Chapter 3
Multiple Linear Regression

April 1, 2010

Example 3.1: The Delivery Time Data

1. Data & Plots

Observation	Delivery Time(y)	Number of Cases(x1)	Distance(x2)
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
7	8.00	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
⋮	⋮	⋮	⋮
24	19.83	8	635
25	10.75	4	150

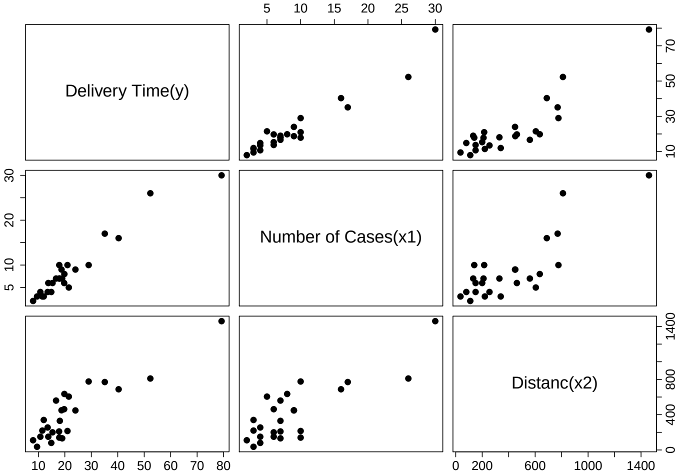
Example 3.1: The Delivery Time Data (cont.)

```
> url <- "https://raw.github.com/dongikjang/regression/master/"
> rfun <- getURL(paste(http, "read.xls2.r",sep=""))
> eval(parse(text=rfun))
> # If OS is Windows then install "xlsReadWrite" package
> # If OS is Mac or Linux then install "gdata" package
>
> library(RCurl)
> tf <- paste(tempfile(), "xls", sep = ".")
> download.file(paste(url, "Dataset/data-ex-3-1.xls", sep=""), tf,
+               method="curl")
  % Total    % Received % Xferd  Average Speed   Time    Time     Time
                                 Dload  Upload   Total   Spent    Left
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:
> data_3.1 <- read.xls2(tf, header=TRUE)
> colnames(data_3.1) <- c("Observation", "Delivery Time(y)",
+                         "Number of Cases(x1)", "Distanc(x2)")
```

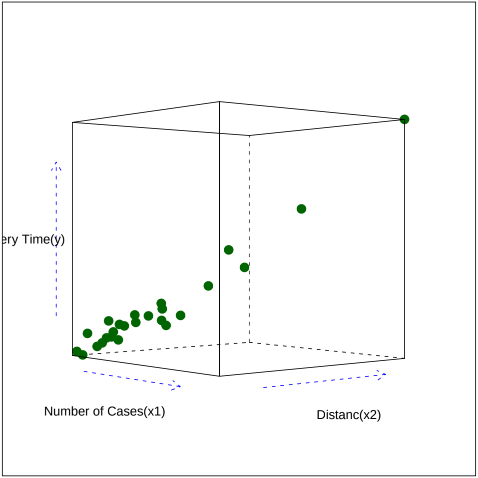
Example 3.1: The Delivery Time Data (cont.)

```
> head(data_3.1)
  Observation Delivery Time(y) Number of Cases(x1) Distanc(x2)
1           1            16.68                   7         560
2           2            11.50                   3         220
3           3            12.03                   3         340
4           4            14.88                   4          80
5           5            13.75                   6         150
6           6            18.11                   7         330
```

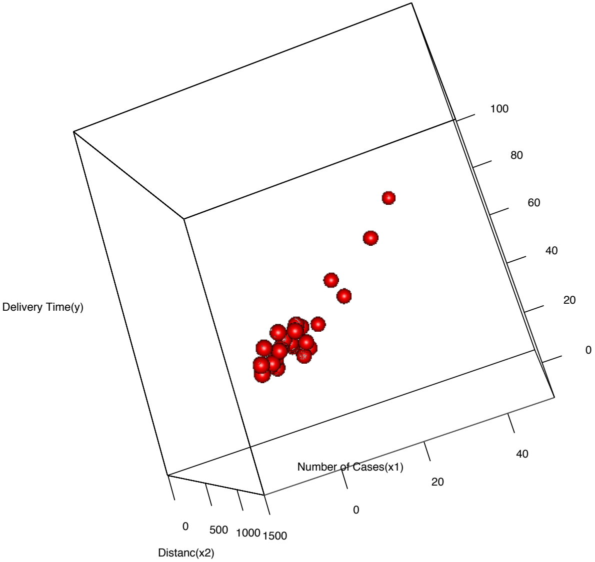
Example 3.1: The Delivery Time Data (cont.)



Example 3.1: The Delivery Time Data (cont.)



Example 3.1: The Delivery Time Data (cont.)



Example 3.1: The Delivery Time Data (cont.)

```
> # scatterplot matrices
> par(mar=c(4.5,5,5,2),cex.main=2, cex.lab=1.5, cex.axis=1.5)
> pairs(data_3.1[,2:4], pch=19, cex=1.5)
>
> # 3d scatter plot
> library(lattice)        #need lattice package
> trellis.par.get("background")$col
[1] "transparent"
> trellis.par.set(theme=col.whitebg())
> par(mar=c(4.5,7,5,2), cex.main=2, cex.lab=1.5, cex.axis=1.5)
> cloud(data_3.1[,2]~data_3.1[,3]*data_3.1[,4], cex=1.5,
+          scales=list(col="blue", lty=2, cex=2),
+          screen=list(x=-90, y=-50, z=0), pch=16,
+          xlab=colnames(data_3.1)[3],
+          ylab=colnames(data_3.1)[4],
+          zlab=colnames(data_3.1)[2])
```

Example 3.1: The Delivery Time Data (cont.)

```
> # interactive 3D scatterplot
> library(rgl)
> plot3d(x=data_3.1[,3], y=data_3.1[,4], z=data_3.1[,2],
+           radius=20, type="s", col=2,
+           xlab=colnames(data_3.1)[3],
+           ylab=colnames(data_3.1)[4],
+           zlab=colnames(data_3.1)[2])
```

Example 3.1: The Delivery Time Data (cont.)

2. Multiple linear regression fit

```
> # linear models fit
> nl <- colnames(data_3.1)
> colnames(data_3.1) <- c("obs", "d_time", "n_case", "dista")
> attach(data_3.1)
> lmfit <- lm(d_time~n_case+dista)
> lmfit

Call:
lm(formula = d_time ~ n_case + dista)

Coefficients:
(Intercept)       n_case        dista
    2.34123      1.61591      0.01438
```

Example 3.1: The Delivery Time Data (cont.)

```
> # summary of fitted model
> (sfit <- summary(lmfit))

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 *
n_case      1.615907   0.170735   9.464 3.25e-09 ***
dista       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
```

Example 3.1: The Delivery Time Data (cont.)

3. Anova table

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

Response: d_time
          Df Sum Sq Mean Sq F value    Pr(>F)
n_case     1 5382.4  5382.4 506.619 < 2.2e-16 ***
dista      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
```

Example 3.1: The Delivery Time Data (cont.)

```
# modification of default anova table of R
> anova2 <- function(x){
+         fit <- anova(x)
+         nrows <- nrow(fit)
+         fit[1,1:2] <- apply(fit[1:(nrows-1), 1:2], 2, sum)
+         fit <- fit[-(2:(nrows-1)), ]
+         fit[1,3] <- fit[1,2]/fit[1,1]
+         fit[1,4] <- fit[1,3]/fit[2,3]
+         rownames(fit)[1] <- "Regression"
+         fit[1,5] <- pf(fit[1,4], fit[1,1], fit[2,1],
+                        lower.tail=FALSE)
+         return(fit)
+ }
```

Example 3.1: The Delivery Time Data (cont.)

```
> # modified anova table
> anova2(lmfit)
Analysis of Variance Table

Response: d_time
           Df Sum Sq Mean Sq F value    Pr(>F)
Regression  2 5550.8 2775.41  261.24 4.687e-16 ***
Residuals  22  233.7   10.62
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Example 3.1: The Delivery Time Data (cont.)

```
> # table 3.3
> ta3.3 <- cbind(d_time, fitted(lmfit), residuals(lmfit))
> colnames(ta3.3) <- c("y", "y_hat", "residual")
> head(round(ta3.3, digits=2))
      y y_hat residual
1 16.68 21.71    -5.03
2 11.50 10.35     1.15
3 12.03 12.08    -0.05
4 14.88  9.96     4.92
5 13.75 14.19    -0.44
6 18.11 18.40    -0.29
```

Example 3.1: The Delivery Time Data (cont.)

4. Estimation of σ^2

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

Response: d_time
          Df Sum Sq Mean Sq F value    Pr(>F)
n_case     1 5382.4  5382.4 506.619 < 2.2e-16 ***
dista      1  168.4   168.4  15.851 0.0006312 ***
Residuals 22  233.7    10.6

> (summary(lmfit)$sigma)^2
[1] 10.62417
```

Example 3.1: The Delivery Time Data (cont.)

5. We can also do it directly using the F-testing formula:

```
> (tss <- sum((d_time-mean(d_time))^2)        #sum of square total
[1] 5784.543
> (sse <- deviance(lmfit))        #sum of square err
[1] 233.7317
> (df.r <- df.residual(lmfit))        #n-p-1
[1] 22
> p <- 2
> (fstat <- ((tss-sse)/p)/(sse/df.r))        #F-statistics
[1] 261.2351
> pf(fstat, p, df.residual(lmfit), lower.tail=FALSE)        #p-value
[1] 4.687422e-16
```

Example 3.1: The Delivery Time Data (cont.)

6. Tests on individual regression coefficients

```
> # F-test
> # Reduced Model (H0 : coefficient of n_dist = 0)
> redfit <- lm(d_time ~ n_case)
> (sse1 <- deviance(redfit))        #SSE of Reduced Model
[1] 402.1338
> (fstat <- (deviance(redfit)-deviance(lmfit))/
+ (deviance(lmfit)/df.residual(lmfit)))
[1] 15.85085
> pf(fstat, 1, df.residual(lmfit), lower.tail=FALSE)
[1] 0.0006312469
> sqrt(fstat)
[1] 3.981313
```

Example 3.1: The Delivery Time Data (cont.)

```
> summary(lmfit)$coef
              Estimate  Std. Error  t value     Pr(>|t|)
(Intercept) 2.34123115 1.096730168 2.134738 4.417012e-02
n_case      1.61590721 0.170734918 9.464421 3.254932e-09
dista       0.01438483 0.003613086 3.981313 6.312469e-04

> # t-test  (H0 : coefficient of n_dist = 0)
> (tstat <- summary(lmfit)$coef[3,3])
[1] 3.981313
> 2*pt(sqrt(fstat), df.residual(lmfit), lower.tail=FALSE)
[1] 0.0006312469
```

Example 3.1: The Delivery Time Data (cont.)

```
> # Using anova function
> anova(redfit, lmfit)
Analysis of Variance Table

Model 1: d_time ~ n_case
Model 2: d_time ~ n_case + dista
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     23 402.13
2     22 233.73  1    168.40 15.851 0.0006312 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Example 3.1: The Delivery Time Data (cont.)

7. Testing equality of regression coefficient

```
> # (Ho: coefficient of n.case = coefficient of dist)
> redfit2 <- lm(d_time ~ I(n_case + dista))
> anova(redfit2, lmfit)
Analysis of Variance Table

Model 1: d_time ~ I(n_case + dista)
Model 2: d_time ~ n_case + dista
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     23 1136.63
2     22  233.73  1    902.89 84.985 5.192e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Example 3.1: The Delivery Time Data (cont.)

8. Test whether one of the coefficients can be set to a particular value

```
> # (Ho : coefficient of dist = 0.5)
> redfit3 <- lm(d_time ~ n_case + offset(0.5*dista))
> anova(redfit3, lmfit)
Analysis of Variance Table

Model 1: d_time ~ n_case + offset(0.5 * dista)
Model 2: d_time ~ n_case + dista
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     23 192155
2     22    234  1    191921 18065 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> (tstat <- (summary(lmfit)$coef[3,1]-0.5)/summary(lmfit)$coef[3,2])
[1] -134.4045
> 2*pt(abs(tstat), df.residual(lmfit), lower.tail=FALSE)
[1] 1.451327e-33
> tstat^2
[1] 18064.58
```

Example 3.1: The Delivery Time Data (cont.)

9. Confdence interval on the regression coefficients

```
> summary(lmfit)$coef
              Estimate  Std. Error  t value     Pr(>|t|)
(Intercept) 2.34123115 1.096730168 2.134738 4.417012e-02
n_case      1.61590721 0.170734918 9.464421 3.254932e-09
dista       0.01438483 0.003613086 3.981313 6.312469e-04
> sfit <- summary(lmfit)
> t.025 <- qt(0.975, df.residual(lmfit))
> c(sfit$coef[2,1] - t.025*sfit$coef[2,2],
+   sfit$coef[2,1] + t.025*sfit$coef[2,2])
[1] 1.261825 1.969990
> confint(lmfit)
                  2.5 %     97.5 %
(Intercept) 0.066751987 4.61571030
n_case      1.261824662 1.96998976
dista       0.006891745 0.02187791
> confint(lmfit, parm='dista', level = 0.95)
            2.5 %     97.5 %
dista 0.006891745 0.02187791
```

Example 3.1: The Delivery Time Data (cont.)

10. Confidence interval estimation of the mean response

```
> x0 <- c(1, 8, 275)
> (y0 <- sum(x0*coef(lmfit)))
[1] 19.22432
> t.025 <- qt(0.975, df.residual(lmfit))
> x <- model.matrix(lmfit)
> xtxi <- solve(t(x) %*% x)
> bm <- sqrt(x0 %*% xtxi %*% x0) *t.025 * summary(lmfit)$sigma
> c(y0-bm, y0+bm)
[1] 17.65390 20.79474
> predict(lmfit, data.frame(n_case=8,dista=275),
+          interval="confidence")
       fit      lwr      upr
1 19.22432 17.65390 20.79474
```

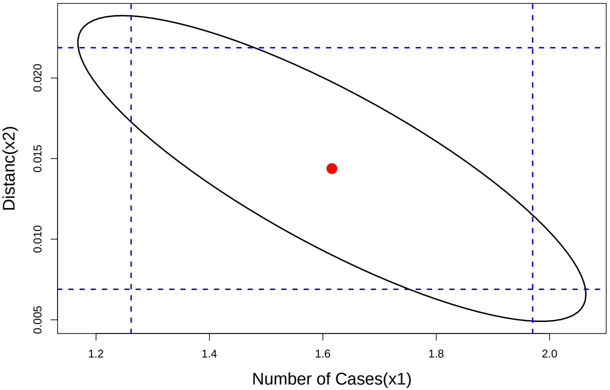
Example 3.1: The Delivery Time Data (cont.)

11. Prediction of new observation

```
> bm <- sqrt(1+x0 %*% xtxi %*% x0) *t.025 * summary(lmfit)$sigma
> c(y0-bm, y0+bm)
[1] 12.28456 26.16407
> x0 <- data.frame(n_case=8, dista=275)
> str(predict(lmfit, x0, se=TRUE))
List of 4
 $ fit           : Named num 19.2
  ..- attr(*, "names")= chr "1"
 $ se.fit        : num 0.757
 $ df            : int 22
 $ residual.scale: num 3.26
> predict(lmfit, x0, interval="confidence")
       fit      lwr      upr
1 19.22432 17.65390 20.79474
> predict(lmfit, x0, interval="prediction")
       fit      lwr      upr
1 19.22432 12.28456 26.16407
```

Example 3.1: The Delivery Time Data (cont.)

12. The joint 95% confidence region for these parameters



Example 3.1: The Delivery Time Data (cont.)

```
> library(ellipse) #need ellipse package
> plot(ellipse(lmfit, c(2,3)), type="l", lwd=2,
+      xlab="Number of Cases(x1)", ylab="Distanc(x2)")
> points(coef(lmfit)[2], coef(lmfit)[3], pch=19, col=2, cex=2)
> abline(v=confint(lmfit)[2,], lty=2, col=4, lwd=2)
> abline(h=confint(lmfit)[3,], lty=2, col=4, lwd=2)
```

Example 3.1: The Delivery Time Data (cont.)