Introduction to Machine Learning, STAT5310
Dr. Daniel Mayo
Class Project: due not later than Dec. 12th, 2019
Allen Baumgarten (very upset Miami Dolphins fan)

Binomial Logistic Regression Applied to Breast Cancer Data

Breast cancer afflicts tens of thousands of women each year. The National Cancer Institute estimates that roughly 268,600 new cases will be diagnosed in 2019 with 41,760 estimated deaths occurring in the same year.[1] These estimated new breast cancer cases will, moreover, account for roughly 15% of all newly diagnosed cancer cases and which is higher than for any other type of cancer besides lung.[2,3] The prevalence of breast cancer is more startling when one realizes that as of 2016 the number of women living with breast cancer was estimated to be just under 3.8 million.

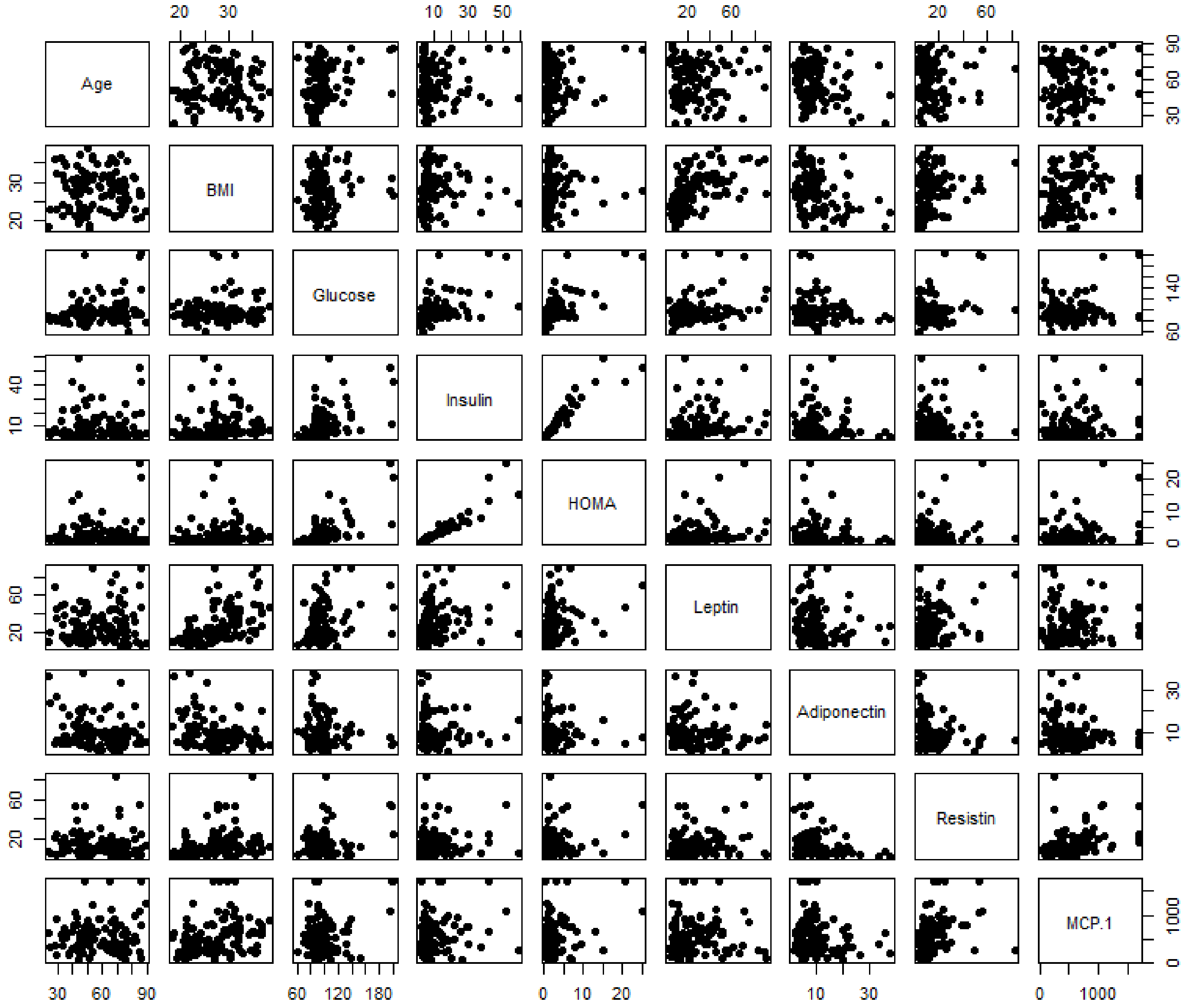
A number of risk factors associated with breast cancer in women include age, personal history of breast cancer, family history of breast cancer, radiation exposure, and other factors.[4]

The UCI Machine Learning Repository of publicly available datasets contains one particular set of data focused on breast cancer incidence.[5] This dataset (ds) contains blood panel findings on 64 patients classified as having breast (invasive vs. non-invasive not specified) and 52 patients without breast cancer, for a total of 116 total patients. Nine quantitative variables are included for analysis, seven of which are findings based on routine blood work and two based on patient clinical information. Because this dataset is not an attempt to randomize samples from a population as such, we cannot draw inferences on the extent to which cancer will be diagnosed, only whether these markers will be associated with people who have cancer. These variables are:

1. Age (in years)
2. BMI (kg/m2)
3. Glucose (mg/dL)
4. Insulin (µU/mL)
5. HOMA (Homeostatic Model Assessment of Insulin Resistance)
6. Leptin (ng/mL)
7. Adiponectin (µg/mL)
8. Resistin (ng/mL)
9. MCP-1 (pg/dL)

[1] "Cancer Stat Facts: Female Breast Cancer," National Cancer Institute Surveillance, Epidemiology, and End Results Program. Accessed 11/26/2019 at: https://seer.cancer.gov/statfacts/html/breast.html.
[2] Ibid.
[3] "U.S. Breast Cancer Statistics," Breastcancer.org. Accessed 11/26/2019 at: https://www.breastcancer.org/symptoms/understand_bc/statistics
[4] "Breast Cancer," May Clinic. Accessed 11/26/2019 at: https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470
[5] "Breast Cancer Coimbra Data Set," UCI Machine Learning Repository. Accessed 11/26/2019 at: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra#

Each of these nine variables is suspected of being associated with breast cancer incidence according to the authors. Age, of course, is the universal confounder in epidemiology and is thought to be associated with most of diseases, including cancer.[6] The next variable, BMI (body-mass index) is believed by some to be associated with cancer but this conclusion is controversial according to some researchers.[7] One variable of interest is Glucose level in the blood. It is well known that cancer cells thrive on sugar and some imaging tests (PET and SPECT scans) will utilize injections of radioactive sugar to locate a tumor.[8] This variable will be of particular interest in our study. High insulin levels are also thought to be associated with cancer.[9] We begin by examining a scatter plot matrix of all predictor variables which, in this ds, are continuous and without factors.



[6] "Breast Cancer," May Clinic. Accessed 11/26/2019 at: https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

[7] "Association between body mass index and breast cancer risk: evidence based on dose-response meta-analysis," Cancer Management Research. Accessed 11/26/2019 at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5783020/

[8] "Molecular and Nuclear Imaging (PET and SPECT)," National Cancer Institute. Accessed 11/26/2019 at: https://imaging.cancer.gov/imaging_basics/cancer_imaging/nuclear_imaging.htm

[9] "High Insulin Levels May Up Breast Cancer Risk," Accessed 11/26/2019 at: https://www.webmd.com/breast-cancer/news/20150115/unhealthy-insulin-levels-may-boost-breast-cancer-risk#1

Predictor Selection

A cursory examination of our predictor variables seems to reveal what could be some positive or negative relationships between some of these variables. Leptin, for example, seems to increase as BMI increases while Adiponectin seems to decrease with increases in BMI. Not surprisingly, Insulin and HOMA are very closely associated with each other, respectively. One of these two variables should probably be dropped to avoid correlated predictor variables. We start with an all-in binomial response model (logit link) with nine all variables present. BMI, Glucose (no surprises here), and Resistin all appear to be statistically significant compared to our Classification response variable.

```
Call:
glm(formula = Classification ~ ., family = binomial, data = breastcancer)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.2992  -0.8548   0.1847   0.7429   2.1632  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)   
(Intercept) -5.6512154  3.3580998  -1.683  0.09240 . 
Age         -0.0233524  0.0156230  -1.495  0.13498   
BMI         -0.1501231  0.0674938  -2.224  0.02613 * 
Glucose      0.1055941  0.0348082   3.034  0.00242 **
Insulin      0.2071782  0.2629802   0.788  0.43081   
HOMA        -0.5978147  1.0898156  -0.549  0.58332   
Leptin      -0.0101709  0.0172662  -0.589  0.55582   
Adiponectin -0.0052619  0.0375568  -0.140  0.88858   
Resistin     0.0585546  0.0298523   1.961  0.04982 * 
MCP.1        0.0006975  0.0008068   0.865  0.38730   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 159.57  on 115  degrees of freedom
Residual deviance: 111.73  on 106  degrees of freedom
AIC: 131.73

Number of Fisher Scoring iterations: 7
```

We have a couple of approaches in selecting a “best” model to use. One approach is to iterate our way through different combinations of predictors and compare which subsets of predictors produce the lowest AIC (Akaike Information Criteria). Another approach would be to simply build a model using the three statistically significant variables we saw earlier. We will attempt both of these approaches in that order.

We will use the AIC as our judging statistic though Joseph Hilbe suggests care in using it. He says, “The AIC is perhaps the most well-known and well used information statistic in current research. What may seem surprising to many readers is that there are a plethora of journal articles detailing studies proving how poor the AIC test is in assessing which of two models is the better fitted. Even Akaike himself later developed another criterion which he preferred to the original. However, it is his original 1973 version that is used by most researchers and that is found in most journals to assess comparative model fit. The traditional AIC statistic is found in two versions:

$AIC = -2L + 2k$ or $-2(L – k)$

or

$AIC = (-2L + 2k)/n$ or $2(L – k)/n$

where L is the log-likelihood model, k is the number of parameter estimates in the model, and n is the number of observations in the model."[10]

Our step-wise approach starts with an all-in model. Notice that the starting AIC of 131.73 below is identical to the all-in model we built earlier, containing all nine variables. Stepping through our variable iterations, we see at bottom that the final minimizing iteration selects BMI, Glucose, Insulin, and Resistin, which is strikingly similar to our original hypothesis of including just BMI, Glucose, and Resistin.

```
Start:  AIC=131.73
Classification ~ Age + BMI + Glucose + Insulin + HOMA + Leptin + Adiponectin + 
Resistin + MCP.1

              Df Deviance    AIC
- Adiponectin  1   111.75 129.75
- HOMA         1   111.98 129.98
- Leptin       1   112.08 130.08
- Insulin      1   112.24 130.24
- MCP.1        1   112.49 130.49
<none>             111.73 131.73
- Age          1   114.01 132.01
- BMI          1   117.11 135.11
- Resistin     1   117.32 135.32
- Glucose      1   120.53 138.53

Step:  AIC=129.75
Classification ~ Age + BMI + Glucose + Insulin + HOMA + Leptin + Resistin + MCP.1

           Df Deviance    AIC
- HOMA      1   112.01 128.01
- Leptin    1   112.13 128.13
- Insulin   1   112.27 128.27
- MCP.1     1   112.52 128.51
<none>          111.75 129.75
- Age       1   114.08 130.07
- BMI       1   117.37 133.37
- Resistin  1   117.80 133.80
- Glucose   1   120.69 136.69

Step:  AIC=128.01
Classification ~ Age + BMI + Glucose + Insulin + Leptin + Resistin + MCP.1

           Df Deviance    AIC
- Leptin    1   112.48 126.48
- MCP.1     1   112.75 126.75
<none>          112.01 128.01
- Age       1   114.33 128.34
- Insulin   1   115.45 129.45
- BMI       1   117.37 131.37
- Resistin  1   118.17 132.17
- Glucose   1   135.51 149.51
```

[10] Hilbe, Joseph M., *Practical Guide to Logistic Regression*, (CRC Press: 2016), 53.

```
Step:  AIC=126.48
Classification ~ Age + BMI + Glucose + Insulin + Resistin + MCP.1

           Df Deviance    AIC
- MCP.1     1   113.57 125.57
<none>          112.48 126.48
- Age       1   114.74 126.74
- Insulin   1   115.68 127.68
- Resistin  1   118.19 130.19
- BMI       1   124.43 136.43
- Glucose   1   135.55 147.55

Step:  AIC=125.57
Classification ~ Age + BMI + Glucose + Insulin + Resistin

           Df Deviance    AIC
- Age       1   115.54 125.54
<none>          113.57 125.57
- Insulin   1   116.97 126.97
- Resistin  1   120.80 130.80
- BMI       1   124.48 134.48
- Glucose   1   135.55 145.55

Step:  AIC=125.54
Classification ~ BMI + Glucose + Insulin + Resistin

           Df Deviance    AIC
<none>          115.54 125.54
- Insulin   1   119.50 127.50
- Resistin  1   123.37 131.37
- BMI       1   126.30 134.30
- Glucose   1   135.68 143.68
Warning messages:
1: glm.fit: fitted probabilities numerically 0 or 1 occurred
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Let’s test our original thesis and run a model with just BMI, Glucose, and Resistin. We see below that it produces an AIC = 127.5 which is better than our all-in model but not quite as good as the best step-wise model above which also includes Insulin.

```
Call:
glm(formula = Classification ~ BMI + Glucose + Resistin, family = binomial,
    data = breastcancer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0464  -0.8604   0.2193   0.8738   1.9186

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.28518    2.03459  -2.598  0.00939 **
BMI         -0.13144    0.04675  -2.812  0.00493 **
Glucose      0.08676    0.02061   4.210 2.56e-05 ***
Resistin     0.07023    0.03051   2.302  0.02135 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 159.57  on 115  degrees of freedom
Residual deviance: 119.50  on 112  degrees of freedom
AIC: 127.5

Number of Fisher Scoring iterations: 6
```

Selected Model

We will thus conclude that our predictor selection process lands us on using four predictors, viz., BMI, Glucose, Insulin, and Resistin, which is:

```
Call:
glm(formula = Classification ~ BMI + Glucose + Insulin + Resistin,
    family = binomial, data = breastcancer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8974  -0.8352   0.1808   0.7714   1.8703

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.58114    2.06646  -2.217 0.026630 *
BMI         -0.15207    0.04941  -3.078 0.002085 **
Glucose      0.07921    0.02108   3.758 0.000171 ***
Insulin      0.06790    0.03889   1.746 0.080837 .
Resistin     0.06963    0.03030   2.298 0.021548 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 159.57  on 115  degrees of freedom
Residual deviance: 115.54  on 111  degrees of freedom
AIC: 125.54

Number of Fisher Scoring iterations: 6
```

Testing Our Model for Accuracy

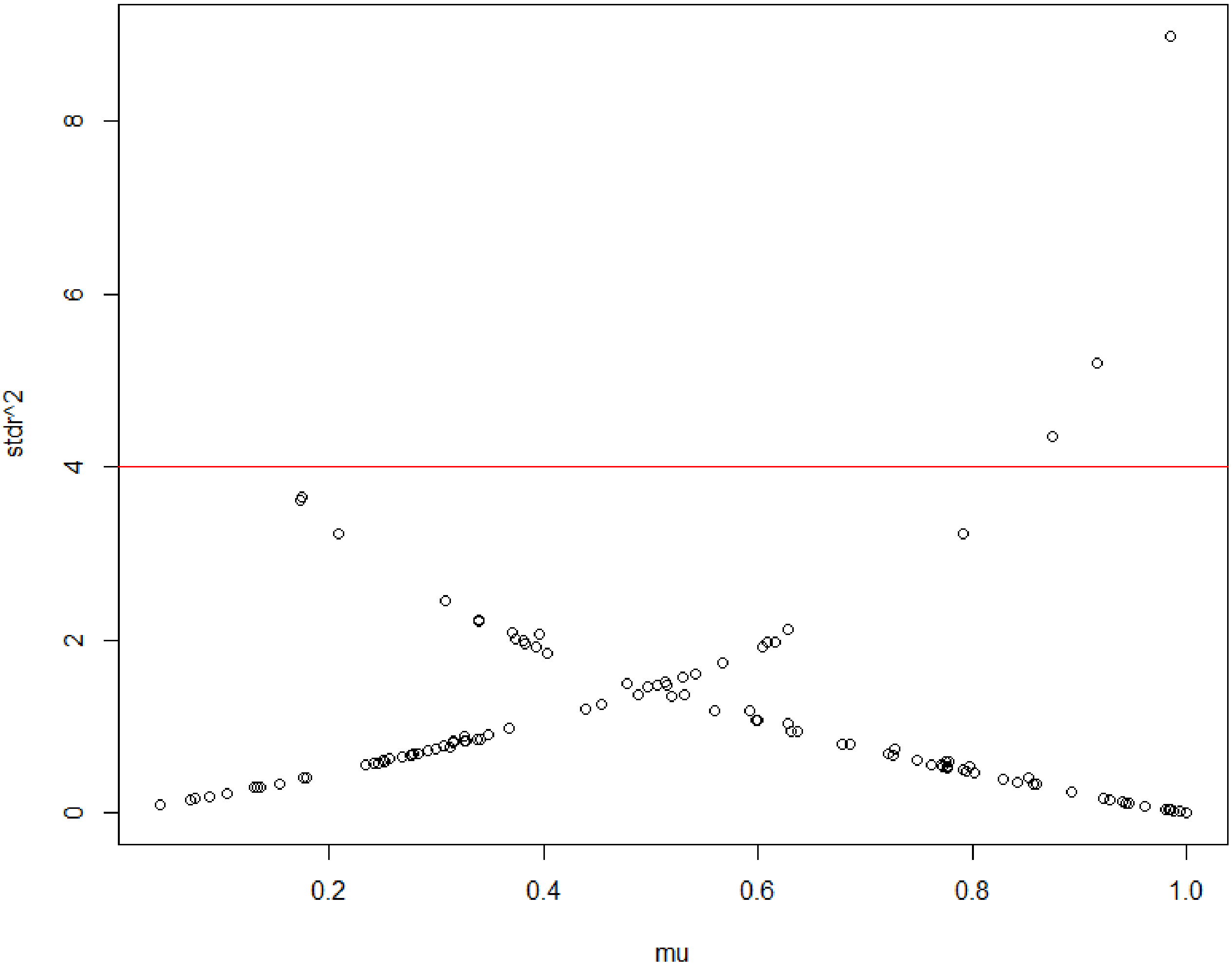
A number of statistical tests are commonly used to test whether a GLM fits well. One measure we can use is the Deviance statistic. Julian Faraway writes that this test statistic essentially tests how well our selected smaller model fits compared to a fuller “saturated” model in which all test data points are “perfectly” predicted while Hilbe states that this is known as the null deviance, or an intercept-only model.[11,12] Our model generates a D = 0.3649361 which is considerably higher than p = .05, indicating our model fits well. Our predictor variables also have p-values with significant fits with the exception Insulin. Insulin should be retained, however, due to its overall contribution to the model fit.

We may also examine standardized residuals. In so doing, we do see that three of them depart significantly at more than 4 standard deviations from the mean response. Given time, we could perhaps remove those three observations, rerun, and test for a tighter fit. A Pearson Goodness of Fit (GOF) statistic

[11] Faraway, Julian J., *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, (Boca Raton: CRC Press, 2006), 32.

[12] Hilbe, 54.

does not perform well with our model as is, indicating perhaps that these outlier points skew it.[13] Another telling problem for this current model could be that the Bayesian Information Criterion is not very close to the AIC.[14] Our AIC = 125.54 while the BIC = 139.31 and if Hilbe's comments hold, we may have some additional hurdles to consider before deploying. For this exercise, however, we will assume our model fits adequately.



Applying Our Model

Our model would suggest that, if this data sample is representative on average of other breast cancer patients sharing similar physiological characteristics, these four variables taken together may point to the presence of breast cancer. Let's take our model coefficients and "predict" on average what such outcomes may look like. Our model says that p(cancer) = $\beta_0 + X\beta_1 + X\beta_2 + X\beta_3 + X\beta_4$ which in this case is:

P(cancer) = -4.581 + BMI(-0.152) + Glucose(0.079) + Insulin(0.068) + Resistin(0.070)

[13] Note: I later dropped observations 8, 38, and 51 just for kicks and the resulting reduced dataset performed better with better-fitting Deviance, Pearson, and Hosmer-Lemeshow statistics.

[14] Hilbe advises that, "The Schwartz Bayesian information criterion (BIC) is the most used BIC test found in the literature…My recommendation is to test models with both [the AIC and BIC]. If the values substantially differ, it is likely the model is mis-specified." 60.

So, for example, our fifth observation with a BMI of 21.37, Glucose of 77, Insulin of 3.2, and Resistin of 12.8, and exponentiating, we can easily accomplish this with some simple code:

```
> nd <- data.frame(BMI = 21.37, Glucose = 77, Insulin = 3.2, Resistin = 12.8)
> predict(model_4, nd, type="response")
```

The probability of a patient having breast cancer based on these four values is on average = 0.349

APPENDIX: R SCRIPTS USED

```
> library(readr)
> library(readxl)
> breastcancer <- read_excel("c:/Users/baumgaral/Data/STAT5310/breast_cancer_data.xlsx")
> head(breastcancer)
# A tibble: 6 x 10
    Age   BMI Glucose Insulin  HOMA Leptin Adiponectin Resistin MCP.1 Classification
  <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl>       <dbl>    <dbl> <dbl>          <dbl>
1    48  23.5      70    2.71 0.467   8.81        9.70     8.00  417.              0
2    83  20.7      92    3.12 0.707   8.84        5.43     4.06  469.              0
3    82  23.1      91    4.50 1.01   17.9        22.4      9.28  555.              0
4    68  21.4      77    3.23 0.613   9.88        7.17    12.8   928.              0
5    86  21.1      92    3.55 0.805   6.70        4.82    10.6   774.              0
6    49  22.9      92    3.23 0.732   6.83       13.7     10.3   530.              0

> pairs(breastcancer[,1:9], pch = 19, lower.panel = NULL)

> all_in_model <- glm(Classification ~ ., data = breastcancer, family = binomial)
> summary(all_in_model)

> library(faraway)
> reduced <- step(all_in_model)

> summary(model_3 <- glm(Classification ~ BMI + Glucose + Resistin, data = breastcancer, family = binomial))
> summary(model_4 <- glm(Classification ~ BMI + Glucose + Insulin + Resistin, data = breastcancer, family =
binomial))
> pchisq(deviance(model_4), df.residual(model_4), lower=FALSE)
[1] 0.3649361

# Print standardized residuals (full model)
> mu <- current_model$fitted.value
> dr <- resid(current_model, type = "deviance")
> hat <- hatvalues(current_model)
> stdr <- dr/sqrt(1-hat)
> plot(mu, stdr^2)
> abline(h = 4, col = "red")

> current_model <- model_4
> current_data <- breastcancer

# Print summary of results and coefficients
> print(summary(current_model))
> writeLines("Coefficients Exponentiated:")
> print(exp(coef(current_model)))
> writeLines("")
> writeLines("")
> writeLines("95% CI's (Wald):")
> print(confint.default(current_model))
> writeLines("")
> writeLines("")
> writeLines("95% CI's (Wald) Exponentiated:")
> print(exp(confint.default(current_model)))
> writeLines("")
> writeLines("")
> writeLines("95% CI's (Profile):")
> print(confint(current_model))
> writeLines("")
> writeLines("")
> writeLines("95% CI's (Profile) Exponentiated:")
> print(exp(confint(current_model)))
> writeLines("")
> writeLines("")
```

```
Coefficients Exponentiated:
> print(exp(coef(current_model)))
(Intercept)         BMI     Glucose     Insulin    Resistin 
 0.01024322  0.85892979  1.08243363  1.07026338  1.07210652 
```

```
> library(COUNT)
> writeLines("AIC and BIC Statistics:")
> print(modelfit(current_model))
> nd <- data.frame(BMI = 21.37, Glucose = 77, Insulin = 3.2, Resistin = 12.8)
> predict(model_4, nd, type="response")
```