CORRELATION ANALYSIS AND LINEAR REGRESSION AND

DESCRIBING THE RELATION BETWEEN TWO
VARIABLE - SCATTER PLOTS
CORRELATION ANALYSIS
LINEAR REGRESSION ANALYSIS

UNIVARIATE AND BIVARIATE

• Univariate — Single variable to be measured for each individual in the study.

• Bivariate — Data in which two variables are measured on an individual.

DESCRIBING THE RELATION BETWEEN TWO VARIABLES

- Correlation analysis are most used statistical procedure used by business decision makers for analyzing the relationship between two variables.
- Common visual tool for two-variable relationship: scatter plots (or scatter diagram)
- Scatter diagram: A two-dimensional plot showing the values for the joint occurrence of two quantitative variables measured on the same individual.
 - Each individual in the data set is represented by a point in the scatter diagram.

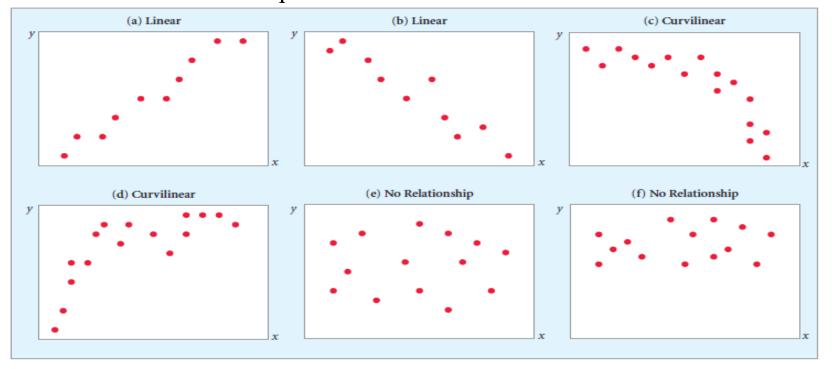
DESCRIBING THE RELATION BETWEEN TWO VARIABLES

- Dependent (response) variable, y a variation wish to explain
 - the response variable is plotted on the vertical axis.
- Independent (explanatory) variable, x used to explain variation in dependent variable.
 - The explanatory variable is plotted on the horizontal axis

CAUTION!!!

Do not connect points when drawing scatter diagram

Two- variable relationship



- (a) and (b) examples of strong *linear* (or straight line) relationships between *x* and *y*
- (c) and (d) examples of relationship between the x and y variable is nonlinear.
- (e) and (f) examples in which there is **no identifiable relationship** between the two variables. This means that as *x* increases, *y* sometimes increases and sometimes decreases but with no particular pattern. no relation between the variables.

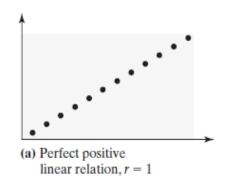
CORRELATION COEFFICIENT

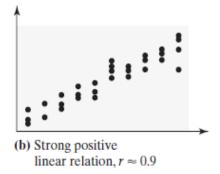


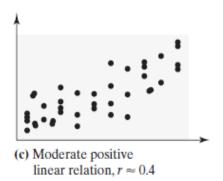
- What it is? A quantitative measure of the strength of the linear relationship between two variables.
- What can we measure?
 - Strength of the relationship: weak, moderate or strong
 - Is indicate through the value of correlation coefficient, r
 - The more correlation differs from 0.0, the stronger the linear relationship between the two variables.
 - Direction of the relationship: positive or negative or no relationship
 - Is indicate through the sign (+ or -) of correlation coefficient, r
 - Significance test for correlation: correlation between variables exists or not
 - Is indicate from the hypothesis testing of ρ (rho).

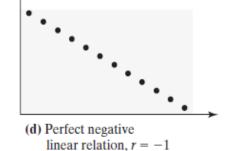
PROPERTIES OF THE LINEAR CORRELATION COEFFICIENT

- 1. The linear correlation coefficient is always between -1 and 1, inclusive. That is, $-1 \le r \le 1$.
- 2. If $r = \pm 1$, then a perfect positive linear relation exists between the two variables. See Figure (a).
- 3. If r = -1, then a perfect negative linear relation exists between the two variables. See Figure (d).
- 4. The closer r is to ± 1 , the stronger is the evidence of positive association between the two variables. See Figures (b) and (c).



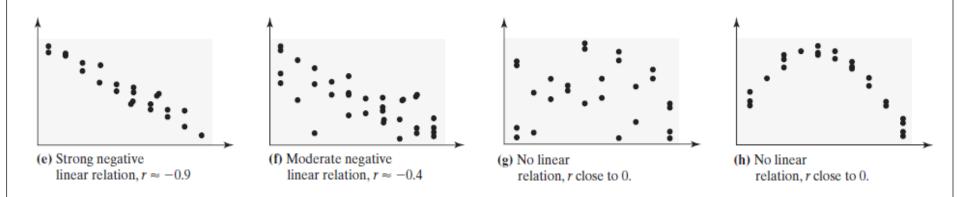






PROPERTIES OF THE LINEAR CORRELATION COEFFICIENT

- 5. The closer *r* is to -1, the stronger is the evidence of negative association between the two variables. See Figures (e) and (f).
- 6. If *r* is close to 0, then little or no evidence exists of a *linear* relation between the two variables. So *r* close to 0 does not imply no relation, just no *linear* relation. See Figures (g) and (h).



CORRELATION COEFFICIENT

$$r = \frac{n\sum xy - \sum x\sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} \quad \text{OR} \quad r = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sqrt{[\sum (x - \overline{x})^2][\sum (y - \overline{y})^2]}}$$

where:

r =Sample correlation coefficient

n =Sample size

x = value of the independent variable

y =value of the dependent variable

also called as Pearson
Product Moment
Correlation

However, because the calculations are rather tedious and long, need to have computer/statistical tool to perform the computation.

SIGNIFICANCE TEST FOR THE CORRELATION

- What it is?: a hypothesis testing to determine whether there exists relationship between the variables.
- Hypothesis: using the ρ (rho) notation

 H_0 : $\rho = 0$ (no correlation)

 H_A : $\rho \neq 0$ (correlation exists)

• Test statistics for correlation:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} \qquad df = n - 2$$

where:

t =Number of standard errors r is from 0

r =Sample correlation coefficient

n =Sample size

SIGNIFICANCE TEST FOR THE CORRELATION

Decision rule: calculate a t value (t- test statistics) and compare it to the critical value, *t* in the t-distribution table. Since it involve with 2 tailed;

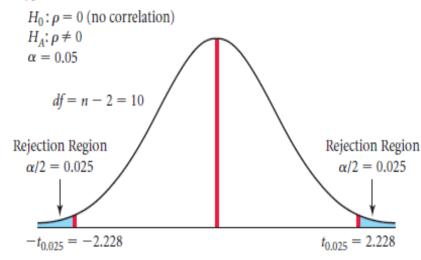
Lower tail

- ightharpoonup if the -t \geq critical value, - $t_{lpha/2}$ therefore accept $\mathrm{H_o}$
- ightharpoonup if the -t < critical value, - $t_{lpha/2}$ therefore reject $H_{
 m o}$

• Upper tail

- \geq if the $t \geq$ critical value, $t_{\alpha/2}$ therefore reject H_0
- \succ if the t < critical value, $t_{lpha/2}$ therefore accept H_o

Hypothesis:



CORRELATION ANALYSIS- EXAMPLE

A student intern at the investment firm of McMillan & Associates was given the assignment of determining whether there is a positive correlation between the number of individual stocks in a client's portfolio and the annual rate of return for the portfolio. The intern selected a simple random sample of 10 client portfolios and determined the number of individual company stocks and the annual rate of return earned by the client on his or her portfolio. To determine whether there is a statistically significant positive correlation between the two variables. The following sample data were obtained:

Number of Stocks	Rate of Return
9	0.13
16	0.16
25	0.21
16	0.18
20	0.18
16	0.19
20	0.15
20	0.17
16	0.13
9	0.11

Steps to solution:

Specify the population parameter of interest:

The intern wishes to determine whether the number of stocks is positively correlated with the rate of return earned by the client. The parameter of interest is, therefore, the population correlation, ρ

Formulate the null and the alternative hypothesis

$$H_0: \rho = 0$$

 $H_A: \rho > 0$ (claim)

Construct the rejection region and decision rule

This is one-tailed test with upper tail (right hand) of sampling distribution. With 0.05 level of significance, the degree of freedom, n-2=10-2=8 degree of freedom. The t-critical value, $t_{0.05}=1.860$

Decision rule: If t > 1.860, then reject H_0 , otherwise accept H_0

Compute the correlation coefficient and test statistics

compute the sample correlation coefficient using software tools r = 0.7796

Compute the test statistic

This problem will use t-test statistics.

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.7796}{\sqrt{\frac{1-0.7796^2}{10-2}}} = 3.52$$

The decision rule: If t > 1.860, then reject H_0 , otherwise accept H_0

Since 3.52 > 1.860, therefore H_0 is rejected

Draw a conclusion

Because the null hypothesis is rejected, the sample data do support that there is a positive linear relationship between the number of individual stocks in a client's portfolio and the portfolio's rate of return.

CORRELATION ANALYSIS- EXAMPLE (R-programming)

The investment firm Harmonic Investments wants to manage the pension fund of a major Chicago retailer. For their presentation to the retailer, the Harmonic analysts want to use correlation analysis to know the relationship between profits and number of employees for 50 Fortune 500 companies in the firm's portfolio. The data for the analysis are contained in the file **Fortune 50**.

Rank	Company	Revenues	Profits	Assets	Stockholders Equity	Market Value	Earnings per share	Total Return to Investors 1997	Employees
1	General Electric	90840	8203	304012	34438	260147.17	2.46	51.09	276000
2	Ford Motor	153627	6920	279097	30734	73923	5.62	57.01	363892
3	General Motors	178174	6698	228888	17506	54243.84	8.62	19.47	608000
4	Exxon	122379	8460	96064	43660	158783.63	3.37	28.33	80000
5	Citicorp	34697	3591	310897	21196	66105.35	7.33	24.85	93700
6	Chase Manhattan Corp.	30381	3708	365521	21742	58150.63	8.03	25.44	69033
7	Travelers Group	37609	3104	386555	20893	69419.71	2.54	79.83	67250
8	Intl. Business Machines	78508	6093	81499	19816	98321.88	6.01	39.64	269465
9	State Farm Insurance Cos.	43956.97	3833.27	103626.18	37635.38	0	0	0	72655
10	American International Group	30519.51	3332.34	163970.69	24001.13	87964.42	4.73	51.13	40000
11	Fannie Mae	27776.9	3055.8	391672.7	13793	67592.92	2.83	54.52	3500
12	AT&T	53261	4638	58635	22647	105878.74	2.84	46.16	128000
13	Philip Morris	56114	6310	55947	14920	102931.62	2.58	25.01	152000
14	Morgan Stanley Dean Witter Discover	27132	2586	302287	13956	46842.41	4.25	81.35	47277
15	Merrill Lynch	31731	1906	292819	8329	29152.16	4.83	81.35	56600
16	BankAmerica Corp.	23585	3210	260159	19837	57839.78	4.32	49.02	77000

CORRELATION ANALYSIS- EXAMPLE (R-programming)

SOLUTION:

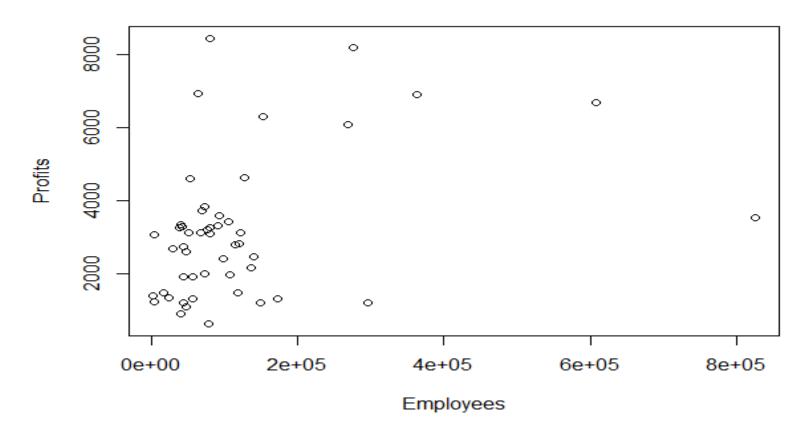
Scatter plot for independent and dependent variable y axis (dependent) – profits; x axis (independent) – employees

Significance test for the correlation and calculate the correlation coefficient (pearson product moment correlation or r-value)

<u>Significance test</u>: To test either is there any relationship between variable "Profits" with variable "Employees".

<u>correlation coefficient value (r-value)</u>: to see the strength and direction of the relationship

> plot(Profits ~ Employees, data=Fortune50)



summary: scatter plot shown fairly linear relationship

Formulate the null and the alternative hypothesis

```
H_0: \rho = 0
                   H_{\Lambda}: \rho \neq 0 (claim)
> with(Fortune50, cor.test(Profits, Employees))
        Pearson's product-moment correlation
data: Profits and Employees
t = 2.7057, df = 48 p-value = 0.009409
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.09504911 0.58308036
sample estimates:
0.3637726
```

correlation coefficient value: r = 0.363 indicate weak positive linear relationship

p-value $< \alpha = 0.05$, therefore, reject H_0 which indicate there is a correlation (relationship) between the variables

SIMPLE LINEAR REGRESSION ANALYSIS

THE REGRESSION MODEL & ASSUMPTIONS

- Statistical method which is used to analyze the relationship → Regression Analysis
- When we have 2 variables → referred to simple regression analysis
- When the 2 variables (independent & dependent) have linear relationship → referred as simple linear regression

THE REGRESSION MODEL & ASSUMPTIONS

 $y = \beta_0 + \beta_1 x + \varepsilon$

where:

y =Value of the dependent variable

 $x = \text{Value } \phi f$ the independent variable

 β_0 = Population's y intercept

 β_1 = Sløpe of the population regression line

 $\varepsilon = R$ and om error term

Linear Component

Random error component
- maybe positive, zero or negative

THE REGRESSION MODEL & ASSUMPTIONS

$$y = \beta_0 + \beta_1 x + \varepsilon$$
 where:
$$y = \text{Value of the dependent variable}$$

$$x = \text{Value of the independent variable}$$

$$\beta_0 = \text{Population's } y \text{ intercept}$$

$$\beta_1 = \text{Slope of the population regression line}$$

$$\varepsilon = \text{Random error term}$$

Population regression coefficient

 β_1 (slope): can be either positive, zero or negative. Example:

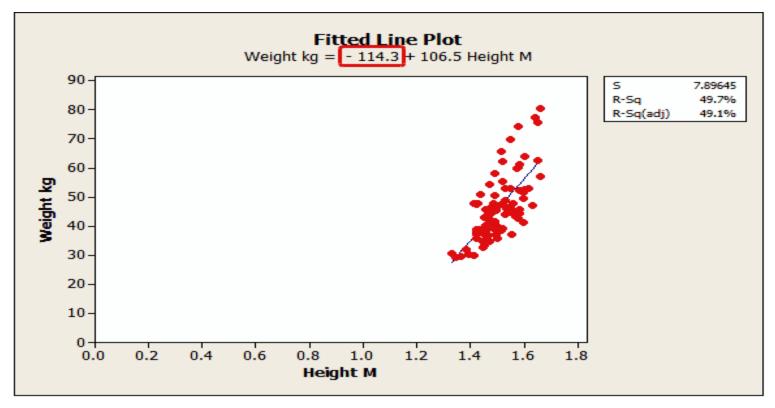
- 1. if $(\beta_1=12)$, it indicate that for a 1 unit increase in x we can expect an average 12 unit increase in y
- 2. if $(\beta_1 = -12)$, it indicate that for a 1 unit decrease in x we can expect an average 12 unit increases in y

INTERPRETATION OF THE Y-INTERCEPT

- The *y*-intercept of any line is the point where the graph intersects the vertical axis. It is found by letting x = 0 in an equation and solving for *y*.
- In general, we interpret a *y*-intercept as the value of the response variable when the value of the explanatory variable is 0.
- To interpret the *y*-intercept, we must first ask two questions:
 - 1. Is 0 a reasonable value for the explanatory variable?
 - 2. Do any observations near x = 0 exist in the data set?

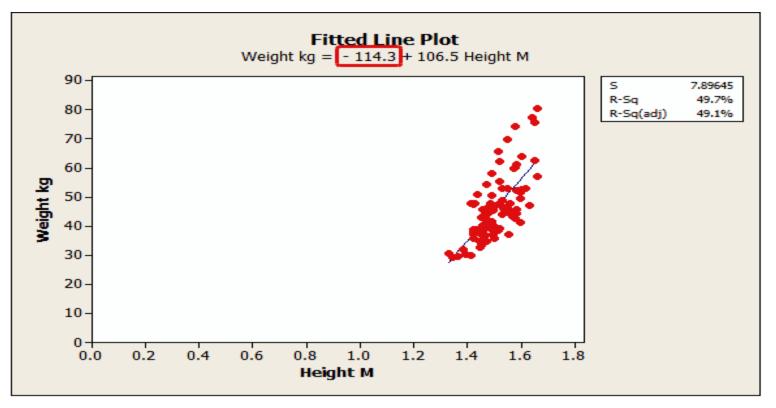
If the answer to either of these questions is NO, we do not interpret the *y*-intercept.

1. Is 0 a reasonable value for the explanatory variable?



If height is zero, the regression equation predicts that weight is -114.3 kilograms! No human can have zero height or a negative weight!

2. Do any observations near x = 0 exist in the data set?

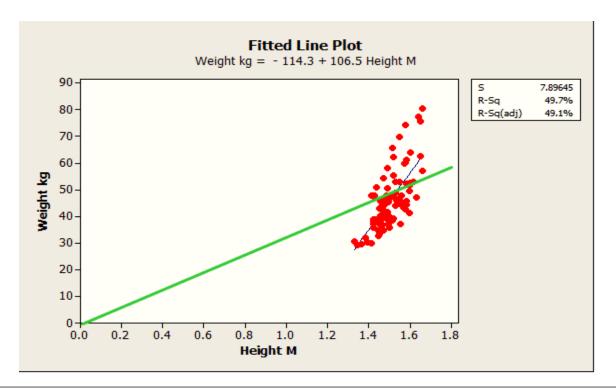


Should NEVER use a regression model to make a prediction for a point that is outside the range of your data because the relationship between the variables might change.

If you didn't collect data in this all-zero range, you can't trust the value of the constant.

INTERPRETATION OF THE Y-INTERCEPT

- Regression constant is generally not worth interpreting.
- But Why Is it Crucial to Include the Constant in a Regression Model?
 - if you don't include the constant, the regression line is forced to go through the origin.

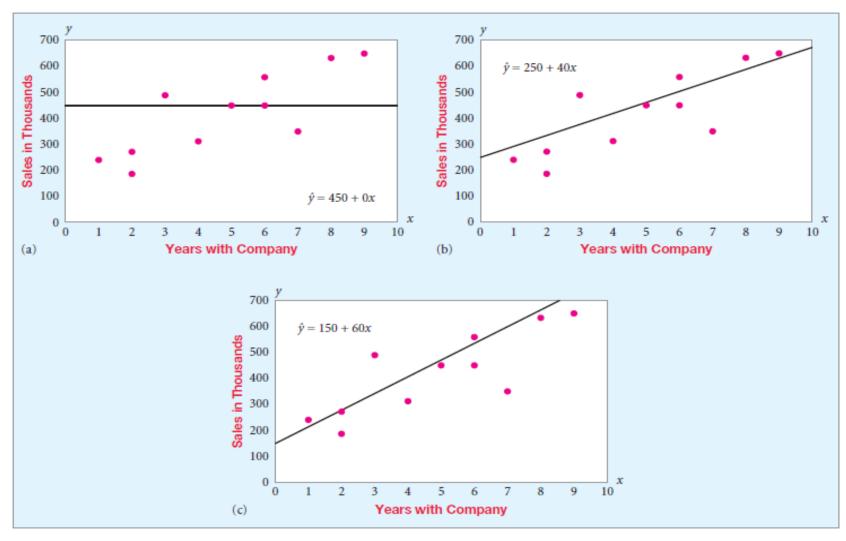


The green line just doesn't fit.

The slope is way off and the predicted values are biased. For the model without the constant, the weight predictions tend to be too high for shorter subjects and too low for taller subjects.

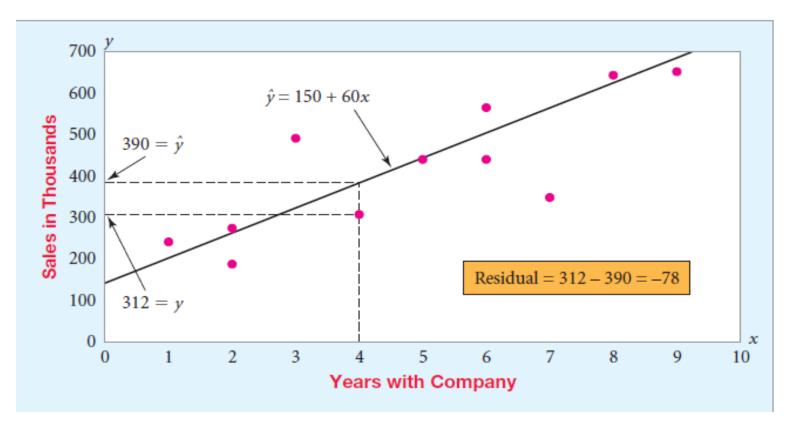
REGRESSION ANALYSIS

- To estimate regression equation (model), we need to identify the value of β_0 and β_1 .
- From the regression model we can identify the regression line.
 - regression line: through the sample data is the best estimate of the population regression line
 - we must establish a criterion for selecting the best line
 - the criteria is called as LEAST SQUARES CRITERION
 - Least Square criterion: criteria to determine a regression line that minimize the sum of squares prediction error
 - Prediction error \rightarrow Residual (the difference between the actual value of the dependent variable (y) and the value predicted (\hat{y}) by the regression model).



Possible regression line - We must establish a criterion for selecting the best line (**least square criterion**)

REGRESSION ANALYSIS



data value with
$$x = 4$$
:
 $\hat{y} = 150 + 60(4) = 390$

RESIDUAL ANALYSIS

 $\hat{y} = 150 + 60x$

Residual

X	ŷ	У	$y - \hat{y}$	$(y-\hat{y})^2$
3	330	487	157	24,649
5	450	445	-5	25
2	270	272	2	4
8	630	641	11	121
2	270	187	-83	6,889
6	510	440	-70	4,900
7	570	346	-224	50,176
1	210	238	28	784
4	390	312	-78	6,084
2	270	269	-1	1
9	690	655	-35	1,225
6	510	563	53	2,809
				$\Sigma = 97,667$

an example of manual calculation for residual

Figure: an example of computer calculation for regression

SUMMARY OUTPUT						
Regression Sta	atistics					
Multiple R	0.8325					
R Square	0.6931					
Adjusted R Square	0.6624					
Standard Error	92.1055					
Observations	12			1		
		SSE =	84834.2947			
ANOVA			,	ē .		
	df	SS	MS	F	Significance F	
Regression	1	191600.6220	191600.6	22.5853	0.0008	
Residual	10	84834.2947	8483.429			
Total	11	276434.9167				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	175.8288	54.9899	3.1975	0.0095	53.3037	298.3539
Years with Midwest	49.9101	10.5021	4.7524	0.0008	26.5100	73.3102

Estimated regression equation is $\hat{y} = 175.8288 = 49.9101(x)$

SIMPLE REGRESSION ANALYSIS- EXAMPLE (R-programming)

The investment firm Harmonic Investments wants to manage the pension fund of a major Chicago retailer. Correlation analysis have been done and it shows that the relationship between profits and number of employees for 50 Fortune 500 companies in the firm's portfolio are positive linear relationship. Next, the Harmonic wants to use simple linear regression analysis for their presentation to the retailer. The data for the analysis are contained in the file **Fortune 50**.

SOLUTION:

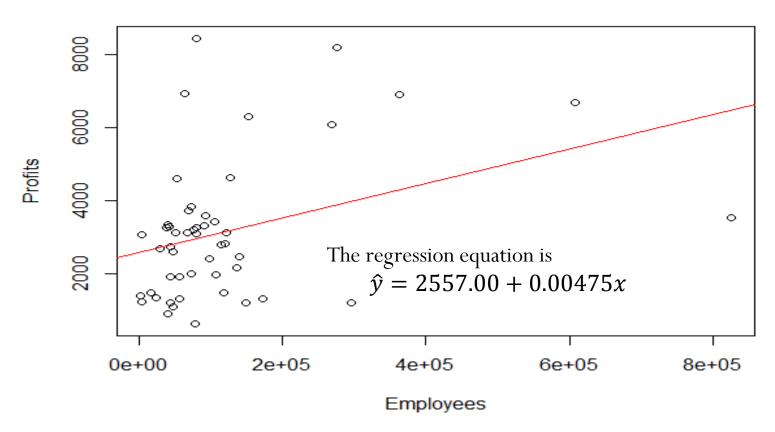
Identify independent (x) and dependent (y) variable

Identify the linear regression equation/model

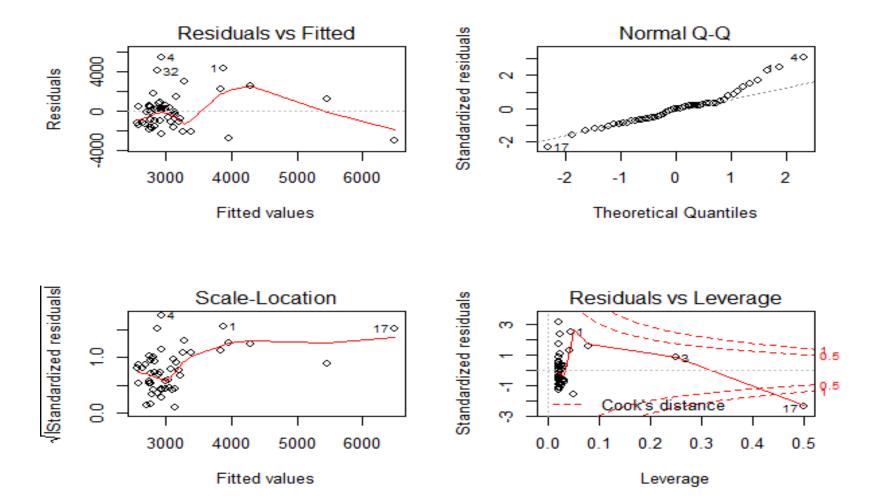
```
> lm.out = lm(Profits ~ Employees, data=Fortune50) # name the output
> 1m. out
call:
lm(formula = Profits ~ Employees, data = Fortune50)
                                              \rightarrow \beta_1 - slope
oefficients:
(Intercept)
               Employees
  2.557e+03
                4.759e-03
> summary(lm.out)
call:
lm(formula = Profits ~ Employees, data = Fortune 0)
Residuals:
    Min
             1Q Median
                              3Q
                                     Max
-2957.1 <u>-1176</u>.0
                   -38.0
                           566.8 5522.4
                                                          → regression
Coefficients:
                                                            coefficient
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.557e+03 /3.283e+02
                                    7.788 4.61e-10 ***
Employees
            4.759e-03 1.759e-03
                                    2.706 0.00941 **
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1794 on 48 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.1323, Adjusted R-squared: 0.1143
F-statistic: 7.321 on 1 and 48 DF, p-value: 0.009409
```

The regression equation is $\hat{y} = 2557.00 + 0.00475x$

Employees versus Profits (millions of dollars)



- > plot(Profits ~ Employees, data=Fortune50, main="Employees versus Profits (millions of dollars)")
- > abline(lm.out, col="red")



- > par(mfrow=c(2,2)) # partition the graphics device
- > plot(lm.out)

REGRESSION ANALYSIS PLOT

- Other graphic display of how good the model fit is can be achieved as follows:
 - Residual vs fitted Points that tend towards being outliers are labeled
 - Normal Q-Q to see our residuals normally distributed
 - Residual vs Leverage Labeled points on this plot represent cases we may want to investigate as possibly having undue influence on the regression relationship.

REGRESSION MODEL DIAGNOSIS

SIGNIFICANCE TEST IN REGRESSION ANALYSIS

- Why need significance test?:
 - Regression coefficient compute from sample data is point estimate. Thus, it subject to sampling error.
 - To diagnose the model.
- There are 3 equivalent significant test in regression analysis:
 - 1. Test for significance of the correlation between *x* and *y*. ---- cover in correlation analysis
 - 2. Test for significance of the coefficient of determination. ---- next slide
 - 3. Test for significance of the regression slope coefficient. ---next next slide ©

- Other diagnostic on the regression line:
 - Perform Residual Analysis on a Regression Model
 - 1. To determine whether a linear model is appropriate to describe the relation between the explanatory and response variables
 - 2. To check for outliers

- Identify influential observations
 - Influential observation → is an observation that significantly affects the least-squares regression line's slope and/or *y*-intercept, or the value of the correlation coefficient.

RESIDUAL ANALYSIS ON A REGRESSION MODEL

- Residual: Is the difference between the observed value of y and the predicted value, \hat{y} .
- Residuals play an important role in determining the adequacy of the linear model. → need to draw residual plot
- If a plot of the residuals against the explanatory variable shows a discernible pattern, such as a curve, then the explanatory and response variable may not be linearly related.

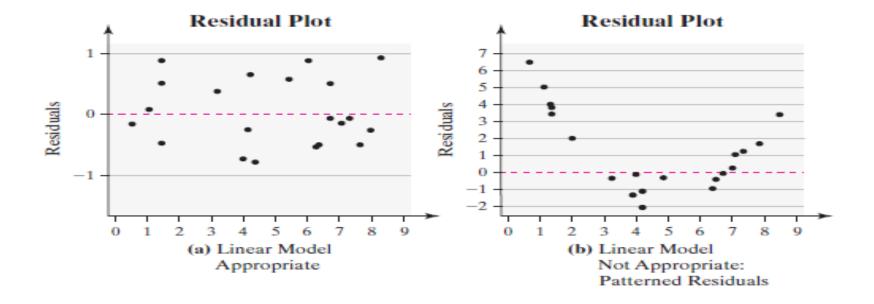


Figure (a) does not show any pattern, so a linear model is appropriate. However, the residual plot in Figure (b) shows a U-shaped pattern, which indicates that a linear model is inappropriate.

- Coefficient of Determination, R^2 :
 - The portion of the total variation in the dependent variable that is explained by its relationship with the independent variable.

Coefficient of Determination, R2

$$R^2 = \frac{SSR}{SST}$$

- R^2 value:
 - The value is between 0 and 1.0
 - R^2 with value 1.0 would respond to a situation in which the regression line would pass through each of the points in the scatter plot.
- How the decision makers use R^2 ?:
 - ullet to indicate how well the linear regression line fits the (x,y) data points.
 - the better the fit, the closer R^2 will be to 1.0
 - R^2 will be close to 0 when there is a weak linear relationship

- Others measurements in Regression Analysis:
 - SST (total sum of squares): used to measure variation in the dependent variable (y).
 - SSR (sum of square regression): used to measure variation attributed to the relationship between the x's and y's
 - SSE (sum square errors): used to measure the variation attributed to the error. It should be minimize

$$SST = SSR + SSE$$

- Hypothesis testing to determine whether the correlation of determination have value of > 0
- Hypothesis: using the ρ^2 (rho) notation

$$H_0: \rho^2 = 0$$

$$H_{A}: \rho^{2} > 0$$

• Test statistics for correlation of determination: using the F-test statistics and compare the critical value from F distribution table.

$$F = \frac{\frac{SSR}{1}}{\frac{SSE}{(n-2)}} \quad df = (D_1 = 1, D_2 = n-2)$$

where:

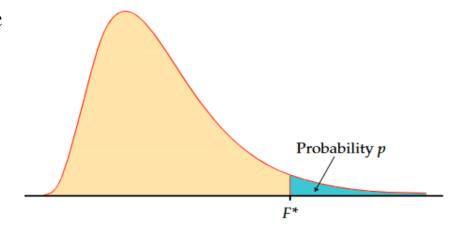
SSR = Sum of squares regression SSE = Sum of squares error

Test Statistic for Significance of the Coefficient of Determination

Decision rule: calculate a F value and compare it to the critical value, F in the f-distribution table.

Since H_A : $\rho^2 > 0$: Upper tail

- Figure if the $F \ge \text{critical}$ value, F_{α} therefore reject H_{α}
- Figure 1. For the F < critical value, F_{α} therefore accept H_{α}



p-value

if the $p \ge$ critical value, α therefore accept H_o if the p < critical value, α therefore reject H_o

```
> lm.out = lm(Profits ~ Employees, data=Fortune50) # name the output
> 1m.out
call:
lm(formula = Profits ~ Employees, data = Fortune50)
Coefficients:
(Intercept)
               Employees |
  2.557e+03
               4.759e-03
> summary(1m.out)
call:
lm(formula = Profits ~ Employees, data = Fortune50)
                                                       p < critical
Residuals:
             1Q Median
    Min
                             3Q
                                    Max
                                                       value, αtherefore
-2957.1 -1176.0 -38.0 566.8 5522.4
                                                       reject H<sub>o</sub>
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.557e+03 3.283e+02 7.788 4.61e-10 ***
Employees
            4.759e-03 1.759e-03 2.706 0.00941 **
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
Residual standard error: 1794 on 48 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.1323, Adjusted R-squared: 0.1143
F-statistic: 7.321 on 1 and 48 DF, p-value: 0.009409
          R^2 value
 correlation coefficient, r = \sqrt{R^2} = \sqrt{0.1323} = 0.3637
```

 $R^2 = r^2$

The F-value and the p-value for testing coefficient of determination, R^2

Decision: P-value $< \alpha = 0.05$, therefore H₀ is rejected

Conclusion:

We reject the null hypothesis and conclude the population coefficient of determination (ρ^2) is greater than zero. This mean the independent variable (x) explains a significant proportion of the variation in the dependent variable (y)

SIGNIFICANCE TEST OF THE REGRESSION SLOPE COEFFICIENT

- Why need to test the regression slope coefficient?
 - we are interested in determining whether the population regression slope coefficient is 0.
 - A slope of 0 would imply that there is no linear relationship between *x* and *y* variables and that the *x* variable
- Hypothesis null and alternative to be tested are:

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

Simple Linear Regression Test Statistic for Test of the Significance of the Slope

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \qquad df = n - 2$$

where:

 b_1 = Sample regression slope coefficient

 β_1 = Hypothesized slope (usually β_1 = 0)

 s_{b_1} = Estimator of the standard error of the slope

Simple Regression Estimator for the Standard Error of the Slope

$$s_{b_1} = \frac{s_{\varepsilon}}{\sqrt{\sum (x - \overline{x})^2}}$$

where:

 s_{b_1} = Estimate of the standard error of the least squares slope

$$s_{\varepsilon} = \sqrt{\frac{SSE}{n-2}}$$
 = Sample standard error of the estimate (the measure of deviation of the actual y-values around the regression line)

SIGNIFICANCE TEST OF THE REGRESSION SLOPE COEFFICIENT

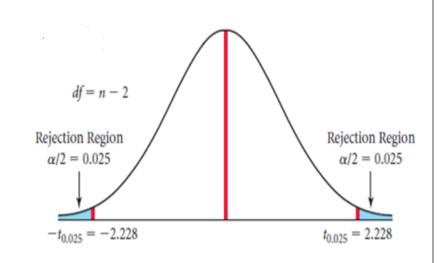
Decision rule: calculate a t value and compare it to the critical value, *t* in the t-distribution table. Since it involve with 2 tailed;

Lower tail

- Figure 1. Figure 1. Figure 1. Figure 2. Figure 1. Figur
- \geq if the -t < critical value , - $t_{lpha/2}$ therefore reject $H_{
 m o}$

Upper tail

- \succ if the $t \ge$ critical value, $t_{\alpha/2}$ therefore reject H_0
- Fig. if the t < critical value, $t_{\alpha/2}$ therefore accept H_0



p-value

if the $p \ge$ critical value, α therefore accept H_o if the p < critical value, α therefore reject H_o

```
> lm.out = lm(Profits ~ Employees, data=Fortune50) # name the output
  > 1m.out
  call:
  lm(formula = Profits ~ Employees, data = Fortune50)
  Coefficients:
  (Intercept)
                 Employees
    2.557e+03
                 4.759e-03
  > summary(1m.out)
  call:
  lm(formula = Profits ~ Employees, data = Fortune50)
  Residuals:
               1Q Median
      Min
                               3Q
                                      Max
  -2957.1 -1176.0 -38.0 566.8 5522.4
  Coefficients:
               Estimate Std. Error t value Pr(>|t|)
  (Intercept) 2.557e+03 3.283e+02
                                     7.788 4.61e-10 ***
              4.759e-03 (1.759e-03) 2.706 0.00941 **
  Employees
  Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '1
  Residual standard error: 1794 on 48 degrees of freedom
    (2 observations deleted due to missingness)
  Multiple R-squared: 0.1323, Adjusted R-squared:
                                                       0.1143
  F-statistic: 7.321 on 1 and 48 DF, p-value: 0.009409
                                          the calculated t-statistics and p-
standard error of the
                                          value for testing whether the
regression slope =
                                          regression slope = 0
0.00176
```

CONFIDENCE INTERVAL ESTIMATE FOR THE REGRESSION SLOPE

- Why need confidence interval estimate for β_1 ?
 - to describe the relationship between the independent variable and the dependent variable.

$$b_1 \pm t s_{b_1}$$
 or equivalently,
$$b_1 \pm t \frac{s_{\varepsilon}}{\sqrt{\sum (x-\overline{x})^2}} \qquad df = n-2$$
 where:
$$s_{b_1} = \text{Standard error of the regression slope coefficient}$$

$$s_{\varepsilon} = \text{Standard error of the estimate}$$

Confidence Interval Estimate for the Regression Slope, Simple Linear Regression

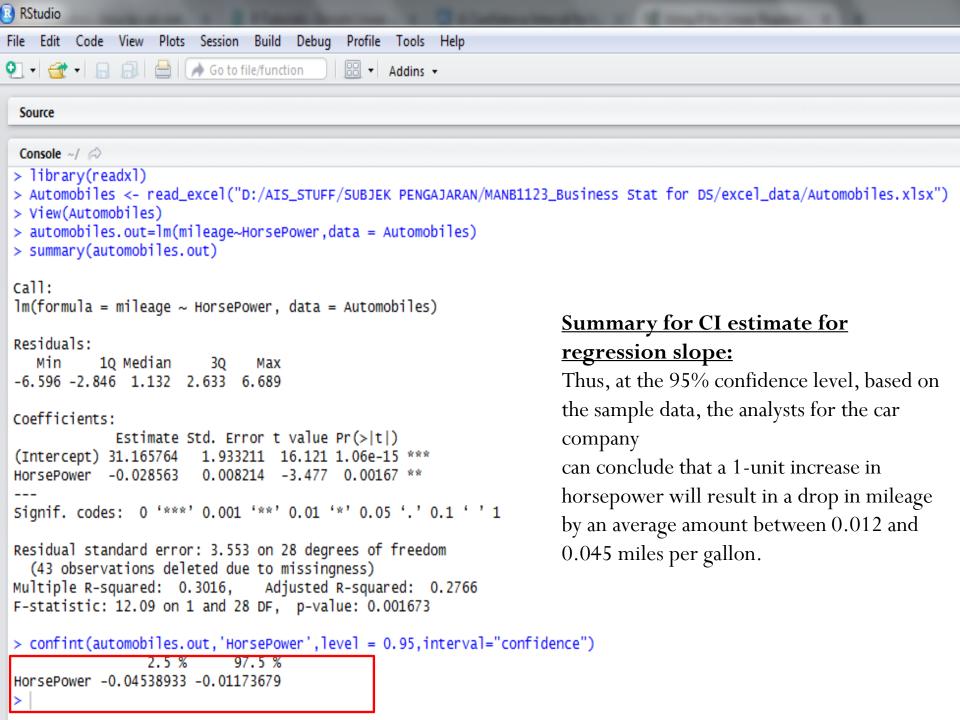
PROBLEM:

In the summer of 2006, gasoline prices soared to record levels in the United States, heightening motor vehicle customers' concern for fuel economy. Analysts at a major automobile company collected data on a variety of variables for a sample of 30 different cars and small trucks. Included among those data were the Environmental Protection Agency (EPA)'s highway mileage rating and the horsepower of each vehicle. The analysts were interested in the relationship between horsepower (x) and highway mileage (y). The data are contained in the file **Automobiles**.

Identify the following:

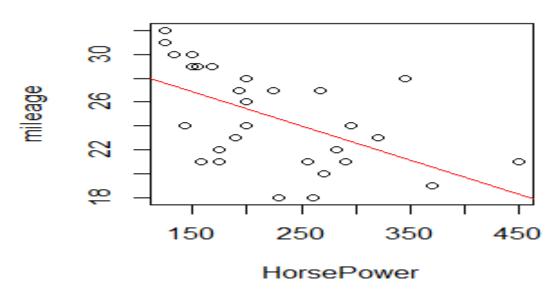
- 1. Correlation coefficient analysis
- 2. Regression coefficient
- 3. Regression equation model
- 4. Regression plot
- 5. Test for significance of the correlation between x and y at 95%
- 6. Test for significance of the coefficient of determination at 95%
- 7. Test for significance of the regression slope coefficient at 95%
- 8. Confidence interval estimate for the regression slope at 95%

```
RStudio
    Edit Code View Plots Session
                              Build Debug Profile Tools Help
🕘 🔻 🚰 🔻 🔒 📋 🌽 Go to file/function
                                         88 ▼ Addins ▼
  Source
 Console ~/ 🖒
 > library(readx1)
 > Automobiles <- read_excel("D:/AIS_STUFF/SUBJEK PENGAJARAN/MANB1123_Business Stat for DS/excel_data/Automobiles.xlsx")
 > View(Automobiles)
 > automobiles.out=lm(mileage~HorsePower,data = Automobiles)
 > summary(automobiles.out)
 call:
 lm(formula = mileage ~ HorsePower, data = Automobiles)
 Residuals:
   Min
           10 Median
                         3Q
                              Max
 -6.596 -2.846 1.132 2.633 6.689
 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 0.008214 -3.477 0.00167 **
 HorsePower -0.028563
 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 Residual standard error: 3.553 on 28 degrees of freedom
   (43 observations deleted due to missingness)
Multiple R-squared: 0.3016, Adjusted R-squared: 0.2766
F-statistic: 12.09 on 1 and 28 DF, p-value: 0.001673
 > confint(automobiles.out, 'HorsePower', level = 0.95, interval="confidence")
                 2.5 %
                            97.5 %
 HorsePower -0.04538933 -0.01173679
```



- > plot(mileage ~ HorsePower, data=Automobiles, main="Mileage versus Horse Power")
- > abline(automobiles.out, col="red")

Mileage versus Horse Power



```
Solution via R-programming
```

```
> with(Automobiles, cor.test(mileage, HorsePower))
        Pearson's product-moment correlation
data: mileage and HorsePower
t = -3.4772, df = 28, p-value = 0.001673
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7592287 -0.2354968
<u>sample estimates:</u>
      cor
-0.549173
```

THANK YOU

Do more exercise!!!!