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 5 due Friday 10/27

Introductions

- One question or thought related to lecture last week (interactions, polynomials, smoothers)

Correlated transformations

Transformations and polynomials are useful, but they often create additional correlation in the model. Suppose we have two increasing functions g and h, and let X_1 and X_2 be i.i.d. continuous random variables.

1. Explain why
$$(g(X_1) - g(X_2))(h(X_1) - h(X_2)) > 0$$

2. Take the expectation of both sides and expand to show that $\text{Cov}(g(X), h(X)) > 0$ for a random variable X with the same distribution as X_1 and X_2.

3. Suppose you have some continuous predictor X and two increasing transformations of X you include in the model. What does this say about the transformed predictors?

4. Suppose you have a strictly positive continuous predictor and you include it as a polynomial. What can you say about the $X, X^2, X^3, ...$ coefficients?

Groups and polynomials on real data

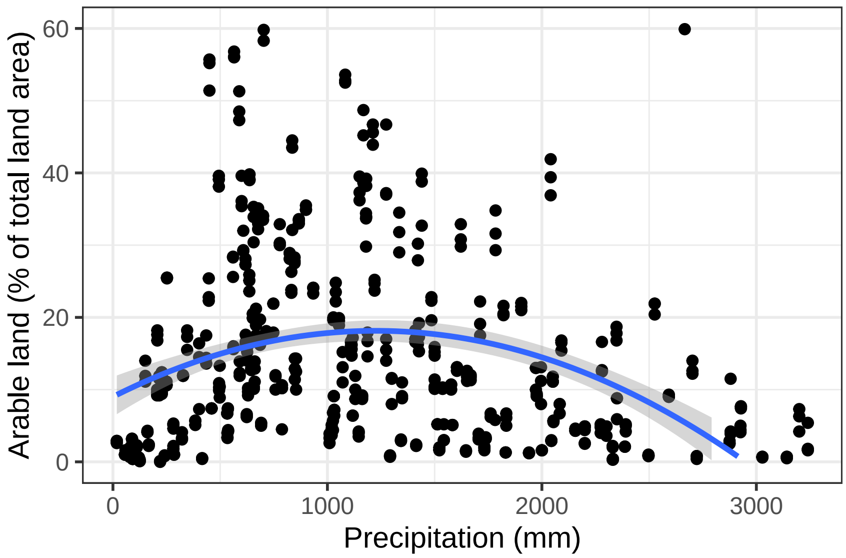
These problems will deal with a dataset of country-level statistics from UNdata, Varieties of Democracy, and the World Bank.

1. With `Northern America` as the reference group, a regression model is fit to predict a country's GDP per capita from its region. Interpret the coefficients.

```
##                           Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)               60429.75   8043.170  7.513176 2.310741e-12
## RegionCaribbean          -42178.39   8743.848 -4.823779 2.912959e-06
## RegionCentral America    -55547.75   9850.831 -5.638890 6.236910e-08
## RegionCentral Asia       -57064.35  10791.045 -5.288121 3.428172e-07
## RegionEastern Africa     -58689.36   8892.059 -6.600199 4.114707e-10
## RegionEastern Asia       -37451.32  10082.647 -3.714433 2.687637e-04
## RegionEastern Europe     -50539.75   9516.807 -5.310578 3.080613e-07
## RegionMelanesia          -50993.95  10791.045 -4.725580 4.499311e-06
## RegionMicronesia         -55750.35  10791.045 -5.166353 6.087637e-07
## RegionMiddle Africa      -56316.31   9666.687 -5.825812 2.442783e-08
## RegionNorthern Africa    -55787.75  10383.688 -5.372634 2.289083e-07
## RegionNorthern Europe    -19061.95   9516.807 -2.002977 4.662394e-02
## RegionOceania            -14338.75  13931.179 -1.029256 3.046888e-01
## RegionPolynesia          -51192.15  10791.045 -4.743947 4.149928e-06
## RegionSouth America      -52447.75   9287.453 -5.647162 5.985999e-08
## RegionSouth-eastern Asia -50754.57   9392.399 -5.403792 1.970253e-07
## RegionSouthern Africa    -55554.35  10791.045 -5.148190 6.626840e-07
## RegionSouthern Asia      -57816.19   9666.687 -5.980973 1.105426e-08
## RegionSouthern Europe    -38273.46   9120.097 -4.196607 4.183567e-05
## RegionWestern Africa     -59333.75   8992.537 -6.598110 4.161998e-10
## RegionWestern Asia       -42989.28   8939.484 -4.808922 3.112268e-06
## RegionWestern Europe      19415.92   9666.687  2.008539 4.602472e-02
```

2. The following 2nd order polynomial regression model predicts the percent of arable land in a country from its average annual precipitation. What is the optimal precipitation for having the most arable land?

```
##                                              Estimate   Std. Error   t value
## (Intercept)                              9.006338e+00 2.561091e+00  3.516602
## poly(`Precipitation (mm)`, 2, raw = TRUE)1  1.485334e-02 4.222079e-03  3.518015
## poly(`Precipitation (mm)`, 2, raw = TRUE)2 -5.911257e-06 1.417983e-06 -4.168778
##                                              Pr(>|t|)
## (Intercept)                              5.560447e-04
## poly(`Precipitation (mm)`, 2, raw = TRUE)1 5.532877e-04
## poly(`Precipitation (mm)`, 2, raw = TRUE)2 4.797093e-05
```

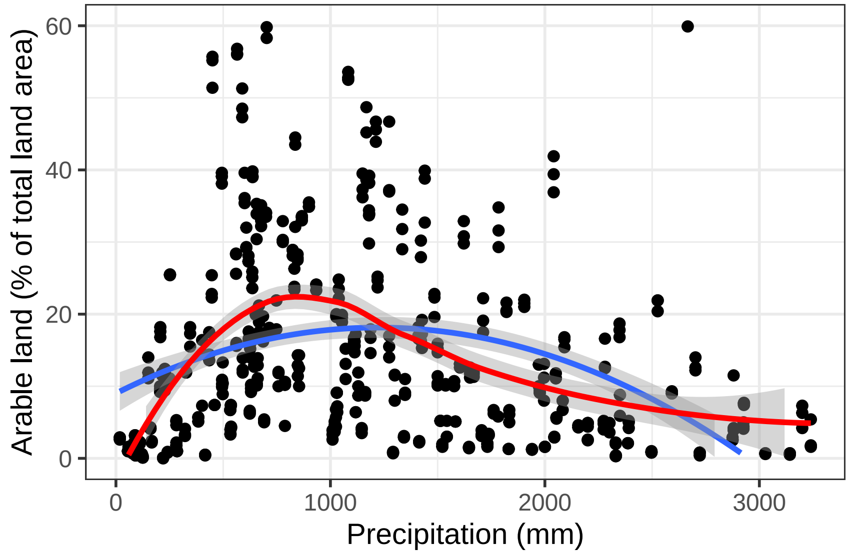


3. Use the previous model to find the probability that a country with x mm annual precipitation will have less than τ percent of its land arable. Recall that

$$T = \frac{Y - \vec{X}_0^T \vec{\hat{\beta}}}{\hat{\sigma}\sqrt{1 + \vec{X}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\vec{X}_0}} \sim t_{n-(p+1)}$$

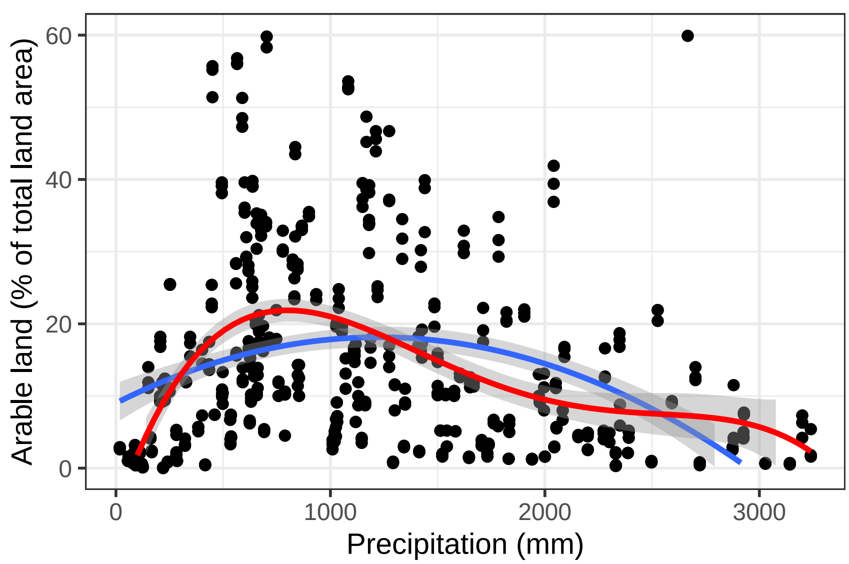
where $\vec{X}_0$ is the new vector of predictors, $\mathbf{X}$ is the matrix of previous predictors, and Y is the new outcome assuming it follows it the previous model.

4. Compare the prediction accuracy of a LOESS model to that of the previous model.




```
##    LM R2 LOESS R2
##    0.103    0.221
```

5. Perform a formal hypothesis test to determine whether a fourth degree polynomial fits the data better than a second degree polynomial.

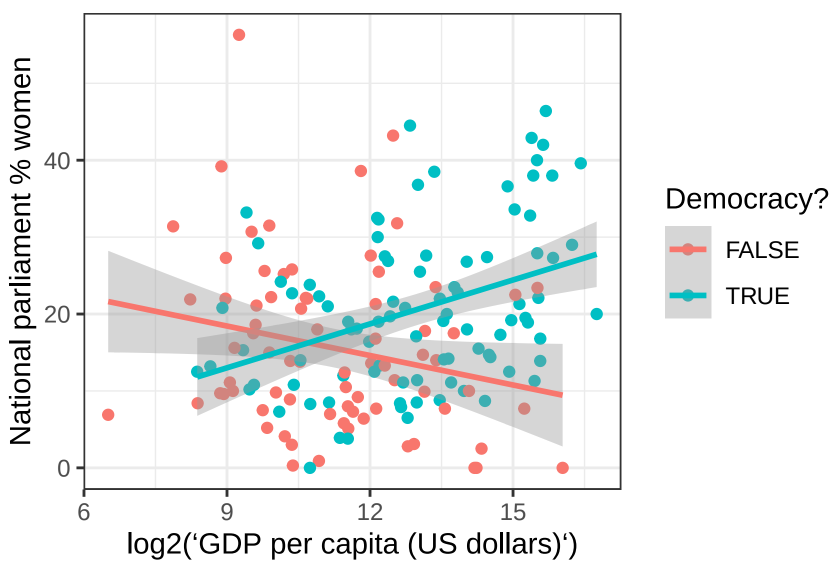



```
## Analysis of Variance Table
##
```

```
## Model 1: `Arable land (% of total land area)` ~ poly(`Precipitation (mm)`,
##     2, raw = TRUE)
## Model 2: `Arable land (% of total land area)` ~ poly(`Precipitation (mm)`,
##     4, raw = TRUE)
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    176 29714
## 2    174 25781  2    3933.3 13.273 4.316e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. Interpret the following model that looks at the proportion of the national parliament that is women as a function of GDP per capita and whether the country is a democracy.

```
##                                                          Estimate Std. Error
## (Intercept)                                             29.937995  6.9554720
## log2(`GDP per capita (US dollars)`)                     -1.277528  0.6094048
## is_democracyTRUE                                       -34.066009  9.8578282
## log2(`GDP per capita (US dollars)`):is_democracyTRUE     3.181353  0.8080104
##                                                          t value     Pr(>|t|)
## (Intercept)                                             4.304236 2.909763e-05
## log2(`GDP per capita (US dollars)`)                    -2.096353 3.762423e-02
## is_democracyTRUE                                       -3.455732 7.028484e-04
## log2(`GDP per capita (US dollars)`):is_democracyTRUE    3.937268 1.227137e-04
```

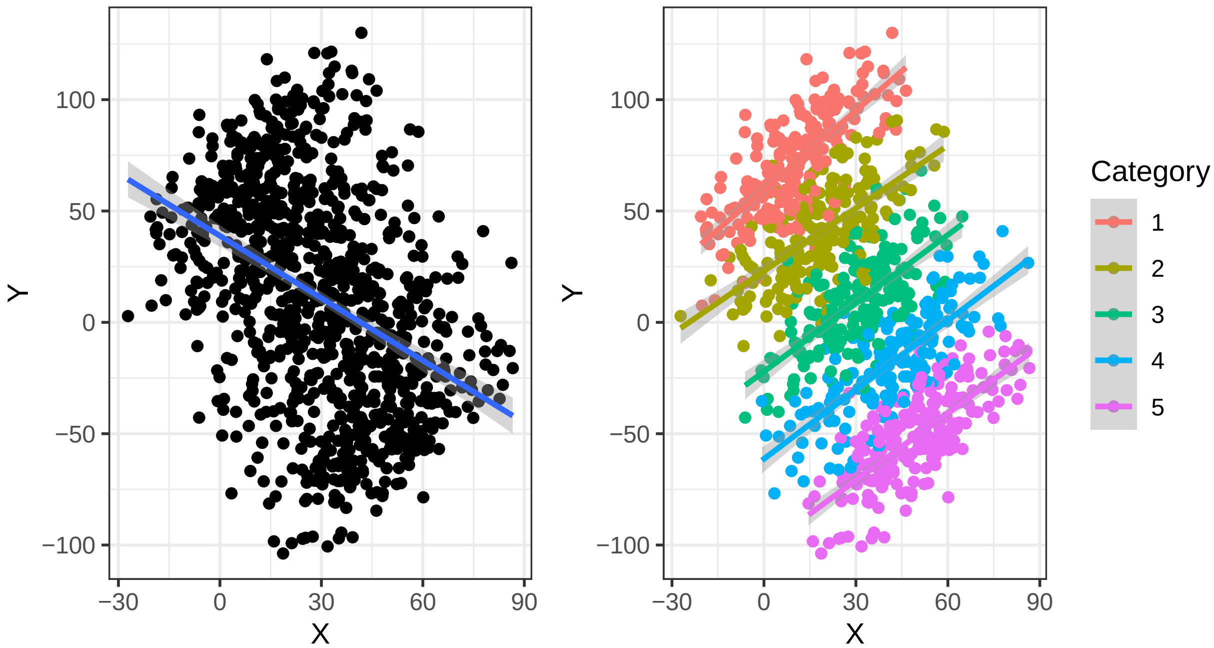


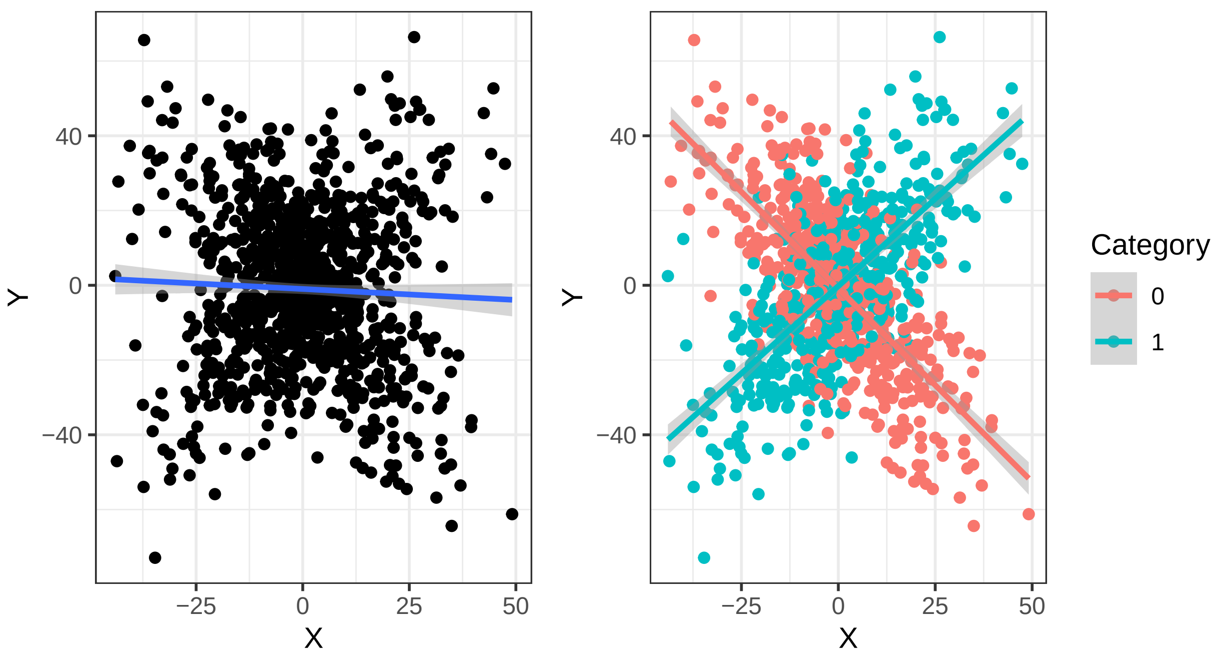
Simpson's simulation

1. For the following data table, write out the design matrix that would be used in the following model: `response ~ category * value`.

Response	Category	Value
12.7	3	5.1
24.7	2	4.9
-4.0	3	2.0
11.2	1	2.2
14.6	1	5.3
17.9	1	7.2
15.4	2	3.0
46.0	2	6.0
47.2	2	5.3
9.3	1	5.0

2. For each of the pairs of plots below, determine what model should be fit to best describe the data (e.g., `response ~ xˆ2 + category`).





3. Name a reason to avoid fitting many interaction terms right from the beginning.

ANOVA as a linear model

Let Y_{ij} be data point j from group i where there are k groups with n_i data points in group i. Imagine we run an ANOVA as well as an F-test for overall significance of a regression model with only the categories as predictors. Recall the original ANOVA F-statistic:

$$\frac{\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2/(n-k)}$$

and the overall regression F-statistic:

$$\frac{\sum_{i,j}(\hat{Y}_{ij} - \bar{Y})^2/p}{\sum_{i,j}(Y_{ij} - \hat{Y}_{ij})^2/(n-p-1)}$$

where p is the number of predictors (not including the intercept in the model).

1. What is p in terms of k?

2. What is $\hat{Y}_{ij}$? Why is this the case?

3. Show that the two F-statistics are equal.