GPPI PPOL501-02 & -06: Fall 2014
Course Notes #17: Bivariate Regression – The Basics
Professor Carolyn Hill

I. FROM ASSOCIATION TO PREDICTION

- We've used the correlation coefficient to answer the questions of (1) whether a relationship between two variables existed (i.e., whether the association was statistically significant; (2) the direction of the association (i.e., the sign of the correlation); and (3) how strong that relationship was (i.e., the magnitude of the absolute value of the correlation).

- What if we want to make a prediction, make a comparison, or assess effects more precisely? The correlation coefficient is less useful for such analyses.

  Question: Using available data, what can we say about the effectiveness of past policies or programs?

  -- Example: Were the welfare reforms of 1996 successful in moving clients off welfare, and increasing their earnings?

  Question: Using available data, what can we say about likely effects of a policy or program proposal?

  -- Example: Should states (or the federal government, or local jurisdictions) implement a policy to encourage housing construction or rehabilitation in suburban and rural areas? What might be the effects on crime rates?

- To answer these questions, we still need some idea of the association between two variables (e.g., urbanization and crime rates; welfare reform and client earnings), but a different type of info about that relationship

II. FURTHER REVIEW OF FUNCTIONAL (LINEAR) RELATIONSHIPS AND REGRESSION

- The correlation coefficient gave us information about the degree to which points were clustered around a straight line…but nothing about the *slope* of that line. Regression analysis will produce this kind of information.

- Step back and think very generally about describing the association between variables:

  $Y = f(X)$

  That is, some variable *Y* is a function of variable *X*. *Y* is somehow determined by, or associated with, or caused by, *X*. (As you know, these are not all the same things, but may be relevant under different circumstances)

- Terminology:

Y	X
Dependent Variable	Independent Variable
Explained Variable	Explanatory Variable
Response Variable	Control Variable
Predicted Variable	Predictor Variable
Outcome Variable	Covariate
Left hand side Variable	Right hand side Variable
Regressand	Regressor

- *Y* may be a function of many different things (*X*s).

  Example: What factors might influence a person's wages besides their IQ?

- For now, we'll build the foundation by considering just *one* explanatory variable:

  Example: Wage = *f* (IQ)

- In this notation, we have not yet defined exactly *how* IQ and wages are related. In other words, we have just specified that there is some association or correlation between the two variables. Exactly how they are related is referred to as the *functional form* of the relationship between *X* and *Y*. There are a number of possibilities for this relationship:

  -- Wages = $(IQ)^2$
  -- Wages = $\sqrt{(IQ)}$
  -- Wages = k*(IQ) (where k is some constant number)
  -- etc……

For example, given a set of data points, I might attempt to draw a “line” that goes through each point:

```
. plot wage iq, encode

    2404 +
         |                                              *
         |
         |                                        *
         |       *  *
         |                             *  *  **                   *
         |                                *
         |                                       **         *    *
    w    |                            *  *                              *
    a    |                          *   2  * *    2**           *
    g    |                            *    *2 * ****  2     *   *  *
    e    |                      2      22 22 *   * 2*2*  ** 4    *
         |           *  *          3*2   22 424 *3 323  *2
         |               *    **2***4**37523 245*5 3 27*32642* *3
         |             2    2** **2**32 3438473940293562722252  *
         |           * 2 2 * 3*2445*5594B08GBH0G9H7G6HA68323624 *3
         | * 2*     2**  32**2*58684D4I5LAIPAZCJEP4LCF69A6A65*4 5 24
         |          2**5*337*6599AF796B4PCSMAICE8H7E9J970935436**3 2 * *
         | **  ** 2 *53535444*25A5D8D7G8F4GJ6HCH6B8F9FH385433242* *2*      *
         |       22*4533*338 4*4848206B6E38E573A56296484634***  ** * *   *
     112 +       **  2****25    3 **2 32* ** * 6   **3 *2 * *  2   *
          +-----------------------------------------------------------------+
               50                       iq                            149

.
end of do-file
```

- Why might I not want to do this?

- To simplify the world (and our lives), we’ll focus for now on *linear relationships*. That is, the relationship between the wages and IQ can be **approximated** by a straight line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \qquad WagesHat = \hat{\beta}_0 + \hat{\beta}_1 IQ$$

- Clearly, assuming that the relationship is linear between these two variables (or any two variables) may be problematic. But it provides an **approximation** of that relationship. It is up to the analyst to assess whether that approximation is a meaningful one.

- But there are many possible straight lines that might be drawn through a set of points…How to choose the “best” one?

- How does an expression of a linear relationship relate to our understanding of social phenomena, or our ability to predict anything about the effects of a policy?

III. LINEAR REGRESSION (ORDINARY LEAST SQUARES)

- WHICH STRAIGHT LINE should we select to describe the relationship between X and Y, given that I assume that the relationship is linear?

****

The regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ is selected *to minimize the sum of the squared distances between each point and the regression line.*

This is the best-fitting line possible given any set of data points. Best-fitting in what sense? Best-fitting in that it minimizes the sum of the squared distances from the actual to the estimated values.

The process of calculating/selecting this particular best-fitting line is referred to as "minimizing the sum of squared residuals" or "minimizing the sum of squared errors."

This process for fitting a line to the data is referred to as Ordinary Least Squares regression, or OLS. It is the workhorse of policy analysis, although much more complicated methods (sometimes more useful, sometimes not) exist and are used. The regression line is sometimes referred to as the "least-squares line" for the reasons above.

$$\text{minimize} \sum_i e_i^2 = \min \sum_i (Y_i - \hat{Y}_i)^2 = \min \sum_i [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2$$

****

- The first term is called "Y-hat" or "predicted Y" and we obtain it for a particular observation by summing $\hat{\beta}_0 + \hat{\beta}_1$*(*X* value for that observation). This predicted value will give a point in the regression line. It is a *best guess* of a value for *Y*, given a value of *X* (i.e., *conditional on X)*

- (See final page of these notes for the derivation of these terms):

$\hat{\beta}_0$ = the intercept or constant term. It is the value of Y when X=0 (i.e., the value of Y where the regression line crosses the Y axis)

Note: The constant may not always be meaningful if the data you are interested in never have a value of X=0

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

$\hat{\beta}_1$ = the slope or coefficient on X: how is Y predicted to change, given a one-unit change in X?

Slope = $\dfrac{\textit{change in Y}}{\textit{change in X}} = \dfrac{\Delta Y}{\Delta X}$.

$$\hat{\beta}_1 = \frac{\text{cov}\,ariance(X,Y)}{\text{var}\,iance(X)} = \frac{s_{xy}}{s_x^2} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{n\sum_i X_i Y_i - \left(\sum_i X_i\right)\left(\sum_i Y_i\right)}{n\sum_i X_i^2 - (\sum_i X_i)^2}$$

-- Note the similarity between the formula for $\hat{\beta}_1$ and the formula for **Pearson's correlation coefficient**: the numerator of each statistic is some measure of how X and Y move together, in relation to their respective means.

-- Also note that a simple regression runs through the pair of means.

<u>The population regression function</u>

- This builds on everything you've learned about statistical inference – that is, you are making some statement about the *population of interest* from the *sample data* that are available.

- We would be able to state the "true" the relationship between *X* and *Y* if we had all the data in the world about *X* and *Y*. As with t-tests, correlations, etc., the "<u>true</u>" model for linear regression is written using Greek letters:

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

  Where β_0 = "beta-naught" is the *true* intercept, or constant term
  β_1 = "beta-one" is the *true* slope, or coefficient on X
  u = u is the *true* residual, or error, or unobserved part of Y

- Because we seldom can get all the data in the world about X and Y, we have to *estimate* the relationship between X and Y. This process is analogous to all the other kinds of statistical inference we've done up to this point in Quant I. The estimated model is usually written with hats over the Greek letters, to show that the parameters are estimates of the true values.

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

  Where $\hat{\beta}_0$ = the *estimated* intercept, or constant term
  $\hat{\beta}_1$ = the *estimated* slope, or coefficient on X
  e = the *estimated* residual, or error

  -- A picture of the "true" population regression function and estimated regression lines (with a link back to why we do hypothesis testing)….

Key Assumptions in Simple Least Squares Regression

SLR1: Linear in the parameters: $Y = \beta_0 + \beta_1 X + u$

SLR2: Random sampling

SLR3: Zero conditional mean: $E(u|X) = 0$

** Think of SLR3 as: *X* and *u* are uncorrelated. What's *u*? Anything else that may affect Y that is not included in the regression model.

** Think some more: do you think this is likely to hold for simple regression models?

SLR4: The values of X must vary in the sample. (think about the formula for β_1 for understanding why variation in X is a necessary condition for estimating the model)

⇒ Under Assumptions **SLR1 through SLR**4, $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$

In other words, **when these assumptions hold**, $\hat{\beta}_0$ is an *unbiased estimator* of β_0, and $\hat{\beta}_1$ is an *unbiased estimator* for β_1

SO…….

- If any of the four previous assumptions fail, then we cannot assume that the regression estimates will produce unbiased estimates of the population parameters.
- Even if all the assumptions hold, this does *not* guarantee that the parameter estimates from a particular estimated model are exactly equal to the population parameters. **WHY NOT?**

SLR5: Equal variances, conditional on X. The conditional variance is the same for each different value of the explanatory variable. (i.e., σ^2 is constant, or the same, for each value of the explanatory variable):

$Var(u|X) = \sigma^2$ (i.e., variance is homoskedastic)

IV. INTERPRETING REGRESSION PARAMETER ESTIMATES

- Is there a relationship between IQ scores and wages?

```
. summarize wage iq

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |      3010    577.2824    262.9583        100       2404
          iq |      2061    102.4498    15.42376         50        149

. summarize wage iq, detail

                            wage
-------------------------------------------------------------
      Percentiles      Smallest
 1%          188            100
 5%          250            100
10%          291            112       Obs                3010
25%          394            117       Sum of Wgt.        3010

50%        537.5                      Mean           577.2824
                        Largest       Std. Dev.      262.9583
75%          709           2083
90%          894           2115       Variance       69147.07
95%         1039           2244       Skewness       1.479585
99%         1442           2404       Kurtosis       7.482434

                             iq
-------------------------------------------------------------
      Percentiles      Smallest
 1%           65             50
 5%           74             51
10%           83             53       Obs                2061
25%           93             54       Sum of Wgt.        2061

50%          103                      Mean           102.4498
                        Largest       Std. Dev.      15.42376
75%          113            144
90%          122            145       Variance       237.8923
95%          126            146       Skewness      -.3112868
99%          136            149       Kurtosis       3.091002


. regress wage iq

      Source |       SS       df       MS              Number of obs =    2061
-------------+------------------------------           F(  1,  2059) =   78.82
       Model |  5288704.88     1  5288704.88           Prob > F      =  0.0000
    Residual |   138159366  2059  67100.2265           R-squared     =  0.0369
-------------+------------------------------           Adj R-squared =  0.0364
       Total |   143448071  2060  69634.9861           Root MSE      =  259.04

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
          iq |   3.285117   .3700311     8.88   0.000     2.559443    4.010792
       _cons |   278.0227   38.33661     7.25   0.000     202.8401    353.2052
------------------------------------------------------------------------------

.
end of do-file
```

The least squares regression line is: Y = 278.02 + 3.29X……… **WAGES** = 278.02 + 3.29**IQ**

- Before interpreting these numbers, you need to test for statistical significance.

  -- Just as we thought about the sampling distribution of sample means and other statistics, we can think about the sampling distribution of the coefficients in OLS regressions – i.e., we can test the coefficient estimates for statistical significance: is it likely that we see the relationship that we did between *X* and *Y* if in fact there were no true relationship?

  -- Run a *t* – test:

  H0: $\beta_{IQ} = 0$
  H1: $\beta_{IQ} \neq 0$

- the standard error of $\hat{\beta}_1$ is: $s_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2}}$ where $\hat{\sigma}^2 = \frac{1}{(n-2)} \sum e_i^2$

- d.f. = n-2
  = 2,061-2 = 2,059

  t-crit (95% confidence level) = ±1.96 t-crit (99.9% confidence level) = ±3.921

- and the test statistic is: $t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = \frac{3.28512 - 0}{0.37003} = 8.88$

- Make a conclusion: |8.88| > |±3.921|. Therefore *reject the null hypothesis* that there is no relationship between IQ and wages.

- The estimate of the relationship between IQ and wages is statistically significant. Is it substantively significant?

- ***If*** coefficient is statistically significant, then the standard interpretation of the slope of an OLS regression line is:

  For a one *X*-unit increase in the *X* variable, the *Y* variable is predicted to (increase / decrease) by $\hat{\beta}_1$ *Y*-units. E.g.:

  For each one-point increase in IQ, wages are expected to increase by 3.29 cents, on average.

- To interpret regression results in a meaningful way, you need to interpret them in the specific context of your study: the plain old vanilla interpretation won't cut it. To do this, you should have a firm grasp on:

  -- the units in which X is measured (here, IQ is measured in points)

  -- the units in which Y is measured (here, wages are measured in cents per hour in 1976)

  -- In interpreting the coefficient(s), try to step back and develop some sense of the magnitude or importance of the estimate.

V. MAKING PREDICTIONS FROM REGRESSION COEFFICIENTS

- So $Y = \hat{\beta}_0 + \hat{\beta}_1 X$ is the regression line that "best" fits the data: **WAGES** = 278.02 + 3.29**IQ**

- We can use this equation to make predictions about the wages, for a given IQ.

- Example: What would we predict wages to IQ=100? How does this compare with the predicted wages when IQ=120?

  Wages (case1) = 278.02 + 3.29***(100)** = 607.02 (*predicted)*
  Wages (case2) = 278.02 + 3.29***(120)** = 672.82 (*predicted)*

  So, we predict that a person with an IQ of 100 will have wages of 607.02 cents in 1976 (i.e., $6.07); and a person with an IQ of 120 will have wages of $6.73.

- In a later example, we'll revisit making predictions, establish the link to how this relates to *errors* in prediction, and talk about R-squared.

VI. ANOTHER EXAMPLE

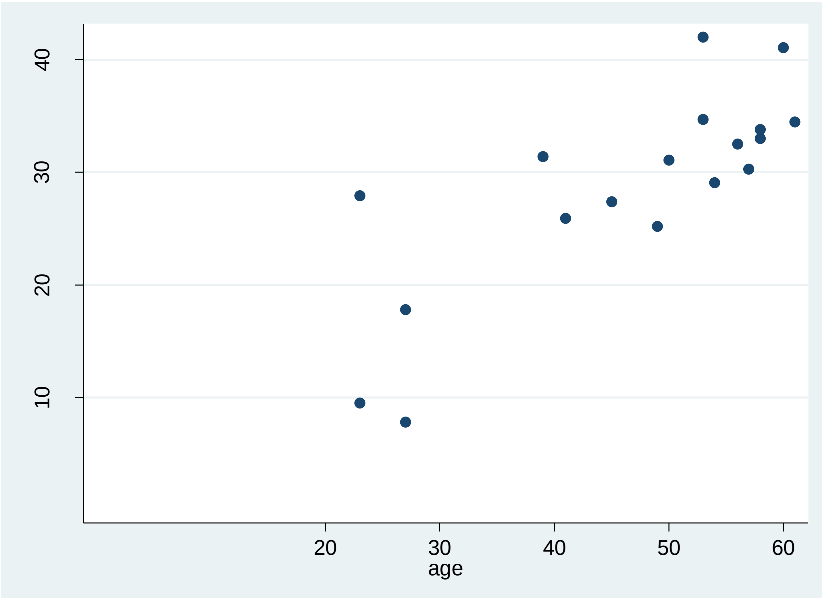
- In the paper, "Total Body Composition by Dual-Photon Absorptiometry" (Am. J. Clinical Nutrition, 1984), Mazess et al. studied methods for quantifying body composition. The body fat and ages of 18 randomly selected adults were recorded:

```
. flist
     |      age     pbodfat |
     |----------------------|
  1. |       23         9.5 |
  2. |       23        27.9 |
  3. |       27         7.8 |
  4. |       27        17.8 |
  5. |       39        31.4 |
     |----------------------|
  6. |       41        25.9 |
  7. |       45        27.4 |
  8. |       49        25.2 |
  9. |       50        31.1 |
 10. |       53        34.7 |
     |----------------------|
 11. |       53          42 |
 12. |       54        29.1 |
 13. |       56        32.5 |
 14. |       57        30.3 |
 15. |       58          33 |
     |----------------------|
 16. |       58        33.8 |
 17. |       60        41.1 |
 18. |       61        34.5 |
     +----------------------+
```

```
. summarize
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        18    46.33333    13.21764         23         61
     pbodfat |        18    28.61111     9.14439        7.8         42
```

Are age and percentage of body fat related? If so, what is the relationship?

```
scatter pbodfat age, ysc(r(0)) xsc(r(0))
```



- We can calculate the least squares regression line that best fits these data:

	A	B	C	D	E	F	
	pbodfat	age	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	C*D	$(X_i - \bar{X})^2$	
	9.5	23	-23.33	-19.11	445.93	544.44	
	27.9	23	-23.33	-0.71	16.59	544.44	
	7.8	27	-19.33	-20.81	402.35	373.78	
	17.8	27	-19.33	-10.81	209.01	373.78	
	31.4	39	-7.33	2.79	-20.45	53.78	
	25.9	41	-5.33	-2.71	14.46	28.44	
	27.4	45	-1.33	-1.21	1.61	1.78	
	25.2	49	2.67	-3.41	-9.10	7.11	
	31.1	50	3.67	2.49	9.13	13.44	
	34.7	53	6.67	6.09	40.59	44.44	
	42.0	53	6.67	13.39	89.26	44.44	
	29.1	54	7.67	0.49	3.75	58.78	
	32.5	56	9.67	3.89	37.59	93.44	
	30.3	57	10.67	1.69	18.01	113.78	
	33.0	58	11.67	4.39	51.20	136.11	
	33.8	58	11.67	5.19	60.54	136.11	
	41.1	60	13.67	12.49	170.68	186.78	
	34.5	61	14.67	5.89	86.37	215.11	
MEAN	28.6	46.3					
SUM					1627.53	2970.00	
Estimate of $\hat{\beta}_1$:							**0.54799**
Estimate of $\hat{\beta}_0$:							**3.22086**

- How does this compare with what STATA calculates?

```
. regress pbodfat age

      Source |       SS       df       MS              Number of obs =      18
-------------+------------------------------           F(  1,    16) =   26.94
       Model |  891.873652     1  891.873652           Prob > F      =  0.0001
    Residual |  529.664093    16  33.1040058           R-squared     =  0.6274
-------------+------------------------------           Adj R-squared =  0.6041
       Total |  1421.53775    17  83.6198674           Root MSE      =  5.7536

------------------------------------------------------------------------------
     pbodfat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .547991   .1055752     5.19   0.000     .3241815    .7718005
       _cons |    3.22086   5.076158     0.63   0.535    -7.540113    13.98183
```

Q1: If a person is zero years old, what is our best estimate of his/her body fat percentage?

Q2: Is there reason to be skeptical about your answer to Q1?

Q3: Interpret the coefficient on AGE (remember to test for statistical significance first!)

Q4: What percentage of body fat would we predict for a person who is 50 years old?

Q5: How far is this prediction from the *actual* body fat percentage for the person in the sample who is 50 years old?

- Why would we want to predict a person's body fat percentage if we know the actual value? Remember: we're in the business of making *inferences about a population of interest from the sample data available to us.* We can use the regression equation (i.e., the linear approximation of the relationship between age and body fat) to make predictions about other cases of interest:

Q6: What percentage of body fat do you predict for a person who is 34 years old? 54 years old?

Q7: What is the predicted difference in body fat for persons who are 20 years apart in age?

VII. R^2

- This idea of error will help motivate why we care about a thing called R^2. Think back to the idea of prediction in regression, and let's take one example:

  What percentage of body fat would I predict, using the regression equation, for a person who is 50 years old?

  3.22 + 0.548*(50) = 30.62 percent body fat

- How far away is this prediction from the *actual* percentage of body fat for the person in the sample who is 50 years old?:

  31.1-30.62 = 0.48. This is the value of the ***prediction error (e)***

  So, $Y = \hat{\beta}_0 + \hat{\beta}_1 X + e$: 31.1 = 3.22 + (0.548*50) + 0.48

  e = $Y - \hat{Y}$

  the residual or error of the regression model: the difference between the <u>actual value</u> of *Y* for that observation and the <u>predicted value</u> of *Y* for a particular observation. The distance between each point and the regression line is equal to the residual or the error for that observation – it is just a measure of how far away the regression line is from the point.

  The regression line does not touch *all* of the points in a scatterplot (if it did, it wouldn't be a straight line). Most of the points do not actually lie on the line.

- Knowing that predictions obtained from the regression line may not exactly correspond to the actual values (i.e., the line that we fit to the data is not an exact fit – it is a *linear approximation*), we need a measure of how well the line fits the data.

- The **R-squared statistic** tells us this. Sometimes it's called the "coefficient of determination"

- R-squared ranges from 0 to 1.00. Zero indicates no explanatory power; 1 indicates perfect fitting of the line to the data.

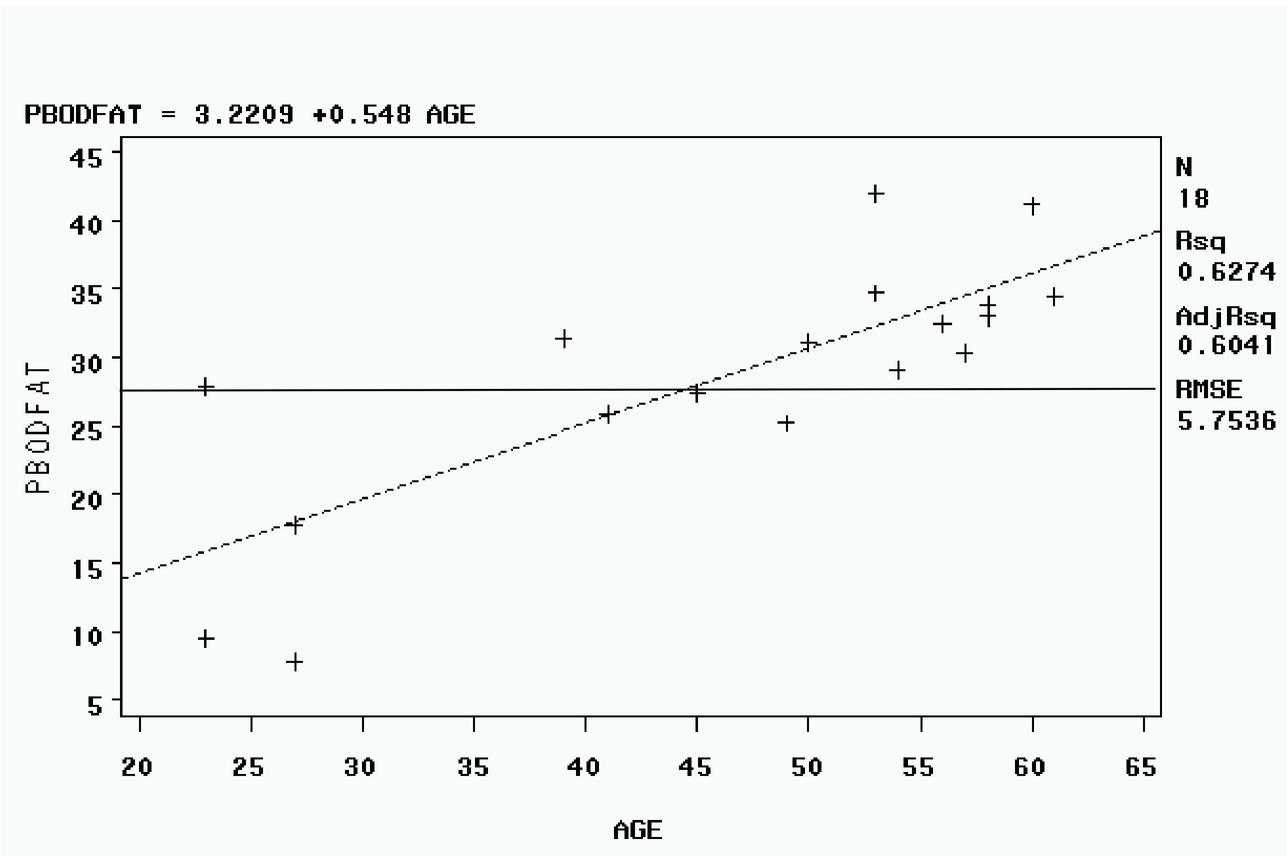
-- The whole idea behind R-squared is to give a single statistic that provides some indication of how well the estimated regression line fits the data. Another more precise way of saying this is that R-squared tells the proportion of the variation in the dependent variable, *Y*, that is explained by the independent variable(s), *X*.

$$R^2 = \frac{Error\ without\ Xs - Error\ with\ Xs}{Error\ without\ Xs} = \frac{\text{explained variation in } Y \text{ from model}}{\text{total variation in } Y} =$$

$$= \frac{(SST - SSR)}{SST} = \frac{\sum (Y_i - \bar{Y})^2 - \sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} =$$

$$= 1 - \frac{\text{unexplained variation}}{\text{total variation}} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

- Let's work through the R-squared stat and try to figure out why it makes any sense:
  - SST = Sum of Squares total = $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$ (i.e., the total variation in *Y*)
  - SSE = *Explained* sum of Squares = $\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$ (i.e., the variation in *Y* that is explained by the regression)
  - SSR = *Residual* Sum of Squares = $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ (i.e., the variation in *Y* that is not explained by the regression)



	A	B	D	E	F	G
			$(Y_i - \bar{Y})^2$	$\hat{Y}_i$	$(\hat{Y}_i - \bar{Y})^2$	$(Y_i - \hat{Y}_i)^2$
	pbodfat	age				
	9.5	23	365.23	15.82	163.49	40.00
	27.9	23	0.51	15.82	163.49	145.81
	7.8	27	433.10	18.02	112.24	104.38
	17.8	27	116.88	18.02	112.24	0.05
	31.4	39	7.78	24.59	16.15	46.34
	25.9	41	7.35	25.69	8.54	0.04
	27.4	45	1.47	27.88	0.53	0.23
	25.2	49	11.64	30.07	2.14	23.74
	31.1	50	6.19	30.62	4.04	0.23
	34.7	53	37.07	32.26	13.35	5.93
	42.0	53	179.26	32.26	13.35	94.78
	29.1	54	0.24	32.81	17.65	13.78
	32.5	56	15.12	33.91	28.06	1.98
	30.3	57	2.85	34.46	34.17	17.28
	33.0	58	19.26	35.00	40.87	4.02
	33.8	58	26.92	35.00	40.87	1.45
	41.1	60	155.97	36.10	56.09	25.00
	34.5	61	34.68	36.65	64.60	4.62
MEAN	28.6	46.3				
SUM			1421.54		891.87	529.66
Estimate of b1:						0.547991
Estimate of b0:						3.22086

- In the bodyfat -age regression, $R^2 = 0.6274$.

  With the linear approximation we estimated, we are able to explain about 63% of the variation in body fat percentage: about 63% of the variation in the body fat percentage is accounted for by the variation in age.

- In the bivariate regression case (i.e., one *X* variable and one *Y* variable), R^2 is equal to the square of the correlation coefficient. This is a special case!

Derivation of the OLS Estimator

Recall that the OLS criterion is: minimize the sum of squared residuals: $\min \sum_i e_i^2 = \min \sum_i [Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i]^2$. Using calculus, we can solve for the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that meet this criterion:

$$
\begin{aligned}
\frac{\partial[\sum_i (e_i^2)]}{\partial \hat{\beta}_0} = \frac{\partial[\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\partial \hat{\beta}_0} \quad &= \quad -2\sum (Y - \hat{\beta}_0 - \hat{\beta}_1 X) = 0 \\
&= \quad \sum (Y - \hat{\beta}_0 - \hat{\beta}_1 X) = 0 \\
&= \quad \sum Y - n\beta_0 - \hat{\beta}_1 \sum X = 0 \\
&= \quad \sum Y = n\beta_0 + \hat{\beta}_1 \sum X \\
&= \quad \overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X} \\
&= \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial[\sum_i (e_i^2)]}{\partial \hat{\beta}_1} = \frac{\partial[\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\partial \hat{\beta}_1} \quad &= \quad -2\sum X(Y - \hat{\beta}_0 - \hat{\beta}_1 X) = 0 \\
&= \quad \sum X(Y - \hat{\beta}_0 - \hat{\beta}_1 X) = 0 \\
&= \quad \sum XY - \hat{\beta}_0 \sum X - \hat{\beta}_1 \sum X^2 = 0 \\
&= \quad \sum XY = \hat{\beta}_0 \sum X + \hat{\beta}_1 \sum X^2 \\
&= \quad \sum XY = [\overline{Y} - \hat{\beta}_1 \overline{X}] \sum X + \hat{\beta}_1 \sum X^2 \\
&= \quad \sum XY = \overline{Y} \sum X - \hat{\beta}_1 \overline{X} \sum X + \hat{\beta}_1 \sum X^2 \\
&= \quad \sum XY - \overline{Y} \sum X = -\hat{\beta}_1 \overline{X} \sum X + \hat{\beta}_1 \sum X^2 \\
&= \quad \sum XY - \overline{Y} \sum X = \hat{\beta}_1 \left[\sum X^2 - \overline{X} \sum X\right] \\
&= \quad \frac{\sum XY - \overline{Y} \sum X}{\sum X^2 - \overline{X} \sum X} = \hat{\beta}_1 \\
&= \quad \ldots. \\
&= \quad \hat{\beta}_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}
\end{aligned}
$$