Chapter 8
Assessing Relationships with Correlation and Regression

Section 8.4
Regression Model ANOVA

School of Business Administration
Portland State University

- Model Based ANOVA

8.4a
Regression Model ANOVA

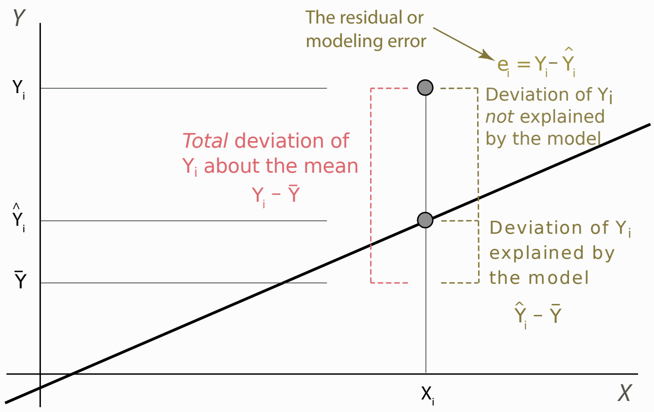
Variability via the Regression Model

More Insight Provided by a Regression Model

Variability of Y explained

- Underlying the meaning of R^2 is the accounting of the variation of the response variable Y provided by the regression model
- We have already explored the variation of Y *not* accounted for by the model in terms of scatter about the regression line as assessed by $\sum e_i^2$
- There is also variation of Y that is accounted for by the model
- And, there is the total variation of Y presented in the early chapters relating to the standard deviation of Y
- All of these concepts are related, as shown next

Key to R^2: Breakdown of Variation of Y



Key to R^2: Breakdown of Variation of Y

- Axes for Response Variable Y and Predictor Variable X
- Paired data values for i^{th} row, $< X_i, Y_i >$, a single point
- Consider the regression line and fitted value of Y_i, $\hat{Y}_i$
- Modeling error or residual is distance of Y_i from $\hat{Y}_i$
- Consider also the total deviation of Y, $Y_i - m_i$
- total deviation = explained deviation + unexplained deviation

Key to R^2: Breakdown of deviation of Y

Explained and unexplained deviation

- Express the result from the previous figure as an equation
- Express the total deviation in Y_i in terms of what is explained and what is not explained by the model
- The relation is additive
- Begin with the total deviation about the mean: $Y_i - m$
- $Y_i - m = Y_i - m$
- Now just add $-\hat{Y} + \hat{Y} = 0$ to the right side of the equation
- $Y_i - m = (Y_i - \hat{Y}) + (\hat{Y} - m)$
- Total deviation =
  Unexplained deviation + Explained deviation
- This result generalizes to the computation of the corresponding sum of squared deviations for each term

Key to R^2: Breakdown of Variation

Understand how much the model does and does not explain

- SSY (or SST): Total variation of Y based on deviation scores,
  $\sum(Y_i - m)^2$
- SSR: Variation of Y accounted for by regression model

  $\sum(\hat{Y}_i - m)^2$
- SSE: Error or residual variation of Y, *not* accounted for by model

  $e_i^2 = \sum(Y_i - \hat{Y}_i)^2$
- For response variable Y, variation that the model explains and does not explain adds up to the total variation of Y

  SSY = SSR + SSE

R: `reg` Output, ANOVA

Breakdown of overall variation of Y

- Return to the ANOVA table from the `reg` output, which contains both SSR and SSE

```
Analysis of Variance

           df    Sum Sq    Mean Sq    F-value    p-value
       Ht  1   1565.226   1565.226      8.363     0.0201
Residuals  8   1497.249    187.156
```

- For this model and data, $SSR = 1565.2$ and $SSE = \sum e^2 = 1497.2$
- The total variation of Y is not given, and so SSY would be computed from addition

$$SSY = SSR + SSE = 1565.2 + 1497.2 = 3062.4$$

- The End