MATH - 4360: Linear Statistical Models

Chapter 3: Multiple Linear Regression

Suthakaran Ratnasingam

Multiple Regression Model

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j + \cdots + \beta_k X_k + \epsilon$$

Thus, the least - squares estimator of β is

$$\hat{\beta} = (X'X)^{-1}_{p\times p}(X'y)_{p\times 1}$$

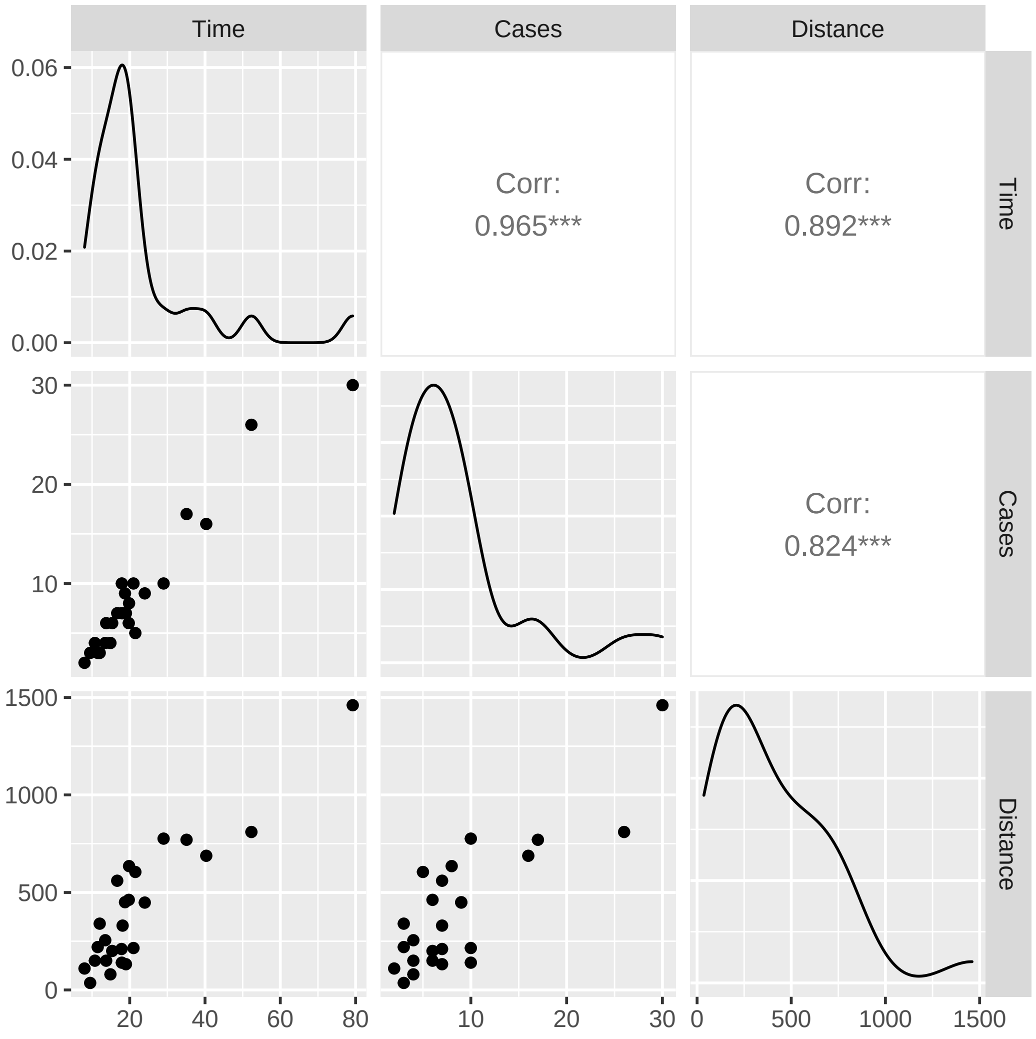
```
rm(list = ls())
# I assume that you have installed the following R packages. If not, please install
# them using the R function: install.packages('package_name')
library(olsrr)
library(ggfortify)
library(ggplot2)
library(tidyverse)
library(car)
library(Rcpp)
library(GGally)
library(leaps)
library(matlib) # enables function inv()
data1 = read.table("D:\\CSUSB\\Fall 2021\\MATH 4360\\RNotes\\ex31.txt", header = TRUE)
head(data1)
```

Example 3.1 Delivery Time Data for Example

```
##    Time Cases Distance
## 1 16.68     7      560
## 2 11.50     3      220
## 3 12.03     3      340
## 4 14.88     4       80
## 5 13.75     6      150
## 6 18.11     7      330
```

Scatterplot matrix for the delivery time data

```
GGally::ggpairs(data1)
```



```r
n = nrow(data1)
Fit1 = lm(Time ~ Cases + Distance, data = data1)
p = length(coef(Fit1))
X = cbind(rep(1, n), data1$Cases, data1$Distance) # X matrix
y = data1$Time

Xt = t(X) # Transpose of X
dim(Xt)
```

```
## [1]  3 25
```

```r
# Lets find XtX matrix
Xt_X = t(X) %*% X
dim(Xt_X)
```

```
## [1] 3 3
```

```r
# Estimate the coefficients
beta_hat = solve(t(X) %*% X) %*% t(X) %*% y
beta_hat
```

```
##            [,1]
## [1,] 2.34123115
## [2,] 1.61590721
## [3,] 0.01438483
```

Using `lm()` function in `R`

```
Fit1 = lm(Time ~ Cases + Distance, data = data1)
summary(Fit1)
```

```
##
## Call:
## lm(formula = Time ~ Cases + Distance, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7880 -0.6629  0.4364  1.1566  7.4197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.341231   1.096730   2.135 0.044170 *
## Cases       1.615907   0.170735   9.464 3.25e-09 ***
## Distance    0.014385   0.003613   3.981 0.000631 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 22 degrees of freedom
## Multiple R-squared:  0.9596, Adjusted R-squared:  0.9559
## F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

```
coef(Fit1)
```

```
## (Intercept)       Cases    Distance
##  2.34123115  1.61590721  0.01438483
```

3.2.4: Estimation of σ^2

The estimate of σ^2 is the residual mean square

$$E\left[\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{(n-p)}\right]=E\left[\frac{e'e}{(n-p)}\right]=E\left[\frac{y'(I-H)y}{(n-p)}\right]=\sigma^2,\quad \text{or}$$

$$E\big[MS_{Res}\big]=\sigma^2$$

where $MS_{Res}=\dfrac{SS_{Res}}{(n-p)}$ is the mean sum of squares due to residual. Thus an unbiased estimator of σ^2 is (that the expected value of MS_{Res} is σ^2, so an unbiased estimator of σ^2 is given by)

$$\hat{\sigma}^2=MS_{Res}=S^2$$

```
p = length(beta_hat)
p
```

```
## [1] 3
```

```
# Method 1
SS_Res = sum((Fit1$residuals)^2)
SS_Res
```

```
## [1] 233.7317
```

```
sigmahat_squared  = SS_Res/(n - p)
sigmahat_squared
```

```
## [1] 10.62417
```

```
# Method 2
y_hat = X %*% solve(t(X) %*% X) %*% t(X) %*% y
e     = y - y_hat
t(e) %*% e / (n - p)
```

```
##          [,1]
## [1,] 10.62417
```

```
# Method 3
sum((y - y_hat) ^ 2) / (n - p)
```

```
## [1] 10.62417
```

```
# Method 4
(summary(Fit1)$sigma)^2
```

```
## [1] 10.62417
```

Variance

The variance of $\hat{\beta}$ can be obtained as the sum of variances of all $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ which is the trace of covariance matrix of $\hat{\beta}$. Thus

$$\begin{aligned}
Var(\hat{\beta}) &= tr(Var(\hat{\beta})) \\
&= \sum_{i=1}^{k} E(\hat{\beta}_i - \beta_i)^2 \\
&= \sum_{i=1}^{k} Var(\hat{\beta}_i) \\
Var(\hat{\beta}) &= Var\big[(X'X)^{-1}X'y\big] \\
&= (X'X)^{-1}X'Var(y)\big[(X'X)^{-1}X'\big]' \\
&= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

```
sigmahat_squared = (summary(Fit1)$sigma)^2
XtX_inv = solve(t(X) %*% X)
XtX_inv
```

```
##               [,1]          [,2]          [,3]
## [1,]  1.132152e-01 -4.448593e-03 -8.367257e-05
## [2,] -4.448593e-03  2.743783e-03 -4.785709e-05
## [3,] -8.367257e-05 -4.785709e-05  1.228745e-06
```

```
var_beta = sigmahat_squared*solve(t(X) %*% X)
var_beta
```

```
##               [,1]          [,2]          [,3]
## [1,]  1.2028170618 -0.0472625981 -8.889514e-04
## [2,] -0.0472625981  0.0291504123 -5.084417e-04
## [3,] -0.0008889514 -0.0005084417  1.305439e-05
```

Analysis of Variance

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$SS_T = SS_R + SS_{Res}$$

In multiple regression, hypothesis testing is

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad vs \quad H_1 : \beta_j \neq 0 \quad \text{at least one } j$$

Further, SS_{Res}/σ^2 follows a χ^2_{n-k-1} distribution and that SS_{Res} and SS_R are independent.

By the definition of an $F-$ statistic

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}}$$

follows the $F_{k,n-k-1}$ distribution.

We reject H_0 if

$$F_0 \geq F_{\alpha,k,n-k-1}$$

Table 1: Analysis of Variance for Significance of Regression in Multiple Regression

Source of variation	Sum of squares	Degrees of freedom	Mean square	F_0
Regression	SS_R	k	$\frac{SS_R}{k} = MS_R$	$\frac{MS_R}{MS_{Res}}$
Residual	SS_{Res}	$(n-k-1)$	$\frac{SS_{Res}}{(n-k-1)} = MS_{Res}$	
Total	SS_T	$(n-1)$		

$$SS_T = SS_R + SS_{Res}$$

$$SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y})^2 = y'y - \hat{\beta}'X'y$$

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SS_R = \hat{\beta}'X'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

```
SS_Res = sum((Fit1$residuals)^2)
SS_Res
```

```
## [1] 233.7317
```

```
y = as.matrix(y, ncol = 1)
Xt_y = Xt %*% y
Xt_y
```

```
##            [,1]
## [1,]    559.60
## [2,]   7375.44
## [3,] 337071.69
```

```
SS_Res = t(y) %*% y -  t(beta_hat) %*% Xt_y
SS_Res
```

```
##           [,1]
## [1,] 233.7317
```

```
SS_T = sum((y - mean(y))^2)
SS_T
```

```
## [1] 5784.543
```

```
SS_R = t(beta_hat) %*% Xt_y  - 1/n*(sum(y))^2
SS_R
```

```
##           [,1]
## [1,] 5550.811
```

```
anova(Fit1)
```

```
## Analysis of Variance Table
##
## Response: Time
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Cases      1 5382.4  5382.4 506.619 < 2.2e-16 ***
## Distance   1  168.4   168.4  15.851 0.0006312 ***
## Residuals 22  233.7    10.6
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova_fit = anova(Fit1)
tab = as.table(cbind(
  'SS' = c('Regression' = sum(anova_fit[1:2, 2]),
           'Residual' = anova_fit[3, 2],
           'Total' = sum(anova_fit[1:3, 2])),
  'Df' = c( sum(anova_fit[1:2, 1]),
            anova_fit[3, 1],
            sum(anova_fit[1:3, 1])),
  'MS' = c( sum(anova_fit[1:2, 2])/sum(anova_fit[1:2, 1]),
            anova_fit[3, 2] / anova_fit[3, 1],
            NA),
  'F-Test' = c( (sum(anova_fit[1:2, 2])/sum(anova_fit[1:2, 1]))/(anova_fit[3, 2] / anova_fit[3, 1]),
                NA,
                NA)
))
round(tab, 4)
```

```
##                     SS        Df        MS    F-Test
## Regression 5550.8109    2.0000 2775.4055  261.2351
## Residual    233.7317   22.0000   10.6242
## Total      5784.5426   24.0000
```

3.3.2: Test of hypothesis on individual regression coefficients

```
summary(Fit1)$coef
```

```
##               Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept) 2.34123115 1.096730168 2.134738 4.417012e-02
## Cases       1.61590721 0.170734918 9.464421 3.254932e-09
## Distance    0.01438483 0.003613086 3.981313 6.312469e-04
```

3.4: Confidence Intervals in Multiple Regression

Let $p = k + 1$. A a $(l - \alpha)100\%$ confidence interval for the regression coefficient $\beta_j, \quad j = 0, 1, 2, \ldots, k$ as

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \;\; \leq \;\beta_j\; \leq \;\; \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

```
Xt_X_inv = solve(t(X) %*% X)
Xt_X_inv
```

```
##               [,1]          [,2]          [,3]
## [1,]  1.132152e-01 -4.448593e-03 -8.367257e-05
## [2,] -4.448593e-03  2.743783e-03 -4.785709e-05
## [3,] -8.367257e-05 -4.785709e-05  1.228745e-06
```

```
# beta_0
t_value = qt(0.975, df = 22)
beta_hat[1,] + c(-1,1)*t_value* sqrt(sigmahat_squared * XtX_inv[1,1])
```

```
## [1] 0.06675199 4.61571030
```

```
# beta_1
beta_hat[2,] + c(-1,1)*t_value* sqrt(sigmahat_squared * XtX_inv[2,2])
```

```
## [1] 1.261825 1.969990
```

```
# beta_2
beta_hat[3,] + c(-1,1)*t_value* sqrt(sigmahat_squared * XtX_inv[3,3])
```

```
## [1] 0.006891745 0.021877908
```

```
#Using R function confint()
confint(Fit1)
```

```
##                   2.5 %     97.5 %
## (Intercept) 0.066751987 4.61571030
## Cases       1.261824662 1.96998976
## Distance    0.006891745 0.02187791
```

3.4.2: Confidence Interval Estimation of the Mean Response

We may construct a confidence interval on the mean response at a particular point, such as $x_{01}, x_{02}, \ldots, x_{0k}$. Define the vector x_0 as

$$x_0 = (1, x_{01}, x_{02}, \cdots, x_{0k})'$$

Then our estimate of $E[Y|x_0]$ for a set of values x_0 is given by the fitted value at this point is

$$\begin{aligned}\hat{y}_0 &= x_0'\hat{\beta} \\ &= \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_k x_{0k}\end{aligned}$$

This is an unbiased estimator of $E(y|x_0)$, since $E(\hat{y}_0) = x_0'\hat{\beta} = E(y|x_0)$, and the variance of $\hat{y}_0$ is

$$Var(\hat{y}_0) = \sigma^2 x_0'(X'X)^{-1}x_0$$

Therefore, a $(l-\alpha)100\%$ confidence interval on the mean response at the point is $x_{01}, x_{02}, \ldots, x_{0k}$ is

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\sigma^2 x_0'(X'X)^{-1}x_0} \;\; \leq \; \beta_j \; \leq \;\; \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\sigma^2 x_0'(X'X)^{-1}x_0}$$

would like to construct a 95% CI on the mean delivery time for an outlet requiring $x_1 = 8$ cases and where the distance $x_2 = 275$ feet.

```
new_obs = data.frame(Cases = c(8), Distance = c(275))
predict(Fit1, newdata = new_obs, interval = "confidence", level = 0.95)
```

```
##        fit     lwr      upr
## 1 19.22432 17.6539 20.79474
```

Prediction Intervals of New Observations

1. The prediction in the multiple regression model has two aspects. 1) Prediction of the average value of study variable or mean response. 2)Prediction of the actual value of the study variable.
2. The particular values of the regressor variables, for example, $x_0' = [1, x_{01}, x_{02}, \ldots, x_{0k}]$ then a point estimate of the future observation y_0 at the point $x_{01}, x_{02}, \ldots, x_{0k}$

$$\hat{y}_0 = x_0'\hat{\beta}$$

We now develop a prediction interval for the future observation y_0. Note that the random variable $\psi = y_0 - \hat{y}_0$

$$\begin{aligned} Var(\psi) &= Var(y_0 - \hat{y}_0) \\ &= Var(y_0) + Var(\hat{y}_0) \\ &= \sigma^2 + \sigma^2 x_0'(X'X)^{-1}x_0 \\ &= \sigma^2\left(1 + x_0'(X'X)^{-1}x_0\right) \end{aligned}$$

Because the future observation y_0 is independent of $\hat{y}_0$. If we use $\hat{y}_0$ to predict y_0, then the standard error of $\psi = y_0 - \hat{y}_0$ is the appropriate statistic on which to base a prediction interval.

A $(1-\alpha)100\%$ prediction interval for this future observation is

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\sigma^2\left[1 + x_0'(X'X)^{-1}x_0\right]} \;\; \leq \; y_0 \; \leq \;\; \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\sigma^2\left[1 + x_0'(X'X)^{-1}x_0\right]}$$

```
new_obs = data.frame(Cases = c(8), Distance = c(275))
predict(Fit1, newdata = new_obs, interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 19.22432 12.28456 26.16407
```

References

- Introduction to Linear Regression Analysis, 5th Edition, by Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining (Wiley), ISBN: 978-0-470-54281-1.
- R Core Team (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
- RStudio Team (2020). RStudio: Integrated Development Environment for R. Boston, MA: RStudio, PBC.