**Multiple Regression**

**Tutorial Questions**

Data for these questions is provided in the Excel File **Multiple** **Regression Tutorial Questions.xlsx**

**Question 1**

The Equal Opportunity Commission is investigating questions around unequal pay rates and discriminatory remuneration in various industries. The Pay Equity tab in the excel workbook contains information on 100 employees from a particular industry. Information includes:

* Salary ($)
* Gender
* Years of Education
* Years of Experience
* Division of employment
* Age

An example for the first 10 employees is shown in the table below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Employee ID** | **Salary (in $)** | **Gender** | **Education (years)** | **Experience (years)** | **Division** | **Age** |
| 1 | 20860 | M | 11 | 4 | Marketing | 28 |
| 2 | 30200 | F | 16 | 1 | Marketing | 26 |
| 3 | 31240 | F | 12 | 1 | Operational | 25 |
| 4 | 36860 | M | 12 | 8 | Finance | 29 |
| 5 | 44760 | F | 14 | 4 | Operational | 27 |
| 6 | 46690 | F | 19 | 3 | Operational | 29 |
| 7 | 47400 | F | 15 | 1 | Marketing | 24 |
| 8 | 47880 | F | 14 | 3 | Marketing | 27 |
| 9 | 50620 | F | 13 | 6 | Finance | 30 |
| 10 | 50690 | F | 14 | 4 | Marketing | 29 |

1. Run a multiple regression analysis looking at the relationship between salary and years of education and years of experience.
2. What proportion of variation in salary can be explained by these two variables?
3. Conduct a test of the overall significance of the model.
4. Test both the Education and Experience variables separately. Do both contribute to explaining the variation in salaries?
5. Write out the estimated equation and interpret all coefficients.



For this question we are estimating: $Y=β\_{0}+β\_{1}X\_{1}+β\_{2}X\_{2}+ε$

i The proportion of variation in Y which can be explained by the independent variables is the Coefficient of Variation. For multiple regression we use the Adjusted R2 which is 0.5743. Hence we say that 57.43% of the variation in salary is explained by educational and work experience.

ii. Test of Overall Significance is the F Test

$$H\_{0}: β\_{1}=β\_{2}=0$$

$$H\_{1}:At least one β\ne 0$$

Decision Rule: Reject Null Hypothesis if p-value < α (which we will assume to be 0.05)

As this is an F test the p-value comes from the ANOVA table and:

P = 3.83E-19 which is zero

Hence we can reject Ho and conclude that the overall model is significant and the two variables education and work experience do significantly explain salary differences.

iii

Testing each of the variables separately are the t-tests.

Education Experience

$$H\_{0}: β\_{1}=0$$

$$H\_{1}:β\_{1}>0$$

We do this as an upper one tail test as we would expect that more education would lead to a higher wage.

Decision Rule: Reject Null Hypothesis if p-value < α

As this is a t test the p-values come from the coefficients table.

p-value = 0.0001/2 = 0.00005 (We divide by 2 as this is a one tail test and p-values in this table are given for two tail)

Hence we can reject the null hypothesis and conclude that educational experience significantly explains variation in salary.

Work Experience

$$H\_{0}: β\_{2}=0$$

$$H\_{1}:β\_{2}>0$$

We do this as an upper one tail test as we would expect that more experience would lead to a higher wage.

Decision Rule: Reject Null Hypothesis if p-value < α

p-value = 0.0000/2 = 0

Hence we can reject the null hypothesis and conclude that work experience significantly explains variation in salary.

iv. $\hat{y}=-13353.69+4494.98x\_{1}+2460.90x\_{2}$

The intercept term, b0 = -13353.69 implies that for a worker with no education and no work experience they would pay the employer $13,353 to work. This is not a sensible outcome, and is driven by the fact that we are extrapolating beyond the range of the data. Minimum education in the data was 11 years.

b1 = 4494.98 implies that for every additional year of education employees can expect their salary to increase by $4494.98 on average. This is in line with expectations.

b2 = 2460.90 implies that for every extra year of work experience employees would expect their salary to increase by $2460.90 on average. Again this is in line with expectations of a positive relationship.

1. In order to answer the question of pay equity, include the variable Gender. Use Male as the reference point i.e. M = 0.
2. To what extent has including Gender increased the explanatory power of the model?
3. Conduct a test on each of the three variables to see if each contributes to the explanatory power of the model.
4. Based on the results of the tests does this mean that there is no difference in salaries between males and females? Is there another test you could suggest to determine if there is a difference?



1. The explanatory power of the model increased from $\overbar{R}^{2}$=0.5743 to 0.5773 or an increase of only 0.003 or 0.3%. This is minimal. It seems that adding Gender has not increased the explanatory power or the model to any large extent.

ASIDE: IT is worth noting here that the p-value for the F test has increased and the Fcalc itself has decreased. While the overall significance of the model has not changed, both of these changes indicate less evidence for rejecting the null hypothesis. The implication here is that adding the additional variable may have slightly increased the explanatory power, but this increase is so small it is overwhelmed by the fact that we have included an additional variable which would always increase R2 even without having any real importance.

Testing each of the variables separately are the t-tests.

Education Experience

$$H\_{0}: β\_{1}=0$$

$$H\_{1}:β\_{1}>0$$

We do this as an upper one tail test as we would expect that more education would lead to a higher wage.

Decision Rule: Reject Null Hypothesis if p-value < α

p-value = 0.0001/2 = 0.00005

Hence we can reject the null hypothesis and conclude that educational experience significantly explains variation in salary. This is the same as before adding Gender

Work Experience

$$H\_{0}: β\_{2}=0$$

$$H\_{1}:β\_{2}>0$$

We do this as an upper one tail test as we would expect that more experience would lead to a higher wage.

Decision Rule: Reject Null Hypothesis if p-value < α

p-value = 0.0000/2 = 0

Hence we can reject the null hypothesis and conclude that work experience significantly explains variation in salary.

Gender

$$H\_{0}: β\_{3}=0$$

$$H\_{1}:β\_{3}<0$$

We do this as a lower one tail test as we are interested in whether or not females earn less.

Decision Rule: Reject Null Hypothesis if p-value < α

p-value = 0.1954/2 = 0.0972

Hence we do not reject the null hypothesis and conclude that the variable gender is not a significant factor in explaining wages.

Based on this result it appears that Gender has no impact on salaries, but in fact that is not necessarily the fact. Instead, this result says that if we include the factors education and work experience, then gender has no impact.

A quick look at salaries for Males and Females shows an average for Males of $97645 and for Females of $77550 which does indicate a difference. A two population Hypothesis Test comparing the average for Males and Females:

Ho: µF ≥ µM

HA: µF < µM

Decision Rule: Reject Null Hypothesis if p-value < α

Here the p-value for one tail is 0.000248

We can reject Ho and conclude Females earn less on average.

The factor at work here is experience. The common explanation for women earning less is that they often spend long periods out of the workforce to take on the duties of raising families, so it would be worthwhile seeing if there is any evidence that females have less experience than males.

If we conduct a two population Hypothesis Test of the average number of years of experience for Males and Females we get:



Ho: µF ≥ µM

H1: µF < µM

Decision Rule: Reject Null Hypothesis if p-value < α

Here the p-value for one tail is 0.000027

We can reject Ho and conclude Females do have less work experience than males.

Given that the regression equation took in to account the experience factor which was considered to be significant, incorporating both variables is not necessary.

Note that this makes no comment on the equity question itself, it simply indicates that there is evidence that the experience women lose is the important factor in explaining the disparity in income equality. In fact this is probably good evidence for why there should be additional support and recognition for taking on caring duties.

1. Include the age variable to the analysis and create a correlation matrix for all four variables. Does age appear to have a significant correlation with Salary? Is it likely to increase the explanatory power of the model?



Age has a correlation of 0.6780 with salary indicating that it could have a significant impact on explaining salary. However, like the example of Gender and Experience (which has a correlation coefficient of 0.39) Age is also significantly correlated with Experience (correlation coefficient of 0.9785). This is likely to have two impacts – firstly that it probably doesn’t add anything above experience, and there is a very high correlation between two independent variables violating one of the key assumptions of the regression model.

1. Run the regression adding in Age.
2. What is the change in the Adjusted R2? Has age added anything?
3. Conduct a t test each of the coefficients to determine if they are significant interpret the coefficients.
4. Are the results of the last two questions consistent with part c.? What is the likely issue?



$\overbar{R}^{2}$= 0.5774 which is basically no change on the previous regression, hence adding age has not increased the explanatory power.

Conduction the tests for each variable we use the common Decision Rule: Reject Ho if p-value < α

Education Experience

$$H\_{0}: β\_{1}=0$$

$H\_{1}:β\_{1}>0$

p-value = 0, hence we reject Ho and Education is still considered significant. b1 = 4427.199 implies that for every additional year of education, salary will increase on average by $4427.20

Work Experience

$$H\_{0}: β\_{2}=0$$

$$H\_{1}:β\_{2}>0$$

p-value = 0.003/2 = 0.0015. Hence we can reject the null hypothesis and conclude that work experience significantly explains variation in salary. b2 = 3466.82 implies that for every additional year of work experience salaries will increase on average by $3466.82.

Gender

$$H\_{0}: β\_{3}=0$$

$$H\_{1}:β\_{3}<0$$

p-value = 0.204/2 = 0.102. Hence we do not reject the null hypothesis and conclude that the variable gender is not a significant factor in explaining wages. b3 = -5353.44 which implies that females earn $5353.44 less than males on average.

Age

$$H\_{0}: β\_{4}=0$$

$H\_{1}:β\_{4}>0$ We would do this as an upper tail test as peoples incomes usually increase as they get older

p-value = 0.313/2 = 0.155. Hence we do not reject the null hypothesis and conclude that the variable age is not a significant factor in explaining wages.

This last outcome is not unexpected because we already saw that age and experience were strongly correlated. Also, there was no increase in the adjusted R2.

b4 = -1107.271 which implies that for every additional year of age, salary would decrease by $1107.27. This result makes no theoretical sense, at least not for people of normal working age. Also, if you look at the correlation between Age and Salary in the correlation matrix it was a positive value r = 0.678 which indicates a positive relationship. This is one of the key problems with Multicollinearity, that the individual impacts of two highly correlated variables are indistinguishable and hence could be badly estimated for both.

1. Run the multiple regression adding in the variables for Division but excluding Age. For Division you will need two Dummy Variables. For the first make Finance = 1 with Marketing and Operations = 0. For the second make Marketing = 1 and the other two = 0. This can all be done in the one regression. Write out the estimated equation. Comment on the strength of the relationship and the significance of each of the independent variables (just interpret p-value). Interpret each of the coefficients of the dummy variables.



The $\overbar{R}^{2}$ has increased slightly to 0.5924 indicating an increase of explanatory power of 1.5%. This is not big, but given we added two variables it may be that one is significant while the other isn’t. This is backed up by the individual t tests of the independent variables. Comparing the p-values to α=0.05 provides us with the result that Education (p=0.00005), Experience (p=0) and Marketing dummy variable (p=0.0307) are all significant variables. Gender (p=0.1132) and Finance (p=0.94) are not.

b4 = 382.9012 which implies that employees in the finance division earn on average $382.90 more than those in operations on average.

b5 = -9358.89 implies that employees in the marketing division earn on average $9,350.89 less than employees in the operational division.

Both are compared to operations as this is the base variable of 0 in all instances.

1. Run the regression again only this time make Operations = 1 for the first dummy variable and marketing = 1 for the second. What do you notice about the results? Does this change the interpretation of the coefficients.



Because we are using the same variables as before we see no change in all of the interpretive statistics - $\overbar{R}^{2}$, F values, t-statistics and p-values. We also see no change in the coefficient estimates for Gender, Education and experience.

The only difference that occurs is in the coefficients of the two dummy variables. Having swapped Finance and Operations as the base variable for one of the Dummy’s we now have to compare the Dummy Variable Coefficients to Finance. Whereas previously the result showed finance employees earned $382.90 more on average than operations, we now see that Operations employees earn $382.90 less than finance. The same result expressed as the opposite.

In relation to Marketing, we previously saw that they earned $9350.89 less than employees in the operational division. When we now compare them to finance we see that they earn $9733.79 less than employees in finance on average. This change incorporates the extra $382.90 that is the difference between the Finance and Operations divisions.

Taking all of this in to account, and considering the individual p-values, the conclusion is that there is no difference in average wages between the Finance and Operational divisions, however there is a significant difference (decrease) in wages for employees in the Marketing division.

1. Exclude all of the non-significant variables and run the regression again. Do all tests, write out the estimated equation and interpret all values now that we have our final model. What salary would you expect for a 38 year old female with 20 years education and 10 years work experience in the marketing division?

Excluding Gender, wages and the dummy variable related to Finance and Operations, we are now going to estimate:

$$Y=β\_{0}+β\_{1}X\_{1}+β\_{2}X\_{2}+β\_{3}X\_{3}+ε$$

Where:

Y = Salary ($)

X1 = Years of education

X2 = Years of work experience

X3 = Dummy Variable for Marketing Division

Running this regression we get:



$\overbar{R}^{2}$ = 0.5944 which implies that 59.44% of the variation in salaries can be explained by variation in education and work experience and the difference of working in the Marketing division. Note that the adjusted value is now closer to the overall R2 having cleaned out the non-significant variables.

Test of Overall Significance – F Test

$$H\_{0}: β\_{1}=β\_{2}=β\_{3}=0$$

$$H\_{1}:At least one β\ne 0$$

Decision Rule: Reject Null Hypothesis if p-value < α

P = 2.18E-19 which is zero

Hence we can reject Ho and conclude that the overall model is significant and the three variables together do significantly explain salary differences.

Testing each of the variables separately are the t-tests.

Education Experience

$$H\_{0}: β\_{1}=0$$

$$H\_{1}:β\_{1}>0$$

Decision Rule: Reject Null Hypothesis if p-value < α

p-value = 0.0002/2 = 0.0001

Hence we can reject the null hypothesis and conclude that educational experience significantly explains variation in salary.

Work Experience

$$H\_{0}: β\_{2}=0$$

$$H\_{1}:β\_{2}>0$$

Decision Rule: Reject Null Hypothesis if p-value < α

p-value = 0.0000/2 = 0

Hence we can reject the null hypothesis and conclude that work experience significantly explains variation in salary.

Dummy Variable – Marketing Division

$$H\_{0}: β\_{3}=0$$

$$H\_{1}:β\_{3}\ne 0$$

Note that this test was done as a two tail test as we potentially have no prior information or expectations about which divisions may or may not pay more.

Decision Rule: Reject Null Hypothesis if p-value < α

p-value = 0.0178

Hence we can again reject the null hypothesis and conclude that there is a significant impact on average salary for working in the finance division.

In conclusion we can say that each of the three variables included individually contribute to explaining the variation in salaries.

Looking at our estimated Equation

$$\hat{y}=-6803.7+4253.13x\_{1}+2468.63x\_{2}-9646.17x\_{3}$$

The intercept term, b0 = -6803.7 implies that for a worker with no education and no work experience they would pay the employer $6,803.70 to work. This is not a sensible outcome, and is driven by the fact that we are extrapolating beyond the range of the data.

b1 = 4253.13 implies that for every additional year of education employees can expect their salary to increase by $4,253.13 on average (other variables remaining constant). This is in line with expectations.

b2 = 2468.63 implies that for every extra year of work experience employees would expect their salary to increase by $2,468.63 on average. Again this is in line with expectations of a positive relationship.

b3 = -9646.17 implies that employees in the marketing division earn $9646.17 on average less than other employees in the organisation.

Testing our Assumptions

Multiple regression is built on the six assumptions:

* $E\left(ε\_{i}\right)=0$
	+ Zero mean. Means we have an unbiased estimator has to be true for line of best fit.
* $V\left(ε\_{i}\right)=σ\_{ε}^{2}$
	+ Constant variance – if not the data may explain some portion of dependent variable better than others – not consistent
* $Cov\left(ε\_{i}, ε\_{j}\right)=0$
	+ Our observations are independent. If not there may be an underlying pattern affecting data in other ways – e.g. seasonality
* $ε\_{i}\~N\left(0, σ\_{ε}\right)$
	+ Errors are normally distributed. This allows us to carry out our various tests
* Errors are independent of the Independent variables
* Independent variables are independent of each other

We have already considered the issues associated with the last assumption through the earlier parts of this analysis. To test the other assumptions we need to look at our residual plots.



* The dummy variable plot does not tell us very much as it is only at two points on the x axis, so we will ignore that.
* The residual plot for Experience has a few outliers but excluding them, the variance of the majority of the results is fairly constant and there appears to be no consistent patterns.
* The residual plot for Education is similar. Taking out the 4 outliers the remaining residuals do not appear to be exhibiting any patters. There is a possibility that the variance for the larger number of years of education is slightly larger than at the very small end, however the difference doesn’t appear great. It may be worth looking at further tests on this variable, but for the moment we can consider the assumptions to have been met.

Prediction

What salary would you expect for a 38 year old female with 20 years education and 10 years work experience in the marketing division?

Having excluded Gender and Age, the only information we need to incorporate here is the information on work experience (x1 = 10), education experience (x2 = 20) plus the fact that the person is working in Marketing (x3 = 1).

$$\hat{y}=-6803.7+4253.13x\_{1}+2468.63x\_{2}-9646.17x\_{3}$$

$$\hat{y}=-6803.7+4253.13(10)+2468.63(20)-9646.17(1)$$

$$\hat{y}=75,454.03$$

Question 2

Sales of single-family houses have been brisk in Big City this year. The worksheet: **Big** **City** contains data on 128 recent sales in Big City. For each sale, the file indicates the neighbourhood (Darra, Lota or Windsor) in which the sold house is located, the area in square metres, whether the house is made of brick, the number of bedrooms, the number of bathrooms, and the selling price. Darra and Lota are suburbs more than 12 km away from the CBD while Windsor is closer to the CBD, and a more prestigious neighbourhood. As a property investment consultant, you have been commissioned by an investor from Big City to explore the data and address the following questions. A portion of the data is presented below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Home** | **Nbhd** | **Size** | **Brick** | **Bedrooms** | **Bathrooms** | **Price** |
| 1 | Lota | 179 | No | 2 | 2 | 571500 |
| 2 | Lota | 203 | No | 4 | 2 | 571000 |
| 3 | Lota | 174 | No | 3 | 2 | 574000 |
| 4 | Lota | 198 | No | 3 | 2 | 473500 |
| … | … | … | … | … | … | … |

1. It is commonly believed that brick houses are more expensive than non-brick houses to build and hence more expensive. Based on this data, conduct a two sample t-test to determine if this belief is true.

Let Population 1 = Brick

Let Population 2 = Non-Brick

We want to test if the average price of Brick homes is more than Non Brick, i.e. µA > µB, or µA - µB > 0. This is what we are testing so it is the alternative hypothesis.

H0: µ1 ≤ µ2 or rather µ1 - µ2 ≤ 0

H1: µ1 > µ2 or rather µ1 - µ2 > 0

This is a one tail test using a level of significance of where we will assume α = 0.05. The test will be conducted using a t distribution because the standard deviations come from the two samples. Because the sample of Brick homes nA = 42 and the sample of non-brick homes nB = 86 are both greater than 30 then we can assume a normal distribution for the sampling distribution of the difference in means.

Decision Rule:

Reject H0 if:

tcalc > t(v, α)

In this instance the decision rule depends on the number of degrees of freedom which depends upon whether or not we believe the population variances are equal or not.

Using the rule of thumb:

If ratio of $\frac{Larger Variance}{Smaller Variance }<3$ we can assume the are approximately equal

Calculating the descriptive statistics using Excel we get:

|  |  |
| --- | --- |
| *Price Non Brick* | *Price Brick* |
|   |   |   |   |
| Mean | 609790.7 | Mean | 738845.2 |
| Standard Error | 12179.31 | Standard Error | 20699.32 |
| Median | 588250 | Median | 737500 |
| Mode | 589000 | Mode | #N/A |
| Standard Deviation | 112946.3 | Standard Deviation | 134146.9 |
| Sample Variance | 1.28E+10 | Sample Variance | 1.8E+10 |
| Kurtosis | -0.19413 | Kurtosis | -0.58821 |
| Skewness | 0.297374 | Skewness | 0.347585 |
| Range | 559000 | Range | 525500 |
| Minimum | 345500 | Minimum | 530500 |
| Maximum | 904500 | Maximum | 1056000 |
| Sum | 52442000 | Sum | 31031500 |
| Count | 86 | Count | 42 |

$\frac{Larger Variance}{Smaller Variance }=\frac{1.8E+10}{1.28E+10}=1.41$ which is < 3 and hence we assume the population variances are equal (σ2A = σ2B) and so our degrees of freedom are v = n1 + n2 – 2 = 42 + 86 – 2 = 126

and so our decision rule becomes:

Reject H0 if:

tcalc > t(12, 0.05)

tcalc > 1.645

$$t\_{calc}=\frac{(\overbar{x}\_{1}-\overbar{x}\_{2})-\left(μ\_{1}-μ\_{2}\right)}{s\_{(\overbar{x}\_{1}-\overbar{x}\_{2})}}$$

where

$$s\_{(\overbar{x}\_{1}-\overbar{x}\_{2})}=\sqrt{s^{2}\left(\frac{1}{n\_{1}}+\frac{1}{n\_{2}}\right)}$$

Where

$$s^{2}=\frac{\left(n\_{1}-1\right)S\_{1}^{2}+\left(n\_{2}-1\right)S\_{2}^{2}}{n\_{1}+n\_{2}-2}$$

$$s^{2}=\frac{\left(41\right)(17995396196)+\left(85\right)(12756867442)}{126}=14461468054.8$$

$$s\_{(1-\overbar{x}\_{2})}=\sqrt{s^{2}\left(\frac{1}{n\_{1}}+\frac{1}{n\_{2}}\right)}=\sqrt{14461468054.8\left(\frac{1}{41}+\frac{1}{85}\right)}=22865.99$$

$$t\_{calc}=\frac{(738845.2-609790.7)-\left(0\right)}{22865.99}=5.64$$

tcalc = 5.64 is greater than the critical value of 1.6445

Hence we reject the null hypothesis and conclude there is sufficient evidence at the 5% level of significance to conclude brick houses do cost more on average than non-brick houses.

This is backed up by the Excel test of differences between means assuming equal variances

|  |
| --- |
| t-Test: Two-Sample Assuming Equal Variances |
|   |   |   |
|  | *Price Non Brick* | *Price Brick* |
| Mean | 609790.6977 | 738845.2381 |
| Variance | 12756867442 | 17995396196 |
| Observations | 86 | 42 |
| Pooled Variance | 14461468068 |   |
| Hypothesized Mean Difference | 0 |   |
| df | 126 |   |
| t Stat | -5.700802338 |   |
| P(T<=t) one-tail | 4.01138E-08 |   |
| t Critical one-tail | 1.657036982 |   |
| P(T<=t) two-tail | 8.02276E-08 |   |
| t Critical two-tail | 1.978970602 |   |

1. Conduct a regression analysis to investigate the factors influencing house prices in Big City, and answer the below questions. For the Neighbourhood variable let Windsor = 1 to test if there is a premium for living in a prestigious suburb.
2. Assess the overall model adequacy.
3. What is the intercept coefficient? Is this value interpretable?
4. Is there a ‘premium’ for a brick house, everything else being equal? Is the result consistent with your answer in part (a)? Discuss.
5. Interpret the coefficients of the Nbhd dummy variables. Discuss its statistical significance.
6. Use the model you have developed to predict how much the investor should be paying for a brick house of 180 square metres in size, with 3 bedrooms and 2 bathrooms in Windsor.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |   |   |   |   |   |
|   |   |   |   |   |   |
| *Regression Statistics* |   |   |   |   |
| Multiple R | 0.8944 |   |   |   |   |
| R Square | 0.7999 |   |   |   |   |
| Adjusted R Square | 0.7917 |   |   |   |   |
| Standard Error | 61317.84 |   |   |   |   |
| Observations | 128 |   |   |   |   |
|   |   |   |   |   |   |
| ANOVA |   |   |   |   |   |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 5 | 1.83343E+12 | 3.66685E+11 | 97.52581 | 6.36313E-41 |
| Residual | 122 | 4.58705E+11 | 3759877161 |   |   |
| Total | 127 | 2.29213E+12 |   |   |   |
|   |   |   |   |   |   |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |   |
| Intercept | 81870.5286 | 52659.1448 | 1.5547 | 0.1226 |   |
| Size | 1855.5273 | 321.3701 | 5.7738 | 0.0000 |   |
| Brick | 97430.7803 | 11769.3416 | 8.2784 | 0.0000 |   |
| Bedrooms | 11402.4126 | 9536.9933 | 1.1956 | 0.2342 |   |
| Bathrooms | 34861.0583 | 12922.3567 | 2.6977 | 0.0080 |   |
| Nbhd | 155230.0014 | 13494.2313 | 11.5034 | 0.0000 |   |

Overall Adequacy of the Model

i.

$\overbar{R}^{2}$ = 0.7917 which implies that 79.17% of the variation in price can be explained by variation in the independent variables, size, building material, number of bedrooms, number of bathrooms and neighborhood. This fairly high, indicating that only 20.83% is unknown or due random variation.

Test of Overall Significance – F Test

$$H\_{0}: β\_{1}=β\_{2}=β\_{3}=β\_{4}=β\_{5}=0$$

$$H\_{1}:At least one β\ne 0$$

Decision Rule: Reject Null Hypothesis if p-value < α

P = is zero

Hence we can reject Ho and conclude that the overall model is significant and the five variables together do significantly explain salary differences.

Having tested the overall adequacy of the model we should also now test each of the individual components to see if they are significant. These are the t tests of each of the five slope coefficients which are summarized in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Hypothesis | P-value (use a=0.05) | Decision |
| Size | $H\_{0}: β\_{1}=0$ $$H\_{1}: β\_{1}>0$$ | 0 | Reject Ho, Size of the property is significant  |
| Brick | $H\_{0}: β\_{2}=0$ $$H\_{1}: β\_{2}>0$$ | 0 | Reject Ho, There is a significant difference in the price of brick v non brick houses |
| Bedrooms | $H\_{0}: β\_{3}=0$ $$H\_{1}: β\_{3}>0$$ | 0.2342/2 = 0.1171 | Do not Reject Ho, which indicates the number of bedrooms is not significantNote this is a strange result as the number of bedrooms is usually considered important and may be worth further consideration |
| Bathrooms | $H\_{0}: β\_{4}=0$ $$H\_{1}: β\_{4}>0$$ | 0..008/2 = 0.004 | Reject Ho and the number of bathrooms is a significant factor |
| Neighbourhood | $H\_{0}: β\_{5}=0$ $$H\_{1}: β\_{5}>0$$ | 0 | Reject Ho and conclude that there is a significant difference in price for houses in premium or prestigious suburbs |

To finalise the adequacy of the model we should check the assumptions. First check is to look at correlations between the variables, and so we construct a correlation matrix:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *Size* | *Brick* | *Bedrooms* | *Bathrooms* | *Nbhd* | *Price* |
| Size | 1 |   |   |   |   |   |
| Brick | 0.0798 | 1 |   |   |   |   |
| Bedrooms | 0.4838 | 0.0464 | 1 |   |   |   |
| Bathrooms | 0.5227 | 0.1720 | 0.4146 | 1 |   |   |
| Nbhd | 0.2508 | 0.1158 | 0.4715 | 0.2859 | 1 |   |
| Price | 0.5530 | 0.4528 | 0.5259 | 0.5233 | 0.7140 | 1 |

There are no very high correlations between the independent variables with the highest correlation being size and bathrooms. Given the correlation for bedrooms is lower this would not seem to be the reason for the non-significant result, however we do see that the individual correlation between bedrooms and price is higher than bathrooms and price which leads to the question of why bathrooms was significant but bedrooms was not.

Next we look at the residual plots:



There are no obvious patterns in any of these, however the variance for bedrooms does seem to change greatly for the 3 and 4 bedroom properties. This may be the issue, but there is no real evidence here.

ii.

The intercept coefficient, b0 = 81870.53 which is the value of price if all other variables are equal to zero. So for a non-brick house in a non-premium suburb that has no bedrooms or bathrooms and is zero size, the price would be $81,870 on average. No this is not really interpretable other than we know that everything increases from there.

iii.

Looking at the variable Brick which is a dummy variable with Brick = 1, the coefficient b1 = 97430.78 which implies that brick houses cost $97,430.78 more on average than non-brick houses.

To some extent this is consistent with part a in that we see there is a higher price on average for Brick houses, although the difference in average here is less ($97,000 compared to $129,000). The difference in these two values could be due to the fact that the regression equation is also adjusting for other factors – for example there may be more brick houses in some suburbs compared to others.

Since in part a we carried out a test of the significant difference using a t-test of differences in means, we should also carry out a test of significance here to see if it is consistent. This was done in the last part of the question and showed a significant difference in the price.

iv.

The neighbourhood dummy variable has a slope coefficient b5 = 155230 which implies that houses in the premium suburb of Windsor cost on average $155,230 more than those in less prestigious suburbs. Again it has a p-value of 0 which implies that this is a significant factor in pricing homes.

v. Use the model you have developed to predict how much the investor should be paying for a brick house of 180 square metres in size, with 3 bedrooms and 2 bathrooms in Windsor.

$$\hat{y}=81870.53+1855.53x\_{1}+97430.78x\_{2}+11402.41x\_{3}+34861.06x\_{4}+155230.23x\_{5}$$

$$\hat{y}=81870.53+1855.53(180)+97430.78(1)+11402.41(3)+34861.06(2)+155230.23(1)$$

$$\hat{y}=732456.39$$

**Question 3**

The personnel manager of a government department, as part of its recruitment program, regularly supplies tertiary students with its latest survey of wage rates and years of tertiary education and service of its current employees. The data is given below:

|  |  |  |
| --- | --- | --- |
| Wage | Education | Service |
| 81 | 5.5 | 12 |
| 45 | 0 | 3 |
| 37 | 1 | 1 |
| 61 | 4 | 10 |
| 57 | 3.5 | 6 |
| 49 | 2 | 20 |
| 73 | 6 | 14 |
| 53 | 3 | 7 |
| 65 | 4.5 | 16 |
| 69 | 5 | 13 |

where Wage = hourly wage rate ($)

Education = years of tertiary education

Service = years of service

Regress Hourly Wage on years of education and service. Conduct a full analysis of the results including interpreting any values necessary.

What hourly wage would you expect for a person with 8 years of tertiary education and 3 years of service? Do you have any concerns about this estimate?

We will be estimating the equation: $Y=β\_{0}+β\_{1}X\_{1}+β\_{2}X\_{2}+ε$

 Where Y = Hourly Wage Rate ($)

X1 = Years of Tertiary Education

X2 = Years of Service

We would expect both education and service to have a positive effect on Y

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |   |   |   |   |   |
|   |   |   |   |   |   |
| *Regression Statistics* |   |   |   |   |
| Multiple R | 0.9348 |   |   |   |   |
| R Square | 0.8738 |   |   |   |   |
| Adjusted R Square | 0.8378 |   |   |   |   |
| Standard Error | 5.4369 |   |   |   |   |
| Observations | 10 |   |   |   |   |
|   |   |   |   |   |   |
| ANOVA |   |   |   |   |   |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 2 | 1433.0774 | 716.5387 | 24.2398 | 0.0007 |
| Residual | 7 | 206.9226 | 29.5604 |   |   |
| Total | 9 | 1640 |   |   |   |
|   |   |   |   |   |   |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |   |
| Intercept | 36.5481 | 3.9616 | 9.2256 | 0.0000 |   |
| Education | 6.3099 | 1.1061 | 5.7048 | 0.0007 |   |
| Service | 0.0669 | 0.3646 | 0.1836 | 0.8595 |   |

The estimated equation is:

$$\hat{y}=36.5481+6.3099x\_{1}+0.0699x\_{2}$$

b0 = 36.5481. This implies that for a new employee with no tertiary education the base wage would be $36.54 per hour.

b1 = 6.3099 implies that for every additional year of tertiary education, the hourly wage increases by $6.31 on average.

b2 = 0.0699 implies that for every additional year of service, the hourly wage increases by 7c. This seems a strange result and not in line with normal expectations.

Looking at the overall adequacy of the model, $\overbar{R}^{2}=0.8378$ which implies that 83.78% of the variation in wages can be explained by variation in education and years of service.

Test of Overall Significance

$$H\_{0}: β\_{1}=β\_{2}=0$$

$$H\_{1}:At least one β\ne 0$$

Decision Rule: Reject Null Hypothesis if p-value < α (which we will set at 0.05)

P = 0.0007 which is less than 0.05

Hence we can reject Ho and conclude that the overall model is significant and the variables education and service explain wages.

Testing each of the variables separately are the t-tests.

Education Experience

$$H\_{0}: β\_{1}=0$$

$$H\_{1}:β\_{1}>0$$

Decision Rule: Reject Null Hypothesis if p-value < α

p-value = 0.0007/2 = 0.00035

Hence we can reject the null hypothesis and conclude that educational experience significantly explains variation in hourly wage.

Service

$$H\_{0}: β\_{2}=0$$

$$H\_{1}:β\_{2}>0$$

Decision Rule: Reject Null Hypothesis if p-value < α

p-value = 0.8595/2 = 0.42975

Hence we can not reject the null hypothesis and conclude that years of service do not explain wages.

Checking the validity of the assumptions we look at the correlation matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | *wage* | *Education* | *Service* |
| wage | 1 |   |   |
| Education | 0.934462 | 1 |   |
| Service | 0.535928 | 0.551508 | 1 |

From this we see that there is a reasonably high correlation between years of education and years of service with r=0.5515. In fact this correlation between service and education is higher than between service and wage (0.5359). This is an issue because, even though service is obviously correlated with wages, the correlation with education means we have an issue with Multicollinearity. The real problem here is that there would not seem to be any reason why years of education and years of service would be correlated so highly. Given this is a very small sample I suspect this is a sampling error issue.



In addition to the issue of multicollinearity, there also seems to be a long run of negative errors in the education residual plot, and in fact it looks as if there are two significant outliers on the high side. Again this is an issue with the sample.

At this point I would say that the current results would need to be accepted only with a great deal of caution and potentially a new sample selected.

With that in mind, if we were to predict an outcome for an employee with 8 years of tertiary education and 3 years of service we would get:

$$\hat{y}=36.5481+6.3099x\_{1}+0.0699x\_{2}$$

$$\hat{y}=36.5481+6.3099(8)+0.0699(3)$$

$$\hat{y}=\$87.24$$

A further issue of concern I would have with this particular estimate is that the value for x1 = 8 is an extrapolation well outside the bounds of our observed data with the previous highest observed observation being 5.5. Hence there is potential for further significant error that we would need to take in to account.

**Question 4**

For each of the following statements indicate if they are true or false and why.

1. When conducting an F test of the overall significance of the model, rejecting the null hypothesis implies that all of the independent variables are significant.

False. The F test of overall significance tests if there is an overall impact on the dependent variable. This could be caused by some or all of the variables included in the regression. The alternative hypothesis is always that At least one independent variable has a significant impact. If we determine there is a significant overall model, we then go to the individual t tests to see which of them are necessary. Note that if the F-Test was not significant then we do not need to do the t-tests as no individual variables will be significant.

1. The value of the adjusted R2 can decrease even if the Coefficient of Determination R2 increases.

True. The adjusted R2 is adjusted for the inclusion of additional variables. If we include an extra variable that has any non zero impact on Y, even though it might be very small, then R2 will increase. If that impact is very small, then R2 can increase slightly, but the adjusted R2 can decrease once we adjust for non-necessary independent variables.

1. The larger the value of the slope coefficient b1 the more likely it is to be significant.

False. As with all tests, it is not the value of the coefficient that is important, but its size relative to the standard deviation (standard error). That is why all tests use the standard error as the normalizing factor on the bottom line.

1. A p-value of 0.5 proves that the null hypothesis is true and there is no relationship between the dependent and independent variables.

False. Anyone who writes prove on the research report or final exam will lose marks. We are trying to determine the result of an analysis based on sample evidence. Wherever there is sample evidence there is uncertainty and the possibility of sampling error. All conclusions are the best based on the evidence, but they are not proof. A p-value of 0.5 leads us not to reject the null hypothesis because there is insufficient evidence to do so. It does not mean that the null hypothesis is true, just that we can’t reject it. Remember, it is like a trial – can we find evidence beyond reasonable doubt.

1. The intercept term can always be interpreted as the starting point for any relationship.

False. In some cases this might be true, if the data set relating to the X variables incorporates the possibility of 0 in the range of the variable. If not, then the intercept is simply an extrapolation of the data, and hence does not indicate a starting point, simple the point at which we would cross the Y axis if we could find data at that point.