

LINEAR MODEL

SP03 3 - 2

OLS REGRESSION

Motivating the model...

A random variable: systematic & stochastic parts:

$$Y_i = \mu + u_i$$

Think of the linear relationship in terms of X influencing the systematic part of Y...

- That is, we want to estimate

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

How?

- Lots of possible estimators...
- Want a “good” estimate...
 - * One which is “close” to the real values
 - * I.e., which minimizes the distance between predicted and actual...
- Could use Minimum Error method
 - * I.e. pick the estimate of $\hat{\beta}$ that minimizes $\sum u$
 - * This isn't optimal, however, since values may “cancel out”...
- OK, use absolute value...
 - * I.e. choose $\hat{\beta}$ to minimize $\sum |u_i|$.
 - * The MAD estimator.
 - * Still not optimal... (we'll discuss why not later on...).
- Could square the errors, to make them positive
 - * Then minimize $\sum u_i^2$
 - * The Least-Squares Estimator
 - * This is what OLS is built around
 - * Has some nice properties...

An Example...

Two data points on two variables X and Y: (1,3) and (2,5). <<GRAPH>>

- Want to estimate the linear relationship between the two of them... (HOW?)
- Y is a random variable...
 - * Set $Y_i = \beta_0 + \beta_1 X_i + u_i$
 - * Need to minimize $\sum u_i^2$

Note that $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

* So for the first observation, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1(1)$

* For the second, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1(2)$

Note also that $u_i = Y - \hat{Y}$
 $= 3 - \hat{\beta}_0 - \hat{\beta}_1$ for the first observation
 $= 5 - \hat{\beta}_0 + 2\hat{\beta}_1$ for the second one

So the sum of squared residuals **S** is

$$\begin{aligned} \mathbf{S} &= u_1^2 + u_2^2 \\ &= (3 - \hat{\beta}_0 - \hat{\beta}_1)^2 + (5 - \hat{\beta}_0 + 2\hat{\beta}_1)^2 \\ &= (9 + \hat{\beta}_0^2 + \hat{\beta}_1^2 - 6\hat{\beta}_0 - 6\hat{\beta}_1 + 2\hat{\beta}_0\hat{\beta}_1) + (25 + \hat{\beta}_0^2 + 4\hat{\beta}_1^2 - 10\hat{\beta}_0 - 20\hat{\beta}_1 \\ &\quad + 4\hat{\beta}_0\hat{\beta}_1) \\ &= 2\hat{\beta}_0^2 + 5\hat{\beta}_1^2 + 6\hat{\beta}_0\hat{\beta}_1 - 16\hat{\beta}_0 - 26\hat{\beta}_1 + 34 \end{aligned}$$

Now we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize this function

* HOW?

* *partial* derivatives w.r.t. $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\begin{aligned} \partial S / \partial \hat{\beta}_0 &= 4\hat{\beta}_0 + 6\hat{\beta}_1 - 16 \\ \partial S / \partial \hat{\beta}_1 &= 10\hat{\beta}_1 + 6\hat{\beta}_0 - 26 \end{aligned}$$

Then what?

* Set both equal to zero and solve to minimize the function...

$$\begin{aligned} 4\hat{\beta}_0 + 6\hat{\beta}_1 - 16 &= 0 \\ 2\hat{\beta}_0 + 3\hat{\beta}_1 - 8 &= 0 \\ 2\hat{\beta}_0 &= -3\hat{\beta}_1 + 8 \\ \hat{\beta}_0 &= -3/2\hat{\beta}_1 + 4 \end{aligned}$$

$$\begin{aligned} 10\hat{\beta}_1 + 6\hat{\beta}_0 - 26 &= 0 \\ 5\hat{\beta}_1 + 3\hat{\beta}_0 - 13 &= 0 \\ 5\hat{\beta}_1 + 3[-3/2\hat{\beta}_1 + 4] - 13 &= 0 \\ 5\hat{\beta}_1 - 9/2\hat{\beta}_1 + 12 - 13 &= 0 \\ 1/2\hat{\beta}_1 - 1 &= 0 \\ \hat{\beta}_1 &= 2 \end{aligned}$$

$$\begin{aligned} 2\hat{\beta}_0 + 3(2) - 8 &= 0 \\ 2\hat{\beta}_0 &= 2 \\ \hat{\beta}_0 &= 1 \end{aligned}$$

So the equation is $Y_i = 1 + 2X_i + u_i$

- Note that if we'd used the "slope" method we'd have gotten:

$$* \hat{\beta}_1 = (5-3)/(2-1) = 2, \text{ and}$$

$$* \hat{\beta}_0 = -2(2) + 5 = 1$$

This is basically what OLS does...

In the more general case with N observations, $Y_i = \beta_0 + \beta_1 X_i + u_i$

* Minimize $\sum u_i^2$

$$\begin{aligned} * \text{ Rewrite as } \quad \mathbf{S} &= \sum u_i^2 = \sum (Y_i - \hat{Y}_i)^2 \\ &= \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\ \text{for each observation } i &= Y_i^2 - 2Y_i \hat{\beta}_0 - 2Y_i \hat{\beta}_1 X_i + \hat{\beta}_0^2 + 2\hat{\beta}_0 \hat{\beta}_1 X_i + \hat{\beta}_1^2 X_i^2 \end{aligned}$$

Then we partially differentiate this equation w.r.t. $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\begin{aligned} \partial S / \partial \hat{\beta}_0 &= 0 - 2Y - 0 + 2\hat{\beta}_0 + 2\hat{\beta}_1 X + 0 \\ &= -2Y - 2\hat{\beta}_0 + 2\hat{\beta}_1 X \\ \text{(summing again...)} &= -2\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\ &= -2\sum(\tilde{u}_i) \end{aligned}$$

$$\begin{aligned} \partial S / \partial \hat{\beta}_1 &= -2\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i \\ &= -2\sum \tilde{u}_i X_i \end{aligned}$$

Setting these equations equal to zero and doing a little algebra yields...

$$\sum Y_i = N\hat{\beta}_0 + \hat{\beta}_1 \sum X_i$$

$$\sum Y_i X_i = \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2$$

These are what is known as the OLS **normal equations**



Solving them simultaneously yields:

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

and

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



This is the general OLS estimator equation:

- Essentially, it is **Cov(X,Y) / Var(X)**

- $\hat{\beta}_0$ is the “intercept”

* Place where the regression line crosses the Y-axis

* Expected value of Y when X = 0

- The fact that we don't have Y in the denominator means that Y isn't “normed out”

* $\hat{\beta}_1$ is expressed in terms (units) of Y (“slope”)

* Common interpretation: the effect on Y of a one-unit change in X

Since $Cov(x, y) = E[(x_1 - \mu_1)(x_2 - \mu_2)] = \frac{1}{n} \sum_{i=1}^n (x - \bar{x})(y - \bar{y})$, and since

$$Var(x) = \frac{1}{n} \sum_{i=1}^n (x - \bar{x})^2,$$

then because $\hat{\beta}_0 = \frac{\sum_{i=1}^n (x - \bar{x})(y - \bar{y})}{\sum_{i=1}^n (x - \bar{x})^2}$, $\hat{\beta}_0 = \frac{Cov(x, y)}{Var(x)}$.

An Example: Supreme Court Voting

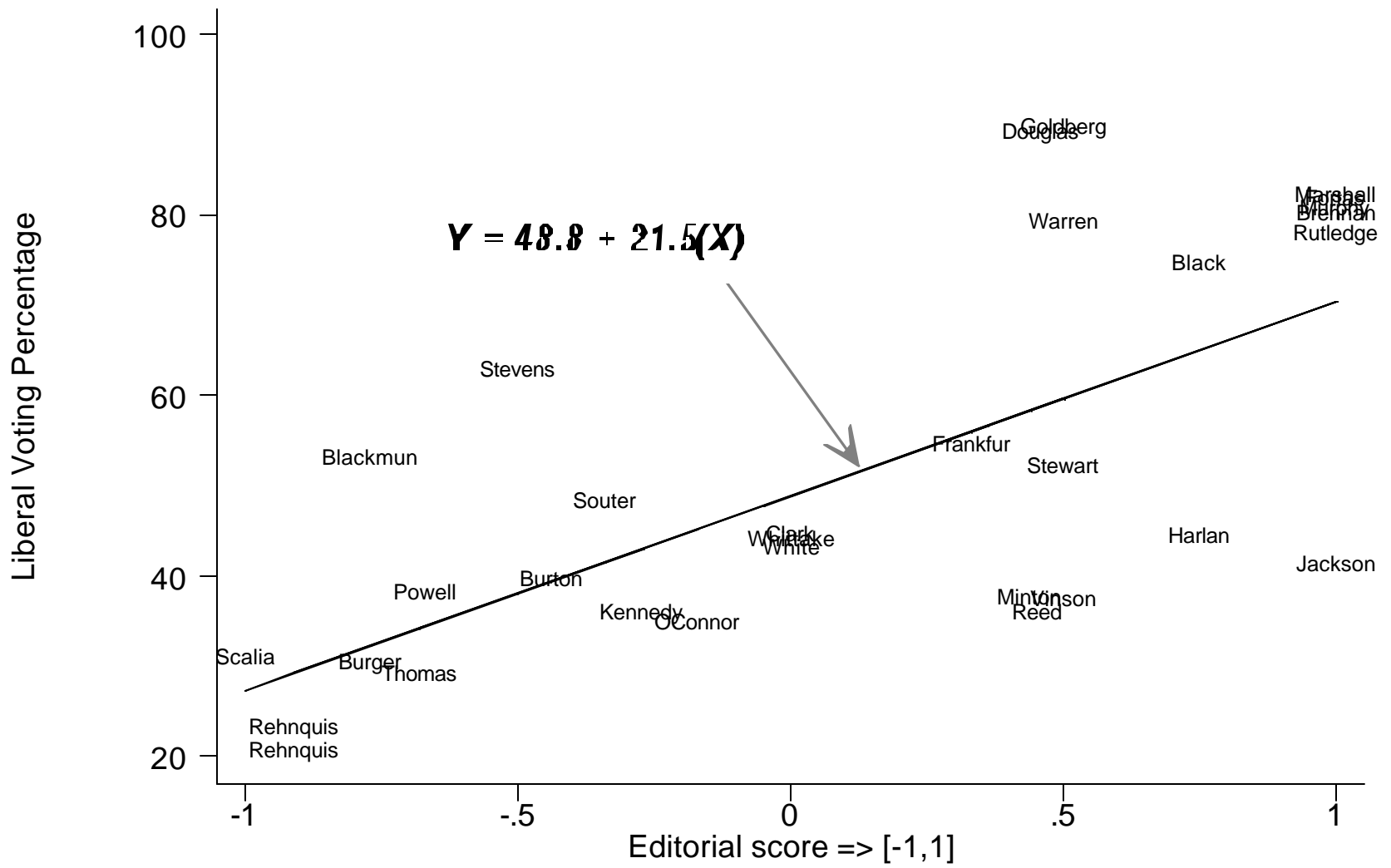
```
. reg civrts score
```

Source	SS	df	MS	Number of obs =	31
Model	6406.51588	1	6406.51588	F(1, 29) =	26.24
Residual	7081.02635	29	244.173322	Prob > F =	0.0000
Total	13487.5422	30	449.584741	R-squared =	0.4750
				Adj R-squared =	0.4569
				Root MSE =	15.626

civrts	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
score	21.54446	4.206044	5.12	0.000	12.94214	30.14679
_cons	48.80994	2.852268	17.11	0.000	42.9764	54.64349

. list name score civrts votehat residual in 1/31

	name	score	civrts	votehat	residual
1.	Black	.75	73.9	64.96829	8.93171
2.	Reed	.45	35.1	58.50495	-23.40495
3.	Frankfurter	.33	53.8	55.91962	-2.119618
4.	Douglas	.46	88.4	58.7204	29.6796
5.	Murphy	1	80	70.35441	9.645593
6.	Jackson	1	40.4	70.35441	-29.9544
7.	Rutledge	1	77.2	70.35441	6.84559
8.	Burton	-.44	38.9	39.33038	-.4303787
9.	Vinson	.5	36.7	59.58218	-22.88218
10.	Clark	0	43.8	48.80994	-5.009945
11.	Minton	.44	36.8	58.28951	-21.48951
12.	Warren	.5	78.5	59.58218	18.91782
13.	Harlan	.75	43.7	64.96829	-21.26829
14.	Brennan	1	79.5	70.35441	9.145593
15.	Whittaker	0	43.3	48.80994	-5.509945
16.	Stewart	.5	51.3	59.58218	-8.282176
17.	White	0	42.4	48.80994	-6.409943
18.	Goldberg	.5	88.9	59.58218	29.31783
19.	Fortas	1	81	70.35441	10.64559
20.	Marshall	1	81.4	70.35441	11.04559
21.	Burger	-.77	29.6	32.22071	-2.620707
22.	Blackmun	-.77	52.3	32.22071	20.07929
23.	Powell	-.67	37.4	34.37515	3.024848
24.	Rehnquist	-.91	19.8	29.20448	-9.404483
25.	Stevens	-.5	62	38.03771	23.96229
26.	OConnor	-.17	34.1	45.14738	-11.04739
27.	Rehnquist	-.91	22.5	29.20448	-6.704482
28.	Scalia	-1	30.2	27.26548	2.93452
29.	Kennedy	-.27	35.1	42.99294	-7.89294
30.	Souter	-.34	47.6	41.48483	6.115172
31.	Thomas	-.68	28.3	34.15971	-5.85971



INFERENCE IN THE CLRM...

Point estimates are nice, but they only get us so far...

- If we want to make *inferences* about the population from which our sample of data is drawn, we need to know the variability (or precision) of our estimates.
- This will also allow us to say some things about the sampling variability properties of those estimates as well...

It can be shown that the variance of our two estimates are:

$$\text{Var}(\hat{\mathbf{b}}_0) = \frac{\sum X_i^2}{N \sum (X_i - \bar{X})^2} \mathbf{s}^2 \quad (1)$$

and

$$\text{Var}(\hat{\mathbf{b}}_1) = \frac{\mathbf{s}^2}{\sum (X_i - \bar{X})^2} \quad (2)$$

while

$$\text{Cov}(\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1) = \frac{-\bar{X}}{\sum_{i=1}^N (X_i - \bar{X})^2} \mathbf{s}^2 \quad (3)$$

respectively.

Likewise, the estimates for the standard errors of $\hat{\mathbf{b}}_0$ and $\hat{\mathbf{b}}_1$ are simply the square roots of the first two equations...



Note several things about these formulas:

- **The variance of both estimates is directly proportional to σ^2 .**

- * All else equal, the more variability there is in the errors, the greater the variability in our estimates of β .
- * Makes sense: Greater variability in the *us* means that our regression isn't "predicting" very well; thus, we can't be as sure that our estimates of β are accurate...

- **The variance of both estimates is inversely proportional to $\sum (X_i - \bar{X})$.**

- * That is, all else (including σ^2) equal, the more variation in X, the more precise our estimates of β .
- * Again, not surprising: The less variability we have in X, the worse job X will do of "explaining" Y (in the limit, if X is constant, it tells us nothing about variability in Y).

- **As N increases, the variability of our estimates will go down.**

- * This is directly shown for $\text{Var}(\beta_0)$.
- * Is also true for $\text{Var}(\beta_1)$, since σ^2 is also a decreasing function of N.

- **The covariance of the two estimates depends on the sign of the mean of X.**

- * The importance of the covariance of estimates will be a bit more obvious later...

Now that we know the variability of our estimates, we can say some useful things about them...

The Gauss-Markov Theorem

“Given the assumptions of the classical linear regression model, the least squares estimators are the minimum variance estimators among the class of unbiased linear estimators. (They are BLUE).”

How do we know this?

- Think about the estimator for \hat{b}_1 :

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

We can make use of some algebraic identities to rewrite this equation as:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} y_i$$

Think of this as a weighted combination of the Y_i s, with the first term acting as the "weight"; call this weight k :

$$\hat{\beta}_1 = \sum k_i Y_i$$

Now, suppose we were to derive some other estimator, using some other "weight" instead; call that weight w :

$$\tilde{\beta}_1 = \sum w_i Y_i$$

For this new weight, we can see what needs to hold for \tilde{b}_1 to be unbiased:

$$\begin{aligned} E(\tilde{b}) &= \sum w_i E(Y_i) \\ &= \sum w_i (b_0 + b_1 X_i) \\ &= b_0 \sum w_i + b_1 \sum w_i X_i \end{aligned}$$





In order for this new estimator to be unbiased, it has to be the case that $\sum w_i = 0$ and $\sum(w_i X_i) = 1$.

The variance of this new estimator is:

$$\begin{aligned} \text{Var}(\tilde{b}_1) &= \text{Var}\left(\sum w_i Y_i\right) \\ &= s^2 \sum w_i^2 \\ &= s^2 \sum \left[w_i - \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} + \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} \right]^2 \quad (4) \\ &= s^2 \sum \left(w_i - \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} \right)^2 + s^2 \left(\frac{1}{\sum (X_i - \bar{X})^2} \right) \end{aligned}$$

- Note that the last term of this last equation is a constant; this means that the estimator with the smallest variability has weights which minimize

$$\sum \left(w_i - \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} \right)^2. \text{ Obviously, this term is minimized when it equals}$$

zero; **that is, when** $w_i = \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$. When this happens, the first term in


(4) drops out, and the variance of our estimator becomes $\frac{s^2}{\sum (X_i - \bar{X})^2}$; i.e.,

that for the OLS estimator.

- We can show a similar property for any given estimator for \hat{b}_0 .
- All this means is that, **among all possible estimators of β , the estimator with the smallest variability / greatest precision is the least-squares estimator.**

Inference...

To get to inference, we can make use of these properties of the CLRM...

- IF we assume that our errors are distributed *normally* (that is, $u_i \sim$ i.i.d. $N(0, \sigma^2)$), then our estimates of β (which are also random variables, and which are functions of the u_i s) will *also be normally distributed*. 
- This means that inference on them is really, really easy...
- For example, since \hat{b}_1 is normally distributed, then we ought to be able to convert this variable to a Z-score...

$$Z = (\hat{b}_1 - \beta_1) / \text{s.e.}(\hat{b}_1)$$

S.E. is the standard error of \hat{b}_1 , i.e. the square root of its variance

This Z variable is distributed as $N(0,1)$, because it is simply the "standardized" version of \hat{b}_1 .


BUT

There's a problem...

- To calculate $\text{s.e.}(\hat{b}_1)$, we need to know σ^2 ...
- i.e. the variance of the errors *in the population*
 - We really never know this...

So instead, we have to use its estimate, $\hat{\sigma}^2$...

How do we estimate $\hat{\sigma}^2$? Fox (section 10.3) shows that an unbiased estimator of σ^2 is:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{N - 2}$$


where \hat{u}_i^2 is simply the estimated u_i from the regression (i.e., the square of the observed minus the expected values).

Plugging $\hat{\sigma}^2$ in for σ^2 in the equation for the variance of \hat{b}_1 :

$$\text{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}$$

Take its square root, to get the estimated standard error of \hat{b}_1 ...

$$\begin{aligned} \text{s.e.}(\hat{\beta}_1) &= \sqrt{\text{Var}(\hat{\beta}_1)} = \sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}} \\ &= \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}} \end{aligned}$$

If we now calculate $\left(\frac{\hat{b}_1 - b_1}{\text{s.e.}(\hat{b}_1)} \right)$ using that formula, we get a statistic:

$$\begin{aligned} \frac{\hat{\beta}_1 - \beta_1}{\text{s.e.}(\hat{\beta}_1)} &= \frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}}} \\ &= \frac{(\hat{\beta}_1 - \beta_1) \sqrt{\sum (x_i - \bar{x})^2}}{\hat{\sigma}} \end{aligned}$$

This follows a **t** distribution, with (N-2) degrees of freedom because:

- * as we noted before, the numerator is a standard normal variable, and
- * the errors \hat{u}_i are distributed as independent standard normal variables, so that
- * the squared errors \hat{u}_i^2 are all independent chi-squared variables with one degree of freedom each; this means that
- * the denominator $\hat{\sigma}^2$ is a sum of (N-2) independent chi-squared variables, and so is itself distributed as chi-squared with (N-2) degrees of freedom. So,
- * We have a standard normal variate over a chi-square variable with (N - 2) d.f., yielding our old friend, the **t** distribution...

- This is where the regression **t-test** for the coefficient estimates comes from...
- One can derive a similar distribution for the estimate of \hat{b}_0 , using \hat{s}^2 there as well.

Given that we know this, how do we test regression hypotheses?...

We can think of hypothesis testing in two ways: INTERVALS and SIGNIFICANCE

- **INTERVALS** involve drawing a “confidence interval” around our estimate of β
 - * Since we know the distribution of the estimate, we can talk about how confident we are that the “true” value is within some interval around that range...
 - * For example, we know that roughly 95% of the “mass” of a **t** -distribution occurs within two standard deviations of its mean

SO

- * Given a confidence interval of 95%, we can be sure that, 95 times out of 100, an interval constructed of two standard deviations around a given point estimate will contain the “true” value β .
- * This is a very nice approach, BUT
- * We tend to want *point estimates*; so most of the time you’ll see:

SIGNIFICANCE TESTS

- Complimentary to confidence intervals...
- In SIGNIFICANCE tests we test a specific hypothesis about the true value of β

In a *t*-test, we don’t know the true value of β_1

- In essence, we “plug in” a value for β_1 into the *t*-test formula, and then evaluate how likely it is that we’d have drawn that particular sample of data given that value

An Example:

A regression of Prez. Popularity on Unemployment % yields:

$$\text{Pop.} = \begin{array}{r} 76 \\ (10) \end{array} - \begin{array}{r} 1.2(\text{UE}) \\ (0.5) \end{array} + u$$

Interval estimation:

Calculate the two-tailed 95% confidence interval for β_0 and β_1

- For β_0 , we have:	lower bound	=	$76 - (1.96 \times 10)$	=	56.4
	upper bound	=	$76 + (1.96 \times 10)$	=	95.6
- For β_1 , we have:	lower bound	=	$1.2 - (1.96 \times 0.5)$	=	0.22
	upper bound	=	$1.2 + (1.96 \times 0.5)$	=	2.18


(Q: Where did the "1.96" come from???)

Hypothesis Testing:

Test: Hypothesis that UE has no impact on popularity

* $t = (1.2 - 0) / 0.5 = 2.4$ ($p < .01$)

* Can *reject* this hypothesis at the .01 level

Test: UE has a 1-to-1 impact on unemployment 

* $t = (1.2 - 1) / 0.5 = 0.2 / 0.5 = \mathbf{0.4}$ ($p > .10$)

* *Cannot reject* this hypothesis at the .10 level

ALWAYS phrased in terms of *rejecting* or *not rejecting* stated hypotheses

- A "null" hypothesis is essentially one of no effect.
- Don't always want to test the "null"; may want to test something else...

READ the recent (September 1999) article by Jeff Gill in *PRQ* about how null hypothesis testing is hosed... (in the POLS 509 course 'readings' folder)

If A then B is highly likely

Not B observed

Therefore A is highly unlikely

If H_0 is true then the data are highly likely to follow an expected pattern

The data do not follow the expected pattern

Therefore H_0 is highly unlikely.

If A then B is highly likely

Not B observed

Therefore A is highly unlikely

If a person is an American then it is highly unlikely she is a member of Congress

The person is a member of Congress

Therefore it is highly unlikely she is an American.

From this simple little example and the resulting absurdity it is easy to see that if the $P(\text{Congress}|\text{American})$ is low (the p-value), it does *not* imply that $P(\text{American}|\text{Congress})$ is also low.

Predictions and Inference

The predicted value for Y given a particular value of X (say, X_k) is just:

$$E(Y|X_k) \equiv \hat{Y}_k = \hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 X_k$$

This is a **point prediction** – a single value equal to the expected value of Y associated with a particular set of values for X_k , $\hat{\mathbf{b}}_0$, and $\hat{\mathbf{b}}_1$.

Note that this prediction depends on the values of the estimates, *which are themselves random variables*. This means that **the point prediction is also a random variable**.

OK, so *what are the properties of this random variable \hat{Y}_k ?*

Well, its easy to show that \hat{Y}_k is an **unbiased** estimator of Y_k :

$$\begin{aligned} E(\hat{Y}_k) &= E(\hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 X_k) \\ &= E(\hat{\mathbf{b}}_0) + E(\hat{\mathbf{b}}_1) X_k \\ &= \mathbf{b}_0 + \mathbf{b}_1 X_k \end{aligned}$$

The next thing we might want to know is *how good a prediction is this?* That is, how much variability is there in this prediction?

Recall that, for two random variables A and B, the variance of their sum is equal to:

$$\text{Var}(A + B) = \text{Var}(A) + \text{Var}(B) + 2\text{Cov}(A, B)$$

This is useful, in that we want to know **Var(\hat{Y}_k)**, which is:

$$\begin{aligned} \text{Var}(\hat{Y}_k) &= \text{Var}(\hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 X_k) \\ &= \text{Var}(\hat{\mathbf{b}}_0) + \text{Var}(\hat{\mathbf{b}}_1) X_k^2 + 2\text{Cov}(\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1) X_k \\ &= \left(\frac{\sum X_i^2}{N \sum (X_i - \bar{X})^2} \mathbf{s}^2 \right) + \left(\frac{\mathbf{s}^2}{\sum (X_i - \bar{X})^2} \right) X_k^2 + 2 \left(\frac{-\bar{X}}{\sum_{i=1}^N (X_i - \bar{X})^2} \mathbf{s}^2 \right) X_k \end{aligned}$$

A little bit of algebra yields a more visually-satisfying representation:

$$\text{Var}(\hat{Y}_k) = \mathbf{s}^2 \left[\frac{1}{N} + \frac{(X_k - \bar{X})^2}{\sum_{i=1}^N (X_i - \bar{X})^2} \right]$$

What does this equation tell us?

- The variability of a prediction **decreases as N increases**.
- The variability of the predicted mean **decreases as the variability of X increases**.
- The variability of the prediction **increases as the value of X at which we are predicting gets farther away from the mean of X** .

Inference

As before, since we typically don't know \mathbf{s}^2 , we replace it with its unbiased estimator $\hat{\mathbf{s}}^2$. The square root of $\text{Var}(\hat{Y}_k)$ is what is known as the standard error of the prediction:

$$s.e.(\hat{Y}_k) = \sqrt{\hat{\mathbf{s}}^2 \left[\frac{1}{N} + \frac{(X_k - \bar{X})^2}{\sum_{i=1}^N (X_i - \bar{X})^2} \right]}$$

As with our actual coefficients, we can use the t distribution to make inferences about the predictions. In particular, we can calculate (say) the 95-percent confidence interval around our prediction \hat{Y}_k as:

$$\hat{Y}_k \pm [1.96 \times s.e.(\hat{Y}_k)]$$

When using predicted values to interpret a regression model, this is an important quantity to discuss, since it tells us how "accurate" (or "precise") the predictions of our model are.

An Example: More Supreme Court Voting

Remember this?

```
. reg civrts score
```

Source	SS	df	MS			
Model	6406.51588	1	6406.51588	Number of obs =	31	
Residual	7081.02635	29	244.173322	F(1, 29) =	26.24	
Total	13487.5422	30	449.584741	Prob > F =	0.0000	
				R-squared =	0.4750	
				Adj R-squared =	0.4569	
				Root MSE =	15.626	

civrts	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
score	21.54446	4.206044	5.12	0.000	12.94214	30.14679
_cons	48.80994	2.852268	17.11	0.000	42.9764	54.64349

What do these results tell us?

- **Total** is the total variability in Y – that is, $\sum_{i=1}^N (Y_i - \bar{Y})^2 = 13487.54$. (i.e., SST)

- **Model** is the model (“explained” or “regression”) sum of squares – that is,

$$\sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2 = 6406.52.$$

- **Residual** is the residual (“unexplained”) sum of squares (i.e., SSR) – that is,

$$\sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \left(\equiv \sum_{i=1}^N \hat{u}_i^2 \right) = 7081.02. \text{ This value, divided by the number of}$$

independent variables k (plus one for the constant term), gives us our estimate

$$\hat{\sigma}^2 \text{ (so, here, } \hat{\sigma}^2 = \frac{\sum_{i=1}^N \hat{u}_i^2}{N - k - 1} = \frac{7081.02}{29} = 244.173, \text{ and the standard error of}$$

our estimate (SEE, here the “**Root MSE**”) is equal to $\sqrt{\frac{RSS}{N - k - 1}}$. Here,

$$\text{that's } \sqrt{\frac{7081.02}{29}} = \sqrt{244.173} = 15.626. \quad \text{📄}$$

Now, we can calculate the variability in X as $\sum_{i=1}^N X_i^2 = 14.26$, and its variability

around its mean as $\sum_{i=1}^N (X_i - \bar{X})^2 = 13.80$. This means that:

- The **variance** estimate for the **constant** term is

$$\begin{aligned} \text{Var}(\hat{\mathbf{b}}_0) &= \frac{\sum X_i^2}{N \sum (X_i - \bar{X})^2} \hat{\mathbf{s}}^2 \\ &= \frac{14.26}{31(13.80)} (244.173) \\ &= 8.14 \end{aligned}$$

and the square root of this is the standard error estimate of the constant term (that is, about 2.85).

- The **variance** estimate for the **slope** term is

$$\begin{aligned} \text{Var}(\hat{\mathbf{b}}_1) &= \frac{\hat{\mathbf{s}}^2}{\sum (X_i - \bar{X})^2} \\ &= \frac{244.173}{13.80} \\ &= 17.69 \end{aligned}$$

and the square root of this is the estimated standard error for the slope coefficient (that is, about 4.21).

- The **covariance** of the estimated slope and intercept is

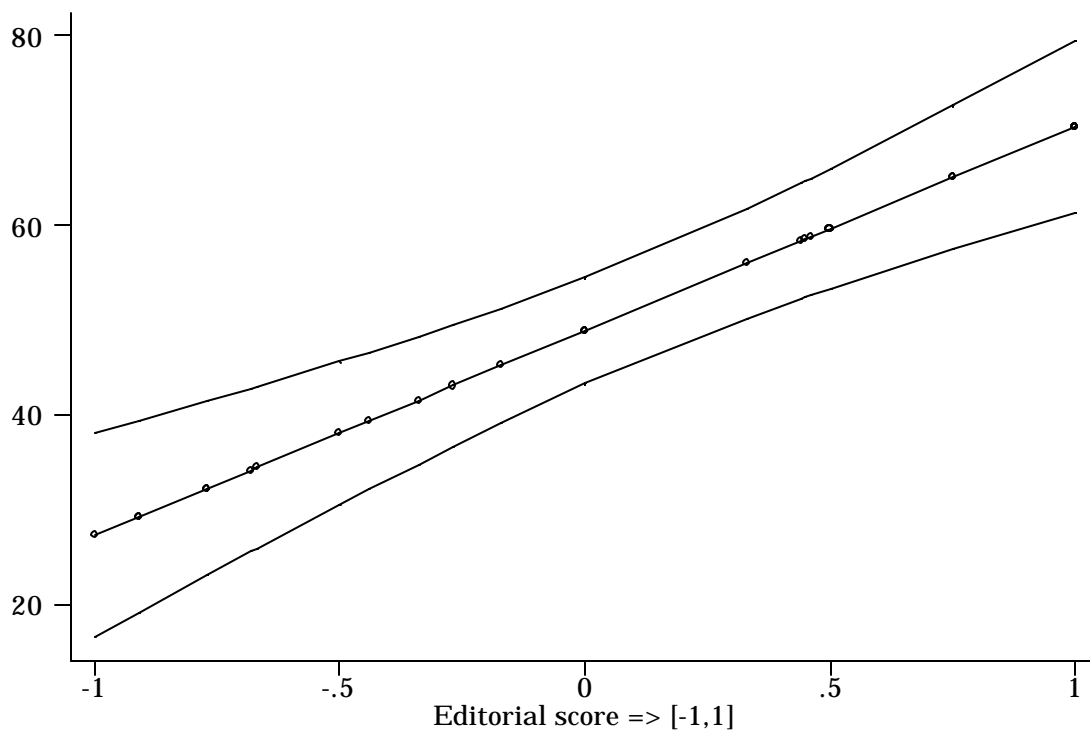
$$\begin{aligned} \text{Cov}(\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1) &= \frac{-\bar{X}}{\sum_{i=1}^N (X_i - \bar{X})^2} \hat{\mathbf{s}}^2 \\ &= \left(\frac{-0.12}{14.26} \right) (244.173) \\ &= -2.07 \end{aligned}$$

Stata will also generate the standard errors of the individual predicted values of Y for us, using the `-predict-` command:

```
. predict votehat  
  
. predict se_pred, stdp
```

The former command gives us the predicted value \hat{Y}_i for all the observations in the data, while the latter gives us the standard error of that prediction. From this, we can draw (say) 95% confidence intervals around our (predicted) regression line:

```
. gen pr_ub=votehat+(1.96*se_pred)  
  
. gen pr_lb=votehat-(1.96*se_pred)  
  
. gra votehat pr_ub pr_lb score, c(l1l1) s(o..) xlab ylab t1(" ")
```



We could also do “out-of-sample” predictions this way as well...