Multiple regression cheat sheet

Developed by Alison Pearce as an attendee of the ACSPRI Fundamentals of Regression workshop in June 2012, taught by David Gow.

Baby Statistics

Mean	μ or $\bar{X}$	$\sum X_i - \bar{X}) = 0$	- Value where the sum of the deviations is equal to zero
Variance	s^2 or σ^2	$s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$	- Larger values = larger spread - Value itself cannot be interpreted easily
Standard deviation	s or σ	$s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$	- In original units of the X variable - Larger std dev = more spread of data
z-scores / standardized scores		$\frac{X_i - \bar{X}}{S_x}$	- Transforms value to have mean = 0 and standard deviation = 1 - Does NOT change the distribution to be normal
Skew	sk	$sk = \frac{\sum z_i^3}{n-1}$	- 0 means distribution is symmetric - Usually a score between -7 and +7 - Positive sk indicates +ve skewed data - Negative sk indicates –ve skewed data
Kurtosis	ku	$ku = \frac{\sum x_i^4}{n-1} - 3$	- If '-3' is included in the formula then the ku of a normal distribution =0
Mean deviations		$(X_i - \bar{X})$	- Used to calculate mean and Z-scores (and then skew and kurtosis)
Squared mean deviations		$(X_i - \bar{X})^2$	- Used for variance and standard deviations
Rule of 2-sigma			- In a normal distribution, ○ 68% will fall within +/- 1 std dev ○ 95% will be within +/- 2 std dev ○ 99.7% will be within +/- 3 std dev

Bivariate Relationships

Covariance	Syx Cov	$$\frac{\sum[X-\overline{X})(Y_i-\overline{Y})]}{n-1}$$	- Extent to which values of 2 variables are associated - Increased association = positive covariance - Less association (ie many mismatched pairs) = negative covariance
Pearsons product moment correlation coefficient	r	$$r=\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2}}$$	- Value between -1 and +1 - 0 = no correlation, +1 = perfect positive correlation, -1 = perfect negative correlation - Symmetric distribution - How well the data points 'hug' the regression line – ie goodness of fit
Regression model		$Y_i = a + bX_i + e_i$	- In SAS the components of the Regression model are called parameter estimates
Line slope / Regression coefficient	b	$$b=\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2}$$	- Least squares method - Interpret as "For each 1 unit increase in X there is a b unit increase in Y" - Is impact of Independent variable on dependent - Assymetric, and can take any value
Line intercept	a b_0	$a = \bar{Y} - b\bar{X}$	- Least squares method - Intercept is the constant in the model
Predicted values	$\hat{Y}$	$\hat{Y} = a + bX$	- Predicted values based on regression line - "fitted value"
Residual	e	$e_i = Y_i - \hat{Y}_i$	- Variation in Y not explained to by changes in X
Standardised regression	b* β	$\beta = b_{yx} \times (\frac{S_x}{S_y})$	- Same as regression coefficient, but unit of measurement is standard deviation
ANOVA			
Total Sum of Squares	TotSS	$\sum(Y_i - \bar{Y})^2$	- Amount of variation in the Y data
Regression sum of squares	RegSS ModSS	$\sum(\hat{Y}_i - \bar{Y})^2$	- Amount of variation in Y explained by our model (variation in X)
Error sum of squares	ErrSS	$\sum(Y_i - \hat{Y}_i)^2 = \sum e_i^2$	- Yi – Y-hat is the error, so formula can be simplified - Variation which is unexplained by the model

Tests of Model Goodness of Fit			
Coefficient of determination/ R^2	R^2	$1-\frac{ErrSS}{TotSS}=\frac{RegSS}{TotSS}=\frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2}$	- Proportion of variation in Y explained by the model - Value between 0 and 1, but usually expressed as a % - "X% of variation in Y can be explained by X" - Most common measure of goodness of fit - Can also use R (square root of R), but not as easy to interpret
Adjusted R^2	$\overline{R^2}$ or R^2_{adj}	$1-\frac{ErrSS/(n-2)}{TotSS/(n-1)}$	- Makes more sense for multivariate analysis, because the degrees of freedom is adjusted for number of variables in model - In bivariate analysis usually similar to R^2, especially when n>100, as differences are very small
Standard Error of the Estimate (Root Mean Standard Error)	SEE RMSE	$\sqrt{\frac{\sum e_i^2}{n-2}}$	- Is the standard deviation of the residuals - Expressed in the units of measurement of the dependent variable - Because it is a standard deviation, if you assume the distribution is normal, then you can use the 2-sigma rule. Ie: able to say we can assume that 68% of values will lie within +/- SEE; 95% of values will be +/- 2xSEE. - Preferred measure of goodness of fit

Statistical Inference

Expected mean of repeated sample means	E	$E(\bar{X}) = \mu$	- Central limit theorem states that if multiple samples are drawn and the mean calculated, the average of these means will be centred around true mean of the population
Test statistic for sample means	t	$t = \frac{\bar{X} - \mu}{SE(X)}$	- Tests if a sample mean, $\bar{X}$ is consistent with an hypothesized value μ
Standard error of the mean	SE($\bar{X}$)	$SE(\bar{X}) = \sqrt{\frac{\sigma^2}{n}}$	- Standard deviation of the sample means from multiple drawn samples
Standard error of the regression coefficient	SE(b)	$SE(b) = \sqrt{\frac{\sigma_e^2}{\sum(X_i - \bar{X})^2}} = \frac{\sum e_i^2}{n-2}$	- Standard deviation of the sample regression coefficient from multiple drawn samples - Requires the σ, which is the population variance, but because we don't / can't know this, we instead use the variance of the residuals of the sample - Reported in the units of the variable of interest
Test statistic for sample regression coefficient	t	$t = \frac{b - \beta}{SE(b)}$	- Tests if a sample regression coefficient, b, is compatible with an hypothesized value, β
Confidence interval	CI	$CI = b \pm SE(b) \times t_{crit}$	- Usually use 95% CI -