LINEAR REGRESSION FORMULA SHEET

Note: This is not a cheat sheet. You cannot use this during the exam.

EQUATIONS TO MEASURE RELATIONSHIPS BETWEEN VARIABLES

Sample mean of x	$\bar{x}$	$\frac{1}{n}\sum x_i$	Average x_i $E(x_i) = \bar{x}, \forall_i$
Sample variance of x	s_x^2 var(x)	$\frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$	Variation in x_i about $\bar{x}$
Sample standard deviation of x	s_x sd(x)	$\sqrt{s_x^2}$	Average distance of x_i from $\bar{x}$
Sample covariance of x and y	s_{xy} cov(x,y)	$\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{S_{xy}}{n-1}$	Measures direction of relationship: if y goes up or down when x goes up or down. Is <u>not</u> predictive!
Sample estimate of the correlation between x and y	r_{xy} corr(x,y)	$\frac{s_{xy}}{s_x s_y} = \frac{\text{cov}(x,y)}{s_x s_y}$	Correlation is the standardized (unit-less) version of covariance. The correlation ranges between -1 and +1, inclusively.
Not quite the variance (uppercase 'S')	S_{xx}	$\sum (x_i - \bar{x})^2 = \sum (x_i - \bar{x}) \cdot x_i$	Used in equations below. This is a mathematical "shorthand"; it doesn't have meaning in and of itself.
Not quite the covariance (uppercase 'S')	S_{xy}	$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x}) \cdot y_i$	Used in equations below. This is a mathematical "shorthand"; it doesn't have meaning in and of itself.

BIVARIATE REGRESSION EQUATIONS

y_i = sample value; $\bar{y}$ = average of the sample values; $\hat{y}$ = value predicted by model; $e_i = y_i - \hat{y}$ = residual

Two-variable linear model		Population model: $y = \alpha + \beta x + \varepsilon$ Sample model: $y = a + bx + e$ Estimated model: $\hat{y} = a + bx$	ε = unobservable errors e = residual errors No error term in the estimate!
OLS estimate for α	a	$\bar{y} - b\bar{x}$	a is the y-intercept
OLS estimate for β	b	$\frac{S_{xy}}{S_{xx}} = \frac{s_{xy}}{s_x^2}$	b is the slope of the regression line. A one-unit change in *x* causes a change of *b* in *y*. *b* is predictive!
Total sum of squared deviations **SST = SSE + SSR**	SST (TSS)	$\sum (y_i - \bar{y})^2$	How much the actual sample points vary from the sample average.
Explained sum of squared deviations **SSE = SST - SSR**	SSE (ESS)	$\sum (\hat{y}_i - \bar{y})^2 = b^2 \cdot \sum (x_i - \bar{x})^2 = b^2 \cdot S_{xx}$	How much the predicted points vary from the sample average.
Residual sum of squared deviations. **SSR = SST - SSE**	SSR (RSS)	$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$	How much the actual sample points vary from the predicted points.
Coefficient of determination	R^2	$\frac{SSE}{SST} = 1 - \frac{SSR}{SST}$	Proportion of the total sum of squares that is explained by *x*; how well *x* predicts *y*. Convert to percentage.
Standard error of the entire regression *An <u>estimate</u> of the population std error.*	se root MSE	$\sqrt{\frac{SSR}{N-(k+1)}}$	N = total observations k = # of independent variables <u>Note</u>: This is **not** the standard error of the population, but an estimate of the population standard error.
Estimated variance of the error	se^2	$\frac{SSR}{N-(k+1)}$	Measure of variance in *y* not explained by the variance in *x* for the entire model.

Variance of the estimator b	Var(b)	$\frac{se^2}{S_{xx}} = \frac{se^2}{(n-1)s_x^2}$	
Standard error of the estimator b	se_b	$\sqrt{\text{var}(b)}$	Standard error of the estimator for the x coefficient – i.e. the width of the estimator's sampling distribution.
Variance of the estimator a	Var(a)	$se^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$	
Standard error of the estimator a	se_a	$\sqrt{\text{var}(a)}$	Standard error of the estimator for the y intercept – i.e. the width of the estimator's sampling distribution.
Using statistical inference, how to tell if b is significantly different from β	t	$\frac{b-\beta}{se_b}$	Test statistic for testing the significance of the predicted x effect on y. Usually, set β=0

MULTIVARIATE REGRESSION EQUATIONS

Multivariate linear model		Population model: $y = \alpha + \beta x + \gamma z + \varepsilon$ Sample model: $y = a + bx + gz + e$ Estimated model: $\hat{y} = a + bx + gz$	ε = unobservable errors e = residual errors No error term in the estimate!
OLS estimate for α	a	$\bar{y} - b\bar{x} - g\bar{z}$	
OLS estimate for β	b	$\frac{S_{xy}S_{zz} - S_{zy}S_{xz}}{S_{xx}S_{zz} - S_{xz}^2}$	Holding constant all other explanatory variables in the model, a one-unit change in *x* causes a change of *b* in *y*. Depends on the relationship between *x* and *z*, and *z* and *y*.
OLS estimate for γ	g	$\frac{S_{zy}S_{xx} - S_{xy}S_{xz}}{S_{xx}S_{zz} - S_{xz}^2}$	Ceteris paribus, a one-unit change in *z* causes a change of *g* in *y*. Depends on the relationship between *z* and *x*, and *x* and *y*
Variance of the estimator b **Precision** This is also known as precision or efficiency: smaller var(b) means higher precision.	Var(b)	$\frac{se^2}{n \cdot s_x^2} \cdot \frac{1}{1 - r_{x,z}^2}$	$r_{x,z}^2$ is the sample correlation between x & z. If x and z are not correlated, we recover the bivariate formula (almost – need 'n-1' instead of 'n'). If x and z are perfectly correlated, the standard error blows up. In general, as x and z approach perfect correlation, se blows up.
Standard error of the estimator b	se_b	$\sqrt{\text{var}(b)}$	See comments above.
Omitted variable bias	E(b)	$\beta + \gamma \cdot \left(\frac{\text{cov}(x,z)}{\text{var}(x)}\right)$	The second term is the effect of *x* on z. It's magnitude of the bias. E(b) = β <u>if</u> γ= 0 or <u>if</u> cov(x,z) = 0 Not solving for numbers, just positive or negative signs which tell you "understated" or "overstated" effect of the omitted variable Note: var(x) is always positive.