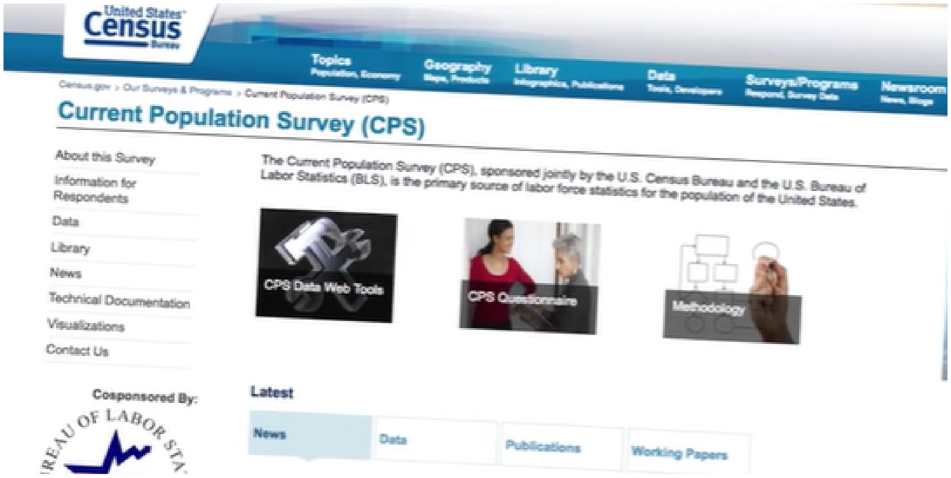
Predicting Wages

In this segment, we will look at 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 MSE and R^2, as well as Out-Of-Sample MSE, and R^2.

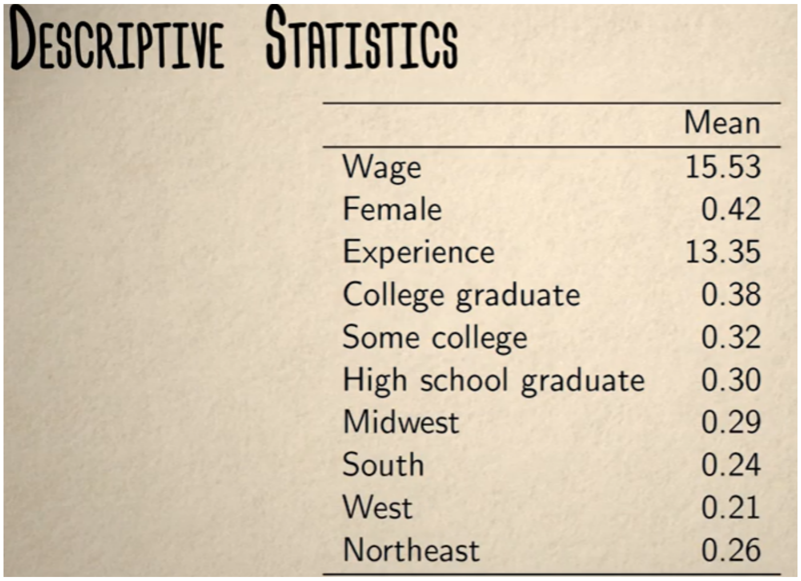
Our data for this example comes from the March supplement of the US Current Population Survey for the year 2012.



The Current Population Survey (CPS), sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics (BLS), is the primary source of labor force statistics for the population of the United States.

We focus on the single workers with education levels equal to high school, some college, or college graduates. The sample size is about 4000. **Our outcome variable Y is hourly wage** and X is a vector of various characteristics of the workers, such as gender, experience, education, and geographical indicators - the independent variables.

The following table shows some descriptive statistics about the dataset.



Descriptive Statistics

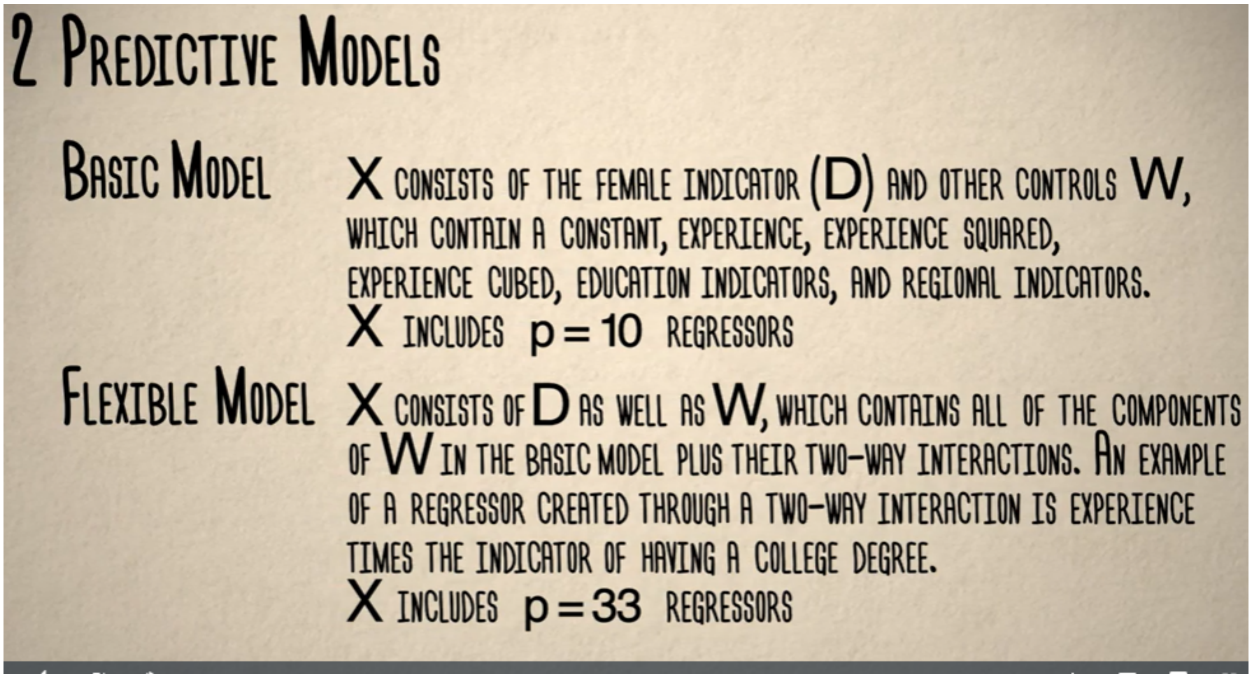
	Mean
Wage	15.53
Female	0.42
Experience	13.35
College graduate	0.38
Some college	0.32
High school graduate	0.30
Midwest	0.29
South	0.24
West	0.21
Northeast	0.26

From the table, we see that the average wage is about $15 per hour, 42% of workers are women and the average experience is over 13 years. “**College graduate**”, “**Some college**” and “**High school graduate**” are variables that capture the highest education level of the workers - 38% of the workers are college graduates, 32% have done some college work as their highest level of education, and 30% of the workers only possess a high school diploma, with no education beyond that.

The remaining variables, “**Midwest**”, “**South**”, “**West**” and “**Northeast**”, also give an idea of the geographical distribution of workers across the United States.

Now, we consider two predictive models - a basic model and a flexible one.

The basic model has 10 regressors.
The flexible model, on the other hand, consists of 33 regressors.



2 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.
X includes $p = 33$ regressors

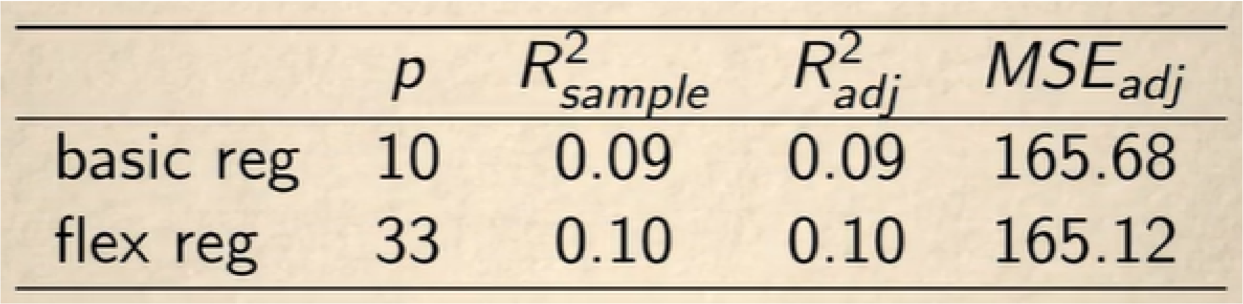
P/N - the ratio of the number of features (in this case, regressors) to the number of samples in the dataset, is an important dataset metric to keep track of for machine learning on tabular datasets.

The P/N ratio for this dataset is quite small since the number of samples in the dataset (N ~ 4000) is quite large. In such scenarios, 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 sample performance.

The following table shows the results for the basic & flexible linear regression models.



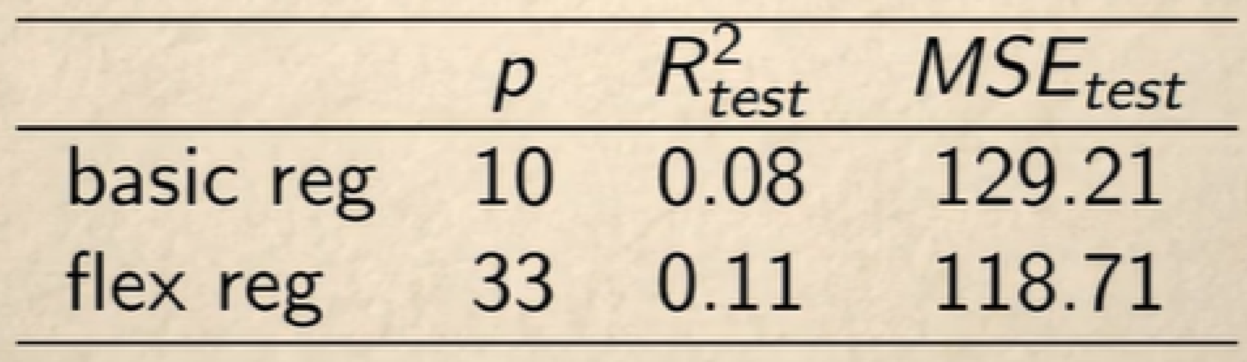
	p	R^2_{sample}	R^2_{adj}	MSE_{adj}
basic reg	10	0.09	0.09	165.68
flex reg	33	0.10	0.10	165.12

As we see from the above table, **the sample and adjusted R^2 are nearly identical** to each other for both the basic and flexible models.

We also see that **the predictive performance** of the basic and flexible regression models **is quite similar**, with the adjusted MSE and R^2 values being not very different from each other either.

The flexible model seems to be performing just a tiny bit better, having a **slightly higher adjusted R^2 and a slightly lower adjusted MSE.**

Next, we report the out-of-sample predictive performance measured by the test MSE and test R^2.



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

In this case, we are reporting 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's 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.

In this segment, using a real example, we have assessed the predicted performance of two linear prediction rules. They both performed similarly, with the flexible model performing slightly better for out-of-sample data.