Homework Assignment 5

Max H. Farrell – Chicago Booth
BUS41100 Applied Regression Analysis

Due at the beginning of class of week 8

1 Pricing Experiment

Return to the pricing experiment we studied in class.

(a) Create the same 2×2 table that we made in class. Use this information to compute the intercept and slope of a **linear regression** of `buy` on `price`. Show your steps. Compare these steps to what we did for logistic regression in class.

(b) Run the logistic regression and obtain the intercept and slope. Compare the slope estimates, linear versus logistic, in sign and magnitude.

The data includes a variable `customerSize` that gives the size of the customer firm (remember, the customers of this business are themselves firms). The sizes are coded 0, 1, 2, for small, medium, and large firms.

(c) Create the 2×2 table of `buy` and `price` for medium sized firms. Compute the slope coefficient of a logistic regression for only this subgroup.

(d) Compare your answer above to the slope coefficient we found in class, which was for all firms.

(e) Use `glm` to run a logistic regression for only medium sized firms. What do the measures of statistical uncertainty tell you for this subgroup? Do the same for large firms.

(f) Include `price` and `customerSize` into one logistic regression, using all firms. That is, fit the model

$$\mathbb{P}[\texttt{buy} \mid \texttt{price}, \texttt{customerSize}] = \frac{e^{\beta_0 + \beta_1 \texttt{price} + \beta_2 \texttt{customerSize}}}{1 + e^{\beta_0 + \beta_1 \texttt{price} + \beta_2 \texttt{customerSize}}}.$$

Interpret the coefficient row for `customerSize`, both the estimates and the measures of uncertainty. Compare to what you found in the previous part, running subgroup-specific models.

(g) Now include the interaction of `price` and `customerSize`. Describe what you find and how it compares to the results from part **(f)**.

(h) Now treat `customerSize` as a factor variable. Run the main effects model, akin to **(f)** and the interactions model, akin to **(g)**. Comment on what you find here and how it compares to those prior parts.

2 Community Crime

The file `CommunityCrime.csv` has violent crime rates for 1994 communities across the US and 25 descriptive variables. Our goal is to find a good-performing parsimonious model to predict the log crime rate. The demographic variables include:

- `householdsize:` mean people per household
- `PctUnemployed:` % of people 16 and over unemployed

- `PctFam2Par`: % of families (w/ kids) having two parents
- `PctRecentImmig`: % immigrated within 3 years
- `PctHousOccup`: % of housing occupied
- `RentMedian`: rental housing – median rent
- `PctUsePubTrans`: % of people who use public transit

The rest of the 25 variables are similar, and can be interpreted from their names in the data.

First split the data into training and testing samples (size and sampling scheme is up to you).

(a) Using the training data, build a model for log crime rate by using forward stepwise selection guided by both AIC and BIC to search through all main effects.

(b) Redo **(a)** allowing for all possible interactions. What has changed?

(c) Still within the training data, use the LASSO to select a model from all main effects and interactions.

(d) Compute the BIC-based model probabilities for all the models found thus far, and the model which includes all main effects. Plot the fit for each model against the true value of log crime rate.

(e) Use the test data to compare out-of-sample MSE performance. Compare your results with what you found in **(d)**.

(f) Both BIC and LASSO rely on an assumption called *sparsity.* A regression is said to be "sparse " if a few variables matter a lot and the rest of the variables don't matter at all (for predicting Y). Formally, if we assume that we have p X variables (these p variables already include interactions, powers, etc) and that

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p,$$

then sparsity requires that most of the $\beta_k = 0$. The model selection problem occurs because we don't know which variables have coefficients of zero. If we knew that, we'd just delete all the useless variables from our data and proceed with the analysis.

But is sparsity always a reasonable assumption? This is kind of up to the user and the application. Sometimes, it makes more sense to assume that all the variables matter, but just a little bit.[1] We'll explore one technique for dealing with this situation called Ridge regression, another type of penalized regression. Ridge regression solves

$$\min \left\{ n^{-1} \sum (Y_i - \boldsymbol{X}_i' \boldsymbol{\beta})^2 + \lambda \sqrt{\sum_{j=1}^{p} |\beta_j|^2} \right\}.$$

(Compare to LASSO and the generic idea of penalized regression from the slides.) Without getting into many details, Ridge shares some of the good aspects of LASSO, but does not perform variable selection at all, it keeps all the variables in the model, but compared to the full model, the coefficients will be different.[2]

Use `cv.glmnet` to fit Ridge regression on all main effects and interactions. Compute the out-of-sample MSE and compare to **(e)**. *(You may want to consult* `?predict.cv.glmnet`*.)*

[1] Another common idea is that the variables matter in groups called "factors", where each contributes just a little bit to its factor, and there are a small number of factors. Any time you hear about factor analysis, principle component analysis, linear discriminant analysis, or singular value decomposition, this is usually what's going on.

[2] If you want to build some intuition on Ridge regression, you can redo the intuitive explanation of LASSO from class. In class, we drew the analogy between the LASSO penalty and a linear budget set, we drew the "diamonds" representing the budget set. To adapt this from ridge regression, just change these diamonds to circles around the point $(0, 0)$. See `http://faculty.chicagobooth.edu/max.farrell/bus41100/lasso_vs_ridge_regression.png` for a picture.

3 Hourly Bike Sharing

From Capital Bikeshare (D.C.'s Divvy) we have *hourly* bike rentals & weather info (the data used in week 5 was a collapsed version of this data). See `bikeSharing.csv`. The data are from 2011 & 2012, and we have the first 19 days of every month.[3] We have the number of rentals from registered and non-registered users (`registered` and `casual`, respectively). The available information is

- `datetime` – the date and hour, formatted as "2011-04-12 23:00:00" for 11PM, April 12, 2011
- `season` – Spring, summer, fall, winter
- `holiday` – Is the day a holiday?
- `workingday` – Is it a work day (not holiday, not weekend)?
- `weather` – coded 1=nice, 2=OK, 3=bad, 4=terrible
- `temp` – degrees Celsius
- `atemp` – "feels like" in Celsius
- `humidity` – relative humidity
- `windspeed`

(a) Examine the data for outlier patterns in the weather and remove them. Give clear justification for each data point removed.

(b) First split the data into training and testing samples (size and sampling scheme is up to you). We won't use the test set until the very end.

(c) Create dummy variables for each hour (24 total). Using your favorite model selection tool, figure out which are the most important hours of the day for predicting bike rental demand. Does the pattern vary between registered and non-registered users? Do the important hours change with the season? *(You may find the commands* `substr` *and* `model.matrix` *useful.)*

(d) Create a small set of dummies for logical groups of hours, such as commuting times, nighttime, etc. For both user types, first considering only main effects and second allowing for all possible interactions, select one model that uses the 24 dummy variables and one that uses your newly created groups.

(e) Use the test data to compare out-of-sample MSE performance of all the models you've considered. Discuss your findings.

4 Inference After Model Selection

This question illustrates conceptual material, and thus it has a lot of exposition.

In class we discussed model selection tools in the context of building a high-quality prediction model, but cautioned that statistical inference (testing, confidence intervals, etc) were unreliable following model selection. Let us review why. In week 2 we showed that our uncertainty regarding b_1 as an estimator of β_1 comes from the fact that if the data were to change, so would our estimate. The standard errors we derived, and those reported by the computer, capture this uncertainty. However, when doing variable selection, an *additional* layer of uncertainty is introduced: the fact that as the data changes, the very model we select may change, above and beyond to the coefficient estimates within the model changing. This second layer of uncertainty is not at all reflected in the standard errors.

[3] This data comes from a data-mining competition: the data for this problem is in fact the training data set, and the rest of the days were the test data, and hence the outcomes `registered` and `casual` were withheld by the company sponsoring the competition.

This question will explore inference after model selection and highlight some of the potential problems using Monte Carlo simulation, just like we did in week 2. The set up will be the simplest possible, just to make things easy and to illustrate the issues.

We have an outcome Y and two predictors, X_1 and X_2, in the standard multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \qquad \varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2). \tag{1}$$

Our goal is to do inference on β_1, the coefficient on X_1, for example forming a confidence interval, *after* deciding if X_2 belongs in the model. If X_2 belongs in the model and we run the *full* model, i.e. `lm(Y ~ X_1 + X_2)`, we get an estimator b_1 that is unbiased and Normally distributed with standard error $s_{b_1,F}$ (the "F" is for "Full"); that is, in math

$$\frac{b_1 - \beta_1}{s_{b_1,F}} \approx \mathcal{N}(0,1). \tag{2}$$

On the other hand, if X_2 does not belong in the model, then we will get a better estimate of β_1 by running the *restricted* model that only includes X_1, i.e. `lm(Y ~ X_1)`, and

$$\frac{b_1 - \beta_1}{s_{b_1,R}} \approx \mathcal{N}(0,1), \tag{3}$$

where, typically $s_{b_1,R} < s_{b_1,F}$, reflecting that the new estimate b_1 is more precise.

To form a confidence interval for β_1, we will perform the following steps.

(1) Somehow decide which model is better, F or R.

(2) Run the selected model, and obtain estimates b_1 and its standard error.

(3) Form the confidence interval $[b_1 \pm 2 \times s_{b_1,M}]$, where b_1 is the estimate of β_1 coming from the chosen model and $s_{b_1,M}$ is the standard error from that model (i.e. either $M = F$ or $M = R$).

We aim to answer questions like: What is the sampling distribution of the estimate b_1 from the algorithm above? Is it the same or different as the one we would expect from week 2? Why? In particular, how does the confidence interval behave?

To answer theses, you will use perform simulation studies from the model in Equation (1), using the code file `homework5-ModelSelectionMonteCarlo.R` from the course website. At the top of the code file, you will see a section where you can change the various parameters of the model:

```
## Sample size
  n <- 100
## set regression coefficients and other parameters
  beta.1 <- 2
  beta.2 <- 2
  x.cov <- 0  #the covariance between the two X variables
  sigmaSquared <- 1  #the variance of the epsilons
```

The code chooses a model based on the partial F test.

(a) Set $\beta_1 = \beta_2 = 2$ and $\mathbb{COV}(X_1, X_2) = 0.5$ and verify numerically that inference for β_1 works well even after model selection. Explain why the sampling distribution of b_1 is not affected by model selection in this case. Is the sampling distribution of b_1 given by Equation (2) or (3)?

(b) Set $\beta_1 = 2$, $\beta_2 = 1/4$, and $\mathbb{COV}(X_1, X_2) = 0$ and verify numerically that inference for β_1 works well even after model selection. Explain why the sampling distribution of b_1 is not affected by model selection in this case. Is the sampling distribution of b_1 given by Equation (2) or (3)?

(c) Set $\beta_1 = 2$, $\beta_2 = 0$, and $\mathbb{COV}(X_1, X_2) = 0.5$ and verify numerically that inference for β_1 works well even after model selection. Explain why the sampling distribution of b_1 is not affected by model selection in this case. Is the sampling distribution of b_1 given by Equation (2) or (3)?

Parts (a), (b), and (c) cover all the extreme cases of when model selection does NOT matter. In most other cases, it will affect inference to some degree. So the lesson from (a), (b), and (c) should NOT be "eh, it's mostly not an issue", but instead, "only in very particular cases can we ignore the problem."

(d) Set $\beta_1 = 2$ and $\beta_2 = 1/4$ like in **(b)**. Now run the code for $\mathbb{COV}(X_1, X_2) = 0.2, 0.5, 0.8$. What is going on? Can you explain what exactly is the distribution you see with $\mathbb{COV}(X_1, X_2) = 0.8$?

(e) Repeat part **(d)** with $n = 500$.

(f) Repeat part **(d)** with $n = 500$ and $\sigma^2 = 10$.

Part (e) makes it seem like this is just a "small sample size" problem. But that is NOT correct. It is true that the problem is a lack of information, and that a bigger n represents more information, but this neglects the "noise" part of the signal-to-noise ratio: the epsilons. For any sample size, the variance could be large enough to make this a problem, i.e. to mask the signal.

The small the signal is, the harder it is to detect, as we show next.

(g) Set $\beta_1 = 2$, $\beta_2 = 0$, and $\mathbb{COV}(X_1, X_2) = 0.5$ as in part **(c)**. Then try increase β_2 in steps of 0.05 and see what happens. Explain the pattern you find.

So far, we have only used the partial F test to select the model. The problem of inference after model selection is more general. To see this, let's try again with AIC and BIC focusing on one of the problematic cases.

(h) Set $\beta_1 = 2$, $\beta_2 = 1/4$, and $\mathbb{COV}(X_1, X_2) = 0.8$ and change the code to pick the model based on **BIC**. Using our discussion of BIC from class, explain what you find relative to part **(d)**.

(i) Set $\beta_1 = 2$, $\beta_2 = 1/4$, and $\mathbb{COV}(X_1, X_2) = 0.8$ and change the code to pick the model based on **AIC**. Explain what you find relative to parts **(d)** and **(h)**.

In this whole problem we just had two X variables. This isn't meant to be realistic; with two variables, you would just include both of them, i.e. not do model selection. The idea is to just illustrate the potential problems. You can imagine how bad the problem can be when you have hundreds or thousands of variables.