Coefficient of Determination, $R^2$

- The Coefficient of Determination ($R^2$) is a measure of how much of the variation in $Y$ is explained by the model.

Regression Sum of Squares: $SSR = \sum (\hat{y}_i - \bar{y})^2$

Error Sum of Squares: $SSE = \sum (y_i - \hat{y}_i)^2$

Total Sum of Squares: $SSTO = \sum (y_i - \bar{y})^2$

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \quad \text{(only true for linear regression)}$$

Goodness of Fit

- Goodness of fit metric: $R^2 = r^2$

$$r = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} \quad R^2 = \frac{\text{cov}^2(X,Y)}{\text{var}(X)\text{var}(Y)}$$

- Also, we can show that:

$$R^2 = 1 - \frac{\text{var}(e)}{\text{var}(Y)} \quad R^2 \text{ is the fraction of the variance of } Y \text{ that is explained by the linear fit}$$

Overall F-Test for Regression

- An overall test for model significance:
  - $H_0$: $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$ (not testing intercept)
  - $H_A$: $\beta_j \neq 0$, for at least one value of $j$
  - $p = \text{number of parameters in the model}$

- If $H_0$ is true, then the model is not useful for explaining the variation in $y$
  - $SSR$ is much smaller than $SSE$
  - $SSE$ is about the same as $SSTO$

Build an ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>SSR</td>
<td>p-1</td>
<td>SSR/(p-1)</td>
<td>MSR/MSE</td>
<td>from F-distribution</td>
</tr>
<tr>
<td>Error</td>
<td>SSE</td>
<td>n-p</td>
<td>SSE/(n-p)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>SSTO</td>
<td>n-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$F = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{MSR}{MSE}$$
For Weighted Regression

- We must include the weights in our sum of squares calculations

\[ \text{Regression Sum of Squares: } SSR = \sum w_i (\hat{y}_i - \bar{y}_w)^2 \]

\[ \text{Error Sum of Squares: } SSE = \sum w_i (y_i - \hat{y}_i)^2 \]

\[ \text{Total Sum of Squares: } SSTO = \sum w_i (y_i - \bar{y}_w)^2 \]

Overall F-Test for Regression

- Note that the coefficient of determination \((R^2)\) is related to this F-statistic:

\[ R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \]

\[ F = \frac{R^2/(p-1)}{(1-R^2)/(n-p)} \]

Training vs. Predicting

- We build a model by fitting it to data
  - We “train” or calibrate a model using a given data set
  - The residual standard deviation is a measure of how well the model fits this data set
  - If we add more fitting parameters to the model, we always get a better fit
- The real test, however, is how well we match a new data set, one not used in the training
  - Called validation of the model
  - The residual standard deviation for validation data will almost always be higher than for the training data

Overfitting

- Adding new model terms always makes the fit better, but can result in fitting noise
Model Scope

- Every model is built and validated over a range of input conditions, called the **scope** of the model
  - When a model is developed, its scope should be clearly specified
  - Prediction and interpretation (the two goals of modeling) should generally be limited to within the scope
  - Extrapolations are sometimes done, but know that uncertainty estimates are no longer valid

Lecture 35: What have we learned?

- What is the coefficient of determination and how is it calculated?
- What is the overall regression F-test and how is it used?
- What are the assumptions inherent in the F-test?
- What is model validation?
- Define model scope