

Chapter 0: Review

Chapter 2: Simple Linear Regression

$$\begin{aligned} \mathbf{E}[\mathbf{y}|\mathbf{x}] &= \mu_{y|x} = \mathbf{E}[\beta_0 + \beta_1 x + \epsilon] = \beta_0 + \beta_1 x & \mathbf{V}[\mathbf{y}|\mathbf{x}] &= \sigma_{y|x}^2 = \mathbf{V}[\beta_0 + \beta_1 x + \epsilon] = \sigma^2 & \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} & \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \mathbf{E}[\hat{\beta}_1] &= \sum_{i=1}^n c_i \mathbf{E}[y_i] = \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i = \beta_1 & \mathbf{V}[\hat{\beta}_1] &= \sum_{i=1}^n c_i^2 (\sigma^2) = \sigma^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}} \\ \mathbf{E}[\hat{\beta}_0] &= \beta_0 & \mathbf{V}[\hat{\beta}_0] &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) = V[\bar{y} - \beta_1 \bar{x}] = V[\bar{y}] + x^2 V[\hat{\beta}_1] - \text{cov}(\bar{y}, \hat{\beta}_1) & \mathbf{c}_i &= \frac{x - \bar{x}}{S_{xx}} \\ \mathbf{SS}_{res} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \epsilon_i^2 & \mathbf{SS}_T &= \sum_{i=1}^n y_i^2 - n\bar{y}^2, n-1 \text{ df} & \mathbf{SS}_{Reg} &= \hat{\beta}_1 S_{xy}, \text{ if df} = 1, \text{ then} = MS_{Res} \\ \mathbf{MS}_{res} &= \sigma^2 = \frac{SS_{res}}{n-2} \end{aligned}$$

Hypothesis Testing (Regression)

$$\begin{aligned} \text{Reject } H_0 \text{ if } |t_0| &\geq t_{\frac{\alpha}{2}, n-2} \text{ where } t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MS_{res}}{S_{xx}}}} & \text{Failing to reject } H_0: \beta_i = 0 \text{ implies no rslshp between } x \text{ and } y. & \mathbf{E}[y_i] &= \beta_1 x + \beta_0 \\ F_0 &= \frac{MS_{Reg}}{MS_{res}} = t_0^2 & \text{Reject if } F_0 > F_{\alpha, 1, n-1} & \text{CI: } \hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2} se(\hat{\beta}_{10}) < \hat{\beta}_{10} < \hat{\beta}_1 + t_{\frac{\alpha}{2}, n-2} se(\hat{\beta}_{10}) & se(\hat{\beta}_1) &= \sqrt{\frac{MS_{res}}{S_{xx}}}, se(\hat{\beta}_0) = \sqrt{V[\hat{\beta}_0]} \\ R^2 &= 1 - \frac{SS_{res}}{SS_T} = \frac{SS_{Reg}}{SS_T} & R_{adj}^2 &= 1 - \frac{SS_{res}(n-1)}{SS_T(n-k-1)} \text{ (penalizes you for adding nonsignificant terms to the model)} \end{aligned}$$

Chapter 3: Multiple Linear Regression

$$\begin{aligned} \mathbf{y} &= \mathbf{x} \times \boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{where } \mathbf{p} = k+1, \mathbf{p} \text{ is the total number of betas (or parameters), } k \text{ is the number of regressor variables.} \\ \epsilon &\sim N(0, \sigma^2 \mathbf{I}) \text{ where } \mathbf{I} \text{ is the identity matrix whatever size} & \mathbf{E}[\mathbf{y}] &= \mathbf{x}\boldsymbol{\beta} & \mathbf{V}[\mathbf{y}] &= \mathbf{V}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{I} & \mathbf{y} &\sim N(\mathbf{x}\boldsymbol{\beta}, \sigma^2 \mathbf{I}) \end{aligned}$$

Least Square Estimate for $\boldsymbol{\beta}$ and σ^2

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^n \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{x}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{x}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{x}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{x}'\mathbf{x}\boldsymbol{\beta} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} & \mathbf{E}[\hat{\boldsymbol{\beta}}] &= \mathbf{E}[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}] = \mathbf{E}[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'(\mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon})] = \boldsymbol{\beta} & \mathbf{V}[\hat{\boldsymbol{\beta}}] &= (\mathbf{x}'\mathbf{x})^{-1}\sigma^2 = \mathbf{c}\sigma^2 & \mathbf{V}[\hat{\beta}_j] &= c_{jj}\sigma^2 & \mathbf{E}[\hat{\beta}_j] &= \beta_j \\ \hat{\beta}_j &\sim N(\beta_j, c_{jj}\sigma^2) & \hat{y} &= \mathbf{x}\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = \mathbf{H}\mathbf{y} & \mathbf{E}[\hat{y}] &= \mathbf{E}[\mathbf{x}\hat{\boldsymbol{\beta}}] = \mathbf{x}\boldsymbol{\beta} & \mathbf{V}[\hat{y}] &= \mathbf{V}[\mathbf{x}\hat{\boldsymbol{\beta}}] = \mathbf{x}\mathbf{V}[\hat{\boldsymbol{\beta}}]\mathbf{x}' = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\sigma^2 = \mathbf{H}\sigma^2 \\ \hat{y} &\sim N(\mathbf{x}\boldsymbol{\beta}, \mathbf{H}\sigma^2) & \hat{y}_j &\sim N(x_j\beta_j, h_{jj}\sigma^2), \text{ where } h_{jj} = \mathbf{x}_j'(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_j & x_j &= [x_{j0}, x_{j1}, \dots, x_{jk}] \text{ and} & \hat{\epsilon} &= \mathbf{y} - \hat{y} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} \\ \hat{\sigma}^2(\text{estimator}) &= \frac{SS_{res}}{n-p} = MS_{res} \text{ where } p = k+1 = \text{the number of parameters (i.e. } \beta_0, \beta_1, \dots, \beta_k) & \mathbf{Cov}[\hat{\boldsymbol{\beta}}] &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \text{ (cov matrix c)} \\ SS_{res}(\mathbf{n} - \mathbf{p}) &= (\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}) = \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{x}\hat{\boldsymbol{\beta}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} & SS_{Reg}(\mathbf{k}) &= \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - \frac{(\sum_{i=1}^n y_i)^2}{n} & SS_T(\mathbf{n} - 1) &= \mathbf{y}'\mathbf{y} - \frac{(\sum_{i=1}^n y_i)^2}{n} \\ MS_{res} &= \frac{SS_{res}}{n-k-1} & MS_{Reg} &= \frac{SS_{Reg}}{k} & MS_T &= \frac{SS_T}{n-1} \\ \text{If } \frac{SS_{res}}{\sigma^2} &\sim \chi_{n-k-1}^2 \text{ and } SS_{res}, SS_{Reg} \text{ are indep, then } F_0 = \frac{\frac{k}{SS_{Res}}}{\frac{n-k-1}{MS_{res}}} = \frac{MS_{Reg}}{MS_{res}} \text{ } \mathbf{F} \text{ statistic} & \text{We reject } H_0 \text{ if } F_0 > F_{\alpha, k, n-k-1} \end{aligned}$$

$$\text{error} = (\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon} \quad \mathbf{E}[MS_{Res}] = \sigma^2 \quad \mathbf{E}[MS_{Reg}] = \sigma^2 + \frac{\boldsymbol{\beta}'^* \mathbf{x}_c' \mathbf{x}_c \boldsymbol{\beta}^*}{k\sigma^2} \text{ where } \boldsymbol{\beta}^* = (\beta_1, \beta_2, \dots, \beta_k) \text{ and } \mathbf{x}_c \text{ is the center}$$

$$\text{Testing Individual Coefficients (Partial Test): If } H_0: \beta_j = 0 \text{ is not rejected then delete it: } t_0 = \frac{\hat{\beta}_j}{\sqrt{\sigma^2 c_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad \text{reject if } |t_0| > t_{\frac{\alpha}{2}, n-k-1}$$

Confidence Intervals

$$\begin{aligned} \sigma^2 \text{ known: } \hat{\beta}_j &\sim N(\beta_j, c_{jj}\sigma^2) \longrightarrow \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}\sigma^2}} \sim N(0, 1) \text{ or, if variance is unknown, } \hat{\beta}_j \sim N(\beta_j, c_{jj}MS_{res}) \longrightarrow \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}MS_{res}}} \text{ or } \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-p} \\ \text{Then the variance estimator is } \hat{\sigma}^2 &= MS_{res} = \frac{SS_{res}}{n-p} \sim \chi_{n-p}^2 & \text{So, the } (1 - \alpha) \text{ confidence interval for } \beta_j & \text{ is } \hat{\beta}_j \pm t_{\frac{\alpha}{2}, n-p} se(\hat{\beta}_j) \\ 100(1 - \alpha)\% \text{ for } \sigma^2: & \frac{(n-2)MS_{res}}{\chi_{\frac{\alpha}{2}, n-2}^2} \leq \sigma^2 \leq \frac{(n-2)MS_{res}}{\chi_{1-\frac{\alpha}{2}, n-2}^2} & \hat{y}_j &\sim N(x_j\beta_j, h_{jj}\sigma^2), \text{ so} & \frac{\hat{y}_j - x_j\beta_j}{\sqrt{h_{jj}\sigma^2}} &\sim N(0, 1) & \frac{\hat{y}_j - x_j\beta_j}{\sqrt{h_{jj}MS_{res}}} &\sim t_{n-p} \text{ } MS_{res} \text{ ests } \sigma^2 \end{aligned}$$

$$\text{A } 1 - \alpha \text{ confidence interval for } \mathbf{E}[y_0|x_0] \text{ is } \hat{y}_0 \pm t_{\frac{\alpha}{2}, n-p} \sqrt{\mathbf{x}_0'(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_0\sigma^2} \text{ or } \hat{y}_0 \pm t_{\frac{\alpha}{2}, n-p} \sqrt{\mathbf{x}_0'(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_0MS_{res}}$$

Chapter 4: Model Testing

$$\text{Properties of residuals: mean } 0, MS_{res} = \sum_{i=1}^n \frac{(\epsilon_i - \bar{\epsilon})^2}{n-p} = \sum_{i=1}^n \frac{\epsilon_i^2}{n-p} = \frac{SS_{res}}{n-p}$$

$$\text{Scaling Residuals: Standardized Residuals: } d_i = \frac{\epsilon_i}{\sqrt{MS_{res}}} \text{ Studentized: } r_i = \frac{\epsilon_i}{\sqrt{MS_{res}(1-h_{ii})}}, \mathbf{V}[\epsilon_i] = \sigma^2(1 - h_{ii}), \mathbf{cov}(\epsilon_i, \epsilon_j) = -\sigma^2 h_{ij}$$

Other model testing: plot x_i and x_j : linear rln means high corr.

Formal test for lack of fit: Assuming everything is tested and ideal, to test for linearity, we use: $SS_{res} = SS_{PE} + SS_{LOF}$

$$F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}} \quad \mathbf{E}[MS_{LOF}] = \mathbf{E}[MS_{PE}] = \sigma^2, \text{ where } m \text{ is num regressors, } n \text{ is num samples} \quad \mathbf{V}[\bar{y}] = \frac{p\sigma^2}{n} \text{ (indpure e)}$$