ACSC/STAT 3740, Predictive Analytics

WINTER 2025

Toby Kenney

Homework Sheet 5

Model Solutions

Standard Questions

1. *A scientist is studying crystal formation, and has written the following conclusion to her paper.*

The aim of this study was to determine the factors influencing size and shape of the formation of crystals of sodium carbonate. Previous research [1] has highlighted temperature, water salinity, humidity and air pressure as key factors. Theoretical considerations [2] have also suggested that potassium concentrations and pH should also be extremely influential.

We collected data from 186 natural crystal formations across 38 sites. For each crystal, we recorded the total mass, and a classification of the shape based on 4 variables: whether the shape was regular or irregular; whether the edges were aligned to cubic or octahedral lattices; the number of sharp edges; and the length of the longest edge. We also recorded a number of variables describing the chemical composition of the crystal, and the various impurities found. For each site, we recorded various conditions of that site: humidity, salinity, temperature, air pressure, and pH.

A preliminary exploration of the data identified 1 outlying site, where the recorded humidity was only 36%, whereas all other sites had humidities in excess of 68%. This site, which contained 4 crystals was removed from the data. One other crystal had a recorded total impurity of over 18%, but was in a site with salinity only 17%. These two measurements are inconsistent, both with the theory [3], and with the other crystals observed, both in that site, and in other sites. We therefore also removed this crystal as an outlier. We also looked at the relation between pairs of predictors and at the relation of each predictor with the size of the crystals. We found that there was a strong linear relation between temperature of the site and potassium concentration of the crystal, and a weak nonlinear relation between pH of the site and purity of the crystal. Crystal size has a heavy-tailed distribution, and given that it is necessarily positive, a log-transformation was strongly suggested. The strongest predictors of crystal size are humidity, which appears to have a linear relation with log crystal size, and temperature, which seems to have a non-linear relation.

We used three modelling frameworks to predict the crystal size. The first was a simple linear model to predict the logarithm of crystal size from the other predictors, with a quadratic term in temperature, and an interaction term between air pressure and salinity. The second is a generalised additive model, including linear terms in pH, temperature and

salinity, and non-linear terms in humidity, potassium concentration and total impurity. This model used a gamma distribution for crystal size with a log-link function. The final model was a mixed effect generalised linear model with a gamma distribution for crystal size. In this model, the site is used as an additional predictor, but instead of estimating the coefficient, the coefficient is treated as a random variable with mean 0 and unknown variance, and the likelihood is obtained by taking the expectation over the distribution of these coefficients. The effect of this is that the random coefficients incorporate all the quantities not measured in each site. Because the aim of the project is to improve our scientific understanding of the factors influencing crystal formation, we needed to choose an interpretable model. We therefore did not consider black-box methods such as random forest. However, we fitted random forest to the data for comparison.

We used 10-fold cross-validation to assess the predictive ability of each model. The cross-validated mean squared errors for estimating log crystal size were 0.467 for the linear model, 0.448 for the generalised additive model and 0.469 for the mixed effect model. These compare with 0.441 for random forest. Based on the similar predictive performances between the fitted models and random forest, the fitted models seem appropriate. We also looked at the deviance residuals for the three models, and found that they showed no relation, either in mean or in variance with the fitted values, and that they approximately follow a normal distribution, indicating that the modelling assumptions are reasonable. Comparing cross-validated log-likelihood for the fitted models, we get 944.4 for the linear model, 965.6 for the GAM and 961.0 for the mixed effect model. This suggests that the GAM is the best fit for the data.

The GAM model fits a significant positive coefficient to pH, a significant negative coefficient to temperature, and a significant non-linear effect for total impurity. These are consistent with previous research. The estimated variance in the mixed effect model is 0.28, indicating that unmeasured factors probably account for approximately 19.2% of the variation in crystal size.

write an abstract for this paper with a word limit of 150 words.

The study examined 186 natural sodium carbonate crystal formations across 38 sites, with the objective of determining the key factors influencing size and shape. We found that the best interpretable model for predicting crystal size was a generalised additive model, with a gamma distribution for crystal size. This model had similar cross-validated MSE to random forest, indicating a good predictive ability. This model indicates that higher pH and lower temperature are associated with larger crystals. In both cases, the predictors have a linear relation with log crystal size. There is also a significant non-linear relation between total impurity and log crystal size. These results are consistent with previous studies.

2. *The following quotes come from a report on the effect of page layout on customer purchase decisions. Where in the report should they be placed? Justify your answers.*

(i)

Online sales make up a huge market, with current estimates of the global market size ranging from 6 trillion to 8 trillion US dollars [1]. It is also a very competitive market with more and more companies competing. It is therefore crucial to make the most of all opportunities. In this report, I examine how good website design can enable companies to do that.

This is clearly from the "Introduction". It is providing the background for the business problem. This same background would also be in the "Executive summary", and it is possible that this quote could be there. However, it may be more brief in the "Executive Summary". For example, it is unusual to include references in the executive summary.

(ii)

Previous research [1] has suggested that having too many frames on the page might reduce purchases by as much as 8%. However, other research [2] suggests this might be a surrogate for small text size. In order to better determine whether this is the case, we included a larger range of sites in our data, with less correlation between these two predictors.

This is clearly from the introduction. It discusses previous research on the topic, and the current state of the field. Similar statements to the first two might be in the conclusion section, but I would expect them to be followed by some discussion of how the results relate to the questions raised by previous research.

(iii)

Because the data from different online retailers is formatted differently, in some cases, we were unable to get reliable recordings of the variables "top.aspect.ratio" and "colour.contrast" for all retailers. To assess the extent to which this unreliability could influence our conclusions, we used the method of [1] to generate other different values for these data, to see how much our other conclusions would change. The results are in Table 1.

This is probably from an appendix. It is checking some technical details about the analysis. This would probably not be included in the main document. If it were included in the main document, it would be in the "Data Analysis" or "Results" section.

(iv)

When we plot title font size against frame border, we see that there are four outliers with large title font size but small borders. Given the high correlation between title font size and frame border in the rest of the data, I decided to remove these outliers for the analysis.

This is clearly from the "Data Exploration" section. It is identifying important data features that need to be dealt with before the analysis.

(v)

Because of the lower cross-validated MSE, we choose the GAM with no interaction terms as our final model. Based on this model, we find that font size of title has a non-linear effect, while number of frames and frame separation have linear effects.

This is probably from the "Results" or "Data Analysis" section. It provides a clear statement of the model chosen and discusses the results from fitting that model. It might also be in the "Executive Summary", but the discussion of cross-validated MSE might be too technical for the "Executive Summary", which may be read by people with limited data analysis experience.

(vi)

We found that as suggested by [1], the font size of the title has a non-linear relation with logistic transformed purchase probability.

This would be in the conclusion. It is comparing the results of the data analysis with results from the literature. It could be in the results section, but it would be more usual to compare results with existing literature in the conclusions section.

(vii)

Based on our model, we would improve sales by 5% if we reduced the number of frames on screen to 4; increased title font size by 1 point; and increased separation between frames by 2 points.

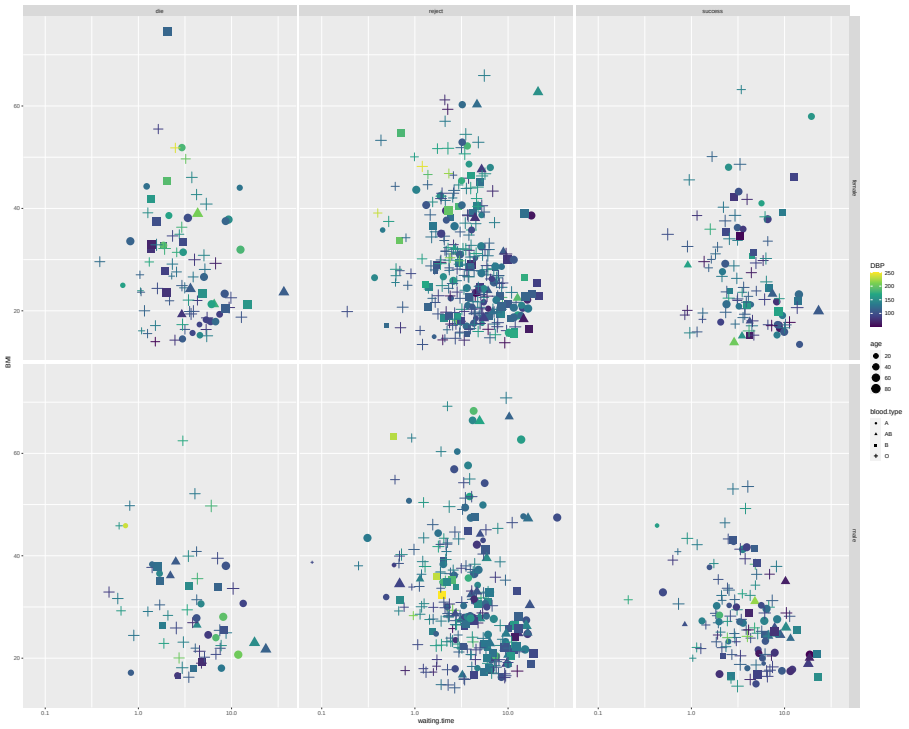
This could be in the "Executive Summary" — it is a very clear actionable statement of the results of the modelling that could be easily understood by readers with no data analysis background. It might also appear in the "Conclusions" section.

(viii)

We see that even after log-transformation, there is a nonlinear relation between colour contrast and average amount spent. Based on this, we will add a quadratic term in colour contrast as an additional predictor to our first model.

This is clearly from the "Data Exploration" section. It is discussing the initial observations from the data, and how they could influence the analysis.

3. *A doctor has analysed the data in file* `HW5Q3.txt`*, and produced the following plot of the results. The data are on predicting the outcome of a kidney transplant operation.*



Write a paragraph to describe the figure and the conclusions drawn from it.

Figure 1 shows the outcomes of the kidney transplant operations. We see that only a minority of operations are successful, with the majority being rejected, and a number of patients dying. We see that the patients for which the operation is a success tend not to have very high BMI, particularly for male patients. Blood type O patients seem to have the highest

chance of success. Higher blood pressure also seems to be associated with worse outcomes. Waiting time, age and patient sex do not appear to significantly effect outcome. Even blood-type O patients with low BMI have a fairly low chance of success.

4. *A pensions company is modelling improvements in mortality. It collects the following data on its policyholders:*

Variable	*meaning*
init.age	*The age of the policyholder at the time of plan initiation.*
init.year	*The year of plan initiation*
death.age	*The age of the policyholder at death (0 if policyholder is still alive)*
death.year	*The year of the policyholder's death*
sex	*The sex of the policyholder*
race	*The race of the policyholder*
income	*The policyholder's income (adjusted for inflation) at time of initiation.*
smoking	*Whether and how much the policyholder smokes at time of initiation.*

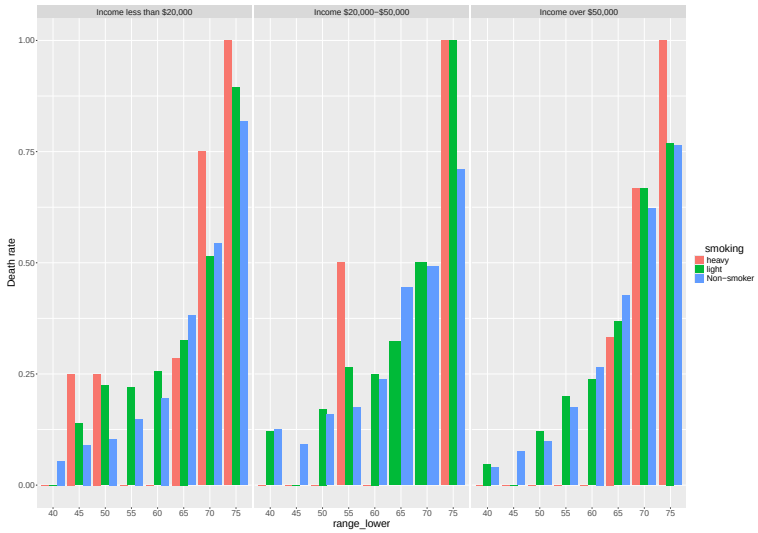
The data are in file `HW5Q4.txt`.

After analysing the data, they have made the following conclusions:

- *Smoking increases an individual's risk of dying at all ages.*
- *Having a higher income increases an individual's life expectancy.*

Make a plot that demonstrates these conclusions for presentation in their report.

One approach is a bar-chart, using a facet grid by income group. This clearly shows higher death rates for lower income groups, but the small number of heavy smokers in each group means the bars have large variance.



We could add error bars, or add another bar to show sample size. It is made with the following code:

```
HW5Q4<-read.table("HW5Q4.txt")
library(dplyr)
library(tidyr)
library(ggplot2)

ggplot(HW5Q4%>%mutate(final.age=ifelse(is.na(death.age),init.age+2022-init.year,death.age
                      range_40_45=init.age<40&final.age>45,
                      range_45_50=init.age<45&final.age>50,
                      range_50_55=init.age<50&final.age>55,
                      range_55_60=init.age<55&final.age>60,
                      range_60_65=init.age<60&final.age>65,
                      range_65_70=init.age<65&final.age>70,
                      range_70_75=init.age<70&final.age>75,
                      range_75_80=init.age<75&final.age>80,
                      income_range=cut(income,breaks=c(0,
                                                       20000,
                                                       50000,
                                                       100000000000000000000),
                                       labels=c("Income less than $20,000",
                                                "Income $20,000-$50,000",
                                                "Income over $50,000"))
                      )%>%
       pivot_longer(10:17,names_to="range",values_to="InRange")%>%
       filter(InRange)%>%
       separate(range,sep="_",into=c("x","range_lower","range_upper"))%>%
       group_by(range_lower,smoking,income_range)%>%
       summarise(no.deaths=sum(!is.na(death.age)&death.age<range_upper),
                 no.inrange=n()),
       mapping=aes(x=range_lower,y=no.deaths/no.inrange,fill=smoking))+
    geom_col(position="dodge")+
    largertextsize+
    facet_wrap(income_range~.)+
    scale_y_continuous(name="Death rate")
```

Another approach is to use the `survminer` package to make a survival plot.

It is made with the following code:

```
arrange_ggsurvplots(
    list(
        ggsurvplot(
            surv_fit(Surv(death.age)~smoking,
                     data=HW5Q4)),
        ggsurvplot(
            surv_fit(Surv(death.age)~income.range,
                     data=HW5Q4%>%
                         mutate(
                             income.range=cut(income,
                                              breaks=c(0,20000,50000,1e10),
                                              labels=c("under $20,000",
                                                       "$20,000-$50,000",
                                                       "over $50,000")))
                     )
        )
    ),nrow=2)
```