STA 5207: Homework 5

Due: Wednesday, October 18 by 11:59 PM

Include your R code as an appendix at the end of your homework. Do not include your code in your answers unless the question explicitly tells you to include your code. Your answers to each exercise should be self-contained without code so that the grader can determine your solution without reading your code or deciphering its output.

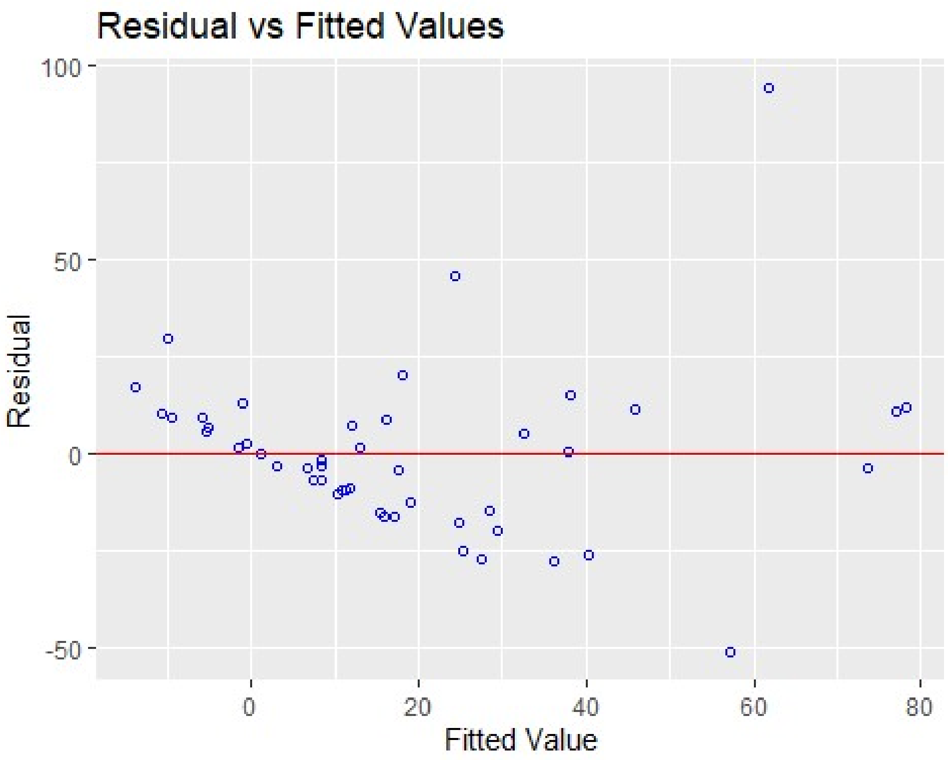
Exercise 1 (Diagnostics for Teenage Gambling Data) [40 points]

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.

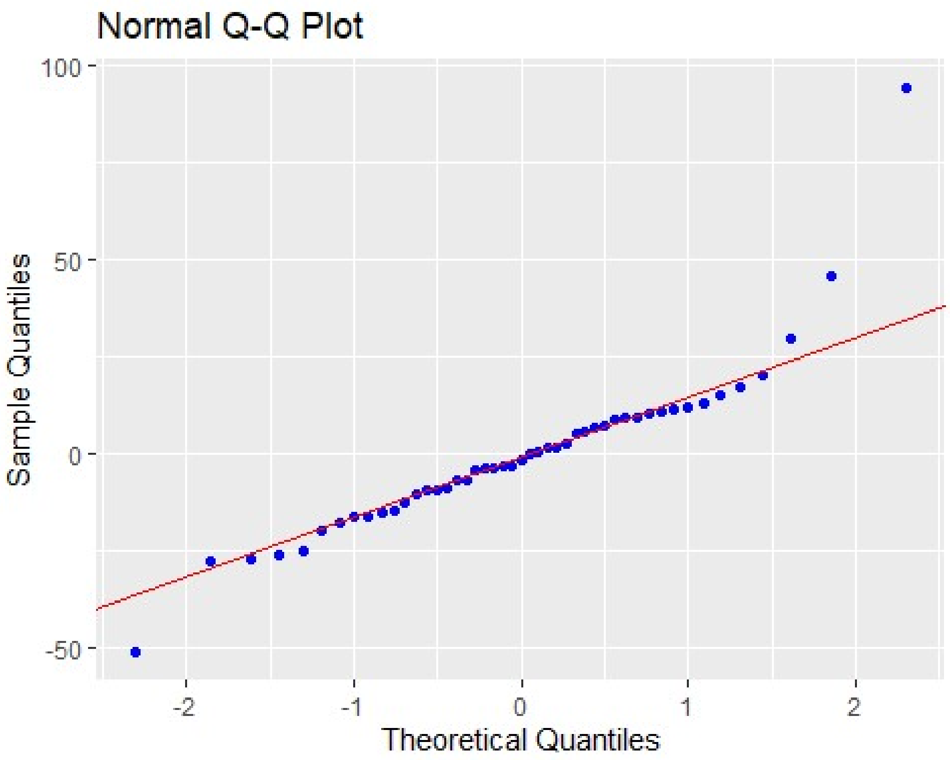
1. (8 points) Check the constant variance assumption for this model using a graphical method and a hypothesis test at the $\alpha = 0.05$ significance level. Do you feel it has been violated? Justify your answer. Include any plots in your response.



–We observe no discernible pattern, and the distribution around zero appears roughly uniform, suggesting that the constant variance assumption for this model remains unchallenged.

–Given that the P-value (0.1693) exceeds the significance level $\alpha = 0.05$, we do not have sufficient evidence to reject the null hypothesis. Consequently, we may conclude that the errors exhibit homoscedasticity, indicating that they have a constant variance.

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.



–I suspect that the normality assumption is violated due to the discrepancies observed at both ends of the plot, despite the majority of the points clustering around the central line.

–Based on the Shapiro test, with a P-value of 0.0000816, we reject the null hypothesis, indicating that the errors do not conform to a normal distribution.

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

–The observations with significant leverage are 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.

– At the significance level of $\alpha = 0.05$, the 24th Observation is identified as the outlier within the dataset.

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

–The 24th and 39th observations are identified as highly influential, with the additional note that the 24th observation is classified as an outlier.

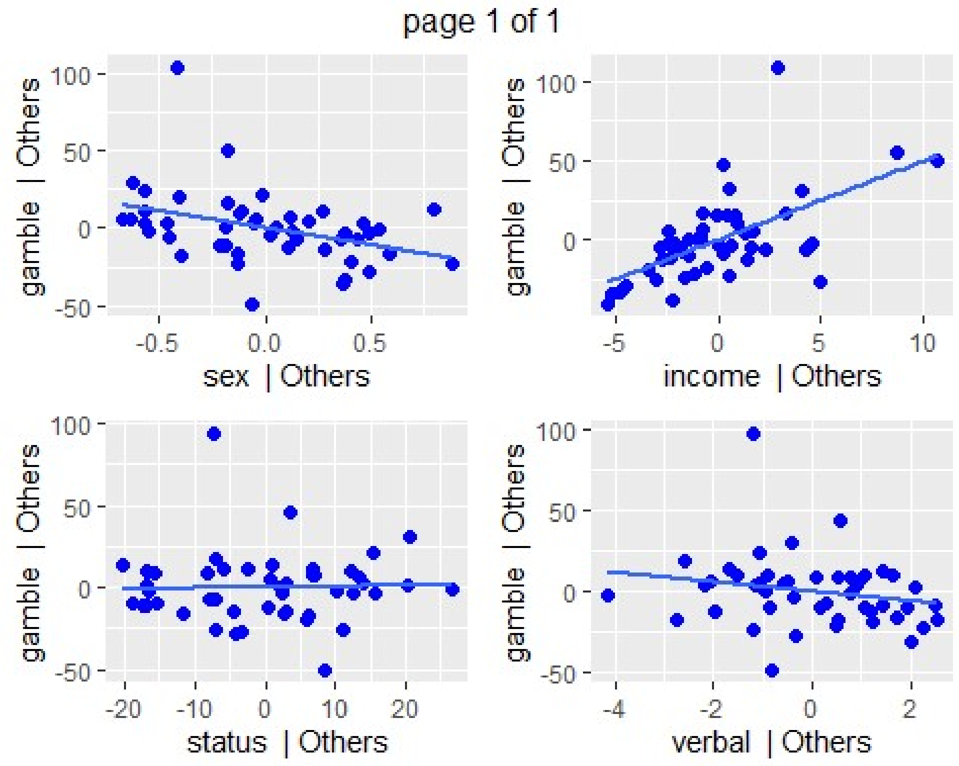
6. (9 points) Fit a model with the high influence points you found in the previous question removed. Perform a hypothesis test at the $\alpha = 0.05$ significance level to check the normality assumption. What do you conclude?

–At the 0.05 significance level, the Shapiro test yields a p-value of 0.23, indicating that the null hypothesis is not rejected. Thus, the errors are assumed to conform to a normal distribution. –Consequently, eliminating influential points is expected to resolve the normality issues observed in the regression.

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.



Based on the observed plots:

– The relationship between gamble and status appears to be non-linear, even after accounting for the effects of other predictors.

– After adjusting for the influences of other predictors, a clear linear relationship is evident between gamble and income.

– Removing the effects of the other predictors reveals a relatively weak linear relationship between gamble and verbal, as well as sex.

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.

- The value of slope parameter is 4.961979.

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?

–The coefficient for `income` in both the model from part 1 and the slope parameter value from part 2 is identical (4.961).

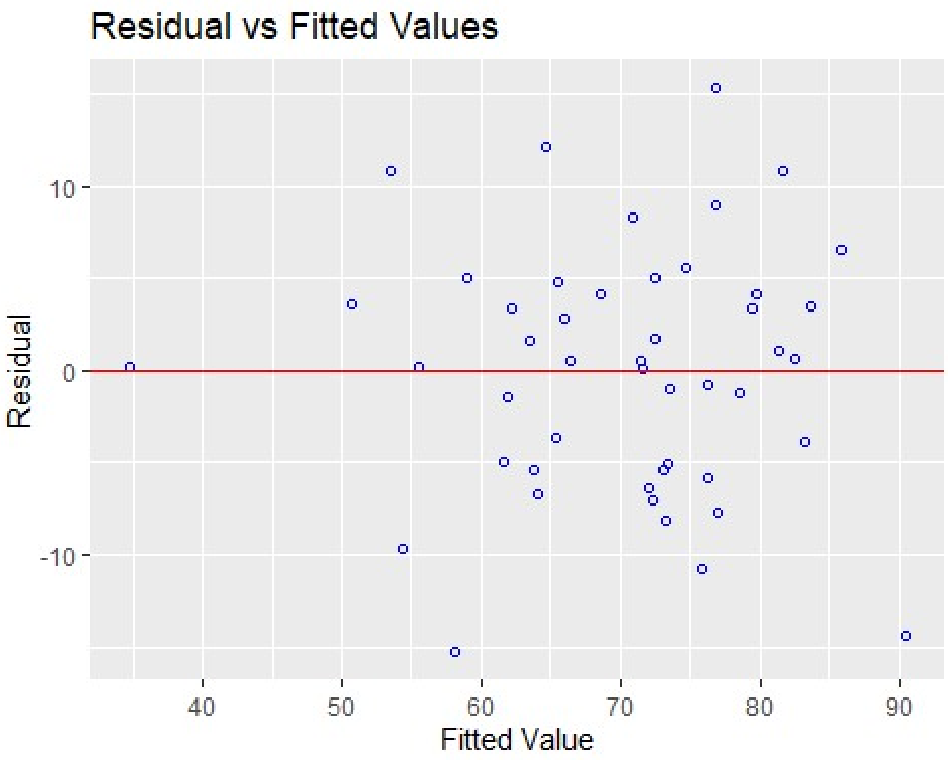
Exercise 3 (Diagnostics for Swiss Fertility Data) [40 points]

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.

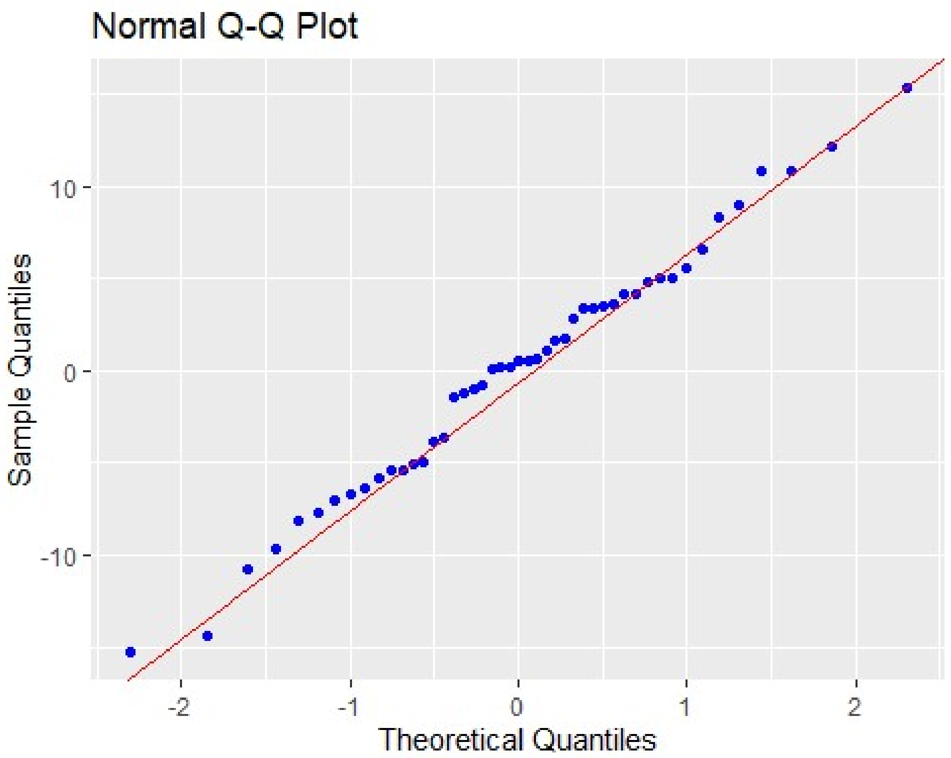
1. (8 points) Check the constant variance assumption for this model using a graphical method and a hypothesis test at the $\alpha = 0.05$ significance level. Do you feel it has been violated? Justify your answer. Include any plots in your response.



–There is no clear discernible pattern, and the distribution around zero appears to be roughly uniform. Thus, the constant variance assumption for this model remains unchallenged.

–Given that the P-value (0.321) exceeds the significance level of $\alpha = 0.05$, we fail to reject the null hypothesis. Consequently, we may infer that the errors exhibit constant variance.

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.



–Given that the majority of the points cluster closely around the line, the Q-Q plots appear to be highly favorable, leading me to conclude that the normality assumption remains intact.

–Since the P-value of the Shapiro test is 0.9318, we fail to reject the null hypothesis, indicating that the errors conform to a normal distribution.

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

– The observations that have high leverage are 19 (La Vallee) and 45 (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.

–All the data points in the data set fall within a consistent range.

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

–The dataset includes five key data points, namely the 6th entry representing Porrentruy, the 37th entry corresponding to Sierre, the 42nd entry denoting Neuchatel, the 46th entry representing Rive Droite, and the 47th entry representing 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.

–The removal of influential observations appears to have had little impact on the regression coefficients.

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.

–After excluding influential data points, the forecast for Sierre remains relatively stable, while the projected fertility measure for Porrentruy increases by approximately 4 units. Additionally, the anticipated fertility levels for Neuchatel, Rive Droite, and Rive Gauche surge by approximately 2, 3, and 3 units, respectively.

Code Appendix

```r
## Solution for Ex 1 Q 1

# dataset load

library(readr)
data(teengamb, package='faraway')


library(olsrr)

cv_model = lm(gamble ~ .,data = teengamb)

ols_plot_resid_fit(cv_model)

## Solution for Ex 1 Q 1

library(lmtest)

bptest(cv_model)


## Solution for Ex 1 Q 2

ols_plot_resid_qq(cv_model)

## Solution for Ex 1 Q 2

shapiro.test(resid(cv_model))

## Solution for Ex 1 Q 3

hatvalues(cv_model)

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

```r
## Solution for Ex 1 Q 4

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

ctof = outlier_cutoff(cv_model, alpha = 0.05)

which(abs(rstudent(cv_model)) > ctof)

## Solution for Ex 1 Q 5

which(cooks.distance(cv_model) > 4 / length(cooks.distance(cv_model)))

## Solution for Ex 1 Q 6

influence_excl=which(cooks.distance(cv_model)<=
                       4/length(cooks.distance(cv_model)))

inflnc_rmv_model = lm(gamble ~ .,data = teengamb,subset = influence_excl)

shapiro.test(resid(inflnc_rmv_model))


## Solution for Ex 2 Q 1


library(readr)
data(teengamb, package='faraway')

var_plts = lm(gamble ~ .,data = teengamb)

library(olsrr)

ols_plot_added_variable(var_plts)


## Solution for Ex 2 Q 2

model_1 = lm(gamble~verbal+status+sex,data = teengamb)

model_2 = lm(income~verbal+status+sex,data = teengamb)

resid(model_1)

resid(model_2)
```

```r
residuals_model = lm (resid(model_1)~resid(model_2))

coef(residuals_model)

## Solution for Ex 2 Q 3


coef(var_plts)


## Solution for Ex 3 Q 1


library(readr)
data(swiss, package='faraway')


library(olsrr)

cv_model = lm(Fertility ~ .,data = swiss)

ols_plot_resid_fit(cv_model)

## Solution for Ex 3 Q 1

library(lmtest)

bptest(cv_model)


## Solution for Ex 3 Q 2

ols_plot_resid_qq(cv_model)

## Solution for Ex 3 Q 2

shapiro.test(resid(cv_model))


## Solution for Ex 3 Q 3

hatvalues(cv_model)

which(hatvalues(cv_model) > 2 * mean(hatvalues(cv_model)))

## Solution for Ex 3 Q 4
```

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

ctof = outlier_cutoff(cv_model, alpha = 0.05)

which(abs(rstudent(cv_model)) > ctof)

## Solution for Ex 3 Q 5

which(cooks.distance(cv_model) > 4 / length(cooks.distance(cv_model)))

## Solution for Ex 3 Q 6

# with influential observations

coef(cv_model)

# without influential observations

non_influence = which(cooks.distance(cv_model) <=
                         4 / length(cooks.distance(cv_model)))

inflnc_rmv_model = lm (Fertility ~ .,data = swiss,subset = non_influence)

coef(inflnc_rmv_model)


## Solution for Ex 3 Q 7

with_influence = subset(swiss,cooks.distance(cv_model) >
                           4 / length(cooks.distance(cv_model)))


predict(cv_model,with_influence)

predict(inflnc_rmv_model,with_influence)
```