

Unit 2. Regression and Correlation

“ ‘Don’t let us quarrel,’ the White Queen said in an anxious tone. ‘What is the cause of lightning?’ ‘The cause of lightning,’ Alice said very decidedly, for she felt quite certain about this, ‘is the thunder-oh no!’ , she hastily corrected herself. ‘I meant the other way.’ ‘It’s too late to correct it,’ said the Red Queen: ‘when you’ve once said a thing, that fixes it, and you must take the consequences.’ “

- Carroll

Menopause heralds a complex interplay of hormonal and physiologic changes. Some are temporary discomforts (e.g., hot flashes, sleep disturbances, depression) while others are long-term changes that increase the risk of significant chronic health conditions, bone loss and osteoporosis in particular. Recent observations of an association between **depressive symptoms** and **low bone mineral density (BMD)** raise the intriguing possibility that alleviation of depression might confer a risk benefit with respect to bone mineral density loss and osteoporosis. However, the finding of an association in a simple (one predictor) linear regression model analysis has multiple possible explanations, only one of which is causal. Others include, but are not limited to: (1) the apparent association is an artifact of the confounding effects of exercise, body fat, education, smoking, etc; (2) there is no relationship and we have observed a chance event of low probability (it can happen!); (3) the pathway is the other way around (low BMD causes depressive symptoms), albeit highly unlikely; and/or (4) the finding is spurious due to study design flaws (selection bias, misclassification, etc).

In settings where multiple, related predictors are associated with the outcome of interest, multiple predictor linear regression analysis allows us to study the joint relationships among the multiple predictors (depressive symptoms, exercise, body fat, etc) and a single continuous outcome (BMD). In this example, we might be especially interested in using multiple predictor linear regression to isolate the effect of depressive symptoms on BMD, holding all other predictors constant (**adjustment**). Or, we might want to investigate the possibility of synergism or **interaction**.

Table of Contents

Topic	Learning Objectives	3
	1. <u>Simple Linear Regression</u>	<u>4</u>
	a. Definition of the Linear Regression Model	4
	b. Estimation	12
	c. The Analysis of Variance Table	20
	d. Assumptions for the Straight Line Regression	25
	e. Hypothesis Testing	28
	f. Confidence Interval Estimation	34
	2. <u>Introduction to Correlation</u>	<u>37</u>
	a. Pearson Product Moment Correlation	37
	b. Hypothesis Test for Correlation	40
	3. <u>Multivariable Regression</u>	<u>42</u>
	a. Introduction, Indicator and Design Variables	42
	b. The Analysis of Variance Table	46
	c. The Partial F Test	48
	d. Multiple Partial Correlation	50
	4. <u>Multivariable Model Development</u>	<u>52</u>
	a. Introduction	52
	b. Example – Human p53 and Breast Cancer Risk.....	53
	c. Guidelines for Multivariable Analyses of Large Data Sets	62
	5. <u>Goodness-of-Fit and Regression Diagnostics</u>	<u>64</u>
	a. Introduction and Terminology.....	64
	b. Assessment of Normality.....	71
	c. Cook-Weisberg Test of Heteroscedasticity	75
	d. Method of Fractional Polynomials	76
	e. Ramsay Test for Omitted Variables	78
	f. Residuals, Leverage, & Cook’s Distance	79

1. Learning Objectives

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

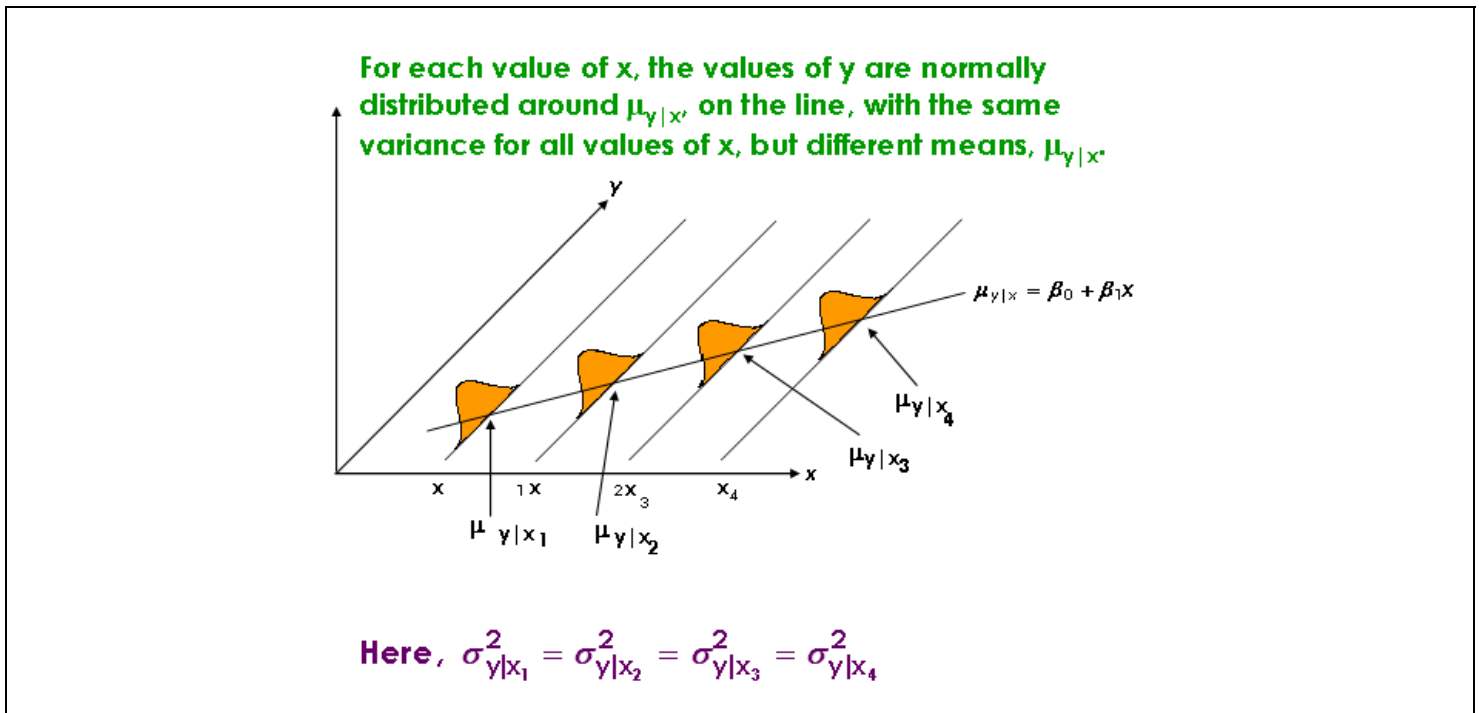
- Explain the concepts of association, causation, confounding, mediation, and effect modification;
- Construct and interpret a scatter plot with respect to: evidence of association, assessment of linearity, and the presence of outlying values;
- State the multiple predictor linear regression model and the assumptions necessary for its use;
- Perform and interpret the Shapiro-Wilk and Kolmogorov-Smirnov tests of normality;
- Explain the relevance of the normal probability distribution;
- Explain and interpret the coefficients (and standard error) and analysis of variance tables outputs of a single or multiple predictor regression model estimation;
- Explain and compare crude versus adjusted estimates (betas) of association;
- Explain and interpret regression model estimates of effect modification (interaction);
- Explain and interpret overall and adjusted R-squared measures of association;
- Explain and interpret overall and partial F-tests;
- Draft an analysis plan for a multiple predictor regression model analysis; and
- Explain and interpret selected regression model diagnostics: residuals, leverage, and Cook's distance.

1. Simple Linear Regression

a. Definition of the Linear Regression Model

Simple Linear Regression

A simple linear regression model is a particular model of how the mean μ (the average value) of **one continuous outcome random variable Y** (e.g. $Y = \text{bone mineral density}$) varies, depending on the value of a single (usually continuous) predictor variable X (e.g. $X = \text{depressive symptoms}$). Specifically, it says that the average values of the outcome variable, as X changes, lie on a straight line (“regression line”).



The estimation and hypothesis testing involved are extensions of ideas and techniques that we have already seen. In linear regression,

- ◆ we observe an outcome or dependent variable “Y” at several levels of the independent or predictor variable “X” (there may be more than one predictor “X” as seen later).
- ◆ A linear regression model assumes that the values of the predictor “X” have been fixed in advance of observing “Y”.
- ◆ However, this is not always the reality. Often “Y” and “X” are observed jointly and are both random variables.

Correlation

Correlation considers the association of **two random** variables, Y and X.

- ◆ 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.

We develop the linear regression model analysis for a simple example involving one predictor and one outcome.

Example.

Source: Kleinbaum, Kupper, and Muller 1988

Suppose we have observations of age and weight for n=11 chicken embryos. The predictor of interest is X=AGE. The outcome of interest is weight. For purposes of illustration, suppose we are interested in two models of weight. In one, the outcome variable is Y=WT. In the other, the outcome is the logarithm of weight, Z=LOGWT.

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_1, Y_1) where X=AGE and Y=WT
 $(X_1, Y_1) = (6, .029) \dots (X_{11}, Y_{11}) = (16, 2.812)$ and
- ◆ equivalently, 11 pairs of (X_1, Z_1) where X=AGE and Z=LOGWT
 $(X_1, Z_1) = (6, -1.538) \dots (X_{11}, Z_{11}) = (16, 0.449)$

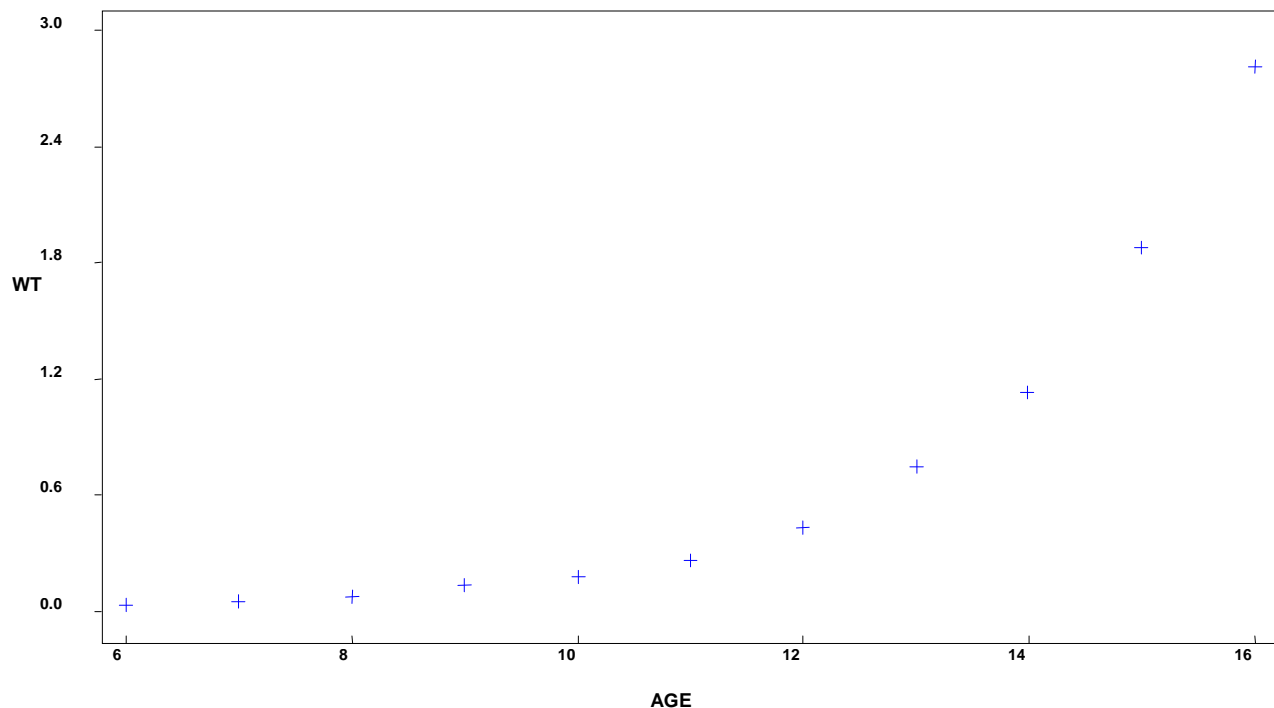


Though simple, it helps to be clear in the research question

- How does weight change with age? Does it change linearly?
- In the language of analysis of variance we are asking the following:
Can the variability in weight be explained, to a significant extent, by variations in age?
- What is a “good” functional form that relates age to weight?

Always begin with a scatter plot of the data! Plot the predictor X on the horizontal and the outcome Y on the vertical. A graph allows you to see things that you cannot see in the numbers alone: range, patterns, outliers, etc. Here, let's begin with a plot of X=AGE versus Y=WT

Scatter Plot of WT vs AGE



What to look for in a scatter plot of the data:

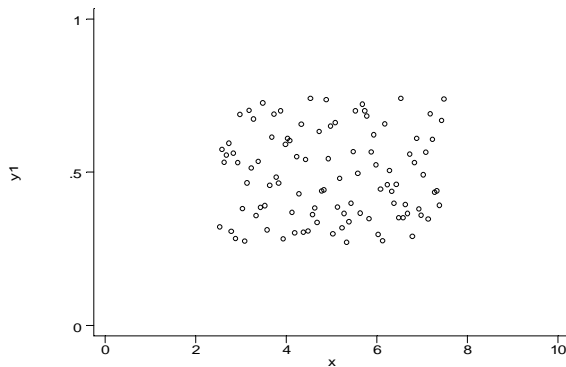
- ◆ The average and median of X
- ◆ The range and pattern of variability in X
- ◆ The average and median of Y
- ◆ The range and pattern of variability in Y
- ◆ The nature of the relationship between X and Y
- ◆ The strength of the relationship between X and Y
- ◆ The identification of any points that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

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

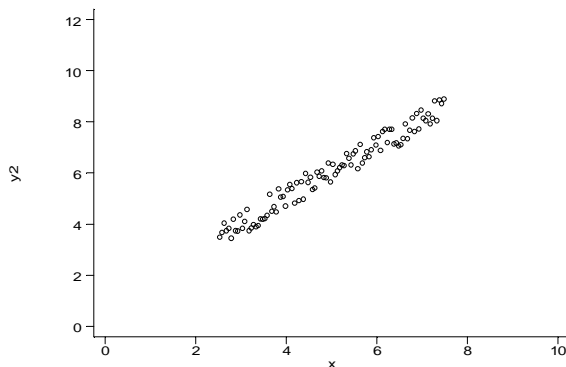
Example – age (X) and weight (Y) of chicken embryos:

- ◆ 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

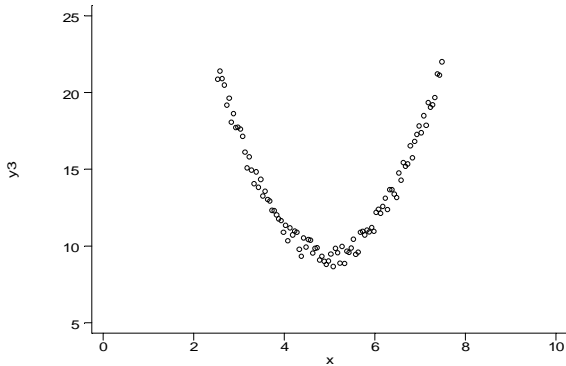
We might have gotten any of a variety of scatter plots:



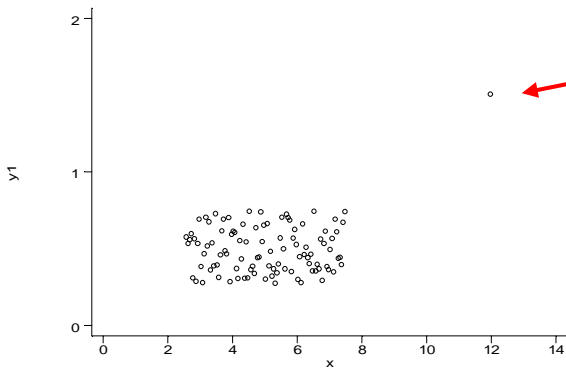
No relationship between X and Y



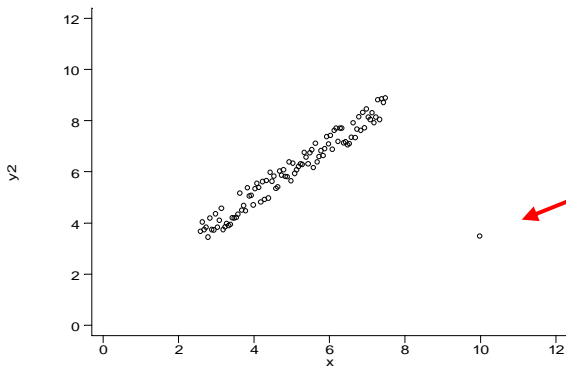
Linear relationship between X and Y



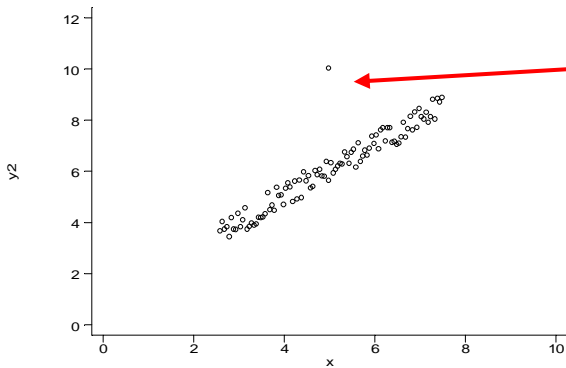
Non-linear relationship between X and Y



Note the arrow pointing to the outlying point
Fit of a linear model will yield estimated slope that is spuriously non-zero.



Note the arrow pointing to the outlying point
Fit of a linear model will yield an estimated slope that is spuriously near zero.

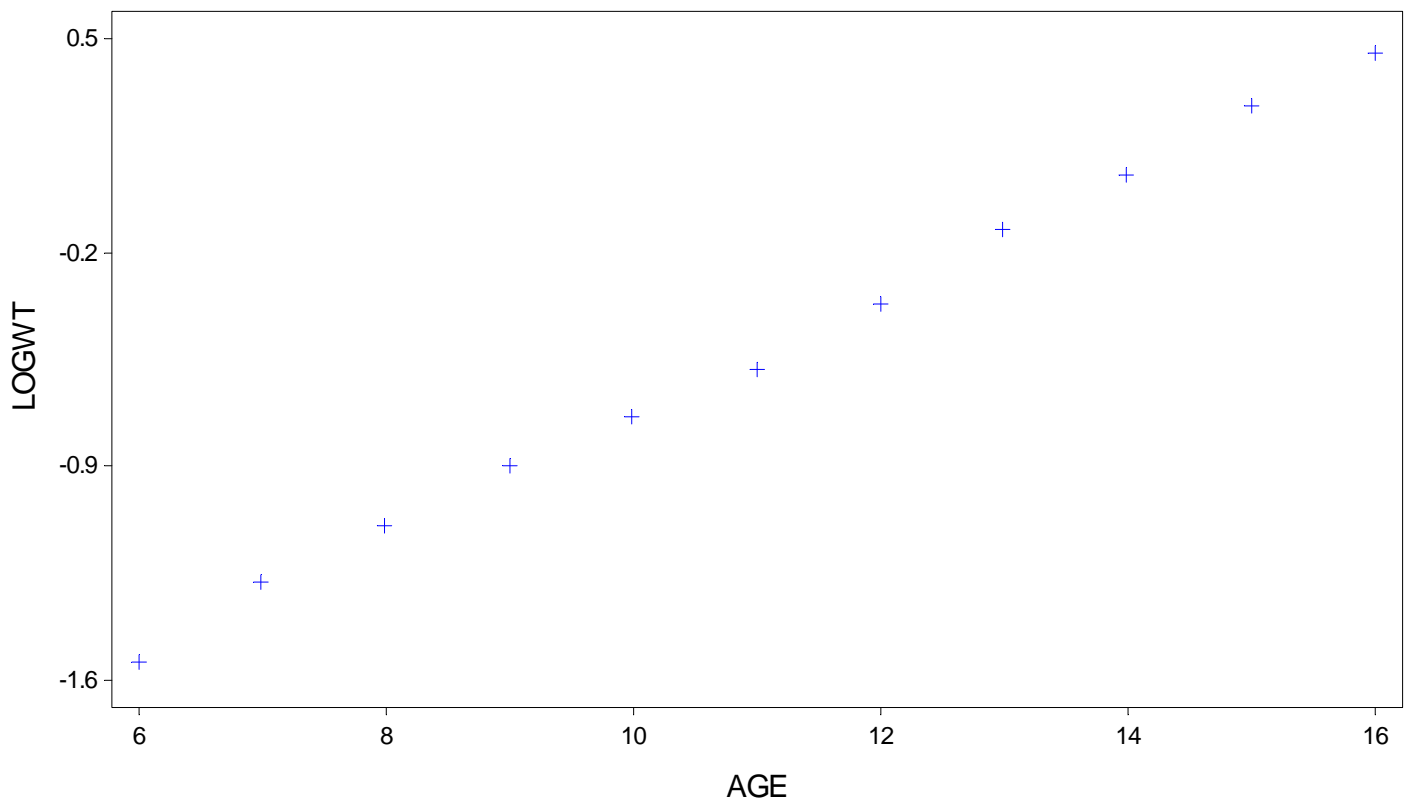


Note the arrow pointing to the outlying point
Fit of a linear model will yield an estimated slope that is spuriously high.

Example, continuous – age (X) and logweight (Z) of chicken embryos:

The X-Y plot on page 6 is rather “bowl” shaped. Here we consider an X-Z scatter plot. It is much more linear looking, suggesting that perhaps a better model relates the logarithm of WT (Z) to AGE:

Scatter Plot of LOGWT vs AGE

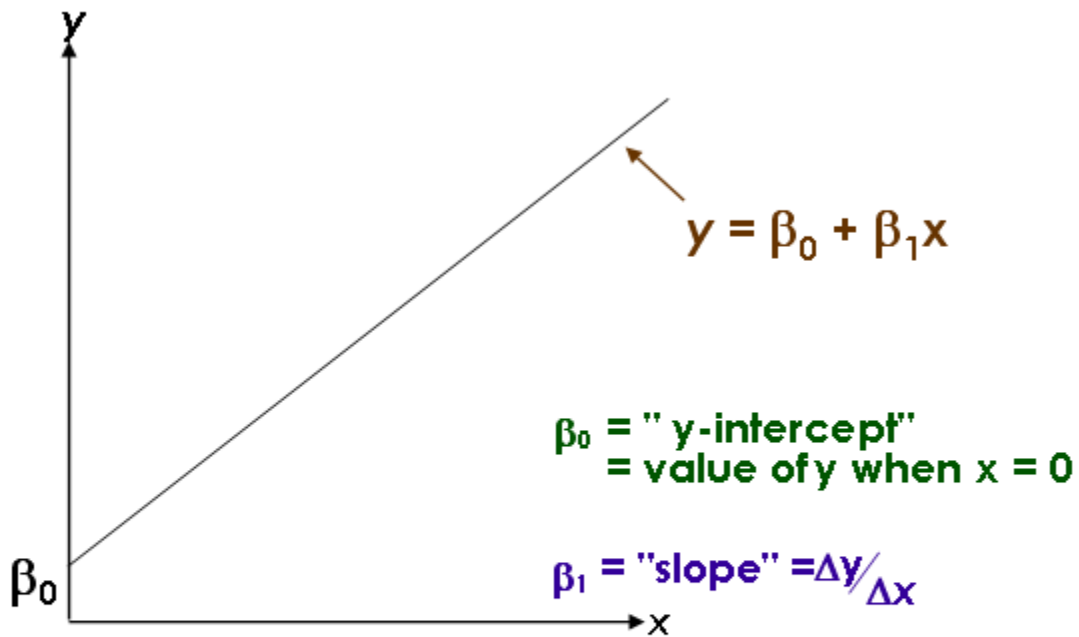


We’ll investigate two models.

- 1) $WT = \beta_0 + \beta_1 \text{ AGE}$
- 2) $\text{LOGWT} = \beta_0 + \beta_1 \text{ AGE}$



[A little review of your high school introduction to straight line relationships](#)



$\beta_0 = \text{"y-intercept"} = \text{value of } y \text{ when } x = 0$

$\beta_1 = \text{"slope"} = \frac{\Delta y}{\Delta x} = (\text{change in } y) / (\text{change in } x)$

Slope > 0	Slope = 0	Slope < 0

Definition of the Straight Line Model

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

Population	Sample
$Y = \beta_0 + \beta_1 X + \varepsilon$	$Y = \hat{\beta}_0 + \hat{\beta}_1 X + e$
$Y = \beta_0 + \beta_1 X$ is the relationship in the population. It is measured with error.	$\hat{\beta}_0$, $\hat{\beta}_1$, and e are our guesses of β_0 , β_1 and ε
ε = measurement error	e = residual
We do NOT know the value of β_0 nor β_1 nor ε	We do have values of $\hat{\beta}_0$, $\hat{\beta}_1$ and e
	The values of $\hat{\beta}_0$, $\hat{\beta}_1$ and e are obtained by the method of <u>least squares estimation</u> .
	To see if $\hat{\beta}_0 \approx \beta_0$ and $\hat{\beta}_1 \approx \beta_1$ we perform <u>regression diagnostics</u> .

A little 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



b. Estimation

There are a variety of methods for obtaining estimates of β_0 and β_1 .

In this course, we will consider two of them, maximum likelihood estimation and least squares estimation.

Maximum Likelihood Estimation - This requires use of a probability distribution model. For example, we might assume that the outcome variable Y is distributed normal, with mean values that lie on the regression line. Maximum likelihood estimation chooses estimates of β_0 and β_1 that, when applied to the data, gives us the largest value possible for the likelihood of the data that was actually observed.

Least Squares Estimation - **NO probability distribution model required here!** Least squares estimation chooses estimates of β_0 and β_1 that yield the smallest total of vertical distances (observed to predicted)

When the outcome variable Y is distributed **normal**,

Maximum Likelihood Estimation = Least Squares Estimation

How Least Squares Estimation works.

Theoretically, we could draw lots of possible lines through the X-Y scatter of the data points. Which one is the “closest”? And what do we mean by “close” anyway? Consider the following:

Y = observed

\hat{Y} = predicted, meaning that $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$(Y - \hat{Y})^2$ = vertical distance between observed outcome and predicted outcome

In least squares estimation, “close” means the following.

- ◆ We’d like the observed Y and its corresponding prediction \hat{Y} to be as close as possible.
- ◆ This is the same as wanting $(Y - \hat{Y})^2$ to be as small as possible
- ◆ It’s not possible to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that it minimizes

$(Y_1 - \hat{Y}_1)^2$ and minimizes individually

$(Y_2 - \hat{Y}_2)^2$ and minimizes individually

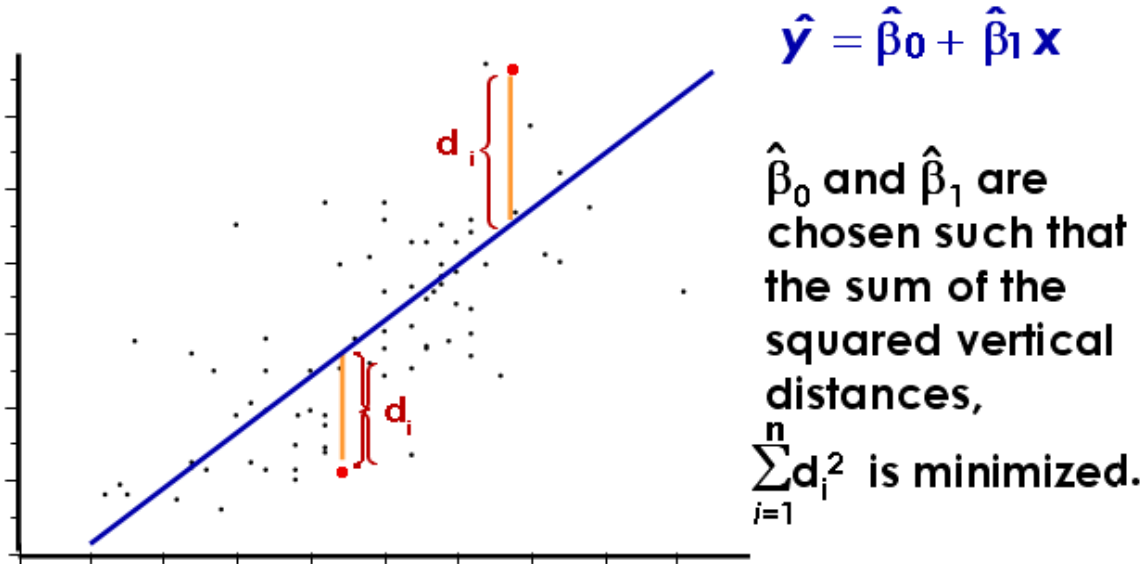
....

$(Y_n - \hat{Y}_n)^2$

- ◆ So, instead, we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ that makes their total as small as possible

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

How Least Squares Estimation works – a picture.



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 $d_i = (y_i - \hat{y}_i)$.

The total (a total of squared differences) that we want to minimize has a variety of names.

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i])^2 \text{ is variously called:}$$

- ◆ residual sum of squares
- ◆ sum of squares about the regression line
- ◆ sum of squares due error (SSE)

For the calculus lover, A little calculus yields the solution for the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

- ◆ Consider $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i])^2$
- ◆ *Step 1:* Differentiate with respect to $\hat{\beta}_1$
Set derivative equal to 0 and solve.
- ◆ *Step 2:* Differentiate with respect to $\hat{\beta}_0$
Set derivative equal to 0, insert $\hat{\beta}_1$ and solve.

β_1 is the unknown slope in the population

- Its estimate is denoted $\hat{\beta}_1$ or b_1

β_0 is the unknown intercept in the population

- Its estimate is denoted $\hat{\beta}_0$ or b_0

How to use some summation calculations to obtain these estimates

Calculate

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

Review. These expressions make use of a special notation called the “summation notation”.

The capitol “S” indicates “summation”.

In S_{xy} , the first subscript “x” is saying $(x - \bar{x})$.

The second subscript “y” is saying $(y - \bar{y})$.

$$S_{xy} = \sum (X - \bar{X})(Y - \bar{Y})$$

S subscript x subscript y

Formulae for Estimated Slope and Intercept

Slope	$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}}$ $= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) / (n-1)}{\sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)} = \frac{\text{c\hat{ov}}(X, Y)}{\text{v\hat{ar}}(X)}$	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$
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	$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$ $= \frac{\text{c\hat{ov}}(X, Y)}{\text{v\hat{ar}}(X)}$	<p>The linear movement in Y with linear movement in X is measured relative to the variability in X.</p> <p>$\hat{\beta}_1 = 0$ says: With a unit change in X, overall there is a 50-50 chance that Y increases versus decreases.</p> <p>$\hat{\beta}_1 \neq 0$ says: With a unit increase in X, Y increases also ($\hat{\beta}_1 > 0$) or Y decreases ($\hat{\beta}_1 < 0$).</p>
Intercept	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$	<p>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.</p>

Illustration in Stata
Command.

```
. regress wt age
```

Partial listing of output – annotations in red

wt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.2350727 = slope = b_1	.0459425	5.12	0.001	.1311437	.3390018
_cons	-1.884527 = intercept = b_0	.5258354	-3.58	0.006	-3.07405	-.695005

The fitted line is therefore $WT^{\hat{}} = -1.88453 + 0.23507 * AGE$

Overlay of Least Squares Line on the X-Y Scatter Plot	
	<p>As we might have guessed, the straight line model may not be the best choice.</p> <p>However ... it's worth noting that the “bowl” shape of the scatter plot does have a linear component.</p> <p>→ So ... without the plot, we might have believed the straight line fit is okay.</p>

Illustration of straight line model fit to Z=LOGWT versus X=AGE.

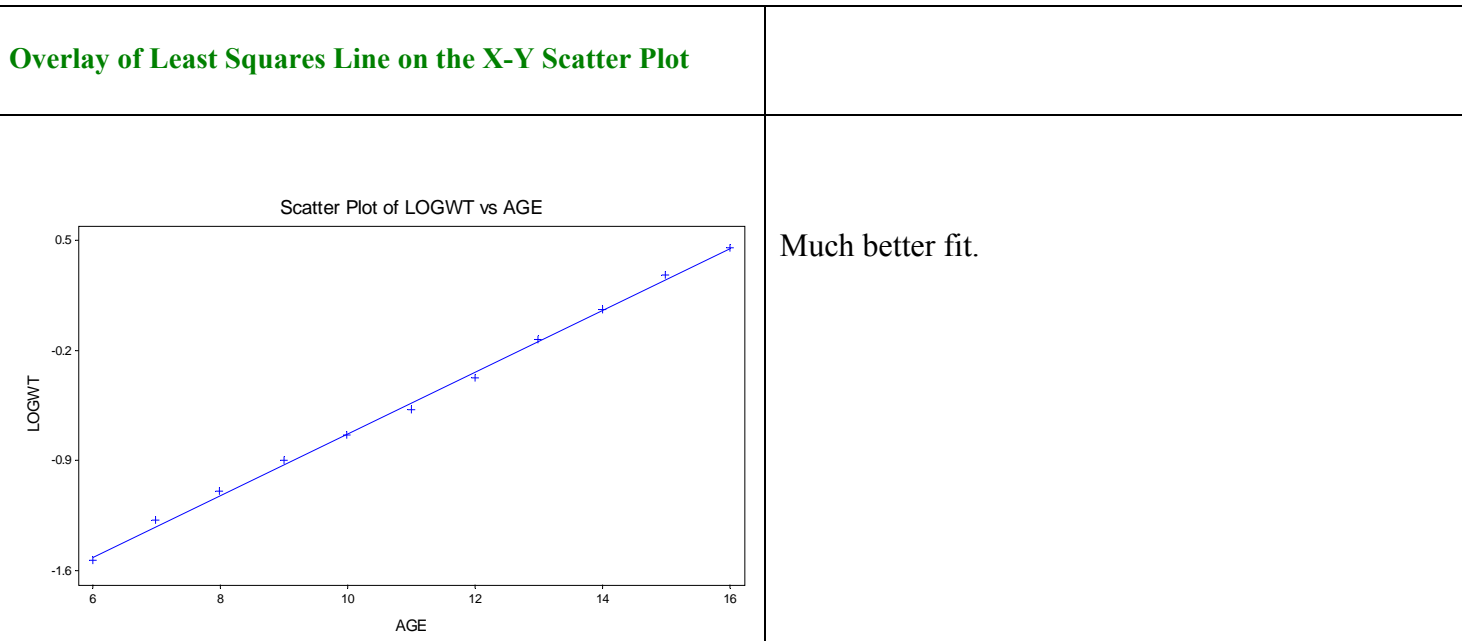
Command.

```
. regress logwt age
```

Partial Listing of Output – Annotations in red..

logwt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.1958909 = slope = b_1	.0026768	73.18	0.000	.1898356 .2019462
_cons	-2.689255 = intercept = b_0	.030637	-87.78	0.000	-2.75856 -2.619949

Thus, the fitted line is $LOGWT = -2.68925 + 0.19589 * AGE$



Now You Try ...

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

Some preliminary calculations have been done for you

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

c. The Analysis of Variance Table

The analysis of variance table is used to assess the explanatory power of the model just fit.

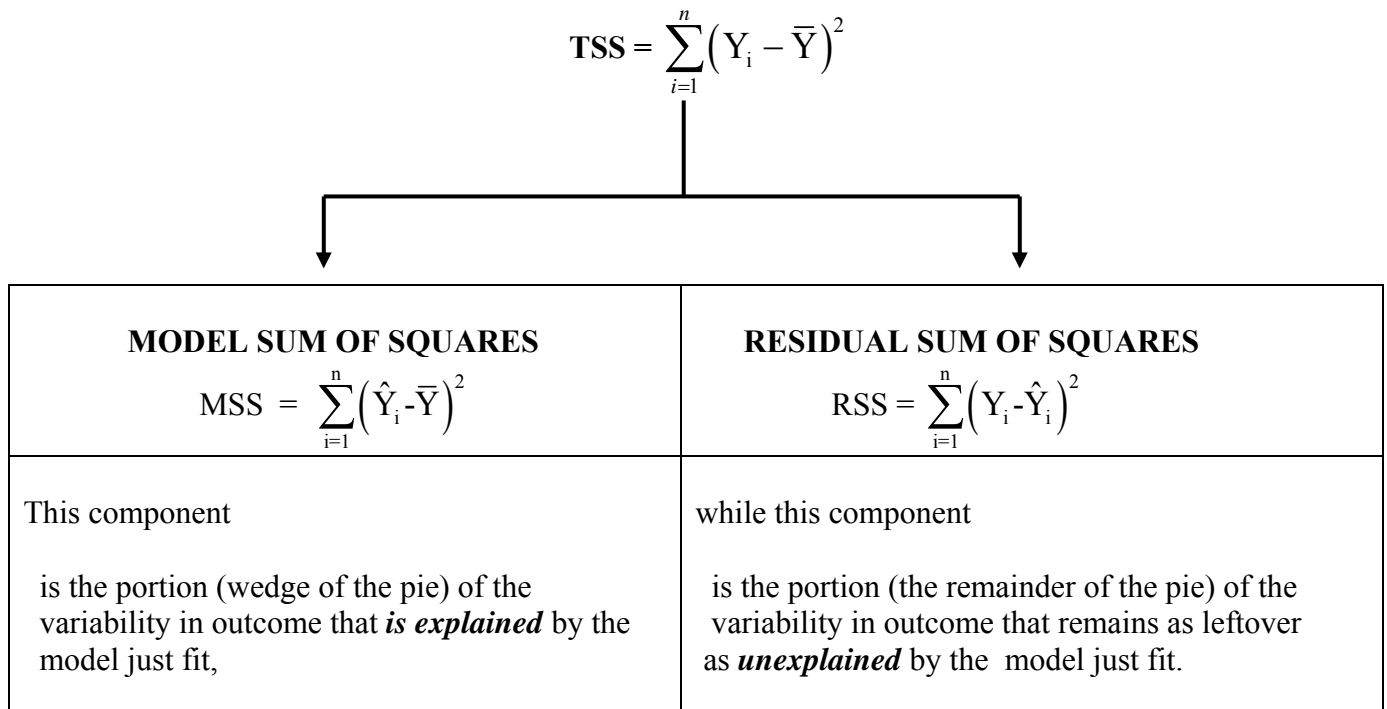
Analysis of variance calculations include sums of squares, degrees of freedom and mean squares.

Sums of squares. An analysis of variance is an analysis of a “*total variability*” that is also called a “*total sum of squares*”:

$$\text{TOTAL SUM OF SQUARES, TSS} = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

- The TSS measures the total variability of the observed outcomes Y about their mean (“the average”).
- TSS is thus 100% of what we are trying to explain (the whole pie) using the model just fit.

In a simple linear regression analysis of variance, the TSS is split into just two components:



In an analysis of variance we compare the portion of the total that is explained by the fitted model with the portion of the total that remains leftover as residual

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

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

Some algebra (not shown) confirms the partition of the total sum of squares into its two components.

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

\uparrow \uparrow \uparrow
 TSS RSS + MSS
 Total sum of squares Residual sum of squares Model sum of squares

Degrees of freedom (df). Each of the three sums of squares (TSS, RSS, and MSS) is a calculation that utilizes every data point, all “n” of them. They differ, however, in the constraints that were also utilized. **Tip!** – **Every time a constraint is placed on the data, a degree of freedom is lost.**

- To start with, the data are a random sample of *mutually independent* outcomes, sample size = n. The key here is “mutually independent”, because it means “*free to vary*”. Thus, to start with, and before any constraints, degrees of freedom = sample size = n.
- TSS:** “1 degree of freedom is lost because there is 1 constraint on the data”
In computing the **total sum of squares**, squared deviations are measured about the sample mean \bar{Y} . There is 1 constraint on the data in fixing \bar{Y} . Thus,
TSS degrees of freedom = (n-1)
- RSS:** “2 degrees of freedom are lost because there are 2 constraints on the data”
In computing the **residual sum of squares**, squared deviations are measured about the predicted values $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Now there are 2 constraints on the data, one for fixing $\hat{\beta}_0$ and the second for fixing $\hat{\beta}_1$. Thus,
RSS degrees of freedom, simple linear regression = (n-2)
- MSS:** Tip – Now we have to think about this as follows: “count one degree of freedom for each regression parameter AFTER the intercept”
In simple linear regression there are two regression **model** parameters, one for the slope and one for the intercept. →
Thus, after the intercept, there is just the regression parameter and it is for the slope.
MSS degrees of freedom = (1)

Mean squares. A sum of squares by itself is *not* a variance estimate because it is a measure of all the variability; eg “all the variability about the mean (TSS)” or “all the variability of the model about the mean (MSS)” or “all the variability of the observations about their associated predicted values (RSS)”. Instead, mean squares are variance estimates. They are defined:

$$\text{mean square} = \text{variance estimate} = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

The analysis of variance in simple linear regression compares the two variance estimates, “**due model**” versus “**due residual**” to assess the explanatory power of the model just fit.

<p>The relationship between X and Y has a <i>linear component</i> with a non-zero slope: $\beta_1 \neq 0$</p>	<p>The relationship between X and Y (if any) has <i>no linear component</i>. $\beta_1 = 0$</p>
<p>A good prediction of Y is the fitted line: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$</p>	<p>A better prediction of Y is the average of the Y’s: $\hat{Y} = \hat{\beta}_0 = \bar{Y}$</p>
<p>Consider the “due model” deviations: $(\hat{Y} - \bar{Y}) = (\hat{\beta}_0 + \hat{\beta}_1 X) - \bar{Y}$ $= \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X - \bar{Y}$ $= \hat{\beta}_1 (X - \bar{X})$</p>	<p>Here, consider the due “due residual” deviations: $(Y - \hat{Y}) = (Y - [\hat{\beta}_0]) = (Y - \bar{Y})$</p>
<p>A straight line relationship is helpful. $MSS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ is relatively large $RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ is relatively small</p>	<p>A straight line relationship is not helpful $MSS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ is relatively small $RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ is relatively large</p>
<p>$\frac{\text{MODEL mean square}}{\text{RESIDUAL mean square}}$ will be large</p>	<p>$\frac{\text{MODEL mean square}}{\text{RESIDUAL mean square}}$ will be small (close to 1)</p>

Summary of Analysis of Variance Terms

1. **TSS:** The “total” or “total, corrected” refers to the variability of Y about \bar{Y}

- ◆ $\sum_{i=1}^n (Y_i - \bar{Y})^2$ 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. **RSS:** The “residual” or “due error” refers to the variability of Y about \hat{Y}

- ◆ $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ 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”.

3. **MSS:** The “model” or “due regression” refers to the variability of \hat{Y} about \bar{Y}

- ◆ $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$ 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”.

Source	df	Sum of Squares	Mean Square
Model	1	MSS = SS(model) = $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	MSS/1 = SS(model) / 1
Residual	(n-2)	RSS = SS(residual) = $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	RSS/(n-2) = SS(residual)/(n-2)
Total, corrected	(n-1)	TSS = SS(total) = $\sum_{i=1}^n (Y_i - \bar{Y})^2$	

Hint – The entry in the “mean square” column is always the sum of squares divided by the degrees of freedom

Be careful!

Analysis of variance answers a limited question.

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

Is this better than using \bar{Y} alone?

Analysis of Variance does NOT address:

- Is the choice of the straight line model correct?
- Would another functional form be a better choice?

Illustration in Stata

Command.

```
. regress logwt age
```

Partial listing of output (now I'm showing you the analysis of variance portion) annotations in red

	Sum of Squares	Degrees of freedom	Mean Square	
Source	SS	df	MS	Number of obs = 11
-----+-----				F(1, 9) = 5355.60= MS (Model)/MS (Residual)
Model	4.22105734= MSS	1	4.22105734= MSS/1	Prob > F = 0.0000
Residual	.007093416= RSS	9	.000788157= RSS/9	R-squared = 0.9983= SS(Model)/SS(total)
-----+-----				Adj R-squared = 0.9981= R ² adjusted for n, #predictors
Total	4.22815076	10	.422815076	Root MSE = .02807= √ MS(Residual)



d. Assumptions for a Straight Line Regression Analysis

See again, page 12. Least squares estimation does not require a probability model. However, if we want to do hypothesis tests or confidence interval estimation or both, then we do need a probability model.

Assumptions of Simple Linear Regression

1. The outcomes Y_1, Y_2, \dots, Y_n are independent.
2. The values of the predictor variable X are fixed and measured without error.
3. At each value of the predictor variable $X=x$, the distribution of the outcome Y is **normal** with mean $= \mu_{Y|X=x} = \beta_0 + \beta_1 X$ and common variance $= \sigma_{Y|x}^2$.

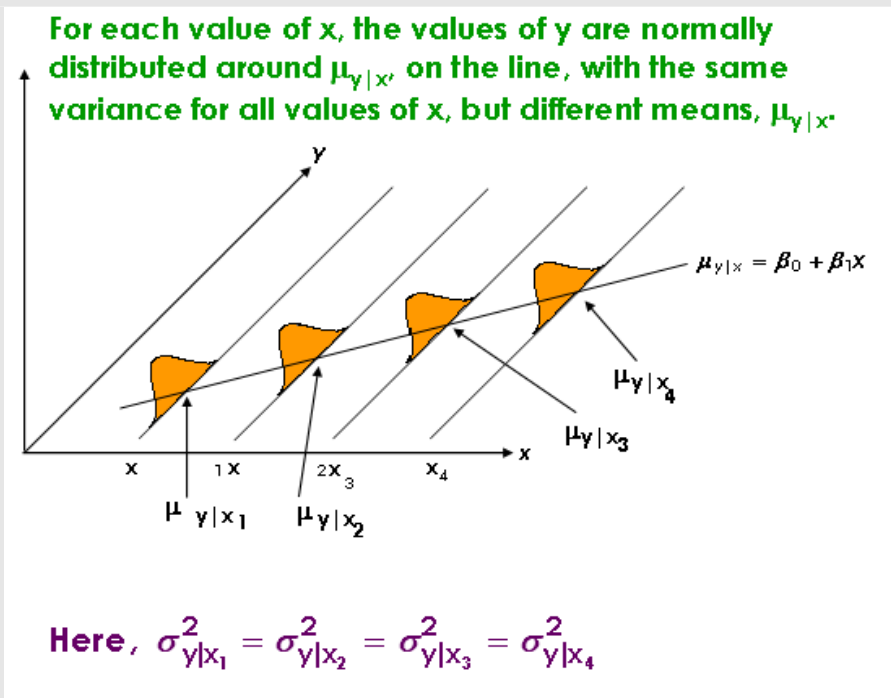
Assumptions 1-3 also mean that, for each individual “i”,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ where}$$

$\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent and identically distributed Normal with

mean $\mu_\varepsilon = 0$ and

variance $= \sigma_\varepsilon^2 = \sigma_{Y|x}^2$



With these assumptions, the comparison of the “due model” versus “due residual” variance estimates is an F-statistic under the null hypothesis of zero slope.

$$F = \frac{\text{mean square (due model)}}{\text{mean square (due residual)}} \quad \text{with } df = 1, (n-2) \mid \text{ null true.}$$

Null Hypothesis true $\beta_1 = 0$	Null Hypothesis not true $\beta_1 \neq 0$
Due model mean square has expected value $\sigma_{Y X}^2$	Due model means square has expected value $\sigma_{Y X}^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$
Due residual mean square, MS(residual), has expected value $\sigma_{Y X}^2$	Due residual mean square, MS(residual), has expected value $\sigma_{Y X}^2$
F = MS(model)/MS(residual) will be close to 1	F = MS(model)/MS(residual) will be LARGER than 1

Illustration in Stata for the model of Y=LOGWT to X=AGE:

```
. regress logwt age
```

Output (another partial listing) - Annotations in red.

	Sum of Squares	Degrees of freedom	Mean Square	
	↓	↓	↓	
Source	SS	df	MS	
-----+-----				Number of obs = 11
Model 4.22105734	1	4.22105734	F(1, 9) = 5355.60= MS (Model)/MS (Residual)	
Residual .007093416	9	.000788157	Prob > F = 0.0000	
-----+-----			R-squared = 0.9983= SS(Model)/SS(total)	
Total 4.22815076	10	.422815076	Adj R-squared = 0.9981= R ² adjusted for n and # predictors	
			Root MSE = .02807= √ MS(Residual)	



This output corresponds to the following.

Source	df	Sum of Squares	Mean Square
Due model	1	$MSS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = 4.22106$	$MSS/1 = 4.22106$
Due residual	$(n-2) = 9$	$RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 0.00709$	$RSS/(n-2) = 0.00078816$
Total, corrected	$(n-1) = 10$	$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2 = 4.22815$	

Stata provides other information, too:

- ◆ **R-SQUARED** = $(MSS)/(TSS)$
This is the percent of the total sum of squares TSS that is explained by the fit of the current model (in this case, the straight line model).
 - **Tip!** 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.
- ◆ **F(1, 9)** = [mean square(model)] / [mean square(residual)]
This is the overall F test introduced on page 26.
= [4.22106] / [0.000788816]
= 5355.60 with df = 1, 9
- **Prob > F** = achieved significance level (p-value)
This is the result of the p-value calculation for the F test. Thus, it is the probability of an F statistic value as extreme or more extreme than the value attained for the observed data under the null hypothesis assumption. In this example, p-value < 0.0001, prompting rejection of the null hypothesis H_0 . We conclude that the fitted line is a statistically significant improvement model of the outcome Y compared to using the simple model of the average Y.
- **Root MSE** = $\sqrt{[\text{mean square (residual)}]}$
This is used in many hypothesis test and confidence interval calculations.

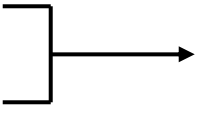
e. Hypothesis Testing

Simple Linear Regression Model: $Y = \beta_0 + \beta_1 X$

While there are more than two statistical tests, there are just two hypothesis test questions in simple linear regression:

1. Does the fit of a straight line explain statistically significantly more of the variability in outcomes than the null model that says there is no systematic relationship between X and Y?

Overall F-test
 t-test of zero slope
 t-test of zero correlation



Tip! These are all equivalent!

2. Given the fit of a straight line relationship, is the intercept statistically significantly different from zero; that is, does the line pass through the origin?

t-test of zero intercept

Overall F-Test

Research Question: Does the fitted regression model explain statistically significantly more of the variability among the outcomes Y than is explained by the average of the Y's?

Assumptions: As before (see page 25).

H₀ and H_A:

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

Test Statistic:

$$F = \frac{\text{mean square(model)}}{\text{mean square(residual)}}$$

$$df = 1, (n-2)$$

Evaluation rule:

Under the null hypothesis, F-statistic values will be close to 1.

Under the alternative hypothesis, $\beta_1 \neq 0$, F-statistic values will tend to 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 =

$$\text{pr} \left[F_{1,(n-2)} \geq \frac{\text{mean square(model)}}{\text{mean square(residual)}} \mid \beta_1=0 \right] = \text{pr} [F_{1,9} \geq 5355.60] \ll .0001$$

Evaluate:

Under the null hypothesis that $\beta_1 = 0$, the chances of obtaining an F-statistic value as (or more) extreme as 5355.60 were less than 1 chance in 10,000. This is a very small likelihood! → Statistical rejection.

Interpret:

The fitted straight line model explains statistically significantly more of the variability in Y=LOGWT than is explained by the average of LOGWT alone

... Stay tuned. Later, we'll see that the analysis does not stop here ...

T-test of Zero Slope

Preliminaries: (1) The overall F test and the test of the slope are equivalent.; (2) the test of the slope uses a t-score approach to hypothesis testing; and (3) it can be shown that $\{ \text{t-score for slope} \}^2 = \{ \text{overall F} \}$

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 $\hat{\beta}_1$

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

Our t-score is therefore:

$$\text{t-score} = \left[\frac{(\text{observed}) - (\text{expected})}{s\hat{e}(\text{observed})} \right] = \left[\frac{(\hat{\beta}_1) - (0)}{s\hat{e}(\hat{\beta}_1)} \right]$$

$$df = (n-2)$$

Illustration in Stata

logwt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.1958909	.0026768	73.18	0.000	.1898356	.2019462
_cons	-2.689255	.030637	-87.78	0.000	-2.75856	-2.619949

$t = (\text{Coef}) / (\text{Std. Err.})$
 $= 0.19589 / 0.00268$

Review- Recall what we mean by a t-score:

T=73.18 says “the estimated slope is estimated to be 73.18 standard error units away from its expected value of zero”.

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

[73.18]² = 5355.31 which is close.

Evaluation rule:

The p-value calculation answers: “Assuming null hypothesis model ($\beta_1 = 0$), what were the chances of obtaining an estimated slope (0.1959) that is as extreme as 73.18 standard error units away (in either direction!) from its expected value of 0?”

Calculations:

For our data, we obtain the two sided p-value =

$$2\text{pr} \left[t_{(n-2)} \geq \left| \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)} \right| \right] = 2\text{pr} [t_9 \geq 73.18] \ll .0001$$

Evaluate:

Under the null hypothesis that $\beta_1 = 0$, the chances of obtaining a t-statistic value as (or more) extreme as 73.18 were less than 1 chance in 10,000. This is a very small likelihood! → Statistical rejection.

Interpret:

The interpretation is the same as for the overall F-test

T-test of Zero Intercept

Tip! This is rarely of interest

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

Assumptions: As before.

H₀ and H_A:

$$H_0: \beta_0 = 0$$

$$H_A: \beta_0 \neq 0$$

Test Statistic:

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

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

Our t-score is therefore:

$$\text{t-score} = \left[\frac{(\text{observed}) - (\text{expected})}{\hat{se}(\text{observed})} \right] = \left[\frac{(\hat{\beta}_0) - (0)}{\hat{se}(\hat{\beta}_0)} \right]$$

$$df = (n - 2)$$

Illustration in Stata

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
logwt						
age	.1958909	.0026768	73.18	0.000	.1898356	.2019462
_cons	-2.689255	.030637	-87.78	0.000	-2.75856	-2.619949

t = (Coef) / (Std. Err.)

= -2.689255 / 0.030637



Evaluation rule:

The p-value calculation answers: “Assuming null hypothesis model ($\beta_0 = 0$), what were the chances of obtaining an estimated intercept (-2.6893) that is as extreme as 87.78 standard error units away (in either direction!) from its expected value of 0?”

Calculations:

For these data the two sided p-value =

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

Evaluate:

Under the null hypothesis that $\beta_0 = 0$, the chances of obtaining a t-statistic value as (or more) extreme as 87.78 were less than 1 chance in 10,000. This is a very small likelihood! → Statistical rejection.

Interpret:

Conclude that the intercept is statistically significantly different from zero or, equivalently, that the straight line relationship does not pass through the origin.

f. Confidence Interval Estimation

Simple Linear Regression Model: $Y = \beta_0 + \beta_1 X$

Confidence intervals are helpful in providing information about the range of possible parameter values that are consistent with the observed data. A simple linear regression analysis might include confidence interval estimation of four parameters: (1) slope: β_1 ; (2) intercept: β_0 ; (3) mean of population for whom $X=x_0$: $\beta_0 + \beta_1 x_0$ and (4) predicted response for an individual with $X=x$: $\beta_0 + \beta_1 x_0$.

In all instances, the confidence coefficient is a percentile of the student t-distribution with $df = (n-2)$.

Parameter	Estimate	SE (Estimate)	Confidence Coefficient
Slope: β_1	$\hat{\beta}_1$	$\sqrt{\text{mean square(residual)} \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}}$	Percentile Student t-df = (n-2)
Intercept: β_0	$\hat{\beta}_0$	$\sqrt{\text{mean square(residual)} \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}$	Percentile Student t-df = (n-2)
Mean: $\beta_0 + \beta_1 x_0$	$\hat{Y}_{X=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$	$\sqrt{\text{mean square(residual)} \left[\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}$	Percentile Student t-df = (n-2)
Prediction: $\beta_0 + \beta_1 x_0$	$\hat{Y}_{X=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$	$\sqrt{\text{mean square(residual)} \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}$	Percentile Student t-df = (n-2)

Review of PubHlth 540! (1) for a 95% CI, the correct percentile is the 97.th percentile; and more generally (2) for a $(1-\alpha)100\%$ CI, the correct percentile is the $(1-\alpha/2)100^{\text{th}}$ percentile

Stata illustration for the model which fits $Y=\text{LOGWT}$ to $X=\text{AGE}$.**How nice – Stata pretty much gives it to you!**

logwt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.1958909	.0026768	73.18	0.000	.1898356 .2019462
_cons	-2.689255	.030637	-87.78	0.000	-2.75856 -2.619949

95% Confidence Interval for the Slope, β_1

1) Estimate = $\hat{\beta}_1 = 0.19589$

2) SE (Estimate) = $se(\hat{\beta}_1) = 0.00268$

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

$$\begin{aligned}
 \text{95\% Confidence Interval for Slope } \beta_1 &= \text{Estimate} \pm (\text{confidence coefficient}) * \text{SE} \\
 &= 0.19589 \pm (2.26)(0.00268) \\
 &= (0.1898, 0.2019)
 \end{aligned}$$

95% Confidence Interval for the Intercept, β_0

1) Estimate = $\hat{\beta}_0 = -2.68925$

2) SE (Estimate) = $se(\hat{\beta}_0) = 0.03064$

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

$$\begin{aligned}
 \text{95\% Confidence Interval for Intercept } \beta_0 &= \text{Estimate} \pm (\text{confidence coefficient}) * \text{SE} \\
 &= -2.68925 \pm (2.26)(0.03064) \\
 &= (-2.7585, -2.6200)
 \end{aligned}$$

Confidence Intervals for Predictions

Stata code. Green=comment, black = command, blue=output

```

. * Confidence Intervals for Fit of Y=LOGWT to X=AGE
. * Obtain conf coeff as 97.5th percentile of Student t w df=9
. display invttail(9,.025)
2.2621572

. regress logwt age
<output not shown>

. * Obtain predicted values yhat
. predict yhat, xb
. * Obtain se for predicted individual sei
. predict sei, stdf
. * Obtain se for predicted mean semean
. predict semean, stdp

. * 95% Confidence Intervals for Individual Predictions
. generate cllowi = yhat - (2.2621572*sei)
. generate cluppi = yhat + (2.2621572*sei)
. list logwt yhat cllowi cluppi
<output shown below>

. * 95% Confidence Intervals for Mean Predictions
. generate cllowm = yhat - (2.2621572*semean)
. generate cluppm = yhat + (2.2621572*semean)
. list logwt yhat cllowm cluppm
<output shown below>
    
```

95% Confidence Intervals for <u>Individual</u> Predictions					95% Confidence Intervals for <u>Mean</u> Predictions				
+-----+ logwt yhat cllowi cluppi +-----+					+-----+ logwt yhat cllowm cluppm +-----+				
1.	-1.538	-1.513909	-1.586824	-1.440994	1.	-1.538	-1.513909	-1.549733	-1.478086
2.	-1.284	-1.318018	-1.388634	-1.247402	2.	-1.284	-1.318018	-1.348894	-1.287142
3.	-1.102	-1.122127	-1.190902	-1.053353	3.	-1.102	-1.122127	-1.148522	-1.095733
4.	-.903	-.9262364	-.9936649	-.8588079	4.	-.903	-.9262364	-.9488931	-.9035797
5.	-.742	-.7303454	-.7969533	-.6637375	5.	-.742	-.7303454	-.7504284	-.7102624
+-----+					+-----+				
6.	-.583	-.5344545	-.6007866	-.4681225	6.	-.583	-.5344545	-.5536029	-.5153061
7.	-.372	-.3385637	-.4051715	-.2719558	7.	-.372	-.3385637	-.3586467	-.3184806
8.	-.132	-.1426727	-.2101013	-.0752442	8.	-.132	-.1426727	-.1653294	-.120016
9.	.053	.0532182	-.0155564	.1219927	9.	.053	.0532182	.0268239	.0796125
10.	.275	.2491091	.1784932	.319725	10.	.275	.2491091	.2182332	.279985
+-----+					+-----+				
11.	.449	.445	.372085	.517915	11.	.449	.445	.4091766	.4808234



2. Introduction to Correlation

a. Pearson Product Moment Correlation

What is a correlation coefficient?

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.

The **Spearman** correlation is a measure of the strength of the *monotone increasing (or decreasing)* relationship between the two random variables.

Formula for the Pearson Product Moment Correlation ρ

- The population parameter designation is rho, written as ρ
- The estimate of ρ , based on information in a sample is represented using r .
- Some preliminaries:

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

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

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

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

Sample Pearson Product Moment Correlation

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

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

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

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

- 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 $\hat{\beta}_1$.

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

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

A little algebra reveals the following interrelationships:

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

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.

b. Hypothesis Test of Zero Correlation

Recall (see page 28) -

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. See page 25.

H₀ and H_A:

$$H_0 : \rho = 0$$

$$H_A : \rho \neq 0$$

Test Statistic:

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

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

$$df = (n - 2)$$

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

Stata illustration for the model which fits Y=LOGWT to X=AGE.

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

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

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

Substitution into the formula for the t-score yields

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

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 (73.18) 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.

Assumptions

The assumptions required are an extension of the ones we saw previously.

1. The separate observations Y_1, Y_2, \dots, Y_n are independent.
2. The values of the predictor variables $X_1 \dots X_p$ are fixed and measured without error.
3. For each vector value of the predictor variable $\underline{X}=\underline{x}$, the distribution of values of Y follows a normal distribution with mean equal to $\mu_{Y|\underline{X}=\underline{x}}$ and common variance equal to $\sigma_{Y|\underline{x}}^2$.
4. The separate means $\mu_{Y|\underline{X}=\underline{x}}$ lie on the line with definition

$$\mu_{Y|\underline{X}=\underline{x}} = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

Indicator Variables (0/1 Predictors)

Indicator variables are commonly used as predictors in multivariable regression models. We let

- 1 = value of indicator when “trait” is present
- 0 = value of indicator when “trait” is not present

- ◆ The estimated regression coefficient β associated with an indicator variable has a straightforward interpretation, namely:
- ◆ β = predicted change in outcome Y that accompanies presence of “trait”

Examples of Indicator Variables

SEXF = 1 if individual is female
0 otherwise

TREAT = 1 if individual received experimental treatment
0 otherwise

Design Variables (Meaningful “sets” of 0/1 predictor variables)

What do you do if you have a nominal predictor with more than 2 possible values? Answer – design variables. Design variables are sets of indicator variables that together define values of nominal variables.

If a nominal variable has k possible values, (k-1) indicator variables are needed to distinguish the entire range of possibilities.

Examples of Design Variables

Suppose a randomized trial seeks to compare medical therapy versus angioplasty versus bypass surgery for the treatment of myocardial infarction. Thus, the original treatment variable TREAT is nominal with 3 possible values:

TREAT = 1 if treatment is medical therapy
 2 if treatment is angioplasty
 3 if treatment is bypass surgery

We cannot put TREAT into a regression model as is because the estimated regression coefficient would be uninterpretable. So TREAT is replaced with a set of 2 design variables. For example, we might include the following set:

TR_ANG = 1 if treatment is angioplasty
 0 otherwise

TR_SUR = 1 if treatment is bypass surgery
 0 otherwise

A set of design variables comprised of (3-1) = 2 indicator variables summarize three possible values of treatment. The reference category is medical therapy.

Subgroup	Value of TR_ANG	Value of TR_SUR
TREAT=1 (“medical”)	0	0
TREAT=2 (“angioplasty”)	1	0
TREAT=3 (“surgery”)	0	1

Guidelines for the Definition of Indicator and Design Variables

1) Consider the choice of the reference group. Often this choice will be straightforward. It might be one of the following categories of values of the nominal variable:

- The unexposed
- The placebo
- The standard
- The most frequent

2) K levels of the nominal predictor \rightarrow (K-1) indicator variables

When the number of levels of the nominal predictor variable = k, define (k-1) indicator variables that will identify persons in each of the separate groups, apart from the reference group.

3) In general (this is not hard and fast), treat the (k-1) design variables as a set.

- Enter the set together
- Remove the set together
- In general, retain all (k-1) of the indicator variables, even when only a subset are significant.

b. The Analysis of Variance Table

The ideas of the analysis of variance table introduced previously (*see page 20*) apply here, as well.

1. **TSS:** “Total” or “total, corrected”

- ◆ $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$ is the variability of Y about \bar{Y}
- ◆ Degrees of freedom = $df = (n-1)$.

2. **MSS:** “Regression” or “due model”

- ◆ $MSS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ is the variability of \hat{Y} about \bar{Y}
- ◆ Degrees of freedom = $df = p = \#$ predictors apart from intercept

3. **RSS:** “Residual” or “due error” refers to the

- ◆ $RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ is the variability of Y about \hat{Y}
- ◆ Degrees of freedom = $df = (n-1) - (p)$

Source	df	Sum of Squares	Mean Square
Model	p	$MSS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	(MSS)/p
Residual	(n-1) - p	$RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	(RSS)/(n-1-p)
Total, corrected	(n-1)	$TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$	

Overall F Test

The overall F test introduced previously (*see page 28*) also applies, yielding an overall F-test to assess the significance of the variance explained by the model. Note that the degrees of freedom is different here; this is because there are now “p” predictors instead of 1 predictor.

$$H_O: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_A: \text{At least one } \beta_i \neq 0$$

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

c. The Partial F Test

The partial F test is a statistical technique for assessing associations while controlling for confounding. It is appropriate only when the two models being compared are “hierarchical”.

What are hierarchical models?

- Hierarchical models are two models of a particular type. The setting is that we are interested in comparing the two models.
- The descriptor “hierarchical” means that all of the predictors in one of the two models (the smaller model) are contained in the larger model.
- For example, suppose we are doing a multiple linear regression analysis of the outcome Y = length of hospital stay. We are interested in comparing two models and they are hierarchical:

Predictors in smaller model = {AGE, SEX }

Predictors in larger model = {AGE, SEX + HISTORY OF FRACTURE}

- “Hierarchical” is satisfied because all of the predictors (e.g. - AGE and SEX) that are contained in the smaller model are contained in the larger model.
- The important point to note is this. The comparison of these two models is an analysis of the nature and significance of the extra predictor, HISTORY OF FRACTURE for the prediction of length of hospital stay, adjusting for (controlling for) all of the variables in the smaller model (AGE, SEX).

Thus, the comparison of the hierarchical models is addressing the following question:

What is the significance of HISTORY OF FRACTURE for the prediction of Y = length of hospital stay, *after controlling for* the effects of AGE and SEX?



Statistical Definition of the Partial F Test

Research Question: Does inclusion of the “extra” predictors explain significantly more of the variability in outcome compared to the variability that is explained by the predictors that are already in the model?

H₀: Addition of $X_{p+1} \dots X_{p+k}$ is of no statistical significance for the prediction of Y after controlling for the predictors $X_1 \dots X_p$ meaning that:

$$\beta_{p+1} = \beta_{p+2} = \dots = \beta_{p+k} = 0$$

H_A: Not

$$F_{\text{PARTIAL}} = \frac{\{ \text{Extra regression sum of squares} \} / \{ \text{Extra regression df} \}}{\{ \text{Residual sum of squares LARGE} \} / \{ \text{Residual df LARGE} \}}$$

$$= \frac{[\text{MSS}(X_1 \dots X_p X_{p+1} \dots X_{p+k}) - \text{MSS}(X_1 \dots X_p)] / [(p+k) - p]}{[\text{RSS}(X_1 \dots X_p X_{p+1} \dots X_{p+k})] / [(n-1) - (p+k)]}$$

Numerator df = $(p+k) - (p) = k$
 Denominator df = $(n - 1) - (p+k)$

<p>H₀ true: The extra predictors are not significant in adjusted analysis</p>	<p>F value = small p-value = large</p>
<p>H₀ false: The extra predictors are significant in adjusted analysis</p>	<p>F value = large p-value = small</p>

d. Multiple Partial Correlation

The concept of a partial correlation is related to that of a partial F test.

- “To what extent are two variables, say X and Y , correlated *after* accounting for a control variable, say Z ”?
 - **Preliminary 1:** Regress X on the control variable Z
 - Obtain the residuals
 - These residuals represent the information in X that is independent of Z
 - **Preliminary 2:** Now regress Y on the control variable Z
 - Obtain the residuals
 - These residuals represent the information in Y that is independent of Z
 - These two sets of residuals permit you to look at the relationship between X and Y , independent of Z .

Partial correlation (X, Y | controlling for Z)

= Correlation (residuals of X regressed on Z , residuals of Y regressed on Z)

If there is more than one control variable Z , the result is a multiple partial correlation

A nice identity allows us to compute a partial correlation by hand from a multivariable model development

- Recall that $R^2 = [\text{model sum of squares}]/[\text{total sum of squares}] = \text{MSS} / \text{TSS}$
- A **partial correlation** is also a **ratio of sums of squares**.
Tip! – A partial F statistic is a ratio of mean squares.

$$\begin{aligned} & \text{Partial Correlation}^2(X, Y | \text{controlling for } Z) \\ &= \frac{\text{MSS Model}(X, Z) - \text{MSS Model}(Z \text{ alone})}{\text{RSS Residual}(Z \text{ alone model})} \end{aligned}$$

The hypothesis test of a **zero partial correlation** is the **partial F test** introduced previously.

Research Question: Controlling for Z, is there a linear correlation between X and Y?

H₀: $\rho_{X,Y|Z} = 0$

H_A: Not

$$\begin{aligned} F_{\text{PARTIAL}} &= \frac{[\text{MSS}(X, Z) - \text{MSS}(X)] / [(2) - 1]}{[\text{RSS}(X, Z)] / [(n-1) - (2)]} \\ &= \frac{\{ \text{Extra regression sum of squares} \} / \{ \text{Extra regression df} = 1 \}}{\{ \text{SS Residual LARGE} \} / \{ \text{df Residual LARGE} \}} \end{aligned}$$

Numerator df = (2) – (1) = 1
 Denominator df = (n – 1) – (2)

Tip! Notice that the denominator of the partial F test contains the residual sum of squares (RSS) for the **large** model, whereas the denominator of the partial correlation contains the residual sum of squares (RSS) for the **small** model!

4. Multivariable Model Development

a. Introduction

Tip! Be careful in the use of such text book approaches as “forward stepwise”, “backward elimination”, “best subsets”, etc.!!

- A detailed discussion of multivariable model development is beyond the scope of this course. However, we introduce the basic ideas and suggest some strategies.
- These ideas will be seen to be applicable in the setting of logistic regression analysis also.
- Appropriate strategies for model development should instead be guided by the goal of the analysis and subject matter considerations.
- There is no single best strategy. For example, our approach might be very different depending on the design of our epidemiological study:

Setting/Goal of Analysis	Priorities in Model Development
Randomized controlled trial. Goal: Does the intervention work?	The treatment variable might be the last variable entered into the model. Thus, we address the question: What is the independent significance of the experimental intervention controlling for all other influences?
Epidemiological description	We might want to retain as few predictors as possible so as to ensure generalizability of our findings
Risk assessment	A public health investigation of hazardous exposures and outcomes might seek to be generous in its identification of possible health risks so as to not “miss” anything.



Characteristics of Analysis Sample (n=68)

	mean (sd)	Range/sd
Age, years	39 (14)	15-75
Age at First Mensis, years	12 (1.4)	9-16
Age at First Pregnancy, years (n=51 Parous only)	23 (6)	15-40.5
P53 Score (valid range: 1 – 6)	3.25	Sd =1.05
	n	
Family History Breast Cancer	20	29%
Hormone Replacement Therapy User	3	4%
Post Menopausal	19	28%
Oral Contraceptive User	56	82%
<u>Parity Status (nominal!)</u>		
Nulliparous	17	25%
Early Parous (≤ 24 years)	32	47%
Late Parous (> 24 years)	19	28%
<u>Number of Pregnancies (ordinal!)</u>		
0	17	25%
1	9	13%
2	24	35%
3 or 4	18	26%

Study Variables:

Variable	Label	Definition/Codings
Outcome, Y p53	P53	continuous
Predictors of interest		
parous	Parity status	1 = ever parous 0 = not
pregnum	Number of pregnancies	0 = 0 pregnancies 1 = 1 pregnancy 2 = 2 pregnancies 3 = 3+ pregnancies
one	0/1 indicator of 1 pregnancy	= 1 if (pregnum=1) 0 otherwise
two	0/1 indicator of 2 pregnancies	= 1 if (pregnum=2) 0 otherwise
threep	0/1 indicator of 2 or more pregnancies	= 1 if (pregnum=3) 0 otherwise
agepreg1	Age at first pregnancy	Continuous, years = "missing" for never parous
early	0/1 indicator first pregnancy at age ≤ 24	1 = yes 0 = no = "missing" for never parous
late	0/1 indicator first pregnancy at age >24	1 = yes 0 = no = "missing" for never parous
Potential Covariates		
agecurr	Current age	continuous, years
agemen	Age at first mensis	Continuous, years
famhx01	0/1 indicator of family history of breast cancer	= 1 if any family hx of breast ca 0 otherwise
menop	0/1 indicator of post-menopause	= 1 if yes 0 otherwise
oc	0/1 indicator of ever used oral contraceptives	= 1 if yes 0 otherwise



Step 1:

Fit one predictor models.

- Retain for further consideration: predictors with crude significance levels for association $p < .25$
- Retain for further consideration: predictors of *a priori* interest that you know want to retain (for example: a blocking variable such as study site in a multicenter randomized trial)
- Report estimated regression coefficients, SE, 95% confidence intervals, p

Example

1 Predictor Model	R ² = % Variance Explained	Significance of Overall F Test	Remark
parous	.13	.0024	Retain for further consideration
one, two, threep*	.21	.0019	Retain for further consideration
agepreg1	.04	.14	Retain for further consideration
early, late	.16	.004	Retain for further consideration
agecurr	.02	.28	-
agemen	.02	.33	-
famhx01	.0003	.90	-
menop	.002	.72	-
oc	.02	.29	-

* Note – I did not assume that Y=p53 is linearly related to PREGNUM. Thus, instead of using PREGNUM as is, this predictor is replaced by three design variables: ONE, TWO and THREEP. The referent group is thus PREGNUM=0, representing “nulliparous”.

Predictor, X	$\hat{\beta} = \frac{\Delta p53}{\Delta X}$	$s\hat{e}(\hat{\beta})$	95% CI	P*
Parous	0.8949	0.2835	0.3287, 1.4611	.0024
One	0.1964	0.3977	-0.6024, 0.9951	.63
Two	0.9693	0.3096	0.3505, 1.5880	.003
Threep	1.1450	0.3596	0.4863, 1.8037	.001
Agepreg1	-0.0351	0.0236	-0.0826, +0.0123	.14
Early	1.0430	0.3007	0.4422, 1.6438	.001
Late	0.6455	0.3333	-0.0203, +1.3112	.057
agecurr	0.0103	0.0095	-0.0086, +0.0292	.28
agemen	-0.0980	0.0990	-0.2957, +0.0998	.33
Famhx01	0.0371	0.2836	-0.5294, +0.6035	.90
Menop	0.1044	0.2877	-0.4701, +0.6790	.72
Oc	0.3689	0.3474	-0.3250, +1.0627	.29

* Note – Significance of t-test for predictor, adjusted for the other design variables.

Step 2: Fit a “step 2” multiple linear model. The predictors in this model are the “candidates” from step 1.

- Retain for further consideration: predictors with adjusted significance levels for association $p < .10$
- Retain for further consideration: predictors of *a priori* interest, regardless of the significance of their crude associations in step 1.

Example - continued

Caution!! – In fitting a multiple predictor model with design variables, especially, care needs to be taken to avoid what is called “overfitting”. Overfitting occurs when two variables have definitions that are equivalent. It also occurs when two or more predictors in the model are collinear. In this example, to avoid overfitting

- (1) Parity will be modeled using TWO and THREEP
- (2) Age at first parity will be modeled using EARLY and LATE

Predictor, X	ADJUSTED $\hat{\beta} = \frac{\Delta p53}{\Delta X}$	$s\hat{e}(\hat{\beta})$	95% CI	P*
Two	0.6951	0.4049	-0.1143, 0.5046	.09
Threep	0.8686	0.4221	0.0249, 1.7123	.04
Early	0.3208	0.4661	-0.6109, 1.2525	.49
late	0.1608	0.4076	-0.6541, 0.9757	.70

*Note – Significance of Wald t-test for predictor, adjusted for the other variables in the model.

Step 3: Fit a “step 3” multiple linear model. The predictors in this model are a subset of the predictors and are the ones with adjusted significance levels $p < .10$

- Compare the “step 2” and “step 3” models using a partial F test.

Step 2 model

Source	SSQ	DF	MSQ = SSQ/DF
Model: <i>two, threep, early, late</i>	15.6625	4	3.9156
Residual:	57.7211	62	0.9310
Total, corrected:	73.3836	66	

Step 3 model

Source	SSQ	DF	MSQ = SSQ/DF
Model: <i>two, threep</i>	15.1812	2	7.5906
Residual:	58.2024	64	0.9094
Total, corrected:	73.3836	66	

$$\begin{aligned}
 \text{Partial } F_{2,62} &= \frac{[\text{SS}_{\text{model}}(4 \text{ predictor model}) - \text{SS}_{\text{model}}(2 \text{ predictor model})] / [(4) - (2)]}{\text{SS}_{\text{residual}}(4 \text{ predictor model}) / [(n-1) - (4)]} \\
 &= \frac{[15.6625 - 15.1812] / [2]}{57.7211 / [62]} \\
 &= 0.2584
 \end{aligned}$$

p-value = Probability $[F_{2,62} \geq 0.26] = .7730$

This is not statistically significant, suggesting that EARLY and LATE are not significant predictors of P53 after adjustment for TWO and THREEP. Consider dropping EARLY and LATE.

Step 4: Investigate **confounding** by considering as possible confounders the extra variables in the “step 2” model that are not in the “step 3” model. For each confounder, **one at a time**, perform a partial F test that compares reduced (small) model = step 3 model with the full (large) model = step 3 model + confounder of interest.

A Suggested Statistical Criterion for Determination of Confounding

A variable Z might be judged to be a confounder of an X-Y relationship if BOTH of the following are satisfied:

- 1) Its inclusion in a model that already contains X as a predictor has adjusted significance level < .10 or < .05; and
- 2) Its inclusion in the model alters the estimated regression coefficient for X by 15-20% or more, relative to the model that contains only X as a predictor.

Example – No evidence of confounding of Y=p53 X=TWO relationship by EARLY or LATE

	Potential Confounder =	
	EARLY	LATE
Significance of 1 df Partial F test	.55	.83
$\hat{\beta}_{\text{WITH confounder (TWO)}} =$.7856	.8992
$\hat{\beta}_{\text{WITHOUT confounder (TWO)}} =$.8986	.8986
$\Delta\hat{\beta} = \left(\frac{ \hat{\beta}_{\text{with confounder}} - \hat{\beta}_{\text{without confounder}} }{\hat{\beta}_{\text{withconfounder}}} \right) * 100$	14.4%	6.7%

Example – No evidence of Confounding of Y=p53 X=THREEP relationship by EARLY or LATE

	Potential Confounder =	
	EARLY	LATE
Significance of 1 df Partial F test	.55	.83
$\hat{\beta}_{\text{WITH confounder (THREEP)}} =$.9588	1.0742
$\hat{\beta}_{\text{WITHOUT confounder (THREEP)}} =$	1.0743	1.0743
$\Delta\hat{\beta} = \left(\frac{ \hat{\beta}_{\text{with confounder}} - \hat{\beta}_{\text{without confounder}} }{\hat{\beta}_{\text{withconfounder}}} \right) * 100$	12.0%	< 1%



Step 5: Investigate effect modification considering as your starting point the “step 4” model.

- Begin with your near final model; this is your “step 4” model
- Create interaction variables
These will be defined as pairwise products of the predictor variables.
- For each interaction variable, one at a time
Perform a partial F test that compares reduced model = step 4 model
full model = step 4 model + interaction variable.

A Suggested Statistical Criterion for Assessment of Interaction

A “candidate” interaction variable might be judged to be worth retaining in the model if **BOTH** of the following are satisfied:

- 1) **The partial F test for its inclusion has significance level $< .05$.**
- 2) **Its inclusion in the model alters the estimated regression coefficient for the main effects by 15-20% or more.**

Example –

The results to this point suggest that a good model is one containing TWO and THREEP. The potential predictors EARLY and LATE were not significant after adjustment for TWO and THREEP. Thus, we might stop here and not explore potential effect modification. However, one of the hypotheses noted in the background (see again page 71), expresses an interest in the possibility that first pregnancy *at earlier age* might influence P53. So we will explore it here. This also has the advantage of illustrating the mechanics of “step 5”.

Step 5 model contains potential modifier

Source	SSQ	DF	MSQ = SSQ/DF
Model: <i>two, threep, early</i>	15.5176	3	5.1725
Residual:	57.8660	63	0.9185
Total, corrected:	73.3836	66	

Step4 model

Source	SSQ	DF	MSQ = SSQ/DF
Model: <i>two, threep</i>	15.1812	2	7.5906
Residual:	58.2024	64	0.9094
Total, corrected:	73.3836	66	

$$\begin{aligned}
 \text{Partial } F_{1,63} &= \frac{[\text{SS}_{\text{model}}(3 \text{ predictor model}) - \text{SS}_{\text{model}}(2 \text{ predictor model})] / [(3)-(2)]}{\text{SS}_{\text{residual}}(3 \text{ predictor model}) / [(n-1)-(3)]} \\
 &= \frac{[15.5176 - 15.1812] / [1]}{57.8660 / [63]} \\
 &= 0.3662
 \end{aligned}$$

p-value = Probability [$F_{1,63} \geq 0.3662$] = .5472 not significant; null is NOT rejected.
 Conclude no evidence of modification of TWO and THREEP by EARLY.

Thus, the final model is

$$\begin{aligned}
 p\hat{3} &= 2.641 + 0.90 * \text{TWO} + 1.07 * \text{THREEP} \\
 \% \text{ variance explained} &= 20.7\% \\
 \text{Significance of Overall F test} &= .0006
 \end{aligned}$$

The significance of the overall F test can be seen from the output below (highlighted in red):

Source	SS	df	MS	Number of obs =	67
Model	15.1811813	2	7.59059063	F(2, 64) =	8.35
Residual	58.2024193	64	.909412802	Prob > F =	0.0006
Total	73.3836006	66	1.11187274	R-squared =	0.2069
				Adj R-squared =	0.1821
				Root MSE =	.95363



c. Guidelines for Multivariable Analysis of Large Data Sets

#1. State the Research Questions.

Aim for a focus that is explicit, complete, and focused, including:

- ◆ Statement of population
- ◆ Definition of outcome
- ◆ Specification of hypotheses (predictor-outcome relationships)
- ◆ Identification of (including nature of) hypothesized covariate relationships

#2. Define the Analysis Variables.

For each research question, note for each analysis variable, its hypothesized role.

- ◆ Outcome
- ◆ Predictor
- ◆ Confounder
- ◆ Effect Modifier
- ◆ Intermediary (also called intervening)

#3. Prepare a “Clean” Data Set Ready for Analysis (Data Management)

For each variable, check its distribution, especially:

- ◆ Completeness
- ◆ Occurrence of logical errors
- ◆ Within form consistency
- ◆ Between form consistency
- ◆ Range

#4. Describe the Analysis Sample

This description serves three purposes:

- 1) Identifies the population actually represented by the sample
- 2) Defines the range(s) of relationships that can be explored
- 3) Identifies, tentatively, the function form of the relationships

Methods include:

- ◆ Frequency distributions for discrete variables
- ◆ Mean, standard deviation, percentiles for continuous variables
- ◆ Bar charts
- ◆ Box and whisker plots
- ◆ Scatter plots

#5. Assessment of Confounding

The identification of confounders is needed for the correct interpretation of the predictor-outcome relationships. Confounders need to be controlled in analyses of predictor-outcome relationships.

Methods include:

- ◆ Cross-tabulations and single predictor regression models to determine whether suspected confounders are predictive of outcome and are related to the predictor of interest.
- ◆ This step should include a determination that there is a confounder-exposure relationship among controls.

#6. Single Predictor Regression Model Analyses

The fit of these models identifies the nature and magnitude of crude associations. It also permits assessment of the appropriateness of the assumed functional form of the predictor-outcome relationship.

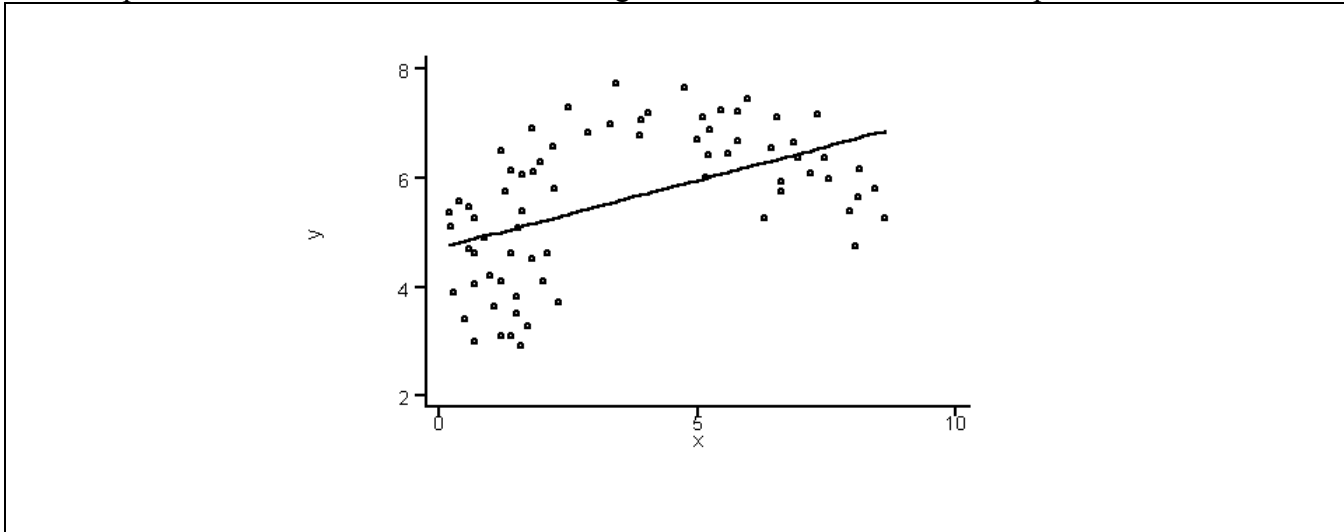
- ◆ Cross-tabulations
- ◆ Graphical displays (Scatter plots)
- ◆ Estimation of single predictor models



5. Goodness-of-Fit and Regression Diagnostics

a. Introduction and Terminology

Neither prediction nor estimation have meaning when the estimated model is a poor fit to the data:



Our eye “tells” us:

- ◆ A better fitting relationship between X and Y is quadratic
- ◆ We notice different sizes of discrepancies
- ◆ Some observed Y are close to the fitted \hat{Y} (e.g. near X=1 or X=8)
- ◆ Other observed Y are very far from the fitted \hat{Y} (e.g. near X=5)

Poor fits of the data to a fitted line can occur for several reasons and can occur even when the fitted line explains a large proportion (R^2) of the total variability in response:

- ◆ The wrong functional form (link function) was fit.
- ◆ Extreme values (outliers) exhibit uniquely large discrepancies between observed and fitted values.
- ◆ One or more important explanatory variables have been omitted.
- ◆ One or more model assumptions have been violated.

Consequences of a poor fit include:

- ◆ We learn the wrong biology.
- ◆ Comparison of group differences aren't "fair" because they are unduly influenced by a minority.
- ◆ Comparison of group means aren't "fair" because we used the wrong standard error.
- ◆ Predictions are wrong because the fitted model does not apply to the case of interest.

Available techniques of goodness-of-fit assessment are of two types:

1. **Systematic** - those that explore the appropriateness of the model itself

*Have we fit the correct model?
Should we fit another model?*

2. **Case Analysis** – those that investigate the influence of individual data points

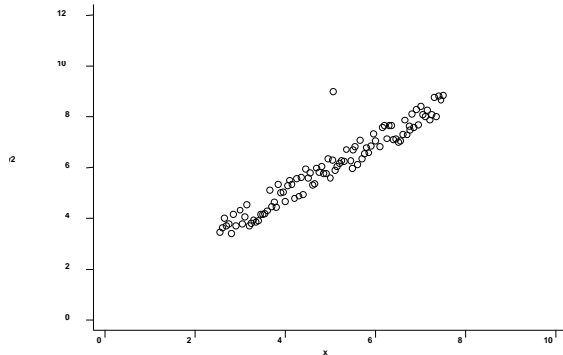
Are there a small number of individuals whose inclusion in the analysis influences excessively the choice of the fitted model?

Goodness-of-Fit Assessment Some Terminology

Systematic Component

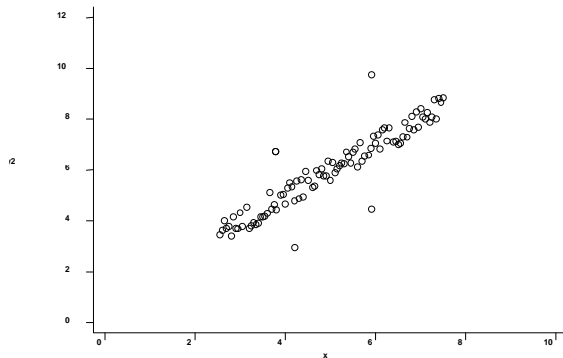
Link:	<p>The functional form (and the assumed underlying distribution of the errors) is sometimes called the link.</p> <p>Example: mean, $\mu = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$ is called the linear link.</p> <p>Example: When μ is a proportion, we might model $\ln [\mu/(1-\mu)] = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$. This is called the logit link.</p>
Normality:	<p>In the linear model regression analysis, we assume that the errors E follow a Normal(0, σ^2) distribution.</p> <p>Recall: The errors ε are estimated by the residuals e.</p>
Heteroscedasticity:	<p>If the assumption of constant variance of the errors E is not true, we say there is heteroscedasticity of errors, or non-homogeneity of errors.</p>

A Feel for Residual, Leverage, Influence
Large residuals may or may not be influential



Large residual
 Low leverage

The large residual effects a large influence.

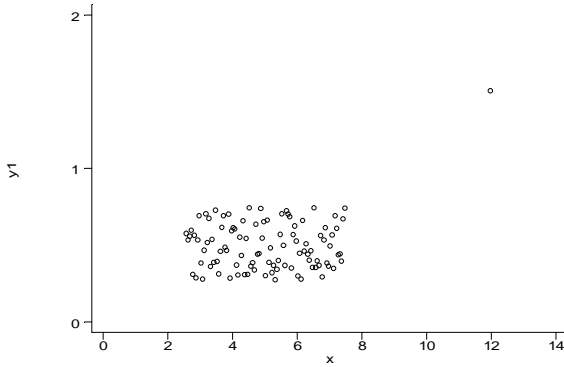


Large residual
 Low leverage

Despite its size, the large residual effects only small influence.

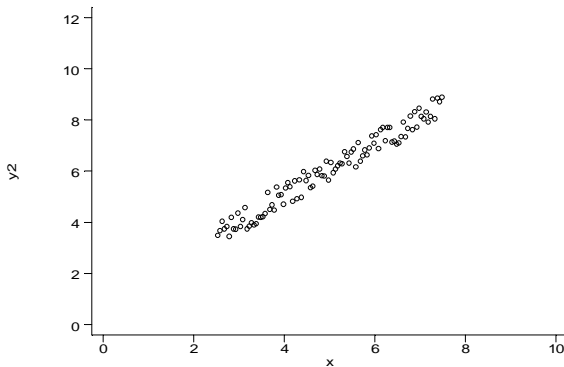
A Feel for Residual, Leverage, Influence

High leverage may or may not be influential



High leverage
Small residual

The high leverage effects a large influence.



High leverage
Small residual

Despite its size, the large leverage effects only small influence.

Thus, case analysis is needed to discover all of:

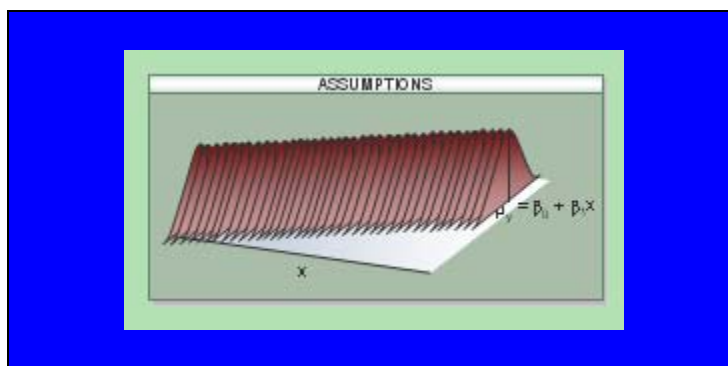
- ◆ high leverage
- ◆ large residuals
- ◆ large influence

b. Assessment of Normality

Recall what we are assuming with respect to normality:

- **Simple Linear Regression:**
At each level “x” of the predictor variable X, the outcomes Y are distributed **normal** with mean $= \mu_{Y|x} = \beta_0 + \beta_1 x$ and constant variance $\sigma_{Y|x}^2$
- **Multiple Linear Regression:**
At each vector level “ $\underline{x} = [x_1, x_2, \dots, x_p]$ ” of the predictor vector \underline{X} , the outcomes Y are distributed **normal** with mean $= \mu_{Y|\underline{x}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$ and constant variance $\sigma_{Y|\underline{x}}^2$

This is what it looks like (courtesy of a picture on the web!)



Violations of Normality are sometimes, but not always, a serious problem

- **When not to worry:** Estimation and hypothesis tests of regression parameters are fairly robust to modest violations of normality
- **When to worry:** Predictions are sensitive to violations of normality
- **Beware:** Sometimes the cure for violations of normality is worse than the problem.

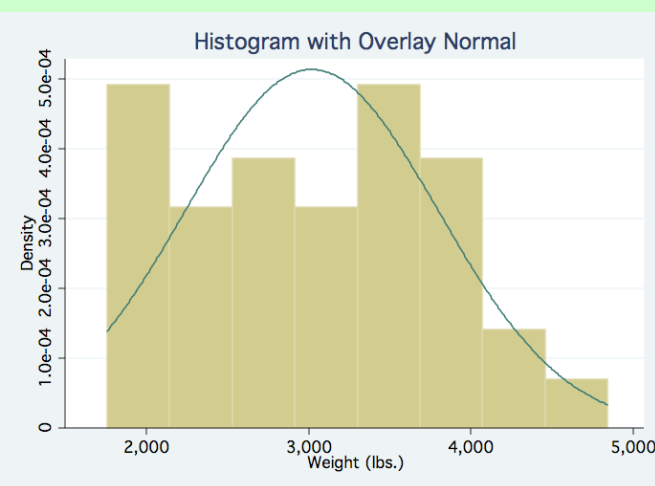
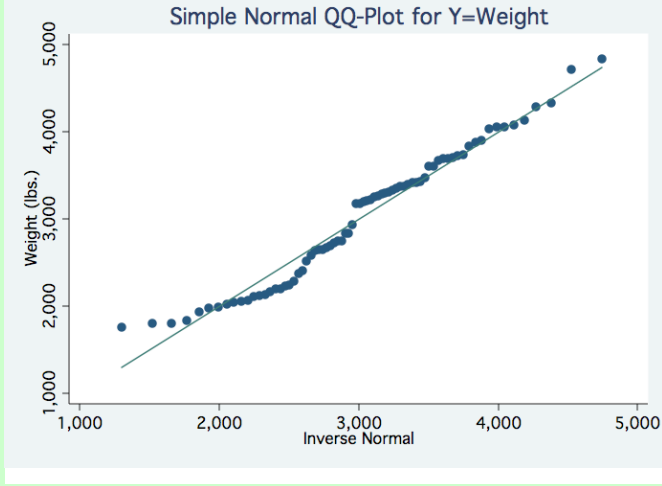
Some graphical assessments of normality and what to watch out for:

Method	What to watch out for:
1. Histogram of outcome variable Y and/or Histogram of residuals	Look for normal shape of the histogram.
2. Histogram of residuals (or studentized or jackknife residuals)	Look for normal shape of the histogram.
3. Quantile quantile plot of the quantiles of the residuals versus the quantiles of the assumed normal distribution of the residuals.	Normally distributed residuals will appear, approximately, linear.

Stata Illustration *(note – This example uses a data set from another source, not this lecture)*

Histogram with overlay normal

Quantile Quantile Plot w reference = Normal

<pre>. histogram weight, normal title ("Histogram with Overlay Normal")</pre>	<pre>. qnorm weight, title("Simple Normal QQ-Plot for Y=Weight")</pre>
 <p>A histogram showing the density distribution of weight (lbs.) with a normal distribution curve overlaid. The x-axis is labeled 'Weight (lbs.)' and ranges from 2,000 to 5,000. The y-axis is labeled 'Density' and ranges from 0 to 5.0e-04. The histogram bars are olive green, and the normal curve is a smooth, light blue line.</p>	 <p>A Quantile-Quantile (QQ) plot for Y=Weight. The x-axis is labeled 'Inverse Normal' and ranges from 1,000 to 5,000. The y-axis is labeled 'Weight (lbs.)' and ranges from 1,000 to 5,000. The plot shows a series of blue dots representing the data points, which closely follow a solid diagonal line representing the normal distribution reference.</p>

Skewness and Kurtosis Statistics for Assessing Normality:

	What to watch out for:
<p>Skewness - symmetry of the curve Standardization of the 3rd sample moment about the mean $m_2 = E \left[(X-\mu)^2 \right]$ $m_3 = E \left[(X-\mu)^3 \right]$</p> <p>What is actually examined is $a_3 = \frac{m_3}{(m_2)^{3/2}}$</p> <p>because it is unitless</p> <p>$a_3 = 0$ indicates symmetry $a_3 < 0$ indicates lefthand skew (tail to left) $a_3 > 0$ indicates right hand skew (tail to right)</p>	<p>Look for skew values between -2 and +2, roughly.</p>
<p>Kurtosis – flatness versus peakedness of the curve Standardization of the 4th sample moment about the mean $m_2 = E \left[(X-\mu)^2 \right]$ $m_4 = E \left[(X-\mu)^4 \right]$</p> <p>Pearson kurtosis is $a_4 = \frac{m_4}{(m_2)^2}$</p> <p>$a_4 = 3$ when distribution is normal $a_4 < 3$ is “leptokurtic” is too little in the tails $a_4 > 3$ is “platykurtic” is too much in the tails</p>	

Hypothesis Tests of Normality and what to watch out for:

Test Statistic	What to watch out for:
<p>1. <u>Shapiro Wilk (W)</u></p> <p>W is a measure of the correlation between the values in the sample and their associated normal scores (for review of Normal Scores, see BE540 Topic 5 – Normal Distribution)</p> <p>W = 1 under normality</p>	<p>Evidence of violation of normality is reflected in</p> <p>W < 1</p> <p>small p-value</p>
<p>2. <u>Kolmogorov-Smirnov (D)</u>. See also <u>Lilliefors (K-S)</u></p> <p>This is a goodness of fit test that compares the distribution of the residuals to that of a reference normal distribution using a chi square test.</p> <p>Lilliefors utilizes a correction</p>	<p>Evidence of violation of normality is reflected in</p> <p>D > 0</p> <p>K-S > 0</p> <p>small p-value</p>

Guidelines

In practice, the assessment of normality is made after assessment of other model assumption violations. The linear model is often more robust to violations of the assumption of normality. The cure, is often worse than the problem. (e.g. – transformation of the outcome variable)

Consider doing a scatterplot of the residuals. Look for

- ◆ Bell shaped pattern
- ◆ Center at zero
- ◆ No gross outliers



c. Cook-Weisberg Test of Heteroscedasticity

Recall what we are assuming with respect to homogeneity of variance:

- **In Simple Linear Regression:**
At each level “x” of the predictor variable X, the outcomes Y are distributed normal with mean = $\mu_{Y|x} = \beta_0 + \beta_1x$ and **constant variance** $\sigma^2_{Y|x}$

Evidence of a **violation** of homogeneity (this is heteroscedasticity) is seen when

- There is increasing or decreasing variation in the residuals with fitted \hat{Y}
- There is increasing or decreasing variation in the residuals with predictor X

Some graphical assessments of homogeneity of variance and what to watch out for:

Method	What to watch out for:
1. Plot Residuals or standardized residuals or studentized residuals on the vertical – versus – Predicted outcomes \hat{Y} on the horizontal	Look for even band at zero
2. Plot Residuals or standardized residuals or studentized residuals on the vertical – versus – Predictor values X	Look for even band at zero

Hypothesis Test of homogeneity of variance is Cook-Weisberg

Cook-Weisberg Test	What to watch out for:
This test is based on a model of the variance as a function of the fitted values (or the predictor X). Specifically, it is a chi square test of whether the squared standardized residuals are linearly related to the fitted values (or the predictor X).	Evidence of violation of homogeneity of variance is reflected in Large test statistic > 0 small p-value

d. The Method of Fractional Polynomials

This method is beyond the scope of this course. However, it's helpful to understand its theory.

Goal: The goal is to select a “good” functional form that relates X to Y from a collection of candidate models. Candidates are lower polynomials and members of the Box-Tidwell family.

Fractional Polynomials: Instead of $Y = \beta_0 + \beta_1 X$, we consider the following:

$$Y = \mathcal{G}_m \left(X; \underline{\beta}, \underline{p} \right) = \beta_0 + \sum_{j=1}^m \beta_j H_j(X) \text{ where}$$

$$H_j(X) = X_j^{p_j} \text{ when } p_j \neq p_{j-1} \text{ and}$$

$$H_j(X) = \ln(X) \text{ when } p_j = p_{j-1}.$$

- ◆ $\underline{\beta}$ is a vector of coefficients length = m
- ◆ \underline{p} is a vector of “fractional powers” of length = m
- ◆ The powers are selected from among { -3, -2, -1, -0.5, 0, 0.5, 1, 2, 3 }

Example: When m=1 the vectors $\underline{\beta}$ and \underline{p} are each of length 1. Suppose $\underline{p} = [1]$

$$Y = \mathcal{G}_m \left(X; \underline{\beta}, \underline{p} \right) = \beta_0 + \sum_{j=1}^m \beta_j H_j(X) = \beta_0 + \beta_1 X$$

Example: When m=2 the vectors $\underline{\beta}$ and \underline{p} are each of length 2. Suppose $\underline{p} = [.5, .5]$

$$Y = \mathcal{G}_m \left(X; \underline{\beta}, \underline{p} \right) = \beta_0 + \sum_{j=1}^m \beta_j H_j(X) = \beta_0 + \beta_1 \sqrt{X} + \beta_2 \ln(X)$$

note – Working with this expression does take practice!

The Method of Fractional Polynomials - Continued

Guidelines

Competing models are assessed using a chi square statistic that compares the likelihoods of the data under each of the two models using what is called a “deviance” statistic.

Don't worry: We will learn about the “deviance” statistic in Unit 5 in the context of the logistic regression model.

Search begins with examination of all models for which $m=1$. We choose the one model in this class that has the smallest deviance.

- ◆ We compare the best $m=1$ model to the specific model for which $m=1$ and $p=[1]$ because the latter is the simple linear model.
- ◆ Thus, we are asking whether it is really necessary to abandon the simple linear model.

Next, we compare the best $m=1$ model to the best $m=2$ model. And so on ...

- ◆ In general, we must choose between two costs:
 - 1) A smaller model has a lower goodness-of-fit but more generalizability
 - 2) A larger model has a higher goodness-of-fit but less generalizability
- ◆ Our goal is to choose the smallest model for which the goodness-of-fit is acceptable.

e. Ramsey Test for Omitted Variables

A fitted model that fails to include an important explanatory variable is problematic.

- ◆ Our understanding of the outcomes is incomplete.
- ◆ Estimated associations may be biased due to confounding.
- ◆ Model assumptions may be violated.

Method of the Ramsey Test

- ◆ H_0 : Predicted values from the fitted model are unrelated to powers of the fitted model, after adjustment for the predictor variables in the model.

$$\text{corr}(\hat{Y}, \hat{Y}^p) = 0$$

- ◆ For example, we might fit the model $\hat{Y} = \beta_0 + \beta_1 \hat{Y} + \beta_2 \hat{Y}^2 + \beta_3 X + \text{error}$ and test the significance of $\hat{\beta}_1$ and $\hat{\beta}_2$.
- ◆ The test statistic is an F statistic.

Guidelines

Evidence of a failure to include one or more explanatory variables is reflected in a large F statistic value.

As a suggestion, do also a scatterplot of the squared standardized residuals versus the leverage values. Omission of an important explanatory variables is suggested by

- ◆ Extreme values
- ◆ Any systematic pattern

f. Residuals, Leverage, and Cook’s Distance

Residuals - There are multiple measures of “residual”.

<p>Ordinary residual $e = (Y - \hat{Y})$</p>	<p>Standardized residual $e^* = \frac{e}{\sqrt{ms(residual)}} = \frac{e}{\sqrt{\hat{\sigma}_{Y x}^2}}$</p>
<p>Studentized residual $e^* = \frac{e}{\sqrt{ms(residual)}\sqrt{1-h}} = \frac{e}{\sqrt{\hat{\sigma}_{Y x}^2}\sqrt{1-h}}$</p>	<p>Jackknife residual, also called Studentized deleted residual $e^* = \frac{e}{\sqrt{ms(residual)_{-i}}\sqrt{1-h}} = \frac{e}{\sqrt{\hat{\sigma}_{Y x}^2}\sqrt{1-h}}$</p>

Which one or ones should we use?

- **Standardized** residuals can be appreciated as we do z-scores.
- **Studentized** residuals are distributed Student’s t (df=n-p-1) when regression assumptions hold.
- **Jackknife** residuals are distributed Student’s t (df=n-p-2) when regression assumptions hold. These also have the advantage of correcting the magnitude of the $\sqrt{MS(residual)}$ when it is otherwise too big because of the effects of influential points.

Leverage, h_i :

Leverage is the distance of a predictor value $X=x$ from the center of the values of the predictor value $X = \bar{x}$. This distance is denoted h_i .

For simple linear regression,
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

For simple linear regression, a “large” leverage value is $h_i \geq \frac{4}{n}$



Cook's Distance, d

Recall from our pictures that neither a large residual alone nor a high leverage determine the influence of an individual data point.

Cook's distance provides a measure of the influence of an individual data point on the fitted model and is a function of the values of both the residual and leverage:

Cook's Distance = Change in estimated regression coefficient
value, expressed in standard error units.

1) For simple linear regression
$$d = \frac{e^2 h}{2s^2(1-h)^2}$$

2) For multivariable linear regression models
$$d_i = \frac{(\hat{\beta}_{-i} - \hat{\beta})' (X'X)(\hat{\beta}_{-i} - \hat{\beta})}{p' s_{Y|X}^2}$$

where

i indexes the individual for which measure of influence is sought

$\hat{\beta}$ = vector of estimated regression coefficients using the entire sample

$\hat{\beta}_{-i}$ = vector of estimated regression coefficients with omission of the
 i^{th} data point

X = matrix of values of the predictor variables

p' = rank (X) = number of predictors + 1

Guidelines

- ◆ For the linear regression model, a “noteworthy” influential data point is one for which $d \geq 1$.

For a multivariable regression model, a “noteworthy” influential data point is one for which $d \geq 2(p+1)/n$ where p = # predictors.