End Term Assignment

Name: Prakhar Rathi

Roll Number: 1810110169

Q1. Do you think that the aptitude test is useful for selecting students for admission to the course, and if so, how would you determine the pass mark?

Solution: There are multiple stages of answering this question. First, we need to encode the text data into numerical variables for the regression analysis. Hence, the field 'course_result' was encoded which had a 1 if the person passed the course and 0 if a person failed the course. This was the dependent variable and the field 'test_score' was the independent variable. We want to identify whether the independent variable 'test_score', which contains the scores for the aptitude test, has an impact in determining whether a person will pass or fail the course. Since, the dependent variable is a binary variable taking values 0 or 1, the equation or the econometric model which establishes the relationship between X and the outcome Y can be seen below. Here b_0 and b_1 are the population parameters whose values we have to estimate to establish a relationship between the dependent and independent variables and u is the disturbance term.

$$P(Y = 1 \mid X) = b_0 + b_1 * X + u$$

The equation that we fit on our data will be written as follows:

$$P(Y = 1 \mid X) = B_0 + B_1 * X$$

The above equation represents the Ordinary Least Squares regression equation for binary dependent variable where Y is the binary dependent variable and X is the independent variable and P(Y=1 | X) is the probability of getting Y = 1 (student passes) given a particular X value. B_0 and B_1 are the linear, unbiased and the most efficient estimates of b_0 and b_1 which have been obtained using the Least Squares Estimates Derivation Method which shows the coefficients as:

$$B1 = \frac{\sum_{i=1}^{n}(Xi - \overline{X})(Yi - \overline{Y})}{\sum_{i=1}^{n}(Xi - \overline{X})^2}$$

$$B0 = \bar{Y} - B1\bar{X}$$

Here, n are the number of observations and for this question n=36.

By doing a regression analysis with the help of STATA, we get the estimates for the parameters of this relationship. The column *'Coef.'* from the screenshot below gives us the estimates of these parameters so our equation can now be written as:

$$P(course\_result = 1 \mid test\_score) = 0.0626239 + 0.0099203 * test\_score$$

```
. reg course_result test_score
```

Source	SS	df	MS	Number of obs =	36
				F(1, 34) =	5.21
Model	1.19154349	1	1.19154349	Prob > F =	0.0289
Residual	7.78067873	34	.228843492	R-squared =	0.1328
				Adj R-squared =	0.1073
Total	8.97222222	35	.256349206	Root MSE =	.47838

course_res~t	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
test_score	.0099203	.0043475	2.28	0.029	.0010851	.0187556
_cons	.0626239	.2188873	0.29	0.777	-.3822086	.5074563

Based on the above data, we can say that if the estimates are not rejected then an increase of 1 mark in the test_score would increase the probability that a student will pass by 0.0099.

Hypothesis Testing and t-tests

The above snippet contains the output of the regression analysis of the 'course_result' on 'test_score'. Even if we get the estimates of the parameters of the relationship, they are still estimates and we need to identify whether they are reliable. I will perform a two-tailed t-test to see whether the relationship exists or not. I will take my null hypothesis (H_0) as the assertion that the course result does not depend upon the test score. My alternate hypothesis (H_1) would then be that test scores do affect the course outcome of a student.

$$H_0: b_1 = 0$$

$$H_1: b_1 \neq 0$$

If I can reject my hull hypothesis then I can establish some relationship between the test score and the course outcome, at least in general terms. I will calculate the t-statistic as follows:

$$t = \frac{B1 - b1}{s.e(B1)} = \frac{0.0099203 - 0}{0.0043475} \approx 2.28$$

Here, s.e is the standard error.

The t-statistic value has been calculated as 2.28 and it can be seen that the STATA output is also the same for this null hypothesis. There are 36 observations in our dataset so the degrees of freedom (n-2) will be equal to 34. The t_{crit} value for 34 degrees of freedom at 5 percent significance level (calculated using a t-table) is 2.03. Since the t statistic value is greater than the critical value, we will reject the null hypothesis at this level and conclude that the test scores do affect the course outcome. On performing the test at 5 percent level, there is also a 5 percent risk of a Type I error so I will perform another test at the 1 percent level. However, when we perform the same test at 1 percent significance level then the

critical value is 2.72 which is larger than the t-statistic value that was calculated, hence, we cannot reject the null hypothesis at this level and we conclude that the test scores have no effect on the course outcome.

The column 'P > |*t*|' provide the p-value for each coefficient. This p-value is the probability of obtaining the corresponding t-statistic value by chance, if the null hypothesis *H0: b1 = 0* were true. Here, the p-value is 0.029 which means that this probability is 2.9% this shows that the null hypothesis would be rejected at the 5 percent level (0.029 < 0.05) but not at the 1 percent level (0.029 > 0.01). By the virtue of p-value, we get the same result as above.

There are certain other factors which contribute to the conclusion of whether the test is reliable or not. The first factor is the R^2 value. The R^2 is the ratio of the explained sum of squares to the total sum of squares or 1 – ratio of the residual sum of squares and the total sum of squares. Essentially, it tells what proportion of the total variance in our dependent variable is explained by the independent variable. In this case, that value is 0.13 which can be translated to say that only 13% of the variation in the course outcomes is actually explained by the test scores which is quite a low percentage. While this is not an absolute measure but such a low R2 score does cast doubts on the relationship between aptitude test scores and the course outcome.

On graphically representing the data, another interesting thing can be seen, a student who got marks around 80 can also fail the course which means that higher marks on the test do not *guarantee* passing in the course.

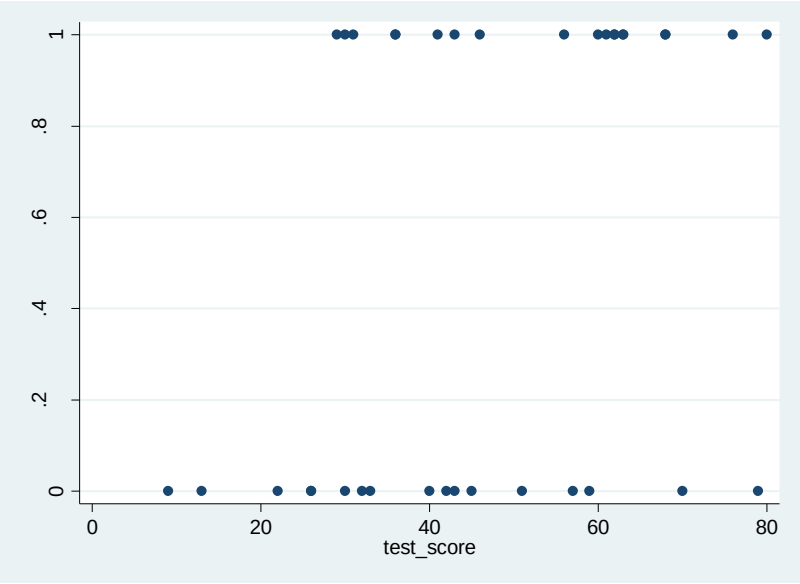

Figure 1: Scatter Plot of the course_result against test_scores

Another thing that can be seen is that it is very hard to draw a line of demarcation in this graph where we can say that "students on the left will fail the course and students on the right will pass" because

there is a lot of region of overlapping where the outcome is uncertain. This can be formalized through the means of a simple table.

Range of aptitude test scores	Course Result (with certainty)	Number of Students
0 – 26	Failed with absolute certainty	5
27 – 79	No certainty of pass/fail	30
80 – 100	Passed with absolute certainty	1

While, **in this sample,** marks below 26 have guaranteed failing and marks above 80 have guaranteed passing, almost 52 marks range and outcome for 30 students (83% of the class) is uncertain. These estimates are only for this sample and our bound to have a certain degree of uncertainty and error but over 80% outcome uncertainty is concerning. Finally, there are two students at the score of 43 out of which one student passed and the other failed. This is adding to the concerns.

Keeping all the above tests and observations in mind, I can conclude that this **aptitude test is unreliable in predicting the course outcome** and hence, **not useful in selecting students for admission** to the course.

However, since I have made the model already, there is a way to estimate the pass mark for this test. This is just an approach to finding the pass mark and in the future if the aptitude test can be made more reliable then we can actually use it. Using the equation below, we can find a test score which approximates the passing probability to 0.5 (50%).

$$P(course\_result = 1 \mid test\_score) = 0.0626239 + 0.0099203 * test\_score$$

$$P(course\_result = 1 \mid test\_score) = 0.5$$

$$0.5 = 0.0626239 + 0.0099203 * pass\_mark$$

$$Pass\_mark = (0.5 - 0.0626239)/0.0099203 = 44.08$$

$$= 45 \text{ (round up to gurantee a prob higher than 0.5)}$$

Note:- This approach would only work if we *accepted* that there is a relation between test score and course result (which we haven't based on this sample data), hence as of now we can't take this as a pass mark. Since, the test is unreliable at the moment so the pass mark is also unreliable.

One-tailed t-test (Additional)

I have added this section just as a thinking exercise and may not be necessary for evaluation. Holding the risk of making a Type I error constant (if the null hypothesis is true), we will have a smaller risk of making a Type II error (if the null hypothesis is false), if we use a one-sided test instead of a two-sided test (*from Dougherty*).

There are often claims that higher aptitude test scores may generate higher chances of passing in a course so if I take my alternative hypotheses as

$$H_0: b_1 = 0$$

$$H_1: b_1 > 0$$

and perform a one-tailed test on the right tail, I get the following results for different levels.

At 5 percent: t-crit (right-tail) = 1.69

At 1 percent: t-crit (right-tail) = 2.44

Since our t-stat = 2.28, we reject the null hypothesis at the 5 percent level since t-stat > t-crit but fail to reject at the one percent level because t-stat < t-crit. Even by switching to a one tailed test, we still get the same results for our inference thereby further strengthening our reasoning.