STATISTICS CHEAT SHEET

GORDON MCDONALD, 2017

1. Least Squares Regression

Trying to estimate the coefficients β in $\mathbf{y} = \mathbf{X}\beta + \epsilon$ is done by

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{1}$$

This is analogous to finding the projection $\mathbf{y}* = \beta\hat{\mathbf{x}} + \mathbf{r}$ of the vector $\mathbf{y}$ on the vector $\mathbf{x}$:

$$\beta = \frac{\mathbf{x} \cdot \mathbf{y}}{\mathbf{x} \cdot \mathbf{x}} \tag{2}$$

If you wish to include a diagonal weight matrix $\mathbf{W}$ this is done by

$$\beta = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} \tag{3}$$

and in the case that $\mathbf{W} = \mathbf{I}$ this reduces to (1). One would ordinarily choose the weight matrix to be e.g. $\mathbf{W} = \mathrm{diag}(1/\sigma_i^2)$ if the uncertanties σ_i for each point y_i are known. Alternatively for survey data for example, $\mathbf{W} = \mathrm{diag}(n_i)$ if the population n_i of each district for which the statistic x_i was measured. In the first case, an estimate of the variance is given by the square residuals, divided by the number of points minus the number of parameters, e.g.

$$\hat{\sigma}^2 = \frac{1}{n-p}|\epsilon|^2 \tag{4}$$

$$= \frac{1}{n-p}\left|\mathbf{y} - \mathbf{X}\hat{\beta}\right|^2 \tag{5}$$

The covariance matrix $\mathbf{\Sigma}$ for β is given by

$$\mathbf{\Sigma} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\hat{\sigma}^2 \tag{6}$$

$$= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\frac{1}{n-p}\left|\mathbf{y} - \mathbf{X}\hat{\beta}\right|^2 \tag{7}$$

where n is the number of data points and p is the number of parameters, i.e. $\nu = n - p$ is the degrees of freedom. An estimate of the variance of each parameter β_i is given by the diagonal entries of $\mathbf{\Sigma}$,

$$\mathrm{Var}(\beta_i) = \mathbf{\Sigma}_{ii} \tag{8}$$

and assuming they are either normally distributed (for large n) or t-distributed for small n, we can work out confidence intervals for the parameter by using either of the following

$$\beta_i = \hat{\beta}_i \pm z^* \sqrt{\mathbf{\Sigma}_{ii}} \tag{9}$$

$$\beta_i = \hat{\beta}_i \pm t^* \sqrt{\mathbf{\Sigma}_{ii}} \tag{10}$$

where

$$p = \operatorname{erf}\left(\frac{z_p^*}{\sqrt{2}}\right) \tag{11}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\frac{-z_p^*}{\sqrt{2}}}^{\frac{z_p^*}{\sqrt{2}}} e^{-t^2} dt \tag{12}$$

$$= \Phi\left(z_p^*\right) - \Phi\left(-z_p^*\right) \tag{13}$$

is the confidence (probability) that the true value will lie in this range for the normal distribution, and t_p^* is defined similarly but for Student's t-distribution.
The variance of a particular estimated new value of $\hat{y}$ (for input vector $\mathbf{x}$) is given by

$$\operatorname{Var}(y) = \mathbf{x}^T.\mathbf{\Sigma}.\mathbf{x} \tag{14}$$

So, to work out the bounds on a given fit, use the covariance matrix thusly

$$\operatorname{Var}(\hat{y}) = \operatorname{diag}\left(\mathbf{X}\mathbf{\Sigma}\mathbf{X}^T\right) \tag{15}$$

so the standard deviation is given by

$$\sigma_{\hat{y}} = \sqrt{\operatorname{diag}\left(\mathbf{X}\mathbf{\Sigma}\mathbf{X}^T\right)} \tag{16}$$

and the $100p\%$-confidence interval is given by

$$\hat{y} \pm t_p^* \sigma_{\hat{y}} \tag{17}$$

which is a function of x.

In the case that you have two or more disjoint subsets of points $(\mathbf{X}_a, \mathbf{y}_a)$ and $(\mathbf{X}_b, \mathbf{y}_b)$ each with estimates for parameters β respectively of β_a and β_b - the correct way to combine these to form an overall estimate is the following

$$\mathbf{X}_{\text{tot}}^T \mathbf{X}_{\text{tot}} = \mathbf{X}_a^T \mathbf{X}_a + \mathbf{X}_b^T \mathbf{X}_b \tag{18}$$

$$\beta_{\text{total}} = \left(\mathbf{X}_{\text{tot}}^T \mathbf{X}_{\text{tot}}\right)^{-1} \left(\mathbf{X}_a^T \mathbf{X}_a \beta_a + \mathbf{X}_b^T \mathbf{X}_b \beta_b\right) \tag{19}$$

Or for an arbitrary number

$$\mathbf{X}_{\text{tot}}^T \mathbf{X}_{\text{tot}} = \sum_i \mathbf{X}_i^T \mathbf{X}_i \tag{20}$$

$$\beta_{\text{total}} = \left(\mathbf{X}_{\text{tot}}^T \mathbf{X}_{\text{tot}}\right)^{-1} \sum_i \mathbf{X}_i^T \mathbf{X}_i \beta_i \tag{21}$$

this has the advantage of saving and or computing with only $p \times p$ symmetric matricies, where p is the number of parameters i.e. the length of β. So you only need to save $(p+2)p$ values. The computation of the inverse matrix will probably be the bottleneck and go as $\mathcal{O}(p^3)$.