1

STAT2008/STAT6038

Multiple Regression Interval Estimation

Multiple Linear Regression Prediction

2

We will often be interested in predicting the value of the response associated with a particular set of predictor values, $x_0 = (1, x_{01}, \ldots, x_{0k})^T$, where we have included the leading one associated with the intercept.

Our best guess for the value of the response at x_0 is:

$$\hat{Y}(x_0) = b_0 + b_1 x_{01} + \ldots + b_k x_{0k} = x_0^T b.$$

Prediction and Confidence Intervals

3

We see that

$$Var\{\hat{Y}(x_0)\} = Var(x_0^T b) = x_0^T Var(b)\{x_0^T\}^T = \sigma^2 x_0^T (X^T X)^{-1} x_0.$$

Therefore, we can see that a $100(1-\alpha)\%$ confidence interval for the expected response associated with the set of predictor values, x_0, is given by:

$$\hat{Y}(x_0) \pm t_{n-p}(1-\alpha/2) s_\epsilon \sqrt{x_0^T (X^T X)^{-1} x_0}.$$

A $100(1-\alpha)\%$ prediction interval can also be calculated as:

$$\hat{Y}(x_0) \pm t_{n-p}(1-\alpha/2) s_\epsilon \sqrt{1 + x_0^T (X^T X)^{-1} x_0}.$$

Example

4

Twenty stands of pine trees were measured in an effort to assess the amount and quality of wood which would be obtained. Each stand's age (AGE), average height of the dominant trees (HD), number of trees (N) and average diameter at 4.5 feet above the ground (MDBH) were obtained. Theory suggests that MDBH may be effectively modelled as:

$$MDBH = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

where $x_1 = HD$, $x_2 = AGE \cdot N$ and $x_3 = HD/N$.

Example

5

Suppose that we wanted to predict the expected value of all pine stands (CI) as well as the value of an individual pine stand's (PI) MDBH for three different types of stands.

The first is a 5 year old stand having 500 trees of an average height of 10 feet.

The second is a 10 year old stand having 600 trees of an average height of 80 feet.

The third is a 25 year old stand having 1000 trees of an average height of 75 feet.

Example

6

```
> pine<-read.csv("pine.csv")
> attach(pine)
> pine
   AGE   HD   N MDBH
1   19 51.5 500  7.0
2   14 41.3 900  5.0
3   11 36.7 650  6.2
4   13 32.2 480  5.2
……
18  16 50.3 730  6.9
19  14 50.5 680  6.9
20  22 57.7 480  7.9
```

Example

7

```
> x1<-HD
> x2<-AGE*N
> x3 <- HD/N
> pine.lm <- lm(MDBH ~ x1 + x2 + x3)
> anova(pine.lm)
Analysis of Variance Table

Response: MDBH
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 6.2070  6.2070 72.0117 2.559e-07 ***
x2         1 2.7846  2.7846 32.3054 3.390e-05 ***
x3         1 0.0148  0.0148  0.1714    0.6844
Residuals 16 1.3791  0.0862
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example

8

```
> x1 <- c(10,80,75)
> x2 <- c(5*500,10*600,25*1000)
> x3 <- c(10/500,80/600,75/1000)
> prd <- predict(pine.lm,as.data.frame(cbind(x1,x2,x3)),se.fit=T)
> CIup <- prd$fit+(qt(0.975,16)*prd$se.fit)
> CIlw <- prd$fit-(qt(0.975,16)*prd$se.fit)
> cbind(CIlw,prd$fit,CIup)
      CIlw                CIup
1 3.279259  3.85699  4.434721
2 9.027290 10.47730 11.927312
3 5.848304  6.57982  7.311336
```

Example

9

```
> pi.std <- sqrt(prd$residual.scale^2+prd$se.fit^2)
> PIup <- prd$fit+(qt(0.975,16)*pi.std)
> PIlw <- prd$fit-(qt(0.975,16)*pi.std)
> cbind(PIlw,prd$fit,PIup)
      PIlw                PIup
1 3.007795  3.85699  4.706186
2 8.899362 10.47730 12.055240
3 5.619364  6.57982  7.540276
```

NB, while R can produce this output with less commands, you will need to understand the elements of the solution presented here

Example

10

The ANOVA table shows that once the first two predictors are included in the model, the third predictor does not add substantially to the model. So we re-fit the model and re-predict without this predictor.

```
> x1 <- HD
> x2 <- AGE*N
> pine.lm <- lm(MDBH ~ x1 + x2)
> anova(pine.lm)
Analysis of Variance Table

Response: MDBH
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 6.2070  6.2070  75.701 1.143e-07 ***
x2         1 2.7846  2.7846  33.961 2.018e-05 ***
Residuals 17 1.3939  0.0820
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example

11

```
> x1 <- c(10,80,75)
> x2 <- c(5*500,10*600,25*1000)
> prd <- predict(pine.lm,as.data.frame(cbind(x1,x2)),se.fit=T)
> CIup <- prd$fit+(qt(0.975,17)*prd$se.fit)
> CLlw <- prd$fit-(qt(0.975,17)*prd$se.fit)
> cbind(CIlw,prd$fit,CIup)
      CIlw                CIup
1 3.279259  3.855267  4.415995
2 9.027290 10.675075 11.682427
3 5.848304  6.535209  7.209772
> pi.std <- sqrt(prd$residual.scale^2+prd$se.fit^2)
> PIup <- prd$fit+(qt(0.975,17)*pi.std)
> PIlw <- prd$fit-(qt(0.975,17)*pi.std)
> cbind(PIlw,prd$fit,PIup)
      PIlw                PIup
1 3.031012  3.855267  4.679522
2 9.500453 10.675075 11.849698
3 5.629662  6.535209  7.440756
```