

POL 502
INTERMEDIATE STATISTICS
Lecture 1

January 27th, 2005

Regression Analysis

What kind of data can we use for regression?

Dependent variable: must be either interval or at least ordinal with many categories.

Independent variable can be ordinal, interval, or dummy.

The only thing that the independent variable cannot be is a nominal variable with more than 2 categories.

Remember the equation of a line from the 8th grade?

$$Y = mX + b$$

We use:

$$Y = \beta_0 + \beta_1 X$$

Y is the dependent variable.

X is the independent variable.

β_0 is the Y intercept.

β_1 is the slope of the line also called the *coefficient*.

The coefficient tells us how much Y will change based on a 1 unit increase in X.

If I know any 3 of these, I can easily calculate the fourth.

Relationship between X and Y is constant over the course of the line.

Meaning the slope is constant.

What kind of relationship is that?

Linear.

Celsius and Fahrenheit, perfect linear relationship.

How can I calculate slope?

$$\frac{\Delta Y}{\Delta X}$$

I can summarize the relationship between Celsius and Fahrenheit by saying:
when its 0 Celsius its 32 Fahrenheit and when its 1 Celsius its 33.8 and on and on.

Or, I can just offer the equation of the relationship.
 $Y=32+1.8X$

What is the coefficient?

It means that for every one unit increase in X, Y will increase by 1.8.

For every one unit change in X (degree Celsius increase), there is a 1.8 increase in Y (Fahrenheit increases).

So, given any X_i , I can predict Y_i .

But social science data never look as good as those lines.

We get data points that are scattered.
A scatter plot reveals a glimpse of the relationship.

Look at the baseball data set in MINITAB.

What are some relationships we could look at or hypothesize about?

We can see the basic relationship looking at these graphs.

Graph \rightarrow plot \rightarrow $Y = \text{win}$, $X = \text{ba}$

This shows the relationship between batting average and winning percentage.

But how do we know as precisely as possible what the relationship is?

We want to fit the best line possible through the mess of points.

Which is the best line?

The one that misses least.

For now, enough to say that a straight line does not go through our points exactly.

So, our Y variable may look to depend on X, but not perfectly.

Y is not only a function of X, it is a function of some ERROR, ε (epsilon)

What is in Epsilon?

All the things that we left out.

Measurement error.

Some information lost by making the relationship linear.

Randomness of the world.

$Y = \beta_0 + \beta_1 X_1 + \varepsilon$ is our basic model.

$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$ tells us what we expect for the i^{th} observation of Y.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

Expands to include additional independent variables.

So what is Y_i ?

β_0 ?

β_1 ?

X_{1i} ?

β_2 ?

X_{2i} ?

β_3 ?

X_{3i} ?

ε_i ?

We can estimate a regression equation in MINITAB for the Baseball data set.

Stat → Regression → Regression → Response = win, predictor = ba, era, fielding.

$$\text{win} = 2.65 + 4.03 \text{ ba} - 0.128 \text{ era} - 2.70 \text{ fielding}$$

This tells us that if I know a team's Batting Average, Earned Run Average, and Fielding Percentage I can predict what their winning percentage will be.

When batting average goes up by .1, winning percentage goes up by .4.

When a team allows 1 more run per game, winning percentage goes down by 0.128.

To keep things a bit simpler, stick to one independent variable for now.

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

We will learn how to estimate β_0 , β_1 etc.... we call these parameters.

Once we estimate β_0 and β_1 we can take a value of X and get our predicted value of Y.

Our estimates of β_0 and β_1 we call Beta-Hats, $\hat{\beta}$

Means that it is a predicted value.

When we use our predicted values of β_0 and β_1 and a value for X_i we get a predicted value for Y_i – which is called \hat{Y} .

\hat{Y} is on the regression line.
Its our best guess of Y given a value of X.

We can estimate an equation from our sample data for the Baseball data set.

$$Win = \beta_0 + \beta_1 BA + \beta_2 ERA + \varepsilon$$

Get the following equation:

$$win = 0.048 + 3.75 ba - 0.122 era$$

$$\hat{\beta}_1 = 3.75$$

$$\hat{\beta}_2 = -0.122$$

We can now make predictions based on these estimates.

We could estimate the winning percentage the Mets should have had based on their ERA and BA.

Given their BA of .249 and their ERA of 4.07, we would expect the Mets to have won what proportion of their games?

$$\widehat{Y}_{Mets} = 0.048 + 3.75(.249) - 0.122(4.07)$$

$$\widehat{Y}_{Mets} = 0.048 + 0.93375 - 0.49654$$

$$\widehat{Y}_{Mets} = 0.485$$

In fact, the Mets' winning percentage was 0.506.

Using just two independent variables we got pretty close.

We can get the predicted winning percentage for every team.
Stat → regression → regression, click “storage” and check “Fits”

Now it will save Y-hats (also known as “fitted values”) in the last column.

Note, there is no error term to deal with here.

We get our best estimates of β_0 , β_1 , β_2 and the value of X_{1i} , X_{2i} and use those to get a prediction.

Remember, we do have a data value of Y_i .

What's the difference between Y_i and \hat{Y}_i -- that is, what is the difference between what we expect to see and what we actually see?

The error:

$$e_i = Y_i - \hat{Y}_i.$$

Note that our model discusses theoretical concepts, $\beta_0, \beta_1, \epsilon_i$

Those are not the same thing as the estimates we get for our data:

$$\hat{\beta}_0, \hat{\beta}_1, e_i$$

The first set of numbers describes our idea about the relationship between the variables. But the second set is produced by our actual observations and will change if our observations change.

Try the Page 20 example in Minitab.

Enter the Heights and Weights for 20 observations.

Review:

Regression analysis explains and predicts Y in terms of independent variables X 's.

We use the data to get estimates about the *parameters* including the *coefficients* that tell us how much Y changes based on a one unit change in X .

We can use these to get predicted values for Y , that is \hat{Y} .

We will try to have the best measured variables and the most important variables but we will never be able to make exactly correct guesses about the values for Y because the world has randomness in it.

Our guesses may be close but there will always be a little bit of error.

Homework:

Studenmund Exercises at the end of Chapter 1: 1, 2(if we didn't cover it in class), 5 & 6.