MATH 523 - ASSIGNMENT 3: SOLUTIONS

1 The first step is to reconfigure the data set to a more conventional format, organized into rows with one row per subject. This is easily achieved using the commands:

```
Food<-read.csv("Food.csv",header=T)
Y<-c(rep(Food[,1],Food[,2]),rep(Food[,1],Food[,3]))
Sex<-as.factor(rep(c("Male","Female"),c(sum(Food[,2]),sum(Food[,3]))))
Food.data<-data.frame(Y,Sex)
```

(a) We have count data, so a Poisson GLM seems initially plausible. However, under the default log link, fitting the model with `Sex` as a factor, we observed evidence of overdispersion (residual deviance 2194.7 on 387 degrees of freedom).

```
> Food.p<-glm(Y~Sex,family=poisson,data=Food.data)
> summary(Food.p)

Call: glm(formula = Y ~ Sex, family = poisson, data = Food.data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.21148    0.01456  220.52   <2e-16 ***
SexMale     -0.44804    0.02300  -19.48   <2e-16 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2582.4  on 388  degrees of freedom
Residual deviance: 2194.7  on 387  degrees of freedom
```

Hence we now attempt the fit of a Negative Binomial model, again using a log link:

```
> library(MASS)
> Food.nb<-glm.nb(Y~Sex,data=Food.data)
> summary(Food.nb)

Call:glm.nb(formula = Y~Sex, data=Food.data, init.theta=4.433616268,
            link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.21148    0.03741  85.855   <2e-16 ***
SexMale     -0.44804    0.05338  -8.393   <2e-16 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

(Dispersion parameter for Negative Binomial(4.4336) family taken to be 1)

    Null deviance: 468.81  on 388  degrees of freedom
Residual deviance: 398.45  on 387  degrees of freedom

              Theta:  4.434
          Std. Err.:  0.383
2 x log-likelihood:  -2857.637
```

This fit seems adequate, as the residual deviance is plausibly small (398.45 on 387 degrees of freedom. The effect of `Sex` is evident: the coefficient `SexMale` is -0.448 (0.053), yielding a Wald statistic of -8.383, that is, strongly significant. For an analysis of deviance, we can used the deviance, that is

$$\frac{D(\boldsymbol{y};\widehat{\mu}_0) - D(\boldsymbol{y};\widehat{\mu}_{\text{SEX}})}{\phi} = \frac{468.81 - 398.45}{1} = 70.36$$

which, when compared against a χ^2_1 distribution is a highly significant result. Thus the expected counts for Males and Females are significantly different. For a global assessment of model adequacy $\widehat{\phi} = 398.45/387 \doteq 1$, indicating no evidence for model inadequacy.

5 MARKS

For an approximate 95 % interval for $\mu_M - \mu_F$, we cannot use the above analysis, as it formulates a model for $\log \mu_F$ and $\log \mu_M$. However, if the `identity` link is used, so that

$$\mu = \begin{cases} \alpha_0 & \texttt{Sex=Female} \\ \alpha_0 + \alpha_1 & \texttt{Sex=Male} \end{cases}$$

the interval can be computed directly.

```
>Food.nbi<-glm.nb(Y~Sex,data=Food.data,link="identity")
>summary(Food.nbi)


Call: glm.nb(formula = Y ~ Sex, data = Food.data, link = "identity",
    init.theta = 4.433616268)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  24.8158     0.9283  26.734  < 2e-16 ***
SexMale      -8.9615     1.1073  -8.093 5.83e-16 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

(Dispersion parameter for Negative Binomial(4.4336) family taken to be 1)

    Null deviance: 468.81  on 388  degrees of freedom
Residual deviance: 398.45  on 387  degrees of freedom

              Theta:  4.434
          Std. Err.:  0.383
 2 x log-likelihood:  -2857.637
```

Using the rule of thumb of "estimate plus or minus two standard errors", we have a confidence interval of $-8.962 \pm 2 \times 1.107 = (-11.176, -6.747)$. Using the `confint` function, which uses a more mathematically precise method, we obtain a similar result.

```
> confint(Food.nbi)
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept)  23.07573 26.720455
SexMale     -11.17966 -6.827767
```

3 MARKS

Notes:

(i) For a completely non-parametric (model-free) interval, we might use the moment-based interval

$$(\overline{y}_M - \overline{y}_F) \pm 2\sqrt{\frac{s_F^2}{n_F} + \frac{s_M^2}{n_M}} = (-11.328, -6.596)$$

which allows for a different variance within each sex subgroup.

(ii) The maximized log-likelihood, and the parameter estimates of for θ_Z (the negative binomial parameter) are left unchanged by the move from log to identity link. This is because one model is a reparameterization of the other: we have that

$$\log \mu_F = \beta_0 \Longleftrightarrow \mu_F = e^{\beta_0} = \alpha_0$$

and

$$\log \mu_M = \beta_0 + \beta_1 \Longleftrightarrow \mu_M = e^{\beta_0+\beta_1} = e^{\beta_0}e^{\beta_1} = \alpha_0(1 + \alpha_1/\alpha_0)$$

say, where $\beta_1 = \log(1 + \alpha_1/\alpha_0)$. Hence

$$\log \mu_M - \log \mu_F = \beta_1 = \log(1 + \alpha_1/\alpha_0) \qquad \mu_M - \mu_F = \alpha_1$$

(b) As the example above indicates, the parameter estimates attached to the coefficients of the components of the linear predictor will **not** in general be equal, even though the model fit will may be the same. This is for several reasons:

- The predictor effects are clearly on a *different scale*; for example, for the null model, we have either

$$\log \mu = \beta_0 \qquad \text{or} \qquad \mu = \alpha_0$$

say.

- In some cases (such as in Q1) there may be a 1-1 relationship between the estimates themselves. This typically happens in the context of **factor** predictors. However, the standard errors will necessarily be different, due to the Delta method result. In general for two link functions g_1 and g_2, there will not be a 1-1 relationship for **continuous** (covariate) predictors.

- In general, from first principles, we have the log-likelihood in the Exponential-Dispersion Family as

$$l(\theta, \phi; \boldsymbol{y}) = \sum_{i=1}^{n} \frac{w_i \left(y_i h(x_i\beta) - b(h(x_i\beta))\right)}{\phi} + \sum_{i=1}^{n} c\left(y_i, \frac{\phi}{w_i}\right) \tag{1}$$

where $h(.)$ incorporates the linear predictor/mean relationship and the link function. It is evident that the first and second derivatives depend on the form of h, and the resulting score equations and Hessian (second derivative) matrix change depending on h. As these derivatives determine the estimates and standard errors, both quantities will vary as h varies.

4 MARKS

2 Here is the code to carry out the simulation and display. The settings I have used are to set $n = 1000$, and then to use β_1 values equally spaced in the range $[-5, 5]$ for 1000 replicate data sets. I then plot boxplots and histograms to demonstrate the difference in results as β_1 varies.

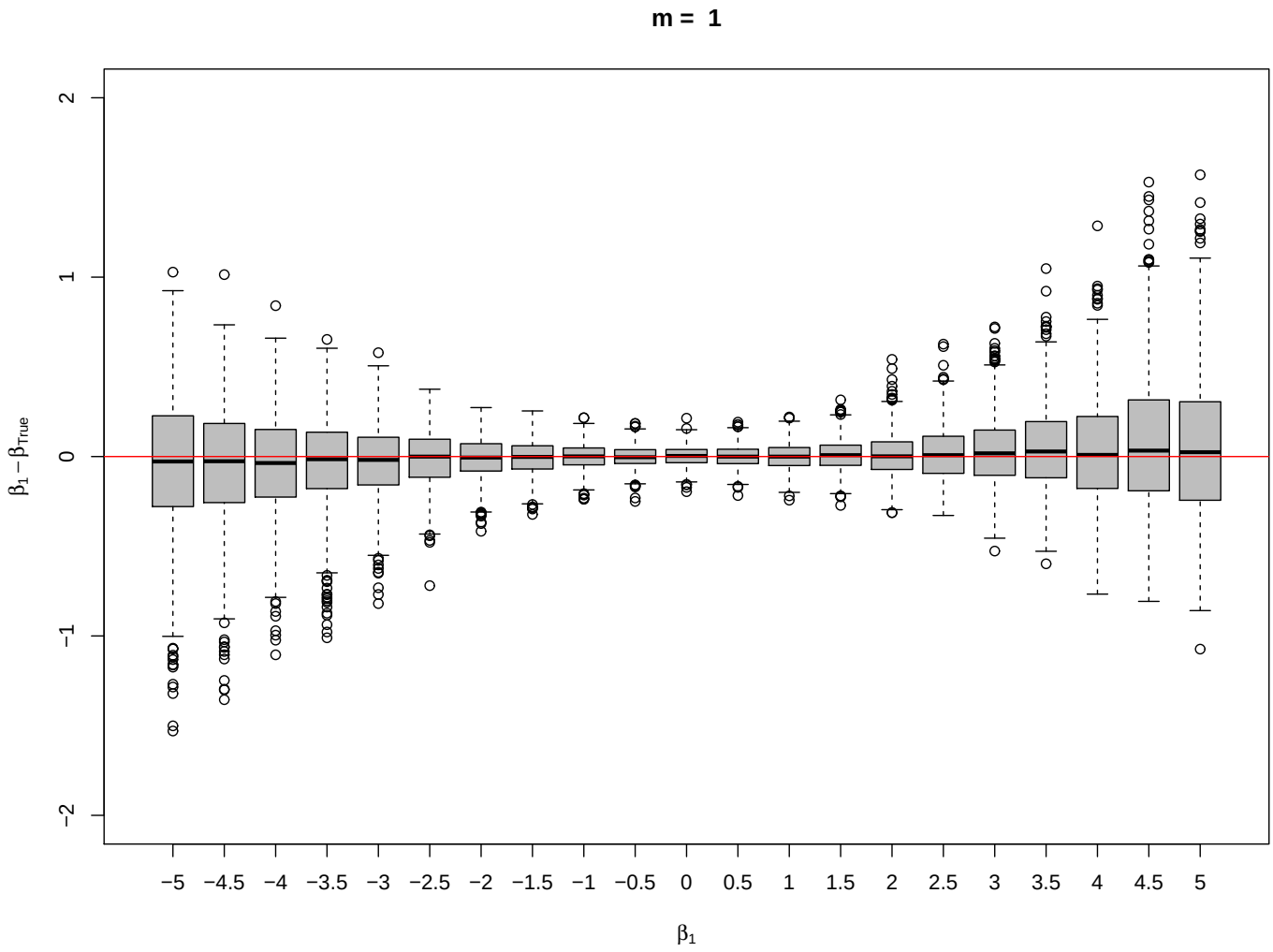


Figure 1: Boxplots of $\widehat{\beta}_1 - \beta_1$ for values of β_1 in the range -5 to 5.

(a) For the ML estimator of β_1: the distributions of the estimators for different values of β_1 presented as boxplots in Figure 2a and as histograms and QQ plots in Figure 2a. Despite some mild skewness for larger values of β_1, the approximate normality is observed. With larger sample sizes, the normal approximation becomes even better.

4 MARKS

(b) For the distribution of the deviance statistic: the distributions of the deviance statistics for different values of β_1 presented as boxplots in Figure 2b. Clearly there is considerable variation in the distribution with β_1.

4 MARKS

Note that I also include similar plots for differing values of m, that is, $m = 1, 3, 5, 10, 100, 1000$. The theory suggests that as m increases, problems associated with the deviance in the $m = 1$ case are alleviated, and this is backed up by the boxplots.

Note that this result on the deviance does not necessarily mean that the analysis of deviance test is invalid; the difference in deviances between two nested models may be still well approximated by a chi-squared distribution.

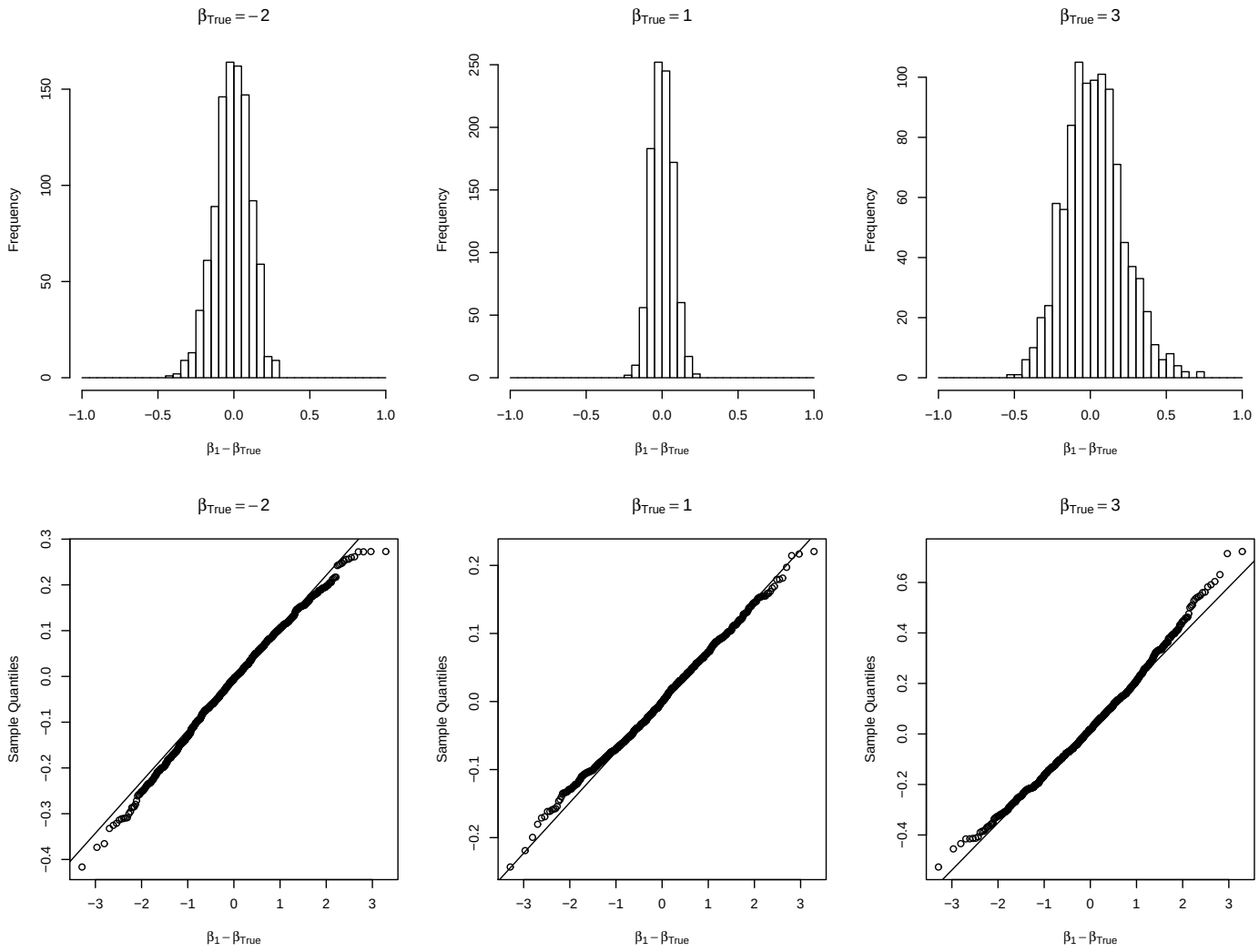


Figure 2: Histograms and QQPlots of $\widehat{\beta}_1 - \beta_1$ for three values of $\beta_1 : -2, 1, 3$.

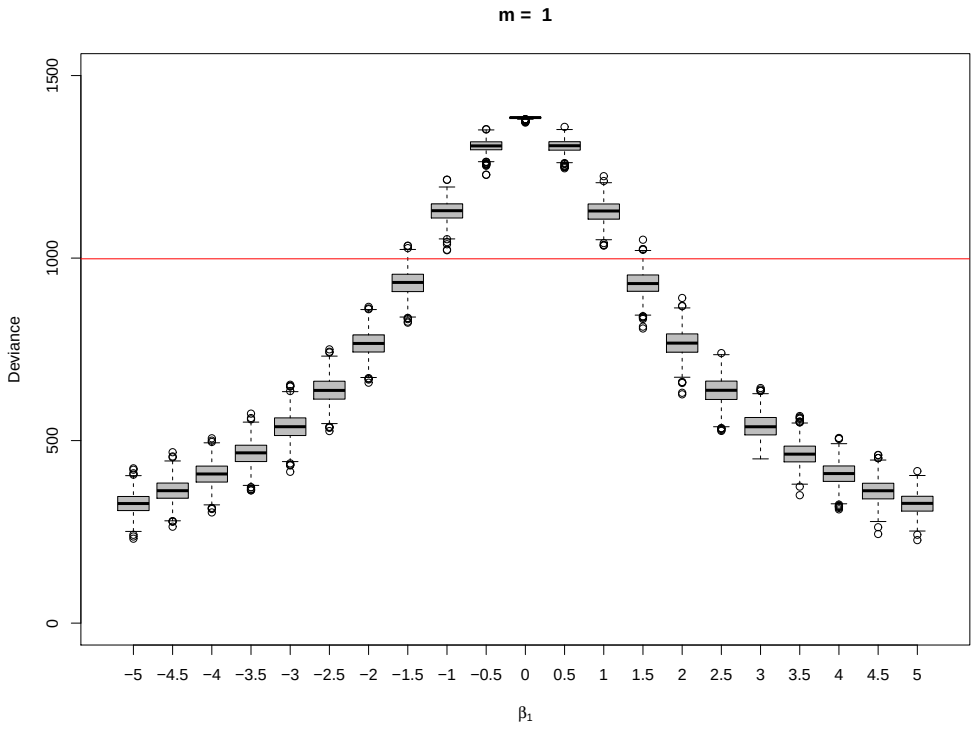


Figure 3: Boxplots of deviance statistic for values of β_1 in the range -5 to 5. Horizontal line at $y = n - 2 = 998$ indicates the expected deviance value under regular asymptotic theory.

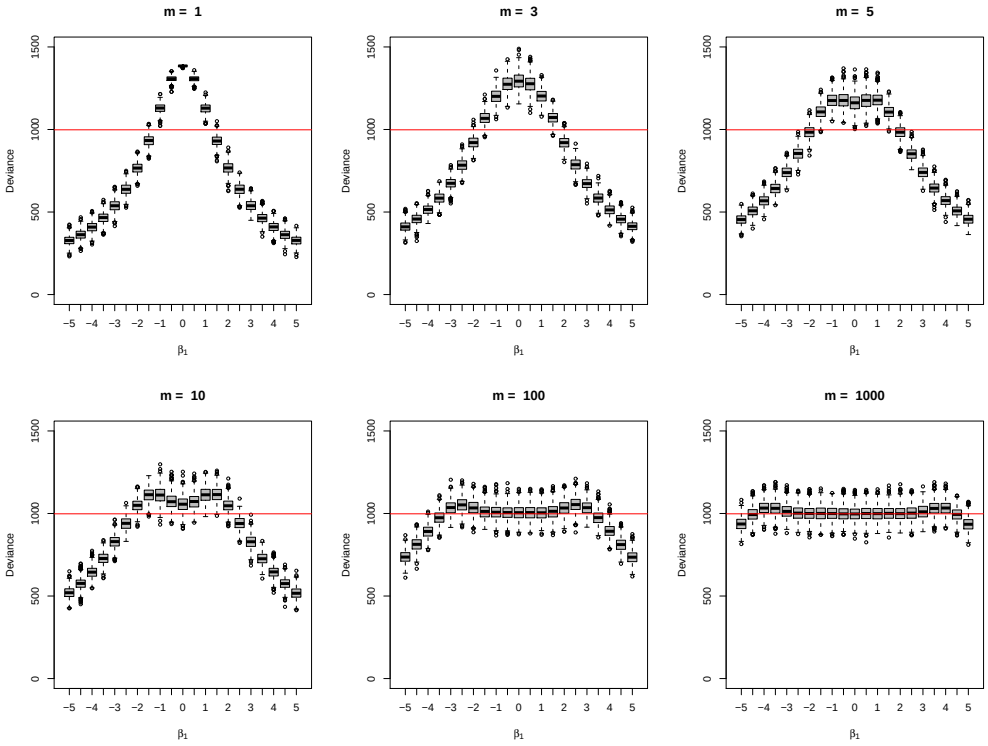


Figure 4: Boxplots of deviance statistics for values of β_1 in the range -5 to 5 for $m = 1, 3, 5, 10, 100, 1000$. Horizontal line at $y = n - 2 = 998$ indicates the expected deviance value under regular asymptotic theory.