STAT 410 - Linear Regression

Lecture 7

Meng Li

Department of Statistics

Sep. 26, 2017

ANOVA

- We have covered marginal tests and confidence intervals for $\boldsymbol{\beta}$ in multiple linear regression.
- How to conduct a test for all $\boldsymbol{\beta}$ at significance level α, for example, the significance test of regression

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \text{ not } H_0?$$

- Possible solution: test each $H_0^j : \beta_j = 0$ at α separately and reject H_0 if we reject any of H_0^j.
- Analysis of variance (ANOVA) provides a general technique to compare multiple population means, and particularly can be used to test significance of regression.

ANOVA in SLR

- Starting with the identity $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$, we have

$$\begin{aligned}\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} \left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right]^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2\underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{=0 \text{ (Fact)}}\end{aligned}$$

- Therefore, we obtain

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.$$

  - $SS_T=\sum_{i=1}^{n}(y_i-\bar{y})^2$ measures the total variability in the observations thus is called the **corrected sum of squares of the observations**.
  - $SS_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ measures to amount of variability in the observations accounted for by the regression line thus is called the **regression or model sum of squares**.
    - In SLR, we have $SS_R=\hat{\beta}_1 S_{xy}$.
  - $SS_{Res}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ is the **residual or error sum of squares**, which measures the residual variation left unexplained by the regression line.
- Using the notation, ANOVA decomposition becomes

$$SS_T=SS_R+SS_{Res}.$$

Coefficient of Determination

- Coefficient of Determination (R^2):

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

- R^2 is the proportion of variation explained by the regressor(s).
- Because $0 \leq SS_{Res} \leq SS_T$, it follows that

$$0 \leq R^2 \leq 1.$$

- Values of R^2 that are close to 1 imply that most of the variability in observations is explained by the regression model.
- In single linear regression, we have $R^2 = r^2$, where r is the correlation coefficient between $(y_1, \ldots, y_n)$ and $(x_1, \ldots, x_n)$:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

ANOVA table

Source of Variation	SS	df	MS	E(MS)	F_0
Regression	SS_R	1	MS_R	$\sigma^2 + \beta_1^2 S_{xx}$	MS_R/MS_{Res}
Residual	SS_{Res}	$n-2$	MS_{Res}	σ^2	
Total	SS_T	$n-1$			

- How to obtain each block in the ANOVA table above?

F test

- The column of E(MS) in the ANOVA table inspires an alternative test for $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.
- Test statistic:

$$F_0 = \frac{\mathrm{MS}_R}{\mathrm{MS}_{Res}} = \frac{SS_R \ / \ 1}{SS_{Res} \ / \ (n-2)} \sim F(1, n-2).$$

  - The ratio of two independent χ^2 random variables is distributed as an F distribution:

$$\frac{\chi^2_{\nu_1}/\nu_1}{\chi^2_{\nu_2}/\nu_2} \sim F_{\nu_1,\nu_2}.$$

  - Under normal assumptions in SLR, we have

$$SS_R \perp\!\!\!\perp SS_{Res}, \quad \frac{SS_R}{\sigma^2} \sim \chi^2_1, \quad \frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-2}.$$

- Reject H_0 if $F_0 > F_{\alpha,1,n-2}$.

- Recall the t test:

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{MS_{Res}/S_{xx}}}.$$

- It leads to

$$t_0^2 = \frac{\hat{\beta}_1^2}{MS_{Res}/S_{xx}} = \frac{\hat{\beta}_1 S_{xy}}{MS_{Res}} = \frac{MS_R}{MS_{Res}} = F_0.$$

- Thus t_0^2 is identical to F_0 in the ANOVA approach.
- In general, if $Z \sim t_m$, then $Z^2 \sim F_{1,m}$.
- Consequently, the t test is equivalent to the F test for $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.
- However, F tests consider only the two-sided alternative while t tests are also applicable for one-sided alternatives.

ANOVA in MLR

Source of Variation	SS	df	MS	E(MS)	F_0
Regression	SS_R	k	MS_R	$\sigma^2 + \frac{\boldsymbol{\beta}^* \mathbf{X}_c' \mathbf{X}_c \boldsymbol{\beta}^*}{k\sigma^2}$	MS_R/MS_{Res}
Residual	SS_{Res}	$n-k-1$	MS_{Res}	σ^2	
Total	SS_T	$n-1$			

- Hypothesis testing of interest:

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \text{ not } H_0$$

- F test:

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}}.$$

- Under H_0, we have $F_0 \sim F_{k,n-k-1}$.
- Reject H_0 if $F_0 > F_{\alpha,k,n-k-1}$.