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 (Inference, linear model assumptions, and intro to multiple regression)

Filling in the lm table

Here's some useful information:

Definitions:

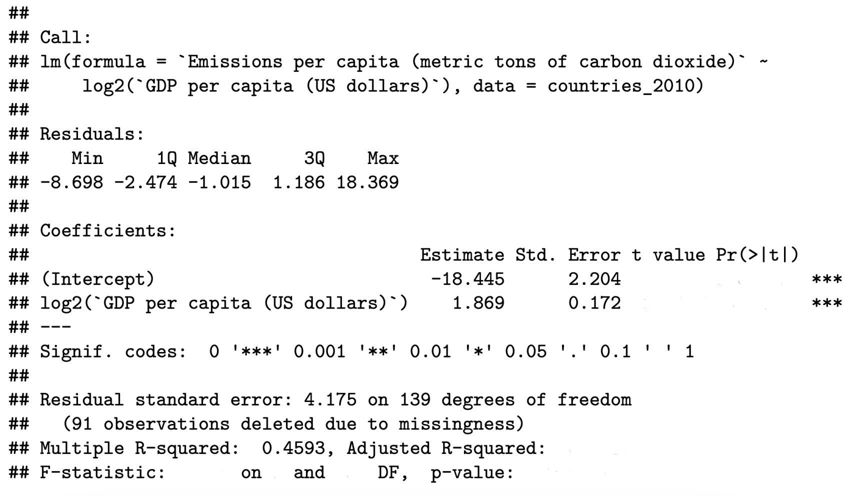
- Sum of squares model (SSM): $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$
- Sum of squares error (SSE): $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
- Sum of squares total (SST): $\sum_{i=1}^n (Y_i - \bar{Y})^2$
- Degrees of freedom for the model with p predictors and an intercept (df_M): p
- Degrees of freedom for 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,

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

which doesn't include 0 as expected.

5. The adjusted R^2.

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

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

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

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

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

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.

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

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.

Regression on real data

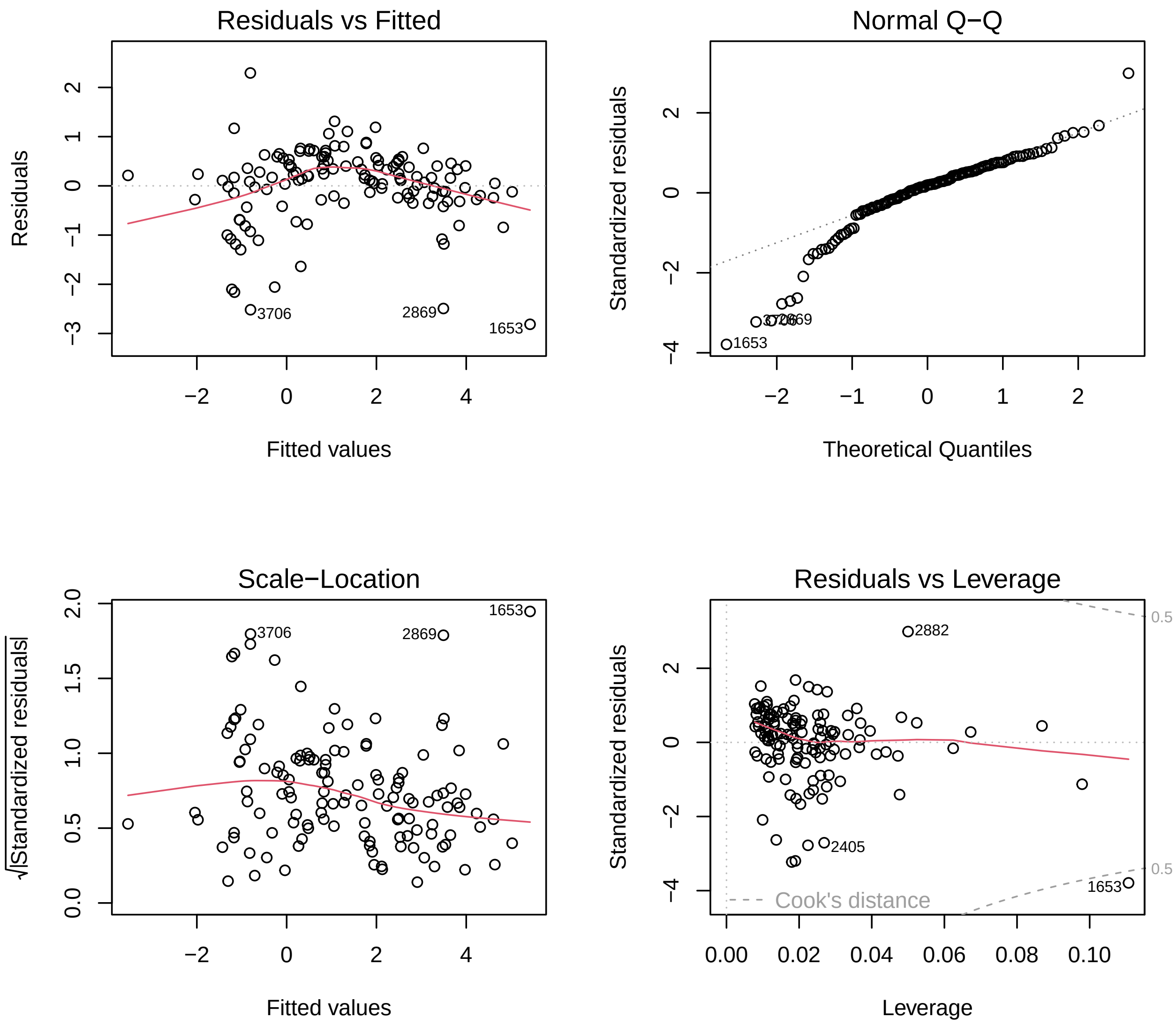
These problems will deal with a dataset of country-level statistics from UNdata and Varieties of Democracy.

1. Using this linear model regressing log emissions per capita on log energy per capita and the log of the number of tourists, interpret the results:

```
##
## Call:
## lm(formula = log2(`Emissions per capita (metric tons of carbon dioxide)`) ~
##     log2(`Supply per capita (gigajoules)`) + log2(`Tourist/visitor arrivals (thousands)`),
##     data = countries_2010[countries_2010$`Emissions per capita (metric tons of carbon dioxide)` >
##         0, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8131 -0.2609  0.1511  0.4590  2.2934
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                  -6.33647    0.36951 -17.148
## log2(`Supply per capita (gigajoules)`)        1.11628    0.05216  21.402
## log2(`Tourist/visitor arrivals (thousands)`)  0.09175    0.03605   2.545
##                                              Pr(>|t|)
## (Intercept)                                    <2e-16 ***
## log2(`Supply per capita (gigajoules)`)         <2e-16 ***
## log2(`Tourist/visitor arrivals (thousands)`)   0.0121 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7869 on 128 degrees of freedom
##   (99 observations deleted due to missingness)
## Multiple R-squared:  0.8452, Adjusted R-squared:  0.8428
## F-statistic: 349.5 on 2 and 128 DF,  p-value: < 2.2e-16
```

Holding the number of tourists constant, a doubling in energy supply per capita (a 1 point change on the log2 scale) is associated with a $2^{1.116} = 2.17\times$ increase in emissions. Holding the energy supply constant, a doubling in tourist arrivals is associated with a $1.07\times$ increase in emissions.

2. Check the assumptions of the model.



- Linearity: The Residuals vs Fitted plot shows that there is no clear pattern to the residuals, so linearity is likely upheld.
- Constant variance: Based on the Scale-Location plot, the residuals are about equal over the fitted values.
- Normality: The Q-Q plot show that the lower tail is larger than expected. The emissions are possibly left skewed because a few countries had already started cutting emissions at this point.
- Independence: This might not be true: countries that had entered into emissions cutting deals by 2010 probably influenced each others' emissions.

3. Uganda has tourism and energy usage data but no emissions data. The following are a 90% confidence interval and a 90% prediction interval for Uganda's log emissions from this data. Identify which is which, and interpret them.

```
##         fit   lwr    upr
## 3638 -0.605 -0.77 -0.439
```

```
##         fit    lwr   upr
## 3638 -0.605 -1.919 0.709
```

The first is the confidence interval because it is narrower; it can be interpreted as an interval for the mean log emissions of countries with energy usage and tourism like Uganda. The second is the prediction interval because it is wider; it can be interpreted as an interval for Uganda's log emissions (or a country with the same energy usage and tourism as Uganda).

4. What we actually care about is Uganda's emissions, not its log emissions. We can exponentiate one of the intervals above to get a valid interval on the original scale, but exponentiating the other would not be valid. Which is which and why?

We cannot exponentiate the confidence interval because that would violate Jensen's inequality:

$$0.95 = P(A \leq E(\log(Y)|X = x) \leq B) = P(e^A \leq \exp(E(\log(Y)|X = x)) \leq e^B) \neq P(e^A \leq E(Y|X = x) \leq e^B)$$

However, we can exponentiate the prediction interval:

$$0.95 = P(A \leq \log(Y) \leq B|X = x) = P(e^A \leq Y \leq e^B|X = x)$$

This exponentiated prediction interval is 0.26 to 1.63 metric tons of carbon dioxide per person in Uganda. For reference, the United States' 2010 emissions per capita were 17.3 metric tons of carbon dioxide per person.