

Week One

[Help Center](#)

introduction

Review of Basic Statistics and Introduction to Regression

During our first week together, we will review some basic statistical concepts including bias, confidence intervals and p-values. We will also begin our introduction into basic regression concepts and correlation. We will end our week with an introduction to STATA statistical software.

There is no homework for this first week, however, please be sure to check out the [Download STATA](#) page and check out some of the tutorial websites listed there to begin familiarizing yourself with STATA. You'll have the opportunity to install STATA later on during this week.

*Please be advised that you will be **unable** to install STATA without the software license code until after we email them on or around Thursday, March 26. They will be sent to the email address you use for logging into Coursera.*

Lectures

Please click on the links below to access the video lectures for this first week

- [Introduction and Course Highlights](#)
- [Central Tendency and Variability](#)
- [Sampling Distribution](#)
- [Bias](#)
- [Confidence Intervals](#)
- [p-Values](#)
- [Regression and Correlation](#)
- [Introduction to STATA](#)

Lecture Material

Please click on the link below to download the slides of the first week

Week One "Review of Basic Statistics and Introduction to Regression"

Conversations

Please join in the conversations around regression analysis in our **community forums** area. You can ask and answer questions and discover insights and help for yourself and others as we come together to encourage each other in our exploration.

Key Terms

Below are definitions of some important terms covered this week:

- **Degrees of Freedom:** The number of values in the calculation of a test statistic that are free to vary. For example, imagine a set of 5 test scores whose mean is 90. You are 'free' to pick the value of the first four test scores. However, once you know the first four values, the final value is fixed. In this example you have 4 degree of freedom, because you were 'free' to choose the first four scores, at which point the final score is fixed.
- **Central Limit Theorem:** A probability theorem which states that the mean of a large number of observations will be approximately normally distributed.
- **Total Sum of Squares:** The total sum of squares is a measure of the total variability of a set of scores around the mean of the dependent variable (outcome).
- **Variance:** A measure of the spread of data, or a measure of variability in data. You can calculate the variance by averaging the squared deviations of each observation in comparison to the mean.
- **Slope of Regression Line:** Coefficient of a predictor/independent variable in a regression model. The slope indicates how the outcome changes with increases in the predictor value. This slope can be interpreted as 'for every unit increase in predictor X, Y will increase or decrease by (slope value)'.

Homework

There is no homework this week.

Please visit the **Download STATA** page to learn more about the statistical software that will be demonstrated in our course homework starting in week two.

You can visit the [homework page](#) to also learn more.

Quiz

After you've gone through the materials for this week please be sure to visit the [quizzes area](#) to complete this week's quiz.



[Creative Commons License](#)

Header image is used and altered with permission from [Kevin Dooley](#) according to its [Creative Commons Attribution 2.0 Generic License](#).

Created Mon 2 Feb 2015 11:48 AM PST

Last Modified Mon 23 Mar 2015 10:57 AM PDT

Week One

[Help Center](#)

introduction

Review of Basic Statistics and Introduction to Regression

During our first week together, we will review some basic statistical concepts including bias, confidence intervals and p-values. We will also begin our introduction into basic regression concepts and correlation. We will end our week with an introduction to STATA statistical software.

There is no homework for this first week, however, please be sure to check out the [Download STATA](#) page and check out some of the tutorial websites listed there to begin familiarizing yourself with STATA. You'll have the opportunity to install STATA later on during this week.

*Please be advised that you will be **unable** to install STATA without the software license code until after we email them on or around Thursday, March 26. They will be sent to the email address you use for logging into Coursera.*

Lectures

Please click on the links below to access the video lectures for this first week

- [Introduction and Course Highlights](#)
- [Central Tendency and Variability](#)
- [Sampling Distribution](#)
- [Bias](#)
- [Confidence Intervals](#)
- [p-Values](#)
- [Regression and Correlation](#)
- [Introduction to STATA](#)

Lecture Material

Please click on the link below to download the slides of the first week

Week One "Review of Basic Statistics and Introduction to Regression"

Conversations

Please join in the conversations around regression analysis in our **community forums** area. You can ask and answer questions and discover insights and help for yourself and others as we come together to encourage each other in our exploration.

Key Terms

Below are definitions of some important terms covered this week:

- **Degrees of Freedom:** The number of values in the calculation of a test statistic that are free to vary. For example, imagine a set of 5 test scores whose mean is 90. You are 'free' to pick the value of the first four test scores. However, once you know the first four values, the final value is fixed. In this example you have 4 degree of freedom, because you were 'free' to choose the first four scores, at which point the final score is fixed.
- **Central Limit Theorem:** A probability theorem which states that the mean of a large number of observations will be approximately normally distributed.
- **Total Sum of Squares:** The total sum of squares is a measure of the total variability of a set of scores around the mean of the dependent variable (outcome).
- **Variance:** A measure of the spread of data, or a measure of variability in data. You can calculate the variance by averaging the squared deviations of each observation in comparison to the mean.
- **Slope of Regression Line:** Coefficient of a predictor/independent variable in a regression model. The slope indicates how the outcome changes with increases in the predictor value. This slope can be interpreted as 'for every unit increase in predictor X, Y will increase or decrease by (slope value)'.

Homework

There is no homework this week.

Please visit the **Download STATA** page to learn more about the statistical software that will be demonstrated in our course homework starting in week two.

You can visit the [homework page](#) to also learn more.

Quiz

After you've gone through the materials for this week please be sure to visit the [quizzes area](#) to complete this week's quiz.



Header image is used and altered with permission from [Kevin Dooley](#) according to its [Creative Commons Attribution 2.0 Generic License](#).

Created Mon 2 Feb 2015 11:48 AM PST

Last Modified Mon 30 Mar 2015 8:39 AM PDT

Week Two

[Help Center](#)

Foundations of Regression Analysis

This week we will start with linear regression for a single explanatory variable. This builds upon the basic statistical concepts required for regression analysis covered during week one.

In this week you will:

- learn to make a scatter plot and fit a regression line using STATA.
- be acquainted with the assumption of homoscedasticity for linear regression.
- learn to draw inference based on hypothesis testing and estimation of confidence interval.

We will end this week learning about prediction interval and also have a chance to see a demonstration on STATA that will be helpful in completing your first homework.

Lectures

Please click on the links below to access the video lectures for this second week

- [Linear Regression I](#)
- [Linear Regression II](#)
- [Assumptions of Linear Regression](#)
- [Hypothesis Testing and Confidence Interval](#)
- [Confidence Interval and Homework](#)

Lecture Material

Please click on the link below to download the slides of the second week

[Week Two "Fundamentals of Regression Analysis"](#)

Conversations

Please join in the conversations around regression analysis in our **community forums** area. You can ask and answer questions and discover insights and help for yourself and others as we come together to encourage each other in our exploration.

Key Terms

Below are definitions of some important terms covered this week:

- **Scatter Plot/Diagram:** Scatter plot is a plot to display values for two variables of the dataset. It is a plot of Y values versus X values. Making a scatter plot should be the first step in regression analysis.
- **Outlier:** An outlier is an observation point that is far from the other observations. Outlier can affect the slope of the regression line. Scatter plots are useful in detecting outliers.
- **Residuals:** Residual is the difference between observed value and the predicted value. Predicted values are obtained by using regression equation.
- **Homoscedasticity:** Homoscedasticity is also known as the homogeneity of variance. It means that the dependent variable exhibits similar amounts of variance across the range of values for an independent variable. This is an important assumption in linear regression analysis.
- **Naïve Model:** The Naïve Model is a model without any explanatory/ predictor variables. This model only contains an intercept value with no other covariates. Therefore, every observation in the dataset will be modeled with the mean outcome. The mean of the Independent variable is the naïve model.
- **Prediction Interval:** Prediction interval is an interval for a random variable that is yet to be observed. This sort of an interval is for an individual future observation and not for the true population mean.

Homework

Navigate to the **Week Two Homework** page to view and download the homework for this week.

Quiz

After you've gone through the materials for this week please be sure to visit the [quizzes area](#) to complete this week's quiz.



[Header image](#) is used and altered with permission from [Kevin Dooley](#) according to its [Creative Commons Attribution 2.0 Generic License](#).

Created Mon 2 Feb 2015 12:13 PM PST

Last Modified Tue 31 Mar 2015 8:58 AM PDT

Week Three

[Help Center](#)

The Correlation Coefficient and the ANOVA table

This week we will learn methods to measure the strength of the relationship between two variables using the correlation coefficient and the coefficient of determination. Furthermore, we will learn how to compute and interpret an analysis of variance (ANOVA) table for straight line regression. The homework will involve making ANOVA tables for several regressions, performing F-tests on these regressions, and interpreting the results.

Lectures

Please click on the links below to access the video lectures for this third week:

- [The Correlation Coefficient: \$r\$](#)
- [The Coefficient of Determination: \$r^2\$](#)
- [Hypothesis Testing of \$r\$](#)
- [Sums of Squares](#)
- [Analysis of Variance: The ANOVA table](#)
- [The F-Test of Significance](#)
- [Reviewing STATA Output - Homework](#)

Lecture Material

Please click on the link below to download the slides of the third week

[Week Three "The Correlation Coefficient and the ANOVA table"](#)

Conversations

Please join in the conversations around regression analysis in our [community forums](#) area. You can ask and answer questions and discover insights and help for yourself and others as we come together to encourage each other in our exploration.

Key Terms

Below are definitions of some important terms covered this week:

- **Correlation Coefficient (r):** Measure of the strength of the relationship between 2 random variables
- **Coefficient of Determination (r^2 / ρ_{xy}^2):** proportion in the Y "explained by" knowledge of X
- **Fisher's Z Transformation:** Method to transform r into a statistic that is approximately normal for the purpose of hypothesis testing

Homework

Please watch the following video, [Homework Highlights from Week Two](#), to review the homework from last week.

Navigate to the [Week Three Homework](#) page to view and download the homework for this week.

Quiz

After you've gone through the materials for this week please be sure to visit the [quizzes area](#) to complete this week's quiz.



[Header image](#) is used and altered with permission from [Kevin Dooley](#) according to its [Creative](#)

Commons Attribution 2.0 Generic License.

Created Mon 2 Feb 2015 12:14 PM PST

Last Modified Mon 6 Apr 2015 9:25 AM PDT

Week Four

[Help Center](#)

polynomials

Polynomial Regression

This week we will introduce the concept of polynomial regression analysis where the k_{th} order of a single predictor is used to model the dependent variable. You will learn how to create and interpret the ANOVA table for polynomial regression based on the techniques taught in week 3.

We will explain the use of hypothesis testing in concluding whether or not you need to make use of polynomial regression. Furthermore, we will illustrate all the concepts using an interesting example which will also be helpful for the homework for this week.

The homework involves fitting a series of models, including models with polynomial terms in order to determine which model best fits the data.

Lectures

Please click on the links below to access the video lectures for this fourth week

- [Polynomial Regression I: Introduction](#)
- [Polynomial Regression II](#)
- [Polynomial Regression III](#)
- [Example: Dose Response Study and Assessing Multicollinearity](#)
- [Example: Potential Energy](#)

Lecture Material

Please click on the link below to download the slides of the fourth week

[Week Four "Polynomial Regression"](#)

Conversations

Please join in the conversations around regression analysis in our [community forums](#) area. You can ask and answer questions and discover insights and help for yourself and others as we come together to encourage each other in our exploration.

Key Terms

Below are definitions of two important terms covered this week:

- **Polynomial Regression:** A type of linear regression model where the relationship between the dependent variable Y , and the independent variable, X is modeled as a k_{th} degree polynomial term.
- **Sums of Squares Due Regression:** The sums of the squared deviations of each predicted value from the mean.

Homework

Please watch the following video, [Homework Highlights from Week Three](#), to review the homework from last week.

Navigate to the [Week Four Homework](#) page to view and download the homework for this week.

Quiz

After you've gone through the materials for this week please be sure to visit the [quizzes area](#) to complete this week's quiz.



[Header image](#) is used and altered with permission from [Kevin Dooley](#) according to its [Creative Commons Attribution 2.0 Generic License](#).

Created Mon 2 Feb 2015 12:14 PM PST

Last Modified Mon 13 Apr 2015 6:36 AM PDT

Week Five

[Help Center](#)

multiple regression

Fundamentals of Multiple Linear Regression Analysis

In previous weeks we covered simple [linear regression](#) and [polynomial regression](#). This week we extend what we've learnt so far and begin exploring multiple linear regression.

We will discover assumptions involved in the multiple linear regression model and also learn various model building techniques. We will also make use of partial F-test, previously covered in [week 4](#), to ascertain if a particular variable has a significant impact on the model under consideration.

Throughout this week's lectures, we will see demonstrations of STATA and you will be taught how to read the STATA output and arrive at a conclusion.

Lectures

Please click on the links below to access the video lectures for this first week

- [Multiple Regression: A Graphical Interpretation and Assumptions I](#)
- [Multiple Regression: Assumptions II and Least Squares Estimation](#)
- [Multiple Regression: Computer Output](#)
- [Multiple Regression: A Step by Step Review](#)
- [Hypothesis Testing I: F-test](#)
- [Hypothesis Testing II: Partial F-test and Homework](#)

Lecture Material

Please click on the link below to download the slides of the fifth week

[Week Five "Multiple Regression"](#)

Conversations

Please join in the conversations around regression analysis in our [community forums](#) area. You can ask and answer questions and discover insights and help for yourself and others as we come together to encourage each other in our exploration.

Key Terms

Below are definitions of some important terms covered this week:

- **Stepwise Forward Selection:** This procedure involves starting with no variables in the model and then adding each variable at a step until R^2 gets as high as possible.
- **Stepwise Backward Strategy:** This strategy involves starting with all possible variables in the model and then deleting each variable at a step as long as the reduction in R^2 is not significant.
- **Tolerance:** Tolerance is a measure used in identifying multicollinearity. It is the reciprocal of the variance inflation factor (VIF) i.e. $\text{Tolerance} = 1/\text{VIF}$. If the tolerance is less than 0.01 then it is an indicator of multicollinearity.

Homework

Please watch the following video, [Homework Highlights from Week Four](#), to review the homework from last week.

Navigate to the [Week Five Homework](#) page to view and download the homework for this week.

Quiz

After you've gone through the materials for this week please be sure to visit the [quizzes area](#) to complete this week's quiz.



Header image is used and altered with permission from [Kevin Dooley](#) according to its [Creative Commons Attribution 2.0 Generic License](#).

Created Mon 2 Feb 2015 12:14 PM PST

Last Modified Sun 19 Apr 2015 9:31 PM PDT

Week Six

[Help Center](#)

interactions

Dummy Variables, Interaction & Methods to Compare Straight Line Regressions

This week we will be learning how to code nominal scaled data through dummy variables. We will also review the statistical interaction, and how to interpret the coefficients for these interaction terms in a regression model. Finally, we will apply these concepts into two different methods of comparing straight-line regressions. The first method will involve fitting separate models and hand calculating a t-statistic, and the second, more elegant method will involve a single regression model and partial F-tests.

This week's homework will involve analyzing the equality of 2 straight-line regressions using a t-test. In this exercise, we'll be testing to see if mean grip strength is the same between 2 groups of people. Our homework will also involve regression analysis using a dummy coded independent variable.

We recommend that you do not watch the final *Week Six Homework Highlights* video until you have finished the homework for week 6, because we will be discussing the answers to the homework in that video.

Lectures

Please click on the links below to access the video lectures for this first week

- [Dummy Variables](#)
- [Statistical Interaction I: Introduction](#)
- [Statistical Interaction II: Graphing](#)
- [Comparing Straight Line Regressions: Method One](#)
- [Comparing Straight Line Regressions: Method One](#)
- [Comparing Straight Line Regressions: Method Two and Homework](#)

Lecture Material

Please click on the link below to download the slides of the sixth week

Week Six "Dummy Variables, Interaction and Methods to Compare Straight Line Regressions"

Conversations

Please join in the conversations around regression analysis in our **community forums** area. You can ask and answer questions and discover insights and help for yourself and others as we come together to encourage each other in our exploration.

Key Terms

Below are definitions of some important terms covered this week:

- **Dummy Variable:** Any variable in a regression equation that takes on a finite number of values; it is used to indicate categories of a nominal scaled variable.
- **Nominal Scaled Data:** Discrete classification of data, in which data are neither measured nor ordered but subjects are merely allocated to distinct categories.
- **Reference Cell Coding:** Method of coding in which the reference cell is coding as 0 (i.e. In the dummy coding of a nominal variable with k categories, the different categories will be coded as follows: 0, 1, ..., (k-1)).
- **Statistical Interaction:** The effect of one predictor variable is dependent on the levels of another predictor variable.
- **Parallel Lines (in Regression):** Regression lines with the same slope.
- **Coincident Lines (in Regression):** Regression lines with the same slope and intercept.

Homework

Please watch the following video, **Homework Highlights from Week Five**, to review the homework from last week.

Navigate to the **Week Six Homework** page to view and download the homework for this week.

After you complete the Week Six Homework, please watch the **Homework Highlights from Week Six** video to review the homework.

Quiz

After you've gone through the materials for this week please be sure to visit the [quizzes area](#) to complete this week's quiz.



Header image is used and altered with permission from [Kevin Dooley](#) according to its [Creative Commons Attribution 2.0 Generic License](#).

Created Mon 2 Feb 2015 12:14 PM PST

Last Modified Tue 28 Apr 2015 11:51 AM PDT

Applied Regression Analysis

Week 1

1. Review of basic statistical concepts
 - Central tendency and variability
 - Sampling distributions
 - Bias
 - Confidence intervals
 - p-values
2. Regression and correlation
3. Introduction to STATA

Stanley Lemeshow, Professor of Biostatistics
College of Public Health, The Ohio State University



THE OHIO STATE UNIVERSITY

Topics to be discussed:

- **Measures of Central Tendency**
- **Measures of Dispersion**
- **Degrees of Freedom**
- **Population Parameters vs Sample Statistics**
- **Sampling Distributions**
 - **Expected Values, Standard Errors**
 - **Unbiased vs Biased Estimators**
 - **Confidence Intervals**
 - **Hypothesis Testing**
 - **p-values**

Measures of Central Tendency

The Population Mean

Given a set of N values, X_1, X_2, \dots, X_N ,
the population mean is computed as:

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

The Sample Mean

Given a set of n values, x_1, x_2, \dots, x_n , their mean, denoted by \bar{x} , is defined by their sum divided by the number of observations, n :

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

e.g.

Suppose the scores on five exams are as follows:

$$X_1 : 90$$

$$X_2 : 80$$

$$X_3 : 95$$

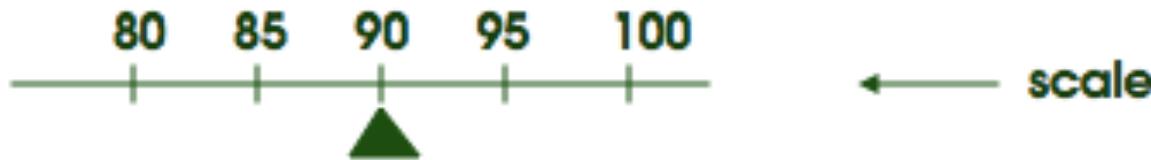
$$X_4 : 85$$

$$X_5 : 100$$

What is the average?

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{5} = \frac{90 + 80 + 95 + 85 + 100}{5} = \frac{450}{5} = 90$$

The mean is the “center of gravity” of the observations.



note also: $\sum(x_i - \bar{x}) = 0$

The Variance and Standard Deviation

The value $\sum_{i=1}^N (x_i - \mu)$ = sum of deviations about the sample mean. This $\equiv 0$.

The value $\sum_{i=1}^N (x_i - \mu) / N$ = mean deviation. This also $\equiv 0$.

We can avoid this difficulty by taking absolute values of the deviations.

- i.e., we can ignore the algebraic signs. Then, we could use the “mean absolute deviation”.

$$\frac{1}{N} \sum_{i=1}^N |x_i - \mu|$$

where $|x_i - \mu|$ is read "the absolute value of $x_i - \mu$ "

We don't use the mean absolute deviation because working with absolute values discourages further mathematical or theoretical treatment.

We also avoided the difficulty by squaring each deviation since, in the process, all negative signs disappear.

$$\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \text{VARIANCE}$$

This provides a MEASURE OF DISPERSION

The value $\sum_{i=1}^N (x_i - \mu)^2$ = sum of square deviations about the sample mean. (This is ≥ 0 .)

The value $\sum_{i=1}^N (x_i - \mu)^2 / N$ = mean square deviation.

(i) if observed values are identical,

$$\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = 0$$

(ii) if observed values are close together,

$$\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \text{ will be small}$$

(iii) if observed values scatter over a wide range,

$$\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \text{ will be correspondingly large.}$$

Computations from sample

x_i	\bar{x}	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
12	12	0	0
6	12	-6	36
15	12	3	9
3	12	-9	81
12	12	0	0
6	12	-6	36
21	12	9	81
15	12	3	9
18	12	6	36
12	12	0	0
$\sum x_i = 120$		☺	228
$\bar{x} = \frac{\sum x_i}{10} = 12$			

check on work

If not = 0, there is an error

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 288$$

and

$$\text{the variance } - s^2 = \frac{\sum_{i=1}^{10} (x_i - \bar{x})^2}{10 - 1} = 32$$

Note that the “sample variance” is defined as:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

where “ n ” is the number of observations in the “sample”.

In the previous discussion N was the number of observations in a “population”.

Q. Why do we use “ $n - 1$ ” in the denominator of the sample variance instead of “ n ”?

A. (i) if we selected many samples from a population, and computed a variance for each, then the average of the sample variance will not equal the population variance, σ^2 , if n is used in the denominator.

In fact, the average of these sample variances will be too small.

i.e., $\sum_{i=1}^k s_i^2 / k < \sigma^2$, where $k = \# \text{ samples drawn}$.

Alternatively, if we make each s^2 larger (i.e., by using $n - 1$ in

the denominator), we would find that $\sum_{i=1}^k s_i^2 / k = \sigma^2$.

- (ii) $n - 1$ represents "degrees of freedom". To calculate the variance, we first calculated \bar{x} . But, given \bar{x} , we have lost a degree of freedom since if you tell me \bar{x} , only $n - 1$ of the n observations are free to vary.

e.g., suppose I tell you that $\bar{x} = 90$ and $n = 5$. Then,

$$x_1 = 80$$

$$x_2 = 85$$

$$x_3 = 90$$

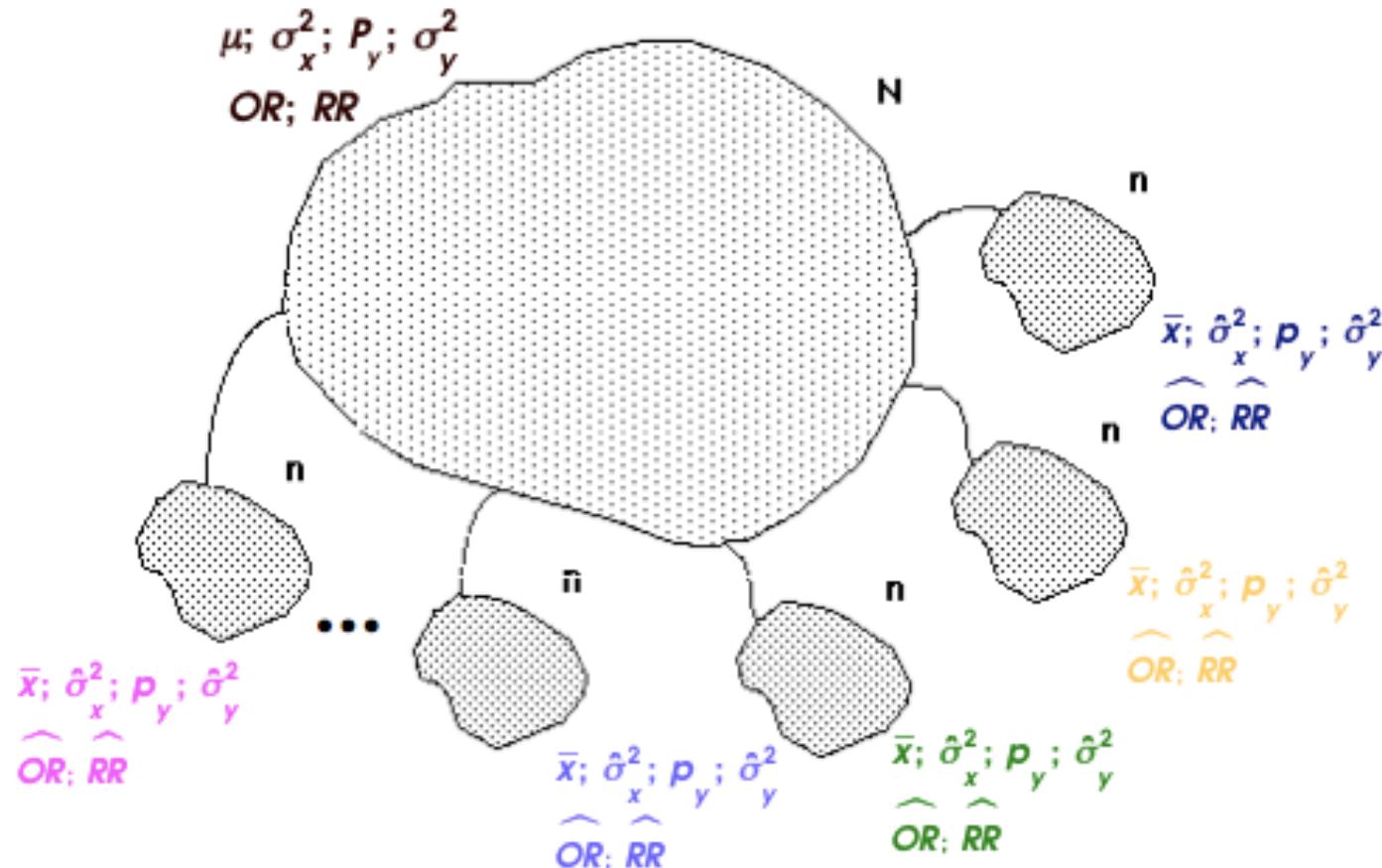
$$x_4 = 95$$

$$x_5 = \leftarrow \text{must} = 100$$

$$\text{since } \bar{x} = \frac{\sum x_i}{5} = 90 \Rightarrow \sum x_i = 450$$

Sampling Distributions

We consider all possible samples that can be generated using a particular sampling plan.



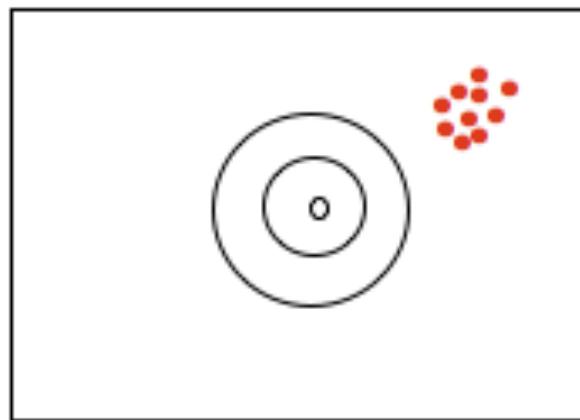
We select 10,000,000 samples from this population, each time estimating the population parameters.

Sample	\bar{x}_t
1	\bar{x}_1
2	\bar{x}_2
3	x_3
:	:
t	\bar{x}_t
:	:
10,000,000	$\bar{x}_{10,000,000}$

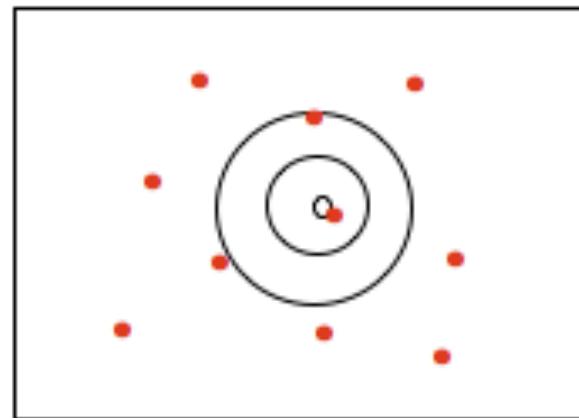
This could be any statistic

If, on average, the sample estimate is equal to the population parameter, then the estimate is said to be “unbiased”.

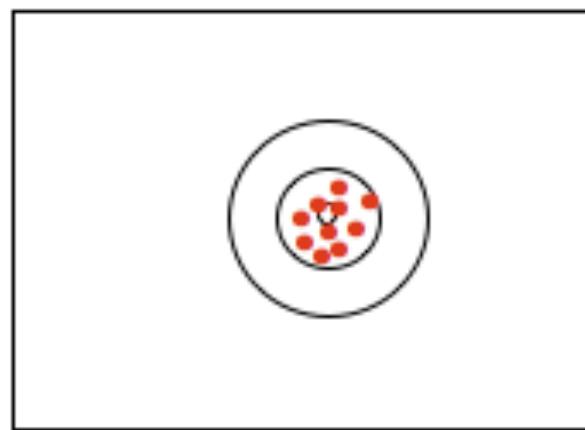
We not only look for estimators that are unbiased, but we also want estimators that have minimum variance.



high bias
low variance



low bias
high variance



low bias
low variance

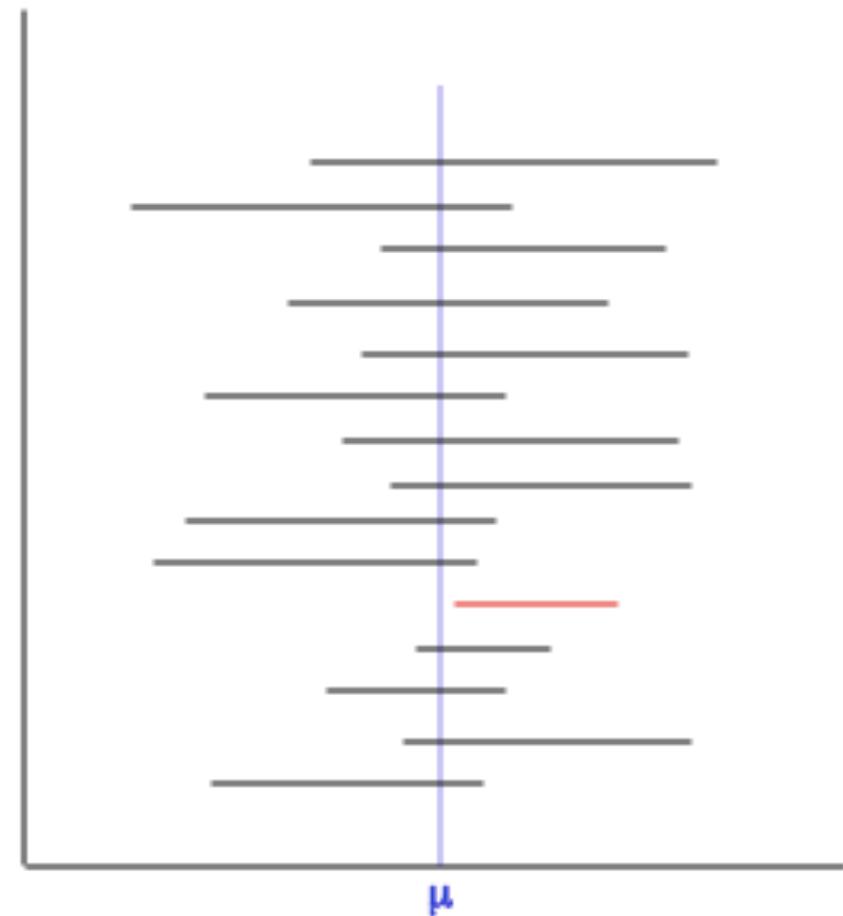
Confidence intervals:

Take the form:

$$\bar{x} - 1.96 \widehat{SE}(\bar{x}) \leq \mu \leq \bar{x} + 1.96 \widehat{SE}(\bar{x})$$

some multiplier

could be any parameter



def: 95% Confidence Interval:

Upon repeated sampling, 95% of intervals constructed in the same way will “cover” the true population parameter.

note: Once an interval is specified, it is either right or wrong

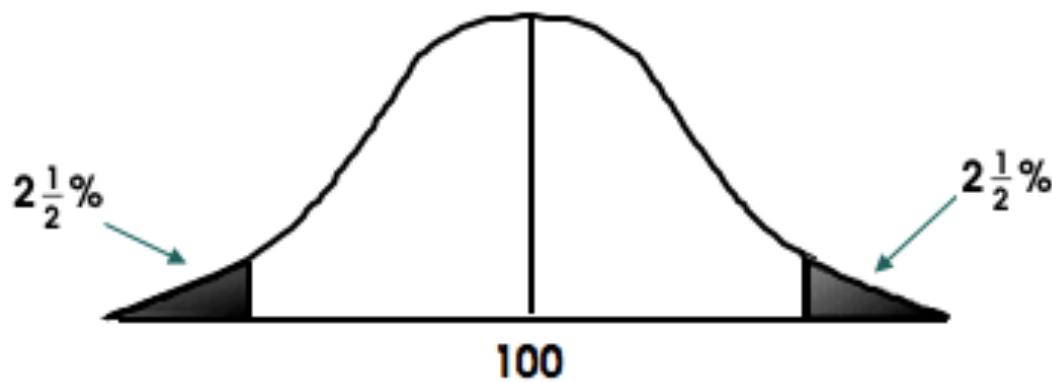
Hypothesis Testing:

$$H_0 : \mu = 100$$

$$H_a : \mu \neq 100$$

Statistician's role: Believe the null hypothesis until the evidence is so strong that the only reasonable response is to reject it.

Sampling distribution of \bar{x} :



The “Central Limit Theorem” tells us that the sampling distribution of the sample mean is normal...irrespective of the distribution of the original data.

If the null hypothesis is true, then the mean of the sampling distribution of the sample mean will = 100

If the sample mean is far from the hypothesized mean, then chances are that the hypothesis is false.

Type I error: Probability of rejecting the H_0 when H_0 is true.

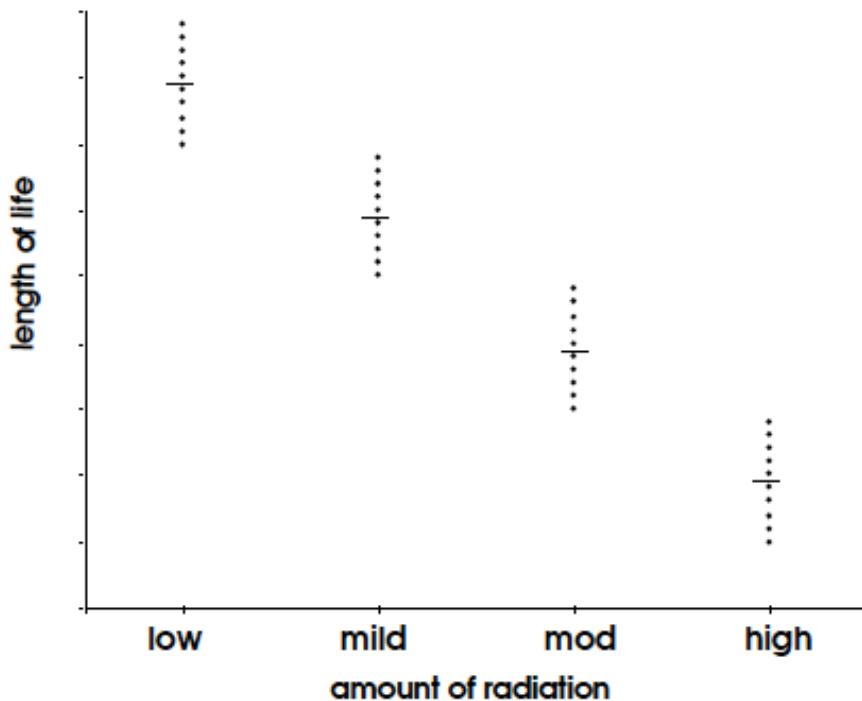
Type II error: Probability of failing to reject H_0 when H_0 is false.

p-value: Probability of observing a result as extreme or more extreme than the one observed given H_0 is true.

- a small p-value (e.g., $p \leq .05$) suggests that the results of a study are “statistically significant”
 - i.e., null hypothesis is probably false
- a large p-value (e.g., $p > .05$) suggests that the results of a study are not significant
 - i.e., there is no evidence to reject the null hypothesis

REGRESSION: considers the frequency distribution of one variable when another is held fixed at each of several levels

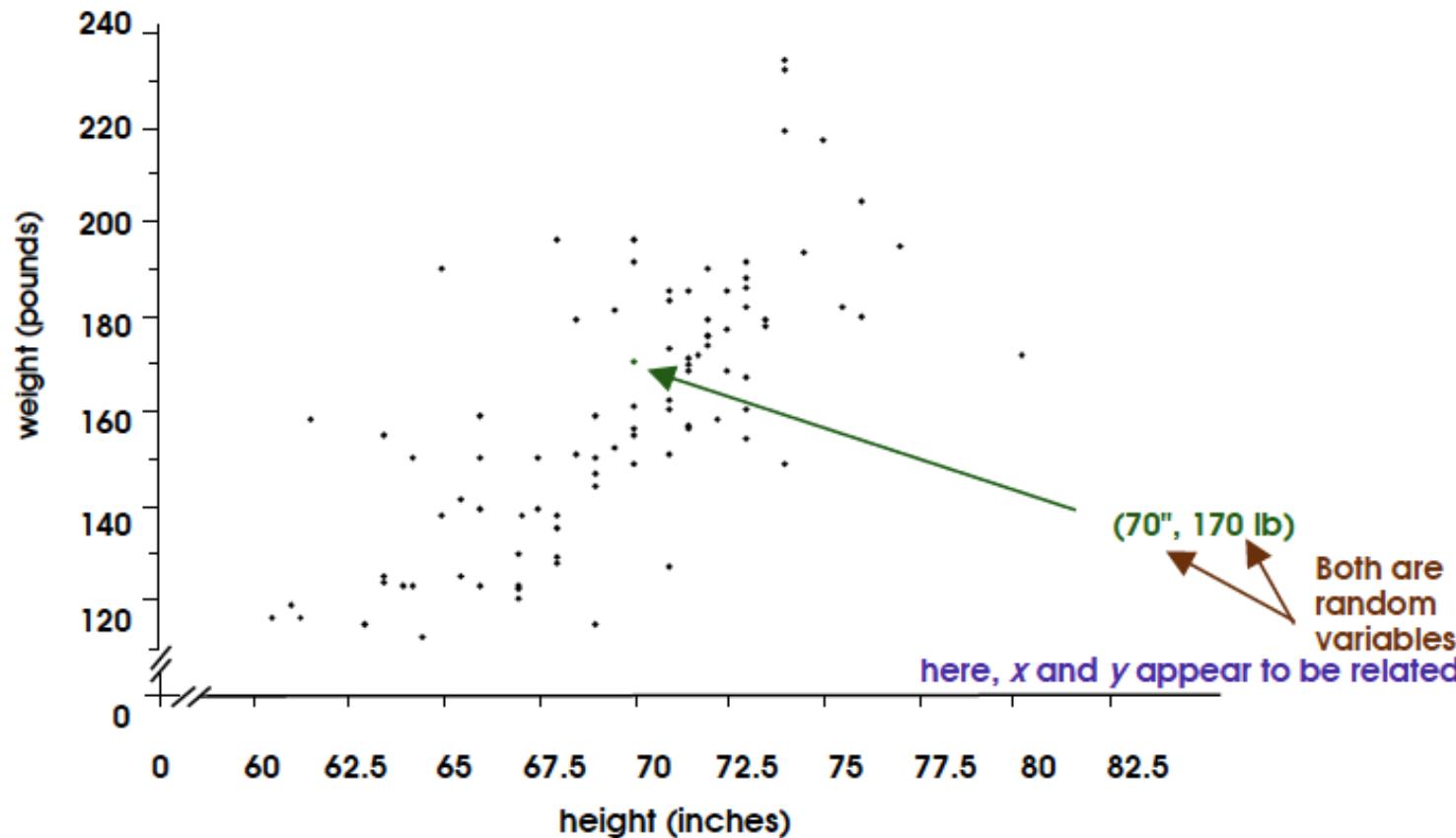
e.g. Suppose we have 40 mice and we expose them to varying amounts of radiation (4 levels)



In this problem, the variability in length of life (y = dependent variable) is studied with respect to particular levels of the amount of radiation (x = independent variable).

CORRELATION: considers the association of two random variables

e.g., Suppose we're studying height and weight

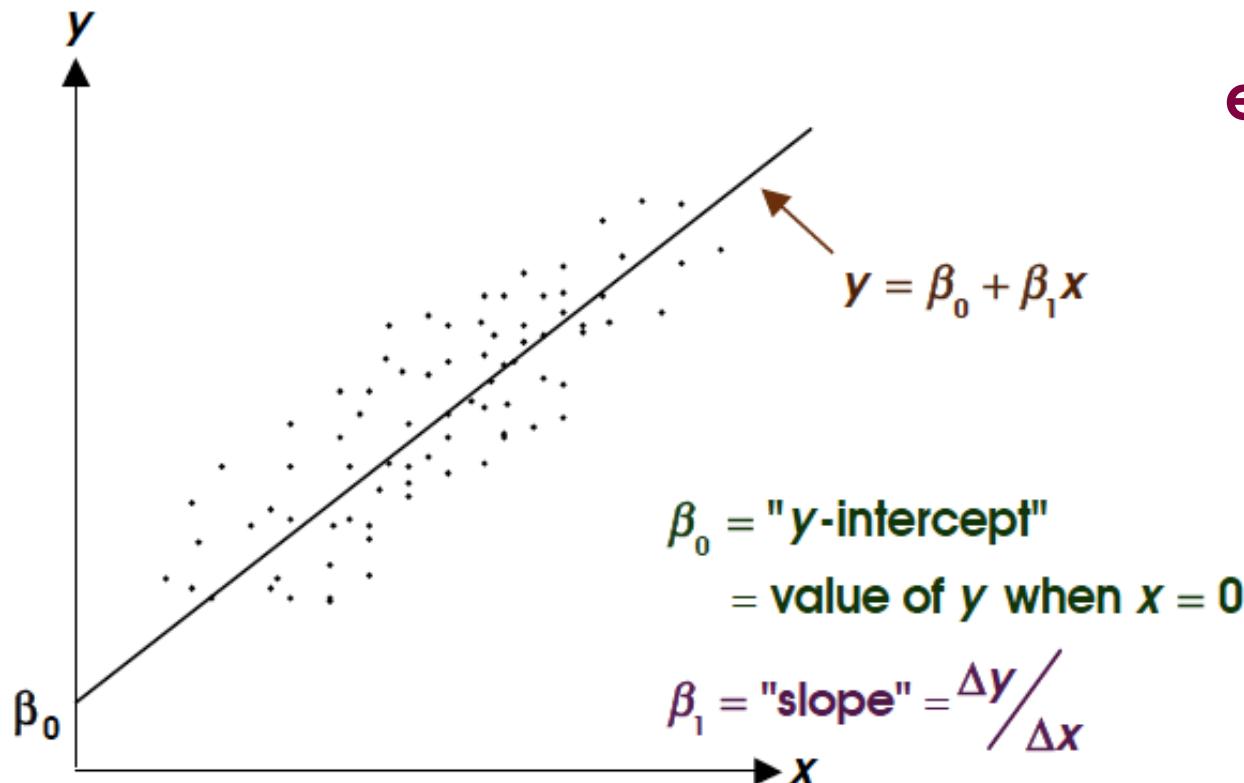


We will use similar techniques for the two types of problems but we should keep the distinction in mind.

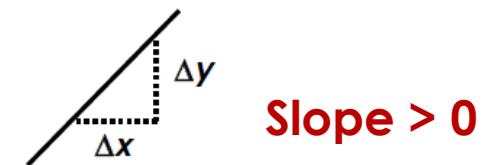
For both the correlation and regression problems it makes sense to describe the relationship between the two variables by fitting a line to the points.

- This line exhibits the "trend" in the data.

Let us review some elementary algebra:



e.g.,



— Slope = 0



Slope < 0

Applied Regression Analysis

Week 2

1. Linear regression I
2. Linear regression II
3. Assumptions for linear regression
4. Hypothesis testing and confidence intervals
5. Homework

Stanley Lemeshow, Professor of Biostatistics
College of Public Health, The Ohio State University



THE OHIO STATE UNIVERSITY

Suppose we have the following observations on systolic blood pressure and age for a sample of 30 individuals:

individual (<i>i</i>)	SBP (<i>y</i>)	AGE (<i>x</i>)	individual (<i>i</i>)	SBP (<i>y</i>)	AGE (<i>x</i>)
1	144	39	16	130	48
2	220	47	17	135	45
3	138	45	18	114	17
4	145	47	19	116	20
5	162	65	20	124	19
6	142	46	21	136	36
7	170	67	22	142	50
8	124	42	23	120	39
9	158	67	24	120	21
10	154	56	25	160	44
11	162	64	26	158	53
12	150	56	27	144	63
13	140	59	28	130	29
14	110	34	29	125	25
15	128	42	30	175	69

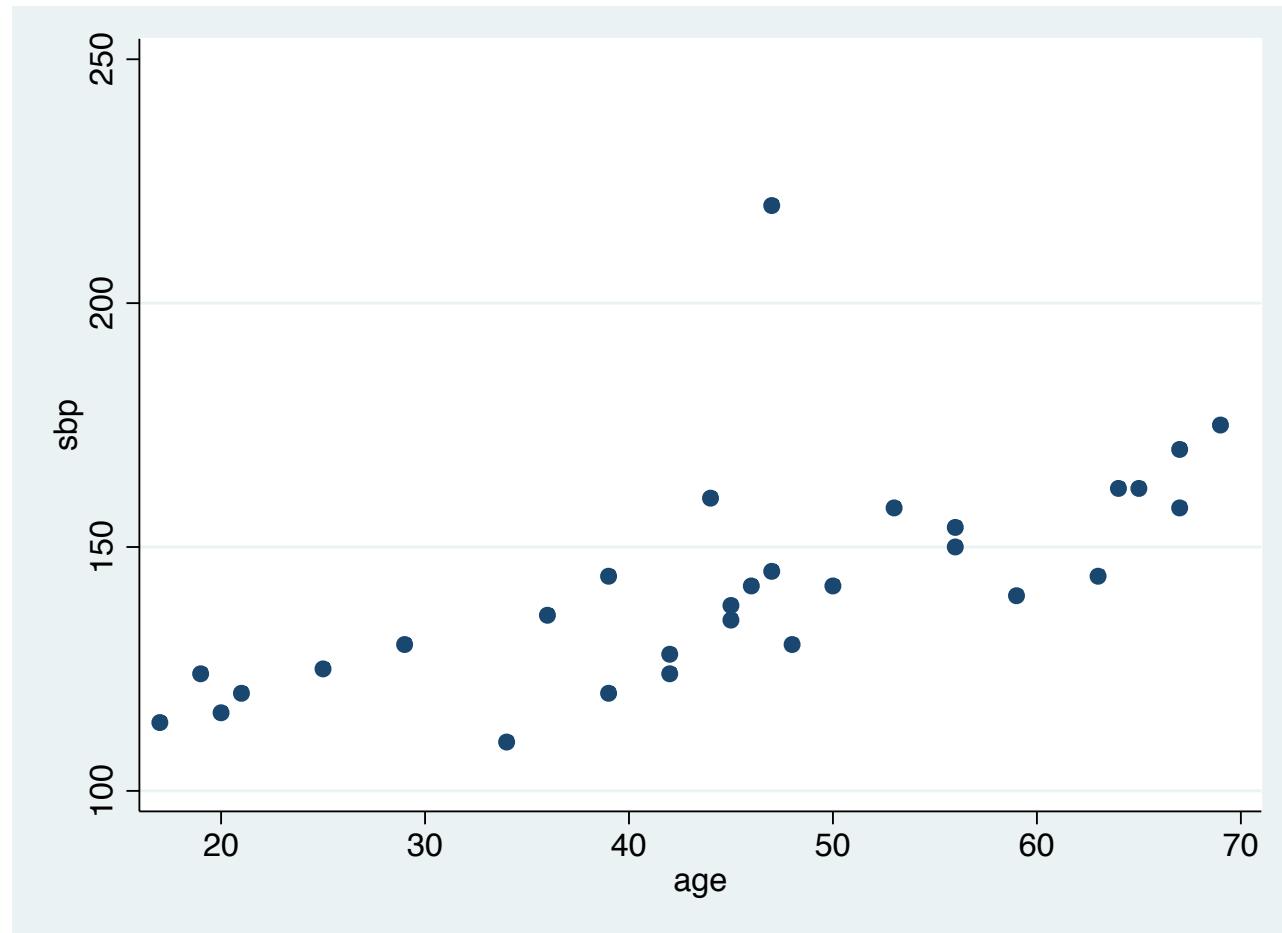
n

note: we have 30 pairs of observations that are denoted by

$$(x_1, y_1), (x_2, y_2), \dots, (x_{30}, y_{30}) = (39, 144), (47, 220), \dots, (69, 175).$$

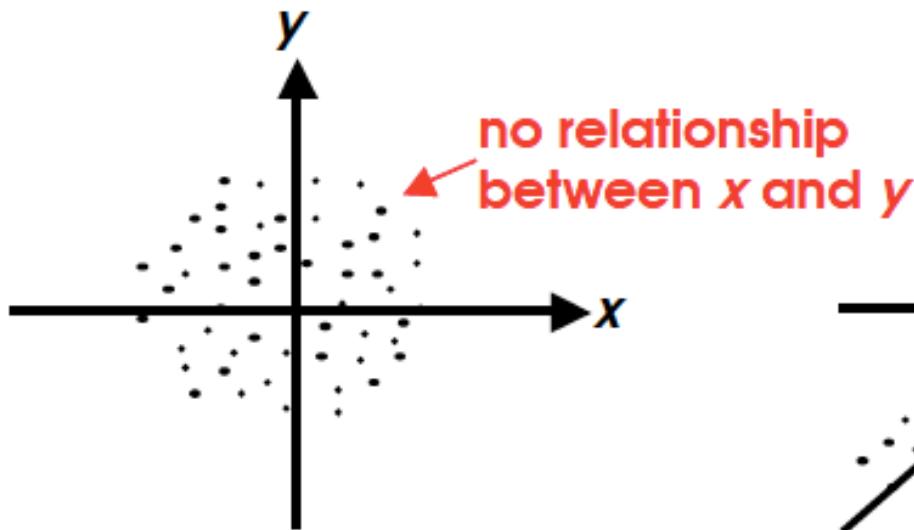
- These pairs may be considered as points in two dimensional space, so that we may plot them on a graph.
 - Such a graph is called a scatter diagram

. `twoway (scatter sbp age)`

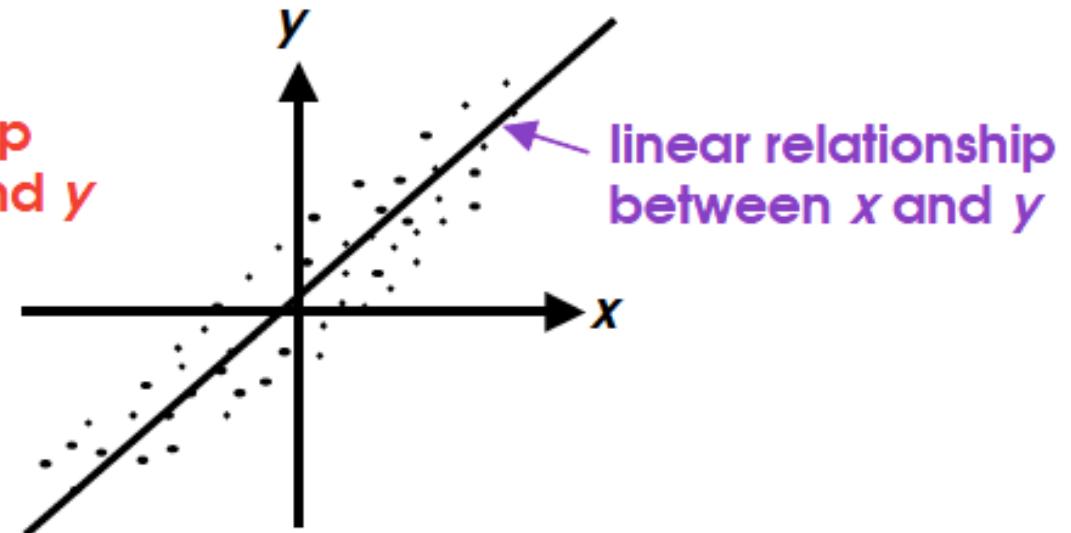


Note: AGE and SBP seem to be related.
How can this relationship be measured?

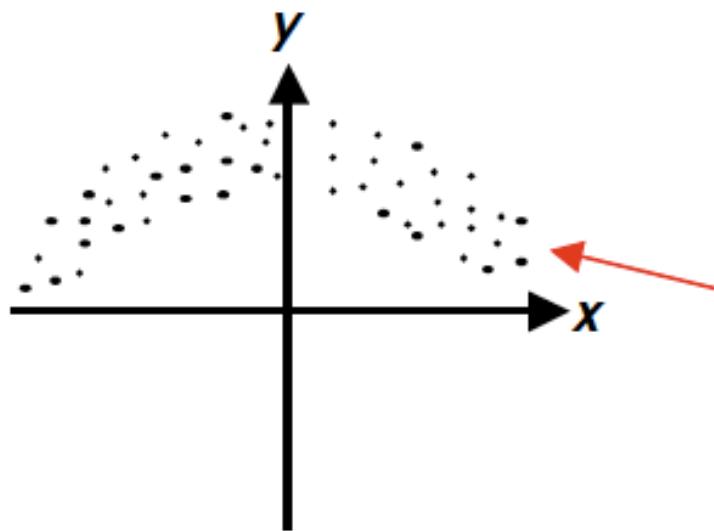
note: Scatter diagrams can take many shapes



no relationship
between x and y



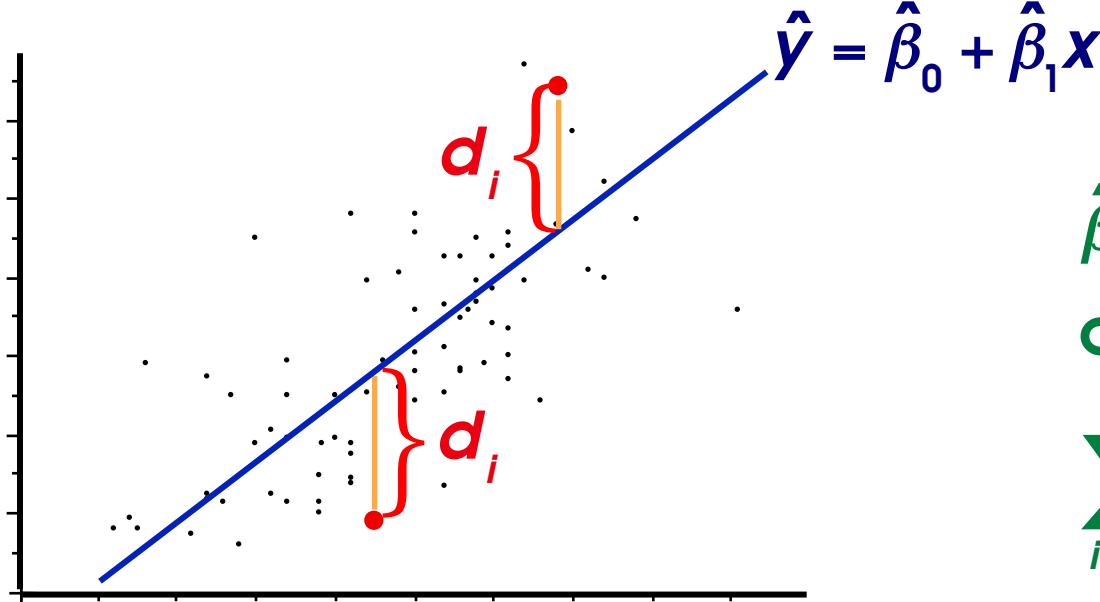
linear relationship
between x and y



non-linear relationship
between x and y

Now, given a set of data, how can we determine the line of regression?

- We are looking for that line that minimizes the vertical distances to the data points



$\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen such that

$\sum_{i=1}^n d_i^2$ is minimized

i.e., we want that line such that minimizes

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

The solution to the best-fit problem is obtained by solving, simultaneously, the following equations:

$$\begin{aligned}\sum y_i &= n\beta_0 + \beta_1 \sum x_i \\ \sum x_i y_i &= \beta_0 \sum x_i + \beta_1 \sum x_i^2\end{aligned} \quad \left\{ \begin{array}{l} \text{come from calculus - first} \\ \text{take derivative w.r.t. } \beta_0 \text{ then} \\ \text{w.r.t. } \beta_1 \text{ and set equal to zero} \end{array} \right.$$

Solving for β_0 and β_1 we obtain

$$\hat{\beta}_1 = \frac{\widehat{\text{Cov}}(x, y)}{\widehat{\text{Var}}(x)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

or
$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

Example

Using the previous data on 30 individuals where we measured

$x = \text{AGE}$

$y = \text{SBP}$

computations result in:

$$n = 30$$

$$\begin{aligned}\bar{y} &= 142.53 \\ \bar{x} &= 45.13\end{aligned}$$

$$\sum_{i=1}^n x_i y_i = 199,576$$

$$\sum_{i=1}^n x_i = 1,354$$

$$\sum_{i=1}^n y_i = 4,276$$

$$\sum_{i=1}^n x_i^2 = 67,894$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{199576 - \frac{(1354)(4276)}{30}}{67894 - \frac{(1354)^2}{30}} = 0.97$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 142.53 - (0.97)(45.13) = 98.71$$

Thus, the equation for this straight line is given by

$$\hat{y} = 98.71 + 0.97x$$

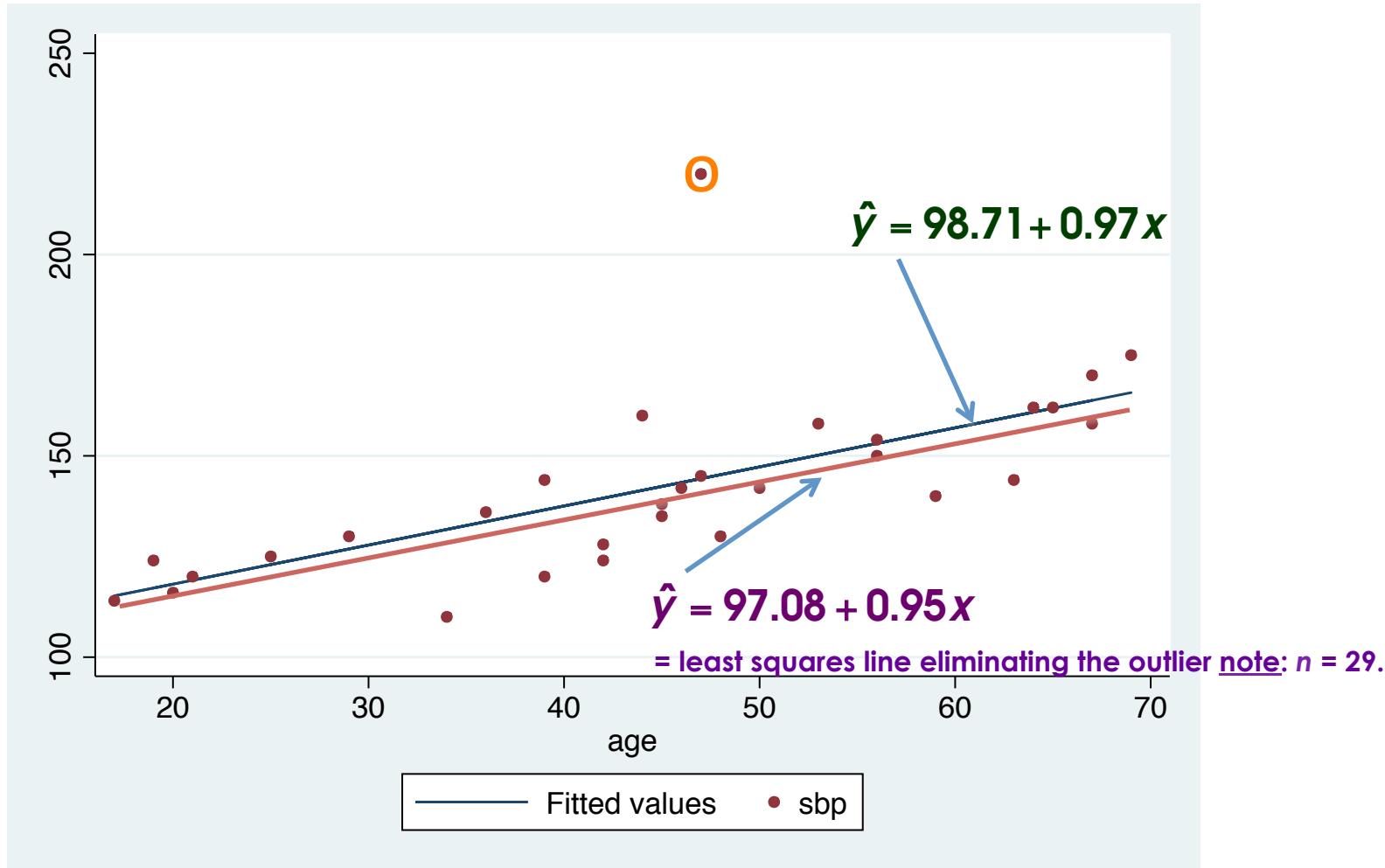
or equivalently by

$$\hat{y} = 142.53 + 0.97(x - 45.13)$$

This line should now be plotted on the scatter diagram.

e.g.,

- . scatter yhat sbp age, c(l .) s(i o)



Now, recall that

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Clearly, if $\text{SSE} = 0 \Rightarrow$ perfect fit

i.e., $y_i = \hat{y}_i$, all i

as the fit gets worse, SSE gets larger

. sum sbp age, detail

sbp

	Percentiles	Smallest		
1%	110	110		
5%	114	114		
10%	118	116	Obs	30
25%	125	120	Sum of Wgt.	30
50%	141		Mean	142.5333
		Largest	Std. Dev.	22.58125
75%	158	162		
90%	166	170	Variance	509.9126
95%	175	175	Skewness	1.291729
99%	220	220	Kurtosis	5.684303

age

	Percentiles	Smallest		
1%	17	17		
5%	19	19		
10%	20.5	20	Obs	30
25%	36	21	Sum of Wgt.	30
50%	45.5		Mean	45.13333
		Largest	Std. Dev.	15.2942
75%	56	65		
90%	66	67	Variance	233.9126
95%	67	67	Skewness	-.2395541
99%	69	69	Kurtosis	2.167069

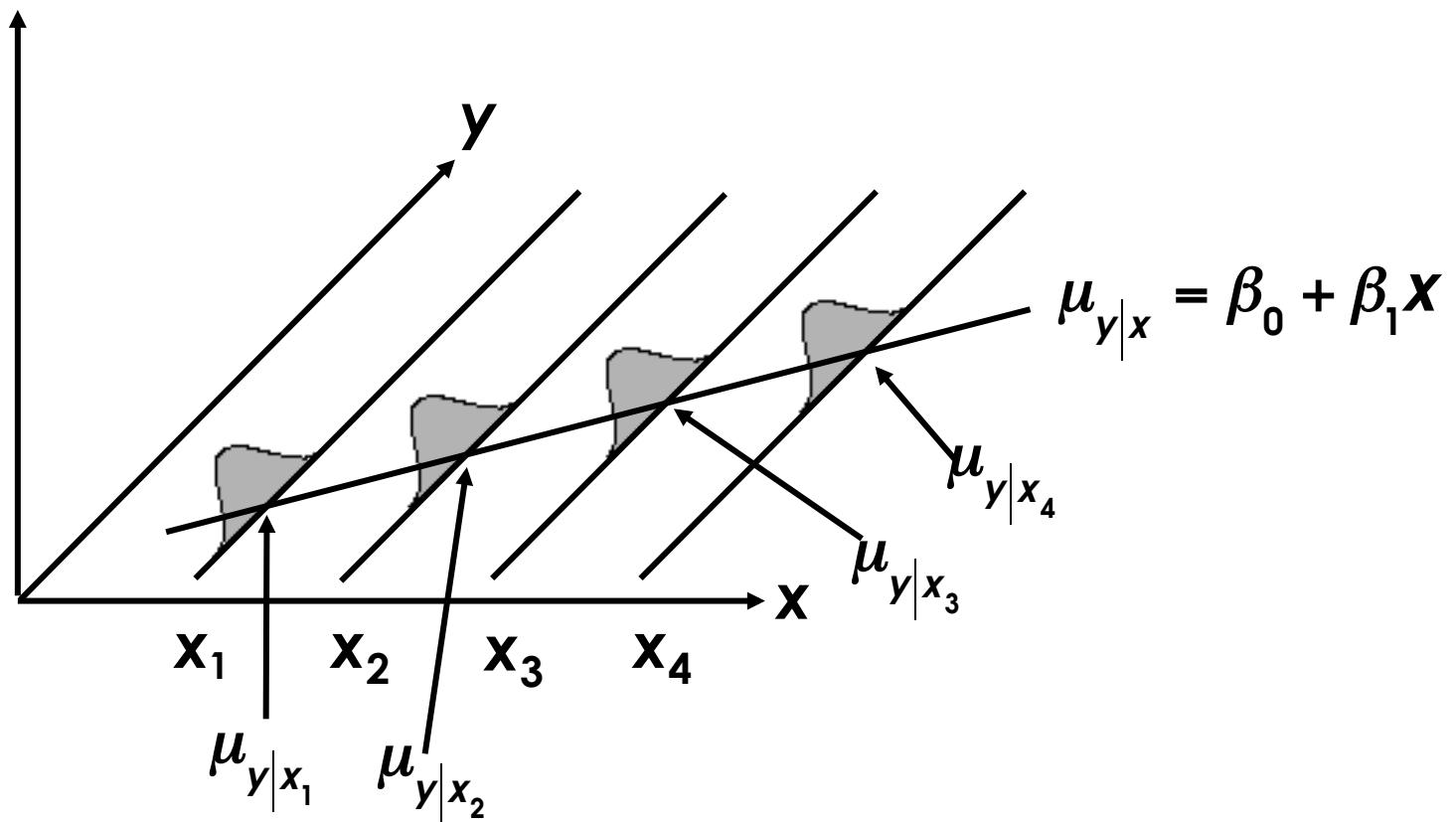
```
. regress sbp age
```

Source	SS	df	MS	Number of obs	=	30
Model	6394.02269	1	6394.02269	F(1, 28)	=	21.33
Residual	8393.44398	28	299.765856	Prob > F	=	0.0001
Total	14787.4667	29	509.912644	R-squared	=	0.4324
				Adj R-squared	=	0.4121
				Root MSE	=	17.314

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.9708704	.2102157	4.618	0.000	.5402629 1.401478
_cons	98.71472	10.00047	9.871	0.000	78.22969 119.1997

Now, one of the assumptions for regression analysis is that of homoscedasticity

(i.e., the variance of y is the same for any x)



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

i.e., $\sigma_{y|x_i}^2$ is the same for all i

We will denote this common value σ^2

i.e., $\sigma_{y|x}^2 = \sigma^2$ for all x .

An estimate of σ^2 is given by the formula

$$s_{y|x}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n-2} (SSE)$$

lose 2 d.f.
one for β_0
one for β_1

$$= \frac{n-1}{n-2} (s_y^2 - \hat{\beta}_1^2 s_x^2)$$

sample variance of y

sample variance of x

where

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

lose 1 d.f. for
and
estimating μ

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}{n-1}$$

in our example

$$s_y^2 = 509.91$$

$$s_x^2 = 233.91$$

$$\hat{\beta}_1 = 0.97$$

$$s_{y|x}^2 = \frac{n-1}{n-2} (s_y^2 - \hat{\beta}_1^2 s_x^2) = \frac{29}{28} (509.91 - .097^2 (233.91))$$

$$s_{y|x}^2 = 299.77$$

$\sqrt{s_{y|x}^2} = s_{y|x}$ is called the "standard error of estimate"

here

$$s_{y|x} = \sqrt{s_{y|x}^2} = \sqrt{299.77} = 17.31$$

Now, if we assume that for any fixed value of x , y has a normal distribution, we can test hypotheses and construct confidence intervals for β_0 or β_1 .

Under this assumption, it can be shown that

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)\right)$$

and

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_x^2(n-1)}\right)$$

Since we don't know σ^2 we estimate it with $s_{y|x}^2$ and use the t -distribution with $n - 2$ degrees of freedom.

First consider β_1

In order to test $H_0 : \beta_1 = \beta_1^0$, where β_1^0 is some hypothesized value for β_1 , the test statistic is

$$t = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{\frac{s_{y|x}}{s_x \sqrt{n-1}}}$$

and

$$t \sim t(n-2)$$

or, setting up confidence intervals for β_1

$$\hat{\beta}_1 - t_{1-\alpha/2}(n-2) \left[\frac{s_{y|x}}{s_x \sqrt{n-1}} \right] \leq \beta_1 \leq \hat{\beta}_1 + t_{1-\alpha/2}(n-2) \left[\frac{s_{y|x}}{s_x \sqrt{n-1}} \right]$$

e.g.

In the current example, suppose we wish to test

$$H_0 : \beta_1 = 0$$

$$\text{vs. } H_a : \beta_1 \neq 0$$

then,

$$t = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{s_{y|x} \sqrt{n-1}} = \frac{0.97 - 0}{17.31 / (15.29) \sqrt{29}} = 4.62$$

and we reject H_0 if $t > t_{.975}(28) = 2.0484$

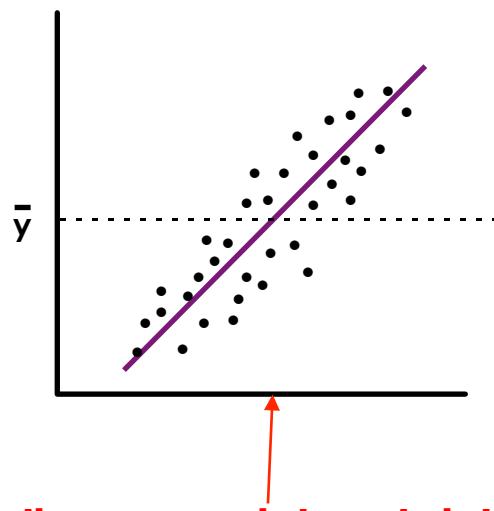
or if $t < t_{.025}(28) = -2.0484$

∴ reject H_0 at $\alpha=.05$

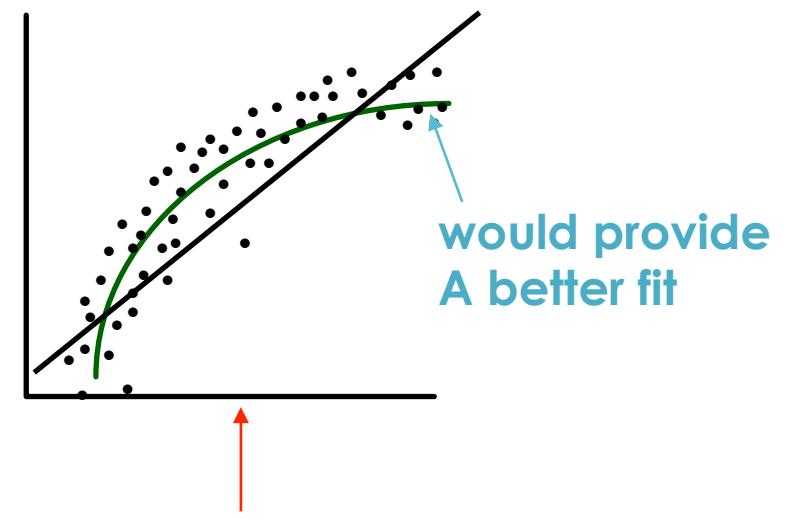
(In fact, $p < .001$)

This means that x provides significant information for the prediction of y . That is, $\hat{y} = \bar{y} + \hat{\beta}_1(x - \bar{x})$ is far better than the naive model for predicting y .

A better model might exist (e.g., one with a curvilinear term), but there is a definite linear component.



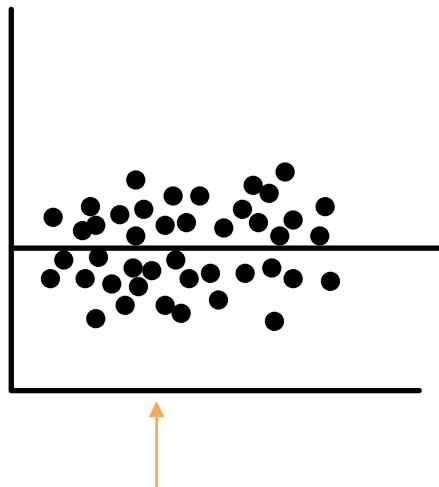
linear model certainly
fits better than $\hat{y} = \bar{y}$



would provide
A better fit

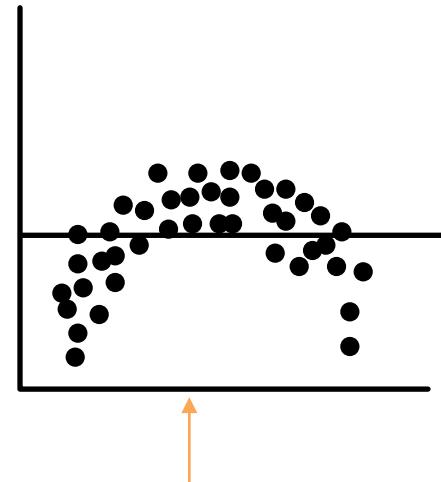
The straight line model may very well represent only a linear approximation to a truly nonlinear relationship

note: if $H_0 : \beta_1 = 0$ is not rejected it means either



x provides little or no help in predicting y

or



The true relationship between x and y is not linear.

* **Important point: whether or not $H_0 : \beta_1 = 0$ is rejected, the straight-line model may not be appropriate. Some other function may better describe the relationship between x and y.**

Now Consider β_0

In order to test $H_0 : \beta_0 = \beta_0^{(0)}$, the test statistic used is

$$t = \frac{\hat{\beta}_0 - \beta_0^{(0)}}{s_{y|x} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}}$$

and

$$t \sim t(n-2)$$

and confidence intervals may be constructed as

$$\hat{\beta}_0 - t_{1-\alpha/2}(n-2) \left[s_{y|x} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}} \right] \leq \beta_0 \leq \hat{\beta}_0 + t_{1-\alpha/2}(n-2) \left[s_{y|x} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}} \right]$$

e.g.,

Continuing this example, to test

$$H_0 : \beta_0 = 75$$

$$\text{vs. } H_\alpha : \beta_0 \neq 75$$

$$t = \frac{\hat{\beta}_0 - \beta_0^{(0)}}{s_{y|x} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}} = \frac{98.71 - 75}{17.31 \sqrt{\frac{1}{30} + \frac{(45.13)^2}{(29)(15.29)}}} = 2.37$$

and again reject H_0 at the $\alpha=.05$ level

here $.02 < p < .05$

and

$$98.71 - 2.0484(17.31) \sqrt{\frac{1}{30} + \frac{(45.13)^2}{29(15.29)^2}} \leq \beta_0 \leq 98.71 + 2.0484(17.31) \sqrt{\frac{1}{30} + \frac{(45.13)^2}{29(15.29)^2}}$$

$$78.23 \leq \beta_0 \leq 119.20$$

Now, if you give me a value of x , I'll give you a confidence interval for $\mu_{y|x}$.

It can be demonstrated that

$$\sigma_{\hat{y}_{x_0}}^2 = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2} \right)$$

and this is estimated by

$$s_{\hat{y}_{x_0}}^2 = s_{y|x}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2} \right)$$

and a $100(1 - \alpha)\%$ confidence interval estimate for $\mu_{y|x}$ is

$$\hat{y}_{x_0} - t_{1-\alpha/2}(n-2)s_{\hat{y}_{x_0}} \leq \mu_{y|x} \leq \hat{y}_{x_0} + t_{1-\alpha/2}(n-2)s_{\hat{y}_{x_0}}$$

Example:

Suppose we want a 90% confidence interval for the mean SBP of 65 year old individuals

$$\begin{aligned}\hat{y}_{x_0} &= \hat{y}_{65} = 142.53 + (0.97)(65 - 45.13) \\ &= 161.80\end{aligned}$$

$$s_{\hat{y}_{65}} = 17.31 \left(\frac{1}{30} + \frac{(65 - 45.13)^2}{(29)(15.29)^2} \right)^{1/2} = 5.24$$

$$161.80 - 1.7011(5.24) \leq \mu_{y|65} \leq 161.80 + 1.7011(5.24)$$

$$152.89 \leq \mu_{y|65} \leq 170.71$$

Suppose we now wish to estimate the response y of a single individual based on the fitted regression function.

It can be demonstrated that the “prediction interval**” (PI) is given by**

$$\hat{Y}_{x_0} - t_{1-\alpha/2} (n-2)s_{y|x} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}} \leq Y_{x_0} \leq \hat{Y}_{x_0} + t_{1-\alpha/2} (n-2)s_{y|x} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}$$

This is not a parameter.
Hence, we use the expression
"PI" rather than "CI".

Note that this is the only difference from the previous expression.

Example

Suppose we want a 90% prediction interval for SBP for an individual whose age is 65

again, $\hat{Y}_{65} = 161.80$

$$s_{y|x} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)s_x^2}} = 17.31 \sqrt{1 + \frac{1}{30} + \frac{(65 - 45.13)^2}{(29)(15.29)^2}} = 18.09$$

note how much larger this is than before (5.24)

$$161.80 - (1.7011)(18.09) \leq Y_{65} \leq 161.80 + (1.7011)(18.09)$$

$$131.03 \leq Y_{65} \leq 192.57$$

prediction interval is much wider than confidence interval was

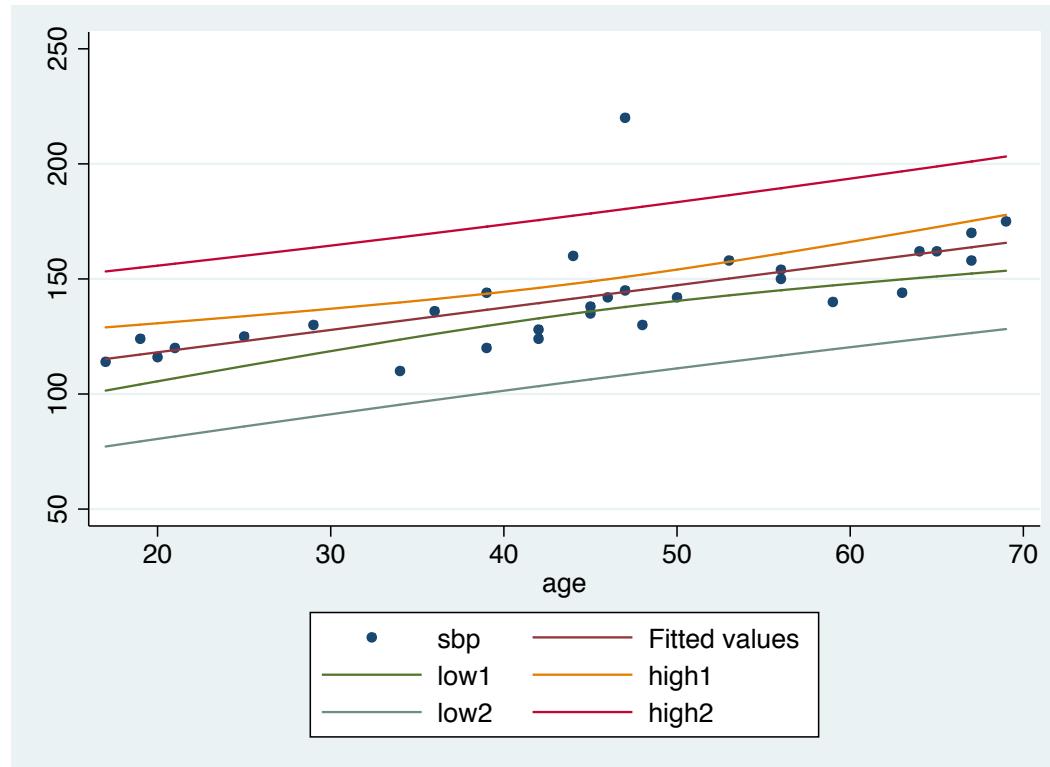
Note that whether we are constructing confidence intervals or prediction intervals, the expressions contain the term

$$\frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}$$

This means that the farther x_0 is from \bar{x} , the larger will be the variance and the wider will be the interval.

Diagrammatically:

```
. predict seyhat, stdp  
. display invtail(28,0.025)  
2.0484071  
. generate low1= yhat-2.0484071* seyhat  
. generate high1= yhat+2.0484071* seyhat  
  
. predict sepred, stdf  
. generate low2= yhat-invtail(28,0.025)* sepred  
. generate high2= yhat+invtail(28,0.025)* sepred  
. scatter sbp yhat low1 high1 low2 high2 age,sort connect(. l l l l l)  
symbol(o i i i i)
```



Hence, we can make more precise estimates for $\mu_{y|x}$ or Y_{x_0} when we are close to \bar{x} . As we move away from \bar{x} our confidence intervals and prediction intervals increase in width.

Applied Regression Analysis

Week 3

1. Homework week 2: highlights
2. Correlation coefficient I
3. Correlation coefficient II
4. Coefficient of determination, r^2
5. The ANOVA table for straight line regression I
6. The ANOVA table for straight line regression II
7. The ANOVA table for straight line regression III
8. Homework

Stanley Lemeshow, Professor of Biostatistics
College of Public Health, The Ohio State University



THE OHIO STATE UNIVERSITY

DEF: The correlation coefficient provides a measure of how two random variables are associated in a sample.

- It is also a measure of the strength of the straight-line relationship between x and y .

$$\begin{aligned}
 r &= \frac{\widehat{\text{cov}}(x, y)}{\sqrt{\widehat{\text{var}}(x) \widehat{\text{var}}(y)}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \\
 &= \frac{\sum_{i=1}^n x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\left[\sum_{i=1}^n x_i^2 - \frac{(\sum x_i)^2}{n} \right] \left[\sum_{i=1}^n y_i^2 - \frac{(\sum y_i)^2}{n} \right]}}^{1/2}
 \end{aligned}$$

\$S_x S_y\$ →

note: since $\hat{\beta}_1 = \frac{\widehat{\text{cov}}(x, y)}{\widehat{\text{var}}(x)} = \frac{s_{xy}}{s_x^2}$

$$\text{and } r = \frac{s_{xy}}{s_x s_y}$$

we have that $r = \frac{s_x}{s_y} \hat{\beta}_1$

Example - For the AGE, SBP data,

$$r = \frac{199,576 - \frac{(1354)(4276)}{30}}{\left\{ \left[67894 - \frac{1354^2}{30} \right] \left[624260 - \frac{4276^2}{30} \right] \right\}^{1/2}} = 0.66$$

or, more simply, since $\hat{\beta}_1 = 0.97$, $r = \frac{15.29}{22.58}(0.97) = 0.66$

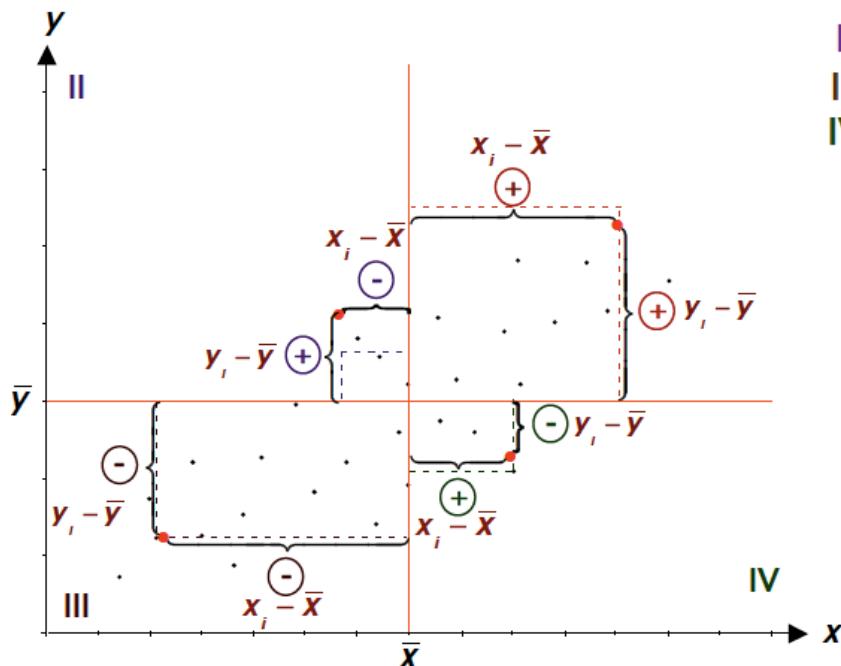
now $-1 \leq r \leq +1$, and r is dimensionless

- i.e., it is independent of the units of measurement of x or y .

finally, r always has the same sign as $\hat{\beta}_1$

Actually, r is the standardized covariance.

Let us motivate what is meant by the covariance between x and y .

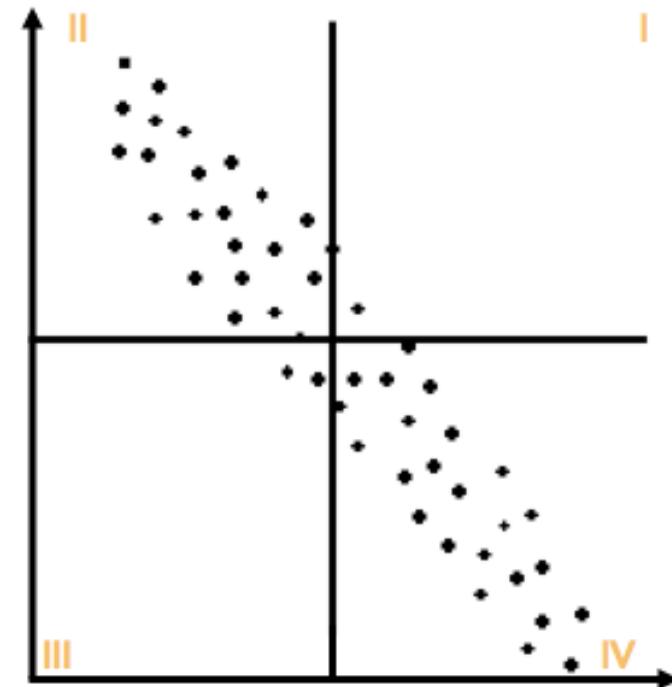
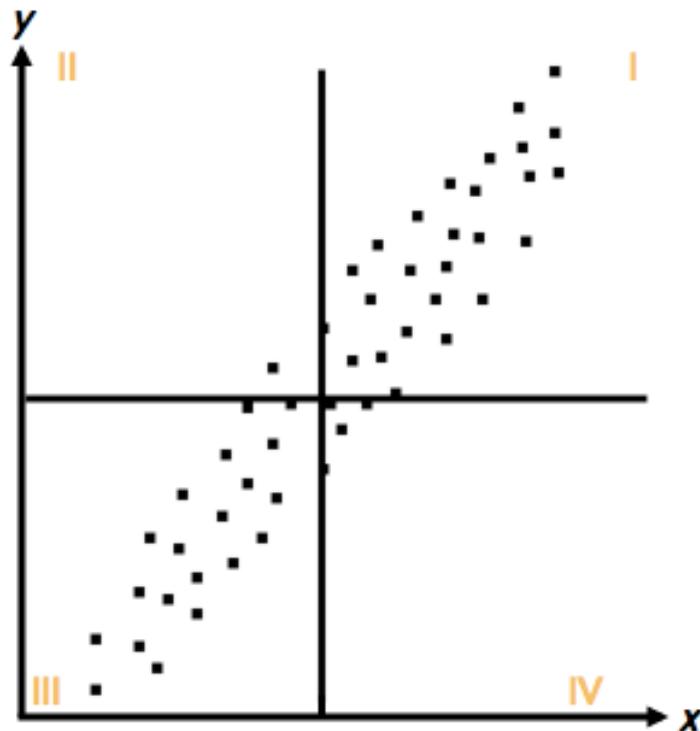


Quadrant	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
I	+	+	+
II	-	+	-
III	-	-	+
IV	+	-	-

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

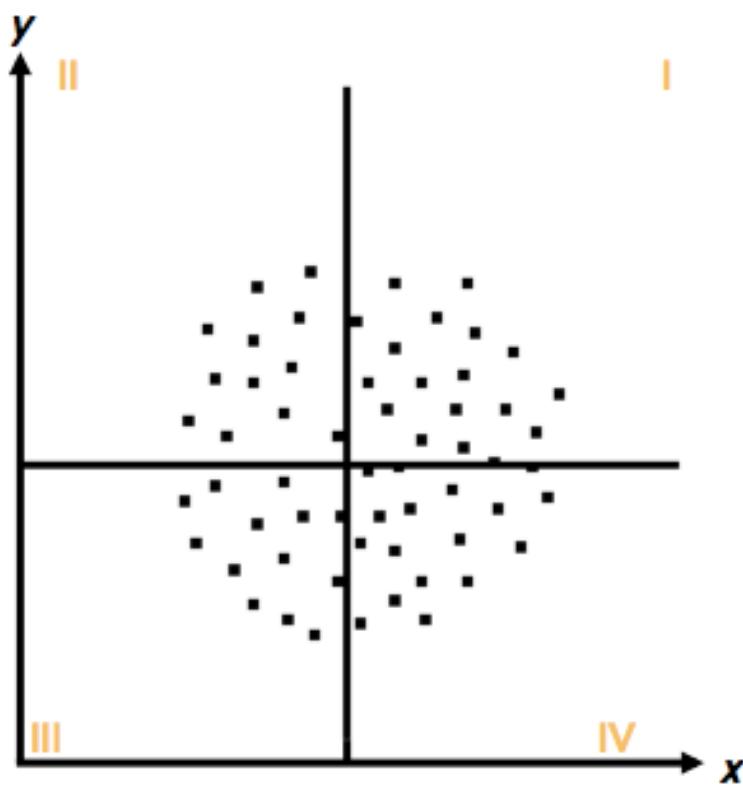
sample covariance: $s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Now, if points look like:



$s_{xy} > 0$ since most points are in QI and QIII
 $\therefore r > 0, \hat{\beta}_1 > 0$

$s_{xy} < 0$ since most points are in QII and QIV
 $\therefore r < 0, \hat{\beta}_1 < 0$

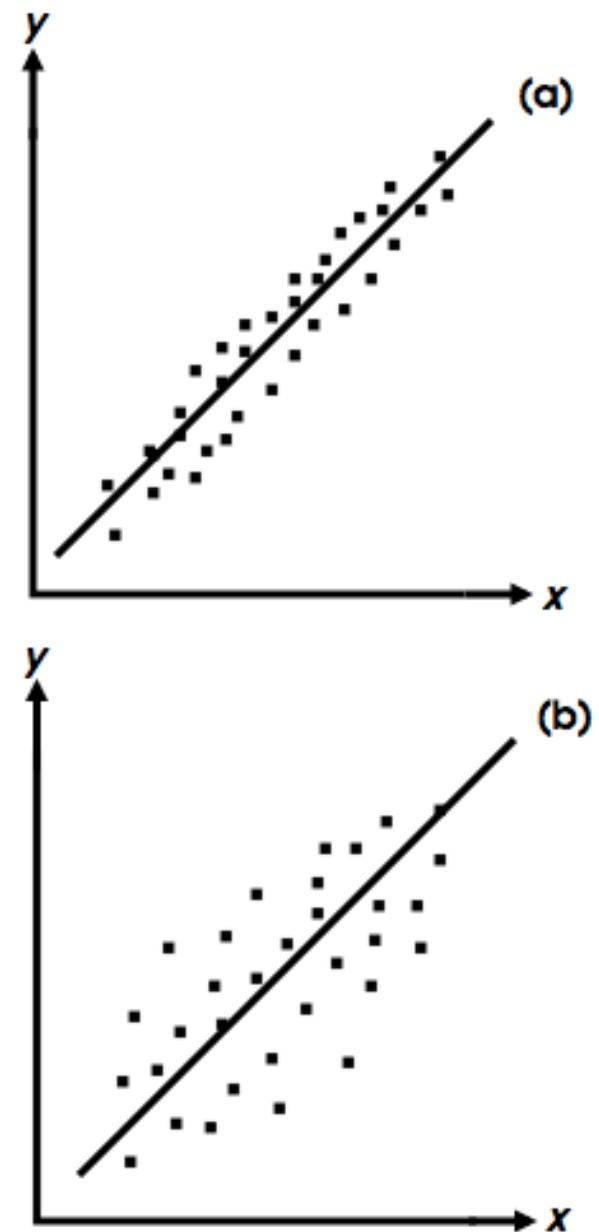


$S_{xy} = 0$ since + points are offset by - points

$$\therefore r = 0, \hat{\beta}_1 = 0$$

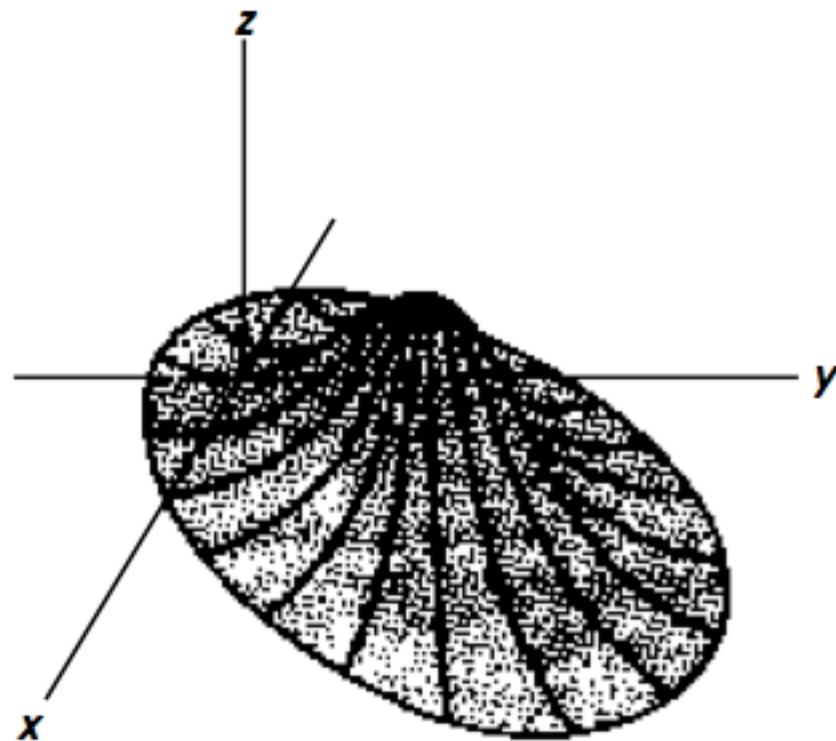
r in (a) is much greater than r in (b) since there are fewer points in QII and IV in (a)

This is true even if though the slopes are identical

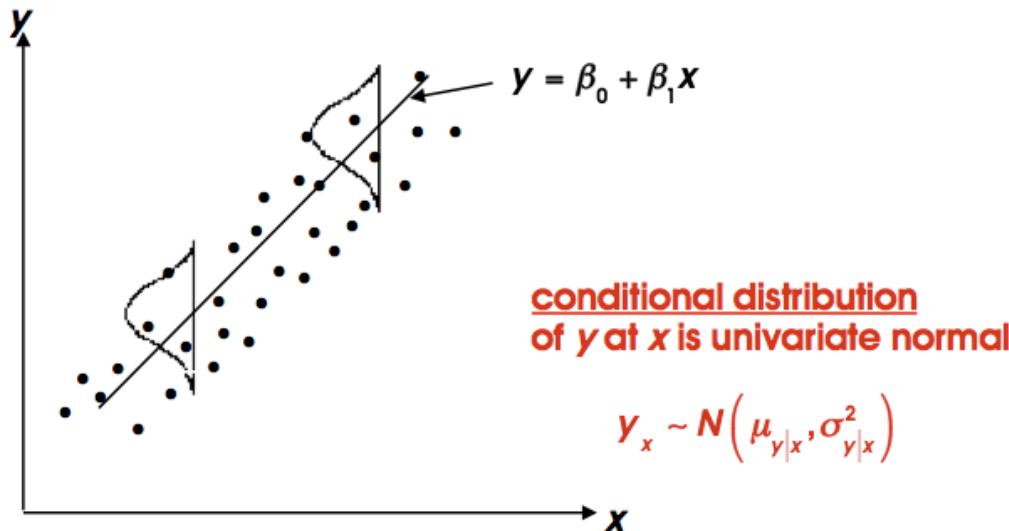


Let us assume that x and y are random variables having the bivariate normal distribution.

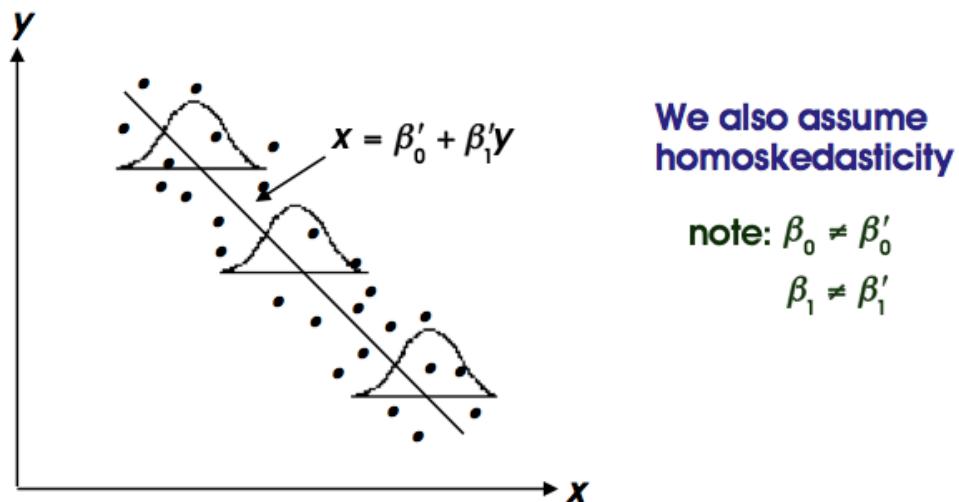
- This distribution has the following appearance:



One property of this distribution is that the distribution of y for any fixed x is normal



and the distribution of x for any fixed y is normal



It follows from statistical theory that

$$\mu_{y|x} = \mu_y + \rho_{xy} \frac{\sigma_y}{\sigma_x} (x - \mu_x)$$

and

$$\sigma_{y|x}^2 = \sigma_y^2 (1 - \rho_{xy}^2)$$

Let $\beta_1 = \rho_{xy} \left(\frac{\sigma_y}{\sigma_x} \right)$ **and** $\beta_0 = \mu_y - \beta_1 \mu_x$

Then $\mu_{y|x} = \underbrace{\beta_0 + \beta_1 \mu_x}_{\mu_y} + \underbrace{\beta_1}_{\rho_{xy} \left(\frac{\sigma_y}{\sigma_x} \right)} (x - \mu_x) = \beta_0 + \beta_1 x$

hence we have the familiar straight-line model

Also, $\mu_{y|x}$ can be estimated by

$$\hat{\mu}_{y|x} = \bar{y} + r \left(\frac{s_y}{s_x} \right) (x - \bar{x})$$

and, since $\hat{\beta}_1 = r \left(\frac{s_y}{s_x} \right) = \frac{s_{xy}}{s_x s_y} \left(\frac{s_y}{s_x} \right) = \frac{s_{xy}}{s_x^2}$

$$\hat{\mu}_{y|x} = \bar{y} + \hat{\beta}_1 (x - \bar{x}) \leftarrow \text{usual least squares line}$$

Thus, the least squares formulae for $\hat{\beta}_0$ and $\hat{\beta}_1$ can be developed by assuming that x and y are random variables having the bivariate normal distribution and by substituting the usual estimates of μ_x , μ_y , σ_x , σ_y , and ρ_{xy} into the expression for $\mu_{y|x}$.

$$\text{Also, } \hat{\sigma}_{y|x}^2 = s_{y|x}^2 = \frac{n-1}{n-2} \left(s_y^2 - \hat{\beta}_1^2 s_x^2 \right) = \frac{n-1}{n-2} s_y^2 (1 - r^2)$$

$$\approx s_y^2 (1 - r^2)$$

Now, as we've seen,

$$\sigma_{y|x}^2 = \sigma_y^2 (1 - \rho_{xy}^2)$$

where σ_y^2 = unconditional variance of y

i.e., it's the variance of y when we know nothing about x

On the other hand,

$$\sigma_{y|x}^2 = \text{conditional variance of } y$$

i.e., it's the variance of y when we know the corresponding value of x

Hence, the reduction in the variance of y due to knowledge of x is:

$$\sigma_y^2 - \sigma_{y|x}^2 = \rho_{xy}^2 \sigma_y^2$$

and

$$\rho_{xy}^2 = \frac{\sigma_y^2 - \sigma_{y|x}^2}{\sigma_y^2}$$

Hence, the squared correlation coefficient is the proportion of the variance of y "explained by" knowledge of x .

When $\rho_{xy} = 0$ this means that $\sigma_{y|x}^2 = \sigma_y^2$

i.e., none of the variance in y is explained
by the regression of y on x .

When $\rho_{xy} = 1$ this means that $\sigma_{y|x}^2 = 0$

i.e., all of the variance in y is explained
by the regression of y on x .

i.e., the
relationship
between y and
 x is perfectly
linear.

Hence, defining $SSY = \sum_{i=1}^n (y_i - \bar{y})^2$

and

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

if the model fits, $SSE \ll SSY$. Then

$$r^2 = \frac{SSY - SSE}{SSY}$$

is a quantitative measure of the improvement in the fit obtained by using x and

$$0 \leq r^2 \leq 1$$

Note:

One should not be led to a false sense of security by considering the magnitude of r rather than of r^2 when assessing the strength of the linear association between x and y .

e.g. when $r = .5$, $r^2 = .25$

$r = .7$, $r^2 = .49$

$r = .3$, $r^2 = .09$

Example

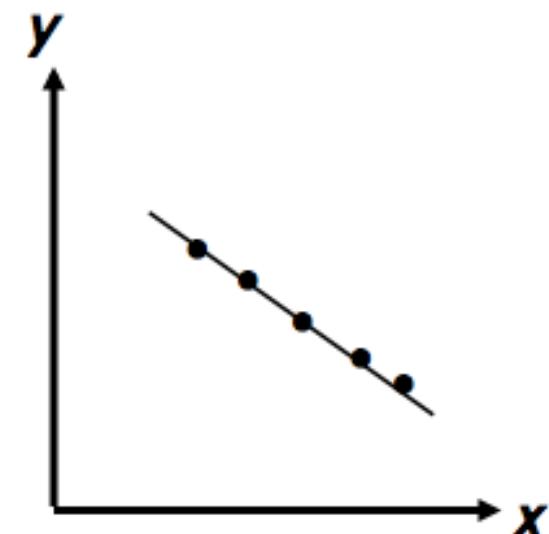
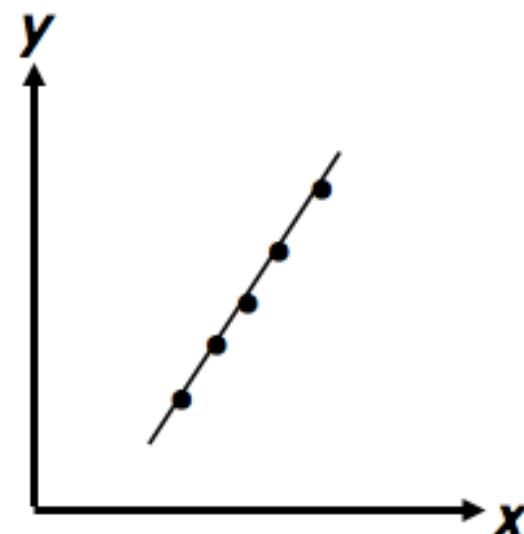
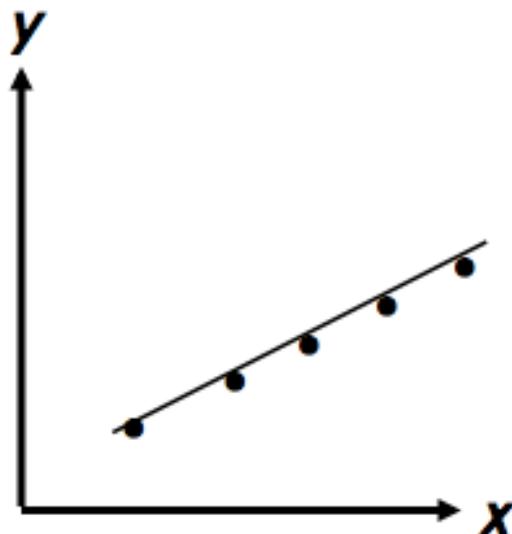
For the age-SBP data

$$r = .66$$

$$r^2 = .44$$

Also note:

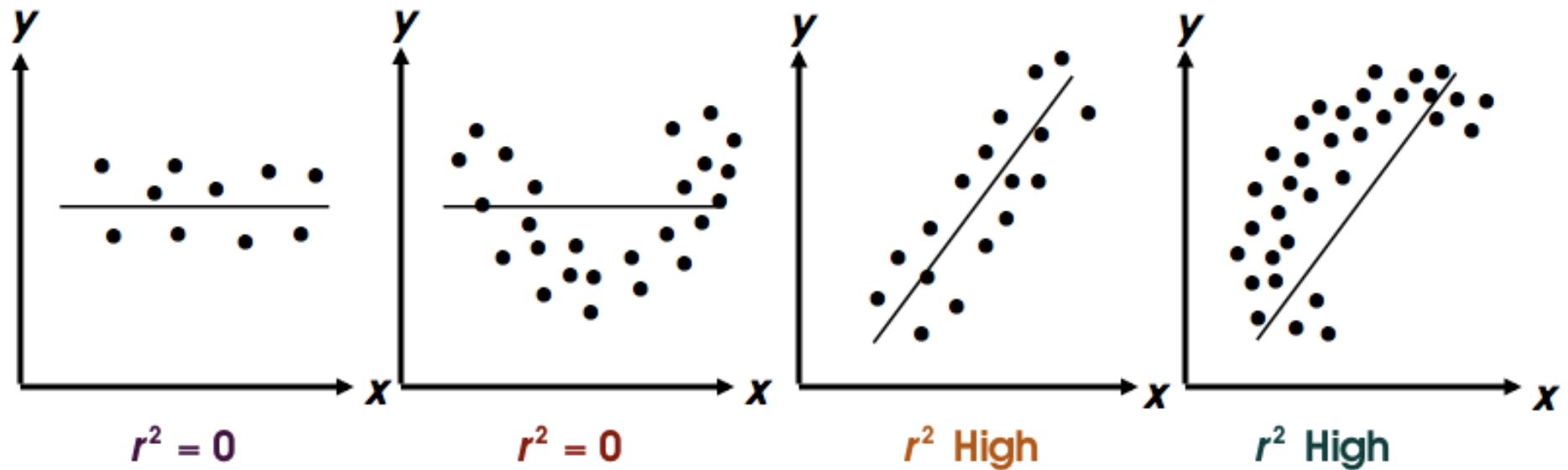
- r^2 is not a measure of the magnitude of the slope of the regression line



In all instances $r^2 = 1$ but the slopes differ widely.

- r^2 is a measure of the clustering of points about the regression line.

r^2 is not a measure of the appropriateness of the straight-line model



$r^2 = 0$
no association
between x and y

$r^2 = 0$
strong nonlinear
association

r^2 High
straight-line model
is appropriate

r^2 High
nonlinear model
may be better

To test $H_0 : \rho_{xy} = 0$
vs. $H_a : \rho_{xy} \neq 0$ (or one-sided)

we may simply test

$H_0 : \beta_1 = 0$
vs. $H_a : \beta_1 \neq 0$ as we learned before

or, we may use

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{and } t \sim t(n-2)$$

Note:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \equiv \frac{\hat{\beta}_1 s_x}{s_{y|x} / \sqrt{n-1}}$$

that we learned before

e.g., in the age-SBP problem,

$$r = .66$$

and

$$t = \frac{.66\sqrt{30 - 2}}{\sqrt{1 - (.66)^2}} = 4.62$$

which is the same value as that obtained in the test for the slope

To test

$$H_0 : \rho_{xy} = \rho_0$$

$$\text{vs. } H_a : \rho_{xy} \neq \rho_0 \quad \text{where } \rho_0 \neq 0$$

we cannot use t as described previously.

Also, $H_0 : \rho_{xy} = \rho_0 (\neq 0)$ is not equivalent to $H_0 : \beta_1 = \beta_1^{(0)}$.

We must consider the distribution of r

- $r \sim$ symmetric only when $\rho_{xy} = 0$
- r is not symmetric when $\rho_{xy} \neq 0$
 - In that case the sampling distribution of r is skewed.

Hence, we cannot use t (that has a normally distributed estimator in the numerator and an estimate of its standard deviation in the denominator).

Fortunately, through an appropriate transformation, r can be changed to a statistic that is approximately normal.

FISHER' S Z TRANSFORMATION

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

note: $z = \text{inv hyp tan } (r)$

has been shown to be approximately normal with

$$E(z) = \frac{1}{2} \ln \left(\frac{1+\rho_{xy}}{1-\rho_{xy}} \right) = z \quad \text{and} \quad \text{var}(z) = \frac{1}{n-3}$$

i.e.,

$$z \sim N \left(z, \frac{1}{n-3} \right)$$

The inverse Fisher transformation is

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

note: $r = \text{hyp tan } (z)$

Example: for the AGE-SBP data

Test: $H_0: \rho_{xy} = .85$

vs: $H_a: \rho_{xy} \neq .85$

$r = 0.66, n = 30$

$$z = \frac{1}{2} \ln \left(\frac{1 + r}{1 - r} \right) = \frac{1}{2} \ln \left(\frac{1 + .66}{1 - .66} \right) = .7928$$

$$Z = \frac{1}{2} \ln \left(\frac{1 + \rho}{1 - \rho} \right) = \frac{1}{2} \ln \left(\frac{1 + .85}{1 - .85} \right) = 1.2561$$

Then

$$z' = \frac{z - Z}{\sqrt{\frac{1}{n-3}}} = \frac{.7928 - 1.2561}{\sqrt{\frac{1}{27}}} = -2.41$$

and reject H_0 if $z' > 1.96$

or if $z' < -1.96 \quad \therefore \text{reject } H_0.$

Similarly, we can get a $(1 - \alpha)\%$ C.I. estimate for ρ_{xy}

$$z - 1.96\sqrt{\frac{1}{n-3}} \leq Z \leq z + 1.96\sqrt{\frac{1}{n-3}}$$

$$.7928 - 1.96\sqrt{\frac{1}{27}} \leq Z \leq .7928 + 1.96\sqrt{\frac{1}{27}}$$

$$.4156 \leq Z \leq 1.1700$$

Using the inverse transformation

$$r_L = \frac{e^{2(0.4156)} - 1}{e^{2(0.4156)} + 1} = \frac{1.296}{3.296} = .393$$

$$r_U = \frac{e^{2(1.1700)} - 1}{e^{2(1.1700)} + 1} = \frac{9.38}{11.38} = .824$$

$$.393 \leq \rho_{xy} \leq .824$$

```
. z_r sbp age          (STB-32: sg51)
(sample correlations, n=30)
      sbp      age
sbp  1.0000
age  0.6576  1.0000

(lower\upper 95% confidence limits)
      sbp      age
sbp  1.0000  0.8229
age  0.3896  1.0000
```

Let

$$(y_i - \bar{y}) = \text{"total" deviation} = a$$

$$(y_i - \hat{y}_i) = \text{"unexplained" deviation} = b$$

$$(\hat{y}_i - \bar{y}) = \text{"explained" deviation} = c$$

Note:

$$(y_i - \bar{y}) = (y_i - \hat{y}_i + \hat{y}_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

$$a = b + c$$

now, squaring both sides of the equation

$$(y_i - \bar{y})^2 = (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

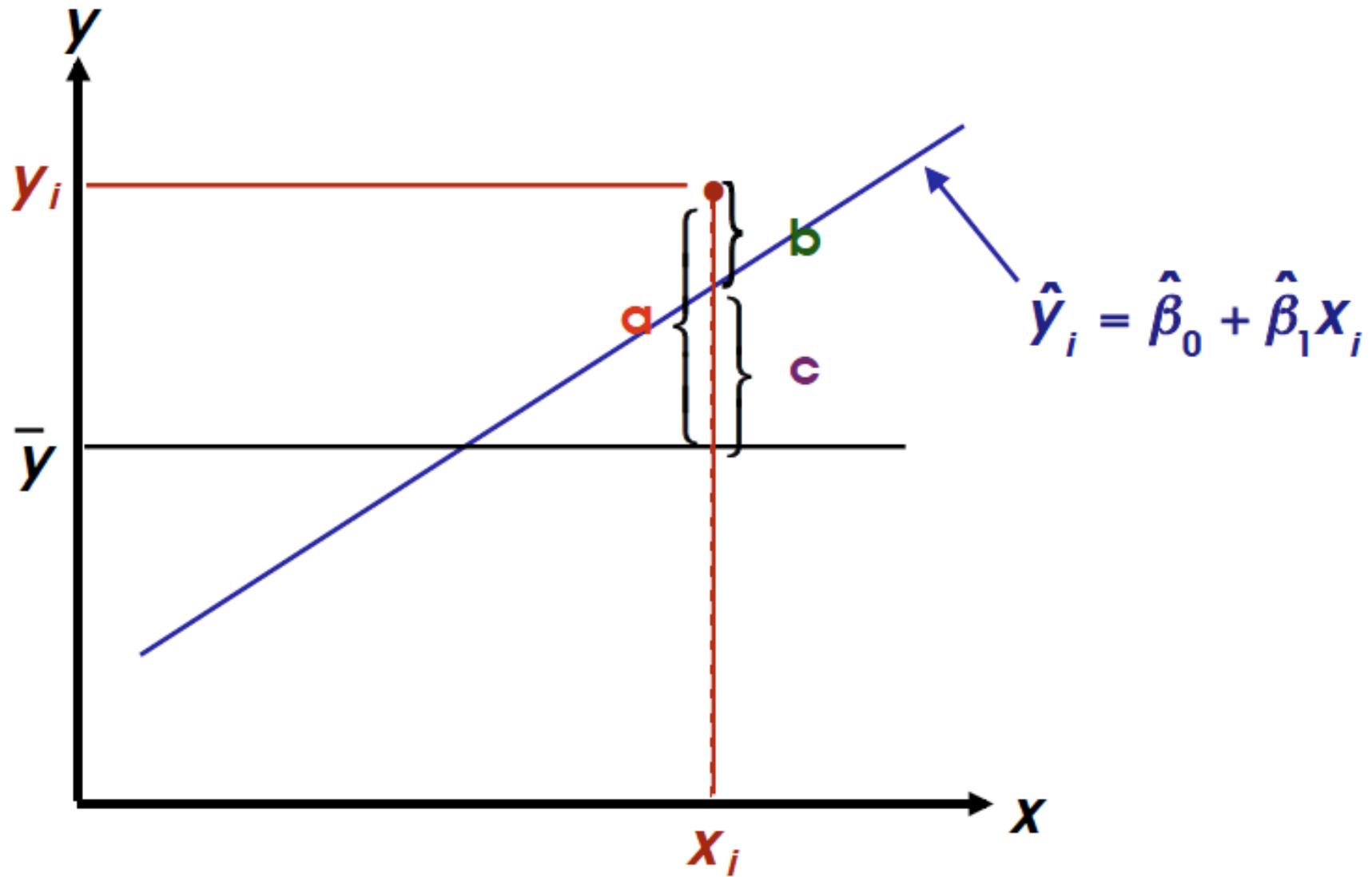
and summing over all n points

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 0$$

\sum cross products = 0

This is called the:
“fundamental equation of regression analysis”

graphically,



Recall:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = SSY \quad \sum_{i=1}^n (y_i - \hat{y}_i)^2 = SSE \quad \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = SSY - SSE$$

Now, for the age-SBP data this can all be summarized in an ANOVA table as follows:

Source	df	SS	MS	EMS	F
Regression (x)	1	$SSY - SSE = 6394.02$	6394.02	$\sigma^2 + c\beta_1^2$	21.33
Residual	28	$SSE = 8393.44$	299.77	σ^2	
Total	29	$SSY = 14787.46$			

Note: $r^2 = \frac{SSY - SSE}{SSY} = \frac{6394.02}{14787.46} = 0.43$

$$\frac{SSE}{df} = MSE = s^2_{y|x} \quad \text{we computed earlier}$$

$s^2_{y|x}$ is an estimate of σ^2 if the true regression model is linear.

$SSY - SSE$ provides an estimate of σ^2 only if the variable x does not help to predict the dependent variable, y (i.e., $\beta_1 = 0$)

MS_1 and MS_2 are independent and, if $H_0 : \beta_1 = 0$ is true, then

$$F = \frac{MS_1}{MS_2} \sim F(1, n-2)$$

This is also a test of $H_0 : \rho_{xy} = 0$

Fortunately this F test is equivalent to the 2-sided t -test discussed previously.

recall:

$$F(1, \nu) = t^2(\nu)$$

so

$$F_{.95}(1, \nu) = (t_{.975}(\nu))^2$$

e.g.,

in the age-SBP example

$$t = 4.62 \quad t^2 = 21.33 = F$$

also, $t_{.975}(28) = 2.05$ and $(2.05)^2 = 4.20 = F_{.95}(1, 28)$

Hence, the two sided critical region:

reject H_0 if $t > t_{.975}(\nu)$
or if $t < t_{.025}(\nu)$

is equivalent to

reject H_0 if $F > F_{.95}(1, \nu)$

Spreadsheets For Computing ANOVA Table In Regression

ID	SBP (y)	AGE (x)	$x - \bar{x}$	$(x - \bar{x})^2$	$y - \bar{y}$	$(y - \bar{y})^2$	$(x - \bar{x}) \times (y - \bar{y})$	\hat{y}	$y - \hat{y}$	$(y - \hat{y})^2$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$	$(y - \hat{y}) \times (\hat{y} - \bar{y})$
1	144	39	-6.13	37.62	1.47	2.15	-9.00	136.58	7.42	55.08	-5.95	35.46	-44.19
2	220	47	1.87	3.48	77.47	6001.08	144.60	144.35	75.65	5723.58	1.81	3.28	137.11
3	138	45	-0.13	0.02	-4.53	20.55	0.60	142.40	-4.40	19.39	-0.13	0.02	0.57
4	145	47	1.87	3.48	2.47	6.08	4.60	144.35	0.65	0.43	1.81	3.28	1.19
5	162	65	19.87	394.68	19.47	378.95	386.74	161.82	0.18	0.03	19.29	372.03	3.45
:	:	:	:	:	:	:	:	:	:	:	:	:	:
25	160	44	-1.13	1.28	17.47	305.08	-19.80	141.43	18.57	344.73	-1.10	1.21	-20.43
26	158	53	7.87	61.88	15.47	239.22	121.67	150.17	7.83	61.30	7.64	58.33	59.80
27	144	63	17.87	319.22	1.47	2.15	26.20	159.88	-15.88	252.16	17.35	300.89	-275.45
28	130	29	-16.13	260.28	-12.53	157.08	202.20	126.87	3.13	9.80	-15.66	245.34	-49.03
29	125	25	-20.13	405.35	-17.53	307.42	353.00	122.99	2.01	4.05	-19.55	382.08	-39.36
30	175	69	23.87	569.62	32.47	1054.08	774.87	165.70	9.30	86.40	23.17	536.92	215.38
SUM	4276	1354	0.00	6783.47	0.00	14787.47	6585.87		0.00	8393.44	0.00	6394.02	0.00
MEAN	142.53	45.13											
VAR	509.91	233.91											

N= 30
 SLOPE= 0.97
 INTERCEPT= 98.71
 CORR= 0.66

SOURCE	DF	SS	MS	F	p
REGRESSION	1	6394.02	6394.02	21.33	0.00
RESIDUAL	28	8393.44	299.77		
TOTAL	29	14787.47			

Applied Regression Analysis

Week 4

1. Homework week 3: highlights
2. Polynomial regression I
3. Polynomial regression II
4. Polynomial regression III
5. Example: Dose-response study / assessing multicollinearity
6. Example: Potential energy
7. Homework

Stanley Lemeshow, Professor of Biostatistics
College of Public Health, The Ohio State University



THE OHIO STATE UNIVERSITY

WEEK 4: POLYNOMIAL REGRESSION

A polynomial of order k in x is an expression of the form

$$y = c_0 + c_1x + c_2x^2 + c_3x^3 + \cdots + c_kx^k$$

where the c 's and k are constants.

When $k = 1$ we had

$$y = c_0 + c_1x \quad \text{straight line}$$

Let us now focus on the 2nd order polynomial ($k = 2$)

$$y = c_0 + c_1x + c_2x^2$$

These are mathematical models.

The statistical model for the $k = 2$ case can be expressed in one of two ways:

$$\mu_{y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$$

mean of y at
a given x

or

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

unknown parameters
(regression coefficients)

Error component

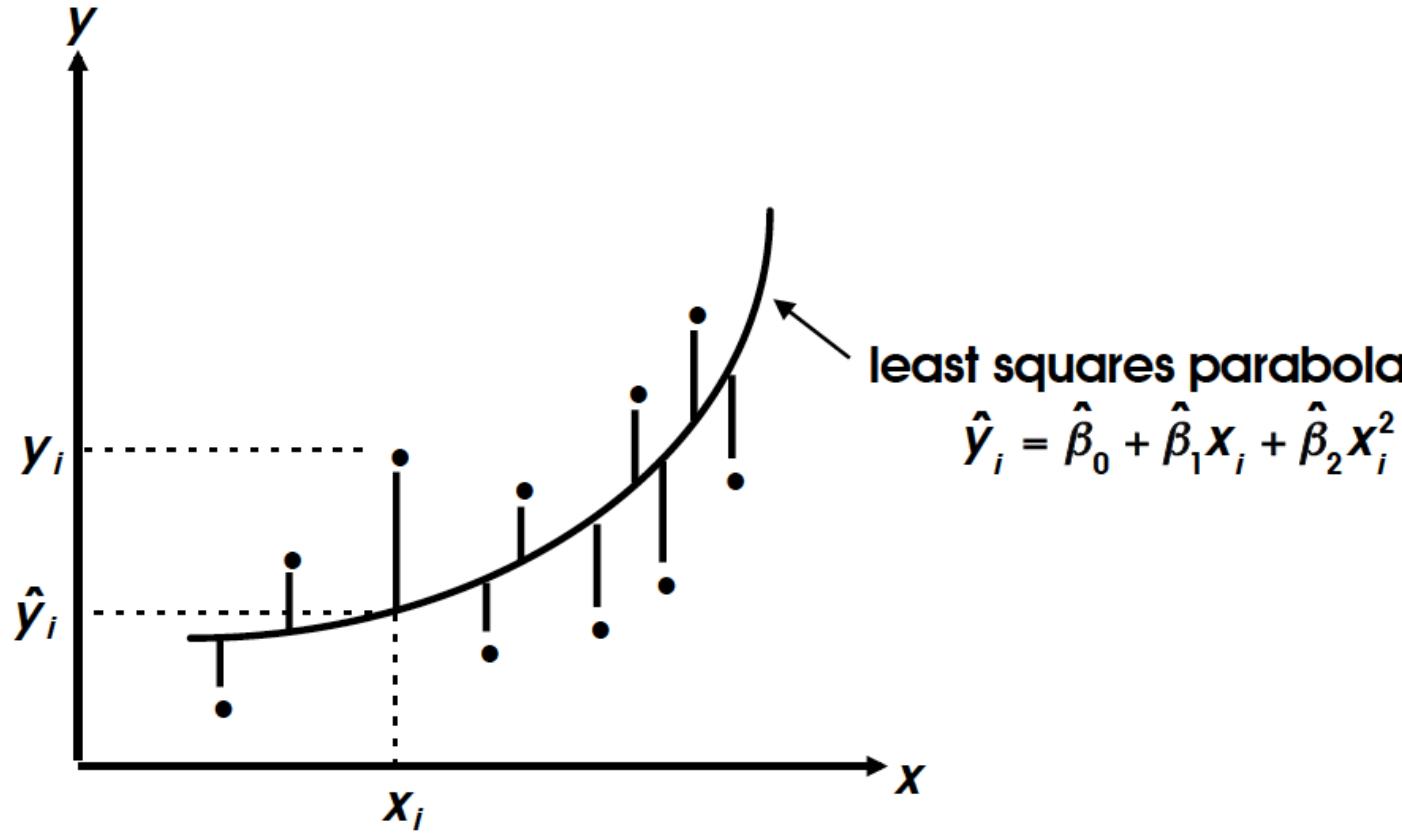
Let us now use the method of least squares to obtain estimates for the regression coefficients in the parabolic (2nd degree) model.

The estimated parabola may be written as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

and

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2)^2$$



$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ are chosen so that the SSE is smaller than for any other choice of β 's.

Instead of presenting here the precise formulas for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ (which can get very complex-particularly when k is large) it is assumed that you will be doing this work via computer.

For the age-SBP example with the outlier removed ($n = 29$) we obtain from the computer:

$$\hat{\beta}_0 = 113.41$$

$$\hat{\beta}_1 = 0.088$$

$$\hat{\beta}_2 = 0.010$$

Hence, the fitted model is

$$\hat{y} = 113.41 + 0.088x + 0.010x^2$$

Recall that for these $n = 29$ individuals, the straight-line model was

$$\hat{y} = 97.08 + 0.95x$$

Now, the essential results based on fitting a 2nd - (or higher) order polynomial model can be summarized in an ANOVA table.

As was true for the 1st order polynomial model,

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

$$SSY = (SSY - SSE) + SSE$$

Total SS = Due regression SS + residual SS

Then the ANOVA Table is

Source	df	SS	MS	F
Due Regression	$k = 2$	$SSY - SSE = 6273.40$	$\frac{SSY - SSE}{k} = 3136.70$	35.37
Due Residual	$n - k - 1 = 26$	$SSE = 2306.05$	$\frac{SSE}{n - 1 - k} = 88.69$	$(p < .001)$
Total	$r^2 = .731$	$n - 1 = 28$	$SSY = 8579.45$	

Now recall that in the straight-line model with the outlier removed we had

Source	df	SS	MS	F
Due Regression	1	6110.10	6110.10	66.81
Due Residual	27	2469.35	91.46	$(p < .0001)$
Total	$r^2 = .712$	28	8579.45	

These tables give rise to the following for the 2nd order polynomial model:

Source	df	SS	MS	F
Regression	1	6110.10	6110.10	66.81 $= \frac{6110.1}{91.46}$
	1	163.30	163.3	1.84 $= \frac{163.30}{88.69}$
Residual	26	2306.05	88.69	
Total	28	8979.45		

computed by subtraction

note: as usual, the residual sum of squares SSE is divided by its degrees of freedom to yield an estimate of σ^2

i.e.,

$$\text{MS residual} = s_{y|x}^2 = \frac{1}{n-3} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

follows general rule = $n - \# \text{ estimated regression coefficients}$

There are 2 basic inferential questions associated with 2nd order polynomial regression:

- (1) Is the overall regression significant?
- (2) Does the 2nd order model explain significantly more than that achieved by the straight-line model?

(1) Test for overall regression

H_0 : There is no significant overall regression using x and x^2

H_a : There is a significant overall regression

we use $F = \frac{\text{MS regression}}{\text{MS residual}}$

and compare this to the $F(2, n-1-k)$

In our example

$F = 35.37$ and $F_{.999}(2, 26) = 9.12 \therefore \text{reject } H_0 (p < .001)$

This F - test is not equivalent to any t - test

[This is true since $F(1, v) = t^2(v)$ but $F(2, v) \neq t$]

We can compute the multiple R^2

R^2 = “squared multiple correlation coefficient”
= proportionate reduction in the error sum of squares
obtained using x and x^2 instead of the naive predictor \bar{y} .

$$R^2 = \frac{\text{SSY} - \text{SSE}(2^{\text{nd}} \text{ order model})}{\text{SSY}} = \frac{\text{Due Reg. SS}}{\text{Total SS}}$$

In our example $R^2 = 0.731$. The F -test also tests

$$\begin{aligned} H_0 &: R^2 = 0 \\ \text{vs } H_a &: R^2 > 0. \end{aligned}$$

As was true for the straight-line model, this one is significant.

(2) Test for the Addition of x^2 Into the Model

H_0 : The addition of the x^2 term to the model does not significantly improve the prediction of y over and above that achieved by the straight-line model.

H_a : it does add to the prediction of y

note:

$r^2 = .712$ for the straight-line model

$R^2 = .731$ for the second order model

*more variation will always be explained by adding extra terms to the model.

The question here is whether the increase

$$= (.731 - .712) = .019$$

represents a significant increase in the variation explained by the additional term.

(i.e., is .019 enough of an increase to warrant adding the x^2 term to the model).

To answer this we compute the extra sum of squares due to the addition of x^2 . This appeared in the ANOVA table under the source heading "Regression $x^2 | x$ ".

$$\text{Extra SS due to adding } x^2 = \text{SS regression}_{(2^{\text{nd}} \text{ order model})} - \text{SS regression}_{(1^{\text{st}} \text{ order model})}$$

In our example,

$$\text{SS regression (straight-line model)} = 6110.10$$

$$\text{SS regression (2nd order model)} = 6273.40$$

$$\text{Extra SS due to adding } x^2 \text{ term} = 6273.40 - 6110.10 = 163.30$$

To test H_0 , we use

$$F = \frac{\text{(Extra SS due to adding } x^2\text{)}/1}{\text{MS residual for 2nd order model}}$$

and this F is compared to the $F(1, n - 1 - k)$

In our example

$$F = \frac{163.30}{88.69} = 1.84$$

and $F_{.90}(1,26) = 2.91$

in fact $.10 < p < .25$

Another way to perform this test is to compute

$$t = \frac{\hat{\beta}_2}{\widehat{\text{SE}}(\hat{\beta}_2)}$$

obtain from computer output

and compare this to a $t(n-1-k)$

```

. use ":Macintosh HD:Desktop Folder:notes1.dta"

. drop if sbp==220
(1 observation deleted)

. regress sbp age

```

Source	SS	df	MS	Number of obs	=	29
Model	6110.10173	1	6110.10173	F(1, 27)	=	66.81
Residual	2469.34654	27	91.4572794	Prob > F	=	0.0000
Total	8579.44828	28	306.408867	R-squared	=	0.7122

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.9493225	.1161445	8.174	0.000	.7110137 1.187631
_cons	97.07708	5.527552	17.562	0.000	85.73549 108.4187

```
. vif
```

Variable	VIF	1/VIF
age	1.00	1.000000
Mean VIF	1.00	

. gen agesq=age*age

. regress sbp age agesq

Source	SS	df	MS	Number of obs	=	29
Model	6273.40168	2	3136.70084	F(2, 26)	=	35.37
Residual	2306.0466	26	88.6940999	Prob > F	=	0.0000
Total	8579.44828	28	306.408867	R-squared	=	0.7312

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.0875433	.6453289	0.136	0.893	-1.238949 1.414036
agesq	.0099368	.0073232	1.357	0.186	-.0051163 .0249899
_cons	113.4097	13.21041	8.585	0.000	86.25533 140.5641

. vif

Variable	VIF	1/VIF
age	31.83	0.031413
agesq	31.83	0.031413
Mean VIF	31.83	

Another Example

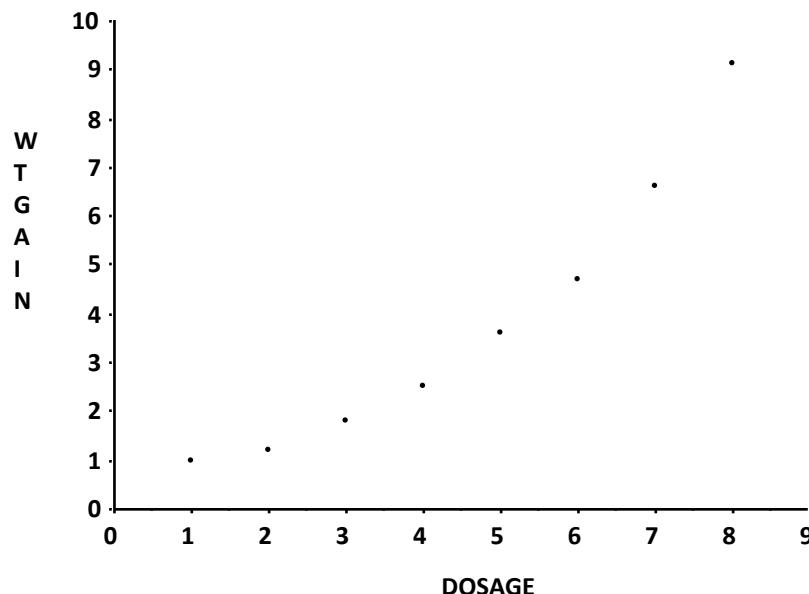
Let x = dose of a certain drug

y = weight gain (in decagrams) after 2 weeks

$n = 8$ laboratory animals were used and each assigned to one of eight dosage levels of the drug.

x (Dosage)	1	2	3	4	5	6	7	8
y (Weight Gain)	1	1.2	1.8	2.5	3.6	4.7	6.6	9.1

Scatter diagram:



If we had fit a straight-line regression to these data we would find

$$\hat{y} = -1.20 + 1.11x$$

ANOVA (straight-line model)

Source	df	SS	MS	F
Regression (x)	1	52.04	52.04	61.95
Residual	6	5.03	0.84	
Total	7	57.07		

note that $r^2 = 0.912$

and $F = 61.95$ is compared to $F_{.999}(1, 6) = 35.51$

i.e., $p < .001$

Let us decide whether or not the addition of the x^2 term significantly improves the prediction of y over and above that achieved via a straight-line.

$$2^{\text{nd}} \text{ order equation: } \hat{y} = 1.35 - 0.41x + 0.17x^2$$

ANOVA

Source	df	SS	MS	F
Regression $\begin{cases} x \\ x^2 x \end{cases}$	1	52.04	52.04	61.95
	1	4.83	4.83	120.75
Residual	5	0.2	0.04	
Total	7	57.07		

$$\leftarrow F(1,6)$$

$$\leftarrow F(1,5)$$

here $R^2 = .997$

We would like to know whether the increase of $(.997 - .912) = .085$ in R^2 represents a significant improvement in the fit.

The test for this is

$$F = \frac{(\text{extra SS due to adding } x^2)/1}{\text{MS residual for 2}^{\text{nd}} \text{ order model}} = \frac{4.83}{0.04} = 120.75$$

and $F_{.999}(1,5) = 47.18$ ($p < .001$)

\therefore reject H_0

Hence, the addition of the x^2 term to the model significantly improves the prediction.

Also, the test of the overall 2nd order model is highly significant.

$$F = \frac{\text{MS regression} (2^{\text{nd}}\text{- order model})}{\text{MS residual } (2^{\text{nd}}\text{- order model})} = \frac{(52.04 + 4.83)/2}{0.04} = 710.88$$

Hence, the straight-line model is not as good as the 2nd order model.

Can the 2nd order model be improved upon?

- let us add the x^3 term to the model and see if it improves the prediction.

ANOVA (3rd order model)

Source	df	SS	MS	F
Regression	x	1	52.040	52.04
	x^2	1	4.830	4.83
	x^3	1	0.140	0.14
Residual	4	0.056	0.014	
Total	7	57.066		

Here $R^2 = .999$

is the increase in $R^2 = (.999 - .997 = .002)$ significant?

H_0 : the addition of the x^3 term is not worthwhile

$$F = \frac{(\text{extra SS due to adding } x^3)/1}{\text{MS residual for 3rd order model}} = \frac{0.14}{.014} = 10.0$$

and $F \sim F(1, 4)$

$$F_{.95}(1, 4) = 7.71$$

$$F_{.975}(1, 4) = 12.22$$

$$.025 < p < .05$$

I still wouldn't add x^3 since

- (1) R^2 for the 2nd order model was very high = .997
- (2) Increase in R^2 was only .002
- (3) Tolerance suggests multicollinearity
- (4) Scatter diagram suggests 2nd order model
- (5) When in doubt use the simplest model
 - this promotes ease of interpretation

Hence the best fitting model is

$$\hat{y} = 1.35 - 0.41x + 0.17x^2$$

with $R^2 = 0.997$

Finally, the computer programs give us the standard errors associated with each β .

Coeff $\hat{\beta}_i$	$s_{\hat{\beta}_i}$
$\hat{\beta}_1 = -.41$	$s_{\hat{\beta}_1} = .141$
$\hat{\beta}_2 = .17$	$s_{\hat{\beta}_2} = .015$

Using these we can compute confidence intervals

$$\hat{\beta}_i - [t_{.975}(n-1-k)]s_{\hat{\beta}_i} \leq \beta_i \leq \hat{\beta}_i + [t_{.975}(n-1-k)]s_{\hat{\beta}_i}$$

95% confidence interval

e.g.,

$$0.17 - (2.571)(.015) \leq \beta_2 \leq 0.17 + (2.571)(.015)$$

$t_{.975}(5)$ $.13 \leq \beta_2 \leq .21$

note that 0 is not in the interval

t -tests can also be constructed in the obvious way

. regress wtgain dose

Source	SS	df	MS	Number of obs	=	8

Model	52.037204	1	52.037204	F(1, 6)	=	62.05
Residual	5.03154917	6	.838591529	Prob > F	=	0.0002

Total	57.0687531	7	8.15267902	R-squared	=	0.9118

wtgain	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	

dose	1.113095	.1413027	7.877	0.000	.7673399	1.458851
_cons	-1.196429	.7135439	-1.677	0.145	-2.942408	.5495503

. vif

Variable	VIF	1/VIF

dose	1.00	1.000000

Mean VIF	1.00	

. regress wtgain dose dosesq

Source	SS	df	MS	Number of obs	=	8
Model	56.8720267	2	28.4360133	F(2, 5)	=	722.73
Residual	.196726451	5	.03934529	Prob > F	=	0.0000
Total	57.0687531	7	8.15267902	R-squared	=	0.9966

wtgain	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dose	-.4136907	.1410916	-2.932	0.033	-.7763782	-.0510031
dosesq	.1696429	.0153035	11.085	0.000	.1303039	.2089819
_cons	1.348215	.276736	4.872	0.005	.6368421	2.059587

. vif

Variable	VIF	1/VIF
dose	21.25	0.047059
dosesq	21.25	0.047059
Mean VIF	21.25	

```
. regress wtgain dose dosesq dosecube
```

Source	SS	df	MS	Number of obs	=	8
Model	57.0129739	3	19.0043246	F(3, 4)	=	1362.82
Residual	.055779265	4	.013944816	Prob > F	=	0.0000
Total	57.0687531	7	8.15267902	R-squared	=	0.9990
				Adj R-squared	=	0.9983
				Root MSE	=	.11809

wtgain	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
dose	.379618	.2632868	1.442	0.223	-.3513834 1.110619
dosesq	-.0383118	.0660419	-0.580	0.593	-.2216734 .1450498
dosecube	.0154041	.0048452	3.179	0.034	.0019516 .0288565
_cons	.585714	.2909724	2.013	0.114	-.222155 1.393583

```
. vif
```

Variable	VIF	1/VIF
dosesq	1116.59	0.000896
dosecube	399.01	0.002506
dose	208.78	0.004790
Mean VIF	574.79	

Applied Regression Analysis

Week 5

1. Homework week 4: highlights
2. Multiple regression
 - Graphical interpretation / Assumptions I
 - Assumptions II / Least squares estimation
 - Computer output
 - Step by step review
3. Hypothesis testing: F-test / partial F-test
4. Homework

Stanley Lemeshow, Professor of Biostatistics
College of Public Health, The Ohio State University



THE OHIO STATE UNIVERSITY

WEEK 5: MULTIPLE REGRESSION

Suppose we wish to predict one variable, y , from k independent variables x_1, x_2, \dots, x_k , $k > 1$

y = "dependent" variable

x_1, x_2, \dots, x_k = "independent" variables

The general form of the regression model for k independent variables is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Regression coefficients
That need to be estimated

Independent variables

error

Note: in the 2nd order model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

if we let

$$\left. \begin{array}{l} x_1 = x \\ x_2 = x^2 \end{array} \right\} \text{here we really have 1 independent variable. } x_2 \text{ is a function of that variable.}$$

Then we can write this as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

In the multiple regression model some of the x_i may be functions of a few basic variables.

It should be noted that, with respect to what has come before

(1) It is sometimes difficult to determine the best choice of model.

- There will sometimes be several reasonable candidates to choose from.

(2) It is difficult (if not impossible) to visualize what the fitted model looks like.

- not possible to plot the data or the model when $k > 3$.

(3) Sometimes the best-fitting model will be difficult to interpret in real-life terms.

(4) Computations can't be done by hand

- high-speed computers are necessary
- reliable packaged computer program is necessary

Example:

$y = \text{weight}$ (WGT)

$x_1 = \text{height}$ (HGT)

$x_2 = \text{age}$ (AGE)

There are $n = 12$ children available, each having a particular kind of nutritional deficiency

The data are:

Child	y WGT	x_1 HGT	x_2 AGE
1	64	57	8
2	71	59	10
3	53	49	6
4	67	62	11
5	55	51	8
6	58	50	7
7	77	55	10
8	57	48	9
9	56	42	10
10	51	42	6
11	76	61	12
12	68	57	9

Many models are possible. For example

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

or

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where $x_3 = x_1^2$

or

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

where $x_3 = x_1^2$, $x_4 = x_2^2$, $x_5 = x_1 x_2$

so, this is equivalent to

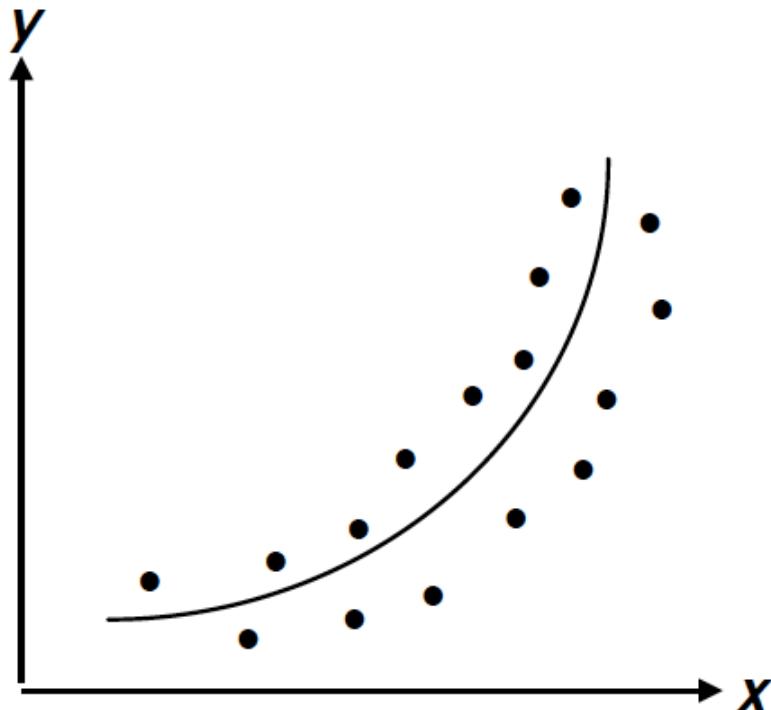
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$$

choice of best model is a topic to be considered later

One reasonable criterion might be to choose the one with the max R^2 .

Graphical Interpretation

If we had a single independent variable our lives would be quite simple (even if we have higher-order polynomial models)



The regression equation is the path described by the mean values of the distribution of y when x is allowed to vary.

When $k \geq 2$ our problems increase significantly

We no longer deal with a line or a curve but, rather, with a hyper surface in $(k + 1)$ - dimensional space.

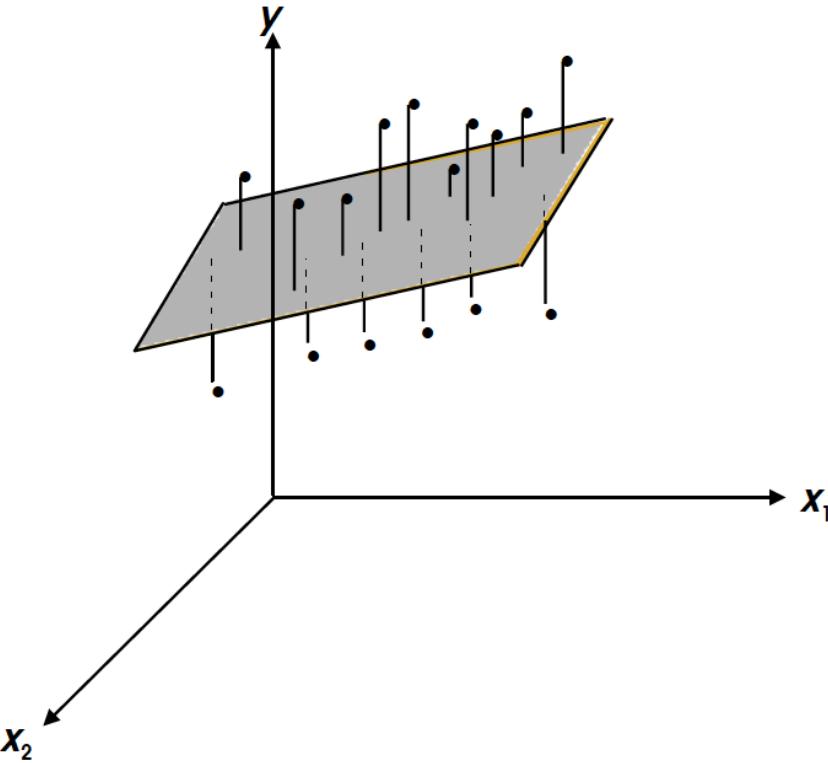
If $k > 2$, we can't plot the scatter of points or the regression equation.

For $k = 2$ we seek the surface in 3-dimensional space that best fits the scatter of points $(x_{11}, x_{21}, y_1), (x_{12}, x_{22}, y_2), \dots, (x_{1n}, x_{2n}, y_n)$

In this case, the regression equation is the surface described by the mean values of y at various combinations of x_1, x_2 .

i.e., at each distinct pair of values x_1 and x_2 there is a distribution of y values with mean $\mu_{y|x_1, x_2}$ and variance $\sigma^2_{y|x_1, x_2}$.

- The simplest curve in two-dimensional space is the straight line.
- The simplest surface in three-dimensional space is a plane that has the statistical model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$



In the three dimensional case, the least squares solution giving the best fitting plane is determined by minimizing the sum of squares of distances between the observed y_i , and the predicted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$ based on the fitted plane.

i.e., minimize $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2$

Assumptions of Multiple Regression

- (1) For each specific combination of x_1, x_2, \dots, x_k , y is a (univariable) random variable with a certain probability distribution.
- (2) The y observations are statistically independent.
- (3) The mean value of y at x_1, x_2, \dots, x_k is a linear function of x_1, \dots, x_k .
i.e., $\mu_{y|x_1, x_2, \dots, x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$
or $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$

Note:

(a) The surface $\mu_{y|x_1, x_2, \dots, x_k} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$ is called the *regression equation* or *response surface* or *regression surface*.

(b) If some of the independent variables are higher-order functions of a few basic independent variables (e.g., $x_3 = x_1^2$, $x_5 = x_1 x_2$) then $\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$ is really nonlinear in the basic variables. Hence we use the term "surface" rather than "plane".

We can use the multiple regression techniques so long as the model is inherently linear in the regression coefficients.

e.g., $\mu_{y|x} = \beta_0 e^{\beta_1 x}$ is inherently linear

since $\ln(\mu_{y|x}) = \ln(\beta_0) + \beta_1 x_1$

$$\mu_{y|x}^* = \beta_0^* + \beta_1 x_1$$

For this we
need nonlinear
regression
procedures

However,

$$\mu_{y|x_1, x_2} = e^{\beta_1 x_1} + e^{\beta_2 x_2}$$

cannot be transformed directly into
a form that is linear in β_1 and β_2

(c) ε is the error component in the model. It is the amount by which any individual's observed response deviates from the response surface.

Assumptions (cont'd)

$$(4) \quad \sigma_{y|x_1, x_2, \dots, x_k}^2 = \text{var}(y|x_1, x_2, \dots, x_k) \equiv \sigma^2$$

i.e., homoskedasticity

In general, mild departures from this assumption will not adversely affect the results.

(5) For any fixed x_1, x_2, \dots, x_k y is normally distributed

i.e.,
$$Y \sim N\left(\mu_{y|x_1, \dots, x_k}, \sigma^2\right)$$

These assumptions are not necessary for obtaining least squares estimates but are necessary for hypothesis testing and other inferential techniques.

Fortunately, usual parametric techniques used in regression analysis are “robust” in the sense that only extreme departures from the assumptions may yield spurious results.

Least Squares Estimates of Parameters

Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$

denote the fitted least squares regression model

The values $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ are chosen so that

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki})^2$$

is smaller than would be the case with any other value of $\hat{\beta}_i$

This minimum sum of squares is generally called the
"residual sum of squares"
"error sum of squares"
"sum of squares about regression"

The $\hat{\beta}_i$, determined with the method of least squares are also the minimum variance unbiased estimates of β_i .

The least-squares regression equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

is that unique linear combination of the independent variables x_1, x_2, \dots, x_k that has maximum possible correlation with the dependent variable.

i.e.,

$r_{y,\hat{y}}$ is greater than $r_{y,\hat{y}'}$ where \hat{y}' is any other linear combination of the x 's

Also note:

- each $\hat{\beta}_i$ is a linear function of the y values
- since y is assumed to be normally distributed, each of the estimators $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ will be normally distributed
- computer programs will provide us with these as well as their estimated variances. t – tests and confidence intervals would be carried out in the usual manner.

Example

For the WGT, HGT and AGE data

$$\text{WGT} = \beta_0 + \beta_1 \text{HGT} + \beta_2 \text{AGE} + \beta_3 \text{AGE}^2 + \varepsilon$$

The least squares estimates are

$$\hat{\beta}_0 = 3.438$$

$$\hat{\beta}_1 = 0.724$$

$$\hat{\beta}_2 = 2.777$$

$$\hat{\beta}_3 = -0.042$$

so

$$\widehat{WGT} = 3.438 + .724(HGT) + 2.77(AGE) - .042(AGE)^2$$

The ANOVA table for this model is:

ANOVA

Source	df	SS	MS	F
Regression	3	$SSY - SSE = 693.06$	231.02	9.47
Residual	8	$SSE = 195.19$	24.40	
Total	11	$SSY = 888.25$		

$$R^2 = 0.7802$$

To get the ANOVA table we use the familiar partitioning

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



Total SS = total variability in y before accounting for
the joint effect of using the independent variables
HGT, AGE, AGE²

Residual SS = SS due error
= amount of y variation left unexplained
after the independent variables have
been used in the regression equation
to predict y .

Regression SS = reduction in variation (or variation explained) due to the
independent variables in the regression equation.

now, to test H_0 : all k independent variables considered together do not explain a significant amount of the variation in y ,

or $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

vs. $H_a : \text{some } \beta_i \neq 0$

we use

$$F = \frac{\text{MS regression}}{\text{MS residual}}$$

and $F \sim F(k, n-1-k)$

The hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

vs. $H_a : \text{not all } \beta_i = 0$

can also be tested by an equivalent expression

$$F = \frac{R^2}{1-R^2} \frac{(n-1-k)}{k}$$

which is also compared to $F(k, n-1-k)$

note:

$$R^2 = \frac{\text{SSY} - \text{SSE}}{\text{SSY}}$$

Example

In the HGT, WGT, AGE example, from the ANOVA table

$$F = \frac{\text{MS regression}}{\text{MS residual}} = \frac{231.02}{24.40} = 9.47$$

and $F_{99}(3,8) = 7.59 \therefore p < .01$

also

$$F = \frac{R^2}{1-R^2} \frac{(n-1-k)}{k} = \frac{.7802}{1-.7802} \frac{(12-1-3)}{3} = 9.47$$

$\therefore \text{reject } H_0$

note:

MS residual = $\frac{1}{n-1-k} \text{SSE} = \frac{1}{n-1-k} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ is an unbiased

estimate of σ^2 under the assumed model.

MS regression is an independent estimate of σ^2 only if H_0 is true. Otherwise it overestimates σ^2 .

Hence we always reject if F gets too large.

```
. gen agesq=age*age
```

```
. regress wgt hgt age agesq
```

Source	SS	df	MS	Number of obs	=	12
Model	693.060463	3	231.020154	F(3, 8)	=	9.47
Residual	195.189537	8	24.3986921	Prob > F	=	0.0052
Total	888.25	11	80.75	R-squared	=	0.7803

wgt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
hgt	.7236902	.2769632	2.613	0.031	.085012 1.362368
age	2.776875	7.427279	0.374	0.718	-14.35046 19.90421
agesq	-.0417067	.4224071	-0.099	0.924	-1.015779 .9323659
_cons	3.438426	33.61082	0.102	0.921	-74.06826 80.94512

```
. vif
```

Variable	VIF	1/VIF
agesq	89.97	0.011115
age	89.68	0.011150
hgt	1.61	0.620927
Mean VIF	60.42	

. regress wgt hgt age

Source	SS	df	MS	Number of obs	=	12
Model	692.822607	2	346.411303	F(2, 9)	=	15.95
Residual	195.427393	9	21.7141548	Prob > F	=	0.0011
Total	888.25	11	80.75	R-squared	=	0.7800
				Adj R-squared	=	0.7311
				Root MSE	=	4.6598

wgt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
hgt	.722038	.2608051	2.768	0.022	.1320559 1.31202
age	2.050126	.9372256	2.187	0.056	-.0700253 4.170278
_cons	6.553048	10.94483	0.599	0.564	-18.20587 31.31197

. vif

Variable	VIF	1/VIF
age	1.60	0.623202
hgt	1.60	0.623202
Mean VIF	1.60	

```
. regress wgt hgt
```

Source	SS	df	MS	Number of obs	=	12
<hr/>						
Model	588.922523	1	588.922523	F(1, 10)	=	19.67
Residual	299.327477	10	29.9327477	Prob > F	=	0.0013
<hr/>						
Total	888.25	11	80.75	R-squared	=	0.6630
				Adj R-squared	=	0.6293
				Root MSE	=	5.4711

wgt	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
<hr/>					
hgt	1.07223	.241731	4.436	0.001	.5336202 1.610841
_cons	6.189849	12.84875	0.482	0.640	-22.43894 34.81864

```
. vif
```

Variable	VIF	1/VIF
<hr/>		
hgt	1.00	1.000000
<hr/>		
Mean VIF	1.00	

The previous ANOVA Table may be presented as follows.

Source		df	SS	MS	F
Regression	x_1	1	588.92	588.92	19.67 ***
	$x_2 x_1$	1	103.90	103.90	4.78
	$x_3 x_1, x_2$	1	0.24	0.24	0.01 *
Residual		8	195.19	24.40	NS
Total	11		888.25		

*** p<.01

* .05<p<.10

Here

$$SS(x_1) =$$

SS explained just using x_1 = HGT alone

$$SS(x_2 | x_1) =$$

extra SS explained by using x_2 = AGE

in addition to x_1 in predicting y

$$SS(x_3 | x_1, x_2) =$$

extra SS explained by using x_3 = AGE²

in addition to x_1 and x_2 in predicting y

Questions:

1. Does $x_1 = \text{HGT}$ alone significantly aid in predicting y ?
2. Does the addition of $x_2 = \text{AGE}$ significantly contribute to the prediction of y after controlling for the contribution of x_1 ?
3. Does the addition of $x_3 = \text{AGE}^2$ significantly contribute to the prediction of y after controlling for the contribution of x_1 and x_2 ?

Let us consider these one at a time

Question 1: Does $x_1 = \text{HGT}$ alone significantly aid in predicting y ?

Fit the straight line model $y = \beta_0 + \beta_1 \times \text{HGT}$

$\text{SS}(x_1) = 588.92 = \text{regression SS for this straight line model}$

$$\begin{aligned}\text{SSE} &= \text{SS}(x_2|x_1) + \text{SS}(x_3|x_1, x_2) + \text{SS resid} \\ &= 103.90 + 0.24 + 195.19 = 299.33\end{aligned}$$

$$\begin{aligned}\text{df} &= \text{df}(x_2|x_1) + \text{df}(x_3|x_1, x_2) + \text{df resid} \\ &= 1 + 1 + 8 = 10\end{aligned}$$

$$\therefore \text{MS resid} = \frac{299.33}{10} = 29.933$$

and

$$F = \frac{\text{MS regression}}{\text{MS resid}} = \frac{588.92}{29.933} = 19.67 \quad \text{as in the table}$$
$$F \sim F(1, 10) \text{ here } p < .01$$

i.e., x_1 contributes significantly to the linear prediction of y

Question 2: Does the addition of $x_2 = \text{AGE}$ significantly contribute to the prediction of y after controlling for the contribution of x_1 ?

To answer this we use a "partial F-test". This test allows for the elimination of variables that are of no help in predicting y and thus enables one to reduce the set of possible independent variables to an economical set of "important" predictors.

To perform a partial F -test concerning a variable x^* , say,
given that x_1, x_2, \dots, x_p are already in the model we:

(1) Compute the "extra SS from adding x^* , given x_1, x_2, \dots, x_p "

- This is placed into the ANOVA table under the source heading "Regression $x^* | x_1, x_2, \dots, x_p$ "

$$\text{Extra SS from adding } x^* \text{ given } x_1, x_2, \dots, x_p = \frac{\text{regression SS when } x_1, x_2, \dots, x_p \text{ and } x^* \text{ are all in the model}}{\text{regression SS when } x_1, x_2, \dots, x_p \text{ and not } x^* \text{ are all in the model}}$$

or

$$\begin{aligned} \text{SS}(x^* | x_1, x_2, \dots, x_p) &= \text{regression SS}(x_1, x_2, \dots, x_p, x^*) \\ &\quad - \text{regression SS}(x_1, x_2, \dots, x_p) \end{aligned}$$

In our example

$$\begin{aligned} \text{SS}(x_2|x_1) &= \text{regression SS}(x_1, x_2) - \text{regression SS}(x_1) \\ &= 692.82 - 588.92 \\ &= 103.90 \end{aligned}$$

$$\begin{aligned} \text{SS}(x_3|x_1, x_2) &= \text{regression SS}(x_1, x_2, x_3) - \text{regression SS}(x_1, x_2) \\ &= 693.06 - 692.82 \\ &= 0.24 \end{aligned}$$

To test

H_0 : The addition of x^* to a model already containing x_1, x_2, \dots, x_p does not significantly improve the prediction of y

we compute

$$F(x^* | x_1, x_2, \dots, x_p) = \frac{SS(x^* | x_1, x_2, \dots, x_p)}{\text{MS residual}(x_1, x_2, \dots, x_p, x^*)}$$

and

$$F(x^* | x_1, x_2, \dots, x_p) \sim F(1, n - p - 2)$$

In our example

$$F(x_2|x_1) = \frac{SS(x_2|x_1)}{\text{MS residual}(x_1, x_2)} = \frac{103.90}{\left(\frac{.24 + 195.19}{1+8} \right)} = 4.78$$

$$F_{.90}(1, 9) = 3.36; F_{.95}(1, 9) = 5.12$$

and

$$F(x_3|x_1, x_2) = \frac{SS(x_3|x_1, x_2)}{\text{MS residual}(x_1, x_2, x_3)} = \frac{0.24}{24.40} = 0.01$$

Hence, the addition of x_2 after accounting for x_1 significantly adds to the prediction of y at the $\alpha = 0.10$ level.

Had we used $\alpha = 0.05$ we would not add x_2 .

Once $x_1 = \text{HGT}$ and $x_2 = \text{AGE}$ are in the model, the addition of $x_3 = \text{AGE}^2$ is superfluous.

There is an alternative (but equivalent) way to perform the partial F -test. That involves a test of

$$H_0 : \beta^* = 0$$

where β^* is the coefficient of x^* in

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \beta^* x^* + \varepsilon$$

Here

$$t = \frac{\hat{\beta}^*}{s_{\hat{\beta}^*}} \quad \begin{array}{l} \leftarrow \text{estimated coefficient} \\ \leftarrow \text{estimated standard error of } \hat{\beta}^* \end{array} \quad \left. \begin{array}{l} \text{printed by} \\ \text{computer programs} \end{array} \right\}$$

$$\left. \begin{array}{l} \text{reject } H_0 \text{ if } t > t_{1-\alpha/2}(n-p-2) \\ \text{or if } t < t_{\alpha/2}(n-p-2) \end{array} \right\} \quad \begin{array}{l} \text{2- sided test of } H_0 \\ H_a : \beta^* \neq 0 \end{array}$$

similarly, one sided tests can be constructed

e.g., for $H_a : \beta^* > 0$ (reject if $t > t_{1-\alpha}(n-p-2)$)

In our example

$$H_0 : \beta_2 = 0$$

$$H_a : \beta_2 \neq 0$$

in the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Then

$$t = \frac{\hat{\beta}_2}{s_{\hat{\beta}_2}} = \frac{2.050}{0.937} = 2.188$$

and $t_{.95}(9) = 1.833$, $t_{.975}(9) = 2.2622$

Hence $.05 < p < .10$ since 2-sided

Note $t^2 = 2.188^2 = 4.79 = \text{partial } F(x_2|x_1)$
in ANOVA table

and

$$t_{1-\alpha/2}^2(9) = F_{1-\alpha}(1, 9)$$

Similarly, when testing

$$\begin{aligned} H_0: \beta_3 &= 0 \\ \text{vs } H_\alpha: \beta_3 &\neq 0 \end{aligned}$$

in the model $y = \beta_0 + \beta_1 x_1 + \beta_3 x_2 + \beta_3 x_3 + \varepsilon$

we compute

$$t = \frac{\hat{\beta}_3}{s_{\hat{\beta}_3}} = \frac{-0.042}{0.422} = -0.0995$$

and

$t^2 = (-0.0995)^2 = .01 = \text{Partial } F(x_3|x_1, x_2)$
in ANOVA table

Applied Regression Analysis

Week 6

1. Homework week 5: highlights
2. Dummy variables
3. Statistical interaction
4. Comparing two straight line regression equations
 - Method 1: fitting separate models
 - Method 2: fitting a single model
5. Homework
6. Homework week 6: highlights

Stanley Lemeshow, Professor of Biostatistics
College of Public Health, The Ohio State University



THE OHIO STATE UNIVERSITY

A dummy variable is any variable in a regression equation that takes on a finite number of values

- used to indicate categories of a nominal scaled variable

The term “Dummy” simply means that the actual values used (e.g., 0,1 or +1, 0, -1) do not describe a meaningful measurement level of the variable

- instead they only “indicate” the categories of interest

These variables allow us to broaden the scope of regression analysis to include ANOVA, ANCOVA, Discriminant Analysis, and these variables will be widely used in logistic regression analysis.

Examples:

2 categories
1 dummy variable

$$\begin{cases} x_1 = \begin{cases} 1 & \text{If patient received Drug A} \\ 0 & \text{otherwise} \end{cases} \\ x_2 = \begin{cases} +1 & \text{If female} \\ -1 & \text{If male} \end{cases} \end{cases}$$

3 categories
2 dummy variables

$$\begin{array}{cc} \frac{x_3}{x_4} & \\ \hline 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{array} \quad \left\{ \begin{array}{l} \text{If treated at hospital A} \\ \text{If treated at hospital B} \\ \text{If treated at hospital C} \end{array} \right.$$

Another way to code a 3 category variable is:

Instead of
0 0


$$\begin{array}{cc} \frac{x_3}{x_4} & \\ \hline -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{array} \quad \left\{ \begin{array}{l} \text{If treated at hospital A} \\ \text{If treated at hospital B} \\ \text{If treated at hospital C} \end{array} \right.$$

Also, you can mix up the placement of the +1's, 0's and -1's.

Important note:

Different dummy variable definitions will lead to coefficients that have different meaning. However, the test procedures involve the same null hypothesis and test statistics.

General Rule

If the nominal independent variable has k categories, then one must define exactly $k-1$ dummy variables to index these categories, provided that the regression model contains a constant term (i.e., β_0).

If the regression model does not contain a constant term, then k dummy variables are needed to index the k categories of interest.

If k dummy variables are used to describe a nominal variable with k categories in a model containing a constant term, then all the coefficients in the model cannot be uniquely estimated.

Let x_1 and x_2 be two independent variables

Let y represent the dependent variable

Q: How do x_1 and x_2 “interact” to affect y ?

There is “no statistical interaction” between x_1 and x_2 if the relationship between x_1 and y is the “same” regardless of the value of x_2 and the relationship between x_2 and y is the “same” regardless of the value of x_1 .

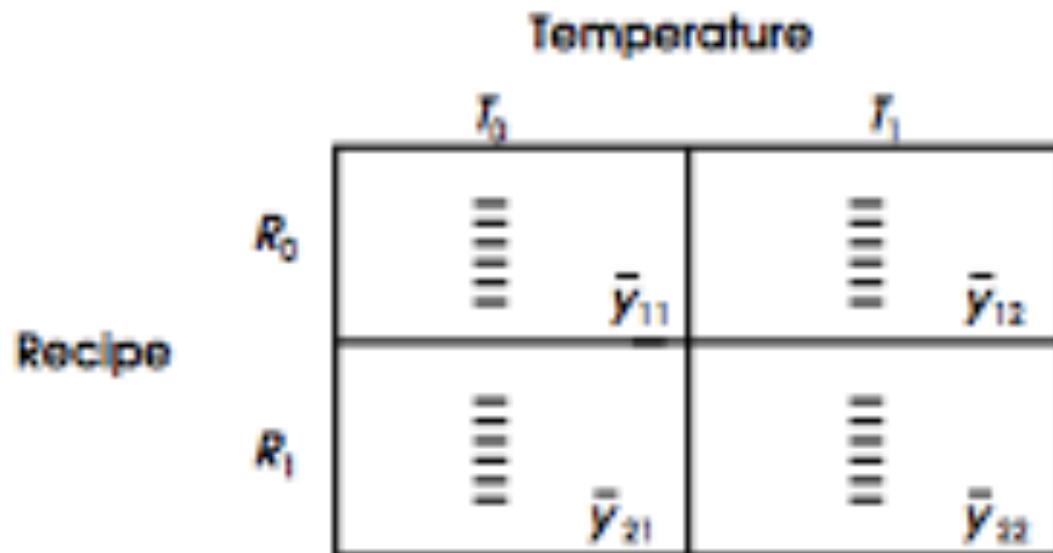
Example

let y = height a cake rises

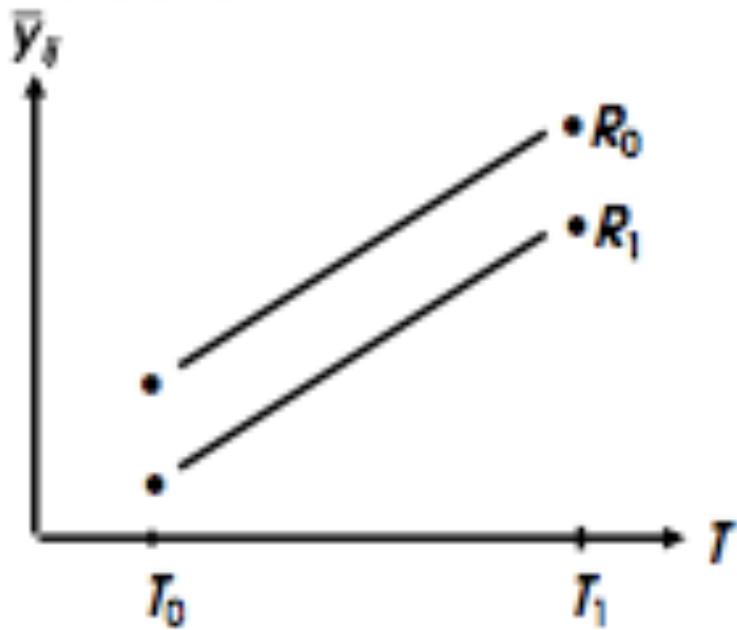
T = temperature \leftarrow two levels: T_0 and T_1

R = recipe \leftarrow two levels: R_0 and R_1

Of interest is to determine how the two independent variables, T and R , jointly affect the height the cake rises.



No interaction:



Here we have no interaction

i.e., The rate of change in Y as a function of temperature is the same regardless of which recipe is used.

i.e., The relationship between Y and T does not, in any way, depend on R .

note: We are not saying that Y and R are unrelated. We are saying that the relationship between Y and T is independent of the relationship between Y and R .

- When this is the case, we say that there is no $T \times R$ Interaction effect.
- This means that we can investigate the effects of T and R on Y independently of one another.

This relationship can be quantified with

$$\mu_{Y|T,R} = \beta_0 + \beta_1 T + \beta_2 R$$

where $T = \begin{cases} 0 \\ 1 \end{cases}$ $R = \begin{cases} 0 \\ 1 \end{cases}$

Here, the change in Y for a one-unit change in $T = \beta_1$,
Regardless of the level of R .

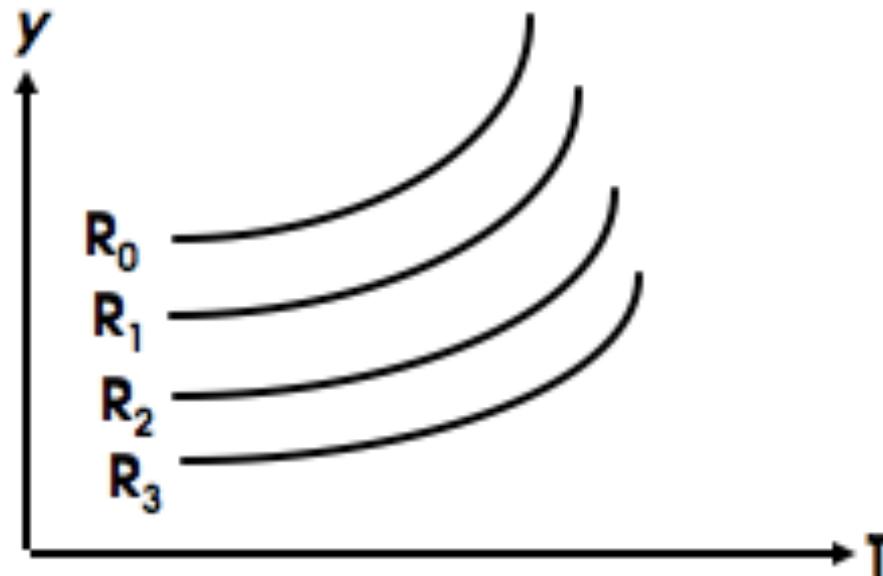
- Changing the level of R has the effect of shifting the straight line either up or down without affecting the value of the slope, β_1 .

$$\text{i.e., } \mu_{Y|T,R_0} = (\beta_0 + \beta_2 R_0) + \beta_1 T = \beta'_0 + \beta_1 T$$

$$\mu_{Y|T,R_1} = (\beta_0 + \beta_2 R_1) + \beta_1 T = \beta''_0 + \beta_1 T$$

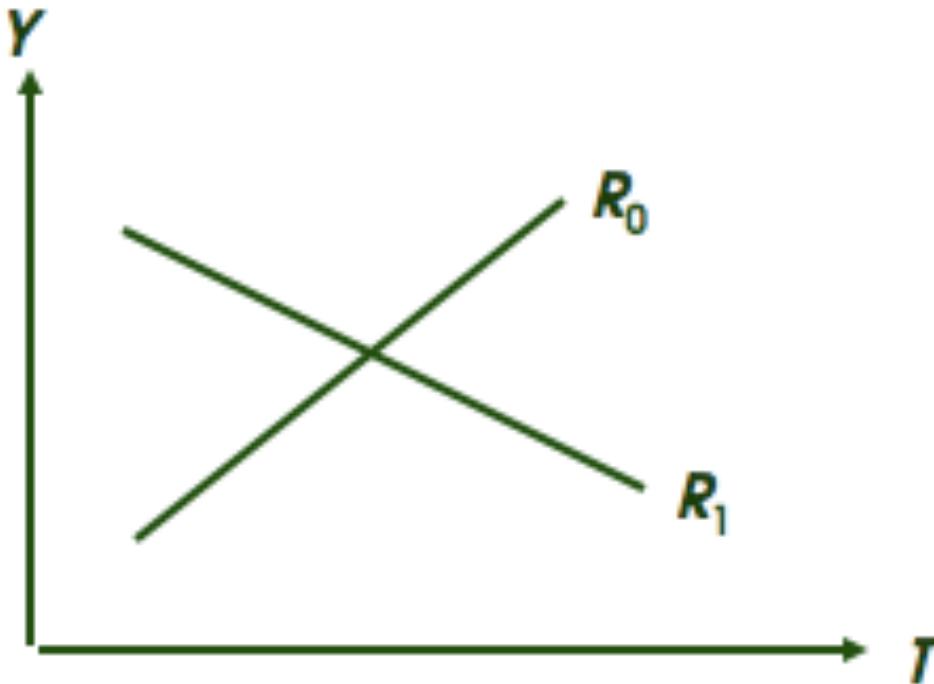
Hence, no interaction is synonymous with “parallelism”

Another example of no interaction is the following:



These response curves (that are non-linear) all have the same general shape, differing from each other only by an additive constant independent of T .

Now, consider the following alternative graph:



Here the relationship between Y and T depends upon R .

- Y decreases with Temperature with Recipe 1
- Y increases with Temperature with Recipe 0.
 - Hence there is an $R \times T$ interaction.

We can model this by: $\mu_{Y|R,T} = \beta_0 + \beta_1 T + \beta_2 R + \beta_{12} TR$

$$\mu_{Y|T,R} = \beta_0 + \beta_1 T + \beta_2 R + \beta_{12} TR$$

Here, the change in the mean value of Y for a 1-unit change in T is equal to

$$\begin{aligned}\mu_{Y|T+1,R} - \mu_{Y|T,R} &= \beta_0 + \beta_1(T+1) + \beta_2 R + \beta_{12}(T+1)R \\ &\quad - (\beta_0 + \beta_1 T + \beta_2 R + \beta_{12} TR) \\ &= \beta_1 + \beta_{12}R\end{aligned}$$

which clearly depends upon the value of R .

The introduction of a product term such as $\beta_{12}TR$ into the model is one way to account for the fact that two such factors as T and R do not operate independently of one another.

In our example, when $R = R_0$, the model can be written as:

$$\mu_{Y|R_0} = (\beta_0 + \beta_2 R_0) + (\beta_1 + \beta_{12} R_0)T$$

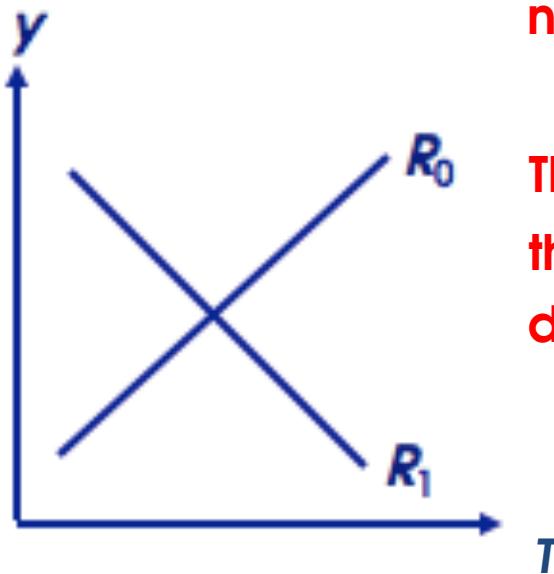
and, when $R = R_1$,

$$\mu_{Y|R_1} = (\beta_0 + \beta_2 R_1) + (\beta_1 + \beta_{12} R_1)T$$

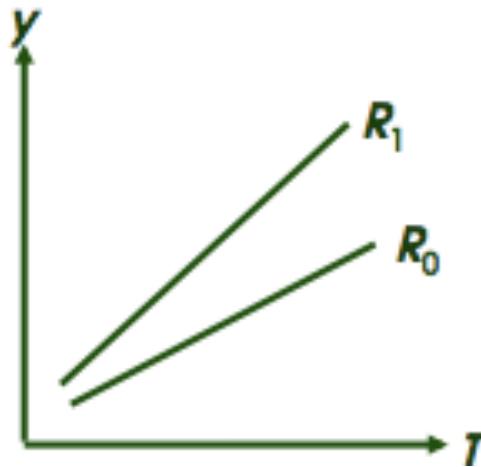
note: the linear effect of T at R_0 is positive $= \beta_1 + \beta_{12} R_0$

the linear effect of T at R_1 is negative $= \beta_1 + \beta_{12} R_1$

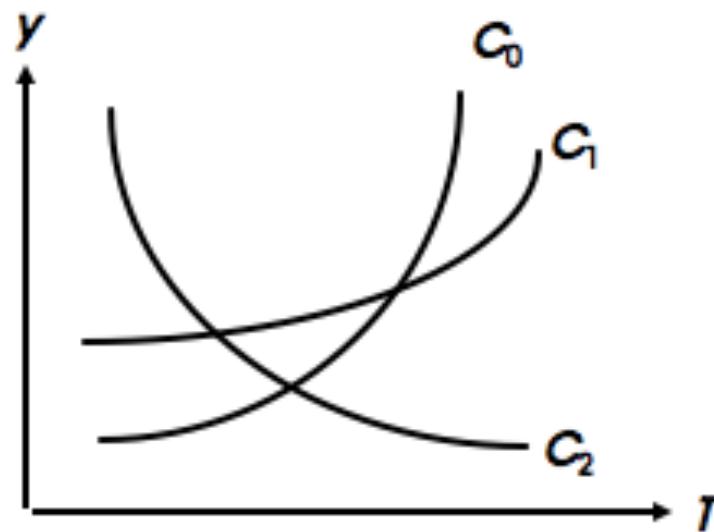
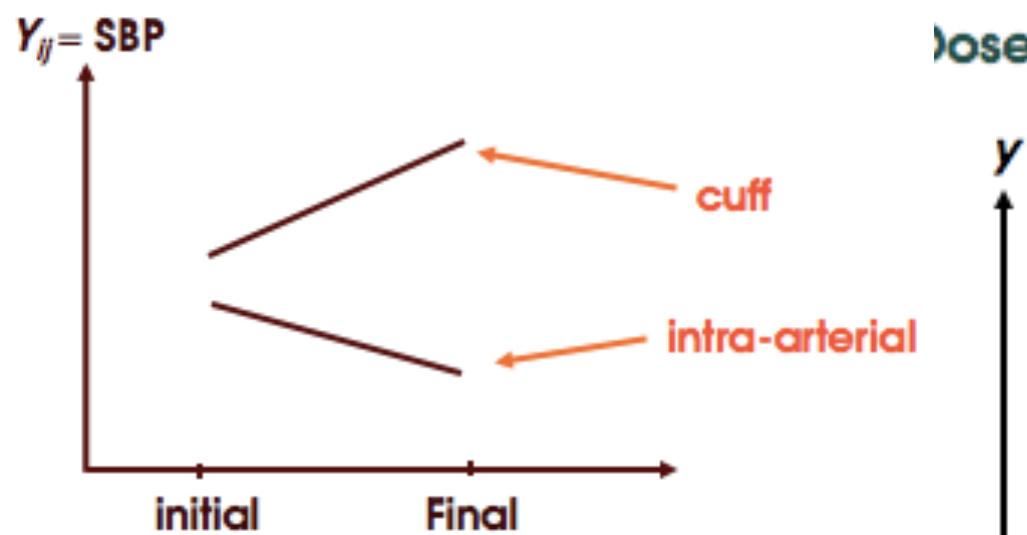
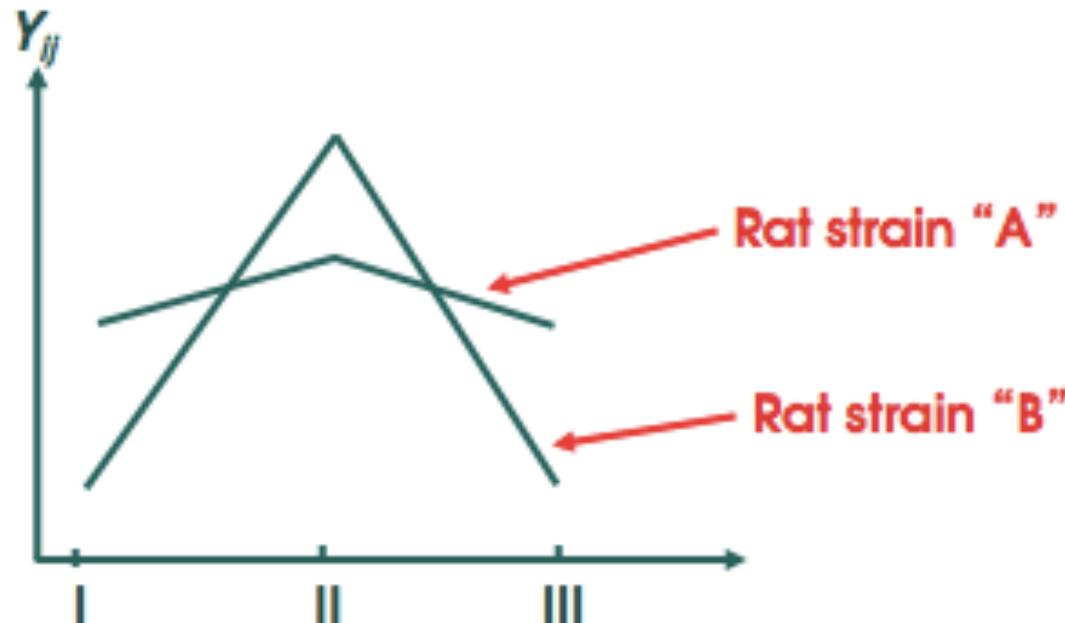
This suggests that the interaction β_{12} is negative since
the slope of the linear relationship between Y and T
decreases as R changes from R_0 to R_1 .



β_{12} positive could occur in a situation as follows:



Other examples of interaction:



Suppose we want to compare males and females w.r.t. their separate straight line regressions of SBP(y) on AGE (x).

Let $n_m = \#(x, y)$ pairs for the males

$n_f = \#(x, y)$ pairs for the females

Two Methods:

Method 1: fit two separate regression equations:

$$\left. \begin{array}{l} y_m = \beta_{0m} + \beta_{1m} x + \varepsilon \\ y_f = \beta_{0f} + \beta_{1f} x + \varepsilon \end{array} \right\} \text{Treats male and female data separately}$$

Then use statistical methods to compare β_{1m} and β_{1f}
or to compare β_{0m} and β_{0f}

Method 2: Define

$$z = \begin{cases} 0 & \text{if male} \\ 1 & \text{if female} \end{cases}$$

for males: $(x_{1m}, y_{1m}, 0), (x_{2m}, y_{2m}, 0), \dots, (x_{n_m m}, y_{n_m m}, 0)$

for females: $(x_{1f}, y_{1f}, 1), (x_{2f}, y_{2f}, 1), \dots, (x_{n_f f}, y_{n_f f}, 1)$

Then fit a single model:

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$$

Note:

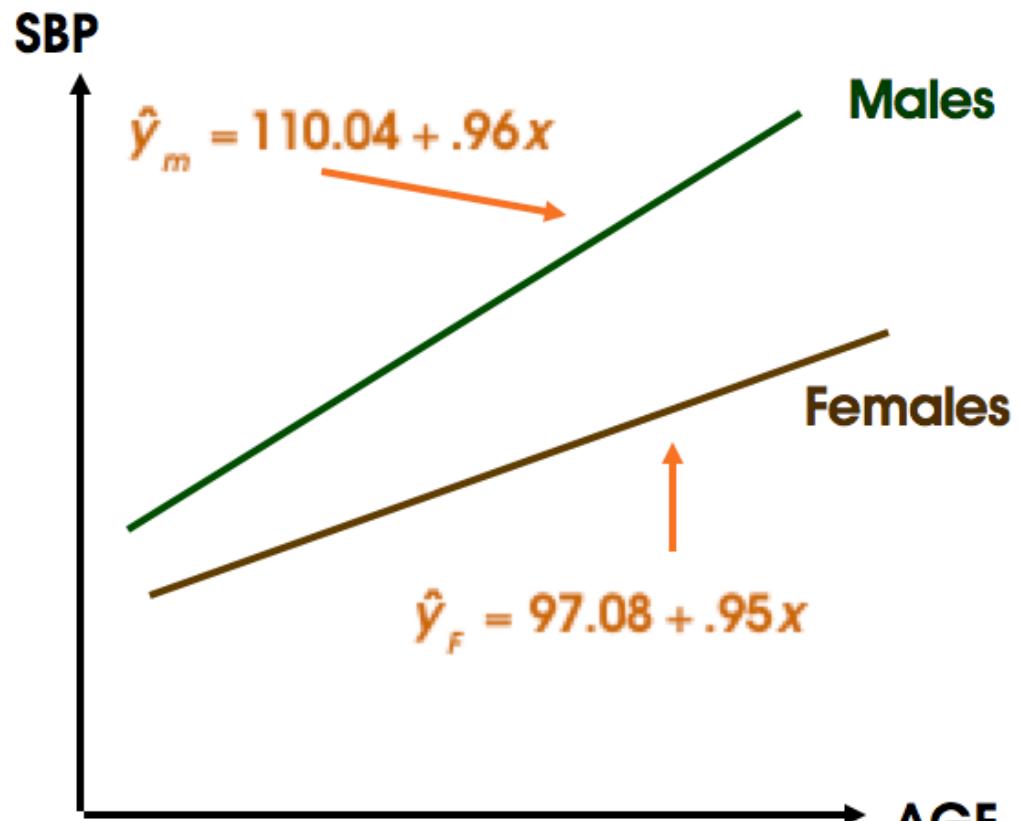
when $z = 0$, $y_m = (\beta_0) + \beta_1 x + \varepsilon = \beta_{0m} + \beta_{1m} x + \varepsilon$

$z = 1$, $y_f = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x + \varepsilon = \beta_{0f} + \beta_{1f} x + \varepsilon$

Hence, this model incorporates two separate regressions within a single model and allows for different slopes (β_1 for males and $\beta_1 + \beta_3$ for females) as well as different intercepts.

EXAMPLE:

	<u>MALES</u>	<u>FEMALES</u>
n	40	29
$\hat{\beta}_0$	110.04	97.08
$\hat{\beta}_1$	0.96	0.95
\bar{x}	46.93	45.07
\bar{y}	155.15	139.86
s_x^2	221.15	242.14
$s_{y x}^2$	71.90	91.46



Here, the separate regressions are $\hat{y}_m = 110.04 + .96x$
 $\hat{y}_f = 97.08 + .95x$

and the combined model is $\hat{y} = 110.04 + .96x - 12.96z - .012xz$

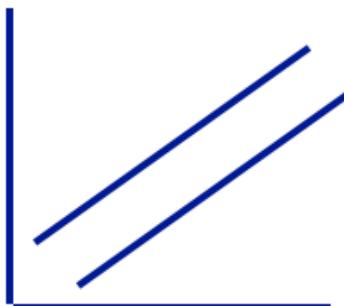
note:

when $z = 0$ (males), $\hat{y} = 110.04 + .96x$

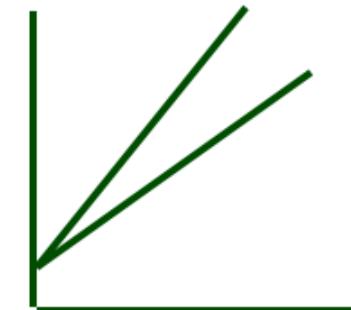
and when $z = 1$ (females), $\hat{y} = 110.04 + .96x$

$$-12.96(1) - .012x = 97.08 + .95x$$

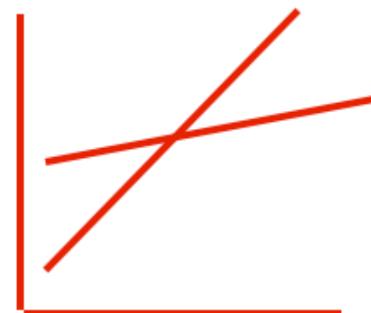
The appearance of the plots can be:



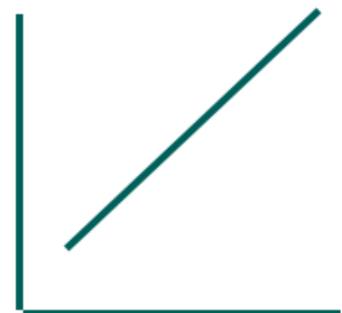
Parallel



Equal Intercepts



Different slopes
and intercepts



Coincident

Now, we are interested in testing

H_0 : The two regression lines are parallel.

or $H_0 : \beta_3 = 0$ (since then $\beta_{1f} = \beta_1 + \beta_3 = \beta_1 + 0 = \beta_1 = \beta_{1m}$)

This can be tested with the usual Partial F -test (or equivalent t -test) for the significance of the addition of the variable xz to the model already containing x and z .

The ANOVA table is:

Source	df	SS	MS	F
Regression (x)	1	14951.25	14951.25	121.27
Residual	67	8260.51	123.29	
Regression (x, z)	2	18009.78	9004.89	114.25
Residual	66	5201.99	78.82	
Regression (x, z, xz)	3	18010.33	6003.44	75.02
Residual	65	5201.44	80.02	

To test $H_0 : \beta_3 = 0$ we compute

$$F(xz|x, z) = \frac{SS_{\text{reg}}(x, z, xz) - SS_{\text{reg}}(x, z)}{MS_{\text{resid}}(x, z, xz)} = \frac{18010.33 - 18009.78}{80.02} = .007$$

and testing this against an $F(1, 65)$ leads to the conclusion that there is no evidence that the two lines are not parallel.

To see whether the two lines coincide we test

$$H_0 : \beta_2 = \beta_3 = 0$$

To do this we use

$$F(xz, z | x) = \frac{[SS_{\text{reg}}(x, z, xz) - SS_{\text{reg}}(x)]}{MS_{\text{resid}}(x, z, xz)}$$
$$= \frac{[18010.33 - 1495.25]}{80.02} = 103.19$$

and, since $F_{.999}(2, 65) = 7.72$, we reject H_0 with $p < .001$.

Hence there is strong evidence that the two lines are not coincident.

- at this point the complete model could be reduced to the form: $y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$

Feedback — Quiz: Week One

[Help Center](#)

You submitted this quiz on **Mon 23 Mar 2015 8:53 PM PDT**. You got a score of **6.00** out of **6.00**.

Question 1

What is the sum of the deviations about the mean?

Your Answer	Score	Explanation
-------------	-------	-------------

1

-1

0

 1.00

Great job. It is always true that the sum of the deviations below the mean will always equal the sum of the deviations above the mean, summing to 0.

-5

5

Total

1.00 /

1.00

Question 2

Which expression below will give you an unbiased estimate of the population variance?

Your Answer	Score	Explanation
-------------	-------	-------------

$\frac{\sum_{i=1}^n x_i^2}{n}$

$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$



1.00

Great job.

$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 2}$$

Total 1.00 / 1.00

Question 3

What theorem tells us that shape is the sampling distribution of the sample mean will be normal?

Your Score Explanation

Answer

Bayes
Theorem

Central Limit Theorem 1.00 Great job. The Central Limit Theorem states that the distribution of a large number of independent and identically distributed variables will be approximately normal.

Burke's
Theorem

None
of the
above

Total 1.00 /
1.00

Question 4

Is the following interpretation of a confidence interval True or False?

Upon repeated sampling, 95% of intervals constructed in the same way will contain the true population parameter.

Your Answer

Score

Explanation

True ✓ 1.00

False

Total

1.00 / 1.00

Question 5

Big p-values ($p>.05$) conclude which one of the following?

Your Answer**Score****Explanation**

Accept the null hypothesis

Fail to reject the null hypothesis ✓ 1.00 Good job! It is incorrect to 'accept' the null hypothesis. Statisticians will always use 'fail to reject the null hypothesis'.

Reject the null hypothesis

None of the above

Total

1.00 /
1.00

Question 6

What does a positive slope indicate?

Your Answer**Score****Explanation**

As X increases, Y increases ✓ 1.00 Great job! A positive slope always indicates that Y is increasing with X. If Y were to decrease with increases in X, this would indicate a negative slope.

No association between X and Y

As X increases, Y decreases

None of the
above

Total 1.00 /
 1.00

Feedback — Quiz: Week Two

[Help Center](#)

You submitted this quiz on **Mon 30 Mar 2015 11:37 AM PDT**. You got a score of **6.00** out of **6.00**.

Question 1

Slope being zero indicates that

Your Answer**Score** **Explanation**

- There is a strong linear relationship between X and Y.
- Slope does not say anything about the relationship between X and Y.
- There is no linear relationship between X and Y.

✓ 1.00 Good job!

Slope being zero indicates that the correlation coefficient between X and Y is zero.

Both correlation coefficient and slope measure linear relationship.

There could be any other relationship between X and Y such as quadratic, exponential etc.

- There is no relationship between X and Y.

Total 1.00 /
1.00

Question 2

The estimate of the intercept term β_0 cannot be negative.

Your Answer	Score	Explanation
-------------	-------	-------------

True

False ✓ 1.00 Great job!

The intercept term can be negative if the mean of the dependent variable (Y) is negative.

Total	1.00 / 1.00
-------	----------------

Question 3

The assumption of homoscedasticity means that the variance of Y is same for all X

Your Answer	Score	Explanation
-------------	-------	-------------

True ✓ 1.00 Nice work!

False

Total	1.00 / 1.00
-------	-------------

Question 4

The null hypothesis for testing the linear relationship between X and Y is

Your Answer	Score	Explanation
-------------	-------	-------------

$H_0 :$
 $\beta_0 \neq 0$

$H_0 :$ ✓ 1.00 Yes, you are right!
 $\beta_1 = 0$

The slope gives us an idea about the linear relationship between X and Y and in the null hypothesis we let the true slope be zero.

$H_0 :$
 $\beta_0 = 0$

$H_0 :$

$\beta_1 \neq 0$

Total	1.00 /
	1.00

Question 5

The test statistic for testing $\beta_1 = 0$ follows t-distribution with degrees of freedom

Your Answer	Score	Explanation
-------------	-------	-------------

n

$n - 2$ ✓ 1.00 Great job!

From n , 2 degrees of freedom are lost in estimating the intercept term and the slope.

$n - 3$

$n - 1$

Total	1.00 /
	1.00

Question 6

For the prediction interval, the farther away X_0 is from \bar{X} , the interval

Your Answer	Score	Explanation
-------------	-------	-------------

Gets wider ✓ 1.00 Yes, you got it!

The prediction interval depends on the term $\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}$.

As the distance between X_0 and \bar{X} increases, the term gets bigger and hence makes the prediction interval wider.

Gets smaller

Stays
the same

Total 1.00 /
 1.00

Feedback — Quiz: Week Three

[Help Center](#)

You submitted this quiz on **Fri 10 Apr 2015 12:24 PM PDT**. You got a score of **6.00** out of **6.00**.

Question 1

The correlation coefficient is **NOT** a measure of:

Your Answer

Score **Explanation**

1. The strength
of a straight line
relationship
between X and Y

2. The
association
between two
random variables
in a sample

3. The
association
between one
random variable
and one fixed
variable in a
sample

4. The
magnitude of the
slope of the
straight line
relationship
between X and Y

5. Options 3 and 4 above.  1.00 Good job.

The correlation coefficient is only a measure of the strength of a relationship between 2 **random** variables (a correlation).

Furthermore, the correlation coefficient can tell us the direction of a slope (if R is negative, the slope will be negative), but **not** the magnitude of the slope (R will

increase if the covariance between X & Y increases- even if the slope remains the same).

6. None of the above

Total 1.00 /
1.00

Question 2

Consider $\rho_{xy}=1$. Which of the following statements are true.

Select all that apply

Your Answer	Score	Explanation
<input checked="" type="checkbox"/> All of the variance in Y is explained by the regression of Y on X	✓ 0.33	With regards to $\rho_{xy}=1$, this statement is true.
<input checked="" type="checkbox"/> The relationship between X & Y is perfectly linear	✓ 0.33	With regards to $\rho_{xy}=1$, this statement is true.
<input checked="" type="checkbox"/> The conditional variance of Y = 0	✓ 0.33	With regards to $\rho_{xy}=1$, this statement is true.

Total 1.00 /
1.00

Question 3

The correlation coefficient is the proportion of the variance in Y explained by X.

(please answer True or False below)

Your Score Explanation
Answer

True

False ✓ 1.00 Good job!

The correlation coefficient squared is the proportion of the variance in Y explained by X

Total 1.00 /
 1.00

Question 4

If r^2 is high, we can assume that there is a strong linear association between X and Y.

(please answer True or False below)

Your Answer	Score	Explanation
-------------	-------	-------------

True

False ✓ 1.00 Good job!

r^2 is not a measure of the appropriateness of the straight-line model.

A non-linear model may better describe the relationship between X and Y, even if the r^2 is large.

Total 1.00 /
 1.00

Question 5

If the mean square due regression greatly exceeds the mean square due residual, then we reject the null and conclude the slope is zero.

(please answer True or False below)

Your Answer	Score	Explanation
-------------	-------	-------------

True

False ✓ 1.00 Good job.

In this case we would reject the null and conclude that there is

sufficient evidence that the slope does **not** equal zero.

Total	1.00 /
	1.00

Question 6

When looking at the STATA output of a regression of Y on X, which of the following could indicate that linear model is significantly better than the naive model?

Select all that apply

Your Answer	Score	Explanation
<input checked="" type="checkbox"/> The confidence interval for the slope	✓ 0.33	
<input type="checkbox"/> The r^2 value	✓ 0.33	The r^2 value does not indicate significance.
<input checked="" type="checkbox"/> The F statistic	✓ 0.33	
Total	1.00 /	
	1.00	

Feedback — Quiz: Week Four

[Help Center](#)

You submitted this quiz on **Mon 13 Apr 2015 9:46 PM PDT**. You got a score of **7.00** out of **7.00**.

Question 1

Which of the following models is an example of a second order polynomial model ($k = 2$)?

Your Answer	Score	Explanation
<input type="radio"/> $y = c_0 + c_1X + c_2X^3$		
<input checked="" type="radio"/> $y = c_0 + c_1X + c_2X^2$	✓ 1.00	Great job, this is correct!
<input type="radio"/> $y = c_0 + c_1X$		
Total	1.00 / 1.00	

Question 2

If you fit a polynomial regression model and obtain an overall p-value of 0.17 for the model, what does this indicate?

Your Answer	Score	Explanation
<input checked="" type="radio"/> This model does not fit better than the naïve model	✓ 1.00	Great job! In this instance, large p-values (>.05) indicate that the current model does not fit any better than the naïve model. If this p-value were significant at $p < .05$, this would indicate a better fit than the naïve model.
<input type="radio"/> This model fits better than the naïve model		
<input type="radio"/> This does not tell us anything about the model		

- This model has the same fit as the naïve model

Total	1.00 /
	1.00

Question 3

If you added any random variable into a model, what would happen to the sums of squares due regression?

Your Answer	Score	Explanation
<input type="radio"/> The sums of squares due regression would decrease		
<input checked="" type="radio"/> The sums of squares due regression would increase	1.00	Nice work! Adding any variable into a model will always increase the sums of squares due regression. Even if the added variable is a series of random numbers, by chance you will always end up artificially increasing the sums of squares due regression.
<input type="radio"/> The sums of squares due regression would remain the same		
<input type="radio"/> We do not have enough information to interpret this		

Total	1.00 /
	1.00

Question 4

What is a partial F test used for in polynomial regression?

Your Answer	Score	Explanation
<input checked="" type="radio"/> To determine if the addition of a new x^2 variable to the model already containing x is significant	✓ 1.00	Yes, you got it!
<input type="radio"/> To determine which model best fits the data		
<input type="radio"/> To determine the amount of variability in Y explained by X		
Total	1.00 / 1.00	

Question 5

How is extra sums of squares calculated in polynomial regression?

Your Answer	Score	Explanation
<input checked="" type="radio"/> By subtracting the due regression sums of squares for the straight line model from the due regression sums of squares for the polynomial model	✓ 1.00	Great job!
<input type="radio"/> By subtracting the total sums of squares for the straight line model from the total sums of squares for the polynomial model		
<input type="radio"/> This cannot be calculated		
Total	1.00 / 1.00	

Question 6

You conduct a partial F-test of a polynomial term and yield an F value of 16. You compare this value to a critical value of F of 2.02. Should the polynomial term be added to the model?

Your Answer	Score	Explanation
<input type="radio"/> No. You do not reject the null so the term is not significant.		
<input type="radio"/> No. You reject the null and the term is not significant.		

- Yes. You reject the null and conclude the term is significant.

✓ 1.00

Yes, this is correct!

Total

1.00 /
1.00

Question 7

Conclusions from t-test and partial f test are exactly equivalent.

(please answer True or False below)

Your Answer	Score	Explanation
-------------	-------	-------------

True ✓ 1.00 Great job, you got it right!

Conclusions for both tests will be equivalent because we know that $t^2 = F$

False

Total	1.00 / 1.00
-------	----------------

Feedback — Quiz: Week Five

[Help Center](#)

You submitted this quiz on **Mon 20 Apr 2015 5:24 AM PDT**. You got a score of **7.00** out of **7.00**.

Question 1

Which of the following is a form of multiple linear regression equation?

Your Answer**Score** **Explanation**

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$ ✓ 1.00 Great job!
This response is correct because there are K different variables

$y = \beta_0 + \beta_1 X_1$

Total 1.00 /
1.00

Question 2

In case of multiple linear regression, the mean value of Y at x_1, \dots, x_k is a linear function of X_1, \dots, X_k .

(please answer True or False below)

**Your
Answer****Score****Explanation**

True ✓ 1.00 Good job!

We know this because

$$\mu_{y/x_1, \dots, x_k} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

False

Total	1.00 /
	1.00

Question 3

Which of the following is not an inherently linear model?

Your Answer	Score	Explanation
<input type="radio"/> $\mu_{y x} = \beta_1 X$		
<input type="radio"/> $\mu_{y x} = \beta_0 e^{\beta_1 X}$		
<input type="radio"/> $\mu_{y x} = \beta_0 + \beta_1 X_1$		
<input checked="" type="radio"/> $\mu_{y x_1,x_2} = e^{\beta_0} + e^{\beta_1 X_1} + e^{\beta_2 X_2}$	✓ 1.00	Great job. This response is correct because it cannot be easily transformed into a linear form.

Total	1.00 /
	1.00

Question 4

The total sum of squares (SST) remains unchanged irrespective of the number of independent variables.

(please answer True or False below)

Your Answer	Score	Explanation
<input checked="" type="radio"/> True	✓ 1.00	Great job! We know this because the SST is the sum of squared deviations about the grand mean.
<input type="radio"/> False		
Total	1.00 /	
	1.00	

Question 5

If the tolerance is 0.0001 then it implies that there is no multicollinearity between the independent variables.

(please answer True or False below)

Your Score Explanation

Answer

True

1.00 Good job!

False

We know this to be false because if the tolerance is less than 0.01, then it indicates multicollinearity. In this case, $0.0001 < 0.01$

Total 1.00 /
 1.00

Question 6

Suppose that $F = 4$ where $F = (\text{MS regression}) / (\text{MS Residual})$. Then R^2 is:

Your Answer Score Explanation

0.64

1.00 Nice work!

We know that $F = R^2 / (1 - R^2)$

So, if we let w represent R^2 , then we can say that $4 = w / (1 - w)$ which we can solve to show that $w = 0.8$

0.2

The given information is
not sufficient to calculate
 R^2

Total	1.00 /
	1.00

Question 7

Suppose the following:

After running some analysis, the result of the partial F-test for a particular variable was: $F = 8.9$ and the result of the t-test for the variable under consideration was: $t\text{-test} = 0.01$.

This statement is:

(please answer True or False below)

Your Answer	Score	Explanation
-------------	-------	-------------

True

False  1.00 Great job!

We know this is false because the partial F-test and its respective t-test are equivalent. $F = t^2$

Total	1.00 /
	1.00

Feedback — Quiz: Week Six

[Help Center](#)

You submitted this quiz on **Tue 28 Apr 2015 11:29 AM PDT**. You got a score of **6.00** out of **6.00**.

Question 1

Which of the following would best be categorized as a dummy variable?

Your Answer	Score	Explanation
<input type="radio"/> Dose of medication (measured in mg)		
<input checked="" type="radio"/> Smoking habits (non-smoker, occasional smoker, frequent smoker)	✓ 1.00	Great job! We know that dummy variables are used to indicate categories of nominal scaled variables, which is the case for smoking habits.
<input type="radio"/> Birth weight (measured in ounces)		
<input type="radio"/> Systolic blood pressure (measured as mmHg)		
Total	1.00 / 1.00	

Question 2

Which of the following is the correct linear regression equation for a model with the predictors **gender** (male/female) and **treatment** (Drug A, Drug B, No Drug), given that all predictors are recoded as dummy variables?

Your Answer	Score	Explanation
<input type="radio"/>		
$E(Y) = \beta_0 + \beta_1(\text{male}) + \beta_2(\text{No Drug}) + \beta_3(\text{Drug A}) + \beta_4(\text{Drug B})$		
<input checked="" type="radio"/> $E(Y) = \beta_0 + \beta_1(\text{male}) + \beta_2(\text{Drug A}) + \beta_3(\text{Drug B})$	✓ 1.00	Nice work!

We know that if a variable contains K categories, then we must define exactly $(k - 1)$ dummy variables to index these categories.

In this situation, only $(2 - 1) = 1$ dummy variable is needed to indicate gender, and $(3 - 1) = 2$ dummy variables are needed to indicate treatment.



$$E(Y) = \beta_0 + \beta_1(female) + \beta_2(male) + \beta_3(No\ Drug) + \beta_4(Drug\ A)$$

Total

1.00 /
1.00

Question 3

If k dummy variables is fit for a nominal variable with k categories in a model containing a constant term, then all the coefficients cannot be uniquely estimated due to collinearity.

(please answer True or False below)

Your Answer	Score	Explanation
-------------	-------	-------------

<input checked="" type="radio"/> True	✓ 1.00	Yes, you answered correctly!
---------------------------------------	--------	------------------------------

The model will not be able to fit due to collinearity between the dummy variables.

False

Total 1.00 /
1.00

Question 4

In a model with dependent variable Y and predictors X & Z , we can assume that there is no interaction if the relationship between X & Z is independent of Y .

(please answer True or False below)

Your Answer	Score	Explanation
-------------	-------	-------------

True

False ✓ 1.00 Great job!

We can, however, assume there is no interaction if the relationship between X & Y is independent of Z

Total 1.00 /
1.00

Question 5

For this question, consider the following regression which is used to compare two separate straight line regressions using the single regression model method:

$$\mu_{y|xz} = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

With regard to the above regression, assuming that β_3 significantly contributes to the model, we can interpret β_2 as the unit change in Y given one unit increase in Z .

(please answer True or False below)

Your Answer	Score	Explanation
<input type="radio"/> True		
<input checked="" type="radio"/> False	1.00	Great job!
		This is false because the β_3 will also contribute to the change in Y when interaction is present.
Total	1.00 / 1.00	

Question 6

For this question, consider the following regression which is used to compare two separate straight line regressions using the single regression model method:

$$\mu_{y|xz} = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

Suppose $\beta_3 = 0$. Which of the following must **always** be true:

Select all that apply

Your Answer	Score	Explanation
<input type="checkbox"/> The two lines are coincident	0.25	The intercept of the the two regression lines are generally not equal and hence are generally not coincident .
<input checked="" type="checkbox"/> The regression lines will be parallel	0.25	$\beta_3 = 0$ means that the interaction term is not significant and so the two regression lines will be parallel .
<input type="checkbox"/> The intercept is equal for X & Y	0.25	The intercept of the the two regression lines are generally not equal and hence are generally not coincident .
<input checked="" type="checkbox"/> The slopes will be equal for X & Y	0.25	Because the two regression lines are parallel, then their slopes are equal .
Total	1.00 / 1.00	

Homework Assignments

Applied Regression Analysis

WEEK 2

The table gives the systolic blood pressure (SBP), body size (QUET), age (AGE), and smoking history (SMK = 0 if nonsmoker, SMK = 1 if a current or previous smoker) for a hypothetical sample of 32 white males over 40 years old from the town of Angina.

Person	SBP	QUET	AGE	SMK
1	135	2.876	45	0
2	122	3.251	41	0
3	130	3.100	49	0
4	148	3.768	52	0
5	146	2.979	54	1
6	129	2.790	47	1
7	162	3.668	60	1
8	160	3.612	48	1
9	144	2.368	44	1
10	180	4.637	64	1
11	166	3.877	59	1
12	138	4.032	51	1
13	152	4.116	64	0
14	138	3.673	56	0
15	140	3.562	54	1
16	134	2.998	50	1
17	145	3.360	49	1
18	142	3.024	46	1
19	135	3.171	57	0
20	142	3.401	56	0
21	150	3.628	56	1
22	144	3.751	58	0
23	137	3.296	53	0
24	132	3.210	50	0
25	149	3.301	54	1
26	132	3.017	48	1
27	120	2.789	43	0
28	126	2.956	43	1
29	161	3.800	63	0
30	170	4.132	63	1
31	152	3.962	62	0
32	164	4.010	65	0

Exercise One

Generate scatter diagrams for each of the following variable pairs:

1. SPB (Y) vs. QUET (X)
2. SBP (Y) vs. SMK (X)
3. QUET (Y) vs. AGE (X)
4. SBP (Y) vs. AGE (X)

Exercise Two

Sketch a line

Using scatter diagrams #1, #3, and #4 that you generated above, use paper and pencil to roughly sketch a line that fits the data reasonably well. Use the [homework forum](#) to share your sketches and comment on the relationships described.

Exercise Three

Comparing Blood Pressure with Smoking History

1. Determine the least-squares estimates of slope (β_1) and intercept (β_0) for the straight-line regression of SBP (Y) on SMK (X).
2. Compare the value of $\hat{\beta}_0$ with the mean SBP for nonsmokers. Compare the value of $\hat{\beta}_0 + \hat{\beta}_1$ with the mean SBP for smokers. Explain the results of these comparisons.
3. Test the hypothesis that the true slope (β_1) is 0.
4. Is the test in part (e) equivalent to the usual two-sample t test for the equality of two population means assuming equal but unknown variances? Demonstrate your answer numerically.

Exercise Four

Comparing Blood Pressure with Body Size

1. Determine the least-squares estimates of slope and intercept for the straight-line regression of SBP (Y) on QUET (X).
2. Sketch the estimated regression line on the scatter diagram involving SBP and QUET.
3. Test the hypothesis of zero slope.
4. Find a 95% confidence interval for $\mu_{y|\bar{x}}$.
5. Calculate 95% prediction bands.
6. Based on the above, would you conclude that blood pressure increases as body size increases?
7. Are any of the assumptions for straight-line regression clearly not satisfied in this example?

Week Two Homework Solutions

[Help Center](#)[Homework Central](#) / [Week Two Homework](#) / Week Two Solutions

If you need help and answers to the exercises of week two, please click below on the selected exercise:

- [Exercise One](#)
- [Exercise Two](#)
- [Exercise Three](#)
- [Exercise Four](#)

Click the button below to return to this week's homework exercise.

[←Homework Page](#)

Created Tue 10 Mar 2015 8:53 AM PDT

Last Modified Mon 30 Mar 2015 10:03 AM PDT

Homework Solutions

Applied Regression Analysis

WEEK 2

Exercise One

Generate scatter diagrams for each of the following variable pairs:

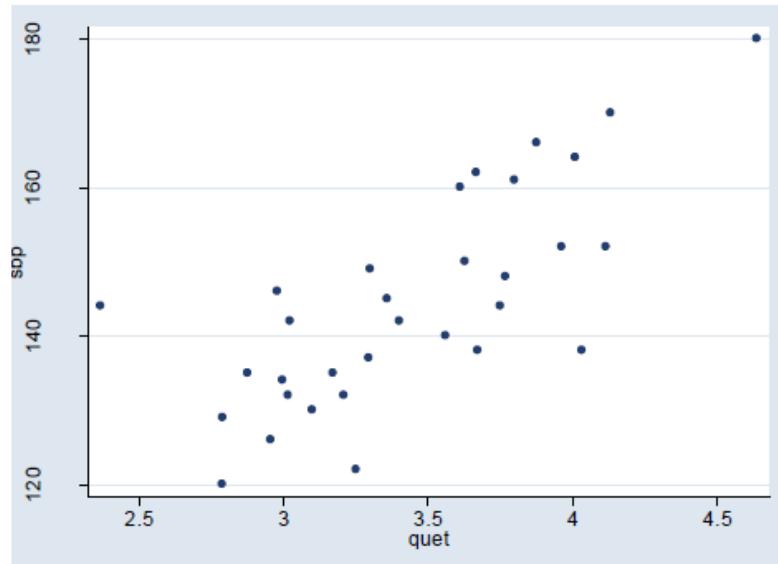
1. SBP (Y) vs. QUET (X)

In the command window, enter ‘.scatter sbp quet’

This will produce the scatterplot of SBP (Y) and QUET (X). The resulting scatterplot displayed should resemble the screenshot depicted below

(Note: Stata is case-sensitive so the variables should be lower-case if that is how they are coded in the dataset).

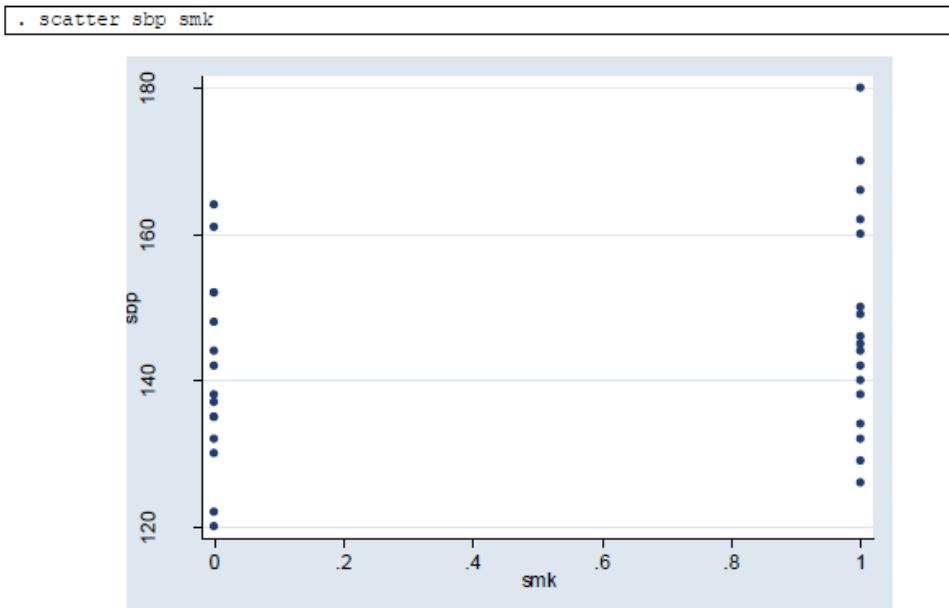
```
.scatter sbp quet
```



2. SBP (Y) vs. SMK (X)

In the command window, enter ‘.scatter sbp smk’.

This will produce the scatterplot of SBP (Y) and SMK (X). The resulting scatterplot displayed should resemble the screenshot depicted below.



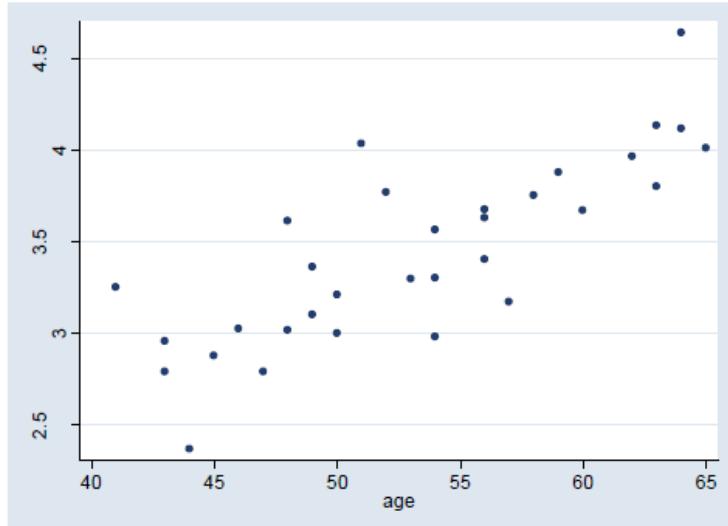
Note: Smoking is a binary variable, therefore we should not expect to see an even scatter of observations as with the other plots containing continuous variables.

3. QUET (Y) vs. AGE (X)

In the command window, enter ‘.scatter quet age’.

This will produce the scatterplot of QUET (Y) and AGE (X). The resulting scatterplot displayed should resemble the screenshot depicted below.

```
. scatter quet age
```

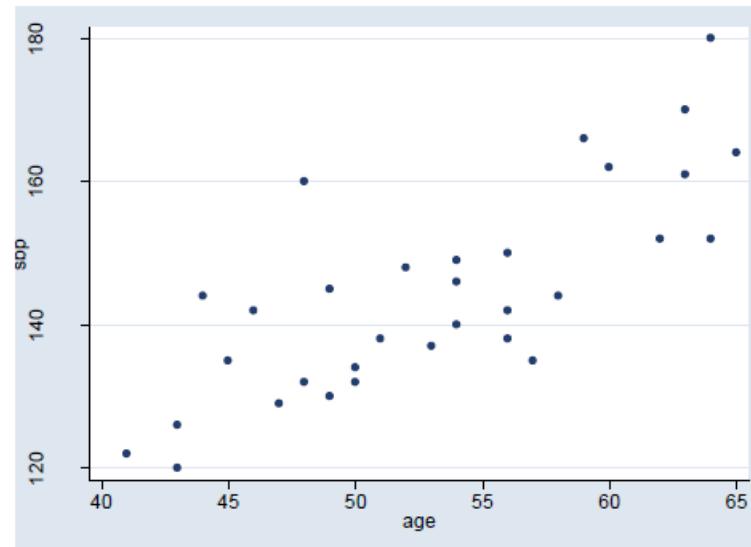


4. SBP (Y) vs. AGE (X)

In the command window, enter ‘.scatter sbp age’.

This will produce the scatterplot of SBP (Y) and AGE (X). The resulting scatterplot displayed should resemble the screenshot depicted below.

```
. scatter sbp age
```



Homework Solutions

Applied Regression Analysis

WEEK 2

Exercise Two

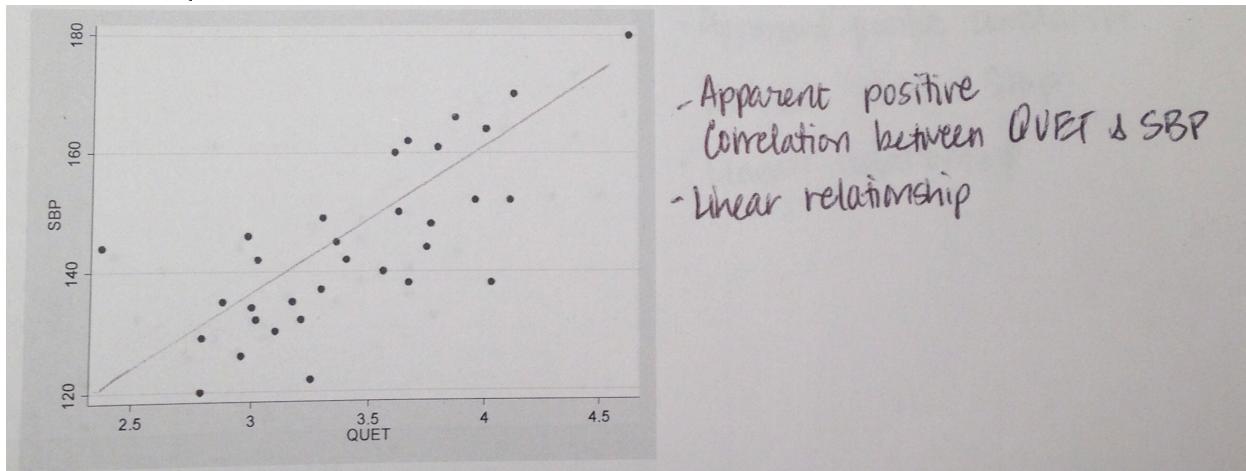
Sketch a line

Using scatter diagrams #1, #3, and #4 that you generated in exercise one, use paper and pencil to roughly sketch a line that fits the data reasonably well.

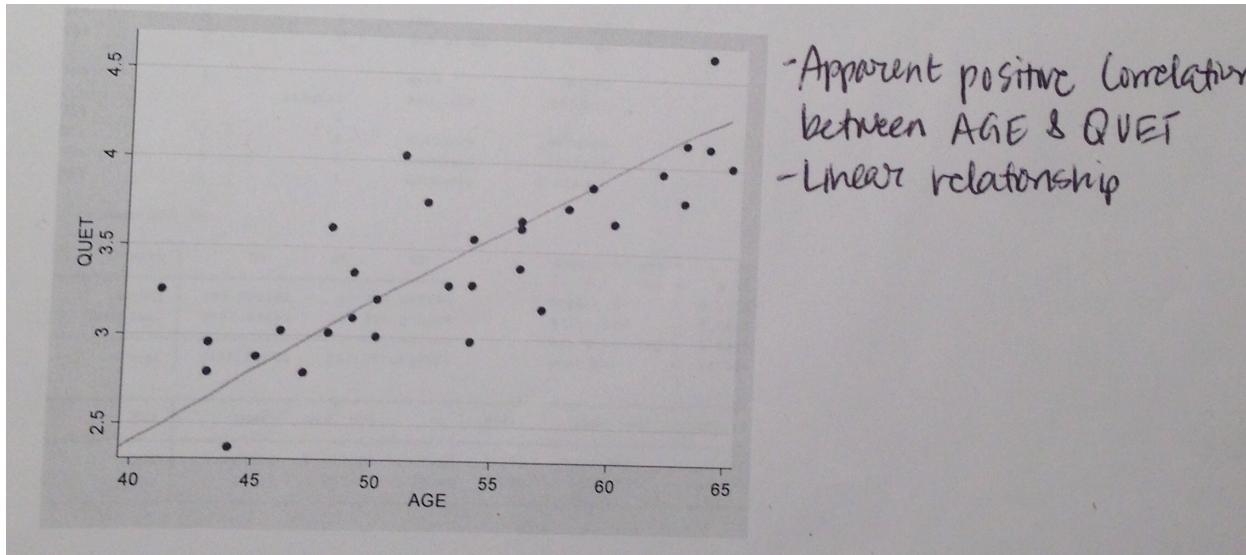
Use the **homework forum** to share your sketches and comment on the relationships described.

Below are examples of how this could look when you hand-draw a line.

#1 – SBP vs. QUET



#3 – QUET vs. AGE



#4 – SBP vs. AGE



Homework Solutions

Applied Regression Analysis

Exercise Three

Comparing Blood Pressure with Smoking History

1. Determine the least-squares estimates of slope (β_1) and intercept (β_0) for the straight-line regression of SBP (Y) on SMK (X).

We can determine the least squares estimates for the parameters in simple linear regression by regressing Y on X. In the command window, enter '.regress sbp smk'. This will produce the output below.

. regress sbp smk					
Source	SS	df	MS	Number of obs = 32	
Model	393.098162	1	393.098162	F(1, 30)	= 1.95
Residual	6032.87059	30	201.095686	Prob > F	= 0.1723
Total	6425.96875	31	207.289315	R-squared	= 0.0612
				Adj R-squared	= 0.0299
				Root MSE	= 14.181
sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
smk	7.023529	5.023498	1.398	0.172	-3.235823 17.28288
_cons	140.8	3.661472	38.454	0.000	133.3223 148.2777

$$\begin{aligned}y &= \beta_0 + \beta_1 x \\&= 140.8 + 7.02(\text{smk})\end{aligned}$$

Note: When entering data into STATA, always list the dependant variable first (sbp, in this case) and then the independent variable (smk, in this case).

The slope for smk is determined by the value listed in the "Coef." column of the table above, and the value for the intercept is determined in a similar fashion in the _cons row.

How can you interpret the slope?

"Under this model, current or previous smokers have an average systolic blood pressure 7.02 mm Hg higher than that of non-smokers"

2. Compare the value of $\hat{\beta}_0$ with the mean SBP for nonsmokers. Compare the value of $\hat{\beta}_0 + \hat{\beta}_1$ with the mean SBP for smokers. Explain the results of these comparisons.

To compare the mean of one variable across different categories of another variable in STATA, you must first sort the data by the second categorizing variable (in this case, smk).

In the command window, enter ‘.sort smk’.

You must then use the ‘sum’ command to get descriptive statistics on your variable of interest (in this case, sbp), but first you must use the ‘by’ command to split the results by smoking status.

In the command window, enter ‘.by smk:sum sbp’.

This will produce the output below.

. sort smk
. by smk:sum sbp
 -> smk = 0
Variable Obs Mean Std. Dev. Min Max
-----+-----
sbp 15 140.8 12.90183 120 164
 -> smk = 1
Variable Obs Mean Std. Dev. Min Max
-----+-----
sbp 17 147.8235 15.21198 126 180

The mean value of SBP for nonsmokers (140.8) is equal to the value of $\hat{\beta}_0$. The mean value of SBP for smokers is 147.82 which is equal to $\hat{\beta}_0 + \hat{\beta}_1$.

In simple linear regression, the intercept can be interpreted as the value for Y when X=0. Given that smoking is a binary variable, and is coded (0,1) (0 for non-smokers, 1 for smokers), then the intercept is the mean value for non-smokers (Y when X=0), and the intercept plus the slope is the mean value for smokers (Y when X=1).

3. Test the hypothesis that the true slope (β_1) is 0.

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

The null hypothesis cannot be rejected, $p = 0.172$. There is not sufficient evidence to conclude that the slope is significantly different from 0.

Note: You can test for the significance of the slope by looking at the p-value for the t-test in the table for the regression in problem 1. The p-value tells us that the probability of rejecting the null when the null is true is 17.2%, which exceeds 5%. Therefore there is insufficient evidence to reject the null.

4. Is the test in part (3) equivalent to the usual two-sample t test for the equality of two population means assuming equal but unknown variances? Demonstrate your answer numerically.

To perform a t-test to compare the mean sbp across the populations in the different smoking categories, enter ‘.ttest sbp, by(smk)’ into the command window.

This will produce the output below.

```
. ttest sbp, by(smk)

Two-sample t test with equal variances

-----+-----+-----+-----+-----+
      Group |     Obs        Mean    Std. Err.    Std. Dev.   [95% Conf. Interval]
-----+-----+-----+-----+-----+
          0 |     15      140.8    3.331237    12.90183    133.6552    147.9448
          1 |     17      147.8235   3.689448    15.21198    140.0022    155.6448
-----+-----+-----+-----+-----+
```

```
combined |     32      144.5313    2.545151    14.39755    139.3404    149.7221
-----+-----+-----+-----+-----+
      diff |      -7.023529    5.023498           -17.28288    3.235823
-----+-----+
Degrees of freedom: 30

Ho: mean(0) = mean(1) = diff = 0

Ha: diff < 0           Ha: diff ~= 0           Ha: diff > 0
      t =    -1.3981       t =    -1.3981       t =    -1.3981
      P < t =    0.0862       P > |t| =    0.1723       P > t =    0.9138
```

The t-test gives the same t-value and p-value as the test for the hypothesis that the true slope, β_1 , is 0.

The p-value for question 3 is the same at the two-sided p-value for the two-sample t-test. In both tests, you are testing whether smoking has a significant impact on systolic blood pressure by determining if the sbp for smokers is significantly different than that of non-smokers.

Homework Solutions

Applied Regression Analysis

WEEK 2

Exercise Four

1. Determine the least-squares estimates of slope and intercept for the straight-line regression of SBP (Y) on QUET (X).

We can determine the least squares estimates for the parameters in simple linear regression by regressing Y on X.

In the command window, enter ‘.regress sbp quet’.

This will produce the output below.

. regress sbp quet					
Source	SS	df	MS		
Model	3537.94585	1	3537.94585	Number of obs = 32	
Residual	2888.0229	30	96.2674299	F(1, 30) = 36.75	
Total	6425.96875	31	207.289315	Prob > F = 0.0000	
sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
quet	21.49167	3.545147	6.062	0.000	14.25151 28.73182
_cons	70.57641	12.32187	5.728	0.000	45.4118 95.74102

$$\hat{\beta}_0 = 70.576$$

$$\hat{\beta}_1 = 21.492$$

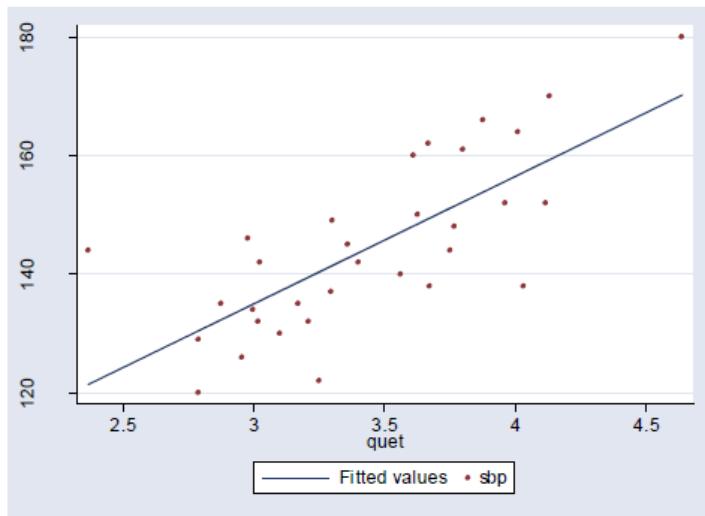
2. Sketch the estimated regression line on the scatter diagram involving SBP and QUET.

In order to fit a regression line in STATA, you must first create a new variable in your dataset for of the predicted y given x under the regression model.

You can do this simply by entering ‘predict yhat’ into the command window. Next, create a scatterplot with a line by entering ‘scatter yhat sbp quet, c(1 .) s(i o)’ into the command window.

The commands ‘c(1 .) s(i o)’ specify that the yhat should be labeled with a line and data points with dots, respectively.

```
. predict yhat  
(option xb assumed; fitted values)  
  
. scatter yhat sbp quet, c(1 .) s(i o)
```



3. Test the hypothesis of zero slope.

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

Reject the null hypothesis, $p < 0.001$. There is sufficient evidence to conclude that the slope is significantly different from 0.

Note: You can test for the significance of the slope by looking at the p-value for the t-test in the table for the regression in problem 1. The p-value tells us that the probability of rejecting the null when the null is true is less than 5%. Therefore there is sufficient evidence to reject the null.

4. Find a 95% confidence interval for $\mu_{y|\bar{x}}$.

To calculate confidence intervals, you need to know the descriptive statistics for the variables, including their mean values and standard deviations.

To get these values, use the ‘sum’ command by entering ‘.sum sbp quet age smk’ into the command window.

Next we can calculate $\mu_{y|\bar{x}}$ by entering the mean value for quet within the regression equation using our previously estimated parameters. The confidence limits about $\mu_{y|\bar{x}}$ can then be estimated using the mean value and standard deviation of x.

. sum sbp quet age smk					
Variable	Obs	Mean	Std. Dev.	Min	Max
sbp	32	144.5313	14.39755	120	180
quet	32	3.441094	.4970781	2.368	4.637
age	32	53.25	6.956083	41	65
smk	32	.53125	.5070073	0	1

$$\hat{y}_{\bar{x}} = 70.57641 + 21.49167 * 3.44 = 144.508$$

$$s_{\hat{y}_{x_0}}^2 = s_{\hat{y}_{\bar{x}}}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2} \right)$$

$$s_{\hat{y}_{\bar{x}}}^2 = s_{\hat{y}_{\bar{x}}}^2 \left(\frac{1}{n} \right) = \frac{96.2674299}{32} = 3.008357$$

$$s_{\hat{y}_{\bar{x}}} = \sqrt{3.008357} = 1.7344616$$

$$95\% \text{ CI: } \hat{y}_{\bar{x}} \pm t_{975}(30) s_{\hat{y}_{\bar{x}}} = 144.508 \pm 2.042 \times 1.7344616 = (140.97, 148.05)$$

Interpretation: We are 95% confident that the true value for the mean value of y is between 140.97 and 148.05 mm Hg.

5. Calculate 95% prediction bands.

For this problem, we are asking for a plot of the prediction bands using STATA, not for hand-calculations. To do this, we must enter 'predict sepred, stdf' into the command window to generate a variable- 'sepred'- for the standard deviation used within the prediction interval.

Next, we can calculate the value for the lower limit of the prediction interval by entering 'generate low=yhat-invtail(30,0.025)*sepred' and the upper limit of the prediction interval by entering 'generate high=yhat+invtail(30,0.025)*sepred' (note: invtail(30,0.025)= $t_{0.975}(30)$).

From here you can create a plot of the prediction intervals with the regression line by entering 'scatter sbp yhat low high, sort connect (. 1 1 1) symbol (o i i i)'. The code and plot below includes both the confidence and prediction intervals, however you only need to graph the prediction intervals for this question.

```
. predict yhat
(option xb assumed; fitted values)

. predict seyhat, stdp

. display invtail(30,0.025)
2.0422724

. generate lowl= yhat-2.0422724* seyhat

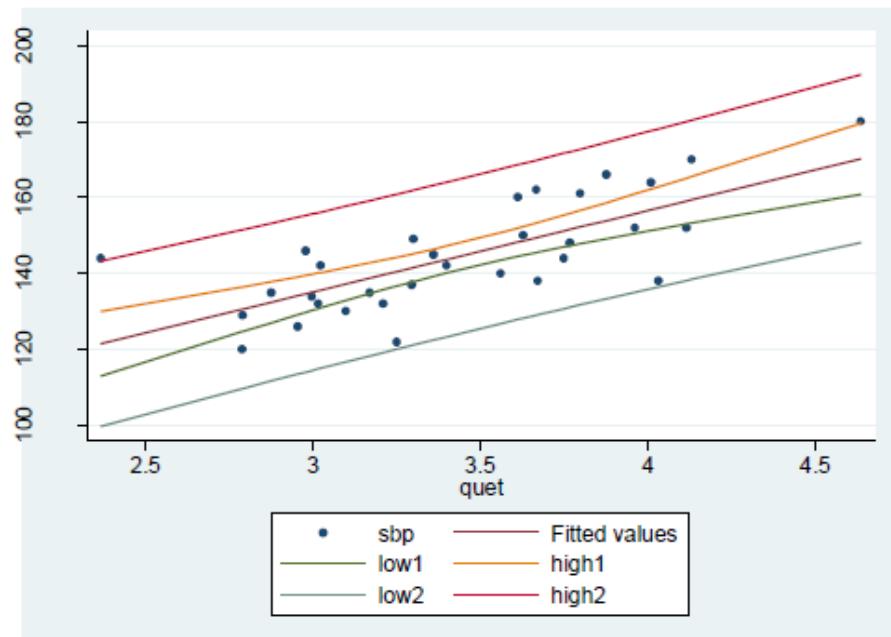
. generate highl= yhat+2.0422724* seyhat

. predict sepred, stdf

. generate low2= yhat-invtail(30,0.025)* sepred

. generate high2= yhat+invtail(30,0.025)* sepred

. scatter sbp yhat lowl highl low2 high2 quet,sort connect(. 1 1 1 1 1)
symbol(o i i i i)
```



- 6. Based on the above, would you conclude that blood pressure increases as body size increases?**

Yes, because the fitted regression line, as well as the confidence and prediction band, appear to have an upward slope.

- 7. Are any of the assumptions for straight-line regression clearly not satisfied in this example?**

Simple Linear Regression Assumptions:

Linearity: SBP and SMK appear to be linearly related based on the above scatterplot

Independence: The study design does not suggest that the observations are not independent

Normality: The variables appear to be normally distributed (there are no significant outliers)

Equal Variance (homoscedasticity): The variances along the regression line appear to remain similar as you move across the line

There are no apparent violations of homoscedasticity, normality, or independence. Formal tests of these assumptions are possible but are not included here.

All results for: Week2 Homework

[Week2 Homework](#)[Search All Forums](#)

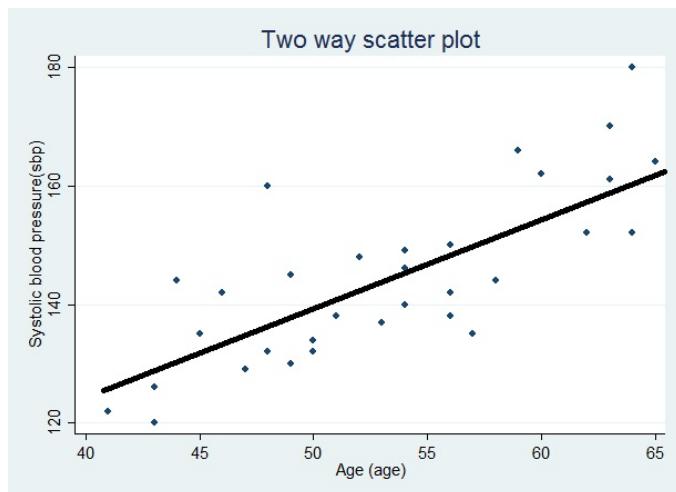
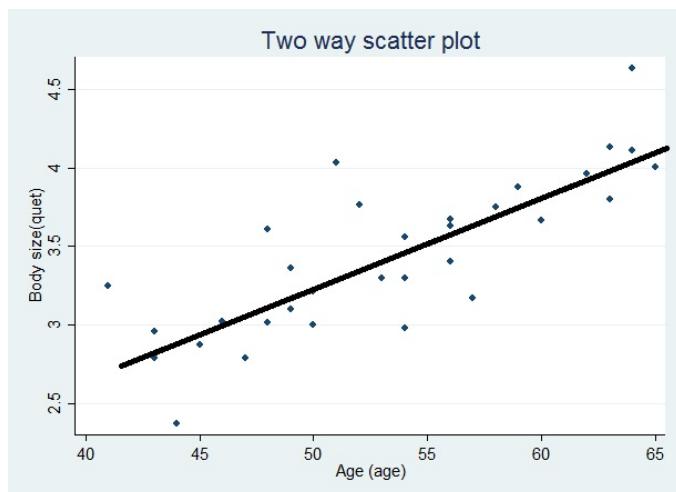
Your search returned approximately **41** results

Ganesh Sharma · 10 days ago 

Posted in [Week 2 homework exercise two](#)

Week 2 homework exercise two

I used microsoft paint to draw lines in the scatter plot ...

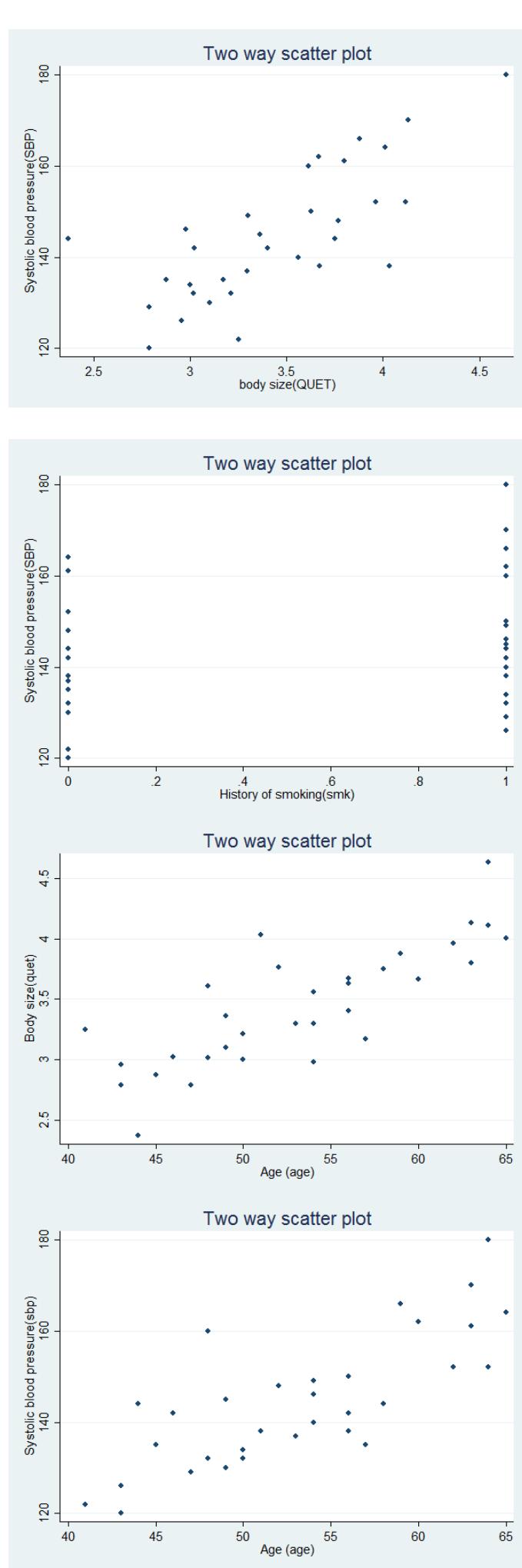


Ganesh Sharma · 10 days ago 

Posted in [Week 2 homework exercise one...](#)

Dear all,

Week 2 homework exercise one..



David Alejandro Ramirez Palacios · 13 days ago

Posted in Week 2 HW Exercise 4

Hi, everyone.

I need a help with excercise 4 because I got a diffent answer in the excercise 4 that Homework solutions. The slope that I calculated was 0.0214917 , but Homework solutions have 21.4917.

Thanks

The screenshot shows the Stata 13.1 Results window. The command history includes:

```
# Command
1 import excel "C:/Users/User...".xlsm
2 TwoWay (scatter SBP QUET) 199
3 twoway
4 twoway (scatter SBP QUET)
5 graph twoway (scatter SBP AGE)
6 graph twoway (scatter SBP SMK)
7 twoway (scatter SBP SMK)
8 graph save Graph "C:/Users...
9 graph save Graph "C:/Users...
10 twoway (scatter QUET AG...
11 graph save Graph "C:/Users...
12 twoway (scatter SBP AGE)
13 graph save Graph "C:/Users...
14 sum SBP SMK
15 sum SBP AGE
16 reg SBP SMK
17 twoway (scatter SBP SMK)
18 sort SMK
19 by SMK: sum SBP
20 reg SBP QUET
```

Summary statistics for SBP:

Variable	Obs	Mean	Std. Dev.	Min	Max
SBP	19	140.8	12.90183	120	164

Summary statistics for SBP:

Variable	Obs	Mean	Std. Dev.	Min	Max
SBP	17	147.8235	15.21198	126	180

Regression output for reg SBP QUET:

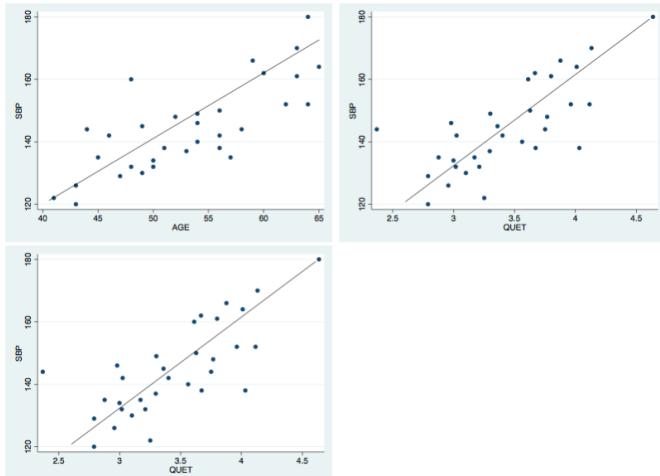
Source	SS	df	MS	
Model	3537.94574	1	3537.94574	Number of obs = 32
Residual	2888.02301	30	96.2674337	F(1, 30) = 36.79
Total	6425.96875	31	207.89315	Prob > F = 0.0000

Output continues with R-squared, Adj R-squared, Root MSE, and the coefficient table:

SBP	Coeff.	Std. Err.	t	P> t	(95% Conf. Interval)
QUET	0.0214917	.0035451	6.06	0.000	.0142515 .0287318
_cons	70.5764	12.30287	-5.73	0.000	45.41179 95.74101

David C. Morris · 11 days ago

Posted in Week Two Homework - Exercise 1



Positive and linear relationships.

Thomas Evans STAFF · 7 days ago

Posted in Different data for lectures & homework?

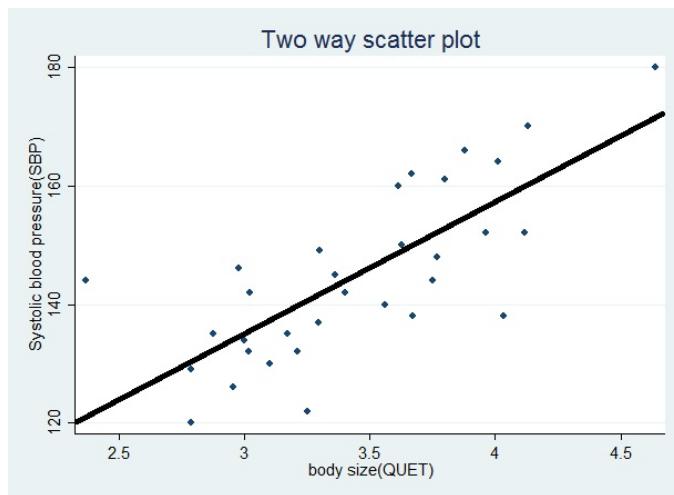
The TAs and professor can correct me if I'm wrong, but I believe that the dataset used in the homework is a bit different from the dataset demonstrated on video 2.1.

The video was demonstrating the process and then the homework had you practice the process in a similar style problem with different numbers. The differing data would most likely produce different plots.

Professor Lemeshow explains a bit about the week 2 homework in this [homework highlights video](#).

Ganesh Sharma · 10 days ago

Posted in Week 2 homework exercise two



Thomas Evans STAFF · 14 days ago

Posted in homework data

Hmm... weird... I was able to copy/paste from Chrome to Excel (Mac) and it transferred properly. Regardless, I've made a CSV of the data and posted it as a download on the [Week Two Homework](#) page.

Hope this helps.

Cheers,

Tom

Ganesh Sharma · 7 days ago

Posted in [Week 2 homework exercise four.](#)

Question 1. · Determine the least-squares estimates of slope and intercept for the straight-line regression of SBP (Y) on QUET (X).

Intercept(B_0) = 70.57641

Slope(B_1) = 21.49167

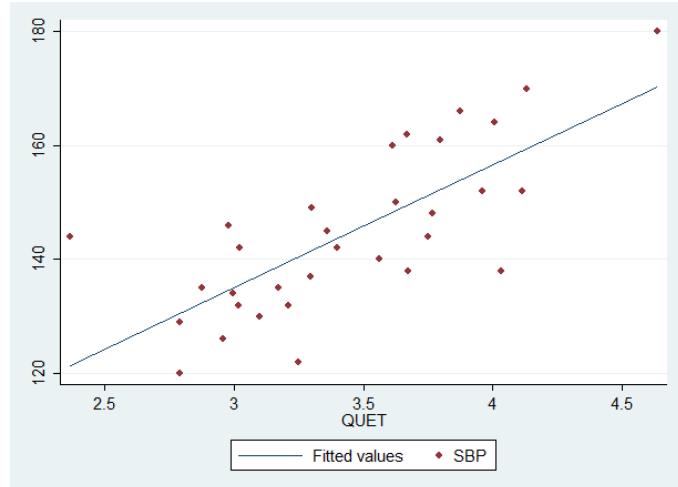
regress sbp quet

Source	SS	df	MS	Number of obs	=	32
Model	3537.94585	1	3537.94585	F(1, 30)	=	36.75
Residual	2888.0229	30	96.2674299	Prob > F	=	0.0000
Total	6425.96875	31	207.289315	Adj R-squared	=	0.5356

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
quet	21.49167	3.545147	6.06	0.000	14.25151 28.73182
_cons	70.57641	12.32187	5.73	0.000	45.4118 95.74102

Question 2

Sketch the estimated regression line on the scatter diagram involving SBP and QUET.



Question 3

We can reject the null hypothesis and accept the alternative hypothesis because p value is less than .05(i.e .000). It means there is significant relationship between body size and systolic blood pressure.

Question 4

Find a 95% confidence interval for $\mu_y|x^-$

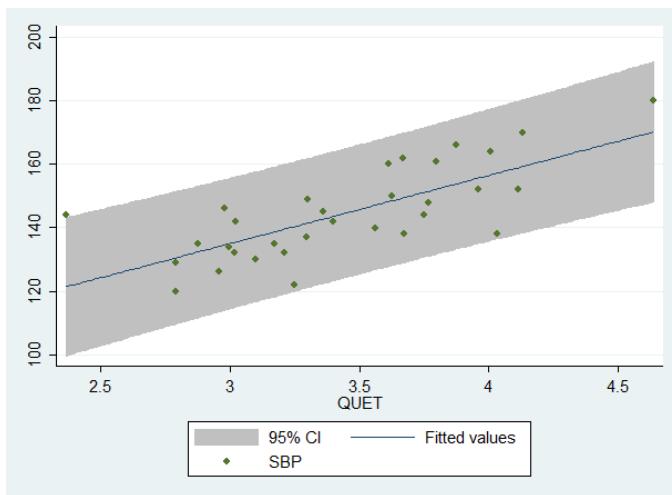
The confidence level for the equation is 140.5073, 148.5552. We are 95% sure that the true value lies between 140.5073, 148.5552.

ci sbp quet age smk yhat

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
sbp	32	144.5313	.545151	139.3404	149.7221
quet	32	3.441094	.0878718	3.261878	3.62031
age	32	53.25	1.229673	50.74206	55.75794
smk	32	.53125	.0896271	.3484544	.7140456
yhat	32	144.5312	1.973011	140.5073	148.5552

Question 5

Calculate 95% prediction bands



Question 6

Based on the above, would you conclude that blood pressure increases as body size increases

We are 95% confident that the systolic blood pressure (sbp) goes up when the body size(quet) increases. The regression including the shaded area goes upward right giving evidence to support the statement.

Question 7. Are any of the assumptions for straight-line regression clearly not satisfied in this example

There are a lot of assumptions that still are not satisfied by the example prior to doing regression analysis. But in my previous submission exercise I mentioned quiet few of them. The assumptions that need to fulfill are independence, normality, linearity and homoeadasticity. Besides that multi collinearity test is also required to skip insignificant variable/s in an attempt to find the causing factor limited to these datasets.

Ganesh Sharma · 8 days ago

Posted in Week 2 homework exercise three

Question 1. Determine the least-squares estimates of slope β_1 and intercept β_0 for the straight-line regression of SBP (Y) on SMK (X).

The histogram check showed the SBP has almost in normal distribution pattern. Besides sapiro-wilk test and sapiro francia showed $p>.05$, indicating normality. Variance was same for SBP for any SMK(homoscedasticity). So, regression is applicable for the situation.

$$\text{Intercept } (\beta_0) = 140.8$$

Slope (β_1) = 7.023529

regress sbp smk

Source	SS	df	MS	Number of obs = 32		
Model	393.098162	1	393.098162	$F(1, 30) = 1.95$		
Residual	6032.87059	30	201.095686	Prob > F = 0.1723		
Total	6425.96875	31	207.289315	R-squared = 0.0612		
				Adj R-squared = 0.0299		
				Root MSE = 14.181		
<hr/>						
sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
smk	7.023529	5.023498	1.40	0.172	-3.235823	17.28288
_cons	140.8	3.661472	38.45	0.000	133.3223	148.2777

There is no significant difference(confidence interval – 3.235823, 17.28288) in SBP between those with and without smoking history.

Question 2.

Question 2. Compare the value of β^0 with the mean SBP for nonsmokers. Compare the value of $\beta^0 + \beta^1$ with the mean SBP for smokers.

```
. sort smk
by smk: sum sbp
```

-> smk = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
sbp	15	140.8	12.90183	120	164

-> smk = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
sbp	17	147.8235	15.21198	126	180

The mean of non-smokers (coded as 0) is 140.8 as against smokers (coded as 1) is 147.8235. As non-smoker is coded as 0, it becomes base category, as intercept(because multiplying the coefficient with 0 becomes 0) and non-smoker includes intercept and slope(1)

While fitting the value of smoking history in straight line equation

i) If smoking = 0

$$Y = 140.8 + 7.023529 * x = 140.8 + 7.023529 * 0 = 140.8$$

ii) If smoking = 1

$$Y = 140.8 + 7.023529*x = 140.8 + 7.023529 * 1 = 147.8$$

Question 3. Test the hypothesis that the true slope B1 is 0 .

Null hypothesis $B1 = 0$ (there is no linear relationship between SBP(Y) and SMK(X))

Alternate hypothesis $B2 \neq 0$ (There is linear relationship between SBP(Y) and SMK(X))

There is no linear relationship ($p = 0.172$) between SBP(Y) and SMK(X)). I think this is non-directional (two sided) hypothesis checking in both positive and negative linear relationship.

Question 4. Is the test in question (3) equivalent to the usual two-sample t test for the equality of two population means assuming equal but unknown variances?

ttest sbp, by (smk)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
0	15	140.8	3.331237	12.90183	133.6552 147.9448
1	17	147.8235	3.689448	15.21198	140.0022 155.6448
combined	32	144.5313	2.545151	14.39755	139.3404 149.7221
diff		-7.023529	5.023498		-17.28288 3.235823
diff = mean(0) - mean(1)			t = -1.3981		
Ho: diff = 0		degrees of freedom = 30			
Ha: diff < 0		Ha: diff != 0		Ha: diff > 0	
Pr(T < t) = 0.0862		Pr(T > t) = 0.1723		Pr(T > t) = 0.9138	

think it is equivalent. Because, the hypothesis mentioned in question 3 is linear relationship. Linear relationship generated numerical values like 140.8 and 147.8.

In independent (two sample) t-test, the hypothesis is expressed like; There is no difference (means) in SBP between smoker and non-smokers (NULL) and there is significant difference as Alternate hypothesis.

Test in question (3) has one categorical and another numerical variable and that is the same for T-test.

Only the output formats are different. Regression covers more with % of variability explained by independent variable in dependent variable and shows more about residual sum of squares and model sum of squares. For t-test, it calculates mean difference between two groups.

Two sample test with equal variance

$$\Pr(|T| > |t|) = 0.1723$$

Mario R. Melchiori · 13 days ago  Posted in [How to import data relavent to this course in Stata](#)

Hi,

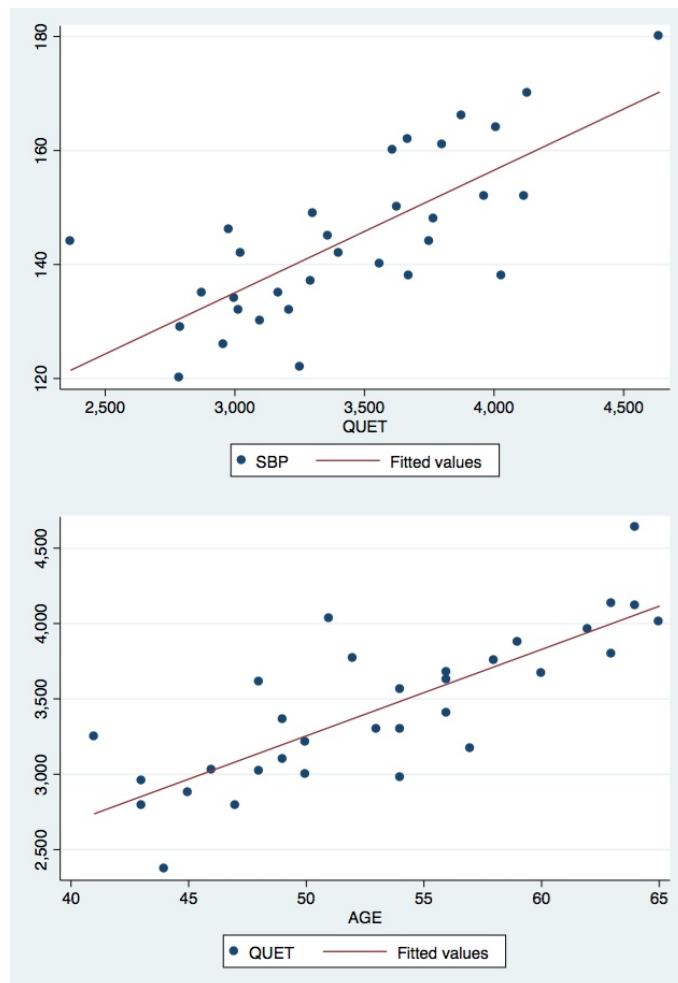
On the Quiz section, you will find out the data. Below, the link:

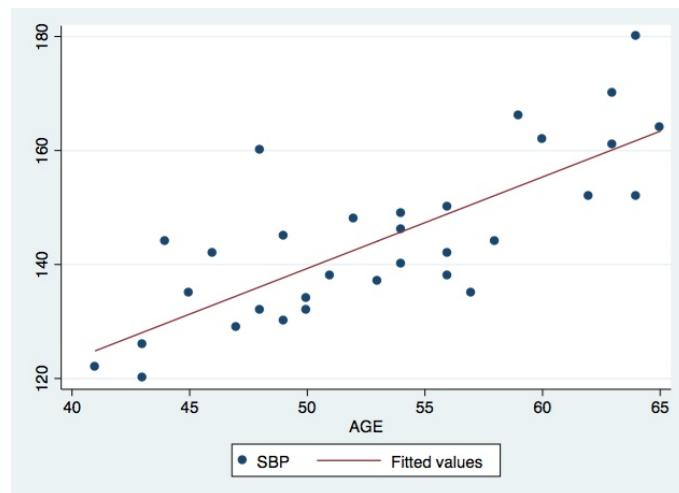
<https://d396qusza40orc.cloudfront.net/appliedregression/Homeworks/week2-HW-data.csv>

I hope it helps.

Daniel Rojas · 5 days ago 

Posted in [Homework-Week Two-Exercise Two](#)





Pegando porte y la vara con la regresion lineal

Costa Rica representing

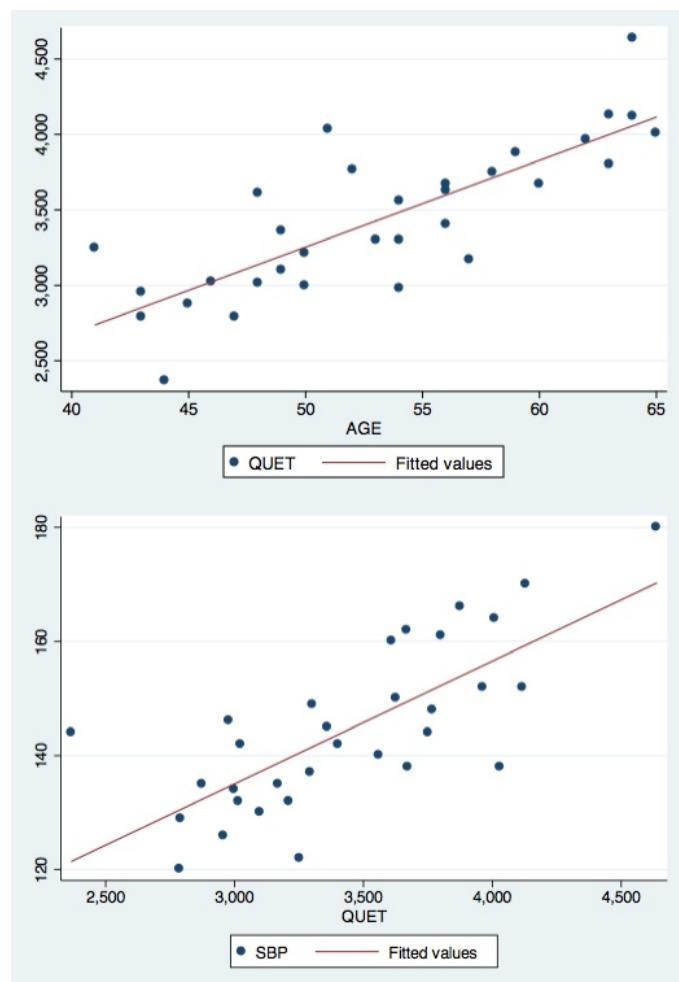
There is a lineal relationship in all these cases

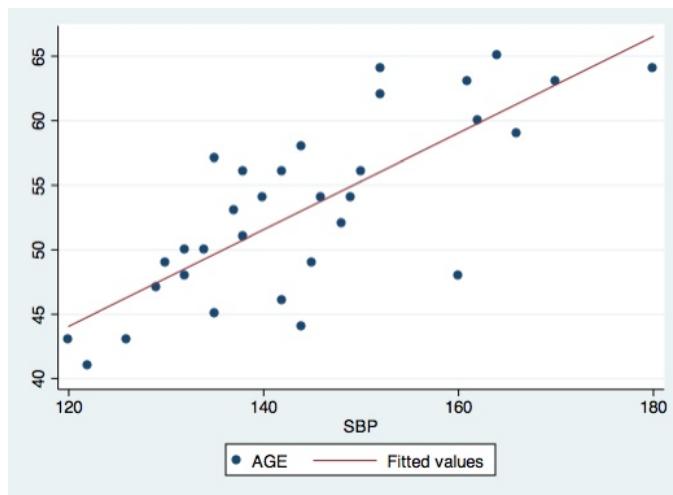
Susana Acosta Gonzalez · 5 days ago

Posted in Week2 HW Exercise 2

A bit late but here's my results:

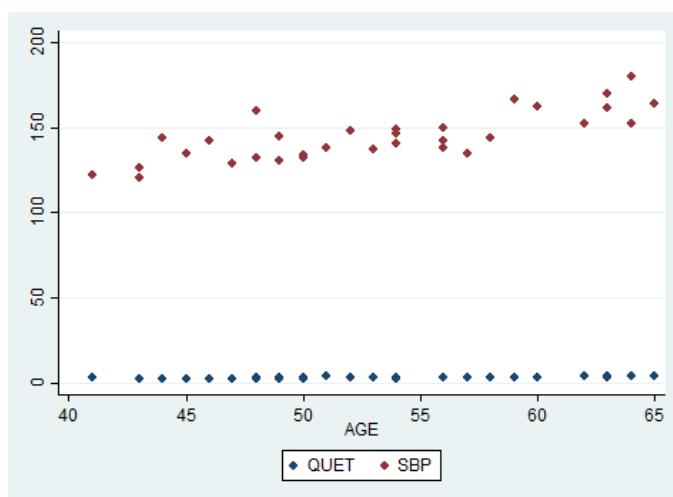
All of the graphs show a positive linear relationship between the variables.





Hagai Levine · 13 days ago

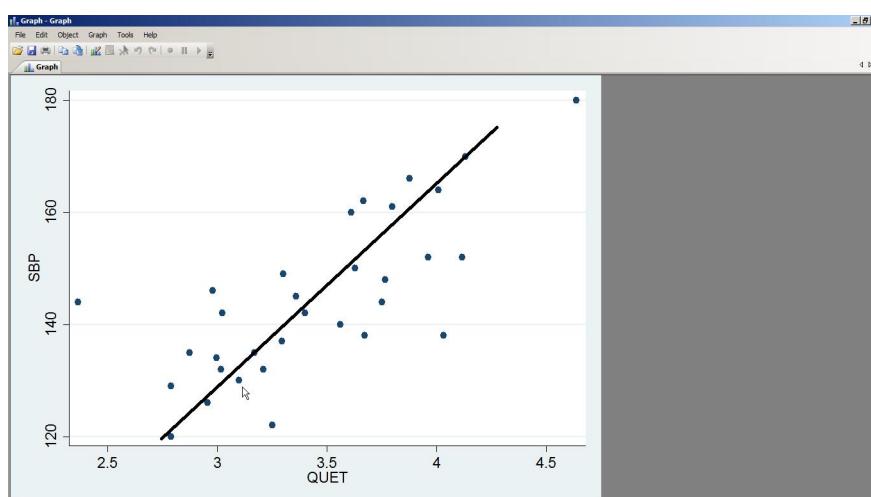
Posted in Week Two Homework - Exercise 1

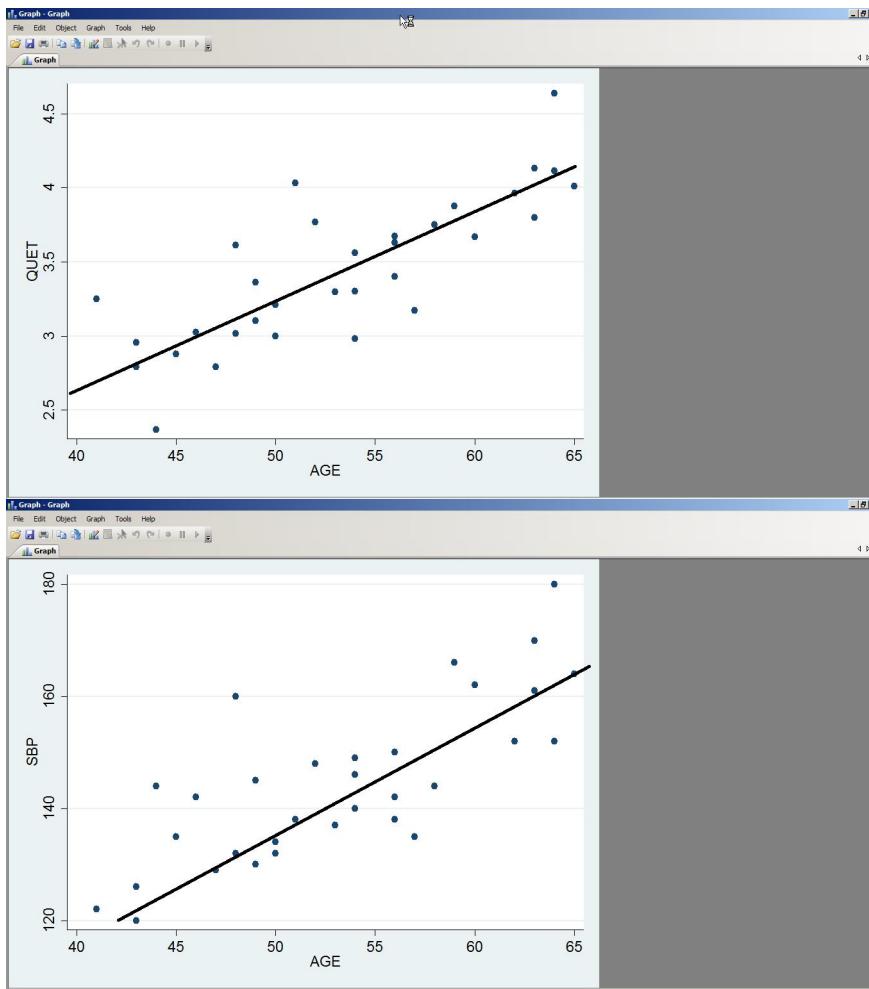


It's nice to see how scaling is important - see what happens if we show sbp and quet in the same graph

Gari de Jesus Jerez Aguero · 3 days ago

Posted in Week Two Homework - Exercise 1





Oscar Granados Cordero · 12 hours ago

Posted in Homework week 3

My results below:

SBP vs SMK

The regression is not significant. $F=1.95$. $p\text{-value} = 0.1723$ much greater than 0,05. $r=0.2474$, $r^2=0.0612$. Only 6% of SBP is explained by the regression with SMK.

SBP vs QUET

The regression is significant. $F=36.75$. $p\text{-value} = 0.0$ less than 0,05 and 0.01. $r=0.7420$, $r^2=0.5506$. 55% de SBP is explained by the regression with QUET.

QUET vs AGE

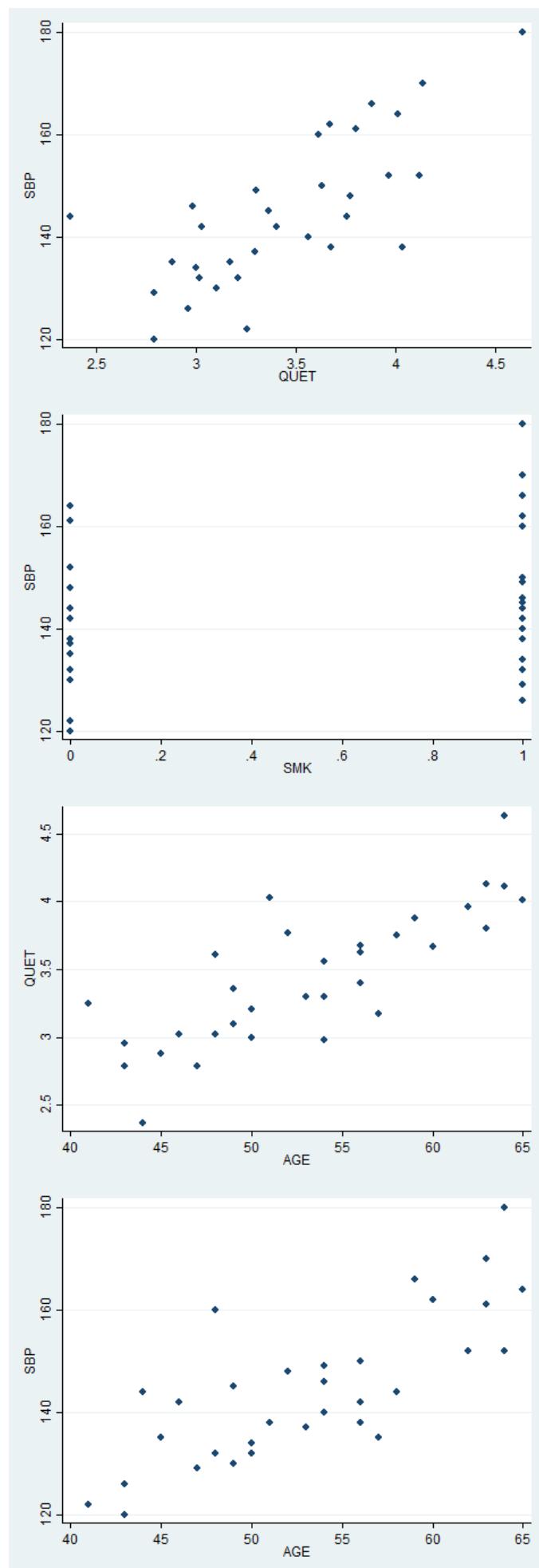
The regression is significant. $F=54.37$. $p\text{-value} = 0.0$ less than 0,05 and 0.01. $r=0.8027$, $r^2=0.6444$. 64.4% de QUET is explained by the regression with AGE.

SBP vs AGE

The regression is significant. $F=45.18$. $p\text{-value} = 0.0$ less than 0,05 and 0.01. $r=0.7752$, $r^2=0.6009$. 60% de SBP is explained by the regression with AGE.

luca balestrini · 4 days ago

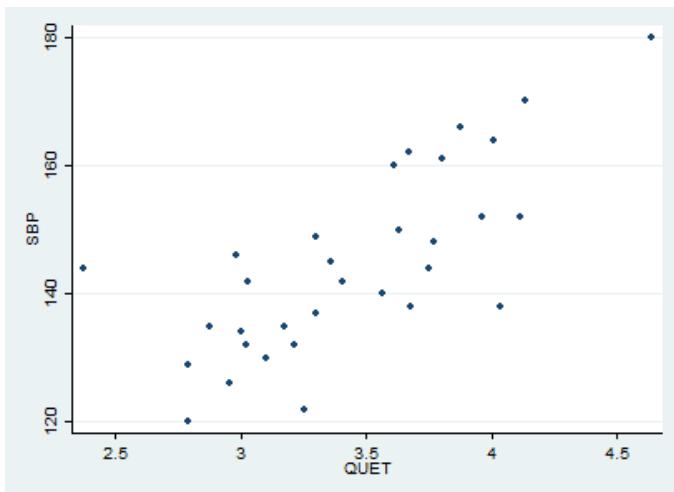
Posted in Week Two Homework - Exercise 1



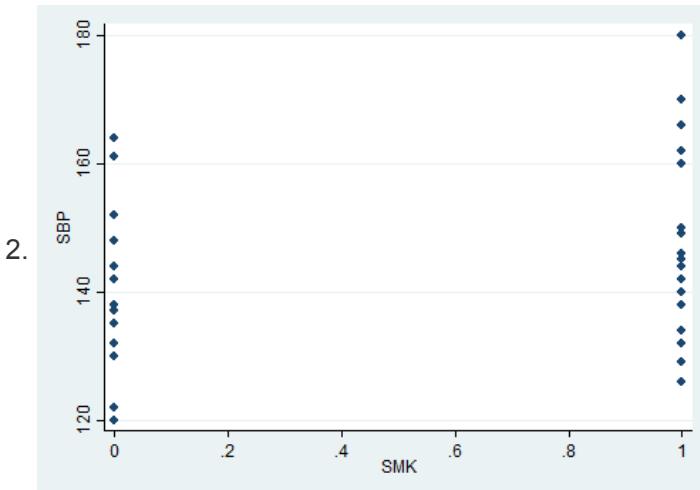
Viorel NITA · 14 days ago

Posted in Week Two Homework - Exercise 1

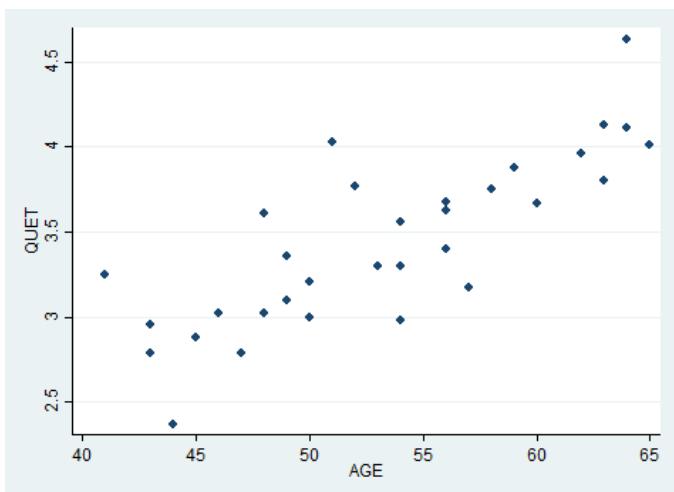
1.



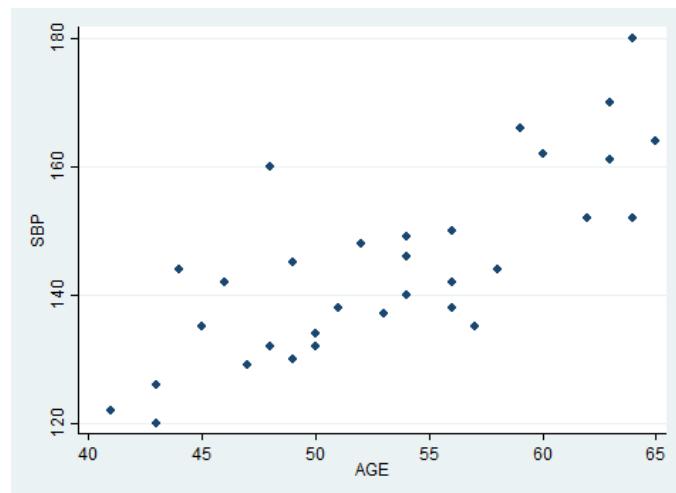
2.



3.

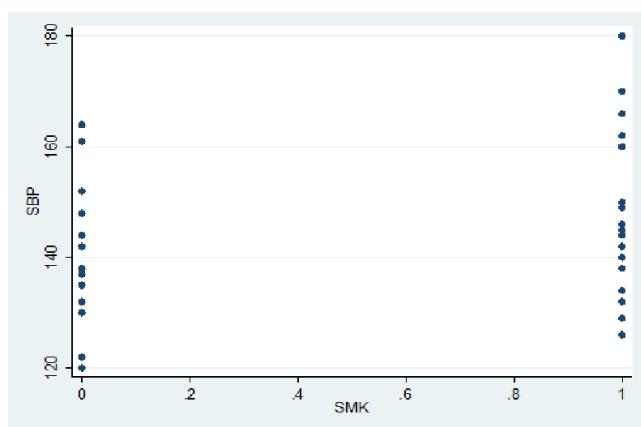
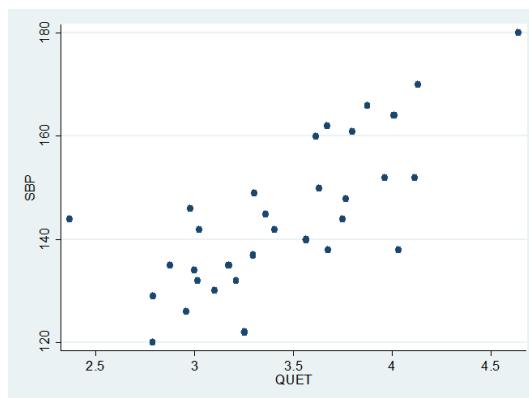


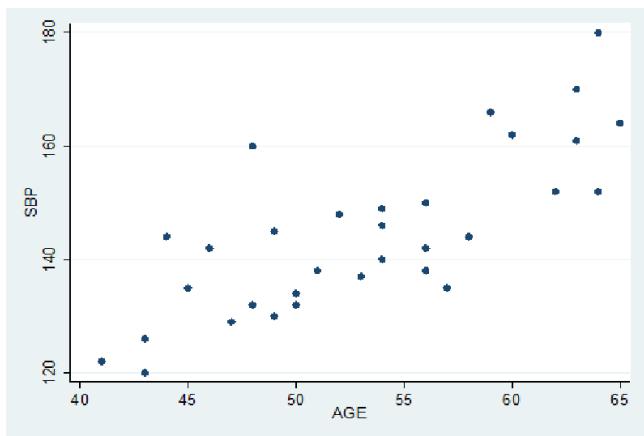
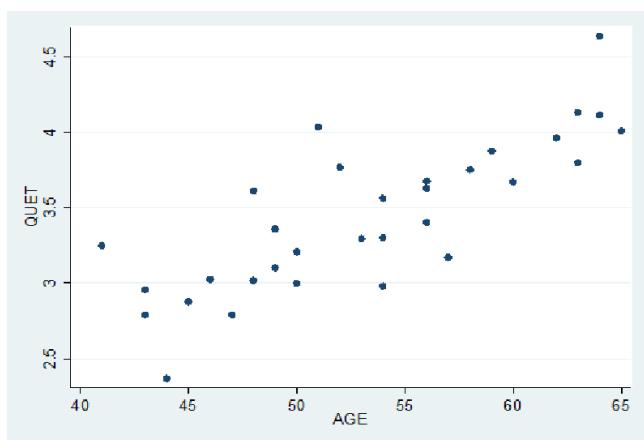
4.



Musa Hasen Ahmed · 12 days ago

Posted in Week Two Homework - Exercise 1





Viorel NITA · 2 days ago

Posted in week 3 homework

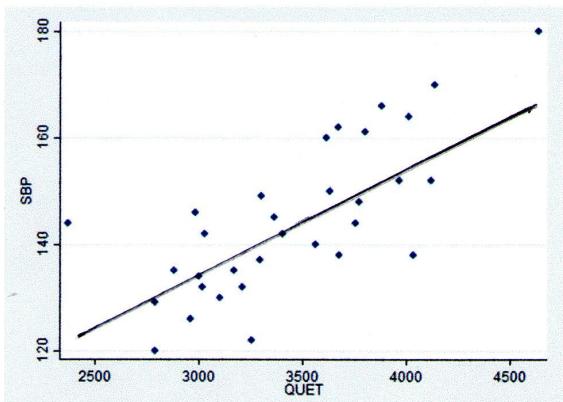
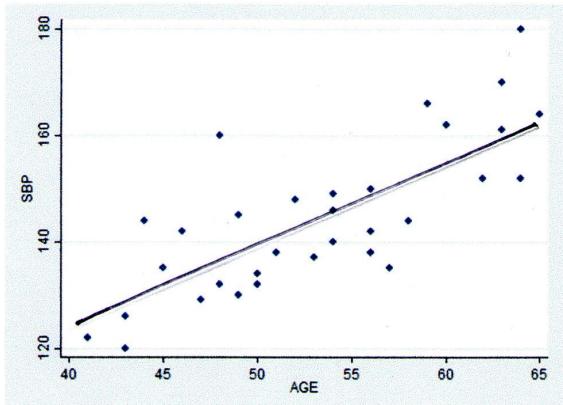
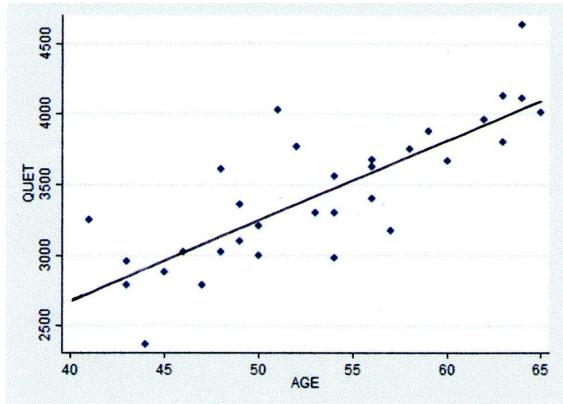
. regress SBP AGE

Source	SS	df	MS	Number of obs	=	32
+-----				F(1, 30)	=	45.18
Model	3861.63037	1	3861.63037	Prob > F	=	0.0000
Residual	2564.33838	30	85.4779458	R-squared	=	0.6009
+-----				Adj R-squared	=	0.5876
Total	6425.96875	31	207.289315	Root MSE	=	9.2454

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
+-----					
AGE	1.6045	.2387159	6.72	0.000	1.116977 2.092023
_cons	59.09162	12.81626	4.61	0.000	32.91733 85.26592
+-----					

Jorge Marques Pontes · 11 days ago

Posted in Week2 HW Exercise 2



I can perceive that there is positive linear relation between the variables in the three graphs..

1 2 3 → Next

Homework Solutions

Applied Regression Analysis

WEEK 2

Exercise One

Generate scatter diagrams for each of the following variable pairs:

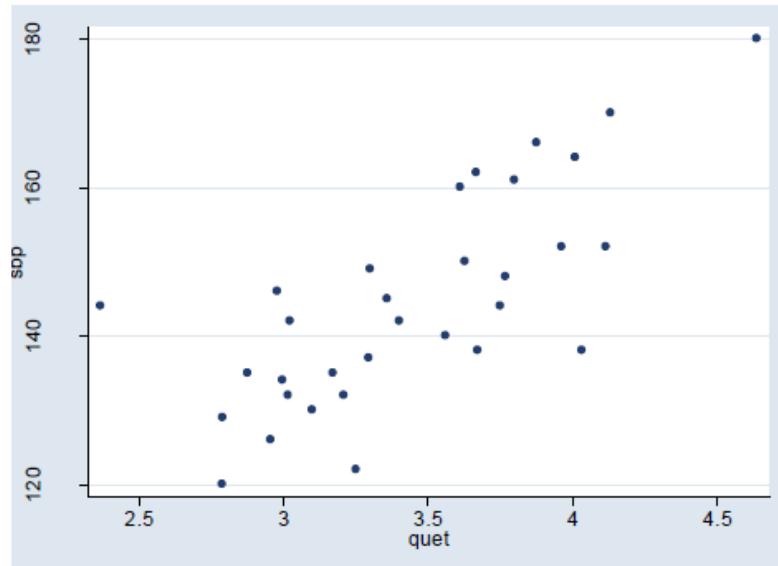
1. SBP (Y) vs. QUET (X)

In the command window, enter ‘.scatter sbp quet’

This will produce the scatterplot of SBP (Y) and QUET (X). The resulting scatterplot displayed should resemble the screenshot depicted below

(Note: Stata is case-sensitive so the variables should be lower-case if that is how they are coded in the dataset).

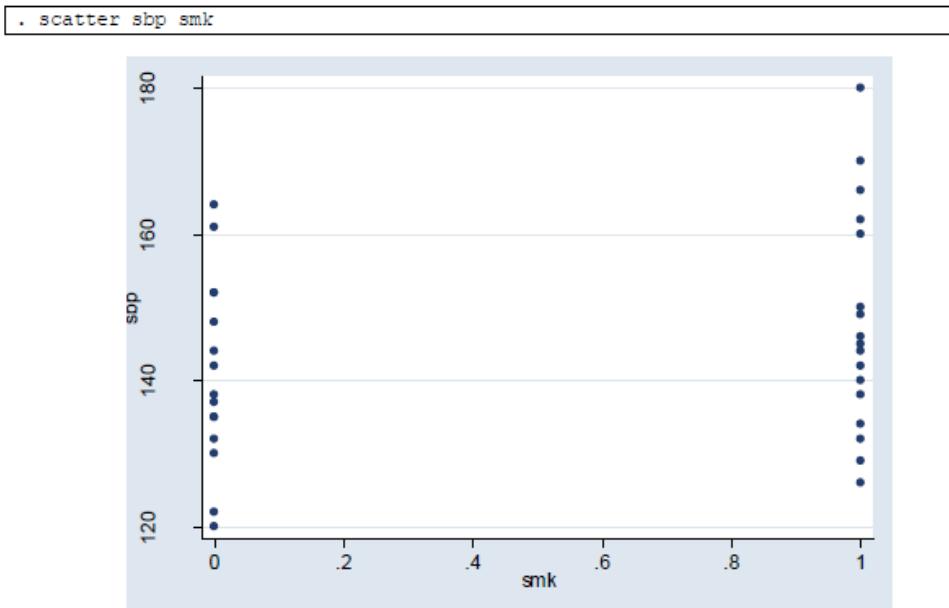
```
.scatter sbp quet
```



2. SBP (Y) vs. SMK (X)

In the command window, enter ‘.scatter sbp smk’.

This will produce the scatterplot of SBP (Y) and SMK (X). The resulting scatterplot displayed should resemble the screenshot depicted below.



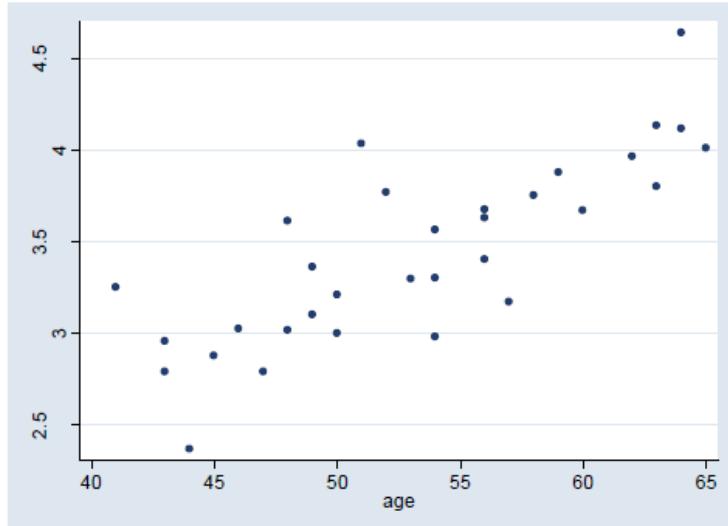
Note: Smoking is a binary variable, therefore we should not expect to see an even scatter of observations as with the other plots containing continuous variables.

3. QUET (Y) vs. AGE (X)

In the command window, enter ‘.scatter quet age’.

This will produce the scatterplot of QUET (Y) and AGE (X). The resulting scatterplot displayed should resemble the screenshot depicted below.

```
. scatter quet age
```

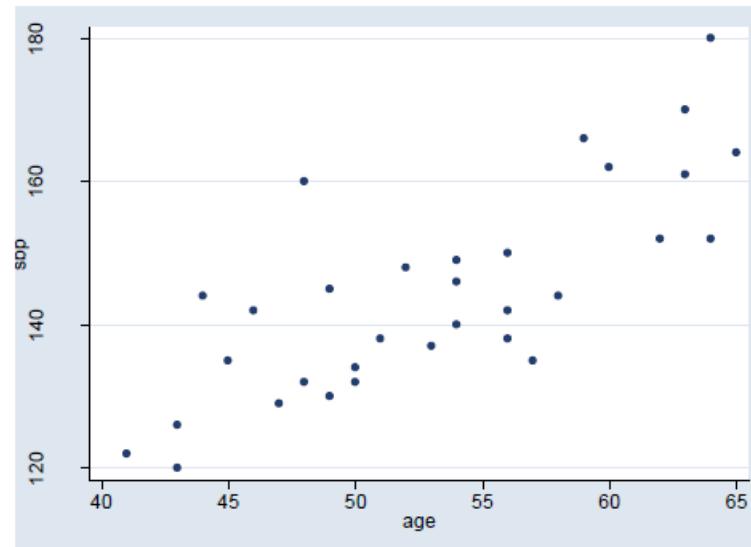


4. SBP (Y) vs. AGE (X)

In the command window, enter ‘.scatter sbp age’.

This will produce the scatterplot of SBP (Y) and AGE (X). The resulting scatterplot displayed should resemble the screenshot depicted below.

```
. scatter sbp age
```



Homework Solutions

Applied Regression Analysis

WEEK 2

Exercise Two

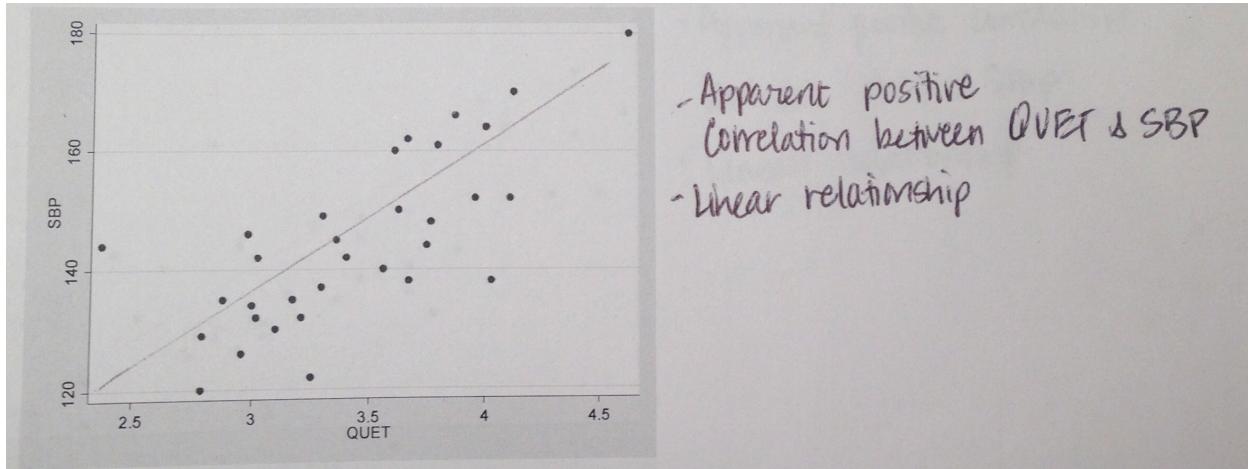
Sketch a line

Using scatter diagrams #1, #3, and #4 that you generated in exercise one, use paper and pencil to roughly sketch a line that fits the data reasonably well.

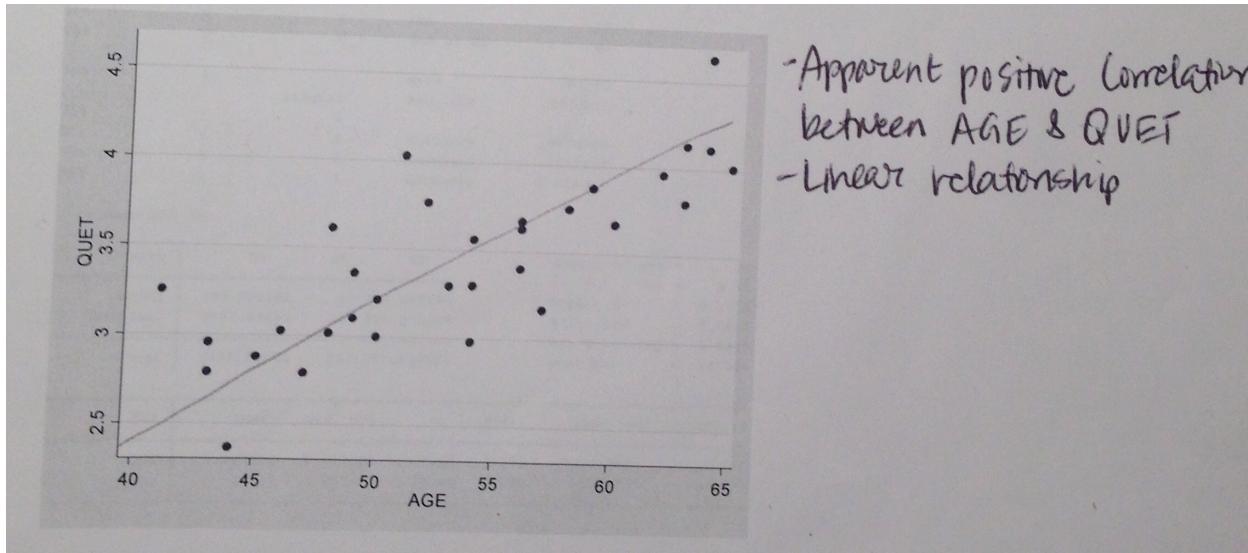
Use the **homework forum** to share your sketches and comment on the relationships described.

Below are examples of how this could look when you hand-draw a line.

#1 – SBP vs. QUET



#3 – QUET vs. AGE



#4 – SBP vs. AGE



Homework Solutions

Applied Regression Analysis

Exercise Three

Comparing Blood Pressure with Smoking History

1. Determine the least-squares estimates of slope (β_1) and intercept (β_0) for the straight-line regression of SBP (Y) on SMK (X).

We can determine the least squares estimates for the parameters in simple linear regression by regressing Y on X. In the command window, enter '.regress sbp smk'. This will produce the output below.

. regress sbp smk					
Source	SS	df	MS	Number of obs = 32	
Model	393.098162	1	393.098162	F(1, 30)	= 1.95
Residual	6032.87059	30	201.095686	Prob > F	= 0.1723
Total	6425.96875	31	207.289315	R-squared	= 0.0612
				Adj R-squared	= 0.0299
				Root MSE	= 14.181
sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
smk	7.023529	5.023498	1.398	0.172	-3.235823 17.28288
_cons	140.8	3.661472	38.454	0.000	133.3223 148.2777

$$\begin{aligned}y &= \beta_0 + \beta_1 x \\&= 140.8 + 7.02(\text{smk})\end{aligned}$$

Note: When entering data into STATA, always list the dependant variable first (sbp, in this case) and then the independent variable (smk, in this case).

The slope for smk is determined by the value listed in the "Coef." column of the table above, and the value for the intercept is determined in a similar fashion in the _cons row.

How can you interpret the slope?

"Under this model, current or previous smokers have an average systolic blood pressure 7.02 mm Hg higher than that of non-smokers"

2. Compare the value of $\hat{\beta}_0$ with the mean SBP for nonsmokers. Compare the value of $\hat{\beta}_0 + \hat{\beta}_1$ with the mean SBP for smokers. Explain the results of these comparisons.

To compare the mean of one variable across different categories of another variable in STATA, you must first sort the data by the second categorizing variable (in this case, smk).

In the command window, enter ‘.sort smk’.

You must then use the ‘sum’ command to get descriptive statistics on your variable of interest (in this case, sbp), but first you must use the ‘by’ command to split the results by smoking status.

In the command window, enter ‘.by smk:sum sbp’.

This will produce the output below.

. sort smk
. by smk:sum sbp
-> smk = 0
Variable Obs Mean Std. Dev. Min Max
-----+-----
sbp 15 140.8 12.90183 120 164
-> smk = 1
Variable Obs Mean Std. Dev. Min Max
-----+-----
sbp 17 147.8235 15.21198 126 180

The mean value of SBP for nonsmokers (140.8) is equal to the value of $\hat{\beta}_0$. The mean value of SBP for smokers is 147.82 which is equal to $\hat{\beta}_0 + \hat{\beta}_1$.

In simple linear regression, the intercept can be interpreted as the value for Y when X=0. Given that smoking is a binary variable, and is coded (0,1) (0 for non-smokers, 1 for smokers), then the intercept is the mean value for non-smokers (Y when X=0), and the intercept plus the slope is the mean value for smokers (Y when X=1).

3. Test the hypothesis that the true slope (β_1) is 0.

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

The null hypothesis cannot be rejected, $p = 0.172$. There is not sufficient evidence to conclude that the slope is significantly different from 0.

Note: You can test for the significance of the slope by looking at the p-value for the t-test in the table for the regression in problem 1. The p-value tells us that the probability of rejecting the null when the null is true is 17.2%, which exceeds 5%. Therefore there is insufficient evidence to reject the null.

4. Is the test in part (3) equivalent to the usual two-sample t test for the equality of two population means assuming equal but unknown variances? Demonstrate your answer numerically.

To perform a t-test to compare the mean sbp across the populations in the different smoking categories, enter ‘.ttest sbp, by(smk)’ into the command window.

This will produce the output below.

```
. ttest sbp, by(smk)

Two-sample t test with equal variances

-----+-----+-----+-----+-----+
      Group |     Obs        Mean    Std. Err.    Std. Dev.   [95% Conf. Interval]
-----+-----+-----+-----+-----+
          0 |     15      140.8    3.331237    12.90183    133.6552    147.9448
          1 |     17      147.8235   3.689448    15.21198    140.0022    155.6448
-----+-----+-----+-----+-----+
```

```
combined |     32      144.5313    2.545151    14.39755    139.3404    149.7221
-----+-----+-----+-----+-----+
      diff |      -7.023529    5.023498           -17.28288    3.235823
-----+-----+
Degrees of freedom: 30

Ho: mean(0) = mean(1) = diff = 0

Ha: diff < 0           Ha: diff ~= 0           Ha: diff > 0
      t =    -1.3981       t =    -1.3981       t =    -1.3981
      P < t =    0.0862       P > |t| =    0.1723       P > t =    0.9138
```

The t-test gives the same t-value and p-value as the test for the hypothesis that the true slope, β_1 , is 0.

The p-value for question 3 is the same at the two-sided p-value for the two-sample t-test. In both tests, you are testing whether smoking has a significant impact on systolic blood pressure by determining if the sbp for smokers is significantly different than that of non-smokers.

Homework Solutions

Applied Regression Analysis

WEEK 2

Exercise Four

1. Determine the least-squares estimates of slope and intercept for the straight-line regression of SBP (Y) on QUET (X).

We can determine the least squares estimates for the parameters in simple linear regression by regressing Y on X.

In the command window, enter ‘.regress sbp quet’.

This will produce the output below.

. regress sbp quet					
Source	SS	df	MS		
Model	3537.94585	1	3537.94585	Number of obs = 32	
Residual	2888.0229	30	96.2674299	F(1, 30) = 36.75	
Total	6425.96875	31	207.289315	Prob > F = 0.0000	
sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
quet	21.49167	3.545147	6.062	0.000	14.25151 28.73182
_cons	70.57641	12.32187	5.728	0.000	45.4118 95.74102

$$\hat{\beta}_0 = 70.576$$

$$\hat{\beta}_1 = 21.492$$

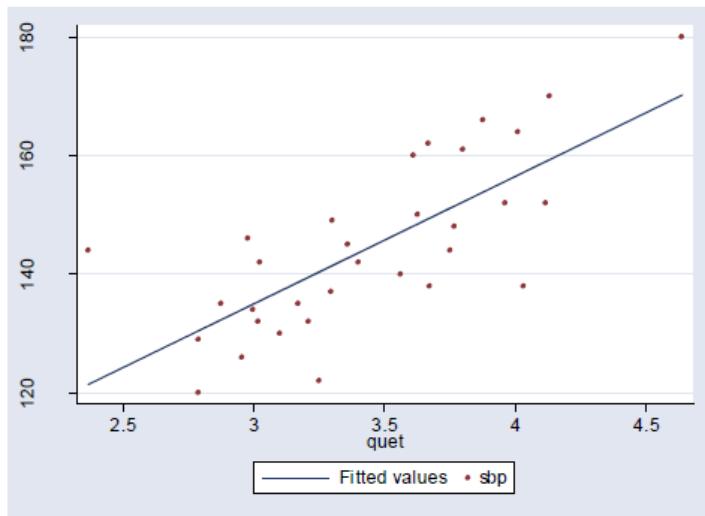
2. Sketch the estimated regression line on the scatter diagram involving SBP and QUET.

In order to fit a regression line in STATA, you must first create a new variable in your dataset for of the predicted y given x under the regression model.

You can do this simply by entering ‘predict yhat’ into the command window. Next, create a scatterplot with a line by entering ‘scatter yhat sbp quet, c(1 .) s(i o)’ into the command window.

The commands ‘c(1 .) s(i o)’ specify that the yhat should be labeled with a line and data points with dots, respectively.

```
. predict yhat  
(option xb assumed; fitted values)  
  
. scatter yhat sbp quet, c(1 .) s(i o)
```



3. Test the hypothesis of zero slope.

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

Reject the null hypothesis, $p < 0.001$. There is sufficient evidence to conclude that the slope is significantly different from 0.

Note: You can test for the significance of the slope by looking at the p-value for the t-test in the table for the regression in problem 1. The p-value tells us that the probability of rejecting the null when the null is true is less than 5%. Therefore there is sufficient evidence to reject the null.

4. Find a 95% confidence interval for $\mu_{y|\bar{x}}$.

To calculate confidence intervals, you need to know the descriptive statistics for the variables, including their mean values and standard deviations.

To get these values, use the ‘sum’ command by entering ‘.sum sbp quet age smk’ into the command window.

Next we can calculate $\mu_{y|\bar{x}}$ by entering the mean value for quet within the regression equation using our previously estimated parameters. The confidence limits about $\mu_{y|\bar{x}}$ can then be estimated using the mean value and standard deviation of x.

. sum sbp quet age smk					
Variable	Obs	Mean	Std. Dev.	Min	Max
sbp	32	144.5313	14.39755	120	180
quet	32	3.441094	.4970781	2.368	4.637
age	32	53.25	6.956083	41	65
smk	32	.53125	.5070073	0	1

$$\hat{y}_{\bar{x}} = 70.57641 + 21.49167 * 3.44 = 144.508$$

$$s_{\hat{y}_{x_0}}^2 = s_{\hat{y}_{\bar{x}}}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2} \right)$$

$$s_{\hat{y}_{\bar{x}}}^2 = s_{\hat{y}_{\bar{x}}}^2 \left(\frac{1}{n} \right) = \frac{96.2674299}{32} = 3.008357$$

$$s_{\hat{y}_{\bar{x}}} = \sqrt{3.008357} = 1.7344616$$

$$95\% \text{ CI: } \hat{y}_{\bar{x}} \pm t_{975}(30) s_{\hat{y}_{\bar{x}}} = 144.508 \pm 2.042 \times 1.7344616 = (140.97, 148.05)$$

Interpretation: We are 95% confident that the true value for the mean value of y is between 140.97 and 148.05 mm Hg.

5. Calculate 95% prediction bands.

For this problem, we are asking for a plot of the prediction bands using STATA, not for hand-calculations. To do this, we must enter 'predict sepred, stdf' into the command window to generate a variable- 'sepred'- for the standard deviation used within the prediction interval.

Next, we can calculate the value for the lower limit of the prediction interval by entering 'generate low=yhat-invtail(30,0.025)*sepred' and the upper limit of the prediction interval by entering 'generate high=yhat+invtail(30,0.025)*sepred' (note: invtail(30,0.025)= $t_{0.975}(30)$).

From here you can create a plot of the prediction intervals with the regression line by entering 'scatter sbp yhat low high, sort connect (. 1 1 1) symbol (o i i i)'. The code and plot below includes both the confidence and prediction intervals, however you only need to graph the prediction intervals for this question.

```
. predict yhat
(option xb assumed; fitted values)

. predict seyhat, stdp

. display invtail(30,0.025)
2.0422724

. generate lowl= yhat-2.0422724* seyhat

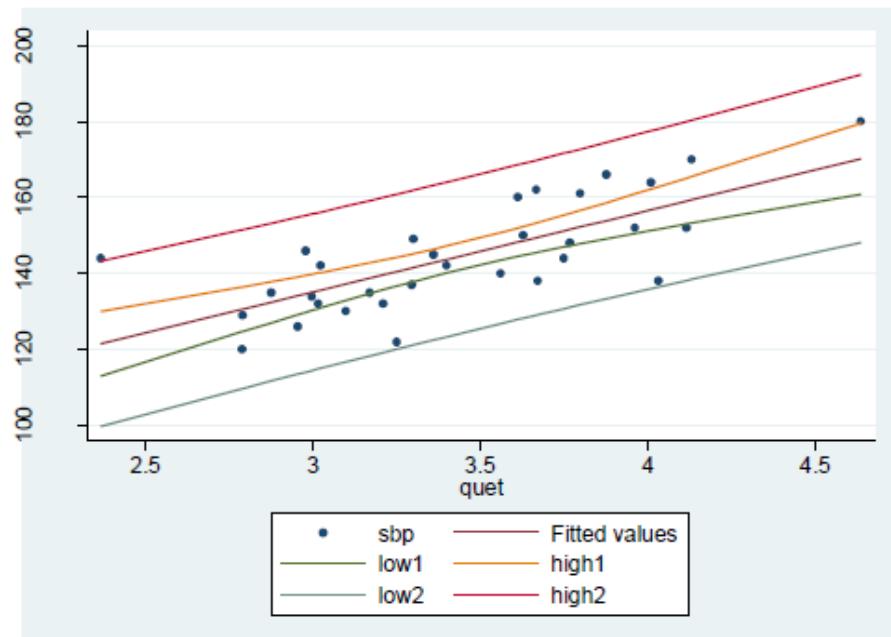
. generate highl= yhat+2.0422724* seyhat

. predict sepred, stdf

. generate low2= yhat-invtail(30,0.025)* sepred

. generate high2= yhat+invtail(30,0.025)* sepred

. scatter sbp yhat lowl highl low2 high2 quet,sort connect(. 1 1 1 1 1)
symbol(o i i i i)
```



- 6. Based on the above, would you conclude that blood pressure increases as body size increases?**

Yes, because the fitted regression line, as well as the confidence and prediction band, appear to have an upward slope.

- 7. Are any of the assumptions for straight-line regression clearly not satisfied in this example?**

Simple Linear Regression Assumptions:

Linearity: SBP and SMK appear to be linearly related based on the above scatterplot

Independence: The study design does not suggest that the observations are not independent

Normality: The variables appear to be normally distributed (there are no significant outliers)

Equal Variance (homoscedasticity): The variances along the regression line appear to remain similar as you move across the line

There are no apparent violations of homoscedasticity, normality, or independence. Formal tests of these assumptions are possible but are not included here.

Week Three Homework Exercise

[Help Center](#)[Homework Central](#) / Week Three Homework /

For this week's homework, you will notice that the table of data is the same one from the [week two homework](#) from last week.

Use the data from the table to determine the ANOVA tables for some regressions specifically noted below. Scroll down below the table to view the exercise. Use the [homework forum](#) to discuss this exercise with your peers and share your discoveries.

To recall, the following table gives the systolic blood pressure (SBP), body size (QUET), age (AGE), and smoking history (SMK = 0 if nonsmoker, SMK = 1 if a current or previous smoker) for a hypothetical sample of 32 white males over 40 years old from the town of Angina.

Person	SBP	QUET	AGE	SMK
1	135	2.876	45	0
2	122	3.251	41	0
3	130	3.100	49	0
4	148	3.768	52	0
5	146	2.979	54	1
6	129	2.790	47	1
7	162	3.668	60	1
8	160	3.612	48	1
9	144	2.368	44	1
10	180	4.637	64	1
11	166	3.877	59	1
12	138	4.032	51	1
13	152	4.116	64	0
14	138	3.673	56	0
15	140	3.562	54	1
16	134	2.998	50	1
17	145	3.360	49	1
18	142	3.024	46	1
19	135	3.171	57	0

20	142	3.401	56	0
21	150	3.628	56	1
22	144	3.751	58	0
23	137	3.296	53	0
24	132	3.210	50	0
25	149	3.301	54	1
26	132	3.017	48	1
27	120	2.789	43	0
28	126	2.956	43	1
29	161	3.800	63	0
30	170	4.132	63	1
31	152	3.962	62	0
32	164	4.010	65	0

If desired, you may download the data for this exercise in this [CSV file](#)

For our homework, complete the following:

Exercise One

Determine the ANOVA tables for the following regressions:

1. SBP (Y) on SMK (X)
2. SBP (Y) on QUET (X)
3. QUET (Y) on AGE (X)
4. SBP (Y) on AGE (X)

Exercise Two

Use the ANOVA tables to perform the F-test for the significance of each straight-line regression.

Exercise Three

Interpret your results.

Use the [homework forum](#) to share your discoveries and discuss this homework with your peers.

You can download all homework assignments for week 3 [here](#)

For help and answers to this week's exercises, click the button below to visit the solutions page.

[Solutions Page](#)

Created Tue 10 Mar 2015 8:17 AM PDT

Last Modified Mon 6 Apr 2015 9:25 AM PDT

Week Three Homework Solutions

[Help Center](#)[Homework Central](#) / [Week Three Homework](#) / Week Three Solutions

If you need help and answers to the exercises of week three, please click below on the selected exercise:

- [Exercise One](#)
- [Exercise Two](#)
- **Exercise Three** requires that you use our [Discussion Forums](#) to share your discoveries and discuss insights with your peers.

Click the button below to return to this week's homework exercise.

[←Homework Page](#)

Created Tue 10 Mar 2015 8:53 AM PDT

Last Modified Thu 9 Apr 2015 12:38 PM PDT

Homework Solutions

Applied Regression Analysis

WEEK 3

Exercise One

Determine the ANOVA tables for the following regressions:

For each of the regressions, type “regress” followed by the Y and X variables, into the command box. From this command, you should see an ANOVA table, as well as a t-statistic for the slope coefficient below, and an F-statistic of model fit in the upper right corner (*See images below*)

1. SBP (Y) on SMK (X)

. regress sbp smk					
Source	SS	df	MS	Number of obs = 32	
Model	393.098162	1	393.098162	F(1, 30) =	1.95
Residual	6032.87059	30	201.095686	Prob > F =	0.1723
Total	6425.96875	31	207.289315	R-squared =	0.0612
				Adj R-squared =	0.0299
				Root MSE =	14.181
<hr/>					
sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
smk	7.023529	5.023498	1.398	0.172	-3.235823 17.28288
_cons	140.8	3.661472	38.454	0.000	133.3223 148.2777

2. SBP (Y) on QUET (X)

. regress sbp quet					
Source	SS	df	MS	Number of obs = 32	
Model	3537.94585	1	3537.94585	F(1, 30) =	36.75
Residual	2888.0229	30	96.2674299	Prob > F =	0.0000
Total	6425.96875	31	207.289315	R-squared =	0.5506
				Adj R-squared =	0.5356
				Root MSE =	9.8116
<hr/>					
sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
quet	21.49167	3.545147	6.062	0.000	14.25151 28.73182
_cons	70.57641	12.32187	5.728	0.000	45.4118 95.74102

3. QUET (Y) on AGE (X)

```
. regress quet age
```

Source	SS	df	MS	Number of obs	=	32
Model	4.93597216	1	4.93597216	F(1, 30)	=	54.37
Residual	2.72371324	30	.090790441	Prob > F	=	0.0000
Total	7.6596854	31	.247086626	R-squared	=	0.6444
				Adj R-squared	=	0.6326
				Root MSE	=	.30131

quet	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.0573642	.0077799	7.37	0.000	.0414755 .0732529
_cons	.3864517	.4176903	0.93	0.362	-.4665857 1.239489

4. SBP (Y) on AGE (X)

```
. regress sbp age
```

Source	SS	df	MS	Number of obs	=	32
Model	3861.63037	1	3861.63037	F(1, 30)	=	45.18
Residual	2564.33838	30	85.4779458	Prob > F	=	0.0000
Total	6425.96875	31	207.289315	R-squared	=	0.6009
				Adj R-squared	=	0.5876
				Root MSE	=	9.2454

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	1.6045	.2387159	6.72	0.000	1.116977 2.092023
_cons	59.09162	12.81626	4.61	0.000	32.91733 85.26592

Homework Solutions

Applied Regression Analysis

WEEK 3

Exercise Two

Use the ANOVA tables to perform the F test for the significance of each straight-line regression.

For each of the following, we are testing whether the slope coefficient is equal to zero. In other words, we are testing if the independent variable contributes significantly to the model. Notice that, because there is only one independent variable in the model, the F-test for the overall model yields the same significance as the t-test for the slope coefficient.

Looking at SBP (Y) on SMK (X)

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

<code>. regress sbp smk</code>					
Source	SS	df	MS	Number of obs = 32	
Model	393.098162	1	393.098162	F(1, 30) = 1.95	
Residual	6032.87059	30	201.095686	Prob > F = 0.1723	
Total	6425.96875	31	207.289315	R-squared = 0.0612	
<hr/>					
	sbp	Coef.	Std. Err.	t	P> t [95% Conf. Interval]
<hr/>					
smk	7.023529	5.023498	1.398	0.172	-3.235823 17.28288
_cons	140.8	3.661472	38.454	0.000	133.3223 148.2777

With F=1.95 and p=0.1723, we fail to reject the null hypothesis. There is not sufficient evidence to conclude that there is a significant straight-line relationship between SBP and SMK.

Looking at SBP (Y) on QUET (X)

<code>. regress sbp quet</code>					
Source	SS	df	MS	Number of obs = 32	
Model	3537.94585	1	3537.94585	F(1, 30) = 36.75	
Residual	2888.0229	30	96.2674299	Prob > F = 0.0000	
Total	6425.96875	31	207.289315	R-squared = 0.5506	
<hr/>					
	sbp	Coef.	Std. Err.	t	P> t [95% Conf. Interval]
<hr/>					
quet	21.49167	3.545147	6.062	0.000	14.25151 28.73182
_cons	70.57641	12.32187	5.728	0.000	45.4118 95.74102

With F=36.75 and p<0.0001, we reject the null hypothesis. There is sufficient evidence to conclude that there is a significant straight-line relationship between SBP and QUET.

Week 3 Homework Assignment

[Subscribe for email updates.](#)

No tags yet. [+ Add Tag](#)

Sort replies by: [Oldest first](#) [Newest first](#) [Most popular](#)

Amos B Robinson · 4 days ago 

Effect of Smoking on Systolic Blood Pressure

Number of obs = 32

F(1, 30) = 1.95

Prob > F = 0.1723

Under the Naive Model the blood pressure is 144. So because the "F" statistic is so low, there is not convincing evidence that smoking has a statistical significance on SBP. This is confirmed on the "F" statistic because the "P" value for the "F" statistic is .1723, which is above the critical value of .05 for a 95% confidence level. Additionally, the R-squared for this regression is 0.0612, which further confirms that the statistical impact of smoking on blood pressure is minimal.

Effect of Body Size on Systolic Blood Pressure

Number of obs = 32

F(1, 30) = 36.75

Prob > F = 0.0000

R-squared = 0.5506

Under the Naive Model the blood pressure is 144. Here the "F" statistic is very large and the associated "P" value is < than .0000. So there is convincing evidence that body size has a statistically significant effect on systolic blood pressure. The high coefficient of determination "R-squared" shows that QUET explains about 55% of SBP.

Effect of AGE on Body Size

Number of obs = 32

F(1, 30) = 54.37

Prob > F = 0.0000

R-squared = 0.6444

The mean body size of this sample is 3.44. However, there is convincing evidence that age does effect body size. The "F" statistic is large at 54.37 and the associated "P" value is less than .0000. The R-squared shows that age explains 64.44% of body size.

Effect of Age on Systolic Blood Pressure

Number of obs = 32

F(1, 30) = 45.18

Prob > F = 0.0000

R-squared = 0.6009

The Naive Model for SBP is 144. The "F" statistic is 45.18, which is pretty large. The corresponding "P" value for this "F" statistic is less than .0000. So there is convincing evidence that age does have

an impact on systolic blood pressure.

4 · flag

Kahsay Tadesse · 3 days ago

well done, great

0 · flag

[+ Comment](#)

Kahsay Tadesse · 3 days ago

am not clear about the culculation on :"Effect of Body Size on Systolic Blood Press

0 · flag

[+ Comment](#)

KK Wong · 3 days ago

1. SBP vs SMK

. reg SBP SMK

Source	SS	df	MS	Number of obs	=	32
Model	393.098162	1	393.098162	F(1, 30)	=	1.95
Residual	6032.87059	30	201.095686	Prob > F	=	0.1723
Total	6425.96875	31	207.289315	R-squared	=	0.0612
				Adj R-squared	=	0.0299
				Root MSE	=	14.181

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
SMK	7.023529	5.023498	1.40	0.172	-3.235823 17.28288
_cons	140.8	3.661472	38.45	0.000	133.3223 148.2777

Given relatively small $F(1,30)=1.95$, high $\text{Prob}>F=0.1723$ and low adj R-squared=0.0299, it suggests to reject the hypothesis that SBP has an association with SMK; ie, the association between SBP & SMK is approx 2.99% as suggested by the adj R-squared. It further supports by the fact that SMK 95% CI contains 0 and its $P>|t|=0.172$ which is significantly away from 0.

2. SBP vs QUET

. reg SBP QUET

Source	SS	df	MS	Number of obs	=	32
Model	3537.94574	1	3537.94574	F(1, 30)	=	36.75
Residual	2888.02301	30	96.2674337	Prob > F	=	0.0000
Total	6425.96875	31	207.289315	R-squared	=	0.5506
				Adj R-squared	=	0.5356
				Root MSE	=	9.8116

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
QUET	21.49167	3.545147	6.06	0.000	14.25151 28.73182
_cons	70.5764	12.32187	5.73	0.000	45.41179 95.74101

Given relatively high $F(1,30)=36.75$, significantly low $\text{Prob}>F=0.0000$ and high adj R-squared=0.5356, it suggests not to reject the hypothesis that SBP vs QUET; ie, the association between SBP & QUET is

approx 53.56% as suggested by the adj R-squared. It further supports by the fact that QUET 95% CI does not contains 0 and its $P>|t|=0.000$.

3. QUET vs AGE

. reg QUET AGE

Source	SS	df	MS	Number of obs =	32
Model	4.93597143	1	4.93597143	F(1, 30) =	54.37
Residual	2.72371329	30	.090790443	Prob > F =	0.0000
Total	7.65968472	31	.247086604	R-squared =	0.6444
				Adj R-squared =	0.6326
				Root MSE =	.30131

QUET	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
AGE	.0573642	.0077799	7.37	0.000	.0414755 .0732529
_cons	.3864519	.4176903	0.93	0.362	-.4665855 1.239489

Given relatively high $F(1,30)=54.37$, significantly low $\text{Prob}>F=0.0000$ and high adj R-squared=0.6326, it suggests not to reject the hypothesis that QUET vs AGE; ie, the association between QUET & AGE is approx 63.26% as suggested by the adj R-squared. It further supports by the fact that AGE 95% CI does not contains 0 and its $P>|t|=0.000$.

4. SBP vs AGE

. reg SBP AGE

Source	SS	df	MS	Number of obs =	32
Model	3861.63037	1	3861.63037	F(1, 30) =	45.18
Residual	2564.33838	30	85.4779458	Prob > F =	0.0000
Total	6425.96875	31	207.289315	R-squared =	0.6009
				Adj R-squared =	0.5876
				Root MSE =	9.2454

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
AGE	1.6045	.2387159	6.72	0.000	1.116977 2.092023
_cons	59.09162	12.81626	4.61	0.000	32.91733 85.26592

Given relatively high $F(1,30)=45.18$, significantly low $\text{Prob}>F=0.0000$ and high adj R-squared=0.5876, it suggests not to reject the hypothesis that SMK vs AGE; ie, the association between SMK & AGE is approx 58.76% as suggested by the adj R-squared. It further supports by the fact that AGE 95% CI does not contains 0 and its $P>|t|=0.000$.

↑ 1 ↓ · flag

+ Comment



Juan C. Trujillo · 2 days ago

I am confused about this case. Could there be a possibility in which the F test turns out to be statistically significant, but the t test for the explanatory variable appears with a high p-value?

Please, explain.

↑ 0 ↓ · flag

Anonymous · 2 days ago

If you are testing the statistical significance of the estimate of a single regression coefficient, the inferential outcome will be the same whether you use the t-test or the F-test.

Mechanically, this is because the F-statistic associated with testing a single coefficient estimate is just squared t-statistic associated with that same estimate.

↑ 0 ↓ · flag

+ Comment

Varalakshmi · 2 days ago 

Amos: I did the F statistic for SBP (Y) on SMK (X)

The Model with 1 d.f parameter because there is one independent variable SMK and the residual which is y and y predicted has 2 d.f one for intercept and slope. As you stated the R squared which is the explained variation to the Total variation is very little. P value so small suggests that there is a significant difference in Smoking history and SBP.

Finding the t-statistic = $r/\sqrt{(1-r^2)/(n-2)}$ = 1.398459 or as is also given in the table below.

We can check that $t^2 = 1.40^2$ is the same as F statistic= 1.95

The lecture was quite informative!

Can someone help me to graph two scatter plots in one - Week 1 homework where we need to graph the residuals as well the given observations?

Source	SS	df	MS	Number
> of obs	=	32		
				$F(1, > 30) = 1.95$
Model	393.098162	1	393.098162	Prob > F = 0.1723
Residual	6032.87059	30	201.095686	R-squared = 0.0612
				Adj R-squared = 0.0299
Total	6425.96875	31	207.289315	Root MSE = 14.181

sbp	Coef.	Std. Err.	t	P> t	[95 % Con f. Interval]
smk	7.023529	5.023498	1.40	0.172	-3.2 35823
	17.28288				
_cons	140.8	3.661472	38.45	0.000	133 .3223
	148.2777				

↑ 0 ↓ · flag

+ Comment

luca balestrini · a day ago 

Source SS df MS Number of obs = 32

	F(1, 30)	= 1.95
Model	393.098162	1 393.098162 Prob > F = 0.1723
Residual	6032.87059	30 201.095686 R-squared = 0.0612
	Adj R-squared	= 0.0299
Total	6425.96875	31 207.289315 Root MSE = 14.181
sbp	Coef.	Std. Err. t P>t [95% Conf. Interval]
smk	7.023529	5.023498 1.40 0.172 -3.235823 17.28288
_cons	140.8	3.661472 38.45 0.000 133.3223 148.2777

2) sbp on quet

Source	SS	df	MS	Number of obs = 32
		F(1, 30)	= 36.75	
Model	3537.94585	1 3537.94585 Prob > F = 0.0000		
Residual	2888.0229	30 96.2674299 R-squared = 0.5506		
		Adj R-squared	= 0.5356	
Total	6425.96875	31 207.289315 Root MSE = 9.8116		
sbp	Coef.	Std. Err. t P>t [95% Conf. Interval]		
quet	21.49167	3.545147 6.06 0.000 14.25151 28.73182		
_cons	70.57641	12.32187 5.73 0.000 45.4118 95.74102		

3) quet on age

Source	SS	df	MS	Number of obs = 32
		F(1, 30)	= 54.37	
Model	4.93597216	1 4.93597216 Prob > F = 0.0000		
Residual	2.72371324	30 .090790441 R-squared = 0.6444		
		Adj R-squared	= 0.6326	
Total	7.6596854	31 .247086626 Root MSE = .30131		
quet	Coef.	Std. Err. t P>t [95% Conf. Interval]		
age	.0573642	.0077799 7.37 0.000 .0414755 .0732529		
_cons	.3864517	.4176903 0.93 0.362 -.4665857 1.239489		

4) sbp on age

Source	SS	df	MS	Number of obs = 32
		F(1, 30)	= 45.18	
Model	3861.63037	1 3861.63037 Prob > F = 0.0000		
Residual	2564.33838	30 85.4779458 R-squared = 0.6009		
		Adj R-squared	= 0.5876	
Total	6425.96875	31 207.289315 Root MSE = 9.2454		
sbp	Coef.	Std. Err. t P>t [95% Conf. Interval]		
age	1.6045	.2387159 6.72 0.000 1.116977 2.092023		
_cons	59.09162	12.81626 4.61 0.000 32.91733 85.26592		

↑ 0 ↓ · flag

[+ Comment](#)

David C. Morris · a day ago 

I got the same results as above. I'm not going to repost here. However, I was thinking about the four analyses we did. Two of them make sense to me: 1) Blood pressure (sbp) and Smoking (smk); and 2) Blood pressure (sbp) and Body size (quet). I know we're just doing the homework to learn how to run/interpret anova tables. However, it got me thinking about what 'quet' really is. The description says 'body size' but looking at the values (range from ~2 to ~4.6) I don't have any reference point for that. What is Quet? How is it measured? The reason I bring this up is I was a little surprised by the high correlation between Quet and Age. The range of age in the sample is 41 to 65. The correlation between them is R-squared .64. Looking at the scatter plot, as Age goes up, so does Quet. I would've expected it to taper off toward the older ages since people tend to lose muscle mass and usually weight the older they get. Does anyone know what QUET is? I did a Google search but only found other data sets with no description.

 0  · flag

[+ Comment](#)

Emilija Nikolic-Djoric · 20 hours ago 

I think that it is Quetelet's index defined as:

$$\text{QUET} = 100 * (\text{weight}/\text{height}^2)$$

 1  · flag

[+ Comment](#)

Emilija Nikolic-Djoric · 19 hours ago 

Week 3-Slide 17-at the bottom n-2 instead n-1?

 0  · flag

[+ Comment](#)

ANCA MINCIU · 18 hours ago 

I have done the test, with same results. To avoid re-posting the same information, I would just add one sentence to each interpretation.

1. Effect of Smoking on Systolic Blood Pressure

The confidence interval in this case contains zero, so smoking is not a good predictor for blood pressure.

2. Effect of Body Size on Systolic Blood Pressure

The confidence interval does not contain zero, so the body size influences the systolic blood pressure.

3. Effect of AGE on Body Size

The confidence interval does not contain zero, so age is a very important predictor for body size.

4. Effect of Age on Systolic Blood Pressure

Age is a very important predictor for systolic blood pressure, taking into consideration that the confidence interval does not contain zero.

↑ 0 ↓ · flag

+ Comment

Alina Denham · 11 hours ago 

Hello, everybody!

Here are my observations:

- Let's test the null hypothesis (the slope equals zero) for SBP on SMK. $F=1.95 (< F_{.95}=4.20)$ and $p=0.1723 (>.05)$. Therefore, we fail to reject the null hypothesis. This means that we do not have sufficient evidence to prove that there is a significant linear relationship between blood pressure and smoking history.
- Let's test the null hypothesis (the slope equals zero) for SBP on QUET. $F=36.75 (>> F_{.95})$ and $p=0.000 (<0.001)$. Therefore, we reject the null hypothesis. This means that we have sufficient evidence to prove that there is a significant linear relationship between blood pressure and body size.
- Let's test the null hypothesis (the slope equals zero) for QUET on AGE. $F=54.37 (>> F_{.95})$ and $p=0.000 (<0.001)$. Therefore, we reject the null hypothesis. This means that we have sufficient evidence to prove that there is significant linear relationship between body size and age.
- Let's test the null hypothesis (the slope equals zero) for SBP on AGE. $F=45.18 (>> F_{.95})$ and $p=0.000 (<0.001)$. Therefore, we reject the null hypothesis. This means that we have sufficient evidence to prove that there is significant linear relationship between blood pressure and age.

↑ 0 ↓ · flag

+ Comment

Lien-yu Yeh · 10 hours ago 

Assume that,

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

According to the rule of test with P-value,

if $\alpha > p\text{-value}$, then we reject H_0 , and if $\alpha < p\text{-value}$, we don't reject H_0 , where α is significant level because $p\text{-value} = \Pr(\text{reject } H_0 \mid H_0 \text{ is true}) = \text{probability of type I error (with sample)}$, when $p\text{-value}$ is large, H_0 probably be true, because there is high probability to make mistake, so we don't reject H_0
when $p\text{-value}$ is small, the probability of type I error is too low, so we reject H_0

I used above concept to test following question, and I assume that significant level is 0.05($\alpha=0.05$)

. regress SBP SMK

Source	SS	df	MS	Number of obs	=	32
				F(1, 30)	=	1.95
Model	393.098162	1	393.098162	Prob > F	=	0.1723
Residual	6032.87059	30	201.095686	R-squared	=	0.0612
				Adj R-squared	=	0.0299
Total	6425.96875	31	207.289315	Root MSE	=	14.18

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
SMK	7.023529	5.023498	1.40	0.172	-3.235823 17.28288
_cons	140.8	3.661472	38.45	0.000	133.3223 148.2777

significant level=0.05 < p-value=0.1723 ,we don't reject H0: β_1 equal to 0,
so we don't have enough evidence to refer that SBP and SMK has significant relationship.

. regress SBP QUET

Source	SS	df	MS	Number of obs	=	32
				F(1, 30)	=	36.75
Model	3537.94574	1	3537.94574	Prob > F	=	0.0000
Residual	2888.02301	30	96.2674337	R-squared	=	0.5506
				Adj R-squared	=	0.5356
Total	6425.96875	31	207.289315	Root MSE	=	9.8116

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
QUET	21.49167	3.545147	6.06	0.000	14.25151 28.73182
_cons	70.5764	12.32187	5.73	0.000	45.41179 95.74101

significant level=0.05 > p-value=0.0000 ,we reject H0: β_1 equal to 0,
so we have enough evidence to refer that SBP and QUET has significant relationship.

. regress QUET AGE

Source	SS	df	MS	Number of obs	=	32
				F(1, 30)	=	54.37
Model	4.93597143	1	4.93597143	Prob > F	=	0.0000
Residual	2.72371329	30	.090790443	R-squared	=	0.6444
				Adj R-squared	=	0.6326
Total	7.65968472	31	.247086604	Root MSE	=	.30131

QUET	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
<hr/>					
AGE	.0573642	.0077799	7.37	0.000	.0414755 .0732529
_cons	.3864519	.4176903	0.93	0.362	-.4665855 1.239489

significant level=0.05 > p-value=0.0000 ,we reject H0: β_1 equal to 0,
so we have enough evidence to refer that QUET and AGE has significant relationship.

. regress SBP AGE

Source	SS	df	MS	Number of obs =	32
<hr/>					
Model	3861.63037	1	3861.63037	Prob > F =	0.0000
Residual	2564.33838	30	85.4779458	R-squared =	0.6009
<hr/>					
Total	6425.96875	31	207.289315	Root MSE =	9.2454

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
<hr/>					
AGE	1.6045	.2387159	6.72	0.000	1.116977 2.092023
_cons	59.09162	12.81626	4.61	0.000	32.91733 85.26592

significant level=0.05 > p-value=0.0000 ,we reject H0: β_1 equal to 0,
so we have enough evidence to refer that SBP and AGE has significant relationship.

↑ 0 ↓ · flag

+ Comment

Walter O. Augenstein · 2 hours ago 

1. SBP(Y) vs. SMK(X)

. regress sbp smk

Source	SS	df	MS	Number of obs =	32
<hr/>					
Model	393.098162	1	393.098162	Prob > F =	0.1723
Residual	6032.87059	30	201.095686	R-squared =	0.0612
<hr/>					
Total	6425.96875	31	207.289315	Root MSE =	14.181

sbp | Coef. Std. Err. t P>|t| [95% Conf. Interval]

	smk	7.023529	5.023498	1.40	0.172	-3.235823	17.28288
_cons		140.8	3.661472	38.45	0.000	133.3223	148.2777

$\text{invFtail}(1, 30, 0.05) = 4.1708768$, but $F = 1.95 \Rightarrow$ we cannot reject the Null hypothesis.

The regression line explains 6% of the total squared variation.

We cannot establish a relationship of sbp to smk.

2. SBP(Y) vs. QUET(x)

. regress sbp quet

Source	SS	df	MS	Number of obs	=	32
				$F(1, 30)$	=	36.75
Model	3537.94585	1	3537.94585		Prob > F	= 0.0000
Residual	2888.0229	30	96.2674299		R-squared	= 0.5506
				Adj R-squared	=	0.5356
Total	6425.96875	31	207.289315		Root MSE	= 9.8116

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
quet	21.49167	3.545147	6.06	0.000	14.25151 28.73182
_cons	70.57641	12.32187	5.73	0.000	45.4118 95.74102

$\text{invFtail}(1, 30, 0.05) = 4.1708768$, and $F = 36.75 \Rightarrow$ we reject the Null hypothesis.

The regression line explains 55% of the total squared variation.

There is a definite linear component to the regression of sbp on quet.

3. QUET(Y) vs. AGE(X)

. regress quet age

Source	SS	df	MS	Number of obs	=	32
				$F(1, 30)$	=	54.37
Model	4.93597216	1	4.93597216		Prob > F	= 0.0000
Residual	2.72371324	30	.090790441		R-squared	= 0.6444
				Adj R-squared	=	0.6326
Total	7.6596854	31	.247086626		Root MSE	= .30131

quet	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.0573642	.0077799	7.37	0.000	.0414755 .0732529
_cons	.3864517	.4176903	0.93	0.362	-.4665857 1.239489

$\text{invFtail}(1, 30, 0.05) = 4.1708768$, and $F = 54.37 \Rightarrow$ we reject the Null hypothesis.

The regression line explains 64% of the total squared variation.
There is a definite linear component to the regression of sbp on age.

4. SBP(Y) vs AGE(X)

. regress sbp age

Source	SS	df	MS	Number of obs	= 32
Model	3861.63037	1	3861.63037	F(1, 30)	= 45.18
Residual	2564.33838	30	85.4779458	Prob > F	= 0.0000
Total	6425.96875	31	207.289315	R-squared	= 0.6009
				Adj R-squared	= 0.5876
				Root MSE	= 9.2454

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	1.6045	.2387159	6.72	0.000	1.116977 2.092023
_cons	59.09162	12.81626	4.61	0.000	32.91733 85.26592

$\text{invFtail}(1, 30, 0.05) = 4.1708768$, and $F = 45.18 \Rightarrow$ we reject the Null hypothesis.

The regression line explains 60% of the total squared variation.

There is a definite linear component to the regression of sbp on age.

↑ 0 ↓ · flag

+ Comment

New post

To ensure a positive and productive discussion, please read our [forum posting policies](#) before posting.

B	I	≡	≡	Link	<code>	Pic	Math		Edit: Rich ▾	Preview

Make this post anonymous to other students

Subscribe to this thread at the same time

Add post

Week Three Homework Assignment

[Subscribe for email updates.](#)

No tags yet. [+ Add Tag](#)

Jesse Booher · a day ago 

Regression 1

. regress sbp smk

Source	SS	df	MS	Number of obs	=	32
Model	393.098162	1	393.098162	Prob > F	=	0.1723
Residual	6032.87059	30	201.095686	R-squared	=	0.0612
Total	6425.96875	31	207.289315	Root MSE	=	14.181

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
smk	7.023529	5.023498	1.40	0.172	-3.235823 17.28288
_cons	140.8	3.661472	38.45	0.000	133.3223 148.2777

The ANOVA output for the first regression shows a weak linear relationships between blood pressure and smoking. A strong relationship would indicate a high F score. While higher than 0, an F score of 1.95 with 1 IV and 30 DF yields a P-Value of 0.172, which is much higher than the 0.05 value we would need to have confidence in this model. Further the R-Square indicates that even if the model were a statistically significant predictor of blood pressure, we would only be able to explain 6% of the variance.

Regression 2

. regress sbp quet

Source	SS	df	MS	Number of obs	=	32
Model	3537.94585	1	3537.94585	Prob > F	=	0.0000
Residual	2888.0229	30	96.2674299	R-squared	=	0.5506
Total	6425.96875	31	207.289315	Root MSE	=	9.8116

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----	-------	-----------	---	------	----------------------

quet	21.49167	3.545147	6.06	0.000	14.25151	28.73182
_cons	70.57641	12.32187	5.73	0.000	45.4118	95.74102

The ANOVA Output for the second regression shows a strong linear relationship between blood pressure and body size. With an F Score of 36.75, 1 IV and 30 DF, we have a P-value much less than 0.05 at 0.000. Coupled with an R-Square score that explains 55% of the variance in blood pressure, we can be confident that the model is a statistically significant predictor of blood pressure.

Regression 3

. regress quet age

Source	SS	df	MS	Number of obs =	32
F(1, 30) = 54.37					
Model	4.93597216	1	4.93597216	Prob > F =	0.0000
Residual	2.72371324	30	.090790441	R-squared =	0.6444
Adj R-squared = 0.6326					
Total	7.6596854	31	.247086626	Root MSE =	.30131

quet	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.0573642	.0077799	7.37	0.000	.0414755 .0732529
_cons	.3864517	.4176903	0.93	0.362	-.4665857 1.239489

The third ANOVA output shows an even stronger linear relationship between body size and age. With an F-Score of 54.37, 1 IV and 30 DF, we have a P-value much less than 0.05 at 0.000. Coupled with an R-Square score that explains 64% of variance in body size, we can be confident that the model is a statistically significant predictor of blood pressure.

Regression 4

. regress sbp age

Source	SS	df	MS	Number of obs =	32
F(1, 30) = 45.18					
Model	3861.63037	1	3861.63037	Prob > F =	0.0000
Residual	2564.33838	30	85.4779458	R-squared =	0.6009
Adj R-squared = 0.5876					
Total	6425.96875	31	207.289315	Root MSE =	9.2454

sdp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	1.6045	.2387159	6.72	0.000	1.116977 2.092023
_cons	59.09162	12.81626	4.61	0.000	32.91733 85.26592

The 4th ANOVA output shows a strong linear relationship between blood pressure and age. With an F-Score of 45.18, 1 IV and 30 DF, we have a P-value much less than 0.05 at 0.000. Coupled with an R-Square that explains 60% of the variance in blood pressure, we can be confident that the model is a statistically significant predictor of age.

As we progress into multiple IVs, I would expect a model with all IVs to show us that age and body size a statistically significant predictors of blood pressure.

↑ 0 ↓ · flag

New post

To ensure a positive and productive discussion, please read our [forum posting policies](#) before posting.

B *I* Link <code> Pic Math Edit: Rich ▾ Preview

Make this post anonymous to other students

Subscribe to this thread at the same time

Add post

Week 3 homework results and analysis

[Subscribe for email updates.](#)

No tags yet. [+ Add Tag](#)

Sort replies by: [Oldest first](#) [Newest first](#) [Most popular](#)



W J Kinder · 5 days ago

SBP (Y) on SMK (X): With n=32, F=1.95 and p=.1723, we fail to reject the null hypothesis. There is not sufficient evidence to conclude that there is a significant linear relationship between SBP and SMK.

SBP (Y) on QUET (X): With n=32, F=36.75 and p=.0000 we reject the null hypothesis and conclude that there is a significant linear relationship between SBP and QUET. The model is SBP= 70.58+21.49*QUET . For each one unit increase in QUET, this model predicts SBP increases 21.49; CI(14.25 to 28.73)

QUET (Y) on AGE (X): With n=32, F= 54.37 and p=.0000 we reject the null hypothesis and conclude that there is a significant linear relationship between QUET and AGE. The model is QUET=.3864+.057*AGE. For every one year increase in age, this model predicts an increase of .057 in QUET, CI(.04 to .07)

SBP (Y) on AGE (X): With n=32, F=45.18 and p=.0000 we reject the null hypothesis and conclude that there is a significant linear relationship between SBP and AGE. The model is SBP= 59.09+1.60*AGE. For each one year increase in age, this model predicts a sbp increase of 1.60, CI(1.11 to 2.09)

Other thoughts:

The dataset included ages from 41 to 65, mean age was 53, so we can not generalize the results to other ages. If a variable for gender would have been included and the sample size was larger, it would be interesting to run the analysis separately by gender to see what differences by gender might exist.

1 · flag

Kathy Padkapayeva · 5 days ago

Nice post, W J Kinder! I'm wondering if professor Lemeshow has the information on gender for this sample, and could share it with us, so that we could see if there are differences by gender here.

Also, it is interesting what is the sum of the percentage of variance in SBP that is explained by the three models (taking the R-squared):

SMK explains about 6% of the variance in SBP

QUET explains about 55% of the variance

AGE explains about 64% of the variance

In total it gets to 125%. If we sum what we have from adjusted R-Squared, it still gets to about 115%. It could mean that I'm wrong in my interpretation of the results, or in my attempt to sum it up. Or maybe it just demonstrates that our independent variables SMK, QUET, and AGE are interrelated themselves? It would be helpful if anyone who is a more mature statistician could provide some explanation on it.

Thank you

↑ 0 ↓ · flag



W J Kinder · 5 days ago



We can't sum up the individual R-squares to get 115%. We can run a multiple regression model using all the variables and then examine the adjusted R square for that multivariable model. When I did that I got a model with adjusted r-square of .73, but QUET was not statistically significant in that model. I think we'll cover that in week5. And you are right the independent variables are correlated with each other and that can create challenges too and I think week 5 will be interesting.

↑ 0 ↓ · flag

+ Comment

New post

To ensure a positive and productive discussion, please read our [forum posting policies](#) before posting.

B	I	≡	≡	Link	<code>	Pic	Math		Edit: Rich ▾	Preview

Make this post anonymous to other students

Subscribe to this thread at the same time

Add post

Week 3 homework thread for R users

[Subscribe for email updates.](#)

[R](#) × [homework](#) × + Add Tag

Sort replies by: [Oldest first](#) [Newest first](#) [Most popular](#)

Jai Broome · 6 days ago

Hi all, I thought I'd create a thread at the start of the week for those of us that are doing the analyses in R. Feel free to post your questions or solutions to this week's assignment

0 · flag

onur miskbay · 6 days ago

Hi folks, I have a question.

Is there a function for showing the summary statistics and the ANOVA table at the same time?

What I would like to have is a nice output like STATA but I need to punch in anova function and summary function seperately for thaat. The anova function does not have any arguments and summary function's arguments did not do what I wanted.

0 · flag

W J Kinder · 5 days ago

This is the R code that I've used so far for week3 homework. It gives same conclusions as STATA but I'd like to improve it.

```
# Read dataset  
mydata<- read.csv("~/A stata MOOC/week3/week3-HW-data.csv")  
# Plot SBP by AGE  
plot(mydata$SBP~mydata$AGE)  
# Same plot with Variable names  
plot(SBP~AGE, data=mydata)  
# Run linear regression model named model1  
model1=lm(SBP~AGE, data=mydata)  
#see what is included in model1  
names(model1)  
#Summarize model1 this gives nice output  
summary(model1)  
# ANOVA for this model  
anova(model1)  
#Show the Confidence interval for slope
```

```
confint(model1)
# Set view for 2 rows and 2 columns for diagnostic graphs
par(mfrow=c(2,2))
#plot diagnostics for regression model1
plot(model1)
#reset view for 1 row 1 column graph
par(mfrow=c(1,1))
#plot residuals for model1
plot (model1$residuals)
#draw horizontal line at zero point
abline(h=0)
```

↑ 0 ↓ · flag

▀ A post was deleted

+ Comment

New post

To ensure a positive and productive discussion, please read our [forum posting policies](#) before posting.

B *I* Link <code> Pic Math Edit: Rich ▾ Preview

Make this post anonymous to other students

Subscribe to this thread at the same time

Add post

Homework Solutions

Applied Regression Analysis

WEEK 3

Exercise One

Determine the ANOVA tables for the following regressions:

For each of the regressions, type “regress” followed by the Y and X variables, into the command box. From this command, you should see an ANOVA table, as well as a t-statistic for the slope coefficient below, and an F-statistic of model fit in the upper right corner (*See images below*)

1. SBP (Y) on SMK (X)

. regress sbp smk					
Source	SS	df	MS	Number of obs = 32	
Model	393.098162	1	393.098162	F(1, 30) =	1.95
Residual	6032.87059	30	201.095686	Prob > F =	0.1723
Total	6425.96875	31	207.289315	R-squared =	0.0612
				Adj R-squared =	0.0299
				Root MSE =	14.181

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
smk	7.023529	5.023498	1.398	0.172	-3.235823 17.28288
_cons	140.8	3.661472	38.454	0.000	133.3223 148.2777

2. SBP (Y) on QUET (X)

. regress sbp quet					
Source	SS	df	MS	Number of obs = 32	
Model	3537.94585	1	3537.94585	F(1, 30) =	36.75
Residual	2888.0229	30	96.2674299	Prob > F =	0.0000
Total	6425.96875	31	207.289315	R-squared =	0.5506
				Adj R-squared =	0.5356
				Root MSE =	9.8116

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
quet	21.49167	3.545147	6.062	0.000	14.25151 28.73182
_cons	70.57641	12.32187	5.728	0.000	45.4118 95.74102

3. QUET (Y) on AGE (X)

```
. regress quet age
```

Source	SS	df	MS	Number of obs	=	32
Model	4.93597216	1	4.93597216	F(1, 30)	=	54.37
Residual	2.72371324	30	.090790441	Prob > F	=	0.0000
Total	7.6596854	31	.247086626	R-squared	=	0.6444
				Adj R-squared	=	0.6326
				Root MSE	=	.30131

quet	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.0573642	.0077799	7.37	0.000	.0414755 .0732529
_cons	.3864517	.4176903	0.93	0.362	-.4665857 1.239489

4. SBP (Y) on AGE (X)

```
. regress sbp age
```

Source	SS	df	MS	Number of obs	=	32
Model	3861.63037	1	3861.63037	F(1, 30)	=	45.18
Residual	2564.33838	30	85.4779458	Prob > F	=	0.0000
Total	6425.96875	31	207.289315	R-squared	=	0.6009
				Adj R-squared	=	0.5876
				Root MSE	=	9.2454

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	1.6045	.2387159	6.72	0.000	1.116977 2.092023
_cons	59.09162	12.81626	4.61	0.000	32.91733 85.26592

Homework Solutions

Applied Regression Analysis

WEEK 3

Exercise Two

Use the ANOVA tables to perform the F test for the significance of each straight-line regression.

For each of the following, we are testing whether the slope coefficient is equal to zero. In other words, we are testing if the independent variable contributes significantly to the model. Notice that, because there is only one independent variable in the model, the F-test for the overall model yields the same significance as the t-test for the slope coefficient.

Looking at SBP (Y) on SMK (X)

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

<code>. regress sbp smk</code>					
Source	SS	df	MS	Number of obs = 32	
Model	393.098162	1	393.098162	F(1, 30) = 1.95	
Residual	6032.87059	30	201.095686	Prob > F = 0.1723	
Total	6425.96875	31	207.289315	R-squared = 0.0612	
<hr/>					
	sbp	Coef.	Std. Err.	t	P> t [95% Conf. Interval]
<hr/>					
smk	7.023529	5.023498	1.398	0.172	-3.235823 17.28288
_cons	140.8	3.661472	38.454	0.000	133.3223 148.2777

With F=1.95 and p=0.1723, we fail to reject the null hypothesis. There is not sufficient evidence to conclude that there is a significant straight-line relationship between SBP and SMK.

Looking at SBP (Y) on QUET (X)

<code>. regress sbp quet</code>					
Source	SS	df	MS	Number of obs = 32	
Model	3537.94585	1	3537.94585	F(1, 30) = 36.75	
Residual	2888.0229	30	96.2674299	Prob > F = 0.0000	
Total	6425.96875	31	207.289315	R-squared = 0.5506	
<hr/>					
	sbp	Coef.	Std. Err.	t	P> t [95% Conf. Interval]
<hr/>					
quet	21.49167	3.545147	6.062	0.000	14.25151 28.73182
_cons	70.57641	12.32187	5.728	0.000	45.4118 95.74102

With F=36.75 and p<0.0001, we reject the null hypothesis. There is sufficient evidence to conclude that there is a significant straight-line relationship between SBP and QUET.

Week 3 Homework Assignment

[Subscribe for email updates.](#)

No tags yet. [+ Add Tag](#)

Sort replies by: [Oldest first](#) [Newest first](#) [Most popular](#)

Amos B Robinson · a month ago 

Effect of Smoking on Systolic Blood Pressure

Number of obs = 32

F(1, 30) = 1.95

Prob > F = 0.1723

Under the Naive Model the blood pressure is 144. So because the "F" statistic is so low, there is not convincing evidence that smoking has a statistical significance on SBP. This is confirmed on the "F" statistic because the "P" value for the "F" statistic is .1723, which is above the critical value of .05 for a 95% confidence level. Additionally, the R-squared for this regression is 0.0612, which further confirms that the statistical impact of smoking on blood pressure is minimal.

Effect of Body Size on Systolic Blood Pressure

Number of obs = 32

F(1, 30) = 36.75

Prob > F = 0.0000

R-squared = 0.5506

Under the Naive Model the blood pressure is 144. Here the "F" statistic is very large and the associated "P" value is < than .0000. So there is convincing evidence that body size has a statistically significant effect on systolic blood pressure. The high coefficient of determination "R-squared" shows that QUET explains about 55% of SBP.

Effect of AGE on Body Size

Number of obs = 32

F(1, 30) = 54.37

Prob > F = 0.0000

R-squared = 0.6444

The mean body size of this sample is 3.44. However, there is convincing evidence that age does effect body size. The "F" statistic is large at 54.37 and the associated "P" value is less than .0000. The R-squared shows that age explains 64.44% of body size.

Effect of Age on Systolic Blood Pressure

Number of obs = 32

F(1, 30) = 45.18

Prob > F = 0.0000

R-squared = 0.6009

The Naive Model for SBP is 144. The "F" statistic is 45.18, which is pretty large. The corresponding "P" value for this "F" statistic is less than .0000. So there is convincing evidence that age does have

an impact on systolic blood pressure.

5 · flag

Kahsay Tadesse · a month ago

well done, great

0 · flag

[+ Comment](#)

Kahsay Tadesse · a month ago

am not clear about the culculation on :"Effect of Body Size on Systolic Blood Press

0 · flag

[+ Comment](#)

KK Wong · a month ago

1. SBP vs SMK

. reg SBP SMK

Source	SS	df	MS	Number of obs	=	32
Model	393.098162	1	393.098162	F(1, 30)	=	1.95
Residual	6032.87059	30	201.095686	Prob > F	=	0.1723
Total	6425.96875	31	207.289315	R-squared	=	0.0612
				Adj R-squared	=	0.0299
				Root MSE	=	14.181

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
SMK	7.023529	5.023498	1.40	0.172	-3.235823 17.28288
_cons	140.8	3.661472	38.45	0.000	133.3223 148.2777

Given relatively small $F(1,30)=1.95$, high $\text{Prob}>F=0.1723$ and low adj R-squared=0.0299, it suggests to reject the hypothesis that SBP has an association with SMK; ie, the association between SBP & SMK is approx 2.99% as suggested by the adj R-squared. It further supports by the fact that SMK 95% CI contains 0 and its $P>|t|=0.172$ which is significantly away from 0.

2. SBP vs QUET

. reg SBP QUET

Source	SS	df	MS	Number of obs	=	32
Model	3537.94574	1	3537.94574	F(1, 30)	=	36.75
Residual	2888.02301	30	96.2674337	Prob > F	=	0.0000
Total	6425.96875	31	207.289315	R-squared	=	0.5506
				Adj R-squared	=	0.5356
				Root MSE	=	9.8116

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
QUET	21.49167	3.545147	6.06	0.000	14.25151 28.73182
_cons	70.5764	12.32187	5.73	0.000	45.41179 95.74101

Given relatively high $F(1,30)=36.75$, significantly low $\text{Prob}>F=0.0000$ and high adj R-squared=0.5356, it suggests not to reject the hypothesis that SBP vs QUET; ie, the association between SBP & QUET is

approx 53.56% as suggested by the adj R-squared. It further supports by the fact that QUET 95% CI does not contains 0 and its P>|t|=0.000.

3. QUET vs AGE

. reg QUET AGE

Source	SS	df	MS	Number of obs	=	32
Model	4.93597143	1	4.93597143	F(1, 30)	=	54.37
Residual	2.72371329	30	.090790443	Prob > F	=	0.0000
Total	7.65968472	31	.247086604	R-squared	=	0.6444
				Adj R-squared	=	0.6326
				Root MSE	=	.30131

QUET	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
AGE	.0573642	.0077799	7.37	0.000	.0414755 .0732529
_cons	.3864519	.4176903	0.93	0.362	-.4665855 1.239489

Given relatively high F(1,30)=54.37, significantly low Prob>F=0.0000 and high adj R-squared=0.6326, it suggests not to reject the hypothesis that QUET vs AGE; ie, the association between QUET & AGE is approx 63.26% as suggested by the adj R-squared. It further supports by the fact that AGE 95% CI does not contains 0 and its P>|t|=0.000.

4. SBP vs AGE

. reg SBP AGE

Source	SS	df	MS	Number of obs	=	32
Model	3861.63037	1	3861.63037	F(1, 30)	=	45.18
Residual	2564.33838	30	85.4779458	Prob > F	=	0.0000
Total	6425.96875	31	207.289315	R-squared	=	0.6009
				Adj R-squared	=	0.5876
				Root MSE	=	9.2454

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
AGE	1.6045	.2387159	6.72	0.000	1.116977 2.092023
_cons	59.09162	12.81626	4.61	0.000	32.91733 85.26592

Given relatively high F(1,30)=45.18, significantly low Prob>F=0.0000 and high adj R-squared=0.5876, it suggests not to reject the hypothesis that SMK vs AGE; ie, the association between SMK & AGE is approx 58.76% as suggested by the adj R-squared. It further supports by the fact that AGE 95% CI does not contains 0 and its P>|t|=0.000.

↑ 2 ↓ · flag

+ Comment



Juan C. Trujillo · a month ago



I am confused about this case. Could there be a possibility in which the F test turns out to be statistically significant, but the t test for the explanatory variable appears with a high p-value?

Please, explain.

↑ 0 ↓ · flag

Anonymous · a month ago



If you are testing the statistical significance of the estimate of a single regression coefficient, the inferential outcome will be the same whether you use the t-test or the F-test.

Mechanically, this is because the F-statistic associated with testing a single coefficient estimate is just squared t-statistic associated with that same estimate.

↑ 1 · flag

+ Comment

Varalakshmi · a month ago 

Amos: I did the F statistic for SBP (Y) on SMK (X)

The Model with 1 d.f parameter because there is one independent variable SMK and the residual which is y and y predicted has 2 d.f one for intercept and slope. As you stated the R squared which is the explained variation to the Total variation is very little. P value so small suggests that there is a significant difference in Smoking history and SBP.

Finding the t-statistic = $r/\sqrt{(1-r^2)/(n-2)}$ = 1.398459 or as is also given in the table below.

We can check that $t^2 = 1.40^2$ is the same as F statistic= 1.95

The lecture was quite informative!

Can someone help me to graph two scatter plots in one - Week 1 homework where we need to graph the residuals as well the given observations?

Source	SS	df	MS	Number
> of obs	= 32			
				F(1,> 30) = 1.95
Model	393.098162	1	393.098162	Prob >> F = 0.1723
Residual	6032.87059	30	201.095686	R-squared = 0.0612
				Adj R-squared = 0.0299
Total	6425.96875	31	207.289315	Root MSE = 14.181

sbp	Coef.	Std. Err.	t	P> t	[95 % Con f. Interval]
smk	7.023529	5.023498	1.40	0.172	-3.2 35823
	17.28288				
_cons	140.8	3.661472	38.45	0.000	133 .3223
	148.2777				

↑ 0 · flag

+ Comment

luca balestrini · a month ago 

Source SS df MS Number of obs = 32

	F(1, 30)	= 1.95			
Model	393.098162	1 393.098162	Prob > F	= 0.1723	
Residual	6032.87059	30 201.095686	R-squared	= 0.0612	
	Adj R-squared			= 0.0299	
Total	6425.96875	31 207.289315	Root MSE	= 14.181	
sbp	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]
smk	7.023529	5.023498	1.40	0.172	-3.235823 17.28288
_cons	140.8	3.661472	38.45	0.000	133.3223 148.2777

2) sbp on quet

Source	SS	df	MS	Number of obs = 32	
		F(1, 30)	= 36.75		
Model	3537.94585	1 3537.94585	Prob > F	= 0.0000	
Residual	2888.0229	30 96.2674299	R-squared	= 0.5506	
	Adj R-squared			= 0.5356	
Total	6425.96875	31 207.289315	Root MSE	= 9.8116	
sbp	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]
quet	21.49167	3.545147	6.06	0.000	14.25151 28.73182
_cons	70.57641	12.32187	5.73	0.000	45.4118 95.74102

3) quet on age

Source	SS	df	MS	Number of obs = 32	
		F(1, 30)	= 54.37		
Model	4.93597216	1 4.93597216	Prob > F	= 0.0000	
Residual	2.72371324	30 .090790441	R-squared	= 0.6444	
	Adj R-squared			= 0.6326	
Total	7.6596854	31 .247086626	Root MSE	= .30131	
quet	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]
age	.0573642	.0077799	7.37	0.000	.0414755 .0732529
_cons	.3864517	.4176903	0.93	0.362	-.4665857 1.239489

4) sbp on age

Source	SS	df	MS	Number of obs = 32	
		F(1, 30)	= 45.18		
Model	3861.63037	1 3861.63037	Prob > F	= 0.0000	
Residual	2564.33838	30 85.4779458	R-squared	= 0.6009	
	Adj R-squared			= 0.5876	
Total	6425.96875	31 207.289315	Root MSE	= 9.2454	
sbp	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]
age	1.6045	.2387159	6.72	0.000	1.116977 2.092023
_cons	59.09162	12.81626	4.61	0.000	32.91733 85.26592

↑ 0 ↓ · flag

+ Comment

David C. Morris · a month ago 

I got the same results as above. I'm not going to repost here. However, I was thinking about the four analyses we did. Two of them make sense to me: 1) Blood pressure (sbp) and Smoking (smk); and 2) Blood pressure (sbp) and Body size (quet). I know we're just doing the homework to learn how to run/interpret anova tables. However, it got me thinking about what 'quet' really is. The description says 'body size' but looking at the values (range from ~2 to ~4.6) I don't have any reference point for that. What is Quet? How is it measured? The reason I bring this up is I was a little surprised by the high correlation between Quet and Age. The range of age in the sample is 41 to 65. The correlation between them is R-squared .64. Looking at the scatter plot, as Age goes up, so does Quet. I would've expected it to taper off toward the older ages since people tend to lose muscle mass and usually weight the older they get. Does anyone know what QUET is? I did a Google search but only found other data sets with no description.

 0  · flag

+ Comment

Emilija Nikolic-Djoric · a month ago 

I think that it is Quetelet's index defined as:

$$\text{QUET} = 100 * (\text{weight}/\text{height}^2)$$

 1  · flag

+ Comment

Emilija Nikolic-Djoric · a month ago 

Week 3-Slide 17-at the bottom n-2 instead n-1?

 0  · flag

+ Comment

ANCA MINCIU · a month ago 

I have done the test, with same results. To avoid re-posting the same information, I would just add one sentence to each interpretation.

1. Effect of Smoking on Systolic Blood Pressure

The confidence interval in this case contains zero, so smoking is not a good predictor for blood pressure.

2. Effect of Body Size on Systolic Blood Pressure

The confidence interval does not contain zero, so the body size influences the systolic blood pressure.

3. Effect of AGE on Body Size

The confidence interval does not contain zero, so age is a very important predictor for body size.

4. Effect of Age on Systolic Blood Pressure

Age is a very important predictor for systolic blood pressure, taking into consideration that the confidence interval does not contain zero.

↑ 1 · flag

+ Comment

Alina Denham · a month ago 

Hello, everybody!

Here are my observations:

- Let's test the null hypothesis (the slope equals zero) for SBP on SMK. $F=1.95 (< F_{.95}=4.20)$ and $p=0.1723 (>.05)$. Therefore, we fail to reject the null hypothesis. This means that we do not have sufficient evidence to prove that there is a significant linear relationship between blood pressure and smoking history.
- Let's test the null hypothesis (the slope equals zero) for SBP on QUET. $F=36.75 (>> F_{.95})$ and $p=0.000 (<0.001)$. Therefore, we reject the null hypothesis. This means that we have sufficient evidence to prove that there is a significant linear relationship between blood pressure and body size.
- Let's test the null hypothesis (the slope equals zero) for QUET on AGE. $F=54.37 (>> F_{.95})$ and $p=0.000 (<0.001)$. Therefore, we reject the null hypothesis. This means that we have sufficient evidence to prove that there is significant linear relationship between body size and age.
- Let's test the null hypothesis (the slope equals zero) for SBP on AGE. $F=45.18 (>> F_{.95})$ and $p=0.000 (<0.001)$. Therefore, we reject the null hypothesis. This means that we have sufficient evidence to prove that there is significant linear relationship between blood pressure and age.

↑ 2 · flag

+ Comment

Lien-yu Yeh · a month ago 

Assume that,

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

According to the rule of test with P-value,

if $\alpha > p\text{-value}$, then we reject H_0 , and if $\alpha < p\text{-value}$, we don't reject H_0 , where α is significant level because $p\text{-value} = \Pr(\text{reject } H_0 \mid H_0 \text{ is true}) = \text{probability of type I error (with sample)}$, when $p\text{-value}$ is large, H_0 probably be true, because there is high probability to make mistake, so we don't reject H_0

when $p\text{-value}$ is small, the probability of type I error is too low, so we reject H_0

I used above concept to test following question, and I assume that significant level is

0.05($\alpha=0.05$)

```
. regress SBP SMK
```

Source	SS	df	MS	Number of obs = 32			
				F(1, 30) = 1.95			
Model	393.098162	1	393.098162	Prob > F = 0.1723			
Residual	6032.87059	30	201.095686	R-squared = 0.0612			
				Adj R-squared = 0.0299			
Total	6425.96875	31	207.289315	Root MSE = 14.18			

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
SMK	7.023529	5.023498	1.40	0.172	-3.235823	17.28288
_cons	140.8	3.661472	38.45	0.000	133.3223	148.2777

**significant level=0.05 < p-value=0.1723 ,we don't reject H0: β_1 equal to 0,
so we don't have enough evidence to refer that SBP and SMK has significant relationship.**

```
. regress SBP QUET
```

Source	SS	df	MS	Number of obs = 32			
				F(1, 30) = 36.75			
Model	3537.94574	1	3537.94574	Prob > F = 0.0000			
Residual	2888.02301	30	96.2674337	R-squared = 0.5506			
				Adj R-squared = 0.5356			
Total	6425.96875	31	207.289315	Root MSE = 9.8116			

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
QUET	21.49167	3.545147	6.06	0.000	14.25151	28.73182
_cons	70.5764	12.32187	5.73	0.000	45.41179	95.74101

**significant level=0.05 > p-value=0.0000 ,we reject H0: β_1 equal to 0,
so we have enough evidence to refer that SBP and QUET has significant relationship.**

```
. regress QUET AGE
```

Source	SS	df	MS	Number of obs = 32			
				F(1, 30) = 54.37			
Model	4.93597143	1	4.93597143	Prob > F = 0.0000			
Residual	2.72371329	30	.090790443	R-squared = 0.6444			
				Adj R-squared = 0.6326			
Total	7.65968472	31	.247086604	Root MSE = .30131			

QUET	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
<hr/>					
AGE	.0573642	.0077799	7.37	0.000	.0414755 .0732529
_cons	.3864519	.4176903	0.93	0.362	-.4665855 1.239489

**significant level=0.05 > p-value=0.0000 ,we reject H0: β_1 equal to 0,
so we have enough evidence to refer that QUET and AGE has significant relationship.**

. regress SBP AGE

Source	SS	df	MS	Number of obs =	32
<hr/>					
Model	3861.63037	1	3861.63037	Prob > F =	0.0000
Residual	2564.33838	30	85.4779458	R-squared =	0.6009
<hr/>					
Total	6425.96875	31	207.289315	Adj R-squared =	0.5876
<hr/>					
Root MSE =	9.2454				

SBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
<hr/>					
AGE	1.6045	.2387159	6.72	0.000	1.116977 2.092023
_cons	59.09162	12.81626	4.61	0.000	32.91733 85.26592

**significant level=0.05 > p-value=0.0000 ,we reject H0: β_1 equal to 0,
so we have enough evidence to refer that SBP and AGE has significant relationship.**

↑ 0 ↓ · flag

+ Comment



Walter O. Augenstein · a month ago



1. SBP(Y) vs. SMK(X)

. regress sbp smk

Source	SS	df	MS	Number of obs =	32
<hr/>					
Model	393.098162	1	393.098162	Prob > F =	0.1723
Residual	6032.87059	30	201.095686	R-squared =	0.0612
<hr/>					
Total	6425.96875	31	207.289315	Adj R-squared =	0.0299
Root MSE =	14.181				

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----	-------	-----------	---	------	----------------------

smk	7.023529	5.023498	1.40	0.172	-3.235823	17.28288
_cons	140.8	3.661472	38.45	0.000	133.3223	148.2777

$\text{invFtail}(1, 30, 0.05) = 4.1708768$, but $F = 1.95 \Rightarrow$ we cannot reject the Null hypothesis.

The regression line explains 6% of the total squared variation.

We cannot establish a relationship of sbp to smk.

2. SBP(Y) vs. QUET(x)

. regress sbp quet

Source	SS	df	MS	Number of obs = 32		
				$F(1, 30) = 36.75$		
Model	3537.94585	1	3537.94585	Prob > F = 0.0000		
Residual	2888.0229	30	96.2674299	R-squared = 0.5506		
				Adj R-squared = 0.5356		
Total	6425.96875	31	207.289315	Root MSE = 9.8116		

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
quet	21.49167	3.545147	6.06	0.000	14.25151 28.73182
_cons	70.57641	12.32187	5.73	0.000	45.4118 95.74102

$\text{invFtail}(1, 30, 0.05) = 4.1708768$, and $F = 36.75 \Rightarrow$ we reject the Null hypothesis.

The regression line explains 55% of the total squared variation.

There is a definite linear component to the regression of sbp on quet.

3. QUET(Y) vs. AGE(X)

. regress quet age

Source	SS	df	MS	Number of obs = 32		
				$F(1, 30) = 54.37$		
Model	4.93597216	1	4.93597216	Prob > F = 0.0000		
Residual	2.72371324	30	.090790441	R-squared = 0.6444		
				Adj R-squared = 0.6326		
Total	7.6596854	31	.247086626	Root MSE = .30131		

quet	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	.0573642	.0077799	7.37	0.000	.0414755 .0732529
_cons	.3864517	.4176903	0.93	0.362	-.4665857 1.239489

$\text{invFtail}(1, 30, 0.05) = 4.1708768$, and $F = 54.37 \Rightarrow$ we reject the Null hypothesis.

The regression line explains 64% of the total squared variation.
There is a definite linear component to the regression of sbp on age.

4. SBP(Y) vs AGE(X)

. regress sbp age

Source	SS	df	MS	Number of obs = 32		
Model	3861.63037	1	3861.63037	$F(1, 30) = 45.18$		
Residual	2564.33838	30	85.4779458	Prob > F = 0.0000		
Total	6425.96875	31	207.289315	R-squared = 0.6009		
				Adj R-squared = 0.5876		
				Root MSE = 9.2454		
<hr/>						
sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	1.6045	.2387159	6.72	0.000	1.116977	2.092023
_cons	59.09162	12.81626	4.61	0.000	32.91733	85.26592

invFtail(1, 30, 0.05) = 4.1708768, and $F = 45.18 \Rightarrow$ we reject the Null hypothesis.

The regression line explains 60% of the total squared variation.

There is a definite linear component to the regression of sbp on age.

↑ 0 ↓ · flag

+ Comment

Erin Dillon · a month ago 

I also got the same results as the responses above and will avoid re-posting the results. Taking this one step further, though, if we know from this homework that age predicts body size (body size increases as people get older) and that both body size and age predict blood pressure, how do we determine if these are both useful predictors? Perhaps body size is irrelevant, except that body size tends to increase as people age. If you maintain a low body size as you get older, will your blood pressure still increase?

If I put both variables in the model at the same time, I get:

. regress sbp quet age

Source	SS	df	MS	Number of obs = 32		
Model	4120.59224	2	2060.29612	$F(2, 29) = 25.92$		
Residual	2305.37651	29	79.4957416	Prob > F = 0.0000		
Total	6425.96875	31	207.289315	R-squared = 0.6412		
				Adj R-squared = 0.6165		
				Root MSE = 8.916		

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
<hr/>					
quet	9.750732	5.402456	1.80	0.081	-1.298531 20.8
age	1.045157	.3860567	2.71	0.011	.2555828 1.834732
_cons	55.32344	12.53475	4.41	0.000	29.687 80.95987

Age remains a significant predictor of blood pressure, but body size falls below the p<.05 mark. So perhaps it's the case that age is the real predictor of blood pressure and both body size and blood pressure increase independently with age.

↑ 0 ↓ · flag

+ Comment

Anonymous · a month ago

why do we call this the Naive Model

↑ 0 ↓ · flag

+ Comment

Anonymous · a month ago

What is the most important factor that determines how the significance to which the independent variable contributes to the model? Is it the p value or the value of the coefficient of independent variable

↑ 0 ↓ · flag

+ Comment

Ronald Ndesanjo · a month ago

. regress SBP QSBT					
Source	SS	df	MS	Number of obs = 32	
Model	3337.34374	1	3337.34374	F:	1, 31
Residual	2688.02318	30	86.2674337	Prob > F	= 0.0000
Total	6025.3675	31	197.118932	R-squared	= 0.1504
				Adj R-squared	= 0.1354
				Root MSE	= 9.1014

. regress QSBT AGE					
Source	SS	df	MS	Number of obs = 32	
Model	4935971.43	1	4935971.43	F:	1, 31
Residual	2727215.29	30	90706.445	Prob > F	= 0.0000
Total	7633686.72	31	247096.634	R-squared	= 0.4524
				Adj R-squared	= 0.4324
				Root MSE	= 302.31

. regress SBP AGE					
Source	SS	df	MS	Number of obs = 32	
Model	3861.43037	1	3861.43037	F:	1, 31
Residual	2644.33037	30	85.1779458	Prob > F	= 0.0000
Total	6505.76074	31	209.52024	R-squared	= 0.1409
				Adj R-squared	= 0.1374
				Root MSE	= 9.2454

. regress SBP AGE					
Source	SS	df	MS	Number of obs = 32	
Model	1.0405	1	1.0405	F:	1, 31
Residual	59.09162	31	1.90458	Prob > F	= 0.0000
Total	60.13162	32	1.8754	R-squared	= 0.0033
				Adj R-squared	= 0.0000
				Root MSE	= 2.0000

(a) The regression does not suggest a strong relationship between smoking and blood pressure ($F=1.95$). Therefore, we don't have evidence to reject the null hypothesis.

(b) The regression indicates a strong influence of body size on blood pressure ($F=36.75$). In other words, as one body grows so is chance for blood pressure. There is evidence to reject the null

hypothesis therefore.

(c) The regression show that with increasing age so is the body size ($F=54.37$). We therefore have evidence to reject the null hypothesis.

(d) Age has got influence on bloods pressure ($F=45.18$). We have evidence to reject the null hypothesis.

↑ 0 ↓ · flag

+ Comment

Sveta Kochergina · 21 days ago 

1) SBP (Y) on SMK (X)

$$F(1, 30) = 1.95$$

that's why β_1 is probably equal to zero. We fail to reject the null hypothesis. There is not sufficient evidence that smoking has a statistically significant effect on SBP

2) SBP (Y) on QUET (X)

$$F(1, 30) = 36.75$$

the F-ratio gets too large, so it's likely that β_1 is not equal to zero. We reject the null hypothesis. There is sufficient evidence that body size has a statistically significant effect on SBP

3) QUET (Y) on AGE (X)

$$F(1, 30) = 54.37$$

the F-ratio gets too large, so it's likely that β_1 is not equal to zero. We reject the null hypothesis. There is sufficient evidence that age has an impact on the body size

4) SBP (Y) on AGE (X)

$$F(1, 30) = 45.18$$

the F-ratio gets too large, so it's likely that β_1 is not equal to zero. We reject the null hypothesis. There is sufficient evidence that age has an impact on SBP

↑ 0 ↓ · flag

+ Comment

Michele Svanera · 16 days ago 

Great explanation from Amos B Robinson! (I avoid to repeat answers)

↑ 0 ↓ · flag

+ Comment

New post

To ensure a positive and productive discussion, please read our [forum posting policies](#) before posting.

B*I*

:=

:=

Link

<code>

Pic

Math

Edit: Rich ▾

Preview

- Make this post anonymous to other students
- Subscribe to this thread at the same time

Add post

Week Four Homework Exercise

[Help Center](#)[Homework Central](#) / Week Four Homework /

The following table represents results from an environmental engineering study of a certain chemical reaction. The concentrations of 18 separately prepared solutions were recorded at different times (three measurements at each of six times). The natural logarithms of the concentrations were also computed. The data are given in the table below.

You will use the data from this table to complete the homework exercises for the week. Scroll down below the table to view those exercises. Use the [homework forum](#) to discuss the exercises with your peers and share your discoveries.

If you need help entering data into STATA, please download our tutorial on [how to enter datasets into STATA](#).

Solution Number <i>(i)</i>	Time (X_i) (hrs)	Concentration (Y_i) (mg/ml)	Ln of Concentration ($\ln Y_i$)
1	6	0.029	-3.540
2	6	0.032	-3.442
3	6	0.027	-3.612
4	8	0.079	-2.538
5	8	0.072	-2.631
6	8	0.088	-2.430
7	10	0.181	-1.709
8	10	0.165	-1.802
9	10	0.201	-1.604
10	12	0.425	-0.856
11	12	0.384	-0.957
12	12	0.472	-0.751
13	14	1.130	0.122
14	14	1.020	0.020
15	14	1.249	0.222
16	16	2.812	1.034
17	16	2.465	0.902
18	16	3.099	1.131

If desired, you may download the data for this exercise in this [CSV file](#)

Use Stata to complete the following exercises:

Exercise One

Generate separate graphs of:

1. Concentration (Y) vs. Time (X)
2. Natural Logarithm of Concentration ($\ln Y$) vs. Time (X)

Exercise Two

Equations and Plotting

Using the output from exercise one, obtain the following:

1. The estimated equation of the straight-line (degree 1) regression of Y on X
2. The estimated equation of the quadratic (degree 2) regression of Y on X
3. The estimated equation of the straight-line (degree 1) regression of $\ln Y$ on X
4. Plots of each of these fitted equations on their respective scatter diagrams.

Exercise Three

Determine and Compare

Determine and compare the proportions of the total variation in Y explained by the straight-line regression on X and by the quadratic regression on X

Exercise Four

F-Tests

1. Carry out F-tests for the significance of the straight-line regression of Y on X
2. Carry out an overall F-test for the significance of the quadratic regression of Y on X and a test for the significance of the addition of x^2 to the model
3. For the straight-line regression of $\ln Y$ on X , carry out F-tests for the significance of the overall regression

Exercise Five

Determine and Compare

- What proportion of the variation in $\ln Y$ is explained by the straight-line regression of $\ln Y$ on X ?
- Compare this result with that obtained in Exercise Three for the quadratic regression of $\ln Y$ on X . Discuss this in the [homework forum](#)

Exercise Six

Examine and Discuss

Use the [homework forum](#) to explain your thoughts on the following:

- A fundamental assumption in regression analysis is variance homoscedasticity. By examining the scatter diagrams constructed in Exercises One & Two, state why taking natural logarithms of the concentrations helps with regard to the assumption of variance homogeneity.
- Do you think the straight-line regression of $\ln Y$ on X is better for describing this set of data than the quadratic regression of Y on X ? Share your thoughts in the [homework forum](#).
- Considering the overall table, what key assumption about the data would be in question if, instead of 18 different solutions, there were only 3 different solutions, each of which was analyzed at the six different time points?

You can download all homework assignments for week 4 [here](#)

For help and answers to this week's exercises, click the button below to visit the solutions page.

[Solutions Page](#)

Created Tue 10 Mar 2015 8:17 AM PDT

Last Modified Mon 13 Apr 2015 6:33 AM PDT

Homework Solutions

Applied Regression Analysis

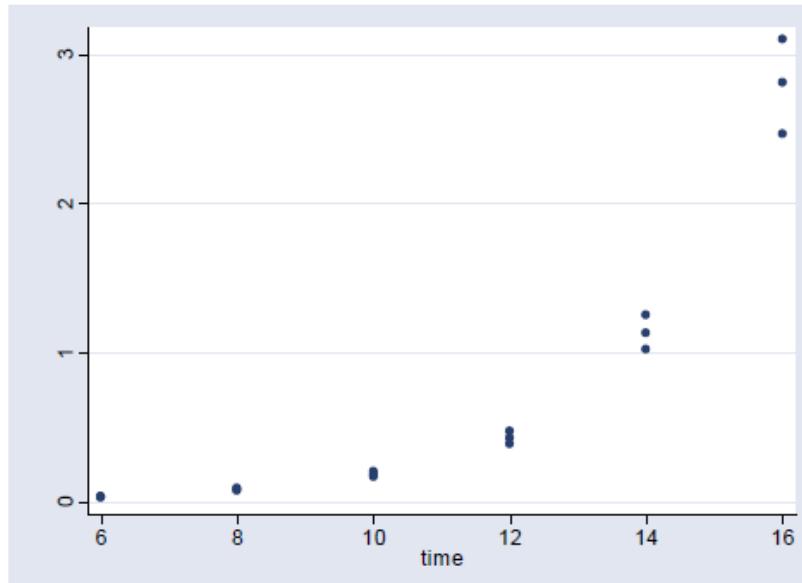
WEEK 4

Exercise One: Generate separate graphs of:

1. Concentration (Y) vs. Time (X)

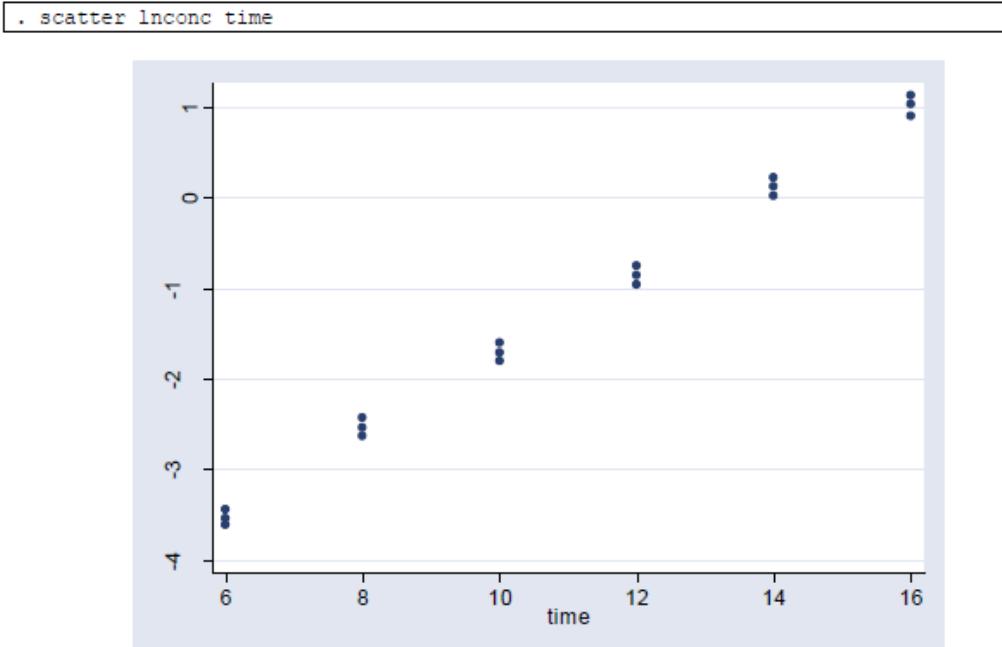
For this question, we are looking for a scatterplot of Concentration vs Time. Type “scatter concen time” into the command window. From this command, you should see the graph shown below:

```
. scatter concen time
```



2. Natural Logarithm of Concentration (lnY) vs. Time (X)

To generate a scatterplot of the log transformation of Concentration vs Time, type “scatter lnconc time” into the command window. From this command, you should see the graph shown below:



As you can see from the graphs above, the log transformation of concentration appears to “linearize” the model.

Homework Solutions

Applied Regression Analysis

WEEK 4

Exercise Two: Equations and Plotting

Using the output from exercise one, obtain the following:

1. The estimated equation of the straight-line (degree 1) regression of Y on X

To generate an estimated equation, we must fit the regression in order to obtain the least squares estimates by typing “regress” into the command window, followed by the dependent and independent variables. From the output, you can obtain the coefficient for the slope (β_1) as well as the intercept (β_0) in the bottom right corner of the output in the “Coef.” column. You can then substitute these values into an equation for the estimated straight line regression (see output below).

After you fit the regression, type “predict yhat” into the command window in order to generate fitted values for Y under the given model. This will generate a new variable called “yhat” (Note: as you continue to fit several regressions, you must change the name of “yhat” for each regression (eg. predict yhat2, predict yhat3, predict yhat4....)).

Otherwise, you will get an error as you have already generated a variable under the same name). Graph the fitted regression line onto the scatterplot of the data by typing “scatter yhat” followed by the independent and independent variables, then type “,connect (1.) symbol (i o)”. This latter command will connect the yhat to form a regression line, while plotting the data as dots.

```
. regress concen time

      Source |       SS          df       MS           Number of obs =      18
-----+---- Model |  12.7054079         1  12.7054079          F( 1,    16) =   43.70
Residual |  4.65200556        16  .290750347          Prob > F =  0.0000
-----+---- Total |  17.3574134        17  1.02102432          R-squared =  0.7320
                               Adj R-squared =  0.7152
                               Root MSE =  .53921

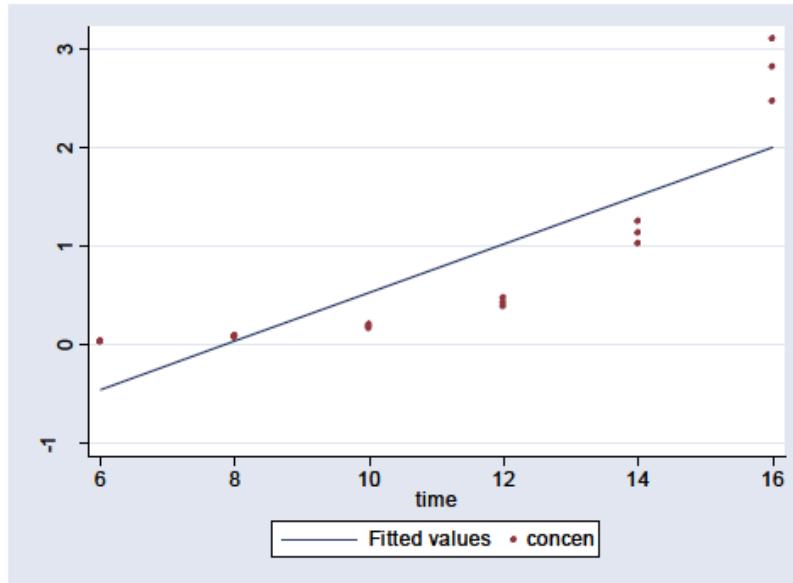
-----+
-----+---- concen |     Coef.    Std. Err.          t      P>|t|      [95% Conf. Interval]
-----+---- time |  .2459714   .0372092       6.610   0.000      .1670914   .3248514
_cons | -1.931797  .4285795      -4.507   0.000     -2.840345  -1.023249
-----+
```

Estimated equation of the straight-line regression:

$$y = -1.931797 + 0.245714 * \text{time}$$

```
. predict yhat
(option xb assumed; fitted values)
```

```
. scatter yhat concen time, connect(1 .)symbol(i o)
```



2. The estimated equation of the quadratic (degree 2) regression of Y on X

Recall that quadratic regressions include both X and X^2 into the model. Thus, you must first generate a variable for X^2 by typing “gen time2=time^2” into the command box. You can then fit the quadratic regression by typing “regress concen time time2” into the command box. Proceed to then substitute these both slope coefficients into the estimated equation for the straight line regression, making sure to include time² as the second variable (see output below).

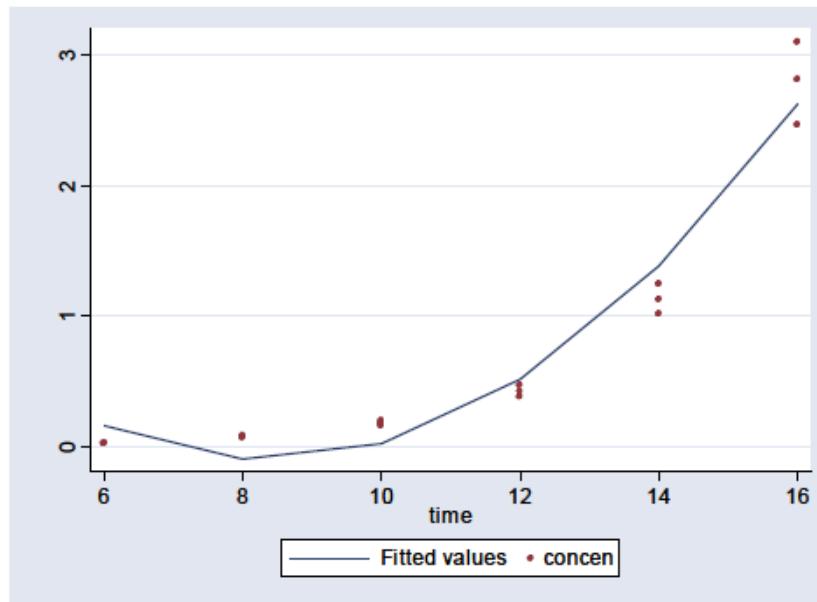
. gen time2=time^2						
. regress concen time time2						
Source SS df MS						Number of obs = 18
Model 16.6104749 2 8.30523743						F(2, 15) = 166.79
Residual .74693859 15 .049795906						Prob > F = 0.0000
Total 17.3574134 17 1.02102432						R-squared = 0.9570
						Adj R-squared = 0.9512
						Root MSE = .22315
concen Coef. Std. Err. t P> t [95% Conf. Interval]						
time -.7810226 .1169892 -6.676 0.000 -1.030379 -.5316661						
time2 .0466815 .0052714 8.856 0.000 .0354458 .0579173						
_cons 3.172052 .6030163 5.260 0.000 1.886754 4.457351						

Estimated equation of the quadratic regression:

$$y = 3.172052 - 0.7810226 \cdot \text{time} + 0.0466815 \cdot \text{time}^2$$

```
. predict yhat2
(option xb assumed: fitted values)

. scatter yhat2 concen time, connect(l .) symbol(i o)
```



alternatively,

```
. twoway (scatter concen time, sort) (qfit yhat2 time, sort)
```

3. The estimated equation of the straight-line (degree 1) regression of lnY on X

```
. regress lnconc time

      Source |       SS          df          MS
-----+---- Model |  42.7523927        1   42.7523927
Residual | .150140289       16   .009383768
-----+---- Total |  42.902533       17   2.52367841

      Number of obs =      18
      F( 1,    16) = 4555.99
      Prob > F = 0.0000
      R-squared = 0.9965
      Adj R-squared = 0.9963
      Root MSE = .09687

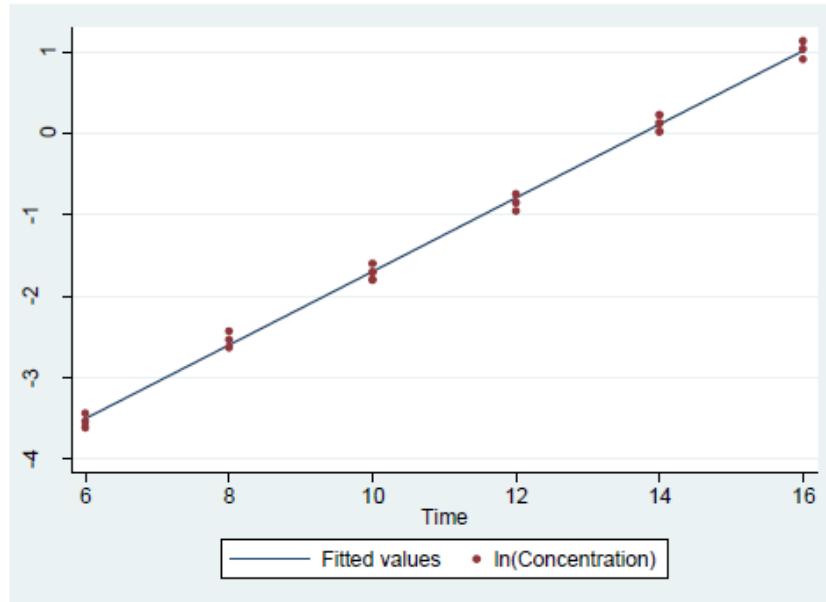
      lnconc |     Coef.    Std. Err.      t    P>|t|      [95% Conf. Interval]
-----+---- time |  .4512015  .0066847    67.498  0.000      .4370307  .4653724
_cons | -6.209981  .0769945   -80.655  0.000     -6.373202  -6.04676
```

Estimated equation of the straight-line regression:

$$\ln(y) = -6.209981 + 0.4512015 * \text{time}$$

```
. predict yhat3
(option xb assumed; fitted values)

. scatter yhat3 lnconc time, connect(l .) symbol(i o)
```



4. Plots of each of these fitted equations on their respective scatter diagrams.

Each scatter plots are located as an image in each of the 3 estimated equations above

Homework Solutions

Applied Regression Analysis

WEEK 4

Exercise Three: Determine and Compare

Determine and compare the proportions of the total variation in Y explained by the straight-line regression on X and by the quadratic regression on X

We can determine the proportion of the total variation in Y explained by the regression models by the R^2 value within the STATA output of each fitted regression.

$$R_{\text{straight}}^2 = 0.73 \quad R_{x^2}^2 = 0.957$$

The quadratic regression explains more of the total variation in Y than the straight-line regression.

Interpretation: 95.7% of the total variation in Y is explained by the quadratic regression on X, whereas only 73% of the total variation in Y is explained by the straight line regression on X.

Homework Solutions

Applied Regression Analysis

WEEK 4

Exercise Four: F-Tests

1. Carry out F-tests for the significance of the straight-line regression of Y on X

$$F_{\text{regression}} = \frac{SS_{\text{reg}} / df}{SS_{\text{res}} / df} = \frac{12.7}{4.65 / 16} = 43.7, \text{ p-value} < 0.001$$

We reject the null ($\beta_1=0$); there is evidence of a significant linear relationship between Y and X.

2. Carry out an overall F-test for the significance of the quadratic regression of Y on X and a test for the significance of the addition of x^2 to the model

For this question, we must calculate the overall F-statistic, as well as the partial F-statistic for the addition of x^2 :

$$\text{Overall Regression } F_{\text{regression}} = \frac{SS_{\text{reg}} / df}{SS_{\text{res}} / df} = \frac{16.6 / 2}{0.75 / 15} = 166.8, \text{ p-value} < 0.001$$

$$\text{Test for addition of } X^2 F_{\text{add } X^2} = \frac{SS_{X|X^2} / df}{SS_{\text{res}} / df} = \frac{3.9}{0.75 / 15} = 79.1, \text{ p-value} < 0.001$$

The overall F-test ($p<0.001$) indicates that we can reject the null ($H_0: \beta_1=\beta_2=0$) and conclude that the quadratic regression of Y on X is significant.

The partial F-test ($p<0.001$) indicates that we can reject the null ($H_0: \beta_2=0$) and conclude that x^2 contributes significantly to the model.

3. For the straight-line regression of $\ln Y$ on X, carry out F-tests for the significance of the overall regression

$$F_{\text{regression}} = \frac{SS_{\text{reg}} / df}{SS_{\text{res}} / df} = \frac{42.75}{0.15 / 16} = 4556, \text{ p-value} < 0.001$$

We reject the null ($\beta_1=0$); there is evidence of a significant linear relationship between $\ln Y$ and X.

Homework Solutions

Applied Regression Analysis

WEEK 4

Exercise Five: Determine and Compare

- What proportion of the variation in $\ln Y$ is explained by the straight-line regression of $\ln Y$ on X ?

Again, we can determine the proportion of the total variation in $\ln Y$ explained by the regression models by the R^2 value within the STATA output of the fitted regression.

$$R^2_{X \text{ on } \ln(Y)} = 0.9965$$

- Compare this result with that obtained in Exercise Three for the quadratic regression of $\ln Y$ on X .

Discuss this in the homework forum

Week Five Homework Exercise

[Help Center](#)[Homework Central](#) / Week Five Homework /

Exercise One

Earlier in the course we studied the multiple regression relationship of SBP (Y) to AGE (X_1), SMK (X_2), and QUET (X_3) using the data in Homework 1 of Week 2.

If you need to refer to the dataset from Week 2 homework, you can download the [CSV file](#).

Three regression models will now be considered:

Model	Independent Variables Used
1	AGE (X_1)
2	AGE (X_1), SMK (X_2),
3	AGE (X_1), SMK (X_2), QUET (X_3)

First, use your computer to generate each of the above models.

Then, complete the following:

A. Use model 3 to determine:

1. What is the predicted SBP for a 50-year old smoker with a quetelet (QUET) index of 3.5?
2. What is the predicted SBP for a 50-year-old non-smoker with a quetelet index of 3.5?
3. For 50-year-old smokers, give an estimate of the change in SBP corresponding to an increase in quetelet index from 3.0 to 3.5.

B. Using the ANOVA tables, compute and compare the R^2 -values for models 1,2, and 3.

C. Conduct (separately) the overall F -tests for significant regression under models 1,2, and 3. Be sure to state your null hypothesis for each model in terms of regression coefficients.

Exercise Two

The accompanying table presents the weight (X_1), age (X_2) and plasma lipid levels of total cholesterol (Y) for a hypothetical sample of 25 patients with hyperlipoproteinemia before drug therapy.

Patient	Total Cholesterol (Y)	Weight (X_1)	Age (X_2)
	(mg/100 ml)	(kg)	(yr)
1	354	84	46
2	190	73	20
3	405	65	52
4	263	70	30
5	451	76	57
6	302	69	25
7	288	63	28
8	385	72	36
9	402	79	57
10	365	75	44
11	209	27	24
12	290	89	31
13	346	65	52
14	254	57	23
15	395	59	60
16	434	69	48
17	220	60	34
18	374	79	51
19	308	75	50
20	220	82	34
21	311	59	46
22	181	67	23
23	274	85	37
24	303	55	40
25	244	63	30

If desired, you may download the data for this exercise in this [CSV file](#)

Complete the following six questions:

1. Generate the separate straight-line regressions of Y on X_1 (model 1) and Y on X_2 (model 2).

2. Generate the regression model of Y on both X_1 and X_2
3. For each of the models in questions 1 and 2, determine the predicted cholesterol level \hat{Y} for patient 4 (with $Y = 263$, $X_1 = 70$, and $X_2 = 30$) and compare these predicted cholesterol levels with the observed value. *Comment on your findings in the [homework forum](#).*
4. Carry out the overall F -test for the two-variable model and the partial F -test for the addition of X_1 to the model, given that X_2 is already in the model.
5. Compute and compare the R^2 -values for each of the three models considered in questions 1 and 2.
6. Based on the results obtained in questions 1-5, what do you consider to be the best predictive model involving either one or both of the independent variables considered? Why? *Discuss your response in the [homework forum](#).*

You can download all homework assignments for week 5 [here](#)

For help and answers to this week's exercises, click the button below to visit the solutions page.

[Solutions Page](#)

Created Tue 10 Mar 2015 8:17 AM PDT

Last Modified Sun 19 Apr 2015 9:32 PM PDT

Homework Solutions

Applied Regression Analysis

WEEK 5

Exercise One

Earlier in the course we studied the multiple regression relationship of SBP (Y) to AGE (X_1), SMK (X_2), and QUET (X_3) using the data in Homework 1 of Week 2. Please refer to the dataset from Week 2 homework if you need to.

Three regression models were considered:

Model	Independent Variables Used
1	AGE (X_1)
2	AGE (X_1), SMK (X_2),
3	AGE (X_1), SMK (X_2), QUET (X_3)

First, use your computer to generate each of the above models.

In order to fit the three models, you have to consider each model separately.

Model 1

Type ‘regress sbp age’ in the command window. From the output, you can obtain the coefficient for the slope (β_1) as well as the intercept (β_0) in the bottom right corner of the output in the “Coef.” column.

Model 1

. regress sbp age					
Source	SS	df	MS	Number of obs = 32	
Model	3861.63038	1	3861.63038	F(1, 30) = 45.18	
Residual	2564.33838	30	85.4779458	Prob > F = 0.0000	
Total	6425.96875	31	207.289315	R-squared = 0.6009	
				Adj R-squared = 0.5876	
				Root MSE = 9.2454	
sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	1.6045	.2387159	6.721	0.000	1.116977 2.092023
_cons	59.09162	12.81626	4.611	0.000	32.91733 85.26592

Model 2

Type ‘regress sbp age smk’ in the command window. From the output, you can obtain the coefficient for β_1 and β_2 as well as the intercept (β_0) in the bottom right corner of the output in the “Coef.” column.

Model 2

. regress sbp age smk					
Source	SS	df	MS	Number of obs = 32	
Model	4689.68423	2	2344.84211	F(2, 29) = 39.16	
Residual	1736.28452	29	59.87188	Prob > F = 0.0000	
Total	6425.96875	31	207.289315	R-squared = 0.7298	
				Adj R-squared = 0.7112	
				Root MSE = 7.7377	

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	1.70916	.2017587	8.471	0.000	1.296517 2.121803
smk	10.29439	2.768107	3.719	0.001	4.632978 15.95581
_cons	48.0496	11.12956	4.317	0.000	25.2871 70.81211

Model 3

Type ‘regress sbp age smk quet’ in the command window. From the output, you can obtain the coefficient for β_1 , β_2 and β_3 as well as the intercept (β_0) in the bottom right corner of the output in the “Coef.” column.

Model 3

. regress sbp age smk quet					
Source	SS	df	MS	Number of obs = 32	
Model	4889.82567	3	1629.94189	F(3, 28) = 29.71	
Residual	1536.14308	28	54.8622529	Prob > F = 0.0000	
Total	6425.96875	31	207.289315	R-squared = 0.7609	
				Adj R-squared = 0.7353	
				Root MSE = 7.4069	

sbp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	1.212715	.3238192	3.745	0.001	.549401 1.876028
smk	9.945568	2.656057	3.744	0.001	4.504882 15.38625
quet	8.592448	4.498681	1.910	0.066	-.6226827 17.80758
_cons	45.10319	10.76488	4.190	0.000	23.05235 67.15404

Then, complete the following:

A. Use model 3 for the following:

- (1) What is the predicted SBP for a 50-year old smoker with a quetelet (QUET) index of 3.5?

The regression equation for the third model is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \text{age} + \hat{\beta}_2 \text{smk} + \hat{\beta}_3 \text{quet}$$

substituting the different value of the predictor variables, we get the predicted value of SBP

for AGE=50, SMK=1 and QUET=3.5

$$\begin{aligned} y &= 45.103 + 1.2127(50) + 9.945568(1) + 8.592448(3.5) \\ &= 145.76 \end{aligned}$$

- (2) What is the predicted SBP for a 50-year-old non-smoker with a quetelet index of 3.5?

Using the same regression equation again

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \text{age} + \hat{\beta}_2 \text{smk} + \hat{\beta}_3 \text{quet}$$

and substituting the value of the predictors, we get the new predicted value of the SBP

$$\begin{aligned} \text{for AGE=50, SMK=0 and QUET=3.5} \\ y &= 45.103 + 1.2127(50) + 8.592448(3.5) \\ &= 135.81 \end{aligned}$$

- (3) For 50-year-old smokers, give an estimate of the change in SBP corresponding to an increase in quetelet index from 3.0 to 3.5.

Making use of the regression equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \text{age} + \hat{\beta}_2 \text{smk} + \hat{\beta}_3 \text{quet}$ again and just changing the value of QUET from 3.5 to 3.0 from the previous question, we will get the predicted value of the SBP. We will then make use of this and take the difference between the previous value and the present value of the predicted SBP to obtain the change due to increase in QUET.

$$\begin{aligned} \text{for AGE=50, SMK=1 and QUET=3.0} \\ y &= 45.103 + 1.2127(50) + 9.945568(1) + 8.592448(3.0) \\ &= 141.46 \end{aligned}$$

Thus the change is SBP = 145.76 - 141.46 = 4.30

- B. Using the ANOVA tables, compute and compare the R^2 -values for models 1, 2, and 3.

Now we will make use of the three outputs of the regression models that we had obtained earlier. Make sure that you match the correct output to the models. The value of R^2 for each of the model can be obtained from the fourth line of the right hand side (RHS) of the outputs.

Model	Independent Variables Used	R^2
1	AGE (X_1)	0.6009
2	AGE (X_1), SMK (X_2),	0.7298
3	AGE (X_1), SMK (X_2), QUET (X_3)	0.7609

The R^2 value increases with each variable added to the model.

- C. Conduct (separately) the overall F tests for significant regression under models 1,2, and 3. Be sure to state your null hypothesis for each model in terms of regression coefficients.

In order to conduct the overall F test, we essentially check to see the null hypothesis that the slope coefficients simultaneously equal to zero. The value of the F statistic for each model can be obtained from the second line of the RHS of the output. The corresponding p-value can be obtained from the third line of the RHS of the output.

Model	Independent Variables Used	H_0	F	p
1	AGE (X_1)	$\beta_{age}=0$	45.18	<0.001
2	AGE (X_1), SMK (X_2),	$\beta_{age}=\beta_{smk}=0$	39.16	<0.001
3	AGE (X_1), SMK (X_2), QUET (X_3)	$\beta_{age}=\beta_{smk}=\beta_{quet}=0$	29.71	<0.001

Homework Solutions

Applied Regression Analysis

WEEK 5

Exercise Two

Complete the following six questions:

1. Generate the separate straight-line regressions of Y on X_1 (model 1) and Y on X_2 (model 2). Which of the two independent variables would you say is the more important predictor of Y ? *Discuss your response in the homework forum.*

We will consider the two simple linear regression models separately. Type 'regress choles weight' in the command window.

Model 1

. regress choles weight					
Source	SS	df	MS	Number of obs = 25	
Model	10231.7262	1	10231.7262	F(1, 23) = 1.74	
Residual	135145.314	23	5875.88321	Prob > F = 0.2000	
				R-squared = 0.0704	
				Adj R-squared = 0.0300	
Total	145377.04	24	6057.37667	Root MSE = 76.654	
choles	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
weight	1.622343	1.229433	1.320	0.200	-.9209323 4.165618
_cons	199.2975	85.81792	2.322	0.029	21.76962 376.8254

Type 'regress choles age' in the command window.

Model 2

. regress choles age					
Source	SS	df	MS	Number of obs = 25	
Model	101932.666	1	101932.666	F(1, 23) = 53.96	
Residual	43444.3743	23	1888.88584	Prob > F = 0.0000	
Total	145377.04	24	6057.37667	R-squared = 0.7012	
				Adj R-squared = 0.6882	
				Root MSE = 43.461	
choles	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	5.320676	.7242909	7.346	0.000	3.822367 6.818986
_cons	102.5751	29.63757	3.461	0.002	41.26516 163.8851

Age is a more important predictor. See the R^2 and the F test.

While making your decision regarding the independent variable, you need to take into consideration the p-value and the R^2 value. *Please discuss your response in the homework forum.*

- Generate the regression model of Y on both X_1 and X_2 .

We will now consider a multiple linear regression model. Type ‘regress choles weight age’ in the command window. From the output, you can obtain the coefficient for β_1 and β_2 as well as the intercept (β_0) in the bottom right corner of the output in the “Coef.” column.

. regress choles weight age					
Source	SS	df	MS	Number of obs = 25 F(2, 22) = 26.36 Prob > F = 0.0000 R-squared = 0.7056 Adj R-squared = 0.6788 Root MSE = 44.111	
Model 102570.815	2	51285.4073			
Residual 42806.2253	22	1945.73752			
Total 145377.04	24	6057.37667			

choles	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
weight	.4173621	.7287761	0.573	0.573	-1.094027 1.928751
age	5.216591	.7572445	6.889	0.000	3.646162 6.78702
_cons	77.98254	52.42964	1.487	0.151	-30.74988 186.715

- For each of the models in questions 1 and 2, determine the predicted cholesterol level (Y) for patient 4 (with $Y=263$, $X_1=70$, and $X_2=30$) and compare these predicted cholesterol levels with the observed value. *Comment on your findings in the homework forum.*

There are three models in total. Two simple linear regression models from question 1 and the multiple linear regression model that we just fit. We have the intercept and slope coefficients from the outputs. Recall that in order to obtain the predicted value we just substitute the value of the predictor variables in the regression equation.

Please discuss your response in the homework forum.

$$Y=263, X_1 = 70 \text{ and } X_2 = 30$$

$$\text{Model 1: } y=199.298+1.62234\text{WEIGT}$$

$$\begin{aligned} &=199.298+1.62234(70) \\ &=312.8618 \end{aligned}$$

$$\text{Model 2: } y=102.575+5.32068\text{AGE}$$

$$\begin{aligned} &=102.575+5.32068(30) \\ &=262.1954 \end{aligned}$$

$$\text{Model 3: } y=77.9825+5.21659\text{AGE}+0.41736\text{WEIGHT}$$

$$\begin{aligned} &=77.9825+5.21659(30)+0.41736(70) \\ &=263.695 \end{aligned}$$

Models 2 and 3 yield predictions very close to the observed cholesterol value of 263 while Model 1 provides a very poor prediction. Model 3 is the closest to the observed value.

4. Carry out the overall F test for the two-variable model and the partial F test for the addition of X_1 to the model, given that X_2 is already in the model.

For this question we will consider the multiple linear regression model. To carry out the overall F test, we will test if the null hypothesis that β_1 and β_2 are simultaneously equal to zero. The F statistic and the p-value for the same can be obtained from the RHS of the output from question 2.

Overall F-Test

$$H_0: \beta_{x1} = \beta_{x2} = 0$$

$$H_A: \text{At least one of the } \beta's \neq 0$$

$$F=26.36, \text{ p-value}<0.001 \text{ (from computer output)}$$

Reject the null hypothesis. There is significant overall regression.

Partial F-Test for the addition of X_1 given that X_2 is already in the model

H_0 : The addition of X_1 (Weight) to the model does not significantly improve the prediction of Cholesterol over and above that achieved by the model containing X_2 (Age).

H_A : The addition of X_1 adds to the prediction of Cholesterol

$$F(x_1 | x_2) = \frac{SS_{reg}(x_1, x_2) - SS_{reg}(x_2)}{MS_{residual}(x_1, x_2)} = \frac{102571 - 101933}{1945.74} = 0.3279, \text{ with 1,22 d.f.}$$

\therefore Not Significant, Fail to reject the null hypothesis. The addition of X_1 to the model already containing X_2 does not add to the prediction of cholesterol.

5. Compute and compare the R^2 -values for each of the three models considered in questions 1 and 2.

This question is similar to question B of Problem 1. We will make use of the three outputs of the regression models that we had obtained earlier. Make sure that you match the correct output to the models. The value of R^2 for each of the model can be obtained from the fourth line of the right hand side (RHS) of the outputs.

Model	R^2
1: WEIGHT	0.0704
2: AGE	0.7012
3: WEIGHT and AGE	0.7056

6. While concluding that a particular model is the best amongst the ones that you have fit, make sure you take into consideration the R^2 value and the corresponding p-value of the model. Please discuss your response in the homework forum.

Week Six Homework Exercise

[Help Center](#)[Homework Central](#) / Week Six Homework /

In a study at the Ohio State University Medical Center, doctors investigated the association between ICU acquired paresis (ICUAP) (or muscular weakness) and subject's admission SOFA score dichotomized at 11 (*sofa11*) (0 for less than 11 and 1 for 11 and greater).

Specifically the investigators wanted to determine if handgrip strength (*max_grip*) could be used as a surrogate for ICUAP. The investigators believed that the number of days the patient was mechanically ventilated (*MV_days*) could affect the relationship between weakness and dichotomized SOFA.

Use the **ICU acquired weakness** dataset to answer the following exercises. You can also find this data within this [CSV file](#).

Exercise One

Using a two-sample t-test, determine whether the two SOFA Score groups differ significantly with respect to ICUAP (as measured by *max_grip*).

Exercise Two

Perform the same analysis from Exercise One (i.e., 2 sample t-test) using a regression analysis.

Exercise Three

Using a single model determine the regression of *max_grip* on *MVdays* for each of the two SOFA score groups.

Exercise Four

Test whether the two SOFA cohorts differ significantly, controlling for mechanically ventilated days.

Exercise Five

Determine and Compare

Test whether the two SOFA cohorts have the same slope. Do they have the same intercept? Are they coincident? Use the [homework forum](#) to share your findings.

You can download all homework assignments for week 6 [here](#)

For help and answers to this week's exercises, click the button below to visit the solutions page.

[Solutions Page](#)

Created Tue 10 Mar 2015 8:17 AM PDT

Last Modified Tue 28 Apr 2015 7:32 AM PDT

Homework Solutions

Applied Regression Analysis

WEEK 6

Exercise One

Using a two-sample t-test, determine whether the two SOFA Score groups differ significantly with respect to ICUAP (as measured by *max_grip*).

Type “`ttest max_grip, by(sofa11)`” in the command window to get the output shown below. The “`by(sofa11)`” command is used to specify the two groups on which we are conducting the t-test.

```
. ttest max_grip, by(sofa11)

Two-sample t test with equal variances
-----+-----+-----+-----+-----+-----+
 Group |     Obs      Mean    Std. Err.    Std. Dev.   [95% Conf. Interval]
 -----+-----+-----+-----+-----+-----+
 0 |     125     15.14    1.241118    13.87612    12.68348    17.59652
 1 |      12     6.166667    2.138512    7.408022    1.459834    10.8735
 -----+-----+-----+-----+-----+-----+
combined |     137    14.35401    1.166686    13.65571    12.04682    16.66121
 -----+-----+-----+-----+-----+-----+
 diff |          8.973333    4.069572           .924972    17.02169
-----+-----+-----+-----+-----+-----+
diff = mean(0) - mean(1)                      t =      2.2050
Ho: diff = 0                                     degrees of freedom =      135
Ha: diff < 0          Ha: diff != 0          Ha: diff > 0
Pr(T < t) = 0.9854          Pr(|T| > |t|) = 0.0291          Pr(T > t) = 0.0146
```

The two sample t-test indicates that the two groups differ significantly ($p = .0291$) with respect to *max_grip*. In fact, we see that the mean strength of patients with SOFA scores less than 11 is 15.14 whereas it's only 6.7 for patients with SOFA scores greater than or equal to 11.

NOTE: We used the two sided alternative hypothesis test.

Homework Solutions

Applied Regression Analysis

WEEK 6

Exercise Two

Perform the same analysis from Exercise One (i.e., 2 sample t-test) using a regression analysis.

In order to perform a regression analysis on the data set, type “`regress max_grip sofall`” in the command window.

Obtain the p-value from the RHS (right hand side) of the output below or from the column next to the t statistic in the table.

p-value of 0.0291 not only indicates that the model is significantly different from the naïve model, but it also indicates that the two groups of SOFA scores differ significantly with respect to `max_grip`.

<code>. regress max_grip sofall</code>						
Source	SS	df	MS		Number of obs	= 137
Model	881.613625	1	881.613625		F(1, 135)	= 4.86
Residual	24479.4667	135	181.329383		Prob > F	= 0.0291
Total	25361.0803	136	186.478532		R-squared	= 0.0348
					Adj R-squared	= 0.0276
					Root MSE	= 13.466

max_grip	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sofall	-8.973333	4.069572	-2.20	0.029	-17.02169	-.924972
_cons	15.14	1.204423	12.57	0.000	12.75802	17.52198

Homework Solutions

Applied Regression Analysis

WEEK 6

Exercise Three

Using a single model determine the regression of *max_grip* on *MVdays* for each of the two SOFA score groups.

To use a single regression model, we need to include the interaction term.

Type “gen MVxSOF = MVdays*sofa11” in the command window to obtain the interaction term.

Now that we have the interaction term, we can use the regress command. Type “regress max_grip sofa11 MVdays MVxSOF” in the command window to obtain the regression output.

```
. gen MVxSOF= MVdays*sofa11

. regress max_grip sofa11 MVdays MVxSOF

      Source |       SS          df       MS
-----+-----
      Model |  3297.36312        3  1099.12104
      Residual |  22063.7172     133   165.89261
-----+-----
      Total |  25361.0803     136   186.478532

      Number of obs =      137
      F(  3,    133) =      6.63
      Prob > F      =  0.0003
      R-squared      =  0.1300
      Adj R-squared =  0.1104
      Root MSE       =    12.88

      max_grip |     Coef.    Std. Err.      t    P>|t| [95% Conf. Interval]
-----+-----
      sofa11 |  -10.30809   6.97879    -1.48   0.142   -24.11187   3.495683
      MVdays |  -.6937794  .1864576    -3.72   0.000   -1.062585  -.3249735
      MVxSOF |   .4083585  .3851152     1.06   0.291   -.3533845  1.170101
      _cons |   21.18421  1.991451    10.64   0.000    17.24519  25.12322
-----+-----

. regress max_grip sofa11 MVdays

      Source |       SS          df       MS
-----+-----
      Model |  3110.84167        2  1555.42084
      Residual |  22250.2386     134   166.046557
-----+-----
      Total |  25361.0803     136   186.478532

      Number of obs =      137
      F(  2,    134) =      9.37
      Prob > F      =  0.0002
      R-squared      =  0.1227
      Adj R-squared =  0.1096
      Root MSE       =    12.886

      max_grip |     Coef.    Std. Err.      t    P>|t| [95% Conf. Interval]
-----+-----
      sofa11 |  -4.315677  4.096519    -1.05   0.294   -12.41788   3.786524
      MVdays |  -.5980555  .1632222    -3.66   0.000   -.9208806  -.2752305
      _cons |   20.35026  1.830419    11.12   0.000    16.73001  23.97051
-----+-----
```

From this output, we see that the interaction between SOFA score group and *MVdays* is not significant so we can use the following model that assumes the groups have the same slope: $\text{max_grip} = 20.35026 - 4.32 \times \text{sofa11} - .60 \times \text{MVdays}$. So, for $\text{sofa11}=0$, $\text{max_grip} = 20.35026 - .60 \times \text{MVdays}$ and, for $\text{sofa11}=1$, $\text{max_grip} = 20.35026 - 4.32 \times 1 - .60 \times \text{MVdays} = 16.03 - .60 \times \text{MVdays}$.

Homework Solutions

Applied Regression Analysis

WEEK 6

Exercise Four

Test whether the two SOFA cohorts differ significantly, controlling for mechanically ventilated days.

From exercise three we know that the interaction term is not significant. Hence, we will fit a model without the interaction term. Type “regress max_grip SOFA MVdays” in the command window. Your output should look like the second part of the output in the previous exercise.

From the output, we see that, after controlling for *MVdays*, there is no difference between the SOFA groups (i.e., $p = .294$). This is the p-value of the t-test which can be obtained from the column next to the t statistic in the table.

Exercise Five

Test whether the two SOFA cohorts have the same slope. Do they have the same intercept? Are they coincident?

From the same output that we used exercise three, we see that, for the model containing sofa11 MVdays and MVxSOF, the interaction term is not significant, so we cannot reject $H_0: \beta_3 = 0$ so we conclude that the lines are parallel. From the model containing sofa11 and *MVdays* (the output that we used for the previous exercise) we see that we cannot reject $H_0: \beta_2 = 0$ as the p-value for its t test is greater than 0.05, so the lines have the same intercept. Therefore we conclude that the lines are coincident.

Working with large data files

Stata requires that the data file you want to analyze fits into memory. This means that working with files approaching the size of memory on your computer can be a challenge. Fortunately, Stata has supplied a number of nice tools for dealing with large data files. We review them here.

describe using

Sometimes you may just want to see what variables are in the large file. You don't need to **use** the entire file just to see a list of variables and their labels. Instead, you can type

```
describe using "bigfile.dta"
```

where "bigfile.dta" is that name of the file you want to describe. Stata will give you all the information about the variables that you would expect from the **describe** command. Ideally, you'll be able to select a subset of variables, or a subset of observations, just by looking at **describe**.

lookfor and lookfor_all

If the big file has a lot of variables, the **describe using** command will give you a lot of text to search. The **lookfor** command will search the variable names and labels for any character string you supply and list the variable names/labels containing that string. If you have several files to search, try **lookfor_all**. This command is available from the SSC archives. It searches through all Stata data files in the current directory (and its subdirectories if you ask for it) for any string you want to find. The string may be in the variable name or label. For example, you may want to find the variable containing the sampling weight, so you try searching for the string "weight". First, change directories (**cd**) to the directory containing the file or files you want to search, then "lookfor" the string:

```
cd "c:\big_file_directory"  
lookfor_all weight, subdir
```

The command lists the name of each file containing that string along with the names of all the variables containing that string in their name or label. It then gives you a clickable link to each file with a match. This command has lots of nice features. See **help lookfor_all** at CPC, or you can download it to your standalone computer with **ssc install lookfor_all**.

use list_of_variables using

You can bring a subset of variables from **bigfile.dta** into memory using this form of the **use** command:

```
use list_of_variables using "bigfile.dta"
```

After looking at the results of **describe using** or **lookfor**, decide which variables you need for your analysis, and list them in the **use** command.

use in

You can bring in a small sample of observations from a large file with this version of the command:

```
use in 1/20 using "bigfile.dta"
```

This allows you to look at a sample of the variables more carefully, perhaps learning more than you could glean from the **describe** command.

use if

Suppose you're only interested in studying people in a certain age range.

```
use if age >= 1 & age < 5 using "bigfile.dta"
```

Of course, you can combine any or all of these features in the same command.

random sample

You might want to test your model on a small number of observations. Selecting those observations randomly can help you get a somewhat more representative set than selecting those from the beginning of the file, for example. You can use the **runiform** function to select any percent of observations you choose. The function returns a value between 0 and 1, so to get a 10% sample, you might use observations when runiform returns values between 0 and 0.1, or any other range of length 0.1, like this:

```
use if inrange(runiform(),0,.1) using "bigfile.dta"
```

Review Again?

Another topic?

A simple program

Writing a simple program

The term "program" has a special meaning in Stata. It is a set of commands that starts with **program define** and ends with **end**. In between you can put any of the commands you're used to using in Stata, and you can use many other commands that are specific to programs. These are discussed in detail in the PDF document "Stata Programming Reference Manual".

The purpose of this page is to give you a simple example of a program and to show you the power of a program to save you time and effort.

Suppose that you want to add household income and assets, variables in the household-level data, to each individual in the household. Using the methods discussed in One-to-many merging you would merge the household data onto the individual data using houseid as the merge key. Now let's suppose you want to do this for 50 countries to look at time and regional differences. It would be cumbersome to copy and paste the code 50 times. But more importantly, if you decided to add some code, you'd have to add it 50 times.

Instead, you can write a program and run it 50 times. Each time you run it you only need to change the name of the country and the year of the survey. If you want to add a command, you only add it once, and it is automatically run on all 50 countries when you run the program. Here's an example:

```
capture program drop mergedata
program define mergedata
    use "c:/data/`1'/'2'/individual.dta",clear
    merge m:1 houseid using "c:/data/`1'/'2'/household.dta", ///
        keepusing(houseid hhincome assets)
    drop if _merge == 2 // households without individuals in sample
    drop _merge
    save "c:/data/`1'/'2'/merged.dta", replace
end

mergedata Peru 2000
mergedata Peru 2005
mergedata "Costa Rica" 2001
mergedata "Costa Rica" 2004
mergedata Brazil 1998
mergedata Brazil 2003
mergedata Brazil 2008
(etc.)
```

Questions:

1. Which part is the "program"? Answer.

2. What does "capture program drop mergedata" do? Answer.

3. There are numbers "1" and "2" scattered around the program. What do they do? Answer.

4. So how do I send values to these local macros `1' and `2'? Answer.

5. Why does "Costa Rica" have quotes around it? Answer.

6. The slashes ("/") in the paths are backward from the way I usually type them in Windows. Is that necessary? Answer.

7. How to I run a program? Answer.

8. What if I have an error in my program? How do I see what values went into `1' and `2'? Answer.

Answers:

1. Each command between and including **program define mergedata** and **end** is part of the program. Here we've chosen to name the program "mergedata", but you can name it almost anything. It's best to avoid names that are already taken, like "merge". You can check whether **help** turns up a command when selecting a name for your program, and then you'll know you need to try a different name.

[Back to question](#)

2. The **capture** command allows you to give a command that might otherwise fail and continue anyway. In this case, the command is **program drop mergedata**. If there is no program in memory already called "mergedata", the command to drop the program from memory will fail. By capturing the error message from that situation, you can continue working.

[Back to question](#)

3. The numbers `1' and `2' are called "local macros". They are temporary variables that hold values you send to them. The first value you send goes into `1' and the second value you send goes into `2'. You can have many more local macros numbered `3', `4', etc. if you need them.

Notice that these numbers have a backward apostrophe to their left side. This character is on the key in the upper left corner of the English keyboard along with the tilde (~) character. The character to the right of the number is an apostrophe (also called "single quote"), which shares a key with the quote ("").

This is one way to send values to a program, but not the only way. Others are discussed in the programming reference.

[Back to question](#)

4. The lines starting with **mergedata Peru 2000** that come after **end** are the commands to run the program. The first word is the new command **mergedata** which we defined above it. Following the command are the two values we want to send to the local macros `1' and `2'.

[Back to question](#)

5. "Costa Rica" has quotes because it has a space in the middle. We want **mergedata** to understand that the value going into local macro `1' is two words. Without the quotes, "Costa" will go into `1', "Rica" will go into `2', and 2001 will have nowhere to go because we didn't put a `3' in our program. [Back to question](#).

6. Stata doesn't normally care which way you type the slashes. In this case, though, where we have a local macro in the file path, we must type the slashes as shown in the example.

[Back to question](#)

7. To run your program you need to first define it to Stata, that is highlight and execute the commands from **capture program drop** through **end**. You will see the commands echoed in the Results Window, but you won't otherwise get any indication from Stata that it has stored your program in memory. Next, you can highlight and execute one line at a time that calls the program (in this case the commands like **mergedata Peru 2000**) so you can check the results one merge at a time. Or, if you're feeling very self-confident, you can execute all 50 of them at once!

[Back to question](#)

8. Debugging a program can be tricky. The best tool available is **set trace on** in combination with **set tracedepth**

```
set tracedepth 1  
set trace on
```

set trace on shows each of your commands followed by the values that were substituted in your local macros. It takes a minute to figure out how to read this, but it's your best tool and well worth that minute of staring at it.

set tracedepth 1 (its smallest value) allows you to see only the substitutions of values in your program. If you forget to set tracedepth, you'll see deep down into all the commands you're calling as the substitutions scroll by in the Results Window. It's slow and not particularly useful.

Put the **set tracedepth** command anywhere before the **set trace** command. You can put the **set trace** command in front of the program, or in front of your first line that calls the program, to have it apply to all lines of the program. If you have a long program and know approximately where your problem is, you can turn trace on for just the one or two lines of code that you think have an error, then turn it off again like this:

```
set trace on  
    merge m:1 houseid using ...  
set trace off
```

That way you have fewer lines of code to sift through in your Results Window.

Once you figure out how to fix your program, remember to **set trace off** so you don't have to look at the extra stuff scrolling by in the Results Window.

[Back to question](#)

[Review again?](#)

[Another topic?](#)

Adding summary statistics to a data file

Adding a statistic to each observation, reducing the file to summary statistics.

Stata has commands that allow you to either add a summary statistic to each observation in memory, or to reduce the file according to values of a group so that each resulting observation is the summary statistic for that group.

```
clear
use "q:\utilities\statatut\exampfac.dta"

/* Add the mean facility age to each observation. */

egen mage= mean(age)
su mage
list facid age mage in 1/5

/* Create a file containing the mean facility age by authority. */

collapse (mean) age, by(authorit)
list
```

Stata offers a large number of statistics with both the egen and collapse commands. See the manual or the on-line help for a full list.

Questions:

1. The **egen** command adds a new variable, in this case called `mage`, to every observation in the data. What happens to the original observations and variables? Answer.
2. The name "egen" stands for "extensions to generate." What's the difference between **generate** and **egen**? Answer.
3. The **collapse** command calculated the mean age for each value of authority. How many observations and variables did the resulting data file in memory contain? Answer.

Answers:

1. The original observations and variables are unchanged. The egen command simply adds another column to the data in memory.

[Back to question](#)

2. The **generate** command has functions, such as "log" below, to create unique values on each observation. The **egen** command has a different set of functions. Some of its functions put unique values on each observation, while others put summary statistics across all observations (or groups) on each observation.

For example, the following generate command would calculate the natural log of the age of each facility in exampfac.dta and add that value to each facility's record:

```
generate lage= log(age)
```

In contrast, the following egen command would calculate the median age of all facilities of each type (authorit), and it would add that value to each facility's record:

```
sort authorit  
by authorit: egen medage= median(age)
```

By the way, you can use **bysort** to combine the above two commands and reduce your typing:

```
bysort authorit: egen medage= median(age)
```

See the manual or online help for **generate**, **egen**, and **functions**

for more information.

[Back to question](#)

3. After the collapse command, the resulting file had 13 observations and 2 variables. The number of observations is determined by the number of distinct values in the by variable. The number of variables is: one for each summary statistic calculated, and one for each by variable.

[Back to question](#)

Review again?

Another topic?

Analyzing data from sample surveys

A sample survey is conducted to obtain information about the characteristics of a population. To reduce the cost and time necessary to collect the data, this task is often handled by selecting a subset (a sample) from the target population of interest to the researchers. The sample design adds certain characteristics to the data that may bias the analysis. The general term for this potential bias is the "sample survey design effect." Methods are available to adjust the analysis to account for some of these characteristics. There should be variables for each observation in your data set to identify each of the characteristics that are described next.

Stata has a well-developed and straightforward set of commands that allow you to adjust your analysis for the sample survey design effect. This section of the tutorial discusses how to use these survey data commands. The discussion is divided into:

- Data Characteristics
- Choosing the Correct Weight Syntax
- Commands to Analyze Survey Data
- Logistic Regression Analysis
- Common Errors and How to Avoid Them

Most of the information in this section on analyzing survey data was provided by Kim Chantala. However, please direct questions and comments to Phil Bardsley as noted below.

Another topic?

Appending data files

Adding observations with the same variables

The need to append doesn't arise often in surveys. Usually it is needed during data entry when new questionnaires arrive in the survey office in batches, are entered, and the resulting files must be combined with files from previous batches of surveys.

For this example we'll simulate the situation in which a new batch of questionnaires has been entered. Facility data from 19 of 20 regions have been entered and combined in one file, named fac19.dta. We've received the remaining facilities' data and need to append it to the original 19.

Copy these commands into a do-file editor and run them.

```
/* Use the original facility file with 19 regions */
clear
use "q:\utilities\statatut\fac19.dta"
tabulate region

/* Append the file with one remaining region (number 7) */

append using "q:\utilities\statatut\newfac.dta"
tabulate region
```

Questions:

1. Why don't we need to sort before appending? Answer.
2. Several notes appear in the Stata log about labels already being defined. Is this a problem? Answer.
3. When using append, the two files should have identical variable names. How would I know if, by mistake, a variable in one file had a different name from a variable in the other file being appended? Answer.

Answers:

1. We don't sort before appending because we do not need to match on any identifiers. We use append when we want to add new observations, so no matching is involved.

[Back to question](#)

2. No, these notes do not indicate a problem with the append. They mean that the file in memory has label definitions with the same names as those in the "using" file on disk. We expect that, since both files have identical variables and, therefore, identical labels defined for their values.

[Back to question](#)

3. The only way to catch this situation is to **describe** both files before the append. They should have the same number of variables. After the append command, the resulting file should also have that number of variables. If the files have the same number before but one more variable after the append, you know that a variable in one of the input files had a different name.

[Back to question](#)

Review again?

Another topic?

Article

Labor-Saving Techniques

Stata offers several techniques that will save you time and reduce the chances that you'll make a mistake in your code out of sheer boredom. The techniques discussed here are:

- Looping Over Variables and Values
- Dummy variables
- A Simple Program

Changing data values

Replacing values, recoding variables, editing data, labelling data.

If you haven't read about relational, logical, and arithmetic operators in the previous page on Groups and Subsets of Data, [click here](#) to read a brief summary.

We'll continue using the 1999 Tanzania Facility Survey data.

```
clear
use "q:\utilities\statatut\exampfac.dta"

/* as a precaution, make a copy of the variable you want to recode */
gen factype2=factype

/* change 8 values of factype2 into 4 */

replace factype2= 1 if factype2 >= 2 & factype2 <= 4
replace factype2= 2 if factype2 == 5
replace factype2= 3 if factype2 == 6
replace factype2= 4 if factype2 == 7 | factype2 == 9

/* the recode command does this more efficiently */

/* recode factype2 2/4=1 5=2 6=3 7 9=4 */

/* change a variable's name */

rename factype2 type

/* give the new variable some labels */

label variable type "Recoded facility type"
/* Click here for more on the label variable command */

label define fac 1 "Hospital" 2 "HealthCenter" 3 "Dispensary" 4 "Other"
label values type fac
/* Click here for more on the label value command */

/* the data file already has a label, but this is how it was done: */

label data "1999 Tanzania Facility Survey"

/* delete the original factype variable */

drop factype

/* delete observations with missing facility type */

drop if missing(type)
browse
edit
```

Questions:

1. The **replace** command changes specific values in an existing variable. What would happen if you reversed the order of these replace commands?

```
replace factype2= 4 if factype2 == 7 | factype2 == 9  
replace factype2= 3 if factype2 == 6  
replace factype2= 2 if factype2 == 5  
replace factype2= 1 if factype2 >= 2 & factype2 <= 4
```

Answer.

2. What happens to missing values of factype2? Answer.
3. How many variable names can you change with one **rename** command? Answer.
4. The **label values** command assigns a set of labels to the values of one variable. If I have several variables needing the same value labels, can I define one label and assign it to all the variables? Answer.
5. How can I change existing variable and value labels? Answer.
6. What if I want to **drop** most of my variables? That's a lot of typing! Answer.
7. What is the difference between **browse** and **edit**? Answer.

Answers:

1. All observations would have a value of 1 for factype2. Remember that each Stata command is executed for every observation before the next command is executed. The **recode** command makes all these changes in a single command, so the order of the changes doesn't matter.

[Back to question.](#)

2. We didn't change the missing values in either the replace or recode examples, so they remain missing.

[Back to question.](#)

3. Only one. Later we'll see how to rename or make just about any other change to a lot of variables easily using the **foreach** command.

Back to question.

4. Yes, for example the last 10 variables, pill-natural, have a yes/no answer format:

```
label def yesno 1 "Yes" 2 "No"
label val pill yesno
label val inject yesno
etc.
label val natural yesno
```

Back to question.

5. To change a variable label, simply type a new **label var** statement:

```
label var factype "Type of facility"
```

To change a value label, you have two choices:

- drop the existing label and recreate it with the changes, or
- create a new label and reassign the variable to the new label.

The **describe** command shows you which variables have value labels attached to them and the names of the labels. Using that command, we see that the variable urbrur has the label urb. Here's how to drop urb and recreate it with new values:

```
label drop urb
label def urb 1 "RURAL" 2 "URBAN" 3 "MIXED"
label val urbrur urb
```

Back to question.

6. Sometimes it's easier to use the **keep** or the **keep if** command than the drop or drop if command.

Back to question.

7. Both commands display the data in spreadsheet format. The edit command allows you to make changes directly in the data, just as you would in a spreadsheet. The browse command allows you to view the data but not to make changes.

Note: the edit command does make entries in the Stata log, but they are not a very useful record of the changes you've made. The best record to keep of data editing uses a case identifier, but the Edit command uses _n, which is relative to the current sort order of the data in memory. **Because Edit leaves you with no record of what you've done, we recommend that you never use it.**

Back to question.

Review again?

Another topic?

Choosing the Correct Weight Syntax

One of the most common mistakes made when analyzing data from sample surveys is specifying an incorrect type of weight for the sampling weights. Only one of the four weight keywords provided by Stata, pweight, is correct to use for sampling weights. The purpose of each type of weight follows.

Sampling or Probability weights: pweight

Stata has a special term **pweight** to specify probability weights. Probability weights are another name for sampling weights. The pweight option causes Stata to use the sampling weight as the number of subjects in the population that each observation represents when computing estimates such as proportions, means, and regressions parameters. A robust variance estimation technique will automatically be used to adjust for the design characteristics so that variances, standard errors and confidence intervals are correct.

Run the following commands to demonstrate the difference between unweighted and weighted results, and to see that Stata automatically uses the robust estimation technique when you use pweights. As before, these data are from the 1999 Tanzania DHS women's survey. Here we predict having 0-2 kids using a woman's education and controlling for her age.

```
clear
use "q:\utilities\statatut\svysamp.dta"
logit twokids age educat
logit twokids age educat [pweight=sampwt]
logit twokids age educat [pweight=sampwt], robust cluster(earea)
```

Note that the coefficient associated with education changes only slightly with the use of pweights. However, the standard error increases quite a bit, with a corresponding decrease in z. Most notably, p increases from .001 to .030 with the addition of weights. If we had not included probability weights, we would have assigned too much importance to the role education plays in the number of children these women have. Adding the cluster(earea) option makes only a slight adjustment in this case, but is recommended.

The discussion below of other weight commands is included as general information. In most cases, these commands are ***NOT APPROPRIATE for use with sample survey data.***

Frequency Weights: fweight

Frequency weights are integers that indicate the number of times the observation was actually observed. It is used when your data set has been collapsed and contains a variable that tells the frequency each record occurred. For example, if the original data was:

x1	x2	y
16	3	1
16	3	1
19	2	0
19	2	0
19	2	0

and the estimation command would be

```
logit y x1 x2
```

then the collapsed data would look like:

x1	x2	y	count
16	3	1	2
19	2	0	3

and the estimation command would be

```
logit y x1 x2 [fweight=count]
```

Do not use fweights to specify sampling weights. Your variance of estimates, p-values and standard errors will be computed incorrectly.

Analytic Weights: aweight

Analytic weights are used when you want to compute a linear regression on data that are observed means. For example, instead of having data that looks like:

group	x	y
1	3	22
1	4	30
2	8	25
2	2	19
2	5	16

suppose the data has been condensed with only the averages being available:

group	x	y	n
1	3.5	26.0	2
2	5.0	20.0	3

and a linear regression could be done by using the command:

```
regress y x [aweight=n]
```

Do not use aweights to specify sampling weights. This is because the formulas that use aweights assume that larger weights designate more accurately measured observations. Conversely, one observation from a sample survey is no more accurately measured than any other observation. Hence, using the aweight command to specify sampling weights will cause Stata to estimate incorrect values of the variance and standard errors of estimates, and p-values for hypothesis tests.

Importance Weights: iweight

Stata has a special weight command, iweight, which can be used by programmers who need to implement their own analytical techniques by using some of the available estimation commands. Special care should be taken when using importance weights to understand how they are used in the formulas for estimates and variance. This information is available in the Methods and Formulas section in the Stata manual for each estimation command. In general, these formulas will be incorrect for computing the variance for data from a sample survey.

Review again?

Another topic?

Combining data files

Introduction to merging

Merging is the process of adding variables from a permanent file on disk to the data in memory. The observations in the two files may be on the same level, for example, they may both be from surveys of the same people who were interviewed at different times. Or, the observations may be from surveys on different levels, such as mothers and their children. In either case, the files have one or more identifying variables in common.

Many people confuse **merge** with **append**. Append combines two files with completely different observations but the same variables, while merge combines files with the same or related observations but different variables. The **append** command is explained fully in a later example.

The **merge** command is simple to use, but there are a wide variety of situations, and corresponding pitfalls, associated with merging. The following three examples cover the more common situations.

One-to-one merging

Match merging

One-to-many merging

Caution (if using a version of Stata prior to 12)! Merging and appending both add data to the data already in Stata's memory. It is easy to ask Stata to put more data in memory than you have allowed room for. Add together the sizes of all the files you want to merge or append before you combine them, **clear** and **set memory** if necessary, then combine the files. If not, you may get the message "No room to add more variables/observations."

Another topic?

Commands to Analyze Survey Data

Stata provides two ways to analyze survey data. After a description of the two ways, there is a table to help you decide which one to choose.

The survey Commands

The preferred way is to use the family of commands that begin with **svy:**. (See **help survey** in Stata for a list of commands that can be run after svy:) These commands were designed especially for analyzing data from sample surveys. Before any of the survey estimation commands can be used, the **svyset** command should be used to specify one or more of the variables that describe the stratification, sampling weight, and/or primary sampling unit variables. You can try **svyset** by running the following commands:

```
clear
use "q:\utilities\statatut\svysamp.dta"
svyset earea [pweight=sampwt], strata(urbrur)
```

In this example from the 1999 Tanzania DHS data, the variable *earea* ("enumeration area") is the PSU, *sampwt* is the probability weight, and *urbrur* (urban-rural) is the stratum identifier.

These values stay in effect until they are cleared or reset. If you save the data, these values are saved with the data and will be in effect the next time you use the data file.

We could now use any of the survey estimation commands. For example, the mean of a variable from the data set could be estimated as follows:

```
svy: mean numkids
(running mean on estimation sample)
Survey: Mean estimation

Number of strata =      2          Number of obs     =    4029
Number of PSUs   =     176          Population size  =    4029
                                         Design df        =     174

+-----+
|           Linearized
|   Mean   Std. Err.    [95% Conf. Interval]
+-----+
numkids |  2.409699   .0617263    2.28787    2.531528
+-----+
```

Stata first reports the names of variables that were defined with the svyset command and some statistics about the data used in the computation. It is a good idea to make sure the names of the variables, the number of strata and the number of PSU's reported are correct. The number of observations with non-missing data (4029) and the size of the population represented by the observations (4029) are also reported (these weights are normalized).

After any of the survey estimation commands, you can use the **test** command to test linear hypotheses and **lincom** to compute linear combinations of estimations. These special commands adjust the test statistics properly for the sample design. For example, to test whether urban women have fewer children than rural women:

```
svy: mean numkids, over(urbrur)
test [numkids]Urban = [numkids]Rural
```

Subpopulation Analysis

When using the svy commands to analyze only a portion of the sample (a sub-population), it is important to analyze the entire data set and to use the **subpop** option to identify those observations you want included in the estimate. This is because Stata needs to have information from every observation in the sample to compute the variance, standard error, and confidence intervals even though only the observations in the sub-sample are needed to compute means, proportions, and regression coefficients.

To use the subpop option, you need to generate a variable that has a value of 1 for the observations in your sub-population and a value of 0 for those that should be excluded. Here is an example where we compute the mean of numkids for the people living on Zanzibar (the variable zanzibar has a value of 1):

```
svy, subpop(zanzibar): mean numkids // CORRECT SUBPOPULATION ANALYSIS

(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      2          Number of obs     =    4029
Number of PSUs   =     176          Population size  =  4.0e+09
                                         Subpop. no. obs =      969
                                         Subpop. size    =  1.0e+08
                                         Design df       =      174

-----+
|           Linearized
|   Mean   Std. Err.   [95% Conf. Interval]
-----+
numkids |  2.858553  .1157252   2.630147   3.086959
-----+
```

Note that the subpopulation number of observations is listed as 969. It's a good idea to check that number to make sure your subpop variable is working as expected.

It would be **incorrect** to use the **if** option to subset the data:

```
svy: mean numkids if zanzibar==1 // INCORRECT - DO NOT DO THIS

(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      2          Number of obs     =    969
Number of PSUs   =     30          Population size  =  1.0e+08
                                         Design df       =      28
                                         WRONG

-----+
|           Linearized
|   Mean   Std. Err.   [95% Conf. Interval]
-----+
numkids |  2.858553  .1037723   2.645985   3.071121
-----+
                                         WRONG      WRONG      WRONG
```

The number of PSUs is incorrect, so the standard error and the confidence interval are also incorrect. Note that the estimate of the mean is the same. This is just one example of how different the results can be when you subset the data. For some variables the difference might

be much smaller or much larger. It is best to always use the subpop option when analyzing a sub-population with the svy command.

Using the pweight and robust cluster() Options

The second way to analyze survey data is to use the estimation commands that allow the pweight and robust cluster options. The estimation commands when used with the pweight and robust cluster options handle the sampling weights and clustering properly. However, there is no option for specifying the stratification variable. As a result, the standard error may be larger than it would be using and svy command.

The following set of commands demonstrates the difference between logit (without stratum) and svylogit (with stratum):

```
clear
use "q:\utilities\statatut\svysamp.dta"
svyset earea [pweight=sampwt], strata(urbrur)
logit twokids age educat [pweight=sampwt], robust cluster(earea)
svy: logit twokids age educat
```

Choosing a method

The following table compares the two methods available for analyzing data from a sample survey:

Method	Strengths	Limitations
The survey commands	<p>test and lincom commands used after estimation adjust the test statistics correctly for the sample design.</p> <p>Can make finite population corrections for without-replacement samples.</p> <p>Option available on svyset command to specify the stratification variable.</p>	The analysis command you want may not support the svy prefix: see help svy_estimation for a current list of those commands
Commands that allow pweight and robust cluster() options.	<p>There may be an estimation command that supports cluster but does not support svy.</p>	<p>Should have at least 40 clusters available.</p> <p>Option for specifying a stratification variable is not available.</p>

It is best to use the survey commands to analyze survey data. These commands incorporate the effect of clustering and stratification as well as the effect of sampling weights when computing the variance, standard error, and confidence intervals, and they allow you to perform analyses on a subset of the data using the subpop option. If the analysis technique you need is not available with the survey commands, then using the estimation commands with pweights and robust cluster() options would be a good choice.

Review again?
Another topic?

Common Errors and How to Avoid Them

Here are some common errors that you can avoid.

Ignoring Clustering

Ignoring clustering and unequal probability of selection of participants in your analyses. This results in biased estimates and false-positive hypothesis test results. *Avoid this error by using the svy commands for your analysis. If your analysis technique is not available with the svy commands, then use a command that allows pweight with the robust cluster() option.*

Using the Wrong Weight Command

Using the wrong weight specification in Stata. *For data from a sample survey, you should use the pweight option to define the sampling weight.* Using any of the other weight options (aweight, fweight, or iweight) can result in incorrect variance, standard errors, confidence intervals, and p-values.

Subsetting the Sample

Subsetting the sample when using the svy commands in Stata. These commands use the Taylor Series approximation for the variance estimation and must be able to correctly count the number of primary sampling units (PSUs) that were originally sampled. Subsetting the data may cause an incorrect number of PSU's to be used in the variance computation formula. *Do not subset the data from a sample survey and always use the subpop option when using the svy commands to do sub-population analysis.*

Stratum with only one PSU detected

You may get an error message when you try to run an svy command: "stratum with only one PSU detected". This happens when observations have values missing for variables in your model, resulting in their being dropped. An entire PSU may disappear as a result of missing values. Use the **svydes** command to identify the problem strata. A common fix is to combine a small stratum with an adjoining stratum. See the manual entry on svydes, or in Stata type **findit svydes**, or see <http://www.stata.com/support/faqs/stat/stratum.html> for details.

What set of observations is Stata analyzing?"

Using a subpop variable does not do the same thing as an -if-. In fact that's why the subpop option was invented. The -svy- commands use the whole dataset to help determine the standard error even if you are only looking at a subset of it (with a subpop var). During the time Stata is analyzing your data, Stata subsets to only those observations where ALL the following variables are non-missing:

- strata (if using one)
- psu (if using one)
- sample weight (if using one)
- subpop (if using one)
- analysis variable(s)**

If any one of them is missing then Stata drops the obs where any of those variables are missing. ** svymean with more than one variable will not subset to obs where all analysis variables are non-missing unless the "complete" option is specified.

Review again?

Another topic?

Data Characteristics

A sample survey is conducted to obtain information about the characteristics of a population. To reduce the cost and time necessary to collect the data, this task is often handled by selecting a subset (a sample) from the set of all measurements (the target population) of interest to the researchers. The methods that are used to select the sample add certain characteristics to the data. These characteristics must be incorporated into your analysis to get estimates concerning the entire population. There should be variables for each observation in your data set to identify each of the characteristics that are described next.

Clustering

A truly random selection of households would involve listing all the households in the country and randomly selecting the desired number of households from that list. Using this approach, each household would have an independent and equal chance of being included in the survey. While this approach is statistically ideal, the cost of first enumerating all households and then visiting the selected households that would be scattered all over the country make this approach impractical.

A more practical approach is cluster sampling. For example, in the Demographic and Health Surveys, "enumeration areas" from the Census or similar national surveys are first selected randomly from a list of all such areas in the country (or within strata if stratification is being used). These areas are often referred to as "clusters" or as "primary sampling units" (PSU's). They may be towns or villages, or they may be census tracts in cities. Generally each cluster contains roughly the same number of households.

The next step is to enumerate (count and label) all the households in the cluster. Then a random-selection process is used to select households within each cluster. This is the sample of households that will be visited for the survey.

While cluster sampling is much more practical, it also means that the households are not statistically independent. Instead, the characteristics of a given household (and its household members) are more like those of other households in the same cluster, and are less like households in other clusters. This effect of a non-independent sampling process, called the "sample survey design effect", shows up in the standard error of estimation statistics (means, regression coefficients). Clustering tends to decrease the size of standard errors, leading to a greater likelihood of rejecting the null hypothesis. In other words, it's more conservative to correct statistically for the design effect.

Stratification

The population can be divided into sections (the strata) that are internally more homogeneous. This may be done in order to over-sample smaller groups in a target population. Examples of strata are region of country, urban/rural residence, or education level. A separate sample is selected from each stratum. Like clustering, the observations within strata are not statistically independent, and adjusting for stratification leads to more conservative inferences about statistical significance.

Sampling Weights

Each observation in the sample is chosen using a method of random selection. An important property of this method is that the probability of selection may not be equal for all members of

the population. The sampling weight for each observation is computed as the inverse of the selection probability. Additional adjustments (such as non-response) may be made to the sampling weights. An observation with a sampling weight of 1000 represents one thousand individuals from the target population while another observation with a sampling weight of 50 represents only fifty individuals. Your analysis technique will need to use the sampling weights to estimate the characteristics of the target population from the reports of the sample. Thus, the sampling weights are needed in computing both the population estimates (such as means and regression coefficients) and their standard errors.

Population Number of PSUs in Each Stratum

Sampling with replacement means that once a PSU was chosen, it remains eligible to be selected again. Without replacement means that once the unit is selected, it is no longer eligible for selection and a finite population correction will need to be made in your analysis. Data sets for without replacement samples will need to have a variable that tells how many PSUs per stratum are in the population. Stata will use the data set to count how many PSUs were selected per stratum and compute a sampling fraction to use in analysis.

If the proportion of PSUs selected from each stratum (the sampling fraction) is small, then your sample can be analyzed as if it was selected with replacement and you do not need this variable. This simplifies your analysis since you can ignore the finite population correction. According to Cochran, "In practice the fpc can be ignored whenever the sampling fraction does not exceed 5% and for many purposes even if it is as high as 10%. The effect of ignoring the correction is to overestimate the standard error of the estimate." (William G. Cochran, Sampling Techniques, 3rd Edition, 1977, John Wiley & Sons)

Questions

What characteristics of the sampling design affect estimates such as totals, means, proportions, and regression coefficients? **Answer:** Sampling weights.

What characteristics of the sampling design affect standard errors, p-values, and confidence intervals? **Answer:** Sampling weights, clustering, and stratification.

Review again?

Another topic?

Data cleaning

Working with do-files, documenting, outliers, duplicate ids.

Unless you're lucky enough to be working with perfectly clean data, you'll need to at least check for outliers and, the bane of all survey work, duplicate identifiers. In the previous section on changing data, the **edit** command was introduced. We don't recommend using it for data cleaning, because it does not keep a record of what you've done, and you can't easily repeat data cleaning steps you've taken.

Stata do-files, which are collections of commands that you write, are an easy way to clean and document your data. Copy and paste the following commands into a text editor (e.g., Wordpad, Notepad, or the Stata do-file editor), and save the file as a *text* (ascii) file named test.do. The suffix should be ".do" (not required, but strongly recommended). The file should be in the same directory that you're using for your Stata session. If you're using a different directory, remember where you saved it.

```
clear
use "q:\utilities\statatut\exampfac.dta"

/* Notice the blank line above this line. Blanks are fine in do-files. */
/* Comment Lines, Like these, can be anywhere in a do-file, including */
/* at the end of a command line. */

// You can also use double-slashes anywhere in a command line
// to write a comment. Everything from there to the end of the
// line will be treated as a comment.

gen factype2= factype /* comment at the end of a command */
gen factype3= factype // this works, too

* Which facilities have an odd value for age?

ta age
list facid facname factype authorit if age == 1998
replace age= 1 if age == 1998

* Look for duplicate values of facid using duplicates command.

duplicates list facid
list facid factype authorit if facid == 1001

* Look for duplicate values of facid using _n.

sort facid
list facid factype authorit if facid == facid[_n-1]
#delimit ;
list facid factype authorit if facid == facid[_n-1]
| facid == facid[_n+1];
#delimit cr

* Drop the duplicate observation where factype and authorit are missing.

drop if facid == 1001 & missing(factype)
```

If you're using the Stata do-file editor, click on the button "Execute (do)" (the icon looks like a page with writing on it and an arrow pointing to the right). If you used another editor, open Stata and type:

```
do test.do
```

or

```
do "driveletter/directory/etc/test.do"
```

depending on where you saved the file test.do. Press Enter and watch the results. When you see --more-- press the Space Bar.

Questions:

- Comments can begin with an asterisk (*) and end with a carriage return (Enter key), or they can begin with two slashes (//) and end with a carriage return, or they can be bracketed by /* */ and span as many lines as needed. All are shown in the example above. What happens if commands are enclosed in comments, like this:

```
/*
ta age
list facid facname factype authorit if age == 1998
replace age= 1 if age == 1998
*/
```

Answer.

- When **ta age** was executed, the screen stopped scrolling and --more-- appeared at the bottom. What does that mean? Answer.
 - What does the **duplicates** command do? Answer.
 - Why are you showing a second way to check for duplicates? Answer.
 - The character **_n** is the sequence number of the current observation in memory. What does **facid[_n-1]** mean? Answer.
 - Why do we need to sort the data before testing for duplicate values of facid? Answer.
 - Why list both observations **_n-1** and **_n+1**? Answer.
 - What if there is more than one nested id variable? Answer.
 - What do the commands **#delimit ;** and **#delimit cr** do? Answer.
-

Answers:

- Everything that's enclosed in comments is treated as a comment and is not executed. This is

a handy way to block out portions of a do-file that you don't want to run.

[Back to question](#)

2. When the results of a Stata command fill up more than one screen, Stata pauses to let you review what's currently on the screen. When you want to continue, press the Space Bar to see another page, or press the Enter key to see another line.

You have other options for how to handle this situation.

- put the command **set more off** in the beginning of your do-file
- click on the green "Clear --more-- Condition" button to see another page
- press Ctrl-Break to cancel the command
- press the "Q" letter key to cancel the command
- click on the "Break" button (red circle with white "X")

The **set more off** command allows the do-file to continue uninterrupted to the finish. This is a good option if you have turned on logging, so you'll have a record of what scrolled by. Use **set more on**(the default) when you need to stop and check each screen.

The other options are useful if you're debugging a do-file by executing one command at a time, or if you're working interactively (not using a do-file). If you Break or "Q" in the middle of a do-file, the do-file execution is canceled along with the command that produced the --more-- condition.

[Back to question](#)

3. The duplicates command counts, lists, tags, or drops duplicates. In this case we're just listing the duplicates so you can explore them further. See "One-to-one merging" for an example of dropping duplicates using the duplicates command. See **help duplicates** for details.

[Back to question](#)

4. Before the duplicates command, you needed to use `[_n]` to check for duplicates. Even though you no longer need it to check for duplicates, it's useful to know about `[_n]` for more advanced programming.

[Back to question](#)

5. You can think of each variable as a column in Excel, and `_n` is the row number. For example, `facid` is the first variable, which would be column A in Excel. To refer to the cell in the third row, you would type `A3` in Excel. Similarly, in Stata `facid[3]` is the value of `facid` for the third observation in memory.

Since Stata executes a command on every observation from top to bottom, `facid[_n]` is the current observation it's looking at, `facid[_n-1]` is the observation before the current observation, and `facid[_n+1]` is the next observation.

[Back to question](#)

6. If duplicate values of `facid` exist, sorting the data will place them next to each other in the data

in memory. So, the array element reference `[_n-1]` will refer to the element immediately prior to the current observation, and the duplicate value will match the current value.

[Back to question](#)

7. Actually, we've asked to list the current observation (`_n`). If the current observation has the same value of `facid` as either the previous or the next observation, it's a duplicate. This method catches 2 or more observations with the same value for any variable.

[Back to question](#)

8. If there is more than one nested id variable, you need to sort in the proper nest order and check all ids for duplicates using the logical operator `&` (ampersand) between them. Suppose there are 3 nested ids:

```
sort id1 id2 id3
list id1 id2 id3  if id1 == id1[_n-1] & id2 == id2[_n-1] & id3 == id3[_n-1]
```

Again, we'd like to point out that the **duplicates** command is much easier to use than this method:

```
duplicates list id1 id2 id3
```

We've shown you the method using `_n` because there are many instances in programming where you'll want to refer to observations before or after a particular case. This implicit array is powerful and well worth understanding and remembering for other applications.

[Back to question.](#)

9. In a do-file, Stata assumes that each command is no more than 1 line long, and that each line ends in a carriage return (when you press the Enter key, a text editor inserts a carriage return symbol). If you want to type a command that is more than one line long, you can use `#delimit ;` to tell Stata to look for a semi-colon instead of a carriage return. From that point on, you must end each command, whether one or more lines long, with a semi-colon. To switch back to carriage return, use `#delimit cr`.

There are other ways to continue a single command across more than one line. One way is to comment out the carriage return - type `/*` at the end of one line, and `*/` at the beginning of the next line (to end the comment):

```
list facid factype authorit if facid == facid[_n-1] /*
 */ | facid == facid[_n+1]
```

Another way is to end a line with `///`, which tells Stata to continue reading the next line as a continuation of this line:

```
list facid factype authorit if facid == facid[_n-1] ///
 | facid == facid[_n+1]
```

There is no one "correct" method, so use the one you prefer.

[Back to question](#)

Review again?

Another topic?

Describing the data

Describing the variables: means, univariates, frequencies, and data types.

Now we'll use some real data. The data are from a health facility survey conducted in Tanzania in 1999. Copy each of these commands into the Command window, press Enter, and observe the results. Note that the letters in bold on each command are acceptable abbreviations. If you see "-more-" at the bottom of your Results screen, press the Space Bar to see a new page of data, or press "q" to quit the command.

```
clear
use "q:\utilities\statatut\exampfac.dta"
describe
summarize
su pill-natural
su fphour*
codebook urbrur facname
tabulate urbrur
tab urbrur, nolabel missing plot
tab factype urbrur
tab1 factype urbrur
tab2 factype urbrur
tab2 factype urbrur, row col cell
```

Questions:

1. The **describe** command lists each variable in Stata's memory. What do the terms "double," "str42," "byte," etc. in the second column refer to? Answer.
2. How do I specify which data type I want to use? Answer.
3. The **summarize** command lists the number of observations, mean, standard deviation, min, and max for a variable. Why is the number of observations different for some variables, and is even 0 for facname? Answer.
4. How can I summarize a specific set of variables? Answer.
5. When is the **codebook** command useful? Answer.
6. The **tabulate** command gives frequencies (counts), and is most useful with categorical variables. What are the two ways to specify a one-way frequency? Answer.

7. What do the **nolabel missing plot** options do on the **tabulate** command? Answer.

8. How can I get two-way frequencies (cross-tabulations)? Answer.

Answers:

1. That is the data type for each variable. Each data type handles a different kind of data. The following table describes the data types used by Stata:

Type	Min	Max	Precision	Bytes	Type
byte	-2 digits	2 digits	2 digits	1	integer
int	-4 digits	4 digits	4 digits	2	integer
long	-9 digits	9 digits	9 digits	4	integer
float	-10^{**38}	10^{**36}	10^{**-8}	4	real
double	-10^{**307}	10^{**307}	10^{**-16}	8	real
str1	1	1		1	string
str2	2	2		2	string
...
str2045	1	2045		2045	string
strL	2000000000	2000000000		2000000000	long string

Prior to Stata version 13, strings were limited to 2045 characters. Starting with Stata 13 a new data type **strL** can hold strings up to 2 billion characters. You can see this and the limits on just about everything else in Stata by typing the command **help limits**, and there's a detailed explanation of data types in the PDF documentation and under **help data types**.

[Back to question.](#)

2. You can specify a data type on the **generate** command:

```
gen byte a=0
```

If you don't specify a data type, by default Stata uses type float (4 bytes). Using an efficient data type reduces the file size. This is important for very large files or for computers with little memory (RAM). The **compress** command selects the most efficient data type after variables have been generated. See **compress** in Miscellaneous Tips and Tricks for details on compress.

[Back to question.](#)

3. The "Obs" column displays the number of non-missing observations for numeric variables.

For string variables, like facname, it is always 0.

Back to question.

4. There are two ways to specify a variable list, both shown in the example:

- **pill-natural** (first variable - last variable)
- **fphour*** (root variable name plus *)

These two methods work with all Stata commands. To use the first method, you need to know the position of each variable in the Stata data file. Use the **describe** command to see those positions, or look for them in the Variables window.

Back to question.

5. The **codebook** command gives univariate statistics about numeric variables, and it is a handy way to get information about string variables.

Back to question.

6. The two ways to get one-way frequencies are:

- **tab factype** (for a single variable)
- **tab1 pill-natural** (necessary for lists of variables)

Another handy command is **fre** written by Ben Jan at the University of Bern. It is not built into Stata, but we have installed it on all terminal servers at CPC. Like all user-contributed Stata commands, it is available for free from the SSC archives at Boston College. You can install it on your desktop or laptop by typing:

ssc install fre

Back to question.

7. These three options give extra information about the variable urbrur:

- **nolabel** displays the numeric values instead of the value labels
- **missing** shows how many observations have missing values
- **plot** gives a graphical comparison of the frequencies

For more information on how Stata handles missing values, see missing values in the

Miscellaneous Tips and Tricks section of this tutorial.

[Back to question.](#)

8. The two ways to get two-way frequencies are:

- tab factype urbrur
- tab2 factype urbrur

These two commands are equivalent.

[Back to question.](#)

Review again?

[Another topic?](#)

Documenting your work

Stata provides several opportunities to document your work. You can describe your variables in more detail than is possible with the variable name using the **label variable** command. You can attach one or more words to each value of a variable using the **label values** command. These variable and value labels will be displayed in the results of Stata commands that support them.

The do-file is the best place to document how you created your variables or cleaned your data. Not only is the code there, allowing you or others to reproduce your work, but you can also add explanations and keep a record of decisions about how decisions were made. You can start with this **do-file template** to create a format that suits your needs.

After creating your analysis file, or if you have a data file for public dissemination, you might want to create a codebook using the **cb2html** command. This command allows you to add much more information to your codebook about each variable, the data set, and even the questionnaire, using the Stata **note** command. Such information might include the algorithm you chose and the decisions you made in creating a new variable. For survey data, they might include the skip instructions to the interviewer to help the data user understand the pattern of missing values. If you are not using CPC's network, you'll need to download this command from the SSC archive:

```
ssc install cb2html
```

Another topic?

Dummy variables

Creating Indicator (Dummy) Variables

Shortcuts to save you time.

Sometimes we need to create a number of indicator, or dummy, variables from a single categorical variable. These indicators usually take the value of 1 if the observation has the attribute and 0 if the observation does not. Most Stata commands support a syntax that creates indicator variables for you automatically. They call that syntax **factor variables**. There's a clear and detailed discussion in Chapter 25 of the Users Guide.

Here's a simple example to give you the flavor of factor variables syntax. Suppose each respondent has a value for their age group in a variable called agegroup that has 5 values. You want to create 5 indicator variables with values 0/1 that describe whether the respondent is in age group 1 or not, age group 2 or not, etc., and use them in a regression on the correlates of body mass index (bmi). The factor variable syntax is:

```
regress bmi i.agegroup
```

It's that simple! Age group 1 is automatically treated as the base level and omitted from the equation.

Factor variables syntax is much more powerful than this simple example illustrates. For example, it will create interactions for you with continuous as well as categorical variables. See the Users Guide for a complete discussion.

Unfortunately, not all commands support the factor variable syntax. Below are some alternatives in case you need to use a command that doesn't support factor variables.

The most obvious way to do this is the **generate** command. Suppose we want to create 5 indicator variables from agegroup, a variable with 5 values:

```
gen age1=0
replace age1= 1 if agegroup == 1
gen age2=0
replace age2= 1 if agegroup == 2
etc.
```

This can be tedious if you're creating a lot of indicator variables. A few shortcuts are available, including **recode**, **autocode**, and **egen**, all of which are discussed in the Users Guide referred to above. Here are a few more alternatives.

The first shortcut is the **forvalues** command. See Looping over variables and values in this tutorial to learn the basics of this command.

```
forvalues n=1/5 {
    gen byte age`n' = 0
    replace age`n' = 1 if agegroup==`n'
```

{}

This use of **forvalues** command simply generates the two commands in the first example 5 times for us, eliminating all that typing and opportunity for error. Note that we've added the data storage type **byte** to the generate command. Since the indicators only contain the values 0 and 1, they easily fit in a single byte of storage, so this option saves megabytes of storage. See Describing the data in this tutorial for an explanation of Stata's storage types.

The second shortcut is the **tabulate** command, which is the easiest to use.

```
tab agegroup, gen(age)
```

The gen option on tabulate creates a new dummy variable for each value of agegroup. It names each dummy using the prefix you assign in parentheses, in this case "age". Note that the dummies are named age1 through age5, which may or may not correspond to their value. However, the values are recorded in the variable labels.

The third shortcut is the **xi** command. This command is really intended to feed indicator variables into another Stata command, such as a regression. It has largely been replaced by factor variables, but it will create dummy variables.

```
rename agegroup age
xi, prefix(i) noomit i.age
```

First we rename agegroup to age so that the indicator variables have a shorter name. In the xi command, the **prefix(i)**option gets rid of Stata's default prefix, "_I", which it adds to each dummy variable name. We use the **noomit** option because xi does not create a dummy for the lowest value (remember, it's designed to feed these dummies into a multivariate procedure, so one category must be dropped). The result is 5 indicator variables named iage_1, iage_2, ..., iage_5.

The fourth shortcut is the **margins** command. This command is used in postestimation after a previously fitted model at fixed values of some covariates. The command is powerful and full of options, and it is much more than a "shortcut". See the full help in the PDF Base Reference.

Review again?

Another topic?

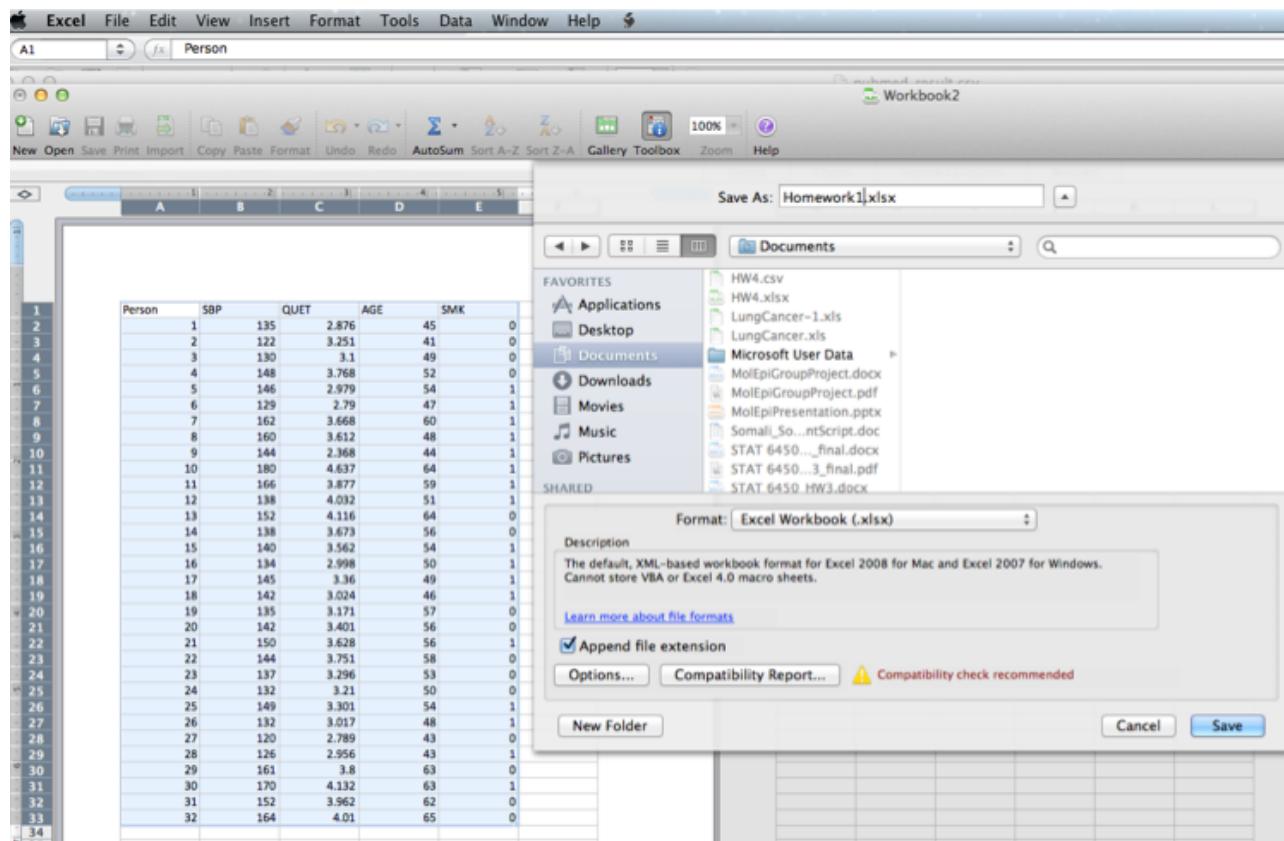
Entering Data into STATA

[Help Center](#)

Here are some steps that you could follow to enter datasets into STATA

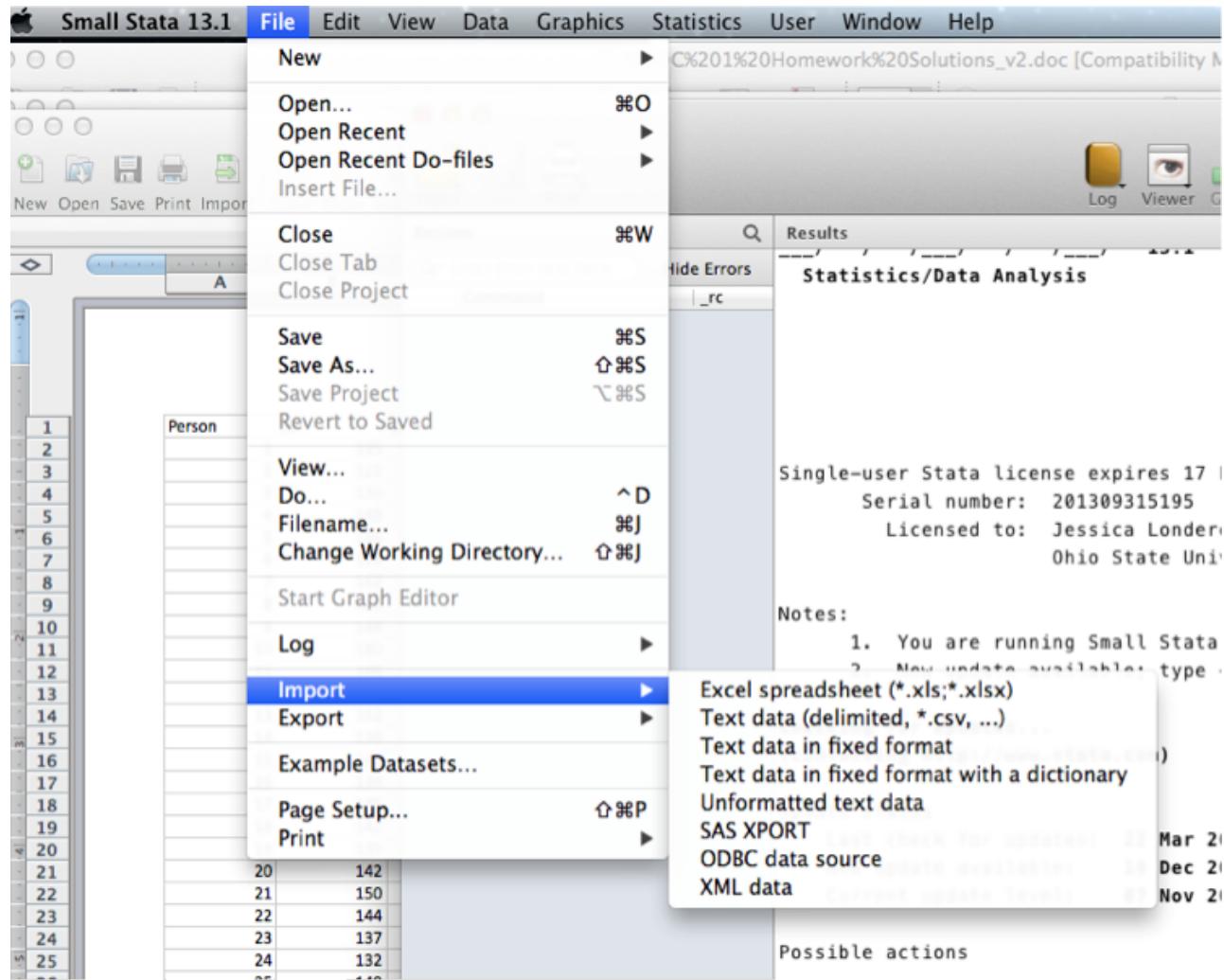
Step One:

Copy and paste the table into a Microsoft Excel spreadsheet. Save the spreadsheet.



Step Two:

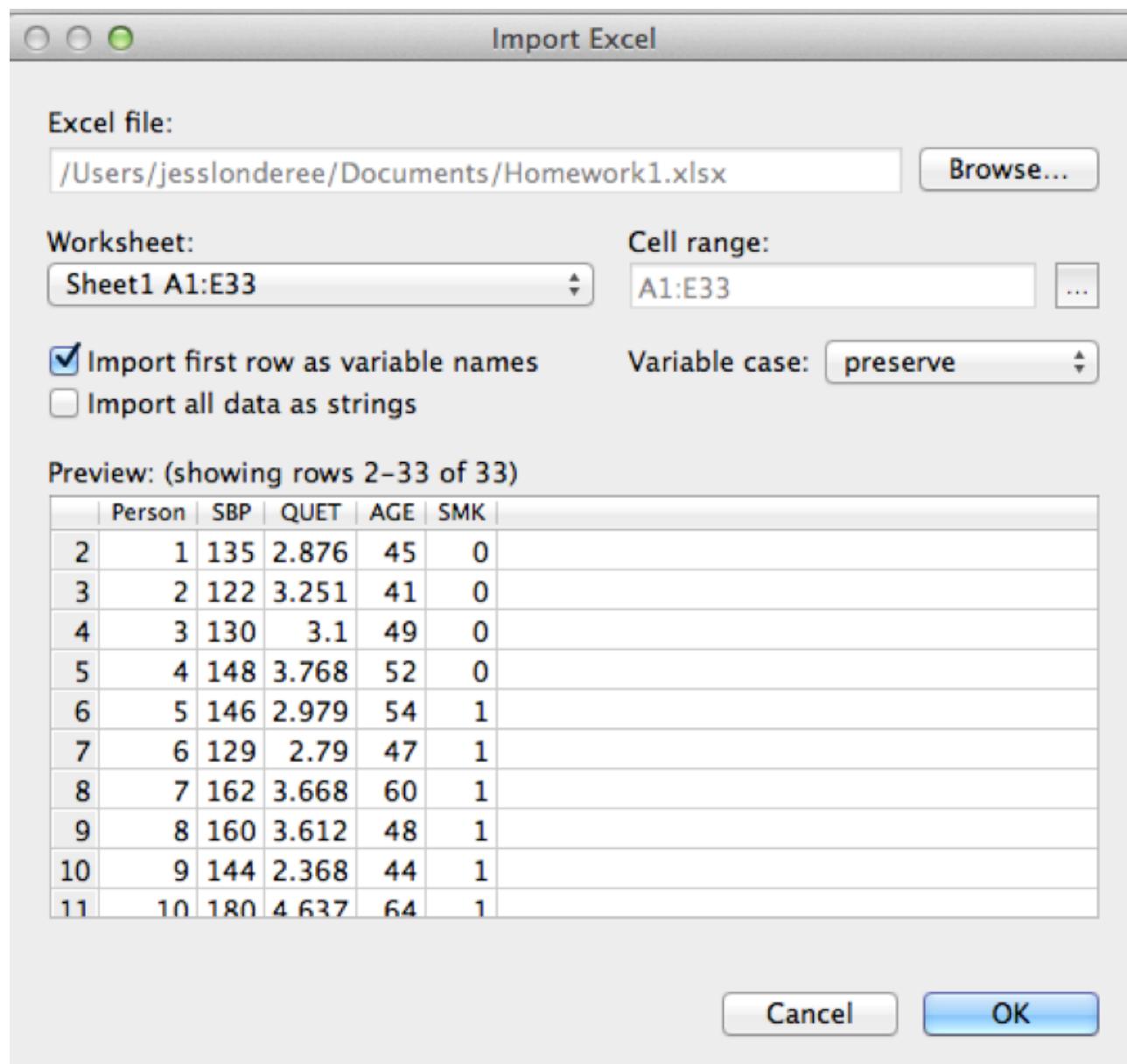
Open STATA, Click on File>Import>Excel Spreadsheet. This will bring you to the "Import Excel" module



Step Three:

To the right of where it says "Excel File", click "Browse" then navigate to find your spreadsheet.

Make sure the box next to "Import first row as variable names" is checkmarked, then click "Ok".



You should be all set to get started!

You can also [download this guide](#).

Created Tue 24 Mar 2015 12:34 PM PDT

Last Modified Tue 24 Mar 2015 12:46 PM PDT

Error messages

Error messages and what they mean

Stata's error messages are usually clear, but it's not always obvious what you can do about them. Here are a few that could use a little explanation.

1. too many values

This error message can result from a 2-way tabulation of variables that each have a lot of values. If you really want to cross-tabulate these two variables, you can do it in pieces, limiting the range of one or the other variable in each tabulation:

```
tab2 var1 var2 if var1 <= 10  
tab2 var1 var2 if var1 > 10 & var1 <= 20  
etc.
```

2. no: data in memory would be lost

If you make any changes in the data currently in memory, then try to exit Stata, you'll get this message. Stata is reminding you that you have not saved the data in memory since making the changes. You can either **save** the data, if you want to keep the changes, or **clear** the data from memory, if you don't want to keep the changes. Then exit.

Review again?

Another topic?

Exporting Stata Results to MS Office

Stata results can be exported in a wide variety of ways. We've divided the approaches into two groups: working with the Results Window and working with estimated parameters directly.

From the Results Window

Copy Table into Excel, Word, or PowerPoint

Suppose you've run `svy:mean` on 3 variables, and Stata has produced a nice table in the Results Window that you want to capture.

. svy, subpop(subpop): mean var1 var2 var3 (running mean on estimation sample)
Survey: Mean estimation
Number of strata = 43 Number of obs = 20607
Number of PSUs = 86 Population size = 261897236
Subpop. no. obs = 18081 Subpop. size = 253890607
Design df = 43

Linearized
Mean Std. Err. [95% Conf. Interval]

var1 289.6824 4.175324 281.2621 298.1028
var2 106.6606 1.311485 104.0157 109.3055
var3 396.343 4.804086 386.6547 406.0314

Copy Table

You can copy the table into Excel or into a Word table and retain much of its formatting:

- Highlight the table in the Results Window.
- Right click on the table and select Copy Table.
- In **Excel**, click on the cell where you want the upper-left corner of the table.
- Right click and select Paste.

You'll need to adjust column widths to see all the text. We've found that copying the upper half of the table separately from the lower half helps Excel select the right number of columns to use.

To copy into **Word**, count the number of rows and columns you'll need, create a table of that size, highlight all the cells in the table, and then paste. In the above example, the table of parameter estimates requires 5 columns and 6 rows if you do not include the top and bottom horizontal lines in your highlighting.

Copy Picture

You can copy the highlighted table as a picture directly into a Word document. It's a perfect snapshot of the table in the Results Window.

- Highlight the table in the Results Window.
- Drag the left border of the Results Window to the right until the highlighted table just fills the width of the Window.
- Right click on the table and select Copy as Picture.
- In **Word** or **PowerPoint**, right click and paste where you want the picture.

You can resize the picture, and you can right click to edit it.

Using the Parameter Estimates

Suppose that we want to make a table of just the means in the above example. We can do this by taking advantage of a very nice feature of Stata. Most Stata commands produce temporary variables containing the key results. These temporary variables will continue to store these values until you use another Stata command that replaces those results, or until you end your Stata session.

You can look at what results have been stored using either **return list** or for estimation commands **ereturn list**. Here's a subset of the estimate parameters stored temporarily after the svy:mean command that produced the table above:

```
ereturn list

scalars:
    e(df_r) = 43
    e(N_strata_omit) = 0
    e(singleton) = 0
    e(census) = 0
    e(N_subpop) = 253890607
    e(N_sub) = 18081
    e(N_pop) = 261897236
    e(N_psu) = 86
    e(N_strata) = 43
    e(N_over) = 1
    e(N) = 20607

(lots of results omitted here)

matrices:
    e(b) : 1 x 3
    e(V) : 3 x 3
    e(_N_subp) : 1 x 3
    e(V_srssub) : 3 x 3
    e(V_srs) : 3 x 3
    e(_N) : 1 x 3

(mores results omitted here)
```

For purposes of this example, we're only interested in the means. The estimates of the means are contained in a matrix called **e(b)**, which has dimensions 1 x 3.

The matrix **e(b)** is temporary. We want to create variables that we can export to Excel, so we first need to put the means into variables that won't go away when we run another estimation command. There are lots of ways to do this. We'll demonstrate one way that uses official Stata commands and is flexible, but it requires you to manage the data a bit more than you might care to do. After that we'll talk about a number of user-contributed commands that work with results from estimation commands.

First, let's verify that we have in fact selected the right matrix of estimates:

```
matrix list e(b)
```

```
e(b)[1,3]
      var1      var2      var3
y1  289.68245  106.6606  396.34305
```

We can compare these values with the means printed in the table above and see that we have the estimates of the means. Next put them into a matrix, and then create a variable for each cell in the matrix:

```
matrix means = e(b)
gen meanvar1= means[1,1]
gen meanvar2= means[1,2]
gen meanvar3= means[1,3]
list meanvar* in 1

+-----+
| meanvar1    meanvar2    meanvar3 |
|-----|
1. | 289.6824   106.6606   396.343 |
+-----+

keep in 1
(21661 observations deleted)
```

Since every observation has the same values for the three means, we can keep just one observation.

At this point we can **browse** the data, copy, and paste it to an empty Excel file template we've created with column and row labels, colors, and all the other formatting we want. We find this to be the easiest approach.

Another approach is to **export** the results to Excel.

```
export excel "c:\tables\means.xlsx", firstrow(variables)
```

Here is an example of how to write a program to build a more complex table with many rows.

putexcel

New in Stata 13 is the command **putexcel**. It reads the parameter estimates described above and writes them directly to an Excel spreadsheet. You can specify the starting cell (upper left) and even write column headers. Here's an example that accomplishes the task described above.

```
svy, subpop(subpop): mean var1 var2 var3
putexcel C4=("Means") B5=matrix(e(b)) using "c:\tables\means.xlsx"
```

As we saw above, the **svy: mean** command puts the means into the matrix **e(b)**. We then instructed **putexcel** to write those 3 means into a row in a spreadsheet, starting with cell B5. We also wrote the label "Means" centered above the 3 means in cell C4. The **putexcel** command is very powerful and flexible. See the PDF help for full details.

User-Contributed Commands

It is a testament to both the cleverness and generosity of the Stata user community that so

many powerful open-source commands exist outside the set shipped by Stata Corporation, and further that the authors maintain these commands and continue to offer improved versions. Most of these commands are archived at Boston College in the SSC (Statistical Software Components) website. To read about and install these commands, you can search the SSC website, or you can use the **ssc** command in Stata.

```
ssc describe parmest
ssc install parmest
```

CPC maintains the most current version of each of the commands listed below. You will need to install them locally if you run Stata on a standalone computer.

The discussion below is brief, intended only to point you to the command you might need. See the help for each command for details.

tabout

The **tabout** command produces publication-quality cross tabulations. Lots of features are available to customize the table. The output file may be in tab-delimited or html format. Tab-delimited format files may be copied into Word and converted into a table using the Table menu (Table, Convert, Text to table). Html format files can be opened in your browser and copied. Inside Word, use Paste Special and paste the table as Formatted Text (rtf). There's a tutorial on the command as well.

outreg

The **outreg** command uses the saved results after an estimation command to create a text file of the results you select. You can then turn this text file into a Word table. This command has lots of formatting features, and you can combine the results from several regressions into a single table. Here's a simple example to get you started.

```
cd "c:\tables\"  
sysuse auto, clear  
regress mpg foreign weight headroom trunk length turn displacement  
outreg using outtab1, replace
```

The output file is tab-delimited and has the extension .out. Open the file in Word, highlight the rows of results, click on Insert, Table, Convert text to table.

outreg2

This is an extension of outreg to enhance its capabilities with multiple models and to provide more format options. Here's an example from the help. Follow the instructions above to convert the results to a Word table.

```
cd "c:\tables\"  
sysuse auto, clear  
regress mpg foreign weight headroom trunk length turn displacement  
outreg2 using outtab2, replace cttop(full)  
regress mpg foreign weight headroom  
outreg2 using outtab2, see
```

esttab

This command is part of the **estout** package. The **esttab** command produces a table of regression results that have been stored by the **eststo** command, or the current results if nothing has been stored. The output table may be tab, csv, rtf, html, or other formats. Tab-delimited format files may be imported into Excel using the Import Wizard. Rtf files go directly into Word. There's a steep learning curve to this powerful command, but if you produce publication-quality tables often that have to look "just so", it's worth the time to study this command. Here's a simple example that outputs to .rtf.

```
eststo: svy, subpop(subpop if childage == 1): mean var1 var2 var3
esttab using "c:\tables\esttab_means.rtf", replace
```

The command gives you a clickable link to esttab_means.rtf in the Results Window. There's a tutorial available for esttab here.

xml_tab

Like esttab, the **xml_tab** command works from stored estimates. The output is in Excel's xml format. This format allows for a very feature-rich table that takes advantage of many of Excel's capabilities. Here's a simple example.

```
svy, subpop(subpop): mean enerbev1 enerfood1 enerpot1
xml_tab e(b), save("d:\statatemp\xml_tab.xml") replace
```

The command gives you a clickable link to xml_tab.xml. You can store multiple sets of estimates using **estimates store** and use **xml_tab** to combine them in a single table.

parmest

The parmest command comes in a package of four modules: parmest, parmby, parmcip, and metaparm. The **parmest** and **parmby** commands put each estimation result into a separate observation in an output dataset. **parmcip** inputs these variables and adds new variables containing confidence intervals and p-values. **metaparm** does a metanalysis on sets of estimation results. You can see an example here.

logout

The **logout** command captures the results that are printed to the Results Window (log) and writes them to Excel, Word, or other formats. The success of the formatting in the output file depends on the complexity of the results table you are exporting. Compared with copying from the log by hand and pasting into a spreadsheet, this approach may not produce as well formatted tables, but it can be automated. Here's an example:

```
logout, save(c:\tables\logout_means) excel replace: ///
svy, subpop(subpop): mean var1 var2 var3
```

This command produces a file call logout_means.xml, and it gives you a clickable link to this file in the Results Window.

Review Again?

Another topic?

Getting help for Stata

There are several ways to get help for Stata commands within Stata:

- help
- menus
- search

The **Help** menu links you to the PDF Documentation, which is the full set of official manuals for Stata. The **help** command gives most or all of the same information, but you must know the command.

The Data, Graphics, and Statistics **menus** build Stata commands for you. They are function-oriented, so you do not need to know the command or syntax.

The **search** command and link in the Help menu allows a keyword search for related words that are not themselves Stata commands. They will point you to help for the related commands. It searches not only the help files that come with Stata but also a wide range of web-based resources at Stata Corp and elsewhere. Type **help search** for details.

Web-based Help

The search command searches most web-based resources for specific commands and keywords. However, if you want more general help, such as a tutorial, a good resource is the UCLA's IDRE Stat website. It has links to resources for many statistical computing programs, including Stata, and has a search feature. You may also want to explore Stata's Resources for learning Stata.

Manuals

CPC no longer maintains current printed copies of the Stata manual set, since the full set is available in PDF in the Help menu. However, we do have other Stata and third-party books that focus on specific topics in Stata programming. These books are shelved in the CPC library and can be checked out.

User-Supplied Stata Commands

Many more commands are available for Stata than are shipped in the official version. Stata Corporation encourages users to develop new commands, and at times adopts these commands in some form into their official version. These include statistical, data management, and graphics commands. If you don't see a command in the official Stata manuals or help, the search command will search the user-supplied command archives. You can also search the main repository at the Boston College Department of Economics:
<http://ideas.repec.org/s/boc/bocode.html>

Review again?

Another topic?

Graphics

Creating simple graphs to visualize your data

Stata produces publication-quality graphs that can be modified to fit publishers' formats, and *schemes* (templates) are available for many journals to automate this process.

This page gives a few examples of graphs that you might want to create, but it by no means does justice to the broad capability of Stata graphics. In addition to the Stata *Graphics* PDF manual and the online help, other sources of information include *A Visual Guide to Stata Graphics* by Michael Mitchell (available from Stata Press), the drop-down Graphics menu in the Windows version of Stata, and a couple of online learning aids:

- www.stata.com/help.cgi?graph
- www.ats.ucla.edu/stat/stata/topics/graphics.htm

Each of the commands below creates an example of a commonly used graph. Use the Tanzania 1999 facility data file, then run the following commands to see the examples.

```
clear
use "q:\utilities\statatut\exampfac.dta"

/* Create Government vs Non-govt authority */

gen govt= .
replace govt= 1 if authorit == 1
replace govt= 0 if authorit > 1 & !missing(authorit)
label define g 0 "Non-govt" 1 "Govt"
label value govt g

/* Count FP staff-hours/week at each facility */

gen fphrs= 0
foreach v of varlist fphour* {
    replace fphrs= fphrs + `v'  if !missing(`v')
}
label value fphrs FP staff-hours/week

/* Count FP methods available at each facility */

gen methods= 0
foreach v of varlist pill-natural {
    replace methods= methods + 1  if `v' == 1
}
label value methods FP methods

/* Histogram of number of methods with normal curve superimposed */

histogram methods, normal

/* Box plot of methods by whether govt or non-govt facility */

graph box methods, by(govt)

/* Scatter plot of FP hours by number of methods */

scatter fphrs methods

/* Add a linear regression line to the scatter plot */

twoway (scatter fphrs methods) (lfit fphrs methods)
```

Another topic?

Groups and subsets of data

Groups and subsets of data, comparison, missing values.

We'll continue using the 1999 Tanzania Facility Survey data. Copy and paste the following commands into the Command window, press Enter, and see what happens. No need to copy the comments (surrounded by /* */). Note the double equal signs (==) in the **tab** and **su** commands.

```
clear
use "q:\utilities\statatut\exampfac.dta"

/* types of government facilities */

tab factype if authorit==1

/* mean age of hospitals */

su age if factype<=4

/* availability of condoms at religious facilities */

tab malecond if authorit==4 | (authorit>=7 & authorit<=9), missing

/* mean age of facilities by urban-rural location */

sort urbrur
by urbrur: su age

/* mean age of facilities that offer family planning by urban-rural location */

by urbrur: su age if pill<.

/* display the first 10 observations in memory */

list factype authorit urbrur in 1/10
```

For details on subgroup processing, see **by** command in the Miscellaneous Tips and Tricks section of this tutorial.

Questions:

1. The **if** option on the **tab** command restricts the command to observations that meet the following qualifications. In this case, the qualification is that authority is 1 (government). What is the difference between

```
authorit=1
```

and

authorit==1

in the **tab** command? Answer.

2. The == sign is called a relational operator. What relational, logical, and arithmetic operators are available in Stata? Answer.
3. The **sort** command puts the observations in memory in ascending order of the values of one or more variables. In this case, the variable is authority. What does the command **sort factype urbrur** do? Answer.
4. The **by** option executes the following command once for each value of the by variable. When would you use **by** instead of **tab2** to see data separately by groups. Answer.
5. What does the phrase **pill<.** mean? Answer.
6. How are missing values stored in Stata data? Answer.
7. The **in** option allows you to specify which specific observations you want. What do the numbers "1" and "10" refer to in this example? Answer.
8. What is another way, using **_n**, to write "in 1/10"? Answer.

Answers:

1. A single equal sign means give the value on the right to the variable on the left, in other words it means "assignment." A double equal sign means check whether the variable on the left has the value on the right, in other words "comparison."

[Back to question.](#)

2. Here is the full list of relational, logical, and arithmetic operators:

==	equal to
>	greater than
>=	greater than or equal to
<	less than
<=	less than or equal to
~	not

```
!    not
&    and
|    or
+    addition
-    subtraction
*    multiplication
/    division
^    power
```

[Back to question.](#)

3. It puts the data in ascending order of factype, and within each value of factype it orders the data in ascending order of urbrur. The **gsort** command allows you to sort in descending as well as ascending order.

[Back to question.](#)

4. The tab2 command works best with categorical data, while the by option works best with continuous variables. For example, if a and b are categorical variables:

```
by a: tab b
```

is the same as:

```
tab2 a b
```

The tab2 command gives more concise output and offers chi-square and other statistics.

For n-way analyses of continuous data, consider the **tabsum**, **tabstat**, or **table** commands instead, such as:

```
tabulate urbrur, summarize(age)
tabstat age, by(urbrur) stats(mean n)
table urbrur, contents(mean age n age)
```

The table command is particularly powerful and handles multiple levels of conditioning variables. See help for details.

[Back to question.](#)

5. It means "null is less than missing."

[Back to question.](#)

6. Missing values are stored as a number larger than the largest allowable value for the data type. So, you need to be careful when using the "greater than" operator: This command includes missing values of age:

```
tab factype if age>=25
```

If you don't want missing, you need to specifically exclude it:

```
tab factype if age>=25 & age<
```

You can specify up to 27 different types of missing values. They are: ".", ".a", ".b", ... ,".z". (During data entry, you can use these to differentiate among Refused, Not Applicable, Don't Know, and other possible reasons for missing values.) These are the largest values allowed by the data type, so you can use "<." to exclude all 27 missing values for a variable. Back to question.

7. The numbers refer to the temporary variable "_n" that Stata creates for each observation in memory. This number is not saved if you save a permanent data file. Furthermore, this number changes if you change the sort order of the data.

[Back to question.](#)

8. You can write:

```
list factype authorit urbrur if _n <= 10
```

[Back to question.](#)

Review again?

Another topic?

Importing and exporting data files

We often find data on the Web in formats other than Stata, or we need to share data with our collaborators who (unfortunately) are not Stata users.

There are several ways to get data into and out of Stata. While this list is probably not exhaustive, it covers most of the formats that we encounter at CPC. Some methods are available from within Stata, while others require shareware or commercial software.

Stata commands

import and export

These are native Stata commands to read and write several data file formats. The commands **import excel** and **export excel** read and write Excel spreadsheets (.xls and .xlsx). The command **export excel** is quite flexible, allowing you to write to a specific cell in a specific sheet in an Excel workbook if you need to. The commands **import sasxport** and **export sasxport** read and write SAS xport (transport) format files and will even read and write value labels from/to a SAS formats.xpf file. See **help import** and **help export** for details. See Exporting Stata Results to MS Office for an example of **export excel**. Here's a simple example of importing an Excel spreadsheet:

```
import excel "C:\tables\workbook.xlsx", firstrow
```

usespss

This command, written by Sergiy Radyakin, is available from the SSC archive. It handles Windows SPSS data files with the .sav suffix (called "system" files in SPSS). It does not handle SPSS portable files with the .por suffix. See **usesas** and **Stat/Transfer** below for software that can deal with SPSS portable files.

outsheet and insheet

The **outsheet** command is a way to write the variable names and values (but no labels) into tab-delimited text files, comma-separated values (.csv) files, or files with other delimiters. Most software will read a text file with one of these delimiters between the values. Stata also reads these text files with the **insheet** command. So, you can export data from something like MySQL to a .csv file and read it into Stata with **insheet**.

xmluse and xmlsave

These commands transfer data between Stata and MS Excel's xml format. (For most purposes you'll probably find the **import excel** and **export excel** commands more useful.) The xml format is a portable text version of Excel's .xls or .xlsx format files. **xmlsave** produces a file that Excel reads directly. In order to use an Excel file in Stata, though, you need to open it in Excel and save it as type xml. There are two xml "Save As" formats in Excel. The "XML spreadsheet 2003 (*.xml)" option seems to work better.

Other software

usesas and savasas

While these are implemented as Stata commands, they are listed here under "Other software" because they both require you to have SAS installed locally as well as Stata. At CPC we have both of these commands installed and they work wonderfully.

If you have a standalone computer with both SAS and Stata installed, you may want to download these commands from the SSC archive. You may need to edit the code to tell Stata where to find the SAS command, but that is all explained in the help and in comments inside the code.

Stat/Transfer

This is not intended as an advertisement for a commercial product, but it really is useful. It converts data files between statistical software, spreadsheets, databases, etc. with ease. If you find yourself needing to import or export data across software formats often, it is worth the investment. It is distributed by Stata Corp as well as by the developer Circle Systems. At CPC, send an email to CPC Help for access to Stat/Transfer.

use13, saveold

At CPC we keep Stata up to date with the most recent version. Currently, that is version 13. You may be working with colleagues who have an earlier version that cannot read version 13 files. When they try to use one of your files, they will get the message that the file is not in Stata format.

One option is the command **use13**, written by Sergiy Radyakin at the World Bank and available from the SSC archives. You can ask your colleagues to install **use13** and run it in place of the **use** command to read your files.

Another option is Stata's **saveold** command. You can use it to save your file in version 12 format, which version 11 can also read, and send this older version of the file to your colleagues.

Review Again?

Another topic?

Introduction to Stata

This tutorial is function-oriented, focusing on the data-management tasks most needed by data analysts working with sample survey data. It works up from basic tasks, such as how to drop variables, to the tasks needed for complex file organization, such as how to reshape and merge data files.

There is also a section on Analyzing Data from Sample Surveys. It explains which sampling weight command to use and whether to use svy or robust cluster to adjust for survey design effects.

These web pages assume that you are using Stata Version 13 for Windows.

If you would like to run the example commands, you need to copy the example Stata data files to your local PC. Click download sample data for instructions. If you're using a computer at the Carolina Population Center, the data are available to you on Q:\temp\statatut\.

See Stata Windows environment below for an orientation to the Windows interface. It also gives you sources of help beyond this tutorial.

Other resources available to help you learn Stata include the UCLA's IDRE Stat website, several introductory guides in the CPC library and others available from Stata Press, and Stata Corporation's Resources for learning Stata.

SAS Users: the SAS User's Guide to Stata may help you make the transition from SAS to Stata.

A simple example

- **input:** putting data into Stata
- **generate:** creating a new variable
- **list (or browse):** viewing the contents of memory
- **save:** saving memory in a permanent Stata-format file
- **log:** capturing the results of Stata commands for printing
- Stata's default actions
- how data are stored in RAM

Using permanent Stata data files

- **clear:** clearing Stata's memory
- **set memory:** allowing enough space for the data
- **use:** copying the file into memory
- **save,replace:** saving changes

Describing the data

- **describe:** names of variables
- **summarize:** the mean, min, and max of variables
- **codebook:** more univariate statistics
- **tabulate:** frequencies and cross-tabulations

- **data types and data storage**

Groups and subsets of data

- **if:** do command for a subset of observations
- **sort:** order observations by the values of a variable
- **by:** do command for groups of observations (requires sort)
- **in:** do command for a range of observations
- relational, logical, and arithmetic operators
- missing values

Changing the data

- **replace:** change the values of a variable
- **recode:** change the values of a variable
- **rename:** change a variable name
- **label:** labeling variables, values, and data files
- **drop:** drop one or more variables
- **drop if:** drop observations conditional on one or more variables
- **edit:** editing the data file directly

Data cleaning

- **do:** storing and executing commands in do-files
- **#delimit:** writing long commands in do-files
- **/* */:** documenting your do-files
- finding and fixing outliers
- **duplicates:** finding duplicate ids

Adding summary statistics to a data file

- **egen:** add summary statistics to each observation
- **collapse:** create file of summary statistics by groups

Combining data files

- **one-to-one:** same observations in each file
- **match merge:** many observations in each file match, but some don't
- **one-to-many:** hierarchical data, analysis at the **lower** level
- **merging summary statistics:** hierarchical data, analysis at the **higher** level
- **appending:** adding observations with the same variables

Reshaping a data file

- **reshape long:** change variables to observations
- **reshape wide:** change observations to variables

Documenting Your Work

- **variable labels**
- **value labels**

- **do-file template**

Graphics

- **histogram** with normal curve fitted to it
- **graph box** plot displayed for two groups
- **scatter plot**
- **twoway** scatter plot with regression line
- other resources for learning graphics in Stata

Analyzing Data from Sample Surveys

- **Data characteristics:** stratification, clustering, sampling weights
- **Choosing the correct weight syntax:** pweight, aweight, fweight, or iweight?
- **Commands to analyze survey data:** svy, robust cluster, subpop
- **Logistic Regression Example:** adjust, svylc, svytest
- **Common errors** and how to avoid them

Labor-Saving Techniques

- **looping over variables and values**
- **dummy variables**
- **a simple program**

Miscellaneous Tips and Tricks

- **getting help for Stata**
- **updating Stata**
- **importing and exporting data files**
- **working with large files**
- **shrinking large data files**
- **error messages** and what they mean
- **missing values** and how to work with them
- **the by command** in detail
- **exporting results** to MS Office
- **the parmesst command:** saving Stata results
- **temporary files**
- **looping:** foreach in detail
- **looping with while**
- **precision** and data storage

Authors: Phil Bardsley, Kim Chantala, and Dan Blanchette

Logistic Regression Analysis

The Logit Model

Logistic Regression is used to model dichotomous (0 or 1) outcomes. This technique models the log odds of an outcome defined by the values of covariates in your model. In addition to covering how to model sub-populations, we will use both the svy commands and the robust cluster commands. The following example comes from The National Longitudinal Study of Adolescent to Adult Health.

Research Question: How is being in the upper quartile of the Vocabulary test score (PVT_Q4) influenced by a boy's grade in English (ENGL_GPA) and Family composition (BIOMAPA)?

Predictive Model:

$$\log \left| \frac{\Pr(PVT_Q4 = 1)}{1 - \Pr(PVT_Q4 = 1)} \right| = b_0 + b_1 AGE_KID + b_2 BIOMAPA + b_3 ENGL_GPA$$

Where

b_0 = Intercept

b_1 = Change in log odds of being in upper quartile for one year increment in age

b_2 = Change in log odds of being in upper quartile for living with Biological Parents

b_3 = Change in log odds of being in upper quartile for increase in one grade level

The model predicted log-odds for the categorical subpopulations will be:

BIOMAPA ENGL_GPA Ln(odds)

0 = No	4 = A	$b_0 + b_1 AGE_KID + 4b_3$
0 = No	3 = B	$b_0 + b_1 AGE_KID + 3b_3$
0 = No	2 = C	$b_0 + b_1 AGE_KID + 2b_3$
0 = No	1 = D/F	$b_0 + b_1 AGE_KID + b_3$
1 = Yes	4 = A	$b_0 + b_1 AGE_KID + b_2 + 4b_3$
1 = Yes	3 = B	$b_0 + b_1 AGE_KID + b_2 + 3b_3$
1 = Yes	2 = C	$b_0 + b_1 AGE_KID + b_2 + 2b_3$
1 = Yes	1 = D/F	$b_0 + b_1 AGE_KID + b_2 + b_3$

We are assuming a model with a common slope for age of the boy, but different intercepts defined by grade in English and living with both biological parents.

The relationship between probability and odds

The odds of an outcome is related to the probability of the outcome by the following relation:

$$\text{odds} = \frac{\text{probability}}{1 - \text{probability}}$$

An odds ratio is just the ratio of the odds of the outcome evaluated at two different sets of values for your covariates. It is easy to show that to test the hypothesis that $p_1 = p_2$ you can test that the hypothesis that an odds ratio comparing group 1 to group 2 is equal to 1. However, you cannot easily put a confidence interval on the difference between the two probabilities.

SVY: LOGIT

The **svyset** command is used to specify the design information for analysis. Use the **strata** keyword to specify the stratification variable (region), the **pweight** keyword to specify the probability weight variable (gswgt1), and specify the primary sampling unit (psuscid).

```
svyset psuscid [pweight=gswgt1], strata(region)
```

The **svy: logit** command states the model being tested. The first variable following **svy: logit** denotes the outcome (pvt_q4) of our model, and the following variables are the covariates. The option **subpop**

is used to specify the sub-population we want to be used to compute parameter estimates. All 18,924 observations are needed for the variance computation because Stata determines the design information (number of primary sampling units) used in the formula variance computation.

```
svy, subpop(male): logit pvt_q4 age_kid biomapa engl_gpa
```

Stata lists the number of observations with no missing values for the variables in the model ($N=17,191$) and has summed the corresponding sample weights to estimate 19,955,620 adolescents in the U.S. are represented by these observations. The number of observations with complete data in the sub-population is 8,366 representing 10,084,117 boys. Note that the number of strata (4) and primary sampling units (132) has been correctly counted.

Survey: Logistic regression

Number of strata	=	4	Number of obs	=	17191
Number of PSUs	=	132	Population size	=	19955620
			Subpop. no. of obs	=	8366
			Subpop. size	=	10084117
			Design df	=	128
			F(3, 126)	=	49.14
			Prob > F	=	0.0000

	Linearized					
	twokids	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age_kid	-.0451845	.0278879	-1.62	0.108	-.1003656	.0099965
biomapa	.4273138	.0820139	5.21	0.000	.2650354	.5895923
engl_gpa	.4258579	.0423055	10.07	0.000	.3421493	.5095664
_cons	-1.886177	.4411884	-4.28	0.000	-2.759144	-1.013211

The **adjust** command can be used to estimate a linear combination of the coefficients estimated for the variables in our model. If you do not specify a value for a variable when using **adjust**, Stata will incorrectly substitute the sample mean rather than an estimate of the population

mean. ***This is because adjust ignores any weights used by the estimation commands.*** (See *Stata Reference Manual, Release 9, Vol 1 A-G*, page 10.) To correctly compute a linear combination, it is necessary to specify a value for all variables in the model. For example, the following statement:

```
adjust age_kid=17 engl_gpa=3, by(biomapa) xb se ci
```

produces an estimate of the log odds of scoring above the 75th percentile for boys at age 17 with a grade of B in English for both categories of living with both biological parents:

```
Dependent variable: pvt_q4      Command: logit
Covariates set to value: age_kid = 17, engl_gpa = 3

Live with |
Bio Mom & |
Dad
0=N/1=Y |      xb       stdp       1b       ub
-----+
  0 | -1.37674  (.100564) [-1.57572  -1.17776]
  1 | -.949427  (.094665) [-1.13674  -.762115]

Key: xb      = Linear Prediction
      stdp    = Standard Error
      [1b , ub] = [95% Confidence Interval]
```

You can also include the **exp** option at the end of the **adjust** command to get exponentiated linear combinations of the coefficients. The **pr** option on **adjust** is not available after using the **svylogit** command.

The **lincom** command can also be used to produce linear combinations of the coefficients:

```
lincom 17*age_kid + 1*biomapa + 3*engl_gpa + _cons
( 1) 17.0 age_kid + biomapa + 3.0 engl_gpa + _cons = 0.0

pvt_q4 | Coef. Std. Err.      t   P>|t|   [95% Conf. Interval]
-----+
(1) | -.9494267 .0946653 -10.03 0.000  -1.136738  -.7621154
```

The results from **lincom** match those from **adjust**. The advantage of using **lincom** is that a hypothesis test can also be performed. For example, suppose you want to compute the odds ratio comparing 17 year-old boys not living with both biological parents to 12 year-old boys living with both biological parents. Assume both boys make the same grade in English. We would want to estimate the difference in log odds for these to:

$$(b_0 + 17*b_1 + GRADE*b_3) - (b_0 + 12*b_1 + b_2 + GRADE*b_3) = 5*b_1 - b_2$$

Since b_1 is the coefficient for AGE_KID and b_2 is the coefficient for BIOMAPA, the **lincom** command would be:

```
lincom 5*age_kid - 1*biomapa
```

This produces the desired difference in log odds:

(1) 5.0 age_kid - biomapa = 0.0					
pvt_q4	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	-.6532364	.1641137	-3.98	0.000	-.9779635 -.3285094

The **or** option can be added to the lincom command to get the odds ratio ($e^{5^*b_1-b_2}$):

```
lincom 5*age_kid - 1*biomapa , or
```

The following table will be printed:

(1) 5.0 age_kid - biomapa = 0.0					
pvt_q4	Odds Ratio	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	.5203589	.085398	-3.98	0.000	.3760762 .7199962

Thus, assuming equal grades in English, the odds of a 17 year-old boy not living with both biological parents is only half that of a 12 year boy who lives with his biological parents.

The **test** command can be used to test joint hypothesis about variables. For example, testing that the coefficient for age_kid and biomapa are both equal to zero can be done with the following stata command:

```
test age_kid biomapa
```

which produces the following output:

```
Adjusted Wald test
( 1) age_kid = 0.0
( 2) biomapa = 0.0

F( 2,    127) =   14.51
Prob > F =     0.0000
```

logit with pweight and robust cluster

Note that we can subset the data (**if male == 1**) when using the **robust cluster()** options in Stata and still have the variance computed with an acceptable technique. The primary sampling unit (psuscid) is used as the argument to the **cluster** option and the sample weights (gswgt1) are specified by [**pweight=gswgt1**].

```
logit pvt_q4 age_kid biomapa engl_gpa if male == 1 [pweight=gswgt1], robust cluster(psuscid)
```

The results and interpretation in the following output are identical to the results from svylogit.

Logistic regression		Number of obs = 8366			
		Wald chi2(3) = 142.44			
		Prob > chi2 = 0.0000			
Log pseudolikelihood = -4429.7883		Pseudo R2 = 0.0384			
(Std. Err. adjusted for 132 clusters in psuscid)					
pvt_q4	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
age_kid	-.0451845	.0277835	-1.626	0.104	-.0996393 .0092702
biomapa	.4273138	.0817512	5.227	0.000	.2670844 .5875433
engl_gpa	.4258579	.0430438	9.894	0.000	.3414936 .5102222
_cons	-1.886177	.4414855	-4.272	0.000	-2.751473 -1.020881

Review again?

Another topic?

Looping over variables and values

We often need to run the same command for a large number of variables. For example, we might want to change the value 9 to missing for 200 variables in a data file. We can type the recode command 200 times, or we can type the **foreach** command once and let it create 200 copies of recode for us.

The **foreach** and **forvalues** commands are convenient ways to save you typing. In addition to the above recode situation, they can also rename a group of variables for us, saving us typing many rename commands. In fact, any command, or set of commands, that you need to repeat over a group of variables, is a likely candidate for one of these labor-saving commands.

Two examples of **foreach** and **forvalues** are shown below.

To see an engaging discussion of the topic by Nicholas Cox of the University of Durham, UK, see this PDF file: How To Face Lists With Fortitude. This article was also published in Stata Journal 2(2), 2002.

```
clear

use "q:\utilities\statatut\examfac2.dta"

su q102_* q103_*
su q102_* q103_*

/* Replace all values of 99 with missing in q102_* and q103_* */

foreach x of varlist q102_* q103_* {
    replace `x'=. if `x' == 99
}

/* Rename q102_* to title* and rename q103_* to fphour* */

forval x=1/20 {
    rename q102_`x' title`x'
    rename q103_`x' fphour`x'
}
exit
```

TIP: To see how each pass through the loop is resolved, do the following.

```
clear
use "q:\utilities\statatut\examfac2.dta"
set tracedepth 1 // only show one level down
set trace on // turn on Stata's trace option

foreach x of varlist q102_* q103_* {
    replace `x'=. if `x' == 99
}
set trace off // return to normal Stata mode
```

Questions:

1. What is the "*" in the foreach and forvalues commands? Answer.
 2. What purpose does the word "varlist" serve in the **foreach** command? Answer.
 3. What is the "x" in each of the commands? Answer.
 4. The **foreach** and **forval** commands have a couple characters I haven't seen before in this tutorial: {} and `. What are they? Answer.
 5. Why does the **forval** command take up more than one line? Answer.
-

Answers:

1. The asterisk (*) after a variable name is a shortcut in Stata. You can use it in any Stata command, not just this one. It tells Stata to look for all variable names that begin with "q102_" and "q103_" and end with anything. We know that they end with the numbers 1-20. It's the same as typing all 40 variable names:

```
.....  
foreach x of varlist q102_1 q102_2 q102_3 ... q103_19 q103_20 {  
.....
```

[Back to question](#)

2. The word "varlist" in the **foreach** command tells Stata that we are referring to a list of existing variables. The foreach command has other options, such as "newlist" for generating a list of new variables.

[Back to question](#)

3. The "x" is itself a variable that stands for each variable in the foreach variable list, or each value in the forvalues number list. In the first example, the command following the "{" is repeated once for each variable in the variable list, substituting the real variable name for the "x" in the replace command. In the second example, "x" is substituted for each number in the list 1/20. Note that it need not be "x" - any variable name will do, but "x" is quick and easy to type.

The net result of the **foreach** command is the same as typing the following 40 times:

```
replace q102_1=. if q102_1==99
replace q102_2=. if q102_2==99
etc.
```

[Back to question](#)

4. The braces {} in the **foreach** and **forval** commands surround the commands that you want to execute for each variable. In the **foreach** command there is one command (replace). But in the **forval** command we have inserted two commands between the braces. The other character is an accent mark. On a standard US keyboard it is on the same key with the tilda (~) on the left side next to the 1 key. This character tells Stata that the character following is a special kind of variable known as a "local macro" in Stata. A macro (local or global) is temporary and does not become part of the data in memory. A macro must be surrounded by an accent on the left and an apostrophe (also called a "single quote") on the right like this: `m'

[Back to question](#)

5. Stata requires separate lines for each part of these commands. Here is how Stata's help lays out the syntax rules:

- the open brace must appear on the same line as "foreach" or "forvalues"
- nothing may follow the open brace except, of course, comments
- the first command to be executed must appear on a new line
- the close brace must appear on a line by itself

Alternatively, we can change the command delimiter to semicolon (;) and put the entire command on one line like this:

```
#delimit ;
forval x=1/20 {; rename q102_`x' title`x'; rename q103_`x' fphour`x'; };
```

Frankly, that's pretty hard to read. Using separate lines and indentation looks much nicer. In fact, if you copy the above lines into your do-file editor and run them, you'll see in the Results window that Stata will improve your code by using separate lines and indentation - at least your log will be easy to read!

[Back to question](#)

[Review again?](#)

Another topic?

Looping with while

while: another way to repeat commands.

The **while** command can be used in much the same way as **foreach** or **forvalues**, but it has greater flexibility. While it's generally used by programmers (writing commands), it is a tool available for do-files as well.

You can copy the following text into a do-file and run it to see how **while** works.

```

clear
use "q:\utilities\statatut\examfac2.dta"
su
list q102_* in 1/10
***** from log *****
    q102_1      q102_2      q102_3      q102_4      q102_5
1. PH NurseB        Other    ClinOff      99      99
> 2. ClinOff          99        99        99        99
> 3. NurseOff    PH NurseB    PH NurseB    MCH Aide    MCH Aide
> 4. ClinOff      ClinOff  Nurse/Midwife  MCH Aide      99
> 5. MCH Aide    NurseAssist    ClinOff      99      99
> 6. Nurse/Midwife  ClinOff  Nurse/Midwife      99      99
> 7. NurseAssist  NurseAssist        99        99        99
> 8. ClinOff      NurseAssist  NurseAssist      99        99
> 9. ClinOff      NurseAssist  NurseAssist      99        99
> 10. Nurse/Midwife  99        99        99        99
>
*****
*/
**
Notice that q102_* vars appear to be character data not numeric.
The value label "title" is associated to these variables formats
the numeric data to look like character data. Click here to see webpages about
variable labels and value labels.
*/
**

/** Say you want to disassociate the value label title from the q102_* vars.
You could type: label value q102_1 ;
               label value q102_2 ;
               label value q102_3 ;
               label value q102_4 ;
               label value q102_5 ;

Or you could use a while command to disassociate
the value label title with the q102_* vars. **/


local i=1
while `i' <= 5 {
    label value q102_`i'
    local i= `i' + 1
}

```

```
/** It is possible to write a while command without using a local macro variable.
Users generally use local macro variables because they are not associated
to the data set. It's not possible for their values to be different from
one observation to the next and they can hold a number or a string of text.

In this example, the while command uses the local macro variable i to step
through the list of variables q102_1 q102_2 q102_3 q102_4 q102_5. `i' is
first evaluated to be 1 and then 2 and so on until the expression `i'<=5 is
not true (when `i'=6). The brackets "{" and "}" enclose all commands to be
processed by the while command. **/


list q102_* in 1/10

/* Create string variables that can hold up to 13 characters. */

local j=1 /* local i=1 would also work. i is not the only local macro variable name allowed */
while `j' <= 5 {
    gen str13 title`j'=
    local j= `j' + 1
}

/** Rename q102_* vars */
local hi= 1
while `hi' <= 5 {
    rename q102_`hi' title`hi'
    local hi= `hi' + 1
}

list title1-title5 in 1/10

local hi= 1
while `hi' <= 10 {
    replace title1= "not missing" if title1 == `hi'
    local hi= `hi' + 1
}

/* More than 1 list of variables can used in while command. */

/* Replace all values of 99 with missing in title1-title5 and q103_* */

su title5-q103_1

local i= 1
while `i' <= 5 {
    replace title`i'= . if title`i' == 99
    replace q103_`i'= . if q103_`i' == 99
    local i= `i' + 1
}

su title5-q103_1

/** Create new variables from existing variables. */
local i=1
quietly while `i' <= 5 {
    gen mch_h`i'= q103_`i' if title`i' == 8
    label var mch_h`i' "MCH Aid's hours"
    local i= `i' + 1
}

/** NOTE: Adding "quietly" to a STATA command tells STATA not to print any output for the command.
All lists of variable have to have the same number of elements.
*****/


list mch_h1-mch_h5 in 11/15

/** NOTE: STATA reads the shorthand varlist of mch_h1-mch_h5 to be all
variables positionally in the data set between mch_h1 and mch_h5 (not
necessarily mch_h1 mch_h2 mch_h3 mch_h4 mch_h5). When using STATA
interactively, the variable list window in the lower left-hand side
shows the list of variables in the data set in the order of their position. **/
```

```
/* Here is an example using while to run the decode command  
with all title1-title5 variables:  
  
local i= 1  
while `i' <= 5 {  
    decode title`i', gen(title`i'c)  
    local `i'= `i'+1  
}  
*****
```

Note: Dan Blanchette contributed this web page, however please direct questions to Phil Bardsley as noted below.

Review again?

Another topic?

Looping: foreach in detail

More examples of foreach and forvalues commands.

Stata processes all observations of the data set for each element in the list of the **foreach** and **forvalues** commands.

This page shows 4 different types of lists that can be used in the **foreach** command: varlist, newlist, numlist, and anylist and how to use the **forvalues** command.

Both **foreach** and **forvalues** commands use "local macro" variables in processing whatever list they loop through. A local macro variable is not a variable in the data set but rather a variable available to Stata for programming purposes. To learn more about local macro variables use Stata's help by searching for "macro". Here's a quick example:

```
. local X= "MyVarName" // create local macro variable X and set it equal to the text MyVarName.
. display "My variable name is: `X'"

"My variable name is: MyVarName"
```

Notice that the local macro variable X is enclosed in a left and right single quote. The left quote is under the tilde (~) key on the left side of your keyboard and the right quote is under the double quote ("") key on the right side of your keyboard. Enclosing a letter or variable name in left and right quotes tells Stata to evaluate it as a local macro variable.

NOTE: the following is a **do-file**. The **foreach** and **forvalues** commands are better suited to a **do-file** than to interactive use of Stata. Highlight and copy all text between the horizontal bars and paste into the Stata interactive do-file window so that you can run this do-file.

```
clear
use "q:\utilities\statatut\examfac2.dta"
summarize

list q102_* in 1/10
*****
      q102_1      q102_2      q102_3      q102_4      q102_5
1.    PH NurseB      Other      ClinOff      99      99
> 2.    ClinOff      99      99      99      99
> 3.    NurseOff    PH NurseB    PH NurseB    MCH Aide    MCH Aide
> 4.    ClinOff    ClinOff    Nurse/Midwife    MCH Aide      99
> 5.    MCH Aide    NurseAssist    ClinOff      99      99
> 6.    Nurse/Midwife    ClinOff    Nurse/Midwife      99      99
> 7.    NurseAssist    NurseAssist      99      99      99
> 8.    ClinOff    NurseAssist    NurseAssist      99      99
> 9.    ClinOff    NurseAssist    NurseAssist      99      99
> 10.   Nurse/Midwife    99      99      99      99
*****
*****
```

```
/***
Notice that q102_* vars appear to be character data not numeric.
The value label "title" is associated to these variables formats
the numeric data to look like character data. Click here to see webpages about
variable labels and value labels ***/

/** use varlist for a list of variables that already exist in a data set */

/** Say you want to disassociate the value label title from the q102_* variables.
You could type: label value q102_1
                label value q102_2
                label value q102_3
                label value q102_4
                label value q102_5

Or you could use a varlist in a foreach command to disassociate
the value label title with the q102_* vars. **/


foreach X of varlist q102_1 q102_2 q102_3 q102_4 q102_5{
    label value `X' /* Notice that the capital letter X is enclosed in left and right quotes.
                      * This tells Stata to evaluate the local macro X. */
    display "label value `X'" // Displays in the results window/log file what commands Stata processed.
                      // This is completely unnecessary but helpful in explaining the looping.
}

list q102_* in 1/10

*****
In both cases the value label title will still be available for use
with other variables or even to be re-used again on the q102_1-q102_5 variables.
If the value label title is not associated to any variable when the data set
is saved then it will be dropped.
*****


/* Now re-assign the title value label to each q102_* variable.
This time use a local macro variable in the foreach command. **/


local titles "q102_1 q102_2 q102_3 q102_4 q102_5"

foreach X of varlist `titles' {
    label val `X' title
}

/** NOTE: The apostrophe on the left is a left apostrophe.
It is on your keyboard next to the "1" key.
When a local macro variable is enclosed in a left
and right apostrophe, it is evaluated as the
contents of that local macro variable. **/


list `titles' in 1/10

/** NOTE: A local macro variable can be used many times throughout a program and
thus can save a lot of typing. It can also help keep a foreach command
from looking very messy, like if you wanted to pass through a foreach
command 10 or more variables that could not be represented in shorthand
(like q103_1-q103_10 can be).

NOTE: `titles' represents the text "q102_1 q102_2 q102_3 q102_4 q102_5".
Local macro variable names can be up to 7 (not 8) characters long in Stata 6.
Stata 7 allows them to be up to 31 characters long.

*****


/* Use newlist for variables that are created/generated by the foreach command. */

/* Create string variables title1c, title2c, title3c, title4c and title5c that can hold up to 13 characters. */
```

```

foreach X of newlist title1-title5{
    gen str13 `X'c= ""
}

/** NOTE: A foreach command like:
foreach X of newlist title* {
    gen str13 `X'c= ""
}
would not work because Stata wouldn't know how many title variables to create. **/


/** Use numlist to insert numbers in place of the local macro var X into a foreach command. **/


/** Use numlist to rename q102_* vars in a foreach command **/


foreach X of numlist 1/5 {
    rename q102_`X' title`X'
    /** The local macro variable X is replaced with
        the numbers 1, 2, 3, 4, 5 ***/
}

list title1-title5 in 1/10

/** NOTE: The above foreach command could have been written with a forvalues command:
* forval X= 1/5 {
*     rename q102_`X' title`X'
* }
* The capital letter X and could be any letter, upper or lower case.
* A good rule of thumb is have all variable names in lowercase.
**/


/** The anylist is a list of words. Shorthand notation like title* or
title1-title5 does not work. **/


foreach X in 1 2 3 4 5 6 7 8 9 10 {
    replace title1c="not missing" if title1==`X'
}

/* The anylist is the most universal type of foreach command since it is not
restricted to pre-existing variables, new variables, or just numbers. */


/* More than 1 list of variables can processed at a time. */

/* Replace all values of 99 with missing in title1-title5 and q103_* */

summarize title5-q103_1

foreach X of varlist title1-title5 q103_* {
    replace `X'= . if `X' == 99
}

summarize title5-q103_1


/** Two or more lists or types of lists can be processed at the same time. **/


/** Create new variables from existing variables in a foreach command. **/


/* First create 2 variable list and store them in local macro variables using the
 * the unab Stata command: */
unab varlist1 : title1-title5
unab varlist2 : q103_1-q103_5

/* Use the display command to see that local macro varlist1 contains the string:
 * q103_1 q103_2 q103_3 q103_4 q103_5 */
di "varlist1"

foreach X of newlist mch_h1-mch_h5 {
    local n= `n' + 1 /* keep count of how many times the loop is processed */
    /* Use the extended macro function ": word # of" to set the local macro variable Y to the n'th variable name */
}

```

```
local Y : word `n' of `varlist1'  
local Z : word `n' of `varlist2'  
gen `X'= `Z' if `Y' == 8  
label var `X' "MCH Aid's hours"  
}  
  
/* ** NOTE: All lists have to have the same number of elements. **/  
  
/* Since all the variables involved have numbers involved in their name, using the forval command would be simpler:  
forvalues X= 1/5 {  
    gen mch_h`X'= q103_1`X' if title`X' == 8  
    label var mch_h1`X' "MCH Aid's hours"  
}  
*****  
list mch_h1-mch_h5 in 23/31  
  
/* ** NOTE: Stata reads the shorthand varlist of mch_h1-mch_h5 to be all  
variables positionally in the data set between mch_h1 and mch_h5 (not  
necessarily mch_h1 mch_h2 mch_h3 mch_h4 mch_h5). When using Stata  
interactively, the variable list window in the lower left-hand side  
shows the list of variables in the data set in the order of their position.  
The describe command also shows variable order. **/
```

Note: Dan Blanchette contributed this web page, however please direct questions to Phil Bardsley as noted below.

Review again?

Another topic?

Match merging

Many observations match, but others don't

In this example, we have data from the 1996 Tanzania Facility Survey that we want to merge with data from the 1999 survey. This will allow us to examine trends. The 1999 survey was not designed as a followup of the 1996 survey, so not all the same facilities were visited in the two surveys. To simplify the programming, we've only included facilities in the 1999 file that were also surveyed in 1996.

For this example, we'll look at trends in the availability of the 4 major contraceptive methods, so the two files contain only those 4 variables in addition to the single identifier: facid.

Prior to Stata 11, the data had to be sorted on the identifying variables before merging. Now you have a choice of syntax for merging. We've shown you both methods below, the old which requires sorting, and the new which does not (Stata will sort the data for you when you use the new syntax). Both the old and new syntax work in Stata 11 and after, but the new syntax is better.

The syntax is the same as the "One-to-one merging" example, and in fact this is one-to-one merging. The difference is that in the earlier example we expected a perfect match, but here we don't.

You can type the following commands into Stata, or copy them into a do-file and run them.

Merge using New Syntax (Stata 11 and later)

```
clear
use "q:\utilities\statatut\merge96.dta"
merge 1:1 facid using "q:\temp\statatut\merge99.dta"
drop _merge
/* Look at the Results Window - the merge is a mess! */
```

Merge using Old Syntax (Stata 10 and earlier)

```
/* Sort the 1999 file and save it temporarily */
clear
use "q:\utilities\statatut\merge99.dta"
sort facid
save "d:\statatemp\temp99.dta"

/* Sort the 1996 file */
clear
use "q:\utilities\statatut\merge96.dta"
```

```

sort facid

/* Merge the 1999 data onto the 1996 data */

merge facid using "d:\statatemp\temp99.dta"

/* Check how well the merge went - we don't expect all "3's" */

tab _merge
drop _merge
su

/* Clean up */

erase "d:\statatemp\temp99.dta"

```

Questions:

1. This merge is a mess! Less than half of the facilities in 1996 have a match in 1999. How can we get rid of the non-matching observations? Answer.
 2. The 1996 and 1999 files each have 5 variables. How many variables do we expect the merged file to have? Answer.
 3. This is a simple example with only 5 variables in each input file. So, it's easy to check whether I've used different variable names in each file. What would happen if I accidentally had a variable in each file with the same name? Answer.
 4. How would I know if I had a variable with the same name on each file? Answer.
 5. How many data files can I merge with one command? Answer.
-

Answers:

1. Only 207 of the facilities matched (_merge==3). The remainder were only in the 1996 "master" file (_merge == 1) except for one facility in the 1999 "using" file (_merge == 2). To get rid of the non-matches, we need to use _merge: **keep if _merge == 3**

You can type this command now and see the result. Stata drops 274 observations from the 1996 data in memory and one observation from the 1999 data. The remaining 207 observations are the matched 1996-1999 facilities. Now we can drop _merge, since we no longer need it.

[Back to question](#)

2. The resulting file should have facid (common to both files), 4 more variables from each input file, and _merge, for a total of **10 variables**
-

[Back to question](#)

3. Stata by default keeps the values of the "master" file when a merge involves variables with the same name in both files. We planned for this default by renaming the 4 method variables to include the survey year in both files. If we had not done that, the values in the 1999 "using" file would have been lost for matching facilities.

The merge command includes an **update** option to override this default. However, you probably won't need that. It's best to always rename your variables so that the master and using files have unique variable names.

[Back to question](#)

4. The merged file would have one fewer variables than you expected. That is the only way you would know; Stata does not issue a warning.

[Back to question](#)

5. In prior versions (8-10) of Stata, you could merge 3 or more files in a single merge command. Frankly, we thought that merging more than 2 files at a time was error prone. There are many ways to miss problems in a merge, and these are compounded when more files are merged at the same time.

So, we're happy to see that in Stata 11 and after you can only merge 2 files together at a time (using the new syntax).

[Back to question](#)

[Review again?](#)

[Another topic?](#)

Merging summary statistics onto each observation

Hierarchical data, analysis at the higher level

This example combines the **collapse** command learned in Adding summary statistics to a data file with the **merge** command covered earlier. We'll use the same data here as the previous example: 503 facilities (which we call "higher" level) in one file, and 2,584 service providers (which we call "lower" level) in another file. But this time we want to combine them in such a way that each observation is a facility, i.e., we want to do our analysis at the higher level.

Our analysis question for this example might be: is the mean number of trained providers at government facilities higher than at private facilities? We need to count trained providers in the provider file (using the **collapse** command) to summarize their data at the facility level, and merge the resulting file onto the facility data.

We use only the new merge syntax here. Please refer to the previous examples of **merge** for Stata 10 and earlier syntax.

Copy these commands into a do-file editor and run them.

```
/* Use the facility file and drop the duplicate. */

clear
use "q:\utilities\statatut\exampfac.dta"
keep facid authorit
duplicates list facid
drop if facid == 1001 & missing(authorit) // get rid of the duplicate
su
save "d:\statatemp\tempfac.dta", replace

/* Collapse the service provider file to count the trained providers per facility */

clear
use "q:\utilities\statatut\exampro.dta"
keep facid bcs ccs
su
replace bcs=0 if bcs==2
replace ccs=0 if ccs==2
sort facid
collapse (sum) bcs ccs, by (facid)
list in 1/10
su

/* Merge the collapsed provider data with the facility data */

merge 1:1 facid using "d:\statatemp\tempfac.dta"

/* Check how well the merge went - we don't expect all "3's" */

keep if _merge == 3
su

/* Recode authorit to give 2 groups: govt and non-govt */
/* Note that the private group is labeled "MarieStopes" */
/* because that's the label on value 2. */
```

```
recode authorit 3/12=2  
ta authorit  
  
/* Run a t-test to check difference in mean training Levels */  
  
ttest bcs, by(authorit)  
  
/* Clean up */  
  
erase "d:\statatemp\tempfac.dta"  
drop _merge
```

Questions:

1. Why recode bcs and ccs to 0/1 before the collapse command? Answer.
 2. After collapsing the provider data to the facility level, we merged a facility-level file with another facility-level file. Could we have merged them without using facid, in other words, could we have done a one-to-one merge instead of a match-merge? Answer.
 3. What were the results of the match-merge? Answer.
 4. Do we need to drop the non-matching observations before running the t-test? Answer.
 5. Could we use **egen** instead of collapse in this situation? Answer.
-

Answers:

1. We need to count the number of trained providers at each facility. The easiest way to do that is to add the people who are trained, so we recode them to 1=trained and 0=not trained.

[Back to question](#)

2. No. In the earlier example, where we merged two files of DHS women's data, the two files originally came from the same combined file. So, we were confident that by sorting first, the women would be in the same order and match one-to-one without using their identifying variables. (We matched on identifiers anyway just to be safe.)

In this case, though, the provider data were collected separately from the facility data. Theoretically, all facilities should be represented in the provider data. But practically, given all the ways field surveys can and do go wrong, it didn't work out that cleanly. Therefore, we needed to match-merge.

[Back to question](#)

3. The results were the same as in our previous example using these two files. 468 matches, 34 facilities with no matching provider data, and 9 (facility-level) provider observations with no matching facilities. These 9 provider observations are collapsed from the 22 providers we saw unmatched in the previous example.

[Back to question](#)

4. No we don't. The t-test, and most other Stata statistical commands, drop observations with missing values on any one of the analysis variables. In this case we have 467 observations in the analysis instead of 468, because 1 facility has a missing value for authority.

In fact, if we were to do this analysis using the **svy** commands, we would need to leave all facilities in the sample. The svy commands correct the standard error for the effects of clustering and stratification, and they require the full set of observations in order to make this correction accurately. See Commands to Analyze Survey Data for more information on this topic.

[Back to question](#)

5. Yes, we could use **egen** to count the number of trained providers, drop the duplicates in the provider file so there was only one observation per facility, and merge the result onto the facility file:

```
/* Use egen to count the trained providers per facility */

clear
use "q:\utilities\statatut\exampro.dta"
keep facid bcs ccs
replace bcs= 0 if bcs == 2
replace ccs= 0 if ccs == 2
sort facid
by facid: egen numbcs= total(bcs)
by facid: egen numccs= total(ccs)
list in 1/10
duplicates drop facid, force
drop bcs ccs
su

/* Merge the collapsed provider data onto the facility data */

merge 1:1 facid using "tempfac.dta"
```

There are a couple of advantages to the egen approach. First, you can easily keep additional variables that may be of interest on the provider file. While this is possible with collapse, it can be cumbersome. Second, you can browse the data after egen to see whether you created the count correctly. Then, if it looks good, you can drop the duplicates.

[Back to question](#)

Review again?

Another topic?

Miscellaneous Tips and Tricks

You might find it useful to browse these topics from time to time to see whether they're useful for your data management work.

- Getting help for Stata
- Updating Stata
- Importing and exporting data files
- Working with large data files
- Shrinking large data files
- Error messages
- Missing values
- The by command in detail
- Exporting Stata Results to MS Office
- The parmesst command
- Temporary files
- Looping: foreach in detail
- Looping with while
- Precision and data storage

Missing values

Missing Values and how to work with them

Stata represents missing values with a "." (period) in its results window. If the file has special missing values, Stata represents them as ".a", ".b", ..., ".z". Missing values are stored as values larger than the largest allowable number for the data type. For example, a variable stored as data type **byte** can take the following values:

You see:	Stored as:	Treated as:
1	1	1
2	2	2
...
100	100	100
.	101	missing
.a	102	missing
.b	103	missing
...
.z	127	missing

Most commands ignore missing values by default. Some commands, such as **tabulate**, have an option to display missing if you want to see how many missing observations there are. Other commands, however, may use missing values in a way that will surprise you. For example, the **replace** command does not ignore missing values. Here is a simple example to demonstrate how replace and recode handle missing values differently. In this example, we have three variables with the values of 1, 2, and missing. We want to change all values of 2 to 1.

```
clear
input a b c
1 1 1
2 2 2
.
.
end
list
replace a=1 if a>1
replace b=1 if b>1 & b<.
recode c 1/max=1
list
```

The first **replace** command changes every value that's greater than 1 to 1. This command does not ignore missing values, so both 2 and missing are changed to 1. This probably is not what we would normally want to do, since missing values should remain missing.

The second **replace** command changes all values greater than 1 but less than missing to 1. In this case values of 2 are changed and missing values are not changed, which is our intention.

The **recode** command automatically ignores missing values, so we don't have to think about it. The results are the same as the second replace command.

Review again?

Another topic?

One-to-many merging

Hierarchical data, analysis at the lower level

In this example, we have data from the 1999 Tanzania Facility Survey. We have data from 503 facilities (which we call "higher" level) in one file, and data from 2,584 service providers (which we call "lower" level) in another file. We want to combine them so that each observation is a service provider, i.e., we want to do our analysis at the lower level.

Our analysis question for this example might be: are providers from government facilities more likely to be trained than those from private facilities?

Copy these commands into a do-file editor and run them.

Merge using New Syntax (Stata 11 and later)

```
clear
use "q:\utilities\statatut\exampfac.dta"
keep facid authorit
merge 1:m facid using "q:\utilities\statatut\exampro.dta", keepusing(facid bcs ccs)
drop _merge
/* Check how well the merge went - we do not expect all "3's" */
```

Merge using Old Syntax (Stata 10 and earlier)

```
/* Use the provider file and sort it */

clear
use "q:\utilities\statatut\exampro.dta"
keep facid bcs ccs
sort facid
save "d:\statatemp\tempro.dta", replace

/* Use the facility file and sort it */

clear
use "q:\utilities\statatut\exampfac.dta"
keep facid authorit
sort facid

/* Merge the provider data onto the facility data */

merge facid using "d:\statatemp\tempro.dta"
/* Check how well the merge went - we don't expect all "3's" */
```

```

tab _merge
keep if _merge == 3
drop _merge
su

/* Recode authority to form government vs private */

recode authorit 3/12=2

/* Run chi-square analysis */

tab authorit bcs, row chi

/* Clean up */

erase "d:\statatemp\tempro.dta"

```

Questions:

1. I used the new Stata syntax, and the merge failed! I got an error message:

variable facid does not uniquely identify observations in the master data

What does this mean, and how can I fix it? Answer.

2. Using the Stata 11 and later syntax, can I first use providers and then merge facilities onto them? Answer.

3. I used the Stata 10 and earlier syntax, and the merge worked, but I got two notes:

variable facid does not uniquely identify observations in the master data
variable facid does not uniquely identify observations in tempro.dta

What does this mean, and do I need to worry about it? Answer.

4. This merge is not too bad - about 98% of the service providers were matched with their appropriate facilities. How many providers had no matching facility data, and how many facilities weren't paired with providers? Answer.

5. Is it possible to tell from the tabulation of _merge that we're getting the correct number of providers in the merge? How about the correct number of facilities? Answer.

Answers:

1. The problem is that the facility ("master") data has duplicates, that is, it contains more than one observation with the same value of facid. The good news is that Stata stopped you from continuing with the merge, getting the wrong answer, and possibly never discovering the error. (Earlier versions of Stata allowed you to continue, but with a warning - see Answer 2.) The bad news is that you should have checked for duplicates before doing this merge!

Earlier, you saw how to check directly for duplicate identifiers:

```
use "q:\utilities\statatut\exampfac.dta", clear
duplicates list facid
```

That is the easiest way to check for duplicates, and it should always be done prior to a merge.

The error message in our results window refers to the master file, which in this case is the facilities. There may be multiple providers from each facility, so we expect duplicates of facid in the provider (using) file. We told Stata to expect duplicate providers by putting "m" on the right side (corresponding to the using file) of "1:m" in the merge command. Similarly, we told Stata to expect unique values of facid in the facility file by putting "1" on the left side (corresponding to the master file).

Using the duplicates list command, we discovered that our facility data has two facilities with facid == 1001. One of these is a mistake, because its value of authorit is missing:

	facid	authorit
12.	1001	Gov
13.	1001	.

To fix the problem we drop that facility and continue with the merge. Back to question.

2. Yes, you can merge facilities and providers together in either order. To merge in the other order, the syntax (after fixing the duplicates problem) is:

```
clear
use "q:\utilities\statatut\exampro.dta"
keep facid bcs ccs
merge m:1 facid using "q:\utilities\statatut\exampfac.dta", keepusing(facid authorit)
```

The terms "1:m" and "m:1" tell Stata what to expect when it tries to match observations in the two files. If it finds what you tell it to expect, your merge will be successful. But, if Stata doesn't find what you expect, it will stop the merge. The only case where it will continue is where you tell it there are many observations with the same identifiers, but in fact there is only 1 instance of each identifier. Since the resulting merged files will be correct, there's no reason for Stata to stop or even to warn you.

Stata also allows a "many-to-many" merge (m:m), but we recommend that you forget you ever heard this. In over 30 years of managing survey data, we have never found an instance where Stata's implementation of this concept fit. In most situations it would result in the wrong answer, and as seen in question 3, that is the case here as well. Back to question.

3. Please read the answer to Question 1 first, then continue here.

As we see here, it's possible to ignore these notes and continue as though the merge worked properly, when in fact we got the wrong answer. (We could even use the new syntax "merge m:m" and continue as though all is well, but the merge would be incorrect.)

Specifically, here's what happens when we ignore the warnings:

As we saw earlier, one of the facilities with facid 1001 is a mistake, because its value of authorit is missing:

	facid	authorit
12.	1001	Gov
13.	1001	.

Our provider data contains 5 observations from facility 1001:

	facid	bcs	ccs
29.	1001	1	1
30.	1001	2	2
31.	1001	2	2
32.	1001	2	1
33.	1001	1	2

When we allow the merge to continue, our mistake will be hidden from us. Look at the result of the merge by facid:

	facid	authorit	bcs	ccs	_merge
12.	1001	Gov	1	1	3
13.	1001	.	2	2	3
521.	1001	.	2	2	3
522.	1001	.	2	1	3
523.	1001	.	1	2	3

Stata matches the first facility, which has a value of "Gov" for authorit, with the first provider. This match is correct. Then it matches the second facility, which has a value of "." (missing) for authorit, with the second provider. It has no new facilities to match with the third provider, so it uses the second facility again, repeating this mistake for the remaining providers from facility 1001.

Despite these mistakes, we think the merge is fine when we see all those 3's in the _merge column. The best way to prevent the error is to check for duplicate facilities **before the merge**. Back to question.

4. 22 providers had no facility data, and 34 facilities had no provider data. The facilities were in memory for the merge (the "master" data), so if they had no matches they received a value of 1 for `_merge`. The providers were on disk (the "using" data), so if they had no matches they received a value of 2 for `_merge`. Back to question.

5. The tabulation of `_merge` tells us that we have 22 providers with no matching facilities (`_merge==2`) and 2562 providers with matching facilities (`_merge==3`). The total is 2584, which is the number of providers in the original file, so all are accounted for.

There's no simple check like this for the number of facilities. Back to question.

Review again?

Another topic?

One-to-one merging

Same observations in each file being merged

The first example is a one-to-one merge of DHS data. The women's file in DHS has more variables than the 2,047 limit in Intercooled Stata. This is not a problem at CPC, where we are running Stata/SE with a limit of 32,767 variables. However, many people run Intercooled Stata, so we are taking a conservative approach in this example. (Type **help limits** to see how many variables your version of Stata can handle.)

Because of this variable limit, we broke the women's file into 2 pieces. Each file has exactly the same women (8,781 observations), but each file has a different set of variables (about 1500 variables in each). When we split the original file in two, we were careful to keep the 3 identifying variables in each file for later merging.

For this example, we selected age, education, and number of children from the first file, and marital status, age at first marriage, and desired family size from the second file. We need to merge these two sample files into a single file for analysis. The identifying variables for this merge are cluster (v001), household within cluster (v002), and line number within household (v003). The order in which the identifying variables are nested determines the sort order.

Prior to Stata 11, the data had to be sorted by the identifying variables before merging. Now you have a choice of syntax for merging. We have shown you both methods below, the old which requires sorting, and the new which does not (starting with version 11 Stata will sort the data for you when you use the new syntax). Both the old and new syntax work in Stata 11, but the new syntax is better for several reasons explained below.

Regardless of which version you are using, it is always a good idea to check your identifiers for duplicates before a merge. These are observations that have the same identifying variables, in this case v001 v002 v003. We expect each combination of these three variables to uniquely identify a woman, but as we see below, two duplicates crept into our data somehow.

Before trying this example, create a folder somewhere on your computer to store intermediate data files. These are files that you need for only for a short time while creating an analysis file, and then you can erase them. At CPC I put them into a folder I named **d:\statatemp**.

Then, copy the following commands into a do-file and submit them.

Check for, examine, and drop duplicates

```
/* Check for duplicates in exampw1 */

clear
use "q:\utilities\statatut\exampw1.dta"
duplicates report v001 v002 v003

/* List the duplicates */

duplicates tag v001 v002 v003, gen(duptag)
list if duptag > 0
```

```
/* Drop the duplicates */

duplicates drop v001 v002 v003, force
drop duptag
save "d:\statatemp\tempw1.dta"

/* Drop duplicates from exampw2 */

clear
use "q:\utilities\statatut\exampw2.dta"
duplicates drop v001 v002 v003, force
save "d:\statatemp\tempw2.dta"
```

Merge using New Syntax (Stata 11 and later)

```
clear
use "d:\statatemp\tempw1.dta"
merge 1:1 v001 v002 v003 using "d:\statatemp\tempw2.dta"

/* Look at the Results Window - should see only matches */

/* Clean up */

drop _merge
```

Merge using Old Syntax (Stata 10 and earlier)

```
/* Sort the first file and save it temporarily */

clear
use "d:\statatemp\tempw1.dta"
sort v001 v002 v003
save "d:\statatemp\tempw1.dta", replace

/* Sort the second file and save it temporarily */

clear
use "d:\statatemp\tempw2.dta"
sort v001 v002 v003
save "d:\statatemp\tempw2.dta", replace

/* Use the first file, merge on the second file */

clear
use "d:\statatemp\tempw1.dta"
merge v001 v002 v003 using "d:\statatemp\tempw2.dta"

/* Check how well the merge went - should see only "3" for _merge */

tab _merge
su

/* Clean up */

erase "d:\statatemp\tempw1.dta"
erase "d:\statatemp\tempw2.dta"
drop _merge
```

Questions:

1. What does the **duplicates** command do? Answer.
 2. The file **tempw1.dta** has 8,779 observations (after dropping two duplicates) and 6 variables. The file tempw2.dta also has 8,779 observations and 6 variables. After merging them, how many observations and variables do you expect the merged file to contain? Answer.
 3. There is a 10th variable on the merged file, named **_merge**. What is its role? Answer.
 4. What does the command **erase** do? Answer.
 5. Why **drop _merge**? Answer.
 6. Is there an easier way to work with intermediate data files? Answer.
-

Answers:

1. The **duplicates** command checks whether the variables you list identify observations in your data uniquely. If more than one observation has the same combination of identifying variables, **duplicates report** will tell you. You can then tag the duplicates and examine them using **duplicates tag**.

To examine the duplicates, you can list them as we show above. Or you might use the Data Editor in browse mode (**browse**), so you don't accidentally change the data:

```
browse if duptag > 0
```

At this point, if you have access to the questionnaires or other information about the raw data, you may be able to resolve the duplicates problem. In this case, though, we do not have access to the questionnaires, so we used **duplicates drop** to arbitrarily drop all but the first instance of each duplicate. (The command drops all but the first instance, given the current sort order of the data.)

[Back to question](#)

2. It is a good idea to anticipate the results of a merge, so that when you see the actual results, you will know immediately whether the merge worked the way you intended it.

In this case, a one-to-one merge, we expect the resulting file to have the same number of observations as the two original files: 8,779. The two files have 3 ID variables in common, so the merged file will have 3 ID variables. The two files have 3 additional variables each, so the

merged file will have 6 variables in addition to the IDs, for a total of **9 variables**.

[Back to question](#)

3. Stata creates the variable **_merge** when you ask it to merge two files. The variable can have 3 different values as follows:

- 1= this observation comes from the "master" file only (the file in memory)
- 2= this observation comes from the "using" file only (the file on disk)
- 3= this observation contains variables from both the master and using files (a match)

In this case, all observations have variables from both the master and using files, so they all have a value of 3 for **_merge**.

[Back to question](#)

4. After sorting the master and using files, we saved each one on disk with the names tempw1.dta and tempw2.dta. When we no longer needed them, we deleted them from disk with the **erase** command.

[Back to question](#)

5. At some time in the future, we may want to merge another file onto this new file. During the merge, Stata will again want to create the variable **_merge**. Since the variable already exists, Stata will stop the merge command and write an error message to the log.

The easiest way to handle this is to always **drop _merge** after looking at it. If you want to keep it around, you can rename it, or you can ask Stata to create it with a different name like "merge1" (using the **generate** option on the merge command).

[Back to question](#)

6. Intermediate data files can be handled in two ways. One way is shown here, where we create a permanent file and then erase it when we no longer need it. The second way is to use Stata temporary files, which Stata automatically erases at the end of your interactive Stata session. This second method is described in temporary files in the Miscellaneous Tips and Tricks section of this tutorial.

It is up to you to decide which way is easier!

[Back to question.](#)

Review again?

Another topic?

Precision and data storage

Two precision issues come up repeatedly when using Stata (and other similar analysis packages). One is how decimal values are represented in the computer's memory. The other is how large an integer you can store in a given Stata data type.

When 0.7 doesn't equal 0.7

Computers use a binary (0's and 1's) system to store decimal numbers. This leads to some inaccuracy, since some decimal values can't be stored exactly in binary. Try this:

```
clear all
set obs 1
gen x= 0.7
list
list if x == 0.7 // 0.7 doesn't equal 0.7
browse
list if x == float(0.7) // now they are equal
```

You'll notice that the command **list if x == 0.7** results in nothing being listed! When you browse the data, you'll see that 0.7 is being stored as the value 0.69999999. Since that value isn't 0.7, your command to list x results in no matches.

The **float** function takes care of this problem - it rounds the value 0.69999999 to 0.7. Many decimal values are stored accurately in binary, for example 0.5, but many are not. Rather than trying to memorize which are and which are not, we suggest always using the float function.

Big integers

The other precision issue has to do with Stata's data types. Stata offers 3 data types for integers (byte, int, and long) and 2 for floating point (float and double). For character data, Stata offers data type string. If you type **help data_types** you'll see a table that lists the 5 numeric data types that Stata uses, and the string data type, along with the minimum and maximum values that can be stored in each data type, and the maximum value that can be stored precisely.

Why data types? When Stata was first being developed, computers had very little random-access memory, and RAM was expensive. So, there was a benefit to storing values in as little memory as possible. While it would be convenient to create all variables with the highest precision available, which for numeric data is type double, this would waste a lot of memory. For example, a typical yes-no (1,2) variable can be stored accurately in a single byte, so storing it in type double would waste 7 bytes per observation. In a data file of survey results with thousands of variables and thousands of observations, this adds up many megabytes of wasted storage.

Now that computer memory is less expensive, we tend to pay less attention to it. But we need to pay attention when creating values in Stata that are relatively large. Typically, this occurs when the researcher decides to create a single numeric identifier out of multiple, nested identifying variables. In the Demographic and Health Survey data, one can often use three variables (the sampling cluster, the household identifier within the cluster, and the person identifier within the household) jointly to identify an individual respondent. For example:

```
duplicates report v001 v002 v003
```

usually demonstrates that these 3 nested variables create a unique identifier (but not always - be sure to check). But, if you try to combine them into a single numeric variable, you may run into trouble:

```
gen id= (v001*1000000) + (v002*1000) + v003
```

This example creates an 8- or 9-digit number, depending on the values of the cluster identifier v001. But Stata will store id **by default in type float**. Data type float begins to lose precision above 7 digits. We might be lucky and create a unique identifier, but probably not.

There are two ways to get around this. First, specify **double** when generating large integers. Data type double can accurately represent integers up to 15 digits:

```
gen double id= (v001*1000000) + (v002*1000) + v003
```

While double is discussed in the data type help page in terms of floating point precision, it works well for integers too, and it's the only way to store big integers precisely.

The other way to do this is using data type string for the identifier. In fact, the DHS data includes a string identifier, called caseid for the individual respondent. Strings are a bit of a pain to work with, but they precisely hold integers up to 244 characters in length.

We recommend that you always check for duplicates when creating a composite identifier. And always use **data type double** for these big integers. Some people even recommend that you always use data type double, which you can do with:

```
set type double, permanently
```

This asks Stata to create all new variables with data type double, greatly reducing your need to worry about precision. When you're finished creating an analysis file, you can **compress** the file. This command asks Stata to decide how each variable can be stored most efficiently.

The Stata Corp. web site offers many FAQ's on topics like this. Here's one on the precision of floating point storage that you might find useful: The accuracy of the float data type

Note: Dan Blanchette contributed this web page, however please direct questions to Phil Bardsley as noted below.

Review again?

Another topic?

Reshaping a data file

Converting variables to observations, observations to variables.

Most population surveys gather information about household members in a household roster. The data are organized with one observation being a household. Information about each member of a household is in different variables on the same observation. For example, the age of the first household member might be in the variable age01, the second in variable age02, and so forth.

We may need to summarize the data about household members. Or we may need to merge the household roster information with other information collected from selected members on different forms. In either case, it would be more convenient if the household roster variables were organized differently, with each observation being a member (instead of a household).

Stata offers the **reshape** command to make this transformation easier. Stata calls the original data format "wide", where household members are represented as variables on a household observation. They call the new format "long", where each household member is a separate observation. So, the command to transform the data from wide to long is **reshape long**, and from long to wide is **reshape wide**.

The example below uses a household roster with up to 10 members. **Browse** the data before and after each reshape to help visualize the transformation.

```
clear
use "q:\utilities\statatut\hhwide.dta"
summarize

/* Reshape from wide form to Long form (each member becomes an observation). */

reshape long age edu rel sex, i(hhid) j(lineno)
summarize

/* Reshape from Long form to wide form (each member becomes a set of 4 variables). */

reshape wide age edu rel sex, i(hhid) j(lineno)
summarize
```

Questions:

1. What do the terms "age", "edu", "rel", and "sex" stand for in the reshape command? Answer.
2. What is the term "i(hhid)"? Answer.
3. What is the term "j(lineno)"? Answer.
4. Why are there 5000 values for lineno in the long file, but only 2564 values for age and the

other variables? Answer.

5. What happens to the variable "lineno" when we convert from long to wide? Answer.

6. The variable "electric" is a household-level variable. We want it to be added unchanged to each observation in the long form. Is that what happens? Answer.

Answers:

1. These are the prefixes for the variable names in the original (wide) form of the data. Reshape works most easily if the variable names have a prefix followed by a number. In this case we have the variable name prefix "age" followed by a number 1-10 corresponding to the household member's order in the roster. If your variable names don't have this prefix-number form, you may want to **rename** them before doing reshape. See the explanation of **foreach** elsewhere in this tutorial for an easy way to rename multiple variables. Alternatively, you may want to explore the advanced features of reshape that allow you work with other naming conventions without renaming the variables. See the Stata manual for details.

[Back to question](#)

2. The reshape command needs a common identifier for each member of a household.

Reshape long assigns this identifier to each household member, while **reshape wide** uses this identifier to assign each member to a household. In this case, that identifier is named hhid in the original data. We would need this identifier in order to collapse the data to the household level or to merge data from other files onto the new long file.

[Back to question](#)

3. The **reshape** command needs an identifier for each individual in the household. Typically, this identifier is the line number in the household roster. The "j lineno)" term tells **reshape long** to create a new variable named lineno. This variable captures the position of the household member in each household observation in the original (wide) file. The **reshape wide** command uses this variable to assign a suffix to the new household-level age, education, relationship, and sex variables it is creating. In **reshape long**, we can choose any name we want for the "j" variable, so we chose lineno, while in reshape wide we must use the variable that actually identifies each individual in the household. We would need this line number, together with the hhid variable, to merge individual-level data with the new long file.

[Back to question](#)

4. **Reshape** creates one new observation for each set of 4 variables in the wide file. We have 10 sets of 4 variables, so we get 10 observations. The wide file has 500 households. Multiplied by 10 members, that's 5,000 new observations in the long file. However, not every household in the wide file has 10 members - some have fewer. Nevertheless, **reshape** creates a new observation for each set of variables, missing or not. The **summarize** command tells us that these 500 households have 2,564 members distributed among them, with anywhere from 1-10 members in each household. We probably would **drop** any observation with all missing data before doing anything further with the long file. The command would be:

```
drop if missing(age)
```

Note: you should not use **edu** to drop observations with missing values, because 8 people are missing education but have values for all the other variables. They shouldn't be dropped.

[Back to question.](#)

5. The variable **lineno** disappears going from long to wide. Its values become the suffix in the variable names **age**, **edu**, **rel**, and **sex**.

[Back to question](#)

6. Yes, the variable **electric** is automatically added to each new observation generated by the **reshape long** command. We do not name it in the **reshape** command, and Stata carries it along unchanged.

[Back to question](#)

Review again?

Another topic?

Shrinking large data files

In an earlier example, we mentioned the **compress** command as a way to reduce the size of Stata data files. It works by choosing the most efficient data type that is necessary to store each variable and still maintain the precision of your data.

When working with very large files, or on computers with limited RAM, you need to compress the data. Below is an example you can try to see how compress works.

```
clear  
use "q:\utilities\statatut\excomp.dta"  
set more off  
compress
```

Questions:

1. All these variables started as data type "float" or as short strings. How much is the memory need reduced when compress changes floats to bytes? Answer.
2. What does **set more off** do? Answer.
3. How can I tell whether compress will make much difference in my RAM requirements? Answer.
4. This is a "Catch-22"! I can't fit my Stata data into memory on my computer, but I can't compress it unless I can get it into Stata. Answer.
5. I have a Windows compression program on my PC. Can I use that instead? Answer.

Answers:

1. A float variable requires 4 bytes, while a byte variable requires 1 byte. See **help data types** in Stata for details on the amount of storage required by each data type.

[Back to question](#)

2. You'll recall from an earlier example that **--more--** shows on the bottom of the Stata results window when a command generates more lines than can fit in that window. When you see **--more--** you need to press the space bar to continue viewing results.

You can tell Stata to continue scrolling the results and not stop when the screen fills up. The command is **set more off**. To turn more back on, use **set more on**. That way you can go get a cup of coffee while Stata compresses your giant survey file.

[Back to question](#)

3. You can look at the current data type of the variables using the **describe** command. If you see a lot of doubles (8 bytes) or floats (4 bytes) and you know your data are mostly 1 or 2 digit values, you'll know that **compress** will make a great deal of difference. If your data are already stored in bytes, **compress** won't help much.

[Back to question](#)

4. There are two ways to get around this problem. You can find another computer with more RAM than is available on yours, and compress the file there. Or, you can use the **varlist** option on the **use** command to bring a subset of variables into memory. This will allow you to split the file into two or more pieces, each of which is small enough to fit in RAM and compress. Then you can recombine them, if necessary, into a single file. An example of this was shown in the discussion of One-to-one merging

. Remember to include the necessary identification variables in the smaller files you create so that you can merge if you have to.

[Back to question](#)

5. Compression programs like WinZip and PKZIP use a different compression method. Stata cannot read data files that have been compressed by those programs, or by Unix commands such as **compress**, **gzip**, or **tar**.

[Back to question](#)

[Review again?](#)

[Another topic?](#)

Temporary files

Saving and using temporary files on disk

In previous examples we often created a file on disk that we needed only for the duration of the example. We erased the file at the end of the example to clean up after ourselves.

Stata supplies the **tempfile** command to get around creating and erasing permanent files. However, temporary files are a bit tricky to work with, and I don't recommend using them in everyday work. They're best suited to programs, that is sets of Stata commands that are executed together. Most researchers work with a few commands at a time in a do-file, which makes temporary files cumbersome. For completeness, however, here is an example of the tempfile command:

```
clear
use "t:\statatut\exampw1.dta"
sort v001 v002 v003
tempfile temp1 /* create a temporary file */
save ``temp1'' /* save memory into the temporary file */
use "t:\statatut\exampw2.dta"
sort v001 v002 v003
merge v001 v002 v003 using ``temp1'' /* use the temporary file */
```

Questions:

1. The name temp1 in the save and merge commands is enclosed in single quotes. Why is that? Answer.
2. Why don't we need to erase temp1 after we're finished with it? Answer.

Answers:

1. The name "temp1" is actually a temporary variable name, which is called a "macro" in Stata. In this case it refers to a temporary file. Looking in the log you'll see that the actual temporary file is named something like:

```
C:\TEMP\ST_040001.tmp
```

The macro "temp1" contains that value. To get that value, we need to put the accent mark (`) in

front of the macro name and a single quote after the macro name.

Note that you must use the "backward" single quote (accent mark on the tilda key) at the beginning of your macro name, and the "forward" single quote (apostrophe on the double-quotes key) at the end of your macro name when you refer to the file.

[Back to question](#)

2. We don't need to erase temp1 because Stata takes care of that for us.

[Back to question](#)

[Review again?](#)

[Another topic?](#)

The by command in detail

More examples of the by command.

This page is intended to show how to use the **by** command to create variables and access specific observations in sub-groups. Sub-group processing breaks down the dataset into multiple mini-datasets. The way to think about it is as if each sub-group is the entire dataset.

The way to access data from different observations is with internal Stata variables **_n** and **_N**. The variable **_n** is equal to the numeric position of an observation and **_N** is equal to the total number of observations. For example, in a dataset of 10 observations, in the first observations **_n** is equal to 1 and **_N** is equal to 10. In the second observation **_n** is equal to 2 and **_N** is equal to 10, etc.

```
v001 _n _N
-----
1002 1 10
1002 2 10
1002 3 10
1002 4 10
1002 5 10
1003 6 10
1003 7 10
1003 8 10
1004 9 10
1004 10 10
```

When using the **by** command, **_n** is equal to 1 for the first obs of the by-group and **_N** is equal to the total number of observations for the by-group. So if a person has 5 observations in the data set and **v001** is the id variable, "**by v001:**" creates by-groups where the first observation in the by-group **_n=1** and **_N=5**. The second observation **_n=2** and **_N=5**, etc. The last observation for that person **_n** will equal **_N**.

Note: if conditions in the command being run by the **by** command do not subset the by group so **_n** and **_N** are not affected by the **if** condition.

```
by v001:
```

makes the data look like this:

```
v001 _n _N
-----
1002 1 5
1002 2 5
1002 3 5
1002 4 5
1002 5 5
1003 1 3
1003 2 3
1003 3 3
1004 1 2
```

```
1004 2 2
```

If you want to keep the first observation per person do:

```
by v001: keep if _n == 1
```

second observation is:

```
by v001: keep if _n == 2
```

second to the last observation is:

```
by v001: keep if _n == _N - 1
```

last observation is:

```
by v001: keep if _n == _N
```

The following is Stata code that plays with this concept.

NOTE: the following is a **do-file**. Highlight and copy all text between the horizontal bars and paste into the Stata interactive do-file editor window so that you can submit this do-file.

```
use "q:\utilities\statatut\examfac2.dta",clear

keep facid factype q102_1-q103_20
** q103_1-q103_20 are # hours/week usually worked **

** facid is the id var **

** q102_1-q102_20 are titles of facility workers **
label list title type

/** The following code figures out how many total hours per week
 * different types of workers at a health facility work. ***/
sort facid

/* Get rid of duplicate ids. */
drop if facid == facid[_n-1]

/* This would also keep the first obs in a duplicate id situation:
 * by facid: keep if facid[_n] == 1 */

/* Make the data set multiple obs per facility. */
reshape long q102_ q103_, i(facid) j(position)

/** Save the data set as a temporary data set in case you want to try another idea. ***/
preserve
gen ttl_h1= 0
replace ttl_h1=q103_ if q103_ != 99 & q102_ == 1

by facid: gen ttl_th1=ttl_h1[1]+ttl_h1[2]+ttl_h1[3]+ttl_h1[4]+ttl_h1[5]+ ///
    ttl_h1[6]+ttl_h1[7]+ttl_h1[8]+ttl_h1[9]+ttl_h1[10]+ ///
    ttl_h1[11]+ttl_h1[12]+ttl_h1[13]+ttl_h1[14]+ttl_h1[15]+ ///
    ttl_h1[16]+ttl_h1[17]+ttl_h1[18]+ttl_h1[19]+ttl_h1[20]

list facid factype q102_ q103_ ttl_h1 ttl_th1 in 1/500 if factype==3

/* or you could use the -forvalues- command */
replace ttl_th1= 0
```

```
forvalues num= 1/20 {  
    by facid: replace ttl_th1= ttl_h1[`num'] + ttl_th1  
}  
  
/* If you wanted totals for all the other titles you'd have  
 * to repeat the above 19 more times. */  
  
/* You could use the -egen- command to sum up work hours per title for each facility  
 * and for commands to reduce the amount of typing. */  
  
clear  
restore  
preserve  
  
foreach var of newlist ttl_h1-ttl_h10 {  
    gen `var'= .  
}  
  
forvalues num = 1/10 {  
    replace ttl_h`num'= q103_ if q103_ != 99 & q102_ == `num'  
    egen ttl_th`num'= sum(ttl_h`num'), by(facid)  
}  
  
// examples of when factype == 3 in the first 500 obs  
list facid factype q102_ q103_ ttl_h1 ttl_th1 in 1/500 if factype == 3  
  
/** Keep only the variables that represent all obs of the facility. **/  
keep facid factype ttl_th1-ttl_th10  
  
/** Keep only the last obs per facid, though any of it's obs would be just as good. **/  
by facid: keep if _n == _N  
  
list in 1/500 if factype == 3
```

This page was contributed by Dan Blanchette, however please direct questions to Phil Bardsley as noted below.

Review again?

Another topic?

The parmest command

The **parmest** command follows any Stata command that produces parameter estimates, and saves the results in a Stata dataset, the Results window, and/or the log file. It saves the results as one observation per parameter in a Stata dataset. You can format the results for printing and publication, e.g. rounding estimates to 2 decimal places.

Parmest does not come with the Stata software but is available through the SSC archive. If you are not on the CPC network, you can download the latest parmest package by typing at a Stata command line:

```
ssc install parmest
```

You need to have an active connection to the internet to do this. If you have a copy of parmest but are unsure how recently it was downloaded, install it again using the "replace" option.

Here's an example showing how to save the results of a regress command:

```
use "t:\statatut\examfac2.dta"
reg age factype authorit urbrur femster
parmest, saving("c:\tables\reg1.dta", replace)
```

You now have a copy of the results saved in a dataset. To list the results in the Results Window in a format that allows you to copy them to Excel:

```
use "c:\tables\reg1.dta", clear
list parm estimate t p, noobs clean

      parm    estimate        t        p
factype   -3.5400133    -4.16595   .00004732
authorit   -.56049792   -1.1524406   .25061116
       urbrur   -2.2121531   -1.4140991   .15899552
       femster   -3.2143824   -.98375474   .32650721
       _cons     51.359101    8.4572685  7.705e-15
```

You can now highlight the table of results, right click, copy table, and paste into Excel.

Type **help parmest** for more information. See Exporting Stata Results to MS Office for another way to put results into a Stata dataset.

[Review Again?](#)
[Another topic?](#)

Updating Stata

How to update Stata

Stata Corp. makes updates to your current version available for free on their website. If you have your own Stata license, such as on a standalone PC or laptop, it's a good idea to make a habit of checking once a month for updates. (At CPC, we keep the network copies up-to-date for you.)

To check whether your copy is up-to-date, first make sure that your computer is connected to the Internet, then click on Check for Updates in the Help menu, or type the command:

```
update query
```

This instructs Stata to first check when you last updated it, then connect to www.stata.com and check whether a newer version is available. If updates are available for you, it will instruct you how to get the updates.

You can also ask Stata to check for updates automatically. Go to the Edit menu and click on Preferences, General Preferences, Internet, and check the box for Enable automatic update checking.

If you haven't updated your copy for some time, or if your connection speed is slow, your connection may time out before all the updates have been downloaded. The default connection period is 5 minutes (300 seconds). If you need to increase this, type:

```
set timeout2 sss
```

where "sss" is the number of seconds. The maximum is 32,000.

If you don't have Internet connectivity for your computer, you can download the updates at CPC, copy them to a flash drive or other portable storage, and install them on your computer at a later date. See Keeping Stata up to date for complete instructions.

Review again?

Another topic?

Using permanent Stata data files

Clearing memory, using a Stata file, saving changes to the file.

In Example1 you created a permanent Stata data file called myfile.dta. This example uses that file. Type each of these commands and observe the results:

```
clear  
use myfile  
drop d  
save myfile  
save myfile,replace
```

Questions:

1. The **clear** command removes data from Stata's memory. Why do we need to use this command before copying a new data file into Stata's memory? Answer.
2. The **use** command copies a Stata data file into Stata's memory. How would you change the **use** command if the Stata data file was not in the present working directory? Assume it was in the following path:

```
e:\student\jdoe\myproject\data\myfile.dta
```

Answer.

3. What does the **drop** command do? Answer.
4. Why do you get the error message "file myfile.dta already exists" when you type the **save** command? Answer.

Answers:

1. Stata can only keep one set of data in memory at a time. In order to protect you from accidentally writing over and destroying your data in memory, Stata requires that you clear memory before either of these tasks.

[Back to question.](#)

2. Change the use command to include the full path:

```
use "e:\student\jdoe\myproject\data\myfile.dta"
```

By the way, if you make a mistake typing a Stata command, especially a long one like this, you don't have to retype the whole command. You can press the **Page Up** key to display your last command, edit it, and press Enter to resubmit it. Or, you can click on the command in the Review window to display it for editing.

[Back to question.](#)

3. The drop command drops one or more variables. We cover it in detail in the next example. It's included here only to make this example a little less artificial.

[Back to question.](#)

4. Stata is warning you that you are about to write over an existing file on disk. It's giving you a chance to think about whether that's what you really should be doing. The replace option reassures Stata that you know what you're doing:

```
save myfile,replace
```

[Back to question.](#)

[Review again?](#)

[Another topic?](#)

Value labels

Value labels: words in place of numbers.

Value labels give more description than the integer values of a variable and can make it more user-friendly.

Stata does *not* allow value labels to be associated to a range of values (eg. 1-10 "low" 11-20 "med" 21-30 "high"). It is possible to get the same results by defining the same value label one-by-one to all integer values in the range. The foreach command may come to your rescue if the range is long. In this case, though, it would make more sense to create a recoded variable with 3 values (1="low", 2="med", 3="high").

Stata does *not* allow value labels to be associated with non-integer data (e.g., 1.5 "low" 2.1 "med" 3.3 "high") or with character strings (e.g., "Yes" "No").

Examples of the **label** command are shown below in a **do-file**.

```
clear
use "q:\utilities\statatut\examfac2.dta"
keep facid q102_1-q102_5

* Confirm that the variables are numeric (in this case "byte").
* Value labels named "title" are associated with the variables.

describe

* Many commands, like browse, show the labels rather than the values.

browse

* To see the values, use the "no label" option.

browse, nolabel

* List all value labels stored in examfac2.dta.

label dir

* Show what the value label title is.

label list title

* These numbers are the actual data that STATA stores for each variable.
* Notice that there is no value label for responses of 99. Stata prints
*   the actual data for un-formatted responses.

* Remove the value label title.

label drop title
browse

* Re-create the title value label.

#delimit ;
label define title
    1 "Doctor"
    2 "AssistMedOff"
    3 "ClinOff"
```

```

        4 "AssistClinOff"
        5 "NurseOff"
        6 "Nurse/Midwife";

* Add more to the title value label.;

label define title
    7 "PH NurseB"
    8 "MCH Aide"
    9 "NurseAssist"
    10 "Other"
    11 "Dr. Doolittle", add;
#delimit cr

* Change the title value label.

label define title 11 "", modify

* Re-assign the title value label to each q102_* variable.

foreach v of varlist q102* {
    label val `v' title
}
browse

```

Questions:

1. How many characters long can a value label be? Answer.
 2. Can I see both the values and the value labels at the same time? Answer.
 3. Can value labels have the same name as the variable? Answer.
-

Answers:

1. The maximum length for a value label is 32,000 characters. See **help limits** for this and almost all other limits in Stata.

[Back to question](#)

2. The **label list** command in the above example gives a kind of codebook for the values and labels. Otherwise, most commands give you either the value or the label but not both. One exception is the **fre** command, written by Ben Jann and available on CPC computers or by download from the SSC archive: **ssc install fre** It works like **tab** but shows both values and value labels in the tabulation.

[Back to question](#)

3. Yes. For example, you could have a value label **gender** assigned to the variable **gender**

[Back to question](#)

[Review again?](#)

[Another topic?](#)

Variable labels

Variable labels: describe your variables.

Variable labels document your data and make it more user-friendly.

Examples of the **label variable** command are shown below in a **do-file**.

```
use "q:\utilities\statatut\examfac2.dta", clear
keep facid q102_1-q102_5
/** show variable labels */
describe
/** remove variable labels */
label variable q102_1 ""
label variable q102_2 ""
label variable q102_3 ""
label variable q102_4 ""
label variable q102_5 ""
describe
/** re-create variable labels */
label var q102_1 "Staff member 1 job title"
label var q102_2 "Staff member 2 job title"
label var q102_3 "Staff member 3 job title"
label var q102_4 "Staff member 4 job title"
label var q102_5 "Staff member 5 job title"
describe
```

Questions:

1. How many characters long can a variable label be? Answer.
2. Can more than one variable be assigned a variable label in one label command? Answer.

Answers:

1. The maximum length for a variable label is 80 characters. See **help limits** for this and almost every other limit in Stata.

[Back to question](#)

2. No. Only one variable can be processed in a label variable command. However, you can use a command like **foreach** or **forvalues** to create labels for many similar variables in a single loop, which saves a lot of typing.

[Back to question](#)

[Review again?](#)

[Another topic?](#)

A simple example

Inputting data, generating a new variable, listing and saving data, log files.

Start Stata by double-clicking on the desktop icon. Type each of these commands in the Stata Command window, one at a time, pressing Enter after each line:

```
input a b c  
1 5 10  
0 8 7  
1 4 6  
end  
generate d=c-b  
list  
browse  
save myfile
```

Questions:

1. The **input** command puts data into Stata's memory. What is the result of the input command? In other words, where are those numbers stored? Answer.
2. What does the **generate** command do? Answer.
3. The **list** and **browse** commands list the contents of Stata's memory. Where are the results of all Stata commands displayed? Answer.
4. The **save** command writes a permanent disk file. Where is the file "myfile" created? Answer.
5. What does the file "myfile" contain? Answer.
6. What are the three most important default actions of a simple Stata program like the one above? Answer.
7. What would you change in the program to make it read one thousand records placed after the input command instead of three? Answer.
8. What would you change in the program if your data records were in a file separate from the program? Assume your data file has the following path:

```
e:\student\jdoe\myproject\data\part1.dat
```

Answer.

9. Describe the result if you typed the generate command before the input command:

```
generate d=c-b  
input a b c  
1 5 10  
0 8 7  
1 4 6  
end
```

Answer.

10. Describe the result if you typed the variable name "C" in caps in the generate command.

```
input a b c  
1 5 10  
0 8 7  
1 4 6  
end  
generate d=C-b
```

Answer.

Answers:

1. The numbers are stored in the desktop PC memory (RAM). There is no file, temporary or permanent, created on disk. The data are stored conceptually as 3 rows (observations) and 3 columns (variables), like an Excel spreadsheet.

[Back to question.](#)

2. It creates a 4th column (variable) named "d" with one value for each observation.

[Back to question.](#)

3. Stata echos each command that you type and displays the results in the Stata Results window. An exception is **browse**, which opens a new window. The **log** command will create a permanent file on disk with the contents of the Stata Results window while you type commands.

```
input a b c  
1 5 10  
0 8 7  
1 4 6  
end  
generate d=c-b  
log using mylog.log  
list  
log close  
save myfile
```

Here the log command creates a file named mylog.log containing the results of the list command but none of the other contents of the Stata Results window.

[Back to question.](#)

4. The save and log commands create files in the present working directory (pwd). This directory is displayed in the lower left corner of the Stata window. You can change it with the cd command. You can also specify a full path name in the save and log commands. Stata automatically adds the suffix ".dta" to files created with the save command.

[Back to question.](#)

5. The save command creates a Stata-format file. It contains the contents of Stata's memory at the time you type the save command. In this case, the file would have 3 obs and 4 variables.

[Back to question.](#)

6. Default actions:

- all file-related commands (save, log) work in the present working directory
- all variable-related commands (generate) affect every observation in Stata's memory
- commands are executed in order from the top down

[Back to question.](#)

7. Nothing.

[Back to question.](#)

8. Replace the **input** command with the **infile** command:

```
infile a b c using "e:/student/jdoe/myproject/data/part1.dat"
```

Note the quotes around the path and file name. Quotes are only necessary if a directory has blank spaces in it, but it's a good habit to get into. Also, note that the "/" (slashes) can be either forward or backward - Stata doesn't distinguish between them.

[Back to question.](#)

9. The generate command results in an error message "c not found", since Stata doesn't yet know about the variable c.

[Back to question.](#)

10. The generate command results in an error message "C not found", since Stata distinguishes between upper and lower case.

[Back to question.](#)

[Review again?](#)

[Another topic?](#)

Degrees of freedom (statistics)

From Wikipedia, the free encyclopedia

In statistics, the number of **degrees of freedom** is the number of values in the final calculation of a statistic that are free to vary.^[1]

The number of independent ways by which a dynamic system can move, without violating any constraint imposed on it, is called *number of degrees of freedom*. In other words, the number of degree of freedom can be defined as the minimum number of independent coordinates that can specify the position of the system completely.

Estimates of statistical parameters can be based upon different amounts of information or data. The number of independent pieces of information that go into the estimate of a parameter is called the degrees of freedom. In general, the degrees of freedom of an estimate of a parameter is equal to the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself (i.e. the sample variance has $N-1$ degrees of freedom, since it is computed from N random scores minus the only 1 parameter estimated as intermediate step, which is the sample mean).^[2]

Mathematically, degrees of freedom is the number of dimensions of the domain of a random vector, or essentially the number of "free" components (how many components need to be known before the vector is fully determined).

The term is most often used in the context of linear models (linear regression, analysis of variance), where certain random vectors are constrained to lie in linear subspaces, and the number of degrees of freedom is the dimension of the subspace. The degrees of freedom are also commonly associated with the squared lengths (or "sum of squares" of the coordinates) of such vectors, and the parameters of chi-squared and other distributions that arise in associated statistical testing problems.

While introductory textbooks may introduce degrees of freedom as distribution parameters or through hypothesis testing, it is the underlying geometry that defines degrees of freedom, and is critical to a proper understanding of the concept. Walker (1940)^[3] has stated this succinctly as "the number of observations minus the number of necessary relations among these observations."

Contents

- 1 Notation
- 2 Residuals
 - 2.1 Linear regression
- 3 Degrees of freedom of a random vector
- 4 Degrees of freedom in linear models
- 5 Sum of squares and degrees of freedom
- 6 Degrees of freedom parameters in probability distributions
- 7 Effective degrees of freedom
 - 7.1 Regression effective degrees of freedom

- 7.2 Residual effective degrees of freedom
- 7.3 General
- 7.4 Other formulations
 - 7.4.1 Alternative
- 8 See also
- 9 References
- 10 Further reading
- 11 External links

Notation

In equations, the typical symbol for degrees of freedom is ν (lowercase Greek letter nu). In text and tables, the abbreviation "d.f." is commonly used. R.A. Fisher used n to symbolize degrees of freedom but modern usage typically reserves n for sample size.

Residuals

A common way to think of degrees of freedom is as the number of independent pieces of information available to estimate another piece of information. More concretely, the number of degrees of freedom is the number of independent observations in a sample of data that are available to estimate a parameter of the population from which that sample is drawn. For example, if we have two observations, when calculating the mean we have two independent observations; however, when calculating the variance, we have only one independent observation, since the two observations are equally distant from the mean.

In fitting statistical models to data, the vectors of residuals are constrained to lie in a space of smaller dimension than the number of components in the vector. That smaller dimension is the number of **degrees of freedom for error**.

Linear regression

Perhaps the simplest example is this. Suppose

$$X_1, \dots, X_n$$

are random variables each with expected value μ , and let

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

be the "sample mean." Then the quantities

$$X_i - \bar{X}_n$$

are residuals that may be considered estimates of the errors $X_i - \mu$. The sum of the residuals (unlike the sum of the errors) is necessarily 0. If one knows the values of any $n - 1$ of the residuals, one can thus find the last one. That means they are constrained to lie in a space of dimension $n - 1$. One says that "**there are $n - 1$ degrees of freedom for errors.**"

An only slightly less simple example is that of least squares estimation of a and b in the model

$$Y_i = a + bx_i + e_i \text{ for } i = 1, \dots, n$$

where x_i is given, but e_i and hence Y_i are random. Let \hat{a} and \hat{b} be the least-squares estimates of a and b . Then the residuals

$$e_i = y_i - (\hat{a} + \hat{b}x_i)$$

are constrained to lie within the space defined by the two equations

$$\begin{aligned} e_1 + \cdots + e_n &= 0, \\ x_1 e_1 + \cdots + x_n e_n &= 0. \end{aligned}$$

One says that **there are $n - 2$ degrees of freedom for error.**

Note about notation: the capital letter Y is used in specifying the model, while lower-case y in the definition of the residuals; that is because the former are hypothesized random variables and the latter are actual data.

We can generalise this to multiple regression involving p parameters and covariates (e.g. $p - 1$ predictors and one mean), in which case the cost in *degrees of freedom of the fit* is p .

Degrees of freedom of a random vector

Geometrically, the degrees of freedom can be interpreted as the dimension of certain vector subspaces. As a starting point, suppose that we have a sample of n independent normally distributed observations,

$$X_1, \dots, X_n.$$

This can be represented as an n -dimensional random vector:

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}.$$

Since this random vector can lie anywhere in n -dimensional space, it has n degrees of freedom.

Now, let \bar{X} be the sample mean. The random vector can be decomposed as the sum of the sample mean plus a vector of residuals:

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \bar{X} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{pmatrix}.$$

The first vector on the right-hand side is constrained to be a multiple of the vector of 1's, and the only free quantity is \bar{X} . It therefore has 1 degree of freedom.

The second vector is constrained by the relation $\sum_{i=1}^n (X_i - \bar{X}) = 0$. The first $n - 1$ components of this vector can be anything. However, once you know the first $n - 1$ components, the constraint tells you the value of the n th component. Therefore, this vector has $n - 1$ degrees of freedom.

Mathematically, the first vector is the orthogonal, or least-squares, projection of the data vector onto the subspace spanned by the vector of 1's. The 1 degree of freedom is the dimension of this subspace. The second residual vector is the least-squares projection onto the $(n - 1)$ -dimensional orthogonal complement of this subspace, and has $n - 1$ degrees of freedom.

In statistical testing applications, often one isn't directly interested in the component vectors, but rather in their squared lengths. In the example above, the residual sum-of-squares is

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \left\| \begin{pmatrix} X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{pmatrix} \right\|^2.$$

If the data points X_i are normally distributed with mean 0 and variance σ^2 , then the residual sum of squares has a scaled chi-squared distribution (scaled by the factor σ^2), with $n - 1$ degrees of freedom. The degrees-of-freedom, here a parameter of the distribution, can still be interpreted as the dimension of an underlying vector subspace.

Likewise, the one-sample t -test statistic,

$$\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)}}$$

follows a Student's t distribution with $n - 1$ degrees of freedom when the hypothesized mean μ_0 is correct. Again, the degrees-of-freedom arises from the residual vector in the denominator.

Degrees of freedom in linear models

The demonstration of the t and chi-squared distributions for one-sample problems above is the simplest example where degrees-of-freedom arise. However, similar geometry and vector decompositions underlie much of the theory of linear models, including linear regression and analysis of variance. An explicit example based on comparison of three means is presented here; the geometry of linear models is discussed in more complete detail by Christensen (2002).^[4]

Suppose independent observations are made for three populations, $X_1, \dots, X_n, Y_1, \dots, Y_n$ and Z_1, \dots, Z_n . The restriction to three groups and equal sample sizes simplifies notation, but the ideas are easily generalized.

The observations can be decomposed as

$$X_i = \bar{M} + (\bar{X} - \bar{M}) + (X_i - \bar{X})$$

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

$$Z_i = \bar{M} + (\bar{Z} - \bar{M}) + (Z_i - \bar{Z})$$

where $\bar{X}, \bar{Y}, \bar{Z}$ are the means of the individual samples, and $\bar{M} = (\bar{X} + \bar{Y} + \bar{Z})/3$ is the mean of all $3n$ observations. In vector notation this decomposition can be written as

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \\ Y_1 \\ \vdots \\ Y_n \\ Z_1 \\ \vdots \\ Z_n \end{pmatrix} = \bar{M} \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} \bar{X} - \bar{M} \\ \vdots \\ \bar{X} - \bar{M} \\ \bar{Y} - \bar{M} \\ \vdots \\ \bar{Y} - \bar{M} \\ \bar{Z} - \bar{M} \\ \vdots \\ \bar{Z} - \bar{M} \end{pmatrix} + \begin{pmatrix} X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \\ Y_1 - \bar{Y} \\ \vdots \\ Y_n - \bar{Y} \\ Z_1 - \bar{Z} \\ \vdots \\ Z_n - \bar{Z} \end{pmatrix}.$$

The observation vector, on the left-hand side, has $3n$ degrees of freedom. On the right-hand side, the first vector has one degree of freedom (or dimension) for the overall mean. The second vector depends on three random variables, $\bar{X} - \bar{M}, \bar{Y} - \bar{M}$ and $\bar{Z} - \bar{M}$. However, these must sum to 0 and so are constrained; the vector therefore must lie in a 2-dimensional subspace, and has 2 degrees of freedom. The remaining $3n - 3$ degrees of freedom are in the residual vector (made up of $n - 1$ degrees of freedom within each of the populations).

Sum of squares and degrees of freedom

In statistical testing problems, one usually isn't interested in the component vectors themselves, but rather in their squared lengths, or Sum of Squares. The degrees of freedom associated with a sum-of-squares is the degrees-of-freedom of the corresponding component vectors.

The three-population example above is an example of one-way Analysis of Variance. The model, or treatment, sum-of-squares is the squared length of the second vector,

$$SSTr = n(\bar{X} - \bar{M})^2 + n(\bar{Y} - \bar{M})^2 + n(\bar{Z} - \bar{M})^2$$

with 2 degrees of freedom. The residual, or error, sum-of-squares is

$$SSE = \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2 + \sum_{i=1}^n (Z_i - \bar{Z})^2$$

with $3(n-1)$ degrees of freedom. Of course, introductory books on ANOVA usually state formulae without showing the vectors, but it is this underlying geometry that gives rise to SS formulae, and shows how to unambiguously determine the degrees of freedom in any given situation.

Under the null hypothesis of no difference between population means (and assuming that standard ANOVA regularity assumptions are satisfied) the sums of squares have scaled chi-squared distributions, with the corresponding degrees of freedom. The F-test statistic is the ratio, after scaling by the degrees of freedom. If there is no difference between population means this ratio follows an F distribution with 2 and $3n - 3$ degrees of freedom.

In some complicated settings, such as unbalanced split-plot designs, the sums-of-squares no longer have scaled chi-squared distributions. Comparison of sum-of-squares with degrees-of-freedom is no longer meaningful, and software may report certain fractional 'degrees of freedom' in these cases. Such numbers have no genuine degrees-of-freedom interpretation, but are simply providing an *approximate* chi-squared distribution for the corresponding sum-of-squares. The details of such approximations are beyond the scope of this page.

Degrees of freedom parameters in probability distributions

Several commonly encountered statistical distributions (Student's t, Chi-Squared, F) have parameters that are commonly referred to as *degrees of freedom*. This terminology simply reflects that in many applications where these distributions occur, the parameter corresponds to the degrees of freedom of an underlying random vector, as in the preceding ANOVA example. Another simple example is: if $X_i; i = 1, \dots, n$ are independent normal (μ, σ^2) random variables, the statistic

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}$$

follows a chi-squared distribution with $n-1$ degrees of freedom. Here, the degrees of freedom arises from the residual sum-of-squares in the numerator, and in turn the $n-1$ degrees of freedom of the underlying residual vector $\{X_i - \bar{X}\}$.

In the application of these distributions to linear models, the degrees of freedom parameters can take only integer values. The underlying families of distributions allow fractional values for the degrees-of-freedom parameters, which can arise in more sophisticated uses. One set of examples is problems where chi-squared approximations based on effective degrees of freedom are used. In other applications, such as modelling heavy-tailed data, a t or F distribution may be used as an empirical model. In these cases, there is no particular *degrees of freedom* interpretation to the distribution parameters, even though the terminology may continue to be used.

Effective degrees of freedom

Many regression methods, including ridge regression, linear smoothers and smoothing splines are not based on ordinary least squares projections, but rather on regularized (generalized and/or penalized) least-squares, and so degrees of freedom defined in terms of dimensionality is generally not useful for these procedures. However, these procedures are still linear in the observations, and the fitted values of the regression can be expressed in the form

$$\hat{y} = Hy,$$

where \hat{y} is the vector of fitted values at each of the original covariate values from the fitted model, y is the original vector of responses, and H is the hat matrix or, more generally, smoother matrix.

For statistical inference, sums-of-squares can still be formed: the model sum-of-squares is $\|Hy\|^2$; the residual sum-of-squares is $\|y - Hy\|^2$. However, because H does not correspond to an ordinary least-squares fit (i.e. is not an orthogonal projection), these sums-of-squares no longer have (scaled, non-central) chi-squared distributions, and dimensionally defined degrees-of-freedom are not useful.

The *effective degrees of freedom* of the fit can be defined in various ways to implement goodness-of-fit tests, cross-validation and other inferential procedures. Here one can distinguish between *regression effective degrees of freedom* and *residual effective degrees of freedom*.

Regression effective degrees of freedom

Regarding the former, appropriate definitions can include the trace of the hat matrix,^[5] $\text{tr}(H)$, the trace of the quadratic form of the hat matrix, $\text{tr}(H'H)$, the form $\text{tr}(2H - HH')$, or the Satterthwaite approximation, $\text{tr}(H'H)^2/\text{tr}(H'HH'H)$. In the case of linear regression, the hat matrix H is $X(X'X)^{-1}X'$, and all these definitions reduce to the usual degrees of freedom. Notice that

$$\text{tr}(H) = \sum_i h_{ii} = \sum_i \frac{\partial \hat{y}_i}{\partial y_i},$$

the regression (not residual) degrees of freedom in linear models are "the sum of the sensitivities of the fitted values with respect to the observed response values",^[6] i.e. the sum of leverage scores.

Residual effective degrees of freedom

There are corresponding definitions of residual effective degrees-of-freedom (redf), with H replaced by $I - H$. For example, if the goal is to estimate error variance, the redf would be defined as $\text{tr}((I - H)'(I - H))$, and the unbiased estimate is (with $\hat{r} = y - Hy$),

$$\hat{\sigma}^2 = \frac{\|\hat{r}\|^2}{\text{tr}((I - H)'(I - H))},$$

or:^{[7][8][9]}

$$\hat{\sigma}^2 = \frac{\|\hat{r}\|^2}{n - \text{tr}(2H - HH')} = \frac{\|\hat{r}\|^2}{n - 2\text{tr}(H) + \text{tr}(HH')}$$

$$\hat{\sigma}^2 \approx \frac{\|\hat{r}\|^2}{n - 1.25\text{tr}(H) + 0.5}.$$

The last approximation above^[8] reduces the computational cost from $O(n^2)$ to only $O(n)$. In general the numerator would be the objective function being minimized; e.g., if the hat matrix includes an observation covariance matrix, Σ , then $\|\hat{r}\|^2$ becomes $\hat{r}'\Sigma^{-1}\hat{r}$.

General

Note that unlike in the original case, non-integer degrees of freedom are allowed, though the value must usually still be constrained between 0 and n .

Consider, as an example, the k -nearest neighbour smoother, which is the average of the k nearest measured values to the given point. Then, at each of the n measured points, the weight of the original value on the linear combination that makes up the predicted value is just $1/k$. Thus, the trace of the hat matrix is n/k . Thus the smooth costs n/k effective degrees of freedom.

As another example, consider the existence of nearly duplicated observations. Naive application of classical formula, $n - p$, would lead to over-estimation of the residuals degree of freedom, as if each observation were independent. More realistically, though, the hat matrix $H = X(X' \Sigma^{-1} X)^{-1}X' \Sigma^{-1}$ would involve an observation covariance matrix Σ indicating the non-zero correlation among observations. The more general formulation of effective degree of freedom would result in a more realistic estimate for, e.g., the error variance σ^2 .

Other formulations

Similar concepts are the *equivalent degrees of freedom* in non-parametric regression,^[10] the *degree of freedom of signal* in atmospheric studies,^{[11][12]} and the *non-integer degree of freedom* in geodesy.^{[13][14]}

Alternative

The residual sum-of-squares $\|y - Hy\|^2$ has a generalized chi-squared distribution, and the theory associated with this distribution^[15] provides an alternative route to the answers provided above.

See also

- Pooled degrees of freedom
- Replication (statistics)
- Sample size
- Statistical model
- Variance

References

1. "Degrees of Freedom" (<http://www.animatedsoftware.com/statglos/sgdegree.htm>). *"Glossary of Statistical Terms"*. Animated Software. Retrieved 2008-08-21.
2. Lane, David M. "Degrees of Freedom" (<http://davidmlane.com/hyperstat/A42408.html>). *HyperStat Online*. Statistics Solutions. Retrieved 2008-08-21.
3. Walker, H. M. (April 1940). "Degrees of Freedom". *Journal of Educational Psychology* **31** (4): 253–269. doi:10.1037/h0054588 (<https://dx.doi.org/10.1037%2Fh0054588>).
4. Christensen, Ronald (2002). *Plane Answers to Complex Questions: The Theory of Linear Models* (Third ed.). New York: Springer. ISBN 0-387-95361-2.
5. Trevor Hastie, Robert Tibshirani, Jerome H. Friedman (2009), *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed., 746 p. ISBN 978-0-387-84857-0, doi:10.1007/978-0-387-84858-7 (<https://dx.doi.org/10.1007%2F978-0-387-84858-7>), [1] (<http://books.google.com/books?>

- id=tVIjmNS3Ob8C&lpg=PA153&dq=degrees%20of%20freedom%20of%20a%20smoother&pg=PA154#v=onepage&q=degrees%20of%20freedom%20of%20a%20smoother&f=false) (eq.(5.16))
6. Ye, J. (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection", *Journal of the American Statistical Association*, 93 (441), 120–131. JSTOR 2669609 (<http://www.jstor.org/stable/2669609>) (eq.(7))
 7. Clive Loader (1999), *Local regression and likelihood*, ISBN 978-0-387-98775-0, doi:10.1007/b98858 (<https://dx.doi.org/10.1007%2Fb98858>), [2] (<http://books.google.com/books?id=D7GgBAfL4ngC&lpg=PP1&pg=PA28#v=onepage&q=degree%20of%20freedom&f=false>) (eq.(2.18), p.30)
 8. Trevor Hastie, Robert Tibshirani (1990), *Generalized additive models*, CRC Press, [3] ([http://books.google.com/books?id=qz9r1Ze1coC&lpg=PR3&dq=Hastie%20T.%20J.%20and%20Tibshirani%20R.%20J.%20\(1990\)%20Generalized%20Additive%20Models%20London%3A%20Chapman%20and%20Hall.&pg=PA54#v=onepage&q=degrees%20of%20freedom&f=false](http://books.google.com/books?id=qz9r1Ze1coC&lpg=PR3&dq=Hastie%20T.%20J.%20and%20Tibshirani%20R.%20J.%20(1990)%20Generalized%20Additive%20Models%20London%3A%20Chapman%20and%20Hall.&pg=PA54#v=onepage&q=degrees%20of%20freedom&f=false)) (p.54) and (eq.(B.1), p.305))
 9. Simon N. Wood (2006), *Generalized additive models: an introduction with R*, CRC Press, [4] (<http://books.google.com/books?id=hr17lZC-3jQC&lpg=PA170&dq=Effective%20degrees%20of%20freedom&pg=PA172#v=onepage&q&f=false>) (eq.(4.14), p.172)
 10. Peter J. Green, B. W. Silverman (1994), *Nonparametric regression and generalized linear models: a roughness penalty approach*, CRC Press [5] (<http://books.google.com/books?id=AIVXozvpLUC&lpg=PA37&dq=generalized%20effective%20degrees%20of%20freedom&pg=PA37#v=onepage&q&f=false>) (eq.(3.15), p.37)
 11. Clive D. Rodgers (2000), *Inverse methods for atmospheric sounding: theory and practice*, World Scientific (eq.(2.56), p.31)
 12. Adrian Doicu, Thomas Trautmann, Franz Schreier (2010), *Numerical Regularization for Atmospheric Inverse Problems*, Springer (eq.(4.26), p.114)
 13. D. Dong, T. A. Herring and R. W. King (1997), Estimating regional deformation from a combination of space and terrestrial geodetic data, *J. Geodesy*, 72 (4), 200–214, doi:10.1007/s001900050161 (<https://dx.doi.org/10.1007%2Fs001900050161>) (eq.(27), p.205)
 14. H. Theil (1963), "On the Use of Incomplete Prior Information in Regression Analysis", *Journal of the American Statistical Association*, 58 (302), 401–414 JSTOR 2283275 (<http://www.jstor.org/stable/2283275>) (eq.(5.19)-(5.20))
 15. Jones, D.A. (1983) "Statistical analysis of empirical models fitted by optimisation", *Biometrika*, 70 (1), 67–88

Further reading

- Bowers, David (1982). *Statistics for Economists*. London: Macmillan. pp. 175–178. ISBN 0-333-30110-2.
- Eisenhauer, J. G. (2008). "Degrees of Freedom". *Teaching Statistics* 30 (3): 75–78. doi:10.1111/j.1467-9639.2008.00324.x (<https://dx.doi.org/10.1111%2Fj.1467-9639.2008.00324.x>).
- Good, I. J. (1973). "What Are Degrees of Freedom?". *The American Statistician* 27 (5): 227–228.

- doi:10.1080/00031305.1973.10479042 (<https://dx.doi.org/10.1080%2F00031305.1973.10479042>).
JSTOR 3087407 (<https://www.jstor.org/stable/3087407>).
■ Walker, H. W. (1940). "Degrees of Freedom". *Journal of Educational Psychology* **31** (4): 253–269. doi:10.1037/h0054588 (<https://dx.doi.org/10.1037%2Fh0054588>). Transcription by C Olsen with errata (http://courses.ncssm.edu/math/Stat_Inst/PDFS/DFWalker.pdf)

External links

- Yu, Chong-ho (1997) Illustrating degrees of freedom in terms of sample size and dimensionality (<http://www.creative-wisdom.com/computer/sas/df.html>)
- Dallal, GE. (2003) Degrees of Freedom (<http://www.tufts.edu/~gdallal/dof.htm>)

Retrieved from "[http://en.wikipedia.org/w/index.php?title=Degrees_of_freedom_\(statistics\)&oldid=649822959](http://en.wikipedia.org/w/index.php?title=Degrees_of_freedom_(statistics)&oldid=649822959)"

Categories: Statistical terminology

-
- This page was last modified on 4 March 2015, at 10:34.
 - Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

Announcement

How to Read the Output From Multiple Linear Regression Analyses

Here's a typical piece of output from a multiple linear regression of homocysteine (LHCY) on vitamin B12 (LB12) and folate as measured by the CLC method (LCLC). That is, vitamin B12 and CLC are being used to predict homocysteine. A (common) logarithmic transformation had been applied to all variables prior to formal analysis, hence the initial L in each variable name, but that detail is of no concern here.

Dependent Variable: LHCY

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	0.47066	0.23533	8.205	0.0004
Error	233	6.68271	0.02868		
C Total	235	7.15337			
Root MSE		0.16936	R-square	0.0658	
Dep Mean		1.14711	Adj R-sq	0.0578	
C.V.		14.76360			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	1.570602	0.15467199	10.154	0.0001
LCLC	1	-0.082103	0.03381570	-2.428	0.0159
LB12	1	-0.136784	0.06442935	-2.123	0.0348

Parameter Estimates.

The column labeled **Variable** should be self-explanatory. It contains the names of the predictor variables which label each row of output.

DF stands for **degrees of freedom**. For the moment, all entries will be 1. Degrees of freedom will be discussed in detail later.

The **Parameter Estimates** are the regression coefficients. The regression equation is

$$\text{LHCY} = 1.570602 - 0.082103 \text{ LCLC} - 0.136784 \text{ LB12}$$

To find the predicted homocysteine level of someone with a CLC of 12.3 and B12 of 300, we begin by taking logarithms. $\log(12.3)=1.0899$ and $\log(300)=2.4771$. We then calculate

$$\begin{aligned} \text{LHCY} &= 1.570602 - 0.082103 \cdot 1.0899 - 0.136784 \cdot 2.4771 \\ &= 1.1423 \end{aligned}$$

Homocysteine is the anti-logarithm of this value, that is, $10^{1.1423} = 13.88$.

The **Standard Errors** are the standard errors of the regression coefficients. They can be used for hypothesis testing and constructing confidence intervals. For example, confidence intervals for LCLC are constructed as $(-0.082103 \pm k \cdot 0.03381570)$, where k is the appropriate constant depending on the level of confidence desired. For example, for 95% confidence intervals based on large samples, k would be 1.96.

The **T** statistic tests the hypothesis that a population regression coefficient is 0 **WHEN THE OTHER PREDICTORS ARE IN THE MODEL**. It is the ratio of the sample regression coefficient to its

standard error. The statistic has the form (estimate - hypothesized value) / SE. Since the hypothesized value is 0, the statistic reduces to Estimate/SE. If, for some reason, we wished to test the hypothesis that the coefficient for LCFC was -0.100, we could calculate the statistic (-0.082103 - (-0.10))/0.03381570.

Prob > |T| labels the **P values** or the **observed significance levels** for the t statistics. The degrees of freedom used to calculate the P values is given by the Error DF from the ANOVA table. The P values tell us whether a variable has statistically significant predictive capability in the presence of the other variables, that is, whether it adds something to the equation. In some circumstances, a nonsignificant P value might be used to determine whether to remove a variable from a model without significantly reducing the model's predictive capability. For example, if one variable has a nonsignificant P value, we can say that it does not have predictive capability in the presence of the others, remove it, and refit the model without it. These P values should not be used to eliminate more than one variable at a time, however. A variable that does not have predictive capability in the presence of the other predictors may have predictive capability when some of those predictors are removed from the model.

The Analysis of Variance Table

The **Analysis of Variance** table is also known as the **ANOVA table** (for ANalysis Of VAriance). There is variability in the response variable. It is the uncertainty that would be present if one had to predict individual responses without any other information. The best one could do is predict each observation to be equal to the sample mean. The amount of uncertainty or variability can be measured by the Total Sum of Squares, which is the numerator of the sample variance. The ANOVA table partitions this variability into two parts. One portion is fitted by (many incorrectly say "explained by") the model. It's the reduction in uncertainty that occurs when the regression model is used to predict the responses. The remaining portion is the uncertainty that remains even after the model is used. The model is considered to be statistically significant if it can account for a large amount of variability in the response.

The column labeled **Source** has three rows, one for total variability and one for each of the two pieces that the total is divided into--**Model**, which is sometimes called **Regression**, and **Error**, sometimes called **Residual**. The **C** in **C Total** stands for **corrected**. Some programs ignore the **C** and label this **Total**. The **C Total Sum of Squares** and **Degrees of Freedom** will be the sum of Model and Error.

Sums of Squares: The total amount of variability in the response can be written $\Sigma (y - \bar{y})^2$, where \bar{y} is the sample mean. (The "Corrected" in "C Total" refers to subtracting the sample mean before squaring.) If we were asked to make a prediction without any other information, the best we can do, in a certain sense, is the sample mean. The amount of variation in the data that can't be accounted for by this simple method of prediction is given by the Total Sum of Squares.

When the regression model is used for prediction, the amount of uncertainty that remains is the variability about the regression line, $\Sigma (y - \hat{y})^2$. This is the Error sum of squares. The difference between the Total sum of squares and the Error sum of squares is the Model Sum of Squares, which happens to be equal to $\Sigma (\hat{y} - \bar{y})^2$.

Each sum of squares has corresponding degrees of freedom (DF) associated with it. Total df is one less than the number of observations, $n - 1$. The Model df is the number of independent variables in the model, p . The Error df is the difference between the Total df ($n - 1$) and the Model df (p), that is, $n - p - 1$.

The **Mean Squares** are the Sums of Squares divided by the corresponding degrees of freedom.

The **F Value** or **F ratio** is the test statistic used to decide whether the model as a whole has statistically significant predictive capability, that is, whether the regression SS is big enough, considering the number of variables needed to achieve it. F is the ratio of the Model Mean Square to the Error Mean Square. Under the null hypothesis that the model has no predictive capability--that is,

that all population regression coefficients are 0 simultaneously--the F statistic follows an F distribution with p numerator degrees of freedom and $n-p-1$ denominator degrees of freedom. The null hypothesis is rejected if the F ratio is large. Some analysts recommend ignoring the P values for the individual regression coefficients if the overall F ratio is not statistically significant, because of the problems caused by multiple testing. I tend to agree with this recommendation with one important exception. If the purpose of the analysis is to examine a particular regression coefficient after adjusting for the effects of other variables, I would ignore everything but the regression coefficient under study. For example, if in order to see whether dietary fiber has an effect on cholesterol, a multiple regression equation is fitted to predict cholesterol levels from dietary fiber along with all other known or suspected determinants of cholesterol, I would focus on the regression coefficient for fiber regardless of the overall F ratio. (This isn't quite true. I would certainly wonder why the overall F ratio was not statistically significant if I'm using the known predictors, but I hope you get the idea. If the focus of a study is a particular regression coefficient, it gets most of the attention and everything else is secondary.)

The **Root Mean Square Error** (also known as **the standard error of the estimate**) is the square root of the Residual Mean Square. It is the standard deviation of the data about the regression line, rather than about the sample mean.

R^2 is the squared multiple correlation coefficient. It is also called the **Coefficient of Determination**. R^2 is the ratio of the Regression sum of squares to the Total sum of squares, $\text{RegSS}/\text{TotSS}$. It is the proportion of the variability in the response that is fitted by the model. Since the Total SS is the sum of the Regression and Residual Sums of squares, R^2 can be rewritten as $(\text{TotSS}-\text{ResSS})/\text{TotSS} = 1 - \text{ResSS}/\text{TotSS}$. Some call R^2 *the proportion of the variance explained by the model*. I don't like the use of the word *explained* because it implies causality. However, the phrase is firmly entrenched in the literature. If a model has perfect predictability, $R^2=1$. If a model has no predictive capability, $R^2=0$. (In practice, R^2 is never observed to be exactly 0 the same way the difference between the means of two samples drawn from the same population is never exactly 0.) R , the multiple correlation coefficient and square root of R^2 , is the correlation between the observed values (y), and the predicted values ($y\hat{}$).

As additional variables are added to a regression equation, R^2 increases even when the new variables have no real predictive capability. The **adjusted-R²** is an R^2 -like measure that avoids this difficulty. When variables are added to the equation, adj- R^2 doesn't increase unless the new variables have additional predictive capability. Where R^2 is $1 - \text{ResSS}/\text{TotSS}$, we have

$\text{adj } R^2 = 1 - (\text{ResSS}/\text{ResDF})/(\text{TotSS}/(n-1))$, that is, it is 1 minus the ratio of (the square of the standard error of the estimate) to (the sample variance of the response). Additional variables with no explanatory capability will increase the Regression SS (and reduce the Residual SS) slightly, except in the unlikely event that the sample partial correlation is *exactly* 0. However, they won't tend to decrease the standard error of the estimate because the reduction in Residual SS will be accompanied by a decrease in Residual DF. If the additional variable has no predictive capability, these two reductions will cancel each other out.

Copyright © 2000 [Gerard E. Dallal](#)

Last modified: 05/23/2012 08:23:08.

R Tutorial

An R Introduction to Statistics

[HOME](#)
[DOWNLOAD](#)
[EBOOK](#)
[SITE MAP](#)
[CONTACT](#)

Significance Test for Linear Regression

Assume that the error term ϵ in the [linear regression model](#) is independent of x , and is [normally distributed](#), with zero [mean](#) and constant [variance](#). We can decide whether there is any [significant relationship](#) between x and y by testing the null hypothesis that $\beta = 0$.

Problem

Decide whether there is a significant relationship between the variables in the linear regression model of the data set [faithful](#) at .05 significance level.

Solution

We apply the `lm` function to a formula that describes the variable eruptions by the variable waiting, and save the linear regression model in a new variable `eruption.lm`.

```
> eruption.lm = lm(eruptions ~ waiting, data=faithful)
```

Then we print out the F-statistics of the significance test with the `summary` function.

```
> summary(eruption.lm)

Call:
lm(formula = eruptions ~ waiting, data = faithful)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2992 -0.3769  0.0351  0.3491  1.1933 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.87402   0.16014  -11.7   <2e-16 ***  
waiting       0.07563   0.00222    34.1   <2e-16 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.497 on 270 degrees of freedom
Multiple R-squared:  0.811,    Adjusted R-squared:  0.811 
F-statistic: 1.16e+03 on 1 and 270 DF,  p-value: <2e-16
```

Answer

As the p-value is much less than 0.05, we reject the null hypothesis that $\beta = 0$. Hence there is a significant relationship between the variables in the linear regression model of the data set [faithful](#).

Note

Further detail of the `summary` function for linear regression model can be found in the R documentation.

```
> help(summary.lm)
```

[< Coefficient of Determination](#)

[up](#)

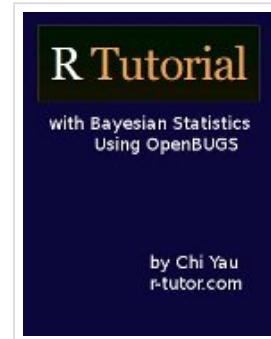
[Confidence Interval for Linear Regression >](#)

Tags: [Elementary Statistics with R](#) [error term](#) [linear regression](#) [significance test](#) [lm](#)

Search this site:

Search

R Tutorial eBook



R Tutorials

[R Introduction](#)

[Elementary Statistics with R](#)

[Qualitative Data](#)

[Quantitative Data](#)

[Numerical Measures](#)

[Probability Distributions](#)

[Interval Estimation](#)

[Hypothesis Testing](#)

[Type II Error](#)

[Inference About Two Populations](#)

[Goodness of Fit](#)

[Analysis of Variance](#)

[Non-parametric Methods](#)

[Simple Linear Regression](#)

[Estimated Simple Regression Equation](#)

[Coefficient of Determination](#)

[Significance Test for Linear Regression](#)

[Confidence Interval for Linear Regression](#)

[Prediction Interval for Linear Regression](#)

```
summary faithful
```

[Residual Plot](#)[Standardized Residual](#)[Normal Probability Plot
of Residuals](#)[Multiple Linear Regression](#)[Logistic Regression](#)[GPU Computing with R](#)

Recent Articles

- [Installing CUDA Toolkit 6.5 on Ubuntu 14.04 Linux](#)

September 3, 2014

- [Installing CUDA Toolkit 6.5 on Fedora 20 Linux](#)

September 3, 2014

- [Hierarchical Linear Model](#)

July 22, 2013

- [Bayesian Classification with Gaussian Process](#)

January 6, 2013

Copyright © 2009 - 2015 Chi Yau All Rights Reserved
Theme design by [styleshout](#) Fractal graphics by [zyzstar](#) Adaptation by [Chi Yau](#)

[To Documents](#)

The F-test for Linear Regression

Definitions for Regression with Intercept

- n is the number of observations, p is the number of regression parameters.
- **Corrected Sum of Squares for Model:** $SSM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$,
also called sum of squares for regression.
- **Sum of Squares for Error:** $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$,
also called sum of squares for residuals.
- **Corrected Sum of Squares Total:** $SST = \sum_{i=1}^n (y_i - \bar{y})^2$
This is the sample variance of the y-variable multiplied by n - 1.
- For multiple regression models, $SSM + SSE = SST$.
- **Corrected Degrees of Freedom for Model:** $DFM = p - 1$
- **Degrees of Freedom for Error:** $DFE = n - p$
- **Corrected Degrees of Freedom Total:** $DFT = n - 1$
Subtract 1 from n for the corrected degrees of freedom.
Horizontal line regression is the null hypothesis model.
- For multiple regression models with intercept, $DFM + DFE = DFT$.
- **Mean of Squares for Model:** $MSM = SSM / DFM$
- **Mean of Squares for Error:** $MSE = SSE / DFE$
The sample variance of the residuals.
- In a manner analogous to Property 10 of [Properties of Random Variables](#), which states that s^2 is unbiased for σ^2 , it can be shown that MSE is unbiased for σ^2 for multiple regression models.
- **Mean of Squares Total:** $MST = SST / DFT$
The sample variance of the y-variable.
- In general, a researcher wants the variation due to the model (MSM) to be large with respect to the variation due to the residuals (MSE).
- **Note:** the definitions in this section are not valid for regression through the origin models. They require the use of uncorrected sums of squares.

The F-test

- For a multiple regression model with intercept, we want to test the following null hypothesis and alternative hypothesis:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

$$H_1: \beta_j \neq 0, \text{ for at least one value of } j$$

This test is known as the overall **F-test for regression**.

- Here are the five steps of the **overall F-test for regression**

- State the null and alternative hypotheses:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

$$H_1: \beta_j \neq 0, \text{ for at least one value of } j$$

- Compute the test statistic assuming that the null hypothesis is true:

$$F = MSM / MSE = (\text{explained variance}) / (\text{unexplained variance})$$

- Find a $(1 - \alpha)100\%$ confidence interval I for (DFM, DFE) degrees of freedom using an F-table or statistical software.
 - Accept the null hypothesis if $F \in I$; reject it if $F \notin I$.
 - Use statistical software to determine the p-value.
- Practice Problem:** For a multiple regression model with 35 observations and 9 independent variables (10 parameters), SSE = 134 and SSM = 289, test the null hypothesis that all of the regression parameters are zero at the 0.05 level.

Solution: DFE = $n - p = 35 - 10 = 25$ and DFM = $p - 1 = 10 - 1 = 9$. Here are the five steps of the test of hypothesis:

- State the null and alternative hypothesis:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

$$H_1: \beta_j \neq 0 \text{ for some } j$$

- Compute the test statistic:

$$F = MSM/MSE = (SSM/DFM) / (SSE/DFE) = (289/9) / (134/25) = 32.111 / 5.360 = 5.991$$

3. Find a $(1 - 0.05) \times 100\%$ confidence interval for the test statistic. Look in the F-table at the 0.05 entry for 9 df in the numerator and 25 df in the denominator. This entry is 2.28, so the 95% confidence interval is $[0, 2.34]$. This confidence interval can also be found using the R function call `qf(0.95, 9, 25)`.
4. Decide whether to accept or reject the null hypothesis: $5.991 \notin [0, 2.28]$, so reject H_0 .
5. Determine the p-value. To obtain the exact p-value, use statistical software. However, we can find a rough approximation to the p-value by examining the other entries in the F-table for (9, 25) degrees of freedom:

Level	Confidence Interval	F-value
0.100	$[0, 0.900]$	1.89
0.050	$[0, 0.950]$	2.28
0.025	$[0, 0.975]$	2.68
0.010	$[0, 0.990]$	2.22
0.001	$[0, 0.999]$	4.71

The F-value is 5.991, so the p-value must be less than 0.005.

- Verify the value of the F-statistic for the [Hamster Example](#).

Technical Details for the Overall F-Test

- If t_1, t_2, \dots, t_m , are independent, $N(0, \sigma^2)$ random variables, then $\sum_{i=1}^m t_i^2$ is a χ^2 (chi-squared) random variable with m degrees of freedom.
- It can be shown that if H_0 is true and the residuals are unbiased, homoscedastic, independent, and normal:
 1. SSE / σ^2 has a χ^2 distribution with DFE degrees of freedom.
 2. SSM / σ^2 has a χ^2 distribution with DFM degrees of freedom.
 3. SSE and SSM are independent random variables.
- If u is a χ^2 random variable with n degrees of freedom, v is a χ^2 random variable with m degrees of freedom, and u and v are independent, then if $F = (u/n)/(v/m)$ has an **F distribution with (n,m) degrees of freedom**. See the F-tables in the [Statistical Tables](#).

- By the previous information, if H_0 is true, $F = [(SSM/\sigma)/DFM]/[(SSE/\sigma)/DFE]$ has an F distribution with (DFM, DFE) degrees of freedom.
- But $F = [(SSM/\sigma)/DFM]/[(SSE/\sigma)/DFE] = (SSM/DFM)/(SSE/DFE) = MSM/MSE$, so F is independent of σ .

The R^2 and Adjusted R^2 Values

- For simple linear regression, R^2 is the square of the sample correlation r_{xy} .
- For multiple linear regression with intercept (which includes simple linear regression), it is defined as $r^2 = SSM / SST$.
- In either case, R^2 indicates the proportion of variation in the y-variable that is due to variation in the x-variables.
- Many researchers prefer the **adjusted R^2 value** = \bar{R}^2 instead, which is penalized for having a large number of parameters in the model:

$$\bar{R}^2 = 1 - (1 - R^2)(n - 1) / (n - p)$$

- Here derivation of \bar{R}^2 : R^2 is defined as $1 - SSE/SST$ or $1 - R^2 = SSE/SST$. To take into account the number of regression parameters p , define the adjusted R-squared value as

$$1 - \bar{R}^2 = MSE/MST,$$

where $MSE = SSE/DFE = SSE/(n-p)$ and $MST = SST/DFT = SST/(n-1)$. Thus,

$$\begin{aligned} 1 - \bar{R}^2 &= [SSE/(n - p)] / [SST/(n - 1)] \\ &= (SSE/SST)(n - 1) / (n - p) \end{aligned}$$

so

$$\begin{aligned} \bar{R}^2 &= 1 - (SSE/SST)(n - 1) / (n - p) \\ &= 1 - (1 - R^2)(n - 1) / (n - p) \end{aligned}$$

The assumptions of the linear regression model

MICHAEL A. POOLE

(*Lecturer in Geography, The Queen's University of Belfast*)

AND PATRICK N. O'FARRELL

(*Research Geographer, Research and Development, Coras Iompair Eireann, Dublin*)

Revised MS received 10 July 1970

ABSTRACT. The paper is prompted by certain apparent deficiencies both in the discussion of the regression model in instructional sources for geographers and in the actual empirical application of the model by geographical writers. In the first part of the paper the assumptions of the two regression models, the 'fixed X' and the 'random X', are outlined in detail, and the relative importance of each of the assumptions for the variety of purposes for which regression analysis may be employed is indicated. Where any of the critical assumptions of the model are seriously violated, variations on the basic model must be used and these are reviewed in the second half of the paper.

THE rapid increase in the employment of mathematical models in planning has led R. J. Colenutt to discuss 'some of the problems and errors encountered in building linear models for prediction'.¹ Colenutt rightly points out that the mathematical framework selected for such models 'places severe demands on the model builder because it is associated with a highly restrictive set of assumptions . . . and it is therefore imperative that, if simple linear models are to be used in planning, their limitations should be clearly understood'.²

These models have also been widely used in geography, for descriptive and inferential purposes as well as for prediction, and there is abundant evidence that, like their colleagues in planning, many geographers, when employing these models, have not ensured that their data satisfied the appropriate assumptions. Thus many researchers appear to have employed linear models either without verifying a sufficient number of assumptions or else after performing tests which are irrelevant because they relate to one or more assumptions not required by the model. Furthermore, many writers, reporting geographical research, have completely omitted to indicate whether any of the assumptions have been satisfied. This last group is ambiguous, and it is clearly not possible, unless the values of the variables are published, to judge whether the correct set of assumptions has been tested or, indeed, to ascertain whether any such testing has been performed at all.

This problem partially arises from certain shortcomings in material which has been published with the specific objective, at least *inter alia*, of instructing geographers on the use of quantitative techniques. All of these sources make either incomplete or inaccurate specifications of the assumptions underlying the application of linear models, although it is encouraging to note that there has been a considerable improvement in the quality of this literature in recent years. Thus, there were four books and two articles published in the early and mid-1960s which may be classified as belonging to this body of literature,³ yet, in five of these six sources, only one of the assumptions of the model is mentioned and, even

in the other, only two are referred to. Many of these writers, it is true, discuss regression analysis too briefly to allow space for a comprehensive treatment of the assumptions of the model, but it is unfortunate that none of them did the same as J. B. Cole and C. A. M. King in 1968 who at least warned of the existence of unspecified dangers in the use of regression analysis.⁴ Except for this qualification, the work of Cole and King is similar to the earlier volumes, for only one of the model's assumptions is mentioned.⁵ However, M. H. Yeates's volume, published in 1968, represents a significant improvement, for three of the assumptions are referred to.⁶ This improvement has since continued, for much the most comprehensive coverage of these assumptions presented so far by a geographer has been that of L. J. King, published in 1969: he, in fact, alludes to each one of the model's seven assumptions.⁷ Nevertheless, L. J. King's account must be criticized for its unsystematic exposition of the assumptions, for its inaccurate or ambiguous treatment of three of them and for its failure to distinguish basic assumptions from rather less critical ones. Further, he fails to discuss either the task of testing to discover whether the assumptions are satisfied or the problem of finding an alternative method to overcome the difficulty encountered when testing reveals that an assumption is not satisfied.

There is a close parallel between this work of L. J. King, directed towards geographers, and that of Colenutt in the field of planning. Both writers have felt it necessary to warn their colleagues in their respective professions that the correct use of the regression model requires that several critical assumptions be satisfied. This at least implies that the model has been used carelessly in the past. In fact, King has explicitly pointed out that geographers have tended to employ correlation and regression analysis without showing sufficient awareness of the technical problems involved, prominent among which are the assumptions on which such analysis is based. The similarity between Colenutt's paper and the account of King even extends to the fact that neither of them presents a fully adequate or accurate account of the assumptions of the model, though it should be pointed out that the work of King is, in this respect, superior to that of Colenutt: thus the latter totally ignores two of the most critical assumptions, and his treatment of a third is in error.

Such inadequate treatment of the topic by planners and geographers suggests the need for a concise review of the assumptions of linear models, especially as the elementary statistical texts, such as those of M. J. Moroney and M. R. Spiegel,⁸ generally concentrate on outlining the computational procedures and ignore the underlying assumptions. Moreover, even the more advanced and specialized sources are rarely comprehensive in their treatment of these assumptions and their implications; and they tend necessarily, too, to submerge the assumptions in the details of the theory of mathematical statistics.

Therefore, the objective of this paper is to bring together, from many of the less elementary sources, material on two major topics relating to what is probably the most frequently applied of the linear models, the regression equation. The two topics on which attention is focused are:

- (1) The fundamental assumptions which must be satisfied if the application of the classical linear regression model is to be totally valid.
- (2) The alternative techniques which may be employed when these assumptions are not satisfied in any specific empirical problem; treatment of this second topic necessarily involves the discussion of tests designed to discover the extent to which the assumptions are satisfied, and an indication of how severe a deviation from each assumption may be tolerated without having to resort to the alternative techniques.

Since the aim is to present a concise review of these topics, theoretical proofs are not presented, nor are the computational procedures outlined; however, references to more detailed sources are provided.

THE CLASSICAL LINEAR REGRESSION MODEL

The assumptions of the model

The general single-equation linear regression model, which is the universal set containing simple (two-variable) regression and multiple regression as complementary subsets, maybe represented as

$$Y = a + \sum_{i=1}^k b_i X_i + u$$

where Y is the dependent variable; X_1, X_2, \dots, X_k are k independent variables; a and b_i are the regression coefficients, representing the parameters of the model for a specific population; and u is a stochastic disturbance-term which may be interpreted as resulting from the effect of unspecified independent variables and/or a totally random element in the relationship specified.⁹

So far, the discussion has proceeded as if there was only one general regression model. It is important to distinguish two distinct models, each of which is expressed in the form of the equation above. The critical difference between these two models concerns the nature of the independent variables, X_i ; in one, the X_i are held constant experimentally at certain fixed values while, in the other, the X_i values are selected at random.¹⁰ Therefore, since Y is random in both models, the 'fixed X ' model is characterized by a contrast between X_i and Y , for the former are fixed, while the latter is random; on the other hand, in the 'random X ' model, either sets of X_i and Y values are selected at random from a multivariate population, or else pairs of X and Y values are drawn at random from a bivariate population. The implication of this last difference is that, since the concept of correction is appropriate only for bivariate or multivariate populations, it follows that correlation analysis is valid only in the case of the 'random X ' model.¹¹

There are several research objectives for which the regression model may be used, but they may be classified into three groups: (i) the computation of point estimates, (ii) the derivation of interval estimates, and (iii) the testing of hypotheses. The assumptions to be satisfied for the proper application of the model vary with the research objective: in particular, the computation of point estimates requires a less restrictive set of assumptions than do the others, and it is therefore proposed to commence by considering such estimates.¹²

Four principal point estimates may be required. First, estimates are generally wanted for a and for the b_i in order to allow the derivation of a regression equation containing specific numerical constants. Secondly, it may be required to predict the expected value of Y corresponding to specific values of X_i ; thirdly, a point estimate of the variance of Y must be computed as an intermediate step in the deriving of interval estimates and in the testing of hypotheses; and fourthly, a point estimate of the correlation coefficient, r , may be obtained. It may be shown that the best (minimum variance) linear unbiased estimates of the regression parameters are derived by applying the least-squares method of computation to the sample data to be analysed. Moreover, the least-squares principle allows the derivation both of the best linear unbiased predictor of the expected value of Y for specific values of X_i and also of unbiased estimates of the variance of Y and of r .¹³

However, these results are conditional upon six critical assumptions being satisfied:

(1) Each value of X_i and of Y is observed without measurement error.¹⁴

Alternatively, this first assumption maybe partially relaxed to state that only X_i must be observed without measurement error but, in this case, the interpretation of u must be expanded to include not only the effect of unspecified independent variables and of an essentially random element in the relationship but also the error incurred in measuring Y .¹⁵ The second assumption is that of linearity.

(2) The relationships between Y and each of the independent variables X_i are linear in the parameters of the specific functional form chosen. Three of the four remaining assumptions relate to the attributes of the disturbance-term, u : the first two of them relate specifically to the nature of the condition distribution of u (i.e., the set of frequency distributions of u , each corresponding to specific values of X_i) and, therefore, by implication, are both concerned with the conditional distribution of Y .¹⁶

(3) Each conditional distribution of u has a mean of zero.

(4) The variance of the conditional distribution of u is constant for all such distributions; this is the homoscedasticity assumption.

(5) The values of u are serially independent; thus the values of u are independent of each other and their covariance is accordingly zero. If the fourth assumption is not satisfied in a specific empirical situation, heteroscedasticity is said to be present in the data, while, if the fifth assumption is not satisfied, autocorrelation is said to be present.

It should be noted that the requisite properties of the conditional distribution of the disturbances need hold-only for certain specific values in the 'fixed X ' model, whereas, in the 'random X ' model, these properties must be satisfied for every possible value of X .¹⁷

All of these five assumptions are critical for both simple and multiple regression, but the sixth of the fundamental assumptions is relevant only to the multiple regression model since it is concerned with the relationships between the independent variables.

(6) The independent variables, X_i , are linearly independent of each other. If this assumption is not satisfied in a specific rose, multicollinearity is said to be present.¹⁸

These six assumptions, which are all critical for point estimation in regression analysis, must also all be satisfied if the model is to be used for the purpose either of interval estimation or of hypothesis testing. However, if the regression model is to be employed for such inferential purposes, then these six assumptions are not sufficient for the valid application of the model, for one further assumption is needed. The precise form of this further assumption differs according to whether the model being employed is of the 'fixed X ' or 'random X ' type:

(7) The fixed X model requires that the conditional distribution of the disturbance-term must be normal in form, which clearly implies that the dependent variable, Y , has a normal conditional distribution. **The random X model requires that both the conditional and marginal distributions of each variable are normal:** this model thus requires not only conditional normality for Y , but also for X_i , and, in addition, the overall frequency distribution of each variable must be normal.

It may be added that, in relation to the calculation of point estimates, the assumption of the normal conditional distribution of the disturbance-term would allow the derivation of maximum likelihood estimates of the regression coefficients. However, these, in fact, turn out to be identical to the least-squares estimates.¹⁹ Therefore, neither form of the normality assumption is necessary for point estimation.²⁰ The reason that it is necessary, on the other

hand, for inferential problems is that the two statistics commonly used, Student's t and Snedecor's F , both require that the data be normally distributed. Even in relation to inferential problems, however, this assumption is not binding, provided that the sample is very large. Interval estimates are most often made when computing confidence intervals for the individual regression coefficients, a and b , and when calculating interval estimates for predicted values of Y corresponding to specific values of X ; in both these cases Student's t is used. Sometimes, however, the aim is to establish confidence regions for both or all regression coefficients simultaneously and, in this case, Snedecor's F is used. The most frequently performed tests relate to hypotheses on the value of individual regression coefficients and on the values of the entire set of such coefficients: when the hypothesized values are zero, this second test is equivalent to testing the significance of the regression model as a whole. For the first of these tests, either the t or F statistics may be used, but the second test requires Snedecor's F .²¹

In the case of the random X model, inferential analysis may also be carried out upon the correlation coefficient. Such inference includes the establishment of confidence intervals about r by means either of David's tables, based on the density function of r , or of Fisher's z transformation and the distribution of Student's t . It also includes both the test of the hypothesis that r equals zero, using Student's t , and also the test of the hypothesis that r equals some specific value other than zero, for which either David's tables or else Student's t with Fisher's z transformation may be employed. Such tests and estimation procedures are applicable not only to simple and multiple correlation coefficients but also to partial correlation coefficients.²²

The geographical literature on the assumptions of the regression model

It is claimed that the rustication for this paper is the inadequate attention given to the assumptions of the regression model in the geographical literature. Exempt for a passing mention given to those books written by geographers which purport to provide instruction on the use of regression analysis, no evidence has yet been presented to demonstrate the inadequacy of the attention given to the assumptions of the model. Therefore, now that each assumption has been described, it is proposed to examine briefly the extent to which each one has been alluded to by geographers when reporting specific applications of regression analysis. In the course of this examination, the comments made in the introduction about the treatment of the assumptions in the literature purporting to give instruction to geographers and planners on the use of the regression model will also be elaborated.

It is rare to find explicit reference to the assumption that there are no measurement errors in the data, though it might be argued that the need for accurate data is so obvious that it is taken for granted: A. H. Robinson and R. A. Bryson are among the few geographers who have referred explicitly to the problem.²³ Clearly, it is difficult to test for measurement error, even though A. R. Hill has conducted an experiment to isolate operator variability in pebble-orientation analysis.²⁴ However, there appears to be little awareness in the geographical literature of the fact that the presence of measurement error in the independent variables is a much more serious problem than its presence in the dependent variable. The exception to this statement is found in the book by L. J. King, for he does refer quite correctly to this measurement error assumption.²⁵

The linearity assumption is the only one which is mentioned by every single geographer giving instruction on the use of the regression model.²⁶ It is therefore no surprise that many

geographers, when using regression analysis, have been conscious of the need for the relationships investigated to be linear if the linear model is to be fitted: H. G. Kariel, and also J. F. Hart and N. E. Salisbury, are examples of geographers who have actually tested for the presence of linearity.²⁷

The assumption relating to the nature of the disturbance-term has been given much less attention in the geographical literature than has the linearity assumption. The assumption that each conditional distribution has a mean of zero appears to have been almost totally ignored, and the homoscedasticity assumption has not been given much more attention. B. J. L. Berry and H. G. Barnum have performed a logarithmic transformation in order to ensure greater homoscedasticity,²⁸ but there are few other explicit references in the geographical literature to testing for homoscedasticity, and it seems reasonable to conclude that most geographers have not verified this assumption of the model.

The third of the assumptions relating to the disturbance-term is that there should be no autocorrelation. Almost all the discussion of this problem in the literature of econometrics and statistics has dealt with the presence of autocorrelation in time-series data. Clearly, however, since geographical analysis is more concerned with spatial variation than with temporal variation, much of the data subjected to regression analysis by geographers refer to cross-sections through time, so the problem of time-series autocorrelation does not arise. But the correlation between values corresponding to successive time-periods, which is such a common feature of time-series data, has its parallel in the analysis of spatial variation at a cross-section through time, for there is frequently correlation between the values of the disturbance-term corresponding to contiguous spatial units. This problem of spatial autocorrelation is even more complex than the temporal autocorrelation problem, because there is more than one dimension involved in the spatial case. For a long time the only contribution on the spatial autocorrelation problem was that of the statistician R. C. Geary,²⁹ but some geographers have recently become aware of the problem: L. Curry has alluded to it, and M. F. Dacey has derived methods to test for the presence of spatial autocorrelation in data measured on a nominal scale.³⁰ However, it still appears to be rare for geographers using regression analysis to test for spatial autocorrelation.

Of all the geographers who have provided instruction on the use of regression analysis in their books, L. J. King is the only one who has mentioned any of these three assumptions relating to the characteristics of the disturbance-term.³¹ He refers to each of the three although, in stating the assumptions that the disturbances have a mean of zero, he fails to state that this relates only to the conditional distribution. This is in contrast to his treatment of the homoscedasticity assumption, which follows immediately after his reference to the zero mean assumption, for he does make it clear that homoscedasticity implies equal variance for all conditional distributions.

The last of the six basic assumptions of the regression model, the absence of multicollinearity, has been recognized by many geographers; and E. N. Thomas, R. A. Mitchell and D. A. Blome and Shue Tuck Wong are examples of writers who have examined their correction matrix for the presence of multicollinearity.³² Moreover, of the instructional sources, two of them, the books by Yeates and L. J. King, mention this assumption.³³ Multicollinearity, in fact, is probably second only to linearity among the six basic assumptions in the frequency with which it is alluded to by geographers using regression analysis. The remaining four of these six assumptions have been referred to much less frequently.

The most curious feature, however, of the geographical literature on the assumptions

of the regression model is that neither the presence of linearity, nor the absence of multicollinearity, nor any of the other basic assumptions of the regression model has been alluded to as frequently by geographers as has the property of normality. It is the presence of normality which has been most often stated by geographers to be a critical assumption of the model and it is normality whose presence has been most often tested for. Yet, as we have seen, this property is relevant only for interval estimation and hypothesis testing: it is not one of the six basic assumptions necessary for the initial point estimation.³⁴

Even more important, geographers, when testing distributions prior to using regression analysis, appear almost invariably to have examined the marginal distribution and to have ignored the conditional distribution. True, this has seldom been made absolutely explicit, for the term 'marginal distribution' never seems to have been employed. However, when geographers} reporting the use of regression analysis, write of testing 'the distribution of the individual variables'³⁵ or of 'normalizing the data by means of log transformation',³⁶ it seems likely that they are referring implicitly to the marginal distribution of the variables concerned. Yet, at least in the case of the fixed X model, the form of the marginal distribution is totally irrelevant. Clearly, in most instances in geographical research, it is, in fact, the random X model which is the appropriate one to employ, and this, of course, does require that the marginal distribution be normal. However, a normal marginal distribution is not alone sufficient, for it is essential that the conditional distribution also be normal.

Those sources purporting to provide instruction on the use of regression analysis are no better, for McCarty and Lindberg and also Yeates fail to state, or even imply, that the conditional distribution should be normal, and L. J. King, in referring to the normality-assumption, does riot make it clear whether he is referring to the marginal or the conditional distribution of the disturbance-term.³⁷ Moreover, none of these sources points out that the normality assumption is relevant only for interval estimation and hypothesis testing: true, King does state that the assumptions of the regression model are of varying importance, depending on whether or not the work has an inferential purpose, but he fails to state which specific assumptions this statement refers to.³⁸

Before concluding this discussion of the extent to which geographers have shown an awareness of the assumptions of the regression model, detailed mention should be made of the paper written on this model by Colenutt for, although directed primarily towards planners, it is published in a source familiar to many geographers. Of the six basic assumptions of the model, four are alluded to by Colenutt, but he omits to state that the conditional distributions of the disturbance-term should each have a mean of zero and a constant variance. Moreover, in relation to the normality assumption, he makes the same error that so many geographers commit by omitting to say that the conditions} distribution of the variables should be normal.³⁹

Thus, although L. J. King and Colenutt have provided much better instruction or-i the assumptions necessary for the valid use of regression analysis than have previous writers in the disciplines of geography and planning, there are deficiencies even in their accounts. Since the geographical and planning sources on the use of the model thus exhibit inadequacies and since the record of application of the model in these disciplines has been rather unsatisfactory, there appears to be some need for a paper whose objectives are to state clearly to geographers the assumptions of the regression model, to indicate how these assumptions may be tested for and to describe alternative models which may be used when certain assumptions are not satisfied.

ALTERNATIVES TO THE CLASSICAL LINEAR REGRESSION MODEL

Since the application of classical linear regression analysis, to be totally valid, requires that so many assumptions are satisfied, it follows that the testing of these assumptions is a critical part of any such analysis. Yet the geographical literature on regression analysis contains few detailed references to such testing: thus the only instructional sources including such references are the volumes by L. J. King and by Cole and C. A. M. King. Both of them contain descriptions of testing for normality, though not in relation to regression analysis specifically,⁴⁰ and L. J. King also describes the tests for autocorrelation devised by Geary and Dacey.⁴¹ Reports on specific empirical applications of regression analysis by geographers often simply state that some assumption has been tested for, without indicating the method used. It is clear, however, from those reports which do specify the actual test used, that geographers have made little use of statistical inference procedures when performing such tests: in fact, most of the assumption-testing which has been done has consisted of the visual inspection of graphs, such as the fractile diagrams used by P. D. La Valle to test for normality and the scattergrams employed by Kariel to test for linearity.⁴²

If testing reveals that a particular assumption of the classical regression model is not satisfied, then some alternative to the straightforward application of this classical model must be resorted to. Those alternative methods, which still basically involve regression analysis, are of two main types: either the input data maybe transformed, most frequently by applying a logarithmic, reciprocal, power or arcsin transformation,⁴³ or else a variation on the classical regression model may be applied. Geographers have frequently used the first of these methods, 'transformation, though almost entirely in order to satisfy the normality and linearity assumptions. Thus almost all the instructional sources refer to the possibility of transforming the data to achieve linearity and D. S. Knos is an example of a geographer who has done this when working on a specific empirical problem.⁴⁴ Far fewer of the instructional sources allude to transformation as a way to achieve normality, but many geographers, such as G. Olsson and A. Persson and La Valle, have, in fact, done this in empirical work.⁴⁵ Berry and Barnum, in contrast, are two of the few geographers who have transformed data in order to ensure greater homoscedasticity.⁴⁶ The second of the alternative ways of satisfying the assumptions of the model, by using a variant on classical regression analysis, appears to have been almost totally ignored by geographers; among the few references of this type are the allusions by P. Haggett and R. J. Chorley to the use of polynomials when the linearity assumption cannot be satisfied, even by transformation.⁴⁷ In this paper, on the other hand, a brief outline will be given of both types of method and of the circumstances in which they maybe used. The topic will be approached by discussing each of the seven assumptions in turn,

Assumptions on measurement error and linearity

The incurring of measurement error in the observation of the independent variables, X , would lead to biased estimates of the regression coefficients if the classical regression model were used.⁴⁸ However, any test of the degree of measurement error is clearly made difficult by the fact that the amount of error is unknown. In fact, for most purposes, it is frequently assumed, at least in econometrics, that measurement error is much less significant than errors resulting from incorrect equation specification, so the former is generally ignored.⁴⁹

The problem of measurement error can also be ignored if the sole objective of the regression analysis is to predict the value of Y corresponding to a given set of X_i values.⁵⁰

Prediction however, is rarely the sole reason for performing regression analysis, and it sometimes happens that serious measurement error is suspected in the data, so that some other method than classical least-squares must be adopted. Since the mathematics of some of these methods are complex, most writers discuss them in relation only to simple regressions⁵¹ and this convention will be followed here. There is quite an easy method which is applicable when measurement error occurs in the independent variable, X , in simple regression, but Y is observed without error. The solution in this case is to reverse the roles of X and Y by using the former as the dependent variable and the latter as the independent variable: having made this adjustment, ordinary least-squares estimation is then valid, provided, of course, that the other assumptions of the model are satisfied.⁵²

It is when serious measurement error occurs in the variables on both sides of the equation that more elaborate methods are required. The first such method is to assume that the measurement errors are serially independent and normally distributed and to use estimates of the error variances as weights in a modified application of the least-squares model.⁵³ The second method is to rank and group the data and, assuming that the errors are serially independent but not necessarily normally distributed, to derive the regression coefficients by manipulating the subgroup means for the respective variables.⁵⁴ The third and final method involves the use of instrumental variables, which are independent of the errors and highly correlated with the true values of the variables, and the manipulation of the deviations of individual values from mean values for both the original variables and the instrumental variables.⁵⁵

Turning from measurement error to equation specification error, it is now necessary to consider tests relating to the linearity assumption of the regression model and to indicate the procedures available when this assumption is significantly violated. Testing for the linearity of a relationship may take one of three forms. Either, having fitted a high-order polynomial function, the regression coefficients for the terms of second or higher order may be tested for significant departures from zero; or, after stratifying the data on the basis of the X values, a regression equation may be calculated for each stratum and the significance of the differences between each of the slope coefficients maybe tested; or else the sequence of residuals, arranged in order of increasing X , maybe tested for randomness.⁵⁶

If the application of any of these tests suggests that the linearity assumption is not satisfied in a specific instance, the input data are generally transformed to yield new data which satisfy this assumption more closely: ordinary linear regression can then be applied to these transformed data.⁵⁷ Alternatively, either the attempt maybe made to fit a higher-order polynomial function to the original data which appear to be linked in a curvilinear relationship,⁵⁸ or else one of several iterative methods of estimating the parameters of other non-linear functions may be used.⁵⁹

Assumptions on the pattern of disturbances

It is impossible to test directly, in any specific empirical example, the validity of the four assumptions relating to characteristics of the disturbances, for these characteristics are unknown because the disturbances are unobservable. However, tests may be carried out on the pattern of the residuals, using this as an estimate of the pattern of disturbances.⁶⁰

The first of these assumptions relating to the nature of the disturbance is that the mean disturbance is zero for each value of X_i . In practice, the principal point is that the residual mean, e , should be independent of X_i . However, the bias introduced into the model when this assumption is not satisfied is small, provided that the residual variance is small.⁶¹ Testing thus involves measuring both the correlation between e and X_i and also the variance of e . If such testing reveals however, that the assumption is so poorly satisfied that a considerable bias is introduced into the estimation of the regression coefficients, then the form of the specification of the relationship between X_i and Y should be changed and an alternative equation used.

The homoscedasticity assumption states that the conditional disturbance distribution should have a variance which is constant for all X_i values but, again, the major requirement in practice is that the residual variance should be independent of X_i . In fact, if the variance of e is not constant, but is independent of X_i , the estimates of the regression coefficients are still unbiased, though the usual methods of statistical inference are invalid. However, if the variance of e is not only subject to variation, but is correlated with X_i , then the estimates of the regression coefficients are seriously biased, and valid inference is also impossible.⁶²

No precise test of homoscedasticity is possible, because the tests available, such as those of Hartley or Bartlett, are highly sensitive to non-normality in the data.⁶³ However, if there does appear to be a correlation between X_i and the variance of e , then either the input data may be transformed in order to try to reduce or eliminate the heteroscedasticity,⁶⁴ or else a modified form of the regression model, weighted regression, may be used. In this modified regression model, weights, which are proportional to the variance of e , are applied to the variables; a frequent special case, used when the variance of e is proportional to X_i , arises when the ratio Y/X_i is used as the dependent variable instead of Y itself and the reciprocal of X_i is the independent variable.⁶⁵

The third assumption states that the errors are serially independent. It may be shown that, although the presence of autocorrelated disturbances does not prevent the derivation of an unbiased estimate of the regression coefficients, it does lead to two serious consequences, especially if the autocorrelation coefficient is high: first, the estimates of the regression coefficients have an unduly large and inaccurately estimated variance, and, secondly, the procedures for statistical inference are inapplicable.⁶⁶

The presence of autocorrelation in one-dimensional data, such as the values corresponding to a time series or to a cross-section through space, may be tested for by means of the Durbin-Watson d statistic: specifically, this tests for the existence of dependence between successive residuals, arrayed in order of temporal or spatial sequence and derived by the application of ordinary least-squares methods.⁶⁷ Testing for the presence of autocorrelation in the case of two-dimensional spatial data is more difficult, but a variant on the Durbin-Watson d statistic, called the contiguity ratio, was developed by Geary: essentially the use of this ratio tests for the similarity of the residuals corresponding to contiguous spatial units.⁶⁸ Dacey, on the other hand, in a suggested alternative way of testing for autocorrelation in two-dimensional data, proposes an extension of the conventional one-dimensional runs test: his method is to reduce the residual-term to a binary variable by distinguishing only between positive and negative residuals and then to test whether there is a significant tendency for contiguous areas to have residuals of the same sign.⁶⁹

If testing reveals that a set of data is autocorrelated, then two types of solution are available. First, since the autocorrelation probably results either from an error in the linearity

specification or from measurement error or from the effect of a variable excluded from the model, the attempt may be made to eliminate it by transforming the data or by introducing further independent variables into the model and then using ordinary least-squares methods.⁷⁰ Secondly, one of the several more complex methods available for the computation of the regression coefficients maybe used, though their application is generally made difficult by the fact that an estimation of, or assumptions about, the form of the autocorrelation function must be made.⁷¹

The last of the assumptions relating to the conditional disturbance distribution is that these distributions should be normal, but, even when the intention is to perform significance tests and establish confidence intervals, this assumption may frequently be relaxed. This is because such statistical inference procedures are not particularly sensitive to departures from normality: if the disturbances are non-normally distributed, the tests and intervals are still approximately correct and, indeed, if the sample is large, the approximations are extremely good.⁷² On the other hand, if the sample is small, it is very difficult to test for normality.

If the assumption of normality does appear to be seriously violated, the data may be transformed to derive more normal condition disturbance distributions. However, the robustness of regression analysis with respect to the assumption of normality and the fact that there is a greater need to satisfy such other assumptions as homoscedasticity and linearity, together have the result that transformations specifically for the purpose of imposing normality are infrequent.⁷³

The assumption of the absence of multicollinearity

The last of the assumptions of the classical linear regression model is that the independent variables, X_i , are linearly independent of each other. If this assumption is not satisfied and the independent variables are thus multicollinear, the result is that the individual regression Coefficients for each variable are not identifiable: in fact, the closer the linear correlation between the independent variables, the less the certainty with which these coefficients may be identified. This imprecision in the estimate of the regression coefficients is generally revealed by the occurrence of high standard errors. However, if the data contain measurement error, it can happen that standard errors are low despite the presence of multicollinearity, and, in this case, confluence analysis (bunch-map analysis) may be necessary to reveal the existence of the multicollinearity.⁷⁴

Because multicollinearity makes the regression coefficients quite unidentifiable, it is important, if the aim is to estimate the regression equation, to reduce it as much as possible. Either further data may be sought,⁷⁵ or certain variables maybe omitted from the model. If the latter solution is adopted, however, care must be taken in interpreting the resulting equation, for it cannot be assumed that an omitted variable has no effect: it is simply that its separate effect could not be isolated.⁷⁶ It maybe added, however, that, if the purpose of the regression analysis is only to predict the value of Y corresponding to a set of X_i values, then multicollinearity is not a serious problem, provided that the intercorrelations continue unchanged into the future.⁷⁷

CONCLUSION

This paper has attempted to summarize the major properties and assumptions of the linear regression model, and has reviewed and commented upon the shortcomings revealed by

some geographers in employing this model. In the case of each of the seven assumptions of the least-squares regression model, methods have been developed to overcome the problems presented when these assumptions are not satisfied in specific empirical situations. However, when alternative models are proposed as the solution, the model developed to overcome any one problem often cannot simultaneously handle other problems too, because it is highly dependent upon the other assumptions being satisfied.⁷⁸ thus one of the methods for overcoming the problem of measurement error depends upon the assumption that these errors are not autocorrelated and have a normal conditional distribution with zero mean and constant variance.⁷⁹ On the other hand, in the case of data transformations, it frequently happens in practice that a transformation which is designed to overcome the problems arising when one of the assumptions is not satisfied, simultaneously solves problems relating to other assumptions.⁸⁰

In addition to indicating methods of overcoming these problems when the assumptions of the simple model are not satisfied it has also been shown that the assumptions vary considerably in their significance. They vary both according to the purpose for which the model is to be used, especially in relation to whether or not any significance testing or derivation of confidence limits is to be done, and according to whether the purpose of the analysis is explanation or prediction: in the case of the latter, it is not essential to satisfy the assumptions of measurement error or multicollinearity. The assumptions also vary in the degree to which they are robust for any particular purpose. In general, however, it may be concluded that the normality, measurement error and zero disturbance-mean assumptions may be given less attention than is necessary in the case of the other four assumptions: it is of paramount importance that the relationships between variables be linear, that the disturbances be homoscedastic and serially independent and, if multiple regression is being performed, that the independent variables are not linearly correlated.

ACKNOWLEDGEMENTS

The authors are indebted to Mr S. B. Essig, Lecturer in Economics, The Queen's University of Belfast, for commenting upon an earlier draft of this paper.

NOTES

1. R. J. COLENUTT, 'Building linear predictive models for urban planning', *Reg. Stud.* 2 (1968), 140
2. *Ibid.*, 140
3. S. GREGORY, *Statistical methods and the geographer* (1963), 167-208; R. G. BARRY, 'An introduction to numerical and mechanical techniques' in F. J. MONKHOUSE and H. R. WILKINSON, *Maps and diagrams: their compilation and construction* (1963), 415-21; P. HAGGETT, *Locational analysis in human geography* (1965), 293-9; R. J. CHORLEY, 'The application of statistical methods to geomorphology' in G. H. DURY (ed.), *Essays in geomorphology* (1966), 340-56, 370-77; C. A. M. KING, *Techniques in geomorphology* (1966), 312-23; H. H. McCARTY and J. B. LINDBERG, *A preface to economic geography* (New Jersey, 1966), 71-81
4. J. B. COLE and C. A. M. KING, *Quantitative geography: techniques and theories in geography* (1968), 263
5. *Ibid.*, 138-46, 150-3, 287-94
6. M. H. YEATES, *An introduction to quantitative analysis in economic geography* (1968), 15-21, 81-2, 50-3, 100-6
7. L. J. KING, *Statistical analysis in geography* (New Jersey, 1969), 117-64
8. M. J. MORONEY, *Facts from figures* (1956), 276-320; M. R. SPIEGEL, *Theory and problems of statistics* (1961), 217-82
9. J. JOHNSTON, *Econometric methods* (1963), 5-6; F. A. GRAYBILL, *An introduction to linear statistical models*, vol. I (1961), 99-104

10. It was with reference to the 'random X ' model that the term 'regression' was originally used (G. SNEDECOR, *Statistical methods* (Ames, Iowa, 1956), 152-3), and some writers still restrict the use of the term in this way (F. A. GRAYBILL, op. cit., 101). However, it has become common to refer to the 'fixed X ' model, too, as a regression model (N. R. DRAPER and H. SMITH, *Applied regression analysis* (1966), 6).
11. F. S. ACTON, *Analysis of straight-line data* (1959), 7; F. A. GRAYBILL, op. cit., 206-7
12. F. A. GRAYBILL, op. cit., 109-10
13. J. JOHNSTON, op. cit., 9-20, 34-6, 108-13; F. A. GRAYBILL, op. cit., 114-17; E. MALINVAUD, *Statistical methods of econometrics* (Amsterdam, 1966), 78-81, 84-6, 97-9
14. E. MALINVAUD, op. cit., 75
15. F. S. ACTON, op. cit., 8; F. A. GRAYBILL, op. cit., 103-4
16. J. JOHNSTON, op. cit., 7-9, 106-8; E. MALINVAUD, op. cit., 73-86, 98-9, 173-4; N. R. DRAPER and H. SMITH, op. cit., 17; F. A. GRAYBILL, op. cit., 108-9, 114-17
17. J. JOHNSTON, op. cit., 25-9, 133; F. A. GRAYBILL, op. cit., 204-6
18. J. JOHNSTON, op. cit., 107-8; E. MALINVAUD, op. cit., 174-6
19. J. JOHNSTON, op. cit., 20-21, 115-6; E. MALINVAUD, op. cit., 86-9; F. A. GRAYBILL, op. cit., 110-14
20. J. JOHNSTON, op. cit., 21-5, 116-27; E. MALINVAUD, op. cit., 77, 89-92; N. R. DRAPER and H. SMITH, op. cit., 59; G. E. V. LESER, *Econometric techniques and problems* (1966), 9
21. J. JOHNSTON, op. cit., 23-5, 36-7, 118-33; E. MALINVAUD, op. cit., 91-2, 99-100, 199-205; F. A. GRAYBILL, op. cit., 120-45; N. R. DRAPER and H. SMITH, op. cit., 18-26, 63-7; F. S. ACTON, op. cit., 23-53
22. F. A. GRAYBILL, op. cit., 208-16
23. A. H. ROBINSON and R. A. BRYSON, 'A method for describing quantitatively the correspondence of geographical distributions', *Ann. Ass. Am. Geogr.* 47 (1957), 388
24. A. R. HILL, 'An experimental test for the field technique of till macrofabric analysis', *Trans. Inst. Br. Geogr.* 45 (1968), 93-105
25. L. J. KING, op. cit., 122-3
26. S. GREGORY, op. cit., 203; R. G. BARRY, op. cit., 417-18; P. HAGGETT, op. cit., 294-6; R. J. CHORLEY, op. cit., 341, 371, 374; C. A. M. KING, op. cit., 312; H. H. McCARTY and J. B. LINDBERG, op. cit., 71-2; J. B. COLE and C. A. M. KING, op. cit., 138; M. H. YEATES, op. cit., 15; L. J. KING, op. cit., 120
27. H. G. KARIEL, 'Selected factors areally associated with population growth due to net migration', *Ann. Ass. Am. Geogr.* 53 (1963), 215; J. F. HART and N. E. SALISBURY, 'Population changes in Middle Western villages: a statistical approach', *Ann. Ass. Am. Geogr.* 55 (1965), 151-2
28. B. J. L. BERRY and H. G. BARNUM, 'Aggregate relations and elemental components of central place systems', *J. reg. Sci.* 4 (1962), 36
29. R. C. GEARY, 'The contiguity ratio and statistical mapping', *Inc. Statist.* 5 (1954), 115-41
30. L. CURRY, 'Quantitative geography, 1967', *Canad. Geogr.* 11 (1967), 268-73; M. F. DACEY, 'A review on measures of contiguity for two and k-color maps', *Tech. Rep.* 2 (Spatial Diffusion Study, Dept. of Geography, Northwestern University, Evanston, Illinois, 1965)
31. L. J. KING, op. cit., 121-3, 157-162
32. E. N. THOMAS, R. A. MITCHELL and D. A. BLOME, 'The spatial behavior of a dispersed non-farm population', *Pap. reg. Sci. Ass.* 9 (1962), 125-6; SHUE TUCK WONG, 'A multivariate statistical model for predicting mean annual flood in New England', *Ann. Ass. Am. Geogr.* 53 (1963), 299
33. YEATES, op. cit., 81; L. J. KING, op. cit., 162-3
34. Two examples of papers, in which variables have been transformed in order to achieve normality, despite the use of the entire population of data, are: H. ALDSKOGIUS, 'Vacation house settlement in the Siljan region', *Geogr. Annalr* 49 B (1967), 78; and G. OLSSON and A. PERSSON, 'The spacing of central places in Sweden', *Pap. reg. Sci. Ass.* 12 (1964) 90
35. M. H. YEATES, 'Some factors affecting the spatial distribution of Chicago land values, 1910-1960', *Econ. Geogr.* 41 (1965), 63
36. E. J. TAFFE, R. L. MORRILL and P. R. GOULD, 'Transport expansion in underdeveloped countries: a comparative analysis', *Geogr. Rev.* 53 (1963), 516
37. H. H. McCARTY and J. B. LINDBERG, op. cit., 72; M. H. YEATES, *An introduction to quantitative analysis*, 20; L. J. KING, op. cit., 121-3
38. L. J. KING, op. cit., 123
39. R. J. COLENUTT, op. cit., 140-1
40. L. J. KING, op. cit., 82; J. B. COLE and C. A. M. KING, op. cit., 130-1
41. L. J. KING, op. cit., 158-60
42. P. D. LA VALLE, 'Some aspects of linear karst depression development in southcentral Kentucky', *Ann. Ass. Am. Geogr.* 57 (1967), 61; H. G. KARIEL, op. cit., 215
43. N. R. DRAPER and H. SMITH, op. cit., 131-4; F. S. ACTON, op. cit., 221-3; J. JOHNSTON, op. cit., 47-50
44. D. S. KNOS, 'The distribution of land values in Topeka, Kansas' in B. J. L. BARRY and D. F. MARBLE (eds.) *Spatial analysis: a reader in statistical geography* (New Jersey, 1968), 271-5

45. G. OLSSON and A. PERSSON, op. cit., 96; P. D. LA VALLE, op. cit., 61
46. B. J. L. BERRY and H. G. BARNUM, op. cit., 36
47. P. HAGGETT, op. cit., 296; R. J. CHORLEY, op. cit., 347-8
48. J. JOHNSTON, op. cit., 148-50; E. MALINVAUD, op. cit., 331-3
49. E. MALINVAUD, op. cit., 362
50. J. JOHNSTON, op. cit., 162-4
51. E. MALINVAUD, op. cit., 362
52. C. E. V. LESER, op. cit., 18
53. J. JOHNSTON, op. cit., 150-62; E. MALINVAUD, op. cit., 335-47
54. J. JOHNSTON, op. cit., 164-5; E. MALINVAUD, op. cit., 359-62
55. J. JOHNSTON, op. cit., 165-6; E. MALINVAUD, op. cit., 347-52
56. E. MALINVAUD, op. cit., 268-71
57. J. JOHNSTON, op. cit., 44-50; N. R. DRAPER and H. SMITH, op. cit., 131-4
58. N. R. DRAPER and H. SMITH, op. cit., 129-30, 150-55; G. SNEDECOR, op. cit., 452-4, 461-71; F. A. GRAYBILL, op. cit., 165-84; F. S. ACTON, op. cit., 193-218
59. N. R. DRAPER and H. SMITH, op. cit., 267-85; E. MALINVAUD, op. cit., 290-314
60. N. R. DRAPER and H. SMITH, op. cit., 86-97; C. E. V. LESER, op. cit., 14-15
61. E. MALINVAUD, op. cit., 258-60
62. Ibid., 254-7
63. C. E. V. LESER, op. cit., 13; F. S. ACTON, op. cit., 89-90
64. F. S. ACTON, op. cit., 90, 219
65. J. JOHNSTON, op. cit., 208-11; E. MALINVAUD, op. cit., 257-58; C. E. V. LESER, op. cit., 13-14
66. J. JOHNSTON, op. cit., 179, 187-92; E. MALINVAUD, op. cit., 433-9
67. J. JOHNSTON, op. cit., 192; E. MALINVAUD, op. cit., 421-5
68. R. C. GEARY, op. cit., 115-41
69. M. F. DACEY, op. cit.
70. C. E. V. LESER, op. cit., 17
71. J. JOHNSTON, op. cit., 179-87, 192-5; E. MALINVAUD, op. cit., 439-45
72. E. MALINVAUD, op. cit., 93
73. F. S. ACTON, op. cit., 220
74. J. JOHNSTON, op. cit., 201-7; C. E. V. LESER, op. cit., 27
75. J. JOHNSTON, op. cit., 207
76. C. E. V. LESER, op. cit., 28
77. J. JOHNSTON, op. cit., 207; C. E. V. LESER, op. cit., 28
78. J. JOHNSTON, op. cit., 147
79. E. MALINVAUD, op. cit., 329
80. F. S. ACTON, op. cit., 221

RÉSUMÉ. *Les hypothèses du modèle de régression linéaire.* Ce qui a inspiré cet exposé, c'est qu'il semble y avoir quelques insuffisances tant dans la discussion du modèle de régression dans les instructions fournies pour les géographes que dans l'application empirique elle-même du modèle par les écrivains de la géographie. Dans la première partie de l'exposé, les hypothèses des deux modèles de régression, le «X fixe» et le «X pris au hasard», sont indiquées en détail, et l'on a aussi indiqué l'importance relative de chacune des hypothèses dans tous les usages où l'on pourrait employer l'analyse de régression. Dans le cas où quelques-unes des hypothèses critiques du modèle portent gravement à faux, il faut employer des variations du modèle fondamental, et ces variations sont examinées dans la deuxième partie de l'exposé.

ZUSAMMENFASSUNG. *Die Annahmen des linearen Regressionsmodells.* Die Abhandlung beschäftigt sich mit scheinbaren Unzulänglichkeiten sowohl in der Besprechung des Regressionsmodells als auch in Lehrmaterial für Geographen und in der tatsächlichen erfahrungsmässigen Anwendung des Modells durch geographische Schriftsteller. Im ersten Teil der Abhandlung sind die Annahmen der zwei Regressionsmodelle, das ‚festgelegte X‘ und das ‚wahllose X‘ in Einzelheiten umrissen und die relative Wichtigkeit von jeder dieser beiden Annahmen, für die verschiedenen Möglichkeiten auf welche die Regressionsanalyse angewandt werden kann, ist ange deutet. Wo eine der kritischen Annahmen des Modells ernstlich übertreten wird, müssen Variationen des Ausgangsmodells benutzt werden und diese werden in der zweiten Hälfte der Abhandlung besprochen.

Lecture Outline (week 10)

Polynomial Regression Models

(quadratic models, Partial F tests, Type 1 SS, Residuals, Partial Correlation)

Multiple Regression Models

Models with One Qualitative and One Quantitative Variables

Polynomial Regression Models

(quadratic models, Partial F tests, Type 1 SS)

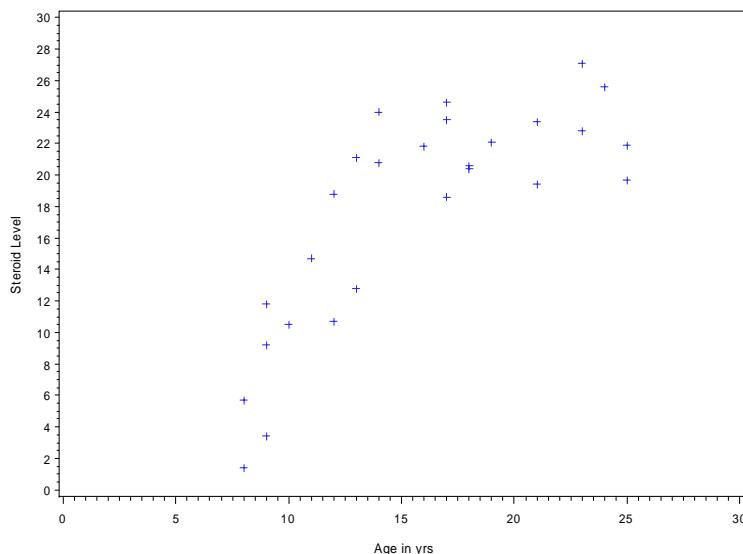
Example 1: Steriod Levels among Women age 8-25

(Problem 8 (p336) Chapter 8 (Kutner, Nachtsheim, neter, and Li (5th edition: Applied Linear Statistical Models) (program esb10p24.sas, esb10p25.sas))

Construct Scatter Plots with Independent Variables

```
FOOTNOTE "&prg";
AXIS1 LABEL=(ANGLE=90 " Steroid Level " ROTATE=0) ORDER=0 to 30 by 2;
* Defines label for y axis;
AXIS2 LABEL=( " Age in yrs" ROTATE=0) ORDER=0 to 30 by 5 ;
PROC GPLOT DATA=d2;
PLOT steroid*age/ vaxis=axis1 haxis=axis2;
TITLE1 "Figure 1. Scatter plot of Steroid by age ";
RUN;
```

Figure 1. Scatter plot of Steroid by age



Source: esb10p24.sas 3/9/2010 by ejs

```

DATA d2;
  SET d1;
  age2=age*age;
  RUN;
PROC SORT DATA=d2;
  BY age;
RUN;
PROC PRINT DATA=d2 (OBS=10) NOOBS;
  VAR steroid age age2;
  TITLE2 "Table 1. Example of Data for Steroids";
  RUN;

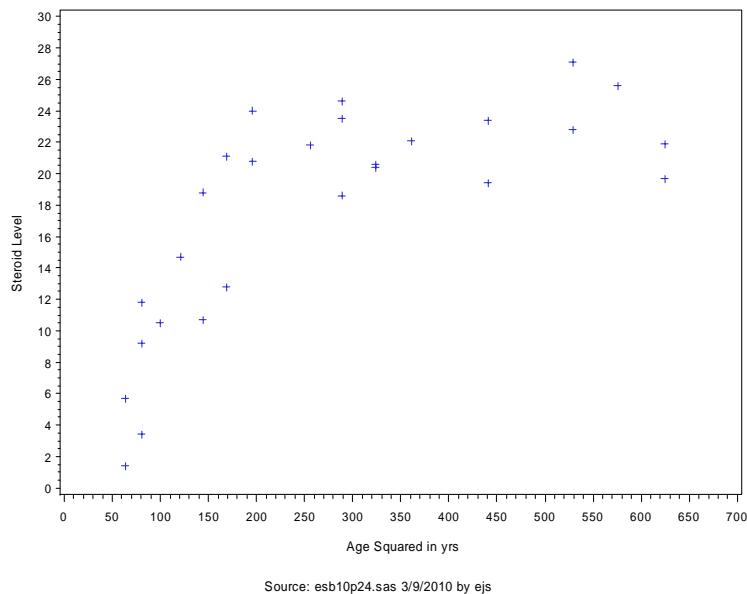
```

Table 1. Example of Data for Steroids

steroid	age	age2
1.4	8	64
5.7	8	64
9.2	9	81
11.8	9	81
3.4	9	81
10.5	10	100
14.7	11	121
10.7	12	144
18.8	12	144
21.1	13	169

Source: esb10p24.sas 3/9/2010 by ejs

Figure 2. Scatter plot of Steroid by age squared



Source: esb10p24.sas 3/9/2010 by ejs

Notes:

This scatter plot doesn't look like a straight line would provide a good fit. Still, it appears that a line would be better than a horizontal line.

This scatter plot does not account for what might be explained by a linear regression with age.

Evaluate Correlation of Steroids with age, age squared.

```
PROC CORR DATA=d2;
  VAR steroid age age2;
  TITLE2 "Table 3. Correlation of Steroids with other variables";
RUN;
```

Table 3. Correlation of Steroids with other variables

The CORR Procedure

3 Variables: steroid age age2						
Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
steroid	27	17.64444	7.02963	476.40000	1.40000	27.10000
age	27	15.77778	5.50058	426.00000	8.00000	25.00000
age2	27	278.07407	181.68060	7508	64.00000	625.00000

Pearson Correlation Coefficients, N = 27
Prob > |r| under H0: Rho=0

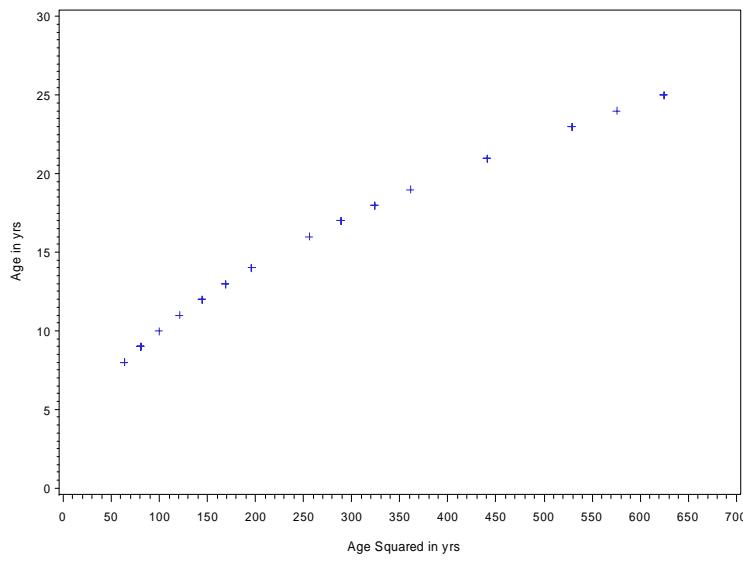
	steroid	age	age2
steroid	1.00000	0.78577	0.71312
		<.0001	<.0001
age	0.78577	1.00000	0.98943
	<.0001		<.0001
age2	0.71312	0.98943	1.00000
	<.0001	<.0001	

Source: esb10p24.sas 3/9/2010 by ejs

Notes:

There is a large correlation of age with age squared, but not a perfect correlation. When two variable are highly correlated, they are called ‘collinear’. A scatter plot illustrates this.

Figure 3. Scatter plot of Age by Age Squared



Source: esb10p24.sas 3/9/2010 by ejs

Fit Regression Models

Model 1: $Y_i = \beta_0 + X_{1i}\beta_1 + E_i$ $X_{1i} = \text{Age}$

Model 2: with polynomial regression, only consider hierarchical models (models where the lower order polynomial terms are included). This means that we would not consider fitting a model with only a quadratic term.

Model 3: $Y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + E_i$ $X_{1i} = \text{Age}$ $X_{2i} = \text{Age squared}$

Table 4. Regression of Steroids with age (Model 1)

Analysis of Variance

Source	DF	Sum of		F Value	Pr > F
		Squares	Mean Square		
Model	1	793.28051	793.28051	40.35	<.0001
Error	25	491.52616	19.66105		
Corrected Total	26	1284.80667			
Root MSE	4.43408	R-Square	0.6174		
Dependent Mean	17.64444	Adj R-Sq	0.6021		

Table 6. Regression of Steroids with age and age squared (Model 3)

Analysis of Variance

Source	DF	Sum of		F Value	Pr > F
		Squares	Mean Square		
Model	2	1046.26586	523.13293	52.63	<.0001
Error	24	238.54081	9.93920		
Corrected Total	26	1284.80667			
Root MSE	3.15265	R-Square	0.8143		
Dependent Mean	17.64444	Adj R-Sq	0.7989		

Notes:

Corrected Total Sums of squares is the same in all models.

The Adjusted R-square is largest with Model 3.

Comparing Models

Comparison of Model 3 with Model 1:

Model 1: $Y_i = \beta_0 + X_{1i}\beta_1 + E_i$ $X_{1i} = \text{Age}$

Model 3: $Y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + E_i$ $X_{1i} = \text{Age}$ $X_{2i} = \text{Age squared}$

Null Hypothesis: There is no difference between Model 3 and Model 1

or $H_0 : \beta_2 = 0$

Alternative Hypothesis: $H_a : \beta_2 \neq 0$

To test this Hypothesis, we use a Partial F-test

Extra Sum of Squares

Define: Regression sum of Squares:

$SSR(X_1)$ = sums of squares explained by including X_1

Examples:

For Model 1: $SSR(X_1) = 793.28$ df=1

For Model 3: $SSR(X_1, X_2) = 1046.26$ df=2

Extra sum of squares:

$$\begin{aligned} SSR(X_2 | X_1) &= SSR(X_1, X_2) - SSR(X_1) \\ &= 1046.26 - 793.28 \\ &= 252.98 \end{aligned}$$

Extra Mean Square: $MSR(X_2 | X_1) = 252.98 / 1 = 252.98$

Partial F-test: $F_{cal} = \frac{MSR(X_2 | X_1)}{MSE} = \frac{252.98}{9.939} = 25.45$ (denominator is MSE for Model 3)

Compare with F with 1 and 24 DF

```

PROC REG DATA=d2;
  MODEL steroid=age age2 /SS1;
  PLOT p.*age steroid*age/OVERLAY;
  TITLE2 "Table 6. Regression of Steroids with age and age squared (Model
3)";
  RUN;

```

Table 6. Regression of Steroids with age and age squared

Dependent Variable: steroid

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1046.26586	523.13293	52.63	<.0001
Error	24	238.54081	9.93920		
Corrected Total	26	1284.80667			
Root MSE		3.15265	R-Square	0.8143	
Dependent Mean		17.64444	Adj R-Sq	0.7989	
Coeff Var		17.86766			

Parameter Estimates						
	Parameter	Standard				
Variable	DF	Estimate	Error	t Value	Pr > t	Type I SS
Intercept	1	-26.32541	5.88154	-4.48	0.0002	8405.81333
age	1	4.87357	0.77515	6.29	<.0001	793.28051
age2	1	-0.11840	0.02347	-5.05	<.0001	252.98535

Source: esb10p24.sas 3/9/2010 by ejs

The partial F-test is $F_{cal} = (-5.05)^2$.

Residuals

Fit Model of Steriods on Age, and Get Residuals

```
PROC REG DATA=d2;
  MODEL steroid=age ;
  OUTPUT OUT=e1  p=yhat  r=yresid;
  TITLE2 "Simple regression model on age";
  RUN;
PROC PRINT DATA=e1 (OBS=10) NOOBS;
  TITLE2 "Table 7. List of Residuals from Reg of Steroids on Age";
  RUN;
```

Table 7. List of Residuals from Reg of Steroids on Age

steroid	age	age2	yhat	yresid
27.1	23	529	24.8970	2.20304
22.1	19	361	20.8802	1.21982
21.9	25	625	26.9054	-5.00535
10.7	12	144	13.8508	-3.15082
1.4	8	64	9.8340	-8.43404
18.8	12	144	13.8508	4.94918
14.7	11	121	12.8466	1.85338
5.7	8	64	9.8340	-4.13404
18.6	17	289	18.8718	-0.27179
20.4	18	324	19.8760	0.52401

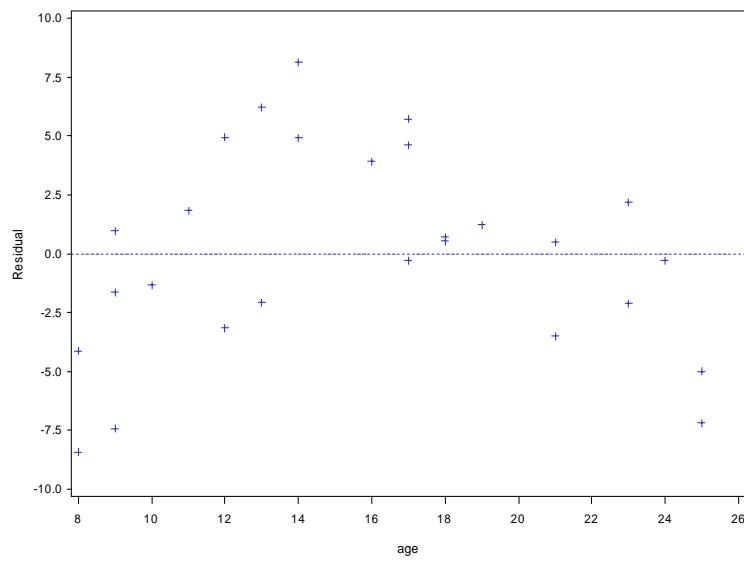
Source: esb10p25.sas 3/23/2010 by ejs

Construct Residual Plots and Studentized Residuals

Studentized Residual: Fit model without the observation, calculate the residual using the observation, divide by the standard error based on fitted model. (If error is normally distributed, residuals should be between -2 and 2.)

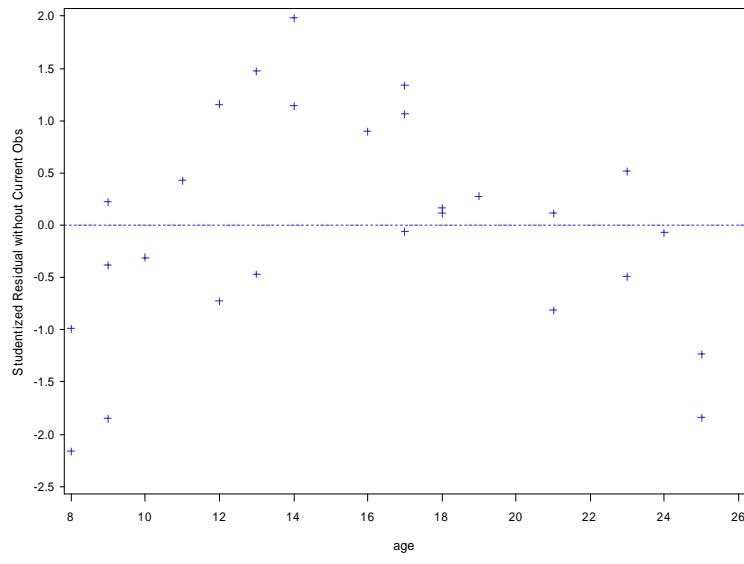
```
FOOTNOTE "&prg";
PROC REG DATA=d2;
  MODEL steroid=age ;
  OUTPUT OUT=e1  p=yhat  r=yresid;
  PLOT r.*age /NOMODEL NOSTAT ;
  PLOT rstudent.*age /NOMODEL NOSTAT ;
  TITLE1 "Figure 3. Residuals from Regression on Age (Model 1)";
  RUN;
FOOTNOTE ;
TITLE1 "&prg" ;
```

Figure 3. Residuals from Regression on Age (Model 1)



Source: esb10p25.sas 3/23/2010 by ejs

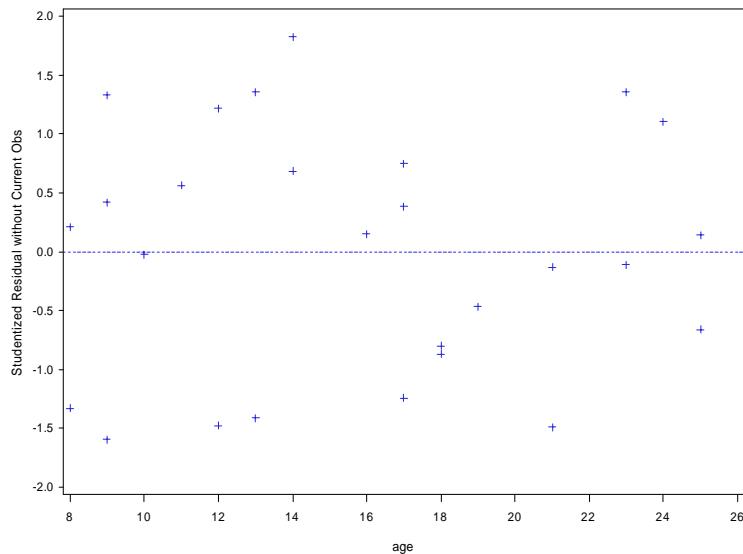
Figure 3. Residuals from Regression on Age (Model 1)



Source: esb10p25.sas 3/23/2010 by ejs

```
FOOTNOTE "&prg";
PROC REG DATA=d2;
  MODEL steroid=age age2;
  OUTPUT OUT=e1 p=yhat r=yresid;
  PLOT rstudent.*age /NOMODEL NOSTAT ;
  TITLE1 "Figure 4. Residuals from Regression on Age and Age2 (Model 3)";
  RUN;
FOOTNOTE ;
TITLE1 "&prg" ;
```

Figure 4. Residuals from Regression on Age and Age2 (Model 3)



Source: esb10p25.sas 3/23/2010 by ejs

Figure 4. Residuals from Regression on Age and Age2 (Model 3)
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1046.26586	523.13293	52.63	<.0001
Error	24	238.54081	9.93920		
Corrected Total	26	1284.80667			

Building Regression Models by new models for Residuals:

Fit Model of Steriods on Age, and Get Y-Residuals

Fit Model of Age2 on Age, and Get X2-Residuals

Fit Model of Y-Residuals on X2-Residuals

```
PROC REG DATA=d2;
  MODEL steroid=age ;
  OUTPUT OUT=e1 p=yhat r=yresid;
  TITLE2 "Table 7. Simple Regression on Age (Model 1)";
  RUN;
PROC REG DATA=d2;
  MODEL age2=age ;
  OUTPUT OUT=e2 r=age2_resid;
  TITLE2 "Table 8. Simple Regression on Age (Model 1)";
  RUN;
DATA d3;
  MERGE e1 (KEEP=yresid age)
    e2 (KEEP=age2_resid age2);
RUN;
PROC PRINT DATA=d3 (OBS=10) NOOBS;
  VAR yresid age2_resid;
  TITLE2 "Table 8. List of residuals ";
RUN;

PROC REG DATA=d3;
  MODEL yresid=age2_resid;
  TITLE2 "Table 9. Regression of Residuals on Age Squared (like Model 3)";
  RUN;
```

Table 8. List of residuals

	age2_
yresid	resid
2.20304	14.9021
1.21982	-22.3770
-5.00535	45.5416
-3.15082	-10.6154
-8.43404	40.1055
4.94918	-10.6154
1.85338	-0.9352
-4.13404	40.1055
-0.27179	-29.0166
0.52401	-26.6968

Table 7. Simple Regression on Age (Model 1)

Dependent Variable: steroid

Source	DF	Sum of		Mean Square	F Value	Pr > F
		Squares				
Model	1	793.28051		793.28051	40.35	<.0001
Error	25	491.52616		19.66105		
Corrected Total	26	1284.80667				
Root MSE		4.43408	R-Square	0.6174		
Dependent Mean		17.64444	Adj R-Sq	0.6021		
Coeff Var		25.13016				

Table 9. Regression of Residuals on Age Squared (like Model 3)

Dependent Variable: yresid Residual

Source	DF	Analysis of Variance		F Value	Pr > F
		Sum of	Mean		
Model	1	252.98535	252.98535	26.51	<.0001
Error	25	238.54081	9.54163		
Corrected Total	26	491.52616			
Root MSE		3.08895	R-Square	0.5147	
Dependent Mean		-3.8323E-15	Adj R-Sq	0.4953	
Coeff Var		-8.06026E16			

Source: esb10p25.sas 3/23/2010 by ejs

Compare to Model 3:

Table 6. Residuals from Regression on Age and Age2 (Model 3)

Dependent Variable: steroid

Source	DF	Analysis of Variance		F Value	Pr > F
		Sum of	Mean		
Model	2	1046.26586	523.13293	52.63	<.0001
Error	24	238.54081	9.93920		
Corrected Total	26	1284.80667			
Root MSE		3.15265	R-Square	0.8143	
Dependent Mean		17.64444	Adj R-Sq	0.7989	
Coeff Var		17.86766			

Variable	DF	Parameter Estimates				
		Parameter	Standard			
Intercept	1	-26.32541	5.88154	-4.48	0.0002	8405.81333
age	1	4.87357	0.77515	6.29	<.0001	793.28051
age2	1	-0.11840	0.02347	-5.05	<.0001	252.98535

Source: esb10p25.sas 3/23/2010 by ejs

Notes:

The accounting of the SS is the same using Tables 7 and 9, as in Table 6. The DF for the Error for Table 6 is correct- The DF in Table 9 does not account for having fit age in the model.

Correlation and Partial Correlation

```
PROC CORR DATA=d2;
  VAR steroid age age2;
  TITLE2 "Table 1. Correlation of Steroids with other variables";
RUN;
```

Table 1. Correlation of Steroids with other variables

Pearson Correlation Coefficients, N = 27

Prob > |r| under H0: Rho=0

	steroid	age	age2
steroid	1.00000	0.78577	0.71312
		<.0001	<.0001
age	0.78577	1.00000	0.98943
	<.0001		<.0001
age2	0.71312	0.98943	1.00000
	<.0001	<.0001	

Source: esb10p25.sas 3/23/2010 by ejs

Partial Correlation: Correlation of Residuals

```
PROC CORR DATA=d3;
  VAR yresid age2_resid;
  TITLE2 "Table 10. Correlation of Residuals of Y on age with age2 with age";
RUN;
```

Table 10. Correlation of Residuals of Y on age with age2 with age

2 Variables: yresid age2_resid

Pearson Correlation Coefficients, N = 27

Prob > |r| under H0: Rho=0

		age2_
	yresid	resid
yresid	1.00000	-0.71742
Residual		<.0001
age2_resid	-0.71742	1.00000
Residual	<.0001	

```

PROC CORR DATA=d2;
  PARTIAL age;
  VAR steroid age2;
  TITLE2 "Table 11. Partial correlation of age squared wih steroids";
  RUN;

```

Table 11. Partial correlation of age squared wih steroids

1 Partial Variables:	age
2 Variables:	steroid age2
Pearson Partial Correlation Coefficients, N = 27	
Prob > r under H0: Partial Rho=0	
	steroid age2
steroid	1.00000 -0.71742 <.0001
age2	-0.71742 1.00000 <.0001

Source: esb10p25.sas 3/23/2010 by ejs

Notes:

Partial correlations allow you to see what variable is most highly correlated after accounting for the previous variable.

Example 2: Body Fat (Y) and its relationship to

X1 triceps skinfold thickness
 X2 thigh circumference
 X3 midarm circumference

What is the best model for estimating Body Fat (CH07TA01) based on these other measures?

(Chapter 7 (p257) (Kutner, Nachtsheim, neter, and Li (5th edition: Applied Linear Statistical Models) (program esb10p26.sas))

7

Dummy-Variable Regression

One of the serious limitations of multiple-regression analysis, as presented in Chapters 5 and 6, is that it accommodates only quantitative response and explanatory variables. In this chapter and the next, I will explain how qualitative explanatory variables, called *factors*, can be incorporated into a linear model.¹

The current chapter begins with an explanation of how a *dummy-variable regressor* can be coded to represent a *dichotomous* (i.e., two-category) factor. I proceed to show how a set of dummy regressors can be employed to represent a *polytomous* (many-category) factor. I next describe how interactions between quantitative and qualitative explanatory variables can be represented in dummy-regression models and how to summarize models that incorporate interactions. Finally, I explain why it does not make sense to standardize dummy-variable and interaction regressors.

7.1 A Dichotomous Factor

Let us consider the simplest case: one dichotomous factor and one quantitative explanatory variable. As in the two previous chapters, assume that relationships are *additive*—that is, that the partial effect of each explanatory variable is the same regardless of the specific value at which the other explanatory variable is held constant. As well, suppose that the other assumptions of the regression model hold: The errors are independent and normally distributed, with zero means and constant variance.

The general motivation for including a factor in a regression is essentially the same as for including an additional quantitative explanatory variable: (1) to account more fully for the response variable, by making the errors smaller, and (2) even more important, to avoid a biased assessment of the impact of an explanatory variable, as a consequence of omitting another explanatory variable that is related to it.

For concreteness, suppose that we are interested in investigating the relationship between education and income among women and men. Figure 7.1(a) and (b) represents two small (idealized) populations. In both cases, the within-gender regressions of income on education are parallel. Parallel regressions imply additive effects of education and gender on income: Holding education constant, the “effect” of gender is the vertical distance between the two regression lines, which—for parallel lines—is everywhere the same. Likewise, holding gender constant, the “effect” of education is captured by the within-gender education slope, which—for parallel lines—is the same for men and women.²

In Figure 7.1(a), the explanatory variables gender and education are unrelated to each other: Women and men have identical distributions of education scores (as can be seen by projecting the points onto the horizontal axis). In this circumstance, if we ignore gender and regress income on education alone, we obtain the same slope as is produced by the separate within-gender

¹Chapter 14 deals with qualitative *response* variables.

²I will consider nonparallel within-group regressions in Section 7.3.

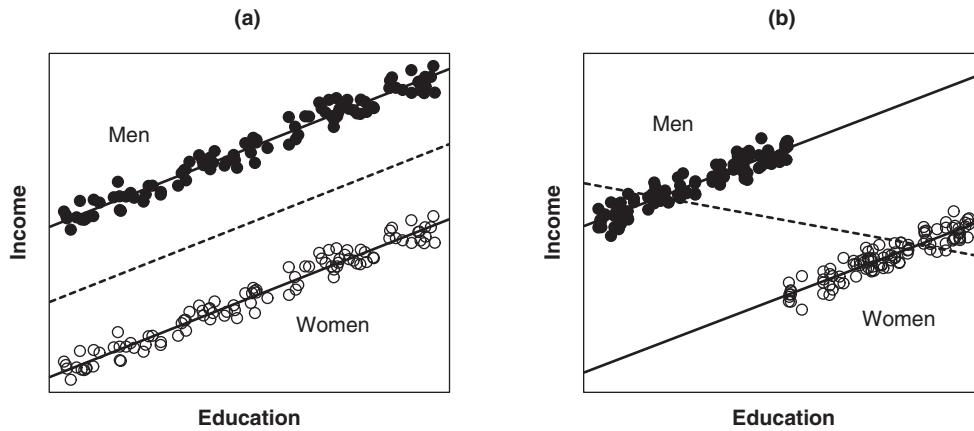


Figure 7.1 Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both (a) and (b), the within-gender (i.e., partial) regressions (solid lines) are parallel. In each graph, the overall (i.e., marginal) regression of income on education (ignoring gender) is given by the broken line.

regressions. Because women have lower incomes than men of equal education, however, by ignoring gender we inflate the size of the errors.

The situation depicted in Figure 7.1(b) is importantly different. Here, gender and education are related, and therefore if we regress income on education alone, we arrive at a biased assessment of the effect of education on income: Because women have a higher average level of education than men, and because—for a given level of education—women's incomes are lower, on average, than men's, the overall regression of income on education has a *negative* slope even though the within-gender regressions have a *positive* slope.³

In light of these considerations, we might proceed to partition our sample by gender and perform separate regressions for women and men. This approach is reasonable, but it has its limitations: Fitting separate regressions makes it difficult to estimate and test for gender differences in income. Furthermore, if we can reasonably assume parallel regressions for women and men, we can more efficiently estimate the common education slope by pooling sample data drawn from both groups. In particular, if the usual assumptions of the regression model hold, then it is desirable to fit the common-slope model by least squares.

One way of formulating the common-slope model is

$$Y_i = \alpha + \beta X_i + \gamma D_i + \varepsilon_i \quad (7.1)$$

where D , called a *dummy-variable regressor* or an *indicator variable*, is coded 1 for men and 0 for women:

$$D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}$$

³That marginal and partial relationships can differ in sign is called *Simpson's paradox* (Simpson, 1951). Here, the marginal relationship between income and education is negative, while the partial relationship, controlling for gender, is positive.

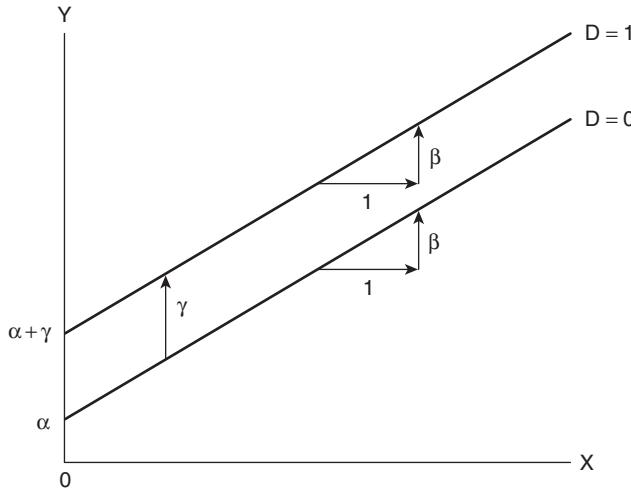


Figure 7.2 The additive dummy-variable regression model. The line labeled $D = 1$ is for men; the line labeled $D = 0$ is for women.

Thus, for women the model becomes

$$Y_i = \alpha + \beta X_i + \gamma(0) + \varepsilon_i = \alpha + \beta X_i + \varepsilon_i$$

and for men

$$Y_i = \alpha + \beta X_i + \gamma(1) + \varepsilon_i = (\alpha + \gamma) + \beta X_i + \varepsilon_i$$

These regression equations are graphed in Figure 7.2.

This is our initial encounter with an idea that is fundamental to many linear models: the distinction between *explanatory variables* and *regressors*. Here, *gender* is a qualitative explanatory variable (i.e., a factor), with categories *male* and *female*. The dummy variable D is a regressor, representing the factor gender. In contrast, the quantitative explanatory variable *education* and the regressor X are one and the same. Were we to transform education, however, prior to entering it into the regression equation—say, by taking logs—then there would be a distinction between the explanatory variable (*education*) and the regressor (*log education*). In subsequent sections of this chapter, it will transpire that an explanatory variable can give rise to several regressors and that some regressors are functions of more than one explanatory variable.

Returning to Equation 7.1 and Figure 7.2, the coefficient γ for the dummy regressor gives the difference in intercepts for the two regression lines. Moreover, because the within-gender regression lines are parallel, γ also represents the constant vertical separation between the lines, and it may, therefore, be interpreted as the expected income advantage accruing to men when education is held constant. If men were *disadvantaged* relative to women with the same level of education, then γ would be *negative*. The coefficient α gives the intercept for women, for whom $D = 0$; and β is the common within-gender education slope.

Figure 7.3 reveals the fundamental geometric “trick” underlying the coding of a dummy regressor: We are, in fact, fitting a regression plane to the data, but the dummy regressor D is defined only at the values 0 and 1. The regression plane intersects the planes $\{X, Y|D = 0\}$ and $\{X, Y|D = 1\}$ in two lines, each with slope β . Because the difference between $D = 0$ and $D = 1$ is one unit, the difference in the Y -intercepts of these two lines is the slope of the plane in the D direction,

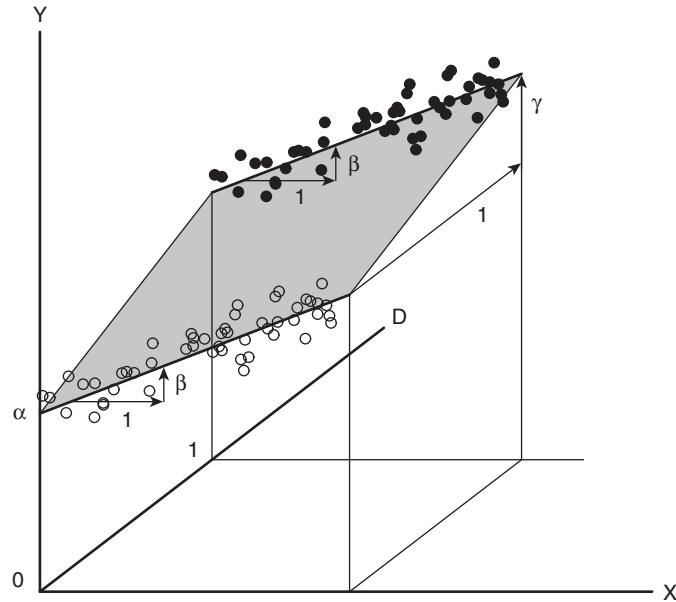


Figure 7.3 The geometric “trick” underlying dummy regression: The linear regression plane is defined only at $D = 0$ and $D = 1$, producing two regression lines with slope β and vertical separation γ . The hollow circles represent women, for whom $D = 0$, and the solid circles men, for whom $D = 1$.

that is γ . Indeed, Figure 7.2 is simply the projection of the two regression lines onto the $\{X, Y\}$ plane.

Essentially similar results are obtained if we instead code D equal to 0 for men and 1 for women, making men the *baseline* (or *reference*) category (see Figure 7.4): The *sign* of γ is reversed, because it now represents the difference in intercepts between women and men (rather than vice versa), but its *magnitude* remains the same. The coefficient α now gives the income intercept for men. It is therefore immaterial which group is coded 1 and which is coded 0, as long as we are careful to interpret the coefficients of the model—for example, the sign of γ —in a manner consistent with the coding scheme that is employed.

To determine whether gender affects income, controlling for education, we can test $H_0: \gamma = 0$, either by a t -test, dividing the estimate of γ by its standard error, or, equivalently, by dropping D from the regression model and formulating an incremental F -test. In either event, the statistical-inference procedures of the previous chapter apply.

Although I have developed dummy-variable regression for a single quantitative regressor, the method can be applied to any number of quantitative explanatory variables, as long as we are willing to assume that the slopes are the same in the two categories of the factor—that is, that the regression surfaces are parallel in the two groups. In general, if we fit the model

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \gamma D_i + \varepsilon_i$$

then, for $D = 0$, we have

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

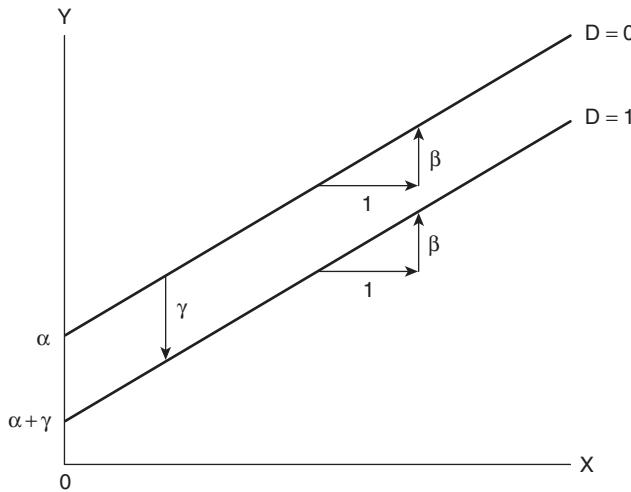


Figure 7.4 The additive dummy-regression model coding $D = 0$ for men and $D = 1$ for women (cf., Figure 7.2).

and, for $D = 1$,

$$Y_i = (\alpha + \gamma) + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

A dichotomous factor can be entered into a regression equation by formulating a dummy regressor, coded 1 for one category of the factor and 0 for the other category. A model incorporating a dummy regressor represents parallel regression surfaces, with the constant vertical separation between the surfaces given by the coefficient of the dummy regressor.

7.2 Polytomous Factors

The coding method of the previous section generalizes straightforwardly to polytomous factors. By way of illustration, recall (from the previous chapter) the Canadian occupational prestige data. I have classified the occupations into three rough categories: (1) professional and managerial occupations, (2) “white-collar” occupations, and (3) “blue-collar” occupations.⁴

Figure 7.5 shows conditioning plots for the relationship between prestige and each of income and education within occupational types.⁵ The partial relationships between prestige and the explanatory variables appear reasonably linear, although there seems to be evidence that the income slope varies across the categories of type of occupation (a possibility that I will pursue in the next section of the chapter). Indeed, this change in slope is an explanation of the nonlinearity in the relationship between prestige and income that we noticed in Chapter 4. These conditioning

⁴Although there are 102 occupations in the full data set, several are difficult to classify and consequently were dropped from the analysis. The omitted occupations are athletes, babysitters, farmers, and “newsboys,” leaving us with 98 observations.

⁵In the preceding chapter, I also included the gender composition of the occupations as an explanatory variable, but I omit that variable here. Conditioning plots are described in Section 3.3.4.

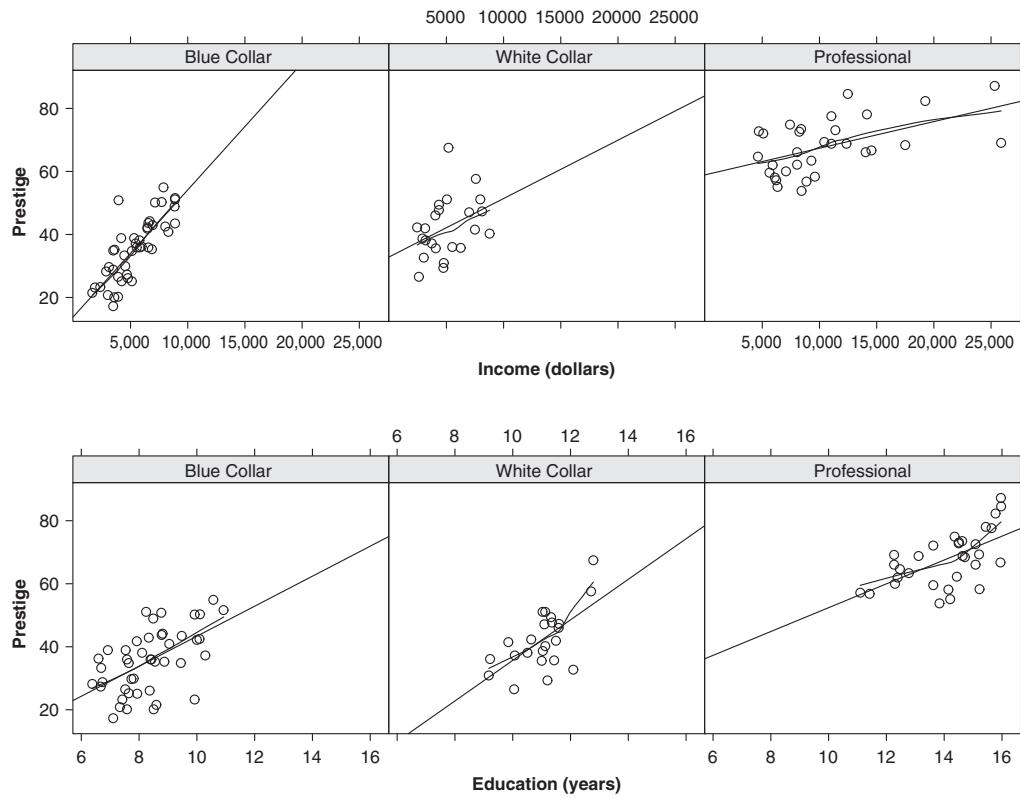


Figure 7.5 Conditioning plots for the relationships between prestige and each of income (top panel) and education (bottom panel) by type of occupation, for the Canadian occupational prestige data. Each panel shows the linear least-squares fit and a lowess smooth with a span of 0.9. The graphs labeled “Professional” are for professional and managerial occupations.

plots do not tell the whole story, however, because the income and education levels of the occupations are correlated, but they give us a reasonable initial look at the data. Conditioning the plot for income by level of education (and vice versa) is out of the question here because of the small size of the data set.

The *three*-category occupational-type factor can be represented in the regression equation by introducing *two* dummy regressors, employing the following coding scheme:

Category	D_1	D_2	
Professional and managerial	1	0	(7.2)
White collar	0	1	
Blue collar	0	0	

A model for the regression of prestige on income, education, and type of occupation is then

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \varepsilon_i \quad (7.3)$$

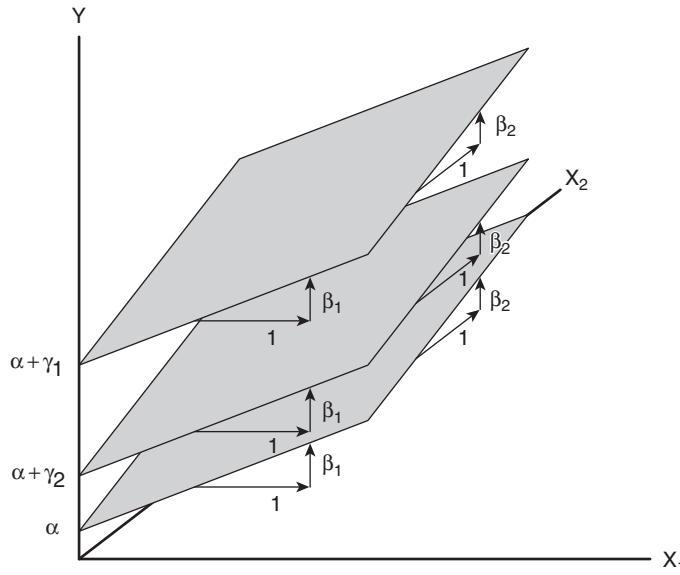


Figure 7.6 The additive dummy-regression model with two quantitative explanatory variables X_1 and X_2 represents parallel planes with potentially different intercepts in the $\{X_1, X_2, Y\}$ space.

where X_1 is income and X_2 is education. This model describes three parallel regression planes, which can differ in their intercepts:

$$\begin{aligned} \text{Professional: } & Y_i = (\alpha + \gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\ \text{White collar: } & Y_i = (\alpha + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\ \text{Blue collar: } & Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \end{aligned}$$

The coefficient α , therefore, gives the intercept for blue-collar occupations; γ_1 represents the constant vertical difference between the parallel regression planes for professional and blue-collar occupations (fixing the values of education and income); and γ_2 represents the constant vertical distance between the regression planes for white-collar and blue-collar occupations (again, fixing education and income). Assuming, for simplicity, that all coefficients are positive, and that $\gamma_1 > \gamma_2$, the geometry of the model in Equation 7.3 is illustrated in Figure 7.6.

Because blue-collar occupations are coded 0 for both dummy regressors, “blue collar” implicitly serves as the baseline category to which the other occupational-type categories are compared. The choice of a baseline category is essentially arbitrary, for we would fit precisely the same three regression planes regardless of which of the three occupational-type categories is selected for this role. The values (and meaning) of the individual dummy-variable coefficients γ_1 and γ_2 depend, however, on which category is chosen as the baseline.

It is sometimes natural to select a particular category as a basis for comparison—an experiment that includes a “control group” comes immediately to mind. In this instance, the individual dummy-variable coefficients are of interest, because they reflect differences between the “experimental” groups and the control group, holding other explanatory variables constant.

In most applications, however, the choice of a baseline category is entirely arbitrary, as it is for the occupational prestige regression. We are, therefore, most interested in testing the null hypothesis of no effect of occupational type, controlling for education and income,

$$H_0: \gamma_1 = \gamma_2 = 0 \tag{7.4}$$

but the individual hypotheses $H_0: \gamma_1 = 0$ and $H_0: \gamma_2 = 0$ —which test, respectively, for differences between professional and blue-collar occupations and between white-collar and blue-collar occupations—are of less intrinsic interest.⁶ The null hypothesis in Equation 7.4 can be tested by the incremental-sum-of-squares approach, dropping the two dummy variables for type of occupation from the model.

I have demonstrated how to model the effects of a three-category factor by coding two dummy regressors. It may seem more natural to treat the three occupational categories symmetrically, coding *three* dummy regressors, rather than arbitrarily selecting one category as the baseline:

Category	D_1	D_2	D_3	
Professional and managerial	1	0	0	(7.5)
White collar	0	1	0	
Blue collar	0	0	1	

Then, for the j th occupational type, we would have

$$Y_i = (\alpha + \gamma_j) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

The problem with this procedure is that there are too many parameters: We have used four parameters ($\alpha, \gamma_1, \gamma_2, \gamma_3$) to represent only three group intercepts. As a consequence, we could not find unique values for these four parameters even if we knew the three population regression lines. Likewise, we cannot calculate unique least-squares estimates for the model because the set of three dummy variables is perfectly collinear; for example, as is apparent from the table in Equation 7.5, $D_3 = 1 - D_1 - D_2$.

In general, then, for a polytomous factor with m categories, we need to code $m - 1$ dummy regressors. One simple scheme is to select the last category as the baseline and to code $D_{ij} = 1$ when observation i falls in category j , and 0 otherwise:

Category	D_1	D_2	...	D_{m-1}	
1	1	0	...	0	
2	0	1	...	0	(7.6)
:	:	:		:	
$m-1$	0	0	...	1	
m	0	0	...	0	

A polytomous factor can be entered into a regression by coding a set of 0/1 dummy regressors, one fewer than the number of categories of the factor. The “omitted” category, coded 0 for all dummy regressors in the set, serves as a baseline to which the other categories are compared. The model represents parallel regression surfaces, one for each category of the factor.

⁶The essential point here is not that the separate hypotheses are of *no* interest but that they are an arbitrary subset of the pairwise differences among the categories. In the present case, where there are three categories, the individual hypotheses represent two of the three pairwise group comparisons. The third comparison, between professional and white-collar occupations, is not *directly* represented in the model, although it is given indirectly by the difference $\gamma_1 - \gamma_2$. See Section 7.2.1 for an elaboration of this point.

When there is more than one factor, and if we assume that the factors have additive effects, we can simply code a set of dummy regressors for each. To test the hypothesis that the effect of a factor is nil, we delete its dummy regressors from the model and compute an incremental F -test of the hypothesis that all the associated coefficients are 0.

Regressing occupational prestige (Y) on income (X_1) and education (X_2) produces the fitted regression equation

$$\hat{Y} = -7.621 + 0.001241X_1 + 4.292X_2 \quad R^2 = .81400$$

$$(3.116) \quad (0.000219) \quad (0.336)$$

As is common practice, I have shown the estimated standard error of each regression coefficient in parentheses beneath the coefficient. The three occupational categories differ considerably in their average levels of prestige:

Category	Number of Cases	Mean Prestige
Professional and managerial	31	67.85
White collar	23	42.24
Blue collar	44	35.53
All occupations	98	47.33

Inserting dummy variables for type of occupation into the regression equation, employing the coding scheme shown in Equation 7.2, produces the following results:

$$\hat{Y} = -0.6229 + 0.001013X_1 + 3.673X_2 + 6.039D_1 - 2.737D_2$$

$$(5.2275) \quad (0.000221) \quad (0.641) \quad (3.867) \quad (2.514)$$

$$R^2 = .83486 \quad (7.7)$$

The three fitted regression equations are, therefore,

$$\begin{aligned} \text{Professional: } \hat{Y} &= 5.416 + 0.001013X_1 + 3.673X_2 \\ \text{White collar: } \hat{Y} &= -3.360 + 0.001013X_1 + 3.673X_2 \\ \text{Blue collar: } \hat{Y} &= -0.623 + 0.001013X_1 + 3.673X_2 \end{aligned}$$

Note that the coefficients for both income and education become slightly smaller when type of occupation is controlled. As well, the dummy-variable coefficients (or, equivalently, the category intercepts) reveal that when education and income levels are held constant statistically, the difference in average prestige between professional and blue-collar occupations declines greatly, from $67.85 - 35.53 = 32.32$ points to 6.04 points. The difference between white-collar and blue-collar occupations is reversed when income and education are held constant, changing from $42.24 - 35.53 = +6.71$ points to -2.74 points. That is, the greater prestige of professional occupations compared with blue-collar occupations appears to be due mostly to differences in education and income between these two classes of occupations. While white-collar occupations have greater prestige, on average, than blue-collar occupations, they have lower prestige than blue-collar occupations of the same educational and income levels.⁷

To test the null hypothesis of no partial effect of type of occupation,

$$H_0: \gamma_1 = \gamma_2 = 0$$

⁷These conclusions presuppose that the additive model that we have fit to the data is adequate, which, as we will see in Section 7.3.5, is not the case.

we can calculate the incremental F -statistic

$$\begin{aligned} F_0 &= \frac{n - k - 1}{q} \times \frac{R_1^2 - R_0^2}{1 - R_1^2} \\ &= \frac{98 - 4 - 1}{2} \times \frac{.83486 - .81400}{1 - .83486} = 5.874 \end{aligned} \quad (7.8)$$

with 2 and 93 degrees of freedom, for which $p = .0040$. The occupational-type effect is therefore statistically significant but (examining the coefficient standard errors) not very precisely estimated. The education and income coefficients are several times their respective standard errors, and hence are highly statistically significant.

7.2.1 Coefficient Quasi-Variances*

Consider a dummy-regression model with p quantitative explanatory variables and an m -category factor:

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \cdots + \gamma_{m-1} D_{i,m-1} + \varepsilon_i$$

The dummy-variable coefficients $\gamma_1, \gamma_2, \dots, \gamma_{m-1}$ represent differences (or *contrasts*) between each of the other categories of the factor and the reference category m , holding constant X_1, \dots, X_p . If we are interested in a comparison between any other two categories, we can simply take the difference in their dummy-regressor coefficients. Thus, in the preceding example (letting $C_1 \equiv \hat{\gamma}_1$ and $C_2 \equiv \hat{\gamma}_2$),

$$C_1 - C_2 = 5.416 - (-3.360) = 8.776$$

is the estimated average difference in prestige between professional and white-collar occupations of equal income and education.

Suppose, however, that we want to know the standard error of $C_1 - C_2$. The standard errors of C_1 and C_2 are available directly in the regression “output” (Equation 7.7), but to compute the standard error of $C_1 - C_2$, we need in addition the estimated sampling covariance of these two coefficients. That is,⁸

$$\text{SE}(C_1 - C_2) = \sqrt{\hat{V}(C_1) + \hat{V}(C_2) - 2 \times \hat{C}(C_1, C_2)}$$

where $\hat{V}(C_j) = [\text{SE}(C_j)]^2$ is the estimated sampling variance of coefficient C_j , and $\hat{C}(C_1, C_2)$ is the estimated sampling covariance of C_1 and C_2 . For the occupational prestige regression, $\hat{C}(C_1, C_2) = 6.797$, and so

$$\text{SE}(C_1 - C_2) = \sqrt{3.867^2 + 2.514^2 - 2 \times 6.797} = 2.771$$

We can use this standard error in the normal manner for a t -test of the difference between C_1 and C_2 .⁹ For example, noting that the difference exceeds twice its standard error suggests that it is statistically significant.

⁸See Appendix D on probability and estimation. The computation of regression-coefficient covariances is taken up in Chapter 9.

⁹Testing all differences between pairs of factor categories raises an issue of simultaneous inference, however. See the discussion of Scheffé confidence intervals in Section 9.4.4.

Although computer programs for regression analysis typically report the covariance matrix of the regression coefficients if asked to do so, it is not common to include coefficient covariances in published research along with estimated coefficients and standard errors, because with $k + 1$ coefficients in the model, there are $k(k + 1)/2$ variances and covariances among them—a potentially large number. Readers of a research report are therefore put at a disadvantage by the arbitrary choice of a reference category in dummy regression, because they are unable to calculate the standard errors of the differences between all pairs of categories of a factor.

Quasi-variances of dummy-regression coefficients (Firth, 2003; Firth & De Menezes, 2004) speak to this problem. Let $\tilde{V}(C_j)$ denote the quasi-variance of dummy coefficient C_j . Then,

$$\text{SE}(C_j - C_{j'}) \approx \sqrt{\tilde{V}(C_j) + \tilde{V}(C_{j'})}$$

The squared relative error of this approximation for the contrast $C_j - C_{j'}$ is

$$\text{RE}_{jj'} \equiv \frac{\tilde{V}(C_j - C_{j'})}{\widehat{V}(C_j - C_{j'})} = \frac{\tilde{V}(C_j) + \tilde{V}(C_{j'})}{\widehat{V}(C_j) + \widehat{V}(C_{j'}) - 2 \times \widehat{C}(C_j, C_{j'})}$$

The approximation is accurate for this contrast when $\text{RE}_{jj'}$ is close to 1, or, equivalently, when

$$\log(\text{RE}_{jj'}) = \log[\tilde{V}(C_j) + \tilde{V}(C_{j'})] - \log[\widehat{V}(C_j) + \widehat{V}(C_{j'}) - 2 \times \widehat{C}(C_j, C_{j'})]$$

is close to 0. The quasi-variances $\tilde{V}(C_j)$ are therefore selected to minimize the sum of squared log relative errors of approximation over all pairwise contrasts, $\sum_{j < j'} [\log(\text{RE}_{jj'})]^2$. The resulting errors of approximation are typically very small (Firth, 2003; Firth & De Menezes, 2004).

The following table gives dummy-variable coefficients, standard errors, and quasi-variances for type of occupation in the Canadian occupational prestige regression:

Category	C_j	$\text{SE}(C_j)$	$\tilde{V}(C_j)$
Professional	6.039	3.867	8.155
White collar	-2.737	2.514	-0.4772
Blue collar	0	0	6.797

I have set to 0 the coefficient (and its standard error) for the baseline category, blue collar. The negative quasi-variance for the white-collar coefficient is at first blush disconcerting (after all, ordinary variances cannot be negative), but it is not wrong: The quasi-variances are computed to provide accurate variance approximations for coefficient *differences*; they do not apply directly to the coefficients themselves. For the contrast between professional and white-collar occupations, we have

$$\text{SE}(C_1 - C_2) \approx \sqrt{8.155 - 0.4772} = 2.771$$

Likewise, for the contrast between professional and blue-collar occupations,

$$C_1 - C_3 = 6.039 - 0 = 6.039$$

$$\text{SE}(C_1 - C_3) \approx \sqrt{8.155 + 6.797} = 3.867$$

Note that in this application, the quasi-variance “approximation” to the standard error proves to be exact, and indeed this is necessarily the case when there are just three factor categories, because there are then just three pairwise differences among the categories to capture.¹⁰

¹⁰For the details of the computation of quasi-variances, see Chapter 15, Exercise 15.11.

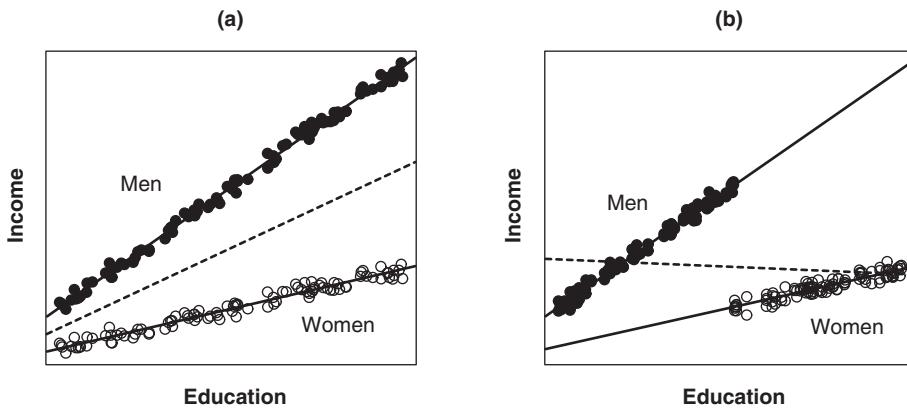


Figure 7.7 Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both cases, the within-gender regressions (solid lines) are not parallel—the slope for men is greater than the slope for women—and, consequently, education and gender interact in affecting income. In each graph, the overall regression of income on education (ignoring gender) is given by the broken line.

7.3 Modeling Interactions

Two explanatory variables are said to *interact* in determining a response variable when the partial effect of one depends on the value of the other. The additive models that we have considered thus far therefore specify the *absence* of interactions. In this section, I will explain how the dummy-variable regression model can be modified to accommodate interactions between factors and quantitative explanatory variables.¹¹

The treatment of dummy-variable regression in the preceding two sections has assumed parallel regressions across the several categories of a factor. If these regressions are *not* parallel, then the factor interacts with one or more of the quantitative explanatory variables. The dummy-regression model can easily be modified to reflect these interactions.

For simplicity, I return to the contrived example of Section 7.1, examining the regression of income on gender and education. Consider the hypothetical data shown in Figure 7.7 (and contrast these examples with those shown in Figure 7.1 on page 121, where the effects of gender and education are additive). In Figure 7.7(a) [as in Figure 7.1(a)], gender and education are independent, because women and men have identical education distributions; in Figure 7.7(b) [as in Figure 7.1(b)], gender and education are related, because women, on average, have higher levels of education than men.

It is apparent in both Figure 7.7(a) and Figure 7.7(b), however, that the within-gender regressions of income on education are not parallel: In both cases, the slope for men is larger than the slope for women. Because the effect of education varies by gender, education and gender interact in affecting income.

It is also the case, incidentally, that the effect of gender varies by education. Because the regressions are not parallel, the relative income advantage of men changes (indeed, grows) with

¹¹Interactions between factors are taken up in the next chapter on analysis of variance; interactions between quantitative explanatory variables are discussed in Section 17.1 on polynomial regression.

education. Interaction, then, is a symmetric concept—that the effect of education varies by gender implies that the effect of gender varies by education (and, of course, vice versa).

The simple examples in Figures 7.1 and 7.7 illustrate an important and frequently misunderstood point: *Interaction* and *correlation* of explanatory variables are empirically and logically distinct phenomena. Two explanatory variables can interact *whether or not* they are related to one another statistically. Interaction refers to the manner in which explanatory variables *combine* to affect a response variable, not to the relationship *between* the explanatory variables themselves.

Interaction and correlation of explanatory variables are empirically and logically distinct phenomena. Two explanatory variables can interact whether or not they are related to one another statistically. Interaction refers to the manner in which explanatory variables combine to affect a response variable, not to the relationship between the explanatory variables themselves.

7.3.1 Constructing Interaction Regressors

We could model the data in Figure 7.7 by fitting separate regressions of income on education for women and men. As before, however, it is more convenient to fit a combined model, primarily because a combined model facilitates a test of the gender-by-education interaction. Moreover, a properly formulated unified model that permits different intercepts and slopes in the two groups produces the same fit to the data as separate regressions: The full sample is composed of the two groups, and, consequently, the residual sum of squares for the full sample is minimized when the residual sum of squares is minimized in each group.¹²

The following model accommodates different intercepts and slopes for women and men:

$$Y_i = \alpha + \beta X_i + \gamma D_i + \delta(X_i D_i) + \varepsilon_i \quad (7.9)$$

Along with the quantitative regressor X for education and the dummy regressor D for gender, I have introduced the *interaction regressor* XD into the regression equation. The interaction regressor is the *product* of the other two regressors; although XD is therefore a function of X and D , it is not a *linear* function, and perfect collinearity is avoided.¹³

For women, model (7.9) becomes

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma(0) + \delta(X_i \cdot 0) + \varepsilon_i \\ &= \alpha + \beta X_i + \varepsilon_i \end{aligned}$$

and for men

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma(1) + \delta(X_i \cdot 1) + \varepsilon_i \\ &= (\alpha + \gamma) + (\beta + \delta)X_i + \varepsilon_i \end{aligned}$$

¹²See Exercise 7.4.

¹³If this procedure seems illegitimate, then think of the interaction regressor as a new variable, say $Z \equiv XD$. The model is linear in X , D , and Z . The “trick” of introducing an interaction regressor is similar to the trick of formulating dummy regressors to capture the effect of a factor: In both cases, there is a distinction between explanatory variables and regressors. Unlike a dummy regressor, however, the interaction regressor is a function of *both* explanatory variables.

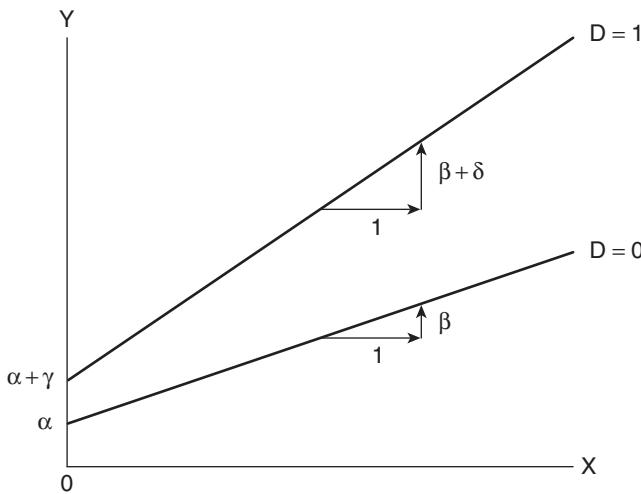


Figure 7.8 The dummy-variable regression model with an interaction regressor. The line labeled $D = 1$ is for men; the line labeled $D = 0$ is for women.

These regression equations are graphed in Figure 7.8: The parameters α and β are, respectively, the intercept and slope for the regression of income on education among women (the baseline category for gender); γ gives the difference in intercepts between the male and female groups; and δ gives the difference in slopes between the two groups. To test for interaction, therefore, we may simply test the hypothesis $H_0: \delta = 0$.

Interactions can be incorporated by coding interaction regressors, taking products of dummy regressors with quantitative explanatory variables. The resulting model permits different slopes in different groups—that is, regression surfaces that are not parallel.

In the additive, no-interaction model of Equation 7.1 and Figure 7.2, the dummy-regressor coefficient γ represents the *unique* partial effect of gender (i.e., the expected income difference between men and women of equal education, regardless of the value at which education is fixed), while the slope β represents the *unique* partial effect of education (i.e., the within-gender expected increment in income for a one-unit increase in education, for both women and men). In the interaction model of Equation 7.9 and Figure 7.8, in contrast, γ is no longer interpretable as the unqualified income difference between men and women of equal education.

Because the within-gender regressions are not parallel, the separation between the regression lines changes; here, γ is simply the separation at $X = 0$ —that is, above the origin. It is generally no more important to assess the expected income difference between men and women of 0 education than at other educational levels, and therefore the difference-in-intercepts parameter γ is not of special interest in the interaction model. Indeed, in many instances (although not here), the value $X = 0$ may not occur in the data or may be impossible (as, for example, if X is weight). In such cases, γ has no literal interpretation in the interaction model (see Figure 7.9).

Likewise, in the interaction model, β is not the unqualified partial effect of education, but rather the effect of education among women. Although this coefficient *is* of interest, it is not necessarily

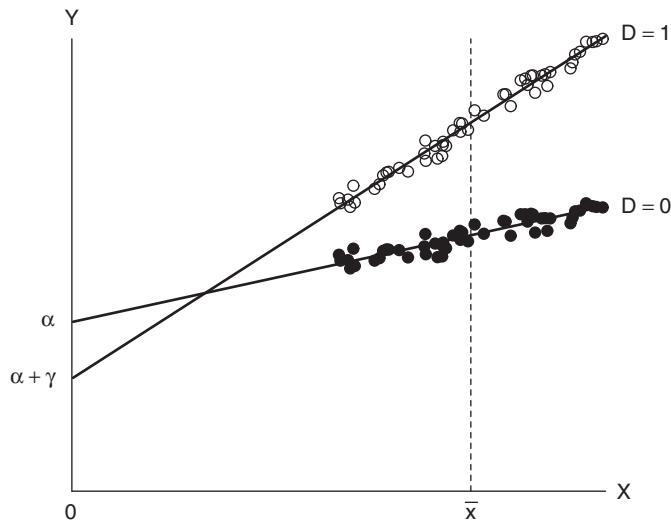


Figure 7.9 Why the difference in intercepts does not represent a meaningful partial effect for a factor when there is interaction: The difference-in-intercepts parameter γ is *negative* even though, within the range of the data, the regression line for the group coded $D = 1$ is *above* the line for the group coded $D = 0$.

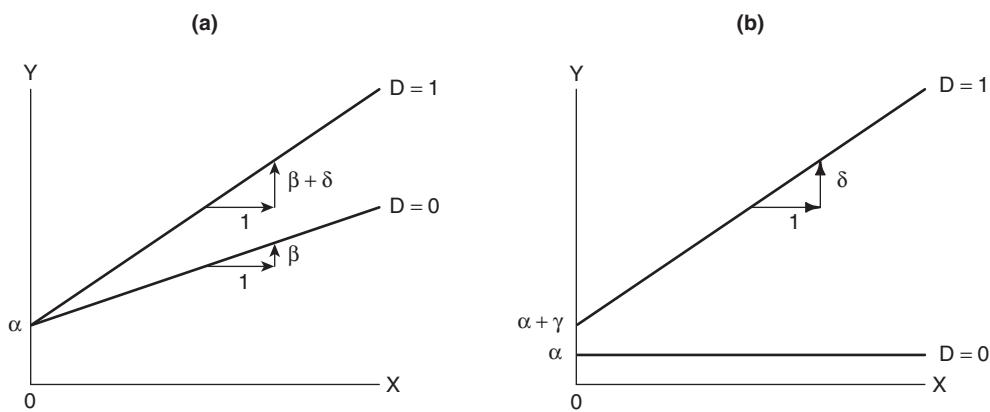


Figure 7.10 Two models that violate the principle of marginality: In (a), the dummy regressor D is omitted from the model $E(Y) = \alpha + \beta X + \delta(XD)$; in (b), the quantitative explanatory variable X is omitted from the model $E(Y) = \alpha + \gamma D + \delta(XD)$. These models violate the principle of marginality because they include the term XD , which is a higher-order relative of both X and D (one of which is omitted from each model).

more important than the effect of education among men ($\beta + \delta$), which does not appear *directly* in the model.

7.3.2 The Principle of Marginality

Following Nelder (1977), we say that the separate partial effects, or *main effects*, of education and gender are *marginal* to the education-by-gender interaction. In general, we neither test nor interpret the main effects of explanatory variables that interact. If, however, we can rule out interaction either on theoretical or on empirical grounds, then we can proceed to test, estimate, and interpret the main effects.

As a corollary to this principle, it does not generally make sense to specify and fit models that include interaction regressors but that omit main effects that are marginal to them. This is not to say that such models—which violate the *principle of marginality*—are uninterpretable: They are, rather, not broadly applicable.

The principle of marginality specifies that a model including a *high-order term* (such as an interaction) should normally also include the “lower-order relatives” of that term (the main effects that “compose” the interaction).

Suppose, for example, that we fit the model

$$Y_i = \alpha + \beta X_i + \delta(X_i D_i) + \varepsilon_i$$

which omits the dummy regressor D , but includes its “higher-order relative” XD . As shown in Figure 7.10(a), this model describes regression lines for women and men that have the same intercept but (potentially) different slopes, a specification that is peculiar and of no substantive interest. Similarly, the model

$$Y_i = \alpha + \gamma D_i + \delta(X_i D_i) + \varepsilon_i$$

graphed in Figure 7.10(b), constrains the slope for women to 0, which is needlessly restrictive.

7.3.3 Interactions With Polytomous Factors

The method for modeling interactions by forming product regressors is easily extended to polytomous factors, to several factors, and to several quantitative explanatory variables. I will use the Canadian occupational prestige regression to illustrate the application of the method, entertaining the possibility that occupational type interacts both with income (X_1) and with education (X_2):

$$\begin{aligned} Y_i = & \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} \\ & + \delta_{11} X_{i1} D_{i1} + \delta_{12} X_{i1} D_{i2} + \delta_{21} X_{i2} D_{i1} + \delta_{22} X_{i2} D_{i2} + \varepsilon_i \end{aligned} \quad (7.10)$$

Note that we require one interaction regressor for each product of a dummy regressor with a quantitative explanatory variable. The regressors $X_1 D_1$ and $X_1 D_2$ capture the interaction between income and occupational type; $X_2 D_1$ and $X_2 D_2$ capture the interaction between education and

occupational type. The model therefore permits different intercepts and slopes for the three types of occupations:

$$\begin{aligned}\text{Professional: } Y_i &= (\alpha + \gamma_1) + (\beta_1 + \delta_{11})X_{i1} + (\beta_2 + \delta_{21})X_{i2} + \varepsilon_i \\ \text{White collar: } Y_i &= (\alpha + \gamma_2) + (\beta_1 + \delta_{12})X_{i1} + (\beta_2 + \delta_{22})X_{i2} + \varepsilon_i \\ \text{Blue collar: } Y_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i\end{aligned}\quad (7.11)$$

Blue-collar occupations, which are coded 0 for both dummy regressors, serve as the baseline for the intercepts and slopes of the other occupational types. As in the no-interaction model, the choice of baseline category is generally arbitrary, as it is here, and is inconsequential. Fitting the model in Equation 7.10 to the prestige data produces the following results:

$$\begin{aligned}\widehat{Y}_i &= 2.276 + 0.003522X_1 + 1.713X_2 + 15.35D_1 - 33.54D_2 \\ &\quad (7.057) \quad (0.000556) \quad (0.927) \quad (13.72) \quad (17.54) \\ &\quad - 0.002903X_1D_1 - 0.002072X_1D_2 \\ &\quad (0.000599) \quad (0.000894) \\ &\quad + 1.388X_2D_1 + 4.291X_2D_2 \\ &\quad (1.289) \quad (1.757) \\ R^2 &= .8747\end{aligned}\quad (7.12)$$

This example is discussed further in the following section.

7.3.4 Interpreting Dummy-Regression Models With Interactions

It is difficult in dummy-regression models with interactions (and in other complex statistical models) to understand what the model is saying about the data simply by examining the regression coefficients. One approach to interpretation, which works reasonably well in a relatively straightforward model such as Equation 7.12, is to write out the implied regression equation for each group (using Equation 7.11):

$$\begin{aligned}\text{Professional: } \widehat{\text{Prestige}} &= 17.63 + 0.000619 \times \text{Income} + 3.101 \times \text{Education} \\ \text{White collar: } \widehat{\text{Prestige}} &= -31.26 + 0.001450 \times \text{Income} + 6.004 \times \text{Education} \\ \text{Blue collar: } \widehat{\text{Prestige}} &= 2.276 + 0.003522 \times \text{Income} + 1.713 \times \text{Education}\end{aligned}\quad (7.13)$$

From these equations, we can see, for example, that income appears to make much more difference to prestige in blue-collar occupations than in white-collar occupations, and has even less impact on prestige in professional and managerial occupations. Education, in contrast, has the largest impact on prestige among white-collar occupations, and has the smallest effect in blue-collar occupations.

An alternative approach (from Fox, 1987, 2003; Fox & Andersen, 2006) that generalizes readily to more complex models is to examine the high-order terms of the model. In the illustration, the high-order terms are the interactions between income and type and between education and type.

- Focusing in turn on each high-order term, we allow the variables in the term to range over their combinations of values in the data, fixing other variables to typical values. For example, for the interaction between type and income, we let type of occupation take on successively the categories blue collar, white collar, and professional (for which the dummy regressors

D_1 and D_2 are set to the corresponding values given in Equation 7.6), in combination with income values between \$1500 and \$26,000 (the approximate range of income in the Canadian occupational prestige data set); education is fixed to its average value in the data, $\bar{X}_2 = 10.79$.

- We next compute the fitted value of prestige at each combination of values of income and type of occupation. These fitted values are graphed in the “effect display” shown in the upper panel of Figure 7.11; the lower panel of this figure shows a similar effect display for the interaction between education and type of occupation, holding income at its average value. The broken lines in Figure 7.11 give ± 2 standard errors around the fitted values—that is, approximate 95% pointwise confidence intervals for the effects.¹⁴ The nature of the interactions between income and type and between education and type is readily discerned from these graphs.

7.3.5 Hypothesis Tests for Main Effects and Interactions

To test the null hypothesis of no interaction between income and type, $H_0: \delta_{11} = \delta_{12} = 0$, we need to delete the interaction regressors $X_1 D_1$ and $X_1 D_2$ from the full model (Equation 7.10) and calculate an incremental F -test; likewise, to test the null hypothesis of no interaction between education and type, $H_0: \delta_{21} = \delta_{22} = 0$, we delete the interaction regressors $X_2 D_1$ and $X_2 D_2$ from the full model. These tests, and tests for the main effects of income, education, and occupational type, are detailed in Tables 7.1 and 7.2: Table 7.1 gives the regression sums of squares for several models, which, along with the residual sum of squares for the full model, $RSS_1 = 3553$, are the building blocks of the incremental F -tests shown in Table 7.2. Table 7.3 shows the hypothesis tested by each of the incremental F -statistics in Table 7.2.

Although the analysis-of-variance table (Table 7.2) conventionally shows the tests for the main effects of education, income, and type before the education-by-type and income-by-type interactions, the structure of the model makes it sensible to examine the interactions first: Conforming to the principle of marginality, the test for each main effect is computed assuming that the interactions that are higher-order relatives of the main effect are 0 (as shown in Table 7.3). Thus, for example, the test for the income main effect assumes that the income-by-type interaction is absent (i.e., that $\delta_{11} = \delta_{12} = 0$), but not that the education-by-type interaction is absent ($\delta_{21} = \delta_{22} = 0$).¹⁵

The principle of marginality serves as a guide to constructing incremental F -tests for the terms in a model that includes interactions.

In this case, then, there is weak evidence of an interaction between education and type of occupation, and much stronger evidence of an income-by-type interaction. Considering the small number of cases, we are squeezing the data quite hard, and it is apparent from the coefficient standard errors (in Equation 7.12) and from the effect displays in Figure 7.11 that the interactions are not precisely estimated. The tests for the main effects of income, education, and type, computed assuming that the higher-order relatives of each such term are absent, are all highly statistically

¹⁴For standard errors of fitted values, see Exercise 9.14.

¹⁵Tests constructed to conform to the principle of marginality are sometimes called “type-II” tests, terminology introduced by the SAS statistical software package. This terminology, and alternative tests, are described in the next chapter.

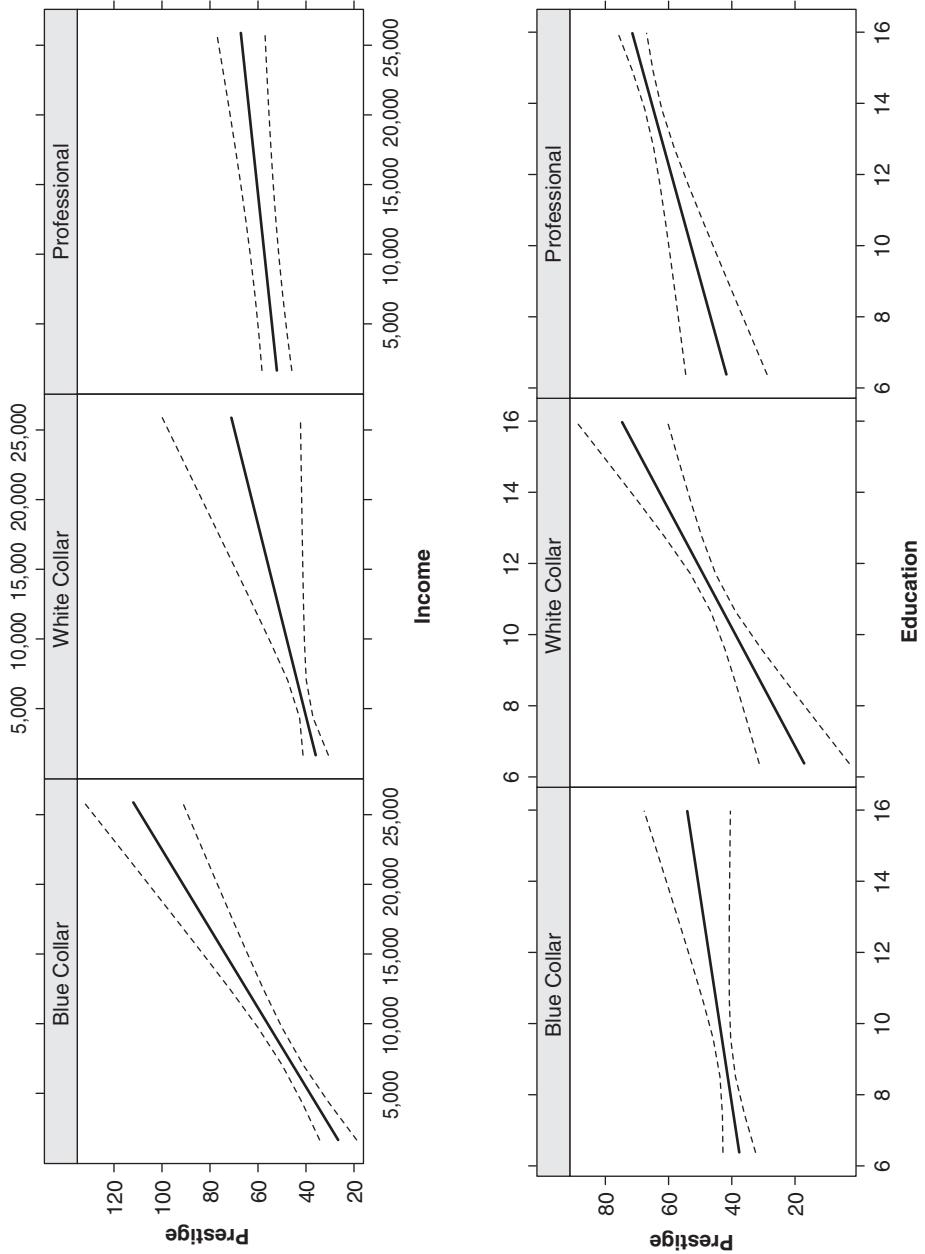


Figure 7.11 Income-by-type (upper panel) and education-by-type (lower panel) “effect displays” for the regression of prestige on income, education, and type of occupation. The solid lines give fitted values under the model, while the broken lines give 95% pointwise confidence intervals around the fit. To compute fitted values in the upper panel, education is set to its average value in the data; in the lower panel, income is set to its average value.

Table 7.1 Regression Sums of Squares for Several Models Fit to the Canadian Occupational Prestige Data

Model	Terms	Parameters	Regression Sum of Squares	df
1	$I, E, T, I \times T, E \times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}$	24,794.	8
2	$I, E, T, I \times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}$	24,556.	6
3	$I, E, T, E \times T$	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{21}, \delta_{22}$	23,842.	6
4	I, E, T	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2$	23,666.	4
5	I, E	α, β_1, β_2	23,074.	2
6	$I, T, I \times T$	$\alpha, \beta_1, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}$	23,488.	5
7	$E, T, E \times T$	$\alpha, \beta_2, \gamma_1, \gamma_2, \delta_{21}, \delta_{22}$	22,710.	5

NOTE: These sums of squares are the building blocks of incremental F -tests for the main and interaction effects of the explanatory variables. The following code is used for “terms” in the model: I , income; E , education; T , occupational type.

Table 7.2 Analysis-of-Variance Table, Showing Incremental F -Tests for the Terms in the Canadian Occupational Prestige Regression

Source	Models Contrasted	Sum of Squares	df	F	p
Income	3–7	1132.	1	28.35	<.0001
Education	2–6	1068.	1	26.75	<.0001
Type	4–5	592.	2	7.41	<.0011
Income \times Type	1–3	952.	2	11.92	<.0001
Education \times Type	1–2	238.	2	2.98	.056
Residuals		3553.	89		
Total		28,347.	97		

Table 7.3 Hypotheses Tested by the Incremental F -Tests in Table 7.2

Source	Models Contrasted	Null Hypothesis
Income	3–7	$\beta_1 = 0 \delta_{11} = \delta_{12} = 0$
Education	2–6	$\beta_2 = 0 \delta_{21} = \delta_{22} = 0$
Type	4–5	$\gamma_1 = \gamma_2 = 0 \delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = 0$
Income \times Type	1–3	$\delta_{11} = \delta_{12} = 0$
Education \times Type	1–2	$\delta_{21} = \delta_{22} = 0$

significant. In light of the strong evidence for an interaction between income and type, however, the income and type main effects are not really of interest.¹⁶

The degrees of freedom for the several sources of variation add to the total degrees of freedom, but—because the regressors in different sets are correlated—the sums of squares do not add to the total sum of squares.¹⁷ What is important here (and more generally) is that sensible hypotheses are tested, not that the sums of squares add to the total sum of squares.

7.4 A Caution Concerning Standardized Coefficients

In Chapter 5, I explained the use—and limitations—of standardized regression coefficients. It is appropriate to sound another cautionary note here: Inexperienced researchers sometimes report standardized coefficients for dummy regressors. As I have explained, an *unstandardized* coefficient for a dummy regressor is interpretable as the expected response-variable difference between a particular category and the baseline category for the dummy-regressor set (controlling, of course, for the other explanatory variables in the model).

If a dummy-regressor coefficient is standardized, then this straightforward interpretation is lost. Furthermore, because a 0/1 dummy regressor cannot be increased by one standard deviation, the usual interpretation of a standardized regression coefficient also does not apply. Standardization is a linear transformation, so many characteristics of the regression model—the value of R^2 , for example—do not change, but the standardized coefficient itself is not directly interpretable. These difficulties can be avoided by standardizing only the response variable and *quantitative* explanatory variables in a regression, leaving dummy regressors in 0/1 form.

A similar point applies to interaction regressors. We may legitimately standardize a quantitative explanatory variable *prior* to taking its product with a dummy regressor, but to standardize the interaction regressor itself is not sensible: The interaction regressor cannot change independently of the main-effect regressors that compose it and are marginal to it.

It is not sensible to standardize dummy regressors or interaction regressors.

Exercises

Exercise 7.1. Suppose that the values -1 and 1 are used for the dummy regressor D in Equation 7.1 instead of 0 and 1 . Write out the regression equations for men and women, and explain how the parameters of the model are to be interpreted. Does this alternative coding of the

¹⁶We tested the occupational type main effect in Section 7.2 (Equation 7.8 on page 129), but using an estimate of error variance based on Model 4, which does not contain the interactions. In Table 7.2, the estimated error variance is based on the full model, Model 1. Sound general practice is to use the largest model fit to the data to estimate the error variance even when, as is frequently the case, this model includes effects that are not statistically significant. The largest model necessarily has the smallest residual sum of squares, but it also has the fewest residual degrees of freedom. These two factors tend to offset one another, and it usually makes little difference whether the estimated error variance is based on the full model or on a model that deletes nonsignificant terms. Nevertheless, using the full model ensures an unbiased estimate of the error variance.

¹⁷See Section 10.2 for a detailed explanation of this phenomenon.

dummy regressor adequately capture the effect of gender? Is it fair to conclude that the dummy-regression model will “work” properly as long as two distinct values of the dummy regressor are employed, one each for women and men? Is there a reason to prefer one coding to another?

Exercise 7.2. Adjusted means (based on Section 7.2): Let \bar{Y}_1 represent the (“unadjusted”) mean prestige score of professional occupations in the Canadian occupational prestige data, \bar{Y}_2 that of white-collar occupations, and \bar{Y}_3 that of blue-collar occupations. Differences among the \bar{Y}_j may partly reflect differences among occupational types in their income and education levels. In the dummy-variable regression in Equation 7.7, type-of-occupation differences are “controlled” for income and education, producing the fitted regression equation

$$\hat{Y} = A + B_1X_1 + B_2X_2 + C_1D_1 + C_2D_2$$

Consequently, if we fix income and education at particular values—say, $X_1 = x_1$ and $X_2 = x_2$ —then the fitted prestige scores for the several occupation types are given by (treating “blue collar” as the baseline type):

$$\begin{aligned}\hat{Y}_1 &= (A + C_1) + B_1x_1 + B_2x_2 \\ \hat{Y}_2 &= (A + C_2) + B_1x_1 + B_2x_2 \\ \hat{Y}_3 &= \quad A \quad + B_1x_1 + B_2x_2\end{aligned}$$

- (a) Note that the *differences* among the \hat{Y}_j depend only on the dummy-variable coefficients C_1 and C_2 and not on the values of x_1 and x_2 . Why is this so?
- (b) When $x_1 = \bar{X}_1$ and $x_2 = \bar{X}_2$, the \hat{Y}_j are called *adjusted means* and are denoted \tilde{Y}_j . How can the adjusted means \tilde{Y}_j be interpreted? In what sense is \tilde{Y}_j an “adjusted” mean?
- (c) Locate the “unadjusted” and adjusted means for women and men in each of Figures 7.1(a) and (b) (on page 121). Construct a similar figure in which the difference between adjusted means is *smaller* than the difference in unadjusted means.
- (d) Using the results in the text, along with the mean income and education values for the three occupational types, compute adjusted mean prestige scores for each of the three types, controlling for income and education. Compare the adjusted with the unadjusted means for the three types of occupations and comment on the differences, if any, between them.

Exercise 7.3. Can the concept of an adjusted mean, introduced in Exercise 7.2, be extended to a model that includes interactions? If so, show how adjusted means can be found for the data in Figure 7.7(a) and (b) (on page 131).

Exercise 7.4. Verify that the regression equations for each occupational type given in Equation 7.13 (page 136) are identical to the results obtained by regressing prestige on income and education *separately* for each of the three types of occupations. Explain why this is the case.

Summary

- A dichotomous factor can be entered into a regression equation by formulating a dummy regressor, coded 1 for one category of the variable and 0 for the other category. A model incorporating a dummy regressor represents parallel regression surfaces, with the constant separation between the surfaces given by the coefficient of the dummy regressor.
- A polytomous factor can be entered into a regression by coding a set of 0/1 dummy regressors, one fewer than the number of categories of the factor. The “omitted” category, coded

0 for all dummy regressors in the set, serves as a baseline to which the other categories are compared. The model represents parallel regression surfaces, one for each category of the factor.

- Interactions can be incorporated by coding interaction regressors, taking products of dummy regressors with quantitative explanatory variables. The model permits different slopes in different groups—that is, regression surfaces that are not parallel.
- *Interaction* and *correlation* of explanatory variables are empirically and logically distinct phenomena. Two explanatory variables can interact *whether or not* they are related to one another statistically. Interaction refers to the manner in which explanatory variables *combine* to affect a response variable, not to the relationship *between* the explanatory variables themselves
- The principle of marginality specifies that a model including a high-order term (such as an interaction) should normally also include the lower-order relatives of that term (the main effects that “compose” the interaction). The principle of marginality also serves as a guide to constructing incremental *F*-tests for the terms in a model that includes interactions, and for examining the effects of explanatory variables.
- It is not sensible to standardize dummy regressors or interaction regressors.