Identify the relationship between referral rate and potential predictors using logistic regression

Elisa Zhang, Shicong Wang

4/8/2022

Introduction

In this report, we focus on implementing a logistic regression model to identify the relationship between potential predictors variables and genetic referral rate. The predictor variables that we will include in the model are cancer stage, ECOG level, sex assigned at birth, ethnicity and age.

The report will include visualizations of the relationship between each potential predictor and referral rate. From the visualizations, we might conclude that there is not obvious relationship between each predictor and the response variable.

After implementing the logistic regression model, the model output is aligned with what the visualizations seem to be telling us–we do not see strong evidence of a relationship between any of the predictors and referral rate.

One potential concern would be multicollinearity - whether the predictor variables are highly correlated. Since the number of observations in the contingency table of ECOG level and Cancer stage is too small to satisfy the rule of thumbs of chi-sqaure test of independence, we include two additional models in the appendix to check the multicollinearity. One model omits ECOG from the predictors, and the other omits stage from the predictors. The result came out that there is no difference between two, which indicates that the two predictor variable is not correlated based on the data we have. In all models, there is no strong evidence of a relationship between the predictors and the response.

Visulization

In this part, we generate the stacked bar plots for each combination of predictor and outcome variable. The error bars in the following figures are calculated using the formula in the appendix.

Notice that in each figure, the dotted line and shaded band in this figure is the overall estimate of the referral rate and 95% confidence interval, while the orange bars are 95% confidence intervals for each proportion.

1. Cancer Stage

The figure below illustrates the relationship between referral rate and cancer stages. There are many overlaps among error bars. The cancer stage might not be related to the referral rate.

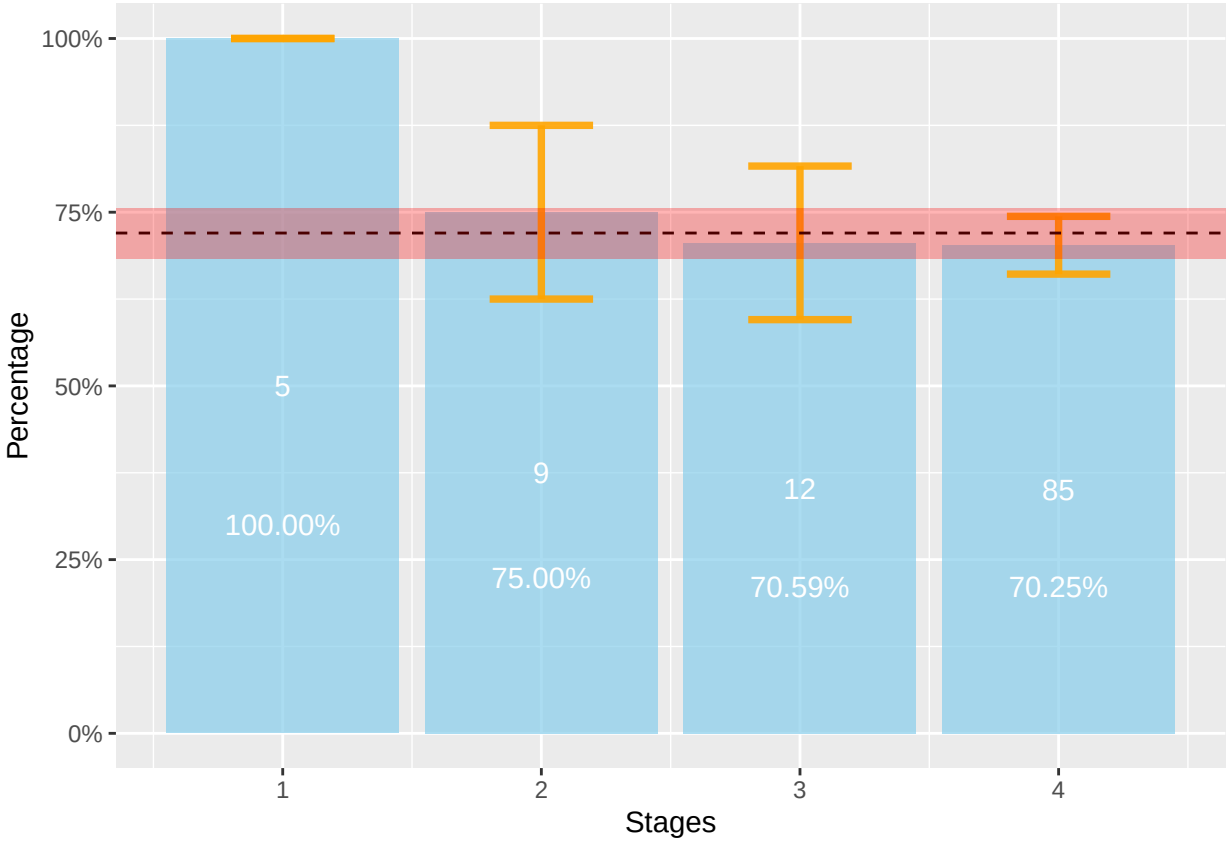


Figure 1: Proportion of Patients Who got referrals in each Cancer Stage

2. ECOG level

The figure below shows the relationship between ECOG status and referral rate. We can conclude that there is no obvious relationship between ECOG status and the referral rate.

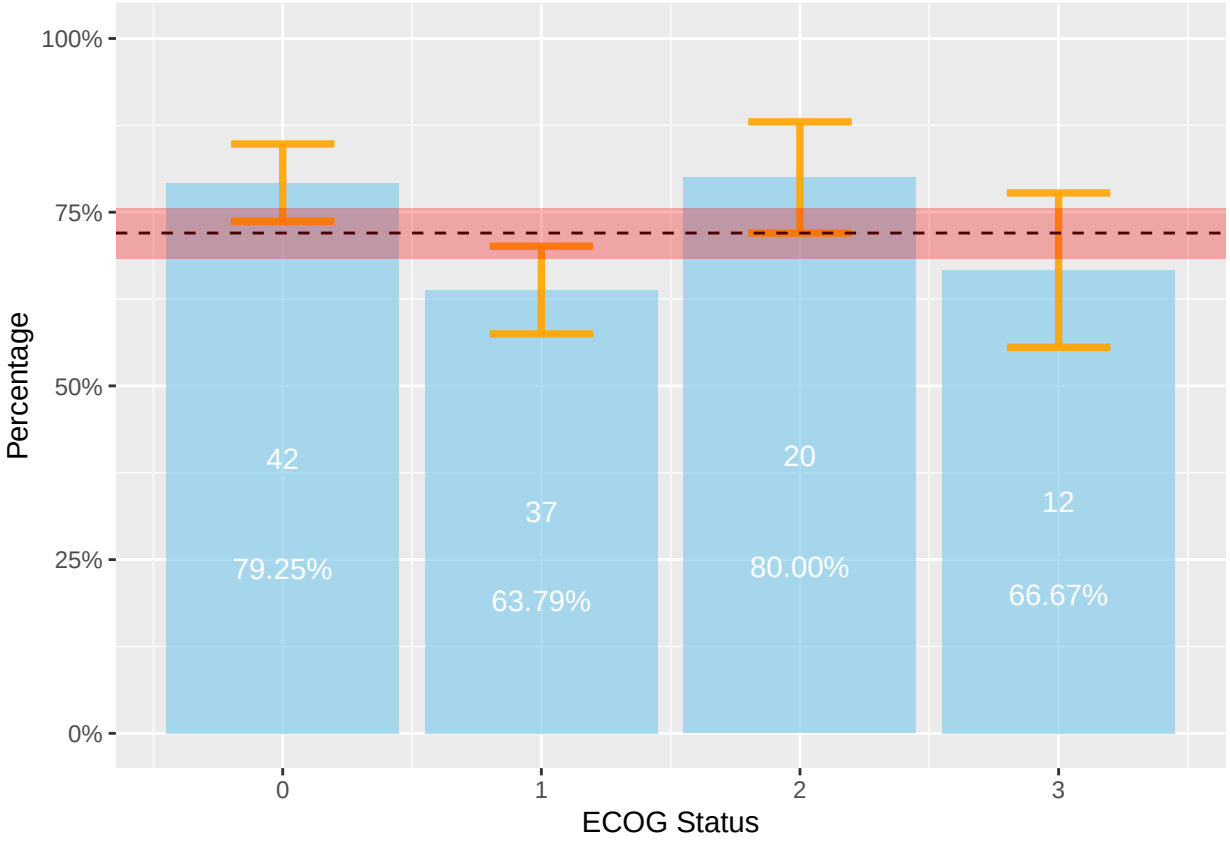
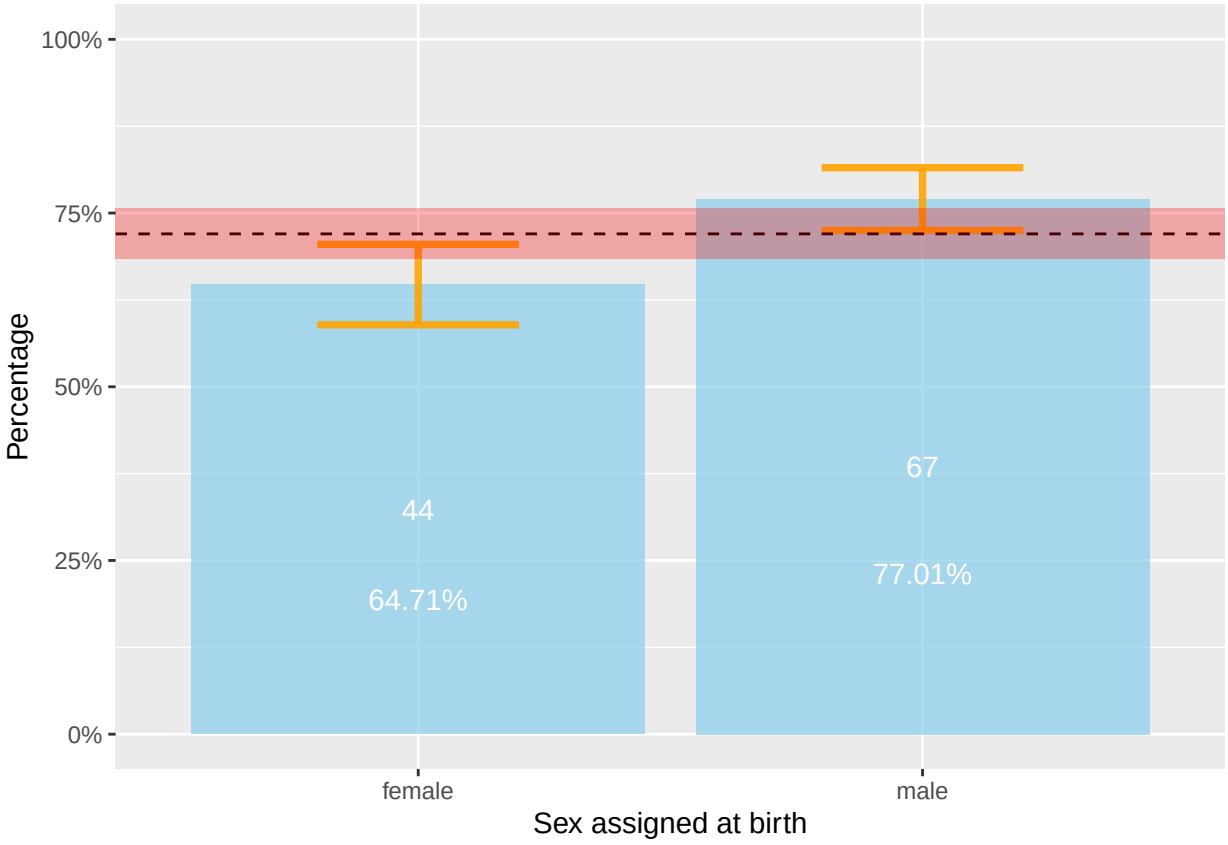


Figure 2: Proportions of Patients Who got referral rate in each ECOG Level

3. Sex Assigned at Birth

From the the plot below, we might conclude that gender is related to the referral rate.



4. Ethnicity

Since some ethnicity only contain few observations, we divide the race into white and not white and use as the new variable. And the figure below does not show that there is an obvious relationship between race and referral rate.

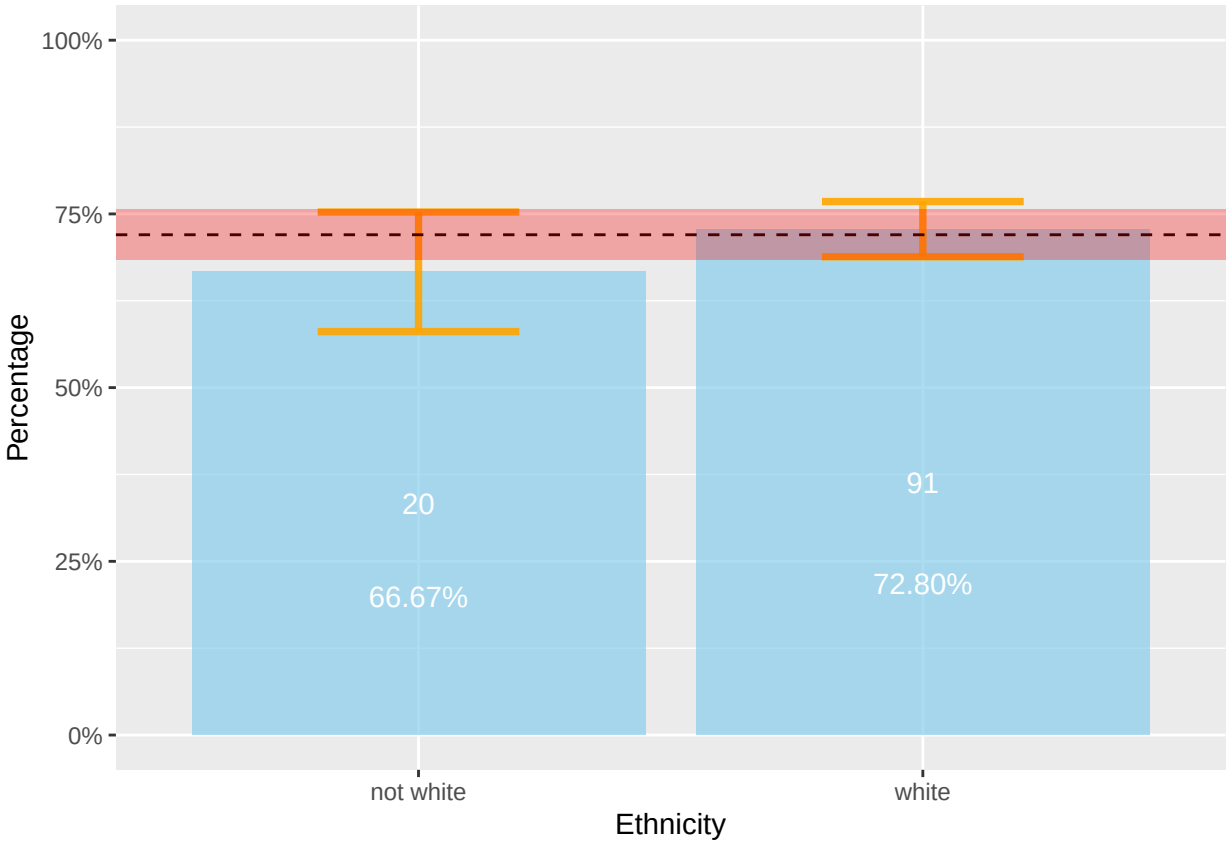


Figure 3: Proportions of Patients Who got referral rate among races

5. Age

We divide Age into groups using quantiles. The first group contains patients from 32 to 62 years old. And the second group contains patients aged from 62 to 68. And the third group include patients from 68 to 75.5 years old. And the last group include patients whose age are in the 4th quantile.

The error bar for individuals in the first quantile, [32.0,62.0], does not overlap with the other bars or the overall referral rate error bars. While this could suggest that the referral rate in this quantile is different, given that we have produced 15 estimates with error bars in this report, it would be unsurprising to see some which does not overlap with the others even if the true referral rate was the same in all groups. In our model, we do not find strong evidence of a relationship between age and referral rates.

```
##   0%  25%  50%  75% 100%
## 32.0 62.0 68.0 75.5 96.0
```

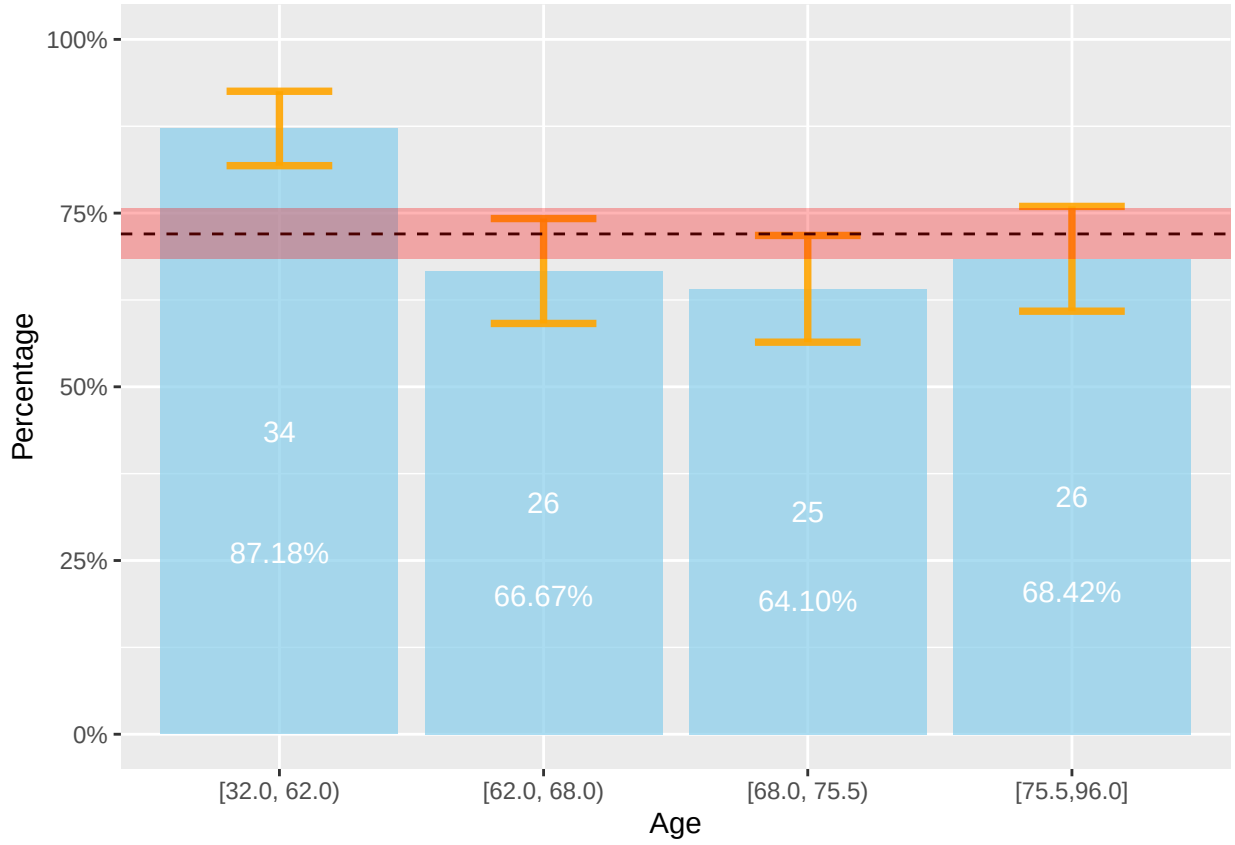


Figure 4: Proportions of Patients Who got referral rate for white/non-white

Model Fit

We will use logistic regression to fit our data since we have binary outcome of whether the patients got a referral from GIM or not. We will fit three models. The full model will include both cancer stage and ECOG level. And the other two will only include either ECOG level or Cancer Stage which are put in appendix.

Below is the full model:

```
model1 <- glm(`Genetics Referrals` ~  `Stage at Dx (#0-4)` +
                `ECOG at Initial (#0-4)` +
                `Age at Dx (#)` +
                `Ethnicity/ Ancestry` + `Sex Assigned
at Birth (m/f)` , family = "binomial", data = dat_1)
summary(model1)
```

```
##
## Call:
## glm(formula = `Genetics Referrals` ~ `Stage at Dx (#0-4)` + `ECOG at Initial (#0-4)` +
##     `Age at Dx (#)` + `Ethnicity/ Ancestry` + `Sex Assigned\nat Birth (m/f)`,
##     family = "binomial", data = dat_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8614  -1.3437   0.7106   0.8358   1.2903
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       3.4843634  1.8067729   1.929   0.0538 .
## `Stage at Dx (#0-4)`             -0.2976799  0.2814298  -1.058   0.2902
## `ECOG at Initial (#0-4)`         -0.0003937  0.1960274  -0.002   0.9984
## `Age at Dx (#)`                  -0.0277649  0.0198487  -1.399   0.1619
## `Ethnicity/ Ancestry`white        0.2280004  0.4502499   0.506   0.6126
## `Sex Assigned\\nat Birth (m/f)`male  0.5031592  0.3687174   1.365   0.1724
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 184.94  on 154  degrees of freedom
## Residual deviance: 178.24  on 149  degrees of freedom
## AIC: 190.24
##
## Number of Fisher Scoring iterations: 4
```

Since the all of p-values are greater than 0.05, we don't have enough confidence to reject the null hypothesis that the variables have no correlations with the dependent variable. Therefore, it's insufficient to infer that these variables contribute significantly to genetic referrals.

Robustness of model results to association between ECOG and cancer stage

In case of the correlation between ECOG level and cancer stage, we consider to extract either of them from the full model and compare them to the full model. However, conclusions have not changed so far.

We consider the following results:

(1) According to the corresponding information from client, we note that many patients even with high cancer stage are asymptomatic. Since ECOG measures well being, it is reasonable that there might not be a relationship between these variables.

(2) By fitting three models, we checked whether the analysis is robust to possible association. Then we find the results of three models are quite similar that all of which are not significant enough. In that case, association between these variables does not seem to have an effect on the model results.

Therefore, we still maintain the full model in this section.

Diagnostic Plot

The diagnostic plot can be used to measure the goodness of fit of the model.

Binned residual plot

The binned residual plot can assess the overall fit of regression models for binary outcomes. Since most of points are within the grey line and there is no obvious pattern for the points, the model does not have much problem.

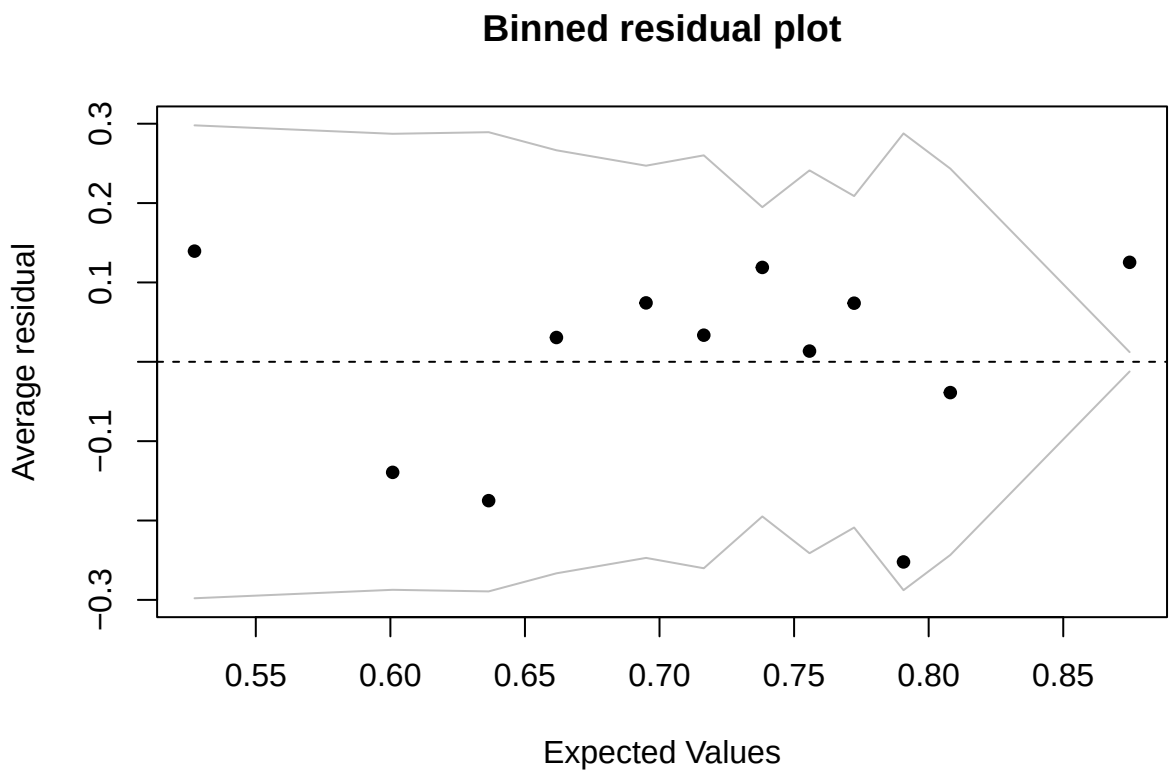


Figure 5: Binned residual plot for full model

Appendix

two sub-models

Model without Cancer Stage

```
model2 <- glm(`Genetics Referrals` ~   `ECOG at Initial (#0-4)`
              + `Age at Dx (#)` +
                `Ethnicity/ Ancestry` +
                `Sex Assigned
at Birth (m/f)` ,
family = "binomial", data = dat_1)
summary(model2)
```

```
##
## Call:
## glm(formula = `Genetics Referrals` ~ `ECOG at Initial (#0-4)` +
##     `Age at Dx (#)` + `Ethnicity/ Ancestry` + `Sex Assigned\nat Birth (m/f)`,
##     family = "binomial", data = dat_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8419  -1.3603   0.7124   0.8382   1.2384
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                            2.29660    1.39086   1.651   0.0987 .
## `ECOG at Initial (#0-4)`              -0.04970    0.19004  -0.262   0.7937
## `Age at Dx (#)`                       -0.02543    0.01943  -1.309   0.1907
## `Ethnicity/ Ancestry`white             0.19515    0.44769   0.436   0.6629
## `Sex Assigned\\nat Birth (m/f)`male    0.52785    0.36674   1.439   0.1501
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 184.94  on 154  degrees of freedom
## Residual deviance: 179.46  on 150  degrees of freedom
## AIC: 189.46
##
## Number of Fisher Scoring iterations: 4
```

Model without ECOG level

```
model3 <- glm(`Genetics Referrals` ~  `Stage at Dx (#0-4)` +
                `Age at Dx (#)` +
                `Ethnicity/ Ancestry` +
                `Sex Assigned
at Birth (m/f)` , family = "binomial", data = dat_1)
summary(model3)
```

```
##
## Call:
## glm(formula = `Genetics Referrals` ~ `Stage at Dx (#0-4)` + `Age at Dx (#)` +
##     `Ethnicity/ Ancestry` + `Sex Assigned\nat Birth (m/f)`, family = "binomial",
##     data = dat_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8613  -1.3438   0.7106   0.8357   1.2903
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          3.48520    1.75846   1.982   0.0475 *
## `Stage at Dx (#0-4)`                -0.29780    0.27537  -1.081   0.2795
## `Age at Dx (#)`                     -0.02778    0.01862  -1.492   0.1357
## `Ethnicity/ Ancestry`white           0.22809    0.44797   0.509   0.6106
## `Sex Assigned\\nat Birth (m/f)`male  0.50323    0.36703   1.371   0.1703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 184.94  on 154  degrees of freedom
## Residual deviance: 178.24  on 150  degrees of freedom
## AIC: 188.24
##
## Number of Fisher Scoring iterations: 4
```

Error bar calculations

During the analysis, we faced a problem when estimating the standard error for the overall referral rate. Note that our response is binary, we deal with the proportion problem as below:

The random variable P' is the sample proportion:

$$P' = \frac{X}{n}$$

where X is the random variable for the number of acceptance, n is the sample size.

The standard deviation is found to be

$$\sigma_{p'} = \sqrt{\frac{p(1-p)}{n}}$$

where p is the probability of acceptance, p' is the sample proportion of acceptance, and n is the size of the sample.

Therefore, the confidence interval for a population proportion become as

$$p = p' \pm \left[Z_{\alpha/2} \sqrt{\frac{p'(1-p')}{n}} \right]$$

where $Z_{\alpha/2}$ is set according to our desired degree of confidence and $\sqrt{\frac{p(1-p)}{n}}$ is the standard deviation of the sampling distribution.