SECTION 4: SIMPLE LINEAR REGRESSION AND SIMPLE LINEAR CORRELATION

Linear Regression and Linear Correlation

- analyze the **relationship** between **quantitative variables**
- **"Simple"** means **only two variables** (x and y) are involved
- In the next section, we will discuss situations where there are more than two variables (multiple linear regression)
- There are other types of regression that are not linear (e.g., curvilinear), but the analysis of these are not dealt with in this course

Scatterplot or scatter diagram

- Illustrates the relationship between two quantitative variables, by placing points on the graph that correspond to the values of both variables **(x, y)** at the same time
- If a the data points in a scatterplot fall (roughly) in a straight line, this indicates a probably linear relationship between the two variables

Simple Linear Regression

- Used to analyze the relationship between two quantitative variables when one variable responds to the other
- **Explanatory variable** (or **predictor variable**) (plotted on ***x***-axis)
  = the variable that may affect the other variable or that can be used to make predictions about the other variable
- **Response variable** (plotted on ***y***-axis)
  = the variable that reacts to or is affected by the explanatory variable. It responds to changes in the predictor variable

Simple Linear Correlation

- Used to analyze the relationship between two quantitative variables when a change in one variable appears to be related to (or associated with) a change in the other, but one is not necessarily responding to the other
- So, the two variables are co-related
- Either variable may be plotted as ***x*** or ***y***
- Therefore, correlation can be applied in a wider variety of situations even when the variables cannot be identified as explanatory and response variables

Notation and Formulas for Quantities Used in Regression and Correlation

Quantity	Defining formula	Computing formula
Sum of Squares of (the deviations in) ***x***	$S_{xx} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$	$S_{xx} = \sum {x_i}^2 - \frac{\left(\sum x_i\right)^2}{n}$
Sum of Squares of (the deviations in) ***y***	$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$S_{yy} = \sum {y_i}^2 - \frac{\left(\sum y_i\right)^2}{n}$
Sum of Products of (the deviations in) ***x*** and ***y***	$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$	$S_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$

4.1 The Linear Regression Model

- Usually, data obtained from a sample of a population do not fall exactly along a straight line
- **Linear regression line** – the "best fit" line that passes through the points and is calculated using the **"least squares criterion"**

Simple Linear Regression Model

Model for the population regression line:

$$\mu(Y \mid X) = \beta_0 + \beta_1 X$$

$[\mu(Y \mid X)$ means "predicted mean of Y at a given $X]$

Where Y is the response variable

X is the explanatory variable

Parameters: β_0 is the y-intercept

β_1 is the slope (change in Y over change in X)

Estimated Model or sample regression line (based on a set of ***n*** data points):

$$\hat{y} = \hat{\mu}(Y \mid X) = \hat{\beta}_0 + \hat{\beta}_1 x$$ ['^' denotes an estimate]

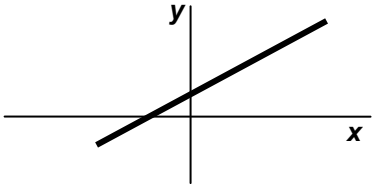
Where $\hat{y}$ is used to denote the y-value predicted by a regression equation

$\hat{\beta}_0$ and $\hat{\beta}_1$ are least squares estimates

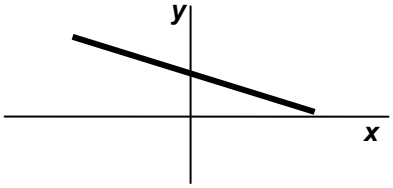
Slope = $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$

y-intercept = $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \dfrac{\sum y}{n} - \hat{\beta}_1 \dfrac{\sum x}{n}$

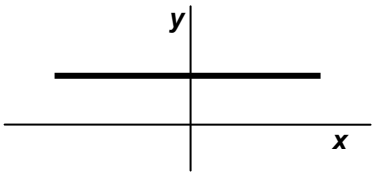
Different types of slopes



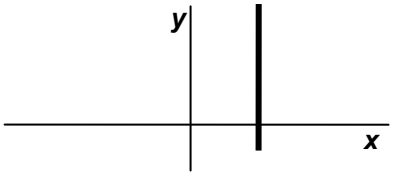
Positive slope ($\beta_1 > 0$)



Negative slope ($\beta_1 < 0$)

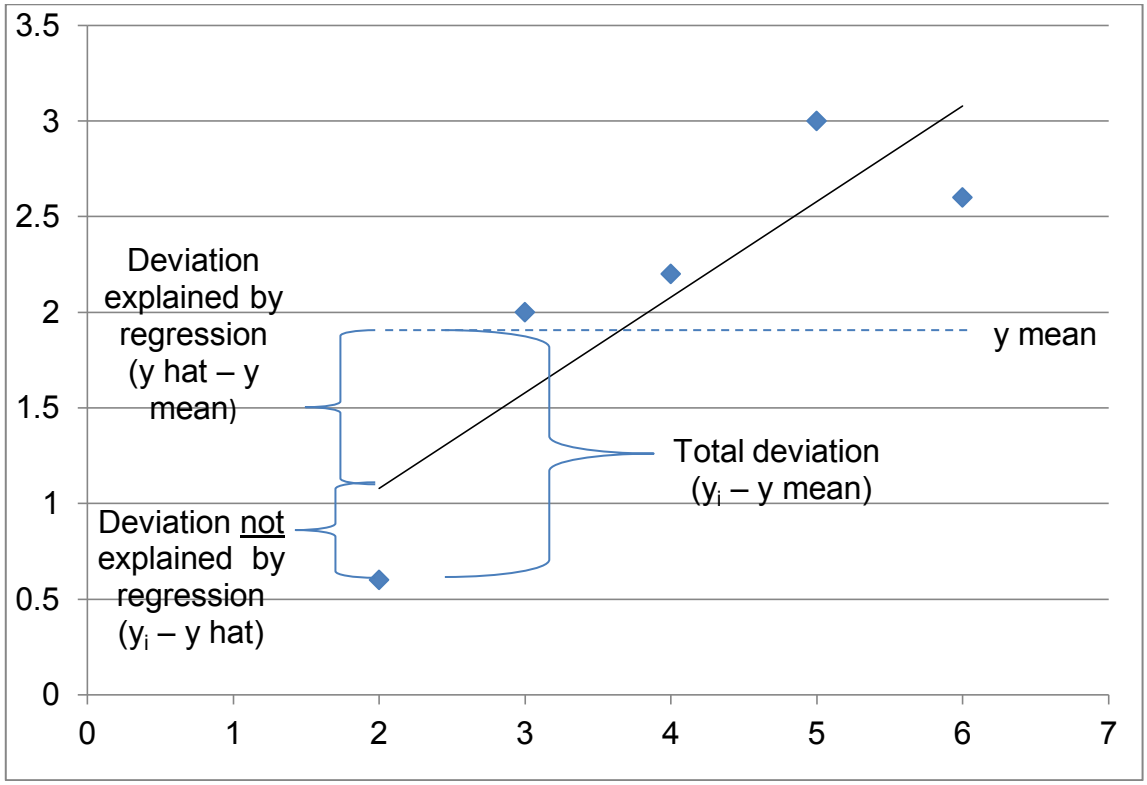


No slope ($\beta_1 = 0$)



Infinite slope ($\beta_1 = \infty$)

Three Sources of Variation (Deviation) in Regression



Three Sums of Squares in Regression
[Representing the three sources of variation or deviation]

Total Sum of Squares (SS_{TOTAL} = Syy)
= total variation in the observed values of the response variable

$$SS_{TOTAL} = S_{yy} = \sum (y_i - \bar{y})^2$$

Regression Sum of Squares (SS_{REGR})
= variation in the observed values of the response variable explained by regression model

$$SS_{REGR} = \sum (\hat{y}_i - \bar{y})^2 = \frac{(S_{xy})^2}{S_{xx}} = \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum (x_i - \bar{x})^2}$$

Error (Residual) Sum of Squares (SS_{ERROR})
= variation in the observed values of the response variable that is <u>not</u> explained by the regression model

$$SS_{ERROR} = SS_{RES} = \sum (y_i - \hat{y}_i)^2$$

$$SS_{ERROR} = SS_{TOTAL} - SS_{REGR}$$

- It is SS_{ERROR} that regression tries to minimize in order to obtain the "best fit" line

Regression identity: $SS_{TOTAL} = SS_{REGR} + SS_{ERROR}$

Analysis of Residuals

Residual = error (e) = vertical distance from the regression line to a data point (may be + or –)

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

Residual Sum of Squares = Error Sum of Squares

= the variation in the observed values of the response variable that is <u>not</u> explained by the regression

$$SS_{\text{RES}} = SS_{Error} = \sum (y_i - \hat{y}_i)^2$$

Least-squares criterion

- Tries to minimize the Residual or Error Sum of Squares (SSE) in order to get the "best fit" line
- Thus, regression tries to minimize the errors due to deviations <u>not</u> explained by the regression equation

Residual Plots and Residual Analysis

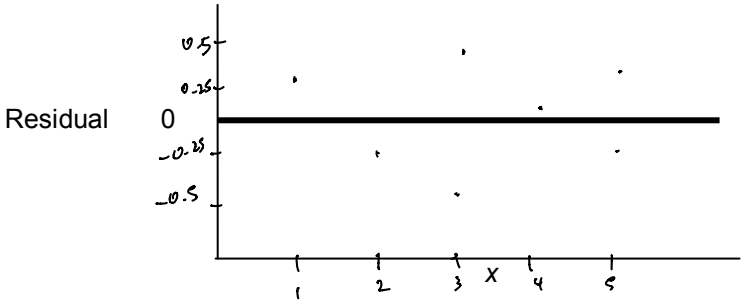
- Can be used for checking whether a certain data set fits the assumptions of regression analysis
- Each data point is plotted as the residual (error) against the corresponding the x-value
- If the assumptions for regression inferences are met, the plot of the residuals against the values of the predictor variable should:
  <u>fall roughly in a horizontal band centered and symmetric about the x-axis</u>

Example of Residual Analysis (heights of the trees of different ages)

Using the linear regression equation ($\hat{y} = -0.08087 + 0.66809x$), we can determine $\hat{y}$ and the residual for every data point and find the Residual Sum of Squares (SSE)

Age (years) x	**Height (m) y**	$\hat{y}$	**Residual (error)** $(y_i - \hat{y}_i)$	**(error)2** $(y_i - \hat{y}_i)^2$
1	0.9	0.587	0.313	0.0980
2	1.0	1.255	-0.255	0.0652
3	1.4	1.923	-0.523	0.2735
3	2.2	1.923	0.277	0.0767
4	2.6	2.591	0.009	0.0001
5	3.0	3.260	-0.260	0.0676
5	3.7	3.260	0.440	0.1936
			Sum = 0	$SS_{\text{RES}} = \sum (y_i - \hat{y}_i)^2 = 0.7747$

Sketch of a residual plot for the heights of trees [plot Error against x]



Standard error of the model (= Common standard deviation of the model)
(= Standard deviation of the residuals)

- Quantifies the amount of scatter around the regression line
- Given by: $s_e = \hat{\sigma} = \sqrt{\frac{\sum e^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SS_{ERROR}}{n-2}} = \sqrt{MS_{ERROR}}$

Prediction: Interpolation and Extrapolation

Interpolation = using the regression equation to make predictions about the response variable, within the range of the observed values of ***x***

- can be reasonably accurate

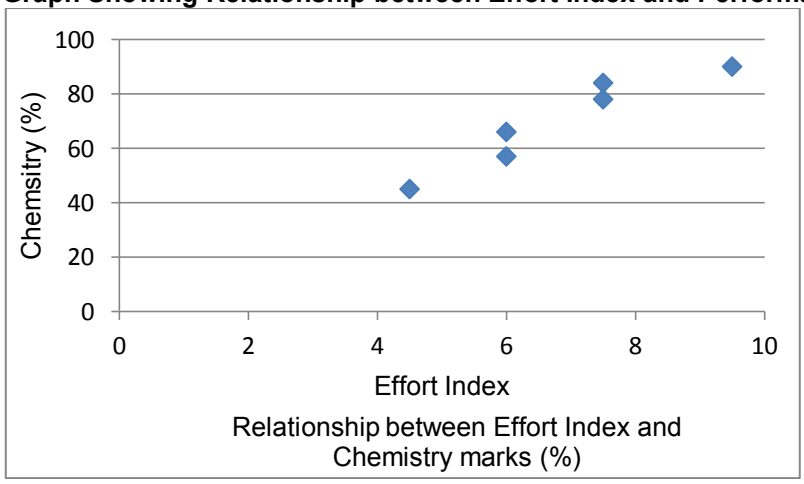
Extrapolation = using the regression equation to make predictions about the response variable, outside the range of the observed values of ***x***

- can sometimes lead to seriously incorrect predictions because the relationship may not hold beyond the observed range
- to avoid errors due to extrapolation, researchers sometimes define the range of the observed values along with a regression equation
  - e.g., $\hat{y} = 24.7 + 3.4x$, 6 ≤ ***x*** ≤ 45

Example of Calculation of the Regression Line
Effort Index (on a scale of 0 – 10) was calculated based on a combination of factors such as attendance in classes, hours per week spent studying and completion of assignments on time. The table below shows the Effort Index and Chemistry and Biology marks (in %) of a random sample of 6 students.

	Halima	John	Jing	Jasmin	Vanessa	Harry
Effort Index	9.5	4.5	7.5	6	7.5	6
Chemistry (%)	90	45	84	66	78	57
Biology (%)	90	56	96	65	81	74

Graph Showing Relationship between Effort Index and Performance in Chemistry



This above graph shows that the relationship is appropriate for linear regression analysis because: (1) linear relationship (2) no significant outliers

Calculation of the regression line:

Table showing calculation of the deviations in ***x*** and ***y*** and the product of the deviations

	Effort index ***x***	Chemistry (%) ***y***	Deviations in ***x*** $(x_i - \bar{x})$	Deviations in ***y*** $(y_i - \bar{y})$	Product of deviations in ***x*** and ***y*** $(x_i - \bar{x})(y_i - \bar{y})$
Halima	9.5	90	2.6667	20	53.334
John	4.5	45	-2.3333	-25	58.3325
Jing	7.5	84	0.6667	14	9.3338
Jasmin	6.0	66	-0.8333	-4	3.3332
Vanessa	7.5	78	0.6667	8	5.3336
Harry	6.0	57	-0.8333	-13	10.8329
Totals	**41**	**420**	**0**	**0**	$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = 140.5$
Mean	**6.833333**	**70**			
SD	**1.722401**	**17.146428**			

Squared deviations in ***x*** $(x_i - \bar{x})^2$	Squared deviations in ***y*** $(y_i - \bar{y})^2$
7.1113	400
5.4443	625
0.4445	196
0.6944	16
0.4445	64
0.6944	169
$S_{xx} = \sum(x_i - \bar{x})^2 = 14.8333$	$S_{yy} = \sum(y_i - \bar{y})^2 = 1470$

Calculate the linear regression equation describing the relationship between Effort Index and performance (%) in Chemistry. Also, interpret the meaning of the slope.

>>>>>>>>>>

Slope $= \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

$= \frac{140.5}{14.8333} = 9.4719$ % patients effort index

y-intercept $= \hat{\beta} = \bar{y} - \hat{\beta}_1 \bar{x}$

$= 70.0 - 9.471976833$

$= 5.27528$

Regression Eq. $\hat{y} = 5.275 + 9.472x$

where $4.5 \leq x \leq 9.5$

>>>>>>>>>>

4.2 Linear Correlation (r) and Coefficient of Determination (R^2)

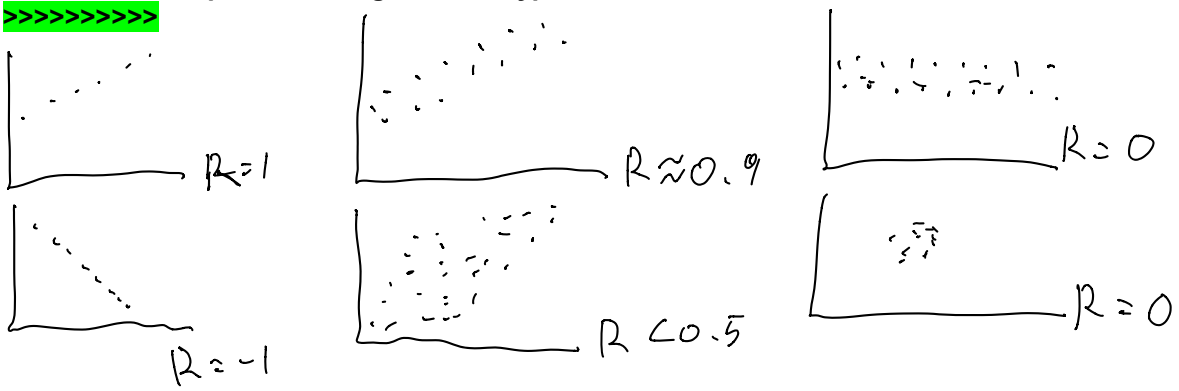
- The most common measure of correlation is the Pearson product-moment correlation coefficient.

Three Aspects of a Relationship between Variables (Correlation and Regression)

- **Direction:**
  - Positive correlation: if ***r*** has + sign, then the slope must have + sign
  - Negative correlation: if ***r*** has — sign, then the slope must have — sign
- **Form:**
  - May be a straight line relationship (linear) or curved (Here we only deal with linear)
- **Strength:** The magnitude of ***r*** indicates the strength of the linear relationship between the two variables.
  - ***r*** close to -1 or 1 indicates a strong linear relationship and the regression equation is very useful for making predictions
  - ***r*** close to 0 indicates no relationship or a weak linear relation and the regression equation is either useless or not very useful for making predictions

Sketch 6 scatterplots showing different types of linear correlation

>>>>>>>>>>



>>>>>>>>>>

Warnings on the Use of the Linear Correlation Coefficient (and linear regression analysis)

- The linear correlation coefficient should only be used when a scatterplot indicates that the data points are scattered roughly about a straight line

Outliers, Leverage and Influential Points

Outlier

- Any data point that does not follow the general pattern of the rest of the data or that stands away from other data points
- May be either due to having a large residual (on the y-scale) or having high leverage (being far away from other points on the x-scale)

Leverage

- Data points whose x-values are far from the mean of x have leverage
- Data points having leverage can have a great affect on the linear regression line
- They can completely change the slope and y-intercept

Influential Points

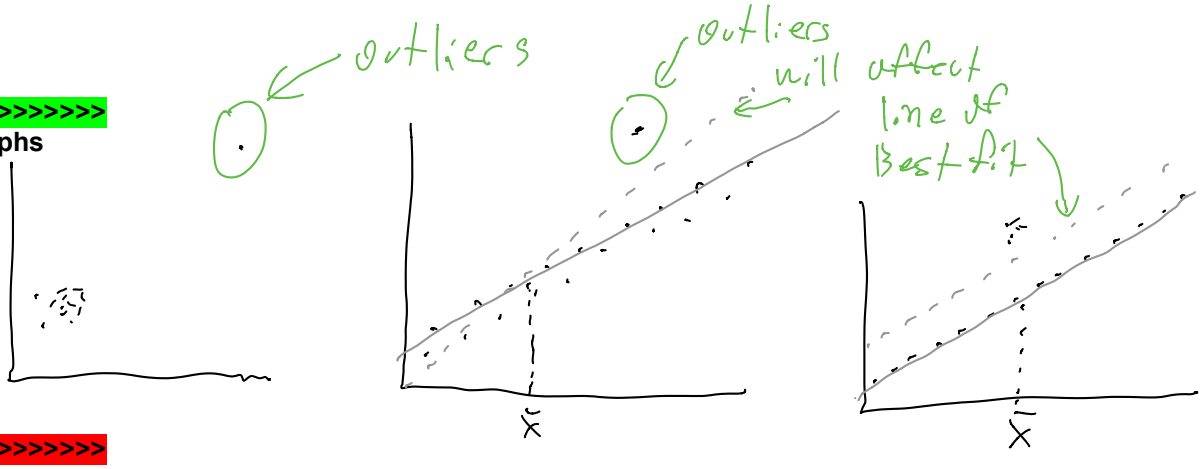
- A data point which, if omitted, results in a very different regression model

Serious outliers and influential observations

- They can make a weak correlation appear to be a strong correlation
- They can change the slope considerably
- They can even make a positive correlation to be calculate as a negative correlation

>>>>>>>>>>

Graphs



>>>>>>>>>>

Correlation versus Prediction/Response

- Correlation between variables does not necessarily mean that one variable affects or can be used to predict the other variable

Calculation of the Correlation Coefficient

The Linear Correlation Coefficient, *r*

For a set of ***n*** data points,

$$r = \frac{\text{covariance of } x \text{ and } y}{(\text{standard deviation of } x) \times (\text{standard deviation of } y)}$$

Where: Covariance of ***x*** and ***y*** = $s_{xy} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$

Thus, $$r = \frac{s_{xy}}{s_x \times s_y} = \frac{\dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}}{\sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n-1}} \times \sqrt{\dfrac{\sum (y_i - \bar{y})^2}{n-1}}}$$

Where s_x and s_y are the sample standard deviations of the x-values and y-values, respectively.

Also, $$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum (x_i - \bar{x})^2\right]\left[\sum (y_i - \bar{y})^2\right]}}$$

Also, $r = \sqrt{\dfrac{SS_{REGR}}{SS_{TOTAL}}}$ (then **add correct sign**, + or –)

Always: $-1 \le r \le 1$

r has no units because they cancel out during calculations

Coefficient of Determination (R^2) and Adjusted R^2

- Adjusted R^2 takes into account (adjusts for) samples size, though it makes little difference in SLR

Coefficient of determination (R^2) = [correlation coefficient]^2

R^2 = the fraction or percentage of variation in the observed values of the response variable that is accounted for by the regression analysis

$$R^2 = r^2 = \frac{\text{Explained variability}}{\text{Total variability}}$$

$$R^2 = \frac{SS_{REGR}}{SS_{TOTAL}} = 1 - \frac{SS_{Error}}{SS_{TOTAL}} = \frac{SS_{TOTAL} - SS_{Error}}{SS_{TOTAL}}$$

Adjusted Coefficient of Determination

$$R^2_{adj} = 1 - \frac{MSE}{MST}$$

Always: $0 \le R^2 \le 1$
OR: $0\% \le R^2 \le 100\%$

This implies that $1 - R^2$ of the variation in the observed values of the response variable are accounted for by other factors, not the explanatory variable used in the regression analysis

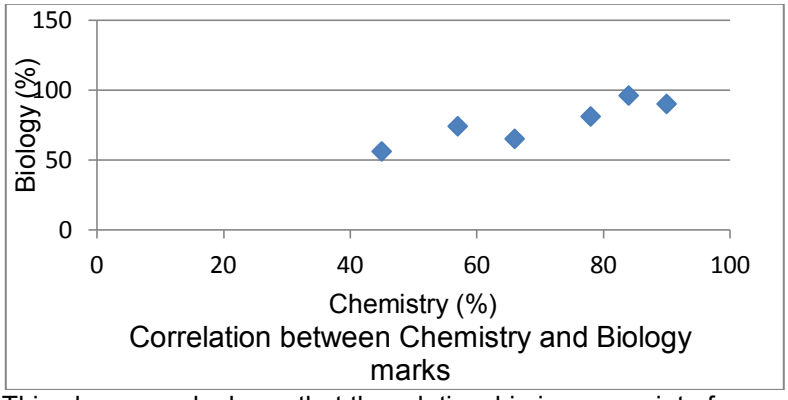
- If $\mathbf{r^2}$ is close to 0, this suggests that the regression equation is not very useful for accounting for the response variable or for making predictions
- If $\mathbf{r^2}$ is close to 1, this suggests that the regression equation is very useful for accounting for the response variable or for making predictions

Relationship Between the Correlation Coefficient and the Coefficient of Determination

- Both are measures for indicating the strength of a linear relationship or the usefulness of the regression equation for making predictions.
- If you have already calculated the coefficient of determination and we know the direction of the relationship (positive or negative), we can calculate the correlation coefficient as:

$$r = \sqrt{R^2}$$ [And then add the appropriate sign, **+** or **-**]

Example: Correlation between Chemistry and Biology marks
Note: Since this is correlation, either variable could be considered as ***x*** or ***y***



This above graph shows that the relationship is appropriate for analysis with linear correlation because: (1) linear relationship (2) no significant outliers

Table showing calculation of the deviations in ***x*** and ***y*** and the product of the deviations

	Chem (%) ***x***	Biol (%) ***y***	Deviations in ***x*** $(x_i - \bar{x})$	Deviations in ***y*** $(y_i - \bar{y})$	Product of deviations in ***x*** and ***y*** $(x_i - \bar{x})(y_i - \bar{y})$
Halima	90	90	20	13	20 x 13 = 260
John	45	56	-25	-21	525
Jing	84	96	14	19	266
Jasmin	66	65	-4	-12	48
Vanessa	78	81	8	4	32
Harry	57	74	-13	-3	39
Totals	**420**	**462**	**0**	**0**	$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = 1{,}170$
Mean	**70**	**77**			
SD	**17.14643**	**15.09967**			

Squared deviations in ***x*** $(x_i - \bar{x})^2$	Squared deviations in ***y*** $(y_i - \bar{y})^2$
400	169
625	441
196	361
16	144
64	16
169	9
$S_{xx} = \sum(x_i - \bar{x})^2 = 1{,}470$	$S_{yy} = \sum(y_i - \bar{y})^2 = 1{,}150$

Covariance of x and y is:

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)} = \frac{1{,}170}{6-1} = 234$$

Correlation coefficient (***r***) is:

$$r = \frac{s_{xy}}{s_x \times s_y} = \frac{234}{17.14643 \times 15.09967} = 0.9038$$

Correlation coefficient can also be calculated as:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum(x_i - \bar{x})^2\right]\left[\sum(y_i - \bar{y})^2\right]}} = \frac{1{,}170}{\sqrt{(1{,}470)(1{,}150)}} = 0.900$$

Does the strong positive correlation indicate that the performance of the students in Chemistry <u>explains</u> their performance in Biology?

4.3 Assumptions for Regression Inferences and Analysis of Residuals

Assumptions (Conditions) for Regression Inferences

1. **Population regression line (<u>linearity</u>):** The relationship between the two variables must be approximately linear. In other words, there are constants β_0 and β_1 such that, for each value x of the predictor variable, the conditional mean of the response variable is $\beta_0 + \beta_1 x$.

2. **Equal standard deviations (homoscedasticity):** The standard deviations of ***y***-values must be approximately the same for all values of ***x***

3. **Normal populations:** For each value of ***x***, the corresponding ***y***-values must be normally distributed)

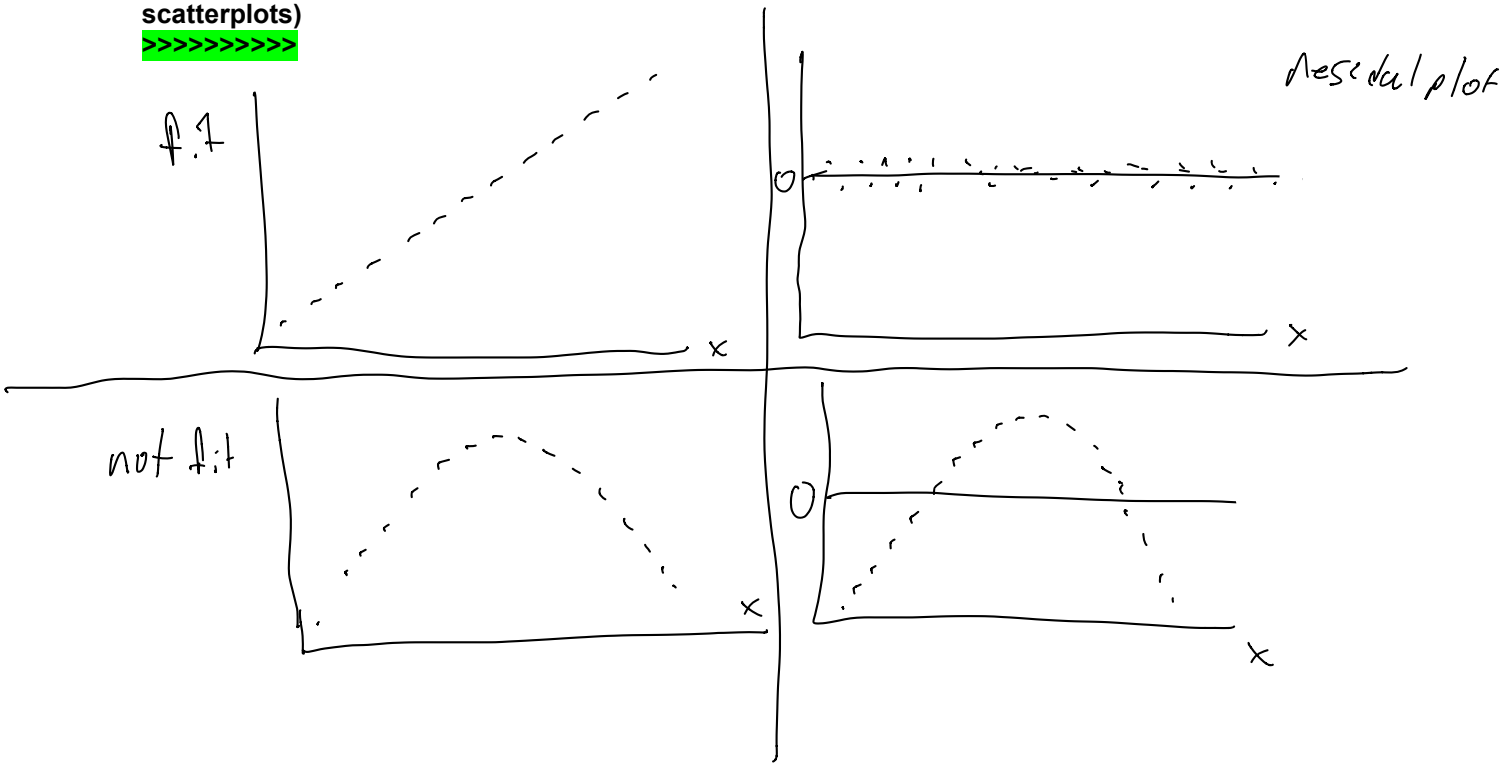
4. **No Serious Outliers:** Significant outliers can drastically change the regression model (just as for correlation)

5. **Independent observations:** The observations of the response variable are independent of one another. This implies that the observations of the predictor variable not need to be independent.

Note: All assumptions (except independence) can be <u>checked</u> by examining a scatterplot and/or residual plot. The assumption for normality is best checked with a NPP.

Illustrations for the first 2 assumptions of regression inferences each with xy scatterplot and a residual plot, for data that fit each assumption and data that do not (for normality, just use xy scatterplots)

>>>>>>>>>>



>>>>>>>>>

4.4 Testing the Significance of the Model using the Regression ANOVA Test

- Regression ANOVA tests the overall significance of the regression model
- In SLR, regression ANOVA can also be used to test for the significance of the slope, since there is only one slope

Mean Squares and F-Statistic in Simple Linear Regression ANOVA

Regression mean square (MS_{Regr}) = regression sum of squares divided by regression df

$$MS_{REGR} = SS_{REGR} / 1$$

Error mean square (MS_{ERROR}) = error sum of squares divided by error df

$$MS_{ERROR} = SS_{Error} / (n-2)$$

F-Statistic (F) $$F = \frac{SS_{REGR}/1}{SS_{Error}/(n-2)} = \frac{MS_{REGR}}{MS_{ERROR}}$$ Where n = number of x,y observations

ANOVA Test for Significance of the Model and the Slope of a Population Regression line

Purpose: To determine whether there is a relationship between two quantitative variables OR to decide whether the slope of the line is significantly different from zero.
Assumptions: see textbox on assumptions above

Step 1: Select appropriate test by checking purpose and assumptions

Step 2: H_0: $\beta_1 = 0$ (There is no relationship between the two variables)
H_a: $\beta_1 \neq 0$ (There is a relationship between the two variables)

Step 3: Calculate the three sums of squares (see page 3) and construct an ANOVA table
ANOVA Table for Simple Linear Regression

Source of variation	SS	df	MS = SS/df	F-statistic
Regression	SS_{REGR}	2 - 1	$MS_{REGR} = SS_{REGR}/1$	$F = MS_{REGR} / MS_{ERROR}$
Error	SS_{Error}	n – 2	$\hat{\sigma}^2 = MS_{ERROR} = SS_{Error}/(n-2)$	
Total	SS_{TOTAL}	n - 1		

$$F = \frac{SS_{REGR}/1}{SS_{Error}/(n-2)} = \frac{MS_{REGR}}{MS_{ERROR}}$$

Step 4: Decide to reject or not reject Ho
***df* =** (numerator degrees of freedom, denominator degrees of freedom)
$df = (1, n-2)$ (Where ***n*** = no. of ***xy*** observations)
If P-value ≤ α, reject H_0 (otherwise, do not reject H_0)
Step 5: Conclusion in terms of the research problem

Note:

1. $MS_{ERROR} = \hat{\sigma}^2 = (\text{standard error of the model})^2 = \text{variance of the model}$
2. $t^2 = F$
3. **The P-value will be the same for both the F-test and the t-test (two-tailed)**

Note: In general, the regression ***df*** is the number of parameters being estimated minus 1. Since here we are estimating two parameters, the y-intercept and slope (β_0 and β_1), the regression ***df*** = 2 – 1 = 1

Computer Output for the Example on Effort Index and Chemistry Performance

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.951477
R Square	0.905308
Adjusted R Square	0.881636
Standard Error	5.899081
Observations	6

model squared

$\sqrt{F} = t$

ANOVA table for Simple Linear Regression					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1330.8034	1330.8034	38.2424	0.003474604
Residual	4	139.19663	34.799157		
Total	5	1470			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.275281	10.739903	0.4911851	0.649026	-24.54347	35.09403
Effort index	9.47191	1.5316692	6.1840442	0.003475	5.21931	13.72451

Example: At the 5% significance level, test for a relationship between Effort Index and performance in Chemistry. [In other words, test whether the slope of the regression line is significant.] The regression line has been shown to be linear. Assume that all other assumptions are met.

>>>>>>>>>>

$H_0: \beta_1 = 0$ (there is no relationship)

$H_A: \beta_1 \neq 0$ (there is a relationship)

$SS_T = SS_{YY} = 1,470$

$SS_R = \frac{(SS_{XY})^2}{SS_{XX}} = \frac{(140.5)^2}{14.8333} = 1330.8064$

$SS_E = SS_T - SS_R$

$= 1470 - 1330.8064$

$= 139.1936$

$df(1, 4)$

$0.001 < p < 0.005$

$F = \frac{SS_R/1}{SS_E/n-2} = \frac{MS_R}{MS_E} = \frac{1330.8064}{34.7984} = 38.242$

>>>>>>>>>>

Comparison between Regression t-test and ANOVA F-test (Applies <u>ONLY</u> to SLR)

F-statistic = (t-statistic)^2 = $(6.1840442)^2$ = 38.24

Two-tailed P-values are always equal for the F-test and t-test

For both the t-test and the F-test, the exact P-value = 0.003475

4.5 Inferences for the Slope of the Population Regression Line

- In SLR, either a Regression t-test OR Regression ANOVA can be used to test the slope
- However, the Regression t-test is more flexible because it is suitable for doing two-tailed tests or one-tailed tests

Regression t-Test for Significance of the Slope of a Population Regression line

Purpose: To determine whether there is a relationship between two quantitative variables OR to decide whether the slope of the line is significantly different from zero.

Assumptions: see textbox on assumptions above

Step 1: Select appropriate test by checking purpose and assumptions

Step 2:

Null hypothesis

H_0: $\beta_1 = 0$ (T**here is no relationship between the two variables**)

Alternative hypotheses

Two-tailed test: H_a: $\beta_1 \neq 0$ **(There is a relationship between the two variables**)

Left-tailed test: H_a: $\beta_1 < 0$ **(There is a negative relationship between the two variables**)

Right-tailed test: H_a: $\beta_1 > 0$ **(There is a positive relationship between the two variables**)

Step 3: Compute the calculated value of the test statistic

$$SS_{TOTAL} = S_{yy} = \sum (y_i - \bar{y})^2$$

$$SS_{REGR} = \sum (\hat{y}_i - \bar{y})^2 = \frac{(S_{xy})^2}{S_{xx}} = \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum (x_i - \bar{x})^2}$$

$$SS_{ERROR} = SS_{TOTAL} - SS_{REGR}$$

Standard error of the model (residual standard deviation) (s_e):

$$\hat{\sigma} = \sqrt{\frac{SS_{ERROR}}{n-2}} = \sqrt{MS_{ERROR}}$$

Standard Error of the Estimate of the Slope:

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

The Regression t-statistic:

$$t = \frac{\hat{\beta}_1}{\hat{\sigma} / \sqrt{S_{xx}}} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

Step 4: Decide to reject or not reject Ho and state the strength of the evidence against Ho.

$df = n - 2$ (Where ***n*** = no. of ***xy*** observations)

If P-value ≤ α, reject H_0; otherwise, do not reject H_0

Step 5: Conclusion in terms of the research problem

Testing the Significance of the Slope using the Regression t-test
At the 5% significance level, test for a relationship between Effort Index and performance in Chemistry. [In other words, test whether the slope of the regression line is significant.] The regression line has been shown to be linear. Assume that all other assumptions are met.

Step 1: Regression t-test is selected because the purpose is to test if the slope is significantly different from 0.

Step 2: H_0: $\beta_1 = 0$ (There is no relationship between performance in Chemistry and Effort index.)
H_a: $\beta_1 \neq 0$ (There is a relationship between performance in Chemistry and Effort index.)

Step 3: Compute the three sums of squares

$$SS_{TOTAL} = S_{yy} = \sum (y_i - \bar{y})^2 = 1,470$$

$$SS_{REGR} = \frac{(S_{xy})^2}{S_{xx}} = \frac{(140.5)^2}{14.8333} = 1330.8064$$

$$SS_{ERROR} = SS_{TOTAL} - SS_{REGR} = 1470 - 1330.8064 = 139.1936$$

>>>>>>>>>>

SE of model

$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{139.1936}{6-2}} = 5.8990$

SE of slope

$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \frac{5.8990}{14.8333} = 1.5316$

$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{9.4719}{1.5316} = 6.184$

$df = 6-2 = 4$

$(0.001 < p < 0.0025) \times 2 = (0.002 < p < 0.005)$

There is a very strong evidence against H_0

Since p-value $\leq \alpha$, reject H_0.

>>>>>>>>>>

4.6 Confidence Interval for the Slope of the Population Regression Line

- The confidence interval for the slope follows the same general formula as for other t-procedures:

Confidence Interval: $Estimate \pm Critical\ Value \times SE(Estimate)$

Confidence Interval for the Slope of the Population Regression Line
[Regression t-Interval Procedure]

Purpose: To find a confidence interval for the slope, $\boldsymbol{\beta_1}$, of the population regression line

Assumptions: The four assumptions for regression inferences

Step 1: For a given confidence level (1 – α), use the t-table to find $t_{\alpha/2}$ in the row for the appropriate ***df,*** where ***df = n – 2***.

Step 2: The endpoints of the confidence interval for $\boldsymbol{\beta_1}$ are:

$$\hat{\beta}_1 \pm t_{\alpha/2} \times \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \quad \Rightarrow \quad \hat{\beta}_1 \pm t_{\alpha/2} \times SE(\hat{\beta}_1)$$

Step 3: Interpret the confidence interval.

Example of Finding a Confidence Interval For the Slope
Calculate a 95% confidence interval for the slope of the regression line for the relationship between Effort Index and Chemistry Performance.

[Previously calculated: $Slope(\hat{\beta}_1) = 9.4719, S_{xx} = 14.8333, \hat{\sigma} = 5.8990$ (Standard error of the model)]

>>>>>>>>>>

For a 95% C.I. $\alpha = 0.05$

① $df = n-2 = 6-2 = 4$, $t_{\alpha/2} = t_{0.025} = 2.776$

$\hat{\beta}_1 \pm t_{\alpha/2} \times SE(\hat{\beta}_1)$

$9.4719 \pm 2.776 \times 1.5316$

$= (5.22, 13.72)$

∴ you get the idea for the conclusion.

>>>>>>>>>>

See Computer Output on page 13

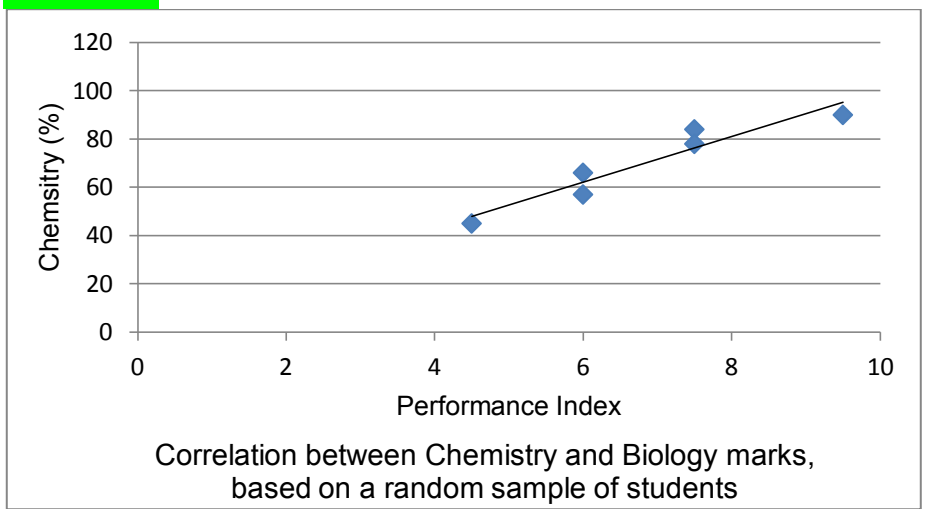
$$\hat{\sigma} = \sqrt{\frac{SS_{ERROR}}{n-2}} = \sqrt{\frac{139.19663}{6-2}} = \sqrt{MS_{ERROR}} = 5.8991$$

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{9.47191}{1.53167} = 6.184$$

Confidence interval (β_1) = (5.21931, 13.7245)

Sketch the confidence interval for the slope:

>>>>>>>>>>



Correlation between Chemistry and Biology marks, based on a random sample of students

>>>>>>>>>>

4.7 Confidence Intervals for Estimation of Mean Response and Predicted Response

Confidence Interval for Mean Response (or Conditional Mean)

- Used to estimate the confidence interval for the mean response of the response variable for a given value of the explanatory variable
- The predicted value of the response variable (y) for a given value of the predictor variable (x) can be determined by simply substituting that value of the given x into the regression equation.

Confidence Interval for the Mean response of *y* for a Given *x*
[Conditional Mean t-interval Procedure]

Purpose: To find a confidence interval for the mean response of the response variable for any given value (***x*_{*p*}**) of the predictor or explanatory variable

Assumptions: The four assumptions for regression inferences.

Step 1: For a given confidence level (1 – α), use the t-table to find $t_{\alpha/2}$ at ***df = n – 2***

Step 2: Compute the point estimate (the predicted value of response variable for a given value of the predictor variable) by using the linear regression equation:

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

Then, calculate the endpoints of the confidence interval by:

$$\hat{y}_p \pm t_{\alpha/2,n-2} \times \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}} \qquad \text{Where } S_{xx} = (n-1)s_x^2$$

Step 3: Interpret the confidence interval.

Prediction Interval

- A confidence interval for predicting all single observations of the response variable at a given value of the explanatory or predictor variable

Prediction Interval (OR Confidence Interval for the prediction of <u>all single observations</u> of the response of *y* for a Given *x*]

Purpose: To find a prediction interval for <u>all single observation responses</u> of the response variable for any given value (x_p) of the predictor or response variable

Assumptions: The four assumptions for regression inferences.

Step 1: For a given confidence level (1 – α), use the t-table to find $t_{\alpha/2}$ at ***df = n – 2***.

Step 2: Compute the point estimate (the predicted value of response variable for a given value of the predictor variable) by using the linear regression equation:

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

Then, calculate the endpoints of the confidence interval by:

$$\hat{y}_p \pm t_{\alpha/2,n-2} \times \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{S_{xx}}}$$

Note also: $S_{xx} = (n-1)s_x^2$

Step 4: Interpret the confidence interval in terms of the research problem.

Example of Finding the Mean Response and Predicted Response

Age (years) x	Height (m) y
1	0.9
2	1.0
3	1.4
3	2.2
4	2.6
5	3.0
5	3.7

Given: $\hat{y} = -0.08087 + 0.66809x$

$S_{xx} = 13.42857$

$\hat{\sigma} = 0.39367$

$\bar{x} = 3.28571$

$\sum x = 23$

Find a 95% confidence interval for the mean height of all 3-year-old trees.
Also, find a 95% prediction interval for the height of a 3-year-old tree.

>>>>>>>>>>

for a 95% C.I, $\alpha = 0.05$ $df = n-2 = 7-2 = 5$ $t_{\alpha/2} = t_{0.025} = 2.571$

$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p = 1.923$

$\hat{y}_p \pm t_{\alpha/2,n-2} \times \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{S_{xx}}}$

$1.923 \pm 2.571 \times 0.39367 \times \sqrt{1+\frac{1}{7}+\frac{(3-3.28571)^2}{13.42857}}$

(0.838, 3.008) m

$S_x = 1.496026$

$S_{xx} = (n-1)S_x^2$

$= (7-1)(1.496026)$

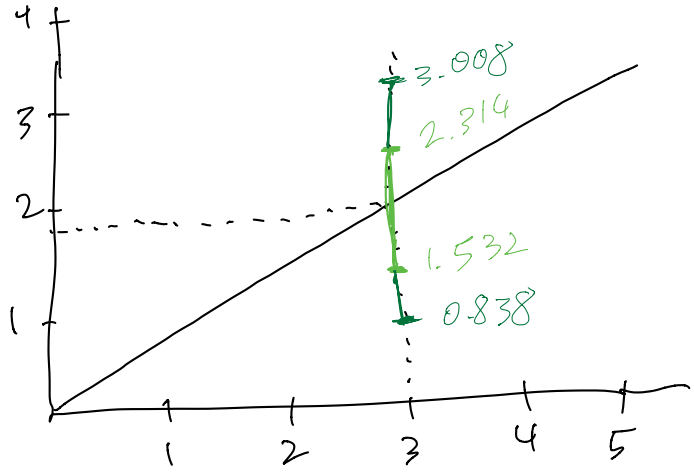
$= 13.42857$

1. we can be 95% confident that the mean height of 3-year old trees is somewhere between 1.532 and 2.314 m

2. any 3-year old tree be somewhere between 0.838 and 3.008 m

Comparison Of Confidence Intervals For the Mean Response and Predicted Response

Graph



>>>>>>>>>>

Note: The prediction interval is always wider than the confidence interval because:

- The estimate of a mean is always close to the population mean, whereas variation in all observed values is more disperse

4.8 Hypothesis Test for Linear Correlation

- A hypothesis test for the significance of the correlation between two quantitative variables
- This hypothesis test can be performed either:
  - When one variable can be identified as the explanatory variable (and regression can also be performed).
  - Or, when neither variable can be considered as the explanatory variable (and regression would not be performed)

Linear Correlation Hypothesis Test

Purpose: To perform a hypothesis test to decide whether two quantitative variables are significantly correlated.

Step 1: Select the appropriate test by checking purpose and assumptions

Step 2: State the null and alternative hypotheses

H_0: $\rho = 0$ (There is no correlation between the two variables.)

Just like regression, correlation can be a two-tailed, left-tailed or right-tailed test, so the alternative hypothesis may be:

Two-tailed: Ha: $\rho \neq 0$; Left-tailed: Ha: $\rho < 0$ (negative); Right-tailed: Ha: $\rho > 0$ (positive)

Note: Rho (ρ) is the Greek letter for population correlation coefficient

Step 3: Compute the correlation coefficient (***r***)

Using $r = \dfrac{s_{xy}}{s_x \times s_y}$ Where $s_{xy} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$

OR $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum (x_i - \bar{x})^2\right]\left[\sum (y_i - \bar{y})^2\right]}}$

OR $r = \sqrt{\dfrac{SS_{REGR}}{SS_{TOTAL}}}$ **(then add correct sign, + or −)**

Step 4: Decide to reject H_0 or not reject H_0 and state the strength of the evidence against H_0
df = n – 2 (where n = no. of ***xy*** observations)

Utilize the Table for the correlation coefficient, ***r***
If the P-value ≤ α, we reject H_0 (otherwise do not reject H_0)

Step 5: Interpretation (conclusion) in words in terms of the research problem being investigated.

Correlation p-value (two-tailed)
p-value of the t-test (two-tailed)
p-value of the F-test

Example of Linear Correlation Hypothesis Test
At the 5% significance level, test whether there was a correlation between performance in Chemistry and Effort index (based on a random sample of 6 students).

Recall: We previously calculated the following (page 5)

Sum of Products of deviations of ***x*** and ***y***: $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = 140.5$

Sum of Squares of deviations in ***x***: $S_{xx} = \sum (x_i - \bar{x})^2 = 14.8333$

Sum of Squares of deviations in ***y***: $S_{yy} = \sum (y_i - \bar{y})^2 = 1470$

>>>>>>>>>>

$H_0: \rho = 0$ i.e. no correlation

$H_A: \rho \neq 0$ i.e correlation

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{140.5}{\sqrt{14.833 \cdot 1470}} = 0.9515$$

$df = n - 2 = 6 - 2 = 4$

There is a strong evidence against H_0 and $\alpha < 0.05$

$\therefore$ reject

Note: The p-value for the correlation test is exactly the same as the p-value of the t-test for the slope and Regression ANOVA F-test.

>>>>>>>>>>

Comparative Examples Correlation Coefficients

1. An education researcher wanted to test (at α = 0.01) whether there is significant correlation between the amount of time per week that high school students spend watching TV and their academic performance. Upon analyzing data obtained from a random sample of 50 students, she found the correlation coefficient ***r*** = – 0.374. What P-value and conclusion did she obtain?

   (a) P < 0.001; significant correlation
   (b) 0.005 < P < 0.01; no significant negative correlation
   (c) P > 0.50; no significant correlation
   (d) 0.005 < P < 0.01; significant negative correlation
   (e) 0.02 < P > 0.01; significant correlation

2. The correlation coefficient between two variables was r = 0.903, sample size = 6. State the P-value and conclude whether the correlation is significant at the 1% significance level.

>>>>>>>>>>

$df = 6 - 2 = 4, \quad 0.01 < p < 0.02$

>>>>>>>>>>

Example on Biotechnology: Using the water fern Azolla to produce Hydrogen Fuel
[Example Combining All Concepts]

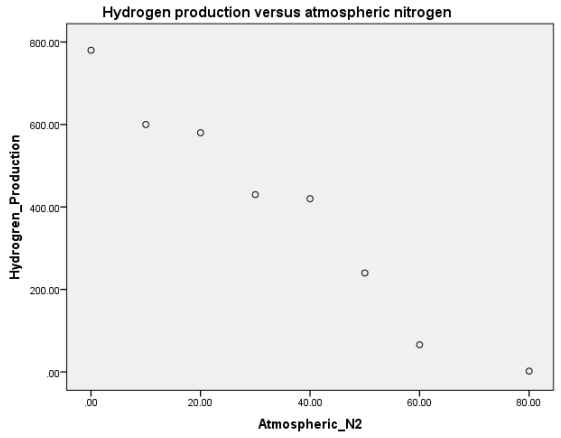
Use information in the Excel output and Residual plot below to answer questions 8 – 12:

Azolla is a water fern that fixes nitrogen and can be used as a biofertilizer on rice. However, experiments in biotechnology have shown that when *Azolla* is grown at reduced levels of atmospheric nitrogen, it produces hydrogen gas, a high energy, non-polluting fuel. Below is incomplete SPSS output showing regression analysis of data from an experiment in which *Azolla* was exposed to atmospheric N_2 levels ranging from 80% to 0% and the production of H_2 (in nmol H_2 g^{-1} fresh weight $hour^{-1}$) was measured.

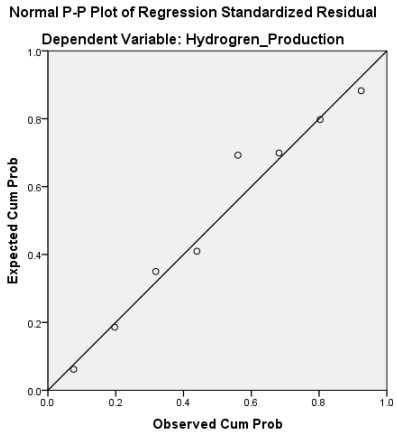
Consider also that n = 8, $\bar{x} = 36.25$ and $S_{XX} = 4,987.5$.

Atmospheric nitrogen (N_2)	Hydrogen Production (nmol H_2 g^{-1} fresh weight $hour^{-1}$)
80.00	2.00
60.00	66.00
50.00	240.00
40.00	420.00
30.00	430.00
20.00	580.00
10.00	600.00
0.00	780.00

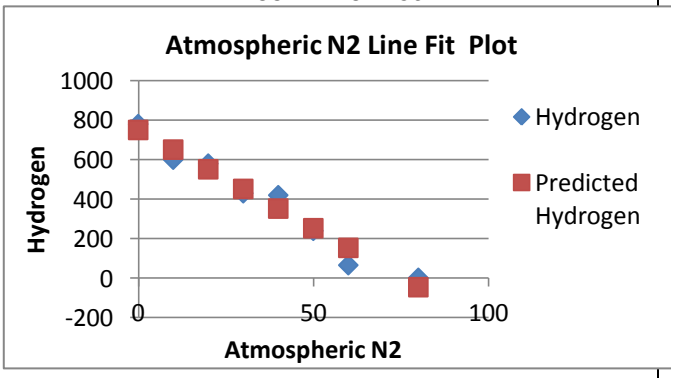
SPSS: xy graph



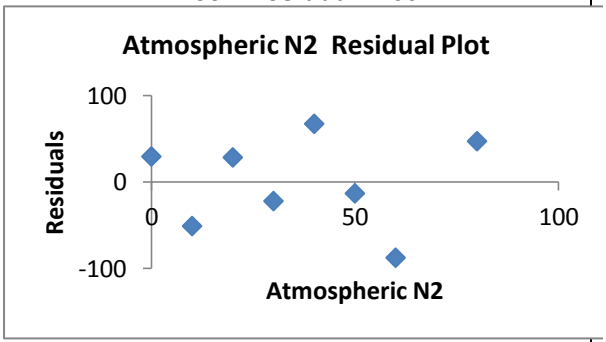
SPSS: Normal Probability Plot



Excel: Line Plot



Excel: Residual Plot



Multiple R

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.981[a]	.962	.956	56.86996

a. Predictors: (Constant), Atmospheric_N2

b. Dependent Variable: Hydrogren_Production

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	493414.347	1	493414.347	152.562	.000[b]
	Residual	19405.153	6	3234.192		
	Total	512819.500	7			

a. Dependent Variable: Hydrogren_Production

b. Predictors: (Constant), Atmospheric_N2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	750.306	35.446		21.168	7.25E-07	663.574	837.038
	Atmospheric_N2	-9.946	.805	-.981	-12.352	1.72E-05	-11.917	-7.976

a. Dependent Variable: Hydrogren_Production

Suppose the numbers highlighted in yellow in the table above were not given

>>>>>>>>>>

(a) According to the regression model what would you predict to be the rate of the hydrogen production at 25% atmospheric nitrogen? Would this be a reliable estimate?

$\hat{Y} = 750.306 - 9.946(x)$

$= 750.306 - 9.946(25)$

$= 501.656$

The predicted rate of H_2 production @ 25% atmos. N_2 is 501.656...

This would be a reliable estimate since 25% is with the Obs. range of x (interpolation)

(b) What was the residual (error) of this regression model at an atmospheric N_2 level of 60%?

$\hat{Y} = 750.306 - 9.946(60)$

$\varepsilon = Y_i - \hat{Y}$

$66 - \hat{Y}$

$66 - 153.546 = -87.546$ n mol $H_2\ y^{-1}$ fresh wt/h

(c) Calculate the linear correlation coefficient for the relationship between atmospheric N_2 and hydrogen production. What is the exact P-value?

$SS_T = SS_R + SS_E$

$= 493,414,397 + 19,405,153 = 512,819,500$

$R^2 = \frac{SS_R}{SS_E} = \frac{493,414,397}{19,405,153} = 0.96216$

$r = -\sqrt{R^2} = -\sqrt{0.96216} = -0.981$

Exact p-value is 1.72×10^{-5}

note: this is a two-tailed test if doing a one-tailed test you would divide the p-value.

(d) What is the standard error of the model?

$\hat{\sigma} = \sqrt{MS_E} = \sqrt{\frac{SS_E}{n-2}} = \sqrt{\frac{19405.153}{8-2}} = 56.86996$

(e) What percentage of variability in hydrogen production is explained by the level of atmospheric N_2?

$R^2 = 0.96216$ calculated in part c)

96.22% of the variability in h_2 production is explained by the level of the atmos N_2 or by the regression model.

(f) At the 1% significance level, test the hypothesis that there is a negative relationship between atmospheric N_2 and hydrogen production. Carry out the most appropriate, showing all steps (give both the exact P-value and the value from the table).

$H_0: \beta_1 = 0$ there is no relationship between N_2 and H_2 production

$H_A: \beta_1 \neq 0$ there is a relationship " " " " "

$t = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{-9.946}{0.805} = -12.355$

$df = n-2 = 8-2 = 6$

$P < 0.0005$ exact p-value is $\frac{1.72 \times 10^{-5}}{2} = 0.0000086$

(g) What are the value of the *F*-statistic and the P-value (both the exact value and the value from the table?

$F = t^2 = (-12.3516)^2 = 152.562$

$df(1, n-2) = (1, 8-2) = (1, 6)$

$p < 0.001$

$p = 0.0000172$

(h) Find the margin of error for a 99% confidence interval for the expected value of hydrogen production at 20% atmospheric nitrogen.

for a 99% C.I, $\alpha = 0.01$

$df = n-2, 8-2 = 6$ $t_{\alpha/2} = t_{0.005} = 3.707$

$ME = t_{\alpha/2} \times \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}$

$= 3.707 \times 56.86996 \sqrt{\frac{1}{8} + \frac{(20 - 36.25)^2}{4987.5}}$

$ME = 88.93$

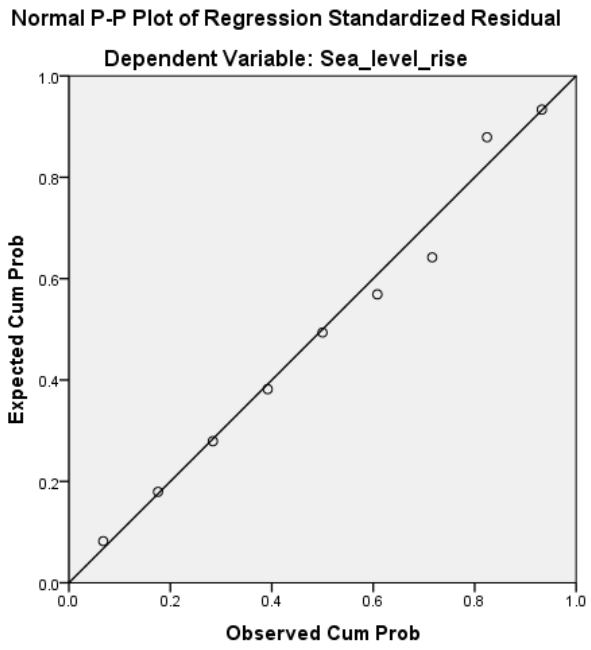
The ME for a 99% C.I. for the expected value of H_2 production @ 20% atmos N_2 is 88.93 nmol H_2/hour.

>>>>>>>>>>

Example on Sea Level Rise in Seychelles (Combing all concepts in the section)

A marine biologist used a tidal gauge to monitor mean annual sea level rise (in *mm*) in Seychelles for a period of 9 years from 2001 to 2009, using the year 2000 as baseline. The data fit the required assumptions. The table below shows incomplete SPSS output from regression analysis. Perform all calculations assuming that the numbers highlighted in yellow are not given.

Year	Sea level rise (mm) above 2000 baseline
2001	2
2002	5
2003	5
2004	9
2005	14
2006	16
2007	16
2008	18
2009	22



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.982[a]	.964	.958	1.405

a. Predictors: (Constant), Year

b. Dependent Variable: Sea_level_rise

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	365.067	1	365.067	184.881	.000[b]
	Residual	13.822	7	1.975		
	Total	378.889	8			

a. Dependent Variable: Sea_level_rise

b. Predictors: (Constant), Year

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-4933.778	363.730		-13.564	.000	-5793.862	-4073.693
	Year	2.467	.181	.982	13.597	.000	2.038	2.896

a. Dependent Variable: Sea_level_rise

(a) At the 1% significance level, test whether there is a relationship between time (in years) and mean annual sea level rise in Seychelles. In other words, test for the significance of the slope of the regression line.

H_0: $\beta_1 = 0$ (There is no relationship between time (in years) and mean annual sea level rise)
H_a: $\beta_1 \neq 0$ (There is a relationship between time (in years) and mean annual sea level rise)

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{2.467}{0.181} = 13.630$$

df = n – 2 = 9 – 2 = 7
P-value = 2 X (P < 0.0005) ==> P < 0.001
Since P ≤ α (0.01), we reject Ho with extremely strong evidence

Interpretation: At the 1% significance level, the data provide sufficient evidence to conclude that the slope of the population regression line is not 0 and therefore there is a relationship between time (in years) and mean annual sea level rise.

(b) Find a 95% confidence interval for the slope of the regression line that relates mean annual sea level rise to time (in years). Use this confidence interval to determine whether the slope is significant.

>>>>>>>>>> $\alpha = 0.05$ at $df = n-2 = 7-2 = df = 5$ $t_{0.025,7} = 2.365$

$\hat{\beta}_1 \pm t^* \times SE(\hat{\beta}_1)$

$2.467 \pm 2.365 \times 0.181$

$(2.039, 2.895)$

we can be 95% ... slope of the regression line is somewhere between 2.039 and 2.895

Since 0 isn't in the C.I, there is significance.

(c) Calculate a 95% confidence interval for the effect of an additional 5 years on the average sea level.

The addition of 5 years is $5 \cdot \beta_1$

$5\hat{\beta}_1 \pm t_{0.025,7} \times 5\,SE(\hat{\beta}_1)$

$\therefore$

$(10.20, 14.48)$

>>>>>>>>>>

(d) What is the standard error of the model?

$$SS_{Error} = SS_{TOTAL} - SS_{REGR} = 378.889 - 365.067 = 13.822$$

$$\hat{\sigma} = \sqrt{MS_{ERROR}} = \sqrt{\frac{SS_{Error}}{n-2}} = \sqrt{\frac{13.822}{9-2}} = \sqrt{1.9746} = 1.405$$

Rise in Global Mean Sea Level

The change in global mean sea level per decade (10 years) from 1930 to 2010 (considering 1930 as baseline year 0, thus covering 8 decades), can be described by the following linear equation: $\hat{y} = 20.07 + 18.35x$. The correlation coefficient relating time and sea level is: r = 0.988703. [Years were coded as decades.]

>>>>>>>>>>

(a) What percentage of variability in global mean sea level is explained by time?

(coeff of determination is R^2

$r^2 = 0.98703^2 = 0.97753$

Thus 97.75% of variability ...

(b) What is the value of the F-statistic for determining whether the relationship between time and sea level is significant?

$$F = \frac{SS_R \times (n-2)}{(SS_T - SS_R)/SS_T} = \frac{(n-2)\,SS_R/SS_T}{(SS_T - SS_R)/SS_T}$$

$$= \frac{(n-2)R^2}{1-R^2}$$

$$= \frac{(8-2)(0.97753)}{1-0.97753}$$

$$= 261.022697$$

(c) What is the standard error of the slope of the regression line?

$$t = \sqrt{F} = \sqrt{261.022697} = 16.1562$$

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \Rightarrow SE(\hat{\beta}_1) = \frac{\hat{\beta}_1}{16.1562} = 1.1361$$

>>>>>>>>>>

4.9 More on Assumptions and Transformations of Data

Variations in notation for the linear regression equation

$$\hat{\mu}\{Y \mid X) = 3.707 - 0.012X$$

Same as: $\hat{y} = 3.707 - 0.012x$

Assumptions (Conditions) for Regression Inferences and Robustness

1. **Linearity:**
   Checking: Scatterplot or residual plot

   Robustness: If no linear relationship exists in the data, don't use linear regression!

   Solution: Consider transformation. If the problem is only with linearity (not with equal standard deviations), try transforming ***x***. If both problems exist, transform ***y*** first.

2. **Equal standard deviations:**

   Checking: Scatterplot or residual plot (but residual plot is best)

   Robustness: The consequences for violating this assumption are critical. When there is lack of equal variability, the resulting standard errors inaccurately estimate their respective parameters, thus confidence intervals and hypothesis tests can be misleading.

   Solution: Consider transformation of ***y***.

3. **Normal populations:**

   Checking: Normal Probability Plot (Look for an increasing linear pattern)

   Robustness: Standard errors are robust to non-normal distributions (*t*-tools). The consequences of violating this assumption are usually minor. The only situation of large concern is when the distributions have long tails, outliers are present, and/or sample sizes are small. In terms of prediction, normality is critical.

   Solution: Consider transformation of ***y***, only if the problem is very serious.

4. **No Serious Outliers:**

   Checking: Scatterplot or Normal Probability Plot

   Robustness: Several Significant outliers can drastically change the regression model (just as for correlation)

   Solution: Consider transformation of ***y***, if there are several serious outliers and sample size is small.

5. **Independent observations:** The observations of the response variable are independent of one another. This implies that the observations of the predictor variable not need to be independent.

   Checking: Evaluate the research design.

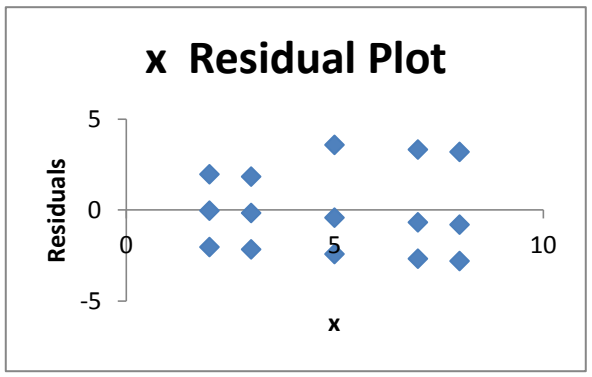
   Robustness: Lack of independence causes inaccurate standard errors.

   Solution: Difficult to solve unless you revise the research design.

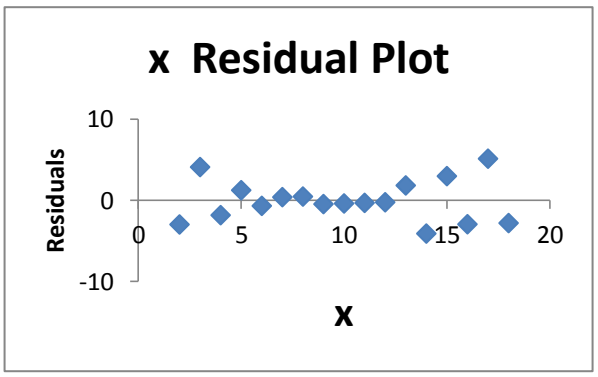
Checking Assumptions with Scatterplots, Residual Plots and Normal Probability Plots

Which assumption or assumptions are violated in the data sets below?

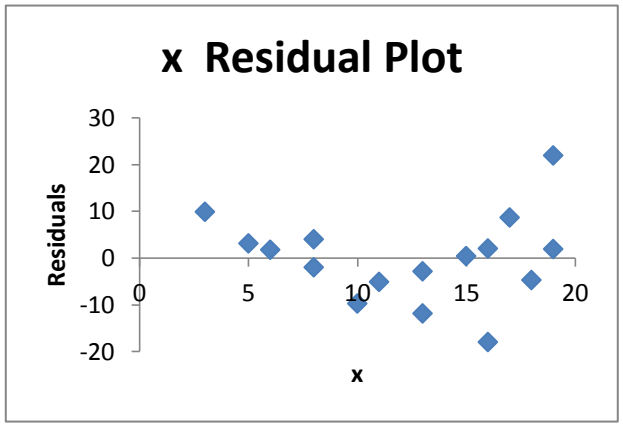
Example 1



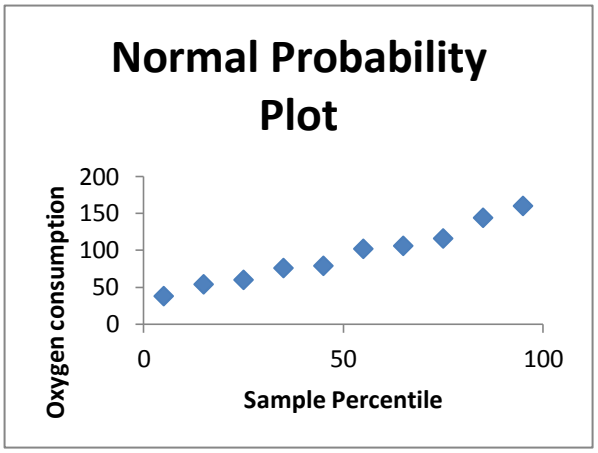
Example 2



Example 3



Example 4



Data for Example 3 above:

x	3	5	6	8	8	10	11	13	13	15	16	16	17	18	19	19
y	4	4	6	9	15	8	16	16	25	35	20	40	50	40	50	70

>>>>>>>>>>

Ex 1: no serious violations

Ex 2: only equal std dev violated

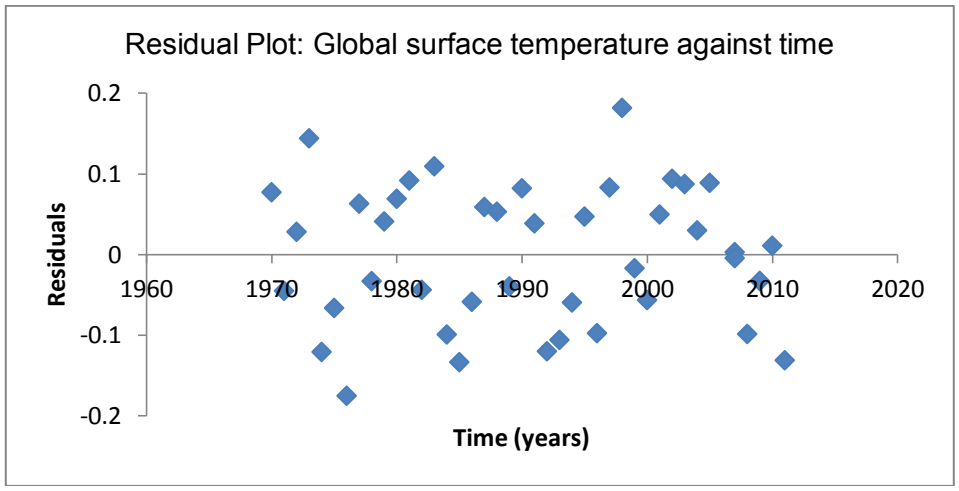
Ex 3: linearity and equal std dev and outliers are violated

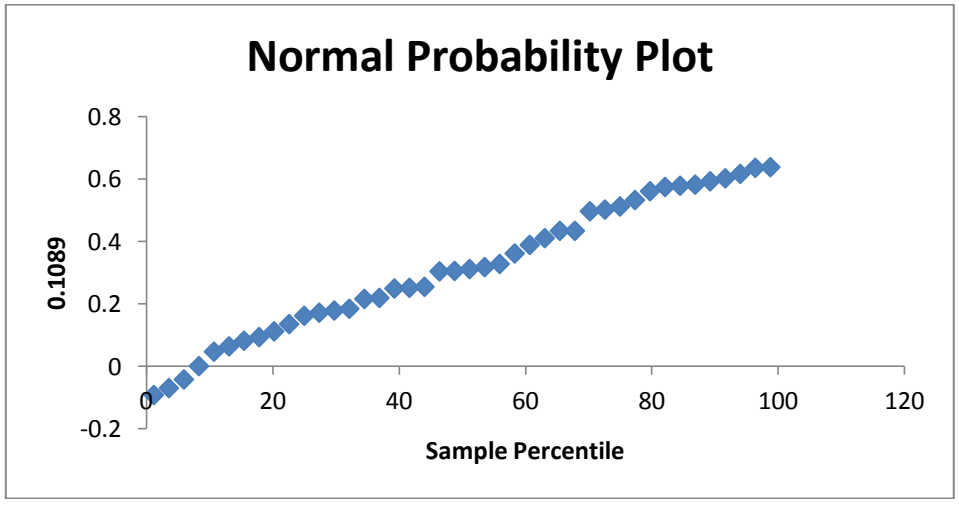
Ex 4: data is normal but normality plot cannot be used to test other violations.

>>>>>>>>>>

Checking the Assumptions of Regression for the Data on Global Surface Temperature Against Time (1969 – 2011)

[Examine the NOAA Graph: 1880 – 2011]





Conclusion:

- The data points in the residual plot fall roughly in a horizontal band centered and symmetric about the x-axis
- The normal probability plot is approximately linear
- Therefore, these data fit all the assumptions of regression analysis

Interpretation of Model Effects in SLR after Log Transformation

***SLR model of Blood Pressure on Age*:** The effect of age on mean blood pressure is measured as the change in the mean blood pressure associated with a 1-unit (one year) increase in age. This effect is measured as β_1 in the model below.

$\mu(bp \mid age) = \beta_0 + \beta_1 age$

This is known as an additive effect.

IN GENERAL, for *Y* vs. *X*:

Additive change of *k* units in *X* → *Additive* change of $k\beta_1$ in the mean of *Y*.
(*X* + *k*)

Suppose (in separate circumstances) that the following natural log transformations were required. Interpretation of the model effect on the original scale will follow.

most common →

i) a natural log transformation was used on the response variable only. (ln *Y* vs. *X*)
ii) a natural log transformation was used on the predictor variable only. (*Y* vs. ln *X*)
iii) a natural log transformation was used on both variables. (ln *Y* vs. ln *X*)

***Case i)*: (ln *Y* vs. *X*)**

IN GENERAL, for ln *Y* vs. *X*:

Additive change of *k* units in *X* → *Multiplicative* change of $e^{k\beta_1}$ in the median of *Y*. (Take antilog of the slope)

c = final - initial

Suppose we had this model:

Model: $\mu(\ln(bp) \mid age) = \beta_0 + \beta_1 age$ → $\hat{\mu}(\ln(bp) \mid age) = 4.481 + 0.010 age$

The additive effect of *age* on $\mu_{\ln(bp)}$ is β_1.

Back Transforming to the Original Scale

Example 1: the additive effect of 1-year (*k* = 1) in *age* is associated with a multiplicative effect of $e^{1(\beta_1)} = e^{\beta_1}$ on *Median*(*bp*). It is estimated that a 1-year increase in *age* is associated with a multiplicative change of $e^{0.010} = 1.01$ in *Median*(*bp*). In other words, the median *bp* at *age* + 1 is estimated to be 1.01 times (1.01 – 1 = 0.01 ==> 1% higher than) the median *bp* at the given *age*.

Example 2: It is estimated that a 5-year increase in *age* is associated with a multiplicative change of $e^{5(0.010)} = 1.051$ in *Median*(*bp*). In other words, the median *bp* at *age* + 5 is estimated to be 1.051 times (1.051 – 1 = 0.051 ==> 5.1% higher than) the median *bp* at the given *age*.

Example 3: The median blood pressure of 55-year-olds will be 1.22 times (1.22 – 1 = 0.22 ==> 22% higher than) the median blood pressure of 35-year-olds ($e^{20(0.010)} = 1.22$).

[Likewise for confidence intervals, take the antilog of the two endpoints.]

***Case ii)*: (*Y* vs. ln *X*)**

IN GENERAL, for *Y* vs. ln *X*:

Multiplicative change by a factor of *k* in *X* (*X* × *k*) → *Additive* change of $\beta_1\ln(k)$ in the mean of *Y*.

Suppose we had this model:

Model: $\mu(bp \mid \ln(age)) = \beta_0 + \beta_1\ln(age)$ → $\hat{\mu}(bp \mid \ln(age))$ = -81.784 + 58.967ln(*age*)

The additive effect of ln(*age*) on μ_{bp} is β_1.

Example 1: A multiplicative change in *age* by a factor of *k* is associated with an additive change of ($\beta_1\ln(k)$) in μ_{bp}. It is estimated that a multiplicative change in *age* by a factor of *k* is associated with an additive change of 58.967ln(*k*) in μ_{bp}.

Example 2: It is estimated that aging from 40 to 50 (*k* = 50/40 = 1.25) is associated with an additive increase in μ_{bp} of 58.967(ln(1.25)) = (58.967)(0.22314) = 13.16.

Example 3: It is estimated that aging from 28 to 35 (*k* = 35/28 = 1.25) is associated with an additive increase in μ_{bp} of 58.967(ln(1.25)) = (58.967)(0.22314) = 13.16.

Example 4: It is estimated that aging from 30 to 50 (*k* = 50/30 = 1.67) is associated with an additive increase in μ_{bp} of 58.967(ln(5/3)) = (58.967)(0.5108) = 30.12.

***Case iii)*: (ln *Y* vs. ln *X*)**

IN GENERAL, for ln *Y* vs. ln *X*:

Multiplicative change by a factor of *k* in *X* (*X* × *k*) → *Multiplicative* change of k^{β_1} in the median of *Y*.

Suppose we had this model:

Model: $\mu(\ln(bp) \mid \ln(age)) = \beta_0 + \beta_1\ln(age)$ → $\hat{\mu}(\ln(bp) \mid \ln(age))$ = 3.332 + 0.426ln(*age*)

The additive effect of ln(*age*) on $\mu_{\ln(bp)}$ is β_1.

Example 1: A multiplicative change in *age* by a factor of *k* is associated with a multiplicative change of k^{β_1} in *Median*(*bp*). It is estimated that a multiplicative change in *age* by a factor of *k* is associated with a multiplicative change of $k^{0.426}$ in *Median*(*bp*).

Example 2: It is estimated that aging from 40 to 50 (*k* = 50/40 = 1.25) is associated with a multiplicative change of $1.25^{0.426}$ = 1.0997 in *Median*(*bp*). That is, the Median blood pressure will be 1.0997 times (1.0997 – 1 = 0.0997 ==> 9.97% higher) what it was at the age of 40.

Example 3: It is estimated that aging from 35 to 55 (*k* = 55/35 = 1.57) is associated with a multiplicative change of $1.57^{0.426}$ = 1.212 in *Median*(*bp*), that is there is a 21.2% increase (1.212 – 1 = 0.212).

Example on Log Transformed Data [Like Case (i) described above]
Suppose the relationship between the annual rate of hip fractures (per 100,000 people) and age follows the following model: $\hat{\mu}(\ln(fractures)\,|\,age) = -2.09 + 0.0912\ age$. Suppose further that a 98% confidence interval for the slope based on the logged data is (0.0723, 0.1101)

(a) For an increase in age from 40 to 50 years old, what would be your interpretation regarding the rate of hip fractures on the original scale?

>>>>>>>>>>

$e^{(2.13)} = e^{(10)(.213)} = e^{(10)(0.0912)} = 2.489$

The median rate of hip fractures / 100,000 ppl. at 50 years of age will be 2.489 times the median rate of hip fractures at 40 years.

(b) What is the estimated rate of hip fractures on the original scale for people that are 80 years old?

$\hat{\mu}(\ln(\text{fractures})\,|\,\text{age}) = -2.09 + 0.0912(80) = 5.206$

$e^{5.206} = 182.363$

It is estimated that the rate of hip fractures for ppl. that are 80 years old is on average 18

(c) Interpret the 98% confidence interval for the slope on the original scale. Also, based on this confidence interval back transformed to the original scale, would you conclude that there is a relationship between the annual rate of hip fractures and age? Explain your answer.

Take the antilog

$(e^{0.703}, e^{0.1101}) = (1.075, 1.116)$

It is estimated with 98% confidence that a 1 year increase in age is associated with a multiplicative change of 1.075 to 1.116 in the median rate of hip fractures per 100,000.

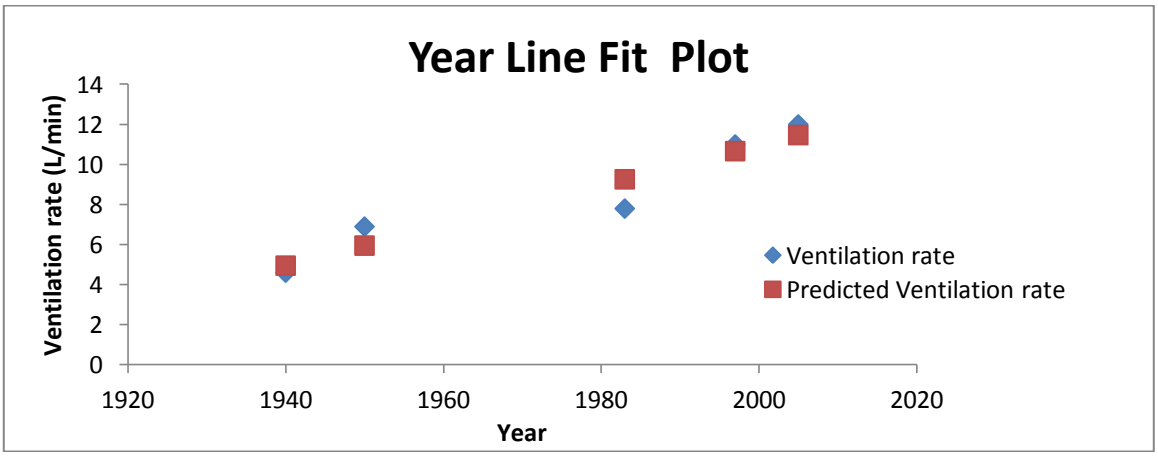
The slope is sig. as 1 isn't in the C.I.

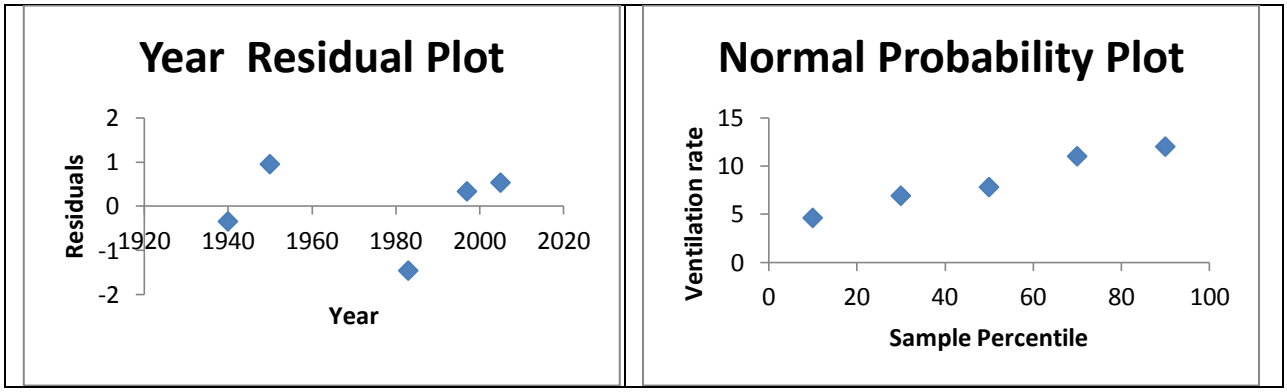
>>>>>>>>>>

Practice Question on Inferences for Regression: Changes in Human Ventilation Rate over Time
A group of medical researchers suspected that various growing health issues such as lack of fitness, obesity and increasing prevalence of diseases such as cancer, diabetes and heart disease may have caused ventilation rates in humans to change over the past few decades. Examination of past records provided data as shown below, using 1940 as a baseline. At the 5% significance level, test whether the data provide sufficient evidence to conclude that there has been significant change in human ventilation rates over time.

(a) Check whether the data fit the assumptions of regression analysis.

Year	Average human ventilation rate (L/min)
1940	4.6
1950	6.9
1983	7.8
1997	11
2005	12
$\bar{x} = 1975$	
$s_x = 28.714108$	





Conclusion regarding the assumptions: These graphs show that the data approximately fit the linear regression model. Though there are few data points, the residual plot (right) shows no noticeable violation of equal standard deviations and linearity, while the normal probability plot shows that the data are approximately normally distributed.

Regression Statistics	
Multiple R	0.950454
R Square	0.903362
Adjusted R Square	0.87115
Standard Error	1.088061
Observations	5

ANOVA table for Regression analysis					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	33.20037	33.20037	28.04375	0.01314
Residual	3	3.55163	1.183878		
Total	4	36.752			

***t*-test for the Significance of the Slope**						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-189.699	37.42242	-5.06912	0.014823	-308.794	-70.6039
Year	0.100334	0.018946	5.295635	0.01314	0.040037	0.16063

Note: Suppose the numbers highlighted in yellow are missing values in the table.

Linear Correlation Coefficient (r) = 0.950

(a) Give the regression equation and interpret the meaning of the slope and the y-intercept in terms of the research problem.

The regression equation is: $\hat{y} = -189.699 + 0.1003x$

The meaning of the slope is that ventilation rates in humans have increased, on average, by 0.1003 L/min per year from 1940 to 2005 (or 1.003 liters per decade).
The meaning of the y-intercept is WOW!!!!

(b) At the 5% significance level, test for the significance of the slope. In other words, determine if there has been a significant change in human ventilation rates over time.

Step 1: Regression t-test is selected because the purpose is to test of the slope is significantly different from 0.

Step 2:
H_0: β1 = 0 (There has been no significant change in human ventilation rates over time or there is no relationship between time and human ventilation rates.)
H_a: β1 ≠ 0 (There has been significant change in human ventilation rates over time or there is a relationship between time and human ventilation rates.)

Step 3:

Standard Error of the Slope:

Given in the computer output as: $SE(\hat{\beta}_1) = 0.018946$

The Regression t-statistic:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0.100334}{0.018946} = 5.296$$

Step 4: Decide to reject or not reject Ho

$df = n - 2 = 5 - 2 = 3$

So, P-value is between (0.005 and 0.01) x 2: 0.01 < P < 0.02
Since P-value ≤ α, reject H_0 with strong evidence.

Step 5: At the 5% significance level, the data provide sufficient evidence to conclude that there has been significant change in human ventilation rates over time.

Alternate Method

The significance of the slope could likewise be tested with ANOVA as follows:

$$F = \frac{SS_{REGR}/1}{SS_{Error}/(n-2)} = \frac{33.20037/1}{3.55163/3} = \frac{33.20037}{1.183878} = 28.04$$

At $df = (1, n-2) = (1,3)$, P-value: 0.01 < P < 0.025

[**Note:** Exact P-value = 0.01314 for both tests]

(c) Calculate a 95% confidence interval for the slope of the regression line.

At the 95% confidence level and df = 5 – 2 = 3, $t_{\alpha/2} = 3.182$

$$\hat{\beta}_1 \pm t_{\alpha/2} \times SE(\hat{\beta}_1)$$

$$0.100334 \pm 3.182 \times 0.018946$$

0.100334 ± 0.060286

Or 0.040 to 0.161 L/min per year

Interpretation: We can be 95% confident that the increase in human ventilation rates over time is somewhere between 0.040 and 0.161 L/min per year.

(d) Calculate a 95% confidence interval for the mean response of variable y (human ventilation rate) in 1990.

At the 95% confidence level and df = 5 – 2 = 3, t_{Crit} = 3.182

The point estimate of variable y in 1990 was:

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

$$\hat{y} = -189.699 + 0.1003(1990) = 9.966 \text{ L/min}$$

The endpoints are given by:

$$\hat{y}_p \pm t_{\alpha/2,n-2} \times \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}$$

Three Methods of Finding S_{xx}

Method 1:

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \Rightarrow \sqrt{S_{xx}} = \frac{\hat{\sigma}}{SE(\hat{\beta}_1)}$$

$$\Rightarrow S_{xx} = \left(\frac{\hat{\sigma}}{SE(\hat{\beta}_1)}\right)^2 = \left(\frac{1.08806}{0.018946}\right)^2 = 3298.15 = 3298$$

Method 2:

$$S_{xx} = (n-1)s_x^2$$
$$= (5-1)(28.714108)^2 = 3298.0000$$

Method 3:

$$S_{xx} = \sum (x_i - \bar{x})^2$$

[Calculate from raw data]

$$9.966 \pm 3.182 \times 1.08806\sqrt{\frac{1}{5} + \frac{(1990-1975)^2}{3298}}$$

9.966 ± 1.793

Or 8.173 to 11.759 L/min

Interpretation: We can be 95% confident that the mean ventilation rate of humans at 30 years after baseline was somewhere between 8.173 and 11.759 L/min.