Statistical Models – November 2016 — Assignment 03

Chouliaras Georgios Christos , Jiayang Zhuo
Group 15

I. Computational Problems

Problem 1

(i) In order to investigate the association among the total number of *seizures* during the eight-week period and the predictors *treatment*, *baseline* and *age* it is appropriate to check firstly the plots between the response and the predictors, so as to do some visual observations. In figure (1) can be seen the plots between the number of seizures against the baseline rate and the age of the patient, as well as the box-plot about whether the treatment which was given to the patient was active or not (placebo). For the first figure (left) it can be observed that as the baseline rate grows, the total number of seizures grows also. The relationship is not perfectly linear, however these two variables are clearly linear dependent, hence it can be assumed that the predictor *baseline rate* might be significant. For the second figure (middle), there is not a visible pattern between the number of epileptic seizures and the age of the patient, but in general age usually affects the health condition , so this predictor might also be significant. However, conclusions about age's significance can't be conducted until a model is fitted to the data. On the third plot (right), there is the box-plot of the total number of seizures by treatment, where while the median is the same for the two cases, there is more variation in the number of epileptic seizures for the patients who took a placebo. This difference between the two cases, indicates that the treatment (being a placebo or not) actually plays a role to the number of seizures. Due to the fact that the response variable (total number of epileptic seizures) is not continuous, but a count, it is advisable to fit a Poisson regression model to these data. After fitting the aforementioned model with the canonical link function , $g(\mu) = log\mu$, and using as predictors all the available variables, the expected number of epileptic seizures was estimated and it is given to the table (1) along with the real number of seizures in order to check model's accuracy.

Figure 1: Informative plots indicating the association between number of seizures and the predictors.

patient	estimate	real	patient	estimate	real
1	18.11	14	30	29.60	28
2	17.71	14	31	14.62	7
3	14.10	11	32	14.94	13
4	18.95	13	33	13.97	19
5	51.76	55	34	17.94	11
6	24.93	22	35	24.11	74
7	18.53	12	36	18.33	20
8	59.20	95	37	14.28	10
9	27.28	22	38	43.68	24
10	16.54	33	39	25.26	29
11	51.67	66	40	13.33	4
12	25.52	30	41	16.76	6
13	17.72	16	42	20.07	12
14	41.13	42	43	36.34	65
15	91.48	59	44	22.04	26
16	39.35	16	45	31.68	39
17	19.85	6	46	12.45	7
18	177.12	123	47	24.68	32
19	21.73	15	48	13.64	3
20	17.73	16	49	20.55	13
21	17.71	14	50	27.04	26
22	13.79	14	51	27.63	10
23	21.24	13	52	34.77	70
24	23.29	30	53	26.38	13
25	48.29	143	54	17.92	15
26	21.22	6	55	17.94	51
27	13.49	10	56	17.15	6
28	33.56	53	57	18.33	0
29	51.25	42	58	18.33	10

Table 1: Estimate for the expected number of epileptic seizures.

From table (1) it can be seen that the accuracy of the model is not so good, since for many patients the estimate is very far from the real number of seizures. This is the case especially for when the number of seizures is very high or very low. It can be explained as follows: Due to the fact that there were not many observations with very low or very high number of seizures, the model cannot produce good estimates for these cases. However, for average number of seizures the accuracy of estimates is better.

Furthermore, approximate 95% confidence intervals for each regression coefficient were calculated and are provided in table (2):

Parameter	Lower Bound	Upper Bound
Intercept	1.66540716	2.21989965
Treatment	-0.25447993	-0.03986753
Baseline	0.02113686	0.02446812
Age	0.01459702	0.03076304

Table 2: Confidence Intervals of the regression coefficients

It can be noticed, that any of the confidence intervals includes zero and this is an indication that all of the predictors are significant. This must be verified though, through the p-values of the summary of the model as well as from the analysis of deviance table.

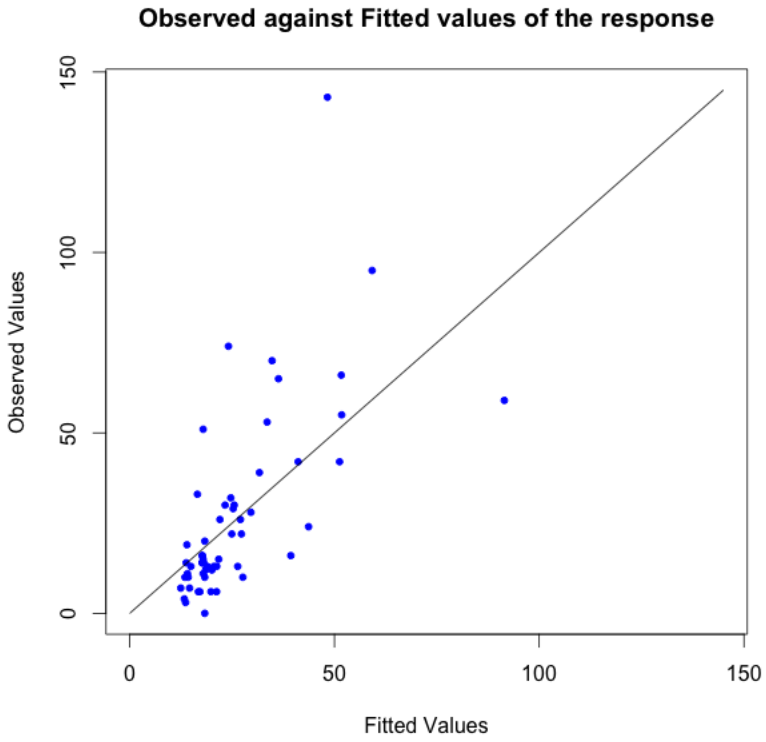


Figure 2: Observed values of the response against the fitted values.

Furthermore, in figure (2) can be seen the plot of the observed values of the response against the fitted values. An extreme outlier (18^{th} patient with observed value 123 and fitted value 177) has been removed from the plot. From the plot it can be observed that although there is a linear relationship between observed and fitted values, the fit is not so good as there are many data points which deviate from the line $y = x$. It can be observed that for large number of epileptic seizures, the deviation from the line is significant. This is due to the fact that there was a very limited number of observations with high number of epileptic seizures, hence the model cannot produce a good fit when the number of epileptic seizures is large.

(ii) In order to evaluate the significance of each predictor, the P-values need to be checked in both the summary of the model as well as the analysis of deviance table. For the analysis of deviance table, a *chi-square* test was conducted so as to check the p-value of each predictor. The *Fischer Scoring* method converged after 5 iterations and the two tables are provided below:

From tables (3) and (4) it can be seen that all the predictors are significant since their p-values are much less than the significance level $\alpha = 0.05$, hence we reject the null hypothesis that the predictors are not significant. Moreover, the p-values in both tables are almost identical with only exception the *Treatment* predictor which has a larger p-value in the summary of the model table, however still lower than the significance level α. Additionally, from table (4) it can be seen that the predictor *Baseline* explains most of the model's deviance, hence this predictor might be the

	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
(Intercept)	1.9427	0.1383	14.05	$2 \cdot 10^{-16}$
Treatment	-0.1472	0.0535	-2.75	0.0060
Baseline	0.0228	0.0008	27.45	$2 \cdot 10^{-16}$
Age	0.0227	0.0040	5.63	$1.85 \cdot 10^{-8}$

Table 3: Table obtained from the summary of the model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			57	1284.72	
Treatment	1	68.26	56	1216.46	$2 \cdot 10^{-16}$
Baseline	1	625.80	55	590.66	$2 \cdot 10^{-16}$
Age	1	31.27	54	559.39	$2.246 \cdot 10^{-8}$

Table 4: Analysis of deviance table with chi-square test.

most significant for the model. Furthermore, using the estimates for the coefficients from table (3) the resulting estimated regression function is:

$$log(\mu_i) = -1.9427 - 0.1472 \cdot \text{Treatment} + 0.0228 \cdot \text{Baseline} + 0.0227 \cdot \text{Age} \tag{1}$$

(iii) The plots of deviance and Pearson residuals against the fitted values can be seen in figure (3). There is a great similarity between the two plots as the points in both plots are spread in a similar way and also in both plots there are two outliers for large fitted values. It can be seen that, the points are not evenly spread around 0 and this might be an indicator for bad fit. However, this visual observation is not enough to assess the goodness of fit and further analysis needs to be conducted.

The *Pearson chi-squared statistic* P is obtained by summing the squared Pearson residuals and is the following:

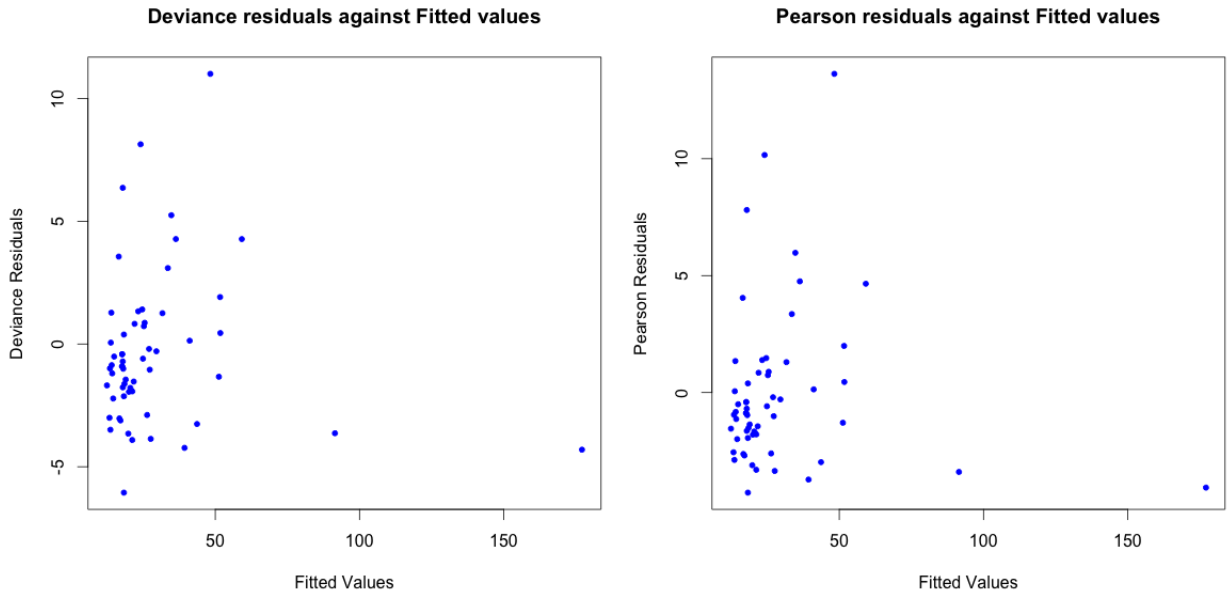
$$\text{P} = 646.1294$$

Similarly the *deviance* D is obtained by summing the squared Deviance residuals:

$$\text{D} = 559.391$$

In order to check the goodness of fit, P/ϕ and D/ϕ need to be compared to $\chi^2_{n-p-1,1-\alpha}$. For the Poisson model $\phi = 1$ and hence, P and D need to be compared to $\chi^2_{54,0.95} = 72.15322$. Due to the fact that, both values are larger than $\chi^2_{54,0.95}$, the null hypothesis that the model fits the data well, is rejected. Moreover, the values of P and D are not approximately equal and this is another indicator of bad quality of fit.

(iv) An interesting investigation is to check whether a different link function could improve the fit. Hence, the data were fitted again with a Poisson model but this time with the link function $g(u) = \sqrt{u}$. After fitting the data with this model, the summary of the model as well as the analysis of deviance table with a chi-square test were obtained again in order to compare the results with the previous model. The aforementioned tables can be seen below:



(a) Plot of the Deviance residuals against the fitted values *(b) Plot of the Pearson residuals against the fitted values*

Figure 3: Deviance and Pearson residuals against fitted values

Parameter	Estimate	Std. Error	z value	P value
Intercept	1.395404	0.349193	3.996	$6.44 \cdot 10^{-5}$
Treatment	-0.556096	0.132350	-4.202	$2.65 \cdot 10^{-5}$
Baseline	0.082097	0.003066	26.774	$2 \cdot 10^{-16}$
Age	0.053133	0.010663	4.983	$6.27 \cdot 10^{-7}$

Table 5: Summary of the model with link function $g(u) = \sqrt{u}$.

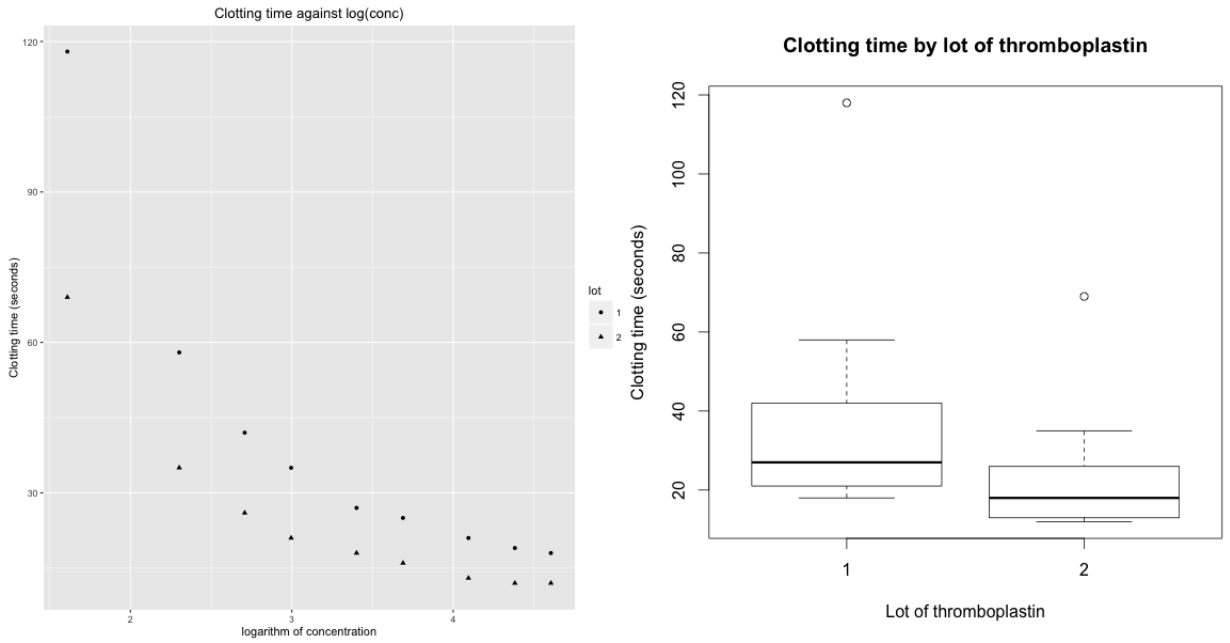
Parameter	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			57	1284.72	
Treatment	1	68.26	56	1216.46	$2.2 \cdot 10^{-16}$
Baseline	1	710.89	55	505.57	$2.2 \cdot 10^{-16}$
Age	1	23.85	54	481.72	$1.042 \cdot 10^{-6}$

Table 6: Analysis of Deviance table for the model with link function $g(u) = \sqrt{u}$.

From a comparison between the new and the former tables it can be seen that again all the predictors are significant since the p-values are again extremely low, specifically *Treatment* before had a p-value 0.00596 which became $2.65 \cdot 10^{-5}$ using the new model. An important observation is that the AIC of the new model is 765.44 which is lower compared to the former AIC = 843.11 and this is an indication that the model with the link function $g(u) = \sqrt{u}$ performs better. Furthermore, the new Pearson chi-squared statistic is $P = 550.73$ and the new deviance is $D = 481.7244$, which are both lower than before. However, P and D are still larger than $\chi^2_{54,0.95}$, so despite the fact that the fit is better than before, it is still a bad fit in general. It needs to noted, that a reason that this bad fit might occurred is the limited number of patients, as both Pearson chi-squared statistic and Deviance require a large number of observations in order for the inference to be valid.

Problem 2

(i) In order to investigate the association between *clotting time* and the predictors *log(conc)* and *lot* it is advisable to observe firstly this relationship visually. The plot of the clotting time against the logarithm of blood concentration using different symbol for each log along with a box-plot of clotting time by lot of thromboplastin can be seen in figure (4).



(a) Plot of clotting time against the logarithm of concentration with different symbol for each lot.

(b) Box-plot of clotting time and the two different lots of thromboplastin.

Figure 4: Informative plots

The relationship between clotting time and the logarithm of concentration in figure (4a) follows an exponential decay. Moreover, it can be observed that the clotting which was induced by the 1^{st} lot of thromboplastin has larger times than the clotting by the 2^{nd} lot, for the same concentration level. This is also noticeable by the box-plot in figure (4b) where it is clear that the clotting times for the first lot are larger and have a bigger variation compared to the clotting times of the second lot. From these observations, a Gamma model with the multiplicative inverse link function might be a good fit for these data. After fitting the aforementioned model to the data, the estimates for the expected clotting time for each individual have been obtained and are provided below, along with the real clotting times in order to evaluate the prediction.

Individual	Estimated time	Real time	Lot
1	140.29	118	1
2	51.45	58	1
3	37.54	42	1
4	31.50	35	1
5	25.67	27	1
6	22.70	25	1
7	19.51	21	1
8	17.74	19	1
9	16.57	18	1
10	55.56	69	2
11	32.99	35	2
12	26.66	26	2
13	23.46	21	2
14	20.07	18	2
15	18.20	16	2
16	16.09	13	2
17	14.87	12	2
18	14.04	12	2

Table 7: Estimates for the expected clotting time for each individual.

It can be seen that the estimated clotting times do not deviate much from the real times and this is an indication that the model fits the data well. However, it needs to be verified by appropriate statistics. Moreover, it can be seen that the model tends to underestimate the time of the clotting which was induced by lot 1 and overestimates the clotting time for the lot 2.

The approximate 90% confidence intervals for each regression coefficient are shown in the table (8).

Parameter	Lower Bound	Upper Bound
Intercept	-0.025283679	-0.01761569
log(conc)	0.015963341	0.01954939
lot	0.007450342	0.01428658

Table 8: Confidence Intervals of the regression coefficients

From this table it is can be seen that zero does not lies in the confidence interval of any of the predictors, thus it implies that both predictors are significant for the model.

The plot of the observed values of the response against the fitted values can be seen in figure (5) along with the line y=x in order to observe deviation from this line. The majority of the points are close to this line, something that indicates that the model fits the data nicely. Moreover, the points for the first and second lot are similarly distributed across the y=x line with only one point of the first lot to deviate from the line in some extent.

(ii) In order to investigate the significance of each predictor the table from the summary of the model was obtained, along with the analysis of deviance table in which an F test was conducted for the significance of each predictor. These tables are provided below.

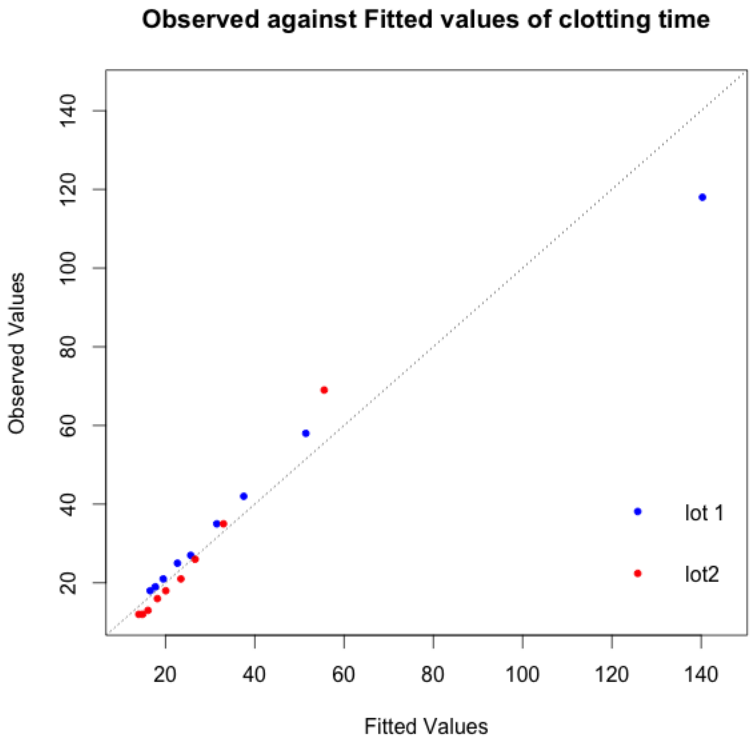


Figure 5: Observed against fitted values of the response with different color for each lot.

Parameter	Estimate	t value	P value
Intercept	-0.021450	-9.808	$6.45 \cdot 10^{-8}$
log(conc)	0.017756	17.361	$2.43 \cdot 10^{-11}$
lot	0.010868	5.574	$5.32 \cdot 10^{-5}$

Table 9: Estimates, t-values and P-values from the table obtained by the summary of the model.

Parameter	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL			17	7.7087		
log(conc)	1	6.6904	16	1.0183	341.598	$9.868 \cdot 10^{-12}$
lot	1	0.7178	15	0.3004	36.651	$2.207 \cdot 10^{-5}$

Table 10: Analysis of Deviance table for the model.

It can be observed that in both tables all the predictors have p-values much lower than the significance level $\alpha = 0.05$ and thus, all the predictors are significant. Moreover, from table (10) it can be seen that the logarithm of the concentration reduces the deviance by 6.6904 compared to the predictor lot which reduced the deviance by only 0.7178, so concentration is more significant predictor, however the p-value of the predictor lot is low enough, hence there is no reason to drop it from the model. Also, the p-values have been calculated manually using t-quantiles, and they are identical to the p-values of table (9) (the calculation can be seen in the R code in the Appendix).

(iii) The Deviance residuals against the Pearson residuals have been plotted in figure (6). It can be seen that there is a linear relationship between the residuals and almost all the points lie on the line y=x, something that indicates that the Pearson and Deviance statistics might be similar. This

can be confirmed by the fact that, the Pearson chi-squared statistic is $P = 0.294$ and the Deviance $D = 0.3$ which are indeed very close, and this is an indication of good fit quality. Although, in order to assess the quality of fit, P/ϕ and D/ϕ need to be compared to $\chi^2_{15,0.95}$. Due to the fact that the model's distribution is Gamma, ϕ is unknown and thus it can be replaced by the estimate $\hat{\phi}$ calculated as:

$$\hat{\phi} = \frac{P}{n-p-1} = \frac{0.294}{15} = 0.196$$

Hence, both $P/\hat{\phi} = 15$ and $D/\hat{\phi} = 15.33887$ are lower than $\chi^2_{15,0.95} = 24.99579$, something that verifies the argument for good fit.

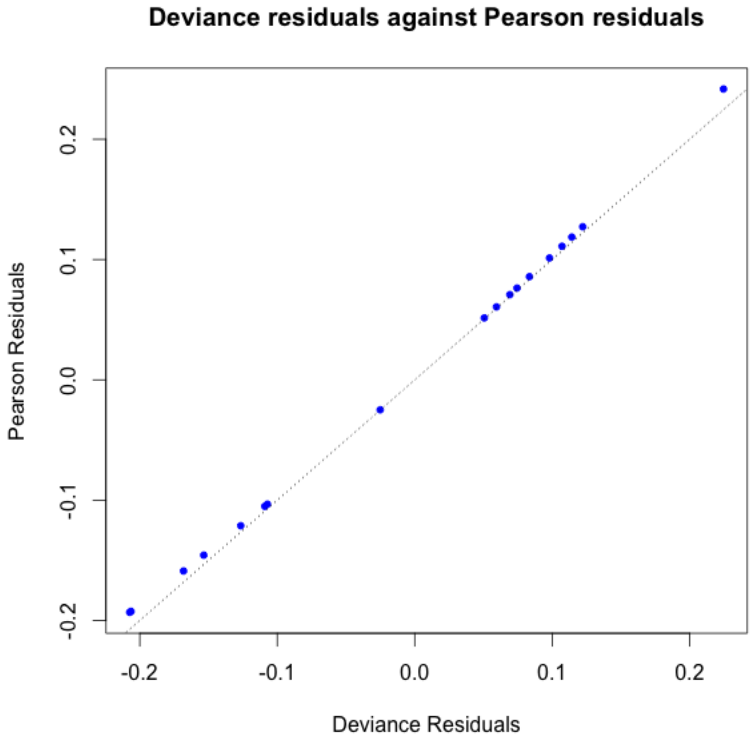
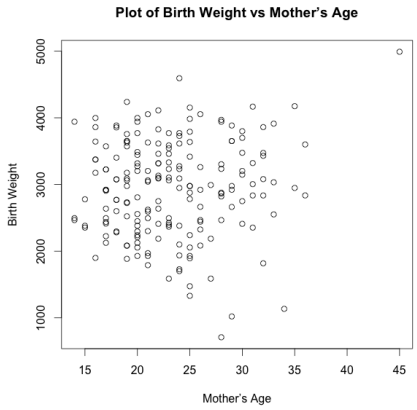


Figure 6: Plot of Deviance residuals against Pearson residuals.

Problem 3

(i) In order to observe how the different predictors affect the infant's weight several informative plots have been created, so as to make some visual observations.



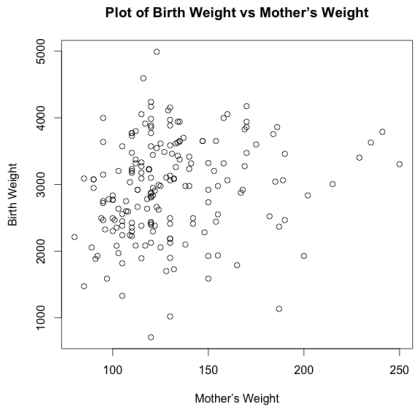
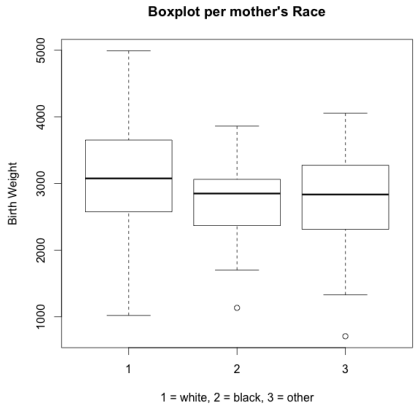


Figure 7: Plot of Birth weight against mother's age *Figure 8: Plot of Birth weight against mother's weight*



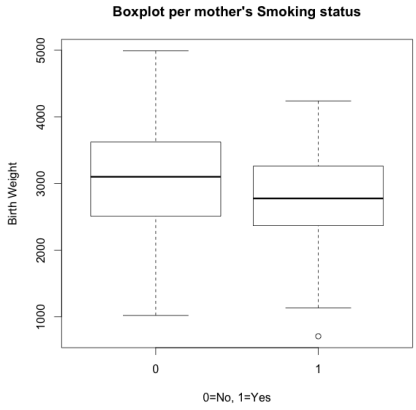


Figure 9: Box-plot per mother's race *Figure 10: Box-plot per mother's smoking status*

Firstly, from figure 7 concerning the mother's *age* if observed carefully, it can be seen that for weights lower than 2500 grams, the birth weight tends to be lower, as the age of the mother gets larger. There is also one outlier in the data for mother's age 45 years old, which might occurred due to the large number of years of the mother. From this plot one cannot conclude much, and further data analysis needs to be conducted in order to investigate the significance of mother's age. Similarly, from figure (8) regarding the relation between birth weight and *mother's weight* there is not any noticeable pattern, except from the fact that the weight of the infant tends to be analogously larger with mother's weight if the mother weights over 200 pounds. This is an indication of a small positive correlation between the weight of the infant and its' mother's weight. From the box-plot in figure (9) it can be seen that there are only a few insignificant differences between the three categories of mother's *race*, hence this predictor might be insignificant for the birth weight. Similarly, from figure (10) the median, as well as the variation for the non smoking

mothers seems to be slightly larger than the smoking mothers. From this it can be assumed that the predictor *smoking status* might by slight significant for the model. Furthermore, from figure (11) concerning the mother's *premature labor history* there was only one observation with three premature labors and only a few with two, hence there are not enough observations to obtain safe conclusions. However, by comparing the cases of 0 and 1 occurrences where there are more data points, it can be seen that mothers who didn't have any premature labor have a larger birth weight than those who had one premature labor. This is an indication that this predictor plays a role to infant's weight. From the box-plot in figure (12) which is about the mother's *hypertension history* it can be observed that the birth weight is larger and with less variation if the mother had no hypertension history. Hence, this predictor might be significant as it affects the birth weight. A very similar box-plot can be seen in figure (13) concerning the mother's *uterine irritability status* where again, the birth weight tends to be larger if the mother had no uterine irritability and this is an indication that this predictor might be significant. Lastly, in the box-plot of figure (14) there are no safe conclusions which can be made as the number of observations for a large number of physician visits is limited. However it can be seen that in general there are not significant fluctuations in the birth weight as the number of visits gets larger. Hence, this is an indication that this factor might be insignificant. However, the significance of each factor has to be checked by checking their p-values.

In order to perform analysis in the data, after fitting a binomial model using the *logit* link function, having as response variable the *Low Birth Weight* and as predictors the variables *MotherAge, MotherWeight, Race, Smoker, PremLabor, Hypertension, UterinIrrit* and *PhysicianVisis* the *Fischer scoring* method converged after 4 iterations to the following estimates:

Parameter	Estimate
Intercept	0.48062320
MotherAge	-0.02954903
MotherWeight	-0.01542428
Race2	1.27225979
Race3	0.88049592
Smoker	0.93884570
PremLabor	0.54333703
Hypertension	1.86330287
UterineIrrit	0.76764814
PhysicianVisits	0.06530183

Table 11: Parameters estimates of the full model.

In order to check the significance of each predictor, an analysis of deviance table was created along with a chi-squared test in order to check the p-values for each predictor. It has to be noted that the predictors which are assumed to be significant are these which reduce significantly the model's deviance. The analysis of deviance table can be seen in table (12).

Figure 11: Box-plot per Premature Labor history

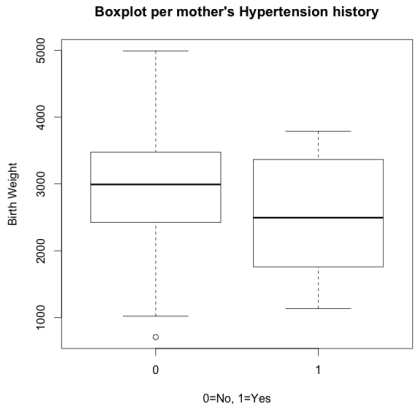


Figure 12: Box-plot per Hypertension history

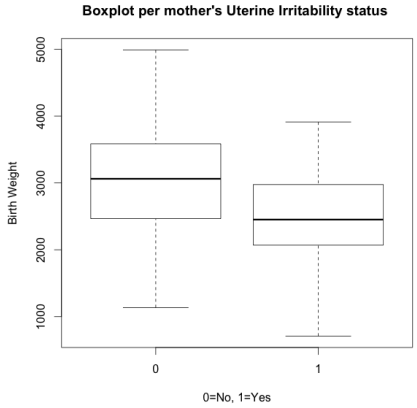


Figure 13: Box-plot per Uterine Irritability Status

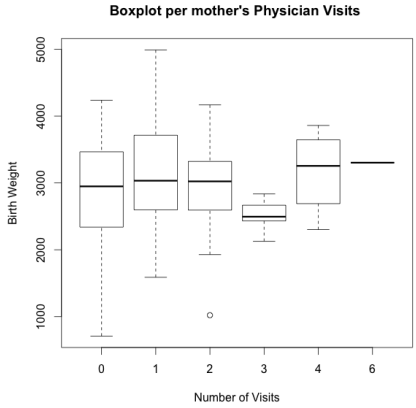


Figure 14: Box-plot per mother's physician visits.

Parameter	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			188	234.67	
Mother Age	1	2.7600	187	231.91	**0.096646**
Mother Weight	1	4.7886	186	227.12	**0.028649**
Race	2	4.4628	184	222.66	0.107381
Smoker	1	8.0834	183	214.58	**0.004467**
Premature Labor	1	3.8993	182	210.68	**0.048307**
Hypertension	1	6.5715	181	204.11	**0.010362**
Uterine Irritability	1	2.6795	180	201.43	0.101647
Physician Visits	1	0.1422	179	201.28	0.706147

Table 12: Analysis of Deviance table for the full model. P-values below 0.1 are shown in bold.

As it can be seen from table (12) the parameters with p-value below 0.1 are: *MotherAge, MotherWeight, Smoker, PremLabor* and *Hypertension*. These predictors are used in order to create a reduced model in which the insignificant parameters *Race*, *UterinIrrit* and *PhysicianVisis* are omitted. It can be observed that for most of the predictors, the smaller the p-value, the larger

the reduction that they produce in the deviance. All the predictors that are omitted, have a low reduction in the deviance. In order to compare the full with the reduced model *Akaike Information Criterion*(AIC) as well as *Bayes Information Criterion*(BIC) were used and their values are provided below:

Model	AIC	BIC
full	221.2848	253.7023
reduced	223.3278	242.7782

Table 13: AIC and BIC values for both models.

It can be seen that AIC for the full model is slightly smaller than in the reduced model. In principle, one should choose the model which minimizes AIC however in this case the difference in the AIC values is very small and also the reduced model has 3 less predictors something that makes it simpler. Moreover, BIC value for the full model is larger than the one for the reduced model, something that indicates that the reduced model is better. This difference between the two criteria lies in the fact that BIC uses $p\ log(n)$ as a penalty term, instead of $2p$ which is used by AIC. This means that BIC punishes the complexity of the model more than AIC and thus, it suggests that the simpler, reduced model is more suitable for modelling the data than the full model. Due to the aforementioned, and given the fact that the difference in AIC values is very small, it can be concluded that the reduced model is the one that should be used for the data, as it has a better balance between performance and complexity than the full model.

(ii) Using the reduced model, the estimated probabilities of low birth weight can be seen in table (14), which can be found in the Appendix due to its extended length. The *Prediction* column in this table was created as follows: if the estimated probability is above 0.5 then the prediction gets the value 1, otherwise it is 0. This column was created in order to be able to compare it with the real values of *Low Birth Weight* and thus, to be able to compute the accuracy of the reduced model.

The accuracy of the reduced model was computed 71.42 % which indicates a relatively good fit. In comparison, the accuracy of the full model was computed 74.07 % which is slightly better than the one of the reduced model. The better predictive performance of the full model is due to the larger number of predictors in the model, however as stated when comparing the AIC values, this difference is not large enough to sacrifice complexity for performance. Moreover, in order to assess the quality of fit, the Pearson residuals are plotted against the Deviance residuals in order to check visually if they are similar. From figure (15) it can be seen that there is a linear relationship between the residuals and the points of the plot follow the y=x line with only a few points to deviate from it. Hence, this graph indicates that the Pearson chi-squared statistic P might have similar value as the deviance D.

To investigate this, the Pearson chi-squared statistic was computed $P = 188.37$, while the deviance $D = 211.32$. The values are close, but not approximately equal, however this is not enough to assess the quality of the fit, since it should be evaluated by comparing P/ϕ and $D\ /\phi$ with $\chi^2_{183,0.95} = 215.56$. Since the distribution is Binomial, $\phi = 1$, hence P and D are compared directly to $\chi^2_{183,0.95}$. Both of them are lower than $\chi^2_{183,0.95}$, thus the H_0 :*the model fits the data well* is not rejected and as a consequence the quality of the fit can be considered good.

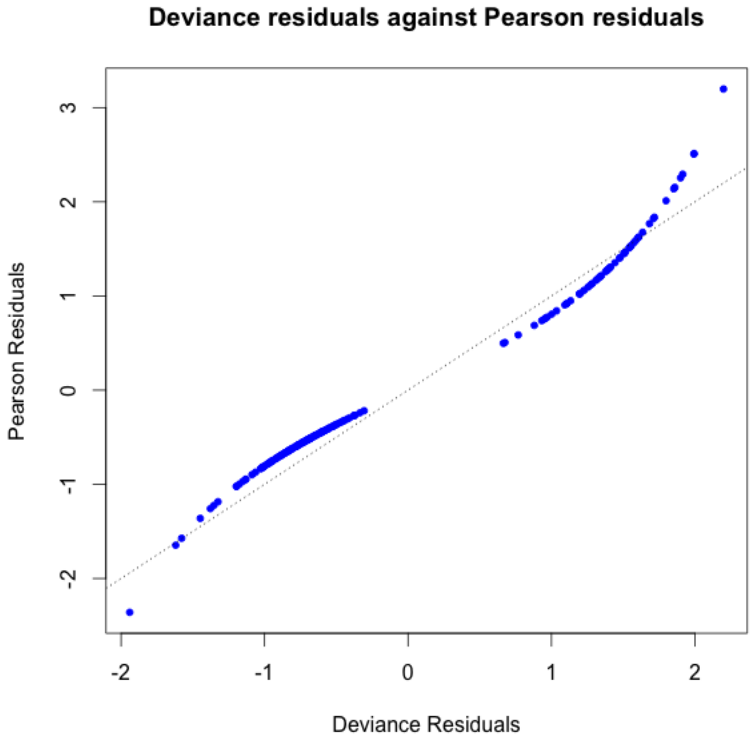


Figure 15: Deviance residuals against Pearson residuals

II. Theoretical Problems

Problem 1

Using the fact that $Y_i = Y_1, \ldots, Y_n$ has probability density function f_i of a one-parameter exponential family of distributions which can be written in *canonical form* along with the definition of expectation the following holds:

$$\mu_i = \mathbb{E}Y_i = \int_{-\infty}^{\infty} y f_i(y, \theta_i) dy \tag{2}$$

$$= \int_{-\infty}^{\infty} y \; exp(\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c(y, \phi/A_i)) \tag{3}$$

But due to the fact that $f_i(y, \theta_i)$ is a probability density function it holds that:

$$\int_{-\infty}^{\infty} f_i(y, \theta_i) dy = 1 \Leftrightarrow \tag{4}$$

$$\int_{-\infty}^{\infty} exp(\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c(y, \phi/A_i)) = 1 \tag{5}$$

Taking derivatives to both sides of equation (4) with respect to θ_i:

$$\int_{-\infty}^{\infty} \frac{y - b'(\theta_i)}{\phi/A_i}\ exp(\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c(y, \phi/A_i)) = 0 \Leftrightarrow$$
$$\int_{-\infty}^{\infty} \frac{A_i}{\phi}(y - b'(\theta_i))f(y, \theta_i)dy = 0 \Leftrightarrow$$
$$\int_{-\infty}^{\infty} \frac{A_i}{\phi} y f(y, \theta_i) - \frac{A_i}{\phi} b'(\theta_i)\ f(y, \theta_i)dy = 0 \Leftrightarrow$$
$$\frac{A_i}{\phi}\int_{-\infty}^{\infty} y f(y, \theta_i) = \frac{A_i}{\phi} b'(\theta_i)\int_{-\infty}^{\infty} f(y, \theta_i)dy \Leftrightarrow$$
$$\int_{-\infty}^{\infty} y f(y, \theta_i)dy = b'(\theta_i)\int_{-\infty}^{\infty} f(y, \theta_i)dy \tag{6}$$

Now, using equations (2), (4), equation (6) becomes:

$$\mathbb{E}Y_i = b'(\theta_i) \cdot 1 \Leftrightarrow$$
$$\mathbb{E}Y_i = \mu_i = b'(\theta_i) \qquad i = 1, \ldots, n \tag{7}$$

□

For the variance of a random variable X holds that:

$$VarX = \mathbb{E}[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx \tag{8}$$

Hence, for $VarY_i$ holds:

$$\begin{aligned}
VarY_i &= \int_{-\infty}^{\infty} (y - \mathbb{E}Y_i)^2 f(y, \theta_i)dy \\
&= \int_{-\infty}^{\infty} (y^2 - 2y\mathbb{E}Y_i + (\mathbb{E}Y_i)^2 f(y, \theta_i)dy \\
&= \int_{-\infty}^{\infty} y^2 f(y, \theta_i)dy - 2\mathbb{E}Y_i \int_{-\infty}^{\infty} y f(y, \theta_i)dy + (\mathbb{E}Y_i)^2 \int_{-\infty}^{\infty} f(y, \theta_i)dy \\
&\overset{(3),(4)}{=\!=} \int_{-\infty}^{\infty} y^2 f(y, \theta_i)dy - 2\mathbb{E}Y_i\mathbb{E}Y_i + (\mathbb{E}Y_i)^2 \\
&= \int_{-\infty}^{\infty} y^2 f(y, \theta_i)dy - 2(\mathbb{E}Y_i)^2 + (\mathbb{E}Y_i)^2 \\
&= \int_{-\infty}^{\infty} y^2 f(y, \theta_i)dy - (\mathbb{E}Y_i)^2 \\
&\overset{(7)}{=} \int_{-\infty}^{\infty} y^2 f(y, \theta_i)dy - (b'(\theta_i))^2
\end{aligned} \tag{9}$$

From equation (6) it holds that:

$$\mathbb{E}Y_i = \int_{-\infty}^{\infty} y f(y, \theta_i)dy = b'(\theta_i) \tag{10}$$

In order to obtain an expression for $b''(\theta_i)$, derivative with respect to θ_i needs to be taken in equation (10):

$$\begin{aligned}
\int_{-\infty}^{\infty} y f'(y,\theta_i)dy = b''(\theta_i) \Leftrightarrow \\
\int_{-\infty}^{\infty} y(\frac{y\theta_i - b'(\theta_i)}{\phi/A_i}) f(y,\theta_i)dy = b''(\theta_i) \Leftrightarrow \\
\int_{-\infty}^{\infty} \frac{y^2}{\phi/A_i} f(y,\theta_i)dy - \int_{-\infty}^{\infty} \frac{yb'(\theta_i)}{\phi/A_i} f(y,\theta_i)dy = b''(\theta_i) \Leftrightarrow \\
\frac{1}{\phi/A_i}\int_{-\infty}^{\infty} y^2 f(y,\theta_i)dy = b''(\theta_i) + \frac{b'(\theta_i)}{\phi/A_i}\int_{-\infty}^{\infty} y f(y,\theta_i)dy \Leftrightarrow \\
\int_{-\infty}^{\infty} y^2 f(y,\theta_i)dy = \frac{\phi}{A_i} b''(\theta_i) + b'(\theta_i)\mathbb{E}Y_i \overset{(10)}{\Leftrightarrow} \\
\int_{-\infty}^{\infty} y^2 f(y,\theta_i)dy = \frac{\phi}{A_i} b''(\theta_i) + (b'(\theta_i))^2 \qquad (11)
\end{aligned}$$

Due to (11), equation (9) becomes:

$$\begin{aligned}
VarY_i = \frac{\phi}{A_i} b''(\theta_i) + (b'(\theta_i))^2 - (b'(\theta_i))^2 \Leftrightarrow \\
VarY_i = \frac{\phi}{A_i} b''(\theta_i) \qquad i = 1, \ldots, n \qquad (12)
\end{aligned}$$

□

Problem 2

The Poisson probability mass function $f(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$ needs to be written in exponential form by using the property $e^{logx} = x$. Hence:

$$f(y;\lambda) = exp(log(f(y;\lambda))) = exp\left(log\left(\frac{e^{-\lambda}\lambda^y}{y!}\right)\right) \qquad (13)$$

Now, using the properties of the logarithm $log(\alpha\beta) = log\alpha + log\beta$ and $log(\frac{\alpha}{\beta}) = log\alpha - log\beta$ equation (13) becomes:

$$\begin{aligned}
exp(log(e^{-\lambda}) + log(\lambda^y) - log(y!)) = exp(-\lambda loge + ylog\lambda - log(y!)) \\
\overset{loge=1}{=\!=} exp\left(\frac{ylog\lambda - \lambda}{1/1} + (-log(y!))\right) \qquad (14)
\end{aligned}$$

By comparing equation (14) with equation (3.9) from the lecture notes the following quantities can be obtained:

$$log\lambda = \theta \qquad (15)$$

$$b(\theta) = \lambda \qquad (16)$$

$$c(y, \phi/A_i) = -log(y!) \qquad (17)$$

$$\phi = 1 \qquad (18)$$

$$A = 1 \qquad (19)$$

It needs to be noted that $\phi = 1$ due to the fact that the distribution is Poisson. From equation (15) it can be seen that the canonical link function is:

$$\eta = \theta = g(\mu) = log(\mu) \tag{20}$$

Now, using equation (3.10) from lecture notes and the fact that $\lambda = e^{\theta}$ which follows from equation (15) it can be seen that for $\mathbb{E}Y$ and $VarY$ holds:

$$\mathbb{E}Y = \mu = b'(\theta) \overset{(16)}{=} (e^{\theta})' = e^{\theta} = e^{log\lambda} = \lambda \tag{21}$$

$$VarY = b''(\theta)\phi/A = (e^{\theta})''1/1 = e^{\theta} = e^{log\lambda} = \lambda \tag{22}$$

Hence it follows that:

$$\mathbb{E}Y = VarY = \lambda \tag{23}$$

Now assuming that one uses Poisson regression with the canonical link, the form of the i-th diagonal element of the weight matrix W is:

$$w_{ii} = \frac{A_i}{[g'(\mu_i)]^2 b''(\theta_i)} \tag{24}$$

But $A_i = 1$, $g'(\mu_i) = \frac{1}{\mu_i} \overset{(21)}{=} \frac{1}{\lambda_i}$ and $b''(\theta_i) \overset{(22)}{=} \lambda_i$ Hence, equation (24) becomes:

$$w_{ii} = \frac{1}{(\frac{1}{\lambda_i})^2 \lambda_i} = \frac{1}{\frac{1}{(\lambda_i)^2}\lambda_i} = \frac{1}{\frac{1}{\lambda_i}} = \lambda_i \tag{25}$$

Problem 3

Deviance D is defined as the scaled log-likelihood-ratio statistic:

$$D = 2\phi[l(\tilde{\theta}) - l(\hat{\theta})] \tag{26}$$

which can be rewritten as:

$$\begin{aligned} D &= 2\phi[log(\mathcal{L}(\tilde{\theta})) - log(\mathcal{L}(\hat{\theta}))] \\ &= 2\phi \left[log\left(\frac{\mathcal{L}(\tilde{\theta})}{\mathcal{L}(\hat{\theta})}\right)\right] \\ &= -2\phi \left[log\left(\frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\tilde{\theta})}\right)\right] \end{aligned} \tag{27}$$

where $\mathcal{L}(\hat{\theta})$ is the likelihood of the fitted model and $\mathcal{L}(\tilde{\theta})$ is the likelihood of the saturated model. Moreover, since there is one parameter π_j and $Y_1, \ldots, n$ take binary values, the underlying data are Bernoulli data:

$$Y_j \sim Bern(\pi_j) \tag{28}$$

But, since $Bernoulli(\pi_j)$ is equal with $Binomial(1, \pi_j)$ for the estimate of the saturated model $\tilde{\pi_j}$ holds that:

$$\begin{aligned} &\hat{Y_j} = 1 \cdot \tilde{\pi}_j \Leftrightarrow \\ &\tilde{\pi}_j = \frac{\hat{Y_j}}{1} = \frac{Y_j}{1} = \begin{cases} 1 & \text{if } Y_j = 1 \\ 0 & \text{if } Y_j = 0 \end{cases} \end{aligned} \tag{29}$$

The likelihood for the saturated model is:

$$\mathcal{L}(\tilde{\pi}_j|Y_j) = \prod_{j=1}^{n} \tilde{\pi}_j^{Y_j}(1-\tilde{\pi}_j)^{1-Y_j} \tag{30}$$

For the individual j if $Y_j = 1$ then from equation (29) $\tilde{\pi}_j = 1$ and:

$$\mathcal{L}(\tilde{\pi}_j|Y_j) = \prod_{j=1}^{n} 1^1 \cdot \underbrace{0^0}_{1} = 1 \tag{31}$$

Likewise, if $Y_j = 0$ then $\tilde{\pi}_j = 0$:

$$\mathcal{L}(\tilde{\pi}_j|Y_j) = \prod_{j=1}^{n} 0^0 \cdot 1^1 = 1 \tag{32}$$

Hence, the likelihood for the saturated model is :

$$\mathcal{L}(\tilde{\pi}_j|Y_j) = \mathcal{L}(\tilde{\theta}) = \prod_{j=1}^{n} 1 = 1 \tag{33}$$

and

$$l(\tilde{\theta}) = log(\mathcal{L}(\tilde{\theta})) = log1 = 0 \tag{34}$$

Hence using equation (27) and the fact that $\phi = 1$ for Binomial data, the formula for the deviance becomes:

$$D = -2log(\mathcal{L}(\hat{\theta})) = -2l(\hat{\theta}) \tag{35}$$

The probability mass function for the Bernoulli distribution is:

$$f(Y_j) = \pi_j^{Y_j}(1-\pi_j)^{1-Y_j} \tag{36}$$

[1] According to R, Google and Matlab, $0^0 = 1$

The probability mass function needs to be reformed to the canonical form in order to obtain the necessary quantities:

$$\begin{aligned} f(Y_j) &= exp(log(\pi_j^{Y_j}(1-\pi_j)^{1-Y_j})) \\ &= exp(log(\pi_j^{Y_j}) + log((1-\pi_j)^{1-Y_j}) \\ &= exp(Y_j log(\pi_j) + (1-Y_j) log((1-\pi_j)) \\ &= exp(Y_j log(\pi_j) + log((1-\pi_j) - Y_j log((1-\pi_j)) \end{aligned} \tag{37}$$

But, using the fact that $\pi_j = \frac{exp(x_j^T\beta)}{1+exp(x_j^T\beta)}$, $log(\pi_j)$ can be written as:

$$\begin{aligned} log(\pi_j) &= log\left(\frac{e^{x_j^T\beta}}{1+e^{x_j^T\beta}}\right) \\ &= log(e^{x_j^T\beta}) - log(1+e^{x_j^T\beta}) \\ &= x_j^T\beta - log(1+e^{x_j^T\beta}) \end{aligned} \tag{38}$$

But,

$$\begin{aligned} 1-\pi_j &= 1 - \frac{e^{x_j^T\beta}}{1+e^{x_j^T\beta}} \\ &= \frac{1+e^{x_j^T\beta} - e^{x_j^T\beta}}{1+e^{x_j^T\beta}} \\ &= (1+e^{x_j^T\beta})^{-1} \Leftrightarrow \\ 1+e^{x_j^T\beta} &= (1-\pi_j)^{-1} \end{aligned} \tag{39}$$

Using this result in equation (38):

$$\begin{aligned} log(\pi_j) &= x_j^T\beta - log((1-\pi_j)^{-1}) \\ &= x_j^T\beta - log(\frac{1}{1-\pi_j}) \\ &= x_j^T\beta - (log(1) - log(1-\pi_j)) \\ &= x_j^T\beta + log(1-\pi_j) \end{aligned} \tag{40}$$

Hence, equation (37) becomes:

$$\begin{aligned} f(Y_i) &= exp(Y_j(x_j^T\beta + log(1-\pi_j)) + log((1-\pi_j) - Y_j log((1-\pi_j)) \\ &= exp(Y_j x_j^T\beta + Y_j log(1-\pi_j) + log((1-\pi_j) - Y_j log((1-\pi_j)) \\ &= exp(Y_j x_j^T\beta + log((1-\pi_j)) \end{aligned} \tag{41}$$

By comparing equation (41) to the canonical form, the necessary quantities are obtained:

$$\theta_j = x_j^T \beta \tag{42}$$
$$b(\theta_j) = -log(1 - \pi_j) \tag{43}$$
$$\phi = 1 \tag{44}$$
$$A_j = 1 \tag{45}$$
$$c(y, \phi/A_j) = 0 \tag{46}$$

The log-likelihood function for the fitted model is given by the formula:

$$l(\hat{\theta}) = l(\hat{\beta}) = \sum_{j=1}^{n} \left(\frac{Y_j \theta_j - b(\theta_j)}{\phi/A_j} + c(Y_j, \phi/A_j) \right) \tag{47}$$

By replacing the quantities from equations (42) - (46) and using the fact that $\hat{\pi}_j = \pi_j(\hat{\beta})$ the formula for the log-likelihood for the fitted model becomes:

$$\begin{aligned} l(\hat{\theta}) &= \sum_{j=1}^{n} \left(\frac{Y_j x_j^T \hat{\beta} + log(1 - \hat{\pi}_j)}{1/1} + 0 \right) \\ &= \sum_{j=1}^{n} Y_j x_j^T \hat{\beta} + \sum_{j=1}^{n} log(1 - \hat{\pi}_j) \\ &= Y^T X \hat{\beta} + \sum_{j=1}^{n} log(1 - \hat{\pi}_j) \end{aligned} \tag{48}$$

Plugging in this result to equation (35) the desired result can be obtained:

$$D = -2 \left[Y^T X \hat{\beta} + \sum_{j=1}^{n} log(1 - \hat{\pi}_j) \right] \tag{49}$$

□

The maximum likelihood estimator of $\hat{\beta} : (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$ is the value obtained by maximizing $l(\hat{\beta})$ with respect to $\hat{\beta}$ and can be found by setting:

$$\frac{\partial l(\hat{\beta})}{\partial \hat{\beta}} = 0 \tag{50}$$

But, for $l(\hat{\beta})$ holds:

$$
\begin{aligned}
l(\hat{\beta}) &= Y^T X\hat{\beta} + \sum_{j=1}^{n} log(1-\hat{\pi}_j) \\
&= Y^T X\hat{\beta} + \sum_{j=1}^{n} log\left(1 - \frac{e^{x_j^T\hat{\beta}}}{1+e^{x_j^T\hat{\beta}}}\right) \\
&= Y^T X\hat{\beta} + \sum_{j=1}^{n} \left(\frac{1+e^{x_j^T\hat{\beta}} - e^{x_j^T\hat{\beta}}}{1+e^{x_j^T\hat{\beta}}}\right) \\
&= Y^T X\hat{\beta} - \sum_{j=1}^{n} log(1+e^{x_j^T\hat{\beta}}) \qquad (51)
\end{aligned}
$$

Now, using equation (50):

$$
\begin{aligned}
\frac{\partial l(\hat{\beta})}{\partial\hat{\beta}} = 0 \Leftrightarrow \\
Y^T X - \sum_{j=1}^{n} log(1+e^{x_j^T\hat{\beta}}) = 0 \Leftrightarrow \\
Y^T X - \sum_{j=1}^{n} \frac{x_j^T e^{x_j^T\hat{\beta}}}{1+e^{x_j^T\hat{\beta}}} = 0 \Leftrightarrow \\
Y^T X - \sum_{j=1}^{n} x_j^T\hat{\pi}_j = 0 \Leftrightarrow \\
Y^T X - X\hat{\pi} = 0 \Leftrightarrow \\
Y^T X = X\hat{\pi} \Leftrightarrow \\
(Y^T X)^T = (X\hat{\pi})^T \Leftrightarrow \\
X^T Y = X^T\hat{\pi} \qquad (52)
\end{aligned}
$$

□

Appendix

The table with the estimates for the probability of low birth weight for each infant, for computational problem 3.(ii):

Infant	Estimated Probability	Prediction	Real
1	0.12	0	0
2	0.10	0	0
3	0.43	0	0
4	0.41	0	0
5	0.45	0	0
6	0.24	0	0
7	0.25	0	0
8	0.34	0	0
9	0.28	0	0
10	0.34	0	0
11	0.35	0	0
12	0.19	0	0
13	0.73	1	0
14	0.35	0	0
15	0.47	0	0
16	0.47	0	0
17	0.38	0	0
18	0.33	0	0
19	0.26	0	0
20	0.30	0	0
21	0.16	0	0
22	0.22	0	0
23	0.05	0	0
24	0.19	0	0
25	0.22	0	0
26	0.11	0	0
27	0.40	0	0
28	0.13	0	0
29	0.18	0	0
30	0.31	0	0
31	0.31	0	0
32	0.61	1	0
33	0.37	0	0
34	0.14	0	0
35	0.20	0	0
36	0.23	0	0
37	0.32	0	0
38	0.29	0	0
39	0.08	0	0
40	0.28	0	0
41	0.18	0	0
42	0.11	0	0

Infant	Estimated Probability	Prediction	Real
43	0.21	0	0
44	0.15	0	0
45	0.51	1	0
46	0.51	1	0
47	0.14	0	0
48	0.23	0	0
49	0.24	0	0
50	0.48	0	0
51	0.65	1	0
52	0.21	0	0
53	0.32	0	0
54	0.36	0	0
55	0.28	0	0
56	0.33	0	0
57	0.40	0	0
58	0.12	0	0
59	0.31	0	0
60	0.29	0	0
61	0.29	0	0
62	0.23	0	0
63	0.25	0	0
64	0.15	0	0
65	0.60	1	0
66	0.25	0	0
67	0.24	0	0
68	0.05	0	0
69	0.50	0	0
70	0.26	0	0
71	0.71	1	0
72	0.19	0	0
73	0.36	0	0
74	0.32	0	0
75	0.37	0	0
76	0.07	0	0
77	0.17	0	0
78	0.36	0	0
79	0.37	0	0
80	0.09	0	0
81	0.21	0	0
82	0.09	0	0
83	0.20	0	0
84	0.24	0	0
85	0.22	0	0
86	0.41	0	0
87	0.31	0	0
88	0.21	0	0
89	0.07	0	0

Infant	Estimated Probability	Prediction	Real
90	0.23	0	0
91	0.19	0	0
92	0.21	0	0
93	0.39	0	0
94	0.85	1	0
95	0.37	0	0
96	0.15	0	0
97	0.12	0	0
98	0.30	0	0
99	0.30	0	0
100	0.14	0	0
101	0.25	0	0
102	0.58	1	0
103	0.40	0	0
104	0.26	0	0
105	0.26	0	0
106	0.20	0	0
107	0.20	0	0
108	0.13	0	0
109	0.40	0	0
110	0.16	0	0
111	0.07	0	0
112	0.27	0	0
113	0.26	0	0
114	0.17	0	0
115	0.22	0	0
116	0.16	0	0
117	0.27	0	0
118	0.17	0	0
119	0.22	0	0
120	0.38	0	0
121	0.16	0	0
122	0.13	0	0
123	0.26	0	0
124	0.22	0	0
125	0.19	0	0
126	0.17	0	0
127	0.14	0	0
128	0.39	0	0
129	0.23	0	0
130	0.10	0	0
131	0.46	0	1
132	0.16	0	1
133	0.40	0	1
134	0.80	1	1
135	0.32	0	1
136	0.14	0	1

Infant	Estimated Probability	Prediction	Real
137	0.30	0	1
138	0.34	0	1
139	0.59	1	1
140	0.63	1	1
141	0.31	0	1
142	0.80	1	1
143	0.23	0	1
144	0.26	0	1
145	0.42	0	1
146	0.28	0	1
147	0.09	0	1
148	0.37	0	1
149	0.30	0	1
150	0.24	0	1
151	0.64	1	1
152	0.32	0	1
153	0.42	0	1
154	0.49	0	1
155	0.18	0	1
156	0.54	1	1
157	0.38	0	1
158	0.49	0	1
159	0.18	0	1
160	0.53	1	1
161	0.45	0	1
162	0.41	0	1
163	0.29	0	1
164	0.20	0	1
165	0.61	1	1
166	0.54	1	1
167	0.47	0	1
168	0.27	0	1
169	0.49	0	1
170	0.34	0	1
171	0.16	0	1
172	0.37	0	1
173	0.39	0	1
174	0.32	0	1
175	0.23	0	1
176	0.35	0	1
177	0.32	0	1
178	0.41	0	1
179	0.55	1	1
180	0.28	0	1
181	0.65	1	1
182	0.30	0	1
183	0.14	0	1

Infant	Estimated Probability	Prediction	Real
184	0.68	1	1
185	0.38	0	1
186	0.38	0	1
187	0.44	0	1
188	0.62	1	1
189	0.74	1	1

Table 14: Estimated probabilities of low birth weight, predictions using threshold 0.5 and real values.

R Code

```
### Exercise 1 ###


library(xtable)


chemo = read.table("chemo.txt",header=TRUE)


chemo$seizures = chemo$SeizuresP1 + chemo$SeizuresP2
 + chemo$SeizuresP3 + chemo$SeizuresP4

#informative plots
plot(chemo$seizures, chemo$Treatment)
plot(chemo$Baseline, chemo$seizures, xlab ='Baseline Rate',
ylab= 'Number of epileptic seizures', main ='Number of seizures vs Baseline Rate',pch =20)
plot(chemo$Age, chemo$seizures, xlab ='Age of patient (years)',
 ylab= 'Number of epileptic seizures',
  main ='Number of seizures vs Age of patient',pch =20)
boxplot(chemo$seizures ~chemo$Treatment, xlab =
'Treatment (0=placebo, 1=active)', ylab ='Number of epileptic seizures',
main ='Number of seizures by Treatment',pch=20)




model1 = glm(seizures~Treatment + Baseline + Age, data = chemo, family = poisson)
summary(model1)
xtable(summary(model1))
beta = coefficients(model1)

#expected number of epilectic seizures for each patient
nseizures = matrix(fitted(model1))
```

```
nseizures
xtable(nseizures)
#95% confidence interval
FisherInv=vcov(model1)
df = model1$df.residual

#intercept
ub1 = beta[1] + qt(0.975,df)*sqrt(FisherInv[1,1])
lb1 = beta[1] - qt(0.975,df)*sqrt(FisherInv[1,1])
ci_intercept = c(lb1,ub1)

#treatment
ub2 = beta[2] + qt(0.975,df)*sqrt(FisherInv[2,2])
lb2 = beta[2] - qt(0.975,df)*sqrt(FisherInv[2,2])
ci_treatment = c(lb2,ub2)


#Baseline
ub3 = beta[3] + qt(0.975,df)*sqrt(FisherInv[3,3])
lb3 = beta[3] - qt(0.975,df)*sqrt(FisherInv[3,3])
ci_baseline = c(lb3,ub3)

#Age
ub4 = beta[4] + qt(0.975,df)*sqrt(FisherInv[4,4])
lb4 = beta[4] - qt(0.975,df)*sqrt(FisherInv[4,4])
ci_age = c(lb4,ub4)

confidence = rbind(ci_intercept, ci_treatment, ci_baseline, ci_age)
confidence
#Pearson chi-squared statistic
P = sum(residuals(model1,"pearson")^2)

#design matrix
X = model.matrix(model1)
#weight matrix
W = data.matrix(diag(model1$weights))

df = model1$df.residual
phi = P/df
If = phi*(t(X) %*% W %*% X)
If_inv = solve(If)
#square root of I diagonals
sq_I = sqrt(diag(If_inv))


#Plot of observed vs fitted values

plot(model1$fitted,chemo$seizures,xlab="Fitted Values",ylab="Observed Values",
```

```
 xlim = c(0,145), ylim = c(0,145),
     main= "Observed against Fitted values of the response", pch=20, cex=1, col="blue")
lines(c(0,145),c(0,145))

#Analysis of Deviance table with chi-square test
anova(model1,test = 'Chisq')
xtable(anova(model1,test = 'Chisq'))

#obtain deviance D
D = deviance(model1)

#residuals
resP = residuals(model1,'pearson')
resD =residuals(model1,'deviance')

#Deviance residuals against fitted values
plot(resD,chemo$seizures,xlab="Deviance Residuals",ylab="Fitted Values",
     main= "Deviance residuals against Fitted values", pch=20, cex=1, col="blue")

#Pearson residuals against fitted values
plot(resP,chemo$seizures,xlab="Pearson Residuals",ylab="Fitted Values",
     main= "Pearson residuals against Fitted values", pch=20, cex=1, col="blue")

#chi-square
qchisq(0.95,df)

#use of sqrt link function
model2 = glm(seizures~Treatment + Baseline + Age, data = chemo,
 family = poisson(link = 'sqrt'))
summary(model2)

P2 = sum(residuals(model2,"pearson")^2)
D2 = sum(residuals(model2,"deviance")^2)
P2
D2
anova(model2,test = 'Chisq')
```

```
### Exercise 2 ###


library(ggplot2)
library(xtable)


clotting = data.frame(conc=c(5,10,15,20,30,40,60,80,100,5,10,15,20,30,40,60,80,100),
lot=factor(c(rep(1,9),rep(2,9))),
time=c(118,58,42,35,27,25,21,19,18,69,35,26,21,18,16,13,12,12))


#informative plots
plot(clotting$time, log(clotting$conc))
plot(clotting$time, clotting$lot)

boxplot(clotting$time ~ clotting$lot, xlab = 'Lot of thromboplastin',
 ylab ='Clotting time (seconds)', main = 'Clotting time by lot of thromboplastin')

model1 = glm(time~log(conc) + lot, data = clotting, family = Gamma(link= 'inverse'))
summary(model1)
beta = coefficients(model1)
#plot of clotting against log(conc)
ggplot(clotting, aes(log(conc),time,shape=lot))+
  geom_point()+
  ggtitle("Clotting time against log(conc)")+
  labs(x="logarithm of concentration",y="Clotting time (seconds)")

#expected clotting time for each individual
times = matrix(fitted(model1))


#90% confidence interval
FisherInv=vcov(model1)
df = model1$df.residual

#intercept
ub1 = beta[1] + qt(0.95,df)*sqrt(FisherInv[1,1])
lb1 = beta[1] - qt(0.95,df)*sqrt(FisherInv[1,1])
ci_intercept= c(lb1,ub1)


#log(conc)
ub2 = beta[2] + qt(0.95,df)*sqrt(FisherInv[2,2])
lb2 = beta[2] - qt(0.95,df)*sqrt(FisherInv[2,2])
ci_log_conc = c(lb2,ub2)
```

```
#lot
ub3 = beta[3] + qt(0.95,df)*sqrt(FisherInv[3,3])
lb3 = beta[3] - qt(0.95,df)*sqrt(FisherInv[3,3])
ci_lot = c(lb3,ub3)

confidence = rbind(ci_intercept,ci_log_conc,ci_lot)
confidence

#Pearson chi-squared statistic
P = sum(residuals(model1,"pearson")^2)
df = model1$df.residual
phi = P/df

phi
#Plot of observed vs fitted values

plot(model1$fitted[1:9], clotting$time[1:9],ylim = c(12,145), xlim = c(12,145),
xlab = "Fitted Values", ylab = "Observed Values",
  main= "Observed against Fitted values of clotting time", pch=20, cex=1, col="blue" )
points(model1$fitted[10:18], clotting$time[10:18],ylim = c(12,145), xlim = c(12,145),
pch =20, col ='red')
abline(0,1,lty=3)
legend("bottomright", legend = c("lot 1","lot2"), pch=c(20,20),
lwd=1,lty=c(NA,NA), col=c("blue","red"),bty = 'n')


#Analysis of deviance table with F test

anova(model1, test ="F")

#manual calculation of p-value using t-quantiles
t_log=coef(model1)[2]/sqrt(FisherInv[2,2])

2*(1-pt(abs(t_log),df))

t_lot=coef(model1)[3]/sqrt(FisherInv[3,3])
2*(1-pt(abs(t_lot),df))




#obtain deviance D
D = deviance(model1)

P/phi
D/phi
#residuals
resP = residuals(model1,'pearson')
```

```
resD =residuals(model1,'deviance')

#Deviance residuals against Pearson residuals
plot(resD,resP,xlab="Deviance Residuals",ylab="Pearson Residuals",
     main= "Deviance residuals against Pearson residuals", pch=20, cex=1, col="blue")
abline(0,1,lty=3)

qchisq(0.95,df)




### Exercise 3 ###

library(xtable)

data = read.table("birthweight.txt",header=TRUE)
#convert int to factors
data$Race = factor(data$Race)
data$Smoker = factor(data$Smoker)
data$Hypertension = factor(data$Hypertension)
data$UterineIrrit = factor(data$UterineIrrit)

#variable low
data$low = data$BirthWeight < 2500
data$low = factor(data$low)

#informative plots


plot(data$MotherAge, data$BirthWeight,
     main="Plot of Birth Weight vs Mothers Age", xlab="Mothers Age", ylab="Birth Weight")

plot(data$MotherWeight, data$BirthWeight,
     main="Plot of Birth Weight vs Mothers Weight",
     xlab="Mothers Weight", ylab="Birth Weight")

boxplot(data$BirthWeight~data$Smoker,
        main="Boxplot per mother's Smoking status",
         xlab="0=No, 1=Yes", ylab="Birth Weight")

boxplot(data$BirthWeight~data$Race,
        main="Boxplot per mother's Race",
        xlab="1 = white, 2 = black, 3 = other", ylab="Birth Weight")

boxplot(data$BirthWeight~data$PremLabor,
        main="Boxplot per mother's Premature Labor history", xlab="Number of Occurences",
```

```
         ylab="Birth Weight")

boxplot(data$BirthWeight~data$Hypertension,
        main="Boxplot per mother's Hypertension history", xlab="0=No, 1=Yes",
         ylab="Birth Weight")

boxplot(data$BirthWeight~data$UterineIrrit,
        main="Boxplot per mother's Uterine Irritability status ", xlab="0=No, 1=Yes",
         ylab="Birth Weight")

boxplot(data$BirthWeight~data$PhysicianVisits,
        main="Boxplot per mother's Physician Visits ", xlab="Number of Visits",
         ylab="Birth Weight")


#fit full model

full = glm(low ~ MotherAge + MotherWeight + Race + Smoker + PremLabor + Hypertension +
UterineIrrit + PhysicianVisits,
data = data, family = binomial(link='logit'))
summary(full)
AIC(full)
BIC(full)
anova(full, test = 'Chisq')


#fit reduced model

reduced = glm(low ~  MotherAge + MotherWeight +
 Smoker + PremLabor + Hypertension , data = data, family = binomial(link='logit'))
summary(reduced)
anova(reduced, test = "Chisq")
AIC(reduced)
BIC(reduced)


#estimates for the probability of low birth weight
prob = matrix(fitted(reduced))
prob

#setting threshold 0.5
pred = matrix(prob > 0.5)
pred = pred * 1


real = matrix(data$BirthWeight < 2500)
real = real * 1
table = cbind(prob,pred,real)
```

```
#accuracy
acc = mean(real == pred)
acc
xtable(table)

#accuracy of the full model
prob2 = matrix(fitted(full))
pred2 = matrix(prob2 > 0.5)
pred2 = pred2 * 1

table2 = cbind(prob2,pred2,real)
acc2 = mean (real == pred2)
acc2

P = sum(residuals(reduced,"pearson")^2)
P
D = deviance(reduced)
D
df = reduced$df.residual


qchisq(0.95,df)

#residuals
resP = residuals(reduced,'pearson')
resD =residuals(reduced,'deviance')

#Deviance residuals against Pearson residuals
plot(resD,resP,xlab="Deviance Residuals",ylab="Pearson Residuals",
     main= "Deviance residuals against Pearson residuals", pch=20, cex=1, col="blue")
abline(0,1,lty=3)
```