Panel 1

Last Time:

Works for nominal + ordinal variables

Chi-Square Test to check if two variables are related.

Hypothesis: Two variables are independent

Conduct Chi-Square test and compute p.

If $p < 0.05$ ⇒ reject Hypothesis and accept alternative

If $p \geq 0.05$ ⇒ do nothing

Note: all expected values are 5 or more.

Does not tell you how strong variables are related.

1

Panel 2

Correlation for Numeric Variables

HS GPA	College GPA
3.8	2.8
3.1	2.2
4.0	3.5
2.5	1.9
3.3	2.5

Q1: Are they related?

Q2: How strong is association

Q3: Make predictions

Compute Correlation Coefficient r:

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

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

2

Panel 3

x	y	x^2	y^2	xy
3.8	2.9	14.44	7.84	$3.8 \cdot 2.9 = 10.64$
3.1	2.2	9.61	4.84	6.82
4.0	3.5	16.00	12.25	14.00
2.7	1.9	6.25	3.61	4.75
3.3	2.5	10.89	6.25	8.25
16.7	12.9	57.19	34.79	44.46

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{5} = 57.19 - \frac{16.7^2}{5} = 1.412$$

$$S_{yy} = 34.79 - \frac{(12.9)^2}{5} = 1.509$$

$$S_{xy} = 44.46 - \frac{16.7 \cdot 12.9}{5} = 1.374$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{1.374}{\sqrt{1.412 \cdot 1.509}} = 0.9446$$

x is independent var (HS)

y is the dependent var (college)

3

Panel 4

The Correlation Coefficient r:

- is always between -1 and 1
- if close to 1 or -1, there is a strong relation between variables
  if close to 0, no relation
- if positive $\Rightarrow$ positive relation (if x gets bigger, y gets bigger)
  if negative $\Rightarrow$ negative relation (if x gets bigger, y gets smaller)

In our example: $r = 0.9$ $\Rightarrow$ strong positive relation between x, y,

i.e. HS GPA (x) is a good predictor of College GPA (y)

4

Panel 5

Scatter Plot

- Plot points (x,y) in a Cartesian coordinate system
- Draw a line that is equally close to all points (may or may not pass through any points) $\Rightarrow$ Least Square Regression Line
- Some high-level math shows that this line has the equation $y = mx + b$, where

$$m = \frac{S_{xy}}{S_{xx}}, \quad b = \bar{y} - m\bar{x}$$

5

Panel 6

Example

Highest year of school completed, father x	Highest year of school completed y	x^2	y^2	xy
12	12			
15	16			
5	7			
16	19			
68	74	670	810	723

$S_{xx} = 74, \quad S_{xy} = 75, \quad S_{yy} = 81 \Rightarrow r = 0.967$

$$m = \frac{S_{xy}}{S_{xx}} = \frac{75}{74} = 1.01, \quad b = \bar{y} - m\bar{x} = \frac{74}{4} - 1.01 \cdot \frac{68}{4} = 18.5 - 17.01 = 1.5$$

$y = 1.01x + 1.5$ is the equation of the least square regression line.

6

Panel 7

We can use that line to make predictions:

$y = 1.01x + 1.5$

If $x = 14 \Rightarrow y = 15.64$ Is this prediction good? Know $r = 0.986$, i.e. good prediction

regression line

Ex:

Simple linear regression results:

Dependent Variable: HIGHEST YEAR SCHOOL

Independent Variable: FATHER HIGHEST YEAR SCHOOL

HIGHEST YEAR SCHOOL = 9.805244 + 0.34786397 FATHER HIGHEST YEAR SCHO

Sample size: 1485

R (correlation coefficient) = 0.4814 some pos. relation

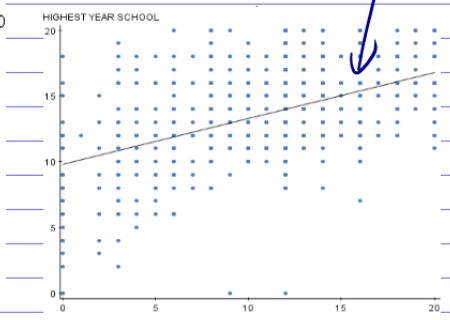
R-sq = 0.23172005

Estimate of error standard deviation: 2.661603

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	9.805244	0.20117672	≠ 0	1483	48.73946	<0.0001
Slope	0.34786397	0.01644814	≠ 0	1483	21.149136	<0.0001

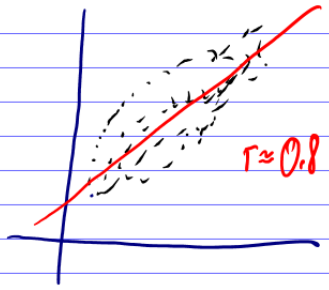
$y = 9.805 + 0.347x$



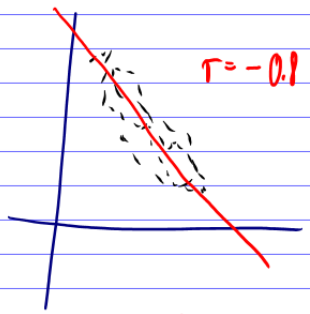
7

Panel 8

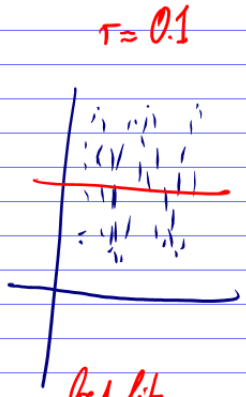
Ex: Scatter Plot



good fit



good fit



bad fit

8

Panel 9

Simple linear regression results:

Dependent Variable: Life Expectancy
Independent Variable: People who read (%)
Life Expectancy = 38.469986 + 0.37040222 People who read (%)
Sample size: 107
R (correlation coefficient) = 0.8436
R-sq = 0.7116605
Estimate of error standard deviation: 5.420809

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Valu
Intercept	38.469986	1.8770674	≠ 0	105	20.494728	<0.000
Slope	0.37040222	0.02300883	≠ 0	105	16.098263	<0.000

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	7615.2866	7615.2866	259.15408	<0.0001
Error	105	3085.4426	29.385168		
Total	106	10700.729			

9