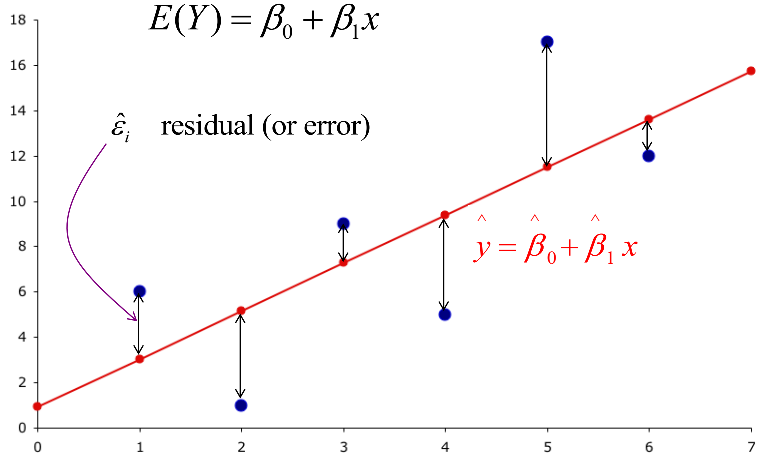
Linear Regression

Model: $Y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim N(0, \sigma)$

$$E(Y) = \beta_0 + \beta_1 x$$



Least Squares Estimation of β_0, β_1

- Least Squares Estimates for β_0 & β_1 minimize the sum of squared errors (SSE)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right)^2$$

Least Squares Estimates (LSEs)

$$\hat{\beta}_1 = \frac{\sum \left(x_i - \bar{x} \right)\left(y_i - \bar{y} \right)}{\sum \left(x_i - \bar{x} \right)^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\sigma}^2 = s^2 = \frac{\sum \left(y_i - \hat{y} \right)^2}{n-2} = \frac{SSE}{n-2} = MSE$$

Inference for the Slope (β_1)

- Standard Error of $\hat{\beta}_1$: $\quad SE_{\hat{\beta}_1} = s / \sqrt{S_{xx}}$
- 2-Sided Test
  - H_0: $\beta_1 = 0$
  - H_A: $\beta_1 \neq 0$

$$t_s = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$$

$RR: \quad |t_s| \geq t_{\alpha/2, n-2}$

$Pvalue: \quad 2P(t \geq |t_{obs}|)$

- (1-α)100% Confidence Interval for β_1

$$\hat{\beta}_1 \pm t_{\alpha/2} SE_{\hat{\beta}_1} \equiv \hat{\beta}_1 \pm t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}}$$

  - Narrower CIs can be achieved through study design

Analysis of Variance in Regression

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

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

Total (SST) **Error** (SSE) **Regression** (SSR)

• $df_{\text{Total}} = n\text{-}1$ $df_{\text{Error}} = n\text{-}2$ $df_{\text{Regression}} = 1$

Analysis of Variance in Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Model	SSR	1	$MSR = SSR/1$	$F = MSR/MSE$
Error	SSE	n-2	$MSE = SSE/(n\text{-}2)$	
Total	$SST=S_{yy}$	n-1		

- Analysis of Variance - F-test
- H_0: $\beta_1 = 0$ H_A: $\beta_1 \neq 0$

$$F_s = \frac{MSR}{MSE}$$

$$RR: \quad F_s \geq F_{\alpha,1,n-2}$$

$$Pvalue: \ P(F \geq F_s)$$

Multiple Regression

- Numeric Response variable (Y)
- p Numeric predictor variables
- Model:

  $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$

- Partial Regression Coefficients: $\beta_i \equiv$ effect on the mean response of increasing the i^{th} predictor variable by 1 unit, **holding all other predictors constant**

Analysis of Variance

- Only changes from SLR are the d.f.
  - $df_{\text{Model}} = p$ $df_{\text{Error}} = n\text{-}p\text{-}1$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Model	SSR	p	$MSR = SSR/p$	$F = MSR/MSE$
Error	SSE	n-p-1	$MSE = SSE/(n\text{-}p\text{-}1)$	
Total	$SST=S_{yy}$	n-1		

$$R^2 = \frac{SST - SSE}{SST} = \frac{SSR}{SST}$$

Testing the Overall Model

- Tests whether **any** of the explanatory variables are associated with the response
- H_0: $\beta_1 = \cdots = \beta_p = 0$
- H_A: Not all $\beta_i = 0$

$$F_s = \frac{MSR}{MSE}$$
$$RR: \qquad F_s \geq F_{\alpha, p, n-p-1}$$
$$Pvalue: P(F \geq F_s)$$

Models with Dummy Variables

- Since <u>genotype</u> is a categorical variable, we need to recode its values in order to include it in the regression model.
- If a categorical variable has *k* levels, need to create *k*-1 dummy variables that take on the values 1 if the level of interest is present, 0 otherwise.
- The baseline level of the categorical variable for which all *k*-1 dummy variables are set to 0
- The regression coefficient corresponding to a dummy variable is the difference between the mean for that level and the mean for baseline group, controlling for all numeric predictors

Analysis of Covariance

- Combination of 1-Way ANOVA and Linear Regression
- Goal: Comparing numeric responses among *k* groups, adjusting for numeric concomitant variable(s), referred to as **Covariate(s)**
- Clinical trial applications: Response is Post-Trt score, covariate is Pre-Trt score
- Epidemiological applications: Outcomes compared across exposure conditions (i.e., genotype), adjusted for other risk factors (age, BMI, sex,...)