R-Studio Project Quan Nguyen

Consider the problem of modeling the price charged for motor transport service. In the early

1980s, several states removed regulatory constraints on the rate charged for intrastate

trucking services. (Florida was the first state to embark on a deregulation policy on July 1,

1980.) Prior to this time, the state determined price schedules for motor transport service with

review and approval by the Public Service Commission. Once approved, individual carriers

we’re not allowed to deviate from these official rates. We want to estimate a model of the

supply price for predicting future prices.

The data employed for this purpose (n = 134 observations) were obtained from a population

of over 27,000 individual shipments in Florida made by major intrastate carriers before and

after deregulation. The shipments of interest were made by one carrier whose

trucks originated from either the city of Jacksonville or Miami. The dependent variable of

interest is y, the natural logarithm of the price (measured in 1980 dollars) charged per

ton-mile.

These data are saved in the TRUCKING file.

1. DISTANCE: Miles travelled (in hundreds)

2. WEIGHT: Weight of product shipped (in 1000 pounds)

3. PCTLOAD: Percent of truck load capacity

4. ORIGIN: City of origin (JAX or MIA)

5. MARKET: Size of market destination (LARGE or SMALL)

6. DEREG: Deregulation in effect (YES or NO)

7. PRODUCT: Product classification (100, 150, or 200) – Value roughly corresponds to

the value-to-weight ratios of the goods being shipped (more valuable goods are

categorized in the higher classification)

1)

**Stepwise regression**

mydata=read.table(file.choose(),header=T)

head(mydata)

install.packages("leaps")

library(leaps)

null.model=lm(DISTANCE~1,data=mydata)

full.model<-lm(DISTANCE~WEIGHT+PCTLOAD +PRODUCT+ LNPRICE, data<-mydata )

step(null.model, scope = list(upper=full.model), data=mydata, direction="both")

Start: AIC=118.35

DISTANCE ~ 1

Df Sum of Sq RSS AIC

+ LNPRICE 1 94.804 224.49 73.143

<none> 319.29 118.348

+ WEIGHT 1 1.549 317.74 119.697

+ PCTLOAD 1 1.476 317.82 119.727

+ PRODUCT 1 0.915 318.38 119.964

Step: AIC=73.14

DISTANCE ~ LNPRICE

Df Sum of Sq RSS AIC

+ PCTLOAD 1 11.873 212.62 67.861

+ WEIGHT 1 11.812 212.68 67.900

<none> 224.49 73.143

+ PRODUCT 1 2.790 221.70 73.467

- LNPRICE 1 94.804 319.29 118.348

Step: AIC=67.86

DISTANCE ~ LNPRICE + PCTLOAD

Df Sum of Sq RSS AIC

<none> 212.62 67.861

+ PRODUCT 1 2.848 209.77 68.054

+ WEIGHT 1 0.068 212.55 69.819

- PCTLOAD 1 11.873 224.49 73.143

- LNPRICE 1 105.201 317.82 119.727

Call:

lm(formula = DISTANCE ~ LNPRICE + PCTLOAD, data = mydata)

Coefficients:

(Intercept) LNPRICE PCTLOAD

16.02533 -1.20522 -0.00899

Fitted model is

Yhat = 16.02533 - 1.20522 - 0.00899 + MARKET

-----------------------------------------------------------------------------------

**Forward selection:**

step(null.model, scope = list(upper=full.model), data=mydata, direction="forward")

Start: AIC=118.35

DISTANCE ~ 1

Df Sum of Sq RSS AIC

+ LNPRICE 1 94.804 224.49 73.143

<none> 319.29 118.348

+ WEIGHT 1 1.549 317.74 119.697

+ PCTLOAD 1 1.476 317.82 119.727

+ PRODUCT 1 0.915 318.38 119.964

Step: AIC=73.14

DISTANCE ~ LNPRICE

Df Sum of Sq RSS AIC

+ PCTLOAD 1 11.8727 212.62 67.861

+ WEIGHT 1 11.8119 212.68 67.900

<none> 224.49 73.143

+ PRODUCT 1 2.7898 221.70 73.467

Step: AIC=67.86

DISTANCE ~ LNPRICE + PCTLOAD

Df Sum of Sq RSS AIC

<none> 212.62 67.861

+ PRODUCT 1 2.84783 209.77 68.054

+ WEIGHT 1 0.06779 212.55 69.819

Call:

lm(formula = DISTANCE ~ LNPRICE + PCTLOAD, data = mydata)

Coefficients:

(Intercept) LNPRICE PCTLOAD

16.02533 -1.20522 -0.00899

Fitted model is

Yhat = 16.02533 - 1.20522 - 0.00899 + MARKET

------------------------------------------------------

**Backward elimination**

step(full.model, data=mydata.data, direction="backward")

step(null.model, scope = list(upper=full.model), data=mydata, direction="forward")

Start: AIC=118.35

DISTANCE ~ 1

Df Sum of Sq RSS AIC

+ LNPRICE 1 94.804 224.49 73.143

<none> 319.29 118.348

+ WEIGHT 1 1.549 317.74 119.697

+ PCTLOAD 1 1.476 317.82 119.727

+ PRODUCT 1 0.915 318.38 119.964

Step: AIC=73.14

DISTANCE ~ LNPRICE

Df Sum of Sq RSS AIC

+ PCTLOAD 1 11.8727 212.62 67.861

+ WEIGHT 1 11.8119 212.68 67.900

<none> 224.49 73.143

+ PRODUCT 1 2.7898 221.70 73.467

Step: AIC=67.86

DISTANCE ~ LNPRICE + PCTLOAD

Df Sum of Sq RSS AIC

<none> 212.62 67.861

+ PRODUCT 1 2.84783 209.77 68.054

+ WEIGHT 1 0.06779 212.55 69.819

Call:

lm(formula = DISTANCE ~ LNPRICE + PCTLOAD, data = mydata)

Coefficients:

(Intercept) LNPRICE PCTLOAD

16.02533 -1.20522 -0.00899

Fitted model is

Yhat = 16.02533 - 1.20522 - 0.00899 + MARKET

2)

y=mydata$DISTANCE

x1=mydata$PCTLOAD

x2=mydata$LNPRICE

Dummies

x3=ifelse(mydata$MARKET=="MEDIUM",1,0)

x4=ifelse(mydata$MARKET=="LARGE",1,0)

Where

X1 = LNPRICE

X2 = PCTLOAD

X3= if Medium if not

X4 = if Large if not

Model 1:

E(y) = B0+B1X1+B2X2+B3X3+B4X4+B5X1X2+B6X1X3+B7X1X4+B8X2X3+B9X2X4+B10X3X4+B9X1^2+B10X2^2+B11X3^2+B12X4^2

Model 2:

E(y) = B0+B1X1+B2X2+B3X3+B4X4+B5X1X2+B6X1X3+B7X1X4+B8X2X3+B9X2X4+B10X3X4

Model 3:

E(y) = B0+B1X1+B2X2+B3X3+B4X4

Model 4:

E(y) = B0+B1X1+B2X2

model1=lm(y ~ x1+x2+x3+x4+ I(x1\*x2) + I(x1\*x3) + I(x1\*x4) + I(x2\*x3) + I(x2\*x4) + I(x3\*x4) + I(x1^2)+ I(x2^2)+ I(x1^3)+ I(x1^4))

model2=lm(y~x1+x2+x3+x4+x1\*x2+x1\*x3+x1\*x4+x2\*x3+x2\*x4+x3\*x4)

model3=lm(y~x1+x2+x3+x4)

model4=lm(y~x1+x2)

summary(model1)

Call:

lm(formula = y ~ x1 + x2 + x3 + x4 + I(x1 \* x2) + I(x1 \* x3) +

I(x1 \* x4) + I(x2 \* x3) + I(x2 \* x4) + I(x3 \* x4) + I(x1^2) +

I(x2^2) + I(x1^3) + I(x1^4))

Residuals:

Min 1Q Median 3Q Max

-2.1592 -0.9126 -0.2507 0.7971 3.1052

Coefficients: (4 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.338e+00 2.490e+01 0.295 0.7687

x1 -2.233e-01 1.039e-01 -2.148 0.0337 \*

x2 7.720e-01 4.389e+00 0.176 0.8607

x3 -1.292e+00 3.691e+00 -0.350 0.7268

x4 NA NA NA NA

I(x1 \* x2) 3.684e-03 5.987e-03 0.615 0.5395

I(x1 \* x3) 1.221e-02 6.596e-03 1.851 0.0666 .

I(x1 \* x4) NA NA NA NA

I(x2 \* x3) 1.166e-01 3.400e-01 0.343 0.7322

I(x2 \* x4) NA NA NA NA

I(x3 \* x4) NA NA NA NA

I(x1^2) 7.667e-03 4.056e-03 1.890 0.0611 .

I(x2^2) -1.026e-01 1.929e-01 -0.532 0.5957

I(x1^3) -1.164e-04 6.854e-05 -1.699 0.0919 .

I(x1^4) 5.643e-07 3.546e-07 1.591 0.1141

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.235 on 123 degrees of freedom

Multiple R-squared: 0.4122, Adjusted R-squared: 0.3644

F-statistic: 8.626 on 10 and 123 DF, p-value: 1.421e-10

summary(model2)

Call:

lm(formula = y ~ x1 + x2 + x3 + x4 + x1 \* x2 + x1 \* x3 + x1 \*

x4 + x2 \* x3 + x2 \* x4 + x3 \* x4)

Residuals:

Min 1Q Median 3Q Max

-2.1368 -0.9547 -0.2805 0.8409 2.9809

Coefficients: (4 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 18.280931 2.596023 7.042 1.06e-10 \*\*\*

x1 -0.051173 0.037036 -1.382 0.1695

x2 -1.417789 0.235871 -6.011 1.82e-08 \*\*\*

x3 -2.106065 3.557617 -0.592 0.5549

x4 NA NA NA NA

x1:x2 0.003649 0.003526 1.035 0.3027

x1:x3 0.011554 0.006673 1.732 0.0858 .

x1:x4 NA NA NA NA

x2:x3 0.200135 0.326120 0.614 0.5405

x2:x4 NA NA NA NA

x3:x4 NA NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.252 on 127 degrees of freedom

Multiple R-squared: 0.3763, Adjusted R-squared: 0.3468

F-statistic: 12.77 on 6 and 127 DF, p-value: 3.015e-11

summary(model3)

Call:

lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:

Min 1Q Median 3Q Max

-2.0155 -0.8503 -0.2174 0.7774 3.4300

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 15.357946 1.657023 9.268 5.23e-16 \*\*\*

x1 -0.008291 0.003298 -2.514 0.0131 \*

x2 -1.163772 0.149070 -7.807 1.71e-12 \*\*\*

x3 0.466860 0.221431 2.108 0.0369 \*

x4 NA NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.258 on 130 degrees of freedom

Multiple R-squared: 0.3561, Adjusted R-squared: 0.3413

F-statistic: 23.97 on 3 and 130 DF, p-value: 2.067e-12

summary(model4)

Call:

lm(formula = y ~ x1 + x2)

Residuals:

Min 1Q Median 3Q Max

-2.1757 -0.9573 -0.2064 0.7640 3.7418

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 16.025334 1.647757 9.726 < 2e-16 \*\*\*

x1 -0.008990 0.003324 -2.705 0.00775 \*\*

x2 -1.205215 0.149699 -8.051 4.38e-13 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.274 on 131 degrees of freedom

Multiple R-squared: 0.3341, Adjusted R-squared: 0.3239

F-statistic: 32.86 on 2 and 131 DF, p-value: 2.711e-12

3)

Model 1 vs Model 2

Model 1: y ~ x1 + x2 + x3 + x4 + I(x1 \* x2) + I(x1 \* x3) + I(x1 \* x4) +

I(x2 \* x3) + I(x2 \* x4) + I(x3 \* x4) + I(x1^2) + I(x2^2) +

I(x1^3) + I(x1^4)

Model 2: y ~ x1 + x2 + x3 + x4 + x1 \* x2 + x1 \* x3 + x1 \* x4 + x2 \* x3 +

x2 \* x4 + x3 \* x4

Res.Df RSS Df Sum of Sq F Pr(>F)

1 123 187.68

2 127 199.15 -4 -11.466 1.8786 0.1184

We see that Test statistic =1.87

P-value = 0.1184

Since 0.1184> Alpha = 0.05, we fail to reject the null hypothesis. This indicates that removing the quadratic terms did not contribute to the prediction of Y

Model 1 is superior to Model 2. We select Model 1

Model 1 vs Model 3

anova(model1,model3)

Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x4 + I(x1 \* x2) + I(x1 \* x3) + I(x1 \* x4) +

I(x2 \* x3) + I(x2 \* x4) + I(x3 \* x4) + I(x1^2) + I(x2^2) +

I(x1^3) + I(x1^4)

Model 2: y ~ x1 + x2 + x3 + x4

Res.Df RSS Df Sum of Sq F Pr(>F)

1 123 187.68

2 130 205.59 -7 -17.908 1.6766 0.1208

We see that Test statistic =1.6766

P-value = 0.1208

Since 0.1208 > Alpha = 0.05, we fail to reject the null hypothesis. This indicates that removing the quadratic terms and interaction terms did not contribute to the prediction of Y

Model 1 is superior to the Model 3. We select Model 1

Model 1 vs Model 4

anova(model1,model4)

Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x4 + I(x1 \* x2) + I(x1 \* x3) + I(x1 \* x4) +

I(x2 \* x3) + I(x2 \* x4) + I(x3 \* x4) + I(x1^2) + I(x2^2) +

I(x1^3) + I(x1^4)

Model 2: y ~ x1 + x2

Res.Df RSS Df Sum of Sq F Pr(>F)

1 123 187.68

2 131 212.62 -8 -24.938 2.0429 0.04668 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

We see that Test statistic =2.0429

P-value = 0.04668

Since 0.04668 <Alpha = 0.05, we reject the null hypothesis. This indicates that removing the quadratic terms and interaction terms and qualitative did not contribute to the prediction of Y

Model 1 is superior to the Model 4. We select Model 1

We select model 1

The result of the preceding tests suggests model 1 is the best model.

**Adequacy checking for model 1**

lm(formula = y ~ x1 + x2 + x3 + x4 + I(x1 \* x2) + I(x1 \* x3) +

I(x1 \* x4) + I(x2 \* x3) + I(x2 \* x4) + I(x3 \* x4) + I(x1^2) +

I(x2^2) + I(x1^3) + I(x1^4))

Residuals:

Min 1Q Median 3Q Max

-2.1592 -0.9126 -0.2507 0.7971 3.1052

Coefficients: (4 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.338e+00 2.490e+01 0.295 0.7687

x1 -2.233e-01 1.039e-01 -2.148 0.0337 \*

x2 7.720e-01 4.389e+00 0.176 0.8607

x3 -1.292e+00 3.691e+00 -0.350 0.7268

x4 NA NA NA NA

I(x1 \* x2) 3.684e-03 5.987e-03 0.615 0.5395

I(x1 \* x3) 1.221e-02 6.596e-03 1.851 0.0666 .

I(x1 \* x4) NA NA NA NA

I(x2 \* x3) 1.166e-01 3.400e-01 0.343 0.7322

I(x2 \* x4) NA NA NA NA

I(x3 \* x4) NA NA NA NA

I(x1^2) 7.667e-03 4.056e-03 1.890 0.0611 .

I(x2^2) -1.026e-01 1.929e-01 -0.532 0.5957

I(x1^3) -1.164e-04 6.854e-05 -1.699 0.0919 .

I(x1^4) 5.643e-07 3.546e-07 1.591 0.1141

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.235 on 123 degrees of freedom

Multiple R-squared: 0.4122, Adjusted R-squared: 0.3644

F-statistic: 8.626 on 10 and 123 DF, p-value: 1.421e-10

**Global F tests**

H0 = B1 = B2 = B3 =B4 = ... B12 = 0

Ha: At least on Bj not equal 0 for j = 1,2,3,4,...,12

Test statistic, F = 8.626

P-value < 1.421e-10

We reject the null hypothesis and conclude that model is highly significant.

R^2adj = 0.3644. The adjusted R-squared value indicates that the model explains approximately 36.44% of the variability in sale

**The standard error of model is 1.235 which is very low.**

**The model is good because the standard error is really low.**