Homework 5

Gianni Spiga

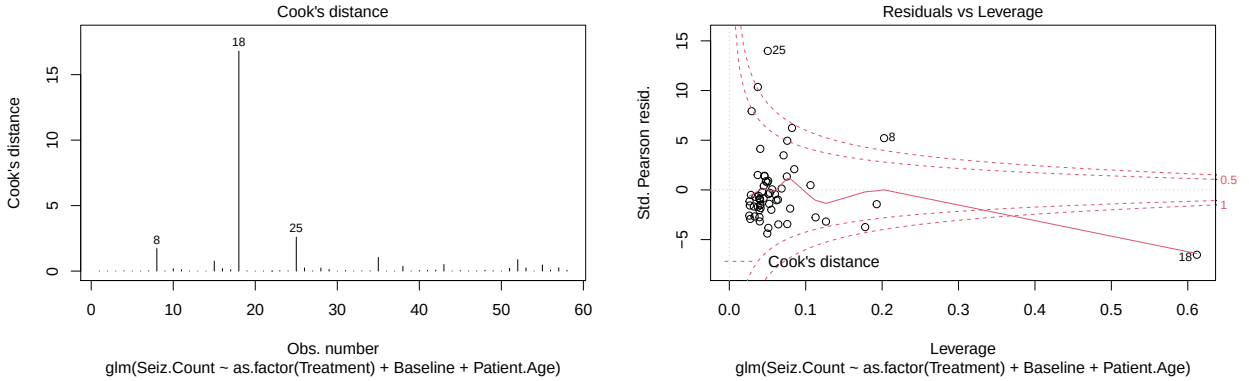
2023-02-14

Problem Set 4

Question 6

a.)

	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



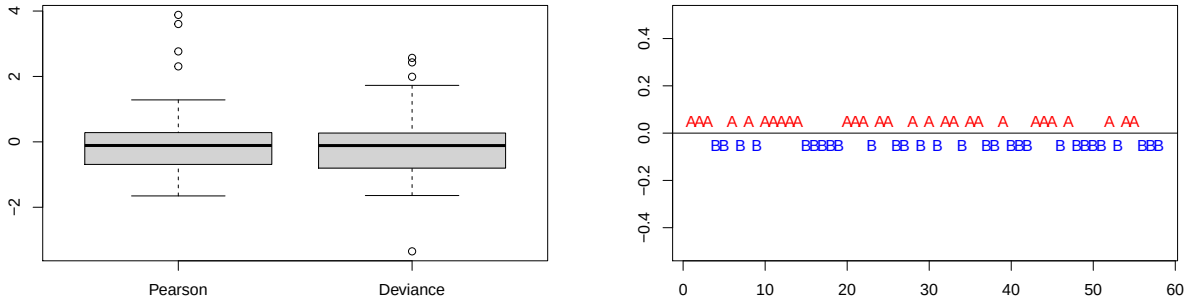
From our first plot, we can see that that observation 18 has the highest Cooks distance, indicating that it is an influential outlier. We can also see more clearly from our second plot that observation 8 is another high leverage outlier, but it is much smaller in reference in to observation 18.

b.)

In our data, we have a very high amout of overdispersion, with a dispersion factor $\sigma^2 = 10.36$, thus we have a violation of the classical poisson assumptions and should find a better model. Overdispersion can arrise when our counts have larger variation than expected.

c.)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8695188	0.4132624	4.523805	0.0000061
as.factor(Treatment)1	-0.1727056	0.1561375	-1.106112	0.2686781
Baseline	0.0305513	0.0035050	8.716412	0.0000000
Patient.Age	0.0158978	0.0125267	1.269113	0.2044008




```
##
##  Runs Test - Two sided
##
## data:  residDev
## Standardized Runs Statistic = 0, p-value = 1
```

After fitting our negative binomial model, we can see from both the comparison of residuals and the runs test that our model is a good fit. Runs test shows no obvious pattern. Our dispersion for the model is now only $\sigma^2 = 1.16$, a huge improvement from the classical Poisson model used before.

Question 7

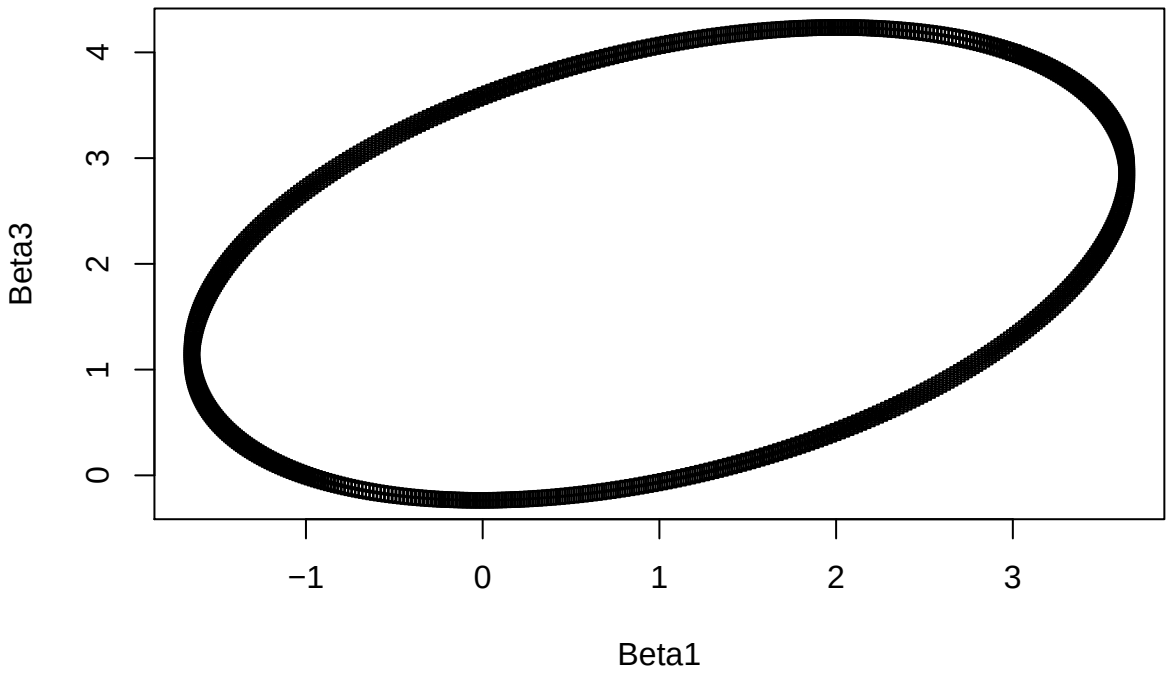
a.)

The most appropriate glm for an approach to this analysis would be the poisson regression, where our response would be the counts of infected individuals, and the snail control index, medical access index, hygeine index, and number of inhabitants would be predictors. This is under the assumption that the counts are distributed Poisson and we do not have over dispersion in the model. Hygeine index would be coded with bad as the baseline and split into two dummy variables: one for medium and the other for good.

b.)

We would expect the number of inhabitants in the village to be a relevant predictor. If the number of inhabitants is larger, we would expect to see a larger number of inhabitants infected. If the predictor is signficant, we would keep the variable encoded as a continuous predictor.

c.)



We could have a 95% confidence ellipsoid such as the one above for the β coefficients for snail control index and access to medical facility variables. We would use the sample coefficients for each respective variable to test a hypothesis involving the two coefficients.

d.)

We would want to test the hypothesis:

$$H_0 : \beta_1 \leq \beta_3 H_a : \beta_1 > \beta_3$$

which could be rewritten as:

$$H_0 : \beta_1 - \beta_3 \leq H_a : \beta_1 - \beta_3 > 0$$

where β_1 is the coefficient for controlling snails and β_3 is the coefficient for medical access. We would create a test statistic that would be the subtraction of the sample coefficients $\hat{\beta}_1$ and $\hat{\beta}_3$, which would be distributed $N(\beta_1 - \beta_3, \Sigma)$

e.)

We would notice that because of this fact, we would have a much larger amount of zero counts in the model than expected, so as a consequence we would have overdispersion. If we were to have a model that fit, we would not have our null counts being fit correctly, as it is very plausable that it will have a different distribution than the non-zero counts.

Question 8

a.)

```
## # weights:  39 (24 variable)
## initial  value 1756.681050
## iter  10 value 1424.322261
## iter  20 value 1219.250892
## iter  30 value 1215.897003
## iter  40 value 1215.580736
## iter  50 value 1213.876622
## iter  50 value 1213.876618
## final  value 1213.876618
## converged
```

	Medium	Good
(Intercept)	-90.1775405	111.4457482
fixed.acidity	0.0108918	0.2158201
volatile.acidity	-3.1314319	-4.9262377
citric.acid	-1.4257650	-0.5423116
residual.sugar	-0.0200410	0.2003092
chlorides	-3.1354382	-11.2235064
free.sulfur.dioxide	0.0237952	0.0265700
total.sulfur.dioxide	-0.0156443	-0.0263690
density	85.4142793	-127.5715383
pH	-0.8547758	-0.7612392
sulphates	2.2519561	5.1914883
alcohol	0.8695651	1.4374749

	Value	Std. Error	t value	pval.wine
fixed.acidity	0.1894909	0.0535620	3.5377875	0.0004035
volatile.acidity	-3.0882525	0.4312538	-7.1611020	0.0000000
citric.acid	-0.8149960	0.4947367	-1.6473330	0.0994896
residual.sugar	0.1256961	0.0395798	3.1757612	0.0014944
chlorides	-5.1314377	1.4464174	-3.5476880	0.0003886
free.sulfur.dioxide	0.0173645	0.0073051	2.3770272	0.0174528
total.sulfur.dioxide	-0.0170623	0.0027653	-6.1700973	0.0000000
density	-119.8533815	1.0078356	-118.9215540	0.0000000
pH	-0.2423291	0.5194272	-0.4665314	0.6408351
sulphates	3.1798018	0.3670429	8.6632983	0.0000000
alcohol	0.8152740	0.0609239	13.3818516	0.0000000
Low|Medium	-110.8748681	1.0318207	-107.4555544	0.0000000
Medium|Good	-107.9737745	1.0449156	-103.3325338	0.0000000

From the fitting of the proportional odds and the baseline odds models, we can see that the fits are very close in accuracy.Our proportional and baseline odds models return an AIC of 2474.01 and 2475.753 respectively. We can see that the predictor "density" has a far greater influence on wine quality than any other predictor, given its very large coefficient. We can notice that in the baseline odds model, the density has a positive effect for medium ratings and a negative effect on good ratings. If we compare this to the proportional odds model, we see that we only have a negative coefficient overall. The significant predictors in these models are all but pH.

b.)

```
## Warning: glm.fit: algorithm did not converge
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-26.5661388	1.164866e+07	-0.0000023	0.9999982
fixed.acidity	0.0000002	1.426898e+04	0.0000000	1.0000000
volatile.acidity	0.0000002	6.763277e+04	0.0000000	1.0000000
citric.acid	-0.0000012	8.098320e+04	0.0000000	1.0000000
residual.sugar	0.0000000	8.246280e+03	0.0000000	1.0000000
chlorides	0.0000003	2.310211e+05	0.0000000	1.0000000
free.sulfur.dioxide	0.0000000	1.197041e+03	0.0000000	1.0000000
total.sulfur.dioxide	0.0000000	4.060545e+02	0.0000000	1.0000000
density	0.0000724	1.188989e+07	0.0000000	1.0000000
pH	0.0000014	1.053050e+05	0.0000000	1.0000000
sulphates	0.0000003	6.377662e+04	0.0000000	1.0000000
alcohol	-0.0000008	1.491513e+04	0.0000000	1.0000000
quality2Good	53.1321353	2.128676e+04	0.0024960	0.9980085

Now using a logistic regression with only two categories in the response, we have a model with a much lower AIC. In this model, we can see that the relevant predictors are volatile acidity, citric acid, chlorides, free sulfur dioxide, total sulfur dioxide, sulphates, and alcohol.

c.)

Given the lower AIC and deviance, it would be best to use the logistic regression model. The logistic regression model also creates a simpler model with more relevant predictors, which we do not see in our multinomial models.

Question 11

a.)

The GLM components would be our response, s, the number of spikes recorded in the time interval. Our predictors would be speed v, acceleration a, and eye movement velocity e. We would use a count regression model as our best fit.

b.)

The best approach would be to remove these observations where there are no counts and form a zero-inflated model. We would then define a parameter α_i to account for structural zeros.

c.)

To incorporate an interaction, we would start by adding a new column to the regression model that is the multiplication of e and v and creating a new predictor variable.

d.)

Use of a classical poisson model would be rare. If we had a large amount of zeros, we would use a zero-inflation model. If overdispersion was present, which is likely, we would use a negative binomial model. If we had both zero-inflation and overdispersion, than we would use a ZINB model. If none of these situations occured, we could fit a classical poisson, assuming it would be a good fit.

Question 12

a.)

We could create a count regression model where we subtract 10 from all the values, and then use a mixture model for all the negative values. We would also need to create a zero-inflation model, since now there would be a larger amount of zeros because we are subtracting 10s from all the misrecordings and the legitimate 10s.

b.)

We could fit the model by testing for over dispersion, to decide whether a Poisson or a Negative Binomial model.

c.)

We would take the proportion of zeros and divide by the average of the proportion of 9s and 11s (before the transformation, -1s and 1s after). This way we could get an estimate of the values of 10s that were actually observed..