Correlation and Regression

Objectives:

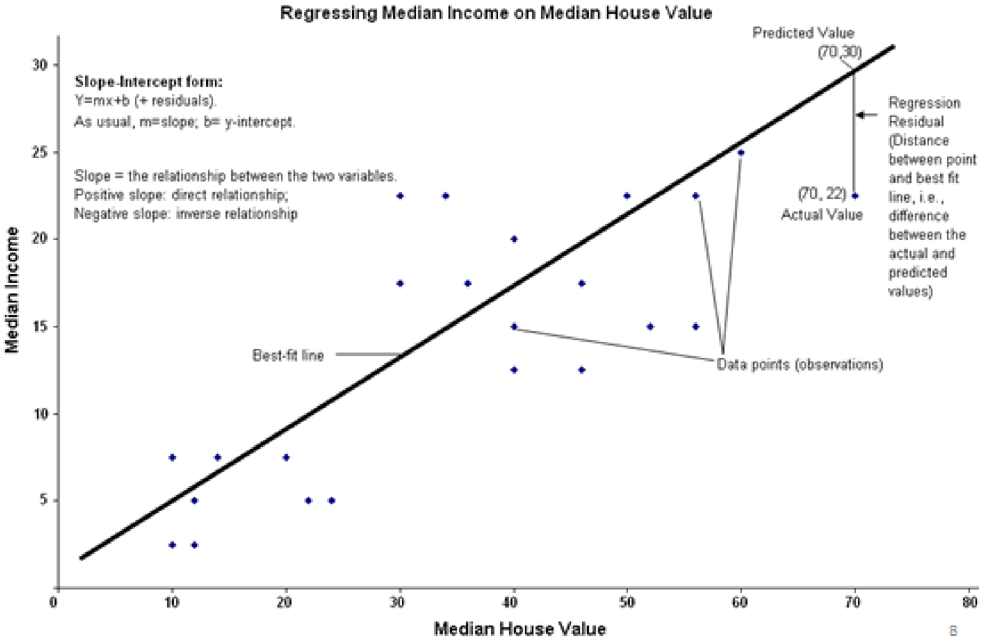
- Understand independent variable and dependent variable
- Calculate correlation coefficient
- Calculate regression parameters
- Evaluate regression model

Regression

- A statistical method used to examine the relationship between a variable of interest (dependent variable) and **one or more** explanatory variables (predictors)
    - Strength of the relationship
    - Direction of the relationship (positive, negative, zero)
    - Goodness of model fit
- Allows you to calculate the amount by which your dependent variable changes when a predictor variable changes by one unit (holding all other predictors constant)
- Often referred to as Ordinary Least Squares (OLS) regression
    - Regression with one predictor is called *simple regression*
    - Regression with two or more predictors is called *multiple regression*
- Just like correlation, if an explanatory variable is a significant predictor of the dependent variable, it doesn't imply that the explanatory variable is a *cause* of the dependent variable

Example

- Assume we have data on median income and median house value in 381 Philadelphia census tracts (i.e., our unit of measurement is a tract)
- Each of the 381 tracts has information on income (call it Y) and on house value (call it X). So, we can create a scatter-plot of Y against X



Correlation and Regression

What is the difference?

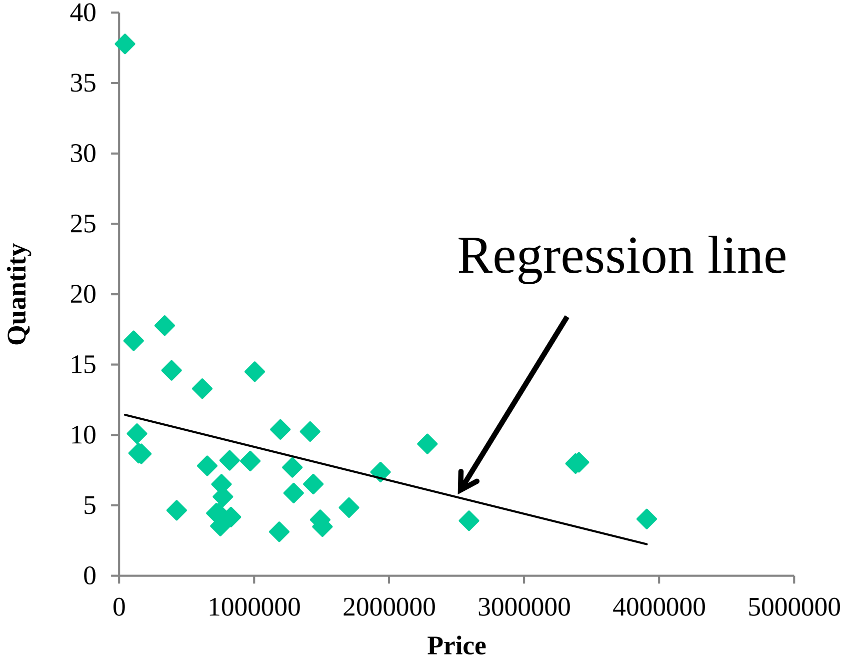
- Mathematically, they are identical.
- Conceptually, very different.

Correlation

- Co-variation
- Relationship or association
- No direction or causation is implied

Regression

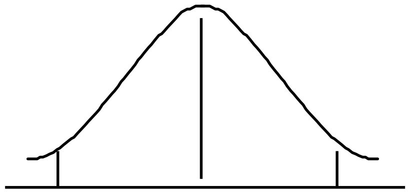
  - Prediction of Y from X
  - Implies, but does not prove, causation

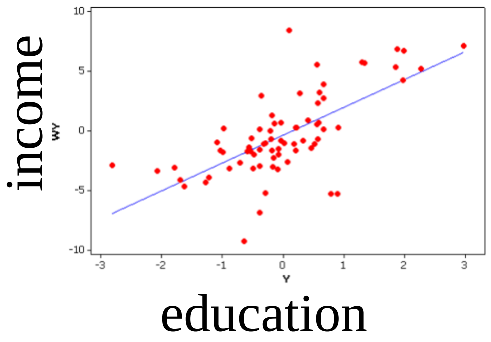


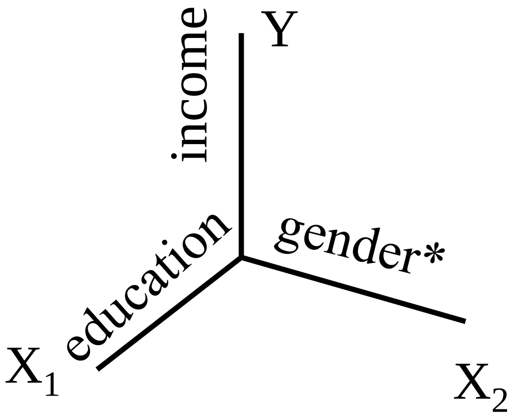
- X (independent variable) ⇨ Y (dependent variable)

Bivariate and Multivariate

- All measures so far have focused on one variable at a time
  - univariate
- Often, we are interested in the association or relationship between two variables
  - bivariate.
- Or more than two variables
  - multivariate







*Gender = male or female

Correlation Coefficient (r)

- Full name is the *Pearson Product Moment correlation coefficient*
- The most common statistic in all of science
- Measures the strength of the relationship (or "association") between two variables e.g. income and education
- Varies on a scale from –1 thru 0 to +1

+1 implies a perfect positive association

As values go up on one, they also go up on the other

- income and education

0 implies no association

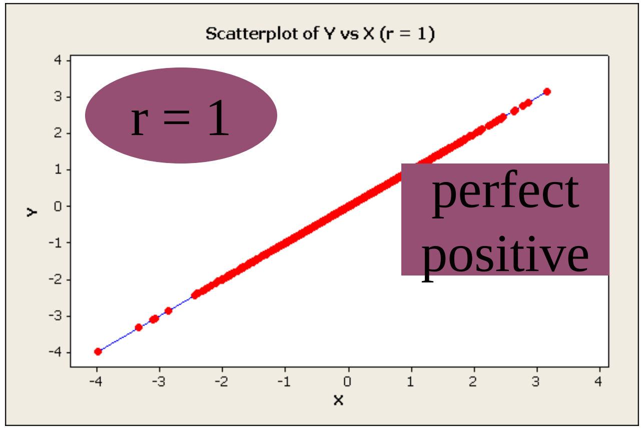
-1 implies perfect negative association

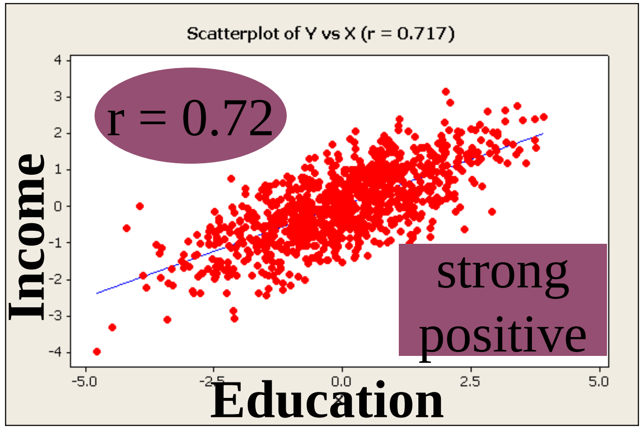
As values go up on one, they go down on the other

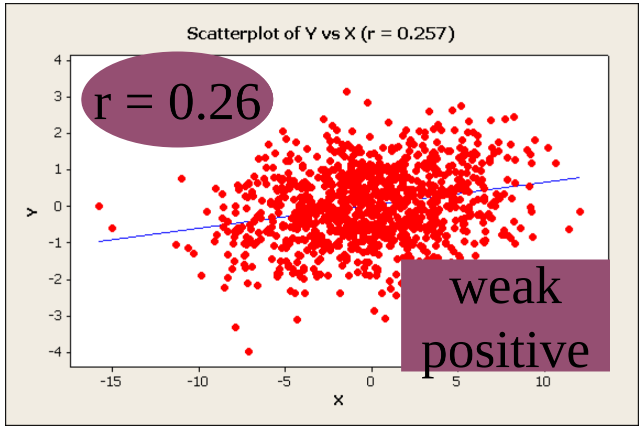
- price and quantity purchased

Examples of Scatter Diagrams and the Correlation Coefficient

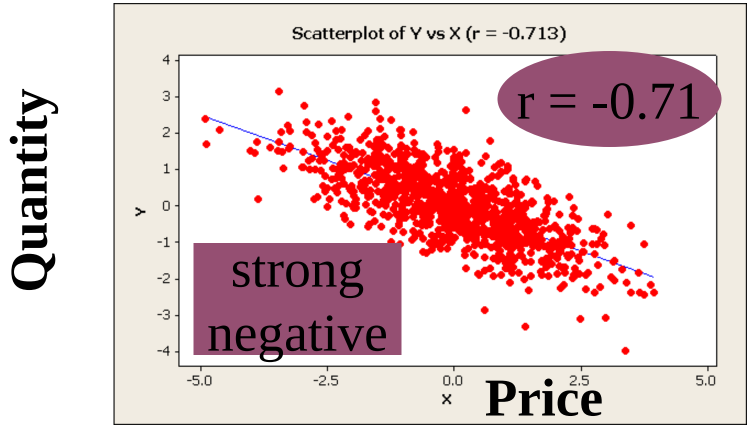
Positive

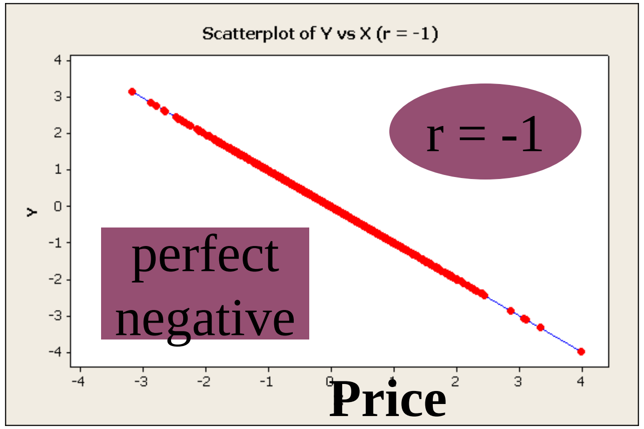






Negative





Pearson Product Moment Correlation Coefficient (r)

“product” is the result of a multiplication

$X * Y = P$

Moments about the mean

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y})}{n \quad S_x S_y}$$

$$S_y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{N}}$$

$$S_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{N}}$$

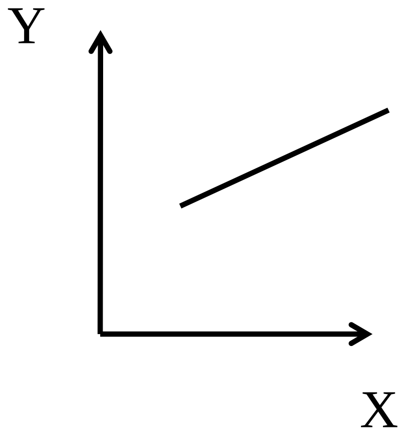
Where Sx and Sy are the standard deviations of X and Y, and $\overline{X}$ and $\overline{Y}$ are the means.

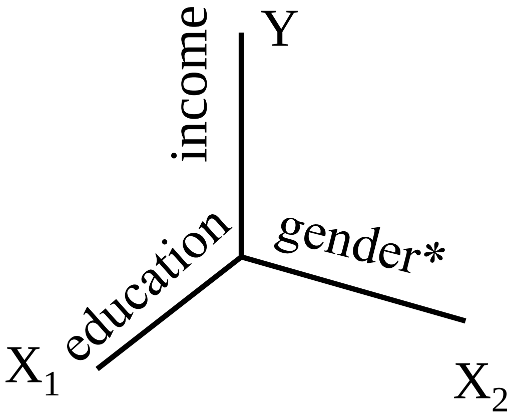
Predictors, Models, and Regression

- Prediction is at the core of building empirical models. We assume that there are drivers and effects. In other words, we want to build predictors of dependent variables *Y* based on the independent variables *X*. Regression techniques are the basis for many prediction methods. There are many types defined according to a variety of criteria and uses.
- The mathematical nature or structure of the predictor determines the type of method, such as linear regression vs. nonlinear regression
- The number of variables determines the dimensionality, simple regression vs. multiple regression;
- The nature of the variables and their explicit variation with time and space determines specific methods, such as Spatial autoregressive vs. autoregressive time series.

Regression

- Simple regression
  - Between two variables
    - One dependent variable (Y)
    - One independent variable (X)
- Multiple Regression
  - Between three or more variables
    - One dependent variable (Y)
    - Two or independent variable (X_1 ,$X_{2\ldots}$)





Simple Linear Least Squares Regression

- Let Y be a random variable defined as the "dependent" or "response" variable, and X another random variable defined as the "independent" or "factor" variable. Assume we have a joint sample x_i, y_i, $i = 1, \ldots, n$ or a set of n paired values of the two variables. This is a **bivariate** or two-variable situation. Denote by Y a **linear least squares** (LLS) estimator of Y from X

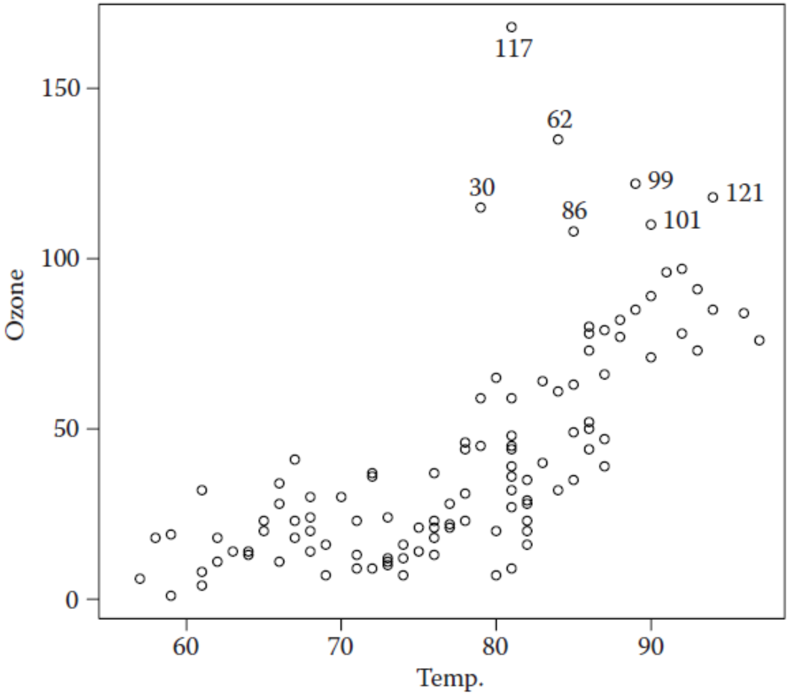


FIGURE 6.1 Scatter plot of x = air temperature and y = ozone concentration with identification of outliers.

$$Y = b_0 + b_1X+\varepsilon$$

Simple Linear Regression

- Concerned with “predicting” one variable (Y - the dependent variable) from another variable (X - the independent variable)

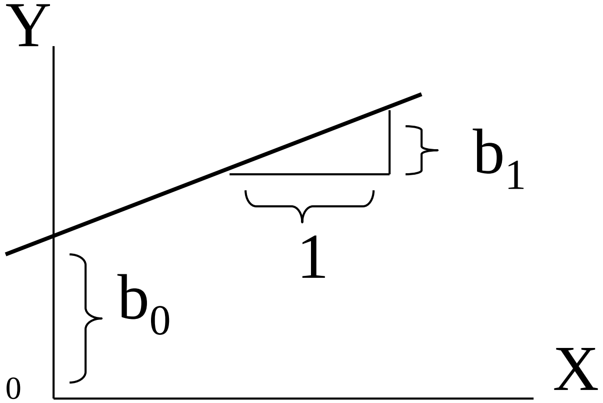
$$Y = b_0 + b_1X+\varepsilon$$

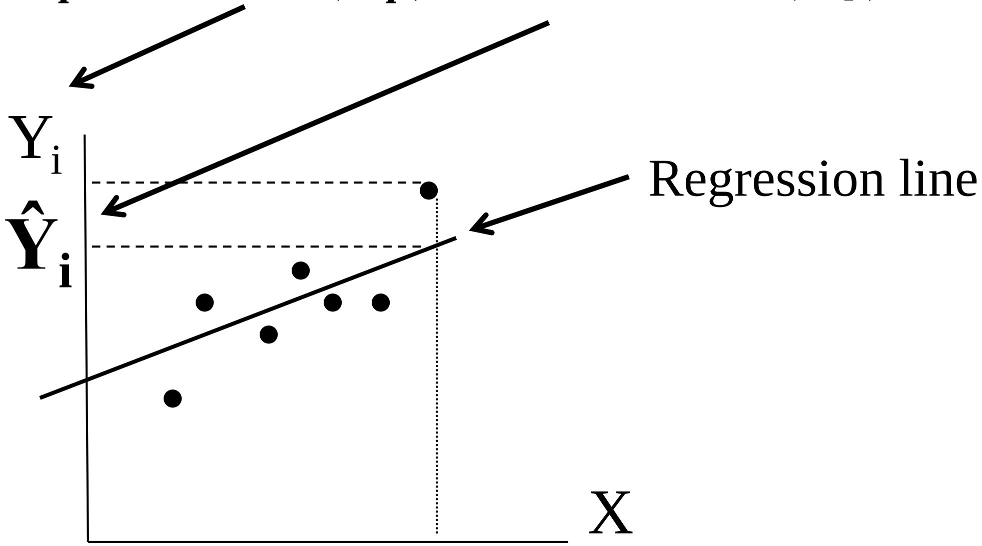
b_0 is the *intercept* —the value of Y when X =0

b_1 is the *regression coefficient* or slope of the line

—the change in Y for a <u>one</u> unit change in X

ε = residual= error = **Y_i-$\hat{Y}_i$ =Actual (Y_i) – Predicted ($\hat{Y}_i$)**





- This is the equation of a straight line with **intercept** *b*0 and **slope** *b*1. For each data point *i*, we have the **estimated** value of *Y* at the specific values *xi*

$$\widehat{y_i} = b_0 + b_1 x_i$$

- The **error** (**residual**) for data point *i* is,

ε

$$e_i = y_i - \widehat{y_i}$$

- And thus another way of writing the relationship of x_i and y_i observations is

$$y_i = b_0 + b_1 x_i + e_i$$

- Take the square and sum over all observations to obtain the **total squared error**

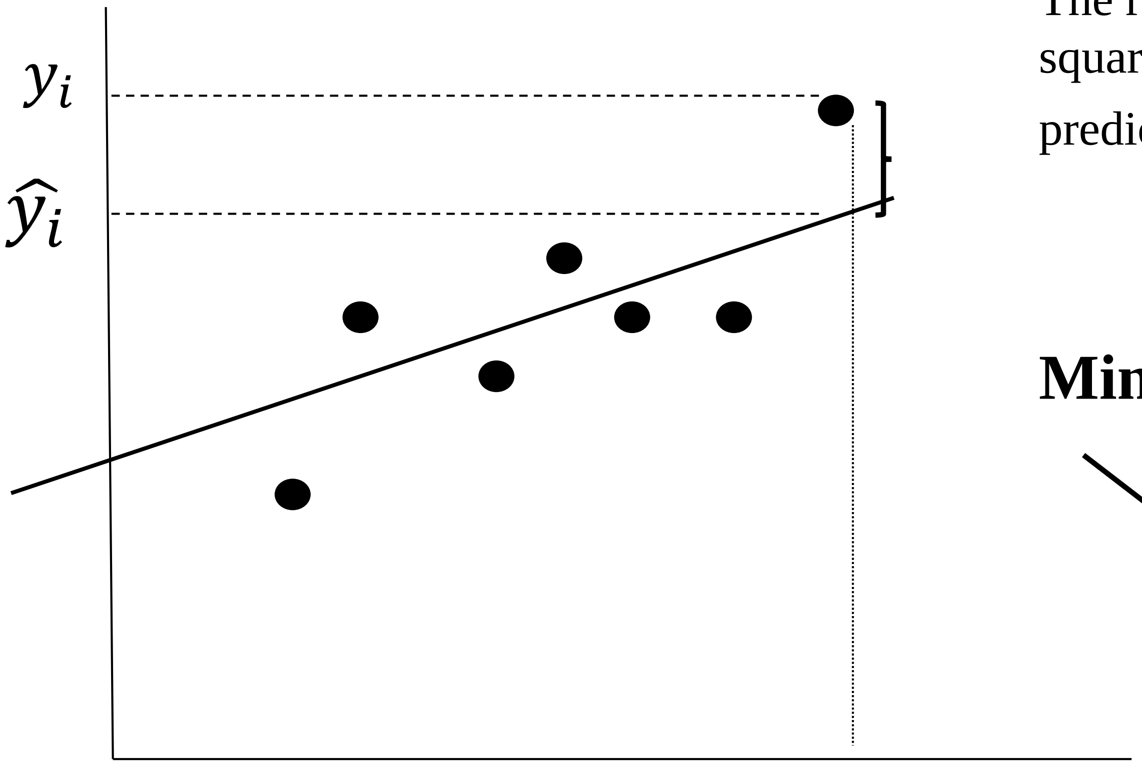
$$q = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

- We want to find the value of the **coefficients** (intercept and slope) b_0, b_1 which minimize the sum of squared errors (over all $i = 1, \ldots, n$). That is to say we want to find b_0, b_1 such that

$$\min_{b_0, b_1} q = \min_{b_0, b_1} \sum_{i=1}^{n} e_i^2 = \min_{b_0, b_1} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

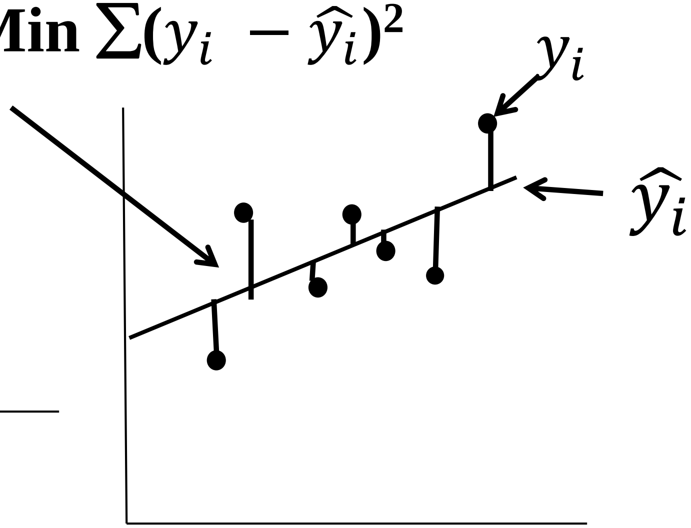
Linear Least Squares (LLS)

--the standard criteria for obtaining the regression line

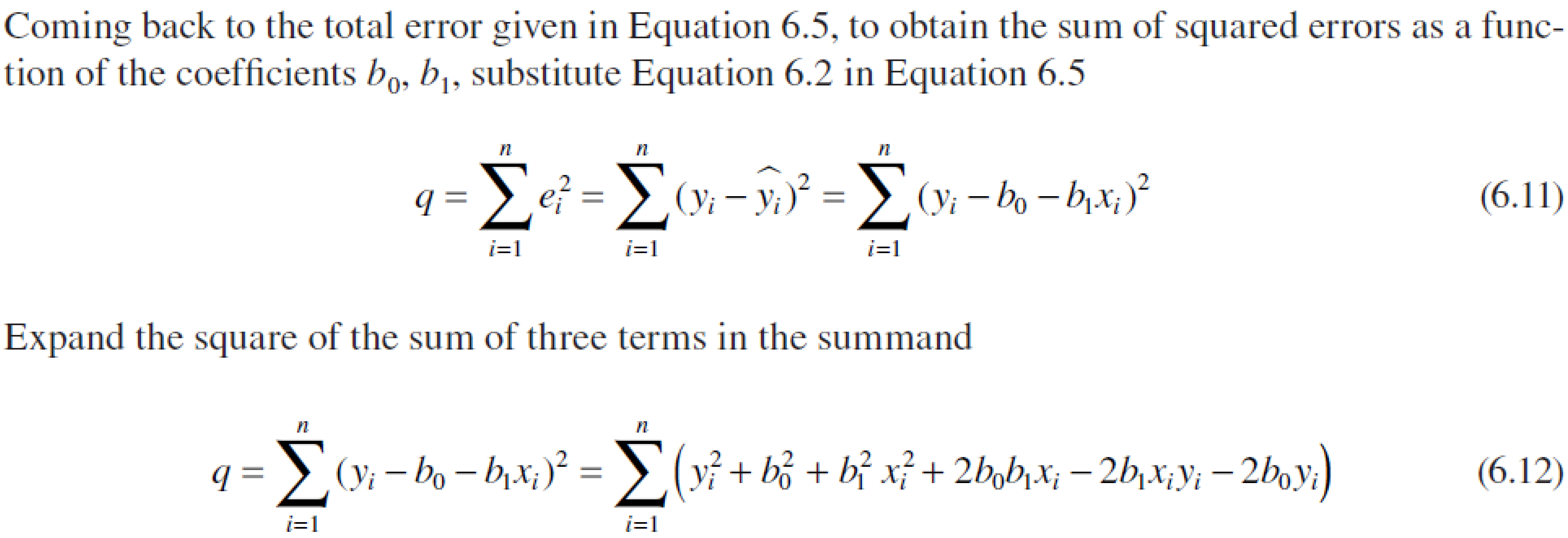


The regression line minimizes the sum of the squared deviations between actual y_i and predicted $\widehat{y_i}$

Min $\sum(y_i - \widehat{y_i})^2$



Calculating Regression Coefficients



Coming back to the total error given in Equation 6.5, to obtain the sum of squared errors as a function of the coefficients b_0, b_1, substitute Equation 6.2 in Equation 6.5

$$q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \tag{6.11}$$

Expand the square of the sum of three terms in the summand

$$q = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n} \left(y_i^2 + b_0^2 + b_1^2\, x_i^2 + 2b_0 b_1 x_i - 2b_1 x_i y_i - 2b_0 y_i \right) \tag{6.12}$$

And now we find partial derivatives of q with respect to b_0, b_1. Start with b_0

$$\frac{\partial q}{\partial b_0} = \frac{\partial}{\partial b_0}\sum_{i=1}^{n}\left(y_i^2 + b_0^2 + b_1^2 x_i^2 + 2b_0 b_1 x_i - 2b_1 x_i y_i - 2b_0 y_i\right)$$

$$= \sum_{i=1}^{n}(2b_0 + 2b_1 x_i - 2y_i) = 2\left(nb_0 + b_1\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i\right) \tag{6.13}$$

Set this derivative to zero

$$\frac{\partial q}{\partial b_0} = 2\left(nb_0 + b_1\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i\right) = 0 \tag{6.14}$$

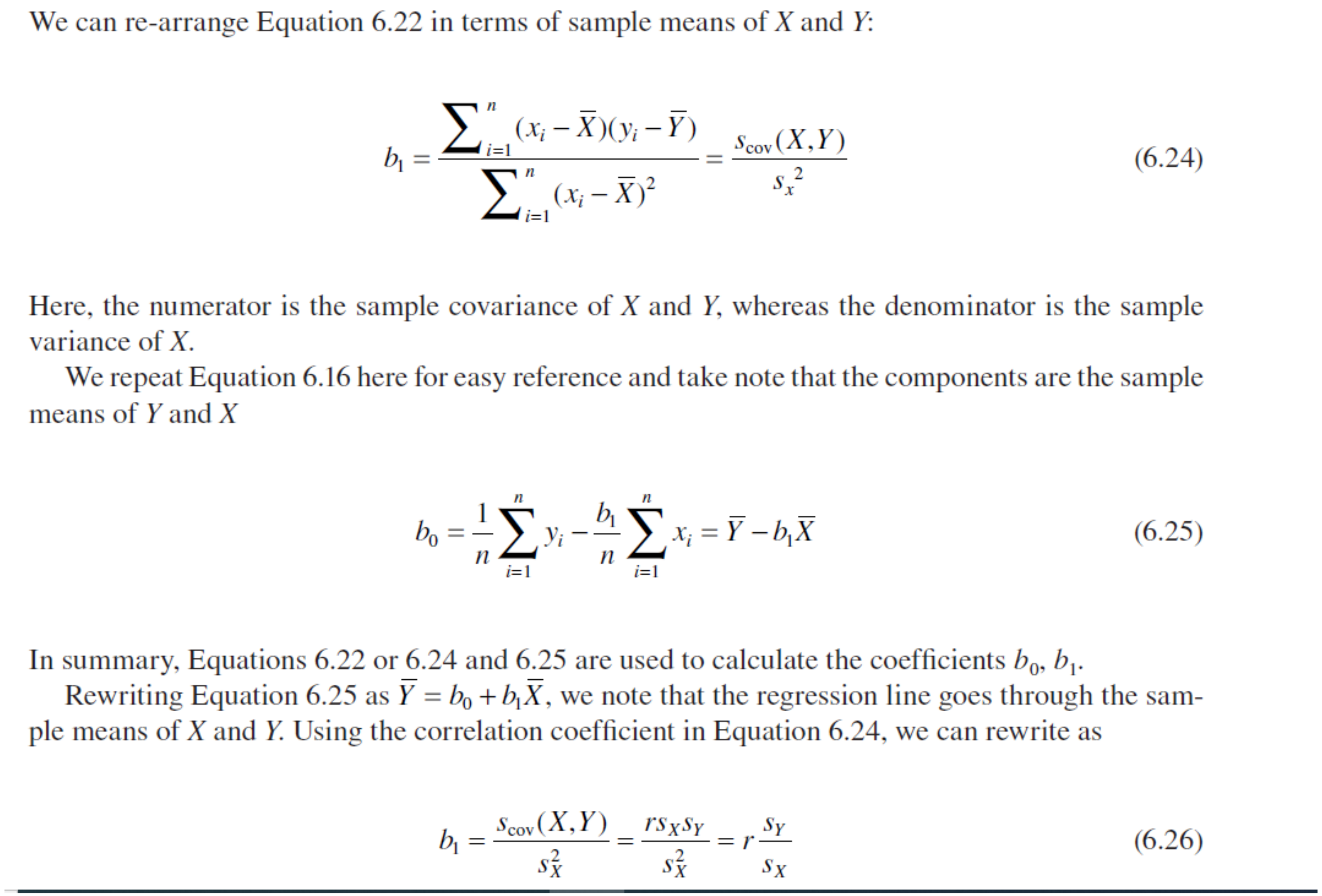
Finally solve for b_1

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \tag{6.22}$$

Once b_1 is calculated using Equation 6.22, then we can calculate b_0 using Equation 6.16 repeated now for easy reference

$$b_0 = -\frac{b_1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} y_i \tag{6.23}$$

Interpreting the Coefficients Using Sample Means, Variances, and Covariance



We can re-arrange Equation 6.22 in terms of sample means of X and Y:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2} = \frac{s_{\text{cov}}(X,Y)}{s_x{}^2} \qquad (6.24)$$

Here, the numerator is the sample covariance of X and Y, whereas the denominator is the sample variance of X.

We repeat Equation 6.16 here for easy reference and take note that the components are the sample means of Y and X

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{b_1}{n}\sum_{i=1}^{n} x_i = \bar{Y} - b_1\bar{X} \qquad (6.25)$$

In summary, Equations 6.22 or 6.24 and 6.25 are used to calculate the coefficients b_0, b_1.

Rewriting Equation 6.25 as $\bar{Y} = b_0 + b_1\bar{X}$, we note that the regression line goes through the sample means of X and Y. Using the correlation coefficient in Equation 6.24, we can rewrite as

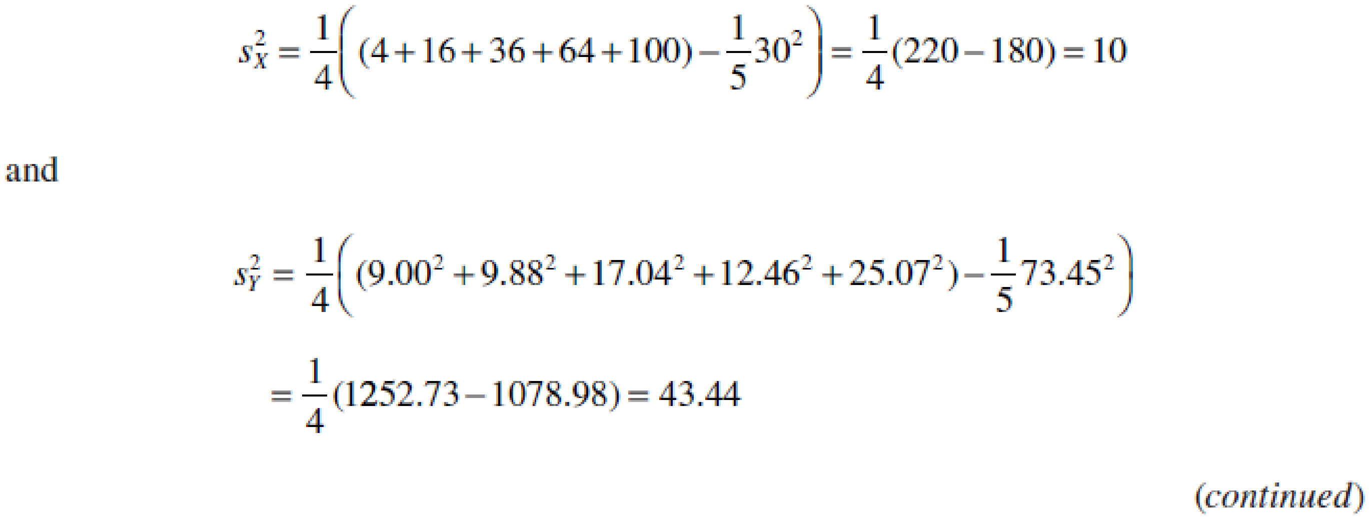
$$b_1 = \frac{s_{\text{cov}}(X,Y)}{s_X^2} = \frac{r s_X s_Y}{s_X^2} = r\frac{s_Y}{s_X} \qquad (6.26)$$

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_0 = -\frac{b_1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} y_i$$

Example

- Let us work out a numerical example with just a few values to illustrate the ideas.
- Suppose we have *n* = 5 pairs of values for *X* and *Y*: (2, 9.00), (4, 9.88), (6, 17.04), (8, 12.46), (10, 25.07). Calculate the sample means *X* = (2 + 4 + 6 + 8 +10)/5 = 6 and *Y* = (9.00 + 9.88 +17.04 +12.46 + 25.07)/5 = 14.69. Calculate the sample variances



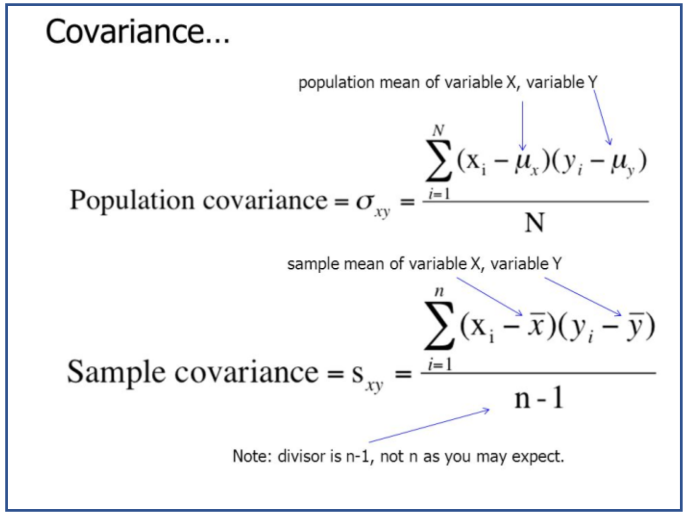
$$s_X^2 = \frac{1}{4}\left((4+16+36+64+100)-\frac{1}{5}30^2\right)=\frac{1}{4}(220-180)=10$$

and

$$s_Y^2 = \frac{1}{4}\left((9.00^2+9.88^2+17.04^2+12.46^2+25.07^2)-\frac{1}{5}73.45^2\right)$$

$$=\frac{1}{4}(1252.73-1078.98)=43.44$$

(continued)



Covariance…

population mean of variable X, variable Y

$$\text{Population covariance} = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

sample mean of variable X, variable Y

$$\text{Sample covariance} = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Note: divisor is n-1, not n as you may expect.

$$cov_{XY} = \sigma_{XY} = \mathrm{E}[XY] - \mu_X \mu_Y$$

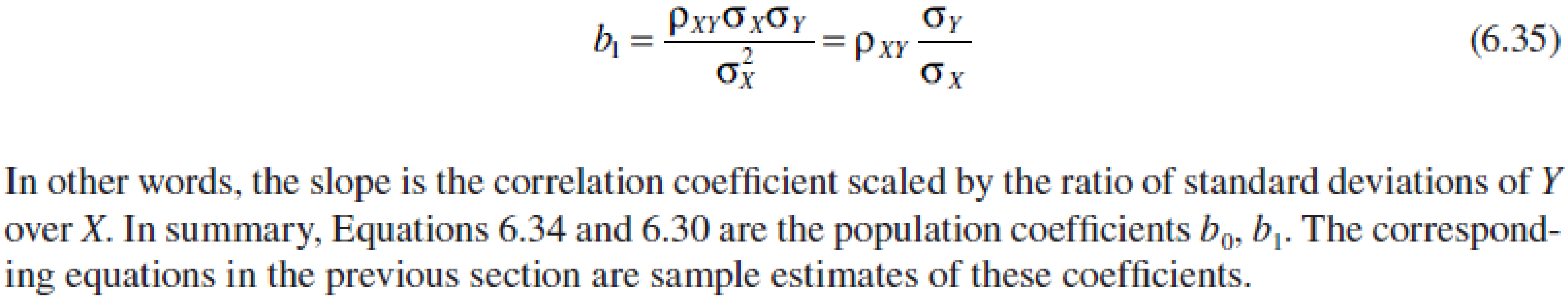
$$S_{xy} = \frac{(\sum_1^n x_i y_i - n\bar{x}\bar{y})}{n-1} = \frac{(\sum_1^n x_i y_i - sum(x)sum(y)/n)}{n-1}$$

Then the sample standard deviations are $s_X = \sqrt{10} = 3.16$ and $s_Y = \sqrt{43.44} = 6.59$. To get the sample covariance of X and Y, first calculate the sum of products $s_{XY} = 2\times 9 + 4\times 9.88 + 6\times 17.04 + 8\times 12.46 + 10\times 25.07) = 510.14$, and then subtract the product of the sums

$$s_{\text{cov}}(X,Y) = \frac{1}{4}\left(510.14 - \frac{1}{5}(30\times 73.45)\right) = 17.32$$

The slope is $b_1 = s_{\text{cov}}(X,Y)/s_X^2 = 17.32/10 = 1.732$ and the intercept is $b_0 = \overline{Y} - b_1\overline{X} = 14.69 - 1.73\times 6 = 4.31$. The equation for the linear predictor is Y = 4.31 + 1.73X. Apply this equation to one of the x values as an illustration, say x_2 = 4, $y_2 = 4.31 + 1.73\times 4 = 11.23$ the squared error for this value of y is $e_2 = (9.88 - 11.23)^2 = 1.82$.

Regression Coefficients from Expected Values



$$b_1 = \frac{\rho_{XY}\sigma_X\sigma_Y}{\sigma_X^2} = \rho_{XY}\,\frac{\sigma_Y}{\sigma_X} \qquad (6.35)$$

In other words, the slope is the correlation coefficient scaled by the ratio of standard deviations of Y over X. In summary, Equations 6.34 and 6.30 are the population coefficients b_0, b_1. The corresponding equations in the previous section are sample estimates of these coefficients.

Interpretation of the Error Terms

There are three important error terms for *Y* in the regression:

- the residual $e_i = y_i - \hat{y}$i
- the difference with respect to the mean $y_i - \bar{Y}$
- the error of the estimate with respect to the mean $\hat{y}_{\mathrm{i}} - \bar{Y}$.

We can relate them by formulating an identity in the following manner:

$$(y_i - \overline{Y}) = (y_i - \widehat{y_i}) + (\widehat{y_i} - \overline{Y})$$

And then sum over all observations, we get

$$\sum(y_i-\overline{Y})^2=\sum(y_i-\widehat{y_i})^2+\sum(\widehat{y_i}-\overline{Y})^2+2\sum(y_i-\widehat{y_i})(\widehat{y_i}-\overline{Y}) \tag{6.38}$$

The cross-product term is zero. Therefore,

$$\sum(y_i-\overline{Y})^2=\sum(y_i-\widehat{y_i})^2+\sum(\widehat{y_i}-\overline{Y})^2 \tag{6.39}$$

In the following, SS denotes "sum of squares." The "total error" SS_T in Y is the sum of squared differences of observations minus the mean

$$SS_T=\sum(y_i-\overline{Y})^2 \tag{6.40}$$

The "residual" or "unexplained" error SS_E is the sum of the squares of the difference between observations and estimated values

$$SS_E=\sum(y_i-\widehat{y_i})^2 \tag{6.41}$$

The “model” or “explained” error SS_M is the sum of squared differences of estimated points minus the mean

$$SS_M = \sum (\widehat{y_i} - \overline{Y})^2 \tag{6.42}$$

Now using Equation 6.39 we write that total error is the sum of model error and the residual, that is to say

$$SS_T = SS_E + SS_M \tag{6.43}$$

And equivalently the residual is the total error minus model error

$$SS_E = SS_T - SS_M \tag{6.44}$$

Please note that when these quantities are divided by n, the number of observations, then we get the mean squares (MS). The $MS_T = s_Y^2$ sample variance of Y, for SS_M the average or mean squared model error is MS_M, and for SS_E the average or mean squared residual error (MS_E).

$$s_Y^2 = MS_E + MS_M \tag{6.45}$$

A common measure of goodness of fit is R^2, which is the ratio of the model error to the total error

$$R^2 = \frac{MS_M}{MS_T} = \frac{MS_M}{MS_E + MS_M} = \frac{MS_M}{s_Y^2} \tag{6.46}$$

When $\mathrm{MS_E}$ (which is minimized by the least squares procedure) is very small, then R^2 approaches 1. Note that R^2 is the fraction (or percent) of variance of Y explained by the regression model. Note that

$$1 - R^2 = \frac{MS_E}{MS_T} = \frac{MS_E}{MS_E + MS_M} = \frac{MS_E}{s_Y^2} \tag{6.47}$$

The difference of R^2 with respect to 1 is the fraction of variance of Y unexplained by the regression.

In addition, it is important to realize that

$$R^2 = \frac{MS_M}{s_Y^2} = \frac{(1/n)\sum\left(b_0 + b_1X - \bar{Y}\right)^2}{s_Y^2} = \frac{(1/n)\sum\left(\bar{Y} - b_1\bar{X} + b_1X - \bar{Y}\right)^2}{s_Y^2} \tag{6.48}$$

The sample mean of Y cancels, b_1 is a common factor, and recognizing the sample variance of X we obtain

$$R^2 = \frac{(1/n)\sum b_1^2(X - \bar{X})^2}{s_Y^2} = \frac{b_1^2(1/n)\sum(X - \bar{X})^2}{s_Y^2} = \frac{b_1{}^2 s_X^2}{s_Y^2} \tag{6.49}$$

And by recalling the expression for b_1

$$R^2 = \frac{\left(\frac{s_{\text{cov}}(X,Y)}{s_X^2}\right)^2 s_X^2}{s_Y^2}$$

Therefore,

$$R^2 = \left(\frac{s_{\text{cov}}(X,Y)}{s_X s_Y}\right)^2 = r^2$$

The square root of R^2 is equal to r, which is the correlation coefficient.

As a numerical example, let us use the data of the example given earlier. Suppose we have $n = 5$ pairs of values for x and y: (2, 9.00), (4, 9.88), (6, 17.04), (8, 12.46), (10, 25.07). Recall the sample mean of y is $\overline{Y} = 14.69$. The total error is

$$SS_T = (9.00 - 14.69)^2 + (9.88 - 14.69)^2 + (17.04 - 14.69)^2$$
$$+ (12.46 - 14.69)^2 + (25.07 - 14.69)^2 = 173.75$$

Then calculate all the predicted values

$$y_1 = 4.31 + 1.73 \times 2 = 7.77$$

$$y_2 = 4.31 + 1.73 \times 4 = 11.23$$

$$y_3 = 4.31 + 1.73 \times 6 = 14.69$$

$$y_4 = 4.31 + 1.73 \times 8 = 18.15$$

$$y_5 = 4.31 + 1.73 \times 10 = 21.61$$

The total residual error is

$$SS_E = (9.00 - 7.77)^2 + (9.88 - 11.23)^2 + (17.04 - 14.69)^2$$
$$+ (12.46 - 18.15)^2 + (25.07 - 21.61)^2 = 53.20$$

The explained error is $SS_M = SS_T - SS_E = 173.75 - 53.20 = 120.55$. Now $R^2 = MS_M / MS_T =$ $120.55/173.75 = 0.69$, which should be equivalent to

$$R^2 = \left(\frac{s_{\text{cov}}(X,Y)}{s_X s_Y} \right)^2 = \left(\frac{17.32}{3.16 \times 6.59} \right)^2 = 0.832^2 = 0.69$$

The correlation coefficient is $r = 0.832$.

Evaluating Regression Models

- Examine how sound is the assumption of **linearity**. We do this by looking at the graph of *Y* vs. *X* (scatter diagrams). If the *yi* points seem to follow a definite nonstraight pattern or curve, then linearity is suspicious even when getting a good R^2
- Look at the statistical significance of the **slope**, or relative magnitude of the MST and the MSE. This we do using ANOVA to test whether the slope is zero.
- Evaluate the **residual error** using residual diagnostic plots
- Examine the **significance** of each coefficient *b*0 and *b*1 being different from zero using *t*-tests.
- Calculate the **residual standard error** or standard deviation of the residuals

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

- Calculate **confidence intervals** to express the reliability of the estimates given by the regression and of the regression coefficients

Number of observations minus *degrees of freedom*
(for simple regression, degrees of freedom = 2)

$$\boxed{s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-k}}}$$

Coefficients of Determination and Nondetermination

- The strength of a *correlation* is often interpreted by calculating the **coefficient of determination**, which is the proportion of variance that two variables share in common.
- The coefficient of determination is computed by squaring r (i.e., r^2).
- If there is a correlation between two variables, then the scores tend to change together in predictable ways.
- This is the shared common variance between the scores which is assumed to be influenced by the same factors (general intelligence, etc.)
- For example, if there is a correlation of $r = .72$ between students' grades in History and English, the coefficient of determination is $r^2 = .52$.
- In other words, 52% of the variability between these grades is shared common variance.

- The **coefficient of nondetermination** tells us the proportion of the variability that is not common variance and that is related to factors that do not influence both sets of scores (different assignments, etc.)

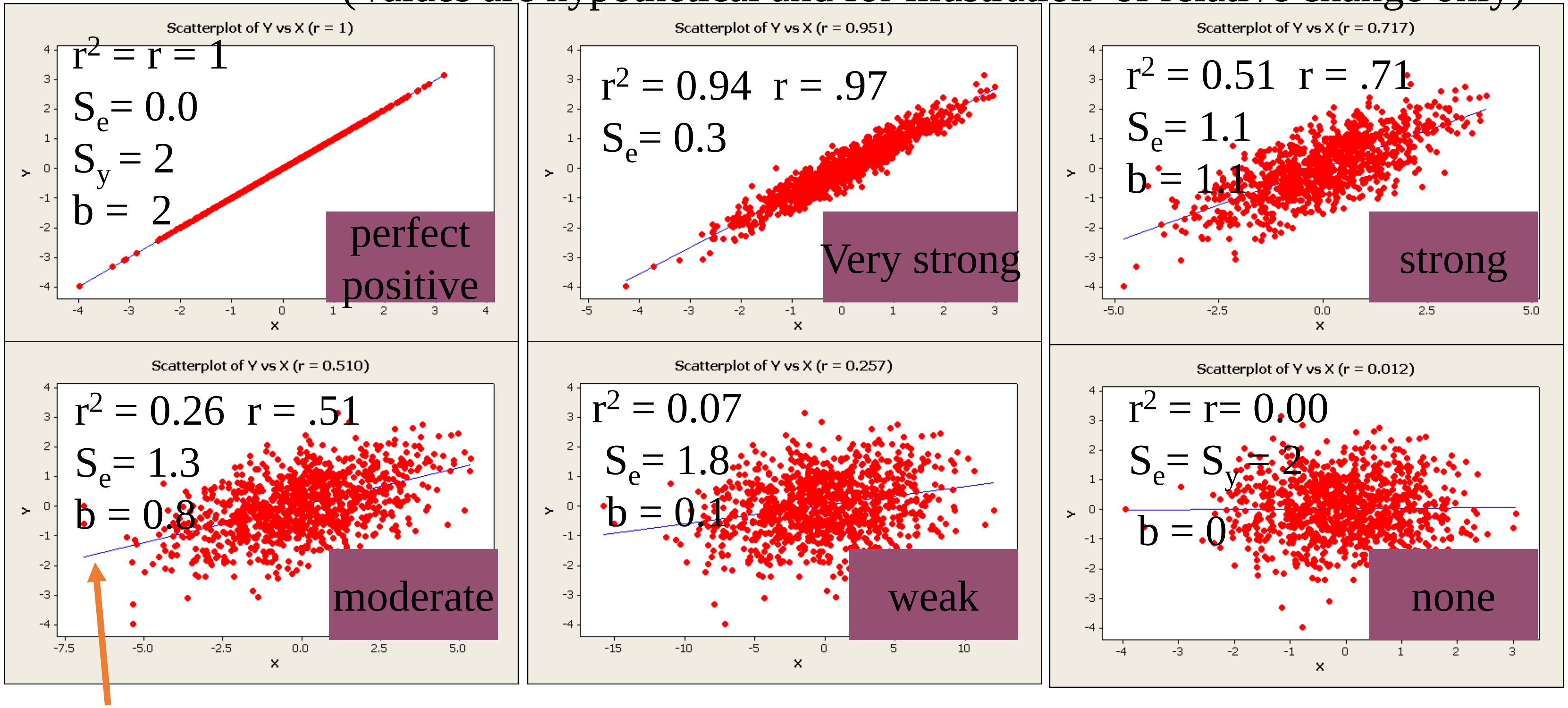
- The coefficient of nondetermination is calculated by subtracting r^2 from 1.

- For the above example, $1 - r^2 = .48$.

- What this value tells us is that 48% of the variance of one variable is *not* explained by the variance of the other.

Coefficient of determination (r^2), correlation coefficient (r), regression coefficient (b), and standard error (S_e)

(Values are hypothetical and for illustration of relative change only)



$r^2 = r = 1$

$S_e = 0.0$

$S_y = 2$

$b = 2$

perfect positive

$r^2 = 0.94$ $r = .97$

$S_e = 0.3$

Very strong

$r^2 = 0.51$ $r = .71$

$S_e = 1.1$

$b = 1.1$

strong

$r^2 = 0.26$ $r = .51$

$S_e = 1.3$

$b = 0.8$

moderate

$r^2 = 0.07$

$S_e = 1.8$

$b = 0.1$

weak

$r^2 = r = 0.00$

$S_e = S_y = 2$

$b = 0$

none

Regression line in blue

As the coefficient of determination gets smaller, the slope of the regression line (b) gets closer to zero.

As the coefficient of determination gets smaller, the standard error gets larger, and closer to the standard deviation of the dependent variable (Y) ($S_y = 2$)

Sample Statistics, Population Parameters and Statistical Significance tests

$Y_i = a + bX_i + \varepsilon_i$ a and b are *sample statistics*

which are estimates of *population parameters* α and β

β (and b) measure the change in Y for a one unit change in X. If $\beta = 0$ then X has no effect on Y, therefore

Null Hypothesis (H_0)**:** in the population $\beta = 0$

Alternative Hypothesis (H_1)**:** in the population $\beta \neq 0$

Thus, we test if our sample regression coefficient, b, is sufficiently different from zero to reject the Null Hypothesis and conclude that X has a statistically significant affect on Y

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Test Statistics in Simple Regression

$$t_{obt} = \frac{M - \mu}{s_M}$$

$$z = \frac{X - \mu}{\sigma}$$

Test statistic for b is distributed according to the *Student's t Distribution* (similar to normal): where s_e^2 is the variance of the estimate, with degrees of freedom = n – 2

$$t = \frac{b}{\mathrm{SE(b)}} = \frac{b}{\sqrt{\frac{s_e^2}{\sum_i (X - \overline{X})^2}}}$$

β = 0

If you need to calculate the standard error of the slope (SE) by hand, use the following formula:

$$\mathrm{SE} = s_{b1} = \mathrm{sqrt}\ [\ \Sigma(y_i - \hat{y}_i)^2 / (n - 2)\ ] / \mathrm{sqrt}\ [\ \Sigma(x_i - x)^2\ ]$$

where y_i is the value of the dependent variable for observation *i*, $\hat{y}_i$ is estimated value of the dependent variable for observation *i*, x_i is the observed value of the independent variable for observation *i*, x is the mean of the independent variable, and n is the number of observations.

Test Statistics in Simple Regression

A test can also be conducted on the *coefficient of determination* (R^2) to test if it is significantly greater than zero, using the *F* frequency distribution.

It is mathematically identical to the t test.

$$R^2 = \frac{(1/n)\sum b_1^2 (X-\bar{X})^2}{s_Y^2} = \frac{b_1^2 (1/n)\sum (X-\bar{X})^2}{s_Y^2} = \frac{b_1{}^2 s_X^2}{s_Y^2}$$

$$\text{F} = \frac{\text{Variance between groups}}{\text{Variance within groups}} = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}}$$

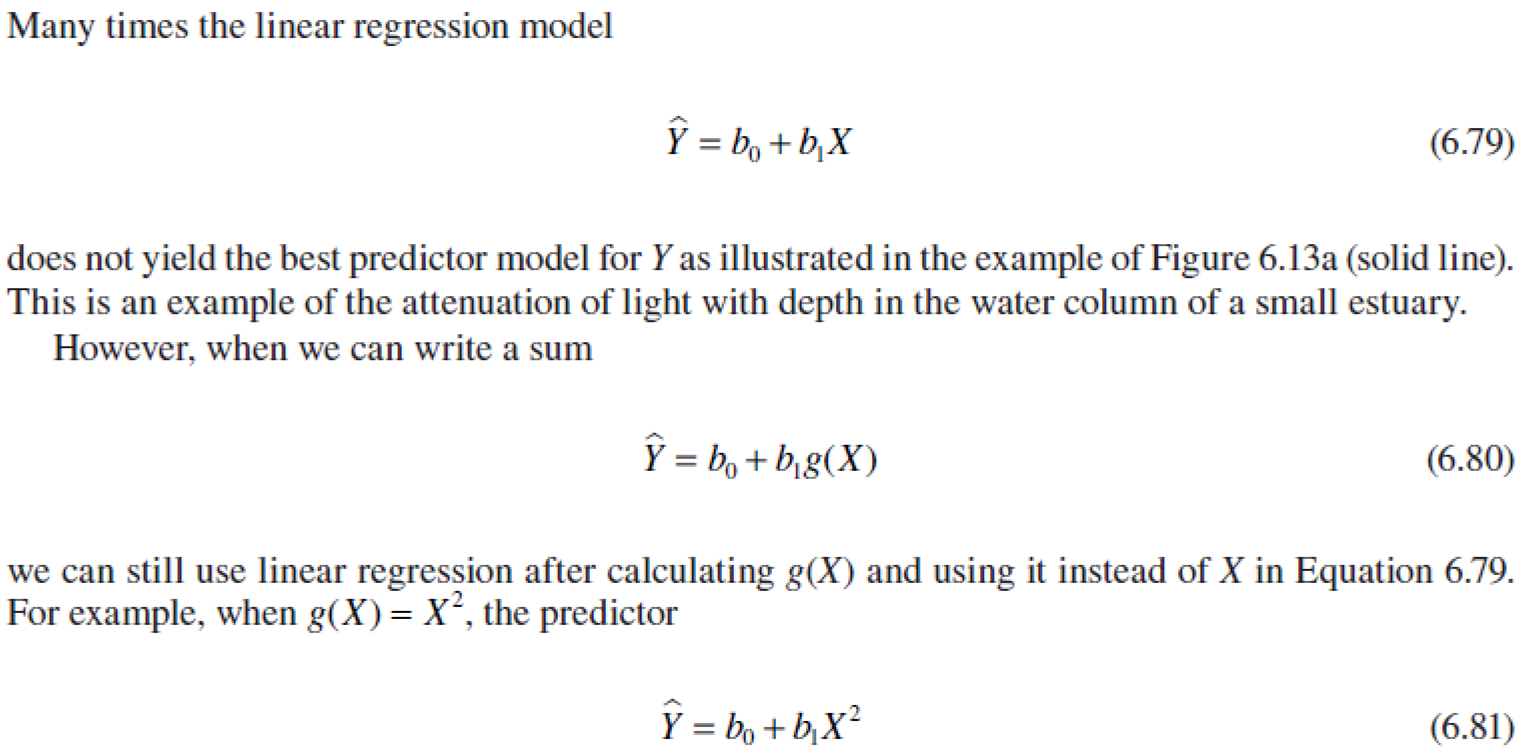
$$\text{MS}_{\text{within}} = \frac{\text{SS within}}{\text{cn - c}}$$

$$\text{MS}_{\text{between}} = \frac{\text{SS between}}{\text{c - 1}}$$

$$\text{F} = \frac{\text{MS between}}{\text{MS within}}$$

$$F = \frac{\text{Regression S.S./d.f.}}{\text{Residual S.S./d.f.}} = \frac{\sum(\hat{y}_i - \bar{Y})^2\ /1}{\sum(y_i\ - \hat{y}_i)^2\ /n-2}$$

Nonlinear Regression



Many times the linear regression model

$$\hat{Y} = b_0 + b_1X \qquad (6.79)$$

does not yield the best predictor model for *Y* as illustrated in the example of Figure 6.13a (solid line). This is an example of the attenuation of light with depth in the water column of a small estuary.

However, when we can write a sum

$$\hat{Y} = b_0 + b_1g(X) \qquad (6.80)$$

we can still use linear regression after calculating $g(X)$ and using it instead of X in Equation 6.79. For example, when $g(X) = X^2$, the predictor

$$\hat{Y} = b_0 + b_1X^2 \qquad (6.81)$$

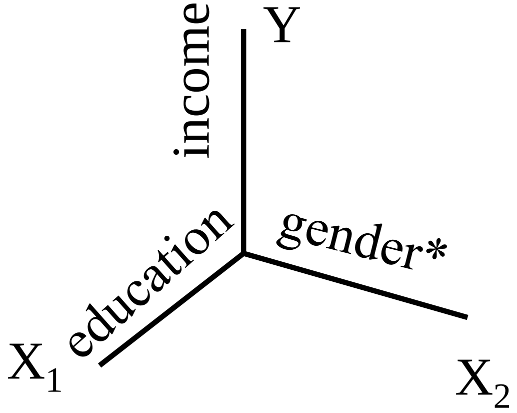
$$\min_{p} q = \min_{p} \sum_{i=1}^{n} e_i^2 = \min_{p} \sum_{i=1}^{n} \left(y_i - f(x_i, p) \right)^2$$

$$\widehat{Y} = b_0 + b_1 X + b_2 X^2 + \cdots + b_m X^m$$

Multiple regression

Simple regression as:

$$Y = b_0 + b_1 X + \varepsilon$$



Multiple regression: Y is predicted from 2 or more independent variables

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m + \varepsilon$$

b_0 is the *intercept* —the value of Y when values of all $X_j = 0$

$b_1 \ldots b_m$ are *partial regression coefficients* which give the change in Y for a one unit change in X_j, all other X variables held constant

m is the number of independent variables

Multiple Linear Regression

$$\widehat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_m X_m$$

$$\widehat{y_i} = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_m x_{mi}$$

$$\min_{\mathbf{b}} q = \min_{\mathbf{b}} \sum_{i=1}^{n} e_i^2 = \min_{\mathbf{b}} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \ldots \\ y_n \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{m1} \\ 1 & x_{12} & x_{22} & \ldots & x_{m2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{mn} \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \ldots \\ b_m \end{bmatrix}$$

In general, for *m* independent variables we get the solution in the special case as

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2 - \cdots - b_m \overline{X}_m \\ \dfrac{s_{\mathrm{cov}(X_1,Y)}}{s_{X_1}^2} \\ \dfrac{s_{\mathrm{cov}(X_2,Y)}}{s_{X_2}^2} \\ \vdots \\ \dfrac{s_{\mathrm{cov}(X_m,Y)}}{s_{X_m}^2} \end{bmatrix}$$

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \mu_Y - b_1 \mu_{X_1} - b_2 \mu_{X_2} \\ \dfrac{\rho(X_1,Y)\sigma_{X_1}\sigma_Y}{\sigma_{X_1}^2} \\ \dfrac{\rho(X_2,Y)\sigma_{X_2}\boldsymbol{\sigma}_{\mathbf{Y}}}{\sigma_{X_2}^2} \end{bmatrix} = \begin{bmatrix} \mu_Y - b_1 \mu_{X_1} - b_2 \mu_{X_2} \\ \dfrac{\rho(X_1,Y)\sigma_Y}{\sigma_{X_1}} \\ \dfrac{\rho(X_2,Y)\sigma_Y}{\boldsymbol{\sigma}_{X_2}} \end{bmatrix}$$

Variable Selection

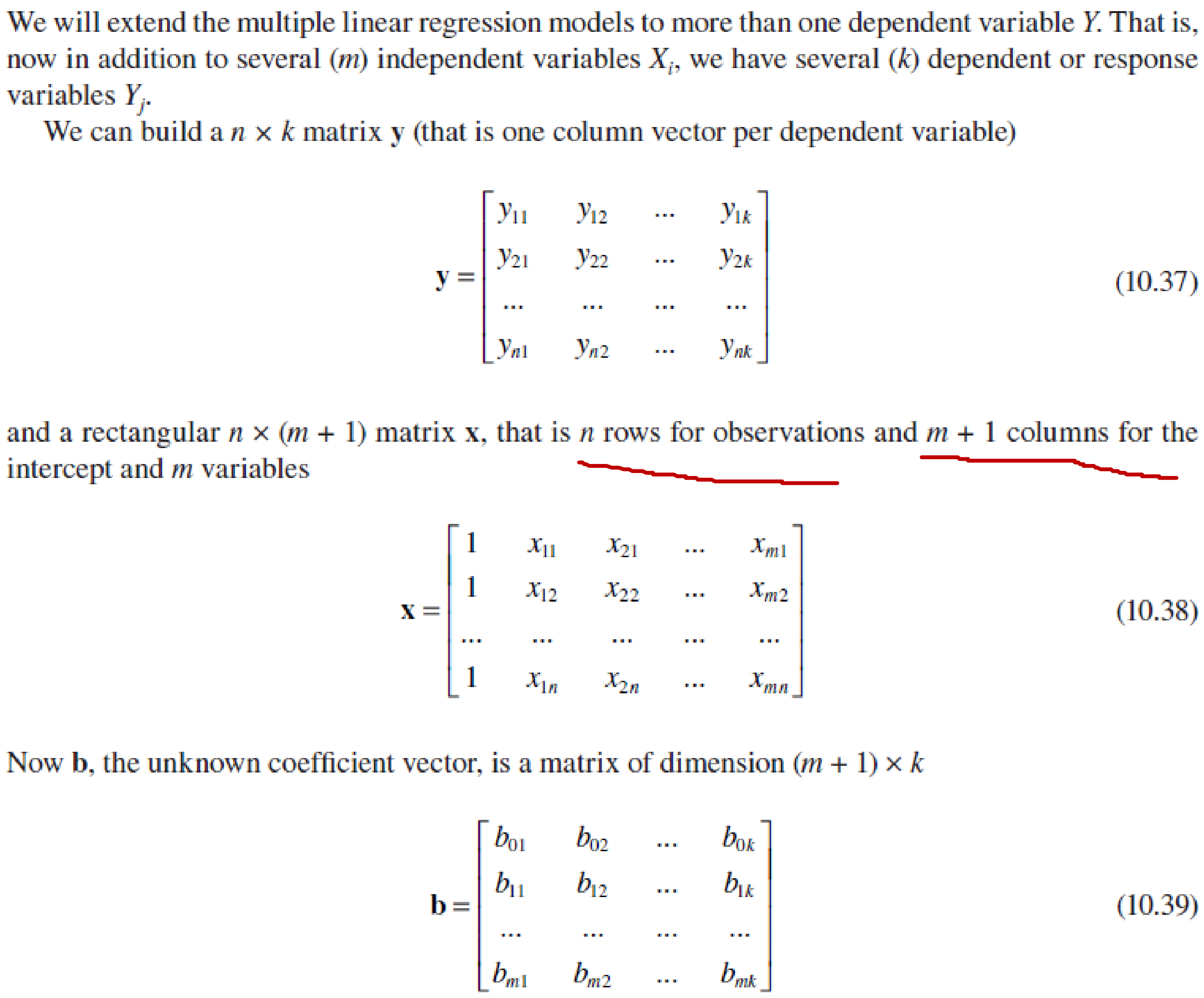
Key issues to consider when applying multiple linear regression is how many variables and which Xi to use (Rogerson, 2001). There are several ways of proceeding.

(1) Backward selection: start by including all variables at once, and then drop variables in sequence without significant reduction of R^2.

(2) Forward selection: we start with the Xi most likely to affect the Y variable, and then add independent variables.

(3) Stepwise selection: Drop and add variables as in forward and backward selection; as we add variables we check to see if we can drop a variable added before. This process can be automated using metrics that describe how good the current selection is. The Mallows’ Cp statistic or the Akaike Information Criterion (AIC) is used to decide whether an X can be dropped or added and as guide to stop the trial and error process.

Multivariate Regression



We will extend the multiple linear regression models to more than one dependent variable Y. That is, now in addition to several (m) independent variables X_i, we have several (k) dependent or response variables Y_j.

We can build a $n \times k$ matrix $\mathbf{y}$ (that is one column vector per dependent variable)

$$\mathbf{y} = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1k} \\ y_{21} & y_{22} & \dots & y_{2k} \\ \dots & \dots & \dots & \dots \\ y_{n1} & y_{n2} & \dots & y_{nk} \end{bmatrix} \tag{10.37}$$

and a rectangular $n \times (m + 1)$ matrix $\mathbf{x}$, that is n rows for observations and $m + 1$ columns for the intercept and m variables

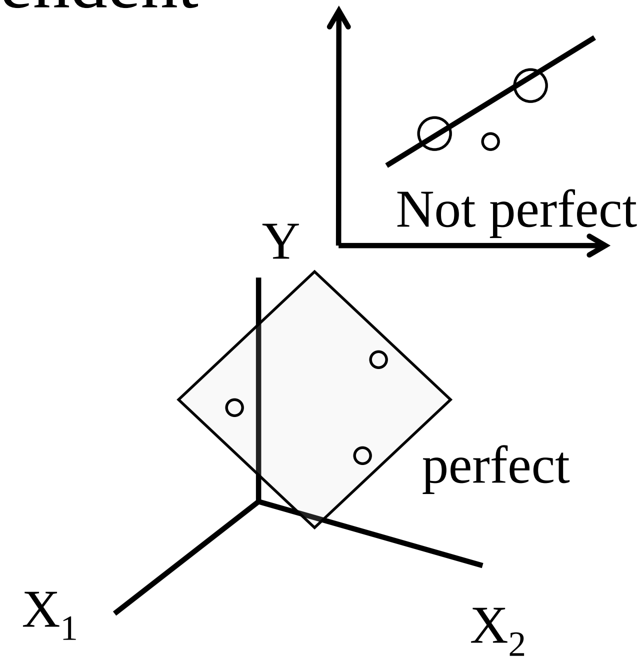
$$\mathbf{x} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{m1} \\ 1 & x_{12} & x_{22} & \dots & x_{m2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{mn} \end{bmatrix} \tag{10.38}$$

Now $\mathbf{b}$, the unknown coefficient vector, is a matrix of dimension $(m + 1) \times k$

$$\mathbf{b} = \begin{bmatrix} b_{01} & b_{02} & \dots & b_{0k} \\ b_{11} & b_{12} & \dots & b_{1k} \\ \dots & \dots & \dots & \dots \\ b_{m1} & b_{m2} & \dots & b_{mk} \end{bmatrix} \tag{10.39}$$

Reduced or Adjusted $\overline{R}^2$

- R^2 will <u>always</u> increase each time another independent variable is included
  - an additional dimension is available for fitting the regression *hyperplane* (the multiple regression equivalent of the regression line)
- Adjusted $\bar{R}^2$ is normally used instead of R^2 in multiple regression



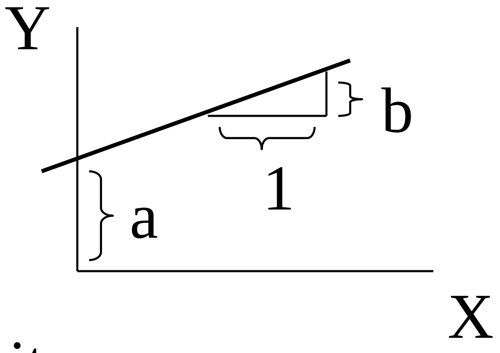
$$\overline{R}^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k}\right)$$

k is the number of coefficients in the regression equation, normally equal to the number of independent variables plus 1 for the intercept.

Interpreting *partial regression coefficients*

- The regression coefficients (b_j) tell us the change in Y for a 1 unit change in X_j, all other X variables "held constant"
- Can we compare these b_j values to tell us the relative importance of the independent variables in affecting the dependent variable?
  - If $b_1 = 2$ and $b_2 = 4$, is the affect of X_2 twice as big as the affect of X_1 ?
- **No, no, no in general!!!!**
- The size of b_j depends on the measurement scale used for each independent variable
  - if X_1 is income, then a 1 unit change is $1
  - but if X_2 is rmb or Euro(€) or even cents (₵)

    1 unit is not the same!
  - And if X_2 is *% population urban,* 1 unit is very different
- Regression coefficients are only directly comparable if the units are all the same: all $ for example



Standardized partial regression coefficients *Comparing the Importance of Independent Variable*s

- How do we compare the relative importance of independent variables?
- We know we cannot use partial regression coefficients to directly compare independent variables unless they are all measured on the same scale
- However, we can use *standardized partial regression coefficients* (also called *beta weights*, *beta coefficients*, or *path coefficients*).
- They tell us the number of standard deviation (SD) unit changes in Y for a one SD change in X)
- They are the partial regression coefficients if we had measured every variable in *standardized form*

$$\boxed{z_i = \frac{(x_i - \bar{x})}{s_X}}$$

$$\beta_{XY_j} = b_j\left(\frac{s_{X_j}}{s_Y}\right)$$

Note the confusing use of β for both standardized partial regression coefficients and for the population parameter they estimate.

Test Statistics in Multiple Regression: *testing each independent variable*

A test can be conducted for each partial regression coefficient b_j to test if the associated independent variable influences the dependent variable. It is distributed according to the *Student's t Distribution* (similar to the normal frequency distribution):

Null Hypothesis $H_o : b_j = 0$

$$t = \frac{b_j}{\mathrm{SE}(b_j)}$$

with degrees of freedom = n – k, where ***k*** is the number of coefficients in the regression equation, normally equal to the number of independent variables plus 1 for the intercept (m+1).

The formula for calculating the standard error (SE) of b_j is more complex than for simple regression , so it is not shown here.

Test Statistics in Multiple Regression *testing the overall model*

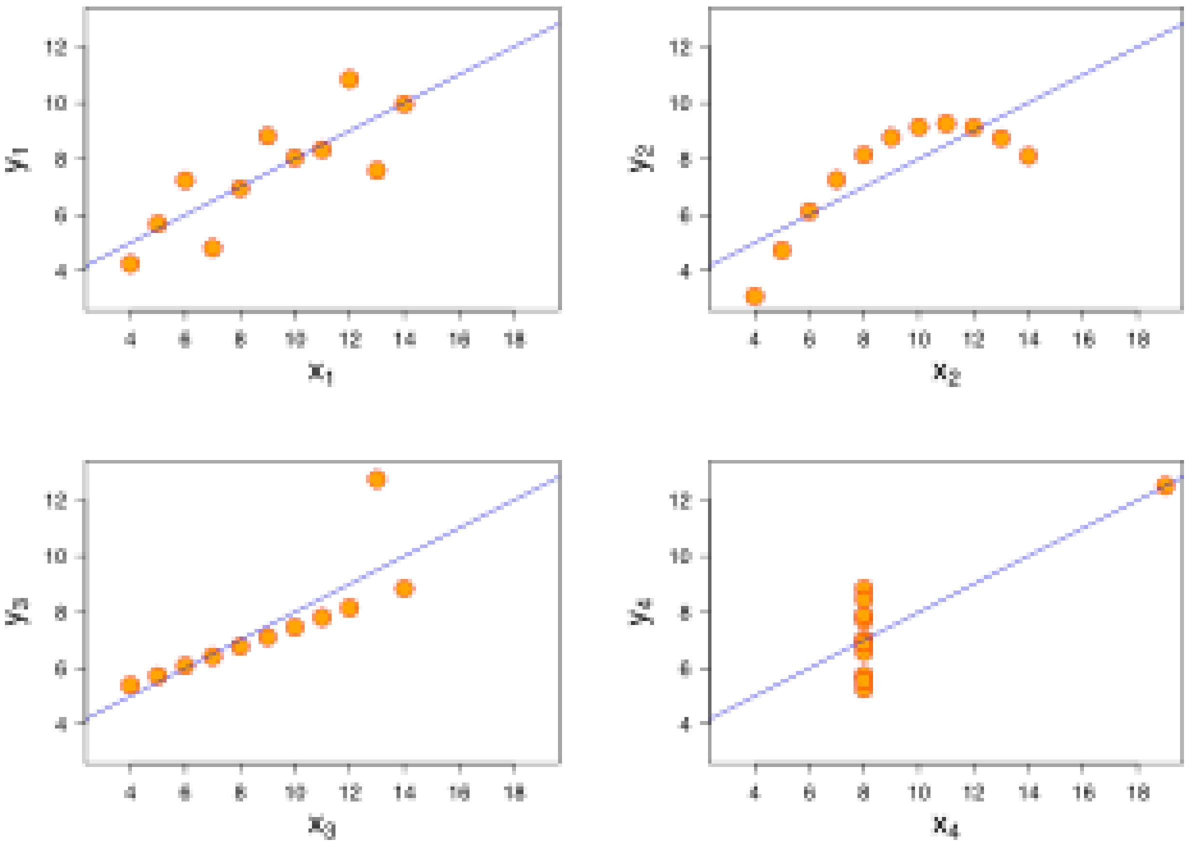
- We test the *coefficient of multiple determination* (R^2) to see if it is significantly greater than zero, using the *F* frequency distribution.
- It is an overall test to see if at least one independent variable, or two or more in combination, affect the dependent variable.
- Does not test if each and every independent variable has an effect

$$F = \frac{\text{Regression S.S./d.f.}}{\text{Residual S.S./d.f.}} = \frac{\sum(\hat{y}_i - \bar{Y})^2\ /k - l}{\sum(y_i\ - \hat{y}_i)^2\ /n - k}$$

- Again, k is the number of coefficients in the regression equation, normally equal to the number of variables (m) plus 1.
- Similar to the F test in simple regression.
  - But unlike simple regression, it is not identical to the t tests.
- It is possible (but unusual) for the F test to be significant but all t tests *not significant.*

Always look at your data

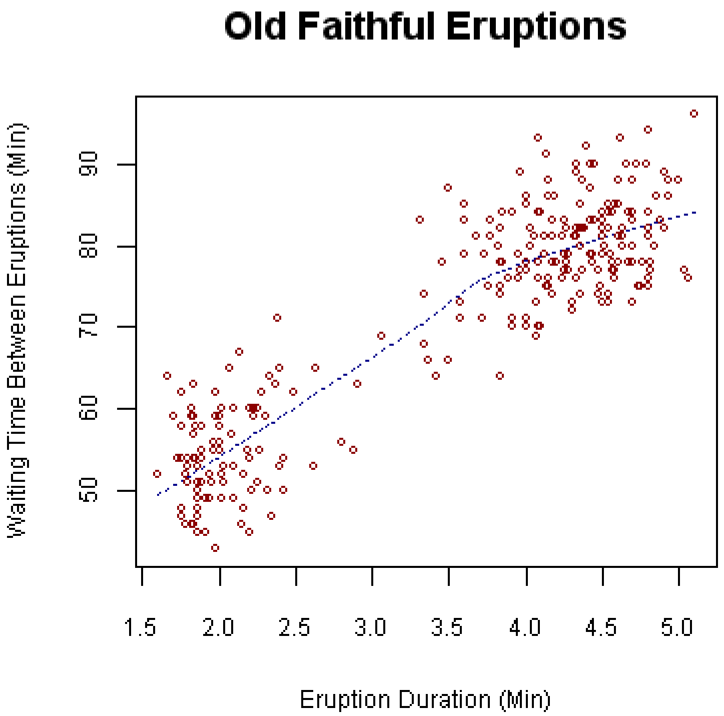
Don’t just rely on the statistics!



Anscombe's quartet
Summary statistics are the
same for all four data sets:

mean (7.5),
standard deviation (4.12),
correlation (0.816)
regression line
($y = 3 + 0.5x$).

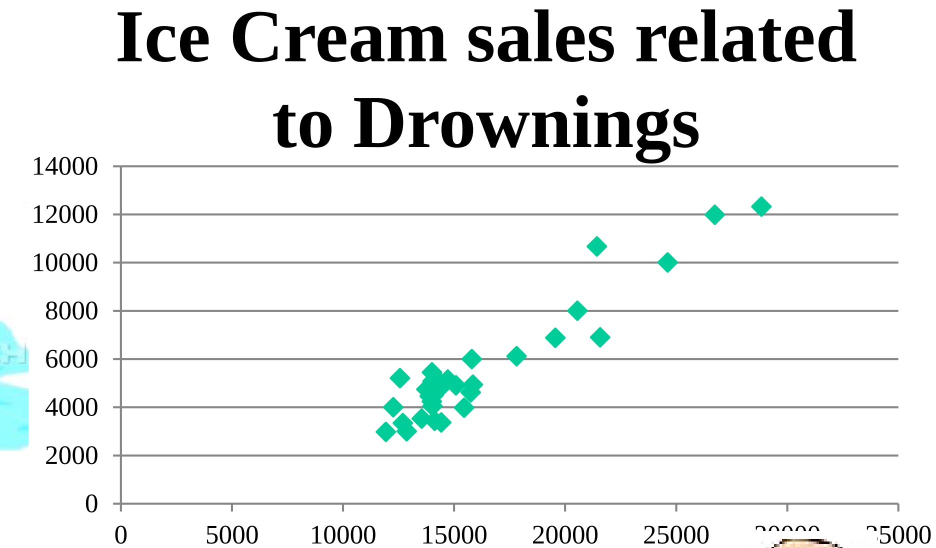
Anscombe, Francis J. (1973). "Graphs in statistical analysis". *The American Statistician* **27**: 17–21.



Real data is almost always more complex than the simple, straight line relationship assumed in regression.

Waiting time between eruptions and the duration of the eruption for the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA. This chart suggests there are generally two "types" of eruptions: short-wait-short-duration, and long-wait-long-duration.
Source: Wikipedia

Spurious relationships


Ice Cream sales related to Drownings

Eating ice cream inhibits swimming ability.
--eat too much, you cannot swim

Omitted variable problem
--both are related to a <u>third variable</u> not included in the analysis

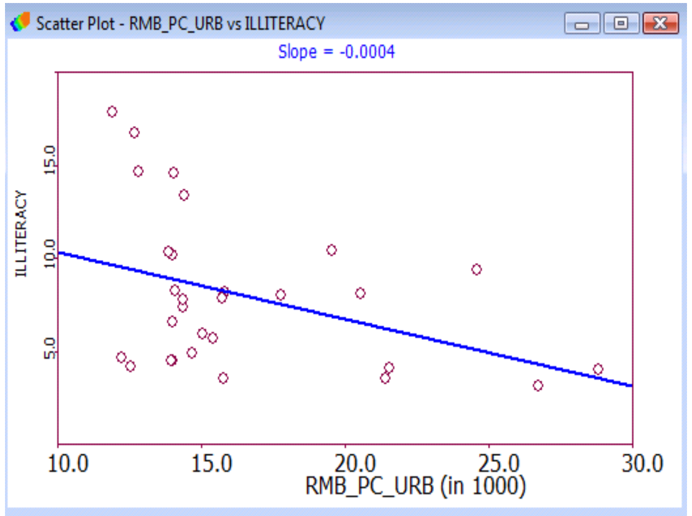
Summer temperatures:
--more people swim
(and some drown)
--more ice cream is sold

Regression does not prove direction or cause!

Income and Illiteracy

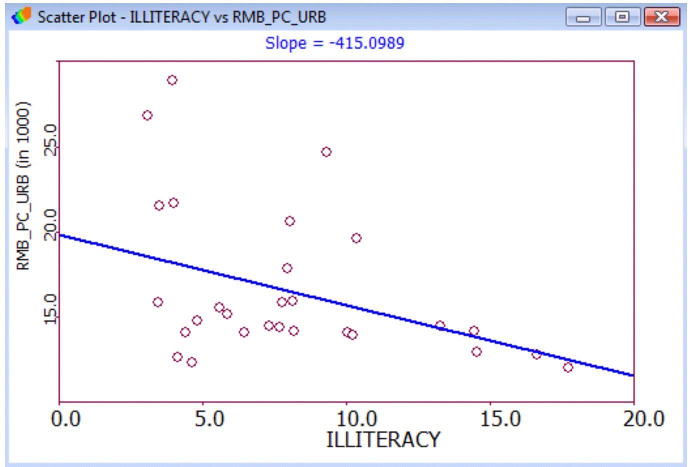
- Provinces with higher incomes can afford to spend more on education, so illiteracy is lower
  - Higher Income>>>>Less Illiteracy
- The higher the level of literacy (and thus the lower the level of <u>il</u>literacy) the more high income jobs.
  - Less Illiteracy>>>>Higher Income
- Regression <u>will not </u>decide!

Illiteracy



Income

Income

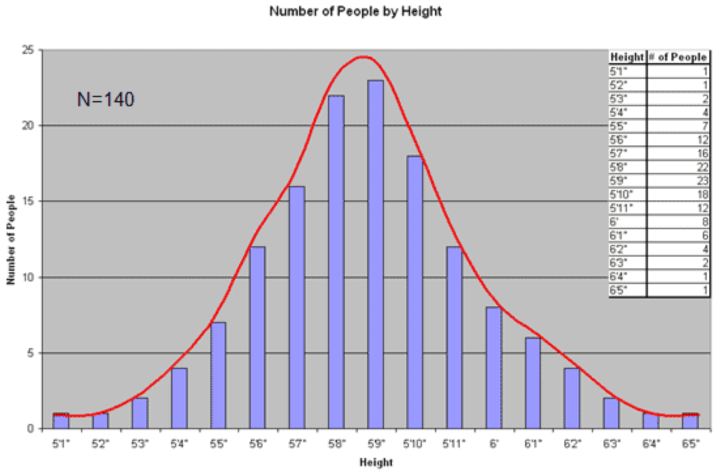


Illiteracy

Some Basic Regression Diagnostics

- The so-called *p-value* associated with the variable
  - For any statistical method, including regression, we are testing some hypothesis. In regression, we are testing the *null hypothesis* that the coefficient (i.e., slope) β is equal to zero (i.e., that the explanatory variable is not a significant predictor of the dependent variable)
  - Formally, the *p-value* is the probability of observing the value of β as extreme (i.e., as different from 0 as its estimated value is) when in reality it equals zero (i.e., when the Null Hypothesis holds). If this probability is small enough (generally, $p<0.05$), we reject the null hypothesis of β =0 for an *alternative hypothesis* of β <>0
    - Again, when the null hypothesis (of β =0) cannot be rejected, the dependent variable is not related to the independent variable.
    - The rejection of a null hypothesis (i.e., when $p <0.05$) indicates that the independent variable is a statistically significant predictor of the dependent variable
  - One p-value per independent variable

- The *sign* of the coefficient of the independent variable (i.e., the slope of the regression line)
    - One coefficient per independent variable Indicates whether the relationship between the dependent and independent variables is positive or negative
    - We should look at the sign when the coefficient is statistically significant (i.e., significantly different from zero)
- *R-squared* Coefficient of Determination): the percent of variance in the dependent variable that is explained by the predictors
    - In the single predictor case, R-squared is simply the square of the correlation between the predictor and dependent variable
    - The more independent variables included, the higher the R-squared
    - Adjusted R-squared: percent of variance in the dependent variable explained, adjusted by the number of predictors
    - One R-squared for the regression model

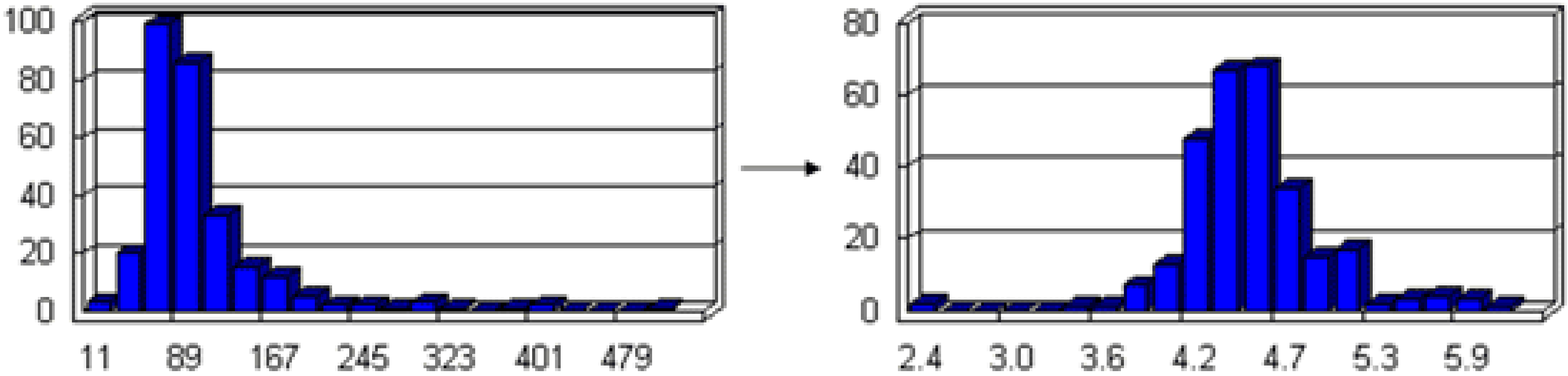


Height	# of People
5'1"	1
5'2"	1
5'3"	2
5'4"	4
5'5"	7
5'6"	12
5'7"	16
5'8"	22
5'9"	23
5'10"	18
5'11"	12
6'	8
6'1"	6
6'2"	4
6'3"	2
6'4"	1
6'5"	1

- **Some (but not all) regression assumptions**
- The dependent variable should be normally distributed (i.e., the histogram of the variable should look like a bell curve)
  - Ideally, this will also be true of independent variables, but this is not essential. Independent variables can also be binary (i.e., have two values, such as 1 (yes) and 0 (no))
- The predictors should not be strongly correlated with each other (i.e., no multicollinearity)
- Very importantly, the observations should be independent of each other. (The same holds for regression residuals). If this assumption is violated, our coefficient estimates could be wrong!
- **General rule of thumb: 10 observations per independent variable**

Data Transformations

- Sometimes, it is possible to *transform* a variable's distribution by subjecting it to some simple algebraic operation.
    - The logarithmic transformation is the most widely used to achieve normality when the variable is *positively skewed* (as in the image on the left below)
    - Analysis is then performed on the *transformed* variable.



Additional Regression Methods

- Logistic regression/Probit regression
    - ✓When your dependent variable is binary (i.e., has two possible outcomes).
        - E.g., Employment Indicator (Are you employed? Yes/No)
- Multinomial logistic regression
    - ✓When your dependent variable is categorical and has more than two categories
        - E.g., Race: Black, Asian, White, Other
- Ordinal logistic regression
    - ✓When your dependent variable is ordinal and has more than two categories
        - E.g., Education: (1=Less than High School, 2=High School, 3=More than High School)
- Poisson regression
    - ✓When your dependent variable is a count
        - E.g., Number of traffic violations (0, 1, 2, 3, 4, 5, etc)

Spatial Autocorrelation and Spatial Regression

- Recall:
    - There is spatial autocorrelation in a variable if observations that are closer to each other in space have related values (Tobler's Law)
    - One of the regression assumptions is independence of observations. If this doesn't hold, we obtain inaccurate estimates of the β coefficients, and the error term β contains spatial dependencies (i.e., meaningful information), whereas we want the error to not be distinguishable from random noise.