> mdata <- read.csv("UN1.csv", header=TRUE)

> mdata

 Locality Fertility PPgdp

1 Afghanistan 6.80 98

2 Albania 2.28 1317

3 Algeria 2.80 1784

4 Angola 7.20 739

5 Argentina 2.44 7163

6 Armenia 1.15 687

7 Australia 1.70 18788

8 Austria 1.28 23260

9 Azerbaijan 2.10 695

10 Bahamas 2.29 14856

11 Bahrain 2.66 12012

12 ……….

>

> plot(mdata$PPgdp,mdata$Fertility, main="Scatterplot Example")



> fit <- lm(mdata$Fertility~mdata$PPgdp)

> summary(fit)

Call:

lm(formula = mdata$Fertility ~ mdata$PPgdp)

Residuals:

 Min 1Q Median 3Q Max

-2.5250 -1.1013 -0.1826 0.9240 4.2817

Coefficients:

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

(Intercept) 3.733e+00 1.331e-01 28.040 < 2e-16 \*\*\*

mdata$PPgdp -8.486e-05 1.174e-05 -7.226 1.15e-11 \*\*\*

---

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

Residual standard error: 1.526 on 191 degrees of freedom

Multiple R-squared: 0.2147, Adjusted R-squared: 0.2106

F-statistic: 52.22 on 1 and 191 DF, p-value: 1.155e-11

>

> plot(mdata$PPgdp,log(mdata$Fertility), main="Scatterplot Example")



> fit1 <- lm(log(mdata$Fertility)~mdata$PPgdp)

> summary(fit1)

Call:

lm(formula = log(mdata$Fertility) ~ mdata$PPgdp)

Residuals:

 Min 1Q Median 3Q Max

-1.05408 -0.28530 0.03513 0.36189 0.89037

Coefficients:

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

(Intercept) 1.194e+00 4.126e-02 28.935 < 2e-16 \*\*\*

mdata$PPgdp -2.745e-05 3.639e-06 -7.542 1.82e-12 \*\*\*

---

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

Residual standard error: 0.4729 on 191 degrees of freedom

Multiple R-squared: 0.2295, Adjusted R-squared: 0.2254

F-statistic: 56.88 on 1 and 191 DF, p-value: 1.824e-12

> plot(log(mdata$PPgdp),log(mdata$Fertility), main="Scatterplot Example")



> fit2 <- lm(log(mdata$Fertility)~log(mdata$PPgdp))

> summary(fit2)

Call:

lm(formula = log(mdata$Fertility) ~ log(mdata$PPgdp))

Residuals:

 Min 1Q Median 3Q Max

-1.11877 -0.18762 0.07041 0.26082 0.90101

Coefficients:

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

(Intercept) 2.70322 0.13538 19.97 <2e-16 \*\*\*

log(mdata$PPgdp) -0.22116 0.01737 -12.73 <2e-16 \*\*\*

---

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

Residual standard error: 0.3962 on 191 degrees of freedom

Multiple R-squared: 0.4591, Adjusted R-squared: 0.4563

F-statistic: 162.1 on 1 and 191 DF, p-value: < 2.2e-16

>

> #Plot1: If you find equally spread residuals around a horizontal line without distinct patterns,

> #that is a good indication you don’t have non-linear relationships.

> #Plot2:shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely?

> #Plot3: check the assumption of equal variance (homoscedasticity). It’s good if you see a horizontal line with equally (randomly) spread points

> Plot4: This plot helps us to find influential cases .Look for cases far beyond the Cook’s distance lines

>

>

> par(mfrow=c(2,2)) # Change the panel layout to 2 x 2

> plot(fit2)

> par(mfrow=c(1,1)) # Change back to 1 x 1



> mdata1 <- read.csv("UN2.csv", header=TRUE)

> mdata1

 logPPgdp logFertility Purban Locality

1 6.614710 1.91692261 22 Afghanistan

2 10.363040 0.82417544 43 Albania

3 10.800900 1.02961942 58 Algeria

4 9.529431 1.97408103 35 Angola

5 12.806348 0.89199804 88 Argentina

6 9.424166 0.13976194 67 Armenia

7 14.197524 0.53062825 91 Australia

8 14.505563 0.24686008 67 Austria

9 9.440869 0.74193734 52 Azerbaijan

> fit3 <- lm(mdata1$logFertility~mdata1$logPPgdp+mdata1$Purban)

> summary(fit3)

Call:

lm(formula = mdata1$logFertility ~ mdata1$logPPgdp + mdata1$Purban)

Residuals:

 Min 1Q Median 3Q Max

-1.05378 -0.16952 0.06835 0.25587 0.89290

Coefficients:

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

(Intercept) 2.592996 0.146864 17.656 < 2e-16 \*\*\*

mdata1$logPPgdp -0.125475 0.019095 -6.571 4.67e-10 \*\*\*

mdata1$Purban -0.003522 0.001884 -1.869 0.0631 .

---

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

Residual standard error: 0.3936 on 190 degrees of freedom

Multiple R-squared: 0.4689, Adjusted R-squared: 0.4633

F-statistic: 83.88 on 2 and 190 DF, p-value: < 2.2e-16

> par(mfrow=c(2,2)) # Change the panel layout to 2 x 2

> plot(fit3)

> par(mfrow=c(1,1)) # Change back to 1 x 1

>



> #EXEMPLUL2

>

> mmm <- read.csv("travel.csv", header=TRUE)

> mmm

 Amount Age Segment C

1 997 44 A 0

2 997 43 A 0

3 951 41 A 0

4 649 59 A 0

5 1265 25 A 0

6 1059 38 A 0

7 837 46 A 0

8 924 42 A 0

9 852 48 A 0

10 963 39 A 0

[ reached getOption("max.print") -- omitted 675 rows ]

> mod<-lm(mmm$Amount~mmm$Age\*mmm$C)

> summary(mod)

Call:

lm(formula = mmm$Amount ~ mmm$Age \* mmm$C)

Residuals:

 Min 1Q Median 3Q Max

-143.298 -30.541 -0.034 31.108 130.743

Coefficients:

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

(Intercept) 1814.5445 8.6011 211.0 <2e-16 \*\*\*

mmm$Age -20.3175 0.1878 -108.2 <2e-16 \*\*\*

mmm$C -1821.2337 12.5736 -144.8 <2e-16 \*\*\*

mmm$Age:mmm$C 40.4461 0.2724 148.5 <2e-16 \*\*\*

---

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

Residual standard error: 47.63 on 921 degrees of freedom

Multiple R-squared: 0.9601, Adjusted R-squared: 0.9599

F-statistic: 7379 on 3 and 921 DF, p-value: < 2.2e-16

>

>

> mod1<-lm(mmm$Amount~mmm$Age)

> summary(mod1)

Call:

lm(formula = mmm$Amount ~ mmm$Age)

Residuals:

 Min 1Q Median 3Q Max

-545.06 -199.03 6.34 198.74 497.39

Coefficients:

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

(Intercept) 957.9103 31.3056 30.599 <2e-16 \*\*\*

mmm$Age -1.1140 0.6784 -1.642 0.101

---

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

Residual standard error: 237.7 on 923 degrees of freedom

Multiple R-squared: 0.002913, Adjusted R-squared: 0.001833

F-statistic: 2.697 on 1 and 923 DF, p-value: 0.1009

>

> anova(mod1, mod)

Analysis of Variance Table

Model 1: mmm$Amount ~ mmm$Age

Model 2: mmm$Amount ~ mmm$Age \* mmm$C

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

1 923 52158945

2 921 2089377 2 50069568 11035 < 2.2e-16 \*\*\*

---

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

>

> #As expected there is very strong evidence against the reduced model in #favour of the full model.

**STEPWISE REGRESSION**

> ndata <- read.csv("bridge.csv", header=TRUE)

> ndata

 Case Time DArea CCost Dwgs Length Spans

1 1 78.8 3.60 82.4 6 90 1

2 2 309.5 5.33 422.3 12 126 2

3 3 184.5 6.29 179.8 9 78 1

4 4 69.6 2.20 100.0 5 60 1

5 5 68.8 1.44 103.0 5 60 1

6 6 95.7 5.40 134.4 5 60 1

7 7 112.3 6.60 173.2 5 180 3

8 8 171.9 7.90 207.9 7 188 2

9 9 177.8 4.88 327.7 9 336 2

10 10 65.8 0.85 56.2 3 25 1

11 11 51.2 1.45 46.8 3 50 1

12 12 59.6 4.10 118.9 6 114 2

> plot(ndata)



The response variable and a number of the predictor variables are highly skewed. There is also evidence of nonconstant variance in the top row of plots. Thus, we need to consider transformations of the response and the five predictor variables.

Using the Box-Cox method to transform the predictor and response variables simultaneously toward multivariate normality, results in values of each *l* close to 0. Thus we shall transform each variable using the log transformation.

> summary(powerTransform(ndata))

bcPower Transformations to Multinormality

 Est.Power Std.Err. Wald Lower Bound Wald Upper Bound

Time -0.1795 0.2001 -0.5716 0.2127

DArea -0.1346 0.0893 -0.3096 0.0404

CCost -0.1762 0.0942 -0.3609 0.0085

Dwgs -0.2507 0.2402 -0.7215 0.2200

Length -0.1975 0.1073 -0.4078 0.0127

Spans -0.3744 0.2594 -0.8828 0.1340

Likelihood ratio tests about transformation parameters

 LRT df pval

LR test, lambda = (0 0 0 0 0 0) 8.121991 6 0.2293015

LR test, lambda = (1 1 1 1 1 1) 283.184024 6 0.0000000

> nndata<-log(ndata)

> plot(nndata)



> par(mfrow=c(2,2)) # Change the panel layout to 2 x 2

> plot(a1)

> par(mfrow=c(1,1)) # Change back to 1 x 1

There is a bad leverage point (i.e., case 22) that requires further investigation.



> a1<-lm(Time~DArea+CCost+Dwgs+Length+Spans, data=nndata)

>

> summary(a1)

Call:

lm(formula = Time ~ DArea + CCost + Dwgs + Length + Spans, data = nndata)

Residuals:

 Min 1Q Median 3Q Max

-0.68394 -0.17167 -0.02604 0.23157 0.67307

Coefficients:

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

(Intercept) 2.28590 0.61926 3.691 0.000681 \*\*\*

DArea -0.04564 0.12675 -0.360 0.720705

CCost 0.19609 0.14445 1.358 0.182426

Dwgs 0.85879 0.22362 3.840 0.000440 \*\*\*

Length -0.03844 0.15487 -0.248 0.805296

Spans 0.23119 0.14068 1.643 0.108349

---

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

Residual standard error: 0.3139 on 39 degrees of freedom

Multiple R-squared: 0.7762, Adjusted R-squared: 0.7475

F-statistic: 27.05 on 5 and 39 DF, p-value: 1.043e-11

Notice that while the overall F-test for model (6.28) is highly statistically significant (i.e., has a very small *p*-value), only one of the estimated regression coefficients is statistically significant (i.e., Dwgs with a *p*-value < 0.001).

Even more troubling is the fact that the estimated regression coefficients for log(DArea) and log(Length) are of the wrong sign (i.e., negative), since longer bridges or bridges with larger area should take a longer rather than a shorter time to design.

> cor(nndata)

 Time DArea CCost Dwgs Length Spans

Time 1.0000000 0.7742171 0.8223730 0.8420257 0.7635266 0.7128162

DArea 0.7742171 1.0000000 0.9092401 0.8011948 0.8841942 0.7815166

CCost 0.8223730 0.9092401 1.0000000 0.8314571 0.8905000 0.7750899

Dwgs 0.8420257 0.8011948 0.8314571 1.0000000 0.7523439 0.6297212

Length 0.7635266 0.8841942 0.8905000 0.7523439 1.0000000 0.8584858

Spans 0.7128162 0.7815166 0.7750899 0.6297212 0.8584858 1.0000000

When two or more highly correlated predictor variables are included in a regression model, they are effectively carrying very similar information about the response variable.

In this situation the overall F-test will be highly statistically significant but very few of the regression coefficients may be statistically significant. Another consequence of highly correlated predictor

variables is that some of the coefficients in the regression model are of the opposite sign than expected.

> vif(a1) ## Multicoliniarity

 DArea CCost Dwgs Length Spans

7.164619 8.483522 3.408900 8.014174 3.878397

A number of these variance inflation factors exceed 5, the cut-off often used, and so the associated regression coefficients are poorly estimated due to multicollinearity.

GOAL: Model *Y* = Time = design time in person-days

Notice above that while the overall F-test for model is highly statistically significant, only one of the estimated regression coefficients is statistically significant (i.e., log(Dwgs) with a p -value < 0.001). Thus, we wish to choose a subset of the predictors using variable selection.

We begin our discussion of variable selection in this example by identifying the subset of the predictors of a given size that maximizes adjusted R-squared.

**Determine which of the five explanatory variables should be included in the regression model using the all possible regressions approach.**

> require(Rcmdr)

> a.glm <- glm(Time ~., data = nndata)

> stepwise(a.glm, trace = T) #Default is BIC

Direction: backward/forward

Criterion: BIC

Start: AIC=41.83

Time ~ DArea + CCost + Dwgs + Length + Spans

 Df Deviance AIC

- Length 1 3.8497 38.097

- DArea 1 3.8564 38.176

- CCost 1 4.0252 40.104

- Spans 1 4.1098 41.039

<none> 3.8436 41.833

- Dwgs 1 5.2972 52.461

Step: AIC=38.1

Time ~ DArea + CCost + Dwgs + Spans

 Df Deviance AIC

- DArea 1 3.8693 34.519

- CCost 1 4.0303 36.354

- Spans 1 4.1647 37.830

<none> 3.8497 38.097

+ Length 1 3.8436 41.833

- Dwgs 1 5.2991 48.671

Step: AIC=34.52

Time ~ CCost + Dwgs + Spans

 Df Deviance AIC

- CCost 1 4.0488 32.754

- Spans 1 4.1658 34.036

<none> 3.8693 34.519

+ DArea 1 3.8497 38.097

+ Length 1 3.8564 38.176

- Dwgs 1 5.3147 44.996

Step: AIC=32.75

Time ~ Dwgs + Spans

 Df Deviance AIC

<none> 4.0488 32.754

+ CCost 1 3.8693 34.519

+ DArea 1 4.0303 36.354

+ Length 1 4.0319 36.372

- Spans 1 4.9975 38.420

- Dwgs 1 8.4478 62.044

Call: glm(formula = Time ~ Dwgs + Spans, data = nndata)

Coefficients:

(Intercept) Dwgs Spans

 2.6617 1.0416 0.2853

Degrees of Freedom: 44 Total (i.e. Null); 42 Residual

Null Deviance: 17.17

Residual Deviance: 4.049 AIC: 27.33

> require(MASS)

> step <- stepAIC(a.glm, direction="backward")

Start: AIC=30.99

Time ~ DArea + CCost + Dwgs + Length + Spans

 Df Deviance AIC

- Length 1 3.8497 29.064

- DArea 1 3.8564 29.142

<none> 3.8436 30.993

- CCost 1 4.0252 31.071

- Spans 1 4.1098 32.006

- Dwgs 1 5.2972 43.428

Step: AIC=29.06

Time ~ DArea + CCost + Dwgs + Spans

 Df Deviance AIC

- DArea 1 3.8693 27.292

<none> 3.8497 29.064

- CCost 1 4.0303 29.128

- Spans 1 4.1647 30.603

- Dwgs 1 5.2991 41.444

Step: AIC=27.29

Time ~ CCost + Dwgs + Spans

 Df Deviance AIC

<none> 3.8693 27.292

- CCost 1 4.0488 27.334

- Spans 1 4.1658 28.616

- Dwgs 1 5.3147 39.576

> a.glm <- glm(Time ~., data = nndata)

> b.glm <- glm(Time ~1, data = nndata)

> step(b.glm, scope=list(lower=b.glm, upper=a.glm), direction="forward")

Start: AIC=88.36

Time ~ 1

 Df Deviance AIC

+ Dwgs 1 4.9975 34.807

+ CCost 1 5.5593 39.601

+ DArea 1 6.8797 49.191

+ Length 1 7.1620 51.000

+ Spans 1 8.4478 58.430

<none> 17.1740 88.358

Step: AIC=34.81

Time ~ Dwgs

 Df Deviance AIC

+ Spans 1 4.0488 27.334

+ CCost 1 4.1658 28.616

+ Length 1 4.3284 30.338

+ DArea 1 4.5218 32.306

<none> 4.9975 34.807

Step: AIC=27.33

Time ~ Dwgs + Spans

 Df Deviance AIC

+ CCost 1 3.8693 27.292

<none> 4.0488 27.334

+ DArea 1 4.0303 29.128

+ Length 1 4.0319 29.146

Step: AIC=27.29

Time ~ Dwgs + Spans + CCost

 Df Deviance AIC

<none> 3.8693 27.292

+ DArea 1 3.8497 29.064

+ Length 1 3.8564 29.142

Call: glm(formula = Time ~ Dwgs + Spans + CCost, data = nndata)

Coefficients:

(Intercept) Dwgs Spans CCost

 2.3317 0.8356 0.1963 0.1483

Degrees of Freedom: 44 Total (i.e. Null); 41 Residual

Null Deviance: 17.17

Residual Deviance: 3.869 AIC: 27.29

> step <- stepAIC(a.glm, direction="both")

Start: AIC=30.99

Time ~ DArea + CCost + Dwgs + Length + Spans

 Df Deviance AIC

- Length 1 3.8497 29.064

- DArea 1 3.8564 29.142

<none> 3.8436 30.993

- CCost 1 4.0252 31.071

- Spans 1 4.1098 32.006

- Dwgs 1 5.2972 43.428

Step: AIC=29.06

Time ~ DArea + CCost + Dwgs + Spans

 Df Deviance AIC

- DArea 1 3.8693 27.292

<none> 3.8497 29.064

- CCost 1 4.0303 29.128

- Spans 1 4.1647 30.603

+ Length 1 3.8436 30.993

- Dwgs 1 5.2991 41.444

Step: AIC=27.29

Time ~ CCost + Dwgs + Spans

 Df Deviance AIC

<none> 3.8693 27.292

- CCost 1 4.0488 27.334

- Spans 1 4.1658 28.616

+ DArea 1 3.8497 29.064

+ Length 1 3.8564 29.142

- Dwgs 1 5.3147 39.576

> library(leaps)

> leaps=regsubsets(Time ~CCost+DArea+Length+Spans+Dwgs,data=nndata, nbest=5)

>

> #To view the ranked models according to the adjusted R- squared criteria and BIC,

>

> plot(leaps, scale="adjr2")

Black indicates that a variable is included in the model, while white indicates that they are not.



**LOGISTIC**

> ndata <- read.csv("Simmons.csv", header=TRUE)

> ndata

> mylogit <- glm(Coupon ~ Spending + Card, data = ndata, family = "binomial")

> summary(mylogit)

Call:

glm(formula = Coupon ~ Spending + Card, family = "binomial",

 data = ndata)

Deviance Residuals:

 Min 1Q Median 3Q Max

-1.6839 -1.0140 -0.6503 1.1216 1.8794

Coefficients:

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

(Intercept) -2.1464 0.5772 -3.718 0.000201 \*\*\*

Spending 0.3416 0.1287 2.655 0.007928 \*\*

Card 1.0987 0.4447 2.471 0.013483 \*

---

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

(Dispersion parameter for binomial family taken to be 1)

 Null deviance: 134.60 on 99 degrees of freedom

Residual deviance: 120.97 on 97 degrees of freedom

AIC: 126.97

Number of Fisher Scoring iterations: 4

> confint(mylogit) #Confidence Intervals for the coefficients estimates

Waiting for profiling to be done...

 2.5 % 97.5 %

(Intercept) -3.35463431 -1.0727051

Spending 0.09640766 0.6049657

Card 0.24476776 1.9982400

> exp(coef(mylogit)) #Estimated odds ratio

(Intercept) Spending Card

 0.1169074 1.4072585 3.0003587

|  |
| --- |
| > anova(mylogit, test="Chisq")Analysis of Deviance TableModel: binomial, link: logitResponse: CouponTerms added sequentially (first to last) Df Deviance Resid. Df Resid. Dev Pr(>Chi) NULL 99 134.60 Spending 1 7.2182 98 127.38 0.007217 \*\*Card 1 6.4103 97 120.97 0.011346 \* ---Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
|

|  |
| --- |
| > #Create  |

 |

> newdata = data.frame(Spending = S, Card = cc)

> newdata

 Spending Card

1 1 1

2 2 1

3 3 1

4 4 1

5 5 1

6 6 1

7 7 1

8 1 0

9 2 0

10 3 0

11 4 0

12 5 0

13 6 0

14 7 0

> predict(mylogit, newdata, type="response")

 1 2 3 4 5 6 7 8 9 10

0.3304837 0.4099058 0.4943225 0.5790640 0.6593897 0.7314947 0.7931244 0.1412763 0.1879957 0.2457437

 11 12 13 14

0.3143631 0.3921804 0.4758906 0.5609777

|  |
| --- |
| > a<-predict(mylogit, newdata, type="response")> > #Table of the estimated probabilities> counts <- data.frame( AmountSpent=c("1", "2", "3", "4", "5", "6", "7", "1", "2", "3", "4", "5", "6", "7"),  CC\_Yes\_No =c("1", "1", "1", "1","1", "1","1", "0", "0", "0", "0","0", "0","0"), Probabability=a )> > counts AmountSpent CC\_Yes\_No Probabability1 1 1 0.33048372 2 1 0.40990583 3 1 0.49432254 4 1 0.57906405 5 1 0.65938976 6 1 0.73149477 7 1 0.79312448 1 0 0.14127639 2 0 0.187995710 3 0 0.245743711 4 0 0.314363112 5 0 0.392180413 6 0 0.475890614 7 0 0.5609777 |
|  |
|

|  |
| --- |
| >  |

 |