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.

```r
library(faraway)
library(olsrr)
```

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

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

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

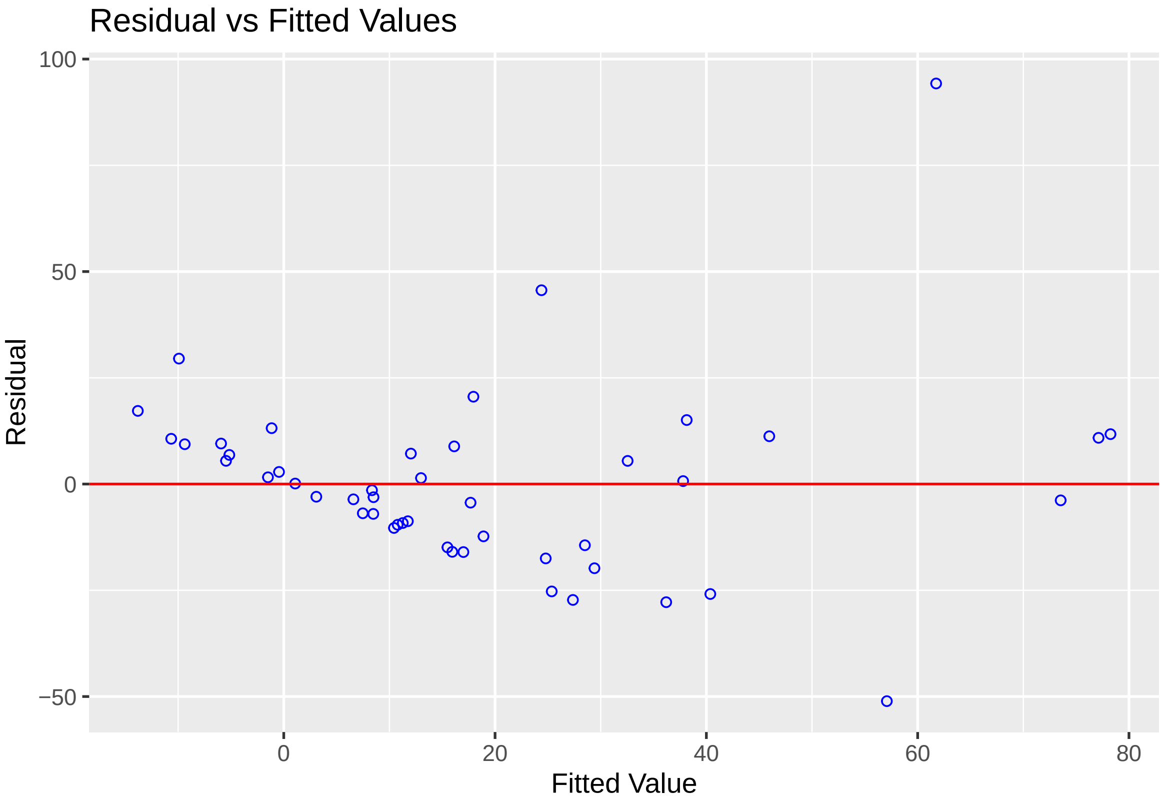
```r
library(lmtest)
```

```
## Loading required package: zoo

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

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.

```
model = lm(gamble ~ ., data = teengamb)
ols_plot_resid_fit(model)
```



```
bptest(model)
```

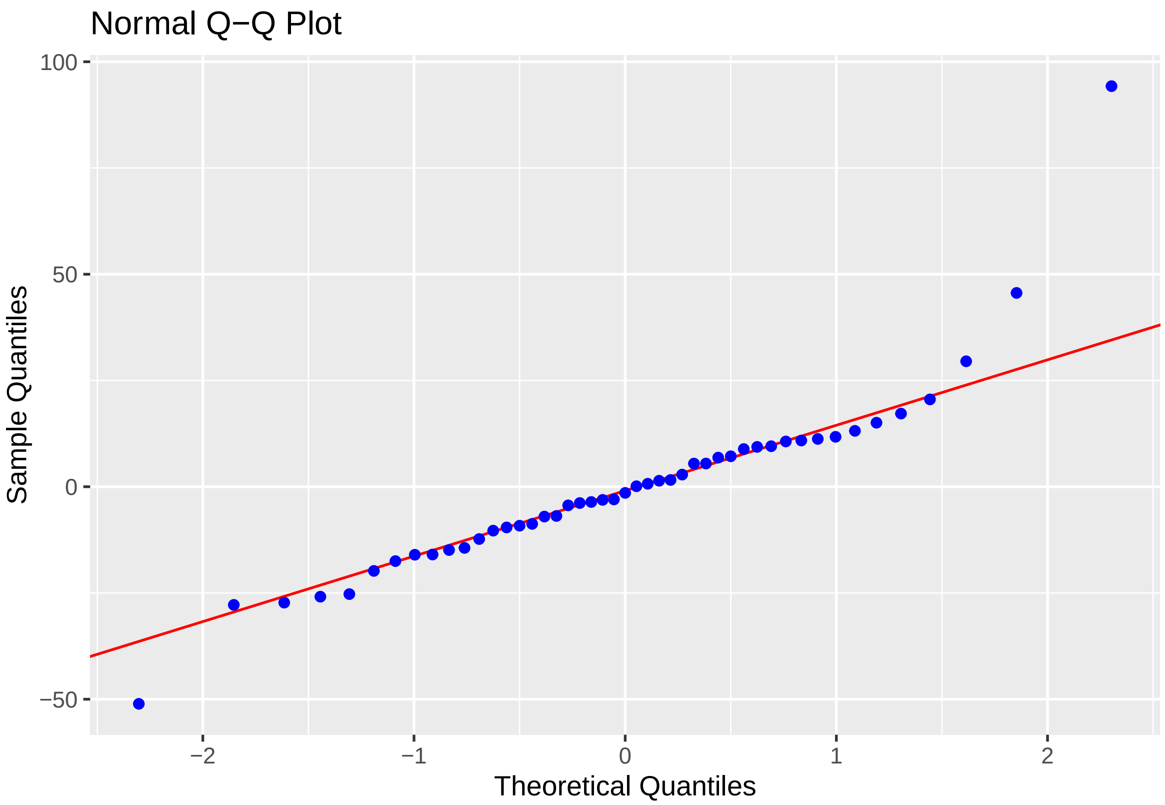
```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 6.4288, df = 4, p-value = 0.1693
```

The null and alternative hypotheses are H_0: Homoscedastic errors and H_1: Heteroscedastic errors. The test statistic is : 6.4288 with a p-value of .1693. We conclude the constant variance assumption is not violated.

From the graphical test, I would believe that the constant variance assumption is violated. However, after performing a hypothesis test, I am convinced that it is simply hard to tell with this dataset.

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 = 0.86839, p-value = 8.16e-05
```

The p-value is 8.16×10^{-5}. We reject the null hypothesis and conclude that the errors are not normally distributed.

I would conclude that the normality assumption is violated. There are two obvious tails on the Q-Q plot, and the hypothesis tests support this conclusion.

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
```

The high leverage points occur at observations 31, 33, 35, and 42

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.

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

# vector of indices for observations deemed outliers.
cutoff = outlier_test_cutoff(model, alpha = 0.05)

which(abs(rstudent(model)) > cutoff)
```

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

There is an 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.

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

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

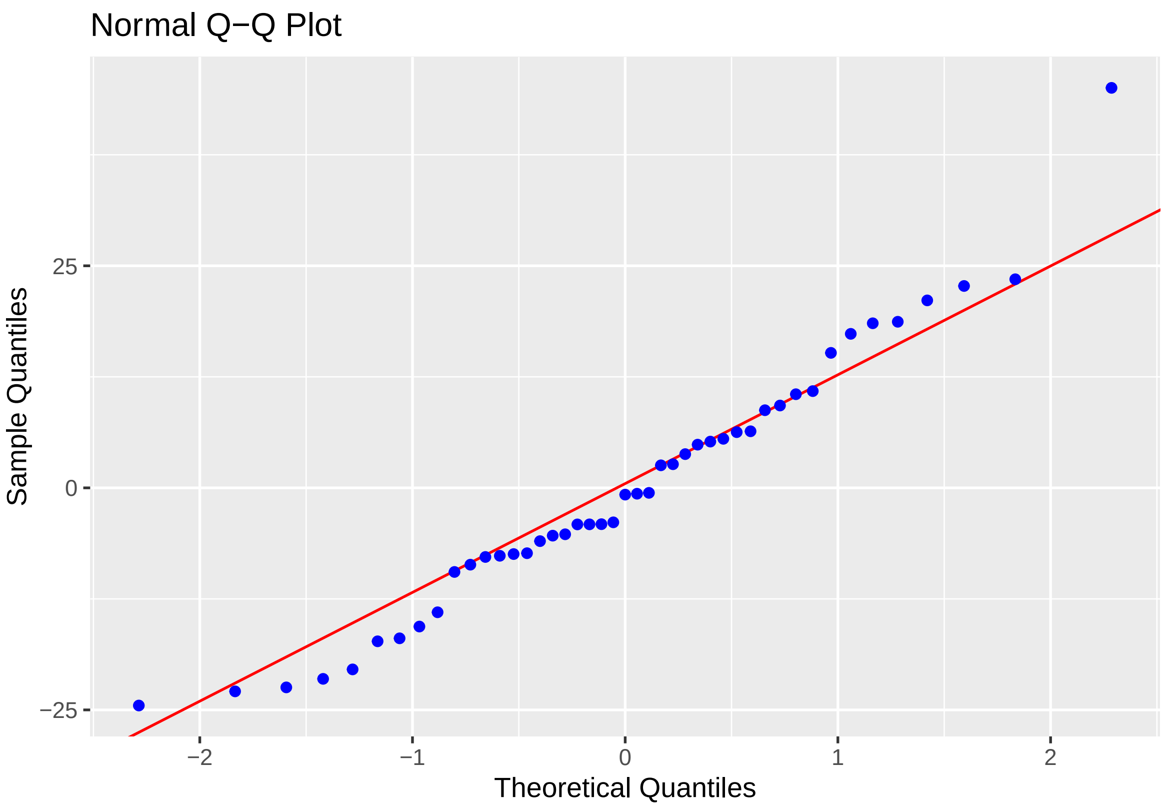
The highly influential points occur at observations 24 and 39.

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?

```
# ids for non-influential observations
noninfluential_ids = which(
    cooks.distance(model) <= 4 / length(cooks.distance(model)))

# fit the model on non-influential subset
fix_model = lm(gamble ~ sex + status + income + verbal,
               data = teengamb,
               subset = noninfluential_ids)

ols_plot_resid_qq(fix_model)
```



```
shapiro.test(resid(fix_model))
```

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

The p-value is .24. We do not reject the null hypothesis and conclude that the errors are normally distributed.

The tails are significantly reduced on the plot, and the hypothesis test supports the conclusion that the high influence points were causing a violation of the normality assumption.

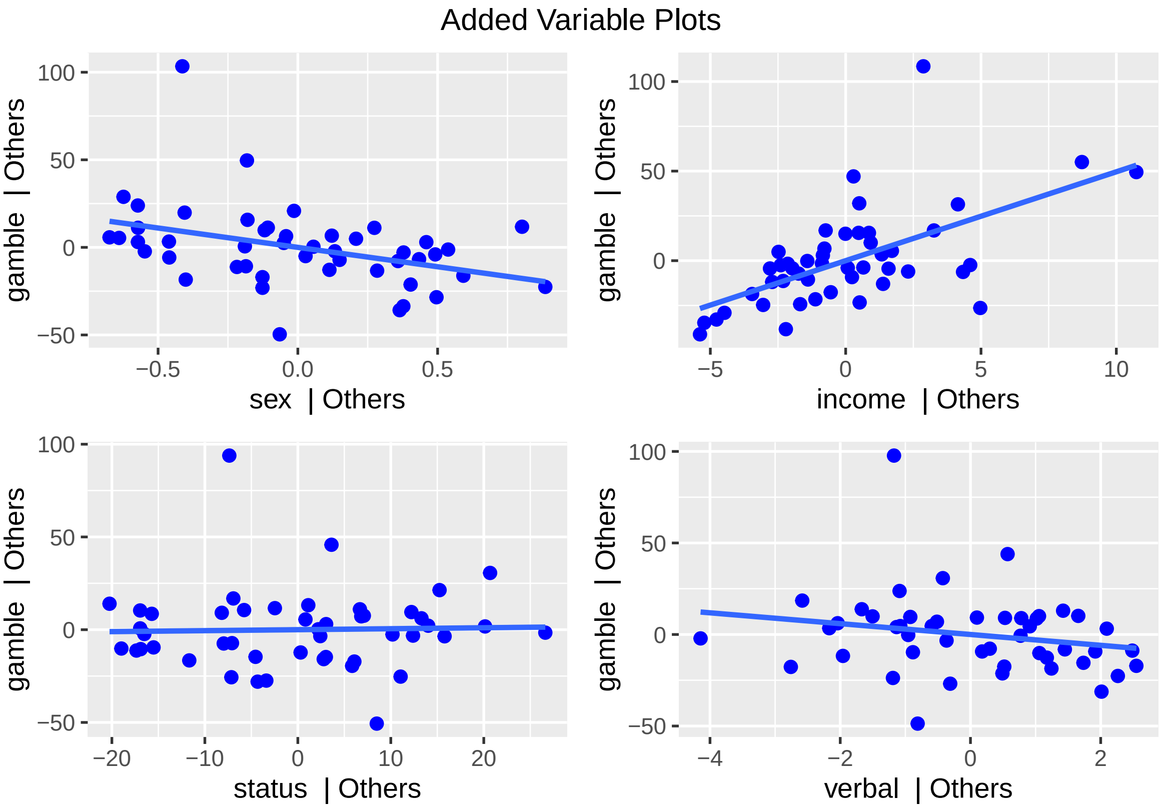
Exercise 2 (Add Variable Plots for the Teenage Gambling Data) [20 points]

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.

```r
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'
```



It appears that each predictor has a linear relationship with the response when they are by themselves in the equation.

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.

```r
model1 = lm(gamble ~ verbal + status + sex, data = teengamb)
model2 = lm(income ~ verbal + status + sex, data = teengamb)
model3 = lm(resid(model1) ~ resid(model2))
summary(model3)$coeff
```

```
##                    Estimate Std. Error       t value     Pr(>|t|)
## (Intercept)   -2.279512e-16  3.1974995 -7.129045e-17 1.000000e+00
## resid(model2)  4.961979e+00  0.9906231  5.008948e+00 8.929015e-06
```

The slope parameter is 4.61979.

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?

```
summary(model)$coeff
```

```
##                  Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  22.55565063 17.1968034  1.3116188 1.967736e-01
## sex         -22.11833009  8.2111145 -2.6937062 1.011184e-02
## status        0.05223384  0.2811115  0.1858118 8.534869e-01
## income        4.96197922  1.0253923  4.8391032 1.791882e-05
## verbal       -2.95949350  2.1721503 -1.3624718 1.803109e-01
```

They are the same.

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

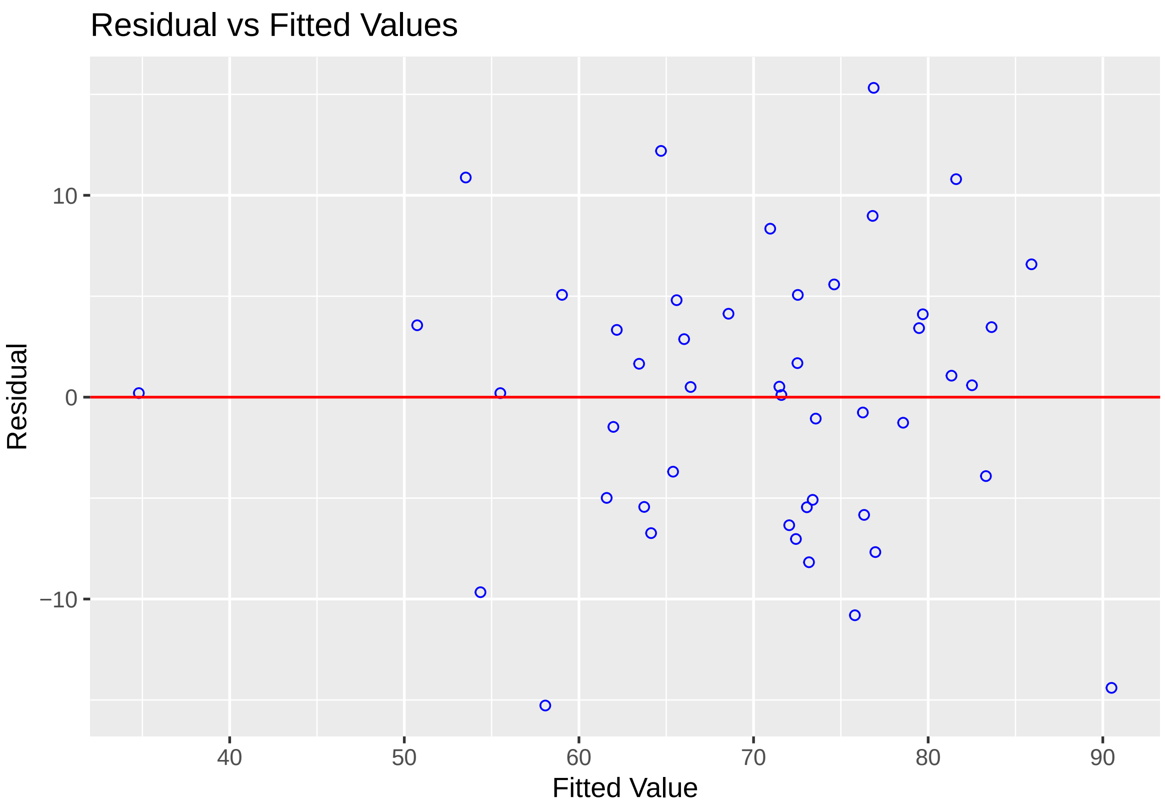
For this exercise we will use the `swiss` data set from the `faraway` package. You can also find the data in `swiss.csv` on Canvas. You can use `?swiss` to learn about the data set. The variables in the data set are

- `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.

```
model = lm(Fertility ~ ., data = swiss)
ols_plot_resid_fit(model)
```



```
bptest(model)
```

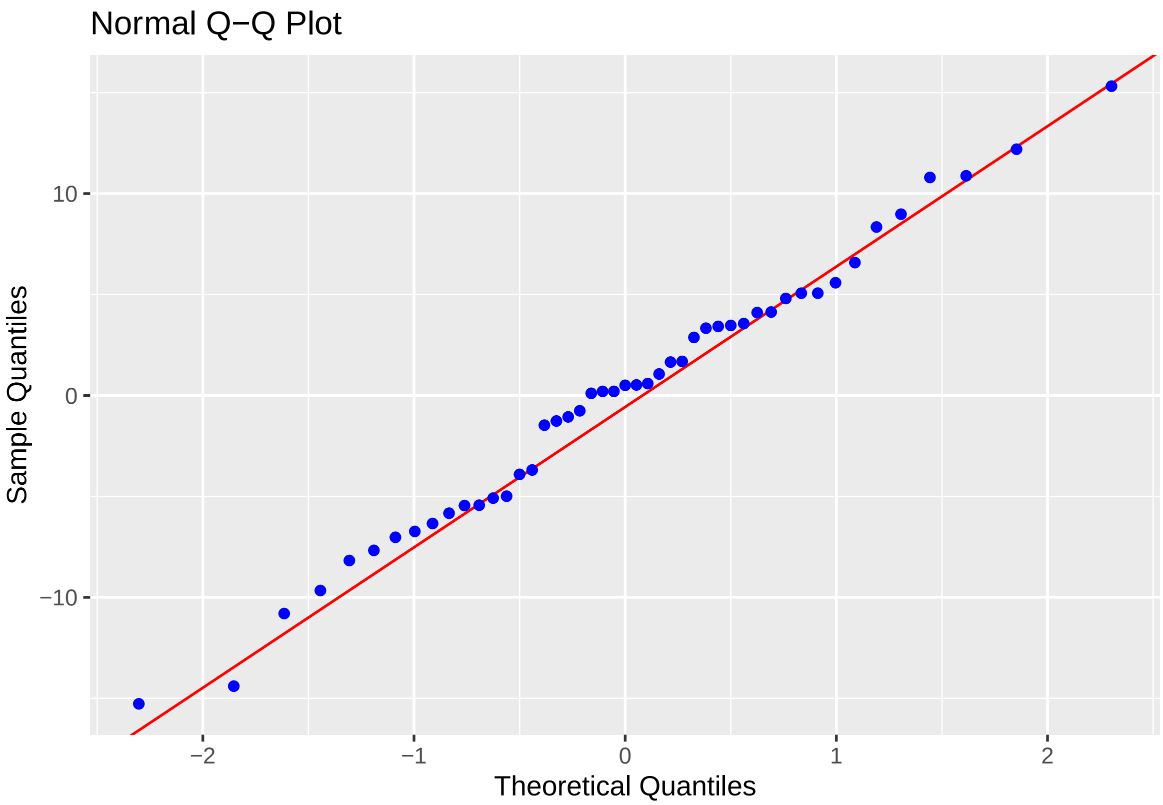
```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 5.8511, df = 5, p-value = 0.321
```

The null and alternative hypotheses are H_0: Homoscedastic errors and H_1: Heteroscedastic errors. The test statistic is : 5.8511 with a p-value of .321. We conclude the constant variance assumption is not violated.

From the graphical test, I would believe that the constant variance assumption is not violated. After performing a hypothesis test, this conclusion is further supported.

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.

```r
ols_plot_resid_qq(model)
```



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

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

The p-value is .9318. We do not reject the null hypothesis and conclude that the errors are normally distributed.

I would conclude that the normality assumption is not violated. There are no obvious tails on the Q-Q plot, and the hypothesis tests support this conclusion.

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

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

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

There is a leverage point in La Vallee at observation 19 and De Geneve at observation 45.

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)
}

# vector of indices for observations deemed outliers.
cutoff = outlier_test_cutoff(model, alpha = 0.05)

which(abs(rstudent(model)) > cutoff)
```

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

There are no outliers.

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

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

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

There is a highly influential point in Porrentruy at observation 6, Sierre at observation 37, Neuchatel Rive at observation 42, Droite Rive at observation 46, and Gauche at observation 47

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

```
coef(model)
```

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

```
# ids for non-influential observations
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
```

There is no significant different in the slopes of both models, except for in examination with the fixed model have twice a downward slope as compared to the standard model.

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)))

# model includes influential observations
predict(model, influential_obs)
```

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

```
# model excludes influential observations
predict(model_fix, influential_obs)
```

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

There is no significant difference in the predictions between the two models.