

EXERCISE SET 5

Exercise 1 (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.

- Derive a formulas for $\text{cov}(X_1, X_2)$ and $\text{var}(X_2)$ in terms of the model parameters.
- Derive a formula for $\text{cov}(Y, X_1)$ in terms of β_1 , q , β_2 and b .
- 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 predictors.
- (optional)** 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 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.

- Perform a linear regression using `statsmodels` with gender and height as predictors.
- Provide interpretations for each regression coefficient (like we did in class for the test score example).
- Which factor, height or gender is more important based on your analysis?
- 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 (Sample distribution): In the notebook from class, we wrote 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.

- 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.
- Is it possible to have large standard errors on all the $\hat{\beta}_i$ values (measured relative to the true values of course), but still have a large (meaning close to one) value of R^2 ? 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.