Announcements

- Make sure to sign in on the google form (linked here)
- Midterm October 11

Slope independent of outcome mean

1. Find the distribution of $\bar{Y}$. Recall that $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

2. Show that $E(\bar{Y}\hat{\beta}_1) = E(\bar{Y})E(\hat{\beta}_1)$. Recall that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

   Hint: Use the fact that $\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{\epsilon}$ and think about what is fixed and what is random.

3. Find the covariance of $\bar{Y}$ and $\hat{\beta}_1$.

4. Apply 7.5.7 from the Stat 110 textbook to show that $\bar{Y}$ and $\hat{\beta}_1$ are independent.

Redundant summary information

Here's a bunch of useful information (also available here, but be careful of what they call p):

Definitions:

- Sum of squares model (SSM): $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$
- Sum of squares error (SSE): $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
- Sum of squares total (SST): $\sum_{i=1}^n (Y_i - \bar{Y})^2$
- Degrees of freedom for model with p predictors and an intercept (df_M): p
- Degrees of freedom for error with p predictors and an intercept (df_E): $n - p - 1$
- Residual standard error: $\sqrt{\text{SSE}/\text{df}_E}$
- R^2: $1 - \text{SSE}/\text{SST}$
- Adjusted R^2: $1 - (1 - R^2)\frac{n-1}{\text{df}_E}$

Facts:

- $\text{SSE} + \text{SSM} = \text{SST}$
- $\hat{\sigma}^2 = \text{SSE}/\text{df}_E$
- Under the null (all coefficients are 0),

$$\frac{\text{SSM}/\text{df}_M}{\text{SSE}/\text{df}_E} \sim F_{\text{df}_M, \text{df}_E}$$

From the partial output above, calculate the following:

1. How many non-NA data points were included.

2. The t-statistics for the intercept and `mad_gdppc` coefficient.

3. The p-values of the two t-tests for the intercept and `mad_gdppc` coefficient being 0.

```
Call:
lm(formula = spi_ospi ~ mad_gdppc, data = countries)

Residuals:
     Min       1Q   Median       3Q      Max
-61.1775  -7.3751   2.7922   9.2159  15.7948

Coefficients:
              Estimate Std. Error t value  Pr(>|t|)
(Intercept) 5.9866e+01 1.2971e+00
mad_gdppc   4.6565e-04 4.6213e-05
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.745 on 151 degrees of freedom
  (41 observations deleted due to missingness)
Multiple R-squared:  0.40205,   Adjusted R-squared:
F-statistic:        on   and 151 DF,  p-value:
```

Figure 1: Lm output with missing information

```r
# TODO: intercept

# TODO: mad_gdppc coefficient
```

4. A 95% confidence interval for the `mad_gdppc` coefficient.

```r
# TODO: CI
```

5. The adjusted R^2.
6. The sum of squares error, the sum of squares total, and the sum of squares model.
7. The f-statistic and p-value for the test that all coefficients are equal to 0.

```r
# TODO: f statistic
```

8. Note that the hypothesis tested in 7 ($H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$) was the same as one of the hypotheses tested in 2. If our framework is consistent, these should give the same answer. Recall from week 2's section that if $T_n \sim t_n$, $T_n^2 \sim F_{1,n}$. Show (numerically) that your calculated t statistic squared is your f statistic, and explain how this shows that the two tests are the same. (Note that this only works because we have a single predictor.)

Regression on real data

This section will deal with a data set of country-level statistics from this source with an explanation of the data encoding found here.

```r
countries <- read.csv("data/countries.csv")
```

1. Fit a linear model to predict the percent of individuals using the internet in a country (`wdi_internet`) from the log of its GDP per capita (`mad_gdppc`), and formally test whether this association is significant. Provide a visual to support your conclusion.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
# TODO: Make linear model

# TODO: Make a plot
```

2. Check the assumptions of the model.

```
# TODO: Visualize assumptions
```

- Linearity:
- Constant variance:
- Normality:
- Independence:

3. Uganda has a GDP per capita listed but no statistic for internet access. Provide a point estimate and 90% prediction interval.

```
# TODO: Provide prediction
```

How bad are correlated residuals?

Let $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ with marginal $\epsilon_i \sim \mathcal{N}(0,1)$ and $\text{Corr}(\epsilon_i, \epsilon_{i+1}) = \rho$ for $i \in \{1, ..., n-1\}$ and $\text{Corr}(\epsilon_i, \epsilon_{i-1}) = \rho$ for $i \in \{2, ..., n\}$ and $\text{Corr}(\epsilon_i, \epsilon_j) = 0$ otherwise. Write a function to use simulation to find the probability of rejecting the null $H_0 : \beta_1 = 0$, the expected value $E(\hat{\beta}_1)$, and the standard deviation $\text{SD}(\hat{\beta}_1)$ in the following situations:

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.3
```

```
nsims = 1000
n = 10
b0 = 1

run_sim = function(nsims, n, p, b0, b1, sorted=FALSE) {
  # Covariance matrix
  Sigma = matrix(0, nrow = n, ncol = n)
  diag(Sigma) <- 1
  for (i in 2:n) {
    Sigma[i, i-1] <- p
    Sigma[i-1, i] <- p
```

```
  }

  pval = vector(length = nsims)
  coef = vector(length = nsims)

  for (i in 1:nsims) {
    # Generate x
    if (sorted) {
      x <- sort(rgamma(n, 3, 2/5))
    } else {
      x <- rgamma(n, 3, 2/5)
    }
    # Generate y with multivariate normal
    y <- b0 + b1 * x + mvrnorm(n = 1, rep(0, n), Sigma)

    # TODO: Get p-value and coefficient
  }

  # TODO: Return probability of rejecting the null, mean, sd
  return()
}
```

1. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$, $\rho = 0$, $\beta_0 = 1$, $\beta_1 = 1$.

```
run_sim(nsims, n, 0, b0, 1, sorted=FALSE)
```

```
## NULL
```

2. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$, $\rho = 0.5$, $\beta_0 = 1$, $\beta_1 = 1$.

```
run_sim(nsims, n, 0.5, b0, 1, sorted=FALSE)
```

```
## NULL
```

3. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$ sorted, $\rho = 0$, $\beta_0 = 1$, $\beta_1 = 1$.

```
run_sim(nsims, n, 0, b0, 1, sorted=TRUE)
```

```
## NULL
```

4. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$ sorted, $\rho = 0.5$, $\beta_0 = 1$, $\beta_1 = 1$.

```
run_sim(nsims, n, 0.5, b0, 1, sorted=TRUE)
```

```
## NULL
```

5. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$, $\rho = 0$, $\beta_0 = 1$, $\beta_1 = 0$.

```r
run_sim(nsims, n, 0, b0, 0, sorted=FALSE)
```

```
## NULL
```

6. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$, $\rho = 0.5$, $\beta_0 = 1$, $\beta_1 = 0$.

```r
run_sim(nsims, n, 0.5, b0, 0, sorted=FALSE)
```

```
## NULL
```

7. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$ sorted, $\rho = 0$, $\beta_0 = 1$, $\beta_1 = 0$.

```r
run_sim(nsims, n, 0, b0, 0, sorted=TRUE)
```

```
## NULL
```

8. $n = 10$, $X_i \sim \text{Gamma}(3, 2/5)$ sorted, $\rho = 0.5$, $\beta_0 = 1$, $\beta_1 = 0$.

```r
run_sim(nsims, n, 0.5, b0, 0, sorted=TRUE)
```

```
## NULL
```

9. What conclusions can you draw?