Project 2

Axel Sjöberg & John Rapp Farnes

8 maj 2019

Contents

1 Introduction

In this paper we bla bla

2 Analysis

2.1 Predicting crime rate based on education level

2.1.1 Seeing if there is relationship

We saw if there was a relationship by plotting bla bla with kernel smoother.

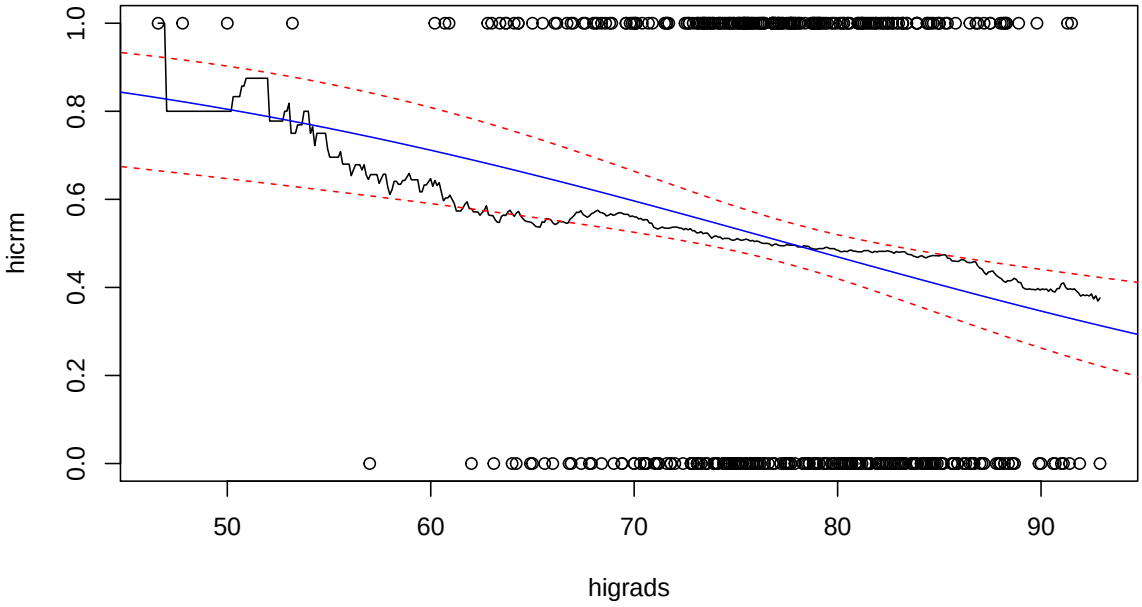

Figure 1: Bla bla

Seems to be relationship!

KOLLA UTAN BÖRJAN

–>

2.1.2 Confidence intervals and significance

Report the beta-estimates together with their confidence intervals and test whether the amount of adults with 12 years in school has a significant effect on the probability of having a higher than median crime rate

Table 1: Test

	Estimate	2.5 %	97.5 %	P-value (%)
β_0	3.866	1.527	6.27	0.044
β_1	-0.050	-0.081	-0.02	0.041

Significance!

2.1.3 Change in odds

Estimate the relative change in the odds (odds ratio) of having a high crime rate, with confidence interval, when the amount of higrads is increased by 1 (percent), and when it is increased by 10 (percent).

If higrads increases 1%, odd decreases by 4.9% If higrads increases 10%, odd decreases by 39.2%

2.1.4 Predict

Use the model to predict the probability, with 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).

Table 2: Test

Higrads	Probability	2.5 %	97.5 %
65	0.6519376	0.5481793	0.7430377
85	0.4087936	0.3415559	0.4796259

Use the model to predict, for each of the counties, whether it would be expected to have a low or a high crime rate (predicted probability below or above 0.5) and calculate the sensitivity and specificity for this model.

Sensitivity is the proportion of the true successes that have been correctly classified as successes (true positive). Specificity is the proportion of the true failures that have been correctly classified as failures (true negatives).

Sensitivity was 55%

Specificity was 58.2%

2.2 Predicting crime rate based on region

Make a cross-tabulation between region and hicrm. Choose as reference region in your regression models the one that has the largest number of counties in it's smallest low/high category. As a tie-breaker, use the other low/high category. Why is this a good idea? Hint: look at how the standard errors for the log odds (ratios) are calculated in this situation.

Table 3: Test

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

Set the reference level to Northeast, makes sense to get low SE

Fit a logistic regression using region as (categorical) covariate and report the beta-estimates together with their confidence intervals. Test whether there are any significant differences between the regions in the probability of having a high crime rate.

Table 4: Test

	Estimate	2.5 %	97.5 %	P-value (%)
β_0	-1.362	-1.867	-0.903	0.000
β_1	0.988	0.383	1.616	0.162
β_2	2.260	1.682	2.874	0.000
β_3	1.811	1.162	2.492	0.000

Significant!

Estimate the odds ratios for having a high crime rate, with confidence interval, for the different regions, compared to the reference region.

Table 5: Test

	OR	2.5 %	97.5 %
South	2.7	1.5	5.0
Midwest	9.6	5.4	17.7
West	6.1	3.2	12.1

Also estimate the probability of having a high crime rate, with confidence interval, for the different regions, including the reference region.

Calculate the sensitivity and specificity for this model. If we are allowed to have either higrads or region as covariate, which one should we choose?

Table 6: Test

	Estimate	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

Sensitivity was 70.5%

Specificity was 66.4%

Table 7: Test

Covariate	Sensitivity	Specificity
Higrads	0.5500000	0.5818182
Region	0.7045455	0.6636364

Would prefer region