EXERCISE SET 4

Exercise 1 (coefficient of determination and p-values)**:** A regression model with two predictors. Fix the parameter values and the distribution of the predictors (e.g. take them both to be standard Normal). Let N denote the number of samples. Now suppose we simulated data sets with increasing N.

(a) How do you expect the p-values to change as N increases? Test your answers by actually performing the experiment.

(b) How do you expect R^2 to change as N increases? Again, test your answers by actually performing the experiment.

(c) Briefly summarize (in your own words) what the implications of these observations are for how we should interpret p-values and R^2.

Exercise 2 (Earnings data revisited)**:** Consider the earnings data. This can be loaded with

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

As in the previous exercise set, you will study the association between earnings and gender, but now using regression with multiple predictors.

(a) Perform a linear regression using `statsmodels` with gender and height as predictors.

(b) Provide interpretations for each regression coefficient (like we did in class for the test score example).

(c) Which factor, height or gender is more important based on your analysis?

(d) Based one the fitted model, predict the chance that someone who is not male and is 5.8ft earns more than a male who is the same height? To get a sense for the importance (or lack-thereof) of the height predictor, compare this to the chance that a male earns more than a non-male (regardless of height).

Exercise 3 (A binary and normal predictor)**:** Consider the a linear regression model

$$Y|(X_1, X_2) \sim \text{Normal}(\beta_0 + \beta_1 X_1 + \beta_2 X_2, \sigma^2)$$

where the two predictors obey

$$X_1 \sim \text{Bernoulli}(q)$$
$$X_2|X_1 \sim \text{Normal}(bX_1, \sigma_{2,1}^2)$$

You can assume $\beta_0 = 0$ for this problem.

(a) Derive a formulas for $\text{cov}(X_1, X_2)$ and $\text{var}(X_2)$ in terms of the model parameters.

(b) Derive formula for the marginal mean (μ_Y) and marginal variance (σ_Y^2).

(c) Derive a formula for $\text{cov}(Y, X_1)$ in terms of β_1, q, β_2 and b.

(d) Explain how the formula you derived in part (b) is related to the equation for $\text{cov}(Y, X_1)$ in the single predictor regression model (page 4 on week 3 notes). In particular, for what parameter values do the two formulas coincide? Your conclusion will be a particular case of what we saw to be true more generally (see week 5 notes) concerning the relationship between β_1 and the covariances in a regression model with two predictions.

(e) The calculations in part (c) allows us to solve an exercise in Chapter 8 in Demidenko's textbook [1], albeit in the more restrictive context of a binary and normal predictor: Is it possible that β_1 and β_2 are **both negative**, yet the (marginal) slope of Y vs. X_1 is **positive**? If so, generate simulated data where this is the case.

Exercise 4 (Exercise vs. weight paradox)**:** The following data has data concerning body weight as a function of exercises intensity.

```
df = pd.read_csv("https://raw.githubusercontent.com
/eugenedemidenko/advancedstatistics/master/RcodeData/simpson.csv")
```

You can check that if we perform a regression using exercise intensity as a predictor and body weight as our response variable, the results suggests that exercise increases body weight, counter to most of our intuition. Using a regression analysis with multiple predictors, try to reconcile this. Explain how this is related to exercise 3 above.

Exercise 5 (Sample distribution)**:** In the colab notebook from class, there is code to generate samples from the sample distribution of $(\hat{\beta}_1, \hat{\beta}_2)$ in the model

$$X_1 \sim \text{Normal}(0, 1).$$
$$X_2 | X_1 \sim \text{Normal}(bX_1, 1 - b^2)$$
$$Y | (X_1, X_2) \sim \text{Normal}(\beta_1 X_1 + \beta_2 X_2, \sigma^2)$$

Specifically, we had a function which takes β_1, β_2 and β_0 as inputs and returns a dataframe where the columns are the samples of $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively. When we plotted the correlation coefficient as a function of b values and estimates the correlation coefficient between $\hat{\beta}_1$ and $\hat{\beta}_2$, it was a decreasing line.

(a) The model is set up so that as we vary b, the correlation between X_1 and X_2 varies, but the overall (marginal) variance in X_2 remains fixed. Show this is true mathematically and then test it with simulations.

(b) What would happen if instead of plotting the correlation coefficient, we plotted $\text{se}(\hat{\beta}_1)$ as a function of b? Would it increase? decrease? neither? Note that both X_1 and X_2 are standardized, so the distribution of X_1 values is not changed when we adjust b. In answering this question, you can either give a geometric intuition, or do a calculation. You should check your answer with simulations, but you still need to provide a detailed explanation.

(c) Is it possible for $\text{se}(\hat{\beta}_i)$ to be large for all the predictors (measured relative to $\hat{\beta}_i$), yet still have a large (meaning close to one) value of R^2? If not, explain why. If so, for what parameter values does this happen? Run simulation(s) to support your answer.

References

[1] Eugene Demidenko. *Advanced statistics with applications in R*, volume 392. John Wiley & Sons, 2019.