R Lab3
Hypothesis Testing of Regression Coefficient
PH 1915, Fall 2013

Ming Yang

UTSPH

November 7, 2013

Table of Contents

Type of tests

- Global test
- Test of a single covariate
- Test a group of covariates

Global test

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ik}\beta_k + \varepsilon_i$$

- Question to answer: does the *entire* set of independent variables contribute significantly to the prediction of **y**?
- An overall test of the significant effect of the regression predictors
- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$
  $H_1 : \beta_j \neq 0$ for at least one j, $j = 1, \cdots, k$
- Reject H_0 implies that at least one of the covariates contributes significantly to the model.

Global test

The ANOVA table

Source of variation	Sum of squares	Degrees of freedom
Regression	$SSR(x_1)$	1
	$SSR(x_2 \mid x_1)$	1
	$\vdots$	$\vdots$
	$SSR(x_k \mid x_{k-1}, x_{k-2}, \cdots, x_1)$	1
	SSR	k
Error	SSE	n-(k+1)
Total	SST	n-1

where

$SSR = SSR(x_1) + SSR(x_2|x_1) + \cdots + SSR(x_k|x_{k-1}, x_{k-2}, \cdots, x_1) = \hat{\beta}'\mathbf{X}'\mathbf{y} - n\bar{y}^2$

$SSE = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y}$ (for the full model)

$SST = \mathbf{y}'\mathbf{y} - n\bar{y}^2$ (stays the same for all models)

Global test

Under the null hypothesis, $SSR/\sigma^2 \sim \chi^2_k$ and $SSE/\sigma^2 \sim \chi^2_{n-k-1}$ are independent. Therefore we have

$$TS = \frac{SSR/k}{SSE/(n-k-1)} \sim F_{k,n-k-1}$$

$p - value = Pr(F_{k,n-k-1} > TS)$.

Global test

Example:

$$mgp_i = \beta_0 + hp_i\beta_1 + wt_i\beta_2 + \varepsilon_i$$

$H_0 : \beta_1 = \beta_2 = 0$, H_1:at least one $\beta \neq 0$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
hp	1	678.37	678.37	100.86	0.0000
wt	1	252.63	252.63	37.56	0.0000
Residuals	29	195.05	6.73		

$$TS = \frac{(678.37 + 252.63)/2}{195.05/29} = 69.21 > F_{2,29,0.95} = 3.33$$

Thus, we reject the null at 0.05 significance level and conclude that at least one β_1 and β_2 is not equal to 0.

Global test

Example cont.

The overall F statistic is also available from the output of `summary()`

```
> summary(fit.all)

Call:
lm(formula = mpg ~ hp + wt, data = mtcars)
Residuals:
   Min     1Q Median     3Q    Max
-3.941 -1.600 -0.182  1.050  5.854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
---

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268, Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

Testing of single β_j

Once we have determined that at least one of the regressors is important, a natural next question might be which one(s)?

Important considerations:

- Is the increase in the regression sums of squares sufficient to warrant an additional predictor in the model?
- Additional predictors will increase the variance of $\hat{y}$ - include only predictors that explain the response (note: we may not know this through hypothesis testing as confounders may not test significant but would still be necessary in the regression model).
- Adding an unimportant predictor may increase the residual mean square thereby reducing the usefulness of the model.

Testing of single β_j

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ij}\beta_j + \cdots + x_{ik}\beta_k + \varepsilon_i$$

- Question to answer: does one particular variable of interest significantly affect the prediction of **y** when the other independent variables presented in the model?
- $H_0 : \beta_j = 0$, $H_1 : \beta_k \neq 0$
- $TS = \frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)} \sim t_{n-k-1}$, reject H_0 if $|TS| > t_{n-k-1,1-\alpha/2}$
- This is a **partial test** because $\hat{\beta}_j$ depends on all of the other predictors x_i, for $i \neq j$, that are in the model. Thus, this is a test of the contribution of x_j given other predictors in the model.

Testing of single β_j

Example cont.:

$$mgp_i = \beta_0 + hp_i\beta_1 + wt_i\beta_2 + \varepsilon_i$$

$H_0 : \beta_2 = 0,\ H_1 : \beta_2 \neq 0$

From the summary of lm $\hat{\beta}_2 = -3.88$, the variance and covariance matrix of the parameter estimates is

```
> vcov(fit.all)
              (Intercept)            hp          wt
(Intercept)  2.5561215917  1.484701e-04 -0.73594515
hp           0.0001484701  8.153566e-05 -0.00376369
wt          -0.7359451464 -3.763690e-03  0.40035167
```

$$TS = \frac{-3.88}{\sqrt{0.40}} = -6.13 < t_{29,0.025} = -2.05$$

Thus, we reject the null and conclude that $\beta_2 \neq 0$.

Testing of a subset of β

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ij}\beta_j + \cdots + x_{ip}\beta_p + \cdots + x_{ik}\beta_k + \varepsilon_i$$

- Often it is of interest to determine whether a group of predictors contribute to predicting y given another predictor or group of predictors that are in the model.
- $H_0 : \beta_j = \cdots = \beta_p = 0$, $H_1 : \beta_l \neq 0$ for at least one $l, l = j, \cdots, p$

Testing of a subset of β

Partition the vector of regression coefficient and **X** matrix as

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}, \mathbf{X} = [\mathbf{X_1} | \mathbf{X_2}]$$

Hypotheses of interest: $H_0 : \boldsymbol{\beta}_2 = 0$ v.s. $H_1 : \boldsymbol{\beta}_2 \neq 0$

The model can be written as $\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta}_1 + \mathbf{X_2}\boldsymbol{\beta}_2 + \varepsilon$

$$SSR(\mathbf{X}) = \hat{\beta}'\mathbf{X}'\mathbf{y} \text{ } (k+1 \text{ degrees of freedom})$$

$$MSE = \frac{\mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y}}{n-k-1}$$

$$\begin{aligned} SSR(\mathbf{X_2}|\mathbf{X_1}) &= SSR(\mathbf{X}) - SSR(\mathbf{X_1}) \\ &= SSE(reduced) - SSE(full) \text{ } (r \text{ degrees of freedom}) \end{aligned}$$

Under H_0

$$TS = \frac{SSR(\mathbf{X_2}|\mathbf{X_1})/r}{MSE} \sim F_{r,n-k-1}$$

Testing of a subset of β

Example cont.

$$mgp_i = \beta_0 + disp_i\beta_1 + hp_i\beta_2 + qsec_i\beta_3 + wt_i\beta_4 + \varepsilon_i$$

```
fit.sub<-lm(mpg~disp+hp+qsec+wt,data=mtcars)
> summary(fit.sub)

Call:
lm(formula = mpg ~ disp + hp + qsec + wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8664 -1.5819 -0.3788  1.1712  5.6468

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.329638   8.639032   3.164  0.00383 **
disp         0.002666   0.010738   0.248  0.80576
hp          -0.018666   0.015613  -1.196  0.24227
qsec         0.544160   0.466493   1.166  0.25362
wt          -4.609123   1.265851  -3.641  0.00113 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.622 on 27 degrees of freedom
Multiple R-squared:  0.8351, Adjusted R-squared:  0.8107
F-statistic: 34.19 on 4 and 27 DF,  p-value: 3.311e-10
```

Testing of a subset of β

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0,\ H_1 : \beta_j \neq 0, j = 1, 2, 3$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
disp	1	808.89	808.89	117.65	0.0000
hp	1	33.67	33.67	4.90	0.0356
qsec	1	6.71	6.71	0.98	0.3321
wt	1	91.15	91.15	13.26	0.0011
Residuals	27	185.64	6.88		

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
wt	1	847.73	847.73	91.38	0.0000
Residuals	30	278.32	9.28		

$$SSR(disp, hp, qsec|wt) = 278.32 - 185.64 = 92.68$$

$$TS = \frac{92.68/3}{6.88} = 4.49 > F_{3,27,0.95} = 2.96$$

Thus we reject the null and conclude that *disp*, *hp* and *qsec* are jointly significant.

Test of the contrast

Many functions in R can be used to test the contrasts.

function	package	description
fit.contrast	{gmodels}	Compute and test arbitrary contrasts for regression objects
contrast.lm	{contrast}	computes one or more contrasts of the estimated regression coefficients
glht	{multcomp}	generalized linear hypothesis test
linear.hypothesis	{car}	Generic function for testing a linear hypothesis

A simple example.

Regression Diagnostics

Frequently used functions provide information used with model diagnostics

fitted.values()	Returns fitted values
residuals()	Returns residuals
rstandard()	Standardized residuals, variance one; residual standardized using overall error variance (9.25)
rstudent()	Studentized residuals, variance one; residual standardized using leave-one-out measure of the error variance (9.26)
qqnorm()	Normal quantile plot
qqline()	Add a line to the normal quantile plot
plot.lm()	Given a lm object it produces six diagnostic plots, selected using the 'which' argument; default is plots 1-3 and 5 1.Residual versus fitted values 2. Normal quantile-quantile plot 3. $\sqrt{\lvert\text{Standardized residuals}\rvert}$ versus fitted values 4. Cook's distance versus row labels 5.Standardized residuals versus leverage along with contours of Cook's distance

plot.lm()	6. Cook's distance versus leverage/(1-leverage) with $\sqrt{\lvert\text{Standardized residuals}\rvert}$ contours
dffits()	Return DFFITS
dfbeta()	Return DFBETAS
covratio()	Return covariance ratio; vector whose *i*th element is the ratio of the determinants of the estimated covariance matrix with and without data point *i*
cooks.distance()	Returns Cook's distance
hatvalues()	Diagonal of the hat matrix
influence.measures()	Returns the previous five measure of influence and flags influential points
lm.influence()	Returns four measures of influence:
hat	Diagonal of the hat matrix, measure of leverage
coefficients	Matrix, whose *i*th row contains the change in the estimated coefficients when the *i*th case is removed
sigma	Vector, whose *i*th element contains the estimated of the residual standard error when the *i*th case is removed
wt.res	Vector of weighted residuals or raw residuals if weights are not set.

Regression Diagnostics

Example:

```
fit<-lm(mpg~wt,data=mtcars)
#influential points are labeled
par(mfrow=c(2,2))
plot(fit) #returns four diagnostics plot (1-3 and 5)
par(mfrow=c(2,3))
plot(fit,which=1:6) #returns all six diagnostic plots

par(ask=T)
plot(residuals(fit),fitted.values(fit))
qqnorm(residuals(fit));qqline(residuals(fit))
plot(cooks.distance(fit),rownames(fit),type="h")

#influence measures
influence.measures(fit)

#extract influential points, uses $is.inf
inf.temp<-influence.measures(fit)
inf.pts<-which(apply(inf.temp$is.inf,1,any))
mtcars[inf.pts,]

#Influence measures
lm.influence(fit)
```

Regression Diagnostics

```
#Extract points that cause the greatest change in the estimates
lm.inf.coef<-lm.influence(fit)$coefficients
lm.inf.pts<-apply(lm.inf.coef[,2,drop=F],2,
+ FUN=function(x)which.max(abs(x)))

lm.inf.coef[lm.inf.pts,]
#this gives the same results with the diagnostic plots

#Get the five points that cause the greatest
#change in the estimates
lm.inf.pts.top5<-apply(lm.inf.coef,2,
+ FUN=function(x)names(rev(sort(abs(x)))[1:5]))
lm.inf.pts.top5
```

Sums of Squares

- Type I
  - Also called "sequential" sum of squares
  - Can be viewed as the reduction in SSE obtained by adding additional term to a fit that already includes the terms listed before it.
  - Pros: a complete decomposition of the predicted SS for the whole model; Preferable when some factors should be taken out before other factors.
  - Cons:Lack of invariance to order of entry into the model; not appropriate for factorial designs.
- Type II
  - The reduction in SSE due to adding the term to the model after all other terms except those that contain it (interaction terms).
  - Pros: Appropriate for model building and natural choice for regression; Most powerful when no interaction; Invariant to the order when the factors are entered to the model.
  - Cons:Not appropriate for factorial designs
- Type III
  - Effect of each variable is evaluated after all other factors have been accounted for.
  - Pros: Appropriate for unbalanced data;
  - Cons: Testing main effects when interactions presence; not appropriate with missing cells.

LRT and F test

The F-test of the null hypothesis $H_0 : \mathbf{C}\beta = \mathbf{t}$ is a likelihood ratio test (LRT) because the F-ratio is a monotone transformation of the likelihood ratio λ.

Proof:

The log-likelihood is given by

$$
\begin{aligned}
\log L(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}SSE(\mathbf{X}) \\
\lambda = -2\log\frac{\max_{H_0} L(\boldsymbol{\beta})}{\max_{H_1 UH_0} L(\boldsymbol{\beta})} &= -2\log\frac{L(\hat{\boldsymbol{\beta}}_1)}{L(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)} \\
&= \frac{SSE(\mathbf{X_1}) - SSE(\mathbf{X_1} + \mathbf{X_2})}{\sigma^2},
\end{aligned}
$$

for a fixed value of σ^2.

Since σ^2 is unknown, we can use $\hat{\sigma}^2_{MLE} = SSE(X_1 + X_2)/n$, then

$$
F = C * \lambda = \frac{[SSE(\mathbf{X_1}) - SSE(\mathbf{X_1} + \mathbf{X_2})]/r}{SSE(\mathbf{X_1} + \mathbf{X_2})/(n-k-1)} \sim F_{r,n-k-1},
$$

where $C = \frac{n-k-1}{nr}$.

References

1. Elizabeth R. Brown, *Introduction to Regression Models*
2. Nicholas Christian, *Statistical Computing in R*
3. Langsrud, ϕ. (2003), ANOVA for Unbalanced Data: Use Type II Instead of Type III Sums of Squares, *Statistics and Computing*, 13, 163-167.