Baseline Linear Model

1.

Fit a linear model using the response and the continuous predictor variables identified previously in your dataset. Report and interpret the estimated coefficients and their associated standard errors and p-values. Report and interpret the adjusted R squared of the model.

```r
# loda data
hours = read.csv('data/hour_clean.csv')
hours = hours[, c("mnth", "workingday", "weathersit", "atemp", "hum", "windspeed", "cnt")]
```

```r
# 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
```

```r
# 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)
```

```
##   workingday  atemp  hum windspeed mnth_2 mnth_3 mnth_4 mnth_5 mnth_6 mnth_7
## 1          0 0.2879 0.81    0.0000      0      0      0      0      0      0
## 2          0 0.2727 0.80    0.0000      0      0      0      0      0      0
## 3          0 0.2727 0.80    0.0000      0      0      0      0      0      0
## 4          0 0.2879 0.75    0.0000      0      0      0      0      0      0
## 5          0 0.2879 0.75    0.0000      0      0      0      0      0      0
## 6          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
```

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

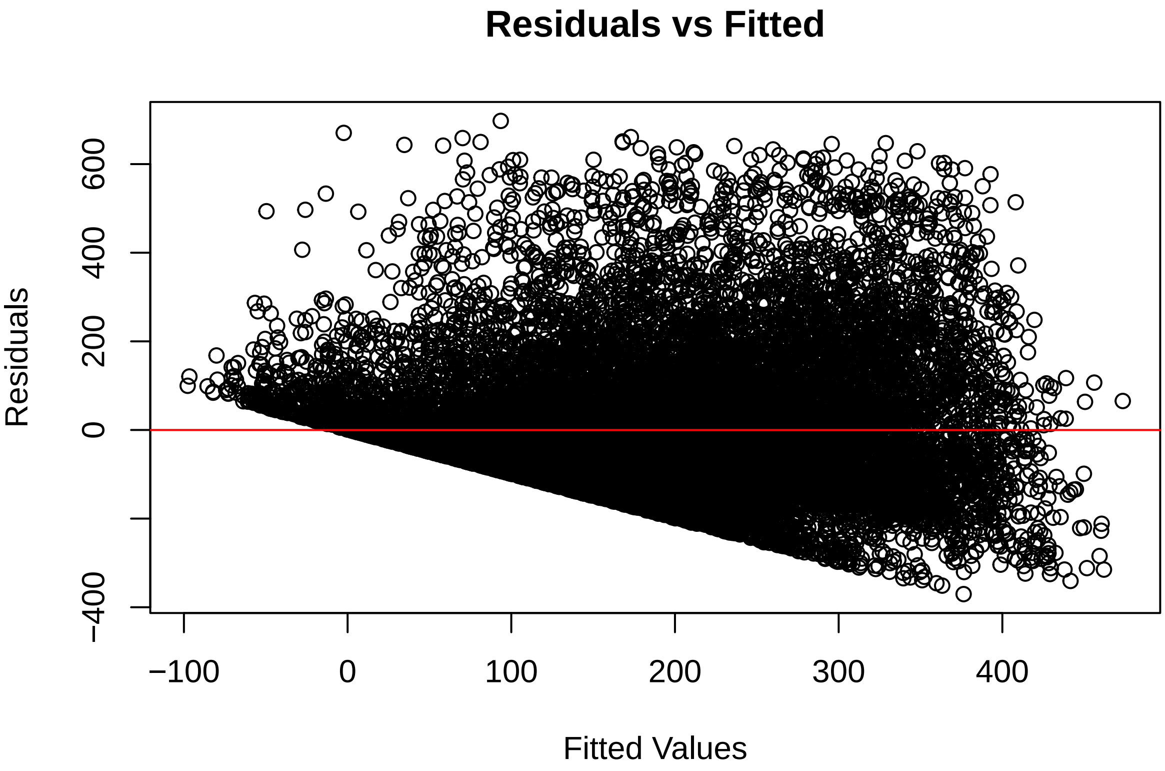
```
## 
## Call:
## lm(formula = cnt ~ ., data = hours)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -370.36 -102.15  -31.39   63.79  697.44 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   112.833      7.688  14.676  < 2e-16 ***
## workingday      3.987      2.519   1.583  0.11352    
## atemp         572.748     13.244  43.246  < 2e-16 ***
## hum          -299.273      7.379 -40.557  < 2e-16 ***
## windspeed      47.057     10.258   4.587 4.52e-06 ***
## mnth_2        -19.475      5.937  -3.280  0.00104 ** 
## mnth_3        -17.856      6.031  -2.961  0.00307 ** 
## mnth_4        -28.823      6.447  -4.471 7.84e-06 ***
## mnth_5        -23.028      7.281  -3.163  0.00157 ** 
## mnth_6        -79.943      7.811 -10.234  < 2e-16 ***
## mnth_7       -118.322      8.432 -14.032  < 2e-16 ***
## mnth_8        -69.810      8.004  -8.722  < 2e-16 ***
## mnth_9         -4.728      7.471  -0.633  0.52685    
## mnth_10        29.371      6.654   4.414 1.02e-05 ***
## mnth_11        25.995      6.030   4.311 1.63e-05 ***
## mnth_12        25.767      5.878   4.384 1.17e-05 ***
## weathersit_2   15.098      2.852   5.295 1.21e-07 ***
## weathersit_3   -1.430      4.764  -0.300  0.76402    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 153.2 on 17260 degrees of freedom
## Multiple R-squared:  0.2893, Adjusted R-squared:  0.2886 
## F-statistic: 413.4 on 17 and 17260 DF,  p-value: < 2.2e-16
```

The linear regression model explains approximately 28.9% of the variation in bike rental counts (Adjusted $R^2 = 0.2886$). The temperature (atemp) and humidity (hum) were the strongest continuous predictors: higher temperatures significantly increased rentals, while higher humidity significantly decreased rentals. Wind speed also had a small but significant positive effect. Seasonal effects were evident, with summer months showing lower rental counts relative to January. Some weather condition categories showed significant differences, though not consistently across all levels.

2.

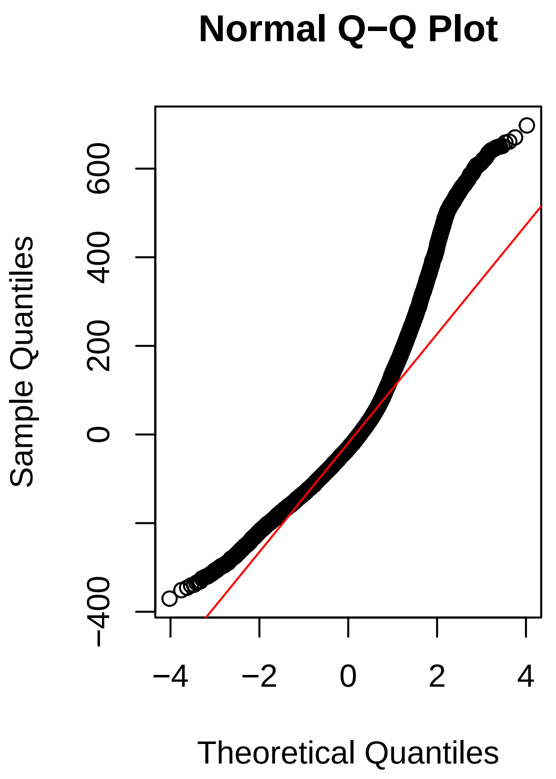
Use diagnostic methods to assess the validity of the standard assumptions in your linear model. Use graphics and offer brief comments on what you observe.

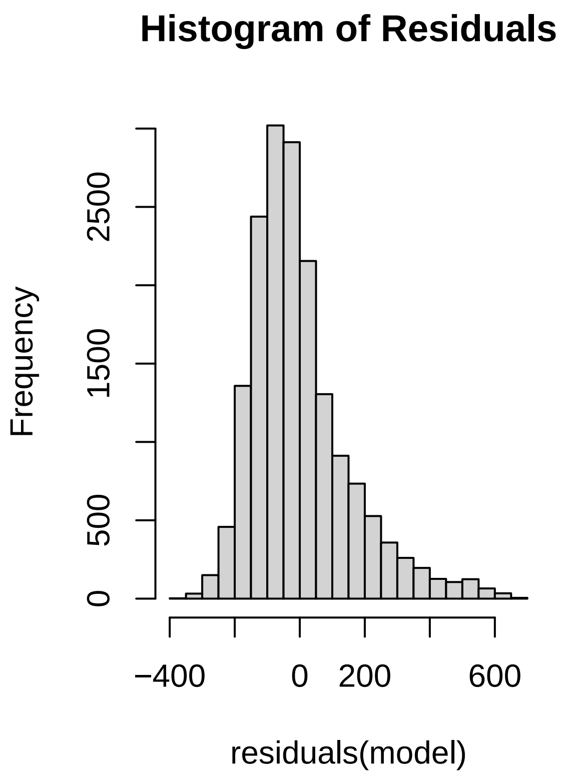
```r
# check for homoscedasticity
par(mfrow = c(1, 1))
plot(model$fitted.values, residuals(model),
     main = "Residuals vs Fitted",
     xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red")
```



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.

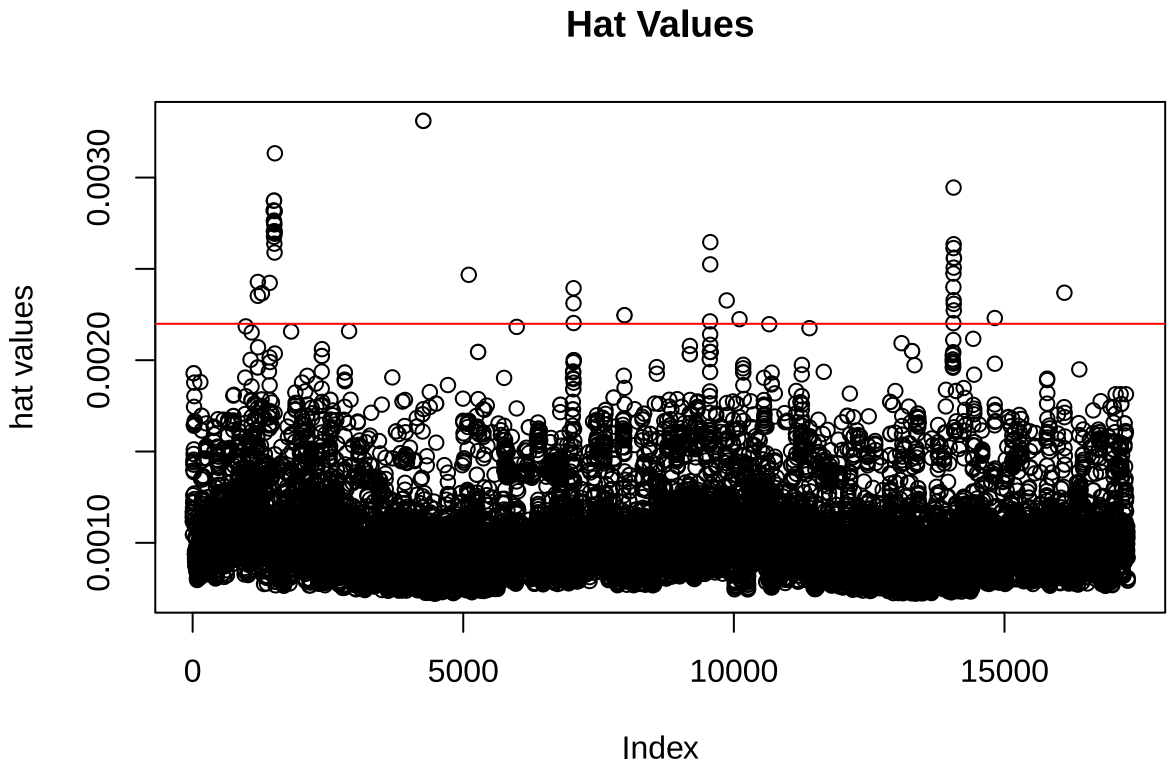
```r
par(mfrow = c(1, 2))
qqnorm(residuals(model));
qqline(residuals(model), col = "red")
hist(residuals(model),
     main = "Histogram of Residuals",
     breaks = 30)
```



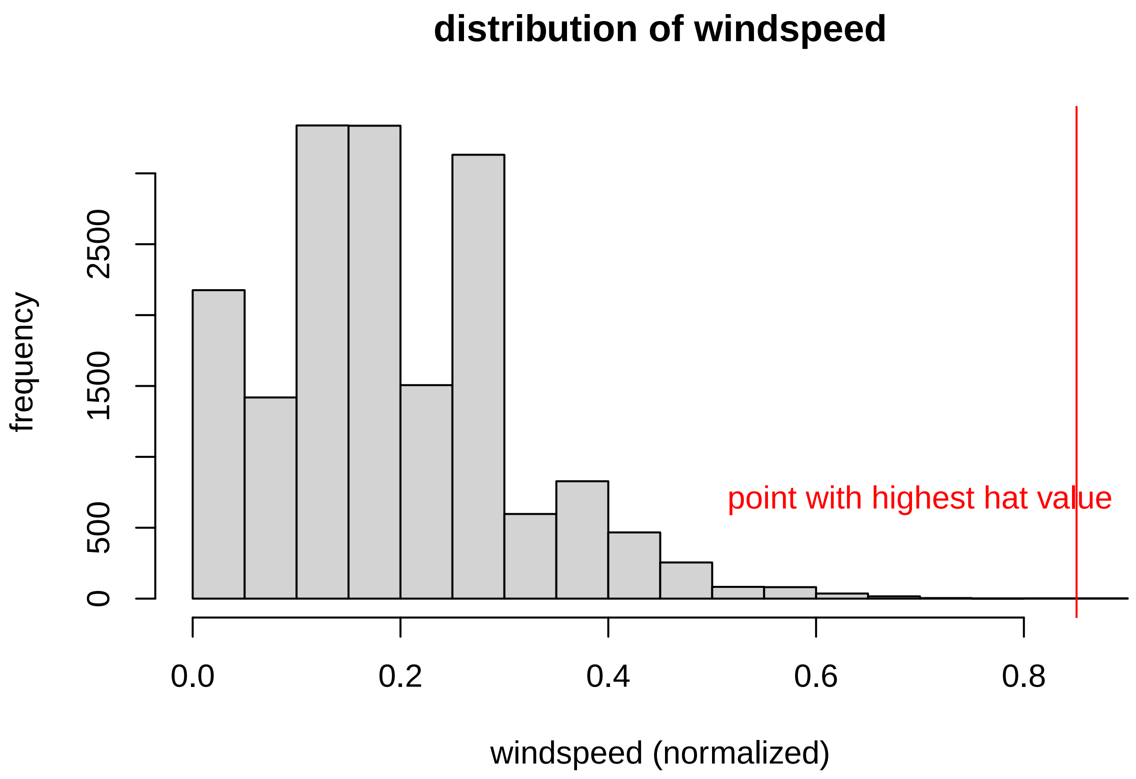


The Q-Q plot shows some deviation from the diagonal, especially in the upper tail and a bit in the lower tail, indicating that the residuals are not perfectly normally distributed. The histogram of residuals show a right skew. These results suggest that the normality assumption may be violated, particularly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

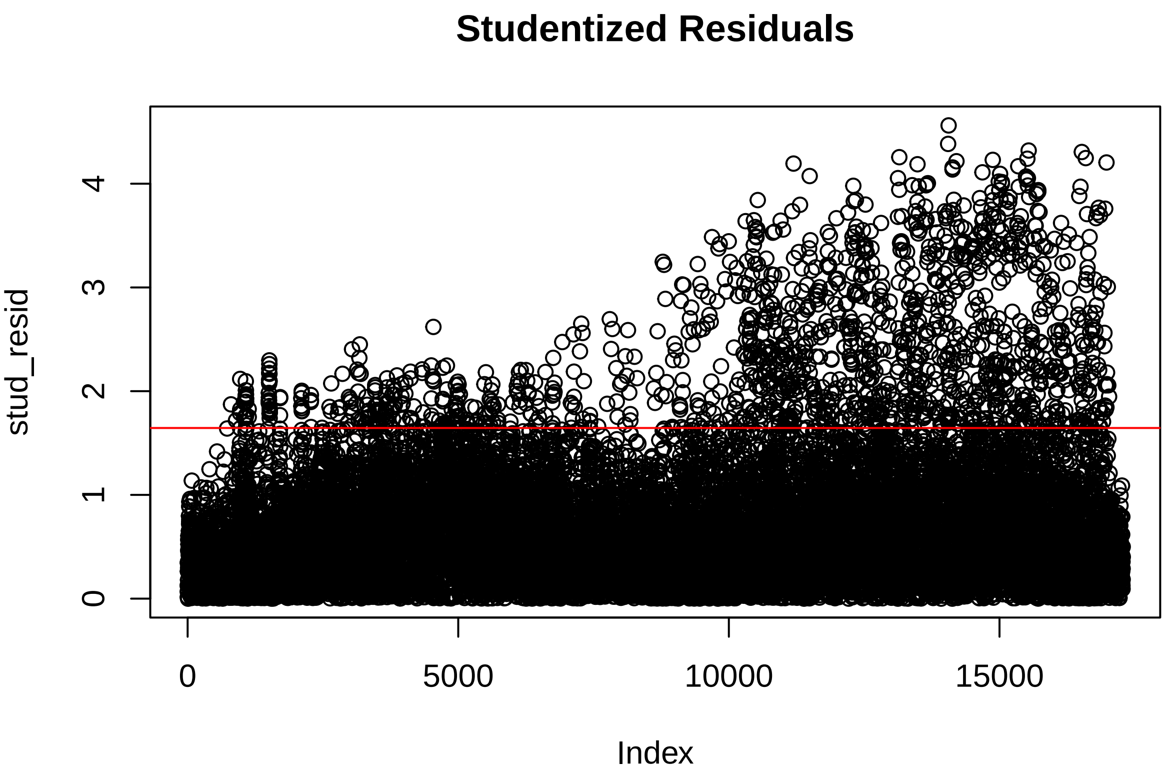


```
# windspeed of the hour with the highest hat value
hist(hours$windspeed,
     xlab="windspeed (normalized)",
     ylab="frequency",
     main="distribution of windspeed")
abline(v=hours[which.max(hat_values), ]$windspeed,
       col='red')
text(x = 0.7,
     y = 500,
     labels = "point with highest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



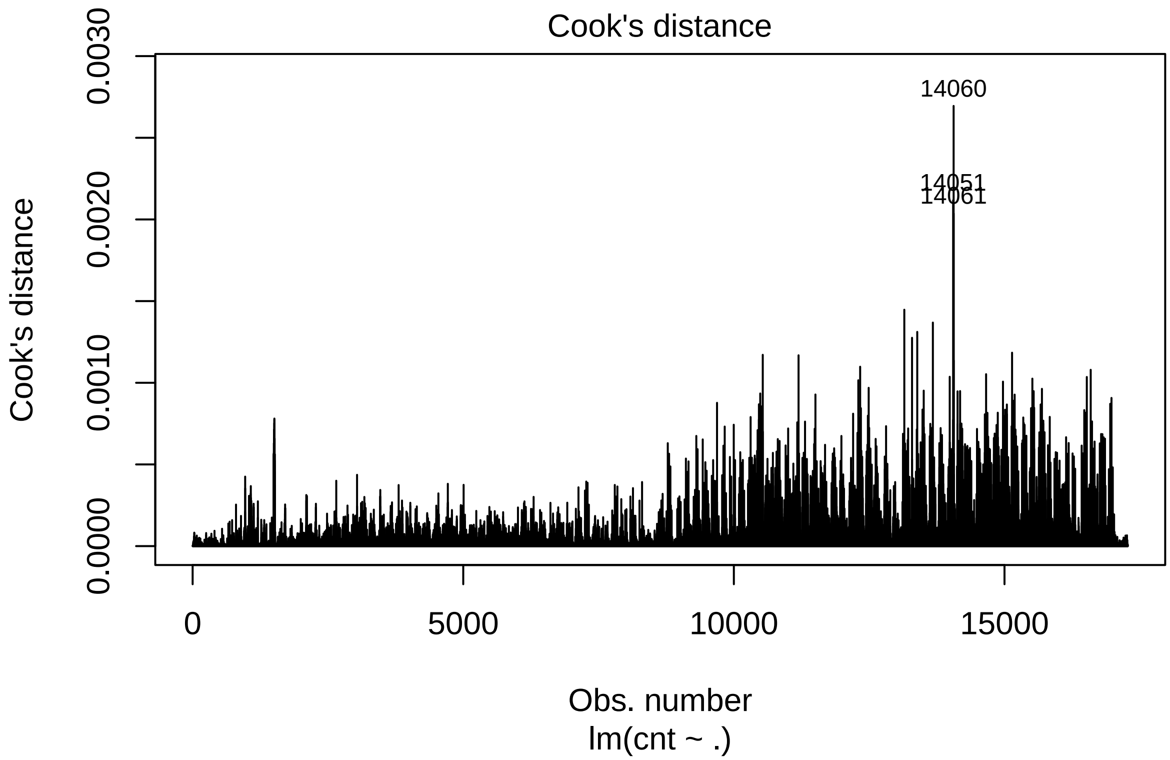
The hat values plot shows that most observations have low leverage, clustered below the cutoff threshold. However, a few points exhibit relatively high leverage, suggesting that these observations have unusual predictor values. For example, the point with the highest hat value shown above has a normalized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

```r
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. 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). This observation may reflect underlying time-dependent factors, specifically a year-dependent shift in rental pattern. Since the linear model did not use year as a feature (we only used month), including the year feature may significantly improve the model performance.

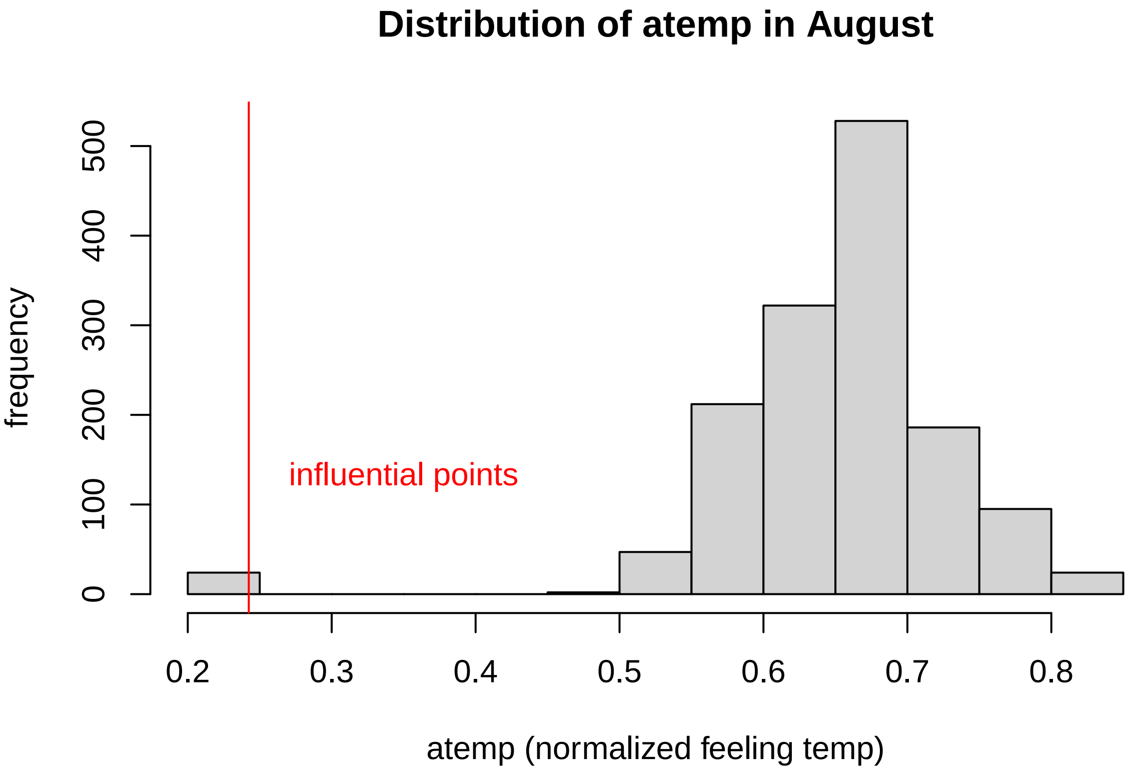
```
# influential points with cook's distance
plot(model, which = 4)
```



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

```
##       workingday  atemp  hum windspeed mnth_2 mnth_3 mnth_4 mnth_5 mnth_6
## 14051          1 0.2424 0.65    0.1343      0      0      0      0      0
## 14060          1 0.2424 0.36    0.3284      0      0      0      0      0
## 14061          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)
```

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

Contribution Statement

Members: Ashley Ho & Mizuho Fukuda

1. We discussed ideas for encoding and transforming the variables before fitting the linear regression. We also observed the outputs together and discussed how to interpret each value. The code and interpretations were written by Mizuho.
2. We discussed each part of the diagnostics and gave our interpretations. The coding and writing was done by both as well (taking turns).