1 Models with Quantitative and Qualitative Predictors (Cont'd)

1.1 ANCOVA Models

- We can include a **blend** of ***both* categorical *and* numerical predictor** variables in a regression model.

  For example, consider a study comparing some response variable for males and females while controlling for age. We could fit the model

  $$Y_i \;=\; \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \tag{1}$$

  where X_1 is the indicator variable

  $$X_{i1} \;=\; \begin{cases} 0 \text{ if the } i\text{th individual is Male} \\ 1 \text{ if the } i\text{th individual is Female} \end{cases} \tag{2}$$

  and

  $$X_{i2} \;=\; \text{The age of the } i\text{th individual (in years)} \tag{3}$$

  The **design matrix**, when there are, say, three **females** and three **males**, will look like this:

  $$\boldsymbol{X} \;=\; \begin{bmatrix} 1 & 1 & 27 \\ 1 & 1 & 34 \\ 1 & 1 & 25 \\ 1 & 0 & 24 \\ 1 & 0 & 33 \\ 1 & 0 & 32 \end{bmatrix}$$

  where the 2nd column indicates **gender** and the 3rd gives the individual's **age**.

- Models such as (1) that include both *categorical* and *numerical* predictors are sometimes called ***analysis of covariance (ANCOVA) models***.*

- **ANCOVA models** are **general linear models**, so the least squares estimates $\boldsymbol{b_0, b_1, \ldots, b_{p-1}}$ of the model parameters are be obtained exactly as before, i.e.

  > **Least Squares Estimates of $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_{p-1}}$ (Matrix Approach)**: The vector of estimated coefficients **b** is obtained by:
  >
  > $$\mathbf{b} \;=\; (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

  and the **(estimated) standard errors** of $b_0, b_1, \ldots, b_{p-1}$ reported by statistical software are:

(Estimated) standard errors: The (estimated) standard errors $\boldsymbol{s\{b_0\}}$, $\boldsymbol{s\{b_1\}}$, $\boldsymbol{\ldots, s\{b_{p-1}\}}$ of $b_0, b_1, \ldots, b_{p-1}$ are the square roots of the diagonal elements of the matrix

$$s^2\{\mathbf{b}\} \;=\; \text{MSE} \cdot (\mathbf{X}^T\mathbf{X})^{-1}\,.$$

*Some people use the term ***analysis of covariance*** to refer specifically to models in which the categorical predictor is of primary interest and the numerical predictors are included in the model to control for them.

1.2 Testing for Equality of Two Lines

- Model (1) can be thought of as specifying **two separate linear relationships** between **age** and the **response** variable, one for **males** and the other for **females**.

  For example, using X_1 and X_2 as defined in (2) and (3), for **males** ($X_1 = 0$) the mean response, from model (1), reduces to

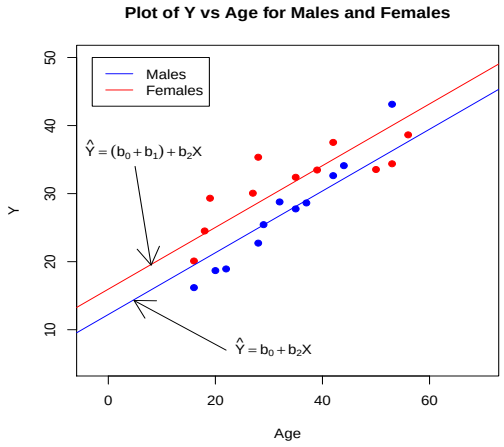
  $$\text{E}(Y) \;=\; \beta_0 + \beta_2 X_2 \tag{4}$$

  whereas for **females** ($X_1 = 1$) the mean response becomes

  $$\text{E}(Y) \;=\; (\beta_0 + \beta_1) + \beta_2 X_2\,. \tag{5}$$

  The two mean responses (4) and (5) are linear functions of X_2 that have the **same slope** (β_2) but **different intercepts** (β_0 for males and $\beta_0 + \beta_1$ for females).

  In the model (1), $\boldsymbol{\beta_1}$ represents the **gap** between the **two lines**, i.e. the "effect" of **gender** on the mean **response**.



If $\beta_1 = 0$ in model (1), the **lines** (4) and (5) for **males** and **females** ***coincide***. Thus the ***t* test** of

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_a : \beta_1 &\neq 0 \end{aligned}$$

reported by statistical software is a **test** of whether the **two lines coincide**.

1.3 Test for Equality of Slopes in Two Lines

- Consider again a model containing **gender** X_1 and **age** X_2, as defined by (2) and (3), but now **with** the **interaction** included in the model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon \tag{6}$$

This model too can be thought of as specifying **two separate linear relationships** between **age** and the **response** variable, one for **males** and the other for **females**.

For example, for **males** ($X_1 = 0$) the mean response, from model (6), is

$$\mathrm{E}(Y) = \beta_0 + \beta_2 X_2 \tag{7}$$

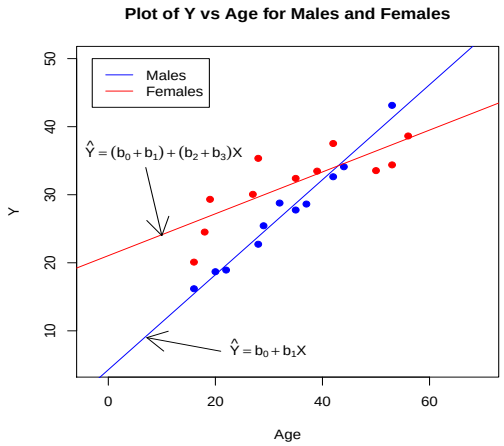
but for **females** ($X_1 = 1$) the mean response is

$$\mathrm{E}(Y) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2 \,. \tag{8}$$

The two mean responses (7) and (8) are linear functions of X_2 that have **different slopes** (β_2 for males and $\beta_2 + \beta_3$ for females) ***and* different intercepts** (β_0 for

males and $\beta_0 + \beta_1$ for females).

The **interaction term** X_1X_2 in the model (6) allows the **two lines** to have **different slopes**.



- **If $\beta_3 = 0$** in model (6) the interaction term disappears and we end up the model (1), in which the **two lines** have the **same slope**.

  Thus the ***t* test** of

$$\begin{aligned} H_0 : \beta_3 &= 0 \\ H_a : \beta_3 &\neq 0 \end{aligned}$$

  reported by statistical software is a **test** of whether the **two lines** have the **same slope**.