STA 5207: Homework 7

Due: Wednesday, November 1 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 (The `divusa` Data Set) [50 points]

For this exercise, we will use the `divusa` data set from the `faraway` package. You can also find the data in `divusa.csv` on Canvas. The data set contains information on divorce rates in the USA from 1920 to 1996. The variables in the data set are

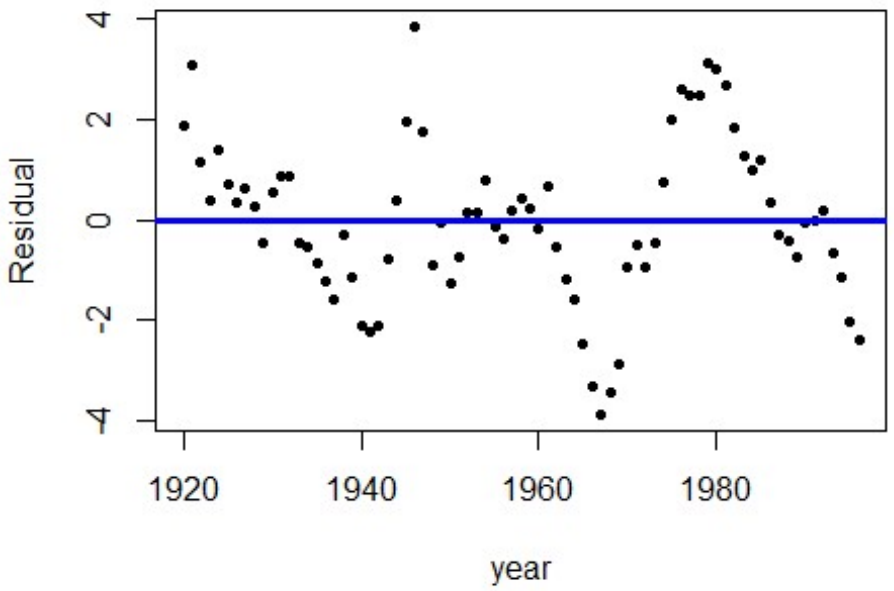
- `year`: the year from 1920-1996.
- `divorce`: divorce per 1000 women aged 15 or more.
- `unemployed`: unemployment rate.
- `femlab`: female participation in labor force aged 16+.
- `marriage`: marriages per 1000 unmarried women aged 16+.
- `birth`: births per 1000 women aged 15-44.
- `military`: military personnel per 1000 population.

In the following exercise, we will model the `divorce` variable in terms of `unemployed`, `femlab`, `marriage`, `birth`, and `military`.

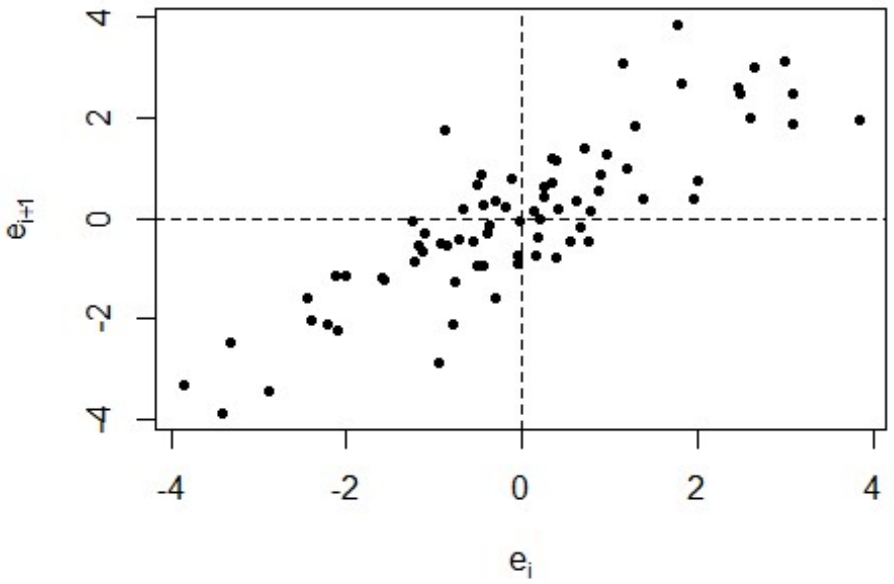
1. (2 points) The variable `year` is not being used in the model, but it shows that the measurements were taken across time. What does this make you suspect about the error term? No output need.

Given that the data measures the same variables over time, it is reasonable to presume that errors are interdependent.

2. (6 points) Fit an OLS regression model with `divorce` as the response and all other variables except `year` as predictors. Check for serial correlation in the errors using a graphical method. Do you feel like the errors are serially correlated? Justify your answer. Include any plots in your response.



The presence of a wavy pattern suggests a clear indication of serial correlation, as opposed to a random scatter of points.



–The presence of a clear linear relationship between successive pairs of residuals in the above plot suggests positive serial correlation, contradicting the assumption of uncorrelated errors.

3. (6 points) Check for the presence of serial correlation in the errors using the Durbin-Watson test. Report the following:
   - The null and alternative hypotheses.
   - The value of the test statistic.
   - The p-value of the test.
   - A statistical decision at the $\alpha = 0.05$ significance level.

- The null hypothesis is -

  $H_0 : \phi = 0$ (uncorrelated errors)

- The alternate hypothesis is -

  $H_1 : \phi \neq 0$ (the errors follow an AR(1) process)

–Given the test statistic value of 0.29988 and a very small p-value of 2.2e-16, the statistical decision at the 0.05 significance level leads to the rejection of the null hypothesis. Consequently, it can be concluded that the errors adhere to an AR(1) process.

4. (10 points) Model the serial correlation with an AR(1) process, meaning that $\Sigma_{ij} = \phi^{|i-j|}$. Use the ML method to estimate the parameters in the GLS fit. Create and report a table with the OLS estimates (model in part 2) and GLS estimates for the slope parameters.

Coefficients	OLS	GLS
unemployed	-0.11125201	0.10764313
femlab	0.38364928	0.31208493
marriage	0.11867431	0.16432630
birth	-0.12995915	-0.04990919
military	-0.02673402	0.01794640

5. (10 points) Perform a t-test at the 5% significance level for each slope parameter for the OLS model in part 2 and the GLS model in part 4. Are there differences between which predictors are significant in the OLS model and which are significant in the GLS model? If so, state the changes.

- According to OLS : femlab,marriage,birth are significant.

GLS indicates that variables such as unemployed, femlab, marriage, and birth are significant. Specifically, while unemployed is deemed significant by GLS, it is not considered as such by OLS.

6. (5 points) For the GLS model in part 4, calculate and report the variance inflation factor (VIF) for each of the predictors using the `vif` function from the `car` package.

Do any of these VIFs suggest we should be cautious about concluding a variable is "not significant" given the other predictors?

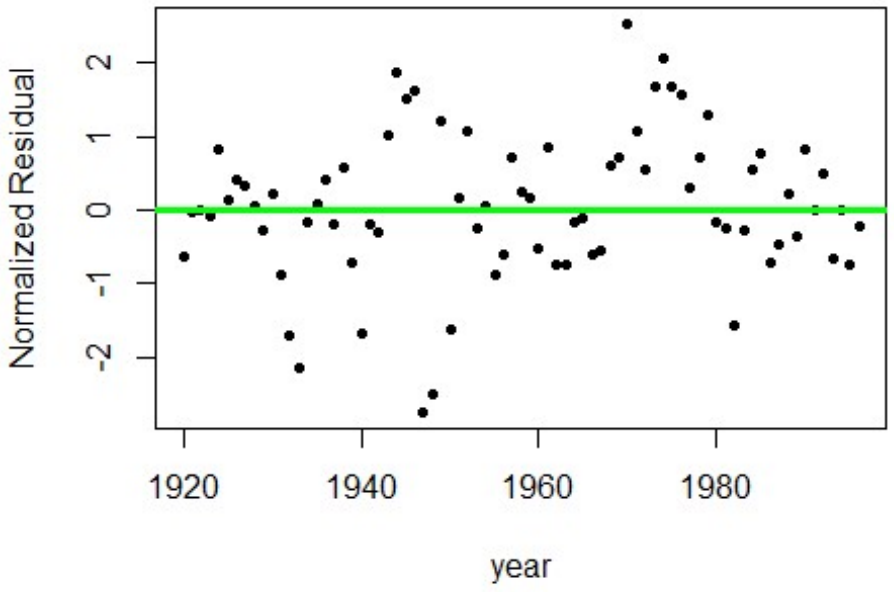
```
## unemployed     femlab   marriage      birth   military
##   1.710203   1.905371   2.624558   1.148642   2.533990
```

–Given that all Variance Inflation Factors (VIFs) are less than 5, there is no cause for concern regarding the conclusion that a variable is "not significant" in light of the other predictors.

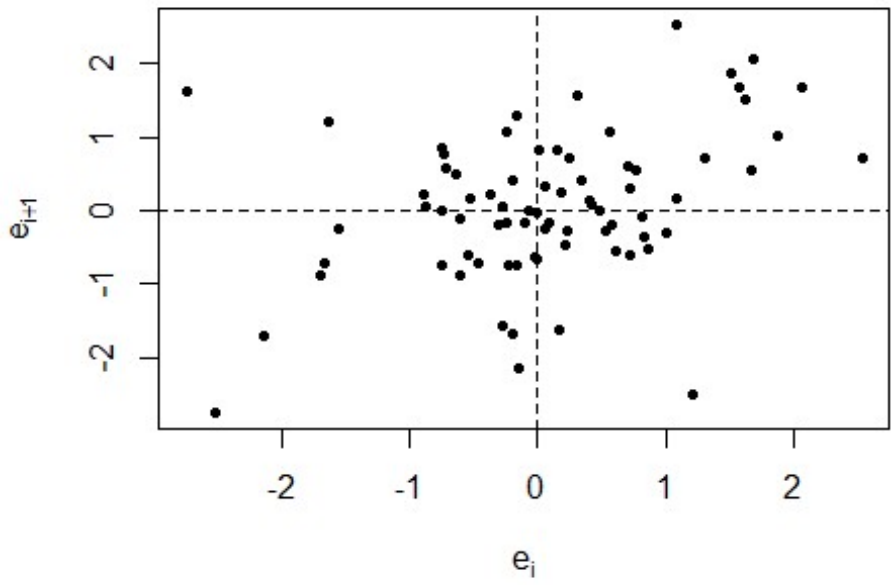
7. (5 points) Report the estimated value of the autocorrelation parameter ϕ and its associated 95% confidence interval. Does the interval indicate that ϕ is significantly different from zero at the 5% significance level?

–The estimated value of phi (ϕ) is 0.9715486, with a 95% confidence interval of (0.6529952, 0.9980179). Since the interval does not include 0, it is inferred that phi (ϕ) is significantly greater than zero at the 5% significance level.

8. (6 points) Check for serial correlation in the normalized errors of the GLS model in part 4 using a graphical method. Do you feel like the normalized errors are serially correlated? Justify your answer. Include any plots in your response.



–The plot has significantly improved, showing a scattered pattern of points around the horizontal line. This suggests that the normalized errors are not correlated sequentially.



–The plot above suggests the absence of a linear relationship, indicating that the normalized errors are not serially correlated.

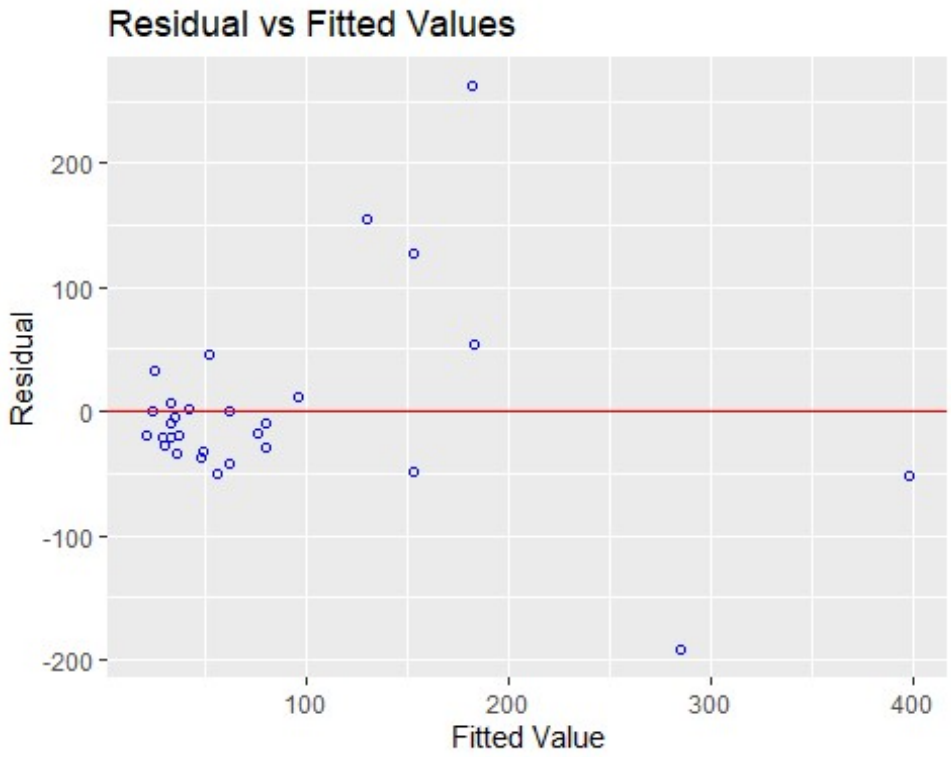
Exercise 2 (The `gala` Data Set) [40 points]

For this exercise, we will use the `gala` data set from the `faraway` package. You can also find the data set in `gala.csv` on Canvas. The data set contains the following variables:

- `Species`: The number of plant species found on the island.
- `Area`: The area of the island (km^2).
- `Elevation`: The highest elevation of the island (m).
- `Nearest`: The distance from the nearest island (km).
- `Scruz`: The distance from Santa Cruz island (km).
- `Adjacent`: The area of the adjacent island (km^2).

In the following exercise, we will model `Species` in terms of `Area`, `Elevation`, and `Nearest`.

1. (5 points) Perform OLS regression with `Species` as the response and `Area`, `Elevation`, and `Nearest` as the predictors. 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.



–The plot depicting the fitted values against the residuals is concerning. The points appear clustered towards the lower end, and the spread widens noticeably as the fitted values increase, indicating a violation of the assumption of constant variance.

–The test statistic is 11.184, leading us to reject the null hypothesis due to the P-value (0.01077) being less than the significance level, set at $= 0.05. Consequently, we can conclude that the errors exhibit heteroscedasticity.

2. (8 points) Perform a regression of the absolute value of the residuals from the model in part 1 against the predictors `Area`, `Elevation`, and `Nearest` using OLS. Report the estimated regression equation using all 3 predictors.

— $|e_i|$ = 5.867 - 0.0361$Area_i$ + 0.1433$Elevation_i$ -0.2557$Nearest_i$

3. (8 points) Perform WLS using the inverse of the squared fitted values from the model in part 2 as weights, i.e, weights $= 1/(\text{fitted values})^2$. Create and report a table with the OLS estimates (model in part 1) and WLS estimates for the slope parameters.

Coefficients	OLS	WLS
Area	0.01908464	0.02237259
Elevation	0.17133627	0.17395271
Nearest	0.07122724	0.40384843

4. (8 points) Perform a t-test at the 5% significance level for each slope parameter for the OLS model in part 1 and the WLS model in part 3. Are there differences between

which predictors are significant in the OLS model and which are significant in the WLS model? If so, state the changes.

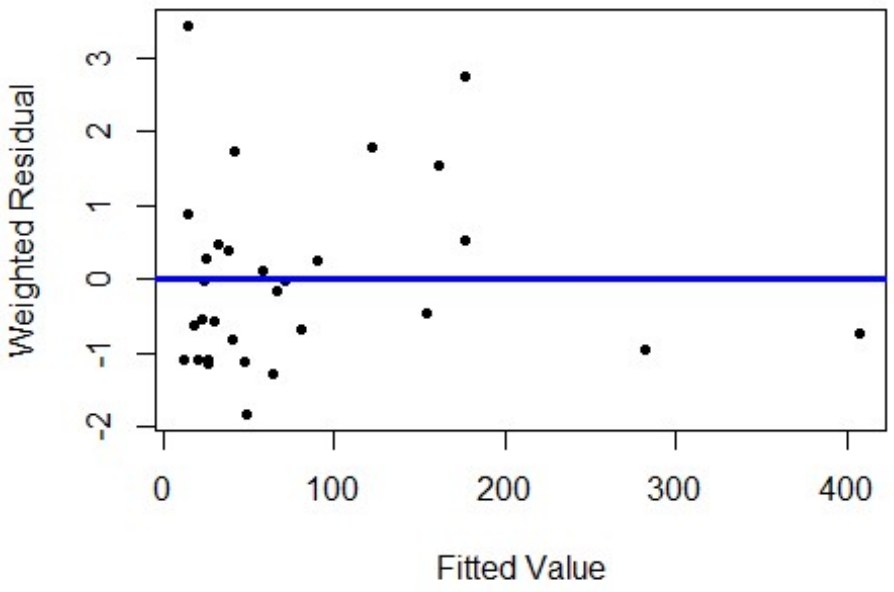
- According to OLS : Only Elevation is significant.
- According to WLS : Elevation and Nearest are significant.
- So Nearest is significant according to WLS but not according to OLS.

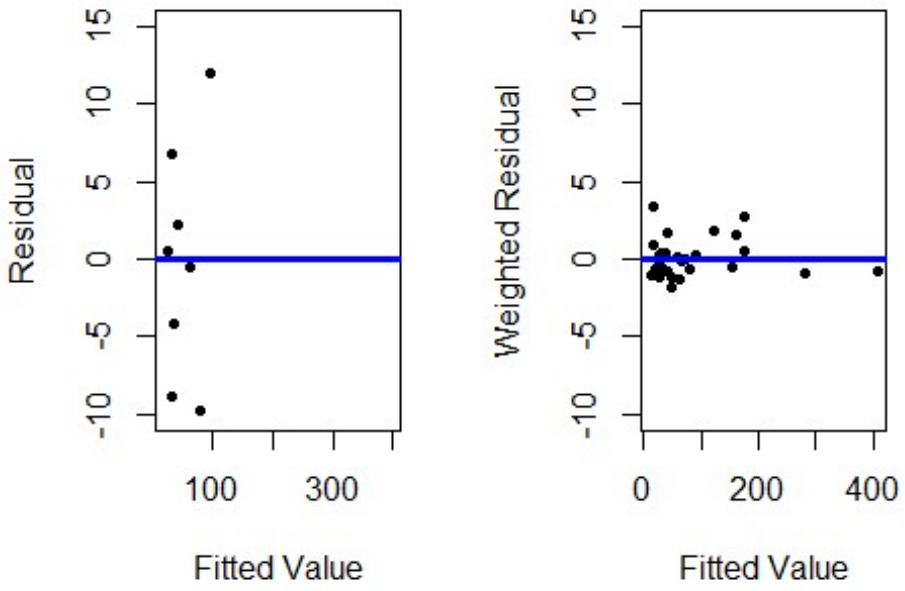
5. (5 points) For the WLS model in part 3, calculate and report the variance inflation factor (VIF) for each of the predictors using the `vif` function from the `car` package. Do any of these VIFs suggest we should be cautious about concluding a variable is "not significant" given the other predictors?

```
##      Area Elevation   Nearest
##  2.154149  2.156878  1.002607
```

- All VIF's are less than 5 , so there is no need to worry about concluding a variable is "not significant" given the other predictor.

6. (6 points) Check the constant variance assumption on the weighted residuals of the WLS model using a a graphical method and a hypothesis test at the $\alpha = 0.05$ significance level. Do you feel that it has been violated? Justify your answer. Include any plots in your response.





–Based on the two plots presented, it appears that the WLS plots exhibit a more favorable pattern, without any discernible trends, and the distribution around zero seems to be relatively consistent. Consequently, it can be inferred that the constant variance assumption for this model remains intact.

–Based on the test statistic of -0.000812, and with the p-value (1) exceeding the significance level of $\alpha = 0.05$, we fail to reject the null hypothesis, thus concluding that the errors are homoscedastic.

Exercise 3 (WLS for Survey Data) [10 points]

For this exercise, we will use the the `chibus` data set, which can be found in `chibus.csv` on Canvas. Each observation in this data set represents a pair of zones in the city of Chicago. The variables in the data set are

- `computed_time`: travel times, computed from bus timetables augmented by walk times from zone centers to bus-stops (assuming a walking speed of 3 mph) and expected waiting times for the bus (= half of the time between successive buses).
- `perceived_time`: average travel times as reported to the U.S. Census Bureau by n travelers.
- `n`: number of travelers per observations for each case.

In the following exercise, we will model `perceived_time` in terms of `computed_time`.

1. (5 points) The variable n is not being used in the model, but it shows that the response is recorded as an average over different groups of size n_i. Based on this observation, what would make for a good choice of weights? No output is needed.

– In above case we can choose weight as $w_i = n_i$.

2. (5 points) Perform WLS with `perceived_time` as the response and `computed_time` as the predictor using the weights you chose in part 1. Report the estimated regression equation for this model.

– $\widehat{perceived\_time} = 2.293 + 1.131 computed\_time_i$

Code Appendix

```
## Solution for Exercise 1 Question 2

# Loading dataset

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


#fitting OLS model
ols_mod = lm(divorce~.-year,data=divusa)

# plotting fitted vs year

plot(resid(ols_mod) ~ year, data = divusa, pch = 20,
     xlab = 'year', ylab = 'Residual')
abline(h=0, lwd=3, col='blue')



## Solution for Exercise 1 Question 2


# plotting pairs of residuals

n = length(resid(ols_mod))
plot(tail(resid(ols_mod), n-1), head(resid(ols_mod), n-1), pch = 20,
     xlab=expression(e[i]), ylab=expression(e[i+1]))

# lines at the x and y axes
abline(h=0, v=0, lty='dashed')


## Solution for Exercise 1 Question 3

library(lmtest)
```

```r
dwtest(ols_mod,alternative = 'two.sided')


## Solution for Exercise 1 Question 4

library(nlme)

gls_mod = gls(divorce ~ . - year,
              correlation = corAR1(form = ~ year),
              method = 'ML', data = divusa)

gls_mod$coefficients


## Solution for Exercise 1 Question 4


ols_mod$coefficients


## Solution for Exercise 1 Question 5

summary(ols_mod)


## Solution for Exercise 1 Question 5

library(nlme)

gls_mod = gls(divorce ~ . - year,
              correlation = corAR1(form = ~ year),
              method = 'ML', data = divusa)

summary(gls_mod)


## Solution for Exercise 1 Question 6

library(car)
library(nlme)

gls_mod = gls(divorce ~ . - year,
              correlation = corAR1(form = ~ year),
              method = 'ML', data = divusa)

car::vif(gls_mod)
```

```r
## Solution for Exercise 1 Question 7

library(nlme)
gls_mod = gls(divorce ~ . - year,
                correlation = corAR1(form = ~ year),
                method = 'ML', data = divusa)

intervals(gls_mod)


## Solution for Exercise 1 Question 8

library(nlme)
gls_mod = gls(divorce ~ . - year,
                correlation = corAR1(form = ~ year),
                method = 'ML', data = divusa)

plot(resid(gls_mod, type = 'normalized') ~ year, data = divusa, pch = 20,
     xlab = 'year', ylab = 'Normalized Residual')

abline(h = 0, lwd = 3, col = 'green')



library(nlme)
gls_mod = gls(divorce ~ . - year,
                correlation = corAR1(form = ~ year),
                method = 'ML', data = divusa)

n = length(resid(gls_mod))
plot(tail(resid(gls_mod, type = 'normalized'), n-1),
     head(resid(gls_mod, type = 'normalized'), n-1),
     pch = 20, xlab=expression(e[i]), ylab=expression(e[i+1]))

# lines at the x and y axes
abline(h=0, v=0, lty='dashed')

## Solution for Exercise 2 Question 1

# Loading dataset

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

library(olsrr)

con_var_model = lm(Species ~ Area+Elevation+Nearest,data = gala)
```

```r
ols_plot_resid_fit(con_var_model)

## Solution for Exercise 2 Question 1

library(lmtest)

bptest(con_var_model)


## Solution for Exercise 2 Question 2

wts_mod = lm(abs(resid(con_var_model))~Area+Elevation+Nearest-
Species,data=gala)

coef(wts_mod)


## Solution for Exercise 2 Question 3

wts_mod = lm(abs(resid(con_var_model))~Area+Elevation+Nearest-
Species,data=gala)

weights = 1 / fitted(wts_mod)^2

sqr_wls_mod = lm(Species ~ Area+Elevation+Nearest, data = gala, weights =
weights)

sqr_wls_mod$coefficients


## Solution for Exercise 2 Question 3

con_var_model$coefficients


## Solution for Exercise 2 Question 4

summary(con_var_model)


## Solution for Exercise 2 Question 4

wts_mod = lm(abs(resid(con_var_model))~Area+Elevation+Nearest-
Species,data=gala)

weights = 1 / fitted(wts_mod)^2
```

```r
sqr_wls_mod = lm(Species ~ Area+Elevation+Nearest, data = gala, weights =
weights)


summary(sqr_wls_mod)


## Solution for Exercise 2 Question 5

library(car)

wts_mod = lm(abs(resid(con_var_model))~Area+Elevation+Nearest-
Species,data=gala)

weights = 1 / fitted(wts_mod)^2

sqr_wls_mod = lm(Species ~ Area+Elevation+Nearest, data = gala, weights =
weights)

car::vif(sqr_wls_mod)


## Solution for Exercise 2 Question 6

wts_mod = lm(abs(resid(con_var_model))~Area+Elevation+Nearest-
Species,data=gala)

weights = 1 / fitted(wts_mod)^2

sqr_wls_mod = lm(Species ~ Area+Elevation+Nearest, data = gala, weights =
weights)

plot(fitted(sqr_wls_mod), weighted.residuals(sqr_wls_mod),
     pch = 20, xlab = 'Fitted Value', ylab = 'Weighted Residual')

abline(h=0, lwd=3, col='blue')


## Solution for Exercise 2 Question 6

par(mfrow = c(1, 2))

# OLS fitted-vs-residual plot
plot(fitted(con_var_model), resid(con_var_model),
     pch = 20, ylim = c(-10, 15),
     xlab = 'Fitted Value', ylab = 'Residual')

abline(h=0, lwd=3, col='blue')
```

```r
# WLS fitted-vs-residual plot

wts_mod = lm(abs(resid(con_var_model))~Area+Elevation+Nearest-
Species,data=gala)

weights = 1 / fitted(wts_mod)^2

sqr_wls_mod = lm(Species ~ Area+Elevation+Nearest, data = gala, weights =
weights)

plot(fitted(sqr_wls_mod), weighted.residuals(sqr_wls_mod),
     pch = 20, ylim = c(-10, 15),
     xlab = 'Fitted Value', ylab = 'Weighted Residual')

abline(h=0, lwd=3, col='blue')


## Solution for Exercise 2 Question 6

library(lmtest)

wts_mod = lm(abs(resid(con_var_model))~Area+Elevation+Nearest-
Species,data=gala)

weights = 1 / fitted(wts_mod)^2

sqr_wls_mod = lm(Species ~ Area+Elevation+Nearest, data = gala, weights =
weights)

bptest(sqr_wls_mod)


## Solution for Exercise 3 Question 2

# Loading dataset

library(readr)
df = read.csv("~/Documents/Applied_Regression_F2023/HomeWork_7/chibus.csv")

wls_mod = lm(perceived_time ~ computed_time, data = df, weights = n)

wls_mod$coefficients
```