RESEARCH SCHOOL OF FINANCE,
ACTUARIAL STUDIES AND STATISTICS
College of Business & Economics, The Australian National University

GENERALISED LINEAR MODELLING

(STAT3015/STAT7030)

Solutions to Assignment 2 for 2015

Question 1 (a)

The *R* model object for my preferred model:

```
> geriatric.glm

Call:  glm(formula = falls ~ strength + balance + gender + training,
    family = poisson)

Coefficients:
(Intercept)     strength      balance       gender     training
   0.489467     0.008566     0.009470    -0.046606    -1.069403

Degrees of Freedom: 99 Total (i.e. Null);  95 Residual
Null Deviance:    199.2
Residual Deviance: 108.8    AIC: 377.3
```

The underlying population model is a log-linear or Poisson regression model, which, for a GLM, consists of a model for the mean or expected value:

$$\mathrm{E}\left[\ln\left(Y_{ij}\right)\right] = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 Z_{ij} + \eta_j$$

where Y is the number of **falls**,

X_1 is the **strength** score (higher values indicate greater stability),
X_2 is the **balance** score (higher values indicate greater strength),
Z is a 0/1 indicator variable for gender (0=female, 1=male),
j indicates the levels of **training** {0 = "education only", 1 = "education + training"},
i = 1, 2, 3, … , 50 for both levels of j; for a total of 50 x 2 = 100 observations, and
$g(\ ) = \ln(\ )$ is the link function, which in this instance is the canonical link for the Poisson family, which is the natural log function (simply log( ) in *S-Plus*).

For a GLM, the underlying population model also includes a model for the variance of the response variable (and hence the errors for the model). In this instance, we are assuming that **falls** follows a Poisson distribution with variance, μ (for Poisson distributed data, the variance should equal the mean) and that following appropriate weighting by the inverse of the variance, the errors (ε_{ij}) will be independently and identically distributed (approximately normally for sufficiently large sample size) with constant dispersion ($\phi = 1$).

As each observation is a single individual (i.e. we are not dealing with a Poisson rate for grouped data with different group sizes), we do not need to specify additional weights (in addition to the default variance weights) to ensure a constant dispersion.

The choice of a Poisson regression with a default log link is fairly obvious considering the nature of the response variable (counts of the number of **falls** in a six month period). No interaction terms were significant, so none have not been included and none of the usual experimenting with transformations of explanatory variables or changing the link function appeared to noticeably improve the residual plot (see part (b) below).

Note that the main effects are not orthogonal and the analysis of deviance table (see part (c) below) can change considerably, depending on the order in which the explanatory variables are included in the model. However, the t statistics in the table of coefficients (also shown in part (c) below for the chosen model), give a fairly consistent story and **training**, **balance** and **strength** appear to be significantly related to the number of **falls**.

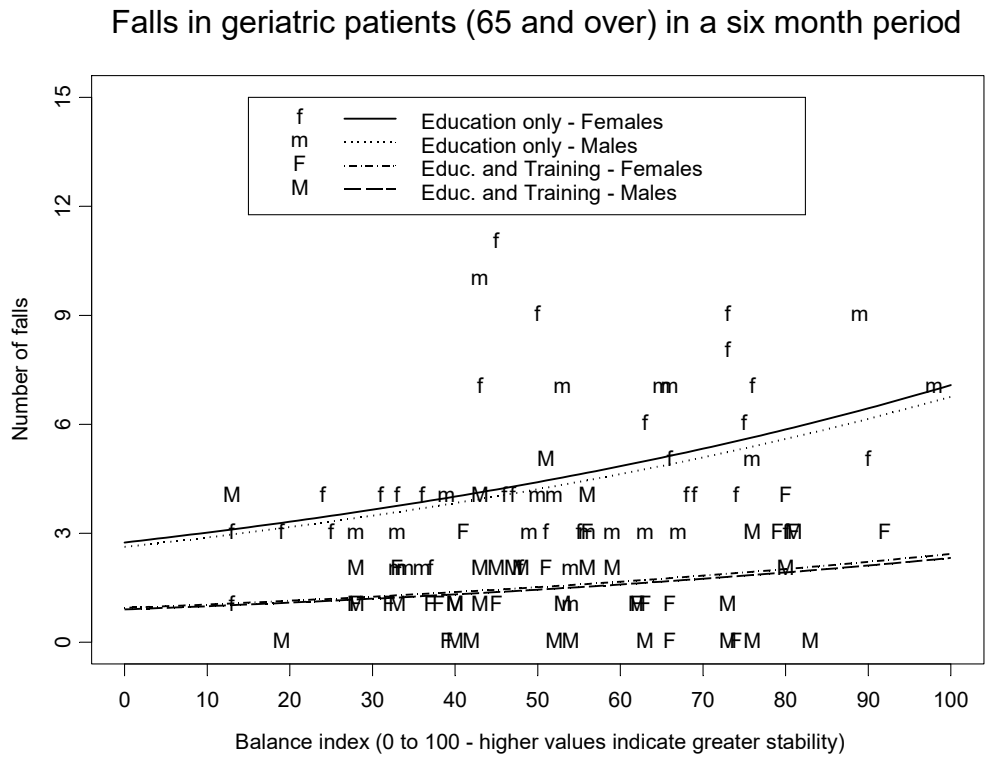
Question 1 (a) continued

Note that I have shown **training** in the above model as a factor variable, as there were 50 observations for each of the two levels of **training**, which suggests the underlying research design for this study has at least one of the features of a designed experiment. On the other hand, I have shown **gender** as a covariate, represented by a 0/1 indicator variable in the model, as there were definitely not equal numbers of observations for each of the four combinations of **gender** and **training**, suggesting that **gender** was not a design factor.

Note that either way, because both **training** and **gender** are 0/1 indicator variables, they are effectively included in the model as treatment coded factor variables with the 0 levels as the reference levels (constraints $Z_0 = 0$, $\eta_0 = 0$) and the corresponding model coefficients (β_3 & η_1) therefore represent the differences in the means between the 1 level and the 0 level of these two factors (i.e. treatment contrasts between levels).

We could exclude gender from the model on the grounds that it is a non-significant observed covariate, however, the wording of part (d) of the question seems to suggest that *a priori* we want a model which examines the effects of **training** on **falls**, controlling for the effects of **gender**, **balance** and **strength**. So, assuming that controlling for all three of **gender**, **balance** and **strength** was part of the research design (which is probably why the researchers decided to collect these covariates in the first place), it is a good idea to include these control variables in the model prior to **training**, so the effects of **training** can be examined after the effects of the control variables have been accounted for. As we are not able to check with the researchers about the details of the design they had in mind, I have decided to make this assumption and retain gender in the final model.

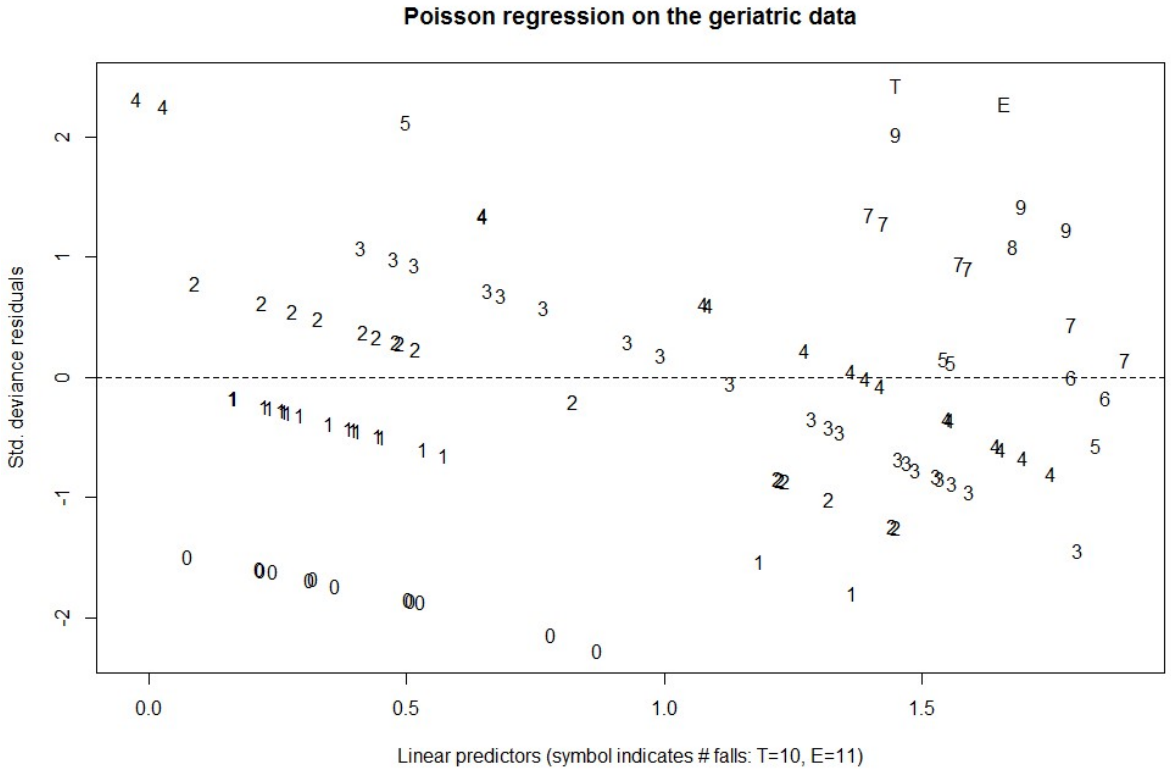
Note that this is not a required part of the assignment, but a plot such as the following illustrates the chosen model on the scale of the original variables and is therefore more readily interpreted by the researchers (you could also do a similar plot with **strength** on the x axis, which is not shown, but which looks similar):



(4 marks)

Question 1 (b)

The plot already presented on the previous page is a good summary plot of the chosen model; however, probably the most appropriate plot for assessing the fit of the GLM is a plot of the standardised deviance residuals against the linear predictors:



On the above plot, I have highlighted which residuals correspond to the different values of **falls**, which makes the reason for the patterns in this plot more obvious. A lot of the pattern is in the horizontal direction (i.e. a function of the explanatory variables), but there is enough systematic variation in the vertical direction to suggest some remaining unexplained correlation between the explanatory variables and the response variable.

Note if you compare the fitted values, rather the above linear predictors (the fitted values on the link transformed scale) it quickly becomes obvious that the model is consistently over-predicting for persons with zero falls and is also under-predicting the higher values of **falls** (8, 9, 10 & 11 (the fitted values range from just under 1 to around 6.6).

The scale of the vertical axis is good, with only 8% of the standardised residuals being more than 2 standard deviations from the mean. The normal quantile plot (not shown), also suggests some skewness in the distribution, but not enough to cause real concern. Taking this, together with fact that there is no evidence of significant under- or over-dispersion (see part (c) below), suggests that the assumption of Poisson distributed data is reasonable.

Given that I have already experimented unsuccessfully with using the existing variables to find another model which has a better version of this plot, I would have to go back to the researchers to see if there might be some other, as yet unrecorded, explanatory variable we could measure and include in the model (for example, the ages of the patients). The chosen model is as probably as good a GLM as we can get with the current data.

(4 marks)

Question 1 (c)

Given the wording of the question and the fact that there is nothing to suggest that we should *a priori* assume either under- or over-dispersion, this issue is best addressed using a two-tailed hypothesis test:

$$\mathrm{H}_0: \quad \phi = 1 \quad \text{vs} \quad \mathrm{H}_\mathrm{A}: \quad \phi \neq 1$$

As a test statistic, we can use the fact the residual deviance (after it has been suitably scaled by dividing by the assumed dispersion of 1, which corresponds to the assumption under the null hypothesis), should have a chi-square distribution with the residual degrees of freedom:

$$\frac{\sum_i d_i^2}{\phi} = \frac{\sum_i d_i^2}{1} = \sum_i d_i^2 \;\sim\; \chi^2_{n-p}$$

Using *R* to calculate the test statistic and a suitable $\alpha = 0.05$ rejection region:

```
> geriatric.glm$deviance
[1] 108.7899
> geriatric.glm$df.residual
[1] 95
> c(qchisq(0.025, geriatric.glm$df.residual),
qchisq(0.975, geriatric.glm$df.residual))
[1]  69.92487 123.85797
```

So as 108.8 lies in the interval (69.9, 123.9), we would not reject the null hypothesis and conclude that the dispersion is not significantly different to 1; i.e. there is no evidence of either under- or over-dispersion. As discussed in part (b), this is consistent with the assumption that the number of **falls** has a Poisson distribution.

(3 marks)

Question 1 (d)

The analysis of deviance table for the model in part (a) is:

```
> anova(geriatric.glm, test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: falls

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        99     199.19
strength  1    4.005        98     195.19  0.045377 *
balance   1    9.091        97     186.10  0.002569 **
gender    1    3.789        96     182.31  0.051601 .
training  1   73.520        95     108.79 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The very small *p*-value associated with the **training** term in the model is considerably smaller than $\alpha = 0.05$ ($\chi^2_1 = 73.5, p = 0.0000$), which indicates that there are significant differences in the expected number of ln(**falls**) between the two levels of **training** , i.e. that we should reject the following null hypothesis in favour of the alternative:

$$\mathrm{H}_0: \eta_{education+training} = \eta_{education\,only} = 0 \;(\text{or } \eta_1 = 0, \text{like} \eta_0) \quad \text{vs} \quad \mathrm{H}_\mathrm{A}: \text{at least one } \eta \neq 0 \quad (\eta_1 \neq 0)$$

Question 1 (d) continued

Including **training** last in the model means we are assessing the effects of **training** as an addition to a model that already includes the other variables, **strength**, **balance** and **gender**; so **training** does have a significant effect controlling for the effects of the other variables.

Note that not all the other variables are significant: **balance** is significant; but **strength** is marginal; and **gender** is not significant at the $\alpha = 0.05$ level. However, as argued in part (a), my interpretation of the research question suggests that we should include these terms in the model anyway, to control for their possible effects on the response.

The results of the hypotheses tests on these control variables are:

strength: reject $H_0 : \beta_1 = 0$ in favour of $H_A : \beta_1 \neq 0, (\chi^2_1 = 4.0, p = 0.0454)$

balance: reject $H_0 : \beta_2 = 0$ in favour of $H_A : \beta_2 \neq 0, (\chi^2_1 = 9.1, p = 0.0026)$

gender: do not reject $H_0 : \beta_3 = 0$ in favour of $H_A : \beta_3 \neq 0, (\chi^2_1 = 3.8, p = 0.0516)$

The table of coefficients gives consistent results to the above tests on the analysis of deviance table (another reason for choosing this particular model):

```
> summary(geriatric.glm)$coef
               Estimate  Std. Error    z value     Pr(>|z|)
(Intercept)  0.489467165 0.336869309  1.4529883 1.462270e-01
strength     0.008565829 0.004312119  1.9864546 4.698287e-02
balance      0.009469987 0.002952922  3.2069881 1.341325e-03
gender      -0.046606063 0.119970256 -0.3884802 6.976607e-01
training    -1.069402551 0.133153890 -8.0313279 9.642328e-16
> qt(0.975, geriatric.glm$df.residual)
[1] 1.985251
```

As each of the terms in the model involves only a single parameter (both of the two factor variables only had 2 levels), the z (or t) tests in the above table of coefficients test the same hypotheses as were tested above and give very similar results.

Note that the signs of the coefficients suggest that **training** (with a negative coefficent) significantly reduces the expected number of ln(**falls**). As ln( ) is a monotonically increasing transformation, a negative coefficient means that as the value of **training** increases from 0 ("education only") to 1 ("training + education"), ln(**falls**) decreases and therefore, the expected number of **falls** also decreases.

The effects of **gender** are very small and as the coefficient is negative this implies slightly fewer, but not significantly fewer, expected falls for males (the 1 level of **gender**) than for females (the 0 level of **gender**). These conclusions about **training** and **gender** are both consistent with the plot of the data on the original scale shown in part (a).

However, the positive coefficients for **balance** and **strength** suggest that an increase in either **balance** and/or **strength** tends to lead to an increase in the number of **falls**. This unusual result is not simply a feature of a model that includes the **training** intervention, because if you fit a model which does not include **training**, then **strength** and **balance** still have positive coefficients.

It might be the case that people who have better ratings on **strength** and **balance** are more active and are therefore more vulnerable to having **falls** (some measure of the activity levels for each patient would be a sensible additional variable), but it is really a question for the researchers who collected these data to consider the underlying science and make sense of these results!

(6 marks)

Question 1 (e)

Using *R* to produce the required predictions from the model in part (a):

```
> new <- data.frame(training=c(0,0,1,1),gender=c(0,1,0,1),
balance=mean(balance),strength=mean(strength))
> new
  training gender balance strength
1        0      0   52.83    60.78
2        0      1   52.83    60.78
3        1      0   52.83    60.78
4        1      1   52.83    60.78
>
> temp <- predict(geriatric.glm, newdata=new, type="link", se.fit=T)
> predictions <- add.ci(geriatric.glm, temp)
>
> exp(predictions$fit)
       1        2        3        4
4.528531 4.322317 1.554253 1.483478
>
> exp(predictions$ci.fit)
     lower    upper
1 3.824871 5.361644
2 3.573308 5.228328
3 1.187007 2.035121
4 1.161083 1.895390
```

Rounding the confidence intervals "outwards" to ensure at least the required precision, we can summarise the results for persons 65 and over, with average strength and balance:

Training	*Gender*	*Average number of falls in 6 months*	*95% Confidence Interval*
Education only	Males	4.5	(3.8, 5.4)
	Females	4.3	(3.5, 5.3)
Education + Training	Males	1.6	(1.1, 2.1)
	Females	1.5	(1.1, 1.9)

These results are consistent with both the discussion in part (d) above and the graph shown in part (a). Given that most of the underlying assumptions about the residuals appear to be satisfied, as discussed in parts (b) and (c) above, I would be reasonably confident in using this model to make this sort of inference (especially around the mean of the data).

However, the slight patterns on the residual plot noted in part (b) do suggest some remaining correlation between the X variables and the response variable, so the model is not perfect. Given that the assumption that the distribution of the response variable is Poisson appears reasonable (on the basis that there was no evidence of under or over-dispersion), I suspect this may be a result of some unobserved heterogeneity (i.e. there might be some other, as yet unrecorded, explanatory variable we could measure and include in our model), however, I would need to discuss this further with the researchers.

(3 marks)

Question 2

Question 2 was deliberately designed to be more of an open question with parts (a) to (e) as more of a general guide to the analysis required in most of the examples covered in this course (including the specific tasks in Question1). As a result, I have not produced model solutions, but the appendix of R commands demonstrates one approach that could be taken.

(20 marks)