Lecture 1: Review and Simulation Methods

Chris Conlon

February 1, 2021

NYU Stern

Introduction

Consider a linear regression:

$$Y_i = X_i'\beta + \varepsilon_i \quad \text{with} \quad \mathbb{E}[\varepsilon_i | X_i] = 0$$

We've discussed the least squares estimator:

$$\widehat{\beta}_{ols} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i'\beta)^2$$

$$\widehat{\beta}_{ols} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}$$

Regression "Fit"

How "well" does this regression perform?

- $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}$: fraction of variance explained by X_i (and the fraction explained by ε_i).
- Alternative: mean squared error (MSE) $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$.
  - This is of course what least-squares is actually minimizing!
- Alternative: root mean squared error (RMSE) $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$.
  - The average distance from a point to the line of best fit.
- Alternative: mean absolute deviation (MAD) $\frac{1}{N}\sum_{i=1}^{N}(|y_i - \hat{y}_i|)$.
  - The average residual.
- Alternative: median absolute error (MAE) median $(|y_i - \hat{y}_i|)$.
  - The median residual (insensitive to outliers).
- If you read enough econometrics papers, you will see enough of these.

Regression "Fit" (continued)

- Nearly all of those measures will improve as we add parameters to the model
- If we choose the model with the lowest $RMSE$ or highest R^2 we will nearly always choose a model with more parameters!
- We might be worried about overfitting: choosing a regression model that fits our particular sample (y_i, x_i) well but might not perform well on a new but similar sample.
- A common solution is penalization

Penalized Regression

$$\min_{\beta} \sum_{i=1}^{N} (y_i - X_i\beta)^2 + f(\beta)$$

Idea if β has too many nonzero elements, or elements are too large – increase the penalty:

- AIC and BIC set $f(\beta)$ as penalty in terms of number of nonzero elements of β the so called L_0 norm.
- Lasso penalizes the L_1 norm $\sum_{k=1}^{K} |\beta_k|$.
- Ridge penalizes the L_2 norm $\sum_{k=1}^{K} |\beta_k|^2$.
- We will talk about penalization later, but this prevents us from selecting models that are "too complicated".

Splitting the Sample

Researchers in Machine learning split their sample into (2-3) parts

Training Data this is where we use (y_i, x_i) and estimate $\widehat{\theta}$

Validation Data this is where we choose which parameters to include, or how much penalization to apply using out of sample fit (MSE, etc).

Test Data this is completely new data where we compare the fit of various approaches (but change nothing).

The goal is to avoid overfitting with too complicated models with low bias but high variance.

Cross Validation

Other ways to evaluate fit involve using hold out samples.

- ie: Estimate $\hat{\theta}$ using 80% of the training data. Use the other 20% (test data) to predict $E[y_i|x_i, \widehat{\theta}] = \hat{y}_i$ and compute MAD, MSE, etc.
- Of course the 20% we initially withheld was arbitrary. So we could repeat the exercise dropping a different 20% each time. Fit the parameters $\widehat{\theta}$ on the training data, and use validation data to select model complexity.
- These are known as folds and the overall procedure is what is known as cross validation. Often 5-fold 10-fold, k-fold, etc.
- An extreme version is LOOCV (leave one out) which is similar to the jackknife.

Monte Carlo Studies

A common technique to assess performance:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

- Draw x_{i1}, x_{i2} from some distribution (say $U[0,1]$).
- Draw u_i from some distribution (with mean zero or subtract $u_i - \overline{u}$).
- Choose some coefficients $(\alpha, \beta_1, \beta_2) = (3, -2, 1)$.
- Calculate y_i.
- Regress y_i on (x_{i1}, x_{i2}) and obtain $\widehat{\beta}$.
- Calculate goodness of fit statistics, standard errors, etc.
- Do this around 1000×

Monte Carlo Studies

What's the point? Now we have a bunch of $\widehat{\beta}_s$ samples:

- How do asymptotic standard errors compare to $Var(\widehat{\beta}_s)$?
- How does R^2 change when we add variance to u_i? to x_i?
- How does variance change with sample size N?
- What is coverage of 95% CI?

Now we try it ourselves.

Thanks!