

Leftovers from Chapter 2 of Studenmund

Example

Given the Y values 1,2,3,4,5,6,7,8,9

I choose a Y_i : What is the best guess about this value of y for a given observation?

Choose the mean, $\bar{Y} = 5$.

What is the Total Sum of Squares for Y?

i.e. What are the square misses I will make if I always guess the mean?

$$-4^2 + -3^2 + -2^2 + -1^2 + 0^2 + 1^2 + 2^2 + 3^2 + 4^2 = 60$$

We don't use the simple distances from the mean because they would cancel out.

$$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2 = 60 \quad (1)$$

Can call this the variation in Y or the Total Sum of Squares for Y (TSS) or SS_{yy} .

Y is our dependent variable.

Our whole research design is about explaining this variation in Y.

Y could be the amount of alcoholic beverages consumed in a week, the number of bills that make it through a committee in a day, the number of days in a month that it rains....

To explain Y, we add some information, an Independent variable, X.

Observations #	Y_i	X_i
1	1	12
2	2	13
3	3	11
4	4	15
5	5	14
6	6	19
7	7	18
8	8	17
9	9	16

We will use these values of X to predict Y.

Assume the relationship is linear of the form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (2)$$

I will estimate $\widehat{\beta}_0, \widehat{\beta}_1$, substitute them into the equation and use the equation to predict values of Y_i .

What is the slope? Rise over the run....

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (3)$$

What is the Y-intercept? Also called the constant.

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X} \quad (4)$$

See where this comes from? Rearranging (2)

Why can we just substitute in the means of X and Y?

We could use any point on the regression line – but we happen to know that the regression line passes through the point where these means meet.

Lets begin by calculating $\widehat{\beta}_1$.

Obs #	Y_i	X_i	$Y_i - \bar{Y}$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$	\hat{Y}	$\hat{Y}_i - \bar{Y}$	$(\hat{Y}_i - \bar{Y})^2$
1	1	12	-4	-3	9	12	2.699	-2.301	5.295
2	2	13	-3	-2	4	6	3.466	-1.534	2.353
3	3	11	-2	-4	16	8	1.932	-3.068	9.413
4	4	15	-1	0	0	0	5	0	0
5	5	14	0	-1	1	0	4.233	-0.767	0.588
6	6	19	1	4	16	4	8.068	3.068	9.413
7	7	18	2	3	9	6	7.301	2.301	5.295
8	8	17	3	2	4	6	6.534	1.534	2.353
9	9	16	4	1	1	4	5.767	0.767	0.588

$$\bar{Y} = 5 \quad \bar{X} = 15$$

$$\sum (\hat{Y}_i - \bar{Y})^2 = 35.298$$

$$\sum_{i=1}^n (X_i - \bar{X})^2 = 60$$

$$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = 46$$

$$\hat{\beta}_1 = 0.767 \quad (46/60)$$

Substitute into (4) all the values we now know:

$$\begin{aligned} \hat{\beta}_0 &= 5 - 0.767(15) \\ &= -6.505 \end{aligned}$$

Giving us our estimated regression equation:

$$\hat{Y}_i = -6.505 + 0.767X_i$$

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

This is the equation that minimizes the Residual Sum of Squares (RSS)

$$\text{RSS} = \sum e_i^2 \quad (5)$$

So now I have an equation to give me a better guess about what Y should be given a value of X....

I can calculate my expectations for each Y_i that is, each \hat{Y}_i .

Calculate the \hat{Y}_i s – see table above.

The sum of $(\hat{Y}_i - \bar{Y})^2$ tells us how much of the distance from the mean we have explained.

Next, the squared distances from the \hat{Y}_i s to the actual observations Y_i s tells us how much we are still missing by.

Obs #	Y_i	\hat{Y}	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$
1	1	2.699	-1.699	2.887
2	2	3.466	-1.466	2.149
3	3	1.932	1.068	1.141
4	4	5	-1	1
5	5	4.233	0.767	0.588

6	6	8.068	-2.068	4.277
7	7	7.301	-0.301	.091
8	8	6.534	1.466	2.149
9	9	5.767	3.233	10.452

$$\sum (Y_i - \hat{Y}_i)^2 = 24.724 \quad \text{THIS IS RESIDUAL SS} \quad (6)$$

$$\sum (\hat{Y}_i - \bar{Y})^2 = 35.298 \quad \text{THIS IS EXPLAINED SS} \quad (7)$$

$$\sum (Y_i - \bar{Y})^2 = 60 \quad \text{THIS IS TOTAL SS} \quad (8)$$

Total Sum of Squares = Explained Sum of Squares + Residual Sum of Squares.

Total = what we explain plus what we do not explained.

$$60 \approx 35.298 + 24.724$$

And what do we call the proportion of what we explain?

$$R^2 = ESS/TSS = 1 - RSS/TSS \quad (9)$$

R^2 tells us how good a fit the regression equation is.

The equation itself can't do this – we can have the same equation with a good or bad fit

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

Multiple Regression

Adding additional independent variables to the regression equation.
One variable will not explain everything for our dependent variable.
We want to enhance our explanation.

Can look at the relationship between two variables when accounting for the presence of a third or fourth....

E.g., part of explaining income is education but what else matters?

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i \quad (4.10)$$

K is the number of independent variables in the model.

How do we decide whether or not to add another independent variable?

(1) Depends partly on the number of cases we have, N.

- (2) Should add some new explanation.
- (3) Fewer is better.

We want to add variables that explain new parts of the dependent variable.

Variance in the dependent variable. movement around the mean.
We want to explain it as best as possible.

Use our best X.

New variables that we add should have new value, not too closely related to variables we are already including.

If Independent Variables are too closely related, **multicollinearity** is a problem.

Why?

Theoretically, model should explain as much as possible with as few variables as possible.

Don't want to include variables twice.

Here is a regression model with 2 independent variables:

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

$$\widehat{\beta}_0 = \bar{Y} - (\widehat{\beta}_1 \bar{X}_1 + \widehat{\beta}_2 \bar{X}_2) \quad (11)$$

$$\widehat{\beta}_1 = \frac{(\sum yx_1)(\sum x_2^2) - (\sum yx_2)(\sum x_1x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2} \quad (12)$$

$$\widehat{\beta}_2 = \frac{(\sum yx_2)(\sum x_1^2) - (\sum yx_1)(\sum x_1x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2} \quad (13)$$

Where the lower case letters y, x_1, x_2 indicate deviations from the mean.

$$y = Y_i - \bar{Y}$$

$$x_1 = X_{1i} - \bar{X}_1$$

$$x_2 = X_{2i} - \bar{X}_2$$

$\widehat{\beta}_1$ is the change in Y given a one unit increase in X_1 and holding all else constant.
 $\widehat{\beta}_2$ is the change in Y given a one unit increase in X_2 and holding all else constant.

These formulae are obviously huge!

Not necessary to calculate by hand the $\widehat{\beta}_1, \widehat{\beta}_2$

But it is useful to look at the formulae and see that each $\widehat{\beta}$ is influenced by the relationship between the X and between Y and the other independent variable in the model.

The closer the relationships are between the Xs, the more problematic are our estimates of $\widehat{\beta}$ s.

The closer the relationships between Y and X_2 , the smaller our estimate of $\widehat{\beta}_1$ becomes.

This implies that if we have two independent variables, X_1 and X_2 ,

Estimating $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ (14)

Will give different values of $\widehat{\beta}_1, \widehat{\beta}_2$ than would estimating

$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$ and $Y_i = \beta_0 + \beta_2 X_{2i} + \varepsilon_i$ separately.

Because (14) takes into account the relationship between X_1 and X_2 .

Adjusted R^2

Even if it has no real explanatory power – *even if it is randomly generated by a computer!* – adding an additional independent variable will never lower R^2 .

If we rely too heavily on R^2 to tell us if our model is good (and we will learn why this is *never* a good idea), we may be tempted to accept additional, irrelevant independent variables.

Adjusted R^2 helps us avoid doing this by pushing down R^2 for each independent variable we add.

$$\overline{R}^2 = 1 - (1 - R^2) \frac{(n-1)}{(n-K-1)} \quad (15)$$

K is the number of additional independent variables beyond 1 and the constant.

Chapter 2 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.

Generate a graph that plots Y and X as well as \hat{Y}_i and $\hat{\beta}$

And Especially this one:

In the Stateenvironment data set choose an interesting dependent variable and 2 independent variables that you think should explain it well.

Run a regression.

What is the regression equation you get?

Interpret the values of $\hat{\beta}$ s that you get.

Overall, how is your model fit?

Chapter 3: Learning to Use Regression Analysis

6 steps in regression analysis.

1. Review the literature and develop the theoretical model.
2. Specify the model: select the independent variables and the functional forms.
3. Hypothesize the expected signs of the coefficients.
4. Collect the data.
5. Estimate and evaluate the equation.
6. Document the results.

1. Review the literature.

The basis of what you are trying to show comes from literature and not from the results. Nothing in the regression analysis tells you that X is influencing Y. Could just as easily be Y influencing X.

Your theory specifies the relationship.

2. Specify the model: select the independent variables and the functional forms.

$$Y = f(X_1, X_2, X_3)$$

The dependent variable and the cases selection are part of your research design.

The operationalization of the independent variables is important.

Which ones to choose should be based on theoretical beliefs.

Functional Form: we will use linear equations almost exclusively.

But you may have reason to think this is not appropriate.

3. Hypothesize the expected signs of the coefficients.

Does the theory tell you to expect a positive or a negative relationship?

Possible that you can't say which, "exploratory research."

$$Y = \beta_0 + \beta_1^+ X_1 + \beta_2^- X_2 + \beta_3^+ X_3 + \varepsilon$$

4. Collect the data.

Try to get as many cases as possible but be sure the sample represents the population you are interested in.

- at some point additional cases aren't very useful.

4B. Discuss the Data

Good reports will have a discussion of the variables separately.

That is, what are the distributions of each. Some examples of each.

Graphs of the distributions.

Graphs of the relationships between variables.

Then go on to estimate the equation and explain the graphs and distribution.

5. Estimate and evaluate the equation.

Use Minitab to estimate the equation and give other details about the hypothesized relationships.

Write out the equation it gives.

6. Document the results.

Presentation of data is very important. Useless if you can't make your results understood.

Don't cut and paste MINITAB output.

Learn to explain the regression results in terms anyone can understand.

Interpret coefficients and hypothesis tests.

Example State Environmental Voting

1. Review the literature and develop the theoretical model.

The key points in Senatorial voting are their constituencies.

There is the voting constituency and the business constituency.

We can make a basic model to explain how senators vote.

Now get more specific:

2. Specify the model: select the independent variables and the functional forms.

How should I operationalize my dependent variable?

How should I operationalize each of the independent variables?

3. Hypothesize the expected signs of the coefficients.

$$Pro - Environment\% = \beta_0 + \beta_1^- Constituents + \beta_2^- Industry$$

Write out each hypothesis and corresponding null hypothesis.
Hypothesis and null hypothesis are mutually exclusive and exhaustive.
Which means?
One and only one must be true.

4. Collect the data.

Though I am using all the data from one year is it still right to say that I am using a representative sample?
I want to make generalizations about the population as a whole.
That population is all Senators in the past and the future.

4B – NOT IN TEXT – DISCUSS THE DATA

5. Estimate and evaluate the equation.

I get the equation for predicting a Senator’s voting.

6. Document the results.

Make the table. We will learn about s.e. and t soon.

MINITAB gives the following output:

The regression equation is
senate = 166 - 1.68 election - 0.0678 energy

Predictor	Coef	SE Coef	T	P
Constant	165.62	30.94	5.35	0.000
election	-1.6848	0.5868	-2.87	0.006
energy	-0.06776	0.02185	-3.10	0.003

S = 21.71 R-Sq = 34.3% R-Sq(adj) = 31.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	11567.0	5783.5	12.27	0.000
Residual Error	47	22151.5	471.3		
Total	49	33718.6			

Source	DF	Seq SS
election	1	7036.4
energy	1	4530.7

Unusual Observations

Obs	election	senate	Fit	SE Fit	Residual	St Resid
2	60.6	10.00	10.22	9.63	-0.22	-0.01 X
7	54.3	17.00	21.13	9.89	-4.13	-0.21 X
8	59.6	20.00	-1.92	13.75	21.92	1.31 X
11	46.6	20.00	65.55	5.43	-45.55	-2.17R
49	56.2	94.00	50.81	3.58	43.19	2.02R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

Table 1: OLS Regression of Senate 1992 Environmental Voting on Election Outcome and Energy Consumption

I. Variable	Coef.	S.e.	t
Election	-1.68	0.59	-2.87*
Energy	-0.07	0.02	-3.10*
Constant	165.62	30.94	5.35*

N=49

R² = 0.343

*Statistically significant at .05 level.

What does all this tell us?

What do we expect from Senators in a state with various values of the independent variables?

My substantive findings:

As energy consumption goes up, environmental voting goes down.

As vote for Republican candidates go up, environmental voting goes down.

Outlying cases are?

Homework:

Studenmund Chapter 3, exercises 1,2,3,9.