Exploratory Data Analysis in R

Sabastian Bouma
2021

The following work is sourced from a 2021 MAST30027 Modern Applied Statistics assignment, for which 30/30 marks were awarded.

Experiment Design: *Components are attached to an electronic circuit card assembly by a wave-soldering process. The soldering process involves baking and preheating the circuit card and then passing it through a solder wave by conveyor. Defects arise during the process, and an experiment was run to try and determine the effect on the number of defects of various aspects of the process. Data: The data is taken from Condra, Lloyd, Reliability Improvement with Design of Experiment. CRC Press, 2001. Full wavesolder data has 48 observations, each of which has the number of defects and seven predictor variables. In this assignment, we will consider only the number of defects (response variable), and four predictor variables, prebake, flux, cooling, temp. The data can be found in the file assignment2 prob1 2021.txt. The dataset has 48 rows representing 48 observations. Each row has entries for:*

- *numDefects: number of defects*
- *prebake: prebake condition - a factor with levels 1 2*
- *flux: flux density - a factor with levels 1 2*
- *cooling: cooling time - a factor with levels 1 2*
- *temp: solder temperature - factor with levels 1 2*

Problem: *We want to determine which factors (prebake, flux, cooling, temp) and two- way interactions are related to the number of defects. Write a report on the analysis that should summarise the substantive conclusions and include the highlights of your analysis: for example, data visualisation, choice of model (e.g., Poisson, binomial, gamma, etc), model fitting and model selection (e.g., using AIC), diagnostic, check for overdispersion if necessary, and summary/interpretation of your final model.*

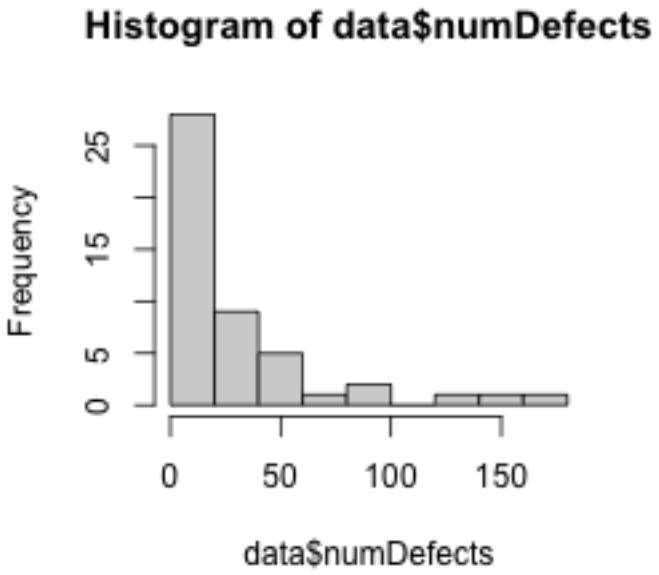
Solution:

Begin by reading in the data and converting the predictors to factos.

```
library(MASS)
data <- read.table(file ="assignment2_prob1_2021.txt", header=TRUE)
dim(data)
names(data)
data$prebake <- factor(data$prebake)
data$flux <- factor(data$flux)
data$cooling <- factor(data$cooling)
data$temp <- factor(data$temp)
```

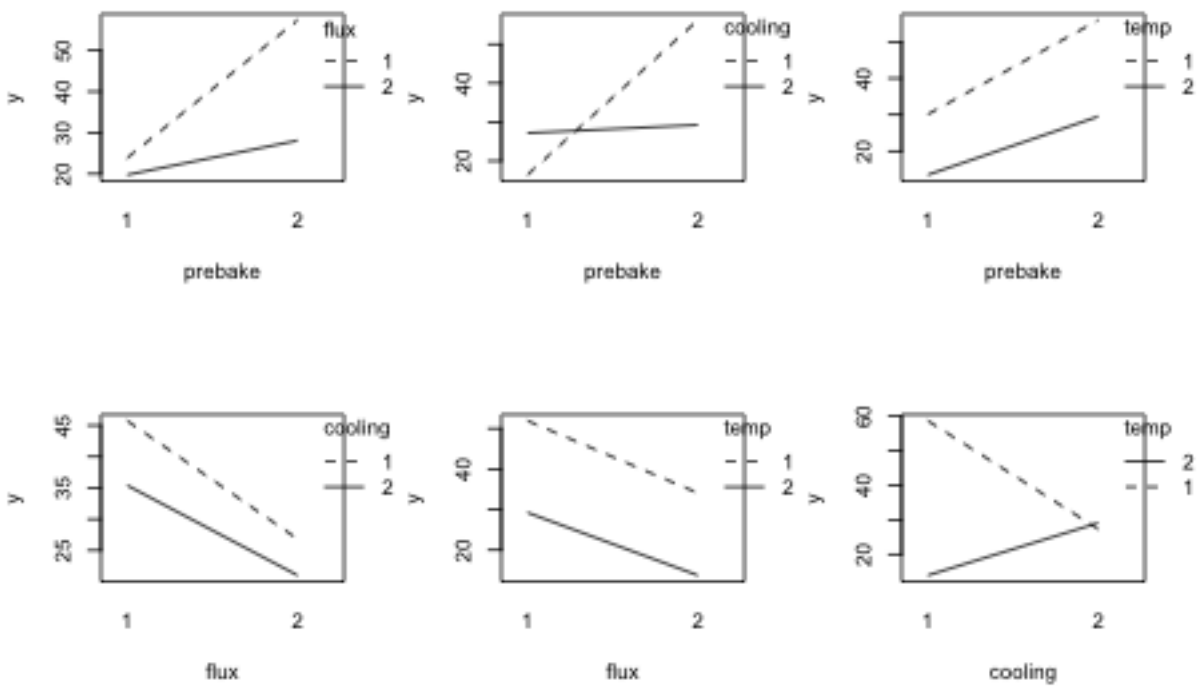
Now let's look at a histogram of the distribution of the response variable, numDefects:

```
hist(data$numDefects)
```



Based on the above histogram, poisson, quasi poisson and negative binomial are candidate distributions for the model due to the large positive skew.
The next step is to look at interaction between predictors.

```
par(mfrow=c(3,2))
with(data, {
  interaction.plot(prebake, flux, numDefects)
  interaction.plot(prebake, cooling, numDefects)
  interaction.plot(prebake, temp, numDefects)
  interaction.plot(flux, cooling, numDefects)
  interaction.plot(flux, temp, numDefects)
  interaction.plot(cooling, temp, numDefects)
})
```



The above interaction plots show clear interaction between the majority of predictors. The most significant interaction appears to be between the predictors (prebake & cooling) and (temp & cooling), with the smallest interaction between flux & cooling.
Now let's try and build a model, first focusing on a poisson distribution. The presence of interaction indicates that it should be included it in the scope of the model. To select the model, forward AIC step-wise selection will be used.

```
model_initial <- glm(numDefects ~ ., family=poisson, data=data)
model <- step(model_initial, direction='forward',scope=. ~ .^2)
summary(model)
```

```
Call:
glm(formula = numDefects ~ prebake + flux + cooling + temp +
    cooling:temp + prebake:cooling + prebake:temp + prebake:flux +
    flux:temp, family = poisson, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.7921  -2.6541  -0.2946   1.3936  13.5042

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.0948     0.5854   6.995 2.65e-12 ***
prebake           2.5860     0.2653   9.747  < 2e-16 ***
flux              0.7467     0.2390   3.124  0.00178 **
cooling          -0.2299     0.2340  -0.983  0.32580
temp             -3.9206     0.3856 -10.167  < 2e-16 ***
cooling:temp      1.7078     0.1244  13.730  < 2e-16 ***
prebake:cooling  -1.4024     0.1231 -11.389  < 2e-16 ***
prebake:temp      0.6640     0.1319   5.036 4.75e-07 ***
```

```
prebake:flux     -0.5172     0.1105  -4.682 2.84e-06 ***
flux:temp        -0.3196     0.1146  -2.788  0.00530 **
---
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:  626.99  on 38  degrees of freedom
AIC: 877.73

Number of Fisher Scoring iterations: 5
```

The forward step-wise selection process ended up on the model containing the following predictors:
prebake + flux + cooling + temp +cooling:temp + prebake:cooling + prebake:temp + prebake:flux +flux:temp.
This model has quite a high residual deviance in comparison to its degrees of freedom. This could be a product of over-dispersion. Compare estimate of dispersion parameter (ϕ) with the assumed dispersion parameter (1).

```
(phi <- sum(residuals(model, type="pearson")^2)/38)

[1] 9.599396
```

This estimate of ϕ is much larger than 1. To attempt to account for this over-dispersion, a quasi-poisson distribution model will now be attempted. We will begin with a our final poisson model and change the family to quasi-poisson.

```
model2 = glm(numDefects ~ prebake + flux + cooling + temp +
    cooling:temp + prebake:cooling + prebake:temp + prebake:flux +
    flux:temp,family=quasipoisson, data = data)
summary(model2)

Call:
glm(formula = numDefects ~ prebake + flux + cooling + temp +
    cooling:temp + prebake:cooling + prebake:temp + prebake:flux +
    flux:temp, family = quasipoisson, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.7921  -2.6541  -0.2946   1.3936  13.5042

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.0948     2.5954   1.578  0.12292
prebake           2.5860     1.1763   2.198  0.03408 *
flux              0.7467     1.0596   0.705  0.48530
cooling          -0.2299     1.0375  -0.222  0.82580
temp             -3.9206     1.7099  -2.293  0.02748 *
cooling:temp      1.7078     0.5515   3.097  0.00367 **
prebake:cooling  -1.4024     0.5460  -2.569  0.01426 *
prebake:temp      0.6640     0.5846   1.136  0.26314
prebake:flux     -0.5172     0.4898  -1.056  0.29765
flux:temp        -0.3196     0.5083  -0.629  0.53324
---
```

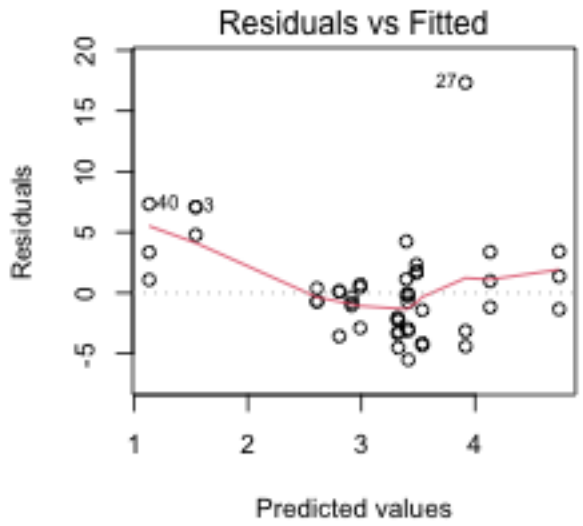
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 19.6588)

    Null deviance: 1450.52  on 47  degrees of freedom
Residual deviance:  626.99  on 38  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

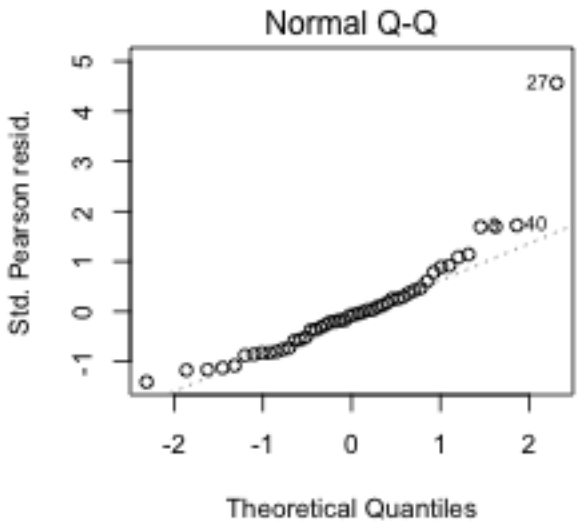
We see that the model now does not assume the dispersion parameter to be 1. Model selection for quasi-poisson is difficult as it is measured by quasi-likelihoods. As such, a negative binomial model will now be attempted.

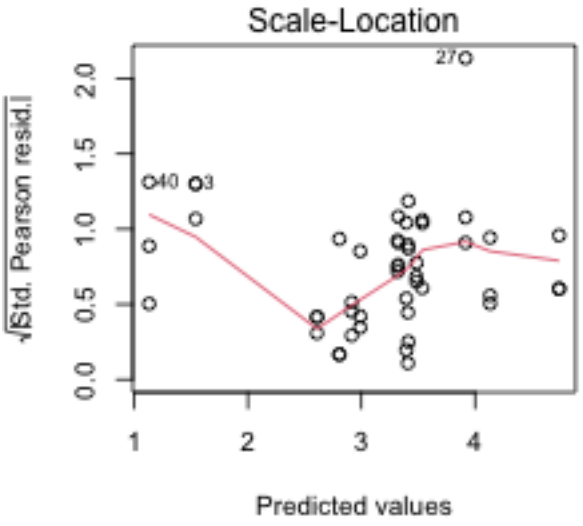
```
model3_init = glm.nb(numDefects ~ ., data=data)
model3 <- step(model3_init, direction='forward',scope=. ~ .^2)
summary(model3)

Call:
glm.nb(formula = numDefects ~ prebake + flux + cooling + temp +
    cooling:temp + prebake:cooling, data = data, init.theta = 2.087379476,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4626  -0.6122  -0.2520   0.2960   3.1105

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.6794     1.4654   3.193 0.001407 **
prebake           1.7750     0.6623   2.680 0.007364 **
flux             -0.5270     0.2087  -2.526 0.011551 *
cooling          -0.2461     0.9051  -0.272 0.785710
temp             -2.2463     0.6626  -3.390 0.000698 ***
cooling:temp      1.1177     0.4176   2.677 0.007436 **
prebake:cooling  -0.9124     0.4175  -2.186 0.028850 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(2.0874) family taken to be 1)

    Null deviance: 90.058  on 47  degrees of freedom
Residual deviance: 53.146  on 41  degrees of freedom
AIC: 418.43

Number of Fisher Scoring iterations: 1


              Theta:  2.087
          Std. Err.:  0.452

 2 x log-likelihood:  -402.427
```

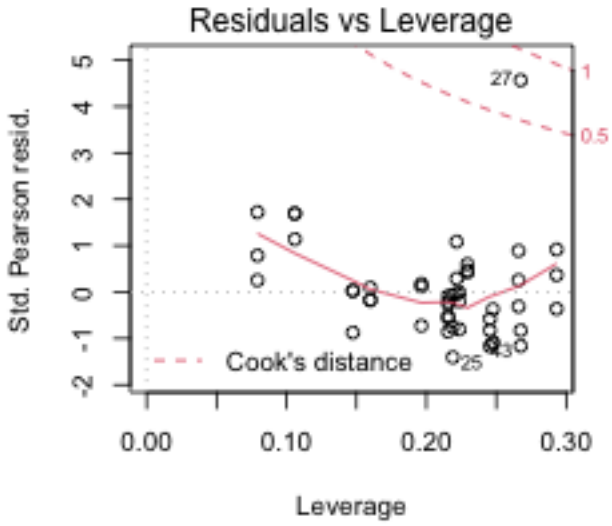
This model has a much lower dispersion parameter, as well as having smaller residual deviance. Now create diagnostic plots for both the quasipoisson and negative binomial models.

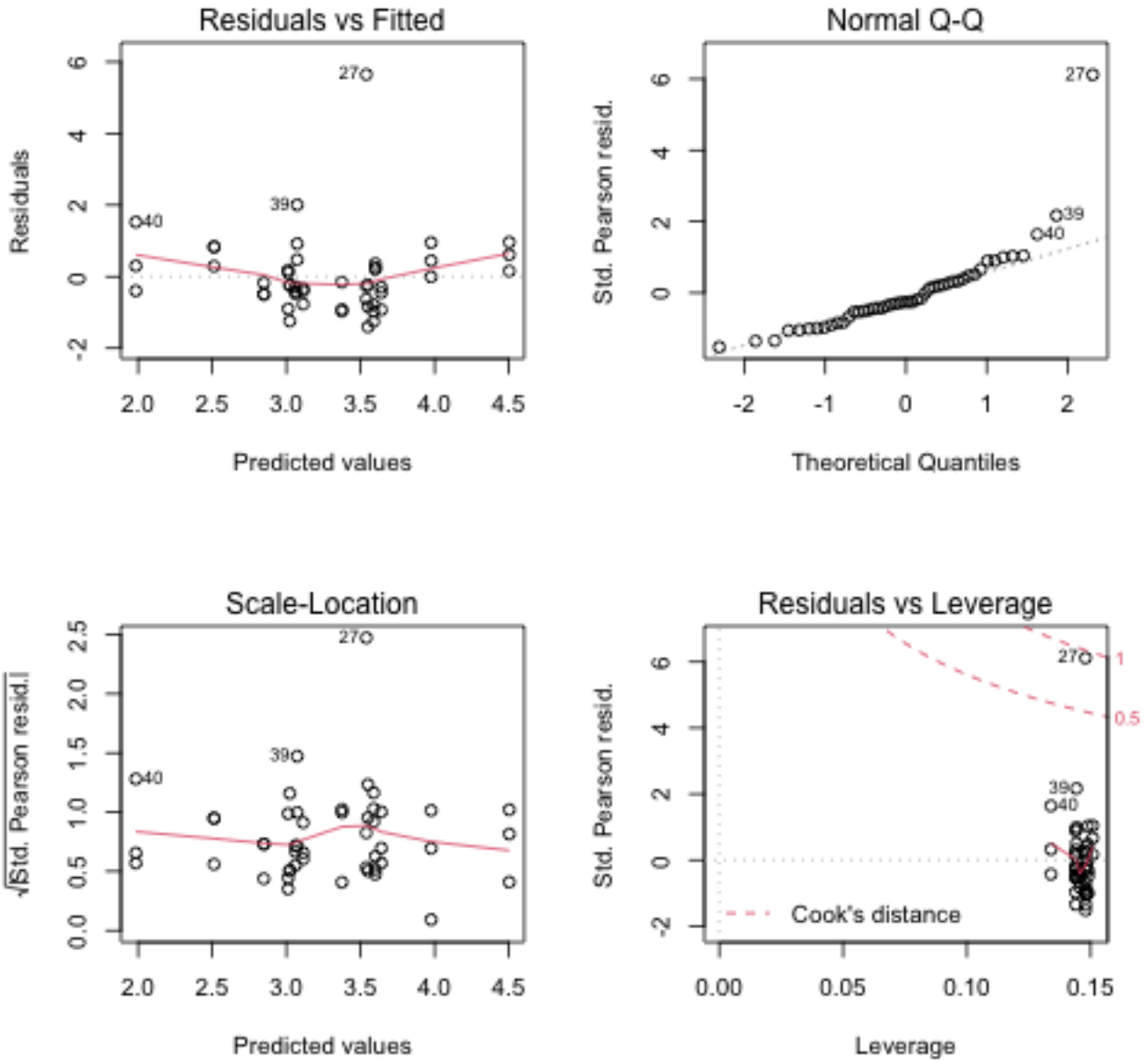
```
par(mfrow=c(2,2))
plot(model2)
par(mfrow=c(2,2))
plot(model3)
```











From the first plot we can see that neither model appears to have an underlying non-linear relationship not described in the model. Plot 2 shows that the residuals for both models are relatively normally distributed, however the 27 point could be of some concern. The 'scale-location' plot indicates a small amount of clustering for both models, however both are still reasonably heteroskedastic. The only noticeable difference between the diagnostic plots of the two models lies in the 'residuals vs leverage' plot. The quasipoisson model has a much higher variance in leverage in comparison to the negative binomial model, with nearly all of the negative binomial leverage points clustered around 0.15, a smaller average leverage in comparison to the quasipoisson.

Taking into account the lower leverage, lower dispersion parameter and smaller residual deviance, the model selected for this data will be the negative binomial model with the following coefficients:

```
round(model3$coefficients, 2)
```

```
(Intercept)         prebake            flux         cooling            temp
       4.68            1.77           -0.53           -0.25           -2.25
```

```
cooling:temp prebake:cooling
        1.12           -0.91
```