MSCI718 2023W Individual Assignment 4 - Logistic Regression

Data Description

- The Wisconsin breast cancer data set contains 569 records derived from the study of pictures of a fine needle aspirate of a breast mass. The first column contains a unique id, and the second column contains the diagnosis. There are no missing values, except the 33rd column which is empty.
- The outcome variable diagnosis has the levels **B (Benign 357 cases)** and **M (Malignant 212 cases)**.
- Columns 3-32 contain continuous numerical variables describing cellular properties as follows:

Ten real-valued features (**radius** (mean of distances from centre to points on the perimeter), **texture** (standard deviation of gray-scale values), **perimeter, 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, fractal dimension** ("coastline approximation" - 1) are computed for **each cell nucleus, and their mean, standard error (SE) and worst values** for each feature are observed, resulting in 30 features.
Diagnosis can be considered as outcome variables and rest of the features can be considered as predictor variables. "**It would be interesting to figure if the type of breast cancer (that is Benign or Malignant) can be predicted based on its features**".
Since the outcome variable (diagnosis) is binary, the objective of this analysis is to create a prediction model with Logistic Regression. The analysis is interpretable with Benign (diagnosis = '0') as the default case.

Building a Model

The correlation plot (shown in Appendix 2) demonstrates that some of the features (such as area, perimeter, and radius) are highly correlated. This is referred as multicollinearity, and it can have an impact on the performance of model.
To address this highly correlated dataset, **three models of features with mean, se, and worst are formed** first, and the VIF[1] of each individual group model is computed, and the variables with high correlation are removed (like perimeter and area), before performing **backward step elimination**[2] on the resulting variable in the groups.
The reason for selecting this method is that the complexity of all subset methods grows exponentially with the number of variables. Since, we have large number of features (30) backward step model is used.
Now, all the variables with significant AIC are concatenated into one model, and after applying backward step elimination for the last time and the **final model with 8 variables (radius_worst, smoothness_mean, concavity_mean, radius_se, texture_se, fractal_dimension_se, smoothness_worst, and concave.points_worst)** with **AIC of 108.81** and Residual deviation of 90.809 on 560 degree of freedom is determined. In Appendix 3 and 8, the model and step results are shown.
Now, Checking the assumptions of the Logistic Regression Model:

Assumptions

1. **Multicollinearity**: The largest value of VIF in the new model is 4.785 (less than 10) and the lowest tolerance (1/VIF) is 0.209 for smoothness_worst, (which is slightly more than 0.2) . Therefore, we proceed ahead with the assumption that there is no collinearity in the data.
2. **Linearity of Logit**: Logit Linearity test[3] (results show in Appendix 4), demonstrates that all the features have p-values more than 0.05. In this case, none of the variables are significant, so we do not reject the assumption of linearity for these variables and continue with assumption of linearity.
3. **Independence Of Errors**: The Durbin-Watson test for independent errors was not significant at the 5% level of significance (d= 1.878, p=0.372). As d is close to 2, we do not reject the null hypothesis[4] and continue with the assumption of independence met.
4. **Incomplete information**: Hosmer-Lemeshow[5] test statistic is not statistically significant (i.e, $p > .05$), suggest there is no evidence of lack of fit. And when the final model is compared with model with more feature, the chi-test statistic (result shown in Appendix 8) is not statistically significant (chi = 1.966, df = 5, p = 0.854), suggest the final model chosen best describe the predictor variables perfectly predict the outcome variable.

[1] VIF (Variance Inflation Factor) to investigate multicollinearity
[2] Eliminate variables that do not contribute significantly to the model and identify the set of predictor variables that best predict the outcome variable
[3] Add in interaction effects with the log of each predictor variable, and see if that is significant
[4] Null hupothesis is that the errors are independent
[5] Hosmer-Lemeshow test is a goodness-of-fit test for logistic regression models and is used to assess whether the model adequately fits the data

5. **Complete Separation**: Visualizing the scatter plots between predictor variables (shown in Appendix 7), it is clear that there is no complete separation in the data and hence this assumption does not get violated.

Checking for Outliers and Influential Points There are 7 residuals are above or below 1.96 standard deviations. As this represents approximately 1.5 % of the observations, which are expected[6], hence we do not consider any of these observations as outliers and continued with all observations included in the model. The max Cook's distance for model is 0.471 which is less than '1'. Hence, we conclude that there are no influential cases in our model.

Analysis

Model Analysis and Interpretation, based on the calculated **co-efficient and their p-values (shown in below table 1), we can reject Null Hypothesis**[7] and conclude that our model is significant. The model is summarized in Appendix 3.
Using confidence intervals, it can be seen that the intercept is between -41.96 and -20.86, which does not overlap one. This means there is a **significant difference between the odds of type of breast cancer being Benign and Malignant in general**, at the 5% level of significance. Also, we can conclude that none of the intervals for features overlap '1', indicating that **all the features have some impact on the diagnosis of breast cancer**, at 5% level of significance.
To interpret the co-efficient, converted them to **odd ratio** (measure of effect size of the feature on the outcome) using exponential and results are shown in table 1. Based on the **Odds ratio**, we observe that `smoothness_worst` has a significant impact on the outcome and `fractal_dimension_se` has lowest impact.

Table 1: Odd-ratio, confidence interval, and estimates of coefficients of model with z-values and p-values

Variable	Odds Ratio (OR)	2.5 %	97.5 %	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	8.279e-14	-41.9632	-20.8591	-30.122	5.3033	-5.67995	<0.001
radius_worst	2.928	0.6947	1.54711	1.07436	0.2142	5.01550	<0.001
smoothness_mean	1.286e-57	-233.258	-42.3103	-130.996	48.1771	-2.71905	0.00655
concavity_mean	2.638e+09	1.4482	42.41813	21.693	10.3747	2.09098	0.03653
radius_se	1.055e+04	4.2742	14.94634	9.26392	2.6026	3.55943	0.00037
texture_se	7.554	0.626	3.56496	2.02201	0.7403	2.73142	0.00631
fractal_dimension_se	1.371e-206	-872.745	-158.816	-474.017	185.4036	-2.55668	0.01057
smoothness_worst	7.886e+46	54.6958	167.763	107.984	28.61091	3.77423	0.00016
concave.points_worst	9.144e+16	11.0346	70.1363	39.0544	14.95191	2.61200	0.00900

Conclusion

The logistic regression model built with eight variables was used to examine the association between predictor variables and the likelihood of breast cancer diagnosis in a sample of patients. Results indicated that **Intercept** (OR = 8.279e-14, p<0.001), **radius_worst** (OR = 2.928, p<0.001), **smoothness_mean** (OR = 1.286e-57, p<0.01), **concavity_mean** (OR = 2.638e+09, p<0.05), **radius_se** (OR = 1.055e+04, p<0.004), **texture_se** (OR = 7.554, p<0.01), **fractal_dimension_se** (OR = 1.371e-206, p<0.05), **smoothness_worst** (OR = 7.886e+46, p<0.01), and **concave.points_worst** (OR = 9.144e+16, p<0.01) were significant predictors of breast cancer diagnosis, after controlling for other variables in the model.
As no confidence interval has 1 lying in its range, direction of the Odd's Interval can be considered reliable. For instance, **If the value of radius_worst increases by 1, the odds of the outcome variable occurring (type of breast cancer being Malignant) increase by a factor of 2.928.**
The model fit the data well with **AIC of 108.81**. However, it is important to note that these results are based on a cross-sectional sample and **cannot be generalized to the population at large**. Future research is needed to confirm these findings and explore other potential predictors of breast cancer diagnosis.

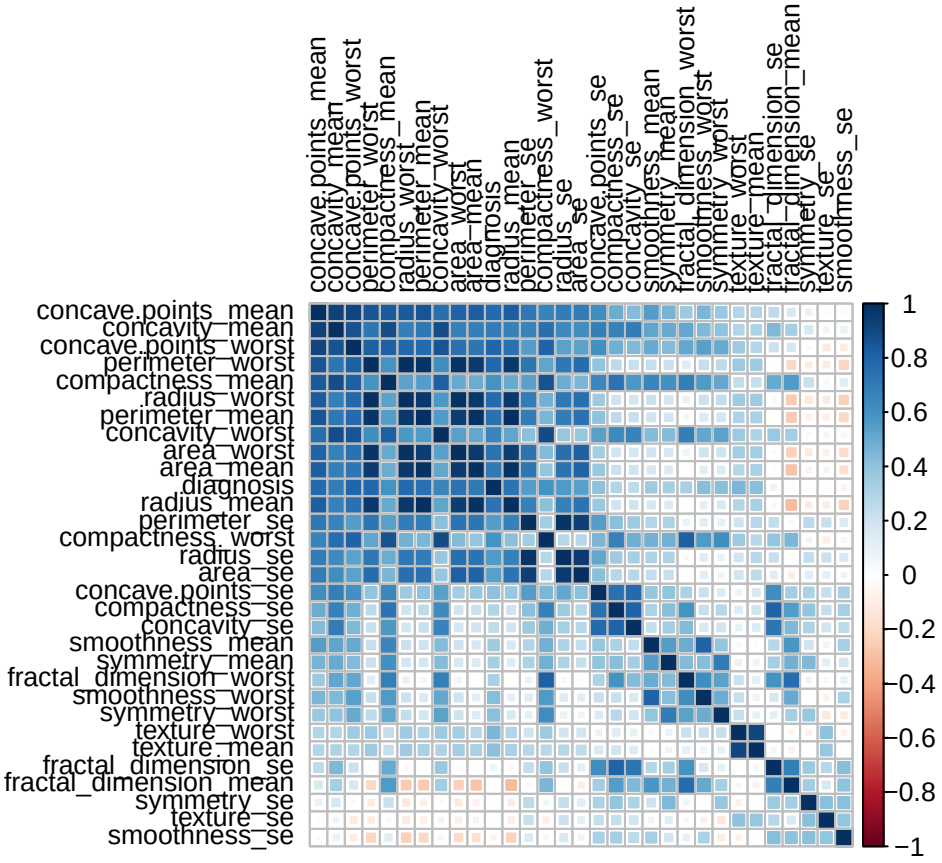
[6] 5% of data is expected to be outside of 2 standard deviations.
[7] **Null Hypothesis is Coefficients of predictor variables in the Logistic Regression model are zero.**

Appendix

Appendix 1: Data Description

```
## 'data.frame':    569 obs. of  31 variables:
##  $ diagnosis              : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ radius_mean            : num  18 20.6 19.7 11.4 20.3 ...
##  $ texture_mean           : num  10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter_mean         : num  122.8 132.9 130 77.6 135.1 ...
##  $ area_mean              : num  1001 1326 1203 386 1297 ...
##  $ smoothness_mean        : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
##  $ compactness_mean       : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
##  $ concavity_mean         : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
##  $ concave.points_mean    : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
##  $ symmetry_mean          : num  0.242 0.181 0.207 0.26 0.181 ...
##  $ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
##  $ radius_se              : num  1.095 0.543 0.746 0.496 0.757 ...
##  $ texture_se             : num  0.905 0.734 0.787 1.156 0.781 ...
##  $ perimeter_se           : num  8.59 3.4 4.58 3.44 5.44 ...
##  $ area_se                : num  153.4 74.1 94 27.2 94.4 ...
##  $ smoothness_se          : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
##  $ compactness_se         : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
##  $ concavity_se           : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
##  $ concave.points_se      : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
##  $ symmetry_se            : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
##  $ fractal_dimension_se   : num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
##  $ radius_worst           : num  25.4 25 23.6 14.9 22.5 ...
##  $ texture_worst          : num  17.3 23.4 25.5 26.5 16.7 ...
##  $ perimeter_worst        : num  184.6 158.8 152.5 98.9 152.2 ...
##  $ area_worst             : num  2019 1956 1709 568 1575 ...
##  $ smoothness_worst       : num  0.162 0.124 0.144 0.21 0.137 ...
##  $ compactness_worst      : num  0.666 0.187 0.424 0.866 0.205 ...
##  $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
##  $ concave.points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
##  $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
##  $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
```

Appendix 2: Correlation of Dataset



Appendix 3: Summary of Final Logistic Model

```
##
## Call:
## glm(formula = diagnosis ~ ., family = binomial(link = "logit"),
##     data = selected_dataset)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.95366  -0.06693  -0.00973   0.00134   2.91893
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -30.1224     5.3033  -5.680 1.35e-08 ***
## radius_worst             1.0744     0.2142   5.015 5.29e-07 ***
## smoothness_mean       -130.9959    48.1771  -2.719 0.006547 **
## concavity_mean          21.6933    10.3747   2.091 0.036530 *
## radius_se                9.2639     2.6026   3.559 0.000372 ***
## texture_se               2.0220     0.7403   2.731 0.006306 **
## fractal_dimension_se  -474.0170   185.4036  -2.557 0.010568 *
## smoothness_worst       107.9841    28.6109   3.774 0.000161 ***
## concave.points_worst    39.0544    14.9519   2.612 0.009001 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.440  on 568  degrees of freedom
## Residual deviance:  90.809  on 560  degrees of freedom
```

```
## AIC: 108.81
##
## Number of Fisher Scoring iterations: 9
```

```
##         radius_worst      smoothness_mean       concavity_mean
##             1.649681             4.676913             3.641580
##            radius_se          texture_se fractal_dimension_se
##             2.215532             1.957207             2.665829
##     smoothness_worst concave.points_worst
##             4.785317             3.350391
```

Appendix 4: Assumption Test results of Logistic Model

```
## [1] "VIF Values:"
```

```
##         radius_worst      smoothness_mean       concavity_mean
##             1.649681             4.676913             3.641580
##            radius_se          texture_se fractal_dimension_se
##             2.215532             1.957207             2.665829
##     smoothness_worst concave.points_worst
##             4.785317             3.350391
```

```
## [1] "Tolerance:"
```

```
##         radius_worst      smoothness_mean       concavity_mean
##            0.6061778            0.2138162            0.2746061
##            radius_se          texture_se fractal_dimension_se
##            0.4513588            0.5109321            0.3751179
##     smoothness_worst concave.points_worst
##            0.2089726            0.2984726
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.06088779      1.878224    0.41
##  Alternative hypothesis: rho != 0
```

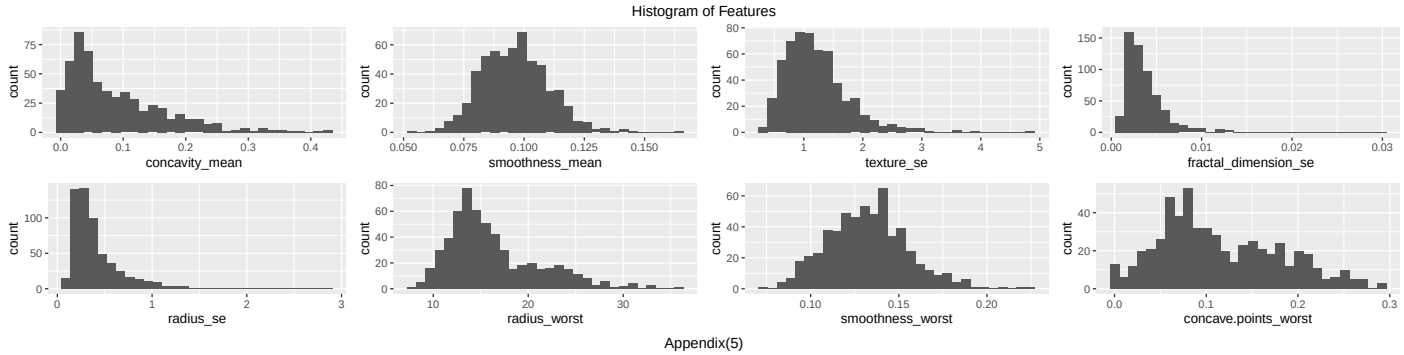
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  fitted(final_model), selected_dataset$diagnosis
## X-squared = 3.1415e-21, df = 98, p-value = 1
```

Linearity Test results

```
##
## Call:
## glm(formula = diagnosis ~ ., family = binomial(link = "logit"),
##     data = selected_dataset_test)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.92704  -0.07692  -0.01098   0.00015   2.81579
##
## Coefficients:
##                                                    Estimate
## (Intercept)                                         -21.9022
## radius_worst                                          2.5496
## smoothness_mean                                     268.8924
## concavity_mean                                       44.7601
## radius_se                                             6.6324
## texture_se                                            6.5790
## fractal_dimension_se                              -1213.7004
## smoothness_worst                                    -47.2688
```

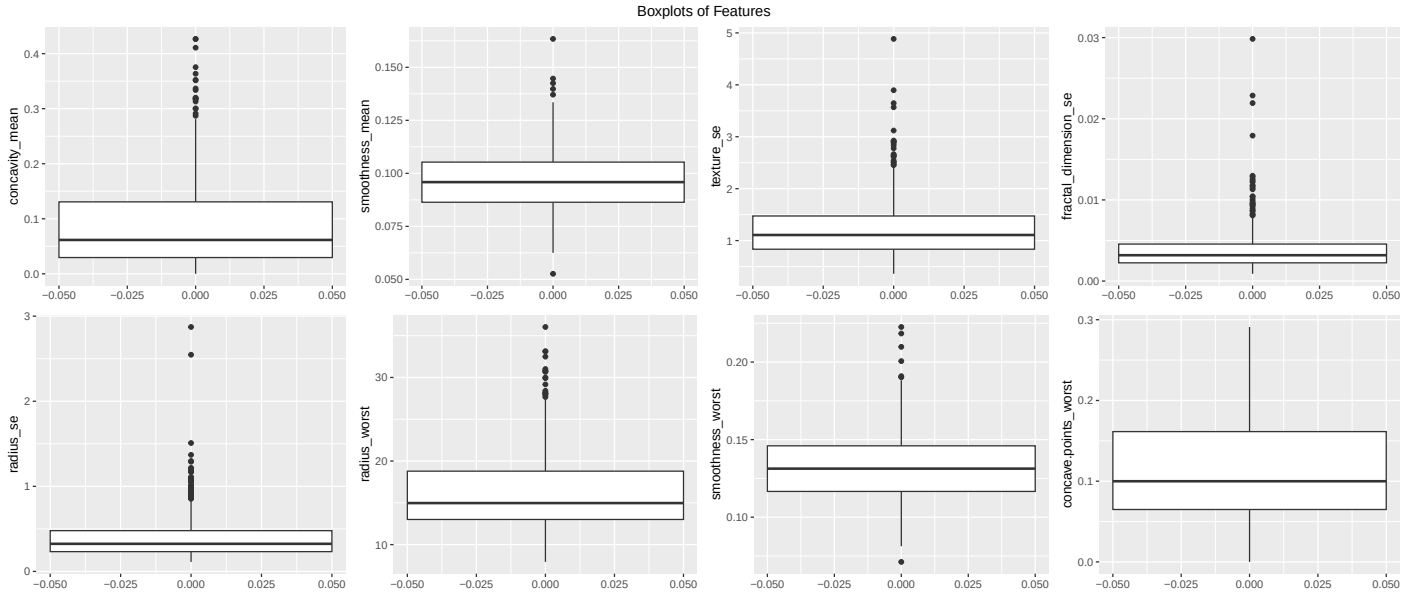
```
## concave.points_worst                                          91.4996
## `log.texture_se <- log(texture_se) * texture_se`               -3.1651
## `log.fractal_dimension_se <- ...`                            -169.5352
## `log.radius_worst <- log(radius_worst) * radius_worst`         -0.3068
## `log.concavity_mean <- log(concavity_mean) * concavity_mean`   18.8298
## `log.radius_se <- log(radius_se) * radius_se`                  30.7466
## `log.concave.points_worst <- ...`                              40.8926
## `log.smoothness_worst <- ...`                                -174.9365
## `log.smoothness_mean <- ...`                                  314.2979
##                                                            Std. Error z value
## (Intercept)                                                   47.3749  -0.462
## radius_worst                                                   9.4353   0.270
## smoothness_mean                                              584.2787   0.460
## concavity_mean                                                31.3767   1.427
## radius_se                                                      4.0408   1.641
## texture_se                                                     3.3770   1.948
## fractal_dimension_se                                        1293.5521  -0.938
## smoothness_worst                                             184.4896  -0.256
## concave.points_worst                                          65.4840   1.397
## `log.texture_se <- log(texture_se) * texture_se`               2.6243  -1.206
## `log.fractal_dimension_se <- ...`                            312.0013  -0.543
## `log.radius_worst <- log(radius_worst) * radius_worst`         2.5034  -0.123
## `log.concavity_mean <- log(concavity_mean) * concavity_mean`  24.1784   0.779
## `log.radius_se <- log(radius_se) * radius_se`                 18.6835   1.646
## `log.concave.points_worst <- ...`                             55.1806   0.741
## `log.smoothness_worst <- ...`                                187.9628  -0.931
## `log.smoothness_mean <- ...`                                 440.4415   0.714
##                                                            Pr(>|z|)
## (Intercept)                                                  0.6439
## radius_worst                                                 0.7870
## smoothness_mean                                              0.6454
## concavity_mean                                               0.1537
## radius_se                                                    0.1007
## texture_se                                                   0.0514 .
## fractal_dimension_se                                         0.3481
## smoothness_worst                                             0.7978
## concave.points_worst                                         0.1623
## `log.texture_se <- log(texture_se) * texture_se`             0.2278
## `log.fractal_dimension_se <- ...`                            0.5869
## `log.radius_worst <- log(radius_worst) * radius_worst`       0.9025
## `log.concavity_mean <- log(concavity_mean) * concavity_mean` 0.4361
## `log.radius_se <- log(radius_se) * radius_se`                0.0998 .
## `log.concave.points_worst <- ...`                            0.4587
## `log.smoothness_worst <- ...`                                0.3520
## `log.smoothness_mean <- ...`                                 0.4755
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 739.14  on 555  degrees of freedom
## Residual deviance:  82.03  on 539  degrees of freedom
##   (13 observations deleted due to missingness)
## AIC: 116.03
##
## Number of Fisher Scoring iterations: 11
```

Appendix 5: Histogram of predictor variables



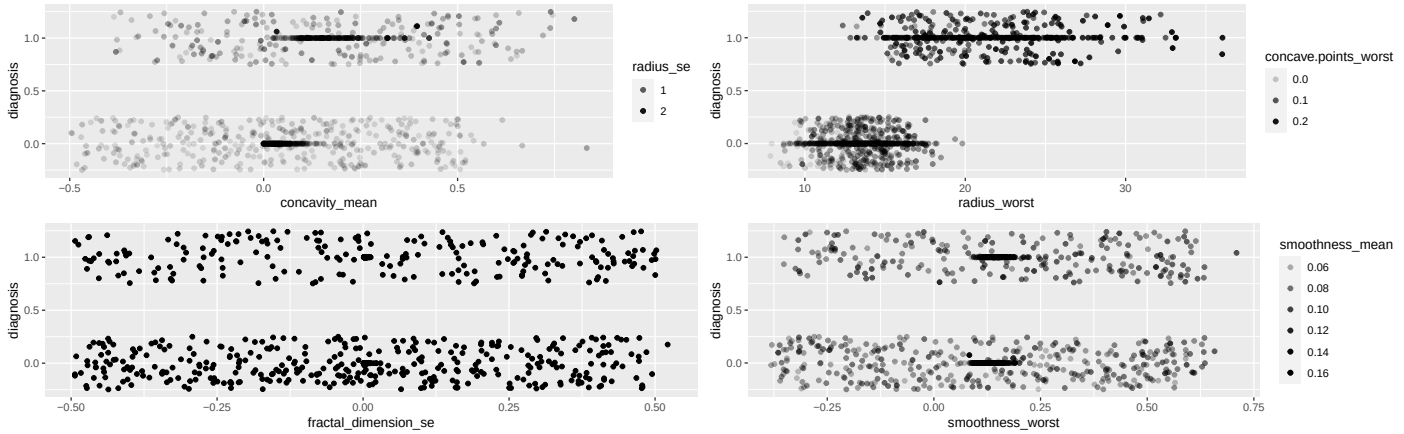
Appendix(5)

Appendix 6: Boxplot of predictor variables

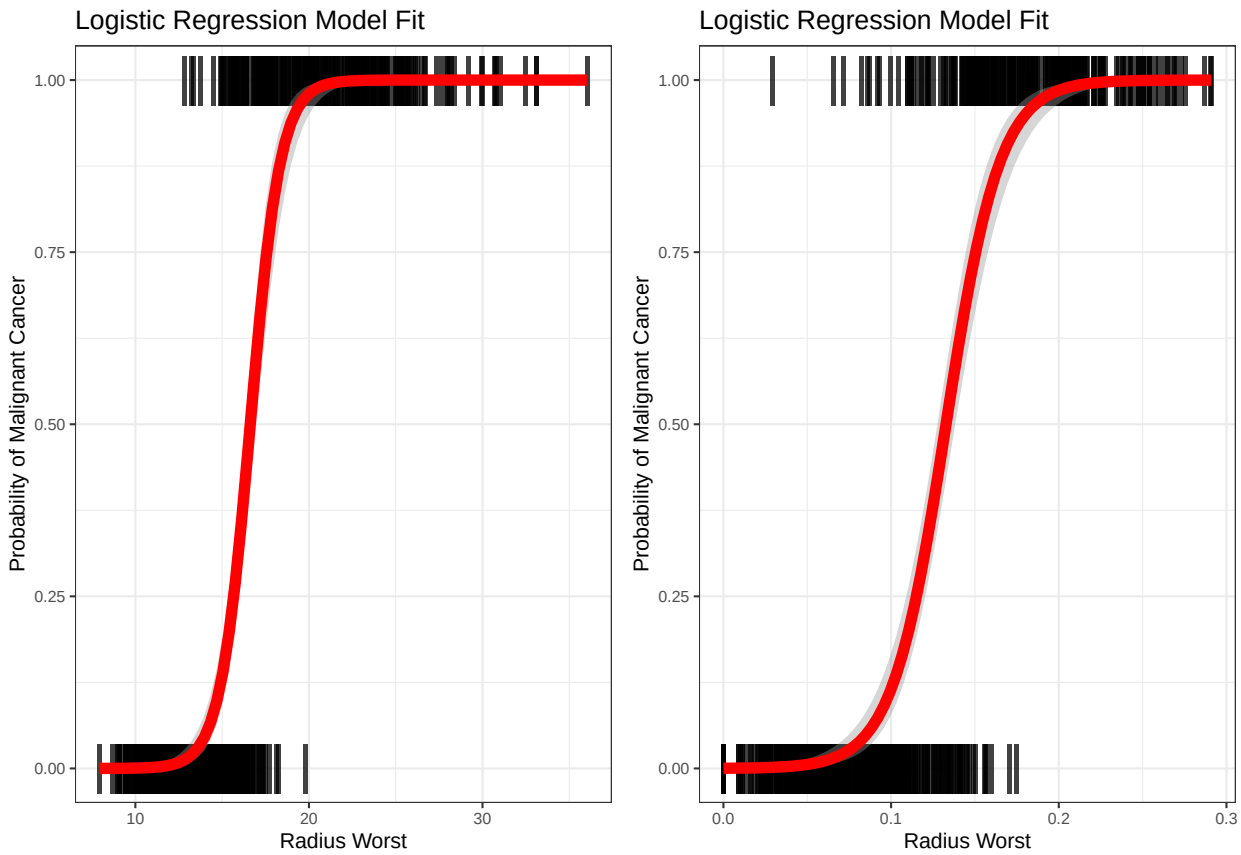


Appendix(6)

Appendix 7: Jitter Scatterplot of predictor variables




```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Appendix 8: Final Model Analysis

```
##        smoothness_mean         concavity_mean    concave.points_mean
##               8.266161               5.381811               6.204316
##          symmetry_mean fractal_dimension_mean              radius_se
##               3.150363               6.334867               3.020334
##             texture_se      concave.points_se            symmetry_se
##               2.014212               5.978406               2.600513
##   fractal_dimension_se          radius_worst      smoothness_worst
##               5.230714               2.999919               5.925580
##   concave.points_worst
##               6.207976
```

```
## Start:  AIC=116.84
## diagnosis ~ smoothness_mean + concavity_mean + concave.points_mean +
##     symmetry_mean + fractal_dimension_mean + radius_se + texture_se +
##     concave.points_se + symmetry_se + fractal_dimension_se +
##     radius_worst + smoothness_worst + concave.points_worst
##
##                          Df Deviance    AIC
## - symmetry_mean           1   88.883 114.88
## - concave.points_mean     1   89.318 115.32
## - symmetry_se             1   89.330 115.33
## - concave.points_se       1   89.570 115.57
## - fractal_dimension_mean  1   89.745 115.75
## - smoothness_mean         1   90.508 116.51
## <none>                        88.843 116.84
## - fractal_dimension_se    1   90.956 116.96
## - concavity_mean          1   93.077 119.08
## - concave.points_worst    1   96.352 122.35
## - texture_se              1   96.876 122.88
## - smoothness_worst        1   99.523 125.52
```

```
## - radius_se               1  100.862 126.86
## - radius_worst            1  105.415 131.41
##
## Step:  AIC=114.88
## diagnosis ~ smoothness_mean + concavity_mean + concave.points_mean +
##     fractal_dimension_mean + radius_se + texture_se + concave.points_se +
##     symmetry_se + fractal_dimension_se + radius_worst + smoothness_worst +
##     concave.points_worst
##
##                          Df Deviance    AIC
## - concave.points_mean     1   89.332 113.33
## - symmetry_se             1   89.455 113.45
## - concave.points_se       1   89.572 113.57
## - fractal_dimension_mean  1   89.770 113.77
## - smoothness_mean         1   90.865 114.86
## <none>                        88.883 114.88
## - fractal_dimension_se    1   90.979 114.98
## - concavity_mean          1   93.123 117.12
## - texture_se              1   96.959 120.96
## - concave.points_worst    1   97.501 121.50
## - smoothness_worst        1   99.939 123.94
## - radius_se               1  101.117 125.12
## - radius_worst            1  105.560 129.56
##
## Step:  AIC=113.33
## diagnosis ~ smoothness_mean + concavity_mean + fractal_dimension_mean +
##     radius_se + texture_se + concave.points_se + symmetry_se +
##     fractal_dimension_se + radius_worst + smoothness_worst +
##     concave.points_worst
##
##                          Df Deviance    AIC
## - symmetry_se             1   89.855 111.86
## - fractal_dimension_mean  1   90.165 112.17
## - concave.points_se       1   90.242 112.24
## - fractal_dimension_se    1   91.208 113.21
## <none>                        89.332 113.33
## - concavity_mean          1   93.562 115.56
## - smoothness_mean         1   94.034 116.03
## - texture_se              1   97.117 119.12
## - concave.points_worst    1   97.798 119.80
## - radius_se               1  101.554 123.55
## - smoothness_worst        1  103.292 125.29
## - radius_worst            1  106.548 128.55
##
## Step:  AIC=111.85
## diagnosis ~ smoothness_mean + concavity_mean + fractal_dimension_mean +
##     radius_se + texture_se + concave.points_se + fractal_dimension_se +
##     radius_worst + smoothness_worst + concave.points_worst
##
##                          Df Deviance    AIC
## - fractal_dimension_mean  1   90.561 110.56
## - concave.points_se       1   90.653 110.65
## - fractal_dimension_se    1   91.709 111.71
## <none>                        89.855 111.86
## - smoothness_mean         1   94.424 114.42
## - concavity_mean          1   94.881 114.88
## - texture_se              1   97.950 117.95
## - concave.points_worst    1   98.383 118.38
## - radius_se               1  102.714 122.71
```

```
## - smoothness_worst        1  103.324 123.32
## - radius_worst            1  106.747 126.75
##
## Step:  AIC=110.56
## diagnosis ~ smoothness_mean + concavity_mean + radius_se + texture_se +
##     concave.points_se + fractal_dimension_se + radius_worst +
##     smoothness_worst + concave.points_worst
##
##                        Df Deviance    AIC
## - concave.points_se     1   90.809 108.81
## <none>                      90.561 110.56
## - concavity_mean        1   95.001 113.00
## - fractal_dimension_se  1   96.697 114.70
## - smoothness_mean       1   97.678 115.68
## - concave.points_worst  1   98.501 116.50
## - texture_se            1   98.839 116.84
## - radius_se             1  102.715 120.72
## - smoothness_worst      1  104.506 122.51
## - radius_worst          1  120.040 138.04
##
## Step:  AIC=108.81
## diagnosis ~ smoothness_mean + concavity_mean + radius_se + texture_se +
##     fractal_dimension_se + radius_worst + smoothness_worst +
##     concave.points_worst
##
##                        Df Deviance    AIC
## <none>                      90.809 108.81
## - concavity_mean        1   95.222 111.22
## - concave.points_worst  1   98.503 114.50
## - texture_se            1   99.043 115.04
## - smoothness_mean       1   99.595 115.59
## - fractal_dimension_se  1  101.267 117.27
## - radius_se             1  103.048 119.05
## - smoothness_worst      1  108.180 124.18
## - radius_worst          1  132.588 148.59

##
## Call:  glm(formula = diagnosis ~ smoothness_mean + concavity_mean +
##     radius_se + texture_se + fractal_dimension_se + radius_worst +
##     smoothness_worst + concave.points_worst, family = binomial(link = "logit"),
##     data = cancer_dataset)
##
## Coefficients:
##          (Intercept)       smoothness_mean        concavity_mean
##              -30.122              -130.996                21.693
##            radius_se            texture_se  fractal_dimension_se
##                9.264                2.022              -474.017
##         radius_worst      smoothness_worst  concave.points_worst
##                1.074              107.984                39.054
##
## Degrees of Freedom: 568 Total (i.e. Null);  560 Residual
## Null Deviance:       751.4
## Residual Deviance: 90.81     AIC: 108.8

## [1] 1.965821

## [1] 5

## [1] 0.8538539
```