Objective: To build a logistic regression model to analyse whether cancer is benign or malignant using features obtained from cell nuclei of a breast mass in the Wisconsin Diagnostic Breast Cancer dataset [1].

Data Summary: The Wisconsin Diagnostic Breast Cancer dataset [1] contains 569 instances spread across 32 attributes (ID, diagnosis, 30 real-valued input features). The attributes describe characteristics of the cell nuclei present in the digitized image of a fine needle aspirate (FNA) of a breast mass. Ten real-valued features are computed for each cell nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The mean, standard error, and "worst" of these features were computed for each image, resulting in 30 features. Out of these I'll be making use of mean values of 3 numeric features (area_mean, smoothness_mean, symmetry_mean) to make a prediction about the binary outcome variable (diagnosis): whether the cancer is benign or malignant.

Predictor variables						Outcome variable		
Variable	**Class**	**Min**	**Median**	**Mean**	**Max**	**Variable**	**Class**	**Values**
Area_mean	Continuous	143.5	551.4	655.7	2501	Diagnosis	Factor	Benign(B): 357 Malignant(M): 212
Smoothness_mean	Continuous	0.062	0.095	0.096	0.163			
Symmetry_mean	Continuous	0.106	0.179	0.181	0.304			

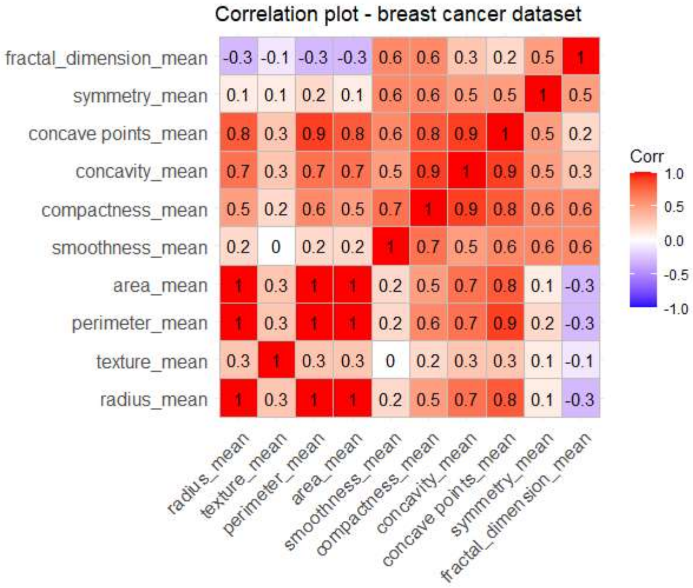
Data Cleaning: There wasn't any missing data for this dataset. I filtered out the dataset to select only the required predictor and outcome variables for further analysis. Our outcome variable: "diagnosis" was originally a character variable; thus, I converted it into a factor variable so that it can be identified as a type of binary variable having 2 levels: B: "Benign" and M: "Malignant".

In the original paper for this dataset [3], the authors mention that "the features are numerically modeled such that larger values will typically indicate a higher likelihood of malignancy". That is why I didn't choose to perform an outlier analysis.

Planning: Since we have 1 factor variable: diagnosis, thus I selected it as my outcome variable for analysis using logistic regression. To select my predictor variables, I built a correlation plot among the mean values of the continuous variables and selected 3 variables having < 0.7 correlation values with each other.

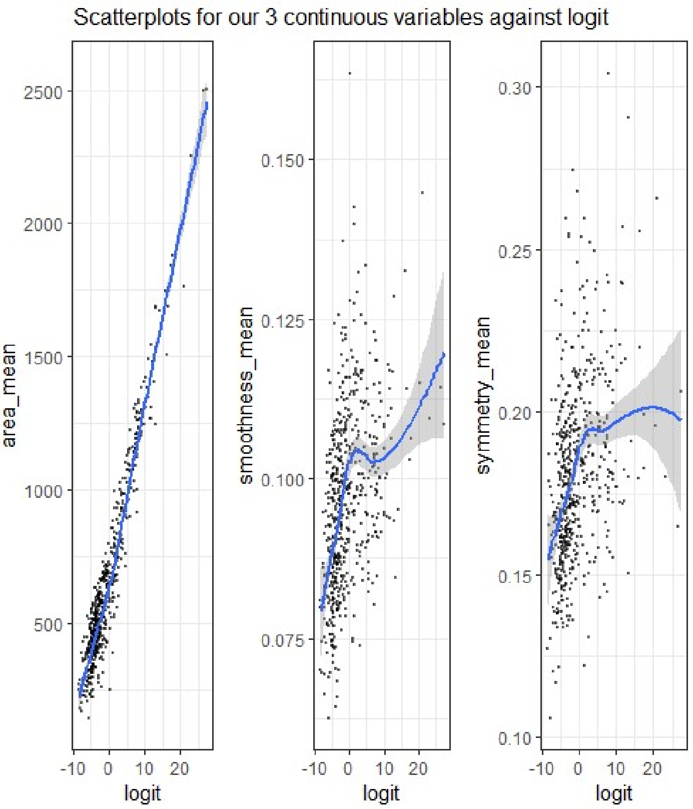
Checking assumptions:

-> Multicollinearity: To check the collinearity among my 3 predictor variables I used Pearson's correlation (since they're continuous variables); and I plotted them against each other for visual inspection. The obtained correlation (pairwise) between them was <0.7 which meant that they were not highly correlated with each other. Also, visual inspection looked good, thus they were selected.



Analysis: Since we're predicting a binary outcome variable from a set of continuous predictor variables thus, I used the glm()[Generalised Linear Model] function to build my logistic regression model. Upon summarizing, we see that all our predictor variables have a p-value<0.05 which meant they significantly affected the outcome. Also, we see there's significant difference between the Null and Residual deviance, which denotes the model is a great fit. Also, having a lower AIC value only makes it better.

<u>Analysing Assumptions</u>:



-> <u>Linearity of independent variables and log-odds</u>: In logistic regression, we check linearity by adding in the interaction effects, x*log(x) of each predictor variable into the model and checking their significance. In our case, none of the x*log(x) variables are significant (p>0.05), so we do not reject the assumption of linearity for these variables. Furthermore, I calculated the log-odds of the dependent variable and plotted it against each predictor variable. Scatterplots displayed fairly linear relationship among them. For instance, odds ratio (e^B) for smoothness_mean was found to be between 1.48e+20 and 9.90e+46 meaning with each unit increase in smoothness_mean, nuclei is significantly more likely to be Malignant (specifically 1.48e+20 to 9.90e+46 times more likely), at 5% level of significance. Also, through the Wald test, we can see all our continuous variables are significant predictors for Malignancy.

-> <u>Multicollinearity</u>: We further inspected multicollinearity by VIF (Variance Inflation factor). The largest VIF was 1.73, much less than 10; the average VIF was 1.54, closer to 1. The lowest tolerance (1/VIF) was 0.57, much greater than 0.1 (which would indicate a serious problem) and 0.2 (which indicates a potential problem). We thus conclude that there is no collinearity in the data.

-> <u>Independence of errors</u>: The Durbin-Watson test for independent errors was significant at the 5% level of significance (d=1.72, p=0). As 'd' is a little away from 2, it shows some degree of positive autocorrelation; and it meant the model could be fine tuned, but since D-W statistic is between 1 and 3, we fail to reject the null hypothesis that the errors are independent and continue with the assumption of independence met.

Conclusion: A logistic regression model was built to predict whether cancer diagnosis is benign or malignant using mean values of area, smoothness, and symmetry features of a cell nucleus present in the digitized image of a fine needle aspirate (FNA) of a breast mass. All assumptions of the logistic model were met.

From this model, we conclude that area, smoothness, and symmetry of the cell nuclei are all significantly involved in predicting whether cancer is benign or malignant. In addition, we can predict the influence of these variables on prediction - for instance, if nothing else is changed, each unit increase in smoothness, nuclei is 1.48e+20 to 9.90e+46 times more likely to be malignant, at 5% level of significance.

References:

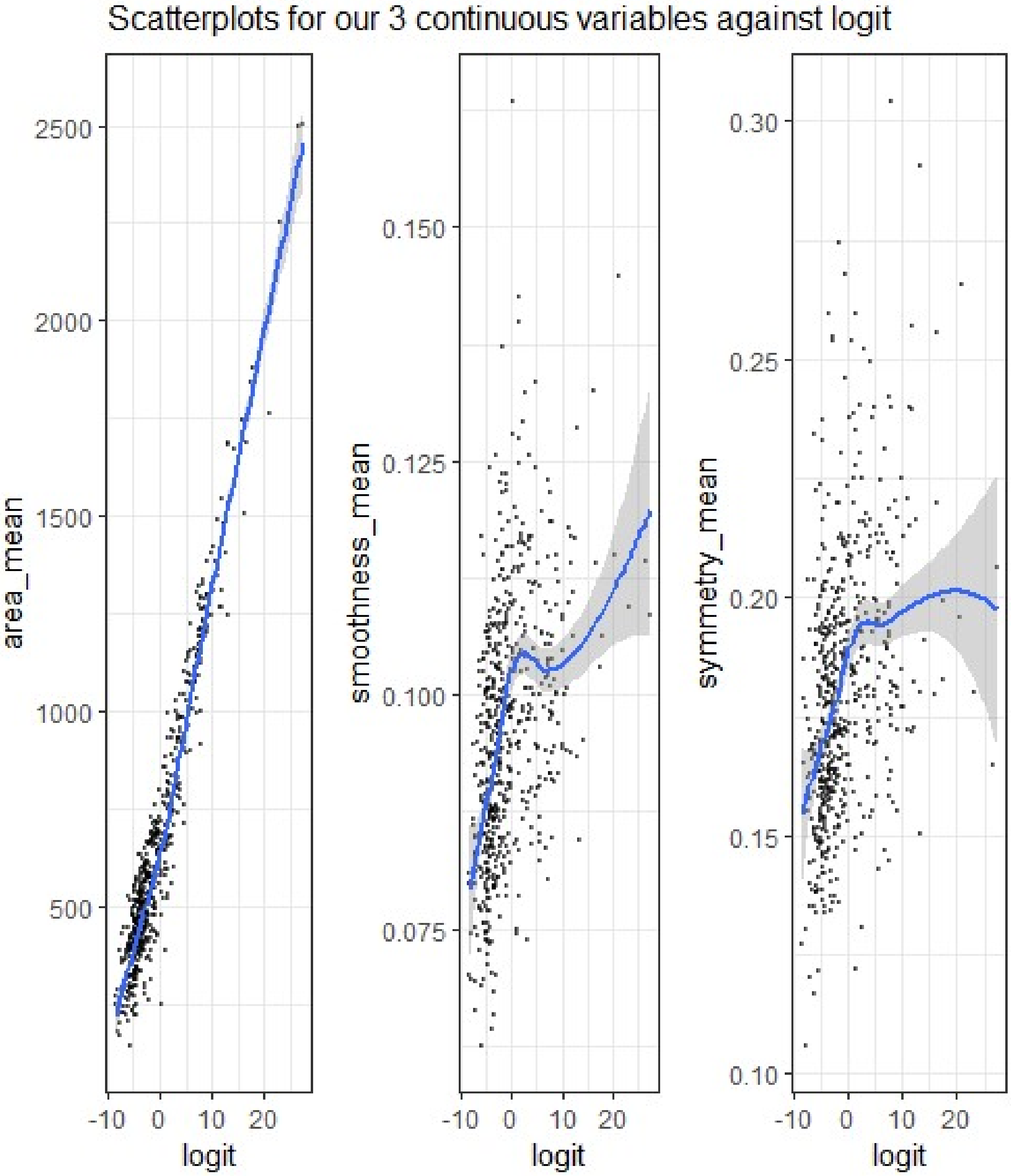
[1] "Breast Cancer Wisconsin (Diagnostic) Data Set," *Kaggle*, Sep. 25, 2016. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

[2] "UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set." https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[3] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," *Proceedings of SPIE*, Jul. 1993, doi: 10.1117/12.148698.

[4] K. Leung, "Assumptions of Logistic Regression, Clearly Explained," *Medium*, Oct. 04, 2022. https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290

Appendix:


Scatterplots for our 3 continuous variables against logit
area_mean
2500
2000
1500
1000
500
-10
0
10
20
logit
smoothness_mean
0.150
0.125
0.100
0.075
-10
0
10
20
logit
symmetry_mean
0.30
0.25
0.20
0.15
0.10
-10
0
10
20
logit