima_weekly_res ~ . - ARR_DELAY - Route - YEAR -
    QUARTER - MONTH - DAY_OF_MONTH - DAY_OF_WEEK - DEP_TIME_BLK -
    ARR_TIME_BLK, data = subset_residuals, weights = wts)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-786.12  -14.07    4.96   23.62  427.57


Residual standard error: 59.46 on 49234 degrees of freedom
Multiple R-squared:  0.4902,    Adjusted R-squared:  0.4823
F-statistic: 61.88 on 765 and 49234 DF,  p-value: < 2.2e-16
```

Figure 9.38: R output for our weighted linear model (using weights as the square root of precipitation)

Bibliography

[America] America, Airlines F.: *Annual U.S. Impact of Flight Delays (NEXTOR report).* https://www.airlines.org/data/annual-u-s-impact-of-flight-delays-nextor-report/

[Athanasopoulos 2018] Athanasopoulos, Rob J Hyndman & G.: *Forecasting Principles and Practice.* OTexts, 2018

[Borders] Borders, Aviation Benefits B.: *Economic Growth.* https://aviationbenefits.org/economic-growth/

[Duffin 2019] Duffin, Erin: *Value added to U.S. GDP, by industry 2018.* https://www.statista.com/statistics/247991/value-added-to-the-us-gdp-by-industry/. Version: September 2019

[NOAA] NOAA: *Climate Data Online: Web Services Documentation.* https://www.ncdc.noaa.gov/cdo-web/webservices/v2

[OpenFlights] OpenFlights: *Airport, airline and route data.* https://openflights.org/data.html

[RStudio] RStudio: *RStudio.* https://rstudio.com/products/rstudio/