EXERCISE SET 3

1. Exercises

Exercise 1 (Testing for normality)**:** Here we consider the dataset that can be loaded with

```
> df = pd.read_csv("https://raw.githubusercontent.com
> /avehtari/ROS-Examples/master/Earnings/data/earnings.csv")
```

(a) Let Y denote the data from the column earn, which contains peoples earnings from this sample of adults in the US. Using this sample, estimate

$$P(Y > \mu_Y + 3\sigma_Y)$$

Recall that μ_Y and σ_Y are respectively the mean and standard deviation of Y.

(b) Based on your results from part (a), do you think the distribution of earnings is accurately captured by a Normal random variable? To help support your claim, make a plot showing samples from a Normal distribution compared to the earnings data. Note that there are different plots you could make which would show this (e.g. histogram, scatterplot). It is somewhat a matter of taste and I am intentionally leaving it up to YOU to decide which one.

(c) Repeat (a) and (b) with data from the height column. Do you think the variation in height data is accurately approximated by a Normal distribution?

Exercise 2 (Swapping response and predictor variables)**:** Consider the linear regression model with Normally distributed predictor:

$$X \sim \text{Normal}(\mu_X, \sigma_X^2)$$
$$Y|X \sim \text{Normal}(\beta_1 X + \beta_0, \sigma^2)$$

The goal of this problem is to understand the distribution of X conditioned on Y. This turns out to be a linear regression model with X as the response variable and Y as the predictor; meaning

$$Y \sim \text{Normal}(\mu_Y, \sigma_Y^2)$$
$$X|Y \sim \text{Normal}(\tilde{\beta}_1 X + \tilde{\beta}_0, \tilde{\sigma}^2)$$

where the parameters $(\mu_Y, \sigma_Y, \tilde{\beta}_1, \tilde{\beta}_0, \tilde{\sigma})$ can be expressed in terms of the original parameters $(\mu_X, \sigma_X, \beta_1, \beta_0, \sigma)$. Note that from class we already know how to express the marginal mean and variance of the response variable (denoted μ_Y and σ_Y respectively) in terms of these parameters (you should review this).

Understanding the relationship between the distribution of X conditioned on Y and that of Y conditioned on X is important because in some applications we have to make a choice about which variable to take as our response and which as our predictor. This exercise will help us understand how the regression parameters depend on this choice. It will also sharpen our understanding of what the covariance means. Note that if we suppose that if $\sigma^2 = 0$, then

$$Y = \beta_1 X + \beta_0 \implies X = \frac{1}{\beta_1} Y - \frac{\beta_0}{\beta_1}$$

so the slope of X vs. Y is $1/\beta_1$.

(a) Derive a formula for $\tilde{\beta}_1$ in terms of β_1, σ^2 and σ_X^2. Note that we already know (why?)

$$\text{cov}(X, Y) = \tilde{\beta}_1 \sigma_Y^2.$$

Double check the units! Hint: Use the formula for the marginal variance of Y and the fact that $\text{cov}(X,Y) = \text{cov}(Y,X)$ (interchanging the role of X and Y doesn't change the covariance).

(b) Discuss what happens to this formula as $\sigma^2 \to 0$ or $\sigma_X^2 \to 0$. Does the behavior of $\tilde{\beta}_1$ in these limits make sense intuitively?

(c) Why is the naive formula $1/\beta_1$ incorrect when $\sigma^2 > 0$? In particular, why can't we simply solve for X in terms of Y to obtain the regression equation? Hint: think about the assumption on ϵ in the linear regression equations.

(d) The quantity

$$\rho_{x,y} = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y}$$

is called the correlation coefficient. Show that

$$\rho_{x,y} = \text{sign}(\beta_1)\sqrt{\beta_1' \beta_1}$$

where $\text{sign}(\beta_1) = 1$ if $\beta_1 > 0$ and -1 otherwise.

Exercise 3 (Election model and prediction)**:** In this exercise you will work with some data on election outcomes and GDP growth. The data can be loaded with

```
> df = pd.read_table("https://raw.githubusercontent.com/avehtari/ROS-Examples/master
> /ElectionsEconomy/data/hibbs.dat",sep="\s+");
```

The columns are as follows

- **year:** Year of the election
- **growth:** A measure of economic growth during the previous four years.
- **vote share:** The vote share in percent for the incumbent.

(a) Fit the data to a linear regression model with economic growth as the predictor and the vote share as the response variable.

(b) Based on your fitted model and neglecting any uncertainty in your estimate, what is your best estimate of the chance that the incumbent will win the election after a period when the economic growth (as measured by the growth variable) is 1.

(c) What is the chance that after a period of economic growth = 2 the incumbent will win by a margin of at least 2%.