STA 5207: Homework 6

Due: Friday, March 1 by 11:59 PM

Include your R code in an R chunks as part of your answer. In addition, your written answer to each exercise should be self-contained so that the grader can determine your solution without reading your code or deciphering its output.

Exercise 1 (Diagnostics for Teenage Gambling Data) [40 points]

For this exercise we will use the `teengamb` data set from the `faraway` package. You can also find that data in `teengamb.csv` on Canvas. You can use `?teengamb` to learn about the data set. The variables in the data set are

- `sex`: 0 = male, 1 = female.
- `status`: Socioeconomic status score based on parents' occupation.
- `income`: in pounds per week.
- `verbal`: verbal score in words out of 12 correctly defined.
- `gamble`: expenditure on gambling in pounds per year.

In the following exercise, use `gamble` as the response and the other variables as predictors. Some of these questions are subjective, so there may not be a "right" answer. Just make sure to justify your answer based on the plots and statistical tests.

1. (8 points) Check the constant variance assumption for this model using a graphical method and a hypothesis test at the $\alpha = 0.05$ significance level. Do you feel it has been violated? Justify your answer. Include any plots in your response.

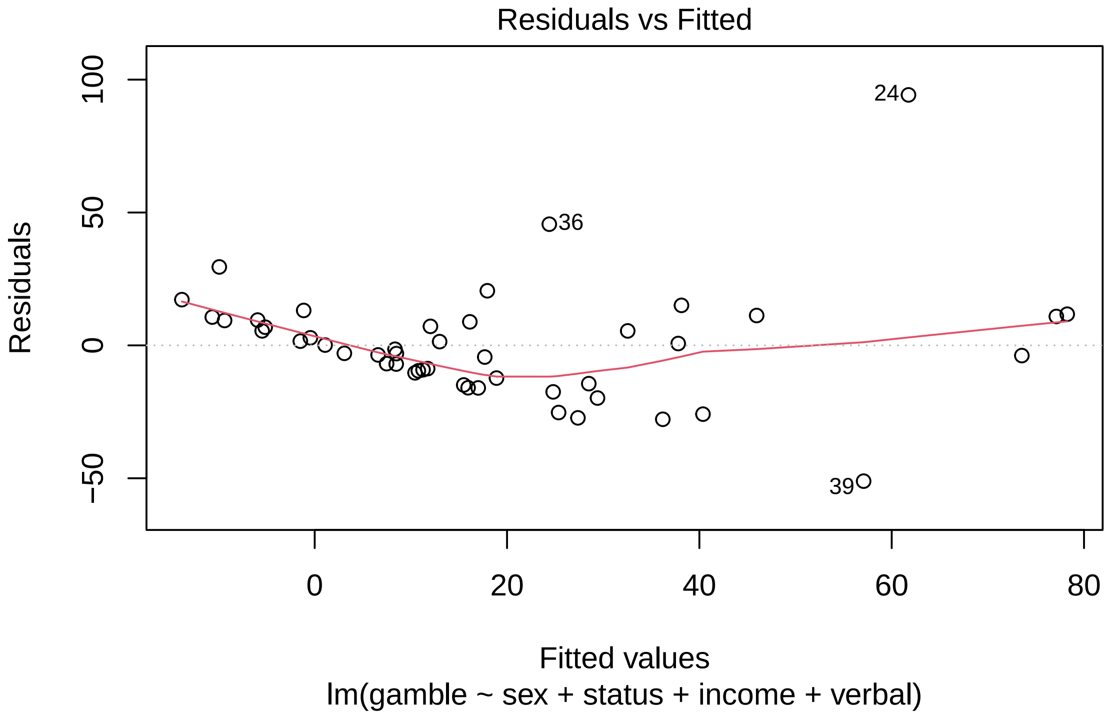
```r
library('faraway')
library('lmtest')
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
data(teengamb, package='faraway')
model <- lm(gamble ~ sex + status + income + verbal, data=teengamb)
plot(model, which=1)
```



```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 6.4288, df = 4, p-value = 0.1693
```

Answer: Since the p-value is 0.169 which is greater than our significance level, we do not reject the null hypothesis at the 0.05 significance level and determine that the constant variance assumption is not violated.

2. (8 points) Check the normality assumption using a Q-Q plot and a hypothesis test at the $\alpha = 0.05$ significance level. Do you feel it has been violated? Justify your answer. Include any plots in your response.

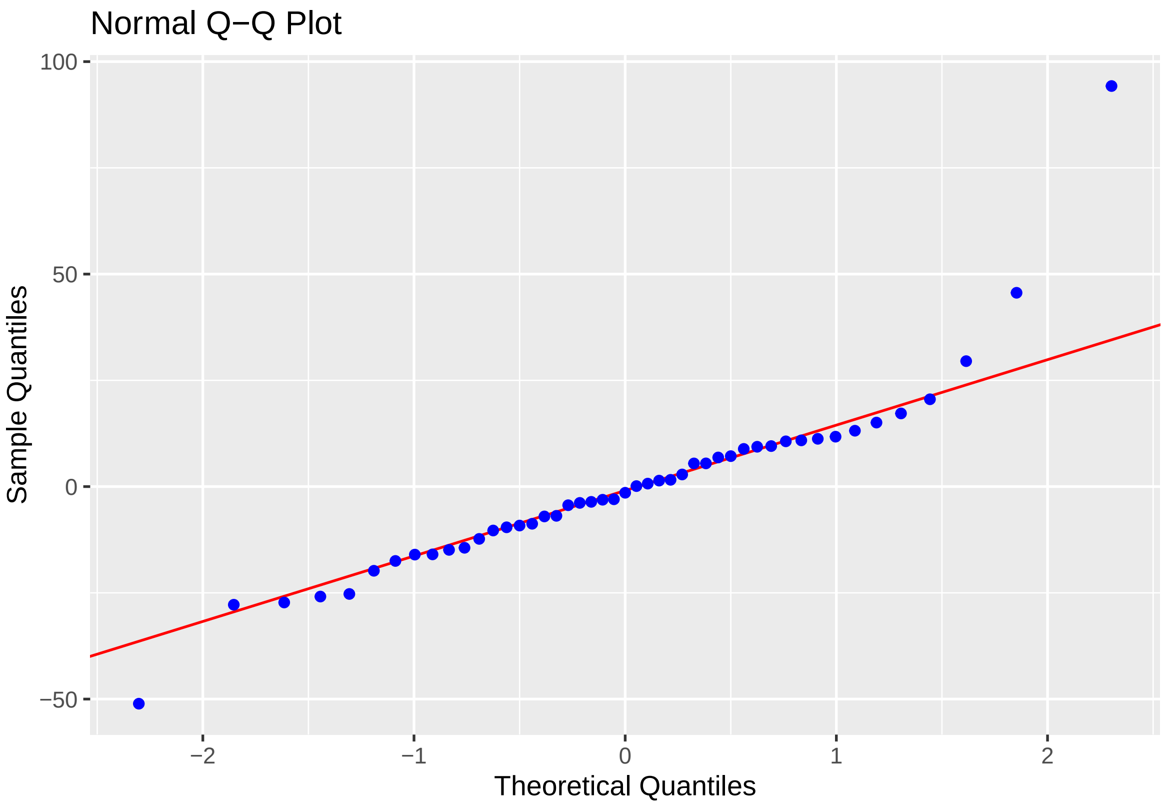
```
library('olsrr')
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:faraway':
##
##     hsb

## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_plot_resid_qq(model)
```



```
shapiro.test(resid(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.86839, p-value = 8.16e-05
```

Answer: Since our Shapiro-Wilk test resulted in a p-value of less than 0.05, we do reject the null hypothesis and say the normality assumption has been violated.

3. (5 points) Check for any high leverage points. Report any observations you determine to have high leverage.

```
which(hatvalues(model) > 2 * mean(hatvalues(model)))
```

```
## 31 33 35 42
## 31 33 35 42
```

Answer: We see there to be 4 observations (31, 33, 35, and 42) which have high leverage points.

4. (5 points) Check for any outliers in the data set at the $\alpha = 0.05$ significance level. Report any observations you determine to be outliers.

```r
outlier_test_cutoff = function(model, alpha = 0.05) {
    n = length(resid(model))
    qt(alpha/(2 * n), df = df.residual(model) - 1, lower.tail = FALSE)
}
cutoff = outlier_test_cutoff(model, alpha = 0.05)
which(abs(rstudent(model)) > cutoff)
```

```
## 24
## 24
```

Answer: We see one outlier at observation 24.

5. (5 points) Check for any highly influential points in the data set. Report any observations your determine are highly influential.

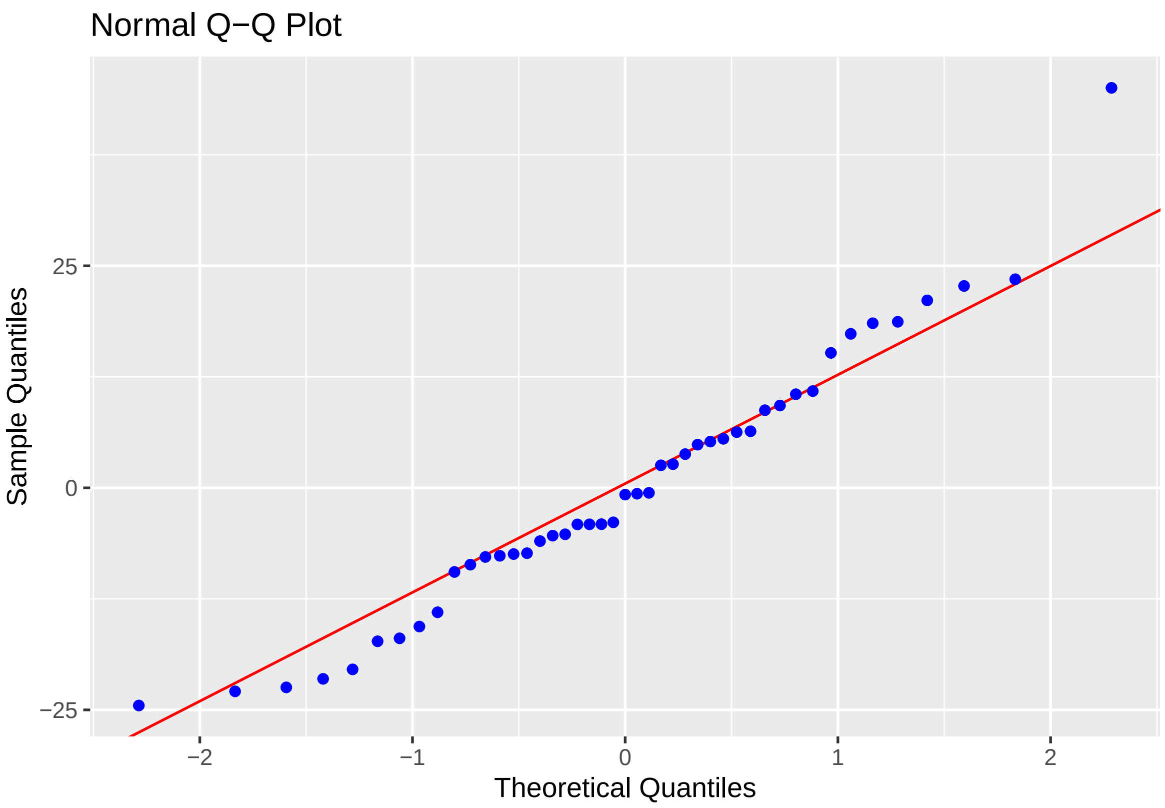
```r
which(cooks.distance(model) > 4 / length(cooks.distance(model)))
```

```
## 24 39
## 24 39
```

Answer: We see that observations 24 and 39 are highly influential points in the data set after using Cook's distance.

6. (9 points) Fit a model with the high influence points you found in the previous question removed. Perform a hypothesis test at the $\alpha = 0.05$ significance level to check the normality assumption. What do you conclude?

```r
noninfluential_ids <- which(
    cooks.distance(model) <= 4 / length(cooks.distance(model)))
model_fix <- lm(gamble ~ sex + status + income + verbal,
                data=teengamb,
                subset = noninfluential_ids)
ols_plot_resid_qq(model_fix)
```



```
shapiro.test(resid(model_fix))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_fix)
## W = 0.96728, p-value = 0.23
```

Answer: Upon removing the influential points, the p-value of our test to check the normality assumption is far greater than our significance level of 0.05. Therefore, we do not rejecct the null hypothesis and conclude that the normality assumption has not been violated.

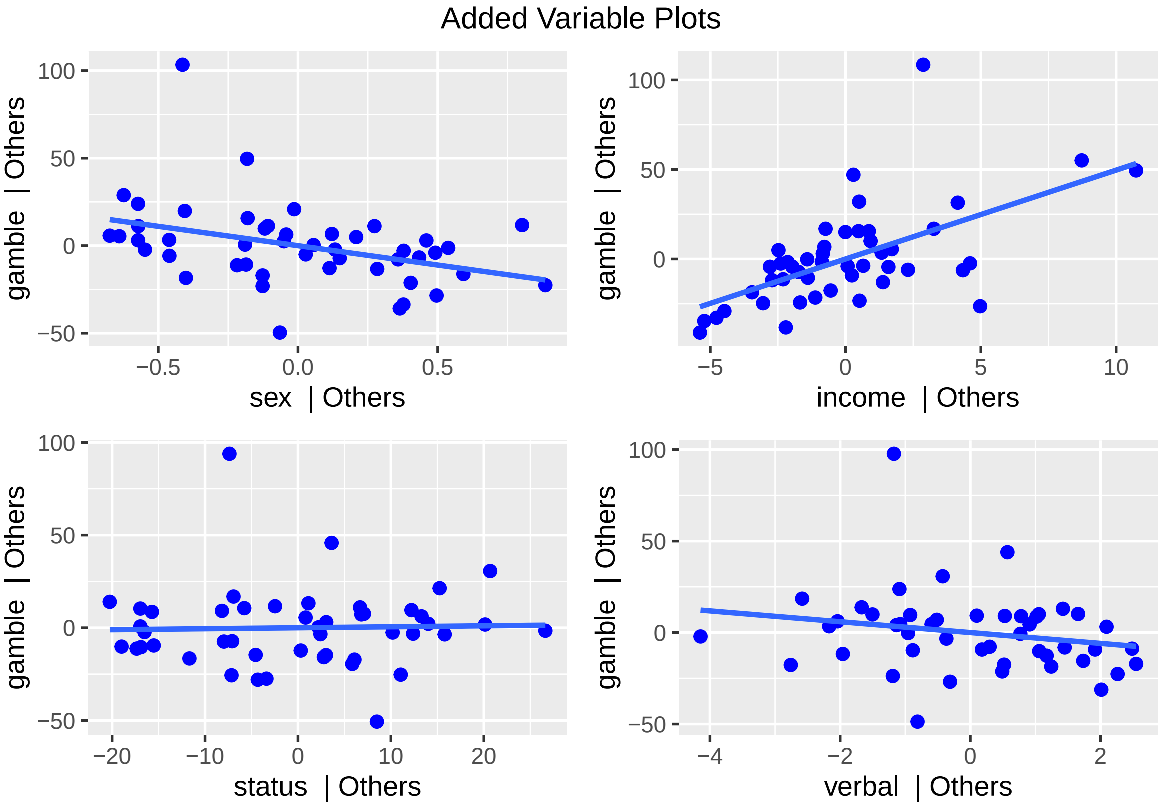
Exercise 2 (Add Variable Plots for the Teenage Gambling Data) [20 points]

For this exercise, we will also use the `teengamb` data set from the `faraway` package. Some of these questions are subjective, so there may not be a "right" answer. Just make sure to justify your answer based on the plots and statistical tests.

1. (8 points) Fit a multiple linear regression model with `gamble` as the response and the other four variables as predictors. Obtain the partial regression plots. For each predictor, determine if it appears to have a linear relationship with the response after removing the effects of the other predictors based on these plots. Include the plots in your response.

```
model <- lm(gamble ~ sex + status + income + verbal, data=teengamb)
ols_plot_added_variable(model)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



Answer: It looks like all predictors have a linear relationship with "gamble" after removing the effects of the other predictors.

2. (8 points) Fit the following two models and obtain their residuals:

   - Model 1: `gamble` ~ `verbal` + `status` + `sex`.
   - Model 2: `income` ~ `verbal` + `status` + `sex`.

   Next fit a simple linear regression model with the residuals of Model 1 as the response and the residuals of Model 2 as the predictor. Report the value of the slope parameter.

```
model1 <- lm(gamble ~ verbal + status + sex, data=teengamb)
model2 <- lm(income ~ verbal + status + sex, data=teengamb)

model3 <- lm(resid(model1) ~ resid(model2))
model3
```

```
##
## Call:
## lm(formula = resid(model1) ~ resid(model2))
##
## Coefficients:
##   (Intercept)  resid(model2)
##    -2.280e-16      4.962e+00
```

Answer: The value of the slope parameter of the model with the residuals of model 1 as the response and the residuals of model 2 as the predictor is 4.962.

3. (4 points) Compare the coefficient of `income` from the model fit in part 1 (`gamble` ~ `verabal` + `status` + `sex` + `income`) to the value of the slope parameter in part 2. Are their values the same or different?

```
coef(model)
```

```
##  (Intercept)          sex       status       income       verbal
##  22.55565063 -22.11833009   0.05223384   4.96197922  -2.95949350
```

```
coef(model3)
```

```
##   (Intercept) resid(model2)
## -2.279512e-16  4.961979e+00
```

Answer: These values are the same, both at 4.961979.

Exercise 3 (Diagnostics for Swiss Fertility Data) [40 points]

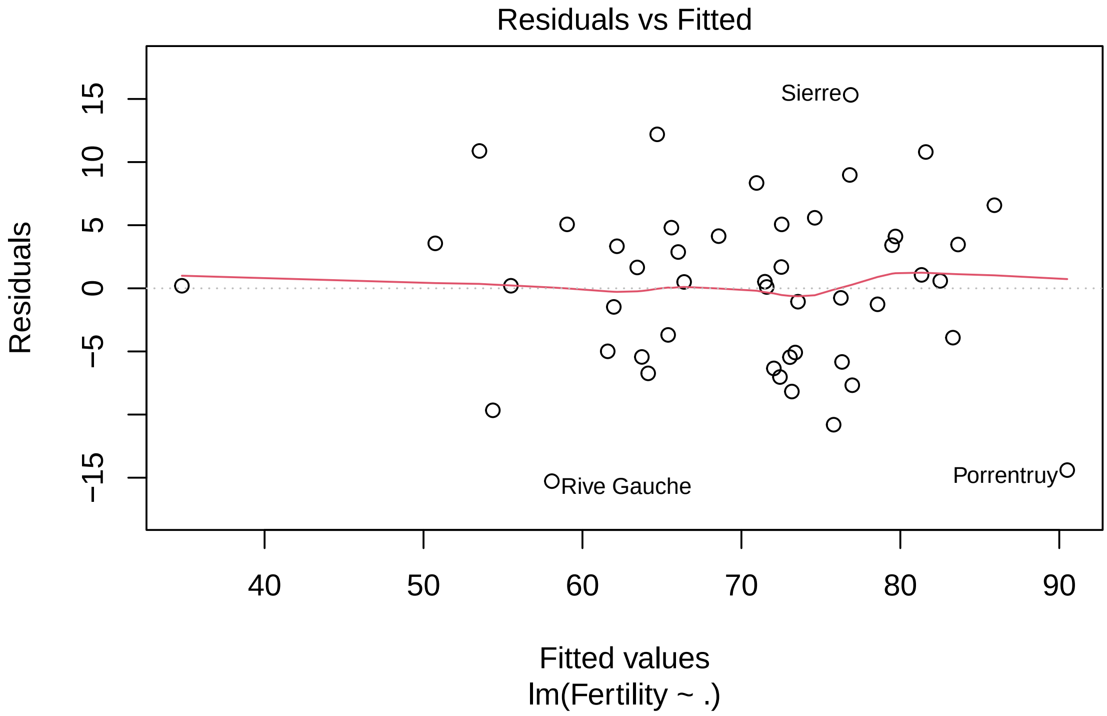
For this exercise we will use the `swiss` data set from the `faraway` package. You can also find the data in `swiss.csv` on Canvas. You can use `?swiss` to learn about the data set. The variables in the data set are

- `Fertility`: a 'common standardized fertility measure'.
- `Agriculture`: proportion of males involved in agriculture as an occupation.
- `Examination`: proportion of draftees receiving the highest mark on army examination.
- `Education`: proportion with education beyond primary school for draftees.
- `Catholic`: proportion 'catholic' (as opposed to 'protestant').
- `Infant.Mortality`: proportion of live births who live less than 1 year.

In the following exercise, use `Fertility` as the response and the other variables as predictors. Some of these questions are subjective, so there may not be a "right" answer. Just make sure to justify your answer based on the plots and statistical tests.

1. (8 points) Check the constant variance assumption for this model using a graphical method and a hypothesis test at the $\alpha = 0.05$ significance level. Do you feel it has been violated? Justify your answer. Include any plots in your response.

```
data(swiss)
model <- lm(Fertility ~ ., data=swiss)
plot(model, which=1)
```



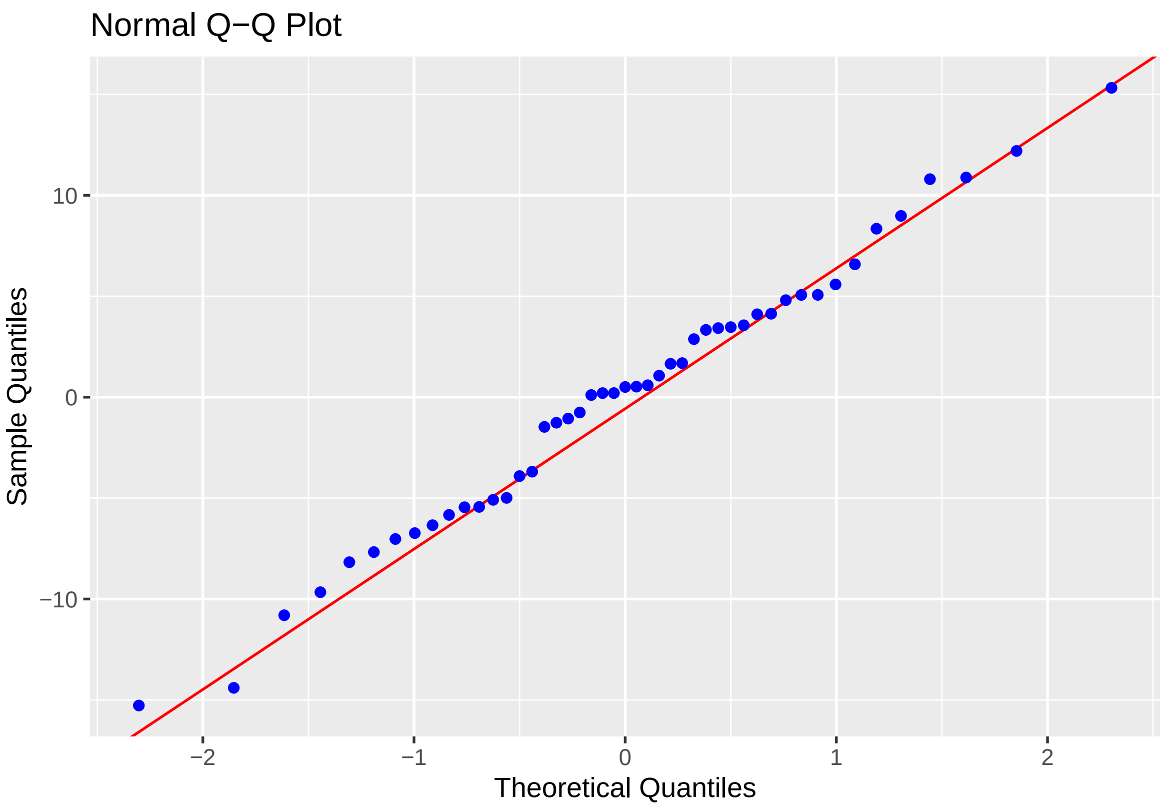
```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 5.8511, df = 5, p-value = 0.321
```

Answer: Since the p-value is .321 which is greater than our significance level, we do not reject the null hypothesis at the 0.05 significance level and determine that the constant variance assumption is not violated.

2. (8 points) Check the normality assumption using a Q-Q plot and a hypothesis test at the $\alpha = 0.05$ significance level. Do you feel it has been violated? Justify your answer. Include any plots in your response.

```
ols_plot_resid_qq(model)
```



```
shapiro.test(resid(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.98892, p-value = 0.9318
```

Answer: Since our Shapiro-Wilk test resulted in a p-value of more than 0.05, we do not reject the null hypothesis and say the normality assumption has not been violated.

3. (4 points) Check for any high leverage points. Report any observations you determine to have high leverage.

```
which(hatvalues(model) > 2 * mean(hatvalues(model)))
```

```
##    La Vallee V. De Geneve
##           19           45
```

Answer: We observe leverage points at La Vallee and V. De Geneve

4. (4 points) Check for any outliers in the data set at the $\alpha = 0.05$ significance level. Report any observations you determine to be outliers.

```
outlier_test_cutoff = function(model, alpha = 0.05) {
    n = length(resid(model))
    qt(alpha/(2 * n), df = df.residual(model) - 1, lower.tail = FALSE)
```

```r
}
cutoff = outlier_test_cutoff(model, alpha = 0.05)
which(abs(rstudent(model)) > cutoff)
```

```
## named integer(0)
```

Answer: We see no outliers in the data set.

5. (4 points) Check for any highly influential points in the data set. Report any observations your determine are highly influential.

```r
which(cooks.distance(model) > 4 / length(cooks.distance(model)))
```

```
##  Porrentruy      Sierre   Neuchatel Rive Droite Rive Gauche 
##           6          37          42          46          47
```

Answer: We see 5 outliers in the data set at observations Porrentruy, Sierre, Neuchatel, Rive Droitte, and Rive Gauche.

6. (6 points) Compare the regression coefficients including and excluding the influential observations. Comment on the difference between these two sets of coefficients.

```r
coef(model)
```

```
##      (Intercept)      Agriculture      Examination        Education 
##       66.9151817       -0.1721140       -0.2580082       -0.8709401 
##         Catholic Infant.Mortality 
##        0.1041153        1.0770481
```

```r
noninfluential_ids = which(
    cooks.distance(model) <= 4 / length(cooks.distance(model)))

# fit the model on non-influential subset
model_fix = lm(Fertility ~ .,
               data = swiss,
               subset = noninfluential_ids)

# return coefficients
coef(model_fix)
```

```
##      (Intercept)      Agriculture      Examination        Education 
##      66.44458475      -0.21819812      -0.50016393      -0.69046520 
##         Catholic Infant.Mortality 
##       0.09846806       1.35767263
```

Answer: The intercepts, and coefficients for agriculture look pretty similar to each other, where as the other predictors have pretty different coefficient values.

7. (6 points) Compare the predictions at the highly influential observations based on a model that includes and excludes the influential observations. Comment on the difference between these two sets of predictions.

```
influential_obs = subset(
    swiss, cooks.distance(model) > 4 / length(cooks.distance(model)))
predict(model, influential_obs)
```

```
##  Porrentruy      Sierre   Neuchatel Rive Droite Rive Gauche
##    90.50011    76.87869    53.51934    54.36209    58.07426
```

```
predict(model_fix, influential_obs)
```

```
##  Porrentruy      Sierre   Neuchatel Rive Droite Rive Gauche
##    94.43980    76.33683    55.89622    57.92582    61.32012
```

Answer: The predictions for Sierre doesn't change nearly at all, but the rest of the observations change quite a bit between the two models.