Eric Slud, Stat 430 Fall 2008

Simple Linear Regression — Formulas & Theory

The purpose of this handout is to serve as a reference for some standard theoretical material in simple linear regression. As a text reference, you should consult either the Simple Linear Regression chapter of your Stat 400/401 (eg the currently used book of Devore) or other calculus-based statistics textbook (e.g., anything titled 'Engineering Statistics'), or a standard book on Linear Regression like

Draper, N. and Smith, H. (1981) **Applied Linear Regression**, 2nd edition. John Wiley: New York.

The model is

$$Y_i = a_0 + b_0 X_i + \epsilon_i, \quad 1 \le i \le n, \qquad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

where the constant real parameters (a_0, b_0, σ_0^2) are unknown. The parameters (a, b) are estimated by maximum likelihood, which in the case of normally distributed *iid* errors as assumed here is equivalent to choosing a, b in terms of $\{(X_i, Y_i)\}_{i=1}^n$ by **least squares**, i.e. to minimize $\sum_{i=1}^n (Y_i - a - bX_i)^2$, which results in:

$$\hat{b} \;=\; \frac{s.cov(\underline{X}, \underline{Y})}{s.var(\underline{X})} \;=\; \frac{s_{XY}}{s_X^2} \quad , \qquad \hat{a} \;=\; \bar{Y} - \hat{b}\bar{X} \tag{2}$$

where

$$\bar{X} \;=\; \frac{1}{n}\sum_{i=1}^n X_i \quad , \qquad \bar{Y} \;=\; \frac{1}{n}\sum_{i=1}^n Y_i \quad , \qquad s_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$$

and

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2 \quad , \qquad s_{XY} = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})$$

The **predictors** and **residuals** for the observed responses Y_i are given respectively by

$$\text{Predictor}_i \;=\; \hat{Y}_i \;=\; \hat{a} + \hat{b}X_i \quad , \qquad \text{Residual}_i \;=\; \hat{\epsilon}_i \;=\; Y_i - \hat{Y}_i$$

The standard (unbiased) estimator of σ^2 is the Mean Residual Sum of Squares (per degree of freedom) given by

$$\hat{\sigma}^2 \;=\; \text{MRSS} \;=\; \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Confidence intervals for estimated parameters are all based on the fact that the least squares estimates $\hat{a}$, $\hat{b}$ and the corresponding predictors of (the mean of) Y_i are linear combinations of the independent normally distributed variables $\epsilon_j,\ j = 1,\ldots,n,$ and the general formula for any sequence of constants $u_j,\ j = 1,\ldots,n,$

$$\sum_{j=1}^{n} u_j\,\epsilon_j \;\sim\; \mathcal{N}(0\,,\ \sigma^2\sum_{j=1}^{n} u_j^2) \tag{3}$$

We use this formula below with various choices for the vector $\underline{u} = \{u_j\}_{j=1}^{n}$.

Under the model (1), with true parameters (a_0, b_0, σ_0^2), we first calculate from (2) that

$$\begin{aligned}\hat{b} - b_0 \;&=\; \frac{\sum_{j=1}^{n}\{Y_j - \bar{Y} - b_0(X_j - \bar{X})\}(X_j - \bar{X})}{\sum_{j=1}^{n}(X_j - \bar{X})^2}\\ &=\; \frac{1}{(n-1)\,s_X^2}\sum_{j=1}^{n}(\epsilon_j + a_0 + b_0\bar{X} - \bar{Y})(X_j - \bar{X}) \;=\; \sum_{j=1}^{n} c_j\,\epsilon_j\end{aligned}$$

(since $\sum_{j=1}^{n}(X_j - \bar{X}) = 0$), where

$$c_j = \frac{X_j - \bar{X}}{(n-1)s_X^2} \quad \text{with sum of squares} \quad \frac{\sum_{j=1}^{n}(X_j - \bar{X})^2}{(n-1)^2\,s_X^4} \;=\; \frac{1}{(n-1)s_X^2}$$

Therefore, by (3), we have

$$\hat{b} - b_0 \;=\; \sum_{j=1}^{n}\frac{X_j - \bar{X}}{(n-1)\,s_X^2}\,\epsilon_j \;\sim\; \mathcal{N}\Big(0,\ \frac{\sigma^2}{(n-1)s_X^2}\Big) \tag{4}$$

Next, using

$$\hat{a} - a_0 \;=\; \bar{Y} - \hat{b}\bar{X} - a_0 \;=\; \frac{1}{n}\sum_{j=1}^{n}(\epsilon_j - (\hat{b} - b_0)\,\bar{X}) \;=\; \sum_{j=1}^{n} u_j\,\epsilon_j$$

where

$$u_j = \frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{(n-1)s_X^2} \quad \text{with sum of squares} \quad \frac{1}{n} + \frac{\bar{X}^2}{(n-1)\,s_X^2}$$

we find

$$\hat{a} - a_0 \;=\; \sum_{j=1}^{n} \Big(\frac{1}{n} - \frac{\bar{X}(X_j - \bar{X})}{(n-1)\,s_X^2}\Big)\,\epsilon_j \;\sim\; \mathcal{N}\Big(0,\, \sigma^2\{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}\}\Big) \tag{5}$$

Similarly,

$$\begin{aligned} \hat{a} - a_0 + (\hat{b} - b_0)\,X_i \;&=\; \sum_{j=1}^{n} \Big(\frac{1}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{(n-1)\,s_X^2}\Big)\,\epsilon_j \\ &\sim\; \mathcal{N}\Big(0,\, \sigma^2\,\Big\{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{(n-1)\,s_X^2}\Big\}\Big) \end{aligned} \tag{6}$$

and finally

$$\begin{aligned} Y_i - \hat{a} - \hat{b}X_i \;&=\; \sum_{j=1}^{n} \Big(\delta_{ji} - \frac{1}{n} - \frac{(X_i - \bar{X})(X_j - \bar{X})}{(n-1)\,s_X^2}\Big)\,\epsilon_j \\ &\sim \mathcal{N}\Big(0,\, \sigma^2\,\Big\{1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{(n-1)s_X^2}\Big\}\Big) \end{aligned} \tag{7}$$

where in the last display we have used the Kronecker delta δ_{ji} defined equal to 1 if $i = j$ and equal to 0 otherwise.

A further item of theoretical background is the expression of the sum of squared errors in a form allowing us to find that it is independent of $\hat{b}$ and is distributed as a constant times χ^2_{n-2}. For that, note first that

$$\text{SSE} \;=\; \sum_{j=1}^{n} (Y_j - \hat{a} - \hat{b}X_j)^2 \;=\; \sum_{j=1}^{n} \Big((Y_j - \bar{Y}) - \hat{b}(X_j - \bar{X})\Big)^2 \tag{8}$$

We used this equation in class to show, by expanding the square in the last summation, that

$$\text{SSE} \;=\; (n-1)\,s_Y^2\,(1 - \hat{r}^2) \quad , \qquad \hat{r} \;=\; \frac{s_{XY}}{s_X\,s_Y} \;= \hat{b}\,\frac{s_X}{s_Y}$$

Continuing with the formula (8) for SSE, we find via (4) that with $u_j = c_j = (X_j - \bar{X})/((n-1)s_X^2)$,

$$\begin{aligned} \text{SSE} &= \sum_{j=1}^{n} (\epsilon_j - \bar{\epsilon} - (\hat{b} - b_0)(X_j - \bar{X}))^2 \\ &= \sum_{j=1}^{n} \Big(\epsilon_j - \bar{\epsilon} - (X_j - \bar{X}) \sum_{k=1}^{n} \frac{X_k - \bar{X}}{(n-1)s_X^2}\, \epsilon_k\Big)^2 \\ &= \sum_{j=1}^{n} (\epsilon_j - \bar{\epsilon})^2 - \frac{1}{(n-1)\, s_X^2} \Big(\sum_{j=1}^{n} \epsilon_j\, (X_j - \bar{X})\Big)^2 \\ &= \mathbf{e}' \Big(I - \frac{1}{n}\mathbf{1}\mathbf{1}' - (n-1)s_X^2 \mathbf{c}\mathbf{c}'\Big)\mathbf{e} \end{aligned} \tag{9}$$

where $\bar{\epsilon} = n^{-1}\sum_{j=1}^{n} \epsilon_j$ and $\mathbf{e}$ denotes the n-vector with components ϵ_j. Since the (jointly) normally distributed variables $\mathbf{c}'\mathbf{e}$ and $\bar{\epsilon} = \mathbf{1}'\mathbf{e}/n$ and the components of $(I - \frac{1}{n}\mathbf{1}\mathbf{1}' - (n-1)s_X^2\mathbf{c}\mathbf{c}')\mathbf{e}$ are uncorrelated, they are actually independent, and the quadratic form (9) can be proved to have the property

$$\frac{\texttt{SSE}}{\sigma_0^2} \;=\; (n-2)\,\frac{\hat{\sigma}^2}{\sigma_0^2} \;\sim\; \chi^2_{n-2} \tag{10}$$

An important aspect of this proof is the observation that the matrix $\mathbf{M} = I - \frac{1}{n}\mathbf{1}\mathbf{1}' - (n-1)s_X^2\mathbf{c}\mathbf{c}'$ is a **projection matrix**, that is, is symmetric and *idempotent*, which means that $M^2 = M$, and the quadratic form (9) is equal to $(\mathbf{M}\mathbf{e})'(\mathbf{M}\mathbf{e})$. The independence of $(\hat{a},\, \hat{b})$ and $\mathbf{M}\mathbf{e}$ is confirmed by checking that the covariances are 0:

$$\text{Cov}(\mathbf{M}\mathbf{e},\, \mathbf{1}'\mathbf{e}) \;=\; \sigma_0^2\, \mathbf{M}\mathbf{1} \;=\; 0 \quad , \qquad \text{Cov}(\mathbf{M}\mathbf{e},\, \mathbf{c}'\mathbf{e}) \;=\; \sigma_0^2\, \mathbf{M}\mathbf{c} \;=\; 0$$

The independence of $(\hat{a}, \hat{b})$ and $\hat{\sigma}^2$ then immediately implies, by definition of the t-distribution, that

$$(\hat{b} - b_0)\,\sqrt{n-1}\,\frac{s_X}{\hat{\sigma}} \;\sim\; t_{n-2} \quad , \qquad \frac{\hat{a} - a_0}{\hat{\sigma}}\Big\{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)\, s_X^2}\Big\}^{-1/2} \;\sim\; t_{n-2} \tag{11}$$

Finally, we turn to the definitions of confidence intervals and the CLM and CLI confidence and prediction intervals constructed and plotted by `SAS`. The main ingredient needed in the justification of these is the result (11) just proved. The confidence and prediction intervals say that each of the following statements has probability $1-\alpha$ under the model (1):

$$b_0 \in \hat{b} \pm t_{n-2,\alpha/2} \frac{\hat{\sigma}}{s_X \sqrt{n-1}} \tag{12}$$

$$a_0 \in \hat{a} \pm t_{n-2,\alpha/2}\, \hat{\sigma} \Big(\frac{1}{n} + \frac{1}{(n-1)s_X^2}\Big)^{1/2} \tag{13}$$

$$a_0 + b_0 X_0 \in \hat{a} + \hat{b} X_0 \pm t_{n-2,\alpha/2}\, \hat{\sigma} \Big(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)s_X^2}\Big)^{1/2} \tag{14}$$

$$Y_i \in \hat{a} + \hat{b} X_i \pm t_{n-2,\alpha/2}\, \hat{\sigma} \Big(1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{(n-1)s_X^2}\Big)^{1/2} \tag{15}$$

The confidence intervals (12) and (13) are exactly as used by `SAS` in determining p-values for the significance of coefficients a, b (in testing the respective null hyptheses that $b = 0$ or that $a = 0$.) The interval (14) is what `SAS` calculates in generating the CLM upper and lower confidence limits that it calculates and plots at location X_i either within `PROC GPLOT` or `PROC REG`. The interval (15) is only a *retrospective* prediction interval within which we should have found Y_i with respect to its predictor $\hat{Y}_i$ but it is **NOT** what `SAS` calculates in generating the CLI upper and lower individual-observation prediction limits at location X_i either within `PROC GPLOT` or `PROC REG`.

Prediction intervals are meant to capture not the observations already seen but rather any new observations Y_i' which would be collected at the previous locations X_i or at brand-new locations X_0 . So we discuss next the corrected formula for prediction interval which SAS actually calculates.

We are interested sometimes, especially as part of diagnostic checking and model-building, in making prediction intervals for values Y_0 (not yet observed) corresponding to values X_0 which were not in fact part of the dataset. The thing to understand is that formula (15) is **not applicable** to this situation because it refers to observations Y_i which were already used as part of the model-fitting that resulted in $\hat{a}$, $\hat{b}$. If $Y_0 = a_0 + b_0 X_0 + \epsilon_0$ with a_0, b_0 the same as before but with $\epsilon_0 \sim \mathcal{N}(0, \sigma_0^2)$ *independent* of all

the data already used, then

$$Y_0 - \hat{a} - \hat{b}X_0 \sim \mathcal{N}\Big(0\,,\, \sigma^2 \Big\{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)\, s_X^2}\Big\}\Big) \tag{16}$$

and with probability $1 - \alpha$,

$$Y_0 \in \hat{a} + \hat{b}X_0 \pm t_{n-2,\alpha}\, \hat{\sigma} \Big(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)\, s_X^2}\Big)^{1/2} \tag{17}$$

So when we make prediction intervals for brand-new points not observed in the dataset, we use formula (17). Once more, to re-cap and correct a previous mis-statement: formula (17) not (15) is the one which `SAS` uses in calculating prediction intervals for `Proc Reg` output files with keywords `L95` or `U95` or which `Proc GPLOT` plots using the `I=RLCLI95` symbol declaration option.