

10. Bivariate regression

LPO.8800: Statistical Methods in Education Research

Sean P. Corcoran

LPO.8800 (Corcoran)

Lecture 10

Last update: November 13, 2023 1 / 77

Last time

- Multivariate distributions
- Measures of association: covariance and correlation
- Scatter diagrams
- Pearson correlation, Spearman correlation
- Effect of variable transformations on correlation
- Hypothesis tests about ρ

LPO.8800 (Corcoran)

Lecture 10

Last update: November 13, 2023 2 / 77

Last time

The sample correlation coefficient r_{xy} is useful for assessing the strength of a linear relationship between two variables:

- Ranges from -1 to +1
- Unit-free

We are often interested, however, in quantifying a relationship in the original units (e.g. *how much* of a change in student outcomes is associated with a unit change in class size?)

Today

Regression is another way of quantifying the relationship between two (or more) variables

- Most widely used tool in social science and policy analysis
- Can be used for descriptive or predictive purposes
- Sometimes useful for causal inference, but only under very strict assumptions (more on this in Lecture 11, and in later courses)

Simple linear regression

Goal: to find the **line of best fit**—the best linear predictor of y given x .
A line is defined by its *slope* (b) and *intercept* (a):

$$y = a + bx$$

In regression analysis these are referred to as the *slope coefficient* and the *constant term*.

Simple vs. multiple linear regression

A **simple regression** involves *two* variables:

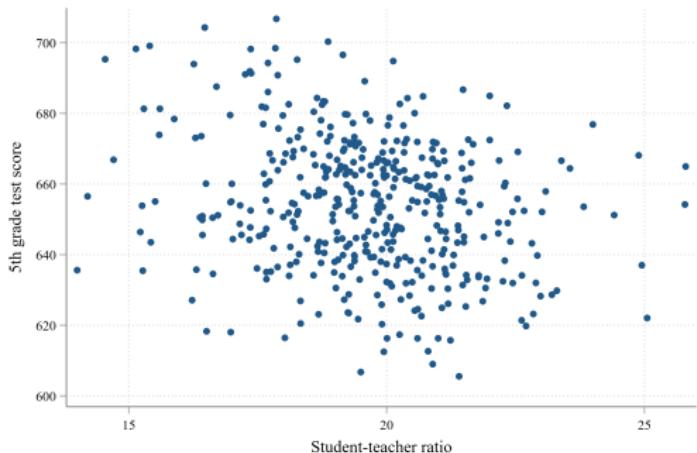
- **response variable (Y)**: often an outcome you are trying to explain or predict (also known as the “dependent” variable)
- **explanatory variable (X)**: often a presumed cause of the outcome, or a predictor of it (sometimes called the “independent” variable—I avoid using this term)

Multiple (multivariate) regression involves *many* variables:

- **response variable (Y)**
- **two or more explanatory variables (X_1, X_2, \dots)**: there is often one of particular interest, plus additional *covariates* or *controls*

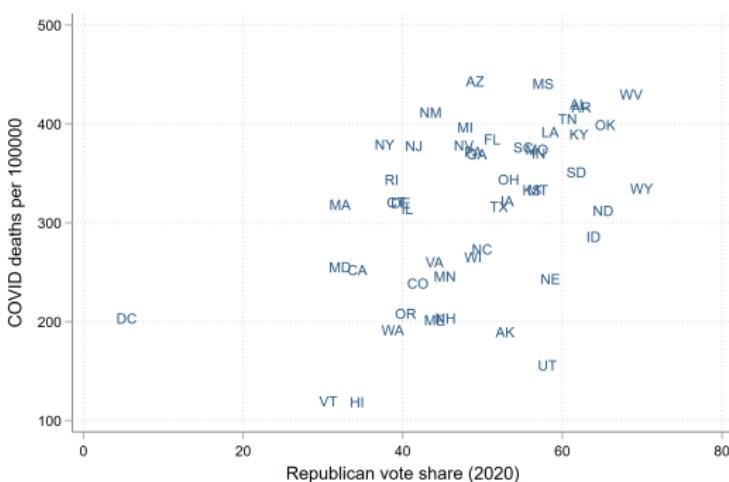
Example 1

From the *caschool* data (California school districts in 1998-99): 5th grade test scores vs. student-teacher ratio



Example 2

COVID death rate and Republican vote share



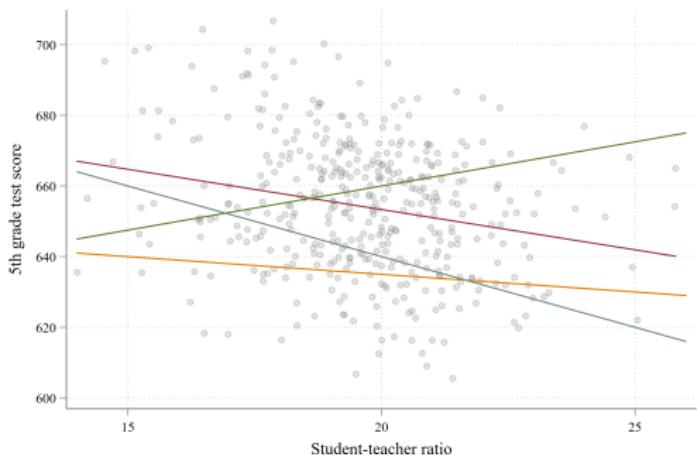
Line of best fit

What makes a particular line the “best fit” or “best predictor” of y given a value of x ? Again, a line is defined by its intercept and slope:

$$y = a + bx$$

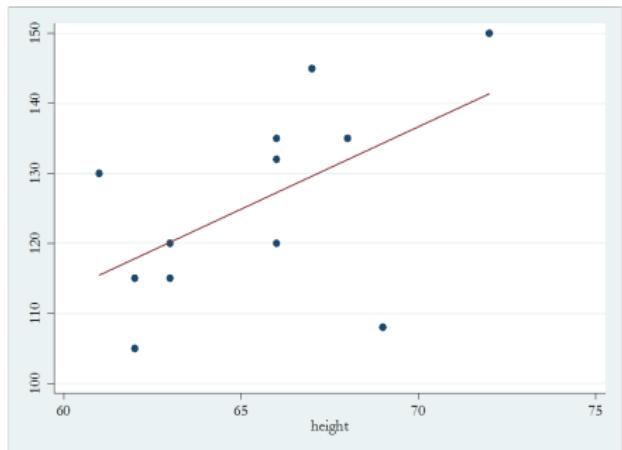
Example 1, cont.

There are lots of candidates for the best fit line, with different values of a and b . Which is the “best”?



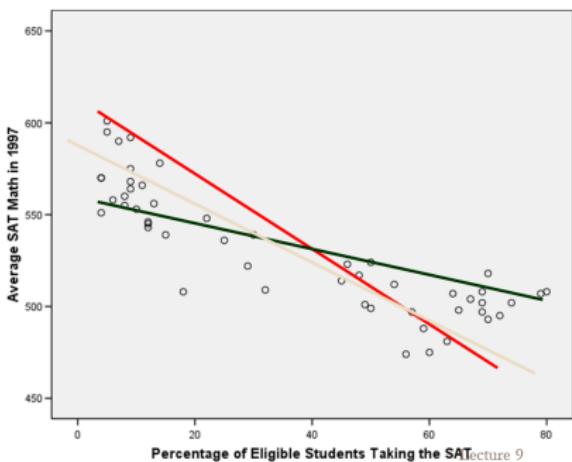
Example 3

Height and weight data



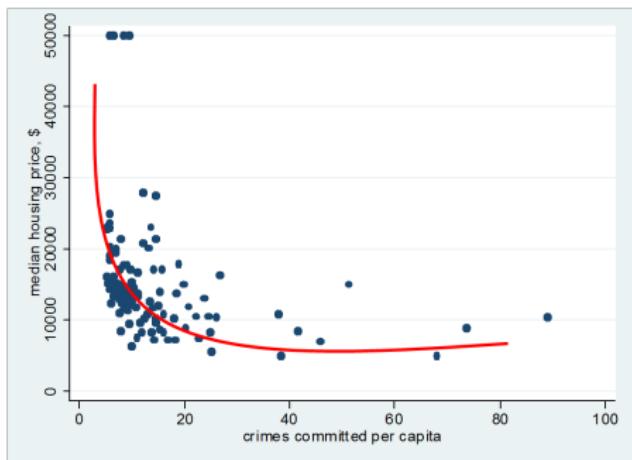
Example 4

State average math SAT scores



Example 5

Crime rate and median house price data; a line is probably not appropriate to describe this relationship



Line of best fit

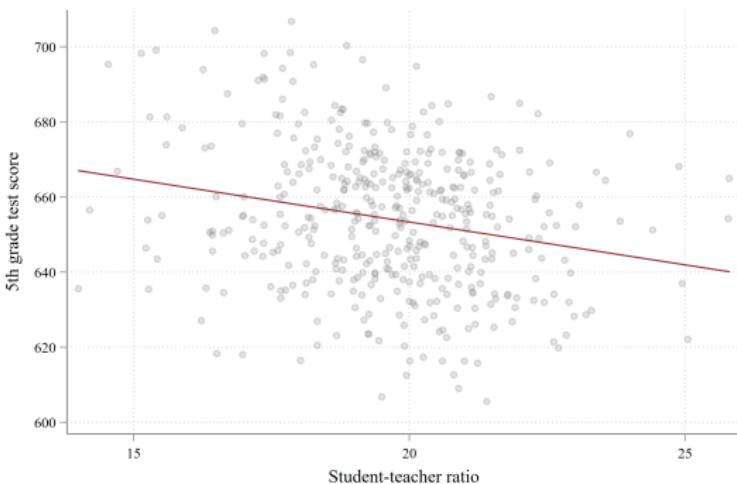
Stata can provide a line of best fit, using two overlaying graphs (scatter and lfit). Using the *caschool* data:

```
twoway (scatter testscr str) (lfit testscr str)
```

It is conventional to put the response variable on the vertical (y) axis, and the explanatory variable on the horizontal (x) axis.

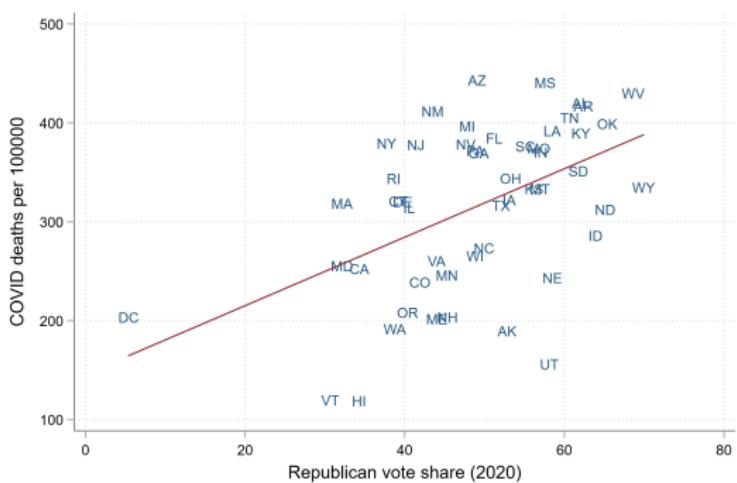
Example 1, cont.

Overlaid graphs scatter and lfit:



Example 2, cont.

Overlaid graphs scatter and lfit:



Line of best fit

In Example 1 the best fit line has an intercept of 698.93, and a slope of -2.28: $\hat{y} = 698.93 - 2.28x$

- The best fit line is also called a **prediction equation**
- \hat{y} is the **predicted value** for y , given x .

We can use the prediction equation to predict y for a specific x value (here, student-teacher ratio):

- Example: suppose $x = 20$ students
- $\hat{y} = 698.93 - (2.28 * 20) = 653.33$
- Predicted 5th grade test score is 653.03 for a class size of 20

Line of best fit

Interpreting the prediction equation: $\hat{y} = 698.93 - 2.28x$

- 698.93: the predicted 5th grade test score when student-teacher ratio is $x = 0$
- -2.28: the predicted *change* in 5th grade test scores when the student-teacher ratio increases by one year.
- Note 5 additional students per teacher would be predicted to change test scores by: $-2.28 * 5 = -11.4$

Least squares

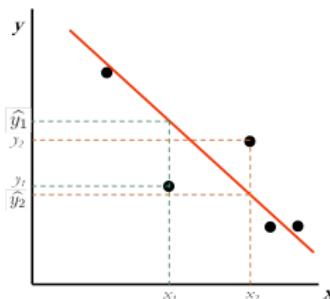
How does one determine the line of “best fit”? Potential criteria:

- Line that passes through as many points as possible?
- Line that minimizes *prediction error*: the gaps (deviations) between the data points and the line?
 - ▶ Problem: positive deviations cancel out negative deviations
 - ▶ Solution: *squared* deviations (compare to variance formula)
 - ▶ Another possibility (rarely used): mean absolute deviation

Least squares

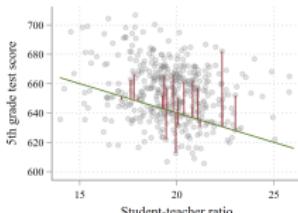
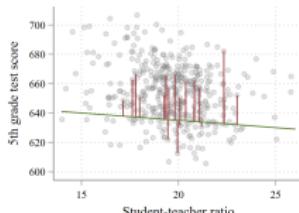
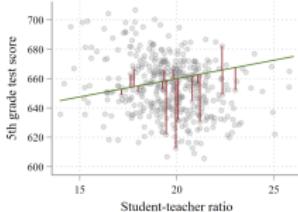
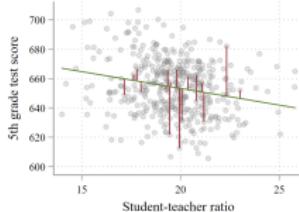
For a given line, we have a set of predictions for y , one for every value of x in the data:

- \hat{y}_1 is the predicted value of y when x is x_1
- \hat{y}_2 is the predicted value of y when x is x_2
- ...and so on, up to \hat{y}_n



Least squares

Four candidate lines for the 5th grade test score data:



Least squares

The line that minimizes the *sum of the squared deviations* between the data points y and the line \hat{y} is the **least squares** or **ordinary least squares (OLS)** regression line:

$$\min_{a, b} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\min_{a, b} \sum_{i=1}^n (y_i - a - bx)^2$$

i.e. choose intercept and slope (a, b) to minimize the sum of the squared deviations $(y_i - \hat{y}_i)$

Least squares

Using calculus it can be shown that the least squares slope (b) and intercept (a) are as follows:

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Least squares

Recall that:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y (n - 1)}$$

Multiply by s_y/s_x :

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y (n - 1)} \frac{s_y}{s_x} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x^2 (n - 1)}$$

$$\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} (n - 1)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = b$$

So: $r_{xy} \frac{s_y}{s_x} = b$

Least squares

$$b = r_{xy} \frac{s_y}{s_x}$$

Notice the slope coefficient will have the same sign (+/-) as the correlation coefficient r_{xy} (re: s_y and s_x are both greater than zero)

Least squares

Notice the least squares solution for a indicates that the point (\bar{x}, \bar{y}) falls on the regression line. That is, line passes through the mean of both variables):

$$\hat{y} = a + b\bar{x}$$

$$\hat{y} = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y}$$

The best prediction of y when $x = \bar{x}$ is \bar{y} .

Least squares

Notice also if $r = 0$, then $b = 0$, and the best prediction of y for *any* value of x is \bar{y} :

$$\hat{y} = a$$

$$a = \bar{y} - b\bar{x} = \bar{y}$$

Put another way, if there is no correlation between y and x , the best prediction of y given x is just \bar{y} .

Regression in practice

To compute the least squares intercept and slope coefficient in Stata use `regress` or `reg yvar xvar` (aka “running a regression”). Examples:

- 5th grade test scores and student-teacher ratio
- Height and weight
- State percent taking SAT and mean math SAT score
- COVID death rates and Republican vote share

Example 1, cont.

From the *caschool* data: 5th grade test scores and student-teacher ratio

. reg testsqr str						
Source	SS	df	MS	Number of obs	=	420
Model	7794.11004	1	7794.11004	F(1, 418)	=	22.58
Residual	144315.484	418	345.252353	Prob > F	=	0.0000
Total	152109.594	419	363.030056	R-squared	=	0.0512
				Adj R-squared	=	0.0490
				Root MSE	=	18.581

testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
str	-2.279808	.4798256	-4.75	0.000	-3.22298	-1.336637
_cons	698.933	9.467491	73.82	0.000	680.3231	717.5428

$$\hat{y} = 698.93 - 2.28x$$

Note: “_cons” refers to the intercept, or **constant term**.

Example 3, cont.

Height and weight data:

. reg weight height						
Source	SS	df	MS	Number of obs	=	12
Model	670.95819	1	670.95819	F(1, 10)	=	4.30
Residual	1558.70848	10	155.870848	Prob > F	=	0.0648
Total	2229.66667	11	202.69697	R-squared	=	0.3009
				Adj R-squared	=	0.2310
				Root MSE	=	12.485

weight	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
height	2.355617	1.135375	2.07	0.065	-.1741566	4.88539
_cons	-28.26327	74.35985	-0.38	0.712	-193.9473	137.4208

$$\hat{y} = -28.263 + 2.356x$$

$$\widehat{\text{weight}} = -28.263 + 2.356 * \text{height}$$

Regression in practice

One can also use descriptive statistics and r to calculate the least squares slope coefficient and intercept:

. corr weight height (obs=12)	
	weight height
weight	1.0000
height	0.5486 1.0000
. tabstat weight height, stat(mean sd n) col(stat)	
variable	mean sd N
weight	125.8333 14.23717 12
height	65.41667 3.315483 12

$$b = r_{xy} \frac{s_y}{s_x}$$

$$b = 0.549 * (14.237 / 3.315) = 2.357$$

$$a = 125.833 - 2.357 * (65.417) = -28.406$$

Example 4, cont.

State average math SAT scores

. reg satm pertak					
Source	SS	df	MS	Number of obs = 51	
Model	40962.5147	1	40962.5147	F(1, 49) = 114.86	
Residual	17474.4657	49	356.621749	Prob > F = 0.0000	
Total	58436.9804	50	1168.73961	R-squared = 0.7010	
				Adj R-squared = 0.6949	
				Root MSE = 18.884	
satm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pertak	-0.9372375	.0874501	-10.72	0.000	-1.112975 -.7614999
_cons	577.1891	4.413822	130.77	0.000	568.3192 586.0591

$$\hat{y} = 577.189 - 0.9372x$$

$$\widehat{\text{SATm}} = 577.189 - 0.9372 * \text{pertak}$$

Regression in practice

Using the regression equation to predict y given a specific value of x , e.g. percent taking SAT of $x = 40$:

- $\hat{y} = 577.189 - 0.9372 * (40) = 539.7$
- Note the value of x used does not have to appear in the actual data.
- What if $x = 40$ *did* appear in the data? Why not use the actual, observed y when $x = 40$ as the prediction?

Regression in practice

Steps for interpreting a regression slope coefficient:

- ① Identify the **explanatory** variable and its units (e.g., height in inches, students per teacher).
- ② Describe a **one-unit increase** in the explanatory variable in everyday language (e.g., one additional student per teacher).
- ③ Identify the **response** or **outcome** variable and its units (e.g., weight in pounds, 5th grade test scores).
- ④ Describe the coefficient as the **change in the outcome** predicted for a one-unit change in the explanatory variable (e.g., an additional student per teacher is predicted to decrease 5th grade test scores by 2.28 points).

Note: be sure your interpretation reflects the appropriate unit of observation (e.g., individual, school, district).

Note: Adapted from Remler & Van Ryzin (2011), chapter 8.

Example 2, cont.

COVID death rate and Republican vote share

```
. reg death_per100k repshare
```

Source	SS	df	MS	Number of obs	=	52
Model	86272.4831	1	86272.4831	F(1, 50)	=	15.73
Residual	274157.456	50	5483.14912	Prob > F	=	0.0002
Total	360429.939	51	7067.25371	R-squared	=	0.2394

death_p~100k	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
repshare	3.461963	.872773	3.97	0.000	1.708947 5.214979

Example 2, cont.

Presenting regression output in publication form

(1)	
Death rate per 100,000	
Republican vote share in 2020	3.46 (0.873)***
Constant	145.896 (44.125)***
R-squared	0.239
N	52

Example 2, cont.

Stata can store regression results using `estimates store` after `regress` (or using the prefix `_eststo:`). Stored regression results can be output to a formatted table or to a file (e.g., `rtf`, `csv`). Examples:

```
_eststo: reg death_per100k repshare  
  
estimates table, b(%4.1f) varlabel varwidth(30) star(0.05  
0.01 0.001) stats(N r2)  
  
esttab using repshare.rtf, se b(3) r2 rtf label
```

These commands come with lots of customizable options.

Regression on a dummy variable

When the explanatory variable is dichotomous (0-1, a “dummy” variable), x only takes on two values. Example: monthly earnings and married vs. unmarried in the `wage2.dta` data.

- $x = 0$: person is unmarried
- $x = 1$: person is married

Regression on a dummy variable

. reg wage married				Number of obs = 935			
Source	SS	df	MS	F(1, 933)	=	17.74	
Model	2848893.49	1	2848893.49	Prob > F		0.0000	
Residual	149867275	933	160629.448	R-squared		0.0187	
Total	152716168	934	163507.675	Adj R-squared		0.0176	
				Root MSE		400.79	
wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]		
married	178.6079	42.41067	4.21	0.000	95.37654	261.8393	
_cons	798.44	40.0786	19.92	0.000	719.7853	877.0947	

There are only two possible predictions:

- When $x = 0$: $\hat{y} = a = 798.44$
- When $x = 1$: $\hat{y} = a + b = 798.44 + 178.61 = 977.05$

Note a is the mean of y for $x = 0$ (unmarried), and b is the *difference in means* for those with $x = 0$ and those with $x = 1$ (married vs. unmarried).

Predicted values and residuals

It is possible to have Stata compute the **predicted values** and **residuals** (prediction errors) after `reg`:

- `predict yhat, xb`
- `predict uhat, resid`

These are called **postestimation** commands:

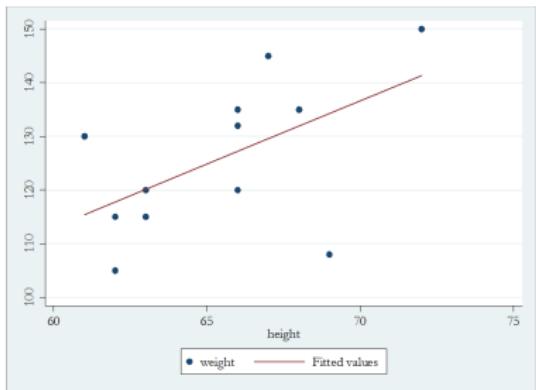
- `xb` refers to the predicted value (\hat{y})
- `resid` refers to the residual (\hat{u} , calculated as $y_i - \hat{y}_i$)

To generate predicted values or residuals only for the estimation sample:

- Add `if e(sample)` to the `predict` command
- Otherwise \hat{y} and \hat{u} will be created for any case in which x is not missing (even if y is missing and the case was not used to estimate the regression)

Example 3, cont.

Height and weight data:



```
twoway (scatter weight height) (lfit weight height)
```

Example 3, cont.

. predict yhat, xb
. predict uhat, resid
. sum yhat uhat
Variable Obs Mean Std. Dev. Min Max
yhat 12 125.8333 7.810006 115.4294 141.3611
uhat 12 -2.99e-08 11.90381 -26.27429 15.43694
. sum weight height
Variable Obs Mean Std. Dev. Min Max
weight 12 125.8333 14.23717 105 150
height 12 65.41667 3.315483 61 72

Things to note:

- Mean of the residuals (\hat{u}) is always zero.
- Mean of the predicted values (\hat{y}) is always \bar{y}

These follow algebraically from least squares.

Stata margins command

The `margins` command can provide predictions from the regression for given levels of x . This command has more advanced applications, but here is a simple prediction for a student-teacher ratio of 20:

```
. margins, at(str=20)
```

Adjusted predictions		Number of obs = 420			
Model VCE : OLS					
Expression : Linear prediction, predict()	at : str = 20				
	Margin	Delta-method			
_cons	653.3368	.9229285	707.90	0.000	651.5226 655.1509

Measuring fit

How well does the regression line fit the data?

- Mechanically, the least squares intercept and slope can be calculated for any set of data points (x, y)
- The line of best fit (OLS) is not necessarily a *good* fit
- Least squares minimizes the sum of the squared residuals, but performs better with some data than others

R^2 , the **coefficient of determination**, is a measure of the goodness of fit.

R^2 is the proportion of the total variation in y from its mean that is “explained” (predicted) by x .

Equivalently, it is the proportionate reduction in the variability of y from its mean that can be achieved using the predicted, rather than the actual, y .

The total variation in y around its mean is the **total sum of squares (TSS)**:

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

Using the predicted y instead of the actual, the **model sum of squares (SSM)** is:

$$SSM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

R^2

The R^2 is therefore:

$$R^2 = \frac{SSM}{TSS}$$

R^2 is **always between 0 and 1**

The model sum of squares (SSM) is sometimes called the “explained” sum of squares.

R^2

The “unexplained” variation in y is the **sum of squared errors (SSE)**, seen before (a.k.a. the residual sum of squares SSR):

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

It makes sense that the R^2 should be related to the SSE, which we aim to minimize when finding the best fit line. In fact, we can write R^2 as:

$$R^2 = 1 - \frac{SSE}{TSS}$$

Or:

$$R^2 = 1 - \frac{SSE}{TSS} = \frac{TSS - SSE}{TSS}$$

Written this way, it's easy to see why R^2 is the proportionate reduction in prediction error:

- TSS would be your SSE if you simply used \bar{y} to predict y .
- SSE is your sum of squared errors after fitting the regression line.
- $(TSS-SSE)$ is your reduction in prediction error.
- $(TSS-SSE)/TSS$ is this reduction expressed as a proportion of the original TSS.

Example 1, cont.

Example using 5th grade test scores and student-teacher ratio:

```
. reg testscr str
```

Source	SS	df	MS	Number of obs	=	420
Model	7794.11004	1	7794.11004	F(1, 418)	=	22.58
Residual	144315.484	418	345.252353	Prob > F	=	0.0000
Total	152109.594	419	363.030056	R-squared	=	0.0512
				Adj R-squared	=	0.0490
				Root MSE	=	18.581

testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
str_cons	-2.279808 698.933	.4798256 9.467491	-4.75 73.82	0.000 0.000	-3.22298 680.3231 -1.336637 717.5428

$$R^2 = \frac{SSM}{TSS} = 0.0512$$

Example 3, cont.

Example using height and weight

. reg weight height				Number of obs = 12			
Source	SS	df	MS	F(1, 10) = 4.30	Prob > F = 0.0648	R-squared = 0.3009	Adj R-squared = 0.2310
Model	670.95819	1	670.95819				
Residual	1558.70848	10	155.870848				
Total	2229.66667	11	202.69697				

weight	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
height	2.355617	1.135375	2.07	0.065	-.1741566 4.88539
_cons	-28.26327	74.35985	-0.38	0.712	-193.9473 137.4208

$$R^2 = \frac{SSM}{TSS} = 0.3009$$

Example 4, cont.

Example using state SAT average scores

. reg satm pertak				Number of obs = 51			
Source	SS	df	MS	F(1, 49) = 114.86	Prob > F = 0.0000	R-squared = 0.7010	Adj R-squared = 0.6949
Model	40962.5147	1	40962.5147				
Residual	17474.4657	49	356.621749				
Total	58436.9804	50	1168.73961				

satm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pertak	-.9372375	.0874501	-10.72	0.000	-.1112975 -.7614999
_cons	577.1891	4.413822	130.77	0.000	568.3192 586.0591

$$R^2 = \frac{SSM}{TSS} = 0.7010$$

Mean squared error (MSE)

A related measure is the **mean squared error (MSE)**:

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2} = \frac{\sum_{i=1}^n \hat{u}_i^2}{n - 2} = \frac{SSE}{n - 2}$$

The MSE is the *average* squared deviation of the predicted y from the actual y (uses $n - 2$ in the denominator). Note the numerator is the sum of squared errors (SSE).

Note: least squares minimizes $\sum \hat{u}_i^2$, so it also minimizes MSE

Root mean squared error (RMSE)

The square root of the MSE is the **root mean squared error (RMSE)** or the “standard error of the estimate”:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{\sum_{i=1}^n \hat{u}_i^2}{n - 2}} = \sqrt{\frac{SSE}{n - 2}}$$

Just as the standard deviation can be interpreted (intuitively, not literally) as the average deviation of y from its mean, the RSME can be interpreted (intuitively, not literally) as the average deviation of y from its *predicted value*

Other properties of fit measures

Some other mathematical properties of fit measures:

- R^2 is the square of the correlation between predicted and actual y (call this R).
- In simple linear regression, $R = |r_{xy}|$ (the absolute value of the correlation between y and x).
- R is positive—a negative correlation between the predicted and actual y 's would not make sense. (You would have a very bad prediction if your predictions moved in the opposite direction of the actual data).

Side note on notation

- Up to this point I have used a and b to represent the intercept and slope of the best fit line.
- We may wish to use sample data on (y, x) to make inferences about their relationship in the *population*. In this case we could denote α and β as the intercept and slope in the population that we are seeking to estimate. a and b (or \hat{a} and $\hat{\beta}$) are then estimators of the population intercept and slope.
- “ $\hat{y} = a + bx$ ” or “ $\hat{y} = \hat{a} + \hat{\beta}x$ ” are sometimes used to describe the *fitted* prediction equation, or the results of OLS.

Conditional mean interpretation

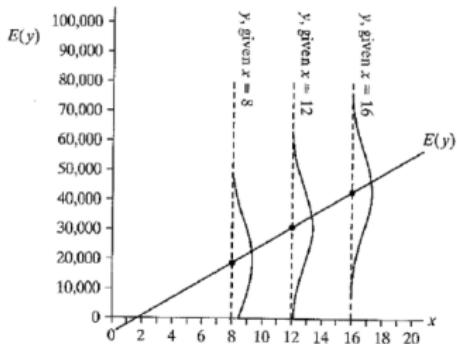
- A linear regression can always be used for description or prediction purposes.
- If the relationship between x and the **mean** of y in the population is **linear** then $E(y|x) = \alpha + \beta x$ (the **population regression function**)
- In this case, our regression is an estimate of how the **conditional mean** of y [$E(y|x)$] varies with x . We use a and b to estimate the population parameters α and β .
- Again, only makes sense if the mean of y is linearly related to x or if a line is a good approximation to the population regression function.
 - ▶ Least squares (minimizing SSE) provides the best approximation to the conditional mean function, even when it is nonlinear.

Conditional mean interpretation

- The standard deviation of y for a given level of x is the **conditional standard deviation**.
- Assume for now that y has a normal distribution at each level of x , and a *constant* standard deviation (σ)
- In this context the RMSE is an estimate of the standard deviation of y given x (i.e., σ)

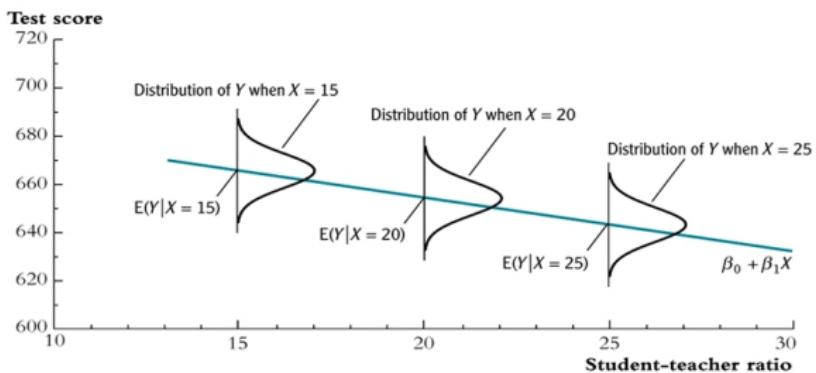
Conditional mean interpretation

Example from Agresti and Finlay: earnings, conditional on years of education



Conditional mean interpretation

Test score vs. student-teacher ratio:



Inferences about β

If we are using a sample of y and x to *estimate* the intercept and slope coefficient in the population, we need to quantify our uncertainty in the same way we did for \bar{x} :

- Confidence intervals for β
- Hypothesis tests about β

To do so we need to know the *sampling distribution* of the slope estimator (b). As was true for \bar{x} , this requires some assumptions:

Inferences about β

- ① The data points (x, y) represent a **random sample** from the population.
- ② In the population, the mean of y conditional on x can be written as:
$$E(y|x) = \alpha + \beta x$$
- ③ The conditional variance of y given x is **constant** (σ^2)—called *homoskedasticity*
- ④ The conditional distribution of y given x is **normal** (this assures our sampling distribution is normal—can be relaxed with a large n).

Inferences about β

To construct confidence intervals and conduct hypothesis tests for β , we need the **standard error** of b . When assumptions 1-3 above hold, this is:

$$se_b = \frac{RMSE}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Note: can also be written

$$se_b = \sqrt{\frac{SSE/(n-2)}{(n-1)s_x^2}}$$

Inferences about β

$$se_b = \sqrt{\frac{SSE/(n-2)}{(n-1)s_x^2}}$$

This expression makes clear what affects the standard error of b :

- SSE (overall fit of the regression line)
- s_x^2 (overall variation in x)
- n (sample size)

Inferences about β

Under the above assumptions, the population slope β will lie within $\pm t_{\alpha/2}$ standard errors of b $(1 - \alpha)\%$ of the time, using a t -distribution with $n - 2$ degrees of freedom. So a $(1 - \alpha)\%$ confidence interval for β is:

$$b \pm t_{\alpha/2} \times se_b$$

This is given in the Stata output (for a default 95% CI). Note the α significance level is not the same thing as the regression intercept, also sometimes denoted using α .

Example 1, cont.

Example using 5th grade test scores:

```
. reg testscr str
```

Source	SS	df	MS	Number of obs	=	420
Model	7794.11004	1	7794.11004	F(1, 418)	=	22.58
Residual	144315.484	418	345.252353	Prob > F	=	0.0000
Total	152109.594	419	363.030056	R-squared	=	0.0512
				Adj R-squared	=	0.0490
				Root MSE	=	18.581
testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
str_cons	-2.279808 698.933	.4798256 9.467491	-4.75 73.82	0.000 0.000	-3.22298 680.3231	-1.336637 717.5428

Inferences about β

Under $H_0 : \beta = 0$, the test statistic $t = b/se_b$ has a t -distribution with $n - 2$ degrees of freedom. We can use this to test the alternative hypothesis that the slope is significantly different from zero.

The t -statistic and (two-sided) p -value are given in the Stata output. In this case we can reject H_0 and conclude that the slope coefficient is statistically different from zero.

Example 1, cont.

Example using 5th grade test scores:

```
. reg testscr str
```

Source	SS	df	MS	Number of obs	=	420
Model	7794.11004	1	7794.11004	F(1, 418)	=	22.58
Residual	144315.484	418	345.252353	Prob > F	=	0.0000
Total	152109.594	419	363.030056	R-squared	=	0.0512
				Adj R-squared	=	0.0490
				Root MSE	=	18.581
testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
str_cons	-2.279808 698.933	.4798256 9.467491	-4.75 73.82	0.000 0.000	-3.22298 680.3231	-1.336637 717.5428

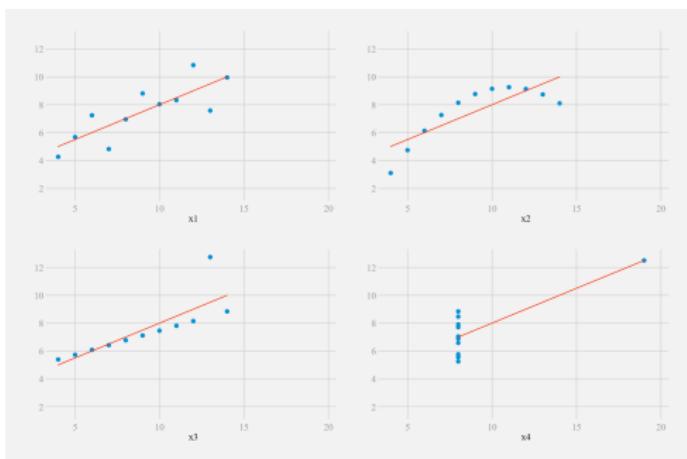
Regression diagnostics

A few tests of basic assumptions underlying the simple linear regression:

- Does the linearity assumption appear to hold?
- Does the variance in the residuals appear *homoskedastic*?

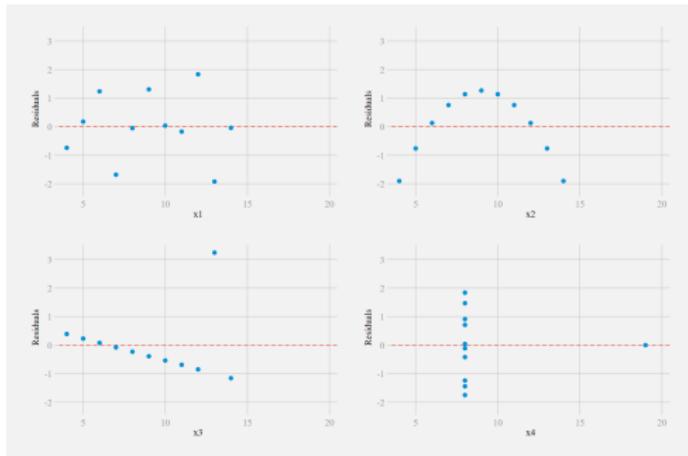
Consider the following four scatterplots (*anscombe.dta*). All four datasets contain $n = 11$ with the same estimated intercept and slope.

Regression diagnostics



Regression diagnostics

A **residual plot** can be revealing: plot \hat{u} against x :



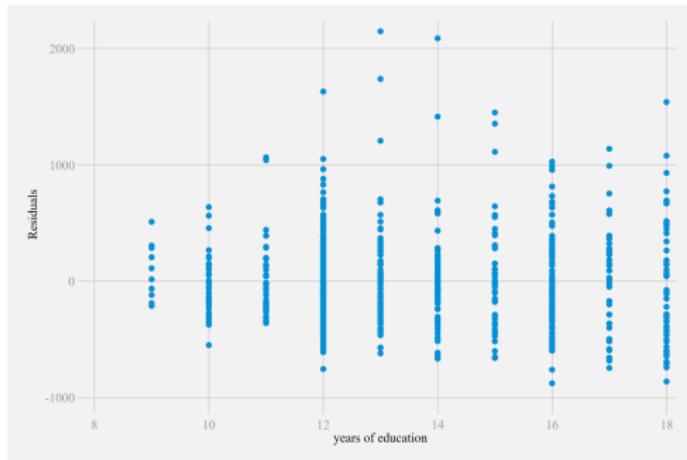
Regression diagnostics

Observations:

- If the relationship is truly linear, the residuals should form a roughly rectangular band around zero.
- Nonlinear relationships will yield a different pattern (e.g. Case 2 above)
- Cases 3 and 4 above illustrate influence of an outlier
- If the vertical spread of the residuals varies with x , the homoskedasticity assumption may be unsupportable (e.g., a "funnel" shape is a common type of *heteroskedasticity*)

Heteroskedasticity

Years of education and monthly earnings (*wage2* data):



Heteroskedasticity

What is the consequence of heteroskedasticity? Most importantly, the standard error calculation for b above is wrong. This means your inference is wrong (confidence intervals, significance tests, etc.) More on this in Regression I.

Influential observations

The estimated intercept and slope in a regression can be heavily influenced by *outliers*. An “influential observation” is one that has a significant effect on the intercept and/or slope.

- Outlying data points with an x value near \bar{x} may affect the intercept but are not likely to affect the slope much.
- Outlying data points with an x value far from \bar{x} that do not stray far from the regression line are unlikely to affect the slope or intercept.

Influential observations

Define $x - \bar{x}$ as *leverage*. Data points with large residuals and high leverage have the most influence on the estimated slope and intercept.

Cook's influence values are an index of the extent to which the data point influences the slope and intercept. To obtain these use `predict varname, cook`. As a rule of thumb, values > 1 have “large” influence.

Next time

- Nonlinear regressions
- Multivariate relationships and “controlling” for other variables (an introduction and preview of Regression I)
- When can regression be considered causal?