JOHNS HOPKINS
BLOOMBERG SCHOOL of PUBLIC HEALTH

Lecture 8

Advanced inference in multiple linear regression

Review of where we left off

1. We have established the multiple linear regressio model:

$$Y_{n\times 1} = X_{n\times(p+1)}\beta_{(p+1)\times 1} + \epsilon_{n\times 1}, \epsilon_{n\times 1} \sim MVN(0_{n\times 1}, \sigma^2 I_{n\times n})$$

2. We know that:

$$\hat{\beta} \text{ satisfies } X'(Y - X\beta) = 0 \text{ and minimizes } \sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

3. We have defined:
   - $\hat{Y} = X\hat{\beta} = HY, \text{ where } H = X(X'X)^{-1}X'$
   - $\hat{R} = Y - \hat{Y} = Y - X\hat{\beta} = (I - H)Y$
4. Then we showed that:
   - $\hat{\beta} \sim MVN(\beta, \sigma^2(X'X)^{-1})$
   - $\hat{Y} \sim MVN(X\beta, \sigma^2 H)$
   - $\hat{R} \sim MVN(0, \sigma^2(I - H))$

Possible inference: single regression coefficient

Target	Estimate ~ Sampling Distn	95% CI for target	Test statistic for H0: Target = 0
β_j	$\hat{\beta}_j \sim N(\beta_j, [\sigma^2(X'X)^{-1})]_{jj})$	$\hat{\beta}_j \pm t \times \hat{se}(\hat{\beta}_j)$	$\frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)}$

Example: inference for single regression coefficient

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.12089    0.50959  21.823  < 2e-16 ***
## age          0.31141    0.09264   3.361 0.000945 ***
## agesp6      -0.27958    0.09441  -2.961 0.003473 **
## ---
```

Possible inference: linear combination of coefficients

Target	Estimate $\sim$ Sampling Distn	95% CI for target	Test statistic for H0: Target = 0
β_j	$\hat{\beta}_j \sim N(\beta_j, [\sigma^2(X'X)^{-1})]_{jj})$	$\hat{\beta}_j \pm t \times \hat{se}(\hat{\beta}_j)$	$\frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)}$
$A\beta$	$A\hat{\beta} \sim N(A\beta, \sigma^2 A(X'X)^{-1}A')$	$A\hat{\beta} \pm t \times \hat{se}(A\hat{\beta})$	$\frac{A\hat{\beta}_j}{\hat{se}(A\hat{\beta}_j)}$

Example: inference for linear combination of coefficients

Target	Estimate $\sim$ Sampling Distn	95% CI for target	Test statistic for H0: Target = 0
β_j	$\hat{\beta}_j \sim N(\beta_j, [\sigma^2(X'X)^{-1})]_{jj})$	$\hat{\beta}_j \pm t \times \hat{se}(\hat{\beta}_j)$	$\frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)}$
$A\beta$	$A\hat{\beta} \sim N(A\beta, \sigma^2 A(X'X)^{-1}A')$	$A\hat{\beta} \pm t \times \hat{se}(A\hat{\beta})$	$\frac{A\hat{\beta}_j}{\hat{se}(A\hat{\beta}_j)}$

Example: inference for linear combination of coefficients

```
cc=complete.cases(select(d,age,arm))
d.cc=filter(d,cc)
d.cc = arrange(d.cc,age)
reg1<-lm(data=d.cc, arm~age+agesp6)
reg1.coeff = reg1$coeff
reg1.vc = vcov(reg1)
```

```
# Define the linear combination of betas
A = matrix(c(0,1,1),nrow=1,ncol=3)
# Estimate the A beta-hat
A %*% reg1.coeff
```

```
##            [,1]
## [1,] 0.03182924
```

```
# What is the statistical variance of the estimate
A %*% reg1.vc %*% t(A)
```

```
##              [,1]
## [1,] 1.985802e-05
```

```
# What is the standard error of the estimate
sqrt(A %*% reg1.vc %*% t(A))
```

```
##             [,1]
## [1,] 0.004456234
```

Example: inference for linear combination of coefficients

```r
# Confirm these values!
summary(glht(reg1, linfct = A))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = arm ~ age + agesp6, data = d.cc)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0 0.031829   0.004456   7.143 2.12e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```r
# 95% CI for beta1 + beta2
A %*% reg1.coeff - qt(0.975,df=summary(reg1)$df[2]) * sqrt(A %*% reg1.vc %*% t(A))
```

```
##             [,1]
## [1,] 0.02303672
```

```r
A %*% reg1.coeff + qt(0.975,df=summary(reg1)$df[2]) * sqrt(A %*% reg1.vc %*% t(A))
```

```
##             [,1]
## [1,] 0.04062177
```

Example: inference for linear combination of coefficients

```r
# Confirm these values!
summary(glht(reg1, linfct = A))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = arm ~ age + agesp6, data = d.cc)
##
## Linear Hypotheses:
##          Estimate Std. Error t value Pr(>|t|)
## 1 == 0 0.031829   0.004456   7.143 2.12e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```r
# Hypothesis test of H0: beta1 + beta2 = 0
test.stat = (A %*% reg1.coeff) / sqrt(A %*% reg1.vc %*% t(A))
test.stat
```

```
##          [,1]
## [1,] 7.142632
```

```r
2 * pt(abs(test.stat),df=summary(reg1)$df[2],lower.tail=FALSE)
```

```
##              [,1]
## [1,] 2.124636e-11
```

Possible inference: Non-linear function of a coefficient

Target	Estimate $\sim$ Sampling Distn	95% CI for target	Test statistic for H0: Target = 0
β_j	$\hat{\beta}_j \sim N(\beta_j, [\sigma^2(X'X)^{-1})]_{jj})$	$\hat{\beta}_j \pm t \times \hat{se}(\hat{\beta}_j)$	$\frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)}$
$A\beta$	$A\hat{\beta} \sim N(A\beta, \sigma^2 A(X'X)^{-1}A')$	$A\hat{\beta} \pm t \times \hat{se}(A\hat{\beta})$	$\frac{A\hat{\beta}_j}{\hat{se}(A\hat{\beta}_j)}$
$g(\beta_j)$	$g(\hat{\beta}_j) \sim N(g(\beta_j), [g'(\beta_j)]^2[\sigma^2(X'X)^{-1}]_{jj})$	$g(\hat{\beta}_j) \pm t \times \hat{se}(g(\hat{\beta}_j))$	$\frac{g(\hat{\beta}_j)}{\hat{se}(g(\hat{\beta}_j))}$

Example: inference for non-linear function of a coefficient

Univariate delta method

Assuming the function g is continuous at its first derivative. The delta method is derived from the first order approximation to Taylor series using Taylor's theorem.

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

In statistical applications, we are interested in finding the distribution of $g(\hat{\theta})$ where $\hat{\theta}$ follows a normal distribution.

Applying the first order Taylor expansion to $g(\hat{\theta})$ about the mean θ, we get:

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta)$$

Then, $E(g(\hat{\theta})) = g(\theta) + g'(\theta)(E(\hat{\theta}) - \theta) = g(\theta) + g'(\theta - \theta) = g(\theta)$ and $Var(g(\hat{\theta})) = g'(\theta)^2 Var(\hat{\theta})$.

Possible inference: non-linear function of coefficients

Target	Estimate ~ Sampling Distn	95% CI for target	Test statistic for H0: Target = 0
β_j	$\hat{\beta}_j \sim N(\beta_j, [\sigma^2(X'X)^{-1})]_{jj})$	$\hat{\beta}_j \pm t \times \hat{se}(\hat{\beta}_j)$	$\frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)}$
$A\beta$	$A\hat{\beta} \sim N(A\beta, \sigma^2 A(X'X)^{-1}A')$	$A\hat{\beta} \pm t \times \hat{se}(A\hat{\beta})$	$\frac{A\hat{\beta}_j}{\hat{se}(A\hat{\beta}_j)}$
$g(\beta_j)$	$g(\hat{\beta}_j) \sim N(g(\beta_j), [g'(\beta_j)]^2[\sigma^2(X'X)^{-1}]_{jj})$	$g(\hat{\beta}_j) \pm t \times \hat{se}(g(\hat{\beta}_j))$	$\frac{g(\hat{\beta}_j)}{\hat{se}(g(\hat{\beta}_j))}$
$g(\beta)$	$g(\hat{\beta}) \sim N(g(\beta), g'(\beta)'[\sigma^2(X'X)^{-1}]g'(\beta))$	$g(\hat{\beta}) \pm t \times \hat{se}(g(\hat{\beta}))$	$\frac{g(\hat{\beta})}{\hat{se}(g(\hat{\beta}))}$

Example: inference for non-linear function of coefficients

Example: non-linear function of coefficients

```r
reg.coeff = reg$coeff
reg.vc = vcov(reg)

# Compute the estimate of g(beta)
g.est = 1 + reg.coeff[3]/reg.coeff[2]
# Define the vector of the derivative of g(beta) wrt beta
g.prime = matrix(c(0,-reg.coeff[3]/reg.coeff[2]^2,1/reg.coeff[2]),nrow=3,ncol=1)
g.prime
```

```
##          [,1]
## [1,] 0.000000
## [2,] 2.883012
## [3,] 3.211236
```

```r
# Compute the variance of g(beta.hat)
g.var = t(g.prime) %*% reg.vc %*% g.prime
g.est
```

```
##    agesp6
## 0.1022112
```

```r
g.est - qt(0.975,df=summary(reg)$df[2]) * sqrt(g.var)
```

```
##            [,1]
## [1,] 0.02689796
```

```r
g.est + qt(0.975,df=summary(reg)$df[2]) * sqrt(g.var)
```

```
##           [,1]
## [1,] 0.1775244
```

Comparing nested MLR models

F-test for nested models; ANOVA method

Define:

- RN
- RE
- Δ

You can show the following results (which we will not do in class):

- $H_E - H_N$ is idempotent with rank s
- $H_E - H_N$ is orthogonal to $(I - H_E)Y$
- $\frac{\Delta'\Delta/s}{R'_E R_E/(n-p-s-1)} \sim \mathscr{F}_{df1=s,df2=n-p-s-1}$

Examples: nested MLR model comparisons

Consider the medical expenditure data you are analyzing for Problem Set 2. Define Y = log(medical expenditures + 1) and let $X_1 = age - 65$ and $X_2 = male$ (indicator 1 = male, 0 = female). Define three models:

Model	Xs	residual df	SS(residual)
A	X_1, X_2	5691	31332.38
B	$X_1, (X_1 - 10)^+, (X_1 - 20)^+, X_2$	5689	31314.59
C	$[X_1, (X_1 - 10)^+, (X_1 - 20)^+] \times X_2$	5686	31299.23

Example 1: nested MLR model comparisons

After adjusting for gender, is the average log expenditure a linear function of age?

H0:

HA:

Model	Xs	residual df	SS(residual)	MS	F
A	X_1, X_2	5691	31332.38		
B	$X_1, (X_1 - 10)^+, (X_1 - 20)^+, X_2$	5689	31314.59	5.50	
Change		2	17.79	8.90	$\frac{8.90}{5.50} = 1.62$

Compute the P-value as: $Pr(\mathscr{F}_{2,5689} > 1.62) = 0.199$.

Example 1: nested MLR model comparisons

```
load("C:\\Users\\Elizabeth\\Dropbox\\Biostat6532020\\Problem Set 2\\nmes.rdata")
d = nmes %>% select(names(.)[c(1,2,3,15)]) %>% filter(.,lastage>=65)
d = mutate(d,
logy = log(totalexp+1),
agec=lastage-65,
agesp1 = ifelse(lastage-75>0, lastage-75,0),
agesp2 = ifelse(lastage-85>0, lastage-85,0)
)
reg0 = lm(logy~agec+male,data=d)
reg1 = lm(logy~agec+agesp1+agesp2+male,data=d)
reg2 = lm(logy~(agec+agesp1+agesp2)*male,data=d)
```

```
# Questoin 1: using anova function
anova(reg0,reg1)
```

```
## Analysis of Variance Table
##
## Model 1: logy ~ agec + male
## Model 2: logy ~ agec + agesp1 + agesp2 + male
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1   5691 31332
## 2   5689 31315  2     17.79 1.6159 0.1988
```

Example 2: nested MLR model comparisons

Is the non-linear relationship of average log expenditures on age the same for males and females? i.e. are the curves parallel?

- Equivalently: Is the difference between the average log expenditure for males and females the same at all ages?

H0:

HA:

Model	Xs	residual df	SS(residual)
A	X_1, X_2	5691	31332.38
B	$X_1, (X_1 - 10)^+, (X_1 - 20)^+, X_2$	5689	31314.59
C	$[X_1, (X_1 - 10)^+, (X_1 - 20)^+] \times X_2$	5686	31299.23

Example 2: nested MLR model comparisons

Model	Xs	residual df	SS(residual)
A	X_1, X_2	5691	31332.38
B	$X_1, (X_1-10)^+, (X_1-20)^+, X_2$	5689	31314.59
C	$[X_1, (X_1-10)^+, (X_1-20)^+] \times X_2$	5686	31299.23

```r
# Question 2:
anova(reg1,reg2)
```

```
## Analysis of Variance Table
##
## Model 1: logy ~ agec + agesp1 + agesp2 + male
## Model 2: logy ~ (agec + agesp1 + agesp2) * male
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1   5689 31315
## 2   5686 31299  3     15.36 0.9301 0.4252
```

Likelihood ratio tests for nested MLR models

Let $loglike_{ext}$ and $loglike_{null}$ be the values of the log likelihoods evaluated at the parameter estimates from the extended and null models, respectively.

Then to test H0:

Compute 2 x $loglike_{ext}$ -2 x $loglike_{null}$ ~ χ, df = s

Examples: nested MLR model comparisons using LRT

```r
# Question 1: by hand
lr.test.stat = as.numeric(2 * logLik(reg1) - 2 * logLik(reg0))
pchisq(lr.test.stat,df=2,lower.tail=FALSE)
```

```
## [1] 0.1985122
```

```r
# Question 1: Using lrtest function
#install.packages(lmtest)
library(lmtest)
```

```r
lrtest(reg0,reg1)
```

```
## Likelihood ratio test
##
## Model 1: logy ~ agec + male
## Model 2: logy ~ agec + agesp1 + agesp2 + male
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   4 -12934
## 2   6 -12933  2 3.2338     0.1985
```

Next time....

- Model checking for MLR models
- Key extensions for MLR models