Announcements

Make sure to sign in on the google form (I send a list of which section questions are useful for which pset questions afterwards)

Pset 3 due Friday 10/6

Introductions

- One question or thought related to lecture last week (Correlation, simple regression, inference)

Slope independent of outcome mean

In this problem, we'll show that the slope in a linear regression ($\hat{\beta}_1$) is independent of the mean outcome ($\bar{Y}$). Suppose we have pairs (X_i, Y_i) for $i \in \{1, ..., n\}$.

1. Recall that in a simple linear regression we assume $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ with X_i known and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The vector $(\bar{Y}, Y_1 - \bar{Y}, Y_2 - \bar{Y}, ..., Y_n - \bar{Y})^T$ has a multivariate Normal distribution. Find the covariance of $\bar{Y}$ and $Y_i - \bar{Y}$.

$$\begin{aligned}
\text{Cov}(\bar{Y}, Y_i - \bar{Y}) &= \text{Cov}(\bar{Y}, Y_i) - \text{Cov}(\bar{Y}, \bar{Y}) \\
&= \text{Cov}(Y_i/n, Y_i) - \text{Var}(\bar{Y}) \\
&= \frac{1}{n}\text{Var}(Y_i) - \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} \beta_0 - \beta_1 X_i + \epsilon_i\right) \\
&= \frac{\sigma^2}{n} - \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(\epsilon_i) \\
&= 0
\end{aligned}$$

2. What does this imply about $\bar{Y}$ and all the $Y_i - \bar{Y}$?

$\bar{Y}$ is independent of all the $Y_i - \bar{Y}$ since uncorrelated implies independent in a multivarite Normal distribution.

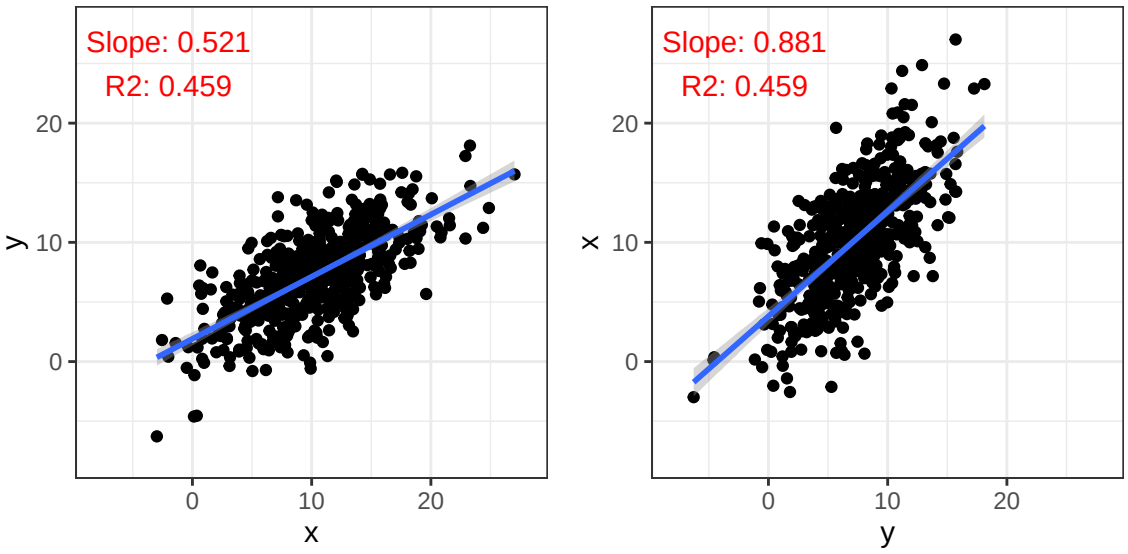
3. What does this say about the relationship between $\bar{Y}$ and $\hat{\beta}_1$? Recall that

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

$\hat{\beta}_1$ is just a function of the $(Y_i - \bar{Y})$s, so it is also independent of $\bar{Y}$.

Rule of thumb

Suppose we have n pairs of (X_i, Y_i) and we regress Y on X to get a slope $\hat{\beta}_1$ and X on Y to get a slope $\hat{\beta}_1'$. At first glance, it might seem like the $\hat{\beta}_1 = 1/\hat{\beta}_1'$. However, as you can see in the plots below, this is wrong.



1. Why is this wrong?

Because we are only trying to minimize the vertical residuals, we end up with non-reciprocal slopes. You can imagine a case where X and Y are independent, so both slopes would be 0, but clearly these are not reciprocals. This can also be viewed as a case of regression to the mean in which an extreme X value predicts a Y value that's not quite as extreme.

2. In the rest of the problem, we'll try to find the proper relationship between the two slopes. Recall that when regressing Y on X, we have

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Consider our simple regression with the estimators

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

and consider the flipped regression estimators

$$\hat{\beta}_1' = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}, \quad \hat{\beta}_0' = \bar{X} - \hat{\beta}_1' \bar{Y}$$

Find an expression for $\hat{\beta}_1'$ in terms of $\hat{\beta}_1$.

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

3. Solve for R^2 in terms of $\hat{\beta}_1$ and $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$. You may use the fact that

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

(See my Stat 111 section 6 notes for why this is the case in simple linear regression.)

$$
\begin{aligned}
R^2 &= 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 \bar{X} - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^n (\hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \hat{\beta}_1^2 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}
\end{aligned}
$$

4. Use this to write an expression for $\hat{\beta}_1'$ in terms of R^2 and $\hat{\beta}_1$.

$$\hat{\beta}_1' = \frac{R^2}{\hat{\beta}_1}$$

Notably, $0 \leq R^2 \leq 1$ for an OLS model, so $0 \leq \hat{\beta}_1' \leq 1/\hat{\beta}_1$. Not only does this give us the relation between the slopes, it does so in a way that uses the two most commonly reported statistics about the model: the estimated slope and the R^2.

Real data linear model

These problems will deal with a dataset of country-level statistics from UNdata and Varieties of Democracy.

1. Suppose we want to know the relationship between log 2010 GDP per capita and the 2010 life expectancy for females at birth. Interpret the following output.

```
##
## Call:
## lm(formula = `Life expectancy at birth for females (years)` ~
##     log2(`GDP per capita (US dollars)`), data = countries_2010)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.3076  -2.2879   0.8095   3.4083  12.2565
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         31.3250     2.2198   14.11   <2e-16 ***
## log2(`GDP per capita (US dollars)`)  3.3499     0.1753   19.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.622 on 207 degrees of freedom
##   (23 observations deleted due to missingness)
## Multiple R-squared:  0.6382, Adjusted R-squared:  0.6365
## F-statistic: 365.2 on 1 and 207 DF,  p-value: < 2.2e-16
```

This shows that a change in log2 GDP per capita of 1 (a doubling of GDP per capita) corresponds to an extra 3.35 years of life expected at birth.

2. Suppose we read this result in a paper but what we actually cared about was the regression of log2 GDP per capita on female life expectancy at birth. What can we conclude about this alternative regression?

Using the result we derived above, our slope would be $R^2/\hat{\beta}_1 = 0.6382/3.3499 = 0.1905$, so an increase in female life expectancy at birth of 1 year is associated with a 0.19 increase in log2 GDP per capita (1.14x multiplicative increase).

Filling in the lm table

Here's some useful information:

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 the model with p predictors and an intercept (df_M): p
- Degrees of freedom for the error with p predictors and an intercept (df_E): $n - p - 1$
- 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),

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

We'll be looking at emissions per capita regressed on log GDP per capita in 2010. For context, average emissions for countries that reported them were 5.27 metric tons of carbon dioxide per person.

```
##
## Call:
## lm(formula = `Emissions per capita (metric tons of carbon dioxide)` ~
##     log2(`GDP per capita (US dollars)`), data = countries_2010)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.698 -2.474 -1.015  1.186 18.369
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -18.445      2.204                 ***
## log2(`GDP per capita (US dollars)`)   1.869      0.172                 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.175 on 139 degrees of freedom
##   (91 observations deleted due to missingness)
## Multiple R-squared:  0.4593, Adjusted R-squared:
## F-statistic:       on   and    DF,  p-value:
```

Figure 1: Lm output with missing information

From the partial output above, calculate the following:

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

$n = \text{df}_E + p + 1 = \text{df}_E + 1 + 1 = 141$

2. The t-statistics for the intercept and the `log2(GDP per capita (US dollars))` coefficient.

$$t = \frac{\text{Estimate}}{\text{Standard error}} \implies t_{\beta_0} = -18.445/2.204 = -8.37$$
$$t_{\beta_1} = 1.869/0.172 = 10.87$$

3. How you would find the p-values of the two t-tests for the intercept and the `log2(GDP per capita (US dollars))` coefficient being 0.

We want the mass that is beyond the t-statistic in the t_{df_E} distribution:

$$p_{\beta_0} = 2 \cdot (1 - F_{t_{139}}(|t_{\beta_0}|)) = 5.5 \times 10^{-14}$$
$$p_{\beta_1} = 2 \cdot (1 - F_{t_{139}}(|t_{\beta_1}|)) = 2.7 \times 10^{-20}$$

where $F_{t_{139}}$ is the t_{139} CDF.

4. A 95% confidence interval for the `log2(GDP per capita (US dollars))` coefficient.

Letting t^* be the 0.975 quantile of the t_{139} distribution,

$$\hat{\beta}_1 \pm t^* \cdot \mathrm{SE}_{\hat{\beta}_1} = 1.869 \pm 1.977 \cdot 0.172 = (1.53, 2.21)$$

which doesn't include 0 as expected.

5. The adjusted R^2.

$$1 - (1 - R^2)\frac{n-1}{\mathrm{df}_E} = 1 - (1 - 0.4593)\frac{140}{139} = 0.4554$$

6. The sum of squares error, the sum of squares total, and the sum of squares model.

$$\mathrm{SSE} = \text{Residual standard error}^2 \cdot \mathrm{df}_E = 4.175^2 \cdot 139 = 2422.857$$
$$\mathrm{SST} = \frac{\mathrm{SSE}}{1 - R^2} = 2422.857/0.5407 = 4480.964$$
$$\mathrm{SSM} = \mathrm{SST} - \mathrm{SSE} = 2058.107$$

7. The f-statistic and p-value for the test that all coefficients are equal to 0.

$$f_{\text{Overall}} = \frac{\mathrm{SSM}/\mathrm{df}_M}{\mathrm{SSE}/\mathrm{df}_E} = \frac{2058.107/1}{2422.857/139} = 118.1$$
$$p_{\text{Overall}} = 1 - F_{1,139}(f_{\text{Overall}}) = 2.7 \times 10^{-20}$$

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.)

The two test statistics are within rounding error of each other: $t^2 = 10.87^2 = 118.2 \approx 118.1 = f$. Under the null, a t-statistic T_n of β_1 has a t_n distribution, so T_n^2 will have an $F_{1,n}$ distribution. Therefore, with the observed t-statistic t_n and $f = t_n^2$,

$$P(|t_n| \geq |T_n|) = P(t_n^2 \geq T_n^2) = P(t_n^2 \geq F_{1,n}) = P(f \geq F_{1,n})$$

where the first and last probabilities give our two p-values.

The full linear model for the image is here:

```
##
## Call:
## lm(formula = `Emissions per capita (metric tons of carbon dioxide)` ~
##     log2(`GDP per capita (US dollars)`), data = countries_2010)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.698 -2.474 -1.015  1.186 18.369
```

```
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -18.445      2.204  -8.367 5.63e-14 ***
## log2(`GDP per capita (US dollars)`)   1.869      0.172  10.866  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.175 on 139 degrees of freedom
##   (91 observations deleted due to missingness)
## Multiple R-squared:  0.4593, Adjusted R-squared:  0.4554
## F-statistic: 118.1 on 1 and 139 DF,  p-value: < 2.2e-16
```

Intuitive F test

Performing an overall F test with the sum of squares as above makes sense when deriving the F test, but the sum of squares involved are cumbersome and unintuitive. Here, we'll create a more intuitive test statistic.

1. Write SSE and SSM in terms of $\hat{\sigma}^2$, df_E, and R^2.

$$\text{SSE} = \hat{\sigma}^2 \cdot \text{df}_E$$
$$\text{SST} = \frac{\text{SSE}}{1-R^2} \implies \text{SSM} = \text{SST} - \text{SSE} = \text{SSE}\left(\frac{1}{1-R^2} - 1\right) = \hat{\sigma}^2 \cdot \text{df}_E \cdot \frac{R^2}{1-R^2}$$

2. Use these to write the F-statistic only in terms of R^2, df_E, and df_M.

$$F = \frac{\text{SSM}/\text{df}_M}{\text{SSE}/\text{df}_E} = \frac{R^2}{1-R^2} \cdot \frac{\text{df}_E}{\text{df}_M}$$

3. Use this to explain how a higher or lower R^2, df_E, and df_M contribute to a more or less significant F test. Why do these make sense?

- Holding df_E and df_M equal, an R^2 closer to 1 gives a larger F-statistic, which makes sense because the model is explaining more of the variability, so we expect the coefficients to be non-zero.
- When df_E is higher (holding the other two equal), the F statistic increases. When the R^2 is the same and df_E is higher, the model is explaining more data points with the same number of predictors, giving us confidence that the coefficients are non-zero.
- When df_M is higher (holding the other two equal), we're using more predictors to get the same explanatory power (R^2), so we expect that these coefficients are not that useful. This drives down the F-statistic, giving us a less significant result.