hwk\_601\_1

Dingxian Cao

September 30, 2015

# ex2

take k=9074463432 mod 35

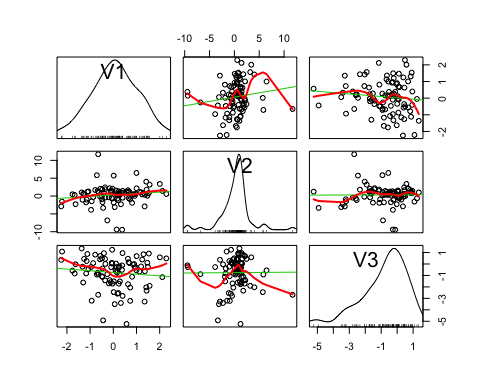
options(digits = 4)  
setwd("/Users/CDX/WISC\_R\_HWK/Regression")  
suppressMessages(library(broom))  
ext2<-read.table("HWK2\_Ext2.txt")  
k<-9074463432 %% 35  
d<-ext2[((k-1)\*90+1):(k\*90),]  
dim(d)

## [1] 90 3

d1<-d[1:30,]  
d2<-d[31:60,]  
d3<-d[61:90,]  
  
knitr::kable(tidy(cor(d)))

|  |  |  |  |
| --- | --- | --- | --- |
| .rownames | V1 | V2 | V3 |
| V1 | 1.0000 | 0.1448 | -0.1085 |
| V2 | 0.1448 | 1.0000 | 0.0067 |
| V3 | -0.1085 | 0.0067 | 1.0000 |

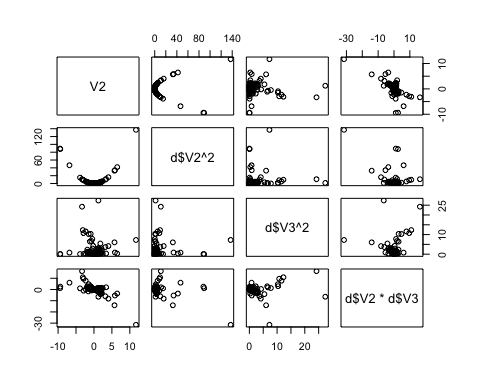
library(car)  
scatterplotMatrix(d,spread=F)



From the scatter plot and the covariance matrix we can see that there are no obvious relationship between V1 and V3, so there shouldn't be a response variable between the two.

But they both have a clear relationship with V2.So,I believe there is V2 is the reponse variable.

plot(cbind(d,d$V2^2,d$V3^2,d$V2\*d$V3)[-c(1,3)])



Since we don't know the true model, we cannot use the model to identify the explantory variables. So I first assume X1=V1.

### model 1

options(digits = 4)  
d11<-as.data.frame(d1)  
names(d11)<-c("x1","y","x2")  
form1<- as.formula("y ~ 1+x1+x1:x2")  
m1<-lm(form1,data=d11);knitr::kable(tidy(m1))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 0.4117 | 0.7985 | 0.5156 | 0.6104 |
| x1 | 0.2452 | 0.8509 | 0.2882 | 0.7754 |
| x1:x2 | 0.4394 | 0.7492 | 0.5866 | 0.5624 |

### model2

d21<-as.data.frame(d2)  
names(d21)<-c("x1","y","x2")  
form2<-as.formula("1/y ~ 1+x1+x2")  
m2<-lm(form2,data=d21);knitr::kable(tidy(m2))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -1.091 | 1.195 | -0.9133 | 0.3692 |
| x1 | 1.155 | 1.046 | 1.1048 | 0.2790 |
| x2 | -1.004 | 1.055 | -0.9521 | 0.3495 |

### model3

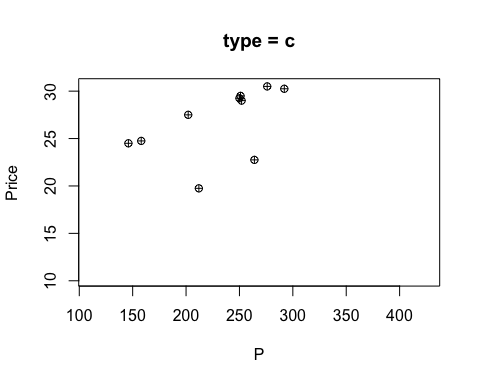
d31<-as.data.frame(d3)  
names(d31)<-c("x1","y","x2")  
form3<-as.formula("y ~ 0+x1+x2+I(x1^2)")  
m3<-lm(form3,data=d31);knitr::kable(tidy(m3))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| x1 | 0.0668 | 0.1938 | 0.3449 | 0.7328 |
| x2 | 0.6623 | 0.1008 | 6.5726 | 0.0000 |
| I(x1^2) | 0.3096 | 0.1221 | 2.5345 | 0.0174 |

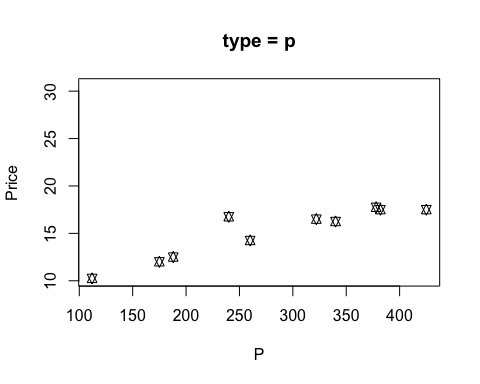
According to the P value of each coefficients, the third model is obviously the best fitted.

# ex3

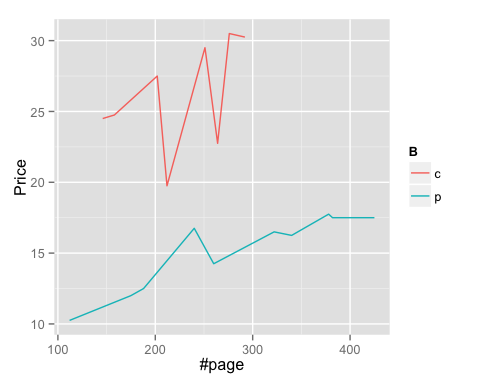
#install.packages("SenSrivastava")  
library(SenSrivastava)  
suppressMessages(library(dplyr))  
y\_range<-range(E1.19$Price)  
x\_range<-range(E1.19$P)  
group<-split(E1.19,E1.19$B)  
par(mfrow=c(1,1))  
plot(group$c[c(2,1)],xlim=x\_range,ylim=y\_range,pch=10,main="type = c")



plot(group$p[c(2,1)],xlim=x\_range,ylim=y\_range,pch=11,main="type = p")



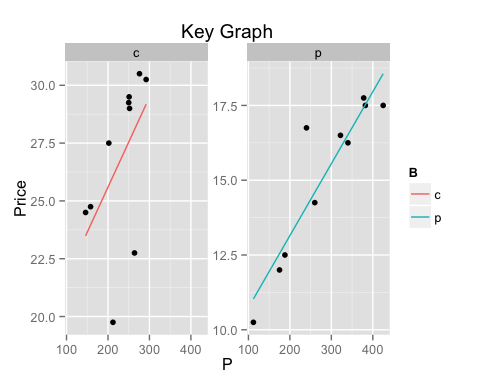
library(ggplot2)  
ggplot(E1.19,aes(x=P,y = Price,group=B,colour=B,shape=B))+ xlab("#page")+geom\_line()



Since their axes are the same, I don't think there is any difference betweent individual and overlaid graph helping me explore the data.

I personally don't want to distroy the structure of original data. But since the x variable varies wider than y variable, I may suggest to scale the x variable by letting each number be divided by their maximum.

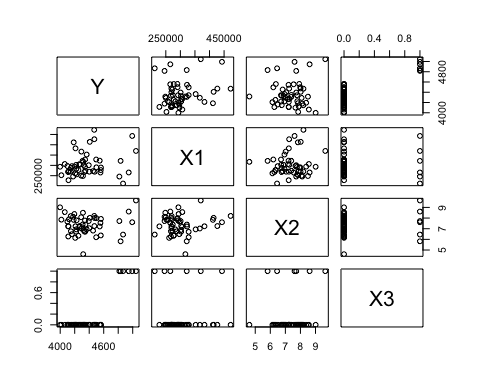
qplot(P,Price,data=E1.19)+  
 geom\_smooth(aes(colour=B),method="lm",se=F)+  
 facet\_wrap( ~ B,scale="free\_y")+  
 labs(title="Key Graph")



This picture clearly shows the linear relationship between price and page's volume in terms of two different kind of books.

# ex4

grocery<-read.table("grocery\_retailer.txt",header = T)  
plot(grocery)



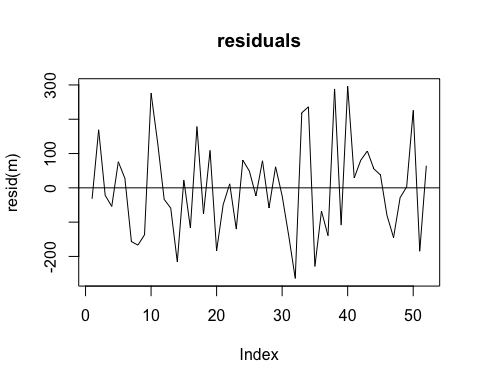
I can get the infomation of pair-wise correlation which can help we varify the signs of coefficients.

m<-lm(Y~X1+X2+X3,data = grocery)  
knitr::kable(tidy(m))

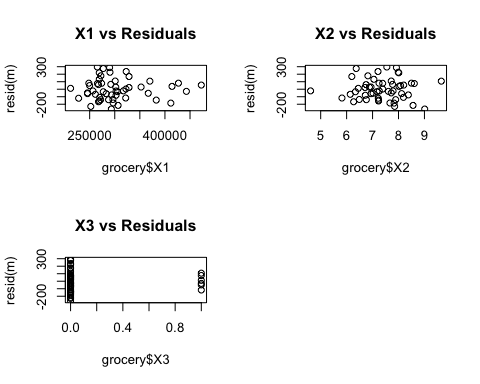
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 4149.8872 | 195.5654 | 21.2199 | 0.0000 |
| X1 | 0.0008 | 0.0004 | 2.1590 | 0.0359 |
| X2 | -13.1660 | 23.0917 | -0.5702 | 0.5712 |
| X3 | 623.5545 | 62.6409 | 9.9544 | 0.0000 |

The coefficient of X3 means that if there happens to be a holiday the total labor hours will be about 623 hours more than usual.

plot(resid(m),type="l",main="residuals")  
abline(h=mean(resid(m)))



layout(matrix(c(1,2,3,4),2,2,byrow = T))  
plot(x=grocery$X1,y=resid(m),main="X1 vs Residuals")  
plot(x=grocery$X2,y=resid(m),main="X2 vs Residuals")  
plot(x=grocery$X3,y=resid(m),main="X3 vs Residuals")



Apparently,there is a pattern between the residuals and independent variable X3.The variance of the error item can change along with X3. So the Gauss-Markov assuption is violated.