ps4 — variability and bias

questions appended with "i" are optional

I — tests, power, and bias

A chemist wants to test whether a new drug has an effect on blood sugar levels. Suppose the mean blood sugar level in the general population is reported to be 95.0 mg/L with a standard deviation of 15.0 mg/L, and suppose that an observed blood sugar level of 100 mg/L would be considered significant for clinical purposes.

1. Suppose the chemist conducts a two-tailed test with $n = 30$ at the 5% level. What is the rejection region?

2. Suppose that the null hypothesis is true, i.e. the drug has no effect on blood sugar levels. Suppose however that the study participants are selected from a hospital, and unbeknown to the chemist, the patients of this hospital have blood sugar levels that are on average 3 mg/L higher than the general population. What is the expected blood sugar level in the biased sample? If the chemist conducted the test in q1 at the 5% significance level, what is the true type I error of the test? Compare this to the type I error rate the chemist believes the test has.

3. Now suppose that the null is false, and that the drug in fact causes blood sugar levels to rise by 6 mg/L on average. What effect does the sample bias in q2 have on the power of the test?

4. Suppose the chemist wants to produce an estimate of the average increase in blood sugar levels caused by the drug. The chemist computes a 95% confidence interval for mean increase in blood sugar level, using a sample size $n = 30$. If the chemist repeated the experiment with 100 different random samples, and assuming the same bias in each sample, what value will the intervals be centered at? Can you guess (very crudely is ok) how many of the intervals might contain the true value?

5. To get a more precise estimate, suppose the chemist had instead used a sample size $n = 500$. Can you guess approximately how many of the intervals might contain the true value?

II — estimator bias

This question will walk you through a simulation in R to understand the bias of an estimator. In the file `ps4.Rmd`, under the header "Estimator Bias", we have provided some code that generates a sample of data to simulate rolling a six-sided die n times. We have also written a function that evaluates the variance of a set of observations using the formula for population variance i.e. $\sigma^2 = \frac{1}{n}\sum_i^n (X_i - \mu)^2$.

1. Generate one sample of data with $n = 15$ and estimate the variance. Do this two ways: using the formula for sample variance (as provided by the `var()` function in R) and using the formula for population variance (as provided by the function we've written). What values does each estimator produce?

2. Repeat the experiment with many samples of data (keeping $n = 15$) and find the expected value of each estimator.

3. Note that the population variance of a uniform distribution between 1 and 6 is $\frac{35}{12} = 2.917$. Based on your results in q2, what is the bias of each estimator? Which estimator has more bias?

4. What is the standard deviation of each estimator? (use the `sd()` function). Which estimator has more variability, the biased one or the unbiased one? Explain your results.

5. Repeat q2, but change n to a much larger value (e.g. $n = 100$ or $n = 1000$). Now what is the expected value of each estimator? Compare the bias and variability of each estimator.

III — confounding variables

In the file `ps4.Rmd`, under the header "Funding Bias", there is a dataset characterizing the amount of government funding received by some members in a hypothetical population. Explore the data—you should see that each member belongs to one of two regional groups, *Austur* and *Vester*. Some members of the Austur group have complained that the Vester members unfairly receive more government funding than they do.

1. Perform a two-sample t-test between the mean funding levels of each regional group. What do you conclude, at the 5% significance level?

2. Construct a simple regression model, using regional group as a categorical predictor for funding. What does the model summary show? Is the predictor significant?

3. For the model in q2, how much of the variability in funding is explained by the model? Is the model overall useful?

4. Now construct a multiple regression model by adding the variables gender and age to the model. Report your results and comment on which of the predictors are useful to the model. How do the presence of the additional predictors affect the coefficient on regional group? Compare this model to the simple one in q2. Which model fits the data better? Why?

5. Based on your results in q4 and q5, what can you conclude about the claim made by the Austur group? What role does age play in this analysis?

6. By examining the relationships between variables in the data (e.g. by making scatterplots or boxplots), provide a possible explanation for why members of the Austur group might have been led to observe the association they claim. Explain why they are/not correct, and what sources of bias may/not have led to this conclusion.

IV — specification bias and log models

In this question we'll use the `gapminder` data to examine misspecified regression models. Make sure you have the gapminder package loaded. We've provided some code in the R file.

1. Fit a linear model predicting life expectancy from GDP per capita, using data from 2007. Report your results. Comment on the model's goodness-of-fit.

2. Use the `autoplot()` function from the `ggfortify` package to make diagnostic plots for the model in q1. Do you notice anything problematic? If so, explain why, taking into consideration the assumptions of least squares regression.

3. Make a scatterplot of life expectancy on GDP per capita. What do you notice about the distribution of data points?

4. Now make a scatterplot on life expectancy on log GDP per capita. Use the `log()` function. What do you notice about this scatterplot? How does it compare to the scatterplot in q3?

5. Now construct a linear model predicting life expectancy from log GDP per capita. Comment on the model's goodness-of-fit, and compare it to the model in q1. Does the result make sense? Explain your reasoning.

6. Make diagnostic plots for the log-transformed regression model. Compare these plots to those in q2. Comment on which of the two models is better, and explain why, referring explicitly to the assumptions of linear regression.

V — logistic regression

In this problem we'll use the endorsements dataset from the `fivethirtyeight` package. This dataset has information on each candidate in primary US elections between 1980 and 2012, and includes variables on party affiliation, proportion of endorsements received, proportion of money raised, proportion of the vote received in the primary, and whether or not that candidate won the primary that year.

In this section we'll build models to predict `won_primary` (whether or not the candidate won). Since this outcome variable is binary, we'll use a technique known as *logistic regression*. The goal in logistic regression is to predict the the *probability* of the outcome variable being a success:

$$\text{logit}\big(\text{P}(Y = 1)\big) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

In R we can do this with `glm(..., family = binomial)`.

1. Use the `glm()` function to build a logistic regression model predicting `won_primary` from `percentage_endorsement_points` and `percentage_of_money`.

In logistic models we typically exponentiate the regression coefficients—this will give the predicted increase in the probability of getting a success ($Y = 1$) for every unit increase in the predictor.

2. Exponentiate the coefficients. Based on your results, if we increase the percentage of endorsement points by 1%, how much does the likelihood of winning increase by, while holding percentage money constant?

3. Does this analysis tell us that receiving more endorsement points causes a candidate to win more votes in their primary? Why or why not?

4. Practically speaking, is it possible in the real world to hold one of these predictors constant while changing the other? Why or why not?

5. Suppose we could enrich this dataset by merging it with others to greatly increase the number of predictor variables and estimate models with almost perfect prediction accuracy. Asides from complicating the interpretation of the money/endorsement coefficients, give one other reason why this might be a bad idea.

Suppose now we want to investigate how party affiliation fits into the picture. To do this we're going to build an *interaction model*, to see how the `party` variable *interacts* with the two predictors we already have. An interaction effect exists if the effect of a predictor variable on a response variable changes depending on the value(s) an interaction variable. Given two predictors X_1 and X_2, and an interaction variable X_3, a logistic interaction model can be expressed as follows:

$$\text{logit}\big(\text{P}(Y = 1)\big) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_3 + \beta_4 X_2 X_3 + \beta_5 X_1 X_2 + \varepsilon$$

where the coefficient β_3 is the interaction effect of X_1 and X_3, and β_4 is the interaction effect of X_2 and X_3. These are the two interaction effects we're interested in. The interaction term between X_1 and X_2 is included as a control.

6*i*. We've given you some code that adds the `party` variable as an interaction term to the model in q1. Use `summary()` to look at the regression output. Since our interaction term is categorical, one of the categories has been chosen as a baseline. Which group is the baseline?

7*i*. The coefficient for the interaction term between the party variable and the money variable tells us the difference between the slopes for money in the baseline group and the non-baseline group. Is this difference statistically significant at the 5% level? What about for endorsements, does the non-baseline party have an endorsement slope that is significantly different at the 5% level?

8*i*. Similar to the F-test for nested linear models, we can do a test between nested logistic models—this is known as a *likelihood ratio test*. Using the `anova()` function, and specifying `test = "LRT"`, compare the interaction model to the simpler model in q1. Do we reject the null hypothesis that the simpler model is sufficient at the 5% level? What about at the 10% level?

Lastly, we've left some code to generate a plot—it shows the two predictor variables on the horizontal and vertical axes, the predicted probability of winning represented by color, and whether or not the candidate actually won represented by the shape of the point. We've also left code to generate a table showing predicted wins vs actual wins.

9*i*. Among candidates who won, what proportion were predicted to win? This is the *true positive rate* (TPR), or *sensitivity*.

10*i*. Among candidates who were predicted to win, what proportion actually won? This is the *positive predictive value* (PPV), or *precision*.

11*i*. Among candidates who were predicted to win, what proportion lost? This is the *false discovery rate* (FDR). ♣