

Linear Regression & Correlation – Prof. Richard B. Goldstein

Single Variable $Y = \alpha + \beta x + \varepsilon$ where $\varepsilon = N(0, \sigma)$

Fitted Regression $\hat{y} = a + bx$ where a and b are found by least squares fit of n data points (x_i, y_i)

Residuals $e_i = y_i - \hat{y}_i$

Sum of Squares $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2$ is minimized

Sums $S_{xx} = \sum x^2 - n\bar{x}^2$ where $\bar{x} = \frac{\sum x}{n}$

$$S_{xy} = \sum xy - n\bar{x}\bar{y}$$

$$S_{yy} = \sum y^2 - n\bar{y}^2$$

Parameter Estimates: $b = \text{est}(\beta) = \frac{S_{xy}}{S_{xx}}$ $a = \bar{y} - b\bar{x}$

$$s_e^2 = \text{est}(\sigma^2) = \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}$$

Confidence Intervals:

$$b - \frac{t_{\alpha/2} s_e}{\sqrt{S_{xx}}} < \beta < b + \frac{t_{\alpha/2} s_e}{\sqrt{S_{xx}}}$$

$$a - \frac{t_{\alpha/2} s_e \sqrt{\sum x_i^2}}{\sqrt{nS_{xx}}} < \alpha < a + \frac{t_{\alpha/2} s_e \sqrt{\sum x_i^2}}{\sqrt{nS_{xx}}}$$

Hypothesis Testing:

$$t = \frac{b - \beta_0}{s(b)} = \frac{b - \beta_0}{s_e / \sqrt{S_{xx}}} \quad \text{with } n - 2 \text{ d.f.}$$

$$t = \frac{b - \alpha_0}{s(a)} = \frac{a - \alpha_0}{s_e \sqrt{\sum x_i^2 / (nS_{xx})}} \quad \text{with } n - 2 \text{ d.f.}$$

CI for mean response $\hat{y}_0 - t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < \mu_{Y|x_0} < \hat{y}_0 + t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$

where $\hat{y}_0 = a + bx_0$

CI for single response y_0

$$\hat{y}_0 - t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < y_0 < \hat{y}_0 + t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Analysis-of-Variance: $SST = SSR + SSE$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

Source of variation	Sum of Squares	Degrees of freedom	Mean square	Computed f
Regression	SSR	1	SSR	SSR/s ²
Error	SSE	n - 2	$s^2 = \frac{SSE}{n - 2}$	
Total	SST	n - 1		

Test for Linearity of Regression for Data with Repeated Observations

x_i has repeated y values y_{ij} for i = 1, 2, ..., k and j = 1, 2, ..., n_i n = $\sum_{i=1}^k n_i$

let $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$ and $\bar{y}_{i\cdot} = \frac{y_{i\cdot}}{n_i}$ then

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^k n_i (\bar{y}_{i\cdot} - \hat{y}_i)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2 = \text{lack of fit error} + \text{Pure error}$$

Source of variation	Sum of Squares	Degrees of freedom	Mean square	Computed f
Regression	SSR	1	SSR	SSR/s ²
Error	SSE	n - 2		
Lack of fit	SSE - SSE (pure)	k - 2	$\frac{SSE - SSE(\text{pure})}{k - 2}$	$\frac{SSE - SSE(\text{pure})}{s^2 (k - 2)}$
Pure error	SSE (pure)	n - k	$s^2 = \frac{SSE(\text{pure})}{n - k}$	
Total	SST	n - 1		

Coefficient of determination $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = \frac{SSR}{SST}$

Correlation Coefficient $r = \text{Est}(\rho) = \sqrt{R^2} = b \sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$

Hypothesis Tests:

H₀: ρ = 0 uses $t = r \sqrt{\frac{n - 2}{1 - r^2}}$ with n - 2 d.f.

H₀: ρ = ρ₀ ≠ 0 uses $z = \frac{\sqrt{n - 3}}{2} \ln \left[\frac{(1 + r)(1 - \rho_0)}{(1 - r)(1 + \rho_0)} \right]$

Multiple Variable

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon \text{ where } \varepsilon = N(0, \sigma)$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum \varepsilon_i^2$$

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Least Squares Normal Eqs:

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y} \Rightarrow \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

$$\text{SSE} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Let $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ \mathbf{C} is a $(k+1) \times (k+1)$ matrix

Analysis-of-Variance:

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

Hypothesis Test of $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$

Source of variation	Sum of Squares	Degrees of freedom	Mean square	Computed f
Regression	SSR	k	$\text{MSR} = \frac{\text{SSR}}{k}$	$f = \frac{\text{MSR}}{\text{MSE}}$
Error	SSE	$n - (k+1)$	$s^2 = \text{MSE} = \frac{\text{SSE}}{n - k - 1}$	
Total	SST	$n - 1$		

s^2 is an unbiased estimate of σ^2

If $\mathbf{1}$ is an $n \times 1$ vector of all 1's, then

$$\text{SST} = \mathbf{y}'\mathbf{y} - \frac{1}{n} \mathbf{y}'\mathbf{1}\mathbf{1}'\mathbf{y}$$

$$\text{SSR} = \mathbf{b}'\mathbf{X}'\mathbf{y} - \frac{1}{n} \mathbf{y}'\mathbf{1}\mathbf{1}'\mathbf{y}$$

$$\text{SSE} = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}$$

$$\text{where } \frac{1}{n} \mathbf{y}'\mathbf{1}\mathbf{1}'\mathbf{y} = \frac{\left(\sum_{i=1}^n y_i \right)^2}{n}$$

CI for mean response

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{x_0' C x_0} < \mu_{Y|x_0, \dots, x_k} < \hat{y}_0 + t_{\alpha/2} s \sqrt{x_0' C x_0} \text{ with } n - k - 1 \text{ d.f.}$$

$$\text{CI for single response } y_0 \quad \hat{y}_0 - t_{\alpha/2} s \sqrt{1 + x_0' C x_0} < y_0 < \hat{y}_0 + t_{\alpha/2} s \sqrt{1 + x_0' C x_0}$$

Hypothesis Test of $H_0: \beta_j = \beta_{j0}$ uses $t = \frac{b_j - \beta_{j0}}{s \sqrt{c_{jj}}}$
 $\beta_j > \beta_{j0}$

Coefficient of determination $R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}$

Adjusted coefficient of deter. $1 - \frac{n-1}{n-k-1} (1 - R^2)$