ENGG 319 F20

Final Report on Course Project

Date of submission: **Dec 7th, 2020**

Group # 56

No	Last Name	First Name	UCID	Section	Email
1	Nguyen	Hao	30088094	L03	hao.nguyen@ucalgary.ca
2	Huynh	Nguyen Gia Hy	30087093	L03	nguyengiahy.huynh@ucalgary.ca
3	Trinh	Phuong	30089842	L01	phuong.trinh@ucalgary.ca
4	Huynh	Ryan	30092355	L03	ryan.huynh@ucalgary.ca

Prediction of the population of Alberta by Year

1. INTRODUCTION:

The population of a fixed influences how policies are made in order to accommodate for that population. The larger the population, the more important decisions are. Alberta is no expectation. We want to see if there is linear relationship the year and the population of Alberta, so that we can predict the population of Alberta in the future. The paired sets of data are outline in TABLE 1.

TABLE 1. Population of Alberta data [1]

Year	Population
1901	73022
1911	374295
1921	588454
1931	731605
1941	796169
1951	939501
1956	1123116
1961	1331944
1969	1463203
1971	1627875
1976	1838035
1981	2237724
1986	2365830
1991	2545553
1996	2696826
2001	2974807
2006	3290350
2011	3645257
2016	4067175

2. CORRELATION ANALYSIS

The ordered pair (x=year, y= total population) are summarized by the scatterplot below (FIGURE 1). The graph indicates an increase in population (y- value) over the period from 1901 to 2016 (x-value) and vice versa showing a positive linear association. The correlation coefficient, r = 0.963, confirms a strong positive relationship. Since r is different from sample to sample, we estimated the population correlation coefficient (ρ). The 95% confidence interval for ρ is (0.9055, 0.9861) which suggest a positive association at 95% confidence and reveals the strength of this relationship can be strong

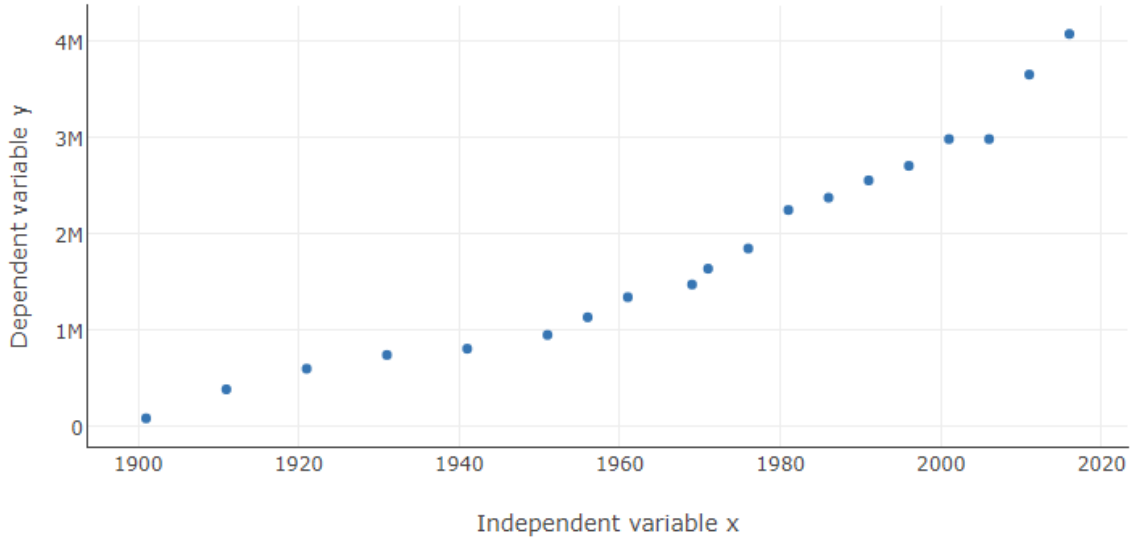


FIGURE 1. A scatterplot showing positive association between year (x) and the population (y).

3. REGRESSION ANALYSIS

3.1 The regression model

Defining the response variable, y = total population, and predictor or independent variable, x=year, the least-squared line that best fits the data is determined to be

Total population = -61261868.76431686 + 32061.71454429 (year)

3.1.1 The regression model

Table 2. Regression model for

Predictor	Regression coefficients	Standard error of coefficients	t - stat
y-intercept	*-61261868.7643*	2201.1872	14.8294
Total population	32061.7145	4330819.3	-14.4005

3.1.2 Interpretation of slope and intercept

The y-intercept indicates the population of Alberta where year is zero, therefore, this y-intercept does not demonstrate a practical interpretation for any time before Alberta was founded.

The slope indicates for every year passed, there is an addition of approximately 32062 new residents in Alberta's population.

The sample slope and sample intercept are dependent on the sample used in the analysis and will vary from sample to sample. Therefore, we have estimated population slope at the 95% confidence level, which is (27998.2511, 37286.3806). That is, we are 95% confident that for every year passed, the number of additional populations is between 27998.2511 and 37286.3806. Similarly, the 95% confidence interval for population intercept is (-71524589, -53250264), which has no interpretation as Alberta was not founded in the year 0 and you cannot have a negative number of people.

3.1.3 Goodness-of-fit statistics

We can see how well the regression model fits the data by analyzing the values of the regression statistics, and by performing a hypothesis test. By doing this, we can determine whether the regression model is good to use for prediction in comparison to the y bar model.

y bar model (yearly population of Alberta): 1,810,273

Regression Statistics: To measure goodness of fit, we look to the regression statistics in Table 3. Looking the coefficient of determination, we can say that the regression model can very strongly predict data points accurate, as it states that the regression predicts fit about 92.73 % of the data perfectly.

Table 3. Regression statistics (Year vs Population of Alberta)

Statistics	Value
Coefficient of determination, r^2	0.92824
Standard error, s	319,598.114

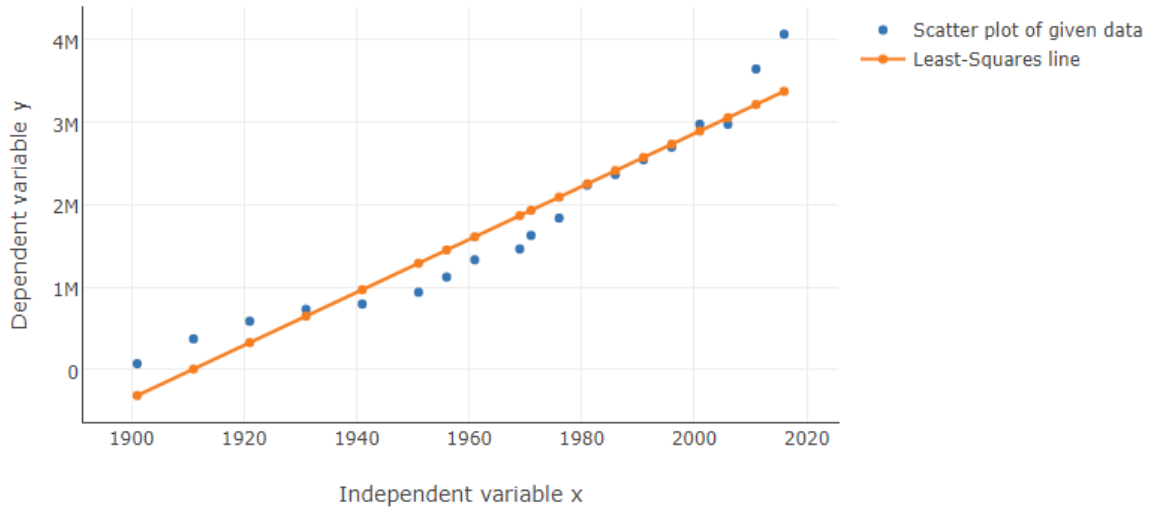


FIGURE 2. The variation of the datapoints around the regression lines indicating how well the line fits the given data.

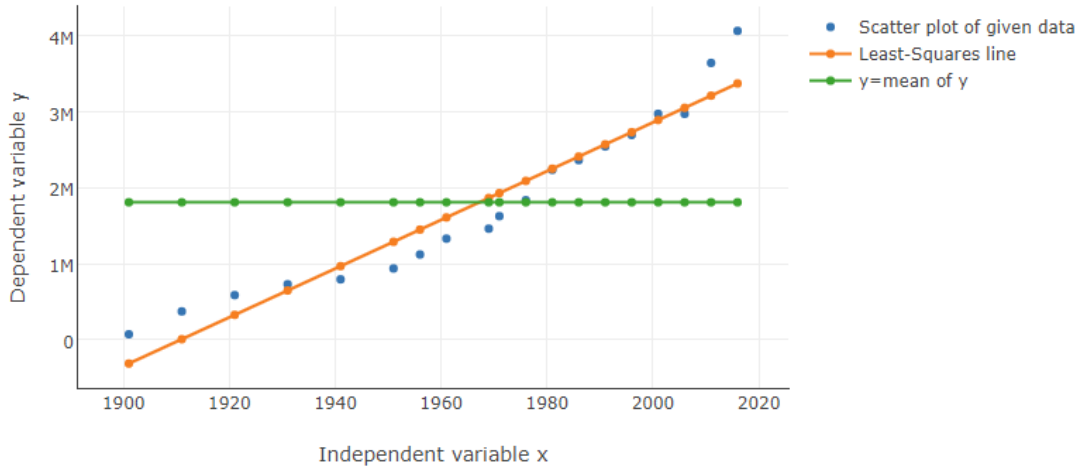


FIGURE 3. The variation of the datapoints around the two models (I) regression model and (ii) $\hat{y} = \bar{y} =$ model (horizontal line).

The standard deviation of y (approx. 1,159,474.068) represents the spread around the sample mean y bar (approx. 1,826,881.105). The standard error of estimate, s (approx. 319,598.114) represents the spread around the regression line. Since s is much less than the standard deviation of y, this tell us that the regression model is superior at prediction than the average model.

Hypothesis test:

To gain further insight, we will perform a hypothesis test to see if the population slope is equal to zero or not. If it is not equal to zero, then that would indicate the regression model is better at prediction than the y bar model.

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

The t statistic for the hypothesis test (population slope is zero) is 14.7246. With 17 degrees of freedom at 5% significance level, is well beyond the scope of the t table. However, the maximum t statistic shown on the t table is 3.965 which would entail a minimum P-value of 0.001. This implies that a t statistic greater than 3.965 (like 14.7246) would entail a P-value smaller than 0.001. Since 0.001 << 0.05, the null hypothesis is rejected at 5% significance level. There, the model is more predicting power than the simple y = y bar model.

3.1.3 Prediction

The linear regression line is important to predict the population of Alberta in the future years based on past trend of population. For example, the predicted population of Alberta in 2040 is estimated 4202898 people.

$$\text{total population} = -62387426.15 + 32642.31581(2040) \sim 4202898$$

However, this number might not be precise since we obtain only an estimated point by plugging-in the x value in the regression model. This point estimate will be different in different sample. We have calculated 95% confidence interval of mean value of y to be (3831145, 4574651). The width of the confidence intervals shows the variation or uncertainty in the prediction.

3.1.4 Checking model assumptions:

We have plotted the residuals against the fitted values of y to check the assumptions of linear models. The plot shows no substantial pattern in the spread of residual values around the mean value (zero), but vertical spread varies a lot over the range of x values. This behaviour is not consistent with the assumptions of linear model.

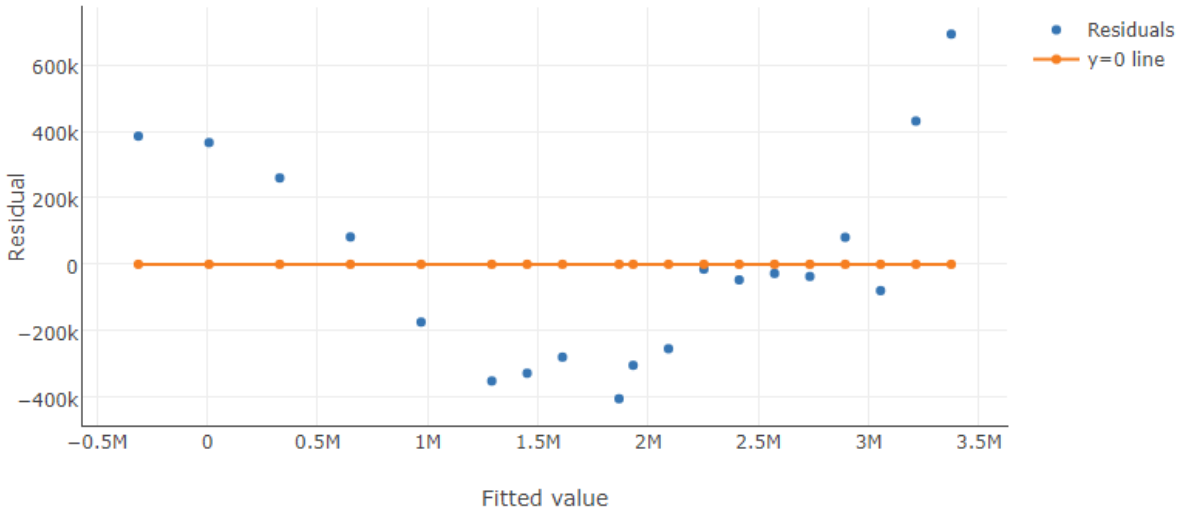


FIGURE 3. Plot of residuals versus fitted value.

4. CONCLUSIONS

There exists a positive linear relationship between the year (independent or predictor variable) and the population of Alberta (response variable). The sample correlation coefficient on a sample of size 19 was 0.963, and the 95% confidence interval was (0.9055, 0.9861), suggesting a stong relationship

The relationship found that can be used to predict the population of Alberta in the future is:

$$\text{Total population} = -61261868.76431686 + 32061.71454429\ (\text{year})$$

Statistical analysis of this regression show that it has statistical significance at 5% significance level. Some population parameters were estimated, revealing the uncertainty in the model prediction.

References

[1] https://en.wikipedia.org/wiki/Demographics_of_Alberta. Accessed on Nov. 24, 2020

[2] https://engg201.pythonanywhere.com/simple_linear_regression. Accessed on Dec. 5, 2020

APPENDIX A: SAMPLE CALCULATIONS

The main part of the report must be typed in MS word and submitted as a pdf file along with the sample calculations. However, the sample calculations can be written by hand or typed in MS word – but in both cases, should be converted pdf pages and appended it to the main report.

Please show the steps of the **detailed calculations (show all steps)** of the following.

1. Sample correlation coefficient (This calculation can be done using Excel spreadsheeted and screenshots can be shown.)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

	A	B	C	D	E	F	G	H	I	J
1		x	y		x-ave x	y-ave y	(x-avex)(y-avey)	POWER(X-AVE X,2)	POWER(Y-AVE Y,2)	
2		1901	73022		-66.21052632	-1753859.105	116123934.4	4383.833795	3076021761114.48	
3		1911	374295		-56.21052632	-1452586.105	81650629.5	3159.623269	2110006393203.59	
4		1921	588454		-46.21052632	-1238427.105	57228368.34	2135.412742	1533701695050.48	
5		1931	731605		-36.21052632	-1095276.105	39660524.23	1311.202216	1199629746760.43	
6		1941	796169		-26.21052632	-1030712.105	27015506.76	686.9916898	1062367443936.01	
7		1951	939501		-16.21052632	-887380.1053	14384898.55	262.7811634	787443451216.85	
8		1956	1123116		-11.21052632	-703765.1053	7889577.233	125.6759003	495285323386.06	
9		1961	1331944		-6.210526316	-494937.1053	3073819.917	38.57063712	244962738166.27	
10		1969	1463203		1.789473684	-363678.1053	-650792.3989	3.202216066	132261764247.80	
11		1971	1627875		3.789473684	-199006.1053	-754128.3989	14.3601108	39603429932.01	
12		1976	1838035		8.789473684	11153.89474	98036.86427	77.25484765	124409367.80	
13		1981	2237724		13.78947368	410842.8947	5665307.285	190.1495845	168791884155.75	
14		1986	2365830		18.78947368	538948.8947	10126566.07	353.0443213	290465911138.06	
15		1991	2545553		23.78947368	718671.8947	17096826.13	565.9390582	516489292284.64	
16		1996	2696826		28.78947368	869944.8947	25045255.65	828.833795	756804119878.70	
17		2001	2974807		33.78947368	1147925.895	38787811.81	1141.728532	1317733859807.38	
18		2006	3290350		38.78947368	1463468.895	56767188.18	1504.623269	2141741205862.27	
19		2011	3645257		43.78947368	1818375.895	79625723.39	1917.518006	3306490894560.01	
20		2016	4067175		48.78947368	2240293.895	109302760	2380.412742	5018916734795.17	
21	**AVERAGE**	1967.210526	1826881.105263	**SUM**			688137813.579	21081.15789	24198842058863.80	
22	POWER(Ave x,2)	3869917.255								
23	**Correlation Coefficient r**	0.9634538137699780								

The value of sample correlation coefficient (r) was to be 0.9634538137699780

2. Sample calculation for 95% CI for population correlation coefficient (ρ)

From question 1, the value of r is 0.9634538137699780

Estimating 95% confidence interval

$$W = \frac{1}{2}\ln\frac{1+r}{1-r}$$

The quantity (W) $=\frac{1}{2}\ln\left(\frac{1+0.96345}{1-0.96345}\right)=1.991942$

$$\sigma_W^2 = \frac{1}{n-3}$$

Sample size n= 19

$$\sigma_W^2 = \frac{1}{19-3} = \frac{1}{16} = 0.0625$$

Square root of variance (σ) $= \sqrt{\sigma_W^2} = \sqrt{0.0625} = 0.25$

95% confidence interval for μw

$$z_{\alpha/2} = z_{0.025} = 1.96$$

$W - z_{\frac{\alpha}{2}}\sigma_W < \mu_W < W + z_{\frac{\alpha}{2}}\sigma_W$

- ⇨ $1.991942 - (1.96)(0.25) < \mu_W < 1.991942 + (1.96)(0.25)$
- ⇨ $1.501942 < \mu_W < 2.481942$

95% confidence interval for ρ

$$\frac{e^{2(1.501942)}-1}{e^{2(1.501942)}+1} < \rho < \frac{e^{2(2.481942)}-1}{e^{2(2.481942)}+1}$$

- ⇨ $0.90549 < \rho < 0.986125$

3. Sample calculation for 95% CI for population intercept (β_0) $\hat{\beta}_0 \pm t_{n-2,\alpha/2} \cdot s_{\hat{\beta}_0}$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Beta 0=-62387426.1537517

t=2.1098

Lower=-71524589

Upper=-53250264

4. Sample calculation for 95% CI for population slop (β_1) $\hat{\beta}_1 \pm t_{n-2,\alpha/2} \cdot s_{\hat{\beta}_1}$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$\alpha = 0.05$

Degree of freedom (n-2) = 19-2=17

Beta 1 =32642.316

t=2.1098

Lower=27998.2511

Upper=37286.3806

Calculations for question 3 and 4

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{(1-r^2)\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-2}}$$

$$= \sqrt{\frac{(1-0.963^2)\,(2419884205\ 8863.80)}{19-2}}$$

$$= 319598.1143$$

$$s_{\hat{\beta}_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$= 319598.1143\sqrt{\frac{1}{19} + \frac{3869917.25}{21081.16}}$$

$$= 4330819.3$$

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \frac{319598.1143}{\sqrt{21081.16}} = 2201.187195$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{68813781 3.579}{21081.16} = 32642.316$$

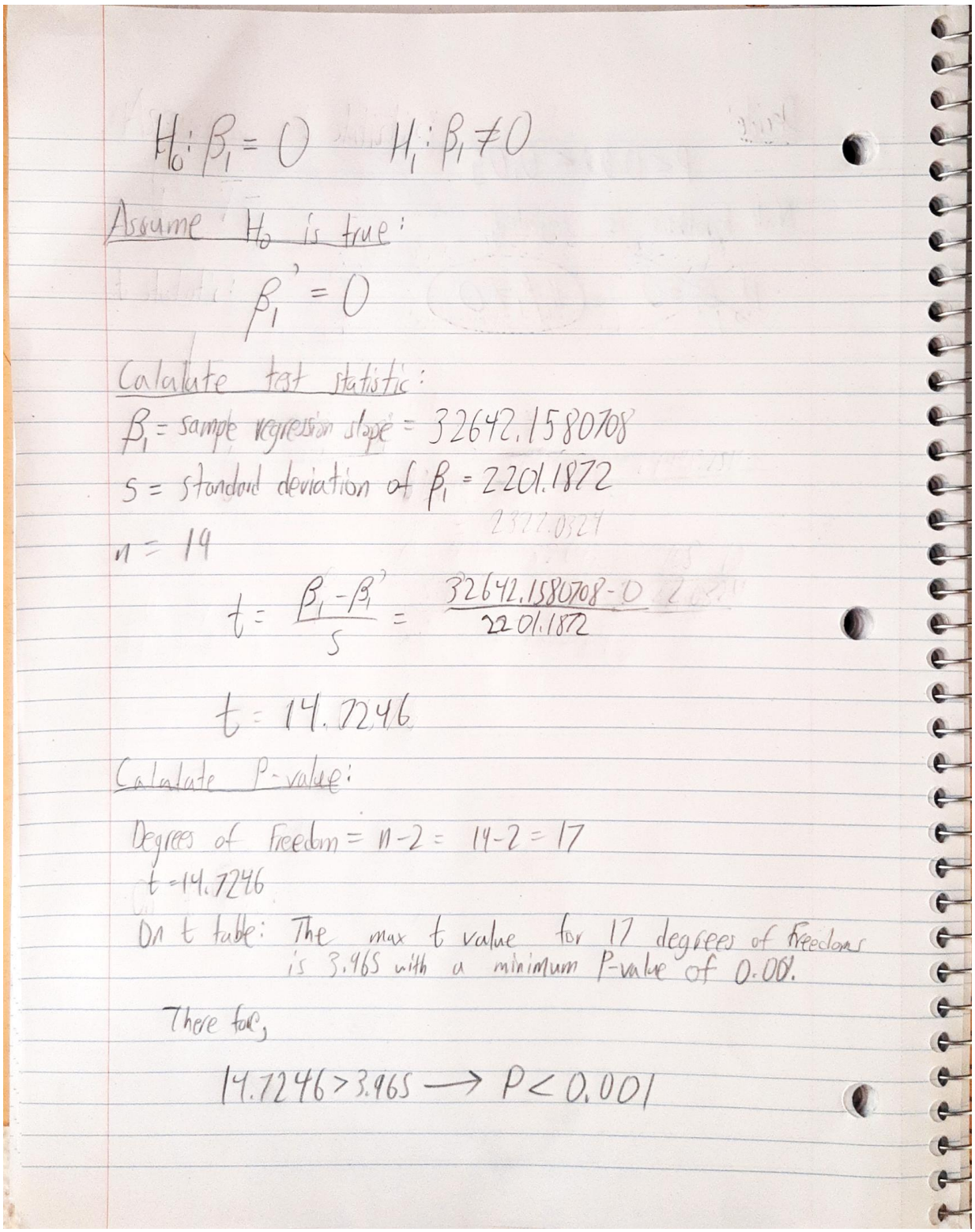
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 1826881.1 - 32642.316 \times 1967.21053$$

$$= -62387426.15$$

$$t_{n-2,\alpha/2} = t_{17,0.025} = 2.1098$$

- $\hat{\beta}_0$
  - Upper: $\hat{\beta}_0 + 2.1098 \times 4330819.3$
    $= -53250264$
  - Lower: $\hat{\beta}_0 - 2.1098 \times 4330819.3$
    $= -71524589$
- $\hat{\beta}_1$
  - $t_{n-2,\alpha/2} = t_{17,0.025} = 2.1098$
  - Lower: $\hat{\beta}_1 - 2.1098 \times 2201.1872 = 27998.2511$
  - Upper: $\hat{\beta}_1 + 2.1098 \times 2201.1872 = 37286.3806$

5. Sample calculation for testing the hypothesis $\beta_1 \neq 0$



$H_0: \beta_1 = 0 \quad H_1: \beta_1 \neq 0$

Assume H_0 is true:

$\beta_1' = 0$

Calculate test statistic:

β_1 = sample regression slope = 32642.1580708

s = standard deviation of β_1 = 2201.1872

$n = 14$

$$t = \frac{\beta_1 - \beta_1'}{s} = \frac{32642.1580708 - 0}{2201.1872}$$

$$t = 14.7246$$

Calculate P-value:

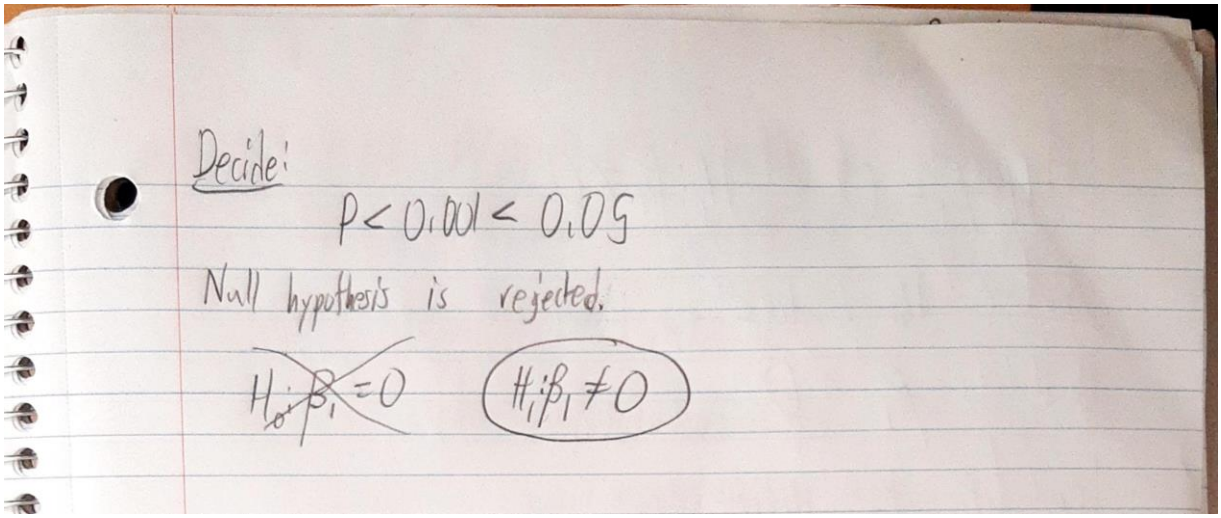
Degrees of Freedom = $n - 2 = 14 - 2 = 12$

$t = 14.7246$

On t table: The max t value for 12 degrees of freedom is 3.965 with a minimum P-value of 0.001.

There for,

$$14.7246 > 3.965 \rightarrow P < 0.001$$



6. Sample calculation for 95% CI for $(\beta_0 + \beta_1 x)$ for a chosen value of x

$t_{n-2,\alpha/2} = t_{17,0.025} = 2.1098$

$x = 2040$

$y = 4202898.0927$

$s = 1159474.098$

$$S_{pred} = s\sqrt{\frac{1}{n} + 1 + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$= 1159474.098\sqrt{\frac{1}{19} + 1 + \frac{(2040-1967.211)^2}{21081.1579}}$$

$= 364953$

→ The lower 95% CI of the population in 2040 is

$\hat{\beta}_0 + \hat{\beta}_1 x - t_{n-2,\alpha/2} S_{pred} = 4202898 - 364953 \times 2.1098 = 3432920$

The upper 95% CI of the population in 2040 is

$\hat{\beta}_0 + \hat{\beta}_1 x + t_{n-2,\alpha/2} \cdot S_{pred} = 4202898 + 364953 \times 2.1098 = 4972876$

→ 95% CI for the population of Alberta in 2040 is (3432920, 4972876)

APPENDIX B: INDIVIDUAL CONTRIBUTIONS

Please add few key words to reflect the major contributions made by each of the members in the group.

No	Last Name	First Name	Major Contributions	Signatures
1	Huynh	Hy	The regression model, interpretation of slope and intercept	Hy Huynh
2	Huynh	Ryan	Goodness-of-fit, hypothesis tests, and related sample calculations	Ryan Huynh
3	Nguyen	Hao	Prediction and Checking Models Assumption, Appendix A Part 6	Hao Nguyen
4	Trinh	Phuong	Appendix A Part 1 to 4	Phuong Trinh