Combined Linear Model

1.

Fit a linear model using the response and the discrete predictor variables identified previously in your dataset. Report and interpret the estimated coefficients and their associated standard errors and p-values. Use diagnostic methods to assess the validity of the standard assumptions. Use graphics and offer brief comments on what you observe.

The discrete variables we decided to use for our model are: - mnth - workingday - weathersit

```
# loda data
hours_clean = read.csv('../data/hour_clean.csv')
hours_dis = hours_clean[, c("mnth", "workingday", "weathersit", "cnt")]
```

```
# one-hot-encode and standardize
library(fastDummies)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# there is only one instance of 4 in weathersit so we replace it with 3.
hours_dis$weathersit[hours_dis$weathersit == 4] = 3

# one-hot-encode
hours_dis = dummy_cols(hours_dis,
                   select_columns = c("mnth", "weathersit"),
                   remove_first_dummy = TRUE,
                   remove_selected_columns = TRUE)
hours_dis = dplyr::select(hours_dis, -cnt, cnt)
head(hours_dis)
```

```
##   workingday mnth_2 mnth_3 mnth_4 mnth_5 mnth_6 mnth_7 mnth_8 mnth_9 mnth_10
## 1          0      0      0      0      0      0      0      0      0       0
## 2          0      0      0      0      0      0      0      0      0       0
## 3          0      0      0      0      0      0      0      0      0       0
## 4          0      0      0      0      0      0      0      0      0       0
## 5          0      0      0      0      0      0      0      0      0       0
## 6          0      0      0      0      0      0      0      0      0       0
##   mnth_11 mnth_12 weathersit_2 weathersit_3 cnt
## 1       0       0            0            0  16
## 2       0       0            0            0  40
## 3       0       0            0            0  32
```

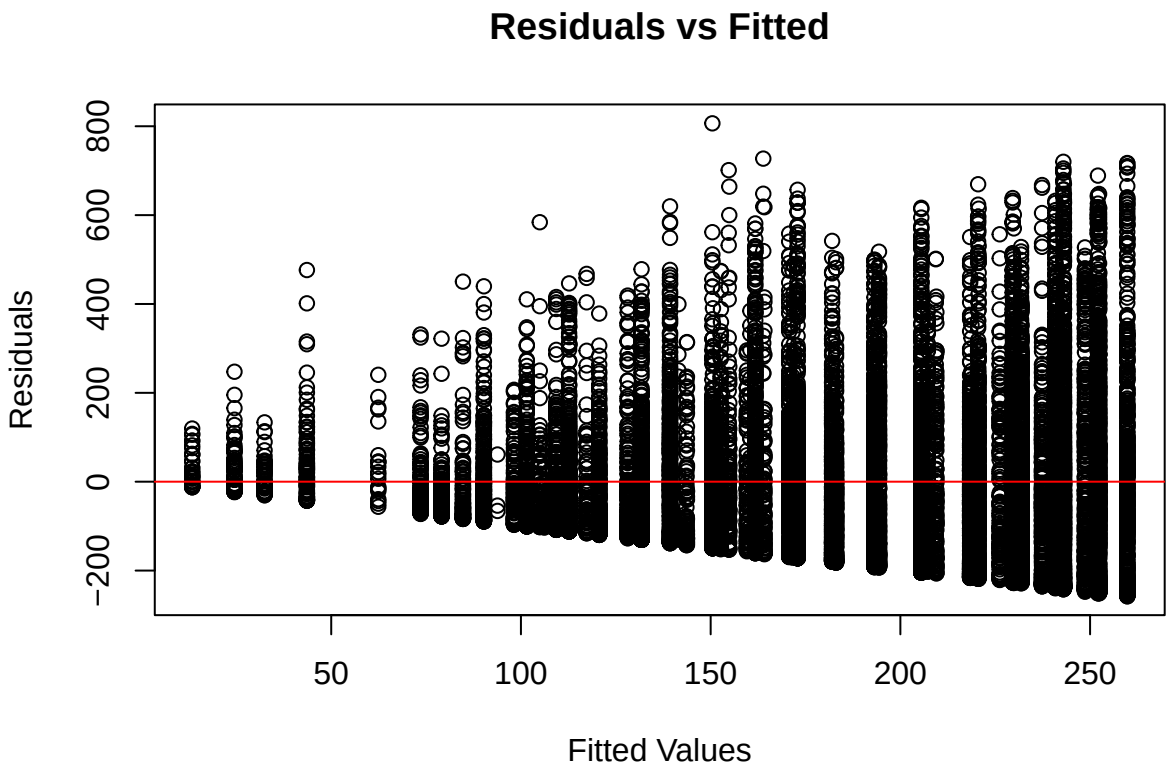
```
## 4       0       0            0            0  13
## 5       0       0            0            0   1
## 6       0       0            1            0   1
```

```
# fit linear model
model_dis = lm(cnt ~ ., data=hours_dis)
summary(model_dis)
```

```
##
## Call:
## lm(formula = cnt ~ ., data = hours_dis)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -257.8 -120.1  -35.7   83.3  806.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   101.549      5.056  20.086  < 2e-16 ***
## workingday     11.150      2.841   3.924 8.72e-05 ***
## mnth_2         19.086      6.648   2.871   0.0041 **
## mnth_3         60.220      6.475   9.301  < 2e-16 ***
## mnth_4         92.818      6.510  14.257  < 2e-16 ***
## mnth_5        128.156      6.456  19.850  < 2e-16 ***
## mnth_6        139.676      6.519  21.425  < 2e-16 ***
## mnth_7        130.407      6.468  20.163  < 2e-16 ***
## mnth_8        139.368      6.524  21.362  < 2e-16 ***
## mnth_9        147.110      6.510  22.599  < 2e-16 ***
## mnth_10       130.263      6.509  20.014  < 2e-16 ***
## mnth_11        80.470      6.509  12.363  < 2e-16 ***
## mnth_12        49.056      6.461   7.592 3.30e-14 ***
## weathersit_2  -22.467      3.070  -7.317 2.64e-13 ***
## weathersit_3  -88.200      4.942 -17.849  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 173.1 on 17263 degrees of freedom
## Multiple R-squared:  0.09246,    Adjusted R-squared:  0.09172
## F-statistic: 125.6 on 14 and 17263 DF,  p-value: < 2.2e-16
```

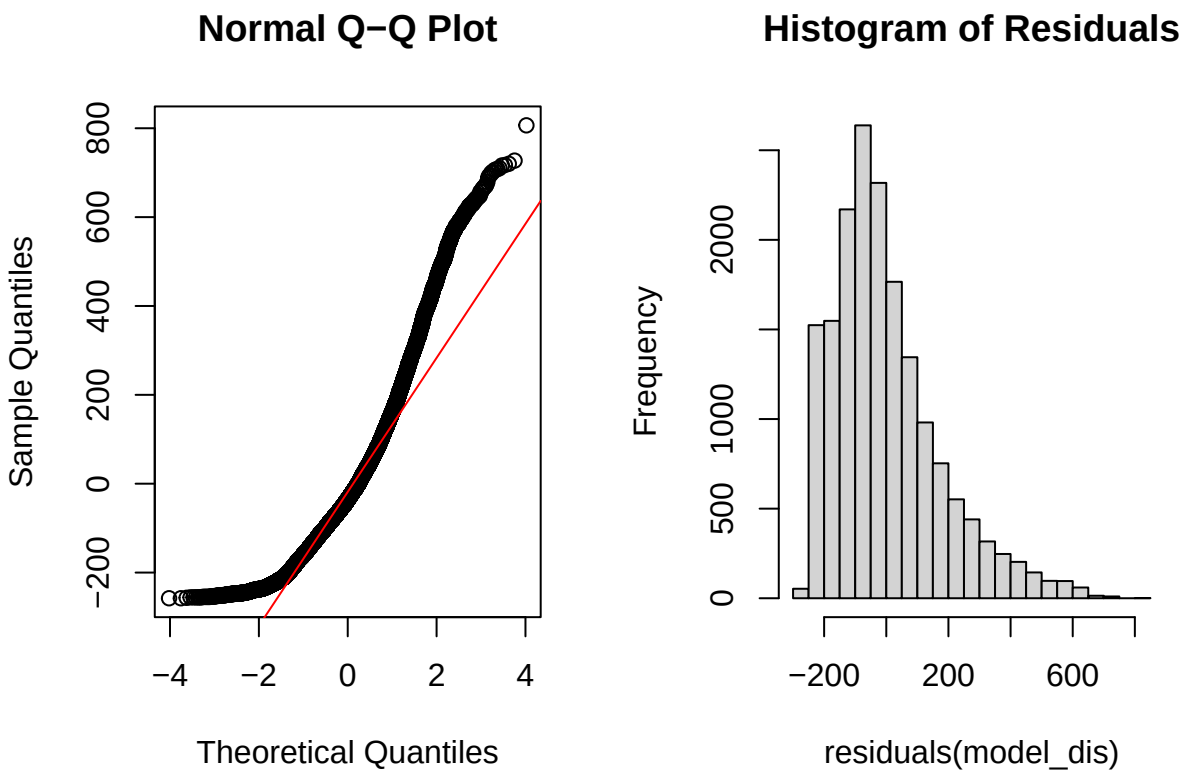
The linear regression model explains approximately 9.17% of the variation in bike rental counts (Adjusted $R^2 = 0.09172$), which indicates a poor model. The coefficients indicate that working days are associated with an increase of about 11 rentals, while summer months like June, July, and August show substantially higher rental counts compared to January. Weather situations also play a significant role: less favorable conditions (conditions 2 and 3) are linked to significant drops in rentals, with light rain reducing counts by about 88. All predictors in the model are statistically significant, indicating they have meaningful associations with rental activity.

```
# check for homoscedasticity
par(mfrow = c(1, 1))
plot(model_dis$fitted.values, residuals(model_dis),
     main = "Residuals vs Fitted",
     xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red")
```



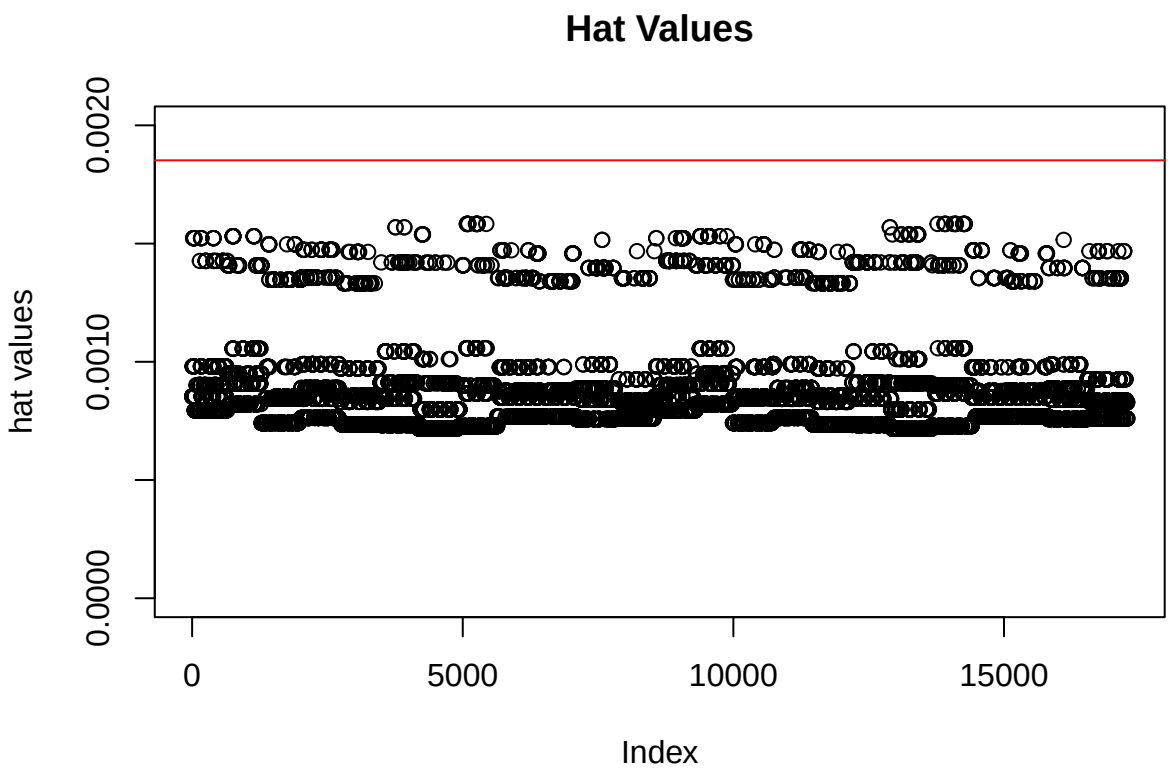
The residuals vs. fitted plot reveals a distinct fan shape, where the spread of residuals increases with higher fitted values. This pattern indicates heteroscedasticity which violates a key assumption of linear regression. This can lead to biased standard errors and unreliable significance tests.

```
par(mfrow = c(1, 2))
qqnorm(residuals(model_dis));
qqline(residuals(model_dis), col = "red")
hist(residuals(model_dis),
     main = "Histogram of Residuals",
     breaks = 30)
```



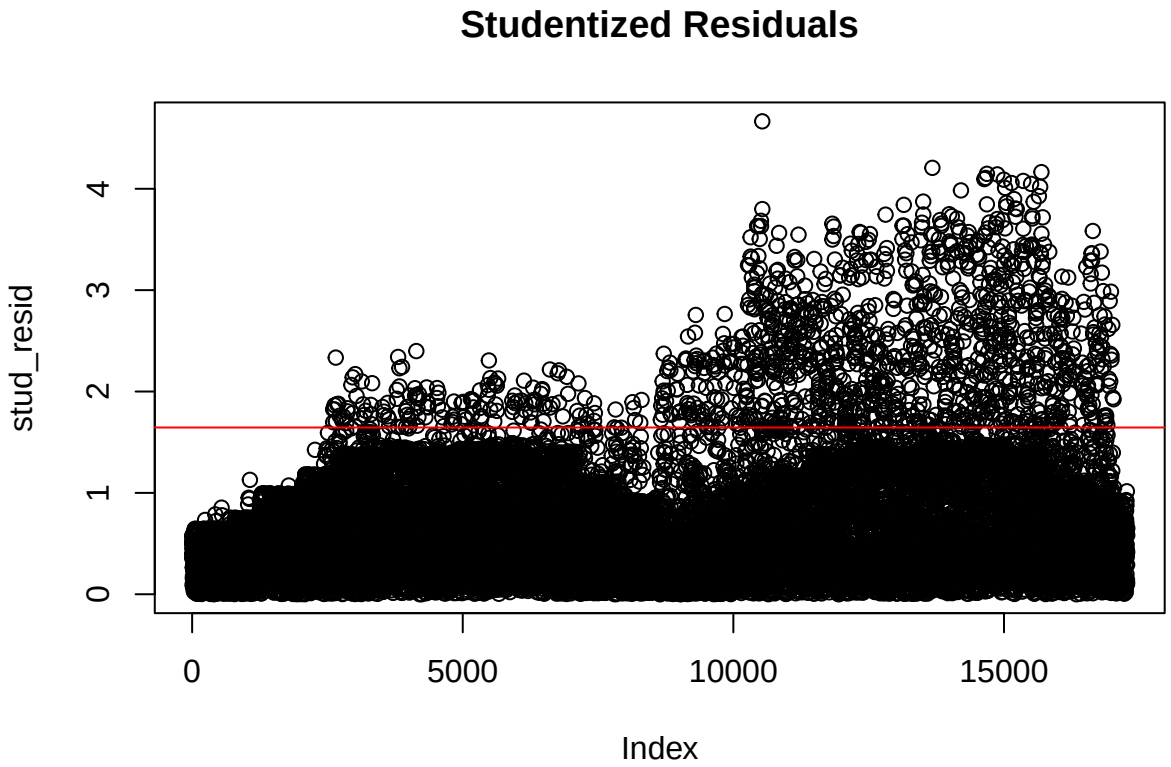
The Q-Q plot shows some deviation from the diagonal, especially in the upper tail and also some in the lower tail, indicating that the residuals are not perfectly normally distributed. The histogram of residuals show a right skew. These results suggest that the normality assumption may be violated, particularly at the extremes.

```
hat_values <- hatvalues(model_dis)
plot(hat_values,
     ylab = "hat values",
     main = "Hat Values",
     ylim = c(0, 0.002))
abline(h = 2 * (ncol(hours_dis) + 1) / nrow(hours_dis),
       col = "red")
```



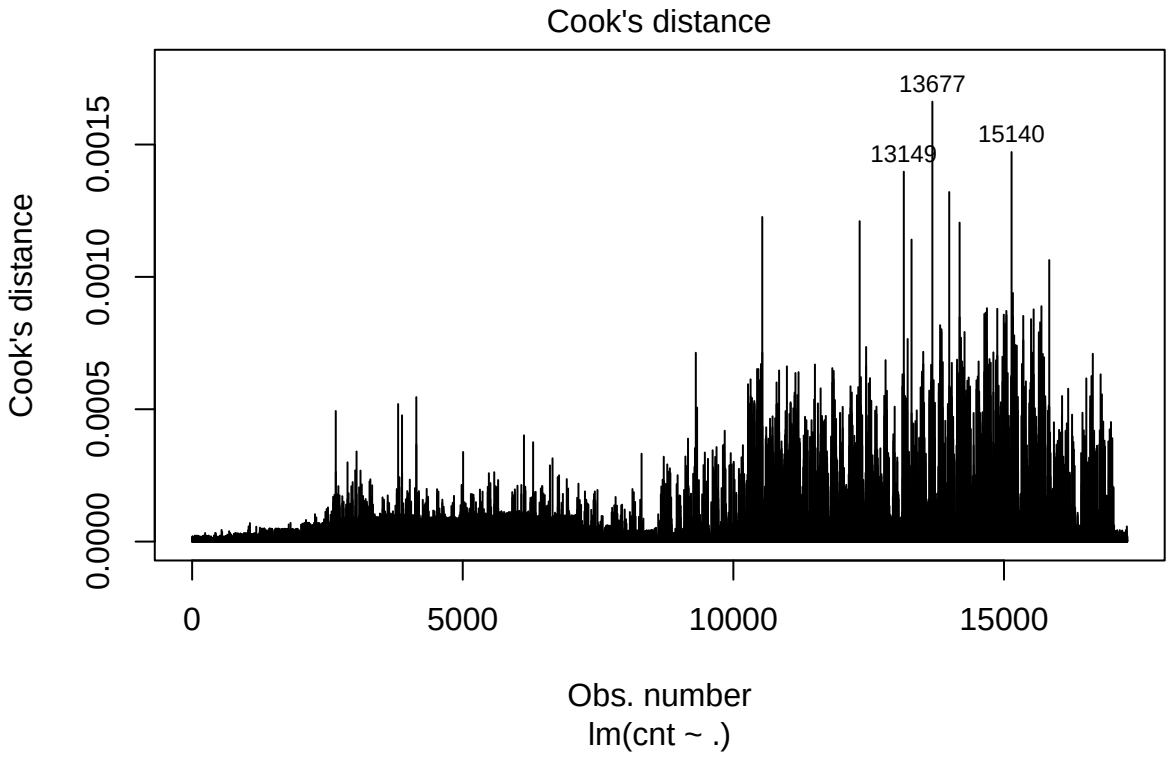
The plot of hat values indicates that leverage is evenly distributed across observations, with no points exceeding the typical high-leverage threshold. This pattern is expected since the model only includes categorical predictors, which group observations into repeated and similar patterns. No single data point exerts disprportionately high influence on its own fitted value, suggesting that the dataset does not contain influential outliers in terms of leverage.

```
stud_resid <- abs(rstudent(model_dis))
plot(stud_resid,
     main = "Studentized Residuals")
abline(h = qt(0.95, df = nrow(hours_dis) - ncol(hours_dis) - 2),
       col = "red")
```



The plot of studentized residuals against observation index shows that residual variance increases as the index increases. Given that the data are sorted chronologically, this suggests that the model performs less consistently in the later part of the dataset, potentially indicating more variation in bike rental counts during the second year (2012).

```
# influential points with cook's distance
plot(model_dis, which = 4)
```

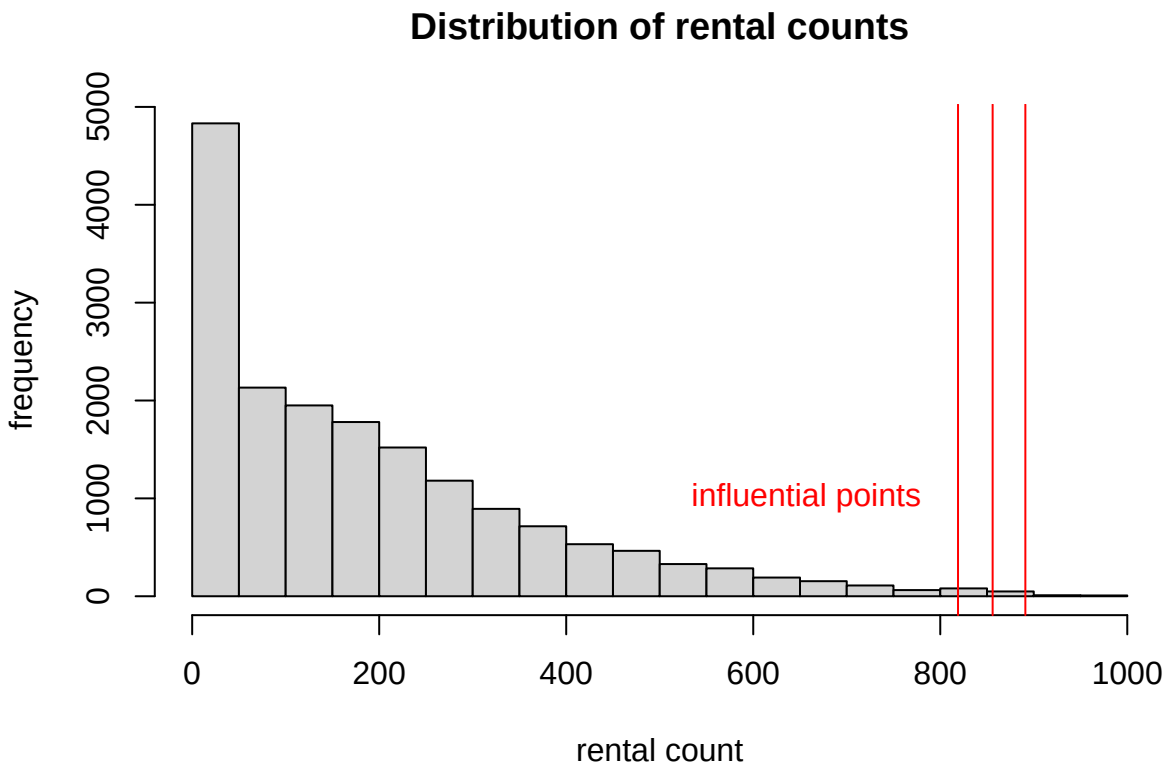



```
# most influential points
hours_dis[c(13149, 13677, 15140),]
```

```
##       workingday mnth_2 mnth_3 mnth_4 mnth_5 mnth_6 mnth_7 mnth_8 mnth_9
## 13149          1      0      0      0      0      0      1      0      0
## 13677          1      0      0      0      0      0      0      1      0
## 15140          1      0      0      0      0      0      0      0      0
##       mnth_10 mnth_11 mnth_12 weathersit_2 weathersit_3 cnt
## 13149       0       0       0            0            1 819
## 13677       0       0       0            0            1 891
## 15140       1       0       0            0            1 856
```

```
hist(hours_dis$cnt,
     main="Distribution of rental counts",
     xlab="rental count",
     ylab="frequency")
abline(v=819, col="red")
abline(v=891, col="red")
abline(v=856, col="red")

text(x = 800,
     y = 1000,
     labels = "influential points",
     pos = 2,
     col = "red")
```



The Cook's Distance plot reveals that indices 13149, 13677, and 15140 have the highest influence on the regression. Since previous diagnostics of hat values did not reveal any unusually high leverage predictor combinations, it is likely that these influential points arise from extremely high response values, i.e. rental counts that are much higher than typical observations. This is supported by the histogram of rental count distribution shown above (the influential points are marked in red). These points may disproportionately affect the model fit and parameter estimates, so it may be worth examining them more closely or considering transformations.

2.

Fit a linear model combining continuous and discrete predictors, including diagnostics. Compare this model to the previous models using only continuous or only discrete variables

```
hours = hours_clean[, c("yr", "mnth", "workingday", "weathersit", "atemp", "hum", "windspeed", "cnt")]
```

```
# one-hot-encode and standardize
library(fastDummies)
library(dplyr)

# these are outliers that affect the scaling
hours$weathersit[hours$weathersit == 4] = 3

hours = dummy_cols(hours,
                   select_columns = c("mnth", "weathersit"),
                   remove_first_dummy = TRUE,
                   remove_selected_columns = TRUE)
hours = dplyr::select(hours, -cnt, cnt)
head(hours)
```

```
##   yr workingday  atemp  hum windspeed mnth_2 mnth_3 mnth_4 mnth_5 mnth_6 mnth_7
## 1  0          0 0.2879 0.81    0.0000      0      0      0      0      0      0
```

```
## 2  0          0 0.2727 0.80    0.0000      0      0      0      0      0      0
## 3  0          0 0.2727 0.80    0.0000      0      0      0      0      0      0
## 4  0          0 0.2879 0.75    0.0000      0      0      0      0      0      0
## 5  0          0 0.2879 0.75    0.0000      0      0      0      0      0      0
## 6  0          0 0.2576 0.75    0.0896      0      0      0      0      0      0
##   mnth_8 mnth_9 mnth_10 mnth_11 mnth_12 weathersit_2 weathersit_3 cnt
## 1      0      0       0       0       0            0            0  16
## 2      0      0       0       0       0            0            0  40
## 3      0      0       0       0       0            0            0  32
## 4      0      0       0       0       0            0            0  13
## 5      0      0       0       0       0            0            0   1
## 6      0      0       0       0       0            1            0   1
```

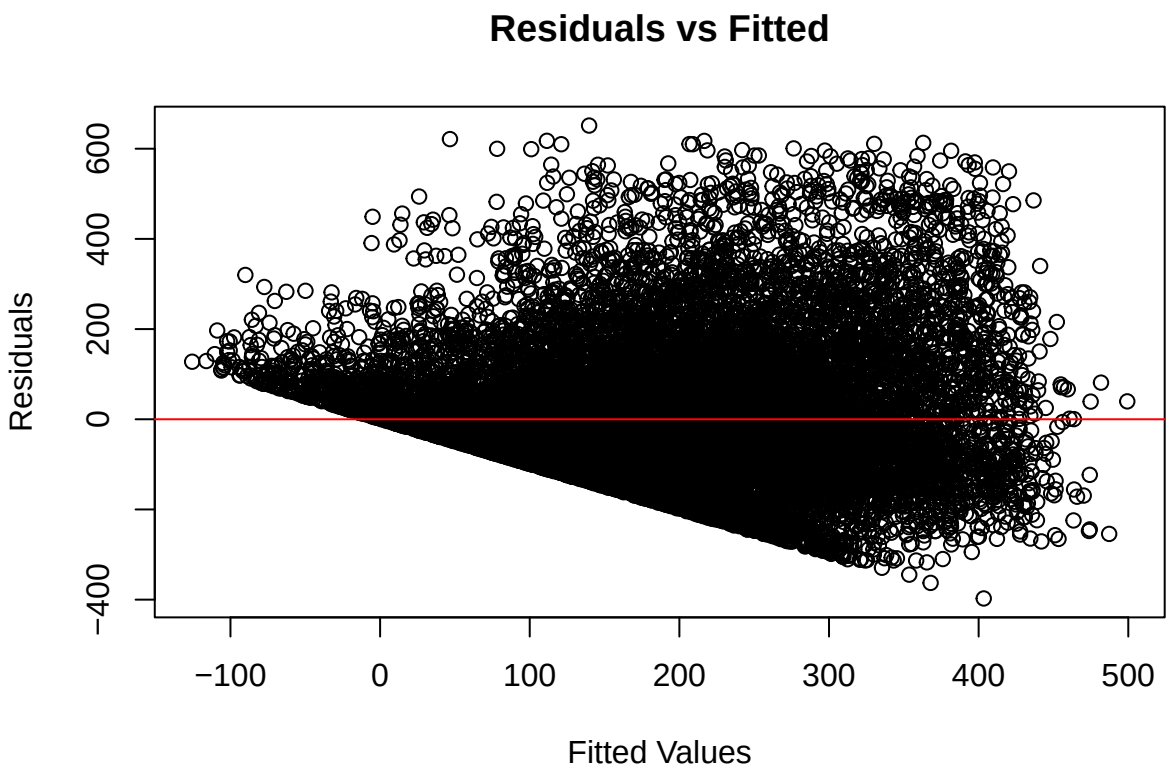
```
# fit linear model
model = lm(cnt ~ ., data=hours)
summary(model)
```

```
##
## Call:
## lm(formula = cnt ~ ., data = hours)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397.40  -99.38  -25.49   69.43  651.31
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    68.351      7.579   9.019  < 2e-16 ***
## yr             75.015      2.277  32.951  < 2e-16 ***
## workingday      4.610      2.444   1.887  0.05924 .
## atemp         543.034     12.878  42.168  < 2e-16 ***
## hum          -281.919      7.177 -39.281  < 2e-16 ***
## windspeed      57.875      9.955   5.813 6.23e-09 ***
## mnth_2        -17.052      5.759  -2.961  0.00307 **
## mnth_3        -12.044      5.852  -2.058  0.03961 *
## mnth_4        -20.593      6.258  -3.291  0.00100 **
## mnth_5        -12.788      7.069  -1.809  0.07049 .
## mnth_6        -65.889      7.589  -8.682  < 2e-16 ***
## mnth_7       -102.514      8.193 -12.512  < 2e-16 ***
## mnth_8        -57.434      7.773  -7.389 1.55e-13 ***
## mnth_9          5.655      7.253   0.780  0.43564
## mnth_10        38.356      6.461   5.937 2.96e-09 ***
## mnth_11        31.435      5.851   5.373 7.87e-08 ***
## mnth_12        29.561      5.703   5.184 2.20e-07 ***
## weathersit_2   12.525      2.767   4.526 6.04e-06 ***
## weathersit_3   -2.816      4.621  -0.609  0.54228
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 148.6 on 17259 degrees of freedom
## Multiple R-squared:  0.3314, Adjusted R-squared:  0.3307
## F-statistic: 475.3 on 18 and 17259 DF,  p-value: < 2.2e-16
```

This regression model, which incorporates both continuous and categorical predictors (as well as the year variable that we decided to add), explains approximately 33% of the variability in hourly bike rental counts.

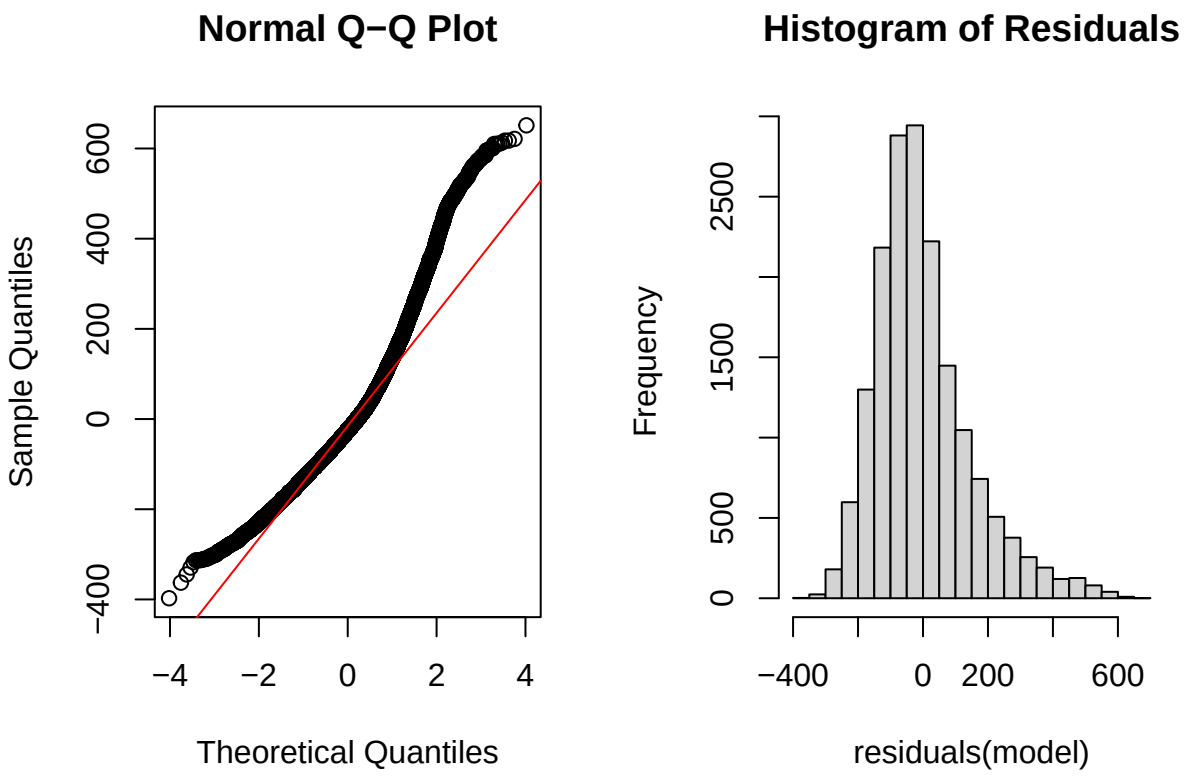
Among the most influential predictors are atemp (positively associated) and hum (negatively associated), indicating weather conditions strongly affect ridership. The year indicator suggests usage increased in 2012 compared to 2011. Interestingly, some summer months (like July and August) show negative associations with rental count, which may reflect unaccounted interactions or overlapping seasonal effects. The weather situation variables showed mixed results, with misty conditions unexpectedly increasing rentals and heavier weather showing no significant impact. Overall, the model performs better than the versions with only categorical variables or only continuous variables, but still leaves considerable variance unexplained, hinting at the need for additional predictors or a different type of model.

```
# check for homoscedasticity
par(mfrow = c(1, 1))
plot(model$fitted.values, residuals(model),
     main = "Residuals vs Fitted",
     xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red")
```



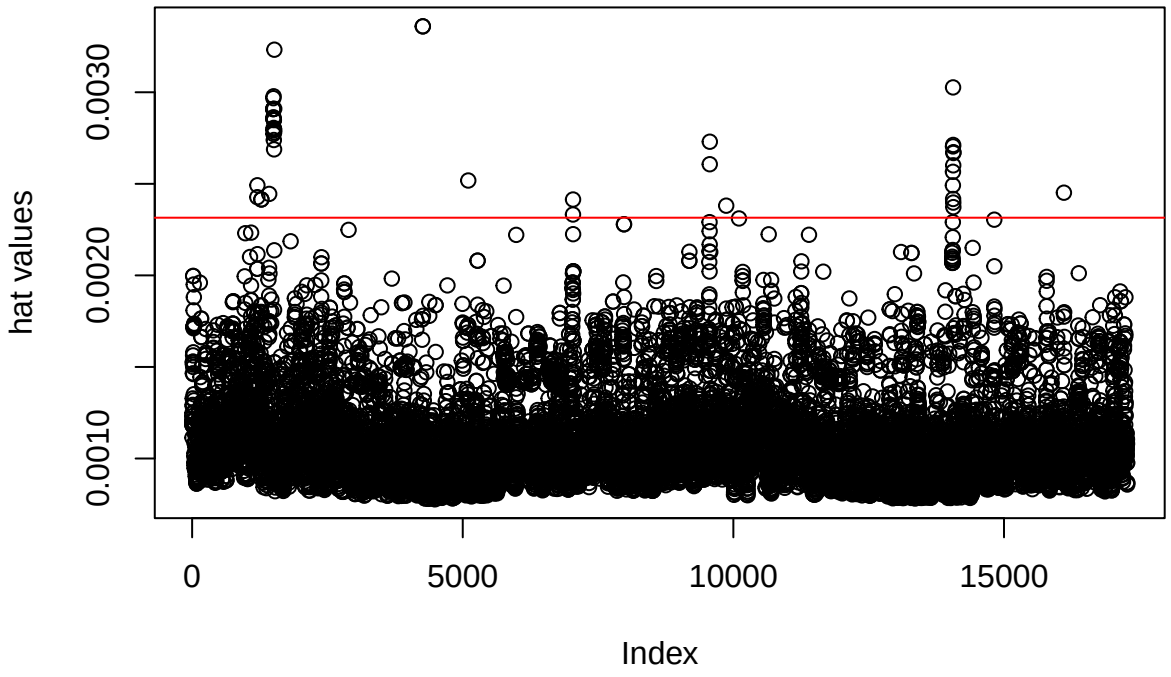
The clear linear boundary observed in the Residuals vs. Fitted plot is likely due to the discrete, non-negative nature of the response variable (cnt). Since bike rentals are count data, applying linear regression, which assumes a continuous response, can produce predictions that do not align with the structure of the data. The plot also shows a distinct funnel shape, where the spread of residuals increases with higher fitted values (even when ignoring the cutoff). This suggests that the assumption of homoscedasticity is violated. The presence of heteroscedasticity indicates that the linear model may not fully capture the relationship between predictors and the response, especially at higher bike rental counts. Further model refinement or transformation of the response variable may be needed to address this issue.

```
par(mfrow = c(1, 2))
qqnorm(residuals(model));
qqline(residuals(model), col = "red")
hist(residuals(model),
     main = "Histogram of Residuals",
     breaks = 30)
```

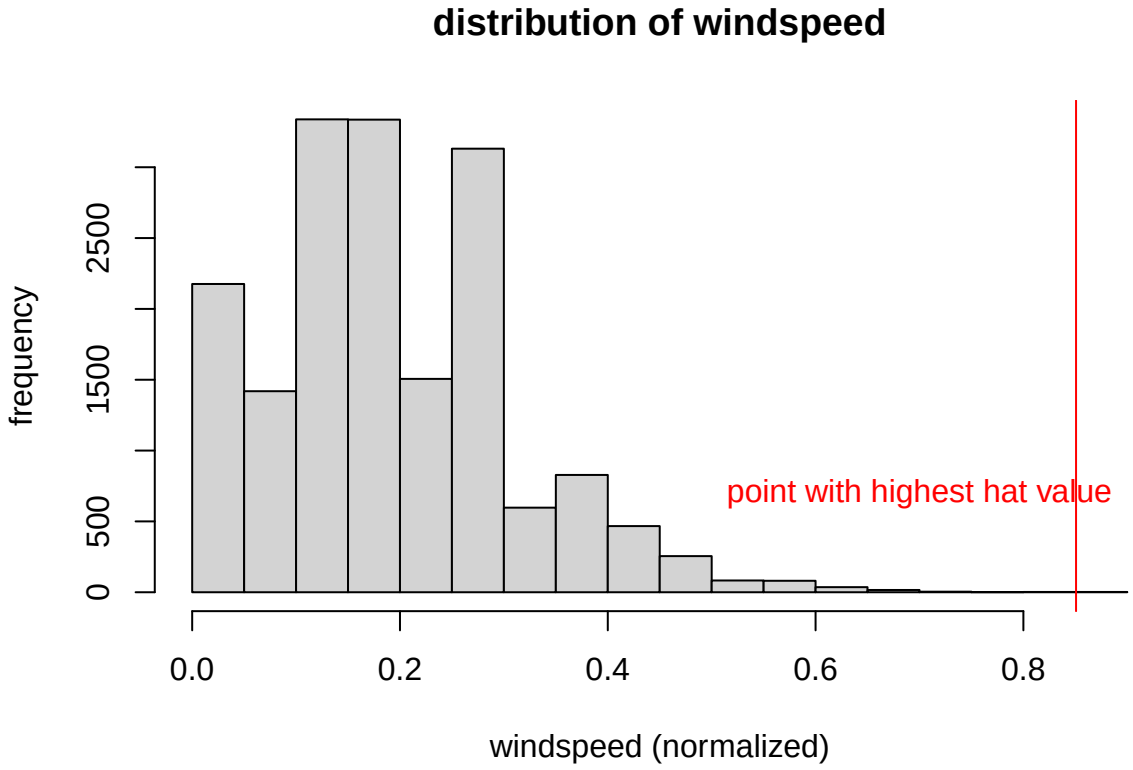


The Q-Q plot shows some deviation from the diagonal, especially in the upper tail and a bit in the lower tail, indicating that the residuals are not perfectly normally distributed. The histogram of residuals show a right skew. These results suggest that the normality assumption may be violated, particularly at the extremes.

```
hat_values <- hatvalues(model)
plot(hat_values,
     ylab = "hat values",
     main = "Hat Values")
abline(h = 2 * (ncol(hours) + 1) / nrow(hours),
       col = "red")
```

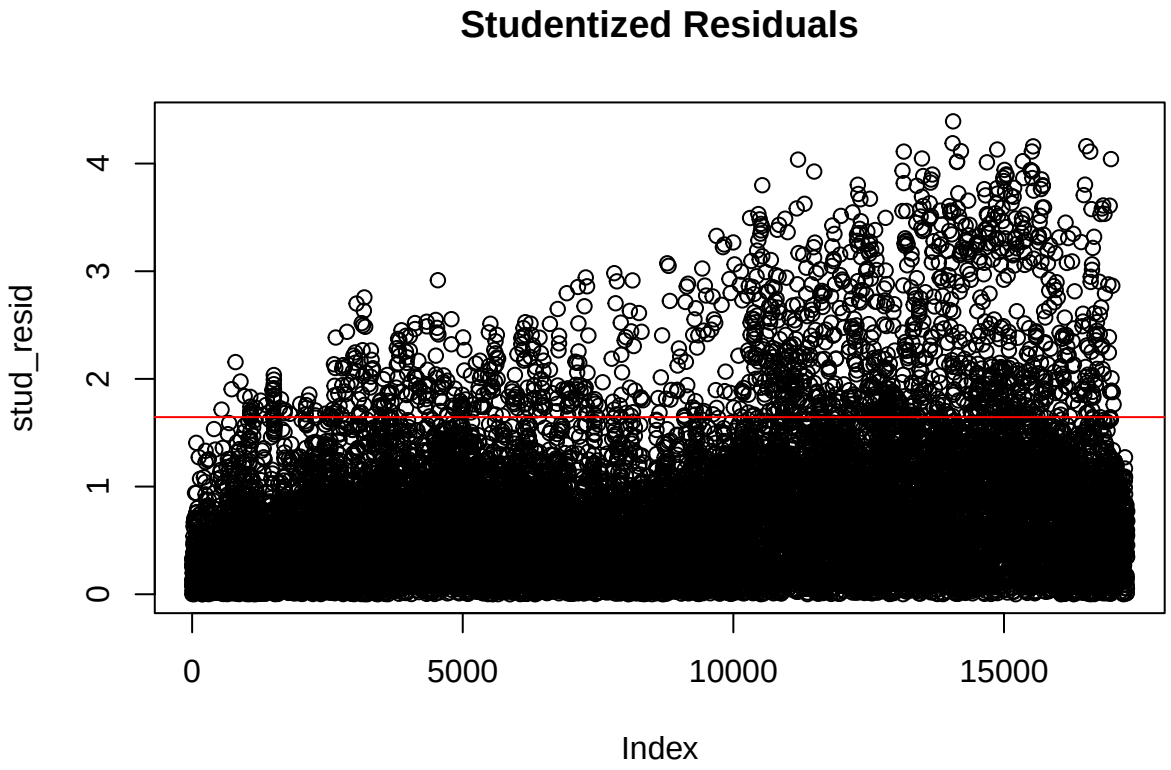



```
# windspeed of the hour with the highest hat value
hist(hours$windspeed,
     xlab="windspeed (normalized)",
     ylab="frequency",
     main="distribution of windspeed")
abline(v=hours[which.max(hat_values), ]$windspeed,
       col='red')
text(x = 0.7,
     y = 500,
     labels = "point with highest hat value",
     pos = 3,
     col = "red")
```



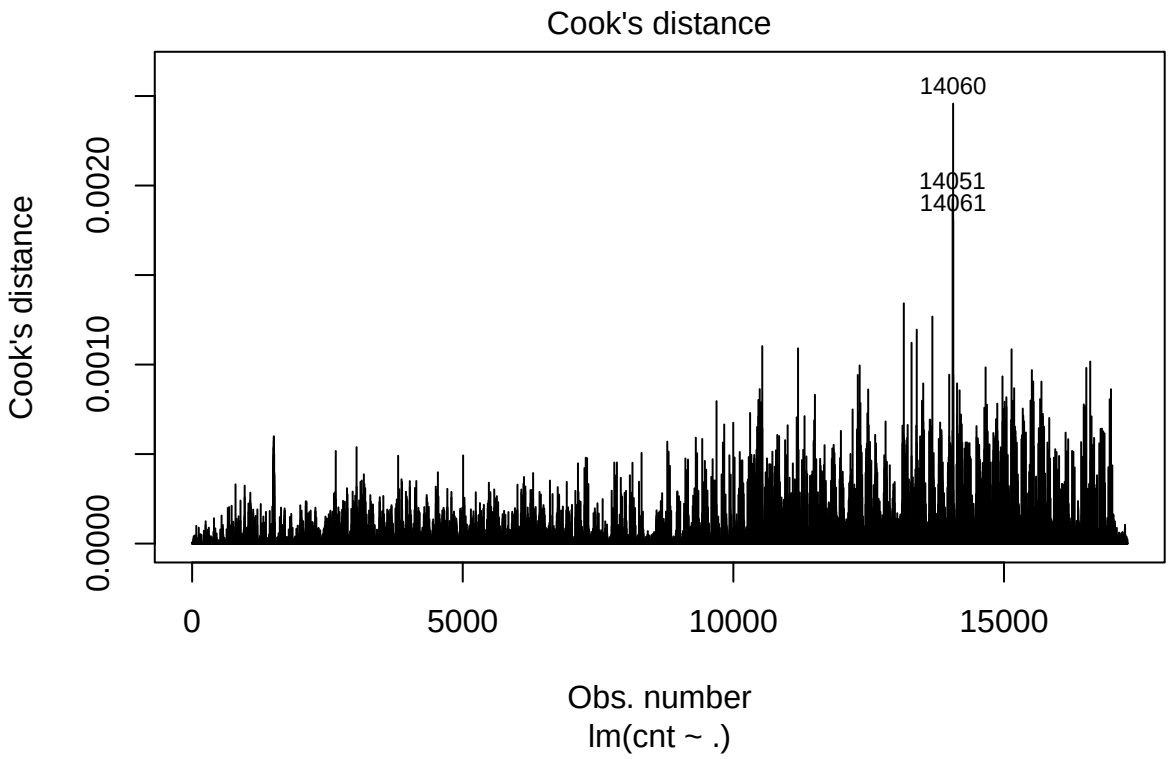
The hat values plot shows that most observations have low leverage, clustered below the cutoff threshold. However, a few points exhibit relatively high leverage, suggesting that these observations have unusual predictor values. For example, the point with the highest hat value shown above has a normalized windspeed of 5.4124, which is extremely high compared to the rest of the observations.

```
stud_resid <- abs(rstudent(model))
plot(stud_resid,
     main = "Studentized Residuals")
abline(h = qt(0.95, df = nrow(hours) - ncol(hours) - 2),
       col = "red")
```



The plot of studentized residuals against observation index reveals that residual variance increases over time. As mentioned previously given that the data are sorted chronologically, this suggests that the model performs less consistently in the later part of the dataset, potentially indicating more variation in bike rental counts during the second year (2012).

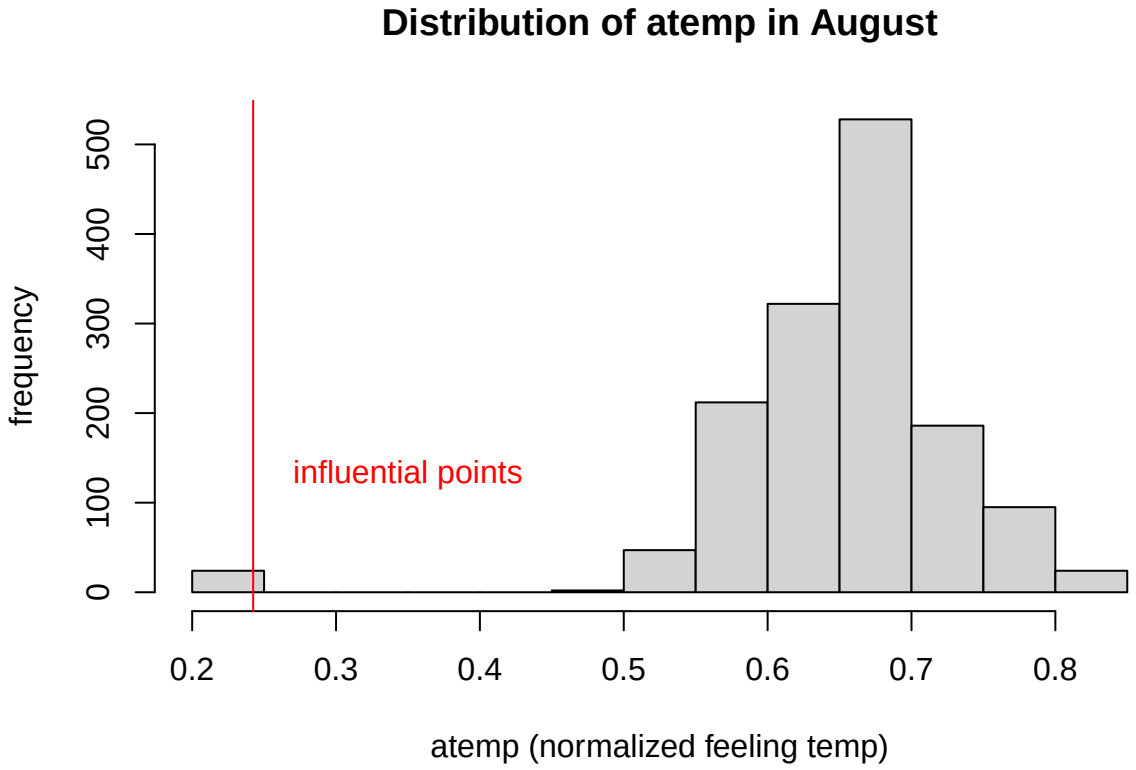
```
# influential points with cook's distance
plot(model, which = 4)
```




```
# most influential points
hours[c(14051, 14060, 14061),]
```

```
##       yr workingday  atemp  hum windspeed mnth_2 mnth_3 mnth_4 mnth_5 mnth_6
## 14051  1          1 0.2424 0.65    0.1343      0      0      0      0      0
## 14060  1          1 0.2424 0.36    0.3284      0      0      0      0      0
## 14061  1          1 0.2424 0.38    0.2537      0      0      0      0      0
##       mnth_7 mnth_8 mnth_9 mnth_10 mnth_11 mnth_12 weathersit_2 weathersit_3
## 14051      0      1      0       0       0       0            0            0
## 14060      0      1      0       0       0       0            0            0
## 14061      0      1      0       0       0       0            1            0
##       cnt
## 14051 668
## 14060 791
## 14061 669
```

```
hist(hours[hours$mnth_8 > 0,]$atemp,
     main="Distribution of atemp in August",
     xlab="atemp (normalized feeling temp)",
     ylab="frequency")
abline(v=0.2424, col="red")
text(x = 0.35,
     y = 100,
     labels = "influential points",
     pos = 3,
     col = "red")
```



The Cook's Distance plot reveals that indices 14051, 14060, and 14061 have the highest influence on the regression. These observations all fall on the same day (2012-08-17). Upon closer inspection, we believe this is possibly a data entry error since all three points happen to have the same exact normalized atemp value of 0.2424, which is extremely unusual in August as shown in the histogram above. In addition, cross-referencing the original data with the actual temperature, we found that the actual normalized temperature on the same day at 9AM was 0.74, which seems consistent with the trend in August. We believe by replacing these atemp values with the actual temperature, we may reduce some error in our model.

```
round(cor(hours[, -ncol(hours)]), 2)
```

```
##                 yr workingday atemp   hum windspeed mnth_2 mnth_3 mnth_4 mnth_5
## yr            1.00       0.00  0.04 -0.08     -0.01   0.01   0.00   0.00   0.00
## workingday    0.00       1.00  0.06  0.02     -0.01   0.00   0.03  -0.01   0.01
## atemp         0.04       0.06  1.00 -0.05     -0.06  -0.29  -0.17  -0.03   0.16
## hum          -0.08       0.02 -0.05  1.00     -0.30  -0.09  -0.06  -0.06   0.10
## windspeed    -0.01      -0.01 -0.06 -0.30      1.00   0.06   0.08   0.11  -0.02
## mnth_2        0.01       0.00 -0.29 -0.09      0.06   1.00  -0.09  -0.09  -0.09
## mnth_3        0.00       0.03 -0.17 -0.06      0.08  -0.09   1.00  -0.09  -0.09
## mnth_4        0.00      -0.01 -0.03 -0.06      0.11  -0.09  -0.09   1.00  -0.09
## mnth_5        0.00       0.01  0.16  0.10     -0.02  -0.09  -0.09  -0.09   1.00
## mnth_6        0.00       0.02  0.28 -0.08     -0.01  -0.09  -0.09  -0.09  -0.09
## mnth_7        0.00      -0.01  0.41 -0.05     -0.06  -0.09  -0.09  -0.09  -0.09
## mnth_8        0.01       0.05  0.31  0.01     -0.06  -0.09  -0.09  -0.09  -0.09
## mnth_9        0.00      -0.01  0.18  0.14     -0.06  -0.09  -0.09  -0.09  -0.09
## mnth_10      -0.01      -0.01  0.00  0.10     -0.05  -0.09  -0.09  -0.09  -0.09
## mnth_11       0.00      -0.01 -0.19  0.00     -0.01  -0.09  -0.09  -0.09  -0.09
## mnth_12       0.00      -0.01 -0.27  0.06     -0.03  -0.09  -0.09  -0.09  -0.09
## weathersit_2  0.01       0.02 -0.07  0.22     -0.05   0.00   0.03   0.00   0.01
## weathersit_3 -0.03       0.03 -0.07  0.31      0.06   0.02   0.01   0.02   0.02
##              mnth_6 mnth_7 mnth_8 mnth_9 mnth_10 mnth_11 mnth_12 weathersit_2
```

```
## yr             0.00   0.00   0.01   0.00   -0.01    0.00    0.00         0.01
## workingday     0.02  -0.01   0.05  -0.01   -0.01   -0.01   -0.01         0.02
## atemp          0.28   0.41   0.31   0.18    0.00   -0.19   -0.27        -0.07
## hum           -0.08  -0.05   0.01   0.14    0.10    0.00    0.06         0.22
## windspeed     -0.01  -0.06  -0.06  -0.06   -0.05   -0.01   -0.03        -0.05
## mnth_2        -0.09  -0.09  -0.09  -0.09   -0.09   -0.09   -0.09         0.00
## mnth_3        -0.09  -0.09  -0.09  -0.09   -0.09   -0.09   -0.09         0.03
## mnth_4        -0.09  -0.09  -0.09  -0.09   -0.09   -0.09   -0.09         0.00
## mnth_5        -0.09  -0.09  -0.09  -0.09   -0.09   -0.09   -0.09         0.01
## mnth_6         1.00  -0.09  -0.09  -0.09   -0.09   -0.09   -0.09        -0.05
## mnth_7        -0.09   1.00  -0.09  -0.09   -0.09   -0.09   -0.09        -0.06
## mnth_8        -0.09  -0.09   1.00  -0.09   -0.09   -0.09   -0.09        -0.04
## mnth_9        -0.09  -0.09  -0.09   1.00   -0.09   -0.09   -0.09         0.02
## mnth_10       -0.09  -0.09  -0.09  -0.09    1.00   -0.09   -0.09         0.02
## mnth_11       -0.09  -0.09  -0.09  -0.09   -0.09    1.00   -0.09         0.00
## mnth_12       -0.09  -0.09  -0.09  -0.09   -0.09   -0.09    1.00         0.06
## weathersit_2  -0.05  -0.06  -0.04   0.02    0.02    0.00    0.06         1.00
## weathersit_3  -0.03  -0.04  -0.03   0.02    0.03   -0.01    0.01        -0.18
##               weathersit_3
## yr                   -0.03
## workingday            0.03
## atemp                -0.07
## hum                   0.31
## windspeed             0.06
## mnth_2                0.02
## mnth_3                0.01
## mnth_4                0.02
## mnth_5                0.02
## mnth_6               -0.03
## mnth_7               -0.04
## mnth_8               -0.03
## mnth_9                0.02
## mnth_10               0.03
## mnth_11              -0.01
## mnth_12               0.01
## weathersit_2         -0.18
## weathersit_3          1.00
```

We do not see any multicollinearity in the predictor variables. The pairwise correlations shown above never exceed 0.41, which indicates very low linear correlations between the variables.

Comparison

- Model 1: Continuous-only predictors (atemp, hum, windspeed)
  - Adjusted $R^2 = 0.2518$
  - atemp and hum are strong predictors; windspeed has a weaker but significant effect.
  - Residuals show heteroscedasticity (funnel shape) and right-skewed distribution, violating key assumptions.
  - Some high-leverage points driven by extreme predictor values (high windspeed).
  - Influential points linked to both extreme windspeed and response values.
- Model 2: Discrete-only predictors (mnth, workingday, weathersit)
  - Adjusted $R^2 = 0.0917$
  - Summer months (July, August) associated with higher rentals; worse weather reduces rentals.
  - Homoscedasticity assumption violated; residuals still skewed and non-normal.
  - Low leverage across all points due to repeated patterns in categorical variables.

  - Influential points arise from extreme rental counts, not predictor outliers.
- Model 3: Combined model (all above + yr)
  - Adjusted $R^2 = 0.3307$
  - Combines strengths of both weather and temporal predictors; yr captures growth from 2011 to 2012.
  - Residual plots still reveal heteroscedasticity and non-normality, especially at higher fitted values.
  - Some high-leverage points tied to unusual inputs (abnormally low atemp in August).
  - Influential points reflect both data entry issues and extreme responses.

Overall, the combined model outperforms the continuous-only and discrete-only models in explaining variation in bike rental counts. However, all models show violations of key linear regression assumptions, particularly heteroscedasticity and non-normal residuals, suggesting that further improvements could be achieved through data transformation, outlier handling, or by using more flexible modeling approaches.

Contribution Statement

Members: Ashley Ho & Mizuho Fukuda

1. We discussed ideas for which discrete variables to include in our data together. We also observed the outputs together and discussed how to interpret each value. The code and interpretations were written by Mizuho.
2. We discussed how to implement and interpret each part together. We talked about the differences in the models and drew the conclusion together. The coding and writing was done by Ashley.