Assignment #9 **Name:_____**
For Practice ONLY **ID:__A00_____**
Write answers in spaces below **Section:_____**

Admission decisions to most university programs in Nova Scotia are based almost completely on the student's high school average. The average is composed of the grades of the best 5 courses that meet the admission requirements. For most business programs, the admission requirement is an average of 70%, based upon English 12, Mathematics 12 and 3 other grade 12 academic subjects.
Below are the grades of 20 students admitted into the B.Comm. program at Saint Mary's University in 2016. All 20 students attending the same high school in the Halifax region.

high school average	70.4	73.6	73.8	75.2	75.75	76.2	77.4	77.5	79.2	80.25
gpa	1.48	1.14	2.2	2.14	3.06	3.40	2.02	3.00	3.92	3.10
high school average	80.6	81.65	82	82.8	85.2	85.6	86.75	86.75	92	94.6
gpa	3.25	3.92	3.16	3.35	4.08	3.72	3.04	4.12	4.12	4.30

The gpa is the grade point average at the end of the fall term.

a) Construct the least squares regression line to predict fall term gpa (Y) based upon the student's high school average (X).

$$\sum X = 1,617.25 \quad \sum X^2 = 131,522.4$$

$$\sum Y = 62.52 \quad \sum Y^2 = 211.1922 \quad \sum XY = 5,141.862$$

$$SS_x = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 131,522.4 - \frac{(1617.25)^2}{20} = 747.5444$$

$$SS_y = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 211.1922 - \frac{(62.52)^2}{20} = 15.75468$$

$$SS_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n} = 5,141.862 - \frac{(1617.25)(62.52)}{20} = 86.3385$$

$$slope = b_1 = \frac{SS_{xy}}{SS_x} = \frac{86.3385}{747.5444} = 0.1155$$

$$intercept = b_0 = \bar{y} - b_1\bar{x} = \frac{62.52}{20} - (0.1155)\frac{1617.25}{20} = -6.2133$$

$$predicted\ gpa = \hat{y} = b_0 + b_1 X = -6.2133 + 0.1155X$$

b) Calculate the coefficient of correlation between high school average and fall term gpa.

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{86.3386}{\sqrt{(747.5444)(15.75468)}} = 0.795575$$

c) Construct a scatter plot of fall term gpa versus high school average and plot the regression line from (a).

d) Based upon your visual examination of the plot and the correlation coefficient found in (b), do you believe that high school average is an effective predictor of fall term performance?

The correlation coefficient is quite high. The scatterplot looks like an upward sloping pattern with the points being reasonably close to the line. Is it an effective predictor? Depends upon what accuracy you are looking for. Some points are almost a full point above or below the line. In one case, the line predicts a C average and the student had a D average. In another it predicted a B average and the student had an A average.

e) Test the hypothesis that high school average and fall term gpa are related at the $\alpha = 0.05$ level of significance. You can either perform the test based upon the slope of the regression line or the correlation coefficient.

H_0: slope of the regression line is zero $\beta_1 = 0$
H_A: slope of the regression line is not zero $\beta_1 \neq 0$

The test statistic is

$$t = \frac{b_1 - \beta_1}{\frac{s_e}{\sqrt{SS_x}}} \; with\; df = n - 2 = 20 - 2 = 18$$

At α = 0.05, we will reject H_0 and conclude that the slope is non-zero if the observed value of t exceeds t(0.025,18) = 2.101 or is less than –t(0.025,18) = -2.101. Note that the alternative says the slope is less than or greater than zero so we have a 2 tailed test.

$$SS_{error} = SS_y - b_1 SS_{xy} = 15.75468 - (0.1155)(86.3385) = 5.78258$$

$$s_e = \sqrt{\frac{SS_{error}}{n-2}} = \sqrt{\frac{5.78258}{20-2}} = 0.5668$$

$$t = \frac{0.1155 - 0}{\frac{0.5668}{\sqrt{747.5444}}} = 5.57$$

Since the observed t score is substantially larger than the critical value of 2.101, there is strong enough evidence to conclude that the slope is non-zero and there is a relationship between high school grades and fall term gpa.

H0: high school average and fall term gpa are unrelated ρ = 0
HA: high school average and fall term gpa are related ρ ≠ 0

The appropriate test statistic to test whether the correlation coefficient is equal to zero is

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}} \; with\; df = n - 2 = 20 - 2 = 18\; degrees\; of\; freedom$$

At α = 0.05, we will reject H_0 and conclude that HS average and gpa are related if the observed value of t exceeds t(0.025,18) = 2.101 or is less than –t(0.025,18) = -2.101. Note that the alternative says the correlation is less than or greater than zero so we have a 2 tailed test.

$$t = \frac{0.795575 - 0}{\sqrt{\frac{1-(0.795575)^2}{20-2}}} = 5.57$$

Since the observed t score is substantially larger than the critical value of 2.101, there is strong enough evidence to conclude that there is a relationship between high school grades and fall term gpa.

Observe that the two ways of testing this hypothesis are equivalent. They both use t scores so we should expect that the critical values and the observed t score will be the same. If they are not, then we made a mistake.

f) Students with an average of 80 received an entrance scholarship of at least $500. Construct a 95% confidence interval for the mean gpa of students with a high school average of 85.

This is a confidence interval for where the line passes above X = 85. It does not tell us what any particular students gpa will be. A confidence interval for $\mu_{y|x}$ is

$$\hat{y} \pm t s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}} \quad where\ t\ has\ n - 2\ degrees\ of\ freedom$$

For 95% confidence, t = 2.101

At x = 85, $\hat{y} = -6.2133 + 0.1155(85) = 3.6042$

The interval is

$$3.6042 \pm (2.101)(0.5668)\sqrt{\frac{1}{20} + \frac{(85-80.8625)^2}{747.5444}}$$

3.6042 ± 0.3215

We can be 95% confident that the mean gpa of all students with a high school average of 85 will be between 3.28 and 3.93.

g) Construct a 95% prediction interval for the gpa of a student with a high school average of 85.

We are asked to predict the gpa of a particular student that has a high school average of 85. This is like predicting where a future observation will fall above x = 85. So we need to take account of both the uncertainty of where the regression line falls as well as the scatter that occurs above and below this line. A prediction interval for a future y value is

$$\hat{y} \pm t s_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_x}} \quad where\ t\ has\ n - 2\ degrees\ of\ freedom$$

The interval is

$$3.6042 \pm (2.101)(0.5668)\sqrt{1 + \frac{1}{20} + \frac{(85-80.8625)^2}{747.5444}}$$

3.6042 ± 1.2335

There is a 95% chance that the student's gpa will fall between 2.37 and 4.84. Since we are familiar with gpa data, we know that it is impossible to have a gpa in excess of 4.30, but our model doesn't know this.

h) What assumptions are required for your above analyses to be valid?

We have assumed that

- There is a linear relationship between high school average and fall term gpa
- The observations are independent of one another
- The distribution of gpa's for any specific high school average is normal. That is the distribution around the line is normal.

- The variation in gpa's, for any specific high school average is constant. That is, the standard deviation of the normal distributions is constant.

i) Below is a plot of the residuals = actual gpa – predicted gpa, versus the high school average. What insights may this plot give you into the validity of your model or how to possibly improve it?

It looks like it might curve in a concave down fashion. The residuals are initially negative, then become positive and then become negative again. This type of pattern suggests that maybe the relationship is slightly curved. Below is a plot with a model whose form is a quadratic

j) In 2016 there were 74 students that were admitted to the B.Comm. program from Nova Scotia high schools outside of Halifax Regional Municipality. Below is the ANOVA table for predicting fall gpa based upon the student's high school average. In what respects do these results differ from what was observed for a Halifax high school?

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.638609
R Square	0.407821
Adjusted R Square	0.399596
Standard Error	0.722773
Observations	74

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	25.90315	25.90315	49.58482	9.26E-10
Residual	72	37.61285	0.522401		
Total	73	63.51599			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-5.50942	1.19298	-4.6182	1.65E-05	-7.88758	-3.13125
high school average	0.10168	0.01444	7.041649	9.26E-10	0.072895	0.130466

The relationship between high school average and gpa is slightly different

Estimated gpa = -5.50942 + 0.10168

The slope is slightly lower suggesting that as HS average increases, the gpa does not increase as fast. This could be simply because it is a different sample of students. If you look at the lower and upper 95% confidence limits on the slope and intercept you will see that they are quite wide and overlap substantially with estimates in part (a).

We also see that the correlation coefficient is somewhat lower at 0.6386 compared to 0.7956 previously. Could this be just random error too? It is beyond what we have covered, but it can be shown that it easily could be due to sampling error and no real difference between urban and rural high schools. Always be careful about reading too much into comparisons of samples if you don't know what the sampling error could be.

k) Explain what the R Square (Coefficient of Determination) represents.

R square is equal to the square of the correlation coefficient, but it has another interpretation if we look at the ANOVA table. As in ANOVA when comparing means, we can partition variation. How much variation is there in gpa when we do not know the high school average? In the above example for rural high schools, the total variation as measured by the sum of squares, is 63.5. But this variation can be partitioned into variation between the overall average and the regression line (SS regression = 25.9) and the variation between the observations and the regression line (SS residual = 37.6). The Regression SS account for 25.9/63.6 = 0.41 or 41% of the variation. This proportion of the SS accounted for by the

model is called the coefficient of determination or R square. It is a useful measure of how useful the model is in explaining what is happening.

Be careful interpreting R square for small samples because it does fluctuate considerably.

l) Explain what the standard error represents. What is the relationship between the "standard error" and MS residual?

The standard error measures the standard deviation of the observations around the regression line. Crudely, you could think of the Emprical Rule and say that 95% of the observations should be within 2 standard errors of the line. Although we use it for tests and confidence intervals, it is often most useful in forecasting future observations and gauging how far off the forecast might be. 2 standard errors is a quick and dirty estimate of accuracy.

MS residual or MS error is equal to the square of the standard error or the sample variance around the regression line.

m) In the above ANOVA table, explain what the coefficients of -5.509 for the intercept and 0.10168 for the high school average represent. Do you have any problems with a negative intercept in this situation?

The slope of 0.10168 indicates that for every 1 point increase in the high school average, we should expect that the mean gpa will increase by 0.10. So students with an average of 90, should have a mean gpa that is approximately 1.0 or one letter grade higher than those with an average of 80 (e.g., A compared to a B).

The intercept indicates that students with a high school average of 0 should have a mean gpa of -5.51? This makes no sense on many levels. You can't have a gpa that is negative, but we would also never admit a student with an average of 0. The model really only applies over a limited range of values, say averages between 70 and 95. Further, we have assumed that the relationship is linear (a straight line). With the Halifax school data we had some indication that the relationship was curved. In general, the model is simply an approximation.

We should not try to interpret the model outside that range of observed cases.

The scatter plot for this larger sample does give some additional insight. Look at the variability in gpa as high school average increases. Since students cannot get a gpa above 4.30, there is a ceiling. From the plot below, it appears that there is less variability in gpa among students with high averages than among those with lower averages. This would violate one of the assumptions of the regression analysis. Dealing with this issue is beyond the scope of this course.