Estimating Flight Delay Time between RDU and MIA Using Historical Carrier Data and Weather Conditions

Martin Olarte

2023-03-06

Introduction

The aviation industry is one of the most critical transportation sectors, providing fast and convenient means of travel for people and goods. On-time performance of flights is a critical aspect of the aviation industry that affects the satisfaction of customers and the reputation of the airlines. The aim of this project is to estimate the delays of flights between RDU (Raleigh-Durham International Airport) and MIA (Miami International Airport) using carrier on-time performance data from the Bureau of Transportation Statistics and weather data from NCDC NOAA datasets.

The rationale behind this project is to provide insights into the factors that contribute to flight delays and to develop a model that can predict the delay of flights. This information can be useful for both airlines to plan their operations and for customers to make informed decisions about their travel plans.

Study Aims

Aim 1

1) To develop a model that predicts the delay of flights between RDU and MIA based on historical carrier performance, scheduled departure time, and weather conditions.

Aim 2

2) To identify the factors that contribute to flight cancellations, such as weather conditions or specific seasonal trends.

Primary Hypotheses

The primary hypothesis is that harsh weather conditions typically result in longer delay times and cancellations for flights between RDU and MIA.

Secondary Hypotheses

The secondary hypotheses are that the relationship between flight delay time / cancellation and weather condition may vary depending on:

- Departure time of day (e.g. early morning vs. night time)
- Flight date (hinting towards seasonality effects)
- Airline-specific factors

Detailed Data and Study Description

Primary Outcome

Table 1: Primary Outcomes

Outcome	Description	Variable Name and Source	Specifications
Departure Delay	Departure delay (minutes), defined as actual departure time - CRS (Computer Reservation System) departure time	DEP_DELAY (carrier.csv)	Minutes (continuous and can be negative)
Cancellation Code	Cancellation codes used by the Bureau of Transportation Statistics (BTS)	CANCELLATION_CODE (carrier.csv)	A=Carrier Caused, B=Weather, C=National Aviation System, D=Security

Additional Variables of Interest

Table 2: Additional Variables of Interest

Variable	Description	Variable Name and Source	Specifications
Unique Carrier Code	Unique carrier code used to identify airlines	OP_UNIQUE_CARRIER (carrier.csv)	Character
Scheduled Departure Time	CRS (Computer Reservation System) departure time	CRS_DEP_TIME (carrier.csv)	Military time (integer from 0 (midnight) to 2359 (11:59pm))
Average Wind Speed	Average daily wind speed (miles per hour)	AWND.<origin or dest> * (weather.csv)	Miles per hour (double)
Precipitation	Precipitation (inches)	PRCP.<origin or dest> * (weather.csv)	Inches (double)
Average Temperature	Average temperature (degrees Fahrenheit)	TAVG.<origin or dest> * (weather.csv)	Degrees Fahrenheit (integer)
Maximum Temperature	Maximum temperature (degrees Fahrenheit)	TMAX.<origin or dest> * (weather.csv)	Degrees Fahrenheit (integer)
Minimum Temperature	Minimum temperature (degrees Fahrenheit)	TMIN.<origin or dest> * (weather.csv)	Degrees Fahrenheit (integer)
Direction of fastest 2-minute wind	Direction of fastest 2-minute wind (degrees)	WDF2.<origin or dest> * (weather.csv)	Degrees (integer from 10 to 360 in intervals of 10)
Direction of fastest 5-minute wind	Direction of fastest 5-minute wind (degrees)	WDF5.<origin or dest> * (weather.csv)	Degrees (integer from 10 to 360 in intervals of 10)
Fastest 2-minute wind speed	Fastest 2-minute wind speed (miles per hour)	WSF2.<origin or dest> * (weather.csv)	Miles per hour (double)
Fastest 5-minute wind speed	Fastest 5-minute wind speed (miles per hour)	WSF5.<origin or dest> * (weather.csv)	Miles per hour (double)

*Data is available for both origin and destination locations on the same day (e.g. TMAX.origin and TMAX.dest)

The complete variable encoding documentation for **weather** data can be found here. Similarly, the complete variable encoding documentation for **carrier** data can be found here.

Inclusion Criteria

Flight data was downloaded on a monthly basis from January 2021 to November 2022 for the state of North Carolina only (includes arrival and departure data).

Exclusion Criteria

The time frame was chosen to gather enough data for statistical significance, but at the same time limiting confounding variables around pre-covid commercial flying patterns. Thus, earlier year data (1987-2019) is still available, but not considered for this analysis.

Study Design

This is an observational study of commercial flights. Each data point represents a flight at a specified date and time, and attached are core weather statistics for the flight day at both origin and destination airports. All available flights between RDU and MIA were selected, but weather information is less reliable due to potential instrumental/human errors, as well as accessibility to certain data points.

Methods

Data Cleaning and Processing

The carrier data was gathered from the BTS website directly, on a monthly basis from January 2021 to December 2022 for the state of North Carolina only (includes arrivals and departures). After binding all monthly data together, and filtering for only flights between RDU and MIA, the weather data (collected from the NCDC NOAA) was joined using the corresponding station IDs for both airports. Weather data was joined twice, once for the origin and once for the destination, and appropriate variable renaming was performed to avoid confusion. Next, irrelevant predictors that would not be available at the time of estimation (other response variables like whether the flight was diverted or arrival information) were immediately discarded for the purpose of this analysis.

After data wrangling, variable selection began by determining NA percentages within each column (12 columns had greater than 70% null values). The table below shows only numerical columns with more than 70% of observations with their respective summary statistics. Some of the weather data was extremely rare, making it difficult to implement in model fitting, and other data was insignificant (e.g. snowing patterns since it has barely ever snowed at the airport locations).

Table 3: Column Summary Statistics

	Minimum	1Q	Median	Mean	3Q	Maximum	Percent NA (%)
CRS_DEP_TIME	530.00	900.00	1408.00	1384.159	1830.00	2230.00	0.000
DEP_DELAY	-19.00	-5.00	-2.00	12.932	7.00	2619.00	2.163
DIVERTED	0.00	0.00	0.00	0.003	0.00	1.00	0.000
AWND.origin	0.45	4.47	6.49	6.846	8.72	20.13	0.000
PRCP.origin	0.00	0.00	0.00	0.155	0.06	5.26	0.000
SNOW.origin	0.00	0.00	0.00	0.004	0.00	1.60	24.202
SNWD.origin	0.00	0.00	0.00	0.008	0.00	2.00	26.161
TAVG.origin	25.00	63.00	75.00	70.297	80.00	88.00	0.000
TMAX.origin	30.00	74.00	83.00	79.656	89.00	102.00	0.000
TMIN.origin	15.00	52.00	68.00	61.882	74.00	84.00	0.000
WDF2.origin	10.00	90.00	130.00	154.247	230.00	360.00	0.000
WDF5.origin	10.00	90.00	140.00	159.598	230.00	360.00	0.522
WSF2.origin	6.90	14.10	16.10	17.039	19.90	47.00	0.000
WSF5.origin	10.10	19.00	21.90	23.309	27.10	64.00	0.522

Exploratory Analyses

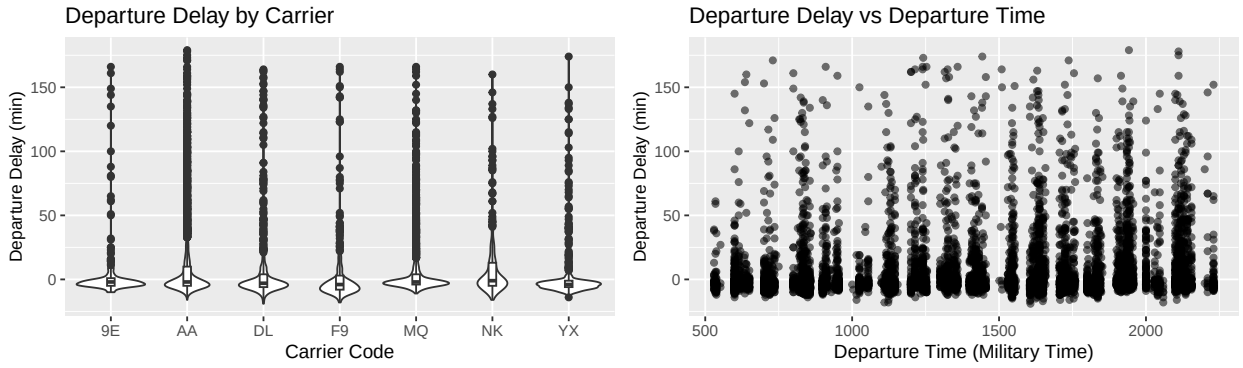

Figure 1: Response Variable EDA

The plots from Figure 1 and 2 are limited to data with a departure delay less than 3 hours or 180 minutes. Only 127 out of the total 7,859 flight observations have departure delays above 3 hours, which is only 1.62% of the data. Thus, it seemed reasonable to fit a model with a heavy-tailed distribution for the error terms. All carriers seem to have a similar highly skewed distribution of delay times, with center around 0 and a long tail towards longer delay times. However, there are differences in variance for outlying values particularly because 4,093 out of the 7,859 flights are from American Airlines. This further motivates the sensitivity analysis with respect to airline segregation for delays and cancellations.

From the EDA, it is evident that pairs of weather covariates from the destination and origin will be highly correlated due to the fact that measurements are daily summaries, which could be an issue looking forward but will be decided when addressing multicollinearity.

From the time series data in Figure 3, which could be interesting to explore further to identify seasonal or weekly changes, we can see that the outlying values are dominating the dataset, and even when filtering extremely for delays of less than 1 hour, fitting the default cubic spline GAM does not indicate any significant trend initially. Thus, this further motivates the use of a robust regression method.

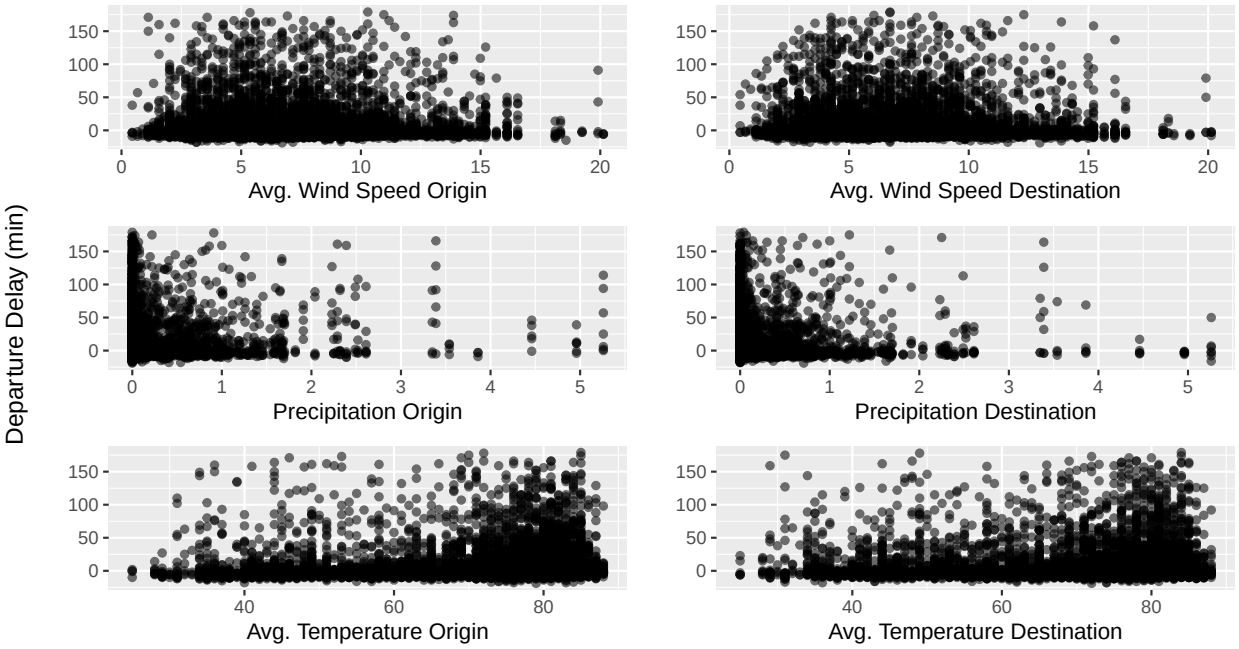

Figure 2: Predictors vs Response Variable

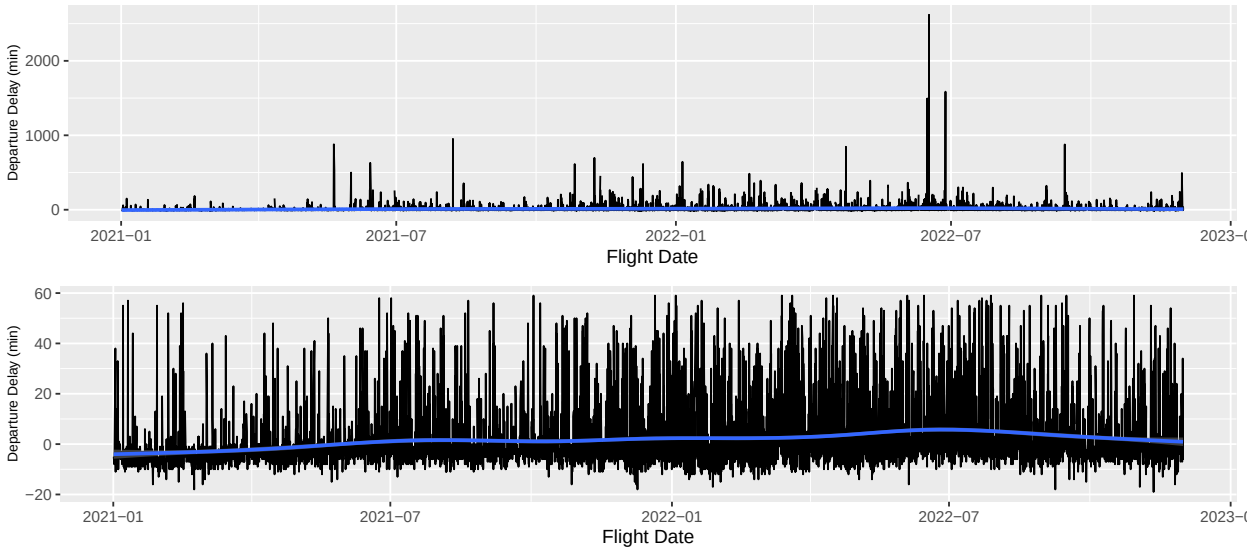

Figure 3: Time Series Plots

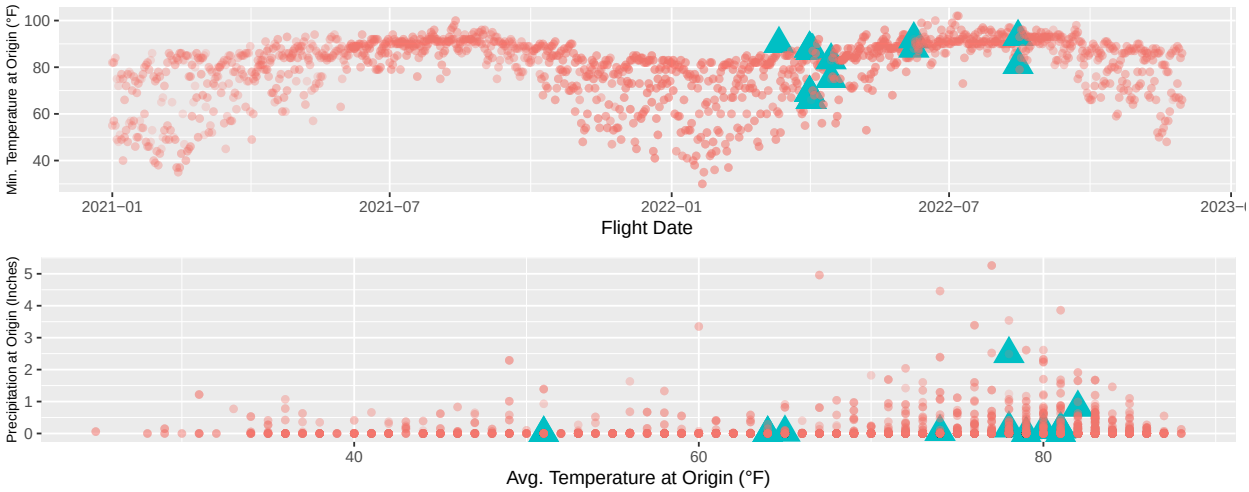

Figure 4: Flight Cancellations

Finally, regarding flight cancellation trends, there were a total of 172 cancelled flights in the dataset, and only 17 of those were weather related (Code C). From Figure 4 it is evident that the cancelled flights due to weather (represented by blue triangles) were only present after 2022 in the data, and there seems to be a correlation with temperature. However, precipitation is not seen as a particularly strong predictor as expected.

Aim 1 - Student's T robust linear regression model for prediction

To achieve Aim 1, we fit a Student's T robust linear regression model using the `rlm()` function from the MASS package (Venables and Ripley (2002)) on only the data from January 2021 to December 2021. The fitted model equation is the same as a traditional linear model ($y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip} + e_i = \mathbf{x}_i^\top \boldsymbol{b} + e_i$). However, this model uses different estimators (M-estimators and MM-estimators (Susanti et al. (2014))) to adjust the weight of each point in an iterative process. Essentially, M-estimators minimize an objective function ($\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^\top \boldsymbol{b})$), where ρ denotes the contribution of each residual. The estimating equations produced by this method can be written as a weighted least-squares problem. However, solving these estimating equations requires an iterative solution, known as iteratively reweighted least-squares (IRLS), because the weights depend on the residuals and the residuals depend on the estimated coefficients (See Appendix A for a detailed description). This robust estimation is much less sensitive to outliers than the traditional Ordinary Least Squares (OLS) method, making it useful for our dataset which had a large number of outlying values for the response variable (Departure Delay). Moreover, the Student's T distribution has a heavy tail to accommodate outlying errors.

The objective and weight functions for four popular M-estimators are given in Table 4: the familiar least-squares estimator; the conservative least absolute deviation estimator; the Huber estimator; and the Tukey bisquare (or biweight) estimator. The Huber and bisquare estimators rely on a tuning constant denoted by k, which can be adjusted to increase resistance to outliers. However, decreasing k to improve resistance to outliers comes at the cost of lower efficiency when the errors are normally distributed. Typically, the tuning constant is selected to provide reasonably high efficiency in the normal case. Specifically, for the Huber and bisquare estimators, a value of $k = 1.345\sigma$ and $k = 4.685\sigma$, respectively, will result in 95% efficiency for normally distributed errors while still offering protection against outliers. As the true value of σ is unknown, we use an estimate of the standard deviation of the errors. In our study, we followed previous literature (Fox (2002)) and adopted a robust measure of spread, namely $\sigma = \text{MAR}/0.6745$, where MAR represents the median absolute residual, instead of the standard deviation of the residuals.

Table 4: Objective and weight functions for least-squares, LAD (or L1), Huber, and bisquare estimators

Method	Objective Function	Weight Function
Least-Squares	$\rho_{LS}(e) = e^2$	$w_{LS}(e) = 1$
Least Absolute Deviations	$\rho_{LAD}(e) = \lvert e \rvert$	$w_{LAD}(e) = \frac{1}{\lvert e \rvert}$
Huber	$\rho_H(e) = \begin{cases} \frac{e^2}{2} & \text{for } \lvert e \rvert \le k \\ k\lvert e \rvert - \frac{k^2}{2} & \text{for } \lvert e \rvert > k \end{cases}$	$w_H(e) = \begin{cases} 1 & \text{for } \lvert e \rvert \le k \\ \frac{k}{\lvert e \rvert} & \text{for } \lvert e \rvert > k \end{cases}$
Bisquare	$\rho_B(e) = \begin{cases} \frac{k^2}{6}(1-(1-(\frac{e}{k})^2)^3) & \text{for } \lvert e \rvert \le k \\ \frac{k^2}{6} & \text{for } \lvert e \rvert > k \end{cases}$	$w_B(e) = \begin{cases} (1-(\frac{e}{k})^2)^2 & \text{for } \lvert e \rvert \le k \\ 0 & \text{for } \lvert e \rvert > k \end{cases}$

Figure 9 in Appendix B compares the objective functions, and the corresponding ψ and weight functions visually. Both the least-squares and Huber objective functions increase without bound as the residual e departs from 0, but the least-squares objective function increases more rapidly. In contrast, the bisquare objective function levels eventually levels off (for $|e| > k$). Least-squares assigns equal weight to each observation; the weights for the Huber estimator decline when $|e| > k$; and the weights for the bisquare decline as soon as e departs from 0, and are 0 for $|e| > k$. Furthermore, LAD has a clear asymptote at $e = 0$, making a weighting approach difficult. Therefore, we decided to focus on Huber and bisquare approaches.

We evaluated the model using the bootstrap method to examine the distribution of coefficient estimates to test hypotheses, which control Type I error probabilities relatively well even when there is heteroscedasticity. As noted in an article provided by Professor Ruczinski from Johns Hopkins Bloomberg School of Public Health, for such a robust model "the R-squared and F-statistics are not given because they cannot be calculated (at least not in the same way)", so "the bootstrap is a general purpose inferential method which is useful in these situations". Therefore, we evaluated the hypotheses by examining the bootstrap distribution of coefficient estimates, with 95% confidence intervals not containing 0 considered evidence of a significant effect.

In addition, we explored interaction terms using residuals to evaluate model fit, and conducted external validation using data from 2022. We also conducted a sensitivity analysis on airport-segregated data by fitting two individual models for RDU and MIA to determine if there were any airport-specific variations.

- $\text{RLM}_{\text{RDU}} : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i$
- $\text{RLM}_{\text{MIA}} : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i$

Finally, we planned to conduct a second sensitivity analysis that focused exclusively on estimating flight delays after the peak of the COVID-19 pandemic by fitting a model solely on data after April 18, 2022, when a federal judge in Florida struck down the U.S. federal transportation mask mandate and all of the major U.S. airlines lifted their pandemic-era mask requirements for domestic flights.

- $\text{RLM}_{\text{post-mask}} : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i$, where i represents flights after April 18, 2022.

Aim 2 - Logistic regression to identify the factors that contribute to flight cancellations

To address Aim 2, we aimed to model the relationship between flight cancellations and weather conditions through a logistic regression to help identify the most important predictors of cancellations. The response variable Cancellation Code is categorical, but we are only interested in weather-related cancellations (Code C), since the other types of cancellations are much more rare and have a sparse nature that makes it difficult to model (e.g. a national security alert). Thus, we removed such records and the response was converted to binary, where 1 represents flights cancelled due to weather and 0 represents any other flight that was not cancelled. We fit a logistic regression model in R software using the `glm()` function with a binomial family.

- The logistic regression model can be expressed as:
  - $logit(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$, where π_i is the probability of a flight being cancelled due to weather, x_{i1} through x_{ip} are predictor variables, and β_0 through β_p are coefficients.
- Since we are interested in interpreting the coefficient results, we will focus on the odds ratio for a predictor variable, which can be calculated as:
  - $OR = e^{\beta}$, where β is the coefficient for that predictor variable.

For our variable selection procedure, we opted for backward selection instead of LASSO regularization, since we were mainly concerned with inference and not necessarily the most parsimonious model. We used AIC and BIC as information criteria to decide whether to include a predictor or not. We checked modeled assumptions such as absence of multicollinearity through the `vif()` function from the `car` package, and to check a linear relationship between explanatory variables and the logit of the response variable we used scatter plots. To evaluate the models after fitting, we used standard goodness-of-fit measures such as the Hosmer-Lemeshow test and the ROC curve (See Appendix C).

Results

Aim 1 - Student's T robust linear regression model for prediction

We conducted a comparison between two robust regression methods (Huber's M-estimator and the biweight estimator) and the use of the inverse hyperbolic sine (IHS) transformation (Aihounton and Henningsen (2020)) to determine which model had better performance in predicting flight delays. The long-tailed errors motivated the use of IHS and the fact that it can transform right-skewed variables that include zero or negative values (compared to log transformations for example). The comparison was done based on two metrics, root mean squared error (RMSE) and mean absolute error (MAE), for both in-sample and out-of-sample predictions. Based on the results from Table 5, we selected the Huber Regression model with IHS transformation of the response for our final analysis as it demonstrated superior predictive performance for our dataset.

Table 5: Model Performance

	In Sample		Out of Sample	
Model	**RMSE**	**MAE**	**RMSE**	**MAE**
Huber Regression	48.58135	14.16398	74.09591	22.26105
Biweight Regression	49.08422	14.15642	74.95829	21.87369
Huber Regression	**48.57331**	**14.29754**	**74.32741**	**22.02406**
Biweight Regression	49.02252	14.13835	75.07267	21.86620

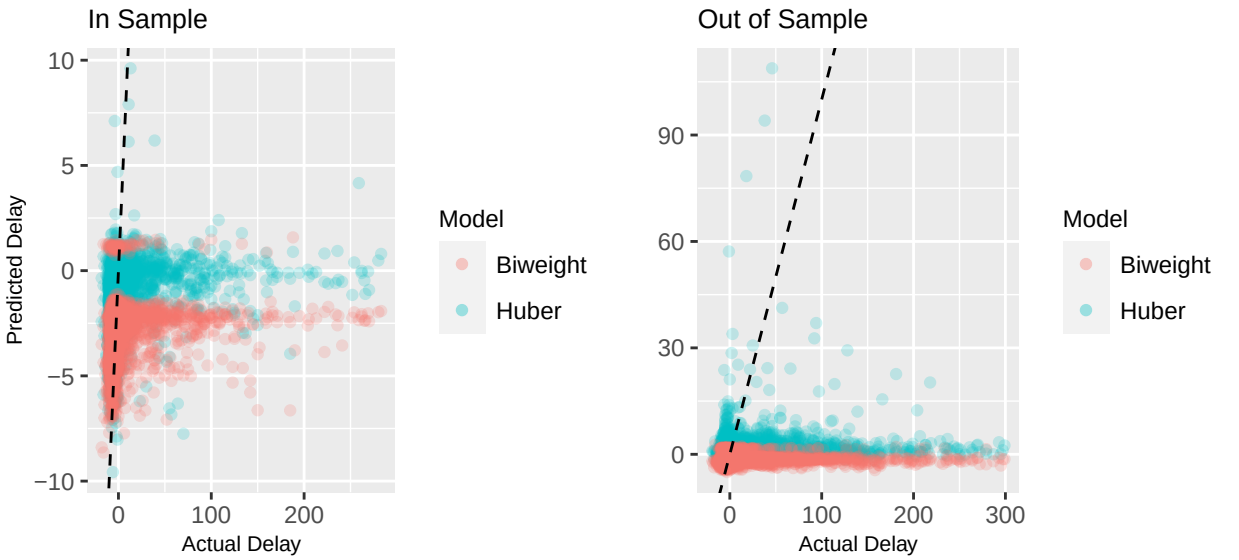

Figure 5: Huber vs Biweight M-estimator performance for Robust Regression

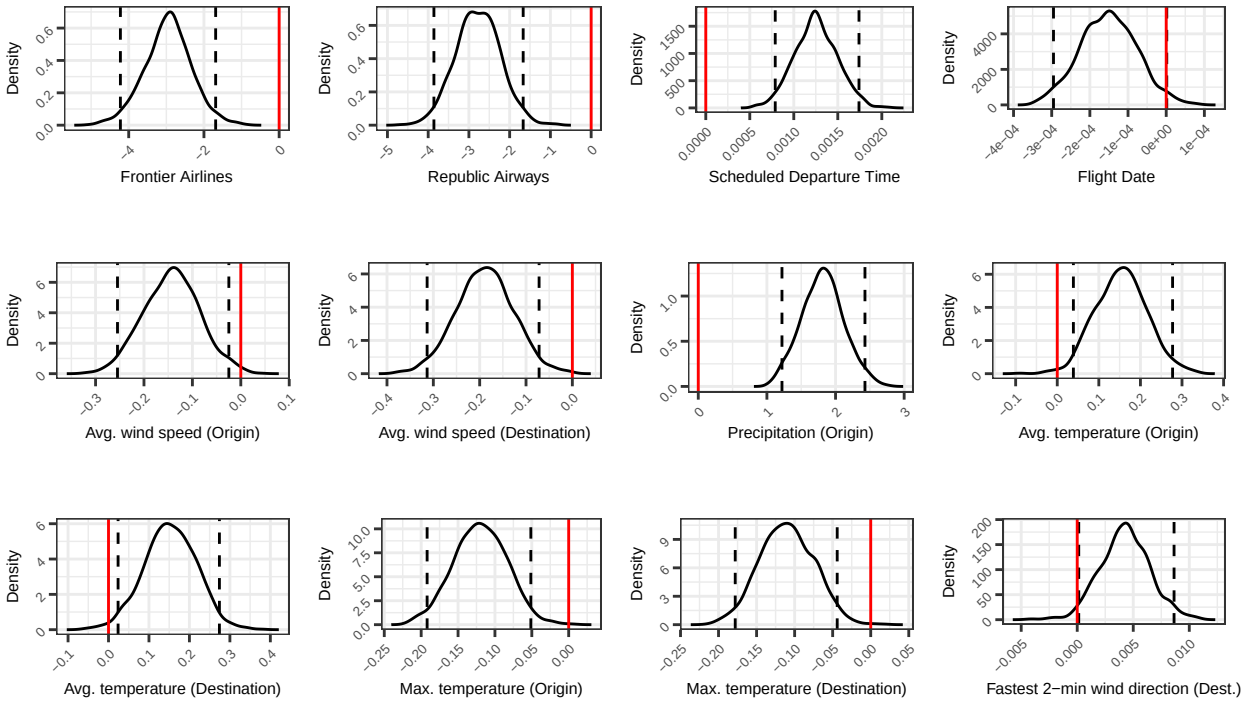

Figure 6: Bootstrap Distribution of Coefficient Estimates

We then applied the percentile bootstrap method for 10000 iterations for the Huber model, randomly sampling the residuals with replacement to determine values of the model coefficients. Figure 6 shows the effects determined to be significant since the constructed 95% empirical confidence intervals (shown as dashed lines) do not contain the value 0. For weather-related predictors, it is surprising to see that both average and maximum daily temperature at both origin and destination are significant, but minimum temperature is not. Moreover, the effect of average and maximum temperature seems contradictory, as average temperature has a positive effect, indicating that a flight during a day with a hotter average temperature tends to have longer delays, but maximum temperature has a negative effect, indicating that a flight during a day with a hotter maximum temperature tends to have shorter delays. Also, precipitation seems to only be influential at the origin of the flight. To further investigate these effects, we explored airport-segregated models, where one model was fit to only flights departing from RDU and one to flights departing from MIA.

All three models agreed on the significance of the coefficients associated to Republic Airways (YX), the flight's scheduled departure time, and the precipitation at the origin. The RDU model agreed with the overall model on the significance of the coefficients associated to Frontier Airlines (F9), the average wind speed at the destination, and the direction of the fastest 2-minute wind on the day of the flight, but had an airport-specific significance assigned to the precipitation at the destination and the direction of the fastest 5-minute wind on the day of the flight. On the other hand, the MIA model agreed on the significance of the coefficients associated to the average wind speed at the origin, the average temperature at the destination, and the maximum temperature at both the origin and destination, but had an airport-specific significance assigned to Spirit Airlines (NK) that was not in the general model.

When attempting to fit a robust linear regression model using data exclusively after April 18, 2022, it was found that the predictor matrix was singular, making it impossible to conduct this analysis. As a result, the post-covid discrepancy could not be examined, and no further conclusions could be drawn about flight delays after the peak of the pandemic. It is important to note that while the sensitivity analysis could not be conducted, the rest of the statistical report's results remain valid and can still provide valuable insights into flight delays during the pandemic. Further research may be necessary to investigate flight delays after the lifting of mask requirements.

Aim 2 - Logistic regression to identify the factors that contribute to flight cancellations

We first fit a logistic regression model on all the available predictors and tried to assess multicollinearity among them. As seen in Table 6, we used $GVIF^{\left(\frac{1}{2\times Df}\right)}$ to make GVIFs comparable across dimensions (Fox and Monette (1992)), and discovered that average temperature in both airports and maximum temperature at the destination airport are the three covariates that are most correlated. Thus, we decided to not include them in the final model.

Table 6: Model Diagnostics - Multicollinearity

	GVIF	**Df**	**GVIF**$^{\left(\frac{1}{2\times \mathbf{Df}}\right)}$
Avg. Temperature (Origin)	49.64793	1	7.046129
Avg. Temperature (Destination)	64.08707	1	8.005440
Max. Temperature (Destination)	27.56088	1	5.249846

The, after executing stepwise backwards selection using AIC, we assessed the model assumptions by examining the linear relationship between the logit of the response variable and predictors. This assumption did not strongly hold for some predictors, but the most insightful relationship was the one with the flight date. From Figure 7, one can see a moderately strong and positive linear relationship, indicating a slow increase on the odds of flight cancellations from 2021 to 2023.

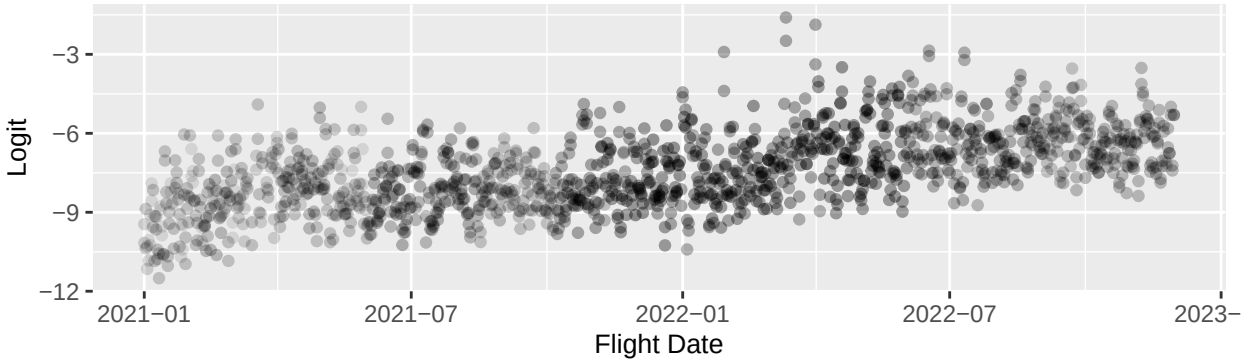

Figure 7: Linear Assumption of Logit vs. Predictor

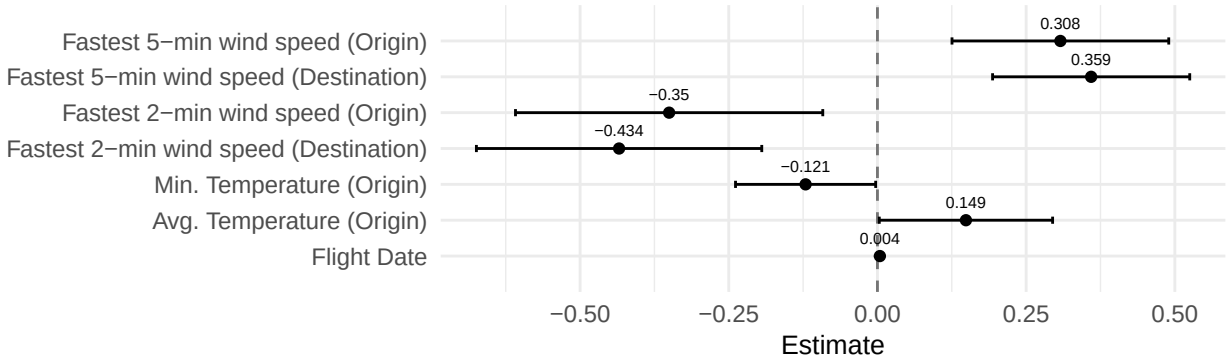

Figure 8: Logistic Regression Final Model Coefficients

Figure 8 describes the result of our logistic regression model, which includes the variables detailed in the stepwise backwards selection process. Our final model included the fastest 2-minute and 5-minute wind speeds for both the origin and destination airports, the minimum and average temperature at the origin airport, and the flight date. The final model had an AIC of 211.1 and a residual deviance of 195.1 on 7770 degrees of freedom. The Hosmer and Lemeshow goodness of fit (GOF) test yielded a p-value of 0.1259, which is acceptable considering our data, and the AUC was calculated to be 0.8318, erring on the side of over fitting perhaps.

The forest plot shows the statistically significant variables at the 0.05 level. The positive coefficients indicate that as the fastest 5-minute wind speed at both airports increase, or the temperature at the origin airport increases, the odds of the flight being canceled also increase. Something that is expected. The small coefficient associated with the date of the flight indicates that seasonality is perhaps not a major predictor in flight cancellation. On the other hand, the negative coefficients suggest that as the minimum temperature at the origin airport increases, or the fastest 2-minute wind speed at both airports increase, the odds of the flight being canceled decrease. This is somewhat counter intuitive, as one might expect that colder temperatures would be associated with more flight cancellations. It's possible that this effect is due to the fact that extremely low temperatures are relatively rare (especially in Raleigh and Miami) and therefore may be associated with other factors that increase the likelihood of flight cancellations. Also, with respect to the fastest 2-minute winds contradicting the effect of the fastest 2-minute winds, this could be because higher wind speeds of this nature can help to clear fog or other weather conditions that might otherwise cause flight cancellations.

In general, these results suggest that weather-related factors play an important role in determining whether flights are canceled. Wind speed and temperature at both the origin and destination airports appear to be particularly important, with different wind speed variables showing different effects. However, the results also suggest that the date of the flight is not as important a factor to consider, negating an assumption that flights tend to be canceled on certain days or during certain periods of the year.

Discussion

The aim of this project was to estimate the delays of flights between RDU and MIA and identify the factors that contribute to flight cancellations. Our primary hypothesis was that harsh weather conditions typically result in longer delay times and cancellations for flights between RDU and MIA. The secondary hypotheses were that the relationship between flight delay time / cancellation and weather condition may vary depending on departure time of day, flight date, and airline-specific factors.

To achieve the first aim, we developed a robust linear model that predicted the delay of flights with less weight on outlying values based on historical carrier performance, scheduled departure time, and weather conditions. The results showed that weather conditions were indeed a significant factor in predicting flight delays, confirming our primary hypothesis, with a few caveats. Additionally, the results showed that departure time of day, flight date, and airline-specific factors were also significant predictors of flight delay, supporting our secondary hypotheses. For the second aim, our logistic regression model was able to discern key predictors that play a role in flight cancellations, and explicitly described how different weather conditions have specific and sometimes counter intuitive effects.

However, the analysis had some limitations. The weather condition dataset used is reliable despite potential minor instrumental errors and human errors, but it is aggregated daily and is missing key data points that could be incredibly useful for this analysis. Moreover, the carrier dataset is limited to November 2022, which may not be representative of flight delays and cancellations in general, especially after the Covid-19 pandemic. Also, The generalizability of the findings for this aim could be hindered considering the potential impact of external factors (such as natural disasters) on flight cancellations. Overall, the number of cancelled flights in the current dataset is also extremely small for a powerful analysis, which calls for the inclusion of older data or perhaps the use of another model made for the most rare events. Despite these limitations, the strength of the robust linear model over traditional OLS method was discussed, which led to meaningful conclusions. Additionally, exploration of individual models for each airport also contributed to the findings.

Appendix

Appendix A - Student's T robust linear model

- Linear regression model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$
- Fitted model: $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip} + e_i = \mathbf{x}_i^\top \boldsymbol{b} + e_i$
- The general M-estimator minimizes the objective function

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^\top \boldsymbol{b})$$

where ρ is a function that gives the contribution of each residual.

Let $\psi = \rho'$ be the derivative of ρ. Differentiating the objective function with respect to the coefficients, $\boldsymbol{b}$, and setting the partial derivatives to 0, produces a system of $p+1$ estimating equations for the p coefficients:

$$\sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i^\top \boldsymbol{b}) \mathbf{x}_i^\top = \mathbf{0}$$

Define the weight function $w(e) = \frac{\psi(e)}{e}$, and let $w_i = w(e_i)$. Then the estimating equations may be written as

$$\sum_{i=1}^{n} w_i (y_i - \mathbf{x}_i^\top \boldsymbol{b}) \mathbf{x}_i^\top = \mathbf{0}$$

Solving the estimating equations is a weighted least-squares problem, minimizing $\sum_{i=1}^{n} w_i^2 e_i^2$. The weights, however, depend upon the residuals, the residuals depend upon the estimated coefficients, and the estimated coefficients depend upon the weights. An iterative solution (called iteratively reweighted least-squares, IRLS) is therefore required:

1. Select initial estimates $\boldsymbol{b}^{(0)}$, such as the least-squares estimates.
2. At each iteration t, calculate residuals $e_i^{(t-1)}$ and associated weights $w_i^{(t-1)} = w[e_i^{(t-1)}]$ from the previous iteration.
3. Solve for new weighted-least-squares estimates

   $$\boldsymbol{b}^{(t)} = (\boldsymbol{X}^\top \boldsymbol{W}^{(t-1)} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(t-1)} \boldsymbol{y}$$

   where $\boldsymbol{X}$ is the model matrix, with $\mathbf{x}_i$ as its ith row, and $\boldsymbol{W}^{(t-1)} = \text{diag}\left\{w_1^{(t-1)}\right\}$ is the current weight matrix.

Steps 2 and 3 are repeated until the estimated coefficients converge.

The asymptotic covariance matrix of $\boldsymbol{b}$ is

$$\mathcal{V}(\boldsymbol{b}) = \frac{E(\psi^2)}{[E(\psi')]^2} (X^\top X)^{-1}$$

Using $\sum[\psi(e_i)]^2$ to estimate $E(\psi^2)$, and $\left[\frac{\sum \psi'(e_i)}{n}\right]^2$ to estimate $[E(\psi')]^2$ produces the estimated asymptotic covariance matrix, $\widehat{\mathcal{V}}(\boldsymbol{b})$ (which is not reliable in small samples).

Appendix B

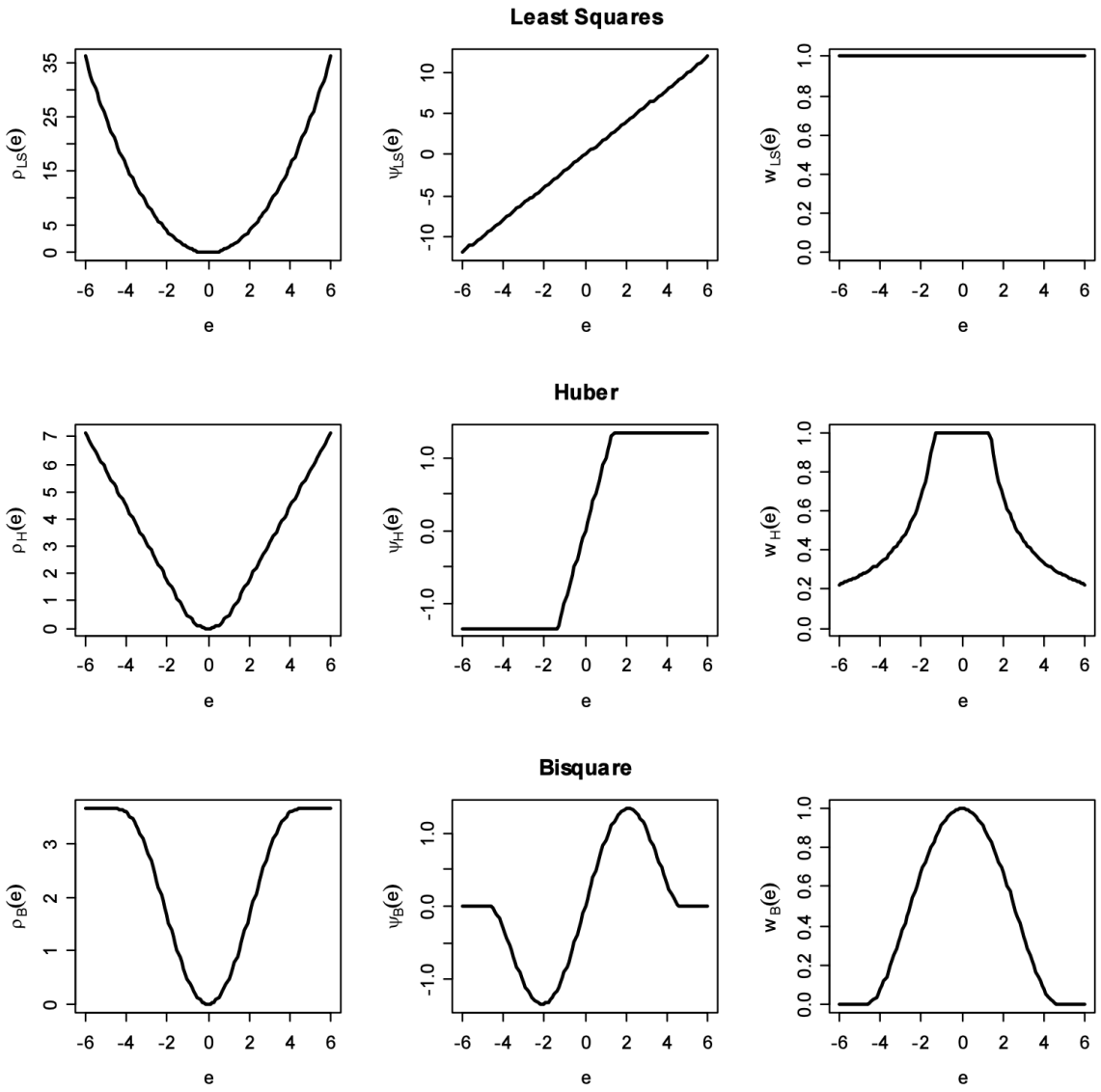

Figure 9: Objective, ψ, and weight functions for the least-squares (top), Huber (middle), and bisquare (bottom) estimators. The tuning constants for these graphs are $k = 1.345$ for the Huber estimator and $k = 4.685$ for the bisquare.

Fox (2002)

Appendix C

- The Hosmer-Lemeshow test can be expressed as:
  - $HL = \sum_{i=1}^{G} \frac{(O_i - E_i)^2}{E_i(E_i/n_i - 1)}$, where O_i and E_i are the observed and expected frequencies in each group, and n_i is the total number of observations in each group.

- To calculate the area under the ROC curve (AUC), we can use the trapezoidal rule:
  - $AUC = \sum_{i=1}^{n-1} \frac{(TPR_{i+1} - TPR_i)(FPR_{i+1} + FPR_i)}{2}$, where TPR is the true positive rate (sensitivity) and FPR is the false positive rate (1 - specificity).

Appendix D

Table 7: Carrier Code to Airline Translation

Unique Carrier Code	Airline
9E	Endeavor Air Inc.
AA	American Airlines Inc.
DL	Delta Air Lines Inc.
F9	Frontier Airlines Inc.
MQ	Envoy Air
NK	Spirit Air Lines
YX	Republic Airways

Appendix E

Data Acquisition

- Carrier data was downloaded from the BTS website directly on February 10, 2023.
- The raw carrier data is stored in `RawData` as monthly csv files and the combined and clean carrier data is available in `Data\carrier.csv`.
- Weather data was ordered from the NOAA website on February 9, 2023 and downloaded after the order was processed.
- The raw weather data is stored in `RawData\weather.csv` and the clean weather data is available in `Data\weather.csv`.
- Contact information for data collection/acquisition: Martin Olarte (mo144@duke.edu)

References

Aihounton, Ghislain B D, and Arne Henningsen. 2020. "Units of Measurement and the Inverse Hyperbolic Sine Transformation." *The Econometrics Journal* 24 (2): 334–51. https://doi.org/10.1093/ectj/utaa032.

Fox, John. 2002. "Robust Regression." Appendix to An R and S-PLUS Companion to Applied Regression. https://web.ipac.caltech.edu/staff/fmasci/home/astro_refs/RobustReg.pdf.

Fox, John, and Georges Monette. 1992. "Generalized Collinearity Diagnostics." *Journal of the American Statistical Association* 87 (417): 178–83. https://doi.org/10.1080/01621459.1992.10475190.

Susanti, Y., H. Pratiwi, S. Sulistijowati H., and T. Liana. 2014. "M ESTIMATION, S ESTIMATION, AND MM ESTIMATION IN ROBUST REGRESSION." *International Journal of Pure and Apllied Mathematics* 91 (3). https://doi.org/10.12732/ijpam.v91i3.7.

Venables, W. N., and B. D. Ripley. 2002. "Modern Applied Statistics with s." https://www.stats.ox.ac.uk/pub/MASS4/.