Project 2

Axel Sjöberg & John Rapp Farnes

14 maj 2019

Contents

1 Introduction

1.1 Background and dataset

The objective of this report is to determine which covarites that can be used to predict if a US county has a low or high crime rate (serious crimes per 1000 inhabitants). The dataset used to do this was county demographic information (CDI) for 440 of the most populous counites in the US 1990-1992. Each county records includes data on the 14 variables listed below in table 1. Counties with missing data has been removed from the dataset.

Table 1: CDI dataset columns

Variable	Description
id	identification number, 1–440
county	county name
state	state abbreviation
area	land area (square miles)
popul	estimated 1990 population
pop1834	percent of 1990 CDI population aged 18–34
pop65plus	percent of 1990 CDI population aged 65 years old or older
phys	number of professionally active nonfederal physicians during 1990
beds	total number of beds, cribs and bassinets during 1990
crimes	total number of serious crimes in 1990
higrads	percent of adults (25 yrs old or older) who completed at least 12 years of school
bachelors	percent of adults (25 yrs old or older) with bachelor's degree
poors	Percent of 1990 CDI population with income below poverty level
unemployed	percent of 1990 CDI labor force which is unemployed
percapitaincome	per capita income of 1990 CDI population (dollars)
totalincome	total personal income of 1990 CDI population (in millions of dollars)
region	Geographic region classification used by the U.S. Bureau of the Census, including Northeast, Midwest, South and West

In order to measure crime rate, another varible called `crm1000` was added, descibing the number of serious crimes per 1000 inhabitants. Using `crm1000`, counties were divided into counties with high or non-high crime rate, where counties with crime rate higher than the median of `crm1000` in the dataset were categorized as having a high crime rate. This crime status of the county was stored in another column called `hircrm`, which takes the value 1 if the county is a high crime county and 0 if it is a low crime county. In this paper, this binary varible will be used as the dependent varible. Similar to crime rate, a variable `phys1000` was also added, measuring the number of physicians per 1000 inhabitants.

1.2 Model

The binary dependent variable was modelled using a logistic regression model. This model assumes that the log-odds of a certain observation i is a linear combination of its covariates $X_{j,i}$ and parameters β_i. As such, the model looks like:

$$\ln \frac{p_i}{1-p_i} = \beta_0 + \sum_j \beta_j \cdot X_{j,i} \tag{1}$$

2 Analysis

2.1 The higrad model

2.1.1 Introduction

The first model considered has `higrads` as the sole covariate. As such, the model becomes:

$$\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_{higrads} \cdot X_{higrads,i} \tag{2}$$

In order to determine if there is a relationship between `hicrm` and `higrads` they are plotted against each other in figure 1. Because `hicrm` is a binary varible it is very difficult to determine if there is a relationship only by the pattern in the plot. In order to clarify this relationship, a kernel smoother was added to the plot, where a smooth line was attained with a bandwidth of 20. In addition, the fitted model along with its 95 % confidence interval are included.

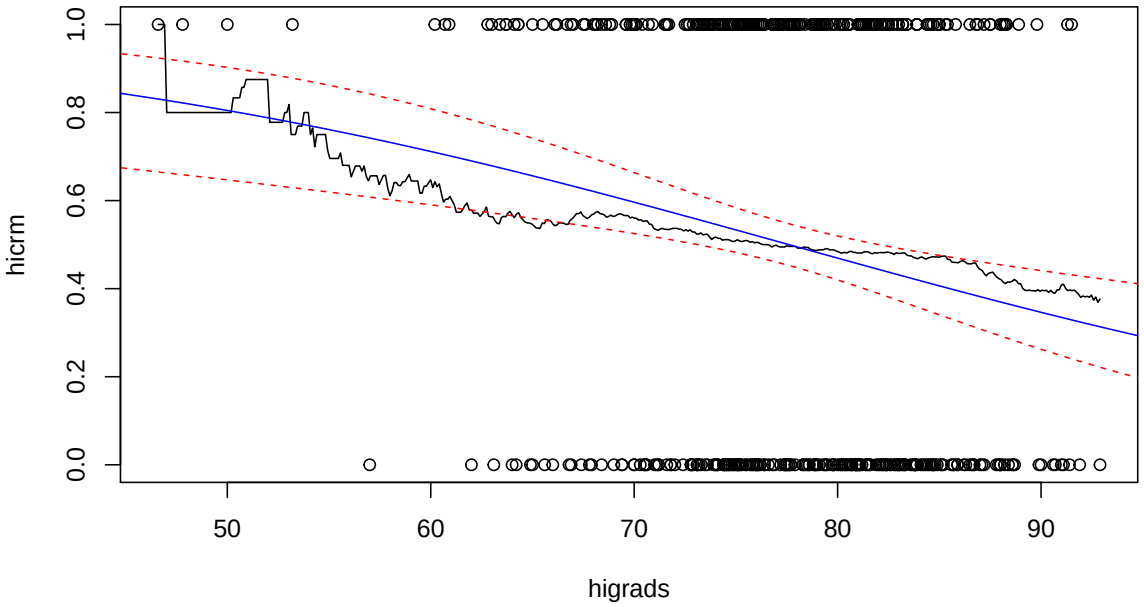

Figure 1: Plot of `hicrm` against `higrads`, including kernel smoothing and prediction of fitted model with 95 % confidence interval

As seen in 1, the kernel curve looks S-shaped, implying that a logistic model may be appropriate to describe the relationship. Further, the S-shape is "downward facing", implying a negative $\beta_{higrads}$, i.e. that the probability of a county being classified as a high crime decreases when the amount of higrads in the county increases. Another thing to note in figure 1 is how few data points exists with `higrads` below 60, meaning that significance is low in this region. In addition, looking at points with `higrads` over 60, the kernel curve and fitted line only cover about 25% - 75%, implying that `higrads` may not predict `hicrm` well.

2.1.2 Fitted model and significance

In order to study the significance of the model, the β values together with their 95 % confidence inteval are presented in table 2.

Table 2: β-values of `higrad` model, with 95 % confidence inteval

	Estimate	2.5 %	97.5 %	P-value
β_0	3.98	1.81	6.25	0.00044
$\beta_{higrads}$	-0.05	-0.08	-0.02	0.00041

As seen in the table, all of the P-values are > 0.05, meaning that that `higrads` has a statistically significant effect on `hicrm`. This effect can be measured by looking at $e^{\beta_{higrads}}$, showing that an increase of 1% in `higrads` decreases odds of `hicrm` by 5%, while an increase of 10% decreases odds of by 40.1%.

2.1.3 Model predictions

Using the higrads model: the probability, with 95 % confidence interval, of having a high crime rate in a county where the amount of higrads is 65 (percent), and where it is 85 (percent) is predicted. The result may be found in table 3.

Table 3: Predictions of `higrads` model

Higrads	Probability (%)	2.5 %	97.5 %
65	65.6	55.9	74.2
85	40.6	34.1	47.6

2.1.4 Model performance analysis

In order to analyze model performance, the sensitivity and specificity of the model was calculated. The sensitivity of a model is the ratio of predicted positives to real positives in the dataset, while the specificity of a model is the ratio of predicted negatives to real negatives in the dataset. As such, the higher the value of the sensitivity and specificity, the better.

The sensitivity of the model was 55.5% and the specificity of the model was 57.3%, implying that the model does not clasify the crime rate of the counties rather successfully.

2.2 The region model

2.2.1 Introduction

Next, a logistic model was adopted based on the `region` covariate. Since `region` is not continous, but categorial, it is modelled using "dummy variables" X_i. In order to implement this effectively, one of the categories is chosen as a reference variable, and the effects of other categories are measured in comparison to it.

In order to determine this reference variable, a cross-tabulation of the data between `region` and `hirm` is studied, see table 4.

Table 4: Cross-tabulation between `region` and `hicrm`

	Low crime	High crime
Northeast	82	21
Midwest	64	44
South	44	108
West	30	47

As a reference region, the one that has the largest number of counties in it's smallest low/high category was chosen. As a tie-breaker, the other low/high category was used. This approach produces the lowest standard error, and therefore highest significance. As seen in table 4, the above given condition results in choosing South as reference region.

Using this reference region, the logistic model becomes

$$\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_{Northeast} \cdot X_{Northeast,i} + \beta_{Midwest} \cdot X_{Midwest,i} + \beta_{West} \cdot X_{West,i} \tag{3}$$

Here, the β coefficients are measured relative to South, while β_0 is log-odds coefficient for South.

2.2.2 Fitted model and significance

The model was fit with the given data set, estimating β_i, shown together with its 95 % confidence interval and P-value in table 5.

Table 5: β-estimates for the `region` model, together with 95 % confidence interval and P-values

	Estimate	2.5 %	97.5 %	P-value
β_0	0.90	0.56	1.26	0.000
$\beta_{Northeast}$	-2.26	-2.87	-1.68	0.000
$\beta_{Midwest}$	-1.27	-1.80	-0.76	0.000
β_{West}	-0.45	-1.03	0.13	0.127

As may be seen in table 5, the P-values for β_{West} is not less than 0.05, indicating a lack of statistical significance in the difference between how South and West predicts `hivrm`.

Next, the odds-ratios for the different categories where determined. The odds-ratios measure the odds of a particular category in relation to the reference category. These may be calculated as $OR_i = e^{\beta_i}$ and may be seen in table 6.

Table 6: Odds-ratios for the `region` model, together with 95 % confidence interval

	OR	2.5 %	97.5 %
Northeast	0.10	0.06	0.19
Midwest	0.28	0.17	0.47
West	0.64	0.36	1.14

As seen in table 6, the odds-ratios are less than 1 for all categories but the reference region. This implies that the odds for all regions are lower compared to the reference region, i.e. that the probability of a high crime rate is lower in all regions compared to the reference region. This can also be seen in table 4.

2.2.3 Model predictions

Using the fitted model, the probabilies of having a high crime rate, with confidence interval, for the different regions was determined, shown in table 7.

Table 7: Probability of high crime rate (%), together with 95 % confidence interval for each of the regions

	Probability (%)	2.5 %	97.5 %
Northeast	20.4	12.6	28.2
Midwest	40.7	31.5	50.0
South	71.1	63.8	78.3
West	61.0	50.1	71.9

2.2.4 Model performance analysis

For the `region`, the sensitivity was 70.5%, while the specificity was 66.4%.

Comparing the two models analyzed so far, the `region` model performs better measured on sensitivity

and specificity, as seen in table 8

Table 8: Comparison of sensitivity and specificity of `higrad` and `region` model

Covariate	Sensitivity (%)	Specificity (%)
Higrads	55.5	57.3
Region	70.5	66.4

2.3 Combined model and comparison

2.3.1 Introduction

Next a model that uses both `higrads` and `region` is analyzed. As such, this model becomes

$$\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_{Northeast} \cdot X_{Northeast,i} + \beta_{Midwest} \cdot X_{Midwest,i} + \beta_{West} \cdot X_{West,i} + \beta_{higrads} \cdot X_{higrads,i} + \epsilon_i \tag{4}$$

2.3.2 Model comparison

In order to compare the models, metrics other than sensitivity and specificity may be studied. Some of these are AIC and BIC, which are penalized-likelihood criteria and Nagelkerke psuedo R^2, which is a measurment that increses up to 1 the better the model fit is. As they are defined, AIC and BIC should be as low as possible for a model to be performant, while psuedo R^2 should be as high as possible.

Comparison between the models in regards to AIC, BIC and Psuedo R^2 are seen in figure 2, while comparison of sensitivity and specificity is seen in table 9.

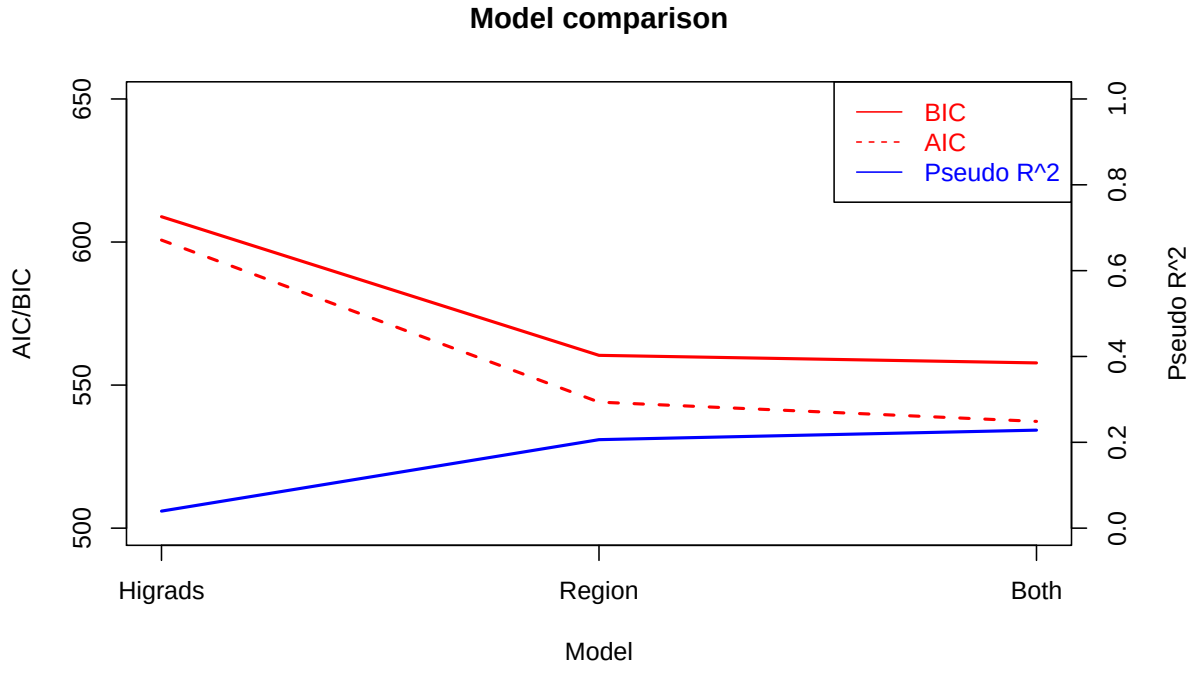

Figure 2: Comparison of AIC and BIC and Nagelkerke psuedo R^2 for the different models

Table 9: Comparison of sensitivity and specificity of models

Covariate	Sensitivity (%)	Specificity (%)
Higrads	55.5	57.3
Region	70.5	66.4
Both	70.5	67.3

As seen in figure 2 and table 9, the combined model with both the covariates performs the best on all the studied metrics.

2.3.3 Combined model performance

Performance of the combined model can be analyzed by studying a QQ-plot (see figure 3) the squared standardized Pearson residuals and the standardized deviance residuals against the linear predictor x^{β} (see figure 5). As well as the Cook's distance against the linear predictor, and against `higrads` and against `region` (see figure 6).

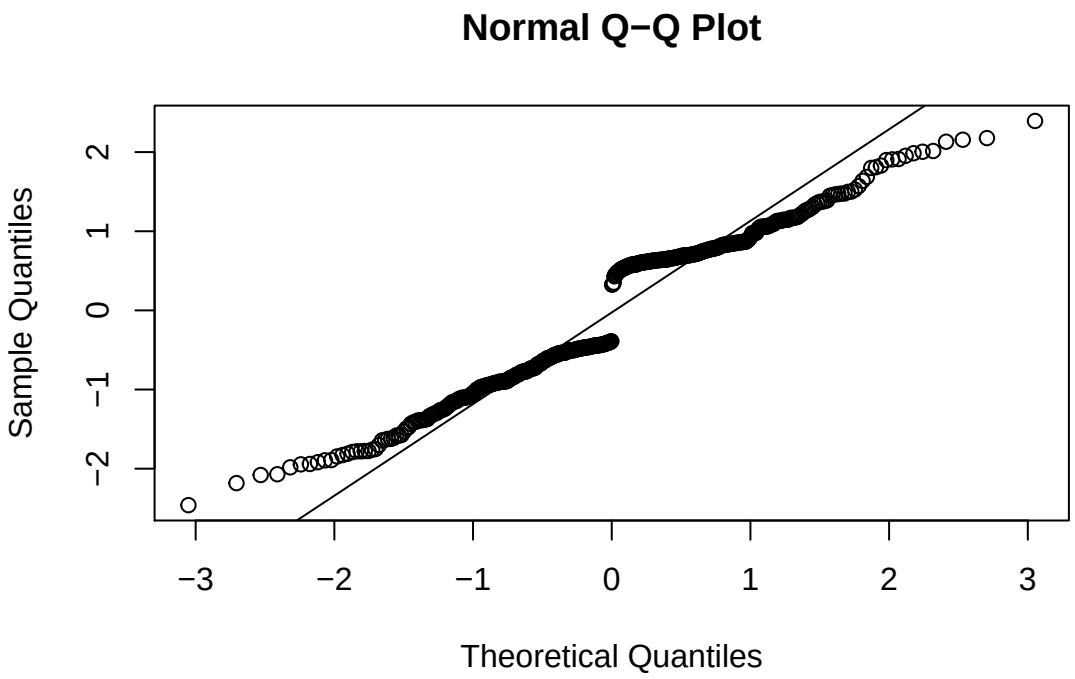

Figure 3: QQ-plot for the combined model

The QQ-plot seem to follow a bimodal distribution.

For a standardized pearson residual to be considered suspiciously large, $|r_i| > |\lambda_{\alpha/2}| \approx 2$. This is true for 8 of the standadized pearson residuals, which can be seen in figures XX and XX, where the standardized pearson residuals are plotted against the linear predictor. For a deviance residual to be considered too large $|r_i| > 2$. This is never the case which is verified by figure XX where the standardized devience residuals are plotted against the linear predictor.

In order to measure how the β-estimates were influenced by indiviual observations Cook's distance for logistic regression was calcultaed and plotted.

As can be seen in diagrams XX, XX, XX the vast majoriy of the the observations is below the horizontal line. The amount of counties with a high Cook's distance is limited to 3-4. The data entry with the largest Cook's distance is found in the South region while the rest of the counties with a high Cook's distance is found in the West region.

Anything alarmin? Any interesting finds?

2.4 Interaction model

2.4.1 Introduction

As a forth model, interaction terms are also considered, building in that the effect of higrads may be different in different regions, where the log-odds in the model includes interaction terms such as $\beta_{Northeast*higrads} \cdot X_{higrads,i}$.

2.4.2 Model performance

The performance of the interaction model compared to the combined model may be analyzed by the likelihood test. This test is similar to a partial F-test, but adapted for logistical regression, using likelihoods since sums of squares are not applicable. The likelihood test results in the interaction model being significantly better than the combined model, with a P-value of 0.019.

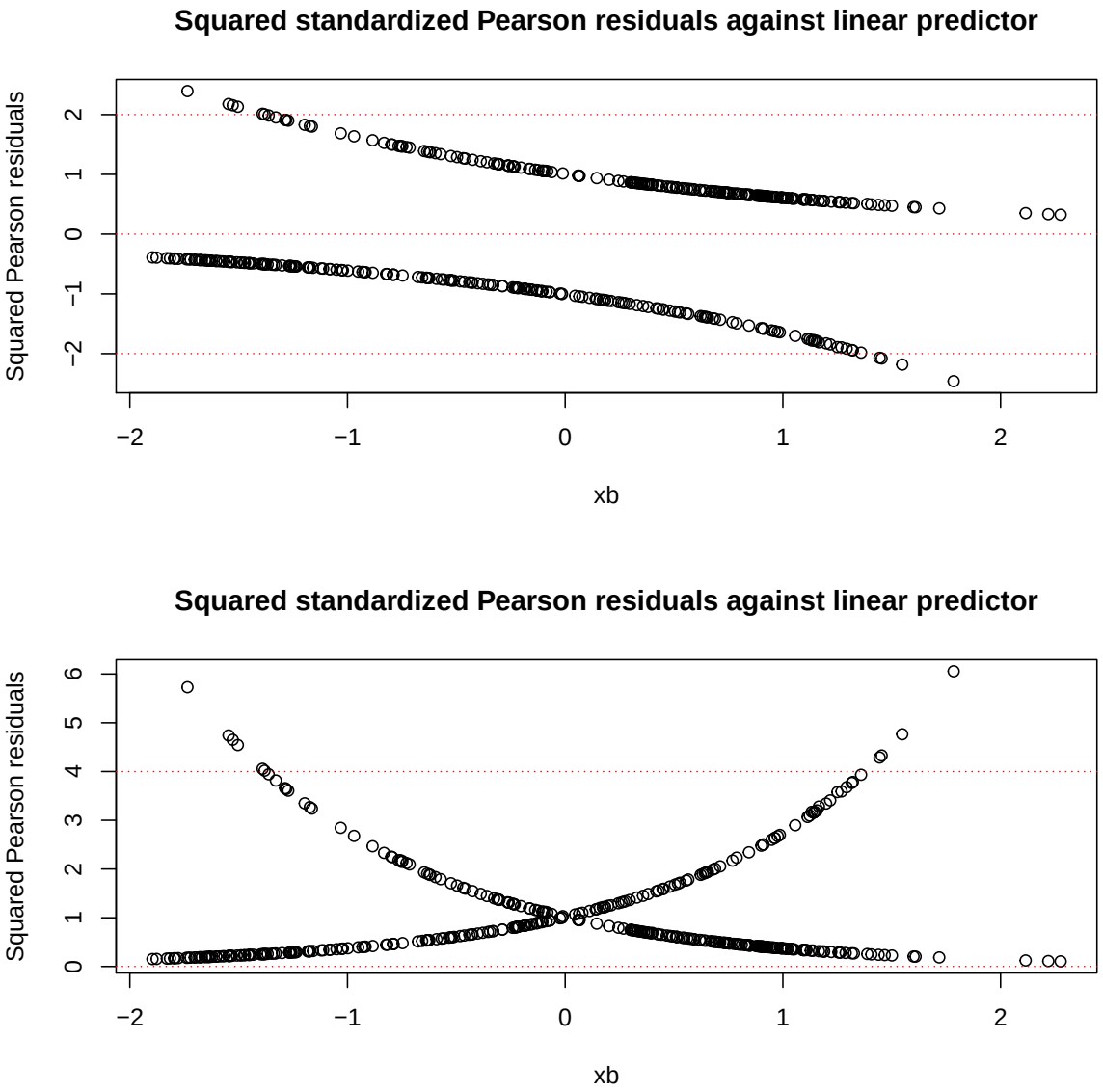

Figure 4: Squared standardized Pearson residuals as well as standardized deviance residuals for the combined model, against the linear predictor x^{β}

In addition, the AIC, BIC, Nagelkerke, sensitivity and specificity is compared to the combined model, in table 10. Performance of the interaction model can be analyzed by studying a QQ-plot (see figure 7) the squared standardized Pearson residuals and the standardized deviance residuals against the linear predictor x^{β} (see figure 8). As well as the Cook's distance against the linear predictor, and against `higrads` and against `region` (see figure 9).

Table 10: Comparison of sensitivity and specificity of models

Covariate	AIC	BIC	Sensitivity (%)	Specificity (%)	Pseudo R2
Combined model	533	566	72	68	0.25
Interaction model	537	558	70	67	0.23

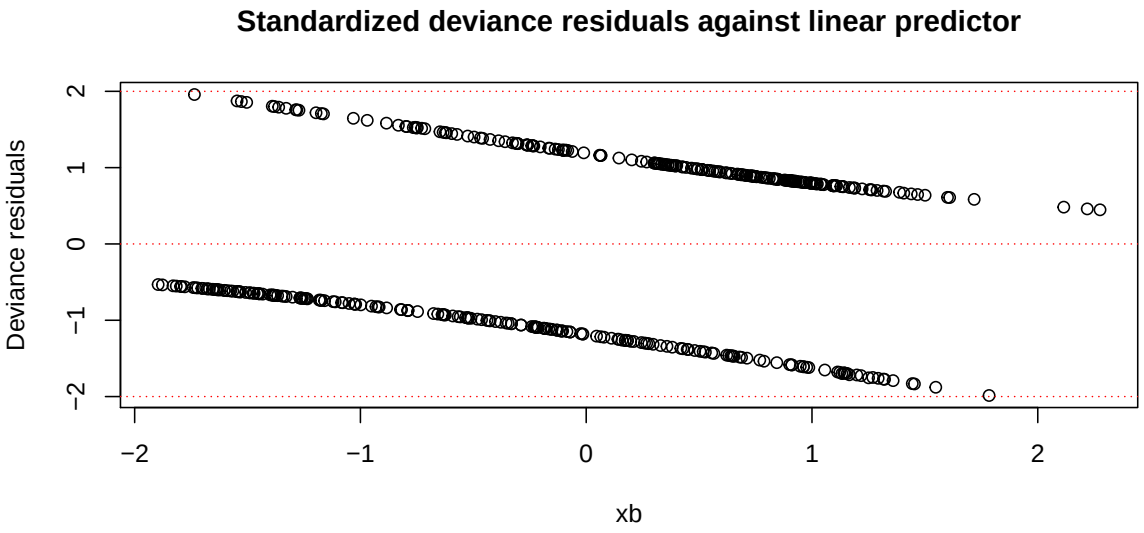

Figure 5: Squared standardized Pearson residuals as well as standardized deviance residuals for the combined model, against the linear predictor x^{β}

Both models perform better on some metrics, while performing worse on other. The interaction model has a worse AIC-value, but a lower BIC-value (KOMMENTERA). The sensitivity, specificity and Pseudo R^2 values are worse for the interaction model.

NÅGOT OM COOKS MM?

VILKEN ÄR BÄST?

2.5 Finding the optimal model

2.5.1 Methology

Next, an attempt to fit an optimal model to predict high crime rates is made, using the previous covariates, as well as `poors` and `pshys1000`. Interaction terms are ignored.

Models of increasingly complexity, adding more covariates are compared to each other on the used metrics, i.e. AIC, BIC, Pseudo R^2, sensitivity and specificity. In addition, the result of automatic selection using

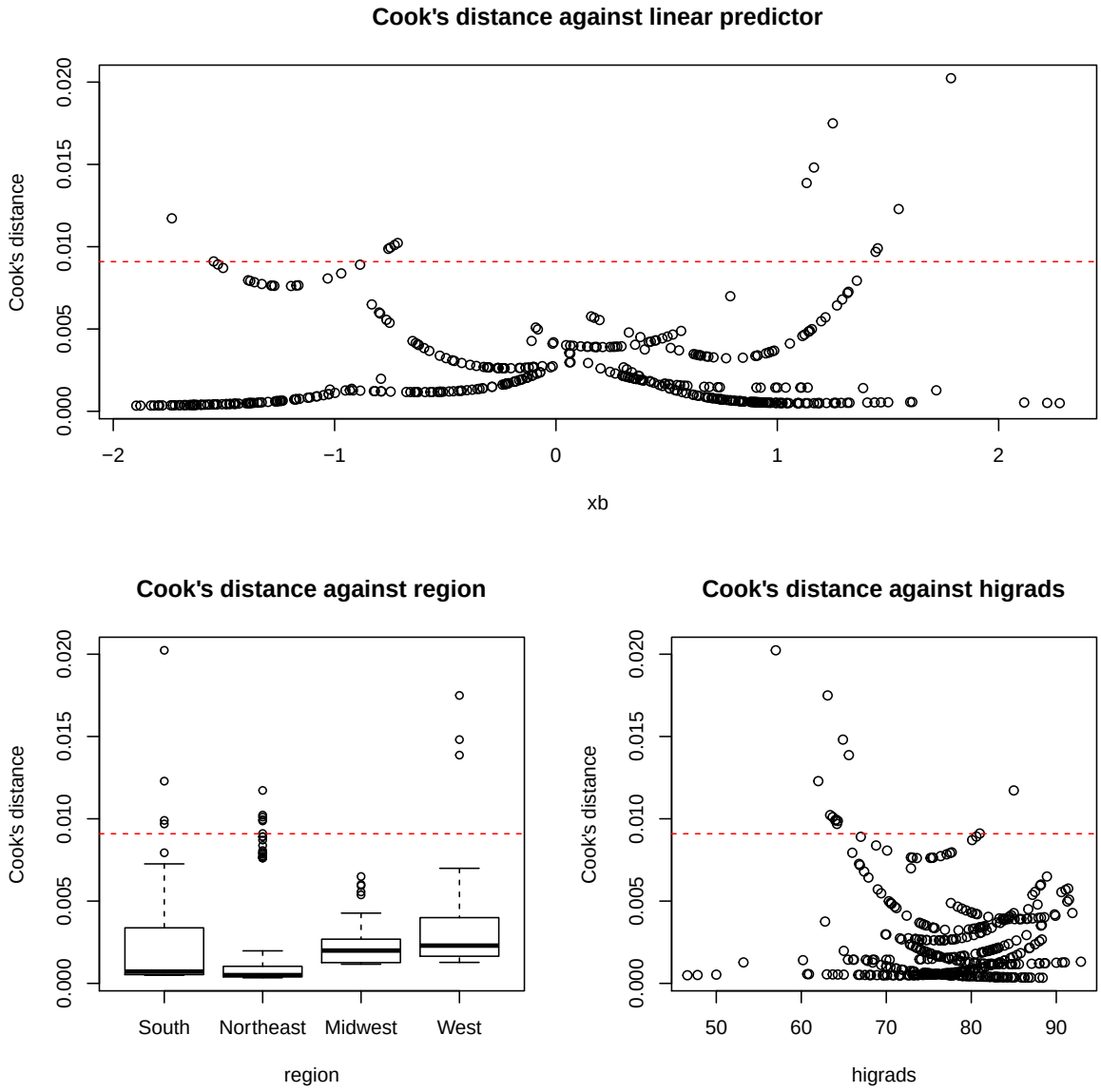

Figure 6: Cook's distance, for the combined model, against linear predictor, region as well as higrads

R `step` function is studyied.

2.5.2 Model comparison

AIC, BIC and Pseudo R^2 of the studied model are shown in figure 10. In addition, table 11 includes sensitivity and specificity for the different models.

Table 11: Comparison of sensitivity and specificity of models. Key: H = `higrads`, R = `region`, Po = `poors`, Phy = `phys1000`

Model	AIC	BIC	Sensitivity (%)	Specificity (%)	Pseudo R2
H	601	609	55	57	0.04
H + R	537	558	70	67	0.23
H + R + Po	494	518	73	75	0.34
H + R + Po + Phy	481	509	74	75	0.37
R + Po + Phy	480	504	75	74	0.37

Model	AIC	BIC	Sensitivity (%)	Specificity (%)	Pseudo R2
H + R + Phy	508	532	72	72	0.30

The results in 10 and 11 show that the `region` + `poors` + `phys1000` model performs best on most of the metrics. This result is also consistent with the `step` algorithm results. As such, this model is considered the **optimal model** for this problem.

2.5.3 Model performance

Performance of the optimal model is then analyzed by studying a QQ-plot (see figure 11) the squared standardized Pearson residuals and the standardized deviance residuals against the linear predictor x^{β} (see figure 12). As well as the Cook's distance against the linear predictor, and against `higrads` and against `region` (see figure 13).

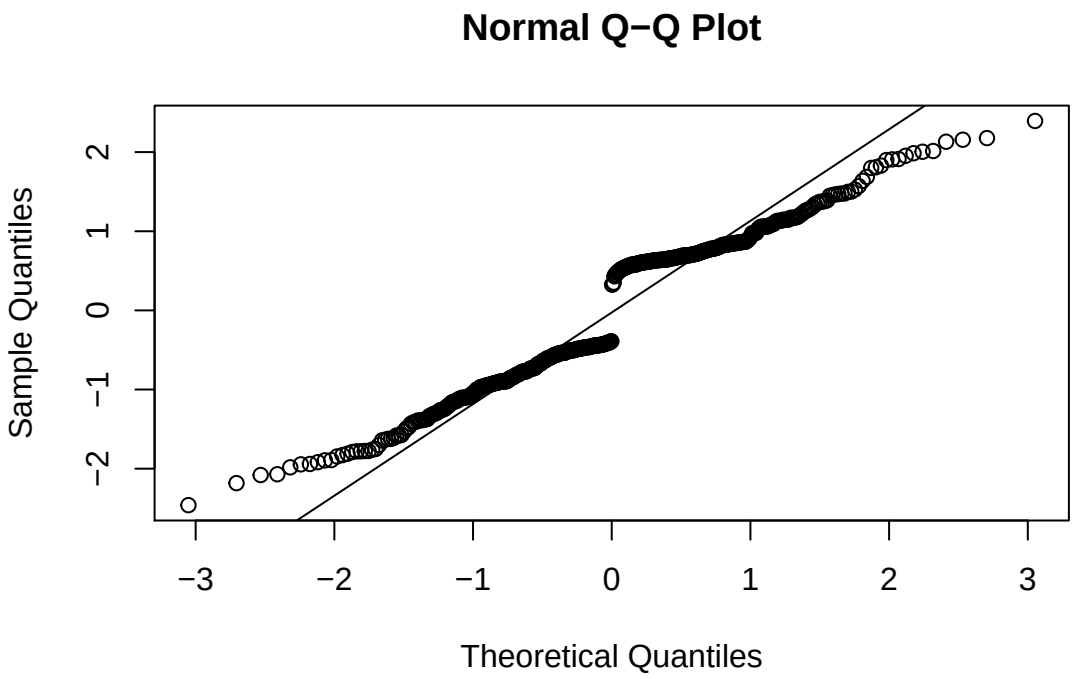

Figure 7: QQ-plot for the integration model

The outlier is Olmsted, see figure 14 for a plot of Cook's distance excluding this point. Table 12 compares the performance of these models.

Table 12: Comparison of sensitivity and specificity of optimal model v.s. optimal model with outlier Olmsted removed

Model	AIC	BIC	Sensitivity (%)	Specificity (%)	Pseudo R2
Optimal	480	504	75	74	0.37
Optimal without outlier	466	491	74	75	0.39

The model with the outlier removed performs better.

2.5.4 Discussion

In order to first get a view on the different covariates and how they relate to each other, they are plotted against eachother in figure 15.

```
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

The optimal model includes the previously studied covariate `region`, but discards the `higrads` covariate. In addition, it includes the new `poors` and `phys1000` covariates. One explaination why `higrads` is not used in the optimal may be seen in figure 15, where there is a high correlation between `higrads` and `poors`. This means that a large part of the correlation between higrads and hicrm is better the effect higrads

This hypothesis is tested in figure 2.5.4, where a linear regression model has been fit. Studying the P-value of the model reveals that the β-values are highly significant. \begin{figure}[h]

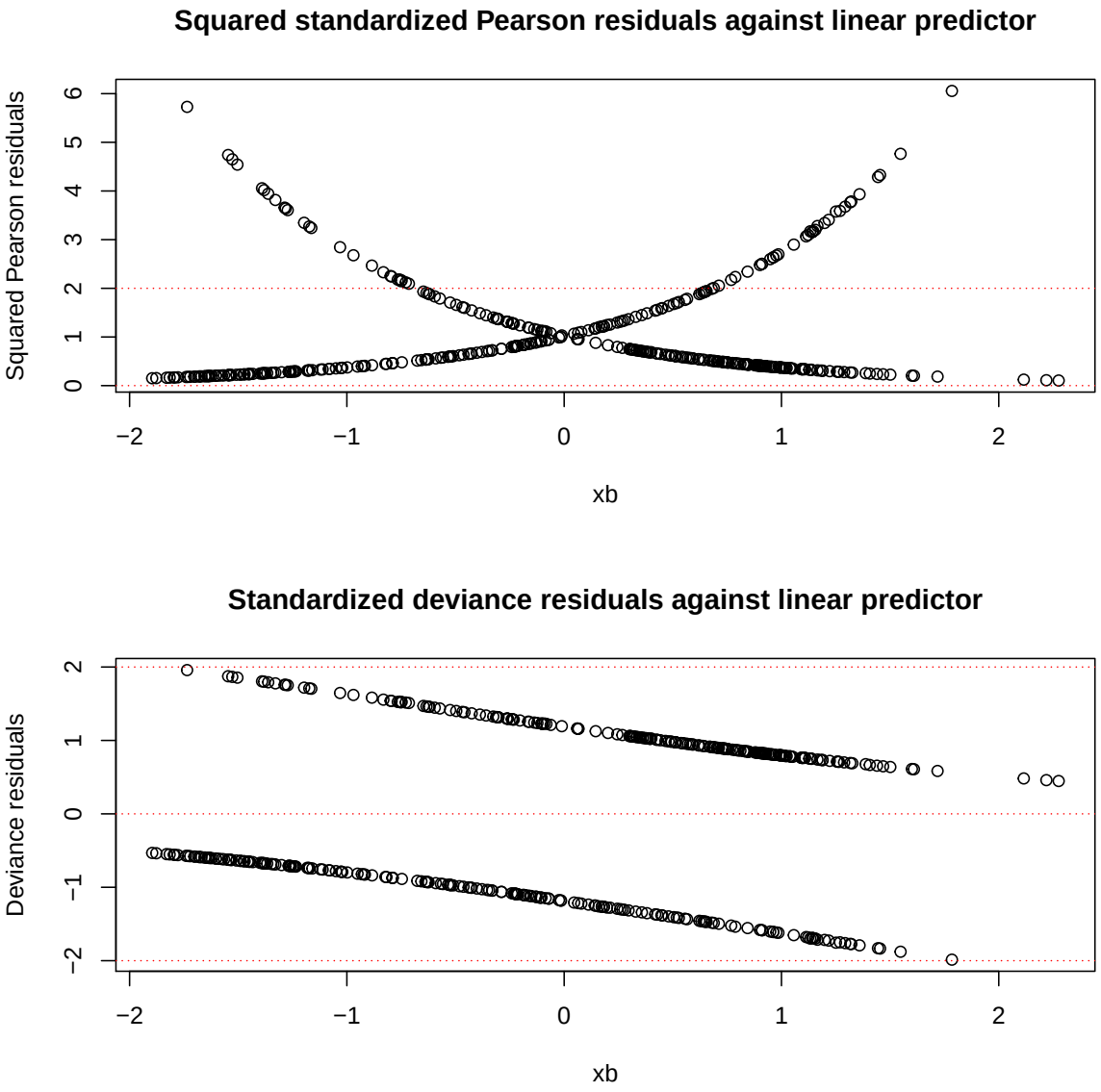

Figure 8: Squared standardized Pearson residuals as well as standardized deviance residuals for the interaction model, against the linear predictor x^β

Regarding `phys1000`, it appears in 15 that it does not have an as clear relationship to the other covariates and therefore provides more information to the model. Looking at how well `phys1000` predicts `hicrm`, seen in figure 17, it seems to follow an approximate S-shape and therefor contributes to the model.

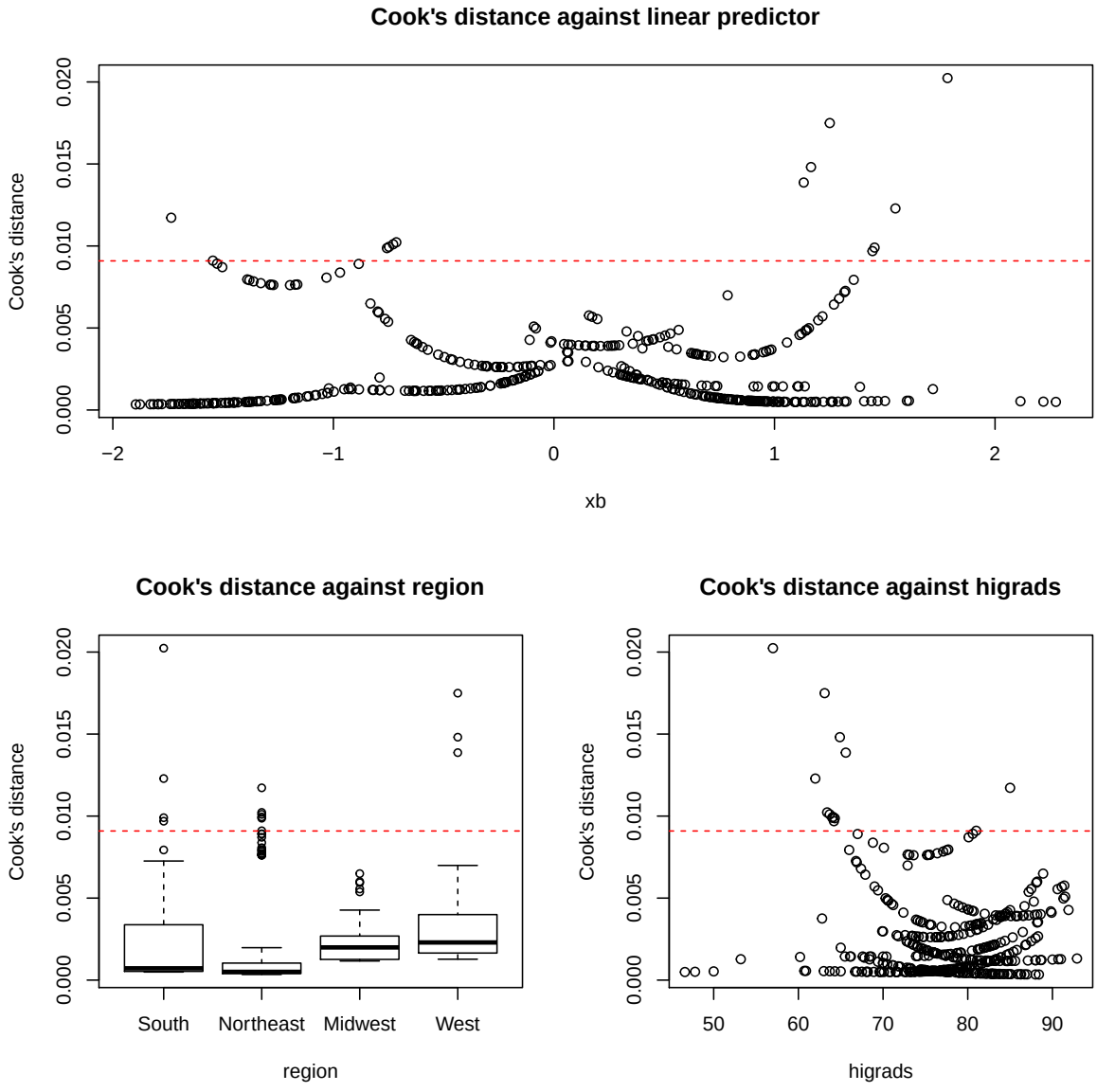

Figure 9: Cook's distance, for the interaction model, against linear predictor, region as well as higrads

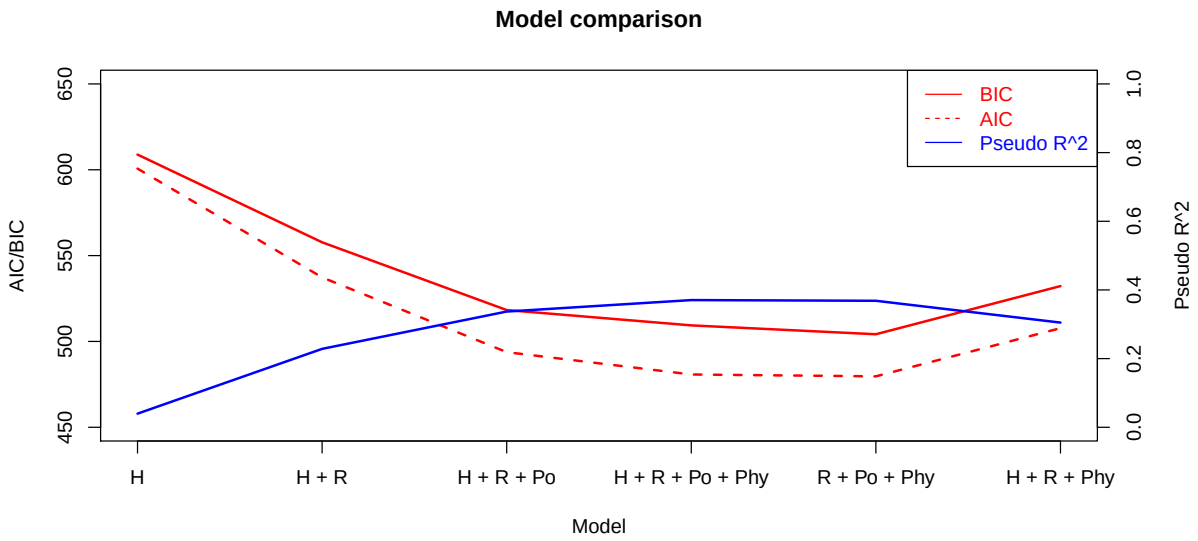

Figure 10: Comparison of AIC and BIC and Nagelkerke psuedo R^2 for the different models. Key: H = `higrads`, R = `region`, Po = `poors`, Phy = `phys1000`

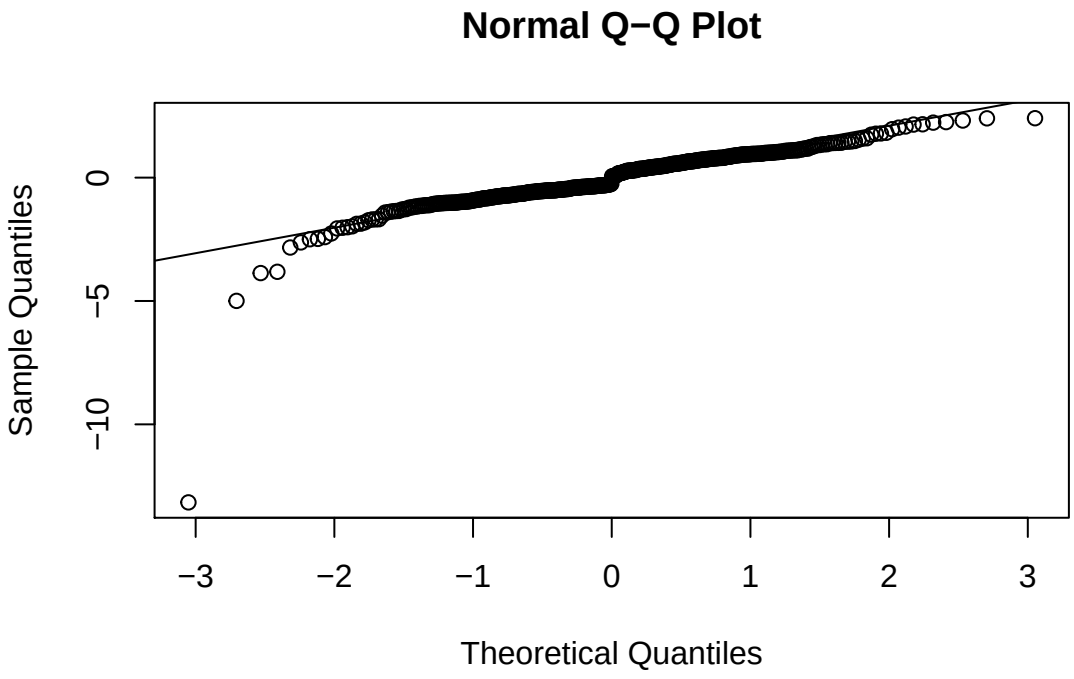

Figure 11: QQ-plot for the optimal model

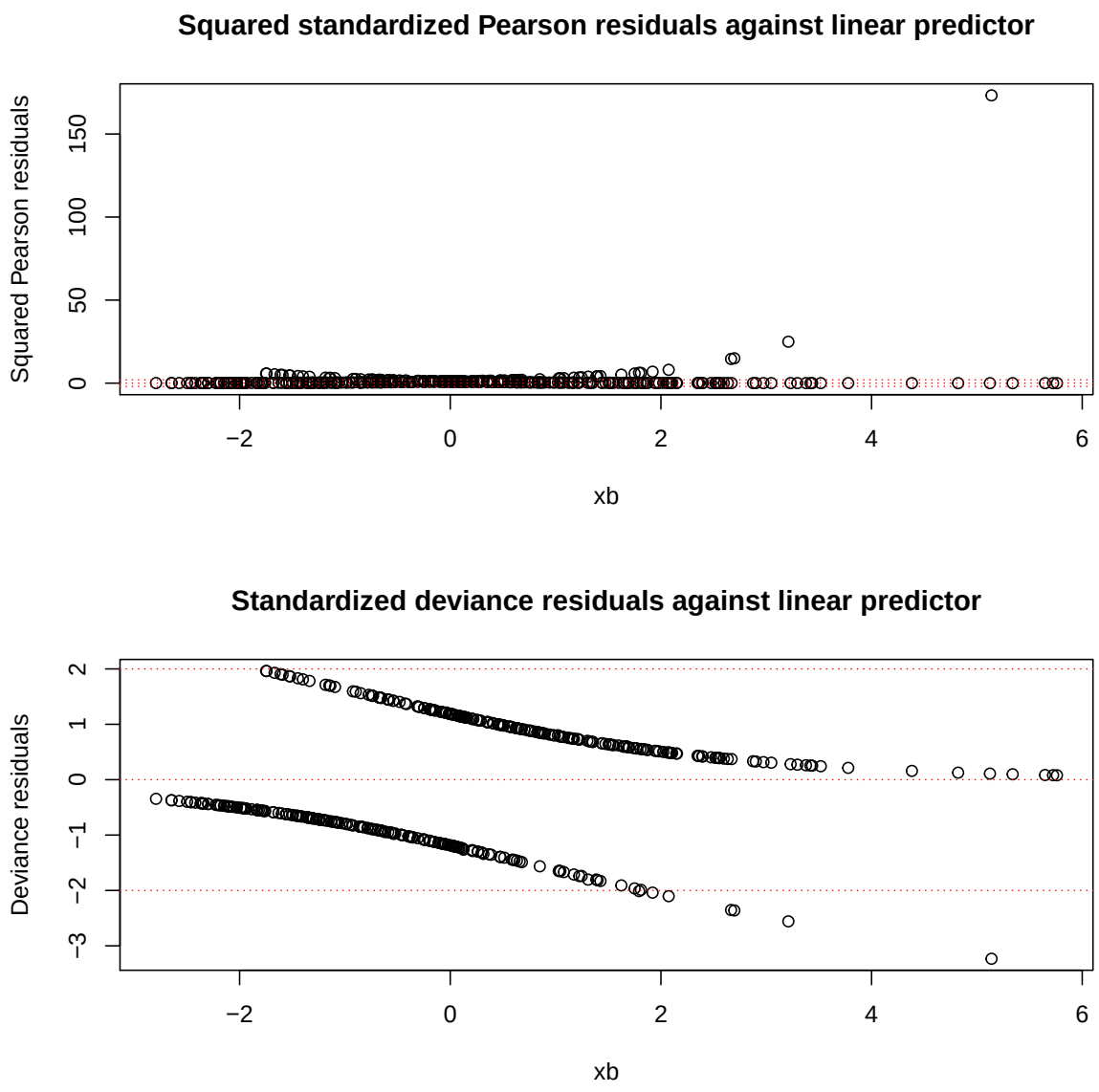

Figure 12: Squared standardized Pearson residuals as well as standardized deviance residuals for the optimal model, against the linear predictor x^{β}

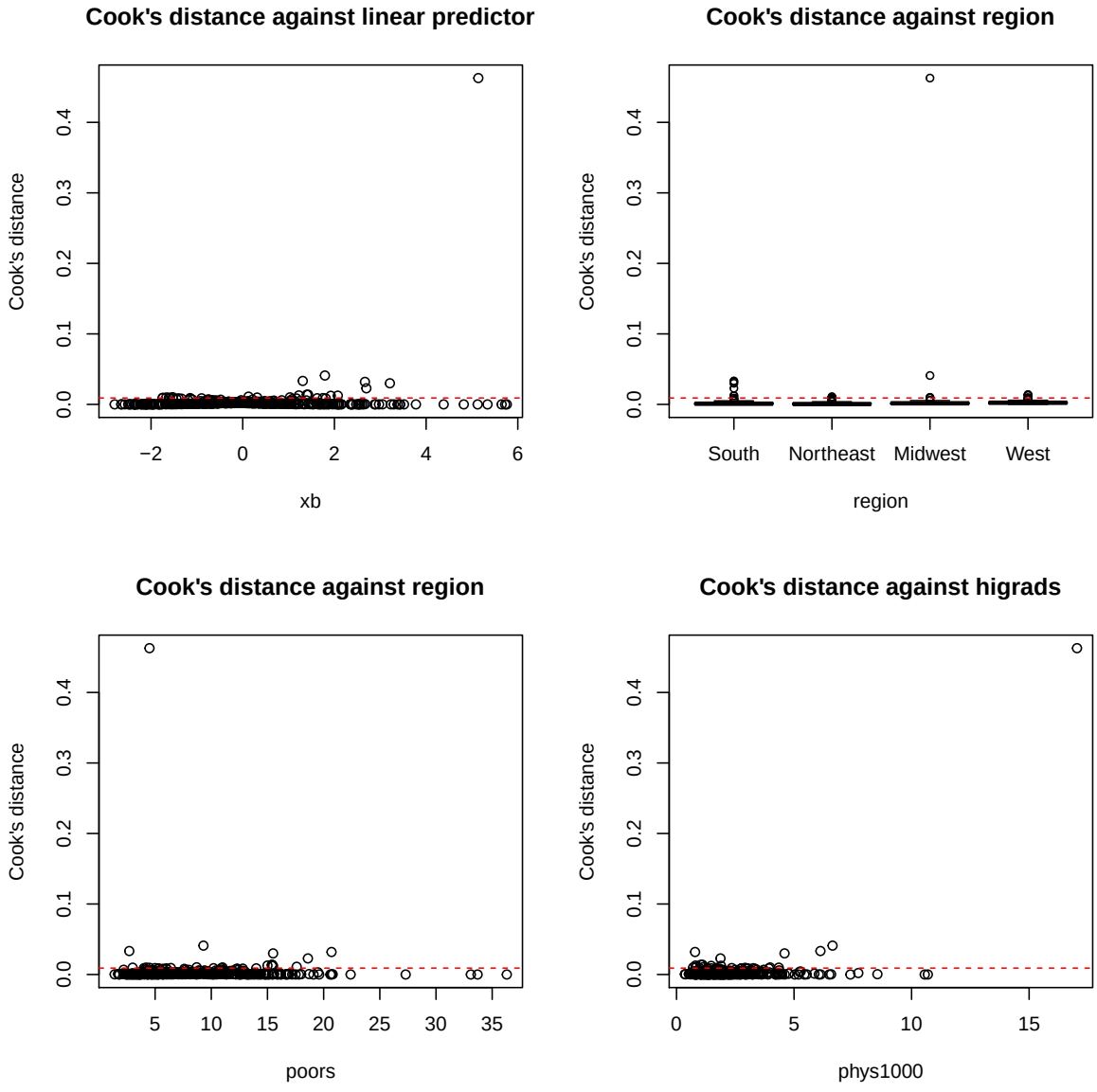

Figure 13: Cook's distance, for the optimal model, against linear predictor, region as well as higrads

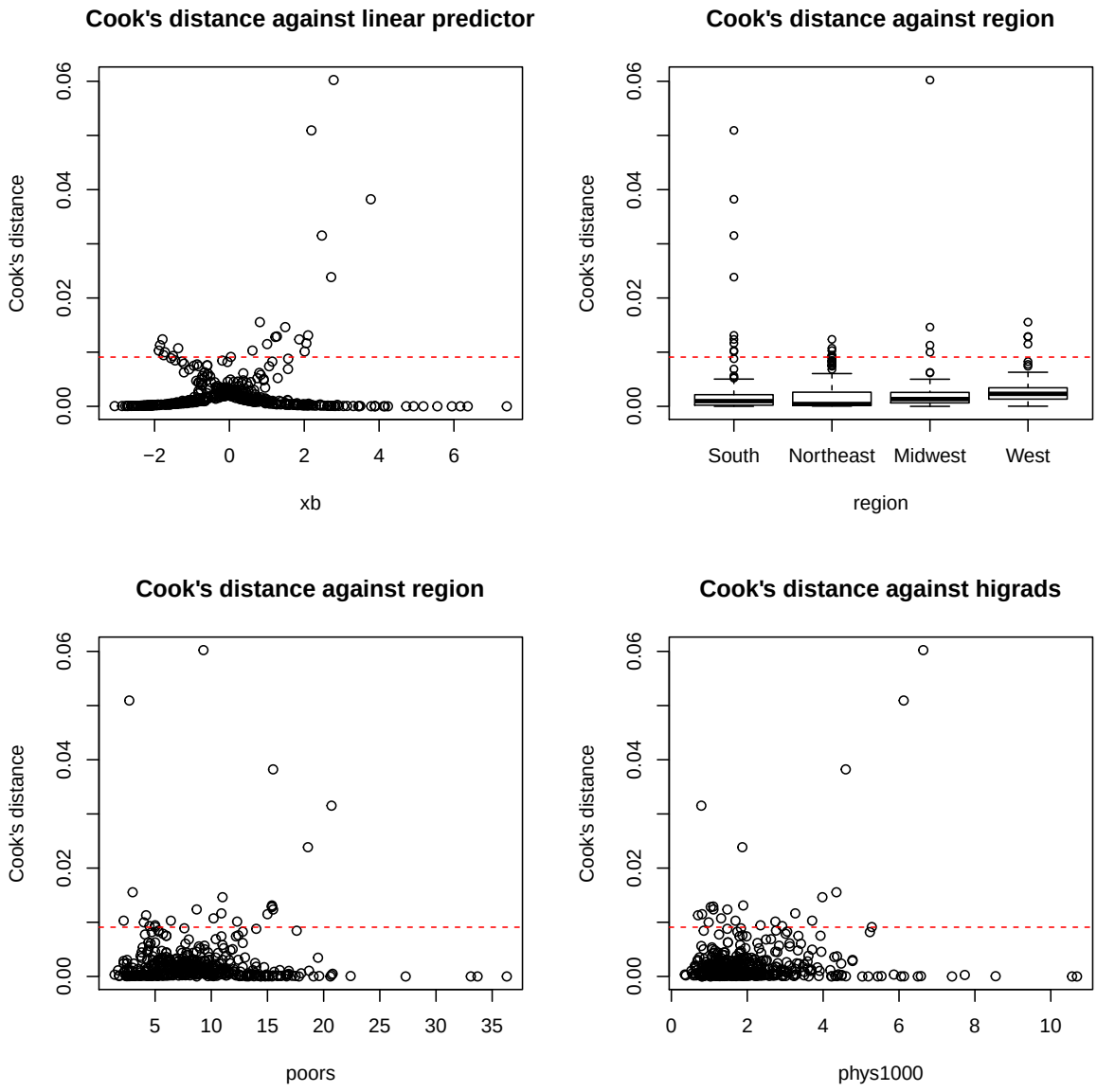

Figure 14: Cook's distance, for the optimal model, against linear predictor, region as well as higrads, excluding outlier Olmsted

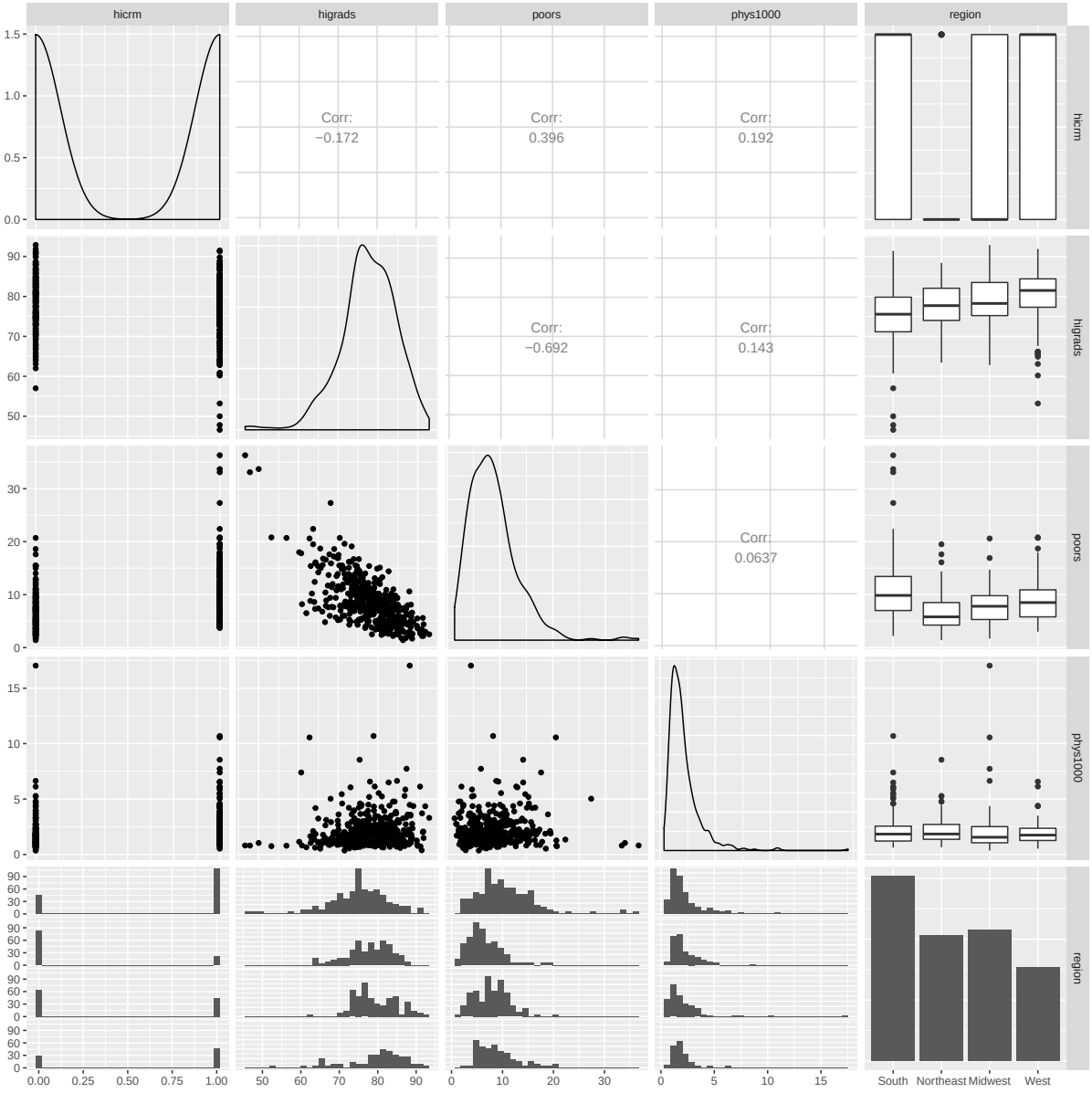

Figure 15: Plot of covariates against eachother

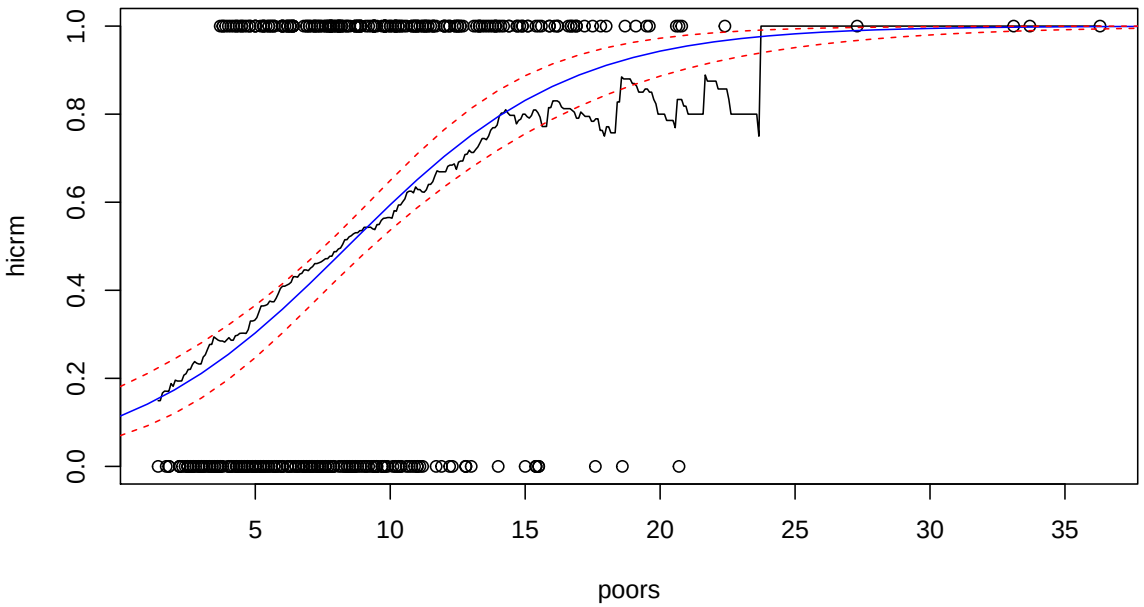

Figure 16: Plot of `hicrm` against `poors`, including kernel smoothing and prediction of fitted model with 95 % confidence interval

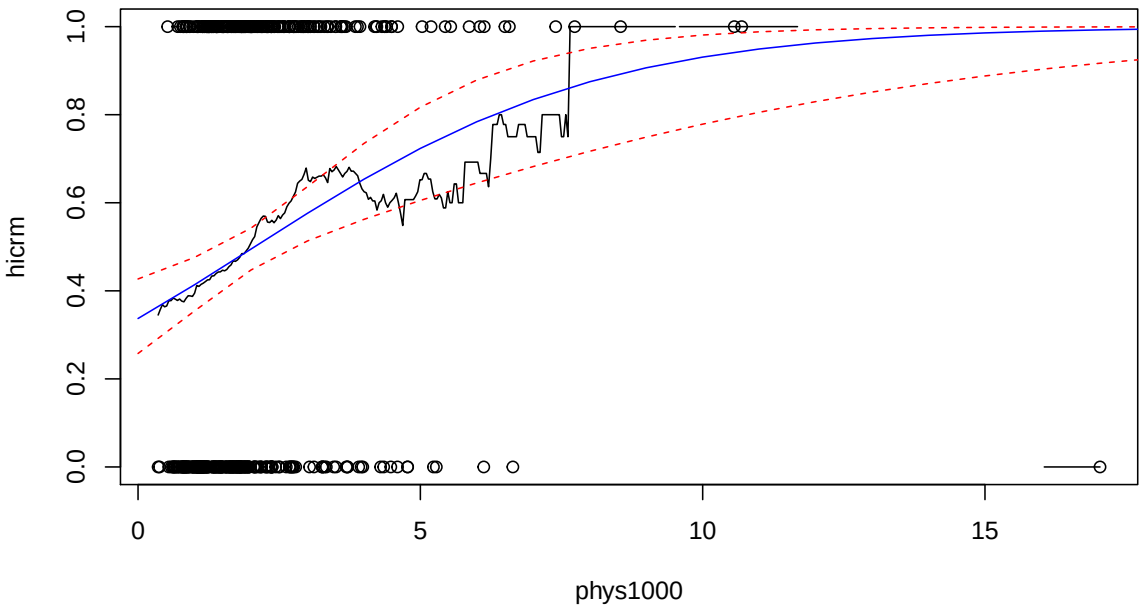

Figure 17: Plot of `hicrm` against `phys1000`, including kernel smoothing and prediction of fitted model with 95 % confidence interval