Predictive Modeling of Breast Cancer Malignancy from FNA-extracted Cell Nuclei Features

Angela Yoon

1 Introduction

Breast cancer is a disease in which malignant cancer cells form in the tissues of the breast. It is the most common cancer in American women except for skin cancers, with 1 in 8 American women expected to develop breast cancer in her lifetime. Fortunately, early detection of breast cancer can raise the 5-year survival rate to as high as 99%, and in the long run, death rates from breast cancer have been declining since about 1990, greatly in part due to early detection.[1] The commonness of the disease and the significance of its early detection motivate the necessity of detection methods that can accurately diagnose malignant cells.

When there is a suspicious area in an individual's breast such as a lump, thickening, or finding revealed through ultrasound scans or MRIs, a procedure called breast biopsy is conducted to evaluate the area to determine whether it is breast cancer. The procedure involves removing a small sample of breast tissue for laboratory testing. Fine-needle aspiration (FNA) biopsy is the simplest among such breast biopsy procedures. It involves directing a very thin needle into the lump to collect a sample of cells.[2] Aside from the other key advantages FNA has over other biopsy modalities—such as simplicity, cost-effectiveness, and low invasiveness and pain—FNA is also important as it is often the only available testing method in medically under-resourced developing countries, which represent more than 80% of the world's population, where imaging and preoperative diagnosis are not readily available and other methods such as core-needle biopsy or histopathology are unaffordable by most.[3] [4]

A review and quality assessment of 46 studies on the accuracy of FNA for evaluating breast lesions, conducted by Yu *et al.*, concludes that FNA reliably diagnoses most benign and malignant breast lesions with high sensitivity and specificity.[5] However, it is worth noting that the clinical use of FNA has been continuously questioned due to high variability in reported results. Moreover, a major disadvantage of FNA is that it may require a more experienced cytopathologist (a trained expert in diagnosis of diseases through studying cells obtained from body fluids) for diagnosis, which is a resource that may not be accessible for many in medically under-resourced areas where FNA is especially useful.[6] [7]

Therefore, the question of accurately predicting breast mass malignancy based on easily interpretable and extractable features from FNA biopsy is an inquiry that can contribute greatly to making early detection of breast cancer more accessible and subsequently to tackling a major public health challenge. To address this question, this project will examine a dataset where FNA biopsy from breast masses were converted to a digitized image and then to a list of numeric values that represent features of the image. Using a Bayesian modeling framework, the project aims to explore ways of how to accurately predict the malignancy of the breast mass in the image based on such features, by doing the following:

1) Employing a general linear model (GLM) to model the relationship between malignancy and image features
2) Assigning a weakly informative normal prior to the coefficients of the linear model
3) Carefully selecting predictor variables with the most predictive power
4) Comparing regressions with different link functions to find the best model in terms of posterior predictivity

[1] https://www.nationalbreastcancer.org/breast-cancer-facts
[2] https://www.mayoclinic.org/tests-procedures/breast-biopsy/about/pac-20384812
[3] https://www.ncbi.nlm.nih.gov/books/NBK470268/
[4] https://acsjournals.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/cncy.21822
[5] https://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-12-41
[6] https://www.ncbi.nlm.nih.gov/books/NBK470268/
[7] https://freida.ama-assn.org/specialty/cytopathology-pth

2 Data

The data used in this project comes from the Kaggle dataset "Breast Cancer Wisconsin (Diagnostic) Data Set".[8] It is publicly available through the UW CS ftp server and can also be found on the UCI Machine Learning Repository. It was initially created by Dr. William H. Wolberg, a physician at the University of Wisconsin Hospital, with 569 digitized images of FNAs of breast masses.

For each image, the dataset records the ID number and diagnosis (whether the breast mass in the image is malignant or benign). Also, ten real-valued features were computed from each one of the cells in the image, and then the mean value, maximum (or "worst"), and standard error was returned for each feature across all cells present in the image. The ten features are 'radius', 'texture', 'perimeter', 'area', 'smoothness', 'compactness', 'concavity', 'concave points', 'symmetry', and 'fractal dimension'. (Refer to Appendix I for full explanation on each variable.) All of the features were numerically modeled such that larger values will typically indicate a higher likelihood of malignancy.[9]

'Diagnosis', a nominal discrete (binary) variable, will be the response variable. The rest are predictor variables; there are a total of 30 of them (3 values for each of the 10 features), all of them continuous. As the goal of this project is to get down to the smallest number of variables that are strong predictors of the response and that minimizes prediction widths, the initial model will include all possible predictors, and then will be reduced to the set of variables that have the most significant effect in explaining the prediction.

The dataset was mutated so that the response has value 1 if malignant (212 observations) and 0 if benign (357 observations). The 'ID' column, which is unnecessary for our analysis, was deleted. Using this finalized dataset, Exploratory Data Analysis was conducted. Paired with relevant literature review, the following takeaways of EDA will be considered in the model fitting process (Refer to Appendix II for relevant visualizations):

1) For all features, the mean and maximum values have a positive linear correlation. The strength of the correlation varies by feature but is strong in general. In the 1992 paper "Nuclear Feature Extraction For Breast Tumor Diagnosis" written by Dr. Wolberg himself among several others, the authors note that extreme (maximum) values are the most intuitively useful, since only a few malignant cells may occur in a given sample.[10]
2) The variables 'radius_worst' and 'perimeter_worst' have a strong positive linear correlation, and 'area_worst' has a strong positive linear correlation with 'radius_worst' squared.
3) It is known that cancer cell nuclei are often misshapen and "blebs" can be observed in its membranes, as opposed to normal cell nuclei with smooth, uniform, spheroid shapes.[11] Based on this knowledge, grouped boxplots were created to visualize how the distributions of maximum radius, symmetry, concavity, concave points, and smoothness differ for benign and malignant response groups. For all visualized predictor variables, the distribution for the malignant group had a notably greater center value, suggesting possibility of correlation between these predictors and the response.

3 Model

We will use regression to understand and predict how the response Y changes over the covariate space for $\boldsymbol{x}$, the vector of predictors. More specifically, we will be using a generalized linear model (GLM) with probability model $Y|\theta \sim Bern(\theta)$, with the following probability mass function,

$$p(y\,;\theta) = \{\begin{matrix} 1-\theta & \text{if } y=0 \\ \theta & \text{if } y=1 \end{matrix}$$

[8] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
[9] https://www.researchgate.net/publication/2512520_Nuclear_Feature_Extraction_For_Breast_Tumor_Diagnosis
[10] https://www.researchgate.net/publication/2512520_Nuclear_Feature_Extraction_For_Breast_Tumor_Diagnosis
[11] https://www.pnas.org/content/early/2013/02/06/1300215110

to reflect that our response $Y \in \{0,1\}$ is a binary variable. The systematic component of the model will take the linear form $\eta = \boldsymbol{x}^T\boldsymbol{\beta} = x_1\beta_1 + \cdots + x_p\beta_p$, where $\boldsymbol{\beta}$ is the vector of coefficients and p is the dimension of $\boldsymbol{\beta}$.

3.1 Prior Model

A weakly informative prior model will be used as our prior on $\boldsymbol{\beta}$. This is because little information is given on how features were calculated, and how they tend to relate to the response. For example, we do not know how smoothness relates with whether a tissue is malignant for cells outside of this dataset, to what degree, and with how much variance. It is hard to make a fair guess too, as we have little information on how cell features were translated into the numeric values in this dataset, or what a unit increase in a feature would mean for a cell. Therefore, we will place little weight on prior knowledge.

We will therefore use a rstanarm default prior that is normal—considering that normal distributions often well reflect distributions of natural or biological observations—and intended to be weakly informative, which looks like the following given our Bernoulli probability model:

1) for coefficients $\beta_1, \dots, \beta_p$,

$$\beta_k \sim Normal(0, \frac{2.5}{s_x}), \text{ or } p(\beta_k) = \frac{1}{\sqrt{2\pi\left(\frac{2.5}{s_x}\right)^2}} e^{-\frac{1}{2\left(\frac{2.5}{s_x}\right)^2}(\beta_k - 0)^2} = \frac{s_x}{\sqrt{12.5\pi}} e^{-\frac{s_x^2}{12.5}\beta_k^2}$$

where $s_x = sd(x)$, or the standard deviation of the covariates.

2) for the intercept β_0,

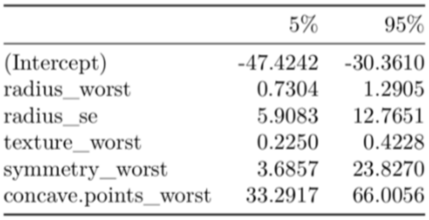
$$\beta_0 \sim Normal(0, 2.5), \text{ or } p(\beta_0) = \frac{1}{\sqrt{2\pi(2.5)^2}} e^{-\frac{1}{2(2.5)^2}(\beta_0 - 0)^2} = \frac{1}{\sqrt{12.5\pi}} e^{-\frac{\beta_0^2}{12.5}}$$

3.2 Variable Selection

To narrow down the predictor variables, we first fit a full model with all variables. At this step, the logit function ($g(\theta) = \log{(\theta/1 - \theta)}$) was used as the link function for our GLM. (Refer to Appendix III for 90% Bayesian uncertainty intervals for all coefficients of the full model.)

Notice that ‘radius_se’, ‘texture_worst’, and ‘symmetry_worst’ are the three variables for which the 90% posterior interval does not include 0. Because having too many variables can confound relationships present in the regression, we cut down variables, making sure to include the three variables that seem to be predictive of the response. For all other variables, based on EDA findings and initial fitting results, we omitted all mean or standard error variables (as mean and maximum values have potential for multicollinearity, and the maximum values have predictive significance over the means and standard errors) and perimeter and area variables (as they relate to radius), and refit the model. (Refer to Appendix III for 90% Bayesian uncertainty intervals for all coefficients of this improved model.)

From this model, we will keep only the variables with coefficient posterior intervals that do not include 0: ‘radius_worst’, ‘radius_se’, ‘texture_worst’, ‘symmetry_worst’, and ‘concave.points_worst’. We refit the model on the five predictors and check the new coefficient confidence intervals, to obtain:



	5%	95%
(Intercept)	-47.4242	-30.3610
radius_worst	0.7304	1.2905
radius_se	5.9083	12.7651
texture_worst	0.2250	0.4228
symmetry_worst	3.6857	23.8270
concave.points_worst	33.2917	66.0056

Table 1: *90% Bayesian uncertainty intervals for coefficients of the logistic regression model with the final set of predictor variables*

3.3 Link Function Selection

Within the same framework for GLM with the Bernoulli probability model and the linear systematic component, we will consider two additional link functions other than the logit function:

1) The probit function, or $g(\theta) = \Phi^{-1}(\theta)$, where Φ is the cumulative distribution function of the standard normal distribution. We will look into this option as there might be a latent (or unobserved) structure related to the malignancy of a breast mass that affects the predictor variables

2) The complementary-log-log function, or $g(\theta) = \log(-\log(1-\theta))$. We will look into this option as the cloglog link is asymmetrical in θ and may be able to better account for the unbalance in the response.

The probit and cloglog regression models were fit on the same selected variables from 3.2. 4 models—the null (logistic) regression model and the linear, probit, and cloglog regression models—were compared through model evaluation.

	elpd_diff	se_diff
logit	0.0	0.0
null	-330.6	10.2

	elpd_diff	se_diff
logit	0.0	0.0
probit	-1.2	2.2

	elpd_diff	se_diff
cloglog	0.0	0.0
logit	-1.0	1.4

Table 2: *Results from using the loo_compare function to compare pairs of models*

First, leave-one-out cross validation was performed on each model, and models were compared based on validation results. From the first table, we can see that the difference in expected log predictive density is much larger than several times the estimated standard error of the difference, indicating that the logistic regression model is expected to have better predictive performance than the null model. This is not the case for the other two comparisons, however, where the logit model is compared to the probit and cloglog models respectively. Considering the small magnitude of elpd differences with respect to the standard error of the differences, and that many problematic observations (232 for probit, 165 for cloglog) were reported while computing the LOO, these results reflect the generally strong predictivity of the logit, probit, and cloglog models but are insufficient evidence to claim that any one is better than the others.

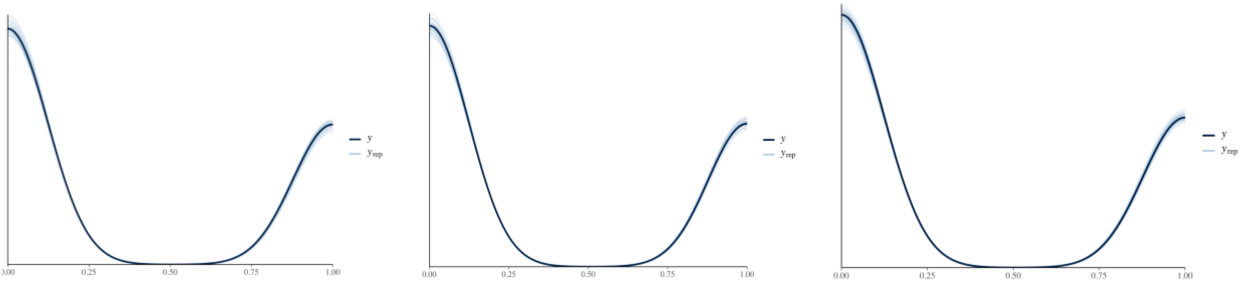
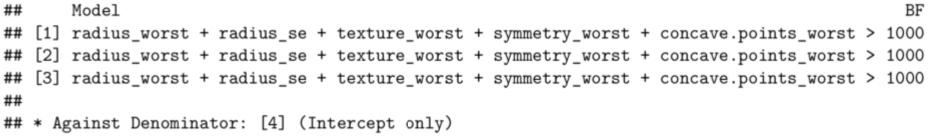


Figure 1: *Graphical posterior predictive check (PPC) results for logit, probit, and cloglog (from left to right) models*

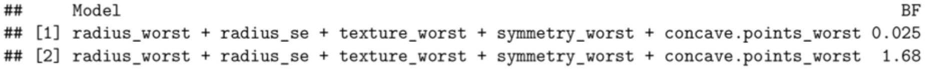
Next, PPC plots were generated using the bayesplot package to see how the distribution of the observed data (dark blue) compares to simulated data from the posterior predictive distributions (light blue). Notice that our response has values 0 and 1; although the distribution is plotted as continuous with a line, the convergence around the tails is what must be observed to see if the simulated data looks like the observed data. Based on this, all three models seem to be a good fit of the data; it is hard to tell if any one of them is better than the others.

Lastly, we examined the Bayes Factor, defined as $BF(\mathcal{M}_1, \mathcal{M}_2) = \frac{[\boldsymbol{y}|\mathcal{M}_1]}{[\boldsymbol{y}|\mathcal{M}_2]}$.



```
##     Model                                                                      BF
## [1] radius_worst + radius_se + texture_worst + symmetry_worst + concave.points_worst > 1000
## [2] radius_worst + radius_se + texture_worst + symmetry_worst + concave.points_worst > 1000
## [3] radius_worst + radius_se + texture_worst + symmetry_worst + concave.points_worst > 1000
##
## * Against Denominator: [4] (Intercept only)
```

For all three models, when the Bayes Factor is computed with the null model at the denominator, the value exceeds 1000, indicating that our regression models are strongly preferred against the null.



```
##     Model                                                                    BF
## [1] radius_worst + radius_se + texture_worst + symmetry_worst + concave.points_worst 0.025
## [2] radius_worst + radius_se + texture_worst + symmetry_worst + concave.points_worst  1.68
```

For the probit and cloglog models, when the Bayes Factor is computed with the logit model at the denominator, $BF < 1$ for probit and $BF > 1$ for cloglog. We can therefore conclude that we should select the cloglog regression model over logit, and logit over probit.

3.4 MCMC Convergence and Mixing Performance

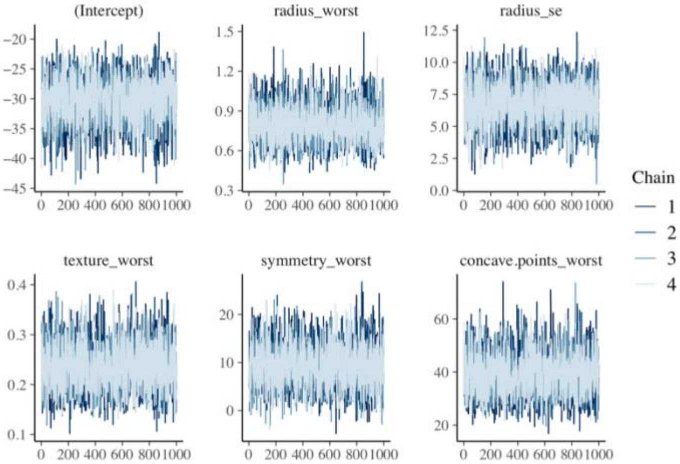
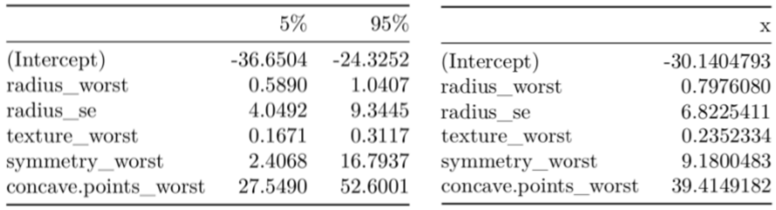


Figure 2: *Trace plots created for the cloglog regression model. Shows the sampled values of the parameters over MCMC iterations.*

As a post-decision examination, we checked MCMC convergence and mixing performance for the cloglog regression model. The trace plots suggest that the model has converged, because for all parameters, the four chains have mixed and there are no clear trends.

4 Results

To analyze the results of the posterior distribution, we will revisit the two main objectives of this Bayesian modeling task.



	5%	95%
(Intercept)	-36.6504	-24.3252
radius_worst	0.5890	1.0407
radius_se	4.0492	9.3445
texture_worst	0.1671	0.3117
symmetry_worst	2.4068	16.7937
concave.points_worst	27.5490	52.6001

	x
(Intercept)	-30.1404793
radius_worst	0.7976080
radius_se	6.8225411
texture_worst	0.2352334
symmetry_worst	9.1800483
concave.points_worst	39.4149182

Table 3: *90% Bayesian uncertainty intervals and mean values of coefficients for the final cloglog regression model*

First of all, we were able to identify a set of features easily extractable from FNA that are strong predictor variables of malignancy, as seen in the coefficient means and posterior intervals above. These features are extractable as long as the FNA-digital image conversion and cell feature computation techniques can be performed, which are well explained in pertinent literature.

One limitation is that due to the usage of the cloglog link function as well as the ambiguity of unit used for the predictors, the interpretation of coefficients is rather difficult. Nevertheless, we can still make the interpretation that high values for maximum nucleus radius, standard error between nucleus radii, maximum texture, maximum concave points, and maximum symmetry of cells are strong predictors of malignancy; this provides more context than many “black box” classification algorithms.

Secondly, we were also able to build a posterior model that accurately predicts malignancy with the above predictors. The cloglog regression model demonstrated good predictive performance as shown by LOO and graphical PCC results and even outperformed other similarly accurate models based on Bayes Factor computations. The cloglog model outperforming the logit model is indeed an interesting result—while the dataset was unbalanced with 212 malignant and 357 benign observations, the unbalance was not extremely significant. This seems to be reflected in the Bayes Factor of the cloglog and logit models returning a value greater but not extremely greater than 1 (1.68). Checking the models’ predictive accuracy with new data from outside our dataset would be helpful in further evaluating the model.

5 Conclusion

With a weakly informative normal prior and cloglog regression model, we were able to construct a model of high predictive performance that predicts breast mass malignancy from five strong predictive features extracted from the breast mass through FNA biopsy. The modeling process and successful outcome show the potential of modeling techniques in aiding early detection of breast cancer in areas that lack medical resources and rely on FNA biopsy for diagnosis.

Works Cited

Breast Cancer Facts. National Breast Cancer Foundation, Inc. https://www.nationalbreastcancer.org/breast-cancer-facts. Accessed May 26, 2021

Breast biopsy. Mayo Clinic. https://www.mayoclinic.org/tests-procedures/breast-biopsy/about/pac-20384812. Accessed May 26, 2021

Casaubon JT, Tomlinson-Hansen S, Regan JP. Fine Needle Aspiration Of Breast Masses. [Updated 2020 Oct 28]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470268/

Cytopathology Help Desk, Cancer Cytopathology. Breast FNA Biopsy Cytology: Current Problems and the International Academy of Cytology Yokohama Standardized Reporting System. https://acsjournals.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/cncy.21822. Accessed May 26, 2021

Yu YH, Wei W, Liu JL. Diagnostic value of fine-needle aspiration biopsy for breast mass: a systematic review and meta-analysis. *BMC Cancer 12*. 41(2012). https://doi.org/10.1186/1471-2407-12-41

Cytopathology. American Medical Association. https://freida.ama-assn.org/specialty/cytopathology-pth. Accessed May 26, 2021

Breast Cancer Wisconsin (Diagnostic) Data Set. Kaggle. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. Accessed May 26, 2021

Street Nick, Wolberg W, Mangasarian O. Nuclear Feature Extraction For Breast Tumor Diagnosis. *Proc. Soc. Photo-Opt. Inst. Eng.* 1993. 10.1117/12.148698. https://www.researchgate.net/publication/2512520_Nuclear_Feature_Extraction_For_Breast_Tumor_Diagnosis

Funkhouser M *et al.* Mechanical model of blebbing in nuclear lamin meshworks. *PNAS*. 2013. https://www.pnas.org/content/early/2013/02/06/1300215110

Appendix

I. Details on "Breast Cancer Wisconsin (Diagnostic) Data Set" variables

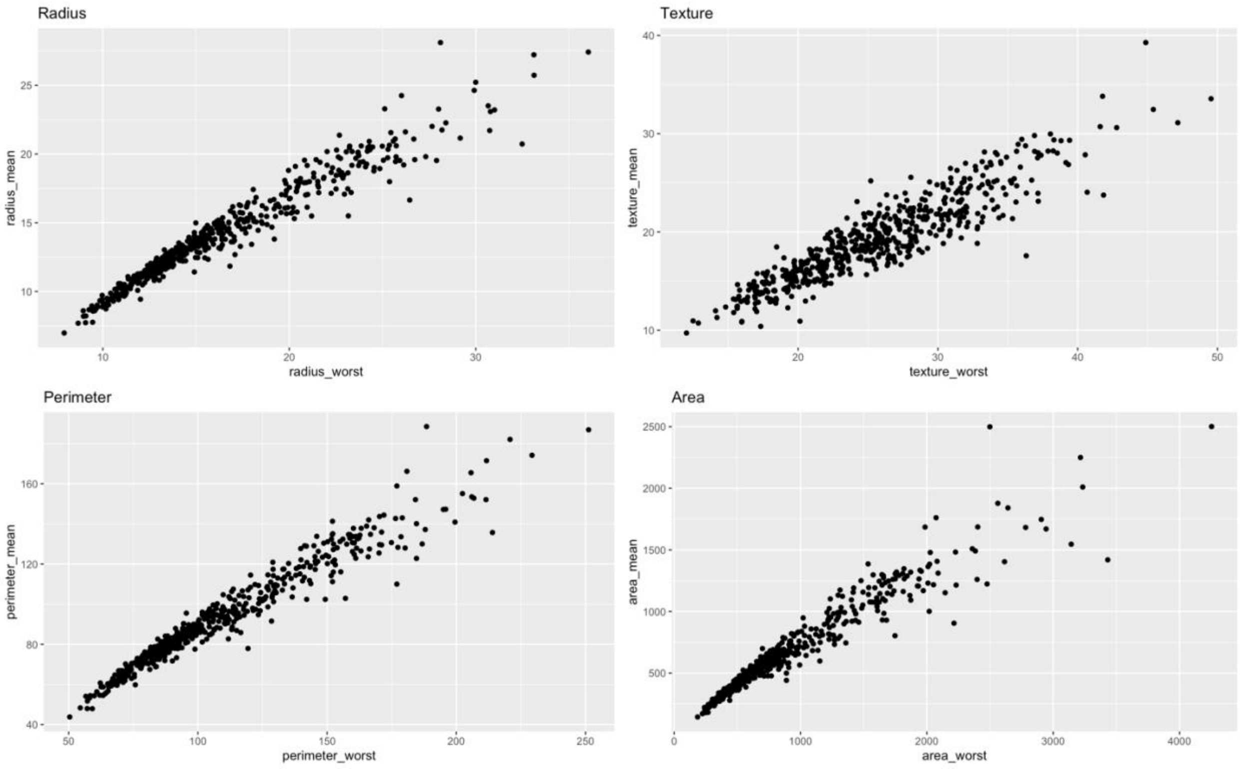
(Most descriptions are retrieved from the Kaggle dataset descriptions; for variables with missing descriptions, Street *et al.* (1992) was referred to for descriptions. The same paper also provides more details on how each feature was computed from digital images.)

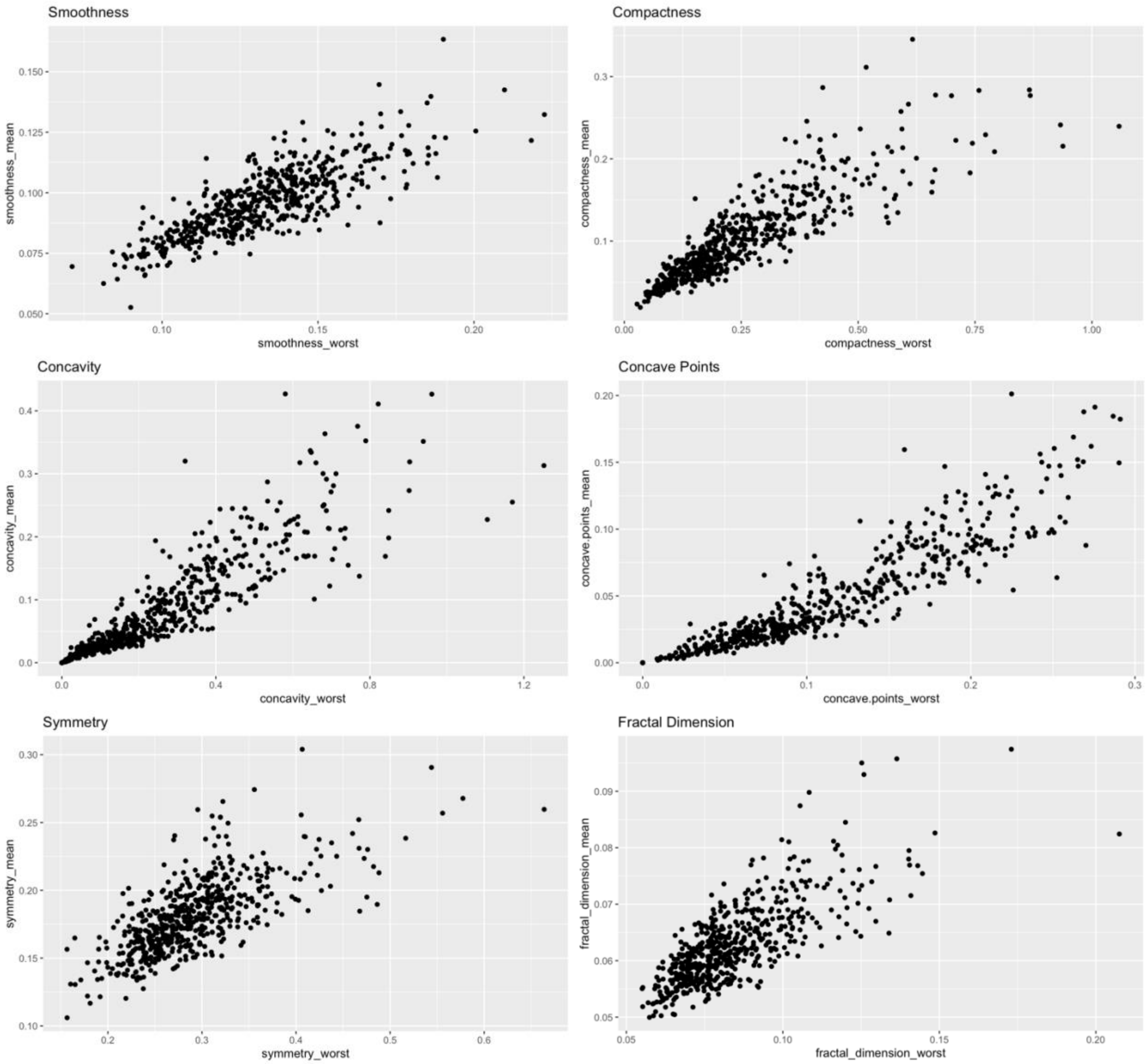
Variable	**Description**
radius	mean of distances from center to points on the nuclear perimeter
texture	standard deviation of gray-scale values
perimeter	nuclear perimeter
area	nuclear area
smoothness	local variation in radius lengths
compactness	perimeter^2 / area – 1.0
concavity	severity of concave portions of the contour
concave points	number of concave portions of the contour
symmetry	length difference between lines perpendicular to the major axis to the cell boundary in both directions
fractal dimension	"coastline approximation" – 1

The maximum, or "worst," value for each feature was computed by getting the mean of the three largest values.

II. Exploratory Data Analysis

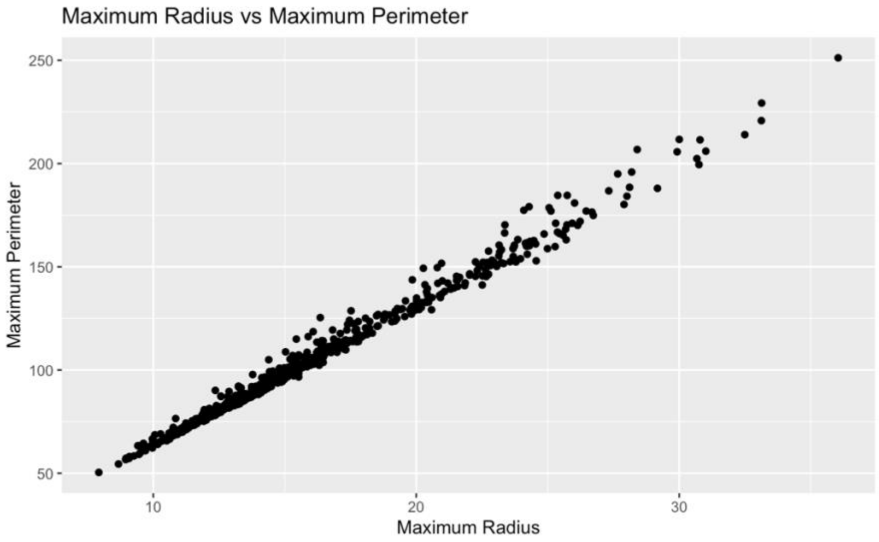
1. Mean vs maximum values for each feature

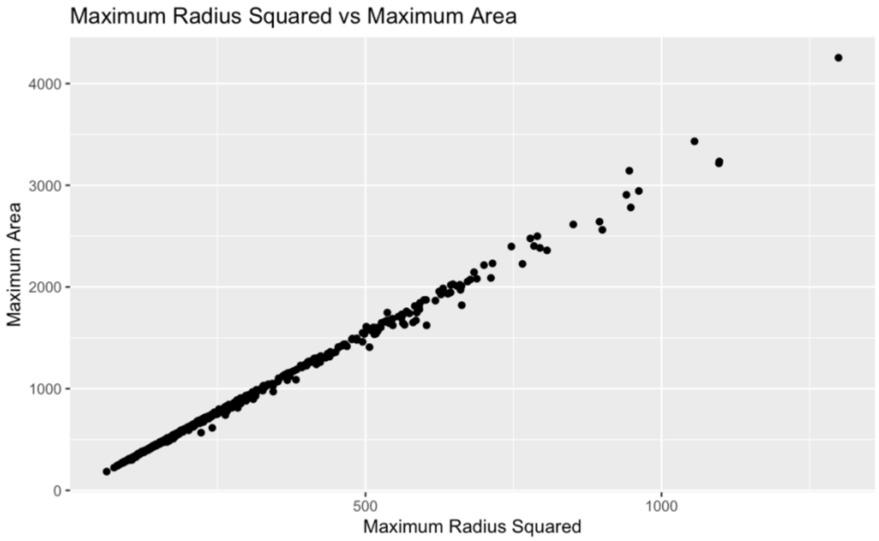




The mean and maximum value for each of the 10 features are plotted against each other as scatterplots, with the maximum ("worst") on the x axis and mean on the y axis.

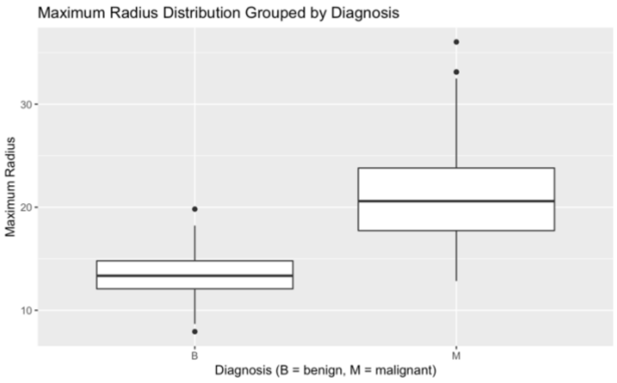
2. Radius vs Perimeter, Radius squared vs Area

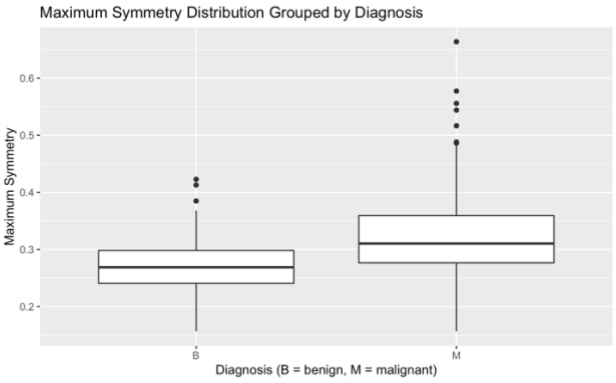


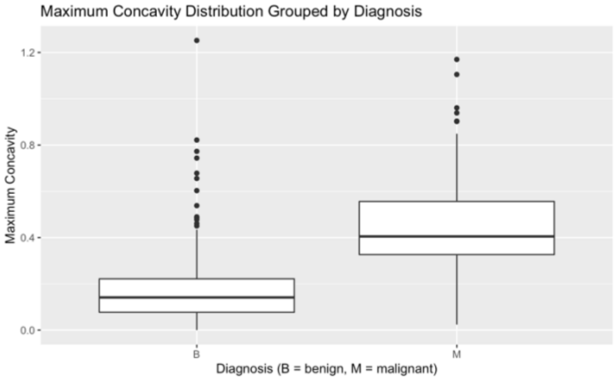


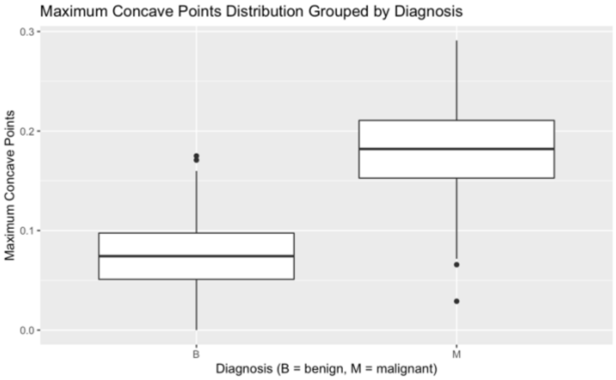
The two relationships are plotted out as scatterplots.

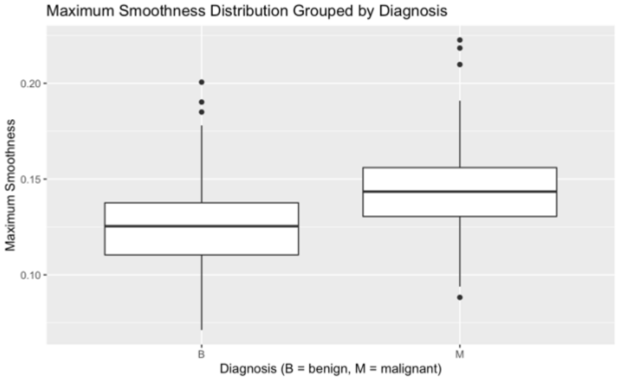
3. Distributions of predictor variables grouped by response






Maximum Concavity Distribution Grouped by Diagnosis

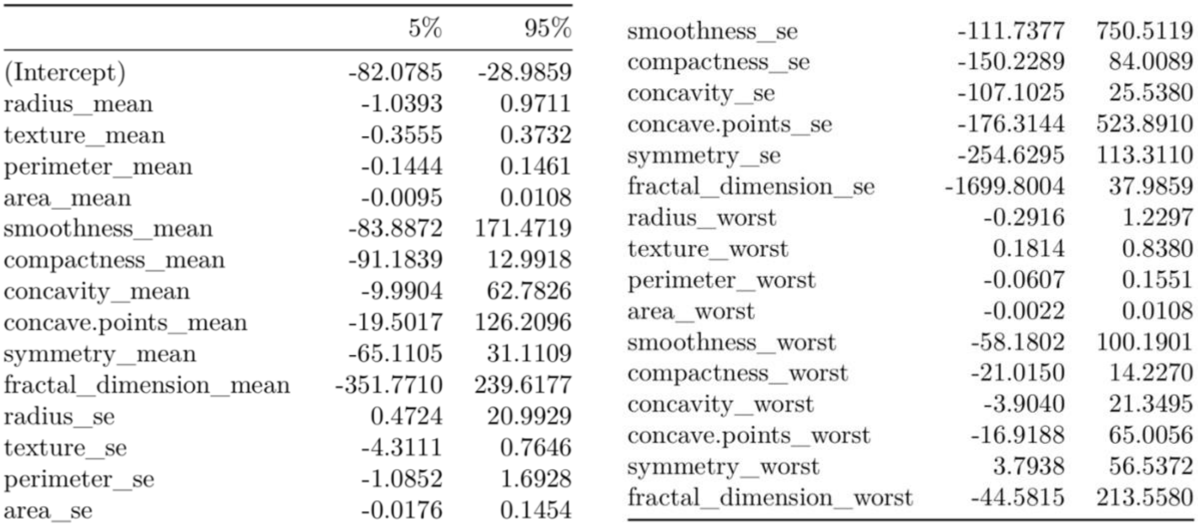

Maximum Concave Points Distribution Grouped by Diagnosis


Maximum Smoothness Distribution Grouped by Diagnosis

The boxplots demonstrate how the distributions of the five predictor variables differ for benign and malignant groups.

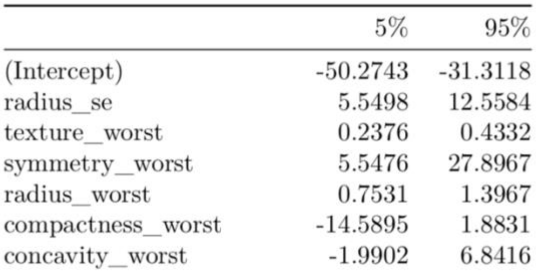
III. Variable selection process

1. 90% Bayesian uncertainty intervals for all coefficients of the full model

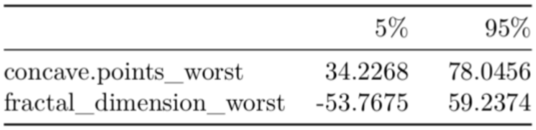


	5%	95%
(Intercept)	-82.0785	-28.9859
radius_mean	-1.0393	0.9711
texture_mean	-0.3555	0.3732
perimeter_mean	-0.1444	0.1461
area_mean	-0.0095	0.0108
smoothness_mean	-83.8872	171.4719
compactness_mean	-91.1839	12.9918
concavity_mean	-9.9904	62.7826
concave.points_mean	-19.5017	126.2096
symmetry_mean	-65.1105	31.1109
fractal_dimension_mean	-351.7710	239.6177
radius_se	0.4724	20.9929
texture_se	-4.3111	0.7646
perimeter_se	-1.0852	1.6928
area_se	-0.0176	0.1454
smoothness_se	-111.7377	750.5119
compactness_se	-150.2289	84.0089
concavity_se	-107.1025	25.5380
concave.points_se	-176.3144	523.8910
symmetry_se	-254.6295	113.3110
fractal_dimension_se	-1699.8004	37.9859
radius_worst	-0.2916	1.2297
texture_worst	0.1814	0.8380
perimeter_worst	-0.0607	0.1551
area_worst	-0.0022	0.0108
smoothness_worst	-58.1802	100.1901
compactness_worst	-21.0150	14.2270
concavity_worst	-3.9040	21.3495
concave.points_worst	-16.9188	65.0056
symmetry_worst	3.7938	56.5372
fractal_dimension_worst	-44.5815	213.5580

2. 90% Bayesian uncertainty intervals for all coefficients of the refit model



	5%	95%
(Intercept)	-50.2743	-31.3118
radius_se	5.5498	12.5584
texture_worst	0.2376	0.4332
symmetry_worst	5.5476	27.8967
radius_worst	0.7531	1.3967
compactness_worst	-14.5895	1.8831
concavity_worst	-1.9902	6.8416



	5%	95%
concave.points_worst	34.2268	78.0456
fractal_dimension_worst	-53.7675	59.2374