Project 2

Axel Sjöberg (ax3817sj-s) & John Rapp Farnes (jo5113fa-s)

16 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 is 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 `hicrm`, which takes the value 1 if the county was a high crime county and 0 if it was a low crime county. In this paper, `hicrm` is 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 well as errors ϵ_i. As such, the model looks like:

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

In contrast to linear regression, the error terms ϵ_i are not assumed to have any particular distribution.

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:

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

In order to determine if there is a relationship between `hicrm` and `higrads` they were plotted against each other. Because `hicrm` is a binary varible it was very difficult to determine if a relationship exists only by the looking at 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 were included. Figure 1 shows the attained plot described above.

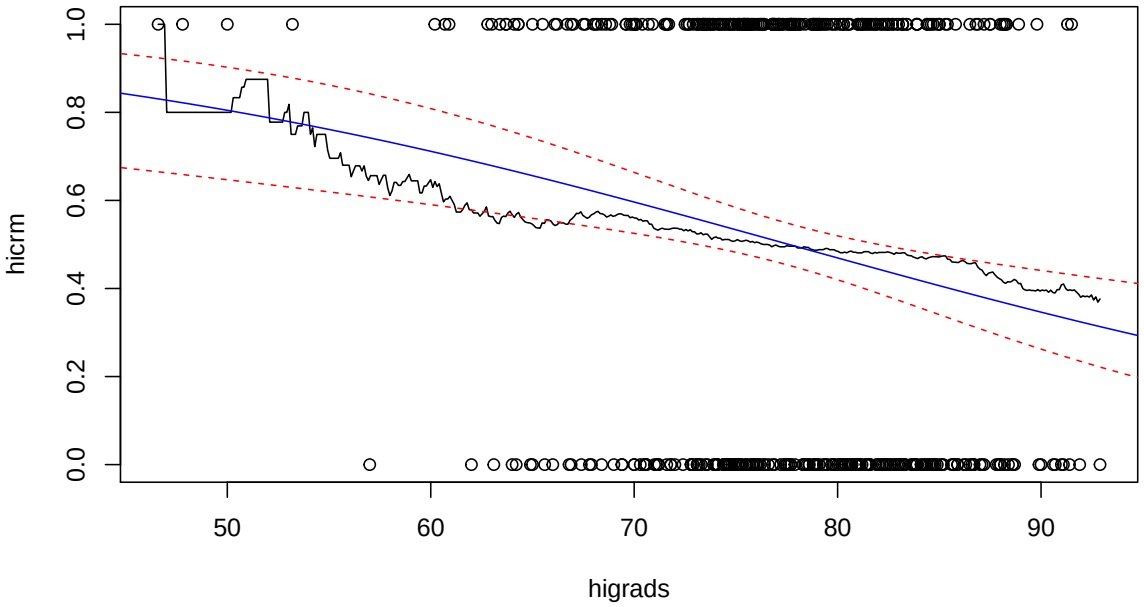

Figure 1: Plot of `hicrm` against `higrads`, including kernel smoothing and prediction of fitted model with 95 % confidence interval

As seen in figure 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 $\beta_{higrads}$, meaning 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 30% - 70%, implying that `higrads` alone 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 2, all of the P-values are < 0.05, meaning both that conties with no `higrads` has greater than 0 probability of having a high crime rate, and 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 a 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) were predicted. The result is presented 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

Looking at table 3, one can see that a county with 65 % high education graduates has greater than 50% probability of having a high crime rate, while a county with 85 % high education graduates has less than 50% probability of having a high crime rate. Further, these results are also statistically significant.

2.1.4 Model performance analysis

In order to analyze model performance, the sensitivity and specificity of the model were 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%, indicating 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 was modelled using "dummy variables" X_i. In order to implement this effectively, one of the categories is chosen as a reference category, and the effects of the other categories are measured in comparison to it.

In order to determine which region to use as reference category, a cross-tabulation of the data between `region` and `hirm` had to be 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 resulted in choosing South as the reference region.

Using this reference region, the logistic model became

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

Here, the β coefficients are measured relative to South, while β_0 was the 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-value for β_{West} was > 0.05, indicating a lack of statistical significance in the difference between how South and West predicts `hicrm`.

Next, the odds-ratios for the different categories were 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 are presented 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 were 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 is however not statistically significant as the confidence interval for the western region's odds ratio, in relation to the refernce category, cover 1. This coincides with the findings in table 5 where the β_{West} P-value was $>$ 0.05.

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` model, 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 used both `higrads` and `region` was analyzed, i.e. the following model:

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

2.3.2 Model comparison

In order to compare the models, metrics other than sensitivity and specificity was studied. These were 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.

A comparison of the models with regards to AIC, BIC and Psuedo R^2 can be seen in figure 2, while a comparison of sensitivity and specificity is found in table 9.

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

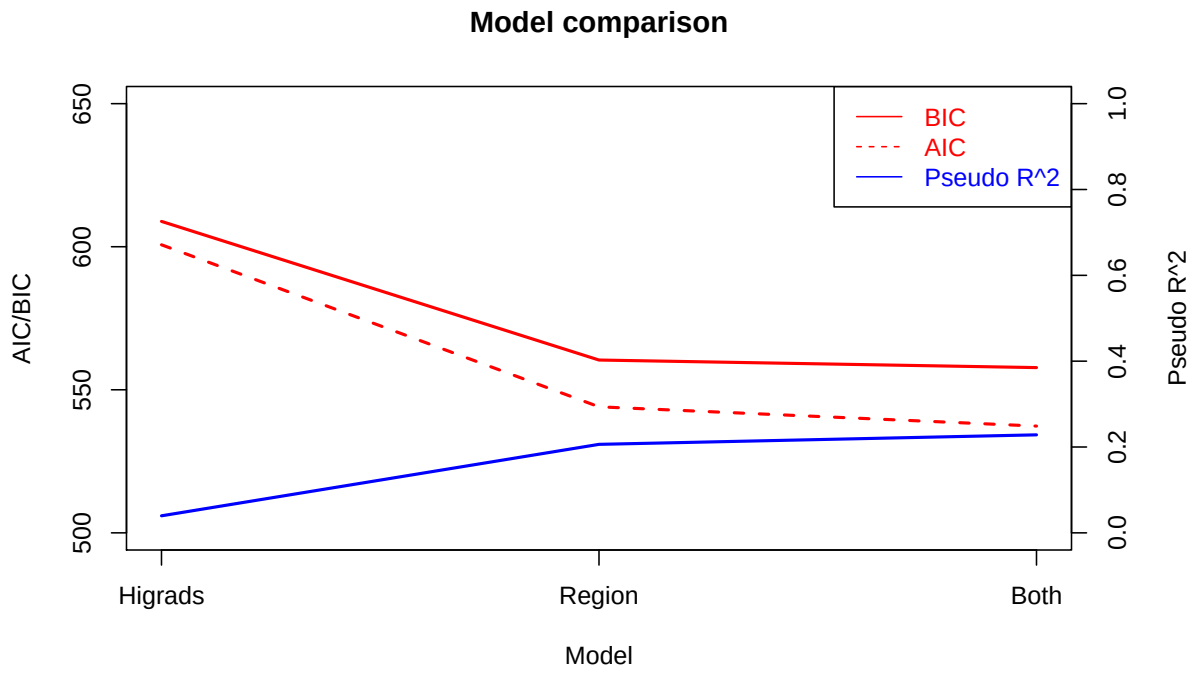

Figure 2: Comparison of AIC and BIC and Nagelkerke psuedo R^2 for the different models

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

2.3.3 Combined model analysis

The combined model was further analyzed by studying the QQ-plot (see figure 3), the squared standardized Pearson residuals and the standardized deviance residuals against the linear predictor $X\beta$ (see figure 4). Furthermore the Cook's distance was plotted against the linear predictor, `higrads` and against `region` was plotted (see figure 5).

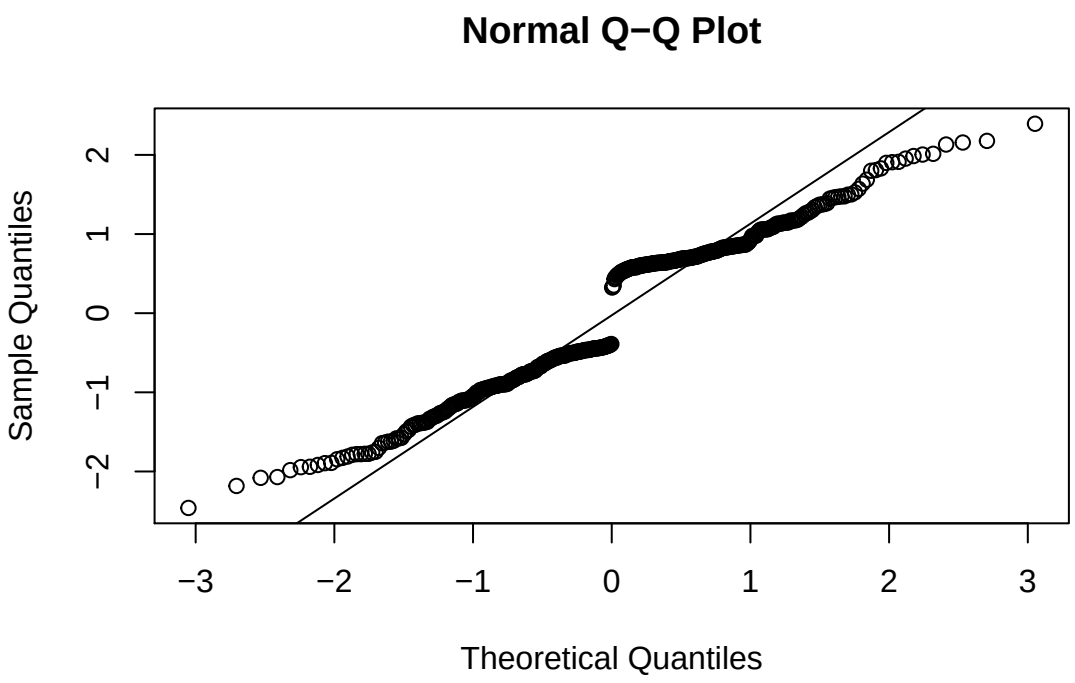

Figure 3: QQ-plot for the combined model

From the QQ-plot it looks like the model residuals follow a bimodal distribution rather than a normal distribution.

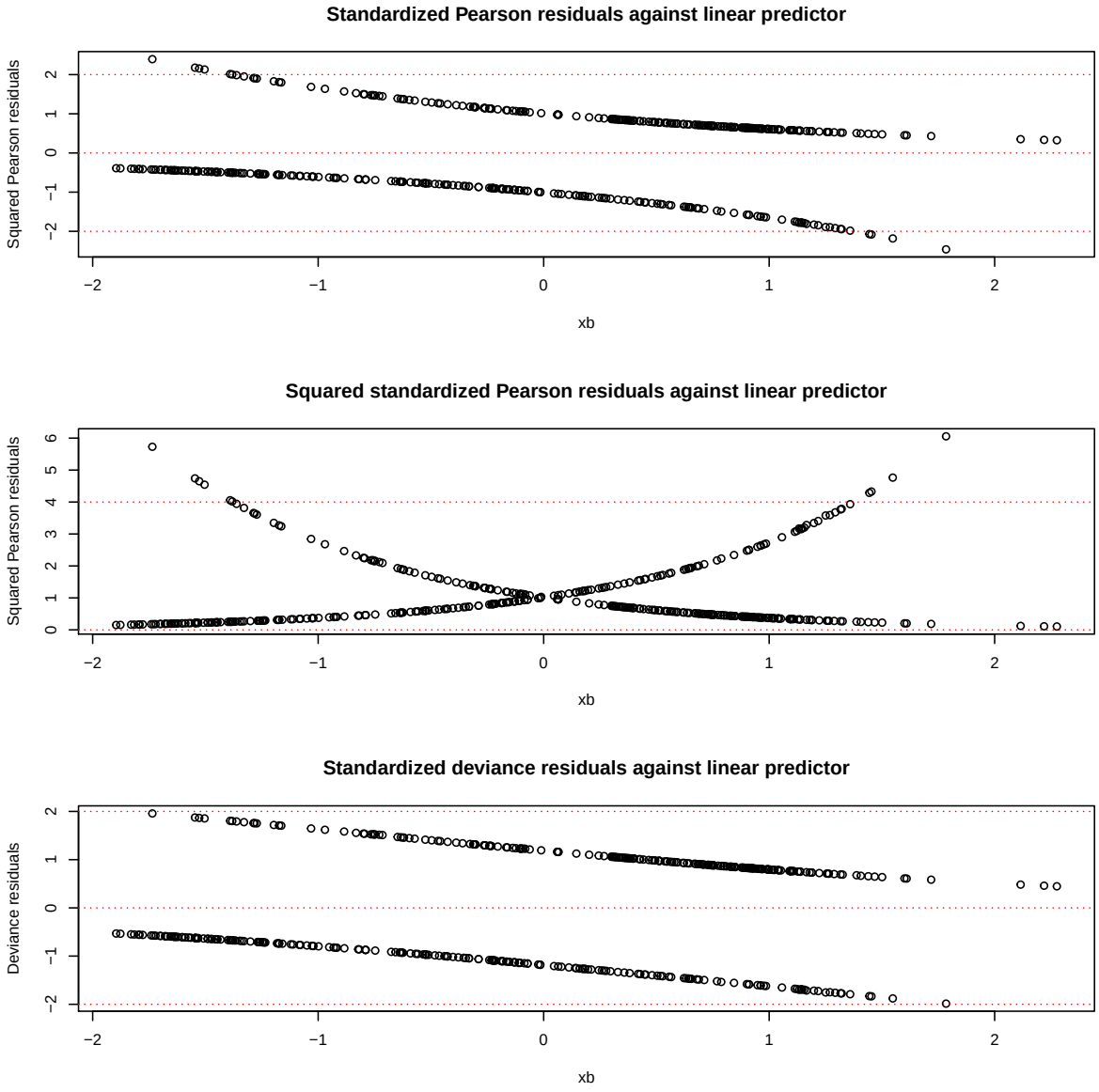

Figure 4: Standardized Pearson residuals as well as standardized deviance residuals for the combined model, against the linear predictor $X\beta$

For a standardized pearson residual to be considered suspiciously large, $|r_i| > |\lambda_{\alpha/2}| \approx 2$. This was true for 8 of the standadized pearson residuals, which can be seen in figure 4, where the standardized pearson residuals were plotted against the linear predictor. For a deviance residual to be considered too large $|r_i| > 2$. This was never the case which is verified by figure 4 where the standardized devience residuals were 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, see figure 5

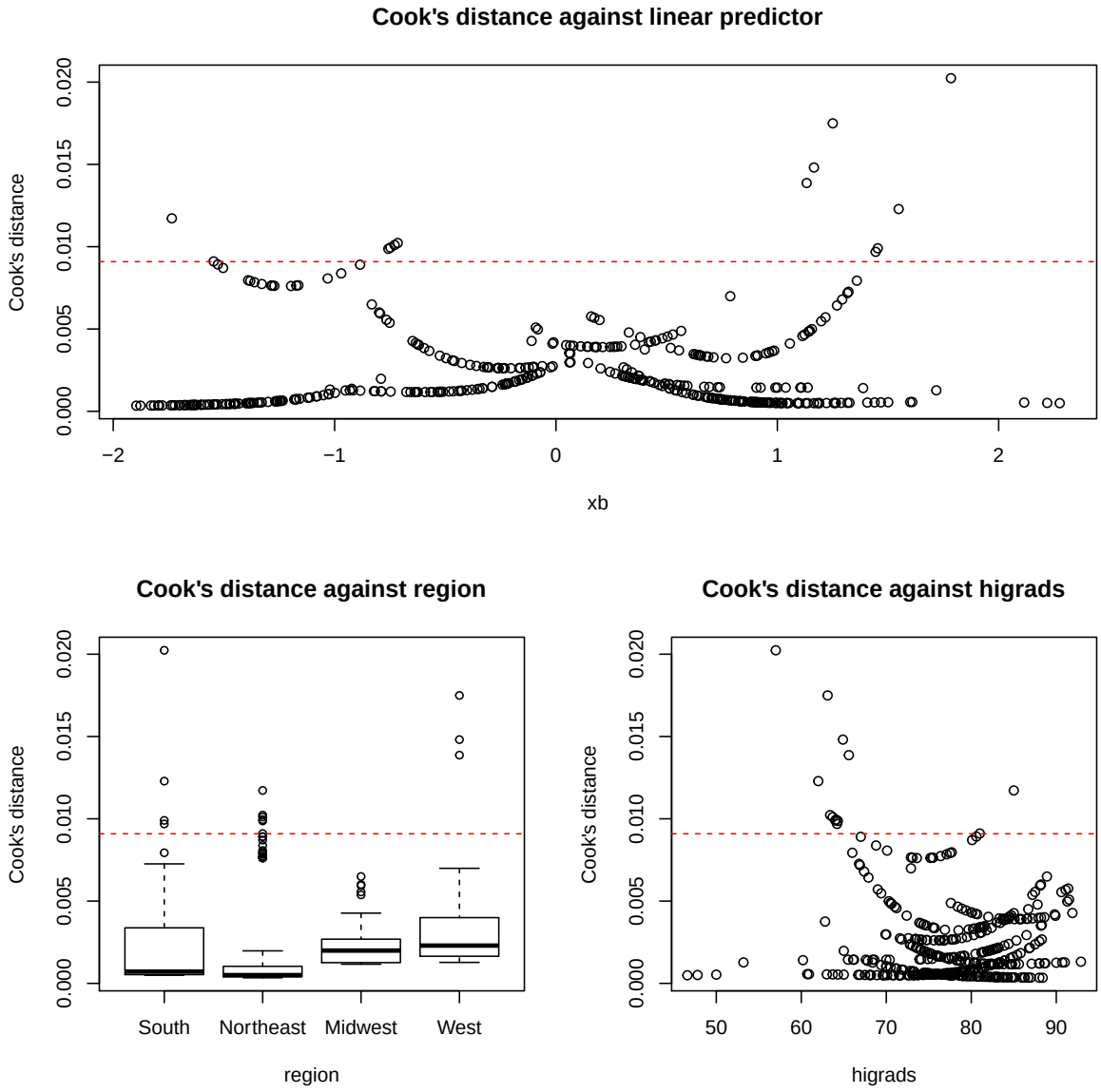

Figure 5: Cook's distance, for the combined model, against linear predictor, region as well as higrads

As can be seen in figure 5 the vast majoriy of the the observations are below the horizontal line. The amount of counties with a high Cook's distance are 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 are found in the West region. The Cook's distance plots give no concern to why the combined model should be amiss.

2.4 Interaction model

2.4.1 Introduction

As a forth model, interaction terms were 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 analysis

The performance of the interaction model compared to the combined model was 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 resulted in the interaction model being significantly better than the combined model, with a P-value of 0.019.

In addition, the AIC, BIC, Nagelkerke, sensitivity and specificity was compared to the combined model, in table 10. The interaction model was analyzed by studying a QQ-plot (see figure 6) the squared standardized Pearson residuals and the standardized deviance residuals against the linear predictor X^{β} (see figure 7). Furthermore the Cook's distance was plotted against the linear predictor, `higrads` and against `region` (see figure 8).

Table 10: Comparison of sensitivity and specificity of models

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

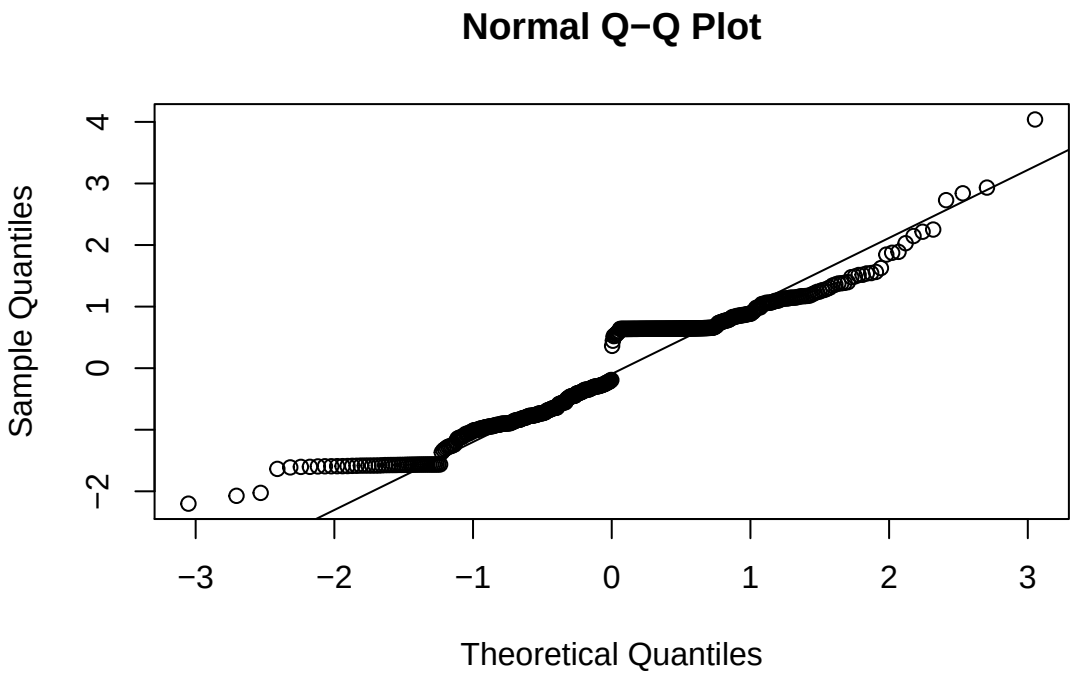

Figure 6: QQ-plot for the integration model

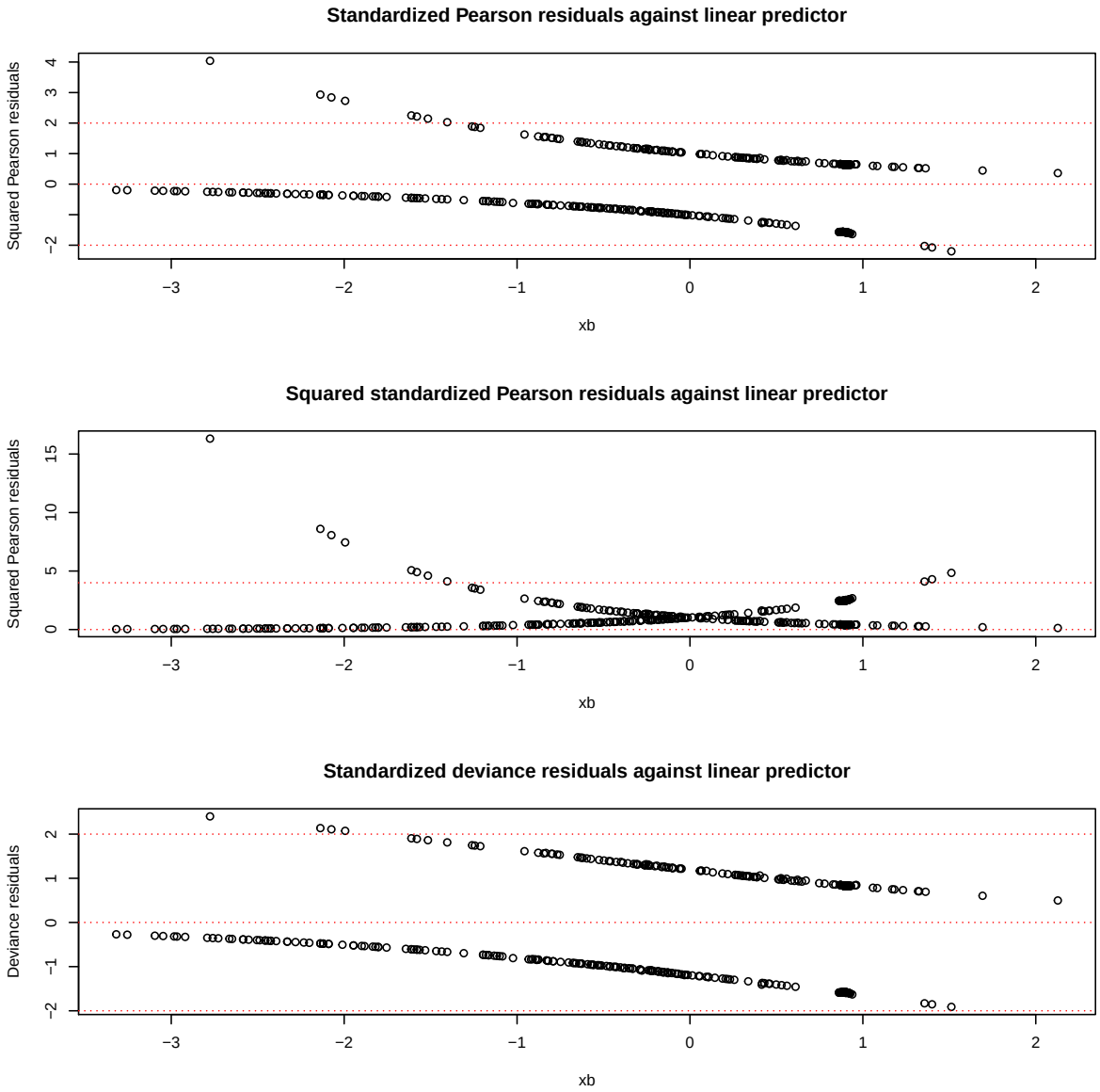

Figure 7: Standardized Pearson residuals as well as standardized deviance residuals for the interaction model, against the linear predictor $X\beta$

The interaction had more Person residuals that may be considered suspisiously large than the combined model. Unlike the combined model it also had standardized devience residuals that are too large.

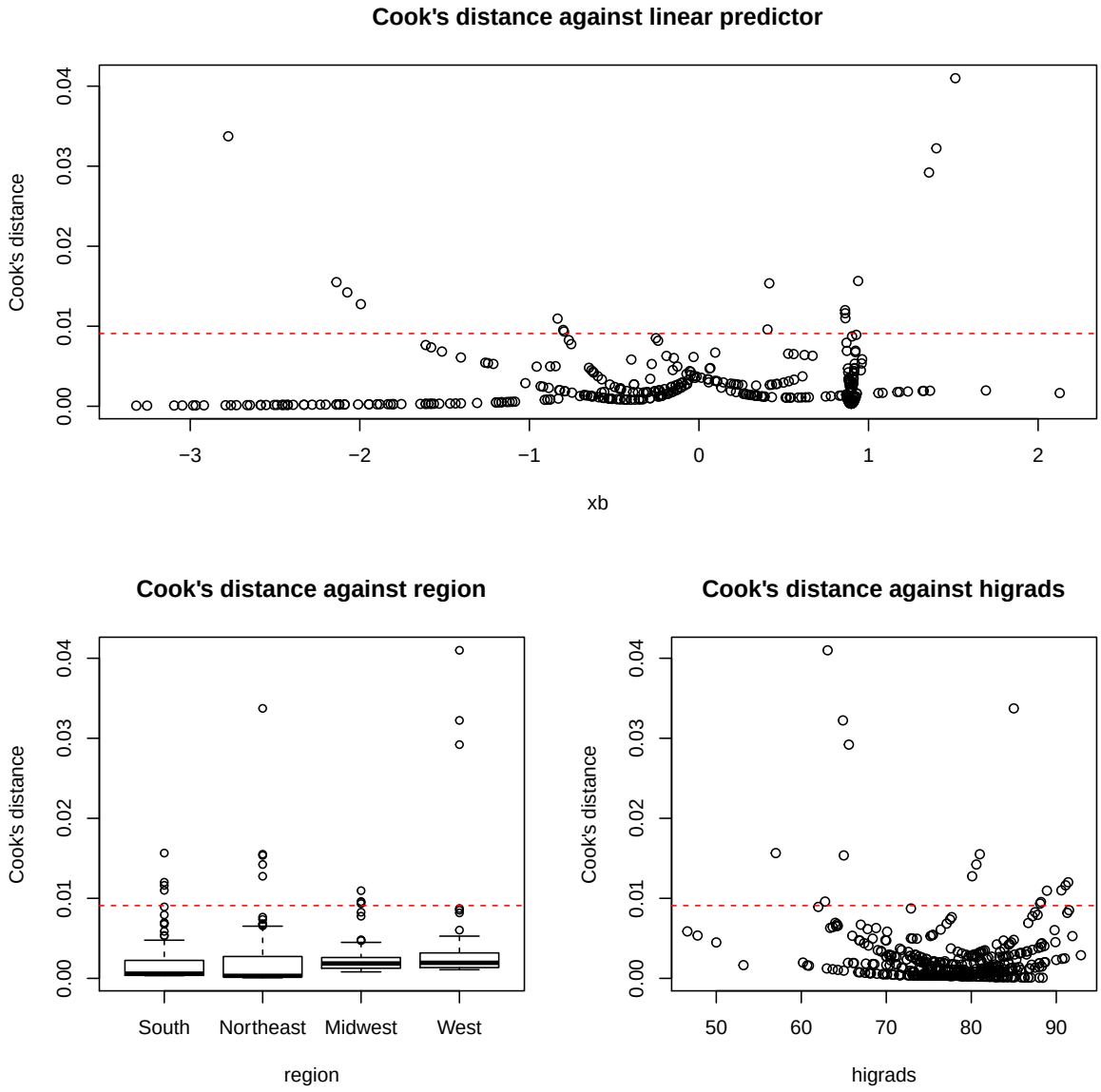

Figure 8: Cook's distance, for the interaction model, against linear predictor, region as well as higrads

Both the interaction and the combined model performed better on some metrics, while performing worse on other. The interaction model had a lower AIC value but at the same time higher BIC-value. This is expected since BIC penalizes larger models more than the AIC value tends to do. The sensitivity, specificity and Pseudo R^2 values were worse for the Combined model. The interaction model performed considerably worse than the combined model on the Cook's distance acount. The interaction model had far more outliers and outliers with a much higher Cook's distance.

As aforementioned, the interaction model had some credibility issues with its residuals. Moreover the interaction model had far more outliers as well as considerably larger Cook's distance than the combined model. At the same time the interaction model outperformed the combined model on likelyhood ratio test, AIC, Sensitivity, Specificity and Pseudo R2. The outliers in the interaction model might therefor not effect the Interaction model to much. In the end, the interaction model is favoured over the combined model.

2.5 Finding the optimal model

2.5.1 Methology

Next, an attempt to fit an optimal model to predict high crime rates was made, using the previous covariates, as well as `poors` and `pshys1000`. When searching for an optimal model, interaction terms were disregarded.

Models of increasing complexity were constructed by adding more covariates. These were then 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 R `step` function was studied.

2.5.2 Model comparison

AIC, BIC and Pseudo R^2 of the studied model are shown in figure 9. In addition, table 11 includes sensitivity and specificity for the different models.

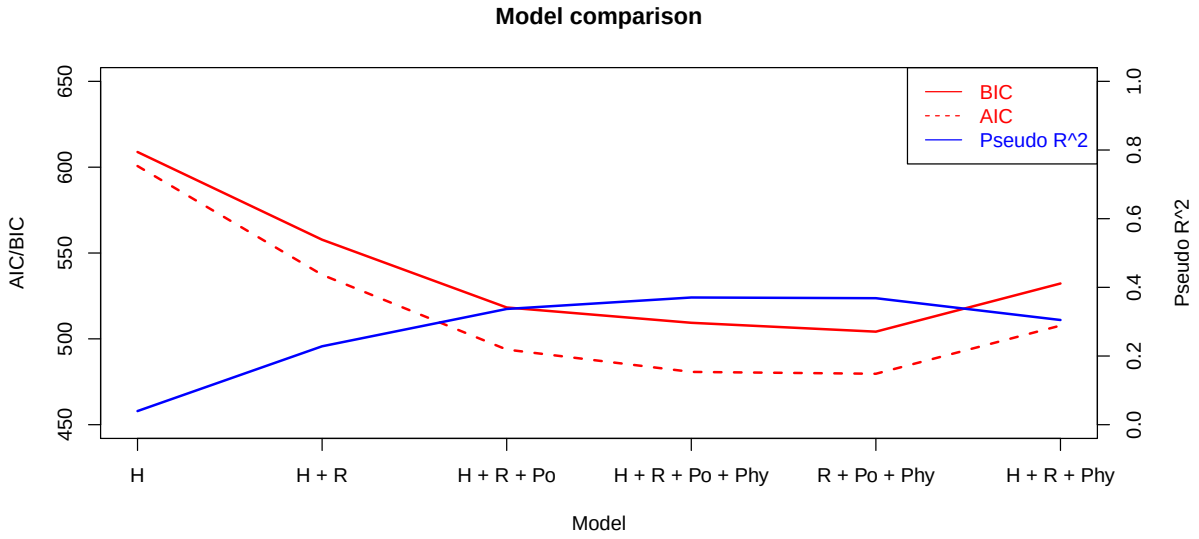

Figure 9: Comparison of AIC and BIC and Nagelkerke psuedo R^2 for the different models. Key: H = `higrads`, R = `region`, Po = `poors`, Phy = `phys1000`

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
H + R + Phy	508	532	72	72	0.30

The results in 9 and 11 show that the `region` + `poors` + `phys1000` model performed the best on most of the metrics. This result is also consistent with the `step` algorithm results. As such, this model was considered the **optimal model** for this problem.

2.5.3 Optimal model analysis

The optimal model was then analyzed by studying a QQ-plot (see figure 10) the squared standardized Pearson residuals and the standardized deviance residuals against the linear predictor $X\beta$ (see figure 11). Furthermore the Cook's distance was plotted against the linear predictor, `higrads` and against `region` (see figure 12).

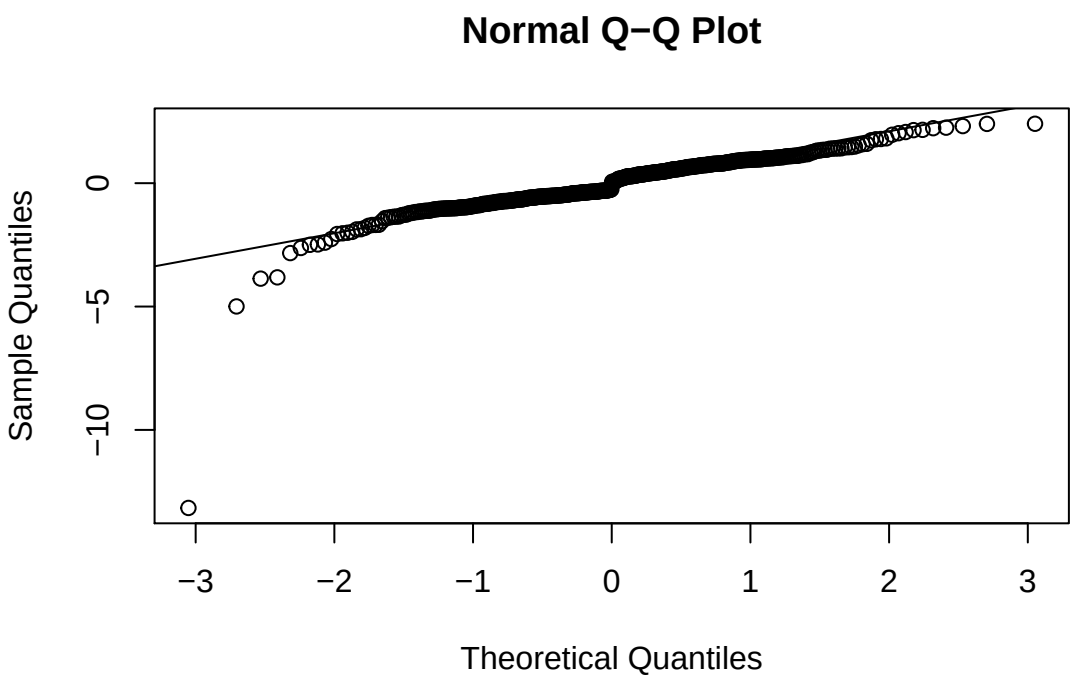

Figure 10: QQ-plot for the optimal model

The QQ plot of the optimal model appear to follow a normal distribution, however skewed to the left.

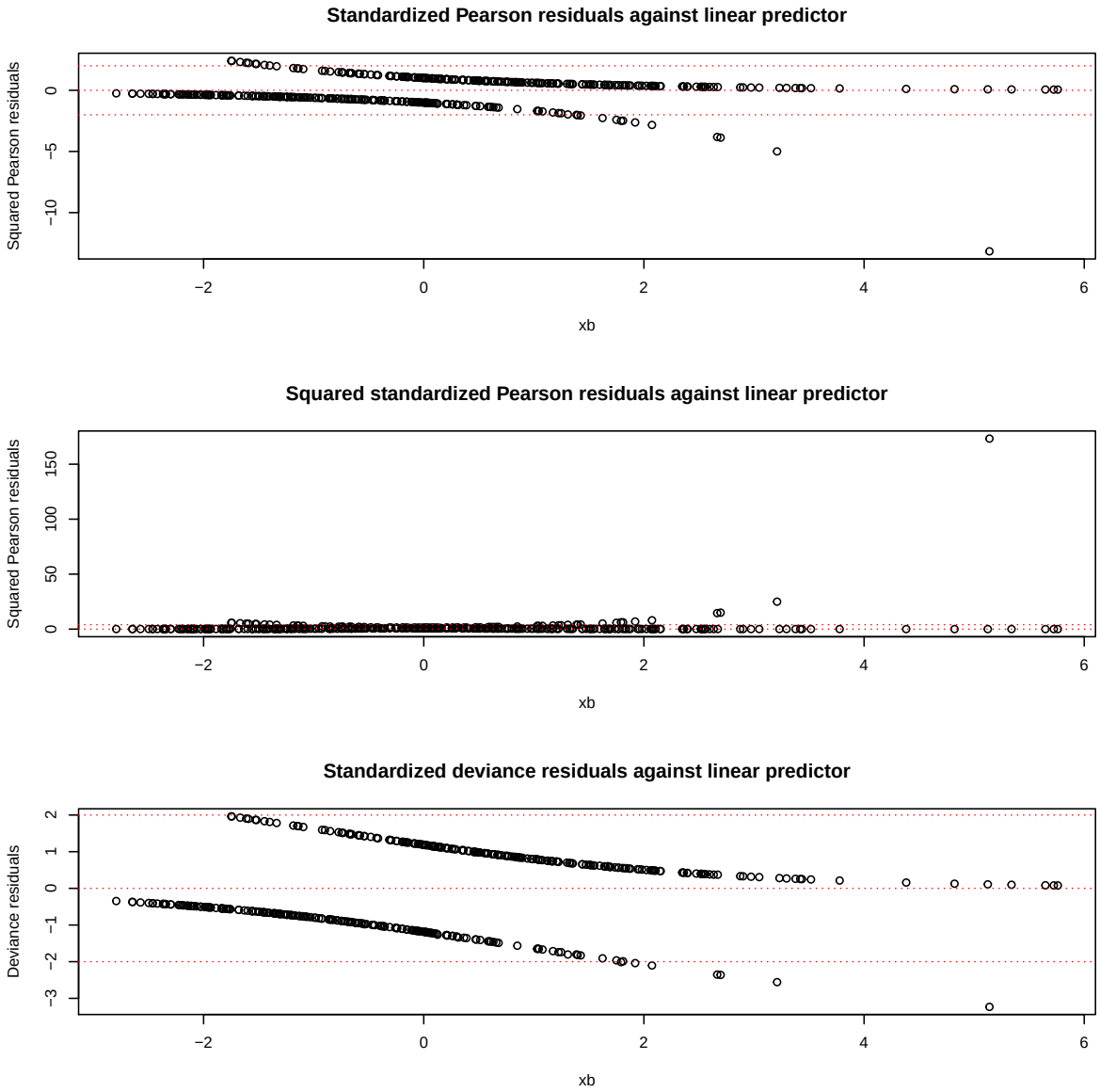

Figure 11: Standardized Pearson residuals as well as standardized deviance residuals for the optimal model, against the linear predictor $X\beta$

There was one particular outlier that seem to generate an exceptionaly large Pearson residual. Furthermore there were a few Pearson Resdiuals that had a suspisiously large value. As for the devience residuals, some of them were too large as well. Disregarding the most extreme outlier, the residuals were not far worse off than the residuals of the interaction model. The most extreme outlier was the Olmsted county. The potential influence of individual observations was adressed by plotting the Cook's disatance, see figure 12.

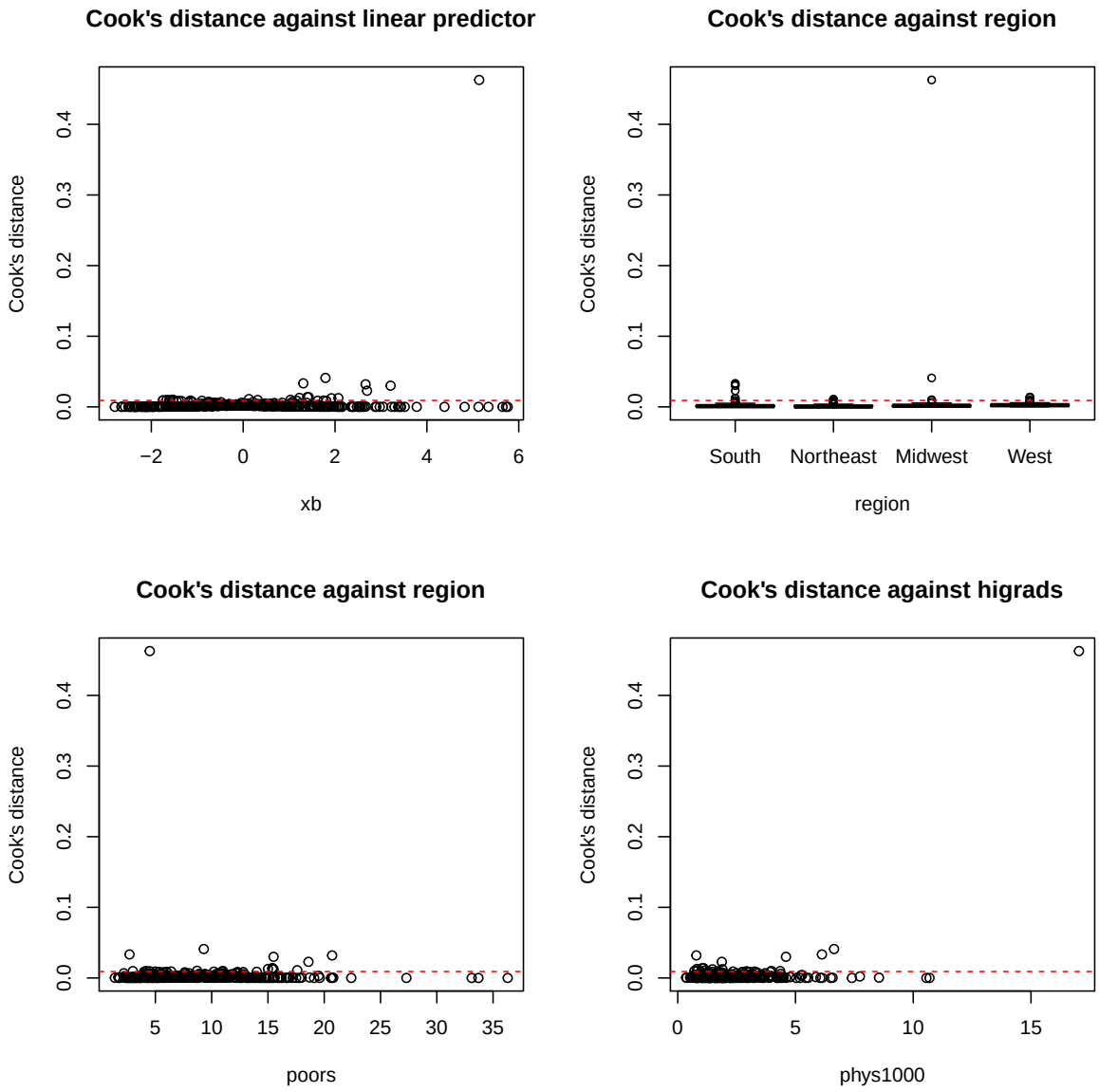

Figure 12: Cook's distance, for the optimal model, against linear predictor, region as well as higrads

The outlier in figure 12 was again Olmsted, see figure 13 for plots of Cook's distance excluding this point. Table 12 compares the performance of these models.

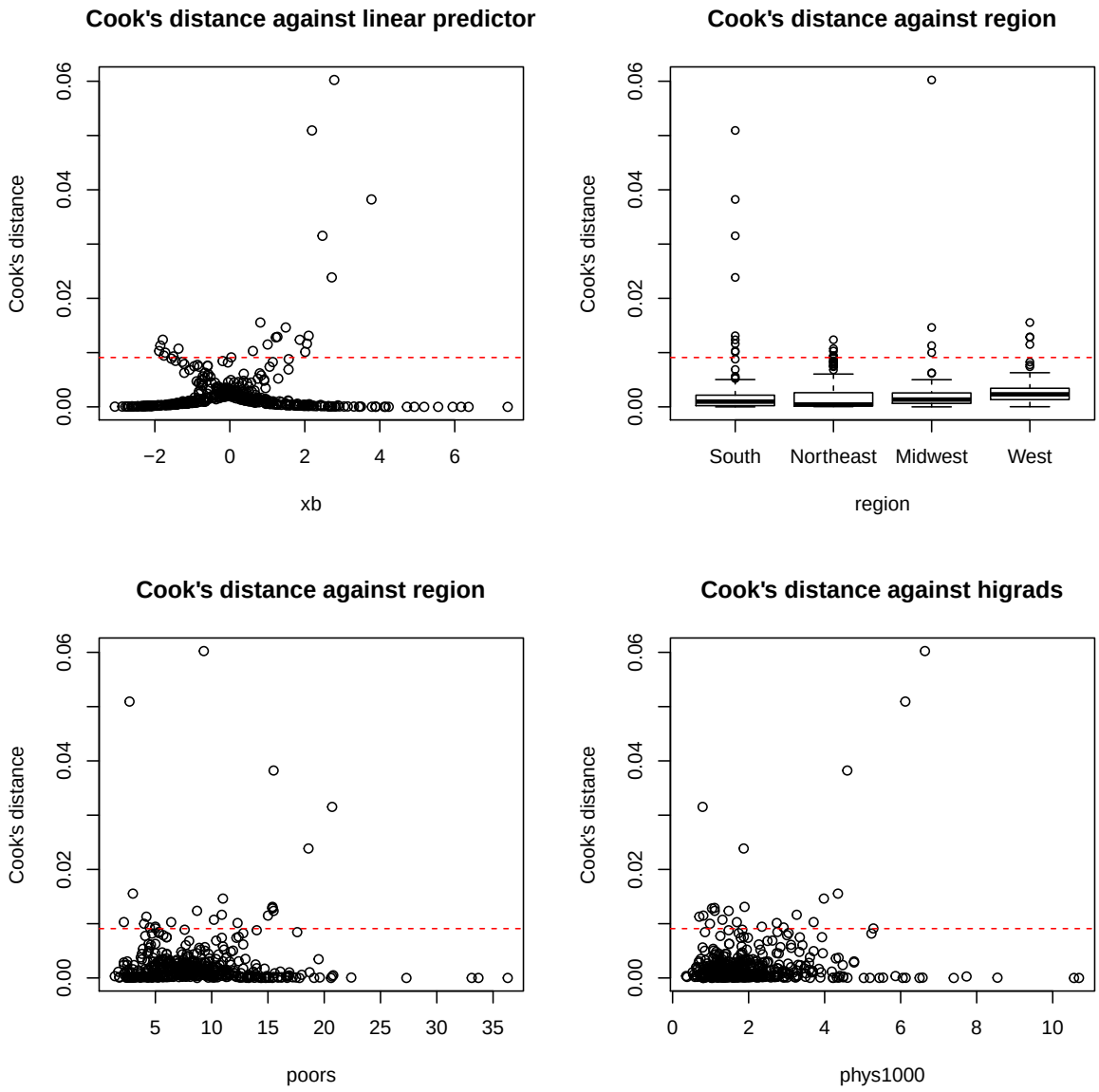

Figure 13: Cook's distance, for the optimal model, against linear predictor, region as well as higrads, excluding outlier Olmsted

Removing the problematic data entry from the CDI data frame, and refitting the optimal model resulted in a substantial improvement in the the Cook's distance plots. Comparing the Cook's distsance plots and resiudal plots without the problematic data entry generated relatively simlar plots as the figure 8. The main difference was the largest two outliers in the figure 12.

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 excluding outlier	466	491	74	75	0.39

The model with the outlier removed performed even better.

2.5.4 Discussion

In order to first get a view on the different covariates and how they relate to each other, they were plotted against eachother in figure 14.

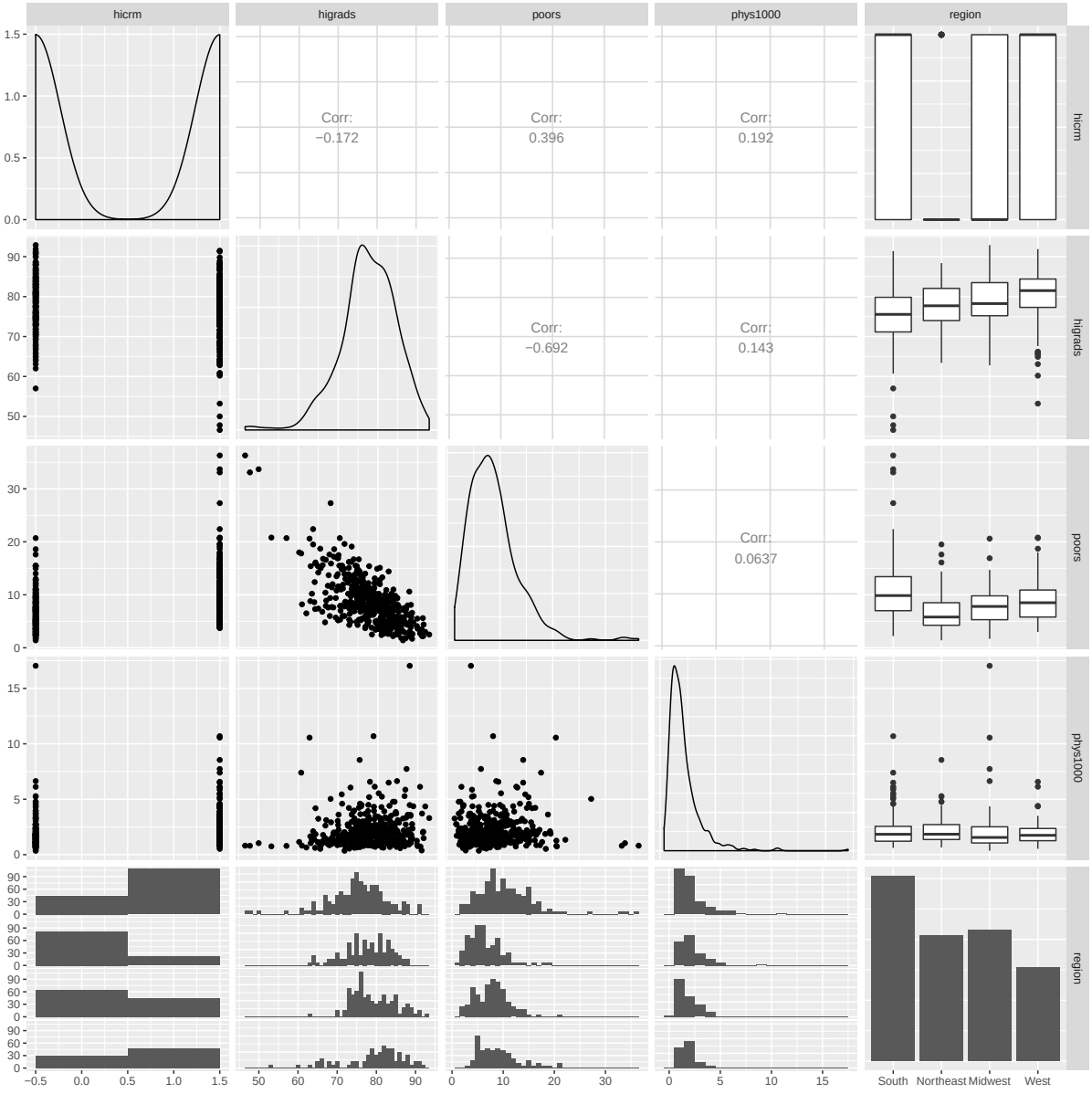

Figure 14: Plot of covariates against eachother

The optimal model included the previously studied covariate `region`, but discarded the `higrads` covariate. In addition, it included the new `poors` and `phys1000` covariates. One explaination why `higrads` was not used in the optimal may be seen in figure 14, where there is a high correlation between `higrads` and `poors`. With increased multicolinearity the standrard error of the coefficents increase. In worst case this can cause some varibles to become insignificant. Problems such as these are not to unlikley to happen when involving both the `region` and `higrads` covariates, which have a correlation of -0.692 bewteen them. This is probably one of the reasons why the optimal model performed better than the interaction model, which had both the `higrads` and `region` as covarites, in the previous section.

This hypothesis was tested in figure 15, where a linear regression model between `poors` and `higrads` was been fit. Studying the P-value of the model revealed that the β-values were highly significant.

Looking at how well `poors` predicts `hicrm` may be seen in figure 16. Here it may be seen that `poors` follows a more distinct S-shape, and that `poors` seems to split the dataset more distinctly between high

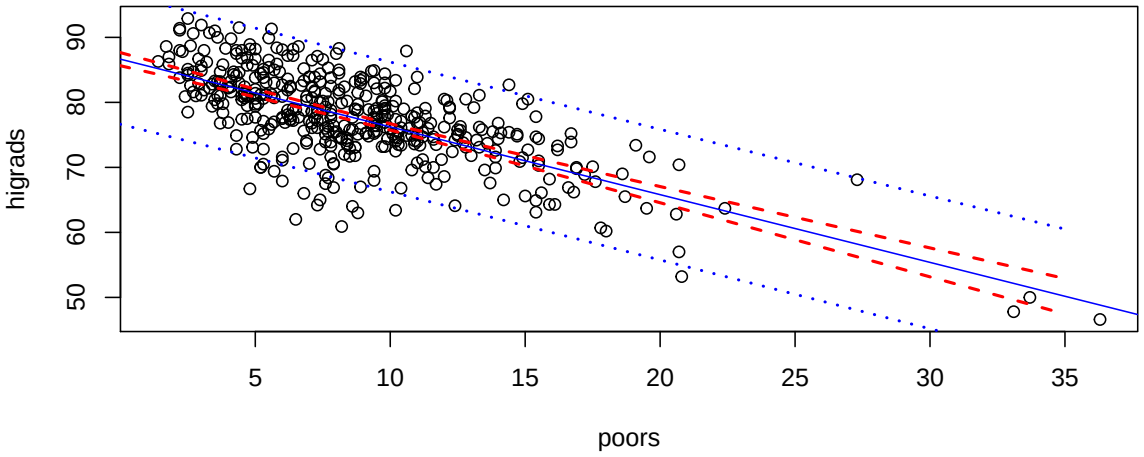

Figure 15: Plot of `higrads` against `poors`, together with linear regression line, with 95 % confidence and prediction intervals

and non-high crime rate, as it varies from ≈ 15, rather than the low separation discussed previously. As such, it seems that `higrads` and `poors` are highly correlated, but that `poors` better predict `hicrm` and is therefore better to use in the model.

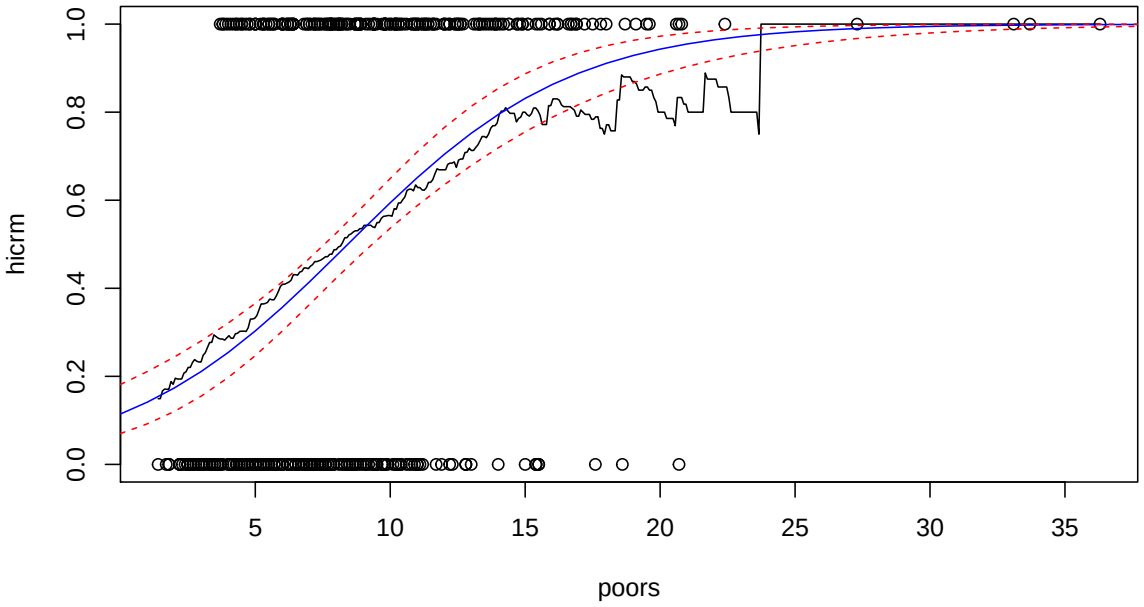

Figure 16: Plot of `hicrm` against `poors`, including kernel smoothing and prediction of fitted model with 95 % confidence interval

Regarding `phys1000`, it appears in 14 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.

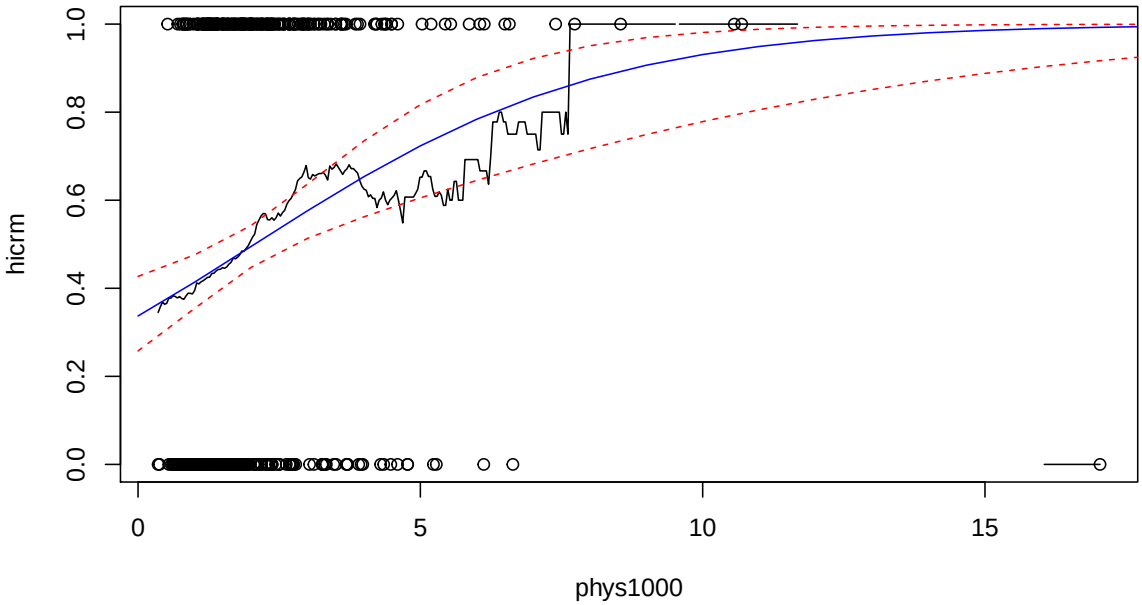

Figure 17: Plot of `hicrm` against `phys1000`, including kernel smoothing and prediction of fitted model with 95 % confidence interval

To Summarize, the optimal model ignores the `higrads` covariate and use `region`, `poors` and `phys1000` as covariates. Discarding the `higrads` covarite probably generated a better model, because of the problems associated with a high covariance between the `higrads` and `poors`. The low covarince among the covarites together with the fact that the `phys1000` and `poors` covarites are good at predicting `hicrm`, meant that the optimal model was better than the interaction model.