Econ 108 FAll 2022
Problem Set 5

This problem set is due at 7pm, Friday, November 4th, 2022.

1. Suppose that using a data set with 10000 observations, a regression of household spending on indicators of broadband access and whether the household has any children produces the following output.

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	6.68508	0.3403		
broadband	0.65285	0.07357		
anychildren	0.08326	0.03570		

   (a) Fill in the last two columns of the above table.

   (b) If you are to test the null hypothesis H_0 : the coefficient on anychildren is negative or zero, against the alternative hypothesis H_1: the coefficient on anychildren is positive, at 1% size. Does the regression output indicate rejection or nonrejection of the null hypothesis?

   (c) How does your answer change if now you want to test the null hypothesis H_0 : the coefficient on anychildren is less than or equal to 0.05, against the alternative hypothesis H_1: the coefficient on anychildren is larger than 0.05, at 1% size.

2. Familiarize with the semiconductor.R code posted on canvas (in the modified_code directory under the files tab) before working on this question.

   (a) Using the spam.csv data, estimate a logit model using all the features to predict the probability of an email being spam. Report the in-sample R^2.

   (b) Plot the histogram of the P-values (excluding the intercept term). Does the histogram of the P-values suggest that some of the features might be useful signals for predicting spam probabilities?

   (c) Implement a Benjamin-Hochberg procedure for controlling the false discovery rate (FDR) at 10%. What is the resulting p-value threshold for declaring that a feature is a statistically significant signal? How many features are discovered by the Benjamin-Hochberg procedure? Examine these features. Do they (at least some of them) make sense as significant features for predicting spams?

(d) Call the model discovered by 10% FDR the cut model. In other words, the cut model includes all the features discovered by the BH correction. Call the initial model using all features the full model. How does the in-sample R^2 of the cut model compared to the in-sample R^2 of the full model?

(e) Implement a 10-fold cross validation method to calculate the out-of-sample R^2 of both the cut model and the full model. Draw a boxplot of the resulting collection of the 10 OOS R^2 for both the cut model and the full model. How do these two out-of-sample R^2 compare to each other? You can compare both the average OOS-R^2 across the 10 folds and their variability.

(f) Run a forward stepwise regression using the spam dataset. How many features are discovered by the forward stepwise regression? Are they a subset or a superset of the features discovered by the FDR method? How many features are discovered by both the stepwise regression and the FDR method?

3. We are very likely to use a final project for the end of quarter evaluation. Teamwork of up to two members will be allowed. You can start looking for a teammate. You can also decide to form a group of your own.