ISYS3401
Information Technology Evaluation

Week 5 Lecture

Dr Vincent Pang

Vincent.Pang@sydney.edu.au

Agenda

- Week 7 Mid-Semester Quiz
- Research Model
- Linear Regression
- Class Activities

Week 7 Mid-Semester Quiz

Venue: Normal Monday Lecture

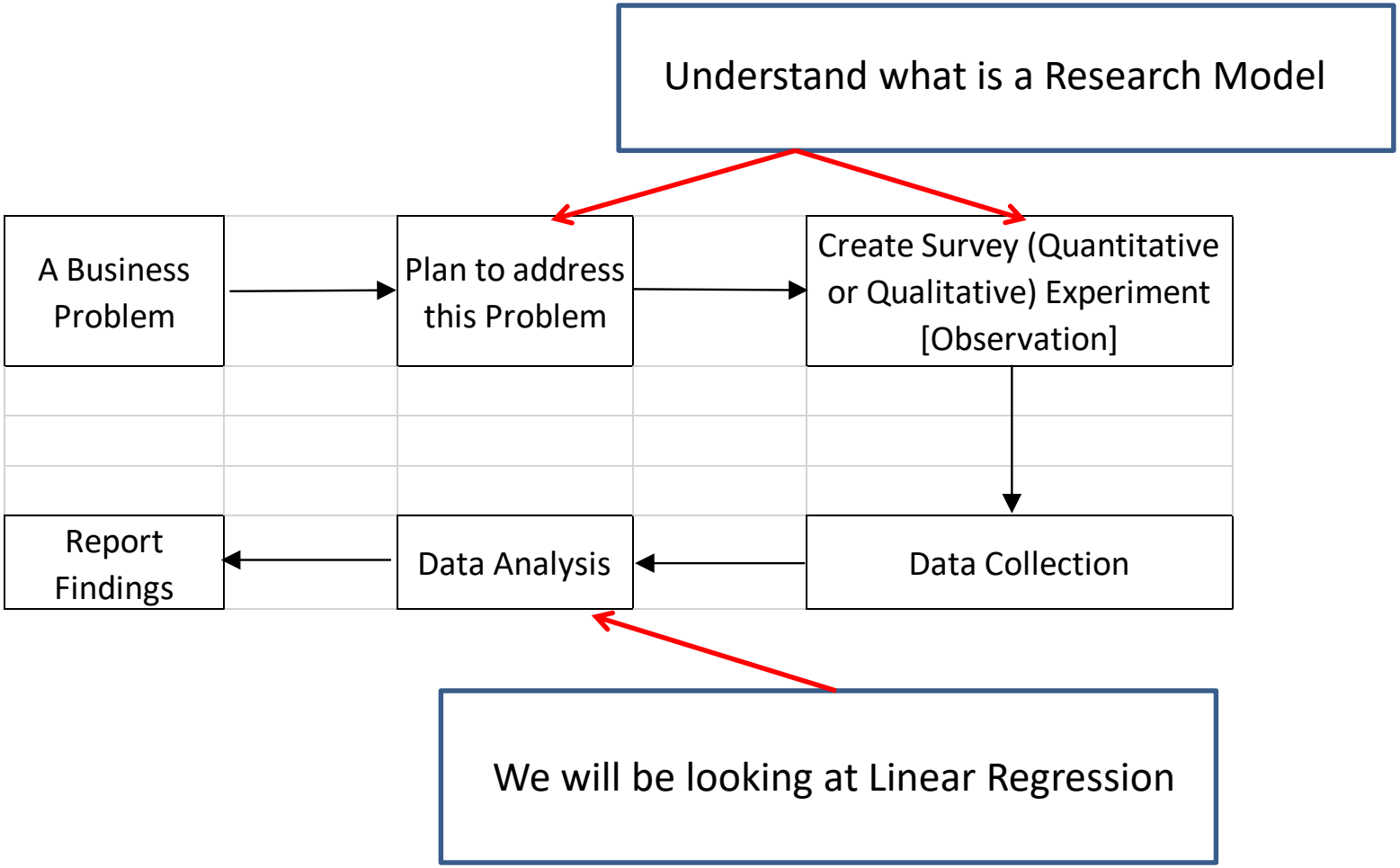
Date: Monday, 8th April, 2019

Time: 12.10pm (1hr and 10mins)

Type: Closed Book

Course Assessment: 15%

This week …


Understand what is a Research Model
A Business Problem
Plan to address this Problem
Create Survey (Quantitative or Qualitative) Experiment [Observation]
Report Findings
Data Analysis
Data Collection
We will be looking at Linear Regression

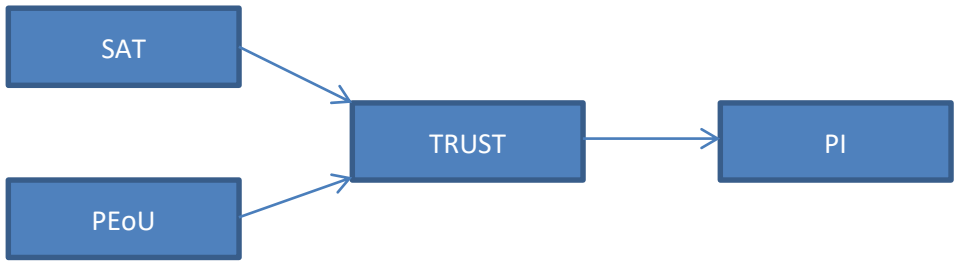
A Research Model (1)

- An academic research model is designed, supported, guided, and more likely based on a theory or theories, i.e. each construct in the model is supported with literature.
- Each construct should have indicators which are associated with survey questions.
- There are many survey models, and for this course, we will focus on the models published in the Information Systems journals..

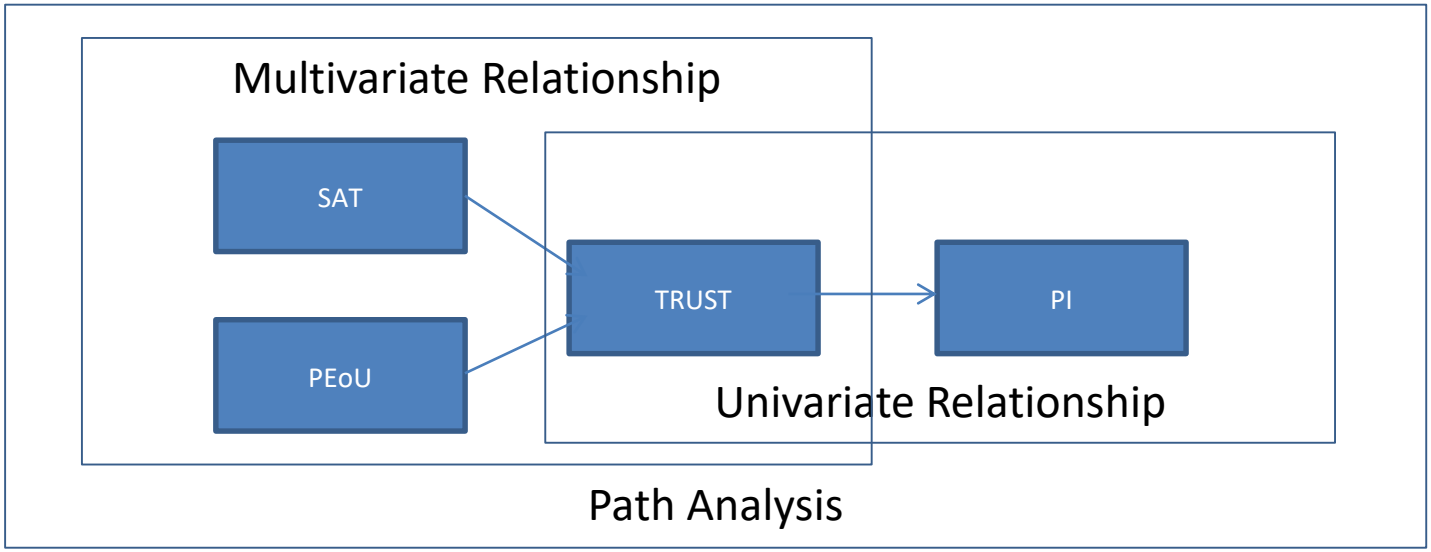
Note: this is not to say models published in other disciplines are incorrect. The expectations and assumptions are sometimes different between disciplines.

A Research Model (2)

- e-Commerce has been found to be one of the most powerful mechanism reforming the world economy in the past 2 decades.
- An important research question in the e-Commerce domain is to study **consumers' trust** in online retailers. Based on prior researches, One may make the following hypotheses for online stores:
  - The **perceived trustworthiness (TRUST)** of the website will in turn affect customers' **Purchase Intention** (**PI**).
  - The trustworthiness perception of the online operation of a company would be positively affected by **customers' SAtisfacTion (SAT)** with the offline operation of the company.
  - The trustworthiness perception of the online operation of a company would be positively affected by the **Perceived Ease of Use (PEoU)** of the website.



Relationships in Research Model

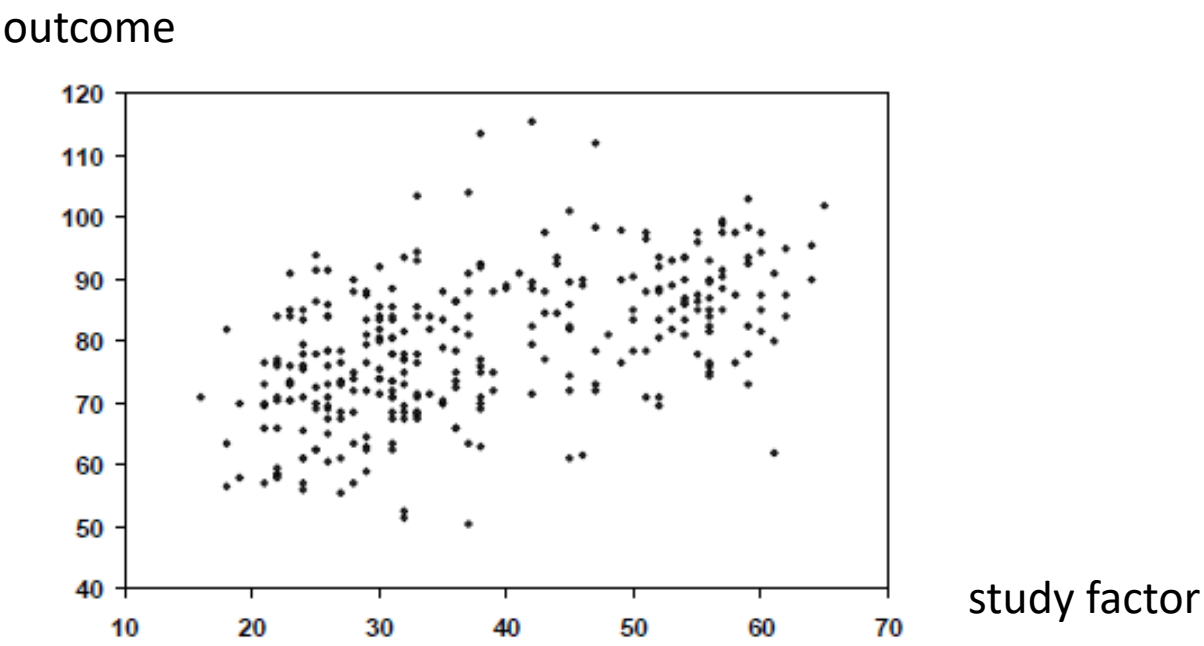


Objectives

1. Describe when to use **Linear Regression line**
2. Explain the assumptions of the linear regression model
3. Interpret an analysis of variance for linear regression
4. Use tables of the F distribution
5. Interpret a regression coefficient and test its statistical significance
6. Interpret a correlation coefficient *r and its square,* coefficient of determination R^2
7. Avoid common pitfalls of linear regression

Linear Regression

- Association – “Two [quantitative] variables are associated if the distribution of one is affected by a knowledge of the other”
- Scatter diagram:



Equation of a straight line

$$y = \alpha + \beta x$$

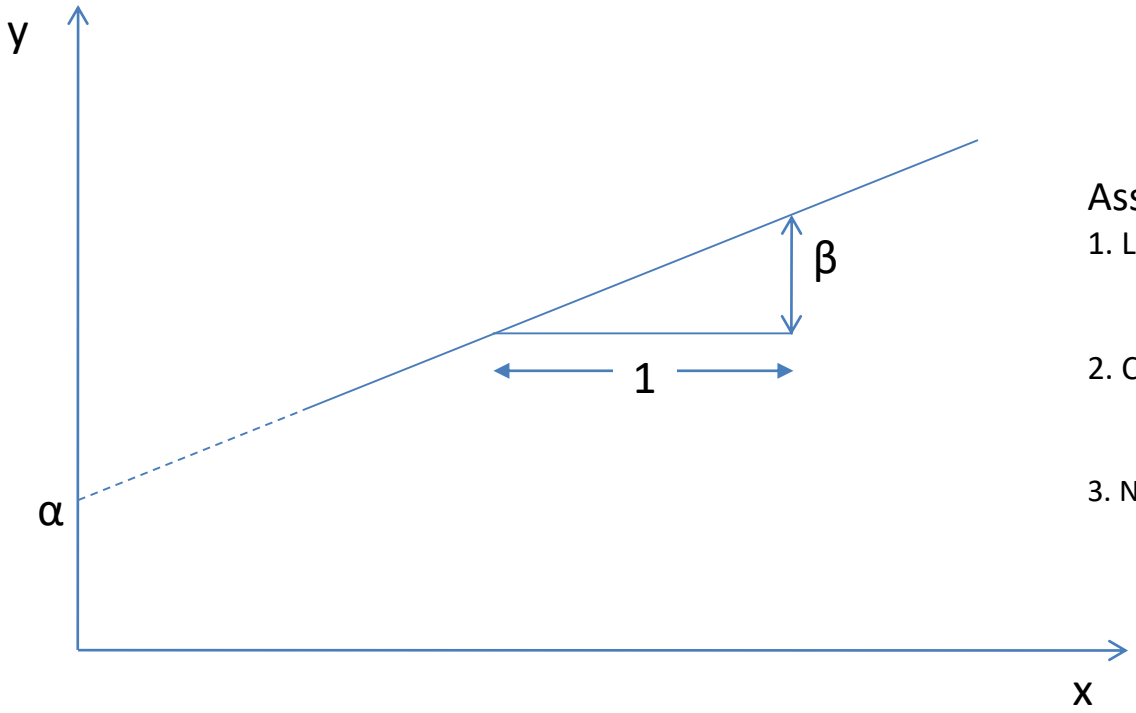
where

y is the outcome variable (dependent)

x is an explanatory or predictor variable (independent study factor)

α is the intercept: the value of y when x=0

β is the slope of the line: the amount y changes for each unit increase in x



Assumptions:

1. Linearity

   $E(y) = \alpha + \beta x$

2. Constant variance (Homoscedasticity)

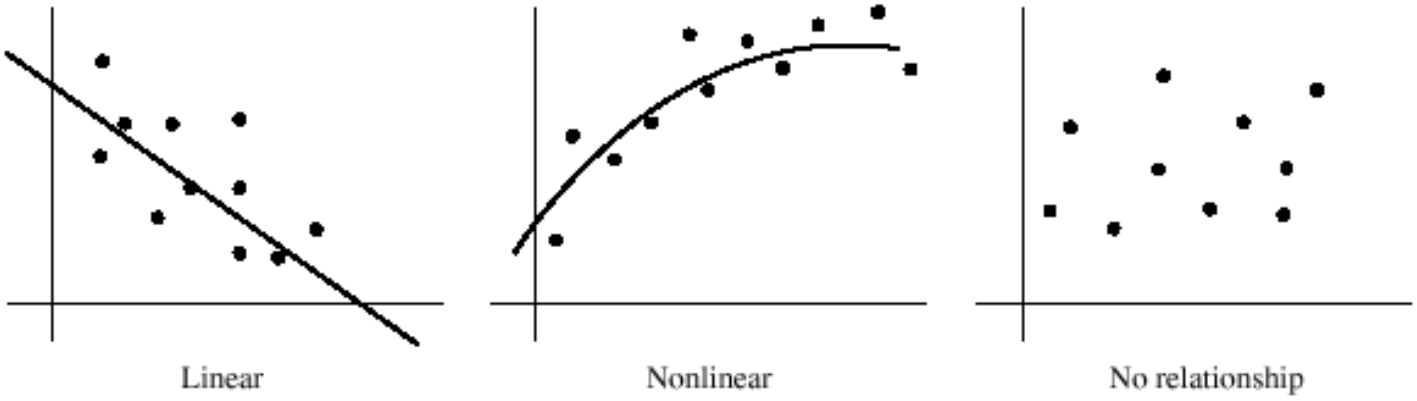
   For all values of *x, SD(y) (= σ) is the same*

3. Normality

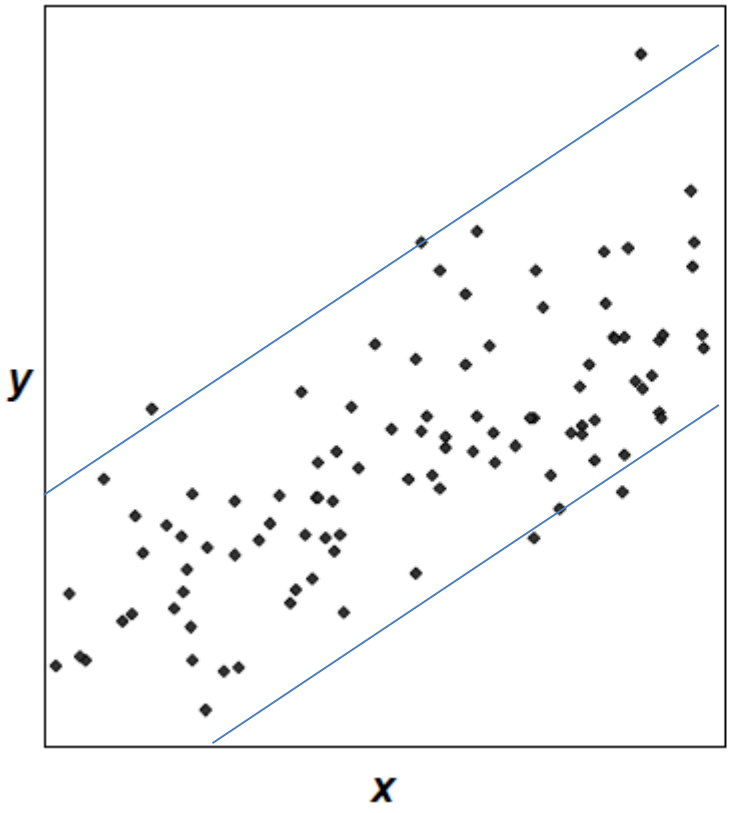
   For a given value of *x, y is Normally distributed*

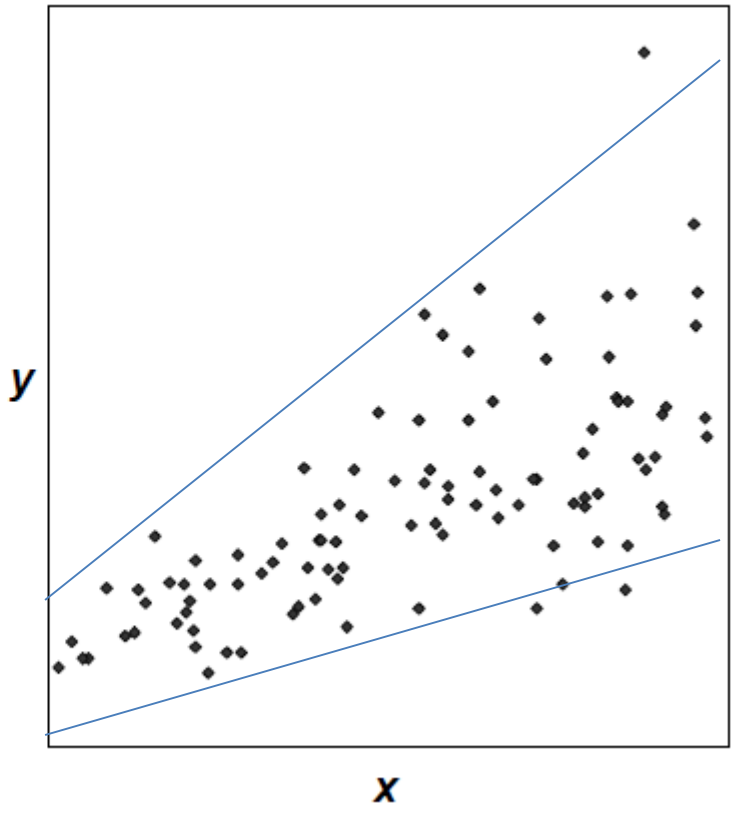
Simple Linear Regression 1: Linearity

- Single independent variable
- Linear relationship

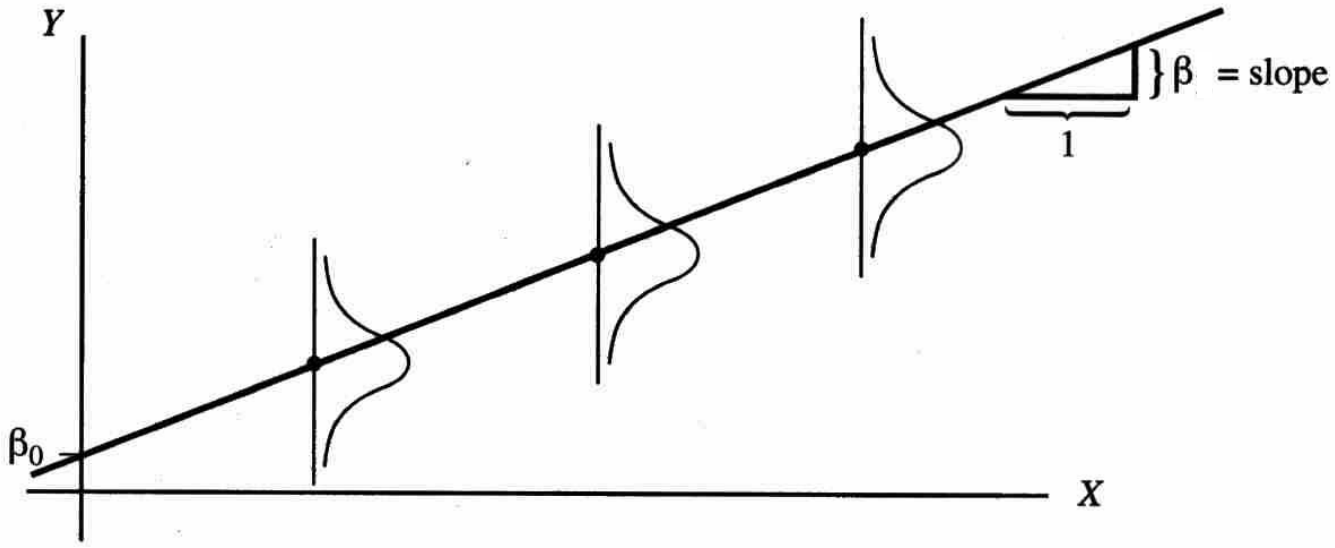


Linear Regression 2: Constant Variance



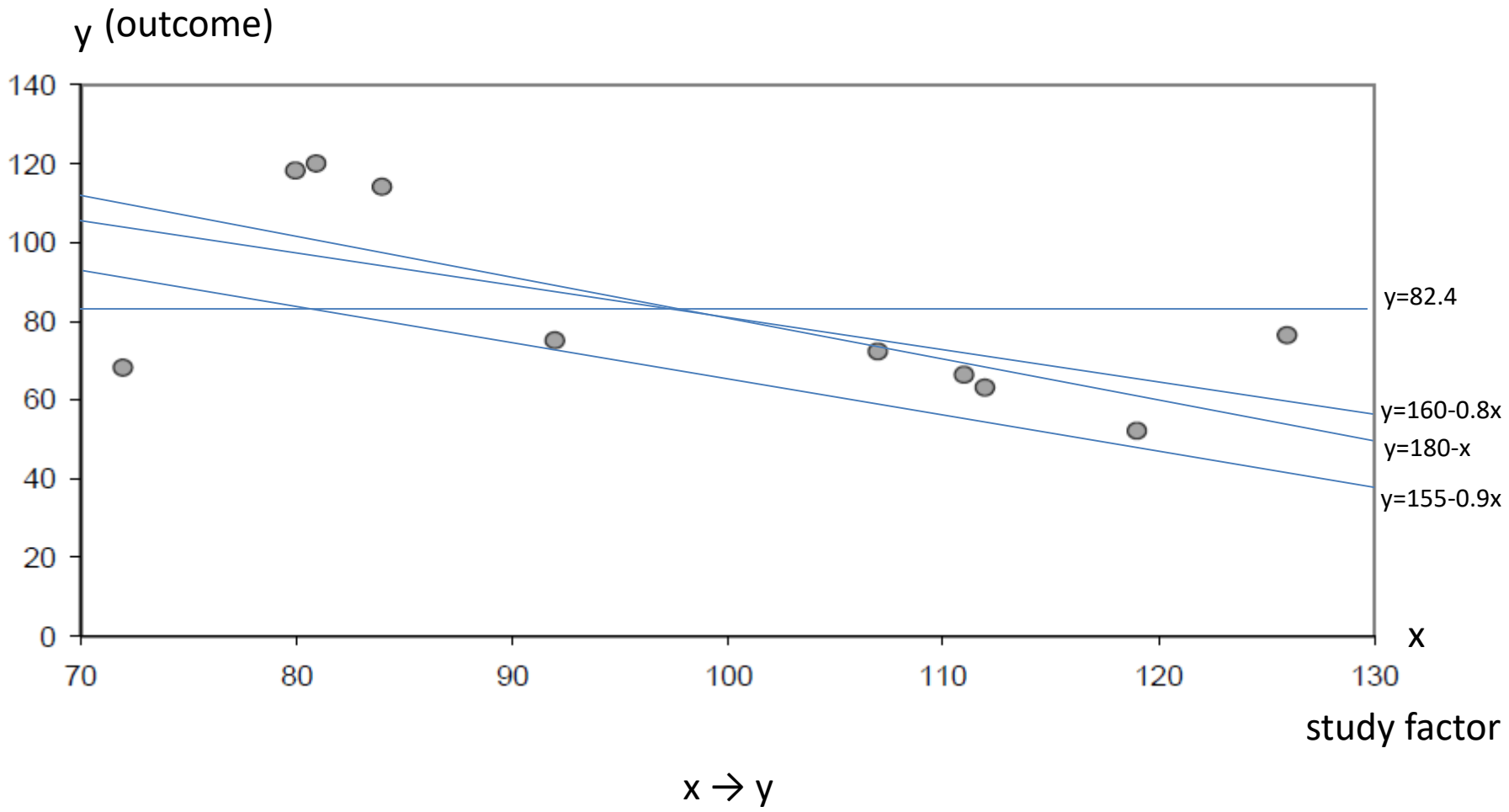


Linear Regression 3: Normality



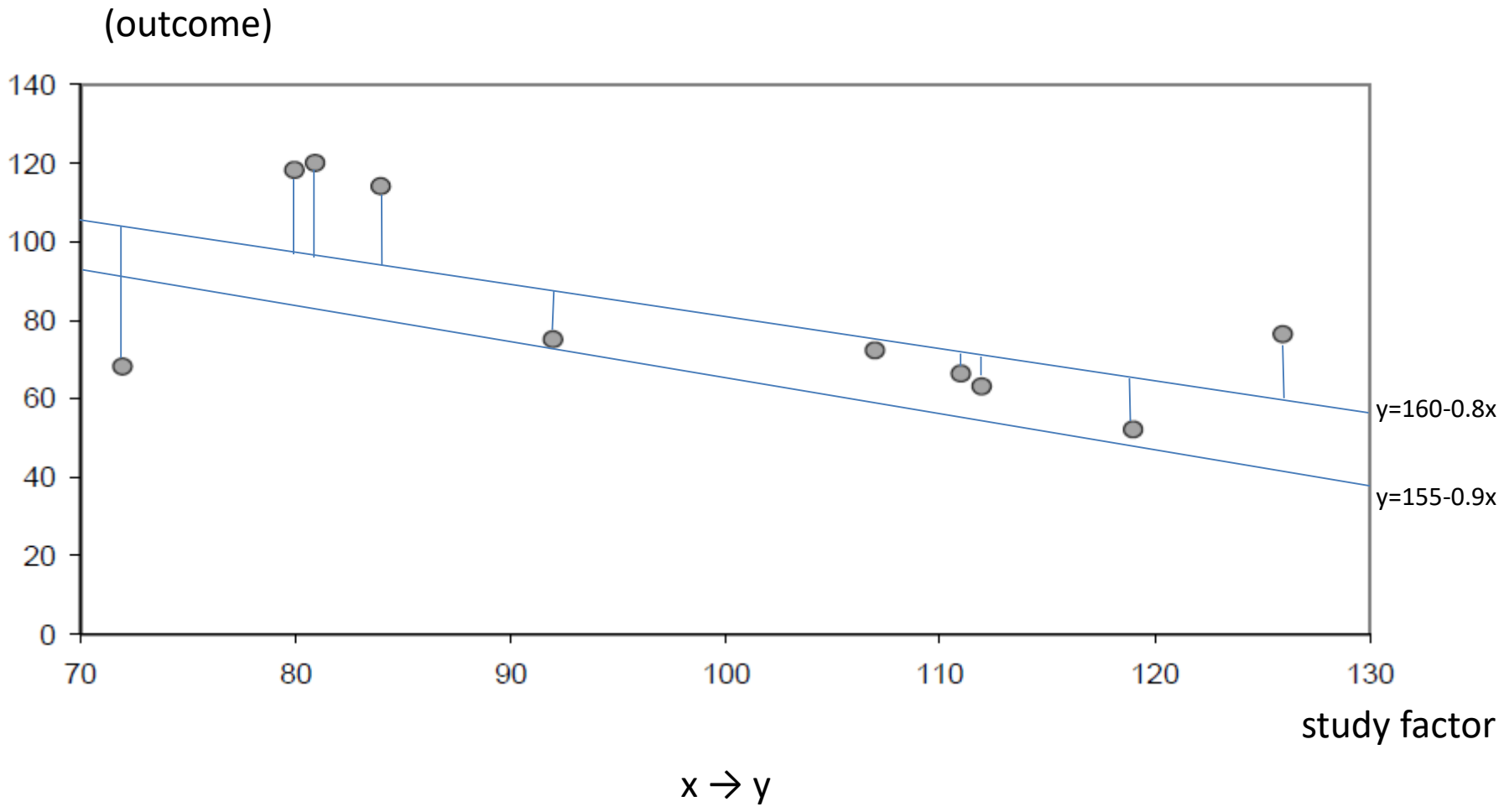
(a) Regression line through population means

How to choose the best line?



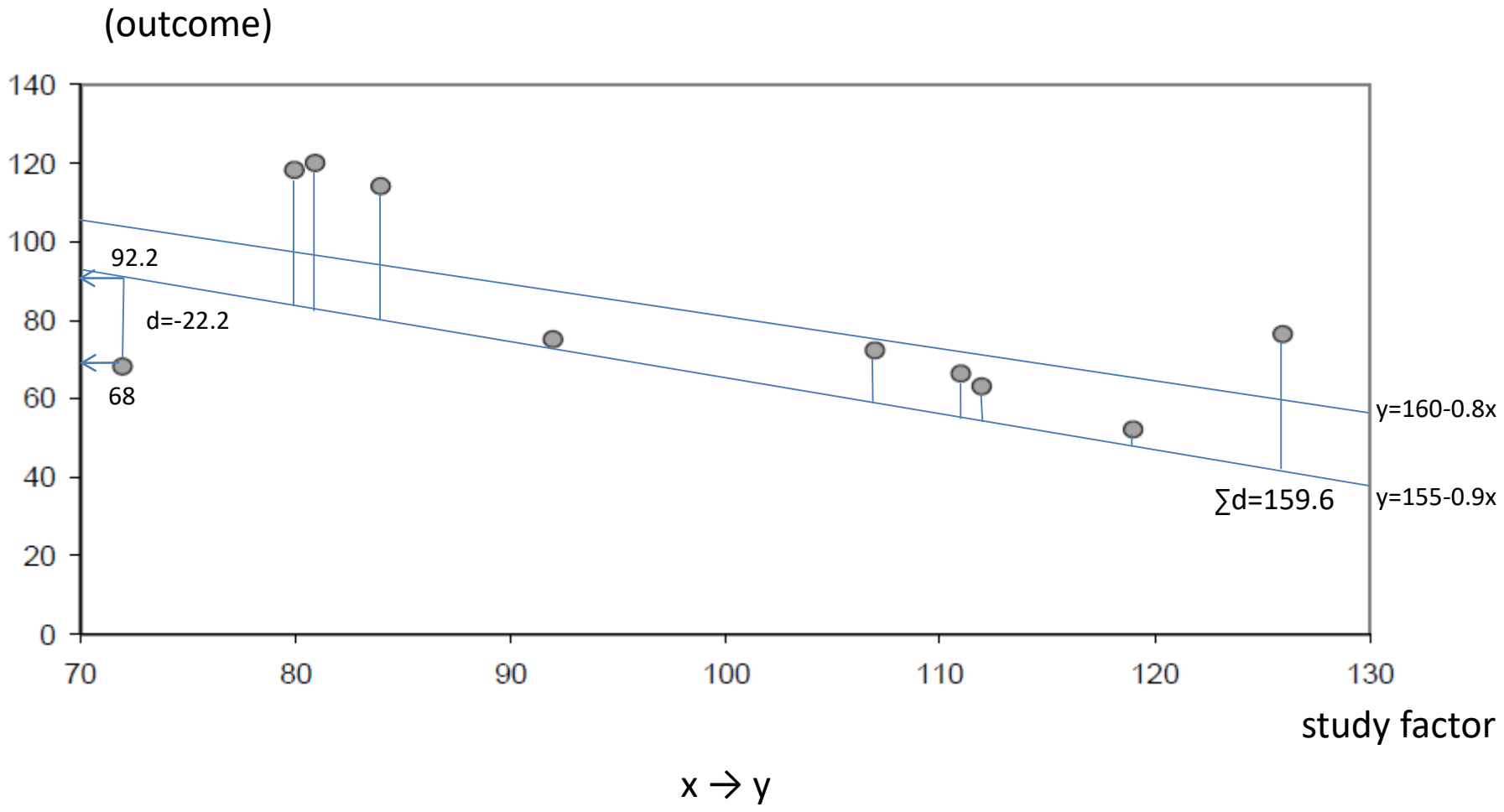
x → y

How to choose the best line?

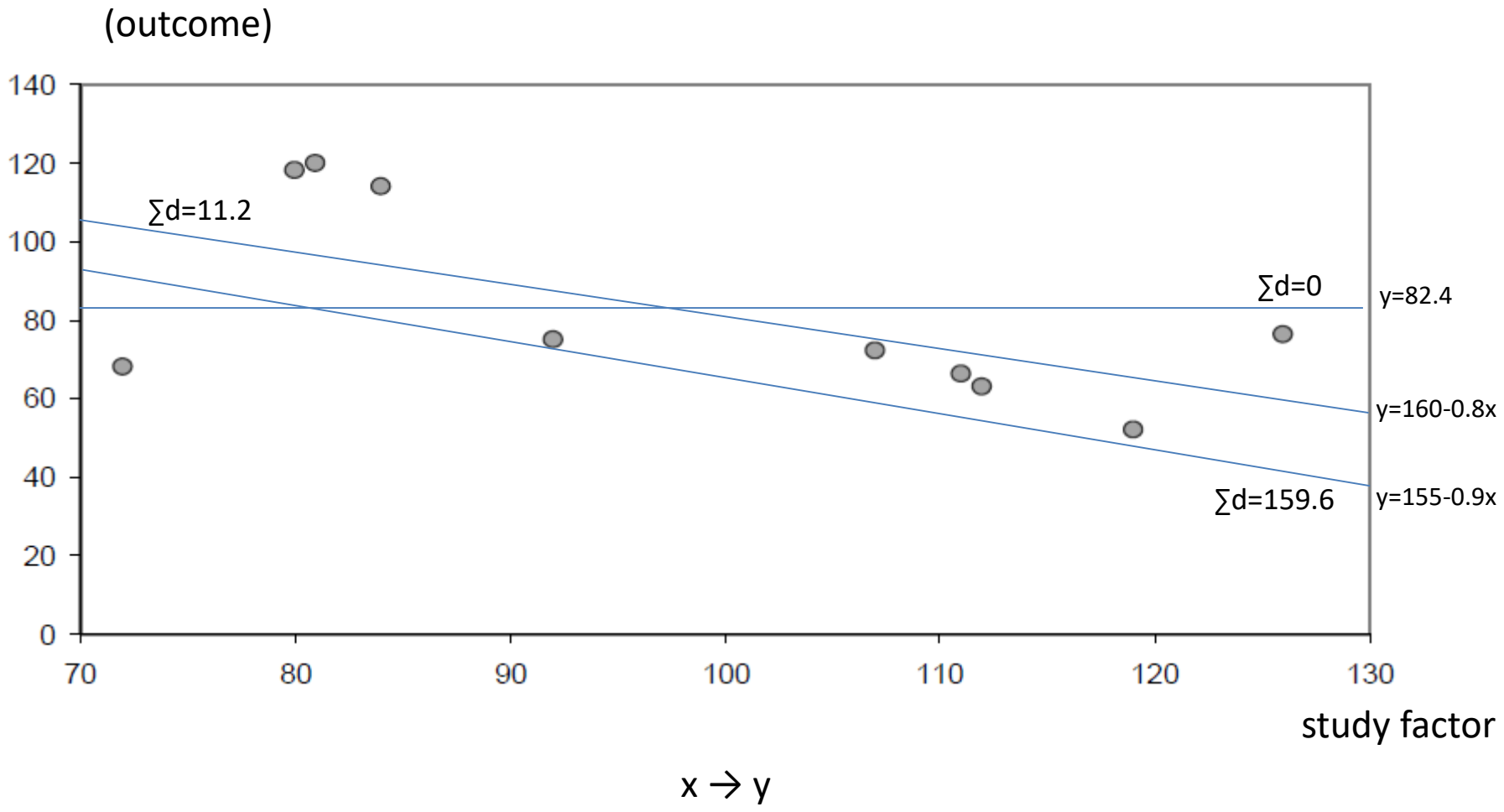


x → y

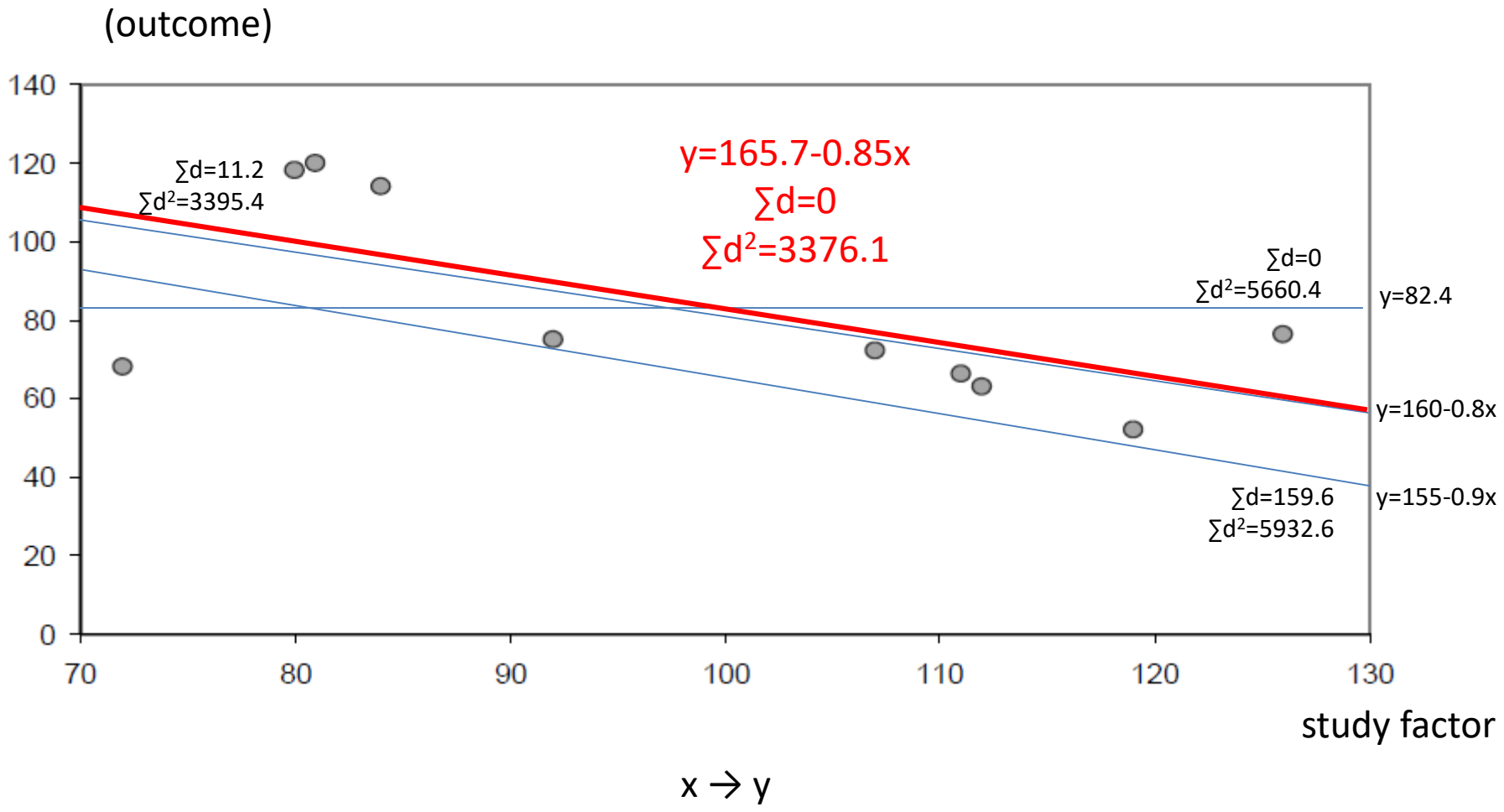
How to choose the best line?



How to choose the best line?

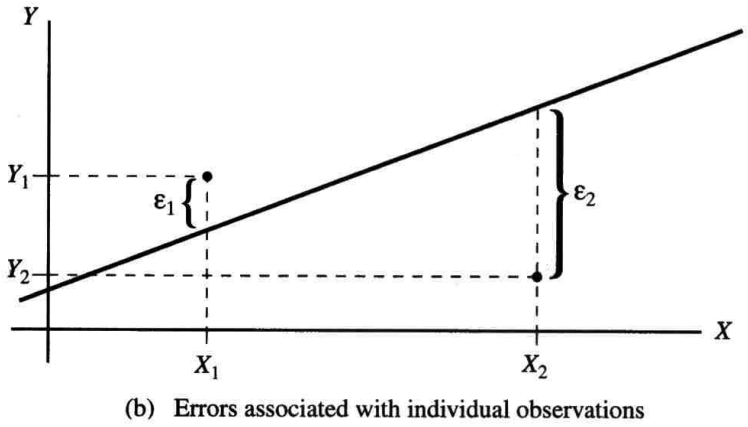


How to choose the best line?



Error terms (Residuals)

$$\varepsilon_i = y_i - \alpha - \beta x_i$$



(b) Errors associated with individual observations

The true "**unobserved**" association

$$E(y) = \alpha + \beta x$$

is estimated by

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

where *a and b are the least squares estimates* of α and β

Observed error

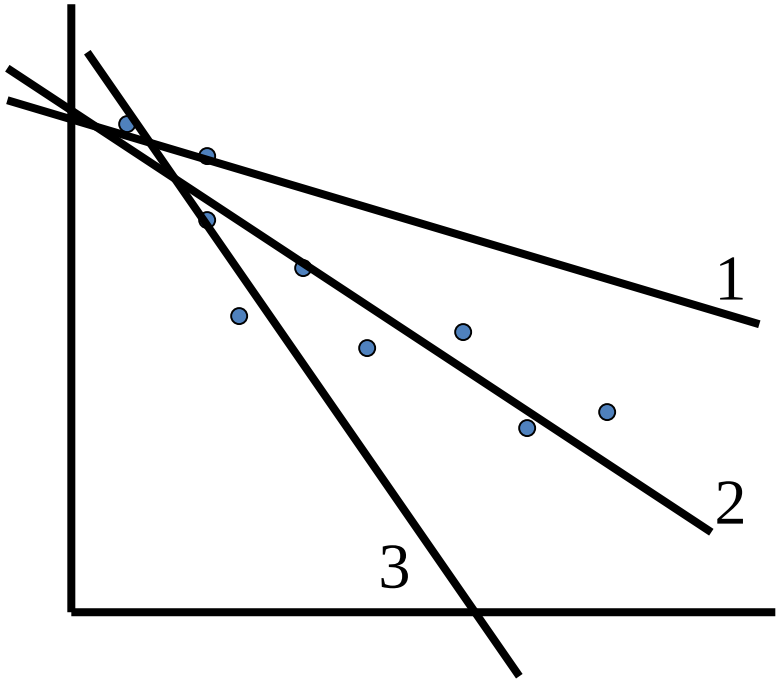
$$e_i = y_i - a - b\, x_i$$

Least Squares Regression

$$\text{minimize} \quad \sum_{i=1}^{n} (y_i - [a + bx_i])^2$$

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

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



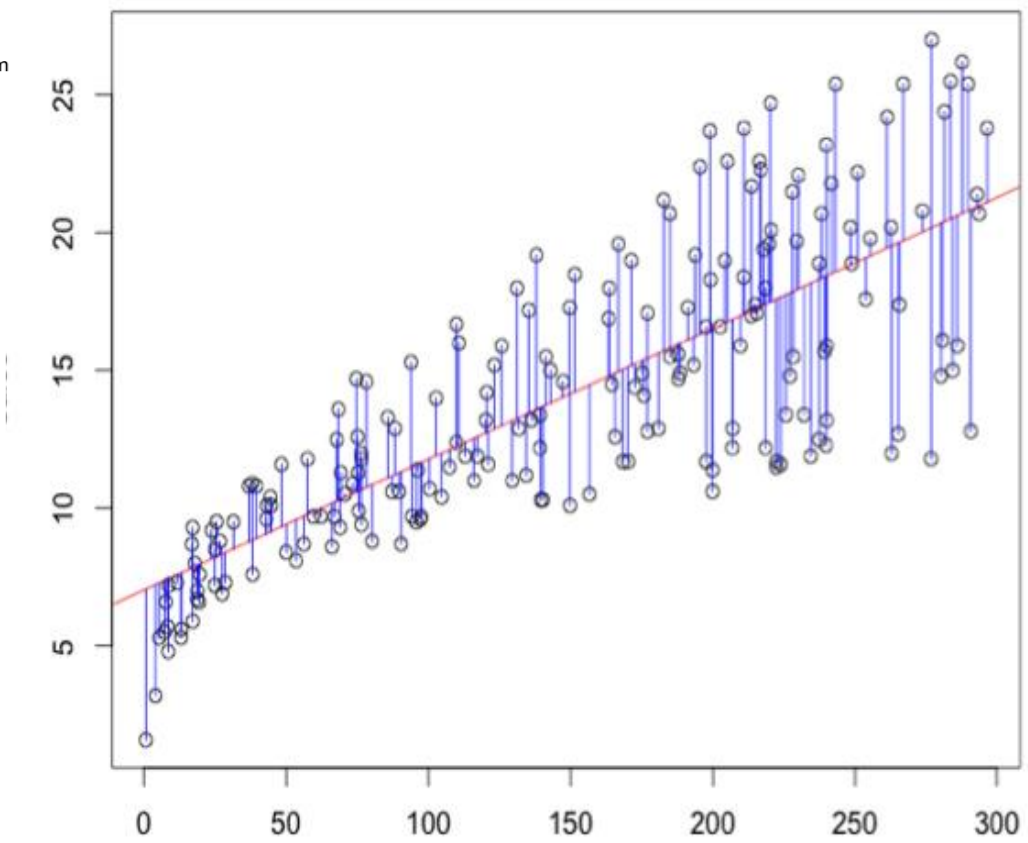
Summary of a Simple Regression

$$y = \alpha + \beta x$$

Y intercept, gradient, error term

$$y = \underbrace{\alpha + \beta x}_{\text{Model}} + \varepsilon_i$$

$$error = Y_{actual} - Y_{predict}$$



Note:
The differences between actual and predicted values of the dependent variable *y.* The difference between Actual Y value and its corresponding value on the regression line…

The goal is to find the best fitting line that would give you the least/minimize the sum of squared errors.

Testing the significance of the slope

- To test H_0: β=0,

  Calculate t = b/SE(b)

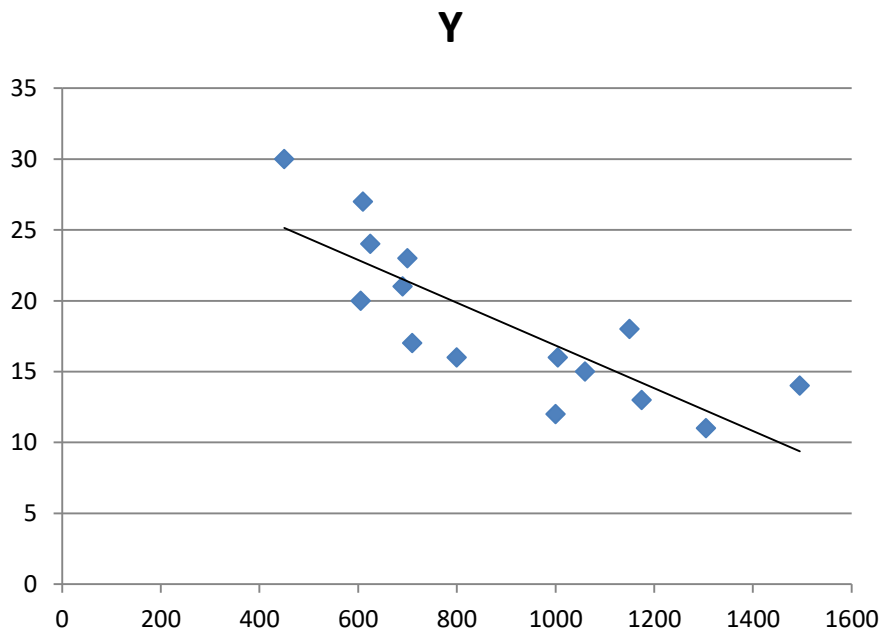
  and refer to a t-distribution with *n - 2 degrees* of freedom

- 95% CI for β:

  $b \pm t_{n-2,\ 0.05} \times SE(b)$

Construct a scatter diagram in EXCEL (Scatter Plot)

- Method 1: Select *Chart/Add Trendline*
- Method 2: Select data series; right click

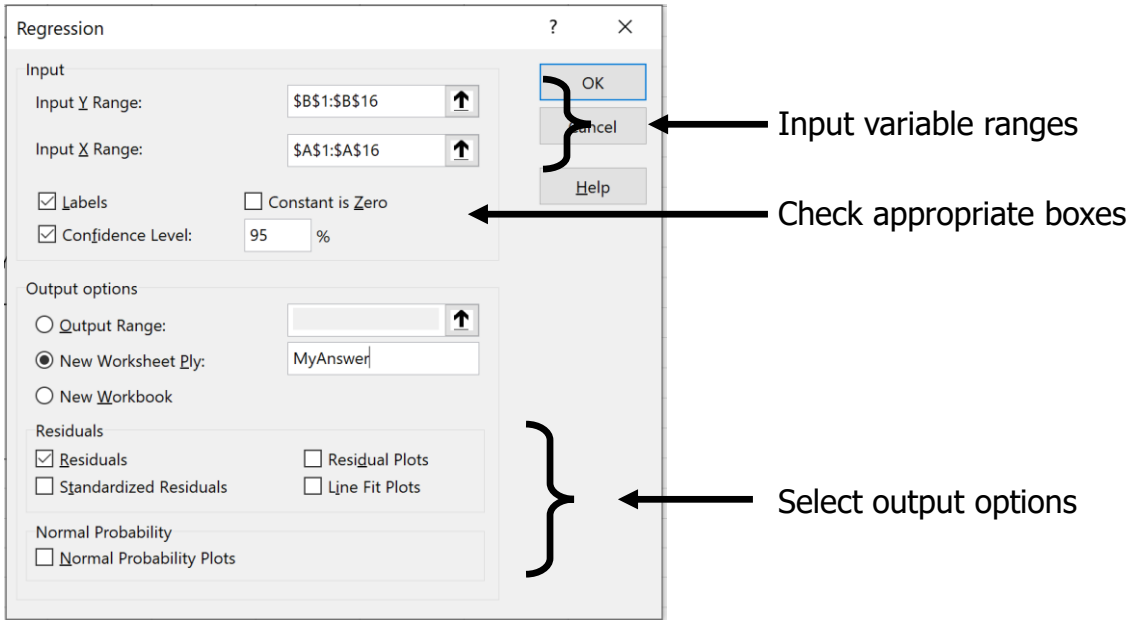


Excel Regression Tool

How does x_1 affect y?

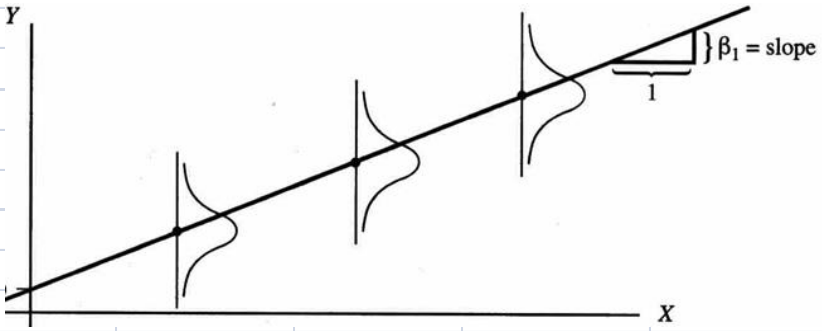
$$y = \alpha + b_1 x_1 + e$$

Excel menu > *Tools* > *Data Analysis* > *Regression*



Interpreting Excel results

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.813680325
R Square	0.662075671
Adjusted R Square	0.636081492
Standard Error	3.380903287
Observations	15



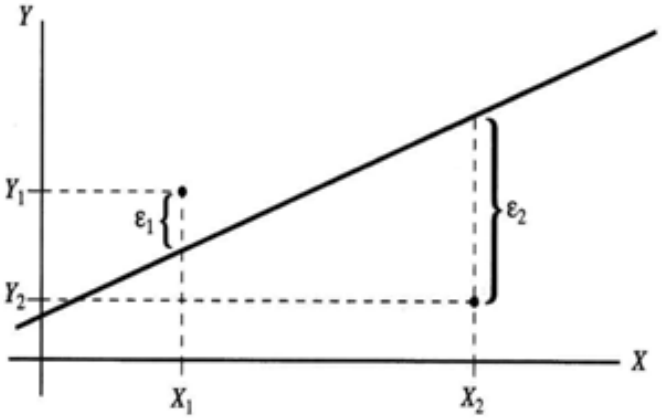
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	291.1367419	291.1367419	25.47015115	0.000223707
Residual	13	148.5965915	11.43050704		
Total	14	439.7333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	31.91255948	2.803609497	11.3826692	3.92706E-08	25.85572941	37.96938955
X	-0.015073871	0.00298682	-5.046796127	0.000223707	-0.021526503	-0.008621239

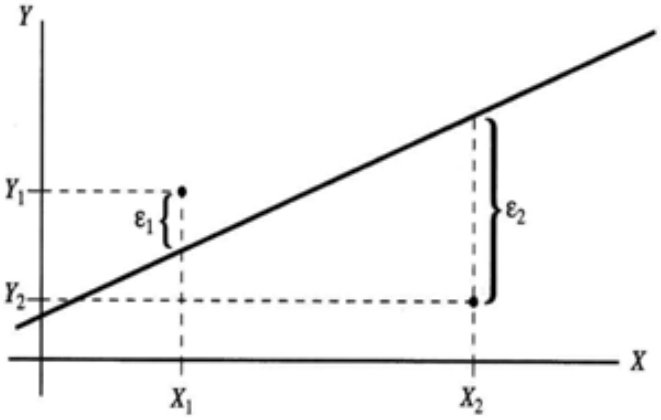
- $t_{13} = -0.0151 / 0.0029 = -5.05, P<0.001$
- 95% CI for β: $b \pm t_{n-2,\ 0.05} \times SE(b)$
  $= -0.0151 \pm 2.16 \times 0.003 \quad = \ -0.022$ to -0.0086

Analysis of Variance (1)



- Measure of variation
  - Variation between the observations and the mean
  - Variation between the predicted values using the regression line and the mean
  - Variation between the individual observations and the predicted values

Analysis of Variance (2)



The true regression

$$E(y) = \alpha + \beta x$$

can be written as

$$y = \alpha + \beta x + \varepsilon$$

where ε is a term for error, and ε is assumed to be normally distributed with mean *0* and variance σ^2

Fitted regression line: $\hat{y} = a + bx$

Residual e

= distance between observed point y and fitted point $\hat{y}$

$= y - \hat{y}$

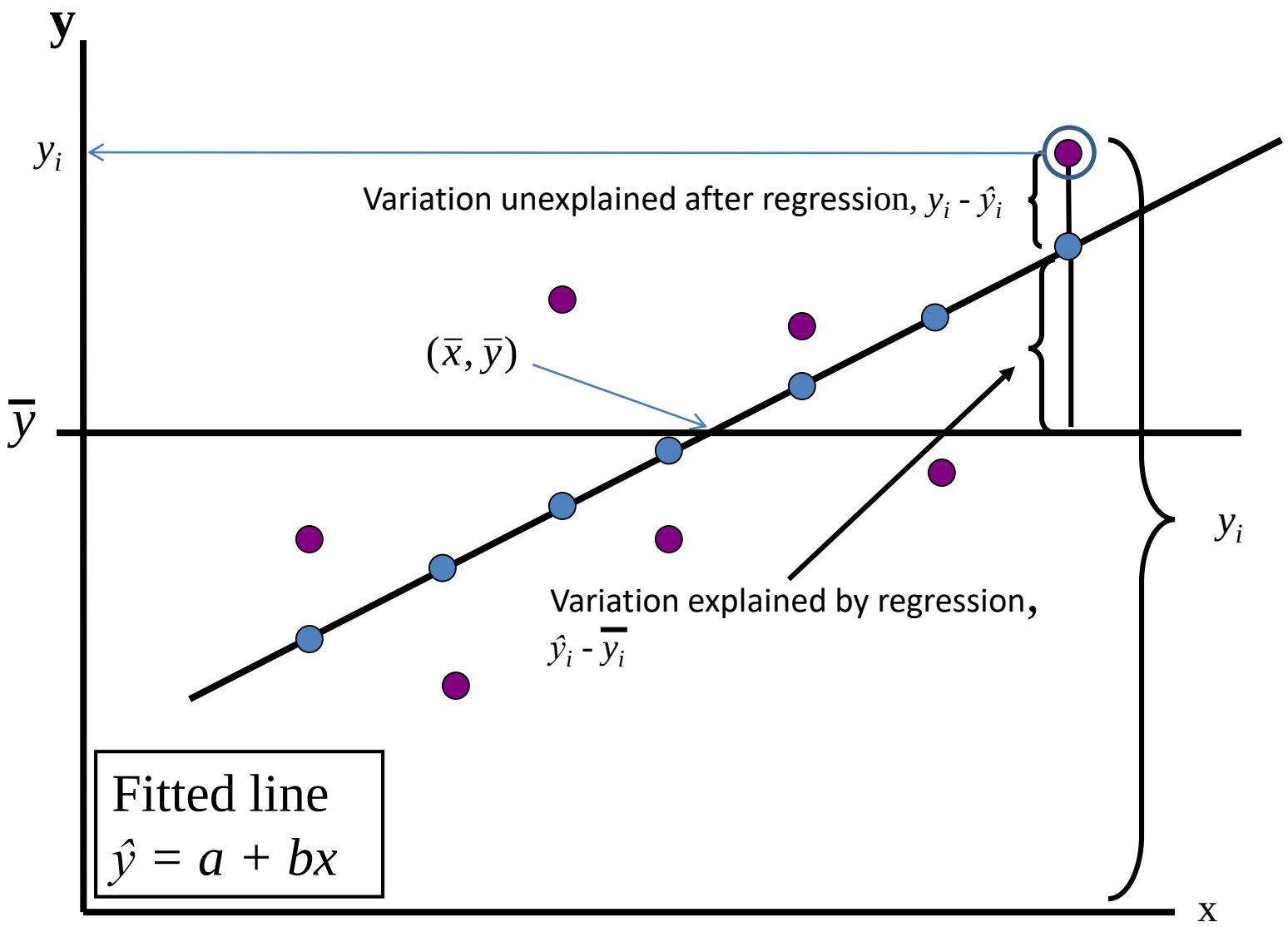
$= y - (a + bx)$

The standard deviation, *s, measures the* scatter of the observations about the fitted line (with n-2 degrees of freedom): $s^2 = \frac{\sum e^2}{n-2}$

Analysis of Variance (3)

- Observed values y_i
- Fitted values $\hat{y}_i$

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



Analysis of Variance (4)

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

- Squaring and summing gives:

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

↓ Total SS ↓ Residual SS or Error SS ↓ Regression SS

	SS_T	=	SS_E	+	SS_R
Degrees of freedom	n-1	=	n-2	+	1

Total Variability (SS_T)	
Variability around regression line (SS_E)	Variability due to regression line (SS_R)

Analysis of Variance (5)

Coefficient of Determination (R^2)

- $R^2 = SS_R/SS_T = (SS_T - SS_E)/SS_T = 1 - SS_E/SS_T$ = coefficient of determination or R^2 (R-Squared): the proportion of variation explained by the independent variable (regression model)

$$0 \leq R^2 \leq 1$$

- If every point in the sample were on the regression line, then 100% of the variability in Y could be explained by the regression equation, $R^2 = 1$.
- The lowest possible value is 0, indicating that X explains 0% of the Y variability.
- Sample correlation coefficient (r)

  $-1 \leq r \leq 1$

  r = 1 => perfect positive correlation

  r = -1 => perfect negative correlation

  r = 0 => no correlation

Analysis of Variance (6)

Standard Error of the Estimate

- Define Mean Square: MS = SS/df
- Error Mean Square (MS_E) = $SS_E/(n-2)$

  an unbiased estimate of the "variance of the errors" about the regression line (s^2, an estimate of σ^2)

- Standard error of the estimate is *s* (measures the spread of data about the line)

Analysis of Variance (7)

Hypothesis Testing

- H_0: There is no association between *y and x* that is, no regression of *y on x* that is, β=0
- A test of the significance of the regression is given by the test statistic:

$$F = \frac{MS_R}{MS_E} \quad with \quad 1 \quad and \quad n-2 \quad df$$

- If H_0 is true, F should be 1
- If H_0 is false, F > 1
- F is compared to the F distribution with 1 and n-2 degrees of freedom
- The F-test/stats of overall significance indicates whether your linear regression model provides a better fit to the data than a model that contains no independent variables. If the p-value is less than the significance level, your sample data provide sufficient evidence to conclude that your regression model fits the data better than the model with no independent variables.

Interpreting Excel results

$F_{1,13}=25.47, P<0.001$

Degrees of freedom

Sum of squares

Mean squares

F=291.14/11.43 = 25.47

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.813680325
R Square	0.662075671
Adjusted R Square	0.636081492
Standard Error	3.380903287
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	291.1367419	291.1367419	25.47015115	0.000223707
Residual	13	148.5965915	11.43050704		
Total	14	439.7333333			

MS_R=291.14/1

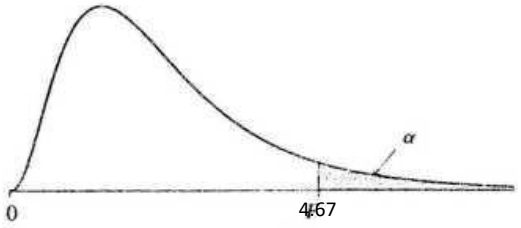
MS_E=148.60/13

n-2

n-1

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	31.91255948	2.803609497	11.3826692	3.92706E-08	25.85572941	37.96938955
X	-0.015073871	0.00298682	-5.046796127	0.000223707	-0.021526503	-0.008621239

xcel results



$\alpha = .05$

df_2	df_1 = 1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

Source: From Table V of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, published by Longman Group Ltd., London, 1974. (Previously published by Oliver & Boyd, Edinburgh.) Reprinted by permission of the authors and publishers.

ın

ʻes

F=291.14/11.43

	F	*Significance F*
.9	25.47015115	0.000223707
14		

MS_R=291.14/1

MS_E=148.60/13

	P-value	*Lower 95%*	*Upper 95%*
)2	3.92706E-08	25.85572941	37.96938955
:7	0.000223707	-0.021526503	-0.008621239

P<0.001

Interpreting Excel results

correlation: r

coefficient of determination: $R^2 = SS_R/SS_T = 291.14/439.73$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.813680325					
R Square	0.662075671					
Adjusted R Square	0.636081492					
Standard Error	3.380903287					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	291.1367419	291.1367419	25.47015115	0.000223707	
Residual	13	148.5965915	11.43050704			
Total	14	439.7333333				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	31.91255948	2.803609497	11.3826692	3.92706E-08	25.85572941	37.96938955
X	-0.015073871	0.00298682	-5.046796127	0.000223707	-0.021526503	-0.008621239

$F_{1,13}=25.47,\ P<0.001 \approx (t_{13})^2$

Reporting: Regression

- Coefficient estimate βs
- P-value of each β estimate
- Test for model significance *F*
- Variation explained R^2

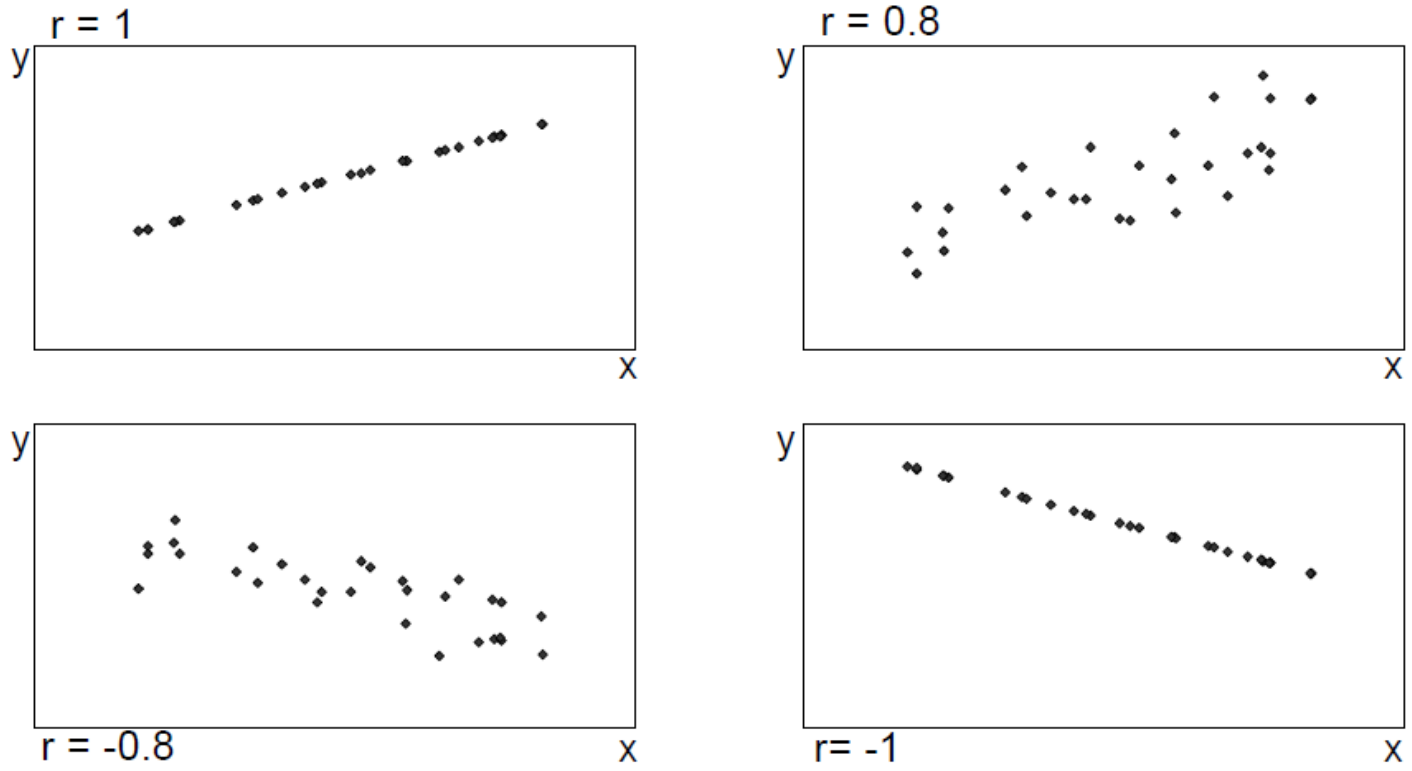
$$y = \alpha + \beta x$$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.813680325					
R Square	0.662075671					
Adjusted R Square	0.636081492					
Standard Error	3.380903287					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	291.1367419	291.1367419	25.47015115	0.000223707	
Residual	13	148.5965915	11.43050704			
Total	14	439.7333333				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	31.91255948	2.803609497	11.3826692	3.92706E-08	25.85572941	37.96938955
X	-0.015073871	0.00298682	-5.046796127	0.000223707	-0.021526503	-0.008621239

- **Conclusion:**
  - Model: *Y* = 31.91 - 0.015*X*
  - There is strong evidence that as x increases y decreases (t_{13} = -5.047, p <0.001). For every 1 unit increase in x, y is expected to decrease by 0.015 units with 95% CI: -0.022, -0.009.
  - The coefficient of determination R^2 is 0.662 – meaning the proportion of variation of y explained by the explanatory variable x is 66.2%.
  - The F statistic is 25.47 (much bigger than 1). Hence, this is strong evidence of association between the explanatory variable x and y. The regression model is found to be statistical significant.

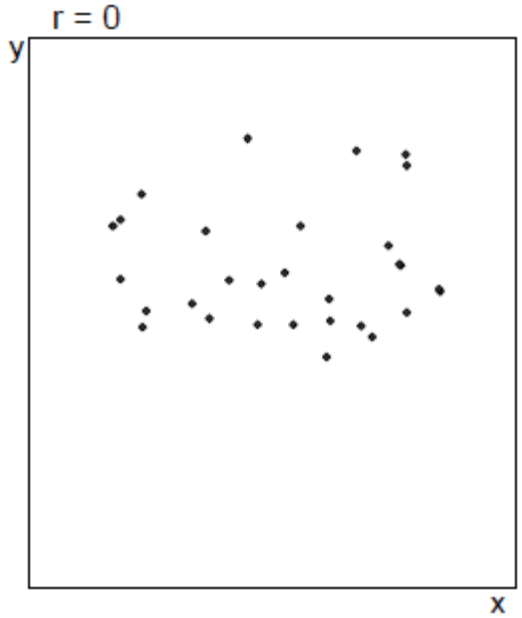
Correlation

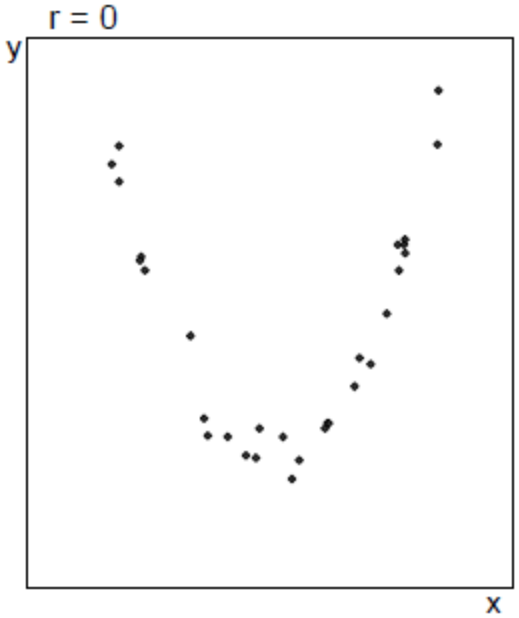
- *r* - product-moment correlation coefficient
- a single measure of association – measures the strength of the straight-line relationship
- combines information about slope and scatter about the line



Correlation

- no association means *r = 0, BUT*
- *r = 0 does not automatically mean no* association

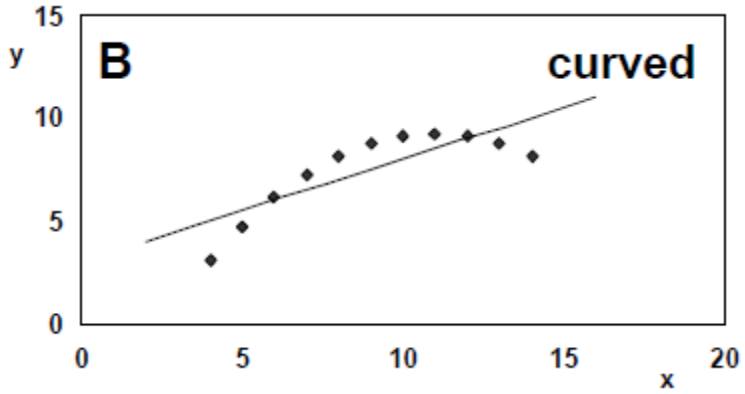


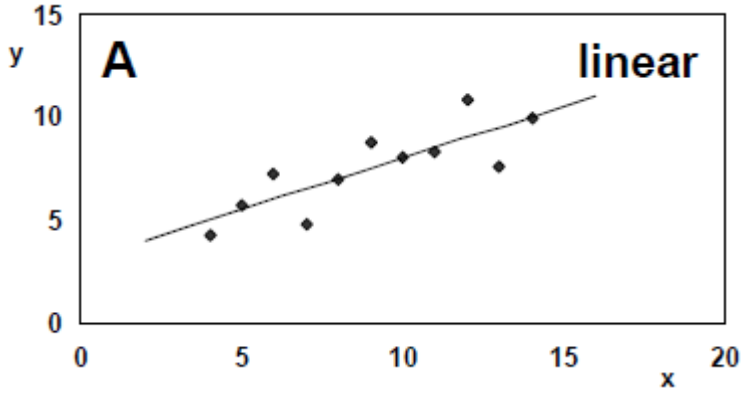


- *Cautions:*
  - Avoid extrapolation beyond the range of the data
  - Always plot the data
  - Association does not imply causation (e.g. 2 variables that are both influenced by time are often associated)

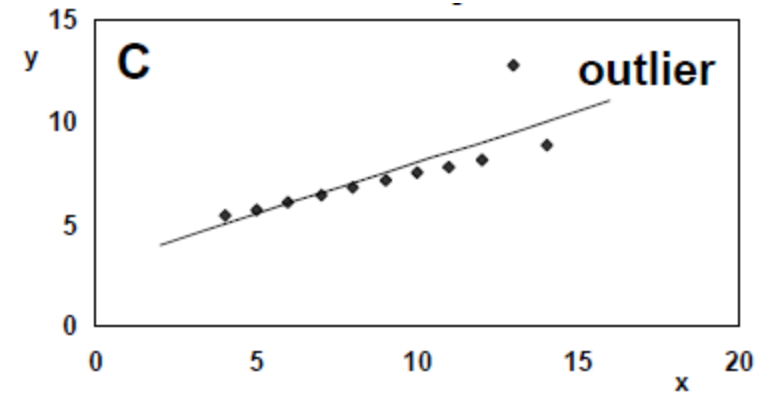
4 different regressions

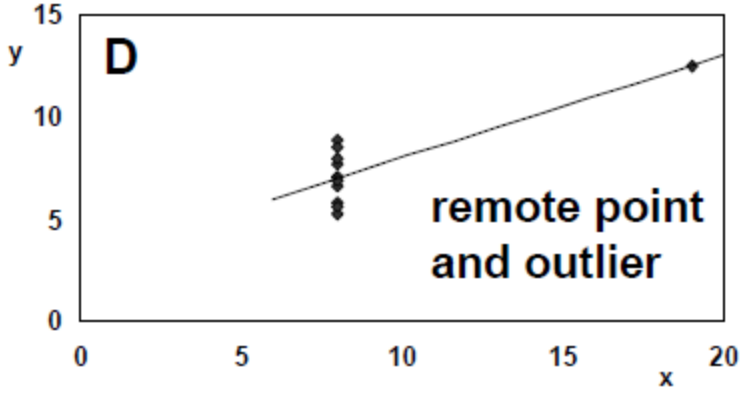
Data from Anscombe, FJ (1973). Graphs in statistical analysis. American Statistician, 27, 17–21





	n	$\bar{x}$	$\bar{y}$	r	Regression equation
A	11	9.0	7.5	0.82	Y = 3 + 0.5x
B	11	9.0	7.5	0.82	Y = 3 + 0.5x
C	11	9.0	7.5	0.82	Y = 3 + 0.5x
D	11	9.0	7.5	0.82	Y = 3 + 0.5x





Some Questions to Ask

- Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Choosing the important variables

- The most direct approach is called **all subsets** or **best subsets** regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size

- However we often can't examine all possible models, since they are 2^p of them; for example when $p = 40$ there are over a billion models! Instead we need an automated approach that searches through a subset of them. We discuss the commonly use approaches

Forward Selection

- Begin with the **null model** — a model that contains an intercept but no predictors
- Fit p simple linear regressions and add to the null model the variable that results in the lowest p-value
- Add to that model the variable that results in the lowest
- p-value amongst all two-variable models
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold

Backward Selection

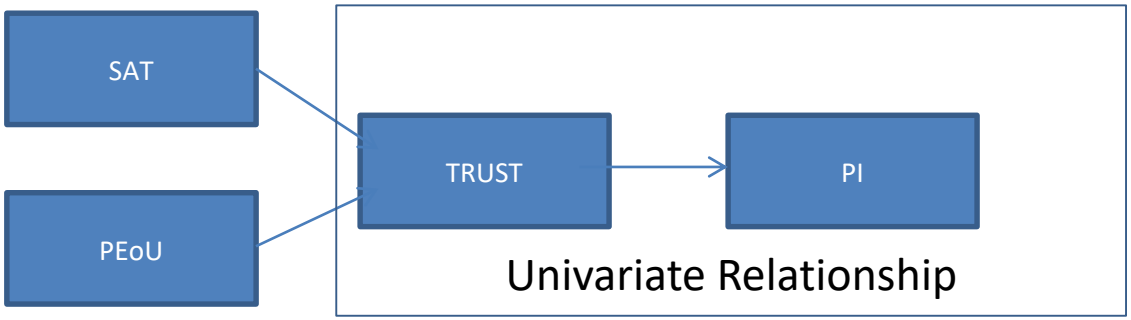
- Start with all variables in the model
- Remove the variable with the largest *p*-value — that is, the variable that is the least statistically significant
- The new ($n - 1$) variable model is fit, and the variable with the largest *p*-value is removed
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant *p*-value defined by some significance threshold

Stepwise Selection

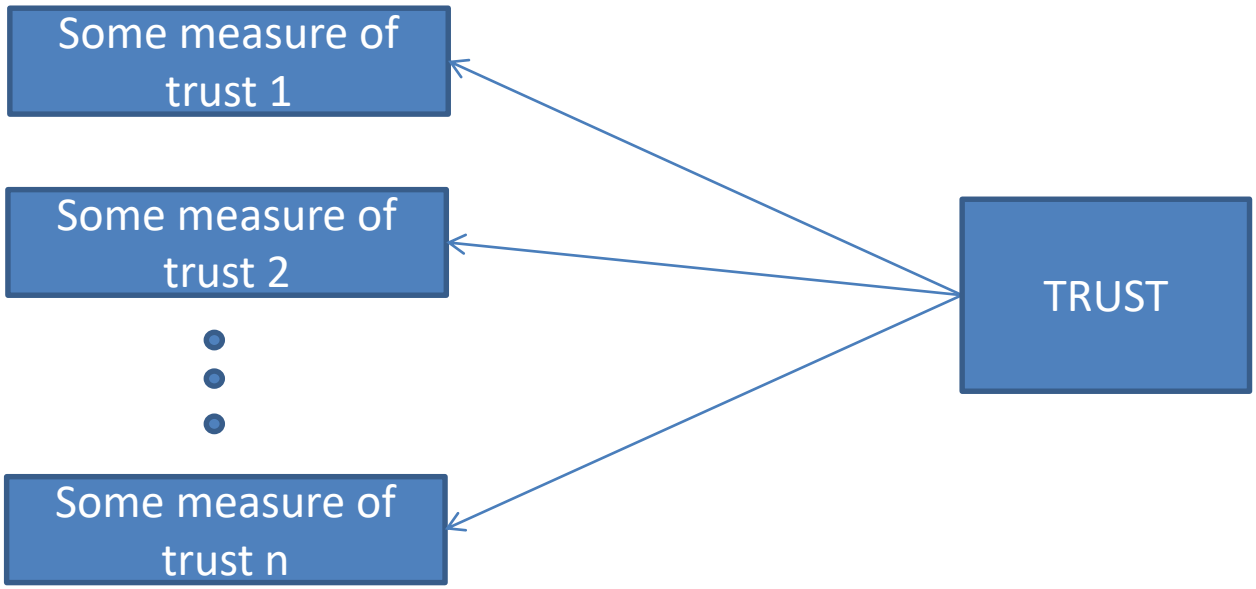
- A hybrid between forward and backward that will have two thresholds *p*-value for entry and removal
- Start similarly to forward selection
- Every time a new variable is added (with *p*-value less than the entry threshold) check all the variables in the model, and remove those with *p*-value higher than removal threshold

Univariate Relationship

- The perceived trustworthiness (Trust) of the website will in turn affect customers' purchase intention (PI).



Non-directly Observable Variables:



Likert Scales are sometimes used

Regression Results
TRUST→PI

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.517318229
R Square	0.26761815
Adjusted R Square	0.260576017
Standard Error	1.151153367
Observations	106

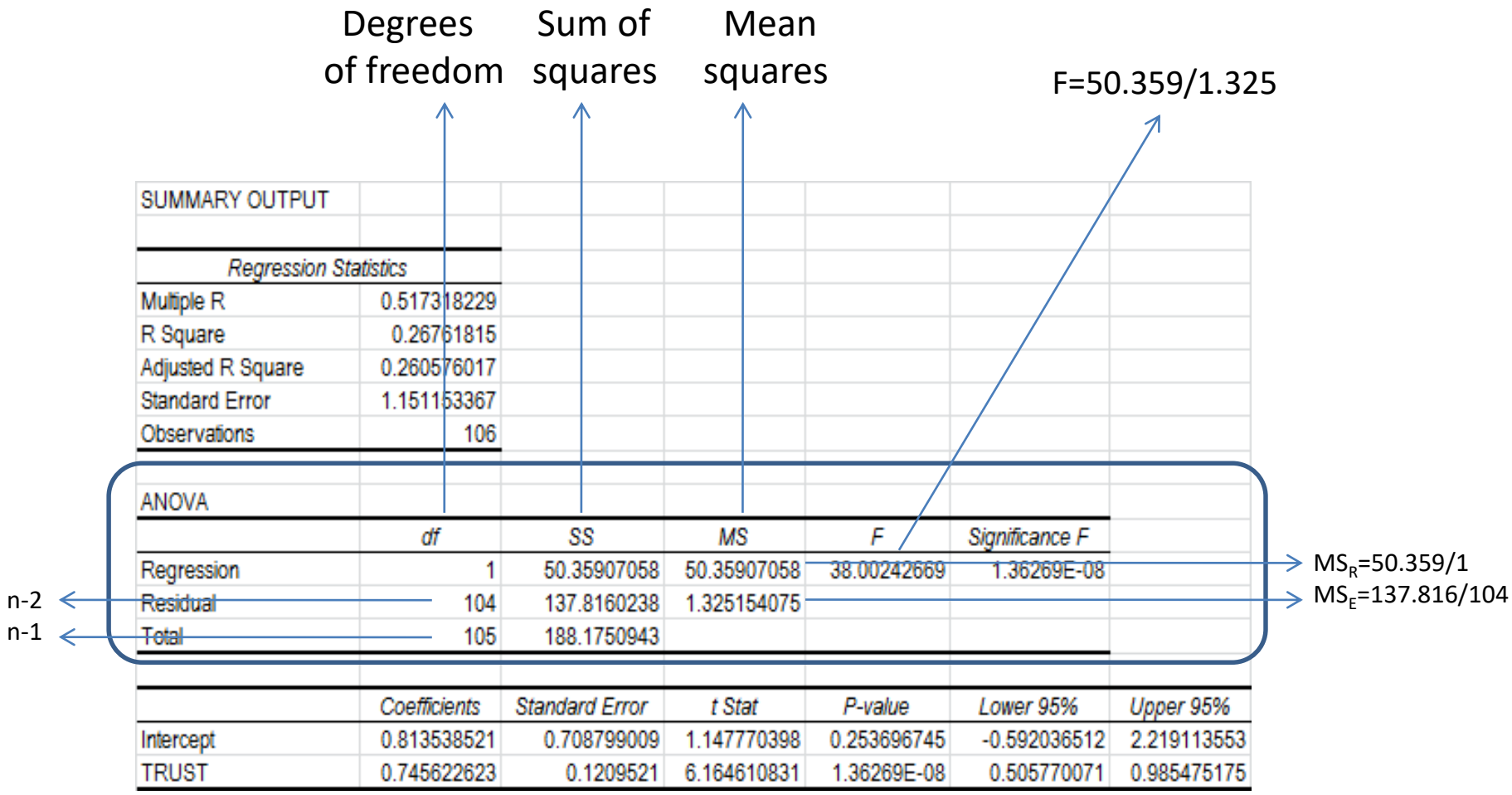
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	50.35907058	50.35907058	38.00242669	1.36269E-08
Residual	104	137.8160238	1.325154075		
Total	105	188.1750943			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.813538521	0.708799009	1.147770398	0.253696745	-0.592036512	2.219113553
TRUST	0.745622623	0.1209521	6.164610831	1.36269E-08	0.505770071	0.985475175

- t_{104} = -0.7456 / 0.1210 = 6.1646, P<0.001
- 95% CI for β: $b \pm t_{n-2,\ 0.05} \times SE(b)$
  = 0.7456 ± 2.16 x 0.121 = 0.5058 to 0.9855

Regression Results
TRUST→PI



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.517318229
R Square	0.26761815
Adjusted R Square	0.260576017
Standard Error	1.151153367
Observations	106

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	50.35907058	50.35907058	38.00242669	1.36269E-08
Residual	104	137.8160238	1.325154075		
Total	105	188.1750943			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.813538521	0.708799009	1.147770398	0.253696745	-0.592036512	2.219113553
TRUST	0.745622623	0.1209521	6.164610831	1.36269E-08	0.505770071	0.985475175

$F_{1,104}$=38.002, P<0.001

Regression Results
TRUST→PI

correlation : r

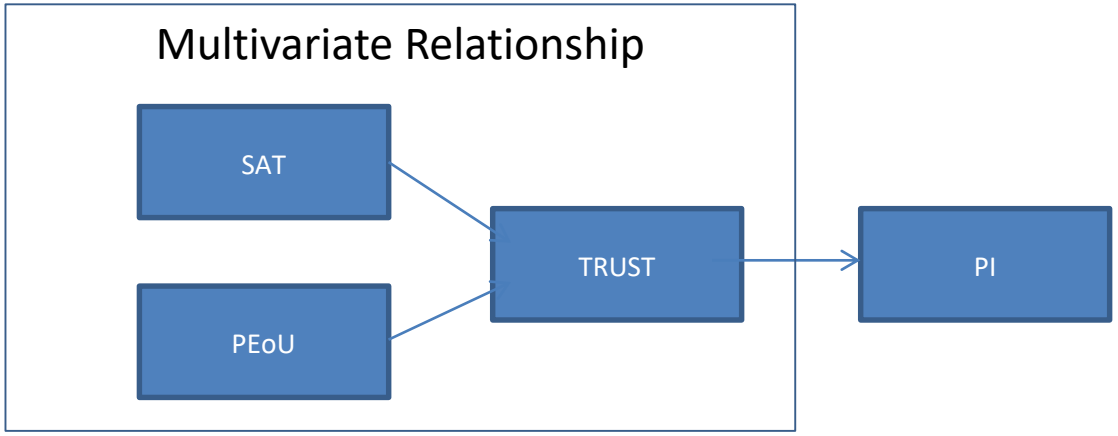
coefficient of determination: $R^2 = SS_R/SS_T = 50.359/188.175$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.517318229					
R Square	0.26761815					
Adjusted R Square	0.260576017					
Standard Error	1.151153367					
Observations	106					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	50.35907058	50.35907058	38.00242669	1.36269E-08	
Residual	104	137.8160238	1.325154075			
Total	105	188.1750943				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.813538521	0.708799009	1.147770398	0.253696745	-0.592036512	2.219113553
TRUST	0.745622623	0.1209521	6.164610831	1.36269E-08	0.505770071	0.985475175

$F_{1,104}=38.0024$, $P<0.001 \approx (t_{104})^2$

Multivariate Relationship

- The **perceived trustworthiness (TRUST)** of the website will in turn affect customers' **purchase intention** (**PI**).
- The trustworthiness perception of the online operation of a company would be positively affected by **customers' satisfaction (SAT)** with the offline operation of the company.
- The trustworthiness perception of the online operation of a company would be positively affected by the **perceived ease of use (PEoU)** of the website.



Model Results

$TRUST = \alpha + b_1 SAT + e$

ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	0.14785014	0.14785014	0.170029773	0.680934096	
Residual	104	90.43365929	0.869554416			
Total	105	90.58150943				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.936284743	0.373682785	15.88589301	1.42248E-29	5.195257758	6.677311729
SAT	-0.037514637	0.090978394	-0.412346667	0.680934096	-0.217928206	0.142898931

$TRUST = \alpha + b_2 PEoU + e$

ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	15.19312472	15.19312472	20.95926285	1.30451E-05	
Residual	104	75.38838471	0.724888315			
Total	105	90.58150943				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3.528652729	0.500129313	7.055480723	1.97738E-10	2.536877533	4.520427926
PEoU	0.385385301	0.084179656	4.57812875	1.30451E-05	0.218453885	0.552316717

$TRUST = \alpha + b_1 SAT + b_2 PEoU + e$

ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	15.47948344	7.739741721	10.61480548	6.43462E-05	
Residual	103	75.10202599	0.729145883			
Total	105	90.58150943				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3.724898982	0.591322079	6.299272621	7.47297E-09	2.552151105	4.89764686
SAT	-0.052247778	0.083371926	-0.626683114	0.532253517	-0.217596331	0.113100774
PEoU	0.387425808	0.084489269	4.585503165	1.27807E-05	0.219861271	0.554990345

Multiple linear regression model

- Multiple linear regression model:

$$y = \alpha + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + e$$

- Predicted model:

$$\hat{y} = \alpha + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

- The b's are called partial regression coefficients.

Excel Regression Tool

How do age and MBA degree affect employee salaries?

$$y = \alpha + b_1 x_1 + b_2 x_2 + e$$

where

y = salary

x_1 = age

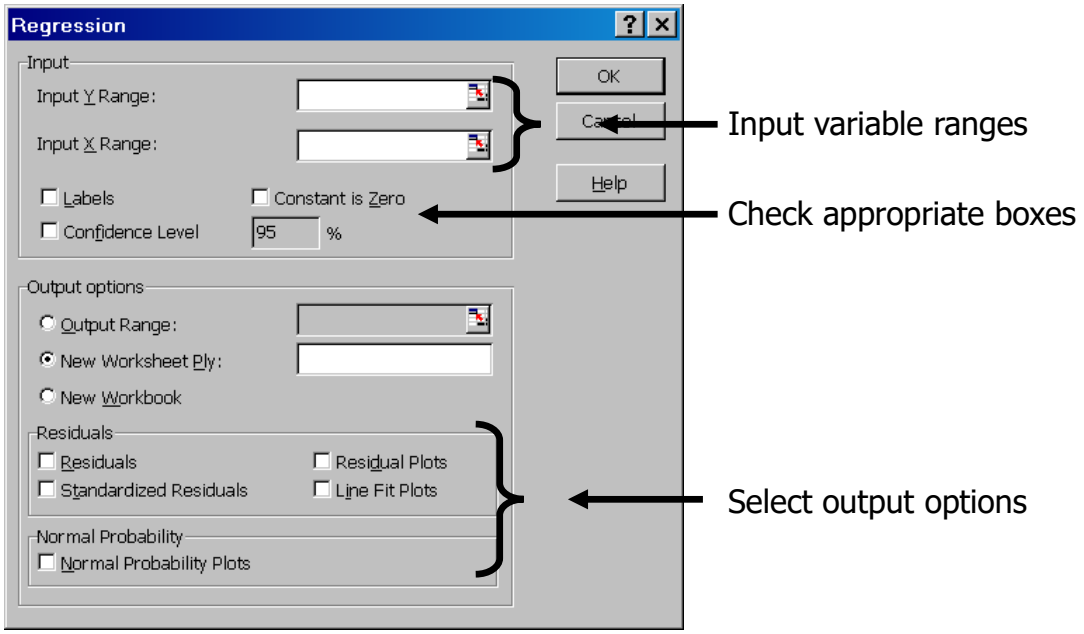
x_2 = MBA indicator (0 = No; 1 = Yes)

Will be done in Tutorial

$$y = \alpha + b_1 Age + b_2 MBA + e$$

	A	B	C	D	E
1	Salary Data				
2					
3	Employee	Salary	Age	MBA	MBA*Age
4	1	28260	25	0	0
5	2	43392	28	1	28
6	3	56322	37	1	37
7	4	26086	23	0	0
8	5	36807	32	0	0
9	6	57119	57	0	0
10	7	48907	45	0	0
11	8	34301	32	0	0
12	9	31104	25	0	0
13	10	60054	57	0	0

Excel menu > *Tools* > *Data Analysis* > *Regression*



Results

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.976118476					
5	R Square	0.952807278					
6	Adjusted R Square	0.949857733					
7	Standard Error	2941.914352					
8	Observations	35					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	5591651177	2795825589	323.0353318	6.05341E-22	
13	Residual	32	276955521.7	8654860.054			
14	Total	34	5868606699				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	893.5875971	1824.575283	0.489751015	0.627650922	-2822.947488	4610.122682
18	Age	1044.146043	42.14128238	24.77727265	1.8878E-22	958.3071326	1129.984954
19	MBA	14767.23159	1351.801764	10.92411031	2.49752E-12	12013.70383	17520.75934

Model

- Salary = 893.59 + 1044.15 Age + 14767.23 MBA
  - No MBA: Salary = 893.59 + 1044.15 Age
  - MBA: Salary = 15660.82 + 1044.15 Age

- The models suggest that the rate of salary increase for age is the same for both groups. However, individuals with MBAs might earn relatively higher salaries as they get older. In other words, the slope of *Age* may depend on the value of *MBA*. Such a dependence is called an interaction.

Interaction Model

$$y = \alpha + b_1 Age + b_2 MBA + b_3 Age*MBA + e$$

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.989321416					
5	R Square	0.978756863					
6	Adjusted R Square	0.976701076					
7	Standard Error	2005.37675					
8	Observations	35					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	5743939086	1914646362	476.098288	5.31397E-26	
13	Residual	31	124667613.2	4021535.91			
14	Total	34	5868606699				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	3902.509386	1336.39766	2.920170772	0.006467654	1176.906869	6628.111903
18	Age	971.3090382	31.06887722	31.26308786	5.23658E-25	907.94361	1034.674466
19	MBA	-2971.080074	3026.24236	-0.98177202	0.333812767	-9143.145501	3200.985353
20	MBA*Age	501.8483604	81.55221742	6.153705887	7.9295E-07	335.5214237	668.1752972

Model Results

- Salary = 3323.11 + 984.25 Age + 425.58 MBA*Age
  - No MBA: Salary = 3323.11 + 984.25 Age + 425.58 (0)*Age
    = 3323.11 + 984.25 Age
  - MBA: Salary = 3323.11 + 984.25 Age + 425.58 (1)*Age
    = 3323.11 + 1409.83 Age

Model building Strategies

- Steps for selecting the appropriate variables in a linear regression
  1. Check your data
     - ✓ missing values
     - ✓ implausible values
  2. Univariate analysis
     - ✓ plot y on each x variable
     - ✓ correlations
  3. Screening for the base model
     - ✓ Select variables to be included in the base model
  4. Fitting the base model
     - ✓ Check large changes in parameter estimates / standard errors
  5. Assessing interactions
     - ✓ Add interaction terms if necessary
  6. Significant variables
  7. Examination of residuals