

Unit 9

Regression and Correlation

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

- Gerald van Belle

At the heart of data exploration is fitting models to data. The data are our “givens” and the statistical model is just that - a model. Once obtained, fitted models have a variety of uses, depending in the extent to which it reveals underlying biology:

-1- “somewhat useful”

The fitted model is a sufficiently good fit to the data that it permits tests of hypotheses such as “the experimental treatment is associated with statistically significant benefit” and it permits assessment of confounding, effect modification, and mediation. The latter are ideas that will be developed in PubHlth 640 Unit 2, ***Multivariable Linear Regression***.

-2- “more useful”

The fitted model can be used to predict the outcomes of future observations. For example, we might be interested in predicting the survival time following surgery of a future patient undergoing coronary bypass surgery.

-3- “most useful”

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...”

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

Table of Contents

Topic		
	1. Unit Roadmap	3
	2. Learning Objectives	4
	3. Definition of the Linear Regression Model	5
	4. Estimation	13
	5. The Analysis of Variance Table	22
	6. Assumptions for the Straight Line Regression	26
	7. Hypothesis Testing	29
	8. Confidence Interval Estimation	35
	9. Introduction to Correlation	40
	10. Hypothesis Test for Correlation	43

Nature



Population/
Sample



Observation/
Data

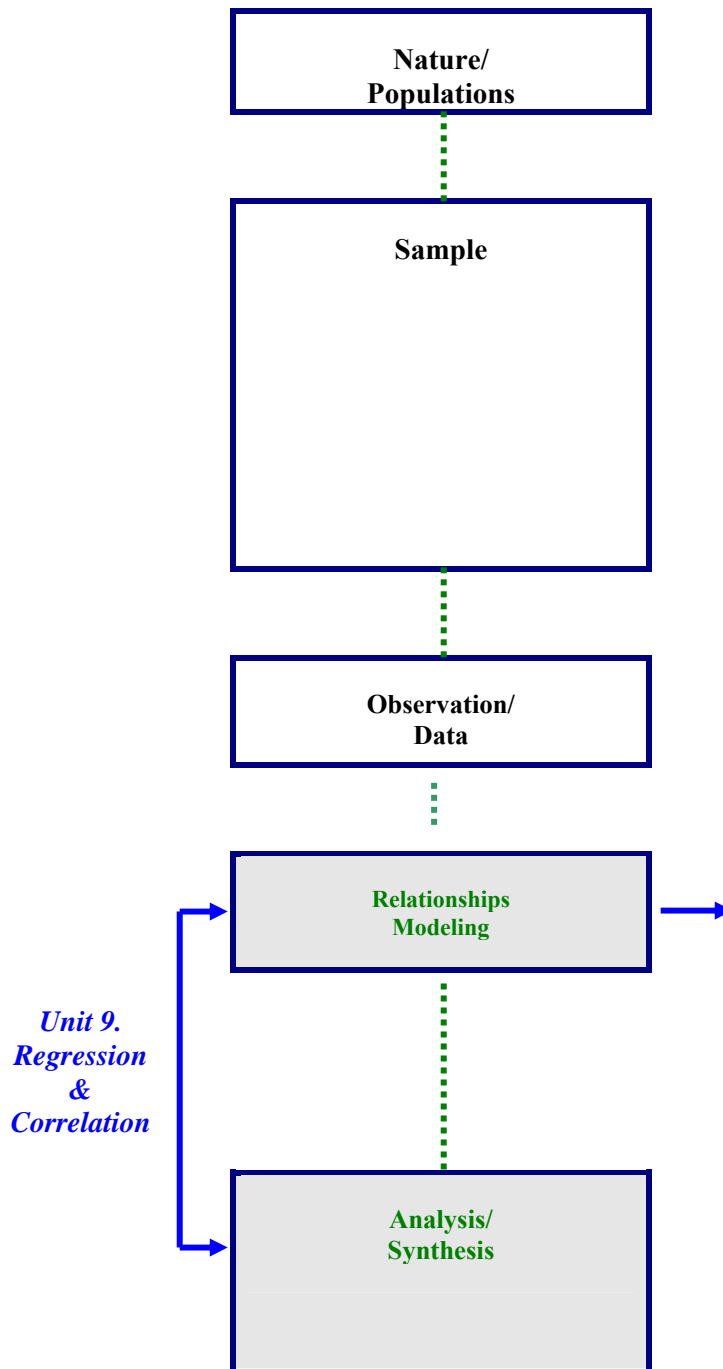


Relationships/
Modeling



Analysis/
Synthesis

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:

$$y = \beta_0 + \beta_1x \quad \text{where}$$

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

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

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

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

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 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 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 confidence interval estimation and statistical hypothesis testing of the Pearson product moment correlation r .

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

3. Definition of the Linear Regression Model

Unit 8 considered two **categorical (discrete)** variables, such as smoking and low birth weight. It was an introduction to chi-square tests of association.

Unit 9 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.”

Linear Regression

Linear regression models the mean $\mu = E [Y]$ of **one random** variable Y as a linear function of one or more other variables (called predictors or explanatory variables) that are treated as fixed. The estimation and hypothesis testing involved are extensions of ideas and techniques that we have already seen. In linear regression,

- ◆ Y is the outcome or dependent variable that we observe. We observe its values for individuals with various combinations of values of a predictor or explanatory variable X. There may be more than one predictor “X”; this will be discussed in PubHlth 640.
- ◆ In simple linear regression the values of the predictor “X” are assumed to be fixed.
- ◆ Often, however, the variables Y and X are both random variables.

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

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

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

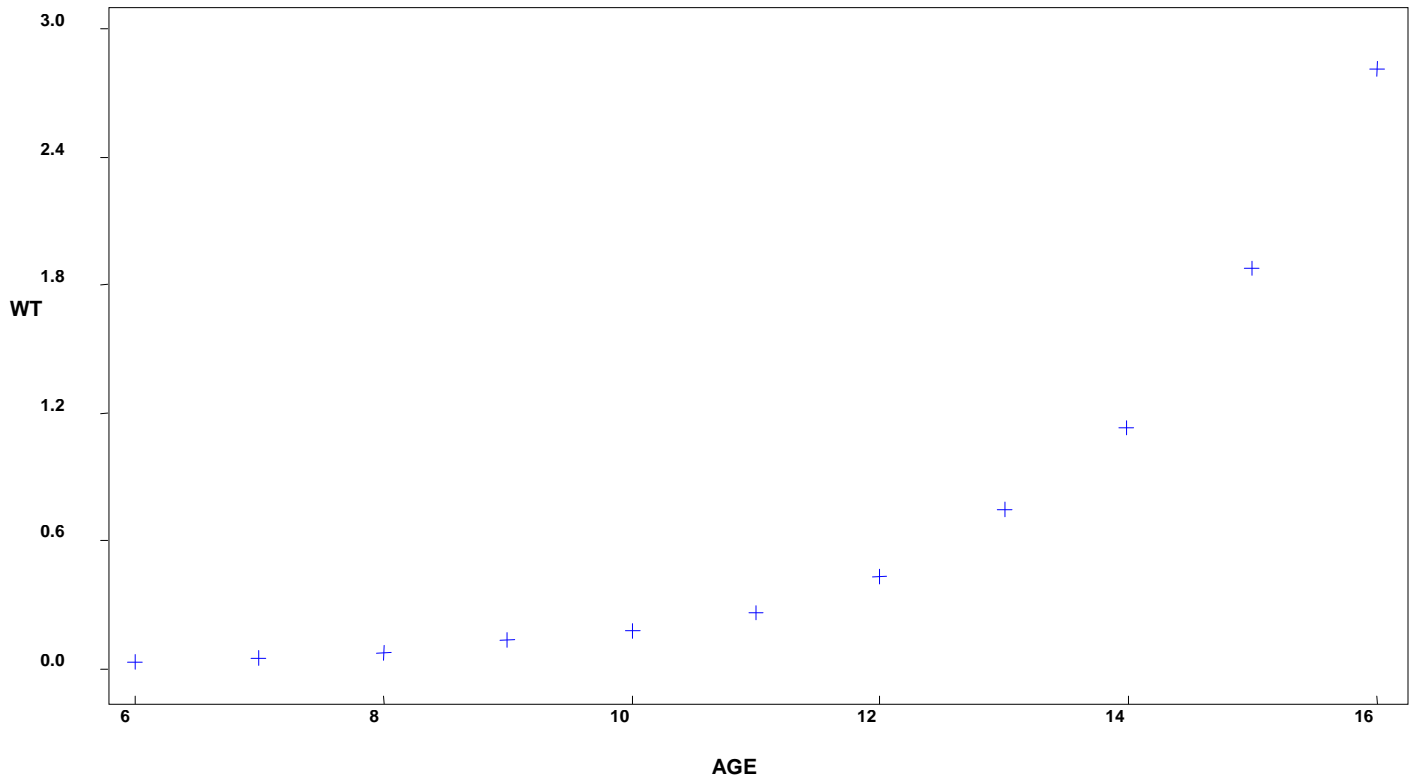
Research question

There are a variety of possible research questions:

- (1) Does weight change with age?
- (2) 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?
- (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

Scatter Plot of WT vs AGE



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

Nature

Population/
Sample

Observation/
Data

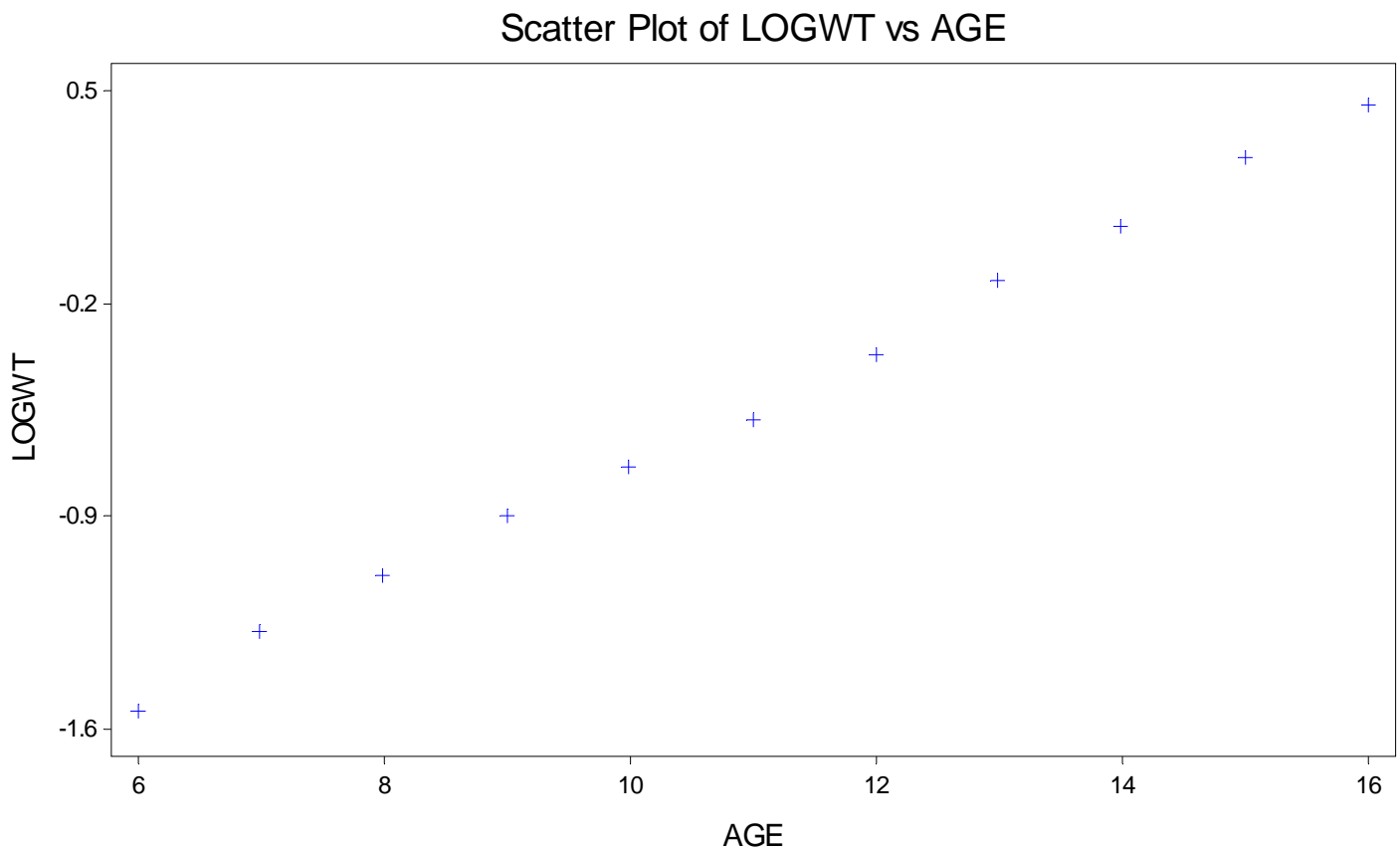
Relationships/
Modeling

Analysis/
Synthesis

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 logarithm of WT (Z) to AGE:



Nature

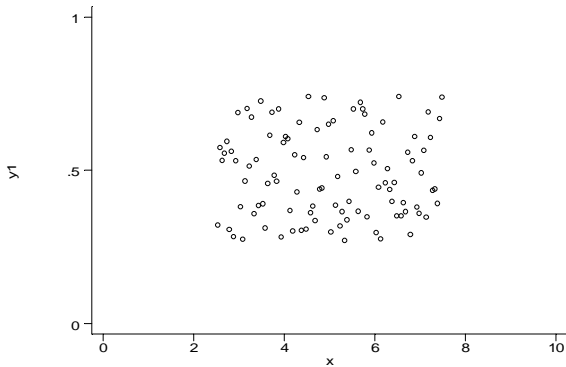
Population/
Sample

Observation/
Data

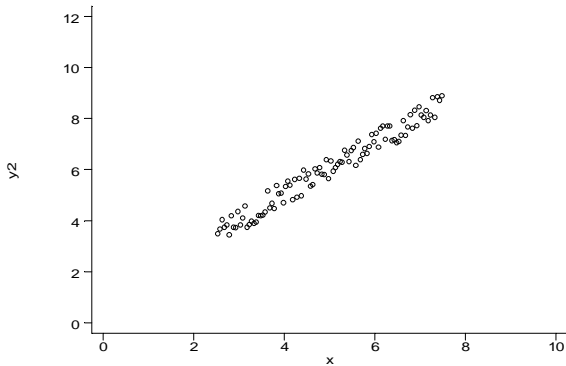
Relationships/
Modeling

Analysis/
Synthesis

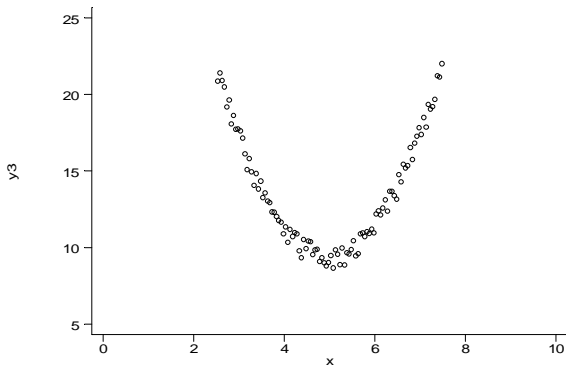
We might have gotten any of a variety of plots.



No relationship between X and Y



Linear relationship between X and Y



Non-linear relationship between X and Y

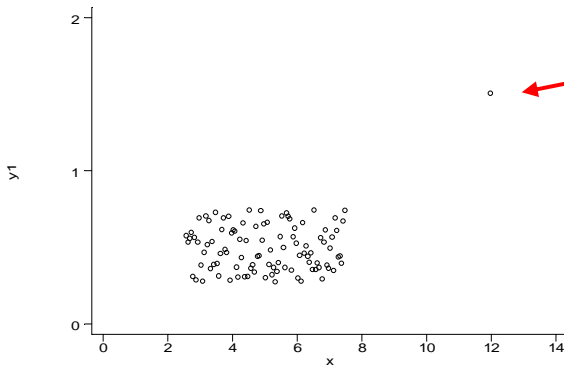
Nature

Population/
Sample

Observation/
Data

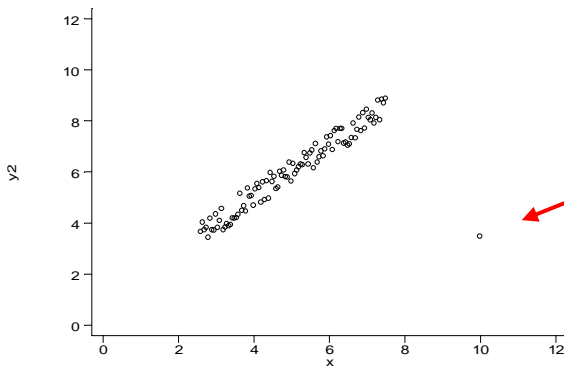
Relationships/
Modeling

Analysis/
Synthesis



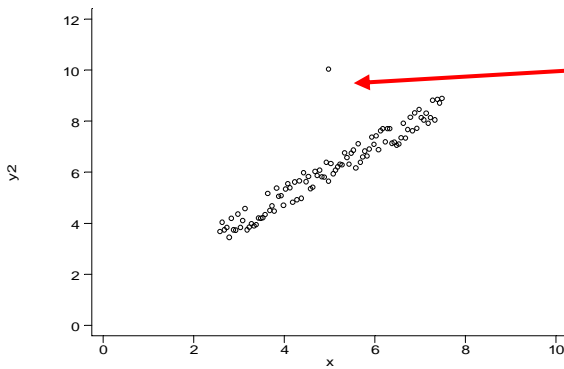
Note the arrow pointing to the outlying point

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



Note the arrow pointing to the outlying point

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



Note the arrow pointing to the outlying point

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

Nature

Population/
Sample

Observation/
Data

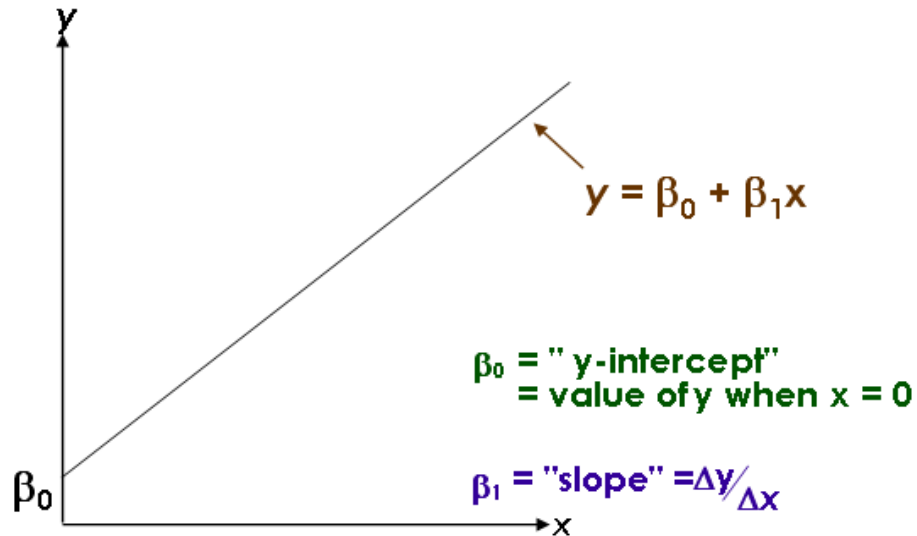
Relationships/
Modeling

Analysis/
Synthesis

Review of the Straight Line

Way back when, in your high school days, you may have learned about the straight line 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: $m \rightarrow \beta_1$
- (2) Intercept: $b \rightarrow \beta_0$



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

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

Slope

Slope > 0	Slope = 0	Slope < 0

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

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. <u>$Y = \beta_0 + \beta_1 X$</u> is measured with <u>error = ε</u>	$\hat{\beta}_0, \hat{\beta}_1$ and e are estimates of β_0, β_1 and ε <u>residual = e</u>
β_0, β_1 and ε are <u>unknown</u>	$\hat{\beta}_0, \hat{\beta}_1$ and e are <u>known</u>
	The values of $\hat{\beta}_0, \hat{\beta}_1$ and e are obtained by the method of <u>least squares estimation</u> . How close did we get? To see if $\hat{\beta}_0 \approx \beta_0$ and $\hat{\beta}_1 \approx \beta_1$ we perform <u>regression diagnostics</u> . <i>Regression diagnostics are discussed in PubHlth 640</i>

Notation

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

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

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.

“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 to choosing a line that is “closest” to the data.

- ◆ d_i
Perhaps we’d like $d_i = [\text{observed } Y - \text{fitted } \hat{Y}] = \text{smallest possible}$.
Note that this is a vertical distance, since it is a distance on the vertical axis.
- ◆ d_i^2
Better yet, perhaps we’d like to minimize the squared difference:
 $d_i^2 = [\text{observed } Y - \text{fitted } \hat{Y}]^2 = \text{smallest possible}$
- ◆ We can’t do this minimization separately for each X-Y pair. That is, it is not possible to choose common values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes

$$d_1^2 = (Y_1 - \hat{Y}_1)^2 \quad \text{for subject 1 *and* minimizes}$$

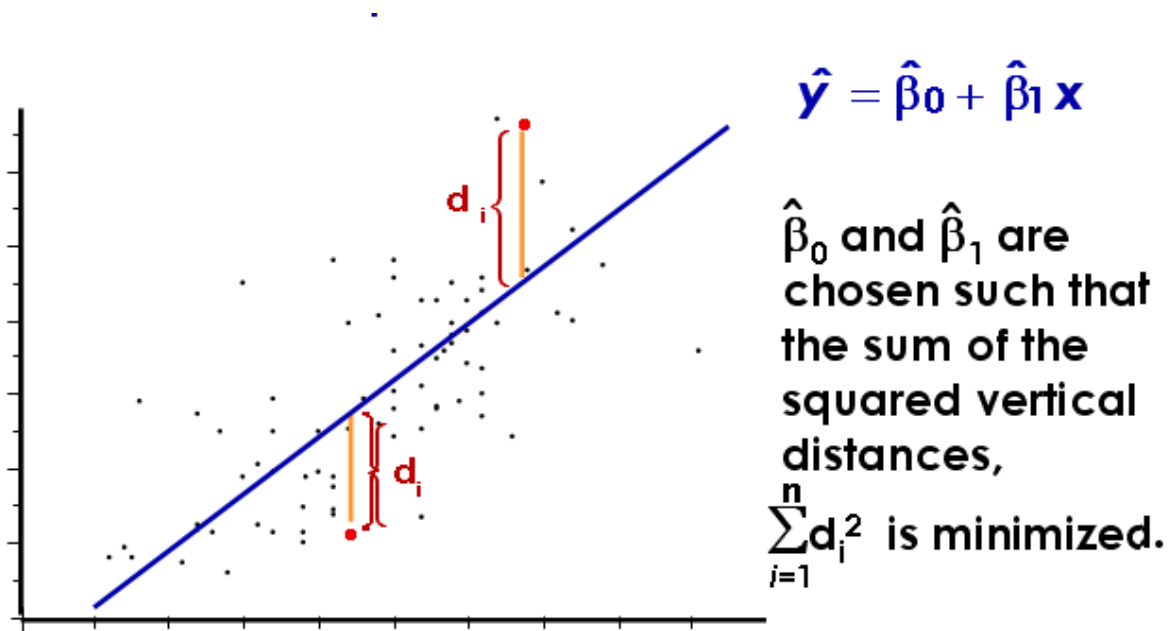
$$d_2^2 = (Y_2 - \hat{Y}_2)^2 \quad \text{for subject 2 *and* minimizes}$$

.... ... *and* minimizes

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

- ◆ So, instead, we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes their total

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



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

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

- ◆ residual sum of squares
- ◆ sum of squares about the regression line
- ◆ sum of squares due error (SSE)
- ◆ $\hat{\sigma}_{Y|X}^2$

Least Squares Estimation Method

In case you're interested

- ◆ Consider $SSE = \sum_{i=1}^n d_1^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - [\hat{\beta}_0 + \hat{\beta}_1 X_i])^2$
- ◆ **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.

Least Squares Estimation Solution

Estimate of Slope $\hat{\beta}_1$ or b_1	$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$
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

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

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

The capital “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

Slope	$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\hat{c}ov(X, Y)}{\hat{v}ar(X)}$	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$
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	$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\widehat{cov}(X, Y)}{\widehat{var}(X)}$	<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</p> <p>$\hat{\beta}_1 = 0$ and $\hat{\beta}_0 = \bar{Y}$ as our best guess of an unknown Y</p>

Illustration in Stata
Y=WT and X=AGE

```
. regress y x
```

Partial listing of output ...

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.2350727	.0459425	5.12	0.001	.1311437	.3390018
_cons	-1.884527	.5258354	-3.58	0.006	-3.07405	-.695005

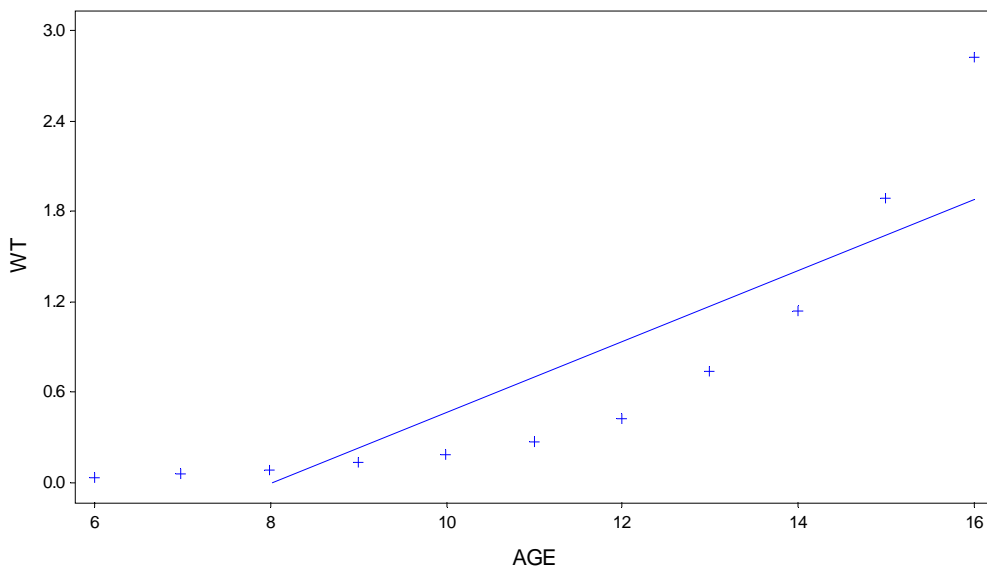
Annotated ...

y = WEIGHT	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x = AGE	.2350727 = b₁	.0459425	5.12	0.001	.1311437	.3390018
_cons = Intercept	-1.884527 = b₀	.5258354	-3.58	0.006	-3.07405	-.695005

The fitted line is therefore $WT = -1.884527 + 0.23507 \cdot AGE$

Let's overlay the fitted line on our scatterplot.

Scatter Plot of WT vs AGE



- ◆ 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 in Stata- continued
Z=LOGWT and X=AGE

```
. regress z x
```

Partial listing of output ...

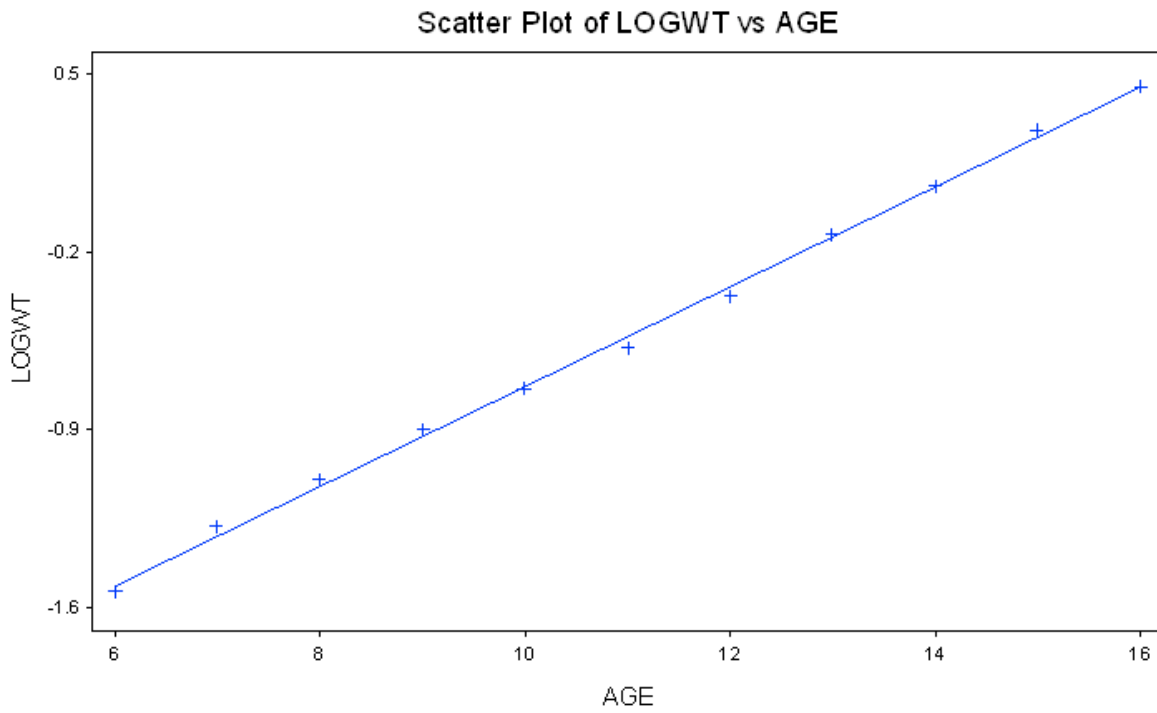
z	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.1958909	.0026768	73.18	0.000	.1898356	.2019462
_cons	-2.689255	.030637	-87.78	0.000	-2.75856	-2.619949

Annotated ...

Z = LOGWT	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x = AGE	.1958909 = b ₁	.0026768	73.18	0.000	.1898356	.2019462
_cons = INTERCEPT	-2.689255 = b ₀	.030637	-87.78	0.000	-2.75856	-2.619949

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

Now the overlay plot looks better:



Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

Now You Try ...

Prediction of Weight from Height

Source: Dixon and Massey (1969)

<u>Individual</u>	<u>Height (X)</u>	<u>Weight (Y)</u>
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	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$	$\hat{\beta}_1 = \frac{863.333}{171.667} = 5.0291$
Intercept	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$	$\hat{\beta}_0 = 141.667 - (5.0291)(63.833)$ $= -179.3573$

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

5. The Analysis of Variance Table

Recall the sample variance introduced in In Unit 1, *Summarizing Data*.

The numerator of the sample variance of the Y data is $\sum_{i=1}^n (Y_i - \bar{Y})^2$.

This same numerator $\sum_{i=1}^n (Y_i - \bar{Y})^2$ is a central figure in regression. It has a new name. Several.

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \text{“total variance of the Y’s”} \\ &= \text{“total sum of squares”}, \\ &= \text{“total, corrected”}, \text{ and} \\ &= \text{“SSY”}. \end{aligned}$$

(Note – “corrected” has to do with subtracting the mean before squaring.)

The analysis of variance tables is all about $\sum_{i=1}^n (Y_i - \bar{Y})^2$ and partitioning it into two components

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

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) 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

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

A closer look...

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

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

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

In the world of regression analyses, we seek to *explain* the total variability $\sum_{i=1}^n (Y_i - \bar{Y})^2$:

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 is LARGE because $(\hat{Y} - \bar{Y}) = (\hat{\beta}_0 + \hat{\beta}_1 X) - \bar{Y}$ $= \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X - \bar{Y}$ $= \hat{\beta}_1 (X - \bar{X})$	Due error is nearly the TOTAL because $(Y - \hat{Y}) = (Y - [\hat{\beta}_0]) = (Y - \bar{Y})$
Due error has to be small	Due regression has to be small
$\frac{\text{due(model)}}{\text{due(error)}}$ will be large	$\frac{\text{due(model)}}{\text{due(error)}}$ will be small

How to Partition the Total Variance

1. 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. 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. The **“regression”** or **“due model”** 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”. It is an example of a variance component.

Source	df	Sum of Squares	Mean Square
Regression	1	$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	SSR/1
Error	(n-2)	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	SSE/(n-2)
Total, corrected	(n-1)	$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$	

Tip! – Mean square = (Sum of squares)/(degrees of freedom,df)

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

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 \bar{Y} ?

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

We are NOT asking:

Is the choice of the straight line model correct?

Would another functional form be a better choice?

We'll use a hypothesis test approach (another "proof by contradiction").

- ◆ Start with 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
- ◆ Recall that we reasoned the following:

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 an value of due(regression)/due(residual) as larger or larger than that observed?

To calculate "chances" we need a probability model.

So far, we have not needed one.

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

6. Assumptions for a Straight Line Regression Analysis

In performing least squares estimation, we did not use a probability model. We were doing geometry. Hypothesis testing requires some assumptions and a probability model.

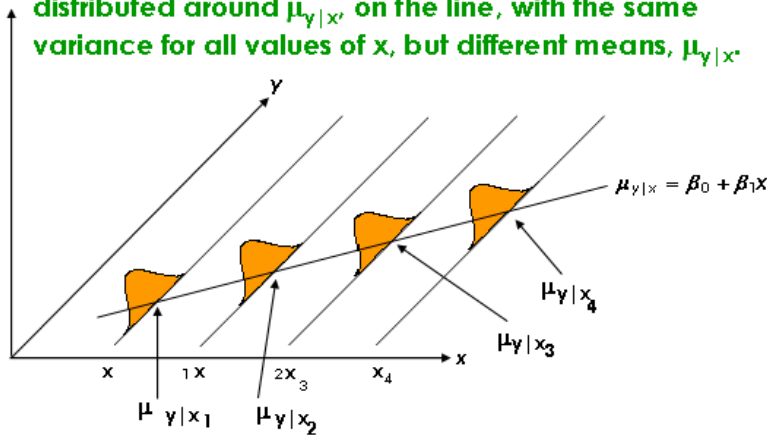
Assumptions

- ◆ The separate observations Y_1, Y_2, \dots, 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 $X=x$

For each value of x , the values of y are normally distributed around $\mu_{y|x}$, on the line, with the same variance for all values of x , but different means, $\mu_{y|x}$.



Here, $\sigma_{y|x_1}^2 = \sigma_{y|x_2}^2 = \sigma_{y|x_3}^2 = \sigma_{y|x_4}^2$

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

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

$\beta_1 = 0$	$\beta_1 \neq 0$
Due model MSR has expected value $\sigma_{Y X}^2$	Due model MSR has expected value $\sigma_{Y X}^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$
Due residual MSE has expected value $\sigma_{Y X}^2$	Due residual MSE has expected value $\sigma_{Y X}^2$
F = (MSR)/MSE will be close to 1	F = (MSR)/MSE will be LARGER than 1

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

Stata illustration with annotations in red.

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
Residual	.007093416	9	.000788157	R-squared =	0.9983 = SSQ(model)/SSQ(TOTAL)
-----+-----				Adj R-squared =	0.9981 = R ² adjusted for n and # of X
Total	4.22815076	10	.422815076	Root MSE =	.02807 = Square root of MSQ(residual)

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

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	1	$SSR = \sum_{i=1}^n (\hat{Z}_i - \bar{Z})^2 = 4.22063$	$SSR/1 = 4.22063$
Error	$(n-2) = 9$	$SSE = \sum_{i=1}^n (Z_i - \hat{Z}_i)^2 = 0.00705$	$SSE/(n-2) = 7/838E-04$
Total, corrected	$(n-1) = 10$	$SST = \sum_{i=1}^n (Z_i - \bar{Z})^2 = 4.22768$	

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

- *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. $F = MSQ(\text{Regression})/MSQ(\text{Residual})$
= $4.22063/0.0007838$
= 5384.94 with $df = 1, 9$

Achieved significance < 0.0001 . Reject H_0 . Conclude that the fitted line is a significant improvement over the average LOGWT.

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

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 \bar{Y} than does \bar{Y} ?

Assumptions: As before.

H_0 and H_A :

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

Test Statistic:

$$F = \frac{msq(\text{regression})}{msq(\text{residual})}$$

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

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 =

$$pr \left[F_{1,(n-2)} \geq \frac{msq(\text{model})}{msq(\text{residual})} \mid \beta_1 = 0 \right] = pr [F_{1,9} \geq 5384.94] \ll .0001$$

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

Evaluate:

Under the null hypothesis that $\beta_1 = 0$, the chances of obtaining a value of F that is so far away from the expected value of 1, with a value of 5394.94, were less than 1 chance in 10,000. This is a very small likelihood!

Interpret:

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

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

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 $\{ \text{t-score for slope} \}^2 = \{ \text{overall F} \}$

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

Assumptions: As before.

H_0 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 $\hat{\beta}_1$

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

Our t-score is therefore:

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

$$df = (n - 2)$$

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 its 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: “What are the chances of obtaining our value of the t or one that is more far away from 0 if we believe the null hypothesis that $\beta_1 = 0$ ”?

Calculations:

For our data, we obtain p-value =

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

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 much better job of explaining the variability in LOGWT than the sample mean.

3) Test of the Intercept, β_0

This addresses the question: Does the 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_0 and H_A :

$$H_o: \beta_0 = 0$$

$$H_A: \beta_0 \neq 0$$

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

Test Statistic:

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

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

Our t-score is therefore:

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

$$df = (n - 2)$$

Again, 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	0.000	.1898356 .2019462
_cons	-2.689255	.030637	-87.78 = -2.689255/.030637	0.000	-2.75856 -2.619949

This t=-87.78 says “the estimated intercept is estimated to be 87.78 standard error units away from its 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: “What are the chances of obtaining our value of the t or one that is more far away from 0 if we believe the null hypothesis that $\beta_0 = 0$ ”?

Calculations:

p-value =

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

Evaluate:

Under the null hypothesis 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.

Interpret:

The inference is that the straight line relationship between $Z = \text{LOGWT}$ and $X = \text{AGE}$ does not pass through the origin.

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

8. Confidence Interval Estimation

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

The confidence intervals here have 3 elements:

- 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 = $\hat{\beta}_1$

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

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

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

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

3) MEAN at $X=x_0$

$$\text{estimate} = \hat{Y}_{X=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

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

4) INDIVIDUAL with $X=x_0$

$$\text{estimate} = \hat{Y}_{X=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

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

Example, continued

Z=LOGWT to X=AGE.

Stata yielded the following fit:

z	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]		
x	.1958909	.0026768	73.18	0.000	.1898356	.2019462	← 95% CI for Slope β_1
_cons	-2.689255	.030637	-87.78	0.000	-2.75856	-2.619949	

95% Confidence Interval for the Slope, β_1

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

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

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

95% Confidence Interval for Slope β_1 = Estimate \pm (confidence coefficient)*SE
 = $0.19589 \pm (2.26)(0.00268)$
 = (0.1898, 0.2019)

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

95% Confidence Interval for the Intercept, β_0

z	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x	.1958909	.0026768	73.18	0.000	.1898356 .2019462
_cons	-2.689255	.030637	-87.78	0.000	-2.75856 -2.619949 ← 95% CI for intercept β_0

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

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

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
 = $-2.68925 \pm (2.26)(0.03064)$
 = $(-2.7585, -2.6200)$

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

Stata Example, continued**Confidence Intervals for MEAN of Z at Each Value of X.**

```

. regress z x
. predict zhat, xb

. ** Obtain SE for MEAN of Z given X
. 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

. ** 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 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

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

Stata Example, continued

Confidence Intervals for INDIVIDUAL PREDICTED Z at Each Value of X.

```
. regress z x
. predict zhat, xb

. ** Obtain SE for INDIVIDUAL PREDICTION of Z at given X
. 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 PubHlth 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

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

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

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

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

Formula for Estimate of Pearson Product Moment Correlation from a Sample

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

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

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 $\hat{\beta}_1$.

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

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.

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

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_0 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 $Z=LOGWT$ to $X=AGE$:

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

Source	SS	df	MS	Number of obs =	11
Model	4.22105734	1	4.22105734	F(1, 9) =	5355.60
Residual	.007093416	9	.000788157	Prob > F =	0.0000
				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

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

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 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.

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis