Chemometrics

MA4605

Week 8. Lecture 15. Linear regression as ANOVA

October 24, 2011

Anova and regression calculations

- The ANOVA table can be used to test not only H_0:$\mu_1 = \mu_2 = ... = \mu_p$ when comparing the means of several groups, but also in regression.
- In the case of means, the null hypothesis means that the group membership does not affect the mean value of the response variable y.
- In the case of regression, the null hypothesis is that the independent variable *x* does not affect y.
- If we were to assume that *y* does not affect *x*, that would mean that the regression line was flat, that it had no slope.
- Therefore, the null hypothesis for the ANOVA table in regression is $H_0 : \beta = 0$ and the alternative hypothesis is $H_a : \beta \neq 0$.

Anova calculations

ANOVA distinguishes between two sources of variation: due to regression and about regression(residual).
If the regression line is flat the two sources of variation are identical.
The regression line is not flat the variability due to regression must be greater than the variability about regression(residual).
This way we decide if the regression slope is statistically significant.

ANOVA uses the fact that

$$y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})$$

which leads to

$$\sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2$$

$\sum (y_i - \overline{y})^2$ is the total sum of squares and measures the total variation among among all the observed y-values.
$\sum (\hat{y}_i - \overline{y})^2$ is the part measuring how much of the variation among the y_i can be explained by the fitted regression line (variation due to regression)
$\sum (y_i - \hat{y}_i)^2$ is the part measuring the size of the deviation from this line (residuals/variation about regression).
If the residuals are small, the line is a good fit for the data.

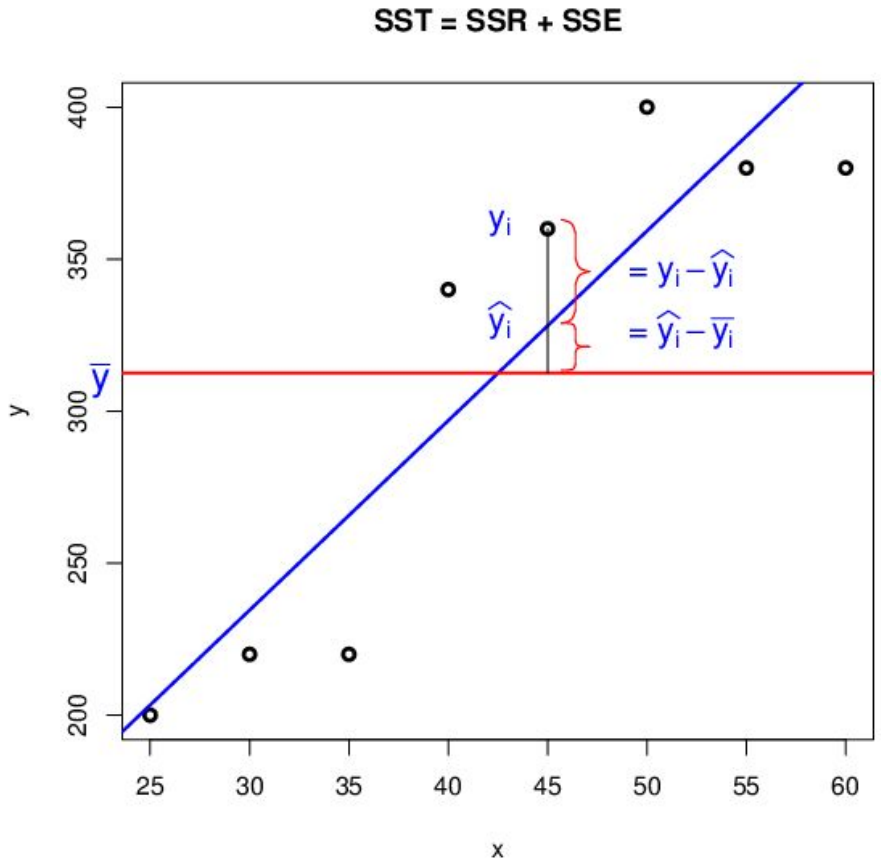
$$\begin{aligned}
\text{Total variation in Y} &= \text{Explained Variation} + \text{Unexplained Variation} \\
\sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y})^2 \\
SST &= SSR + SSE
\end{aligned}$$

where

SST = total sum of squares

SSR = sum of squares due to regression (Explained Variation)

SSE = sum of squares due to error (Unexplained Variation)


SST = SSR + SSE
y
x
200
250
300
350
400
25
30
35
40
45
50
55
60
y_i
$\hat{y}_i$
$\bar{y}$
$= y_i - \hat{y}_i$
$= \hat{y}_i - \bar{y}_i$

Anova table for regression

Source	Sum of Squares (SS)	DF	Mean Squares (MS)	F
Total	TSS= $\sum (y_i - \overline{y})^2$	n-1		
Regression	SSR=$\sum (\hat{y}_i - \overline{y})^2$	1	MSR=$\frac{SSR}{1}$	F=$\frac{MSR}{MSE}$
Error	SSE=$\sum (y_i - \hat{y}_i)^2$	n-2	MSE=$\frac{SSE}{n-2}$	

If the test statistic F is significantly large, i.e. larger than qf(0.95,1,n-2), we reject the null hypothesis $H_0 : \beta = 0$ and assume that the slope is not zero. The p-value from the ANOVA table is the same as the p-values from the t-test for the slope variable. The t-test for $\beta \neq 0$ always agrees with the F-test for $\beta \neq 0$.

Example.

Standard aqueous solutions of fluorescein are examined in a fluorescence spectrometer, and yield the following fluorescence intensities:

Fluorescence intensities	2.1	5.0	9.0	12.6	17.3	21.0	24.7
Concentration	0	2	4	6	8	10	12

> y<- c(2.1, 5.0, 9.0, 12.6, 17.3, 21.0, 24.7)
> x<- c(0, 2, 4, 6, 8, 10, 12)
> model <- lm(y∼x)
> summary(model) provides the t-test for $H_0 : \beta = 0$
> anova(model) provides the F-test for $H_0 : \beta = 0$

Example R output

```
> anova(model)
Analysis of Variance Table
Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 417.34  417.34  2227.5 8.066e-08
Residuals  5   0.94    0.19
> summary(model)
Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   1.5179     0.2949   5.146   0.00363
x             1.9304     0.0409  47.197  8.07e-08
```

Both p-values for testing $\beta = 0$ equal 8.066e-08 which is less than 0.05, hence we reject the null hypothesis $H_0 : \beta = 0$.