Notes on regression and correlation

1. Linear regression models the relationship between a dependent variable (DP) —the variable that is being predicted-- and one or more independent variable (IP) ---explanatory variables denoted X_1, X_2, etc.

2. R: coefficient of correlation: describes the strength of the relationship between the variables.

   $$r = \frac{Sxy}{\sqrt{Sxx \times Syy}}$$

   $$Sxy = \sum (x-\bar{x})(y-\bar{y})$$

   $$Sxx = \sum (x-\bar{x})^2$$

   $$Syy = \sum (y-\bar{y})^2$$

   Note: TI83/84: In order for LinReg test to display r and r^2, press 2^{nd}, then zero (Catalog), scroll down and select DiagnosticOn. To restore Lists, STAT, 5, SetUpEditor.

3. R^2 : coefficient of determination: the proportion of the total variation in the independent variable Y that is explained or accounted for, by the variation of the IP, X.

   $$r^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} = \frac{SSR}{SST}$$

4. T-test for coefficient of correlation: (given by TI & Casio graphing calculators under LinearRegTtest. Not given in the excel regression summary output)): Null, H_0 : rho=0 Alternative, H_a: rho≠0

   $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

5. S_{yx}: The standard error of the estimate: a measure of the dispersion of the observed values around the regression line. In excel summary output appears as Standard error. It is also equal to the square root of the mean square of errors, which also appear in the excel regression output (ANOVA table):

   $$\sqrt{MSE}$$

   If Syx is small, the data is closed to the regression line. In Calculators, appears as *se* or *s*, under the LinearRegTtest. We expect low Syx, high R in a good regression model.

6. In linear regression each y_hat (predicted y) is the mean of the normal distribution of predicted y obtained by a given x, when the experiment is repeated. The standard deviation of the distribution is the standard error of the estimate; then: y_hat ±Syx account for about 68% of the observations; y_hat ±2Syx, account for about 95% of the observations, etc; following the empirical rule.

7. Confidence interval reports the mean Y for a given X (CI estimate population parameters). This is the CI for the mean value of y at $x=x_p$.

   $$\hat{Y} \pm t_{n-2} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i-\bar{X})^2}{\sum(X_i-\bar{X})^2}}$$

   t_{crit} is based on n-2 df.

8. Prediction intervals reports the range of values of Y for a particular value of X: This is the PI for an individual new value of y at $x=x_p$. (A 1 is added inside the root):

   $$\hat{y} \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_{xx}}}$$

Multiple regression:

9. Adjusted R^2: in order to balance the effect that the number of independent variables (regressors) has on R. (R-Squared tends to overestimate the strength of the association between the variables). It appears on excel summary output.

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1} \right]$$

10. Despite their widely use of R^2 and R^2adj, they are only sample statistics. It is risky to judge the usefulness of a model on the basis of these values. That is why it is necessary to conduct a test of hypothesis involving all Beta parameters in the model. This is the F, global test: H_0: Beta1=Beta2=Beta3 etc versus H_a: at least one of the coefficient is nonzero.

$$F = \frac{MSR}{MSE}$$

11. Testing individual regression coefficients: It is a *t* test, where S_b is the standard error for the coefficient. See excel summary output. The p_{value} determines whether or not the null hypothesis can be rejected (H_0: $\beta=0$):
$t = \beta / s_b$

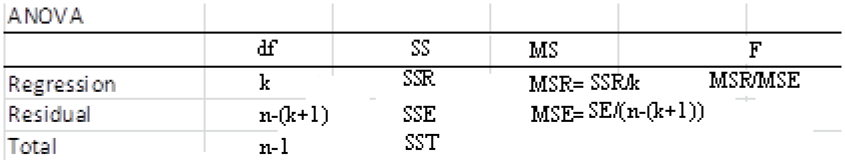
12. Confidence interval for coefficients, β, in the regression equation. In a regression model each coefficient estimates the change in the mean response per unit increase in X.
The closed interval around the population regression coefficient is calculated by: $\beta \pm t_{crit}$. SE (SE is the standard error of the coefficient given by excel regression coefficients table).
The confidence interval is useful to assess the reliability of the estimate of the coefficient. The wider the confidence interval, the less precise the estimate is.

13. Variance inflation factor: R^2_j refers to the coefficient of determination when a given IP is used as DP.

$$VIF_i = \frac{1}{1-R_i^2}$$

This is how minitab help defines VIF: Variance inflation factors measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. Use to describe how much multicollinearity (correlation between predictors) exists in a regression analysis. Multicollinearity is problematic because it can increase the variance of the regression coefficients, making them unstable and difficult to interpret.
If VIF = 1, predictor are not correlated; VIF<5, moderately correlated; VIF>5, highly correlated.
If the Independent variables are highly correlated themselves, the IP should be removed from the model.

14. Excel ANOVA table:



ANOVA				
	df	SS	MS	F
Regression	k	SSR	MSR= SSR/k	MSR/MSE
Residual	n-(k+1)	SSE	MSE= SE/(n-(k+1))	
Total	n-1	SST		