Analyzing Linear Regression with EXCEL

This example is based on 27 college students. The independent variable (x) is SAT score and the dependant variable (y) is GPA. We are interested in understanding if a student's GPA can be predicted using their SAT score

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.440925
R Square	0.194415
Adjusted R Square	0.162192
Standard Error	0.443401
Observations	27

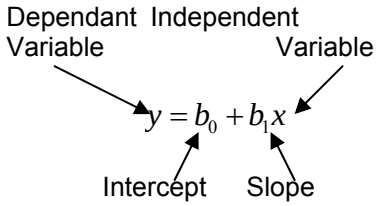
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression (SSR)	1	1.186183794	1.186184	6.033352	0.02133
Residual (SSE)	25	4.915110725	0.196604		
Total (SST)	26	6.101294519			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.355046	0.551989845	2.454838	0.021399	0.218201	2.49189	0.218201216	2.491889929
SAT SCORE	0.001412	0.000574954	2.456288	0.02133	0.000228	0.002596	0.000228113	0.002596391

Linear Regression Analysis

- Regression Line General Form:



- The sample slope, b_1, is given in the "*Coefficients"* column and the "SAT SCORE" row
  - b_1 measures the estimated change in y as a result of a one unit change in x
  - In this example, GPA increases by .00141 for every one point increase in SAT score.
- The intercept, b_0, is given in the "*Coefficients"* column and the "Intercept" row
  - b_0 is the estimated value of y when x is 0
- For this example, the regression line is: $y = 1.355 + .00141x$

Evaluating the Fitness of the Model Using Regression Statistics

- Multiple R – This is the correlation coefficient which measures how well the data clusters around our regression line. The closer this value is to 1, the more "linear" the data is. That is, we could use SAT SCORE to predict to predict GPA. If this value is close to 0 there is no linear relationship between our variables.
- R Square – This is the coefficient of determination. This measures the percentage of variation in the dependant variable that can be explained by the linear relationship between x and y. That is, how accurate the linear regression model is at predicting the GPA of a student based on their SAT Score.

- The total variation in y (dependant variable) is made up of two parts:
  - SST – total variation in y
  - SSR – variation explained by the linear relationship between x and y
  - SSE – variation associated with other factors besides the linear relationship
  - SST = SSE + SSR
  - Based on these, $R^2 = \frac{SSR}{SST}$

Evaluating the Fitness of the Model Using Confidence Intervals

- To determine how accurate our model is, we find a confidence interval for the slope of our regression line.
- If the confidence interval contains 0, then we have significant evidence to believe that there is not a linear relationship between x and y
- The (1-α) confidence interval is calculated using:
  (point estimate) ± (*t*-critical value)(standard error)
  with n-2 degrees of freedom
- For 95% confidence intervals, we can read the end-points straight from the output summary
  - The lower endpoint is given in the "*Lower 95%"* column and the "SAT SCORE" row
  - The upper endpoint is given in the "*Upper 95%"* column and the "SAT SCORE" row
  - In this example, we are 95% confident that the true increase in GPA for a point increase in SAT score is between .000228 and .002596.
  - Since 0 is not contained in our interval, we have reason to believe there is a linear relationship between GPA and SAT scores at the .05 level of significance

Evaluating the Fitness of the Model Using Hypothesis Testing

- We hope to answer the question, "Does a linear relationship exist between x and y?"
  - If there is no linear relationship between these variables, b_1=0. If there is a linear relationship, then $b_1 \neq 0$
- Determine if there is overwhelming evidence at the α=.05 level of a linear relationship between GPA and SAT score
  - $H_0 : \beta_1 = 0$
  - $H_\alpha : \beta_1 \neq 0$
  - α=.05
  - We'll use a *t*-test with n-2 degrees of freedom. Our critical values are 2.06 and -2.06
  - Our sample test statistic is given in the "*t Stat"* column and the "SAT SCORE" row
    - *t*-test statistic = 2.46
  - Since 2.46 > 2.06 we reject H_0 in favor of the alternative hypothesis
  - Therefore, there is overwhelming evidence at the α=.05 level that $\beta_1 \neq 0$. So we find it reasonable to believe that there **is** a linear relationship between GPA and SAT score. The strength of this relationship can be analyzed using our correlation coefficient and coefficient of determination.
- The simplest method to test any hypothesis is the *P*-value method
  - The *P*-value is the probability of observing a test statistic more extreme than what we observed (assuming that the null hypothesis is true)
  - The *P*-value is given in the "*P-value"* column and the "SAT SCORE" row
  - The null hypothesis is rejected if the *P*-value < α
  - In this example, *P*-value=.021 < α, therefore we reject the null hypothesis (just as above)