EEMB 146 Lab Assignment 8

Samantha Chen

5/20/2021

Question 1

Provide some background: What is ozone? Why might some of these variables affect its levels? Where does this data come from? 3-4 sentences is fine. You are required to cite at least one source when writing this response.

Ozone is a highly reactive gas that when, it resides in the stratosphere, plays a critical role in absorbing solar ultraviolet radiation (Staehelin et. al, 2001). Although it is an important barrier from the Sun, high levels of ozone can become a health hazard to humans (Burrows, 2016). Factors like temperatures and solar radiation can dramatically affect ozone levels because it changes the chemical make-up of ozone and can cause ozone depletion (Allen, 2004). Much of this data of Ozone levels comes from organizations such as NASA and EPA (US Environmental Protection Agency).

Citations: Staehelin, J., Harris, N. R. P., Appenzeller, C., and Eberhard, J. (2001), Ozone trends: A review, Rev. Geophys., 39(2), 231– 290, doi:10.1029/1999RG000059.

Burrows, L. (2019, March 15). The complex relationship between heat and ozone. Harvard Gazette. https://news.harvard.edu/gazette/story/2016/04/the-complex-relationship-between-heat-and-ozone/

Environmental Protection Agency. (2020, June 23). What is Ozone? EPA. https://www.epa.gov/ozone-pollution-and-your-patients-health/what-ozone.

Allen, J. (n.d.). NASA GISS: Research Features: Ozone and Climate Change. NASA. https://www.giss.nasa.gov/research

Question 2

Read in the air quality data and visualize it. What are your potential predictors? What are some potential random factors? Do you think any of these predictors might interact with each other? Remove any missing data to 'clean up' the dataset.

```
# loading in airquality data from R database
airqual <- airquality
str(airqual) # 6 variables, 153 obsevations
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
```

```
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
# cleaning up data
airqual$Month <- as.factor(airqual$Month)
airqual$Day <- as.factor(airqual$Day)
airqual_sub <- na.omit(airqual)
```

Some potential predictors of ozone are solar radiation, wind, and temperature. Some potential random factors are month and day. I think solar radiation and temperature will interact with each other because temperature in a given area is influenced by the amount of sunlight present in that area.

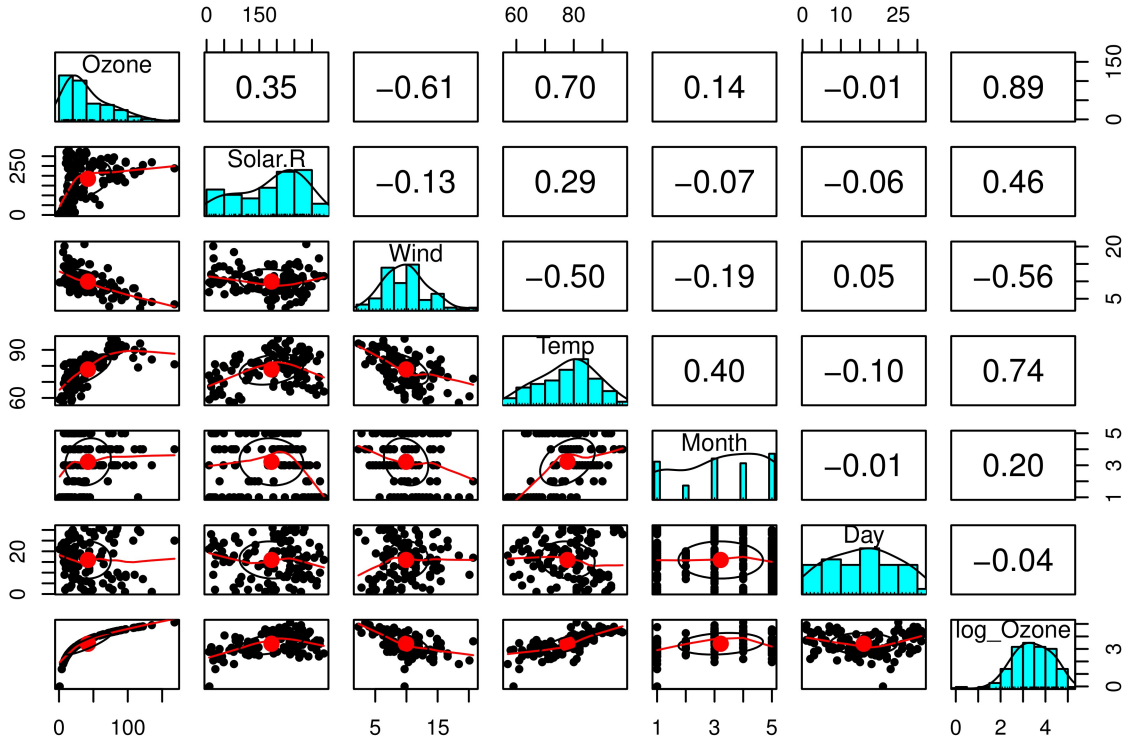
Question 3

Check your data for collinearity, and check whether predictors have a linear relationship with the response variable. Transform any data that violates assumptions (remember, not all predictors need to be normal as long as the model residuals are normal). Check for any major outliers. Describe what you find.

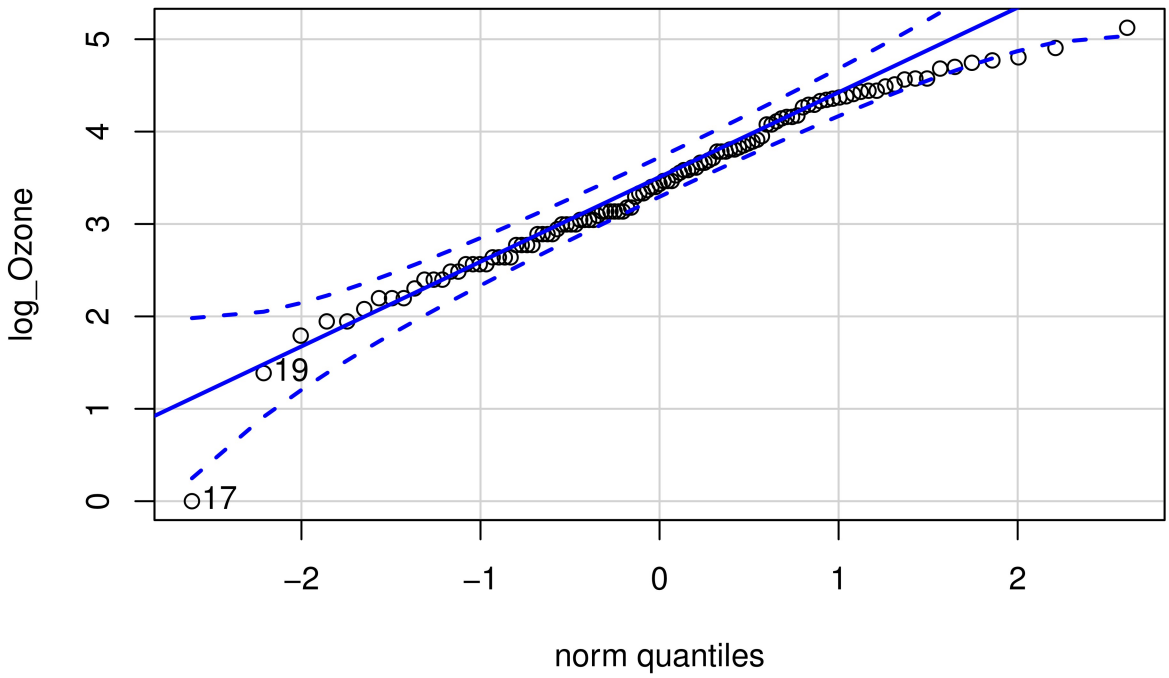
```
# visualizing data
#pairs.panels(airqual, lm = TRUE, cor = T)

# checking normality of solar radiation data
with(airqual, shapiro.test(Solar.R))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Solar.R
## W = 0.94183, p-value = 9.492e-06
```

```
# transforming Ozone data
airqual_sub$log_Ozone <- log(airqual_sub$Ozone)
pairs.panels(airqual_sub)
```




```
with(airqual_sub, qqPlot(log_Ozone))
```




```
## [1] 17 19
```

```
with(airqual_sub, shapiro.test(log_Ozone))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log_Ozone
## W = 0.9712, p-value = 0.01669
```

The histograms for the predictor variables, wind and temperature, look normally distributed but solar radiation does not. I checked Solar.R's QQplot and the data points were within the confidence bands but was not straight which suggests non-normal data. However, when I tried transforming the data it became even more skewed and non-normal so I conducted a shapiro wilk test to see if it is normal (null = data is normally distributed, alternative = data is not normally distributed). The Shapiro Wilk Test produced a p-value smaller than alpha = 0.05 (p = 9.492e-06) so I can reject the null hypothesis that the data is normally distributed. But because transformations made the data even less normal I will assume normality using the central limit theorem since sample size = 153.

The histogram for my response variable, Ozone, was right skewed so I log transformed it and looked at its QQPlot. Most of the data fell within the confidence bands and there are only two ouliers present. I conducted a shapiro wilk test on the transformed Ozone data and although I got a p-value less than alpha = 0.05 (p = 0.01669) I am going to assume normality using the central limit theorem.

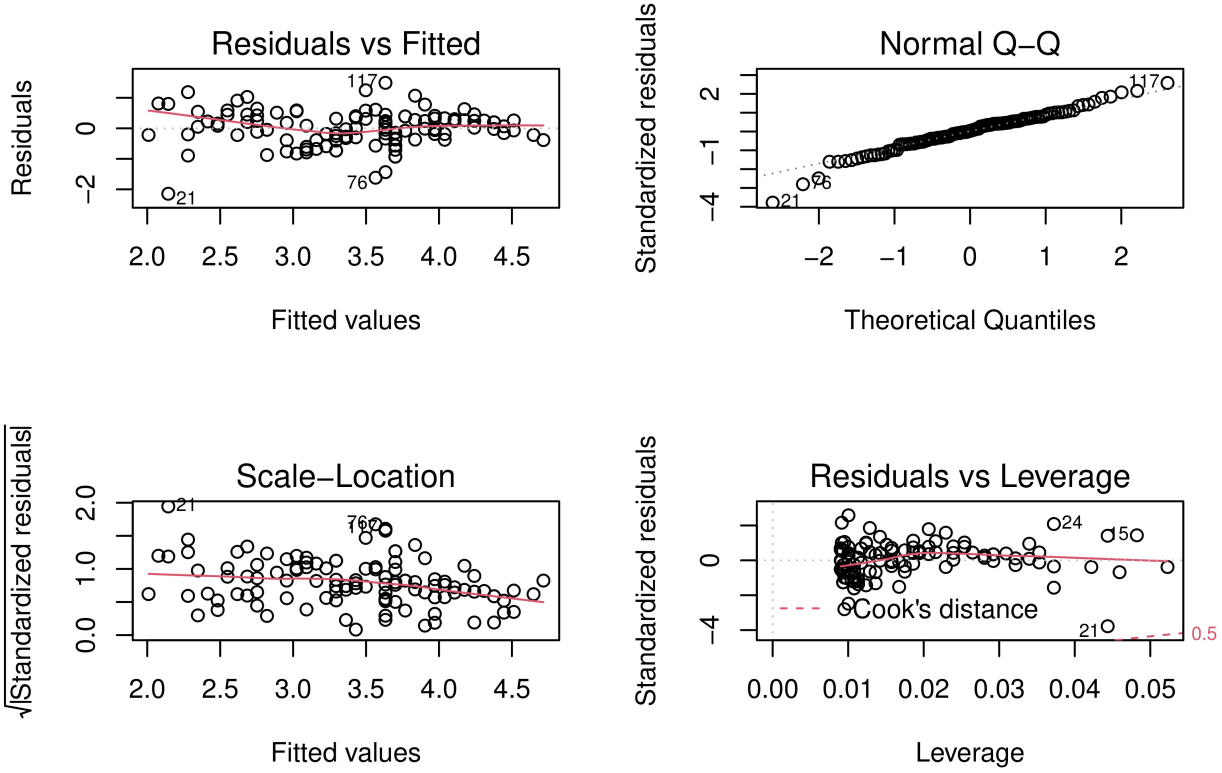
According to the scatterplots, solar radiation, wind, and temperature all have a linear relationship with the transformed response variable. Temperature and wind is especially correlated with log_Ozone. There is

some colinearity between wind and temperature, as well as between temperature and month but the R value is below 0.60 so it will not significantly impact my analysis.

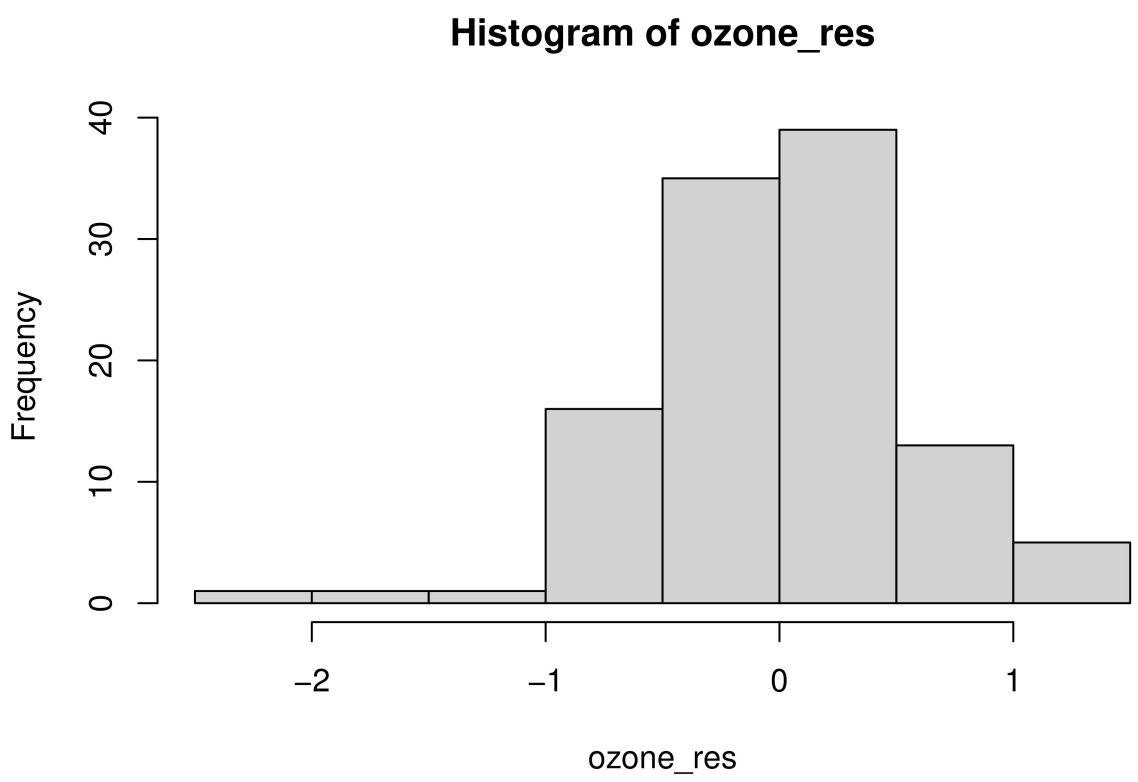
Question 4

Fit a model with one predictor. Check its residuals for normality.

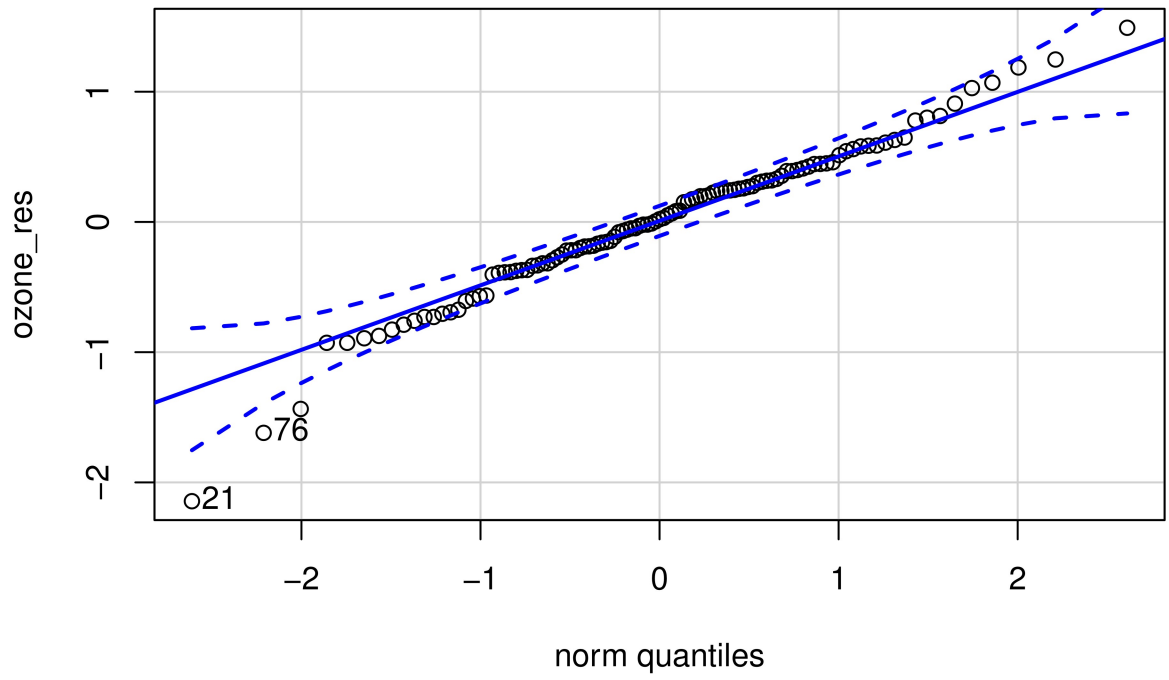
```
# single variable linear model with temperature
fit_1var <- lm(log_Ozone ~ Temp, data = airqual_sub)
par(mfrow = c(2,2))
plot(fit_1var)
```




```
# checking temp residuals
ozone_res = fit_1var$residuals
par(mfrow = c(1,1))
hist(ozone_res)
```




```
qqPlot(ozone_res)
```




```
## 21 76
## 17 45
```

```
shapiro.test(ozone_res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ozone_res
## W = 0.97667, p-value = 0.04867
```

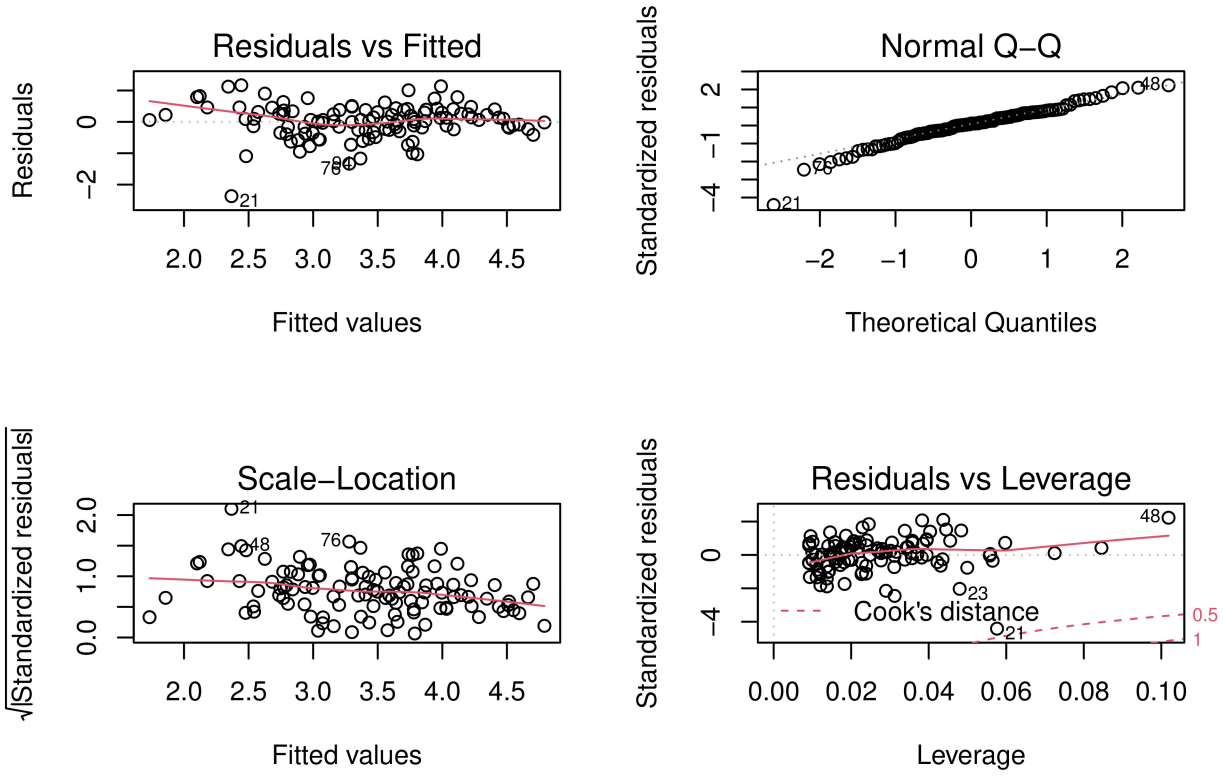
The linear model with only one predictor variable, temperature, has a QQplot that looks normal (most data points follow the straight line) and a residuals vs. fitted graph that does not have a distinct pattern which indicates homogeneity of variance.

The residuals produced a histogram that looked normally distributed (bell-shape), but does not look symmetrical. I checked its qqPlot and most of the data points fell within the confidence bands which indicates normality. The shapiro wilk test produced a p-value equal to alpha = 0.05 (p = 0.04867) which means I marginally fail to reject the null hypothesis that the data is normally distributed. I can also use the central limit theorem (since there is more than 50 datapoints) to assume normality.

Question 5

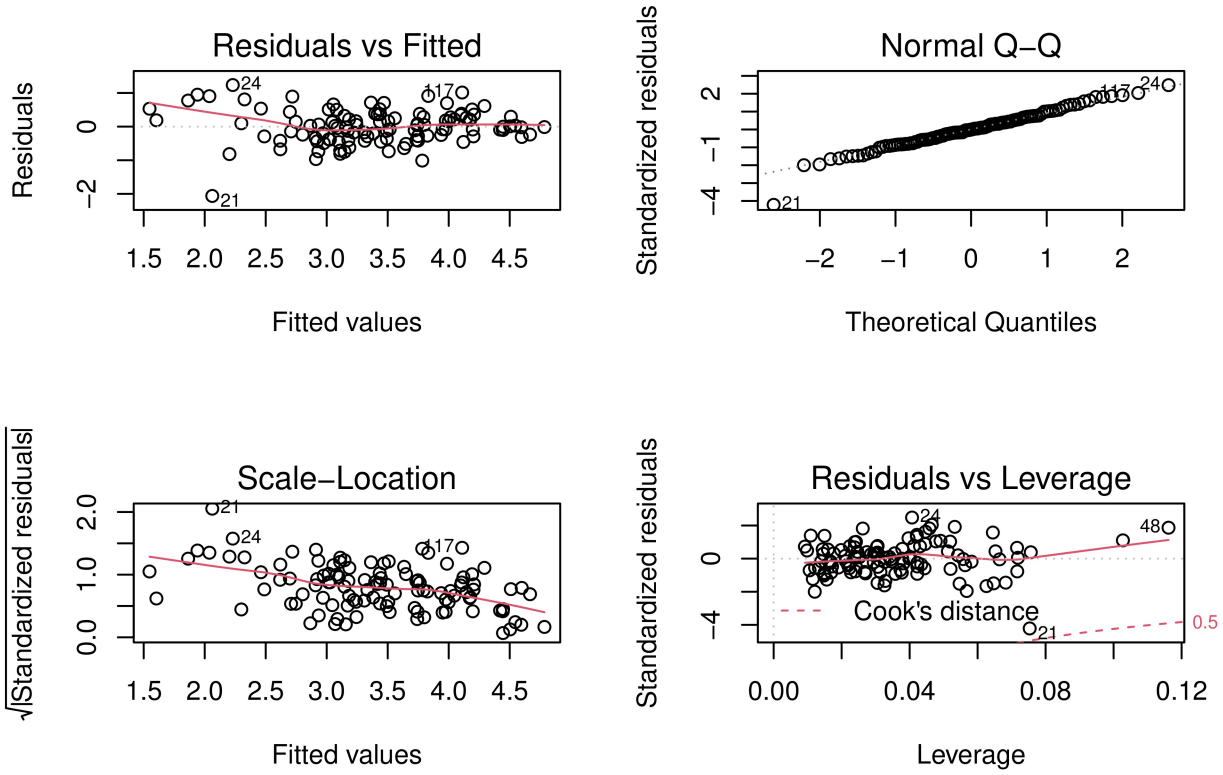
Fit at least two other models with different combinations of predictors. State whether you are including random effects or interaction terms.

```
# multiple variable linear model (temp, wind)
fit_2var <- lm(log_Ozone ~ Temp + Wind, data = airqual_sub)
par(mfrow = c(2,2))
plot(fit_2var)
```



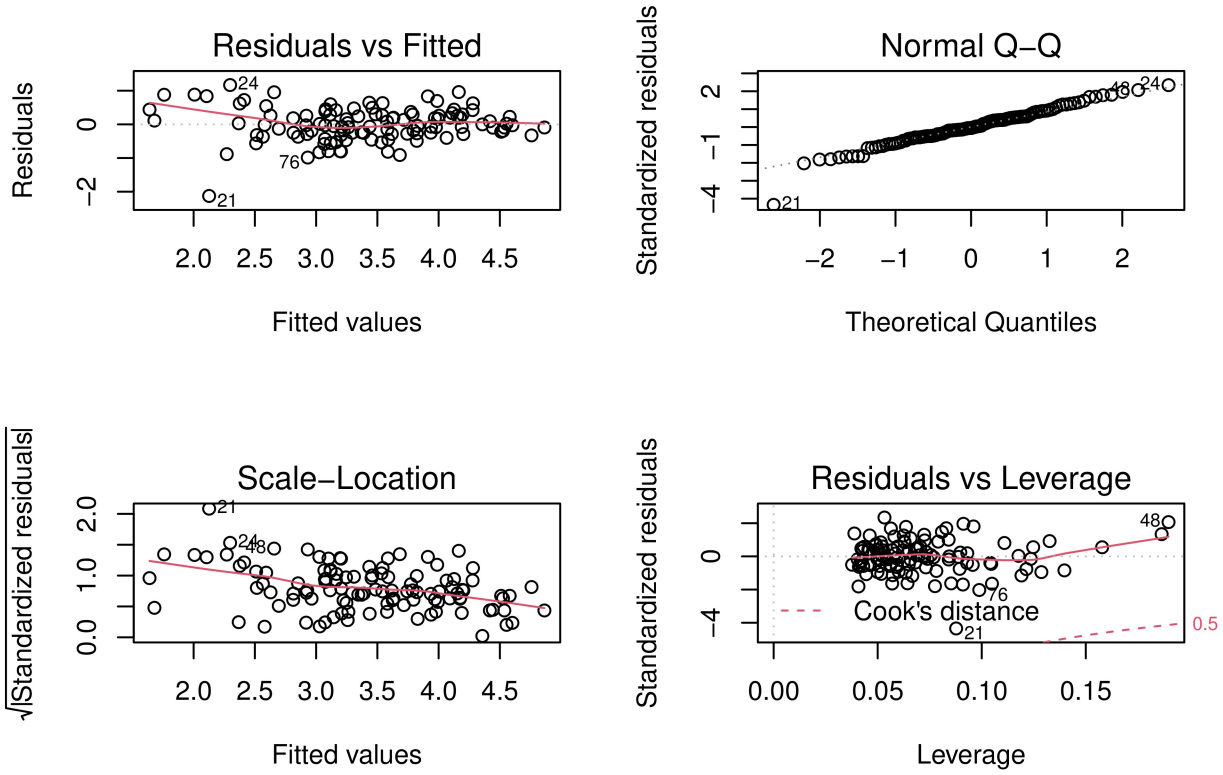
In the linear model with 2 variables I included wind and temperature which are terms that showed interaction in the correlation scatterplots (r = -05).

```
# multiple variable linear model (temp, wind, solar radiation)
fit_3var <- lm(log_Ozone ~ Temp + Wind + Solar.R, data = airqual_sub)
par(mfrow = c(2,2))
plot(fit_3var)
```



In the linear model with 3 variables I included all 3 predictor variables (temp, wind, solar radiation).

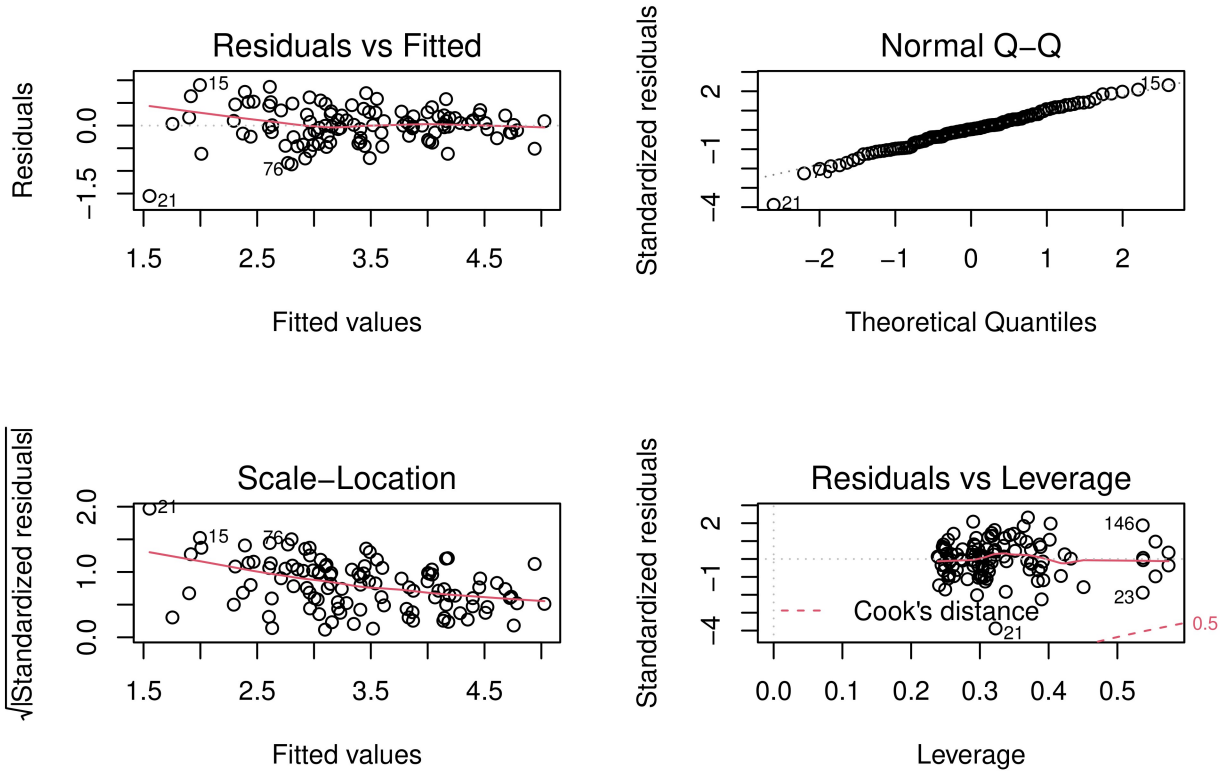
```
# multiple variable linear model (temp, wind, solar.R, Month)
fit_4var <- lm(log_Ozone ~ Temp + Wind + Solar.R + Month,
               data = airqual_sub)
par(mfrow = c(2,2))
plot(fit_4var)
```



In the linear model with 4 variables I included a variable with random effects (Month). I chose to use Month because it had a relatively high correlation with temperature (r = 0.4).

```
# multiple variable linear model with all variables
fit_full <- lm(log_Ozone ~ Temp + Wind + Solar.R +
                 Month + Day, data = airqual_sub)
par(mfrow = c(2,2))
plot(fit_full)
```

```
## Warning: not plotting observations with leverage one:
##   55, 93
```



In this final linear model I included all 5 variables (temp, wind, solar radiation, month, and day).

Question 6

Create a table of results to compare your models using AIC, BIC, and adjusted R-squared. Include the degrees of freedom of each model in the table so that you can consider parsimony in your final decision.

```
#AIC of each model
result <- AIC(fit_1var, fit_2var, fit_3var, fit_4var, fit_full)

# adding other metrics to table
models <- list(fit_1var, fit_2var, fit_3var, fit_4var, fit_full)
result$BIC <- sapply(models, BIC)

model_summary <- lapply(models, summary)

# loop for extracting R^2 and adj R^2 value for each model
for(i in 1:length(models)){
  result$rsq[i] <- model_summary[[i]]$r.squared
  result$adj_rsq[i] <- model_summary[[i]]$adj.r.squared
}
```

```
kable(result, digits = 2, align = "c")
```

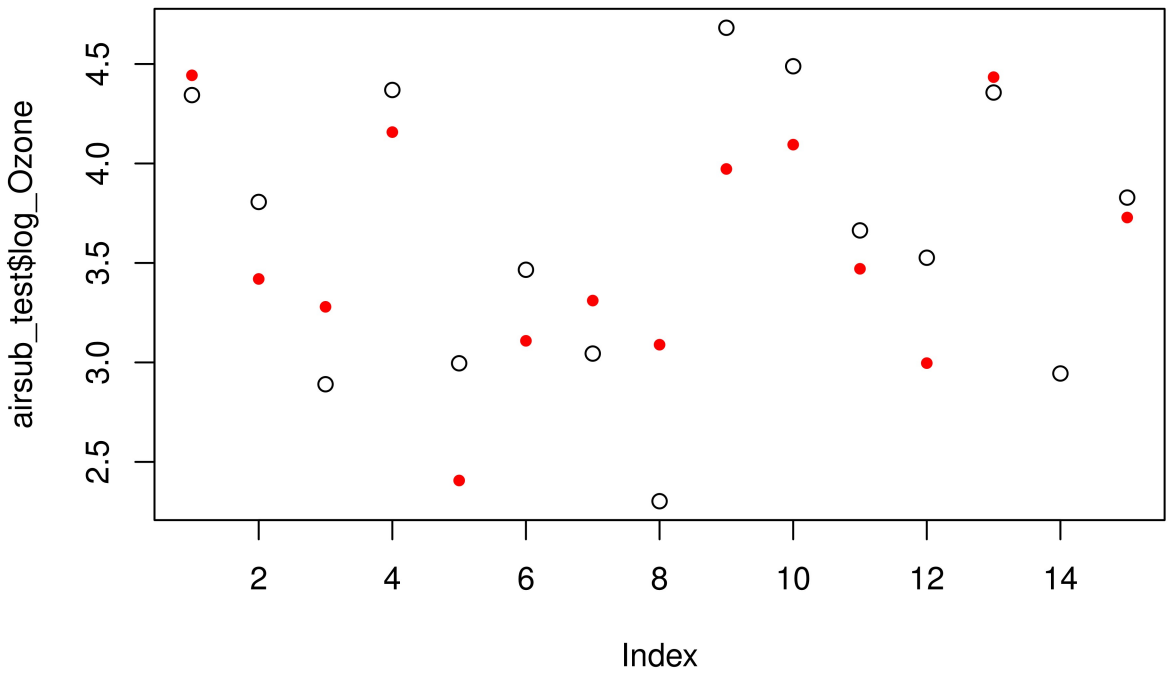
	df	AIC	BIC	rsq	adj_rsq
fit_1var	3	198.21	206.34	0.55	0.55
fit_2var	4	188.20	199.04	0.60	0.59
fit_3var	5	170.83	184.37	0.66	0.66
fit_4var	9	176.32	200.71	0.67	0.65
fit_full	39	186.49	292.16	0.79	0.68

The smallest AIC value is 170.83 from fit_3var and the smallest BIC value is 184.37 also from fit_3var. Fit_full has the highest adjusted R^2 value which means the fit_full model explains a larger portion of the data (this makes sense since it's a model using all the variables of the data).

Question 7

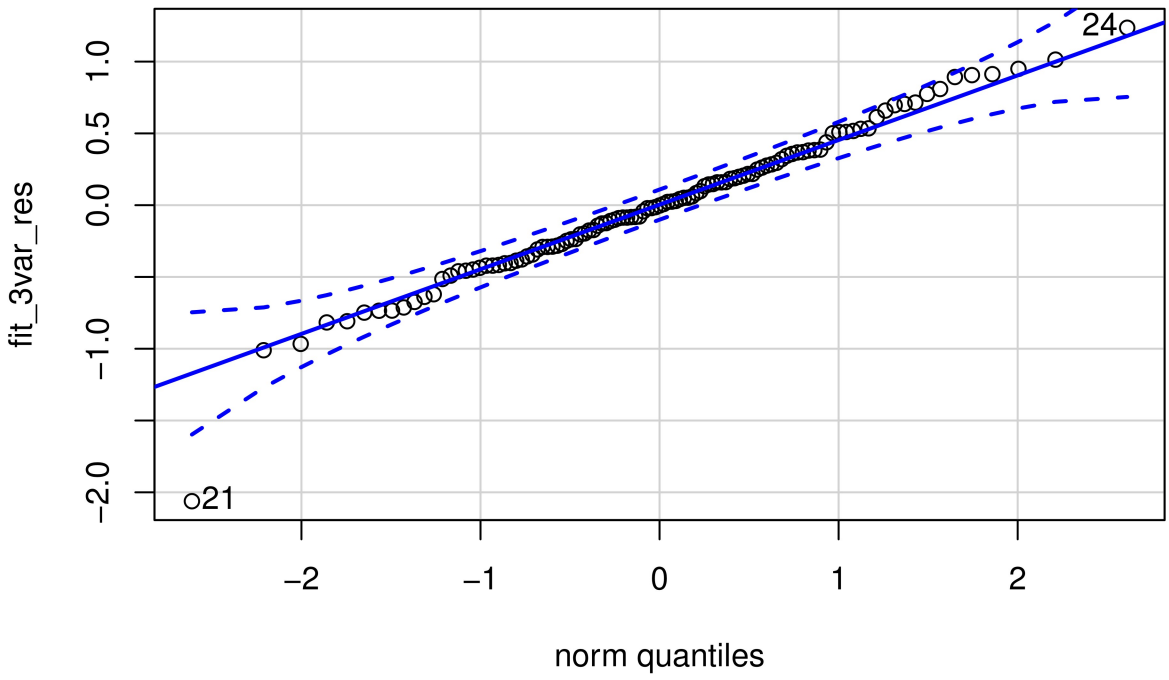
Choose the best model and explain why this was your choice. Check its residuals for normality. How well does this model fit the data? How well does it predict ozone levels?

```
# separating data into training set and test set
splitter <- sample(1:nrow(airqual_sub), 15, replace = F)
airsub_train <- airqual_sub[-splitter,]
airsub_test <- airqual_sub[splitter,]

# fitting final model
fit_3var_split <- lm(log_Ozone ~ Temp + Wind + Solar.R, data = airsub_train)

# predicted values for test data
prediction <- predict(fit_3var_split, airsub_test)

# plotting test data values
plot(airsub_test$log_Ozone, pch = 1)
# plotting model predictions
points(prediction, pch = 20, col = "red")
```




```
# checking residuals for fit_3var
fit_3var_res = fit_3var$residuals
hist(fit_3var_res)
```


```
qqPlot(fit_3var_res)
```




```
## 21 24
## 17 20
```

```
shapiro.test(fit_3var_res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit_3var_res
## W = 0.97749, p-value = 0.05726
```

The model that I chose is fit_3var which includes all three predictor variables (wind, temperature, and solar radiation) and does not include the random effects of Month and Day. I chose this model because it had the lowest AIC value (170.83) and the lowest BIC value (184.37) while staying parsimonious compared to the other models with lower AIC/BIC values.

A Shapiro Wilk test was done on the model's residual and produced a p-value larger than alpha = 0.05 (p = 0.05726) which means I fail to reject the null hypothesis that the data is normal. The qqPlot of the residuals also fell within the confidence bands so I can confidently say the residuals are normal. The adjusted R^2 value was not the highest but it still told me that 66% of the data can be explained with this model (adj R^2 = 0.66). So, the model is a good fit for the data.

As you can see from the prediction graph above, the model fits the data quite well, with predictions (red dots) following a similar pattern with the test data values. When I refreshed the code, it continued to produce prediction values that have a similar pattern to the test data values.