STAC67: Regression Analysis

Lecture 14

Sohee Kang

Mar. 4, 2021

Example: BodyFat Data

- The data consists of 20 females whose age are between 25 and 30 years old.
- Variables in the data set are:
  - y = amount of body fat (percentage) x_1 = triceps skinfold thickness,
  - x_2 = thigh circumference x_3 = midarm circumference

```r
Data = read.table("bodyfat.txt")
names(Data) = c("X1", "X2", "X3", "Y")
Data[1:3, ]
```

```
##     X1   X2   X3    Y
## 1 19.5 43.1 29.1 11.9
## 2 24.7 49.8 28.2 22.8
## 3 30.7 51.9 37.0 18.7
```

```r
fit = lm(Y~X1 + X2 +X3, data=Data)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7263 -1.6111  0.3923  1.4656  4.1277
##
## Coefficients:
```

BodyFat Example

- Model 1: regression of Y on X1:

```r
anova(lm(Y~X1, data=Data))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27  44.305 3.024e-06 ***
## Residuals 18 143.12    7.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Model 2: regression of Y on X2:

```r
anova(lm(Y~X2, data=Data))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## X2         1 381.97  381.97  60.617 3.6e-07 ***
## Residuals 18 113.42    6.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

BodyFat Example

- Model 3: regression of Y on X1 and X2:

```r
anova(lm(Y~X1+X2, data=Data))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27 54.4661 1.075e-06 ***
## X2         1  33.17   33.17  5.1284    0.0369 *
## Residuals 17 109.95    6.47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Model 4: regression of Y on X1, X2, X3:

```r
anova(lm(Y~X1+X2+X3, data=Data))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27 57.2768 1.131e-06 ***
## X2         1  33.17   33.17  5.3931   0.03373 *
## X3         1  11.55   11.55  1.8773   0.18956
## Residuals 16  98.40    6.15
## ---
```

BodyFat Example

- Test for regression coefficients

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

$$H_0 : \beta_3 = 0 \quad vs \quad H_a : \beta_3 \neq 0$$

- Full Model:
- Reduced Model:

$$F =$$

BodyFat Example

- Test for regression coefficients

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

$$H_0 : \beta_2 = \beta_3 = 0 \quad vs \quad H_a :$$

- Full Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$
- Reduced Model: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

$$F = \frac{\frac{SSE(X_1) - SSE(X_1, X_2, X_3)}{(4-2)}}{MSE_F} = \frac{\frac{143.12 - 98.40}{2}}{6.15}$$

$$= \frac{\frac{SSR(X_1, X_2, X_3) - SSR(X_1)}{2}}{MSE} = \frac{\frac{SSR(X_3 | X_2, X_1) + SSR(X_2 | X_1)}{2}}{MSE}$$

$$Fval = qf(0.01, 2, 16)$$

7.6 Multicollinearity and its Effects

1. Uncorrelated Predictor Variables

- Example: two predictor variables are perfectly uncorrelated.

Case	X_1 (Crew size)	X_2 (Bonus pay)	Y (Crew Productivity)
1	4	2	42
2	4	2	39
3	4	3	48
4	4	3	51
5	6	2	49
6	6	2	53
7	6	3	61
8	6	3	60

Models	$\hat{\beta}_1$	$\hat{\beta}_2$
$Y = \beta_0 + \beta_1 + \varepsilon$	5.375	
$Y = \beta_0 + \beta_2 X_2 + \varepsilon$		9.250
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$	5.375	9.250

- Extra Sum of Squares

$SSR(X_1 \mid X_2)$	$SSR(X_1)$	$SSR(X_2 \mid X_1)$	$SSR(X_2)$
231.125	231.125	171.125	171.125

BodyFat Example Revisited

```r
attach(Data)
cor(cbind(X1, X2, X3))
```

```
##           X1        X2        X3
## X1 1.0000000 0.9238425 0.4577772
## X2 0.9238425 1.0000000 0.0846675
## X3 0.4577772 0.0846675 1.0000000
```

- Effects on Regression Coefficients

Models	$\hat{\beta}_1$	$\hat{\beta}_2$
$Y = B_0 + B_1 X_1 + \varepsilon$	0.8572	
$Y = B_0 + B_2 X_2 + \varepsilon$		0.8565
$Y = B_0 + B_1 X_1 + B_2 X_2 + \varepsilon$	0.2224	0.6594
$Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + \varepsilon$	4.334	-2.857

BodyFat Example Revisited

- Inflated variability of estimators

Models	$SE(\hat{\beta}_1)$	$SE(\hat{\beta}_2)$
$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	0.1288	
$Y = \beta_0 + \beta_2 X_2 + \varepsilon$		0.1100
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$	0.3034	0.2912
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$	3.016	2.582

3. Effect of Multicollinearity

$(X'X)^{-1}$ must exist $\begin{cases} ① \ n \gg p \\ ② \text{ extreme multicollinearity DNE} \end{cases}$

- When the multicollinearity is not strong, i.e., $(X'X)^{-1}$ esitst, we can still use the model to make prediction.
- However, the multicollinearity will result in instability of estimated coefficients, i.e,
- The interpretation of the coefficients is difficult. (Remedy: we can use ridge regression)

Example: Cobb-Douglas Production Function (General Linear Hypothesis Testing)

- Cobb and Douglas (1928) proposed a multiplicative production function: Quantitiy Produced (Y), and the independent variables are: Capital (X_1) and Labor(X_2). Data is from US production data from 1899-1922
- All variables were transformed to log:

$$Y^* = log(Y), X_1^* = log(X_1), \text{and} \;\; X_2^* = log(X_2)$$

```r
cobb <-read.table("cobbdoug1.dat",header=F,
      col.names=c("year","Q.index","K.indx","L.indx"))
attach(cobb)

log.Y <- log(Q.index)
log.K <- log(K.indx)
log.L <- log(L.indx)

head(cobb)
```

```
##   year Q.index K.indx L.indx
## 1 1899     100    100    100
## 2 1900     101    107    105
## 3 1901     112    114    110
```

Example

$$H_0 : \beta_1 + \beta_2 = 1$$

Recall: $F^* = \dfrac{(K'\hat{\beta} - m)' [K'(X'X)^{-1}K] (K'\hat{\beta} - m)/K}{MSE}$ $K=1$

$K' = [\,0 \quad 1 \quad 1\,]$ $\qquad m = 1$

$(X'X)^{-1} =$ $\qquad \underset{\sim}{\hat{\beta}} =$

```r
mod1 <- lm(log.Y ~ log.K + log.L)
#summary(mod1)
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: log.Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## log.K      1 1.49156 1.49156 441.280 1.402e-15 ***
## log.L      1 0.10466 0.10466  30.964 1.601e-05 ***
## Residuals 21 0.07098 0.00338
## ---
```

Example (Continued)

$$(K'(X'X)^{-1}K)^{-1} =$$

$$K'\underset{\sim}{\hat{\beta}} - m =$$

$$F =$$

```
#### Matrix Form
n <- length(log.Y)
Y <- log.Y
X <- cbind(rep(1, n),log.K,log.L)
p.prime = dim(X)[2]

Kp <- matrix(c(0,1,1),ncol=3)
m <- 1

beta.hat <- solve(t(X) %*% X) %*% t(X) %*% Y
Yhat <- X %*% beta.hat
e <- Y - Yhat
SSE <- sum(e^2)

solve(Kp %*% solve(t(X)%*%X) %*% t(Kp))
```

```
##           [,1]
## [1,] 0.4064116
```

Example (R codes)

```r
beta.hat
```

```
##               [,1]
##         -0.1773097
## log.K    0.2330535
## log.L    0.8072782
```

HW: $H_0: \beta_1 + \beta_2 = 1$

$$t = \frac{(\hat{\beta}_1 + \hat{\beta}_1) - 1}{SE(\hat{\beta}_1 + \hat{\beta}_2)} \sim t(n-p')$$

from var covar

```r
vcov(mod1)
```

```
##             (Intercept)        log.K        log.L
## (Intercept)  0.18861045  0.019984179 -0.059546854
## log.K        0.01998418  0.004036028 -0.008383119
## log.L       -0.05954685 -0.008383119  0.021047093
```

```r
Q <- t(Kp %*% beta.hat - m) %*% solve(Kp %*% solve(t(X)%*%X) %*% t(Kp)) %*% (Kp %*% beta.hat - m)

F.star = as.numeric(Q/(SSE/(n-p.prime)))
F.star
```

```
## [1] 0.1955836
```

```r
pf(F.star, 1, n-p.prime, lower.tail=F)
```

```
## [1] 0.6628307
```

Ch. 8 Regression Models for Quantitative and Qualitative Predictors

Qualitative Variables as Predictors

- We often wish to use categorical (or qualitative) variables as covariates in a regression model (e.g., gender, marital status, political afflication,. . .)
- For such **binary variable** (dummy variable), it is easy to incluce them in the model.

A single Binary Predictor

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad X_i = \begin{cases} 1 \\ 0 \end{cases}$$

- A response variable, Y, a single binary variable X (coded as 0 or 1)
- The least square estimates are:

$$\hat{\beta}_0 = \bar{Y}_0 \qquad \hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$$

Regression with a Single Binary predictor

When $x=0$ $\quad \hat{Y} = \hat{\beta}_0 = \bar{Y}_0$

when $x=1$ $\quad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 = \bar{Y}_0 + (\bar{Y}_1 - \bar{Y}_0) = \bar{Y}_1$

Furthermore, we can write residuals as

$$e_i = \begin{cases} Y_i - \bar{Y}_0 & \text{if } x_i = 0 \\ Y_i - \bar{Y}_1 & \text{if } x_i = 1 \end{cases}$$

Then MSE, $\hat{\sigma}^2$, in a two class situation becomes a "pooled" estimator of the variance (S_p^2). Similar to BF-test.

Pooled t-test (Under assumption $\sigma_1^2 = \sigma_2^2$)

$H_0: \mu_1 = \mu_2$

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t\left(\underbrace{(n_1-1)+(n_2-1)}_{n-2}\right)$$

Regression with a Single Binary predictor

So $H_0: \mu_1 = \mu_2 \iff H_0: \beta_1 = 0$

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\sigma^2}{S_{xx}}}}$$

Given $\frac{1}{S_{xx}} = \frac{1}{n_1} + \frac{1}{n_2}$

$$MSE = \sigma^2 = S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

$$\frac{1}{S_{xx}} = \frac{1}{\sum (x-\bar{x})^2} = \frac{1}{\sum x_i^2 - n\bar{x}^2}$$

$x_i = \begin{cases} 1 \\ 0 \end{cases}$

$$= \frac{1}{n_1 - \frac{n_1^2}{n}} = \frac{n}{n_1(n-n_1)}$$

$$= \frac{n}{n_1 n_2} = \frac{1}{n_1} + \frac{1}{n_2}$$

Single Factor with k >2 levels

- Many categorical variables have more than 2 levels
- We need to create **dummy variables**
- dummy variable is a binary variable coded as 0's and 1's
- A dummy variable for level j of a categorical variable is defined as:

$$D_i = \begin{cases} 1 & \text{if } x_i = j \\ 0 & \text{if } x_i \neq j \end{cases}$$

- Note that we only need (k-1) dummy variables with k levels
- The level for when we do not produce a dummy variable is called base category or control group
- It is the level to which all other levels are compared
- It does not matter which level we designate as the base category, effectively all other categories will be compared to this base category.

Single Factor with k>2 levels

- Example) X = student's major taking STAC67

1. Econ 2. Math 3. CS

X	D_1	D_2
Econ	0	1
Math	1	0
CS	0	0

— base category

- Suppose we select $X_i = k$ as the base category, then it is easy to show that the least squre estimates are:

$$\hat{B}_0 = \bar{Y}_k \qquad \hat{B}_j = \bar{Y}_j - \bar{Y}_k$$

- We are interested in testing if the mean of the response is the same for each group: $H_0: B_1 = B_2 = \dots = B_{k-1} = 0 \iff H_0: \mu_1 = \mu_2 = \dots = \mu_k$

Can be done using F-test

- This techinique is usually called One-Way Anova

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

One Continuous and one Categorical Covariates

- X_1: Continuous covariate $\quad$ X_2: binary covariate

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$= \begin{cases} \beta_0 + \beta_2 X_2 + \varepsilon & \text{if } X_1 = 0 \\ (\beta_0 + \beta_1) + \beta_2 X_2 + \varepsilon & \text{if } X_1 = 1 \end{cases}$$

- So the linear model can be thought of as two linear models with

  diffient intercepts, but the same slope of the quantitative variable.

- **Interpretation**

  $\hat{\beta}_1$: Same as before

  $\hat{\beta}_2$: The change in intercept of the line when comparing $X_2 = 0$ or $X_2 = 1$ with everything else constant.

  $H_0 : \beta_2 = 0$

  whether change in intercept is the same.

Example

Y: Speed of innovation, X_1: size of a insurance firm, X_2: type of firm

0 : mutual

1 : stock

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

$$E(Y) = \begin{cases} B_0 + B_1 X_1 \\ (B_0 + B_2) + B_1 X_1 \end{cases}$$

Initial data:

	Y	X_1	X_2
1	17	151	Mutual
2	26	92	Mutual
3	21	175	Mutual
⋮	⋮	⋮	⋮
10	16	238	Mutual
11	28	164	Stock
12	15	272	Stock
13	11	295	Stock
14	38	68	Stock
⋮	⋮	⋮	⋮
20	14	246	Stock

Recoded data:

	Y	X_1	X_2
1	17	151	0
2	26	92	0
3	21	175	0
⋮	⋮	⋮	⋮
10	16	238	0
11	28	164	1
12	15	272	1
13	11	295	1
14	38	68	1
⋮	⋮	⋮	⋮
20	14	246	1

R codes

```r
Innovation = read.table("Table8-2.txt", header=F,  col.names=c("Y","X1","X2"))
fit = lm(Y~X1 + X2, data=Innovation)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = Innovation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6915 -1.7036 -0.4385  1.9210  6.3406
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.874069   1.813858  18.675 9.15e-13 ***
## X1          -0.101742   0.008891 -11.443 2.07e-09 ***
## X2           8.055469   1.459106   5.521 3.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.221 on 17 degrees of freedom
## Multiple R-squared:  0.8951, Adjusted R-squared:  0.8827
## F-statistic:  72.5 on 2 and 17 DF,  p-value: 4.765e-09
```

R codes

```
library(ggplot2)
ggplot(data=Innovation, aes(x=X1, y=Y, color=X2, shape=factor(X2))) + geom_point() + geom_smooth(method='lm', f
```

```
## `geom_smooth()` using formula 'y ~ x'
```