

POL 502
January 31st, 2005

Lecture 2 – From Chapter 2 in Studenmund

Ordinary Least Squares

We discussed last time that even though we can get good predictions of Y using X and regression techniques, our predictions (\hat{Y} s) will not be exactly right.

There is always some error, e.

The best regression equation comes when we *minimize* those errors.

That is, the best regression equation is the one that misses the least.

(Draw dots about a regression line)

Label the dots, Y_1, Y_2, Y_3, \dots

For each there is a corresponding \hat{Y} .

And remember what \bar{Y} is?

Mean of Y =

$$\frac{1}{n} \sum_{i=1}^n Y_i$$

Variance of Y, S^2

$$\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

So we have variation in Y, our dependent variable.

And what is the purpose of our research?
- to explain why Y varies.

There is movement about the mean for Y and we use values of X to predict a slope that tells us what to expect for values of Y.

Among the infinite number of lines, the best one will be, the one that misses the least.

What are the misses?

The distance from the line to the observations.

e

To minimize the spread, we think of the vertical distances between points
This distance is the error in our prediction.

Remember that our regression equation gives us the equation of a line.

To get the \hat{Y} for any case we use this equation – so all \hat{Y} s fall on the regression line.

Obviously, the observations do not.

We want to minimize this error.

So that the positives and negatives don't cancel each other out, we want the squared distances.

Called the Residual Sum of Squares

$$RSS = \sum (Y_i - \hat{Y}_i)^2$$

and $Y_i - \hat{Y}_i$ is simply the error, e.

$$\text{So, } RSS = \sum_{i=1}^n e_i^2$$

This is an aggregate measure of how much the line's predicted Y, \hat{Y} , differs from the actual observed values of Y_i .

The regression line is the line with the smallest RESIDUAL Sum of Squares.

We call this LEAST SQUARES.

And this technique, Ordinary Least Squares (OLS)

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

Again, β_1 is the **regression coefficient**

Tells us how much of an increase in Y we expect for a one unit increase in X.

It is:

$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

Once I have an estimate of $\hat{\beta}_1$ I can use it and the means of X and Y to get the constant, $\hat{\beta}_0$.

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Remember that:

$$SS_{xx} = \sum (X_i - \bar{X})^2$$

$$SS_{yy} = \sum (Y_i - \bar{Y})^2$$

$$SS_{xy} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

$$\text{So, } \beta_1 = \frac{SS_{xy}}{SS_{xx}}$$

Lets use a rigged data set.

height weight

60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150

Which is the dependent variable?

This can be easily done in MINITAB

Take a few minutes and calculate β_0 and β_1 for the height/weight example.

What will we need to figure out?

$$X_i \quad Y_i \quad (X_i - \bar{X}) \quad (X_i - \bar{X})^2 \quad (X_i - \bar{X})(Y_i - \bar{Y})$$

What is SS_{xx} ?

SS_{xy} ?

Note that the regression line always passes through the mean of X and Y.

What does this line mean?

If someone is height of zero?

If someone is height of 50 what is their predicted weight?

Height of 60?

β_1 is positive.

positive slope, positive relationship.

If it is negative, if it is zero.

Our hypothesis is about β_1 – slope in population – is it significantly different from zero?

We will get to this in a while.

How good is the fit?

We can have the same equation with a good fit or bad fit

- equation itself doesn't tell us about the fit.

Think about how much we are getting right and how much we are getting wrong.

Think about regression line.

Gives prediction of Y_i based on value of X_i .

Without X, we would just predict \bar{Y} .

Regression line allows us to improve upon that.

Draw mean of Y and regression line.

Distance from \bar{Y} to \hat{Y} is *explained*.

Distance from \hat{Y} to Y_i is error - *residual*.

Explained: $\frac{ESS}{TSS}$

$$\text{ExplainedSS} = \sum (\hat{Y} - \bar{Y})^2$$

$$\text{ResidualSS} = \sum (Y_i - \hat{Y})^2$$

$$\text{TotalSS} = \sum (Y_i - \bar{Y})^2$$

(Total SS is the same thing as SS_{YY})

We can make a table of these, called ANOVA table

$\frac{RSS}{TSS}$ is the proportion of error relative to the total spread.

$$\begin{aligned} R^2 &= \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \\ &= 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \end{aligned}$$

for this data set,

$$R^2 = 6000/10,426 = .58$$

Decomposition of Variance in Y.

Y varies around its mean.

$$Y_i - \bar{Y}$$

Can separate this into two parts:

- (1) $\hat{Y}_i - \bar{Y}$: the difference between the estimated value of Y and the mean value of Y.
- (2) $Y_i - \hat{Y}_i$: the difference between the actual value of Y and the estimated value of Y.

$$(1) + (2) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) = -\bar{Y} + Y_i = Y_i - \bar{Y}$$

EXAMPLE

Perform regression analysis to find the equation that tells us the relationship between the % spent on arms and the number of conflicts.

$$\bar{X} = 35$$

$$\bar{Y} = 2.5$$

	X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	Y_i	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
Canada	10	-25	625	1	-1.5	37.5
sweden	5	-30	900	0	-2.5	75
us	35	0	0	3	.5	0
ussr	40	5	25	4	1.5	7.5
argentina	30	-5	25	2	-.5	2.5
iran	50	15	225	4	1.5	22.5
israel	70	35	1225	5	2.5	87.5
japan	5	-30	900	0	-2.5	75
iraq	80	45	2025	6	3.5	157.5
brazil	25	-10	100	0	-2.5	25
			<u>6050</u>			<u>490</u>

$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$= \frac{490}{6050}$$

$$= 0.081$$

$$\bar{Y} = \beta_0 + .081\bar{X}$$

$$2.5 = \beta_0 + .081(35)$$

$$\beta_0 = -0.335$$

$$Y_i = -.335 + .081X_i$$

This tells us that the number of conflicts is equal to $-.335 + .081$ times the percentage spent on arms.

Positive relationship.

How many conflicts would we expect from a country that spends 50 percent on arms?

Substitute 50 for X and solve = 3.715

That spends 0?

-0.335

For each case I can get a prediction – based on the amount spent on arms, how many conflicts would I predict for each country?

Again, for each case I just substitute the X value into my regression equation.

These are my \hat{Y} s .

	Actual # of Conflicts	Predicted # of Conflicts (\hat{Y} s)
Canada	1	0.475
sweden	0	0.070
us	3	2.500
ussr	4	2.905
argentina	2	2.095
iran	4	3.715
israel	5	5.335
japan	0	0.070
iraq	6	6.145
brazil	0	1.690

If I didn't know anything about X, and you asked me to guess the value of the dependent variable for a case, my best guess would be the mean of Y, 2.5.

Look how much better my predictions are now.

The amount my predictions improve compared to just guessing the mean of Y, is the explained portion.

The amount I am still wrong by is the residual.

Draw the graph.

Regression line passes through the mean of X and Y.

Calculate R^2
= ESS/TSS

$$\text{ExplainedSS} = \sum (\hat{Y} - \bar{Y})^2$$

$$\text{ResidualSS} = \sum (Y_i - \hat{Y})^2$$

$$\text{TotalSS} = \sum (Y_i - \bar{Y})^2$$

	$\hat{Y} - \bar{Y}$	$(\hat{Y} - \bar{Y})^2$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
Canada	-2.02479	4.0998	0.52479	0.27541
Sweden	-2.42975	5.9037	-0.07025	0.00493
US	0.00000	0.0000	0.50000	0.25000
USSR	0.40496	0.1640	1.09504	1.19912
Argent	-0.40496	0.1640	-0.09504	0.00903
Iran	1.21488	1.4759	0.28512	0.08130
Israel	2.83471	8.0356	-0.33471	0.11203
Japan	-2.42975	5.9037	-0.07025	0.00493
Iraq	3.64463	13.2833	-0.14463	0.02092
Brazil	-0.80992	0.6560	-1.69008	2.85638

The Explained Sum of Squares (what MINITAB calls the “regression”) = 39.686.

The Residual Sum of Squares (MINITAB calls this the “residual error”)= 4.814.

The Total Sum of Squares = 44.5.

$$R^2 = \frac{ESS}{TSS} = \frac{39.686}{44.5} = 89.2$$

Or, we can calculate it another way:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{4.814}{44.5} = 89.2$$

This means the using X we have explained 89.2% of the variance in Y.

Try all this in MINITAB.

HOMEWORK

Exercises from Chapter 2:

1, 4, 6a-c, 10, 12 (for calculus explanation of OLS that might help if you are familiar with calculus).

Enter the data from the extended example above into MINITAB.

Run the regression and confirm the results we obtained by hand.