Announcements

- The midterm is on Thursday
- It will be in class and similar in format to the old exams on Smartsite
- Bring a non-graphing calculator, something to write with and a scantron sheet (UCD 2000)
- There will be a formula sheet (it's posted on Smartsite so you can see what is on it)
- It will cover everything up to and including univariate data transformation (Chapters 1 through 4)
- I have office hours this afternoon from 2pm to 5pm (no office hours on Thursday after the exam)
- Problem Set 3 is posted and will graded.

Notation for Bivariate Data

- The choice of which variable is our independent variable and which variable is our dependent variable depends on what kind of causality we have in mind
- Causality is assumed to run from X to Y
- The direction of causality is typically clear from our economic theory but often can't be tested
- Our methods/statistics typically capture *associations*, not causal relationships
- Need some sort of *experiment* to determine causation (change X holding other things constant)

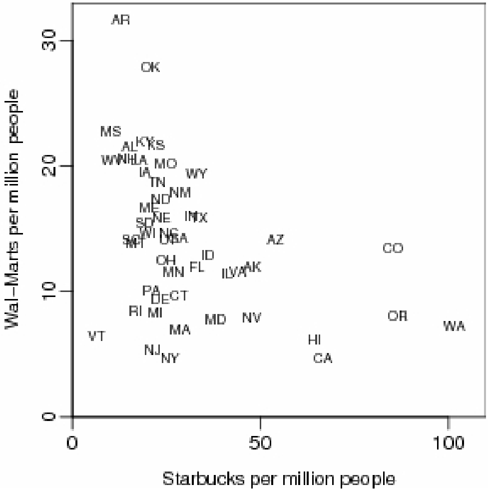
Visual Representations of Bivariate Data

- The most common way to depict bivariate data is with a scatter plot
- Each observation is single point on the graph
- x values are given by the horizontal axis, y values are given by the vertical axis
- In Excel, select the columns containing your x and y values and choose 'Scatter' from the Insert menu
- A trend line can be added by right clicking on a data point on the graph and selecting 'Add trendline...'
- We'll go through an example using data on life expectancy and GNP (gnp-life-expectancy.csv). To Excel...

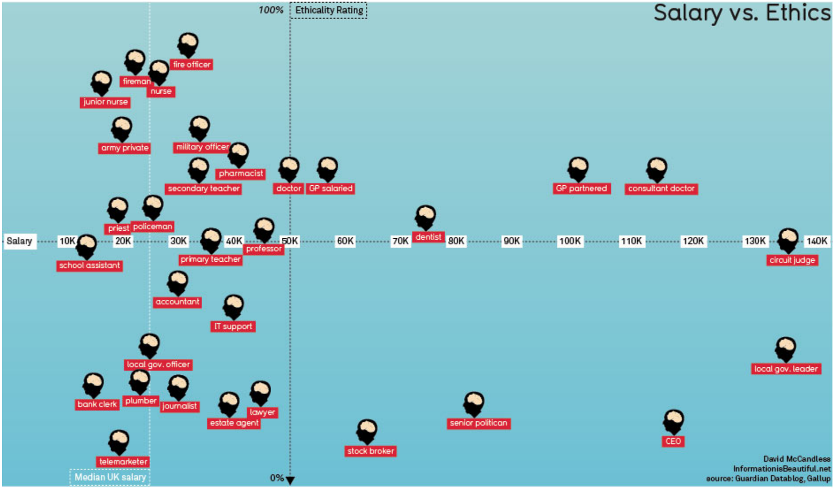
Interpreting Scatter Plots

- The most basic thing we can see on a scatter plot is whether there is a positive or negative relationship between the two variables (or no relationship)
- We can also see how strong the relationship is by how closely the datapoints follow a line
- Including the trendline can help pick out the sign of a very weak relationship
- Sometimes the relationship between two variables is much easier to see on the graph if you transform one or both of the variables ($\ln(x)$, $\sqrt{y}$, etc.) and by adjusting the scales
- Take note of any obvious extreme outlier points, often times these can be a result of incorrectly coded data or unobserved values being coded as 99 or something similar

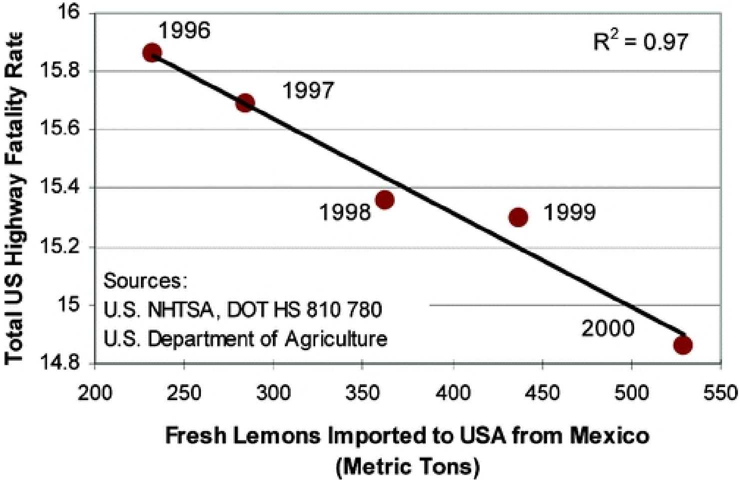
Scatter Plot Examples



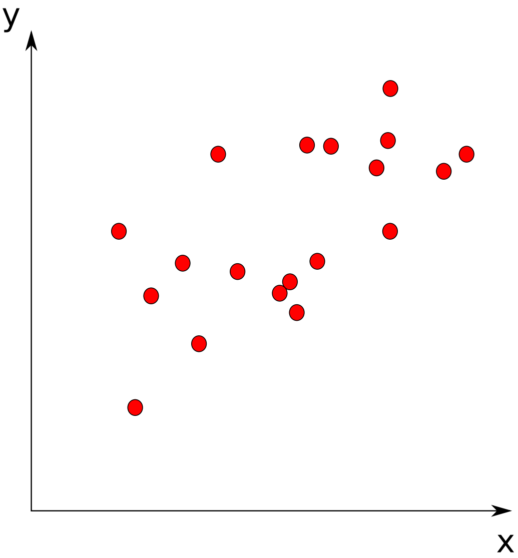
Scatter Plot Examples



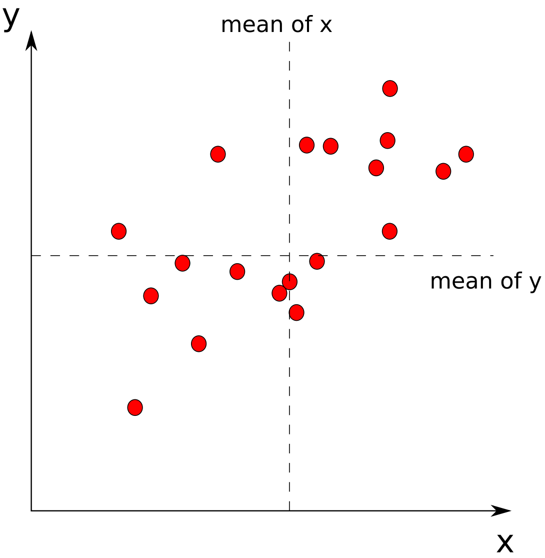
Scatter Plot Examples



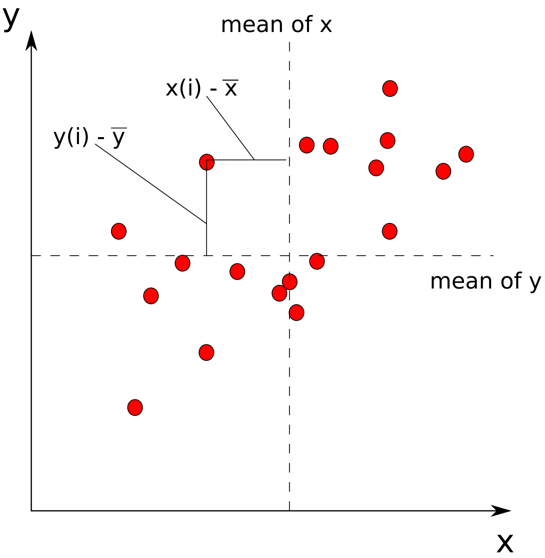
From a Scatter Plot to Descriptive Statistics



From a Scatter Plot to Descriptive Statistics



From a Scatter Plot to Descriptive Statistics



- We want a statistic that captures whether the data points lie along a positive or negative line and how close they are to that line
- One possibility: the **sample covariance**

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

- Any point that lies in the upper-right or lower-left quadrants will be a positive term in the sum
- Any point that lies in the lower-right or upper-left quadrants will be a negative term in the sum
- The sign of the covariance tells us the sign of the relationship between the variables

Covariance and Correlation

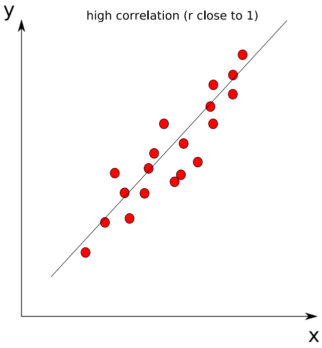
- A problem with the covariance is its magnitude
- The covariance could be large just because x and y tend to be large numbers
- We want a statistic that can tell us not only the sign of a relationship but also the strength of the relationship
- The **sample correlation** provides a standardized version of variance:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y}$$

Interpreting Correlation

- The advantage of correlation is that it is bounded between -1 and 1
- Two variables are *perfectly correlated* if r_{xy} equals -1 or 1
- Two variable are *positively correlated* if $r_{xy} > 0$ and *negatively correlated* if $r_{xy} < 0$
- The larger the magnitude of the correlation, the stronger the relationship between x and y

Interpreting Correlation



Calculating Covariance and Correlation

- You could do it yourself in Excel by calculating all of the relevant sums
- Easier approach is to let Data Analysis do it for you
- To get the sample covariance, start by using the 'Covariance' option under 'Data Analysis'
- This will produce a table of variances and covariances but they will be slightly off
- Excel's covariance function divides by n, not by $n-1$
- You need to multiply result by $\frac{n}{n-1}$
- To get the sample correlation, use the 'Correlation' option under 'Data Analysis'
- To Excel for some examples using data on health and days of missed work (health-habits.csv)...

The Regression Line

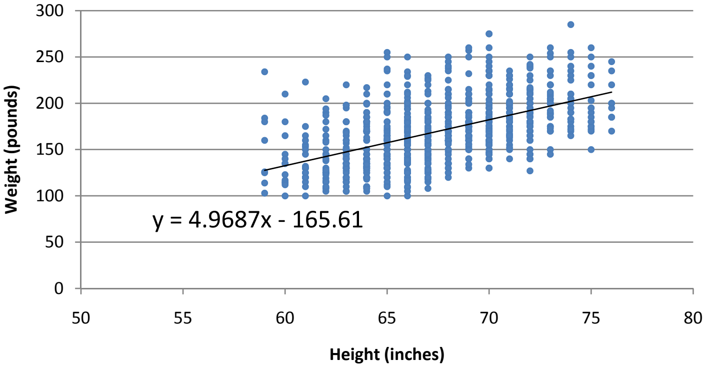
- Correlation is an improvement over covariance, but it still doesn't tell us everything
- In particular, it doesn't tell us how how large the change in y associated with a change in x is
- We would like to know:

$$\frac{\Delta y}{\Delta x}$$

- This is what a *regression line* gives us
- The regression line:

$$\hat{y}_i = b_1 + b_2 x_i$$

The Regression Line

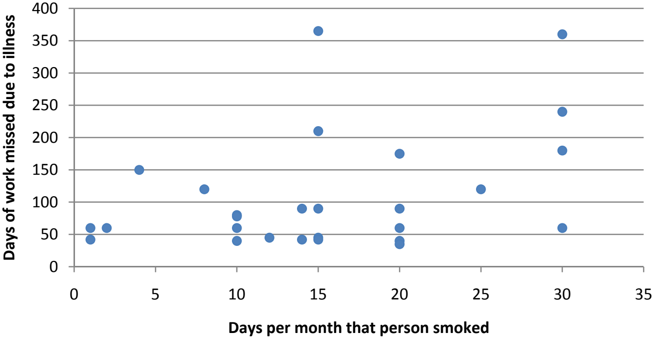


Interpreting the Regression Line

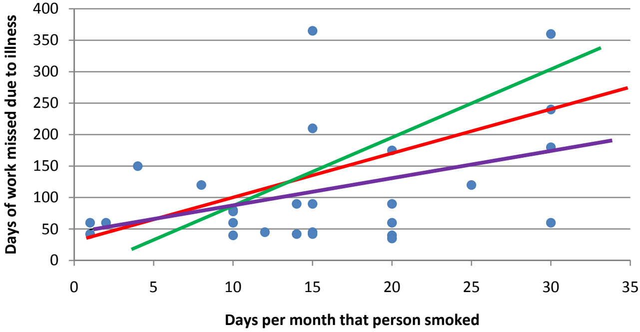
$$\hat{y}_i = b_1 + b_2 x_i$$

- $\hat{y}_i$: predicted value for Y for individual i
- x_i: observed value of X for individual i
- b_1: intercept (predicted value of Y when X equals 0)
- b_2: slope (predicted ΔY for a one unit increase in X)

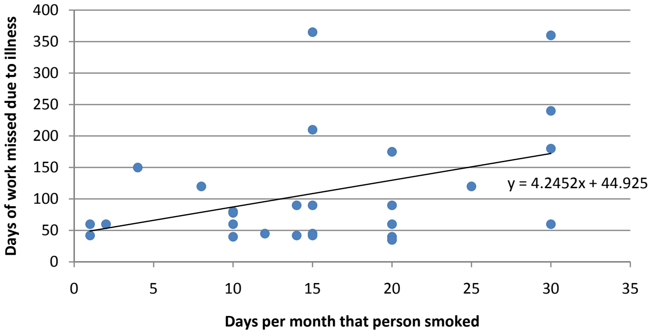
Which Regression Line?



Which Regression Line?



Which Regression Line?



Which Regression Line?

- There are many plausible lines that can be drawn through the data points
- Each different line would give us a different result for the relationship between X and Y
- We should choose the line that gives us the 'best fit'
- We'll define 'best fit' as minimizing the average distance of all of the data points from the regression line

A More Formal Approach to the 'Best Fit'

- Remember that the regression line gives us a predicted value $\hat{y}_i$ based on the observed value of x_i:

$$\hat{y}_i = b_1 + b_2 x_i$$

- The actual value of y_i will rarely be exactly equal to $\hat{y}_i$
- We'll call the difference between the true and predicted value of y_i the residual, ε_i:

$$\varepsilon_i = y_i - \hat{y}_i$$

- We want to choose the regression line such that the residuals are as small as possible

A More Formal Approach to the 'Best Fit'

- How about minimizing the sum of the residuals (or the average of the residuals)?
- No good, if we have big positive residuals and big negative residuals, we may have a bad fit even though the sum (or average) of the residuals could be zero
- We care about the magnitude of the residuals
- What can we do to focus on magnitudes? Square the residuals:

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

A More Formal Approach to the 'Best Fit'

- Now we have a way to define our best fit
- We want to choose b_1 and b_2 to minimize the average of the squared residuals:

$$\min_{b_1, b_2} \sum (y_i - \hat{y}_i)^2$$

- Replacing $\hat{y}$ with the equation for the regression line makes this:

$$\min_{b_1, b_2} \sum (y_i - b_1 - b_2 x_i)^2$$

- This is just a calculus problem that we could solve by taking derivatives with respect to b_1 and b_2 and setting them equal to zero

A More Formal Approach to the 'Best Fit'

- If you work through the math, you come up with the following two equations giving b_1 and b_2:

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_1 = \bar{y} - b_2\bar{x}$$

- Notice that the first equation looks very similar to our variance and covariance formulas, we can rewrite b_2 as:

$$b_2 = \frac{s_{xy}}{s_{xx}} = r_{xy}\sqrt{\frac{s_{yy}}{s_{xx}}}$$

Calculating the Regression Line

- To calculate b_2 and b_1 yourself:
  1. Calculate the covariance of X and Y using the covariance function in Excel
  2. Calculate the variance of X using the variance function in Excel
  3. Calculate b_2 by dividing the covariance of X and Y by the variance of X
  4. Calculate b_1 by subtracting $\bar{x}$ times b_2 you just found from $\bar{y}$ ($\bar{x}$ and $\bar{y}$ can be calculated with the average function in Excel)
- To have Excel calculate b_2 and b_1, use 'Regression' from the 'Data Analysis' choices
- Back to Excel and the health and missed work data to try regressing weight on height...

Assessing How Good the Fit Is

- We found the best fit for the regression line (according to our definition)
- This doesn't mean that we have a perfect fit; many data points will not be on the line
- We would like to know just how good the fit is, how well does the line fit the data?
- To answer this, we can use either the **standard error of the regression** or the **R-squared**

The Standard Error of the Regression

- Think back to the residuals: $y_i - \hat{y}_i$
- One way to check how good the fit is is to see how big the residuals are on average
- This is what the standard error of the regression does:

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

- The smaller the standard error of the regression is, the closer the fitted values are to the actual data for y

The R-Squared

- The standard error of the regression depends on the units that Y is measured in
- The R^2 provides a standardized measure of how good the fit is
- The idea behind the R^2 is to determine how much of the observed variation in y can be explained by the regression on x
- To do this, we need to measure the total variation in y and the amount of the variation that isn't explained by the regression
- These two measures are the **total sum of squares** and the **error (or residual) sum of squares**, respectively

The R-Squared

- The total sum of squares:

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

- The error sum of squares:

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

- The R-squared:

$$R^2 = 1 - \frac{ESS}{TSS}$$

The R-Squared

- The R^2 will always be between 0 and 1
- An R^2 of 1 means a perfect fit, x perfectly predicts y
- An R^2 of 0 means no fit, variation in x can't explain any of the variation in y
- One interpretation of the R^2 value is that it is the percentage of the variation in y explained by variation in x
- With a little algebra, you can show that R^2 is the square of r_{xy}
- The higher the correlation of two variables, the greater the R^2 will be

Regressing Wages on Education

SUMMARY OUTPUT: Weight as dependent variable

Regression Statistics	
Multiple R	0.532681203
R Square	0.283749264
Adjusted R Square	0.282871505
Standard Error	29.49983204
Observations	818

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	281318.8979	281318.8979	323.2658446	3.84342E-61
Residual	816	710115.9139	870.2400905		
Total	817	991434.8117			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-165.605738	18.65570156	-8.87695044	4.30095E-18	-202.224555	-128.986921
height	4.968722683	0.276353423	17.97959523	3.84342E-61	4.426275353	5.511170013

Assessing the R-squared

- In general, we'd like R^2 to be large but a low R^2 doesn't necessarily mean we have nothing of interest
- R^2 will tend to be high when:
  - Looking at certain time series data in economics
  - Looking at data from controlled experiments (especially in the physical sciences)
  - When the outcome is only dependent on a handful of observable variables
- R^2 will tend to be low when:
  - Looking at certain cross-sectional data in economics (especially wages, employment outcomes, productivity, etc.)
  - Looking at data where there are important but unobservable variables
  - Looking at poorly measured data