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

	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

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

test it:

1. Null hypothesis: $H_0 : \hat{\beta}_G = 0$
2. Alternative hypothesis: $H_a : \hat{\beta}_G \neq 0$

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.

Problem 3

(i)

```
Call:
glm(formula = cbind(leukemia = prob3$leukemia, other = prob3$other) ~ 
    prob3$radiation, family = binomial(link = "logit"))

Deviance Residuals: 
[1]  0  0  0  0  0  0

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -3.3699     0.2821 -11.947  < 2e-16 ***
prob3$radiation1to9      -0.3189     0.5334  -0.598   0.5499    
prob3$radiation10to49    -0.0379     0.5350  -0.071   0.9435    
prob3$radiation50to99     0.6184     0.6589   0.939   0.3480    
prob3$radiation100to199   1.3222     0.6015   2.198   0.0279 *  
prob3$radiation200plus    2.7638     0.4067   6.795 1.08e-11 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5.4351e+01  on 5  degrees of freedom
Residual deviance: 1.3323e-14  on 0  degrees of freedom
AIC: 33.665

Number of Fisher Scoring iterations: 4
```

INTERPRET ON ODDS SCALE, SO USE EXPOENENTIATED COEFFICIENTS

(iii)

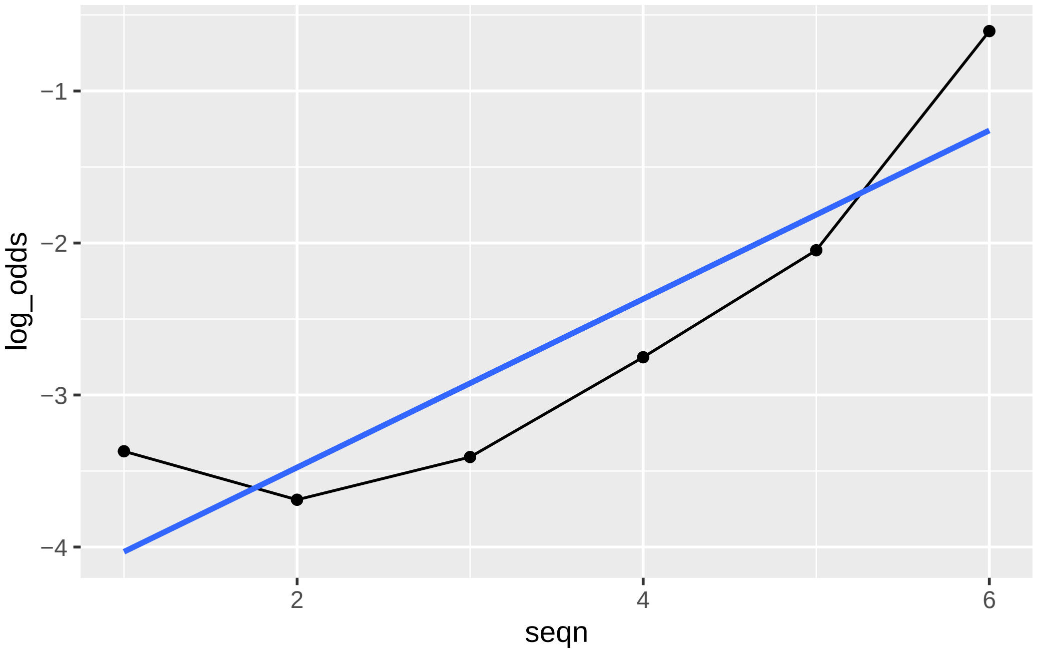
```
Call:
glm(formula = cbind(leukemia = prob3$leukemia, other = prob3$other) ~
    prob3$radiation_midpoint, family = binomial(link = "logit"))

Deviance Residuals:
       1         2         3         4         5         6
 0.67399  -0.41184  -0.41877  -0.08743  -0.42526   0.20237

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -3.565875   0.212254 -16.800  < 2e-16 ***
prob3$radiation_midpoint  0.011624   0.001487   7.819 5.31e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54.3509  on 5  degrees of freedom
Residual deviance:  1.0287  on 4  degrees of freedom
AIC: 26.694

Number of Fisher Scoring iterations: 4
```

COMPARE to model prob3_glm

Prefer this one:

becuase claerly lienar trend is better

Avoid multiple comparisons



(iv)

(v)

```
prob3$radiation_midpoint
                3.160739
```

```
Call:
glm(formula = cbind(leukemia = prob3$leukemia, other = prob3$other) ~
    prob3$radiation_midpoint, family = binomial(link = "logit"))

Deviance Residuals:
       1         2         3         4         5         6
 0.67399  -0.41184  -0.41877  -0.08743  -0.42526   0.20237

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -3.565875   0.212254 -16.800  < 2e-16 ***
```

```
prob3$radiation_midpoint  0.011624   0.001487   7.819 5.31e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54.3509  on 5  degrees of freedom
Residual deviance:  1.0287  on 4  degrees of freedom
AIC: 26.694

Number of Fisher Scoring iterations: 4


    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.003468 0.006062 0.006955 0.006951 0.007807 0.011623
```