MSG500-MVE190 Linear Statistical Models

Project: U.S. county demographic information

Margareta Carlerös & Devosmita Chatterjee

January 10th 2019

(23 pages)

1 The data

The dataset obtained consisted of information about the 440 most populous counties in the US in 1990-1992. In total there were 17 variables.

2 Aim

To predict the number of crimes per 1000 population, crm_1000 (calculated with the crimes and popul variables as 1000*(crimes/popul)), using multivariate linear regression (MLR) and Poisson regression or negative binomial regression.

3 The variables

3.1 Initial variable selection and creation of new variables

Firstly, variables id and county were excluded since they are unique identifiers. The categorical variable state consisted of 48 unique levels, some levels with only one observation, meaning that it would not be useful, and it was therefore omitted. The categorical variable region, however, consisted of 4 levels each with between 77-152 observations. Upon plotting crm_1000 against this variable it was found that level 3=South and level 4=West showed very similar crime rates (see Appendix, fig. 1). These two variables were therefore combined as a baseline level.

Given that the response variable was crimes per 1000 population, suitable covariates should describe different characteristics of these 1000 typical people in each county. The pop1834, pop65plus, higrads, bachelors, poors and unemployed variables were all expressed as percentages and all describe these 1000 typical people in different ways so they were all included for further analysis. The percapitaincome variable was also included since it describes the average income per person while the totalincome variable was excluded as it depends on the population of the county. Similarly, the variables phys and beds are both counts and therefore also depend on the population of the county. Therefore, these were scaled by 1000 population just like the response variable crm_1000. These new variables, phys_1000 and beds_1000, described the healthcare resources that were available to a typical set of 1000 people in the county. The original phys and beds variables were omitted.

Lastly, the variables area and population were combined into a new measure, pop_density (calculated popul/area), describing the population density (population per square mile). In total 11 variables were kept for further analysis.

3.2 Variable transformations

The variable transformations are summarized in table 1. All variables required transformation except the higrads variable (see figures 2-7 in Appendix). The same transformation of the covariates was used in both multivariate linear regression and in Poisson / negative binomial regression. Any mentions of the covariates hereafter refer to the transformed versions of these. The response was only transformed in the multivariate linear regression, where a square root transformation was used. A log transformation was also explored for the response variable but yielded more outliers

when plotted (see figures 2-7 in Appendix) and slightly inferior predictive performance in later analyses. Therefore, only results using the square root transformation of the response will be reported and discussed here.

Table 1: Summary of variable transformations.

Type of variable	Original variable	Transformed variable
Response	crm_1000	sqrt(crm_1000)
Covariates	pop_density	log(pop_density)
	pop1834	log(pop1834)
	pop65plus	log(pop65plus)
	phys_1000	log(phys_1000)
	beds_1000	sqrt(beds_1000)
	higrads	-
	bachelors	log(bachelors)
	poors	log(poors)
	unemployed	log(unemployed)
	percapitaincome	log(percapitaincome)

4 Training and testing data

One county has markedly higher crime rate than the other counties (see figure 1). This observation, Kings county in New York state with a crime rate of nearly 300 per 1000 population was dropped from the dataset. The remaining 439 observations in the dataset were randomly split into a training set (351 observations) and a testing set (88 observations). The training set was used for model selection and building. Given the small size of the dataset, 80% of the observations were allocated to the training set to allow for more data to train on (than if e.g. only 50% of the observations had been used for training). The testing set was used for model evaluation and for comparing the performance of the models obtained by multivariate linear regression and Poisson / negative binomial regression. In order to make the "random" allocation of observations into training and testing sets reproducible, the R function set.seed() was used with argument 1.

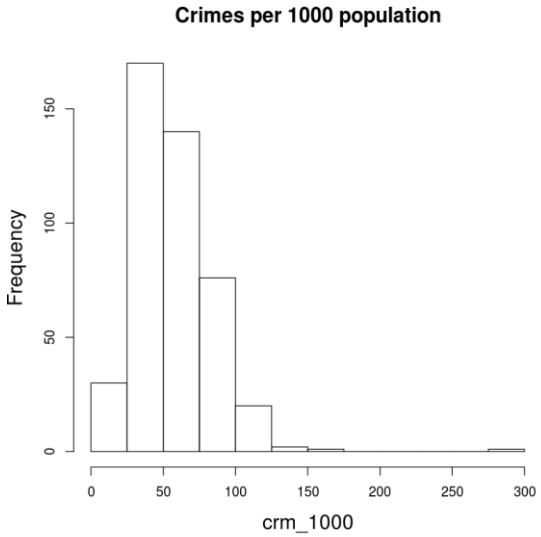

Figure 1: Distribution of response variable crm_1000 showing extreme outlier at crm_1000=296.

5 Methods

5.1 Multicollinearity of covariates

To identify possible collinearity between the continuous variables three different methods were used. Correlations between the variables were calculated and visualized using ggpairs() (R package GGally) and as well as with a heatmap where distances between variables were defined as 1-correlation using heatmap.2() from the package gplots. heatmap.2() also displayed a dendrogram allowing for clusters of covariates to be identified. Furthermore, the variance inflation factor (VIF) was calculated by first constructing the full model and then applying the vif() function from the R cars package. The VIF is defined as the ratio of the variance of a parameter estimate when the full model is fitted divided by the variance of that parameter estimate in a model which only contains the parameter estimate [1, p.101]. It is known that in pairs of highly collinear variables at least one of the variables will have a parameter estimate with inflated variance. The VIF therefore helps to identify one of the variables in such pairs. The smallest possible VIF is 1 which means there is no collinearity, while a value above 5 indicates potential issues with collinearity between that specific covariate and at least one other covariate in the full model.

5.2 Multivariate linear regression: model selection

Multivariate linear regression models were built using the square root of crm_1000 as response and different subsets of the remaining covariates with R function lm(). Three different model selection techniques were used to choose the best covariates to include in the model. The first two were based on backward search using the step() function in R. This algorithm started out with the full model and then eliminated the least useful variable in each step, according to some criterion, until a stopping criterion was reached. By default this function uses Akaike's Information Criterion (AIC) to decide which variables to eliminate and when the algorithm is stopped. The AIC includes a penalty term for large models, 2*p, where p is the number of parameters in the model. An alternative criterion is the Bayesian Information Criterion (BIC) which instead used the penalty term p*log(n), with n being the number of observations the model is built with. The BIC therefore penalizes larger models more, in general leading to the selection of smaller models. Model selection using backward search together with AIC and backward search with BIC were explored here.

The final model obtained by backward search may however just be a local optimum. Another concern is that the algorithm does not take into consideration if the model generalizes well to new data. An alternative to backward search is to perform an exhaustive search which is what a method referred to as "all subsets regression" does. For this method the data available for model building is split into a "training set" and a "test set". By enumerating all possible models (including an intercept only model), training each on the "training set" and subsequently using each fitted model to predict on the "test set", the predictive performance of the different models can be compared. Predictive performance is defined in terms of the estimated predictive mean squared error (pMSE), which is calculated as

$$pMSE(m)=\Sigma_i(y_i^{new}-\hat{y}(m)_i)^2/n,$$

where m is a specific model, y_i^{new} is an observation in the "test set" and $\hat{y}(m)_i$ is the corresponding prediction by model m given the information about the covariates available in the "test set", and n is the number of observations in the "test set". The model with the best predictive performance (i.e. with the lowest pMSE) is selected.

However, model selection is dependent on which observations end up in the "training set" and "test set". To limit this effect and to generate a model that better generalizes to new data a procedure called k-fold cross-validation was used. In this method the entire dataset that is available for building the model is split up into k so called folds. One fold is used as the "test set" while the remaining k-1 folds are used as a "training set". By repeating this a total of k times with each of the folds serving as "test set" one time, an average of the pMSE of each model across all folds can be calculated. Thus, the model with the lowest pMSE should be the one that demonstrates both good predictive performance as well as being able to generalize well to new data. Given the small size of the dataset provided, "all subsets regression" was performed in a 10-fold cross-validation setting, in order to allow for more data to be used in fitting the model.

5.3 Multivariate linear regression: model diagnostics and detection of outliers

After fitting the selected models to the training data, suspected outliers were identified using plots of leverage, studentized residuals, Cook's distance and DFBETAs. Leverage values were obtained in R by applying the hatvalues() function to the model. Observations with high leverage are those that are far from the center of the X-space. Any values larger than 2*p/n, with p the number of parameters in the model and n the number of observations in the training data, were regarded as suspected outliers. Studentized residuals are defined as

$$r_i^* = e_i / (s_{(i)} * \mathrm{sqrt}(1-h_{ii})),$$

where e_i the residual of observation i, $s_{(i)}^2$ is the variance estimate from a regression where observation i has been excluded and h_{ii} is the leverage. The studentized residuals were calculated using the rstudent() function. Studentized residuals have a t_{n-1-p} distribution which approaches a standard normal distribution for large n. Using this approximation, any observations having an absolute value of the corresponding studentized residual above 2 were flagged as potential outliers. Cook's distance values, defined as

$$D_i = (r_i^2/p) * (h_{ii}/(1-h_{ii}))$$

where r_i is the standardized residual, p the number of parameters in the model and h_{ii} the leverage, were generated using the cooks.distance() function. The Cook's distance measures the influence that an observation has on the parameter estimates. A cutoff of 4/n was used to identify suspected outliers. DFBETAs on the other hand measure the influence that an observation has on each parameter estimate, and thereby gives more detailed information than the Cook's distance. DFBETA values for each parameter estimate larger than 2/sqrt(n) were considered as suspected outliers.

Since there was some overlap between the outliers identified by DFBETAs and those identified by leverage, studentized residuals and Cook's distance, these last three diagnostic tools were used to flag observations as outliers. The criteria that was used was that at least 2 of the methods should identify the observation as a suspected outlier for it to be considered an outlier and subsequently removed from the training data for that model. After removal of the outliers the models were retrained.

5.4 Multivariate linear regression: adjusted R^2 and t-test

In order to measure how well the linear regression model captured the total variability in the training data, the adjusted R^2 was used. The reason for not using R^2, defined as 1 – SS(Error)/SS(Total), with $SS(Error)=\Sigma_i(y_i-\hat{y}_i)^2$ and $SS(Total)=\Sigma_i(y_i-\bar{y})^2$, was that this measure increases by adding covariates. The adjusted R^2 , defined as 1- ((n-1)* SS(Error))/((n-p)*SS(Total)), takes the number of covariates, p, in the model into account.

The t-test was used to test the null hypothesis that a parameter estimate was equal to zero against the alternative hypothesis that it was non-zero. The t-tests for each parameter estimate were performed at a significance level of α=0.05.

5.5 Poisson / negative binomial regression with offset

Since our response variable, crm_1000, was a rate and not a count, Poisson and negative binomial regression could not be applied directly. However, if using a log link function, log(crm_1000)=log(crimes/(popul/1000)), can be rewriten as log(crimes) - log(popul/1000), where crimes is a count and log(popul/1000) is called the offset term. Thus, by using the offset term we can model rates using Poisson and negative binomial regression as

$$\log(\mu) = \log(t) + X\beta,$$

with t=popul/1000. In R, a Poisson regression model was created using glm() with crimes as response and setting offset=log(popul/1000), while negative binomial regression model was created using glm.nb() (MASS package) with crimes as response and the term offset(log(popul/1000)) added as a covariate.

5.6 Poisson regression: testing for over- or underdispersion

In modeling the counts using Poisson regression the assumption is made that the expectation of the conditional mean and variance of the response is the same. However, often the variance is higher (overdispersion) which warrants use of a negative binomial regression instead. To test for overdispersion, the dispersiontest() (R package AER) was used. A value that is significantly greater than 1 is indicative of overdispersion.

5.7 Poisson / negative binomial regression: model selection

An initial model was created using the poors variable as covariate, since this variable was consistently included in the multivariate linear regression models. Additional covariates were added using forward search where the likelihood ratio test was used as a criterion to decide whether the model should be extended with a variable or not. The likelihood ratio test compares nested models where the null hypothesis is that the reduced model is sufficient. The test is based on the difference in deviance, where the deviance is a measure of the likelihood that our data was generated by a specific model at the maximum likelihood estimate of the parameters. Larger models have smaller deviance and a larger likelihood. Thus, a large difference in deviance between the reduced model and the extended model, indicates that using the extended model should be used. Under the null hypothesis of the likelihood ratio test, the difference in deviance has a chi-square distribution with k degrees of freedom, where k is the difference in the number of parameters between the models. The significance level used was α=0.05.

The covariate that resulted in the largest increase in the difference in deviance (i.e. reduced the deviance most when added) was included in the model at each step of the forward search. The final model was obtained when none of the remaining covariates yielded a significant likelihood ratio test. This model was then fitted to the training data.

5.8 Poisson / negative binomial regression: model diagnostics and detection of outliers

The diagnostic methods considered were standardized Pearson residuals and Cook's distance. Standardized Pearson residuals were accessed using influence()$pear.res on the model in R. The standardized Pearson residual is defined as

$$r_i^{p*} = r_i^p / \text{sqrt}(1-h_{ii}),$$

with r_i^p being Pearson's residual and h_{ii} the leverage. Any residuals whose absolute value was larger than 2 were considered suspected outliers. The Cook's distance was calculated using cooks.distance(). For generalized linear models the Cook's distance is defined as

$$C_i = (r_i^p / (1-h_{ii})^2) / (h_{ii}/(\text{ø}*p)),$$

where r_i^p and h_{ii} were defined as before, p the number of parameters in the model and ø the dispersion parameter specific to the model. Observations with a corresponding Cook's distance larger than 4/n were flagged as suspected outliers. In order to consider an observation as an outlier it was required that the observation should be a suspected outlier in both plots. After removal of the outliers from the training data the model was refitted.

5.9 Poisson / negative binomial regression: Wald test

The Wald test was used to test the null hypothesis that a parameter estimate was equal to zero against the alternative hypothesis that it was non-zero, just like the t-test in multivariate linear regression. The distribution of the Wald test statistic is a standard normal distribution. The Wald test was performed at a significance level of $\alpha=0.05$ for each parameter estimate.

5.10 Comparison of model performance

The predictive performance of the models was evaluated by having these predict on the test data set. The predicted values were "reverse transformed" so that they were all expressed in terms of crm_1000 (rather than e.g. sqrt(crm_1000)) to facilitate comparison between models. pMSE values were then calculated from these predictions and the observed crm_1000 from the test data. The predicted crime rates were also plotted against the observed for each model.

6 Results and discussion

6.1 Multicollinearity of covariates

High positive correlation was found between higrads and bachelors, where bachelors was highly correlated to several other variables. In addition, poors and percapitaincome exhibited a high negative correlation (see fig. 2 and 3). The VIF for bachelors and percapitaincome were both above 5 (see table 2), suggesting a problem with multicolllinearity with these variables or variables highly correlated with these. Therefore, all three methods indicated that one of bachelors and higrads as well as one of poors and percapitaincome should be omitted in order to reduce multicollinearity issues.

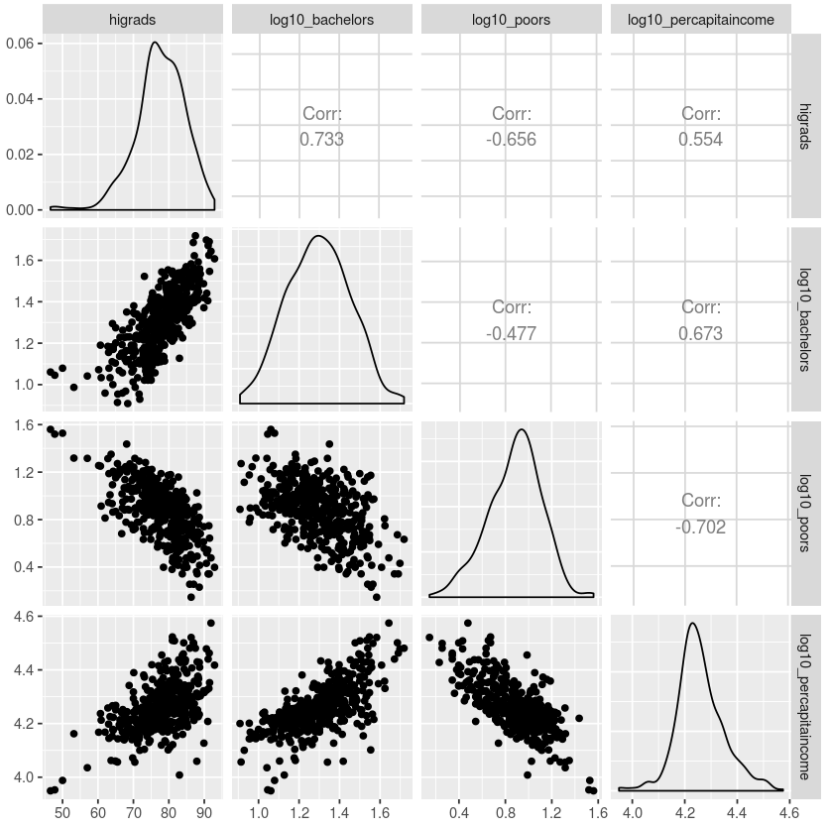

Figure 2: Scatterplots and correlations between highly collinear variables.

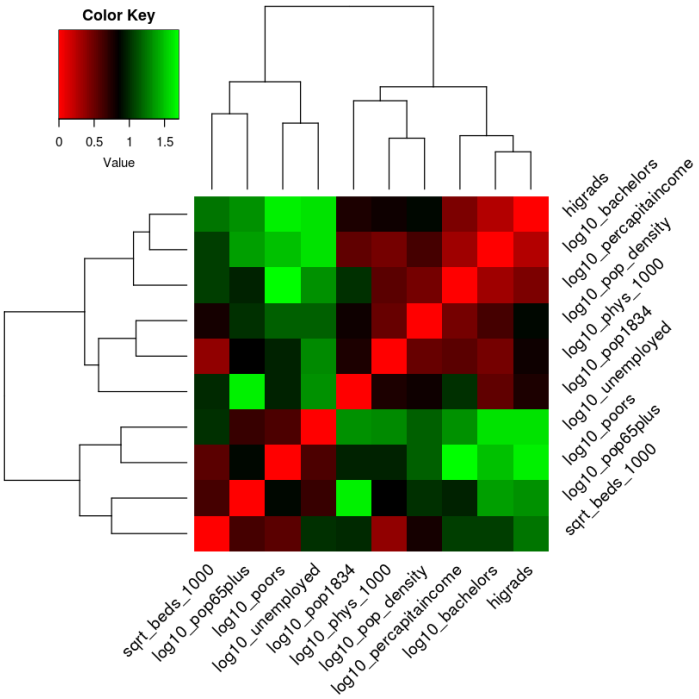

Figure 3: Heatmap of distances between covariates defined as 1-correlation including dendrogram.

Table 2: Variance inflation factors (VIF) for the covariates.

(Transformed) covariate	VIF
higrads	3.36
pop1834	2.62
pop65plus	2.59
phys_1000	4.49
beds_1000	3.03
bachelors	6.32
poors	4.46
unemployed	1.98
percapitaincome	5.92
pop_density	1.78

The variable higrads showed a negative correlation with the crime rate, i.e. higher proportions of high school-educated individuals in the populations appears to protect against crime. On the contrary, the bachelors variable shows a slight positive correlation with crime rate. Given that bachelors had a high correlation with many other variables (more so than higrads), it was decided to omit bachelors.

Poors had a strong positive correlation with crime rate. The poors variable says something about the income distribution of the 1000 typical individuals in the county that we are predicting crime rate amongst. The percapitaincome, however, only provides information about the average income of these 1000 people, so it is therefore less informative than poors. This led to the exclusion of the percapitaincome variable, while poors was kept.

The 9 variables that remained were pop1834, pop65plus, phys_1000, beds_1000, higrads, poors, unemployed, pop_density and region.

6.2 Multivariate linear regression: model selection

The multivariate regression models selected are shown in table 3. Backward search with AIC and "all subsets regression" with 10-fold cross-validation both selected model 1, while backward search with BIC selected model 2. The common variables were phys_1000, poors, pop_density and region, while model 1 in addition selected pop65plus, beds_1000 and unemployed.

Table 3: Multivariate regression models selected. Model 1 was selected both by backward search with AIC and "all subsets regression" with 10-fold cross-validation. Model 2 was selected by backward search with BIC.

Name	Model
Model 1	$y_i = \beta_0 + \beta_{pop65plus}X_{i,pop65plus} + \beta_{phys\_1000}X_{i,phys\_1000} + \beta_{beds\_1000}X_{i,beds\_1000} + \beta_{poors}X_{i,poors} + \beta_{unemployed}X_{i,unemployed} + \beta_{pop\_density}X_{i,pop\_density} + \beta_{region1}X_{i,region2} + \epsilon_i$
Model 2	$y_i = \beta_0 + \beta_{phys\_1000}X_{i,phys\_1000} + \beta_{poors}X_{i,poors} + \beta_{pop\_density}X_{i,pop\_density} + \beta_{region1}X_{i,region2} + \epsilon_i$

6.3 Multivariate linear regression: model diagnostics and detection of outliers

14 observations were identified as outliers for model 1 and 2 (table 4 and figures 8-11 and figures 12-14 in Appendix) , with 11 outliers being common to both sets of outliers.

6.4 Multivariate linear regression: adjusted R^2 and t-test

Upon removal of the outliers the adjusted R^2 of model 1 increased from 0.5862 to 0.6441, while it increased from 0.5794 to 0.6362 for model 2 (table 4). Removing the outliers seems to have

removed some of the unexplained variation in the training data. The consistently higher adjusted R^2 value of model 1 suggests that this model better explains the variation in the training data and that model 2 may be overly simplistic. However, a simpler and thereby more rigid model may will be less prone to overfitting and may generalize better to new data.

Table 4: Adjusted R^2 values of multivariate regression models before and after removal of outliers and the number of outliers removed.

Model name	Adj. R^2 before outlier removal	Adj. R^2 after outlier removal	Number of outliers removed
Model 1	0.5862	0.6441	14
Model 2	0.5794	0.6362	14

All parameter estimates for model 1 were significantly different from zero by t-test before outlier removal (table 5). The beds_1000 variable had a p-value just below the 0.05. However, after outlier removal the beds_1000 parameter was no longer significantly different from zero (table 5). There may be several reasons for this. Firstly, the outliers present in the training data could have influenced the model selection process. It was attempted to remove the outliers that were common to model 1 and 2 from the training data followed by reselecting and retraining the model, removing outliers and refitting the model again. However, this did not remedy issues with non-significant parameter estimates. Another likely cause for the non-significant parameter estimate in model 1 may be that by removing observations in the training data, the variance of the parameter estimates increases, making the estimates more prone to appear non-significant. For the other parameter estimates (except region 2), the reduction of noisiness of the data may have outweighed the smaller size of the training data meaning that the net effect was reduction of parameter estimate variance.

Table 5: Parameter estimates and p-values from the t-test of multivariate linear regression model 1 before and after removal of 14 outliers.

	Before outlier removal		After outlier removal	
Parameter	**Estimate**	**p-value**	**Estimate**	**p-value**
(Intercept)	2.5845	0.00018	2.86800	5.13e-06
log10_pop65plus	-1.1679	0.03805	-1.29922	0.01113
log10_phys_1000	1.1643	0.00400	1.85840	5.96e-07
log10_beds_1000	0.4228	0.04935	0.07489	0.70644
log10_poors	2.1671	1.49e-07	2.52875	3.68e-11
log10_unemployed	1.2619	0.02677	1.44979	0.00548
log10_pop_density	1.0646	2.46e-13	0.97208	1.34e-13
region1	-1.9311	< 2e-16	-1.76596	< 2e-16
region2	-1.1026	4.44e-11	-0.81265	6.15e-08

The model 2 parameter estimates were, however, significant by t-test both before and after removal of outliers (table 6). All parameter estimates were either as or more significant after outlier removal, indicating that the noise in the training data had been reduced. This may also reflect the fact that a simpler model is not as sensitive to shifts in the training data as the more complex model 1 is.

Table 6: Parameter estimates and p-values from the t-test of multivariate linear regression model 2 before and after removal of 14 outliers.

	Before outlier removal		After outlier removal	
Parameter	**Estimate**	**p-value**	**Estimate**	**p-value**

(Intercept)	2.3371	4.99e-07	2.4707	3.98e-09
log10_phys_1000	1.3221	7.27e-06	1.4834	9.69e-08
log10_poors	2.8597	< 2e-16	2.8903	< 2e-16
log10_pop_density	1.1006	3.72e-14	1.0188	7.18e-15
region 1	-1.8244	< 2e-16	-1.7331	< 2e-16
region 2	-1.0152	4.90e-11	-0.8779	3.70e-10

Considering the parameter estimates after outlier removal in model 1 and 2, region1 (Northeast) and region2 (Midwest) had fewer crimes relative to the baseline regions (West and South) if all other variables were kept constant. In model 1, having a higher percentage of the population being over the age of 65 also seemed to protect against high crime rates. Unemployed showed a positive correlation with the response as did beds_1000 (although this was not significant). Variables phys_1000, poors and pop_density, that were present in both model 1 and 2, all showed a positive correlation with crime rate.

6.5 Poisson regression: testing for over- or underdispersion

The Poisson regression model with offset and initial covariate poors exhibited highly significant overdispersion with a value of 2396.142 ($p < 2.2e-16$). Overdispersion was also evident when including other covariates in the Poisson regression model (data not shown). Therefore, it was decided to model the crime rate using a negative binomial regression model with offset in order to account for the larger variance.

6.6 Negative binomial regression: model selection

The parameters selected by the forward search were (in sequence) poors, region, pop_density and phys_1000. Interestingly, these are the same parameters as were selected in model 2, reinforcing the relevance of these covariates in predicting crm_1000. The negative binomial model with offset including these covariates will hereafter be referred to as model 3.

6.7 Negative binomial regression: model diagnostics, detection of outliers and the Wald test

In model 3, 9 outliers were identified (see figure 15 in Appendix for plots). Removing these from the training data and subsequently refitting the model resulted in only slightly more significant parameter estimates (table 7). Regions 1 and 2 appeared to be significantly protective against crime relative to baseline by the Wald test. The remaining variables were all significantly positively correlated with crime by the Wald test.

Table 7: Negative binomial regression model obtained by forward selection. Wald test.

	Before outlier removal		**After outlier removal**	
Parameter	**Estimate**	**p-value**	**Estimate**	**p-value**
(Intercept)	2.64408	< 2e-16	2.60586	< 2e-16
log10_poors	0.77743	< 2e-16	0.80567	< 2e-16
region1	-0.52011	< 2e-16	-0.49674	< 2e-16
region2	-0.26525	3.72e-10	-0.23554	3.63e-10
log10_pop_density	0.29335	1.02e-13	0.29284	< 2e-16
log10_phys_1000	0.32899	6.26e-05	0.34508	1.43e-06

7 Evaluation and comparison of model performance

The model with the best predictive performance on this test data set was model 1, based on multivariate linear regression, with a pMSE of 264.61 (table 8). The negative binomial regression model, model 3, was inferior to the other two models with a pMSE of 277.31 when tested on this test data. Interestingly, model 2 (pMSE 268.19) and 3 are based on the same covariates, but for this data the assumptions made by the multivariate linear regression model, e.g. Gaussian errors, may be more appropriate than the assumptions made by the negative binomial regression model.

Table 8: pMSE values of the three final models when predicting on test data.

Model	pMSE
Multivariate linear regression – model 1	264.61
Multivariate linear regression – model 2	268.19
Negative binomial regression model – model 3	277.31

Figure 4 shows that, while there are some differences in pMSE, the predictions produced by the models appear to be similar. This indicates that, while model 1 performed the best here, there could be some value to each of the models. A different test data may cause different models to be favored. In general, a smaller model such as model 2 and 3, is more robust to perturbations in the test data while more complex models tend to overfit to training data which may result in high variance of the predictions on new data. On the other hand, a smaller model may fail to explain factors that are important in determining the response, i.e. they may be biased.

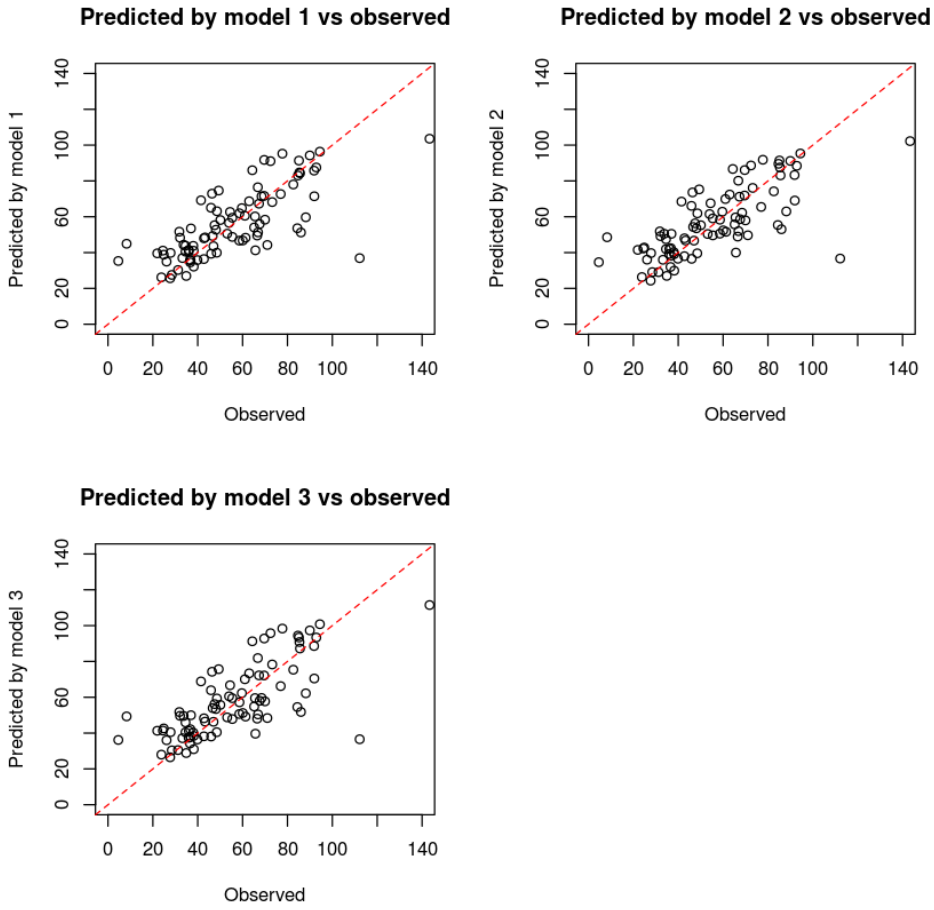

Figure 4: Predicted versus observed crm_1000 using test data and models 1 (top left), model 2 (top right) and model 3 (bottom left).

8 Conclusion

Despite the question of bias-variance tradeoff, all models contain the same four core covariates: poors, region, pop_density and phys_1000. Interestingly, two of these covariates, pop_density and phys_1000, were engineered from the original variables. This demonstrates the importance of carefully crafting relevant variables. For this dataset a multivariate linear regression model with 7 covariates appeared to be the most appropriate.

References

[1] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, 1st ed.,New York NY: Springer, 2013.

Appendix

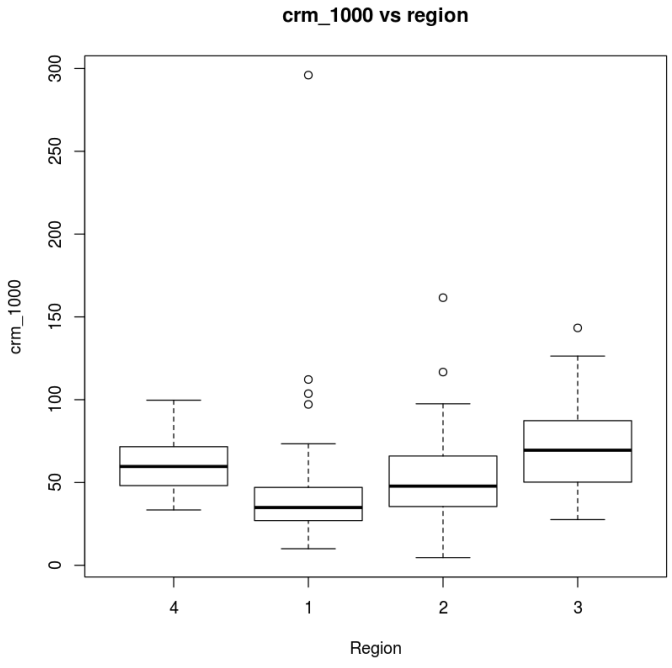

Figure 1: Boxplot of response crm_1000 versus original region variable with level 4 set as baseline.

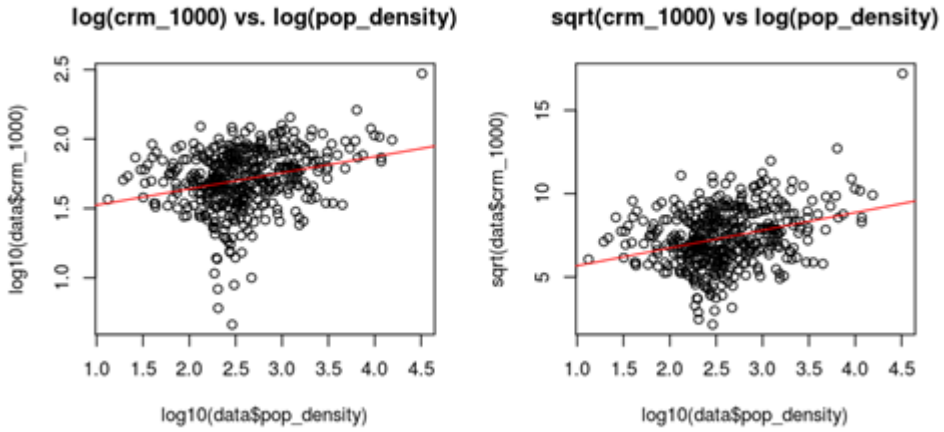

Figure 2: Plots of the transformed response log(crm_1000) (left) and sqrt(crm_1000) (right) versus log(pop_density). The red line denotes the simple linear regression model fitted using the corresponding transformed response and transformed covariate.

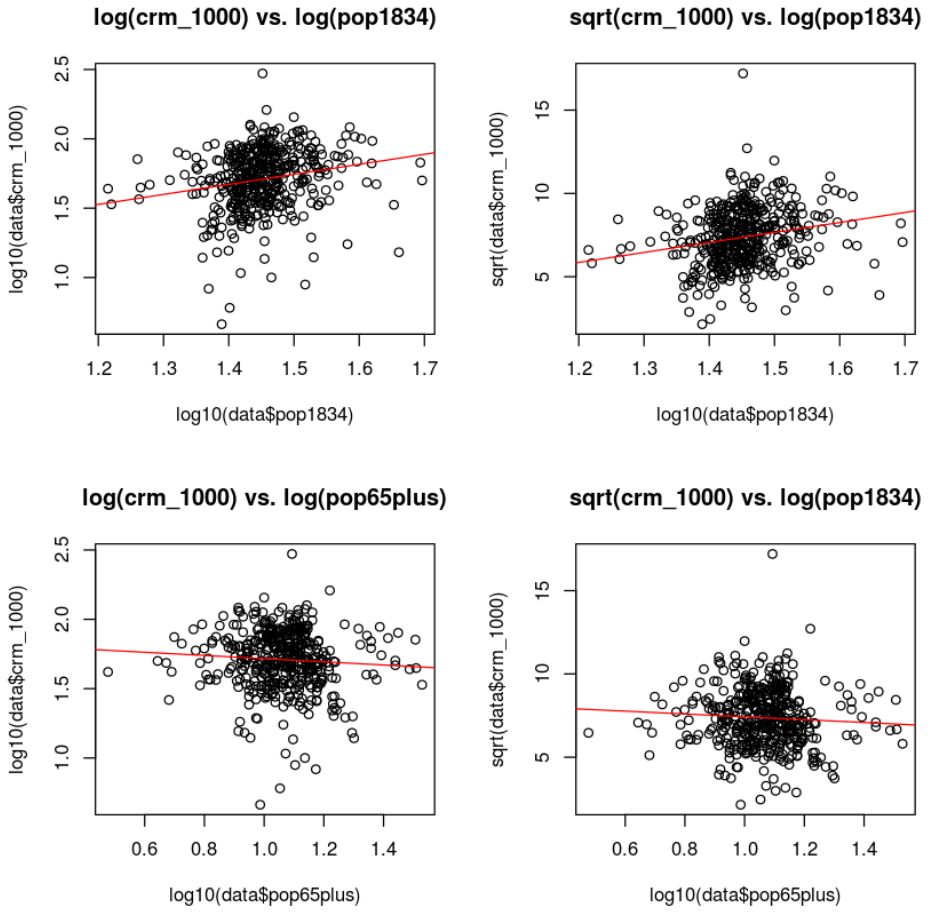

Figure 3: Plots of the transformed response log(crm_1000) (top left) and sqrt(crm_1000) (top right) versus log(pop1834) as well as the transformed response log(crm_1000) (bottom left) and sqrt(crm_1000) (bottom right) versus log(pop65plus). The red line denotes the simple linear regression model fitted using the corresponding transformed response and transformed covariate.

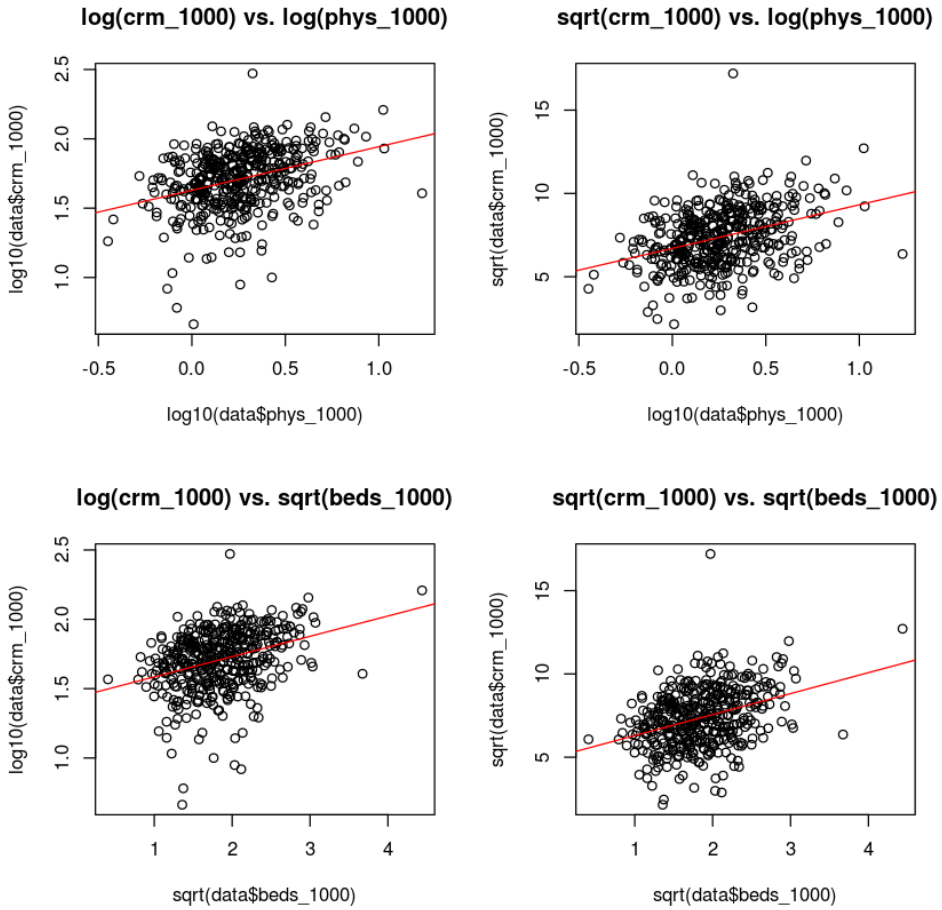

Figure 4: Plots of the transformed response log(crm_1000) (top left) and sqrt(crm_1000) (top right) versus log(phys_1000) as well as the transformed response log(crm_1000) (bottom left) and sqrt(crm_1000) (bottom right) versus sqrt(beds_1000). The red line denotes the simple linear regression model fitted using the corresponding transformed response and transformed covariate.

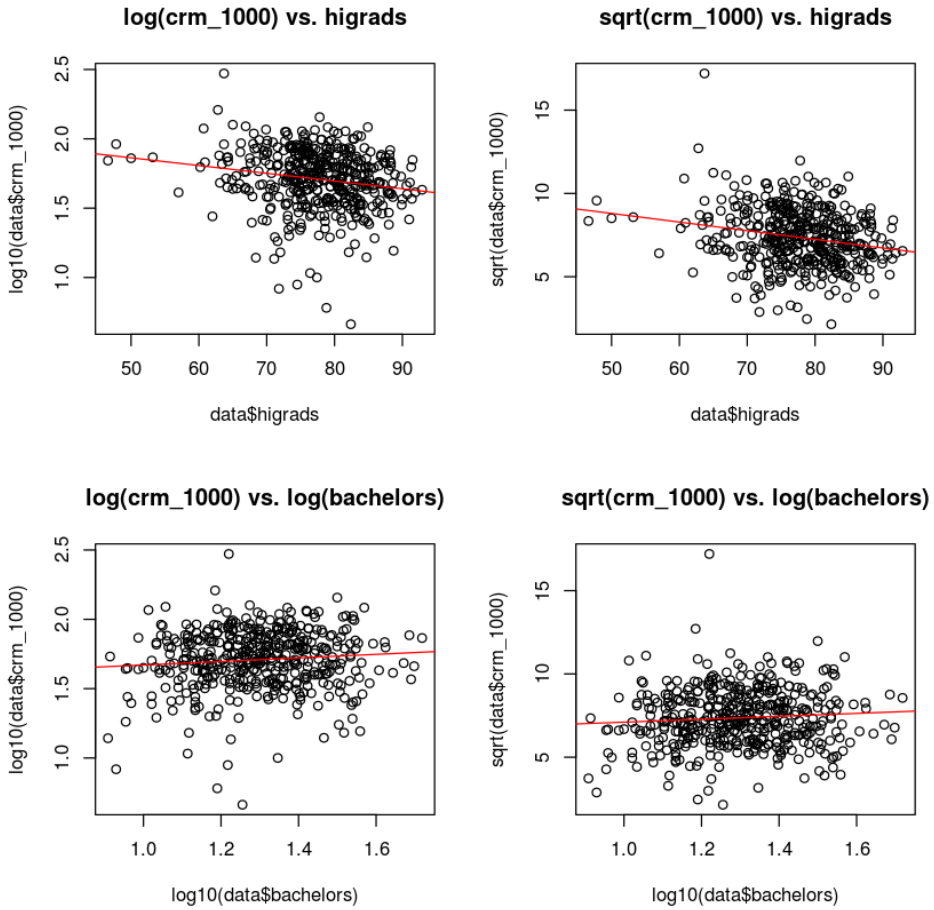

Figure 5: Plots of the transformed response log(crm_1000) (top left) and sqrt(crm_1000) (top right) versus higrads as well as the transformed response log(crm_1000) (bottom left) and sqrt(crm_1000) (bottom right) versus log(bachelors). The red line denotes the simple linear regression model fitted using the corresponding transformed response and (transformed) covariate.

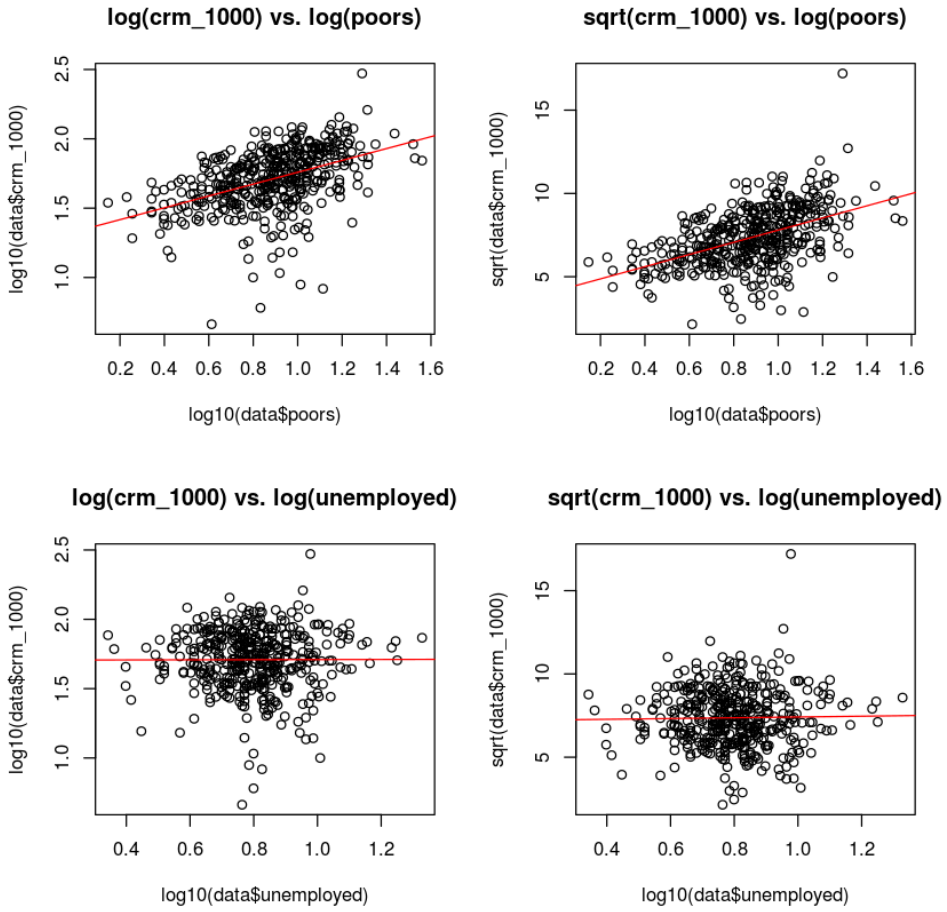

Figure 6: Plots of the transformed response log(crm_1000) (top left) and sqrt(crm_1000) (top right) versus log(poors) as well as the transformed response log(crm_1000) (bottom left) and sqrt(crm_1000) (bottom right) versus log(unemployed). The red line denotes the simple linear regression model fitted using the corresponding transformed response and transformed covariate.

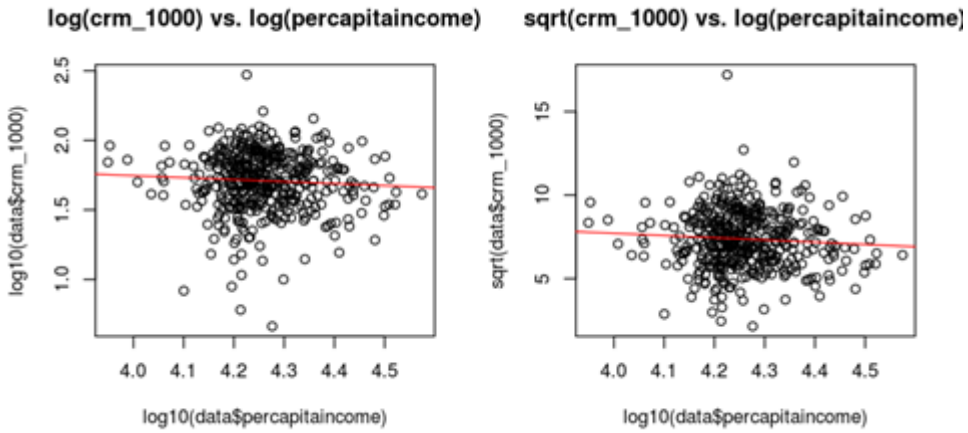

Figure 7: Plots of the transformed response log(crm_1000) (left) and sqrt(crm_1000) (right) versus log(percapitaincome). The red line denotes the simple linear regression model fitted using the corresponding transformed response and transformed covariate.

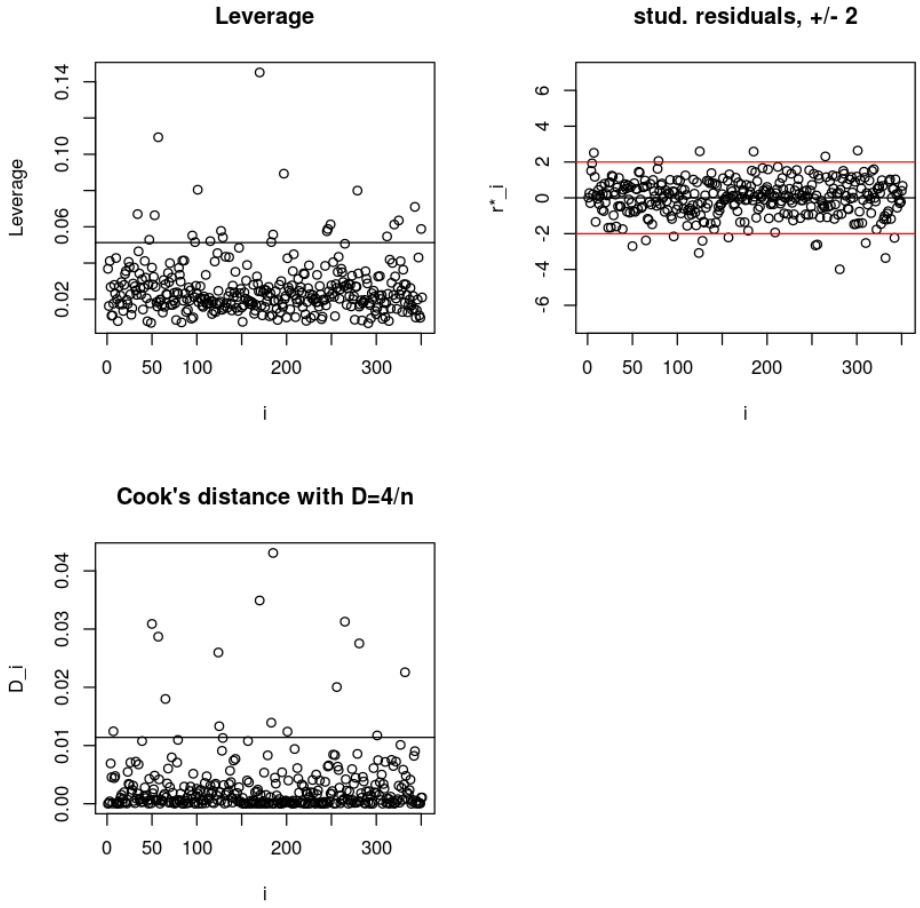

Figure 8: Leverage (top left), studentized residuals (top right) and Cook's distance (bottom left) plots for model 1.

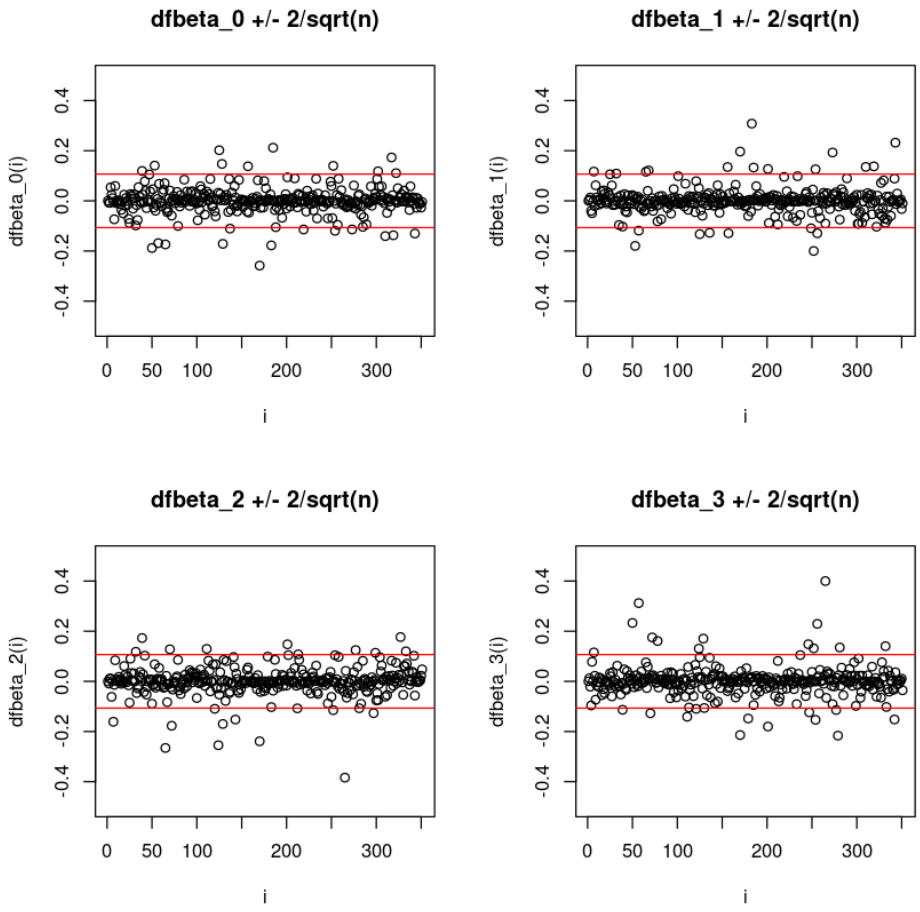

Figure 9: $DFBETA_0$, $DFBETA_1$, $DFBETA_2$ and $DFBETA_3$ plots for model 1.

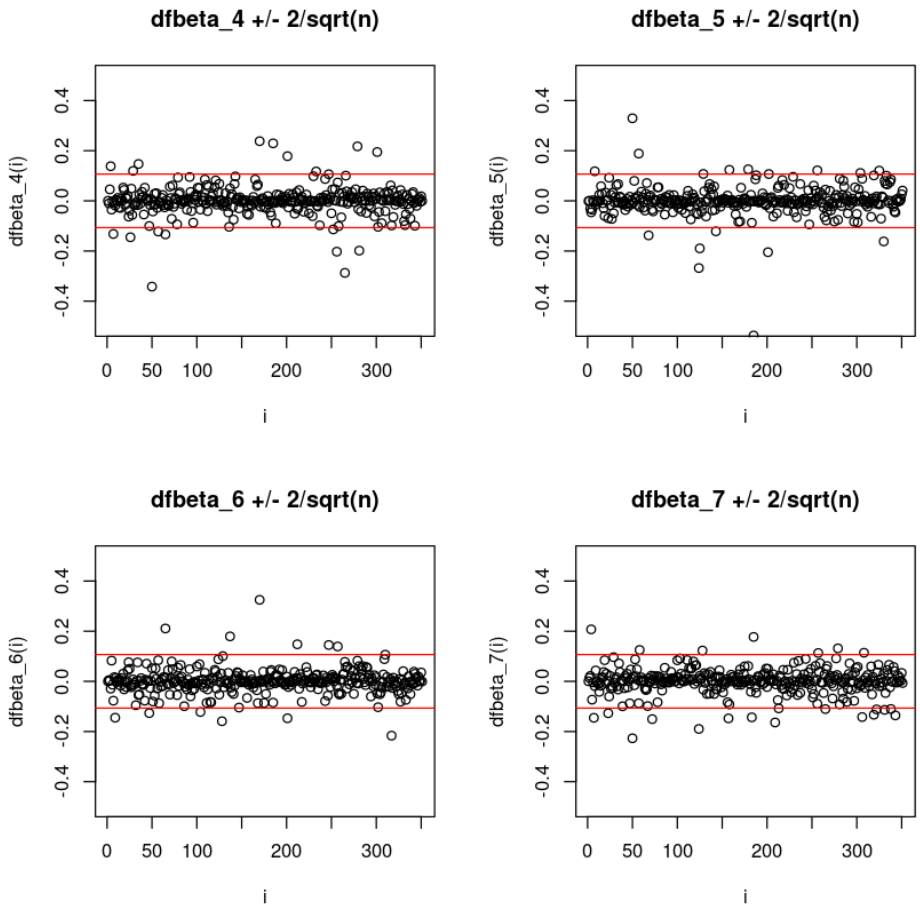

Figure 10: DFBETA$_4$, DFBETA$_5$, DFBETA$_6$ and DFBETA$_7$ plots for model 1.

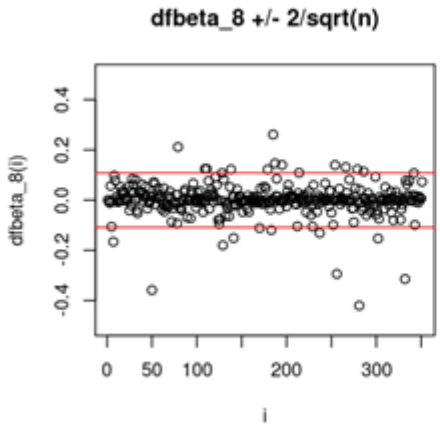

Figure 11: DFBETA$_8$ plot for model 1.

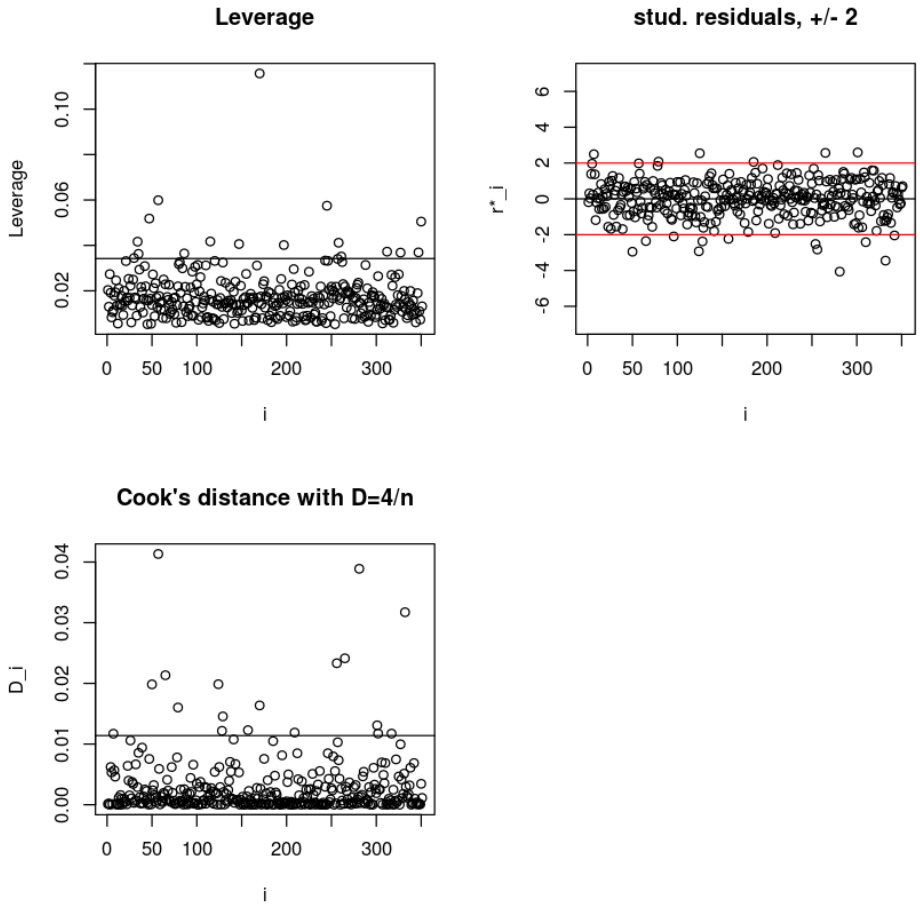

Figure 12: Leverage (top left), studentized residuals (top right) and Cook's distance (bottom left) plots for model 2.

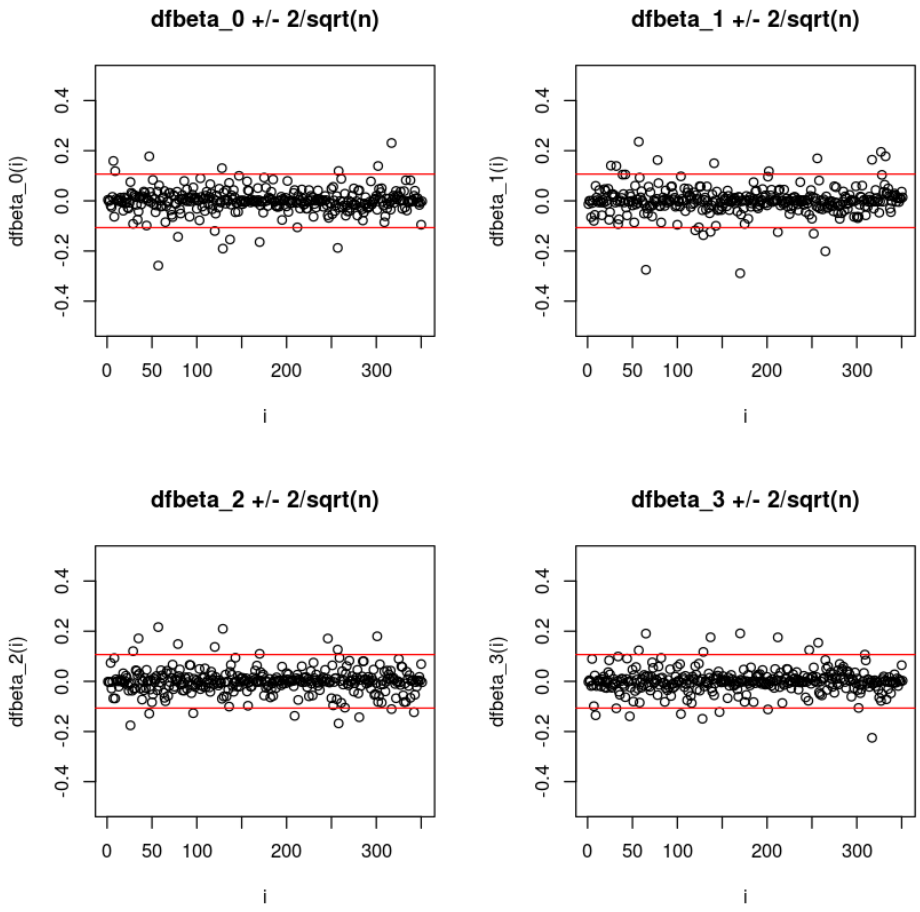

Figure 13: $DFBETA_0$, $DFBETA_1$, $DFBETA_2$ and $DFBETA_3$ plots for model 2.

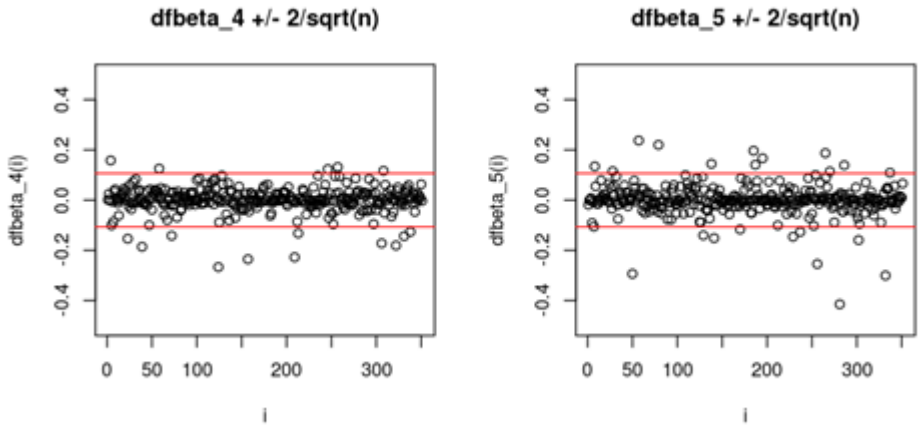

Figure 14: $DFBETA_4$ and $DFBETA_5$ plots for model 2.

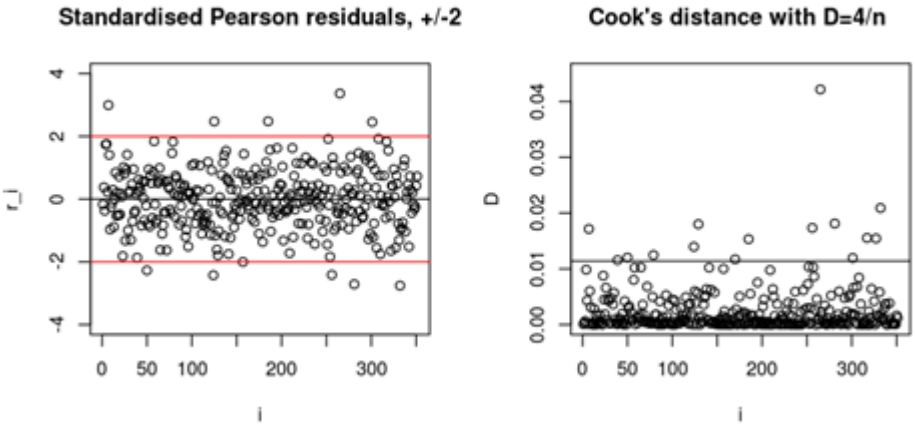

Figure 15: Standardized Pearson residuals (left) and Cook's distance (right) plots for model 3.