Homework 3

Denis Ostroushko

Probem 1

Logistic regression model for fitted data is given by:

$$logit(P(Cancer = Yes)) = -7 + 0.1 * A + 1.2 * S + 0.3 * R + 0.2 * R * S$$

YS conditional odds ratio equation

Conditional YS odds ratio is presented when we compare $R = 1$ to $R = 0$ and let S be a variable in the resulting odds ratio. Then, varying levels of smoking will further change odds ratio for $R = 1$ vs $R = 0$.

$$OR(R|S = s) = \frac{\frac{P(R=1|S=s)}{1-P(R=1|S=s)}}{\frac{P(R=0|S=s)}{1-P(R=0|S=s)}} =$$

Odds ratio for both numerator and denominator simplify to a single exponential term. We hold A constant while adjusting for it in our comparison. We let S = s be an arbitrary value of S that takes on value 0 or 1.

$$\frac{exp(-7 + 0.1 * A + 1.2 * s + 0.3 * 1 + 0.2 * s * 1)}{exp(-7 + 0.1 * A + 1.2 * s + 0.3 * 0 + 0.2 * s * 0)} =$$

$$exp(0.3 + 0.2 * s)$$

This odds ratio is the compares the effects of race on the likelihood of having cancer, while adjusting for smoking. For black smokers, we have the highest chance of getting cancer, and white non-smokers have the lowest chance of getting cancer.

More precisely, black non-smokers are $exp(0.3) = 1.3499$ times more likely to have cancer, while black smokers are $exp(0.3 + 0.2) = 1.6487$ times more likely to have cancer, after adjusting for other variables.

YR conditional odds ratio equation

Conditional YR odds ratio is presented when we compare $S = 1$ to $S = 0$ and let R be a variable in the resulting odds ratio. Then, varying levels of smoking will further change odds ratio for $S = 1$ vs $S = 0$.

$$OR(S|R = r) = \frac{\frac{P(S=1|R=r)}{1-P(S=1|R=r)}}{\frac{P(S=0|R=r)}{1-P(S=0|R=r)}} =$$

$$= \frac{exp(-7 + 0.1 * A + 1.2 + 0.3 * r + 0.2 * 1 * r)}{exp(-7 + 0.1 * A + 1.2 * 0 + 0.3 * r + 0.2 * 0 * r)}$$

$$exp(1.2 + 0.2 * r)$$

So, smokers are $exp(1.2) = 3.3201$ times more likely to have cancer when compared with non-smokers, after adjusting for other variables. Additionally, black smokers are $exp(1.2 + 0.2) = 4.0552$ times more likely to have cancer, after adjusting for other variables.

MORE TO FINISH THE PROBLEM

Problem 2

Stage 3 model summary table given in Table 1

Table 1: Summary of the model

	Estimate	Std..Error	z.value	Pr...z..
(Intercept)	0.3908113	0.0845813	4.620538	0.0000038
Eyes	-2.3960414	0.3878916	-6.177090	0.0000000
Pyes	-1.0994964	0.1786745	-6.153627	0.0000000
GMale	0.3088840	0.1458203	2.118252	0.0341538
Eyes:Pyes	1.7998744	0.5129536	3.508844	0.0004501

change all E, P, G,etc… to their real names for easier reading

Effect of G

interpret effects of 0.308884 and use 0.1458203 to get confidence intervals if independent then their interaction must be non-significant

According to comments from TA we also need to state every

test it:

1. Null hypothesis: $H_0 : \hat{\beta}_G = 0$

1.1 Mull hyp in english

2. Alternative hypothesis: $H_a : \hat{\beta}_G \neq 0$

2.1 Alt hyp in english

3. Z statistic: $(\frac{\hat{\beta}-0}{se(\hat{\beta})}) = 2.1183$
4. P-value: 0.0342
5. Conclusion: There is enough statistical evidence to conclude that the effect

Independence of E and P

if independent then their interaction must be non-significant

test it:

1. Null hypothesis: $H_0 : \hat{\beta}_{E\ and\ P} = 0$
2. Alternative hypothesis: $H_a : \hat{\beta}_{E\ and\ P} \neq 0$
3. Z statistic: $(\frac{\hat{\beta}-0}{se(\hat{\beta})}) = 3.5088$
4. P-value: 5×10^{-4}
5. Conclusion: Effects of E and P are not independent of each, as evidenced by the low p-value and big z-statistic. Therefore, we can conclude that effects of variable E have varying effects on the outcome M, depending on the levels of variable P, after adjusting for other variables.

Table 2: Obseved rates of leukemia by radiation exposure group

Radiation Level	Midpoint	N Observations	Leukemia Proportion
0	0.0	391	3.32%
1to9	5.0	205	2.44%
10to49	29.5	156	3.21%
50to99	74.5	50	6%
100to199	149.5	35	11.43%
200plus	249.5	51	35.29%

Problem 3

(i)

Before fitting the model we can evaluate proportion of leukemia cases in the population for each level of radiation exposure. We can see that the presence of leukemia is similar for the lowest three levels, and starts to increase fast the higher exposure levels go.

We can fit the model with exposure levels as categorical predictor. The model will provide comparisons of each level to the baseline level, which had no exposure. We can then use pairwise comparisons to evalaute a large family of comparisons.

Radiation Level	Estimate	Odds Ratio	Std. Error	Z-value	P-value
Intercept	-3.370	0.034	0.282	-11.947	0.000
radiation1to9	-0.319	0.727	0.533	-0.598	0.550
radiation10to49	-0.038	0.963	0.535	-0.071	0.944
radiation50to99	0.618	1.856	0.659	0.939	0.348
radiation100to199	1.322	3.752	0.602	2.198	0.028
radiation200plus	2.764	15.860	0.407	6.795	0.000

- Looking at a small set of comparisons of each level of exposure to the referense level, people who got 100-200 and 200+ units of exposure are at a much higher chance of developing leukemia, compared with those who had no exposure to radiation.
- We are not able to conclude that the log odds of the event are different from the baseline for other levels of exposure.
- The odds of having leukemia for those who got between 100 and 199 units of radiation when compared with those who got no exposure are 3.7519 times higher, bounded by the (1.0105, 11.3511) 95% confidence interval.

Confidence interval does not include 1, having all values consistently above 1, therefore we have enough statistical evidence to conclude that those who got between 100 and 199 units of exposure are at a much higher risk of developing leukemia.

- Similarly, the odds of developing leukemia for those who got 200+ units of exposure compared to baseline level are 15.8601 times higher, bounded by the (7.2108, 35.917) 95% confidence interval.

In the next section we evaluate pairwise comparisons of levels of exposure.

(ii)

We can perform 15 pairwise comparisons for our model.

Family Wise Error Rate is given as $P(At\ least\ one\ false\ positive) = 1 - P(no\ false\ rejections) = 1 - 0.95^{15} = 0.5367$.

All possible comparisons are given below, the table is ordered from the lowest to the highest p-value:

Contrast	Estimate	Odds Ratio	P-value
0 - 200plus	-2.764	0.063	0.000
1to9 - 200plus	-3.083	0.046	0.000
10to49 - 200plus	-2.802	0.061	0.000
50to99 - 200plus	-2.145	0.117	0.016
100to199 - 200plus	-1.442	0.237	0.165
1to9 - 100to199	-1.641	0.194	0.174
0 - 100to199	-1.322	0.267	0.238
10to49 - 100to199	-1.360	0.257	0.374
1to9 - 50to99	-0.937	0.392	0.810
0 - 50to99	-0.618	0.539	0.937
50to99 - 100to199	-0.704	0.495	0.951
10to49 - 50to99	-0.656	0.519	0.952
0 - 1to9	0.319	1.376	0.991
1to9 - 10to49	-0.281	0.755	0.998
0 - 10to49	0.038	1.039	1.000

Without controling for the multiple comparison, or multiple testing, error, we have 4 statistically significant comparisons at the 0.05 significance level.

Holm-Bonferroni Adjustment

In order to control FWER for 15 comparisons we need to use the Holm - Bonferroni Stepdown Procedure. We present a table with contrasts one more time, including Holm adjusted p-values this time.

Contrast	Estimate	Odds Ratio	P-value	Holm Adjusted P-value
0 - 200plus	-2.764	0.063	0.000	0.000
1to9 - 200plus	-3.083	0.046	0.000	0.000
10to49 - 200plus	-2.802	0.061	0.000	0.000
50to99 - 200plus	-2.145	0.117	0.016	0.186
100to199 - 200plus	-1.442	0.237	0.165	1.000
1to9 - 100to199	-1.641	0.194	0.174	1.000
0 - 100to199	-1.322	0.267	0.238	1.000
10to49 - 100to199	-1.360	0.257	0.374	1.000
1to9 - 50to99	-0.937	0.392	0.810	1.000
0 - 50to99	-0.618	0.539	0.937	1.000
50to99 - 100to199	-0.704	0.495	0.951	1.000
10to49 - 50to99	-0.656	0.519	0.952	1.000
0 - 1to9	0.319	1.376	0.991	1.000
1to9 - 10to49	-0.281	0.755	0.998	1.000
0 - 10to49	0.038	1.039	1.000	1.000

Comments:

- How FWER adjustments prevents us from concluding one of the differeces is there

moving on

Benjamini-Hochberg Adjustment

In order to control FDR for 15 comparisons we need to use the Benjamini-Hochberg Adjustment Procedure. We present a table with contrasts one more time, including Holm adjusted p-values this time.

Comments:

- Confirm that obserbed differences withiut adjustments are kind of true differences.

move on

(iii)

First, let's fit the model that uses a midpoint radiation exposure as a continuous predictor of the log-odds of developing leukemia. Being able to estimate and extrapolate chances of

Contrast	Estimate	Odds Ratio	P-value	BH Adjusted P-value
0 - 200plus	-2.764	0.063	0.000	0.000
1to9 - 200plus	-3.083	0.046	0.000	0.000
10to49 - 200plus	-2.802	0.061	0.000	0.000
50to99 - 200plus	-2.145	0.117	0.016	0.058
100to199 - 200plus	-1.442	0.237	0.165	0.435
1to9 - 100to199	-1.641	0.194	0.174	0.435
0 - 100to199	-1.322	0.267	0.238	0.511
10to49 - 100to199	-1.360	0.257	0.374	0.702
1to9 - 50to99	-0.937	0.392	0.810	1.000
0 - 50to99	-0.618	0.539	0.937	1.000
50to99 - 100to199	-0.704	0.495	0.951	1.000
10to49 - 50to99	-0.656	0.519	0.952	1.000
0 - 1to9	0.319	1.376	0.991	1.000
1to9 - 10to49	-0.281	0.755	0.998	1.000
0 - 10to49	0.038	1.039	1.000	1.000

developing leukemia given some other levels of radiation exposure is a big advantage over a model with categorical predictors.

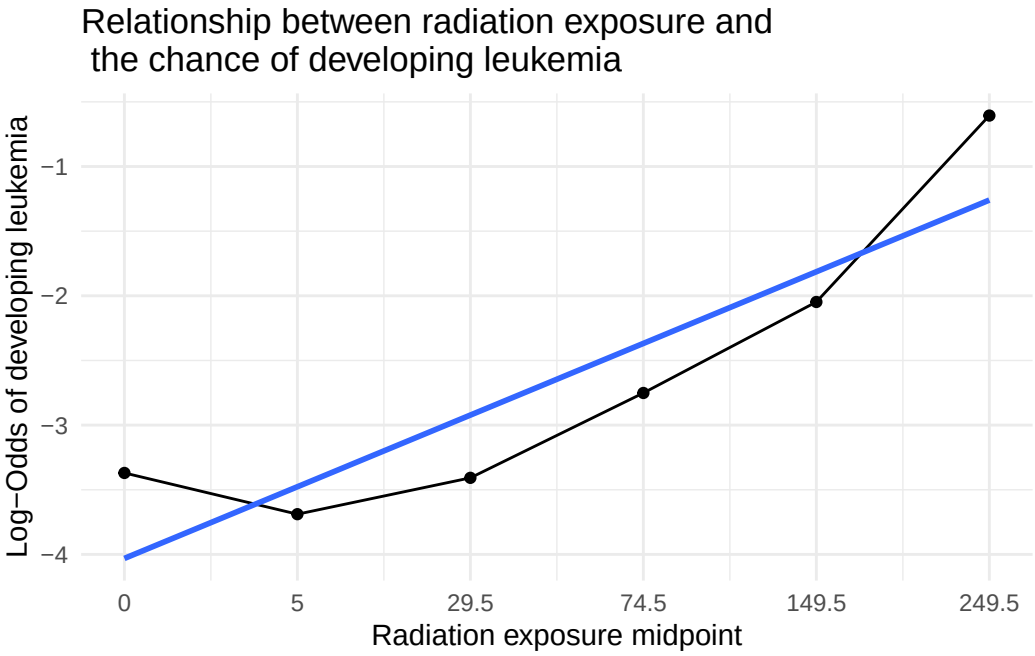
Model is given

Radiation Level	Estimate	Std. Error	Z-value	P-value
Intercept	-3.566	0.212	-16.800	0
radiation_midpoint	0.012	0.001	7.819	0

- We can see that the p-value is low and Z-statistic value is large, so we have enough statistical evidence that radiation exposure is linearly related to the chance of developing leukemia. As radiation exposure levels increase, the chance gets higher.

We should use a linear model over a categorical model for a number of reasons. First, as stated above, is the ability to extrapolate the odds of developing leukemia at levels other than those presented in the data, but without going too far outside of the model scope. For example, we should not use this model to estimate the proportion of leukemia among other levels at 500 units of radiation exposure.

Another reason is the relationship between midpoints and log odds of developing leukemia. I took observed proportions of leukemia from the data and transformed them to the log-odds scale.



- As we can see, there a clear linear trend that can be modeled using a regression approach with a continuous predictor.

Deviance Comparison

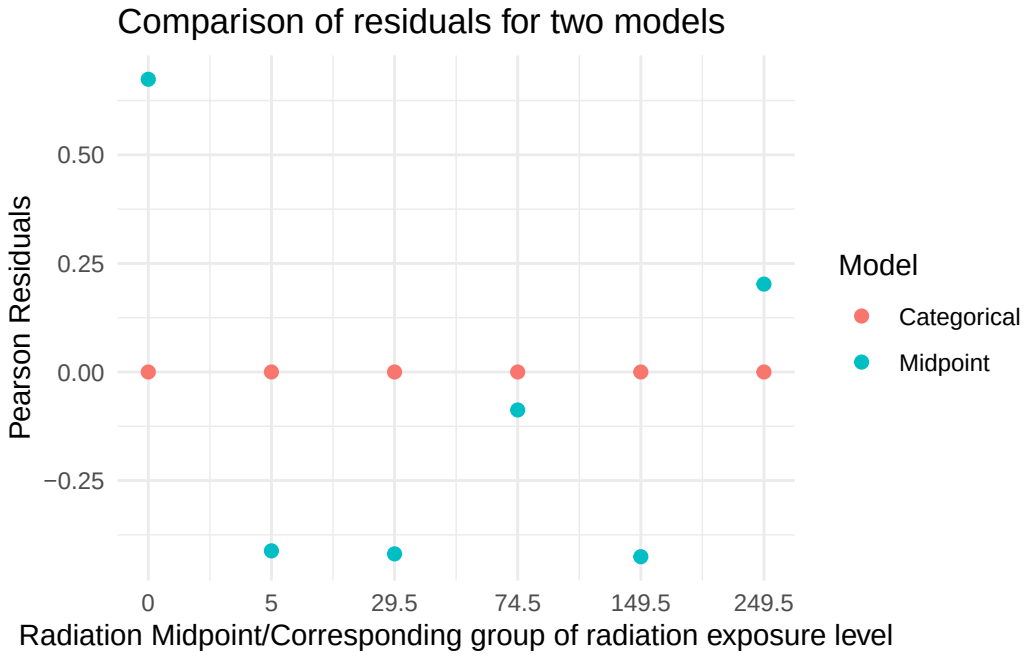
We can also compare the two models using a likelihood ratio test.

model	Resid..Df	Resid..Dev	Df	Deviance	Pr..Chi.
Categorical	0	0.00	NA	NA	NA
Midpoint	4	1.03	-4	-1.03	0.91

- Evidently, a model with a continuous predictor has lower deviance, however, the drop in deviance is not statistically significant, so we can't conclude that using a continuous predictor lower deviance drastically.

Residual Plots

We can finally compare residuals for the two models. I use `R` to calculate and store pearson residuals for two models, and display residual for each model at each respective level. Note that I used radiation midpoint on the x-axis. However, there is a one-to-one relationship between bucketed levels of ration exposure and a midpoint for the bucket. Therefore, it is appropriate to use midpoint as a label.



- As we can see, a model with categorical predictors essentially does not produce any residuals. This is reasonable because we fit a model using 6 observations, and fit one proportion per observation.
- However, a model with a midpoint continuous predictor shows behavior that is more expected of a regression model. We can see how our fit tends to underestimate log-odds of getting leukemia.
- Indeed, we saw of the figure above that, perhaps, a fit with the second order exponential may be more appropriate.

(iv)

Given a logistic regression model with an intercept and one predictor, odds ratio for a one unit change in predictor X are given by $e^{\hat{\beta}_1*((x+1)-x)} = e^{\hat{\beta}_1}$

Therefore, for two values of X that are more than one unit apart, denoted as W_1 and W_2 are given by $e^{\hat{\beta}_1*(W_1-W_2)}$

Using a full notation and including intercepts we have $e^{\hat{\beta}_0-\hat{\beta}_0+\hat{\beta}_1*(W_1-W_2)}$

Now we can take a ratio of odds ratios, keeping intercepts in the notation. We denote levels of X from the first odds ratio as W_1 and W_2, and levels of X from the second odds ratio as Z_1 and Z_2. In each case we compare odds for level with subscript 1 to level with subscript 2.

We then take a ratio of odds ratio for $W's$ to odds ratio for $Z's$. Estimator is given below:

$$e^{\hat{\beta}_0(1-1-1+1)+\hat{\beta}_1*(W_1-W_2-Z_1+Z_2)}$$

We can see that algebraic signs follow a patter, and we have one real number as a multiplier for each model parameters. Note that a real number for $\hat{\beta}_0$ is zero, however, it is convenient to keep it there as for the derivation of a matrix form calculation.

Therefore, as per Jared's tip, a ratio of odds ratios is a function of four values of X, and can be represented as

$$e^{\mathbf{a}^T\hat{\beta}}$$

, where $\mathbf{a}^T = [1-1-1+1, W_1 - W_2 - Z_1 + Z_2]$

So, the ratio of the odds of having leukemia comparing a radiation level of '100 to 199' and a radiation level of '50 to 99' is given by $e^{\hat{\beta}_0(1-1-1+1)+\hat{\beta}_1*(100-199-50+99)} = 0.559218$

(v)

In order to obtain a confidence interval for the ratio of odds ratio we can take two approaches:

1. Calculate confidence interval for $\mathbf{a}^T\hat{\beta}$, and then exponentiate the interval
2. Use the delta method to calculate $Var(e^{\mathbf{a}^T\hat{\beta}})$ and then calculate the 95% confidence interval using a standard error of the odds scale directly.

Logistic regression model estimates are asymptotically normally distributed as a result of MLE estimation, so we can directly apply the delta method, and obtain another normally distributed random variable.

For my own reference, I will use both methods, and validate that the results indeed match.

Transformation of confidence interval bounds

Using Jared's tip, we can calculate $Var(\mathbf{a}^T\hat{\beta})$ directly by taking $a^T\mathbf{V}a$ where V is a variance-covariance matrix of the fitted logistic regression model. Variance-covariance estimates are given for model estimates including the intercept.

Using `R` output we estimate $Var(\mathbf{a}^T\hat{\beta}) = 0.005525$

Then, the 95% confidence interval on the original scale is $\mathbf{a}^T\hat{\beta} \pm 1.96 * \sqrt{Var(\mathbf{a}^T\hat{\beta})}$ = -0.581215 $\pm$1.96$*$ 0.07433

Taking exponential of interval end point given us a confidence interval for the ratio of odds ratios. The 95% confidence interval is (0.483404, 0.646923)

Delta method

Since a^T are constants, and $\widehat{\beta}$ is a random variable that contain a sampling distribution and variance, we define $g(\widehat{\beta}) = e^{a^T * \widehat{\beta}}$ in order to find $Var(e^{a^T * \widehat{\beta}})$

First step is to take a derivative with respect to $\widehat{\beta}$. I will use notation that is specific to our problem rather than a general form.

$\frac{d}{d\widehat{\beta}} g(\widehat{\beta}) = -50 * e^{-50 * \widehat{\beta}_1}$, therefore

$$Var(e^{a^T * \widehat{\beta}}) = Var(e^{-50 * \widehat{\beta}}) = (-50 * e^{-50 * \widehat{\beta}_1})^2 * Var(\widehat{\beta}) = 0.001728$$

So, we can use a ratio of odds ratio of on the odds scale and calculate confidence interval for this estimate using a square root of variance we just obtained.

So, the 95% confidence interval is given by: 0.559218 $\pm$1.96$*$ 0.041567, giving us (0.477747, 0.640689)

Conclusion

Both methods produce *almost* identical confidence intervals. An advantage of the delta method is knowing variance of the ratios of odds ratios, which gives us a standard error. Sometimes, it is easier to interpret and report this quantity to non-statisticians and other professionals.