SOME ADDITIONAL PROBLEMS

Exercise 1 (Variance of correlated random variables)**:** Prove that for two random variables X and Y

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + \text{cov}(X, Y)$$

Exercise 2 (More on correlation)**:** Let

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

and let σ_y^2 and σ_x^2 be the marginal variances of Y and X respectively. As in class, we use ρ^2 to denote the squared correlation coefficient between X and Y; that is,

$$\rho^2 = 1 - \frac{\text{var}(Y|X)}{\text{var}(Y)}. \tag{1}$$

Recall that ρ is has a sign determined by the regression coefficient; that is,

$$\rho = \text{sign}(\beta_1)\sqrt{1 - \frac{\text{var}(Y|X)}{\text{var}(Y)}} \tag{2}$$

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

(a) Prove that

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \tag{3}$$

Hint: Use $\text{var}(Y)$ in as the common denominator to write Equation 1 as a single fraction, then use the formula for σ_y^2 in terms of β_1, σ^2 and σ_x^2 to simplify the numerator. At the end, notice that the sign of $\text{cov}(X, Y)$ is the same as β_1.

(b) Based on Equation 1, we can see that ρ is related to the fraction of variation in Y which is explained by X. Provide a brief explanation of why Equation 3 makes sense for a measure of correlation between two random variables.

(c) Now suppose the predictor follows a Normal distribution:

$$X \sim \text{Normal}(\mu_x, \sigma_x^2).$$

Using the formula derived on Exercise 3 Problem 5 (a), show that

$$\rho = \text{sign}(\beta_1)\sqrt{\beta_1 \beta_1'} \tag{4}$$

where β_1' is the $E[X|Y]$ vs. Y slope (also see Exercise set 3 for the definition of β_1').

(d) Show that

$$\beta_1 = \beta_1' \frac{\sigma_y^2}{\sigma_x^2} \tag{5}$$

Hint: You might have already shown this along the way to deriving the result of Exercise 3 Problem 5 (a), depending on how you approached that problem. If not, the starting point to deriving this formula is very similar to that problem.

(e) In words, Equation 5 is saying that the $E[Y|X]$ vs. X slope is the $E[X|Y]$ vs. Y slope multiplied by the ratio of the variances. Explain in words why this makes sense. To do so, you should be thinking about what certain quantities are very large or small, for example, if β_1 is very large but β_1' is small. It may also help to imagine what the data generated by different parameters would look like and/or perform some simulations.

Exercise 3 (Coin flipping puzzle)**:** Consider the problem "Random Bias" on page 192 of Pete Winkler's puzzle book, which is available in a pdf here.

(a) Formulate the puzzle in terms of a conditional probability model by writing down the number of flips in terms of a conditional distribution.

(b) Use simulations to estimate the answer to Pete's question and show that it agrees with his solution.

(c) (**challenge**) In the model you formulated in part (a), is the marginal distribution of the number of heads (meaning marginalized over p) binomial?

Exercise 4 (Mother and daughter heights)**:** In this problem you will work with some data containing mother and daughter heights. It can be loaded by the code

```
> df = pd.read_csv("https://raw.githubusercontent.com/avehtari/ROS-Examples/master
> /PearsonLee/data/Heights.txt",sep=' ', lineterminator='\n')
```

We will model the relationship between daughter heights (X) and mother heights (Y) with a linear regression model

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

We will assume these are the heights of both mother and daughters as adults.

(a) Fit the data to the linear regression model above using `statsmodels`. Take note of $\hat{\beta}_1$ and the value of R^2. What fraction of variation in women heights is explained by their mother's heights?

(b) Now flip the role of X and Y – that is, let the mother's heights be the predictor and the daughter heights be the response variable. Perform a regression again for this case and take note of the new regression coefficient $\hat{\beta}_1'$ and the R^2. R^2 should be the same as for part (a) (do you see why based on the formula for R^2 and/or its interpretation?). You should notice however that the regression coefficients are slightly different. Is this to be expected?

(c) A reasonable assumption is that the distribution of women's heights is constant from generation to generation, and therefore the distribution of X and Y are the same, therefore the marginal means and variance of Y are the same as for X: $\mu_y = \mu_x$ and $\sigma_y^2 = \sigma_x^2$. In this case, based on Equation 5, we would expect $\beta_1 = \beta_1'$. Show that the discrepancy between $\hat{\beta}_1$ and $\hat{\beta}_1'$ is explained by differences in the variance of mother and daughter heights.

Exercise 5 (The sample distribution of test statistic with unknown σ)**:** This exercise is related to hypothesis testing for a clinical trial example. Recall that if μ_T and μ_C are the average measurement of a quantity (e.g. blood pressure) for those in the control and treatment groups respectively. We are interested in whether $\mu_T = \mu_C$, or equivalently, whether $\beta_1 = \mu_C - \mu_T = 0$. To test this, we consider the test statistics

$$\hat{T} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$$

We are interested in the p-value, which is the probability that, if β_1 was zero, we would obtain a test statistic whose absolute value (when using the two-sided p-value) is greater than $|\hat{T}|$. It is therefore important to understand the distribution of the test statistic, at least if we are interested in better understanding p-values. In class we assumed that se$(\hat{\beta}_1)$ is known, the test statistic $\hat{T}$ has a Normal distribution. This is because in this case se$(\hat{\beta}_1)$ is known. If it is not then the sample distribution of $\hat{T}$ is influenced by randomness in both $\hat{\beta}_1$ and se$(\hat{\beta}_1)$.

(a) Generate 10000 replicates of the clinical trial model assuming the null hypothesis is true ($\beta_1 = 0$) with $N = 10$ people, and $\sigma = 1$. For each replicate compute a test statistic $\hat{T}_0$.

(b) Make a histogram of $\hat{T}_0$. Then compute the mean and standard deviation of $\hat{T}_0$ over all you replicates. Compare the histogram to a Normal random variable with this mean and variance. You can either do this by generating samples from a normal distribution and plotting the histogram on the same plot, or using the `pdf` function from `scipy`. Do they look different? You should experiment with different values of N to see how this changes the distribution.

(c) If you add the line of code `ax.semilogy()` after your plotting commands it will display your plot with a logarithmic scale on the y axis. These means that a distance of 1 unit on the y axis corresponds to a factor of 10 increases, so for example, going from 0.01 to 0.1 is the same distance as 10 to 100. It will therefore magnify the distribution towards the tail (large values of $\hat{T}_0$) where probabilities are small. What do you notice when you do this?

(d) What are the implications for the observations you made in the previous problems for two sided p-value

$$p_v = P(|\hat{T}_0| > |\hat{T}||\hat{T})$$

where $\hat{T}$ is the test statistic we get from data. In particular, if we do not know σ (and hence se$(\hat{\beta}_1)$), will we obtain a larger or smaller p-value? How does it depend on what test statistics we got? You can test your answer by making up a real test statistic $\hat{T}$ and computing the probability $|\hat{T}_0| > |\hat{T}|$.

Logistic regression questions. The following questions are to be done after watching the logistic regression video.

Exercise 6 (Understanding logistic regression with binary predictor)**:** Consider the logistic regression model where the predictor also follows a Bernoulli distribution.

$$X \sim \text{Bernoulli}(q) \tag{6}$$

$$Y|X \sim \text{Bernoulli}\left(\frac{1}{1+e^{-\beta_0+\beta_1 X}}\right) \tag{7}$$

(a) Compute cov(X, Y) in terms of the given model parameters: β_0, β_1 and q.

(b) Compare this formula to the formula for covariance in the context of linear regression. Do we see qualitatively similar relationships between the covariance, β_1, β_0 and the variance of X? It may help to plot the formula you derived in part (a) in Python.

Exercise 7 (On the relationship between logistic regression and linear regression with a binary predictor)**:** Consider the model

$$\begin{aligned} X &\sim \text{Bernoulli}(q) \\ Y|X &\sim \text{Normal}(\beta_0 + \beta_1 X, \sigma^2) \end{aligned} \tag{8}$$

Either using simulations or math, determine whether this becomes a logistic regression model if we look at $X|Y$. That is, is it true or not that

$$X|Y \sim \text{Bernoulli}\left(\frac{1}{1+e^{-\beta_0'+\beta_1' Y}}\right)?$$

for some choices of β_0' and β_1'. If not what is the difference between the logistic regression model with Normally distributed predictor the model given in Equation 8 expressed as a conditional model for X given Y?

Exercise 8 (Washington post data)**:** Consider the following data set form the Washington Post

```
> data= pd.read_csv("https://raw.githubusercontent.com/washingtonpost
> /data-police-shootings/master/v1/fatal-police-shootings-data.csv")
```

(a) Describe an appropriate application of logistic regression to this data set. You should state
- What is the response variable?
- What is the predictor?
- What conclusions could we draw from performing the logistic regression and why are they important?
- What do you expect to find?

(b) Now perform the logistic regression you just described and comment on the results.

Exercise 9 (Parent earnings and high school graduation)**:** This exercise is taken from [1, Exercise 13.9], although I've changed the wording. We are interested understanding the relationship between parent income (measured in units of \$$10k$ and denoted X) and the whether someone will graduate from high school (denoted Y with $Y = 1$ for someone who graduated). This is a natural candidate for a logistic regression model for $Y|X$. Now suppose we are told that for those whose parents have zero income, the chance to graduate highschool is 27%, while for those whose parents earn \$$60k$ the chance to graduate is 88%. What are you best estimates of β_0 and β_1 based on these numbers?

References

[1] Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and other stories*. Cambridge University Press, 2020.