Project 2: LOGISTIC REGRESSION
FMSN30
Lunds Tekniska Högskola

Astrid Dymling
Nils Romanus Myrberg
Martin Nybleus

May 2022

Abstract

In this report we investigate the relationship between personal characteristics and the probability of ending up in the hospital. Using a AIC step scheme we arrive at a logistic regressor using the covariates age, health status, sex, civil status and household income. The covariates that are associated with an increased risk of ending up in the hospitals are: age, "bad" or "somewhere in between" health status and being married, divorced or widowed. Household income, being a woman and the square of age are associated with a decreased risk of ending up in the hospital. The AUC score of this model is 0.67.

Part 1. Introduction to logistic regression

a)

By examining the proportions of at least one hospital day, depending on health, in table 1, it looks like the group with good health has the smallest risk of ending up in hospital. For further analysis we fit a logistic regression model to the data, using "good" health as reference category. The resulting β-estimates, together with the corresponding odds ratios, $\exp \beta$, comparing the other categories with the reference, as well as and 95 % confidence intervals for the odds ratios, are all presented in table 2. McFadden's adjusted pseudo R-squared, as well as AIC and BIC for the model are presented in table 3.

	0	1
Good	0.87	0.13
Bad	0.74	0.26
Somewhere in between	0.76	0.24

Table 1: Proportions for at least one hospital day depending on health status.

	(Intercept)	health: Bad	health: Somewhere in between
β	0.1325	0.1243	0.1050
I_β	(0.11989692 : 0.1450456)	(0.09130718 : 0.1572638)	(0.08360723 : 0.1263015)
$\exp \beta$	1.141646	1.132339	1.110660
$I_{\exp \beta}$	(1.127381,1.156092)	(1.095606,1.170304)	(1.087202,1.134624)

Table 2: A logistic regression model for "good"/"bad"/"somewhere in between" health categories as predictors for having at least one hospital day. β-estimates, the corresponding odds ratios comparing the other categories with the reference, as well as and 95 % confidence intervals.

MacFadden's adjusted pseudo R^2	AIC	BIC
0.02223817	5302.04	5328.785

Table 3: McFadden's adjusted pseudo R-squared, as well as AIC and BIC for the model.

In order to determine whether this model is significantly better than a model with only an intercept (null model) one may use a global likelihood ratio(LR) test with the null hypothesis H_0 : $\boldsymbol{\beta_1} = \mathbf{0}$. Where $\boldsymbol{\beta_1}$ denotes the coefficients associated with the health status. The LR test is done by comparing the difference in deviance between the full- and null model with the value of the $\chi^2(df)$ distribution. Where df denotes the degrees of freedom. The result of the LR test, displayed in table 8, shows that we may reject H_0

H_0	Distribution	χ^2-value	P-value	Conclusion
$\beta_1 = 0$	$\chi^2(3)$	246.19	2.2×10^{-16}	Reject H_0

Table 4: Result of LR test

Using our fitted model we can calculate the predicted probabilities of having at least one day in hospital for each of the three health categories, with 95 % confidence intervals. These results are displayed in table 5. The results seem reasonable when comparing the predicted probabilities, in table 5, with the cross-tabulation in table 1.

	Prediction	Lower	Upper
Good	0.1324713	0.1196401	0.1453025
Bad	0.2567568	0.2256470	0.2878665
Somewhere in between	0.2374256	0.2198226	0.2550287

Table 5: Proportions for at least one hospital day depending on health status.

b)

Hosp (probability of having at least 1 hospital day) was plotted against age in table 1, together with a moving average. As expected, hosp increases with age up until about the age of 75. The drop that occurs after that can probably be explained by people who are admitted to the hospital at that age are at increased risk of dying. This results in an elderly group above the age of around 80 that are relatively healthy and less likely to spend days in the hospital.

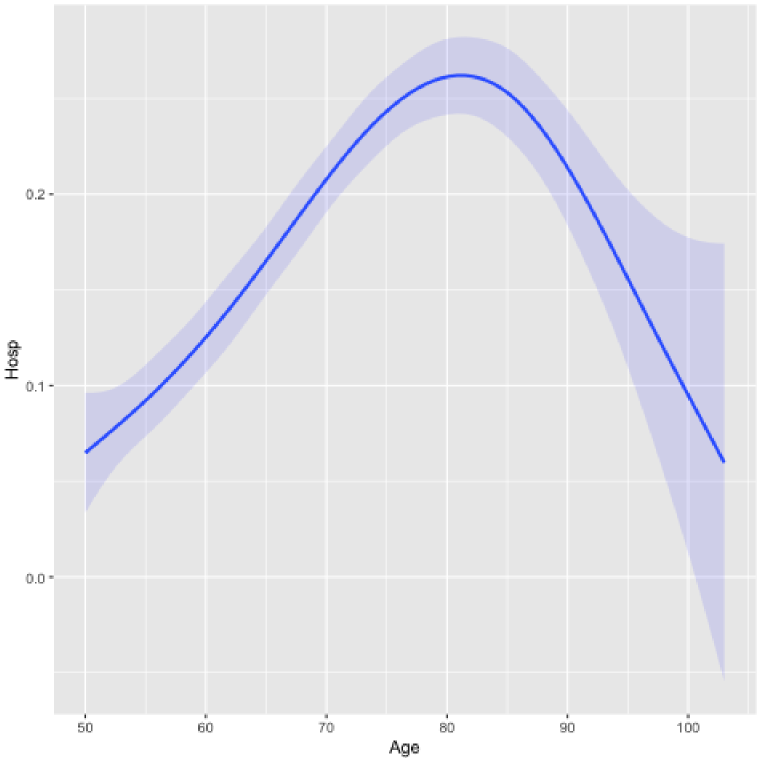

Figure 1: Moving average probability of hospitalization against age

Using age as a predictor we fit a logistic regression model, yielding the parameter estimates and odds ratios, with corresponding confidence bands, in table 6. The McFadden's adjusted pseudo R^2, AIC and BIC for the new model is found in table 7. Using a LR test we determine if age has a significant impact on the probability of having at least one day in hospital. The null hypothesis H_0, the distribution of the test statistic when H_0 is true, the observed value of the test statistic, the P-value and the conclusion are found in table 8. As shown in table 8, unsurprisingly, age has a significant impact.

	(Intercept)	Age
β	-0.1483275	0.0047424
I_β	(-0.205140633 : 0.091514403)	(0.003927902 : 0.005556923)
$\exp\beta$	0.8621487	1.0047537
$I_{\exp\beta}$	(0.8145328,0.9125482)	(1.0039356,1.0055724)

Table 6: A logistic regression model using age as predictor for having at least one hospital day. In the table β-estimates, the corresponding odds ratios, as well as and 95 % confidence intervals for the odds ratios, are all presented.

MacFadden's adjusted pseudo R^2	AIC	BIC
0.02379906	5291.588	5311.647

Table 7: McFadden's adjusted pseudo R-squared, as well as AIC and BIC for the model using age as predictor

H_0	Distribution	χ^2-value	P-value	Conclusion
$\beta_1 = 0$	$\chi^2(1)$	254.65	2.2×10^{-16}	Reject H_0

Table 8: Result of LR test for the model using age as a predictor

How the odds of having at least one day in hospital changes when age increases by 1 year is captured by the the age odds ratio estimate. This yearly change rate, with corresponding confidence bands, is found in table 6. By the same token one may also describe how the odds changes when age increases by 5 years by computing $\delta_5 := \exp(\beta_{age} \times 5) = 1.0047537$, with corresponding confidence bands $I_{\delta 5} = (1.019834; 1.028174)$. Using the new model we can compute the predicted probabilities and their 95 % confidence interval and plot them together with the moving average yielding figure 2.

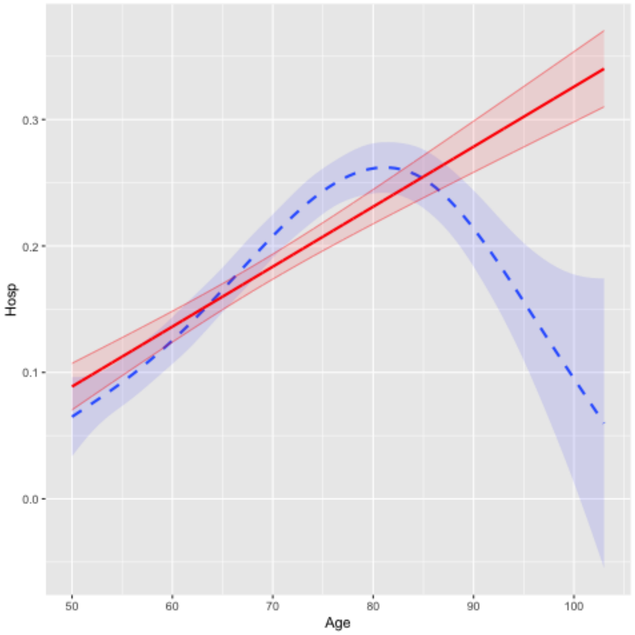

Figure 2: (Blue) Moving average probability of hospitalization against age with corresponding confidence bands. (Red) Predicted probability of hospitalisation using age as a predictor with corresponding confidence bands.

As shown in figure 2 the models does a fine job of capturing the monotonous relationship between the probability of hospitalisation and age before the age of 80. However, the model fails to capture the non-monotonous decrease after 80 due to it's simplicity.

c)

To model the non-monotonous relationship we introduce a model with a quadratic term $I(age^2)$. The beta-estimates, the exponential beta-estimates and their confidence intervals are presented in table 9.

	(Intercept)	age	I(age^2)
β	-1.1702758982	0.0347357029	-0.0002137272
I_β	(-1.4808463604 : -0.8597054359)	(0.0257367709 : 0.0437346350)	(-0.0002775907 : -0.0001498637)
$\exp\beta$	0.3102813	1.0353460	0.9997863
$I_{\exp\beta}$	(0.2274451, 0.4232867)	(1.0260708, 1.0447051)	(0.9997224, 0.9998501)

Table 9: The beta-estimates, the exponential estimates and their confidence intervals for the quadratic model.

McFadden's adjusted pseudo R2, AIC and BIC for the quadratic model are shown in table 10.

MacFadden's adjusted pseudo R^2	AIC	BIC
0.03172046	5250.698	5277.443

Table 10: McFadden's adjusted pseudo R2, AIC and BIC for the quadratic model.

To determine whether the quadratic term is significant or not, a likelihood ratio test was performed, with the nullhypothesis $H_0 : \beta_2 = 0$. The result of this LR test is displayed in table 11.

H_0	Distribution	χ^2-value	P-value	Conclusion
$\beta_2 = 0$	$\chi^2(1)$	42.89	5.791×10^{-11}	Reject H_0

Table 11: Result of partial LR test for the model using the square of age as a predictor

The relative change in the odds when the age increases by one year can be expressed as

$$\%\Delta OR := \frac{e^{\beta_0+\beta_1(x+1)+\beta_2(x+1)^2}}{e^{\beta_0+\beta_1 x+\beta_2 x^2}} = e^{\beta_1+\beta_2(2x+1)}. \tag{1}$$

The estimates of the odds ratios for age = 50, 75 and 100 years are listed in table 12.

Age	50	75	100
%ΔOdds ratio	1.013236	1.002466	0.9918103

Table 12: The odds ratio for different choices of age.

Using the new model we can compute the predicted probabilities and their 95 % confidence interval and plot them together with the moving average, and predicted probabilities using the linear model. A plot showing the predicted probabilities is displayed in figure 3.

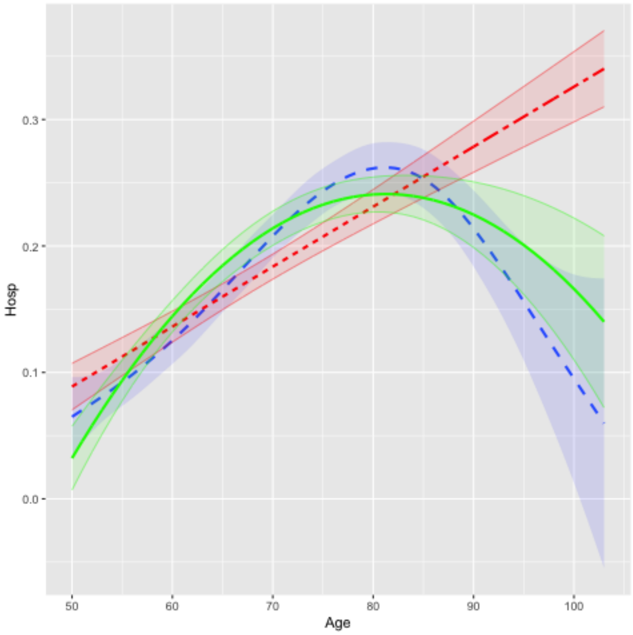

Figure 3: (Blue) Moving average probability of hospitalization against age with corresponding confidence bands. (Red) Predicted probability of hospitalisation using age as a predictor with corresponding confidence bands. (Green) Predicted probability of hospitalisation using age, and the square of age, as a predictors with corresponding confidence bands.

As shown in figure 3 the inclusion of a quadratic term increases the models ability to capture the non-monotonous relationship between age and hospitalisation probability, compared to the model without the quadratic term.

Part 2 Variable selection

In this section we introduce more variables: sex, civil status, exercise frequency, and average work week. For these variables we picked female, married, sometimes, and other/does not work, as reference variables. This because they had the highest n in their respective group, see tables 13 to 16.

	n
Male	2655
Female	3266

Table 13: Sex frequency table.

	n
unmarried	603
married	3540
divorced/seperated	470
widow/widower	1308

Table 14: Civil status frequency table.

	n
practically none	1522
sometimes	2197
once a week	836
twice a week	1009
rather strenuously at least twice a week	357

Table 15: Exercise frequency table.

	n
1-19 hours	73
20–34 hours	509
35–97 hours	1274
farmer or self-employed	381
other, does not work	3684

Table 16: Work week frequency table.

b)

The continuous variables were plotted against each other (see figure 4) and their correlations were calculated (see table 17). The variables are somewhat correlated but there are no problematic correlations, and none exceed a pairwise correlation of $|\rho| \leq 0.7$. However, inc_tot and inc_hh, unsurprisingly, might be slightly problematic.

	age	inc_hh	inc_tot
age	1.00	-0.56	-0.38
inc_hh	-0.56	1.00	0.64
inc_tot	-0.38	0.64	1.00

Table 17: Correlation coefficients for the continuous variables.

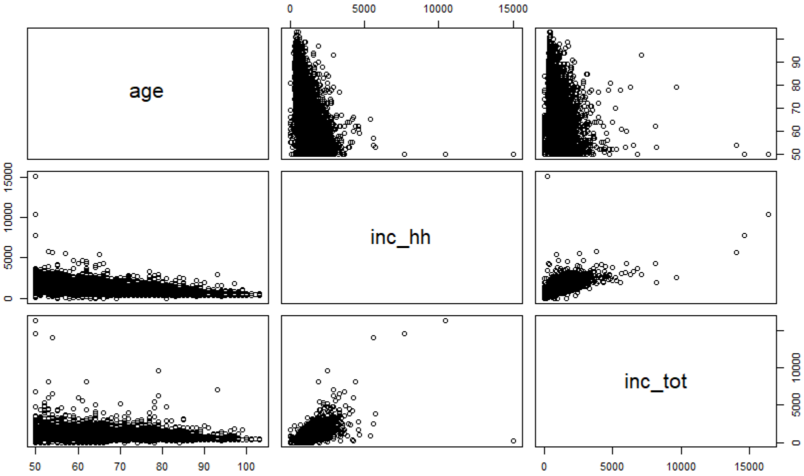

Figure 4: The continuous variables plotted agianst each other.

c)

Using the new categorical variables we fit a new logistic regressor yielding the performance metrics in table 18.

MacFadden's adjusted pseudo R^2	AIC	BIC
0.06166373	5120.572	5254.297

Table 18: McFadden's adjusted pseudo R2, AIC and BIC for the model with all variables.

To get a better view of if the new variables are significant, we perform a perform a suitable test of the significance of all variables in the model. We also test for age and age squared at the same time. The results of the repeated LR tests are displayed in table 19

H_0	Distribution	χ^2-value	P-value	Conclusion
$\boldsymbol{\beta} = \mathbf{0}$	$\chi^2(20)$	333.88	2.2×10^{-16}	Reject H_0
$\boldsymbol{\beta}_1 = \mathbf{0}$	$\chi^2(18)$	58.467	2.02×10^{-15}	Reject H_0
$\beta_2 = 0$	$\chi^2(17)$	53.045	3.3×10^{-13}	Reject H_0
$\boldsymbol{\beta}_3 = \mathbf{0}$	$\chi^2(18)$	11.292	0.01	Reject H_0
$\boldsymbol{\beta}_4 = \mathbf{0}$	$\chi^2(16)$	6.8911	0.1418	Don't reject H_0
$\boldsymbol{\beta}_5 = \mathbf{0}$	$\chi^2(16)$	2.5191	0.6412	Don't reject H_0
$\beta_6 = 0$	$\chi^2(19)$	3.3606	0.06677	Don't reject H_0
$\beta_7 = 0$	$\chi^2(19)$	0.416	0.519	Don't reject H_0
$\beta_8 = 0$	$\chi^2(19)$	25.654	4.1×10^{-7}	Reject H_0
$\beta_9 = 0$	$\chi^2(19)$	21.801	3.0×10^{-6}	Reject H_0
$\beta_8 = \beta_9 = 0$	$\chi^2(18)$	33.041	6.7×10^{-8}	Reject H_0

Table 19: Result of one global LR test and several partial LR tests where $\boldsymbol{\beta}$ denotes *all* the coefficients, $\boldsymbol{\beta}_1$ coefficients associated with health, β_2 the coefficient associated with sex, $\boldsymbol{\beta}_3$ the coefficients associated with civil status, $\boldsymbol{\beta}_4$ the coefficients associated with exercise, $\boldsymbol{\beta}_5$ the coefficients associated with normal working hours per week, β_6 the coefficient associated with household disposable income, β_7 the coefficient associated with individual income, β_8 the coefficient associated with age and β_9 the coefficient associated with age squared.

The results of the LR tests in table 19 show that we may reject that age, the square of age, health, sex and civil are insignificant. However exercise, normal working hours per week, household- and individual income might be insignificant.

d)

When performing a stepwise selection using AIC as criterion, we used the null model as starting model and lower scope and the full model as upper scope. The variables were included and excluded in the following order, see table 20. Notably, work_norm is first included and then excluded by the stepwise selection. As seen in table 20, the exclusion of work_norm happens after age-squared is included. Interestingly, the box plot in figure 5 shows the somewhat quadratic relationsship between age and work_norm. This is a probable cause as to why work_norm becomes redundant.

The McFadden adjusted pseudo R squared, AIC and BIC of the new model is found in table 26. The parameter- and odds estimates, with corresponding confidence bands, are displayed in table 22.

Step	Variable
1	+work_norm
2	+health
3	+sex
4	+age
5	+age^2
6	-work_norm
7	+civil_status
8	+inc_hh

Table 20: The selected variables, in order of inclusion and exclusion. + Denotes addition of feature and - denotes removal.

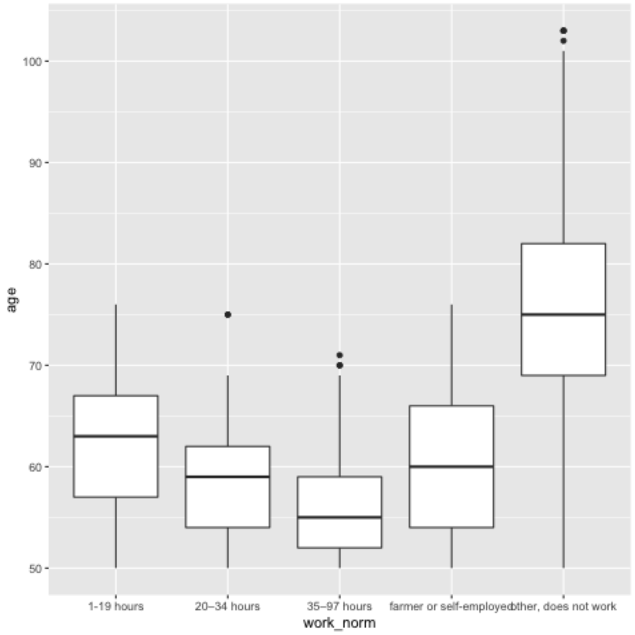

Figure 5: Box plot of age against normal working hours

MacFadden's adjusted pseudo R^2	AIC	BIC
0.06	5221	5288

Table 21: McFadden's adjusted pseudo R2, AIC and BIC for the resulting model from the AIC step

	β	I_β	$\exp\beta$	$I_{\exp\beta}$
(Intercept)	-11.90	[-14.6604 -9.21976]	0.000007	[0.0000004 0.0001]
health: Bad	0.655777	[0.4356 0.87246]	1.926640	[1.5459358 2.3928]
health: Somewhere in between	0.601330	[0.4503 0.75235]	1.824543	[1.5687547 2.1220]
sexFemale	-0.576120	[-0.7223 -0.43059]	0.562075	[0.4856196 0.6501]
age	0.269446	[0.1959 0.34495]	1.309239	[1.2164263 1.4119]
I(age^2)	-0.001702	[-0.0022 -0.00119]	0.998300	[0.9977799 0.9988]
civilst: married	0.410640	[0.1513 0.67857]	1.507782	[1.1633111 1.9711]
civilst: divorced/seperated	0.130042	[-0.2273 0.48309]	1.138876	[0.7967202 1.6211]
civilst: widow/widower	0.259057	[-0.0124 0.53756]	1.295708	[0.9876769 1.7118]
inc_hh	-0.000225	[-0.0004 -0.00005]	0.999775	[0.9995922 1.0000]

Table 22: Parameter- and odds ratio estimates for the resulting model from AIC step scheme with corresponding confidence bands.

In order to examine if the new model generated from the AIC step scheme if different from the model that includes all parameters we perform a LR test between these two. The results of the test is displayed in table 23. This test examines if the models are significantly different and the null hypothesis therefore becomes $H_0 : \boldsymbol{\beta}^* = \mathbf{0}$ where $\boldsymbol{\beta}^*$ denotes the coefficients corresponding the variables that are in the full model but not in the resulting model from the AIC step scheme.

H_0	Distribution	χ^2-value	P-value	Conclusion
$\boldsymbol{\beta}^* = \mathbf{0}$	$\chi^2(10)$	120	2.2×10^{-16}	Reject H_0

Table 23: Result of LR test between model that includes all variables and the reduced model generated by AIC step scheme. $\boldsymbol{\beta}^*$ denotes the coefficients corresponding the variables that are in the full model but not in the resulting model from the AIC step scheme

e)

In the same manner as in 2d) we performed a stepwise selection of the variables, but using BIC as criterion instead of AIC. The selected variables in their included order is shown in table 24.

Step	Variable
1	+age
2	$+age^2$
3	+health
4	+sex

Table 24: The selected variables, in order of inclusion and exclusion. + Denotes addition of feature and - denotes removal.

	β	I_β	$\exp\beta$	$I_{\exp\beta}$
(Intercept)	-13.256029	[-15.584 -10.440]	0.000002	[0.0000002 0.00003]
age	0.295609	[0.225 0.370]	1.345611	[1.2529263 1.44808]
I(age^2)	-0.001841	[-0.002 -0.001]	0.998141	[0.9976332 0.99864]
health: Bad	0.675168	[0.457 0.890]	1.964363	[1.5787299 2.43557]
health: Somewhere in between	0.618284	[0.468 0.768]	1.855740	[1.5975506 2.15557]
sex: Female	-0.567500	[-0.707 -0.429]	0.566941	[0.4933435 0.65118]

Table 25: Parameter- and odds ratio estimates for the resulting model from BIC step scheme with corresponding confidence bands.

MacFadden's adjusted pseudo R^2	AIC	BIC
0.06	5226	5266

Table 26: McFadden's adjusted pseudo R2, AIC and BIC for the resulting model from the AIC step

In order to examine if the new model generated from the BIC step scheme if different from the model that includes all parameters we perform a LR test between these two. The results of the test is displayed in table 27. This test examines if it is beneficial to include the parameters in the original model and the null hypothesis therefore becomes $H_0 : \boldsymbol{\beta}^* = \mathbf{0}$ where $\boldsymbol{\beta}^*$ denotes the coefficients corresponding the variables that are in the full model but not in the resulting model from the BIC step scheme.

H_0	Distribution	χ^2-value	P-value	Conclusion
$\boldsymbol{\beta}^* = \mathbf{0}$	$\chi^2(6)$	133	2.2×10^{-16}	Reject H_0

Table 27: Result of LR test between model that includes all variables and the reduced model generated by BIC step scheme. $\boldsymbol{\beta}^*$ denotes the coefficients corresponding the variables that are in the full model but not in the resulting model from the BIC step scheme

The set of variables included in the model generated by the BIC step scheme is a subset of the variables included in the model generated by the AIC step scheme. The models are hence nested. To test if these two models are significantly different we can perform an LR test with these two models. This test examines if it is beneficial to include all the parameters in the AIC model and the null hypothesis therefore becomes $H_0 : \boldsymbol{\beta}^* = \mathbf{0}$ where $\boldsymbol{\beta}^*$ denotes the coefficients corresponding the variables that are in the AIC model but not in the BIC model. The result of this test is found in table 28

H_0	Distribution	χ^2-value	P-value	Conclusion
$\boldsymbol{\beta}^* = \mathbf{0}$	$\chi^2(6)$	12.9	0.012	Reject H_0

Table 28: Result of LR test between the AIC model and the BIC model. $\boldsymbol{\beta}^*$ denotes the coefficients corresponding to the variables that are in the AIC model but not in the BIC model

Using the model generated by the BIC step scheme we may calculate the predicted probabilities, with confidence intervals, and plot them using age on the x-axis with different colour confidence ribbons for the health categories. This probabilites are shown in figure 6.

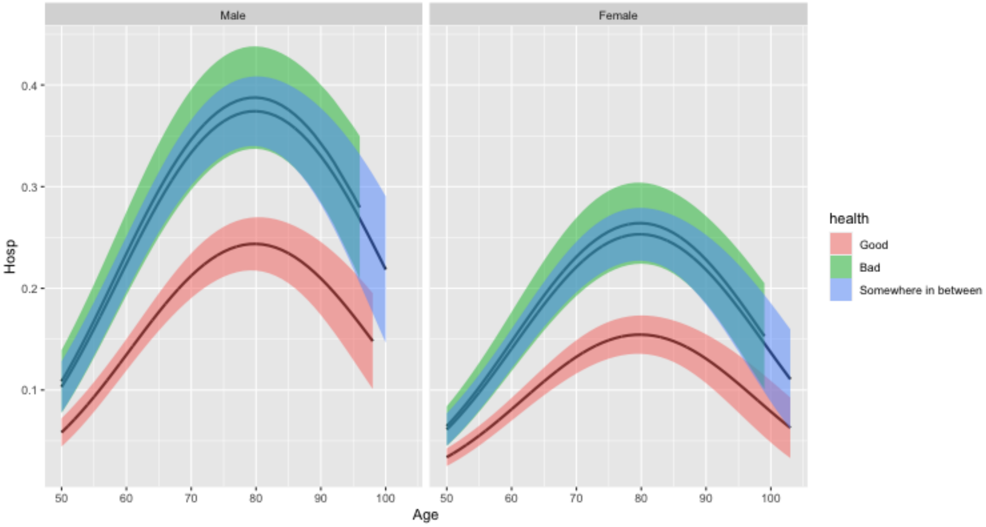

Figure 6: Predicted probabilities by age using the BIC step generated model for different sexes and health categories.

Figure 6 shows that it might not be necessary to separate between all three health categories as "bad" and "somewhere in between" gives roughly the same probabilities. A fact that strengthens this observation is that fitting the model using "bad" as reference category produces roughly the same parameter estimates. Instead, we create a new binary health variable "new_health" that maps 2 or 3 to the same category "not good" and 1 to "good". Using the BIC step scheme with the new health variable we get the parameter- and odds ratio estimates found in table 29. The AIC, BIC and McFadden adjusted pseudo R squared for the new model are displayed in table 30. The resulting model over performs the old in terms of AIC and BIC but not in terms of R squared.

	β	I_β	$\exp\beta$	$I_{\exp\beta}$
(Intercept)	-12.965283	[-15.575 -10.432]	0.000002	[0.0000002 0.00003]
age	0.296465	[0.225 0.370]	1.345096	[1.2524846 1.44747]
new_health: not good	0.675168	[0.492 0.772]	1.881258	[1.6357042 2.16472]
sex: Female	-0.567303	[-0.706 -0.429]	0.567053	[0.4934476 0.65130]
I(age^2)	-0.001857	[-0.002 -0.001]	0.998145	[0.9976372 0.99864]

Table 29: Parameter- and odds ratio estimates, with corresponding confidence bands, for the resulting model from BIC step scheme using a binary health category.

MacFadden's adjusted pseudo R^2	AIC	BIC
0.05751703	5115.024	5155.142

Table 30: McFadden's adjusted pseudo R2, AIC and BIC for the resulting model from the BIC step scheme using a binary health variable.

Part 3 Influential observations

a)

The leverages for the modified BIC-model are plotted against age, separately for each combination of sex and health categories, in figure 7. The observation with the highest leverage comes from an old woman in good health, see the triangle in mentioned figure. Generally it seems older individuals, females, and those in good health have the highest leverage.

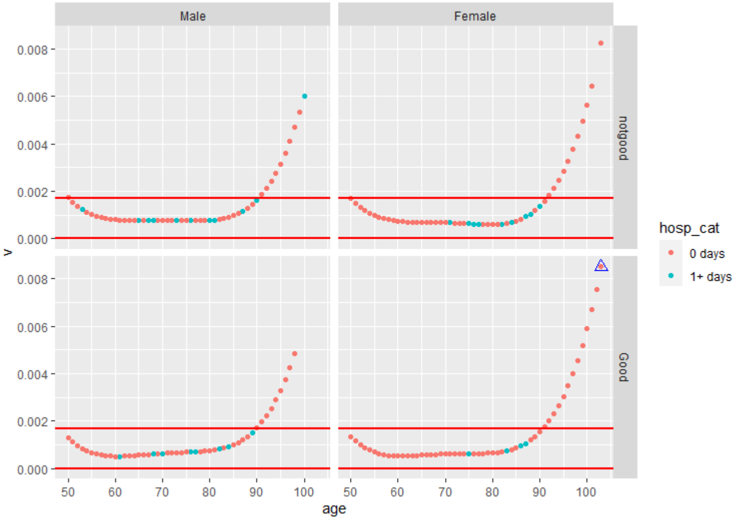

Figure 7: The leverages for the modified BIC-model are plotted against age, separately for each combination of sex and health categories. The individual with the highest leverage is marked with a triangle. Reference lines at $y = 0$ and $y = 2(p + 1)/n$.

b)

In figure 8 the standardized deviance residuals for the modified BIC-model were plotted against the linear predictor. There seems to be a linear correlation between the residuals and the predictor.

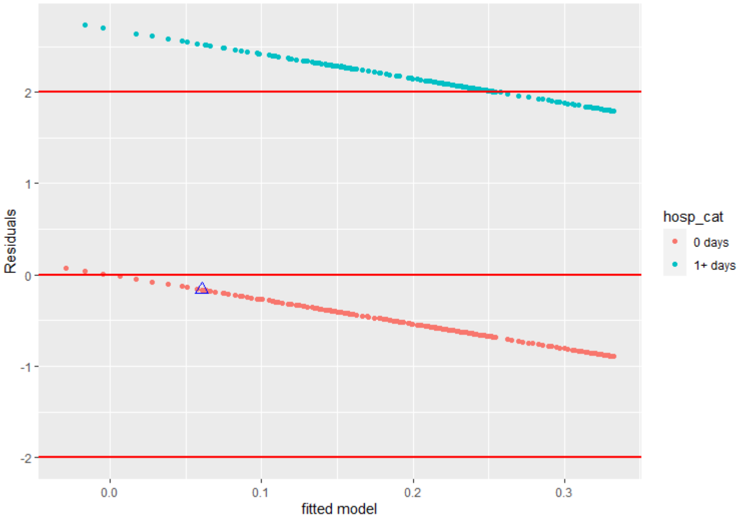

Figure 8: The standardized deviance residuals for the modified BIC-model plotted against the linear predictor. Reference lines at $y = 0$ and $y = \pm 2$.

c)

The Cook's distance was computed for the modified BIC-model and is shown in figure 9 plotted against age, for each combination of health and sex. The observations with the highest leverage and the highest Cook's distance are highlighted in the plot. Note that the observation with the highest leverage, blue triangle, has a rather small Cook's distance.

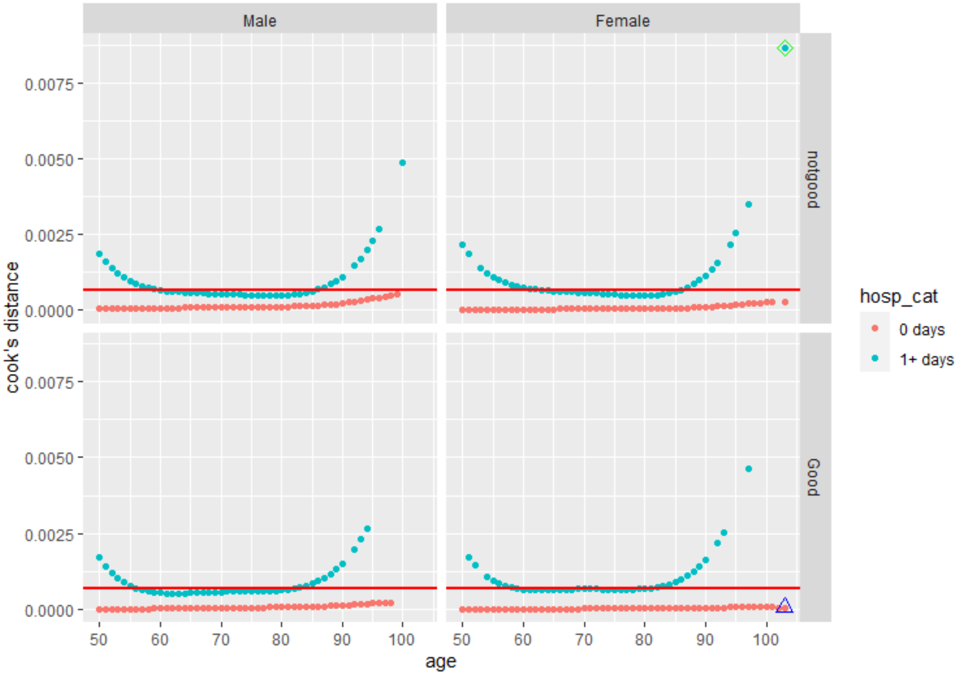

Figure 9: Cook's distance plotted against age, for each combination of health and sex. The observation with the highest leverage and the highest Cook's distance are highlighted with a blue triangle and green diamond, respectively. Reference lines at $y = 4/n$.

d)

The DFBETAS for the modified BIC-model were plotted against each parameter, separately for each combination of sex and health categories, in figures 10 to 13 below. Notably, the high leverage observation identified in 3(a) affected no parameters to a large degree (it is between the two reference lines in every plot), while the high Cook's distance-observation affected both the age and age-squared parameters.

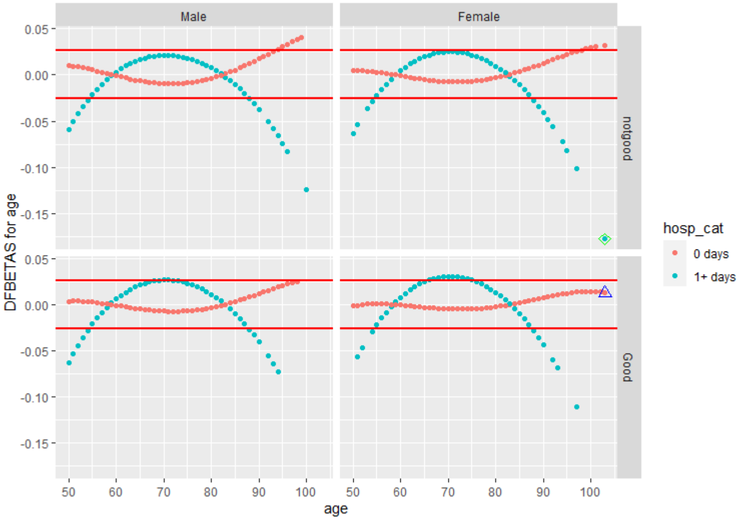

Figure 10: The DFBETAS plotted against age, for each combination of the categories health and sex. The observation with the highest Cook's distance is highlighted with a green diamond and the observation with the highest leverage is highlighted with a blue triangle. Reference lines are at $\pm 2/\sqrt{n}$ where n is the number of data points.

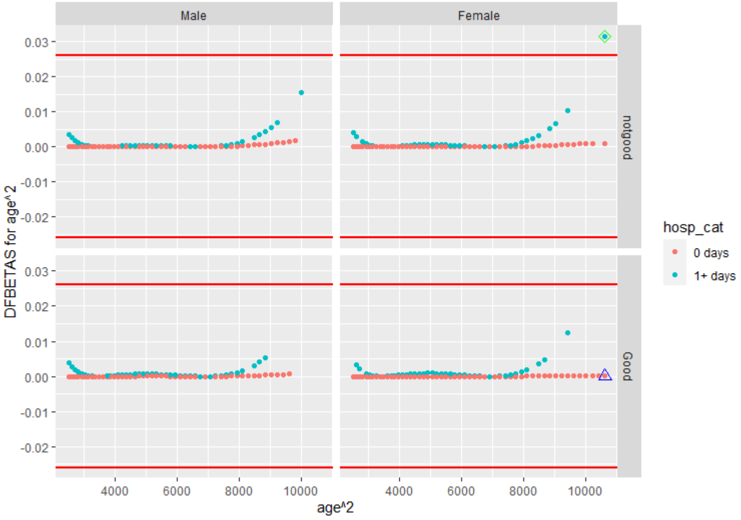

Figure 11: The DFBETAS plotted against age-squared, for each combination of the categories health and sex. The observation with the highest Cook's distance is highlighted with a green diamond and the observation with the highest leverage is highlighted with a blue triangle. Reference lines are at $\pm 2/\sqrt{n}$ where n is the number of datapoints.

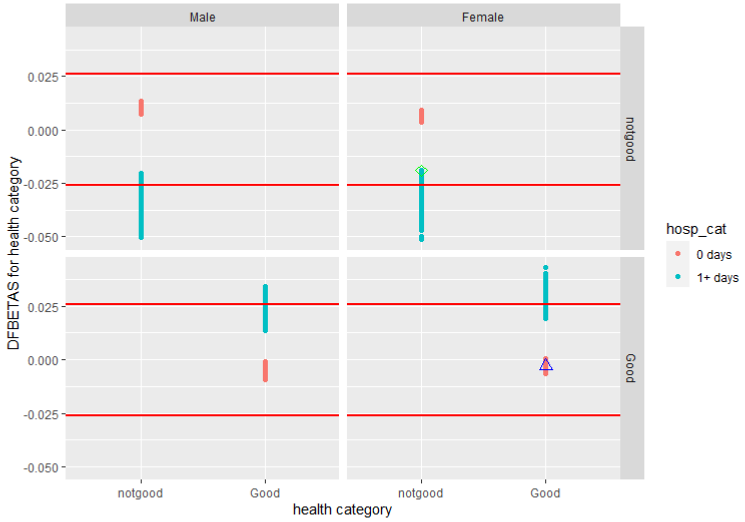

Figure 12: The DFBETAS plotted against health, for each combination of the categories health and sex. The observation with the highest Cook's distance is highlighted in green and the observation with the highest leverage is highlighted in blue. Reference lines are at ±$2/\sqrt{n}$ where n is the number of data points.

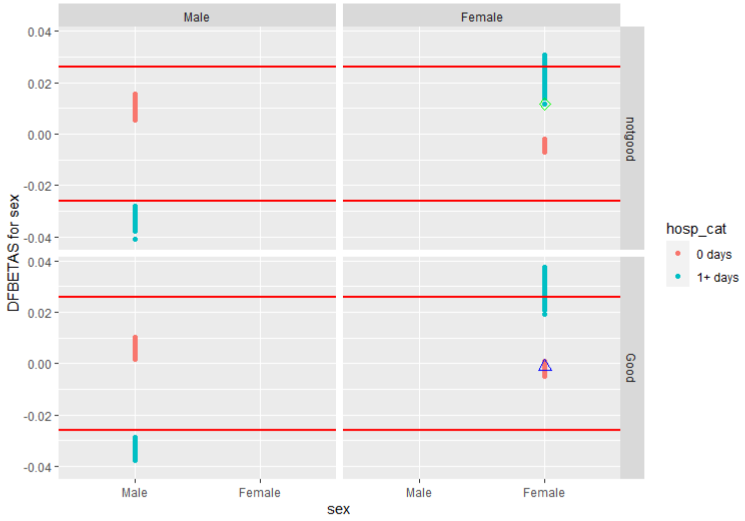

Figure 13: The DFBETAS plotted against sex, for each combination of the categories health and sex. The observation with the highest Cook's distance is highlighted with a green diamond and the observation with the highest leverage is highlighted with a blue triangle. Reference lines are at $\pm 2/\sqrt{n}$ where n is the number of data points.

Part 4 Goodness-of-fit

a)

Using a threshold value of 0.5, the observations of the AIC-model declared in 2 d) were classified according to table 31. Failure refers to "no days in hospital" and Success to "At least one day in hospital".

True (Y_i)	Failure ($\hat{p}_i \leq 0.5$)	Success ($\hat{p}_i > 0.5$)	Total
Failure ($Y_i = 0$)	4869	0	4869
Success ($Y_i = 1$)	1052	0	1052
Total	5921	0	5921

Table 31: Confusion Matrix for the AIC-model.

We have success rate 0 since no observations were found past the threshold of 0.5. Different measures of performance are shown in table 32. An accuracy of 80% might sound good, however our model is simply predicting 'negative' for all samples. I.e it is a rather unskilled classifier. This is shown in the sensitivity and precision metric beeing 0 using the threshold of 0.5.

Sensitivity	Specificity	Accuracy	Precision
0	1	0.8223273	0

Table 32: Performance measures

b)

In order to compare all the models we have assembled we can plot the ROC curves and inspect the area under these curves (AUC). The ROC curves for all models are displayed in figure 14 and the AUC scores in table 33

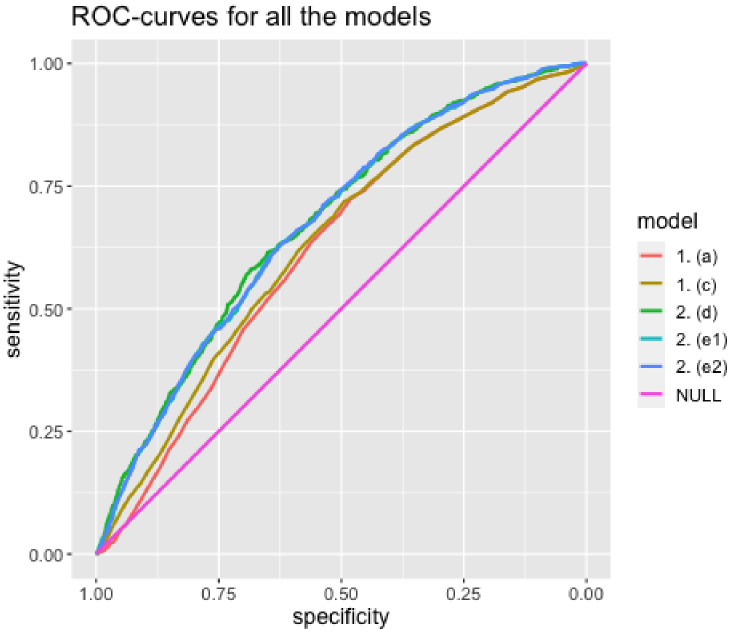

Figure 14: Roc curves for 1(a) Health, 1(c) Age squared, 2(d) AIC, 2(e1) BIC-3 with three health categories, 2(e2) BIC-2 with two health categories and (NULL) the null model.

	Model	AUC	Lower	Upper
1	NULL	0.50	0.50	0.50
2	1. (a)	0.62	0.60	0.64
3	1. (c)	0.63	0.61	0.65
4	2. (d)	0.67	0.66	0.69
5	2. (e1)	0.67	0.65	0.69
6	2. (e2)	0.67	0.65	0.69

Table 33: AUC scores for models for 1(a) Health, 1(c) Age squared, 2(d) AIC, 2(e1) BIC-3 with three health categories, 2(e2) BIC-2 with two health categories and (NULL) the null model.

To further compare our models we perform multiple pairwise DeLong's test comparing the AUC for model 2(d) AIC against each of the other models. The results of these tests are displayed in table 34. The results show that the 2(d) AIC is significantly different compared to (NULL), 1.(a) and 1.(c), however we may not reject that 2(d) has equal performance compared to 2(e1) and 2(e2). These results are quite unsurprising just from a visual inspection of figure 33 as these curves look rather similar.

H_0	Z-value	P-value	Conclusion
$\theta_{2(d)} = \theta_{NULL}$	-19.995	0.00000000000000022	Reject H_0
$\theta_{2(d)} = \theta_{1.(a)}$	-6.7567	0.00000000001412	Reject H_0
$\theta_{2(d)} = \theta_{1.(c)}$	-5.327	0.00000009985	Reject H_0
$\theta_{2(d)} = \theta_{2.(e1)}$	-1.4877	0.1368	Don't reject H_0
$\theta_{2(d)} = \theta_{2.(e2)}$	-1.3739	0.1695	Don't reject H_0

Table 34: Result of pairwise DeLong's tests comparing the AUC for model 2(d) AIC against each of the other models. θ denotes the AUCs.

c)

After experimenting with different values on p lower than 0.5, p= 0.21 was found to be the optimal value where the specificity and the sensitivity are approximately equal. The resulting confusion matrix is shown in table 35.

True (Y_i)	Failure ($\hat{p}_i \leq 0.21$)	Success ($\hat{p}_i > 0.21$)	Total
Failure ($Y_i = 0$)	3086	1783	4869
Success ($Y_i = 1$)	394	658	1052
Total	3480	2441	5921

Table 35: Confusion Matrix for the AIC-model with the optimal p.

The different performance measures when using the new p is presented in table 36.

Sensitivity	Specificity	Accuracy	Precision
0.6254753	0.6338057	0.6323256	0.2695617

Table 36: Performance measures with optimal p.

Comparing the confusion matrix in table 31 and the one in table 35 we can see that the number of correctly classified observations decreased while the number of false positives and false negatives increased in total. That explains the decrease in accuracy that we observed in table 36. Still, this classifier is far superior since it is no longer taking the approach of simply predicting ”negative” for each sample.

d)

In order to test the goodness of fit for model 2.(d) AIC we can perform a Hosmer-Lemeshow (HL) test. Visualisations of the HL tests are displayed in figures 15 and 16. The quantitative results of the HL test can be viewed in table 37. The plots in figure 16 line up with the results of the HL tests in table 37. In the cases where we reject the null hypothesis of models giving the correct probabilities the number of observed observations fall further away from the expected ones. In the cases we are not rejecting the null hypothesis the line representing the observed number of observations lies closer to the expected number of observations. Contrary to what we typically do with a hypothesis test, we want to *not* reject the null hypothesis as this implies that our model is correct.

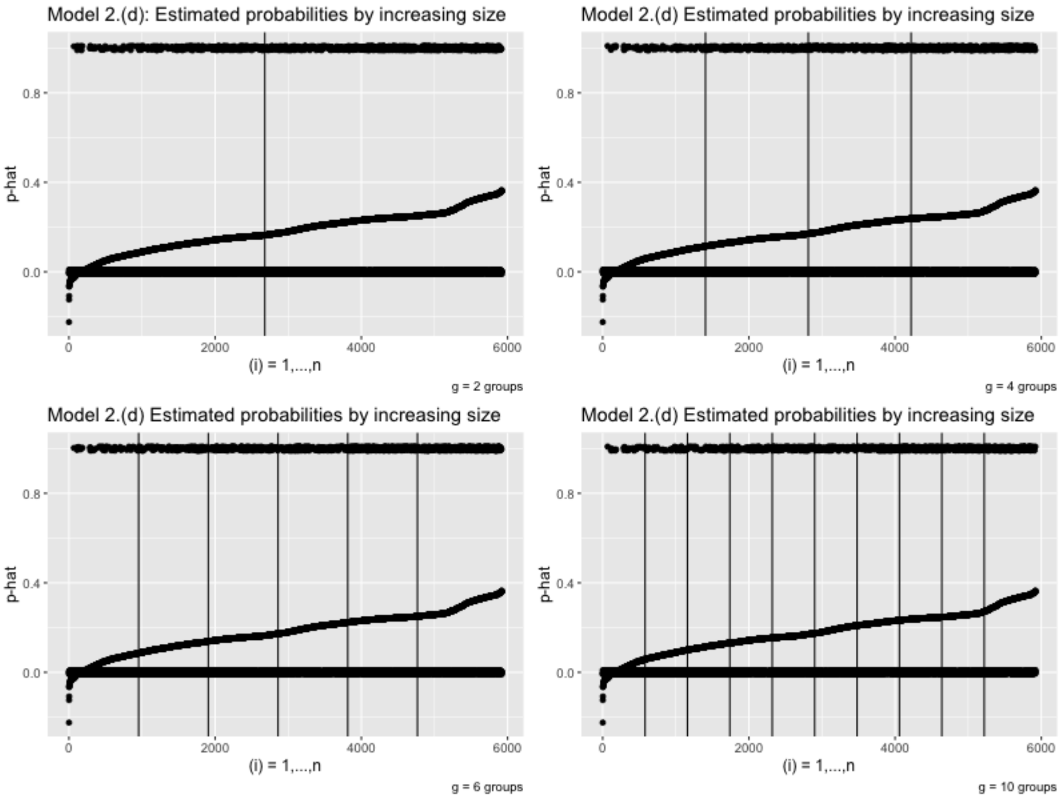

Figure 15: Estimated probabilities by increasing size for model 2(d) AIC using 2,4,6 and 10 groups.

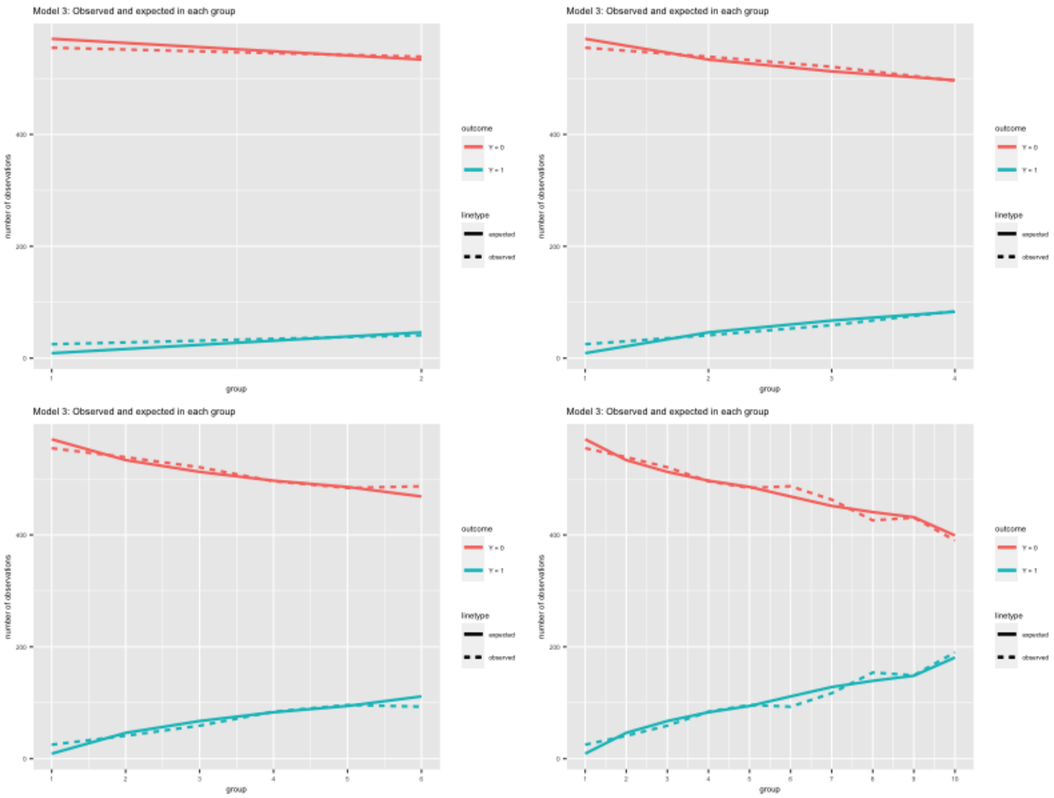

Figure 16: Observed- and expected number of observations against number of groups for model 2(d) AIC using 2,4,6 and 10 groups.

H_0	#groups	Distribution	χ^2-value	P-value	Conclusion
The model gives	10	$\chi^2(8)$	38.694	0.000005602	Reject H_0
correct probabilities	6	$\chi^2(4)$	9.3981	0.05188	Don't reject H_0
	4	$\chi^2(2)$	1.9769	0.3722	Don't reject H_0
	2	$\chi^2(0)$	0.12703	0.00000000000000022	Reject H_0

Table 37: Results from Hosmer-Lemeshow goodness of fit test for for model 2.(d) AIC using 2,4,6 and 10 groups

e)

By examining the AUC-scores in table 33 we have three obvious candidates for the best model. The 2(d) AIC model, 2(e1) BIC-3 with three health categories and 2(e2) BIC-2 with two health categories. The 2(d) AIC model however has the highest lower band value and the tightest confidence band. This model has furthermore been subject more rigorous diagnostics using the DL- and HL tests, with successful results. We can therefore consider 2(d) AIC model to be the best model, however, 2(e1) BIC-3 2(e2) BIC-2 might be equally good. One may also note from table 33 that none of our classifiers are especially good as an AUC of 0.67 is not that much larger than the no-skill classifier AUC score of 0.5. A logistic regressor using the available data might simply not be optimal.