M5MS03 - Applied Statistics; Coursework 2-Autumn 2018

Question A: The sodium data-set

An extract of the dataset, `Sodium`, is shown below.

```
  sodium brand bottle
1   24.4     1      1
2   22.6     1      2
3   23.8     1      3
4   22.0     1      4
5   24.5     1      5
6   22.3     1      6
```

Table 1: Head of the `Sodium` dataset.

The dataset consisted of two predictors, `brand` and `bottle`, and one response, `sodium`. It contained $n = 48$ observations. The experiment design was complete and balanced: there were equal numbers of specimens for each level of `bottle`, and each level of `brand` appeared the same number of times across the levels of `bottle`.

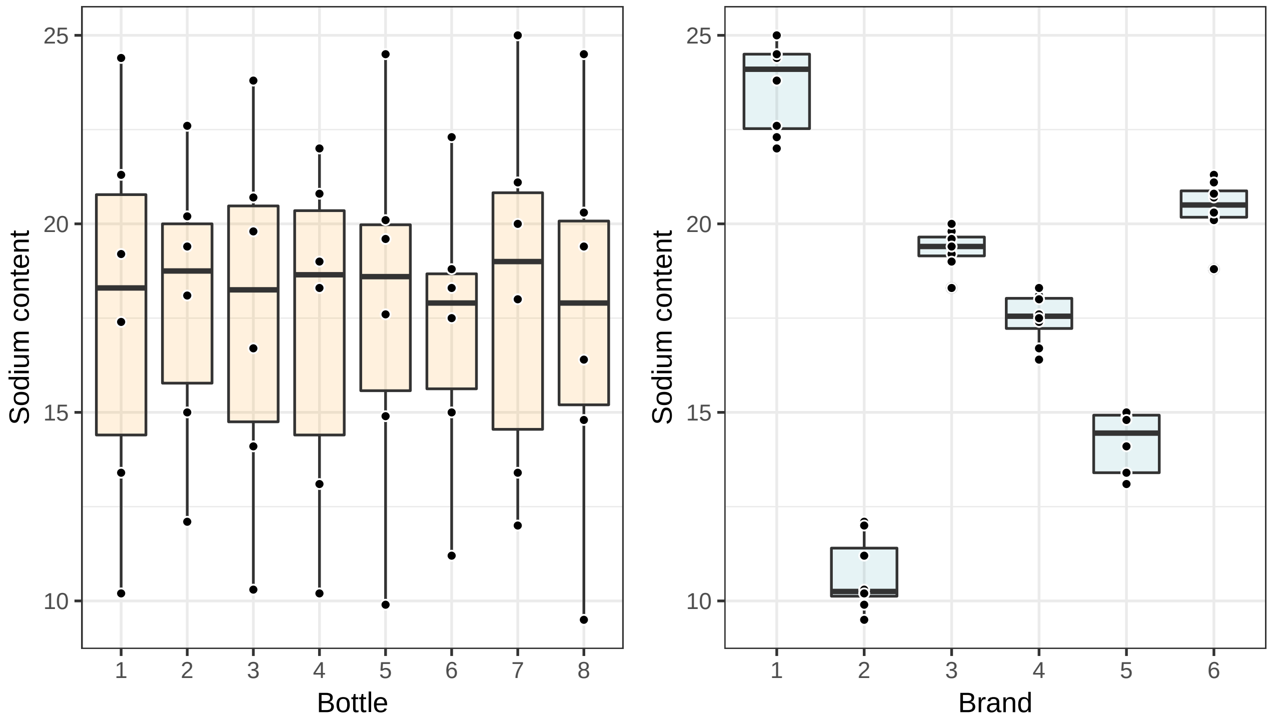

Figure 1: Boxplots for `sodium` values as a function of `bottle` level (left) and`brand` level (right).

Question A1

Question A1 requested that we fit a two-factor fixed-effects model. This model, which will be referred to as Model Ω, can be expressed as

$$y_{ijk} = \alpha + \gamma_i + \delta_j + (\gamma_i \delta_j) + \epsilon_{ijk} \tag{Model Ω}$$

α corresponds to the mean `sodium` value across all observations, a baseline against which to assess the sodium content at the various bottle and brand levels. γ_i is the mean deviation in sodium level w.r.t. this baseline for units of `bottle` $= i$ and `brand`$= 1$. Similarly, δ_j is the mean difference in sodium level for units of `bottle`$= 1$ and `brand`$= j$[1]. ϵ_{ijk} is the error associated with response k of bottle-brand

[1] The semantics of these coefficients apply specifically to the MLE/OLS values.

combination ij. $(\gamma_i\delta_j)$ is shorthand for an interaction term coefficient. Because this is a fixed-effects model, α, γ_i, and δ_i are fixed values while ϵ_{ijk} is a random variable that has distribution $\mathcal{N}(0, \sigma^2)$.

The model of Equation 1 required that $1 + (8 - 1) + (6 - 1) + (8 - 1) \times (6 - 1) = 48$ parameters were estimated, which was equal to the number of observations, n. It therefore had enough degrees of freedom to produce a zero-RSS fit, i.e. the 48 instances of Equation 1 that each correspond to an observation form a system of linear equations with a unique solution[2]. Fitting the model according to the OLS criterion,

```
> Sodium.lm0 <- lm(sodium ~ brand*bottle, data=Sodium)
```

demonstrated that R's output corroborates the conclusion that the model would be over-fit:

```
> sum(Sodium.lm0$residuals^2)
[1] 0
```

Model Ω was appropriate for neither descriptive nor predictive inference. Interpreted literally, the parameter estimates associated with it suggest that all variation in the sodium content of the food products can be attributed to `brand` and `bottle`, which is bizarre and patently untrue. Furthermore, the values for the interaction terms are misleading since they were estimated using only individual observations. We are also unable to provide confidence intervals for our parameter estimates because it is not possible to estimate the error's variance (i.e. the model suggests that $\sigma^2 = 0$).

Question A2

To resolve the issues with Model Ω, the model needed to contain fewer parameters. A straightforward way to do this was to omit the interaction terms and consider the set of nested models

$$y_{ijk} = \alpha + \epsilon_{ijk} \qquad \text{(Model } \alpha\text{)}$$
$$y_{ijk} = \alpha + \gamma_i + \epsilon_{ijk} \qquad \text{(Model } \alpha\gamma\text{)}$$
$$y_{ijk} = \alpha + \delta_j + \epsilon_{ijk} \qquad \text{(Model } \alpha\delta\text{)}$$
$$y_{ijk} = \alpha + \gamma_i + \delta_j + \epsilon_{ijk} \qquad \text{(Model } \alpha\gamma\delta\text{)}$$

Table 1 contains MLE coefficients for each of these models. Table 2 supplements these estimates with p-values, providing the probability that the values of these estimates would be observed under the null hypothesis that each coefficient is equal to zero. The p-values strongly suggest that

- the mean sodium content of all observations, α, is non-zero,
- that choice of `brand` level affects sodium content, i.e. $\delta_i \neq 0$ for all i,
- that it is not possible to conclude that `bottle` has a non-zero effect on sodium content from this data. It cannot be reasonably concluded that $\gamma_j \neq 0$ for any j.

The model can also be evaluated using an ANOVA table, as is provided in Table 2. This values in this table support the conclusions drawn above.

Based on the results of this analysis, Model $\alpha\delta$ seems preferable. The mean response is non-zero and `brand` is a predictive factor at all levels. By contrast, if `bottle` level affects `sodium` content, the effect is too small to be identified from this data set. It is therefore difficult to justify including it in the

[2] This is in contrast to the case where $n > p$ such that they create an overdetermined system of equations.

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
brand	5	854.5292	170.9058333	240.060366	4.822069e-26
bottle	7	5.1525	0.7360714	1.033912	4.255925e-01
Residuals	35	24.9175	0.7119286	NA	NA

Table 2: Two-way ANOVA for Model $\alpha\gamma\delta$.

model at the moment. Naively basing model choice on residual sum-of-squares (RSS) would suggest that Model $\alpha\gamma\delta$ is better, since its RSS is smaller than that of Model $\alpha\delta$[3]. This difference is purely as a result of the fact that the Model $\alpha\gamma\delta$ has more degrees of freedom and is not representative of the model's relative performance on prediction.

In terms of diagnostics, confidence intervals for the Model $\alpha\delta$ coefficient estimates are provided in Figure 2. Quantile-quantile plots are presented in Figure 2

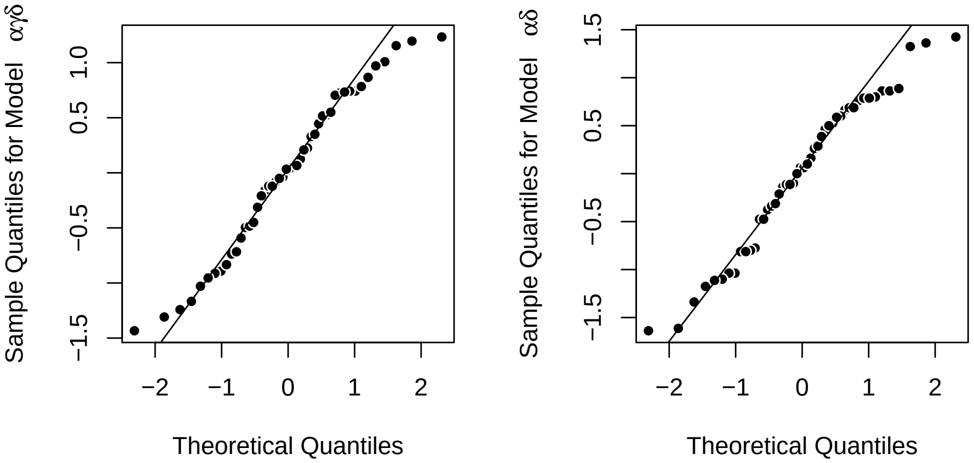

Figure 2: Quantile-quantile plots for Model $\alpha\gamma\delta$ (left) and Model $\alpha\delta$ (right).

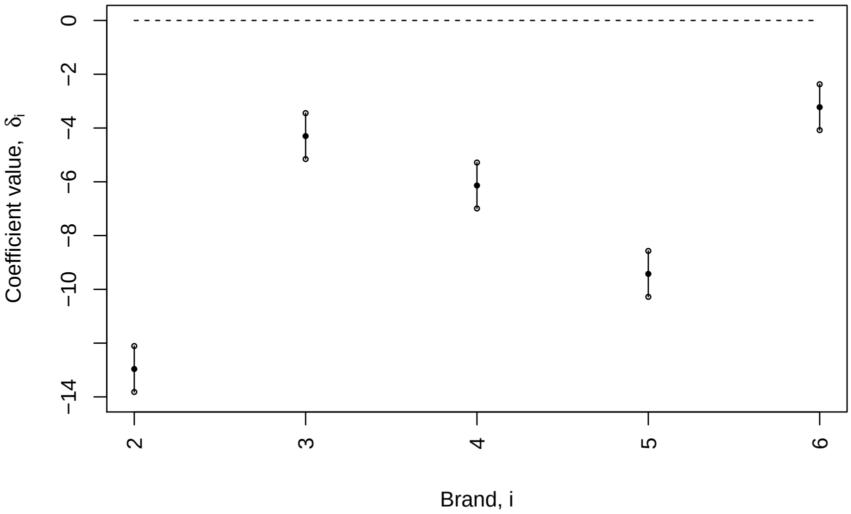

Figure 3: MLE coefficient values, δ_i, and 95% confidence intervals.

[3] A similar conclusion might be drawn from the coefficient of determination, R^2, for either model.

[5] This foonote warns against the hazards of multiple testing when evaluating significance.

	Model.a	Model.ad	Model.ag	Model.agd
(Intercept)	17.6	23.6	17.6	23.7
brand2	NA	-13.0	NA	-13.0
brand3	NA	-4.3	NA	-4.3
brand4	NA	-6.1	NA	-6.1
brand5	NA	-9.4	NA	-9.4
brand6	NA	-3.2	NA	-3.2
bottle2	NA	NA	0.3	0.3
bottle3	NA	NA	-0.1	-0.1
bottle4	NA	NA	-0.4	-0.4
bottle5	NA	NA	0.1	0.1
bottle6	NA	NA	-0.5	-0.5
bottle7	NA	NA	0.6	0.6
bottle8	NA	NA	-0.2	-0.2

Table 3: MLE coefficient values for Models α, $\alpha\delta$, $\alpha\gamma$, and $\alpha\gamma\delta$. Notice that the `brand` coefficients are larger than the `bottle` coefficients. To indicate which of the coefficients in this table are plausibly non-zero, a table of p-values for their associated t-statistics are presented below[5].

	Model.a	Model.ad	Model.ag	Model.agd
(Intercept)	0	0	0.00000	0.00000
brand2	NA	0	NA	0.00000
brand3	NA	0	NA	0.00000
brand4	NA	0	NA	0.00000
brand5	NA	0	NA	0.00000
brand6	NA	0	NA	0.00000
bottle2	NA	NA	0.92688	0.61104
bottle3	NA	NA	0.97560	0.86516
bottle4	NA	NA	0.87845	0.39819
bottle5	NA	NA	0.96584	0.81212
bottle6	NA	NA	0.86401	0.34465
bottle7	NA	NA	0.82573	0.22628
bottle8	NA	NA	0.95122	0.73430

Table 4: p-values for the MLE coefficients of Table 1 under a two-sided t-test, rounded to the fifth decimal place.

Question B: The *chmiss* dataset

This question involved the `chmiss` dataset, the entries of which correspond to postcode districts in 1970s Chicago. The response variable, `involact`, was a count of the number of government-mandated high-risk home insurance policies per 100 housing units in the district at the time[6]. The dataset's predictor variables had the following meanings:

- `race` - racial composition in percent minority,
- `fire` - fires per 100 housing units,
- `theft` - thefts per thousand population,
- `age` - percent of housing units built before 1939,
- `income` - median family income in thousands of dollars

The dataset consisted of $n = 47$ districts, however only $n_c = 27$ did not contain any missing records. The question invited us to identify the two most influential explanatory variables, imputing or omitting for the missing records as we saw fit. Several imputation methods were considered, however for the purposes of analysis only omission and median imputation were used. Other methods considered included nearest-neighbours impuration and spatial imputation, i.e. using the median of the districts that bordered the one with the missing value. No district was missing more than one value and as Table 4 shows, the missing values were distributed across all five features as well as the reponse, `involact`.

Censorship could have been correlated with one variable, which would affect the conclusions that could be drawn from the analysis. For example, it seemed plausible that record-keeping in run-down districts was more difficult, possibly meaning they were more prone to censorship. If such a correlation existed, then censored entries would be clustered in that variable's dimension. The lattice plot in Figure 1 was used to identify whether censored datapoints were unusual in any respect. Fitting a logistic regression model was also considered, to see whether any variable was predictive of censorship, but was not accomplished because of time limitations. The lattice plot allowed two censored outliers to be identified, one with an extreme `involact` value and one with an extreme `theft` value. District 60621 contained the extreme `involact` value and had missing `theft` data, while 60607 had the highest theft rate and was missing its `fire` data. Otherwise, there was little evidence indicating that a particular record was more likely to be censored as a result of its uncensored contents.

```
    race     fire    theft      age involact   income 
       4        2        4        5        3        2 
```

Table 5: Number of missing values per feature in the `chmiss` dataset, of a maximum of 47.

Broadly speaking, the correlation coefficients and plots of Figure 1 suggested that `theft` was least likely to be predictive of `involact`, while the other features - `race`, `fire`, `age`, and `income` - were more plausible candidates for most-influential explanatory variable. Nonlinear transformations of the covariates were not included in the model since they could not be motivated from Figure 1's plots. Interaction terms were also omitted since the additional parameters would inflate the variance of parameter estimates and would complicate later analysis.

A Normal linear model was used in both cases, which was of the form

$$y = \mathrm{X}\,\beta + \epsilon \tag{1}$$

with $y \in \mathbb{R}^n$ denoting a random vector corresponding to `involact` values[7], X representing the matrix of covariate values, $\beta \in \mathbb{R}^p$ denoting a vector of coefficients, and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ denoting response variation attributable to unobserved random variables.

[6] Government-mandated implying that these were policies which insurance companies would otherwise refuse to provide.

[7] In reality the response is a positive integer, not a real number.

The results in the sections that follow indicate that `fire` and `race` are the most-influential explanatory variables. The statistics obtained support what an analyst might conclude by inspecting the plots of each covariate against `involact` shown in Figure 1.

Model fit with values omitted

```
Call:
lm(formula = involact ~ race + fire + theft + age + income, data = Chmiss)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53370 -0.16325 -0.07015  0.12615  0.66316

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.116483   0.605761  -1.843 0.079475 .
race         0.010487   0.003128   3.352 0.003018 **
fire         0.043876   0.010319   4.252 0.000356 ***
theft       -0.017220   0.005900  -2.918 0.008215 **
age          0.009377   0.003494   2.684 0.013904 *
income       0.068701   0.042156   1.630 0.118077
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3382 on 21 degrees of freedom
Multiple R-squared:  0.7911,        Adjusted R-squared:  0.7414
F-statistic: 15.91 on 5 and 21 DF,  p-value: 1.594e-06
```

Model fit with values imputed as median

```
Call:
lm(formula = involact ~ race + fire + theft + age + income, data = Chmiss)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00820 -0.17790  0.00386  0.20956  0.80602

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.124370   0.497138   0.250  0.80370
race         0.006764   0.002593   2.609  0.01262 *
fire         0.027126   0.009004   3.013  0.00442 **
theft       -0.002882   0.002675  -1.077  0.28763
age          0.006348   0.003041   2.087  0.04310 *
income      -0.029689   0.031055  -0.956  0.34467
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3765 on 41 degrees of freedom
Multiple R-squared:  0.6715,        Adjusted R-squared:  0.6315
F-statistic: 16.76 on 5 and 41 DF,  p-value: 5.321e-09
```

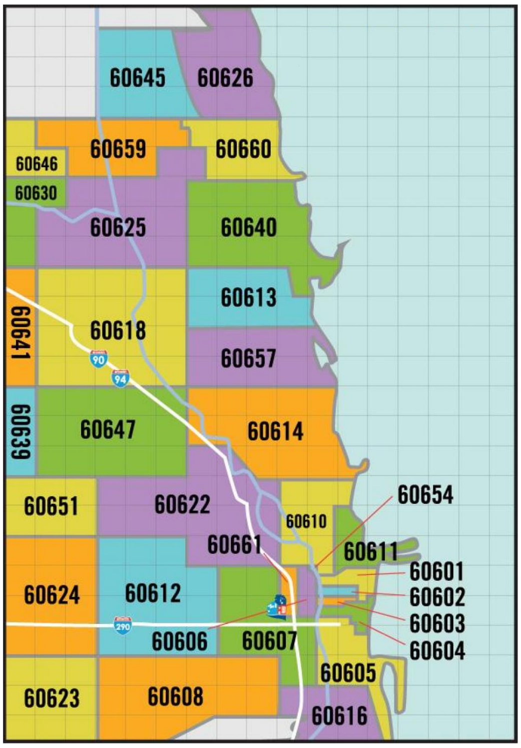

Figure 4: Postcode map of Chicago, courtesy of `http://nl.maps-chicago.com/postcode-kaart-chicago`.

Model fit with values imputed as median, `theft` outlier excluded

```
Call:
lm(formula = involact ~ race + fire + theft + age + income, data = Chmiss)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.86918 -0.20899 -0.03172  0.21421  0.79851 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.009394   0.499784  -0.019  0.98510   
race         0.007125   0.002573   2.769  0.00847 **
fire         0.031816   0.009478   3.357  0.00174 **
theft       -0.008136   0.004528  -1.797  0.07989 . 
age          0.007255   0.003069   2.364  0.02304 * 
income      -0.014512   0.032455  -0.447  0.65718   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3718 on 40 degrees of freedom
Multiple R-squared:  0.6862,	Adjusted R-squared:  0.647 
F-statistic:  17.5 on 5 and 40 DF,  p-value: 3.699e-09
```

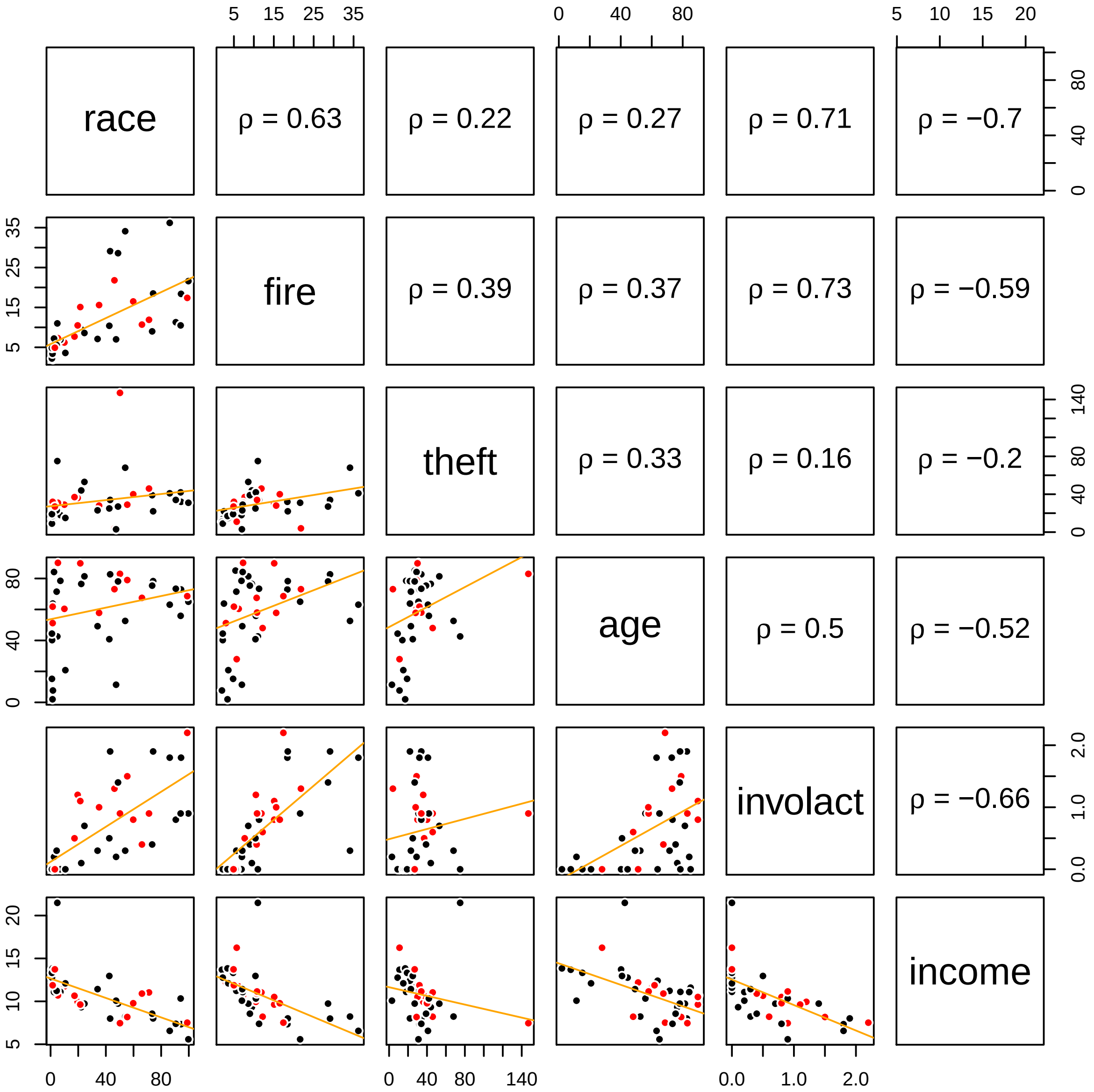

Figure 5: Scatterplot matrix for `chmiss` data with missing value entries highlighted in red. The correlation coefficient is calculated with `NA` entries omitted for both variables.