# Logistic Regression: Fracture Prediction

Alexander Odermatt

July 14, 2024

## 1 Introduction

Fractures can result in grave, painful injuries often leading to disability and sometimes even premature death. It takes multiple weeks for fractures to heal, which can trigger other undesired complications during the healing process.

A common disease whose phenotype include an increased susceptibility to fractures and is among others diagnosed through decreased bone mineral density (bmd), one of the variables considered in this study, is osteoporosis. It is estimated that between 2010 and 2019 the treatment gap, meaning the percentage of individuals that did not receive treatment even though they have the disease, for women went up by 27%. In addition, it is predicted that the number of fractures per year will increase by more than 37% between 2019 and 2034.

Further, fractures generate a considerable amount of healthcare costs (around 4.5% of total healthcare spending), which puts pressure on the health care system. [1, 4, 3]

These findings result in a motivation to better understand and predict fractures, both from a financial and health perspective.

## 2 Exploratory Data Analysis

The dataset contains 169 observations each representing an individual, with 9 variables, notably:

1. id: unique identifier [nominal]
2. age [continuous]
3. sex [binary]
4. fracture: has individual had a fracture [binary]
5. weight in kg [continuous]
6. height in cm [continuous]
7. medication [ordinal]
8. waiting time [discrete]
9. bmd: bone mineral density [continuous]

From figure 1 it becomes visible that the distributions of the continuous variables from the dataset exhibit symmetry. However, figure 2 indicates that the distribution for age and height have minimally short tails.

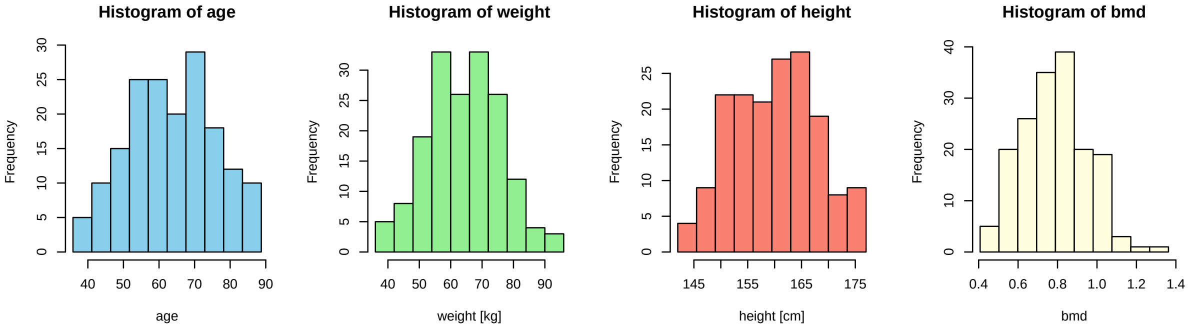


Figure 1: Histograms for age, weight, height and bone mineral density.

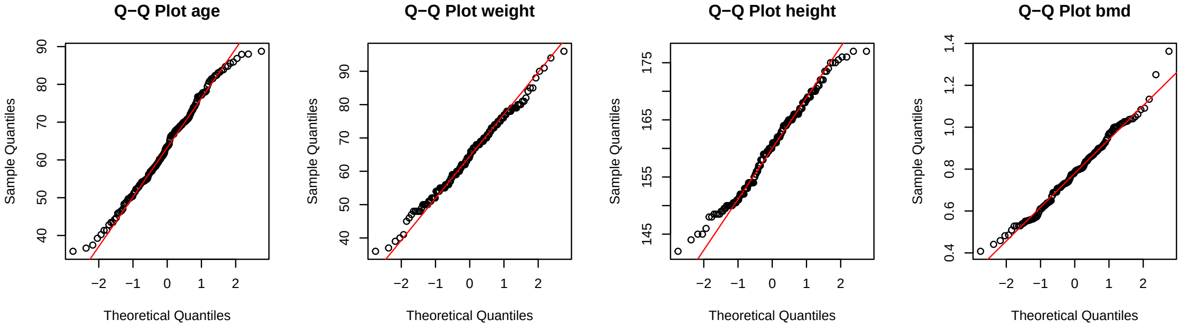


Figure 2: Normal Q-Q plots for age, weight, height and bone mineral density.

For the following model selection only the below variables are considered:

| Variable | Summary |
|---|---|
| age | Min: 35.81, Median: 63.49, Mean: 63.63, Max: 88.75, $\sigma$: 12.36 |
| bmd | Min: 0.408, Median: 0.786, Mean: 0.783, Max: 1.362, $\sigma$: 0.17 |
| bmi | Min: 15.43, Median: 24.96, Mean: 25.20, Max: 38.54, $\sigma$: 4.41 |
| sex | F: 83 / M: 86 |
| fracture | Yes: 50 / No: 119 |

Where the body mass index is calculated as follows: $BMI = weight \times \left(\frac{height}{100}\right)^{-2}$. The pairwise plot in figure 3 indicates that for these variables bmi and bmd seem to have some positive association, age and bmd a slightly negative association, while bmi over age seems to be more or less constant. This gets supported by the linear correlation coefficients displayed in figure 4 as well as the much more clearly visible negative correlation between fracture and bone mineral density, hinting at the fact that the risk of bone fracture increases considerably as the bmd decreases. This is in line with the fact that bmd is used to diagnose osteoporosis and measure treatment effectiveness [3].

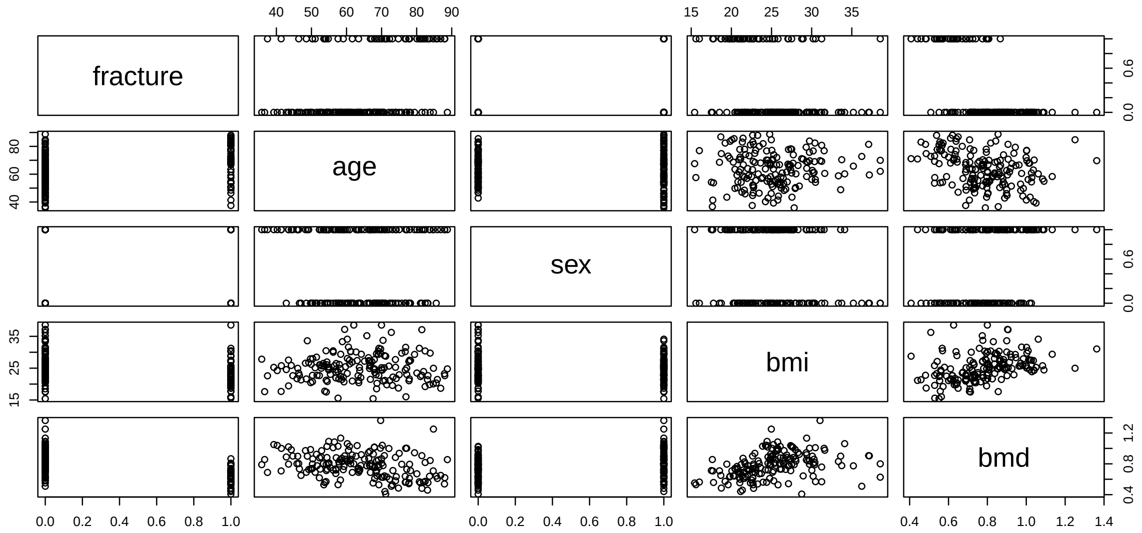
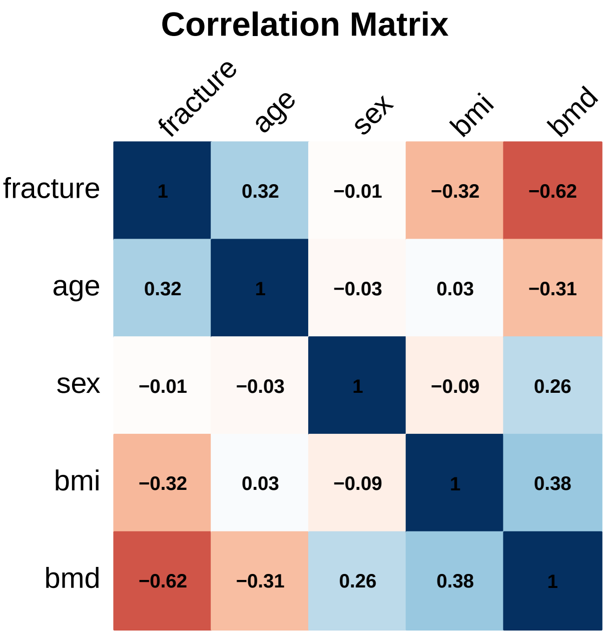


Figure 3: Pair plot of the different variables considered for the logistic regression model.



| | fracture | age | sex | bmi | bmd |
|---|---|---|---|---|---|
| fracture | 1 | 0.32 | −0.01 | −0.32 | −0.62 |
| age | 0.32 | 1 | −0.03 | 0.03 | −0.31 |
| sex | −0.01 | −0.03 | 1 | −0.09 | 0.26 |
| bmi | −0.32 | 0.03 | −0.09 | 1 | 0.38 |
| bmd | −0.62 | −0.31 | 0.26 | 0.38 | 1 |

Figure 4: Correlation matrix of the different variables considered for the logistic regression model.

## 3 Logistic Regression Model

To predict the fracture probability $p$ from predictors $x_i$ the logistic regression model, also called logit model, can be used. The equation representing the logit model is given in equation (1), where the coefficients $\beta_i$ are found through model fitting.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{1}$$

Model fitting is done through maximizing the *Likelihood Function* given in equation (2). This function calculates the product of the probabilities of observing the data points given the specified coefficients. Maximizing this function by changing coefficient values essentially leads to the coefficients which best explain the observed data.

$$\mathcal{L}(\beta_0, \beta_1, \ldots, \beta_k) = \prod_{i=1}^{n} \left[p_i^{y_i} \cdot (1-p_i)^{1-y_i}\right] \tag{2}$$

Thus a fitted model which receives values for its predictors will yield a probability for the binary outcome.

## Method

To decide what predictors should be in the final logistic regression model one needs to compare fitted models with different predictors and predictor-interaction terms. But comparing all possible models to each other is not really feasible as it leads to a combinatorial explosion. Further, an evaluation criterion to rank and compare models to each other needs to be established.

The necessary comparisons can be significantly reduced by using *step-wise selection*. From a specified starting point step-wise selection then explores both, forward and backwards steps, meaning adding predictors and removing predictors, and thus does not run the risk of stopping short in either direction. In addition, three different starting points where tested (models (3), (4), (5)), such that in the case that all starting points converge on the same model a more robust argument can be made. The three starting points include the scope "extremities" as well as one model that is an intermediate with one interaction term and one individual term. The allowed scope for the step-wise selection process spans from the complex model (3) containing all interaction terms as well as individual terms to the simple, constant model (4).

$$\text{Startpoint 1:}\quad \text{fracture} \sim \text{age} \times \text{sex} \times \text{bmi} \times \text{bmd} \tag{3}$$
$$\text{Startpoint 2:}\quad \text{fracture} \sim 1 \tag{4}$$
$$\text{Startpoint 3:}\quad \text{fracture} \sim \text{bmi} \times \text{age} + \text{sex} \tag{5}$$

Note: Models with interaction terms implicitly also contain simple predictor terms for the predictors contained in the interaction terms. E.g. bmi $\times$ age $=$ bmi $+$ age $+$ bmi $\times$ age

Further, to compare the fitted models two evaluation criteria in the context of step-wise selection are tested:

- Akaike Information Criterion: $AIC = -2log(max\ likelihood) + 2 \times \#params$
- Bayes Information Criterion: $BIC = -2log(max\ likelihood) + log(\#obs) \times \#params$

For both criteria lower numbers indicate a better model. As the criteria consider the maximum likelihood (as in equation (2)) as well as the number of parameters in the model, they give a measure of how accurate the model is on the data without running the risk of over fitting as adding parameters gets punished by the criteria.

The Cook's Distance as in formula (6) of the found models are then analyzed to see if any data point has an undue influence on the estimated coefficients. In that case the data points may be removed and the model may be refitted.

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_{j(i)} - \hat{Y}_j)^2}{(p-1) \times \text{MSE}} \tag{6}$$

To compare two models fitted on the same data, where one model has one predictor less than the other, a Likelihood Ratio Test (LRT) can be performed to determine if the predictive quality added by the extra predictor significantly improves the fit. Another valuable analysis for comparing different models is the ROC curve analysis, illustrating a models performance at a binary classification task. From the ROC curve another metric, the area under curve (AUC), can be derived. The AUC measures the overall performance of the model, with values closer to 1 indicating better predictive ability [2].

## Model Convergence & Evaluation

The ensuing findings can only be deemed valid under the following assumptions:

- binary outcome
- independent observations
- linear relation between logit and linear predictor
- no multicollinearity
- no outliers

The assumptions of binary outcome, non-mulitcolinearity and no outliers are supported by the EDA findings in figures 3 and 4.

While for each criteria respectively the step-wise selection process converged to the same model for all three starting points, the converged model differs depending on what criterion is used. With the Akaike Information Criterion the algorithm converged on the model given in (7), while with the Bayes Information Criterion the algorithm converged on the model given in (8). For the rest of this project the models will be referred to as AIC model and BIC model, hinting at the criterion that was used to converge to them. As $log(\#obs) = log(169) \approx 5.13$ it makes sense that if any model has less parameters it is the one found using BIC.

$$\text{AIC Model:} \quad \text{fracture} \sim \text{sex} + \text{bmd} \tag{7}$$

$$\text{BIC Model:} \quad \text{fracture} \sim \text{bmd} \tag{8}$$

An LRT performed on the two models yields a $\chi^2$-test p-value $= 0.0485 < 0.05$, indicating that the additional parameter in the AIC model introduces valuable predictive quality.

However it can also be a good idea to examine the Cook's Distance of the different observations within the models. This has been plotted in figure 5. It becomes apparent that certain data points are above the commonly chosen cutoff of $\frac{4}{\#\text{obs}}$. Thus the two models (7) and (8) are again fitted but on a reduced dataset, not containing the data points at the indices given numerically in figure 5. These reduced datasets are slightly different as they do not contain the same observations (# 62 is unique to the AIC model, while # 107 is unique to the BIC model).

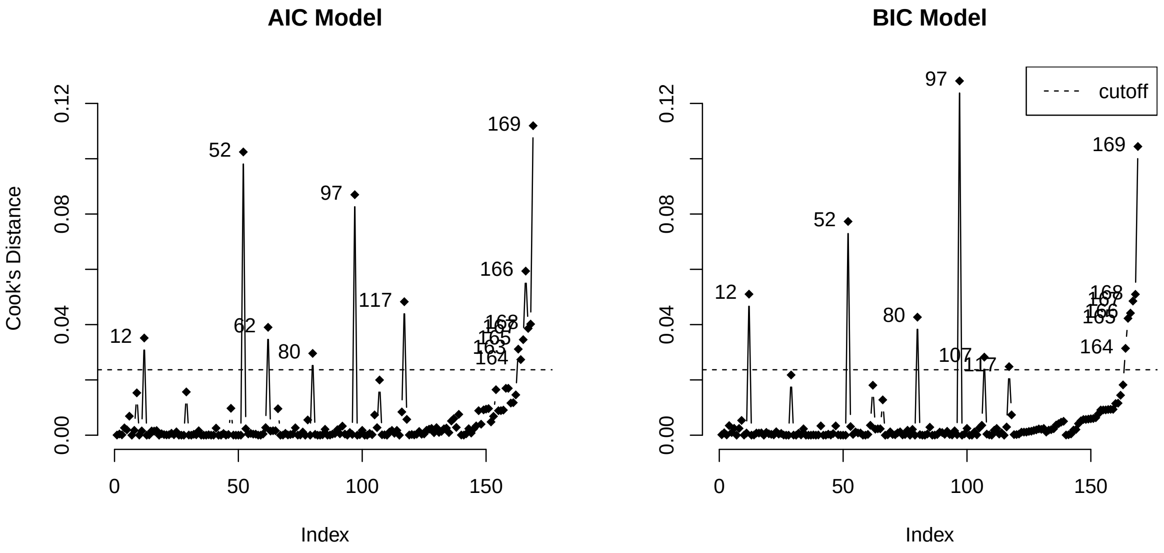


Figure 5: Cook's Distance plots for both models. The indices for data points above the cutoff ($\frac{4}{\#\text{obs}}$) are shown on the plot.

Now two things need to be evaluated:

1. Which model should be chosen?
2. On what data should the model be fitted?

This leaves us with four possibilities. A histogram for each of those possibilities showing the fracture prediction probability for actual fracture and non-fracture cases can be seen in figure 6. For both models the histogram displays accumulations at high and low probabilities for fracture and non-fracture cases respectively. It is clearly visible that those accumulations are much more pronounced when the models have been fitted on the reduced dataset. This in itself only indicates that these models are more decisive but does not necessarily indicate that they are better. However when looking at the ROC curve analysis in figure 7 it can be seen that the models fitted on the reduced dataset exhibit a more pronounced turn in the upper left corner, which can be associated to a higher performing model.

Thus it has become clear that the models should be fitted on their respective reduced dataset, but which between the two performs better? To determine this we can look at different metrics such as the AUC given in table 1 and the performance indicators given in the confusion matrices in table 3. A strong indicator is that with 0.99 the AUC of the AIC model fitted on the reduced dataset is higher than the one of the BIC model. This is further supported by the lower number of false negatives and thus higher accuracy.

Thus the final logistic regression model with coefficients is:

$$\text{Reduced AIC Model with coeffs:} \quad \widehat{\text{fracture}} = 31.12 + 2.20 \cdot \text{sex} - 46.88 \cdot \text{bmd} \tag{9}$$

Note: feminine → sex = 0; masculine → sex = 1

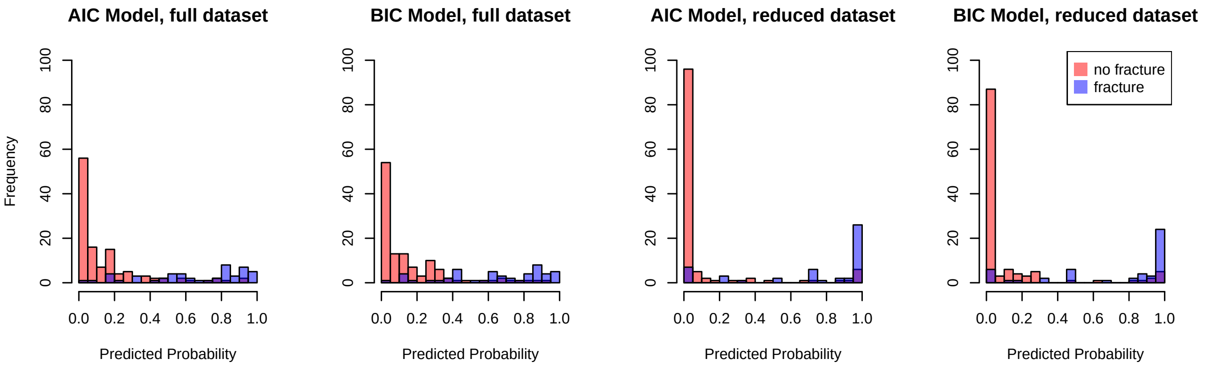


Figure 6: Histograms showing the prediction probabilities of the two models fitted on the complete and their reduced datasets for fracture and non-fracture cases respectively.

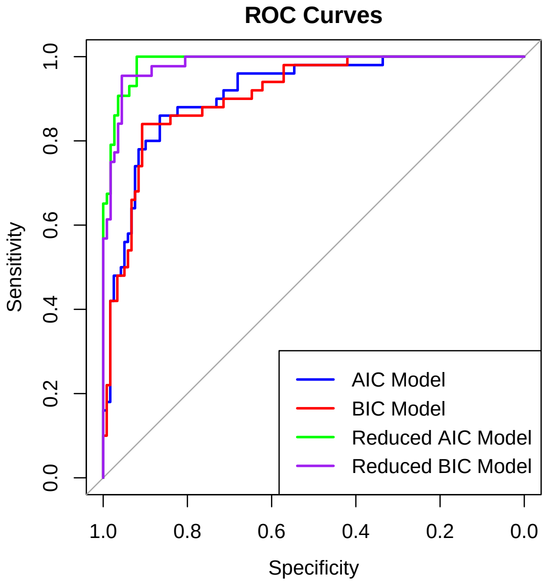


Figure 7: ROC curves for the different models.

| Model | AUC |
|---|---|
| AIC Model | 0.91 |
| BIC Model | 0.91 |
| Reduced AIC Model | 0.99 |
| Reduced BIC Model | 0.98 |

Table 1: AUC values for the different models

# 4 Conclusion

Through step-wise selection and multiple starting points two promising models could be determined. From there, based on the various analyses such as LRT, Cook's Distance, ROC curve and AUC, it was possible to clearly select the best model and determine on what dataset it should be fitted. Especially the Cook's Distance, the ROC curve and the AUC metric were influential in finding the final fitted model. The AIC model fitted on its reduced dataset seems to be the best performing model as it yields the best AUC and accuracy. Thus the resulting final logistic regression model contains the predictors given in (7) fitted on the dataset without the data points at the indices given in the left plot of figure 5. This results in the fitted model given in the equation (9). The signs of the coefficients indicate that men are slightly more prone to obtaining a fracture and a greater bone mineral density decreases fracture risk.

As an expansion of the project one could try out different cutoff values for the Cook's Distance to determine outliers and compare the resulting fitted models to each other.

For other studies one might reflect upon the question if false positives or false negatives have worse implications and adjust the model accordingly in order to increase sensitivity or specificity respectively. However in this project this is not considered as assigning weights to false negatives against false positives is considered to be outside the scope of this project.

# References

[1] Switzerland report.pdf. https://www.osteoporosis.foundation/sites/iofbonehealth/files/scope-2021/Switzerland%20report.pdf.

[2] atmathew. Evaluating Logistic Regression Models | R-bloggers, August 2015.

[3] NIAMS Science Communications and Outreach Branch. Bone Mineral Density Tests: What the Numbers Mean. https://www.niams.nih.gov/health-topics/bone-mineral-density-tests-what-numbers-mean, May 2023.

[4] Benjamin Caldwell. Everything You Need to Know About Fractures and Fracture Healing. https://northazortho.com/ask-the-expert/everything-you-need-to-know-about-fractures-and-fracture-healing/, May 2020.

# Appendix

| **Reference** | **Predicted AIC** | | **Predicted BIC** | |
|---|---|---|---|---|
| | **no fracture** | **fracture** | **no fracture** | **fracture** |
| **no fracture** | 110 | 9 | 110 | 9 |
| **fracture** | 13 | 37 | 16 | 34 |
| **Accuracy** | 0.87 | | 0.852 | |

Table 2: Confusion Matrix and accuracy for the AIC and BIC model.

| **Reference** | **Predicted AIC reduced** | | **Predicted BIC reduced** | |
|---|---|---|---|---|
| | **no fracture** | **fracture** | **no fracture** | **fracture** |
| **no fracture** | 109 | 10 | 109 | 10 |
| **fracture** | 11 | 39 | 16 | 34 |
| **Accuracy** | 0.876 | | 0.846 | |

Table 3: Confusion Matrix and accuracy for the AIC and BIC model fitted on the respective reduced dataset.