STAT6038 week 3 lecture 9

Rui Qiu

2017-03-10

Overall F test for a regression model The multiple regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon, \epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

The overall F test tests:

H_0: all slope coefficients $\beta_1, \beta_2, \beta_3, \ldots, \beta_k = 0$ (implies the mean or null model $Y = \beta_0 + \epsilon$ is sufficient)

against

H_a: at least one $\beta_j \neq 0$ ($j = 1, 2, \ldots, k$) (i.e. we need at least one of the terms involving the X variables in the model)

In SLR this becomes:
$H_0 : \beta_1 = 0$
against
$H_a : \beta_1 \neq 0$
(This is the same hypothesis as the t-test on the slope coefficient)

It is true that the F-test is equivalent to this tests about mean coefficients. But really the F test is a test about variance components (which is why it appears in ANOVA table).

So in terms of the variance model:

1. STEP I (Hypothesis): $H_0 : \frac{\sigma^2_{Y|X}}{\sigma^2} = 1$ vs $H_a : \frac{\sigma^2_{Y|X}}{\sigma^2} > 1$
2. STEP II (Test statistics): $F = \frac{MS_{regression}}{MS_{residual/Error}} \sim F_{k,n-p}$, where $p = k + 1$. [Note that: for SLR $\sim F_{1,n-2}, k = 1$]
3. STEP III (Decision rule): $\alpha = 0.05$, reject H_0 if observed $F > F_{1,n-2}(0.95)$
4. STEP IV (Calculations, 1-tail): $\alpha = 0.05$, observed $F = 48.1$, calculated p-value is 0.000002 in R we use qf(0.95,1,17).
5. STEP V (Conclusion): So, as observed $F = 48.1 \gg F_{1,17}(0.95) = 4.45$ OR as $p = 0.000002 \ll \alpha = 0.05$, reject H_0 in favor of H_a.

Mean Interpretation: the model involving the X variable (there is only 1 here) is superior to a null model.

Variance Interpretation: the proportion of the variance in Y explained by the larger model (involving X) is significantly larger than the error variance.

Why would we bother doing this? Why not just use t-test? Truly, for SLR, they are the same. But:

T-test on the (1) slope coefficient (in SLR)

1. STEP I: $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$
2. STEP II: $t = \frac{b_1 - 0}{se(b_1)} \sim t_{n-2}$ where $n-2$ is the residual or error df.

   Also we use $var(\beta_1) = \frac{\sigma^2}{s_{xx}}$ to estimate $se(b_1)$, so $se(b_1) = \frac{\sigma}{\sqrt{s_{xx}}}$

   σ is unknown, so estimate using $\hat{\sigma}^2 = s^2 = MSE$

   $\hat{\sigma} = s = RSE$ residual SE or RMSE root MSE.
3. STEP III: $\alpha = 0.05$, reject H_0 if observed $t < t_{n-2}(0.025)$ or $t > t_{n-2}(0.075)$
4. STEP IV: check distribution plot.
5. STEP V: $p < \alpha = 0.05$, so reject H_0 and conclude $beta_1 \neq 0$ i.e. there is a relationship between Y(protein) and X(gestation).

It is NOT a coincidence that the p-values for two tests (overall F and t-test) were the same!

$$E[MS_{regression}] = \sigma^2 + \beta_1^2 s_{xx}. \text{ [see QS of Tutorial 1]}$$

$$E[MS_{error} = \sigma^2$$

So

$$F = \frac{MS_{regression}}{MS_{error}} \implies \frac{\sigma^2 + \beta_1^2 s_{xx}}{\sigma^2} = 1 + \frac{\beta_1^2}{1 + t^2}$$