1 General Linear Test Approach

- We turn now to an alternative viewpoint of the regression model F test.
- Define the ***full model*** to be

$$Y_i \;=\; \beta_0 + \beta_1 X_i + \epsilon_i$$

and the ***reduced model*** to be

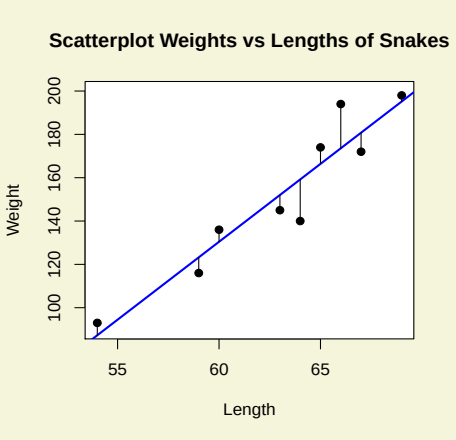
$$Y_i \;=\; \beta_0 + \epsilon_i \,.$$

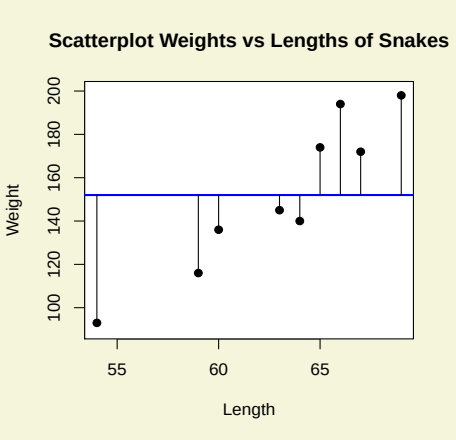
Note that the reduced model is a *special case* of the full model (for which $\beta_1 = 0$). The *general linear F test* can be viewed as a **decision between these two models**.

- Define **SSE(F)** and **SSE(R)** to be the error sums of squares after fitting the **full** and **reduced models**, respectively.
- The full model will explain more of the variation in $Y_1, Y_2, \ldots, Y_n$ than the reduced model, so

$$\text{SSE(F)} \;\leq\; \text{SSE(R)}$$

Example 1.1 For the data on lengths (X) and weights (Y) of nine female snakes, the fitted full and reduced models are show below.





SSE(F) measures the variation depicted in the left plot and **SSE(R)** the variation in the right plot.

- A large difference between these two error sums of squares (**SSE(F)** and **SSE(R)**) provides strong evidence against

$$H_0 : \beta_1 = 0$$

in favor of

$$H_a : \beta_1 \neq 0$$

- To test the above hypotheses, the ***general linear F test approach*** uses the ***F* test statistic**

$$F = \frac{(\text{SSE(R)} - \text{SSE(F)})/(df_R - df_F)}{\text{SSE(F)}/df_F} \tag{1}$$

where $\boldsymbol{df_R}$ and $\boldsymbol{df_F}$ are the degrees of freedom for the error sums of squares in the reduced and full models, respectively:

$$\begin{aligned} df_R &= n - 1 \\ df_F &= n - 2 \end{aligned}$$

> **Fact 1.1** It can be shown that when H_0 is true,
>
> $$F \sim F(df_R - df_F, df_F)$$
>
> i.e. the test statistic (1) follows an F distribution with numerator and denominator degrees of freedom $df_R - df_F$ and df_F, respectively.

P-values are upper tail areas to the **right** of the observed F value under the $F(df_R - df_F, df_F)$ distribution.

- It can also be shown that

$$\text{SSE(R)} = \text{SSTO}$$

and

$$\text{SSE(F)} = \text{SSE}$$

so

$$\text{SSE(R)} - \text{SSE(F)} = \text{SSTO} - \text{SSE} = \text{SSR}$$

and thus

$$F = \frac{(\text{SSE(R)} - \text{SSE(F)})/(df_R - df_F)}{\text{SSE(F)}/df_F} = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{\text{MSR}}{\text{MSE}}. \tag{2}$$

From (2) we see that the general linear F test statistic is the **same** as the regression model F test statistic.

(We'll see later, though, that the general linear F test can be applied more broadly than the regression model F test.)

2 The Coefficient of Determination R^2

- The ***coefficient of determination***, denoted $\boldsymbol{R^2}$, measures the strength of the linear association between X and Y. It's defined as

> **Coefficient of Determination**:
>
> $$R^2 \;=\; \frac{\text{SSR}}{\text{SSTO}} \;=\; 1 - \frac{\text{SSE}}{\text{SSTO}}$$

  R^2 is interpreted as the **proportion of variation** in $\boldsymbol{Y_1, Y_2, \ldots, Y_n}$ that can be **explained** by the **linear relationship between** $\boldsymbol{Y}$ and $\boldsymbol{X}$.

- Note that

$$\mathbf{0} \;\leq\; \boldsymbol{R^2} \;\leq\; \mathbf{1}$$

  Values of R^2 **close to one** indicate:

  - ▷ A large proportion of the variation in Y is explained by X.
  - ▷ There's a strong linear association between these two variables.
  - ▷ The linear model fits the data well.

  Values of R^2 **close to zero** indicate

  - ▷ Only a small proportion of the variation in Y, if any, is explained by X.
  - ▷ There's only a weak (linear) association, or none at all.
  - ▷ The linear model doesn't fit the data very well.

- In the context of simple linear regression, it can be shown that

$$R^2 \;=\; r^2\,,$$

  where $\boldsymbol{r}$ is the ***sample correlation***, i.e.

$$r \;=\; \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}\,.$$

3 Model-Checking Diagnostics

- To assess the adequacy of the linear regression model, we use plots of the *residuals* to check:
  - ▷ Linearity of the relationship between Y and X.
  - ▷ The normality assumption for the error terms ϵ.
  - ▷ The independence assumption for the error terms.
  - ▷ The constant variance assumption for the error terms.
  - ▷ The presence or absence of outliers.

3.1 Semi-Studentized Residuals

- Sometimes before plotting residuals it's useful to **standardize** them, for example to identify outliers.
- The expected value of each residual is zero since

$$\mathrm{E}(e_i) \;=\; \mathrm{E}(Y_i - \hat{Y}_i) \;=\; \mathrm{E}(Y_i) - \mathrm{E}(\hat{Y}_i) \;=\; 0\,.$$

- But the residuals **aren't independent** since they sum to zero. They're **almost** independent, though, especially when n is large, so we'll carry on as if they were independent.
- We can (approximately) **standardize** the residuals by dividing each one by $\sqrt{\mathrm{MSE}}$. These are denoted $\boldsymbol{e_i^*}$ and called ***semi-studentized residuals***:

$$e_i^* \;=\; \frac{e_i - 0}{\sqrt{\mathrm{MSE}}}$$

  Large e_i^*'s (larger than about four) should be investigated as possible outliers.

3.2 Some Useful Plots for Checking Model Assumptions

- Some good plots for checking assumptions are:
  - ▷ For checking the **linearity** assumption and the **constant variance** assumption:
    - * Scatterplot of residuals versus X (curved patterns indicate non-linearity).
    - * Scatterplot of residuals versus fitted (predicted) values $\hat{Y}$ (curved patterns indicate non-linearity, increasing spread indicates non-constant variance).
  - ▷ For checking **normality** of the errors:

    * Histogram of residuals (should be roughly bell-shaped).
    * Normal probability plot of residuals (should follow a straight line).
  - ▷ For checking **whether more predictors should be included** in the model:
    * Scatterplot of residuals versus other potential predictors not included in the model (a linear or curved pattern suggests the predictor should be included in the model).
  - ▷ For checking **independence** of the errors (and deciding whether time should be included in the model as a predictor):
    * Scatterplot of residuals versus the time order in which the observations were recorded (if nearby points tend to be more alike than points farther apart, either the errors are dependent or time should be included in the model).

- **Comments**:
  - ▷ All of the above plots are also useful for identifying outliers.
  - ▷ Any of the above plots could also be made using studentized residuals.
  - ▷ We can perform hypothesis tests for normality, independence, constant variance, outliers, etc. using the residuals. See the textbook.

3.3 Some Remedies

- If assumptions for the linear regression model aren't met, some possible remedies are:
  - ▷ If the **linearity** assumption isn't met:
    * Use a different model (e.g. polynomial regression).
    * Perform a transformation of the X variable (e.g. the log of X).
  - ▷ If the **normality** assumption isn't met:
    * Perform a transformation of the Y variable (e.g. the log of Y).
    * If the distribution of Y is known (e.g. Poisson), use a model appropriate for that distribution (e.g. Poisson regression).
  - ▷ If the **constant variance** assumption isn't met:
    * Perform a transformation of the Y variable (e.g. the log).

- Sometimes violations of more than one assumption occur simultaneously (e.g. the relationship is nonlinear *and* the variance isn't constant). Often a transformation of Y (e.g. the log) will remedy both.

- On the other hand, sometimes a remedy for one violation of an assumption (e.g. taking the log of Y to meet the normality assumption) can introduce violations of another assumption (e.g. linearity). In this case, it may be necessary to apply more than one remedy simultaneously (e.g. transform both X *and* Y).
- See the textbook for specific transformations to use in different situations.