Notes: Confidence & Prediction Intervals For OLS Regression

Trevor Dick

1 Overview

The purpose of these notes is to consolidate ideas around measuring error and finding confidence and/or prediction regions when obtaining the ordinary least squares estimator. The material is best suited for those who have sufficient vector calculus, linear algebra, and statistics knowledge. What is not explicitly defined or derived in these notes is assumed to be known or easily calculated. References may be added in the future if needed.

2 Setting up the Problem

Let m be the number of samples, and n be the number of parameters.

1. For any regression problem we want to obtain the optimal parameters $\hat{\beta}$, for the model function $\vec{y} = \vec{f}(\vec{x}, \vec{\beta})$
2. Obtaining variances can be important for a number of things, but here we will focus on the role variances play in prediction and confidence interval calculations:
   (a) Estimate population standard deviation of $\mathbf{y}$ with sample standard deviation ($\vec{e}$ is the vector of residuals).
   $$\sigma^2 \approx s^2 = MSE = \frac{SSE}{\nu} = \frac{\vec{e}^T\vec{e}}{m-n}$$
      1) When f is linear, s^2 is unbiased
      2) Non-linear, s^2 is biased, however, as $m \longrightarrow \infty$ then the bias approaches 0.
   (b) Variance of parameters $\vec{\beta}$; use if the covariance matrix, given by the least squares process used to find $\hat{\beta}$, is not provided. Where A and J both $m \times n$ matrices and J is the Jacobian of f with respect to $\vec{\beta}$.
      1) Linear (exact):
         - Non-weighted, $Var(\vec{\beta}) = \sigma^2(A^TA)^{-1}$

- Weighted, $Var(\vec{\beta}) = \sigma^2 BB^T$ (see Appendix)

2) Non-linear (approx): $Var(\vec{\beta}) \approx \sigma^2 (J^T J)^{-1}$

(c) Variance of the model function $f(\tilde{x}, \vec{\beta})$, at any new point $\tilde{x}$ (see following sections for details):

1) Linear (exact): $Var(f) = A_{\tilde{x}} Var(\vec{\beta}) A_{\tilde{x}}^T$

2) Non-linear (estimate): $Var(f) \approx J_{\tilde{x}} Var(\vec{\beta}) J_{\tilde{x}}^T$

3. Being able to calculate the critical t-value, for a given percent $100(1-\alpha)\%$ is also used in the prediction and confidence interval formulation: $t_{1-\frac{\alpha}{2},\nu} = \Phi^{-1}(1-\frac{\alpha}{2})$, where Φ is the CDF of the students t-distribution.

4. Prediction and confidence intervals are both given by, $\hat{y} \pm t_{1-\frac{\alpha}{2},\nu} \sqrt{Var(\hat{y})}$, for the optimal prediction $\hat{y}$.

1) Prediction interval: $Var(\hat{y}) = Var(f) + \sigma^2$

2) Confidence interval: $Var(\hat{y}) = Var(f)$

Note: the confidence intervals will always be smaller than prediction intervals across all values of x.

3 Model Concepts

In many cases the random variables or vectors and their operators can be formally summed to yield new random variables or operators.

Guiding paradigm:

$$\textbf{pred} = \textbf{model} + \textbf{actual}$$

Note: 'actual' could be actual errors, for example.

4 Common Steps

4.1 Defining the Model

Let $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}$ be the set of m observations (random vectors) that describes the set of points on which regression fitting techniques will be performed. The vector of n fit parameters is given by $\vec{\beta} = (\beta_1, ..., \beta_n)$.

- Model: $\mathbf{y} = f(\mathbf{x}, \vec{\beta})$
- Prediction (hat denotes optimality aka best fit): $\hat{y} = f(x, \hat{\beta})$

4.2 Optimality

Best curve fit will be assumed to be found under the following standard least squares condition (minimize the sum of square residuals), where $\vec{e}$ is the vector of residuals and W is a positive definite matrix ($W = I$ for unweighted least squares):

$$\min_{\vec{\beta}} \vec{e}^T(\vec{\beta}) W \vec{e}(\vec{\beta}) = \min_{\vec{\beta}} \left[(\vec{y} - \vec{f}(\vec{x}, \vec{\beta})) \cdot W (\vec{y} - \vec{f}(\vec{x}, \vec{\beta})) \right] \tag{1}$$

4.3 Linear vs. Non-linear

For the sake of consistent vector notation, and the remainder of these notes, the vectors $\vec{y}$ and $\vec{x}$ hold the m observations, which are still elements of the random vector $(\mathbf{x}, \mathbf{y})$.

1) Linear, where $\{f_i(x)\}_i$ is a set of linearly independent basis functions:

$$\vec{y} = A(\vec{x}) \cdot \vec{\beta} = \begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_n(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(x_m) & f_2(x_m) & \cdots & f_n(x_m) \end{bmatrix}_{m \times n} \cdot \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}_{n \times 1} \tag{2}$$

Note: When $f_1(x) = 1$ is the constant basis function then β_1 is value for the intercept.

- Special Case, a line ($y = \beta_1 + \beta_2 x$) in $\mathbb{R}^2$, where $f_2(x) = x$:

$$\vec{y} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \tag{3}$$

- Note:

$$\vec{y} = A(\vec{x}) \cdot \vec{\beta} = \sum \beta_i \vec{f_i}(\vec{x})$$

2) Non-linear (general):

$$\vec{y} = f(\vec{x}, \vec{\beta}) \tag{4}$$

Note: cannot generally be expressed as matrix products, as in the linear case, except in a few special cases, for example, when f is a bilinear form, scalar function, ($f = \vec{\beta}^T A \vec{\beta}$).

- Maclaurin Series first order approximation: define the $m \times n$ Jacobian matrix of $\vec{f}$ as the total derivative with respect to $\vec{\beta}$, evaluated at fixed

$\vec{x}$:

$$J = J(\vec{\beta}) = J_{\vec{x}}(\vec{\beta}) = D_\beta \vec{f} = \begin{bmatrix} \frac{\partial f(x_1,\vec{\beta})}{\partial \beta_1} & \cdots & \frac{\partial f(x_1,\vec{\beta})}{\partial \beta_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(x_m,\vec{\beta})}{\partial \beta_1} & \cdots & \frac{\partial f(x_m,\vec{\beta})}{\partial \beta_n} \end{bmatrix}_{m \times n} \tag{5}$$

$$\Longrightarrow$$

$$\vec{f}(\vec{x},\vec{\beta}) \approx J(\vec{0}) \cdot \vec{\beta} + \vec{f}(\vec{x},\vec{0}) \tag{6}$$

Notice when $\vec{f}(\vec{x},\vec{0}) = \vec{0}$, then this is now the linear case where $J_{\vec{x}} = A(\vec{x})$, and A is the best linear approximation of f!

– Second order approximation, where H is the Hessian tensor (like an array of m Hessian matrices, which when m=1 reduces to the standard Hessian matrix):

$$\vec{f}(\vec{x},\vec{\beta}) \approx J \cdot \vec{\beta} + \frac{1}{2}\vec{\beta}^T H \vec{\beta} \tag{7}$$

Note: use Einstein notation to resolve the tensor products.

4.4 More Details on Errors and Variances

- Let ϵ be the error for the model, recall our model paradigm:

$$\hat{y} = \vec{f}(\vec{x},\hat{\beta}) + \epsilon \tag{8}$$

When $\epsilon \sim \mathcal{N}(0,\sigma^2)$, then the OLS best estimator for $\vec{\beta}$ and the MLE best estimator are the same.

- Residuals $\vec{e} = (e_1, \cdots, e_m)$:

$$e_i = y_i - \hat{y}_i = y_i - f(x_i,\hat{\beta}) \tag{9}$$

- Sum of Squared Errors (also RSS):

$$SSE = \vec{e}^T\vec{e} \tag{10}$$

Weighted case:

$$SSE = \vec{e}^T W \vec{e}$$

- Mean Squared Error, let $\nu = m - n$ be the number of degrees of freedom:

$$MSE = \frac{SSE}{\nu} \tag{11}$$

Weighted linear case only, let $M = I - P_A$ and $P_A = A(A^TWA)^{-1}A^TW$:

$$MSE = \frac{\vec{e}^T W \vec{e}}{tr(WM)}$$

To prove this is unbiased see the Appendix.

- Estimate population variance (σ^2) of error through sample variance s^2:

$$s^2 = MSE = \frac{SSE}{\nu} \tag{12}$$

1) Linear, unbiased estimator:

$$\sigma^2 \approx s^2 \tag{13}$$

2) Non-linear, biased estimator, however as $m \to \infty$ bias approaches 0:

$$\sigma^2 \approx s^2 \tag{14}$$

Note: If using the maximum likelihood estimation, MLE, to find population variance of the errors, this too is a biased estimator. Define $\hat{\sigma}^2$ to be MLE estimate:

$$s^2 = \frac{m}{\nu}\hat{\sigma}^2$$

$$\iff$$

$$\hat{\sigma}^2 = \frac{SSE}{m}$$

- Variance of $\vec{\beta}$, $n \times n$ matrix, where $\sigma^2_{\beta_i} = [Var(\vec{\beta})]_{ii}$:

1) Linear (exact):

$$Var(\vec{\beta}) = \sigma^2 BB^T \tag{15}$$

2) Non-linear (estimate):

$$Var(\vec{\beta}) \approx \sigma^2 (J^T J)^{-1} \tag{16}$$

*TODO use Hessian to show a second order version

- Variance of the model function $f(\tilde{x}, \vec{\beta})$, at any new point $\tilde{x}$:

1) Linear (exact):

$$Var(f) = A_{\tilde{x}} Var(\vec{\beta}) A^T_{\tilde{x}} \tag{17}$$

Note: in weighted case only variance of β changes since the objective F is modified and not the model function f

2) Non-linear (estimate):

$$Var(f) \approx J_{\tilde{x}} Var(\vec{\beta}) J^T_{\tilde{x}} \tag{18}$$

- Error Propagation estimate by uncertainty δf (special case of general estimate where all β_i are uncorrelated and the variance matrix is just a diagonal matrix):

$$Var(f) \approx \delta f^2 \approx \sum_{i=1}^{m} (\frac{\partial f}{\partial \beta_i})^2 \delta \beta_i^2 \tag{19}$$

Note: $\delta\beta_i$ is the uncertainty for β_i and is estimated by σ_{β_i}, coming from the diagonals of the variance matrix, if individual measurements (observations in the parlance of physics) are not available.

5 Confidence Interval

The confidence interval is defined for variance of the prediction, when considering variance only coming from variance of the model:

$$Var(\hat{y}) = Var(f) \tag{20}$$

The confidence interval is defined:

$$\hat{y} \pm (t_{1-\frac{\alpha}{2},\nu})\sqrt{Var(\hat{y})} \tag{21}$$

6 Prediction Interval

The prediction interval is defined for variance of the prediction, when considering variance coming from both variance of errors and the model:

$$Var(\hat{y}) = Var(f) + \sigma^2 \tag{22}$$

The prediction interval is defined:

$$\hat{y} \pm (t_{1-\frac{\alpha}{2},\nu})\sqrt{Var(\hat{y})} \tag{23}$$

7 Uncertainties of Measurement

The uncertainty of a measurement, denoted δz for some variable z, can be computed in the case where standard deviation is unknown, like in the case of regression curve fitting. This is achieved through the same critical score used in the students t-test. The uncertainty of a measurement is then defined as $\delta z = t_{1-\frac{\alpha}{2},\nu}\sqrt{Var(z)}$.

7.1 Students t-distribution

For ν degrees of freedom the probability distribution function (pdf) is defined:

$$f_\nu(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{24}$$

The cumulative distribution function (cdf) is defined, in general, for continuous pdf's, let Φ be the cdf:

$$\Phi(x) = \int_{-\infty}^{x} f_\nu(t)dt \tag{25}$$

7.2 Critical t-value (2 sided t-test)

Define $100(1-\alpha)\%$ to be the percent confidence. Let $p = 1 - \frac{\alpha}{2}$, where p is the percentile of the probability distribution.

Then define $t_{1-\frac{\alpha}{2},\nu}$ to be the critical t-value when the following condition of probability calculation is satisfied:

$$Pr(-t_{1-\frac{\alpha}{2},\nu} < t < t_{1-\frac{\alpha}{2},\nu}) = 1 - \alpha \tag{26}$$

$$\iff$$

$$Pr(t < t_{1-\frac{\alpha}{2},\nu}) = p$$

$$\iff$$

$$\Phi(t_{1-\frac{\alpha}{2},\nu}) = 1 - \frac{\alpha}{2}$$

$$\iff$$

$$t_{1-\frac{\alpha}{2},\nu} = \Phi^{-1}(1 - \frac{\alpha}{2}) = \Phi^{-1}(p) \tag{27}$$

Appendix

A Useful Equations

- $A\beta = y \Rightarrow WA\beta = Wy \Rightarrow A^TWA\beta = A^TWy \Rightarrow$

$$\beta = (A^TWA)^{-1}A^TWy$$

- Define $B = (A^TWA)^{-1}A^TW$
- $A\beta$ is a projection of y onto $cols(A)$, therefore, we can define the projection matrix $P_A : \mathbb{R}^m \longrightarrow col(A) \subset \mathbb{R}^n$:

$$P_A = A(A^TWA)^{-1}A^TW$$

- It is a simple check that $P_A^2 = P_A$
- $MP_A = (I - P_A)P_A = 0$
- $\hat{y} = A\hat{\beta} = P_Ay$
- $e = y - \hat{y} = y - P_Ay$, so it is useful to define $M = I - P_A$, thus $e = My$
- Ideal state is the paradigm given earlier where $y = \hat{y} + \epsilon$. In actuality $y = \hat{y} + e$. Thus, but substitution we can conclude that the "smallest residuals" are when $e = \epsilon$.
- Consider of the two cases (ideal and actual) for the starting equation $e = My$, where $y = \hat{y} + \epsilon$, then

$$e = M(\hat{y} + \epsilon) = MP_Ay + M\epsilon = M\epsilon$$

- Define the objective $F = e^TWe$

B Variances

- For a random vector $\mathbf{v}$:
  - $Var(v) = E[vv^T] - E[v]E[v]^T$
  - For constant matrix A, $Var(Av) = AVar(v)A^T$
- By definition $Var(y) = \sigma^2$
- $Var(\beta) = Var(By) = BVar(y)B^T = \sigma^2BB^T$
- $Var(\hat{y}) = Var(P_Ay) = \sigma^2P_AP_A^T$
- $Var(e) = Var(y - \hat{y}) = Var((I - P_A)y) = Var(My) = \sigma^2MM^T$

C Expected Values

- For a random vector $\mathbf{v}$:
  - For a constant matrix A, $E[Av] = AE[v]$
  - For a random matrix A, $tr(E[A]) = E[tr(A)]$ (citation or proof needed)
- $E[\epsilon]$ is whatever the user defined it to be, typically 0 when taking errors to be normally distributed. In general lets denote $E[\epsilon] = \overline{\epsilon}$
- For a constant c, $E[c] = c$
- $E[\beta] = E[\hat{beta} + \epsilon] = E[\hat{\beta}] + E[\epsilon] = \hat{\beta} + \overline{\epsilon}$
- $E[\hat{y}] = E[A\hat{\beta}] = A\hat{\beta}$
- $E[e] = E[M\epsilon] = M\overline{\epsilon}$

D Expectation of Sum of Square Residuals

Recall that $e = M\epsilon$. Now we can derive the following for the expectation of the general case of a weighted sum of square residuals. Since $E[F]$ is the expectation of a constant, the use of the trace is employed at any point it becomes convenient since $tr(c) = c$.

$$\begin{aligned}
E[e^T We] &= E[\epsilon^T M^T WM\epsilon] \\
&= E[tr(\epsilon^T M^T WM\epsilon)] \\
&= E[tr(WM\epsilon\epsilon^T M^T)] \\
&= E[tr(WMVar(\epsilon)M^T)] \\
&= E[tr(WM\sigma^2 IM^T)] \\
&= \sigma^2 E[tr(WMM^T)] \\
&= \sigma^2 E[tr(M^T WM)] \\
&= \sigma^2 E[tr(W - WP_A - P_A^T W + P_A^T WP_A)] \\
&= \sigma^2 E[tr(W) - 2tr(WP_A) + tr(P_A^T WP_A)] \\
&= \sigma^2 E[tr(W) - 2tr(WP_A) + tr(WP_A)] \\
&= \sigma^2 E[tr(W) - tr(WP_A)] \\
&= \sigma^2 E[tr(WM)] \\
&= \sigma^2 tr(WM)
\end{aligned} \tag{28}$$

Note: When $W = I$ then $tr(WM) = tr(I - P_A) = tr(I) - tr(P_A) = m - n$. This is because I is an $m \times m$ matrix. As for the trace of the projection matrix we will simply use the fact that the trace of a projection matrix is equal to the rank of that matrix.

E Derivation for $Var(f)$

Recall $f(\tilde{x}, \vec{\beta}) \approx J_{\tilde{x}}\vec{\beta}$, at some new point $\tilde{x}$. Let $\tilde{J} = J_{\tilde{x}}$ represent the $1 \times n$ matrix of partial derivatives of f at the new point $\tilde{x}$, then:

$$Var(f) \approx Var(\tilde{J}\vec{\beta}) = \tilde{J}Var(\vec{\beta})\tilde{J}^T \tag{29}$$

F Derivation for $Var(f)$ in Linear Model Example

When f is linear then $f(\tilde{x}, \vec{\beta}) = A_{\tilde{x}}\vec{\beta}$. Recall that for the set of m observations in the example when the model is a line:

$$A = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix}$$

Again we let $A_{\tilde{x}} = \tilde{A}$ be a $1 \times n$ matrix, and let $f(\tilde{x}, \vec{\beta}) = \tilde{A}\vec{\beta}$:

$$\tilde{A} = \begin{bmatrix} 1 & \tilde{x} \end{bmatrix} \tag{30}$$

$$Var(f) = Var(\tilde{A}\vec{\beta}) = \tilde{A}Var(\vec{\beta})\tilde{A}^T = \tilde{A}(\sigma^2(A^TA)^{-1})\tilde{A}^T \tag{31}$$

However, for the variance of parameters $\vec{\beta}$, A is used since the sample set of size m is used in fitting the model parameters. Whereas, the variance of the model function is going to be taken at some new point $\tilde{x}$ and so we use $\tilde{A}$.

$$
\begin{aligned}
Var(f) &= \sigma^2 \tilde{A}(A^T A)^{-1}\tilde{A}^T \\
&= \sigma^2 \begin{bmatrix}1 & \tilde{x}\end{bmatrix} \begin{bmatrix} m & \sum_{i=1}^m x_i \\ \sum_{i=1}^m x_i & \sum_{i=1}^m x_i^2 \end{bmatrix}^{-1} \begin{bmatrix}1 \\ \tilde{x}\end{bmatrix} \\
&= \sigma^2 \begin{bmatrix}1 & \tilde{x}\end{bmatrix} \left(m \begin{bmatrix} 1 & \frac{1}{m}\sum_{i=1}^m x_i \\ \frac{1}{m}\sum_{i=1}^m x_i & \frac{1}{m}\sum_{i=1}^m x_i^2 \end{bmatrix} \right)^{-1} \begin{bmatrix}1 \\ \tilde{x}\end{bmatrix} \\
&= \frac{\sigma^2}{m} \begin{bmatrix}1 & \tilde{x}\end{bmatrix} \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix}^{-1} \begin{bmatrix}1 \\ \tilde{x}\end{bmatrix} \\
&= \frac{\sigma^2}{m} \frac{1}{\overline{x^2} - \bar{x}^2} \begin{bmatrix}1 & \tilde{x}\end{bmatrix} \begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix}1 \\ \tilde{x}\end{bmatrix} \\
&= \frac{\sigma^2}{m} \frac{1}{Var(\vec{x})} \begin{bmatrix}1 & \tilde{x}\end{bmatrix} \begin{bmatrix} \overline{x^2} - \bar{x}\tilde{x} \\ -\bar{x} + \tilde{x} \end{bmatrix} \\
&= \frac{\sigma^2}{mVar(\vec{x})} (\overline{x^2} - \bar{x}\tilde{x} + \tilde{x}(\tilde{x} - \bar{x})) \\
&= \frac{\sigma^2}{mVar(\vec{x})} (\overline{x^2} - 2\bar{x}\tilde{x} + \tilde{x}^2) \\
&= \frac{\sigma^2}{mVar(\vec{x})} (\overline{x^2} - \bar{x}^2 + \bar{x}^2 - 2\bar{x}\tilde{x} + \tilde{x}^2) \\
&= \frac{\sigma^2}{mVar(\vec{x})} (Var(\vec{x}) + (\tilde{x} - \bar{x})^2)
\end{aligned}
$$

For simplicity let $x = \tilde{x}$, then:

$$
Var(f) = \frac{\sigma^2}{m} \left(1 + \frac{(x - \bar{x})^2}{Var(\vec{x})} \right) \tag{32}
$$

In this way the variance is a function of any point x.

Note: If rewritten, $Var(f) = \sigma^2(\frac{1}{m} + \frac{(x-\bar{x})^2}{\sum_{i=1}^m (x_i - \bar{x})^2})$, where $\bar{x}$ is the average of $\vec{x}$.