Colin White

HOMEWORK ANALYSIS #8 - BIKE SHARING

Bike sharing systems are new generation of traditional bike rentals where the process from membership, rental and return back has become automatic. Through these systems, users are able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 Bike sharing programs around the world which is composed of over 500,000 bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

The Bike-sharing rental process is highly correlated with environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc. can affect the volume of rentals. This dataset is composed from the two-year historical data corresponding to years 2011 and 2012 from the Capital Bike share system in Washington D.C. The daily counts of the number of bikes used was extracted and then the corresponding weather and seasonal information was added. The dataset Bikes.csv contains the following variables:

Variable Name	Description
Season	season
yr	Year
holiday	Was the day a holiday (Y/N)?
workingday	Was the day a working day (Y/N)?
weathersit	Weather
temp	Temperature
hum	Humidity
windspeed	Windspeed
cnt	Number of bikes rented

CEO's hope to use the covariate information to predict the number of bikes rented based on environmental and season settings.

1. In your own words, summarize the overarching problem. Discuss how statistical modeling will be able to answer the posed questions.

The goal of this analysis is to *predict* how many bikes be rented based on various covariates that we have which are year, season, holiday, working day, weather, temperature, humidity, and wind speed. The model will take counts as a response, asses relationships between bike rental counts, and we will use these estimated covariate/count relationships to predict bike rental counts with

uncertainty.

2. Explain why traditional multiple linear regression methods are not (perfectly) suitable for this problem (even if they are approximately suitable).

Bike rental counts (the response variable) are counting number; therefore, a model that allows non-count values would be unsuitable. Thus, the normality assumption would not be correct, and we will instead assume that the data follow a Poisson distribution. There will likely be violations of the MLR model assumptions (equal variance, linearity)

3. Use variable selection to choose which variables to use in a Poisson regression model for cnt. Provide a justification of your choice in using AIC, BIC or CV. What factors do you find are important in predicting the number of bikes rented?

We are using AIC because it a derived as an approximation for cross-validation, and our primary goal is to *predict* bike counts. We are also using exhaustive method because it is better when it is possible to use. Using this, we select every covariate in the model: season, year, holiday, working day, weathersit, temp, hum, and wind speed.

4. Write out a Poisson regression model (using Greek letters) that includes your chosen covariates. Describe and justify any assumptions that you use in writing out your model.

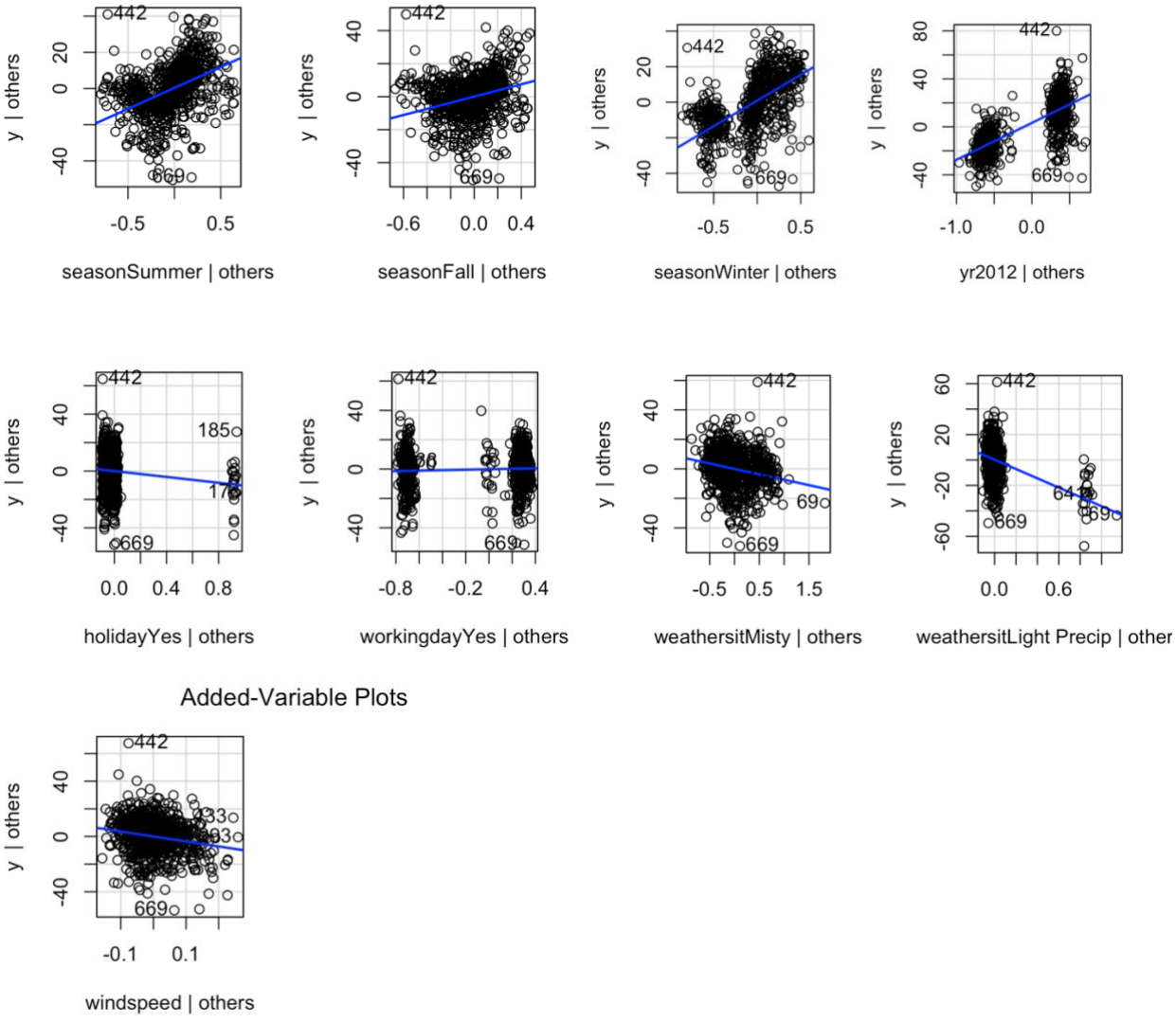
Let Y_i be the counts of bikes rented and x_i thru x_p be the covariates (including dummy variables) associated with season, yr, holiday, working day, weathersit, temp, hum, and wind speed.

$$Y_i \sim^{ind} \text{Poisson}(\mu_i)$$

$$\text{Log}(\mu_i) = \text{Beta}_0 + \sum_{j=1}^{p} x_{ij} * \text{Beta}_j$$

For this model I assume:
1. We assume that there is a log-linear relationship between the mean and the covariates. We asses below using...
2. We assume that the data is independent because we believe that the covariates explain the mean of counts and make the daily counts independent of one another. This may not be true because these data represent time-series.
3.We assume that the response variable come from a Poisson distribution because they are counts.

To assess the log-linearity assumption, we look at the added variable plots. The linearity assumption for most of the variable looks good enough. However, there appear nonlinearities in the AV-Plots for season. That said, because these season variables are dummy variables, there is really no other way we can put these in the models.


y | others
seasonSummer | others
seasonFall | others
seasonWinter | others
yr2012 | others
holidayYes | others
workingdayYes | others
weathersitMisty | others
weathersitLight Precip | other
Added-Variable Plots
windspeed | others

5. Fit the corresponding Poisson regression model and give a 95% confidence interval for each effect therein. Interpret at least one (but not the intercept) of these intervals in the context of the problem and NOT as log-effects.

We fit the model using `bestglm`. We provide confidence intervals below.

	2.5 %	97.5 %

seasonSummer (1.404, 1.417)
seasonFall (1.280, 1.294)
seasonWinter (1.583, 1.595)

yr2012 (1.579, 1.587)
holidayYes (0.841, 0.854)
workingdayYes (1.023, 1.028)

weathersitMisty (0.900, 0.906)
weathersitLight Precip (0.476, 0.486)
temp (3.354, 3.430)

hum (0.774, 0.791)
windspeed (0.554, 0.5720

So, if it is a “holiday” all else constant, we are 95% conflict that the mean of counts (Bikes sold) increases by a factor of between (0.841, 0.854).

6. Predict the average number of bikes rented (and construct a 95% confidence interval for the predicted mean) for the following day: season=“Spring”, yr=“2012”, holiday=“No”, workingday=“Yes”, weathersit=“Misty”, temp=0.34, hum=0.80, windspeed=0.18. Interpret your interval in context.

So the predicted mean number of bikes rented on a day described by problem 6 would be 3066.86. The 95\% confidence interval for this mean is (3055.592, 3078.170). In other words, we are 95% confident that when season is spring, yr="2012", hohliday="No", workingday="Yes", weathersit="Misty", temp=0.34, hum=0.80, windspeed=0.18, the mean counts (mean for all days with these characteristics) will be between (3055.592, 3078.170)

Appendix:

#This code was derived from in class examples, TA hours, and stack exchange

---
title: "Untitled"
author: "Colin White"
date: "4/8/2021"
output:
  pdf_document: default
  html_document:
    df_print: paged

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, cache = TRUE)
```

```{r}
library(bestglm)
library(car)

setwd("~/Desktop/1A School/1A Winter 2021/STAT330/HW9")

load("Bikes.RData")

head(bikes)

class(bikes$yr)
```
##1
##2
##3
```{r}
best_model = bestglm(bikes, family = poisson, method = "exhaustive", IC = "AIC")$BestModel

summary(best_model)
```

##4
```{r}
avPlots(best_model, layout = c(1,2))
```
##5
```{r}
library(knitr)
CI = confint(best_model)
kable(round(exp(CI),3))
```

```
##6
```{r}
pred_data = data.frame(season = "Spring", yr="2012", holiday="No", workingday="Yes",
weathersit="Misty", temp=0.34, hum=0.80, windspeed=0.18)

preds_se = predict.glm(best_model, newdata = pred_data,se.fit = TRUE)

CI_mean = preds_se$fit + c(-1,1) * qnorm(.975) * preds_se$se

exp(preds_se$fit)

exp(CI_mean)
```
```