Regression Analysis

The Basics

$$\mathbf{Y} = \beta_0 + \beta_1 x + \epsilon, \qquad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}, \qquad \hat{\beta}_0 = \overline{\mathbf{Y}} - \hat{\beta}_1 \overline{\mathbf{X}} = \overline{\mathbf{Y}} - \frac{S_{XY}}{S_{XX}} \overline{\mathbf{X}},$$

where $S_{XX} = \sum_i (\mathbf{X}_i - \overline{\mathbf{X}})^2$ and $S_{XY} = \sum_i (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{Y}_i - \overline{\mathbf{Y}})$.

Fixed Design: If x_i is fixed, $\sum_i (x_i - \bar{x}) = 0$ and therefore, $S_{XX} = \sum_i (\mathbf{X}_i - \overline{\mathbf{X}})^2 = \sum_i (x_i - \bar{x})^2 = S_{xx}$,

and $S_{XY} = \sum_i (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{Y}_i - \overline{\mathbf{Y}}) = \sum_i (x_i - \bar{x})(\mathbf{Y}_i - \overline{\mathbf{Y}}) = \sum_i (x_i - \bar{x})\mathbf{Y}_i - \overline{\mathbf{Y}} \cancel{\sum_i (x_i - \bar{x})} = \sum_i (x_i - \bar{x})\mathbf{Y}_i$.

The Gradient

$$\mathbf{E}(\hat{\beta}_1) = \mathbf{E}\left(\frac{S_{XY}}{S_{XX}}\right) = \frac{\mathbf{E}(S_{XY})}{S_{xx}} = \frac{\mathbf{E}(\sum_i (x_i - \bar{x})\mathbf{Y}_i)}{S_{xx}} = \frac{\sum_i (x_i - \bar{x})\mathbf{E}(\mathbf{Y}_i)}{S_{xx}} = \frac{\sum_i (x_i - \bar{x})\mathbf{E}(\beta_0 + \beta_1 x_i)}{S_{xx}}$$

$$= \cancel{\frac{\beta_0}{S_{xx}} \sum_i (x_i - \bar{x})} + \frac{\beta_1}{S_{xx}} \sum_i x_i (x_i - \bar{x}) = \frac{\beta_1}{S_{xx}} \left(\sum_i x_i (x_i - \bar{x}) - \sum_i \bar{x}(x_i - \bar{x}) \right) = \cancel{\frac{\beta_1}{S_{xx}} \sum_i (x_i - \bar{x})^2} = \beta_1.$$

$$\mathbf{Var}(\hat{\beta}_1) = \mathbf{Var}\left(\frac{S_{XY}}{S_{XX}}\right) = \mathbf{Var}\left(\frac{\sum_i (x_i - \bar{x})\mathbf{Y}_i}{S_{xx}}\right) = \mathbf{Var}\left(\sum_i \frac{x_i - \bar{x}}{S_{xx}} \mathbf{Y}_i\right) = \sum_i \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 \mathbf{Var}(\mathbf{Y}_i)$$

$$= \frac{\mathbf{Var}(\mathbf{Y}_i)}{S_{xx}^2} \sum_i (x_i - \bar{x})^2 = \frac{\sigma^2}{S_{xx}^2} S_{xx} = \frac{\sigma^2}{S_{xx}}.$$

$$\boxed{\hat{\beta}_1 \sim \mathbf{N}\left(\beta_1, \frac{\sigma}{\sqrt{S_{xx}}}\right).} \qquad \sqrt{S_{xx}} \cdot \frac{\hat{\beta}_1 - \beta_1}{\sigma} \sim \mathbf{N}(0,1), \quad \sqrt{S_{xx}} \cdot \frac{\hat{\beta}_1 - \beta_1}{S} \sim t_{n-2}.$$

The Intercept

$$\mathbf{E}(\hat{\beta}_0) = \mathbf{E}\left(\overline{\mathbf{Y}} - \hat{\beta}_1 \overline{\mathbf{X}}\right) = \mathbf{E}\left(\overline{\mathbf{Y}}\right) - \mathbf{E}\left(\hat{\beta}_1 \overline{\mathbf{X}}\right) = \mathbf{E}\left(\frac{\sum_i \mathbf{Y_i}}{n}\right) - \mathbf{E}\left(\hat{\beta}_1 \bar{x}\right) = \mathbf{E}\left(\sum_i \frac{\beta_0 + \beta_1 x_i}{n}\right) - \bar{x}\mathbf{E}\left(\hat{\beta}_1\right)$$

$$= \frac{n\beta_0}{n} + \beta_1 \sum_i \frac{x_i}{n} - \bar{x}\beta_1 = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0.$$

$$\mathbf{Var}(\hat{\beta}_0) = \mathbf{Var}\left(\overline{\mathbf{Y}} - \hat{\beta}_1 \overline{\mathbf{X}}\right) = \mathbf{Var}\left(\frac{\sum_i \mathbf{Y_i}}{n}\right) - \bar{x}^2 \mathbf{Var}\left(\hat{\beta}_1\right) = \frac{\sum_i \mathbf{Var}(\mathbf{Y_i})}{n^2} - \bar{x}^2 \frac{\sigma^2}{S_{xx}}$$

$$= \frac{n\sigma^2}{n^2} - \bar{x}^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} - \frac{\bar{x}^2}{S_{xx}}\right).$$

$$\boxed{\hat{\beta}_0 \sim \mathbf{N}\left(\beta_0, \sigma\sqrt{\frac{1}{n} - \frac{\bar{x}^2}{S_{xx}}}\right).} \qquad \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} - \frac{\bar{x}^2}{S_{xx}}}} \sim \mathbf{N}(0,1), \quad \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} - \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}.$$

Test Statistic

$t_0 = \sqrt{S_{xx}} \cdot \dfrac{\hat{b_1}}{S}$, where $\hat{b_1}$ is an estimate of $\hat{\beta}_1$.

To test the null hypotheses $H_0 : \beta_2 = 0$ against the alternative hypothese $H_a : \beta_2 \leq 0$,

H_0 is to be rejected if $\hat{b_1} \notin \left[-t\dfrac{S}{\sqrt{S_{xx}}}, \ t\dfrac{S}{\sqrt{S_{xx}}}\right]$, where $t = t_{n-1,1-\frac{\alpha}{2}}$. i.e. $\left|\hat{b_1}\right| > t\dfrac{S}{\sqrt{S_{xx}}} = \left|\dfrac{\hat{b_1}}{t_0}\right|$.

$\boxed{\text{Reject } H_0 \text{ if } |t_0| > t_{n-1,1-\frac{\alpha}{2}}.}$ i.e. $t_{n-1,1-\frac{\alpha}{2}} \in [-|t_0|, |t_0|]$. Let $T \sim t_{n-1,1-\frac{\alpha}{2}}$.

The p-value (probability of being wrong) $p = 1 - \mathbf{P}(T \in [-|t_0|, |t_0|]) = 2 \times \mathbf{P}(T > |t_0|)$.