DRAFT SOLUTIONS TO STOR 556 MIDTERM 1

RICHARD SMITH, FEBRUARY 25 2019

Question 1:

a. The initial regression model was "lm1" which was then reduced using the "step" command to "lm2", as follows:

```
Call:
glm(formula = low ~ lwt + race + smoke + ptl + ht + ui, family = binomi
al,
    data = birthwt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8915  -0.8121  -0.5471   1.0389   2.1140

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.803644   1.041394  -0.772  0.44029
lwt         -0.012947   0.006555  -1.975  0.04823 *
race         0.469131   0.211799   2.215  0.02676 *
smoke        0.948172   0.395139   2.400  0.01641 *
ptl          0.491451   0.341378   1.440  0.14998
ht           1.833160   0.690234   2.656  0.00791 **
ui           0.747935   0.459942   1.626  0.10392
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 205.21  on 182  degrees of freedom
AIC: 219.21

Number of Fisher Scoring iterations: 4
```

Note that two of the eight original variables (age and ftv) are not significant under this model, and two others (ptl and ui0) do not appear statistically significant. Another acceptable solution would be if you omitted those variables.

b. All of the variables in the above model are factor variables except for "lwt", but the distribution of lwt is already close to normal (could illustrate this with a histogram or QQ plot), so there does not appear to be a strong case for transformation. One change you could make is to recode "race" as a factor variable (it has three levels, 1/2/3 representing white/black/other, but they are coded as an integer variable). However this would be a fairly small change in the model so I have not pursued it here (you will get credit if you did this).

The one thing that really does seem to make a difference is to include quadratic/interaction terms, e.g.

```
> lm3=glm(low~(age+lwt+race+smoke+ptl+ht+ui+ftv)^2,family=binomial,birthwt)
```

```
> lm4=step(lm3,trace=0)
> summary(lm4)

Call:
glm(formula = low ~ age + lwt + smoke + ptl + ht + ui + ftv +
    age:smoke + age:ptl + age:ftv + lwt:smoke + ptl:ui, family = binomial,
    data = birthwt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2032  -0.7100  -0.4653   0.8778   2.2099

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.66572    1.80405   0.369 0.712115
age          0.05105    0.06696   0.762 0.445820
lwt         -0.02659    0.01099  -2.420 0.015504 *
smoke       -5.36576    2.69399  -1.992 0.046398 *
ptl          5.78531    2.66296   2.173 0.029817 *
ht           1.80748    0.74821   2.416 0.015704 *
ui           1.86388    0.57541   3.239 0.001199 **
ftv          3.65960    1.08001   3.388 0.000703 ***
age:smoke    0.14949    0.09021   1.657 0.097505 .
age:ptl     -0.19022    0.10813  -1.759 0.078549 .
age:ftv     -0.16568    0.04960  -3.340 0.000837 ***
lwt:smoke    0.02143    0.01445   1.483 0.138101
ptl:ui      -2.00861    0.78387  -2.562 0.010395 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 183.39  on 176  degrees of freedom
AIC: 209.39

Number of Fisher Scoring iterations: 5
```

The AIC of the final model (209.39) is substantially less than that of the first model (219.21) and the residual deviance also suggests a good fit.

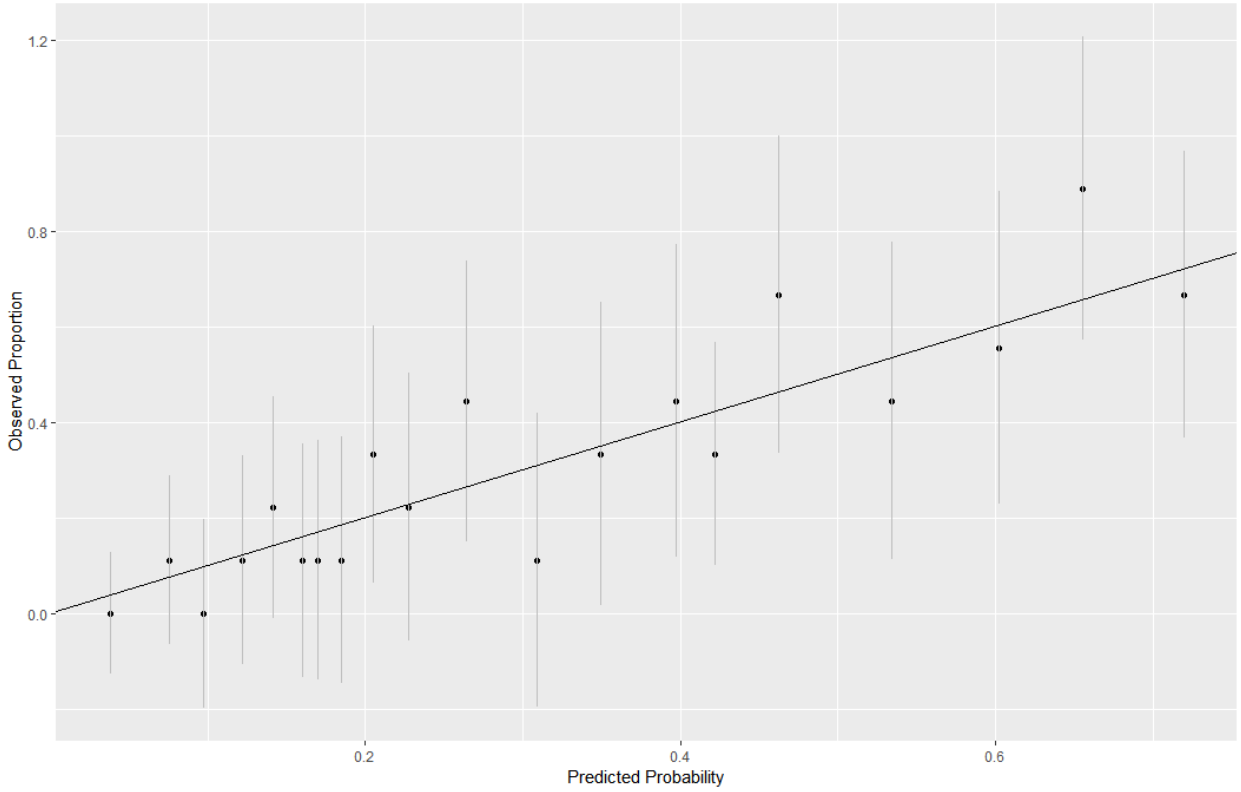
c. The Hosmer-Lemeshow test can be implemented as follows:

```
> library(ResourceSelection)
ResourceSelection 0.3-4        2019-01-08
> hoslem.test(lm4$y,lm4$fitted, g = 20)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  lm4$y, lm4$fitted
X-squared = 14.647, df = 18, p-value = 0.6861
```

This indicates no departure from the assumed distribution (I tried alternative values of g, the number of groups, but none showed a statistically significant result). As for a graphical

Illustration, there are various things you could do, but here is a sequence of code designed to reproduce Figure 2.9 of the course text (again based on a split into 20 groups):

```
> library(ggplot2)
> library(dplyr)
> linpred=predict(lm4)
> predprob=predict(lm4,type='response')
> birthwt=mutate(birthwt,predprob=predict(lm4,type='response'))
> gdf=group_by(birthwt,cut(linpred,breaks=unique(quantile(linpred,(1:20)/21))
))
> hldf=summarise(gdf,low=sum(low),ppred=mean(predprob),count=n())
> hldf=mutate(hldf,se.fit=sqrt(ppred*(1-ppred)/count))
> ggplot(hldf,aes(x=ppred,y=low/count,ymin=low/count-2*se.fit,ymax=low/count+
2*se.fit))+geom_point()+
+     geom_linerange(color=grey(0.75))+geom_abline(intercept=0,slope=1)+xlab(
'Predicted Probability')+
+     ylab("Observed Proportion")
```



d. Use the "predict.glm" command as follows: Note that this is calculating the predicted probability on the "link" scale (I could have explicitly inserted "type=link") as this is the scale of the linear model and not the scale of predicted probability.

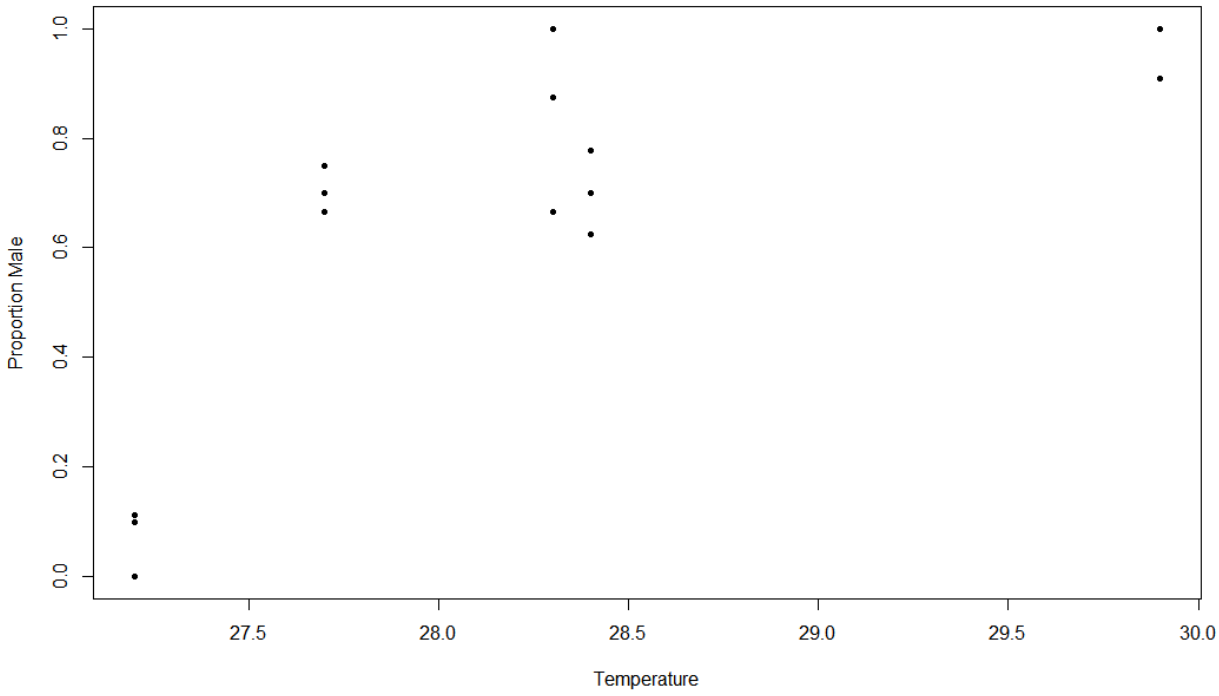
```
> newd=data.frame(age=31, lwt=160, race=2, smoke=0,ptl=1, ht=0, ui=0, ftv=1)
> predict(lm4,newdata=newd,se.fit=T)
```

```
$fit
[1] -3.59437

$se.fit
[1] 1.130546

$residual.scale
[1] 1
```

The predicted value has a mean -3.59437, standard error 1.13056, so a 95% confidence interval (assuming a normal underlying distribution) would be -3.59437 plus or minus 1.96*1.13056, which leads to (-5.810268,-1.378472). Applying the "ilogit" function to the central value and each of the two endpoints, we conclude that the predicted probability is 0.0267, with a 95% confidence interval from 0.0030 to 0.2013.

Question 2:

a. For the plot, see below.



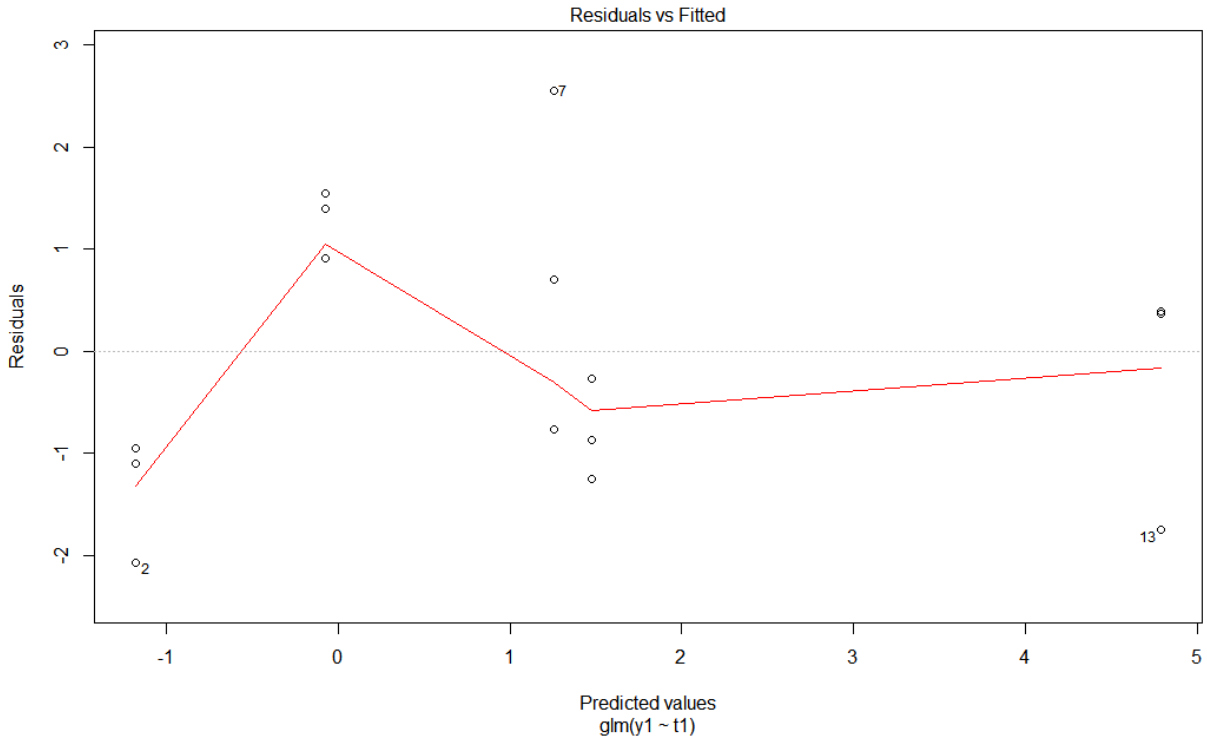
Clearly, the biggest jump is between the lowest temperature (27.2) and the other four, though there's a small linear increase over the last four temperatures as well.

b. The linear fit is as follows:

```
> y1=cbind(turtle$male,turtle$female)
> t1=turtle$temp
> lm1=glm(y1~t1,family=binomial)
>
> summary(lm1)
```

```
Call:
glm(formula = y1 ~ t1, family = binomial)

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 ***
t1            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:
```

The linear term is highly significant (as we would expect) but a residuals v. fitted values plot shows a clear lack of fit, for example,



The residuals of the first group are all below the zero line, while those of the second group are all above, indicating a lack of fit.

c. Including a quadratic term clearly improves the fit of the model, for example

```
> t2=t1^2
> lm3=glm(y1~t1+t2,family=binomial)
> summary(lm3)

Call:
glm(formula = y1 ~ t1 + t2, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6703  -0.8875  -0.4194   0.9481   2.2198

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -677.5950   268.7984  -2.521   0.0117 *
t1            45.9173    18.9169   2.427   0.0152 *
t2            -0.7745     0.3327  -2.328   0.0199 *
---
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: 20.256  on 12  degrees of freedom
AIC: 51.15

Number of Fisher Scoring iterations: 4
```

The t1 and t2 are both statistically significant with a p-value slightly over 0.01, and the model is a satisfactory fit as judged by deviance: for example

```
> 1-pchisq(20.256,12)
[1] 0.06239564
```

shows a p-value (0.06) that indicates acceptance of the null hypothesis (that the model is correct) but is perhaps small enough to justify further investigation. At this point you could fit a quasibinomial model which indicates a moderate dispersion (1.4) but the increase in p-values makes neither t1 nor t2 significant. So it’s unclear what to do at this point.

d. The question is asking you to compare the theoretical with the observed variation within each group of three observations at the same temperature. You can decide for yourself what measure of variation to use, but variance is the most obvious so I’ll stick with that. Here’s a small self-contained piece of R code (you don’t have to do it this way but it’s one possible way of answering the question):

```
> nums=turtle$male+turtle$female
> props=turtle$male/(turtle$male+turtle$female)
> probs=lm3$fitted.values
> vars=probs*(1-probs)/nums
> V=matrix(nrow=5,ncol=2)
> for(i in 1:5){
+     V[i,1]=mean(vars[3*(i-1)+1:3])
+     V[i,2]=var(props[3*(i-1)+1:3])
```

```
+ }
```

The first two lines compute the number of turtles and the proportion of males within each row of data. The third line computes the fitted probability for each row and the fourth converts that to a theoretical variance, using the formal p(1-p)/n for the variance for proportions from the Binomial distribution. The remaining code computes a matrix V where, for each row,

- the first column is the mean of the three theoretical variances for that group of observations (they are not identical because the sample sizes vary within each group);
- the second column is the observed sample variance.

The resulting V matrix looks like this:

```
> V
            [,1]        [,2]
[1,] 0.015058343 0.003744856
[2,] 0.032625718 0.001759259
[3,] 0.014904138 0.028356481
[4,] 0.013959688 0.005835905
[5,] 0.005329255 0.002754821
```

In four out of the five rows (excepting row 3), the observed variance is smaller than the theoretical variance, which implies no evidence of overdispersion.

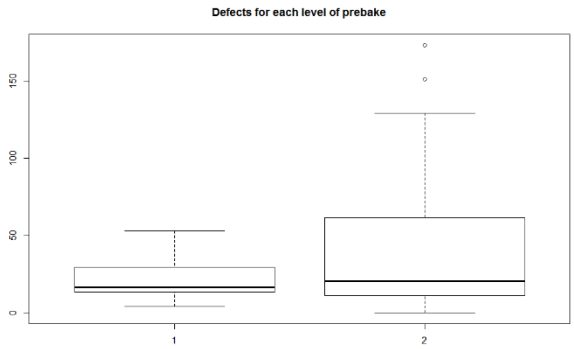
This is a little hard to reconcile with the result of the quasibinomial fit which does indicate mild overdispersion (though not "statistically significant" when judged by the deviance test). It may be that the conclusion is unduly influenced by row 3 which has the largest observed variance, or maybe it is saying indirectly that even the quadratic model does not fit the data too well. Either way, I think some variant of the following conclusion is the most appropriate: there is no overall evidence for overdispersion, and therefore, "lm3" (which includes a quadratic term in temperature) is probably the best model overall.

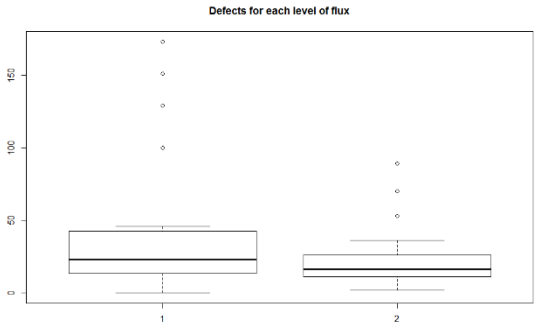
Question 3:

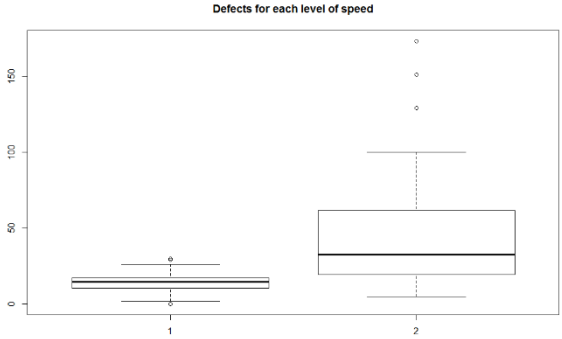
a. The question hints at some recoding, because each row of data is really three rows (each of y1, y2, y3 is an independent response to the corresponding set of covariates). Here is one (by no means unique) way of recoding:

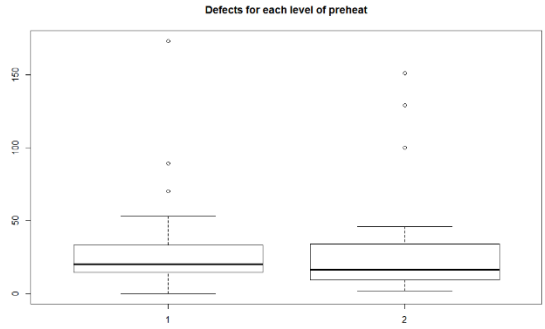
```
> data(wavesolder)
>
> covars=data.frame(rbind(wavesolder[,4:10],wavesolder[,4:10],wave
solder[,4:10]))
> y=c(wavesolder$y1,wavesolder$y2,wavesolder$y3)
> plot(covars$prebake,y,main='Defects for each level of prebake')
```

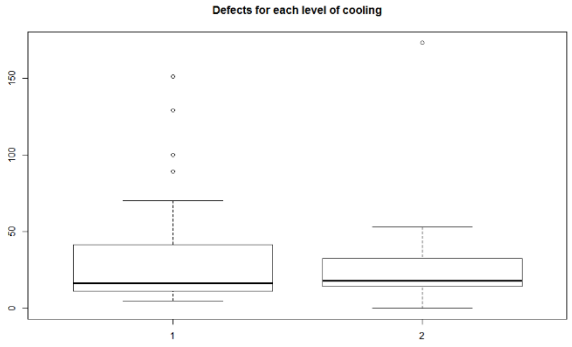
This can be repeated with "prebake" replaced in succession by each of the other covariates. The resulting plots are shown below. There is clear evidence of a dependency on prebake, speed and

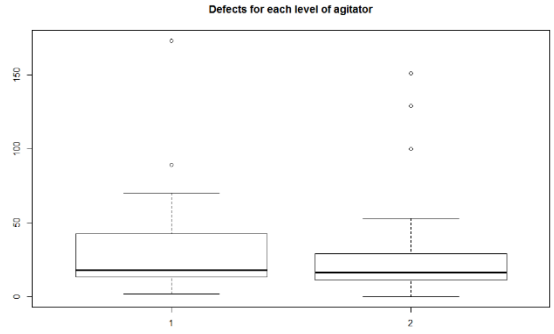

Defects for each level of prebake
0
50
100
150
1
2

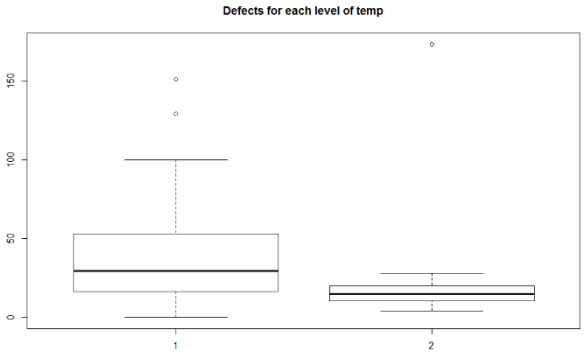

Defects for each level of flux
0
50
100
150
1
2


Defects for each level of speed
0
50
100
150
1
2


Defects for each level of preheat
0
50
100
150
1
2


Defects for each level of cooling
0
50
100
150
1
2


Defects for each level of agitator
0
50
100
150
1
2


Defects for each level of temp
0
50
100
150
1
2

temp, but the others are less clear-cut.

b. The mean, variance and variance/mean ratio for y1, y2, y3 in each of the 16 rows of the original data are as follows:

```
 Mean  Variance Ratio
  23.00   79.00   3.43
  10.33   36.33   3.52
  18.33    8.33   0.45
  43.67    4.33   0.10
  15.33    2.33   0.15
  14.33   14.33   1.00
  39.33  152.33   3.87
  10.00   31.00   3.10
  14.33  210.33  14.67
  15.00   91.00   6.07
  73.33 7470.33 101.87
 126.67  654.33   5.17
  12.33    5.33   0.43
  12.00   75.00   6.25
  70.67  324.33   4.59
  17.33   80.33   4.63
```

In 11 of the 16 rows the variance is greater than the mean, suggesting overdispersion, though the evidence is dominated by a few extreme rows (in particular, row 11) which suggests vulnerability to outliers.

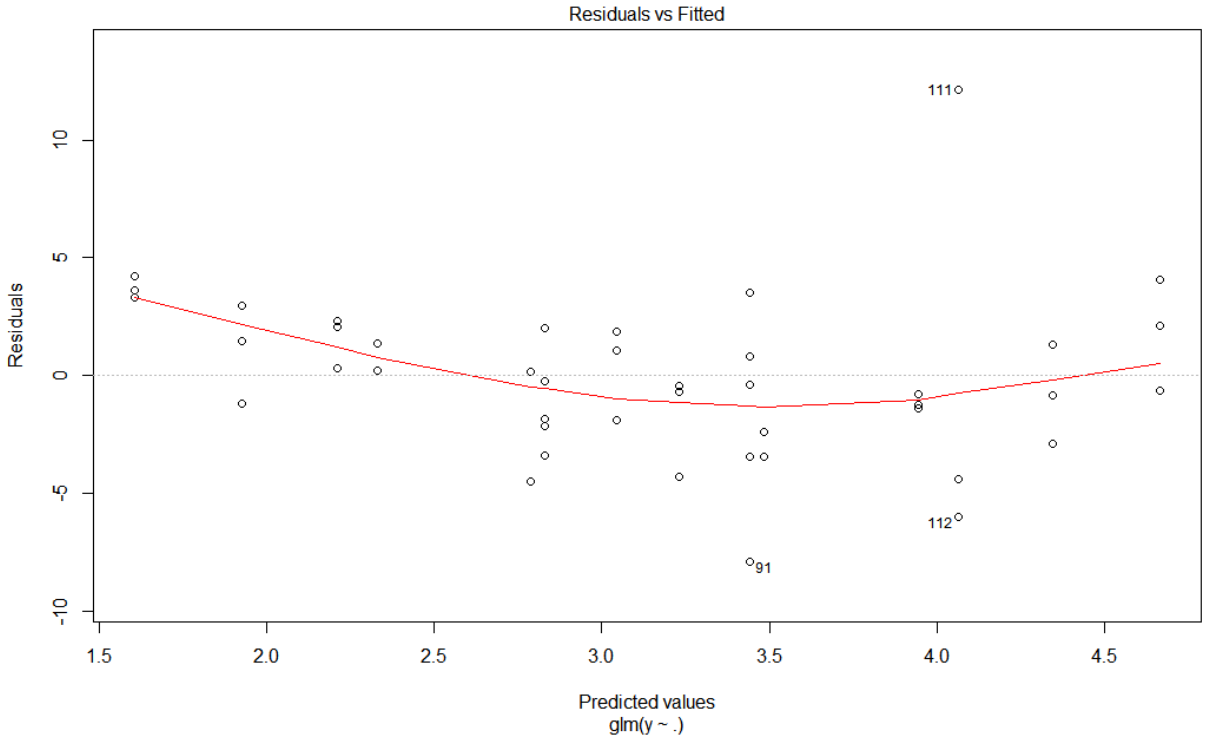
c. See the following:

```
> lm1=glm(y~.,family=poisson,covars)
> summary(lm1)

Call:
glm(formula = y ~ ., family = poisson, data = covars)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.9074  -2.1202  -0.4121   1.5501  12.1486

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.04282    0.08560  35.548  < 2e-16 ***
prebake2     0.65124    0.05442  11.968  < 2e-16 ***
flux2       -0.57667    0.05438 -10.604  < 2e-16 ***
speed2       1.22875    0.06113  20.100  < 2e-16 ***
preheat2    -0.16645    0.05343  -3.115  0.00184 **
cooling2    -0.16187    0.05313  -3.047  0.00231 **
agitator2   -0.08974    0.05263  -1.705  0.08819 .
temp2       -0.69816    0.05537 -12.609  < 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: 1450.52  on 47  degrees of freedom
Residual deviance:  491.78  on 40  degrees of freedom
```

```
AIC: 738.52

Number of Fisher Scoring iterations: 5
```

This analysis suggests that all the covariates are significant, with the possible exception of "agitator."

d. See the following:



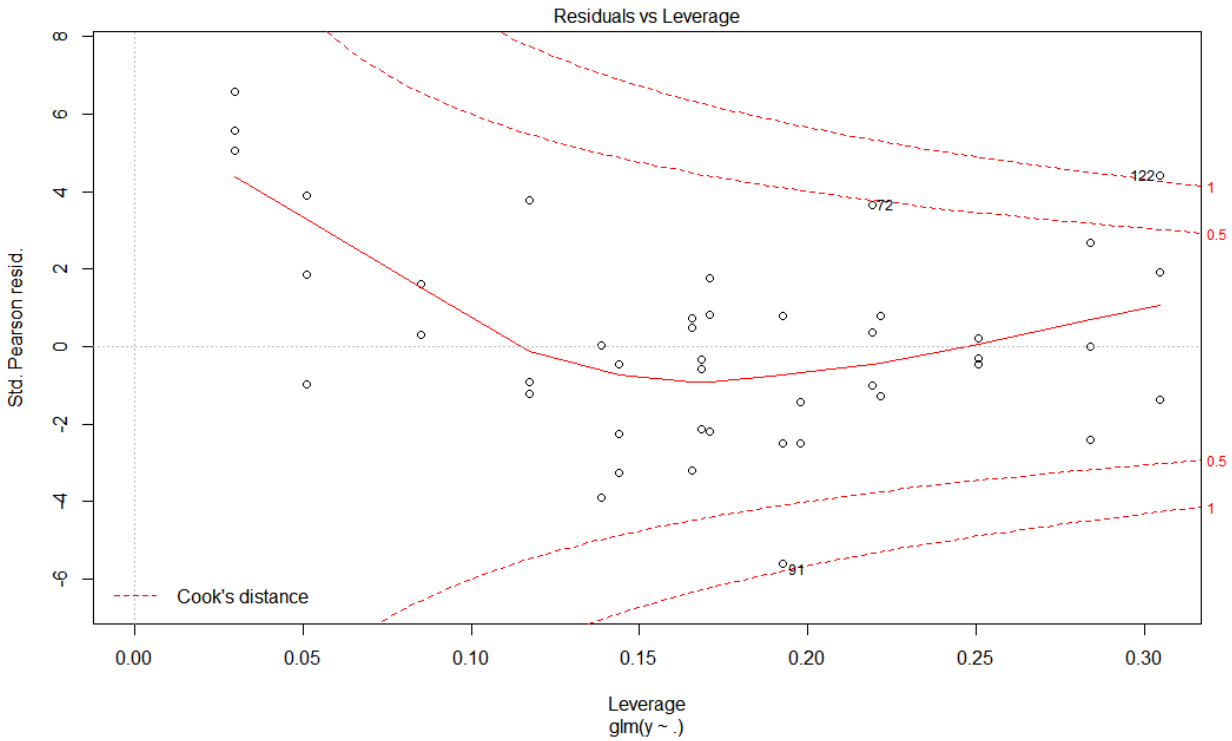
There is clear evidence of an outlier in observation 111. (Although not shown here, both the QQ plot of residuals and the residuals v. leverage plot confirm this.) Although it's not immediately clear from the original dataset which one is being labelled observation 111, a quick tabulation confirms that it is in fact observation y2 from row 11 (173 – the second largest value in the dataset is 151).

e. One way to refit the model without this data point is as follows:

```
> wts=rep(1,48)
> wts[y>160]=0
> lm2=glm(y~.,family=poisson,weights=wts,covars)
> summary(lm2)

Call:
glm(formula = y ~ ., family = poisson, data = covars, weights = wts)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.1213  -1.2433  -0.0027   0.9152   4.8080
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.11004    0.08732  35.615  < 2e-16 ***
prebake2     0.42013    0.05785   7.262 3.81e-13 ***
flux2       -0.33593    0.05781  -5.811 6.20e-09 ***
speed2       1.05490    0.06261  16.850  < 2e-16 ***
preheat2     0.03288    0.05705   0.576   0.5644
cooling2    -0.39813    0.05793  -6.872 6.32e-12 ***
agitator2    0.10100    0.05609   1.801   0.0717 .
temp2       -0.98669    0.06292 -15.680  < 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: 1137.32  on 46  degrees of freedom
Residual deviance:  251.74  on 39  degrees of freedom
AIC: 491.49

Number of Fisher Scoring iterations: 5
```

This time it looks as though neither “preheat” not “agitator” is significant. As for whether the new model fits the data, it looks as there are still some points of high influence, see for example:

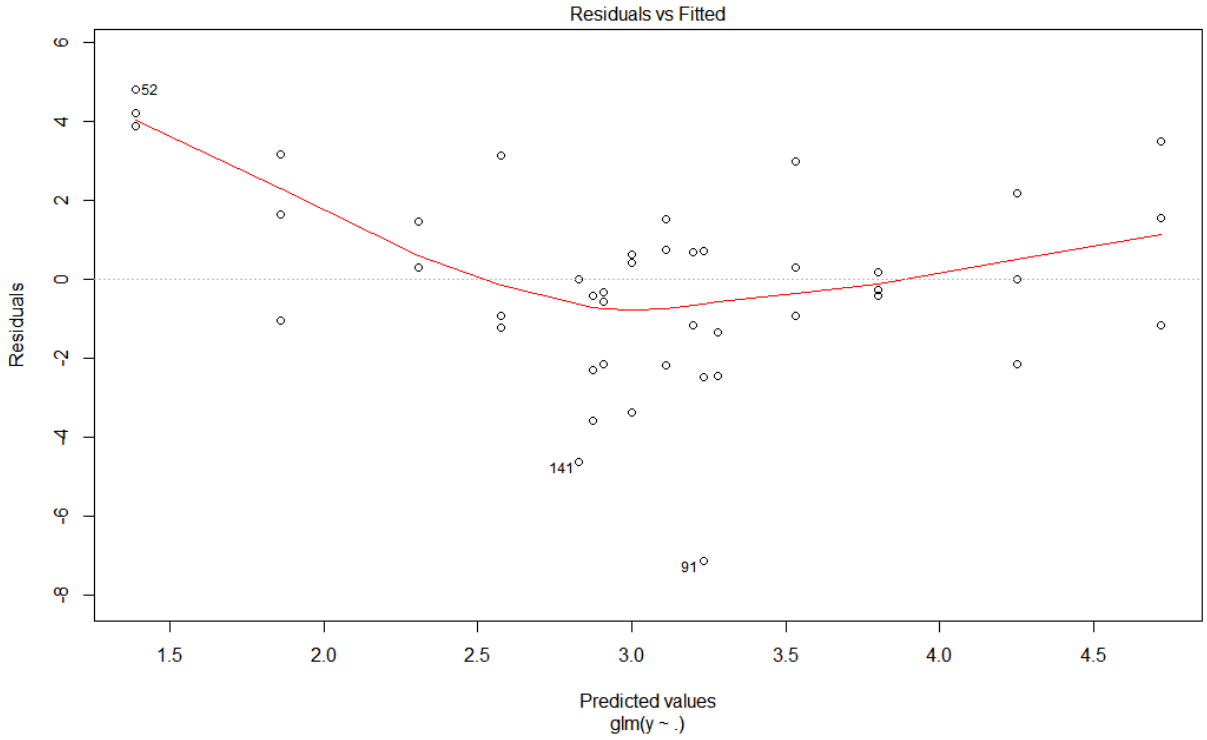


f. Now we fit a quasipoisson model, as follows:

```
> lm3=glm(y~.,family=quasipoisson,weights=wts,covars)
> summary(lm3)
```

```
Call:
glm(formula = y ~ ., family = quasipoisson, data = covars, weights = wts)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.1213  -1.2433  -0.0027   0.9152   4.8080

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.11004    0.22504  13.820  < 2e-16 ***
prebake2     0.42013    0.14909   2.818  0.00755 **
flux2       -0.33593    0.14897  -2.255  0.02982 *
speed2       1.05490    0.16134   6.538 9.28e-08 ***
preheat2     0.03288    0.14702   0.224  0.82420
cooling2    -0.39813    0.14930  -2.667  0.01109 *
agitator2    0.10100    0.14454   0.699  0.48885
temp2       -0.98669    0.16216  -6.085 3.95e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 6.641214)

    Null deviance: 1137.32  on 46  degrees of freedom
Residual deviance:  251.74  on 39  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Note the estimated dispersion parameter 6.64 which, combined with the suspect residual deviance value, implies that overdispersion really is important in this case. The leverage plot (not shown) shows much less evidence of points of high influence, but there is maybe still some evidence of lack of fit in the residuals v. fitted values plot, shown below. In particular, there is visual evidence of curvature in the plot and outliers at the bottom end.



If we "drop1" using the F test, we get

```
> drop1(lm3,test='F')
Single term deletions

Model:
y ~ prebake + flux + speed + preheat + cooling + agitator + temp
         Df Deviance F value    Pr(>F)
<none>        251.74
prebake   1   305.33  8.3029  0.006404 **
flux      1   285.87  5.2879  0.026914 *
speed     1   574.23 49.9600 1.728e-08 ***
preheat   1   252.07  0.0515  0.821701
cooling   1   299.97  7.4719  0.009373 **
agitator  1   254.99  0.5030  0.482398
temp      1   530.75 43.2249 8.267e-08 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Warning message:
In summary.glm(object, dispersion = NULL) :
  observations with zero weight not used for calculating dispersion
```

This really just confirms the same thing – preheat and agitator are not significant, the rest are.

g. Overall summary: The Poisson model in its original form did not fit the data because of the one extreme outlier and also because of the possibility of overdispersion. Dropping the outlier led to an improved fit but there still appeared to be influential datapoints as judged by the Cook statistic. Allowing for overdispersion led to a further improved model but it is still not clear it is the best model. In this analysis, preheat and agitator were not significant and a possible further refinement would be to drop these two terms from the model and refit it (however, the question did not ask us to do that). For minimizing the number of defects, it appears we do best with prebake and speed at level 1, flux, cooling and temp at level 2. This is consistent with the original plots once the outlier is deleted.

Notes added after grading:

Question 1 part e: An alternative solution is to use "type=response" instead of "type=link", in which case you will get the predicted value directly as a probability, rather than logit probability, together with the corresponding SE. You can do the same thing, with predicted value plus or minus 1.96 times the standard error to get a confidence interval, but in this case, the confidence interval is less realistic, for example, the lower confidence bound is below 0. For this and other reasons, I actually do think the above is a better solution, but after consultation with the TA, we agreed that both solutions would receive credit (and actually, there are other cases where I have given you confidence intervals that go outside the interval 0 to 1, an example being the plot on page 3 of these solutions).

If you're wondering how the SE for the response scale relates to that for the link scale, the answer is that a Taylor expansion is used to approximate the link->response transformation as locally linear, followed by using the standard errors on the link scale (the way they are calculated). This is sometimes known as the delta method. However, I'm not expecting you to know the details of that.

Question 2 part d: this question did seem to cause a lot of confusion but I don't know why! – I asked the TA to be lenient with the grading but you had to make some reasonable attempt at the question to get credit. Column 2 of the V matrix I calculated is just the column of sample variances – that shouldn't be too hard to figure (though we would also accept the SD, or indeed some other measure of variability if you said explicitly what it was). Column 1 was derived from the theoretical formula for variance of the binomial distribution, or in other words,

If Y~Binom(n,p), Var(Y)= np(1-p) or equivalently Var(Y/n)=p(1-p)/n.

The formula stated the second way is the variance formula for sample proportions. For this problem, it's a little more complicated than that because the value of n varies slightly between different samples at the same temperature, so there are actually three different theoretical variances for each temperature. My solution to that (shown above) was to average over the three theoretical variances, but I would accept any reasonable attempt to address this issue.

Question 3 part b: apparently some students interpreted "within each group of three replicates" to mean one calculation of mean and variance for all the y1's, a second for all the y2's, a third for all the y3's. Unfortunately, that's not what the question asks for.