ECON 308: Econometrics
Assignment 4

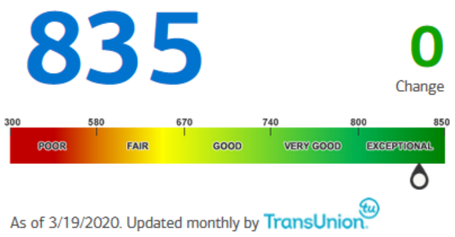
Complete each problem to the best of your ability and submit in class on Tuesday, November 23. You are encouraged to collaborate with other students, but you should turn in the problem solutions individually. Your writeup should include 1) written/typed responses to the questions, including regression tables where needed, 2) the code you ran to generate them (your do-file), and 3) any graphs produced.

1. What is the justification for using LOOCV to estimate the predictive accuracy of a model?

2. Under what circumstances is K-fold cross-validation preferable to LOOCV for model selection? When might AIC and BIC be preferable to either?

3. Predicting Credit Default: If you've ever used credit of any kind, or maybe even if you haven't, you have *credit scores*. These are numerical scores created by companies like TransUnion and Experian to measure your ability to generate profits for lenders. Using proprietary algorithms, they take information about your financial accounts, payment behavior, borrowing behavior, defaults, etc. to generate a summary score that assesses your ability to pay back money you've borrowed. These scores are not only used by lenders, but also landlords and (sometimes) employers (to determine if you are a person who should have a job). They look something like this:

   Figure 1



As of 3/19/2020. Updated monthly by TransUnion

   Here, we're going to use the model validation and variable selection tools we've learned about to formulate models for predicting consumer defaults, using data from 30,000 individuals in Taiwan.

   (a) Download the dataset `credit_default.dta` here:

   `https://www.dropbox.com/s/jfbi5f4zl7nq6db/credit_default.dta?dl=1`.

   (b) Create a do-file and load the dataset. The dataset includes 30,000 observations and 25 variables. They are:

- id: A unique identifier.
- limit_bal: Amount of credit given (i.e., credit limit).
- sex: Gender (1 = male; 2 = female).
- education: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- marriage: Marital status (1 = married; 2 = single; 3 = others).
- age: Age (in years).
- pay_0: History of past payment. The repayment status in September, 2005.
- pay_2,…, pay_6: Repayment status as of August, July,…, April, respectively.[1]
- bill_amt1,…, bill_amt6: Amount of bill statement in September, August,…, April, respectively.
- pay_amt1,…, pay_amt6: Amount of payment in September, August,…, April, respectively.
- defaultpaymentnextmonth: Indicator for whether or not the person defaulted in October.

The last variable will be the outcome we'll try to predict using all of the other variables (except id).

(c) Drop any observations with a 0 for `education` or `marriage`.

(d) Generate encoded versions of the categorical variables `sex`, `education`, `marriage`.[2] You can also create encoded versions of the `pay_` variables, if you'd rather treat them as categorical (rather than continuous, numerical variables).

(e) Generate some variables that you think might be helpful in predicting default, e.g., payment as a share of bill amount each month, or monthly bill amount as a share of credit limit, etc.

(f) Create two different models to predict default payment using subsets of the available predictors. Feel free to include any interactions or nonlinearities you like (though keep the outcome as an indicator variable). Compare them using in-sample RMSE, AIC, and BIC. Which seems to work better? Why do you think that is?

(g) Using `runiform()`, generate a random variable that is uniformly distributed between 0 and 1. Then, generate a variable that equals 1 if the previous variable is greater than 0.5. I.e.,

```
gen rando = runiform()
gen ind = (rando>0.5)
```

This will effectively split your sample in half.

[1] From the codebook, the measurement scale for the repayment status is: -1 = pay duly, 1 = payment delay for one month, 2 = payment delay for two months,…, 8 = payment delay for eight months, 9 = payment delay for nine months and above. This goes for pay_0 as well. -2 isn't mentioned as a possible value, but it shows up in the data, as does 0. I'm guessing 0 indicates on-time (after posting) payment, while -1 and -2 indicate pre-posting on-time payment. Just a guess though. This is not unusual when you're downloading random datasets online.

[2] Since these are numeric, Stata won't let you encode them. So, turn them into strings first. For example: `tostring sex, replace`. Then you can encode them as normal. Why does Stata make you do this? I have no idea.

(h) Estimate your two previous models on just one-half of the sample. Then, use `predict` to generate predicted values of the outcome for the entire sample. Compare the MSE of your predictions on the portion of the sample you used to estimate the model (the training MSE) to the MSE computed using the rest of the data (the test MSE). Which model worked best on each?

(i) Finally, use the `lasso linear` command to perform variable selection on a rich model with many predictors. What are some predictors that appear to be important? What cross-validated MSE do you get?

4. Consider the following hypothetical data on potential outcomes and treatment status D (1 for treated, 0 for untreated).

Student	Y_1	Y_0	δ	D
1	8	9		0
2	9	5		0
3	5	6		1
4	8	5		1
5	7	2		1
6	1	1		0

(a) Fill in the individual (unit) treatment effects (δ).

(b) Find the average treatment effect (ATE).

(c) Find the average treatment effect for the treated (ATT). Why might this differ from ATE?

(d) Find the average treatment effect for the untreated (ATU).

(e) Find the naive average treatment effect ($NATE$).
(Recall: $NATE = ATE + \mathrm{E}[Y^0 \mid D = 1] - \mathrm{E}[Y^0 \mid D = 0] + (1 - P(D = 1)(ATT - ATU)$.)

(f) Compute the overall bias in the $NATE$ and decompose this into selection bias and heterogeneous treatment effect bias.

(g) Under what assumption does $NATE = ATE$? State this formally using potential outcomes notation.