STAT 8051 Week 2 Lab

Yu Yang

Linear Model Basic Analysis

Use the **fuel2001** dataset in **alr4** as an example.

```
library(alr4)
data("fuel2001")
dat <- data.frame(Tax = fuel2001$Tax, Dlic = fuel2001$Drivers/fuel2001$Pop*1000,
                  Income = fuel2001$Income/1000,
                  logMiles = log(fuel2001$Miles),
                  Fuel = fuel2001$FuelC/fuel2001$Pop*1000)
# View(dat)
summary(dat)
```

```
##       Tax             Dlic            Income          logMiles
##  Min.   : 7.50   Min.   : 700.2   Min.   :20.99   Min.   : 7.336
##  1st Qu.:18.00   1st Qu.: 864.1   1st Qu.:25.32   1st Qu.:10.507
##  Median :20.00   Median : 909.1   Median :27.87   Median :11.276
##  Mean   :20.15   Mean   : 903.7   Mean   :28.40   Mean   :10.914
##  3rd Qu.:23.25   3rd Qu.: 943.0   3rd Qu.:31.21   3rd Qu.:11.634
##  Max.   :29.00   Max.   :1075.3   Max.   :40.64   Max.   :12.614
##       Fuel
##  Min.   :317.5
##  1st Qu.:575.0
##  Median :626.0
##  Mean   :613.1
##  3rd Qu.:666.6
##  Max.   :842.8
```

Fit a linear model using all variables.

```
m0 <- lm(Fuel ~ ., data = dat)
summary(m0)
```

```
...
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 154.1928   194.9062   0.791 0.432938
## Tax          -4.2280     2.0301  -2.083 0.042873 *
## Dlic          0.4719     0.1285   3.672 0.000626 ***
## Income       -6.1353     2.1936  -2.797 0.007508 **
## logMiles     26.7552     9.3374   2.865 0.006259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.89 on 46 degrees of freedom
## Multiple R-squared:  0.5105, Adjusted R-squared:  0.4679
## F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.331e-07
...
```

Interpretation of coefficients

(1) Fitted equation

$Fuel = 154.1928 + -4.2280 * Tax + 0.4719 * Dlic + -6.1353 * Income + 26.7552 * logMiles$

(2) Significance of variables

Check the corresponding p-values.

Estimate of the covariance matrix of β

Estimate of σ

```
(sigma.hat <- summary(m0)$sigma)
```

```
## [1] 64.89122
```

Estimate of $Cov(\beta) = \sigma^2(X^TX)^{-1}$

```
vcov(m0)
```

```
##               (Intercept)          Tax         Dlic        Income
## (Intercept) 37988.41145 -120.0960793 -17.18034682 -251.85813715
## Tax          -120.09608    4.1213916   0.02357820    0.17952544
## Dlic          -17.18035    0.0235782   0.01651570    0.05006761
## Income       -251.85814    0.1795254   0.05006761    4.81202826
## logMiles    -1173.39232    0.9734094   0.03281593    6.07625002
##                  logMiles
## (Intercept) -1.173392e+03
## Tax          9.734094e-01
## Dlic         3.281593e-02
## Income       6.076250e+00
## logMiles     8.718655e+01
```

Estimate of $(X^TX)^{-1}$.

```
summary(m0)$cov.unscaled  # directly using the summary output
```

```
##              (Intercept)           Tax          Dlic        Income
## (Intercept)  9.021511563 -2.852049e-02 -4.079999e-03 -5.981143e-02
## Tax         -0.028520492  9.787506e-04  5.599366e-06  4.263381e-05
## Dlic        -0.004079999  5.599366e-06  3.922159e-06  1.189009e-05
## Income      -0.059811427  4.263381e-05  1.189009e-05  1.142763e-03
## logMiles    -0.278657939  2.311659e-04  7.793148e-06  1.442992e-03
##                  logMiles
## (Intercept) -2.786579e-01
## Tax          2.311659e-04
## Dlic         7.793148e-06
## Income       1.442992e-03
## logMiles     2.070512e-02
```

```
vcov(m0)/(sigma.hat^2)  # using the covariance matrix
```

```
##              (Intercept)           Tax          Dlic        Income
## (Intercept)  9.021511563 -2.852049e-02 -4.079999e-03 -5.981143e-02
## Tax         -0.028520492  9.787506e-04  5.599366e-06  4.263381e-05
## Dlic        -0.004079999  5.599366e-06  3.922159e-06  1.189009e-05
## Income      -0.059811427  4.263381e-05  1.189009e-05  1.142763e-03
## logMiles    -0.278657939  2.311659e-04  7.793148e-06  1.442992e-03
```

```
##                  logMiles
## (Intercept) -2.786579e-01
## Tax          2.311659e-04
## Dlic         7.793148e-06
## Income       1.442992e-03
## logMiles     2.070512e-02
```

Hypothesis testing and confidence interval

Case 1: consider only one coefficient, β_i

Two-sided test

$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$

T-test: use the p-value in `summary(m0)`.

F-test:

```
m1 <- update(m0, ~ .-Tax, data = dat)
anova(m0, m1)
```

```
...
## Model 1: Fuel ~ Tax + Dlic + Income + logMiles
## Model 2: Fuel ~ Dlic + Income + logMiles
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     46 193700
## 2     47 211964 -1    -18264 4.3373 0.04287 *
...
```

One-sided test

$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 < 0$

T-test: use the p-value in `summary(m0)`, p-value = 0.042873 / 2 = 0.0214365.

$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 > 0$

T-test: use the p-value in `summary(m0)`, p-value = 1- 0.042873 / 2 = 0.9785635.

Confidence interval

Use the `Estimate` and `Std.Error` in `summary(m0)`. Or use `confint()`.

```
beta1.hat <- summary(m0)$coefficients[2, 1]
beta1.se <- summary(m0)$coefficients[2, 2]

df <- 51-5  # n-p
#uses t quantiles
c(beta1.hat-qt(0.975, df)*beta1.se, beta1.hat+qt(0.975, df)*beta1.se)
```

```
## [1] -8.3144050 -0.1415614
```

```
confint(m0)[2, ]
```

```
##      2.5 %     97.5 %
## -8.3144050 -0.1415614
```

```
# uses normal quantiles
c(beta1.hat-qnorm(0.975)*beta1.se, beta1.hat+qnorm(0.975)*beta1.se)
```

```
## [1] -8.206947 -0.249019
```

```
confint.default(m0)[2, ]
```

```
##     2.5 %    97.5 %
## -8.206947 -0.249019
```

Case 2: consider the linear combination of the coefficients, $a^T\beta$

$H_0: \beta_1 = \beta_2 + \beta_3 \quad vs. \quad H_1: \beta_1 > \beta_2 + \beta_3$

Note $a^T\hat{\beta} \sim N(a^T\beta, a^T\sigma^2(X^TX)^{-1}a)$ and $\frac{a^T\hat{\beta}}{se(a^T\hat{\beta})} \sim_{H_0} t_{n-p}$.

```
a <- c(0, 1, -1, -1, 0)
(estimate <- t(a)%*%summary(m0)$coefficient[, 1])
```

```
##          [,1]
## [1,] 1.435477
```

```
(se <- t(a)%*%vcov(m0)%*%a)
```

```
##          [,1]
## [1,] 8.643864
```

```
(t.value <- estimate / se)
```

```
##           [,1]
## [1,] 0.1660689
```

```
(p.value <- pt(t.value, df, lower.tail = FALSE))  # cannot reject H0
```

```
##           [,1]
## [1,] 0.4344153
```

```
# confidence interval
c(estimate - qt(0.975, df)*se, estimate + qt(0.975, df)*se)
```

```
## [1] -15.96372  18.83467
```