

## Ch. 3: Correlation & Regression

- Exploring relationships between 2 variables
- Scatterplots
- Linear Regression
- Exercise 03
- Correlation
- Other Correlation Coefficients

226

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Bivariate relationships

- “is factor A related to factor B”?
- Methods of analysis:
  - Anecdotal / Clinical -- often forms the basis for further systematic research & data collection
  - Numerically -- check values & % at extremes
  - Visually -- scatterplots
    - easy to see relationships and problems w/ data
    - hard to prove / test
  - Statistically -- correlation & regression
    - hard to detect problems w/ data
    - easy to test hypothesis

227

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Anecdotal / Clinical

- Many interesting findings in psychology first originate from non-scientific approaches
- “Intuition” that something is related through experiencing multiple situations
- Pattern recognition
- Human brains are both excellent and terrible pattern recognizers
- Problems -- faulty memory, confirmation biases, prejudice, etc...
- First step after a “gut” feeling is to begin collecting data.

228

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Simple numerical analysis

- Simplify the situation by using Categorical variables (or reducing Continuous variables to Categorical variables)
- Use extreme cases to maximize effect
- Compute percentages in a 2x2 matrix
- Do the results suggest an effect?
- Compute Chi-square statistic to judge significance

229

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Example

- “I think there is brain dysfunction in HIV disease” as measured by neuropsychological testing
- Medical status: control vs. HIV+ symptomatic
- NP test results: normal vs. impaired

		Medical Status	
		Control	HIV+
NP Status	Normal	85%	52%
	Impaired	15%	48%

230

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Issues

- Pro: Results are very easy to understand from a human or clinical point of view.
- Con: dividing continuous variables into two values throws away a lot of data and statistical power
- Graphical and Statistical methods should be used as well.

231

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Scatterplots

- Graph two variables in relation to each other on two-dimensional X, Y axis
- Easy to see relations between data
- Easy to see problems with data
- Hard to prove or determine if an apparent relationship is “significant”
- Hard to interpret data clinically or in “common sense” terms

232

Psychology 402 - Fall 2014 - Dr Michael Deiter

## Scatterplots

x	y
10	8.04
8	7.58
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68



233

Psychology 402 - Fall 2014 - Dr Michael Deiter

## Linear Regression

- Assume that two variables are related, and that this relationship is linear -- model the data by a simple straight line for the data.
- For any given data set, we pick the line that best “fits” our data
- Similar terms: linear regression, fitting a line, finding the trend, creating a trendline, best fit line, etc.
- Residuals = difference between prediction and actual value
- Linear Regression minimizes the square of the residuals, often called “Ordinary Least Squares”

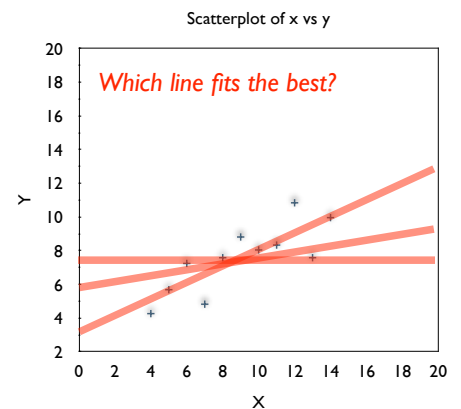
234

Psychology 402 - Fall 2014 - Dr Michael Deiter

## Linear Regression

Equation:  
 $y = 3.0 + 0.5x$

Correlation  
 $r_{x,y} = 0.816$

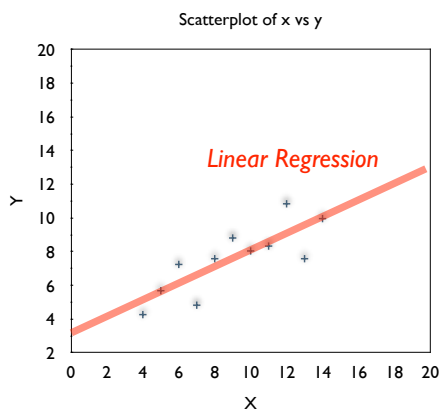


235

Psychology 402 - Fall 2014 - Dr Michael Deiter

## Anscombe's Quartet I

x	y
10	8.04
8	7.58
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

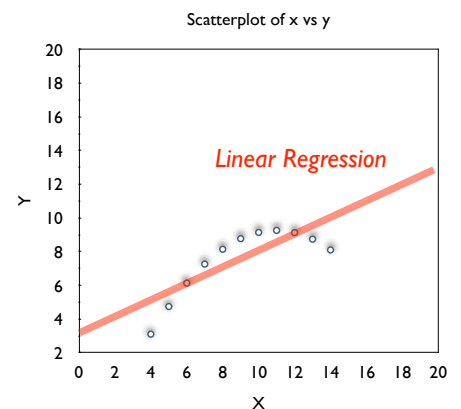


236

Psychology 402 - Fall 2014 - Dr Michael Deiter

## Anscombe's Quartet II

x	y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.13
7	7.26
5	4.74

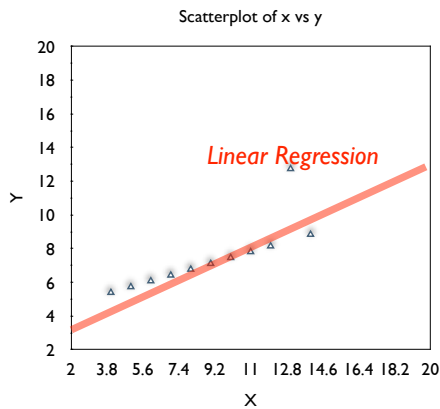


237

Psychology 402 - Fall 2014 - Dr Michael Deiter

## Anscombe's Quartet III

x	y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

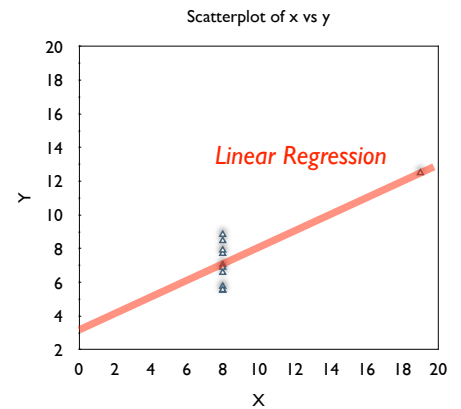


238

Psychology 402 - Fall 2014 - Dr Michael Daler

## Anscombe's Quartet IV

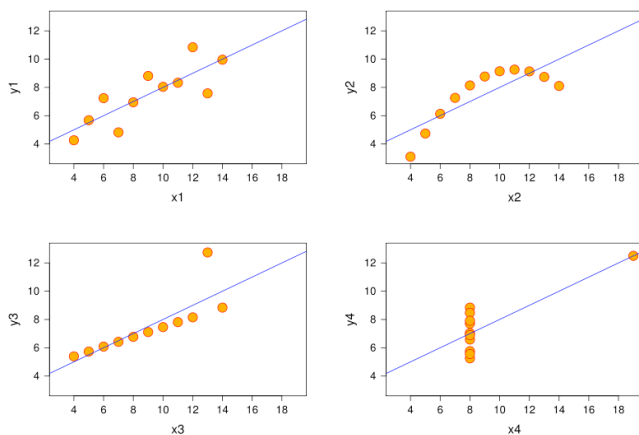
x	y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.52
19	12.5
8	5.56
8	7.91
8	6.89



239

Psychology 402 - Fall 2014 - Dr Michael Daler

## Anscombe's Series 1-4



240

Psychology 402 - Fall 2014 - Dr Michael Daler

## Anscombe's Quartet Summary

- The 4 series of data, though very different, have identical linear regression equations and identical correlations
- Each series has a Quantitative correlation, but it's clear (visually) that the relationships are Qualitatively different
- Each series should probably be handled differently, through techniques such as:
  - Trimmed Least Squares
  - Robust regression
  - Graph Your Data!

241

Psychology 402 - Fall 2014 - Dr Michael Daler

## Linear Regression Equation

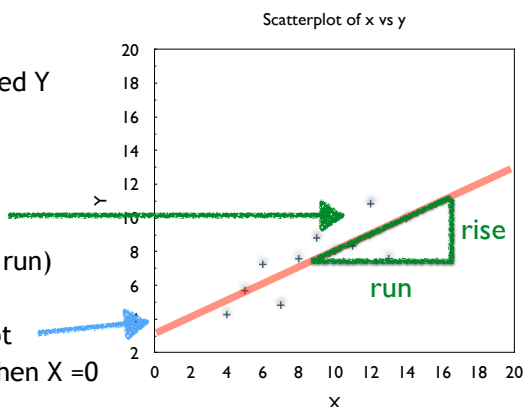
$$Y' = a + bX$$

$Y'$  = predicted Y

X = actual X

b = slope  
DY/DX  
(rise over run)

a = intercept  
Y value when X = 0



242

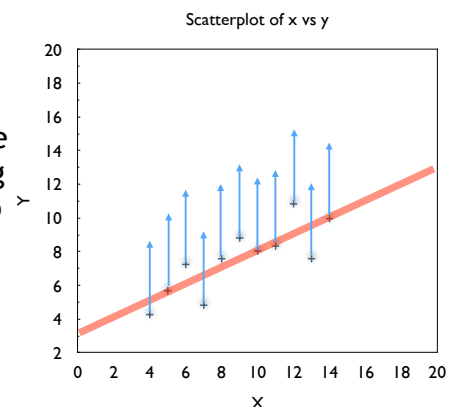
Psychology 402 - Fall 2014 - Dr Michael Daler

## Adjusting Intercept

$$Y' = a + bX$$

a = intercept

Intercept can be raised by adding same amount to all Ys



243

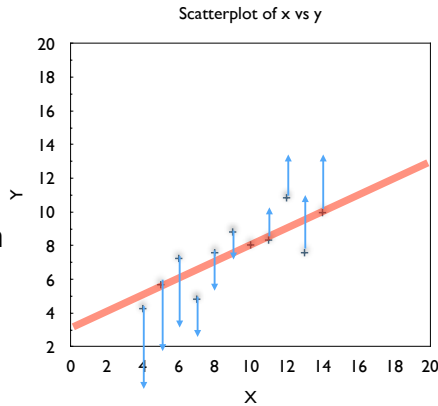
Psychology 402 - Fall 2014 - Dr Michael Daler

## Adjusting Slope

$$Y' = a + bX$$

$b$  = slope

Slope change by adding more to right side, subtracting from left side



244

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Residuals in Linear Regression

- $X$  : independent variable
- $Y$  : dependent variable
- Model: predict  $Y$  from  $X$
- $Y'$  : ( $Y$  prime) = predicted  $Y$
- $Y' = a + bX$
- Prediction is (usually) incorrect. Difference between predicted ( $Y'$ ) and actual ( $Y$ ) is called a "Residual" =  $(Y - Y')$
- Calculation of best fit line minimizes the sum of the squared residuals  $\sum(Y - Y')^2$

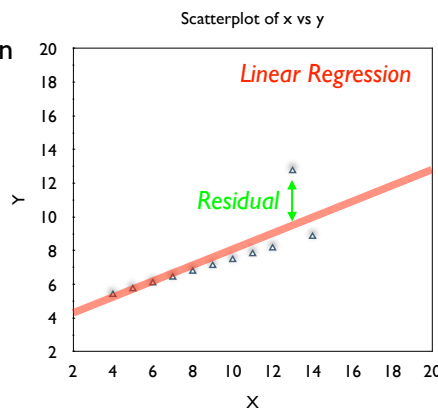
245

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Residuals in Linear Regression

Residual is difference between actual  $Y$  and predicted  $Y'$  ( $Y - Y'$ )

Graphically it is equal to how far away (vertically) a point is from the linear regression line



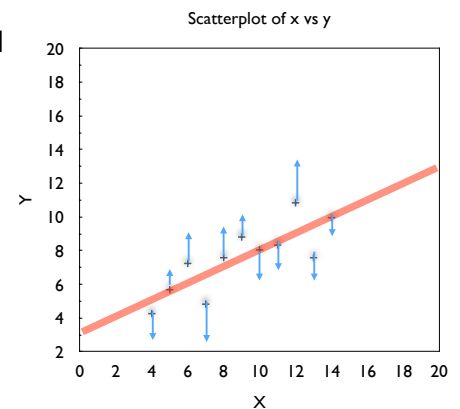
246

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Adjusting Residuals

Residuals (error) can be increased by moving  $Y$  values further away from the regression line.

If above line, move higher.  
If below line, move lower.



247

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Residuals, Variance, $R^2$

- Residual =  $(Y - Y')$
- Squared residual =  $(Y - Y')^2$
- Sum of squared residuals =  $\sum(Y - Y')^2$ 
  - Linear regression minimizes this value
- SSR is hard to interpret
- $R^2$ 
  - $R^2 = 1 - (SSR/SST)$
  - Coefficient of Determination
  - Unexplained Variance
  - Ranges from 0 to 1 (often 0% to 100%)

248

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Standard Error of Estimate

- Residual =  $(Y - Y')$
- Standard Deviation of residuals
  - measure of "average" error
  - aka "Standard Error of Estimate"

254

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Correlation (r) Pearson's r

- Pearson's Product-Moment Correlation
- Measures the strength of the linear relationship between two variables
- Ranges between -1.0 and +1.0
- Is a special case of linear regression, when both X and Y have been turned into Z scores.
- r is transitive (correlation between X and Y is same as correlation between Y and X)
- $R^2$  = "explained variance" is the proportion of variation in the data explained by the model.
- $R^2$  ranges from 0 to 1.0 (0% to 100%)

256

Psychology 402 - Fall 2014 - Dr Michael Doherty

## Regression vs. Correlation

	Linear Regression	Correlation
Scores	Raw	Z
Mean, Std Dev	sample means sample Std Dev	0 1
Equation	$Y' = a + bX$	$Y' = r X$
Slope	b = change in Y per change in X	r = correlation coefficient
Slope	meaningless	R
Transitive?	no	yes, R

257

Psychology 402 - Fall 2014 - Dr Michael Doherty

## Other Correlation Coefficients

- Continuous (interval & ratio): Pearson's r
- Ordinal (Ranked): A B C D... 1st, 2nd, 3rd...
  - Spearman's Rho: correlation between two ordinal / ranked variables.
- Dichotomous (yes/no, one/zero, T/F, Male/Female, Pass/Fail...)
  - True vs. Artificial?

258

Psychology 402 - Fall 2014 - Dr Michael Doherty

## Continuous vs. Dichotomous

Type of X / Type of Y	Continuous	Artificial Dichotomous	True Dichotomous
Continuous	Pearson r	Biserial r	Point biserial r
Artificial Dichotomous	Biserial r	Tetrachoric r	Phi
True Dichotomous	Point biserial r	Phi	Phi

259

Psychology 402 - Fall 2014 - Dr Michael Doherty

## Correlation : Issues

- Technical / Calculation :
  - Non-normal distribution
  - Non-linear data and relationships
  - Outliers, data errors
  - Restricted Range
  - Shrinkage
- Interpretation:
  - Correlation  $\neq$  Causation
  - Third variable explanations

260

Psychology 402 - Fall 2014 - Dr Michael Doherty

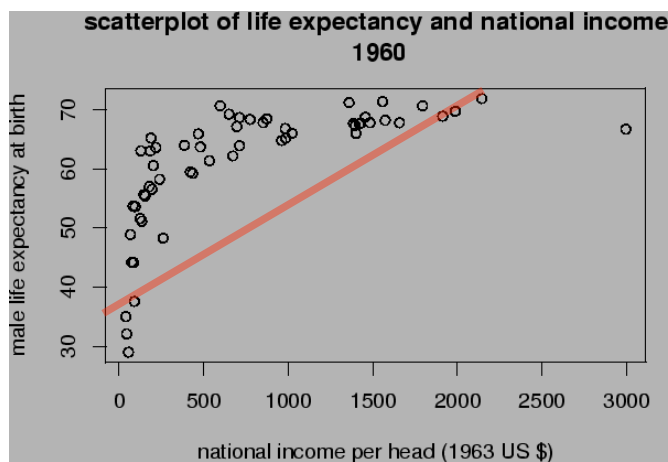
## Non-linearity

- Linear regression & Correlation assume a linear relationship between X and Y
- Are most real-world relationships linear?
- Examples of non-linearity
- Solutions:
  - Intentionally restrict range of X
  - Rank variables then use Spearman's Rho
  - Transform variables (log, root, square, cube, etc.)
  - Use higher-order (polynomial) curve fitting, such as  $Y = a + bX + cX^2 + dX^3 \dots$

261

Psychology 402 - Fall 2014 - Dr Michael Doherty

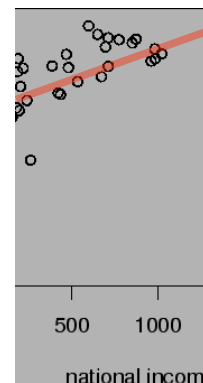
## Life expectancy / national income



262

Psychology 402 - Fall 2014 - Dr Michael Dwyer

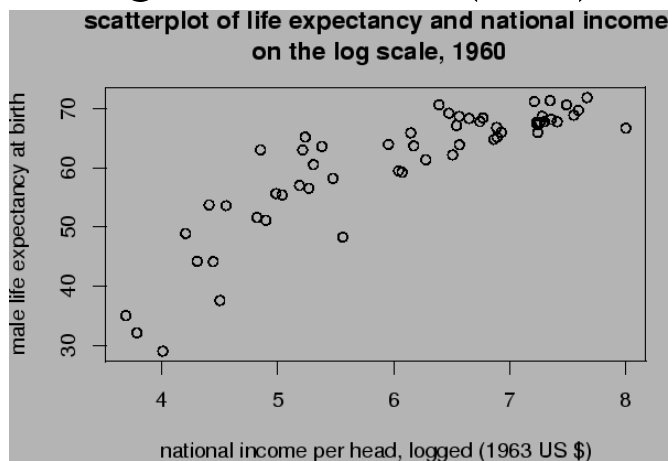
## restrict range of X



263

Psychology 402 - Fall 2014 - Dr Michael Dwyer

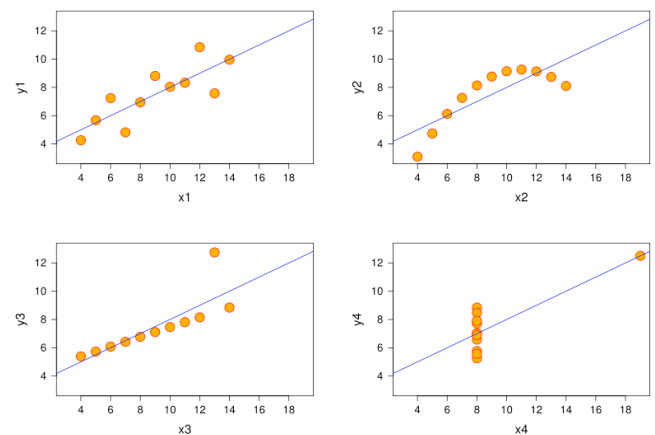
## log transform X (or Y)



264

Psychology 402 - Fall 2014 - Dr Michael Dwyer

## Outliers & Data Errors?



265

Psychology 402 - Fall 2014 - Dr Michael Dwyer

## Correlation = Causation?

- A relationship (linear or otherwise) between X and Y tells us nothing about whether X causes Y
- Lack of correlation between X and Y does not mean that X doesn't cause Y
- Sleeping with your shoes on is correlated with waking up with a headache
- Ice cream sales are positively related to increase in drowning

266

Psychology 402 - Fall 2014 - Dr Michael Dwyer

## Shrinkage

- Least-squares regression attempts to fit the data set presented to it by reducing the observed residuals.
- This data set contains random errors.
- Thus, the parameters (equations) estimated for the linear regression line (and correlation coefficient) and residuals usually be higher than would be found in a separate data set.
- This reduction is called "Shrinkage"
- Cross-validation is best way to deal with it

267

Psychology 402 - Fall 2014 - Dr Michael Dwyer

## Cross Validation

- Step 1: With a given data set, compute the linear regression line that fits this data.
- Step 2: Apply this linear regression equation to a different data set.
- Step 3: Calculate the observed error in step 2. This is typically higher than seen in step 1, and a much better measure of fit.
- Note: sometimes you may artificially “create” two data sets by splitting a single data set in half.

268

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Hypothesis Testing

- All parameters (equations) we estimate from data have inherent error
- How do we know if a given estimate is correct?
- How big is the error likely to be (confidence intervals)?
- Inferential Statistics - covered later
  - Formulas to calculate probability, confidence intervals.
  - Higher N is better
  - “statistical significance” not the same as “clinical significance”

269

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Statistical and Clinical Significance

- These two terms are often confused and have very different meanings
- Statistical Significance: changes in DV are very unlikely to have been the result of random effects or chance. Often expressed as a P value ( $p < .01$ , or less than 1% chance to see these effects under  $H_0$ )
- Clinical Significance : changes in DV are large enough to matter; the change was not trivial. If we accept  $H_1$ , the conclusion is that  $H_1$ 's effect size is important.
  - depends on context. often evaluated in terms of cost/benefit or risk/benefit tradeoff

270

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Significance

- Example 1:
  - Two dice, Roll each once
  - Results: get a 3 and a 5
- Example B:
  - Two dice, Roll each 100 times
  - Results: Die A = 3.0, Die B = 3.10
- Example C:
  - Two dice, Roll each 100 times
  - Results: Die A = 3.0, Die B = 5.0

271

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Lies, damned lies, and statistics

- Statistical significance (P) is a function of...
  - Errors of measurement (E)
  - Effect Size (D)
  - Sample Size (N)
- $p \sim E / (D \times N)$

272

Psychology 402 - Fall 2014 - Dr. Michael Deiter

## Reporting Results

- “Men had higher IQ than women. Results were statistically significant  $p < .001$ ”
- Effect Size (D)
- P-value
- Probability of Type I error ( $\alpha$ )
- Probability of Type II error ( $\beta$ )

273

Psychology 402 - Fall 2014 - Dr. Michael Deiter