Soci708 – Statistics for Sociologists
Module 2 – Looking at Data: Relationships¹

François Nielsen

University of North Carolina
Chapel Hill

Fall 2009

¹Adapted in part from slides from courses by Robert Andersen (University of Toronto) and by John Fox (McMaster University)
Most interesting statistical data analysis in the social sciences examines *relationships* between or among variables:

- Are national rates of infant mortality related to economic development?
- Do women in the paid labour force earn less than men? What about women and men with equal qualifications?
- Do people who read more books more likely to wear glasses than people who read fewer books?
- In U.S. states that have the death penalty, are African Americans convicted of murder more likely to be executed than whites convicted of murder?
In a survey of 300 men in Ohio respondents were asked whether they wore glasses and about the amount of reading they did. Results were as follows:

<table>
<thead>
<tr>
<th>Wears glasses?</th>
<th>Amount of reading</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above average</td>
<td>64.4</td>
<td>35.6</td>
<td>100</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>38.1</td>
<td>61.9</td>
<td>100</td>
<td>126</td>
<td></td>
</tr>
<tr>
<td>Below average</td>
<td>30.7</td>
<td>69.3</td>
<td>100</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>42.0</td>
<td>58.0</td>
<td>100</td>
<td>300</td>
<td></td>
</tr>
</tbody>
</table>

What relationship does the table show?

Two variables measured on the same cases are associated if knowing the value of one of the variables tells us something about the values of the other variable that we would not know without this information.
Introduction

Relationships between two Quantitative Variables

- We will look first at methods for examining relationships between quantitative variables, such as individuals’ education and their income.
- **Scatterplots**
  - Graphs for examining the relationship between two quantitative variables
- **Correlation Coefficient**
  - Measures the degree to which two quantitative variables are related (if their relationship is well summarized by a straight line)
- **Least-Squares Regression**
  - A method for finding a straight line to summarize the relationship between two quantitative variables
When we are interested in the relationship between two variables, it is usually because one variable is thought to affect, partly determine, or influence the other.

We may expect, for example, that individuals’ education will have an influence on their income.

In this case we say that income is the *response variable* (also called *dependent variable*) and education is the *explanatory variable* (also called *independent variable* or the *predictor*).

Distinguishing between explanatory and response variables is easiest in experimental research, where the explanatory variable is directly manipulated by the researcher. In observational research this is typically more difficult.

Which of reading books and wearing glasses affect the other?
Introduction
Explanatory and Response Variables

- Sometimes (but less often) we are interested in the relationship between two variables without identifying one as the explanatory variable and the other as the response.
  - For example, verbal and mathematical scores on an achievement test.

- Occasionally, we want to use the explanatory variable to predict the response variable without necessarily assuming the explanatory variable causes the response.
  - Example?
As usual in data analysis, it is best to start by examining the data with a graph.

- In fact, we should look at the distribution of each variable separately before examining their relationship.

The best way to examine the relationship between two quantitative variables is with a *scatterplot*.

- In a scatterplot, the values of one variable appear (the explanatory variable, if there is one) appear on the horizontal (or $x$) axis; the values of the other variable (the response variable) appear on the vertical (or $y$) axis.
- Each individual observation in the data is plotted as a point on the graph according to its values $(x_i, y_i)$ for the two variables.
Scatterplots
Examples: Canadian Occupational Prestige Data (from John Fox’s car package for R)

- For several example scatterplots I use data on 102 Canadian occupations.
- There are four variables for each case; the first three variables are from the 1971 Census of Canada.
  1. the average number of years of education for individuals in the occupation;
  2. the average income, in dollars, for individuals in the occupation;
  3. the percentage of individuals in the occupation who are women; and
  4. the average prestige rating of the occupation in a social survey of Canadians conducted in the mid-1960s.
    - The prestige rating ranges from 14.2 for newsboys to 87.8 for physicians and surgeons (university teachers are second at 84.6).
Scatterplots

The Prestige Data from John Fox’s car Package for R

> # in R
> library(car)
> data(Prestige)
> attach(Prestige)
> # a few cases from the Prestige dataset
> some(Prestige)

<table>
<thead>
<tr>
<th>education</th>
<th>income</th>
<th>women</th>
<th>prestige</th>
<th>census</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>PURCHASING.OFFICERS</td>
<td>11.42</td>
<td>8865</td>
<td>9.11</td>
<td>56.8</td>
<td>1175</td>
</tr>
<tr>
<td>CIVIL.ENGINEERS</td>
<td>14.52</td>
<td>11377</td>
<td>1.03</td>
<td>73.1</td>
<td>2143</td>
</tr>
<tr>
<td>TYPISTS</td>
<td>11.49</td>
<td>3148</td>
<td>95.97</td>
<td>41.9</td>
<td>4113</td>
</tr>
<tr>
<td>BOOKKEEPERS</td>
<td>11.32</td>
<td>4348</td>
<td>68.24</td>
<td>49.4</td>
<td>4131</td>
</tr>
<tr>
<td>REAL.ESTATE.SALESMEN</td>
<td>11.09</td>
<td>6992</td>
<td>24.44</td>
<td>47.1</td>
<td>5172</td>
</tr>
<tr>
<td>FIREFIGHTERS</td>
<td>9.47</td>
<td>8895</td>
<td>0.00</td>
<td>43.5</td>
<td>6111</td>
</tr>
<tr>
<td>FARM.WORKERS</td>
<td>8.60</td>
<td>1656</td>
<td>27.75</td>
<td>21.5</td>
<td>7182</td>
</tr>
<tr>
<td>TEXTILE.WEAVERS</td>
<td>6.69</td>
<td>4443</td>
<td>31.36</td>
<td>33.3</td>
<td>8267</td>
</tr>
<tr>
<td>TEXTILE.LABOURERS</td>
<td>6.74</td>
<td>3485</td>
<td>39.48</td>
<td>28.8</td>
<td>8278</td>
</tr>
<tr>
<td>BOOKBINDERS</td>
<td>8.55</td>
<td>3617</td>
<td>70.87</td>
<td>35.2</td>
<td>9517</td>
</tr>
</tbody>
</table>

9/147
Scatterplots
Example of scatterplot in R

> # in R
> library(car)
> data(Prestige)
> attach(Prestige)
> # plot(x,y)
> plot(education, income,
>     col="red", cex=1.5)
Scatterplots

Interpreting a Scatterplot

▶ In any graph we look for an overall pattern and for striking deviations from that pattern.

▶ Overall pattern:
  ▶ Are there distinct clusters of observations in the data?
    ▶ Clustering could indicate the presence of a categorical explanatory variable
  ▶ Is there a systematic relationship between variables, and if so what is its form? (Linear, non-linear?)
  ▶ What is the direction of the relationship? (Positive, negative?)
  ▶ What is the strength of the relationship? (How closely do points follow a simple form?)

▶ Deviations:
  ▶ Are there outliers?
    ▶ An outlier in a scatterplot is a point that is far away from the general pattern of the data.
    ▶ Even if an observation is not an outlier on either variable considered individually, it may be an outlier if it is not in line with the overall pattern of the data
    ▶ Outliers can weaken or distort relationships as measured by the correlation coefficient and least-squares regression.
Here the data seem to cluster into two or more groups.

Often a plot like this indicates the data are divided into categories of a categorical explanatory variables.

In this example, one cluster corresponds to more developed, and the other to less developed societies.
Scatterplots
Interpreting a Scatterplot: Identifying Outliers

▶ Even though neither of its $x$ or $y$ values is unusual, this point is an outlier.
  ▶ It stands apart from the general pattern of the rest of the data
  ▶ Omitting this case would strengthen the relationship between the two variables
Scatterplots
Interpreting a Scatterplot: Form of Association

- Is there any apparent relationship or association between the two variables?
  - Is the relationship best described by a straight line or by a curve?
    - In the former case, the relationship is described as *linear*
- Linear relationships are particularly simple, allowing the use of the *correlation coefficient* and *linear (least squares) regression*
  - Simple curves can also be accommodated under least squares, but we will not discuss this further in this course
Scatterplots
Interpreting a Scatterplot: Form of Association

- Notice that the pattern in the top scatterplot is well described by a single straight line. We call this a linear relationship
  - As $x$ goes up, $y$ goes up by about the same amount on average throughout the range of $x$

- The bottom scatterplot shows a nonlinear relationship
Scatterplots
Interpreting a Scatterplot: Direction of Association

▶ Is there a positive or negative association between the two variables?
  ▶ In a *positive association* large values of one variable tend to occur with large values of the other variable, and small values of one with small values of the other.
    ▶ That is, *as x goes up, y also goes up*
  ▶ In a *negative association* large values of one variable tend to occur with small values of the other.
    ▶ That is, *as x goes up, y goes down*
Scatterplots
Interpreting a Scatterplot: Direction of Association

Positive Relationship

Negative Relationship

No Relationship
There is a strong linear relationship if the points fit closely along a straight line.

The relationship is weak if there is a great deal of variation of the response (or dependent) variable around the simple form. In other words, we try to assess the vertical distances of each y-value from a possible line through the data – if on average they appear small, there is a strong linear relationship.

It is possible to have a very strong relationship that is not linear. For now, however, we will focus on linear relationships. (Some discussion of nonlinear relationships will come later.)
Interpreting a Scatterplot: Strength of Association

No Relationship

Moderate Relationship

Strong Relationship
Scatterplots
Adding a Categorical Variable: Coded Scatterplot

- By using different plotting symbols or colors to label its categories, you can show the effect of a categorical explanatory variable on a scatterplot.
- In the following scatterplot, I plot prestige against income for 102 occupations in the Prestige dataset in car.
  - I want to see the impact of type of occupation on the relationship between education and income, so I assign a different symbol to three occupation types: professionals, blue collar and white collar

```r
> library(car)
> data(Prestige)
> attach(Prestige)
> par(pty="s")
> scatterplot(prestige~income|type,
>             boxplots=FALSE,smooth=FALSE,cex=1.5)
```
By distinguishing between occupation types, we can see that the data are indeed clustered according to the type.

Professionals typically have higher prestige, and income has less affect on prestige level than for other occupation types.
Scatterplot Smoothers
Smoothing a Scatterplot with LOWESS

- Example of LOWESS
The Correlation $r$

Formula for the Correlation Coefficient

- The *correlation coefficient* measures the strength and direction of the linear relationship between two variables $x$ and $y$.
  - It is also called the *Pearson correlation*, after Karl Pearson.
- The correlation coefficient is defined by the formula

$$
r = \frac{1}{n - 1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)
$$

- The formula is complicated so it helps to unpack it:
  - $\bar{x}$ and $\bar{y}$ are the means of $x$ and $y$
  - $s_x$ and $s_y$ are the standard deviations of $x$ and $y$
  - $\left( \frac{x_i - \bar{x}}{s_x} \right)$ and $\left( \frac{y_i - \bar{y}}{s_y} \right)$ are therefore the standardized scores (z-scores) for $x$ and $y$ for case $i$.
    - Remember that this does not imply that $x$ and $y$ are normally distributed.
The Correlation $r$

Formula for the Correlation Coefficient (cont’d)

- The correlation formula repeated:

$$r = \frac{1}{n - 1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

- The product $\left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$ will be positive when both $x$ and $y$ are above their means or when both are below their means.

- The product is negative when one variable is above its mean and the other below.

- The capital sigma ($\sum$) sign means “sum up these products over all cases”; then divide result by $n - 1$.

- When there is a positive relationship between $x$ and $y$, the correlation is therefore positive; when there is a negative relationship, the correlation is negative.
The Correlation $r$

How Points Contribute to the Correlation

- Signs of the products $(x_i - \bar{x})(y_i - \bar{y})$

- In this example, most of the points are in the positive quadrants (upper right, lower left), so the correlation between $x$ and $y$ is positive.
The Correlation $r$
Calculating $r$ in R

- Using the Prestige dataset, look at the relationship between prestige level and average education for the 102 occupations.

- As always, start with a scatterplot to check the relationship is linear, then calculate correlation:

  ```r
  # in R
  > data(Prestige)
  > attach(Prestige)
  > plot(education, prestige, cex=1.5, col="red")
  > abline(lm(prestige~education), col="red")
  > cor(education, prestige)
  [1] 0.8501769
  ```
Correlation

Correlations for Canadian Occupations ($N = 102$)

- prestige by education

\[ r = .850 \]

- prestige by income

\[ r = .715 \]

- prestige by women

\[ r = -.118 \]
Properties of Correlations

- Note that $x$ and $y$ enter the formula for the correlation in the same (symmetrical) way.
  - The correlation, therefore, does not depend upon one variable being identified as the response and the other as the explanatory variable.
- The correlation $r$ measures the strength and direction of the relationship:
  - It ranges from $-1$ (a perfect negative linear relationship) to $+1$ (a perfect positive linear relationship). In both cases, all the cases fit perfectly along a straight line.
  - A correlation of $r = 0$ indicates that there is no linear relationship between the variables.
  - Values between $-1$ and $0$, and between $0$ and $+1$, indicate linear relationships of varying strengths.
Properties of Correlations

▶ The correlation is appropriate for *quantitative variables* and qualitative variables with only two categories (dichotomies).
▶ The correlation is fully meaningful only when the relationship between two quantitative variables is linear.
  ▶ When the relationship is nonlinear, $r$ still measures the strength of the *linear component* of the relationship.
  ▶ Some types of nonlinear relationships – even quite strong nonlinear relationships – produce correlations of zero.
▶ $r$, like the mean and standard deviation, can be strongly affected by one or a small number of outliers.
▶ It *does not* assume a causal relationship.
Properties of Correlations

*r* as an Effect Size Index

- Because the correlation uses standardized variables, it does not change when we change the units of measurement of the variables – for example, from dollars to thousands of dollars, or from kilos to pounds.
  - The correlation *r* itself has no units; it is dimensionless.
- Because *r* is dimensionless, *r* values can be meaningfully compared across studies using different units.
  - Cohen (1992)\(^2\) proposed that *r* represents a useful *standardized effect size index* to evaluate the strength of the relationship between two variables.
  - He suggested interpreting values of *r* of .10, .30 and .50 as representing *small*, *medium*, and *large* effects or associations, respectively.

The Correlation $r$

From Fox (1997, Figure 5.3 p.92)

(a) $r = 0$

(b) $r = 0$

(c) $r = -1$

(d) $r = .4$

(e) $r = -.8$

Figure 5.3. Scatterplots illustrating different levels of correlation: $r = 0$ in both (a) and (b); $r = -1$ in (c); $r = .4$ in (d); and $r = -.8$ in (e). All of the datasets have $n = 50$ observations.
Linear regression summarizes a linear relationship between a response variable $y$ and one or more explanatory variables $x$

Unlike correlation, regression assumes a causal direction

- i.e., one of the variables, $y$, is the response variable and the other variables(s), $x$, is (are) explanatory
- The regression of $y$ on $x$ will usually give a different regression line than the regression of $x$ on $y$

Least-squares regression is the most common form of regression analysis – it is versatile and adaptable.

- Simple regression: one explanatory variable
- Multiple regression: two or more explanatory variables
A straight line can be represented by an equation

\[ y = a + bx \]

where

- \( a \), called the *y-intercept* of the line, represents the \( y \)-value corresponding to an \( x \)-value of 0.
- \( b \), called the *slope* of the line, indicates how much \( y \) changes when \( x \) is increased by 1 (i.e., one unit).
  - If \( b \) is *positive*, then the value of \( y \) *increases* as \( x \) increases;
  - if \( b \) is *negative*, then the value of \( y \) *decreases* as \( x \) increases;
  - if \( b = 0 \), then the line is horizontal – the value of \( y \) *does not change* as \( x \) changes.
Fitting a Line to Data

Equation of a Straight Line

From Bob Andersen’s slides
The Least-Squares Regression Line
How Should We Fit a Line to a Scatterplot?

- Unless the linear relationship between $y$ and $x$ is perfect – never the case for real data – no line will go through all of the points in the scatterplot.
  - When the linear relationship between $y$ and $x$ is strong, it is possible to fit a line “by eye” to the scatterplot of the data.
  - When the relationship is weaker it is more difficult to see where the line would fit.
- We therefore need a method of fitting a line to the scatter of points that doesn’t depend upon subjective judgment.
  - The line should come “as close to the points as possible”.
A line that comes close to the data allows us to “predict” values of \( y \) for specific values of \( x \).

We take as an example the relationship of prestige with average years of education for the 102 Canadian occupations. The next slide shows a scatterplot of prestige by education with the least-squares regression line \( \hat{y} = -10.733 + 5.361x \) already drawn in.

- We will explain where this line comes from in a moment.

Take the occupation COMMERCIAL.ARTISTS.

- Education is 11.09 years.
- To find the predicted or fitted value of \( y \) for this occupation, go up the line above 11.09 years, then go over to the \( y \)-axis to find the corresponding value of \( y \), called \( \hat{y} \).
- Or, using the equation above, find

\[
\hat{y} = -10.733 + 5.361 \times 11.09 = 48.71
\]
The Least-Squares Regression Line

How Should We Fit a Line to a Scatterplot?

- The residual $e_i$ of an observation is the vertical distance between the observed value $y_i$ and the predicted value $\hat{y}$, called “y-hat”, given by the formula $\hat{y}_i = a + bx_i$.

- Observed prestige for Commercial Artists is 57.2.

- Thus the residual $e_i$ is given by $y_i - \hat{y}_i = 57.2 - 48.71 = 8.49$

- This is a positive residual: the occupation enjoys greater prestige than that predicted by education.
The Least-Squares Regression Line

How Should We Fit a Line to a Scatterplot?

- Other points in the data will have positive or negative residuals, unless they are situated exactly on the regression line \( \hat{y} = -10.73 + 5.36x \) where \( x \) represents education.

- The graph to the right shows the residuals \( e_i = y_i - \hat{y}_i \) for three occupations.
The Least-Squares Regression Line
How Should We Fit a Line to a Scatterplot?

- Again, the predicted value of \( y \) is represented by \( \hat{y} \) (called “y-hat”)
- The predicted and observed \( y \)-values usually differ, i.e.

\[
y_i = \hat{y}_i + e_i
\]

- For each observation, the difference between the observed and predicted \( y \)-value, representing the “error” in prediction for that observation, is called the \textit{residual} (literally, what is left over):

\[
\text{residual} = \text{observed value} - \text{predicted value}
\]
\[
e_i = y_i - \hat{y}_i
\]

- We want to find the regression line that minimizes the residuals over all the observations.
The Least-Squares Regression Line
How Should We Fit a Line to a Scatterplot?

▶ When there are just two points minimizing the residuals is easy:
  ▶ Simply pass a line between the two points.
▶ When there are many points, however, there are several different ways one could proceed.
▶ The method of least-squares finds the line with the *smallest possible sum of squared residuals*:
  ▶ That is, choose $a$ and $b$ to minimize

$$\sum \text{residuals}_i^2 = \sum e_i^2$$

▶ The residuals are *squared* before adding them up to prevent positive residuals (corresponding to points above the line) from cancelling out negative residuals (points below the line). Squaring makes all of the residuals positive.
Another way to make all of the residuals positive is to take their absolute values:

That is, choose $a$ and $b$ to minimize

$$\sum |\text{residuals}_i| = \sum |e_i|$$

This approach has advantages – for example, it produces values of $a$ and $b$ that are more resistant to outliers than those produced by least-squares regression – but it is more difficult mathematically.
The Least-Squares Regression Line
Finding the LS Coefficients

- The mathematics of finding the least-squares line is not of direct concern to us, but the solution is.
- The least-squares line has the equation:

\[ \hat{y} = a + bx \]

with slope

\[ b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = r \frac{s_y}{s_x} \]

and intercept

\[ a = \bar{y} - b\bar{x} \]

where

- \( r \) is the correlation between \( y \) and \( x \);
- \( s_y \) is the standard deviation of the response variable \( y \);
- \( s_x \) is the standard deviation of the explanatory variable \( x \), and
- \( \bar{y} \) and \( \bar{x} \) are the means of the two variables.
The Least-Squares Regression Line

Discovery of the Method of Least Squares by Adrien Legendre in 1805

- Legendre developed the LS method in the context of reconciling astronomical observations subject to error (next slide).
- Friedrich Gauss also claimed the discovery, but Legendre published first.
The Least-Squares Regression Line

Legendre’s (1805) Appendix (Stigler 1986, Figure 1.5 p.58)
The Least-Squares Regression Line
Finding the LS Coefficients

Calculating the least-squares coefficients \( a \) and \( b \) according to these formulas is a lot of work, even when the number of observations \( n \) is not very large.

Fortunately the computer will do this for us.
The Least-Squares Regression Line
Finding the LS Coefficients

- We continue with the regression of prestige (\(y\)) on education (\(x\)).
- Starting with the correlation, standard deviations, and means of the two variables,

\[
\begin{align*}
    r &= .85018 \\
    \bar{x} &= 10.7380 \\
    \bar{y} &= 46.8333 \\
    s_x &= 2.7284 \\
    s_y &= 17.2045
\end{align*}
\]

we find \(b\) and then \(a\) as in the next slide.
The Least-Squares Regression Line
Finding the LS Coefficients

1. Start by finding the slope, $b$:

$$b = r \frac{s_y}{s_x}$$

$$= .85018 \times \frac{17.2045}{2.7284}$$

$$= 5.3610$$

2. Next, find the intercept, $a$:

$$a = \bar{y} - b\bar{x}$$

$$= 46.8333 - 5.3610 \times 10.7380$$

$$= -10.7331$$

3. The fitted regression equation is therefore:

$$\hat{y} = a + bx$$

$$\hat{y} = -10.7331 + 5.3610x$$
Interpreting the Regression Line

Plotting the Regression Line in R

```r
> library(car)
> data(Prestige)
> attach(Prestige)
> plot(education, prestige,
      cex=1.5, col="red")
> abline(lm(prestige~education),
      col="red")
> model1<-lm(prestige~education)
> model1
```

Call:
`lm(formula = prestige ~ education)`

Coefficients:
```
(Intercept)    education
  -10.732       5.361
```

---

**Code Explanation**

1. **library(car)**: Loads the `car` package, which provides tools for regression analysis.
2. **data(Prestige)**: Loads the Prestige dataset into the environment.
3. **attach(Prestige)**: Attaches the Prestige dataset to the environment, allowing variables to be directly accessed by name.
4. **plot(education, prestige, cex=1.5, col="red")**: Plots the relationship between `education` and `prestige` with customized `cex` (character expansion) and color.
5. **abline(lm(prestige~education), col="red")**: Adds the regression line to the plot in red.
6. **model1<-lm(prestige~education)**: Fits a linear model where `prestige` is predicted by `education`.
7. **model1**: Displays the model summary, showing the coefficients for the intercept and `education` variable.
Interpreting the Regression Line
Plotting the Regression Line in Stata

.g. twoway (scatter prestige education)
   (lfit prestige education,
    xtitle("Education")
    ytitle("Prestige")
    xsize(2) ysize(2) )

.g. cor prestige education
(obs=102)

<table>
<thead>
<tr>
<th></th>
<th>prestige</th>
<th>education</th>
</tr>
</thead>
<tbody>
<tr>
<td>prestige</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>education</td>
<td>0.8502</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Interpreting the Regression Line

Regression Output in Stata

```
. * in Stata
. * use the File/Open ... menu to locate and open Prestige data set
. use "D:\soci708\data\data_from_car_Stata\Prestige.dta", clear
. reg prestige education

Source | SS    df  MS
-------------+------------------------------ F(  1, 100) = 260.75
Model      | 21608.4361  1  21608.4361 Prob > F = 0.0000
Residual  | 8286.99 100  82.8699 R-squared = 0.7228
-------------+------------------------------ Adj R-squared = 0.7200
Total      | 29895.4261 101 295.994318 Root MSE = 9.1033

------------------------------------------------------------------------------
prestige | Coef.  Std. Err.  t    P>|t|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
education |   5.360878    .3319882  16.15  0.000     4.702223    6.019533
   _cons   |  -10.73198    3.677088  -2.92  0.004    -18.02722   -3.436743
------------------------------------------------------------------------------
```

50 / 147
Interpreting the Regression Line

Plotting the Regression Line

- To plot the regression line on the scatterplot, we only need two points on the line.
- For example, we can find the $\hat{y}$-values corresponding to $x$-values of 6 and 16:

  For $x = 6$:
  \[
  \hat{y} = -10.733 + 5.361 \times 6 = 21.433
  \]

  For $x = 16$:
  \[
  \hat{y} = -10.733 + 5.361 \times 16 = 75.043
  \]

- Connecting the points (6, 21.433) and (16, 75.043) locates the least-squares line.
- Two points that are always on the least-squares line are $(0, a)$ and $(\bar{x}, \bar{y})$. 
Interpreting the Regression Line

Interpreting the Intercept, $a = -10.73$

- Recall that the intercept is where the line passes through the $y$-axis
  - Interpreted literally, then, the intercept is the predicted prestige score ($-10.73$) for an occupation with 0 years of education

- In this instance, we should not interpret the value of $a$ literally, because:
  - None of the 102 occupations in the dataset has less than 6 years of education
  - Prestige scores cannot be negative (they can range only from 0 to 100).

- In other words, the intercept in this case helps us draw the line but nothing else.
Interpreting the Regression Line

The slope, $b = 5.361$

- The slope tells us how much on average $y$ changes as $x$ increases by one unit.
- In this case, then, $b = 5.361$ tells us that each additional year of education is accompanied on average by an increase of a bit more than 5 prestige points.
  - This is a descriptive statement about the association between prestige and education.
  - We may or may not be willing to give the slope coefficient a causal interpretation:
    - “Increasing average education by one year causes the prestige of the occupation to rise by more than 5 points.”
- Because it tells us how $y$ changes with $x$, we are usually more interested, substantively, in the slope $b$ than in the intercept $a$. 
Regression vs. Correlation

- The slope $b$ of the least-squares regression line and the correlation $r$ are related by the equations

$$b = r \frac{s_y}{s_x} \quad \text{and, conversely} \quad r = b \frac{s_x}{s_y}$$

- The correlation and slope are similar in certain respects and different in others:
  - When $r = 0$, indicating that there is no linear relationship between $y$ and $x$, then $b = 0$ as well.
  - If $x$ and $y$ are both standardized variables (in which case $s_x = s_y = 1$), then $b = r$
Regression vs. Correlation

The Two LS Lines

- The correlation coefficient $r$ does not depend upon which variable is treated as the response and which as the explanatory variable.
- The slope $b$ does depend upon which variable is treated as the response.
  - If $x$ is regressed on $y$ rather than vice-versa (i.e., if $x$ is treated as the response), then
    \[ b_{x \text{ on } y} = r \frac{s_x}{s_y} \]
    which is usually different from $b_{y \text{ on } x}$.
- There are two least-squares regression lines – one for the regression of $y$ on $x$, and the other for the regression of $x$ on $y$.
- Unless $r = 1$, these two regression lines are different.
Regression vs. Correlation
The Two LS Lines

- The plot shows the two regression lines:

1. Prestige $y$ regressed on education $x$:
   \[ \hat{y} = -10.733 + 5.361x \]

2. Education $x$ regressed on prestige $y$:
   \[ \hat{x} = 4.42 + .13y \]
Regression vs. Correlation

But $r$ is the same as $b$ for Standardized $x$ and $y$

- The slope $b$ of the regression of $z_y$ on $z_x$ is equal to the correlation $r$.
- Thus one interpretation of $r$ is that an increase of 1 SD in one variable ($x$ or $y$) is associated with an increase of $r$ SD in the other.
Understanding $r^2$

- The square of the correlation coefficient ($r^2$) has a special interpretation in least-squares regression:
  - Recall the regression residuals, also called $e_i$, which give the differences between observed and predicted response values,
    \[ \text{residual}_i = e_i = y_i - \hat{y}_i \]
  - The sum of squared residuals represents the variation of $y$ around the regression line:
    \[ \sum \text{residual}^2 = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 \]
    This quantity is called SSE for \textit{sum of squares – error}.
  - The \textit{total variation} of $y$ around its mean (ignoring the regression line) is:
    \[ \sum (y_i - \bar{y})^2 \]
    This quantity is called SSTO for \textit{sum of squares – total}. 

58/147
Understanding $r^2$

- The difference between these two measures of variation is the amount of variation accounted for by the regression of $y$ on $x$:

  \[
  \text{“explained” variation} = \text{total variation} - \text{residual variation}
  \]

  \[
  = \sum (y_i - \bar{y})^2 - \sum \text{residual}^2
  \]

  \[
  = \sum (\hat{y}_i - \bar{y})^2
  \]

  This last quantity is called SSR for \textit{sum of squares – regression}.

- The squared correlation $r^2$ or \textit{r-squared} expresses explained variation of $y$ as a proportion of total variation of $y$.

  \[
  r^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{SSR}{SSTO}
  \]

  \[
  = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum \text{residual}_i^2}{\sum (y_i - \bar{y})^2}
  \]
Understanding $r^2$

- When there is a perfect linear relationship between $y$ and $x$, the residuals are all zero; thus the sum of squared residuals is zero and $r^2 = 1$.
- When there is no linear relationship between $y$ and $x$, the explained variation is 0, and $r^2 = 0$.
- In the regression of occupational prestige on education, $r = 0.85018$, and thus the regression accounts for

$$r^2 = 0.85018^2 = 0.7228$$

or about 72 percent of the variation in prestige scores.
Understanding $r^2$

Standard Error of the Regression

- The standard deviation of the residuals, called the *standard error of the regression*, or $s_e$, is another measure of fit for the least squares line.

- Measured in the units of the dependent variable, it represents an “average” of the residuals:

$$s_e^2 = \frac{\sum e_i^2}{n-2} = \frac{SSE}{n-2}$$

Therefore, $s_e = \sqrt{\frac{\sum e_i^2}{n-2}}$

- For example, in the regression of prestige on education $s_e = 9.103$ (see Stata printout at “root MSE”).
  - This means that on average, predicting income from education results in an error of just over 9 prestige points.
Need to Use Graphs

The Anscombe Data

- Anscombe’s (1973) contrived four data sets to show the importance of using graphs in data analysis rather than simply looking at numerical outputs.

- Anscombe’s four sets are constructed to have exactly the same regression of $y$ on $x$ and the same correlation:

  $$\hat{y} = 3.0 + 0.5x$$
  $$r = .82$$

- $\bar{x}$, $\bar{y}$, $s_x$ and $s_y$ are all the same in the four datasets as well.

- Yet, the linear least-squares regression is a good summary of the relationship between $x$ and $y$ only for the first data set.

- None of the problems with the datasets is clear from the fitted regression equation and correlation, and none (but the last) is clear from looking at the numerical data.
Need to Use Graphs

The Anscombe Data

- The linear regression line adequately summarizes the relationship only in graph (a).
- In graph (b) the relationship is nonlinear.
- In graph (c) there is an outlier.
- In graph (d) the least-squares line “chases” the influential observation.
Need to Use Graphs
R Program for Previous Slide

```r
> # Data from CD ROM in Moore & McCabe (2006, Table 2.6)
> Anscombe<-read.table("ta02_006.txt", header=TRUE)
> attach(Anscombe)
> par(pty="s") # Make plot square
> plot(x, y1, cex=1.5, main="Accurate summary")
> abline(lm(y1~x))
> plot(x, y2, cex=1.5, main="Conceals curvilinearity")
> abline(lm(y2~x))
> plot(x, y3, cex=1.5, main="Drawn to outlier")
> abline(lm(y3~x))
> plot(x4, y4, cex=1.5, main="Chases outlier")
> abline(lm(y4~x4))
```
Residuals

Residuals and Residual Plots

- Recall that a residual is the difference between an observed value of the response variable and the value predicted by the regression line. That is

\[
\text{residual} = \text{observed } y - \text{predicted } y
\]

\[
= y - \hat{y}
\]

- Residuals from the least-squares line have the special property that their mean is always zero.

- A residual plot is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of the regression line, detecting problems such as:
  - unequal variance of $e$ for different $\hat{y}$;
  - outliers and influential observations;
  - nonlinearity in the relationship of $y$ with $x$. 
Residuals
Residuals and Residual Plots

▶ The two plots below examine the relationship between field and lab measurements of depth of defects in the Trans-Alaska oil pipeline (IPS6e p. 90).

▶ One can see that variance of the residuals increases with the size of the defect, a condition called *heretoskedasticity*.
Outliers and Influential Observations

Influential Data

- The least-squares line is a good summary of the relationship between $y$ and $x$ when the relationship is in fact linear and when the data are well behaved.
  - But the least-squares line can sometimes be markedly affected by outlying data.
- In regression analysis, an outlier is a point far away from the general pattern of the data.
  - It is a point whose $y$ value is unusual compared to other points with similar $x$-values.
- Points with unusual $x$-values, when they are out of line with the rest of the data, can be influential, in the sense that their inclusion in the dataset can markedly alter the regression line.
  - Like the mean, standard deviation, and correlation, therefore, the least-squares regression line is not resistant to unusual data.
The scatterplot on the next slide, showing reported and measured weight in kg, is for 181 young men and women engaged in regular exercise.

- The data were collected by Caroline Davis, a psychologist who studies eating disorders.
- If the subjects are unbiased reporters of their weight, then the regression line should be approximately by $\hat{y} = x$ (that is, an intercept of 0 and a slope of 1).
- When the outlying point at the right is omitted, the least-squares line is close to the line of unbiased reporting (the red dashed line).
- In this case, the influential outlier represents an error in recording the data.
Outliers and Influential Observations

Influential Data: Reported and Measured Weight (Davis Data)

- Using the Davis data in the car package, we see clearly that observation 12 is influential.
- The model including observation 12 does a poor job of representing the trend in the data; the model excluding observation 12 does much better.
- The regression output on the next slide confirms this.
Outliers and Influential Observations
Influential Data: Regression of Reported on Measured Weight

| Model of repwt including all cases | Estimate | Std. Error | T value | P(>|t|) |
|-----------------------------------|----------|------------|---------|---------|
| Intercept                         | 15.759   | 2.498      | 6.308   | .000    |
| weight                            | 0.753    | 0.037      | 20.484  | 0.000   |

Residual standard error: 7.584 on 181 df

$R^2 = 0.699$

| Model of repwt excluding observation #12 | Estimate | Std. Error | T value | P(>|t|) |
|-----------------------------------------|----------|------------|---------|---------|
| Intercept                              | -.928    | .861       | -1.078  | 0.283   |
| height                                 | 1.014    | 0.013      | 78.926  | 0.000   |

Residual standard error: 2.318 on 180 df

$R^2 = 0.972$
Outliers and Influential Observations

Influential Data: Davis Data Example R Script

```r
> par(pty="s")
> data(Davis)
> attach(Davis)
> davis.mod1 <- lm(repwt~weight)
> summary(davis.mod1)
> # omitted output

> plot(weight, repwt, main="Davis data")
> identify(weight, repwt, row.names(Davis))

> # observation 12 returned as outlier
> abline(davis.mod1, lty=1, col=1, lwd=3)
> davis.mod2 <- update(davis.mod1, subset=-12)
> abline(davis.mod2, lty=2, col=2, lwd=3)
> legend(locator(1), lty=1:2, col=1:2, lwd=3,
+ legend=c("All cases", "Outlier excluded"))
> summary(davis.mod2)
> # omitted output
```
The Restricted-Range Problem
Validity of SAT and GPA for College Performance

- The *criterion validity* of a measure is the correlation of the measure with the variable it is intended to measure or predict.

- Combining several studies the College Board reported the following correlation between explanatory variables and overall GPA of college students:

<table>
<thead>
<tr>
<th>SAT Math and Verbal grades</th>
<th>SAT plus grades</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r = 0.36 )</td>
<td>( r = 0.42 )</td>
</tr>
<tr>
<td>( r = 0.52 )</td>
<td></td>
</tr>
</tbody>
</table>

- Note that the proportion of variance explained is at most 27% \( (0.52^2 = 0.27) \).
The Restricted-Range Problem

Validity of SAT and GPA for College Performance

Why the low proportion of explained variance? Because of the restricted-range problem:

- Individual colleges do not admit students with the full range of ability.
- Because of this restriction of range $r$ and $r^2$ are lower than they would be if the full range could be observed.

An interesting discussion of the restricted-range issue is in:

Lurking Variables and the Question of Causation

Lurking Variables

- This section will discuss the problem of lurking variables and the question of causation. Textbook materials:
Cautions About Correlation & Regression

Summary and Recommendations

- **Look at the data**
  - We must make sure that a linear trend makes sense and that there are no influential cases

- **Extrapolation**
  - It is not safe to predict outside the range of $x$ values observed in the data

- **Lurking Variables**
  - Can have an important effect on the relationship between $x$ and $y$

- **Association is not Causation**
  - Statistical association does not prove causation

- **Averaged Data**
  - Relationships using aggregate data are sometimes misleading
The real power of regression is its extension to models that include two or more explanatory variables.

Two motivations for adding explanatory variables:

1. Decreasing the size of the residuals, thereby accounting for more of the variation of the response variable; predictions of $y$ become more accurate.
2. Allowing us to hold statistically constant these additional variables, producing a potentially more accurate assessment of the effect of $x$ on $y$.
   - We can assess the potential role of *lurking variables* (when we can measure them).

I shall primarily consider the case of two predictor variables, $x_1$ and $x_2$. 
Introduction to Multiple Regression

Least-Squares Fit for Multiple Regression

- Recall that simple regression specified the equation of the line as:

\[ \hat{y} = a + bx \]

- We can extend this idea to two predictors, where the linear equation will now specify a flat plane rather than a line:

\[ \hat{y} = b_0 + b_1 x_1 + b_2 x_2 \]

- As in simple regression, we want the values of the coefficients \( b_0, b_1, \) and \( b_2 \) that minimize the sum of squared residuals, \( \sum e_i^2 \), where \( e_i = y_i - \hat{y}_i \).

  - We let a computer program find the coefficients for us.
Introduction to Multiple Regression
The Multiple Regression Plane

- $b_1$ and $b_2$ represent the partial slopes for $x_1$ and $x_2$ respectively.
- For each observation, the values of $x_1$, $x_2$ and $y$ are plotted in 3-dimensional space.
- The regression plane is fit by minimizing the sum of the squared residuals.
- $e_i$ (the residual) is now the vertical distance of the observed value $y$ from the fitted value $\hat{y}$ on the plane.

From Fox (1997, Figure 5.5 p.98)
Interpreting the Regression Coefficients

- The equation of the regression plane is

\[ \hat{y} = b_0 + b_1 x_1 + b_2 x_2 \]

where

- \( b_0 \) is the y-intercept of the plane
- \( b_1 \) is the slope of the plane in the \( x_1 \) direction, holding \( x_2 \) constant
- \( b_2 \) is the slope of the plane in the \( x_2 \) direction, holding \( x_1 \) constant

- The slope \( b_1 \) can also be interpreted as the average change in \( y \) associated with a one-unit increase in \( x_1 \), holding the value of \( x_2 \) constant.

- Likewise, the slope \( b_2 \) can be interpreted as the average change in \( y \) associated with a one-unit increase in \( x_2 \), holding the value of \( x_1 \) constant.

- Still another way would be to say that \( b_1 \) represents the “effect” of \( x_1 \) on \( y \), “controlling” for \( x_2 \) – but the causal language may not be appropriate.
Interpreting the Regression Coefficients

\( R^2 \) and Multiple Correlation \( R \)

- As in simple regression, the fit of the model may be summarized by the \( R^2 \):

\[
R^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum \text{residual}_i^2}{\sum (y_i - \bar{y})^2}
\]

where

\[
\hat{y} = a + b_1x_1 + b_2x_2
\]

- \( R^2 \) represents the proportion of the variation in \( y \) accounted for by its linear regression on \( x_1 \) and \( x_2 \).
- The positive square root \( R \) of \( R^2 \) is called the coefficient of multiple correlation or coefficient of determination.
Interpreting the Regression Coefficients

Measures of Model Fit: $R^2$ and $s_e$

Another terminology for $R^2$ is to use the notation:

\[ SST \text{ (total)} = \sum (y_i - \bar{y})^2 \]
\[ SSR \text{ (explained)} = \sum (\hat{y}_i - \bar{y})^2 \]
\[ SSE \text{ (residual)} = \sum \text{residual}_i^2 \]

Then $R^2 = SSR/SST$ or $R^2 = 1 - SSE/SST$.

The standard error $s_e$ of the regression can also be calculated for multiple regression as:

\[ s_e = \sqrt{\frac{SSE}{n - k - 1}} \]

where $n$ is the number of cases and $k$ is the number of explanatory variables (i.e., $k = 2$ in this case). $s_e$ represents the “typical” distance of an observation from the regression plane.
Using Multiple Regression
Case Study: Stata Program for Prestige Model

.* in Stata
.replace income=income/1000
income was int now float
(102 real changes made)

.reg prestige education income

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 102</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>23856.5752</td>
<td>2</td>
<td>11928.2876</td>
<td>F( 2, 99) = 195.55</td>
</tr>
<tr>
<td>Residual</td>
<td>6038.85087</td>
<td>99</td>
<td>60.9984937</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>29895.4261</td>
<td>101</td>
<td>295.994318</td>
<td>R-squared = 0.7980</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.7939</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 7.8102</td>
</tr>
</tbody>
</table>

| prestige | Coef.    | Std. Err. | t     | P>|t|     | [95% Conf. Interval] |
|----------|----------|-----------|-------|---------|---------------------|
| education| 4.137444 | .348912   | 11.86 | 0.000   | 3.445127 4.829762  |
| income   | 1.361166 | .2242121  | 6.07  | 0.000   | .9162804 1.806051  |
| _cons    | -6.847778| 3.218977  | -2.13 | 0.036   | -13.23493 -.4606292|

82 / 147
Case Study: Prestige Model

We regress prestige \((y)\) on income \((X1)\) and education \((X2)\)

\[
\hat{y} = b_0 + b_1 x_1 + b_2 x_2 \\
= -6.85 + 1.36 \times \text{income} + 4.14 \times \text{education}
\]

Interpretation:

- \(b_1 = 1.36\) (slope for income): An increase of $1000 in income, holding education constant, is accompanied on average by an increase of 1.36 prestige points.
- \(b_2 = 4.14\) (slope for education): An increase of one year of education, holding income constant, is associated on average by an increase of 4.14 prestige points.
Case Study: Prestige Model – Marginal and Partial Relationships

- When prestige is regressed on *income and education*, the income coefficient is $b_1 = 1.36$.
  - Recall that in multiple regression $b_1$ represents a partial relationship.
- When prestige is regressed on *income alone*, the income coefficient is $b = 2.90$.
  - In simple regression $b$ represents a *marginal* relationship.
- Note that $b$ is more than twice as high for the marginal relationship.
  - Because income and education levels are substantially correlated ($r = .58$), when education is omitted, and therefore left free to vary rather than held constant, part of its effect on prestige is absorbed by income.
Using Multiple Regression

Standardized Coefficients: Stata Program with `beta` Option

```
. * Income transformed earlier into 1000 dollars units
. reg prestige education income, beta

Source | SS       df       MS               Number of obs = 102
--------+-------------------------------------------------------------
Model   | 23856.5752    2   11928.2876         F(  2,    99) = 195.55
Residual| 6038.85087    99  60.9984937        Prob > F = 0.0000
--------+-------------------------------------------------------------
Total   | 29895.4261   101  295.994318       R-squared = 0.7980
         |                   |                Adj R-squared = 0.7939
         |                   |                Root MSE = 7.8102

------------------------------------------------------------------------------
prestige | Coef.  Std. Err.     t    P>|t|   Beta
--------+------------------------------------------------------------------------
education|   4.137444   .348912     11.86  0.000   .6561537
income   |    1.361166   .224212     6.07  0.000   .3359243
_cons    |  -6.847778   3.218977    -2.13  0.036           
------------------------------------------------------------------------------
```

. * Income transformed earlier into 1000 dollars units
. reg prestige education income, beta
Using Multiple Regression

Interpretation of the Standardized Coefficients

- The *standardized regression coefficient* $b_k^*$ of an explanatory variable $x_k$ (sometimes called *beta*) is calculated as

  $$b_k^* = b_k \frac{s_{x_k}}{s_y}$$

- $b_k^*$ represents the change in $z_y$ (standardized $y$) associated with a one unit increase in $z_{x_k}$ (standardized $x_k$).
- For example, in the regression of prestige on education and income:
  - $b_1^* = .656$: An increase of one SD in education, holding income constant, is accompanied on average by an increase of .656 SD in prestige.
  - $b_2^* = .336$: An increase of one SD in income, holding education constant, is accompanied on average by an increase of .336 SD in prestige.
- We can say that the effect of education is almost twice as large as the effect of income in explaining occupational prestige.
- Note that standardized coefficients permit comparing the effects of explanatory variables that are in different metrics.
The next two slides show a common way of presenting the results of multiple regression analysis in research reports or manuscripts submitted for publication.

The first table shows descriptive statistics (mean, SD, min, max, and bivariate correlations). The second table shows the regression results.

You can use your word processor or a spreadsheet program such as Excel to construct the tables, or you can use LaTeX as I did.

The overarching goal is to make the tables sufficiently self-explanatory that the reader can understand the “story” of the analysis at a glance, without having to read the text of the paper.

“t ratios” and significance levels (stars) are discussed later in the course.
# Table 1. Basic Statistics for the Regression of Occupational Prestige on Occupational Characteristics for 102 Canadian Occupations

<table>
<thead>
<tr>
<th></th>
<th>Prestige</th>
<th>Education</th>
<th>Income (1000's)</th>
<th>Pct women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prestige</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>0.850</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>0.715</td>
<td>0.578</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Pct women</td>
<td>-0.118</td>
<td>0.062</td>
<td>-0.441</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Mean:
- Prestige: 46.833
- Education: 10.738
- Income: 6.798
- Pct women: 28.979

Std. Dev.:
- Prestige: 17.204
- Education: 2.728
- Income: 4.246
- Pct women: 31.725

Min:
- Prestige: 14.8
- Education: 6.38
- Income: 0.611
- Pct women: 0

Max:
- Prestige: 87.2
- Education: 15.97
- Income: 25.879
- Pct women: 97.51
### Table 2. Unstandardized Coefficients for the Regression of Occupational Prestige on Occupational Characteristics for 102 Canadian Occupations (t Ratios in Parentheses)

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Std Coef</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$-10.732^{**}$</td>
<td>$-6.848^*$</td>
<td>$-6.794^*$</td>
<td>$- - -$</td>
</tr>
<tr>
<td></td>
<td>($-2.919$)</td>
<td>($-2.127$)</td>
<td>($-2.098$)</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>$5.361^{***}$</td>
<td>$4.137^{***}$</td>
<td>$4.187^{***}$</td>
<td>$0.664$</td>
</tr>
<tr>
<td></td>
<td>($16.148$)</td>
<td>($11.858$)</td>
<td>($10.771$)</td>
<td></td>
</tr>
<tr>
<td>Income (1000’s)</td>
<td>$- - -$</td>
<td>$0.224^{***}$</td>
<td>$1.314^{***}$</td>
<td>$0.324$</td>
</tr>
<tr>
<td></td>
<td>($6.071$)</td>
<td>($4.729$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pct women</td>
<td>$- - -$</td>
<td>$- - -$</td>
<td>$-.009$</td>
<td>$-.016$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>($-.293$)</td>
<td></td>
</tr>
</tbody>
</table>

| $R^2$            | .723             | .794             | .798             |          |
| $s_e$            | 9.103            | 7.810            | 7.846            |          |

*** $p < .001$; ** $p < .01$; * $p < .05$ (2-tailed tests)
Using Multiple Regression
Case Study: Development Score – Identifying Lurking Variables with Multiple Regression

12 children between the ages of 3 and 11 were tested for cognitive development (D-score).³

<table>
<thead>
<tr>
<th></th>
<th>OBS</th>
<th>DSCORE</th>
<th>AGE</th>
<th>BOY</th>
<th>BOY$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000</td>
<td>8.610</td>
<td>3.330</td>
<td>0.000</td>
<td>G</td>
</tr>
<tr>
<td>2</td>
<td>2.000</td>
<td>9.400</td>
<td>3.250</td>
<td>0.000</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>3.000</td>
<td>9.860</td>
<td>3.920</td>
<td>0.000</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>4.000</td>
<td>9.910</td>
<td>3.500</td>
<td>0.000</td>
<td>G</td>
</tr>
<tr>
<td>5</td>
<td>5.000</td>
<td>10.530</td>
<td>4.330</td>
<td>1.000</td>
<td>B</td>
</tr>
<tr>
<td>6</td>
<td>6.000</td>
<td>10.610</td>
<td>4.920</td>
<td>0.000</td>
<td>G</td>
</tr>
<tr>
<td>7</td>
<td>7.000</td>
<td>10.590</td>
<td>6.080</td>
<td>1.000</td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>8.000</td>
<td>13.280</td>
<td>7.420</td>
<td>1.000</td>
<td>B</td>
</tr>
<tr>
<td>9</td>
<td>9.000</td>
<td>12.760</td>
<td>8.330</td>
<td>1.000</td>
<td>B</td>
</tr>
<tr>
<td>10</td>
<td>10.000</td>
<td>13.440</td>
<td>8.000</td>
<td>0.000</td>
<td>G</td>
</tr>
<tr>
<td>11</td>
<td>11.000</td>
<td>14.270</td>
<td>9.250</td>
<td>1.000</td>
<td>B</td>
</tr>
<tr>
<td>12</td>
<td>12.000</td>
<td>14.130</td>
<td>10.750</td>
<td>1.000</td>
<td>B</td>
</tr>
</tbody>
</table>

Using Multiple Regression
Simple Regression of D-Score on BOY (=1 for boy, 0 for girl) in SYSTAT

Pearson Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>DSCORE</th>
<th>AGE</th>
<th>BOY</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSCORE</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>0.957</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>BOY</td>
<td>0.600</td>
<td>0.647</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Simple Linear Regression

Dep Var: DSCORE  N: 12  Multiple R: 0.600  Squared multiple R: 0.360
Adjusted squared multiple R: 0.296  Standard error of estimate: 1.671

<table>
<thead>
<tr>
<th>Effect</th>
<th>Coefficient</th>
<th>Std Error</th>
<th>Std Coef</th>
<th>Tolerance</th>
<th>t</th>
<th>P(2 Tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>10.305</td>
<td>0.682</td>
<td>0.000</td>
<td>.</td>
<td>15.109</td>
<td>0.000</td>
</tr>
<tr>
<td>BOY</td>
<td>2.288</td>
<td>0.965</td>
<td>0.600</td>
<td>1.000</td>
<td>2.372</td>
<td>0.039</td>
</tr>
</tbody>
</table>

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum-of-Squares</th>
<th>df</th>
<th>Mean-Square</th>
<th>F-ratio</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>15.709</td>
<td>1</td>
<td>15.709</td>
<td>5.629</td>
<td>0.039</td>
</tr>
<tr>
<td>Residual</td>
<td>27.910</td>
<td>10</td>
<td>2.791</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Using Multiple Regression

Scatterplot of D-Score on Age, Coded for Sex (G = Girl, B = Boy)
Using Multiple Regression

Using Multiple Regression to Control for Lurking Age

Multiple Regression of D-Score on BOY and AGE

Dep Var: DSCORE  N: 12  Multiple R: 0.958  Squared multiple R: 0.917
Adjusted squared multiple R: 0.899  Standard error of estimate: 0.634

<table>
<thead>
<tr>
<th>Effect</th>
<th>Coefficient</th>
<th>Std Error</th>
<th>Std Coef</th>
<th>Tolerance</th>
<th>t</th>
<th>P(2 Tail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>6.927</td>
<td>0.506</td>
<td>0.000</td>
<td>.</td>
<td>13.697</td>
<td>0.000</td>
</tr>
<tr>
<td>BOY</td>
<td>-0.126</td>
<td>0.480</td>
<td>-0.033</td>
<td>0.581</td>
<td>-0.262</td>
<td>0.799</td>
</tr>
<tr>
<td>AGE</td>
<td>0.753</td>
<td>0.097</td>
<td>0.979</td>
<td>0.581</td>
<td>7.775</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum-of-Squares</th>
<th>df</th>
<th>Mean-Square</th>
<th>F-ratio</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>40.002</td>
<td>2</td>
<td>20.001</td>
<td>49.765</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual</td>
<td>3.617</td>
<td>9</td>
<td>0.402</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Using Multiple Regression
Case Study: Jasso’s Marital Coital Frequency Study


▶ Jasso uses panel data to estimate age and period effects – controlling for cohort effects – on frequency of sexual relations for married couples from 1970 to 1975.

▶ Frequency is recorded as number of times the couple has had sex in the previous four weeks.

▶ Major Findings:

▶ Negative period effect – frequency declines from 1970 to 1975.
▶ Wife’s age had a positive effect on frequency.
▶ Both findings differ significantly from previous research in the area.
Using Multiple Regression


- Kahn and Udry claim to find two kinds of problems in Jasso’s study:
  - Failure to check the data for influential outliers:
    - Four cases were seemingly miscoded 88 (must be missing data – coded 99 – since no other value was higher than 63 and 99.5% were less than 40).
    - There were 4 additional serious outliers.
  - Jasso missed an important interaction between length of marriage and wife’s age.

- Dropping the 8 miscodes and outliers (from a sample of more than 2000) and adding the interaction drastically changes the findings (next slide).
## Using Multiple Regression

### Determinants of Marital Coital Frequency (Times in Previous 4 Weeks)

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>−.72***</td>
<td>−.67***</td>
<td>−3.06**</td>
<td>−.08</td>
</tr>
<tr>
<td>Log Wife’s Age</td>
<td>27.61**</td>
<td>13.56</td>
<td>29.49</td>
<td>−1.62</td>
</tr>
<tr>
<td>Log Husband’s Age</td>
<td>−6.43</td>
<td>7.87</td>
<td>57.89</td>
<td>−5.23</td>
</tr>
<tr>
<td>Log Marital Duration</td>
<td>−1.50***</td>
<td>−1.56***</td>
<td>−1.51*</td>
<td>1.29</td>
</tr>
<tr>
<td>Wife Pregnant</td>
<td>−3.71***</td>
<td>−3.74***</td>
<td>−2.88***</td>
<td>−3.95*</td>
</tr>
<tr>
<td>Child Under 6</td>
<td>−.56**</td>
<td>−.68***</td>
<td>−2.91***</td>
<td>−.55**</td>
</tr>
<tr>
<td>Wife Employed</td>
<td>.37</td>
<td>.23</td>
<td>.86</td>
<td>.02</td>
</tr>
<tr>
<td>Husband Employed</td>
<td>−1.28**</td>
<td>−1.10**</td>
<td>−4.11***</td>
<td>−.38</td>
</tr>
</tbody>
</table>

| $R^2$                | .0475   | .0612   | .2172   | .0411   |
| $N$                  | 2062    | 2055    | 243     | 1812    |

**Model 1:** Jasso’s original results  
**Model 2:** 4 miscodes & 4 outliers dropped  
**Model 3:** Marital duration ≤ 2  
**Model 4:** Marital duration > 2

Adapted from Kahn, J.R. and J.R. Udry (1986, Table 1 p.735)
Regression with Categorical Explanatory Variables

- Regression analysis can be extended to include categorical response variables and categorical explanatory variables.
  - The former is outside the scope of this course because it no longer uses least squares estimation, but the latter is possible in least squares regression.
- We can explore both main effects and interaction effects with respect to a categorical independent variable.
  - A main effect corresponds to a difference in the intercepts of the regression lines fitted to the different categories of the independent variable.
  - If there are interaction effects, the regression lines differ across categories in terms of both intercepts and slopes.
Regression with Categorical Explanatory Variables

- Linear regression can be extended to accommodate categorical variables using *indicator ("dummy") variables* as regressors.

- A categorical variable is represented by an indicator (dummy) regressor $D$, (coded 1 for one category, 0 for the other):

$$\hat{Y}_i = a + b_1X_i + b_2D_i$$

- This fits two regression lines with the same slope but different intercepts. The coefficient $b_2$ represents the constant separation between the two regression lines:

$$\hat{Y}_i = a + b_1X_i + b_2(0) = a + b_1X_i$$
$$\hat{Y}_i = a + b_1X_i + b_2(1) = (a + b_2) + b_1X_i$$
Regression with Categorical Explanatory Variables

Figure (a) is an example of when failing to account for a categorical variable (gender) does not produce significantly different results, either in terms of the intercept or the slope.

In Figure (b) the dummy regressor captures a significant difference in intercepts. Failing to include gender gives a negative slope for the relationship between education and income (the dotted line) when in fact it should be positive for both men and women.

From Fox (1997, Figure 7.1 p.137)
Indicator (dummy) regressors are easily extended to independent variables with more than two categories. As with the two-category case, one of the categories is a reference group.

Therefore, there is *one less* dummy regressors than there are categories.

Below education is divided into *four* categories represented by *three* regressors, $D_1, D_2$ and $D_3$

<table>
<thead>
<tr>
<th>Category</th>
<th>$D_1$</th>
<th>$D_2$</th>
<th>$D_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>College degree</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Some college</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>High school</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Less than h. s.</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Regression with Categorical Explanatory Variables

- The model with one quantitative predictor $x_1$ (age) takes the following form:

$$\hat{y}_i = a + b_1 x_i + b_2 D_{i1} + b_3 D_{i2} + b_4 D_{i3}$$

- This produces four parallel regression lines:

  College degree \quad \hat{Y}_i = (a + b_2) + b_1 x_i

  Some college \quad \hat{Y}_i = (a + b_3) + b_1 x_i

  High school \quad \hat{Y}_i = (a + b_4) + b_1 x_i

  Less than h. s. \quad \hat{Y}_i = a + b_1 x_i

- These lines are *different only in terms of their intercepts* – i.e., $b_2, b_3,$ and $b_4$ tell us the effects of each education level compared to “less than high school”, when holding age constant.
Regression with Categorical Explanatory Variables
Prestige and Occupation Type

- Below is the output from a model regressing prestige on income and type
- Type is a three category variable measured by two indicators (the reference category is “blue collar”)

```r
> library(car)
> data(Prestige)
> attach(Prestige)
> model1<-lm(prestige~income+type)
> model1

Call:
  lm(formula = prestige ~ income + type)

Coefficients:
(Intercept) income typeprof typewc
  27.997057  0.001401  25.055474  7.167155
```

102 / 147
Regression with Categorical Explanatory Variables
Prestige and Occupation Type

Coefficients:
(Intercept)     income     typeprof     typewc
27.997057       0.001401   25.055474   7.167155

- This is a main effects model – i.e., no interaction terms – so the lines differ in terms of intercepts only
  - In other words, the income slope (.0014) is the same for each occupation type
  - The “intercept” coefficient is the intercept for the reference category only (blue collar). To find the intercepts for the other occupation types we add their coefficients to the blue collar intercept

Intercepts for each occupation type:

<table>
<thead>
<tr>
<th>Occupation Type</th>
<th>Intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue Collar</td>
<td>27.997</td>
</tr>
<tr>
<td>Professional</td>
<td>27.997 + 25.055 = 53.05</td>
</tr>
<tr>
<td>White Collar</td>
<td>27.997 + 7.167 = 35.16</td>
</tr>
</tbody>
</table>
Regression with Categorical Explanatory Variables
Prestige and Occupation Type

- Although the difference in intercept model seems to capture the trend in the data better than one without the dummy regressors for occupation type (the solid black line), there appears to be a difference in slopes that is not accounted for.
> library(car)
> # the car library contains a "fancy"
> # scatterplot function that plots lines
> # conditional on a categorical variable
> scatterplot(prestige~income|type, smooth=FALSE, reg.line=FALSE, cex=1.5)
> # I’ve put smooth=FALSE and reg.line=FALSE
> # because otherwise I’ll have lines with
> # different slopes. I prefer to specify the lines
> # myself from the regression output:
> abline(53.05, .001401, col="green", lwd=2, lty=2)
> abline(27.997, .001401, col="red", lwd=2)
> abline(35.16, .001401, col="blue", lwd=2, lty=3)
> abline(lm(prestige~income), lwd=2)
> legend(10000, 35, "All occupations", col=1, lty=1, lwd=2)
Regression with Categorical Explanatory Variables

Interaction Effects

- Dummy variables test for differences in *level* but not in *slope*
- We may, however, be interested in whether slopes differ. For example, we may want to test whether *age effects on income* are different for men and women.
- When the partial effect of one variable depends on the value of another (a *moderating variable*), the two variables are said to interact
- We could fit separate regressions for men and women, but this does not allow for a formal statistical test of the differences
  - Adding interaction terms to the model allows us to formally test for differences in slope
Differences in slope can be captured in a single regression model by including interaction terms which are the product of the regressors for the two variables.

The interaction regressor in the model below is $XD$:

$$\hat{Y}_i = a + b_1X_i + b_2D_i + b_3(X_iD_i)$$

The parameters $a$ and $b_1$ are the intercept and slope for the reference group that was coded 0 (women).

The intercept for the other group (men) is $a + b_2$; the slope is $b_1 + b_3$. 
Regression with Categorical Explanatory Variables

Interaction Effects

- We can write out the equations as follows:

for $D = 0$ (women):

$$\hat{Y}_i = a + b_1 X_i + b_2(0) + b_3(X_i \times 0)$$
$$= a + b_1 X_i$$

for $D = 1$ (men):

$$\hat{Y}_i = a + b_1 X_i + b_2(1) + b_3(X_i \times 1)$$
$$= (a + b_2) + (b_1 + b_3)X_i$$

- Unlike in the non-interaction model the regression lines are not parallel so we cannot interpret $b_2$ as the unqualified partial effect of gender controlling for age (it is the effect only at $X = 0$).

- Similarly, $b_1$ is not the unqualified effect of age controlling for gender (it is the effect only among women).
Regression with Categorical Explanatory Variables

Interaction Effects: Prestige Data

In this example Occupational prestige is regressed on income and type of occupation (three category variable with “blue collar” as the reference), including an interaction between the two predictors

```r
> library(car)
> data(Prestige)
> attach(Prestige)
> model1<-lm(prestige~income*type)
> model1

Call:
  lm(formula = prestige ~ income * type)

Coefficients:
                Estimate
  (Intercept)    13.904517
      income      0.004023
     typeprof     45.019022
    typewc       18.980739
     income:typeprof -0.003178
    income:typewc   -0.002171
```

Regression with Categorical Explanatory Variables

Interaction Effects: Prestige Data

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>income</th>
<th>typeprof</th>
<th>typewc</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>13.904517</td>
<td>0.004023</td>
<td>45.019022</td>
</tr>
<tr>
<td>income:typeprof</td>
<td>-0.003178</td>
<td></td>
<td></td>
</tr>
<tr>
<td>income:typewc</td>
<td>-0.002171</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From the regression output above, we can find the individual regression lines of prestige regressed on income for each occupation type simply by summing together the appropriate coefficients:

<table>
<thead>
<tr>
<th>Occupation Type</th>
<th>Intercept</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue Collar</td>
<td>13.905</td>
<td>0.0040</td>
</tr>
<tr>
<td>Professional</td>
<td>13.905 + 45.019 = 58.92</td>
<td>0.0040 − 0.0032 = 0.0008</td>
</tr>
<tr>
<td>White collar</td>
<td>13.905 + 18.981 = 32.89</td>
<td>0.0040 − 0.0022 = 0.0018</td>
</tr>
</tbody>
</table>
Regression with Categorical Explanatory Variables

Interaction Effects: Prestige Data

From the graph we see that there is an interaction between income and occupational type:

- Effect of income on prestige is strongest for bc, less for wc, and least for prof.

> scatterplot(prestige~income|type, smooth=FALSE, cex=1.5)
Data Analysis for Two-Way Tables

Outline

- Start with a discussion of looking at simple relationships between categorical dependent and independent variables
  - Converting to proportions, percentages and ratios
  - Contingency tables (cross-tabulation)
  - Partial tables for multivariate relationships
    - Problem of Simpson’s Paradox
- Revisit multiple regression, considering categorical explanatory variables (dummy variables)
  - Look at both main effects (difference in intercepts) and interaction effects (difference in slopes)
- Question of causation
  - Establishing causation
Converting Raw Numbers to Proportions, Percentages or Ratios

- Raw numbers (i.e., simply the number of cases in a particular category) can often be difficult to interpret, especially if there are several categories and an awkward number of total cases.

- Percentages and proportions (and ratios) are also much easier to compare both with variables and across variables.

- Cautions:
  - We must be careful not to over-interprete percentages (or proportions) based on small total sample sizes.
  - Tables containing proportions or percentages should always contain the sample size (n).
The Two-Way Table

Proportions

▶ It is often useful to convert raw numbers to proportions, especially when comparing categories
▶ The formula to find a proportion is simply:

\[ p = \frac{\text{count in category}}{n \ (\text{total sample size})} \]

▶ If there are 60 students in the class, 46 of whom are women, what is the proportion of women? Here count = 46 and \( n = 60 \), therefore

\[ p = \frac{46}{60} = .77 \]

▶ In other words, the proportion of women in the class is .77.
▶ Since the total equals 1, the proportion of men is \( 1 - .77 = .23 \)
The Two-Way Table

Percentages

- A percentage gives us the same information as a proportion except we now multiply the proportion by 100:

\[
\text{count in category} \div n \times 100
\]

- Continuing from the proportions example, if there are 60 students in the class, 46 of whom are women, what is the percentage of women? Here count = 46 and \( n = 60 \), therefore

\[
p = \frac{46}{60} \times 100 = 77\%
\]

- In other words, 77% of the class are women

- Since the total for percentages equals 100, the percentage of men is \( 100 - 77 = 23\% \)
The Two-Way Table

Ratios

- Ratios allow us to directly compare the relative number of cases in one category compared to another:

\[
\text{Ratio of } a \text{ to } b = \frac{\text{count in category } a}{\text{count in category } b}
\]

- If there are 46 women and 14 men, the ratio of women to men is:

\[
\text{Ratio women:men} = \frac{\text{count in category } a}{\text{count in category } b} = \frac{46}{14} = 3.28
\]

- In other words, there are more than 3 times as many women as men in the class.
Bar charts should be used over pie charts – research indicates that people are better able to judge the relative difference in size of straight lines than pie shapes.

```r
type <- c("bc", "prof", "wc")
plot(type, main="Bar chart for occupation type")
pie(table(type), main="Pie chart for occupation type")
```
The Two-Way Table

An Alternative to the Bar graph & Pie Chart: the Dot Chart

► The dot chart is an alternative to the bar graph or pie chart. It is especially useful for displaying multivariate tables (see later)

```
> par(pty="s")
> dotchart(table(type), main="Dot chart for occupation type")
```
The Two-Way Table

Critique of pie chart by Cleveland (1994, Figures 4.19 p. 262 and 4.20 p. 263)

4.19 PIE CHART. The pie chart falls in the category of a pop chart—a graphical method used frequently in the mass media and certain business presentations but far less in science and technology. Both table look-up and pattern perception are less efficient for pie charts than for dot plots.

4.20 DOT PLOT. The data from Figure 4.19 are graphed by a dot plot. Patterns emerge that cannot be decoded from Figure 4.19.
The Two-Way Table
Contingency Tables

- Also called Cross-Tabulations
- Display relationships between categorical variables
  - Remember: With qualitative variables we talk of relationships or associations, but NOT correlations.
- Cells of the table represent the number of observations that fall simultaneously into a particular combination of two categories of the two variables.
- Tables can be presented in several ways:
  1. Raw counts or frequencies
  2. Percentages (or proportions) of total $N$
  3. Percentages (or proportions) of column $Ns$
  4. Percentages (or proportions) of row $Ns$
Joint and Marginal Distributions

Parts of the Contingency Table

- Marginal Distributions
  - The last row and column of the table.
  - The distribution of each variable taken separately.

- Conditional Distributions
  - These cells tell us about the relationship between the variables.

- Row and Column Variables
  - No rule identifies which variable (the explanatory variable, or the response variable) is the row variable and column variable.
  - Tables can be constructed either way, but we must be sure to percentage the table correctly.
Joint and Marginal Distributions

Contingency Table Using Frequencies

- Here gender is the explanatory variable, in this case the column; “frequent binge drinking”, the row variable, is the response variable.
- The table could also be set up in the opposite direction – direction only matters when calculating percentages.

<table>
<thead>
<tr>
<th>Frequent binge drinker</th>
<th>Men</th>
<th>Women</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>1,630</td>
<td>1,684</td>
<td>3,314</td>
</tr>
<tr>
<td>No</td>
<td>5,550</td>
<td>8,232</td>
<td>13,782</td>
</tr>
<tr>
<td>Total</td>
<td>7,180</td>
<td>9,916</td>
<td>17,096</td>
</tr>
</tbody>
</table>

- The “Total” row and “Total” column contain the marginal distributions of the variables.
- The numbers in the cells inside the table (i.e., those aside from the total row and column) represent the *joint frequencies* of the two variables.
The joint distribution of the variables is given by the proportions (or percentages) of cases that have a given value of one variable and a given value of the other variable.

The joint distribution is obtained by dividing all frequencies by the total number of cases ($n$):

<table>
<thead>
<tr>
<th>Frequent binge drinker</th>
<th>Men</th>
<th>Women</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>.095</td>
<td>.099</td>
<td>.194</td>
</tr>
<tr>
<td>No</td>
<td>.325</td>
<td>.482</td>
<td>.806</td>
</tr>
<tr>
<td>Total</td>
<td>.420</td>
<td>.580</td>
<td>1.000</td>
</tr>
</tbody>
</table>

The joint distribution consists of the proportions in the four inner cells. These proportions add up to 1.

The “Total” row and “Total” column contain the marginal distributions of the variables.
To see relationships between variables one calculates the *conditional distribution* of the response variable given values of the explanatory variable.

General Principle: Calculate percentages *within categories of the explanatory variable* to make comparisons between (or among) categories of the explanatory variable.

- If the explanatory variable is the column variable in the table, then percentages add to 100 vertically.
- If the explanatory variable is the row variable, then percentages add to 100 horizontally.
Describing Relationships: Conditional Distributions

- We can facilitate the comparison of the binge drinking habits of men with women by calculating the percentage of respondents falling in both categories of binge drinking habits for each gender.

<table>
<thead>
<tr>
<th>Frequent binge drinker</th>
<th>Men (%)</th>
<th>Women (%)</th>
<th>Total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>22.7</td>
<td>17.0</td>
<td>19.4</td>
</tr>
<tr>
<td>No</td>
<td>77.3</td>
<td>83.0</td>
<td>80.6</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

- Since the explanatory variable (gender) is the column variable, we percentage down the columns.

- If gender were the row variable, we would percentage across the rows instead.
Comparing the conditional distributions of binge drinking for men and for women reveals the association between binge drinking and gender:

- The percentage of binge drinkers is 22.7% among men but only 17.0% among women
- This suggests that men have a greater propensity than women to binge drinking
Describing Relationships: Conditional Distributions

Relative Risk (RR) and Odds Ratio (OR)

- Two quantities that are commonly used to express the effect size of gender on binge drinking are:
  - The relative risk (RR) of binge drinking for men, relative to women, is
    \[
    RR = \frac{P_m}{P_w} = \frac{0.227}{0.170} = 1.34
    \]
    Men have a 34% greater risk of binge drinking than women.
  - The odds ratio (OR) of binge drinking for men, relative to women, is
    \[
    OR = \frac{P_m/(1 - P_m)}{P_w/(1 - P_w)} = \frac{P_m(1 - P_w)}{P_w(1 - P_m)} = \frac{0.227(1 - 0.170)}{0.170(1 - 0.227)} = 1.43
    \]
    The odds of binge drinking for men is 43% higher than for women.
  - The OR is widely used to report the effect of a risk factor in logistic regression.
Simpson’s Paradox

Multivariate Tables

- Bivariate tables show only *marginal relationships*
- Multivariate tables show *partial* relationships by using *partial* tables
  - *Partial tables* are simply bivariate tables broken down by categories of a control variable
  - Like in multiple regression, these tables allow us to statistically “hold a variable constant”
- Multivariate tables are illustrated later in connection with Simpson’d Paradox
Simpson’s Paradox

Representation with Divided Bar Chart (Cleveland 1994, Figure 4.21 p.266)

4.21 DIVIDED BAR CHART. A divided bar chart is used to show the percentage of the vote for three candidates in the 1984 New York Democratic primary election. The Mondale values are graphed by position along a common scale, but the Hart values and the Jackson values are not and our visual decoding of these latter two sets of values is less accurate than for the Mondale values.
4.22 MULTIWAY DOT PLOT. The data from Figure 4.21 are graphed by a multiway dot plot. Now the Hart values and the Jackson values are encoded by position along a common scale. Now we can perceive a Hart age pattern.
Simpson’s Paradox

- Observed associations can be misleading if we do not control for other variables that may be related.
- As we learned from regression, partial and marginal relationships can be different.
- When partial and marginal associations differ in direction, it is called Simpson’s paradox.
  - In other words, a lurking variable changes the direction of the relationship, leading to completely different conclusions.
Simpson’s Paradox
Example: Death Penalty and Race

▶ The data are on whether convicted murderers in US states get the death penalty, according to the race of the defendant and the race of the victim (Radelet, 1981)

▶ If we ignore the race of the victim, and look at the marginal association between race and the death penalty we see that white defendants were slightly more likely than blacks to get the death penalty:

<table>
<thead>
<tr>
<th>Race</th>
<th>Death Penalty</th>
<th>Total (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>white</td>
<td>11.9</td>
<td>88.1</td>
</tr>
<tr>
<td>black</td>
<td>10.2</td>
<td>89.8</td>
</tr>
</tbody>
</table>
Simpson’s Paradox
An example: Death Penalty and Race

- Results differ if we control for race of the victim
- We now see that blacks were generally more likely than whites to get the death penalty, both for white victims and for black victims

<table>
<thead>
<tr>
<th>White Victim</th>
<th>Death Penalty</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>Yes</td>
<td>No</td>
<td>Total (N)</td>
</tr>
<tr>
<td>white</td>
<td>12.6</td>
<td>87.4</td>
<td>100.0% (151)</td>
</tr>
<tr>
<td>black</td>
<td>17.5</td>
<td>82.5</td>
<td>100.0% (63)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Black Victim</th>
<th>Death Penalty</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>Yes</td>
<td>No</td>
<td>Total (N)</td>
</tr>
<tr>
<td>white</td>
<td>0.0</td>
<td>100.0</td>
<td>100.0% (9)</td>
</tr>
<tr>
<td>black</td>
<td>5.8</td>
<td>94.2</td>
<td>100.0% (103)</td>
</tr>
</tbody>
</table>

- We also notice that if we control for race of the defendant that the death penalty was much more likely if the victim was white
Simpson’s Paradox

Another Example: UC Berkely Sex Bias Case – Wikipedia at “Simpson’s Paradox”

- Bias against women in graduate school admissions (Fall 1973)

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th></th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applicants</td>
<td>% admitted</td>
<td>Applicants</td>
<td>% admitted</td>
</tr>
<tr>
<td>8442</td>
<td>44</td>
<td>4321</td>
<td>35</td>
</tr>
</tbody>
</table>

- Breaking down by department removes bias

<table>
<thead>
<tr>
<th>Major</th>
<th>Men</th>
<th></th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applicants</td>
<td>% admitted</td>
<td>Applicants</td>
<td>% admitted</td>
</tr>
<tr>
<td>A</td>
<td>825</td>
<td>62</td>
<td>108</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>63</td>
<td>25</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37</td>
<td>593</td>
</tr>
<tr>
<td>D</td>
<td>417</td>
<td>33</td>
<td>375</td>
</tr>
<tr>
<td>E</td>
<td>191</td>
<td>28</td>
<td>393</td>
</tr>
<tr>
<td>F</td>
<td>272</td>
<td>6</td>
<td>341</td>
</tr>
</tbody>
</table>
Using Two-Way Tables

Would You Allow and Antireligionist to Teach in College or University?

```
. tab degree colath, chi2

<table>
<thead>
<tr>
<th></th>
<th>ALLOW</th>
<th>ANTI-RELIGIONIST TO TEACH</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS HIGHEST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LT HIGH SCHOOL</td>
<td>105</td>
<td>158</td>
</tr>
<tr>
<td>HIGH SCHOOL</td>
<td>524</td>
<td>397</td>
</tr>
<tr>
<td>JUNIOR COLLEGE</td>
<td>91</td>
<td>53</td>
</tr>
<tr>
<td>BACHELOR</td>
<td>239</td>
<td>74</td>
</tr>
<tr>
<td>GRADUATE</td>
<td>108</td>
<td>22</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1,067</strong></td>
<td><strong>704</strong></td>
</tr>
</tbody>
</table>

Pearson chi2(4) = 112.4130    Pr = 0.000
```
Using Two-Way Tables
Allow Antireligionist to Teach in College or University? Row Percentages

```
.tab degree colath, chi2 row nofreq

<table>
<thead>
<tr>
<th></th>
<th>ALLOW ANTI-RELIGIONIST TO TEACH</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS HIGHEST</td>
<td>DEGREE</td>
</tr>
<tr>
<td>LT HIGH SCHOOL</td>
<td>39.92</td>
</tr>
<tr>
<td>HIGH SCHOOL</td>
<td>56.89</td>
</tr>
<tr>
<td>JUNIOR COLLEGE</td>
<td>63.19</td>
</tr>
<tr>
<td>BACHELOR</td>
<td>76.36</td>
</tr>
<tr>
<td>GRADUATE</td>
<td>83.08</td>
</tr>
<tr>
<td>Total</td>
<td>60.25</td>
</tr>
</tbody>
</table>

Pearson chi2(4) = 112.4130  Pr = 0.000
```
### Using Two-Way Tables

Allow Antireligionist to Teach in College or University? Table for publication combines two Stata runs (frequencies only and row percentages)

<table>
<thead>
<tr>
<th>Highest degree</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than High School</td>
<td>39.92</td>
<td>60.08</td>
<td>100.00</td>
<td>263</td>
</tr>
<tr>
<td>High School</td>
<td>56.89</td>
<td>43.11</td>
<td>100.00</td>
<td>921</td>
</tr>
<tr>
<td>Junior College</td>
<td>63.19</td>
<td>36.81</td>
<td>100.00</td>
<td>144</td>
</tr>
<tr>
<td>Bachelor</td>
<td>76.36</td>
<td>23.64</td>
<td>100.00</td>
<td>313</td>
</tr>
<tr>
<td>Graduate</td>
<td>83.08</td>
<td>16.92</td>
<td>100.00</td>
<td>130</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>60.25</td>
<td>39.75</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>1067</td>
<td>704</td>
<td></td>
<td>1771</td>
</tr>
</tbody>
</table>

**SOURCE:** General Social Survey 1998
Using Two-Way Tables
In Module 9 We Will Look at: Goodness of Fit and Significance of Association

### Jury Pool Age Distribution

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Observed</th>
<th>Census</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-19</td>
<td>23</td>
<td>0.061</td>
</tr>
<tr>
<td>20-24</td>
<td>96</td>
<td>0.150</td>
</tr>
<tr>
<td>25-29</td>
<td>134</td>
<td>0.135</td>
</tr>
<tr>
<td>30-39</td>
<td>293</td>
<td>0.217</td>
</tr>
<tr>
<td>40-49</td>
<td>297</td>
<td>0.153</td>
</tr>
<tr>
<td>50-64</td>
<td>380</td>
<td>0.182</td>
</tr>
<tr>
<td>65+</td>
<td>113</td>
<td>0.102</td>
</tr>
<tr>
<td>Total</td>
<td>1336</td>
<td>1.000</td>
</tr>
</tbody>
</table>

### Child Birth Defects by Mother’s Diabetic Status
Among Pima Indian Mothers

<table>
<thead>
<tr>
<th>Data</th>
<th>1+ Defect</th>
<th>No Defect</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nondiabetic</td>
<td>31</td>
<td>754</td>
<td>785</td>
</tr>
<tr>
<td>Prediabetic</td>
<td>13</td>
<td>362</td>
<td>375</td>
</tr>
<tr>
<td>Diabetic</td>
<td>9</td>
<td>38</td>
<td>47</td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>1154</td>
<td>1207</td>
</tr>
</tbody>
</table>

- **Goodness of fit**: Are the age distributions of the jury pool & of the population (according to census) different?

- **Statistical significance**: Is there a statistically significant association between diabetic status of mother & birth defect?
Cautions About Correlation and Regression

Simulating Curvilinear Relationship in Stata

. * in Stata
. clear all
. * next statement needed to set up data frame
. drawnorm xn, n(100)
(obs 100)
. gen x=-0.5+runiform()
. gen x2=x^2
. gen y=x^2
. twoway (scatter y x, xsize(3) ysize(3)
title("Plot of y=x^2"))(lfit y x)
. cor y x
(obs=100)

<table>
<thead>
<tr>
<th></th>
<th>y</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>0.0364</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

. cor y x2
(obs=100)

<table>
<thead>
<tr>
<th></th>
<th>y</th>
<th>x2</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Lurking Variables and the Question of Causation

Lurking Variables

▶ A lurking variable is a variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.
  ▶ E.g., College Board study of 15,941 high school graduates found strong correlation between how much math minority students took in high school and their later success in college. Does this finding mean that “math is the gatekeeper for success in college”?
    ▶ What might be lurking variables?
  ▶ Association does not imply causation. An association between an explanatory variable $x$ and a response variable $y$, even if very strong, is not in itself good evidence that changes in $x$ actually causes changes in $y$. 
Lurking Variables and the Question of Causation

The Question of Causation

- Goal in many studies is to establish causation, i.e. that change in the explanatory variable causes change in the response variable.
- But even strong association is insufficient to establish causation, because association can result from mechanisms other than a causal effect.
- Association between \( x \) and \( y \) can result from (a) causation, (b) common response (to a third variable), and (c) confounding.

![Diagram showing causation, common response, and confounding](image)
Lurking Variables and the Question of Causation
Explaining Association: Causation

▶ E.g., a study of Mexican American girls aged 9 to 12 years recorded BMI for both the girls and their mothers. Also recorded hours of TV, minutes of physical activity, intake of several kinds of foods. The strongest correlation was between daughter BMI and mother BMI ($r = 0.506$).
  ▶ This may represent direct causation, through heredity.
  ▶ But effect of mother’s BMI is only part of the explanation of the variation in daughter’s BMI ($r^2$ is only $0.506^2 = 0.256$).
  ▶ But other factors may affect daughter’s BMI. (Which?)
▶ In any case, it goes to show that you should choose your parents carefully!
Lurking Variables and the Question of Causation
Explaining Association: Common Response

- In *common response* the observed association between variables $x$ and $y$ is explained by a lurking variable $z$.
  - Both $x$ and $y$ change in response to change in $z$.
  - The common response creates an association between $x$ and $y$ even though there may be no direct causal link between $x$ and $y$.

- Examples:
  - SAT scores are associated with college GPA because students who are smart and hard-working tend to have both high SAT scores and high GPA.
  - Stock market return is associated with amount of money invested in stock mutual funds because investor sentiment affects both variables.
Lurking Variables and the Question of Causation
Explaining Association: Confounding

- Two variables are *confounded* when their effects on a response variable cannot be distinguished from each other.
  - The confounded variables may be either explanatory variables or lurking variables.

- Examples:
  - In the mother-daughter BMI association, effect of genes and environment are confounded.
  - People who attend religious service regularly live longer.
    - These people are also less likely to smoke, more likely to exercise, and less likely to be overweight. Effects of these good habits are confounded with effect of religious attendance.
Lurking Variables and the Question of Causation

Establishing Causation

- The strongest evidence for causation is provided by *experimentation*.
  - The researcher sets the value of $x$ and records response $y$.
- When experimentation is not possible, evidence for causation includes (e.g., smoking and cancer):
  - Association is strong.
  - Association is consistent across studies in different contexts.
  - Response is dose-dependent (higher dose $\rightarrow$ higher response)
  - Alleged cause precedes the effect in time.
    - Lung cancer among women rose with 30 years lag.
  - Alleged cause is plausible.
    - Tar causes cancer in animal experiments.
Appendix: Regression With Indicators

Regression With Indicators in Stata

. * in Stata
. clear all
. * Use menus to Open data set
. use "D:\soci708\data\data_from_car_Stata\Prestige.dta", clear
. tab type

<table>
<thead>
<tr>
<th>type</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>4</td>
<td>3.92</td>
<td>3.92</td>
</tr>
<tr>
<td>bc</td>
<td>44</td>
<td>43.14</td>
<td>47.06</td>
</tr>
<tr>
<td>prof</td>
<td>31</td>
<td>30.39</td>
<td>77.45</td>
</tr>
<tr>
<td>wc</td>
<td>23</td>
<td>22.55</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 102   | 100.00  |

. gen type_bc=0
. replace type_bc=. if type=="NA"
(4 real changes made, 4 to missing)
. gen type_prof=0
. replace type_prof=. if type=="NA"
(4 real changes made, 4 to missing)
. replace type_prof=1 if type=="prof"
(31 real changes made)
. gen type_wc=0
. replace type_wc=. if type=="NA"
(4 real changes made, 4 to missing)
. replace type_wc=1 if type=="wc"
(23 real changes made)
Appendix: Regression With Indicators

Regression With Indicators in Stata (Cont’d)

. gen incxprof=income*type_prof
   (4 missing values generated)
. gen incxwc=income*type_wc
   (4 missing values generated)

. reg prestige income type_prof type_wc incxprof incxwc

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 98</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>23487.6344</td>
<td>5</td>
<td>4697.52687</td>
<td>F( 5, 92) = 88.94</td>
</tr>
<tr>
<td>Residual</td>
<td>4859.2407</td>
<td>92</td>
<td>52.8178337</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>28346.8751</td>
<td>97</td>
<td>292.235825</td>
<td>R-squared = 0.8286</td>
</tr>
</tbody>
</table>

| prestige | Coef.     | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-----------|-----------|-----------|-------|-----|-----------------------|
| income    | .0040235  | .000553   | 7.28  | 0.000 | .0029253 - .0051217  |
| type_prof | 45.01902  | 4.29074   | 10.49 | 0.000 | 36.49724 - 53.5408   |
| type_wc   | 18.98074  | 5.342102  | 3.55  | 0.001 | 8.370863 - 29.59061  |
| incxprof  | -.0031783 | .0006047  | -5.26 | 0.000 | -.0043792 - .0019774 |
| incxwc    | -.0021712 | .00097    | -2.24 | 0.028 | -.0040976 - .002448  |
| _cons     | 13.90452  | 3.167179  | 4.39  | 0.000 | 7.614227 - 20.19481  |