**37252 Regression and Linear Models**

**Lab 4: Multiple Linear Regression II**

This lab is marked out of 15.

Please save your file in PDF format with name

**37252\_Lab4\_Surname\_FirstName**

**Due: 12 noon Wednesday 3 April 2024**

In this week’s lab we look at automated variable selection techniques. The data are company sales figures of 25 sales persons, each responsible for a sales district, and available in **37252\_Lab4\_data.csv** which can be downloaded from Canvas.

The variables we now consider are summarised in the table below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Description** |
| $$sales$$ | response | unit district sales |
| $$time$$ | predictor | tenure of salesperson |
| $$potent$$ | predictor | unit industry sales |
| $$adv$$ | predictor | advertising expenditure |
| $$share$$ | predictor | market share last 4 years |
| $$sharechg$$ | predictor | change in market share over previous 4 years |
| $$accts$$ | predictor | number of accounts assigned to sales person |
| $$workload$$ | predictor | average workload per account |
| $$rating$$ | predictor | manager’s rating of sales person (1,2,…,7) |

We start by inspecting the relationships between the variables.

> salesdat <- read.csv("~/2024\_37252/Labs/Lab4/37252\_Lab4\_data.csv")

> pairs(~ ., data = salesdat)



We will also look at the correlation between variables.

> library('Hmisc')

> sales\_cor <- as.matrix(cbind(salesdat$sales, salesdat$time, salesdat$potent, salesdat$adv, salesdat$share, salesdat$sharechg, salesdat$accts, salesdat$workload, salesdat$rating))

> colnames(sales\_cor)<-c("sales", "time", "potent", "adv", "share", "sharechg", "accts", "workload", "rating")

> sales\_cor\_res<-rcorr(sales\_cor, type="pearson")

> round(sales\_cor\_res$r,3)

 sales time potent adv share sharechg accts workload rating

sales 1.000 0.623 0.598 0.596 0.484 0.489 0.754 -0.117 0.402

time 0.623 1.000 0.454 0.249 0.106 0.251 0.758 -0.179 0.101

potent 0.598 0.454 1.000 0.174 -0.211 0.268 0.479 -0.259 0.359

adv 0.596 0.249 0.174 1.000 0.264 0.377 0.200 -0.272 0.411

share 0.484 0.106 -0.211 0.264 1.000 0.085 0.403 0.349 -0.024

sharechg 0.489 0.251 0.268 0.377 0.085 1.000 0.327 -0.288 0.549

accts 0.754 0.758 0.479 0.200 0.403 0.327 1.000 -0.199 0.229

workload -0.117 -0.179 -0.259 -0.272 0.349 -0.288 -0.199 1.000 -0.277

rating 0.402 0.101 0.359 0.411 -0.024 0.549 0.229 -0.277 1.000

> round(sales\_cor\_res$P,3)

 sales time potent adv share sharechg accts workload rating

sales NA 0.001 0.002 0.002 0.014 0.013 0.000 0.577 0.046

time 0.001 NA 0.023 0.230 0.613 0.225 0.000 0.391 0.631

potent 0.002 0.023 NA 0.405 0.312 0.195 0.016 0.212 0.078

adv 0.002 0.230 0.405 NA 0.201 0.064 0.338 0.188 0.041

share 0.014 0.613 0.312 0.201 NA 0.685 0.046 0.087 0.911

sharechg 0.013 0.225 0.195 0.064 0.685 NA 0.110 0.163 0.004

accts 0.000 0.000 0.016 0.338 0.046 0.110 NA 0.341 0.272

workload 0.577 0.391 0.212 0.188 0.087 0.163 0.341 NA 0.180

rating 0.046 0.631 0.078 0.041 0.911 0.004 0.272 0.180 NA

1. Based on correlations, chose two predictors as candidate predictors for a multiple regression **[2 marks]**. Why have you chosen these **[2 marks]**?

Predictors selected for model: $acts$ and $adv$ **[2 marks]**.

These variables have medium-strong correlation with$sales$**[1 mark]***,* but have quite weak correlation with each other **[1 mark]**.

We will now run the forward selection technique.

1. Write down the model selected at each step of the forward procedure **[4 marks]**.

> library('olsrr')

> mod1 <- lm(sales~., data = salesdat)

> ols\_step\_forward\_p(mod1, p\_val = 0.1, details = T)

Step 1. $\hat{sales}=709.324+21.722×accts$

**[1 mark]**

Step 2. $\hat{sales}=50.287+19.048×accts+0.277×adv$

**[1 mark]**

Step 3. $\hat{sales}=-327.240+15.554×accts+0.216×adv+0.022×potent$

**[1 mark]**

Step 4. $\hat{sales}=-1441.938+9.214×accts+0.175×adv+0.038×potent$

$$+190.144×share$$

**[1 mark]**

**(c)**Write down the model selected at each step of the backward procedure **[4 marks]**.

> ols\_step\_backward\_p(mod1, p\_val = 0.11, details=T)

Step 1. $\hat{sales}=-1507.834+2.010×time+0.037×potent+0.151×adv$

$$+199.023×share+290.853×sharechg$$

$$+5.551×accts+19.794×workload$$

$$+8.190×rating$$

**[1 mark]**

Step 2. $\hat{sales}=-1485.898+1.974×time+0.037×potent+0.152×adv$

$$+198.307×share+295.864×sharechg$$

$$+5.610×accts+19.900×workload$$

**[1 mark]**

Step 3. $\hat{sales}=-1165.486+2.269×time+0.038×potent+0.141×adv$

$$+221.604×share+285.107×sharechg$$

$$+4.378×accts$$

**[1 mark]**

Step 4. $\hat{sales}=-1113.793+3.612×time+0.042×potent+0.129×adv$

$$+256.956×share+324.533×sharechg$$

**[1 mark]**

We could also run the stepwise procedure but, in this case, this is results in identical steps as the forward procedure.

**(d)**Describe one way in which the final model of the forward procedure is superior to that of the backward model **[1 mark]**.

The forward model is superior because residuals appears closer to normal **[1 mark]**. Also, the forward model does not have a predictor with beta coefficient p-value greater than 0.05.

Forward:



Backward:



**(e)**Describe two ways in which the final model of the backward procedure is superior to that of the forward model **[2 marks]**.

The backward model is superior because it has higher adjusted R-squared (0.893 versus 0.881) **[1 mark]**. It also has a DW stat closer to 2 (1.762 versus 1.596) **[1 mark]**.

> durbinWatsonTest(mod\_forward)

 lag Autocorrelation D-W Statistic p-value

 1 0.1855081 1.596162 0.208

 Alternative hypothesis: rho != 0

> durbinWatsonTest(mod\_backward)

 lag Autocorrelation D-W Statistic p-value

 1 0.1014294 1.761861 0.338

 Alternative hypothesis: rho != 0