Overall F test for a regression model

The multiple regression model

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

the overall F test tests:

H_0: all slope coefficients $\beta_1, \beta_2, \beta_3 \dots \beta_k = 0$

(implies the mean or null model $Y = \beta_0 + \varepsilon$ is sufficient)

against

H_A: at least one $\beta_j \neq 0$ $(j = 1, 2, \dots k)$

(ie we need at least one of the terms involving the X variables in the model)

In SLR this becomes

$$H_0 : \beta_1 = 0$$

vs $$H_A : \beta_1 \neq 0$$

(same hypotheses as the t-test on the slope coefficient)

It is true that the overall F test is equivalent to this test about mean coefficients, but really the F test is a test about variance components (which is why it appears in the ANOVA table)

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

Step I $H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$

Step II $t = \dfrac{b_1 - 0}{se(b_1)} \sim t_{n-2}$ ← residual or error df

$se(b_1)$: $Var(\beta_1) = \dfrac{\sigma^2}{S_{xx}}$ $\quad se(b_1) = \dfrac{\sigma}{\sqrt{S_{xx}}}$

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

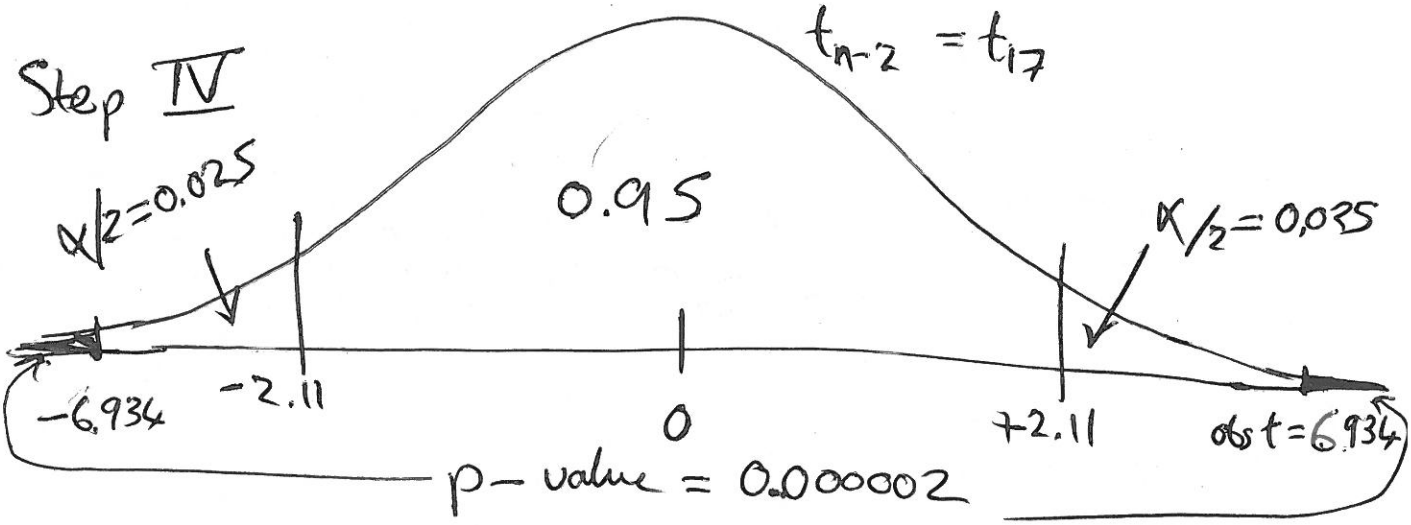
$\hat{\sigma} = s = RSE$ residual SE or RMSE root mse

Step III $\alpha = 0.05$, reject H_0 if observed t ($\alpha/2$)

ie $< t_{n-2}(0.025)$

or $> t_{n-2}(0.975)$

Step IV



p-value = 0.000002

Step V $p < \alpha = 0.05$, so reject H_0 & conclude $\beta_1 \neq 0$

ie there is a relationship between Y (protein) and X (gestation)

So, in terms of the variance model:

Step I (Hypotheses)

$$H_0: \frac{\sigma^2_{Y|X}}{\sigma^2} = 1 \quad \text{vs} \quad H_A: \frac{\sigma^2_{Y|X}}{\sigma^2} > 1$$

Step II (Test statistic)

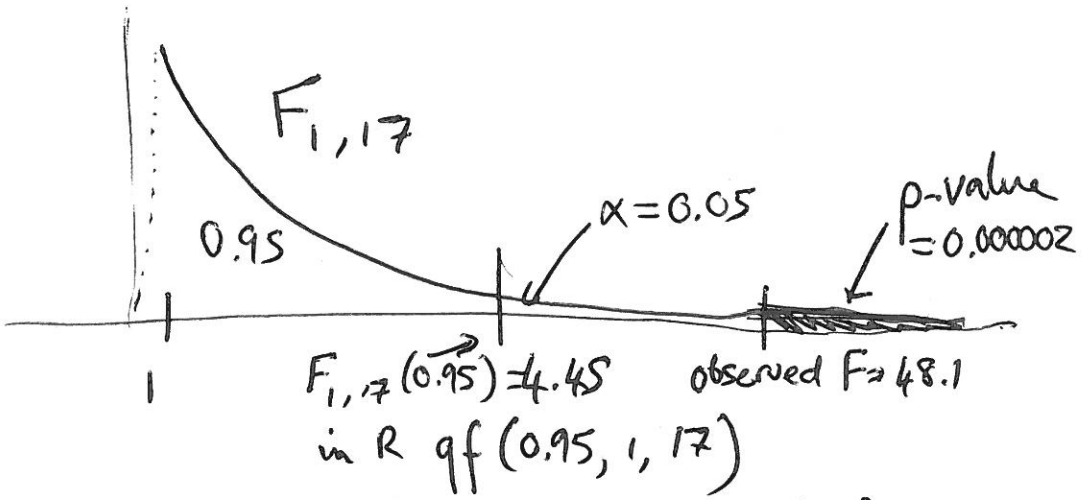
$$F = \frac{MS_{Regression}}{MS_{Residual/Error}} \sim F_{k, n-p}, \quad p = k+1$$

[For SLR $\sim F_{1, n-2}$, $k = 1$]

Step III (Decision Rule)

$\alpha = 0.05$, reject H_0 if observed $F > F_{1, n-2}(0.95)$

Step IV (Calculations)



in R qf(0.95, 1, 17)

Step V (Conclusion)

So, as observed $F = 48.1 >> F_{1,17}(0.95) = 4.45$

OR as $p = 0.000002 << \alpha = 0.05$

reject H_0 in favour 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 the error variance

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

It can be shown (see page 14 of the lecture notes):

$$MS_{Regression} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{1} = b_1^2 S_{xx}$$

$1 \leftarrow k=1$ for SLR

So, for SLR

$$F = \frac{MS_{Regression}}{MS_{Error}} = \frac{b_1^2 S_{xx}}{s^2} = \frac{b_1^2}{s^2/S_{xx}} = \left[\frac{b_1}{se(b_1)}\right]^2$$

$$= t^2$$

In general on $F_{1,\nu} \equiv t_\nu^2$

$1 \leftarrow n=1$

this works for SLR

→ it will not work in general for multiple regression, when $k > 1$