STAT6038 week 4 lecture 10

Rui Qiu

2017-03-15

The overall F-test (in the ANOVA table for SLR & also the last line in the summary output) and the **t-test** on the slope coefficient (for gestation, the X variable) are equivalent for simple linear regression (SLR) and they both answer the question

"Is Y (protein) related to X (gestation)?"

They are not the only inferences we can make using the summary output.

Say the (a priori) question had been "do the protein levels increase by more than 0.02 mg/mL for each of gestation?

- Step I: $H_0 : \beta_1 = 0.02, H_A : \beta_1 > 0.02$
- Step II: $t = \frac{\hat{\beta_1} - \beta_1 | H_0}{se(\hat{\beta_1}}$
  $se(\hat{\beta_1})$ these are still the values given in the summary output.
- Step III: $\alpha = 0.05$, reject H_0 if observed $t > t_{17}(0.95)$, use code qt(...).
- Step IV: Put 0.95 in the lower tail, $\alpha = 0.05$ to right. so $t_{17}(0.95) = 1.79$ (theoretical t-statistics), but the observed t-statistics is 0.86, with p-value 0.2.
  So **not reject** H_0.
- Step V: As observed $t = 0.086 \not> 1.74 = t_{17}(0.95)$
  OR
  as $p = 0.20 > \alpha = 0.05$
  DO NOT REJECT H_0 & conclude that the expected increase in protein levels is of the order of 0.02 mg/mL for each additional week of gestation, but is NOT significantly greater than that.

T-test on the intercept coefficient

- Step I: $H_0 : \beta_0 = 0$ vs $H_A : \beta_0 \neq 0$

- Step II: $t = \frac{\hat{\beta_0} - 0}{se(\hat{\beta_0}} \sim t_{n-2}$
  where se is $s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}$

- Step III: $\alpha = 0.05$; reject H_0 if observed $t < t_{17}(0.025)$ or observed $t > t_{17}(0.975)$.

- Step IV: two tail t-test. two sides each have a tail with $\frac{\alpha}{2} = 0.025$
  In fact, we have observed $t = 2.42 > 2.11 = t_{17}(0.025)$
  What about p-value approach? p-value is 0.027.

- Step V: As $p = 0.027 < \alpha = 0.05$
  reject H_0 and conclude $H_A : \beta_0 \neq 0$.

The plot shows that $\hat{\beta_0}$ (the intercept) is not zero.

But the (possible) true relationship could be exponential(?), as it passes through our **range of data** as well.

So **regression models are at test a good "local linear approximation".**