Lecture8 Handout

Elizabeth Colantuoni

2/15/2021

I. Objectives

Upon completion of this session, you will be able to do the following:

1. Derive and apply distribution results for linear contrasts in the linear model
2. Discuss the approach to making inferences about non-linear functions of the linear regression parameters
3. Derive and apply variance decompositions (ANOVA) for the linear model; i.e. model F tests

II. Main points from Lectures 6 and 7

In this section, we review the main points from Lectures 6 and 7.

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

III. Inference about unknown parameters (β) and functions of them

We can make inferences about unknown parameters (β) and function of them using the information from the table below. NOTE: the t-statistic for computing the 95% confidence interval is given by $t_{n-(p+1),0.025}$. NOTE: the $\hat{se}$ values replace unknown parameters with estimates, i.e. replace σ^2 with $\hat{\sigma}^2$.

			Test statistic for
Target	Estimate $\sim$ Sampling Distn	95% CI for target	H0: Target = 0
β_j	$\hat{\beta}_j \sim N(\beta_j, [\sigma^2(X'X)^{-1})]_{jj})$	$\hat{\beta}_j \pm t \times \hat{se}(\hat{\beta}_j)$	$\frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)}$
$A\beta$	$A\hat{\beta} \sim N(A\beta, \sigma^2 A(X'X)^{-1}A')$	$A\hat{\beta} \pm t \times \hat{se}(A\hat{\beta})$	$\frac{A\hat{\beta}_j}{\hat{se}(A\hat{\beta}_j)}$
$g(\beta_j)$	$g(\hat{\beta}_j) \sim N(g(\beta_j), [g'(\beta_j)]^2[\sigma^2(X'X)^{-1}]_{jj})$	$g(\hat{\beta}_j) \pm t \times \hat{se}(g(\hat{\beta}_j))$	$\frac{g(\hat{\beta}_j)}{\hat{se}(g(\hat{\beta}_j))}$
$g(\beta)$	$g(\hat{\beta}) \sim N(g(\beta), g'(\beta)'[\sigma^2(X'X)^{-1}]g'(\beta))$	$g(\hat{\beta}) \pm t \times \hat{se}(g(\hat{\beta}))$	$\frac{g(\hat{\beta})}{\hat{se}(g(\hat{\beta}))}$
$\mu_i = E(Y_i \mid X_i)$	$\hat{Y}_i \sim N(\mu_i, \sigma^2[H]_{ii})$	$\hat{Y}_i \pm t \times \hat{se}(\hat{Y}_i)$	$\frac{\hat{Y}_i}{\hat{se}(\hat{Y}_i)}$
$\mu(x_0) = E(Y \mid x_0)$	$x_0'\hat{\beta} \sim N(x_0'\beta, \hat{\sigma}^2 x_0'(X'X)^{-1}x_0)$	$x_0'\hat{\beta} \pm t \times \hat{se}(x_0'\hat{\beta})$	$\frac{x_0'\hat{\beta}}{\hat{se}(x_0'\hat{\beta}}$

IV. Delta Method

The “delta method” or “linearization method” is used for obtaining the large-sample distribution of a function of a given statistic, where the distribution of the given statistic is known to be (or asymptotically to be) a normal distribution.

A. Univariate case

In the univariate case, the delta method provides the following:

Assuming that $\hat{\theta} \sim N(\theta, \sigma^2)$, then $g(\hat{\theta}) \sim N(g(\theta), [g'(\theta)]^2\sigma^2)$.

1. Example: univariate case

A random variable Y follows a log-normal distribution if $log(Y) \sim N(\mu, \sigma^2)$. From this log-normal distribution, you can show that the mean of Y is $E(Y) = e^{\mu + \frac{\sigma^2}{2}}$ and the median of Y is e^{μ}.

Suppose that Y = (medical expenditures+1) follows a log-normal distribution (see Problem Set 2) and that you construct the following model for persons 65 years or older:

$$log(Y_i) = \beta_0 + \beta_1(age_i - 65) + \beta_2(age_i - 75)^+ + \beta_3(age_i - 85)^+ + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

.

Note: this is NOT the model you are fitting in Problem Set 2, you are modeling the expenditures directly.

Suppose we want to estimate the median expenditures for 65 year-olds. From the model, we see that β_0 is the mean of the distribution of $log(Y_i)$ for 65 year-olds. So that the median expenditures for 65 year-olds is given by $e^{\beta_0} - 1$.

We know that $\hat{\beta}_0 \sim N(\beta_0, \sigma^2[(X'X)^{-1}]_{11})$, where I have assumed that β_0 is the first element of the $(p+1) \times 1$ vector β.

We will apply the univariate delta method to generate the distribution for $g(\beta_0) = e^{\beta_0} - 1$. To complete the process, we have to take the derivative of $g(\beta_0)$ with respect to β_0, which is e^{β_0}.

Then by the delta method, $g(\hat{\beta}_0) = e^{\hat{\beta}_0} - 1 \sim N(e^{\beta_0} - 1, (e^{\beta_0})^2[\sigma^2(X'X)^{-1}]_{11})$.

To create a 95% confidence interval for the median expenditures for 65 year-olds is given by:

$$(e^{\hat{\beta}_0} - 1) \pm t \times e^{\hat{\beta}_0}\sqrt{[\hat{\sigma}^2(X'X)^{-1}]_{11}}$$

```r
load("nmes.rdata")
d = nmes %>% select(names(.)[c(1,2,15)]) %>% 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)
)
## Fit the regression
reg = lm(logy~agec+agesp1+agesp2,data=d)
## Save the estimate of beta0 and var(hat-beta0)
beta0 = coef(reg)[1]
var.beta0 = vcov(reg)[1,1]
c(beta0,var.beta0)
```

```
## (Intercept)
## 6.265282252 0.004507285
```

```r
## Apply the delta method:
g.est = exp(beta0)-1
g.var = exp(beta0)^2 * var.beta0
c(g.est,g.var)
```

```
## (Intercept) (Intercept)
##      524.99     1247.01
```

```r
## Generate a 95% CI for exp(beta0)-1
g.est - qt(0.975,df=summary(reg)$df[2])*sqrt(g.var)
```

```
## (Intercept)
##     455.763
```

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

```
## (Intercept)
##     594.217
```

2. Derivation of 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,

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

$$Var(g(\hat{\theta})) = g'(\theta)^2 Var(\hat{\theta})$$

B. Multivariate case

You can extend the delta method to the multivariate case where the function g operates on a vector (e.g. β) instead of a univariate statistic (e.g. $\mu(age_i = 65)$ from prior example). The main difference in the application is that when you take the derivative of g it is taken with respect to the vector, i.e. the derivative is a vector not a scalar.

In applications of the multivariate delta method within regression analysis, you will be starting with your assumption that $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$. Then the sampling distribution for $g(\hat{\beta})$ is approximately:

$$N(g(\beta), g'(\beta)'[\sigma^2(X'X)^{-1}]g'(\beta))$$

1. Example multivariate case

Consider the example Lab 4, where we separately estimated the monthly growth rate of arm circumference for children under 6 months and for children 6 months old or older. Recall the regression model:

$$arm_i = \beta_0 + \beta_1 age_i + \beta_2(age_i - 6)^+ + \epsilon_i$$

where ϵ_i iid $N(0, \sigma^2)$.

The monthly growth rate of arm circumference is β_1 and $\beta_1 + beta_2$ for children under 6 months and for children 6 months old and older, respectively.

Now suppose you want to generate a 95% confidence interval for the relative growth rate comparing the growth rate for children 6 months old and older to that of children under 6 months, i.e.

$$g(\beta) = \frac{\beta_1 + \beta_2}{\beta_1} = 1 + \frac{\beta_2}{\beta_1}$$

Our goal is to derive the sampling distribution for $g(\hat{\beta})$. To apply the delta method, we need:

- $\hat{\beta} \sim N(\beta, \sigma^2(X^{\prime}X)^{-1})$
- $g^{\prime}(\beta)$ with respect to β_0, β_1 and β_2. Taking the derivative of $g(\beta)$ we get: 0, $\frac{-\beta_2}{\beta_1^2}$ and $\frac{1}{\beta_1}$, respectively.

```r
load("C:\\Users\\Elizabeth\\Dropbox\\Biostat6532020\\Lecture34\\NepalAnthro.rdata")
d = nepal.anthro %>% select(names(.)[1:16]) %>% filter(.,num==1)
d = mutate(d,
agesp6=ifelse(age-6>0, age-6,0)
)
cc=complete.cases(select(d,age,arm))
d.cc=filter(d,cc)
d.cc = arrange(d.cc,age)
reg<-lm(data=d.cc, arm~age+agesp6)
summary.lm(reg)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 11.1208946 0.50958653 21.823369 1.079068e-52
## age          0.3114066 0.09264467  3.361300 9.453805e-04
## agesp6      -0.2795773 0.09441211 -2.961244 3.472606e-03
```

```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
# Compute the 95% CI for g(beta)
reg.coeff[2]
```

```
##       age
## 0.3114066
```

```r
reg.coeff[2]+reg.coeff[3]
```

```
##        age
## 0.03182924
```

```r
g.est
```

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

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

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

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

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

V. Testing MLR models assuming normally distributed errors

Assume you have a MLR model. We have shown how to conduct statistical inference for a single regression parameter, i.e. how to test $H_0 : \beta_j = 0$. But what if our research question requires that we test a set of regression coefficients?

Consider two models that contain the following explanatory variables:

- Null model: $X_1, X_2, ..., X_p$
- Extended model: $X_1, X_2, ..., X_p, X_{p+1}, X_{p+2}, ..., X_{p+s}$

So that the extended model contains the explanatory variables from the null model plus a set of s additional explanatory variables. We say that these models are "nested models" since the null model is contained within the extended model.

The goal then is to test $H_0 : \beta_{p+j} = 0$ for all j = 1, 2,...,s vs. $H_A :$ at least one $\beta_{p+j} \neq 0$.

There are several approaches to conduct the test of H_0 vs. H_A.

A. F-test for nested models

Define $R_N = (I - H_N)Y$ and $R_E = (I - H_E)Y$ as the residuals from the least squares fit of the null and extended models, respectively.

Further, define $\Delta = R_N - R_E = (H_E - H_N)Y$ to be the difference in the residuals comparing the null and extended models.

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

1. Example F-test for nested models

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

NOTE: Model C contains the main terms for each X plus the interaction of X_2 with all the X_1 variables.

QUESTION 1: After adjusting for gender, is the average log medical expenditures roughly a linear function of age?

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: $Pr(\mathscr{F}_{2,5689} > 1.62) = 0.199$.

QUESTION 2: 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 expenditures for males and females the same at all ages?

You do:

```
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)
c(summary(reg0)$df[1:2],sum(residuals(reg0)^2))
```

```
## [1]     3.00  5691.00 31332.38
```

```
c(summary(reg1)$df[1:2],sum(residuals(reg1)^2))
```

```
## [1]     5.00  5689.00 31314.59
```

```
c(summary(reg2)$df[1:2],sum(residuals(reg2)^2))
```

```
## [1]     8.00  5686.00 31299.23
```

```
# Question 1: by hand
q1.f = ((sum(residuals(reg0)^2) - sum(residuals(reg1)^2))/2)/(sum(residuals(reg1)^2)/summary(reg1)$df[2
pf(q1.f,2,summary(reg1)$df[2],lower.tail=FALSE)
```

```
## [1] 0.1987943
```

```
# Question 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
```

```
# 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
```

C. Likelihood ratio test for nested models

Let $loglike_{null}$ and $loglike_{ext}$ be the values of the log likelihood evaluated at the maximum likelihood estimates for the null and extended models, respectively.

Then we can test the nested hypothesis given in Section 4 by computing:

$$2 \times loglike_{ext} - 2 \times loglike_{null} \sim \chi^2_{df=s}$$

1. Example likelihood ratio test for nested models

Answer the same questions from Section B using likelihood ratio tests.

```
# 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
```

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

```
## Warning: package 'lmtest' was built under R version 3.6.2
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
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
```

```
# Question 2:
lrtest(reg1,reg2)
```

```
## Likelihood ratio test
##
## Model 1: logy ~ agec + agesp1 + agesp2 + male
## Model 2: logy ~ (agec + agesp1 + agesp2) * male
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   6 -12933
## 2   9 -12931  3 2.7936     0.4246
```