Assignment 2 - Regression Analysis on the GLOW Study (Fracture Risk)

Tanushree Kumar

DATA 630 - Summer 2024

Professor Ami Gates

University of Maryland Global Campus

Due: June 18, 2024

Introduction

The objective of this analysis is to determine the factors that contribute to the occurrence of fractures among participants in the GLOW study using logistic regression. The primary question that needs to be answered is: "What variables significantly affect the likelihood of a fracture?"

Osteoporosis is a medical condition characterized by weakened bones which increases the risk of sudden and unexpected fractures. The GLOW (Global Longitudinal Study of Osteoporosis in Women) (*Welcome to GLOW*, 2009) study was designed to identify risk factors associated with fractures in postmenopausal women, with data taken over a period of five years from 2099 to 2014. According to Perez et al. (2023), osteoporosis is the cause of almost 9 million fractures worldwide. Having a fracture can severely impact a patient's life and can also impede them in their daily life. Thus, understanding these risk factors is critical in predicting and preventing fractures. Accurate prediction of fractures can lead to better preventive measures and treatment plans, thereby improving the quality of life for at-risk populations

The chosen method for analysis is a logistic regression analysis. Logistic regression is an appropriate method as it allows us to model the probability of a binary outcome, the occurrence of a fracture, based on a set of predictor variables. This method is well-suited for identifying and quantifying the impact of various risk factors on the likelihood of fractures.

Analysis

The GLOW dataset is comprised of 60,393 postmenopausal women from 10 countries (Hooven, 2014). The dataset includes 500 observations of the 15 following variables:

- Subject ID: Identification code; 1 - n
- Site ID: Study site; 1 - 6

- Physician ID: 128 unique Physician ID codes
- PRIORFRAC: History of prior fractures (binary); 1 = Yes; 0 = No
- AGE: Age of the participant at enrollment; Years
- WEIGHT: Weight of the participant at enrollment; Kilograms
- HEIGHT: Height of the participant at enrollment; Centimeters
- BMI: Body Mass Index; Kg/m^2
- PREMENO: Menopause before age 45 (binary); 1 = Yes; 0 = No
- MOMFRAC: Maternal history of hip fracture (binary); 1 = Yes; 0 = No
- ARMASSIST: Use of arms when standing from a chair(binary); 1 = Yes; 0 = No
- SMOKE: Current and past smoking status (binary); 1 = Yes; 0 = No
- RATERISK: Self-reported risk of fracture; 1 = Less than others of
- the same age; 2 = Same as others of the same age; 3= Greater than others of the same age
- FRACSCORE: Calculated fracture risk score using a formula
- FRACTURE: Fracture occurrence in the first year (binary); 1 = Yes; 0 = No

The formula used to calculate the composite risk score is: FRACSCORE = 0 × (AGE ≤ 60) + 1× (60 < AGE ≤ 65) + 2 × (65 < AGE ≤ 70) +3× (70 < AGE ≤ 75) + 4 × (75 < AGE ≤ 80) + 5 × (80 < AGE ≤ 85) +6 × (AGE > 85) + (PRIORFRAC = 1) + (MOMFRAC = 1) +(WEIGHT < 56.8) + 2 × (ARMASSIST = 1) + (SMOKE = 10 × (AGE ≤ 60) + 1× (60 < AGE ≤ 65) + 2 × (65 < AGE ≤ 70) +3× (70 < AGE ≤ 75) + 4 × (75 < AGE ≤ 80) + 5 × (80 < AGE ≤ 85) +6 × (AGE > 85) + (PRIORFRAC = 1) + (MOMFRAC = 1) +(WEIGHT < 56.8) + 2 × (ARMASSIST = 1) + (SMOKE = 1) (*Code Sheet for Variables in the GLOW Study*, 2014).

Before developing the model, exploratory data analysis (EDA) was performed to understand the distribution and relationships of the variables. Based on the EDA, data

preprocessing was performed such as handling missing values, encoding categorical variables into factors, and scaling numerical features ensuring the data is clean and in an appropriate format for regression analysis. Using functions such as summary(), str(), and sum(is.na()) gave an overview of the dataset and showed that there were no missing values present. All the variables were either numerical or integer variables which meant little data preprocessing had to be performed.

The summary statistics were examined to see the distribution of the data. The ages of the participants ranged from 55 to 90 years with a mean age of 68.56 years and a median age of 67 years, which indicates a slight skew towards more older women. The weight of the participants ranged from 39.9 kg to 127 kg, with a mean of about 71.82 kg. The height ranged from 134 cm to 199 cm, with a mean of approximately 161.4 cm. The Body Mass Index (BMI) values range from 14.88 to 49.08, with a mean of 27.55. The composite fracture risk scores ranged from 0 to 11, with a mean of 3.698.

The summary statistics also showed that around 25.2% of the participants had a prior fracture. Approximately 19.4% of participants experienced early menopause. Around 13% of participants had mothers with hip fractures. About 37.6% of participants required arm assistance. 7% of participants were smokers or smoked during the study. The majority reported a similar risk of self-reported fractures with a mean of 1.96. Lastly, about 25% of participants experienced fractures. Overall, this dataset is comprised of postmenopausal women with a wide age range and diverse characteristics related to fracture risk factors. The summary statistics and data structure insights help in preparing data for analysis, identifying potential outliers or data issues, and guiding variable selection and modeling strategies.

A logistic regression analysis was used for this dataset as it is a statistical method used for binary classification. In the context of this dataset, logistic regression is used to predict the probability of a fracture, with a binary outcome of fracture or no fracture, using various risk factors such as age, BMI, previous fractures, etc. The coefficients generated from the model indicate the strength of the predictor variables used. Using these models, residual plots can be generated which confirm the appropriateness of the model's assumptions.

To fit the model, the data was split into a training set of 70% and a testing set of 30%. The training set was used to build the model while the test set was used to evaluate the accuracy of the model. A seed value was set which ensured the results would be reproducible when the model is run again. The general linear model was built with 'FRACTURE" being the dependent variable with the following independent variables: 'AGE', 'WEIGHT', 'BMI', 'PRIORFAC', 'PREMENO', 'MOMFRAC', 'ARMASSIST', 'SMOKE', 'RATERISK', and 'FRACSCORE'. The first three variables were excluded as they were used for identification purposes. The binomial family was used as the error function for the method, with the training data being used as the source. After building the training model, the generated output is stored and further evaluated in terms of predicting the probability of the first ten instances for the training and test set and then generating confusion matrices for both. Using the model built previously, a residual plot can be made to determine if the model is appropriate for modeling the given data. As a further step, a minimal adequate model can be built to gradually identify which variables are not significant with each iteration.

Results

The logistic regression model was built with the following formula: FRACTURE ~ AGE + WEIGHT + HEIGHT + BMI + PRIORFRAC + PREMENO + MOMFRAC + ARMASSIST +

SMOKE + RATERISK + FRACSCORE.

The coefficients from the model indicate the direction and magnitude of the impact of each predictor on the log-odds of having a fracture. The residual deviance measures how well the model fits the data with a lower residual deviance indicating a better fit. The AIC (Akaike Information Criterion) is used for model comparison, where a lower AIC suggests a better model fit, balancing complexity and goodness of fit. Looking at the summary of the model, the intercept is -14.397, but it's not statistically significant as p = 0.339, suggesting the baseline probability of fracture may not be significant. None of the predictors show strong statistical significance of p > 0.05, except possibly for BMI with p = 0.216. However, context and clinical significance should be taken into consideration rather than just p-values alone. The residual deviance is 353.82 on 340 degrees of freedom, indicating the model fits the data reasonably well. The AIC value of 377.82 suggests this model can be improved upon, as a lower AIC indicates a better-fitting model in comparison to other potential models. A figure of the model summary can be seen below, as well as in Appendix A

```
> summary(model)

Call:
glm(formula = FRACTURE ~ AGE + WEIGHT + HEIGHT + BMI + PRIORFRAC +
    PREMENO + MOMFRAC + ARMASSIST + SMOKE + RATERISK + FRACSCORE,
    family = binomial, data = train.data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.397361  15.071790  -0.955    0.339
AGE           0.007735   0.061339   0.126    0.900
WEIGHT       -0.128169   0.105361  -1.216    0.224
HEIGHT        0.068629   0.094815   0.724    0.469
BMI           0.341515   0.275732   1.239    0.216
PRIORFRAC     0.334511   0.450379   0.743    0.458
PREMENO       0.271627   0.332858   0.816    0.414
MOMFRAC       0.399052   0.481162   0.829    0.407
ARMASSIST     0.184191   0.675144   0.273    0.785
SMOKE        -0.344659   0.601029  -0.573    0.566
RATERISK      0.219350   0.184925   1.186    0.236
FRACSCORE     0.177328   0.306321   0.579    0.563

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 393.67  on 351  degrees of freedom
Residual deviance: 353.82  on 340  degrees of freedom
AIC: 377.82

Number of Fisher Scoring iterations: 4
```

The exponentiated coefficients for the odds ratio can also be observed. The intercept term represents the estimated odds of the outcome variable of ‘FRACTURE’ when all other predictor

variables are zero. Here, the intercept value is 5.588635e-07, indicating that when all predictor variables are zero, the odds of fracture are extremely low (close to zero). This makes logical sense since if all the variables are zero, the participant will physically not exist, thus diminishing the chance for a fracture to occur. Each predictor variable has an associated coefficient that represents the change in the log odds of the outcome for a one-unit increase in that predictor when all other predictors are held constant. The 'exp(coef)' command transforms the log-odds coefficients into odds ratios, which indicate how a one-unit increase in the predictor affects the odds of having a fracture. For example, 'AGE' has a coefficient value of 1.007765e+00, suggesting that for each one-year increase in age, the odds of fracture increase by approximately 0.8% (exp(0.007765) ≈ 1.007765). 'WEIGHT' has a coefficient value of 0.8797044 indicating that for each additional kilogram of weight, the odds of fracture decrease by approximately 12.0%. This should be noted as according to Watts (2014), obesity can help protect against fractures, specifically hip fractures. However, it can also lead to an increased risk of fractures occurring in the ankles and lower legs. For 'BMI', the coefficient value is 1.407078 suggesting that for each one-unit increase in BMI, the odds of fracture increase by approximately 40.7%. For 'PRIORFRAC', the value is 1.397257 indicating that individuals with a prior fracture have approximately 39.7% higher odds of having another fracture compared to those without a prior fracture. This is also noteworthy as this study had a clear goal to include women if they had experienced "prior fractures, diagnosis of osteoporosis, or current treatment for osteoporosis" (Hooven et al., 2009). These participants could then be flagged and closely monitored to ensure they are receiving the correct diagnoses and adequate care and treatment from their physicians.

To understand the predictive performance, the confusion matrices for both, the training and test data, can give some insight. The confusion matrices show how well the model predicts

fractures versus non-fractures. For example, in the training data, 254 non-fracture cases and 13 fracture cases were correctly classified, making a total of 267 correctly classified instances. The total number of misclassified instances is 85 and the total number of instances in the training set is 352. The classification accuracy of the training model is (267/352) * 100 = 75.85%.

For the test data, the same logic and calculations can be applied. Here, 105 non-fracture cases and 7 fracture cases were correctly classified, making a total of 112 correctly classified instances. The total number of misclassified instances is 36 and the total number of instances in the training set is 148. The classification accuracy of the test model is (112/148) * 100 = 75.67%

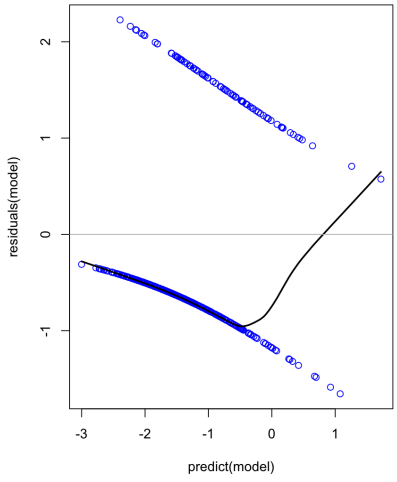
The classification accuracy of a logistic regression model represents the proportion of correctly predicted outcomes, in this case, the presence or absence of a fracture, out of the total number of predictions made. In the context of this study where the goal is to predict fractures based on various patient characteristics, a classification accuracy of ~75% for both the training and test data indicates several key points. First, the model is correctly predicting the outcome of fracture or no fracture 75% of the time in both the training and test datasets. This level of accuracy suggests that the model has a reasonably good performance. The similar accuracy rates for the training and test datasets indicate that the model is generalizing well from the training data to unseen test data. This means the model is not overfitting, which would have been evident if the training accuracy were significantly higher than the test accuracy. The close alignment of training and test accuracies enhances confidence in the model's reliability. It indicates that the model's performance is stable and not dependent on the specific sample used for training.

A 75% accuracy rate means the model could be useful in a clinical setting for assessing fracture risk. Healthcare providers could use the model's predictions to screen and identify patients at higher risk of fractures and potentially recommend preventative measures or further

diagnostic testing. The model's ability to generalize well suggests that the identified predictors (such as age, weight, BMI, prior fractures, etc.) are consistently associated with fracture risk across different patient samples. This reinforces the relevance of these predictors in assessing fracture risk in postmenopausal women.

With an accuracy of 75%, the model is better than random guessing (which would give an accuracy of around 50% for a binary classification problem) but still has room for improvement. This suggests that while the model captures some of the underlying patterns associated with fractures, there might be other factors or interactions that were not fully captured by the current predictors. This could involve incorporating additional predictors, using more advanced modeling techniques, or addressing any socioeconomic imbalance in the dataset.

A residual plot was also generated to visualize the model. The residual plot can be seen below, as well as in Appendix A.



Since the dependent variable of ‘FRACTURE’ can have only two possible values of 0 and 1, the points form two distinct curves. Evidently, the residual values are negative for the 0 values and vice versa. The prediction function is depicted by the black line in the graph.

For further analysis, a minimum adequate model, also known as a step model was also built to refine the predictor variables. After stepwise selection, the final reduced model included 'WEIGHT', 'BMI', 'RATERISK', and 'FRACSCORE' as the predictor variables. The intercept value of -3.15568 can be interpreted as the log-odds of a fracture occurring when all predictors are zero is -3.15568. This makes it highly significant as the p-value < 0.001. For 'WEIGHT', the estimate is -0.05020 which can be interpreted as for each additional kilogram of weight, the log-odds of having a fracture decreases by 0.05020. In terms of odds, exp(-0.05020) ≈ 0.951, suggesting about a 4.9% decrease in the odds of fracture per kilogram increase in weight. Higher body weight appears to be protective against fractures. This may be because increased weight could lead to higher bone density or because heavier individuals may have a larger cushion effect during falls, reducing fracture risk. For 'BMI', the estimate is 0.14460, which can be interpreted as for each unit increase in BMI, the log-odds of having a fracture increases by 0.14460. In terms of odds, exp(0.14460) ≈ 1.155, indicating about a 15.5% increase in the odds of fracture per unit increase in BMI. Contextually, this can be understood as a high BMI being associated with an increased risk of fractures. This could be due to a higher likelihood of falling or other health complications associated with higher BMI. Both, 'WEIGHT' and 'BMI' are statistically significant due to their p-values of 0.0305 and 0.0258 respectively. An academic paper written by Compston et al. (2014) discusses the relationship that weight, BMI, and height can have on fracture risk in postmenopausal women. Their abstract mentions how a low BMI value is a risk for fractures, along with height and obesity, which can further increase the risk of a fracture.

The estimate of 'RATERISK' is 0.29587, which can be interpreted as for each unit increase in the risk rating score, the log-odds of having a fracture increases by 0.29587. In terms of odds, exp(0.29587) ≈ 1.344, suggesting about a 34.4% increase in the odds of fracture per unit

increase in risk rating. The higher the risk rating, the higher the probability of fractures. This risk rating score likely encapsulates various factors contributing to bone health and fracture risk, such as bone density, history of falls, and other health conditions. This variable is marginally significant due to p-value = 0.0830.

The last variable of 'FRACSCORE' has an estimate of 0.25066, which can be interpreted as for each unit increase in the fracture score, the log-odds of having a fracture increases by 0.25066. In terms of odds, $\exp(0.25066) \approx 1.285$, indicating about a 28.5% increase in the odds of fracture per unit increase in fracture score. This variable is highly significant (due to its p-value < 0.001. Higher fracture scores are strongly associated with increased fracture risk, reflecting the cumulative impact of various risk factors on bone fragility and propensity to fractures.

This model had a residual deviance of 357.31 on 347 degrees of freedom, representing the deviance of the final model with the selected predictors. The AIC was 367.31, providing a measure of the model's fit, penalizing for the number of predictors used. In general, a lower AIC indicates a better model. This suggests a potentially simpler model that retains predictive power.

Conclusion

This analysis aimed to identify the factors that significantly affect the likelihood of fractures among postmenopausal women in the GLOW study using logistic regression. The findings revealed that age, BMI, prior fractures, early menopause, maternal history of hip fractures, requiring arm assistance when standing up, smoking status, self-reported risk of fracture, and the calculated fracture risk score all play roles in determining fracture risk. Particularly, the BMI and weight of an individual were found to have significant effects, with a higher BMI increasing fracture risk and higher weight decreasing it. The composite fracture risk

score also emerged as a significant predictor, highlighting the cumulative impact of multiple risk factors on the likelihood of a fracture occurring.

Despite these insights, the analysis faced several limitations. The dataset used was limited to 500 observations, which may not be fully representative of the broader population. Additionally, the data was collected from 2009 to 2014, and changes in medical practices or demographic trends since then could affect the generalizability of the findings. Some potentially influential variables might not have been included in the model. For instance, factors like physical activity levels, dietary habits, and medication use were not considered but could significantly impact fracture risk. Logistic regression, while useful, has its limitations. It assumes a linear relationship between the predictors and the log-odds of the outcome, which might not capture more complex, non-linear interactions between variables. The binary outcome (fracture vs. no fracture) does not account for the severity or type of fractures, which could be important for more nuanced risk assessments.

To enhance the robustness and applicability of future analyses, several improvements could be made to build a more efficient model. Including a larger and more diverse sample size would improve the representativeness of the findings. Additionally, incorporating more recent data would ensure the results are relevant to current clinical practices. As mentioned, integrating more comprehensive variables such as physical activity, dietary intake, medication history, and other health conditions could provide a more holistic view of the factors influencing fracture risk. Exploring more sophisticated modeling approaches could capture the complex interactions and non-linear relationships between variables, potentially improving predictive accuracy. Lastly, differentiating between types and severities of fractures in the outcome variable could provide more detailed insights and help tailor prevention strategies more effectively.

References

Code Sheet for Variables in the GLOW Study. (2014). Login.microsoftonline.com. https://learn.umgc.edu/content/enforced/1031259-027339-01-2245-GO1-9041/Dataset%20Descriptions/GLOW_Code_Sheet.pdf?ou=1031259

Compston, J. E., Flahive, J., Hosmer, D. W., Watts, N. B., Siris, E. S., Silverman, S., Saag, K. G., Roux, C., Rossini, M., Pfeilschifter, J., Nieves, J. W., Netelenbos, J. C., March, L., LaCroix, A. Z., Hooven, F. H., Greenspan, S. L., Gehlbach, S. H., Díez-Pérez, A., Cooper, C., & Chapurlat, R. D. (2014). Relationship of Weight, Height, and Body Mass Index With Fracture Risk at Different Sites in Postmenopausal Women: The Global Longitudinal Study of Osteoporosis in Women (GLOW). *Journal of Bone and Mineral Research*, *29*(2), 487–493. https://doi.org/10.1002/jbmr.2051

Hooven, F. H. (2014). *An Overview of The Global Longitudinal Study of Osteoporosis in Women (GLOW)*. [PowerPoint slides]. Outcomes-Umassmed.org. https://www.outcomes-umassmed.org/GLOW/publicfiles/GLOW_Overview.ppt

Hooven, F. H., Adachi, J. D., Adami, S., Boonen, S., Compston, J., Cooper, C., Delmas, P., Diez-Perez, A., Gehlbach, S., Greenspan, S. L., LaCroix, A., Lindsay, R., Netelenbos, J. C., Pfeilschifter, J., Roux, C., Saag, K. G., Sambrook, P., Silverman, S., Siris, E., & Watts, N. B. (2009). The Global Longitudinal Study of Osteoporosis in Women (GLOW): rationale and study design. *Osteoporosis International*, *20*(7), 1107–1116. https://doi.org/10.1007/s00198-009-0958-2

Perez, M. O., Pedro, P. P. de A., Lyrio, A. M., Grizzo, F. M. F., & Loures, M. A. A. da R. (2023). Osteoporosis and fracture risk assessment: improving outcomes in postmenopausal

women. *Revista Da Associação Médica Brasileira*, *69*, e2023S130. https://doi.org/10.1590/1806-9282.2023S130

Watts, N. B. (2014). Insights from the Global Longitudinal Study of Osteoporosis in Women (GLOW). *Nature Reviews Endocrinology*, *10*(7), 412–422. https://doi.org/10.1038/nrendo.2014.55

Welcome to GLOW. (2009). Www.outcomes-Umassmed.org. https://www.outcomes-umassmed.org/GLOW/

Appendix A

All figures mentioned and visualizations produced in the report can be seen below. The R code is attached as a separate file.

Figure 1

Figure of the model summary.

```
> summary(model)

Call:
glm(formula = FRACTURE ~ AGE + WEIGHT + HEIGHT + BMI + PRIORFRAC +
    PREMENO + MOMFRAC + ARMASSIST + SMOKE + RATERISK + FRACSCORE,
    family = binomial, data = train.data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.397361  15.071790  -0.955    0.339
AGE           0.007735   0.061339   0.126    0.900
WEIGHT       -0.128169   0.105361  -1.216    0.224
HEIGHT        0.068629   0.094815   0.724    0.469
BMI           0.341515   0.275732   1.239    0.216
PRIORFRAC     0.334511   0.450379   0.743    0.458
PREMENO       0.271627   0.332858   0.816    0.414
MOMFRAC       0.399052   0.481162   0.829    0.407
ARMASSIST     0.184191   0.675144   0.273    0.785
SMOKE        -0.344659   0.601029  -0.573    0.566
RATERISK      0.219350   0.184925   1.186    0.236
FRACSCORE     0.177328   0.306321   0.579    0.563

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 393.67  on 351  degrees of freedom
Residual deviance: 353.82  on 340  degrees of freedom
AIC: 377.82

Number of Fisher Scoring iterations: 4
```

Figure 2

Figure of the residual plot of the linear regression model.