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

$$P(Y > \mu_y + 2\sigma_y)$$

where $\mu_y = E[Y]$ and $\sigma_y = \text{var}(Y)$ are the mean and standard deviation of the earnings (these will need to be estimated along the way).

(b) Based on your results from part (a), do you think the distribution of earnings is accurately captured by a Normal random variable?

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

Exercise 2 (Computation with normal variables)**:** Let

$$Z_1 \sim \text{Normal}(0, 1^2)$$
$$Z_1 \sim \text{Normal}(1, 4)$$

be independent.

(a) Using the estimates of Normal probabilities in class, compute the following:
- $P(Z_1 > 2)$
- $P(Z_1 + Z_2 < 6)$
- $P(Z_1 - Z_2 > 4)$

(b) Check your answers to part (a) with Monte Carlo simulations in python.

(c) (**ungraded – for additional practice**) Complete following exercises from [1]: 3.3.1, 3.3.2, 3.3.10

Exercise 3 (Central limit Theorem)**:** Suppose

$$U_i \sim \text{Uniform}(u_0 - L, u_0 + L), \quad i = 1, \dots, N$$

Note that, by symmetry, $E[U_i] = u_0$.

(a) Using simulations, confirm that [1]

$$\text{var}(U_i) = \frac{L^2}{3}.$$

In particular, make a plot of $\text{var}(U_i)$ as a function of L.

(b) Now consider the sum

$$S_N = \sum_{i=1}^{N} U_i$$

Use the CLT to approximate $P(S_N > u_0 + L/4)$ in terms of the cumulative distribution of a Normal random variable.

(c) Based on the binomial example in class, how does ratio between the actual probabilities and Normal approximation, that is,

$$\text{error} = \left| \frac{P(S_N > u_0 + L/4)}{\text{Prob. from Normal approximation}} - 1 \right|,$$

depend on L, u_0 and N? In particular, you should state whether the error will increase, decrease or stay the same when each of these numbers are increased with the other 2 fixed. Explain your reasoning.

(d) Test your predictions for part (b) with simulations.

(e) (**optional challenge**). Suppose that L depends on N via $L = h(N)$ where $h(N) \to \infty$ as $N \to \infty$. Therefore, the distribution of X_i depends on N and the CLT no longer applies. Find sequences $a_1, a_2, \dots$ and $b_1, b_2, \dots$ such that

$$P\left(\frac{S_N - a_N}{b_N} < z\right) \to P(Z < z)$$

for

$$Z \sim \text{Normal}(0, 1).$$

Exercise 4 (Model with conditional variance)**:** Consider the model

$$X \sim \text{Bernoulli}(q)$$
$$Y|X \sim \text{Normal}(a, X + 2(1 - X))$$

(a) Is this a linear regression model (for the variable Y) as defined in class? Are X and Y independent?

[1] If you know calculus you should be able to derive this.

(b) Compute $\mathrm{cov}(X, Y)$.

(c) Confirm your answer by making a plot of the covariance as a function of q from a samples of 10000 (x, y) points.

Exercise 5 (Swapping response and predictor variables $\star$)**:** Consider the conditionally normal model introduced in class

$$X \sim \mathrm{Normal}(\mu_x, \sigma_x^2)$$
$$Y|X \sim \mathrm{Normal}(\beta_1 X + \beta_0, \sigma_{y|x}^2)$$

This is a regression model for Y. The goal of this problem is to understand the distribution of X conditioned on Y. That is, we would like to understand the corresponding regression model for X. This 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 we infer depend on this choice. It will also sharpen our understanding of what the covariance really means.

For some additional motivation, suppose that there is no noise in $Y|X$ (meaning $\sigma_{y|x}^2 = 0$). 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 $1/\beta_1$. We could try adding a normal random variable $Z \sim \mathrm{Normal}(0, \sigma_{y|x}^2)$ to represent the noise in $Y|X$ and then solve this again. This would lead us to to

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

It is tempting to conclude that $Y|X$ follows a Normal distribution with mean $Y/\beta_1 - \beta_0/\beta_1$ and variance $\sigma_{y|x}^2/\beta_1^2$. This is however false – see part (c). In this problem you will derive the correct formula.

(a) Based on the formula for covariance derived in class, we know

$$\mathrm{cov}(X, Y) = \beta_1' \sigma_y^2.$$

where β_1' is the regression slope on X vs. Y and σ_y^2 is the marginal variance of Y. eUsing(1) $\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$ (interchanging the role of X and Y doesn't change the covariance) and (2) the marginal variance of Y is $\sigma_y^2 = \beta_1^2 \sigma_x^2 + \sigma_{y|x}^2$, derive a formula for β_1'.

(b) Using the result of part (a), show that when $\sigma_{y|x}^2 \to 0$ we retrieve the "naive" formula $\beta_1' = 1/\beta_1$.

(c) Why is the naive formula $1/\beta_1$ incorrect when $\sigma_{y|x}^2 > 0$? In particular, why can't we simply solve for X in terms of Y to obtain the regression equation? Hint: does $Y|Z$ have the same distribution as Y?

Exercise 6 (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 either incumbent will win by a margin of 2%.

References

[1] John Tabak. *Probability and statistics: The science of uncertainty*. Infobase Publishing, 2014.