LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN INSTITUT FÜR STATISTIK

Bachelor Thesis
in Statistics
at Ludwig-Maximilians-Universität München
Faculty for Mathematics, Informatics and Statistics
Department Statistics

Comparison of Several Statistical Models and Data Mining Methods for Predicting the Loss-To-Follow-Up of a Tb Study

presented by
Uwe Pritzsche

Advisor:	PD Dr. Christian Heumann
Examiner:	PD Dr. Christian Heumann
Editing timeframe:	13.09.2013 - 10.11.2013

Declaration

I hereby declare that I have written this bachelor thesis without help from others, using only the sources listed in the bibliography.

Munich, 10.11.2013

…………………………………………..
Uwe Pritzsche

Abstract

The goal of this Bachelor thesis is to derive the influential predictors of the loss-to-follow-up (LTFU) of an infant Tuberculosis (Tb) study. Several statistical classification models and data mining methods are compared in order to determine the best prediction model. In particular these approaches are classical logistic regression, discriminant analysis, regularized regression with different lasso penalties, classification trees, Boosting (with trees and regressions), Random Forests, Support Vector Machines and Neural Nets. To honestly assess the method performance, 50-fold random splits into training and test data are used. The best methods create an AUC (area under the curve) of approximately 0.7, which justifies the confidence in the predictive power of the data and the used covariates. But no dominant method, that highly outperformed the others, results. Thus, to assess the important covariates, a synthesis of the best prediction models is given, with substantial interpretation due to an extended logistic regression comprising interactions and non-parametric modeling of covariate effects. This approach suggests that higher mother's age prevents from LTFU. The same holds for mothers already having a HDSS-ID (health and demographic surveillance system identification) at enrolment. Salaried working mothers without a HDSS-ID should be avoided; here Farmers are preferred (regarding mother's occupation). Furthermore, families with lower socio-economic status regarding housing type or education level should be favored, as well as mothers with less additional children for whom it is advantageous to be recruited as soon as possible (lower infant's age). These results can help in creating better retention strategies for future trials. From a methodic point of view, it is important that classical logistic regression was among the best models, which not only justifies the application in similar data situations, but helps in deriving the right conclusion, as this method supports detailed interpretation of covariate influence.

Contents

1 Introduction and Motivation

Tuberculosis is the deadliest infectious disease worldwide. Therefore the importance of developing new and improved vaccination is evident. For the supporting clinical trials a low loss-to-follow-up (LTFU) is inevitable and represents an important factor for the quality of the clinical trial, as a high LTFU might lead to underpowered studies and biased results.
Direct prediction of LTFU as well as identification of influential variables for LTFU can help in excluding potential subjects in advance, or in developing accurate retention strategies.

Above task is a classical classification problem, which is typically modeled by logistic regression or discriminant analysis. However, in recent years several new methods for such prediction tasks were invented. Some of them extend the logistic regression regarding the important step of covariate selection. But there are also some new methods that aim direct prediction and originate from the machine learning or data mining community and have proven to perform extremely well in classification problems. This promising fact encourages their application in predicting the LTFU of a study, even though some of them are black boxes, i.e. interpretation of covariate effects is limited.
On the other hand, the classical logistic regression is highly embedded in a statistical inference framework. Furthermore it allows flexible modeling of covariate effects including interactions and non-parametric (or additive) effects. Therefore logistic regression provides highly interpretable models, which makes it a favorable modeling technique not only in the statistical community.
When it comes to the selection of influential variables, the classical approach with logistic regression is just to compare p-values or conduct a selection based on procedures, which discriminate between models by means of a criterion, e.g. AIC. For this task some new methods were invented in the past decade. These are subsumed under the term *regularization*. One of these methods is represented by penalization of a model's likelihood. Different penalties result in different characteristics of the regularization. E.g. the *lasso* penalty might shrink the logistic regression parameters of non-influential covariates to exactly zero, depending on the extent of penalization, which results in an implicit variable selection or, in case of a categorical variable, an identification of relevant categories. In contrast, the alternative *group-lasso* tries to keep categories of the same categorical predictor together, which means that either all parameters of the dummy coded covariate are zero, or all are non-zero.
Another method which also shrinks regression parameters is *Boosting*. This approach was originally invented for pure classification in the machine learning community, but was then adapted to regression models. The original version uses classification trees as a so-called *base-learner* for classification purpose. The adaption for regression models exchanges the tree with a linear regression model.
A very handy tool for classification in terms of intuitive usability are *CART*s (classification and regression trees). This classical method subsequently splits the data by binary decisions. As a result the user just has to follow the branches of the tree to conduct the prediction for a new case, which makes this approach somewhat attractive for users, even though it shows high variability/instability.
The latter problem can be circumvented by *Random Forest*, which let several trees vote for the resulting class. As a cost, the easy decision path of CARTs gets lost.
Another widely used classification method in the machine learning community are *Support Vector Machines.* This technique is very successful in separating the data in the space of basis functions, build from the predictors, by maximizing the margin between points of the corresponding two target classes.
For already a long time *Neural Nets* are playing a very important role for classification problems in the data mining field. By pushing some "hidden" modeling layers between the input variables and the target, especially nonlinear variable influence can be automatically fitted.

The aim of this thesis is to compare all above mentioned methods in terms of their LTFU prediction performance for an infant Tb study. If one technique would outperform the classical logistic regression, it could be used to provide priceless low LTFU rates for forthcoming trials. But as it will be seen in forthcoming chapters, no clearly superior method yields. This not only encourages to elaborately analyzing the study data with the highly interpretable logistic regression, but also puts quite good confidence in its results. Furthermore the outcome of the other prediction methods can be used to better assess the importance of the predictors.

The first chapter comprises an explanation of the study data, followed by an extensive descriptive analysis of used variables. This is already accompanied by an initial exploration of covariate influence. Chapter 2 starts with presenting the results of the logistic regression. A first variable selection is conducted, followed by an extended design, using non-parametric fitting procedures, as well as interaction modeling. The final regression model provides the deepest insight into data relations. The following subsequent chapters introduce all used prediction methods theoretically. This is facilitated by applying the methods directly to the study data and, if meaningful, presenting results. Chapter 4 prepares the comparison of the predictive power of all methods by first explaining how to sensibly measure classifier performance. Then the results of the evaluation are presented and assessed together with a listing of used R packages and functions together with their parameter settings. The final synthesis of covariate influence is followed by a summary in the last chapter.

2 Initial Analysis

2.1 The Study

The analyzed data originate from an infant tuberculosis study, conducted with newborn of age 0-6 weeks in Kenya, lasting 2009-2011. It was not the real Phase III efficacy trial, but a preparatory study in order to measure Tb incidence. Furthermore it was intended to help in collecting information and experience regarding the willingness of the population to participate and the LTFU rate.

The relevant covariates comprise characteristics of the child like age, sex or place of birth (Health facility or Home), etc. and attributes of the mother at baseline, e.g. mother's age, education level, etc. (see also Table 2.1). The target event LTFU is given if – roughly speaking – 3 planned visits were missed. The planned visits are one after 6 weeks, plus additional every 4 months for 1-2 years. In order to keep LTFU rate low, several supporting actions were undertaken, like transport reimbursement, reminder visits, etc.

More than 20% of enrolled children were LTFU (or "Not Retained"). A usual target in clinical trials is less than 10%. Therefore creating effective prediction models and identification of influential variables can be crucial in attempting to reach this target.

2.2 Data Transformation

The original dataset has 2900 observations/patients. The analysis is done on a reduced dataset of 2695 subjects, which comprises all patients without the ones who died. The exclusion of died subjects is due to the fact, that in such cases, it cannot be decided whether a mother would have been retained or not if the children wouldn't have died. For an overview of used variables see Table 2.1.

Missings (NAs) and nominations of "Other" (for categorical variables) are imputed in the following way: For metric variables the median value is used, for categorical variables the mode category respectively. For the NAs this is done in order to keep as many data as possible. Possible "Other" categories of nominal variables are very sparsely filled (4 cases for *PlaceOfBirth*, 12 cases for *HousingType*) and therefore do not offer a reasonable category on their own. A complete case dataset was also kept aside, and the final analysis was also processed for this data. It should already be mentioned that no considerable differences to the analysis with imputed data resulted.
Interestingly, most missings show a special pattern, which can also be seen in Table 2.1. The missing values for some variables completely correspond to the missings of other variables. E.g. records with missings for *MomEducationLevel* show also missings for *MomOccupation*, *MothersOwn*, *Residence* and *MomAge*. The same happens for *ReceivedAnteNtlCare* and *HIVResultsAs*. This is probably due to the fact, that the corresponding variables are recorded on the same, electronically based, CRF (case report form). So it might be, that these CRFs were either not used, or overlooked for some patients, or not saved.

Furthermore the metric variables are centered. These centered variables are used instead of the original ones, starting with chapter 2.4, i.e. after descriptive analysis is completed. The centering not only supports numerical stability, but also gives the regression intercepts a meaningful interpretation as the effect of an "average patient".

For categorical variables, the reference category (assigned numerical value 1, see Table 2.1) is mostly chosen as the mode. Furthermore, for variables with more than 2 categories, the internal coding also

Variable	**Description**	**Total**	**Retained**	**Not Retained**
		Mean (Sd; Min-Max; NAs) / n (%)	Mean (Sd; Min-Max; NAs) / n (% of total)	Mean (Sd; Min-Max; NAs) / n (% of total)
Retained	Subject retained?			
	1 = Retained	2090 (77.6)		
	2 = Not Retained	605 (22.4)		
*Age**	Infant's age (days)	11.7 (10.8; 0-66; 0)	11.3 (10.7; 0-48; 0)	12.9 (11.3; 0-66; 0)
*WeightHeight**	Infant's birth weight/height (g/cm)	65.6 (10.3; 34.9-118.4; 0)	65.7 (10.2; 34.9-118.4; 0)	65.5 (10.8; 39.1-110.9; 0)
*Temperature**	Infant's birth temperature (°C)	36.5 (0.4; 34.2-39.9; 0)	36.5 (0.4; 34.8-39.9; 0)	36.5 (0.4; 34.2-38.4; 0)
*MomAge**	Mother's age (years)	25.6 (6.8; 9.9-51.9; 26)	26.4 (7; 9.9-51.9; 17)	22.9 (5.4; 13-50.8; 9)
PlaceofEnrolment	Place of enrolment			
	1 = Home	2600 (96.5)	2019 (77.7)	581 (22.3)
	2 = Health Facility	93 (3.5)	71 (76.3)	22 (23.7)
	NA	2 (0.1)	0 (0)	2 (100)
PlaceOfBirth	Place of birth			
	1 = Home	1685 (62.5)	1350 (80.1)	335 (19.9)
	2 = Health Facility	986 (36.6)	725 (73.5)	261 (26.5)
	3 = Other	2 (0.1)	1 (50)	1 (50)
	NA	22 (0.8)	14 (63.6)	8 (36.4)
Sex	Infant's sex			
	1 = Male	1370 (50.8)	1064 (77.7)	306 (22.3)
	2 = Female	1325 (49.2)	1026 (77.4)	299 (22.6)
InfantsDelivered	Number of infants delivered			
	1 = Singleton	2616 (97.1)	2022 (77.3)	594 (22.7)
	2 = Twins	79 (2.9)	68 (86.1)	11 (13.9)
MomEducationLevel	Mother's education level			
	1 = None	97 (3.6)	83 (85.6)	14 (14.4)
	2 = Primary	2184 (81)	1742 (79.8)	442 (20.2)
	3 = Secondary	352 (13.1)	231 (65.6)	121 (34.4)
	4 = Tertiary	36 (1.3)	17 (47.2)	19 (52.8)
	NA	26 (1)	17 (65.4)	9 (34.6)
MomOccupation	Mother's occupation			
	1 = Salaried worker	1538 (57.1)	1131 (73.5)	407 (26.5)
	2 = Farming	983 (36.5)	836 (85)	147 (15)
	3 = Labor	67 (2.5)	43 (64.2)	24 (35.8)
	4 = Business	63 (2.3)	47 (74.6)	16 (25.4)
	5 = Fishing	18 (0.7)	16 (88.9)	2 (11.1)
	NA	26 (1)	17 (65.4)	9 (34.6)
HousingType	Housing type			
	1 = Mud	1767 (65.6)	1442 (81.6)	325 (18.4)
	2 = Semi-permanent	488 (18.1)	366 (75)	122 (25)
	3 = Permanent	408 (15.1)	261 (64)	147 (36)
	4= Other	6 (0.2)	4 (66.7)	2 (33.3)
	NA	26 (1)	17 (65.4)	9 (34.6)
ReceivedAnteNtlCare	Mother received antenatal care?			
	1 = Yes	2386 (88.5)	1847 (77.4)	539 (22.6)
	2 = No	264 (9.8)	206 (78)	58 (22)
	NA	45 (1.7)	37 (82.2)	8 (17.8)
HIVResultsAs	HIV test result			
	1 = Non-reactive	2283 (84.7)	1745 (76.4)	538 (23.6)
	2 = Reactive	352 (13.1)	296 (84.1)	56 (15.9)
	3 = Indeterminent	15 (0.6)	12 (80)	3 (20)
	NA	45 (1.7)	37 (82.2)	8 (17.8)
MothersOwn	Mother’s own children (additional)			
	1 = ≤3 children	1408 (52.2)	1105 (78.5)	303 (21.5)
	2 = >3 children	1261 (46.8)	968 (76.8)	293 (23.2)
	NA	26 (1)	17 (65.4)	9 (34.6)

continued …

Residence	Residence			
	1 = Temporary	1537 (57)	1097 (71.4)	440 (28.6)
	2 = Permanent	1132 (42)	976 (86.2)	156 (13.8)
	NA	26 (1)	17 (65.4)	9 (34.6)
*: For metric variables centered versions are derived, getting a suffix "c" in their name				

Table 2.1: *Description of used variables. For metric variables the mean, standard deviation (sd), minimum and maximum value and number of missings (NA) are listed. For categorical variables the total count and percentage is given. For the strata (Retained, Not Retained) values the latter is relative to the total count in the specific category of the categorical variable (% of total).*

follows the total frequency count, with more frequent categories coming first (e.g. *MomOccupation*). Only for the variables recorded on an order scale this is different, and the coding is due to the natural ordering (e.g. *MomEducationLevel*). Sometimes both assignment strategies accidentally coincide (e.g. for *HousingType*).

To prevent sparsely filled categories, MomOccuption="Small business" and "Business owner (e.g. duka)" were merged to "Business", also "Skilled labor (e.g. carpenter)" and "Unskilled labor (e.g. construction worker)" to "Labor".

Further variables (beside the ones listed in Table 2.1) were available, but disregarded due to high rate of missings (e.g. *Father's education*) or obvious lack of influence (e.g. *Head circumference*). One comment must be given to the special variable *Visits*. This variable records whether a mother came for an unscheduled visit, and can be shown to be highly influential in the sense that, having such a visit, prevents from LTFU. This variable is disregarded, because of two reasons: First it is not known at enrolment time and therefore does not help in choosing promising mothers-child combinations. Secondly it is kind of confounded with the target variable, as retained subjects just have more time to conduct an unscheduled visit, just because they did not already drop out of the trial. This can be underlined by comparing the average time difference between enrolment and first unscheduled visit, which is much higher for retained patients.

2.3 First Descriptive Analysis

Table 2.1 gives an overview of statistical characteristics of variables used in the analysis. It is visually split in target variable (*Retained*), metric and categorical covariates. The overview still lists the number of missing values (NAs) and the number of occurrences of category "Other". Both are imputed (see also chapter 2.2) in forthcoming analyses, starting with Figure 2.1 and Figure 2.2.

The characteristics of the used variables can be summarized as follows:

Target Variable

- *Retained*: A total of 605 (22.4%) patients are not retained (LTFU). This is high above the "target" of 10%.

Metric Covariates (see also Figure 2.1)

- *Age*: The infant's age at enrolment is recorded in days. As it can be seen in Figure 2.1, the distribution is heavily right skewed for both strata ("Retained", "Not Retained"), which is due to the enrolment strategy: Potential mothers were asked to participate already before birth and are immediately contacted after birth, so that the typical enrolment age is probably low. The maximum age is 66. Even though this is, together with another child aged 54, outside the eligibility criterion of

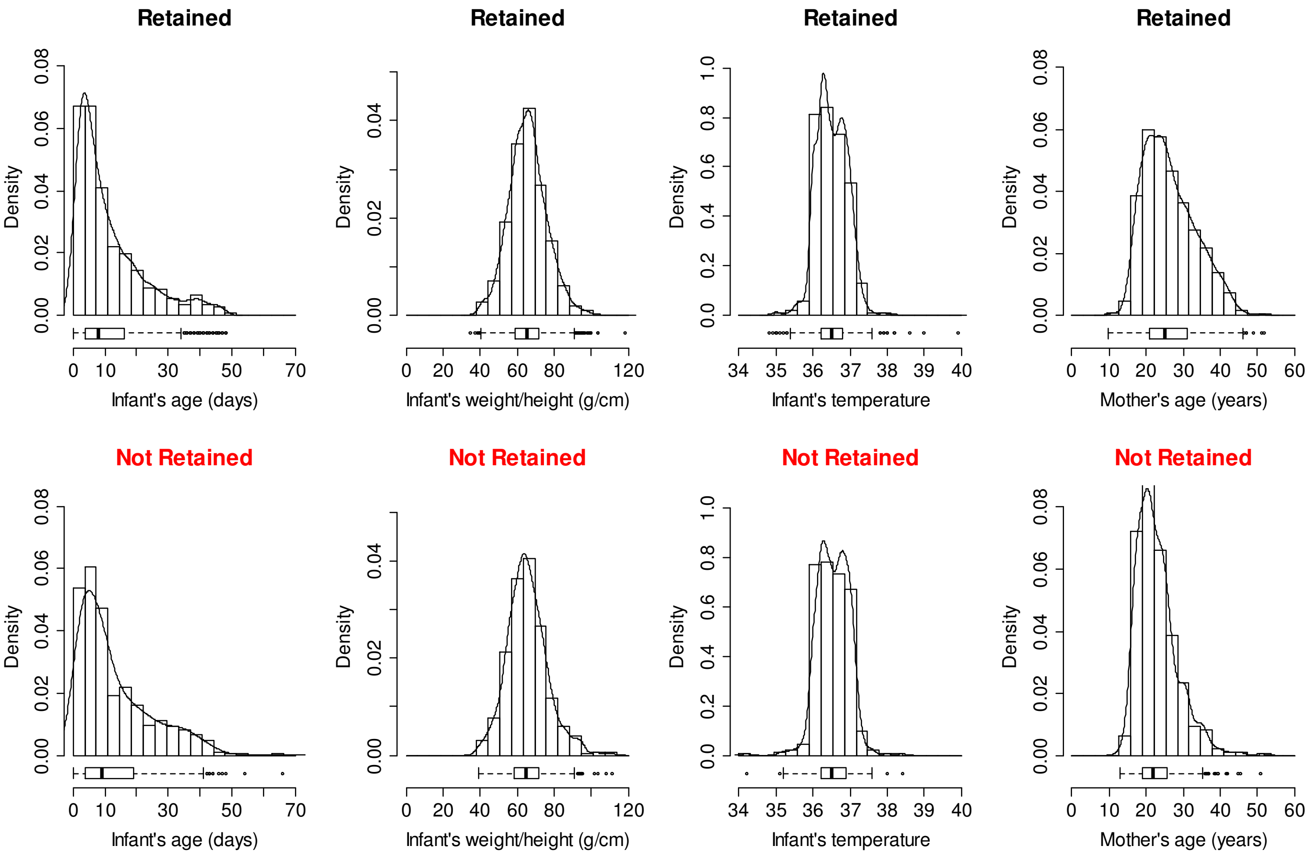


Figure 2.1: *Distribution of metric variables, split by target (with missing values already imputed).*

0-6 weeks (0-49 days) it is not assumed that these outliers are due to miscoding and therefore kept for analysis. The mean age for LTFU infants is just slightly (especially when accounting the standard deviation) above the one for retained subjects (12.9 to 11.3, see Table 2.1).

- *WeightHeight*: This variable measures the weight:height ratio in g/cm and is an indicator of the constitution of the infant. For both strata, similar symmetric shapes of the distribution (see Figure 2.1) result, as well as statistical metrics. The broad range of values (34.9 - 118.4) seems quite high, as two children with same height but weight:height ratios at the edges of this band would differ in weight by a factor of more than 3. But as this is not due to an extreme outlier (see Figure 2.1), it just hints at the highly variable constitution of the study population.
- *Temperature*: The temperature of the infant at birth might indicate a possible illness and therefore an influence on LTFU. Albeit, the marginal (without controlling for all other covariates) metrics and distribution plots do not suggest this effect.
- *MotherAge*: The average mother is 25.6 years. The maximum of 51.9 years, as well as the minimum of 9.9 years, are uncommon but not impossible. Figure 2.1 shows a right-skewed distribution for both strata. For the LTFU stratum the distribution is narrower and obviously more located at younger ages.

Categorical Covariates (see also Figure 2.2)

- *PlaceOfEnrolment*: Mostly mothers are enrolled at home. This also holds for both strata.
- *PlaceOfBirth*: Interestingly, almost 2/3rd of births occurring at home and not at a health facility. For the latter the fraction of patients not retained is slightly higher (26.5% to 19.9%) than for home births.

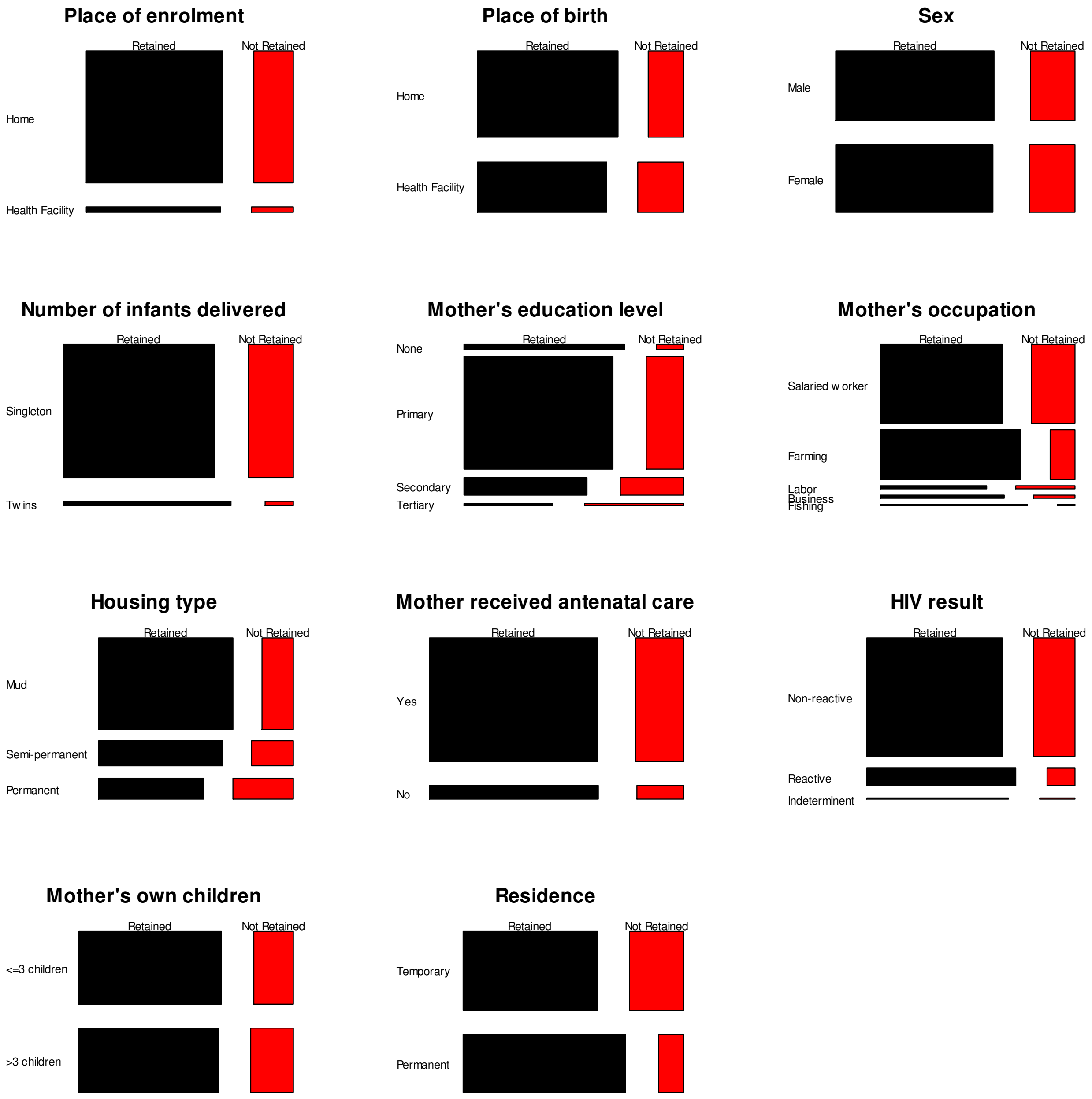


Figure 2.2: *Distribution of categorical variables, split by target (NAs (missings) and "Other" nominations already imputed).*

- *Sex*: 50.8% of enrolled children are male. This slight imbalance is usual and can be due to biological factors (see www.prb.org). The strata show equally distributed portions.
- *InfantsDelivered*: Only 2.9% of patients come as twins and are less probable to LTFU (13.9% to 22.7%) than singleton children. But the low portion of twins narrows the importance of this variable.
- *MomEducationLevel*: Nearly 85% of enrolled mothers have at most just a primary education level. Remarkably, the rate of LTFU subjects clearly increases with higher level of education, ranging from "None" over "Primary" and "Secondary" to "Tertiary" (14.4% – 20.2% – 34.4% – 52.8%)
- *MomOccupation*: Most mothers are salaried workers (57.1%). And 36.5% are working as farmers. Interestingly, none of the mothers is unemployed or characterizes herself as house wife, even though these categories ("Not working", "House wife") were available. Mothers with seemingly more

ambitious jobs (“salaried worker”, “labor”, “business” in comparison to “farmer” and “fisher”) have higher LTFU rates. This corresponds with above described education level influence as well as with the next categorical variable (in terms of the socio-economic status influence).

- *HousingType*: Nearly 2/3rd of enrolled mothers are living in a mud house and have a lower LTFU rate than mothers with at least a semi-permanent (mix of mud and cement) house.
- *ReceivedAnteNtlCare:* Most mothers received antenatal care (88.5%) and do not differ, regarding LTFU, to other mothers that haven’t received it.
- *HIVResultsAs:* Mothers, who did not have a HIV test, are merged with the ones whose test was non-reactive and comprise 85% of study population in total. They have a higher LTFU rate (23.6%) than mothers with a positive result (15.9%).
- *MothersOwn:* Nearly half of enrolled children have more than 3 siblings and are retained as often as children with less.
- *Residence:* This variable needs some explanation: It describes whether a mother has a HDSS-ID (*Residence*=”Permanent”). The health and demographic surveillance system (HDSS) identification is given every four months for people who have moved into the HDSS area (see also www.cdckemri.org). Mothers, who at the time of study had no HDSS identification, were given a temporary one (*Residence*=”Temporary”). The latter show a clearly higher rate of “Not Retained” participants (28.6% in comparison to 13.8%).

2.4 Explorative Analysis

As seen above, Figure 2.1 and Figure 2.2 provide a first impression about potential influential covariates. This exploration can be extended in order to show the influence on the main statistical metric for binary outcome, the odds ratio. As the relevant event is a patient that is not retained, here the odds measures the probability for a patient to be not retained, divided by the probability to be retained. The corresponding ratio is then the quotient relative to the reference category in case of a categorical covariate. For a metric covariate the odds for the minimum covariate value can be set to 1 to get a reference (see e.g. Figure 2.3).

2.4.1 Marginal Metrics

In Figure 2.3 and Figure 2.4 the crude, or marginal odds ratios (without controlling for other covariate influence), are shown together with 95% confidence intervals. These are calculated by a logistic regression with just one influential variable.

Figure 2.3 shows ascending odds ratios for higher infant’s age and temperature, and a very slight decrease for weight:height ratio. For the latter, the confidence interval would also comprise a constant line at 1. The same holds for infant’s temperature, even though the constant would not be at 1. Nevertheless, this makes the marginal effect also questionable. The crude odds ratio for mother’s age is remarkable, especially when noticing that the visual effect of such odds ratio plots depend on the reference. For example, if the reference for mother’s age is taken at the maximum age, the odds ratio would rise from 1 to a value of nearly 40.
The same effect must be kept in mind when viewing Figure 2.4. E.g. using “Permanent” instead of “Temporary” as the reference category for *Residence* would show an odds ratio of 2.5, visually seeming to be more far away from the reference value of 1.

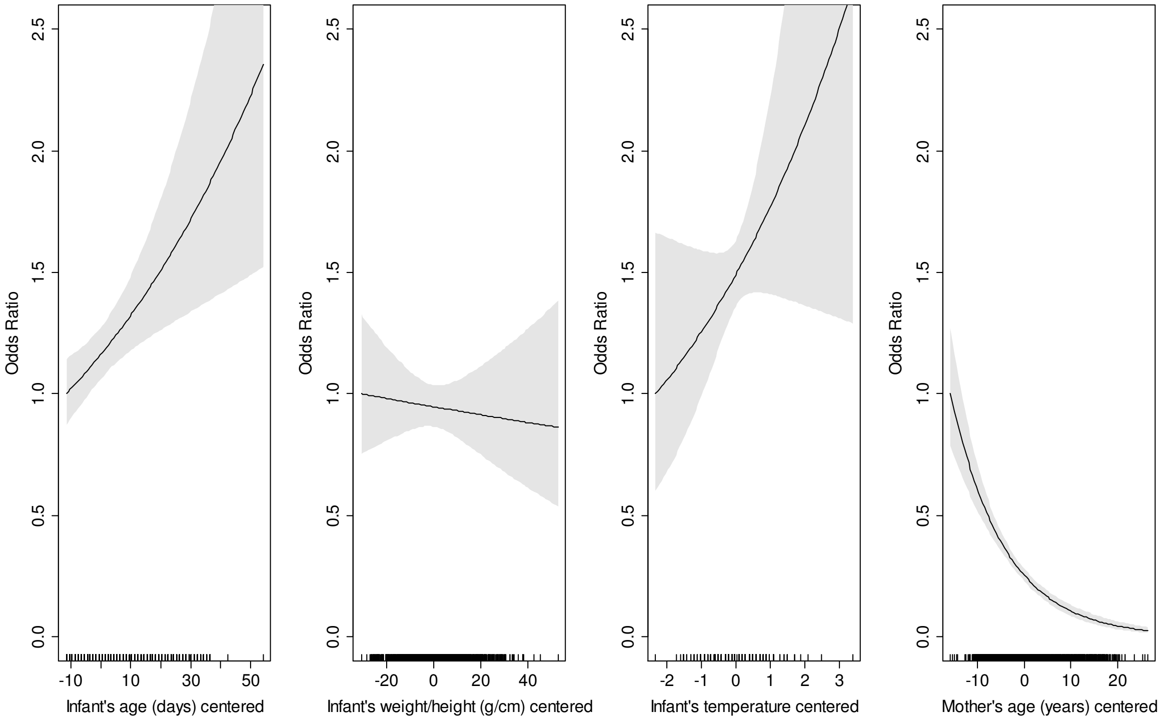
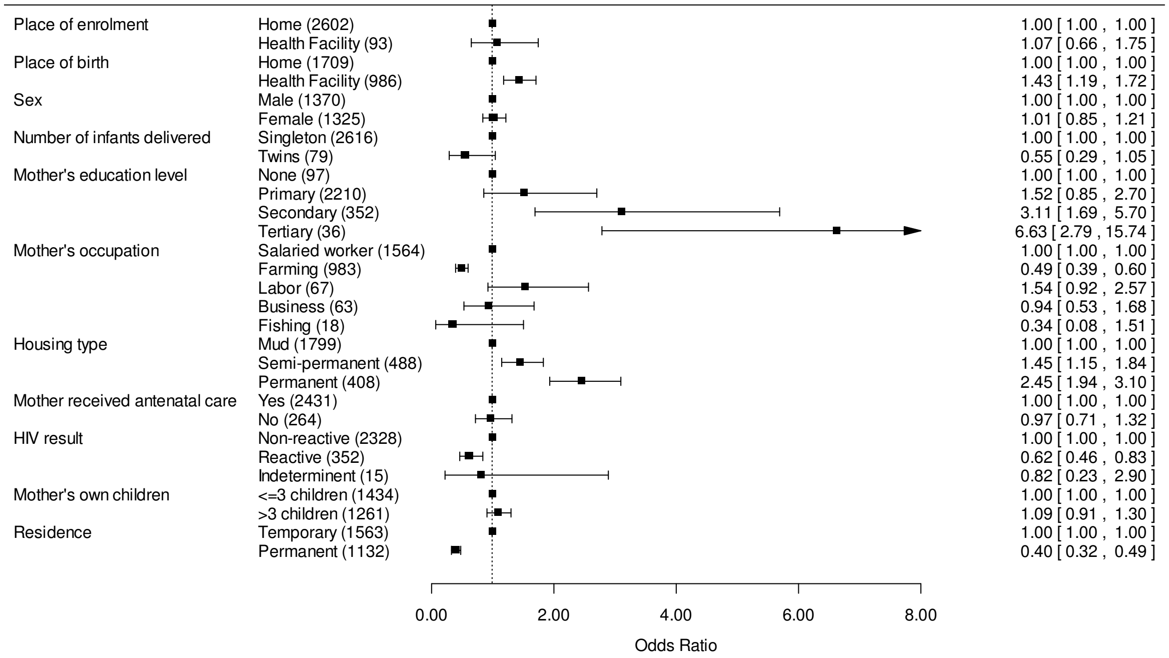


Figure 2.3: *Crude odds ratio effects of centered metric variables with 95% confidence intervals (pointwise). The odds ratio for the minimum input value is scaled to 1. Additional ticks showing input x-axis values.*



Place of enrolment	Home (2602)	1.00 [1.00 , 1.00]
	Health Facility (93)	1.07 [0.66 , 1.75]
Place of birth	Home (1709)	1.00 [1.00 , 1.00]
	Health Facility (986)	1.43 [1.19 , 1.72]
Sex	Male (1370)	1.00 [1.00 , 1.00]
	Female (1325)	1.01 [0.85 , 1.21]
Number of infants delivered	Singleton (2616)	1.00 [1.00 , 1.00]
	Twins (79)	0.55 [0.29 , 1.05]
Mother's education level	None (97)	1.00 [1.00 , 1.00]
	Primary (2210)	1.52 [0.85 , 2.70]
	Secondary (352)	3.11 [1.69 , 5.70]
	Tertiary (36)	6.63 [2.79 , 15.74]
Mother's occupation	Salaried worker (1564)	1.00 [1.00 , 1.00]
	Farming (983)	0.49 [0.39 , 0.60]
	Labor (67)	1.54 [0.92 , 2.57]
	Business (63)	0.94 [0.53 , 1.68]
	Fishing (18)	0.34 [0.08 , 1.51]
Housing type	Mud (1799)	1.00 [1.00 , 1.00]
	Semi-permanent (488)	1.45 [1.15 , 1.84]
	Permanent (408)	2.45 [1.94 , 3.10]
Mother received antenatal care	Yes (2431)	1.00 [1.00 , 1.00]
	No (264)	0.97 [0.71 , 1.32]
HIV result	Non-reactive (2328)	1.00 [1.00 , 1.00]
	Reactive (352)	0.62 [0.46 , 0.83]
	Indeterminent (15)	0.82 [0.23 , 2.90]
Mother's own children	<=3 children (1434)	1.00 [1.00 , 1.00]
	>3 children (1261)	1.09 [0.91 , 1.30]
Residence	Temporary (1563)	1.00 [1.00 , 1.00]
	Permanent (1132)	0.40 [0.32 , 0.49]

Figure 2.4: *Crude odds ratios of categorical variables with 95% confidence intervals. The reference category has value „1.00 [1.00, 1.00]". The number of patients of each category is listed in brackets after the category. The x-axis is cut at a value of 8; confidence intervals spreading beyond are marked with an arrow on the right side.*

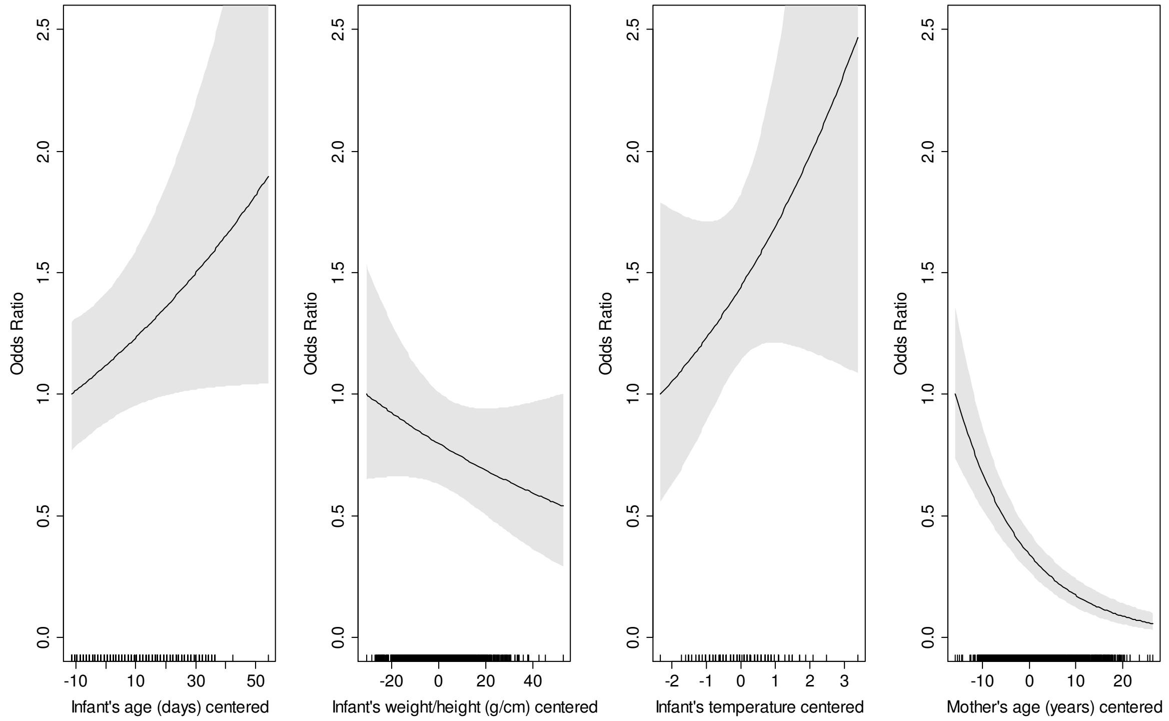


Figure 2.5: *Adjusted odds ratio effects of centered metric variables with 95% confidence intervals (pointwise), see also Figure 2.3. In fact one sees the predicted odds ratio for the corresponding covariate holding the values of the other covariates at typical values (mean value for metric, mode category for categorical covariates).*

Furthermore it can be seen that the width of the confidence intervals corresponds to the category cell frequencies (listed in brackets behind the category name). All in all, the odds ratios confirm the first visual influence analysis from chapter 2.3. Additionally it gives a better visual comparison of the effects of the different covariates and quantifies the uncertainty of the effects through the confidence intervals; thereby showing that the Number of infants delivered is not significant for instance.

In summary, the following categorical variables indicate an influence on LTFU, due to their marginal odds ratios: Place of birth, Mother’s education level (categories “Secondary” and “Tertiary”), Mother’s occupation (category “Farming”), Housing type, HIV result (category “Reactive”), Residence.

2.4.2 Adjusted Metrics

In order to get a direct comparison with the odds ratios, controlled for all other covariates, the results of the logistic regression (chapter 3.1) are already shown in Figure 2.5 and Figure 2.6.
For the odds ratio effects of the metric variables, it can be seen that the absolute value of the effects change slightly for infant’s age, weight:height ratio and temperature.
Together with the broadening confidence intervals only mother’s age keeps to be influential. This effect is more reduced than it visually seems: Making the same consideration as above and exchange the reference with the maximum age, the odds ratio would be less than 20 (in comparison with a value of 40 in a marginal setting, see Figure 2.3).

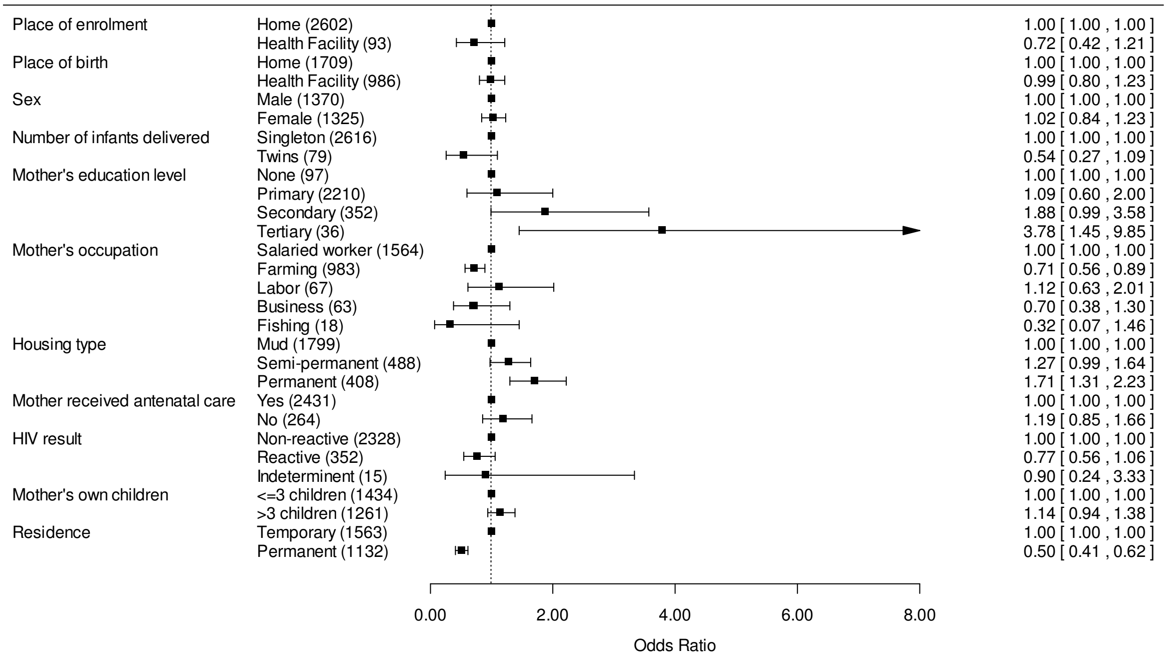


Figure 2.6: *Adjusted odds ratios of categorical variables with 95% confidence intervals. The reference category has value „1.00 [1.00, 1.00]", see also Figure 2.4.*

The odds ratios for the categorical variables also change when controlling for all covariates. The most remarkable differences to the marginal analysis are:

- *Place of birth*: Odds ratio is nearly disappearing (0.99 [0.80, 1.23] in comparison to 1.43 [1.19, 1.72]).
- *Mother's education level*: Odds ratios decrease. For "Secondary" to a non-significant value of 1.88 [0.99, 3.58] (compared to 3.11 [1.69, 5.70] crude odds ratio). The same holds for "Tertiary" (3.78 [1.45, 9.85], marginal: 6.63 [2.79, 15.74]) but which is still significant.
- *Mother's occupation*: Effect for "Farming" is also lowered from 0.49 [0.39, 0.60] to 0.71 [0.56, 0.89].
- *Housing Type*: Effects decreasing with "Semi-permanent" being no more significant (1.27 [0.99, 1.64], marginal: 1.45 [1.15, 1.84]).
- *HIV result*: "Reactive" is no more significant: 0.77 [0.56, 1.06] (marginal: 0.62 [0.46, 0.83])

Altogether, the influential categorical covariates are: Mother's education level ("Tertiary"), Mother's occupation ("Farming"), Housing type ("Permanent") and Residence.

3 Prediction Methods

3.1 Logistic Regression

3.1.1 Basic Modeling

Logistic regression (cf. Fahrmeir (2013)) is the most popular method to model classification problems. Here the expected value of the Bernoulli distributed target is modeled. In order to face the problem that the linear predictor $\eta = \boldsymbol{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ can take any value, but the expected value is restricted to [0;1], both are connected by a link function. In case of the *logit* link this results into:

$$log\left(\frac{\pi}{1-\pi}\right) = \eta \quad \Leftrightarrow \quad \frac{\pi}{1-\pi} = exp(\beta_0) * exp(\beta_1 x_1) * \ldots * exp(\beta_k x_k) \tag{3.1}$$

Using the logit link allows modeling the odds ratio as multiplicative model for the exponentiated covariate effects. The logs of the odds are called logits.

Applying logistic regression to the study data (using all covariates) yield the results shown in Table 3.1. A discussion of the effects was already done in chapter 2.4.2.

Checking model prerequisites

A critical assessment of model assumptions and a check for outliers and influential observations has to be done as well.

Due to the fact that Mother's occupation, Mother's education level and Housing type might be correlated (as they characterize the socioeconomic status), a potential collinearity problem, which would lead to highly increased confidence intervals, must be checked. Typically the variance inflation factor (VIF) is assessed. This figure quantifies the increase of variance of a covariate due to the linear dependence with other variables. All VIFs were below 2, which is highly below the rule-of-thumb critical value of 10 (see Fahrmeir (2013)).

Influential observations and outliers can be identified with *Cook's distance*. This figure characterizes the influence of the ith observation on the estimated $\widehat{\boldsymbol{\beta}}$-vector of effects by calculating the Mahalanobis distance of the leave-one-out estimation $\widehat{\boldsymbol{\beta}}_{(i)}$ to $\widehat{\boldsymbol{\beta}}$ (cf. Tutz (2012). A plot of these distances can be accompanied by an additional graph contrasting the leverage (*hatvalues*) of observations with the studentized residuals (*rstudent*), see Figure A.1, Appendix A (cf. Faraway (2005)). Even though no extreme observations could be identified, the most conspicuous ones are checked for uncommon values, e.g. extreme in the sense of lying on the edge of the value ranges. A *dfbetas* plot (cf. SAS9.2OnlineDoc) helps in identifying the relevant variables, as it basically shows the scalar Cook's distance for one covariate (see Figure A.2). When checking observation variable values, no abnormalities popped up.

Unfortunately there is no possibility to provide a powerful goodness-of-fit test. The usual tests, based on deviance and Pearson's χ_P^2, need sufficient cell counts for applying a fixed cell asymptotic, which is not given in case of metric predictors (see also Tutz (2012)). An alternative Hosmer-Lemeshow test (H$_1$: lack of fit) result in a p-value of 0.76. It is well known, that this test has moderate power and therefore favors H$_0$ (no lack of fit). Nevertheless the high p-value suggests that there is no crucial lack of fit.

Usually for logitstic regression also the rigid assumption of data variance being $\pi * (1 - \pi)$ must be proven. Often this is violated, which means that the data shows *overdispersion*. But for cell counts=1 as it is the case for the Tb data due to the metric covariates, no overdispersion can occur.

Covariate	exp(β)	95% CI	p-value
(Intercept)	0.27	[0.15; 0.50]	<0.001
Agec	1.01	[1.00; 1.02]	0.059
WeightHeightc	0.99	[0.98; 1.00]	0.195
Temperaturec	1.17	[0.93; 1.48]	0.179
MomAgec	0.93	[0.92; 0.95]	<0.001
PlaceOfEnrolmentHealth Facility	0.72	[0.42; 1.21]	0.215
PlaceOfBirthHealth Facility	0.99	[0.80; 1.23]	0.931
SexFemale	1.02	[0.84; 1.23]	0.876
InfantsDeliveredTwins	0.54	[0.27; 1.09]	0.086
MomEducationLevelPrimary	1.09	[0.60; 2.00]	0.771
MomEducationLevelSecondary	1.88	[0.99; 3.58]	0.055
MomEducationLevelTertiary	3.78	[1.45; 9.85]	0.006
MomOccupationFarming	0.71	[0.56; 0.89]	0.003
MomOccupationLabor	1.12	[0.63; 2.01]	0.697
MomOccupationBusiness	0.70	[0.38; 1.30]	0.264
MomOccupationFishing	0.32	[0.07; 1.46]	0.142
HousingTypeSemi-permanent	1.27	[0.99; 1.64]	0.065
HousingTypePermanent	1.71	[1.31; 2.23]	<0.001
ReceivedAnteNtlCareNo	1.19	[0.85; 1.66]	0.311
HIVResultsAsReactive	0.77	[0.56; 1.06]	0.106
HIVResultsAsIndeterminent	0.90	[0.24; 3.33]	0.873
MothersOwn>3 children	1.14	[0.94; 1.38]	0.196
ResidencePermanent	0.50	[0.41; 0.62]	<0.001

Table 3.1: *Odds ratio results of logistic regression. Reference categories for categorical covariates are not listed. Naming of categorical covariates follows the rule "VariablenameCategoryname". Effects of categorical variables correspond to values shown in Figure 2.6. Significant p-values (α=0.05) are shaded.*

Even though there might be two patients with exactly the same covariate values, this very unlikely case can be disregarded when accounting for overdispersion.

For checking the linearity assumption of the metric predictors it is not helpful to plot the usual residual plots (residuals vs. predicted values, residuals vs. values of one covariate) as the residuals can take only two values for a binary target. In case of low cell counts (due to metric covariates) these plots show confusing curved lines, corresponding to the limited number of observed responses (see Faraway (2006)). A much more convenient way for testing non-linearity, is to model the influence of the metric variables in a non-parametric manner, i.e. applying a generalized additive model. This approach is introduced in more detail below (chapter 3.1.2). For now, only the results shown in Figure 3.1 are discussed. When assessing the nonlinear influence, it is important to recognize also the confidence intervals in the logit plot. As a rule of thumb, one can still assume a linear influence if a straight line fits into these intervals. As it can be seen, this is only not possible for *MomAgec* but not in a dramatic manner. Thus for now, a model with just linear effects is assumed to be sufficient. Adapting the model due to this issue is invented in the next chapter 3.1.2.

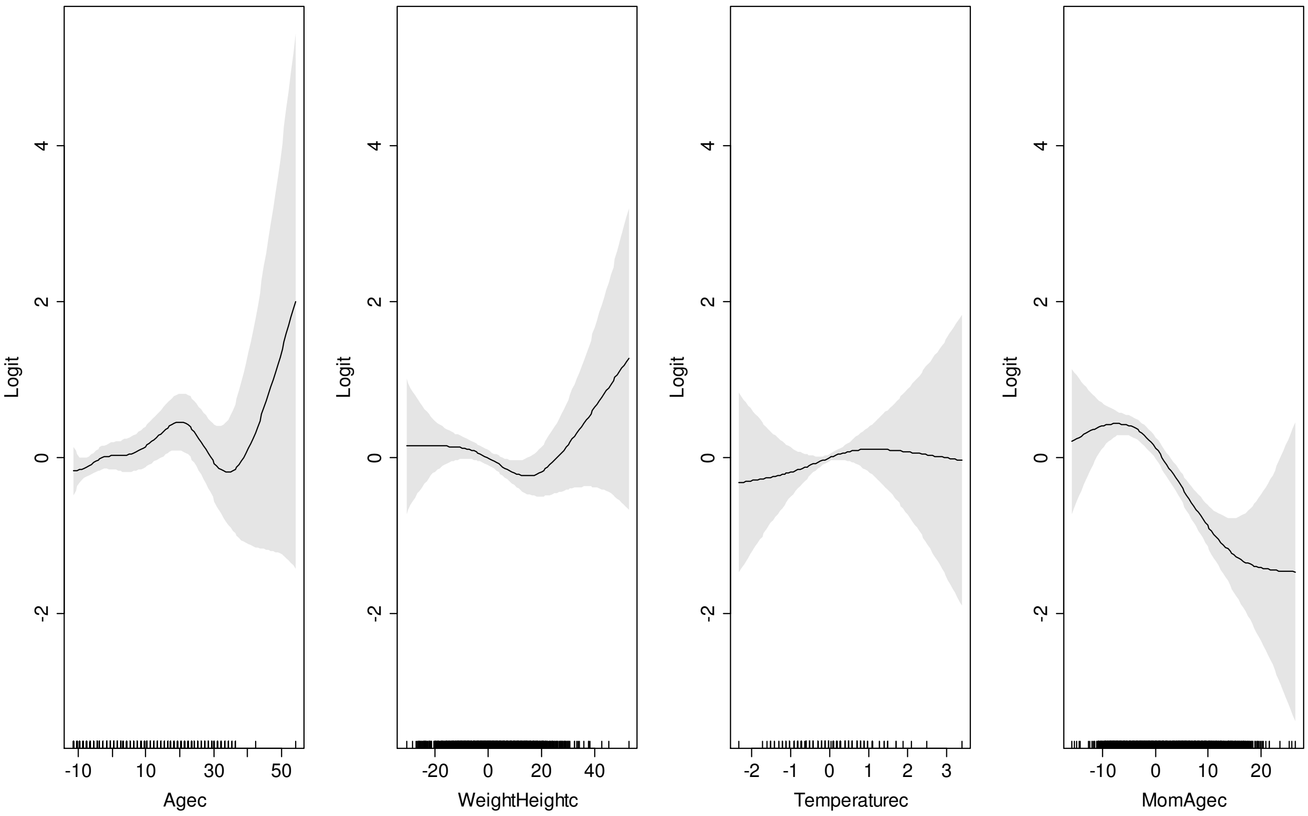


Figure 3.1: *Additive modeling of (centered) metric covariate influence (P-splines with 20 knots). Shaded regions are bayesian pointwise 95% confidence intervals (see Wood (2006)). Additional ticks showing input x-axis values.*

Selection procedure

In order to reduce the number of covariates, an often used approach is variable selection through *AIC* (Akaike Information Criterium) comparison. Three different strategies are usually applied: *Forward*, *Backward* and *Stepwise* selection (cf. Fahrmeir (2013)).

If the number of covariates is moderate, also an exhaustive search can be processed by calculating AICs for all possible models (regarding the combination of covariates). Each of these strategies applied to the Tb data result in the following variable selection: *InfantsDelivered*, *MomEducationLevel*, *MomOccupation*, *HousingType*, *MothersOwn*, *Residence*, *Agec*, *MomAgec* . This corresponds to the p-values listed in Table 3.1, in the sense that these variables represent the ones with the lowest p-values, but are not necessarily significant.

3.1.2 Extended modeling

As it can be seen in chapter 4.2, above main effects regression is already under the best prediction models. This encourages extending the model by 2 approaches: First by additive modeling of effects as already indicated in above chapter, secondly by inserting interactions.

Modeling additive terms

In order to flexibly adapt a possible nonlinear covariate influence, nonparametric regression techniques can be used. Basically a covariate x is therefore replaced by several *basis functions* $B_m(x)$.

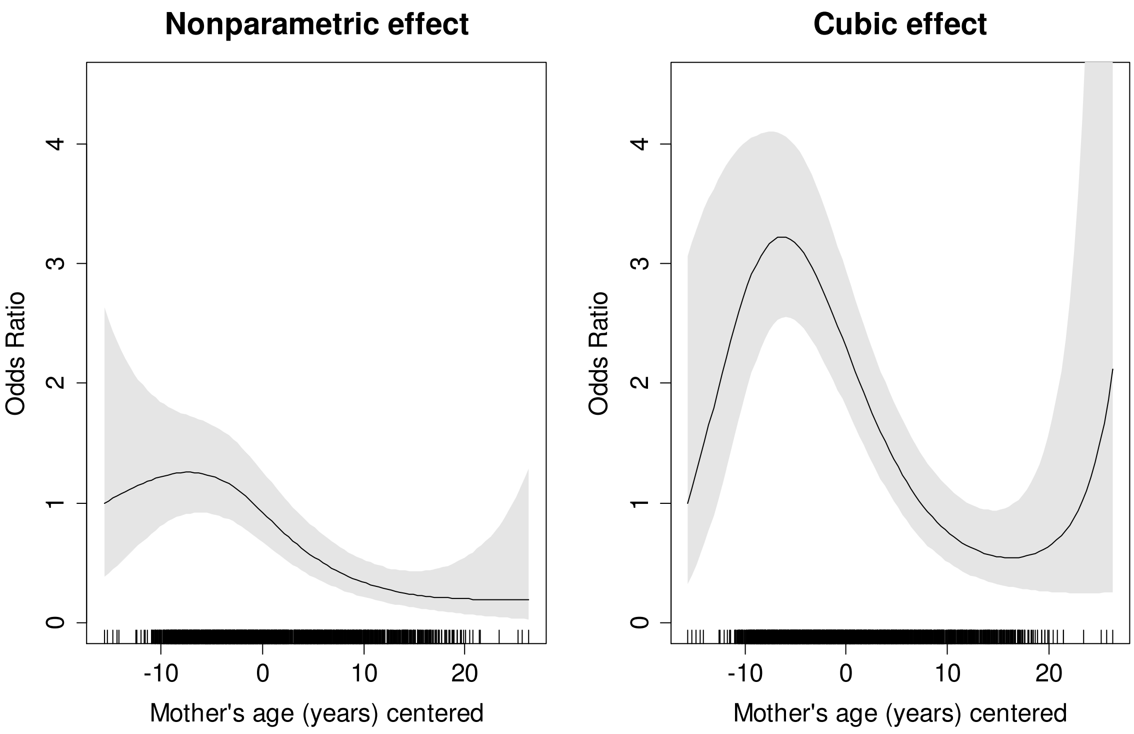


Figure 3.2: *Effects for Mother's age for different modeling approaches.*

Different function "categories" can create this basis extension. So-called *B-Splines* are special polynom pieces, which are connected in a continuous manner. Every piece is associated with one knot from of a set of equidistant knots, distributed over the range of x, in the sense that the spline is non-null only in a local neighborhood of its knot. Usually a high number, e.g. 20, of B-Splines (or knots) is chosen in conjunction with a penalization of the β-parameters in order to prevent overfitting. The penalization parameter can be determined by cross-validation (see also chapter 4.1.2 for an explanation of cross-validation).

The results for this approach, applied to the Tb data, are shown in Figure 3.1. A likelihood ratio test against a linear influence can be executed for each metric covariate in order to decide which influence is not linear. It must be noticed that the resulting p-values are not exact due to several approximations for the test and because of the uncertainty introduced by the estimation of the penalization parameter (see Wood (2006)). The p-values should therefore just be treated as a rule of thumb, in the sense that only highly significant tests indicate nonlinearity. Taking this into account, only *MomAgec* should be modeled non-parametrically.

In order to guarantee asymptotic inference, it can be tried to transfer the additive effect of *MomAgec* into a parametric model. The shape of the effect (see Figure 3.1) suggests a cubic parameterization on the level of the linear predictor. But the difference in the final effects (see Figure 3.2) is too drastic (even though the effective degrees of freedom for the nonparametric effect are just slightly different from 3), so the additive is model is kept.

Modeling interactions

It is also important to test the data for interactions of covariates.

Candidates for interactions were derived from:

- Conspicuous marginal interaction plots (see Figure A.3 and Figure A.4).

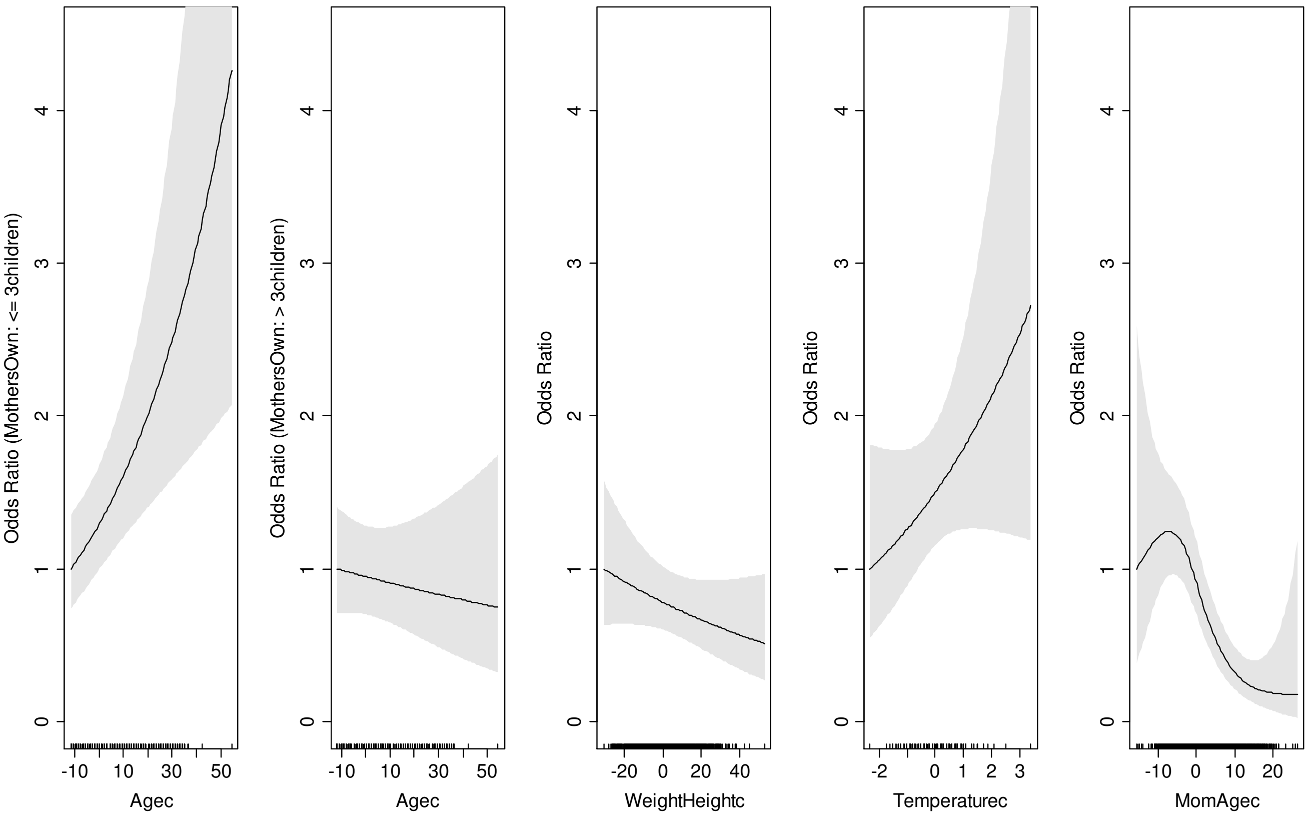


Figure 3.3: *Odds ratio effects of metric variables for final logistic regression model with 95% confidence intervals (pointwise).*

- All combinations of significant variables from the main effects logistic regression, as such variables often show significant interaction effects.
- All combinations of variables of a classification tree (see chapter 3.4), because this method automatically detects interactions.

This ended up in nearly 50 interaction candidates and a likelihood ratio test was conducted for each.

In order to consider the multiple test problem, a Bonferroni adaption of the α-error ($\alpha = 0.05/50 = 0.001$) was made. It must be noticed, that this might not be sufficient to keep the confidence level, as the candidates were derived by kind of “data snooping”. Therefore only interactions with a p-value clearly below 0.001 are taken into account. Furthermore the candidate *PlaceOfBirth:InfantsDelivered* was disregarded because of low importance due to low cell count for *InfantsDelivered*=”Twins”. Eventually the following interactions were included in the final model: *Agec:MothersOwn* + *MomOccupation:Residence*

Final model

As is can be seen in Figure 3.3 (compared to Figure 2.5) Infant’s age become a remarkable effect for mothers with less children (“<=3 children”). Furthermore the decreasing effect of Mother’s age starts not before *MomAgec*=-5, which corresponds to an age of roughly 20.

Regarding the categorical variables, also some remarkable changes occur, due to the interaction of *MomOccupation:Residence*. When comparing Figure 3.4 with Figure 2.6, one can see that the relevant (in terms of cell count) category *MomOccuption*=“Farming” does prevent from LTFU (compared to

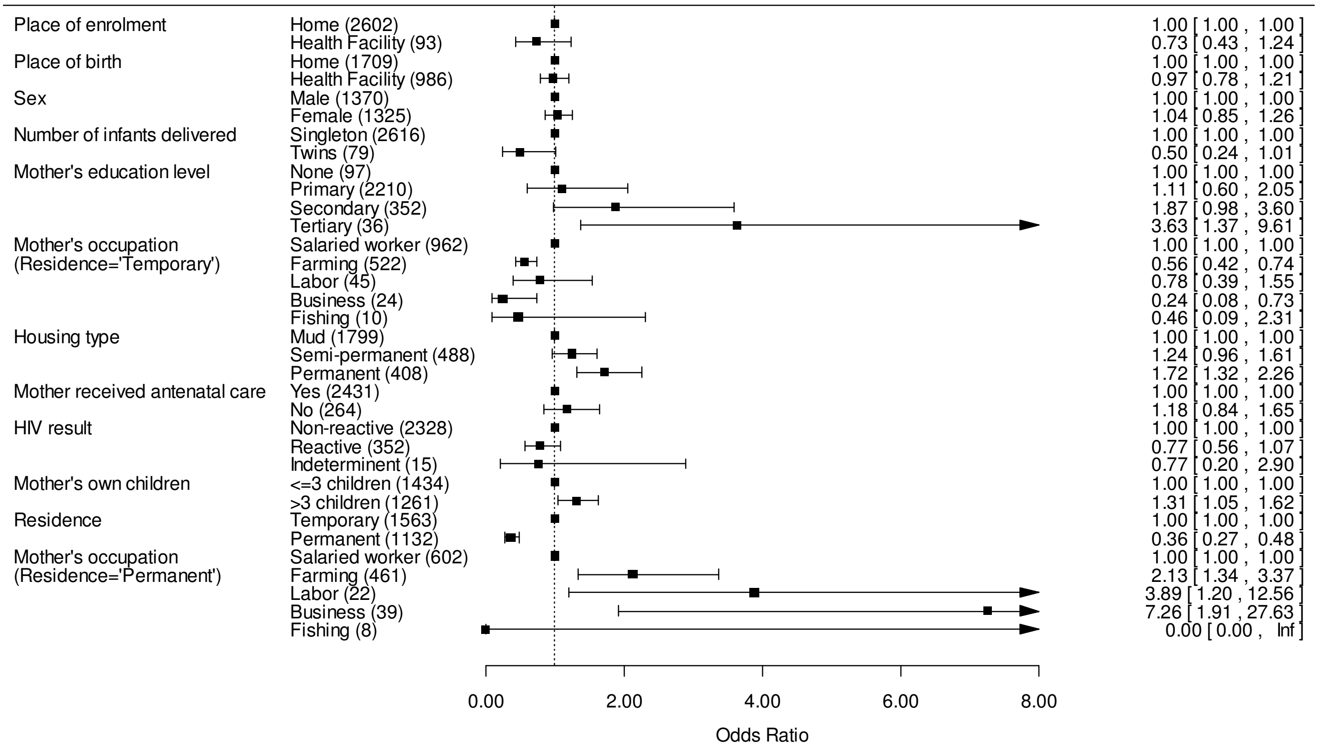


Figure 3.4: *Odds ratio effects of categorical variables for final logistic regression model with 95% confidence intervals.*

MomOccuption="Salaried worker") only for "temporary" mothers whereas for "permanent" mothers this is reversed. Notice that *Residence*="Permanent" itself prevents from LTFU.
Furthermore *MomEducationLevel* ("Tertiary") stays significant, but due to low cell count, it is not "practically significant" in contrast to *HousingType* ("Permanent"). And *MothersOwn* becoming significant is possibly due to the interaction with *Agec*.
A tabular listing of effect values, together with confidence intervals for the final (full) model is attached in appendix A (Figure A.4).
The AIC is remarkably reduced compared to the main effects linear logistic regression model (2611 to 2637). An additional stepwise AIC selection drops variables *PlaceOfEnrolment*, *PlaceOfBirth*, *Sex*, *ReceivedAnteNtlCare* and *HIVResultsAs*. But improvement in AIC is slight (ΔAIC=6). Nonetheless the final assessment of classifier performance in chapter 4.2 is also done with the reduced model (even though this is not crucial for identifying the most important covariates).

Last but no least it should be mentioned that applying this model to complete case data does not yield any relevant changes. Also conducting the model prerequisite tests from chapter 3.1.1 do not result in any new abnormalities with the extended model.

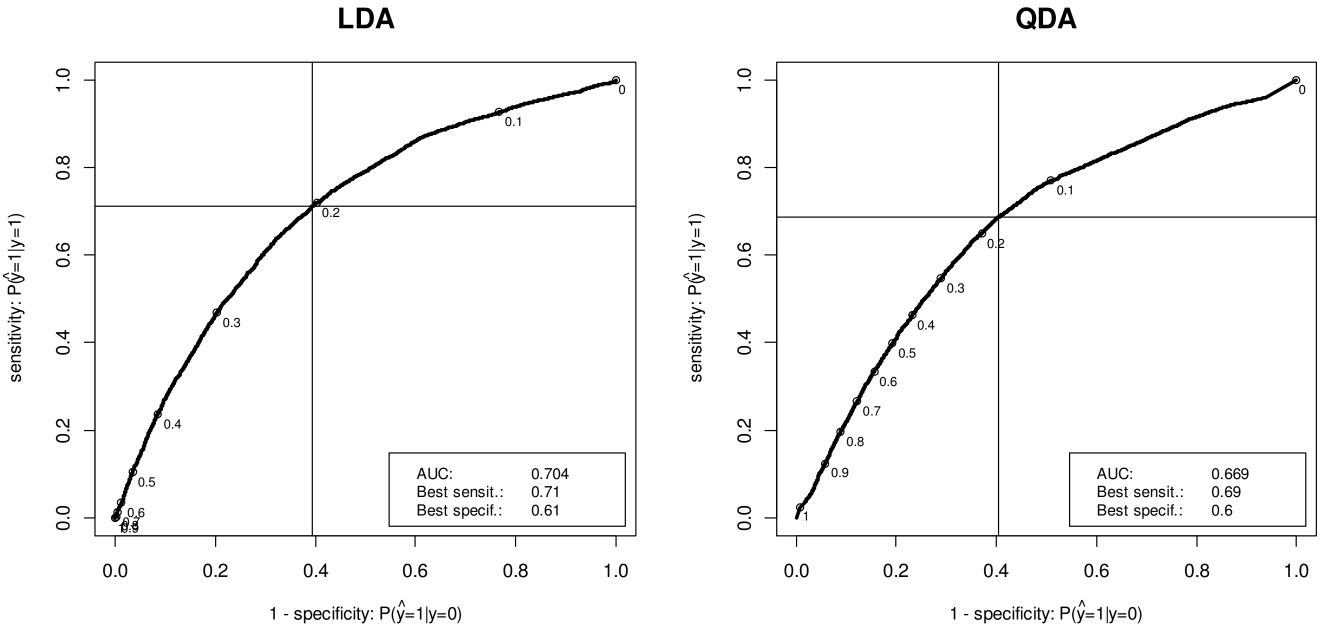


Figure 3.5: *ROC curves (cumulated) for LDA and QDA created by the design discussed in chapter 4.2. Points indicate probability cutoff-values for predicting "Not Retained". Reference lines cross axis at "best" prediction (where sensitivity+specificity is maximized). For a detailed explanation of the ROC curve, see chapter 4.1.1.*

3.2 Discriminant Analysis

In discriminant analysis (cf. Fahrmeir (1996)) the statistical approach is very different to logistic regression, even though both often yield similar results. This can be explained by the fact, that a logistic regression model with logit link can be motivated by linear discriminant analysis (see Tutz (2012)). Basically the method can be summarized as follows:

The estimation of class membership is based on the Bayesian decision rule, i.e. in case of a binary target, choose $y = 1$ if

$$\begin{aligned} p(y = 1|\boldsymbol{x}) &\geq p(y = 0|\boldsymbol{x}) \\ \Leftrightarrow f(\boldsymbol{x}|y = 1)\, p(y = 1) &\geq f(\boldsymbol{x}|y = 0)\, p(y = 0) \\ \Leftrightarrow log\big(f(\boldsymbol{x}|y = 1)\big) + log\big(p(y = 1)\big) &\geq log\big(f(\boldsymbol{x}|y = 0)\big) + log\big(p(y = 0)\big) \end{aligned} \tag{3.2}$$

Here the second row results from Bayes formula and the third from the monotonic character of the log function.

The distribution of $\boldsymbol{x}$ (having dimension p) in each class $k \in \{0; 1\}$ is chosen as multivariate Gaussian:

$$\frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma_k})}} \exp\left(- \frac{\left(\boldsymbol{x} - \boldsymbol{\mu_k}\right)^T \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu_k}\right)}{2} \right) \tag{3.3}$$

Applying this to (3.2) results in the discriminant function which defines the separating hyperplane:

$$\begin{aligned} &-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_1})^T\boldsymbol{\Sigma_1^{-1}}(\boldsymbol{x}-\boldsymbol{\mu_1})-\frac{1}{2}\log(\det(\boldsymbol{\Sigma_1}))+\ log\big(p(y=1)\big) \\ \geq\quad &-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_0})^T\boldsymbol{\Sigma_0^{-1}}(\boldsymbol{x}-\boldsymbol{\mu_0})-\frac{1}{2}\log(\det(\boldsymbol{\Sigma_0}))+\ log\big(p(y=0)\big) \end{aligned} \tag{3.4}$$

Choosing equal covariance matrices for both classes $\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_0}$ results in linear (LDA), different matrices in quadratic discriminant analysis (QDA), as the separating hyperplane is linear in $\boldsymbol{x}$ in the first case and quadratic in the second case respectively (e.g. see Fahrmeir (1996)). The unknown parameters in (3.4) can be estimated by the data.

The same linear decision rule (LDA) also arises from *Fisher's* discriminant approach, which is nonparametric and based on a heuristic argument: minimizing intra-class variance and simultaneously maximize inter-class variance. This fact theoretically supports the experience, that the linear discriminant analysis is robust against violation of the Gaussian distribution assumption (which is obvious for categorical covariates). Hastie (2009) suggests giving these two simple techniques (LDA and QDA) always a try, as they perform very well in classification problems. This might be due to stable simple decision boundaries based on Gaussian models, which implies less variance. The latter is less valid for QDA and might be the reason why QDA, although it is more flexible, is often outperformed by LDA (cf. Tutz (2012)). This also holds for the analyzed Tb data (see Figure 3.5).

3.3 Lasso Regularization

Regularization or *shrinkage* methods can function as an alternative to classical variable selection (cf. Tutz (2012)). The main idea here is to penalize the log-likelihood $l(\boldsymbol{\beta})$ of the regression problem:

$$l_p(\boldsymbol{\beta}) = \sum\nolimits_{i=1}^{n} l(\boldsymbol{\beta}) - \frac{\lambda}{2} J(\boldsymbol{\beta}) \tag{3.5}$$

Using $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2$ results in the well-known *ridge* penalization, which was originally invented to cope with singularity problems in linear regression estimation. This yields smoothly shrinked parameters β_j (depending on the shrinkage parameter λ).

But when using the *lasso* penalty $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$ an implicit variable selection can occur, as shown in Figure 3.6, which presents the solution of equation (3.5) for different values of λ. By selecting just the variables which have non-zero coefficients for the tuning parameter λ chosen by 10-fold cross-validation (see also chapter 4.1.2), some covariates are dropped. Obviously the solution of (3.5) depends on the scaling of the covariates. Therefore the variables are standardized before estimation. This has the positive side effect, that the plot of lasso paths provide a useful visual tool to assess the importance of the influential variables, as their β's are directly comparable.

It can be seen, that in order of importance Mother's age, Residence, Housing type ("Permanent"), Mother's education level ("Secondary" and "Tertiary") and Mother's occupation ("Farming") are the Top 5, regarding influence. Interestingly, the coefficient for *MomEducationLevel*="Primary" decreases to 0 after first being the most important category of mother's education level.

Other penalties can be used. E.g. the *adaptive lasso* uses $J(\boldsymbol{\beta}) = \sum_{j=1}^{p} w_j |\beta_j|$ (with $w_j = 1/|\widehat{\beta}_j|^{\nu}$ and $\widehat{\beta}_j$ least square estimate). The advantage of this more complicated penalty is, that it yields consistent β_j-estimates (cf. Hastie (2009)). The resulting paths in dependence of the shrinkage parameter λ (instead of shrinkage factor $||\beta||/max||\beta||$ like in Figure 3.6) are compared with ordinary lasso in Figure 3.7. There is no obvious difference, which in this case supports the assumption of consistent estimates, even with

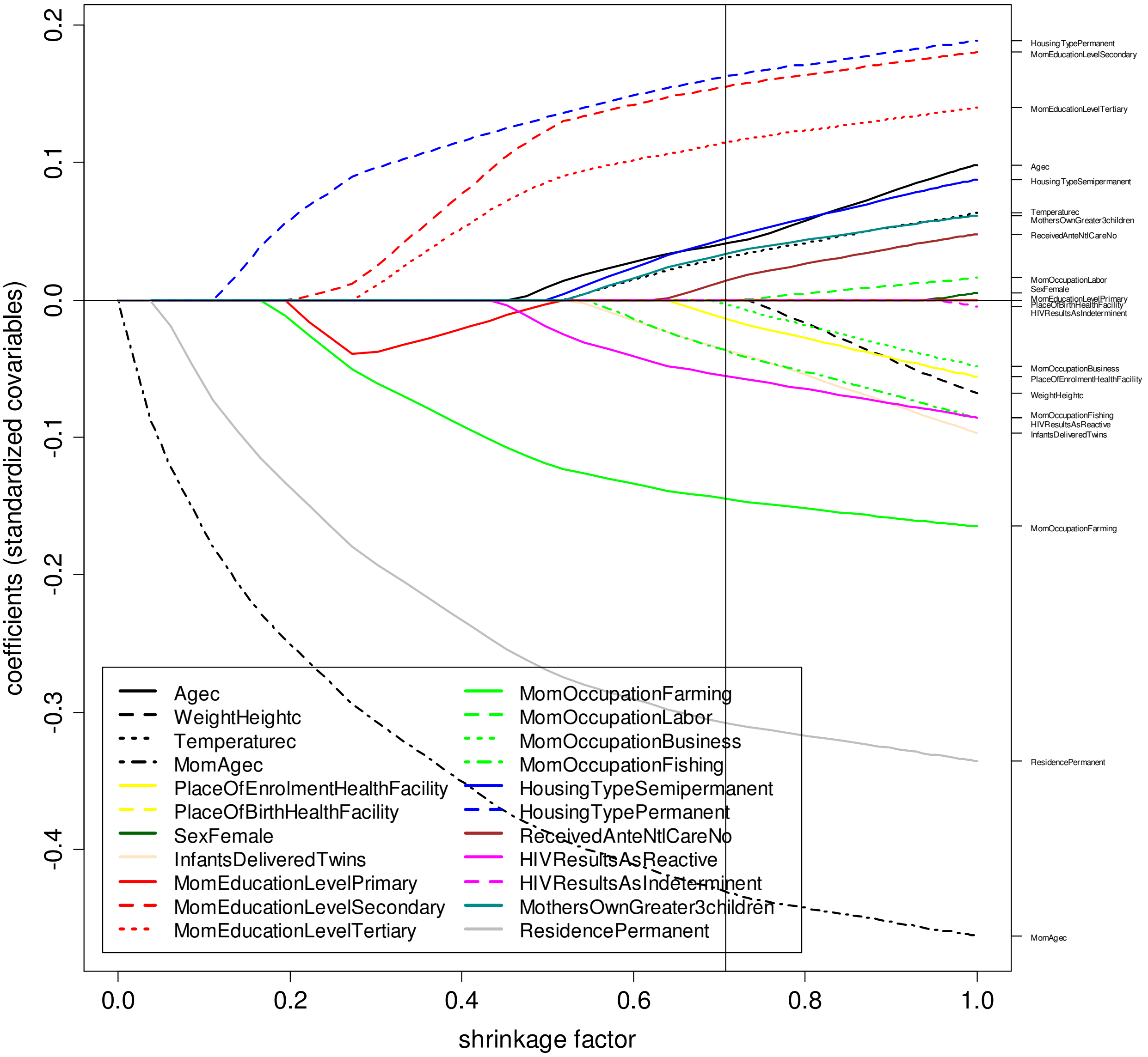


Figure 3.6*: Lasso paths for Tb data. Metric variables are drawn in black. Categories belonging to same categorical variable have same color. On the x-axis the shrinkage factor $\|\beta\|/max\|\beta\|$ is listed, rather than the shrinkage parameter λ (both are linked in a non-linear way). The vertical axis indicates the values of the coefficients for standardized covariables, which are therefore directly comparable in terms of importance. Vertical reference line shows optimal shrinkage, resulting from 10-fold cross-validation.*

the standard lasso. Furthermore this plot shows that the visual impression is highly dependent on whether the shrinkage parameter or factor is listed on the x-axis.

The lasso also explicitly selects and therefore splits off the important categories of a nominal variable. This might not be the wished behavior if one wants to keep all categories, in terms of originating equally from one covariate. The *group lasso* can accomplish this by using another penalty, which encourages sparsity by favoring either $\beta_{js} \neq 0$ or $\beta_{js} = 0$ for all categories s=1..k of covariable x_j. As it is shown in Figure 3.8, this results in keeping categories of the same variable “together”. Dependent on the optimal shrinkage factor, this might lead to changed importance interpretation of specific categories, which is not the case in a crucial manner for the Tb data (see Figure 3.8).

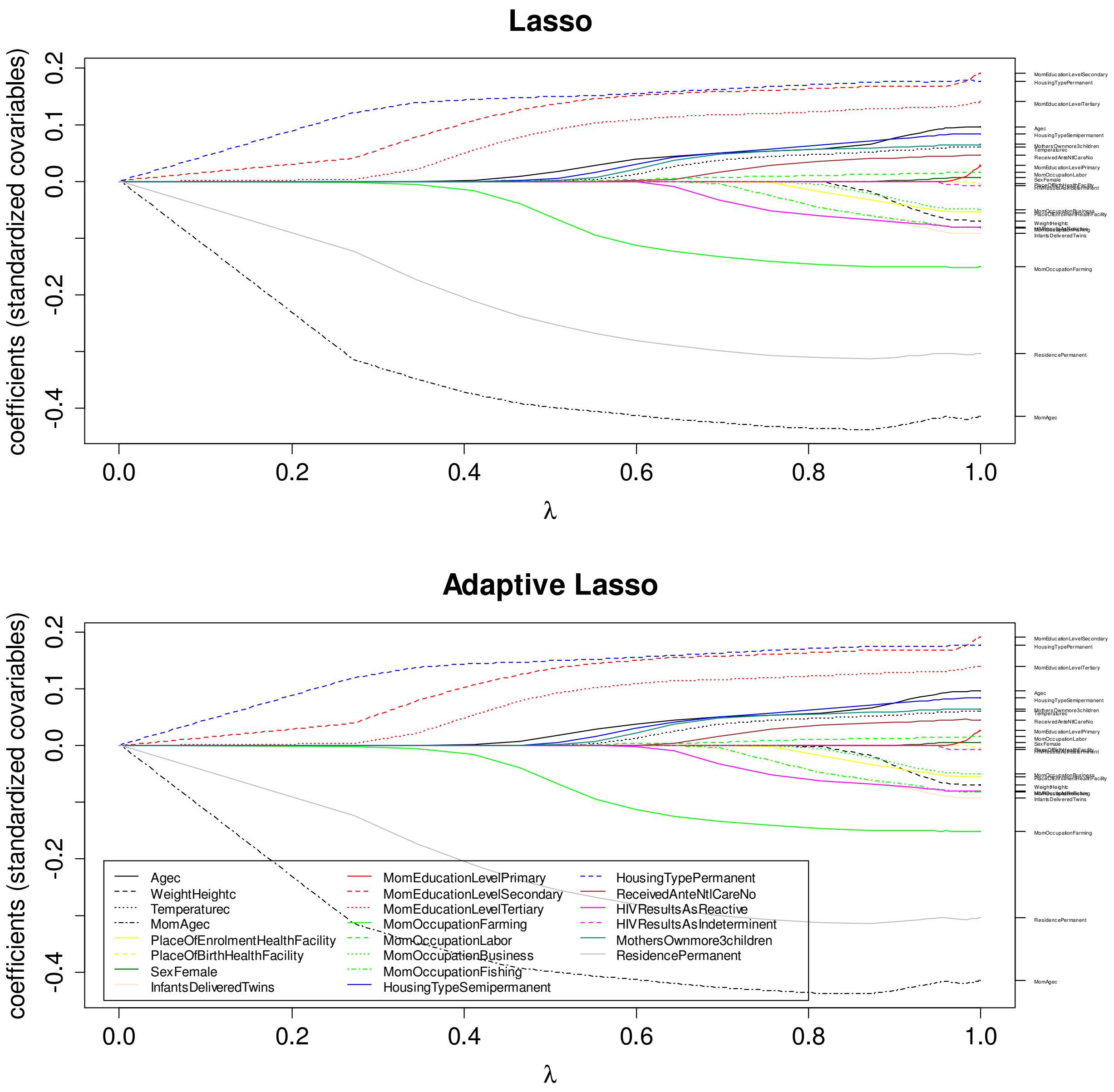


Figure 3.7*: Comparison of lasso and adaptive lasso paths for Tb data. Presentation differs to Figure 3.6 (x-axis indicate shrinkage parameter λ and not shrinkage factor $\|\beta\|/max\|\beta\|$)*

At last is should be noted, that the *lasso* has limitations in case of highly correlated predictors. It then tends to select just one of the correlated variables as a representative (cf. Hastie (2009)).
This can be circumvented by using a penalty which is “between” *lasso* and *ridge* and called *elastic net*: $J(\boldsymbol{\beta}) = (1-\alpha)\sum_{j=1}^{p}|\beta_j| + \alpha\sum_{j=1}^{p}\beta_j^2$. Regression with this penalty shows the grouping effect, i.e. coefficients of highly correlated variables tend to be equal. It should just be noticed, that no grouping effect occurred when testing different values of α between 0 and 1 for the Tb data, which also corresponds to lack of collinearity (cf. chapter 3.1.1). Thus, together with the results of the adaptive lasso comparison and the group lasso this confirms above mentioned Top 5 “consistent” rates of the ordinary lasso.

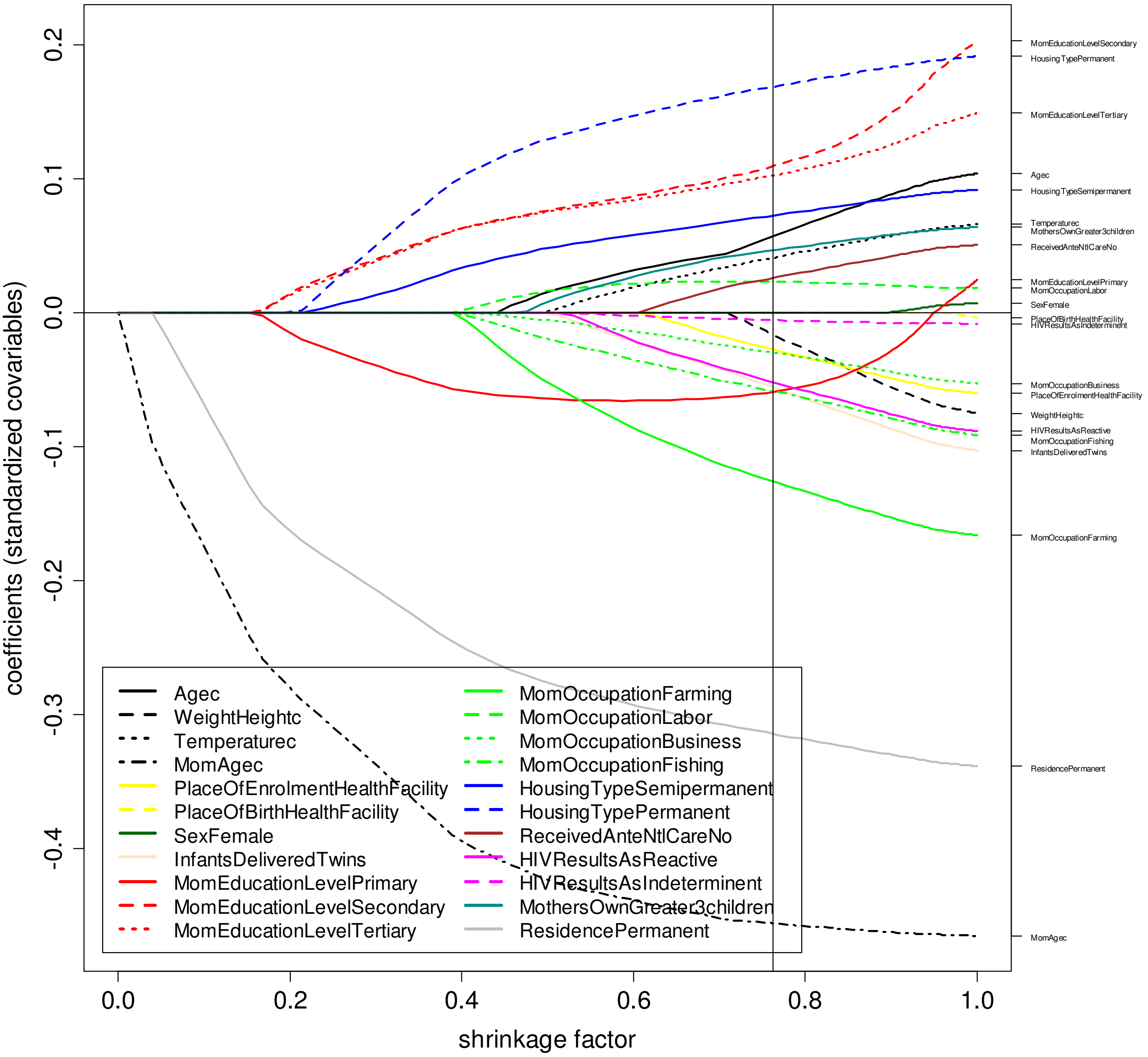


Figure 3.8: *Group-lasso paths. Presentation is equivalent to Figure 3.6.*

3.4 CART (Classification and Regression Tree)

In CARTs (see Tutz (2012)) a series of binary data splits segments the data into a tree, which provides an intuitive instrument to understand the classification process. At each split a criterion is evaluated for all covariates and corresponding possible value partitioning. The best covariate-partitioning combination is finally used to define the split. For binary classification problems (i.e. classification trees) the split-criterion is usually based on minimizing an impurity measure like the Gini index (2p(1-p), with p equal to the proportion of one class), deviance (-plog(p)-(1-p)log(1-p)) or misclassification error (1-max(p,1-p)) (see e.g. Hastie (2009)). This process can be stopped if a node has fewer observations than a defined value, or if the splitting criterion is above or below a threshold. But usually a tree is fully grown and pruned again due a complexity criterion, which can be optimized by cross-validation.

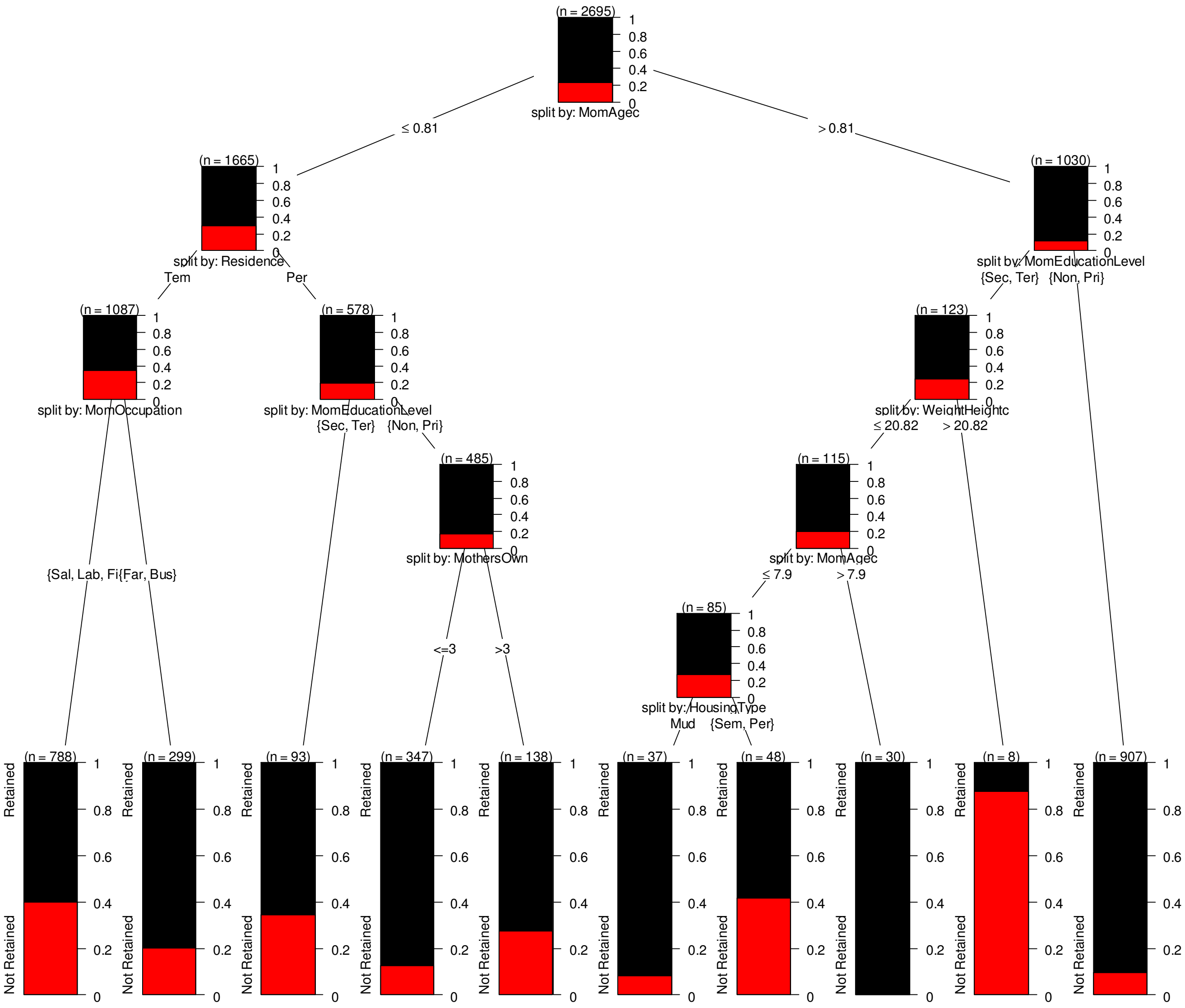


Figure 3.9: C*ART based on conditional inference framework. Category labels are abbreviated (but might still overlap). P-values are not shown: They are ≤0.001 for all splits except for the second MomAgec (0.044) and the HousingType (0.04) splits*

Alternatively the split can be based on conditional inference, where the association of each covariate to the target is measured by a p-value, corresponding to a test for the null hypothesis of independence between covariate and target (see help for *ctree*-function in R-package *party*). Stopping is then due to non-significance of a global independence hypothesis test. This approach circumvents the problem of selection bias, which occurs for example for binary outcomes when using the Gini index as split criterion, as variables with lower number of categories are preferred (see Tutz (2012)).

A general advantage of CARTs is the automatic detection of interactions, which naturally results from consecutive splits of different variables. This is also used to suggest possible interactions for the regression approach in chapter 3.1.2. Furthermore important predictors are automatically selected. For the Tb data a tree based on the conditional independence framework was build (see Figure 3.9) and Mother's age, Residence, Mother's education level, Mother's occupation, Weight:height ratio, Mother's own and Housing type result as influential covariates.

One disadvantage of CARTs is the instability because of small changes in data might result in a very different series of splits. The high variance of CARTs AUC metric of for the Tb data shown in chapter 4.2 is a direct consequence of this issue.
When looking at the tree terminal nodes in Figure 3.9, the most promising ones, in terms of getting a low LTFU rate together with a high enrolment count, are node number 4 and 10 (counted from the left),. A possible strategy suggested by these nodes could be: Choose older (than the average) mothers with low education level; for younger mothers the same holds but just for “permanent” (*Residence*) mothers. Also for younger mothers the rule would extend to: the fewer children they have the better. For the actual Tb data this would result in a total count of 1254 patients with a LTFU below 10%.

3.5 Boosting

3.5.1 Boosting with Trees

Boosting (cf. Hastie (2009)) is, like *Random Forests* (see chapter 3.6), an ensemble method, which means that several classifiers vote for the predicted class. It originates in the data mining field; its most popular member is *AdaBoost*, where basically a weak classifier, i.e. one that is just a little better than guessing, is repeatedly processed on the training data, which itself is adapted at each run. The misclassified observations in each run then get a higher weight in the following run. Finally the prediction is a weighted majority vote of the classifiers of all runs, with the weight proportional to the misclassification rate for the whole training data.

It can be shown, that AdaBoost can also be approximated by means of a loss function

$$L(f) = \sum_{i=1}^{N} L\big(y_i, f(\boldsymbol{x}_i)\big), \tag{3.6}$$

which is minimized with respect to the predicting function *f*. In case of using CARTs as a classifier, *f* can be expressed as a sum of trees: $f(\boldsymbol{x}) = \sum_{m=1}^{M} T(\boldsymbol{x}; \boldsymbol{\theta}_m)$ (with $\boldsymbol{\theta}_m$ denoting the parameters of the tree), and the minimization is reached in a forward stagewise manner, i.e. by repeatedly solving

$$\widehat{\boldsymbol{\theta}}_m = arg \min_{\boldsymbol{\theta}_m} \sum_{i=1}^{N} L\left(y_i, f_{m-1}(\boldsymbol{x}_i) + T\big(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}_m\big)\right) \tag{3.7}$$

yielding $f_m(\boldsymbol{x}) = f_{m-1}(\boldsymbol{x}) + T\big(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_m\big)$ in each step.

The loss function for the AdaBoost is the exponential loss $L(y, f) = e^{-yf}$, with the binary target *y* coded as $\{-1; 1\}$, like it is usual in the machine learning field.
Alternatively the minimization of (3.6) can be solved with *Gradient Boosting*. With this approach just pseudo residuals

$$r_{im} = -\left[\frac{\partial L\big(y_i, f(\boldsymbol{x}_i)\big)}{\partial f(\boldsymbol{x}_i)}\right]_{f=f_{m-1}} \tag{3.8}$$

are fitted by $T\big(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_m\big)$ in each step and the predicting function is updated using a learning rate 0<ν<1 by

$$f_m(\boldsymbol{x}) = f_{m-1}(\boldsymbol{x}) + \nu * T(\boldsymbol{x}; \boldsymbol{\theta}_m) \tag{3.9}$$

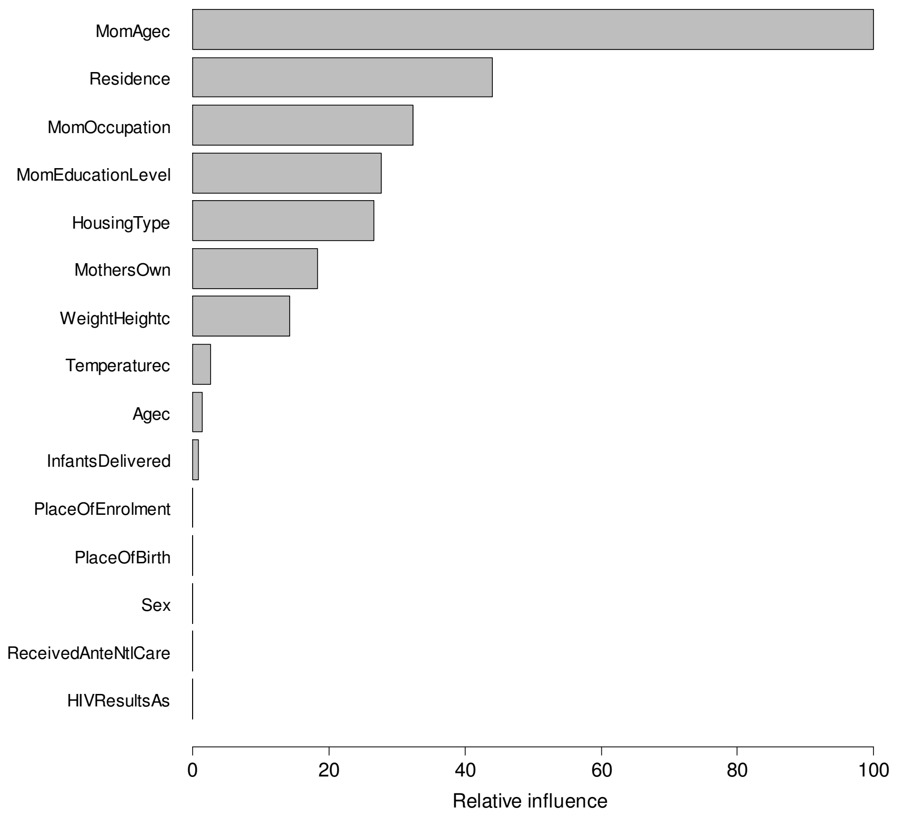


Figure 3.10: *Relative influence of variables in Tree-Boosting. The most important predictor (MomAgec) is scaled to 100.*

The *learning* parameter ν should be sufficiently small, which prevents from overfitting. $T(\boldsymbol{x};\boldsymbol{\theta}_m)$ is then also denoted as a *weak base learner*.

The loss function and the base learner determine the type of Boosting. Alternative loss function to the exponential loss e^{-yf} are e.g. binomial deviance: $log\left(1+e^{-2yf}\right)$, squared error: $(y-f)^2$ or hinge loss: $(1-yf)_+$. As already noted, the exponential loss exactly yields the AdaBoost. But from a statistical perspective this is not the ideal loss function. Using e.g. binomial deviance loss might improve the original AdaBoost in case of noisy settings (see Hastie (2009)), as it represents a more robust loss, which gives outliers less weight. This loss function is also used for all Boosting algorithms applied to the Tb data. In contrast, the hinge loss is used in conjunction with Support Vector Machines (see also chapter 3.7).

Unfortunately the final model is a black box, which means that the influence of a single covariate cannot be directly rated. But a variable importance plot can be derived in the following way: For each variable the sum of the improvements, regarding the impurity criterion, in each node where the interesting variable is used, can be calculated. Averaging over all trees used in Boosting, provides a relative importance metric. Figure 3.10 shows this for the Tb data (using trees with a maximum of 6 terminal nodes). It can be seen that *MomAgec* is by far the most important variable, followed by *Residence*, *MomOccupation*, *MomEducationLevel* and *HousingType*, completing the Top 5.

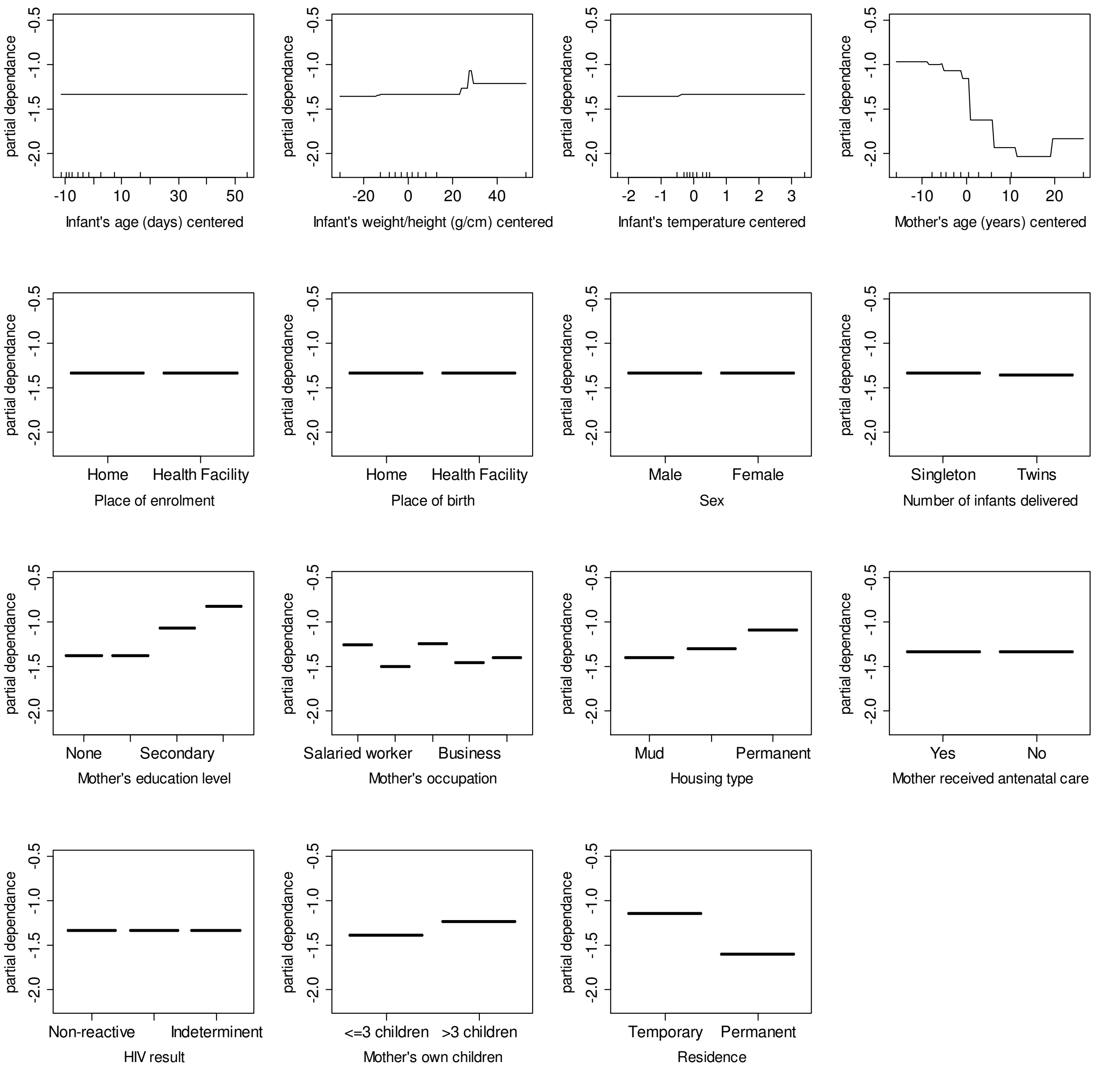


Figure 3.11: *Tree-Boosting: Partial dependence plots. Ticks are showing input x-axis deciles for metric variables.*

But this still hides how e.g. the course of the influence of *MomAgec* is, or which categories of a nominal covariate are important. For this problem a generic approach exists (see Hastie (2009)), which can be applied to any black box. The univariate (but controlled for other covariates) partial dependence of the classifier $f(\mathbf{X})$ on the kth covariate X_k can be defined as the expected value $E_{X_{\setminus k}} f(X_k, \mathbf{X}_{\setminus k})$ with $\mathbf{X}_{\setminus k}$ denoting the covariable vector without the kth variable.This metric can be estimated at every value of X_k by $\frac{1}{N}\sum_{i=1}^{N} f\left(X_k, \mathbf{x}_{\setminus k_i}\right)$. Usually this requires a high number of fitting processes and is computationally intensive (except for Tree-Boosting, see Hastie (2009)). For the Tb data an appropriate plot is presented in Figure 3.11. Interestingly, the dependencies actually correspond to the logistic regression model (see chapter 3.1.2), most striking for Mother's age (cf. Figure 3.3).

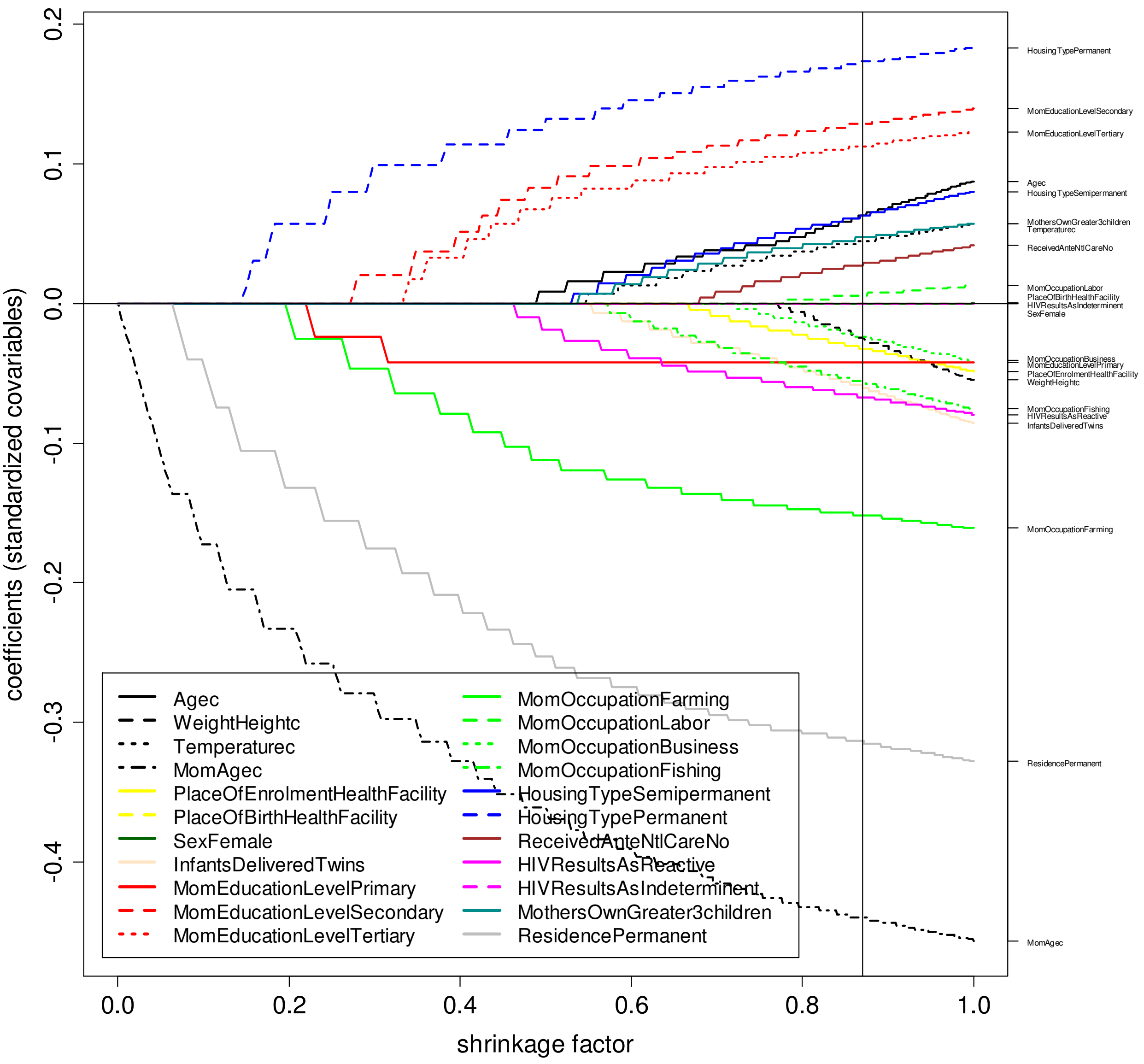


Figure 3.12: *GLM-Boosting paths. Vertical line shows optimal shrinkage due to 10-fold cross-validation.*

3.5.2 Boosting with Regression Models

The above mentioned squared error loss is used in case of quantitative targets and, together with Gradient Boosting and a linear predictor as base learner, just results in iteratively fitting of the regression residuals. For binary targets (with sufficient cell counts) this can be adapted by applying the logit transformation to the target, resulting in the *LogitBoost* (cf. Tutz (2012)).

For the latter a more generalized approach exists, the *likelihood Boosting*. This technique allows extending the boosting idea, which basically is forward stagewise fitting with weak base learners, to all generalized regression models. Here the weak learner just updates an offset of the linear predictor in

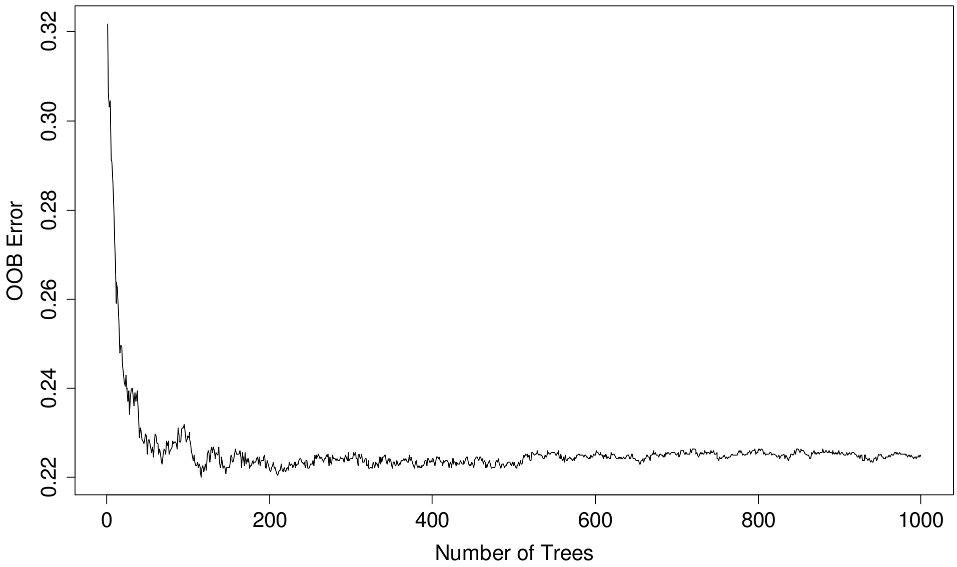


Figure 3.13: *OOB (out-of-bag) error depending on number of trees for a Random Forest on Tb data.*

each Boosting step, which is therefore a standard generalized regression fit including a constant/offset (see Tutz (2012)): $\mu_i = h\left(\hat{\eta}^{(l)}(\boldsymbol{x}_i) + \eta(\boldsymbol{x}_i, \gamma)\right)$, with $\hat{\eta}^{(l)}(\boldsymbol{x}_i)$ as predictor from previous step. Interestingly, if the fitting step just updates the coefficient of the “best” covariate, an implicit selection procedure, similar to the lasso regularization (see chapter 3.3) results. Moreover, it can be shown (see Hastie (2009)), that the resulting Boosting path equals the lasso path in case all lasso coefficients increase monotonically (as Boosting paths are monotone due to construction). This can be seen in Figure 3.12 which shows the Boosting paths for the Tb data. The paths are very similar to the lasso of Figure 3.6, except for the categories of *MomEducationLevel*, as the lasso path for “Primary” is not monotonic.

3.6 Random Forest

CARTs (see chapter 3.4) can be extended to Random Forests (cf. Hastie (2009)), which represent an ensemble method, i.e. several trees, which vote for the most popular class, are processed. This helps in coping with the already mentioned high variance of trees and makes Random Forest one the best predictors in many studies (see Tutz (2012)). Basically the Random Forest algorithm is as follows: Repeatedly draw bootstrap samples of the data. Then build a tree on each sample, consisting of the best variable-split-point combination of m , at each split point, randomly (!) chosen predictor variables out of the p covariates. In case of a binary target, the final prediction is the majority vote from the trees. If a probability value for the target event is needed, the average over the individual class votes can be taken.

The number m of predictors in each tree is a tuning parameter, but is often chosen according to the recommendation of $\sqrt{p}$ in classification problems (cf. Hastie (2009)). The number of trees must also be tuned, but because Random Forests seldom overfit as a result of the tree-number (see Hastie (2009)), it is adequate to use a sufficient high number of trees. This number can be derived by applying the method just once on the whole dataset, as Random Forests have kind of a build-in cross-validation by

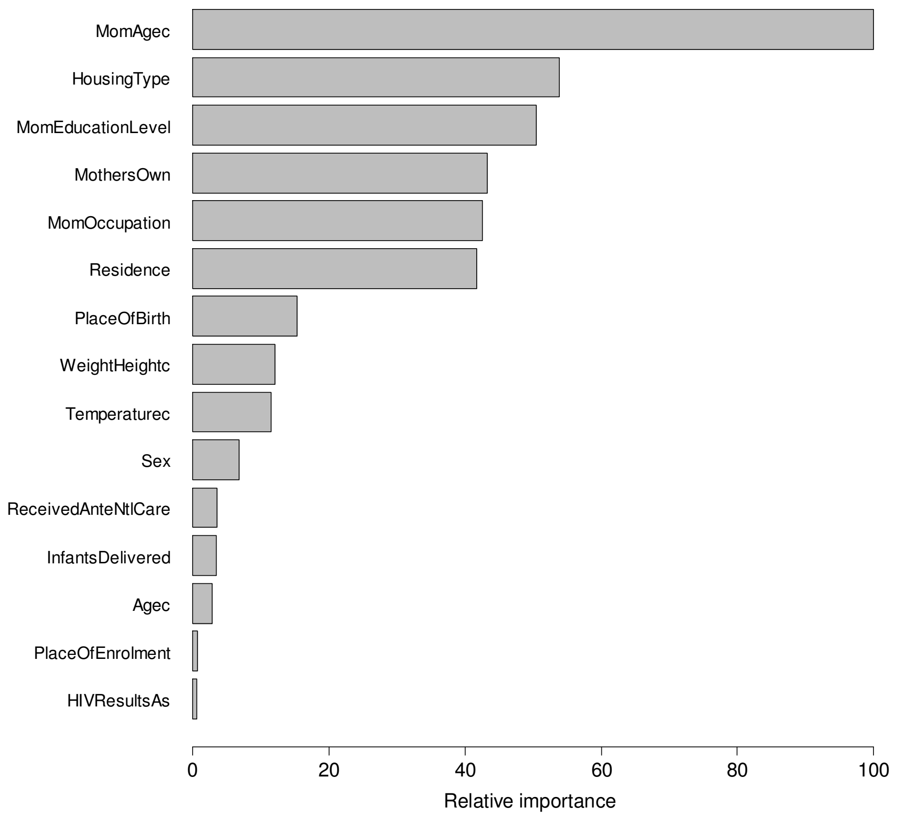


Figure 3.14: *Random forest importance plot. The most important predictor (MomAgec) is scaled to 100.*

calculating the *OOB* (*out-of-bag*) error on the bootstrap samples (see Hastie (2009)). E.g. see Figure 3.13, which shows that the method stabilizes at roughly 200 trees. Therefore a value of 500 might be used for the measurement of classifier performance (see chapter 4.2).

Unfortunately the influence of the individual predictors gets lost. But a variable importance plot can be drawn, in the same way as for Boosting models (see chapter 3.5), and is presented in Figure 3.14.
It shows that Mother's age is again the most important variable. Then a group with *HousingType*, *MomEducationLevel*, *MothersOwn*, *MomOccupation* and *Residence* follows with each having approximately "half the importance of Mother's age". It must be stated that the internal ordering for this group, regarding importance, might change for a second run with different bootstrap samples.

3.7 Support Vector Machine

The *Support Vector Machine* (SVM) classifier (cf. Hastie (2009)) originated from the machine learning community. The idea behind is to maximize the feature space margin between two classes (see Figure 3.15). It can be shown that the signed distance of any point x_i to a separating hyperplane defined by $\boldsymbol{x}'\boldsymbol{\beta} + \beta_0 = 0$ is $\frac{1}{||\boldsymbol{\beta}||}(\boldsymbol{x}_i'\boldsymbol{\beta} + \beta_0)$, see Hastie (2009). If the target is coded as $y_i \in \{-1,1\}$, with $y_i = 1$ having positive distance from the hyperplane (or identifying the interesting event respectively), the idea results in the maximization problem:

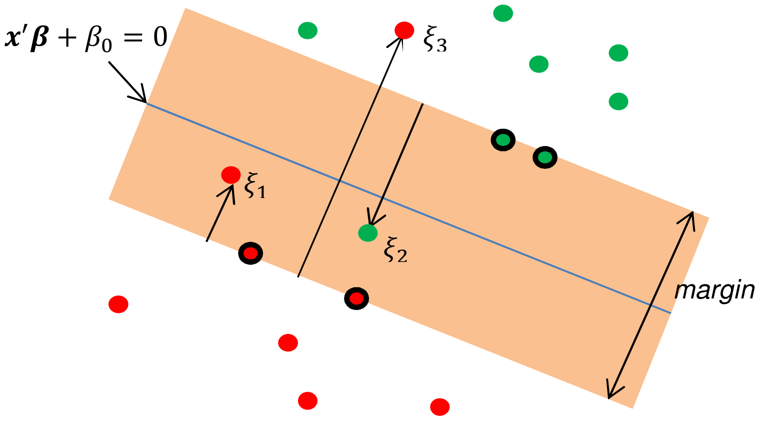


Figure 3.15: *Idea of SVM including slack variables. Support vectors are marked with black border. Separating hyperplane lies in the middle of the margin*

$$\max_{\boldsymbol{\beta},\beta_0} \left\{ y_i \frac{1}{\|\boldsymbol{\beta}\|} (\boldsymbol{x}_i'\boldsymbol{\beta} + \beta_0) \right\} \tag{3.10}$$

This can be equivalently formulated as:

$$\min_{\boldsymbol{\beta},\beta_0} \|\boldsymbol{\beta}\| \text{ , } with\ constraint: y_i(\boldsymbol{x}_i'\boldsymbol{\beta} + \beta_0) \geq const. > 0 \tag{3.11}$$

Choosing the constant arbitrarily as 1 and because of $\|\boldsymbol{\beta}\| > 0$, an equivalent formulation is:

$$\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2} \|\boldsymbol{\beta}\|^2 \text{ , } with\ constraint: y_i(\boldsymbol{x}_i'\boldsymbol{\beta} + \beta_0) \geq 1 \tag{3.12}$$

Usually the data is not completely separable in this way, as there are always points on the false side of the hyperplane. Therefore so-called *slack* variables are invented and the minimization problem (3.12) is relaxed in order to allow some overlap:

$$\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2} \|\boldsymbol{\beta}\|^2, \ with\ constraints: y_i(\boldsymbol{x}_i'\boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i \text{ , } \xi_i > 0, \sum \xi_i \leq const. \tag{3.13}$$

This is usually written in the form:

$$\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2} \|\boldsymbol{\beta}\|^2 + C \sum_{i=1}^{N} \xi_i \text{ , } with\ constraints: y_i(\boldsymbol{x}_i'\boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i \text{ , } \xi_i > 0 \tag{3.14}$$

C is the “cost” parameter and influences how large the separating margin is. Therefore it is a tuning parameter, which is usually derived by cross-validation.

Equation (3.14) represents a convex minimization problem, i.e. it has one and only one solution. As the constraints imply inequalities, the Lagrangian function, which has to be minimized, is accompanied by so-called *Karush-Kahn-Tucker* conditions (e.g. see Bishop (2006)). One of the derivatives of the Lagrangian, together with the Karush-Kahn-Tucker conditions, results in $\boldsymbol{\beta} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$, with nonzero α_i only for the observations for which the constraint is exactly met, i.e. the points lying on the boundary of the margin. So $\boldsymbol{\beta}$ is determined exclusively by these support vectors, which explains the name of the method. This also means that points, lying inside the correct region, do not have an influence on the separation!

The whole approach can be extended by enlarging the feature space with basis expansions like splines. Basically this corresponds to the additive modeling of the linear predictor in regression problems (see also chapter 3.1.2). So $\boldsymbol{x}_i$ is replaced by a set of basis functions $(\boldsymbol{x_i}) \to \big(h_1(\boldsymbol{x}_i), \ldots, h_M(\boldsymbol{x}_i)\big)$. The separating hyperplane can then be written as

$$f(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x})'\boldsymbol{\beta} + \beta_0 = \sum_{i=1}^{N} \alpha_i y_i\, \boldsymbol{h}(\boldsymbol{x})'\boldsymbol{h}(\boldsymbol{x_i}) + \beta_0 \tag{3.15}$$

Thus the solution just depends on the so-called *kernel* $\boldsymbol{K}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \boldsymbol{h}(\boldsymbol{x})'\boldsymbol{h}(\widetilde{\boldsymbol{x}})$ which can be computed very cheaply for a special choice of $\boldsymbol{h}$. Popular members are the radial kernel: $\boldsymbol{K}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = exp(-\gamma \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|^2)$ (which is also used for measuring the SVM performance for the Tb data in chapter 4.2) or the dth-degree polynomial $\boldsymbol{K}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = (1 + \boldsymbol{x}'\widetilde{\boldsymbol{x}})^{\boldsymbol{d}}$. For d=2 the corresponding basis functions are $h_1(\boldsymbol{x}) = 1, h_2(\boldsymbol{x}) = \sqrt{2}x_1,$ $h_3(\boldsymbol{x}) = \sqrt{2}x_2,\ h_4(\boldsymbol{x}) = x_1^2,\ h_5(\boldsymbol{x}) = x_2^2,\ h_6(\boldsymbol{x}) = \sqrt{2}x_1x_2$ for instance.

Remarkably the SVM can also be expressed as a penalization method (see Hastie (2009)). It can be shown that

$$\min_{\boldsymbol{\beta}, \beta_0} \sum_{i=1}^{N} [1 - y_i f(\boldsymbol{x_i})]_+ + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \tag{3.16}$$

(with subscript “+” indicating the positive part) has the same solution as (3.14). So, from a technical aspect it is the same as the regularization methods from chapter 3.3 with a different loss function called the hinge loss: $(y, f) = [1 - yf]_+$.

3.8 Neural Net

Originally *Neural Networks* (cf. Hastie (2009) and Tutz (2012)) tried to mimic the function of the human brain. In Figure 3.16 each unit represent a neuron and the connections between them stand for the synapses. Technically they are just nonlinear statistical models. This is explained in the following for the single hidden layer Neural Net.

If the hidden units z_l are given by a nonlinear transformation (or activation function) $z_l = \phi\Big(\alpha_{l0} + \sum_{j=1}^{p} \alpha_{lj} x_j\Big)$, as well as the output unit $= \tilde{\phi}\big(w_0 + \sum_{l=1}^{k} w_l z_l\big)$, the combination of these terms yield:

$$y = \tilde{\phi}\left(w_0 + \sum_{l=1}^{k} w_l\, \phi\Big(\alpha_{l0} + \sum_{j=1}^{p} \alpha_{lj} x_j\Big)\right) \tag{3.17}$$

In the neural net context the intercepts α_{l0} and w_0 are called “biases” and the parameters α_{lj} and w_l “weights”. For the activation functions ϕ and $\tilde{\phi}$ usually the *sigmoid* (or *logistic*) function $\phi(a) = 1/(1 + e^{-a})$ is used. Alternatives are sometimes Gaussian radial basis functions (see Tutz (2012)). Notice that with choosing $\tilde{\phi}$ as sigmoid and ϕ as the identity, just an ordinary logistic regression model results.

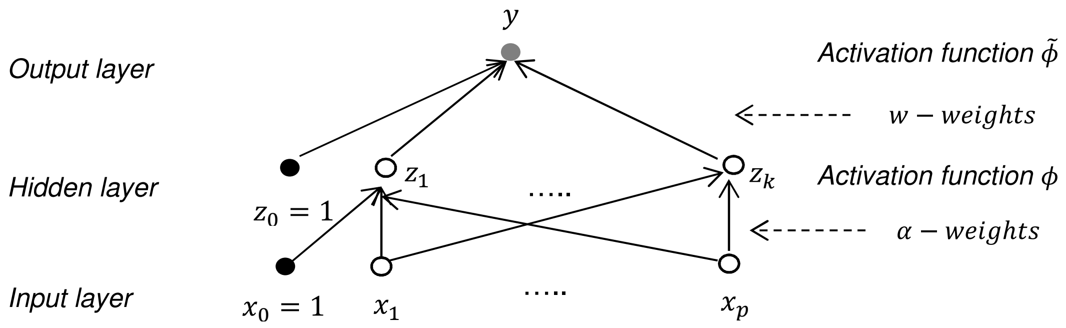


Figure 3.16: *Idea of Neural Nets with one binary output and one hidden layer.*

The estimation of the parameters in the binary target case is usually done by minimizing the cross entropy $R(\boldsymbol{\alpha}, \boldsymbol{w}) = -\sum_{i=1}^{N} y_i \, log\pi_i + (1 - y_i)log(1 - \pi_i)$ which is nothing else than the negative log-likelihood. The minimization problem is solved by *gradient descent* (see Hastie (2009)) and called *back-propagation*. Unfortunately it is not a convex problem, i.e. several local minima might exist. Therefore several starting points must be tested.

As Neural Nets provide a very flexible modeling tool, the global minimum often overfits. This might be faced by *early stopping* (before global minimum) or usage of regularization methods which is called “weight decay” in case of Neural Nets: $R(\boldsymbol{\alpha}, \boldsymbol{w}) + \lambda * J(\boldsymbol{\alpha}, \boldsymbol{w})$, with $J(\boldsymbol{\alpha}, \boldsymbol{w}) = \sum_l w_l^2 + \sum_{lj} \alpha_{lj}^2$.
The parameter λ functions as a tuning parameter as usual. When using regularization, the number of hidden units can be set to a high level, somewhere between 5 and 100; the weight decay then prohibits overfitting (see Hastie (2009)). For the Tb data, a single hidden layer model together with 20 hidden units chosen which results in decays of roughly 1 (see chapter 4.2), which hints to a moderate regularization. Usually the number of hidden layers represents a design question and should be guided by background knowledge, which makes Neural Nets with more than one hidden unit often kind of an art.
At last it must be noticed that Neural Nets are also black boxes regarding the assessment of variable interpretation. This is due to the fact that the predictors enter the model by nonlinear variable combinations, which makes it extremely difficult to quantify their influence. Nonetheless, because of the nonlinear approach, they provide a powerful tool for classification that might outperform other methods.

4 Comparison of Method-Performance

4.1 Measuring Classifier Performance

4.1.1 Metrics for Classifier Performance

Several metrics for measuring classifier performance exist (cf. Tutz (2012)). Some of them can be expressed in terms of a loss function. E.g. the 0-1 loss: $L_{01}(y, \hat{y}) = I(y \neq \hat{y})$ (with I as the indicator function) has the expected value $E\ L_{01}(y, \hat{y}) = P(y \neq \hat{y}|x) = 1 - P(y = \hat{y}|x)$, which is the probability of misclassification and can be estimated by the *misclassification error*:

$$\frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \hat{y}_i) \tag{4.1}$$

This error is minimized by the Bayes classifier, which is $\hat{y} = \text{argmax}_{k=0,1} P(y = k|x)$ (or equivalently $\hat{y} = 1\ if\ P(y = 1|x) > 0.5$) in the binary target case.

If the prediction results in a probability for the event $\hat{\pi}$, a more precise metric represents the quadratic or *Brier score*. Here the loss function is the quadratic loss applied to the real values y and the estimated probabilities $\hat{\pi}$: $L_2(y, \hat{\pi}) = (y - \hat{\pi})^2$. This metric has an interesting expected value: $E\ L_2(y, \hat{\pi}) = L_2(\pi, \hat{\pi}) + var(y)$ (see Tutz (2012)). The last term depends only on the true probability; therefore the whole metric is a measure for the distance of the true probability π to the estimate $\hat{\pi}$. It can be estimated by:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\pi}_i)^2 \tag{4.2}$$

In order to understand why the sheer misclassification error can be misleading in contrast to the Brier score regarding the discriminative power of a method, the following example is given: In the next chapter it can be seen that the misclassification rate of the logistic regression is approximately 22%. But this is also the overall percentage of LTFU. Therefore, if applying the naive classification rule of assigning a probability of $\hat{\pi}$=22% to each patient or equivalently a y=0 ("Not Retained"), would also end up in a 22% misclassification rate. One says that this rule calibrates perfectly but has poor shapness. On the other hand the Brier score would be 0.22^2+0.78^2=0.66, which is much worse than the 0.16 which e.g. results from logistic regression. So the Brier score accounts for both, calibration and sharpness.

The mostly used method for assessing discriminative capability (or sharpness) is represented by the *ROC* (receiver operating characteristic) curve (see e.g. Sachs (2006)), see Figure 4.1. It plots the *sensitivity*: $P(\hat{y} = 1|y = 1)$ (or equivalently true positive rate) against "1-*specificity*": $P(\hat{y} = 1|y = 0)$ (or equivalently false positive rate), depending on the cutoff value $\hat{\pi}_{cutoff}$ for assigning $\hat{y} = 1$. It is much more informative than the misclassification rate or the Brier score, as it provides the discriminative power for all possible cutoff values. The best prediction in terms of maximizing sensitivity+specificity is the curve point which lies as far as possible in the upper left corner. If the importance of sensitivity and specificity are taken unequal, this obviously would change.

In order to summarize the performance of a classification method into one metric, the AUC (area under the curve) of the ROC curve can be given.

Notice that, when the classification method is uninformative, i.e. not better than guessing, the ROC curve would be a straight diagonal from the lower left to the upper right corner, yielding an AUC of 0.5.

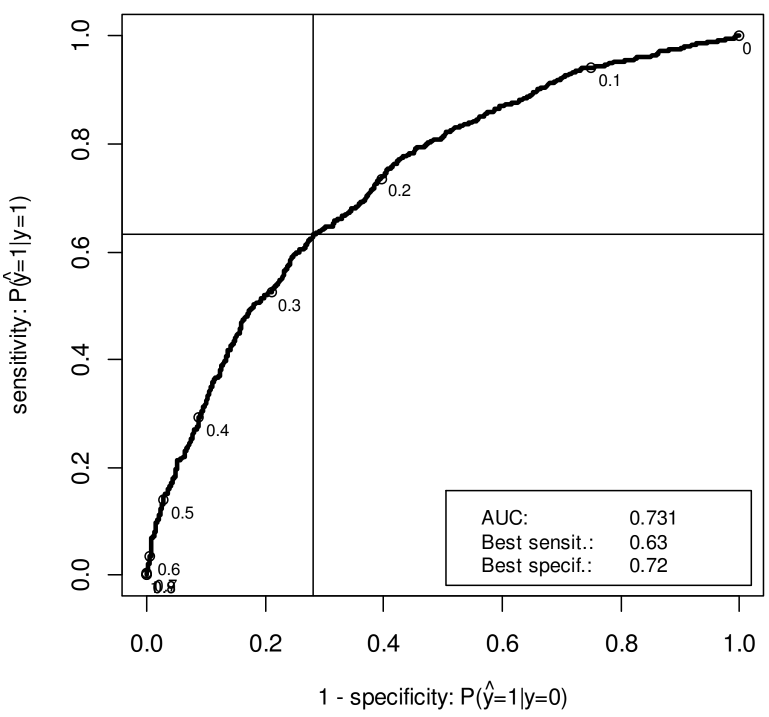


Figure 4.1: *ROC curve (for final logistic regression on Tb data). Points indicate probability cutoff-values for predicting "Not Retained". Reference lines cross axis at "best" prediction (where sensitivity+specificity is maximized).*

4.1.2 Honest Assessment of Classifier Performance

When deriving above metrics from the whole data, the misclassification and the Brier score would be underestimated and the AUC overestimated respectively. This happens as the methods try to fit the data as good as possible and would therefore also fit the noise in the data. Applying the resulting model to new data would show much lower performance. This overfitting is getting worse, when a tuning parameter has to be adapted by the data as well.

One approach to assess the generalization capability of a model is represented by the AIC metric (see also chapter 3.1.1), as it provides an estimate of the test error (see Hastie (2009)). Unfortunately the AIC is not available for all used methods. A better strategy, which is also applicable to every predictive model, is to split the data into training and test data. The model is trained on the training data and the predictive power assessed on the test data, providing a "honest" *out-of-sample error*. Possible tuning parameters can be derived by cross-validation on the training data. Therefore the training data itself is split (in case of 5-fold cross-validation) into 5 equal-sized bunches. Then, for each value of the tuning parameter, out of a predefined list of possible sensible values, the model is repeatedly (5 times in total) trained on 4 of the 5 bunches and evaluated by means of a criterion (misclassification rate, AUC, etc.) on the remaining bunch. The parameter value with results in the best average criterion value is taken as tuning value.

The derivation of the tuning parameter can alternatively based on a third data split; the validation data. The terms "test" and "validation" are sometimes interchanged in literature.

In the next chapter the different methods are compared, regarding their predictive capability, by 50-times 3:1 random split into training and test data, with possible tuning parameters evaluated by cross-validation on the training data. The resulting average (test data) AUC for the best models is around 0.7, whereas this metric derived from the whole data can get up to a value of e.g. AUC=1 for Random Forests. This underlines the importance of a correct estimation strategy for the generalization capabilities.

4.2 Performance Comparison of Classifier Methods

Figure 4.2 shows misclassification rate, Brier score and AUC for 50 random 3:1 splits into training and test data.

Used methods together with applied R-packages and parameters are described in the following. If a tuning parameter must be derived, the default evaluation criterion of the used method is taken. This criterion might differ from one method to another; actually sometimes the R help does not clearly state which criterion (deviance, misclassification rate, AUC, etc.) is used. The list or range of tested tuning parameter values was usually suggested by some pretests.

- *reg_main*: Logistic regression, just with all main effects.
- *reg_mainbestaic*: Same as *reg_main,* but relevant variables were selected by stepwise AIC search on training data (function *stepwise*, package *MASS*). For the whole data stepwise and exhaustive AIC result in the same model and because an exhaustive search takes a long time, it was decided to process just a stepwise search. This approach was tested as it represents a typical procedure used in the field.
- *reg_final*: Logistic regression with non-parametric modeling of *MomAgec* and additional interaction effects *Agec:MothersOwn* and *MomOccupation:Residence.*
- *reg_finalreduced:* Same as *reg_final*, but with reduced set of covariates (which was suggested by an AIC selection on the whole data, see chapter 3.1.2).
- *lda*: Linear discriminant analysis with all covariates (function *lda*, package *MASS*)
- *qda*: Quadratic discriminant analysis with all covariates (function *qda*, package *MASS*). For some runs estimation problems occurred. In such cases no simulation was processed.
- *lasso*: Logistic regression with all covariates and lasso regularization (function *cv.glmnet*, package *glmnet*). Tuning parameter λ is selected by 5-fold cross-validation on training data.
- *grplasso*: Same as lasso, but with group lasso regularization for categorical variables with more than 2 categories (function *cv.gglasso*, package *gglasso*: data must be scaled before processing). Tuning parameter selected in same manner as for *lasso*.
- *cart*: Classification tree based on conditional inference framework (function *ctree*, package *party*). It was decided not to use the classical tree, as the used function *rpart* from package *rpart* resulted in just a root node.
- *boosttree_mboost*: Boosting with trees (based on conditional inference framework) of fixed depth=6 terminal nodes (function *blackboost*, package *mboost*). Tuning parameter (number of Boosting iterations, max=1000) is selected by 5-fold cross-validation on training data. Learning parameter is fixed at ν=0.01.
- *boosttree_gbm*: Analogous to *boosttree_mboost* but with “standard” classification trees (function *gbm*, package *gbm*).

list continued…

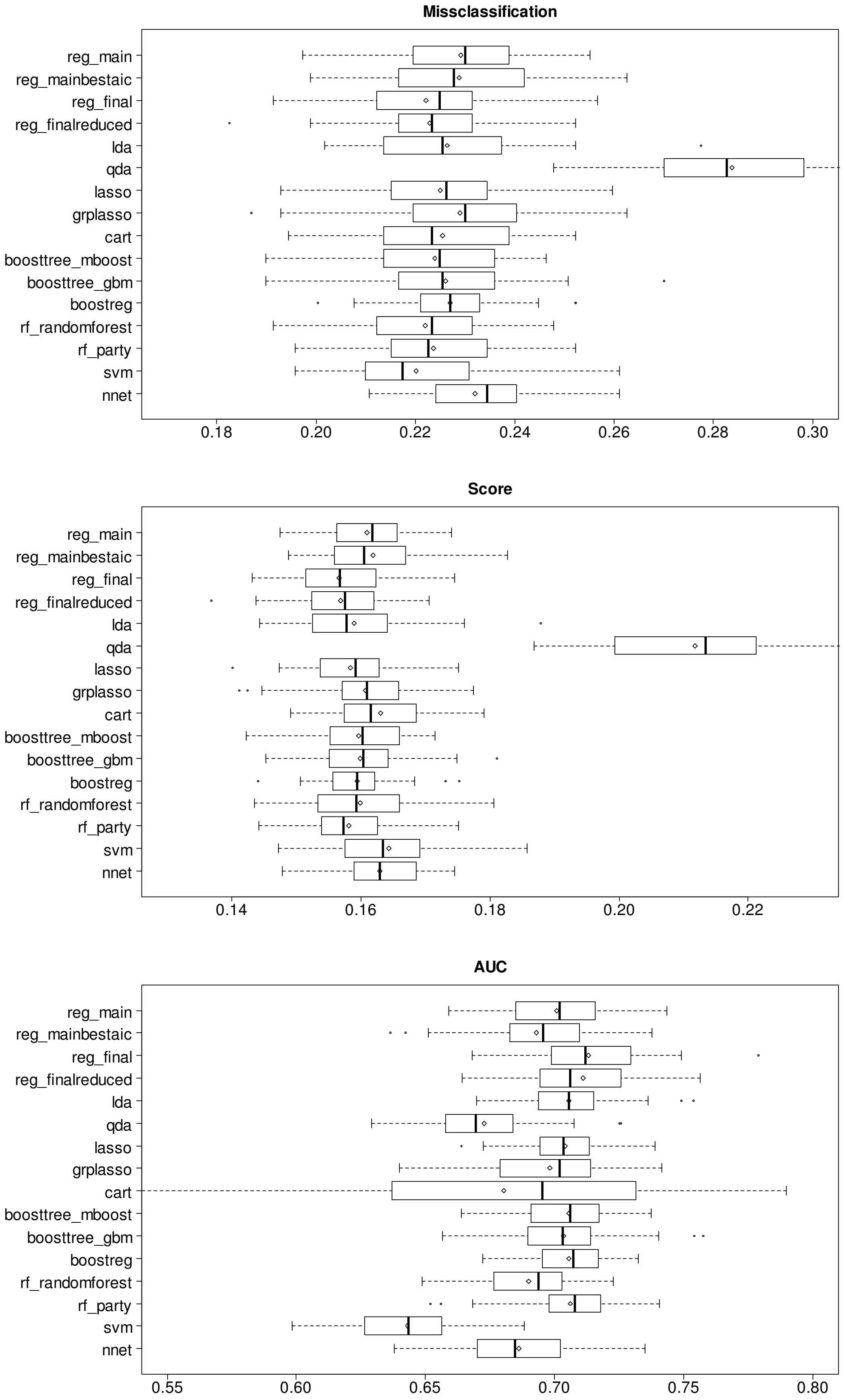


Figure 4.2: *Misclassification rate, Brier score and AUC for the different prediction models (Diamond indicate mean value). For the first two metrics a lower value is better, for the AUC a higher.*

- *boostreg*: Boosting with logistic regression (function *glmboost*, package *mboost*). Tuning (max=500) and learning parameter (ν=0.1) determined as in *boostee_mboost*.
- *rf_randomforest*: Random Forest using “standard” trees as base learners (function *randomForest*, package *randomForest*). Number of randomly chosen predictors is fixed at 3, which is also the default for the used covariate structure. A maximum of 500 trees is processed.
- *rf_party*: Same as *rf_randomforest* but with conditional inference trees as base learners (function *cforest*, package *party*)
- *svm*: Support Vector Machine with Gaussian radial kernel (function *kvsm*, package *kernlab*, kernel parameter internally optimized: *kpar=”automatic”*). Cost parameter C is chosen from the values (0.005, 0.5, 5, 10, 20, 50) by 5-fold cross validation (function *tune*, package *e1071*). For some runs estimation problems occurred; in such cases no simulation was processed.
  Initially the function *svm* from package *e1071* was used, but resulted in absolutely bad predictions for some simulations, i.e. much worse than guessing, which might be a software bug.
- *nnet*: Neural net with one hidden layer and 20 hidden units (function *nnet*, package *nnet*). Weight decay is determined by 5-fold cross validation (function *tune*, package *e1071*). Range of possible decay values was (0, 0.5, 1, 1.5, 2).

Figure 4.2 indicates that SVM, Neural Net and QDA show lowest performance: QDA is highly outperformed by LDA for all 3 metrics, which is a typical behavior (cf. chapter 3.2). SVM shows best performance for misclassification rate. But this is an artifact, as described in chapter 4.1.1, and stands in contradiction to the (more sensitive) Brier score and also the AUC, which show that the SVM, regarding the predictive performance, is performing worse. This is confirmed by the ROC curve in Figure 4.3.
The CART performance confirms the high variability mentioned in chapter 3.4.
Taking variance into account, all other prediction methods show similar performance regarding the three metrics. Also the shapes of their ROC curves are comparable (see Figure 4.3), which indicates that the discriminative capability is also similar for different cutoff values for predicting LTFU.
The final logistic regression model performs best regarding the AUC, but the difference to other methods is so slight, that it might dissapear for a second simulation. Nevertheless this fact justifies the results and effect interpretation of the logistic regression, even though the main effects probably dominate the interactions, as the main effects model perform reasonably well.
Furthermore the AUC of approximately 0.7 for the best methods not only puts quite good confidence in the predictive power of the data, but also indicate that this might be an upper limit for it, as different methods result in this value.

4.3 Final Assessment of Variable Importance

Mother’s age and Residence are the most influential variables for LTFU. The standardized coefficients for the lasso confirm this for the whole lasso path and therefore also for the main effects model (shrinkage parameter = 1). Both variables are also at the top of the Boosting importance plot (Figure 3.10) and occur as split variables for the CART (with a p-value of <0.001), see Figure 3.9. Only in case of the Random Forest, *Residence* is also comparable to other covariates, regarding relative importance.
In order to give an estimate of the influence, the results of the final logistic model can be consulted: The odds ratio for LTFU for “permanent” mothers (42% of all mothers) is 0.36 [0.27, 0.48] times the LTFU of “temporary” mothers on average (see Figure 3.4). To give an impression of the influence of Mother’s age on LTFU, a 30 years (approx. 3rd quartile) old mother can be compared with a 20 years (approx. 1st quartile) old one (as the odds ratio starts to decrease at Mother’s age ≈ 20, see Figure 3.3). Here the odds ratio is lowered by a factor of 0.58 on average for the older mother.

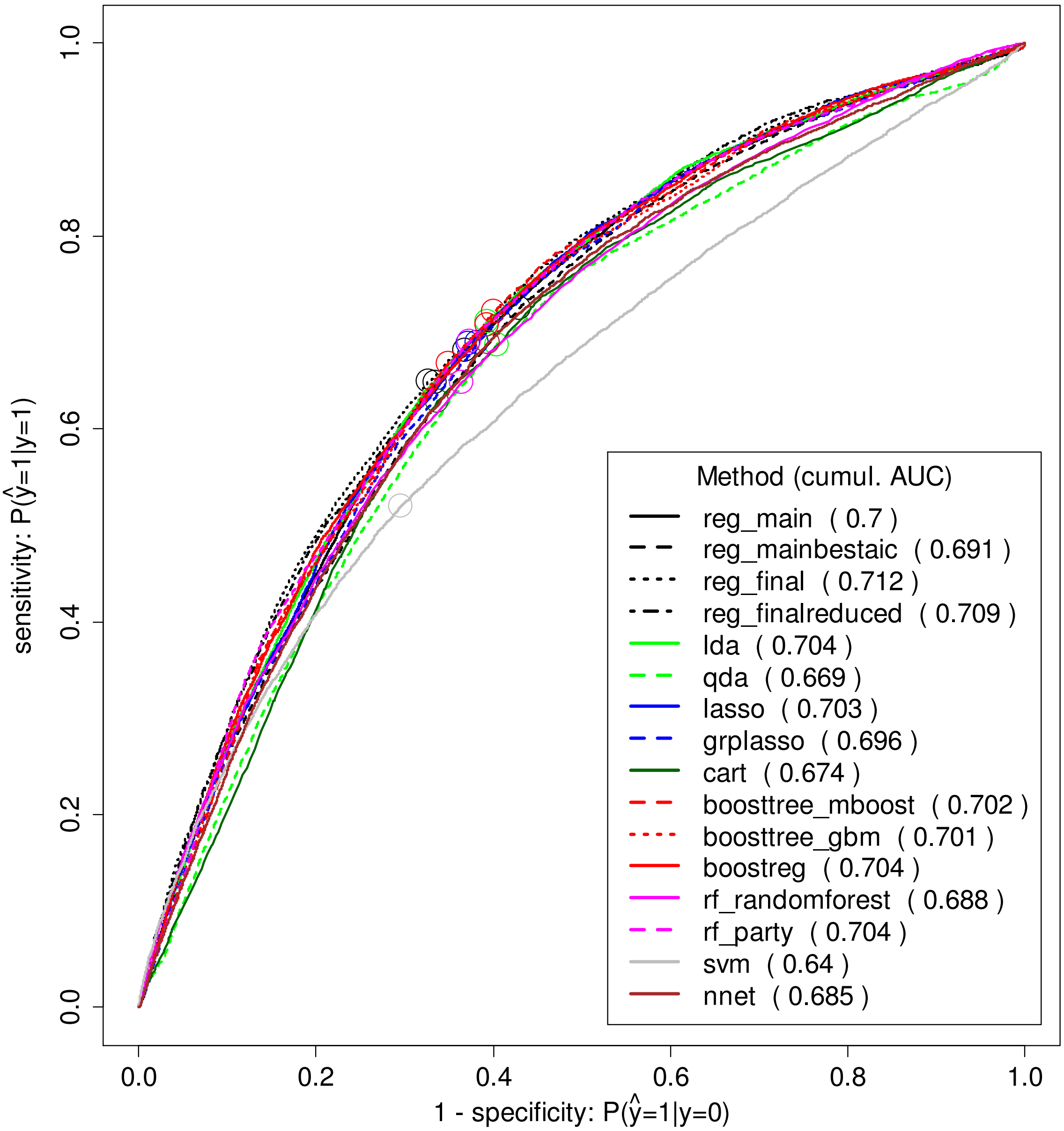


Figure 4.3: *Cumulative (over all simulations) ROC curves for the different prediction models. Dots indicate "best" prediction (maximizing sensitivity+specificity). The listed AUCs might differ slightly from the mean AUCs of* Figure 4.2 *as the cumulated version puts predictions of all simulations "into one pot".*

Furthermore *HousingType* occurs under the Top 5 in every plot or listing that allows relevance comparison. Regarding the coefficients of the final regression model, the odds ratio rises by a factor of 1.24 [0.96, 1.61] for families living in semi-permanent houses (18% of all families) and 1.72 [1.32, 2.26] for "permanent" houses (15% of all families) compared to "mud" houses (66% of all families).
This corresponds, in terms of the socio-economic status, to the higher odds ratio (for being "not retained") for better educated mothers. Even though *MomEducation*="Secondary" (13% of all mothers) is not significant in the final regression model, its remarkable coefficient of 1.84 should not be disregarded as *MomEducation* is an important variable for Lasso regularization, all Boosting methods, CART and

Random Forest. In contrast the significant category “Tertiary” *(1% of all mothers)* has no practical relevance, due to the low cell count and therefore low recruitment capability.

For Mother’s occupation, also listed under Top 5 in all importance plots or listings, the only relevant category (in terms of cell count) is “Farming” (37% of all mothers) compared to the reference “Salaried worker” (57% of all mothers). In the boosttree importance plot *MomOccupation* already comes at third place and the random forest importance plot shows it is under the Top 6, which might also be due to the interaction with *Residence*, detected in the final logistic regression model. Notice that all tree based methods can find interactions implicitly. Interestingly, Farmers have a lower odds ratio (resulting from logistic regression) compared to Salaried Workers only for Residence=”Temporary” (0.56 [0.42, 0.74]) whereas for “permanent” Farmers this is reversed (2.13 [1.34, 3.37]).

MothersOwn is relatively high rated in Random Forests and shows slight significance in logistic regression. Regarding the final regression model its importance might be due to the interaction with Age, even though the variable itself is significant (1.31 [1.05; 1.62]): For Mothers with less children (<= 3 children, 52% of all mothers) the odds ratio for LTFU increases e.g. for a baby enrolled with an age of 30 days compared to 0 days by 70% (see Figure 3.3). But this interaction is not detected by the boosttree, random forest or CART method. Furthermore *Agec* does not pop up as influential in any method (except for logistic regression), which makes this interaction at least questionable. The good performance of the main effects logistic model supports this tentativeness.

All other variables do not seem to have an important impact on the LTFU rate.

Summarizing all above results, a recruiting advice can be (in order of importance): Enroll older mothers having a HDSS-ID (Residence=”Permanent”) and avoid salaried workers without. Generally families with lower socio-economic status regarding housing type or education level are preferred as well as mothers with less additional children, for whom it might be advantageous to be recruited as soon as possible (lower infant’s age).

5 Summary

An analysis of 2695 patients (originally 2900, with died patients excluded) of an infant's Tuberculosis (Tb) study, conducted with newborn 0-6 weeks old, was performed in order to best predict the loss-to-follow-up (LTFU), which will help in creating appropriate retention strategies for future trials.
Potential covariates are characteristics of the child like *Age*, *Sex*, *WeightHeight* (Infant's birth weight/height ratio), *Temperature* (Infant's birth temperature), *PlaceOfBirth* and of the mother like *PlaceOfEnrolment*, *InfantsDelivered* (Number of infants delivered: "Singleton" or "Twins"), *MomEducationLevel* (Mother's education level), *MomOccupation* (Mother's occupation), *HousingType* (Housing type ="Mud", "Semi-permanent" or "Permanent"), *ReceivedAnteNtlCare* (Mother received antenatal care?), *HIVResultsAs* (HIV test result), *MothersOwn* (Mother's own additional children), *Residence* (HDSS-ID status: Health and demographic surveillance system identification). The binary target variable is *Retained*="Retained" or "Not Retained" (approx. 20% of total).
Values of 91 observations with missings or nomination of "Other" category are imputed with median (for metric variables) or mode (for categorical variables) values. Metric variables are additional centered.

A first descriptive analysis was conducted to search for outliers and unusual observations. No conspicuous observations were found.
Adjusted (controlled for all covariates) odds ratios for all variables were derived, with significant p-values already indicating the relevant predictors increasing LTFU rate: Lower *MomAge* (p-value <0.001), *MomEducationLevel* ("Tertiary" compared to "None", p-value=0.006), *MomOccupation* ("Salaried worker" compared to "Farming", p-value=0.003), *HousingType* ("Permanent" compared to "Mud", p-value<0.001) and *Residence* ("Temporary" compared to "Permanent", p<0.001). An AIC analysis additionally suggested the variables *Age*, *InfantsDelivered* and *MothersOwn.*

An extended logistic regression comprising also Interactions and non-parametric modeling of metric covariates with splines resulted in additional remarkable effects: Interactions *Age:MothersOwn* and *MomOccupation:Residence* and a non-parametric effect for *MomAge* advises that higher Infant's age negatively influences LTFU only for *MothersOwn<=3 children*; *MomOccuption* effect reverses for Residence="Permanent" mothers and *MomAge* reduces LTFU rate not until an age of approximately 20.

In the following several statistical models and data mining methods for classification were introduced and applied to the Tb data. In particular these are discriminant analysis, regularized regression with different lasso penalties, CART ("classification and regrssion trees"), Boosting (with trees and regressions), Random Forests, Support Vector Machines and Neural Nets. Noticeable results are represented e.g. by the main effects lasso paths, which provide a visual tool for comparing variable importance (also in dependence of shrinked coefficient values): *MomAge*, *Residence*, *HousingType*, *MomEducationLevel* and *MomOccupation* (in order of importance) again popped up here. These variables are also selected by a classification tree. Applying Boosting with trees not only confirms the importance of aforementioned predictors but also their dependency nature with the target (through partial depedance plots). The former also holds for Random Forest prediction of Tb data.
A final assessment is given (chapter 4.3), which also rates the practical significance of the relevant predictors for LTFU.

The central analysis of this thesis from a methodical standpoint is represented by the comparison of the performance of above mentioned classifier methods. Therefore misclassification rate, Brier score and AUC (area under the curve) are assessed by 50 fold 3:1 splits into training and test data. The models are trained on the training split with possible tuning parameters derived by 5-fold cross-validation and finally tested on the test split. This approach allows a honest assessment of classifier performance. Used R-functions and packages together with parameter settings are listed for every method. The final logistic regression model performs best regarding AUC (most informative metric according predictive capability), but are just slightly better than several other methods, except quadratic discriminant analysis, Support Vector Machine and Neural Net, which all perform worse. For the latter two methods this result came relatively unexpected as both have high reputation in machine learning community. A more detailed modeling, e.g. by testing other kernels for SVM or a more ambitioned hidden layer architecture for the Neural Net might help.

Approximately an average AUC of 0.7 resulted for the best prediction methods which confirms the predictive power of the data (and methods). Therefore the aforementioned interpretation of predictors based on the extended logistic regression modeling best describes the underlying model, even though the main effects might dominate the discriminative capability, as a pure main effects model could compete.

A Supporting Plots and Tables

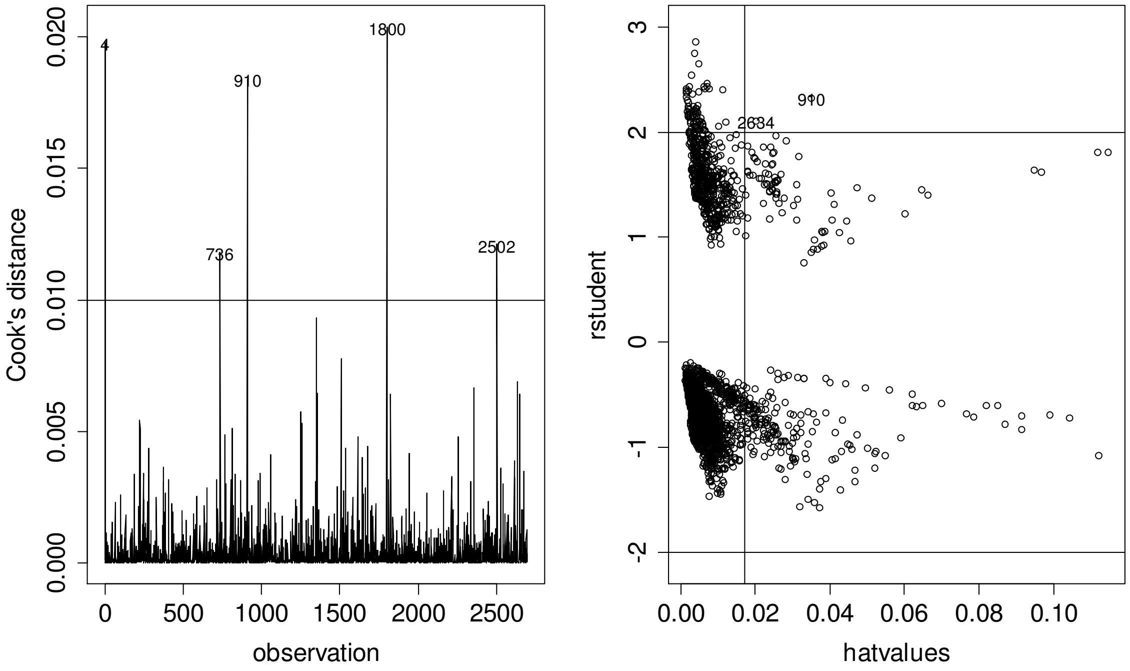


Figure A.1: *Searching influential observations. As cutoff values (reference lines) for Cook's distance are controversy discussed in literature, an arbitrary value of 0.01 is taken just in order to pick some observations with highest distance. For the hatvalues a cutoff of 2 * #coefficients / #observations is taken. The ±2 cutoff for the studentized residuals are oriented at the linear model, even though the residuals of logistic regression usually are not gauss-distributed.*

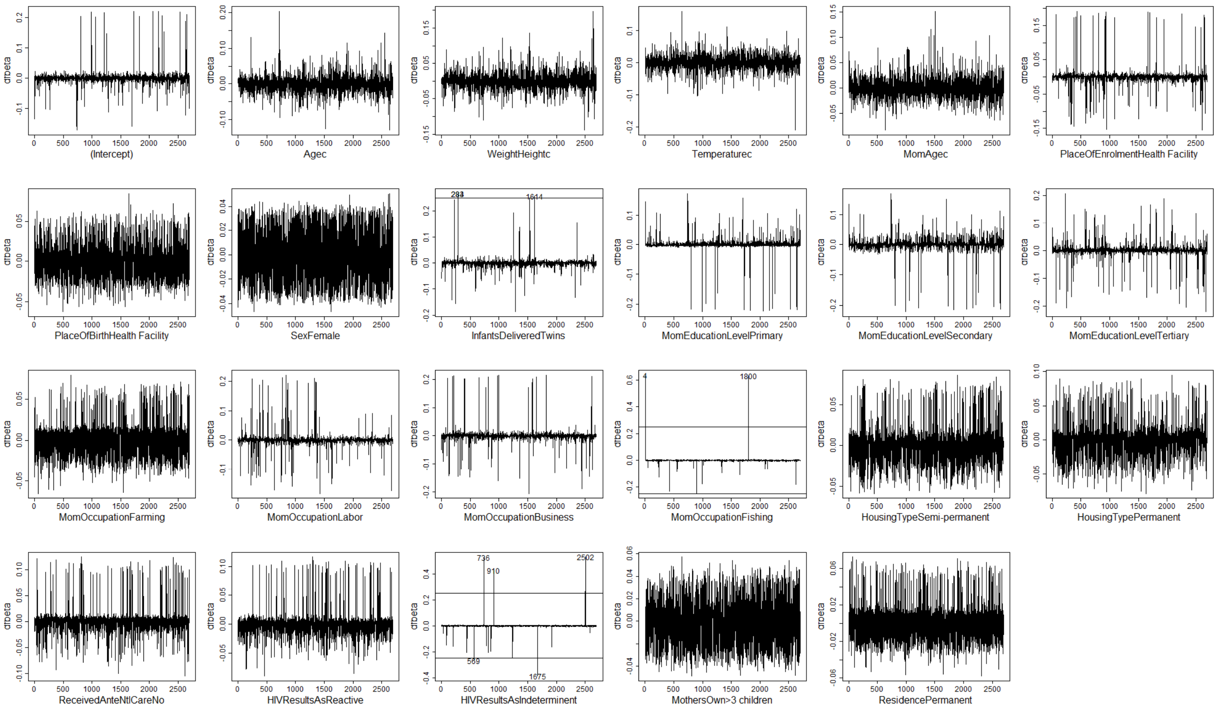


Figure A.2: *Dfbetas plot. Cutoff values for reference lines are chosen like in Figure A.1.*

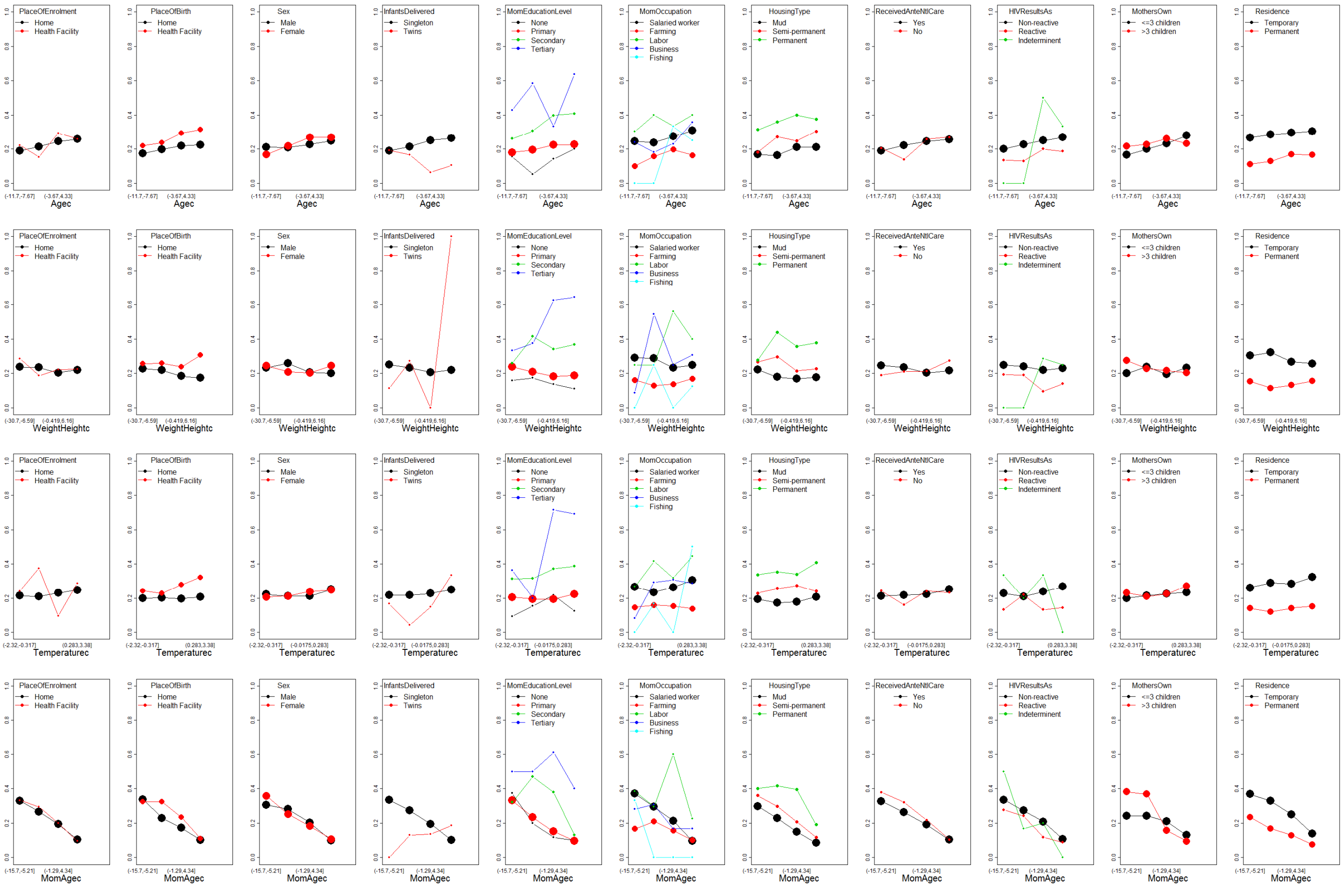


Figure A.3: Marginal interactions of metric with categorical covariates (y-Axis show percentage "Retained"). Metric variables are discretized in quartile bins. Size of points represent cell count of corresponding category and therefore indicate importance.

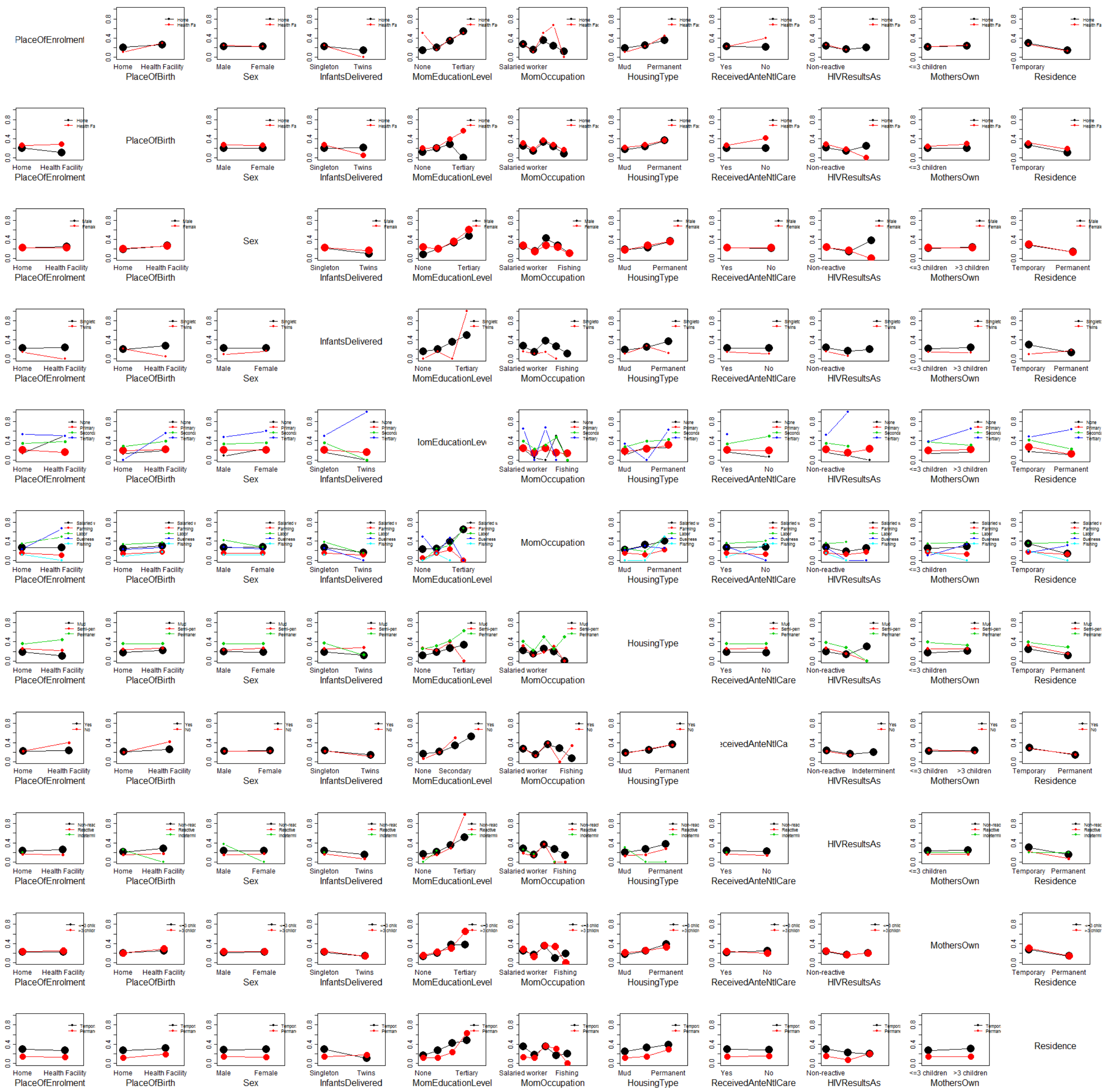


Figure A.4: Marginal interactions of categorical covariates (y-Axis show percentage “Retained”). Size of points represent cell count of corresponding category and therefore indicate importance.

Covariate	exp(β)	95% CI	p-value
(Intercept)	0.27	[0.14; 0.50]	<0.001
Agec	1.02	[1.01; 1.04]	0.001
WeightHeightc	0.99	[0.98; 1.00]	0.168
Temperaturec	1.19	[0.95; 1.50]	0.137
s(MomAgec)	n.a.	n.a.	<0.001
PlaceOfEnrolmentHealth Facility	0.73	[0.43; 1.24]	0.241
PlaceOfBirthHealth Facility	0.97	[0.78; 1.21]	0.799
SexFemale	1.04	[0.85; 1.26]	0.723
InfantsDeliveredTwins	0.50	[0.24; 1.01]	0.055
MomEducationLevelPrimary	1.11	[0.60; 2.05]	0.736
MomEducationLevelSecondary	1.87	[0.98; 3.60]	0.059
MomEducationLevelTertiary	3.63	[1.37; 9.61]	0.009
MomOccupationFarming	0.56	[0.42; 0.74]	<0.001
MomOccupationLabor	0.78	[0.39; 1.55]	0.472
MomOccupationBusiness	0.24	[0.08; 0.73]	0.012
MomOccupationFishing	0.46	[0.09; 2.31]	0.349
HousingTypeSemi-permanent	1.24	[0.96; 1.61]	0.097
HousingTypePermanent	1.72	[1.32; 2.26]	<0.001
ReceivedAnteNtlCareNo	1.18	[0.84; 1.65]	0.340
HIVResultsAsReactive	0.77	[0.56; 1.07]	0.124
HIVResultsAsIndeterminent	0.77	[0.20; 2.90]	0.699
MothersOwn>3 children	1.31	[1.05; 1.62]	0.017
ResidencePermanent	0.36	[0.27; 0.48]	<0.001
Agec:MothersOwn>3 children	0.97	[0.96; 0.99]	0.003
MomOccupationFarming:ResidencePermanent	2.13	[1.34; 3.37]	0.001
MomOccupationLabor:ResidencePermanent	3.89	[1.20; 12.56]	0.023
MomOccupationBusiness:ResidencePermanent	7.26	[1.91; 27.63]	0.004
MomOccupationFishing:ResidencePermanent	0.00	[0.00; Inf]	1

Table A.1: *Odds ratio results of final logistic regression (notation follows Table 3.1, s(MomAgec) indicates the additive effect for Mother's age).*

Bibliography

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*, Springer, New York

Fahrmeir, L. & Hamerle, A. & Tutz, G. (1996). *Multivariate statistische Verfahren*, Walter de Gruyter, Berlin

Fahrmeir, L. & Kneib,T. & Lang, S. & Marx, B. (2013). *Regresssion. Models, Methods and Applications*, Springer, Heidelberg

Faraway, J.J (2005). *Linear models with R*, Chapman&Hall/CRC, Boca Raton, FL.

Faraway, J.J (2006). *Extending the linear models with R: generalized linear, mixed effects and nonparametric regression models*, Chapman&Hall/CRC, Boca Raton, FL.

Hastie, T. & Tibshirani, R. & Friedman, J. (2009*). The Elements of Statistical Learning. Data Mining, Inference and Prediction. Second edition*, Springer, New York

Sachs, L. & Hedderich, J. (2006). *Angewandte Statistik. Methodensammlung mit R*, Springer, Berlin

SAS9.2OnlineDoc. *http://support.sas.com/documentation/92/index.html*

Tutz, G. (2012). *Regression for Categorical Data*, Cambridge University Press, New York

Wood, S.N. (2006). *Generalized Additive Models. An introduction with R*, Chapman&Hall/CRC, Boca Raton, FL.