The objective of this project is to gain some practice in conducting simple linear regression using Excel and gaining a deeper understanding on some key concepts.

Problem 1:

Grade Point average: The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). The results of the study are available in the file gradepoints.txt

Assume that the simple regression model is appropriate. Then

a) Obtain the least squares estimates of the offset and the slope and state the estimated regression function.

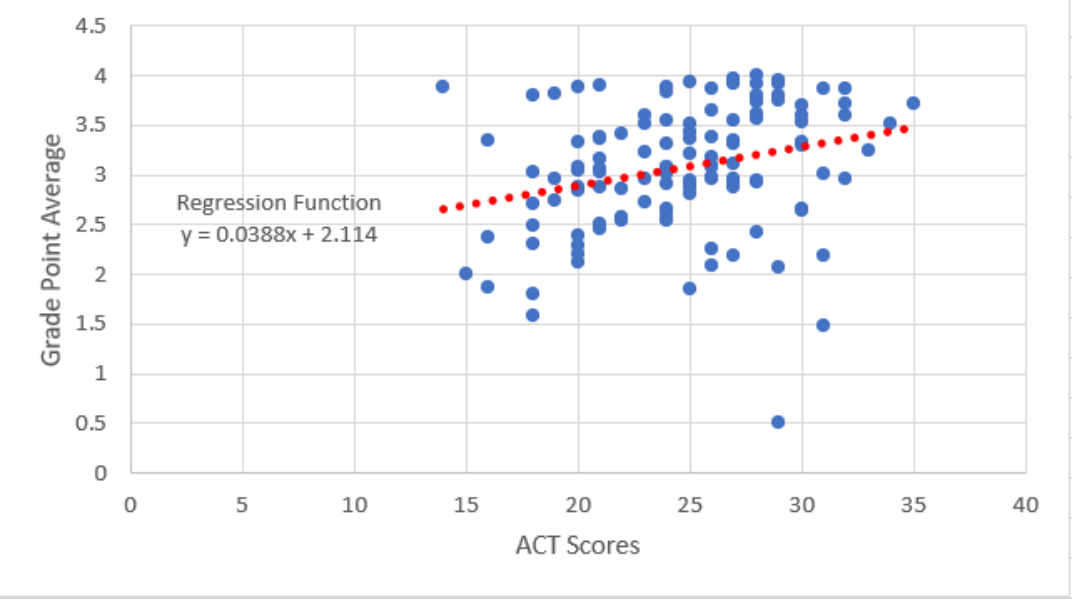
The least square estimate offset is 2.114 and the slope is 0.039. The estimated regression function is

y = 0.039x + 2.114

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.269481803							
R Square	0.072620442							
Adjusted R Square	0.064761293							
Standard Error	0.623125037							
Observations	120							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	3.587845899	3.587846	9.240243	0.002916604			
Residual	118	45.8176078	0.388285					
Total	119	49.4054537						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	2.114049287	0.320894835	6.587982	1.3E-09	1.478590151	2.749508423	1.273902675	2.954195899
X Variable 1	0.038827127	0.012773019	3.039777	0.002917	0.013533071	0.064121182	0.005385614	0.07226864

b) Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?

The estimated regression function appears to fit well with the data because the data is concentrated very closely to the regression function. In addition, it also fits well as some of the data that is scattered across the chart seem to also be near the trend.



c) Obtain a point estimate of the mean freshman GPA for students with ACT test score X=30

3.284 since y=0.039(30) + 2.114

d) What is the point estimate change in the mean response when the entrance test score increases by one point?

When the test score increases by 1 point, the point estimate change is 0.039

e) Obtain the residuals. Do they sum to 0?

When all the residuals are obtained, they do indeed sum up to 0.

f) Estimate $\sigma\sigma 2$. What is the unit in which $\sigma\sigma$ is measured?

ACT (X)		*GPA (Y)*	
Standard Deviation	4.47206549	Standard Deviation	0.644338
Sample Variance	19.99936975	Sample Variance	0.415172

g) Obtain a 99% confidence interval for $\beta\beta_1$? Interpret your confidence interval? Why might the director of admissions be interested in whether the confidence interval contains 0?

The director of admissions may be interested in whether the confidence interval contains a 0 because if there was a 0 included then there would be no evidence of a linear relationship between GPA and ACT scores.

Lower 99.0%	*Upper 99.0%*
1.273902675	2.954195899
0.005385614	0.07226864

h) Mary obtained a score of 28 on the entrance test. Predict her freshman GPA and interpret your results.

```
=2.114+0.0388*(28)
=1.96*(0.6271)
```

3.2004 ± 1.229116

1.971284	<	Estimated GPA	<	4.429516

Problem 2:

CDI Data: The dataset provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the dataset has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

Refer to the CDI data. The number of active physicians in a CDI (Y) is expected to be related to the total population, number of hospital beds, and total personal income. Assume that a simple regression model is appropriate for each of the three predictor variables.

a) Regress the number of active physicians in turn on each of the three predictor variables (one at a time). State the estimated regression function.

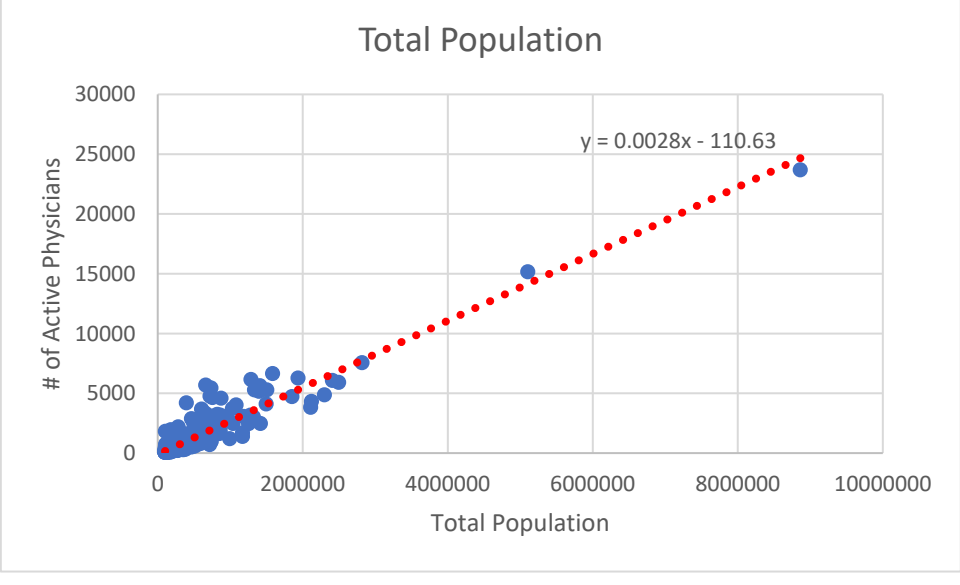
Total Population: Y = 0.0028x – 110.63

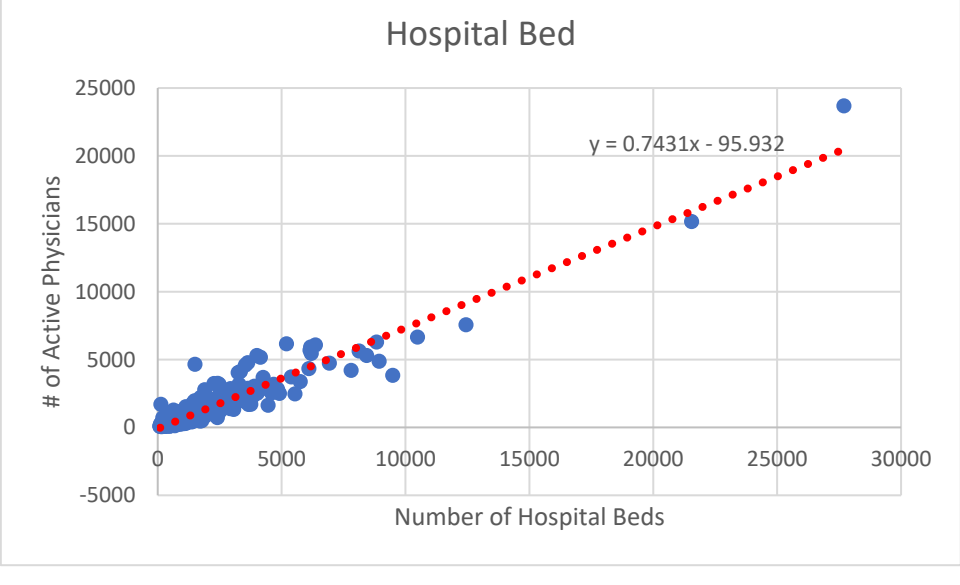
Number of Hospital Beds: Y = 0.7431x – 95.932

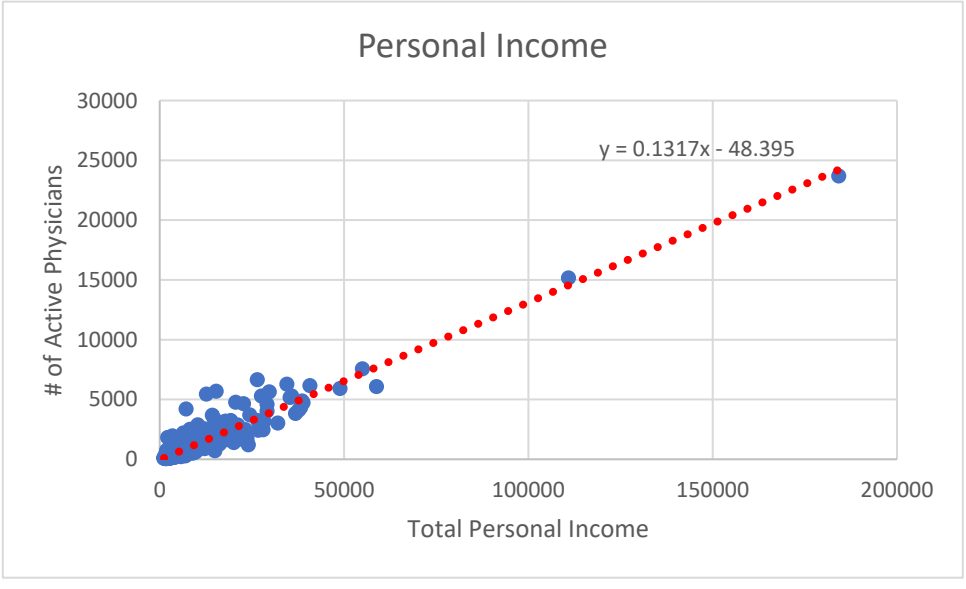
Total Personal Income: Y = 0.1317x – 48.395

b) Plot the three estimated regression functions and data on three plots. Does a linear regression relation appear to provide a good fit for each of the three predictor variables?

The linear regression relation appears to provide a good fit for each of the three predictor variables as most of the data points are concentrated near the regression function. In addition, they are also good fits for each of the three predictor variables as most of the data points that are scattered seem to trend near the regression function as well.







c) Calculate the total sum of square errors, the residual sum of square errors, and the coefficient of determination. Which regression equation is the best fit?

Total sum of square errors: 1,406,206,299

Residual sum of square errors:

- **Total Population: 163,025,135.15**
- **Hospital Beds: 135,864,044.99**
- **Personal Income: 142,148,254.43**

Coefficient of determination:

- **Total Population: 0.8841**
- **Hospital Beds: 0.9034**
- **Personal Income: 0.8989**

All of the regression equations are best fit.