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)

Redundant summary information

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 model with p predictors and an intercept (df_M): p
- Degrees of freedom for error with p predictors and an intercept (df_E): $n - p - 1$
- 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 it 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 `log2(GDP per capita (US dollars))` coefficient.

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

3. How you would find the p-values of the two t-tests for the intercept and `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, so 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:

```
## [1] 141
```

```
##
## 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
```

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.

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.

We want to test $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ where β_1 is the association between log GDP per capita and percent of individuals with access to internet in a country. We get a slope of 20.97 and a t-statistic of 24.7 for that slope with a p-value of less than $2.2 \times 10^{-16} < \alpha = 0.05$, so we reject the null and conclude there is an association between log GDP per capita and percent of individuals with access to internet in a country (go figure!).

```
library(ggplot2)

lm1 <- lm(wdi_internet ~ log(mad_gdppc), countries)
summary(lm1)

ggplot(countries, aes(x=log(mad_gdppc), y=wdi_internet)) +
  geom_point() +
  geom_smooth(method='lm', formula= y~x)
```

2. Check the assumptions of the model.

```
par(mfrow=c(2,2))
plot(lm1)
```

- 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, there might be slightly more variance in residuals for countries with GDPs near the world-wide median, but the variance is about constant. (The Residuals

vs Fitted plot makes it look like there is more variance in the middle, but note that there's also more data there in the first place.)
- Normality: The Q-Q plot show that the lower tails are slightly lower than expected with the normal assumption, but overall the normal assumption fits very well.
- Independence: This is questionable: even given GDP, it's possible that internet use in a region is correlated because companies able to set up and maintain the infrastructure might work across multiple countries in a region.

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

```
log(countries[countries$cname=="Uganda",]$mad_gdppc)
predict(lm1, newdata=countries[countries$cname=="Uganda",],
        interval = c("prediction"), level = 0.90)
```

And do confidence interval

F test from R^2 (bottom of lecture 8)