Multiple Linear Regression (Part III)

Lecture 6

Alexandra Chronopoulou

ILLINOIS

COLLEGE OF LIBERAL ARTS & SCIENCES
Department of Statistics
101 Illini Hall, MC-374
725 S. Wright St.
Champaign, IL 61820-5710

Learning objectives

In this lecture we will:

- discuss hypothesis testing in MLR.
- define partial F-tests.
- introduce permutation tests.

Testing predictors in MLR

Hypothesis testing in MLR

(a) Testing a single predictor/coefficient.

(b) Testing multiple (a subset) predictors/coefficients.

(c) Testing all predictors/coefficients.

(d) Other hypothesis tests.

Testing a Single Predictor (Coefficient)

- Suppose you have a p predictors in your regression model and you want to test the hypothesis[1]:

$$H_0 : \beta_j = c \ \ vs. \ \ H_\alpha : \beta_j \neq c$$

  - The t-test statistic we use is:

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

    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

Remark: Degrees of Freedom of a t-test

The *degrees of freedom* of a t-test are determined by the denominator of the estimated variance $\hat{\sigma}^2$. Consider the following situations:

- *In STAT 400:* Test for $\theta = \alpha$, where $Z_1, \ldots, Z_n \sim \mathcal{N}(\theta, \sigma^2)$

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

- *In SLR:* Test for $\beta_1 = c$, we have

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

- *In MLR with p predictors (including the intercept)*: Test for $\beta_j = c$,

$$\frac{\hat{\beta}_j - c}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j - c}{\hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}} \sim T_{n-p}, \hat{\sigma}^2 = \frac{RSS}{n-p}$$

Overall F-test in MLR

Testing all predictors

$$\begin{cases} H_0 : \beta_2 = \beta_3 = \ldots = \beta_p = 0 \\ H_a : \quad \beta_j \neq 0, \quad \text{for some } j,\, j = 2, \ldots, p \end{cases}$$

- *Under the Null hypothesis*, the test statistic:

$$\begin{aligned} F &= \frac{FSS(X_2, \ldots, X_p)}{p-1} \div \frac{RSS(X_2, \ldots, X_p)}{n-p} \\ &= \frac{MS(Reg)}{MS(Error)} \sim F_{p-1, n-p} \end{aligned}$$

  Large values of F lead to conclusion H_α.

- This is the *overall F test* of whether or not there is a regression relation between the response variable Y and the set of X variables.

ANOVA Table for the overall F-test

Source	df	SS	MS	F-test
Regression	$p-1$	FSS	FSS$/(p-1)$	MS(Reg)/MSE
Error	$n-p$	RSS	RSS$/(n-p)$	
Total	$n-1$	TSS		

`Bike Shares` Example

– We started our analysis with the full model:

$$
\begin{aligned}
(\text{BikeShares})_\text{i} = \beta_0 + \beta_1\, (\text{RealTemp})_\text{i} + \beta_2\, (\text{FeelsLikeTemp})_\text{i} \\
+ \beta_3\, (\text{Humidity})_\text{i} + \beta_4\, (\text{WindSpeed})_\text{i} + \varepsilon_\text{i}
\end{aligned}
$$

Using abstract notation, the model we considered can be written:

$$y_i = \beta_0 + \beta_1\, X_{i1} + \beta_2\, X_{i2} + \beta_3\, X_{i3} + \beta_4\, X_{i4} + \varepsilon_i$$

– We want to test the hypothesis that the response `(BikeShares)` is independent of `(Temperature)`, that is the variables `(RealTemp)`, i.e. X_1, and `(FeelsLikeTemp)`, i.e. X_2. In other words, the alternatives are:

$$
\begin{cases}
H_0 : \beta_1 = \beta_2 = 0 \\
H_\alpha : \text{ not both } \beta_1 \text{ and } \beta_2 \text{ equal } 0
\end{cases}
$$

Test whether Several $\beta_k = 0$

- We fit a reduced model. This implies to *remove* the columns corresponding to variables `RealTemp` and `FeelsLikeTemp` from the design matrix:

$$y_i = \beta_0 + \beta_3 \, X_{i3} + \beta_4 \, X_{i4} + \varepsilon_i$$

- How can we compare the results from the two fitted models? Essentially, the hypothesis for the β's is equivalent to the following:

$$\begin{cases} H_0 : & \text{The reduced model is adequate (Temp is not needed)} \\ H_\alpha : & \text{The full model is required} \end{cases}$$

Test whether Several $\beta_k = 0$

General Linear Test

$$\begin{cases} H_0 : Y_i = \beta_0 + \beta_1 X_{i3} + \beta_4 X_{i4} + \varepsilon_i \text{ Reduced Model} \\ H_\alpha : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i \text{ Full Model} \end{cases}$$

$$F = \frac{RSS(X_3, X_4) - RSS(X_1, X_2, X_3, X_4)}{(n-3) - (n-5)} \div \frac{RSS(X_1, X_2, X_3, X_4)}{n-5}$$

$$\sim F_{2, n-5}$$

Partial F test

- In general, consider the following *partition* of the design matrix into two sub-matrices $\mathbf{X_1}$ and $\mathbf{X_2}$, that is

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

- The corresponding *partition* of the regression parameter is:

$$\beta^T = (\beta_1^T, \beta_2^T)$$

  where β_1 is $(p - q) \times 1$ and β_2 is $q \times 1$.

- This partition is used to test the hypothesis:

$$\begin{cases} H_0 : \beta_2 = \mathbf{0}, \quad \text{i.e.,} \quad \mathbf{y} = \mathbf{X}_1 \beta_1 + \textbf{error} \\ H_\alpha : \beta_2 \neq \mathbf{0}, \quad \text{i.e.,} \quad \mathbf{y} = \mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \textbf{error} \end{cases}$$

Partial F test

- To test this hypothesis, the test statistic is:

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

  where $RRS_0 =$ Residual sum of squares for the model under H_0; $RRS_a =$ Residual sum of squares for the model under H_α.
- 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 test statistic is large, that is, the variation missed by the reduced model, when being compared with the error variance, is significantly large.

Partial F test in R: Bike Shares Example

- In R, the partial F test calculation is done using the `anova (.)` function:

```
bikeshare.mlr.full = lm(cnt ~ t1 + t2+ hum + wind_speed, data=newbikeshares.reg )
bikeshare.mlr.reduced = lm(cnt ~ hum + wind_speed , data=newbikeshares.reg )

anova(bikeshare.mlr.reduced, bikeshare.mlr.full)
```

```
## Analysis of Variance Table
##
## Model 1: cnt ~ hum + wind_speed
## Model 2: cnt ~ t1 + t2 + hum + wind_speed
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1  17411 1.6103e+10
## 2  17409 1.5250e+10  2 853010396 486.88 < 2.2e-16 ***
## ---
## 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(bikeshare.mlr.full$res^2)
# You can also compute it with
# rss.full = deviance(bikeshare.mlr.full)
rss.reduced = sum(bikeshare.mlr.reduced$res^2)
# rss.reduced = deviance(bikeshare.mlr.reduced)
Fstat = (rss.reduced - rss.full)/2/(rss.full/17409)
Fstat
```

```
## [1] 486.8763
```

```
1-pf(Fstat, 2, 17409)
```

```
## [1] 0
```

Examples of F-tests

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

$$\begin{cases} H_0 : \mathbf{y} = \mathbf{1}_n\alpha + \textbf{error} \\ H_\alpha : \mathbf{y} = \mathbf{X}_{n\times p}\beta + \textbf{error} \end{cases}$$

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

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

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

Examples of F-tests (Cont.)

- Testing a subset of predictors:

$$\begin{cases} H_0 : \mathbf{y} = \mathbf{X}_1\beta_1 + \textbf{error} \\ H_\alpha : \mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \textbf{error} \end{cases}$$

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

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

$$\begin{cases} H_0 : \mathbf{y} = \mathbf{X}_1\alpha + \textbf{error} \\ H_\alpha : \mathbf{y} = \mathbf{X}\beta + \textbf{error} \end{cases}$$

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 α is $(p-1) \times 1$.

Comments

(a) For testing whether a single β_k equals zero, two equivalent test statistics are available: the t test statistic and the F linear test statistic. When testing whether several β_k are equal to zero, only the general linear test statistic F is available.

(b) The general linear test statistic for testing whether several X variables can be dropped from the general linear regression model can be expressed in terms of the coefficients of multiple determination for the full and reduced models. Denoting these by R_F^2 and R_R^2 respectively, we have:

$$F = \frac{R_F^2 - R_R^2}{df_R - df_F} \div \frac{1 - R_F^2}{df_F}$$

Note that this test statistic is not appropriate when the full and reduced regression models do not contain the intercept term β_0.

Permutation Tests

Hypothesis Testing when data is not Normal

Question

- How do we test hypotheses in MLR when the distribution of the data is not Normal?
- How do we test hypotheses in MLR, when the distribution of the data is unknown?

Answer

- We use the so-called *permutation tests*.

- A test statistic is a function of the data; denote it $g(data)$.
- The test statistic tends to take *extreme* values under the alternative hypothesis H_α.

Procedure to conduct a permutation test

1. Form the test statistic $g(data)$ which tends to take extreme values under the alternative hypothesis.
2. Evaluate the test statistic on the observed data, denoted by g_0.
3. Find the distribution of $g(data)$, when data are generated from H_0.
4. Calculate the p-value, that is the following probability:

$$\mathbb{P}\left(g(data) \text{ is more extreme than the observed } g_0 \middle| \text{ data } \sim H_0\right)$$

Calculation of the p-value

p-value

$$\mathbb{P}\left(g(data) \text{ is more extreme than the observed } g_0 \middle| \text{ data } \sim H_0\right)$$

- If the distribution of the data under the H_0 is not normal, how can we compute the p-value?
- We can generate data from H_0 and then calculate the p-value for the corresponding test statistic, using the *Monte Carlo method*.

- Suppose the pdf (or pmf) of a random variable Y does not have a simple form, therefore it is not easy to calculate $\mathbb{E}(Y)$ explicitly.
- But suppose it is easy to write a short **R** script to generate such a random variable, i.e. simulate it.
- We can obtain an approximation of $\mathbb{E}(Y)$ as follows:
  1. Generate $N = 1000$ samples from this distribution, $Y_1, \ldots, Y_N$,
  2. Approximate the mean by

$$\mathbb{E}(Y) \approx \frac{1}{N} \sum_{i=1}^{N} Y_i$$

That is, population mean $\approx$ sample mean (when N is large).

Monte Carlo Method (Cont.)

– This method also works if we want to approximate the expected value of a *function* of a random variable:

$$\mathbb{E}\left(f(Y)\right) \approx \frac{1}{N}\sum_{i=1}^{N} f(Y_i)$$

Examples

(a) We can use MC to compute the variance of a r.v.

$$Var(Y) = \mathbb{E}\left(Y^2\right) - \left(\mathbb{E}\left(Y\right)\right)^2$$

(b) or a probability, since

$$\mathbb{P}\left(Y > a\right) = \mathbb{E}\left(\mathbf{1}_{\{Y>a\}}\right),$$

where $\mathbf{1}_{\{.\}}$ is the indicator function.

R code for the `Bike Shares` data set

- Under H_0, `RealTemp` and `FeelsLikeTemp` are not useful in explaining the variation in the response `BikeShares`.
- Every pair of values `RealTemp`, `FeelsLikeTemp` for a particular share can be assigned to any other share, assuming that these variables do not have an effect on the response (`sample(.)` function).
- We can fit a full MLR model for each permutation and obtain many values of the F-statistic under H_0 (loop in code).
- We can estimate the p-value associated to the observed data using the Monte Carlo method.

Permutation test in R: Code

H_0: Variables `RealTemp` and `FeelsLikeTemp` (columns 3 and 4 in the data frame) are not significant

```
n.iter = 2000;
fstats = numeric(n.iter);
for(i in 1:n.iter){
  newbikes = newbikeshares.reg;
  newbikes[, c(3,4)] = newbikeshares.reg[sample(17414), c(3,4)];
  ge = lm(cnt ~ t1 + t2+ hum + wind_speed, data=newbikes);
  fstats[i] = summary(ge)$fstat[1]
}

# Estimated p-value
length(fstats[fstats > summary(bikeshare.mlr.full)$fstat[1]])/n.iter
```

The estimated p-value is the probability of observing a value as extreme as g_0 (F-stat of the observed data for the full model).