Lab 1, Short Question

Contents

1 Political ideology (30 points)

These questions are based on Question 14 of Chapter 3 of the textbook "Analysis of Categorical Data with R" by Bilder and Loughin.

> An example from Section 4.2.5 examines data from the 1991 U.S. General Social Survey that cross-classifies people according to
>
> - Political ideology: Very liberal (VL), Slightly liberal (SL),Moderate (M), Slightly conservative (SC), and Very conservative (VC)
> - Political party: Democrat (D) or Republican (R)
> - Gender: Female (F) or Male (M).
>
> Consider political ideology to be a response variable, and political party and gender to be explanatory variables. The data are available in the file pol_ideol_data.csv.

1.1 Recode Data (2 points)

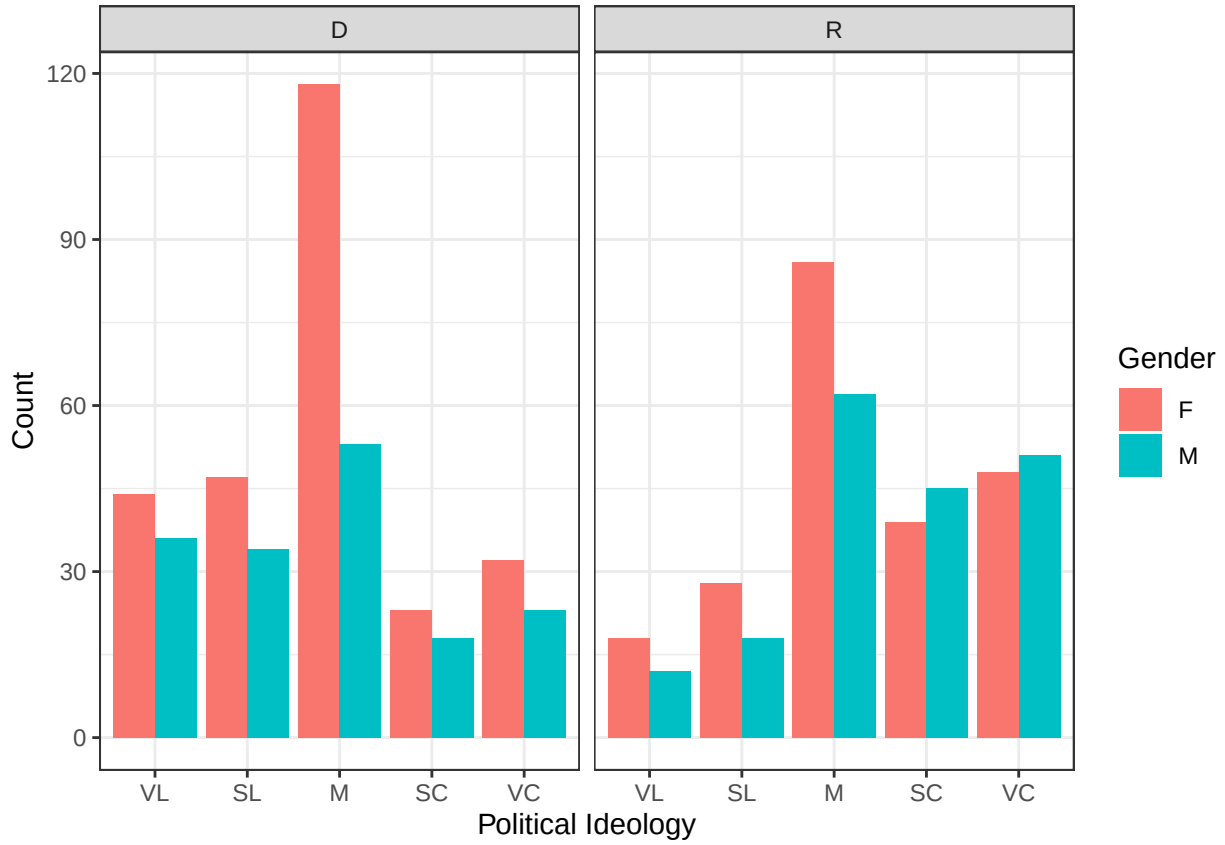
Use the factor() function with the ideology variable to ensure that R places the levels of the ideology variable in the correct order.

```
pol_ideol_data$ideol <- factor(pol_ideol_data_unfactored$ideol,
                               levels = c("VL", "SL", "M", "SC", "VC"),
                               ordered=T)
pol_ideol_data$gender <- factor(pol_ideol_data_unfactored$gender,
                                levels= c("F", "M"), ordered= F)
pol_ideol_data$party <- factor(pol_ideol_data_unfactored$party,
                               levels= c("D", "R"), ordered= F)
pol_ideol_data$count <- pol_ideol_data_unfactored$count
```

1.2 Test for Independence (5 points)

Analyze the relationships between political ideology and political party and gender using basic visualizations. Afterward, generate a contingency table and assess the independence of political ideology from political party and gender.

```
ggplot(pol_ideol_data, aes(x = ideol, y = count, fill = gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(~ party) +
  labs(x = "Political Ideology", y = "Count", fill = "Gender")
```



From the plots, it looks like people with 'Moderate' political ideology are a majority in both the parties and both the genders. Although females seem to be more moderate than males. The trend

also looks like VL and SL are higher in Democrats while SC and VC are higher in the Republican party.

```
# construct contingency table
ideol_gender.table <- xtabs(count ~ ideol + gender, data = pol_ideol_data)
ideol_party.table <- xtabs(count ~ ideol + party, data = pol_ideol_data)


# Chi sq test for independence, or assocstats() from vcd package, or summary()
chisq.test(x = ideol_party.table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ideol_party.table
## X-squared = 60.905, df = 4, p-value = 1.872e-12
```

```
chisq.test(x = ideol_gender.table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ideol_gender.table
## X-squared = 10.732, df = 4, p-value = 0.02975
```

The null and alternative hypothesis for the Chi-squared test are as below:

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$$
$$H_a : \pi_{ij} \neq \pi_{i+}\pi_{+j}$$

The Chi-squared test to assess the independence of political ideology and political party shows a significant p-value. We can reject the null hypothesis and say that there is evidence against independence.

For political ideology and gender, the p-value is significant again, but not as strong as between political ideology and party. Therefore, we reject they null hypothesis again.

Both gender and party have significant association with political ideology, with party having a much lower p-value meaning a stronger association.

1.3 Regression analysis (5 points)

Estimate a multinomial regression model and ordinal (proportional odds) regression model that both include party, gender, and their interaction. Perform Likelihood Ratio Tests (LRTs) to test the importance of each explanatory variable.

Also, test whether the proportional odds assumption in the ordinal model is satisfied. Based on this test and other results, which model do you think is more valid?

```r
# multinomial regression
mod.fit.multi <- multinom(formula = ideol ~ party + gender + party:gender,
                          data = pol_ideol_data, weight=count)
```

```
## # weights:  25 (16 variable)
## initial  value 1343.880657
## iter  10 value 1231.244704
## iter  20 value 1229.548447
## final  value 1229.543342
## converged
```

```r
summary(mod.fit.multi)
```

```
## Call:
## multinom(formula = ideol ~ party + gender + party:gender, data = pol_ideol_data,
##     weights = count)
##
## Coefficients:
##    (Intercept)    partyR     genderM partyR:genderM
## SL  0.06598601 0.3758637 -0.12315074      0.0867552
## M   0.98652431 0.5774673 -0.59976058      0.6779778
## SC -0.64869284 1.4219096 -0.04442702      0.5929326
## VC -0.31838463 1.2992041 -0.12968265      0.5957616
##
## Std. Errors:
##    (Intercept)    partyR   genderM partyR:genderM
## SL   0.2097724 0.3677971 0.3181097      0.5756306
## M    0.1766421 0.3136662 0.2790125      0.4944619
## SC   0.2573076 0.3839323 0.3867020      0.5799046
## VC   0.2323285 0.3610630 0.3538841      0.5518725
##
## Residual Deviance: 2459.087
## AIC: 2491.087
```

```r
# ordinal regression
mod.fit.ord <- polr(formula = ideol ~ party + gender + party:gender,
                    data = pol_ideol_data, weight= count)
summary(mod.fit.ord )
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = ideol ~ party + gender + party:gender, data = pol_ideol_data,
##     weights = count)
```

```
##
## Coefficients:
##                   Value Std. Error t value
## partyR           0.7562     0.1659  4.5593
## genderM         -0.1431     0.1820 -0.7861
## partyR:genderM   0.5091     0.2550  1.9965
##
## Intercepts:
##       Value    Std. Error t value
## VL|SL  -1.5521   0.1332   -11.6560
## SL|M   -0.5550   0.1157    -4.7965
## M|SC    1.1647   0.1226     9.5009
## SC|VC   2.0012   0.1364    14.6666
##
## Residual Deviance: 2470.15
## AIC: 2484.15
```

One multinomial regression model:

$$log(\hat{\pi}_{SL}/\hat{\pi}_{VL}) = 0.066 + 0.38partyR - 0.12genderM + 0.09partyR : genderM$$

One ordinal regression model:

$$logit(\hat{P}(Y \leq j)) = \hat{\beta}_{j0} - 0.76partyR + 0.14genderM - 0.5partyR : genderM$$

```
Anova(mod.fit.multi, test="LRT")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: ideol
##              LR Chisq Df Pr(>Chisq)
## party          60.555  4  2.218e-12 ***
## gender          8.965  4    0.06198 .
## party:gender    3.245  4    0.51763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(mod.fit.ord, test="LRT")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: ideol
##              LR Chisq Df Pr(>Chisq)
## party          56.847  1  4.711e-14 ***
## gender          0.843  1    0.35864
## party:gender    3.992  1    0.04571 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null and the alternative hypothesis for the multinomial model's LRT test are as below:

$$H_0 : \beta_{jr} = 0, j = 2, 3, .., J$$

$$H_a : \beta_{jr} \neq 0$$

From the multinomial regression model, it looks like the gender and the interaction between gender and party are insignificant, while party is statistically significant, given that gender and the interaction between gender and party are in the model.

The null and the alternative hypothesis for the ordinal model's LRT test are as below:

$$H_0 : \beta_r = 0$$

$$H_a : \beta_r \neq 0$$

From the ordinal regression model, it looks like gender is the only variable which is statistically insignificant. `party` variable is statistically significant and the interaction between party and gender is marginally significant, given the presence of other variables in the model. This is in contrast to the LRT test of the multinomial model in which only the party variable is statistically significant.

```
mod.fit.ord.po <- vglm(formula = ideol ~ party + gender + party:gender,
                       data = pol_ideol_data, weights=count,
                       family = cumulative(parallel = TRUE))
summary(mod.fit.ord.po)
```

```
##
## Call:
## vglm(formula = ideol ~ party + gender + party:gender, family = cumulative(parallel = TRUE),
##     data = pol_ideol_data, weights = count)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1   -1.5521     0.1335 -11.623  < 2e-16 ***
## (Intercept):2   -0.5550     0.1170  -4.742 2.11e-06 ***
## (Intercept):3    1.1647     0.1234   9.440  < 2e-16 ***
## (Intercept):4    2.0012     0.1368  14.626  < 2e-16 ***
## partyR          -0.7562     0.1669  -4.531 5.88e-06 ***
## genderM          0.1431     0.1794   0.798   0.4251
## partyR:genderM  -0.5091     0.2541  -2.004   0.0451 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3]), logitlink(P[Y<=4])
##
## Residual deviance: 2470.15 on 73 degrees of freedom
##
## Log-likelihood: -1235.075 on 73 degrees of freedom
```

```
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##         partyR        genderM partyR:genderM
##      0.4694436      1.1538080      0.6010255
```

```
mod.fit.ord.po@coefficients
```

```
##  (Intercept):1  (Intercept):2  (Intercept):3  (Intercept):4         partyR
##     -1.5520829     -0.5549909      1.1646520      2.0012148     -0.7562071
##        genderM partyR:genderM
##      0.1430678     -0.5091179
```

```
mod.fit.ord.npo <- vglm(formula = ideol ~ party + gender + party:gender,
                        data = pol_ideol_data, weights=count,
                        family = cumulative(parallel = FALSE))
summary(mod.fit.ord.npo)
```

```
##
## Call:
## vglm(formula = ideol ~ party + gender + party:gender, family = cumulative(parallel = FALSE)
##     data = pol_ideol_data, weights = count)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept):1      -1.6094     0.1651  -9.746  < 2e-16 ***
## (Intercept):2      -0.6424     0.1295  -4.961 7.01e-07 ***
## (Intercept):3       1.3350     0.1515   8.809  < 2e-16 ***
## (Intercept):4       1.9810     0.1886  10.505  < 2e-16 ***
## partyR:1           -0.8035     0.2963  -2.712  0.00670 **
## partyR:2           -0.6822     0.2104  -3.242  0.00119 **
## partyR:3           -0.9181     0.2050  -4.478 7.54e-06 ***
## partyR:4           -0.7105     0.2495  -2.848  0.00440 **
## genderM:1           0.3409     0.2507   1.360  0.17390
## genderM:2           0.3476     0.2042   1.702  0.08866 .
## genderM:3          -0.2364     0.2356  -1.004  0.31560
## genderM:4          -0.1677     0.2935  -0.572  0.56763
## partyR:genderM:1   -0.6136     0.4609  -1.331  0.18308
## partyR:genderM:2   -0.6844     0.3300  -2.074  0.03807 *
## partyR:genderM:3   -0.2231     0.3096  -0.721  0.47120
## partyR:genderM:4   -0.1146     0.3738  -0.307  0.75921
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3]), logitlink(P[Y<=4])
##
## Residual deviance: 2459.087 on 64 degrees of freedom
##
## Log-likelihood: -1229.543 on 64 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##         partyR:1         partyR:2         partyR:3         partyR:4
##        0.4477612        0.5054945        0.3992740        0.4913793
##        genderM:1        genderM:2        genderM:3        genderM:4
##        1.4062500        1.4157119        0.7894737        0.8455772
## partyR:genderM:1 partyR:genderM:2 partyR:genderM:3 partyR:genderM:4
##        0.5414141        0.5044029        0.8000631        0.8917482
```

```
mod.fit.ord.npo@coefficients
```

```
##    (Intercept):1    (Intercept):2    (Intercept):3    (Intercept):4
##       -1.6094379       -0.6424321        1.3350011        1.9810015
##         partyR:1         partyR:2         partyR:3         partyR:4
##       -0.8034952       -0.6822181       -0.9181073       -0.7105389
##         genderM:1        genderM:2        genderM:3        genderM:4
##        0.3409266        0.3476325       -0.2363888       -0.1677358
## partyR:genderM:1 partyR:genderM:2 partyR:genderM:3 partyR:genderM:4
##       -0.6135708       -0.6843800       -0.2230646       -0.1145715
```

```
tran.LR <- deviance(mod.fit.ord.po) - deviance(mod.fit.ord.npo)
df <- mod.fit.ord.po@df.residual - mod.fit.ord.npo@df.residual
p.value <- 1 - pchisq(q = tran.LR, df = df)
data.frame(tran.LR, df, p.value)
```

```
##    tran.LR df   p.value
## 1 11.06338  9 0.2713899
```

The hypotheses for the test of the proportional odds assumption are as below:

$$H_0 : \beta_{1r} = ... = \beta_{J-1,r} \text{ for } r = 1, ..., p$$

and

$$H_a : \text{Not all equal}$$

The LRT results in a statistic of 11.06 and a p-value of 0.27. Thus, there is not sufficient evidence to indicate the proportional odds assumption has been violated. Since the assumption has not been violated, the ordinal (proportional odds) regression model is more valid (as it is efficient and parsimonious with respect to coefficients).

1.4 Estimated probabilities (5 points)

Compute the estimated probabilities for each ideology level given all possible combinations of the party and gender levels.

```
newdata <- expand.grid(party = levels(pol_ideol_data$party),
                       gender = levels(pol_ideol_data$gender))

pi.hat.ord <- predict(object = mod.fit.ord, type = "probs", newdata=newdata)
pi.hat.ord
```

```
##           VL         SL         M        SC        VC
## 1 0.17478325 0.18992129 0.3974724 0.1187475 0.1190756
## 2 0.09043739 0.12184718 0.3884304 0.1757193 0.2235657
## 3 0.19639087 0.20206353 0.3886812 0.1080000 0.1048645
## 4 0.06450371 0.09295524 0.3531387 0.1960087 0.2933937
```

1.5 Contingency table of estimated counts (5 points)

Construct a contingency table with estimated counts from the model. These estimated counts are found by taking the estimated probability for each ideology level multiplied by their corresponding number of observations for a party and gender combination.

For example, there are 264 observations for gender = "F" and party = "D". Because the multinomial regression model results in $\hat{\pi}_{VL} = 0.1667$, this model's estimated count is $0.1667 \times 264 = 44$.

- Are the estimated counts the same as the observed? Conduct a goodness of fit test for this and explain the results.

```
total_counts <- with(pol_ideol_data, tapply(count, list(party, gender), sum))
total_counts
```

```
##     F   M
## D 264 164
## R 219 188
```

```
estimated_counts_ord <- sweep(pi.hat.ord, 2, total_counts, "*")
```

```
## Warning in sweep(pi.hat.ord, 2, total_counts, "*"): length(STATS) or dim(STATS)
## do not match dim(x)[MARGIN]
```

```r
# Construct table to evaluate goodness of fit
c.table <- xtabs(formula = count ~ party + ideol + gender, data = pol_ideol_data)
obv.data <- ftable ( x = c.table, row.vars = c("gender" , "party"),
                     col.vars = "ideol")
obv.data
```

```
##               ideol  VL  SL   M  SC  VC
## gender party
## F      D             44  47 118  23  32
##        R             18  28  86  39  48
## M      D             36  34  53  18  23
##        R             12  18  62  45  51
```

Looks like our predicted (multinomial) and observed values are the same.

```r
chisq.test(p = obv.data, estimated_counts_ord)
```

```
##
##  Pearson's Chi-squared test
##
## data:  estimated_counts_ord
## X-squared = 79.107, df = 12, p-value = 6.108e-12
```

Looking at the results of our Chisq test, we have a statistically significant result and therefore we reject the null hypothesis that the two samples come from a common distribution.

1.6 Odds ratios and confidence intervals (8 points)

To better understand relationships between the explanatory variables and the response, compute odds ratios and their confidence intervals from the estimated models and interpret them.

```r
# Odds ratio
round(exp(-mod.fit.ord$coefficients), 2)
```

```
##         partyR        genderM partyR:genderM
##           0.47           1.15           0.60
```

The estimated odds of VL vs. SL or M or SC or VC is 0.47 times as large for Republicans compared to Democrats, holding the other variables constant.

The estimated odds of VL vs. SL or M or SC or VC is 1.15 times as large for Males compared to Females, holding the other variables constant, but this term is not significant.

```r
# Confidence Interval
coef.beta <- confint(mod.fit.ord, level = 0.95)
```

```
## Waiting for profiling to be done...
```

```
##
## Re-fitting to get Hessian
```

```r
round(exp(-coef.beta),2)
```

```
##                 2.5 % 97.5 %
## partyR           0.65   0.34
## genderM          1.65   0.81
## partyR:genderM   0.99   0.36
```

With 95% confidence interval, the odds of political ideology being below a particular level change by 0.34 to 0.65 times for Republicans vs. Democrats, holding the other variables constant.

W271 Group Lab

Bike Share Demand

Anson Quon, Nayan Ganguli, Zach Zimmerman

Contents

1 Introduction (5 points)

In major cities around the world, transportation by bike has an ever-growing importance. It offers a healthy, environmentally friendly alternative way of commuting for locals and leisure for tourists. In dense urban environments with limited living space, the bike rental industry offers an affordable alternative to private bike ownership. Understanding the factors that drive hourly bike rentals can help optimize services to meet user demands more effectively.

In this study we will be analyzing the hourly demand for bike rentals in one of the most densely populated urban areas in the world: Seoul, South Korea. Our data encompasses a one-year period from December 1, 2017 to November 30, 2018. The data set records the date, weather and atmospheric conditions, hour of day, as well as whether the day was a holiday or the service was functioning.

Our initial hypothesis is that bike rental demand increases during the holiday evenings in the season of summer. We believed that tourism and school holidays during summer on top of with people returning from work would lead to increased bike rental demand during the evenings. To test this hypothesis, we will create and compare various regression models. This analysis aims to aid in better planning of resource allocation for the bike-sharing program to provide a stable supply of rental bikes at all hours of the day.

2 Data (20 points)

2.1 Description (5 points)

The data set we will be using in our analysis is Seoul Bike Sharing Demand set from the UCI machine learning repository, DOI **10.24432/C5F62R.** This data set is licensed for full usage under the Creative Commons Attribution License 4.0 International. The data comes from the Seoul Bike Sharing System, a citywide un-manned bike-sharing system designed to provide affordable easy access to bike transport to pedestrians. Here, the population is the set of all possible bike rental events, with random variables linked to weather and time contributing to the observed outcomes. Samples are produced by gathering the sum total of all bike rental events within a single hour on a specific day.

The data set contains 8760 entries of 13 features, with no missing values. The features are a mix of integer, continuous, categorical, and binary values. The outcome variable (hourly bike rentals) is an integer. The integer features are snowfall (cm), rainfall (mm), humidity (%), visibility (10m), and hour of day (24h time). The continuous features are temperature (C), wind speed (m/s), dew point temperature (C), and solar radiation (MJ/m^2). There is one categorical feature, season, and two binary features, holiday and functioning day.

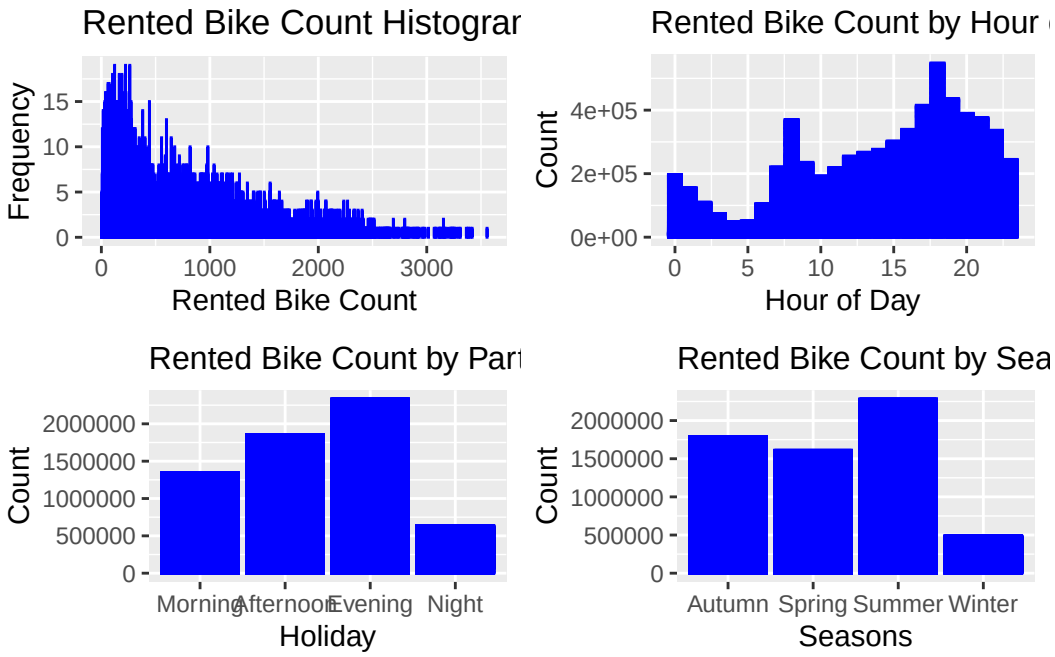
2.2 EDA (15 points)

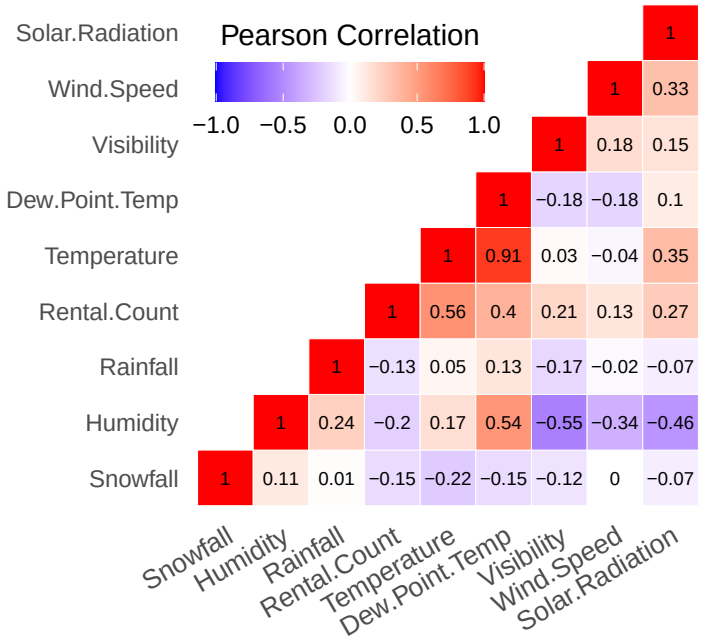
We first looked at the head of the dataset to examine its structure. We observed that the categorical and binary variables along with the date were represented by strings and needed to be factored. Before factoring variables, we verified the claim that the dataset had no missing values.

Next we explored the distribution of our response variable. We noticed a significant number of entries with a response of 0. One of the columns in our dataset is Functioning Day, whether or not bikes were available for rental that day. We confirmed that for all non-functioning days, the total number of bike rentals is 0. Conversely, we confirmed it was always a non-functioning day for entries with a rented bike count of 0. This is an example of complete separation, a variable that can perfectly predict the response variable. This could lead to convergence issues and bias when training our regression model. As such, we chose to omit data taken on non-functioning days. Since the goal of our model is to predict the number of bikes rented per hour, our analysis is more meaningful if we stick exclusively to days that bikes are available for rental. Including the other features on entries taken during non-functioning days would affect the bias of our model towards predicting no bikes being rented. As such, we opted to drop the rows where Functioning.Day was "No". This resulted in 295 rows dropped, a 3.34% loss of data. We do not expect such loss negatively affect the inference power of our model. The bar plot of the filtered data shows a right-skewed Poisson-like distribution.

We then visualized some variable relationships. First, since our goal was to predict hourly bike rentals let's look at the bike rental counts by hour. We observed peak times in the morning and evening. This intuitively corresponds rather well with the standard 9-5 working hours. However, we wanted to bin the hours into a lower-dimensional variable. Allowing every hour to be its own level of a categorical variable risks issues like overfitting, interpretability issues from too many parameters, and computational burden. We chose to bin hours into four even parts of the day: 6-12 (morning), 12-18 (afternoon), 18-24 (evening), and 24-6 (night). Another complex feature to consider was

the date. We saw that grouping by day of the week may have some effect on the frequency of bike rentals (plot not shown). Grouping by month (not shown) also shows an effect, but this was considered redundant as we already had the "season" column. We opted to include the day of the week for our analysis.





To explore relationships between the response variable and numerical data, we used a correlation matrix (shown above) and visualized with scatter plots. Scatter plots are not shown for brevity, but all indicated some relationship between the response variable and explanatory variables.

3 Model Development (40 points)

3.1 Poisson regression (10 points)

Table 1: Stargazer Table for Model 1

	Dependent variable:
	Rental.Count
SeasonsSpring	-0.213^{***}
	(0.001)
SeasonsSummer	0.112^{***}
	(0.001)
SeasonsWinter	-1.405^{***}
	(0.002)
Part.Of.DayAfternoon	0.323^{***}
	(0.001)
Part.Of.DayEvening	0.551^{***}
	(0.001)
Part.Of.DayNight	-0.745^{***}
	(0.002)
HolidayNo Holiday	0.090^{***}
	(0.002)
Constant	6.604^{***}
	(0.002)
Observations	8,465
Log Likelihood	−1,065,890.000
Akaike Inf. Crit.	2,131,795.000
Note:	*p<0.1; **p<0.05; ***p<0.01

	LR Chisq	Df	Pr(>Chisq)
Seasons	1349569.213	3	0
Part.Of.Day	1114864.132	3	0
Holiday	1711.225	1	0

To test our initial hypothesis, we performed a poisson regression on the data, linking the rental counts to our explanatory variables through an exponential function to ensured predicted responses are positive and greater than 0. Our exploratory data analysis showed that the response variable was

Table 2: Confidence Intervals for Poisson Model 1

	2.5 %	97.5 %
(Intercept)	734.5391252	741.3554063
SeasonsSpring	0.8062149	0.8096554
SeasonsSummer	1.1162350	1.1206187
SeasonsWinter	0.2445794	0.2461355
Part.Of.DayAfternoon	1.3777414	1.3838698
Part.Of.DayEvening	1.7304987	1.7378589
Part.Of.DayNight	0.4734949	0.4763270
HolidayNo Holiday	1.0892326	1.0986685

approximately Poisson distributed, justifying our choice of model. We performed a likelihood ratio test using the `Anova` function from the `car` package, and examined the confidence intervals for the exponential of the coefficients. These are interpreted as the multiplicative effect of a 1-unit increase in that variable holding all other variables constant. None of the confidence intervals contained 1, indicating all the anticipated multiplicative effects were statistically different than 1.

Our base case for this and following models is `Autumn` for `Seasons`, `Morning` for `Part.Of.Day` and `Holiday` for the variable `Holiday`. We see that compared to this base case, `SeasonsSpring`, `SeasonsWinter` and `Part.Of.DayNight` reduce the mean estimate of rented bike count. The remaining explanatory variables increase the expected estimate compared to the base case.

Looking at the variables of interest from our hypothesis and at the output of the multiplicative effects, for `SeasonsSummer`, holding all other variables as constant, the expected count of rented bikes increases by 11.8% compared to the autumn season. Likewise, for part of the day, our expected count of rented bikes increases by 73.4% in the evening compared to the mornings. Lastly, our expected count of rented bikes increases by 9.4% when there is no holiday compared to when there is a holiday, contrary to our belief about holidays causing an increase in bike rentals. This indicates we possibly underestimated the daily commuter usage of the bike sharing system.

3.2 Model Comparison (10 points)

To capture any potential omitted variable bias, we introduced all the explanatory variables present in the data set into a secondary model, except dew point temperature since the correlation matrix showed it had a correlation coefficient of 0.91 with temperature. We observed all explanatory variables are statistically significant through an likelihood ratio test and generating confidence intervals (not shown for brevity).

We then introduced quadratic and interaction terms to a third model. We chose to use the quadratic terms of the weather data, and modeled interactions between the Part of Day and season, whether it was a weekday or weekend, and the interaction between holiday and season. We also introduced a new binary variable for whether it was a weekday or weekend. An LRT and the confidence intervals (not shown) indicated all terms except for `Weekday_WeekendWeekend` are statistically significant. However, the interactions between part of day and weekend is statistically significant, meaning that the effect of`Weekday_WeekendWeekend` is significant only when considered in combination with `Part.Of.Day`.

Table 3: Stargazer Table for Models 2 and 3

	Dependent variable:	
	Rental.Count	
	(1)	(2)
Temperature	0.033*** (0.0001)	0.084*** (0.0002)
I(Temperature^2)		−0.002*** (0.00001)
Humidity	−0.009*** (0.00003)	0.014*** (0.0001)
I(Humidity^2)		−0.0002*** (0.00000)
Wind.Speed	0.022*** (0.0005)	0.082*** (0.001)
I(Wind.Speed^2)		−0.015*** (0.0003)
Visibility	−0.00002*** (0.00000)	0.0001*** (0.00000)
I(Visibility^2)		−0.00000*** (0.000)
Solar.Radiation	−0.093*** (0.001)	0.182*** (0.002)
I(Solar.Radiation^2)		−0.083*** (0.001)
SeasonsSpring	−0.190*** (0.001)	0.024*** (0.006)
SeasonsSummer	−0.215*** (0.001)	−0.028*** (0.006)
SeasonsWinter	−0.960*** (0.002)	−0.566*** (0.008)
Rainfall	−0.559*** (0.002)	−0.644*** (0.003)
I(Rainfall^2)		0.019*** (0.0001)
Snowfall	−0.118*** (0.002)	0.044*** (0.005)
I(Snowfall^2)		−0.030*** (0.001)
HolidayNo Holiday	0.176*** (0.002)	0.234*** (0.003)
Part.Of.DayAfternoon	0.046*** (0.001)	0.012*** (0.002)
Part.Of.DayEvening	0.369*** (0.001)	0.362*** (0.002)
Part.Of.DayNight	−0.728*** (0.002)	−0.661*** (0.003)
Weekday_WeekendWeekend		−0.003 (0.002)
Part.Of.DayAfternoon:Weekday_WeekendWeekend		0.015*** (0.003)
Part.Of.DayEvening:Weekday_WeekendWeekend		−0.005** (0.002)
Part.Of.DayNight:Weekday_WeekendWeekend		−0.051*** (0.003)
SeasonsSpring:Part.Of.DayAfternoon		0.149*** (0.003)
SeasonsSummer:Part.Of.DayAfternoon		0.116*** (0.003)
SeasonsWinter:Part.Of.DayAfternoon		−0.324*** (0.004)
SeasonsSpring:Part.Of.DayEvening		0.036*** (0.003)
SeasonsSummer:Part.Of.DayEvening		0.298*** (0.003)
SeasonsWinter:Part.Of.DayEvening		−0.396*** (0.004)
SeasonsSpring:Part.Of.DayNight		−0.100*** (0.004)
SeasonsSummer:Part.Of.DayNight		0.239*** (0.004)
SeasonsWinter:Part.Of.DayNight		−0.068*** (0.006)
SeasonsSpring:HolidayNo Holiday		−0.256*** (0.006)
SeasonsSummer:HolidayNo Holiday		−0.172*** (0.006)
SeasonsWinter:HolidayNo Holiday		0.311*** (0.007)
Constant	6.778*** (0.004)	5.574*** (0.006)
Observations	8,465	8,465
Log Likelihood	−748,807.300	−614,626.400
Akaike Inf. Crit.	1,497,645.000	1,229,329.000

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4: Information Criterion for Fitted Poisson Models

Model	AIC	BIC	AICc
model_poisson_1	2131795	2131852	2131795
model_poisson_2	1497645	1497750	1497645
model_poisson_3	1229329	1229597	1229329

We compared the generated models using three information criterion tests: AIC, BIC, and corrected AIC (AICc). With the lowest information criterion across all three tests, our third model minimizes the mean square error of prediction, accounting for both the error in estimating the coefficients and the variability of the data. Despite AICc's typically higher penalty for small samples, its effect aligns with AIC due to our large sample size (8465 records). With the lowest BIC, the third model is also considered the closest to the true underlying model, assuming the true model is among the models being examined. Since the third model has the lowest AIC, BIC, and AICc values, we consider it our best model. On the contrary, model 2 and model 1 do not fit the data as well as `models_poisson_3` does since their AIC, BIC, and AICc values are consistently higher than those of model 3, with model 1 being 'worse' than model 2.

3.3 Model Assessment (10 points)

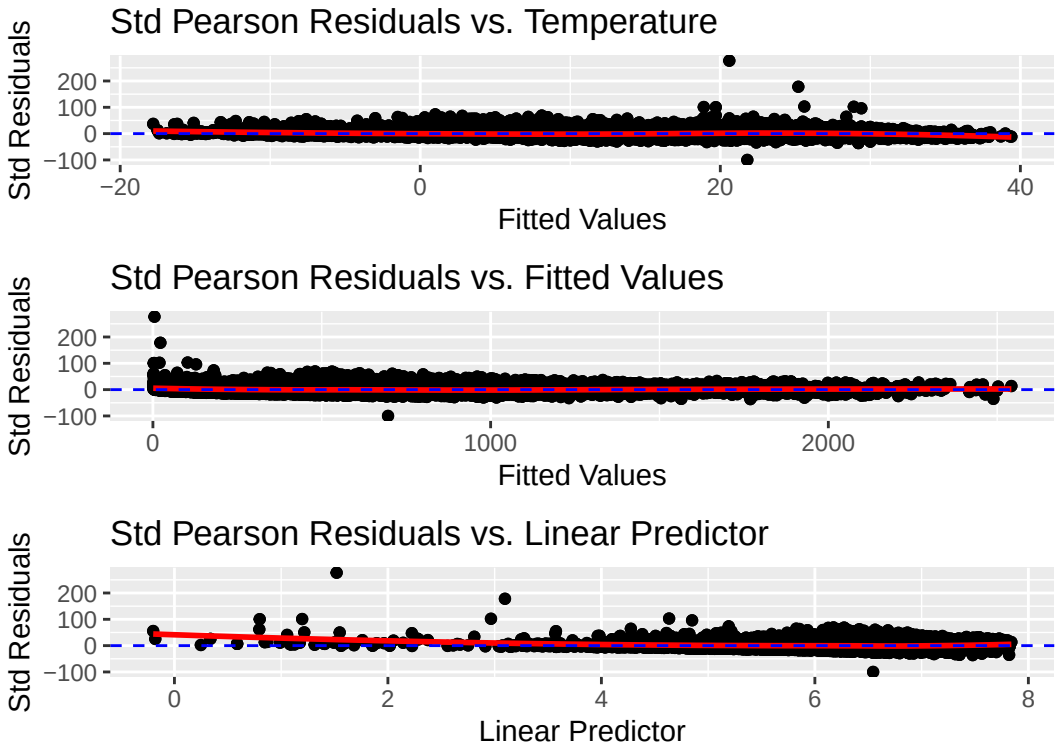
Since our best model is model 3, we will do further assessment comparing to this model.

Poisson regression requires four assumptions to be met - (1) Independent and Identically Distributed (IID) Data, (2) Distribution of response follows Poisson Distribution, (3) Mean of distribution is linked to explanatory variables in a linear fashion, (4) Link relates to explanatory variables in a linear fashion.

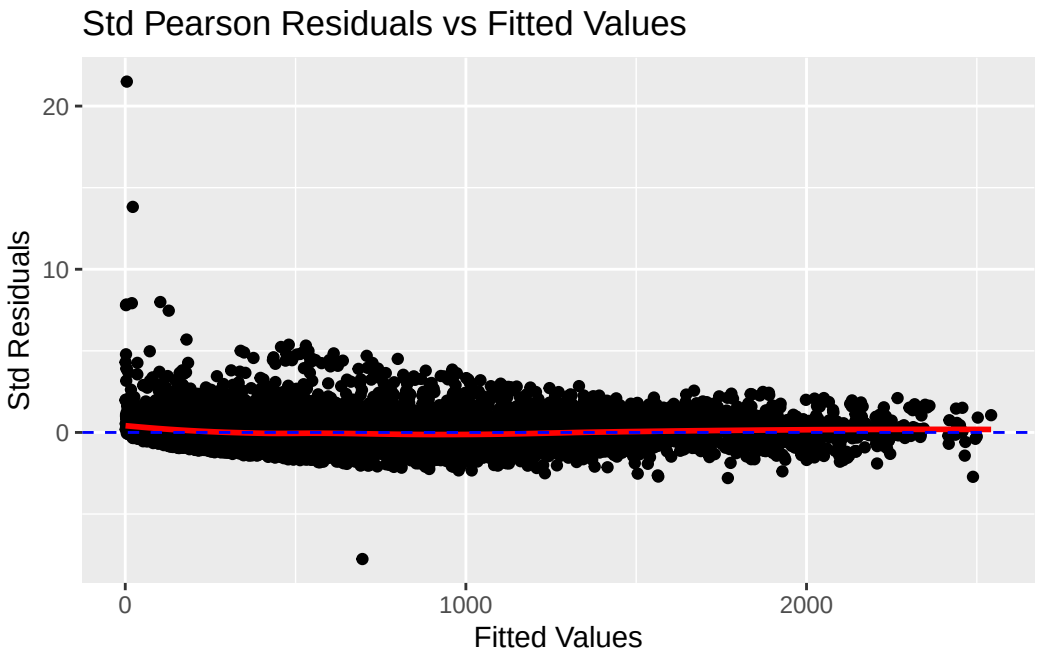
Each record in our dataset corresponds to hourly bike rental event counts from December 1, 2017, to November 30, 2018, accompanied by weather data for the same intervals. Given the nature of weather, which tends to be stable over short periods within a specific geographic area, the characteristics of one record could imply similarities in adjacent records. Consequently, this temporal and spatial consistency in weather conditions suggests that records might exhibit dependence, particularly for those closely timed within the same day. Thus, the IID assumption is potentially compromised.

Our earlier EDA showed the distribution of the number of bike rental events is right skewed and non-negative, making the Poisson distribution appropriate.

The plot of the Standardized Pearson Residuals against the Temperature shows consistent variance, showing that our model uses the appropriate transformations of Temperature that fits the data well. The plot of the Standardized Pearson Residuals against the Fitted Values shows roughly the same variance. Since there are no patterns or curvature in the plot, the link function fits well. The plot of the Standardized Pearson Residuals against the Linear Predictor values shows roughly the same variance. Since there is no significant curvature in the plot, the link function fits well.

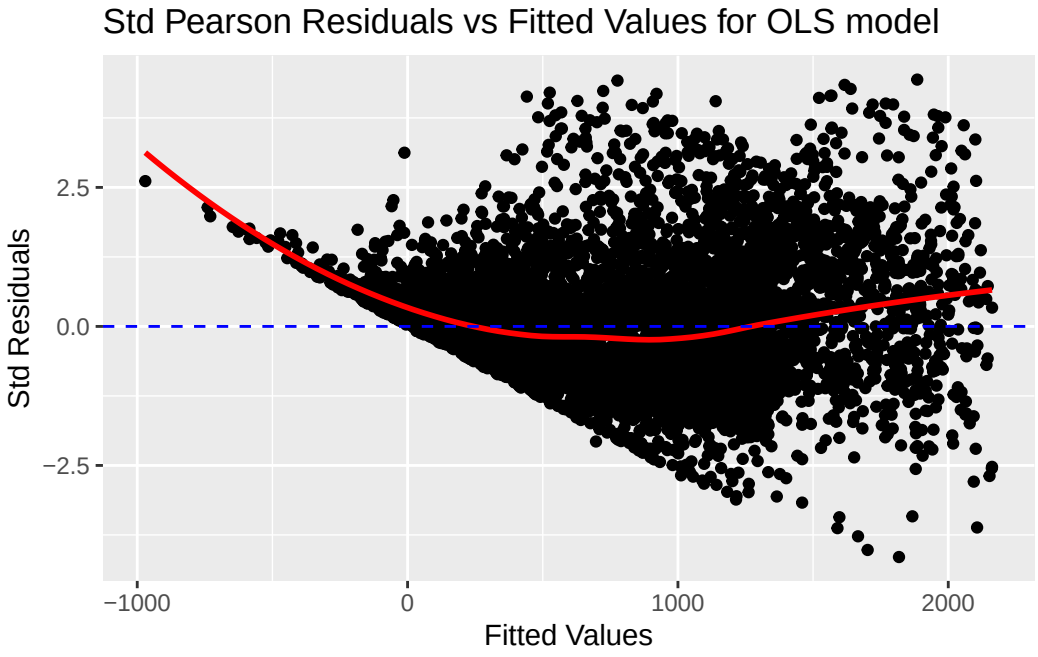


We assess goodness of fit through the Pearson Statistic and Residual Deviance. We obtain small p-values for both, indicating that the null hypothesis that the model is correct should be rejected. Indeed, applying a dispersion test gives an estimated dispersion of 164 with a p-value of nearly 0, indicating that the data is overdispersed and the model does not fit well. To account for overdispersion, we will fit a quasi-poisson model, which allows for the variance to be a function of the mean, thus accommodating the observed overdispersion. This approach modifies the variance assumption to be proportional to the mean, enabling the model to better handle the extra variability in the data. By using a quasi-likelihood approach, the model provides flexibility in modeling count data with overdispersion, offering more accurate standard errors and significance tests for the coefficients without specifying a full probability model. This adjustment aims to improve the goodness of fit and ensure that our conclusions about the relationship between the predictors and the response variable are robust. Our quasi-poisson model had higher deviance than the poisson model 3, and did not show a visible difference with model 3 for the plot of residuals vs fitted values. This indicates that the quasi-poisson model may not be the right choice to account for overdispersion here. We attempted to fit a negative binomial model as well, but were unable to achieve convergence with our model estimation.



3.4 Alternative Specification (10 points)

For comparison, we also tried an OLS regression model with the same explanatory variables as model 3. The model had an R^2 of 0.65, indicating a poor linear fit. The plot of the residuals vs the fitted values shows clear curvilinear properties, indicating that the fit of the OLS model is very poor. Additionally, OLS model responses are not constrained to nonnegative numbers the way a linked poisson regression is, so we observe predictions of negative bike rentals which is obviously nonsensical. Clearly, a poisson regression is the right choice in this case.



4 Conclusion (5 points)

Our study employed statistical models to examine the link between Seoul's bike rental activity and environmental/temporal factors, aiming to identify demand spikes to assist in logistical planning of bike supply. We initially hypothesized increased rentals during holiday evenings in summer. Utilizing Poisson regression due to the count nature and distribution of rental data, statistically significant effects were found for holidays, parts of the day, and seasons. Analysis confirmed peaks in summer and evenings, but contrary to our initial hypothesis, holidays had a reductive effect. We then developed more advanced models incorporating additional variables and quadratic/interaction terms. Using three information criterion, we identified the third model as the most appropriate. Overdispersion was identified, suggesting the need for model refinement or alternative approaches. We fitted a Quasi-Poisson regression model to address this. However, the Quasi-Poisson model's residual deviance was higher than that of the selected Poisson model, indicating that it may not have been the right choice to account for overdispersion. We attempted to also fit a negative binomial model but were unable to achieve convergence. Additionally, an OLS model was fitted to the data to examine whether Poisson was the right choice. The OLS model performed very poorly, indicating that it Poisson was likely the best choice for this count response study.

Our model indicated that the most significant effect on bike demand was the time of day, the weather, and the season. It appears that bike demand is not heavily influenced by whether it is a holiday or weekend, indicating that it is mostly used for commuting purposes in Seoul. Demand is reduced in the winter and at night, and is nearly halved on rainy days or at night or in winter, holding all other variables constant. Warmer, sunnier days also increase the expected mean bike rental. It seems that in general, bike demand is highest in the evenings when there is still sunlight (e.g. summer evenings). This can be intuited as biking home in the twilight being more desirable than biking home in darkness, as winter months get darker much sooner in the day. Rainy evenings being associated with decreased mean bike rentals can also be intuited as dangerous commuting conditions that people may avoid. The results seem generally consistent with the statement that "demand for commuter bikes is higher on nicer days with plenty of light".

The model faces several limitations, including overdispersion, where the variance exceeds the mean, challenging the model's assumption of equal mean and variance. Independence of observations is violated due to likely temporal and spatial correlations among rentals, reducing reliability. Increased complexity from added variables complicates interpretation and risks overfitting. Its context-specific findings limit generalizability to other locales without adjustments. Moreover, the linear nature of Poisson regression may not capture the real-world non-linear relationships between variables, suggesting that alternative models like Negative Binomial regression might provide better fit and accuracy, albeit at the cost of increased model complexity and potentially further reduced interpretability.