Required material, Sections 4.1-4.6 in the slides, Lectures 9-13.

This sheet is split into three sections:

- Section A: introductory question(s)
- Section B: core questions
- Section C: further question(s) - maybe some of these are slightly harder but all of them are on material that is part of the course

Undergraduates: only answers to Section B should be handed in for marking.
Solutions will be provided for all questions.

Section A

1. We have 300 observations on a binary variable `y`. For each observation we also observe the values of continuous variables `x1` and `x2`. The linear predictor η is specified as,

$$\eta = \beta_0 + \beta_1 \texttt{x1} + \beta_2 \texttt{x2} + \beta_3 (\texttt{x2})^2.$$

Some edited output for a fitted Bernoulli GLM with canonical link function is given below.

```
> bern.glm <- glm(y ~ x1 + x2 + I(x2^2), family=binomial)
> summary(bern.glm)
Coefficients:
            Estimate Std. Error
(Intercept)   -0.099      0.262
x1             0.323      0.154
x2             0.454      0.263
I(x2^2)       -0.172      0.096
    Null deviance: 272.12  on 199  degrees of freedom
Residual deviance: 263.76  on 196  degrees of freedom
```

(a) Test the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ at the 5% level.

(b) Calculate the estimated effect of a unit change in `x1` on the odds ratio.

(c) Calculate a 95% confidence interval for the effect of a unit change in `x1` on the odds ratio. What conclusion can you draw from this confidence interval?

(d) At the value `x2` $= 0.5$, calculate the estimated effect of a positive unit change in `x2` on the odds ratio.

Section B

2. The data below are the `babyfood` data described in Faraway (2006, Chapter 2). For each gender and feed type a certain total number ($m_i =$ `disease`$[i] +$ `healthy`$[i]$, $i = 1, \ldots, 6$) of children were followed for their first year of life, and the number developing some form of respiratory disease recorded.

```r
bf <- data.frame(disease=c(77,19,47,48,16,31),
    healthy=c(381,128,447,336,111,433),
    gender=c('M','M','M','F','F','F'),
    feed=c('Bottle','Suppl','Breast','Bottle','Suppl','Breast'))
bf
```

```
##   disease healthy gender   feed
## 1      77     381      M Bottle
## 2      19     128      M  Suppl
## 3      47     447      M Breast
## 4      48     336      F Bottle
## 5      16     111      F  Suppl
## 6      31     433      F Breast
```

```r
summary(glm(cbind(disease,healthy) ~ gender + feed,
            family=binomial, data=bf))
```

```
##
## Call:
## glm(formula = cbind(disease, healthy) ~ gender + feed, family = binomial,
##     data = bf)
##
## Deviance Residuals:
##       1        2        3        4        5        6
##  0.1096  -0.5052   0.1922  -0.1342   0.5896  -0.2284
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.9253     0.1286 -14.971  < 2e-16 ***
## genderM       0.3126     0.1410   2.216   0.0267 *
## feedBreast   -0.6693     0.1530  -4.374 1.22e-05 ***
## feedSuppl    -0.1725     0.2056  -0.839   0.4013
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26.37529  on 5  degrees of freedom
## Residual deviance:  0.72192  on 2  degrees of freedom
## AIC: 40.24
##
## Number of Fisher Scoring iterations: 4
```

(a) Calculate the odds for disease in male and female infants for each feed type. Give a 95% CI for the log-odds ratio for disease in breast against bottle fed infants and test for a difference in log-odds at 5%.

(b) Write down the linear predictor for the null model and carry out a test for any explanatory relation between the response (`disease, healthy`) and the variables `gender` and `feed`.

(c) The residual deviance for the model with linear predictor `1 + gender` is 20.899. Test for an effect due to `feed`.

(d) Show that the model with an interaction between feed and gender (so with a linear predictor of `feed*gender = 1 + feed + gender + feed:gender`) is just the saturated model, and give the residual deviance for this model(!). Test for an interaction between feed and gender at 5%.

(e) Using Problem Sheet 3, Question 3, adapt the IRLS R-code given in lectures to compute the `Estimate` and `Std. Error` columns of the R-output above.

Some hints -- to generate the design matrix:

```
X <- model.matrix(cbind(disease,healthy) ~ gender + feed)
```

and to set up the weight matrix (given mu):

```
m <- disease + healthy
W <- diag(c(mu*(m-mu)/m))
```

3. Brown, B.W. (1980) describes a study of 53 prostate cancer patients. Five binary predictor variables ('xray', 'acid', 'stage', 'aged' and 'grade') were measured before surgery. For these variables 0 indicates the risk factor is absent, and 1 indicates that it is present. The patients then had surgery to determine whether there was nodal involvement (`NI = 1`) or not (`NI = 0`) in the cancer. The aim was to find out which predictor variables were most important. The table below shows the first and last 3 rows of the data.

```
   NI xray acid stage aged grade
1  1    1    1    1     0     1
2  1    1    1    1     0     1
3  1    1    1    1     0     1
...
51 1    0    1    0     0     1
52 0    1    1    0     0     0
53 0    1    0    0     0     0
```

(a) Specify a generalised linear model for these data using 'NI' as the binary response, and 'xray', 'acid', 'stage', 'aged', and 'grade' as binary explanatory variables, using the canonical link function for your model.

(b) Define the saturated model for the GLM you gave in (a), and calculate the residual deviance of the model of (a) as a function of its parameter MLEs.

(c) Fitting the Bernoulli GLM with

```
NI ~ 1 + xray + acid + stage + aged + grade
```

gave the following residual deviances:

```
          Resid. Dev
Intercept  70.25
xray       60.93
acid       54.79
stage      49.18
aged       48.76
grade      47.61
```

`Resid. Dev` in row j is the residual deviance for the GLM with a linear predictor including all the variables in rows 1 to j. For example, the GLM with linear predictor

```
NI ~ 1 + xray + acid
```

has residual deviance of 54.79.

(i) Give the null deviance and test $H_0 : \beta_2 = \beta_3 = \ldots = \beta_6 = 0$.

(ii) Calculate the AIC for each of these nested models and carry out model selection using the AIC values.

(iii) Test $H_0 : \beta_5 = \beta_6 = 0$.

(iv) Explain why we should not test for goodness of fit by directly comparing the residual deviance (of the final model) to a χ^2 variate.

4. The data for this question are in the file `aids.csv`, which you can find in the Canvas 'Problem sheets' module, see also the link below. *(Poisson GLM, data from P.D. Baxter lecture notes 'Generalised Linear Models by Example')*.

The columns of this data set are:

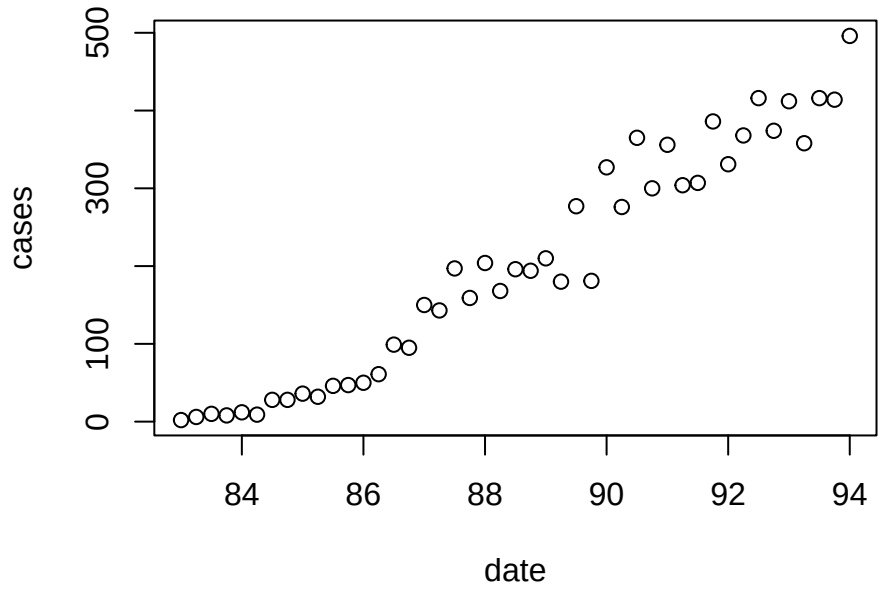
- `cases`, the number of AIDS cases in a given 3 month period (a 'quarter') from Jan 1983 to Mar 1994 as reported by the Public Health Laboratory Service, Communicable Disease Surveillance Unit, London
- `qrt`, the quarter of the year
- `date` the year in fractions of a quarter.

The aim of the analysis is to see if there are any seasonal effects, to see if an exponential increase in the number of aids cases over the period is a reasonable model.

```
aids <- read.csv("http://www.stats.ox.ac.uk/~laws/SB1/data/aids.csv")
head(aids)

##   cases qrt  date
## 1     2   1 83.00
## 2     6   2 83.25
## 3    10   3 83.50
## 4     8   4 83.75
## 5    12   1 84.00
## 6     9   2 84.25

aids$qrt <- as.factor(aids$qrt)
plot(cases ~ date, data=aids)
```



The levels of `qrt` are $1, 2, 3, 4$. Consider fitting a model with y_i = `cases[i]`, $x_{i,2}$, $x_{i,3}$ and $x_{i,4}$ dummy indicator variables for levels 2, 3 and 4 of `qrt[i]`, and $x_{i,5}$ = `date[i]`. The linear predictor is

$$\eta_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5},$$

and the stochastic part of the GLM is $Y_i \sim \text{Poisson}(\mu_i)$.

(a) Give a brief mathematical explanation of what it means to use a log link function, or a square-root link function.

(b) Using R, and the `family=poisson` option to `glm()`, fit the model. Try the log and square-root link functions (use `glm(cases ~ date + qrt, data=aids, family=poisson(link=sqrt))` to get the square-root link function). Which link function provides a better fit?

(c) Is `qrt` a significant explanatory variable? Briefly interpret the estimated parameters of the regression.

(d) (i) Inspect the standardised residuals. Is there evidence of misfit?
(ii) Perform the χ^2 test of goodness of fit, and argue whether this test is appropriate here.
(iii) Are there any points of large influence?

Section C

5. Consider logistic regression with n binary observations $y_1, \ldots, y_n$ and linear predictors $\eta_i = \mathbf{x}_i^T \beta$, $i = 1, \ldots, n$. Show that the likelihood can be written as

$$L(\beta; y) = \prod_i \left(\frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)} \right)^{y_i} \left(\frac{1}{1 + \exp(\mathbf{x}_i^T \beta)} \right)^{1-y_i}.$$

Consider the data $y = (0, 0, 0, 0, 1, 1, 1, 1)^T$ and design matrix $X = (\mathbf{1}, x)$ with $p = 2$ variables, where $\mathbf{1}$ is a column vector of 1s and $x = (-4, -3, -2, -1, 1, 2, 3, 4)^T$.

(a) Explain why the MLE is not defined.

(b) Can you generalise this result to $p > 2$?