JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Simple Linear Regression

John McGready
Johns Hopkins University

Topics

Simple linear regression

Review of equation of a line

Least squares equation

Interpreting results

Estimation and inference in regression

Correlation

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section A

Motivation for Linear Regression

Simple linear regression

- *The t-test compares means in two populations*
- *ANOVA compares means amongst more than two populations with one test*
- *Linear Regression allows for the populations compared to be defined by a continuous "grouping variable"*

Example

Association between Hb levels (g/dL) and packed cell volume (PCV) in 21 patients 20-67 years old

- *A snippet of the data:*

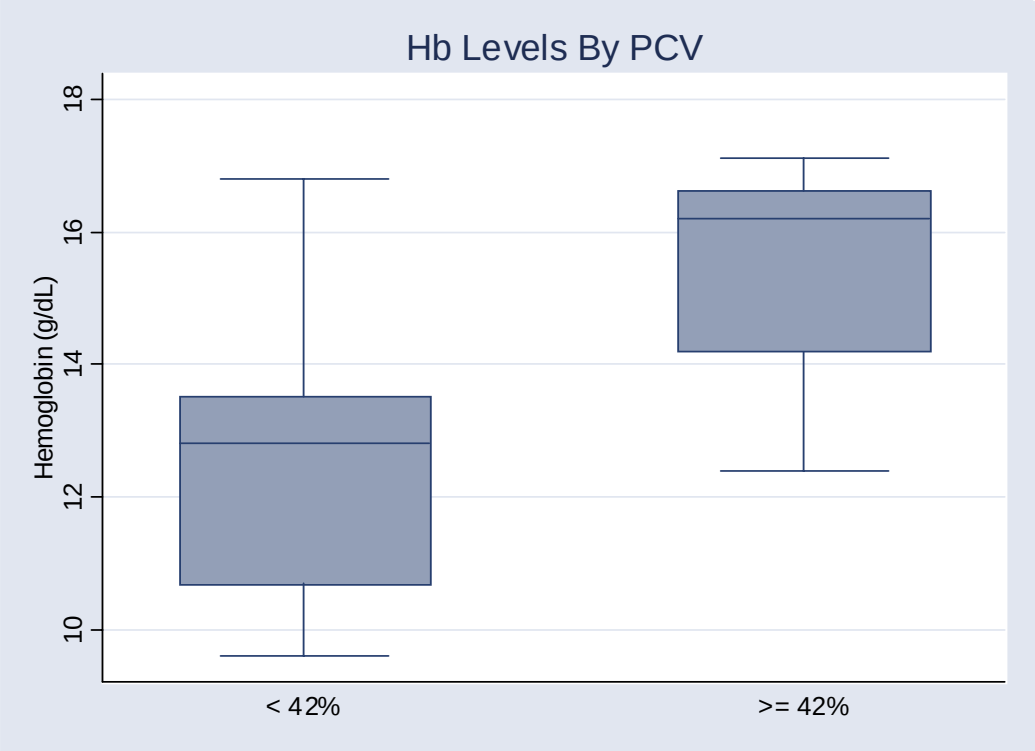
	Hb	PCV
1.	12	35
2.	10.7	39
3.	12.4	47
4.	14.2	53
5.	13.1	30
6.	10.5	30
7.	9.6	25
8.	12.5	33
9.	13.5	35
10.	13.9	40

Example TLC and Height

So how to see if mean Hb varies with PCV?

- *Could dichotomize PCV at median and compare Hb between two height groups using a two-sample t-test*
- *Could multi-categorize subjects into say three PCV groups, say PCV tertiles:*
  - Could do ANOVA to test for differences in Hb and follow-up with (up to) three t-tests/confidence intervals

Boxplot—Hemoglobin levels by PCV group



Continued

Mean hemoglobin levels by PCV group

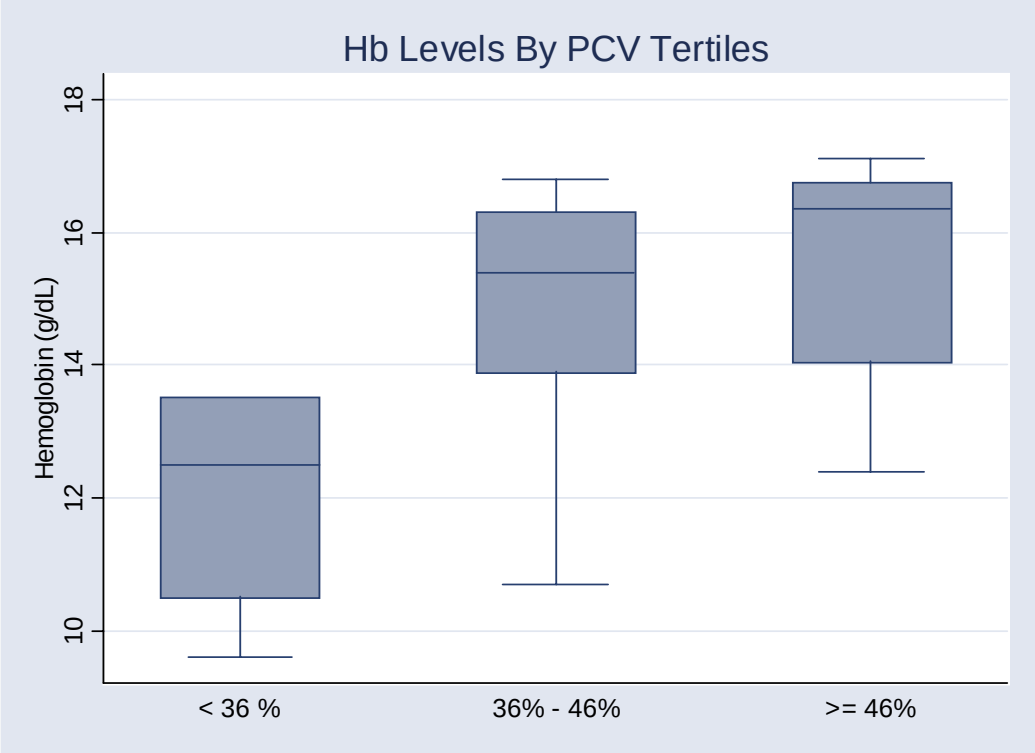
PCV Group	n	Mean Hb	SD
< 42%	10	12.6	2.0
≥ 42%	11	15.5	1.5

Continued

Results from a t-test

```
. ttest Hb, by( PCV_group) unequal

Two-sample t test with unequal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      10       12.61    .6548027    2.070668    11.12873    14.09127
       1 |      11    15.53636    .4462517    1.480049    14.54205    16.53067
---------+--------------------------------------------------------------------
combined |      21    14.14286     .500836    2.295119    13.09813    15.18758
---------+--------------------------------------------------------------------
    diff |           -2.926364    .7924059                -4.604809   -1.247918
------------------------------------------------------------------------------
Satterthwaite's degrees of freedom:  16.1635

                     Ho: mean(0) - mean(1) = diff = 0

     Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
       t =  -3.6930                  t =  -3.6930                  t =  -3.6930
   P < t =   0.0010          P > |t| =   0.0019          P > t =   0.9990
```

Boxplot—hemoglobin levels by PCV tertiles



Continued

Mean hemoglobin levels by PCV group

PCV Group	n	Mean Hb	SD
< 36%	7	12.1	1.5
36% - 46%	6	14.8	2.2
≥ 46%	8	15.5	1.7

Continued

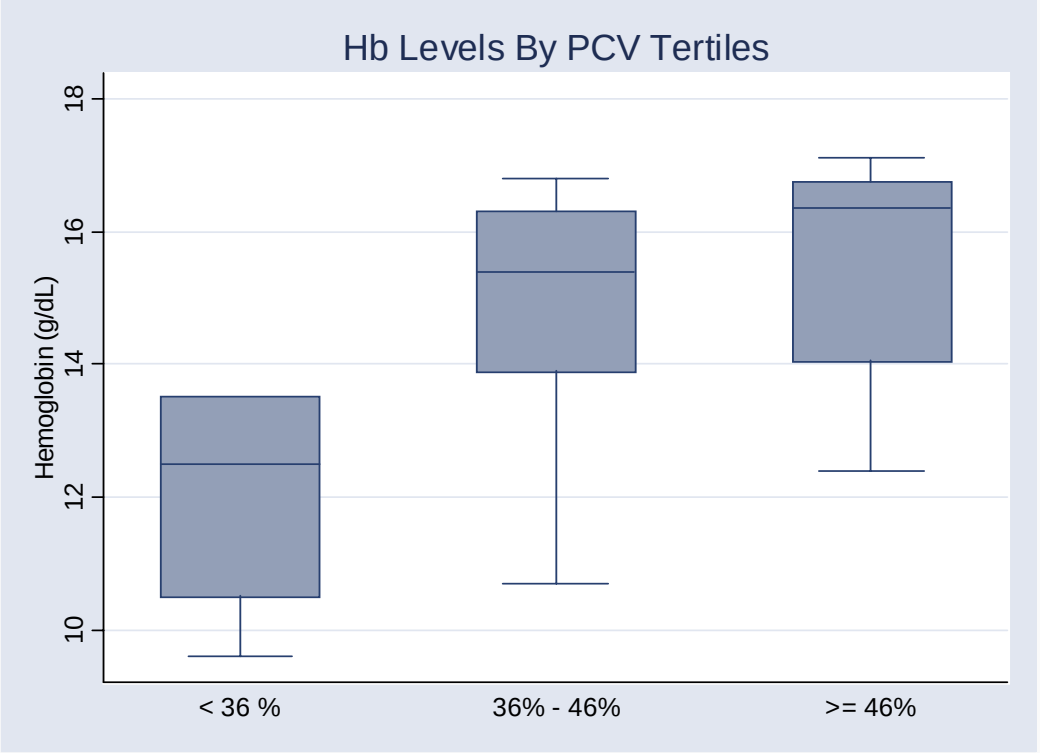
Results from ANOVA testing for mean Hb differences between PCV tertiles: p = .006

So there is a statistical association—but to get estimates of magnitude we would now have to estimate three mean differences and confidence intervals

The precision would be poor, as there are a small number of observations within each PCV tertile

Continued

Recall—hemoglobin levels by PCV tertiles



Can We Do This Without Categorizing PCV?

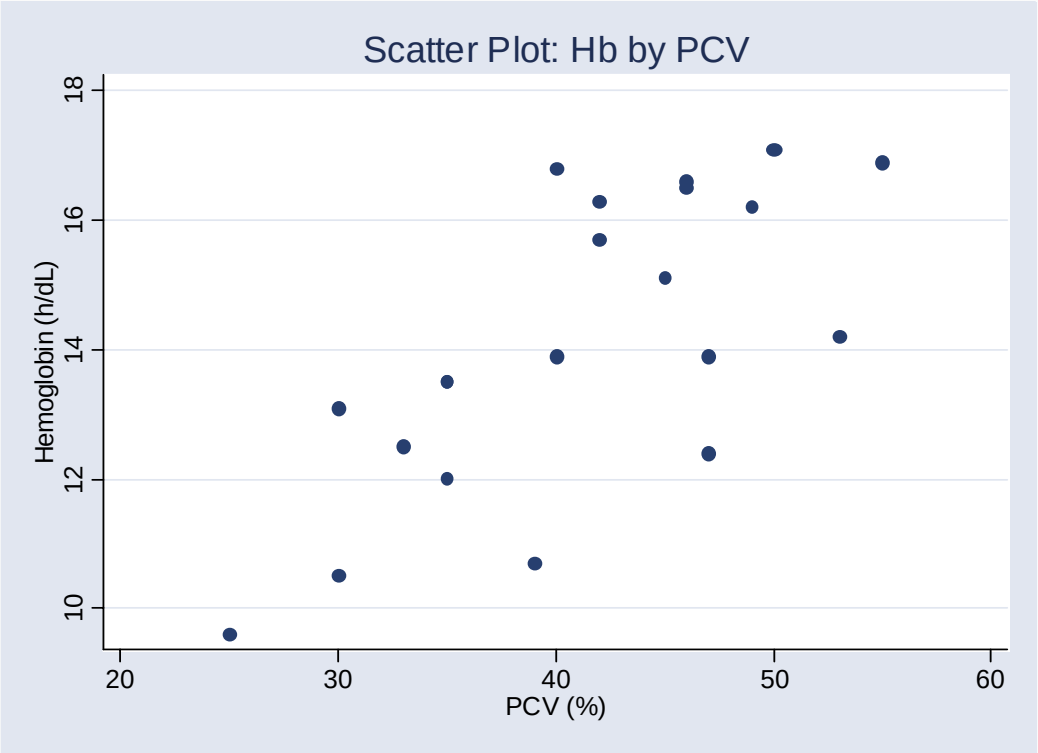
Generally speaking, there appears to be a positive association between Hb and PCV—is there anyway we could estimated the magnitude of this association without breaking the continuous measure of PCV into arbitrary subgroups?

If we could get an overall estimate, we could estimate it using all 21 observations and get better precision

Continued

Can We Do This Without Categorizing PCV?

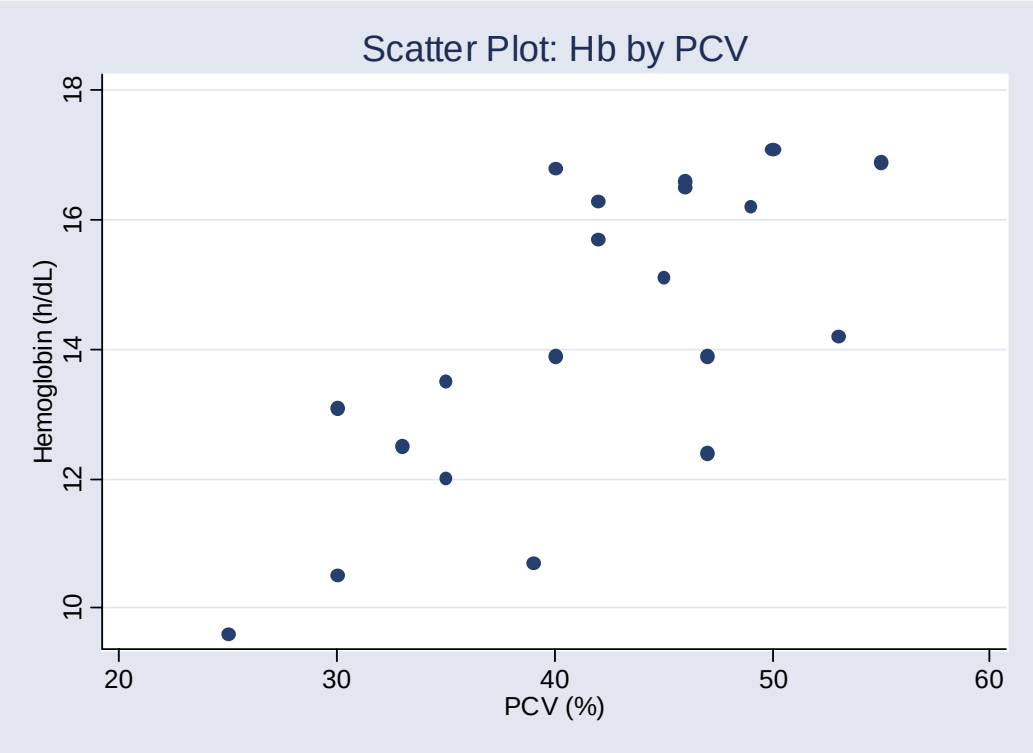
A scatterplot (two dimensional graph) of Hb versus PCV allows us to visualize Hb/PCV relationship without categorizing PCV



Continued

Can We Do This Without Categorizing PCV?

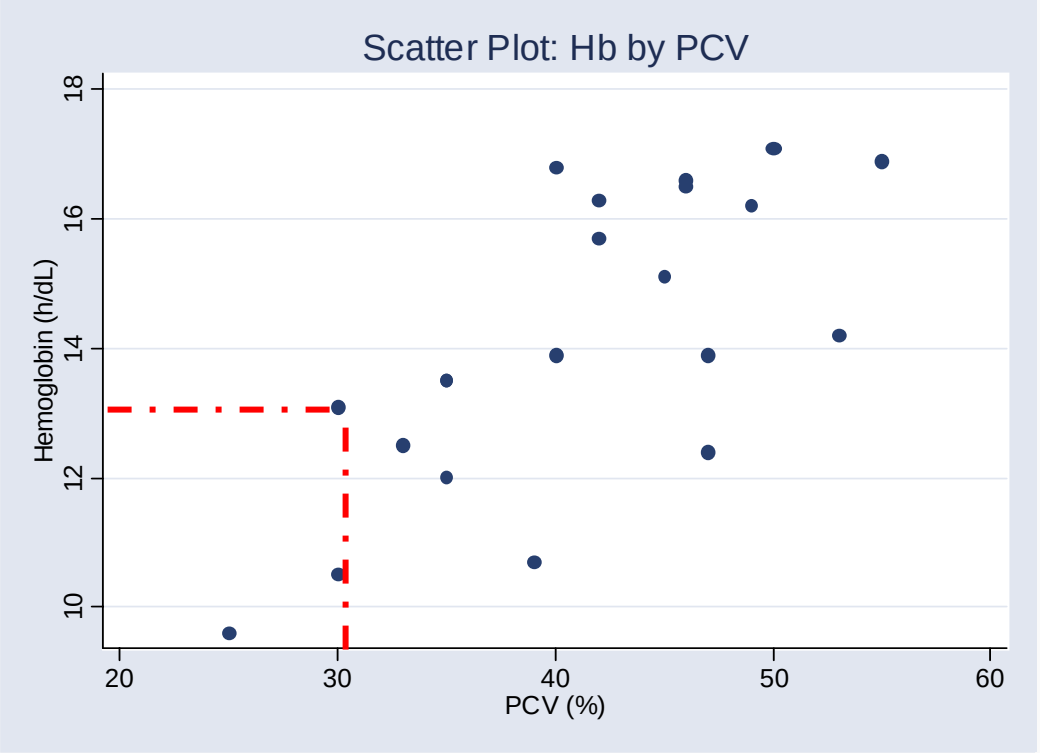
Each point represents an Hb/PCV combination for one subject—there are 21 points



Continued

Can We Do This Without Categorizing PCV?

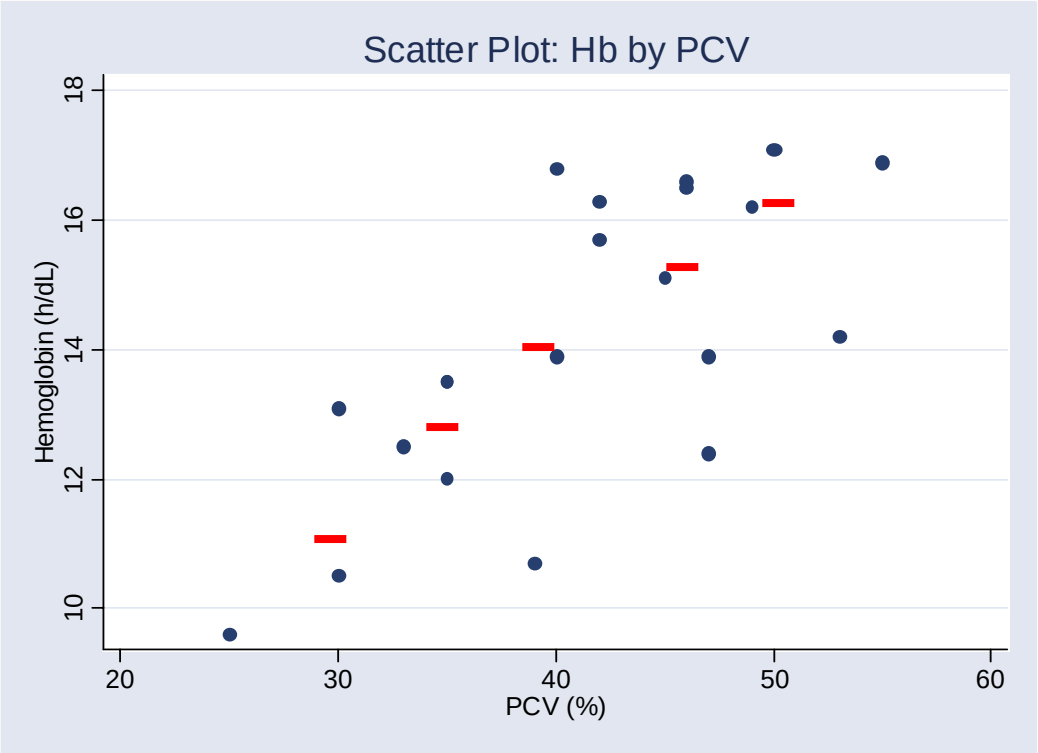
Example—subject with PCV of 30%, Hb of 13 g/dL



Continued

Can We Do This Without Categorizing PCV?

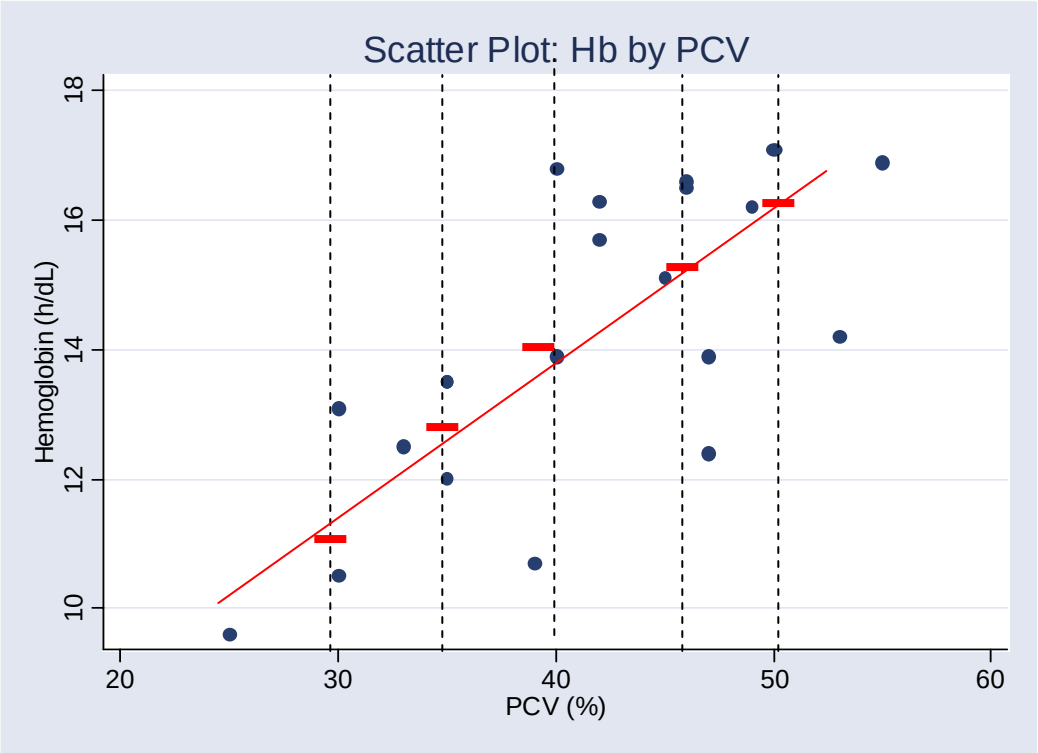
Is there some inherent structure in the relationship between Hb means and TLC that we can exploit?



Continued

Can We Do This Without Categorizing PCV?

Is there some inherent structure in the relationship between Hb means and TLC that we can exploit?



Linear Regression

Linear regression is a general method for estimating/describing association between a continuous outcome variable (dependent) and potentially multiple predictors (continuous, binary, etc.) in one equation—the equation of a line

Continued

Linear Regression

Linear regression allows researcher to . . .

- ★ *Estimate magnitude, strength, and significance of the relationship between a continuous outcome and a predictor from a sample*
- ★ *Develop an equation to predict the outcome value given values of the predictor(s) for observations not in sample*

Continued

Linear Regression

Linear regression allows researcher to . . .

- ★ *Estimate magnitude, strength, and significance of the relationship between a continuous outcome and a predictor from a sample using a single line*
- ★ *Line can be estimated using all of the data*
- ★ *We only need estimate two quantities to fully specify the line (we review equation of a line in the next section)*

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section B

Equation of a Line

Recall, from algebra, there are two values which uniquely define any line

- *y-intercept—where the line crosses the y-axis (when $x = 0$)*
- *Slope—the "rise over the run"—how much y changes for each one unit change in x*

Continued

Recall, from algebra, there are two values which uniquely define any line

$y = mx + b$

- ★ *b = y-intercept*
- ★ *m = slope*

Continued

Of course, statisticians must have their own notation!

$$y = b_o + b_1 x$$

- b_o = *y-intercept*
- b_1 = *slope*

Continued

The Equation of a Line

And of course, not all statisticians are in agreement!

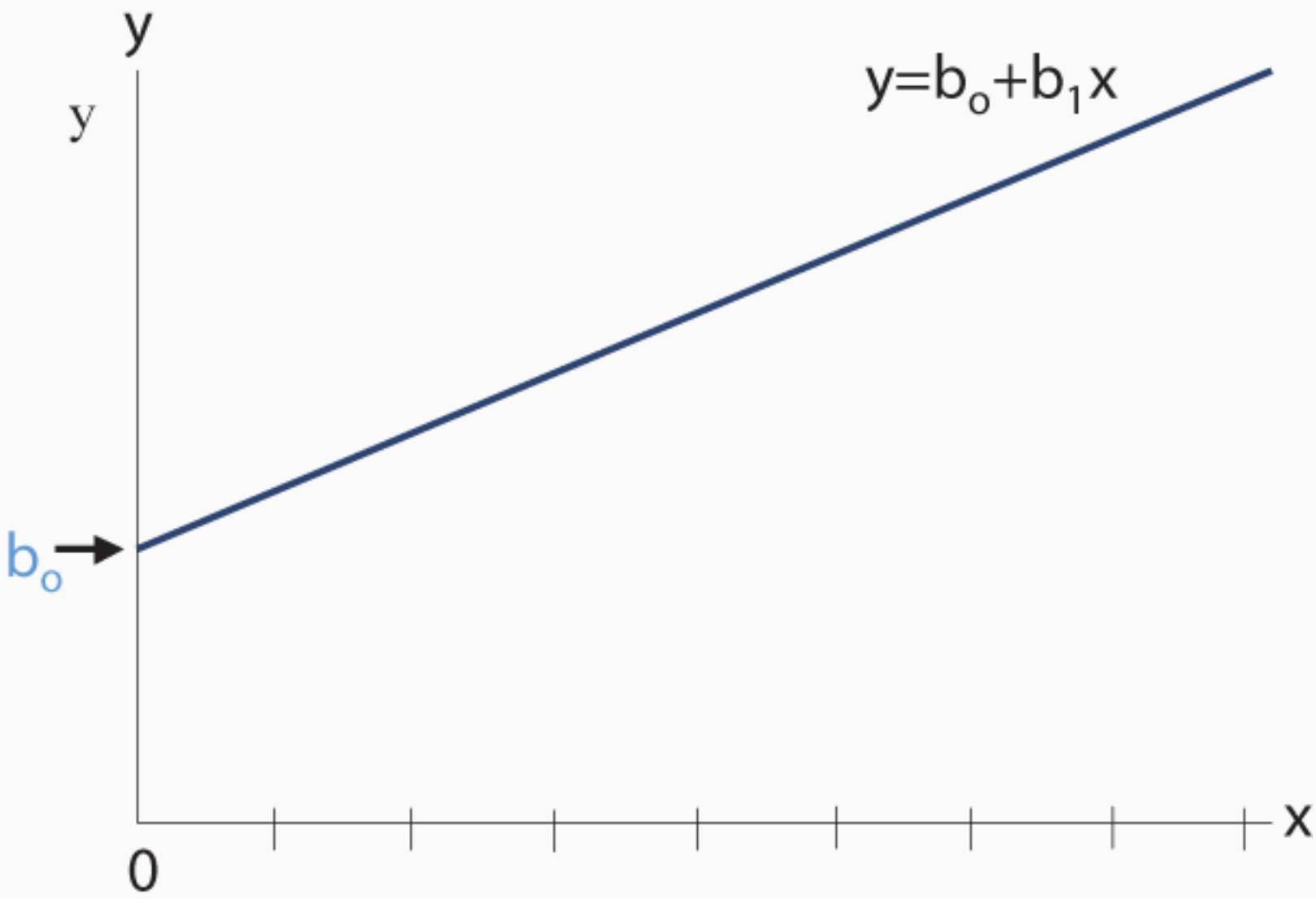
- $y = a + bx$
- $y = \alpha + \beta x$
- $y = \beta_0 + \beta_1 x$

The Intercept, b_o

The intercept b_o is the value of y when x is 0

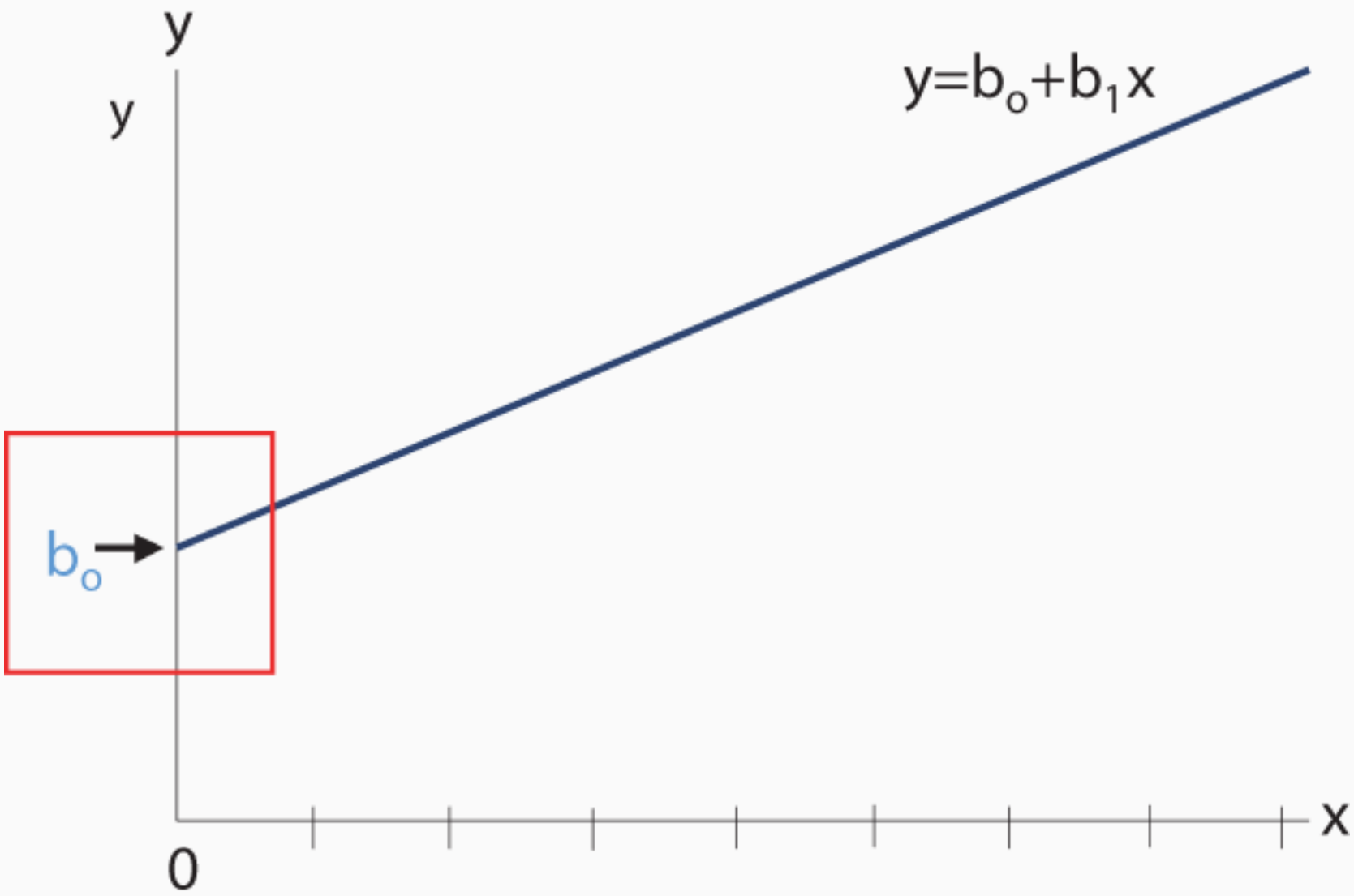
- ★ *It is the point on the graph where the line crosses the y (vertical) axis*

The Equation of a Line



Continued

The Equation of a Line

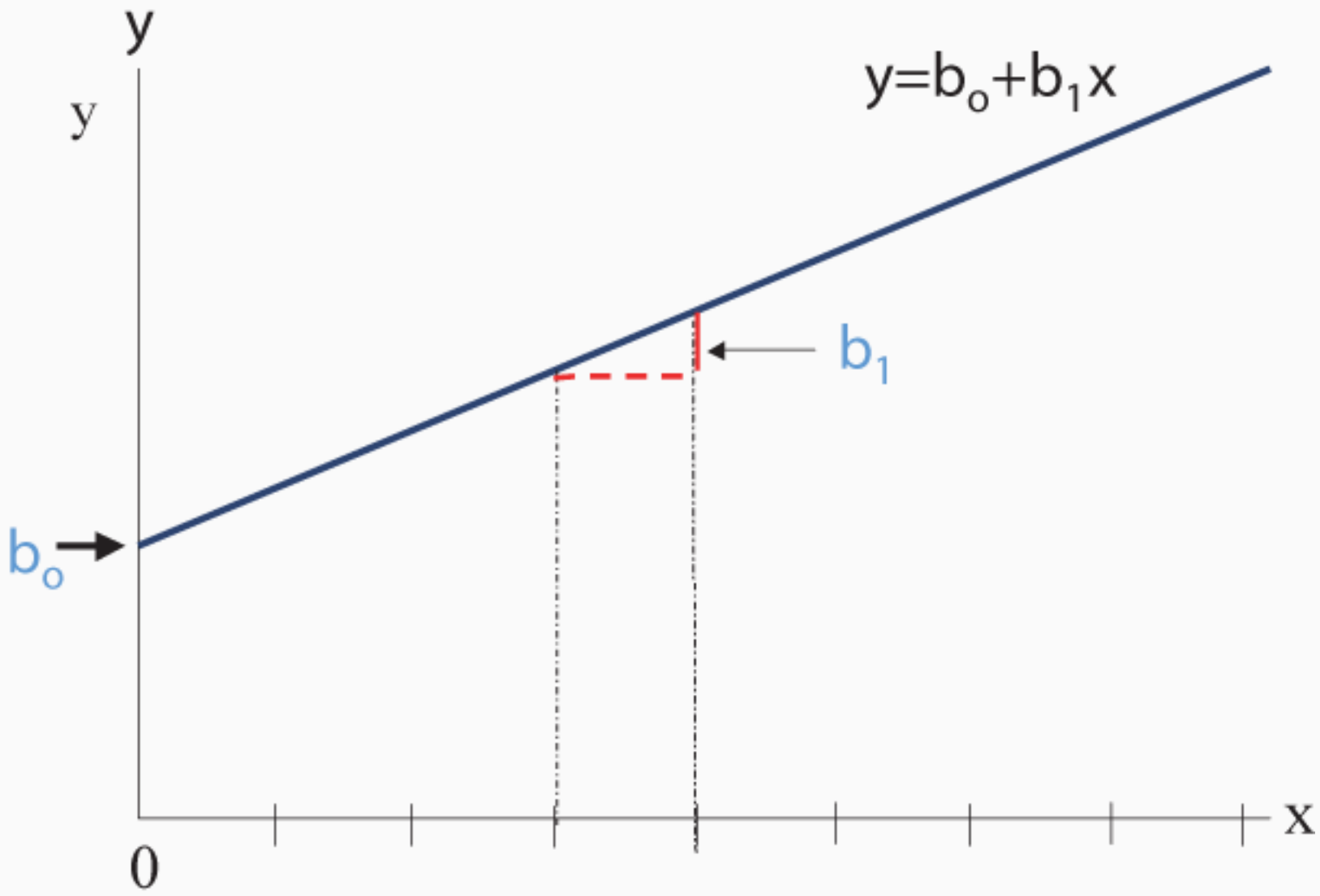


Continued

The Slope, b_1

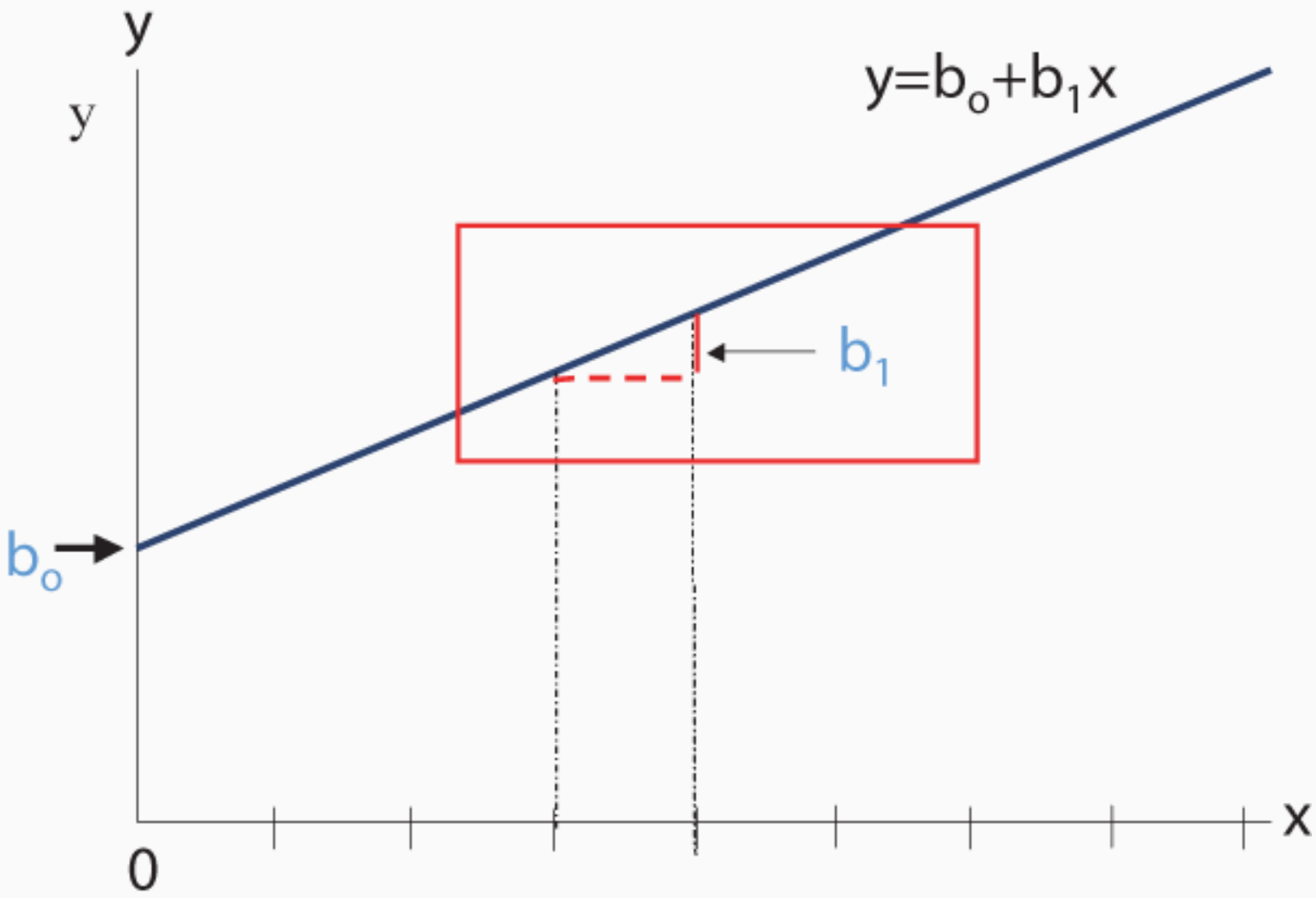
The slope b_1 is the change in y corresponding to a unit increase in x

The Equation of a Line



Continued

The Equation of a Line



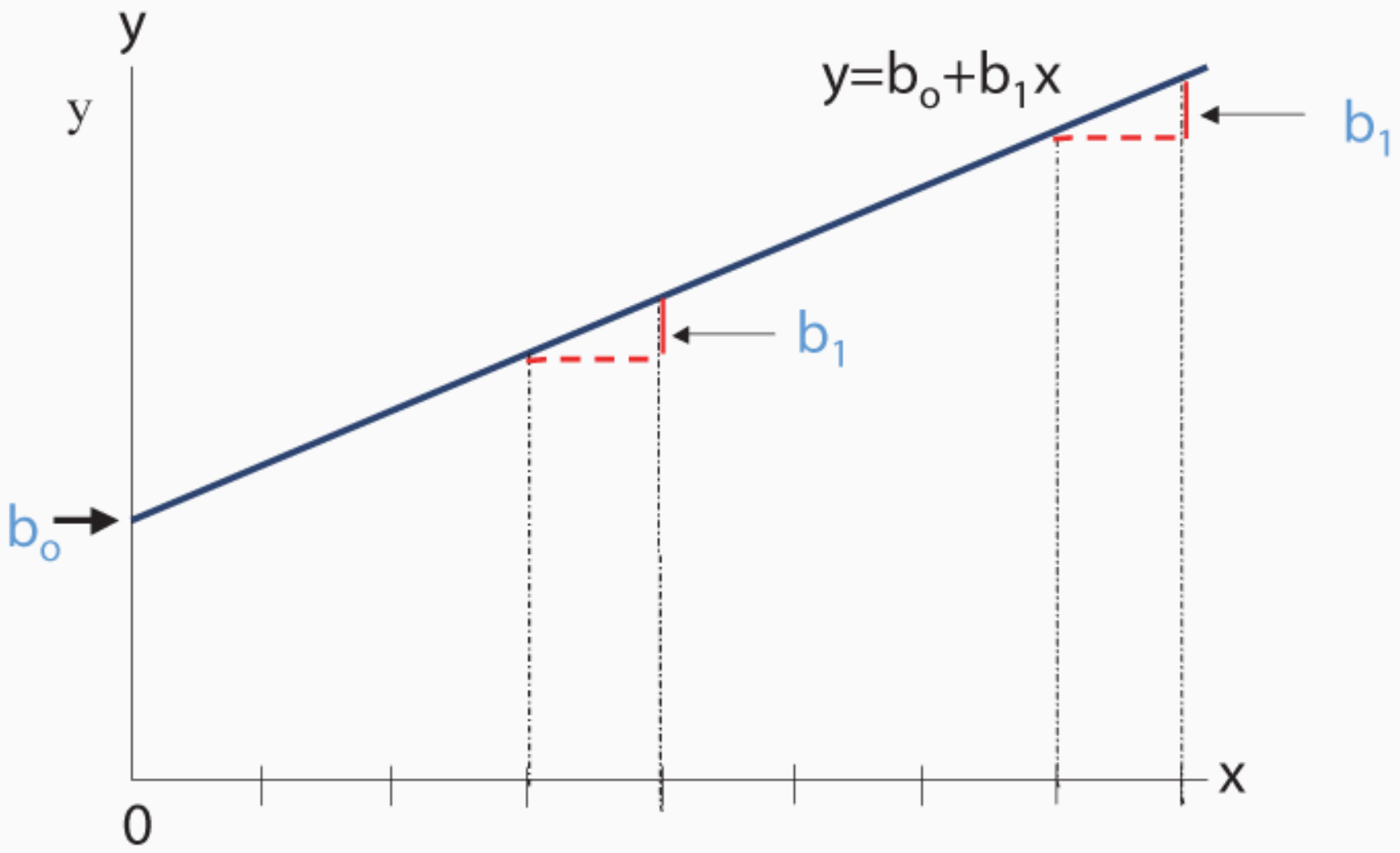
Continued

The Slope, b_1

The slope b_1 is the change in y corresponding to a unit increase in x

This change is the same across the entire line

The Equation of a Line



Continued

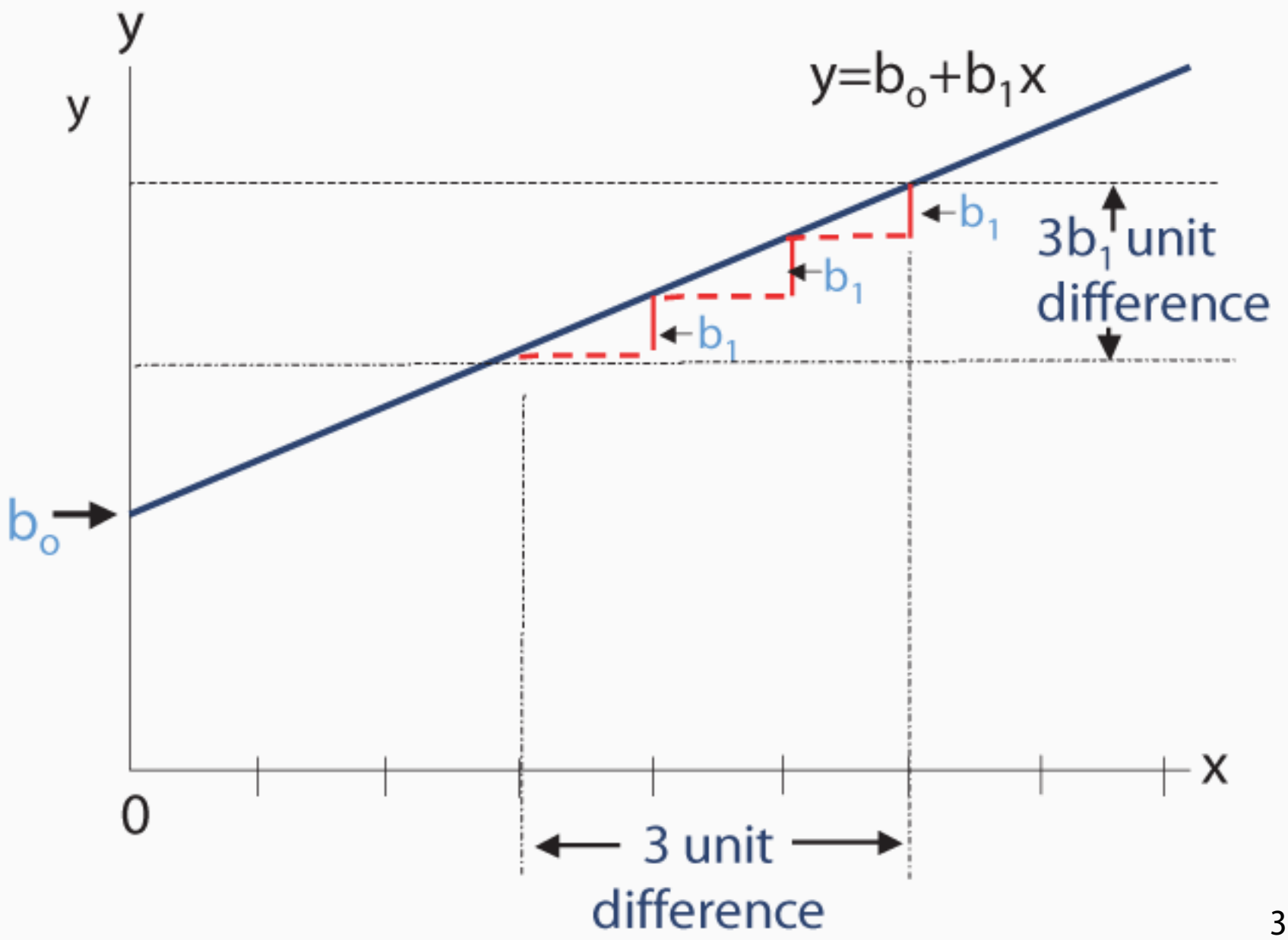
The Slope, b_1

The slope b_1 is the change in y corresponding to a unit increase in x

All information about the difference in the y-value for two differing values of x is contained in the slope!

For example: two values of x three units apart will have a difference in y values of 3* b_1

The Equation of a Line



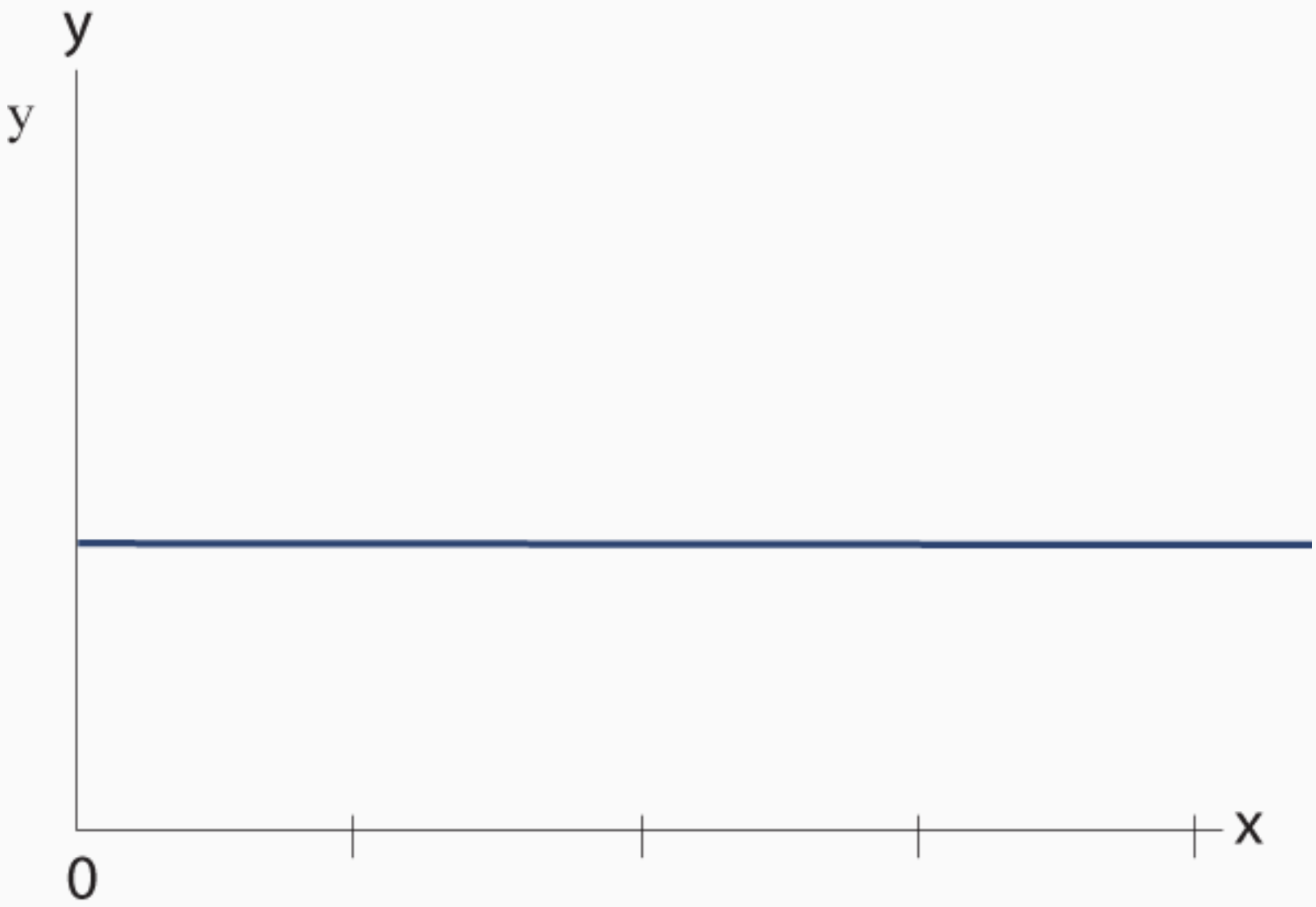
The Slope, b_1

The slope b_1 is the change in y corresponding to a unit increase in x

The slope b_1 gives information about the magnitude and direction of the association between y and x

If slope $b_1 = 0$, indicates that there is no association: (i.e., the values of y are the same regardless of the values of x)

The Slope, $b_1 = 0$



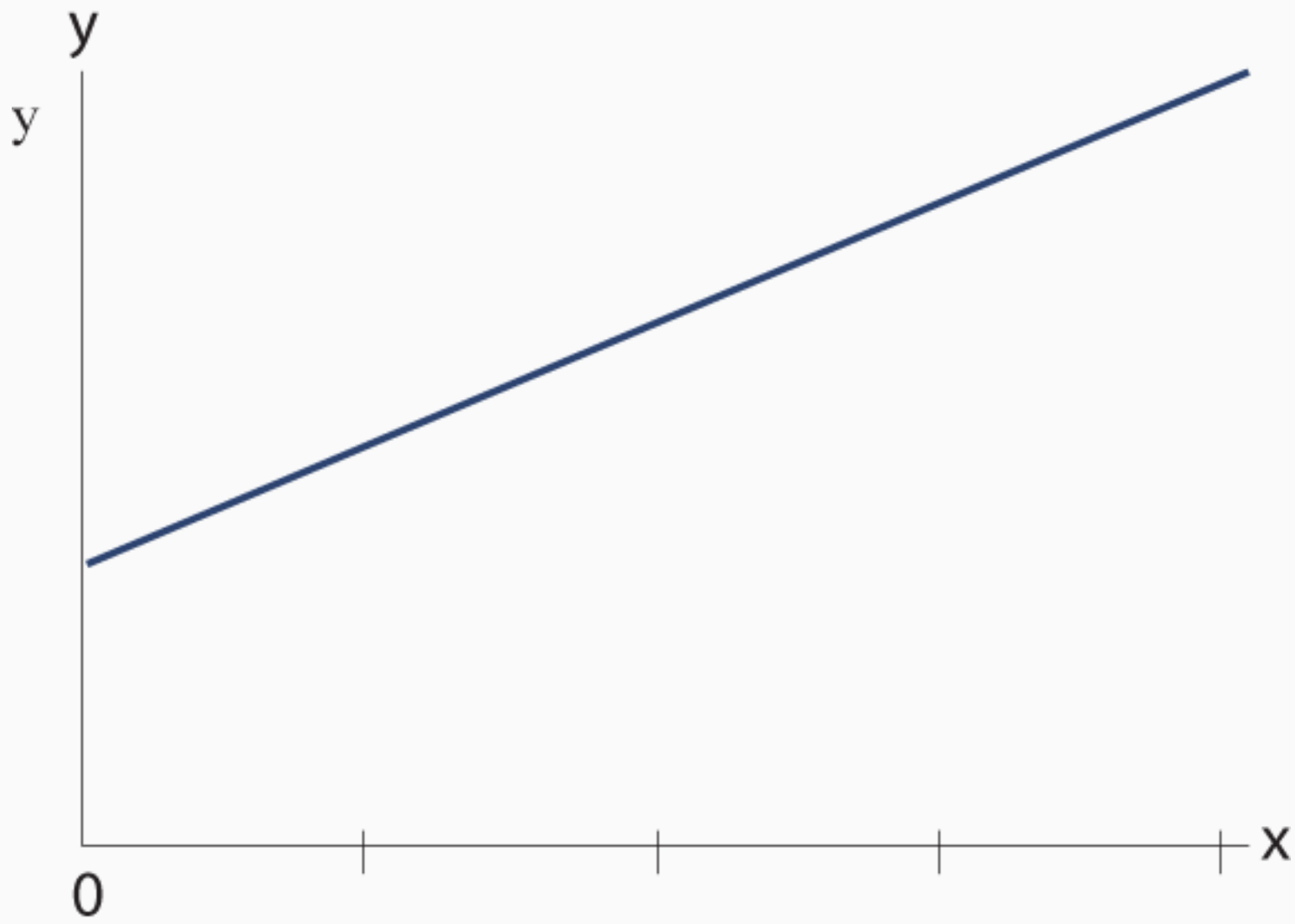
The Slope, b_1

The slope b_1 is the change in y corresponding to a unit increase in x

The slope b_1 gives information about the magnitude and direction of the association between y and x

If slope $b_1 > 0$, indicates that there is a positive association: i.e., value of y increases as value of x increases

The Slope, $b_1 > 0$



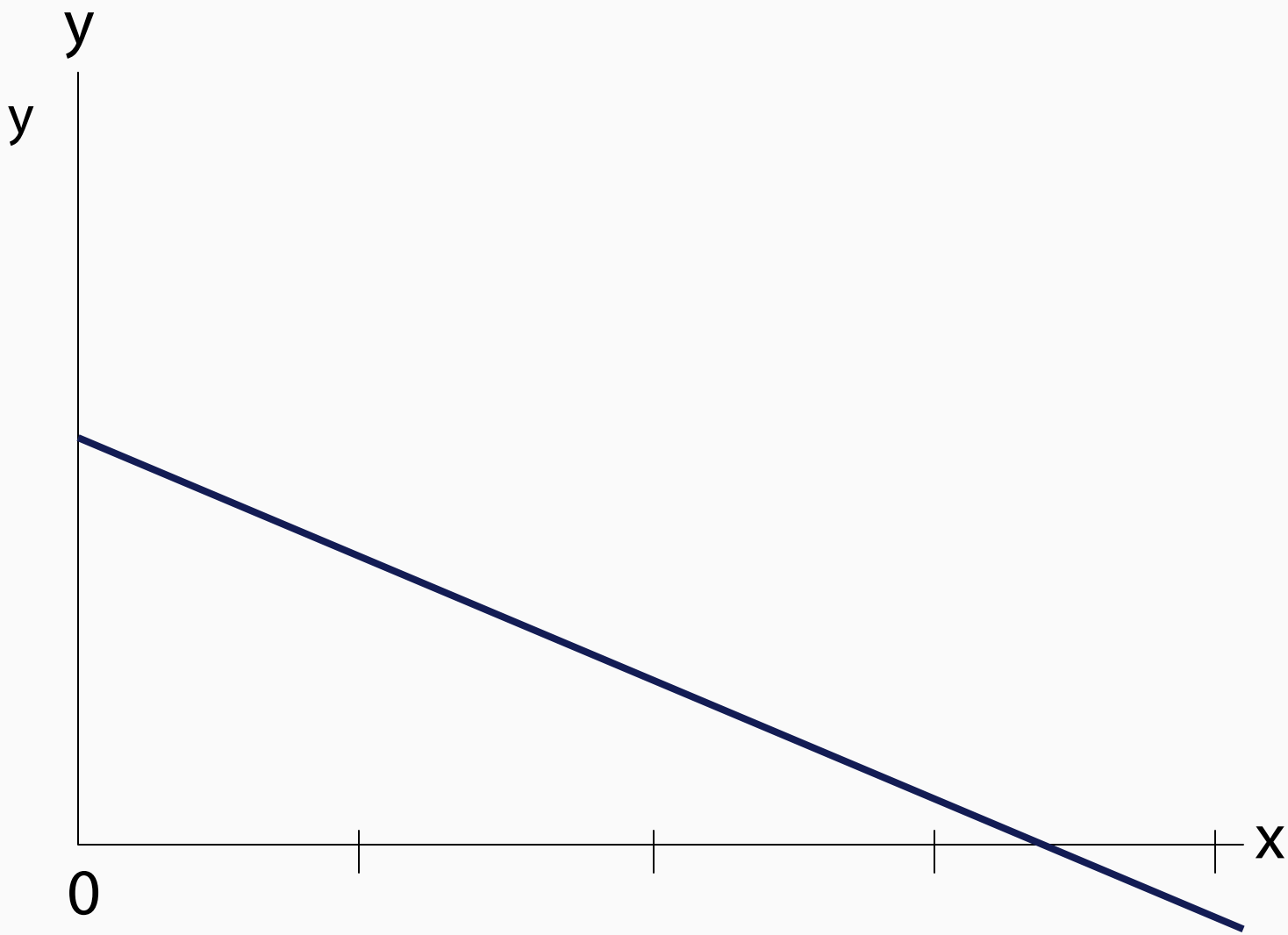
The Slope, b_1

The slope b_1 is the change in y corresponding to a unit increase in x

The slope b_1 gives information about the magnitude and direction of the association between y and x

If slope $b_1 < 0$, indicates that there is a negative association: i.e., value of y decreases as value of x increases

The Slope, $b_1 < 0$



The Equation of a Line

In linear regression situations, points don't fit exactly to a line

We estimate a line that relates the mean of an outcome y to a predictor x

$$E[y] = \hat{b}_0 + \hat{b}_1 x$$

- *E[y] =estimated "expected" (mean) value of y*
- $\hat{b}_0$ = *estimated y-intercept*
- $\hat{b}_1$ = *estimated slope*

Continued

The Equation of a Line

$\hat{b}_0$ and $\hat{b}_1$ are called estimated regression coefficients

These two quantities are estimated using the data

- ★ *Line estimated is line that "fits the data best"*

Many times the equation just written as:

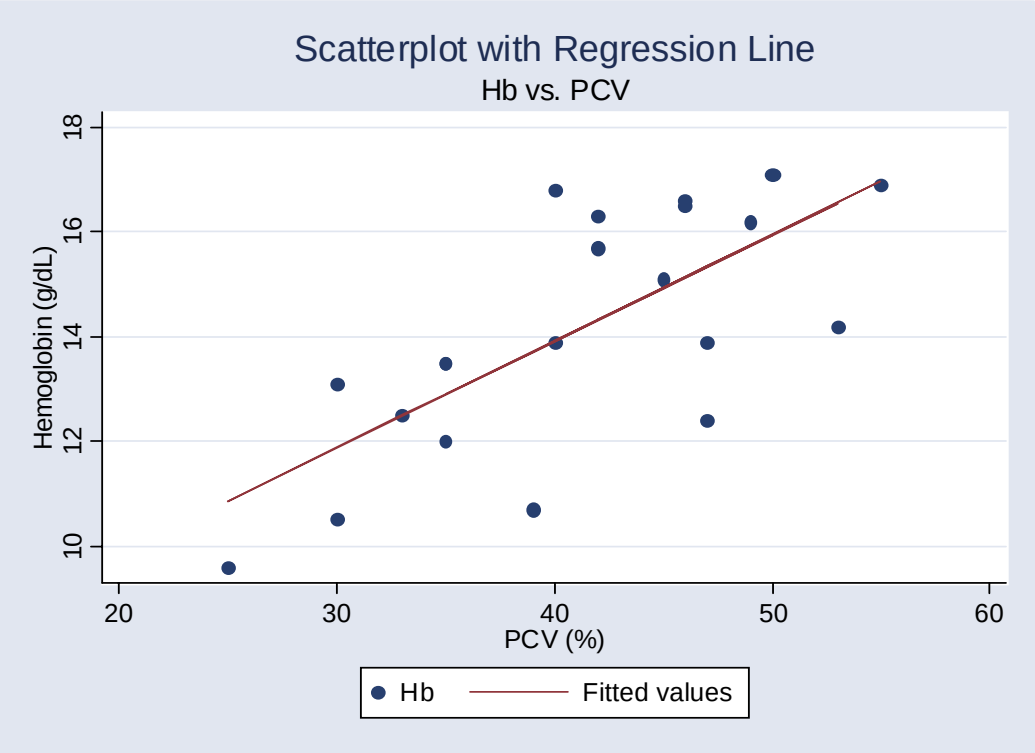
$$y = \hat{b}_0 + \hat{b}_1 x$$

or

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

Example: Hemoglobin and Packed Cell Volume

Scatterplot, with regression line



Continued

Example: Hemoglobin and Packed Cell Volume

Equation of regression line relating estimated mean Hb to PCV (%)

- *E[y] = 5.8 + .2*PCV*
- *In the above equation:*

$$\hat{b}_0 = 5.8$$

$$\hat{b}_1 = 0.2$$

$$x = PCV$$

Continued

Example: Hemoglobin and Packed Cell Volume

Interpretation of estimated y-intercept $\hat{b}_0$

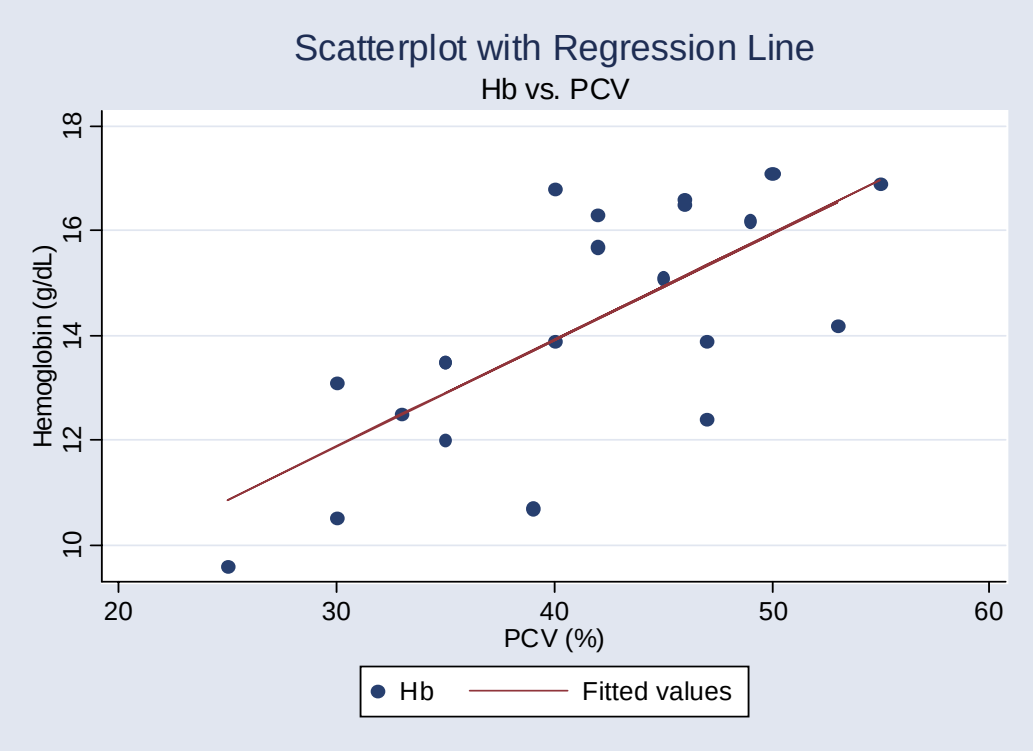
- ★ *Expected value of hemoglobin level for subjects with PCV of 0*
- ★ *In other words—estimated mean hemoglobin level for a group of subjects with PCV of 0*

Does this make sense given our data?

Continued

Example: Hemoglobin and Packed Cell Volume

Notice, range of PCV values in this data set is 25% to 55%—there are no subjects with PCV of 0%



Continued

Example: Hemoglobin and Packed Cell Volume

It is the case in many data situations that the predictor x has range that does not include 0

y-intercept is mathematically necessary to specify equation of line, but in such situations makes little sense scientifically

Important—even though a line goes on "forever," we can only use regression line to describe relationship between y and x for the range of x values in our sample!

Continued

Example: Hemoglobin and Packed Cell Volume

Interpretation of estimated slope $\hat{b}_1$

- ★ *Expected change in hemoglobin levels for one unit (percent) increase in PCV*
- ★ *In other words—estimated mean difference in hemoglobin levels for subjects who differ by one percent in PCV(%)*

So in this example, two groups of subjects who differ by percent in PCV would have on average a difference in hemoglobin levels of 0.2 g/dL

Continued

Example: Hemoglobin and Packed Cell Volume

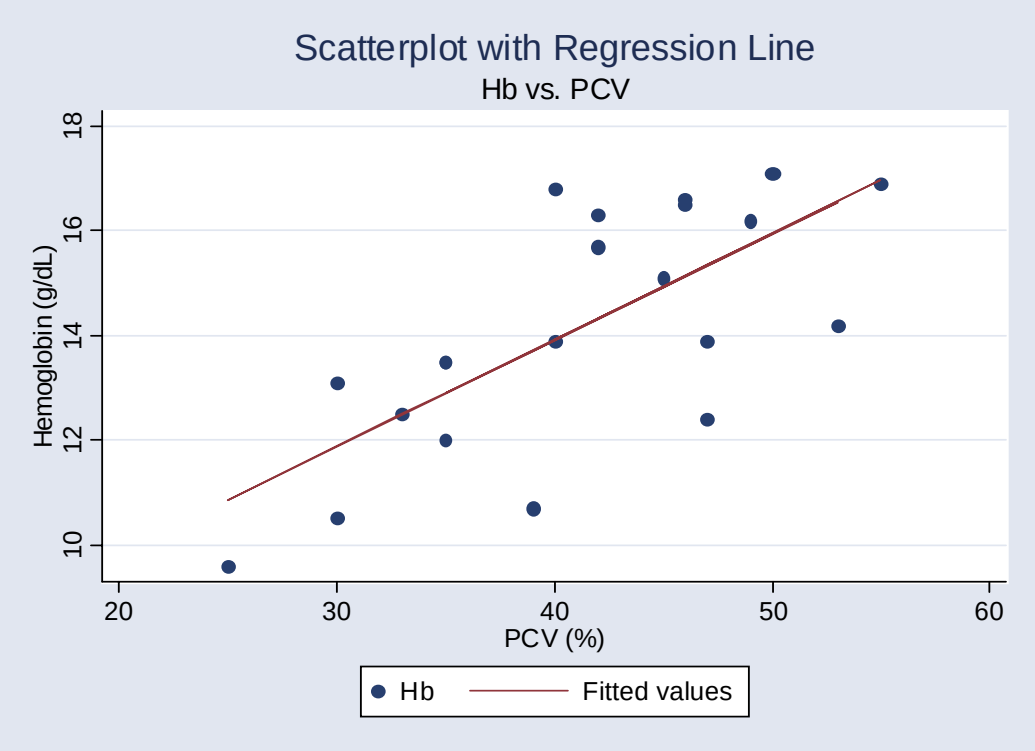
So in this example, two groups of subjects who differ by percent in PCV would have on average a difference in hemoglobin levels of 0.2 g/dL

- ★ *This difference is the same across entire range of PCV values!*

Continued

Example: Hemoglobin and Packed Cell Volume

Notice, range of PCV values in this data set is 25% to 55%



Continued

Example: Hemoglobin and Packed Cell Volume

So for example:

- *Estimated mean difference in Hb level for subjects with PCV of 26% compared to subjects with PCV of 25 %: 0.2 g/dl*
- *Estimated mean difference in Hb level for subjects with PCV of 43% compared to subjects with PCV of 42 %: 0.2 g/dl*
- *Estimated mean difference in Hb level for subjects with PCV of 55% compared to subjects with PCV of 54 %: 0.2 g/dl*

Continued

Example: Hemoglobin and Packed Cell Volume

This one quantity, the estimated slope $\hat{b}_1 = 0.2$, contains all information about mean differences in hemoglobin for groups of subjects who differ in PCV levels

- ★ *How could the slope be used to estimated the mean difference in Hb levels for a groups of individuals with PCV of 50% compared to a group of individuals with PCV of 42%?*
- ★ *In other words, what is the expected difference in mean HB for two groups who PCVs differ by 8%?*

Regression Coefficients

Generally, when interpreting regression coefficients in scientific context, the estimated slope is where we focus our attention

- ★ *Tells us direction/magnitude of association*
- ★ *Gives information about mean difference in y for groups who differ in x values*

As we saw, the intercept alone does not necessarily have a useful scientific interpretation

- ★ *However, it is necessary to specify full regression equation*

Equation of Line

We can also use resulting regression equation to estimate a mean y for a given x

So in the Hb/PCV example, we can use results to estimate the mean Hb level for a subject (group) of subjects given their PCV(%)

- ★ *Recall equation: E[y] = 5.8 + .2*PCV*

Equation of Line

So, for example, how could we estimate Hb for a subject who was not in our sample, but has a measured PCV of 30%?

- ★ *"Plug" PCV=30 into equation*

$$
\begin{aligned}
E[y] &= 5.8 + .2*PCV \\
&= 5.8 + .2*30 \\
&= 5.8 + 6 \\
&= 11.8 \text{ g/dL}
\end{aligned}
$$

Equation of Line

This can be very useful when x is easier to measure than y

- ★ *Do a study with a sample to estimate regression equation*
- ★ *Use equation to predict y for subsequent observations with measure x*

Important—can only use equation to predict y for observations whose x values are in same range as sample data!

Summary : Linear Regression

Linear regression is a statistical method used to estimate the mean of an outcome y as a function of a predictor x via an equation of a line

The slope from an estimated regression equation estimates magnitude and direction of relationship between y and x and contains all information about estimating mean differences in y between groups who differ in value of x

The resulting regression equation can also be employed to estimate the mean of y for an observation(s) with a known x value

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section B

Practice Problems

1. Recall the results from the regression of Hemoglobin levels (Hb) on packed cell volume percentage (PCV) for 21 subjects, with PCV values ranging from 25% to 55%. The resulting regression equation is as follows:

   E[y] = 5.8 + .2*PCV , where E[y] is estimated mean hemoglobin level

   Using these regression results, can you answer the following questions?

Continued

a. Estimate the mean difference in hemoglobin levels for a group of subjects with PCV of 42% compared to a group with 41%.

b. Estimate the mean difference in hemoglobin levels for a group of subjects with PCV of 12% compared to a group with 11%.

c. Estimate the mean difference in hemoglobin levels for a group of subjects with PCV of 42% compared to a group with 37%.

Continued

d. Estimate the mean hemoglobin level for a group of subjects with PCV of 30%.

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section B

Practice Problem Solutions

1. Recall the results from the regression of Hemoglobin levels (Hb) on packed cell volume percentage (PCV) for 21 subjects, with PCV values ranging from 25% to 55%. The resulting regression equation is as follows:

   E[y] = 5.8 + .2*PCV , where E[y] is estimated mean hemoglobin level (g/dL)

   Using these regression results, can you answer the following questions?

Continued

a. Estimate the mean difference in hemoglobin levels for a group of subjects with PCV of 42% compared to a group with 41%.

   As the two groups differ by one unit of PCV, the estimated mean difference in HB for the group with higher PCV (42) to the group with lower PCV (41) would just be the slope estimate from the regression of Hb on PCV: .2 g/dL

Continued

b. Estimate the mean difference in hemoglobin levels for a group of subjects with PCV of 12% compared to a group with 11%.

A trick question—you were probably tempted to answer the same thing here, but if you remember from an earlier slide, the range of observed PCV values in this data set was 25% to 55%, and hence the regression results are not necessarily applicable to Hb/PCV comparisons outside of this range.

Continued

c. Estimate the mean difference in hemoglobin levels for a group of subjects with PCV of 42% compared to a group with 37%.

As the two groups differ by five units of PCV, the estimated mean difference in HB for the group with higher PCV (42) to the group with lower PCV (37) would just be the slope estimate from the regression of Hb on PCV: 5*.2 g/dL= 1 g/dL

Continued

d. Estimate the mean hemoglobin level for a group of subjects with PCV of 30%.

Note—this is legal, since 30% is within the range of observed PCV values. We can just plug x = 30 into our equation relating Hb to PCV:

$$E[y] = 5.8 + .2*PCV$$

$$E[y] = 5.8 + .2*30 = 5.8+6 = 11.8.$$

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section C

The Simple Linear Regression Model: Estimating the Regression Equation—Accounting for Uncertainly in the Estimates

Example: Hemoglobin and Packed Cell Volume

Recall, from last section, the estimated regression line relating estimated mean Hb to PCV (%) was . . .

- *E[y] = 5.8 + .2*PCV*
- *Where . . .*

$$\hat{b}_0 = 5.8$$
$$\hat{b}_1 = 0.2$$
$$x = \text{PCV}$$

Continued

Example: Hemoglobin and Packed Cell Volume

Where did this equation come from?

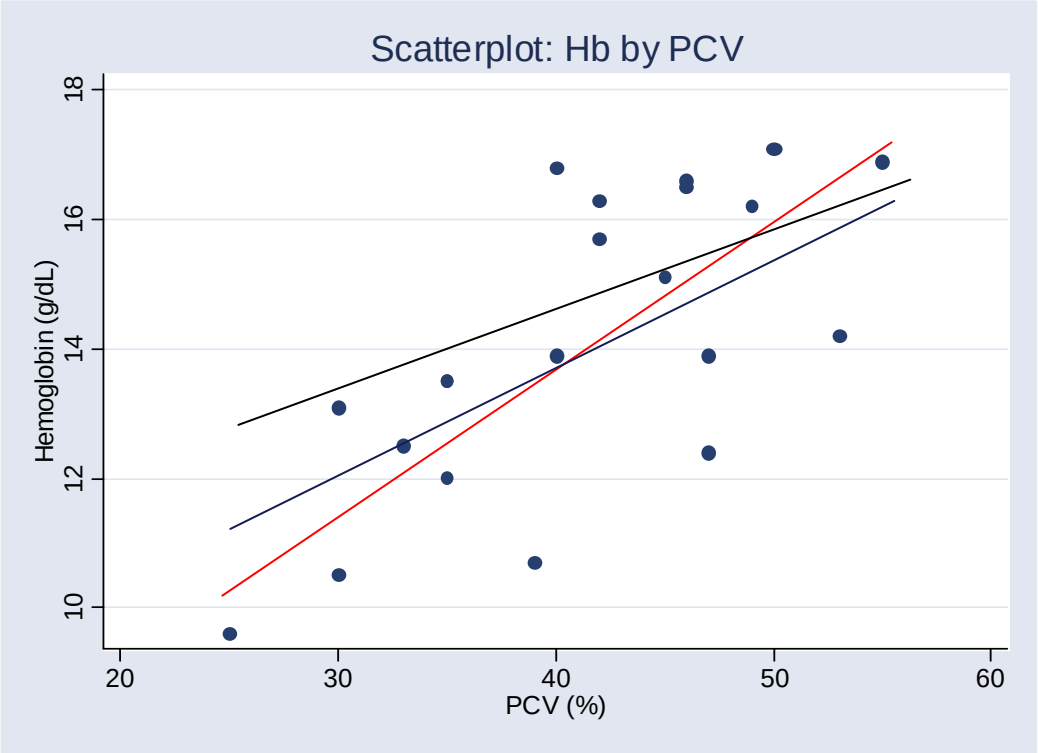
The computer actually did the work, and I'll show you how to use the computer to estimate the line in a few slides

There must be some algorithm for systematically estimating a regression line given a set of data: otherwise, every analyst would end up with a different line for the same data set!!

Continued

Example: Hemoglobin and Packed Cell Volume

Without a systematic way of doing this, everyone would choose a different line!



The Equation of a Line

In linear regression situations, points don't fit exactly to a line

We estimate a line that relates the mean of an outcome y to a predictor x

$$E[y] = \hat{b}_0 + \hat{b}_1 x$$

- *E[y] =estimated "expected" (mean) value of y*
- $\hat{b}_0$= *estimated y-intercept*
- $\hat{b}_1$= *estimated slope*

Simple Linear Regression Model

Points don't fall exactly on line—so each observed value is a combination of its estimated mean and a discrepancy

- *Observed value = mean + discrepancy*
- *Sometimes this is written as:*

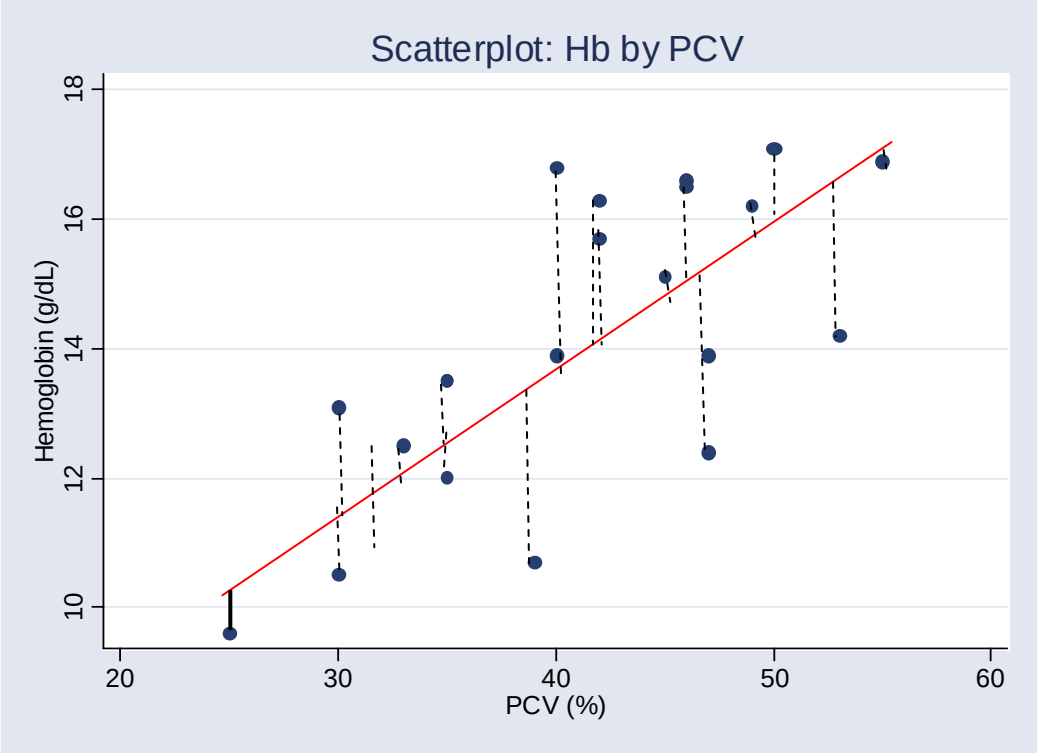
$$y = E[y] + \varepsilon \text{ , or}$$

$$y = \hat{b}_0 + \hat{b}_1 x + \varepsilon$$

Where E[y] = estimated mean, y = actual observed value, ε = discrepancy

Example: Hemoglobin and Packed Cell Volume

Graphic showing "discrepancies" between each observation and regression estimate—we have 21 observations and hence 21 discrepancies



Continued

Example: Hemoglobin and Packed Cell Volume

Regression line is the line that gets "closest" to all of the points

In other words, an estimated regression line is a line that minimizes total discrepancies between the observations and the regression line

Algorithm does this by choosing slope and intercept values that minimize sum of squared discrepancies, i.e. . . .

$$\sum_{I=1}^{21} \varepsilon_i^{\,2}$$

Continued

Example: Hemoglobin and Packed Cell Volume

In other words, the estimated regression line has values of the y-intercept $\hat{b}_0$ and slope $\hat{b}_1$ that minimize:

$$\sum_{I=1}^{21}(y_i - (\hat{b}_0 + \hat{b}_1 x))^2$$

This method is called "least squares estimate" of regression equation

Continued

Example: Hemoglobin and Packed Cell Volume

Obviously, this shouldn't be done by hand

Stata will do the work for us!

Data needs to first be entered in Stata—here is a snippet of the hemoglobin/PCV data

```
. list  Hb PCV in 1/

     +------------+
     |   Hb   PCV |
     |------------|
  1. |   12    35 |
  2. | 10.7    39 |
  3. | 12.4    47 |
  4. | 14.2    53 |
  5. | 13.1    30 |
     |------------|
  6. | 10.5    30 |
  7. |  9.6    25 |
  8. | 12.5    33 |
  9. | 13.5    35 |
 10. | 13.9    40 |
     +------------+
```

Continued

"regress" command in Stata will perform least squares estimation and give us the estimated slope and intercept values

Syntax

- ★ *regress y x*

So with hemoglobin/PCV data, type:

- ★ *regress Hb PCV*

Continued

Stata results: a lot of information!

```
. regress  Hb PCV

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    19) =   19.80
       Model |  53.7803079     1  53.7803079            Prob > F      =  0.0003
    Residual |  51.5711174    19  2.71426934            R-squared     =  0.5104
-------------+------------------------------           Adj R-squared =  0.4846
       Total |  105.351425    20  5.26757126            Root MSE      =  1.6474

------------------------------------------------------------------------------
          Hb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         PCV |   .2033502   .0456835     4.45   0.000     .1077335    .298966
       _cons |    5.77645   1.913624     3.02   0.007     1.771188    9.78171
------------------------------------------------------------------------------
```

Continued

Where is the intercept and slope amidst all of this output?

- ★ *Both in 2nd table, in column labeled "Coef."*

```
. regress  Hb PCV

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    19) =   19.80
       Model |  53.7803079     1  53.7803079            Prob > F      =  0.000
    Residual |  51.5711174    19  2.71426934            R-squared     =  0.510
-------------+------------------------------           Adj R-squared =  0.484
       Total |  105.351425    20  5.26757126            Root MSE      =  1.647

------------------------------------------------------------------------------
          Hb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         PCV |   .2033502   .0456835     4.45   0.000     .1077335    .298966
       _cons |    5.77645   1.913624     3.02   0.007     1.771188    9.78171
------------------------------------------------------------------------------
```

Continued

Intercept always in row labeled "_cons" for "constant"

```
. regress  Hb PCV

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    19) =   19.80
       Model |  53.7803079     1  53.7803079            Prob > F      =  0.000
    Residual |  51.5711174    19  2.71426934            R-squared     =  0.510
-------------+------------------------------           Adj R-squared =  0.484
       Total |  105.351425    20  5.26757126            Root MSE      =  1.647

------------------------------------------------------------------------------
          Hb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         PCV |   .2033502   .0456835     4.45   0.000     .1077335    .298966
       _cons |    5.77645   1.913624     3.02   0.007     1.771188    9.78171
------------------------------------------------------------------------------
```

Continued

Slope always in row labeled with name of predictor (*x*) variable

```
. regress  Hb PCV

      Source |       SS       df       MS              Number of obs =       2
-------------+------------------------------           F(  1,    19) =    19.8
       Model |  53.7803079     1  53.7803079           Prob > F      =   0.000
    Residual |  51.5711174    19  2.71426934           R-squared     =   0.510
-------------+------------------------------           Adj R-squared =   0.484
       Total |  105.351425    20  5.26757126           Root MSE      =   1.647

------------------------------------------------------------------------------
          Hb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         PCV |   .2033502   .0456835     4.45   0.000     .1077335    .298966
       _cons |    5.77645   1.913624     3.02   0.007     1.771188    9.78171
------------------------------------------------------------------------------
```

Example: Hemoglobin and Packed Cell Volume

So the equation given by Stata output is equation we have been using all along:

- *E[y] = 5.8 + .2*PCV*

$$\hat{b}_0 = 5.8$$

$$\hat{b}_1 = 0.2$$

Recall, we are estimating this information based on a sample of only 21 subjects—need to account for uncertainty in our sample estimates

```
. regress  Hb PCV

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =   19.81
       Model |  53.7803079     1  53.7803079            Prob > F      =  0.0003
    Residual |  51.5711174    19  2.71426934           R-squared     =  0.5105
-------------+------------------------------           Adj R-squared =  0.4847
       Total |  105.351425    20  5.26757126           Root MSE      =  1.6475

------------------------------------------------------------------------------
          Hb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         PCV |   .2033502   .0456835     4.45   0.000     .1077335    .2989668
       _cons |    5.77645   1.913624     3.02   0.007     1.771188    9.781712
------------------------------------------------------------------------------
```

Continued

Stata gives standard errors for the intercept and slope estimates

```
. regress  Hb PCV

      Source |       SS       df       MS              Number of obs =       2
-------------+------------------------------           F(  1,    19) =    19.8
       Model |  53.7803079     1  53.7803079           Prob > F      =  0.000
    Residual |  51.5711174    19  2.71426934           R-squared     =  0.510
-------------+------------------------------           Adj R-squared =  0.484
       Total |  105.351425    20  5.26757126           Root MSE      =  1.647

------------------------------------------------------------------------------
          Hb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         PCV |   .2033502   .0456835     4.45   0.000     .1077335    .298966
       _cons |    5.77645   1.913624     3.02   0.007     1.771188    9.78171
------------------------------------------------------------------------------
```

Accounting for Uncertainty from Sampling Variability

We estimated the regression line based on the information on 21 subjects:

- ★ *There is a "true" regression line that related hemoglobin to PCV for the population of all subjects, with true intercept* b_o *and true* b_1

Had we taken a different sample of 21 subjects we would most likely get different estimates of b_o and true b_1

Continued

Accounting for Uncertainty from Sampling Variability

Standard error of estimated y-intercept estimates how close on average this estimate based on 21 observations gets to "true" intercept

In our example . . .

- $se(\hat{b}_0) = 1.9$

Continued

Accounting for Uncertainty from Sampling Variability

Standard error of estimated slope estimates how close on average this estimate based on 21 observations gets to "true" slope

In our example . . .

- $se(\hat{b}_1) = 0.045$

Continued

Accounting for Uncertainty from Sampling Variability

We can use the estimates coupled with their standard errors to create confidence intervals and perform a hypothesis test

We will focus our efforts on doing this for the slope, as the intercept is not usually scientifically interesting or useful

Sampling Variability of Slope

Luckily, the story remains the same

Statistical theory tells us how an estimate of a slope of regression line based on a sample of fixed sizes behaves relative to the truth

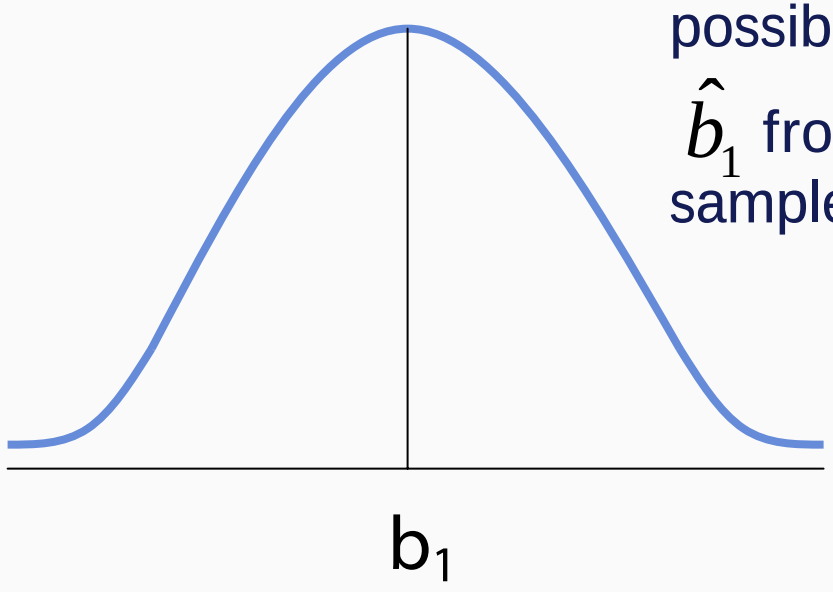
It's the same story as we saw with differences in sample means:

- ★ *t- distribution if n "small" (≤ 60)*
- ★ *Normal distribution if n large (> 60)*

Continued

Sampling Variability of Slope

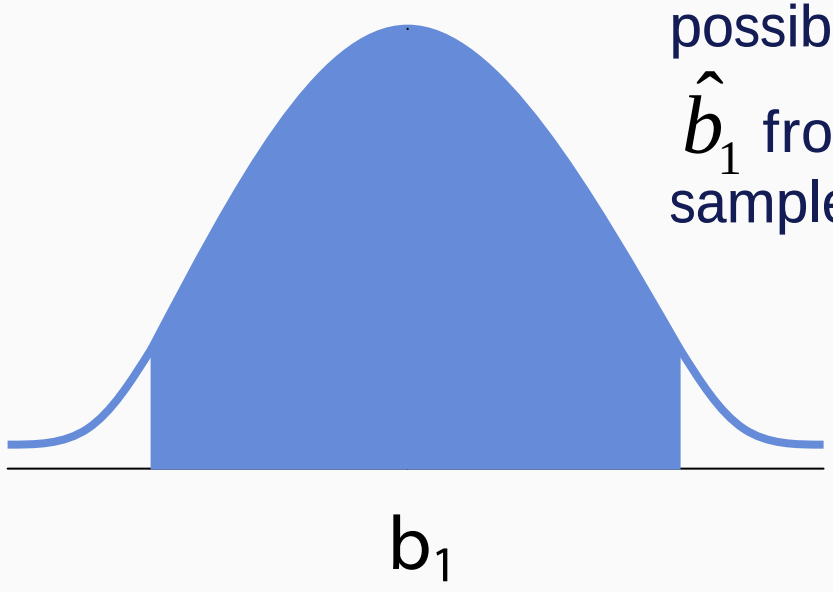
Estimates of slope from different samples of same size vary around true slope via a known distribution



Continued

Sampling Variability of Slope

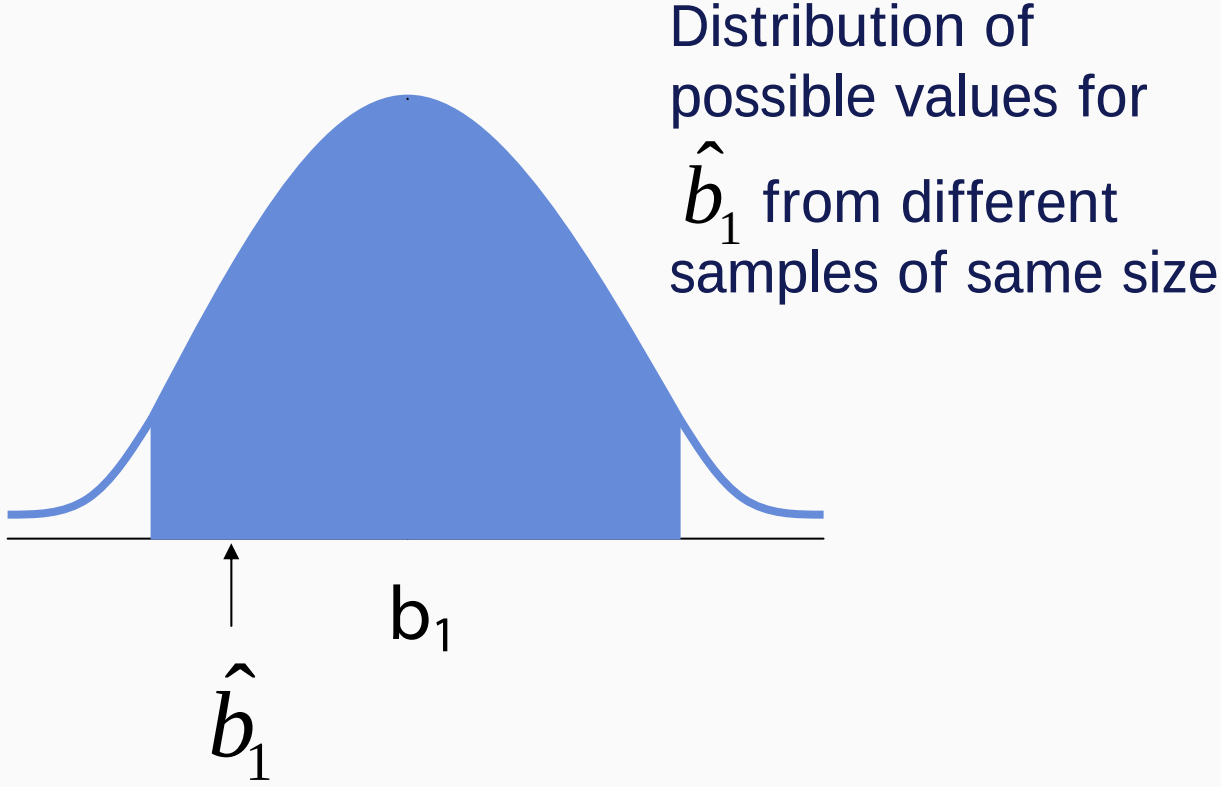
Most possible values for $\hat{b}_1$ (95%) will fall within about two standard errors of truth



Continued

Sampling Variability of Slope

We will only observe one estimate $\hat{b}_1$ but most of the time this will be within about two standard errors of truth



Continued

Hence, we can create a 95% confidence by going about two standard errors in either direction from our estimate

Technically speaking, a 95% CI for b_1 is given by:

$$\hat{b}_1 \pm t_{n-2} se(\hat{b}_1)$$

Where t_{n-2} is the number of standard errors we have in either direction from the center of a t distribution to cut off 95% under the curve

Continued

Sampling Variability of Slope

If n large (> 60) we can use:

$$\hat{b}_1 \pm 2se(\hat{b}_1)$$

Example: Hemoglobin and Packed Cell Volume

In the Hb/PCV example the estimated relationship between mean Hb level and PCV was given by:

- *E[y] = 5.8 + .2*PCV, with* $\hat{b}_1 = 0.2$
- *The standard error of this estimate was:*

$$se(\hat{b}_1) = 0.045$$

Continued

Example: Hemoglobin and Packed Cell Volume

Putting these results together gives a 95% CI for b_1 of :

$$\hat{b}_1 \pm t_{19} se(\hat{b}_1)$$

$$= 0.2 \pm 2.1*(.0046)$$

$$= 0.2 \pm .097$$

$$(0.10, 0.30)$$

Luckily, Stata will do the work for us, so we don't have to run to a t-table!

```
. regress  Hb PCV

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =   19.81
       Model |  53.7803079     1  53.7803079            Prob > F      =  0.0003
    Residual |  51.5711174    19  2.71426934            R-squared     =  0.5105
-------------+------------------------------           Adj R-squared =  0.4847
       Total |  105.351425    20  5.26757126            Root MSE      =  1.6475

------------------------------------------------------------------------------
          Hb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         PCV |   .2033502   .0456835     4.45   0.000     .1077335   .2989668
       _cons |    5.77645   1.913624     3.02   0.007     1.771188   9.781712
------------------------------------------------------------------------------
```

Interpreting Result of 95% CI

So, the estimated slope is 0.2 with 95% CI 0.10 to 0.30

How to interpret results?

- ★ *Based on a sample of 21 subjects, we estimated that PCV(%) is positively associated with hemoglobin levels*
- ★ *We estimated that a one-percent increase in PCV is associated with a 0.2 g/dL increase in hemoglobin on average*
- ★ *Accounting for sampling variability, this increase could be as small as 0.10 g/dL, or as large as 0.3 g/dL in the population of all subjects*

Continued

Interpreting Result of 95% CI

In other words:

- ★ *We estimated that the average difference in hemoglobin levels for two groups of subjects who differ by one-percent PCV to be 0.2 g/dL on average (higher PCV group relative to lower)*
- ★ *Accounting for sampling variability, mean difference could be as small as 0.10 g/dL, or as large as 0.3 g/dL in the population of all subjects*

Continued

Interpreting Result of 95% CI

Remember we said that we could average hemoglobin levels between groups of subjects who differ by more than one unit of PCV (%) just by taking multiples of the slope?

Example—mean difference in Hb levels for a group of individuals with PCV of 50% compared to a group of individuals with PCV of 42%?

- ★ *This is an eight unit difference in PCV*
- ★ *Estimate mean difference in Hb using regression results given by 8* $\hat{\beta}_1$*

Continued

So if we were to compare mean hemoglobin levels for group with 50% PCV to 42%, the estimated mean difference would be 8*0.2 = 1.6 g/dL

How can we get a 95% CI for the above mean difference?

- *Same logic applies*

$$8\hat{b}_1 \pm 2se(8\hat{b}_1)$$

- *Luckily,*

$$se(8\hat{b}_1) = 8se(\hat{b}_1)$$

Continued

Interpreting Result of 95% CI

So in other words, to get 95% CI for $8b_1$, we just multiply endpoints of 95% CI for b_1 by 8

The 95% CI for b_1 was (0.10 g/dL, 0.30 g/dL)

The 95% CI for $8b_1$ is given by (8*0.10, 8*0.30) = (0.8 g/dL to 2.4 g/dL)

Continued

Interpreting Result of 95% CI

Interpretation:

- ★ *We estimated that the average difference in hemoglobin levels for a group of subjects with PCV of 50% relative to a group with PCV of 42% to be 1.6 g/dL*
- ★ *Accounting for sampling variability, this mean difference could be a small as 0.8 g/dL, or as large as 2.4 g/dL*

Hypothesis Testing with the Regression Slope

Recall, the estimated slope relating mean Hb to PCV was 0.20, with 95% CI 0.10 to 0.30

The 95% CI does not include 0, indicating that the relationship is positive at the population level

How could we perform a formal hypothesis test for this relationship and get a p-value?

Continued

Hypothesis Testing with the Regression Slope

If we are interested in whether an association exists at the population level, we are interested in formally testing:

$$H_o: b_1 = 0$$

$$H_a: b_1 \neq 0$$

Recall, a slope of 0 would indicate that the mean of **y** (Hb) does not depend on **x** (PCV): that is, mean values of **y** is same for all levels of **x**

Continued

Hypothesis Testing with the Regression Slope

Same old story

- *Start by pretending* H_o *is true, i.e.,* $\hat{b}_1 = 0$
- *Figure out how far our estimate* $\hat{b}_1$ *is from 0 in terms of standard errors:*

$$t = \frac{\hat{b}_1 - 0}{se(\hat{b}_1)}$$

- *Figure out if result is "unusual" when* $b_1 = 0$ *by comparing this distance to the sampling distribution and getting a p-value*

Example: Hemoglobin and Packed Cell Volume

In the Hb/PCV example, the estimated slope was $\hat{b}_1 = 0.2$, with estimated standard error

$$se(\hat{b}_1) = 0.046$$

So our estimate is $t = \frac{0.2 - 0}{0.046} = \frac{0.2}{0.046} \approx 4.4$ standard errors away from 0

We would compare this to a t-distribution with 19 degrees of freedom to get a p-value

Continued

Example: Hemoglobin and Packed Cell Volume

Again, Stata will do the work for us, so we don't have to run to a t-table!

```
. regress  Hb PCV

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    19) =   19.80
       Model |  53.7803079     1  53.7803079            Prob > F      =  0.0003
    Residual |  51.5711174    19  2.71426934            R-squared     =  0.5104
-------------+------------------------------           Adj R-squared =  0.4846
       Total |  105.351425    20  5.26757126            Root MSE      =  1.6474

------------------------------------------------------------------------------
          Hb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         PCV |   .2033502   .0456835     4.45   0.000     .1077335    .298966
       _cons |    5.77645   1.913624     3.02   0.007     1.771188    9.78171
------------------------------------------------------------------------------
```

Continued

Example: Hemoglobin and Packed Cell Volume

So the p-value for testing:

- ★ *Is very small ($p < .001$):*

$$H_o: b_1 = 0$$

$$H_a: b_1 \neq 0$$

- ★ *At the α=.05 level, we would reject* H_o *in favor of* H_a *and conclude that there is an association between hemoglobin and packed cell volume*

We already knew $p < .05$ because 95% CI for b_1 did not include 0

So, Putting it All Together…

Based on a sample of 21 subjects, we estimated a statistically significant ($p < .001$) positive association between hemoglobin (Hb) levels and percent of packed cells (PCV)

We estimated that the average difference in hemoglobin levels for two groups of subjects who differ by one-percent PCV to be 0.2 g/dL on average (higher PCV group relative to lower)

Accounting for sampling variability, mean difference could be a small as 0.10 g/dL, or as large as 0.2 g/dL in the population of all subjects

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section C

Practice Problems

1. Consider the general form of the simple linear regression equation:

$$E[y] = \hat{b}_0 + \hat{b}_1 x_1$$

Continued

a. What is the most general interpretation of the slope $\hat{b}_1$ estimate in this equation?

b. What substantive question is the following hypothesis test asking?

$$H_o: b_1 = 0$$
$$\text{vs. } H_a: b_1 \neq 0$$

Continued

2. Consider the following results from a SLR regressing weight (lbs) on height (in) for a sample of 12 nutritionally deficient children. The results, below, are missing some key information.

```
. regress wt ht

  Source |       SS       df       MS                  Number of obs =      12
---------+------------------------------                F(  1,    10) =   19.67
   Model |  588.922523     1  588.922523                 Prob > F      =  0.0013
Residual |  299.327477    10  29.9327477                 R-squared     =  0.6630
---------+------------------------------                Adj R-squared =  0.6293
   Total |      888.25    11       80.75                 Root MSE      =  5.4711

------------------------------------------------------------------------------
      wt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      ht |    1.07223    .241731
   _cons |   6.189849   12.84875
```

Continued

a. Write out the regression equation described by these results.

b. What is the interpretation of the coefficient for height?

c. Construct a 95% CI for the coefficient b_1.

Continued

d. What can be said about the p-value for testing

$$H_o: b_1 = 0$$

$$H_a: b_1 \neq 0?$$

e. What would the predicted weight be for a child who is 56 inches tall (assume it is in the range of observed heights)?

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section C

Practice Problem Solutions

1. Consider the general form of the simple linear regression equation:

$$E[y] = \hat{b}_0 + \hat{b}_1 x_1$$

Continued

a. What is the most general interpretation of the slope $\hat{b}_1$ estimate in this equation?

$\hat{b}_1$ is the estimated expected change in **y** for a one unit increase in $\mathbf{x}_1$.

Continued

b. What substantive question is the following hypothesis test asking?

$$H_o: b_1 = 0$$
$$\text{vs. } H_a: b_1 \neq 0$$

Does knowing x_1 add to our knowledge about the mean value of y?

Continued

2. Consider the following results from a SLR regressing weight (lbs) on height (in) for a sample of 12 nutritionally deficient children. The results, below, are missing some key information.

```
. regress wt ht

  Source |       SS       df       MS                  Number of obs =      12
---------+------------------------------                F(  1,    10) =   19.67
   Model |  588.922523     1  588.922523                 Prob > F      =  0.0013
Residual |  299.327477    10  29.9327477                 R-squared     =  0.6630
---------+------------------------------                Adj R-squared =  0.6293
   Total |      888.25    11       80.75                 Root MSE      =  5.4711

------------------------------------------------------------------------------
      wt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      ht |    1.07223    .241731
   _cons |   6.189849   12.84875
```

Continued

a. Write out the regression equation described by these results.

```
. regress wt ht

  Source |       SS       df       MS                  Number of obs =      12
---------+------------------------------                F(  1,    10) =   19.67
   Model |  588.922523     1  588.922523                 Prob > F      =  0.0013
Residual |  299.327477    10  29.9327477                 R-squared     =  0.6630
---------+------------------------------                Adj R-squared =  0.6293
   Total |      888.25    11       80.75                 Root MSE      =  5.4711

------------------------------------------------------------------------------
      wt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      ht |    1.07223    .241731
   _cons |   6.189849   12.84875
```

This is the estimated intercept, $\hat{b}_0$.

Continued

a. Write out the regression equation described by these results.

```
. regress wt ht

  Source |       SS       df       MS                  Number of obs =      12
---------+------------------------------                F(  1,    10) =   19.67
   Model |  588.922523     1  588.922523                Prob > F      =  0.0013
Residual |  299.327477    10  29.9327477                R-squared     =  0.6630
---------+------------------------------                Adj R-squared =  0.6293
   Total |      888.25    11       80.75                Root MSE      =  5.4711

------------------------------------------------------------------------------
      wt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      ht |     1.07223    .241731
   _cons |    6.189849   12.84875
```

This is the estimated slope, $\hat{b}_1$.

Continued

a. Write out the regression equation described by these results.

$$y = \hat{b}_0 + \hat{b}_1 x_1$$

$$y = 6.2 + 1.1x_1$$

Where y = mean weight, $\mathbf{x}_1$ = height

Continued

b. What is the interpretation of the coefficient estimate for height?

If two children differ by one inch in height, the taller child will weigh 1.1 pounds more on average, than the shorter child.

Continued

c. Construct a 95% CI for the coefficient b_1.

Recall, the formula:

$$\hat{b}_1 \pm t_{n-2} SE(\hat{b}_1)$$

Continued

c. Construct a 95% CI for the coefficient b_1.

Here we want the appropriate t-value for computing a 95% CI from a t-distribution with 10 degrees of freedom. From the table on page 521 of Altman, the value we need is 2.28.

Continued

c. Construct a 95% CI for the coefficient b_1.

Hence, a 95% CI for the coefficient b_1:

$$\hat{b}_1 \pm 2.28 SE(\hat{b}_1)$$

$$1.07 \pm 2.28*(.24)$$

$$1.07 \pm 0.55$$

So the 95% CI for is (0.52, 1.62).

Continued

d. What can be said about the p-value for testing

$$H_o: b_1 = 0$$
$$H_a: b_1 \neq 0$$

As the 95% CI for b_1 does not include 0, the p-value for testing the above hypothesis will be < .05.

Continued

e. What would the predicted weight be for a child who is 56 inches tall?

Just plug 56 cm (for x_1) into the regression equation!

$$y = 6.2+1.1(56)$$

$$= 6.2+ 61.6$$

$$= 67.8 \text{ lbs.}$$

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section D

Measuring the Strength of a Linear Association

How Close Do the Points Fall on the Line?

Recall, in the Hemoglobin/PCV example we found a statistically significant positive association between hemoglobin and PCV

We estimated the magnitude of this association via the slope of the regression line, which was chosen by taking the line that got "closest" to the points (least squares estimation)

Can we quantify how close the points got to the line? This would give us some sense of how much PCV "explains" about hemoglobin in our study

Example: Hemoglobin and Packed Cell Volume

Based on a sample of 21 subjects, we estimated a statistically significant ($p < .001$) positive association between hemoglobin (Hb) levels and percent of packed cells (PCV)

We estimated that the average difference in hemoglobin levels for two groups of subjects who differ by one-percent PCV to be 0.2 g/dL on average (higher PCV group relative to lower)

Accounting for sampling variability, mean difference could be as small as 0.10 g/dL, or as large as 0.2 g/dL in the population of all subjects

How Close Do the Points Fall on the Line?

It is not measured by the slope

Slope measures magnitude (size of relationship)

Slope does not measure “closeness” of points to the line

Continued

How Close Do the Points Fall on the Line?

There is another quantity estimated in linear regression, called the "coefficient of determination," usually referred to as R^2 (R-squared)

R^2 is an estimate of the percent of variability in y that has been "explained" by x

- ★ *In our example,* R^2 *estimates the percent of variability in hemoglobin explained by PCV*

Example: Hemoglobin and Packed Cell Volume

In the regression relating hemoglobin to PCV, the estimated R^2 was .51 or 51%

- ★ *Interpretation: in our sample, subject's PCV values explained 51% of the variability in subject's hemoglobin values*

This, like the slope, is an estimate and subject to sampling error: however, it is not standard practice to compute a confidence interval for this quantity

Continued

Example: Hemoglobin and Packed Cell Volume

What does this mean?

- ★ *Interpretation—in our sample, subject's PCV values explained 51% of the variability in subject's hemoglobin values*

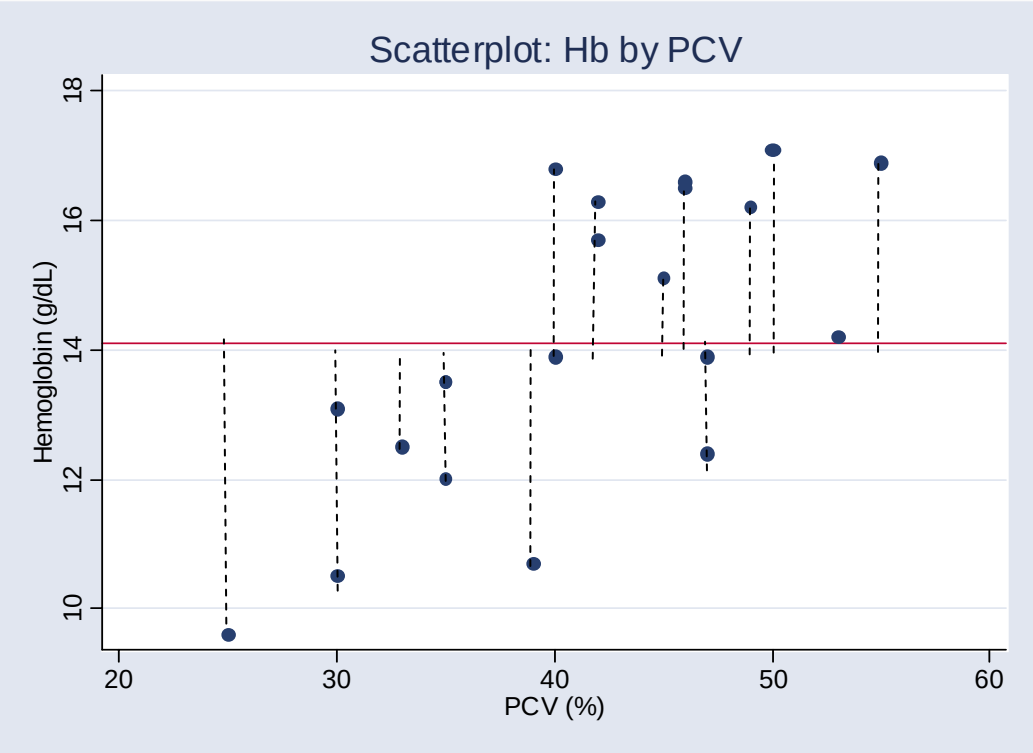
Suppose we only had 21 hemoglobin level measurements, and ignored information about PCV—the total variability in these 21 measures would be as follows:

$$\sum_{i=1}^{21}(y_i - \bar{y})^2$$

Continued

Example: Hemoglobin and Packed Cell Volume

Graphically speaking:



$\bar{y} = 14.1 g/dL$

Continued

Example: Hemoglobin and Packed Cell Volume

Now suppose we utilized the PCV values too, and fit a linear regression to relate mean hemoglobin to PCV: E[y] = 5.8 + .2*PCV

Now, the estimate will differ for different values of PCV—the total variation of the 21 observed hemoglobin levels around their means predicted by the regression equation is given by:

$$\sum_{i=1}^{21}(y_i - E[y])^2$$

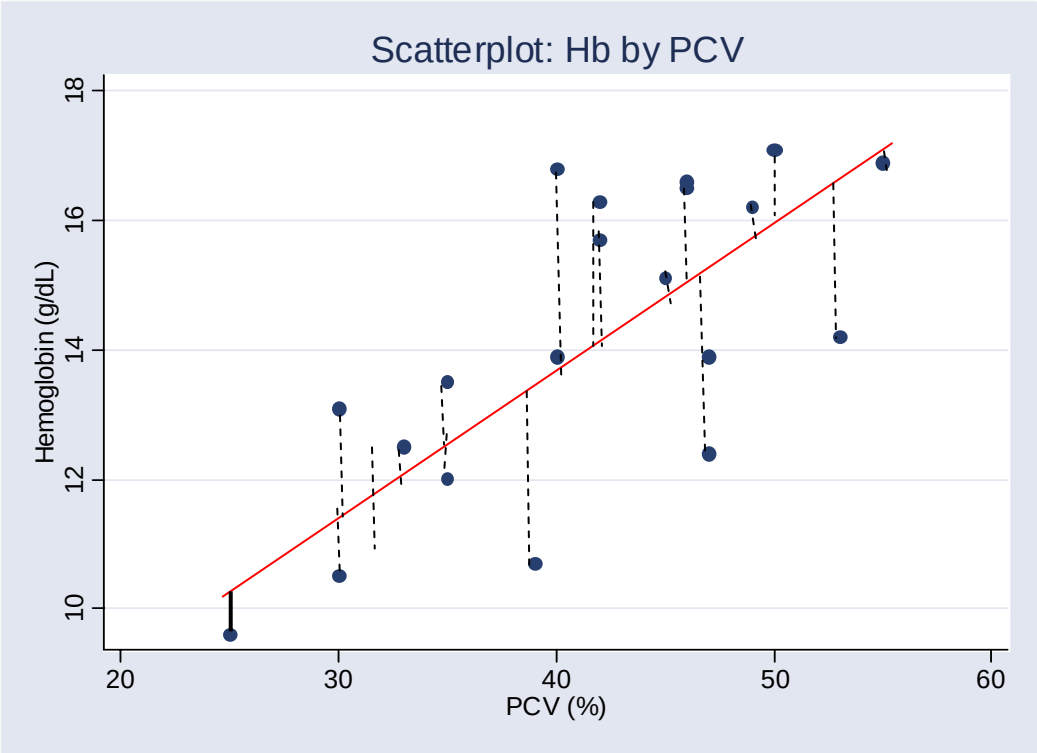
or

$$\sum_{i=1}^{21}(y_i - (5.8 + 0.2 * PCV))^2$$

Continued

Example: Hemoglobin and Packed Cell Volume

Graphically speaking:



Continued

Example: Hemoglobin and Packed Cell Volume

These are the discrepancies between the observed hemoglobin values and their corresponding estimated means given PCV, as estimated by the regression equation

These discrepancies are called "residuals"

You can think of the total discrepancy as the sum of these (squared) discrepancies, given by the formulas . . .

$$\sum_{i=1}^{21}(y_i - E[y])^2$$

or

$$\sum_{i=1}^{21}(y_i - (5.8 + 0.2 * PCV))^2$$

Continued

Example: Hemoglobin and Packed Cell Volume

R^2 is given by

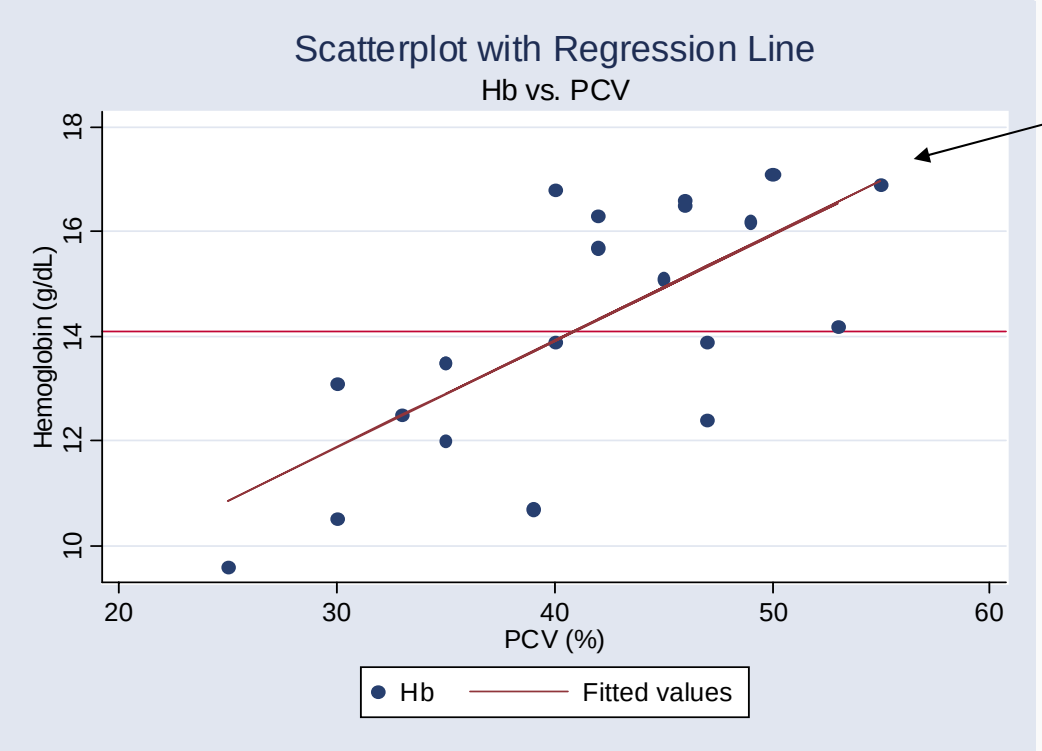
$$1 - \frac{\sum_{i=1}^{21}(y_i - E[y])^2}{\sum_{i=1}^{21}(y_i - \bar{y})^2}$$

The percentage of variability in *y* explained by *x*

Continued

Example: Hemoglobin and Packed Cell Volume

Graphically speaking—R^2 measures how much better a job we do estimating different means of *y* for different values of *x*, as opposed to estimating one overall mean



Regression line

$\bar{y} = 14.1 g/dL$

R^2 with Stata

I showed you a formula and pictures to give an explanation of where R^2 comes from and its interpretation

However, you will never have to compute by hand—Stata will do the work for you!

Continued

R^2 in Stata appears in upper right hand corner of regression output

```
. regress  Hb PCV

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =   19.81
       Model |  53.7803079     1  53.7803079            Prob > F      =  0.0003
    Residual |  51.5711174    19  2.71426934           R-squared     =  0.5105
-------------+------------------------------           Adj R-squared =  0.4847
       Total |  105.351425    20  5.26757126           Root MSE      =  1.6475

------------------------------------------------------------------------------
          Hb |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         PCV |   .2033502   .0456835     4.45   0.000     .1077335    .2989668
       _cons |    5.77645   1.913624     3.02   0.007     1.771188    9.781712
------------------------------------------------------------------------------
```

The Correlation Coefficient, r

A close cousin to R^2, the coefficient of determination is *r*, the correlation coefficient

Measures the direction and strength of the linear association between x and y

Just as it seems, $r = \sqrt{R^2}$ with one important characteristic—r takes the same sign as the slope

The Correlation Coefficient

R^2 is a number between 0 and 1

The closer R^2 is to 1, the closer absolute value of r is to 1

The correlation coefficient is between -1 and +1

- ★ $r > 0$ *Positive association*
- ★ $r < 0$ *Negative association*
- ★ $r = 0$ *No association*

Example: Hemoglobin and Packed Cell Volume

In the regression of hemoglobin on PCV we've been using, $R^2 = .51$, and the estimated slope $\hat{b}_1$ is 0.20, a positive number

In this scenario, $r = \sqrt{0.51} = .71$

Given r, you can always get R^2: given R^2, you can get r if you know direction of association (need to look at slope)

Continued

Example: Hemoglobin and Packed Cell Volume

So in our example, PCV explains about 51% of the variability in hemoglobin, and r = .71

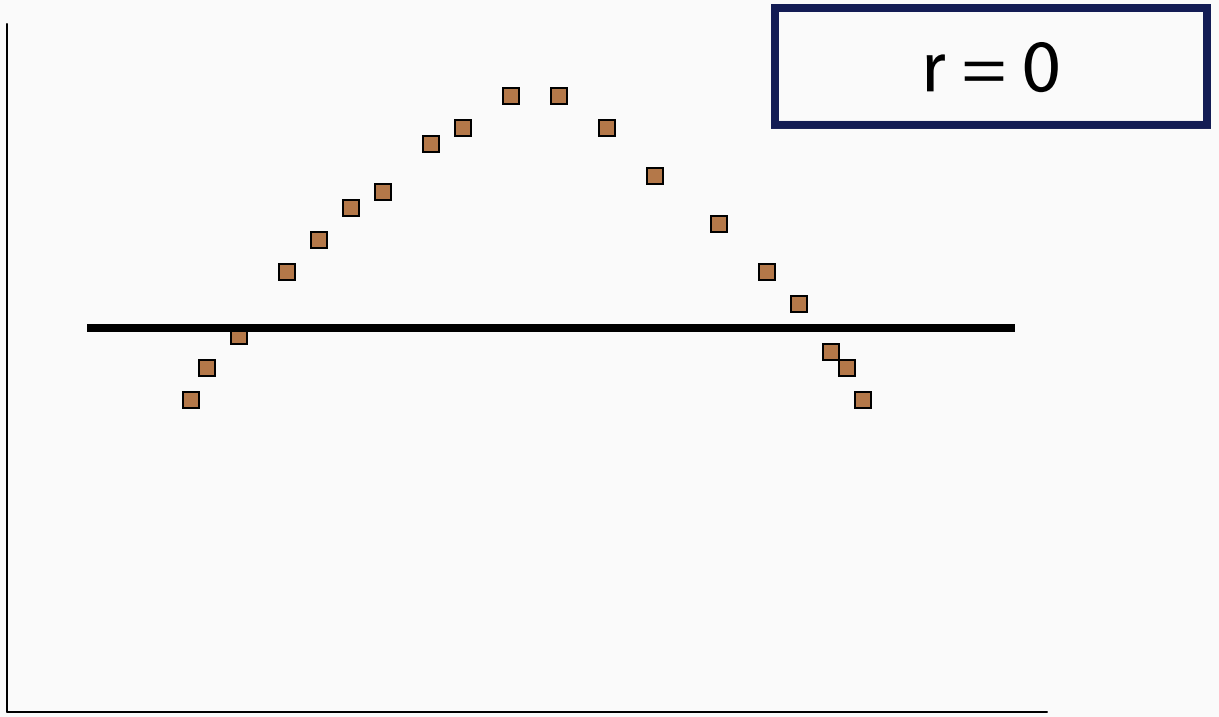
Is this good?

- ★ *Not necessarily good or bad*
- ★ *Roughly 49% of the variability is not explained by PCV- perhaps some of this could be explained by additional information (person's age, smoking status, etc..)*
- ★ *We will see how to use more than one predictor to explain behavior of an outcome in the next section (multiple regression)*

It is important to note that ***r*** and R^2 measures strength of linear association

There could be a strong non-linear relationship between ***y*** and ***x***, and ***r*** and R^2 may not catch it

Continued



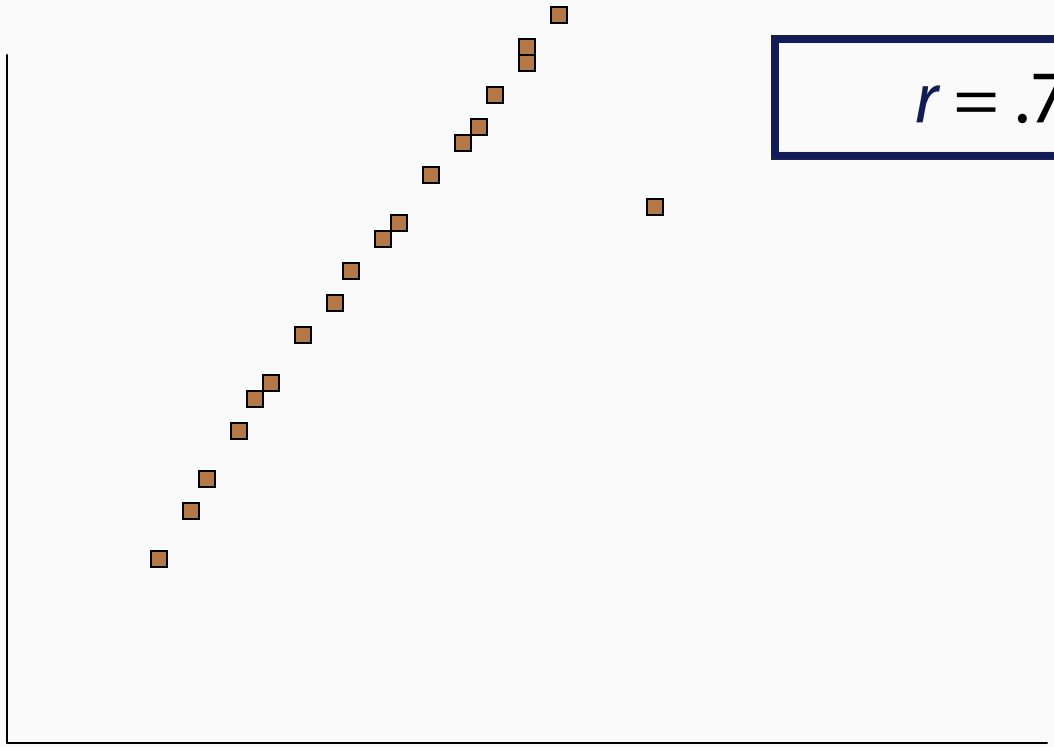
No Linear Correlation!

Continued

Outliers can really affect ***r*** and R^2

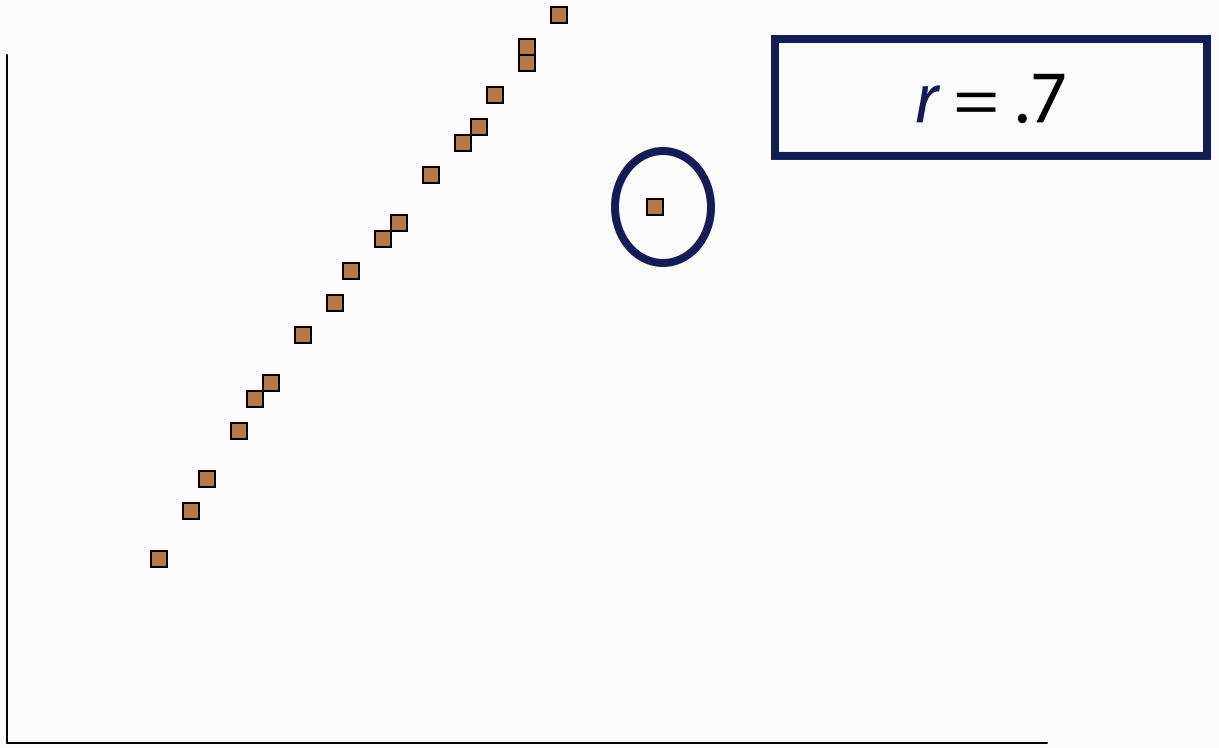
One extreme point can change ***r*** sizably

Continued



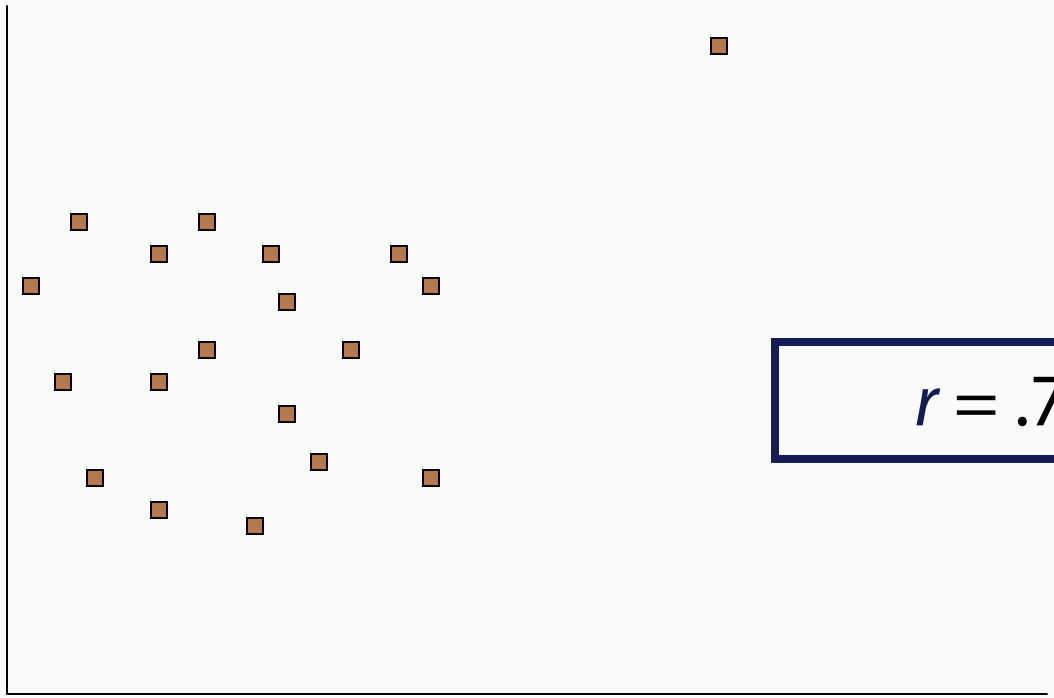
$r = .7$

Continued



Continued

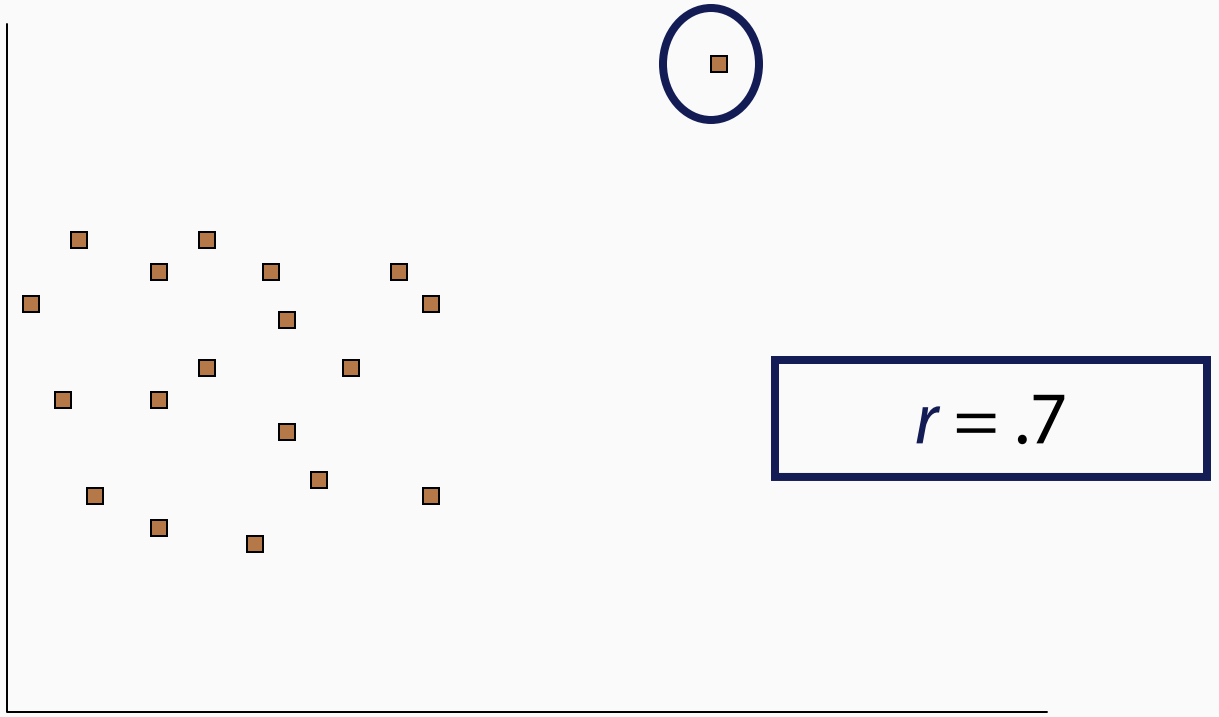
Correlation



$r = .7$

Continued

Correlation



$r = .7$

The Moral

A picture is worth . . .

Always look at your scatterplot when trying to interpret correlation coefficients

In fact—always look at your data before using regression to see if it makes sense to use!

Slope (B^1) vs. the Correlation Coefficient (r)

Both indicate direction of association (positive or negative)

Slope $\hat{b}_1$ is the estimated expected change in ***y*** per unit increase in ***x***

Larger slopes *do not* necessarily mean a *stronger* linear association and smaller slopes do not necessarily mean weaker linear association

The slope depends on the units data is in

Continued

Slope (B^1) vs. the Correlation Coefficient (r)

The correlation coefficient measures how close points fall on a linear line

Estimate of ***r*** (and hence R^2) does not depend on units data is in

Continued

Slope (B^1) vs. the Correlation Coefficient (r)

For example, suppose we measure Hemoglobin in mg/dL instead of g/dL, and regressed on PCV (%)

```
. regress Hb_milg PCV

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =   19.81
       Model |  53780314.3     1  53780314.3           Prob > F      =  0.0003
    Residual |  51571114.1    19  2714269.16           R-squared     =  0.5105
-------------+------------------------------           Adj R-squared =  0.4847
       Total |   105351428    20  5267571.42           Root MSE      =  1647.5

------------------------------------------------------------------------------
     Hb_milg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         PCV |   203.3502    45.6835     4.45   0.000     107.7335    298.9668
       _cons |    5776.45   1913.624     3.02   0.007     1771.188    9781.711
------------------------------------------------------------------------------
```

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section D

Practice Problems

1. Consider the general form of the simple linear regression equation:

$$y = \hat{b}_0 + \hat{b}_1 x_1$$

Continued

a. Suppose we computed ***r***, the correlation coefficient from the above regression? What different information about the relationship between ***y*** and ***x*** does this convey, as compared to $\hat{b}_1$?

Continued

b. If $|r|$ is small, does this guarantee the relationship between y and x is not so strong?

Continued

c. Sketch "mock scatterplots" of fake data illustrating the following four possible scenarios:

1. Small $|\hat{b}_1|$, small $|\mathbf{r}|$. (both close to 0)
2. Small $|\hat{b}_1|$, large $|\mathbf{r}|$. ($|\mathbf{r}|$ close to 1)
3. Large $|\hat{b}_1|$, small $|\mathbf{r}|$.
4. Large $|\hat{b}_1|$, large $|\mathbf{r}|$.

Continued

2. Consider the following results from a SLR regressing weight (lbs) on height (in) for a sample of 12 nutritionally deficient children. The results, below, are missing some key information.

```
. regress wt ht

  Source |       SS       df       MS                  Number of obs =      12
---------+------------------------------                F(  1,    10) =   19.67
   Model |  588.922523     1  588.922523                 Prob > F      =  0.0013
Residual |  299.327477    10  29.9327477                 R-squared     =  0.6630
---------+------------------------------                Adj R-squared =  0.6293
   Total |      888.25    11       80.75                 Root MSE      =  5.4711

------------------------------------------------------------------------------
      wt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      ht |    1.07223    .241731
   _cons |   6.189849   12.84875
```

Continued

a. What is the correlation coefficient, ***r***, measuring the degree of linear correlation between weight and height?

Continued

b. What is the interpretation of the R-squared value for this regression?

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section D

Practice Problem Solutions

1. Consider the general form of the simple linear regression equation:

$$y = \hat{b}_0 + \hat{b}_1 x_1$$

Continued

a. Suppose we computed ***r***, the correlation coefficient from the above regression? What different information about the relationship between ***y*** and ***x*** does this convey, as compared to $\hat{b}_1$?

The correlation coefficient ***r*** gives information about the strength of the linear relationship between ***y*** and x^1, whereas the slope b^1 gives information about the magnitude of this relationship. Both ***r*** and b^1 give information about the direction of the relationship, and will have the same sign.

b. If $|r|$ is small, does this guarantee the relationship between y and x is not so strong?

No. Recall, r measures the degree of linear relationship—it is possible to have a strong relationship that is not linear in nature, and hence get a value of r close to 0.

Continued

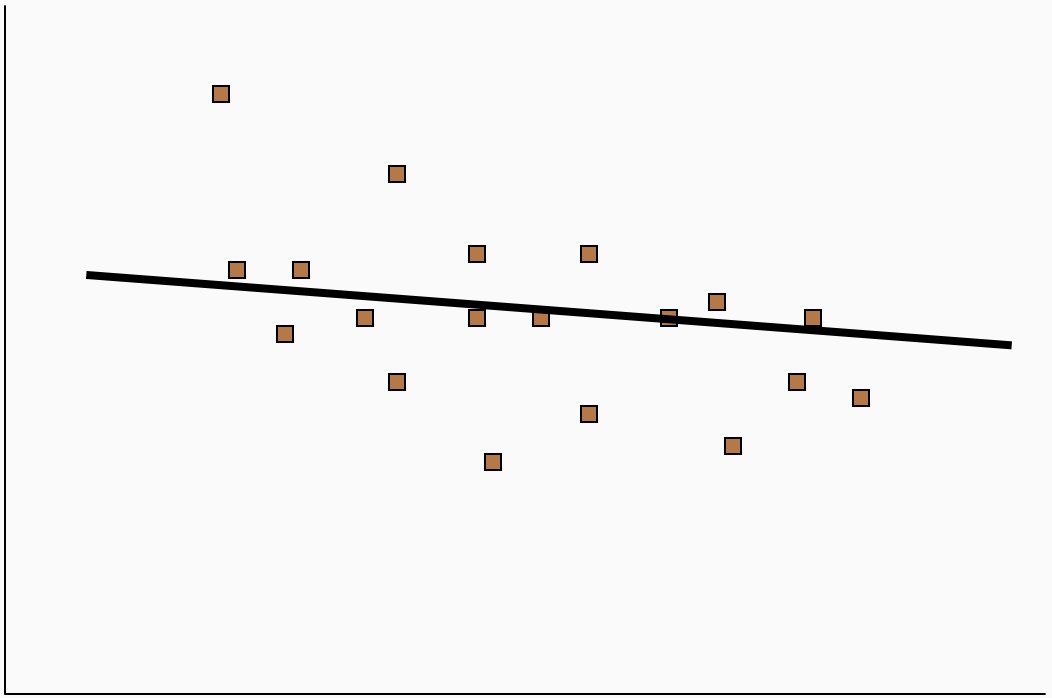
c. Sketch "mock scatterplots" of fake data illustrating the following four possible scenarios:

1. Small $|\hat{b}_1|$, small $|\mathbf{r}|$. (both close to 0)
2. Small $|\hat{b}_1|$, large $|\mathbf{r}|$. ($|\mathbf{r}|$ close to 1)
3. Large $|\hat{b}_1|$, small $|\mathbf{r}|$.
4. Large $|\hat{b}_1|$, large $|\mathbf{r}|$.

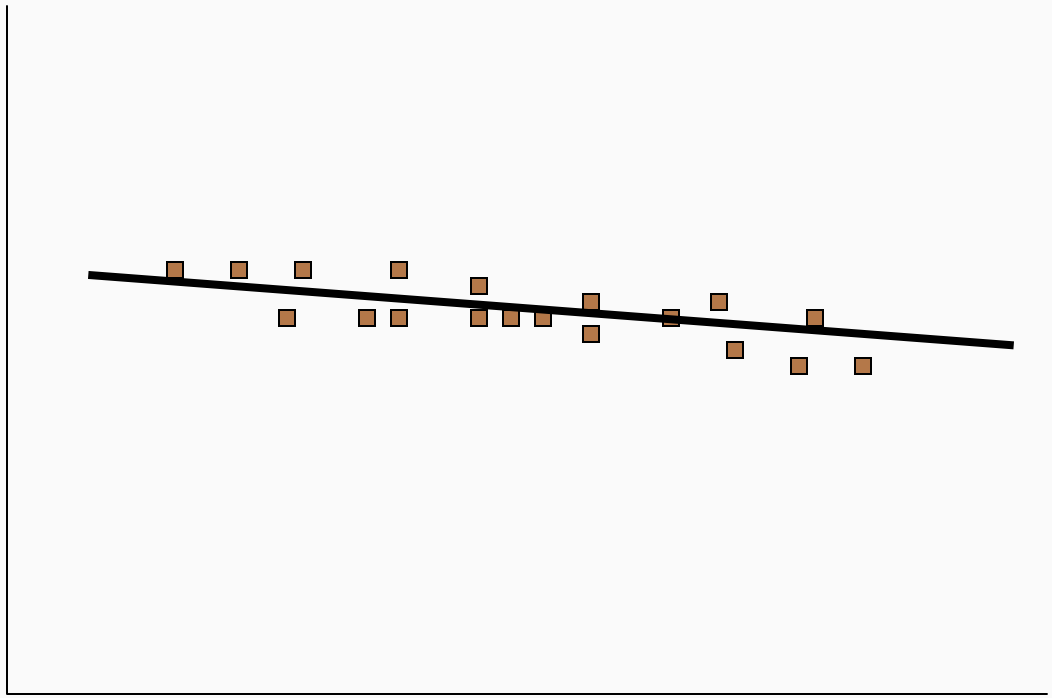
Continued

Practice Problem Solutions

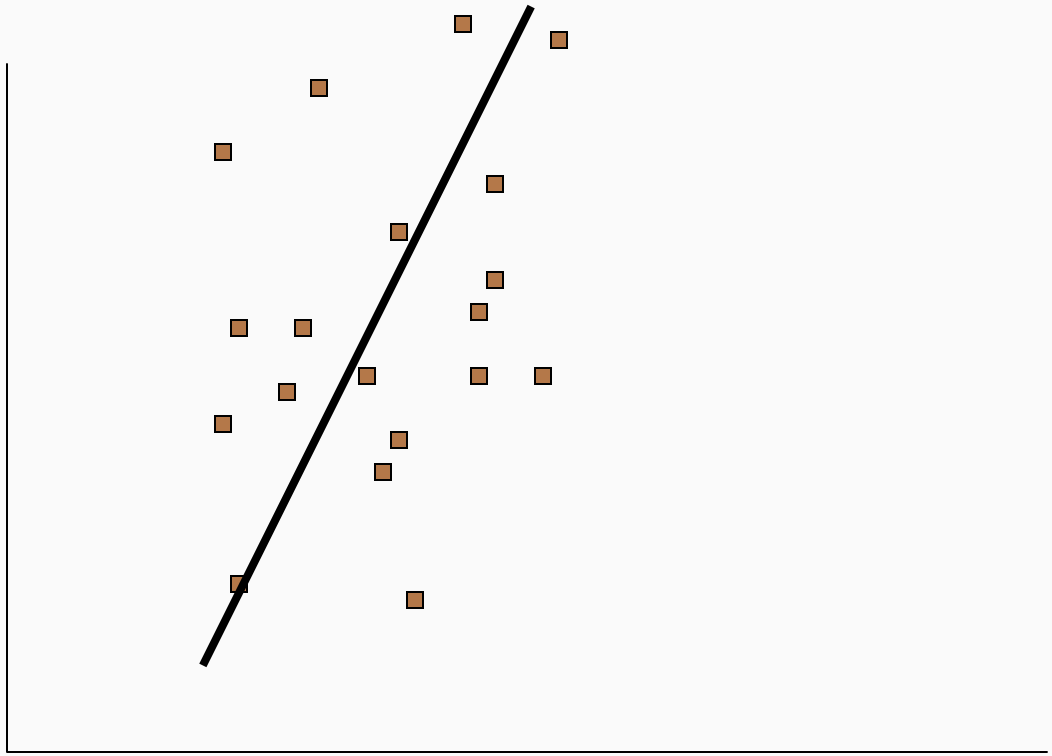
1. Small $|\hat{b}_1|$, small $|\mathbf{r}|$. (both close to 0)



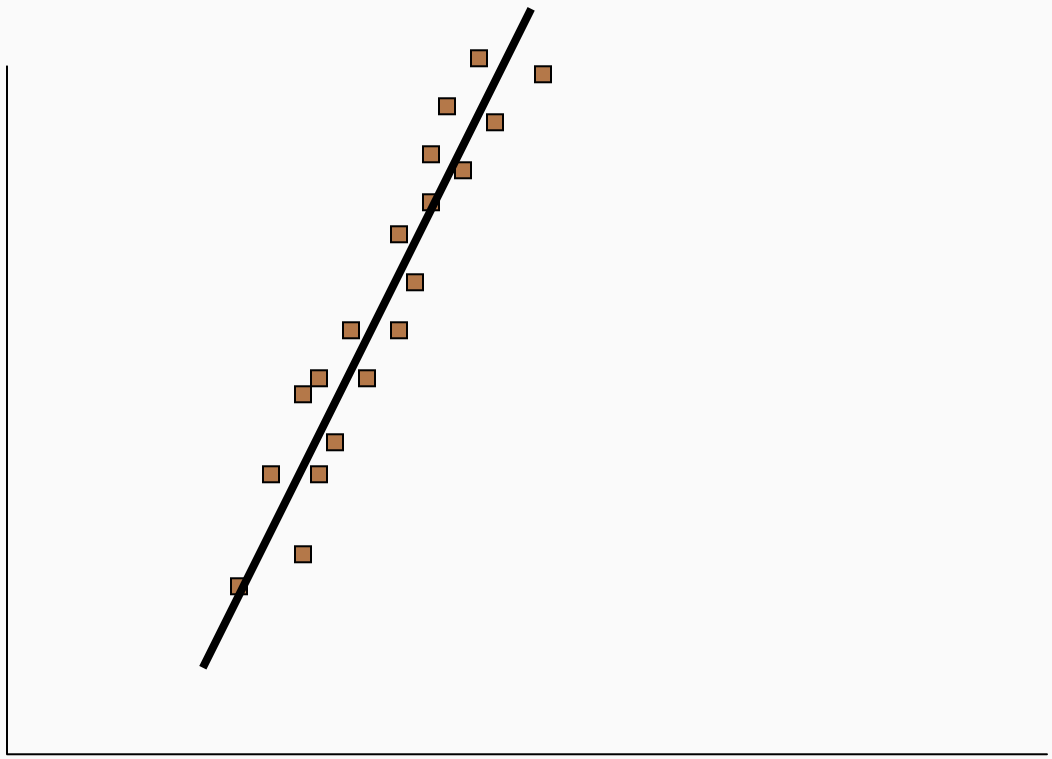
2. Small $|\hat{b}_1|$, large |**r**|. (|**r**| close to 1)



3. Large $|\hat{b}_1|$, small $|\mathbf{r}|$.



4. Large $|\hat{b}_1|$, large $|\mathbf{r}|$.



2. Consider the following results from a SLR regressing weight (lbs) on height (in) for a sample of 12 nutritionally deficient children. The results, below, are missing some key information.

```
. regress wt ht

  Source |       SS       df       MS                  Number of obs =      12
---------+------------------------------                F(  1,    10) =   19.67
   Model |  588.922523     1  588.922523                Prob > F      =  0.0013
Residual |  299.327477    10  29.9327477                R-squared     =  0.6630
---------+------------------------------                Adj R-squared =  0.6293
   Total |      888.25    11       80.75                Root MSE      =  5.4711

------------------------------------------------------------------------------
      wt |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      ht |    1.07223    .241731
   _cons |   6.189849   12.84875
```

Continued

a. What is the correlation coefficient, ***r***, measuring the degree of linear correlation between weight and height?

Recall, ***r*** is equal in absolute magnitude to $\sqrt{R^2}$, and the sign of ***r*** corresponds to the sign of $\hat{b}_1$.

So for this data set, the R-squared value is 0.66, and hence $r = \sqrt{0.66} = \sqrt{R^2} = \pm.81$. As $\hat{b}_1$ is positive, $r = 0.81$.

Continued

b. What is the interpretation of the R-squared value for this regression?

The R-squared value is 0.66 or 66%. Roughly 66% of the original variability in children's weight is "explained" by the children's height.

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section E

The T-Test as a Linear Regression Model

Linear Regression with a Binary X

So far we've been looking at a linear regression example where our predictor is continuous

Linear regression can handle predictors that are binary as well

- ★ *Linear regression is a method for relating the mean of a continuous outcome to a single predictor—this predictor need not be continuous*

The t-test we learned about in 611 is a special case of simple linear regression where x is binary

Continued

Linear Regression with a Binary X

Example—a sample of 312 patients with Primary Biliary Cirrhosis (PBC) studied between 1972 and 1984 at the Mayo Clinic

Patient information collected included bilirubin levels and histologic stage of disease (1-3 and 4)

Suppose we wanted to assess whether the mean bilirubin levels were different by stage of disease groups (1-3 vs. 4)

Continued

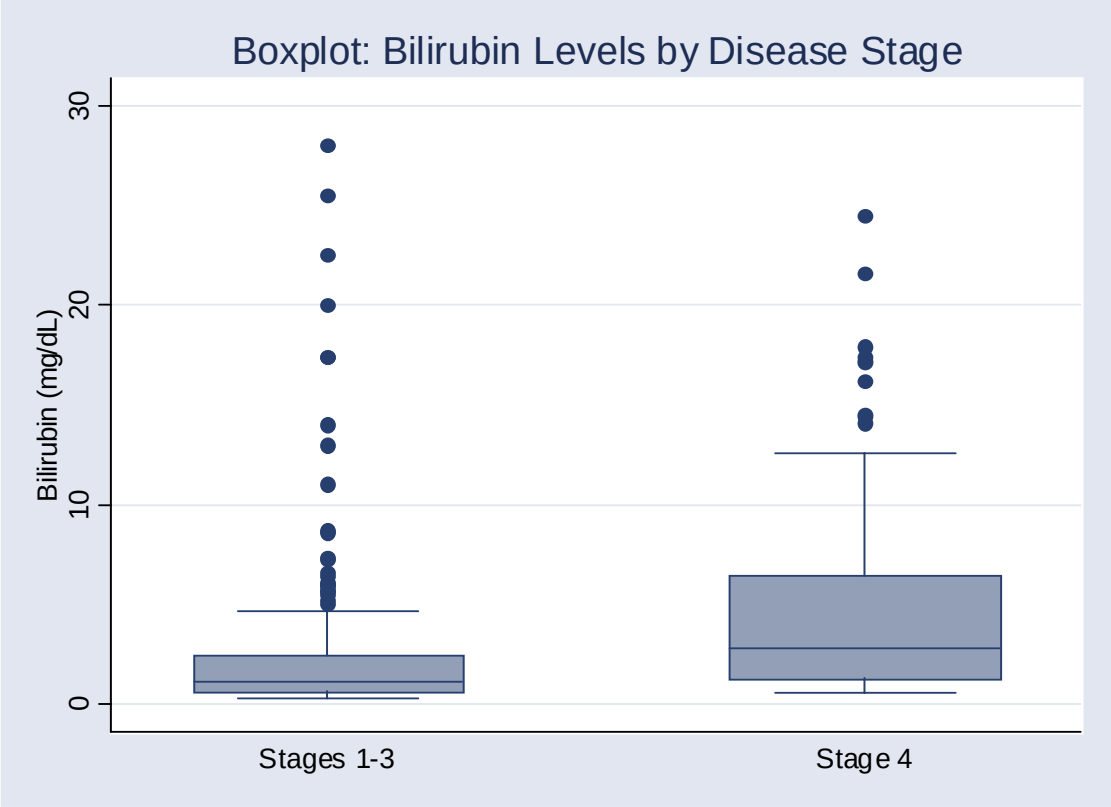
Linear Regression with a Binary X

Data for first ten subjects

- *"bil" is bilirubin level in mg/dL*
- *"hstage4" is binary indicator*
  - 1 if stage 4 disease
  - 0 id stages 1-3

	bil	hstage4
1.	1.1	0
2.	.8	0
3.	.5	0
4.	.6	0
5.	.5	0
6.	.9	0
7.	.6	0
8.	1.2	0
9.	.6	0
10.	1	0

Boxplots of bilirubin levels by disease stage



Mean bilirubin levels by disease stage

Stage	n	Mean bilirubin	SD
1-3	203	2.5	.28
4	109	4.7	.48

Results from a t-test

```
. ttest bil,by( hstage4)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Stages 1 |     203    2.501478    .2835828    4.040433    1.942316     3.06064
 Stage 4 |     109    4.661468    .4837083    5.050063    3.702674    5.620262
---------+--------------------------------------------------------------------
combined |     312     3.25609    .2564786    4.530315    2.751437    3.760742
---------+--------------------------------------------------------------------
    diff |            -2.15999    .5246684                -3.192352   -1.127628
------------------------------------------------------------------------------
Degrees of freedom: 310

                Ho: mean(Stages 1) - mean(Stage 4) = diff = 0

     Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
       t =  -4.1169                 t =  -4.1169                 t =  -4.1169
   P < t =   0.0000             P > |t| =   0.0000             P > t =   1.0000
```

Results

Patients in stages one through three of PBC disease have statistically significantly lower bilirubin levels on average than patients in stage four of the disease

The stage one through three patients have average bilirubin level of 2.15 mg/dL lower than the stage four patients (95% CI : 1.12 mg/dL – 3.19 mg/dL lower)

Could also report that stage four patients have average bilirubin level 2.15 mg/dL higher than stage one through three patients

Could This be Done as a Regression?

We could also do this as a regression!

- *Same set up*

$$E[y] = \hat{b}_0 + \hat{b}_1 x$$

- *Where:*
- *E[y] =estimated expected (mean) bilirubin*
- $\hat{b}_0$ = *estimated y-intercept*
- $\hat{b}_1$ = *estimated slope*
- *x = disease stage (1 = stage 4, 0 = stages 1–3)*

Continued

Could This be Done as a Regression?

Notice, we only have two values of x!

If we talk about $\hat{b}_1$ as the estimated change in mean of y for one-unit increase in x, we have only one possible "one-unit increase" in x—from 0 to 1

So $\hat{b}_1$ is a comparison of means of y between subjects with x = 1 and x = 0

Continued

Could This be Done as a Regression?

Recall the equation

$$E[y] = \hat{b}_0 + \hat{b}_1 x$$

- *So for subjects with x = 1:*

$$E[y] = \hat{b}_0 + \hat{b}_1 * 1 = \hat{b}_0 + \hat{b}_1$$

- *So for subjects with x = 0:*

$$E[y] = \hat{b}_0 + \hat{b}_1 * 0 = \hat{b}_0$$

Continued

Could This be Done as a Regression?

So $\hat{b}_1$ is the estimated mean difference in y for subjects with x = 1 compared to subjects with x = 0

In the bilirubin/stage of PBC disease example,
$\hat{b}_1$ is an estimate of the mean difference in bilirubin levels for stage four patients compared to stage one through three patients

Continued

Could This be Done as a Regression?

If we fit this regression in Stata:

```
. regress bil hstage4

      Source |       SS       df       MS              Number of obs =     312
-------------+------------------------------           F(  1,   310) =   16.95
       Model |  330.880709     1  330.880709            Prob > F      =  0.0000
    Residual |  6052.00773   310  19.5226056            R-squared     =  0.0518
-------------+------------------------------           Adj R-squared =  0.0488
       Total |  6382.88844   311   20.523757            Root MSE      =  4.4184

------------------------------------------------------------------------------
         bil |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hstage4 |    2.15999   .5246684     4.12   0.000     1.127628    3.192352
       _cons |   2.501478   .3101136     8.07   0.000     1.891284    3.111672
------------------------------------------------------------------------------
```

Continued

Could This be Done as a Regression?

Here is $\hat{b}_1$ and a 95% CI for true mean difference (b_1)

```
. regress bil hstage4

      Source |       SS       df       MS              Number of obs =     312
-------------+------------------------------           F(  1,   310) =   16.95
       Model |  330.880709     1  330.880709            Prob > F      =  0.0000
    Residual |  6052.00773   310  19.5226056            R-squared     =  0.0518
-------------+------------------------------           Adj R-squared =  0.0488
       Total |  6382.88844   311   20.523757            Root MSE      =  4.4184

------------------------------------------------------------------------------
         bil |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hstage4 |    2.15999   .5246684     4.12   0.000     1.127628    3.192352
       _cons |   2.501478   .3101136     8.07   0.000     1.891284    3.111672
------------------------------------------------------------------------------
```

Results

Patients in stages four of PBC disease have statistically significantly higher bilirubin levels on average than patients in stages one through three of the disease

The stage four patients have an average bilirubin level of 2.15 mg/dL higher than the stage one through three patients (95% CI : 1.12 mg/dL–3.19 mg/dL higher)

Could also report that stage one through three patients have average bilirubin level 2.15 mg/dL lower than stage four patients

Comparison with Results from T-Test

Notice, this is exactly the result we get from a t-test, except the comparison is in the opposite direction

```
. ttest bil,by( hstage4)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Stages 1 |     203    2.501478    .2835828    4.040433    1.942316     3.06064
 Stage 4 |     109    4.661468    .4837083    5.050063    3.702674    5.620262
---------+--------------------------------------------------------------------
combined |     312     3.25609    .2564786    4.530315    2.751437    3.760742
---------+--------------------------------------------------------------------
    diff |            -2.15999    .5246684               -3.192352   -1.127628
------------------------------------------------------------------------------
Degrees of freedom: 310

                 Ho: mean(Stages 1) - mean(Stage 4) = diff = 0

     Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
       t =  -4.1169                 t =  -4.1169                 t =  -4.1169
   P < t =   0.0000             P > |t| =   0.0000             P > t =   1.0000
```

Could this be Done as a Regression?

We also get, as a bonus, a R^2 value!

```
. regress bil hstage4

      Source |       SS       df       MS              Number of obs =     312
-------------+------------------------------           F(  1,   310) =   16.95
       Model |  330.880709     1  330.880709           Prob > F      =  0.0000
    Residual |  6052.00773   310  19.5226056           R-squared     =  0.0518
-------------+------------------------------           Adj R-squared =  0.0488
       Total |  6382.88844   311   20.523757           Root MSE      =  4.4184

------------------------------------------------------------------------------
         bil |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hstage4 |    2.15999   .5246684     4.12   0.000     1.127628    3.192352
       _cons |   2.501478   .3101136     8.07   0.000     1.891284    3.111672
------------------------------------------------------------------------------
```

Continued

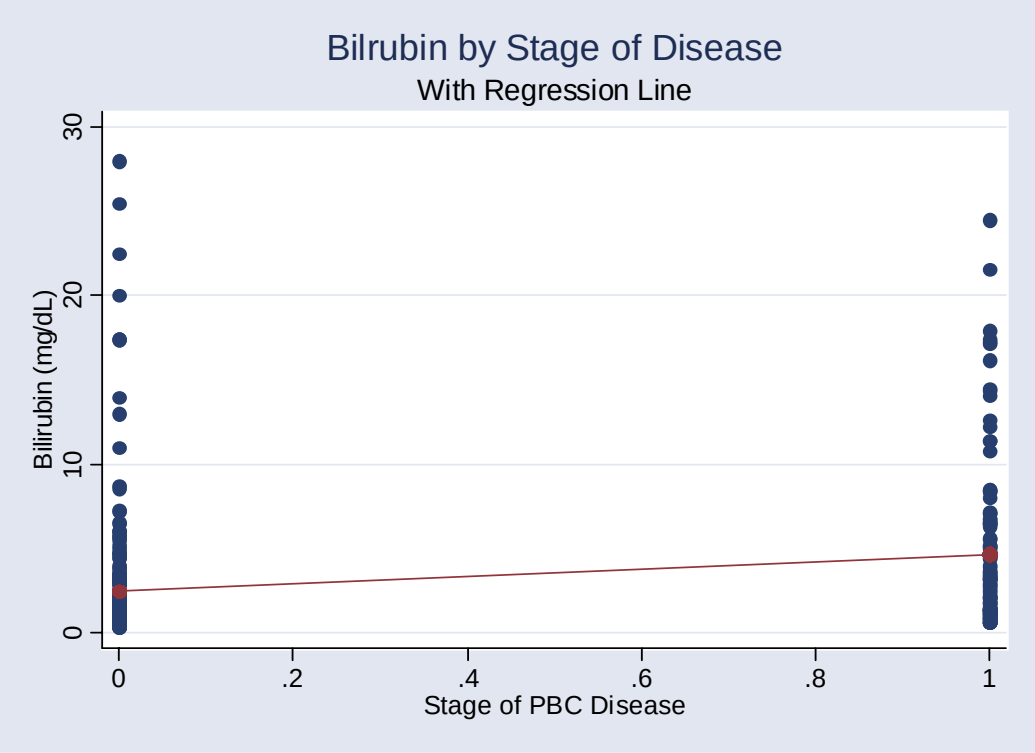
Could this be Done as a Regression?

In this situation, the intercept also gives some useful information:

```
. regress bil hstage4

      Source |       SS       df       MS              Number of obs =     312
-------------+------------------------------           F(  1,   310) =   16.95
       Model |  330.880709     1  330.880709           Prob > F      =  0.0000
    Residual |  6052.00773   310  19.5226056           R-squared     =  0.0518
-------------+------------------------------           Adj R-squared =  0.0488
       Total |  6382.88844   311   20.523757           Root MSE      =  4.4184

------------------------------------------------------------------------------
         bil |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hstage4 |    2.15999   .5246684     4.12   0.000     1.127628    3.192352
       _cons |   2.501478   .3101136     8.07   0.000     1.891284    3.111672
------------------------------------------------------------------------------
```

Where is the "Line"?

So a t-test is a special case of linear regression—so where is the "line"?



Continued

So a t-test is a special case of linear regression—so where is the "line"?