Homework 5

Animesh Sengupta

10/11/2022

```
setwd("/Users/animeshsengupta/Work Directory/DACSS/STAT625/Homeworks")
library(alr4)  # loads the installed package into the workspace so you can use it
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
library(summarytools)
library(ggplot2)
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(splines)
```

Answer 5.1

5.1.1

The mean function is given as

$$E(Y|U_2, ..., U_d) = \beta_0 + \beta_2 U_2 + ... + \beta_j Uj$$

From the given equation, since U is a dummy variable equal to 1 in the jth level. So from the above equation if j=1, then there are no levels for U in 1st level to be 1, hence all the levels will have value to be 0. Hence for level1 we can modify the above equation as :

$$E(Y|U_1) = \beta_0 + 0$$

Since for the jth level all other values of U become 0 other than the jth level we can decompose the mean function to :

$$E(Y|U_j) = \beta_0 + \beta_j U_j = 1$$

Answer 5.5

Answer 5.5.1

$$Y \sim A + B + A : B$$

5.5.2

$$\beta_0 = u_{11}\beta_1 = u_{21} - u_{11}\beta_2 = u_{12} - u_{11}\beta_3 = u_{13} - u_{11}\beta_4 = u_{11} + u_{22} - u_{21} - u_{12}\beta_5 = u_{11} + u_{22} - u_{21} - u_{13}$$

5.5.3 Since we are only fitting for the model Y ~ A+B , we dont have mixed interaction terms between A and B, hence in the above equation we can remove the b4 and b5 since there are no mixed interactions considered in the new model. Hence the B in terms of U can be written as:

$$\beta_0 = u_{11}\beta_1 = u_{21} - u_{11}\beta_2 = u_{12} - u_{11}\beta_3 = u_{13} - u_{11}$$

5.8

5.8.1

```
colnames(cakes)
```

```
## [1] "block" "X1"    "X2"    "Y"
```

```
cl<-lm(Y~X1+X2+I(X1^2)+I(X2^2)+X1:X2,data=cakes)
summary(cl)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + X1:X2, data = cakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4912 -0.3080  0.0200  0.2658  0.5454
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.204e+03  2.416e+02  -9.125 1.67e-05 ***
## X1           2.592e+01  4.659e+00   5.563 0.000533 ***
## X2           9.918e+00  1.167e+00   8.502 2.81e-05 ***
## I(X1^2)     -1.569e-01  3.945e-02  -3.977 0.004079 **
## I(X2^2)     -1.195e-02  1.578e-03  -7.574 6.46e-05 ***
## X1:X2       -4.163e-02  1.072e-02  -3.883 0.004654 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4288 on 8 degrees of freedom
## Multiple R-squared:  0.9487, Adjusted R-squared:  0.9167
## F-statistic:  29.6 on 5 and 8 DF,  p-value: 5.864e-05
```

As per the model summary , we can verify that the significance level of all polynomials terms are less than 0.005.

5.8.2

```
colnames(cakes)
```

```
## [1] "block" "X1"    "X2"    "Y"
```

```
cl<-lm(Y~X1+X2+I(X1^2)+I(X2^2)+X1:X2+block,data=cakes)
summary(cl)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + X1:X2 + block,
##     data = cakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4525 -0.3046  0.0200  0.2924  0.4883
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.205e+03  2.542e+02  -8.672 5.43e-05 ***
## X1           2.592e+01  4.903e+00   5.287 0.001140 **
## X2           9.918e+00  1.228e+00   8.080 8.56e-05 ***
## I(X1^2)     -1.569e-01  4.151e-02  -3.779 0.006898 **
## I(X2^2)     -1.195e-02  1.660e-03  -7.197 0.000178 ***
## block1       1.143e-01  2.412e-01   0.474 0.650014
## X1:X2       -4.163e-02  1.128e-02  -3.690 0.007754 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4512 on 7 degrees of freedom
## Multiple R-squared:  0.9503, Adjusted R-squared:  0.9077
## F-statistic: 22.31 on 6 and 7 DF,  p-value: 0.0003129
```

When adding the block term in our linear model , we get an interaction term of block1 in our mean function with an estimate. Although the significance level of that estiamte is very high i.e 0.65. Relatively in the previous model and this model , the estimate values hasnt changed and remains the same in both the models.

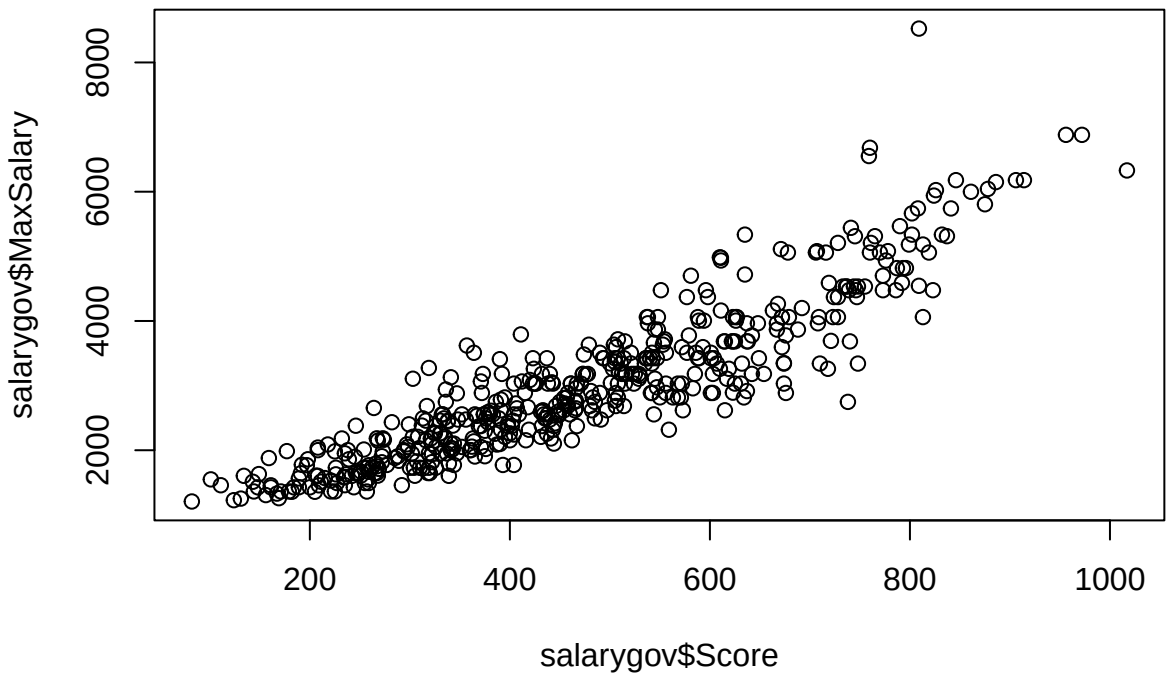
Answer 5.9

5.9.1

```
colnames(salarygov)
```

```
## [1] "JobClass"  "NW"        "NE"        "Score"     "MaxSalary"
```

```
plot(salarygov$Score,salarygov$MaxSalary)
```



As visualised by the plot, a simple linear regression might not be enough to predict accurately the score given the maxsalary because we can see an upwards curve trend. Also we can see a lot of variability at right while a lot many points are consolidated at right.

5.9.2

```
sgl0=lm(MaxSalary~Score,salarygov)
sgl1=lm(MaxSalary~bs(Score,4),salarygov)
sgl2=lm(MaxSalary~bs(Score,5),salarygov)
sgl3=lm(MaxSalary~bs(Score,10),salarygov)
plot(salarygov$Score,salarygov$MaxSalary,pch=".")

abline(sgl1,col = "red")
```

```
## Warning in abline(sgl1, col = "red"): only using the first two of 5 regression
## coefficients
```

```
abline(sgl2,col = "green")
```

```
## Warning in abline(sgl2, col = "green"): only using the first two of 6 regression
## coefficients
```

```
abline(sgl3,col = "blue")
```

```
## Warning in abline(sgl3, col = "blue"): only using the first two of 11 regression
## coefficients
```

5.11

5.11.1

```
colnames(MinnLand)
```

```
## [1] "acrePrice"    "region"       "improvements" "year"         "acres"
## [6] "tillable"     "financing"    "crpPct"       "productivity"
```

```
ml1<-lm(log(acrePrice)~year+region+year:region+financing,data=MinnLand)
summary(ml1)
```

```
##
## Call:
## lm(formula = log(acrePrice) ~ year + region + year:region + financing,
##     data = MinnLand)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8723 -0.2722  0.0131  0.2547  2.7200
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2.097e+02  6.321e+00 -33.177  < 2e-16 ***
## year                        1.079e-01  3.150e-03  34.259  < 2e-16 ***
## regionWest Central         -1.745e+00  9.248e+00  -0.189   0.8504
## regionCentral               3.564e+01  8.714e+00   4.089 4.35e-05 ***
## regionSouth West           -5.634e+01  1.010e+01  -5.579 2.46e-08 ***
## regionSouth Central         1.904e+01  9.714e+00   1.960   0.0500 .
## regionSouth East            5.956e+01  1.050e+01   5.670 1.45e-08 ***
## financingseller_financed -9.511e-02  1.126e-02  -8.448  < 2e-16 ***
## year:regionWest Central     1.266e-03  4.609e-03   0.275   0.7836
## year:regionCentral         -1.720e-02  4.343e-03  -3.959 7.55e-05 ***
## year:regionSouth West       2.862e-02  5.033e-03   5.686 1.32e-08 ***
## year:regionSouth Central -8.842e-03  4.841e-03  -1.827   0.0678 .
## year:regionSouth East    -2.902e-02  5.236e-03  -5.544 3.00e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4853 on 18687 degrees of freedom
## Multiple R-squared:  0.5534, Adjusted R-squared:  0.5531
## F-statistic:  1929 on 12 and 18687 DF,  p-value: < 2.2e-16
```

```
confint(ml1)["financingseller_financed",]
```

```
##       2.5 %      97.5 %
## -0.11717460 -0.07304286
```

5.11.2

1. Seller financing lowers sale prices : Since the estimate of the financingseller_financed is negative as per the aboce summary, we can say that it may imply a certain causation. Although it doesnt mean that seller financing is the only reason to lower the sale price and is the only determinant for lower sale.

2. Seller financing is more likely on lower-priced property transactions: Since the model assigned a negative estiimate for financingseller_financed then it means that it has seen observation where the lower priced property is seller financed, but since there are other estimators and terms involved it may not be the only sole reason.

5.19.

5.19.1

Wew have to assume that i belongs to (len,amp,load) and dummy variable

$$U_{ij}$$

for 3 levels of j i.e 1,2,3 hence we can write first order of mean function as:

$$E(log(cycles)|First-order) = \beta_0 + \sum_{i=1}^{3}\sum_{j=1}^{3}\beta_{ij}U_{ij}$$

$$E(log(cycles)|Second-order) = \beta_0 + \sum_{i=1}^{3}\sum_{j=1}^{3}\beta_{ij}U_{ij} + \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=i+1}^{3}\beta_{ijk}U_{ij}U_{kj}$$

5.19.2

If the load is increased from middle to higher level , it means that the jth factor increases from 1 to 2 for i=3. This means that the differnce in response will be estimator of i,j=3,3 minus i,j=3,2 i.e

$$\beta_{33} - \beta_{32}$$

for the second order similarly we will compute the change in interaction terms when i=1 and 2 while changing the level j from 2 to 3 i.e

$$\beta_{33} - \beta_{32} + \beta_{133} - \beta_{132} + \beta_{233} - \beta_{232}$$