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.

$$\text{Cov}(Z, Z^n) = E(Z^{n+1}) - E(Z)E(Z^n)$$

Because the LOTUS function for an odd power of a standard normal is odd, it integrates to 0, so $E(Z^{n+1}) = E(Z) = 0$, and the covariance and therefore the correlation are 0.

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 $n \geq 0$.

$$\text{Cov}(Z, Z^n) = E(Z^{n+1}) - E(Z)E(Z^n) = E(Z^{2k})$$

for some positive integer k since n is odd. Then, applying the Stat 110 result, $E(Z^{2k}) = \frac{(2k)!}{2^k k!} > 0$, so dividing it by the square root of the product of standard deviations will still give a positive number.

3. Find $\text{Cov}(X, X^2)$. When will this be positive? When will this be negative? (Hint: Consider standardizing X.)

First, note that

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

Then, for a standard normal Z,

$$\begin{aligned}
\text{Cov}(X, X^2) &= \text{Cov}(\sigma Z + \mu, (\sigma Z + \mu)^2) \\
&= \sigma \text{Cov}(Z, \sigma^2 Z^2 + 2\sigma\mu Z) \\
&= \sigma(E(\sigma^2 Z^3 + 2\sigma\mu Z^2) - E(Z)E(\sigma^2 Z^2 + 2\sigma\mu Z)) \\
&= \sigma(\sigma^2 E(Z^3) + 2\sigma\mu E(Z^2)) \\
&= 2\sigma^2\mu
\end{aligned}$$

Because σ^2 will always be positive, the covariance will have the same sign as μ. This makes sense because values of X that are larger in magnitude will correspond to values of X^2 that are larger in magnitude, so the only question is whether the always-positive X^2 will match to X that are usually positive or usually negative.

4. What implication does this have for fitting linear models with a Normal predictor and its squared term?

If the predictor is not centered at 0, a model that uses the predictor and its squared term will potentially have severe multicollinearity, changing slope coefficients significantly from models without the 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.

```
countries$ht_region <- as.factor(countries$ht_region)
countries$ht_region <- relevel(countries$ht_region, ref = "5")
summary(lm(mad_gdppc~ht_region, countries))$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  44737.85   3245.468 13.784714 1.138814e-28
## ht_region1  -26145.84   4380.094 -5.969241 1.581743e-08
## ht_region2  -32542.94   4758.790 -6.838490 1.767067e-10
## ht_region3  -13641.67   4758.790 -2.866625 4.729813e-03
## ht_region4  -39547.73   3974.870 -9.949440 2.552450e-18
## ht_region6  -19846.87   7135.111 -2.781578 6.085807e-03
## ht_region7  -27780.51   6119.713 -4.539512 1.129875e-05
## ht_region8  -39281.01   7135.111 -5.505312 1.508058e-07
## ht_region10 -31275.03   7680.179 -4.072174 7.428271e-05
```

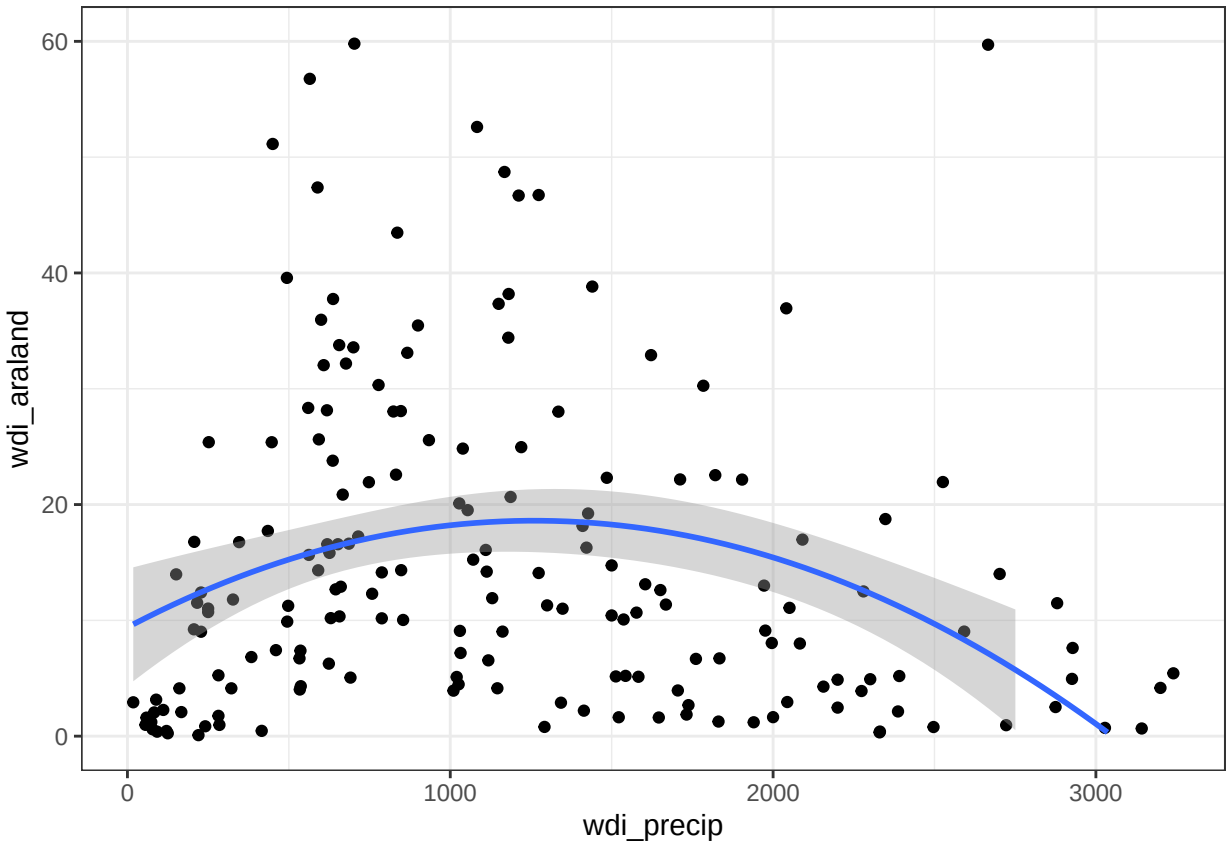
North America and Western Europe have an average GDP per capita of 44737.85 dollars (and this GDP per capita is significantly different from 0). Eastern Europe has an average GDP per capita of $44737.85 - 26145.84 = 18592.01$ dollars. (And so on...) Sub-Saharan Africa has the lowest average GDP per capita at $44737.85 - 39547.73 = 5190.12$ dollars, and South Asia has a GDP per capita that is nearly as low. All regions (except North America and Western Europe) have GDP per capita values that are significantly different from North America and Western Europe's.

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)

lm2 <- lm(wdi_araland~poly(wdi_precip, 2, raw = TRUE), countries)
summary(lm2)
```

```
##
## Call:
## lm(formula = wdi_araland ~ poly(wdi_precip, 2, raw = TRUE), data = countries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.801  -9.218  -3.469   6.259  52.559
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       9.399e+00  2.558e+00   3.674 0.000317 ***
## poly(wdi_precip, 2, raw = TRUE)1  1.461e-02  4.230e-03   3.454 0.000692 ***
## poly(wdi_precip, 2, raw = TRUE)2 -5.797e-06  1.421e-06  -4.080 6.82e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.04 on 175 degrees of freedom
##   (16 observations deleted due to missingness)
## Multiple R-squared:  0.0983, Adjusted R-squared:  0.08799
## F-statistic: 9.538 on 2 and 175 DF,  p-value: 0.000117
```

```
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()
```



Setting the derivative of the fit line to 0 gives $\frac{\hat{\beta_1}}{-2\hat{\beta_2}} \approx 1260$. Therefore, the estimated proportion of arable land increases with increasing precipitation up to 1260 mm per year, after which it declines. This makes sense because areas with either very little rainfall or a lot of rainfall are unlikely to have much productive farmland.

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.

```
threshold = 10

C = c(1, 1500, 1500^2)

# Using the null distribution, we can standardize the threshold and find the density below it
pt((threshold - C %*% coef(lm2)) /
     (sqrt(t(C) %*% vcov(lm2) %*% C + summary(lm2)$sigma^2)),
   lm2$df.residual)
```

```
##             [,1]
## [1,] 0.2645349
```

The probability of a country with 1500 mm annual precipitation having below 10% arable land is 0.264. This answer is based on a multivariate extension of the prediction interval in which

$$\hat{Y}|\vec{C} \sim \mathcal{N}\left(\vec{C}^T\vec{\beta}, \sigma^2[\vec{C}^T(\mathbf{X}^T\mathbf{X})^{-1}\vec{C}+1]\right)$$

When we only have our sample parameters, we use the fact that

$$\frac{\vec{C}^T\hat{\vec{\beta}}}{\sqrt{\hat{\sigma}^2[\vec{C}^T(\mathbf{X}^T\mathbf{X})^{-1}\vec{C}+1]}} \sim t_{n-p-1}$$

and we can use t-statistics of the form

$$\frac{\tau - \vec{C}^T\hat{\vec{\beta}}}{\sqrt{\hat{\sigma}^2[\vec{C}^T(\mathbf{X}^T\mathbf{X})^{-1}\vec{C}+1]}}$$

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

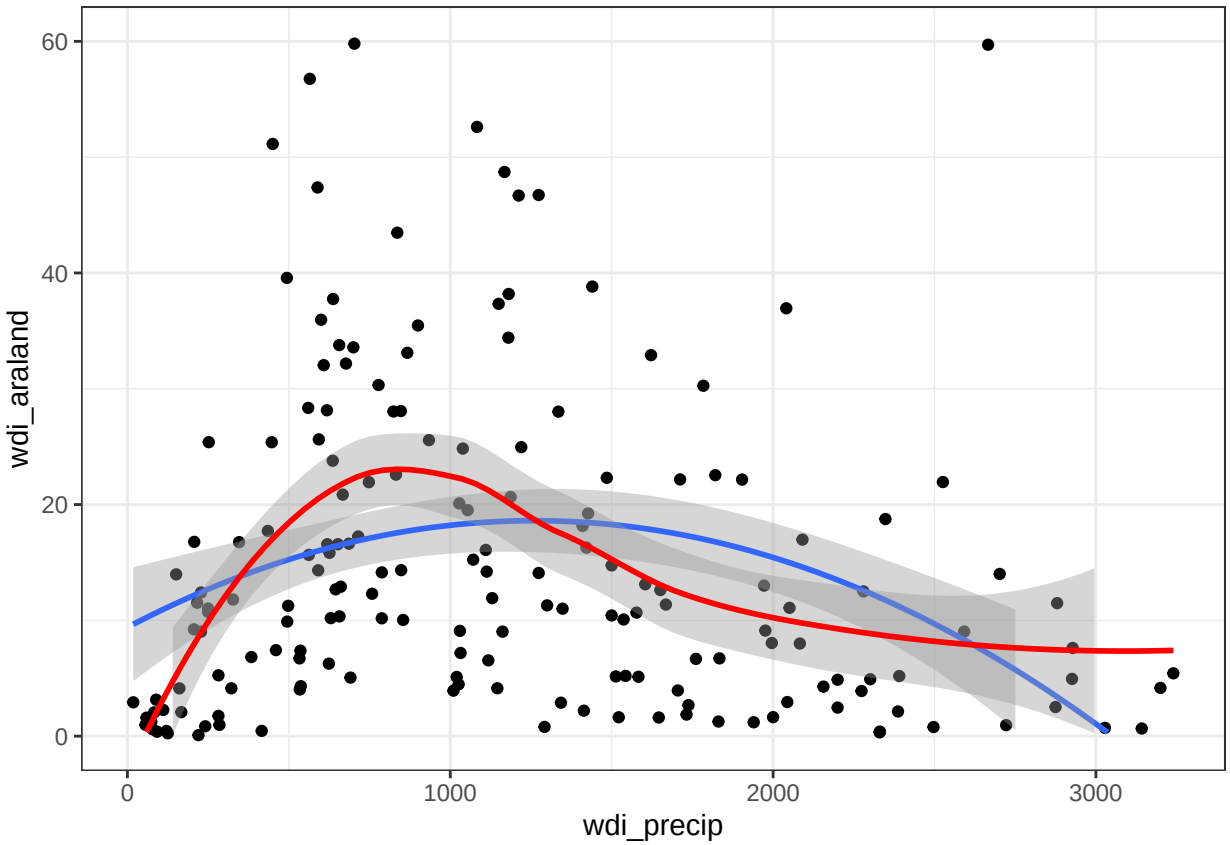
```
loess_model <- loess(wdi_araland~wdi_precip, countries)

# LM Rsq
1-sum(residuals(lm2)^2)/
  sum((countries[!is.na(countries$wdi_araland) & !is.na(countries$wdi_precip),]$wdi_araland -
         mean(countries[!is.na(countries$wdi_araland) & !is.na(countries$wdi_precip),]$wdi_araland))^2)
```

```
## [1] 0.09829558
```

```
# Loess Rsq
1-sum(residuals(loess_model)^2)/
  sum((countries[!is.na(countries$wdi_araland) & !is.na(countries$wdi_precip),]$wdi_araland -
         mean(countries[!is.na(countries$wdi_araland) & !is.na(countries$wdi_precip),]$wdi_araland))^2)
```

```
## [1] 0.2221675
```

```
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()
```



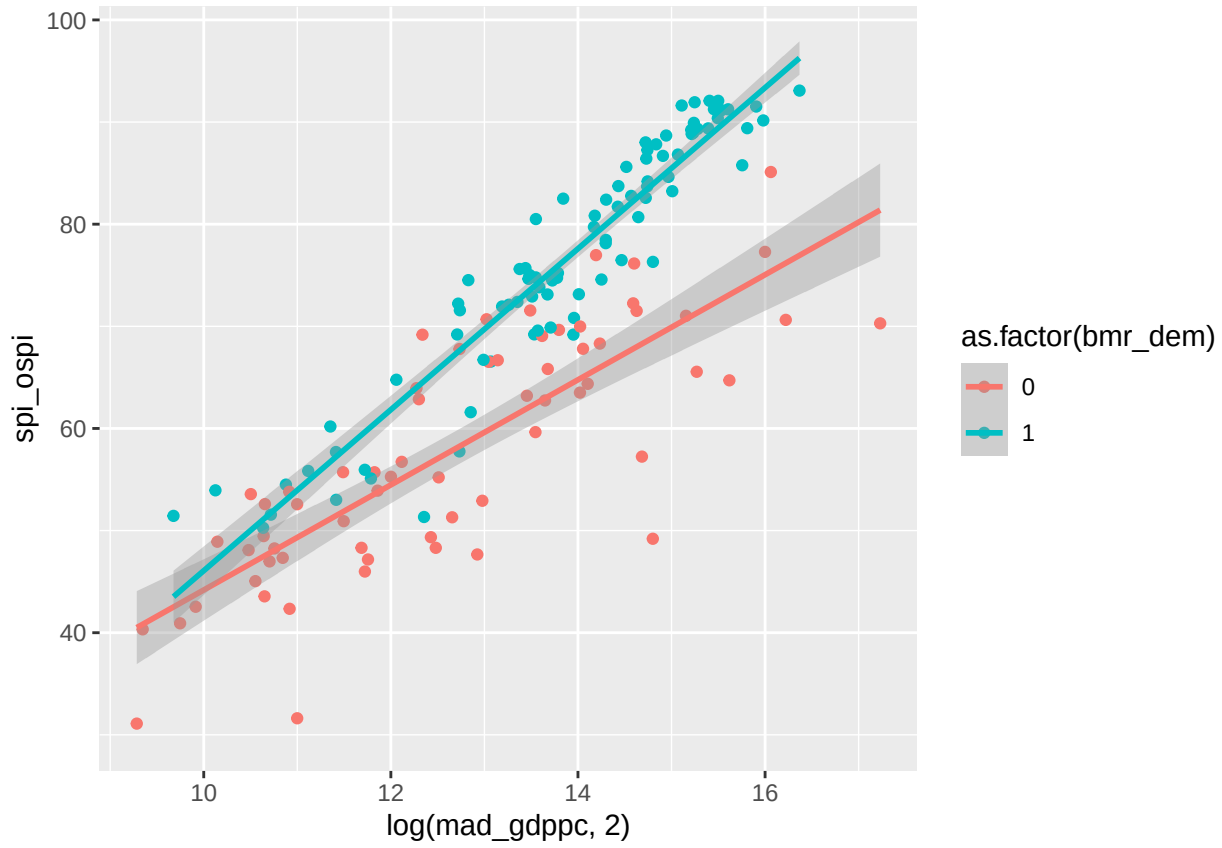
The LOESS model's R^2 value is more than double the linear model's R^2, so it is explaining much more of the variance, particularly in the 0-1000 mm precipitation range.

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.

```
lm3 <- lm(spi_ospi~log(mad_gdppc, 2) * as.factor(bmr_dem), countries)
summary(lm3)$coefficients
```

```
##                                        Estimate Std. Error   t value
## (Intercept)                           -7.268244  4.6416244 -1.565884
## log(mad_gdppc, 2)                      5.144979  0.3625680 14.190384
## as.factor(bmr_dem)1                  -25.480291  7.2101792 -3.533933
## log(mad_gdppc, 2):as.factor(bmr_dem)1  2.737121  0.5360962  5.105653
##                                          Pr(>|t|)
## (Intercept)                          1.194972e-01
## log(mad_gdppc, 2)                    1.854312e-29
## as.factor(bmr_dem)1                  5.456183e-04
## log(mad_gdppc, 2):as.factor(bmr_dem)1 9.921285e-07
```

```
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)
```




```
log(median(countries$mad_gdppc, na.rm = T))
```

```
## [1] 9.392261
```

At a log GDP per capita of 0 (an essentially useless interpretation), being a democracy is significantly associated (p-value $< 10^{-3}$) with a 25.5 point lower OSPI than non-democracies. However, a doubling in GDP per capita is associated with a 5.14 point increase in OSPI for non-democracies while it is associated with a 7.88 point increase in OSPI for democracies. Thus, at a GDP per capita of \$11,995 (the median of the data set), being a democracy is associated with a $2.74 \cdot \log(11995.19, 2) - 25.5 = 11.6$ point higher OSPI than non-democracies with the same GDP per capita.

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.

```
lm4 <- lm(spi_ospi~log(mad_gdppc, 2) + as.factor(bmr_dem), countries)
anova(lm4, lm3)
```

```
## Analysis of Variance Table
##
## Model 1: spi_ospi ~ log(mad_gdppc, 2) + as.factor(bmr_dem)
## Model 2: spi_ospi ~ log(mad_gdppc, 2) * as.factor(bmr_dem)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    150 5032.9
## 2    149 4283.5  1     749.4 26.068 9.921e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In an ESS F-test of $H_0 : \beta_{\log(\text{GDP per capita,2})\cdot\text{Democracy}} = 0$ (the interaction term provides no predictive power) vs $H_a : \beta_{\log(\text{GDP per capita,2})\cdot\text{Democracy}} \neq 0$ (the interaction term provides some predictive power), we get an F-statistic of 26.068 for an $F_{1,149}$ distribution, yielding a p-value of $< 10^{-6}$. Therefore, we reject the null and conclude that the interaction between democracy status and log GDP per capita adds predictive ability to the model. (Note that when we have a nested model where the only thing that changes is a single term, the ESS F-test should have the same significance as the t-test for that term by itself, as it does here.)

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?

We have k groups, but one will be set as the baseline (intercept), so we have $p = k - 1$ predictors not including the intercept.

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

The estimate $\hat{Y}_{ij} = \bar{Y}_i$ because the only non-zero values in $\vec{X_{ij}}$ for all $(\vec{X_{ij}}, Y_{ij})$ in group i are a 1 in the i^{th} position if $i \neq 1$ and a single 1 in the 1^{st} position. Therefore, the predicted value will have to be the same for all Y_{ij} in group i, and the prediction that minimizes the squared differences with respect to these observations is the mean of the group $\bar{Y}_i$.

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

Using the p we found earlier and the fact that $\hat{Y}_{ij} = \bar{Y}_i$,

$$\begin{aligned}\frac{\sum_{i,j}(\hat{Y}_{ij} - \bar{Y})^2/p}{\sum_{i,j}(Y_{ij} - \hat{Y}_{ij})^2/(n-p-1)} &= \frac{\sum_{i,j}(\hat{Y}_{ij} - \bar{Y})^2/(k-1)}{\sum_{i,j}(Y_{ij} - \hat{Y}_{ij})^2/(n-k)} \\ &= \frac{\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2/(n-k)}\end{aligned}$$

Simpson's simulation

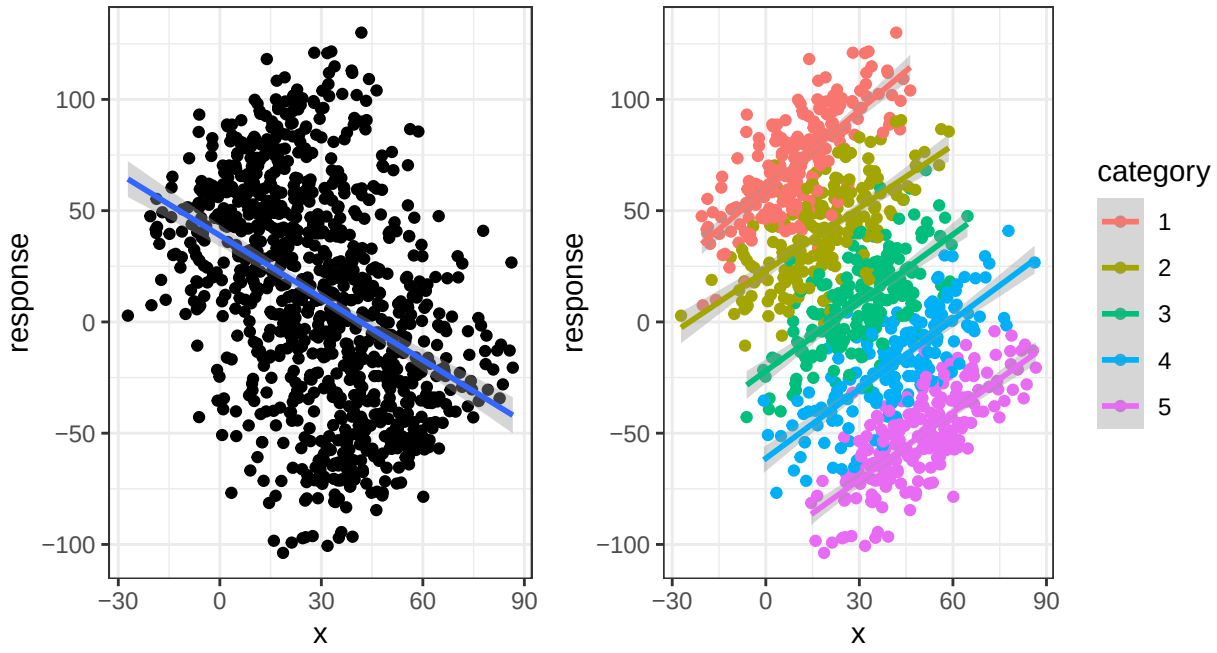
1. For the following data table, write out the design matrix that would be used in the following model: `response ~ category * value`.

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

```
lm_sim <- lm(Response ~ Category * Value, df)
model_matrix_out = model.matrix(lm_sim)
attr(model_matrix_out, "assign") <- NULL
attr(model_matrix_out, "contrasts") <- NULL

print(model_matrix_out)
```

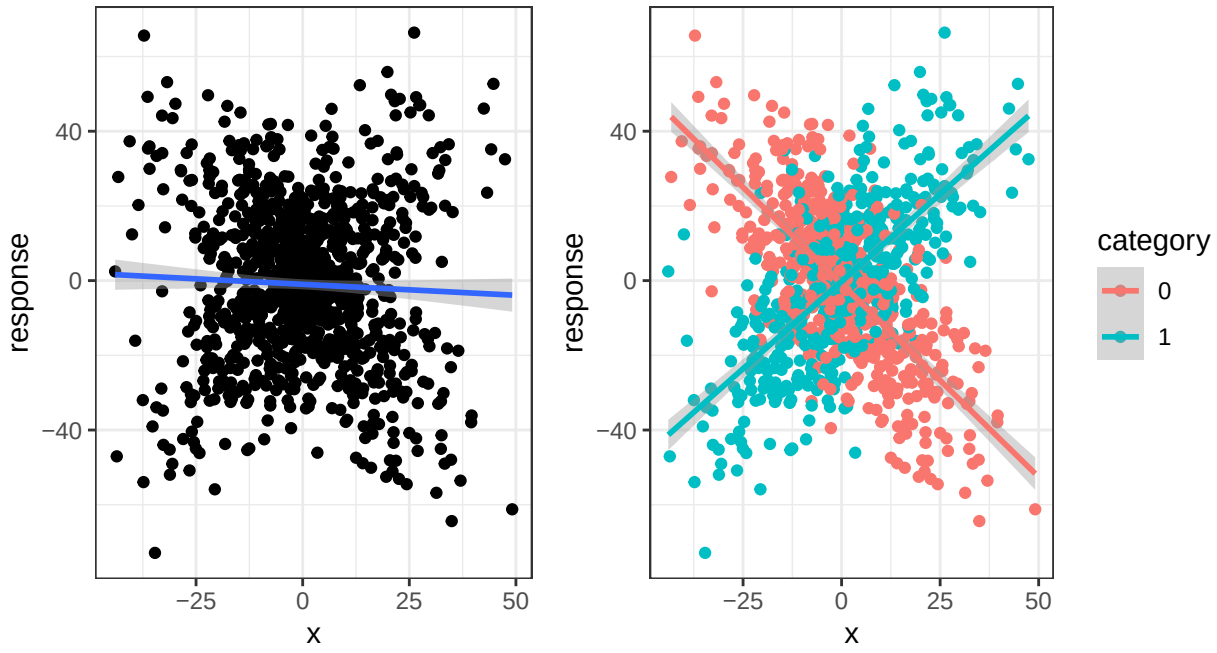
```
##    (Intercept) Category2 Category3    Value Category2:Value Category3:Value
## 1            1         0         1 5.125949        0.000000        5.125949
## 2            1         1         0 4.898613        4.898613        0.000000
## 3            1         0         1 2.031917        0.000000        2.031917
## 4            1         0         0 2.213955        0.000000        0.000000
## 5            1         0         0 5.348074        0.000000        0.000000
## 6            1         0         0 7.238691        0.000000        0.000000
## 7            1         1         0 2.962757        2.962757        0.000000
## 8            1         1         0 5.980036        5.980036        0.000000
## 9            1         1         0 5.333669        5.333669        0.000000
## 10           1         0         0 5.003350        0.000000        0.000000
```

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.




```
##                 Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)   60.927677 1.06427462  57.24808 5.902750e-317
## x              1.037964 0.03170177  32.74153 4.223252e-160
## category2    -39.774997 1.50142295 -26.49153 1.975928e-117
## category3    -83.485683 1.62395431 -51.40889 3.100370e-282
## category4   -122.327181 1.81854828 -67.26639  0.000000e+00
## category5   -163.597216 1.91984109 -85.21394  0.000000e+00
```

Because each group appears to have a similar slope but distinct intercept, we should use `response ~ x + category`.



```
##                Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) -0.8698512 0.68345942  -1.272718  2.034151e-01
## x           -1.0342164 0.04379353 -23.615733  2.859664e-98
## category1    0.6646968 0.97803741   0.679623  4.969011e-01
## x:category1  1.9686853 0.06218601  31.658008 9.201750e-153
```

Because each group appears to have a different slope, we should use `response ~ x * category`.

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

Fitting more terms (especially with little predictive power) increases overfitting and increases your estimated standard error (since $\hat{\sigma}^2 = \text{SSE}/(n-p-1)$). Therefore, you will lose significance for the things that actually matter, and you will reduce the ability of your model to predict on new data.