Introductory Statistics Tutorial
Chapter 12 – Relationships between Quantitative Variables: Regression and Correlation

Section A: The Straight Line Graph

1. The equation of a line is of the form $y = \beta_0 + \beta_1 x$, where β_0 is the y-intercept and β_1 is the slope of the line. Give the values of β_0 and β_1 for the following lines.

(a) $y = 5 + 3x$

$\beta_0 =$

$\beta_1 =$

(b) $y = 10 - 14x$

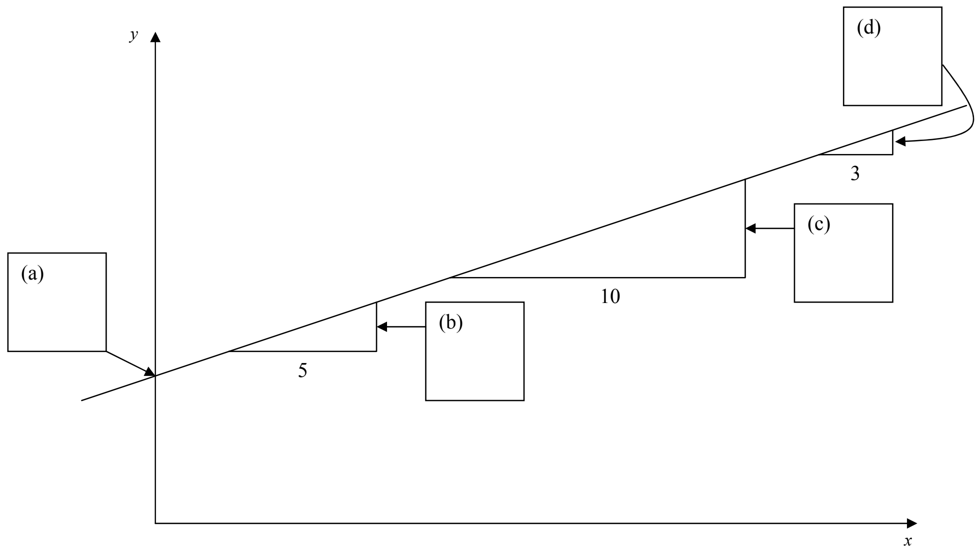
$\beta_0 =$

$\beta_1 =$

2. **(a)** Give the equation of a line that has a slope of 2 and a y-intercept of –3.

(b) Give the equation of a line that has a slope of –4 and a y-intercept of 7.

3. The equation of the line shown on the graph below is $y = 6 + 2.5x$.
Use this equation to fill in the four boxes on the graph below.

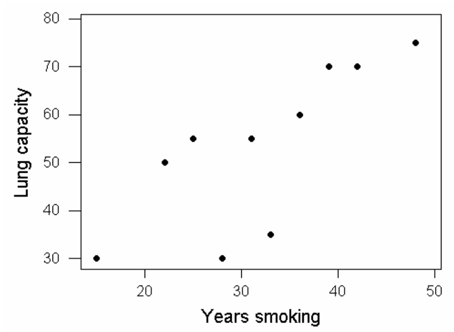


Section B: Regression

1. Observations on lung capacity, measured on a scale of 0 – 100, and the number of years smoking were obtained from a sample of emphysema patients. One of the uses of the data is to use the number of years smoking to predict lung capacity. The data is shown in the table below. A scatter plot, residual plot, Normal probability plot and *Excel* output are also shown.

Patient	1	2	3	4	5	6	7	8	9	10
Number of years smoking	25	36	22	15	48	39	42	31	28	33
Lung capacity	55	60	50	30	75	70	70	55	30	35

Scatter plot



***Excel* regression output**

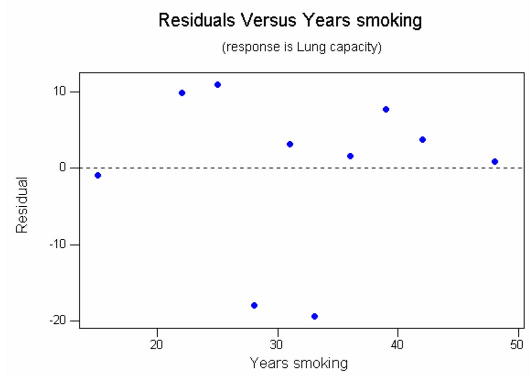
SUMMARY OUTPUT

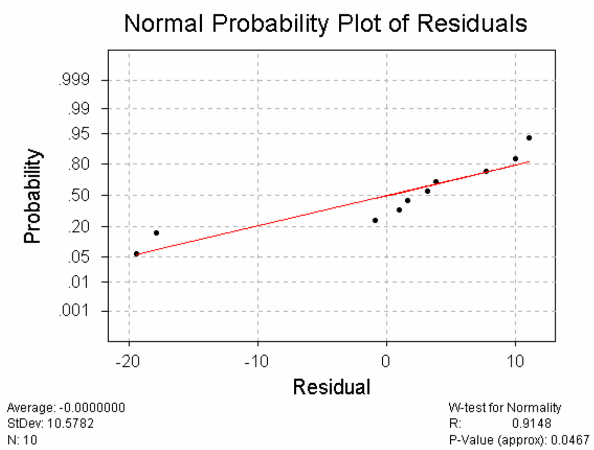
Regression Statistics	
Multiple R	0.773802257
R Square	0.598769933
Adjusted R Square	0.548616175
Standard Error	11.21989008
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1502.912533	1502.912533	11.93868522	0.008627995
Residual	8	1007.087467	125.8859334		
Total	9	2510			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	11.23788345	12.59660909	0.892135603	0.398359228	-17.80996799	40.28573489
Years smoking	1.309157259	0.37889037	3.455240255	0.008627995	0.435433934	2.182880583





(a) Write the equation of the least-squares regression line.

(b) Use the least-squares regression line to predict the lung capacity of an emphysema patient who has been smoking for 30 years.

(c) Patient 1 had smoked for 25 years and had a lung capacity of 55. Calculate the residual (prediction error) for this observation.

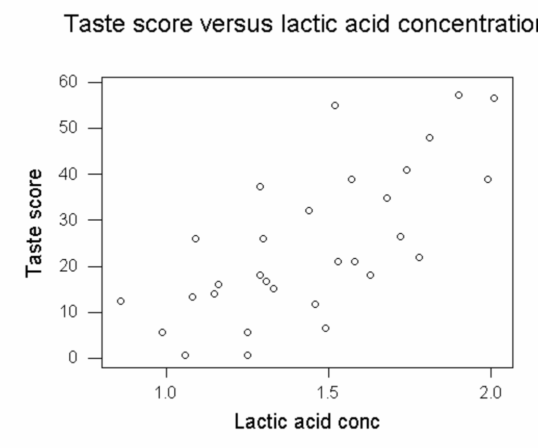
(d) Comment on the appropriateness of using a linear regression model for this data.

(e) Assume that it is appropriate to use a linear regression model for this data. (**Note:** This may not be true.) Carry out a statistical test to see if there is any evidence of a relationship between lung capacity and years of smoking. State the hypotheses and interpret the test. If there is evidence of a relationship (i.e. an effect of years of smoking on lung capacity), then describe the size of the effect.

(f) **(i)** Find the sample correlation coefficient from the *Excel* output.

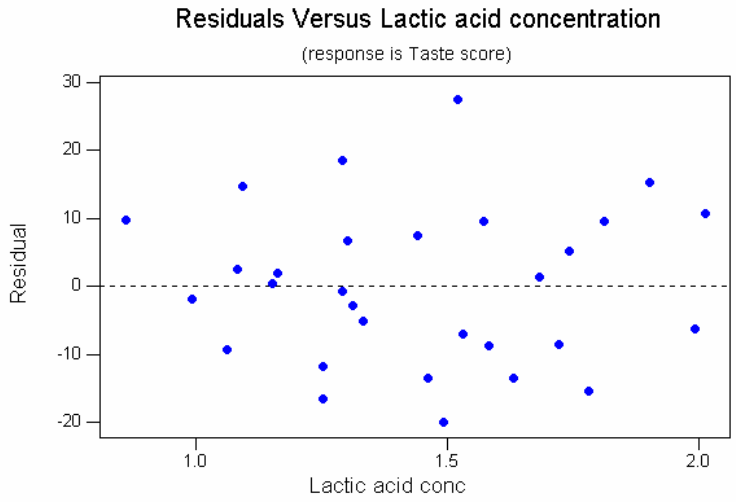
(ii) What does *Excel* call it?

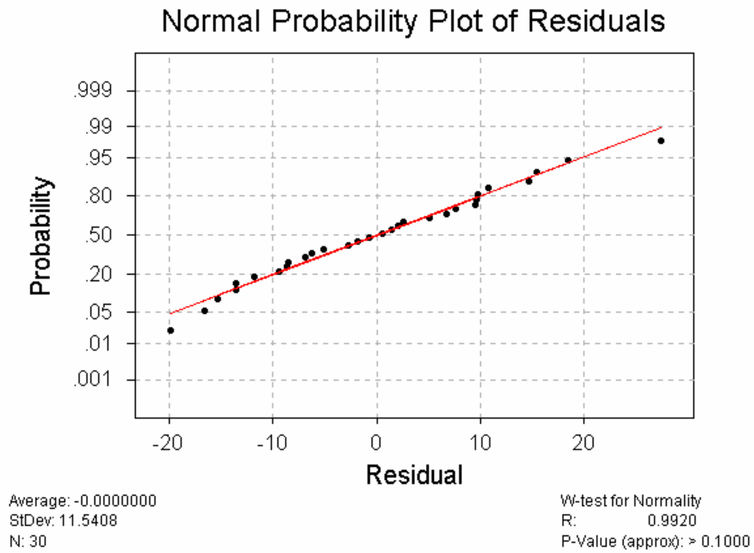
2. A study of cheddar cheese from Latrobe Valley investigated the effect on the taste of cheese of various chemical processes that occur during the aging process. One of the aims of the study was to see if the lactic acid concentration could be used to predict the taste score (a subjective measure of taste). Observations were made on 30 randomly selected samples of mature cheddar cheese. A linear regression model is fitted to the data. A scatter plot, residual plot and a Normal probability plot are given below, along with a Normality test and some MINITAB output.



Regression Analysis

```
The regression equation is
Taste score = - 29.9 + 37.7 Lactic acid conc

Predictor        Coef       StDev         T        P
Constant       -29.86       10.58     -2.82    0.009
Lactic a       37.720       7.186      5.25    0.000

S = 11.75      R-Sq = 49.6%     R-Sq(adj) = 47.8%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      3800.4      3800.4     27.55    0.000
Residual Error    28      3862.5       137.9
Total             29      7662.9
```



(a) One of the observations had a lactic acid concentration of 1.46 and a taste score of 11.6. Calculate the residual for this observation.

(b) Comment on the appropriateness of using a linear regression model for this data.

(c) Assume that it is appropriate to use a linear regression model for this data. (**Note:** This may not be true.) Carry out a statistical test to see if there is any evidence of a relationship between taste score and lactic acid concentration. State the hypotheses and interpret the test. If there is evidence of a relationship (i.e. an effect of lactic acid concentration on taste score), then describe the size of the effect. (Note: For a 95% confidence interval with df = 28, the t-multiplier is 2.048.)

(d) The researcher wanted to predict the taste score of a cheddar cheese with a lactic acid concentration of 1.8 and used MINITAB to produce the following output.

```
Predicted Values

     Fit  StDev Fit         95.0% CI             95.0% PI
   38.04       3.35   (   31.18,    44.90)  (   13.02,    63.05)
```

Use the MINITAB output to interpret the following:

(i) The "Fit" value of 38.04.

(ii) The 95% confidence interval.

(iii) The 95% prediction interval.

Part of the MINITAB output is repeated below to help you answer parts (e) and (f).

(e) The fitted least-squares regression line indicates that for each increase of 0.05 in lactic acid concentration we expect that, on average, the taste score will:

(1) increase by approximately 1.9 units.

(2) decrease by approximately 28.0 units.

(3) increase by approximately 37.7 units.

(4) increase by approximately 18.9 units.

(5) decrease by approximately 29.9 units.

(f) The fitted least-squares regression line can be used to predict taste scores for samples of mature cheddar from the Latrobe Valley. Cheese that has a lactic acid concentration of 1.30 has a predicted taste score of:

(1) 24.5

(2) 19.2

(3) 49.0

(4) 78.9

(5) 25.9

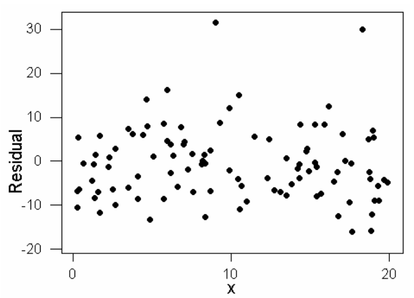
Regression Analysis

```
The regression equation is
Taste score = - 29.9 + 37.7 Lactic acid conc

Predictor        Coef       StDev          T        P
Constant       -29.86       10.58      -2.82    0.009
Lactic a       37.720       7.186       5.25    0.000
```

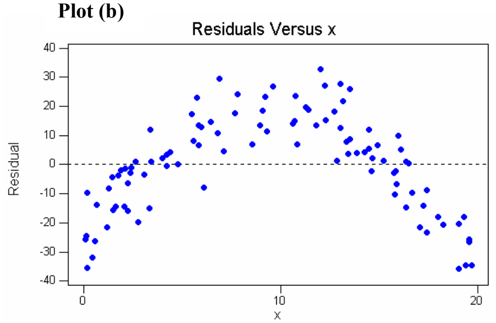
Section C: Old Exam Questions

Questions 1 and 2 refer to the following set of residual plots.

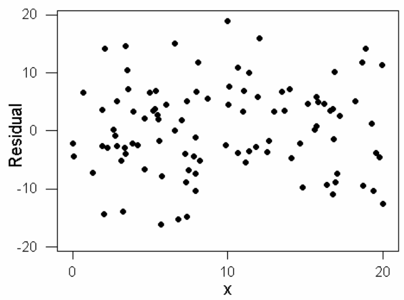
Plot (a)



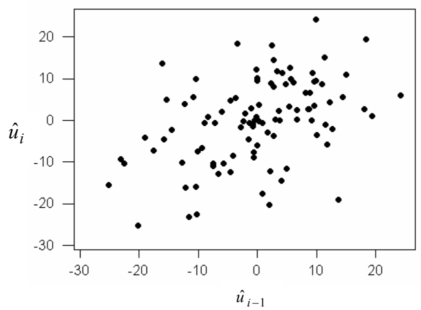
Plot (b)



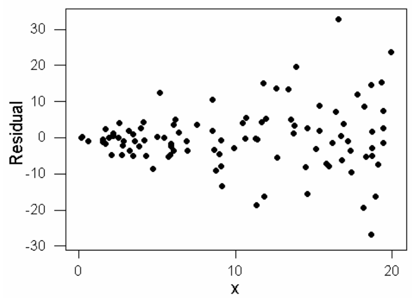
Plot (c)



Plot (d)



Plot (e)



1. Which **one** of the plots does **not** indicate problems with the assumptions underlying the linear regression model?

(1) (a)
(2) (b)
(3) (c)
(4) (d)
(5) (e)

2. Which **one** of the plots indicates that the variability of the error term is **not** independent of x?

(1) (a)
(2) (b)
(3) (c)
(4) (d)
(5) (e)

3. Which **one** of the following statements regarding the sample correlation coefficient, r, is **false**?

(1) The value of r is an indication of the strength of linear association between the two variables.
(2) In the interpretation of r, one variable is always treated as the response variable and the other as the explanatory variable.
(3) A value of r near 1 does not necessarily mean there is a causal relationship between the two variables.
(4) The value of r must be between -1 and 1 inclusive.
(5) The value of r may be near 0 when there is a non-linear relationship between the two variables.

4. Which **one** of the following statements is **not** an assumption of the linear regression model?

(1) The relationship between the X variable and the Y variable is linear.
(2) All random errors are independent.
(3) The X-values are Normally distributed.
(4) The standard deviation of the random errors does not depend on the X-values.
(5) For any X-value, the random errors are Normally distributed with a mean of 0.

5. Which **one** of the following statements regarding linear regression analysis is **false**?

(1) The two main components of a regression relationship are ‘trend’ and ‘scatter’.
(2) Using regression techniques, we can never determine whether a causal relationship exists between two variables.
(3) Outliers in the values of the explanatory variable can have a big influence on the fitted regression line.
(4) Lines fitted to data using the least-squares method do not allow us to reliably predict the behaviour of Y outside the range of x-values for which we have collected data.
(5) The least-squares regression technique minimises the sum of the squared prediction errors.