Unit 12
Simple Linear Regression and Correlation

"Assume that a statistical model such as a linear model is a good first start only"

- Gerald van Belle

Is higher blood pressure in the mom associated with a lower birth weight of her baby? Simple linear regression explores the relationship of ***one continuous outcome*** (Y=birth weight) with ***one continuous predictor*** (X=blood pressure). At the heart of statistics is the fitting of models to data followed by an examination of how the models perform.

-1- "somewhat useful"
A fitted model is somewhat useful if it permits exploration of hypotheses such as "higher blood pressure during pregnancy is associated with statistically significant lower birth weight" and it permits assessment of confounding, effect modification, and mediation. These are ideas that will be developed in BIOSTATS 640 Unit 5, ***Normal Theory Regression.***

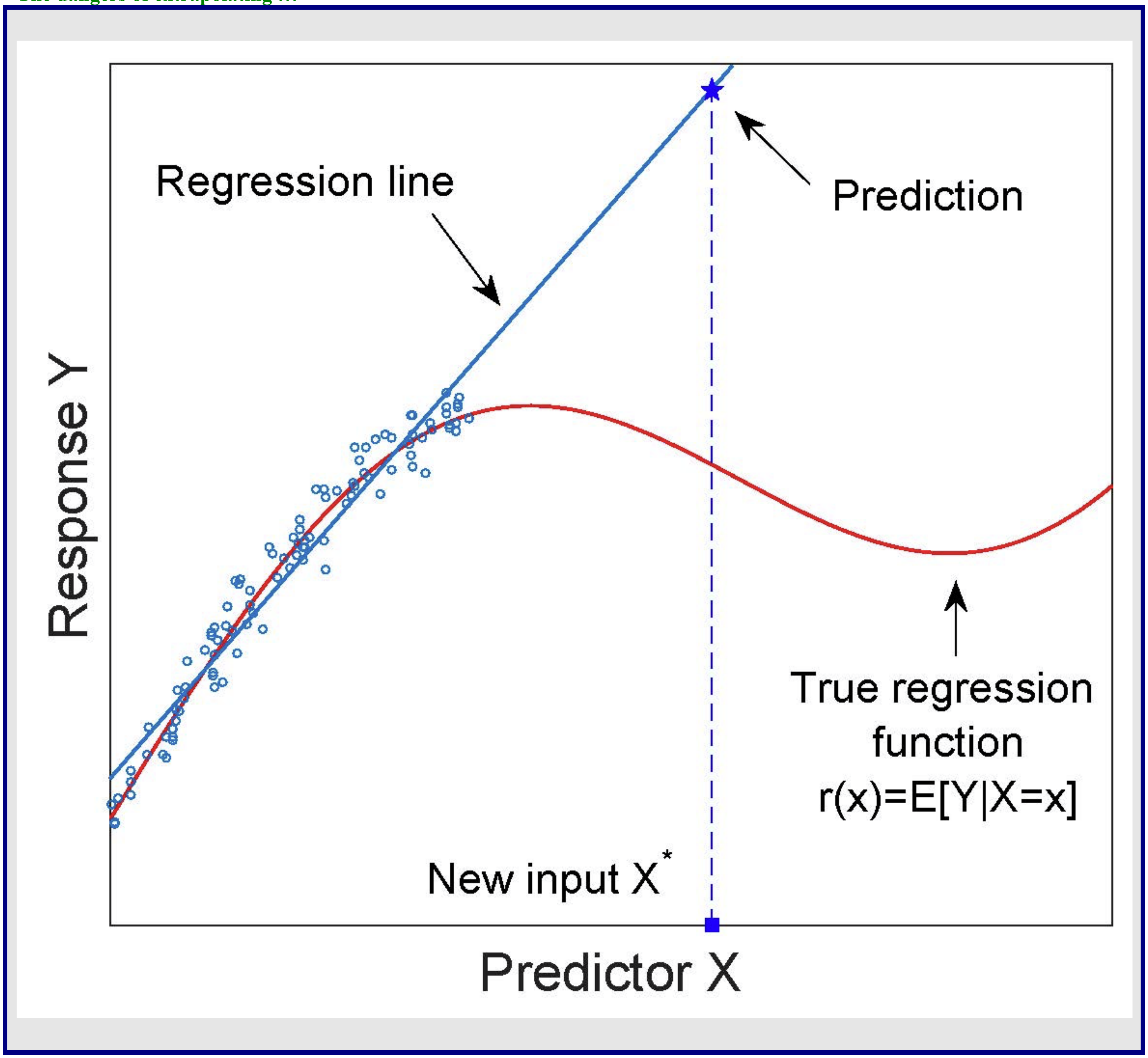
-2- "more useful"
The fitted model is more useful if it can be used to predict the outcomes of future observations. For example, we might be interested in predicting the birth weight of the baby born to a mom with systolic blood pressure 145 mm Hg.

-3- "most useful"
Sometimes, but not so much in public health, the fitted model derives from a physical-equation. An example is Michaelis-Menton kinetics. A Michaelis-Menton model is fit to the data for the purpose of estimating the actual rate of a particular chemical reaction.

Hence – ***"A linear model is a good first start only…"***

Cheers!

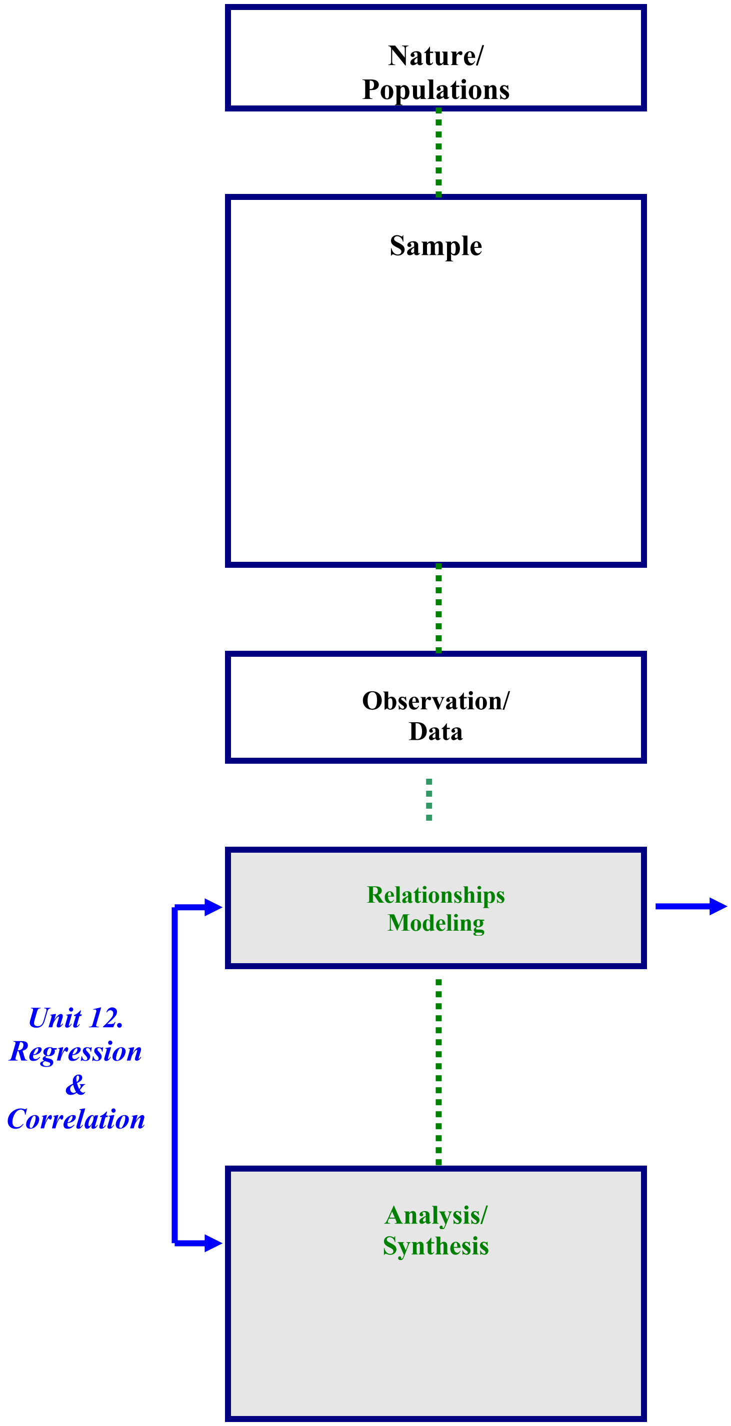
The dangers of extrapolating …



Source: Stack Exchange.

Table of Contents

1. Unit Roadmap



Simple linear regression is used when there is one *response (dependent, Y)* variable and one *explanatory (independent, X)* variables and both are *continuous*.

Examples of explanatory (independent) – response (dependent) variable pairs are height and weight, age and blood pressure, etc

-1- A simple linear regression analysis begins with a scatterplot of the data to "see" if a straight line model is appropriate:

$$y = \beta_0 + \beta_1 x \qquad \text{where}$$

Y = the response or dependent variable
X = the explanatory or independent variable.

-2- The sample data are used to estimate the parameter values and their standard errors.

β_1 = slope (the change in y per 1 unit change in x)
β_0 = intercept (the value of y when x=0)

-3- The fitted model is then compared to the simpler model $y = \beta_0$ which says that y is not linearly related to x.

Nature ______ **Population/ Sample** ______ **Observation/ Data** ______ **Relationships/ Modeling** ______ **Analysis/ Synthesis**

2. Learning Objectives

When you have finished this unit, you should be able to:

- Explain what is meant by independent versus dependent variable and what is meant by a linear relationship;
- Produce and interpret a scatterplot;
- Define and explain the intercept and slope parameters of a linear relationship;
- Explain the theory of least squares estimation of the intercept and slope parameters of a linear relationship;
- Calculate by hand the least squares estimation of the intercept and slope parameters of a linear relationship;
- Explain the theory of the analysis of variance of simple linear regression;
- Calculate by hand the analysis of variance of simple linear regression;
- Explain, compute, and interpret R^2 in the context of simple linear regression;
- State and explain the assumptions required for estimation and hypothesis tests in regression;
- Explain, compute, and interpret the overall F-test in simple linear regression;
- Interpret the computer output of a simple linear regression analysis from a package such as R, Stata, SAS, SPSS, Minitab, etc.;
- Define and interpret the value of a Pearson Product Moment Correlation, r ;
- Explain the relationship between the Pearson product moment correlation r and the linear regression slope parameter; and
- Calculate by hand the confidence interval estimation and statistical hypothesis testing of the Pearson product moment correlation r.

3. Definition of the Linear Regression Model

Unit 11 considered two **categorical (*discrete*)** variables, such as smoking (yes/no) and event of low birth weight (yes/no). It was an introduction to chi-square tests of association.

Unit 12 considers two **continuous** variables, such as age and weight. It is an introduction to **simple linear regression** and **correlation.**

A wonderful introduction to the intuition of linear regression can be found in the text by Freedman, Pisani, and Purves (Statistics. WW Norton & Co., 1978). The following is excerpted from pp 146 and 148 of their text:

> "How is weight related to height? For example, there were 411 men aged 18 to 24 in Cycle I of the Health Examination Survey. Their average height was 5 feet 8 inches = 68 inches, with an overall average weight of 158 pounds. But those men who were one inch above average in height had a somewhat higher average weight. Those men who were two inches above average in height had a still higher average weight. And so on. On the average, how much of an increase in weight is associated with each unit increase in height? The best way to get started is to look at the scattergram for these heights and weights. The object is to see how weight depends on height, so height is taken as the independent variable and plotted horizontally …
>
> ***… The regression line is to a scatter diagram as the average is to a list. The regression line estimates the average value for the dependent variable corresponding to each value of the independent variable."***

The simple linear regression model.

Consider that there is an overall distribution of Y. It has an overall mean $\mu = E[Y]$ and an overall variance $\sigma_Y^2 = Var[Y]$. Next, consider that this overall distribution is made up of subpopulations of Y, one at each level of X (for example – the distribution of Y=weight for children with X=height=50" and the distribution of Y=weight for children with X=height = 51"). We might want to know: how does the distribution of Y change, depending on which level of X we are talking about? These are called the conditional distribution of Y at X.

Modeling the mean of Y. In simple linear regression, the way in which the mean $\mu_x = E[Y\ for\ the\ sub-population\ with\ X = x] = E[Y \mid X = x]$ changes as X changes is modeled linearly: $\mu_x = \beta_0 + \beta_1 x$.

Modeling an individual observation of Y. If we have observations of Y for the subpopulation for which X=x, we are thus saying that each observed Y=y is modeled as a departure (error) from its subpopulation-specific mean as follows:

$$y = [mean] + [error\ in\ observing\ mean]$$

$$= [\mu_{X=x}] + [error]$$

$$= [\beta_0 + \beta_1 x]$$

Variance of Y within each subpopulation defined by X=x. At each value of X, the variance of Y (we call this the conditional variance of Y) is $\sigma^2_{Y|X}$. In simple linear regression, we make the assumption that this conditional variance is the same for all subpopulations defined by X ("homogeneity of error variance").

Correlation

Correlation considers the association of **two random** variables.

- The techniques of estimation and hypothesis testing are the same for linear regression and correlation analyses.
- Exploring the relationship begins with fitting a line to the points.

Development of a simple linear regression model analysis

Example.

Source: Kleinbaum, Kupper, and Muller 1988
The following are observations of age (days) and weight (kg) for n=11 chicken embryos.

WT=Y	AGE=X	LOGWT=Z
0.029	6	-1.538
0.052	7	-1.284
0.079	8	-1.102
0.125	9	-0.903
0.181	10	-0.742
0.261	11	-0.583
0.425	12	-0.372
0.738	13	-0.132
1.13	14	0.053
1.882	15	0.275
2.812	16	0.449

Notation

- The data are 11 pairs of (X_i, Y_i) where X=AGE and Y=**WT**
  $(X_1, Y_1) = (6, .029)$ $\cdots$ $(X_{11}, Y_{11}) = (16, 2.812)$ and
- This table also provides 11 pairs of (X_i, Z_i) where X=AGE and Z=**LOGWT**
  $(X_1, Z_1) = (6, -1.538)$ $\cdots$ $(X_{11}, Z_{11}) = (16, 0.449)$

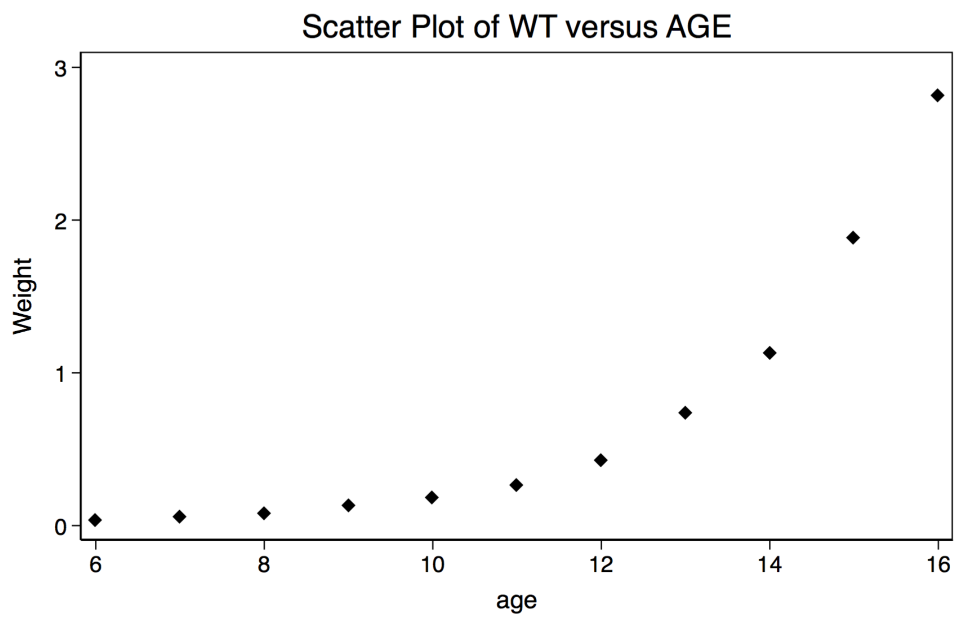
Research question
There are a variety of possible research questions:

(1) Does weight change with age?

(2) Can the variability in weight be explained, to a significant extent, by variations in age?

(3) What is a "good" functional form that relates age to weight?

Tip! Begin with a Scatter plot. Here we plot X=AGE versus Y=WT



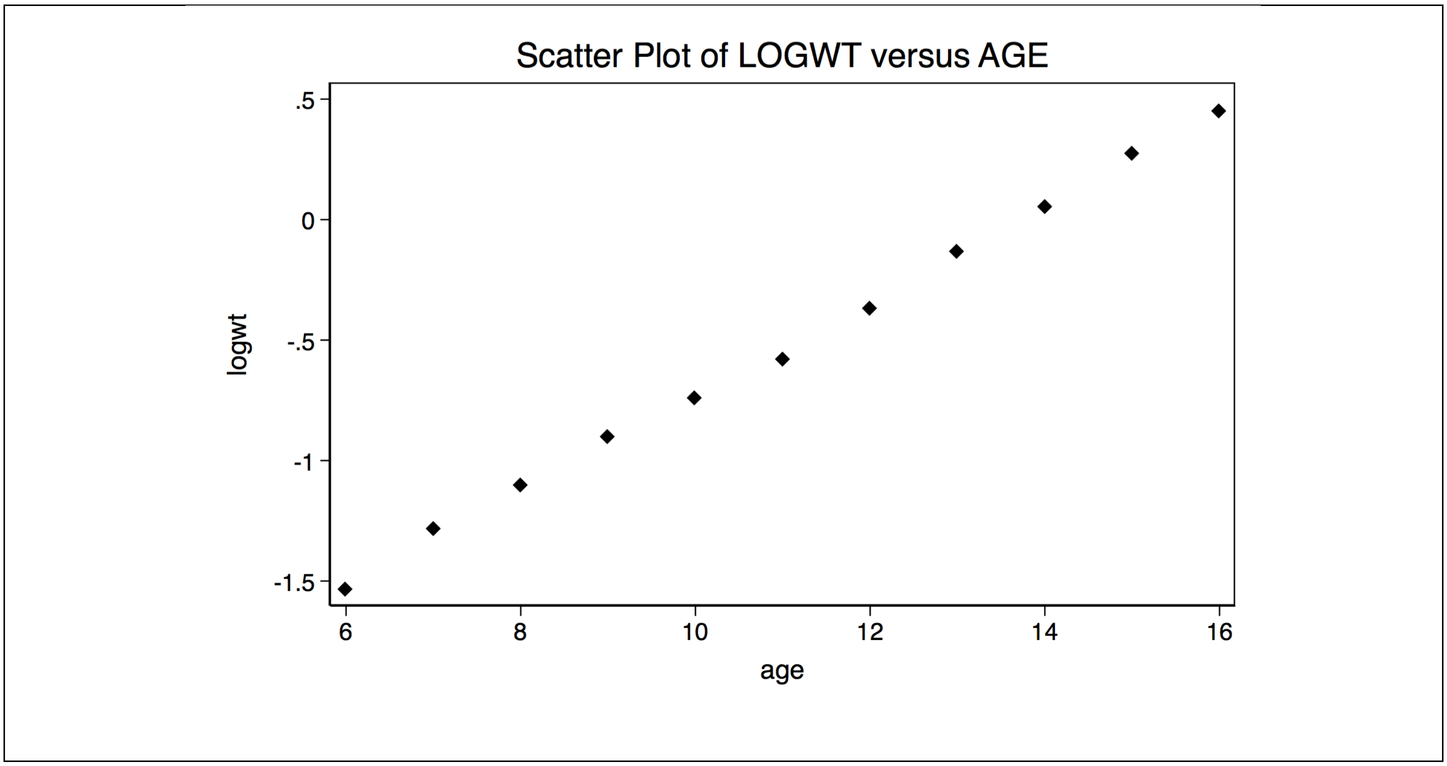
We check and learn about the following:

- The average and median of X
- The range and pattern of variability in X
- The average and median of Y
- The range and pattern of variability in Y
- The nature of the relationship between X and Y
- The strength of the relationship between X and Y
- The identification of any points that might be influential

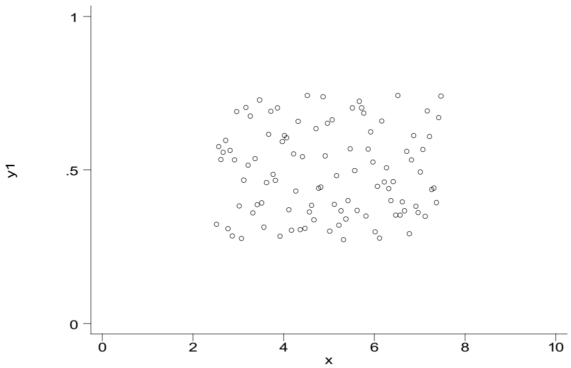
Example, continued

- The plot suggests a relationship between AGE and WT
- A straight line might fit well, but another model might be better
- We have adequate ranges of values for both AGE and WT
- There are no outliers

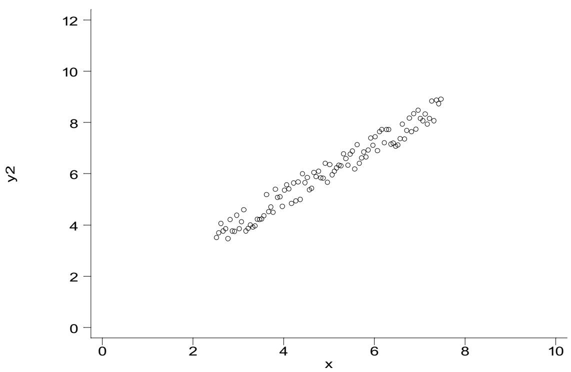
The "bowl" shape of our scatter plot suggests that perhaps a better model relates the <u>logarithm of WT</u> (Z=LOGWT) to AGE:



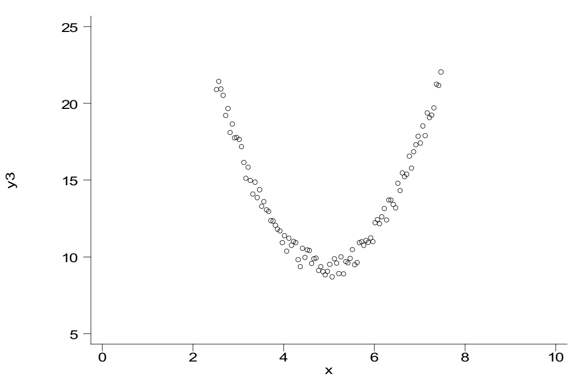
We might have gotten any of a variety of plots.



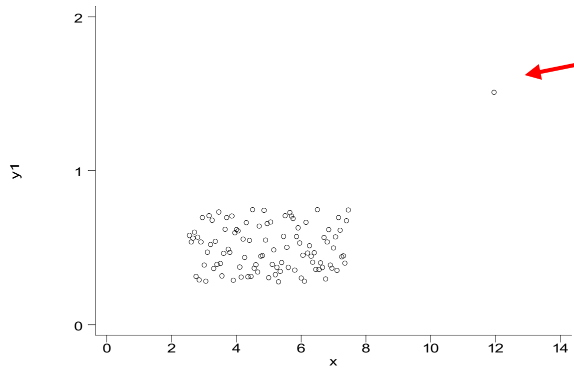
No relationship between X and Y



Linear relationship between X and Y

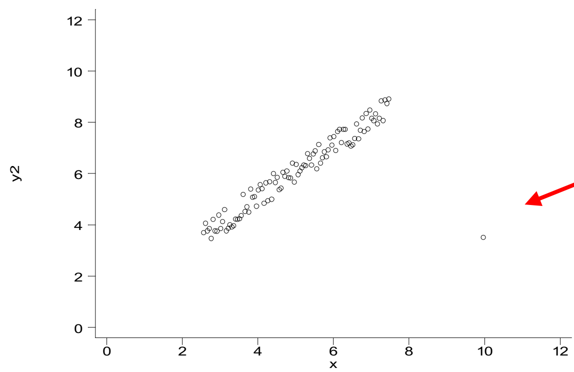


Non-linear relationship between X and Y



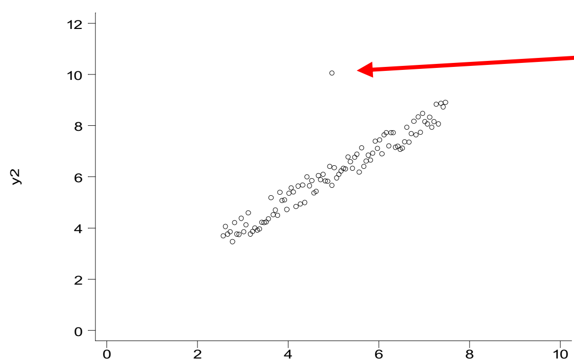
Note the outlying point

Here, a fit of a linear model will yield an estimated slope that is spuriously non-zero.



Note the outlying point

Here, a fit of a linear model will yield an estimated slope that is spuriously near zero.



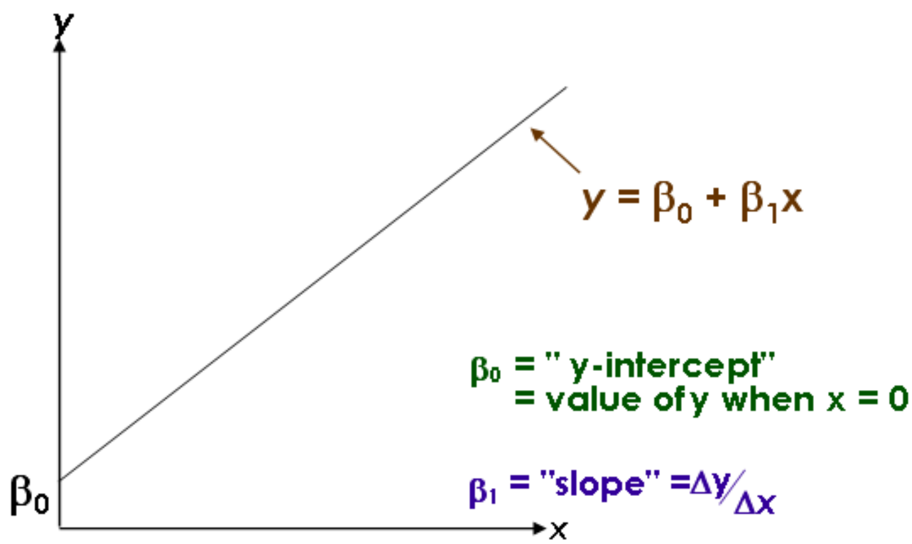
Note the outlying point

Here, a fit of a linear model will yield an estimated slope that is spuriously high.

Review of the Straight Line

Way back when, in your high school days, you may have been introduced to the straight line function, defined as **"y = mx + b"** where m is the slope and b is the intercept. Nothing new here. All we're doing is changing the notation a bit:

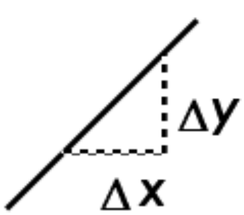
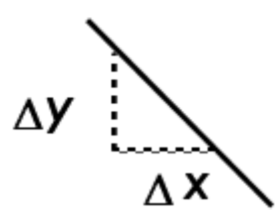
(1) Slope : m → β_1
(2) Intercept: b → β_0



β_0 = "y-intercept" = value of y when x = 0

β_1 = "slope" = $\Delta y/\Delta x$ = (change in y)/(change in x)

Slope

Slope > 0	Slope = 0	Slope < 0
		

Definition of the Straight Line Model

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

Population	Sample
$Y = \beta_0 + \beta_1 X + \varepsilon$	$Y = \hat{\beta}_0 + \hat{\beta}_1 X + e$
$Y = \beta_0 + \beta_1 X + \varepsilon$ = relationship in the population. $Y = \beta_0 + \beta_1 X$ is measured with error ε defined $\varepsilon = [Y] - [\beta_0 + \beta_1 X]$	What are $\hat{\beta}_0$, $\hat{\beta}_1$ and e? They are our estimates of β_0, β_1 and ε These estimates are also sometimes written as b_0, b_1, and e e = residual is the difference between the observed and the estimated model $e = [Y] - [\hat{\beta}_0 + \hat{\beta}_1 X]$
β_0, β_1 and ε are all unknown!!	We obtain the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and e by the method of least squares estimation.
	$\hat{\beta}_0$, $\hat{\beta}_1$ and e are known How close did we get? To see if $\hat{\beta}_0 \approx \beta_0$ and $\hat{\beta}_1 \approx \beta_1$ we perform regression diagnostics. *Regression diagnostics are discussed in BIOSTATS 640*

Notation … sorry …

Y = the outcome or dependent variable
X = the predictor or independent variable

μ_Y = The expected value of Y for all persons in the population
$\mu_{Y|X=x}$ = The expected value of Y for the sub-population for whom X=x

σ_Y^2 = Variability of Y among all persons in the population
$\sigma_{Y|X=x}^2$ = Variability of Y for the ***sub***-population for whom X=x

4. Estimation

Least squares estimation is used to obtain guesses of β_0 and β_1.

When the outcome = Y is distributed normal, least squares estimation is the same as maximum likelihood estimation. **Note – If you are not familiar with "maximum likelihood estimation", don't worry. This is introduced in BIOSTATS 640.**

"Least Squares", "Close" and Least Squares Estimation

Theoretically, it is possible to draw many lines through an X-Y scatter of points. Which to choose? "Least squares" estimation is one approach (fyi – there are others) to choosing a line that is a good fit to the data.

- **d_i =** [observed Y - fitted $\hat{Y}$] for the i^{th} person
  Perhaps we'd like d_i = [observed Y - fitted $\hat{Y}$] = smallest possible.
  Note that this is a vertical distance, since it is a distance on the vertical axis.

- $d_i^2 = \left[Y_i - \hat{Y}_i \right]^2$
  Better yet, perhaps we'd like to minimize the squared difference:
  d_i^2 = [observed Y - fitted $\hat{Y}$]^2 = smallest possible

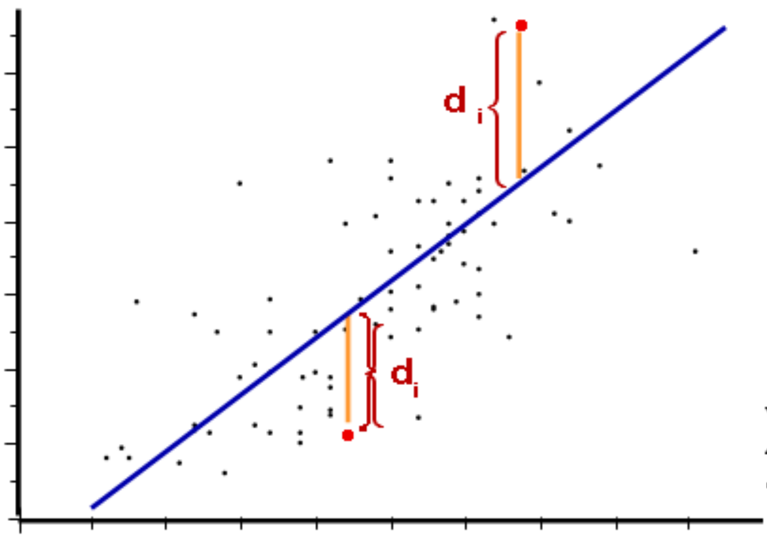
- ***Glitch.*** We can't minimize each d_i^2 separately. In particular, it is not possible to choose common values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes

  $d_1^2 = \left(Y_1 - \hat{Y}_1\right)^2$ for subject 1 ***and*** minimizes
  $d_2^2 = \left(Y_2 - \hat{Y}_2\right)^2$ for subject 2 ***and*** minimizes
  …. … ***and*** minimizes

  $d_n^2 = \left(Y_n - \hat{Y}_n\right)^2$ for the nth subject

- So, instead, we choose values for $\hat{\beta}_0$ and $\hat{\beta}_1$ that, upon insertion, minimizes the total

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 X_i\right]\right)^2$$



For each observed value x_i, we have an observed y_i, and the "predicted" value $\hat{y}_i$, on the line. The vertical distances $d_i = (y_i - \hat{y}_i)$.

$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 X_i\right]\right)^2$ **has a variety of names:**

- residual sum of squares, SSE or SSQ(residual)
- sum of squares about the regression line
- sum of squares due error (SSE)

Least Squares Estimation of the Slope and Intercept

In case you're interested, a little bit of calculus ….

- Consider SSE $= \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 X_i\right]\right)^2$

- ***Step #1**:* Differentiate with respect to $\hat{\beta}_1$

  Set derivative equal to 0 and solve for $\hat{\beta}_1$.

- ***Step #2**:* Differentiate with respect to $\hat{\beta}_0$

  Set derivative equal to 0, insert $\hat{\beta}_1$ and solve for $\hat{\beta}_0$.

Least Squares Estimation Solutions

Note – the estimates are denoted either using Greek letters with a caret or with Roman letters

Estimate of Slope $\hat{\beta}_1$ **or** b_1	$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$
Intercept $\hat{\beta}_0$ **or** b_0	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

A closer look …

Some very helpful preliminary calculations

- $S_{xx} = \sum\left(X-\bar{X}\right)^2 = \sum X^2 - N\bar{X}^2$
- $S_{yy} = \sum\left(Y-\bar{Y}\right)^2 = \sum Y^2 - N\bar{Y}^2$
- $S_{xy} = \sum\left(X-\bar{X}\right)(Y-\bar{Y}) = \sum XY - N\bar{X}\bar{Y}$

Note - These expressions make use of a "summation notation", introduced in Unit 1.

The capitol "S" indicates " summation".
In S_{xy}, the first subscript "x" is saying $(x-\bar{x})$.
The second subscript "y" is saying $(y-\bar{y})$.

$$S_{xy} = \underbrace{\sum}_{\textbf{S}}\underbrace{\left(X-\bar{X}\right)}_{\textbf{subscript x}}\underbrace{(Y-\bar{Y})}_{\textbf{subscript y}}$$

Slope	$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2} = \dfrac{\widehat{\text{cov}}(X,Y)}{\widehat{\text{var}}(X)}$	$\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$
Intercept	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$	
Prediction of Y	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ $= b_0 + b_1 X$	

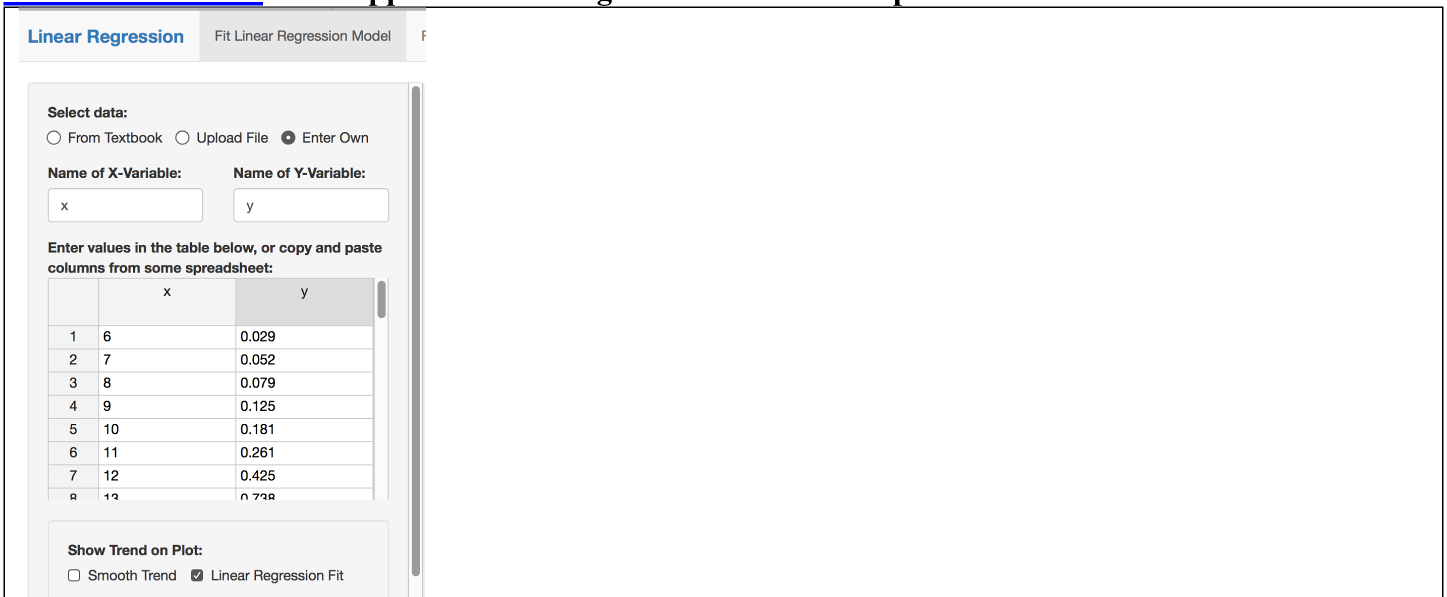
Do these estimates make sense?

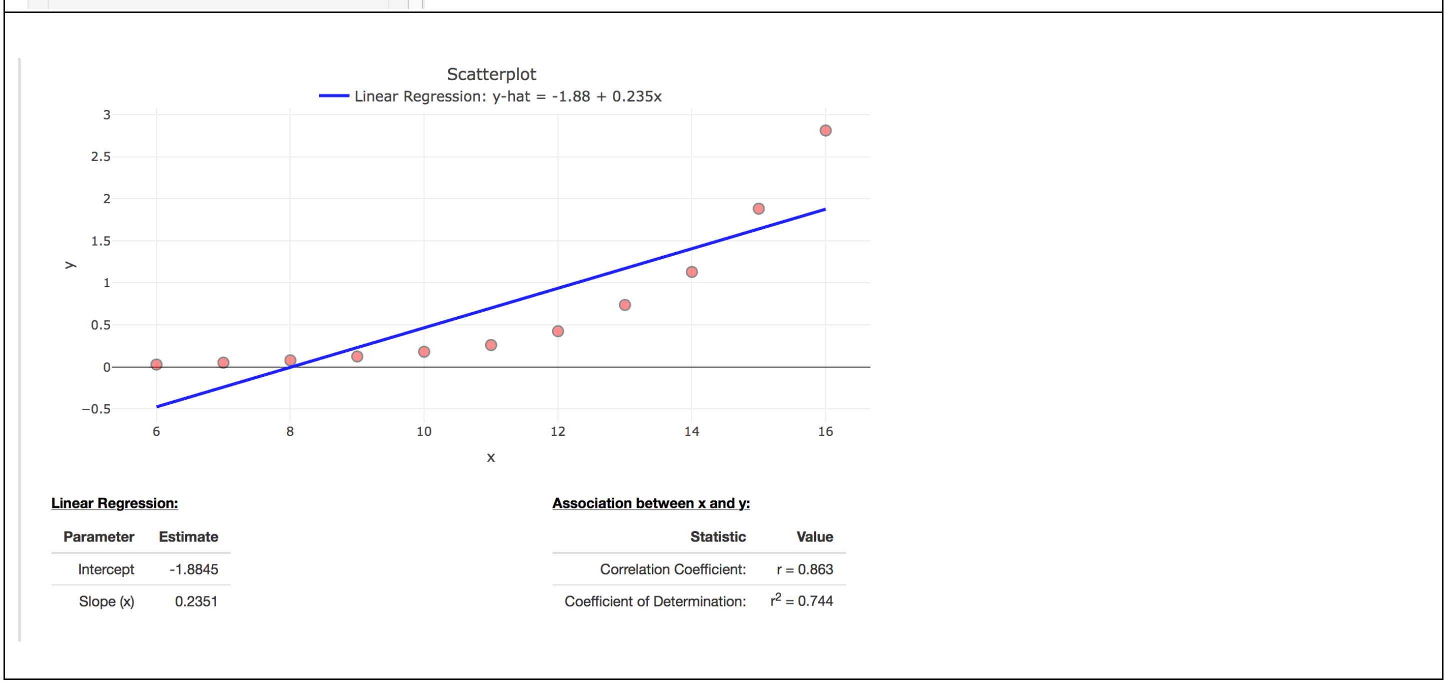
Slope	$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\hat{\text{cov}}(X,Y)}{\hat{\text{var}}(X)}$	The linear movement in Y with linear movement in X is measured relative to the variability in X. $\hat{\beta}_1 = 0$ says: With a unit change in X, overall there is a 50-50 chance that Y increases versus decreases $\hat{\beta}_1 \neq 0$ says: With a unit increase in X, Y increases also ($\hat{\beta}_1 > 0$) or Y decreases ($\hat{\beta}_1 < 0$).
Intercept	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$	If the linear model is incorrect, or, if the true model does not have a linear component, we obtain $\hat{\beta}_1 = 0$ and $\hat{\beta}_0 = \bar{Y}$ as our best guess of an unknown Y

ILLUSTRATION of Model Estimation: Y=WT and X=AGE

ArtofStat **(great for learning and practicing on your own)**

www.artofstat.com > webapps > Linear Regression > At left drop down Enter Data: "Enter Own"





The fitted line is therefore
y = -1.8845 + 0.2351*x.
Key: For each ONE unit (1 day) increase in x, y is estimated to increase by 0.2351 units (kg)

R Users

```
setwd("/Users/cbigelow/Desktop/")              # Set the working directory to desktop
rm(list=ls())                                  # Clear current workspace
options(scipen=1000)                           # Turn off scientific notation
options(show.signif.stars=FALSE)               # Turn off display of significance stars
```

Input data: Copy/paste from Excel -> table -> data frame

```
datatable=read.table(text="
y_wt    x_age   z_logwt
0.029   6.000   -1.538
0.052   7.000   -1.284
0.079   8.000   -1.102
0.125   9.000   -0.903
0.181   10.000  -0.742
0.261   11.000  -0.583
0.425   12.000  -0.372
0.738   13.000  -0.132
1.130   14.000  0.053
1.882   15.000  0.275
2.812   16.000  0.449",header=TRUE)
dataset <- as.data.frame.matrix(datatable)
```

Fit Simple Linear Regression: Dependent=y_wt Predictor=x_age

```
fit1 <- lm(y_wt ~ x_age, data=dataset)
summary(fit1)

##
## Call:
## lm(formula = y_wt ~ x_age, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5113 -0.3593 -0.1061  0.2657  0.9354
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.88453    0.52584  -3.584 0.005895
## x_age        0.23507    0.04594   5.117 0.000631
##
## Residual standard error: 0.4818 on 9 degrees of freedom
## Multiple R-squared:  0.7442, Adjusted R-squared:  0.7158
## F-statistic: 26.18 on 1 and 9 DF,  p-value: 0.0006308
```

Here (similar to the artofstat output), the fitted line is
y_wt = -1.8845 + 0.2351*x_age.
Key: For each ONE unit (1 day) increase in x_age, y_wt is estimated to increase by 0.2351 units (kg)

Stata Users

To enter the data into Stata I did the following: 1) I entered the data into an excel file, saved all columns as "numeric" ; 2) In Stata, I initialized the variables to missing using the command generate; 3) Click on data editor; and 4) Cut/paste from Excel.

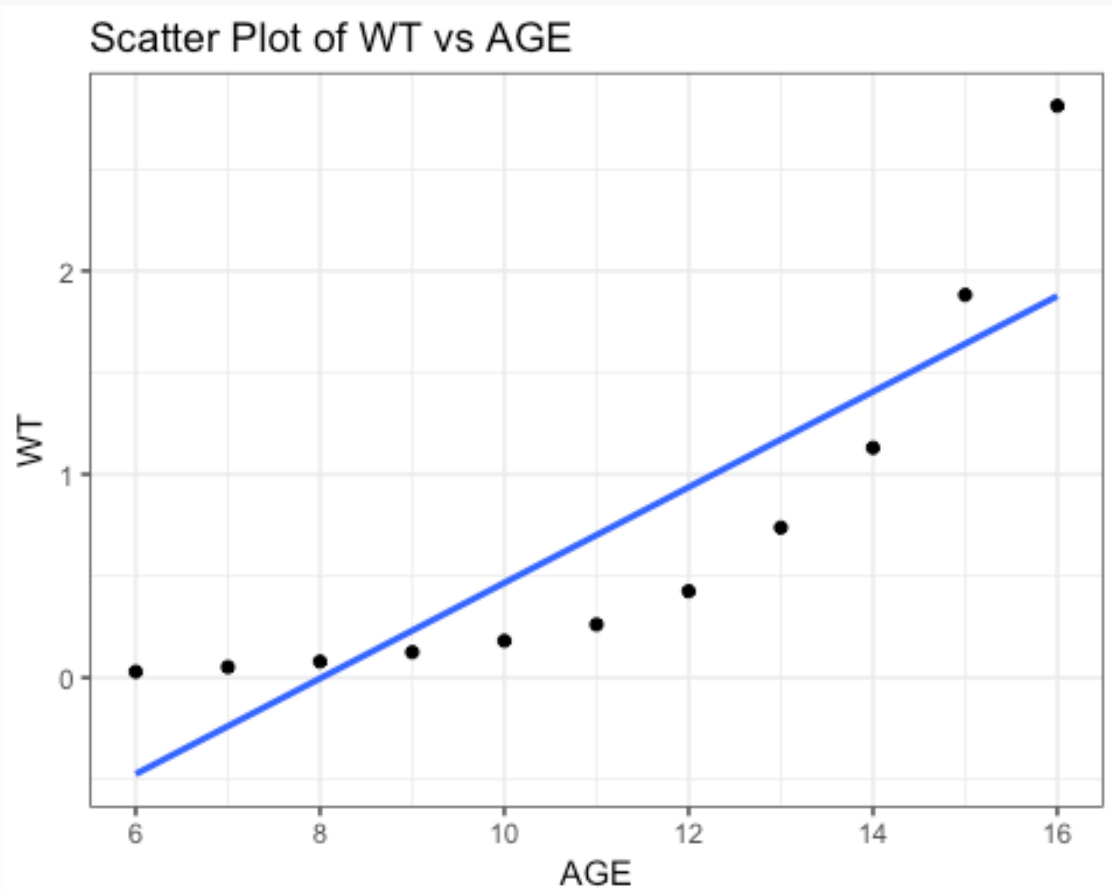
```
. generate y_wt=.
. generate x_age=.
. generate z_logwt=.
. *(3 variables, 11 observations pasted into data editor)


. * Regress Dependent=y_wt on Predictor=x_age
. regress y_wt x_age

      Source |       SS           df       MS      Number of obs   =        11
-------------+----------------------------------   F(1, 9)         =     26.18
       Model |  6.07851058         1  6.07851058   Prob > F        =    0.0006
    Residual |  2.08960166         9  .232177962   R-squared       =    0.7442
-------------+----------------------------------   Adj R-squared   =    0.7158
       Total |  8.16811224        10  .816811224   Root MSE        =    .48185

------------------------------------------------------------------------------
        y_wt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       x_age |   .2350727   .0459425     5.12   0.001     .1311437    .3390018
       _cons |  -1.884527   .5258354    -3.58   0.006     -3.07405    -.695005
------------------------------------------------------------------------------
```

Here (also similar to the artofstat ouput), the fitted line is
y_wt = -1.8845 + 0.2351*x_age.
Key: For each ONE unit (1 day) increase in x_age, y_wt is estimated to increase by 0.23507 units (kg)

ILLUSTRATION of Plot of Scatter with Overlay Fit: Y=WT and X=AGE

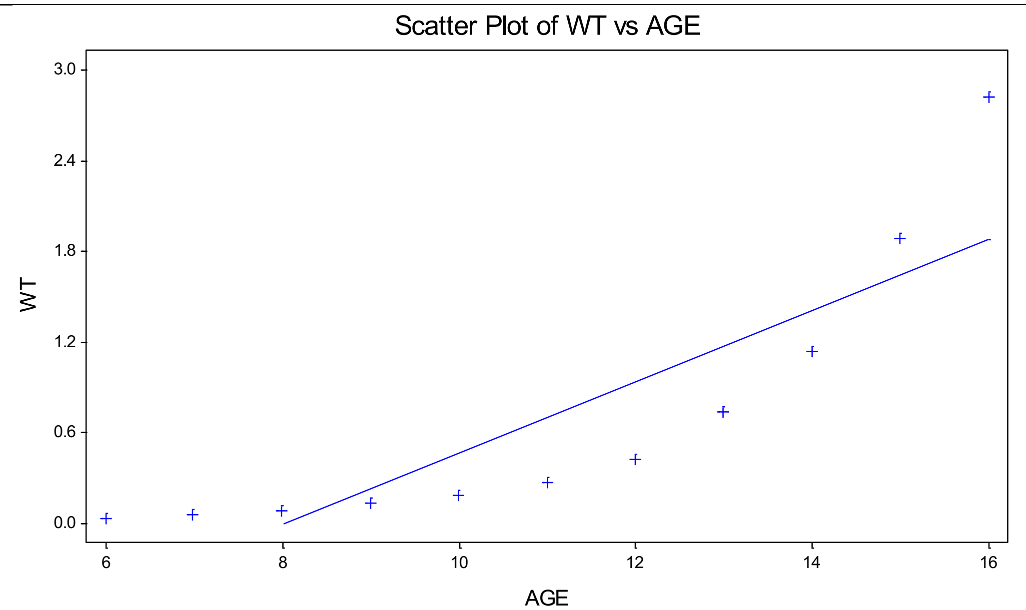
R Users

```
#  ONE TIME ONLY:  Remove comment (#) to install package ggplot2 if you have not already done this
#  install.packages("ggplot2")

library(ggplot2)
p <- ggplot(dataset, aes(x=x_age,y=y_wt)) +
     geom_smooth(method=lm, se=FALSE) +
     geom_point() +
     xlab("AGE") +
     ylab("WT") +
     ggtitle("Scatter Plot of WT vs AGE") +
     theme_bw()
p
```



Stata Users

```
. graph twoway (lfit y_wt x_age) (scatter y_wt x_age, msymbol(+)), title("Scatter Plot of WT vs
AGE")ytitle("WT") xtitle("AGE") legend(off)
```



- As we might have guessed, the straight-line model may not be the best choice.
- The "bowl" shape of the scatter plot does have a linear component, however.
- Without the plot, we might have believed the straight-line fit is okay.

ILLUSTRATION of Model Estimation: Z=LOGWT and X=AGE

R Users

```
# Fit Simple Linear Regression: Dependent=z_logwt Predictor=x_age

fit2 <- lm(z_logwt ~ x_age, data=dataset)
summary(fit2)

##
## Call:
## lm(formula = z_logwt ~ x_age, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.04854 -0.01787  0.00400  0.02168  0.03402
##
## Coefficients:
##              Estimate Std. Error t value           Pr(>|t|)
## (Intercept) -2.689255   0.030637  -87.78 0.0000000000000164
## x_age        0.195891   0.002677   73.18 0.0000000000000840
##
## Residual standard error: 0.02807 on 9 degrees of freedom
## Multiple R-squared:  0.9983, Adjusted R-squared:  0.9981
## F-statistic:  5356 on 1 and 9 DF,  p-value: 0.00000000000008399
```

The fitted line is
z_logwt = -2.6892 + 0.1959*x_age.
Key: For each ONE unit (1 day) increase in x_age, z_logwt is estimated to increase by 0.1959

Stata Users

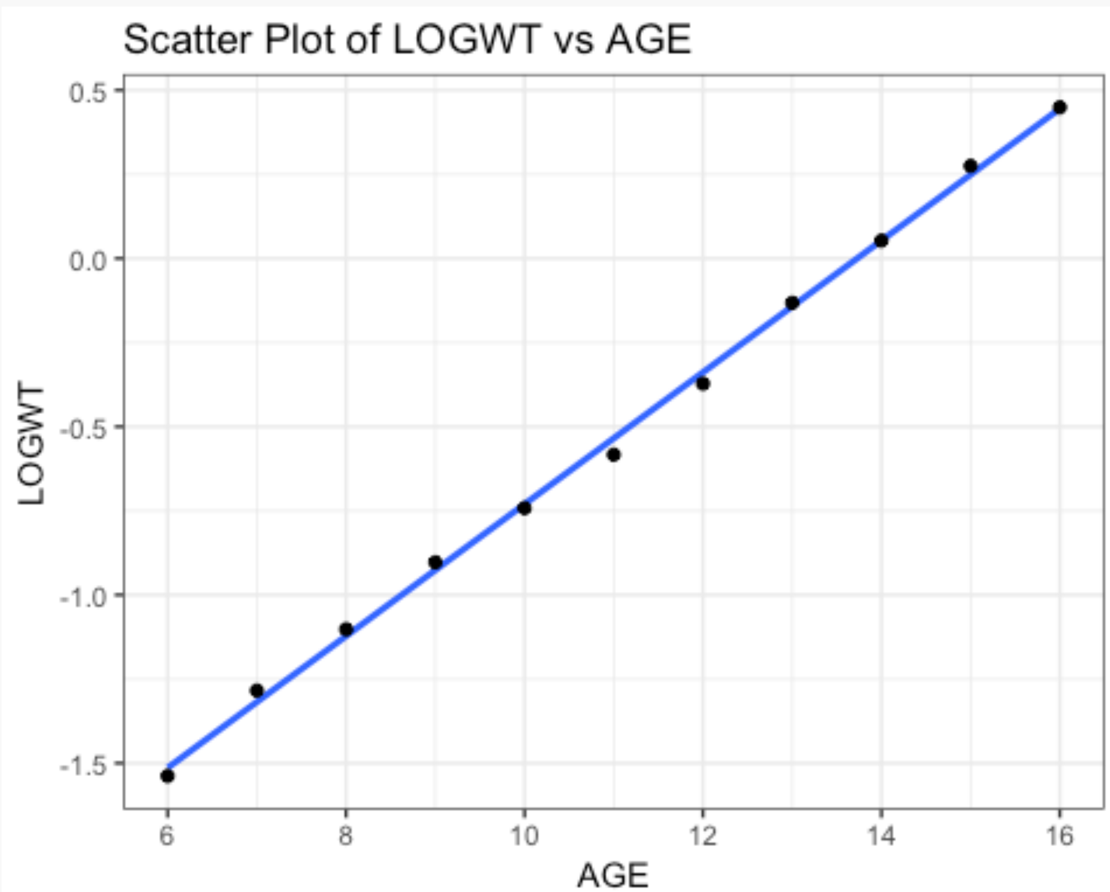
```
. * Regress Dependent=z_logwt on Predictor=x_age
. regress z_logwt x_age

      Source |       SS           df       MS      Number of obs   =        11
-------------+----------------------------------   F(1, 9)         =   5355.60
       Model |  4.22105734         1  4.22105734   Prob > F        =    0.0000
    Residual |  .007093416         9  .000788157   R-squared       =    0.9983
-------------+----------------------------------   Adj R-squared   =    0.9981
       Total |  4.22815076        10  .422815076   Root MSE        =    .02807

------------------------------------------------------------------------------
     z_logwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       x_age |   .1958909   .0026768    73.18   0.000     .1898356    .2019462
       _cons |  -2.689255    .030637   -87.78   0.000     -2.75856   -2.619949
------------------------------------------------------------------------------
```

The fitted line is
z_logwt = -2.6892 + 0.1959*x_age.
Key: For each ONE unit (1 day) increase in x_age, z_logwt is estimated to increase by 0.1959

ILLUSTRATION of Plot of Scatter with Overlay Fit: Z=LOGWT and X=AGE

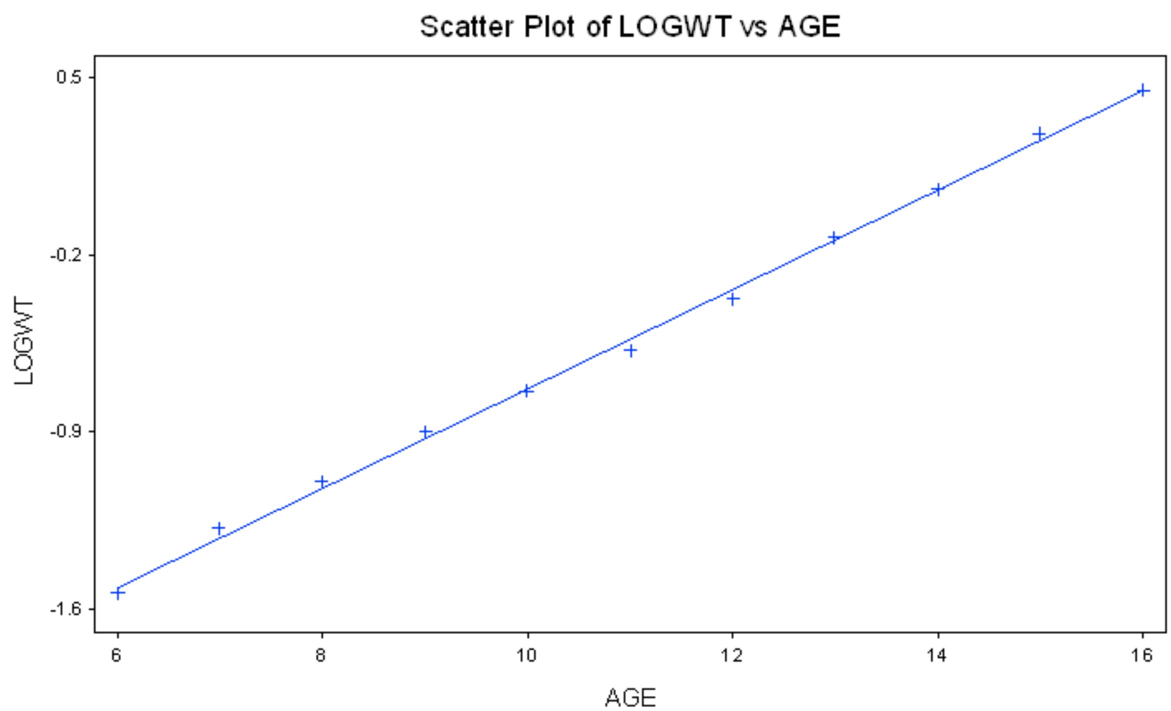
R Users

```
library(ggplot2)
p <- ggplot(dataset, aes(x=x_age,y=z_logwt)) +
     geom_smooth(method=lm, se=FALSE) +
     geom_point() +
     xlab("AGE") +
     ylab("LOGWT") +
     ggtitle("Scatter Plot of LOGWT vs AGE") +
     theme_bw()
p
```



Stata Users

```
. graph twoway (lfit z_logwt x_age) (scatter z_logwt x_age, msymbol(+)), title("Scatter Plot of LOGWT vs
AGE") ytitle("LOGWT") xtitle("AGE") legend(off)
```



- Better!
- From here on, we'll consider dependent = LOGWT versus predictor = AGE

For the brave – Try doing the calculations by hand …

Prediction of Weight from Height

Source: Dixon and Massey (1969)

Individual	Height (X)	Weight (Y)
1	60	110
2	60	135
3	60	120
4	62	120
5	62	140
6	62	130
7	62	135
8	64	150
9	64	145
10	70	170
11	70	185
12	70	160

Preliminary calculations

$\bar{X} = 63.833$	$\bar{Y} = 141.667$
$\sum X_i^2 = 49{,}068$	$\sum Y_i^2 = 246{,}100$
$\sum X_i Y_i = 109{,}380$	$S_{xx} = 171.667$
$S_{yy} = 5{,}266.667$	$S_{xy} = 863.333$

Slope	$\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$	$\hat{\beta}_1 = \dfrac{863.333}{171.667} = 5.0291$
Intercept	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$	$\hat{\beta}_0 = 141.667 - (5.0291)(63.8333)$ **= -179.3573**

5. Analysis of Variance and Introduction to R^2

Analysis of Variance

One goal (by no means the only one!) is to explain the variability in our outcomes Y.

The outcomes are comprised of our data on the dependent variable Y. In the example on page 7, the outcomes were the weights $y_1 = 0.029$, $y_2 = 0.052$, …. , $y_{11} = 2.812$. In fitting a simple linear regression of these weights in the predictor X=age, our goal (*one of them*) was to learn if some of the variability in weights could be explained by age.

The variability in our outcomes Y that we seek to explain is called the *"total sum of squares in Y"*, also called the *"total sum of squares, corrected"*.

Total Variability "to be explained"
Total Sum of Squares

$$\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$$

Key – this is the total variability of the individual observed y about their average $\overline{y}$

- **Features**
  - This is the numerator of the sample variance of the Y's
  - Because there's no division by anything, you can think of it as a measure of total scatter
  - The "noisiness" of the outcomes, if you will

- **This quantity goes by several names and notations** (sorry!)
  - "Total sum of squares"
  - " Total sum of squares, corrected"
  - SSY
  - SST

The analysis of variance starts with this total sum of squares = $\sum_{i=1}^{n}\left(\mathrm{Y}_i - \bar{\mathrm{Y}}\right)^2$ and, like a (delicious) pie, partitions (carves) it into components (wedges).

In simple linear regression, the total is partitioned into just 2 components (wedges of the pie):

1. **Due residual** (the individual Y about the individual prediction $\hat{\mathrm{Y}}$)
2. **Due regression** (the prediction $\hat{\mathrm{Y}}$ about the overall mean $\bar{\mathrm{Y}}$)

Here is the partition (Note – Look closely and you'll see that both sides are the same)

$$\left(Y_i - \bar{Y}\right) = \left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \bar{Y}\right)$$

Some algebra (not shown) reveals a nice partition of the total variability.

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

Total Sum of Squares = Due Error Sum of Squares + Due Model Sum of Squares

Simple Linear Regression Analysis of Variance
The Total Variability in Y is Partitioned into 2 Components

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

Total Sum of Squares = Due Error Sum of Squares + Due Model Sum of Squares

Total Sum of Squares	Due Error Sum of Squares	Due Model Sum of Squares
total variability of y about average $\bar{y}$	**variability of y about predicted $\hat{y}$**	**variability of predicted $\hat{y}$ about average $\bar{y}$**

Example.
Consider again the data on page 7 and, specifically, the simple linear model where outcomes Y=WT were modeled linearly in X=AGE. The output of these model fits were the following (depending on R or Stata):

R Users
Note: The command anova( ) does not produce display of the total sum of squares.

```
# fit1 <- lm(y_wt ~ x_age, data=dataset)
summary(fit1)
anova(fit1)

Analysis of Variance Table
Response: y_wt
          Df Sum Sq Mean Sq F value    Pr(>F)
x_age      1 6.0785  6.0785   26.18 0.0006308
Residuals  9 2.0896  0.2322
```

KEY:
Due Model Sum of Squares = $\sum(\hat{y}-\overline{y})^2$ = x_age Sum Sq = 6.0785
Due Error Sum of Squares = $\sum(y-\hat{y})^2$ = Residuals Sum Sq = 2.0896

Stata Users

```
regress y_wt x_age

      Source |       SS           df       MS      Number of obs   =        11
-------------+----------------------------------   F(1, 9)         =     26.18
       Model |  6.07851058         1  6.07851058   Prob > F        =    0.0006
    Residual |  2.08960166         9  .232177962   R-squared       =    0.7442
-------------+----------------------------------   Adj R-squared   =    0.7158
       Total |  8.16811224        10  .816811224   Root MSE        =    .48185
```

KEY:
Due Model Sum of Squares = $\sum(\hat{y}-\overline{y})^2$ = Model SS = 6.07851058
Due Error Sum of Squares = $\sum(y-\hat{y})^2$ = Residual SS = 2.08960166
Total Sum of Squares, corrected = $\sum(y-\overline{y})^2$ = Total SS = 8.16811224

A closer look…

Total Sum of Squares = Due Model Sum of Squares + Due Error Sum of Squares

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2=\underbrace{\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}_{\text{due model sum of squares}}+\underbrace{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}_{\text{due error sum of squares}}$$

- $\left(Y_i-\bar{Y}\right)$ = **deviation of** Y_i **from** $\bar{Y}$ **that is to be explained**
- $\left(\hat{Y}_i-\bar{Y}\right)$ = **"due model", "signal", "systematic", "due regression"**
- $\left(Y_i-\hat{Y}_i\right)$ = **"due error", "noise", or "residual"**

We seek to *explain* the total variability $\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$ **with a fitted model:**

What happens when $\beta_1 \neq 0$?	What happens when $\beta_1 = 0$?
A straight-line relationship is helpful	A straight-line relationship is not helpful
Best guess is $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X$	Best guess is $\hat{Y}=\hat{\beta}_0=\bar{Y}$
Due model "sum of squares" tends to be LARGE because $\left(\hat{Y}-\bar{Y}\right)=(\left[\hat{\beta}_0+\hat{\beta}_1 X\right]-\bar{Y})$ $=\bar{Y}-\hat{\beta}_1\bar{X}+\hat{\beta}_1 X-\bar{Y}$ $=\hat{\beta}_1\left(X-\bar{X}\right)$	Due error "sum of squares" tends to be nearly the TOTAL because $\left(Y-\hat{Y}\right)=\left(Y-\left[\hat{\beta}_0\right]\right)=\left(Y-\bar{Y}\right)$
Due error "sum of squares" has to be small	Due regression "sum of squares" has to be small
→ $\dfrac{\text{due(model)}}{\text{due(error)}}$ will be large	→ $\dfrac{\text{due(model)}}{\text{due(error)}}$ will be small

Partitioning the Total Variance
and all things sum of squares and mean squares

1. *The total "pie" is what we are partitioning and it is, simply, the variability in the outcome.* **Thus, the "total" or "total, corrected" refers to the variability of Y about $\overline{Y}$**
   - $\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$ is called the "total sum of squares"
   - Degrees of freedom = df = (n-1)
   - Division of the "total sum of squares" by its df yields the "total mean square"

2. *One "piece of the pie" what the model explains.* **The "regression" or "due model" refers to the variability of $\hat{Y}$ about $\overline{Y}$**
   - $\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2=\hat{\beta}_1^2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$ is called the "regression sum of squares"
   - Degrees of freedom = df = 1
   - Division of the "regression sum of squares" by its df yields the "regression mean square" or "model mean square". It is an example of a variance component.

3. *The remaining, "other piece of the pie" what's left over after we've explained what we can with our model.*

   The "residual" or "due error" refers to the variability of Y about $\hat{Y}$
   - $\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$ is called the "residual sum of squares"
   - Degrees of freedom = df = (n-2)
   - Division of the "residual sum of squares" by its df yields the "residual mean square".

Source	**df**	**Sum of Squares** A measure of variability	**Mean Square = Sum of Squares / df** A measure of average/typical/mean variability
Regression **due model**	**1**	$SSR=\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2$	**msq(model) = SSR/1**
Residual **due error**	**(n-2)**	$SSE=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$	**msq(residual) = SSE/(n-2) =** $\hat{\sigma}^2_{Y\mid X}$
Total, corrected	**(n-1)**	$SST=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$	

Be careful! The question we may ask from an analysis of variance table is a limited one.

Does the fit of the straight-line model explain a significant portion of the variability of the individual Y about $\overline{Y}$?

Is this fitted model better than using $\overline{Y}$ alone?

We are NOT asking:

Is the choice of the straight line model correct? ***nor***
Would another functional form be a better choice?

We'll use a hypothesis test approach (another "proof by contradiction" reasoning just like we did in Units 8-10).

- Assume, provisionally, the "nothing is going on" null hypothesis that says $\beta_1 = 0$ ("no linear relationship")
- Use least squares estimation to estimate a "closest" line
- The analysis of variance table provides a comparison of the due regression mean square to the residual mean square
- Where does least squares estimation take us, vis a vis the slope β_1?
  If $\beta_1 \neq 0$ Then due (regression)/due (residual) will be LARGE
  If $\beta_1 = 0$ Then due (regression)/due (residual) will be SMALL
- Our p-value calculation will answer the question:
  If the null hypothesis is true and $\beta_1 = 0$ truly, what were the chances of obtaining a value of due (regression)/due (residual) as larger or larger than that observed?

To calculate "chances of extremeness under some assumed null hypothesis" we need a null hypothesis probability model! But did you notice? So far, we have not actually used one!

R^2

R^2 is the proportion (%) of the total variability in Y that is explained by the fitted model

R^2

Coefficient of Determination

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = \frac{\text{Due Model Sum of Squares}}{\text{Total Sum of Squares}} = \frac{\text{SSR}}{\text{SST}}$$

Note – This is often multiplied by 100 and expressed as a %

- **Features**
  - R^2 is the percent of the total variability that is explained by the model just fit
  - It is a proportion

- Special Case: Simple Linear Regression
  - $\sqrt{R^2} = r =$ Pearson product moment correlation
  - r is a measure of linear association
  - More on this ahead in Section 9. *Introduction to Correlation*

R Users

```
fit1 <- lm(y_wt ~ x_age, data=dataset)
summary(fit1)

##
## Residual standard error: 0.4818 on 9 degrees of freedom
## Multiple R-squared:  0.7442, Adjusted R-squared:  0.7158
## F-statistic: 26.18 on 1 and 9 DF,  p-value: 0.0006308
```

KEY:

Due Model Sum of Squares = $\sum(\hat{y}-\bar{y})^2$ = x_age Sum Sq = 6.0785

Due Error Sum of Squares = $\sum(y-\hat{y})^2$ = Residuals Sum Sq = 2.0896

R Squared = [Model Sum of Squares] / [Total Sum of Squares] = 6.0785 / [6.0785 + 2.0896] = 0.7442

Stata Users

```
regress y_wt x_age

      Source |       SS           df       MS      Number of obs   =        11
-------------+----------------------------------   F(1, 9)         =     26.18
       Model |  6.07851058         1  6.07851058   Prob > F        =    0.0006
    Residual |  2.08960166         9  .232177962   R-squared       =    0.7442
-------------+----------------------------------   Adj R-squared   =    0.7158
       Total |  8.16811224        10  .816811224   Root MSE        =    .48185
```

KEY:

Due Model Sum of Squares = $\sum(\hat{y}-\bar{y})^2$ = Model SS = 6.07851058

Due Error Sum of Squares = $\sum(y-\hat{y})^2$ = Residual SS = 2.08960166

Total Sum of Squares, corrected = $\sum(y-\bar{y})^2$ = Total SS = 8.16811224

R Squared = [Model Sum of Squares] / [Total Sum of Squares] = 6.07851058 / 8.16811224 = 0.7442

6. Assumptions for a Straight-Line Regression Analysis

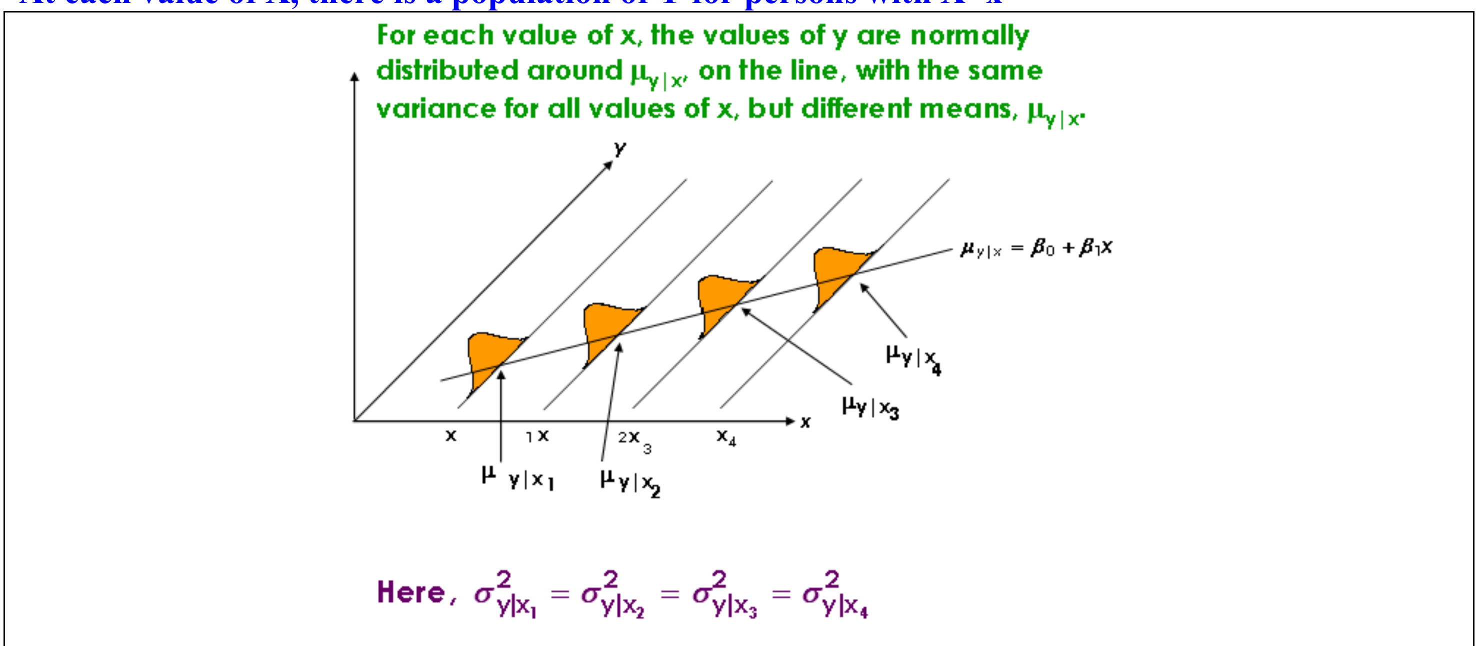
In performing least squares estimation, we did not use a probability model. We were doing geometry. Confidence interval estimation and hypothesis testing require some assumptions and a probability model. Here you go!

Assumptions for Simple Linear Regression

- **The separate observations $Y_1, Y_2, \cdots, Y_n$ are independent.**
- **The values of the predictor variable X are fixed and measured without error.**
- **For each value of the predictor variable X=x, the distribution of values of Y follows a normal distribution with mean equal to $\mu_{Y|X=x}$ and common variance equal to $\sigma_{Y|x}^2$.**
- **The separate means $\mu_{Y|X=x}$ lie on a straight line; that is –**

$$\mu_{Y|X=x} = \beta_0 + \beta_1 X$$

At each value of X, there is a population of Y for persons with X=x



With these assumptions, we can assess the significance of the variance explained by the model.

$$F = \frac{\text{mean square(model)}}{\text{mean square(residual)}} = \frac{\text{msq(model)}}{\text{msq(residual)}} \quad \text{with df} = 1, (n-2)$$

When $\beta_1 = 0$	When $\beta_1 \neq 0$
Mean square model, msq(model), has expected value $\sigma_{Y\|X}^2$	Mean square model, msq(model), has expected value $\sigma_{Y\|X}^2 + \beta_1^2 \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$
Mean square residual, msq(residual), has expected value $\sigma_{Y\|X}^2$	Mean square residual, msq(residual), has expected value $\sigma_{Y\|X}^2$
F = msq(model)/msq(residual) tends to be **close to 1**	F = msq(model)/msq(residual) tends to be **LARGER than 1**

We obtain the analysis of variance table for the model of Z=LOGWT to X=AGE:

R Users: Annotations highlighted in yellow.

ANOVA Table: Dependent=z_logwt Predictor=x_age

```
fit2 <- lm(z_logwt ~ x_age, data=dataset)
anova(fit2)

## Analysis of Variance Table
##
## Response: z_logwt
##           Df Sum Sq Mean Sq F value              Pr(>F)
## x_age      1 4.2211  4.2211  5355.6 0.00000000000008399
## Residuals  9 0.0071  0.0008

summary(fit2)


##
## Call:
## lm(formula = z_logwt ~ x_age, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.04854 -0.01787  0.00400  0.02168  0.03402
##
## Coefficients:
##              Estimate Std. Error t value           Pr(>|t|)
## (Intercept) -2.689255   0.030637  -87.78 0.0000000000000164
## x_age        0.195891   0.002677   73.18 0.0000000000000840
##
## Residual standard error: 0.02807 on 9 degrees of freedom          = Square root of MSQ(residual)
## Multiple R-squared:  0.9983, Adjusted R-squared:  0.9981         R² = SSQ(model)/SSQ(TOTAL)
## F-statistic:  5356 on 1 and 9 DF,  p-value: 0.00000000000008399   F = MSQ(model)/MSQ(residual)
                                                                     =   4.2211/0.0008
```

Stata Users: Annotations in red.

```
. * Regress Dependent=z_logwt on Predictor=x_age
. regress z_logwt x_age


      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) = 5355.60  = MSQ(model)/MSQ(residual)
       Model |  4.22105734     1  4.22105734           Prob > F      =  0.0000  = p-value for Overall F Test
    Residual |  .007093416     9  .000788157           R-squared     =  0.9983  = SSQ(model)/SSQ(TOTAL)
-------------+------------------------------           Adj R-squared =  0.9981  = R² adjusted for n and # of X
       Total |  4.22815076    10  .422815076           Root MSE      =  .02807  = Square root of MSQ(residual)
```

This output corresponds to the following.
Note – In this example our dependent variable is actually Z, not Y.

Source	Df	Sum of Squares	Mean Square
Regression ***due model***	1	$SSR = \sum_{i=1}^{n}\left(\hat{Z}_i - \bar{Z}\right)^2 = 4.22063$	msq(model) = SSR/1 = 4.22063/1 = 4.22063 ***You might see msq(model) = msr***
Residual ***due error***	(n-2) = 9	$SSE = \sum_{i=1}^{n}\left(Z_i - \hat{Z}_i\right)^2 = 0.00705$	msq(residual) = SSE/(n-2) = 0.00705/9 = 0.00078 ***You might see msq(residual) = mse***
Total, corrected	(n-1) = 10	$SST = \sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2 = 4.22768$	

Other information in this output:

- **R-SQUARED** = [(Sum of squares regression)/(Sum of squares total)]
  = proportion of the "total" that we have been able to explain with the fit
  = "percent of variance explained by the model"

  - ***Be careful!*** As predictors are added to the model, R-SQUARED can only increase. Eventually, we need to "adjust" this measure to take this into account. See ADJUSTED R-SQUARED.

- We also get an overall F test of the null hypothesis that the simple linear model does not explain significantly more variability in LOGWT than the average LOGWT. F = MSQ (Regression)/MSQ (Residual)

  = 4.22063/0.0007838
  = 5384.94 with df =1, 9

p-value = achieved significance < 0.0001. This is a highly unlikely outcome! → Reject H_O. Conclude that the fitted line explains statistically significantly more of the variability in Z=LOGWT than is explained by the intercept-only null hypothesis model.

7. Hypothesis Testing

Straight Line Model: $Y = \beta_0 + \beta_1 X$

1) Overall F-Test

Research Question: Does the fitted model, the $\hat{Y}$, explain significantly more of the total variability of the Y about $\overline{Y}$ than does $\overline{Y}$? **A bit of clarification here, in case you're wondering. When the null hypothesis is true, at least two things happen: (1) $\beta_1 = 0$ and (2) the correct model (the null one) says $Y = \beta_0 + \text{error}$. In this situation, the least squares estimate of β_0 turns out to be $\overline{Y}$ (that seems reasonable, right?)**

Assumptions: As before.

H_O and H_A:

$$H_O: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

Test Statistic:

$$F = \frac{msq(regresion)}{msq(residual)}$$
$$df = 1,(n-2)$$

Evaluation rule:

When the null hypothesis is true, the value of F should be close to 1. Alternatively, when $\beta_1 \neq 0$, the value of F will be LARGER than 1.

Thus, our p-value calculation answers: "What are the chances of obtaining our value of the F or one that is larger if we believe the null hypothesis that $\beta_1 = 0$"?

Calculations:

For our data, we obtain p-value =

$$\text{pr}\left[F_{1,(n-2)} \geq \left| \frac{\text{msq(model)}}{\text{msq(residual)}} \right| b_1 = 0\right] = \text{pr}\left[F_{1,9} \geq 5384.94\right] << .0001$$

Evaluate:

Assumption of the null hypothesis that $\beta_1 = 0$ has led to an extremely unlikely outcome (F-statistic value of 5394.94), with chances of being observed less than 1 chance in 10,000. The null hypothesis is rejected.

Interpret:

We have learned that, at least, the fitted straight line model does a much better job of explaining the variability in Z = LOGWT than a model that allows only for the average LOGWT.

… later … (BIOSTATS 640, Intermediate Biostatistics), we'll see that the analysis does not stop here …

R Users

```
# ANOVA Table: Dependent=z_logwt Predictor=x_age
fit2 <- lm(z_logwt ~ x_age, data=dataset)
anova(fit2)

## Analysis of Variance Table
##
## Response: z_logwt
##           Df Sum Sq Mean Sq F value              Pr(>F)
## x_age      1 4.2211  4.2211  5355.6 0.000000000000008399
## Residuals  9 0.0071  0.0008

summary(fit2)
---  some output not shown --

## F-statistic:  5356 on 1 and 9 DF,  p-value: 0.000000000000008399        F = MSQ(model)/MSQ(residual)
```

Stata Users

```
. * Regress Dependent=z_logwt on Predictor=x_age
. regress z_logwt x_age

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,      9) = 5355.60   = MSQ(model)/MSQ(residual)
       Model |  4.22105734     1  4.22105734           Prob > F      =  0.0000   = p-value for Overall F Test
    Residual |  .007093416     9  .000788157           R-squared     =  0.9983
-------------+------------------------------           Adj R-squared =  0.9981
       Total |  4.22815076    10  .422815076           Root MSE      =  .02807
```

2) Test of the Slope, β_1

Notes -
The overall F test and the test of the slope are equivalent. The test of the slope uses a t-score approach to hypothesis testing It can be shown that { t-score for slope }^{2} = { overall F }

Research Question: Is the slope $\beta_1 = 0$?

Assumptions: As before.

H_O and H_A:

$$H_O : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

Test Statistic:

To compute the t-score, we need an estimate of the standard error of $\hat{\beta}_1$

$$S\hat{E}\left(\hat{\beta}_1\right) = \sqrt{msq(residual)\left[\frac{1}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\right]}$$

Our t-score is therefore:

$$t - score = \left[\frac{(observed)-(\exp ected)}{\hat{se}(\exp ected)}\right] = \left[\frac{(\hat{\beta}_1)-(0)}{\hat{se}(\hat{\beta}_1)}\right]$$

$$df = (n-2)$$

We can find this information in our Stata output. Annotations are in **red**.

```
------------------------------------------------------------------------------
           z |      Coef.   Std. Err.      t = Coef/Std. Err.   P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .1958909   .0026768    73.18 = 0.19589/.002678  0.000     .1898356    .2019462
       _cons |  -2.689255    .030637   -87.78                    0.000     -2.75856   -2.619949
------------------------------------------------------------------------------
```

Recall what we mean by a t-score:

t=73.38 says "the estimated slope is estimated to be 73.38 standard error units away from the null hypothesis expected value of zero".

Check that { t-score }^{2} = { Overall F }:

[73.38]^2 = 5384.62 which is close.

Evaluation rule:

When the null hypothesis is true, the value of t should be close to zero.
Alternatively, when $\beta_1 \neq 0$, the value of t will be DIFFERENT from 0.

Here, our p-value calculation answers: "Under the assumption of the null hypothesis that $\beta_1 = 0$, what were our chances of obtaining a t-statistic value 73.38 standard error units away from its null hypothesis expected value of zero"?

Calculations:

For our data, we obtain p-value =

$$2\,pr\left[t_{(n-2)} \geq \left| \frac{\hat{\beta}_1 - 0}{\hat{se}\left(\hat{\beta}_1\right)} \right| \right] = 2\,pr\left[t_9 \geq 73.38 \right] << .0001$$

Evaluate:

Under the null hypothesis that $\beta_1 = 0$, the chances of obtaining a t-score value that is 73.38 or more standard error units away from the expected value of 0 is less than 1 chance in 10,000.

Interpret:

The inference is the same as that for the overall F test. The fitted straight line model does a statistically significantly better job of explaining the variability in LOGWT than the sample mean.

R Users

```
# TEST OF SLOPE: Dependent=z_logwt Predictor=x_age
fit2 <- lm(z_logwt ~ x_age, data=dataset)
summary(fit2)

--  some output not shown --
##
## Coefficients:
##              Estimate Std. Error t value           Pr(>|t|)
## (Intercept) -2.689255   0.030637  -87.78 0.0000000000000164
## x_age        0.195891   0.002677   73.18 0.0000000000000840
```

Stata Users

```
. * TEST OF SLOPE: Dependent=z_logwt Predictor=x_age
. regress z_logwt x_age

--- some output not shown --
------------------------------------------------------------------------------
     z_logwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       x_age |   .1958909   .0026768    73.18   0.000     .1898356    .2019462
       _cons |  -2.689255    .030637   -87.78   0.000     -2.75856   -2.619949
```

3) Test of the Intercept, β_0

This addresses the question: Does the straight-line relationship passes through the origin? It is rarely of interest.

Research Question: Is the intercept $\beta_0 = 0$?

Assumptions: As before.

H_O and H_A:

$$H_O : \beta_0 = 0$$
$$H_A : \beta_0 \neq 0$$

Test Statistic:

To compute the t-score for the intercept, we need an estimate of the standard error of $\hat{\beta}_0$

$$S\hat{E}\left(\hat{\beta}_0\right) = \sqrt{msq(residual)\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\right]}$$

Our t-score is therefore:

$$t - score = \left[\frac{(observed) - (\exp ected)}{s\hat{e}(\exp ected)}\right] = \left[\frac{\left(\hat{\beta}_0\right) - (0)}{s\hat{e}\left(\hat{\beta}_0\right)}\right]$$
$$df = (n - 2)$$

R Users

```
# TEST OF INTERCEPT: Dependent=z_logwt Predictor=x_age
fit2 <- lm(z_logwt ~ x_age, data=dataset)
summary(fit2)

--  some output not shown --
##
## Coefficients:
##              Estimate Std. Error t value           Pr(>|t|)
## (Intercept) -2.689255   0.030637  -87.78 0.0000000000000164
## x_age        0.195891   0.002677   73.18 0.0000000000000840
```

Stata Users

```
. * TEST OF INTERCEPT: Dependent=z_logwt Predictor=x_age
. regress z_logwt x_age

--- some output not shown --
------------------------------------------------------------------------------
     z_logwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       x_age |   .1958909   .0026768    73.18   0.000     .1898356    .2019462
       _cons |  -2.689255    .030637   -87.78   0.000     -2.75856   -2.619949
------------------------------------------------------------------------------
```

Here, **t = -87.78** says "the estimated intercept is estimated to be 87.78 standard error units away from its null hypothesis expected value of zero".

Evaluation rule:

When the null hypothesis is true, the value of t should be close to zero.
Alternatively, when $\beta_0 \neq 0$, the value of t will be DIFFERENT from 0.

Our p-value calculation answers: "Under the assumption of the null hypothesis that $\beta_0 = 0$, what were our chances of obtaining a t-statistic value 87.78 standard error units away from its null hypothesis expected value of zero"?

Calculations:

p-value =

$$2\,pr\left[t_{(n-2)} \geq \left|\frac{\hat{\beta}_0 - 0}{\hat{se}\left(\hat{\beta}_0\right)}\right|\right] = 2\,pr\left[t_9 \geq 87.78\right] << .0001$$

Evaluate:

Under the null hypothesis that the line passes through the origin, that $\beta_0 = 0$, the chances of obtaining a t-score value that is 87.78 or more standard error units away from the expected value of 0 is less than 1 chance in 10,000, again prompting statistical rejection of the null hypothesis.

Interpret:

The inference is that there is statistically significant evidence that the straight-line relationship between Z=LOGWT and X=AGE does ***not*** pass through the origin.

8. Confidence Interval Estimation
Straight Line Model: $Y = \beta_0 + \beta_1 X$

The confidence intervals here have the usual 3 elements (for review, see again Units 8, 9 & 10):

1) Best single guess (estimate)
2) Standard error of the best single guess (SE[estimate])
3) Confidence coefficient: This will be a percentile from the Student t distribution with df=(n-2)

We might want confidence interval estimates of the following 4 parameters:

(1) Slope
(2) Intercept
(3) Mean of subset of population for whom $X=x_0$
(4) Individual response for person for whom $X=x_0$

---

1) SLOPE **estimate =** $\hat{\beta}_1$

$$\hat{se}\left(\hat{b}_1\right) = \sqrt{\text{msq(residual)}\frac{1}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}} = \sqrt{(\text{mse})\frac{1}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}}$$

---

2) INTERCEPT **estimate =** $\hat{\beta}_0$

$$\hat{se}\left(\hat{b}_0\right) = \sqrt{\text{msq(residual)}\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\right]} = \sqrt{(\text{mse})\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\right]}$$

---

3) MEAN at X=x$_0$ **estimate =** $\hat{Y}_{X=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$$\hat{se} = \sqrt{\text{msq(residual)}\left[\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]} = \sqrt{(\text{mse})\left[\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]}$$

---

4) INDIVIDUAL with X=x$_0$ **estimate =** $\hat{Y}_{X=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$$\hat{se} = \sqrt{\text{msq(residual)}\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]} = \sqrt{(\text{mse})\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]}$$

---

1) Confidence Interval for SLOPE
Z=LOGWT to X=AGE.

R Users

```
fit2 <- lm(z_logwt ~ x_age, data=dataset)
confint(fit2, level=.95)

##                  2.5 %     97.5 %
## (Intercept) -2.7585602 -2.6199489
## x_age        0.1898356  0.2019462
```

Stata Users

```
. regress z_logwt x_age

 --- some output not shown –

------------------------------------------------------------------------------
     z_logwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       x_age |   .1958909   .0026768    73.18   0.000     .1898356    .2019462
       _cons |  -2.689255    .030637   -87.78   0.000     -2.75856   -2.619949
------------------------------------------------------------------------------
```

By Hand

<u>95% Confidence Interval for the Slope, β_1</u>

1) Best single guess (estimate) = $\hat{\beta}_1 = 0.19589$

2) Standard error of the best single guess (SE[estimate]) = $se\left(\hat{\beta}_1\right) = 0.00268$

3) Confidence coefficient = 97.5th percentile of Student t = $t_{.975,df=9} = 2.26$

95% Confidence Interval for Slope β_1 = Estimate ± (confidence coefficient)*SE

= 0.19589 ± (2.26)(0.00268)
= (0.1898, 0.2019)

2) Confidence Interval for INTERCEPT
Z=LOGWT to X=AGE.

R Users

```
fit2 <- lm(z_logwt ~ x_age, data=dataset)
confint(fit2, level=.95)

##                  2.5 %     97.5 %
## (Intercept) -2.7585602 -2.6199489
## x_age        0.1898356  0.2019462
```

Stata Users

```
. regress z_logwt x_age

 --- some output not shown –

------------------------------------------------------------------------------
     z_logwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       x_age |   .1958909   .0026768    73.18   0.000     .1898356    .2019462
       _cons |  -2.689255    .030637   -87.78   0.000     -2.75856   -2.619949
------------------------------------------------------------------------------
```

By Hand

1) Best single guess (estimate) = $\hat{\beta}_0 = -2.68925$

2) Standard error of the best single guess (SE[estimate]) = $se\left(\hat{\beta}_0\right) = 0.03064$

3) Confidence coefficient = 97.5th percentile of Student t = $t_{.975, df=9} = 2.26$

95% Confidence Interval for Slope β_0 = Estimate $\pm$ (confidence coefficient)*SE

= -2.68925 $\pm$ (2.26)(0.03064)

= (-2.7585,-2.6200)

3) Confidence Interval for MEANS
Z=LOGWT to X=AGE.

R Users

```
# Confidence Intervals for MEAN at each value of X Dependent=z_logwt Predictor=x_age

age=c(6,7,8,9,10,11,12,13,14,15,16)
ymean <- as.data.frame.matrix(predict(fit2,data.frame(x_age = age), level = 0.95, interval = "confidence")
)
CI_MEANS <- data.frame(age, ymean)
CI_MEANS$predicted <- CI_MEANS$fit
CI_MEANS$lwr_mean <- CI_MEANS$lwr
CI_MEANS$upr_mean <- CI_MEANS$upr
CI_MEANS$fit <- NULL
CI_MEANS$lwr <- NULL
CI_MEANS$upr <- NULL
CI_MEANS

##    age   predicted    lwr_mean    upr_mean
## 1    6 -1.51390909 -1.54973251 -1.47808568
## 2    7 -1.31801818 -1.34889408 -1.28714228
## 3    8 -1.12212727 -1.14852155 -1.09573300
## 4    9 -0.92623636 -0.94889308 -0.90357965
## 5   10 -0.73034545 -0.75042849 -0.71026242
## 6   11 -0.53445455 -0.55360297 -0.51530612
## 7   12 -0.33856364 -0.35864667 -0.31848060
## 8   13 -0.14267273 -0.16532944 -0.12001601
## 9   14  0.05321818  0.02682391  0.07961246
## 10  15  0.24910909  0.21823319  0.27998499
## 11  16  0.44500000  0.40917659  0.48082341
```

Stata Users

```
. regress z_logwt x_age
. * save fitted values xb (this is internal to Stata) to a new variable called zhat
. predict zhat, xb
. ** Obtain SE for MEAN of Z at each X (this is internal to Stata) to a new variable called semeanz
. predict semeanz, stdp
. ** Obtain confidence coefficient = 97.5th percentile of T on df=9
. generate tmult=invttail(9,.025)
. **  Generate lower and upper 95% CI limits for MEAN of Z at Each X
. generate lowmeanz=zhat -tmult*semeanz
. generate highmeanz=zhat+tmult*semeanz
. list x z zhat lowmeanz highmeanz, clean
          x        z        zhat    lowmeanz   highmeanz
  1.      6   -1.538   -1.513909   -1.549733   -1.478086
  2.      7   -1.284   -1.318018   -1.348894   -1.287142
  3.      8   -1.102   -1.122127   -1.148522   -1.095733
  4.      9    -.903   -.9262364   -.9488931   -.9035797
  5.     10    -.742   -.7303454   -.7504284   -.7102624
  6.     11    -.583   -.5344545   -.5536029   -.5153061
  7.     12    -.372   -.3385637   -.3586467   -.3184806
  8.     13    -.132   -.1426727   -.1653294    -.120016
  9.     14     .053    .0532182    .0268239    .0796125
 10.     15     .275    .2491091    .2182332     .279985
 11.     16     .449        .445    .4091766    .4808234
```

4) Confidence Interval for INDIVIDUAL PREDICTIONS Z=LOGWT to X=AGE.

R Users

```
# Confidence Intervals for INDIVIDUAL PREDICTION at each value of X Dependent=z_logwt Predictor=x_age

age=c(6,7,8,9,10,11,12,13,14,15,16)
yindividual <- as.data.frame.matrix(predict(fit2,data.frame(x_age = age), level = 0.95, interval = "predic
tion"))
CI_IND <- data.frame(age, yindividual)
CI_IND$predicted <- CI_IND$fit
CI_IND$lwr_individual <- CI_IND$lwr
CI_IND$upr_individual <- CI_IND$upr
CI_IND$fit <- NULL
CI_IND$lwr <- NULL
CI_IND$upr <- NULL
CI_IND

##    age   predicted lwr_individual upr_individual
## 1    6 -1.51390909    -1.58682410    -1.44099408
## 2    7 -1.31801818    -1.38863407    -1.24740230
## 3    8 -1.12212727    -1.19090183    -1.05335271
## 4    9 -0.92623636    -0.99366491    -0.85880782
## 5   10 -0.73034545    -0.79695334    -0.66373757
## 6   11 -0.53445455    -0.60078662    -0.46812247
## 7   12 -0.33856364    -0.40517152    -0.27195575
## 8   13 -0.14267273    -0.21010127    -0.07524418
## 9   14  0.05321818    -0.01555638     0.12199274
## 10  15  0.24910909     0.17849320     0.31972498
## 11  16  0.44500000     0.37208499     0.51791501
```

Stata Users

```
. regress z_logwt x_age
. *Save fitted values to a new variable called zhat
. predict zhat, xb
. ** Obtain SE for INDIVIDUAL PREDICTION of Z at given X (internal to Stata) to a new variable sepredictz
. predict sepredictz, stdf
. ** Obtain confidence coefficient = 97.5th percentile of T on df=9
. generate tmult=invttail(9,.025)
. **  Generate lower and upper 95% CI limits for INDIVIDUAL PREDICTED Z at Each X
. generate lowpredictz=zhat-tmult*sepredictz
. generate highpredictz=zhat+tmult*sepredictz
. ***  List Individual Predictions with 95% CI Limits
. list x z zhat lowpredictz highpredictz, clean
              x         z        zhat   lowpred~z   highpre~z
  1.          6    -1.538   -1.513909   -1.586824   -1.440994
  2.          7    -1.284   -1.318018   -1.388634   -1.247402
  3.          8    -1.102   -1.122127   -1.190902   -1.053353
  4.          9     -.903   -.9262364   -.9936649   -.8588079
  5.         10     -.742   -.7303454   -.7969533   -.6637375
  6.         11     -.583   -.5344545   -.6007866   -.4681225
  7.         12     -.372   -.3385637   -.4051715   -.2719558
  8.         13     -.132   -.1426727   -.2101013   -.0752442
  9.         14      .053    .0532182   -.0155564    .1219927
 10.         15      .275    .2491091    .1784932     .319725
 11.         16      .449        .445     .372085     .517915
```

9. Introduction to Correlation

Definition of Correlation

A correlation coefficient is a measure of the association between two paired random variables (e.g. height and weight).

The **Pearson product moment correlation**, in particular, is a measure of the strength of the ***straight-line*** relationship between the two random variables.

Another correlation measure (not discussed here) is the **Spearman correlation**. It is a measure of the strength of the ***monotone increasing (or decreasing)*** relationship between the two random variables. The Spearman correlation is a non-parametric (meaning model free) measure. It is introduced in BIOSTATS 640, *Intermediate Biostatistics*.

Formula for the Pearson Product Moment Correlation ρ

- Population product moment correlation = **ρ**
- Sample based estimate = **r**.
- Some preliminaries:

(1) Suppose we are interested in the correlation between X and Y

(2) $\hat{\text{cov}}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n\text{-}1)} = \frac{S_{xy}}{(n\text{-}1)}$ This is the covariance(X,Y)

(3) $\hat{\text{var}}(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n\text{-}1)} = \frac{S_{xx}}{(n\text{-}1)}$ and similarly

(4) $\hat{\text{var}}(Y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{(n\text{-}1)} = \frac{S_{yy}}{(n\text{-}1)}$

Formula for Estimate of Pearson Product Moment Correlation from a Sample

$$\hat{\rho} = r = \frac{\hat{\text{cov}}(x,y)}{\sqrt{\hat{\text{var}}(x)\hat{\text{var}}(y)}}$$

$$= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

If you absolutely have to do it by hand, an equivalent (more calculator/excel friendly formula) is

$$\hat{\rho} = r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}\sqrt{\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]}}$$

- The correlation r can take on values **between 0 and 1 only**
- Thus, the correlation coefficient is said to be **dimensionless** – it is independent of the units of x or y.
- **Sign** of the correlation coefficient (positive or negative) = **Sign** of the estimated slope $\hat{\beta}_1$.

There is a relationship between the slope of the straight line, $\hat{\beta}_1$, and the estimated correlation r.

Relationship between slope $\hat{\beta}_1$ and the sample correlation r

Tip! This is very handy…

Because $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ **and** $r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$

A little algebra reveals that

$$r = \left[\frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} \right] \hat{\beta}_1$$

Thus, beware!!!

- **It is possible to have a very large (positive or negative) r might accompanying a very non-zero slope, inasmuch as**
  - **A very large r might reflect a very large S_{xx}, all other things equal**
  - **A very large r might reflect a very small S_{yy}, all other things equal.**

10. Hypothesis Test of Correlation

The null hypothesis of zero correlation is equivalent to the null hypothesis of zero slope.

Research Question: Is the correlation $\rho = 0$? Is the slope $\beta_1 = 0$?

Assumptions: As before.

H_O and H_A:

$$H_O : \rho = 0$$
$$H_A : \rho \neq 0$$

Test Statistic:

A little algebra (not shown) yields a very nice formula for the t-score that we need.

$$t - score = \left[\frac{\text{r} \sqrt{\text{(n-2)}}}{\sqrt{1 - \text{r}^2}} \right]$$

$$df = (n-2)$$

We can find this information in our output. Recall the first example and the model of Z=LOGWT to X=AGE:

The Pearson Correlation, r, is the $\sqrt{\text{R-squared}}$ in the output.

```
      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) = 5355.60
       Model |  4.22105734     1  4.22105734           Prob > F      =  0.0000
    Residual |  .007093416     9  .000788157           R-squared     =  0.9983
-------------+------------------------------           Adj R-squared =  0.9981
       Total |  4.22815076    10  .422815076           Root MSE      =  .02807
```

Pearson Correlation, $r = \sqrt{0.9983} = 0.9991$

Substitution into the formula for the t-score yields

$$t-score=\left[\frac{r\sqrt{(n\text{-}2)}}{\sqrt{1-r^2}}\right]=\left[\frac{.9991\sqrt{9}}{\sqrt{1\text{-}.9983}}\right]=\left[\frac{2.9974}{.0412}\right]=72.69$$

Note: The value .9991 in the numerator is $r=\sqrt{R^2}=\sqrt{.9983}=.9991$

This is very close to the value of the t-score that was obtained for testing the null hypothesis of zero slope. The discrepancy is probably rounding error. I did the calculations on my calculator using 4 significant digits. Stata probably used more significant digits - cb.