MAST30027: Modern Applied Statistics

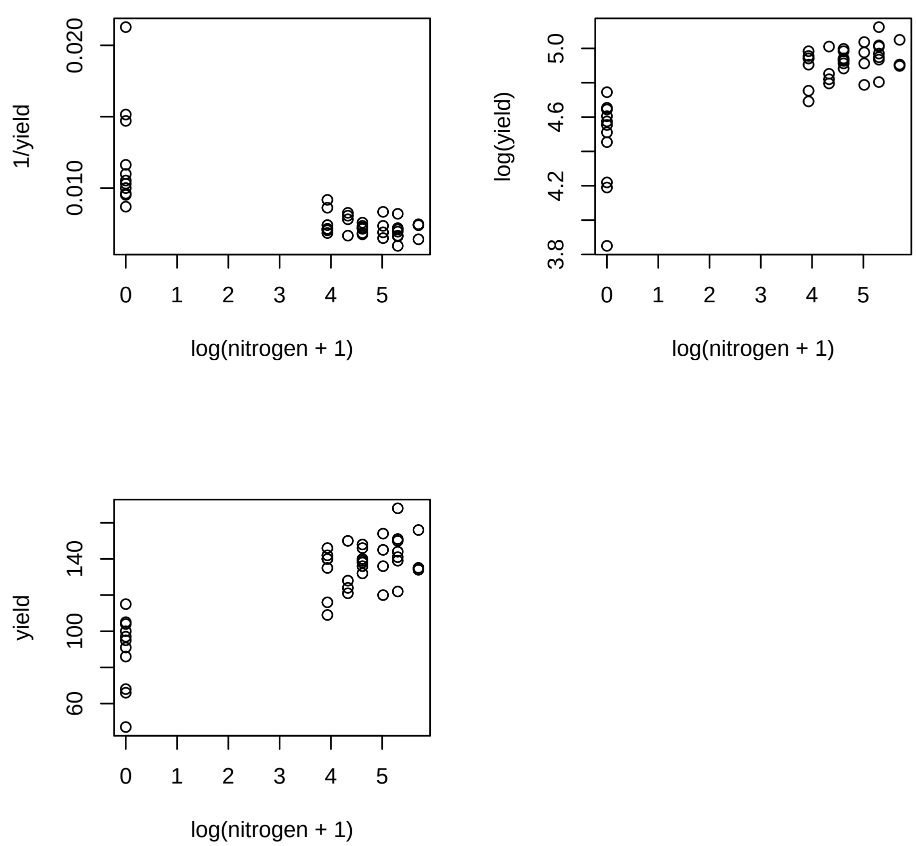
Week 4 Lab

1. The `cornnit` dataset in the `faraway` package contains data on the effect of nitrogen on the yield of corn. Fit a gamma regression to this data, using the `glm` command. You will need to pay attention to the choice of link function (inverse, identity or log), and consider transforming the predictor variable (your first step should be to plot the data).

   Solution: As suggested we plot the data first, using different link functions. It was found that taking a log transform of the nitrogen variable improves the linearity in all cases (note that we add a small constant before taking the log because nitrogren has zero values).

   We suppose that the mean behaves like $g^{-1}(\eta)$, where in this case $\eta = \beta_0 + \beta_1 \log(1 + x)$ and x is the level of nitrogen. Thus a plot of $g(y)$ against $\log(1 + x)$ should look (vaguely) linear.

```
> library(faraway)
> data(cornnit)
> par(mfrow=c(2,2))
> plot(1/yield ~ log(nitrogen+1), data=cornnit)
> plot(log(yield) ~ log(nitrogen+1), data=cornnit)
> plot(yield ~ log(nitrogen+1), data=cornnit)
```



   In all three plots there is an undesirable gap in the observed nitrogen values. We can reduce this a little by using the transform $\log(\text{nitrogen} + k)$ for larger k, but this impinges on the linearity.

   Of the three I think the plot of yield against $\log(\text{nitrogen} + 1)$ looks most linear, but the other two are not unreasonable.

```
> gmod3 <- glm(yield ~ log(nitrogen+1), data=cornnit, family=Gamma(link="identity"))
```

(a) Extract the Pearson residuals $\frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}$ from the fitted model using the `residuals(glmfit, type="pearson")`, then use them to estimate the dispersion parameter ϕ. Check that your answer agrees with the summary output from your model. You can find "Dispersion parameter for Gamma family taken to be ..." in the summary output.

Solution: From the summary we see the dispersion parameter is estimated to be 0.01810, which we can reproduce using Pearson's chi-squared statistic. Note that the model has 42 d.f.

```
> summary(gmod3)

Call:
glm(formula = yield ~ log(nitrogen + 1), family = Gamma(link = "identity"),
    data = cornnit)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.57604  -0.07789   0.02067   0.07948   0.26927

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         88.875      3.571   24.89  < 2e-16 ***
log(nitrogen + 1)   10.337      1.009   10.24 5.46e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.01810187)

    Null deviance: 2.40614  on 43  degrees of freedom
Residual deviance: 0.87603  on 42  degrees of freedom
AIC: 381.71

Number of Fisher Scoring iterations: 4

> (phihat <- sum(residuals(gmod3, "pearson")^2)/42)

[1] 0.01810169
```

(b) Suppose your fitted model is `gmod`, then the command `anova(gmod, test="F")` will compare your model against the null model, using an F test. Using the deviances and dispersion estimates reported by `summary(gmod)`, check that the F statistic reported by the `anova` function is correct.

Solution:

```
> anova(gmod3, test="F")

Analysis of Deviance Table

Model: Gamma, link: identity

Response: yield

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
NULL                                 43    2.40614
log(nitrogen + 1)  1   1.5301        42    0.87603 84.528 1.297e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> model_dev <- .87603
> null_dev <- 2.40614
> (F_statistic <- (null_dev - model_dev)/phihat)
[1] 84.52857
```

2. The `dvisits` data in the `faraway` package comes from the Australian Health Survey of 1977–78 and consist of 5190 observations on single adults, where young and old have been oversampled.

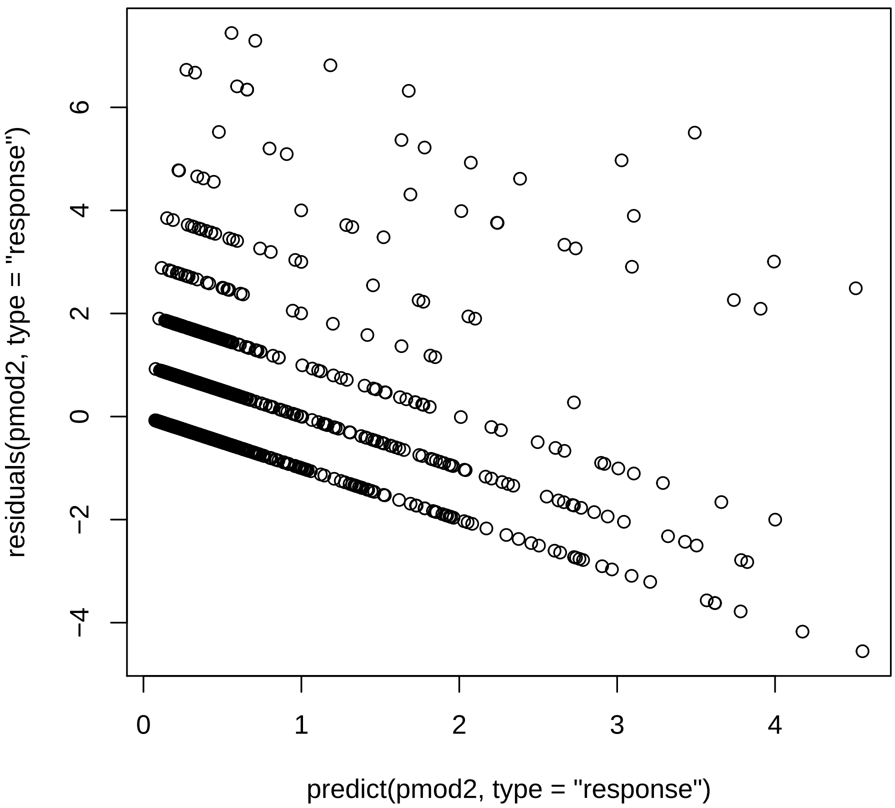
   (a) Build a Poisson regression model with `doctorco` as the response and `sex, age, agesq, income, levyplus, freepoor, freerepa, illness, actdays, hscore, chcond1` as possible predictor variables. Select a model using stepwise model selection based on the AIC. Considering the deviance of the selected model, does this model fit the data? (i.e., is this model adequate?)

   Solution: Using stepwise model selection based on the AIC, we end up with the model `doctorco sex + age + income + levyplus + freepoor + illness + actdays + hscore`. The deviance of 4385.5 is clearly not significant given that we have 5181 degrees of freedom, though note that the responses are not that large, so the deviance may not be close to a chi-squared distribution.

```
> data(dvisits)
> pmod <- glm(doctorco ~ sex + age + agesq + income + levyplus + freepoor
+              + freerepa + illness + actdays + hscore + chcond1,
+              family=poisson, data=dvisits)
> pmod2 <- step(pmod, scope=~., trace=0)
> summary(pmod2)

Call:
glm(formula = doctorco ~ sex + age + income + levyplus + freepoor +
    illness + actdays + hscore, family = poisson, data = dvisits)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0180  -0.6811  -0.5772  -0.4916   5.6590

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.072446   0.100191 -20.685  < 2e-16 ***
sex          0.167591   0.055604   3.014 0.002578 **
age          0.437894   0.137070   3.195 0.001400 **
income      -0.203978   0.084206  -2.422 0.015420 *
levyplus     0.087156   0.053501   1.629 0.103304
freepoor    -0.465788   0.176364  -2.641 0.008265 **
illness      0.196366   0.017603  11.155  < 2e-16 ***
actdays      0.127994   0.004905  26.097  < 2e-16 ***
hscore       0.032854   0.009961   3.298 0.000973 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5634.8  on 5189  degrees of freedom
Residual deviance: 4385.5  on 5181  degrees of freedom
AIC: 6735

Number of Fisher Scoring iterations: 6

> pchisq(deviance(pmod2),5181, lower.tail=FALSE)
[1] 1
```

(b) Extract the response residuals $y_i - \hat{\mu}_i$ from the fitted model using the `residuals(glmfit, type="response")`, then plot the response residuals against the fitted values. Why are there lines of observations on the plot? **Solution:** The lines appear because the response reisuals are given by $y_i - \hat{\mu}_i$ and y_i only takes on finitely many values. Each line corresponds to a different possible value.

```
> plot(predict(pmod2, type="response"), residuals(pmod2, type="response"))
> table(dvisits$doctorco)

   0    1    2    3    4    5    6    7    8    9
4141  782  174   30   24    9   12   12    5    1
```



(c) Starting from the Poisson regression model with `doctorco` as the response and `sex, age, agesq, income, levyplus, freepoor, freerepa, illness, actdays, hscore, chcond1` as possible predictor variables, reduce the model as much as possible using backward elimination with a critical p-value of 5%.

Solution: Using backward elimination and chi-squared tests we end up with the model `doctorco   sex + age + income + freepoor + illness + actdays + hscore`, which is slightly smaller than the model achieved using the AIC and forward-backward elimination (just missing levyplus).

Note that the `step` function uses the AIC, so we have to use `drop1` instead. Here I just give the final step, which shows that we don't need to drop any more variables.

```
> pmod3 <- glm(doctorco ~ sex + age +income + freepoor + illness + actdays
+              + hscore, family=poisson, data=dvisits)
> drop1(pmod3, scope=~., test="Chisq")

Single term deletions

Model:
```

```
doctorco ~ sex + age + income + freepoor + illness + actdays +
    hscore
         Df Deviance    AIC    LRT  Pr(>Chi)
<none>        4388.1 6735.7
sex       1   4398.2 6743.8  10.14  0.001453 **
age       1   4398.2 6743.7  10.06  0.001518 **
income    1   4392.5 6738.1   4.43  0.035274 *
freepoor  1   4397.4 6742.9   9.27  0.002335 **
illness   1   4508.9 6854.5 120.82 < 2.2e-16 ***
actdays   1   4956.5 7302.1 568.41 < 2.2e-16 ***
hscore    1   4398.4 6744.0  10.31  0.001322 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(d) What sort of person would be predicted to visit the doctor the most under your selected model?

Solution Using a log link we have $\mu = e^\eta$, so we wish to maximise $\eta = \mathbf{x}^T\beta$. Looking at the coefficients this means female; as old as possible; no income; not entitled to free health care; very ill in the past two weeks; many days of reduced activity in the last two weeks; and a high hscore.

```
> pmod3$coefficients

(Intercept)         sex         age      income    freepoor     illness
-2.05196250  0.17552865  0.43353243 -0.17105283 -0.49632492  0.19600786
    actdays      hscore
 0.12779329  0.03243268
```

(e) For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0,1,2, etc. times.

Solution:

```
> dim(dvisits)

[1] 5190   19

> lambda <- exp(predict(pmod3, dvisits[5190,]))
> dpois(0:9, lambda)

 [1] 8.451821e-01 1.421623e-01 1.195608e-02 6.703505e-04 2.818878e-05
 [6] 9.482888e-07 2.658420e-08 6.387927e-10 1.343087e-11 2.510129e-13
```