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)

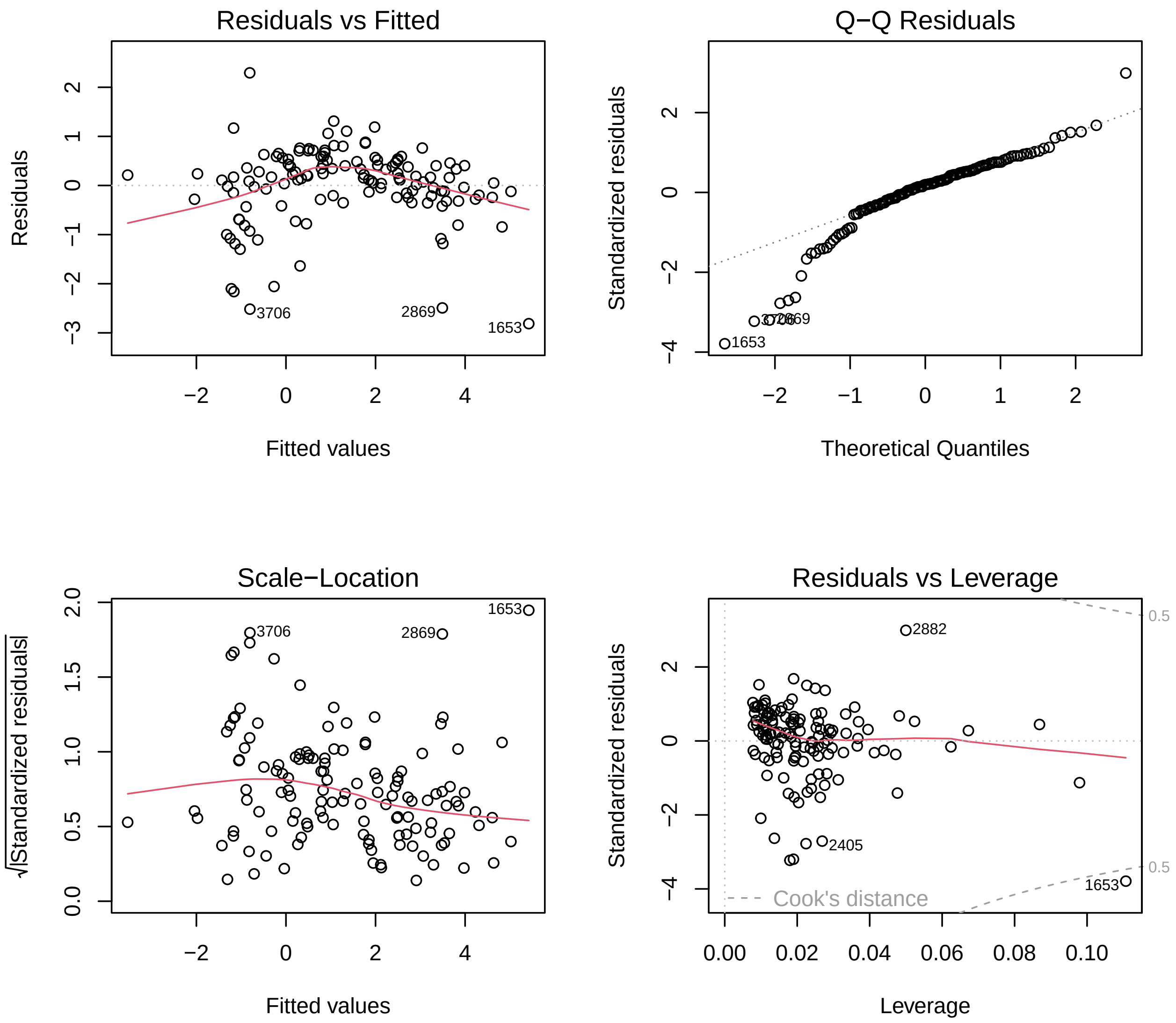
Regression on real data

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
```

2. Check the assumptions of the model.


Residuals vs Fitted
Residuals
Fitted values
−3
−2
−1
0
1
2
−2
0
2
4
3706
2869
1653
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
−4
−2
0
2
−2
−1
0
1
2
3706
2869
1653
Scale−Location
√|Standardized residuals|
Fitted values
0.0
0.5
1.0
1.5
2.0
−2
0
2
4
1653
3706
2869
Residuals vs Leverage
Standardized residuals
Leverage
−4
−2
0
2
0.00
0.02
0.04
0.06
0.08
0.10
2882
2405
1653
0.5
0.5
Cook's distance

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?

Coefficient correlation

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

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

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.

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
```

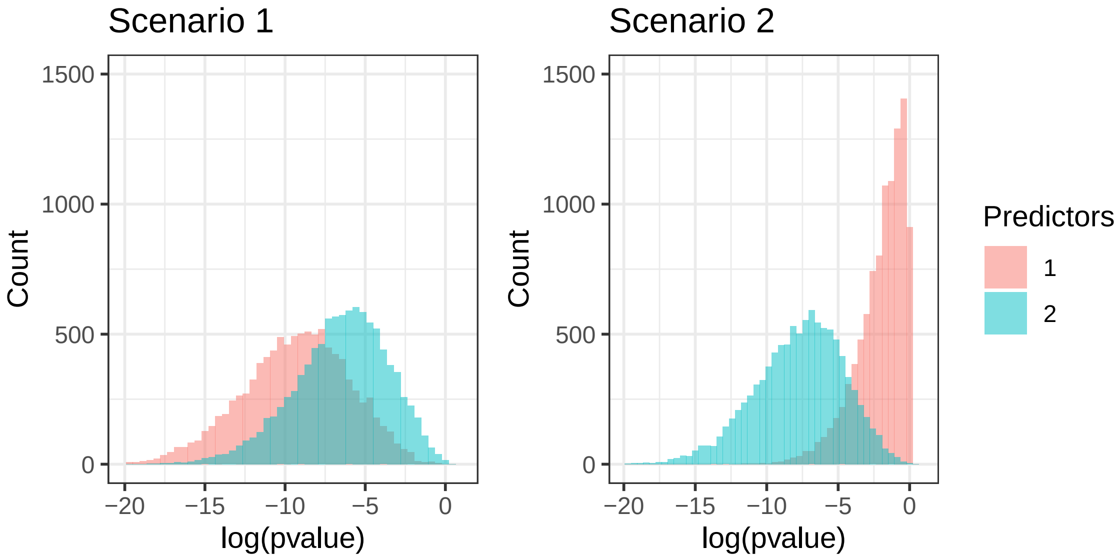
```
##                          Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)           -6.33646928 0.36950869 -17.148363 5.775412e-35
## log2(`Energy supply`)  1.11627506 0.05215828  21.401683 4.116626e-44
## log2(Tourists)         0.09175356 0.03604693   2.545391 1.210220e-02
```

```
##                       (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
```

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

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.

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?

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

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

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?

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
```

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

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

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?

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

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