MLR: Inference on Independent Variables

BIOS 6611

CU Anschutz

Week 10

1. Motivation
2. Overall F-test (All Variables)
3. Partial F-test (Group of Variables)
4. t-test (Single Variable)
5. Appendix: Function to Create ANOVA Table

Motivation

Framingham Heart Study Introduction

The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. We will focus on a subset of the baseline data that is made available from the NHLBI's BioLINCC (Biologic Specimen and Data Repository Information Coordinating Center).

We will focus on the data to evaluate the relationship of total cholesterol with BMI, systolic blood pressure, and diastolic blood pressure with a multiple linear regression model:

```
dat_all <- read.csv('frmgham2_baseline_subset.csv')

# remove cases with missing data
col_vec <- c('TOTCHOL','BMI','SYSBP','DIABP')
dat <- dat_all[complete.cases(dat_all[,col_vec]),]

mod_full <- glm(TOTCHOL ~ BMI + SYSBP + DIABP, data=dat)
```

Inference About Independent Variables

Considering our multiple linear regression model to predict cholesterol (Y) based on BMI (X_1), SBP (X_2), and DBP (X_3). Our expected regression model would be

$$E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

This leads to three potential questions we may wish to address relating to the independent variables (IVs):

1. **Overall test**: taken collectively, does the *entire* set of IVs contribute significantly to the prediction of Y?
2. **Test for a group of variables**: does the addition of some *group* of IVs of interest add significantly to the prediction of Y over and above that achieved by the other IVs already present in the model?
3. **Test for a single variable**: does the addition of *one* particular IV of interest add significantly to the prediction of Y over and above that achieved by other IVs already present in the model?

Full and Reduced Models

The hypothesis tests we will be discussing can be interpreted as a comparison between two models:

- the **full (or complete) model** which includes all predictors of interest
- the **reduced model** which is some subset of the full model

These tests represent the context with **nested models**, where we are comparing some nested subset in the reduced model to some full model. The complete model will simplify to the reduced model if the null hypothesis is true.

For notation, let

- n be the number of observations
- p be the number of IVs in the *reduced model*
- k the number of IVs *removed* from the full model
- $p + k$ the number of IVs in the *full model*

Overall F-test (All Variables)

Overall Test

Taken collectively, does the *entire* set of independent variables contribute significantly to the prediction of Y?

In the context of our multiple linear regression model this would be generally stated as:

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$

H_1 : at least one of the $\beta_k \neq 0$

In terms of our full and reduced model, we have

- Full Model: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$
- Reduced Model: $E(Y) = \beta_0$

We can test this using the **overall F-test**.

Overall F-test

We previously saw the F-test in simple linear regression and as part of the ANOVA table.

Recall, the F statistic is the ratio of the variability explained by our regression model (MS_{Model}) and the variability remaining after fitting our regression line (MS_{Error}):

$$F = \frac{MS_{Model}}{MS_{Error}} = \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2_{Y|X}} \sim F_{k,n-k-1}$$

If the null hypothesis is true, $\hat{\sigma}_0^2$ estimates $\sigma^2_{Y|X}$, otherwise it estimates some quantity larger than $\sigma^2_{Y|X}$.

$\hat{\sigma}^2_{Y|X}$ estimates $\sigma^2_{Y|X}$ whether or not the null hypothesis is true.

Overall F-test Example by Hand

In previous lectures we partitioned the variance of our regression model and noted how we could calculate all the terms that ultimately can be presented in an ANOVA table. These formulas allow us to calculate the overall F-test by hand.

We have also discussed how we can program our own ANOVA table for linear regression models (see the *Appendix* of this lecture for a refresher):

```
linreg_anova_func(mod_full)
```

Source	Sums of Squares	Degrees of Freedom	Mean Square	F-value	p-value
Model	383338	3	127779.32	66.98	$<$0.001
Error	8317099	4360	1907.59		
Total	8700437	4363			

Based on the ANOVA table, we have $F = 66.98$ and $p < 0.001$. Therefore we reject H_0 and conclude that at least one of our β's is not 0.

Overall F-test Example in R

We can also leverage R to directly compare a full and reduced model using the anova function:

```
mod_null <- glm( TOTCHOL ~ 1, data=dat) # fit reduced model

anova( mod_full, mod_null, test='F' ) # specify test='F' for F-test
```

```
## Analysis of Deviance Table
##
## Model 1: TOTCHOL ~ BMI + SYSBP + DIABP
## Model 2: TOTCHOL ~ 1
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1      4360    8317099
## 2      4363    8700437 -3  -383338 66.985 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Partial F-test (Group of Variables)

Test for a Group of Variables

Does the addition of some *group* of independent variables of interest add significantly to the prediction of Y over and above that achieved by the other independent variables already present in the model?

In the context of our multiple linear regression model this could be generally stated as:

$H_0 : \beta_1^* = \beta_2^* = ... = \beta_k^* = 0$

H_1 : at least one of the $\beta_k^* \neq 0$

In terms of our full and reduced model, we would then have

- Full Model:
  $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \beta_1^* X_1^* + ... + \beta_k^* X_k^*$
- Reduced Model: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$

We can test this using the **partial F-test**.

Partial F-test

The partial F-test evaluates if the error sum of squares for the full model is significantly reduced compared to the reduced model without the k independent variables. The F statistic for this test can be calculated as

$$F = \frac{[SS_{model}(full) - SS_{model}(reduced)]/k}{MS_{error}(full)} \sim F_{k,n-p-k-1}$$

The overall F-test is really a special case of the partial F-test, since the model with no predictors has $SS_{model}(reduced) = 0$ and $p = 0$.

NOTE: the result of our partial F-test does not necessarily mean *all* variables in the group are significant predictors. Perhaps a more parsimonious model is adequate.

Partial F-test Example by Hand

Let's test if the addition of the blood pressure variables are meaningful. Here we are specifically testing:

$H_0 : \beta_2 = \beta_3 = 0$

H_1 : at least one blood pressure beta coefficient is not 0

We can fit the reduced model in R and generate the ANOVA table for our hand calculations on the next page:

```
mod_red <- glm(TOTCHOL ~ BMI, data=dat)

linreg_anova_func(mod_red)
```

Partial F-test Example by Hand

Reduced model:

Source	Sums of Squares	Degrees of Freedom	Mean Square	F-value	p-value
Model	130939.6	1	130939.63	66.65	$<$0.001
Error	8569497.4	4362	1964.58		
Total	8700437.1	4363			

From the full model ANOVA earlier, SS_{Model}=383338 and MS_{Error}=1907.59. For this problem, our critical value is $F_{k,n-p-k-1} = F_{2,4364-1-2-1} = F_{2,4360} = 2.998$.

$$F = \frac{[SS_{model}(full) - SS_{model}(reduced)]/k}{MS_{error}(full)}$$

Partial F-test Example in R

We can also leverage the anova function from earlier for a partial F-test:

```
anova(mod_full, mod_red, test='F')
```

```
## Analysis of Deviance Table
##
## Model 1: TOTCHOL ~ BMI + SYSBP + DIABP
## Model 2: TOTCHOL ~ BMI
##   Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
## 1      4360    8317099
## 2      4362    8569497 -2  -252398 66.156 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

t-test (Single Variable)

Test for a Single Variable

Does the addition of *one* particular independent variable of interest add significantly to the prediction of Y over and above that achieved by other independent variables already present in the model?

In the context of our multiple linear regression model this could be generally stated as:

$H_0 : \beta_1^* = 0$

$H_1 : \beta_1^* \neq 0$

In terms of our full and reduced model, we would then have

- Full Model: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \beta_1^* X_1^*$
- Reduced Model: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$

We can test this using the **t-test**.

t-test Refresher

Recall, the t-statistic for our beta coefficients is

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim t_{n-(p+k)-1}$$

In simple linear regression we noted that the t-test and overall F-test had the connection that $t^2 = F$.

In multiple linear regression the t-test will give the same result as the partial F-test for the addition of a single variable (i.e., $k = 1$).

t-test Example from Regression Output

```
round(summary(mod_full)$coefficients,4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 171.3603     5.3325 32.1353   0.0000
## BMI           0.6913     0.1746  3.9600   0.0001
## SYSBP         0.3609     0.0478  7.5549   0.0000
## DIABP        -0.0015     0.0905 -0.0171   0.9864
```

Does the addition of *BMI* add significantly to the prediction of total cholesterol over and above that achieved by SBP and DBP?

Does the addition of *SBP* add significantly to the prediction of total cholesterol over and above that achieved by BMI and DBP?

Does the addition of *DBP* add significantly to the prediction of total cholesterol over and above that achieved by BMI and SBP?

Appendix: Function to Create ANOVA Table

linreg_anova_func R Function Code

```r
linreg_anova_func <- function(mod, ndigits=2, p_ndigits=3, format='kable'){
### Function to create an ANOVA table linear regression results from lm or glm
# mod: an object with the fitted model results
# ndigits: number of digits to round to for most values, default is 2
# p_digits: number of digits to round the p-value to, default is 3
# format: desired format output (default is kable): "kable" for kable table, "df" for data frame as table

  # extract outcome from the object produced by the glm or lm function
  if( class(mod)[1] == 'glm' ){ y <- mod$y }
  if( class(mod)[1] == 'lm' ){ y <- mod$model[,1] }

  ybar <- mean(y);  yhat <- predict(mod);  p <- length(mod$coefficients)-1;  n <- length(y)
  ssm <- sum( (yhat-ybar)^2 );  sse <- sum( (y-yhat)^2 );  sst <- sum( (y-ybar)^2 )
  msm <- ssm/p;  mse <- sse/(n-p-1)
  f_val <- msm/mse;  p_val <- pf(f_val, df1=p, df2=n-p-1, lower.tail=FALSE)

  # Create an ANOVA table to summarize all our results:
  p_digits <- (10^(-p_ndigits))
  p_val_tab <- if(p_val<p_digits){paste0('<',p_digits)}else{round(p_val,p_ndigits)}
  anova_table <- data.frame( 'Source' = c('Model','Error','Total'),
                             'Sums of Squares' = c(round(ssm,ndigits), round(sse,ndigits), round(sst,ndigits)),
                             'Degrees of Freedom' = c(p, n-p-1, n-1),
                             'Mean Square' = c(round(msm,ndigits), round(mse,ndigits),''),
                             'F Value' = c(round(f_val,ndigits),'',''),
                             'p-value' = c(p_val_tab,'',''))

  if( format == 'kable' ){
    library(kableExtra)
    kbl(anova_table, align='lccccc', escape=F,
        col.names=c('Source','Sums of Squares','Degrees of Freedom','Mean Square','F-value','p-value')) %>%
      kable_styling(bootstrap_options = "striped", full_width = F, position = "left")
  }else{ anova_table }
}
```