Homework 3

Gianni Spiga

2023-01-30

Contents

Problem Set 2

Question 4

a.) Repeat the Poisson regression fit for the Melanoma data and obtain Pearson and deviance residuals.

Below is the Poisson model for our melanoma data, as well as the first 5 residual values for both Pearson and Deviance residuals.

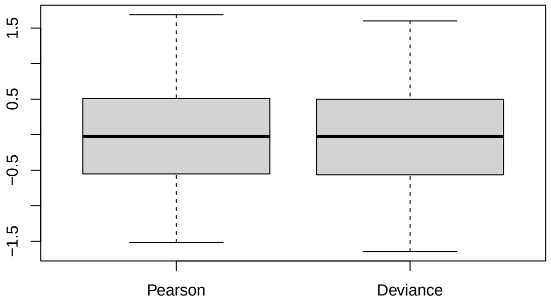
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-76.1070019	6.0198323	-12.6427114	0.0000000
years	0.0405879	0.0030747	13.2005736	0.0000000
sunspotnumber	0.0005740	0.0005996	0.9573381	0.3383966

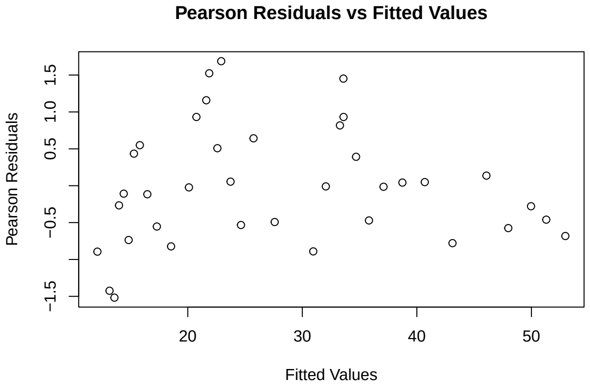
Pearson	Deviance
-0.8938730	-0.9369332
-1.4239931	-1.5368132
-1.5174483	-1.6449550
-0.2663841	-0.2696430
-0.1080065	-0.1085248

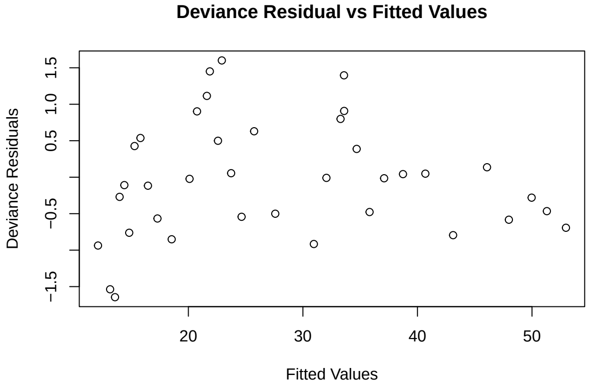
b.)

We first print the summary statistics for both residuals. We can see that they are quite similar in measures of summary statistics.

Pearson	Deviance
Min. :-1.517448	Min. :-1.64495
1st Qu.:-0.553228	1st Qu.:-0.56623
Median :-0.022783	Median :-0.02280
Mean :-0.009038	Mean :-0.03128
3rd Qu.: 0.508303	3rd Qu.: 0.49962
Max. : 1.687693	Max. : 1.60077




Pearson Residuals vs Fitted Values


Deviance Residual vs Fitted Values

Based on the plots, we can see that the deviance and pearson residuals are very similar to each other, thus indicating that our model is a good fit.

c.)

There is no indication of a lack of fit in the reisdual plots.

d.)

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	36	209.42068	NA
years	1	186.0507659	35	23.36991	0.0000000
sunspotnumber	1	0.9106512	34	22.45926	0.3399417

e.)

```
## Start:  AIC=215.83
## totalincidence ~ years + sunspotnumber
##
##                 Df Deviance    AIC
## - sunspotnumber  1   23.370 214.74
## <none>              22.459 215.83
## - years          1  206.330 397.70
##
## Step:  AIC=214.74
## totalincidence ~ years
##
##                 Df Deviance    AIC
## <none>              23.370 214.74
## + sunspotnumber  1   22.459 215.83
## - years          1  209.421 398.79
```

f.)

Based off the stepwise regression, we would pick the model with the only predictor being years, since it has the lowest AIC. We can conclude that the best way to model the incidence of melanomas is to track the year it was acquired for each patient.

Question 5

See written work

Problem Set 3

Question 1, 3, 4

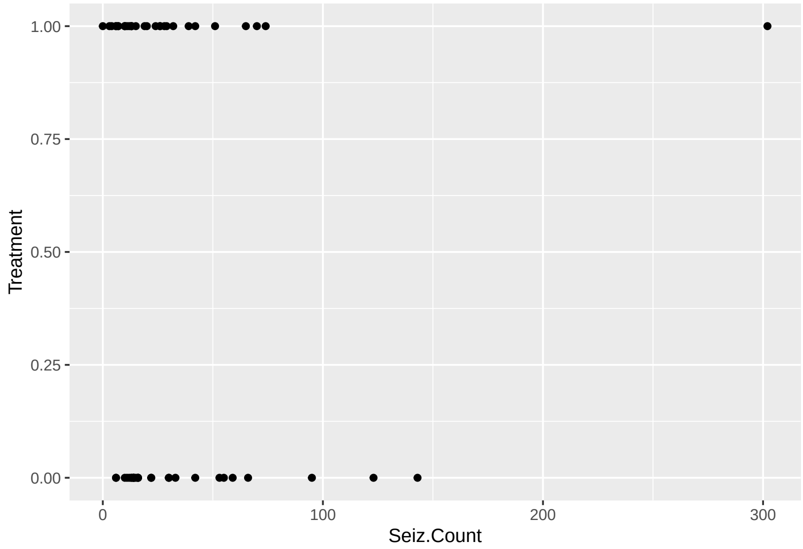
See written work

Question 7

Treatment	Baseline	Patient.Age	Seiz.Count
0	11	31	14
0	11	30	14
0	6	25	11
0	8	36	13
0	66	22	55
0	27	29	22

Above we can see a small preview of the cleaned data set. Before we begin analysis, we will check if the data has any outliers.

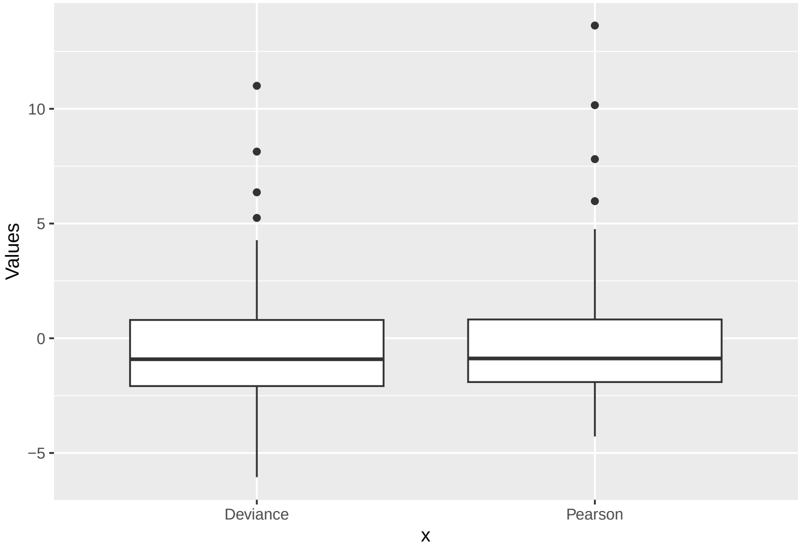
```
##
## Attaching package: 'plotly'

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

## The following object is masked from 'package:MASS':
##
##     select

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

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



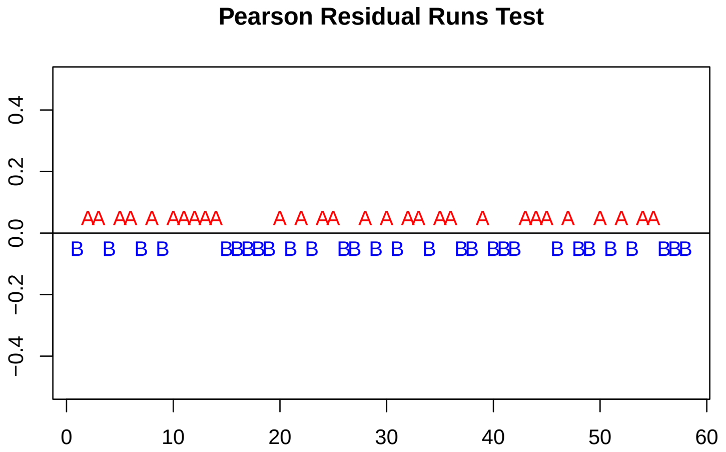
	Treatment	Baseline	Patient.Age	Seiz.Count
49	1	151	22	302

We can see that subject 49 has an abnormally high amount of seizures, so we will remove this observation from the data and begin analysis with our Poisson Regression model.

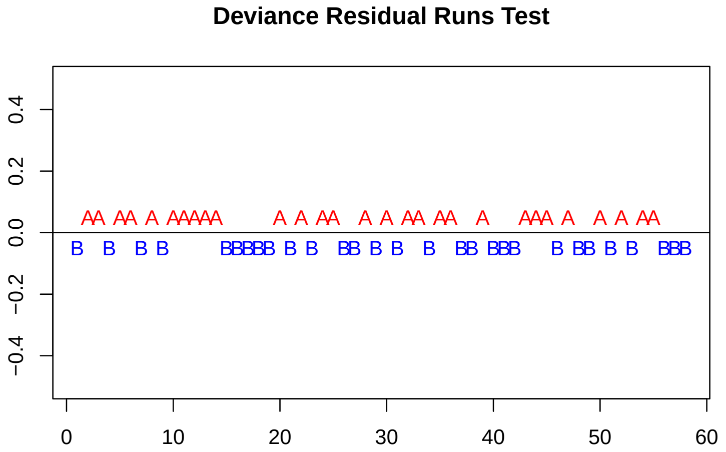
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.9426534	0.1382858	14.048109	0.000000
as.factor(Treatment)1	-0.1471737	0.0535225	-2.749753	0.005964
Baseline	0.0228025	0.0008308	27.446863	0.000000
Patient.Age	0.0226800	0.0040317	5.625471	0.000000



```
##
##  Runs Test - Two sided
##
## data:  chemo.res.P
## Standardized Runs Statistic = 1.3247, p-value = 0.1853
```



```
##
##  Runs Test - Two sided
##
## data:  chemo.res.D
## Standardized Runs Statistic = 1.3247, p-value = 0.1853
```



We can see that all of our predictors in the Poisson model are highly significan, meaning that with a null hypothesis of $\beta_i = 0$, we would reject. We can see from the coefficients that Baseline and Patient Age increase the log expected count of total seizures while the treatment type decreases the log expected count of seizures.

We will next check the goodness-of-fit comparing Deviance and Pearson residuals. From the boxplot above, we can see that our Pearson and Deviance Residuals are very similar, indicating that our model is a good fit. From the Runs Test, we can see that there is no indication of a non-random pattern, causing us to fail to reject the null hypothesis that the sequence of residuals was produced in a random manner.

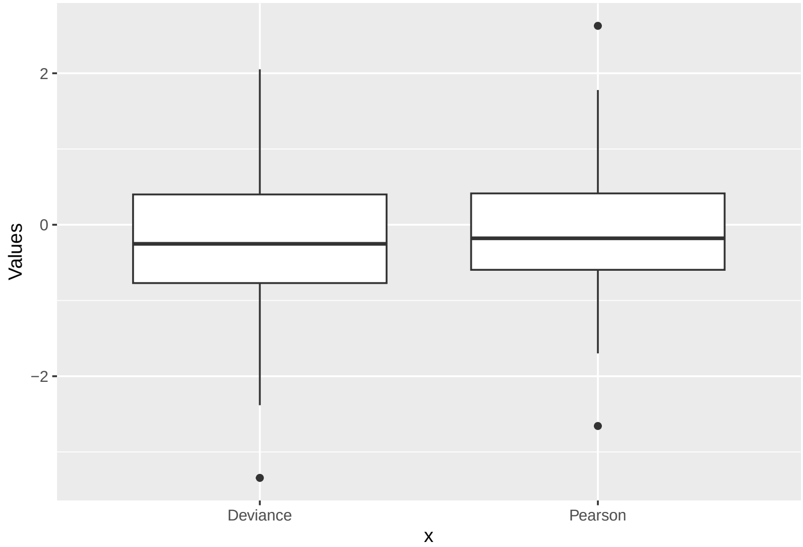
Question 9

a.)

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	52	69.50926	188.1853
- as.factor(Race)	1	1.430665	53	70.93992	187.6159
- as.factor(Sex)	1	1.622000	54	72.56192	187.2379

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	64	322.52677	NA
Dust	1	221.962599	63	100.56417	0.0000000
as.factor(Race)	1	1.054008	62	99.51016	0.3045858
as.factor(Sex)	1	5.966920	61	93.54324	0.0145767
as.factor(Smoker)	1	10.726026	60	82.81721	0.0010564
EmpLength	1	13.307956	59	69.50926	0.0002643

When we perform a stepwise logistic regression with the AIC, we get a best model including Dustiness, Smoker Status, and Employment length. When we find the best model via a deviance table, we get a model that includes all the same variables as before as well as the Sex of the worker. Let's observe the residuals of the smaller model.



We will pick the smaller model as our fit, since the deviance table is inconsistent since it depends on the order of variables, as well as the p-value for the additional variable is the highest.

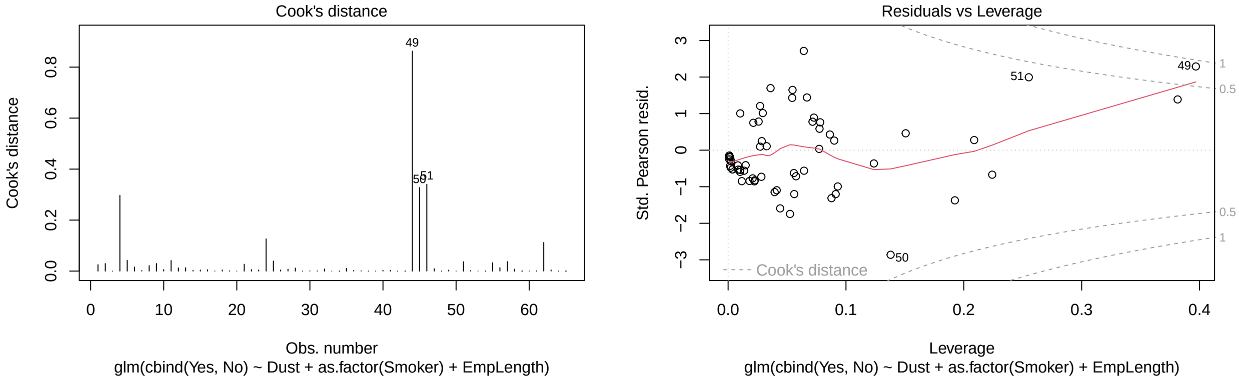
b.)

We would be curious in if if being a non-smoker status decreases chance of illness, so we would be interested in finding out if this rate was strictly greater than 0. We would test:

$$H_0 : \beta_{smoker} \geq 0 H_a : \beta_{smoker} < 0$$

We can calculate this p-value by dividing the two tailed p-value by two, giving us a p-value of 0.000164, leading us to reject H_0.

c.)



From our first plot, we can see that that observation 49 has the highest Cooks distance, indicating that it is an influential outlier. We can also see more clearly from our second plot that observation is another high leverage outlier.

d.)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.0270131	0.5580928	-1.8402193	0.0657361
Dust	-1.4417773	0.3203495	-4.5006385	0.0000068
as.factor(Smoker)2	-1.3831580	1.0964589	-1.2614772	0.2071370
EmpLength	0.5135283	0.2446207	2.0992840	0.0357919
Dust:as.factor(Smoker)2	0.7141178	0.5499181	1.2985893	0.1940849
Dust:EmpLength	-0.0602434	0.1387046	-0.4343286	0.6640498
as.factor(Smoker)2:EmpLength	0.0407968	0.5020888	0.0812540	0.9352399
Dust:as.factor(Smoker)2:EmpLength	-0.1792553	0.2559577	-0.7003319	0.4837201

$$H_0 : A\beta = 0; A \in R^{5x6}, \beta \in R^6 H_a : A\beta = 0; A \in R^{5x6}, \beta \in R^6$$

The degrees of freedom in this example would be 5. Based on the summary table, we should not have any interactions in this model, as none of them are significant. With this model, our smoker variable (main effect, not interaction), is also not significant.

Provided Question 3

See written work.

Appendix

```r
library(knitr)

mela <- read.table("melanoma.txt", header = TRUE)

glm.mela <-
  glm(totalincidence ~ years + sunspotnumber,
      data = mela,
      family = poisson())
sum.mela <- summary(glm.mela)
kable(sum.mela$coefficients)

residPear <- residuals(glm.mela, type = "pearson")
residDev <- residuals(glm.mela, type = "deviance")

residData <- data.frame(residPear, residDev)
names(residData) <- c("Pearson", "Deviance")
kable(head(residData, 5))
kable(summary(residData))

# Boxplots of both residual types
boxplot(residData)

# Residuals vs fitted values
plot(
  glm.mela$fitted.values,
  residData[, 1],
  main = "Pearson Residuals vs Fitted Values",
  xlab = "Fitted Values",
  ylab = "Pearson Residuals"
)

plot(
  glm.mela$fitted.values,
  residData[, 2],
  main = "Deviance Residual vs Fitted Values",
  xlab = "Fitted Values",
  ylab = "Deviance Residuals"
)
# Deviance table
kable(anova(glm.mela, test = "Chisq"))
library(MASS)
scope <- list(upper = ~years + sunspotnumber, lower = ~1)
mela.step <- stepAIC(glm.mela, trace = TRUE, scope = scope)
chemo <- read.table("chemo.dat")
names(chemo) <-
  c(
    "ID",
    "1st.Period",
    "2nd.Period",
    "3rd.Period",
    "4th.Period",
    "Treatment",
    "Baseline",
    "Patient.Age"
  )
chemo["Seiz.Count"] <- chemo[, 2] + chemo[, 3] + chemo[, 4] + chemo[, 5]

chemo <- chemo[, -c(1:5)]
# Preview our cleaned dataset
kable(head(chemo))
library(ggplot2)
library(plotly)

# Scatter of Treatment vs Seiz.Count
ggplot(data = chemo, aes(x = Seiz.Count, y = Treatment)) + geom_point()

# One point with a total amount of seizures over 300
kable(chemo[which(chemo["Seiz.Count"] >= 300),])

# Remove the outlier
chemo.out <- which(chemo["Seiz.Count"] >= 300)
chemo <- chemo[-49, ]
chemo.fit <-
  glm(
    Seiz.Count ~ as.factor(Treatment) + Baseline + Patient.Age,
    data = chemo,
    family = poisson()
  )
chemo.fit.sum <- summary(chemo.fit)
kable(chemo.fit.sum$coefficients)

# Check residuals
chemo.res.P <- residuals(chemo.fit, type = "pearson")
chemo.res.D <- residuals(chemo.fit, type = "deviance")

ggplot() + geom_boxplot(aes(x = "Pearson", y = chemo.res.P)) + geom_boxplot(aes(x = "Deviance", y = chem

# Runs test
library(lawstat)
runs.test(y = chemo.res.P, plot.it = TRUE)
title(main = 'Pearson Residual Runs Test')
runs.test(y = chemo.res.D, plot.it = TRUE)
title(main = 'Deviance Residual Runs Test')
lung <- read.table("lung.dat", header = T)

lung.fit <-
  glm(
    cbind(Yes, No) ~ Dust + as.factor(Race) + as.factor(Sex) + as.factor(Smoker) + EmpLength,
    data = lung,
    family = "binomial"
  )

lung.step <- stepAIC(lung.fit, trace = F)
lung.step.fit <-
  glm(
    cbind(Yes, No) ~ Dust + as.factor(Smoker) + EmpLength,
    data = lung,
    family = "binomial"
  )
kable(lung.step$anova)

kable(anova(lung.fit, test = "Chi"))
lung.resid.P <- residuals(lung.step.fit, type = "pearson")
lung.resid.D <- residuals(lung.step.fit, type = "deviance")

ggplot() + geom_boxplot(aes(x = "Pearson", y = lung.resid.P)) + geom_boxplot(aes(x = "Deviance", y = lun
#summary(lung.step.fit)
# Find leverage points
lung.lev <- hatvalues(lung.step.fit)
lung.cook <- cooks.distance(lung.step.fit)
plot(lung.step.fit, which = c(4, 5))
lung.int <-
  glm(
    cbind(Yes, No) ~ (Dust * as.factor(Smoker) * EmpLength) * (Dust * as.factor(Smoker) * EmpLength),
    data = lung,
    family = "binomial"
  )
kable(summary(lung.int)$coefficients)
```

Written Work