STAC67: Regression Analysis

Lecture 12

Review

$$SST = Y'Y - \frac{1}{n} Y'JY$$
$$= Y'(I - \frac{1}{n}J)Y$$

Sohee Kang

$$SSR = \hat{B}'X'Y - \frac{1}{n} Y'JY$$
$$= Y'(H - \frac{1}{n}J)Y$$

Feb. 25, 2021

$$SSE = Y'(I - H)Y$$

$$E(e'e) = (n-(p+1))\sigma^2$$

$$MSE = E\left(\frac{e'e}{n-(p+1)}\right) = \sigma^2$$

Outline of the Lecture

- Introduce the adjusted R^2
- Example and Inference in Multiple Regression
- General linear hypothesis testing
- Testing a Subset of Coefficients

Unbiased estimator of σ^2

Exercise: Show that $E(\underset{\sim}{e}'\underset{\sim}{e}) = (n - p')\sigma^2$.

Hint: use that $tr(\boldsymbol{P}) = p'$ (without proof) and the quadratic formulation of $\underset{\sim}{e}'\underset{\sim}{e}$.

Exercise:Show that the estimator

$$\text{MSE} = s^2 = \frac{\underset{\sim}{e}'\underset{\sim}{e}}{n - p'}$$

is an unbiased estimator of σ^2.

Coefficient of multiple determination

$$R^2 = \frac{SSR}{SST}$$

- Fraction of the variation in Y explained by the model (i.e. its linear relationship with $X_1, \ldots, X_p$)
- We have $0 \le R^2 \le 1$

Exercise Show that

1. $R^2 = 0$ when $\widehat{\beta}_k = 0$ for each $k = 1, \ldots, p$.
2. $R^2 = 1$ when $\widehat{Y}_i = Y_i$ for each $i = 1, \ldots, n$, i.e. when all the observations fall on the fitted regression surface.

① If $\hat{\beta}_k = 0$ $\quad \hat{Y}_i = \beta_0 = \bar{Y}$ $\quad$ Then $SSR = \sum(\hat{Y}_i - \bar{Y})^2 = \sum(\bar{Y}-\bar{Y})^2 = 0$ $\quad R^2 = \frac{0}{SST} = 0$

② If $\hat{Y}_i = Y_i$ $\quad$ Then $SSR = \sum(Y_i - \bar{Y})^2 = \sum(Y_i - \bar{Y})^2 = SST$

$R^2 = \frac{SST}{SST} = 1$

Adjusted coefficient of multiple determination

- Adding more X to the model , increases R^2.
- Adjusted R^2: modified measure that accounts for the number of variables in the model.

- Adjusted coefficient of multiple determination:

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p'}}{\frac{SST}{n-1}} = 1 - \left(\frac{n-1}{n-\underbrace{p'}_{p+1}}\right)\frac{SSE}{SST}$$

  - R^2_{adj} does not have the same interpretation as R^2.
  - R^2_{adj} may decrease when we add a new variable because the decrease in SSE is greater than compensation of degrees of freedom
  - R^2_{adj} useful for selecting explanatory variables.

Exercise: mpg example

- Construct the ANOVA table, compute R^2 and R^2_{adj}.
- Verify your results with R.

Hint: Use $\underset{\sim}{Y}'\underset{\sim}{Y} = \sum_{i=1}^{n} Y_i^2 = 237665.9$

$$SST = Y'(I - \tfrac{1}{n}J)Y$$

$$SSR = Y'(H - \tfrac{1}{n}J)Y$$

$$SSE = Y'(I-H)Y$$

$$R^2 = \frac{SSR}{SST} \qquad R^2_{adj} = 1 - \frac{n-1}{n-p'}\frac{SSE}{SST}$$

mpg example: R output

```
autompg = read.csv("autompg.csv", header=T)
library(dplyr)
autompg = autompg %>% select(mpg, wt, year)
fit = lm(mpg ~wt + year, data=autompg)
sum.Y2 = t(autompg$mpg)%*%autompg$mpg
sum.Y2
```

```
##          [,1]
## [1,] 237665.9
```

```
n = dim(autompg)[1]
X = cbind(rep(1, n), autompg$wt, autompg$year)
J = matrix(1, ncol=n, nrow=n)
Y= autompg$mpg

SST = t(Y)%*%Y - 1/n*t(Y)%*%J%*%Y
H = X%*% solve(t(X)%*%X)%*%t(X)
I = diag(rep(1, n))
SSR = t(Y)%*%(H - 1/n*J)%*%Y
SSE = t(Y)%*%(I - H)%*%Y

c(SST, SSR, SSE)
```

```
## [1] 23761.672 19205.026  4556.646
```

```
anova(fit)
```

```
## Analysis of Variance Table
##
```

Multiple Regression Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i,$$

- ϵ_i are i.i.d normally distributed mean 0 and common variance, σ^2.

Inference about a single regression parameter

Simple linear regression:

We had

$$\frac{\sigma^2}{S_{xx}}$$

$$\widehat{\beta}_1 \sim N(\beta_1;\ Var(\beta_1))$$

and proved $\frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \sim t(n-2)$

where $s(\widehat{\beta}_1)$ is the standard error of $\widehat{\beta}_1$

$s(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$ $\quad \hat{\sigma}^2 = \frac{SSE}{n-2}$

Multiple linear regression:

We have

$(X'X)^{-1}\sigma^2$

$$\widehat{\beta}_j \sim N(\beta_j;\ Var(\beta_j))$$

and we could prove

$$\frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t(n-p')$$

where $s(\widehat{\beta}_j)$ is the standard error of $\widehat{\beta}_j$

$s(\hat{\beta}_j) = \hat{\sigma}\sqrt{(X'X)^{-1}_{(j+1,j+1)}}$

$\hat{\sigma}^2 = \frac{e'e}{n-p'}$

Inference about the mean response, a prediction of a new observation

similarly, we could show that

Given X_0 $\quad \hat{Y}_0 = X_0 B$ $\quad$ $\dfrac{\hat{Y}_0 - E(\hat{Y}_0)}{s(\hat{Y}_0)} \sim t(n-p')$

$s(\hat{Y}_0) = \hat{\sigma}\sqrt{(X_0(X'X)^{-1}X_0')}$

and $\quad \dfrac{\hat{Y}_{pred} - Y_0}{s(\hat{Y}_{pred})} \sim t(n-p')$

$s(\hat{Y}_{pred}) = \hat{\sigma}\sqrt{1 + X_0(X'X)^{-1}X_0'}$

Note: Hypothesis tests and confidence intervals for a single regression parameter, a mean response, and a prediction of a new observation are constructed exactly as for simple linear regression.

mpg Example

Variance Covariance Matrix = vcov(fit)

- Compute a 95 % confidence interval for β_1.

$$\hat{\beta}_1 \pm t_{0.975}(390-3)\, s(\hat{\beta}_1)$$

- Compute a point estimate and a 95% confidence interval for the expected mpg in automobiles with the wegiht of 2811 and the model year of 1976.

$$\hat{Y}_0 \pm t_{0.975}(387)\, s(\hat{Y}_0) \qquad X_0 = [1 \quad 2811 \quad 76]$$

- Compute a point estimate and a 95% prediction interval for the mpg in a new car with the wegiht of 2811 and the model year of 1976.

$$\hat{Y}_0 \pm t_{0.95}(387)\, s(\hat{Y}_{pred}) \qquad X_0 = [1 \quad 2811 \quad 76]$$

R codes

```r
confint(fit)
```

```
##                    2.5 %       97.5 %
## (Intercept) -22.548083086 -6.727200803
## wt           -0.007057296 -0.006212456
## year          0.663633861  0.859170049
```

```r
new.data =  data.frame(wt=2811, year=76)
predict(fit, new.data, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 24.57827 24.22949 24.92705
```

```r
predict(fit, new.data, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 24.57827 17.82281 31.33373
```

Distribution of Quadratic Forms

- We will use the following result (without proof):
- If $\underset{\sim}{Z} \sim N(\underset{\sim}{\mu}, \boldsymbol{V}\sigma^2)$ for a nonsingular matrix $\boldsymbol{V}$, then

  1. A quadratic form $\underset{\sim}{Z}'(\boldsymbol{A}/\sigma^2)\underset{\sim}{Z}$ is distributed as a noncentral chi-square distribution with
     1. $df = r(A)$ degrees of freedom, where $r(\cdot)$ is the rank
     2. noncentral parameter $\Omega = (\underset{\sim}{\mu}'\boldsymbol{A}\underset{\sim}{\mu})/2\sigma^2$

     if $\boldsymbol{AV}$ is idempotent.
     If $\Omega = \boldsymbol{0}$, then $\underset{\sim}{Z}'(\boldsymbol{A}/\sigma^2)\underset{\sim}{Z}$ is distributed as a $\chi^2_{r(A)}$.
  2. $\underset{\sim}{Z}'\boldsymbol{A}\underset{\sim}{Z}$ and $\underset{\sim}{Z}'\boldsymbol{B}\underset{\sim}{Z}$ are independent if $\boldsymbol{AVB} = \boldsymbol{0}$.
  3. $\underset{\sim}{Z}'\boldsymbol{A}\underset{\sim}{Z}$ and $\boldsymbol{B}\underset{\sim}{Z}$ are independent if $\boldsymbol{BVA} = \boldsymbol{0}$.

Exericse

- We can show that $\frac{1}{n}\boldsymbol{J}$, $\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}$, and $\boldsymbol{I} - \boldsymbol{H}$ are idempotent and pairewise orthogonal (i.e. the product of each pair gives $\mathbf{0}$).

Distribution of SSR/σ^2

$\frac{SSR}{\sigma^2} = Y'\frac{(H - \frac{1}{n}J)}{\sigma^2}Y$ rank $(H - \frac{1}{n}J) = p'-1$

- $\frac{SSR}{\sigma^2}$ is distributed as a noncentral chi-squre with $p' - 1$ degrees of freedom

Distribution of SSE/σ^2

$\frac{SSE}{\sigma^2} = Y'\frac{(I-H)}{\sigma^2}Y$ rank $(I-H) = n-p'$

- $\frac{SSE}{\sigma^2}$ is distributed as a $\chi^2_{n-p'}$

Independence of SSR/σ^2 and SSE/σ^2

- $\frac{SSR}{\sigma^2}$ and $\frac{SSE}{\sigma^2}$ are independent

1) $\frac{1}{n} J \cdot \frac{1}{n} J = \frac{1}{n^2} \cdot J \cdot J$

$= \frac{1}{n^2} \cdot n J = \frac{1}{n} J$

2) $(H - \frac{1}{n} J)'(H - \frac{1}{n} J) = H'H - H'\frac{1}{n} J - \frac{1}{n} J' H + \frac{1}{n^2} J' J$

$= H - \frac{1}{n} J \cancel{- \frac{1}{n} J + \frac{1}{n} J}$

3) $(I - H)'(I - H) = I - H \cancel{- H + H}$

$Y \sim N(X\beta, I\sigma^2)$, I is non-singular

Testing that all the coefficients except β_0 are null

- Suppose the null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$
- Under H_0,SSR/σ^2 is distributed as a central chisquare distribution $\sim \chi^2_{p'-1}$
- Therefore, under H_0

$$\frac{MSR}{MSE} = \frac{SSR/(p'-1)}{SSE/(n-p')} \sim F(p'-1, n-p)$$

- We just constructed the F-test.

F-test

$F =$

- Two sided-test

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \dots \beta_p = 0 \\ H_a : \text{at least one of the } \beta_k, k = 1, \dots, p \text{ is not } 0 \end{cases}$$

- Test statistic

$$F^* = \frac{MSR}{MSE}$$

- **Decision rule**:
  - Reject H_0 if $F^* > F_{1-\alpha;p'-1,n-p'}$
  - Do not reject H_0 if $F^* \leq F_{1-\alpha;p'-1,n-p'}$

where $F_{1-\alpha;p'-1,n-p'}$ is the $1-\alpha$-percentile of a $F(p'-1, n-p')$ distribution

never divide α by 2 for F

- $p(F > F^*) < \alpha$, reject H_0

mpg Example

- Use a hypothesis test (significance level $\alpha = 5\%$) to test whether there is a linear relation between the response variable and the explanatory variables in the mpg example.
- Verify your results with R.

$H_0 : B_1 = B_2 = 0$

H_1 : At least one not equal to 0

$$F^* = \frac{MSR}{MSE} = 815.55$$

Since $F^* > F_{0.95}$, reject H_0.

qf(0.95, 3-1, 390-3)

P val = 1-pf(815.55, 2, 387)

General linear hypothesis testing

$$\begin{cases} H_0 : \boldsymbol{K}'\underset{\sim}{\beta} = \underset{\sim}{m} \\ H_a : \boldsymbol{K}'\underset{\sim}{\beta} \neq \underset{\sim}{m} \end{cases}$$

for a $p' \times k$ nonsingular matrix $\boldsymbol{K}$ and a $k \times 1$ vector $\underset{\sim}{m}$.

Example: In the mpg example, we may be interesting in testing

$$H_0 : \beta_0 = 0 \quad \text{and} \quad 2\beta_1 + \beta_2 = 15$$

This is equavalent to testing

$H_0 : K'\beta = m$ with $K = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 15 \end{bmatrix}$

$$\begin{bmatrix} 1 & 0 & 0 \\ 70 & 1 & 0 \\ 10 & 0 & 1 \\ 10 & 0 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

General linear hypothesis testing

- Two sided-test

$$\begin{cases} H_0 : \boldsymbol{K}'\underset{\sim}{\beta} = \underset{\sim}{m} \\ H_a : \boldsymbol{K}'\underset{\sim}{\beta} \neq \underset{\sim}{m} \end{cases}$$

for a $p' \times k$ nonsingular matrix $\boldsymbol{K}$ and a $k \times 1$ vector $\underset{\sim}{m}$.

- Test statistic

$$F^* = \frac{(\boldsymbol{K}'\underset{\sim}{\widehat{\beta}} - \underset{\sim}{m})' \left[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}\right]^{-1} (\boldsymbol{K}'\underset{\sim}{\widehat{\beta}} - \underset{\sim}{m})/k}{s^2}$$

- **Decision rule**:
  - Reject H_0 if $F^* > F_{1-\alpha;k,n-p'}$
  - Do not reject H_0 if $F^* \leq F_{1-\alpha;k,n-p'}$

where $F_{1-\alpha;k,n-p'}$ is the $1-\alpha$-percentile of a $F(k, n-p')$ distribution

Exercise (Modification from Mahinda's Final Exam)

The following information (i.e. $(X'X)^{-1}$, $\underset{\sim}{\widehat{\beta}}$, error sum of squares (SSE)) were obtained from a study of the relationship between plant dry weight (Y), measured in grams and two independent variables, percent soil organic matter (X_1) and kilograms of supple-mental nitrogen per $1000m^2(X_2)$ based on a sample of n = 7 experimental fields . The regression model included an intercept.

$$(X'X)^{-1} = \begin{pmatrix} 1.7995972 & -0.0685472 & -0.2531648 \\ -0.0685472 & 0.0100774 & -0.0010661 \\ -0.2531648 & -0.0010661 & 0.0570789 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 51.5697 \\ 1.4974 \\ 6.7233 \end{pmatrix}$$

$= \hat{\beta}$

SSE = 27.5808

Test the null hypothesis of $H_0 : \beta_2 = 0.5\beta_1$ vs $H_a : \beta_2 \neq 0.5\beta_1$

$\hat{\sigma}^2 = \frac{SSE}{n-p'} = \frac{27.5808}{7-3}$

$0.5\beta_1 - \beta_2 = 0 \quad K = [0 \; 0.5 \; -1] \quad m = 0$

$F^* = \frac{(K'\hat{\beta})'(K'(X'X)^{-1}K)^{-1}(K'\hat{\beta}) / 1}{\hat{\sigma}^2}$

Chapter 7: Multiple Regresssion II

Testing a Subset of Coefficients

- We may want to test if some but not all the coefficents are 0.
- We define full model to be:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_p X_p + \epsilon_i$$

- Suppose the null hypothesis we want to test is:

$$\begin{aligned} &H_0 : \beta_{k+1} = \beta_{k+2} = \ldots = \beta_p = 0 \\ &H_a : \end{aligned}$$

- Then we can define the reduced model to be:

- From the full model, we get SSE (SSE_F) and MSE (MSE_F) with the degrees of freedom:
- From the reduced model, we get (SSE_R) with degrees of freedom:

Testing a subset of coefficients

$$F =$$

Further Decomposition of Sum of Squares

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_p X_p + \epsilon_i$$

- Series of submodels (or reduced models)

(X_1) :
(X_1, X_2) :
$\vdots$
$(X_1, X_2, \ldots, X_p)$:

Decomposition of sum of squares

- For each model, $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \ldots + \beta_p X_{ip} + \epsilon_i$
- We can calculate its SST

and SSR$(X_1, \ldots, X_p)$ and SSE$(X_1, \ldots, X_p)$

Decomposition of sum of squares

- For any model:

$$SST = SSE(X_1, \dots, X_p) + SSR(X_1, \dots, X_p)$$

$SSR(X_1, \dots, X_p) =$

$SST =$

- Decompostion of degrees of freedom

Interpretation of SSE and SSR

- $SSE(X_1)$:

- $SSR(X_1)$:

- $SSE(X_1, X_2)$:

- $SSR(X_1, X_2)$:

- $SSR(X_2|X_1)$