```
---
title: "Problem Set 8"
subtitle: "Due date: 2 December"
name: "Clare Porter"
format:
  html:
    self-contained: true
toc: true
editor: visual
execute:
  echo: true
---
```

Please upload your completed assignment to the ELMs course site (under the assignments menu). Remember to include an annotated script file for all work with R and show your math for all other problems (if applicable, or necessary). Please also upload your completed assignment to the Github repository that you have shared with us. *We should be able to run your script with no errors.*

Total points: 30

```{r}
library(ggplot2)
library(dplyr)
library(broom)
library(knitr)
```

Question 1

Points: 5

For the following regression equation, $\hat{Y} = 8.5 + 6x + \epsilon$, the standard error for β_0 is 2.5, the standard error for β_1 is 3.5, and the sample size is 2000. Find the t-statistic, 95% confidence interval, and p-value (using a two-tailed test) for β_1.

Is β_1 statistically significant at the 0.05-level with a two-tailed test? Why or why not?

No, B1 is not statistically significant at the 0.05-level with a two-tailed test. We conclude B1 is not statistically significant given that the t-statistic (1.71) is less than the critical value of 1.96, the p-value (0.08) is greater than the 0.05 significance level, and the 95% confidence interval (-0.86, 12.86) includes 0.

```{r}
# input values
b0 <- 8.5
b1 <- 6
b0_SE <- 2.5
b1_SE <- 3.5
dof <- 2000-2
# t-statistic =  B1/SEB1 = 6/3.5 = 1.714
t_stat <- b1 / b1_SE
t_stat
# calculate t-statistic at the 95% level
t_stat_95 <- qt(0.025, df = dof, lower.tail = F)
t_stat_95
# calculate 95% CI
CI_upper <- b1 + b1_SE*t_stat_95
CI_lower <- b1 - b1_SE*t_stat_95
c(CI_lower, CI_upper)
# p-value
```

```
p_value <- 2 * (1 - pt(abs(t_stat), ss))
p_value
```

Question 2

Points: 5

Suppose you estimate an OLS regression and retrieve a R^2 value of 0.45. If the Total Sum of Squares (TSS) from that regression equals 4,700, what is the value for the Residual Sum of Squares (RSS)?

Given an R\^2 of 0.45 and a Total Sum of Squares from that regression of 4,700, the Residual Sum of Squares is 2,585.

```{r}
r2 <- 0.45
tsos <- 4700
# r2 = 1 - rss/tsos
rss <- (1-r2)*tsos
rss
```

Question 3

Points: 5

Suppose you estimate a bivariate regression with a sample size of 102 and obtain a regression coefficient (β_1) of 5.0. What is the largest standard error that β_1 could have and still be statistically significant (i.e., reject the null hypothesis of no relationship) at the 0.05 level with a one-tailed test?

The largest standard error that a B1 of 5.0 could have and still be statistically significant at the 0.05 level with a one-tailed test is 3.01.

```{r}
# input values
b1 <- 5
dof <- 102-2
# calculate t-statistic at the 95% level
t_stat_95 <- qt(0.05, df = dof, lower.tail = F)
t_stat_95
# max SE
# t_stat = b1 / b1_SE
b1_SE <- b1/t_stat_95
b1_SE
```

Question 4

Points: 5

```{r}
gapminder_df <- wbstats::wb_data(c("NY.GDP.PCAP.CD", "SP.DYN.LE00.IN"),
                                 start_date = 2019, end_date = 2019) |>
  dplyr::rename(gdp_per_cap = "NY.GDP.PCAP.CD", life_exp = "SP.DYN.LE00.IN")
```

Using the `gapminder_df` data set read in above, produce a scatterplot of the variables `gdp_per_cap` and `life_exp` (with `life_exp` as the dependent variable on the y-axis). Fit a regression line to the scatterplot. Describe the relationship illustrated. Note any suspected outliers, if any (a visual inspection will suffice for this question).

::: callout-note

The variable `gdp_per_cap` measures each country's GDP per capita (representing their individual wealth), and `life_exp` indicates the number of years individuals within that country born that year are expected to live (representing their health).
:::

The initial scatter plot of life expectance and GDP per capita shows a positive relationship. However, the relationship does not appear very linear. By taking the log of GDP per capita, the scatter plot shows a more linear positive relationship: life expectancy increases as the log of GDP per capita increases. When visually inspecting the transformed data, we can see at least one clear outlier which has extremely low life expectancy and low logged GDP per capita (around 31 years in life expectancy, and just over 6 in logged gdp per capita).

```{r}
# plot the data
ggplot(gapminder_df, aes(x = gdp_per_cap, y = life_exp)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_smooth(method = "lm", color = "red") +
  labs(x = "GDP per Capita", y = "Life Expectancy",
       title = "Scatterplot of Life Expectancy vs GDP per Capita") +
  theme_minimal()

# this data does not look like it has a very linear relationship, so lets take the log of gdp per cap
gapminder_df <- gapminder_df %>%
  mutate(log_gdp = log(gdp_per_cap))

ggplot(gapminder_df, aes(x = log_gdp, y = life_exp)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_smooth(method = "lm", color = "red") +
  labs(x = "Log of GDP per Capita", y = "Life Expectancy",
       title = "Scatterplot of Life Expectancy vs Log GDP per Capita") +
  theme_minimal()
```

Question 5

Points: 10

Estimate a bivariate regression with `life_exp` as the dependent variable and `gdp_per_cap` as the independent variable and report the results in a professionally formatted table. In as much detail as possible, describe (and interpret) the regression results.

The bivariate regression of the gapminder data, with `life_exp` as the dependent variable and `gdp_per_cap` as the independent variable has an intercept (B0) of 67.739 and a slope for GDP per capita of 0.0000161. This means that the estimated life expectancy when GDP per capita is zero is about 70 years old, and that for every \$10,000 increase in GDP per capita about 2 months in estimated life expectancy is gained. The t-statistics for both estimates are very large, and the p-values extremely small, indicating that GPD per capita is a highly statistically significant predictor of life expectancy, and there is a positive relationship between the two.

Since we determined in Q4 that a linear relationship better explains `life_exp` as the dependent variable and `log_gdp` as the independent variable, we should instead fit the bivariate regression to a logged GDP per capita. Using the code below, we can find now that the intercept is 32.666 and the slope for `log_gdp` is 4.524. The t-statistics remain very high (17.209 and 21.582) and the p-values very low (0 and 0), again confirming that the relationship is highly statistically significant and that we can reject the null hypothesis of life expectancy being unrelated to log GDP.

```{r}
# fit bivariate regression
model <- lm(life_exp ~ gdp_per_cap, data = gapminder_df)
```

```
# clean up the data
tidy_model <- tidy(model) %>%
  mutate_if(is.numeric, round, 3)

# display table
kable(tidy_model, caption = "Bivariate Regression of Life Expectancy on GDP per Capita")

## Lets wash and repeat but with log gdp per capita
# fit bivariate regression
model2 <- lm(life_exp ~ log_gdp, data = gapminder_df)

# clean up the data
tidy_model2 <- tidy(model2) %>%
  mutate_if(is.numeric, round, 3)

# display table
kable(tidy_model2, caption = "Bivariate Regression of Life Expectancy on logged GDP per
Capita")
```