Module 2 – Regression and Prediction

CASE STUDY ACTIVITY TUTORIAL

Case Study 1: Predicting Wages 1

Regression 1.4. Case Study: Predicting Wages

V. Chernozhukov
MIT

June 19, 2016

Predicting Wages

Our goals are

1. Predict wages using various characteristics of workers.
2. Assess the predictive performance using adjusted MSE and R^2, and out-of-sample MSE and R^2.

Predicting Wages

Our goals are

1. Predict wages using various characteristics of workers.
2. Assess the predictive performance using adjusted MSE and R^2, and out-of-sample MSE and R^2.

In this segment, we will examine a real example, where we will predict workers' wages using a linear combination of workers' characteristics, and we will assess the predictive performance of our prediction rules using the adjusted mean squared error and r-squared as well as out-of-sample MSE and R^2.

Data

- Data is from the March Supplement of the U.S. Current Population Survey, year 2012.
- Focus on the single (never married) workers with education levels equal to high-school, some college, or college graduates.
- The sample is of size $n \approx 4{,}000$.
- The outcome Y is hourly wage, and X are various characteristics of workers.

Data

- Data is from the March Supplement of the U.S. Current Population Survey, year 2012.
- Focus on the single (never married) workers with education levels equal to high-school, some college, or college graduates.
- The sample is of size $n \approx 4,000$.
- The outcome Y is hourly wage, and X are various characteristics of workers.

Our data comes from the March Supplement of the U.S. Current Population Survey, year 2012. We focus on the single (never married) workers with education levels equal to high-school, some college, or college graduates. The sample size is about 4, 000.
Our outcome variable Y is hourly wage, and our X 's are various characteristics of workers such as gender, experience, education, and geographical indicators.

Descriptive Statistics

	Mean
Wage	15.53
Female	0.42
Experience	13.35
College graduate	0.38
Some college	0.32
High school graduate	0.30
Midwest	0.29
South	0.24
West	0.21
Northeast	0.26

Regression 1.4

Case Study

Descriptive Statistics

Descriptive Statistics

	Mean
Wage	15.53
Female	0.42
Experience	13.35
College graduate	0.38
Some college	0.32
High school graduate	0.30
Midwest	0.29
South	0.24
West	0.21
Northeast	0.26

The following table shows some descriptive statistics... From the table we see that average wage is about 15 dollars per hour... 42% of workers are women... average experience is 13 years;...38% are college graduates... 32% have done some college work, and 30% hold only high school diploma. You can also see geographical distribution of workers across major geographical regions of the states.

Predictive Models

- **Basic Model:** X consists of the female indicator (D) and other controls W, which contain a constant, experience, experience squared, experience cubed, education indicators, and regional indicators. X includes $p = 10$ regressors.

- **Flexible Model:** X consists of D as well as W, which contains all of the components of W in the basic model plus their two-way interactions. An example of a regressor created through a two-way interaction is experience times the indicator of having a college degree; another example is the indicator of having a high-school diploma times the indicator of working in the "north-east" region. X includes $p = 33$ regressors.

Predictive Models

- **Basic Model:** X consists of the female indicator (D) and other controls W, which contain a constant, experience, experience squared, experience cubed, education indicators, and regional indicators. X includes $p = 10$ regressors.
- **Flexible Model:** X consists of D as well as W, which contains all of the components of W in the basic model plus their two-way interactions. An example of a regressor created through a two-way interaction is experience times the indicator of having a college degree; another example is the indicator of having a high-school diploma times the indicator of working in the "north-east" region. X includes $p = 33$ regressors.

We consider two predictive models, basic and flexible.

In the basic model, regressors X consist of the female indicator D and other controls W, which include a constant, experience, experience squared, experience cubed, education and regional indicators. The basic model has 10 regressors in total.

In the flexible model, regressors consist of all regressors in the basic model PLUS their two-way interactions or products. An example of a regressor created by a 2-way interaction is the experience variable times the indicator of having a college degree; another example is the indicator of having a high-school diploma times the indicator of working in the "north-east" region. The flexible model has 33 regressors.

Performance of Predictive Models

- Since p/n is small, the sample linear regression should approximate the population linear regression well.
- We expect the sample R^2 to agree with adjusted R^2 and be a good measure of out-of-sample performance.

Performance of Predictive Models

Performance of Predictive Models

- Since p/n is small, the sample linear regression should approximate the population linear regression well.
- We expect the sample R^2 to agree with adjusted R^2 and be a good measure of out-of-sample performance.

Given that p/n is quite small here, the sample linear regression should approximate the population linear regression quite well.
Accordingly, we expect the sample R^2 to agree with the adjusted R^2 and they should both provide a good measure of out-of-sample performance

Assessing Predictive Performance 1

	p	R^2_{sample}	R^2_{adj}	MSE_{adj}
basic reg	10	0.09	0.09	165.68
flex reg	33	0.10	0.10	165.12

We conclude that the performance of the basic and flexible model are about the same, with the flexible model being just slightly better (slightly higher R^2_{adj} and lower MSE_{adj}).

Assessing Predictive Performance 1

Assessing Predictive Performance 1

	p	R^2_{sample}	R^2_{adj}	MSE_{adj}
basic reg	10	0.09	0.09	165.68
flex reg	33	0.10	0.10	165.12

We conclude that the performance of the basic and flexible model are about the same, with the flexible model being just slightly better (slightly higher R^2_{adj} and lower MSE_{adj}).

The following table shows the results for the basic and flexible linear regressions. The sample and adjusted R^2 are close to each other for both basic and flexible models. We also see that the predictive performance of the basic and flexible regression models is quite similar, with adjusted MSEs and R^2 not being very different from each other. The flexible model is performing just a tiny bit better, having slightly higher adjusted R^2 and slightly lower adjusted MSE.

Assessing Predictive Performance 2

	p	R^2_{test}	MSE_{test}
basic reg	10	0.08	129.21
flex reg	33	0.11	118.71

Here we report results for one random split of the data in 2 halves, and see that the flexible rule works just slightly better.

Note that these numbers vary across different data splits, so we can average results over several data splits.

By looking at results for several splits, we conclude that the basic and flexible model perform about the same.

Assessing Predictive Performance 2

Assessing Predictive Performance 2

	p	R^2_{test}	MSE_{test}
basic reg	10	0.08	129.21
flex reg	33	0.11	118.71

Here we report results for one random split of the data in 2 halves, and see that the flexible rule works just slightly better.

Note that these numbers vary across different data splits, so we can average results over several data splits.

By looking at results for several splits, we conclude that the basic and flexible model perform about the same.

Next we report the out-of-sample predictive performance measured by the test MSE and test R^2. Here we report the results for one random split of the data into the training and testing sample.... The numbers reported actually vary across different data splits, so it is a good idea to average the results over several data splits. By looking at the results for several splits, we can conclude that the basic and flexible models perform about the same.

Conclusion

- Using a real example, we have assessed predictive performance of two linear prediction rules.
- Next we will proceed to discuss the Inference Problem.

Conclusion

- Using a real example, we have assessed predictive performance of two linear prediction rules.
- Next we will proceed to discuss the Inference Problem.

In this segment... using a real example, we have assessed predictive performance of two linear prediction rules. They both performed similarly, with the flexible rule performing slightly better out-of-sample. Next ... we will proceed to discuss the inference problem.

Module 2 – Regression and Prediction

THANK YOU

Case Study 1: Predicting Wages 1