Announcements

- Make sure to sign in on the google form (linked here)
- Pset 5 due October 21 at 5 pm

Normal interactions

Let $Z \sim \mathcal{N}(0,1)$ and $X \sim \mathcal{N}(\mu, \sigma^2)$.

1. Show that $\text{Corr}(Z, Z^n) = 0$ for all even whole numbers n.
2. Show that $\text{Corr}(Z, Z^n) > 0$ for all odd whole numbers n. You may use the useful fact from the Stat 110 book (page 284) that $E(Z^{2n}) = \frac{(2n)!}{2^n n!}$ for integers $n \geq 0$.
3. Find $\text{Cov}(X, X^2)$. When will this be positive? When will this be negative? (Hint: Consider standardizing X.)
4. What implication does this have for fitting linear models with a Normal predictor and its squared term?

Island of Misfit Toys

This section will deal with a data set of country-level statistics from this source with an explanation of the data encoding found here.

```
countries <- read.csv("data/countries.csv")
```

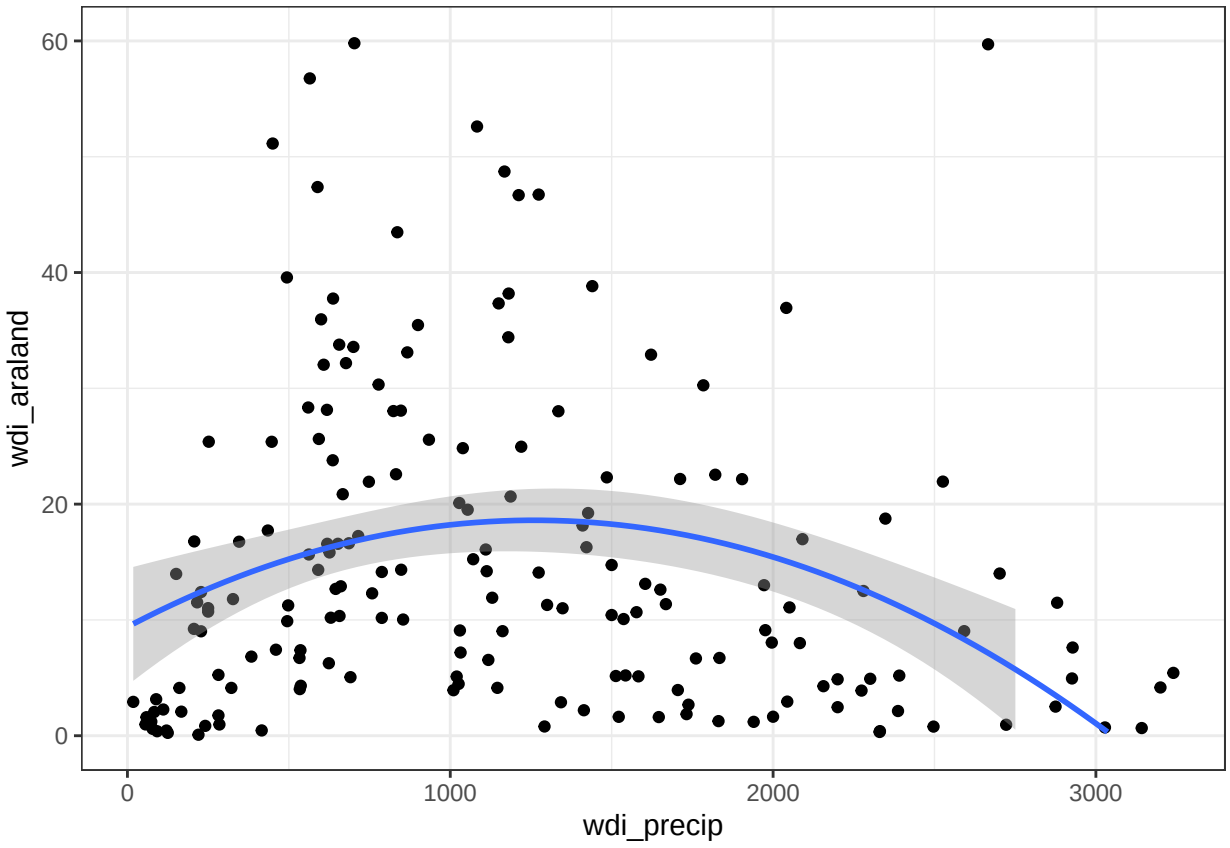
A few useful columns:

- `mad_gdppc`: GDP per capita
- `ht_region`: Country's region of the world: Eastern Europe (1), Latin America (2), North Africa & the Middle East (3), Sub-Saharan Africa (4), Western Europe and North America (5), East Asia (6), South-East Asia (7), South Asia (8), Pacific (9), Caribbean (10)
- `wdi_araland`: Arable land (% of land area)
- `wdi_precip`: Average annual precipitation (mm per year)
- `spi_ospi`: Overall social progress index on 0-100 scale
- `bmr_dem`: Binary democracy measure

1. Using `relevel` to set Western Europe and North America as the reference group, fit a regression model to predict a country's GDP per capita from its region. Interpret the coefficients.
2. Build a 2nd order polynomial regression model to predict the proportion of arable land in a country from its average annual precipitation. Interpret the output and provide a visual.

```
library(ggplot2)

# TODO: Get summary of 2nd order polynomial model

ggplot(countries, aes(x=wdi_precip, y=wdi_araland)) +
  geom_point() +
```

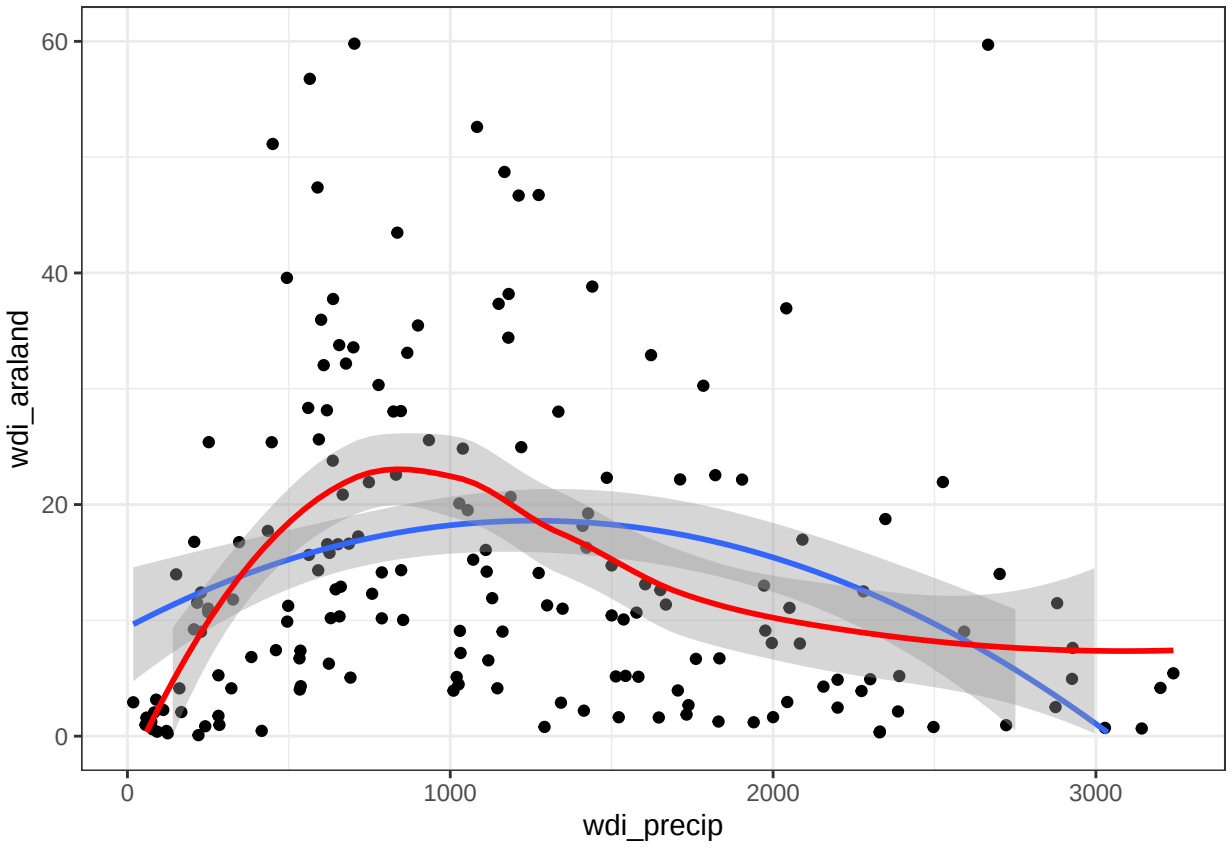
```
  stat_smooth(method = "lm",
              formula = y ~ poly(x, 2)) +
  ylim(0, 60) +
  theme_bw()
```



3. Use the previous model to find the probability that a country with 1500 mm annual precipitation per year on average will have less than 10% of its land arable.

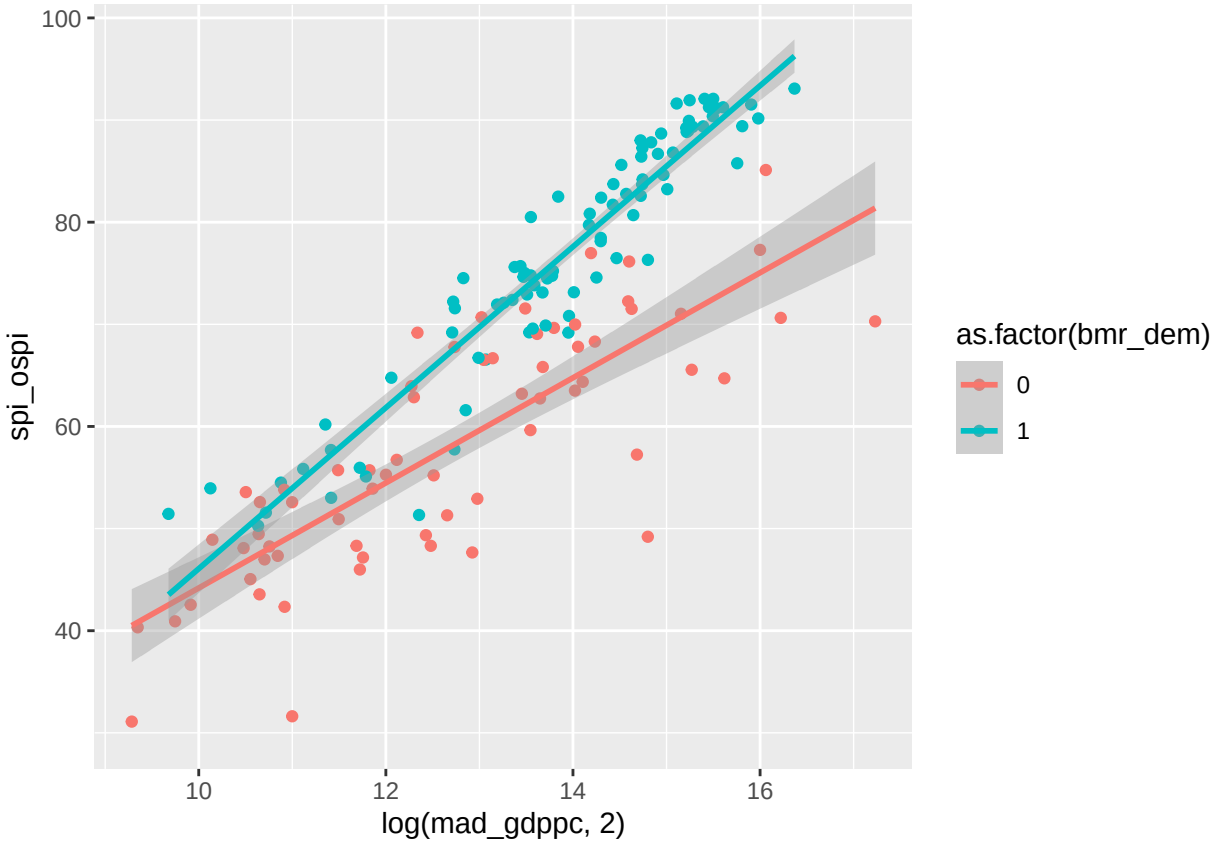
4. Fit a LOESS model for the same data and compare its prediction accuracy to that of the previous model.

```
# TODO: Fit LOESS model

# TODO: LM Rsq

# TODO: Loess Rsq

ggplot(countries, aes(x=wdi_precip, y=wdi_araland)) +
  geom_point() +
  stat_smooth(method = "lm",
              formula = y ~ poly(x, 2)) +
  stat_smooth(method="loess",
              formula = y ~ x,
              col="red") +
  ylim(0, 60) +
  theme_bw()
```



5. Fit a model to predict a country's overall social progress index from the log of its GDP per capita, its democracy status, and the interaction of the two. Interpret the coefficients of the model.

```
# TODO: Fit model

ggplot(countries, aes(x=log(mad_gdppc, 2), y = spi_ospi, col = as.factor(bmr_dem))) +
  geom_point() +
  stat_smooth(method="lm",
              formula = y~x)
```



6. Perform a formal hypothesis test to determine whether the previous model performs significantly better at predicting the overall social progress index than a model without the interaction term.

```
# TODO: Fit model and perform a test
```

Everything is just 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).

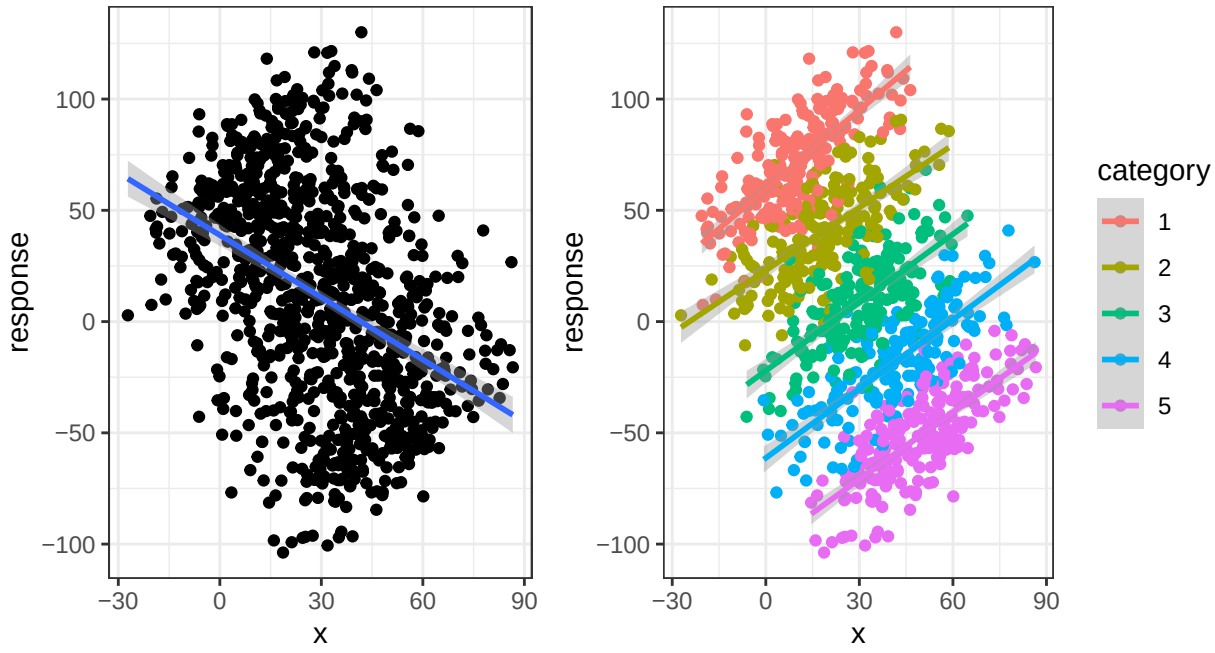
1. What is p in this case?
2. What is $\hat{Y}_{ij}$? Why is this the case?
3. Show that the two F-statistics are equal.

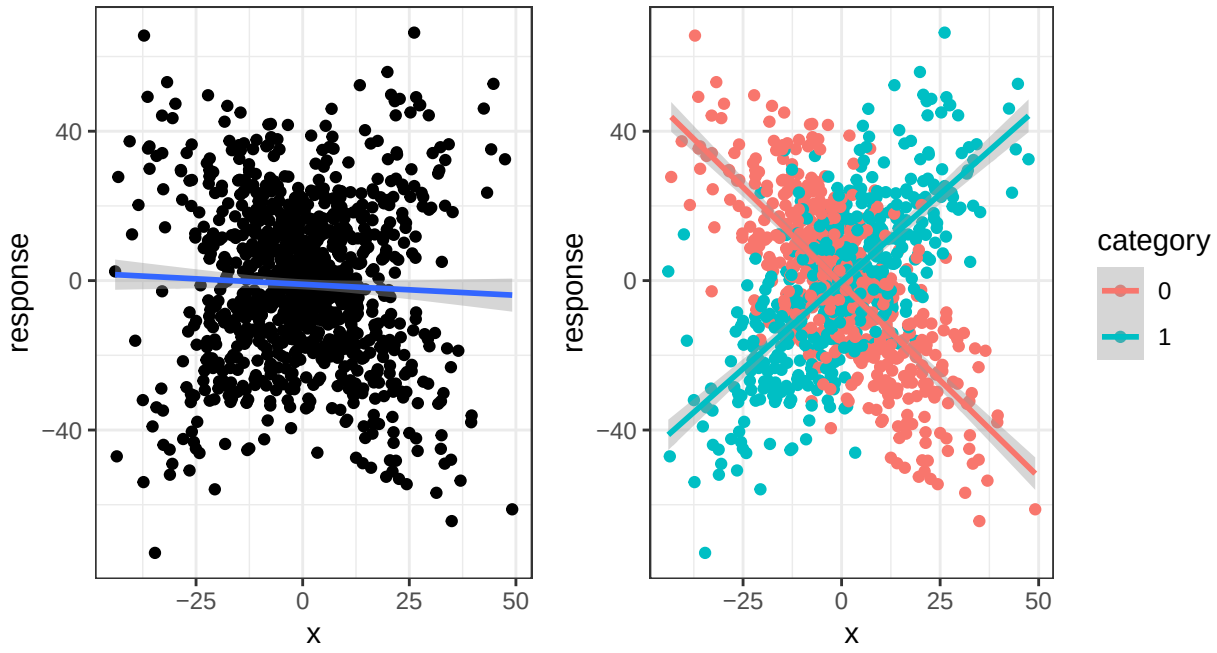
Simpson's simulation

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
12.780339	3	5.125949
24.721573	2	4.898613
-3.930666	3	2.031917
11.217700	1	2.213955
14.694621	1	5.348074
17.980544	1	7.238690
15.176966	2	2.962757
45.851668	2	5.980036
47.415309	2	5.333670
9.360024	1	5.003350

2. Without looking at the code that generated the data, for each of the pairs of plots below, determine what model should be fit to best describe the data.





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