MATH 535 EXAM II

Gustavo Esparza

4/13/2019

Set-Up

Before beginning the problem, we will import our data and only keep the variables of interest. That is, we will keep (Gender,BMI,Age,Race,Tobacco,Diabetes,Artery Disease,Cancer,Albumin,Operation Length) as predictors and Anastomotic Leak as our response. In addition, we can note that the categorical variable of "Race" is not defined consistently throughout the data. Thus, we will alter the dataset to have a consistent binary output.

Question 1

We want to articulate the risks of Anastomotic Leaking associated with BMI in the context of all other potential predictors. For this reason, we will build a Logistic Regression Model utilizing all predictors for the time being.

Now, we can create an initial model utilizing all of our predictors and respective response. Thus, we will have the following model:

$$y = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$$

Where y = Anastamotic Leak, x_1 = Gender (Male = 1, Female = 0), x_2 = BMI, x_3 = Age, x_4 = Race (AA =1, W = 0), x_5 = Tobacco Usage (1= Yes, 0 = No), x_6 = Diabetes (1= Yes, 0 = No), x_7 = Artery Disease (1=Yes, 0 = No), x_8 = Cancer (1= Yes, 0 = No), x_9 = Albumin, and x_{10} = Operation Length.

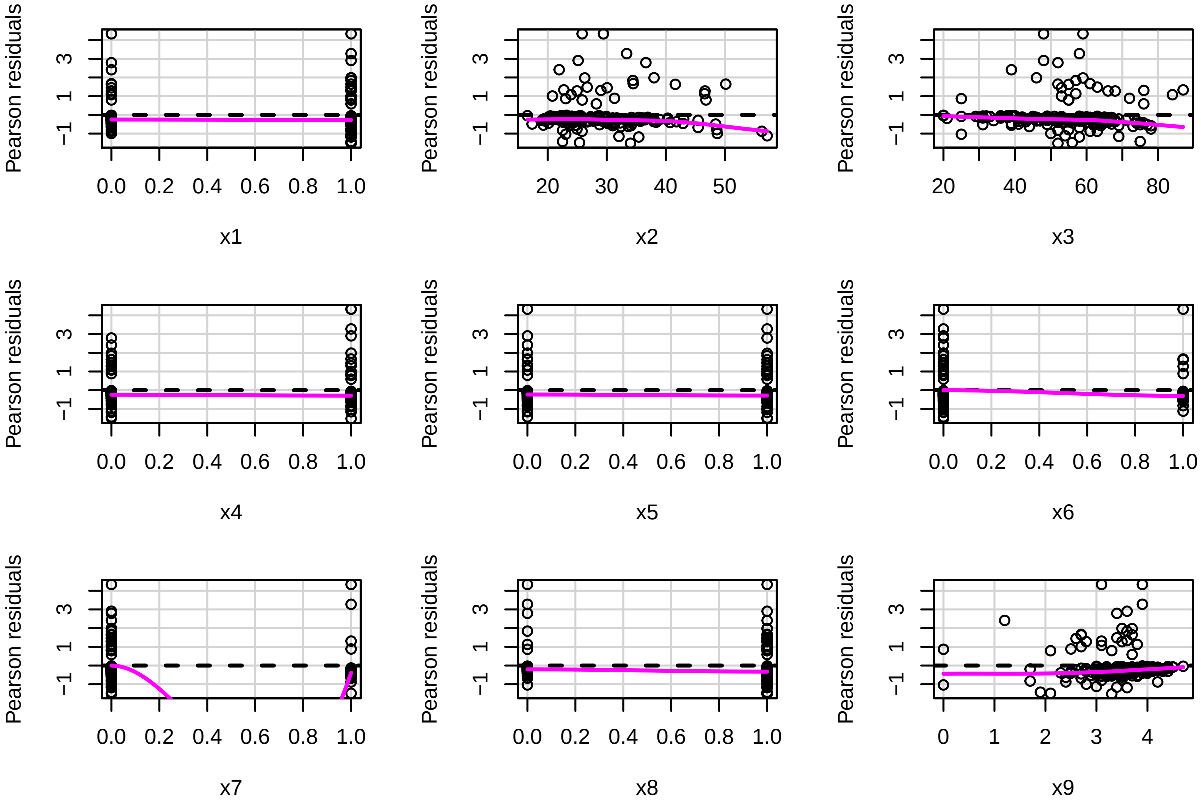
Note: The CAR library was not recognizing our categorical variables due to their labeling. Thus, we will alter them to a simple binary response of (0,1). For the GENDER variable, we will assign Male = 1 and Female = 0. For the RACE variable, we will assign AA = 1 and W = 0.

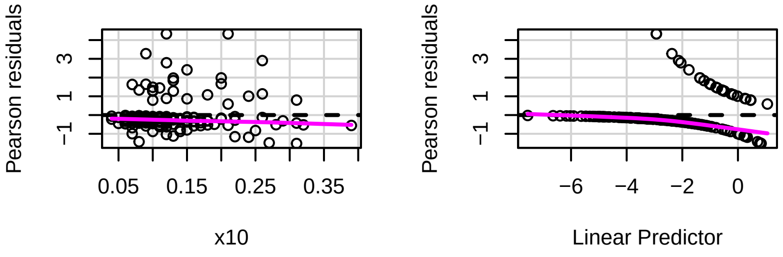
Now, let's take a look at a summary of our model. We want to view the Summary as well as the residual Plots because we want to first understand if we have a valid model for analyzing the risk of Anastomotic Leaking associated with BMI in the context of all other predictors.

```
##
## Call:
## glm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
##     x10, family = "binomial", data = C_data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5455  -0.4899  -0.3060  -0.1282   2.4431
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.15583    2.17879  -3.284 0.001022 **
## x1           0.82283    0.51895   1.586 0.112839
## x2           0.09486    0.03227   2.940 0.003282 **
## x3           0.08437    0.02646   3.188 0.001431 **
## x4           0.23469    0.51959   0.452 0.651494
## x5           1.02614    0.57343   1.789 0.073541 .
## x6          -0.85182    0.64633  -1.318 0.187529
## x7          -0.68212    0.73646  -0.926 0.354334
## x8           0.59864    0.59499   1.006 0.314348
## x9          -1.36235    0.36698  -3.712 0.000205 ***
## x10          7.05757    3.60032   1.960 0.049965 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 148.35  on 178  degrees of freedom
## Residual deviance: 111.56  on 168  degrees of freedom
## AIC: 133.56
##
## Number of Fisher Scoring iterations: 6
```

From our summary, we can initially see that some of our predictors (x4,x6,x7,x8 and to a lesser extent x1,x5) do not appear to be statistically significant, according to their p-value. For the time being, we will leave all variables in the model and refine our variable selection at a later time.

Now, we will take a look at the Residual Plots of our data.





```
##     Test stat Pr(>|Test stat|)
## x1     0.0000          1.00000
## x2     1.9020          0.16786
## x3     0.0056          0.94010
## x4     0.0000          1.00000
## x5     0.0000          1.00000
## x6     0.0000          1.00000
## x7     0.0000          1.00000
## x8     0.0000          1.00000
## x9     0.0181          0.89288
## x10    3.8207          0.05062 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the residual plots, we can see that our response variable has the desired shape for a Logistic Regression Model. There is a slight deviation from a horizontal line for the residuals, but we will be forgiving in this specific problem. In addition, most of our predictors have a horizontal residual line running through the x-axis, which is what we expect to see for Logistic regression. We can note that BMI and Age deviate from the x-axis , implying that linearity is not consistently held as the these respective values increase. We could attempt to fix these issues via transformations, but for now we will resume with our full model.

ANALYZING BMI NUMERICALLY

Now, we will analyze the effect that BMI has on our response, Risk of Anastomotic Leaking, in the context of all other variables. We will again be using the summary of our full model in order to perform any interpretation of the change in odds.

For Logistic Regression, a one unit increase in x_j is an $e^{\hat{\beta}_j}$ change in the odds of$y$, in the context of all other predictors.

In addition, we can also say that when holding all else constant, a one unit increase in x_j is on average an $e^{\hat{\beta}_j \pm 1.96 * se\left(\hat{\beta}^j\right)}$ change in y.

Utilizing these two definitions, we can provide inference for the risk of Anastomotic Leaking following a Colectomy associated with BMI.

Thus, here is the one unit change in odds and odds confidence interval for the one unit increase in BMI in the context of all other predictors:

```
##
## Change in Odds From a One Unit Increase in BMI =  1.099507
```

```
##
## Confidence Interval = [ 1.032126  ,  1.171287 ]
```

We can observe that a one unit increase in BMI changes the odds of experiencing Anastomotic Leaking by a factor of 1.099507. We can also see that the average odds confidence interval contains values ranging from 1.03 and 1.17. Thus, we can say that we are confident that an increase in BMI does in fact increase the risks of Anastomotic Leaking following a Colectomy.

Now that we have articulated the risks of Anastomotic Leaking in a numerical sense of BMI, we can also explain the risks in terms of categories of BMI class.

ANALYZING BMI CATEGORICALLY

In order to analyze the effects of BMI in a categorical sense, we must first create a new categorical variable for BMI and assign levels to the respective BMI values.

Next, we can develop a new model that replaces BMI numerically with the new Categorical BMI. Here is the summary for that specific model.

```
##
## Call:
## glm(formula = y ~ x1 + x2_catBMI + x3 + x4 + x5 + x6 + x7 + x8 +
##     x9 + x10, family = "binomial", data = C_data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5161  -0.5226  -0.3186  -0.1465   2.4529
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -5.11645    1.81890  -2.813 0.004909 **
## x1                             0.82533    0.53168   1.552 0.120589
## x2_catBMIobese                 2.06802    0.79342   2.606 0.009148 **
## x2_catBMIoverweight            0.33552    0.69193   0.485 0.627744
## x2_catBMIseverely overweight  1.09743    0.98657   1.112 0.265981
## x3                             0.08327    0.02658   3.133 0.001733 **
## x4                             0.34738    0.53320   0.651 0.514728
## x5                             0.96946    0.55800   1.737 0.082319 .
## x6                            -0.69770    0.65746  -1.061 0.288598
## x7                            -0.50950    0.72603  -0.702 0.482826
## x8                             0.64320    0.58948   1.091 0.275219
## x9                            -1.26409    0.36658  -3.448 0.000564 ***
## x10                            7.15818    3.53367   2.026 0.042794 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 148.35  on 178  degrees of freedom
## Residual deviance: 113.49  on 166  degrees of freedom
## AIC: 139.49
##
## Number of Fisher Scoring iterations: 6
```

From our summary, the new BMI predictor is given in three respective levels of BMI class. The estimate provided can be used to determine the odds ratio of experiencing Anastomotic Leaking for the three unhealthy BMI classes VS a healthy BMI, in the context of all other present predictors.

Here is the Odds Ratio for Overweight VS Healthy BMI, as well as its' corresponding confidence interval:

```
##
## Odds Ratio for Overweight VS Healthy BMI =  1.39867
```

```
##
## Confidence Interval = [ 0.3603549  ,  5.428749 ]
```

From our Odds Ratio, we can say that a patient with an "Overweight" BMI is about 1.4 times more likely to experience Anastomotic Leaking when compared to an equivalent patient with a "Healthy" BMI.

We can also note that the confidence interval for the Odds Ratio ranges from .36 to 5.4. This is a pretty wide interval that contains values that are both less than and greater than 1. This is likely due to the fact that we are no longer considering a one unit increase in BMI but rather a one category increase. We can see that a majority of our interval is greater than one, but we are not completely confident that an Overweight BMI will always increase the risk of Anastomotic leaking, but we can infer that it will for a majority of cases.

Here is the Odds Ratio for Severely Overweight vs Healthy BMI

```
##
## Odds Ratio for Severely Overweight VS Healthy BMI =  2.996454
```

```
##
## Confidence Interval = [ 0.4333314  ,  20.72025 ]
```

From our Odds Ratio, we can say that a patient with a "Severely Overweight" BMI is about 3 times more likely to experience Anastomotic Leaking when compared to an equivalent patient with a "Healthy" BMI.

We can also note that our confidence interval is even wider than the previous OR interval. We can infer that this is due to us comparing BMI categories that have two degrees of seperation rather than one. We can still see that a majority of the values in the interval are beyond 1, and the wider interval tells us that the risk of Anastomotic Leaking is far greater for Severely Overweight patients than for Healthy individuals. We should still keep the reservation regarding the possibility of no change or even a reduction in odds since our confidence interval does contain 1.

Here is the Odds Ratio for Obese VS Healthy BMI

```
##
## Odds Ratio for Obese VS Healthy BMI =  7.90914
```

```
##
## Confidence Interval = [ 1.670154  ,  37.45433 ]
```

From our Odds Ratio, we can say that a patient with an "Obese" BMI is about 7.9 times more likely to experience Anastomotic Leaking when compared to an equivalent patient with a "Healthy" BMI.

Again, we can see that the Confidence Interval is much wider than the previous two intervals. Thus, we can again infer that since we are comparing Healthy and Obese pateints we can expect to see a much wider margin of change in odds when comparing the risk of Anastomatic Leaking. We can also note that this interval no longer contains values less than one, so we have greater confidence that Obese patients will risk Anastomatic Leaking far greater than healthy patients.

Question 2

Significant Predictors

In order to decide which variables are potentially significant, we will take another look at the summary of our full model.

```
##
## Call:
## glm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
##     x10, family = "binomial", data = C_data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5455  -0.4899  -0.3060  -0.1282   2.4431
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.15583    2.17879  -3.284 0.001022 **
## x1           0.82283    0.51895   1.586 0.112839
## x2           0.09486    0.03227   2.940 0.003282 **
## x3           0.08437    0.02646   3.188 0.001431 **
## x4           0.23469    0.51959   0.452 0.651494
## x5           1.02614    0.57343   1.789 0.073541 .
## x6          -0.85182    0.64633  -1.318 0.187529
## x7          -0.68212    0.73646  -0.926 0.354334
## x8           0.59864    0.59499   1.006 0.314348
## x9          -1.36235    0.36698  -3.712 0.000205 ***
## x10          7.05757    3.60032   1.960 0.049965 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 148.35  on 178  degrees of freedom
## Residual deviance: 111.56  on 168  degrees of freedom
## AIC: 133.56
##
## Number of Fisher Scoring iterations: 6
```

From the summary of our general model, we can note that the p-values for X4 (Race), X6 (Diabetes), X7 (Artery Disease), and X8 (Cancer) are large enough to question whether they are vital to the model or not.In addition, several variables were shown to have residuals that appeared to violate the assumptions for a Logistic regression Model. Since we are using a Logistic Regression Model, we can no longer rely on R^2 as a way to measure the strength of our model. However, we can utlize a cost-benefit method (such as AIC) that will inform us on whether or not a predictor's presence in the model is worth its' respective Information Cost.

Initially, the approach of removing certain predictors one by one and measuring AIC compared to a full model was taken. Although this approach did result in showing which predictors were very significant to the model, it ultimately proved to be quite cumbersome and, depending on the order of variable selection, a bit varying in efficacy. So, instead we will run a stepwise AIC function that considers various models containing our predictors and response, and derive a final working model that minimizes AIC. Thus, our model retains the most statistically significant and beneficial predictors without overcompensating in Information Cost. The following is the result from our function:

```
## Start:  AIC=133.56
## y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10
##
##        Df Deviance    AIC
## - x4    1   111.77 131.77
## - x7    1   112.47 132.47
## - x8    1   112.61 132.60
## - x6    1   113.43 133.43
## <none>      111.56 133.56
## - x1    1   114.16 134.16
## - x5    1   114.99 134.99
## - x10   1   115.32 135.32
## - x2    1   120.51 140.51
## - x3    1   123.62 143.62
## - x9    1   125.75 145.75
##
## Step:  AIC=131.77
## y ~ x1 + x2 + x3 + x5 + x6 + x7 + x8 + x9 + x10
##
##        Df Deviance    AIC
## - x7    1   112.59 130.59
## - x8    1   112.89 130.89
## - x6    1   113.51 131.51
## <none>      111.77 131.77
## - x1    1   114.36 132.36
## - x5    1   115.07 133.07
## - x10   1   116.01 134.01
## - x2    1   120.51 138.51
## - x3    1   123.62 141.62
## - x9    1   125.88 143.88
##
## Step:  AIC=130.59
## y ~ x1 + x2 + x3 + x5 + x6 + x8 + x9 + x10
##
##        Df Deviance    AIC
## - x8    1   113.93 129.93
## - x6    1   114.46 130.46
## <none>      112.59 130.59
## - x1    1   114.67 130.67
## - x5    1   115.57 131.57
## - x10   1   116.75 132.75
## - x2    1   120.68 136.68
## - x3    1   123.63 139.63
## - x9    1   126.10 142.10
##
## Step:  AIC=129.93
## y ~ x1 + x2 + x3 + x5 + x6 + x9 + x10
##
##        Df Deviance    AIC
## <none>      113.93 129.93
## - x6    1   115.95 129.95
## - x5    1   116.15 130.15
## - x1    1   116.39 130.39
## - x10   1   118.71 132.71
## - x2    1   122.49 136.49
## - x3    1   128.93 142.93
## - x9    1   130.12 144.12
```

```
##
## Call:  glm(formula = y ~ x1 + x2 + x3 + x5 + x6 + x9 + x10, family = "binomial",
##     data = C_data1)
##
## Coefficients:
## (Intercept)           x1           x2           x3           x5
##    -6.30447      0.77174      0.08909      0.08276      0.78186
##          x6           x9          x10
##    -0.86507     -1.38995      7.71134
##
## Degrees of Freedom: 178 Total (i.e. Null);  171 Residual
## Null Deviance:       148.3
## Residual Deviance: 113.9     AIC: 129.9
```

We can see that the AIC Step function iterates through various combinations of our available predictors and takes into consideration the Deviance and AIC for each model. Finally, we are presented with a final model that consists of Statistically Significant predictors of Anastomotic Leaking following a colectomy.

Our reduced model is defined as:

$$y = x_1 + x_2 + x_3 + x_5 + x_6 + x_9 + x_{10}$$

Where y and our x values are as previously defined. That is Gender, BMI, Age, Tobacco Usage, Diabetes, Albumin, and Operation Length are the predictors that will be used for our reduced model.

We can see that BMI remains a significant predictor, as expected. The numerical predictors of Age,Albumin and Operation Length remain in the model as well. We can deduce that, in addition to their p-values being relatively small, these numerical predictors would be expected to play a big part in experiencing Anastomotic Leaking.

The categorical predictors of Gender,Tobacco Usage, and Diabetes also were kept in the reduced model. For gender, we can theorize that there is a difference in how men and women may react to this type of operation. Also, Tobacco Usage and Diabetes make sense when considering the possible consequences that such characteristics could have on the human body.

We should also focus on the predictors that were left out of the model: CAD/PAD,Cancer and Race.

For CAD/PAD and Cancer, we would expect these two diseases to play some part in the risk of Anaostomotic Leaking. However, it is important to note that Diabetes is another disease that did make it into the reduced model. Thus, there is a high likelihood that the three predictors are highly correlated to one another, and the AIC step model chose the predictor which contributed most to the model. Therefore we are left with Diabetes as our remaining predictor from the three disease predictors.

We will provide a correlation plot of our categorical variables to investigate our assumption.

Table 1: Correlation of Colectomy Data

	Gender	BMI	Age	Race	Tobacco	Diabetes	Artery	Cancer	Albumin	Length	Leak
Gender	1.000	-0.113	-0.049	0.118	0.061	0.011	0.154	0.005	0.025	0.118	0.093
BMI	-0.113	1.000	0.051	-0.055	-0.176	0.267	0.105	0.094	0.075	0.089	0.124
Age	-0.049	0.051	1.000	-0.010	-0.126	0.227	0.243	0.248	0.193	-0.091	0.146
Race	0.118	-0.055	-0.010	1.000	-0.037	0.146	0.122	0.092	-0.087	0.129	0.039
Tobacco	0.061	-0.176	-0.126	-0.037	1.000	-0.166	0.042	-0.221	-0.055	-0.047	0.052
Diabetes	0.011	0.267	0.227	0.146	-0.166	1.000	0.157	0.082	-0.037	-0.014	-0.015
Artery	0.154	0.105	0.243	0.122	0.042	0.157	1.000	0.056	-0.052	0.010	0.047
Cancer	0.005	0.094	0.248	0.092	-0.221	0.082	0.056	1.000	-0.172	0.117	0.188
Albumin	0.025	0.075	0.193	-0.087	-0.055	-0.037	-0.052	-0.172	1.000	-0.177	-0.243
Length	0.118	0.089	-0.091	0.129	-0.047	-0.014	0.010	0.117	-0.177	1.000	0.192
Anastomotic	0.093	0.124	0.146	0.039	0.052	-0.015	0.047	0.188	-0.243	0.192	1.000

We can see that CAD/PAD does have a correlation that is higher with two predictors (Age and Diabetes) than with our Response (Anastomotic Leaking). We can also observe that Cancer has a higher correlation with Tobacco Usage than it does with the response. These incidents of collinearity as well as the dropped predictors having large p-values are a good hypothesis for the reason why they can be dropped from the model and still reduce AIC in the process.

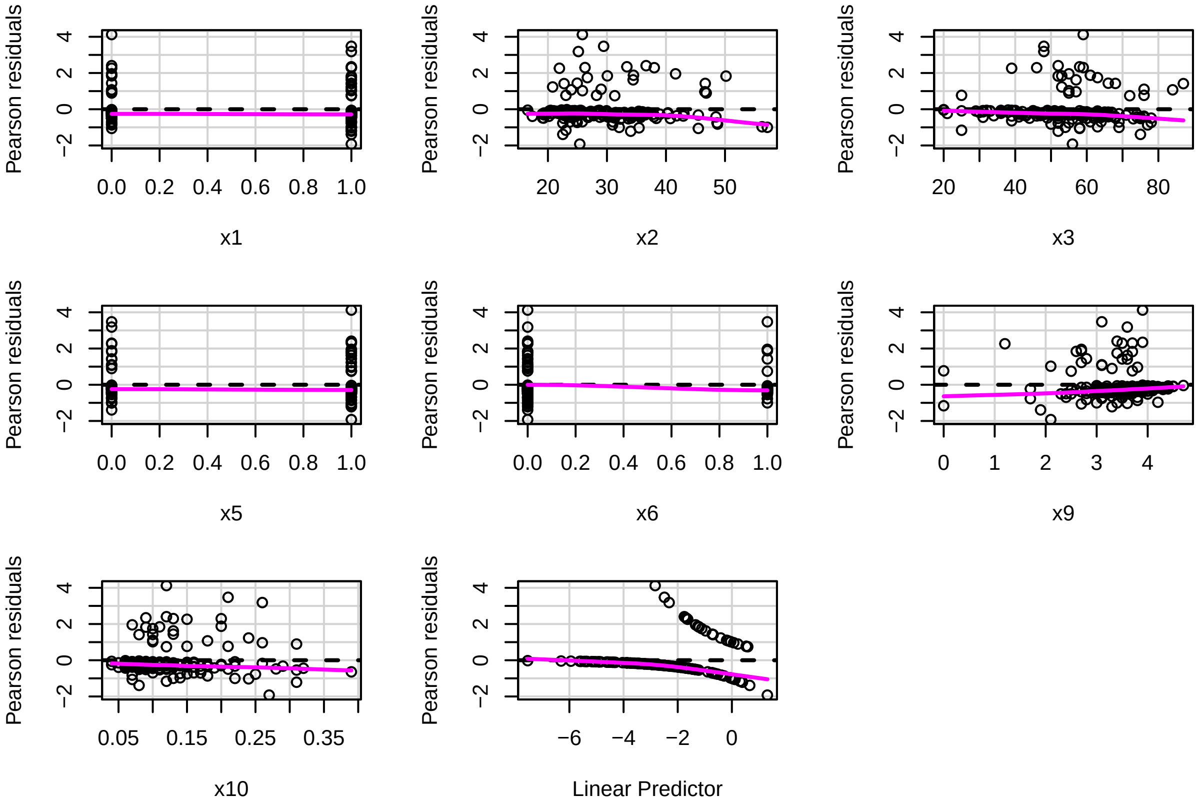
As for Race, there are no other predictors that can easily be associated with Race to infer correlation. In addition, we are unsure if Race would really make a difference in the risk of Anastomotic Leaking. Looking into the data, we can see that the ratio of AA to W patients that experienced Anastomotic Leaking is nearly 50:50. Thus, the race of the patient should not make any difference in the overall risk.

Now, we have our model consisting of possibly statistically significant predictors. We will now provide the summary for our final model:

```
##
## Call:
## glm(formula = y ~ x1 + x2 + x3 + x5 + x6 + x9 + x10, family = "binomial",
##     data = C_data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7600  -0.5141  -0.3220  -0.1457   2.4042
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.30447    1.95529  -3.224 0.001263 **
## x1           0.77174    0.49998   1.544 0.122699
## x2           0.08909    0.03041   2.930 0.003389 **
## x3           0.08276    0.02350   3.522 0.000428 ***
## x5           0.78186    0.53601   1.459 0.144655
## x6          -0.86507    0.63472  -1.363 0.172911
## x9          -1.38995    0.35815  -3.881 0.000104 ***
## x10          7.71134    3.46733   2.224 0.026148 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 148.35  on 178  degrees of freedom
## Residual deviance: 113.93  on 171  degrees of freedom
## AIC: 129.93
##
## Number of Fisher Scoring iterations: 6
```

We are able to view our paramter estimates, standard errors and the reduced AIC value for this model. We can utilize these parameter estimates to discuss how each predictor affects the risk of experiencing Anastomotic Leaking. Before jumping into aalyzing the risk of A.L. associated with each predictor, we will review our Residual Plots to ensure the model is still valid.



```
##     Test stat Pr(>|Test stat|)
## x1     0.0000          1.00000
## x2     1.5629          0.21124
## x3     0.0508          0.82162
## x5     0.0000          1.00000
## x6     0.0000          1.00000
## x9     0.1820          0.66967
## x10    3.2534          0.07128 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that some of the deviation from the zero line is still occuring for some of our predictors, but we will choose to continue on without making transformations to resolve the issue of non-linearity. We would ultimately like to see our response have a strictly horizontal line that averages out to a zero value for the residuals, but for the time being we will accept our model and make corresponding inference.

Analyzing Predictor Influence

Now, we can discuss how each predictor contributes to the Risk of Anastomotic Leaking in the context of all other utilized variables.

We will begin with **x1: Gender**.

```
##
## Change in Odds From a Change in Gender from Female to Male =  2.163537
```

```
##
##  Confidence Interval = [ 0.8120281  ,  5.764446 ]
```

We can observe that a change in gender from Female to Male changes the odds of experiencing Anastomotic Leaking by a factor of 2.163537. We can also see that the average odds confidence interval contains values ranging from .81 to 5.7. Thus, we can say that we are 95% confident that Males do in fact risk Anastomotic Leaking following a Colectomy moreso than Females (by a factor of about 2.16). We should note that the Confidence Interval does contain values that are less than one, but a large majority of values still reside above one. Thus, we can say that there is a large amount of evidence to say that being Male increases the risk of Anastomotic Leaking, compared to Female patients.

For x2: BMI

```
##
## Change in Odds From a One Unit Increase in BMI =  1.093183
```

```
##
##  Confidence Interval = [ 1.029936  ,  1.160315 ]
```

We can observe that a one unit increase in BMI changes the odds of experiencing Anastomotic Leaking by a factor of 1.093. We can also see that the average odds confidence interval contains values ranging from 1.02 and 1.16. Thus, we can say that we are confident that an increase in BMI does in fact increase the risks of Anastomotic Leaking following a Colectomy. We can also note that our Odds Ratio value and corresponding Confidence Interval are not that different from the original full model's values. Thus, we can say that BMI has a similar magnitude of effect for both scenarios.

For x3: Age

```
##
## Change in Odds From a One Unit Increase in Age =  1.086286
```

```
##
## Confidence Interval = [ 1.037386  ,  1.13749 ]
```

We can observe that a one unit increase in Age changes the odds of experiencing Anastomotic Leaking by a factor of 1.086. We can also see that the average odds confidence interval contains values ranging from 1.03 and 1.13. Thus, we can say that we are confident that an increase in Age does in fact increase the risks of Anastomotic Leaking following a Colectomy.

For x5: Tobacco usage

```
##
## Change in Odds From a Change in Tobacco Usage from No to Yes =  2.185541
```

```
##
## Confidence Interval = [ 0.7643613  ,  6.249127 ]
```

We can observe that a change in Tobacco Usage from No to Yes changes the odds of experiencing Anastomotic Leaking by a factor of 2.18. We can also see that the average odds confidence interval contains values ranging from .76 to 6.24. Thus, we can say that we are confident that Smokers do in fact risk Anastomotic Leaking following a Colectomy moreso than Non-Smokers (by a factor of about 2.18). We should note that the Confidence Interval does contain values that are less than one, but a large majority of values still reside above one. Thus, we can say that there is a large amount of evidence to say that being a Smoker increases the risk of Anastomotic Leaking, compared to Non-Smoking patients.

For x6: Diabetes

```
##
## Change in Odds From a Change in Diabetes from No to Yes =  0.4210234
```

```
##
## Confidence Interval = [ 0.1213446 , 1.460804 ]
```

We can observe that a change in Diabetes from No to Yes changes the odds of experiencing Anastomotic Leaking by a factor of .42. We can also see that the average odds confidence interval contains values ranging from .12 to 1.4. We should note that the Confidence Interval does contain a good proportion of values that are greater than one,so for this particular variable we are not as confident that a patient having diabetes will consistently lower the risk of Anastomotic Leaking.

For x9: Albumin

```
##
## Change in Odds From a One Unit Increase in albumin =  0.2490874
```

```
##
## Confidence Interval = [ 0.123448  ,  0.5025964 ]
```

We can observe that a one unit increase in Albumin changes the odds of experiencing Anastomotic Leaking by a factor of .25. We can also see that the average odds confidence interval contains values ranging from .12 and .50. Thus, we can say that we are confident that an increase in Albumin does in fact decrease the risks of Anastomotic Leaking following a Colectomy.

For x10: Operation Length

```
##
## Change in Odds From a One Unit Increase in Operation Length =  2233.544
```

```
##
## Confidence Interval = [ 2.497741  ,  1997291 ]
```

We can observe that a one unit increase (one day) in Operation Length changes the odds of experiencing Anastomotic Leaking by a factor of 2233.544. Thus, we can equivalently say that a one minute increase in Operation Length changes the odds of experiencing Anastomotic Leaking by a factor of about 1.55 We can also see that the average odds confidence interval contains values ranging from 2.5 to 1997291. Thus, we can say that we are confident that an increase in Operation Length does in fact increase the risks of Anastomotic Leaking following a Colectomy.

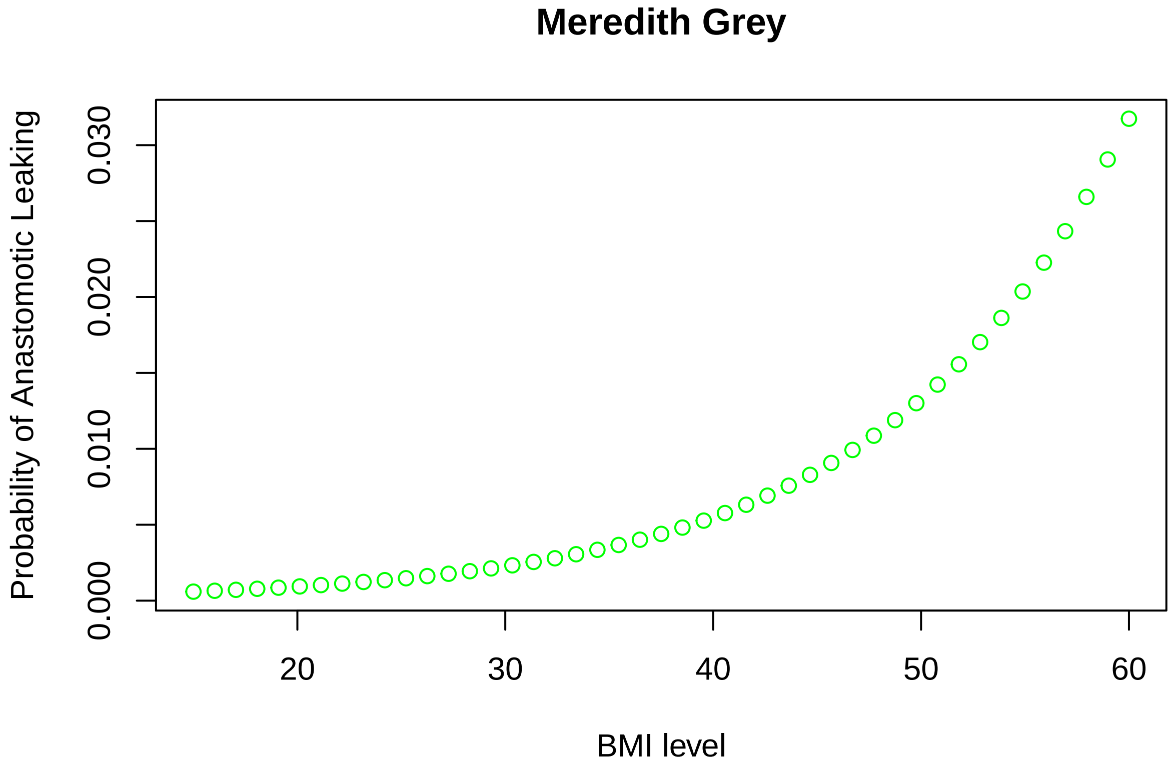
NOTE: It is important to remember that the unit increase for Operation Length is measured in Days. However, our data shows Operation Lengths that are at most about 2/5 of a day in length. We should take this into considertion when considering the magnitude that operation length plays into the risk of Anastomotic Leaking.

Problem 3

For these case studies, we will be ignoring any traits provided that are not included in our reduced model. Essentially, we will assume that the value of Race,PAD, and Cancer are zero and do not contribute towards the risk of Anastomotic Leaking.

Case 1: Meredith Grey

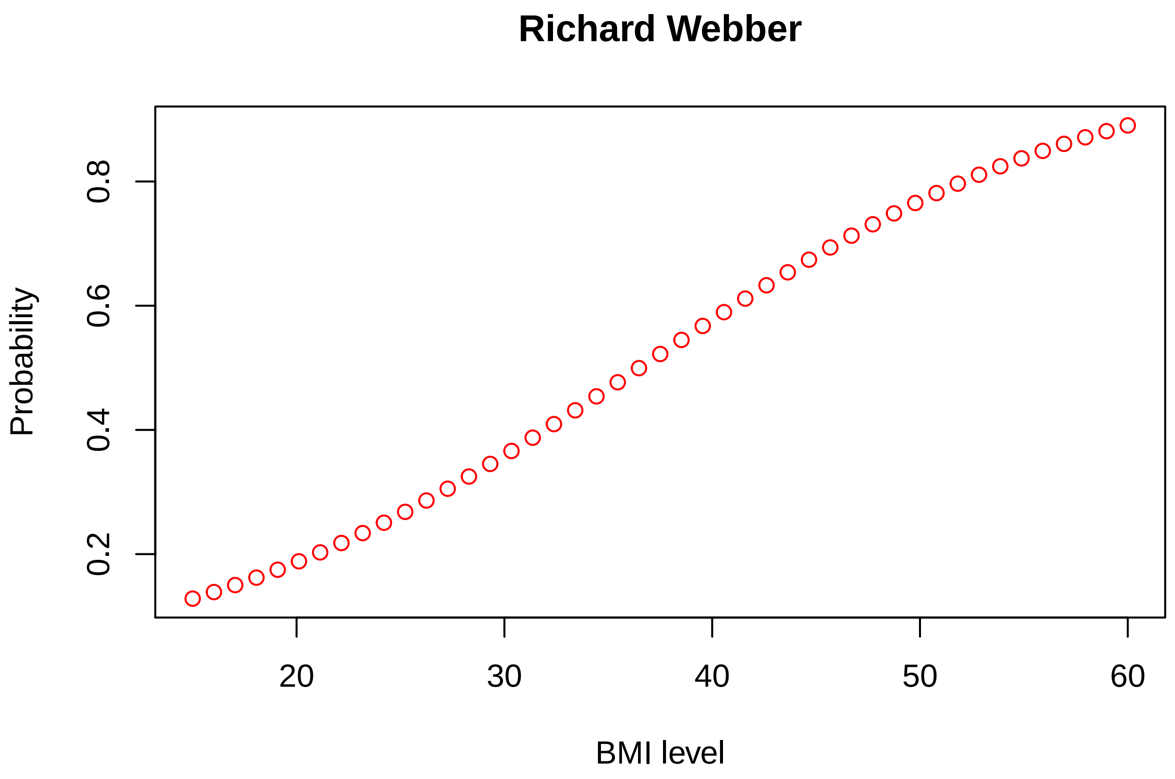
For our first case, we will input our data values and plot the Probability of Anastomotic Leaking (y) over an interval of BMI values ranging from Healthy to Obese.



Thus we can initially observe that an increase in BMI does increase the Probability of Anastomotic Leaking for this case. However, we can note that the probabilities presented in the graph are quite low (ranging from about 0 to .03). Thus, we can infer that a younger female who doesn't use tobacco, doesn't have diabetes, has a somewhat high Albumin value, and had a relatively quick operation will have a significantly low risk of Anastomotic Leaking regardless of how high their BMI value is. It is worth noting, however, that a high BMI still increases this risk in a somewhat exponential manner.

Case 2: Richard Webber

For our second case, we will input our data values and plot the Probability of Anastomotic Leaking (y) over an interval of BMI values ranging from Healthy to Obese.



We can still observe that an increase in BMI does increase the Probability of Anastomotic Leaking for this case. Here, we can note that the probabilities presented in the graph are quite high (ranging from about .2 to .8). Thus, we can infer that an older male who uses tobacco, has diabetes, has a low Albumin value, and had a longer operation will have a significantly high risk of Anastomotic Leaking that is increased significantly as the BMI level is increased. Even for a healthy BMI level, there is a 20% chance of Anastomotic Leaking. In addition, there is an 80% chance of A.L. when the patient has an Obese BMI.Thus, we can see that the BMI value for this patient has a much larger impact than for the previous case study.

Appendix

SET-UP

LOADING LIBRARIES:

```
library(readr)
library(car)
library(MASS)
```

Cleaning Data:

```
C_data= read.csv("/Users/gustavo/Desktop/MATH 535/535 Exam 2/colon.csv",
                 header=TRUE,stringsAsFactors = FALSE)

#Get rid of unwanted factors
C_data1 = C_data
C_data1 = C_data1[,-1:-4] #PI,ICD9,CPT, and Procedure are not needed
C_data1 = C_data1[,-2:-3] #Don't need Weight or height, using BMI
C_data1 = C_data1[,-10:-11] #Not necessary since Operation Length is being used.
C_data1 = C_data1[,-12] #Get rid of

#Get Race to be a true Binary variable
C_data1[C_data1$Race !="AA",4]="W"
```

Problem 1

Assigning Variables:

```
x1 = C_data1$Gender #categorical (Binary)
x2 = C_data1$BMI #numerical
x3 = C_data1$Age #numerical
x4 = C_data1$Race #categorical (Binary)
x5 = C_data1$Tobacco #categorical (binary)
x6 = C_data1$DM #categorical
x7 = C_data1$CAD.PAD #categorical
x8 = C_data1$Cancer #categorical
x9 = C_data1$Albumin..g.dL. #numerical
x10 = C_data1$Operative.Length #numerical
y = C_data1$Anastamotic.Leak #categorical
```

Making Binary Variables:

```
#CREATING BINARY VALUES FOR CATEGORICAL VARIABLES

x1[x1=="Male"]=1
x1[x1=="Female"]=0;
x1=as.numeric(x1)

x4[x4=="AA"]=1
x4[x4=="W"]=0;
x4=as.numeric(x4)
```

Summary of Model:

```
model = glm(y ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10, family = "binomial", data=C_data1)
summary(model)
```

Residuals of Model:

```
#RESIDUAL PLOTS FOR GENERAL MODEL
model = glm(y ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10, family = "binomial", data=C_data1)
residualPlots(model)
```

Numerical BMI Interpretation

```
beta_BMI= model$coefficients[3]
ste_BMI= summary(model)$coefficients[3,2]
OR_BMI = exp(beta_BMI)

#Confidence interval
z_star = 1.96
L = exp(beta_BMI - z_star*ste_BMI)
U = exp(beta_BMI + z_star*ste_BMI)
cat("\nChange in Odds From a One Unit Increase in BMI = ",OR_BMI,"")
cat("\n")
cat("\nConfidence Interval = [",L," , ",U,"]")
```

Categorical BMI Code

```
C_data1$cat_BMI = "healthy"

C_data1$cat_BMI[C_data1$BMI>=30 & C_data1$BMI<35] = "overweight"
C_data1$cat_BMI[C_data1$BMI>=35 & C_data1$BMI<40] = "severely overweight"
C_data1$cat_BMI[C_data1$BMI>=40] = "obese"
x2_catBMI = C_data1$cat_BMI
```

Summary of BMI Categorical:

```
model_categorical = glm(y ~ x1+x2_catBMI+x3+x4+x5+x6+x7+x8+x9+x10,
                        family = "binomial", data=C_data1)
summary(model_categorical)
```

Odds of Overweight VS healthy:

```
#odds of Overweight vs Healthy
beta_Overweight=model_categorical$coefficients[4]
Ste_Overweight=summary(model_categorical)$coefficients[4,2]


OR_Overweight = exp(beta_Overweight)

#Confidence interval
z_star = 1.96
L = exp(beta_Overweight - z_star*Ste_Overweight)
U = exp(beta_Overweight + z_star*Ste_Overweight)
```

```
cat("\nOdds Ratio for Overweight VS Healthy BMI = ",OR_Overweight,"")
cat("\n")
cat("\nConfidence Interval = [",L," , ",U,"]")
```

Odds Ratio for Severely Overweight vs Healthy BMI:

```
#odds of Severely Overweight vs Healthy
beta_Severe=model_categorical$coefficients[5]
Ste_Severe=summary(model_categorical)$coefficients[5,2]

OR_Severe = exp(beta_Severe)

#Confidence interval
z_star = 1.96
L = exp(beta_Severe - z_star*Ste_Severe)
U = exp(beta_Severe + z_star*Ste_Severe)


cat("\nOdds Ratio for Severely Overweight VS Healthy BMI = ",OR_Severe,"")
cat("\n")
cat("\nConfidence Interval = [",L," , ",U,"]")
```

Odds Ratio for Obese VS Healthy BMI:

```
#odds of Obese vs Healthy
beta_Obese=model_categorical$coefficients[3]
Ste_Obese=summary(model_categorical)$coefficients[3,2]

OR_Obese= exp(beta_Obese)

#Confidence interval
z_star = 1.96
L = exp(beta_Obese - z_star*Ste_Obese)
U = exp(beta_Obese + z_star*Ste_Obese)

cat("\nOdds Ratio for Obese VS Healthy BMI = ",OR_Obese,"")
cat("\n")
cat("\nConfidence Interval = [",L," , ",U,"]")
```

Problem 2

Step AIC Function:

```
stepAIC(model)
```

Correlation Plot:

```
Gender=x1
BMI=x2
Age=x3
Race=x4
Tobacco=x5
Diabetes=x6
Artery=x7
Cancer=x8
```

```
Albumin=x9
Length=x10
Anastomotic=y

categories= cbind(Gender,BMI,Age,Race,Tobacco,Diabetes,Artery,Cancer,Albumin,Length,Anastomotic)
category=cor(categories)


colnames(category) =
c("Gender", "BMI", "Age", "Race", "Tobacco", "Diabetes", "Artery", "Cancer", "Albumin", "Length", "Leak
knitr::kable(category, digits = 3, caption = "Correlation", align = 'c')
```

Reduced Model Summary:

```
model_final = glm(y ~x1+x2+x3+x5+x6+x9+x10, family = "binomial", data=C_data1)
summary(model_final)
```

Reduced Model Residuals:

```
residualPlots(model_final)
```

X1(GENDER) Interpretation:

```
beta_gender = model_final$coefficients[2]
ste_gender = summary(model_final)$coefficients[2,2]

OR_gender = exp(beta_gender)

#Confidence interval
z_star = 1.96
L = exp(beta_gender - z_star*ste_gender)
U = exp(beta_gender + z_star*ste_gender)

cat("\nChange in Odds From a Change in Gender from Female to Male = ",OR_gender,"")
cat("\n Confidence Interval = [",L," , ",U,"]")
```

X2(BMI) Interpretation:

```
beta_BMI = model_final$coefficients[3]
ste_BMI = summary(model_final)$coefficients[3,2]

OR_BMI = exp(beta_BMI)

#Confidence interval
z_star = 1.96
L = exp(beta_BMI - z_star*ste_BMI)
U = exp(beta_BMI + z_star*ste_BMI)


cat("\nChange in Odds From a One Unit Increase in BMI = ",OR_BMI,"")
cat("\n Confidence Interval = [",L," , ",U,"]")
```

X3 (AGE) Interpretation:

```
beta_age = model_final$coefficients[4]
ste_age = summary(model_final)$coefficients[4,2]

```

```
OR_age = exp(beta_age)

#Confidence interval
z_star = 1.96
L = exp(beta_age - z_star*ste_age)
U = exp(beta_age + z_star*ste_age)

cat("\nChange in Odds From a One Unit Increase in Age = ",OR_age,"")
cat("\nConfidence Interval = [",L," , ",U,"]")
```

X5 (TOBACCO) Interpretation:

```
beta_tobacco = model_final$coefficients[5]
ste_tobacco = summary(model_final)$coefficients[5,2]

OR_tobacco = exp(beta_tobacco)

#Confidence interval
z_star = 1.96
L = exp(beta_tobacco - z_star*ste_tobacco)
U = exp(beta_tobacco + z_star*ste_tobacco)

cat("\nChange in Odds From a Change in Tobacco Usage from No to Yes = ",OR_tobacco,"")
cat("\nConfidence Interval = [",L," , ",U,"]")
```

X6 (DIABETES) Interpretation:

```
beta_diabetes = model_final$coefficients[6]
ste_diabetes = summary(model_final)$coefficients[6,2]

OR_diabetes = exp(beta_diabetes)

#Confidence interval
z_star = 1.96
L = exp(beta_diabetes - z_star*ste_diabetes)
U = exp(beta_diabetes + z_star*ste_diabetes)

cat("\nChange in Odds From a Change in Diabetes from No to Yes = ",OR_diabetes,"")
cat("\nConfidence Interval = [",L,",",U,"]")
```

X9 (ALBUMIN) Interpretation:

```
beta_albumin = model_final$coefficients[7]
ste_albumin = summary(model_final)$coefficients[7,2]

OR_albumin = exp(beta_albumin)

#Confidence interval
z_star = 1.96
L = exp(beta_albumin - z_star*ste_albumin)
U = exp(beta_albumin + z_star*ste_albumin)

cat("\nChange in Odds From a One Unit Increase in albumin = ",OR_albumin,"")
cat("\nConfidence Interval = [",L," , ",U,"]")
```

X10 (OPERATION LENGTH) Interpretation:

```
beta_length = model_final$coefficients[8]
ste_length = summary(model_final)$coefficients[8,2]

OR_length = exp(beta_length)

#Confidence interval
z_star = 1.96
L = exp(beta_length - z_star*ste_length)
U = exp(beta_length + z_star*ste_length)

cat("\nChange in Odds From a One Unit Increase in Operation Length = ",OR_length,"")
cat("\nConfidence Interval = [",L," , ",U,"]")
```

Problem 3

Case 1: Meredith Grey

```
BMI = seq(15,60,length=45)
Case1 = predict(model_final,newdata=data.frame(x1=0,x2=BMI,x3=35, x5=0, x6 = 0,
                    x9=4.2, x10=0.0625), type="response")

plot(BMI,Case1, xlab="BMI level", ylab="Probability of Anastomotic Leaking",
     main="Meredith Grey",
     col="Green")
```

Case 2: Richard Webber

```
BMI = seq(15,60,length=45)
Case2 = predict(model_final,newdata=data.frame(x1=1,x2=BMI,x3=62, x5=1, x6=1,
                    x9=2.8, x10=.1458), type="response")

plot(BMI,Case2, xlab="BMI level", ylab="Probability",
     main="Richard Webber",
     col="Red")
```