STAT 425

Multiple Linear Regression. Part 3

Hypothesis testing in MLR

Testing a single predictor: Suppose you have a certain number of predictors in your regression model and you want to test the hypothesis[1]:

$$H_0 : \beta_j = c \;\; vs. \;\; H_a : \beta_j \neq c$$

- We use the t-test statistic:

$$t = \frac{\hat{\beta}_j - c}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j - c}{\hat{\sigma}\sqrt{[(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}]_{jj}}} \sim T_{n-p}$$

  under the null hypothesis H_0

- p-value$= 2\times$ the area under the curve of a T_{n-p} distribution more extreme than the observed statistic.
- The p-value returned by the *lm* function command is for $c = 0$.

[1] The test result might vary depending on which other predictors are included in the model

Different t-tests

We've learned various t-tests in class and each seems to have a different degree of freedom. How can I find out the correct df for a t-test?

All t-tests we've encountered so far involve an estimate of the error variance σ^2. The df of a t-test is determined by the denominator of $\hat{\sigma}^2$.

- $Z_1, \dots, Z_n \sim \mathsf{N}(\theta, \sigma^2)$. To test $\theta = a$, we have

$$\frac{\hat{\theta} - a}{\mathsf{se}(\hat{\theta})} = \frac{\bar{Z} - a}{\sqrt{\hat{\sigma}^2 / n}} \sim T_{n-1}, \quad \hat{\sigma}^2 = \frac{\sum_i (Z_i - \bar{Z})^2}{n-1}.$$

- For SLR, to test $\beta_1 = c$, we have

$$\frac{\hat{\beta}_1 - c}{\mathsf{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - c}{\hat{\sigma}/\sqrt{\mathsf{Sxx}}} \sim T_{n-2}, \quad \hat{\sigma}^2 = \frac{\mathsf{RSS}}{n-2}.$$

- For MLR with p predictors (including the intercept), to test $\beta_j = c$,

$$\frac{\hat{\beta}_j - c}{\mathsf{se}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - c}{\hat{\sigma}\left[(\mathbf{X}^t\mathbf{X})^{-1}\right]_{jj}} \sim T_{n-p}, \quad \hat{\sigma}^2 = \frac{\mathsf{RSS}}{n-p}.$$

F-test and ANOVA table

Testing all predictors: Suppose we want to the test hypothesis:

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_p = 0 \;\; vs. \;\; H_a : \;\; \beta_j \neq 0$$

for some j, $j = 2, \ldots, p$.
Under the Null hypothesis, the test statistic:

$$F = \frac{MS(Reg)}{MS(Error)} \sim F_{p-1,n-p}$$

All *F-test* components can be organized in the ANOVA table, where $TSS = FSS + RSS$

ANOVA Table for the overall *F-test*

Source	df	SS	MS	F
Regression	$p-1$	FSS	FSS$/(p-1)$	MS(reg)/MS(err)
Error	$n-p$	RSS	RSS$/(n-p)$	
Total	$n-1$	TSS		

Savings example

- Suppose we start our analysis with the full model:

$$y_i = \beta_1 + \beta_2 pop15_i + \beta_3 pop75_i + \beta_4 dpi_i + \beta_5 ddpi_i + e_i$$

- We want to test the hypothesis that *saving* is independent of age
- We fit a reduced model. This implies to remove the columns corresponding to variables *pop15* and *pop75* from the design matrix:

$$y_i = \beta_1 + \beta_4 dpi_i + \beta_5 ddpi_i + e_i$$

- How can we compare the results from the two fitted models?

Savings example (Cont.)

We want to test the hypothesis:

$$H_0 : \text{The reduced model is adequate (age is not needed)}$$
$$H_a : \text{The full model is required}$$

Under H_0 we assume that the following model is correct:

$$y_j = \beta_1 + \beta_4 dpi + \beta_5 ddpi + e_i$$

We consider the following partition of the design matrix into two sub-matrices $\mathbf{X_1}$ and $\mathbf{X_2}$:

$$\mathbf{X}_{n\times p} = (\mathbf{X}_{1n\times(p-q)}, \mathbf{X}_{2n\times q})$$

The corresponding partition of the regression parameter is:

$$\boldsymbol{\beta}^t = (\boldsymbol{\beta}_1^t, \boldsymbol{\beta}_2^t)$$

where $\boldsymbol{\beta}_1$ is $(p-q)\times 1$ and $\boldsymbol{\beta}_2$ is $q\times 1$ This partition is used to test the hypothesis:

$$\begin{aligned} H_0 :& \boldsymbol{\beta}_2 = \mathbf{0}, \quad \text{i.e.,} \quad \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \textbf{error} \\ H_a :& \boldsymbol{\beta}_2 \neq \mathbf{0}, \quad \text{i.e.,} \quad \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \textbf{error} \end{aligned}$$

Partial F test

We use the test statistic:

$$F = \frac{(RSS_0 - RSS_a)/q}{RSS_a/(n-p)} \sim F_{q,n-p}$$

where RSS_0 = Residual sum of squares for the model under H_0;
RSS_a = Residual sum of squares for the model under H_a.

- Numerator: variation in the data not explained by the reduced model, but explained by the full model.
- Denominator: variation in the data not explained by the full model (i.e., not explained by either model), which is used to estimate the error variance.
- Reject H_0, if *F-stat* is large, that is, the variation missed by the reduced model, when being compared with the error variance, is significantly large.

Partial F test calculation using the **anova (.)** function

```
anova(reducedmodel, fullmodel)
```

```
## Analysis of Variance Table
##
## Model 1: sr ~ dpi + ddpi
## Model 2: sr ~ pop15 + pop75 + dpi + ddpi
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     47 824.72
## 2     45 650.71  2       174 6.0167 0.004835 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null hypothesis that the reduced model is correct.

Partial F test calculation using summary outputs from the two models

```
#
rss.full=sum(fullmodel$res^2)
# rss.full=deviance(fullmodel) # you can use "deviance" to extract RSS
rss.reduced=sum(reducedmodel$res^2)
#rss.reduced = deviance(reducedmodel)
Fstat=(rss.reduced-rss.full)/2/(rss.full/45)
Fstat
```

```
## [1] 6.016652
```

```
1-pf(Fstat, 2, 45)
```

```
## [1] 0.004834923
```

Examples of F-tests

- Example 1: Testing all predictors (The default F-test returned by the function lm(.)):

$$
\begin{aligned}
H_0 : & \mathbf{y} = \mathbf{1}_n \alpha + \textbf{error} \\
H_a : & \mathbf{y} = \mathbf{X}_{n \times p} \boldsymbol{\beta} + \textbf{error}
\end{aligned}
$$

- Example 2: Testing one-predictor (the F-test that is equivalent to the t-test $(H_0 : \beta_j = 0)$):

$$
\begin{aligned}
H_0 : & \mathbf{y} = \mathbf{X}[, -\mathbf{j}]_{n \times (p-1)} \alpha + \textbf{error} \\
H_a : & \mathbf{y} = \mathbf{X}_{n \times p} \boldsymbol{\beta} + \textbf{error}
\end{aligned}
$$

where $\mathbf{X}[, -\mathbf{j}] = \mathbf{X}$ without the j-th column, and $\boldsymbol{\alpha}$ is $(p-1) \times 1$

Examples of F-tests (Cont.)

- Example 3: Testing a subset of predictors:

$$
\begin{aligned}
H_0 :& \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + error \\
H_a :& \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + error
\end{aligned}
$$

  where $(\mathbf{X_1}, \mathbf{X_2})$ is a partition of matrix $\mathbf{X}$

- Example 4: Testing a sub-space (For example $H_0 : \beta_2 = \beta_3$)

$$
\begin{aligned}
H_0 :& \mathbf{y} = \mathbf{X}_1\alpha + \mathbf{error} \\
H_a :& \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{error}
\end{aligned}
$$

  where $\mathbf{X}_1$ is a $n \times (p-1)$ matrix that is almost the same as $\mathbf{X}$, but replaces the 2nd and 3rd columns of $\mathbf{X}$ by their sum, and $\boldsymbol{\alpha}$ is $(p-1) \times 1$.