Homework 4

Week 4 Model Assumptions

NAME

DATE

Contents

The data for this homework are from the study below:

Kim, S. E., Kim, H. N., Cho, J., Kwon, M. J., Chang, Y., et al. (2016) Correction: Direct and indirect effects of five factor personality and gender on depressive symptoms mediated by perceived stress. *PLOS ONE, 11*: e0157204.

The `hw4data.csv` file contains several variables, but the ones you'll need in this assignment are

- `Stress`: Total perceived stress score from self-reported stress questionnaire
- `A`: Total score on Agreeableness from the Revised NEO Personality Inventory

Question 1

Part a) [0.5 pts]

Read in the dataset and run an Ordinary Least Squares (OLS) multiple regression analysis using Agreeableness (`A`) to predict Perceived Stress (`Stress`). Write the regression equation below.

```r
# Code
hw4data <- readr::read_csv("hw4data.csv")

mod1 <- lm(Stress ~ A, data = hw4data)
summary(mod1)
```

```
##
## Call:
## lm(formula = Stress ~ A, data = hw4data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.740  -4.990  -1.496   3.233  26.632
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.067980   0.477069  46.257   <2e-16 ***
## A           -0.098510   0.009885  -9.965   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6.74 on 3948 degrees of freedom
## Multiple R-squared:  0.02454,    Adjusted R-squared:  0.02429
## F-statistic: 99.31 on 1 and 3948 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{Stress}}_i = 22.07 - 0.099 \times \text{Agreeableness}_i$$

Part b) [0.5 pts]

Redo the regression analysis, using `A` to predict *log-transformed* `Stress`. Write the regression equation in the space below.

```r
# Code
hw4data$log_stress <- log(hw4data$Stress)

mod1b <- lm(log_stress ~ A, data = hw4data)
summary(mod1b)
```

```
##
## Call:
## lm(formula = log_stress ~ A, data = hw4data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77855 -0.26771 -0.02222  0.23780  1.03102
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.050783   0.025625  119.05   <2e-16 ***
## A           -0.005564   0.000531  -10.48   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.362 on 3948 degrees of freedom
## Multiple R-squared:  0.02706,    Adjusted R-squared:  0.02682
## F-statistic: 109.8 on 1 and 3948 DF,  p-value: < 2.2e-16
```

$$\widehat{\log(\text{Stress}_i)} = 3.05 - 0.0056 \times \text{Agreeableness}_i$$

Part c) [1 pts]

Re-run the regression model from part a (`Stress` predicted by `A`), but this time using Weighted Least Squares (WLS) regression using inverse-variance weights.

```r
# Code
abs_resids <- abs(residuals(mod1))
preds <- fitted(mod1)
wmod <- lm(abs_resids ~ preds)

variances <- fitted(wmod)^2
weights <- 1/variances

wls_mod1 <- lm(Stress ~ A, data = hw4data, weights = weights)
summary(wls_mod1)
```

```
##
## Call:
## lm(formula = Stress ~ A, data = hw4data, weights = weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8886 -0.9462 -0.2876  0.6139  5.3950
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.988265   0.491348   44.75   <2e-16 ***
## A           -0.096826   0.009931   -9.75   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.289 on 3948 degrees of freedom
## Multiple R-squared:  0.02351,    Adjusted R-squared:  0.02326
## F-statistic: 95.06 on 1 and 3948 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{Stress}_i} = 21.99 - 0.097 \times \text{Agreeableness}_i$$

Part d) [1 pt]

Create a scatter plot with `A` on the x-axis and `Stress` on the y-axis. Add the following fit lines to the graphs

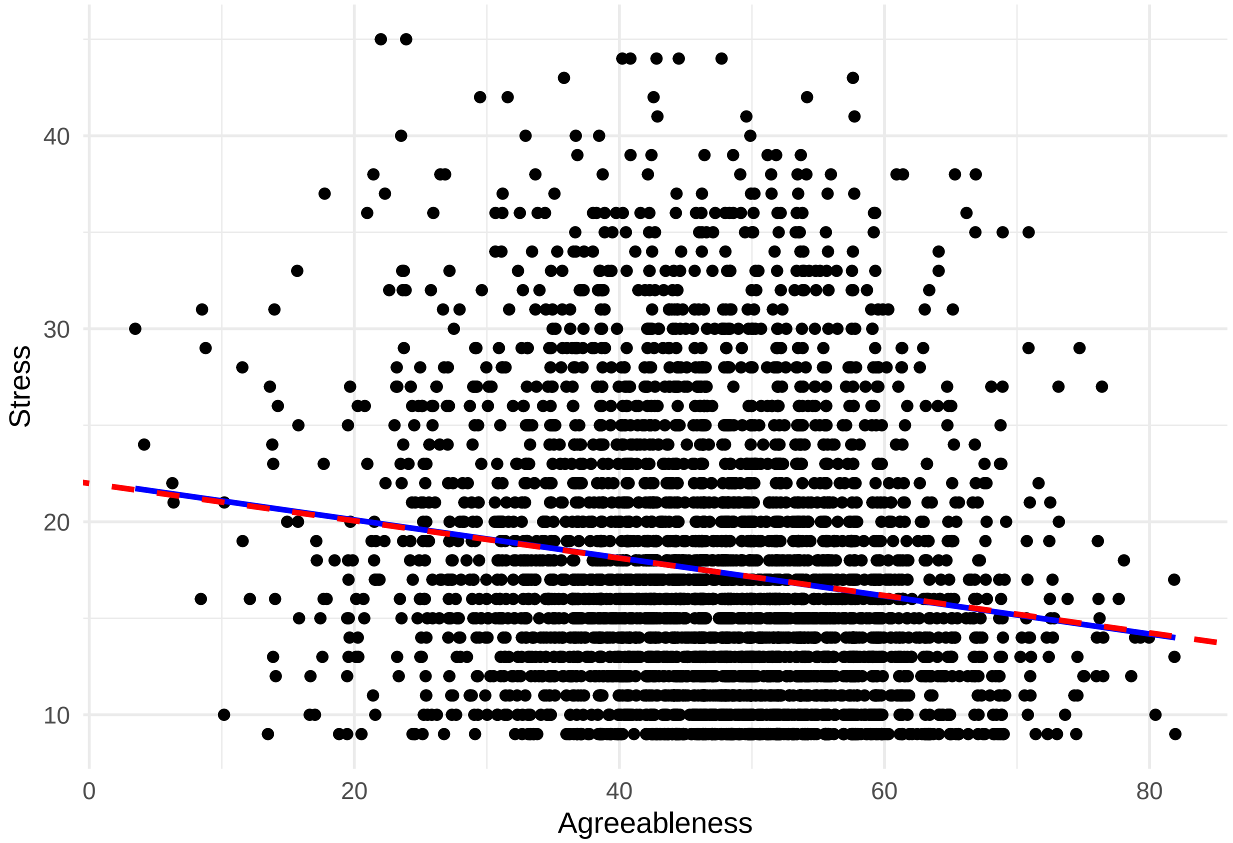
- A fit line based on the OLS regression
- A fit line on the WLS regression.

Make sure the fit lines have distinct colors

```r
# Code

hw4data$ols_fitted <- predict(mod1, hw4data)
hw4data$wls_fitted <- predict(wls_mod1, hw4data)

hw4data |>
  ggplot(aes(x = A, y = Stress)) +
  geom_point() +
  geom_smooth(method = "lm", se = F, fullrange = T, color = "blue") +
  geom_abline(intercept = wls_mod1$coefficients[1],
              slope = wls_mod1$coefficients[2],
              color = "red", size = 1, linetype = "dashed") +
  labs(x = "Agreeableness", y = "Stress") +
  theme_minimal()
```



Question 2

Part a) [1 pt]

For the three models you fit in Question 1:

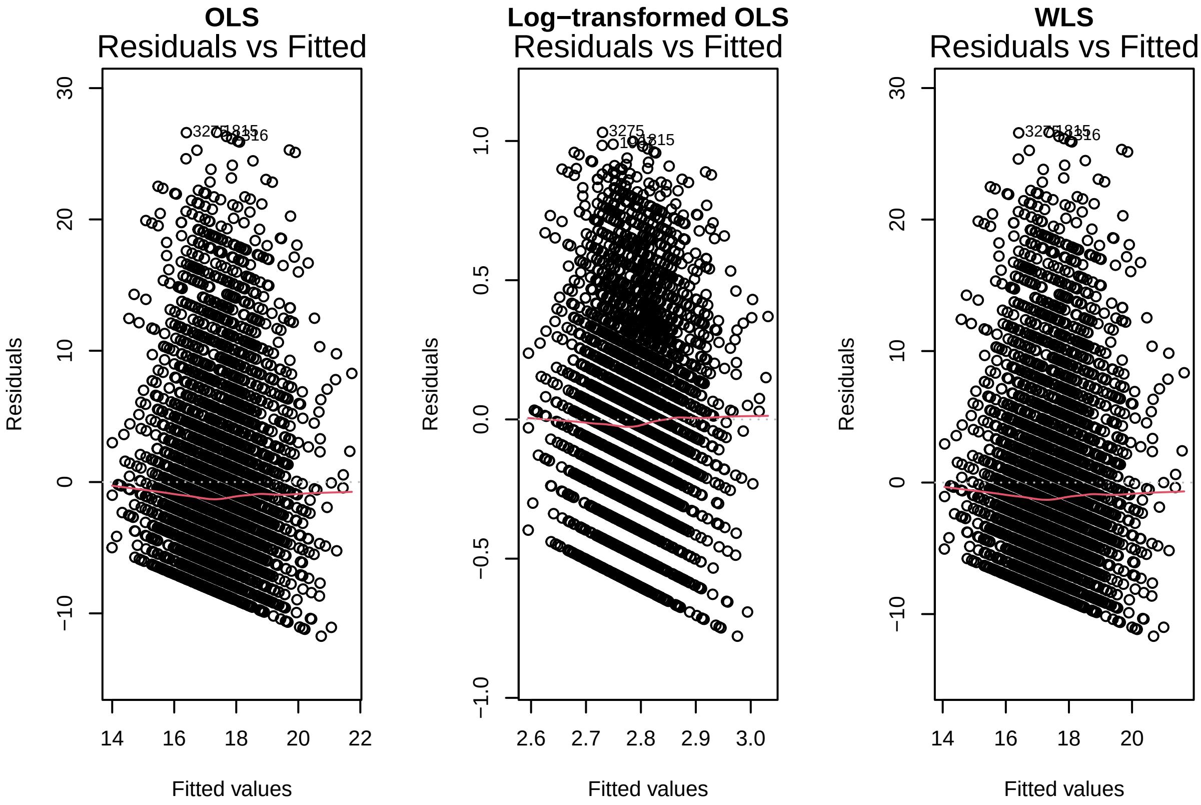
i. Create a side-by-side-by-side graph of the Fitted vs Residual plots
ii. Create a side-by-side-by-side graph of the Residuals Q-Q plot

For each graph, comment on what you see.

Hint: use the `par()` function with the `mfrow` argument like we did in lab.

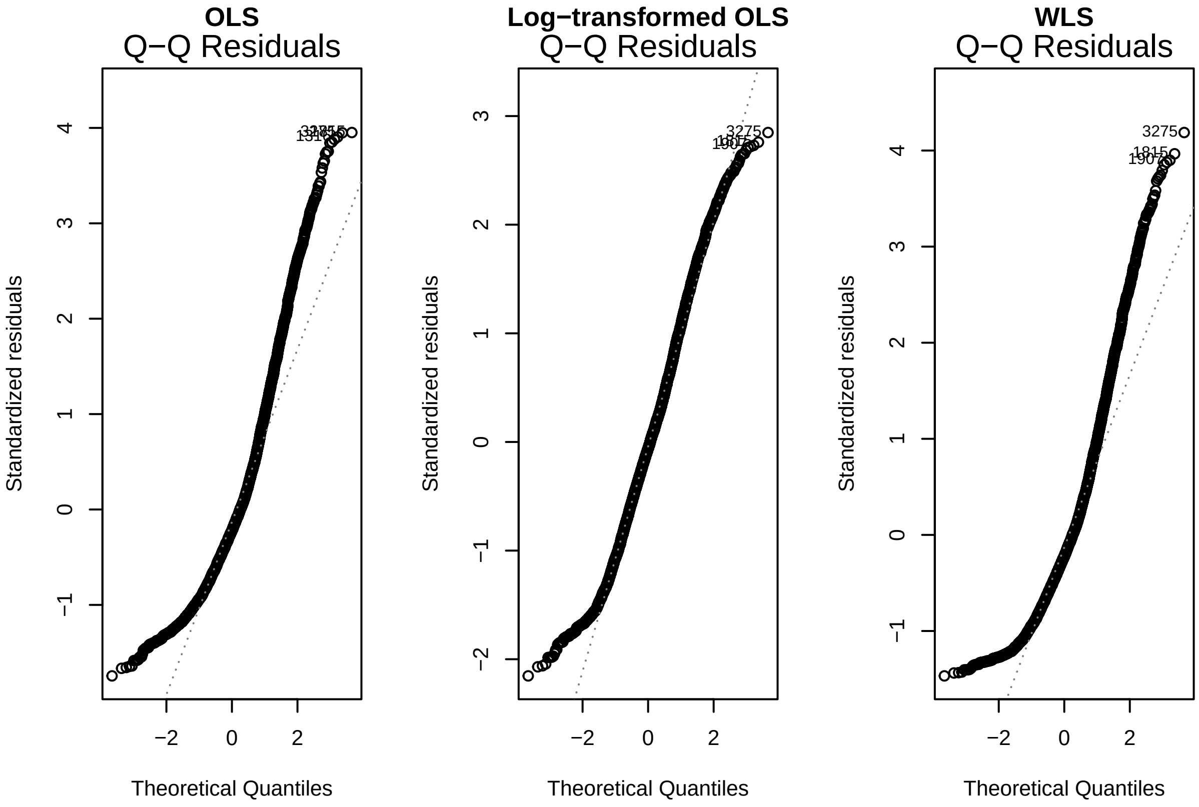
```r
# plot 1

par(mfrow = c(1, 3))
plot(mod1, which = 1, main = "OLS")
plot(mod1b, which = 1, main = "Log-transformed OLS")
plot(wls_mod1, which = 1, main = "WLS")
```



- The model in which Stress is log-transformed may be preferred because of its more horizontal straight line in the Residuals vs. Fitted plot. Deviations found in the other two models indicate potential non-linearity.

```r
# plot 2
par(mfrow = c(1, 3))
plot(mod1, which = 2, main = "OLS")
plot(mod1b, which = 2, main = "Log-transformed OLS")
plot(wls_mod1, which = 2, main = "WLS")
```



- Similar to the previous answer, the plot of the log-transformed OLS model improves the assumption of normality of residuals as they fall approximately along a straight diagonal line. The standard OLS and WLS models display heavier tails.

Part b) [0.5 pts]

Perform White's Test for homogeneity of variance for the original untransformed OLS model, for the OLS model with log-transformed `Stress`, and for the WLS model. Which, if any, model meets this assumption the best?

```r
# Code
models <- list("ols" = mod1, "log_ols" = mod1b, "wls" = wls_mod1)

purrr::map_dfr(models,  white, .id = "model") |>
  kableExtra::kbl(longtable = T, booktabs = T)
```

model	statistic	p.value	parameter	method	alternative
ols	9.917917	0.0070202	2	White's Test	greater
log_ols	2.265623	0.3221263	2	White's Test	greater
wls	9.945442	0.0069243	2	White's Test	greater

- According to the results of White's Test, we can reject the null hypothesis of homogeneity of variance for the standard OLS and WLS models, but not for the log-transformed OLS model. So, for the log-transformed OLS model, White's test did not indicate heteroscedasticity $\chi^2(2) = 2.27$, $p = 0.322$.

Part c) [0.5 pts]

Perform Shapiro-Wilk tests on the residuals for the original untransformed OLS model, for the OLS model with *log-transformed* `Stress`, and for the WLS model. Which, if any, model meets this assumption the best?

```r
# Code
residuals_list <- list()

for (m in 1:length(models)) {

  residuals_list[[names(models[m])]] <- as.vector(models[[m]]$residuals)

}

purrr::map_dfr(residuals_list, ~broom::tidy(shapiro.test(.x)), .id = "model") |>
  kableExtra::kbl(longtable = T, booktabs = T)
```

model	statistic	p.value	method
ols	0.9198791	0	Shapiro-Wilk normality test
log_ols	0.9838072	0	Shapiro-Wilk normality test
wls	0.9196608	0	Shapiro-Wilk normality test

- None of the models meet the assumption of normality of residuals according to the results obtained with the Shapiro-Wilk test. For all three tests we rejected the null hypothesis that the residuals were normally distributed.

Question 3 [5 pts]

Think about the data that you work with in your own research (i.e., some predictor(s) and outcome(s)), and respond to the following questions regarding assumptions in linear regression. Make sure to explain your reasoning. One to two sentences for each question is ok, but please try not to write more than four.

1. Validity: Are your data appropriate in helping you answer your research question(s)?

   - School type can be a valid but simplistic predictor of academic achievement. Many other factors could confound the relationship and need to be considered.

2. Linearity and additivity: Do you think the relationship between your predictor(s) and outcome variable(s) can adequately be described by a linear and additive relationship?

   - The relationship between school type and academic achievement is unlikely to be strictly linear. School type is a categorical variable, so we'll have to use techniques like dummy coding within the regression model, which assumes that the effect of each school type on achievement is distinct.

3. Independence of errors: would you expect observations to be independent or dependent (i.e., correlated)?

   - Observations could be dependent, especially because students are nested within schools. Hierarchical or mixed-effects modeling might be more appropriate to account for the clustering of students within schools.

4. Equal residual variances: would it be reasonable to expect homoscedasticity ?

    - Heteroscedasticity is possible since variability in academic achievement might differ across school types.

5. Normality: could you reasonably expect your residuals to be normally distributed?

    - Standardized tests often approximate normality, so normality of residuals is expected to hold.