Statistical Models and Computing

Yaniv Bronshtein

3/1/2022

Import the necessary libraries

```
library(MASS)
library(pid)
```

```
## Registered S3 method overwritten by 'DoE.base':
##   method           from
##   factorize.factor conf.design
```

```
library(hnp)
```

Problem 1.

We want to test the effect of light level and amount of water on the yield of tomato plants. Each potted plant receives one of three levels of light (1 = 5 hours, 2 = 10 hours, 3 = 15 hours) and one of two levels of water (1 = 1 quart, 2 = 2 quarts). The yield, in pounds, is recorded. The results are as follows: **Read in the data**

```
path1 <- '/Users/yanivbronshtein/Coding/Rutgers/Statistical_Computing_Repo/data/hw2_q1_data.txt'
q1_df <- read.table(path1,header=TRUE)
lm_fit <- lm(Yield~., data=q1_df)
```

Now let us print the summary object

```
summary(lm_fit)
```

```
##
## Call:
## lm(formula = Yield ~ ., data = q1_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6667 -0.9514 -0.2778  1.3750  4.1944
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.7222     2.3703  -0.727  0.47866
## Light         5.9167     0.7257   8.153 6.82e-07 ***
## Water         4.7778     1.1851   4.031  0.00109 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.514 on 15 degrees of freedom
## Multiple R-squared:  0.8465, Adjusted R-squared:  0.826
## F-statistic: 41.36 on 2 and 15 DF,  p-value: 7.868e-07
```

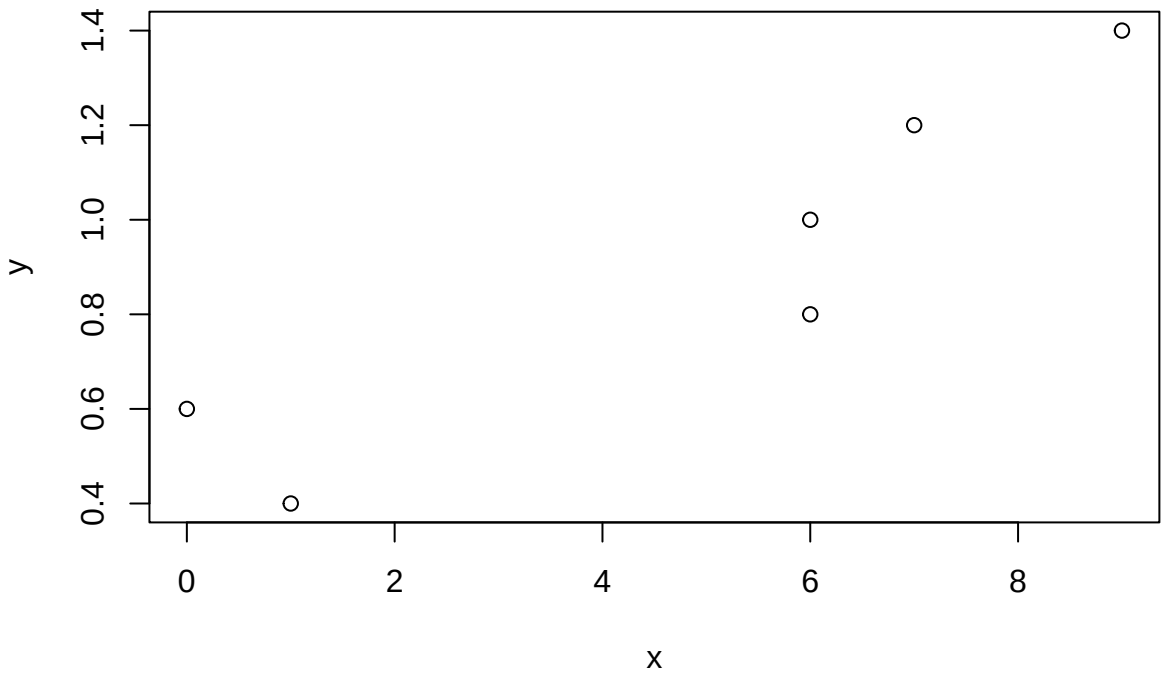
Light is and Water are both significant factors in determining yield because the p-value is less than 0.05 This supports the hypothesis that increasing light exposure and water exposure will lead to higher yield # Problem 2
The following data responses y are generated from a regular Poisson model with a single covariate variable x.

(a). Please write down the Poisson model for this data set, stating all requirements.

```
path2 <- '/Users/yanivbronshtein/Coding/Rutgers/Statistical_Computing_Repo/data/hw2_q2_data.txt'
q2_df <- read.table(path2, header=TRUE)
```

Plot the data

```
plot(x=q2_df$y, y=q2_df$x, xlab='x', ylab='y')
```



(c).

```
get_dev_res <- function(y, mu){
  return(sign(y - mu) * sqrt(2 * y * log(y/mu) - (y-mu)))
}

get_dev_res(1, 0.37)
```

```
## [1] 1.165549
```

Problem 3

Knight & Skagen collected the data shown in the table(and in data frame eagles) during a field study on the foraging behavior of wintering Bald Eagles in Washington State, USA. The data concern 160 attempts by one (pirating) Bald Eagle to steal a chum salmon from another (feeding) Bald Eagle. The abbreviations used are L=Large, S = small A = adult I = immature Report on factors that explain the success of the pirating attempt and give a prediction formula for the probability of success

```
q3_df <- eagles
eagles_glm <- glm(y/n ~ P*A + V,
                  data=q3_df,
                  family=binomial(link='logit'),
```

```
                  weights=n)

summary(eagles_glm)
```

```
##
## Call:
## glm(formula = y/n ~ P * A + V, family = binomial(link = "logit"),
##     data = q3_df, weights = n)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
## -0.02320   0.32133  -0.02030   0.31926   0.15434  -0.15279  -0.27823   0.04506
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.8977     0.4480   2.004  0.04507 *
## PS           -3.4605     1.1287  -3.066  0.00217 **
## AI           -0.3590     0.5986  -0.600  0.54870
## VS            5.4324     1.3602   3.994  6.5e-05 ***
## PS:AI        -3.6614     1.6279  -2.249  0.02450 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 128.26734  on 7  degrees of freedom
## Residual deviance:   0.33274  on 3  degrees of freedom
## AIC: 23.073
##
## Number of Fisher Scoring iterations: 6
```

Based on the summary object, at 0.1% significance code, the pirating eagle will have a higher success rate against a small sized feeding eagle(VS) which is why the coefficient in front of this term is the highest in magnitude and positive. Conversely, we see that at 1% significance code, if the pirating eagle is small, its success rate will decrease, explaining why the coefficient in front of (PS) is large in magnitude and negative. Finally, we see that the age of the pirating Eagle on its own is neither a significant contributor nor detractor. However, at the 5% significance code, when the pirating eagle is both immature and small, the effect of (AI) has an even greater adverse affect on the success rate then (PS) alone.

Problem 4.

A marketing research firm was engaged by an automobile manufacturer to conduct a pilot study to examine the feasibility of using logistic regression for ascertaining the likelihood that a family will purchase a new car during the next year. A random sample of 33 suburban families was selected. Data on annual family income (X1, in thousand dollars) and the current age of the oldest family automobile (X2, in years) were obtained. A follow-up interview conducted 12 months later was used to determine whether the family actually purchased a new car (Y = 1) or did not purchase a new car (Y = 0) during the year.

Read in Problem 4 data

```
q4_df <- read.table('/Users/yanivbronshtein/Coding/Rutgers/Statistical_Computing_Repo/data/Stat567_hw2_p
colnames(q4_df) <- c('Y', 'X1', 'X2')
```

(a) Find the maximum likelihood estimates of $\beta_0, \beta_1, and \beta_2$. State the fitted response function.

```
q4_glm <- glm(Y~., data=q4_df, family=binomial('logit'))
mle_estim <- q4_glm$coefficients

b0 <- mle_estim[[1]]; b1 <- mle_estim[[2]]; b2 <- mle_estim[[3]]
cat("Beta 0:", b0, " Beta 1:",b1, " Beta 2:", b2)
```

```
## Beta 0: -4.739309  Beta 1: 0.06773256  Beta 2: 0.5986317
```

Response Function

$1/(1+\exp(-(b0+(b1*50)+(b2*3))))$

```
response_function <- function(B,X){
  temp <- exp(B[1] + B[2]*X[1] + B[3]*X[2])

  return(as.numeric(temp/(1+temp)))
}
```

b) Obtain $\exp(\beta_1)$ and $\exp(\beta_2)$ and interpret these numbers.

```
cat("exp(Beta 1):", exp(b1),"\n", "exp(Beta 2):", exp(b2))
```

```
## exp(Beta 1): 1.070079
##  exp(Beta 2): 1.819627
```

Taking the exponential of our logit function gives us an understanding of our independent variables on the odds ratio. In our case exp(Beta 2) is almost twice as likely to occur as opposed to not occur. Thus, the age of the car has a higher impact on the likelihood of a sale than the income of the potential buyer.

(c) What is the estimated probability that a family with annual income of \$50 thousand and an oldest car of 3 years will purchase a new car next year?

```
prob <- response_function(B=c(b0,b1,b2), X=c(50.0,3.0))
prob
```

```
## [1] 0.6090245
```

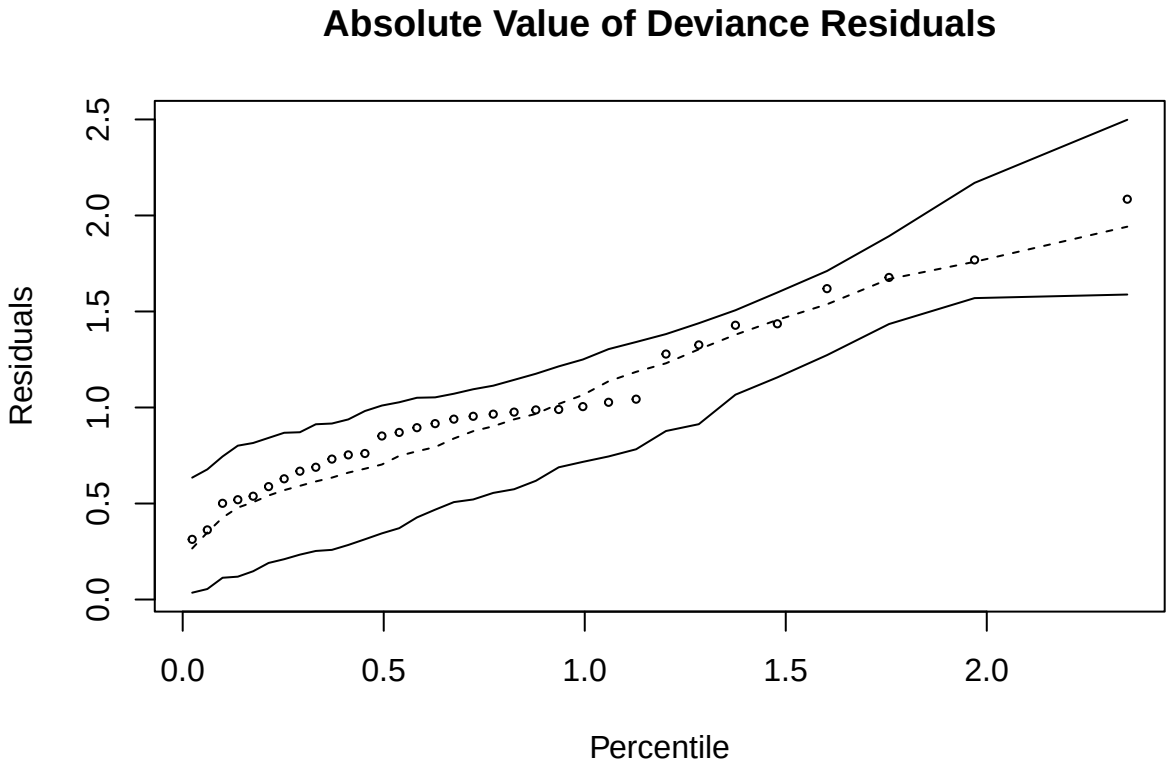
(d) Obtain the deviance residuals and present them in an index plot. Do there appear to be any outlying cases?

```
dev <- resid(q4_glm, type="deviance")
dev <- unname(dev)
plot(dev,ylab="Deviance Residuals",main="Index Plot of Deviance Residuals")
```

(e) Construct a half-normal probability plot of the absolute deviance residuals. Do any cases here appear to be outlying?

```
hnp(q4_glm,xlab="Percentile",main="Absolute Value of Deviance Residuals")
```

```
## Binomial model
```



```r
K <- function(elem){
  return((0.05-elem)/.22)
}

arr <- c(-1.43,-.95,-.19,.02,.14,.83,1.35,1.46, 2.62)
for(elem in arr){
  cat(K(elem),'\n')
}
```

```
## 6.727273
## 4.545455
## 1.090909
## 0.1363636
## -0.4090909
## -3.545455
## -5.909091
## -6.409091
## -11.68182
```

Problem 2: The following data responses y are generated from a regular Poisson model with a single covariate variable x:

x	y
0.4	1
0.6	0
0.8	6
1.0	6
1.2	7
1.4	9

(a). Please write down the Poisson Model for this dataset, stating all requirements.

Mean Model: $\mu_i = E(y_i) = g(\beta_0 + \beta_1 x_i) = e^{\beta_0 + \beta_1 x_i}$

Distribution: $y_i \sim \text{Poisson}(\mu)$ has density $f(y|\mu) = \frac{e^{-\mu}\mu^y}{y!}$ for $y = 0, 1, 2, \ldots$

Likelihood function: $L(\beta|y) = \prod_{i=1}^{n} f(y_i)$

Log-likelihood function: $\ell(\beta|y) = \sum_{i=1}^{n} \log(f(y_i)) = \sum_{i=1}^{n} \{y_i \log \mu_i - \mu_i - \log(y_i!)\}$

Take Derivative and set to 0

$$\frac{\partial \ell(\beta|y)}{\partial \beta} = \sum_{i=1}^{n} \{ y_i \frac{\partial \log \mu_i}{\partial \beta} - \frac{\partial \mu_i}{\partial \beta}$$

(b). Calculate the MLE estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_1 and β_2, and then provide the variance estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Solve $\begin{cases} \sum_{i=1}^{n} (y_i - e^{\beta_0 + \beta_1 x_i}) = 0 \quad ① \\ \sum_{i=1}^{n} (y_i - e^{\beta_0 + \beta_1 x_i}) x_i = 0 \quad ② \end{cases}$

Let us solve ①:

$$(1 - e^{\beta_0 + 0.4\beta_1}) + (0 - e^{\beta_0 + 0.6\beta_1}) + (6 - e^{\beta_0 + 0.8\beta_1}) + (6 - e^{\beta_0 + 1.0\beta_1})$$
$$+ (7 - e^{\beta_0 + 1.2\beta_1}) + (9 - e^{\beta_0 + 1.4\beta_1}) = 0$$

$$29 = e^{\beta_0 + 0.4\beta_1} + e^{\beta_0 + 0.6\beta_1} + e^{\beta_0 + 0.8\beta_1} + e^{\beta_0 + 1.0\beta_1} + e^{\beta_0 + 1.2\beta_1} + e^{\beta_0 + 1.4\beta_1}$$

$$= e^{\beta_0} (e^{0.4\beta_1} + e^{0.6\beta_1} + e^{0.8\beta_1} + e^{1.0\beta_1} + e^{1.2\beta_1} + e^{1.4\beta_1})$$

Now, let us solve ②:

$$\left(0.4 - 0.4e^{\beta_0+0.4\beta_1}\right)+\left(0-0.6e^{\beta_0+0.6\beta_1}\right)+\left(4.8-0.8e^{\beta_0+0.8\beta_1}\right)$$

$$+\left(6-e^{\beta_0+\beta_1}\right)+\left(8.4-1.2e^{\beta_0+1.2\beta_1}\right)+\left(12.6-1.4e^{\beta_0+1.4\beta_1}\right)=0$$

$$32.2=e^{\beta_0}\left(0.4e^{0.4\beta_1}+0.6e^{0.6\beta_1}+0.8e^{0.8\beta_1}+e^{\beta_1}+1.2e^{1.2\beta_1}+1.4e^{1.4\beta_1}\right)$$

Divide ① by ②

$$\frac{29}{32.2}=\frac{e^{0.4\beta_1}+e^{0.6\beta_1}+e^{0.8\beta_1}+e^{\beta_1}+e^{1.2\beta_1}+e^{1.4\beta_1}}{0.4e^{0.4\beta_1}+0.6e^{0.6\beta_1}+0.8e^{0.8\beta_1}+e^{\beta_1}}$$

$$32.2e^{0.4\beta_1}+32.2e^{0.6\beta_1}+32.2e^{0.8\beta_1}+32.2e^{\beta_1}+32.2e^{1.2\beta_1}+32.2e^{1.4\beta_1}=$$
$$29\left(0.4e^{0.4\beta_1}\right)+29\left(0.6e^{0.6\beta_1}\right)+29\left(0.8e^{0.8\beta_1}\right)+29\left(e^{\beta_1}\right)$$

$$\Downarrow$$

$$20.6e^{0.4\beta_1}+14.8e^{0.6\beta_1}+0.9e^{0.8\beta_1}+3.2e^{\beta_1}-2.6e^{1.2\beta_1}-8.4e^{1.4\beta_1}=0$$

$$\Downarrow$$

$$e^{\beta_1}\approx 2.16604 \Longrightarrow \hat{\beta}_1=\log(7.16604)=1.969$$

$29 = e^{\beta_0} * (43.8367)$

$\hat{\beta}_0 = -0.413$

Now write log likelihood

$$L(y|\mu) = \prod_{i=1}^{n} \frac{e^{-\mu} \mu_i^{y_i}}{y_i!}$$

$$\ell(y|\mu) = \sum_{i=1}^{n} \log\left(\frac{e^{-\mu} \mu_i^{y_i}}{y_i!}\right)$$

$$= \sum_{i=1}^{n} \log\left(\frac{e^{-\mu_i} \mu_i^{y_i}}{(y_i)!}\right)$$

$$= \sum_{i=1}^{n} -\mu_i + y_i \log \mu_i - \log(y_i!)$$

In our case, we have params β_0, β_1 and x

$$\ell(y_i | \beta_0, \beta_1, x_i) = \sum_{i=1}^{n} -e^{\beta_0 + \beta_1 x_i} + \sum_{i=1}^{n} y_i \cdot (\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \log(y_i!)$$

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{n} -e^{\beta_0 + \beta_1 x} + \sum_{i=1}^{n} y_i$$

Second Partial Derivative Now

$$\frac{\partial^2 \ell}{\partial \beta_0^2} = \sum_{i=1}^{n} -e^{\beta_0 + \beta_1 x_i}$$

$$I_{\beta_0} = -E\left(\sum_{i=1}^{n} -e^{\beta_0 + \beta_1 x_i}\right) = \sum_{i=1}^{n} e^{\beta_0 + \beta_1 x_i} = e^{0.77} + e^{0.77} + e^{1.16} + e^{1.56} + e^{1.95} + e^{2.34}$$

$$= 28.967 \qquad Var(\hat{\beta}_0) = I_{\beta_0}^{-1} = 0.035$$

Now take derivative w.r.t β_1

$$\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^{n} -x_i e^{\beta_0 + \beta_1 x_i} + \sum_{i=1}^{n} x_i y_i$$

$$\frac{\partial^2 \ell}{\partial \beta_1^2} = \sum_{i=1}^{n} -x_i^2 e^{\beta_0 + \beta_1 x_i}$$

$$I_{\beta_1} = -E\left(\sum_{i=1}^{n} -x_i^2 e^{\beta_0 + \beta_1 x_i}\right) = \sum_{i=1}^{n} x_i^2 e^{\beta_0 + \beta_1 x_i} =$$

$$(0.4)^2 e^{0.37} + (0.6)^2 e^{0.77} + (0.8)^2 e^{1.16} + e^{1.56} + (1.2)^2 e^{1.95} + (1.4)^2 e^{2.34}$$

Thus $I_{\beta_1} = 38.278$

$Var(\hat{\beta}_1) = I_{\beta_1}^{-1} = 0.026$

(c). The values of the linear predictor are 0.37, 0.77, 1.16, 1.56, 1.95, 2.34 for the 6 observations. Please compute the deviance residuals and draw the index plot of the deviance residuals.

$$r_i = \mathrm{sgn}(y_i - \hat{\mu}_i) * \sqrt{2 * y_i * \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)}$$

$$r_1 = \mathrm{sgn}(1 - 0.37) * \sqrt{2 * 1 * \log\left(\frac{1}{0.37}\right) - (1 - 0.37)}$$
$$= 1.165549$$

$$r_2 = \mathrm{sgn}(0 - 0.77) * \sqrt{2 * 0 * \log\left(\frac{0}{0.77}\right) - (0 - 0.77)}$$
$$= -0.8774964$$

$$r_3 = \mathrm{sgn}(6 - 1.16) * \sqrt{2 * 6 * \log\left(\frac{6}{1.16}\right) - (6 - 1.16)}$$
$$= 3.85747$$

$$r_4 = \mathrm{sgn}(6 - 1.56) * \sqrt{2 * 6 * \log\left(\frac{6}{1.56}\right) - (6 - 1.56)}$$
$$= 3.424162$$

$$r_5 = \mathrm{sgn}(7 - 1.95) * \sqrt{2 * 7 * \log\left(\frac{7}{1.95}\right) - (7 - 1.95)}$$
$$= 3.583731$$

$$r_6 = \mathrm{sgn}(9 - 2.34) * \sqrt{2 * 9 * \log\left(\frac{9}{2.34}\right) - (9 - 2.34)}$$
$$= 4.193725$$

(d) Draw a partial residual plot to study the linearity of the covariate variable x (show your calculation)

$$r_i^K = r_i^D + B_1 x_i$$

$r_1^K = -0.394 + 1.969 * 0.4 = 0.3936$

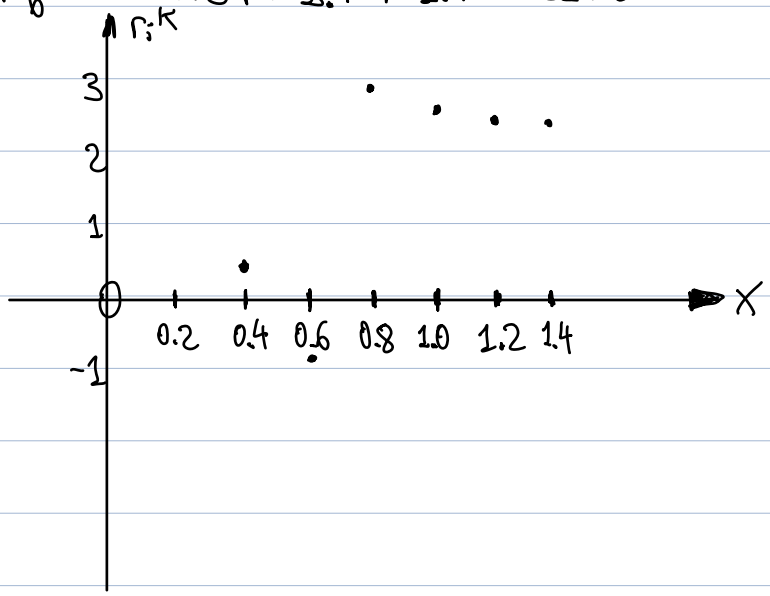
$r_2^K = -2.078 + 1.969 * 0.6 = -0.8966$

$r_3^K = 1.400 + 1.969 * 0.8 = 2.9752$

$r_4^K = 0.547 + 1.969 * 1.0 = 2.516$

$r_5^K = -0.011 + 1.969 * 1.2 = 2.3518$

$r_6^K = -0.439 + 1.969 * 1.4 = 2.3176$

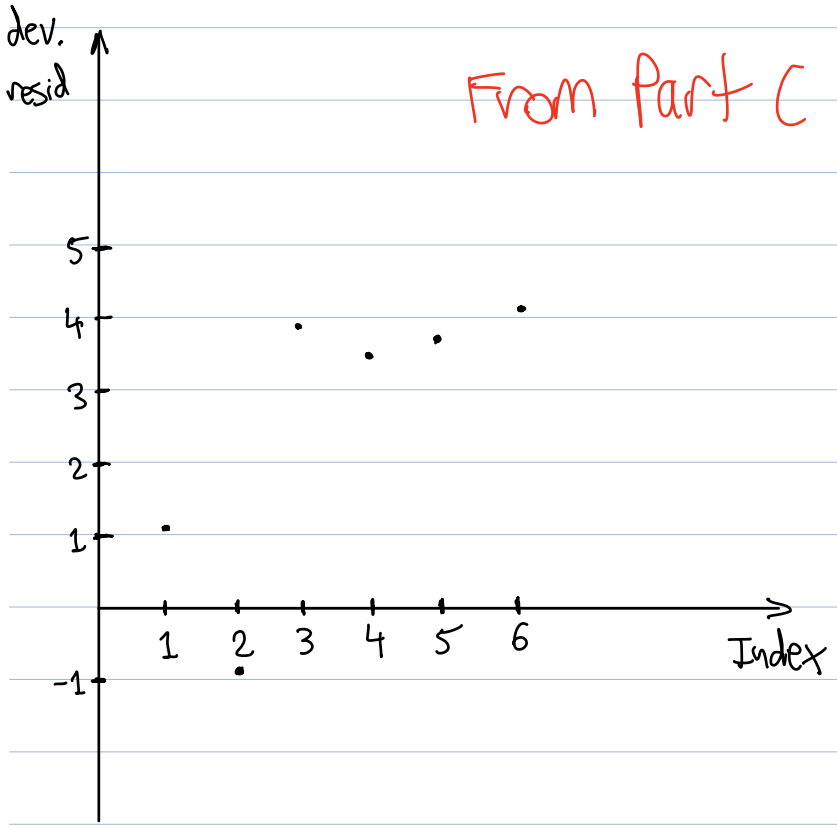


(e). $D_m(F) = (0.394)^2 + (2.078)^2 + (1.400)^2 + (0.547)^2 + (0.011)^2 + (0.439)^2 = 6.925$

$D_M(R)$

y	μ	n_i^D
1	$e^{-0.413}$	$2\left(1 * \log\left(\frac{1}{e^{-0.413}}\right) - 1 + e^{-0.413}\right)$
0	$e^{-0.413}$	$2\left(0 * \log\left(\frac{0}{e^{-0.413}}\right) - 0 + e^{-0.413}\right)$
6	$e^{-0.413}$	$2\left(6 * \log\left(\frac{6}{e^{-0.413}}\right) - 6 + e^{-0.413}\right)$
6	$e^{-0.413}$	$2\left(6 * \log\left(\frac{6}{e^{-0.413}}\right) - 6 + e^{-0.413}\right)$
7	$e^{-0.413}$	$2\left(7 * \log\left(\frac{7}{e^{-0.413}}\right) - 7 + e^{-0.413}\right)$
9	$e^{-0.413}$	$2\left(9 * \log\left(\frac{9}{e^{-0.413}}\right) - 9 + e^{-0.413}\right)$

$\sum n_i^D = 88.257$



From Part C

Problem 4:

(c). $\pi = \dfrac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}$

$= \dfrac{e^{-4.739309 + 0.06773256 x_1 + 0.5986317 x_2}}{1 + e^{-4.739309 + 0.06773256 x_1 + 0.5986317 x_2}}$

Problem 5: Given the following data points:

−1.43	−0.95	−0.19	0.02	0.14	0.83	1.35	1.46	2.62
1	2	3	4	5	6	7	8	9

Compute the kernel density estimate $\hat{f}(x)$ at point $x=0.05$. Use the rectangular kernel $K(t)$ with bin width $h=0.22$. Here $K(t)=\frac{1}{2}$ if $|t|\leq 1$, and it equals 0 if $|t|>1$.

$$\hat{f}_N(x;h)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right) \qquad K(z)=\frac{1}{2}\mathbf{1}_{\{-1<z<1\}}$$

$$\hat{f}_N(x=0.05;\,h=0.22)=\frac{1}{9*0.22}\left[K\left(\frac{0.05-(-1.43)}{0.22}\right)+K\left(\frac{0.05-(-0.95)}{0.22}\right)\right.$$

$$+K\left(\frac{0.05-(-0.19)}{0.22}\right)+K\left(\frac{0.05-0.02}{0.22}\right)+K\left(\frac{0.05-0.14}{0.22}\right)+$$

$$\left.K\left(\frac{0.05-0.83}{0.22}\right)+K\left(\frac{0.05-1.35}{0.22}\right)+K\left(\frac{0.05-1.46}{0.22}\right)+K\left(\frac{0.05-2.62}{0.22}\right)\right]$$

$$=\frac{1}{9*0.22}\left[K(6.73)+K(4.55)+K(1.09)+K(0.136)+K(-0.409)+K(-3.545)\right.$$

$$\left.+K(-5.91)+K(-6.409)+K(-11.682)\right]=0$$

$$=0.5050505\left[0+0+0+1+1+0+0+0+0\right]$$

$$=\boxed{1.01}$$