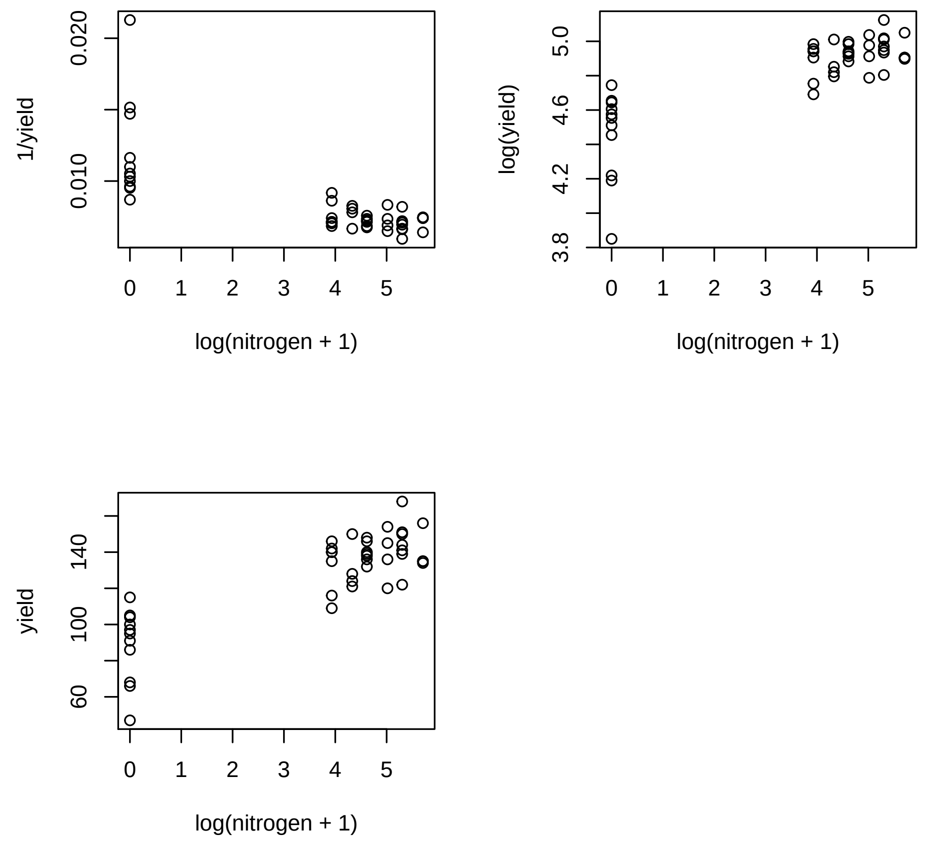
MAST30027: Modern Applied Statistics

Week 5 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, usnig the `glm` command. You will need to pay attention to the choice of link function, and consider transforming the predictor variable (your first step should be to plot the data).

   Solution: As suggested we plot the data first. A log transform of the nitrogen variable improves the linearity (note that we add a small constant before taking the log because nitrogren has zero values).

```
> 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)
```

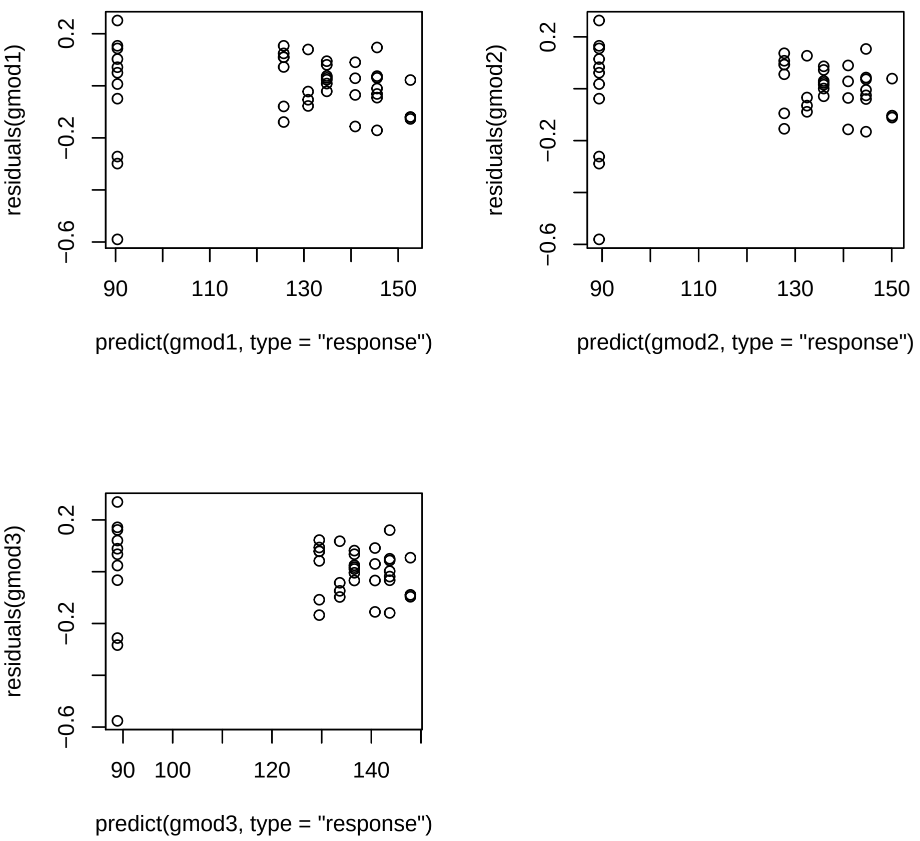


   In all three plots there is an undesirable gap in the observed nitrogen values. We can reduce this a little by using the transform log(nitrogen + k) for larger k, but this impinges on the linearity.

   In the first plot there is noticably more variance when nitrogen is zero, but this is not necessarily a problem, as in a gamma model the variance is proportional to the mean squared. This means that when using the inverse link, small values of $\eta = \mathbf{x}^T\beta$ will correspond to large means.

Of the three I think the plot of yield against log(nitrogen + 1) looks most linear, but the other two are not unreasonable. Accordingly we will try all three link functions are compare the residuals.

```
> par(mfrow=c(2,2))
> gmod1 <- glm(yield ~ log(nitrogen+1), data=cornnit, family=Gamma(link="inverse"))
> plot(predict(gmod1,type="response"), residuals(gmod1))
> gmod2 <- glm(yield ~ log(nitrogen+1), data=cornnit, family=Gamma(link="log"))
> plot(predict(gmod2,type="response"), residuals(gmod2))
> gmod3 <- glm(yield ~ log(nitrogen+1), data=cornnit, family=Gamma(link="identity"))
> plot(predict(gmod3,type="response"), residuals(gmod3))
```



There is not much difference between these plots. In all three cases there is more variation for responses with zero nitrogen, which we don't want, but there's not much we can do about this. If forced to choose, I would go for the identity link, which I will use for the remainder of the question. If we look at the AIC for each model we also see that it is smallest for the model with the identity link (just)

```
> gmod1$aic

[1] 383.7435

> gmod2$aic

[1] 382.4205

> gmod3$aic

[1] 381.7124
```

(a) Extract the Pearson residuals from the fitted model using the `residuals` function, then use them to estimate the dispersion parameter. Check that your answer agrees with the summary output from your model.

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.

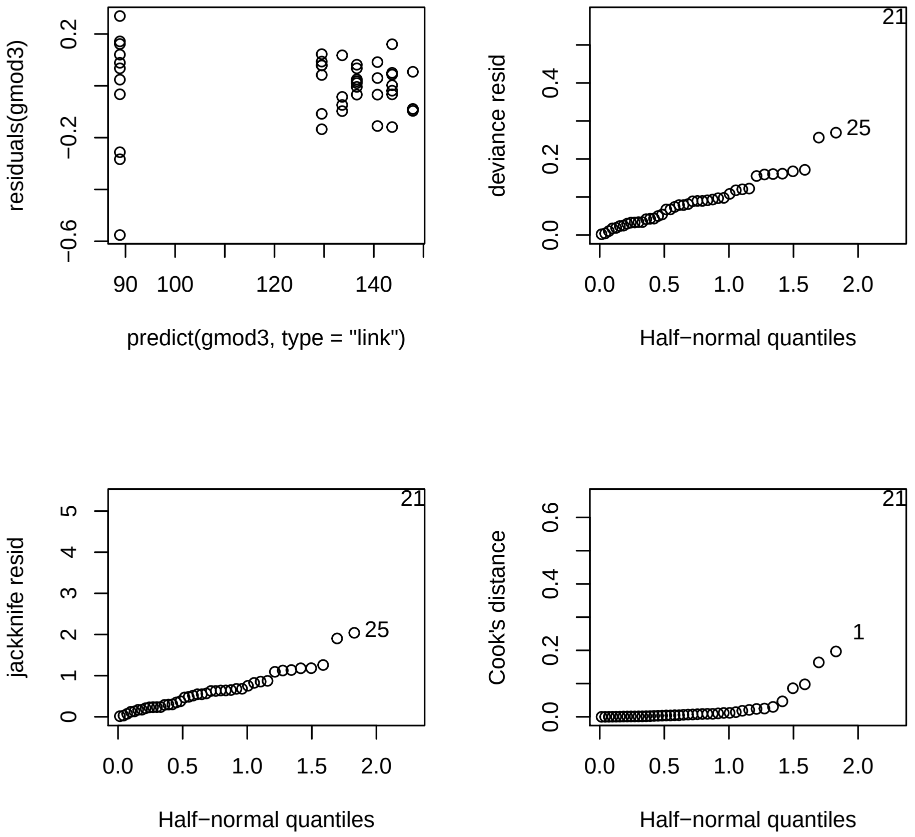
Solution:

```
> 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
```

(c) Now do some diagnostic plots. Can you identify a potential outlier?

Solution: We have already observed more variation than we would like when nitrogen is zero. It also looks like point 21 could be an outlier.

```
> par(mfrow=c(2,2))
> plot(predict(gmod3, type="link"), residuals(gmod3))
> halfnorm(residuals(gmod3), ylab="deviance resid")
> halfnorm(rstudent(gmod3), ylab="jackknife resid")
> halfnorm(cooks.distance(gmod3), ylab="Cook's distance")
```



(d) Fit a linear model to the `cornnit` data.

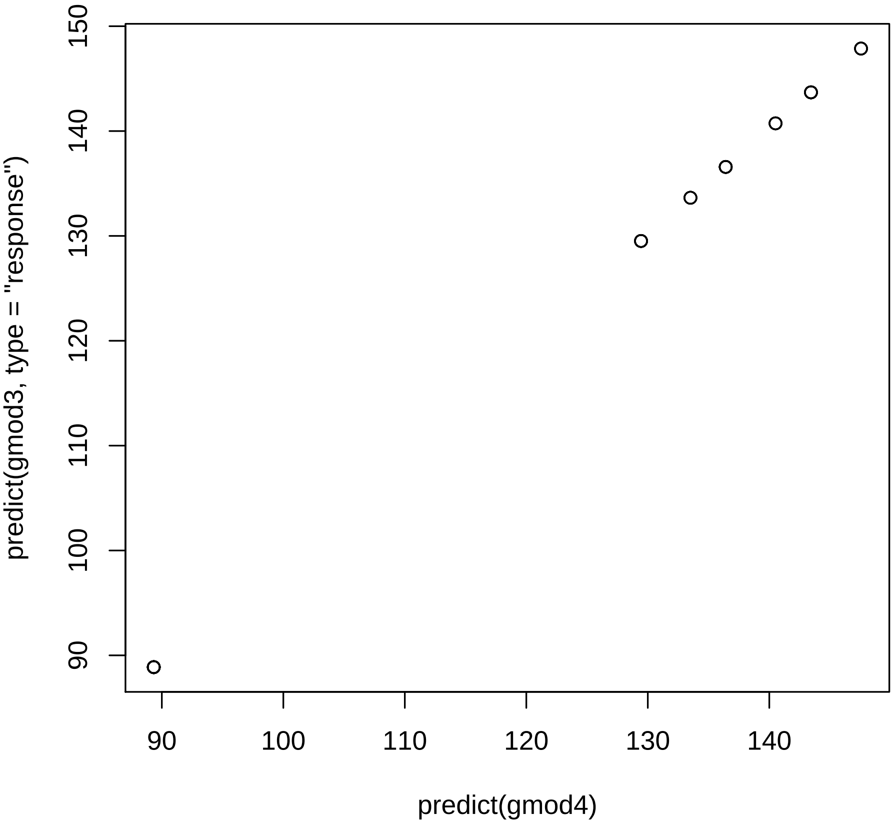
Which do you prefer, the linear model or the gamma model, and why?

Solution: A gamma variable with a large mean looks a lot like a normal, so we expect a linear model to look a lot like our gamma model, and it does. Given this, we may as well go with the linear model.

```
> gmod4 <- lm(yield ~ log(nitrogen+1), data=cornnit)
> summary(gmod4)

Call:
lm(formula = yield ~ log(nitrogen + 1), data = cornnit)

Residuals:
    Min      1Q  Median      3Q     Max 
-42.335 -10.261   2.126  10.558  25.665 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         89.335      4.227   21.13  < 2e-16 ***
```

```
log(nitrogen + 1)   10.201      1.017   10.03 1.03e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.34 on 42 degrees of freedom
Multiple R-squared:  0.7055,       Adjusted R-squared:  0.6985
F-statistic: 100.6 on 1 and 42 DF,  p-value: 1.025e-12
```

```
> par(mfrow=c(1,1))
> plot(predict(gmod4), predict(gmod3, type="response"))
```



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` and `chcond2` as possible predictor variables. Considering the deviance of this model, does this model fit the data?

   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.

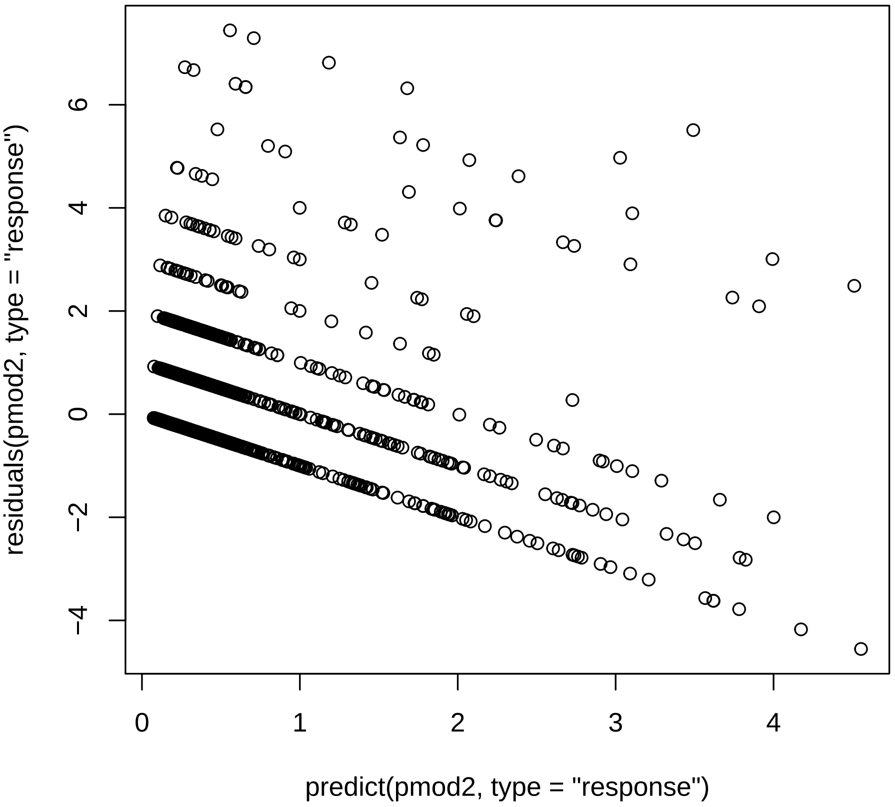
```
> data(dvisits)
> pmod <- glm(doctorco ˜ sex + age + agesq + income + levyplus + freepoor
+              + freerepa + illness + actdays + hscore + chcond1,
+              family=poisson, data=dvisits)
```

```
> 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
```

(b) 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 $y_i - g(\eta_i)$ 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) Use backward elimination with a critical p-value of 5% to reduce the model as much as possible.

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 a 10% significance level, so we have to do the final step manually.

```
> pmod3 <- step(pmod, scope=~., direction="backward", test="Chisq", trace=0)
> summary(pmod3)

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
```

```
> pmod4 <- glm(doctorco ~ sex + age +income + freepoor + illness + actdays
+               + hscore, family=poisson, data=dvisits)
> drop1(pmod4, 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.

```
> pmod4$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(pmod4, dvisits[5190,]))
> dpois(0:9, lambda)
```

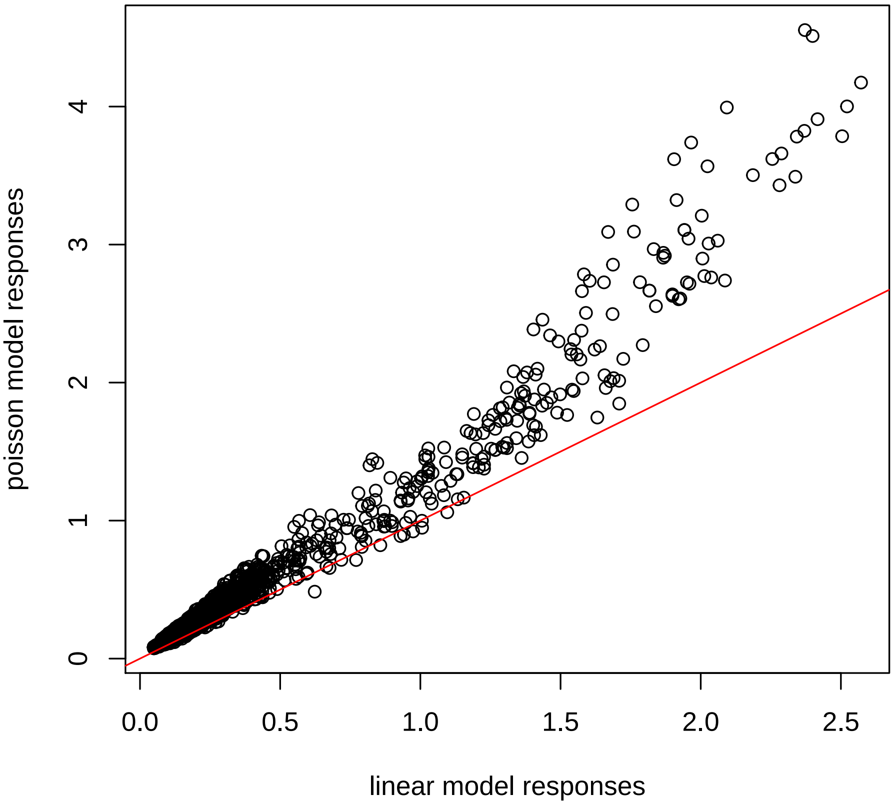
```
 [1] 8.451821e-01 1.421623e-01 1.195608e-02 6.703505e-04 2.818878e-05
 [6] 9.482888e-07 2.658420e-08 6.387927e-10 1.343087e-11 2.510129e-13
```

(f) Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how the Gaussian and Poisson models differ.

Solution: We get a better fit by taking the log of the response (offset by 0.1, as the response can take zero values). The resulting linear model produces fitted values a lot like those of the poisson model.

Note that the mean of a log-normal random variable is given by $\exp(\mu + \sigma^2/2)$. Thus if Y is log-normal, and we estimate the mean μ and variance σ^2 of $\log(Y)$, then our estimate for $\mathbb{E}Y$ is $\exp(\hat{\mu} + \hat{\sigma}^2/2)$.

```
> mod <- lm(log(doctorco + .1) ~ sex + age + agesq + income + levyplus
+           + freepoor + freerepa + illness + actdays + hscore + chcond1,
+           data=dvisits)
> mod2 <- step(mod, scope=~., trace=0)
> mod2si2 <- deviance(mod2)/mod2$df.residual
> plot(exp(predict(mod2) + mod2si2/2) - .1, predict(pmod3, type="response"),
+      xlab="linear model responses", ylab="poisson model responses")
> abline(0, 1, col="red")
```



Although the linear model does surprisingly well, its fitted values are all a little smaller than the corresponding fitted values for the poisson model. The most important difference between how these two models are fitted is their variance structure. The poisson model assumes that $\mathrm{Var}\,Y \propto \mathbb{E}Y$ and the linear model assumes that $\mathrm{Var}\,\log Y$ and hence $\mathrm{Var}\,Y$ is constant. Thus the linear model will be giving too much weight to large responses.

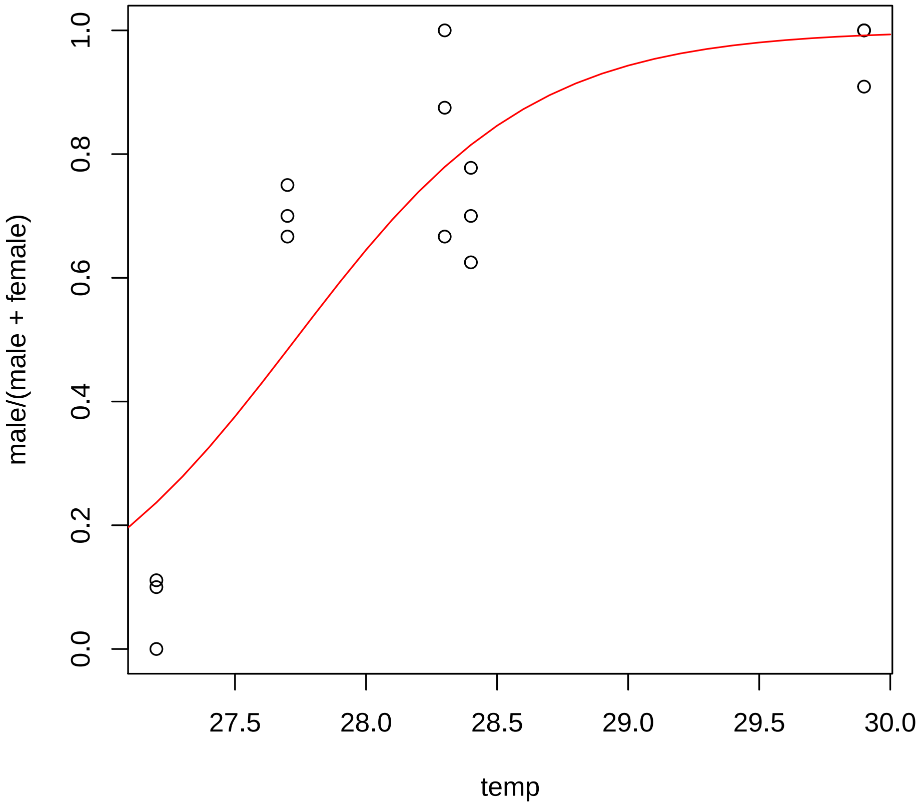
3. Incubation temperature can affect the sex of turtles. An experiment was conducted with three independent replicates for each temperature and the number of male and female turtles born was recorded. The data can be found in the `turtle` dataset in the `faraway` package.

   Check for evidence of overdispersion in a binomial model for the sex of the turtle.

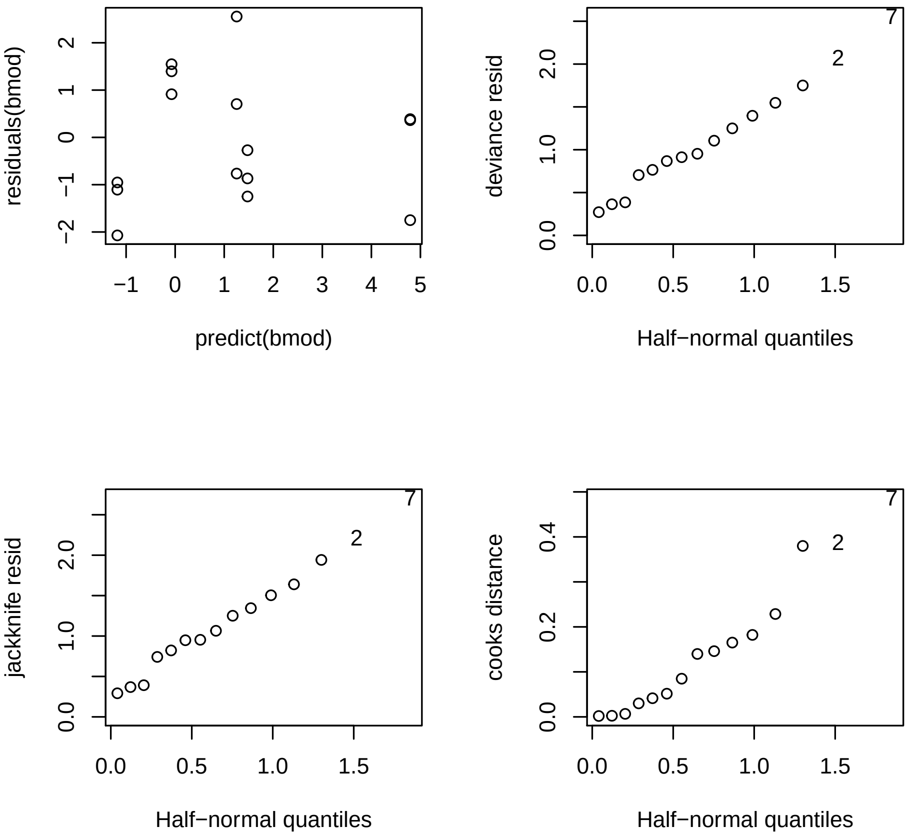
   What problems can arise if you ignore overdispersion?

   Solution: We fit a binomial regression and estimate the dispersion ϕ.

```
> data(turtle)
> with(turtle, plot(temp, male/(male+female)))
> bmod <- glm(cbind(male, female) ~ temp, data=turtle, family=binomial)
> summary(bmod)

Call:
glm(formula = cbind(male, female) ~ temp, family = binomial,
    data = turtle)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0721  -1.0292  -0.2714   0.8087   2.5550

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -61.3183    12.0224  -5.100 3.39e-07 ***
temp          2.2110     0.4309   5.132 2.87e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.508  on 14  degrees of freedom
Residual deviance: 24.942  on 13  degrees of freedom
AIC: 53.836

Number of Fisher Scoring iterations: 5

> t <- seq(27, 30, .1)
> lines(t, ilogit(-61.31 + 2.211*t), col="red")
> (phihat <- sum( residuals(bmod, type="pearson")^2 )/13)

[1] 2.018641
```



```
> par(mfrow=c(2,2))
> plot(predict(bmod), residuals(bmod))
> halfnorm(residuals(bmod), ylab="deviance resid")
> halfnorm(rstudent(bmod), ylab="jackknife resid")
> halfnorm(cooks.distance(bmod), ylab="cooks distance")
```



The fit isn't brilliant, but there is no evidence of outliers. $\hat{\phi}$ is a little bit larger than 1, indicating possible overdispersion. Moreover, if we perform a chi-squared test for model sufficiency using the deviance, we get a significant result, indicating that there is something left unexplained.

```
> pchisq(24.942, 13, lower.tail=F)

[1] 0.02349208
```

If we do not account for overdispersion, then our tests for variable significance will be too sensitive. That is, they may indicate a variable is significant when it really isn't. Q. 3 on the Week 4 was a good example of this: when the dispersion was included the interaction no longer appeared significant. Similarly, if we do not account for overdispersion, confidence intervals for parameter estimates will be too small

4. In a binomial model we assume that any given observation is from a bin(m, p) distribution, for some m and p. That is, we count the number of successes from m i.i.d. bernoulli(p) trials. There are two main ways that overdispersion can arise: the trials are not identically distributed, or the trials are not independent.

   A common way in which we can have heterogeneous trials is via clustering. Suppose that we have m trials split into h clusters of size $k = m/h$, and that the probability of success for a trial in the i-th cluster is p_i. Now suppose that p_i is random, with $\mathbb{E}p_i = p$ and $\operatorname{Var} p_i = \tau^2 p(1-p)$. Let the number of successes from cluster i be Z_i and let the total number of successes be $Y = Z_1 + \cdots + Z_h$.

   Show that

   (a)

$$\mathbb{E}Y = mp$$

(b)

$$\operatorname{Var} Y = (1 + (k-1)\tau^2)mp(1-p)$$

Hint: $\operatorname{Var} Y = \mathbb{E}\operatorname{Var}(Y|X) + \operatorname{Var}\mathbb{E}(Y|X)$.

Thus Y is overdispersed, relative to a binomial.

Solution:

$$\begin{aligned}
\mathbb{E}Y &= \sum_{i=1}^{h} \mathbb{E}Z_i = \sum_{i=1}^{h} \mathbb{E}\mathbb{E}(Z_i|p_i) \\
&= \sum_{i=1}^{h} \mathbb{E}kp_i = \sum_{i=1}^{h} kp = mp
\end{aligned}$$

By independence $\operatorname{Var} Y = \sum_i \operatorname{Var} Z_i$, and

$$\begin{aligned}
\operatorname{Var} Z_1 &= \mathbb{E}\operatorname{Var}(Z_1|p_1) + \operatorname{Var}\mathbb{E}(Z_1|p_1) \\
&= \mathbb{E}kp_1(1-p_1) + \operatorname{Var} kp_1 \\
&= kp - k(\tau^2 p(1-p) + p^2) + k^2\tau^2 p(1-p) \\
&= kp(1-p)(1+(k-1)\tau^2).
\end{aligned}$$

Multiplying by h gives the result.

5. Show that if you put $\operatorname{Var} y = \phi\mu(1-\mu)$, where $\mu = \mathbb{E}y \in [0,1]$, then the log quasi-likelihood is

$$Q(\mu; y) = \frac{1}{\phi}\left(y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu) + c\right),$$

where c does not depend on μ or ϕ

Solution:

$$\begin{aligned}
Q &= \int_y^\mu \frac{y-t}{\phi v(t)} dt \\
&= \frac{1}{\phi}\int_y^\mu \frac{y-t}{t(1-t)} dt \\
&= \frac{1}{\phi}\left(\int_y^\mu \frac{y}{t(1-t)} dt - \int_y^\mu \frac{1}{1-t} dt\right) \\
&= \frac{1}{\phi}\left(\left[y \log \frac{t}{1-t}\right]_y^\mu + [\log(1-t)]_y^\mu\right) \\
&= \frac{1}{\phi}\left(y \log \frac{\mu}{1-\mu} + \log(1-\mu) - y \log \frac{y}{1-y} - \log(1-y)\right)
\end{aligned}$$