**Case Study Deadline:** Wednesday 30th September 2020 by 23:55 PM

**Case Study Deliverables:**

*One professionally written case study report per group, inclusive of the following*:

1. One-page *executive summary*, highlighting main features of your analysis, modelling, and comments on results obtained.
2. Main case study report, not exceeding *15* one-sided pages, including a succinct understanding of the problem, regression models, graphical representations (residual plots, P-P plots, etc.), results, and conclusions. Specific recommendations for the different questions listed below must also be included in the report.
3. Show all the calculations in the report to receive full credit.

|  |  |
| --- | --- |
| **Q1** | 1. Why do you disagree with Jack’s comments about the uselessness of the regression due to the low R-squared? 2. Can you think of a situation in which a useless regression has a high R-squared? 3. There are techniques to determine the validity of a regression model—in particular, whether the relationship is linear and the error terms display equal variance (homoskedasticity). Does the regression in Table 1 violate either of these two assumptions? Justify your answer. |
| **Solution** | **a) Why do you disagree with Jack’s comments about the uselessness of the regression due to the low R-squared?**  R-square / Adjusted R-squared is one of the checks to assess the usefulness of the model . But R square is not the only metric to measure the validity of the model.  To test the validity of the model following tests must be performed   1. F test to check the validity of the model. Looking at the F statistics value, model is valid at **α=0.05**      1. T-test to check the existence of statistically significant relationship between response variable and individual explanatory variables. In this case at least two variables are statistically significant at **α=0.05**      1. Residual analysis to check that variance of residual is constant for all values of Xi. (Homoscedasticity). The residual plot shows that the values are mostly centered around 0. We conclude that there is Homoscedasticity      1. Test to check if residuals follow a normal distribution. This can be done using a pp plot by plotting observed cum probability against expected cum probabilities. The plot shows that though there is some deviation at the edge, the errors are mostly normal. (The deviation could be because of some outliers)      1. Multicolinearity test- to check if there is high correlation between the independent variables. Multicolinearity can be tested using Variance Inflation Factor(VIF). In this model the VIF < 4 for all variables.     **b) Can you think of a situation in which a useless regression has a high R-squared?**  R square/adjusted r square can be spurious i.e. High R-square does not necessarily mean that the Regression model is valid. Seemingly unrelated can show a high r square if the correlation between the variables is high.  **c) There are techniques to determine the validity of a regression model—in particular, whether the relationship is linear and the error terms display equal variance (homoskedasticity). Does the regression in Table 1 violate either of these two assumptions? Justify your answer**  **Homoscedasticity:**  The residual plot shows that the values are mostly centered around 0. We conclude that there is Homoscedasticity. There is no visible pattern in the error terms in the plot and therefore we can conclude that relationship is linear    **Normality of Residuals:**  Normality of residuals can be tested using a **pp plot** by plotting observed cum probability against expected cum probabilities. The plot shows that though there is some deviation at the edge, the errors are mostly normal. (The deviation could be because of some outliers) |

|  |  |
| --- | --- |
| **Q 2** | (a) Estimate the excess return (RET) of the funds that Bob and Putney currently manage. Assume that Princeton’s average composite SAT score is 1355, while Ohio State’s is 1042. Between Bob and Putney, who is expected to obtain higher returns *at their current funds* and by how much?  (b)Between Bob and Putney, who is expected to obtain higher returns *if hired* by AMBTPM and by how much? |
| **Solution** | **a) Estimate the excess return (RET) of the funds that Bob and Putney currently manage. Assume that Princeton’s average composite SAT score is 1355, while Ohio State’s is 1042. Between Bob and Putney, who is expected to obtain higher returns *at their current funds* and by how much?**  Using Mulitple Linear Regression ( and selecting only statistically significant variables at **α=0.05**), we get the following equation  **RET** = **-2.64216** - **GRI** \***2.11046** + **SAT** \* **0.005735**  It follows from the above equation, that Excess return is dependent on SAT and GRI. For the 2 Fund managers with same GRI, Excess Return(RET) varies with SAT of the college. Summarizing the RET calculation for Bob and Putney using the above equation we get   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | **College** | **SAT** | **GRI** | **Equation** | **Return** | | **Bob** | Ohio | 1042 | 1 | -2.64216 - 1 \*2.11046 + (1042) \* 0.005735 | 1.2232499999999993 | | **Putney** | Princeton | 1355 | 1 | -2.64216 - 1 \*2.11046 + (1355) \* 0.005735 | **3.018304999999999** |   The table shows that Putney will generate a higher return and the return would be  **1.7950549999999996**% higher on an average. at **α = 0.05**  **Important Note:**  Regression does not establish a **causal relationship**. Therefore, it is important to understand that the only difference between the Bob and Putney is the Average SAT of the college they graduated from. The MBA degree is not statistically significant at **α = 0.05.** Although the MLR equation does suggest that Putney would generate higher return on an average for the given data, it would be wise to look for more concrete evidence and not conclude based solely on the equation provided.  **b) Between Bob and Putney, who is expected to obtain higher returns *if hired* by AMBTPM and by how much?**  if hired by AMBTPM, there is no change in the above variables. On joining the **AMBTPM,** There will be change in TEN variable( TEN = 0 ). But at **α = 0.05,** TEN is not statistically significant . Therefore, based on the equation, we can conclude that **Putney** would outperform Bob after joining AMBTPM and will on an average generate a higher return of **1.75%**  at **α = 0.05**  **Important Note:**  The regression equation does favour Putney. However regression does not establish **causal relationship**. i.e. overall SAT score of the college might not be the reason for higher return. Not possible to conclusively determine whether Bob or Putney would generate higher return with the given data |

|  |  |
| --- | --- |
| **Q 3** | 1. Can you prove at the 5 percent significance level that if Bob had attended Princeton instead of Ohio State, then the return of his current fund would be greater? 2. Can you prove at the 10 percent level of significance that if Bob were managing a growth fund instead of a growth and income fund, then he would achieve at least 1 percent higher average returns? |
| **Solution** | **a)**  Using Mulitple Linear Regression ( and selecting only statistically significant variables at **α=0.05**), we get the following equation  **RET** = **-2.64216** - **GRI** \***2.11046** + **SAT** \* **0.005735**  It follows from the above equation, that Excess return is dependent on SAT and GRI. If **GRI** is held constant, the Excess Return is varies with the SAT of the college. Princeton has a Higher SAT average than Ohio State. Substituting the SAT values, we get the following Returns   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **College** | **SAT** | **GRI** | **Equation** | **Return** | | Ohio | 1042 | 1 | -2.64216 - 1 \*2.11046 + (1042) \* 0.005735 | 1.2232499999999993 | | Princeton | 1355 | 1 | -2.64216 - 1 \*2.11046 + (1355) \* 0.005735 | **3.018304999999999** |   As per the table(and from the **MLR equation**) Bob would generate a higher return if he had attended Princeton  **Important Observations:**   * It is important to note that Regression only establishes **association relationship** and does not establish a **causal relationship** between the dependent and independent variables. Therefore, it is not accurate to conclude that Bob would generate higher returns had he attended Princeton. * There is no variable in the given data that can be used to show an association between the University and Excess returns. An Association does exist between SAT scores and Excess Return. However students score SAT without help of universities. Therefore, the above equation does not conclusively prove that Bob would have generated higher return had he attended Princeton   b)  Using Multiple Linear Regression ( and selecting only statistically significant variables at **α=0.1**), we get the following equation  **RET** = **-2.64216** - **GRI** \***2.11046** + **SAT** \* **0.005735**  Keeping SAT contant, we get the following results for Growth and Growth and Income funds   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Fund type** | **SAT** | **GRI** | **Equation** | **Return** | | Growth | **Const** | 0 | -2.64216 - 0 \*2.11046 + (const) \* 0.005735 | -2.64216 + const | | Growth & Income | **Const** | 1 | -2.64216 - 1 \*2.11046 + (const) \* 0.005735 | **-4.75262 + const** |   Therefore, if Bob were managing only 'Growth' he would have generated a return **2.11046%** higher on average. Thus he would have generated at least 1% higher average returns |

|  |  |
| --- | --- |
| **Q 4** | 1. Does the regression in Table 1 provide strong evidence for the claim that fund managers with MBAs perform worse than managers without MBAs? What is being held constant in this comparison? Discuss. 2. It has been suggested that fund managers without MBAs get higher expected returns because they invest in riskier stocks. If this were true, what effect would including an independent variable, Beta (with higher values corresponding to higher levels of systematic risk in the fund’s portfolio), have on the coefficient of MBA in the regression of Table 1? |
| **Solution** | **a)**   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | **Coeff** | **Se** | **T** | **P** | **Beta Weight** | | **MBA** | –0.18065 | 0.756644 | –0.2387 | 81.139% | –0.0103 |   Test of statistical significance of MBA  H0 : β = 0, No relationship between MBA and Ret  HA: β != 0, There is relationship between MBA and Ret  P-Value = 0.81  According to the P-value , we retain the Null Hypothesis. i.e there is no relationship between MBA and returns.  It is wrong to conclude that Candidates with MBA perform worse than candidates with MBA or vice versa as MBA is not statistically relevant according to the above regression  **b)**  If an independent variable β(Systemic risk) is added to the regression equation, then there could be 3 scenarios   |  |  |  |  | | --- | --- | --- | --- | | **Scenario** | β(α=0.05) | MBA(α=0.05) | Remarks | | Scenaio 1 | Statistically not significant | Statistically not significant | β values do not have any influence on MBA coefficient | | Scenario 2 | Statistically significant | Statistically not significant | β values do not have any influence on MBA coefficient. As MBA is not part of regression equation | | **Scenario 3** | **Statistically significant** | **Statistically significant** | **If both are statistically relevant , that would mean that as β goes up the value of MBA would come down, as it is theorized that candidates with MBA would prefer assets with low β** | |

|  |  |
| --- | --- |
| **Q 5** | 1. What is the lowest level of significance at which you can prove that the manager’s age has a negative impact on his or her fund’s performance holding the type of the fund, the manager’s education, and years of experience at the fund constant? 2. A survivorship bias is thought to be present in analyzing fund manager performance in which a younger manager’s survival in the industry is more closely linked to his/her performance than an older manager’s survival. In other words, if a new manager does not perform successfully, he or she is not tolerated in the industry for long, but a more experienced manager may be forgiven a year or two of poor performance. Would the presence of this survivorship bias dampen or exacerbate the effect seen in Part (a)? |
| **Solution** | **a)**   |  |  |  | | --- | --- | --- | |  | **Coefficient** | **P-Value** | | **AGE** | –0.068893 | 10.006% |   Coefficient of Age is negative at a significance value of **α = 0.10006.** Therefore , we can say that age has negative impact on the ret at the lowest significance of **~ 0.1.**  **b)**  Presence of Survivor would exacerbate the effect, because only younger managers who generated higher returns would be tolerated in the industry. Naturally, returns generated by the younger managers would be higher in the final regression equation. i.e. only best performing younger managers would survive and only successful managers' would be included in the data. |

|  |  |
| --- | --- |
| **Q 6** | (a) “Streamline” the regression given in Table 1, that is, eliminate all variables that are not significant at the 15 percent level. Write down the new regression equation and check whether the specification satisfies the assumptions of linearity and homoskedasticity.  (b)Compare the coefficient of AGE in the new and the old regressions. What can explain the sign (direction) of the *change* in this estimator? Discuss. |
| **Solution** | **a)**  At 15% significance level  **RET** = **GRI** \* -2.1110+ **SAT** \* 0.0062 + **AGE** \*-0.0960 -2.5839  **Homoscedasticity:**    There is some clustering, this shows that variance error is not constant  **Check for Normality**    Although most residuals are normally distributed , there are some deviations at the edges  b)   |  |  |  | | --- | --- | --- | |  | **Old** | **New** | | Age | –0.068893 | -0.0960 |   The coefficient of AGE is lower in the new regression equation compared to the older regression equation and the AGE is statistically significant variable at α = 0.05. There is no change in the **sign of the estimator**. This shows that coefficient of AGE is stable when a new variable is added or removed. Change in the direction of the sign or large change in the value of coefficients on adding/removing a variable could be a symptom of **multi-collinearity.**  But this is not observed here. |
| **Q7** | (a) Run a regression to compare the average returns of growth and growth and income funds. Which type of fund yields a greater average return? Discuss whether or not the assumption of homoskedasticity is satisfied in this specification. What are the implications of this finding regarding the estimated difference in the average returns and the significance of the difference?  (b) Redo the analysis in part a using Excel’s TTEST function instead of a regression. Using this technique, can you prove at the 5 percent level of significance that the average returns of growth and growth and income funds differ? |
| **Solution** | **(a)**  Regression equation with only GRI as explanatory variable.  RET = -2.3119 \* **GRI** + 0.3959    The Model is statistically significant because p-value of the model is less than 0.05. This shows that average return for Growth Funds is higher on average compared to Growth and Income funds. But this hypothesis can only be accepted after checking the validity of regression for ex: check for homoscedasticity, Normality of error..  But condition of **Homoskedasticity is not satisfied.**  This implies that **variance of residuals is not constant** and hence the Regression equation is not valid    **b)**    The above 2 sample T test shows that **Return of Growth funds** differ from **Growth and Income Funds.** (p-value = 0.00136). |
| **Q8** | (a) You receive the prospectus of a growth fund started in the current year by a Princeton alum. What is the estimated RET (excess return relative to the return of the benchmark market portfolio) for this fund?  (b) Are you confident that this fund will “beat the market”, that is, provide a return in excess of that of the benchmark market portfolio? Which standard error do we have to use in order to answer this question? |
| **Solution** | **The excess return is given by the following equation**  **RET =** -2.2785\* **GRI** **+** 0.0060\***SAT-**6.4611  **Variable Values:**   |  |  | | --- | --- | |  | **Values** | | **SAT** | Avg SAT score of princeton. (Data not provided) | | **GRI** | 0 (for Growth fund) |     Therefore, the excess Return would increase with every additional SAT score of the college.  **(b)**  From the regression equation, only SAT score is relevant. (GRI =0)  To calculate if the fund will beat the benchmark, we calculate the **Prediction interval** using the Standard Error (Se) of the **Entire Model.** This is because Although GRI = 0 , we still have the intercept.    Standard error of the model **Se = 8.41**  Minimum SAT score to **"Beat the market"** is given by  0 = SAT \* 0.0060\***SAT-**6.4611  Solving for SAT, SAT = 1076.85  i.e. at SAT > 1076.85, the **RET > 0** on average.  At Minimum SAT score of 1076.85 the prediction would have positive and negative value.  i.e at SAT score of 1076.85, the probability of 'beating the market' is given by  1 - NORMDIST( 0,0, **8.41, True) = 0.5**  Probability of beating the market is **greater than 50 %**. The confidence value will increase with the SAT score.  **(c)**    The regression equation is given by  From above, the Minimum **SAT** score to beat the market ( to E(Y|X) > 0 ) = **1076.85**  Since RET is normally distributed around 0. The probability of exceeding the Excess return by **1.5** is given by  1 - NORMDIST(1.5, Mean , Se, TRUE)  1 - NORMDIST(1.5, 0, , **8.410, TRUE) = 0.429**  Therefore the probability to beat the market by 1.5% is greater than **42.9%**  **Standard Error** of E(Y|X) is used |

|  |  |
| --- | --- |
| **Q 9** | Suppose that you gain access to a much larger sample of random observations of the same variables that you have in the current dataset. Do you expect that any of your answers to Parts (a)–(c) of Question 8 will change, and if so, how? Discuss |
| **Solution** | There are only 2 possible  There would be no significant change in the answers with a larger dataset because  1> The SAT and GRI are the two variables that are used in the calculation. Both the variables are statistically significant with the current dataset. As a result there would be no significant change in the coefficients with additional samples.  2> The data for GRI and SAT is not skewed and is reasonably balanced.      3> Question 8-c depends on the Standard Error of E(Y|X) . This value would not change significantly for larger values of n. So the values would not change for larger values of n |

|  |  |
| --- | --- |
| **Q 9** | a)Based on the dataset, can you prove at the 5 percent level of significance that among fund managers with the same educational background and same experience with the same fund, those managing growth and income funds are, on average, older?  b) Using the regression developed in part a, provide an 80 percent confidence interval for the average age difference between managers who graduated from the same college in the United States and have managed a growth fund for the same number of years, but differ in whether or not they have an MBA . Are the (otherwise comparable) managers with MBAs younger or older, on average? Discuss (conjecture) why this is the case. |
| **Solution** | **a)**  GRI variable is not statistically significant. **Therefore it cannot be proved at 5%** significance that among fund managers with same educational background and same experience , those managing Growth and Income funds are older.    **b)**    **Regression equation for only growth Funds**  **AGE = SAT\***0.0109 **+ TEN\***0.7762 **-** 1.7134**\*MBA +** 29.1297  Comparing two managers with the same Education, tenure and Managing growth funds, the regression equation will reduce to  **AGE = -**1.7134**\*MBA +** 29.1297  Calculating the confidence interval of managers with and without MBA     |  |  |  | | --- | --- | --- | |  | Lower Level | Upper Level | | **With MBA** | 26.47 | 28.36 | | **Without MBA** | 28.185 | 30.07 |   Managers with MBA are on an average younger than Managers without MBA at 80% confidence. |