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 4 due Friday 10/13

Midterm on Tuesday 10/17

Introductions

- Names
- One question or thought related to lecture last week (Inference in multiple regression, linear regression through matrices, assumptions)

Regression on real data

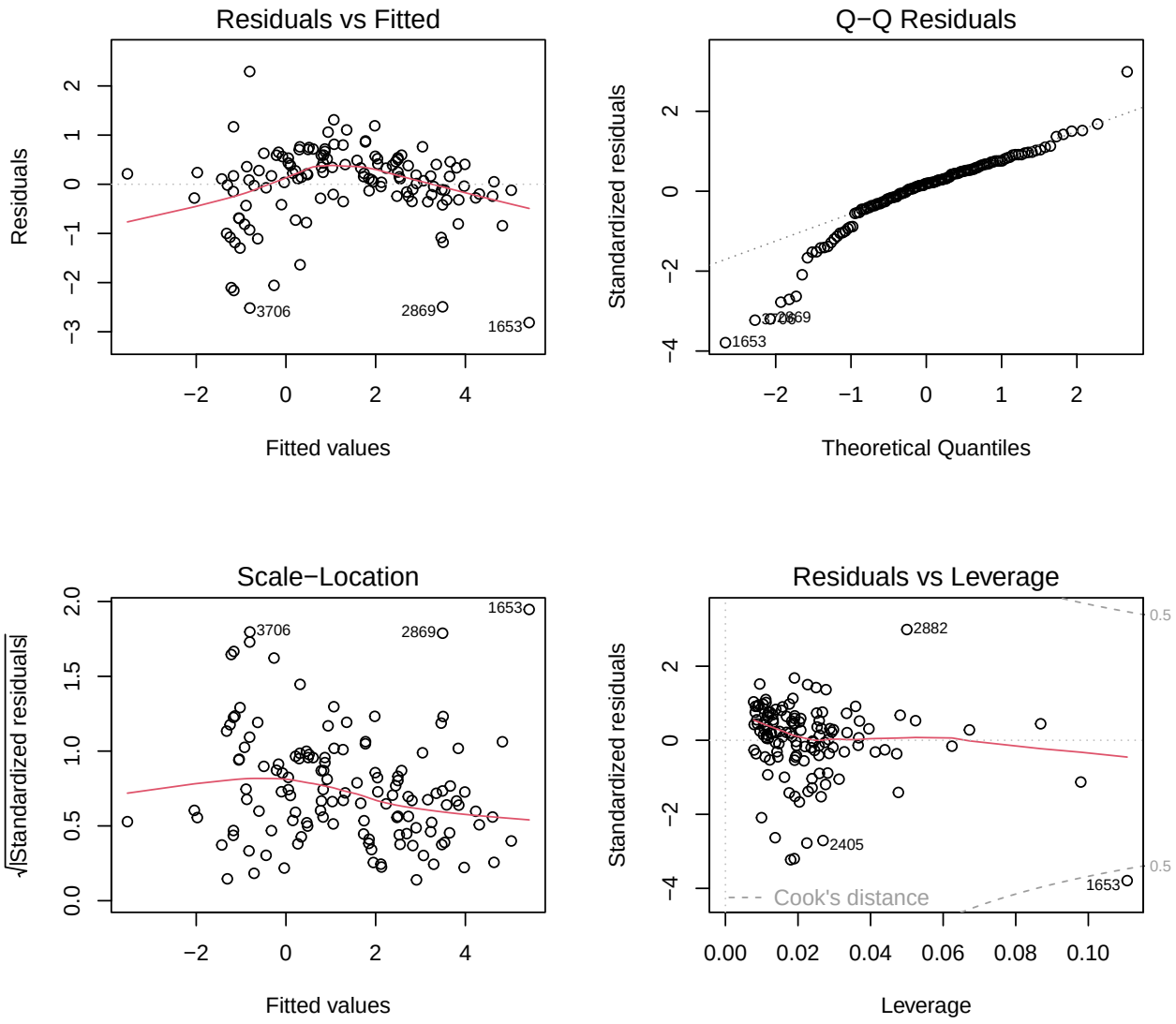
These problems will deal with a dataset of country-level statistics from UNdata and Varieties of Democracy.

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:

```
##                                            Estimate Std. Error    t value
## (Intercept)                               -6.33646928 0.36950869 -17.148363
## log2(`Supply per capita (gigajoules)`)     1.11627506 0.05215828  21.401683
## log2(`Tourist/visitor arrivals (thousands)`)  0.09175356 0.03604693   2.545391
##                                               Pr(>|t|)
## (Intercept)                               5.775412e-35
## log2(`Supply per capita (gigajoules)`)    4.116626e-44
## log2(`Tourist/visitor arrivals (thousands)`) 1.210220e-02
```

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.

Coefficient correlation

Recall that our sampling distribution of $\vec{\hat{\beta}}$ is

$$\vec{\hat{\beta}} \sim \text{MVN}(\vec{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

We usually estimate the variance-covariance matrix with $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$, but covariances in the $\hat{\beta}_i$ are hard to interpret. Instead, it would be better to know the correlations.

1. Let $\mathbf{\Sigma} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. How can we create a correlation matrix from this? You should index into $\mathbf{\Sigma}$ in your answer.

Copy $\mathbf{\Sigma}$ into a new matrix $\mathbf{\Sigma}'$. Then, for each i in $1, ..., p+1$, divide the i^{th} row in $\mathbf{\Sigma}'$ by $\sqrt{\mathbf{\Sigma}_{i,i}}$. Now, divide the i^{th} column in $\mathbf{\Sigma}'$ by $\sqrt{\mathbf{\Sigma}_{i,i}}$. This is now the correlation matrix since the entry corresponding to $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j)$ is divided by $\sqrt{\text{Var}(\hat{\beta}_i)\text{Var}(\hat{\beta}_j)}$.

2. Three models were fit to predict emissions per capita:

- Only energy supply per capita
- Only tourist/visitor arrivals
- Both energy supply per capita and tourist/visitor arrivals.

A correlation matrix for the coefficients is shown for the last model. Explain the large drop in the tourist/visitor arrivals coefficient from model 2 to model 3. Note that in the original data the energy supply per capita and tourist/visitor arrivals are slightly positively correlated.

```
##                         Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)           -5.610742 0.29308328 -19.14385 2.022591e-40
## log2(`Energy supply`)  1.173298 0.04752678  24.68710 4.300151e-52
```

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    -3.9419615 0.76185675 -5.174151 8.380139e-07
## log2(Tourists)  0.4711983 0.06706709  7.025776 1.045734e-10
```

```
##                           Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)           -6.33646928 0.36950869 -17.148363 5.775412e-35
## log2(`Energy supply`)  1.11627506 0.05215828  21.401683 4.116626e-44
## log2(Tourists)         0.09175356 0.03604693   2.545391 1.210220e-02
```

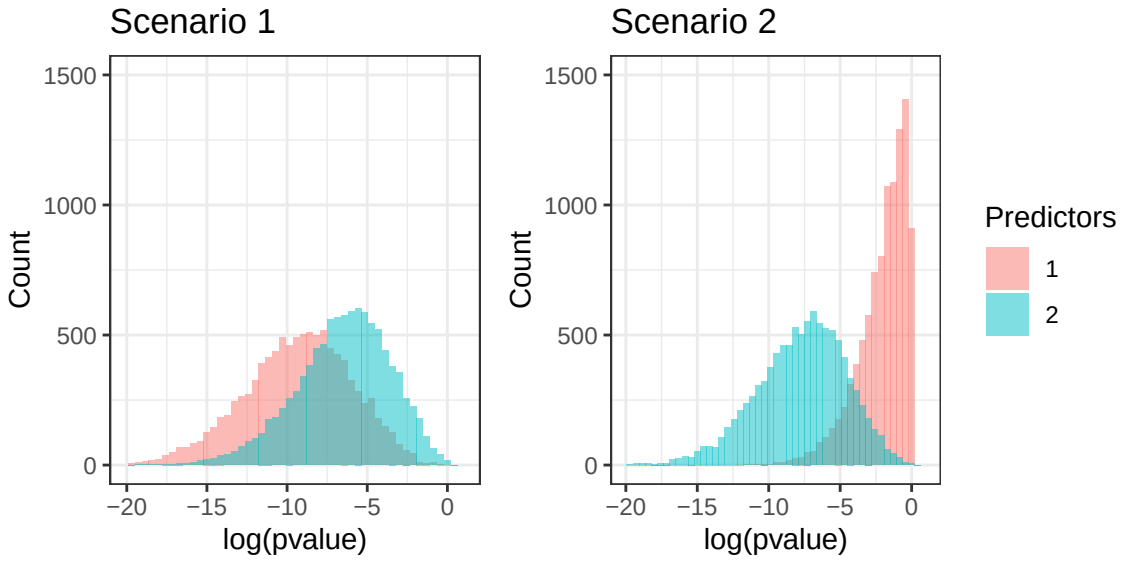
```
##                       (Intercept) log2(`Energy supply`) log2(Tourists)
## (Intercept)                 1.000                -0.285         -0.667
## log2(`Energy supply`)      -0.285                 1.000         -0.506
## log2(Tourists)             -0.667                -0.506          1.000
```

On their own, energy supply per capita and tourist/visitor arrivals are both positively associated with emissions, which makes sense. In the last model, the coefficients for these two variables are quite negatively correlated, meaning that as one increases the other tends to decrease. This explains why both coefficients are slightly less than in their separate models, and the tourist/visitor arrivals coefficient is shrunk more because the energy supply per capita is much more strongly associated with emissions.

3. Consider the following simulation. We will generate data from the model $Y_i = \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. (We'll have $\beta_1 = \beta_2$, but, of course, the linear model doesn't know this.)

- First, the columns $\vec{X}_1$ and $\vec{X}_2$ will be correlated, and we will fit either a regression just on $\vec{X}_1$ or a regression on both $\vec{X}_1$ and $\vec{X}_2$.
- Second, we will make $\vec{X_1}$ and $\vec{X_2}$ uncorrelated but make $\vec{X_2}$ have a very large variance, and we will again test models with and without $\vec{X_2}$.

We'll record the p-value of the test $H_0 : \beta_1 = 0$ each time.



Explain the p-value trends in the missing-predictor models. Reference the equation for the variance-covariance matrix as necessary.

In the first simulation, when the predictors are correlated but one is missing, some of the predictive ability of $\overrightarrow{X_2}$ is absorbed by $\overrightarrow{X_1}$, causing it to be more significant than when $\overrightarrow{X_2}$ is also included. In the second simulation, missing $\overrightarrow{X_2}$ results in a large amount of unexplained variation, driving up $\hat{\sigma}^2$ in the variance-covariance matrix and therefore increasing the apparent variance of $\hat{\beta}_1$, making it less significant.

4. Consider the design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & a \end{bmatrix}$$

What do the rows represent? What do the columns represent? Why should the first column be all 1s?

The rows are individual data observations. The columns are predictors. The first column is all 1s because every observation uses the same intercept.

5. Find the variance of $\hat{\beta}_1$ as a function of a and σ^2.

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 3 & 2+a \\ 2+a & a^2+2 \end{bmatrix} \implies (\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{3a^2+6-(4+4a+a^2)} \begin{bmatrix} a^2+2 & -2-a \\ -2-a & 3 \end{bmatrix}$$
$$= \frac{1}{2a^2-4a+2} \begin{bmatrix} a^2+2 & -2-a \\ -2-a & 3 \end{bmatrix}$$
$$= \frac{1}{2(a-1)^2} \begin{bmatrix} a^2+2 & -2-a \\ -2-a & 3 \end{bmatrix}$$

We can obtain the variance of $\hat{\beta}_1$ by extracting the bottom right entry and multiplying by σ^2 to get

$$\text{Var}(\hat{\beta}_1) = \frac{3\sigma^2}{2(1-a)^2}$$

6. What does this say about how the variance of $\hat{\beta}_1$ changes with a? Why does this make sense?

The variance is highest when $a = 1$ but decreases as a is further from 1. This makes sense because values of a that are close to 1 don't give much information relative to the points we already have.

Contrast test and limiting cases

Recall the set-up for a contrast test: $H_0 : \vec{C}^T\vec{\beta} = \gamma_0$ vs. $H_a : \vec{C}^T\vec{\beta} \neq \gamma_0$. Under the null, the following random variable has a $t_{n-(p+1)}$ distribution.

$$T = \frac{\vec{C}^T\vec{\hat{\beta}} - \gamma_0}{\hat{\sigma}\sqrt{\vec{C}^T(X^TX)^{-1}\vec{C}}}$$

1. Name two situations in which we would take γ_0 to be 0. What would the contrast vectors be in these cases?

- We could take γ_0 to be 0 if we wanted to see whether a single predictor had any effect. In this case, the contrast vector would be 0s except for a single 1 at the index of the predictor we care about.
- We could also take γ_0 to be 0 if we wanted to see whether the difference in the predicted outputs from two sets of predictors was significant. In this case, the contrast vector would be the first set of predictors minus the second.

2. Perform a formal contrast test based on the energy supply per capita plus tourists/visitors model to determine whether the mean emissions for countries like Seychelles is significantly different from the mean emissions for countries like Madagascar (two East African island countries).

```
## Seychelles:
```

```
##   (Intercept) Energy supply      Tourists
##         1.000         6.066         7.451
```

```
## Madagascar:
```

```
##   (Intercept) Energy supply      Tourists
##         1.000         3.170         7.615
```

```
## Test:
```

```
##       t.stat      p.value           df
## 2.087828e+01 4.861677e-43 1.280000e+02
```

To determine whether emissions in countries like Seychelles are higher than in countries like Madagascar, we want to test $H_0 : (6.066 - 3.170)\beta_1 + (7.451 - 7.615)\beta_2 = 0$ vs. $H_a : (6.066 - 3.170)\beta_1 + (7.451 - 7.615)\beta_2 \neq 0$. The resulting t-statistic is 20.9 based on 128 degrees of freedom, resulting in a p-value of 4.9×10^{-43}, so we reject the null and conclude that countries like Seychelles are very likely to have higher average emissions than countries like Madagascar.

3. Name two cases in which a contrast test should give the same result as another test.

- When two contrast vectors differ by 1 in a single predictor, we expect to see the same result as when running a t-test on that predictor itself.
- When the model contains only a single categorical predictor, a test between a vector with the intercept and category as 1s versus a vector with only the intercept as 1 should be the same as running a pooled t-test.

Linear model variances

Let $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ where $\epsilon_i \sim [0, \sigma^2]$ i.i.d. That is, the residuals are centered at 0 and are i.i.d., but they are not Normal. Under some commonly met regularity conditions, it can be shown that

$$\frac{1}{\sigma}(\mathbf{X}^T\mathbf{X})^{1/2}(\vec{\hat{\beta}} - \vec{\beta}) \xrightarrow{d} \text{MVN}(\vec{0}, \mathbf{I_{p+1}})$$

1. Suppose we have a consistent estimator for σ (we have some $\hat{\sigma}$ such that $\hat{\sigma} \xrightarrow{p} \sigma$). In the original multivariate Normal convergence statement, we don't know σ^2, but we still want to say something about convergence. How can we use the consistent estimator instead?

We can simply replace σ with $\hat{\sigma}$. By the continuous mapping theorem, $\frac{\sigma}{\hat{\sigma}} \xrightarrow{p} \frac{\sigma}{\sigma} = 1$, so by Slutsky's theorem

$$\frac{1}{\hat{\sigma}}(\mathbf{X}^T\mathbf{X})^{1/2}(\vec{\hat{\beta}} - \vec{\beta}) = \frac{\sigma}{\hat{\sigma}}\frac{1}{\sigma}(\mathbf{X}^T\mathbf{X})^{1/2}(\vec{\hat{\beta}} - \vec{\beta}) \xrightarrow{d} \text{MVN}(\vec{0}, \mathbf{I_{p+1}})$$

2. Find the approximate distribution of $\vec{\hat{\beta}}$ for large n.

By rearranging the equation above, we get

$$\vec{\hat{\beta}} \dot{\sim} \text{MVN}(\vec{\beta}, \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1})$$

which is the same as the exact distribution (with σ^2 instead) when the ϵ_i are Normal.

3. This result indicates that one of the linear model assumptions does not matter much with large n. Which assumption is this?

The Normality assumption does not matter much with large n.