Lecture 8

Example 1

$n = 100$

$x_i \overset{iid}{\sim} N(0,1)$ for $i = 1, \dots, 100$

$e_i \overset{iid}{\sim} N(0,1)$

$Y_i = x_i + e_i$ $(\beta_0 = 0, \ \beta_1 = 1)$

```
> x <- rnorm(100)
>  e <- rnorm(100)
>  y <- x+e
>  Model1 <- lm(y~x)
>  Model2 <- lm(y~x + I(x^2))
>  Model3 <- lm(y~x + I(x^2) + I(x^3))
>
> summary(Model1)
```

practice problem.

For simple regression $Y \sim x$ and $H_0: \beta_1 = 0$ v.s $H_A: \beta_1 \neq 0$ we have $(t\text{-statistic})^2 = F\text{-statistic}$

$(t\text{-statistic})^2 \overset{H_0}{\sim} (t\text{-dist with } df = n-2)^2 \overset{\text{by def.}}{=} \left(\dfrac{N(0,1)}{\sqrt{\chi^2_{n-2}/(n-2)}}\right)^2$ (ind.)

$F\text{-statistic} \overset{H_0}{\sim} F\text{-dist}(df = 1, n-2) \overset{\text{by definition}}{=} \dfrac{\chi^2_1/1}{\chi^2_{n-2}/(n-2)}$ (ind.)

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3898 -0.7832  0.1945  0.7032  3.0001

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01262    0.10737   0.118    0.907
x            1.07479    0.10654  10.088   <2e-16 ***
```

→ $H_0: \beta_1 = 0$ v.s $H_A: \beta_1 \neq 0$

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.056 on 98 degrees of freedom
Multiple R-squared:  0.5094,	Adjusted R-squared:  0.5044
F-statistic: 101.8 on 1 and 98 DF,  p-value: < 2.2e-16
```

→ $H_0: E[Y|x] = \beta_0$ v.s $H_A: E(Y|x) = \beta_0 + \beta_1 x$

```
> summary(Model2)

Call:
lm(formula = y ~ x + I(x^2))
```

$Y \sim x + x^2$ mean function $E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.3976 -0.7857  0.1962  0.6969  3.0042

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.006729   0.136642   0.049    0.961
x           1.071566   0.116515   9.197 7.28e-15 ***
I(x^2)      0.006380   0.090755   0.070    0.944
```

$E(Y|x) = \beta_0 + \beta_2 x^2$ ⇕ → $H_0: \beta_1 = 0$ v.s $H_A: \beta_1 \neq 0$

→ $H_0: \beta_2 = 0$ v.s $H_A: \beta_2 \neq 0$ ⇕ $E(Y|x) = \beta_0 + \beta_1 x$

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.061 on 97 degrees of freedom
Multiple R-squared:  0.5095,	Adjusted R-squared:  0.4993
F-statistic: 50.37 on 2 and 97 DF,  p-value: 9.954e-16
```

$97 = 100 - 3$

→ $H_0: E[Y|x] = \beta_0$ v.s $H_A: E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$

⇕ $\beta_1 = \beta_2 = 0$

```
> summary(Model3)

Call:
lm(formula = y ~ x + I(x^2) + I(x^3))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.4012 -0.7753  0.1866  0.6933  2.9917

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.001833   0.141756   0.013    0.990
x            1.092728   0.191718   5.700 1.32e-07 ***
I(x^2)       0.013926   0.106066   0.131    0.896
I(x^3)      -0.009714   0.069672  -0.139    0.889
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.067 on 96 degrees of freedom
Multiple R-squared:  0.5096,	Adjusted R-squared:  0.4942
F-statistic: 33.25 on 3 and 96 DF,  p-value: 7.993e-15
```

$H_0: E[Y|x] = \beta_0$

$H_A: E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

96 DF: " 100 - 4

Example 2

for $i = 1, \dots, n = 100$

$x_i \sim N(0,1)$, $e_i \sim N(0,1)$

$y_i = 2x_i + e_i$ ($\beta_0 = 0$, $\beta_1 = 2$)

```
> x <- rnorm(100)
> y <- 2*x+rnorm(100)
>
> Model1 <- lm(y ~ x)
> Model2 <- lm(y ~ x + I(x^2))
> Model3 <- lm(y ~ x + I(x^2) + I(x^3))
>
> summary(Model2)

Call:
lm(formula = y ~ x + I(x^2))

Residuals:
     Min       1Q   Median       3Q      Max
-2.16954 -0.71778 -0.09017  0.71397  2.56037

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04546    0.12146   0.374    0.709
x            1.98096    0.09620  20.593   <2e-16 ***
I(x^2)       0.10624    0.06872   1.546    0.125
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9518 on 97 degrees of freedom
Multiple R-squared:  0.8258,	Adjusted R-squared:  0.8223
F-statistic:   230 on 2 and 97 DF,  p-value: < 2.2e-16
```

rnorm(100): e_i

I(x^2) 0.125 → $H_0: \beta_2 = 0$ v.s $H_A: \beta_2 \neq 0$.

```
>
> anova(Model1, Model2)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x + I(x^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     98 90.043
```

← $H_0: E[Y|x] = \beta_0 + \beta_1 x$ v.s $H_A: E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$.

$\Leftrightarrow \beta_2 = 0$ $\Leftrightarrow \beta_2 \neq 0$

✓ Ho Model 1

HA Model 2

Step 1

$100 - 2 = df_{H_0}$ RSS_{H_0}

2.1653/1 Ho

100-3 = df_{HA} RSSHA F-statistic = ———— / 87.878/97 ~ F-dist (1, 97)

```
Step 2     97 87.878  1     2.1653 2.3901 0.1254
>
```

$df_{H_0} - df_{HA}$ $RSS_{H_0} - RSS_{HA}$ ← p-value

```
> summary(Model3)

Call:
lm(formula = y ~ x + I(x^2) + I(x^3))

Residuals:
     Min       1Q   Median       3Q      Max
-1.82899 -0.75960 -0.02243  0.74387  2.49038

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09454    0.12562   0.753    0.454
x            2.17362    0.16557  13.128   <2e-16 ***
I(x^2)       0.05192    0.07826   0.663    0.509
I(x^3)      -0.07269    0.05098  -1.426    0.157
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9468 on 96 degrees of freedom
Multiple R-squared:  0.8295,	Adjusted R-squared:  0.8241
F-statistic: 155.6 on 3 and 96 DF,  p-value: < 2.2e-16

>
> anova(Model1, Model3)
Analysis of Variance Table

Ho Model 1: y ~ x
HA Model 2: y ~ x + I(x^2) + I(x^3)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
Step 1     98 90.043
Step 2     96 86.055  2    3.9879 2.2244 0.1137
>
```

Model 1: → $E[Y|x] = \beta_0 + \beta_1 x$

Model 2: → $E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.

F-statistic = $\frac{3.9879/2}{86.055/96} \overset{H_0}{\sim}$ F-dist (2, 98)

df_{H_0} df_{HA} RSS_{HA} RSS_{H_0} $df_{H_0} - df_{HA}$ $RSS_{H_0} - RSS_{HA}$ p-value

```
>
> anova(Model3)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
x          1 414.54  414.54 462.4504 <2e-16 ***
I(x^2)     1   2.17    2.17   2.4155 0.1234
I(x^3)     1   1.82    1.82   2.0332 0.1571
Residuals 96  86.06    0.90
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

$y \sim x + x^2 + x^3$ (Type I Anova)

$df_{H_0} - df_{HA}$, $RSS_{H_0} - RSS_{HA}$, $(RSS_{H_0} - RSS_{HA})/(df_{H_0} - df_{HA})$

$(RSS_{H_0} - RSS_{HA})/(df_{H_0} - df_{HA})$ ← 414.54 / $\hat\sigma^2$ estimated from Model 3 $y \sim x + x^2 + x^3$ ← 0.90 $\overset{H_0}{\sim}$ F-dist (1, 96)

x → $H_0: E[Y|x] = \beta_0$ v.s $H_A: E[Y|x] = \beta_0 + \beta_1 x$

I(x^2) → $H_0: E[Y|x] = \beta_0 + \beta_1 x$ v.s $H_A: E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$

I(x^3) → $H_0: E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$ v.s $H_A: E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.

df Model 3 = 100 − 4, RSS Model 3, $\hat\sigma^2$ estimated from Model 3

2.17 / 0.9 $\overset{H_0}{\sim}$ F-dist (1, 96)

```
> library(car)
> Anova(Model3)
Anova Table (Type II tests)

Response: y
          Sum Sq Df  F value Pr(>F)
x        154.501  1 172.3558 <2e-16 ***
I(x^2)     0.395  1   0.4401 0.5086
I(x^3)     1.823  1   2.0332 0.1571
Residuals 86.055 96
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

$RSS_{H_0} - RSS_{HA}$, $df_{H_0} - df_{HA}$

$(RSS_{H_0} - RSS_{HA})/(df_{H_0} - df_{HA})$ / $\hat\sigma^2$ estimated from Model 3 $\overset{H_0}{\sim}$ F-dist (1, 96)

x → $H_0: E[Y|x] = \beta_0 + \beta_2 x^2 + \beta_3 x^3$ v.s $H_A: E[Y|x] = \underbrace{\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3}_{\text{Model 3}}$

I(x^2) → $H_0: E[Y|x] = \beta_0 + \beta_1 x + \beta_3 x^3$ v.s H_A: model 3

I(x^3) → $H_0: E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$ v.s H_A: model 3

RSS of Model 3, df of RSS of Model 3

$\hat\sigma^2$ from model 3 = $\frac{86.055}{96}$