Lecture 11

Wednesday, June 12, 2019 2:35 PM

We're now able to fit a MLR model (i.e., estimate the βs), we can interpret the relationships (via the β's), we can test hypotheses about the β's and we can calculate CI's and PI's for expected/predicted response values.

We should probably decide whether the model fits the data well... We'll do this with an Analysis of Variance (ANOVA).

Analysis of Variance (ANOVA)

Variability in the response variable (the y values) can be quantified by how far they deviate from their mean. Measures of variation in the y's typically contain the term

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

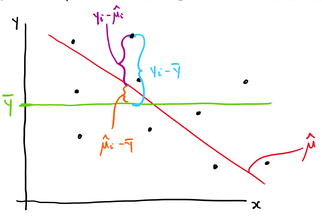
which we define as the **total sum of squares**

* Notice $s_y = \sqrt{\frac{SST}{n-1}}$

The objective of ANOVA is to partition the total variation in y (SST) into two parts:
(1) Variation accounted for by the model
(2) Variation not accounted for by the model

Notice that $y_i - \bar{y} = (y_i - \hat{\mu}_i) + (\hat{\mu}_i - \bar{y})$, $i = 1, 2, \ldots, n$

Although this is true no matter how many explanatory variables we have, we'll visualize it in SLR case.



Let's now look at this partition in terms of the sums of squares:

$$\begin{aligned} SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} [(y_i - \hat{\mu}_i) + (\hat{\mu}_i - \bar{y})]^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 + \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2 + 2\sum_{i=1}^{n} (y_i - \hat{\mu}_i)(\hat{\mu}_i - \bar{y}) \\ &= \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 + \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2 \\ &\equiv SSE + SSR \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)(\hat{\mu}_i - \bar{y}) &= \sum_{i=1}^{n} e_i (\hat{\mu}_i - \bar{y}) = \sum_{i=1}^{n} e_i \hat{\mu}_i - \bar{y} \sum_{i=1}^{n} e_i \\ &= \vec{e}^T \hat{\mu} - \bar{y}\, \vec{e}^T \vec{1} \\ &= 0 - \bar{y} \times 0 \quad \text{since } \vec{e} \text{ is orthogonal to both } \hat{\mu} \text{ and } \vec{1} \\ &= 0 \end{aligned}$$

The **sum of squared errors** (SSE) $\left(= \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2\right)$ are the sum of squared residuals, and this reflects variation in the response that the model could not explain.
* SSE is sometimes referred to as "residual sum of squares"

The **regression sum of squares** (SSR) $\left(= \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2\right)$ measures the amount of variation in the response that the model is able to explain.

This variation decomposition is typically displayed in an ANOVA table.

Source	Sum of Squares	df	Mean Squares	F
Regression	SSR	p	$MSR = \frac{SSR}{p}$	$F_0 = \frac{MSR}{MSE}$
Error	SSE	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$	
Total	SST	$n-1$		

$\hat{\sigma}^2$

* $df(SSR) + df(SSE) = df(SST)$

* the table also contains "mean squares" which are the sums of squares divided by their respective degrees of freedom

test statistic associated with the test overall significance in the linear regression (will come back to this)

To answer the question "Does the model fit the data well?" we can use the **coefficient of determination** (R^2). This metric quantifies the proportion of response variation accounted for by the model.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad \in [0, 1]$$

because $SST = SSR + SSE$

Intuitively larger values are better than small values. The more variation the model explains the larger SSR will be (and hence the smaller SSE will be). What this means practically is that the fitted values lie close to true values and hence our model fits the data well.

PGA Example

Source	df	Sum. Sq.	Mean Sq.	F
Regression	1	2049.0	2049.0	115.31
Error	194	3447.3	17.77	
Total	195	5496.3		

$$R^2 = \frac{2049.0}{5496.3} = 0.3728$$

Therefore, the model (driving distance) explains 37.28% of the variation in the response (driving accuracy).