

Homework 3

Denis Ostroushko

Problem 1

Logistic regression model for fitted data is given by:

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

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 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 is:

	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: $H_0 : \hat{\beta}_{E \text{ and } P} = 0$
2. Alternative hypothesis: $H_a : \hat{\beta}_{E \text{ and } P} \neq 0$
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.

Problem 3

(i)

INTERPRET ON ODDS SCALE, SO USE EXPOENENTIATED COEFFICIENTS

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

(ii)

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

Model is given

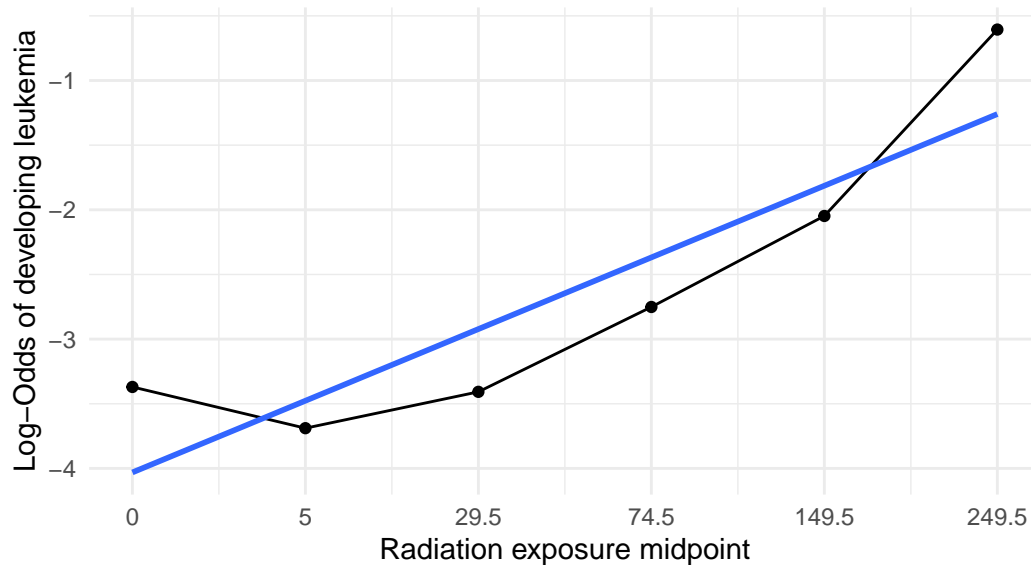
Radiation Level	Estimate	Std. Error	Z-value	P-value
Intercept	-3.566	0.212	-16.800	0
radiation_midpoint	0.012	0.001	7.819	0

- 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.

Relationship between radiation exposure and the chance of developing leukemia



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

Deviance Comparison

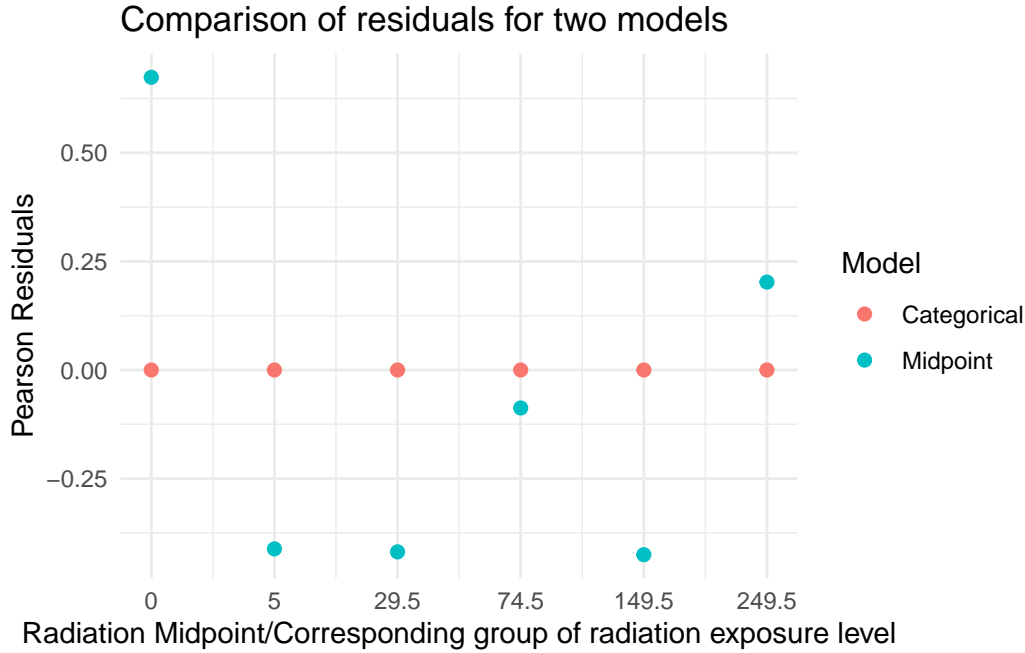
We can also compare the two models using a likelihood ratio test.

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 patten, 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 $\mathbf{a}^T \mathbf{V} \mathbf{a}$ 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 $\hat{\beta}$ is a random variable that contain a sampling distribution and variance, we define $g(\hat{\beta}) = e^{a^T * \hat{\beta}}$ in order to find $Var(e^{a^T * \hat{\beta}})$

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

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

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

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.