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:

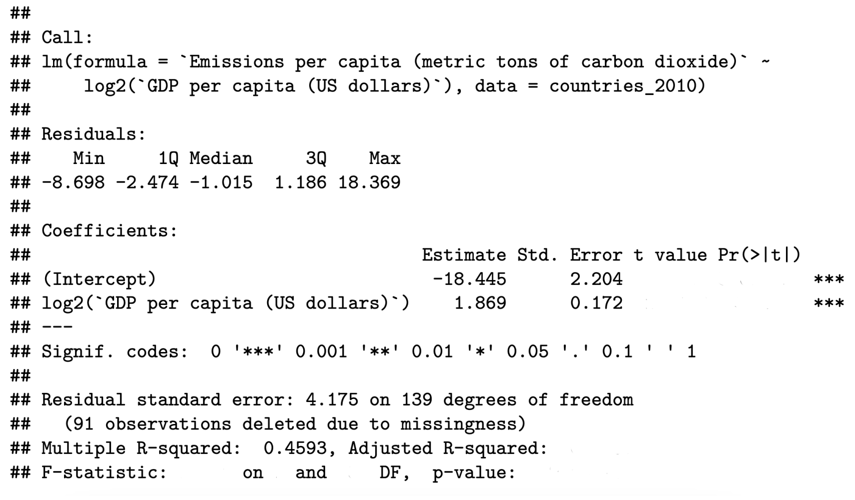
- Sum of squares model (SSM): $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
- Sum of squares error (SSE): $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- Sum of squares total (SST): $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- Degrees of freedom for the model with p predictors and an intercept (df_M): p
- Degrees of freedom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

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

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.

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

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.

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.)

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.

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

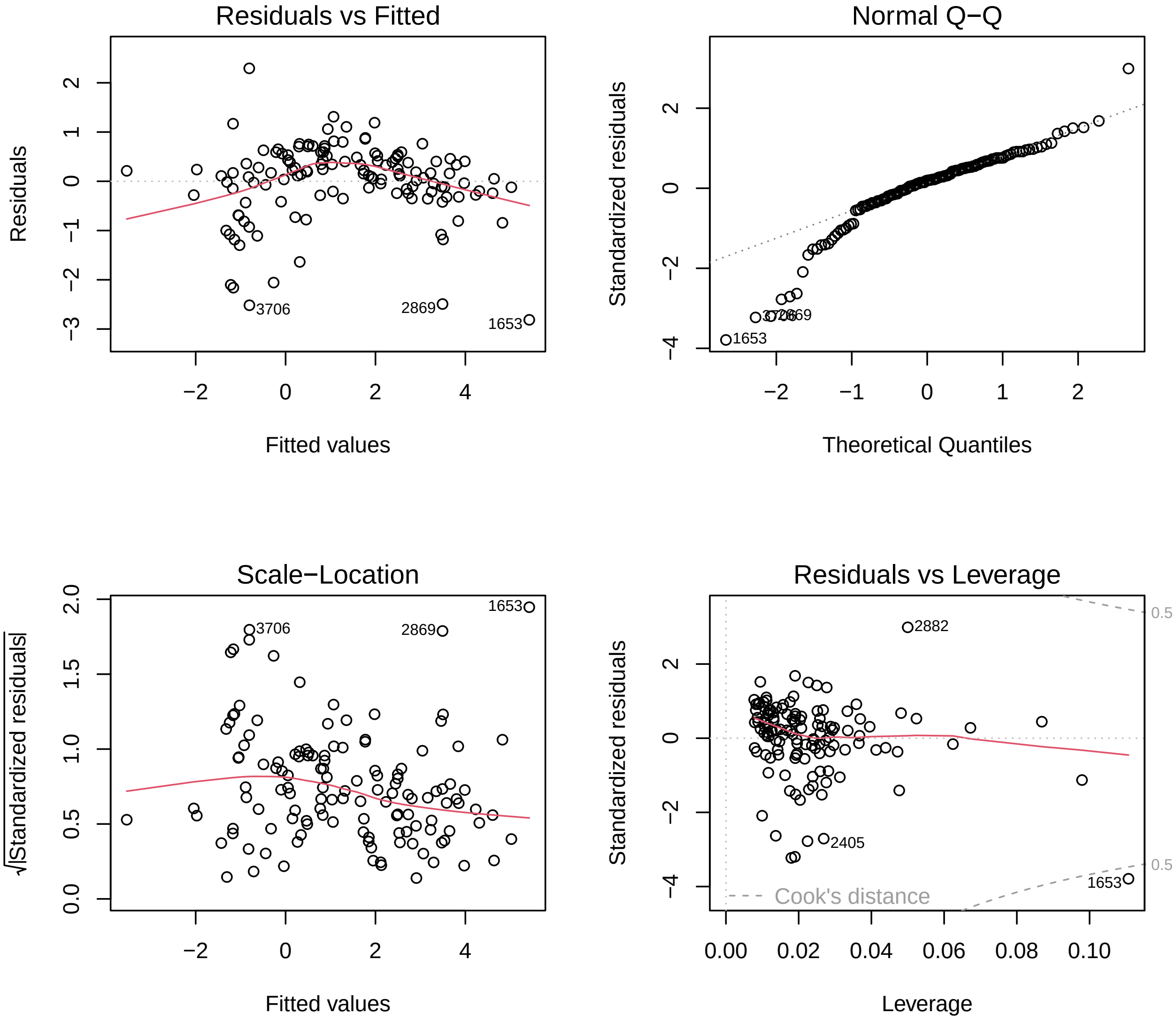
Regression on real data

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
```

2. Check the assumptions of the model.



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
```

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?