Alex Jones

Molstad

STA4504

2/20/2022

STA4504 Exam 1

1.

a. Using the formula in the exam:

$$\text{RR} = \frac{\text{P}(\text{Recovered} = \text{yes} \mid \text{Treatment} = \text{B}, \text{Mutation} = \text{yes})}{\text{P}(\text{Recovered} = \text{yes} \mid \text{Treatment} = \text{A}, \text{Mutation} = \text{yes})}$$
$$= \frac{225/(225+195)}{197/(197+223)}$$
$$= 1.136$$

We see the relative risk between Treatment B and Treatment A when controlling for mutation is 1.136, meaning that those with Treatment B have a 13.6% better chance of recovery than those with Treatment A.

b. Using the formula for constructing a 95% Wald confidence interval in difference of proportions:

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{0.975}\widehat{SE}$$
$$\text{where } \widehat{SE} = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_{1+}} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_{2+}}}$$

And plugging in the values found from the table:

$$\left(\frac{197}{420} - \frac{301}{420}\right) \pm 1.96\sqrt{\frac{0.469(0.531)}{420} + \frac{0.717(0.283)}{420}}$$

We find our interval to be (-0.312, -0.183). As the interval does not contain 0, we can say at $\alpha = 0.05$ that we have convincing evidence that there is a difference in proportion for recovery between patients with or without the gene mutation, specifically that the presence of the genetic mutation strongly decreases the chance of recovery.

c. Using the formula from above, and plugging in the new values:

$$\left(\frac{225}{420} - \frac{242}{420}\right) \pm 1.96\sqrt{\frac{0.536(0.464)}{420} + \frac{0.576(0.424)}{420}}$$

Our new interval is (-0.108, 0.0267). As this does contain 0, we do not have enough evidence to say there is a difference in true proportion of recoveries, which when contrasted with the interval found in part (b), shows that the genetic mutation specifically decreases effectiveness in Treatment A, instead of decreasing effectiveness across the board.

d. Using the formula for finding the 95% Wald Interval for the log-odds ratio:

$$\log(\hat{\theta}) \pm z_{0.975}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}$$

$$\text{where } \hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

And plugging our values from the table in:

$$\log\left(\frac{197(195)}{225(223)}\right) \pm 1.96\sqrt{\frac{1}{197}+\frac{1}{223}+\frac{1}{225}+\frac{1}{195}}$$

We find our log-interval to be (-0.538, 0.004). Further exponentiating this, we get the odds-ratio interval to be (0.584, 1.004). This interval catches the "no-effect value" of 1, so at $\alpha = 0.05$ we do not have enough evidence to conclude that there truly is a difference, however since the interval barely catches 1 and is largely under that value, there is some evidence to believe that with the genetic mutation, the odds of recovery with Treatment B are higher than that those with Treatment A.

e. No, since the researchers collected equal numbers of cases with and without gene mutations, as well as equal cases for treatments within that factor, we do not have a valid way of computing a difference in proportion of recoveries between mutations. However, we can construct an estimate for odds ratio if needed.

f. With the null hypothesis $H_0: \mu_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j}$; for all i, j, and the alternative hypothesis $H_A: \mu_{ij} \neq n\hat{\pi}_{i+}\hat{\pi}_{+j}$; for some i, j, we will be conducting a Pearson's chi-squared test for independence at $\alpha = 0.05$. Importing the table in R, and using the `chisq.test()` function, we find that that the test statistic is equal to 42.67, which under a null distribution of chi-squared with 1df, leads to a p-value of 6.48e-11. This leads us to reject the null hypothesis, showing that we have convincing evidence to conclude that gene mutation and recovery are NOT independent.

g. Again, we will use a Pearson's chi-squared test, but instead of using the table generated from assuming independence, we will instead use the following expected frequencies (drawn from the probabilities in the null hypothesis as well as the total sample size):

X=MUTATION\Y=RECOVERY	**YES**	**NO**
YES	107.8	539
NO	539	970.2

Then, hand-calculating the test statistic this time:

$$= \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

$$= \frac{(80-107.8)^2}{107.8} + \frac{(771-539)^2}{539} + \frac{(261-539)^2}{539} + \frac{(1044-970.2)^2}{970.2}$$

Our test statistic is computed to be 256.026. Using R, we find the corresponding p-value (on a chi-squared distribution with 1df) to be 1.26e-57. With this, we reject the null hypothesis: we have convincing evidence to conclude that the at least one of the probabilities stated in the null hypothesis is not correct.

h. Using the formula for standardized residuals:

$$r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-\hat{\mu}_{i+})(1-\hat{\mu}_{+j})}}$$

And plugging them in for each cell (NOTE: the computed row and column probabilities were derived from the marginal probabilities found in the table describing the null hypothesis. I have reservations as to whether this is valid because the probabilities in the table itself are not independent e.g., $0.05 \neq 0.3 * 0.3$):

$$r_{11} = \frac{80-107.8}{\sqrt{107.8(1-0.3)(1-0.3)}}$$

$$r_{21} = \frac{771-539}{\sqrt{539(1-0.7)(1-0.3)}}$$

$$r_{12} = \frac{261-539}{\sqrt{539(1-0.3)(1-0.7)}}$$

$$r_{21} = \frac{1044-970.2}{\sqrt{970.2(1-0.7)(1-0.7)}}$$

The standardized residual values are as follows, organized in the below table:

X=MUTATION\Y=RECOVERY	**YES**	**NO**
YES	-3.825	-26.130
NO	21.806	7.898

We see that the observed counts for the cases where a patient did not have the mutation and recovered were much higher than expected, and the observed counts where a patient did have the mutation and not recover much lower. This seems to signify that patients with the genetic mutation fare much worse than expected and patients without the mutation fare much better.

i. One thing that wasn't looked at earlier in the problem was the odds ratio between Treatment A and Treatment B when the genetic mutation was NOT present (the inverse was looked at in part (d)). Using the formula for the odds ratio interval:

$$\log\left(\frac{301(178)}{242(119)}\right) \pm 1.96\sqrt{\frac{1}{301}+\frac{1}{178}+\frac{1}{242}+\frac{1}{119}}$$

The odds ratio interval is found to be (1.396, 2.480), indicating Treatment A is significantly better for patients without the genetic mutation. With this information, as well as the random sample of patients with the disease showing that $\frac{80+261}{2156} \approx 16\%$ of the patients have the genetic mutation. I would go with Treatment A over Treatment B. We didn't find convincing evidence to conclude that Treatment B was significantly better than treatment A even when patients had the mutation, while the inverse was found for patients without the mutation, and the low population of patients with the genetic mutation leads to Treatment A being the better option.

2.

a. Importing the data and producing a logistic regression model using the required coefficients, we find $\hat{\beta}_3 = 0.412$ with $\widehat{\text{SE}}_{\beta_3} = 0.414$. Using the formula for the 95% Wald confidence interval:

$$\hat{\beta}_3 \pm z_{0.975}\widehat{SE}_{\beta_3}$$

The interval is found to be (-0.400, 1.224). This interval captures 0, therefore we do not have enough evidence to conclude that the inclusion of $\hat{\beta}_3$ is necessary to produce an accurate model, and in fear of overfitting the model, I would recommend dropping it from the regression model.

b. Since replacing x_1 with $1 - x_1$ is a transformation that replaces all 0s with 1s and vice versa, the coefficient estimates associated with x_1 would be inverted, and with the following output:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.11486    0.19557  -0.587  0.55699
x1          -0.62524    0.22506  -2.778  0.00547 **
x2           0.71748    0.32502   2.208  0.02728 *
x3           0.08182    0.18539   0.441  0.65897
x1:x2        0.41228    0.41437   0.995  0.31975
---
```

The new model comes out to be:

$$\text{logit}\big(\pi(x)\big) = -0.115 + 0.652x_1 + 0.717x_2 - 0.412x_1x_2 + 0.082x_3$$

c. I returned to this question after completing the rest of Question 2, so for the goodness-of-fit test I will be using the data from the grouped model created in part (f). Using the summary output from the grouped model (again shown later in the question), we get a test statistic of 3.9907, which when following a chi-squared distribution with 3df, yields a p-value of 0.262. At $\alpha = 0.05$, we fail to reject the null hypothesis: we do not have convincing evidence to conclude that our model is fit well to the data.

d. The saturated model is fitted in such a way that each sample is given a coefficient. Thus, the model would look like this:

$$\text{logit}\big(\pi(x)\big) = 1 - \sum_{n=2}^{500} 1 * \mathrm{I}_{y_n=0}$$

$$\text{Where I} = \begin{cases} 1 & \text{if } y_n = 0 \\ 0 & \text{otherwise} \end{cases}$$

To dispel any confusion as to what is written, the function of y_i looks to see if the sample data for y_i is zero, and if it is, subtracts from the intercept, which is y_1.

e. Because, in the ungrouped model, y can only take values 0 or 1, the likelihood function of the ungrouped data is a linear model. Contrast this with the grouped data model, where y can take up to n_j values, a polynomial function is generally used for the likelihood. Below is an example of the difference in functions, where the ungrouped likelihood takes value $Y_{ij} = 1$, and the grouped likelihood takes value $Y_1 = 67$ (I used 67 because that's n_1 from the grouped model, making the proportion 1).

f. The relevant data for the grouped data model is shown below:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.11486    0.19557  -0.587  0.55699
x1          -0.62524    0.22506  -2.778  0.00547 **
x2           0.71748    0.32502   2.208  0.02728 *
x3           0.08182    0.18539   0.441  0.65897
x1:x2        0.41228    0.41437   0.995  0.31975
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35.8084  on 7  degrees of freedom
Residual deviance:  3.9907  on 3  degrees of freedom
AIC: 49.284
```

i. Fitting a logistic regression model onto the grouped data, we come up with the same maximum likelihood coefficients. This is because the likelihood function for both the ungrouped and grouped data end up with the same maximum likelihood estimator.

ii. However, the deviances are much different than the ungrouped model. Since the grouped data's saturated model only has seven coefficients and the ungrouped data's saturated model is fit to every single data point, the deviance in the grouped model is expected to be much lower.

g. To conduct a likelihood ratio test for $H_0 : \beta_3 = \beta_4 = 0$, a model without those terms must be fitted. After that, by comparing deviances we come up with a test statistic that follows a chi-squared distribution with 2df. With a test statistic of 1.204 for both the grouped and ungrouped models, we have a p-value of 0.5476. With such a high p-value, at $\alpha = 0.05$ we fail to reject the null hypothesis: we do not have enough evidence to conclude that β_3 or β_4 are necessary inclusions in the regression model.

3.

a. To test for marginal association, we will use the likelihood ratio test using a logit regression model with just SNPcat as a variable:

$$\text{logit}\big(\pi(x)\big) = 0.01645 + 0.638(\text{SNPcat1}) + 0.821(\text{SNPcat2})$$

We will be testing the null hypothesis $H_0: \beta_1 = \beta_2 = 0$ versus the alternative $H_A: \beta_1 or\ \beta_2 \neq 0$, where β_i represents the change in odds ratio of cancer from the presence of 1 or 2 of the reference alleles. Comparing deviances with the null model, we compute a test statistic of 32.032, which, following a chi-squared distribution with 2df, leads to a p-value of 1.107e-07. At $\alpha = 0.05$, we reject the null hypothesis: we have convincing evidence to conclude there exists some association between cancer and presence of the reference alleles.

b. For questions 3(b)-3(g), unless another model is stated to have been used, we will be using the summary output for the logistic regression model using SNPcat, smoking, and age, shown below:

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.061461   0.306277  -0.201   0.8410
as.factor(SNPcat)1  0.674276   0.162442   4.151 3.31e-05 ***
as.factor(SNPcat)2  0.837808   0.197916   4.233 2.30e-05 ***
age                -0.001812   0.003914  -0.463   0.6434
smoking             0.007072   0.002032   3.481   0.0005 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1650.2  on 1200  degrees of freedom
Residual deviance: 1605.3  on 1196  degrees of freedom
AIC: 1615.3
```

Conducting a Wald test for significance of smoking when conditioned upon Age and SNPcat, we will be using the model:

$$\text{logit}\big(\pi(x)\big) = -0.061 + 0.674(\text{SNPcat1}) + 0.838(\text{SNPcat2}) - 0.0018(\text{age}) + 0.0071(\text{smoking})$$

The reason I will be using this model is that SNPcat and age have been accounted for in the model, and the test will see if smoking still needs to be accounted in this model. This means the null hypothesis will be $H_0: \beta_4 = 0$, where β_4 indicates the change in odds ratio corresponding to a unit increase in smoking, and the alternative hypothesis will be $H_A: \beta_4 \neq 0$. Using the summary output's coefficient for the smoking variable as well as its standard error, we find the test statistic equivalent to be 3.481. With the null distribution being normal with mean 0 and variance 1, the corresponding p-value is 0.005, which at $\alpha = 0.05$ allows us to reject the null hypothesis. We have convincing evidence to conclude that the smoking variable does change the log-odds ratio for cancer.

c. Using the formula for producing a 95% confidence interval for the predicted value:

$$(\hat{\alpha} + \beta_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \beta_4 x_4) \pm z_{0.975}\widehat{\text{SE}}_{fit}$$

(note: the standard error is not simply the sums of the standard errors of each coefficient, covariances must also be accounted for)

And using the `predict()` function in R to find the point-estimate and the fitted standard error, we find the interval for the log-odds ratio to be (-03.910, -0.022). Noting the fact that log-odds can be converted into probability using

$$\pi(x) = \frac{\exp\,(\text{logit(x)})}{1 + \exp\,(\text{logit(x)})}$$

The values from the interval are plugged in and we find our interval for probability of cancer given the parameters to be (0.403, 0.495).

d. No, since smoking and age are both count data, a goodness of fit test can't be used for them.

e. As β_3 can be interpreted as the unit change in log-odds ratio, we will instead exponentiate $10\beta_3$:

$$\exp\big(10(-0.001812)\big)$$

And come up with the value 0.982. This indicates the odds ratio decreases to about 98.2% of its original odds when increasing age by 10. Since the logit regression model is based off odds ratios, we can expect the same decrease in odds for all x, $x + 10$.

f. The relevant probit regression model is shown below:

$$\pi(x) = \Phi[-0.024 + 0.416(\text{SNPcat1}) + 0.520(\text{SNPcat2}) - 0.00127(\text{age}) + 0.00429(\text{smoking})]$$

Comparing the difference in odds ratios between age 60 and age 50, we find the odds-ratio to be 0.979. This however, unlike the logit model, will change depending on what value x, $x + 10$ you use as the probit model is based on the normal distribution and not the odds ratio.

g.

i. $\text{logit}\big(\pi(x)\big) = -0.061 + 0.674(\text{SNPcat1}) + 0.838(\text{SNPcat2}) - 0.0018(\text{age}) + 0.0071(\text{smoking})$

ii. $\text{logit}\big(\pi(x)\big) = -0.034 + 0.482(\text{SNPnum}) - 0.0019(\text{age}) + 0.0071(\text{smoking})$

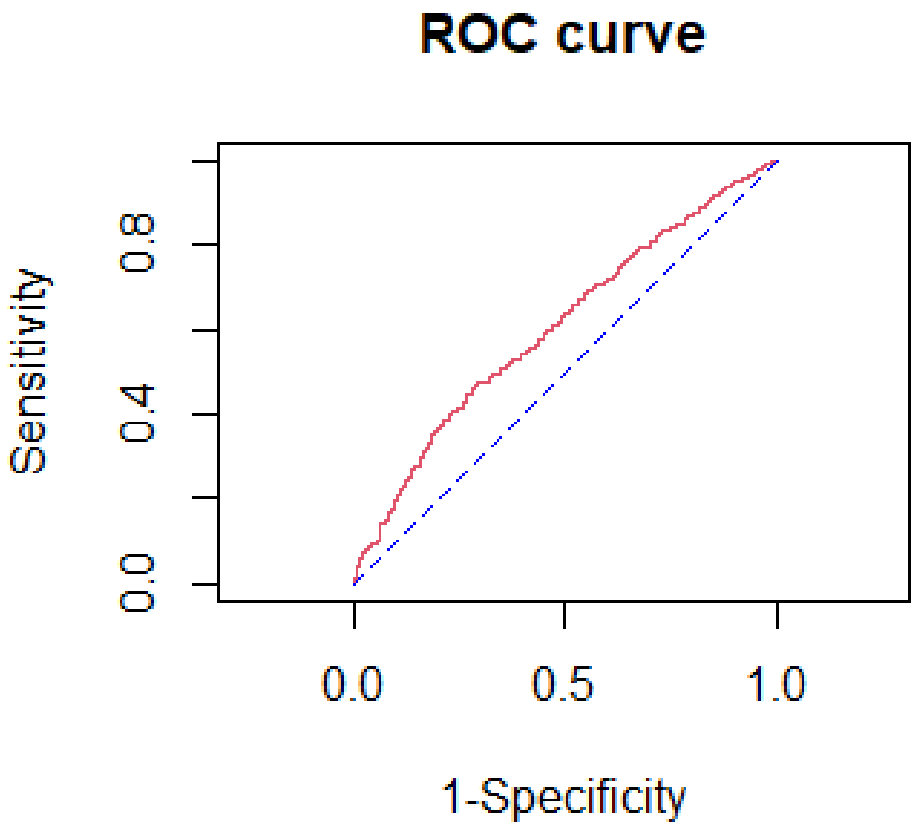
To make model (i) look similar to model (ii), the SNPcat1 and SNPcat2 coefficients would need to be made equivalent in order to correspond to the same unit increase as exhibited in model (ii).

h.

MODEL	AIC	BIC
SNPcat	1624.149	1639.422
SNPcat, age	1625.998	1646.361
SNPcat, age, smoking	1615.255	1640.709
SNPnum	1624.574	1634.756
SNPnum, age	1626.392	1641.665
SNPnum, age, smoking	1615.344	1635.708

Looking at both the AIC and BIC values, I would choose the last model even though it's not the lowest in either category. The reason this one is the best is because it has nearly the same predictive power in terms of AIC as the "SNPcat, age, smoking" model but because it has one less parameter, and the BIC is much better, to the point where it's almost as low as the "SNPnum" model.

i. The ROC Curve for the selected model is plotted on the next page:



This model predicts better than just random guessing. This is evidenced by the fact the red curve (indicating the regression model) is above the blue dotted line, which depicts the curve if only random guessing was involved. This is also evidenced by the curve's Concordance Index equals 0.606, which is stronger than the 0.5 that would occur if random guessing had happened.

j. Using code in R to calculate CCR for $\pi_0 = 0.5$, we find the classification rate to be equal to 0.578. Since the value is fairly close to 0.5 (the CCR if we were to randomly guess), I am not confident that this estimate would work on a new dataset, considering the fact that despite possibly being overfit with 3 parameters, it is not that much greater of an estimator than a coin toss.