Chapter 1

INTRODUCTION

1.1 Background

Bicycling has becoming an increasingly important transportation mode because of its: 1) non-motorized transportation nature which reduces greenhouse gas emissions, 2) ability to relieve traffic congestion and parking difficulties, 3) improved safety of all roadway users, 4) benefits to personal physical wellness, and 5) affordability and accessibility to a broader group of people. As a result, policy makers and urban planners have been consistently promoting the share of bicycling as a major substitute of automobiles in urban transportation. The city of Seattle has set forth goals to increase amount and mode share of bicycle riding in its Bicycle Master Plan (BMP). The BMP calls for a citywide network of trails, protected bicycle lanes, and neighborhood greenways, and proposes an addition of more than 35 miles of protected bicycle lanes and more than 50 miles of neighborhood greenways throughout the city in the next five years [**?**].

Increasing levels of bicycle traffic create the need to quantify the use of these travel modes and their interaction with vehicles in transportation corridors [**?**]. Reliable bike count estimations/predictions are essential to determine/justify whether current corridor designs are working well and for future design of appropriate infrastructure that can accommodate bicycles, pedestrians and vehicles safely. In order to better develop policy and improve relevant infrastructure planning to induce more bicycling, a robust understanding of the factors that impact bicycle volume is necessary. With the advent of automated bike counters that are now installed in multiple sites across the City of Seattle, we are now in a good position to adopt a quantitative approach to investigate the relationship of bike volume and various factors such as weather, seasons, holidays, etc. One of the objectives of this thesis is to provide an extensive study on the impact of different factors on daily bike volume.

A robust understanding of bike volumes can be useful for policy makers, urban planners, and researchers to understand safety, travel behavior, and development needs. More specifically, it has the following impacts [**?**] :

- Determine existing travel patterns and demand;
- Identify corridors where current use and potential for increased use is high;
- Track trends over time;
- Evaluate the effectiveness of programs and/or facilities to promote biking;
- Identify locations for bicycle facility improvements and design appropriate treatments;
- Measure demographic changes as facilities that increase user comfort and attract a wider range of pedestrians and bicyclists are developed;
- Assess future bicycle travel demand;
- Make informed transportation decisions to prioritize bicycle improvement projects;
- Appropriately allocate resources for active transportation.

1.2 Objectives and contribution of this thesis

There are three major objectives that are addressed in this thesis:

1. **Provide an extensive study on the relationship between various factors (such as weather, season, etc) and daily bike ridership counts at Fremont Bridge in the CIty of Seattle.** A systematic approach is adopted to select the factors that have most signficant impact on daily bike counts. Also, the nonlinear relationship between weather and bike counts that has been observed in the literature [] is explicitly examined and accounted for. The resulted model is interpreted both qualitively and quantitatively.

2. **Develop a predictive model to estimate daily bike volume based on weather forecast and other temporal factors.** One benefit of this predicative model is to help transportation administrator to make better-informed decisions/preparations in case of inclement weather or special holidays that would result in siginificant change in bike counts. It is also believed to be able to help reveal multiple aspects of planning implications, such as ability to measure return on investment in bicycle facilities and help identify locations for new facilities.

3. **Understand rider behavior at the aggregated level and determine the 'real' bicycle trip trend after excluding the effect of weather and other temporal/seasonal variations.** This study is useful to help determine the current travel patterns/demand, and justify the effectiveness of current and future investments in facilities to promote biking.

The main contribution of this thesis are as follows:

1. **A square root transformation is applied to address the heteroscedascity in the bike counts data, and counterfactual simulations are conducted to interprate model results**. Even though there is obvious heteroscedasticity (i.e., varying variance) in the bike count data, the log transformed linear regression model is predominatly used in the literature [], possibly due to its convenience in model interpretation (one percent change in the independent variable results in certain fixed amount of percent change in the dependent variable). The square root transformation, well known for its ability to stablize variance [], however does not offer straightforward interpretation as with the log-linear model. To address this, counterfactual simulations [] are conducted to help interprate the model and visualize the effect of changing one varialbe while controling others.

2. **Explicitly examine and quantify the nonlinear relationship between weather factors between daily bike volume**. It has been noticed that there are nonlinear relationships between weather and bike volume []. However, such nonlinearity has not been explicitly modeled and their effects are mostly analyzed using exploratory data analysis. In this thesis, we adopt an

Gaussian Additive Mixture approach to model the nonlinear relationship and investigate its impact on the variable selection and model specification.

3. **Interaction terms are considered as prediction variables in the regression model**. There are limited number of papers that have included interaction terms in their modeling. In those do consider, attentions are only devoted to certain interactions between humidity and temperature [], or temperature and wind speed []. In this thesis, we extend interaction terms to include combination of weather and temporal variables (e.g., precipitation probability and weekend) and compare their goodness-of-fit metrics.

4. **To better capture the autocorrelation in the bike count time series, an Autoregressive Integrated Moving Average (ARIMA) model is fitted and used to predict bike volumes in the future**. With the exception of very few papers [], the majority in the literature assume the daily bike ridership count is independent. In order to develop a good predictive bike volume model, we fit an ARIMA model to account for possible autocorrelation in the daily bike counts. The accuracy of the resulting model is validated using historical data.

1.3 Organization of the Thesis

The remainder of this thesis is organized as follows: Chapter 2 surveys existing literatures on bicycle volume modeling. Methodology of the current study appears in Chapter 3. Detailed analysis and results are presented in Chapter 4, including exploratory data analysis, variable selection, model fitting and interpretation, prediction and trend analysis. Conclusions and future directions are provided in Chapter 5.

Chapter 2

LITERATURE REVIEW

Bicycle ridership in cities is useful for practitioners and researchers to understand safety, travel behavior, and development impacts. Therefore the relationship between bicycle volume and various factors, with the goal to build a predictive model based on this relationship, has been of great interest to researchers over the last decade Among others, weather variables such as temperature and precipitation have long been known to have significant impacts on bike travel demand and travel experience [**?**], since cyclists are fully exposed to outdoor weather conditions. Following the pioneering work by Hanson et al in 1977 [**?**], there have been many studies that attempt to explore the relationship between various weather factors and bicycle volume counts (e.g. [**?**, **?**, **?**, **?**, **?**]). With the advent of automatic bike counter that can continuously record the passing bike counts at specific locations, more statistically sophisticated models can be developed to account for more complicated scenarios and offer more explanatory/predicative power. In the following, we first summarize existing literature from three perspectives: data source of bike count, explanatory variable selection, and modeling methodology. Then we review the results of six papers that are most relevant to this thesis. Limitations of existing literature are discussed at the end of this Chapter.

2.1 Data source

There are two types of data sources than are commonly used to explore the relationship between weather factors and bike travel behavior: 1) Travel survey/census data have been either specifically designed to suit the purpose of the study [**?**] or were broad based travel surveys where data on all travel activity was collected using a successive sample approach over an extended period of time [**?**]. Bike count data obtained through this type of source is typically used to explain influencing factors such as physical, demographical and socio-economic factors on mode choice [**?**, **?**];

2) Observational travel data that is collected either manually or automatically. The manually collected data is usually collected for a specific purpose of the study [**?**] and over a short period of time. On the other hand, the automatically recorded data is continuously collected by automatic data collection equipment over a long period of time [**?**, **?**, **?**, **?**]. One obvious advantage of the automatic data collection systems is that they usually provide a longer time series of data, which will allow modeling of greater variation in weather/temporal parameters [**?**], whereas special purpose surveys are likely to either be of short duration or rely on respondent's recall of their behavior in the past, which is likely more prone to errors [**?**]. Supplemented with weather, temporal and other continuous factors, the automatically collected count data is suitable to develop predicative statistical models to forecast bike volume.

2.2 Variable Selection

A literature review accompanying a recent report by [**?**] identified eleven primary indicators. These included time of day [**?**], season [**?**], population and employment densities [**?**, **?**], land-use mix [**?**], bicycle facility type [**?**], traffic volume [**?**], rain and temperature [**?**, **?**], income [**?**], and age [**?**]. This section outlines the key points made in the literature that are relevant to some of most important variables.

Because cyclists are fully exposed in outdoor condition, weather variables play an crucial role in cyclists' decisions to ride. Research has found the variability for counts has a positive association with high temperature and low precipitation [**?**, **?**]. Meanwhile, as suggested by [**?**] and [**?**], the effects of precipitation and temperature on bicycle volumes are nonlinear. For example, bicycle traffic can decrease in both very cold and very hot weather as noted by [**?**]. In [**?**], the author also concluded that the optimal condition for bicycle usage occurs at approximately 25°C with no rainfall. Apart from the usual temperature and rain variables, [**?**] finds humidity and additional precipitation variables including the presence of rain in the morning and/or during the previous three hours to be significant too. According to [**?**], a 43% to 50% increase in ridership could be expected when the temperature doubled; however, the temperature had a negative effect when it was higher than 28°C and humidity was greater than 60%. Other comparative studies are also available where

bicycle counts are conducted in different cities [], and different sensitives to weather are examined [**?**]. As for longitudinal studies, [**?**] finds increased variability for counts conducted in the later months of the year. [**?**] conclude that morning peak hours from 6 AM to 9 PM accounts for a consistent 95% of the total bicycle volumes by hourly count data.

2.3 Existing literature on weather factor selection

In [**?**], a model is developed to use deviations in daily weather conditions from average conditions to predict deviations in daily cyclist totals from the average daily total. **Add a couple of sentences to evaluate its strength and weakness**

2.4 Modeling Methods

The simple linear regression model has been used in several applications [**?**, **?**]. Other modeling approaches include [**?**] which develops a count model and [**?**] which develops a time-series model. [**?**] also uses a Poisson model to statistically confirm many of the factors thought to influence cyclists. The work by [**?**] adopts a similar time-series approach while incorporating an autoregressive integrated moving average (ARIMA) analysis.

A summary of findings from the literature are presented in Table 2.4. These literature together suggest an opportunity for further model development base around long-term automated counts utilizing appropriate statistical methods. How seasonal factors influence bicycle flow needs to be examined in data that last more than a year. One limitation present in much of the past literature is that few discuss goodness of fit of their modeling. Further a model that can better describe and forecast the bicycle count in longitudinal form is necessary to be developed. Models for count data with better estimation methods offer some promise.

Source	Variable(s) identified	Methods
Fields-2012	Average daily temperature; Total weekly precipitation	Identify patterns through scatter plots; No explicit model is established.
Gallop-2012	Temperature, Relative humidity, wind speed, visibility, fog, precipitation	Use ARIMA to account for serial correlation patterns
Griswold-2011	Nearby population and employment density, proximity to downtown/freeway, age, education level, income, etc.	Log linear ordinary least squares regression is used to estimate a bicycle count model
Helbich-2014	Daily maiximum air temperature, daily average wind speed, daily precipitation	Place-specific associations of weather conditions are explored through geographically weighted logit models
Hunt-2007	Descriptive variables indicating lane use, secured parking, level of experience, etc.	Logistic model of cycling-related choices
Jones-2010	Length of bicycle network, employment density, population density	Standard ordinary least squares regression
Lewin-2011	Max temp, rain flag, snow flag, weekend flag, over 90 flag	Standard linear regression model
McCahil-2008	logarithmic choice measure, population density, worker density	A new space syntax theory is used to evaluate and predict the bicycle volume throughout a network
Miranda-Moreno-2011	Temperature, percent humidity, rain presence, rain presence in prev. 3hrs, warm & humid, morning rain	Both log-linear model and negative binomial model are tested
Niemeier-1996	Morning flag, rain flag, high temp flag, location variable, season variable	A Poisson count model is assumed and fitted
Nosal-2014	Temperature, percent humidity, rain flag, rain prev. 3hrs, am rain, pm rain	The relationship is analyzed using a log-linear regression model
Parkin-2008	Gender, car ownership, hilliness, off-road routes proportion	A logistic regression model of relevant socio-economic and physical variables is estimated.
Pinjari-2009	Household density, employment density, fraction of commercial land area, demographic factors including proportion of population that are seniors and proportion of population by race	The model system takes the form of a joint mixed Multinomial LogitMultiple Discrete-Continuous Extreme Value (MNL-MDCEV) structure
Rose-2011	Temperature, rainfall, holiday flag, school season flag, day of the week	Weather and other effects examined using an aggregate model of daily ridership
Thomas-2009	Temperature, sunshine, precipitation, wind force, cycle path use	A bi-level structure is developed with the upper level being a log-linear model and the lower level being a linear model

Chapter 3

METHODOLOGY

In order to discern the relationship between bicycle counts and several identified weather, seasonal, and temporal factors, we developed a statistical model that attempts to predict daily bicycle counts from these other factors. The major contributions of this thesis are: 1) to provide a detailed analysis on the impact of various weather factors on bike counts, and 2) to quantitatively investigate the nonlinear relationship between weather factors and bike counts. This section describes the methods and procedures we used to collect and process the raw datasets (including both bike counts and weather variables), as well as the statistical tools used for bike count modeling (including linear regression, general linearized model, Gaussian Additive Mixture model, and ARIMA time series analysis).

3.1 Study Location

The bike facility of interest is the Fremont Bridge in the City of Seattle. The Fremont bridge crosses the Lake Washington Ship Canal and links the Fremont with the Queen Ann neighbourhood. The reason for picking Fremont bridge as our study location is: 1) A permanent, automatic bike counter is installed at the Fremont bridge, which provides continuous bike counts; 2) As opposed to other recreational facilities, the Fremont Bridge represents one of the busiest utilitarian facility in the City of Seattle, which is the subject of this thesis; 3) The bike counter on Fremont bridge was first installed in October 2012, the earliest site among the nine bike counters in Seattle, and therefore provides a rich dataset of a little more than 3 years; 4) It's relatively close to the University of Washington, and connects the northen part of Seattle to its downtown area. 5) It captures a substantial amount of bicycle traffic due to its status as one of only five facilities that carry bicyclists across the canal separating the northern and southern halves of Seattle.

SDOT has nine bike counters (four of which also count pedestrians) located on neighborhood greenways, multi-use trails, at the Fremont Bridge and on SW Spokane Street. The counters help create a ridership baseline that can be used to assess future years and make sure the right amount of resources are invested so that the goal of quadrupling ridership by 2030 could be achieved [**?**]

While only 3 percent of downtown Seattles 200,000 daily commuters now bicycle, the number of bike commuters has increased 18 percent since 2010, according to a survey done for the Downtown Seattle Association [], the citys Department of Transportation and King County Metro.

3.2 Data collection, processing and description

3.2.1 Bike Count Data

Bicycle counts were collected at Fremont bridge continuously by the City of Seattle using an in-sidewalk counter manufactured by EcoCounter (see Figure 3.1). When a bicycle passes over an induction loop embedded in the sidewalk on either side of the Fremont Bridge, the counter registers the bicycle. Data from this equipment has been used in a wide range of studies, and when operating properly, the absolute error of these counters has been shown to be below 4% [**?**]. Bicyclists may legally choose to ride in the roadway instead of the sidewalk, and would thus not be detected by the counter. However, we believe these crossings are rare at this location due to the design of the facility, which directs bicyclists to enter the sidewalk, and from our own experience riding and observing other riders. The counters upload data once a day at 5 am, which is then aggregated into 15 minute intervals by the City of Seattle, and are made available to the public via the City of Seattle's data portal [**?**, **?**].

The bike count data used in this study cover a period of three years spanning from October 31, 2012 to October 30, 2015. The continuous bike count data is aggregated into daily counts. Note that in the literature, there is also studies on bike count modeling using hourly bike count. However, the daily bike counts are favored in this study because: 1) it carries less autocorrelation than the hourly data; 2) it's intuitive and simple to interprate, 3) we believe for bicyclists commuting to work, they make the decision of riding based on the daily weather, as opposed to the recreational

Figure 3.1: EcoCounter on the Fremont Bridge

riders is more likely to make deicions on a hourly basis.

3.2.2 Weather Data

Weather data are collected by a variety of sources and are aggregated by Forecast.io. These data are available through the company's web services API [**?**]. Historical daily summaries are available for a range of weather variables including several specifically important to our model such as precipitation, daily minimum and maximum temperatures, sunrise, and sunset, etc.

We downloaded and processed these data programmatically using the R programming language along with several add-on packages [**?**, **?**, **?**, **?**, **?**]. Bicycle counts were aggregated by day, and then joined to weather data by date.

3.2.3 Temporal Data

In addition to the variables collected from the above-mentioned two sources, we were also interested in controlling for holidays and whether or not the nearby University of Washington was in session. These data were collected and coded manually from the National Holiday calendar as well as the University of Washington's historic academic calendar.

3.3 Exploratory Data Analysis

Figure 3.2 provides a visual summary of the processed counts data. Some apparent outliers are visible at the rightmost portion of the histogram. The two highest counts occurred the Monday and Tuesday preceding the beginning of National Bike to Work Month. And the third highest count occurred on National Bike to Work Day.

3.4 Ordinary Least Squares Regression

The ordinary least squares (OLS) regression is one of the most widely used techniques in Statistics to estimate the unknown parameters. The goal is to minimize the differences between the observed responses and the predicted response given by the linear approximation of the data. The basic

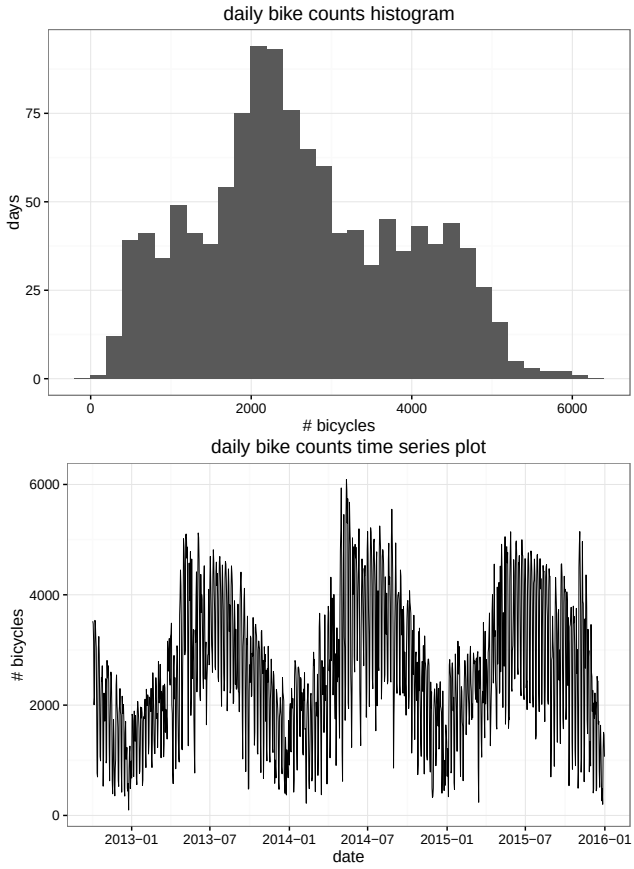

Figure 3.2: Descriptive visualization of bicycle counts dataset.

linear regression model assumes the following structure:

$$Y = \alpha + \beta X + \varepsilon, \tag{3.1}$$

where Y is the response variable (or observations, dependent variable), and X is the predictors (or regressors, independent variables), α and β are the unknown parameters to estimated, and ε is the unobserved scalar random variables (errors) which accounts for the discrepancy between the actual observed responses Y and the predicted responses $\alpha + \beta X$.

The OLS techinque offers a mathematically convinient tool to estimate the linear regression model parameters α and β. There are many available software packages to provide solutions to the linear problem (3.1), i.e., the `lm` function in `R`. However, to properly apply the OLS estimators, certain assumptions need to be checked beforehand, such as: 1) No linear dependence (no multicolinear), 2) Strict Exogeneity ($E[\varepsilon|X] = 0$), 3) Homoscedasticity ($E[\varepsilon^2|X] = \sigma^2$), and 4) Normality (the error term has a normal distribution). For more details on OLS regression models, readers are referred to [].

The goodness-of-fit of the considered model is often evaluated with the R^2 (R squared, or the coefficient of determination) and adjusted R^2. The R^2 measures the percentage of the response variable variation that is explained by a linear model, or equivalently

$$R^2 = 1 - \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}}. \tag{3.2}$$

The adjusted R^2 adds a correction for the number of estimated parameters to guard against overfitting. Other important goodness-of-fit criterions that are used in this study are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Details on the latter two can be found in [].

3.5 Generalized Linear Model (GLM)

3.6 Generalized Additive Mixture (GAM) Model

The generalized additive model (GAM) is a generalized linear model in which the linear response variable depends on unknown smooth functions of the independent variables [**?**]. The goal is to

provide characterization about these smooth functions. GAM was first proposed in [**?**] to blend the properties of generalized linear models with additive models.

Following a similar approach with the GLMs, an exponential family distribution (could also be Poisson, normal, negative binomial, etc) is specified for response Y along with a link function g relating the expected value of Y to the predictors X_i such as

$$g(E[Y]) = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_m(X_m)$$

The function $f_i(X_i)$ may be specified parametric functions (e.g., polynomial) or may be specified non-parametrically, or semi-parametrically, simply as 'smooth functions', to be estimated by non-parametric means. The nonparametric GAM provides a very general modular estimation method capable of using a wide variety of smoothing methods to estimate the $f(X)$. The advantage of non-parametric models is that they are easy and efficient to fit, while the disadvantage is the inability to control the complexity of the model (degree of smoothness of $f(X)$), which often gives rise to problems with interpretation. Overall, a well calibrated GAM is likely to perform better than nearly any other model type, if the dataset is large enough and its behavior is complex enough. However, it could also have problems of overfitting as the number of smoothing parameters increases.

In this thesis, we use the GAM model to explore the potential relationship between the dependent variable bike counts and weather factors (i.e., temperature and precipitation). It is noticed in other studies [] that the temperature has a positive effect on ridership when its below, and a negative impact when higher than. Also, the temperature squared is often used in the literature [] to account for the nonlinearity without good explanation. The GAM provides a good starting point to investigate such nonlinear relationship since it explicitly models the dependent variable on a smooth function of the independent variable. Consequently, resulting non-parametric smooth function provides valuable insight on how to include nonlinear terms in the recommended model. In this thesis, the `gam` in `R` is used to fit generalized additive models, specified by giving a symbolic description of the additive predictor. `gam` uses the backfitting algorithm [] to combine different smoothing or fitting methods. The default built-in nonparametric smooth splines are used to fit the model.

3.7 Autoregressive Integrated Moving Average (ARIMA) Model

Chapter 4
ANALYSIS AND RESULTS

In this chapter, we first explore the variables that have the most significant impact on the bike counts. A thorough analysis is conducted on weather factors to extract a subset of variables from the original available datasets. Second, various Y transformations are investigated so that the residual of the resulting model is most similar to a standard Normal distribution. Third, different model specifications are considered and evaluated against common goodness-of-fit criterions. Model interpretations and cross validations are provided as well. Lastly, the recommended models are used for a few applications, such as bicycle counts predication, and trend analysis, etc.

4.1 Variable Selection

The original weather dataset retrieved from `Forecast.io` contains 36 variables including 35 potential predicators. In order to select an initial subset of variables for our analysis, we adopt a systematic approach which consists of the following perspectives.

4.1.1 Insight from the literature

University of Washington in-session status were selected to represent seasonality. We also deemed the University of Washington variable important in part because of the Fremont Bridge's proximity and connection via the Burke Gilman Trail to the University of Washington. We also felt that this variable was a suitable proxy for the "school season," which more broadly captures whether or not other local schools are in session. The academic calendars of the various local schools do not align perfectly, however they still overlap substantially with the University of Washington, which is itself the largest educational institution in the region.

Inclusion of the holiday variable was an attempt to account for some low outlier counts. Upon

inspection of the dataset, Christmas and Thanksgiving in particular had very low counts of bicycles relative to the days preceding and following. Relatedly, but not accounted for by any variable in our model, are some of the high outlier counts. Upon inspection, some of the highest counts were observed on National Bike to Work Day and on the day of the Fremont Solstice Parade, which typically draws large numbers of bicyclists as participants and spectators. The omission of such a variable is justified based on the few occurrences of high outlier counts, and our desire for this model to only include variables that could be collected or straightforwardly adapted to other locations.

Daily maximum temperature, measured in Fahrenheit, was chosen to represent temperature (rather than, for example, substituting or adding daily minimum temperature) in part to retain simplicity in the model, in part because there is relatively little daily temperature variation in Seattle due to the moderating effect of large water bodies, and in part because maximum temperature better reflects the conditions during daylight hours when most bicycle trips would occur. This simplification may not be warranted for other locations that experience greater temperature variation than Seattle. The squared daily maximum temperature is also included to account for potential nonlinear relationship.

Maximum precipitation, which measures the maximum inches of precipitation that occurred in any hour throughout the day, was chosen rather than average precipitation based on the notion that bicyclists might make travel decisions based on a likely worst case scenario. This assumption is slightly more problematic than our assumptions about temperature, in that we do expect bicyclists to be at least somewhat sensitive to average conditions or conditions observed at their time of departure. As in the case of temperature, this simplification would be less justifiable in locations that experience greater daily variation in precipitation or in locations that have a predictable pattern of precipitation during certain hours.

Precipitation probability

Day of the week was added due to its presence in the literature, as well as an apparent weekly pattern revealed visually by zooming into the timeseries plot. These data were coded as a set of Boolean dummy variables, excluding Sunday as the reference category.

The final variable, the day number, was included so that we could test for a linear trend in bicycling volumes. We created this variable by sequentially numbering (1–720) the observed counts by day during the study period.

4.1.2 Correlation Scatterplot

We then selected the predictors with high absolute correlation and draw several scatterplot matrix to explore their pair-wise relationship. Because of the large number of highly correlated variables, we split them into three groups: one contains all the variables that relate to precipitation; one with all the variables that relate to temperature; and the last group contains the rest of the variables that's highly correlated with bike count.

In the scatter matrix Figure 4.1, one can see the correlation between the probability of precipitation and the response count is -0.452, which is relatively stronger comparing to the other predictors. Furthermore, there exist high correlations between "precipProbability" and other variables such as "precipIntensityMax", "precipIntensity", and "humidity". Therefore, to avoid collinearity we decided to keep "precipProbability" in our model and drop "precipIntensity", "precipIntensityMax" and "humidity". In addition, the correla- tion between "dewpoint? and our response is also relatively high, and the correlation between "dewpoint" and "precipProbability" is relatively small. Thus, we decided to keep "dewpoint" in the pool of covariates as well.

Notice that for "precipIntensity" and "precipIntensityMax", their distribution is highly right skewed. One could also explore the relationship of bike count verses the log transformed variables.

In the scatterplot matrix Figure 4.2, one can see the correlations between each covariate and the response are high, and the correlations between any two covariates are more than 0.9. Therefore, including just one of them should be adequate, and we decided to keep only "temperatureMax" because it has the highest correlation with bike counts.

We also draw box plots for each categorical variable against the response daily count of bicycle. From Figure 4.3, we see that across different levels, the means of bike count for all 6 categorical variables are not the same.

The initial variable set we used for model testing is summarized in Table 4.1. Note `MaxTempSq`

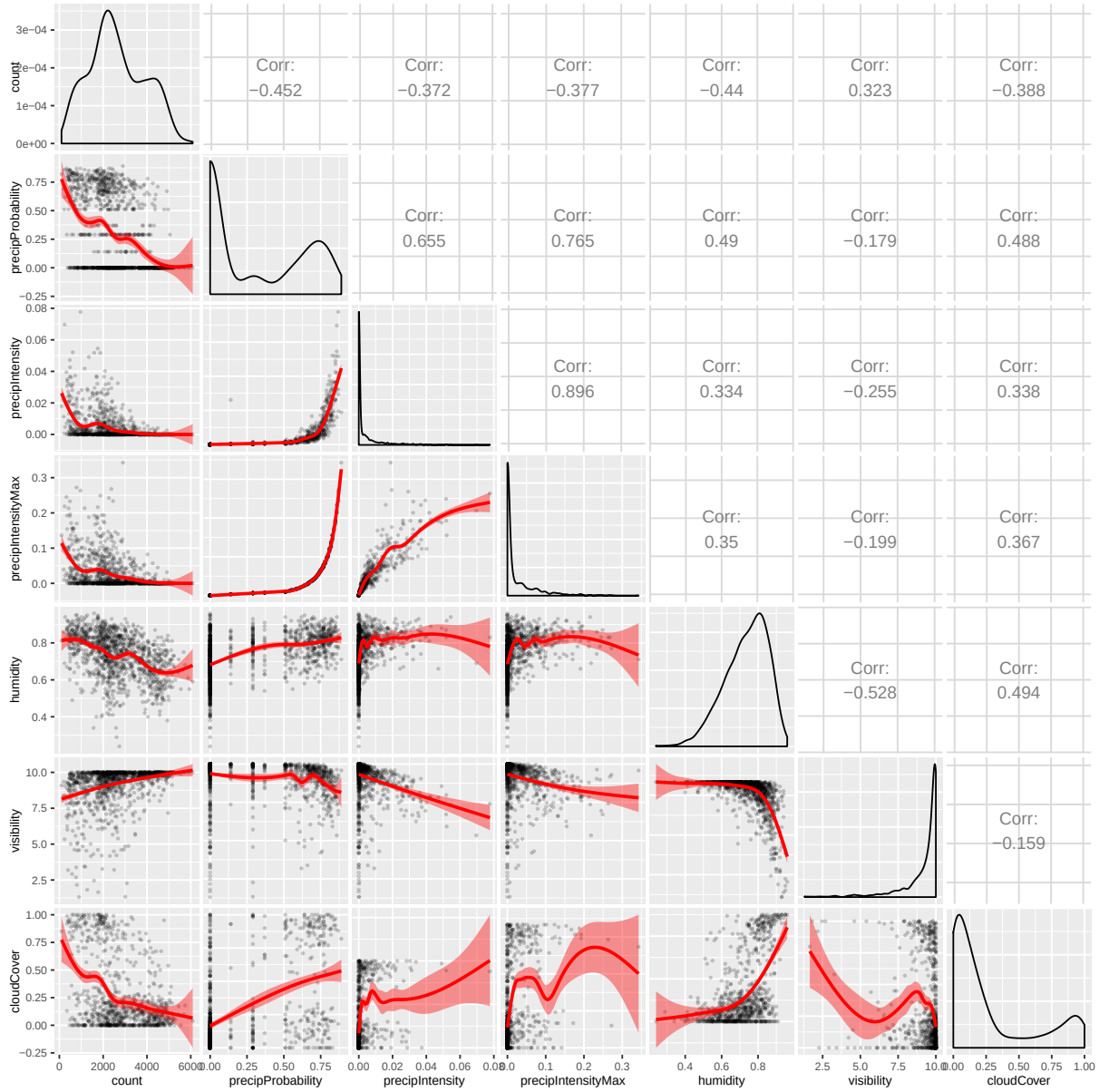

Figure 4.1: Scatterplot Matrix of covariates in the Group of Precipitation and the Response

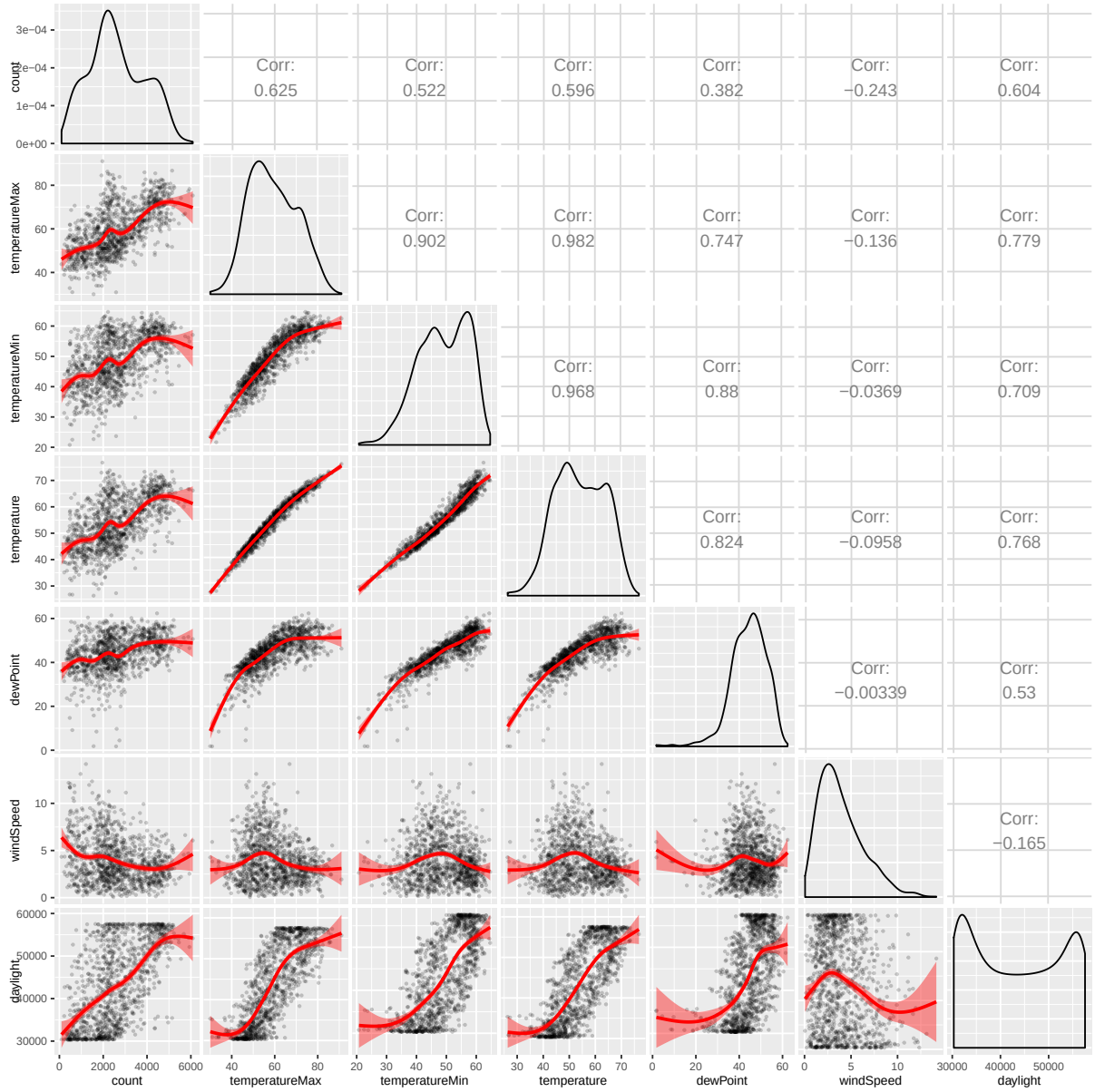

Figure 4.2: Scatterplot Matrix of covariates in the Group of Temperature and the Response

Table 4.1: Variables included in model specification

Variable	Description
Count*	Number of bicycles per day
MaxTempSq	The square of maximum temperature for the day
PrecipProb	Probability of precipitation for a given day
Daylight	Time from dawn to dusk in hour
dow	Day of the week dummy variable (from Mon. to Sun.)
Holiday	The day was a holiday as recognized by UW (TRUE/FALSE)
Weekend	The day is a weekend (TRUE/FALSE)
UW	The University of Washington was in session (TRUE/FALSE)
Season	Season indicator spring, summer, fall, and winter
icon	General weather classification, such as “clear-day”, “cloudy”, “foggy”, “rainy”, “windy”, “partly-cloudy-day”, “partly-cloudy-night” (7 levels)

is included to account for potential nonlinearity for temperature.

4.1.3 Model Fitting

In this section, several candidate model specifications are considered and fitted to evaluate the impact of different weather factors/transformations on bike counts. The base model is chosen to be as follows:

4.2 Y transformation

In this section, we investigated different forms of the transformation of dependent variable Y, so that the fitted model will have a similar distribution of Gaussian. To make fair comparison, we

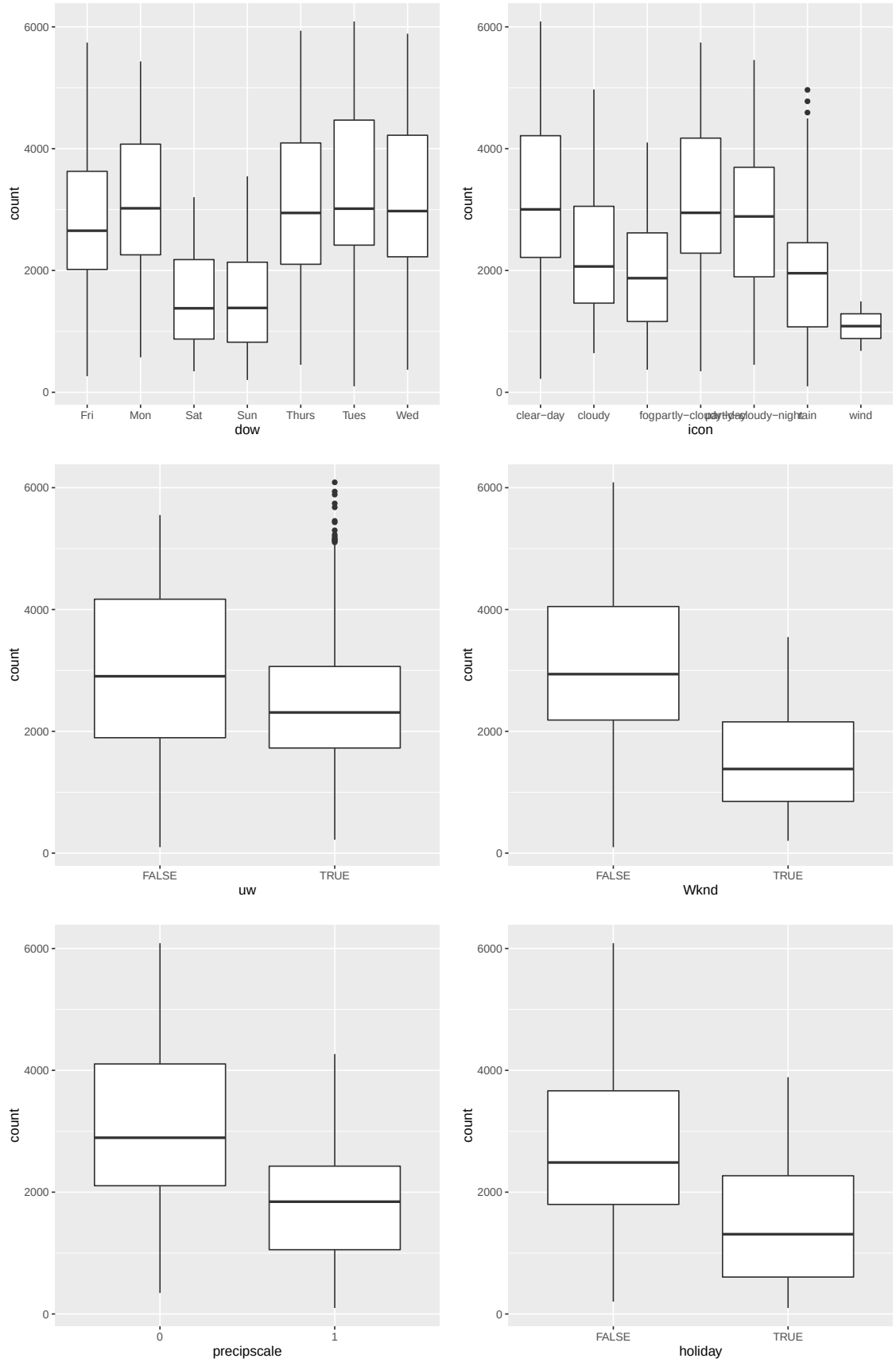

Figure 4.3: Boxplots of “dow”, “icon, “uw”, “wknd”, “prepscale”, and “holiday” versus “count”

fixed the independent variables to be `TempMaxSq`, `Holiday`, `PrecipProb`, `Weekend`, and `Daylight`. The following common transformations of Y are considered:

No Transformation on Y:

$$\text{Count}_t = \beta_0 + \beta_1 \text{TempMaxSq}_t + \beta_2 \text{Holiday}_t + \beta_3 \text{PrecpProb}_t + \beta_4 \text{Weekend}_t + \beta_5 \text{Daylight}_t + \varepsilon \tag{4.1}$$

Log Y:

$$\log(\text{Count}_t) = \beta_0 + \beta_1 \text{TempMaxSq}_t + \beta_2 \text{Holiday}_t + \beta_3 \text{PrecpProb}_t + \beta_4 \text{Weekend}_t + \beta_5 \text{Daylight}_t + \varepsilon. \tag{4.2}$$

Square Root Y:

$$E[\sqrt{\text{Count}_t}|X] = \beta_0 + \beta_1 \text{TempMaxSq}_t + \beta_2 \text{Holiday}_t + \beta_3 \text{PrecpProb}_t + \beta_4 \text{Weekend}_t + \beta_5 \text{Daylight}_t. \tag{4.3}$$

Poisson:

$$\text{Count}_t = \beta_0 + \beta_1 \text{TempMaxSq}_t + \beta_2 \text{Holiday}_t + \beta_3 \text{PrecpProb}_t + \beta_4 \text{Weekend}_t + \beta_5 \text{Daylight}_t + \varepsilon, \tag{4.4}$$

where Count_t is assumed to follow a Poisson distribution.

Log-linked Gassuian (glmlog):

$$\log(E[\text{Count}_t]) = \beta_0 + \beta_1 \text{TempMaxSq}_t + \beta_2 \text{Holiday}_t + \beta_3 \text{PrecpProb}_t + \beta_4 \text{Weekend}_t + \beta_5 \text{Daylight}_t, \tag{4.5}$$

where $\log(\text{Count}_t)$ is assumed to follow a standard Gaussian distribution.

4.2.1 Model residual analysis

To evaluate the impact of different transformations on Y, the following visualization plots are used for comparison among models (4.1)-(4.5).

1. **Residual vs. Fitted plot:** A plot of the residuals against the fitted values should show no pattern. If a pattern is observed, there may be 'heteroscedasticity' in the errors. That is,

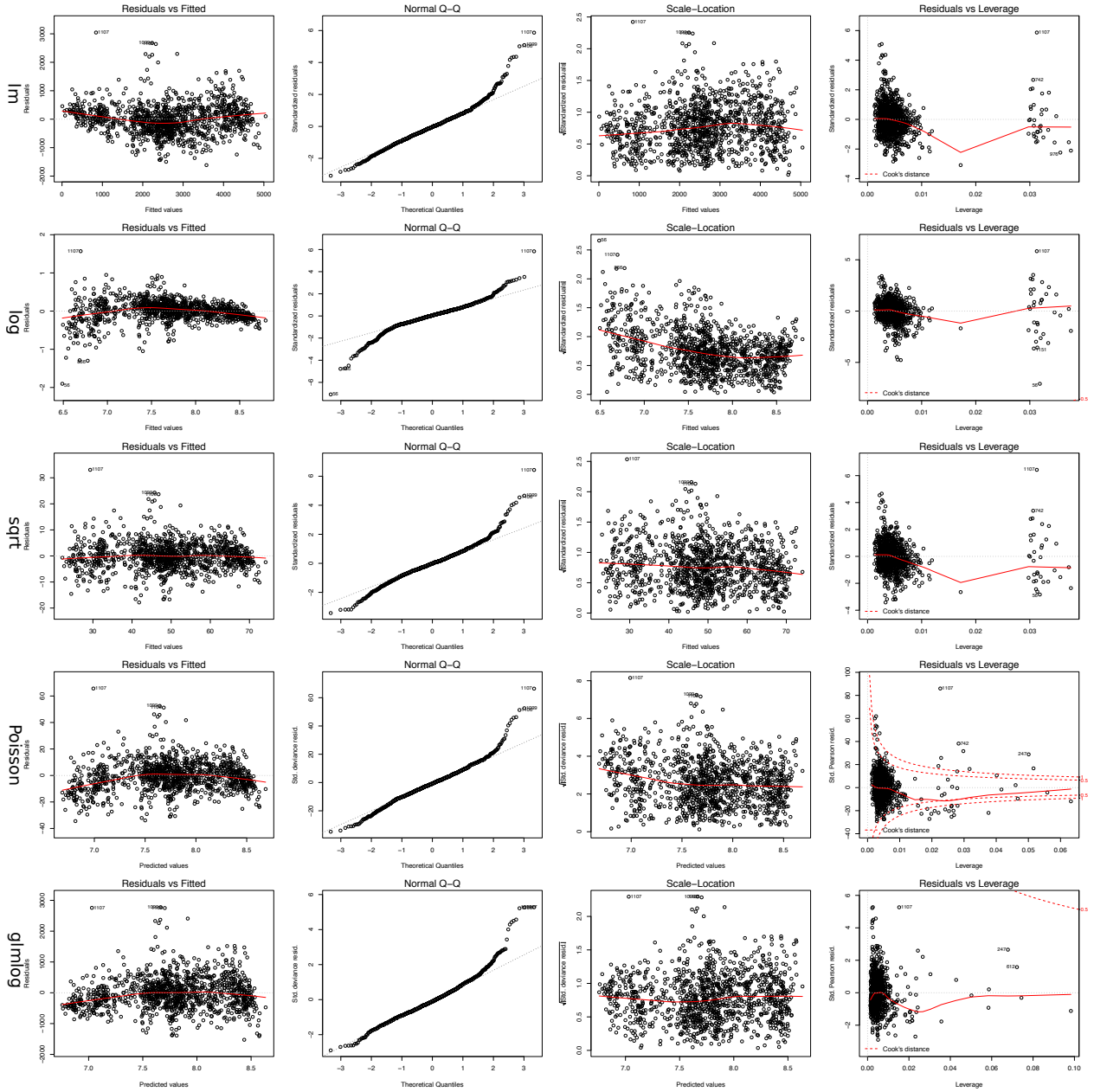

Figure 4.4: Scatterplot Matrix of covariates in the Group of Temperature and the Response

the variance of the residuals may not be constant. This indicates a transformation of the dependent variable may be required. It is shown in Figure 4.4 that the model with original bike count as dependent variable has a clear varying variance in its residual: the residual becomes larger when the bike counts to be predicted become larger. Among the four common transformations we tested (log, sqrt, Poisson and glmlog), the square root transformation results in the best residual plot, which remains constant for all fitted values.

2. **Normal Q-Q plot:** The Normal Quantile-Quantile (Q-Q) plot represents an informal graphical test of the hypothesis that a data sequence is normally distributed. That is, if the points on a normal Q-Q plot are reasonably well approximated by a straight line, the Gaussian data hypothesis is plausible, while marked deviations from linearity provide evidence against this hypothesis [**?**]. It is shown in Figure 4.4 that all models under consideration don't have a perfectly Normal distributed residuals, as is required by the linear regression models. For the square root transformation, the residual error has a slightly heavy tail on the right side.

3. **Scale-Location plot:** It depicts the square root of the absolute values of the residuals against the fitted values, with a lowess curve helpfully overlaid. The scale-location plot is often used to check if the data possesses homoscedasticity, which is required for the linear regression model. Ideally if the modeling data has homoscedasticity, the lowess curve is expected to be flat, not sloped, and the square root of residuals should be approximately evenly distributed along the lowess curve. Using this criterion, it is shown in Figure 4.4 that the square root transformed model has the least heteroscedasticity because its lowess curve is relatively flat and the residual didn't show a obvious pattern.

4. **Residual vs. Leverage plot:** This plot depicts the standardized residuals against the leverage for each point in the data series. The Cook's Distance is also shown in the plot. This plot is mainly used to identify extreme points and possible outliers in the data series that could shift the regression line significantly. The further out in the X or Y axis, the more leverage or standardized residual the point in dataset has. More details on the Residual vs. Leverage

plot could be found in [**?**]. From Figure 4.4 it can be seen that there are a few points in the dataset that has large leverage on the regression lines (indicated by points to the far right side of the plot). The Poisson model has relatively larger residuals with quite a few points with big Cook's distance. This indicate the Poisson model might not be a good fit. For each model, there is one point has big positive error (1107). This outlier point corresponds to the 'City Bike to Work Day' where there is more than six thousand bike passing the Fremont Bridge in one day. This suggests proper data cleaning is required (i.e., removing outliers) to best capture the relationship between utilitarian bike ridership and weather/temporal factors.

In summary, after exploration of different transformations of the dependent variable, the squared root of bike count appears to give the best properties of the resulting model: its residual has an approximate constant variance; it follows an approximate Gaussian distribution; it satisifies the homoscedasticity assumption of the error terms. Therefore, we use the square root transformed bike count in our following analysis.

4.3 Model Specification

In this section, we investigate five different model specifications and conduct a detailed performance comparison by means of goodness-of-fit, cross validation, and their predictive power.

The following five models are considered in this section:

- **Model 0**

$$\log(\text{Count}_t) = \beta_0 + \beta_1 \text{TempMaxSq}_t + \beta_2 \text{Holiday}_t + \beta_3 \text{PrecpProb}_t + \beta_4 \text{Weekend}_t + \beta_5 \text{Daylight}_t + \beta_6 \text{UW}_t + \varepsilon. \tag{4.6}$$

In this model, the log of bike count is used as Y, and the independent variables are: squared max temperature, holiday indicator, precipitation probability, weekend indicator, daylight hours, and "UW in session" indicator. Model 0 serves as the "base model". All the following models differ from the base model by changing or adding one independent variable.

- **Model 1**

$$\log(\text{Count}_t) = \beta_0 + \beta_1 \text{TempMaxSq}_t + \beta_2 \text{Holiday}_t + \beta_3 \text{PrecpProb}_t + \beta_4 \text{Weekend}_t + \beta_5 \text{Season}_t + \beta_6 \text{UW}_t + \varepsilon. \tag{4.7}$$

Model 1 replaces the daylight hours in Model 0 by the Season indicator.

- **Model 2**

$$\log(\text{Count}_t) = \beta_0 + \beta_1 \text{TempMaxSq}_t + \beta_2 \text{Holiday}_t + \beta_3 \text{PrecpProb}_t + \beta_4 \text{Weekend}_t + \beta_5 \text{Daylight}_t + \beta_6 \text{Icon}_t + \beta_7 \text{UW}_t + \varepsilon. \tag{4.8}$$

Model 2 add one more independent variable, the weather icon, to Model 0.

Table 4.2: Model Specification

Name	Model Formula
Model 0	log(count) $\sim$ temperatureMaxSq + holiday + precipProbability + Wknd+ daylight +uw
Model 1	log(count) $\sim$ temperatureMaxSq + holiday + precipProbability + Wknd+ season +uw
Model 2	log(count) $\sim$ temperatureMaxSq + holiday + precipProbability + Wknd+ daylight+ icon +uw
Model 3	log(count) $\sim$ temperatureMaxSq + holiday + precipProbability*Wknd + daylight +uw
Model 4	log(count) $\sim$ temperatureMaxSq + holiday + precipProbability + dow + daylight +uw

- **Model 3**

$$\log(\text{Count}_t) = \beta_0 + \beta_1 \text{TempMaxSq}_t + \beta_2 \text{Holiday}_t + \beta_3 \text{PrecpProb}_t + \beta_4 \text{Weekend}_t + \beta_5 \text{PrecpProb}_t \times \text{Weekend}_t + \beta_6 \text{Daylight}_t + \beta_7 \text{UW}_t + \varepsilon. \quad (4.9)$$

Model 3 differs from Model 0 by allowing interactions between Precipitation Probability and the Weekend indicator variable.

- **Model 4**

$$\log(\text{Count}_t) = \beta_0 + \beta_1 \text{TempMaxSq}_t + \beta_2 \text{Holiday}_t + \beta_3 \text{PrecpProb}_t + \beta_4 \text{dow}_t + \beta_5 \text{Daylight}_t + \beta_6 \text{UW}_t + \varepsilon. \quad (4.10)$$

Model 4 replaces the weekend indicator in Model 0 by the dow indicator.

A summary of the model specification used in this analysis is provided in Table 4.2. Note that Model 0 serves as the base model in this analysis.

4.3.1 Fitted model and goodness-of-fit

We summarize the model fitting results (including all the regression coefficients) as well as their goodness-of-fit criterion in Table 4.3 and Table 4.4.

Table 4.3: Model Comparison Results

	Dependent variable: log(BikeCounts)		
	Model 0	Model 1	Model 2
TempMaxSq	0.0002*** (0.00001)	0.0002*** (0.00001)	0.0002*** (0.00001)
Holiday	−0.798*** (0.046)	−0.822*** (0.046)	−0.804*** (0.045)
PrecipProb	−0.493*** (0.025)	−0.524*** (0.025)	−0.322*** (0.041)
Weekend	−0.811*** (0.017)	−0.815*** (0.017)	−0.811*** (0.017)
daylight	0.00002*** (0.00000)		0.00002*** (0.00000)
spring		0.061*** (0.022)	
summer		0.064** (0.029)	
winter		−0.199*** (0.025)	
cloudy			−0.115** (0.056)
fog			0.023 (0.038)
partly-cloudy-day			0.018 (0.023)
partly-cloudy-night			0.054* (0.032)
rain			−0.139*** (0.030)
wind			−0.206 (0.184)
UW	0.202*** (0.018)	0.178*** (0.020)	0.200*** (0.018)
const.	6.764*** (0.047)	7.381*** (0.047)	6.759*** (0.049)
Observations	1,157	1,157	1,157
R^2	0.819	0.815	0.825
Adjusted R^2	0.818	0.814	0.823
AIC	173.7066	203.1459	149.7224
BIC	214.1353	253.6818	220.4726
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4.4: Model Specification Results Ctnd

	Dependent variable: log(BikeCounts)		
	Model 0	Model 3	Model 4
TempMaxSq	0.0002^{***} (0.00001)	0.0002^{***} (0.00001)	0.0002^{***} (0.00001)
Holiday	-0.798^{***} (0.046)	-0.796^{***} (0.044)	-0.793^{***} (0.045)
PrecipProb	-0.493^{***} (0.025)	-0.359^{***} (0.027)	-0.491^{***} (0.024)
Weekend	-0.811^{***} (0.017)	-0.678^{***} (0.021)	
Mon			0.135^{***} (0.028)
Sat			-0.677^{***} (0.028)
Sun			-0.737^{***} (0.028)
Thu			0.089^{***} (0.028)
Tue			0.150^{***} (0.028)
Wed			0.146^{***} (0.028)
daylight	0.00002^{***} (0.000)	0.00002^{***} (0.000)	0.00002^{***} (0.000)
uw	0.202^{***} (0.018)	0.201^{***} (0.018)	0.205^{***} (0.018)
PrecipProb:Wknd		-0.464^{***} (0.048)	
const.	6.764^{***} (0.047)	6.727^{***} (0.045)	6.654^{***} (0.049)
Observations	1,157	1,157	1,157
R^2	0.819	0.832	0.826
Adjusted R^2	0.818	0.831	0.824
AIC	173.7066	86.6983	139.2741
BIC	214.1353	132.1806	204.9707
Note:			*p<0.1; **p<0.05; ***p<0.01

All fitted models confirm that the bicycle counts have a strong postive relationship with temperature, and a negative relationship with precipitation probability. During weekend and holiday, there is a significant drop in the bike counts. This makes sense since the Fremont bridge is primary a utilitarian bike facility. The daylight, or the seasons, are included to account for the seasonality in the bike count time series. With longer daylight, a larger daily bike count is observed.

Among five models that are considered, Model 3 has the highest R^2, adjusted R^2, and the smalledst AIC and BIC values. It could be due to the fact in Model 3, the interaction item (PrecipProb:Wknd) is included to account for the nonlinear relationship between bike counts and those two factors: rather than assuming the bike count depends on the precipitation probability and weekend indicator in an independent fashion, Model 3 assumes the combination of the two jointly impact on bike ridership. This makes sense sincde for utilitarian bicyclist, it is very likely if it rains AND it is weekend, they won't go out biking, whereas in the weekdays, they will still travel even if it's rainy. For simplicity, the weekend indicator is included to substitute the day of week variable. A R^2 value of 0.832 indicates that 83.2% of the variance in the bike count datasets is explained by the model fitted values.

To do: include model interpretation? Make model recommendation?

4.4 Generalized Additive Model (GAM)

To better explore the nonlinear relationship between the bike counts and weather factors such as temperature and precipitation, a generalized additive model (GAM) is fitted to help understand the relationship. The general GAM methodology is described in Section 3.6. Recall that a GAM has the following form:

$$\log(E[\text{Count}_t]) = \beta_0 + \beta_1 f_1(\text{TempMaxSq}_t) + \beta_2 \text{Holiday}_t + \beta_3 f_2(\text{PrecipProb}_t) + \beta_4 \text{Weekend}_t + \beta_5 \text{Daylight}_t. \tag{4.11}$$

where f_1 and f_2 are two nonlinear smooth functions of `TempMaxSq` and `PrecipProb`, respectively. A log link function is used to connect the expected bike counts with the independent variables. In this study, the `gam` function in the `mgcv` package of R is used to fit the GAM model. By

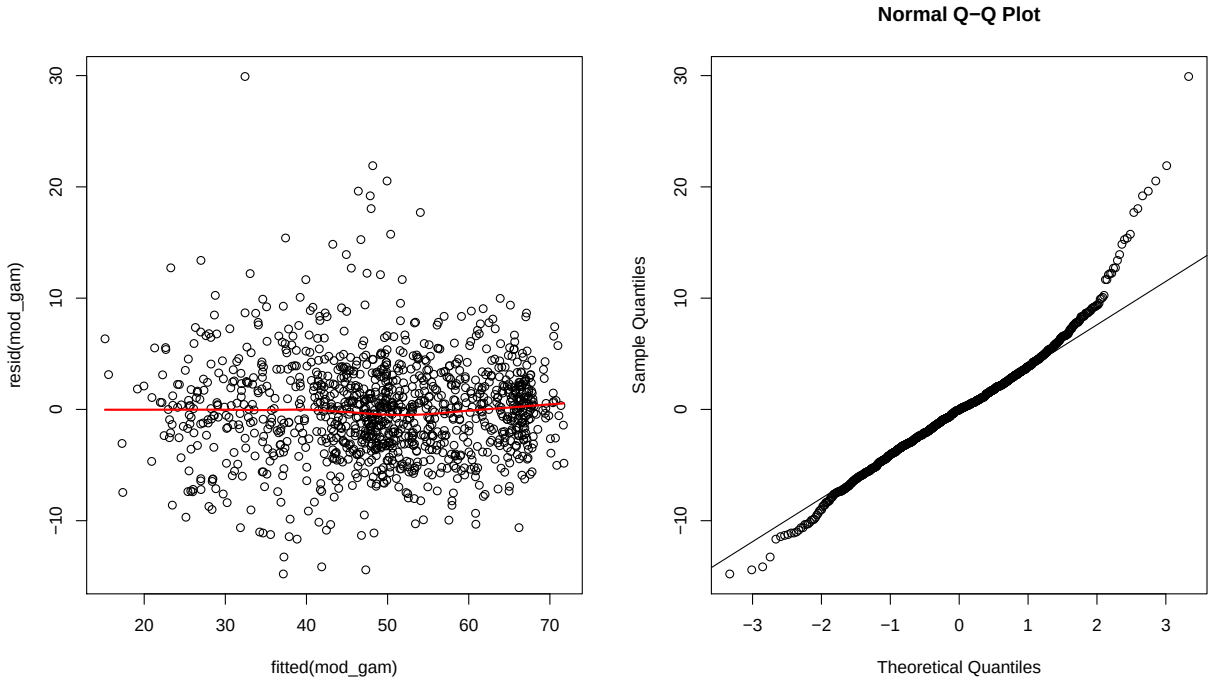

Figure 4.5: Residual plot and qq-plot for GAM fit.

default, f_1 and f_2 will take the form of the built-in nonparametric smooth splines. The smoothing parameter is chosen automatically using cross-validation.

The fitted GAM model is summarized in Table **??**. **Add one paragraph on the GAM model summary**. As seen in Figure 4.5, there is no discernable pattern in the Residual vs. Fitted value plot. Furthermore, the Normal Q-Q plot suggest that the residual follows an approximately Gaussian distribution.

We are interested to see the form of the smooth function f_1 and f_2 of temperature and precipitation probability. Figure 4.6 depicts the relationship between temperature/precipitation probability and bike counts while holding all other variables in the model constant. This allows to evaluate the nature of the influence of the respective predicator, in this case temperature and precipitation, on the dependent variable bike counts.

As seen in Figure 4.6, the smooth function of `temperatureMax` clearly follows an approximate nonlinear relationship. When the maximum daily temperature is below 40°F, the bike count will approximately remain the same; when it is in the range of 40°F and 75°, the bike counts monotically increases as the temperature rises. However, when the max temperature is above 75°F,

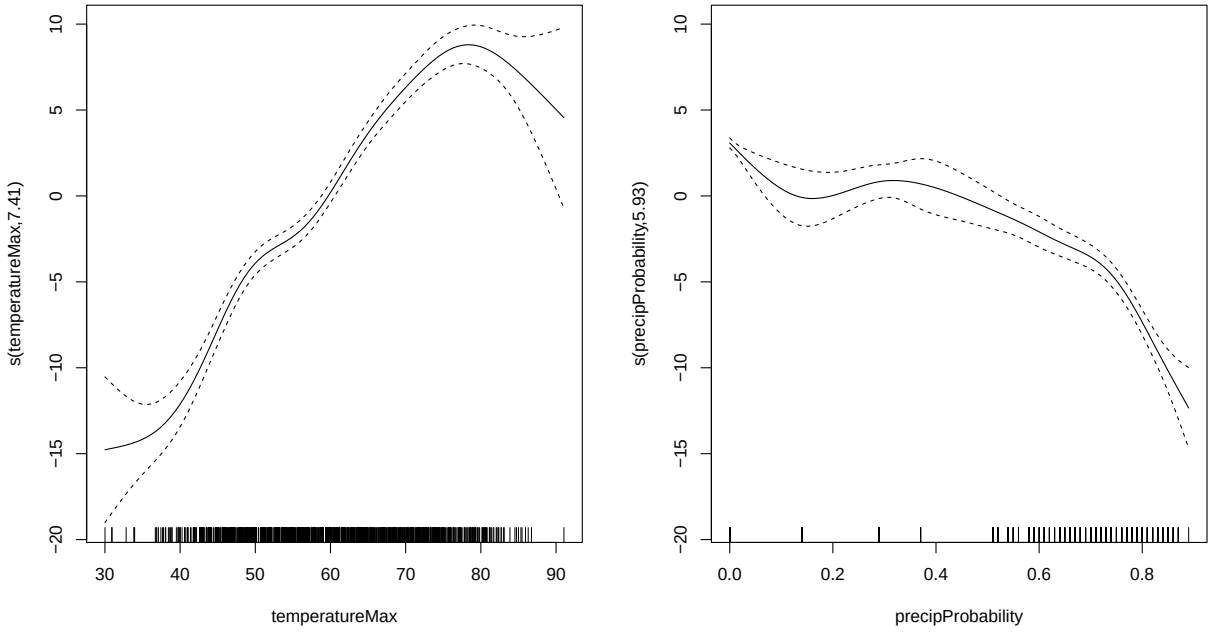

Figure 4.6: Smooth functions of temperatureMax and precipProbability

the bike counts will start drop. It is consistent with finding that is reported in other studies []. The use of GAM helps one to determine the important turning points. Such information can be used to set the threshold for a temperature scale variable for modeling.

Similarly, when the precipitation probability is smaller than 0.3, there is a relatively small negative impact with respect to the bike counts. The bike count follows a nearly linear relationship with a negative slope when the precipitation probability is between 0.3 and 0.7. When the precipitation probability is higher than 0.7, the bike count will experience the fastest drop. This finding is interesting in the sense that the utilitarian biker in Seattle will generally ignore the precipitation factor when make biking decisions unless it will rain with a very high probability. It also suggests three levels of precipitation scales could be used as an alternative to the continuous precipitation probability in the bike count model.

4.5 ARIMA Analysis

In this section, the Autoregressive Integrated Moving Average (ARIMA) model is used to account for possible autocorrelation relationship in the bike count time series. A review on the general ARIMA methodolgoy is provided in Section 3.7.

Following the Box-Cox approach, we first conducted the square root transformation to the dependent variable and fit a linear regression model with the same model specification as in Model 0. The standard Augemented Dickey-Fuller test and KPSS test are applied to the residuals of the fitted model. Their results are summarized in the following Table **??**. Also shown in the top plot in Figure 4.7, there is no discernable trend or seasonality in the residual of responses. Combining all these results together, we conclude that the residual of the fitted model is stationary and no differencing is required.

Add the stationary test table!!!!

The ACF, PACF plots are then investigated to examine any possible autocorrelation in the model residuals, as is shown in Figure 4.7. The ACF and PACF plots suggest there is still significant autocorrelation that needs to be accounted for. Since there are two peaks before the PACF falls below the significance threshold, a second order autoregressive term will be included in the final model.

Next, we use `auto.arima` function in the `forecast` package in `R` to automatically determine the best ARIMA model order. The chosen ARIMA model is of the order $(2,0,1)$, which is consistent of our intuition from the PACF plot.

Finally, we implement the ARIMA$(2,0,1)$ model using `arima` in the `forecast` package in `R`. For comparison, the ARIMA$(1,1,1)$ is also implemented and their results are summarized in Table 4.5 with model coefficients as well as goodness-of-fit indexes. Note that we include ARIMA$(1,1,1)$ because it is the output of `auto.arima` if we explicitly allow it to explore the model spaces where differencing is permitted.

As is shown in Table 4.5, ARIMA$(2,0,1)$ has a better performance compared to ARIMA$(1,1,1)$ as it has: 1) higher log likelihood, and 2) smaller AIC values. Also, the absence of differencing

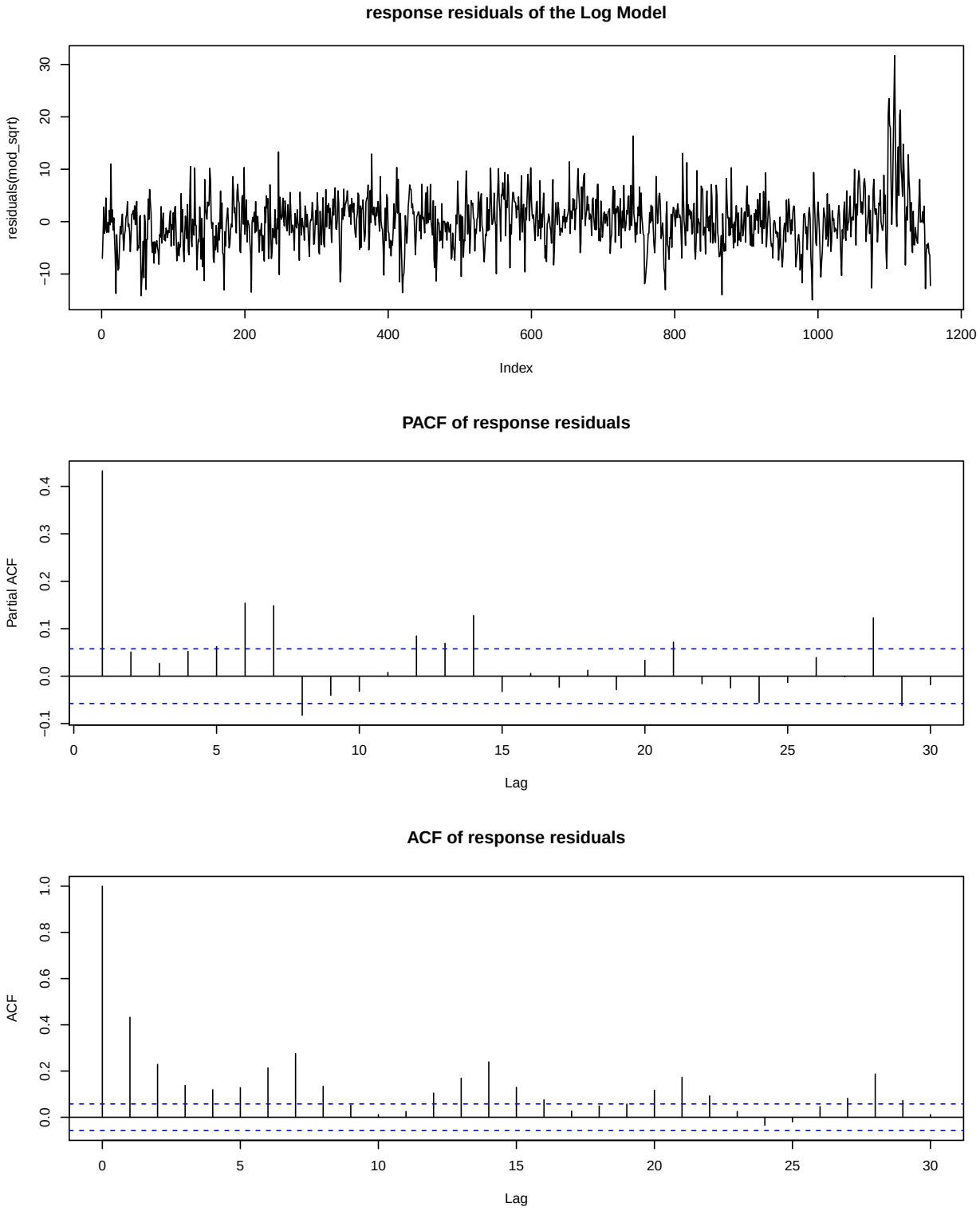

Figure 4.7: From top to bottom: the residual plots, ACF and PACF of the sqrt model residual.

step is preferred because of the simplicity.

Table 4.5: ARIMA Results

	Dependent variable: log(BikeCounts)	
	ARIMA(2,0,1)	ARIMA(1,1,1)
ar1	1.269^{***} (0.038)	0.355^{***} (0.033)
ar2	-0.303^{***} (0.039)	
ma1	-0.879^{***} (0.021)	-0.944^{***} (0.017)
intercept	8.252^{***} (0.136)	
temperatureMaxSq	0.0001^{***} (0.00004)	0.0001^{***} (0.00003)
holiday	-0.681^{***} (0.040)	-0.675^{***} (0.019)
precipProbability	-0.481^{***} (0.024)	-0.479^{***} (0.023)
Wknd	-0.790^{***} (0.016)	-0.787^{***} (0.013)
uw	0.111^{***} (0.040)	0.137^{***} (0.019)
daylight	0.00002	0.00002
Observations	1,157	1,156
Log Likelihood	45.906	37.183
σ^2	0.054	0.055
Akaike Inf. Crit.	-69.812	-56.366
Note:		*p<0.1; **p<0.05; ***p<0.01

4.6 Prediction

In this section, we are interested to evaluate the predicative power of the following models: lm model, log model, sqrt model, poisson model, glmlog model, arima log model and arima lm model.

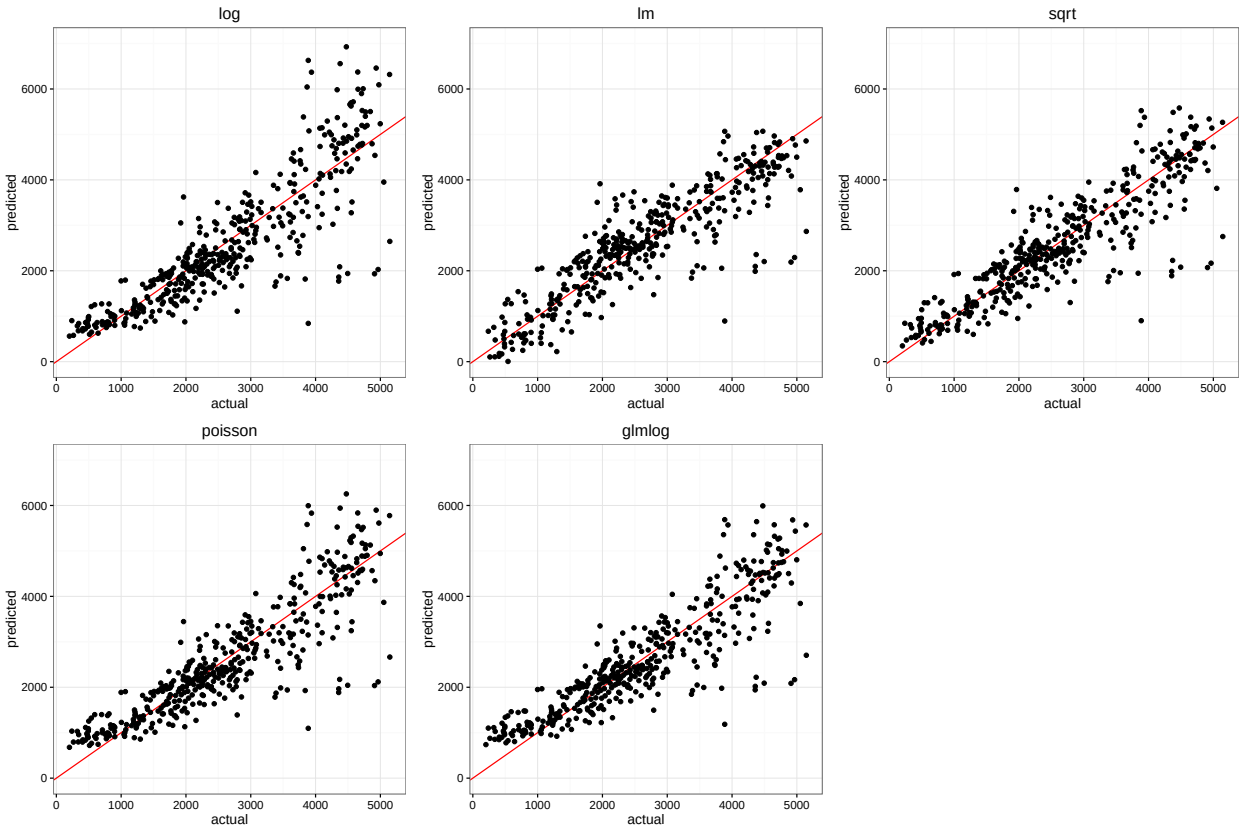

Figure 4.8: Actual vs predicted plots

To aid the comparison, the following analysis is performed: 1) prediction intervals, 2) confidence intervals, 3) prediction vs actual plots,

First of all, we use the entire 3 years' data to train the model and plot the actual data against fitted values, as in Figure 4.8.

Second, we use the first 2 years' data to train the model and then use the trained model to predict the 3rd year's bike counts. The actual vs predicted values plots are shown in Figure 4.9

For comparison of the predicative power, we also include the arima model for the lm model as well. See Figure 4.10 for cross validation and Figure 4.11 for fitted values.

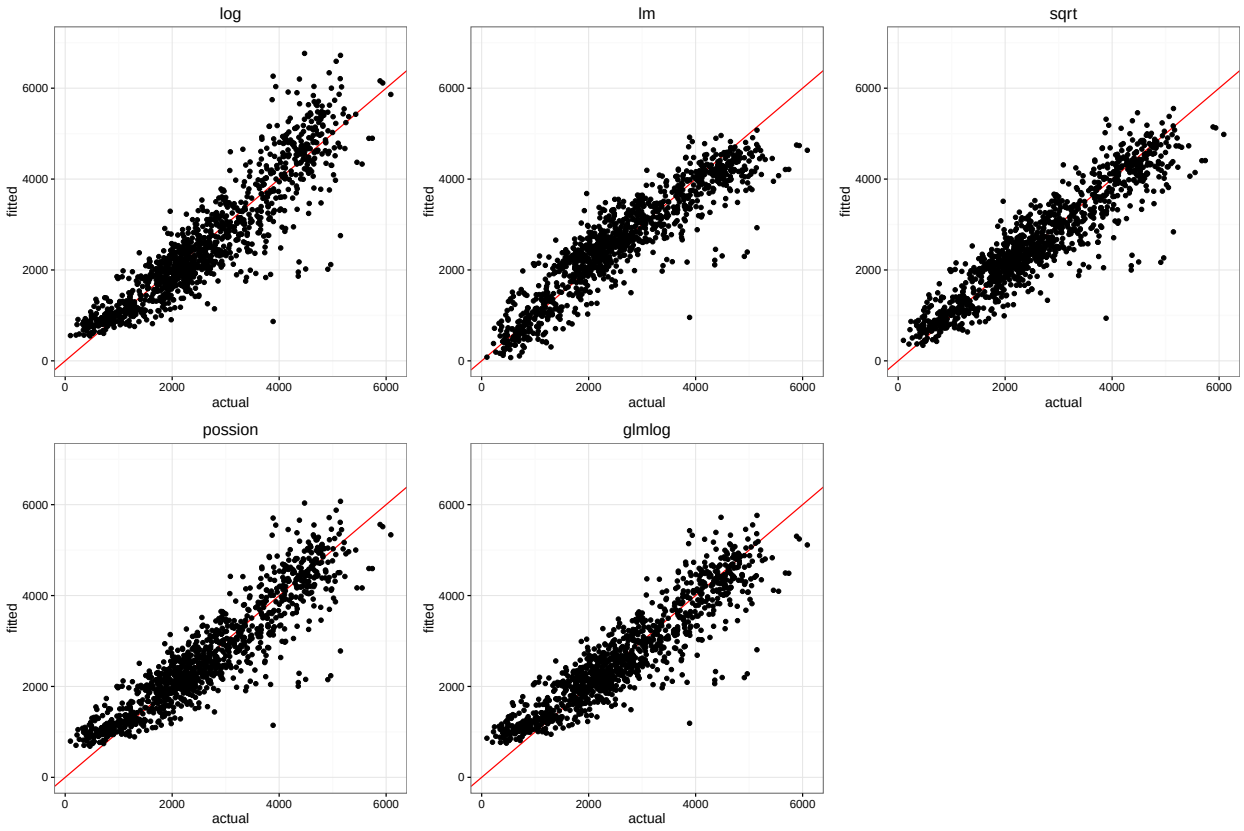

Figure 4.9: Actual vs fitted plots

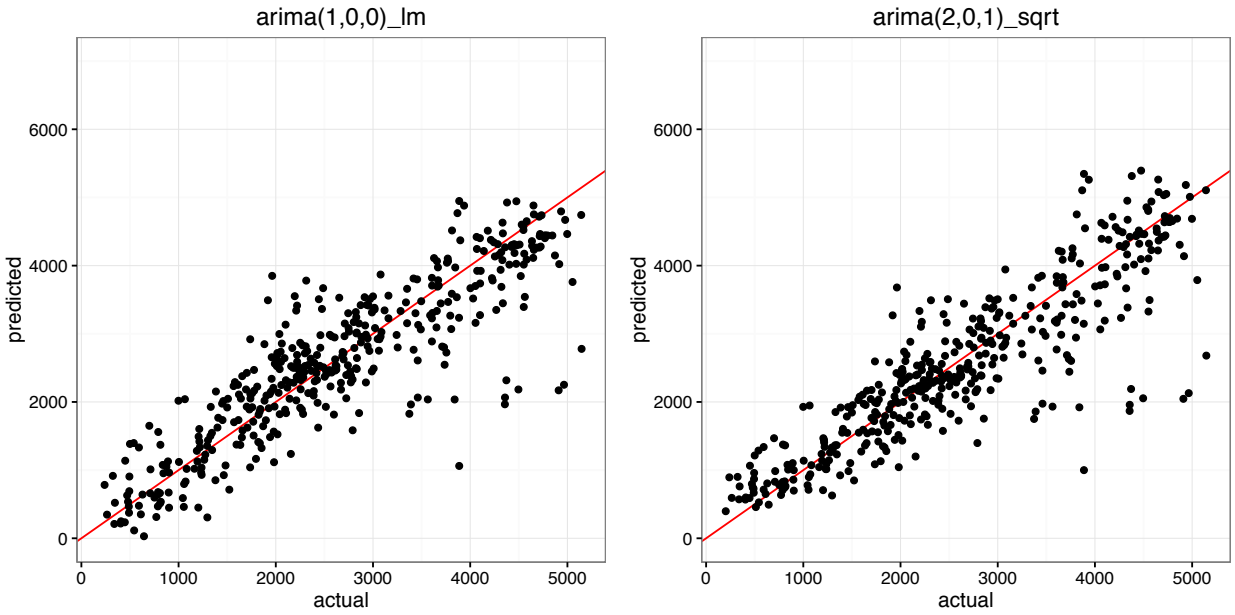

Figure 4.10: Actual vs Predicted plots

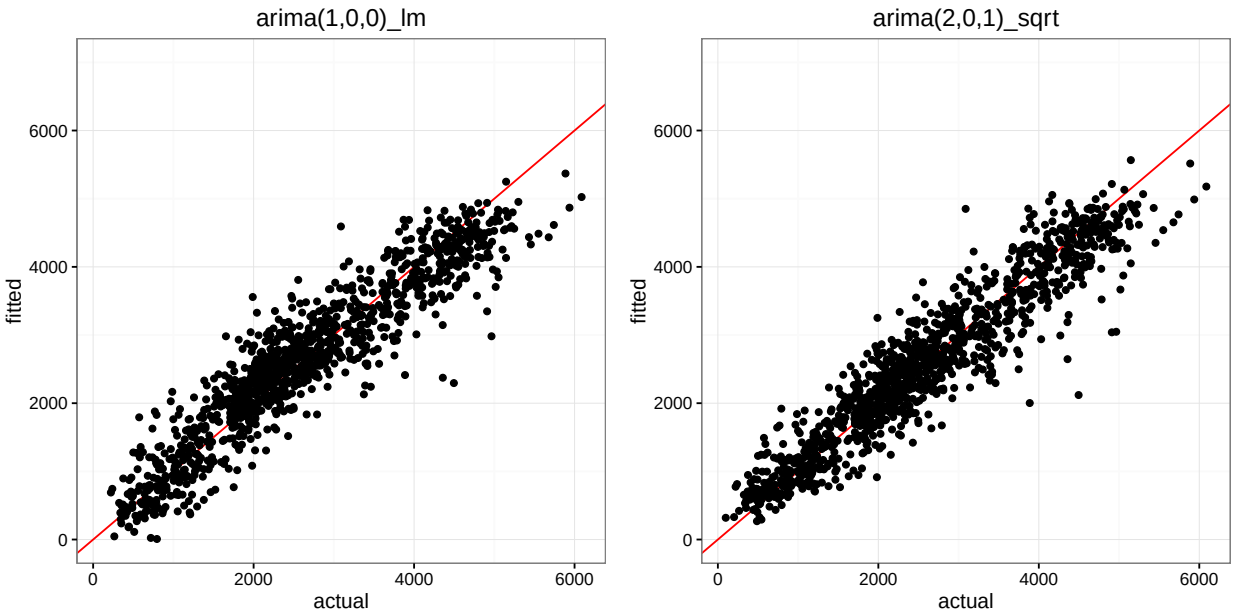

Figure 4.11: Actual vs Fitted plots

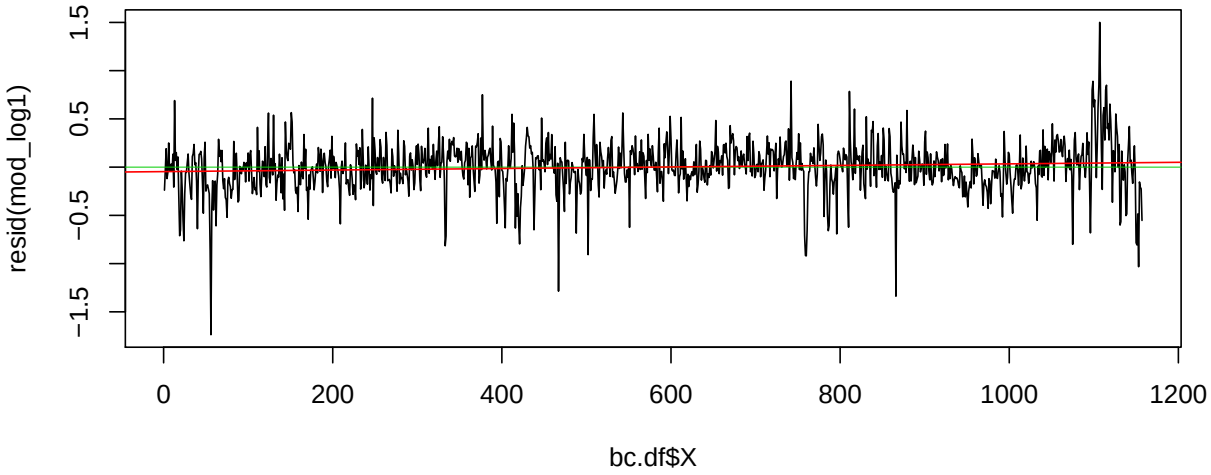

Figure 4.12: Bike count trend after removing weather and seasonal factors. Note the green line is zero, red line is the linear model fit to the residual error of log model.

4.7 Trend Analysis

We are also interested to reveal trend from the data. More specifically, we want to ask this question: after removing all weather and seasonality factors, does the bike count have an upward trend in the past three years? To answer this question, we fit a linear model to the residual of the log model. As is shown in Figure 4.12, there is a slight upward trend of the bike count in the past three years. Since we are looking at the residual error, the weather and other seasonal factors have already been removed. The red line (linear fit to the residual) has a positive slope indicates the bike count is increasing. However, such increase is very small considering the noise in daily bike count.

BIBLIOGRAPHY

Appendix A

WHERE TO FIND THE FILES

VITA

Jim Fox is a Software Engineer with UW Information Technology at the University of Washington. His duties do not include maintaining this package. That is rather an avocation which he enjoys as time and circumstance allow.

He welcomes your comments to `fox@uw.edu`.