ANalysis Of VAriance Table

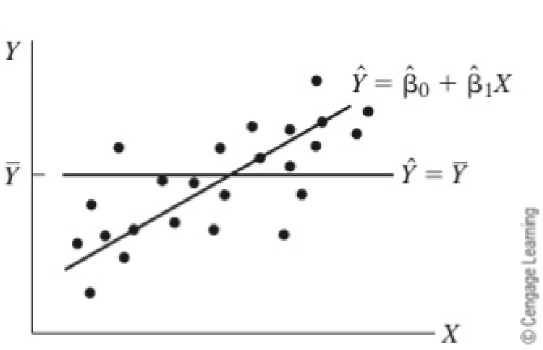

FIGURE 6.4 Predictions of Y using and not using X

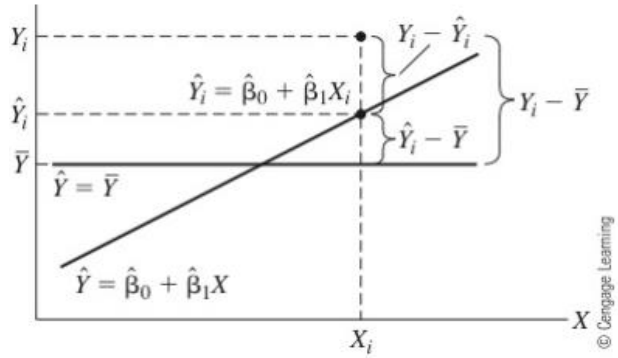

FIGURE 7.1 Variation explained and unexplained by straight-line regression

- Decomposition of the Total SS

$$\sum_i \left[Y_i - \bar{Y} \right]^2 = \sum_i \left[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \right]^2$$

$$= \sum_i (Y_i - \hat{Y}_i)^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + (\hat{Y}_i - \bar{Y})^2$$

$$= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2$$

$$SS(Total) = SS(Error) + SS(\text{Regression})$$

$$SST = SSE + SSR$$

- ANOVA Table

Source	**d.f.**	**Sum of Squares**	**Mean Square**	**F**
Regression (X)	*1*	*SS(Regression)*	*MSR=SSR / 1*	*MSR / MSE*
Residual/Error	*n-2*	*SS(Error)*	*MSE=SSE / (n-2)*	
Total	*n-1*	*SS(Total)*		

Testing the assumption of Linearity

- **Lack of Fit (LOF) test:** possible when there are repeated observations at different *x*-levels
  - Split SSE into 2 parts
    - Amount due to variation in *y*-values for a fixed *x*-value ("pure" experimental error)
    - Amount due to non-linear terms not included in the model ("lack of fit")
      $SSE = SS_{PureError} + SS_{LOF}$
  - If the linear regression model is correct $S^2_{Y|X} = MSE$ gives an unbiased estimate of σ^2
    If the linear regression model is NOT correct, $S^2_{Y|X}$ gives a biased estimate (inflated)
  - Suppose there are k different x-levels with repeated observations:
    At x_1 we have $y_{11}, y_{12}, \ldots , y_{1\,n1}$ (n_1 observations with variance s^2_1)
    At x_2 we have $y_{21}, y_{22}, \ldots , y_{2\,n2}$ (n_2 observations with variance s^2_2)
    …
    At x_k we have $y_{k1}, y_{k2}, \ldots , y_{2\,nk}$ (n_k observations with variance s^2_k)
    - $n_1 + n_2 + \cdots + n_k = n_{PE}$ observations for pure error $(df_{PE} = n_{PE} - k)$
    - Pooling $s^2_1, s^2_2, \ldots, s^2_k$ gives $s_p^2 = \sum_{i=1}^{k} \frac{(n_i - 1)s_i^2}{n_{PE} - k} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_i)^2}{n_{PE} - k} = \frac{SS(PureError)}{df_{PE}}$

Source	d.f.	Sum of Squares	Mean Square	F
Regression	*1*	*SSR*	*SSR / 1*	*MSR / MSE*
Error(Residual)	*n-2*	*SSE*	*SSE / (n-2)*	
Lack of Fit	$n\text{-}2\text{-}df_{PE}$	*SSE-SS(PE)*	$SS(LOF) / df_{LOF}$	
Pure Error	df_{PE}	*SS(PE)*	$SS(PE) / df_{PE}$	
Total	*n-1*	*SSY*		

  - Hypotheses:
    H_0: There is no lack of fit (the regression is linear in *x*)
    H_A: There is lack of fit (the regression is not linear in *x*)