

POL 602
Basics of OLS Regression
December 1st, 6th and 8th, 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. For very few ordered categories we get into *ordered logit and probit*; for a binary dependent variable we have *logistic regression or probit*; for a few unordered categories we have *multinomial logit and probit* all of which are estimated using maximum likelihood. For now we will stick to OLS regression.

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.

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 – hence linear.

This means that the slope is constant.

For example: Celsius and Fahrenheit have a 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.

$$Y = 32 + 1.8X$$

and say: Given any X_i , I can predict Y_i .

But social science data never look as this.

We get data points that are scattered.

A scatter plot reveals a glimpse of the relationship.

Look at the baseball data set in Stata.

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

We can see the basic relationship looking at these graphs.

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.

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

To describe the dependent variable for individual i :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

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 each of

Y_i

β_0

β_1

X_{1i}

β_2

X_{2i}

β_3

X_{3i}

ε_i

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 the parameters β_0, β_1 .

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, $\hat{\beta}$ – they are predicted values.

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}_i .

\hat{Y}_i is our best guess of Y given a value of X_i and falls on the regression line.

Note, there is no error term here.

We get our best estimates of β_0 and β_1 and the value of X_i 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 population parameters that for us are only theoretical concepts: $\beta_0, \beta_1, \varepsilon_i$

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

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

The first set of numbers describes the relationship between the variables in the population, but the second set is produced by our actual observations and will change if our observations change.

(Draw dots about a regression line)

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

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

And remember what \bar{Y} is?

$$\text{Sample mean of } Y = \frac{1}{n} \sum_{i=1}^n Y_i$$

Sample 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.

So have 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, the errors: e_i

To minimize the spread, we think of the vertical distances between our expected and our predicted observations – this distance is the error in our prediction.

The goal of Ordinary Least Squares is to minimize this error.

So that the positives and negatives don't cancel each other out, we want the squared distances and call them the Residual Sum of Squares:

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

This is $\sum_{i=1}^n e_i^2$

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)

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

We call β_1 the *regression coefficient*.

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

It is:

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

and, $\beta_0 = \bar{Y} - \beta_1 \bar{X}$

Useful Sum of Squares:

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

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

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

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 Stata.

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.

Negative slope indicates a negative relationship and a slope of zero indicates no relationship.

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

We get to this in a bit.

How good is the fit?

The equation itself doesn't tell us about the fit. – we can have the same equation with a good fit or bad fit.

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: $\frac{ESS}{TSS}$

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

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

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

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

We can make a table of these, called ANOVA table

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

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

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
			6050			490

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

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

$$= 0.081$$

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

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

$$\beta_0 = -0.335$$

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

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

Draw the graph.

Regression line passes through the mean of X and Y .

$$\text{Calculate } R^2 = \frac{ESS}{TSS}$$

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

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

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

Example – Explaining 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.

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

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

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

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

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

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

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

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 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 explain.

$$60 \approx 35.298 + 24.724$$

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

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \quad (9)$$

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:

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

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

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

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

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

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

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

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

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

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

These formulae are too big to bother calculating $\hat{\beta}_1$ and $\hat{\beta}_2$ by hand.

But it is useful to look at the formulae and see that each $\hat{\beta}$ 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 X s, the more problematic are our estimates of $\hat{\beta}$ s.

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

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

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

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

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

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

The Classical Assumptions

If we want the estimates we get from OLS to be accurate, we need several assumptions to hold.

We will break many of them, so we want to understand the consequences and learn how to spot the problems and when possible avoid them.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

There is some error – the random component that we are not explaining.

That error or the constant or both, includes all the information we are not using.

If I explain winning percentage using only batting, the error term includes the useful information of pitching.

The error tells us a great deal about our model and it should hold certain properties. These make up most of our regression assumptions.

The Assumptions are:

1. Linearity: The regression model is linear in the coefficients, is correctly specified, and has an additive error term.
2. The error term has a zero population mean.
3. All explanatory variables are uncorrelated with the error term.
4. Observations of the error term are not correlated with each other (no serial correlation).
5. The error term has a constant variance across all X 's (no heteroskedasticity).
6. No perfect multicollinearity. The X 's are different from each other, and preferably not too closely related.
7. The error term is normally distributed – this is not required but is usually assumed.

Lets go through these 1 at a time.

1. Linearity: The regression model is linear in the coefficients, is correctly specified, and has an additive error term.

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

The coefficients all predict a linear relationship.

If we predict non-linear relationships, we can transform the equation.

What are some non-linear relationships?

Diminishing Marginal Utility – a economics term meaning that the value of each additional unit goes down.

For example, the impact on income of the first year of education is greater than the second year which is greater than the third year and so on.

Correctly specified means we are including all the relevant X 's

ε_i is not a function of any variable.

2. The error term has a zero population mean.

If the error term is truly random, we should see that in several ways.

ε_i is distributed randomly around the regression line.

For each X_i , we are just as likely to miss above as below, on average we will land on the line.

Thus, the distribution of the error term has a mean of 0 which is equivalent to saying its expected value is 0.

Here we can put in assumption 7 also, that the error term is normally distributed. It is more likely to be closer to the mean than far away.

3. All explanatory variables are uncorrelated with the error term.

If the error term is randomly generated and distributed, it shouldn't be related to the X variables.

If they were related, we might believe X is causing variation in Y that is actually caused by ε_i .

This is something that we should test – and it can be tested fairly easily.

We ask Stata to save the regression residuals, the errors.

Then we check the correlations between the independent variables and the residuals.

4. Observations of the error term are not correlated with each other (no serial correlation).

If the fact that the ε_i from one observation is positive increases the probability that the ε_i from another observation is positive (or negative) the two are correlated.

This is especially a problem with data that is studied over time.

Each observation is supposed to be randomly selected, but *Time Series* data are sequential and may depend on the previous day to a certain extent.

5. The error term has a constant variance across all X 's (no heteroskedasticity).

For any value of X , the variance in ε_i is the same.

We call this homoskedasticity.

Homoskedasticity means that our line is fitting just as well at any point on the line.

Not a better fit in one spot versus another spot.

This can be checked by making a graph of the residuals against X .

6. No perfect multicollinearity. The X 's are different from each other, and preferably not too closely related.

Remember, as our independent variables get more closely related:

$$\frac{(\sum yx_1)(\sum x_2^2) - (\sum yx_2)(\sum x_1x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

becomes 0/0 and regression is impossible.

As X_1 and X_2 get more and more alike, our estimates are problematic.

This can be checked by looking at the correlations between our independent variables.

7. The error term is normally distributed – this is not required but is usually assumed.

The Sampling Distribution of $\hat{\beta}$

Remember the sampling distribution of sample means?

What was it?

It was the distribution of means when we took multiple samples.

That was useful for a single variable.

Now we are interested in two or multiple variables.

We want to think about the $\hat{\beta}$ that we generate as coming from a population of $\hat{\beta}$ s.

Our $\hat{\beta}$ came from taking one sample out of the population and seeing the relationship between the variables.

Had we taken another sample, we would have gotten a different $\hat{\beta}$.

Think about many many samples and the distribution of $\hat{\beta}$ – this is what we mean by “Sampling Distribution of $\hat{\beta}$ ”

If all of the above assumptions hold, we get an “**unbiased estimator**”

Our $\hat{\beta}$ is an “estimator” in that we use it to estimate the relationship between X and Y. An unbiased estimator is one whose sampling distribution has as its expected value, the true value of β .

$$E(\hat{\beta}_k) = \beta_k$$

If we could take an infinite number of samples and get a $\hat{\beta}$ for each, they would be normally distributed.

We only get one sample and therefore one $\hat{\beta}$, so we want to know that it is unbiased.

What else would we like?

Variance of distribution should be as small as possible for an unbiased estimator. This is called *efficiency*.

We can decrease the variance by increasing our sample size.

As the size of our sample approaches the size of the population, we get closer and closer to the actual value of β and the variance is smaller and smaller – *consistency*.

So, we have this theoretical sampling distribution of $\hat{\beta}$ s.

We use the estimate we get, $\hat{\beta}$, to tell us what we believe the population value is. We also want to know how much variance there may be in this estimate.

We want to know the variance of the $\hat{\beta}$.

Use the Standard Deviation of the Sampling Distribution of $\hat{\beta}$. This is called the Standard Error.

For one independent variable:

$$SE(\hat{\beta}) = \sqrt{\frac{\sum e_i^2 / (n-2)}{\sum (X_i - \bar{X})^2}}$$

For two independent variables:

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sum e_i^2 / (n-3)}{\sum (X_{1i} - \bar{X}_1)^2 (1 - r_{12}^2)}}$$

SE is roughly average miss for $\hat{\beta}$.

If all of our assumptions hold our estimator is **BLUE**.
“Best (minimum variance) Linear Unbiased Estimator.”

Our OLS estimates will have the following properties:

1. They are unbiased.

$$E(\hat{\beta}) = \beta$$

2. They are minimum variance.

The distribution of the coefficient estimates around the true parameter values is as tight as possible.

3. They are consistent. As the sample size increases the estimates converge on the true population parameters.

4. They are normally distributed.

The $\hat{\beta}_s \sim N(\beta, VAR[\beta])$.

Hypothesis Testing

We know how to get an estimate of the relationship in the sample, $\hat{\beta}$.

But statistical inference is what we are really interested in.

The weights of 9 students or the conflicts of 10 countries are not what we want to explain.

We want to be able to make generalizable statements about the variables in the entire population.

The regression model for the entire population is:

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

(We switch to Greek when talking about the model as a whole.)

We can't be sure about this equation, we only use our sample to guess what it is.

In our sample, we estimate $\hat{\beta}_0$ & $\hat{\beta}_1$.

For different samples, we are going to get different values for $\hat{\beta}_0$ & $\hat{\beta}_1$.

We want to know how our $\hat{\beta}_0$ & $\hat{\beta}_1$ would be distributed around β_0 & β_1 .

This is how we test our hypothesis.

When we make a hypothesis in regression, what are we hypothesizing?

About β_1 : usually that it is positive or negative or just non-zero.

If there is no relationship between two variables, $\beta_1=0$.

We use our sample to generate $\hat{\beta}_1$ and then use that to tell us what β_1 might be.
 If we can conclude it is positive or negative or non-zero, we reject null hypothesis of no relationship.

Our null Hyp: $\beta_1=0$

If we say $\beta_1 \neq 0$, two tailed hypothesis.

If we say $\beta_1 > 0$ or $\beta_1 < 0$, we are making an one tail hypothesis.

We use the t -distribution to test whether the observed $\hat{\beta}_1$ differs significantly from the population parameter β_1 .

Remember, the t distribution is like the normal but not quite.
 with high n , the same,
 with low n , fatter tails.

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE_{\hat{\beta}_1}}$$

Where β_1 is the null hypothesis' value, usually 0, giving us:

$$t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$$

Degrees of freedom for $t = N - K - 1$

$SE_{\hat{\beta}_1}$ is the estimate of the standard error of the regression coefficient.

Also, theoretically, it is the standard deviation of the Sampling Distribution of $\hat{\beta}_1$.

95% Confidence intervals for β_1 :

$$\hat{\beta}_1 \pm t_{.025}(SE_{\hat{\beta}_1})$$

$$\hat{\beta}_0 \pm t_{.025}(SE_{\hat{\beta}_0})$$

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sum e_i^2 / (n - 2)}{\sum (X_i - \bar{X})^2}}$$

The more error there is, the higher the standard error is.

The more cases there are the lower the standard error is.
The more variance there is in X, the lower the standard error is.

SE is average miss for $\hat{\beta}$.

$$SE(\hat{\beta}_0) = \sqrt{\sum e_i^2 / n - 2} * \sqrt{\frac{\left(\frac{1}{n} + \bar{X}\right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

Our hypothesis is about β – slope in population.

Our hypothesis test is about $\hat{\beta}_1$ - is it significantly different from zero.

Steps of a hypothesis test?

1. Advance a research hypothesis: H_A and Null hypothesis: H_0 .

Our null hypothesis and alternative hypothesis are complements – meaning that one and only one must be true.

We concentrate on the null hypothesis knowing that if we prove it is wrong, we are thereby supporting our alternative hypothesis.

Make the hypothesis directional if possible.

If you use a level of significance other than .05 – justify why you are doing so.

This value is the probability of being wrong by chance alone due to sampling error.

Since we know that it is possible to have an *unrepresentative* sample we want to take into account its possibility.

We can quantify exactly how possible it is that our results come from an unrepresentative sample.

This is what our probability level is.

It tells us the probability that we could be wrong by chance alone.

If we set our level at .05 (which is standard) and find a relationship exists we are still admitting that there is a 5% chance that although we found a relationship in the sample, the relationship might not exist in the population.

The lower the probability level that we set the lower the chance that we are wrong by chance alone.

Thus, at the .01 level there is only a 1% chance that the relationship we find in the sample isn't true of the population as a whole and at the .001 level there is only a 0.1% chance.

Why don't we always try for the highest level of probability?

Because there is a trade-off: The lower our probability level the harder it is to say that we have found something. But the harder a goal we set for ourselves, the better the value of attaining it is.

For example: We might be able to say that we accept our hypothesis at the .05 level – thereby allowing the 5% chance that we are wrong in doing so. But the relationship in our sample would have to be stronger to be able to accept the hypothesis at the .01 level.

2. Choose the test that is appropriate based upon what data you have available.

E.g. If data are arrayed in tabular form and we are looking at 2 variables, we want to know if the population is the same as the sample, Chi-square test is appropriate.

If we can do a regression, then t is appropriate.

3. Calculate the test statistic.

This will depend on which statistic we are using.

4. Locate the appropriate critical value.

To get the critical value you will need to know:

- A. The statistic you are interested in
- B. The degrees of freedom (based on the size of your sample and sometimes the number of independent variables)
- C. The desired probability level (usually .05).
- D. Whether your hypothesis is one tailed (directional – positive or negative) or two tailed (non-directional).

5. Compare the test statistics you calculated with the appropriate critical value.

6. If test statistic < critical value → “Fail to reject the null hypothesis.” You cannot say that the results are statistically significant.

7. If test statistic > critical value → “Reject the null hypothesis and provisionally accept the alternative hypothesis.”

8. Assign a level of probability to your findings.

E.g. "Fail to reject the null hypothesis at the .05 level."

"Reject the null hypothesis (.05 level or whichever level you choose) and provisionally accept the alternative hypothesis."

That is, you find your hypothesis to be statistically significant at the .05 level.

Another way of saying this is that there is a 5% chance that you have made a Type I error.

Which is?

Finding a relationship where none truly exists. Falsely rejecting the null.

A Type II error?

Failing to find a true relationship. Falsely accepting the null.