MEIGUODAIXIECOM

BSB123 Data Analysis

QUESTION ONE

Remuneration for CEO's has been in the press lately. A recent study of CEO salaries was designed to identify the factors that might explain different salary level. The first factor considered was the size of the company in terms of annual sales.

Information was provided by twenty (20) CEO's on their annual salary ($000) and annual sales for their company ($million).

The table below gives the result of a regression analysis undertaken to examine the relationship between these two variables.

Salary ($000)	Sales ($M)
813	90.8
899	283.1
925	198.3
977	255.4
1002	382.2
1004	199.6
1018	266.7
1022	178.9
1038	160.2
1073	143
1208	311
1217	700.1
1228	411.4
1231	388.8
1240	385.7
1254	255.4
1254	155.6
1460	476.7
1531	703.4
1597	697.1

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.760
R Square	0.577
Adjusted R Square	0.553
Standard Error	140.767
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	486270.281	486270.3	24.540	0.0001
Residual	18	356675.577	19815.31		
Total	19	842945.858			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	867.537	65.044	13.338	0.000
Sales ($M)	0.849	0.171	4.954	0.000

a. Define the population model and state the assumptions underlying the model (2)

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Where

Y = CEO Salary (measured in $000)
X_1 = Sales ($M)
ε = Residual or error term

For simple regression there are 5 assumptions related to the error term:

$E(\varepsilon_i) = 0$

- Zero mean. Means we have an unbiased estimator has to be true for line of best fit.

$V(\varepsilon_i) = \sigma_\varepsilon^2$

- Constant variance

$Cov(\varepsilon_i, \varepsilon_j) = 0$

- Our error terms are independent of each other

$\varepsilon_i \sim N(0, \sigma_\varepsilon)$

- Errors are normally distributed. This allows us to carry out our various tests
- Errors are independent of the Independent variables

b. State the estimated equation and interpret the coefficients (2)

$$\hat{y} = 856.537 + 0.849x_1$$

b_0 = 856.537. This is the intercept which is the value of Y if X = 0. This implies for a company with $0 sales the CEO would earn on average $856,537. None of our sales data was close to 0 so this is an extrapolation outside of the range of the data which would have little meaning.
b_1 = 0.849 which is the change in Y for a one unit increase in X. Hence for every additional $1m in sales the CEO could expect their income to increase by $849 on average.

c. Test the significance of the relationship between salary and sales (2)

Testing the significance of an individual variable is the t-test

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

Decision Rule:

Reject Ho if p-value < α (which we will set at 0.05)

p-value = 0 for this regression, and hence we can reject the null hypothesis and conclude that there is a significant relationship between CEO salaries and sales.

Additional data was collected on three extra variables, the number of employees, total capital investment for the company ($million), and whether or not the company was primarily involved in manufacturing. The data and the associated regression output are given below.

d. To what extent does the model explain the variation in salaries. Which statistic did you use to find this value and why? (2)

For this question we use the R^2 Adjusted because we are now looking at multiple independent variables. The R^2 adjusted adjusts the coefficient of variation to take in to account variables that may add no explanatory power of the model – that is it adjusts for coincidental correlation.

$$\bar{R}^2 = 0.843$$

This implies that 84.3% of the variation in CEO salaries can be explained by variation in the four independent variables Sales, Number of Employees, Capital and type of industry (Manufacturing or not).

e. State the estimated equation and interpret all coefficients. (3)

$$\hat{y} = 822.364 + 0.154x_1 + 0.126x_2 + 0.266x_3 - 86.205x_4$$

b_0 = 822.364 implies for a company with $0 sales, no employees, no capital in a non-manufacturing industry the CEO would earn on average $822.364. While this again has little sense here, the value may be considered an estimate of the base salary obtained by all CEO's with real salary going up from there depending on the circumstances of the company.
b_1 = 0.154 which implies for every additional $1m in sales, CEO salaries would increase by $154 on average all other variables remaining constant
b_2 = 0.126 implies that for every additional employee in the company CEO salaries would increase by $126 on average all other variables remaining constant
b_3 = 0.266 implies that for every additional $1m in capital the company manages, the CEO salary would increase by $266 on average all other variables remaining constant
b_4 = -86.205 which implies CEO's in manufacturing industries earn, on average, $86,205 less than their counterparts in non-manufacturing industries.

f. Conduct all tests to determine the significance of the overall model and which of the independent variables are significant factors in explaining the variation in salaries. Do these results make sense? (7)

The first test in any multiple regression is the F test of overall significance:

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
$H_A: At\ least\ on\ \beta\ \neq 0$

Decision Rule:

Reject Ho if p – value < α (We will assume α = 0.05)

p-value from ANOVA table: p = 0

Hence we can reject the null hypothesis and we can say there is sufficient evidence to conclude that the variables sales, employees, capital and industry together have a significant impact in explaining the variation in Incomes.

Secondly we test each of the Individual Variables (using prior expectations). These are t-tests.

For all t-tests the decision rule is:

Reject Ho if the p-value < α (0.05)

Sales

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

p-value = 0.362/2 = 0.181

Hence we can not reject Ho and conclude that Sales is not a significant factor in explaining variations in CEO Salaries. This is the opposite of what we found in our earlier test. This may be because Sales is captured in other variables, or it may be we have a problem with the assumptions in the data.

<u>Number of Employees</u>

H_0: $\beta_2 = 0$
H_A: $\beta_2 > 0$ (we would expect CEO's to earn more the larger the company)

p-value = 0.

Hence we can reject Ho and conclude that the number of employees does significantly explains variation in CEO salaries.

Capital

H_0: $\beta_3 = 0$
H_A: $\beta_3 > 0$ (We would expect CEO's to earn more if they are managing more capital)

p-value = 0.157/2 = 0.0775

Hence we can not reject Ho and conclude that the amount of capital is not a significant factor in explaining the variation in CEO salaries.

Industry (Manufacturing = 1)

H_0: $\beta_4 = 0$
H_A: $\beta_4 \neq 0$ (This is done as a two tail test as I have no previous expectations of whether or not CEO's in Manufacturing earn more than others)

p-value = 0.063

Hence we can not reject Ho and conclude that there is no evidence that Manufacturing CEO's earn a different salary to counterparts in other industries.

These results do not make sense in that only one of the four variables was significant. Although that can happen, and the full explanatory power can come from one variable, the fact that Sales went from being significant to not significant is an indicator that something else might be happening here. For this reason we need to do some addition checks.

g. Consider the correlation matrix provided below. Does this raise any concerns with the above results? Explain. Are there any other checks that you would do? (2)

The correlation matrix raises many concerns.

1. There are four high correlations between independent variables: sales and employees r = 0.6922; sales and capital r – 0.684; Employees and Capital r = 0.568; employees and industry r=0.493. This implies that there is significant multicollinearity between all four variables within the regression which makes it difficult to determine the individual contribution of each variable.
2. The correlation coefficient between Industry (Manufacturing) and Salary is 0.1985. This implies a positive relationship, or, that manufacturing CEO's earn more. Our coefficient of this variable in the estimates, however, was negative which is a direct indication that the estimates are wrong which is no doubt down to the multicollinearity.

In addition to the correlation matrix test of multicollinearity we should also be looking at the residual plots to determine if any of the other assumptions about the error terms have been violated – in particular the

assumptions regarding the constant variance of the errors and the error terms being independent of each other (assumptions 2 and 3 from earlier).

h. What would be the first adjustment you would make to the model provided above and why? (2)

There are a number of adjustments which could be considered, however, in the first instance, since the number of employees is the one variable which is highly correlated with all others we should remove this from the analysis and re-estimate.

Salary ($000)	Sales ($M)	Employees (Number)	Capital ($m)	Manufacturing (1 = yes)
813	90.8	295	91	1
899	283.1	505	107.6	0
925	198.3	417	28.8	0
977	255.4	2182	10	0
1002	382.2	654	181.8	0
1004	199.6	1986	11.4	0
1018	266.7	2013	36.6	1
1022	178.9	1154	16	0
1038	160.2	849	45.2	0
1073	143	1765	94.6	1
1208	311	1984	82.8	0
1217	700.1	2989	140.6	1
1228	411.4	2875	73.4	0
1231	388.8	2986	134.2	1
1240	385.7	2299	454.2	0
1254	255.4	3432	141.6	1
1254	155.6	2375	91.4	0
1460	476.7	4417	195.4	1
1531	703.4	4863	576.8	1
1597	697.1	3300	322.2	0

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.936
R Square	0.876
Adjusted R Square	0.843
Standard Error	83.459
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	738464.941	184616.235	26.505	0.000
Residual	15	104480.917	6965.394		
Total	19	842945.858			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	822.364	41.590	19.773	0.000
Sales ($M)	0.154	0.163	0.944	0.360
Employees (Number)	0.126	0.023	5.489	0.000
Capital ($m)	0.266	0.179	1.489	0.157
Manufacturing (1 = yes)	-86.205	42.917	-2.009	0.063

	Salary ($000)	*Sales ($M)*	*Employees (Number)*	*Capital ($m)*	*Manufacturing (1 = yes)*
Salary ($000)	1				
Sales ($M)	0.7595	1			
Employees (Number)	0.8848	0.6922	1		
Capital ($m)	0.6818	0.6840	0.5680	1	
Manufacturing (1 = yes)	0.1985	0.2042	0.4393	0.1939	1

QUESTION TWO

Considering the information given in Question One suppose we are interested in whether or not there is a significant difference in the salaries of CEO's for Manufacturing and Non-Manufacturing companies. The following table includes some summary statistics on the salaries of the 20 CEO's surveyed.

	Manufacturing	Non Manufacturing
Average Salary	1199.41	1116.24
Standard deviation	233.06	197.605
Sample size	8	12

a. Conduct a **test** at the 5**% level of significance** to determine if there is a **difference in the average** salaries of CEO's of Manufacturing and non-manufacturing companies. (5)

Test with a 5% level of significance implies a hypothesis test. Difference in average (not more or less) implies a two tail test. Difference in average also indicates clearly this is a comparison between the two populations so the parameter being estimated is $\mu_A - \mu_B$

A – Manufacturing
B – Non Manufacturing

State Hypotheses

H_0: $\mu_A - \mu_B = 0$
H_A: $\mu_A - \mu_B \neq 0$

Check for Normality

In this problem both sample sizes are < 30 and we are not told that the populations from which the data is selected is normal. Hence it is necessary for us to make the assumption that both samples were selected from normally distributed populations so that the sampling distributions of each are normal, and the sampling distribution of the difference in means is normally distributed.

Decision Rule:

Reject Null Hypothesis if:

$$t_{calc} < -t_{(v,\alpha/2)} \text{ OR } t_{calc} > t_{(v,\alpha/2)}$$

$\alpha = 0.05$ (given)

To determine the number of degrees of significance, v, we need to determine if the variances are equal or not.

Using the rule of thumb:

$\frac{Larger\ Variance}{Smaller\ Variance} = \frac{233.06^2}{197.605^2} = 1.39$ and hence we can assume the population variances are equal.

Using that, the degrees of freedom:

$v = n_A + n_B - 2 = 8 + 12 - 2 = 18$

Using tables this will give a students t score $t_{(18,0.025)} = 2.101$

Hence our decision rule becomes:

Reject Null Hypothesis if:

$t_{calc} < -2.101$ OR $t_{calc} > 2.101$

Calculations

$$t_{calc} = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{s_{(\bar{x}_A - \bar{x}_B)}}$$

where

$$s_{(\bar{x}_A - \bar{x}_B)} = \sqrt{s^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

Where

$$s^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}$$

$$s^2 = \frac{(7)233.06^2 + (11)197.605^2}{18} = 44985.77$$

$$s_{(\bar{x}_A - \bar{x}_B)} = \sqrt{(44985.77)\left(\frac{1}{8} + \frac{1}{12}\right)} = 96.809$$

$$t_{calc} = \frac{(1199.41 - 1116.24) - (0)}{96.809} = 0.859$$

Make the Decision

Since the $t_{calc} = 0.859$ intside the critical values of ±2.101 we can not reject the null hypothesis and conclude there is insufficient evidence to say there is a difference in CEO salaries between manufacturing and non manufacturing companies.

b. How do these results compare to those in Question 1. Comment. (2)

In one way the results are similar in that for both tests we found there was no difference in the salaries of CEO's in the different industry groups. However the results are different in that for Q1 the coefficient of the Manufacturing dummy variable implied that Manufacturing CEO's earned less than their counterparts, but in this test we saw the average salary for manufacturing CEO's was more (1199 to 1116). This supports what we found with the correlation matrix.