General Linear Model - Apple data

Dominika Bakalarz

1 Introduction

This report presents an analysis of "Apple" data concerned with the growth of Alicyclobacillus Acidoterrestris CRA7152 in Apple Juice. Its goal is to determine factors affecting the presence of CRA7152 growth. Following factors were measured during the experiment: pH level, Brix level (sugar content), the temperature of the juice and Nisin level (food preservative), 74 data samples were obtained.

2 Exploratory Data Analysis

There are four potentially explanatory variables available: ph, brix, temperature, nisin. All variables are continuous.

As we can see in Figure 1, CRA7152 Growth is present more often in samples with high pH level, low Nisin level, medium level temperature and low Brix level. These conclusions will be explored further and verified by suitable experiments in the next part of the report.

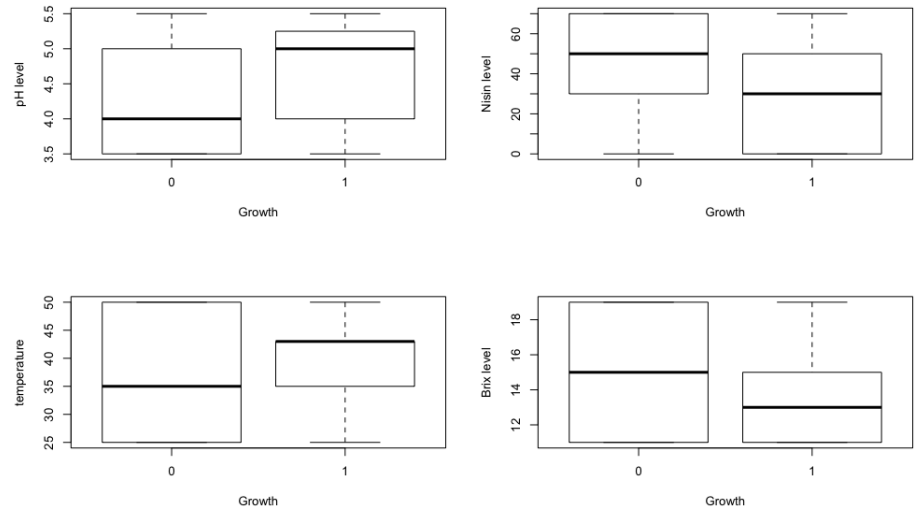

Figure 1: Boxplots presenting the dependence between CRA7152 growth (0 - absent, 1 - present) and each explanatory variable

We have checked that in 32 samples growth is present, and in the remaining 42 samples it is absent, so we have balanced data. Therefore, when plotting Figure 2, we decide to scale the data, so that the heights of red and blue bars represent percentages of all samples with growth present and absent, respectively. That is, in each of the four plots in Figure 2, all blue bars add up to 1 and all red bars add up to 1. Plots in Figure 2 confirm the conclusion drawn from the boxplots in Figure 1. Moreover, we can see that each explanatory variable has four levels, and approximately a quarter of all samples can be found at each level, so the data set is balanced with respect to all variables.

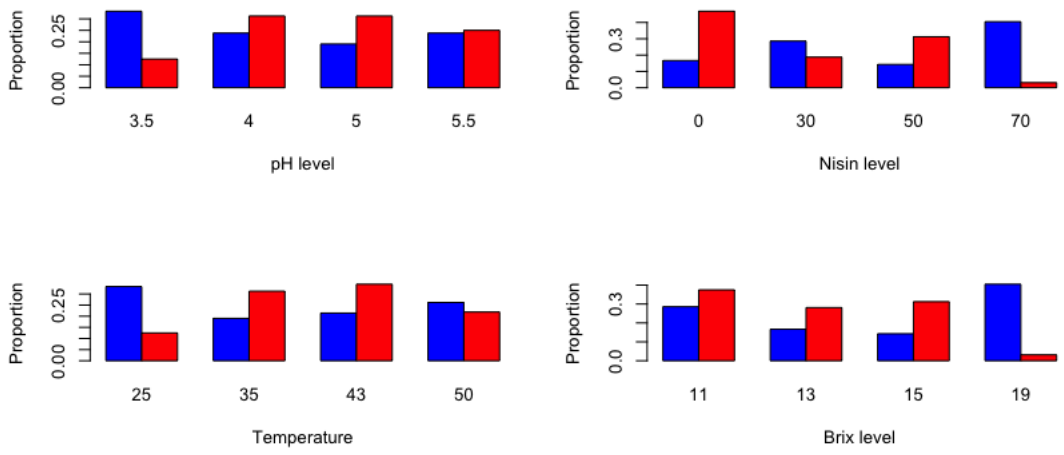

Figure 2: Bar plots presenting growth against each of the explanatory variables. Blue bars represent samples no growth while red bars represent samples with growth present. On each plot, both red and blue bars add up separately to 1.

3 Analysis of the Data

3.1 Univariate logistic regression

3.1.1 pH level

We performed univariate logistic regression for each of the explanatory variables, starting with the pH level, with the canonical link function. Now if for $i = 1, 2, ..., 74$ y_i =growth[i] and $\boldsymbol{x_i}$ = (1,ph[i]) (so we have an intercept and pH level), then we model $y_i \sim$ Bernoulli(π_i) with π_i the probability of growth present, as a function of pH level. The observation model for y_i has the following probability mass function:

$$f(y_i|\theta_i, \psi) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \exp(y_i \log(\pi_i/(1-\pi_i)) + \log(1-\pi_i))$$

so $\theta_i = \log(\pi/(1-\pi)$ is the natural parameter, $\psi = 1$, $\kappa(\theta_i = \log(1+e^{\theta_i})$ and $c = 0$. We have $\mu_i = E(Y_i) = \pi_i$ and the linear predictor is $\eta_i = \beta_1 + \beta_2 x_{i,2}$. We decide to use the logistic link (which is the canonical link in this case):

$$\log(\mu_i/(1-\mu_i)) = \eta_i$$

Thus the probability of growth being present is a logistic function of pH level.
We fit this model to the data.

```
Call:
glm(formula = growth ~ ph, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2323  -0.9839  -0.9075   1.2083   1.4739

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0787     1.3951  -1.490    0.136
ph            0.4012     0.3042   1.319    0.187

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 101.230  on 73  degrees of freedom
Residual deviance:  99.461  on 72  degrees of freedom
AIC: 103.46

Number of Fisher Scoring iterations: 4
```

Table 1: Summary of the model growth $\sim$ ph

As we can see in Table 1, the p-value for the pH level is 0.187, so it is not significant. The estimated change in the log odds when we include the ph term is $\hat{\beta_2} = 0.4012$, what which is an increase of a factor of exp(0.4012) $\backsimeq$ 1.49.

3.1.2 Brix level

Now we modify the model by writing $\boldsymbol{x_i} = (1, brix[i])$, so the probability of growth being present is now a logistic function of Brix level. We fit this new model to the data.

```
Call:
glm(formula = growth ~ brix, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3856  -1.1194  -0.6335   1.1851   1.8464

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.20101    1.25661   2.547  0.01085 *
brix        -0.24763    0.08951  -2.766  0.00567 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 101.230  on 73  degrees of freedom
Residual deviance:  92.353  on 72  degrees of freedom
AIC: 96.353

Number of Fisher Scoring iterations: 4
```

Table 2: Summary of the model growth $\sim$ brix

As we can see in Table 2, the p-value for the Brix level is 0.00567, so it is very significant (at 5% significance level). The estimated change in the log odds when we include the brix term is $\hat{\beta_2}$ = -0.24763, what which is an increase of a factor of exp(-0.24763) $\backsimeq$ 0.78.

3.1.3 Temperature

We modify the model again by writing $\boldsymbol{x_i} = (1, temp[i])$, so the probability of growth being present is now a logistic function of the temperature. We fit this new model to the data.

```
Call:
glm(formula = growth ~ temperature, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2056  -1.1182  -0.9082   1.2377   1.4730

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.41091    1.03436  -1.364    0.173
temperature  0.02954    0.02598   1.137    0.256

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 101.230  on 73  degrees of freedom
Residual deviance:  99.913  on 72  degrees of freedom
AIC: 103.91

Number of Fisher Scoring iterations: 4
```

Table 3: Summary of the model growth $\sim$ temp

As we can see in Table 3, the p-value for the temperature is 0.256, so it is not significant. The estimated change in the log odds when we include the temperature term is $\hat{\beta_2} = 0.02954$, what which is an increase of a factor of $\exp(0.02954) \backsimeq 1.02$.

3.1.4 Nisin level

We modify the model again by writing $\boldsymbol{x_i} = (1, nisin[i])$, so the probability of growth being present is now a logistic function of Nisin level. We fit this new model to the data.

```
Call:
glm(formula = growth ~ nisin, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5226  -1.0521  -0.6634   0.8677   1.8011

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.782731   0.412282   1.899  0.05763 .
nisin       -0.031210   0.009951  -3.137  0.00171 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 101.230  on 73  degrees of freedom
Residual deviance:  90.039  on 72  degrees of freedom
AIC: 94.039

Number of Fisher Scoring iterations: 4
```

Table 4: Summary of the model growth $\sim$ nisin

As we can see in Table 4, the p-value for the Nisin level is 0.00171, so it is very significant (at 5% significance level). The estimated change in the log odds when we include the nisin term is $\hat{\beta_2} =$

-0.03121, what which is an increase of a factor of exp(-0.03121) ≃ 0.97.

To summarize, we have found that Brix and Nisin level are very significant explanatory variables, while pH level and temperature are not significant at all (but their interactions with other terms might be significant and also they might become significant in more complex models, we will explore this in the next part of the report).

3.2 Model selection

We will use the GLM set up from the previous section, but we will vary the design matrix X, in order to find the best model. We start with fitting a full model with four-way interactions (and lower order terms):
(M1) growth ∼ ph : brix : temperature : nisin

```
Call:
glm(formula = growth ~ ph * nisin * temperature * brix, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8799  -0.2452   0.0000   0.0000   2.2519

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               4.419e+03  3.138e+05   0.014    0.989
ph                       -1.168e+03  8.967e+04  -0.013    0.990
nisin                    -8.618e+01  6.277e+03  -0.014    0.989
temperature              -1.532e+02  1.255e+04  -0.012    0.990
brix                     -2.843e+02  1.652e+04  -0.017    0.986
ph:nisin                  2.278e+01  1.793e+03   0.013    0.990
ph:temperature            4.119e+01  3.587e+03   0.011    0.991
nisin:temperature         2.958e+00  2.511e+02   0.012    0.991
ph:brix                   7.322e+01  4.719e+03   0.016    0.988
nisin:brix                5.540e+00  3.304e+02   0.017    0.987
temperature:brix          9.308e+00  6.607e+02   0.014    0.989
ph:nisin:temperature     -7.960e-01  7.173e+01  -0.011    0.991
ph:nisin:brix            -1.427e+00  9.439e+01  -0.015    0.988
ph:temperature:brix      -2.448e+00  1.888e+02  -0.013    0.990
nisin:temperature:brix   -1.794e-01  1.321e+01  -0.014    0.989
ph:nisin:temperature:brix  4.720e-02  3.776e+00   0.013    0.990
```

Table 5: Summary of the model (M1)

We received a warning message saying "glm.fit: fitted probabilities numerically 0 or 1 occurred", which means that the model has over fit and made some dangerous extreme assumptions. This conclusion is confirmed by the model summary shown in Table 5. All p-values are in the range 0.98-1.00 and the coefficient estimates are extremely small. We decide to ignore this model, as it is not feasible.

We try fitting a model with all three-way interactions (and lower order terms):
(M2) growth ∼ ph : brix : temperature + ph : brix : nisin + ph : nisin : temperature + nisin : brix : temperature
We again get the same warning, and the estimates are again extremely small, but the p-values look much better now (the summary is presented in Table 6). However, we probably can drop some terms from this model.

We try fitting a model with all two-way interactions (and lower order terms):
(M3) growth ∼ ph : brix + ph : temperature + ph : nisin + brix : nisin + nisin : temperature + brix : temperature

We will use the likelihood ratio test to compare the last two models. Model (M2) has deviance

```
Call:
glm(formula = growth ~ (ph + nisin + temperature + brix)^3, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.43619  -0.55015  -0.06620   0.04099   2.31878

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.278e+02  7.127e+01   1.793  0.07298 .
ph                    -3.315e+01  1.848e+01  -1.794  0.07283 .
nisin                  2.743e+00  1.084e+00   2.530  0.01142 *
temperature           -8.144e+00  3.154e+00  -2.582  0.00981 **
brix                  -6.123e+00  4.661e+00  -1.314  0.18897
ph:nisin              -6.657e-01  3.015e-01  -2.208  0.02726 *
ph:temperature         2.222e+00  8.414e-01   2.641  0.00826 **
ph:brix                1.459e+00  1.100e+00   1.327  0.18451
nisin:temperature      1.322e-02  1.698e-02   0.779  0.43612
nisin:brix            -1.999e-01  7.630e-02  -2.619  0.00881 **
temperature:brix       4.225e-01  1.772e-01   2.385  0.01710 *
ph:nisin:temperature  -6.499e-03  4.416e-03  -1.472  0.14106
ph:nisin:brix          5.009e-02  1.970e-02   2.542  0.01101 *
ph:temperature:brix   -1.128e-01  4.308e-02  -2.618  0.00885 **
nisin:temperature:brix 5.605e-04  9.899e-04   0.566  0.57125
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 101.230  on 73  degrees of freedom
Residual deviance:  39.525  on 59  degrees of freedom
AIC: 69.525

Number of Fisher Scoring iterations: 9
```

Table 6: Summary of the model (M2)

39.525 and 59 degrees of freedom (Table 6), while model (M3) has deviance 66.05 and 63 degrees of freedom (Table 7). So we test 66.05 - 39.525 against chi-squared distribution with 4 degrees of freedom, the p-value is $2.479506 * 10^{-5}$, so we have to reject model (M2).

By inspecting the summary of model (M2) we notice that none of the terms nisin:temp, nisin:temp:ph, nisin:temp:brix is significant, so we decide to remove nisin:temp interaction from the model. That is, we are fitting the following model (including lower order terms):
(M4) growth $\sim$ ph : brix : temperature + ph : brix : nisin

We will use the likelihood ratio test to compare (M2) and (M4). Model (M4) has deviance 45.517 and 62 degrees of freedom (Table 8), so we test the LRT statistic 45.517-39.525 against chi-squared distribution with 3 degrees of freedom, the p-value is 0.11, so it is not significant, therefore there is no evidence against model (M4).

By inspecting the summary of model (M4) presented in Table 8 we notice that the only terms that are not significant (at 5% significance level) are intercept, brix, ph and ph:brix, but ph:brix:temp and ph:brix:nisin are significant, therefore we cannot simplify this model further, and we choose (M4) as our final model for now.

3.3 Goodness of fit

The goodness of fit test is not applicable to Bernoulli models.

We proceed to outlier analysis. As we can see on Figure 3, point 49 has a deviance residual greater than 2 and it has high leverage (as it is far away from other points), and it also has a large influence, as its Cook's distance exceeds the threshold $\frac{8}{n-2p} = 0.16$. Therefore it is an outlier and we decide to

```
Call:
glm(formula = growth ~ (ph + nisin + temperature + brix)^2, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5106  -0.6798  -0.2729   0.5922   2.0208

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -18.936371  11.889229  -1.593   0.1112
ph                  3.419990   2.500139   1.368   0.1713
nisin               0.059490   0.092571   0.643   0.5205
temperature         0.112959   0.230763   0.490   0.6245
brix                1.118380   0.770519   1.451   0.1467
ph:nisin           -0.052429   0.024277  -2.160   0.0308 *
ph:temperature      0.078262   0.057072   1.371   0.1703
ph:brix            -0.247863   0.149740  -1.655   0.0979 .
nisin:temperature  -0.001741   0.001667  -1.045   0.2962
nisin:brix          0.013902   0.007033   1.977   0.0481 *
temperature:brix   -0.023363   0.014733  -1.586   0.1128
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 101.23  on 73  degrees of freedom
Residual deviance:  66.05  on 63  degrees of freedom
AIC: 88.05

Number of Fisher Scoring iterations: 6
```

Table 7: Summary of the model (M3)

```
Call:
glm(formula = growth ~ (ph + nisin + temperature + brix)^2, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5106  -0.6798  -0.2729   0.5922   2.0208

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -18.936371  11.889229  -1.593   0.1112
ph                  3.419990   2.500139   1.368   0.1713
nisin               0.059490   0.092571   0.643   0.5205
temperature         0.112959   0.230763   0.490   0.6245
brix                1.118380   0.770519   1.451   0.1467
ph:nisin           -0.052429   0.024277  -2.160   0.0308 *
ph:temperature      0.078262   0.057072   1.371   0.1703
ph:brix            -0.247863   0.149740  -1.655   0.0979 .
nisin:temperature  -0.001741   0.001667  -1.045   0.2962
nisin:brix          0.013902   0.007033   1.977   0.0481 *
temperature:brix   -0.023363   0.014733  -1.586   0.1128
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 101.23  on 73  degrees of freedom
Residual deviance:  66.05  on 63  degrees of freedom
AIC: 88.05

Number of Fisher Scoring iterations: 6
```

Table 8: Summary of the model (M4)

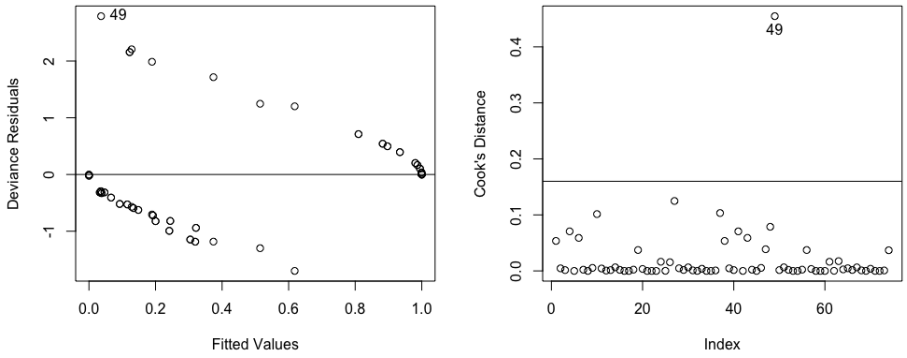

Figure 3: On the left, we can see deviance residuals plotted against fitted values, and on the right, we can see Cook's distances for all data points.

remove it.

We fit again model (M4) to a new data set (with point 49 being removed), and we discover that the model is not valid anymore. We receive the warning message signalizing overfitting and all estimates and p-values are extremely small. We decide to remove all three-way interactions, that is we are fitting again model M3 but to a new data set. Estimates look reasonable now, but none of p-values is significant, so this model is also not valid. We decide to remove all interaction terms, that is, we are fitting a model:
(M5) growth $\sim$ ph + temperature + nisin + brix

```
Call:
glm(formula = growth ~ (ph + nisin + brix + temperature), family = binomial,
    data = apples2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6440  -0.6823  -0.3382   0.8403   2.0040

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.77413    2.24395  -0.345  0.73011
ph           0.97133    0.40127   2.421  0.01549 *
nisin       -0.03812    0.01234  -3.088  0.00201 **
brix        -0.37196    0.12381  -3.004  0.00266 **
temperature  0.06486    0.03380   1.919  0.05498 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 99.536  on 72  degrees of freedom
Residual deviance: 68.532  on 68  degrees of freedom
AIC: 78.532
```

Table 9: Summary of the model (M5) - final model

This model looks plausible, as we can see in Table 9, all terms except temperature are significant at 5% significance level, and the p-value for temperature is 0.05498, so it is very close to 0.05. We will test if we can drop the temperature from this model. That is, we will compare model (M5) with the model:
(M6) growth $\sim$ ph + nisin + brix

Model (M5) has residual deviance 68.532 and 68 degrees of freedom, while model (M6) has residual deviance 72.539 and 69 degrees of freedom. Therefore to test if we can drop temperature we compare the LRT statistic 72.539 - 68.532 against chi-squared distribution with 1 degree of freedom. The p-value is 0.045, so it is significant (at 5% significance level) therefore we will not drop the temperature. Thus, we will use (M5) as our final model.

3.4 Outlier analysis

In Figure 4 we can see four plots that will help us to assess the goodness of fit for our final model. The top left plot presents leverage of each point divided by p/n, as we can see no points are above the threshold 2, so there no points with unusually high leverage, it is a good sign. The top right plot presents deviance residuals, which behave in a way expected for a Bernoulli distribution. The bottom left plot presents the Cook's distance, no point exceeds the $8/(n-2p)$ threshold, so there are no points with unusually high influence, another good sign. And the bottom right plot presents working residuals, everything looks fine. For Bernoulli distribution we cannot expect residuals to be Normal, therefore we have not plotted any qq-plots.

To summarize, there is no evidence for misfit. The final model fits the data (without point 49) very well.

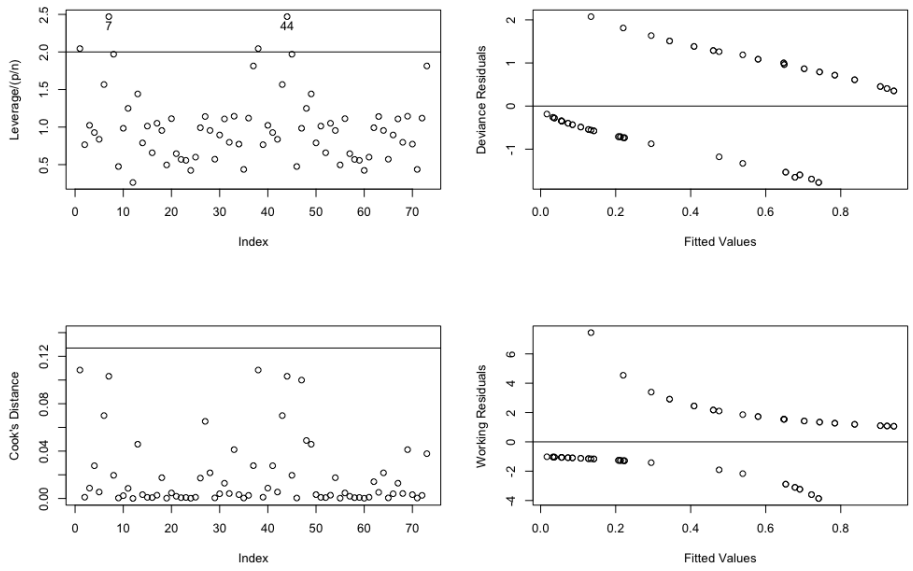

Figure 4: Top left plot presents leverage of each point divided by p/n with threshold 2 marked, top right plot presents deviance residuals, bottom left plot presents Cook's distance of each plot and the threshold $8/(n-2p)$, bottom left plot presents working residuals.

3.5 The GLM setup and interpreation

The GLM setup for the final model: Similarly to the univariate case, let
$y_i = growth[i]$
$\boldsymbol{x_i} = (x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, x_{i,5}) =$ (1, ph[i], brix[i], temperature[i], nisin[i]) for $i = 1, 2, ..., 74$

Then we model $y_i \sim \text{Bernoulli}(\pi_i)$ with π_i the probability of growth present, as a function of explanatory variables. The observation model for y_i has following probability mass function:

$$f(y_i|\theta_i, \psi) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \exp(y_i \log(\pi_i/(1-\pi_i)) + \log(1-\pi_i))$$

so $\theta_i = \log(\pi/(1-\pi)$ is the natural parameter, $\psi = 1$, $\kappa(\theta_i = \log(1+e^{\theta_i})$ and $c = 0$. We have $\mu_i = E(Y_i) = \pi_i$, and the linear predictor is $\eta_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5}$. We use the logistic link (which is the canonical link in this case):

$$\log(\mu_i/(1-\mu_i)) = \eta_i$$

3.6 Interpretation of the fitted model

The estimated change in the log odds when we include the ph term is $\hat{\beta_2} = 0.97133$, what which is an increase of a factor of exp(0.97133) $\backsimeq$ 2.64.

The estimated change in the log odds when we include the nisin term is $\hat{\beta_3}$ = -0.03812, what which is an increase of a factor of exp(-0.03812) $\backsimeq$ 0.96.

The estimated change in the log odds when we include the brix term is $\hat{\beta_4}$ = -0.37196, what which is an increase of a factor of exp(-0.37196) $\backsimeq$ 0.69.

The estimated change in the log odds when we include the temperature term is $\hat{\beta_5} = 0.06486$, what which is an increase of a factor of exp(0.06486) $\backsimeq$ 1.07.

4 Summary

We have identified point 49 as an outlier and decided to remove it. The best model for this data is growth $\sim$ ph + temperature + nisin + brix.

5 Attachments

R code is attached on the next page.