Homework #3

March 15, 2017

1.

(a)
$$\hat{\beta}_0 = logit(\hat{P}(Y_i = 1|S_i = 0, P_i = 0)) = logit(20/50)$$
$$\hat{\beta}_0 + \hat{\beta}_1 = logit(\hat{P}(Y_i = 1|S_i = 1, P_i = 0)) = logit(30/50)$$
$$\hat{\beta}_0 + \hat{\beta}_2 = logit(\hat{P}(Y_i = 1|S_i = 0, P_i = 1)) = logit(10/50)$$
$$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = logit(\hat{P}(Y_i = 1|S_i = 1, P_i = 1)) = logit(32/50)$$

$$\hat{\beta}_0 = -0.4054651$$
$$\hat{\beta}_1 = 0.8109302$$
$$\hat{\beta}_2 = -0.9808293$$
$$\hat{\beta}_3 = 1.150728$$

(b) $\exp(\hat{\beta}_0)$: estimated odds of COPD for a non-smoker who does not live in highly polluted area.
$\exp(\hat{\beta}_2)$: estimated odds ratio of COPD comparing a subject living in highly polluted area and one living in not highly polluted area, when both of them are non-smokers.
$\exp(\hat{\beta}_3)$: Odds ratio of COPD between smoking status in highly polluted area is estimated to be $\exp(\hat{\beta}_3) = 3.16$ times higher than that in not highly polluted area.

(c) Full model: $logit(\pi_i) = \beta_0 + \beta_1 S_i + \beta_2 P_i + \beta_3 S_i P_i$. Deviance is 0.
Reduced model: $logit(\pi_i) = \beta_0 + \beta_1 S_i$. Deviance is 5.0013.
LRT test statistics: $5.0013 \sim \chi^2_2$, p-value is 0.08203166. Fail to reject H_0 at $\alpha = 0.05$.

2.

(a) If $\beta_1 = \beta_2$, then $\pi_1 = \pi_2$.
Then $\phi = 1$.

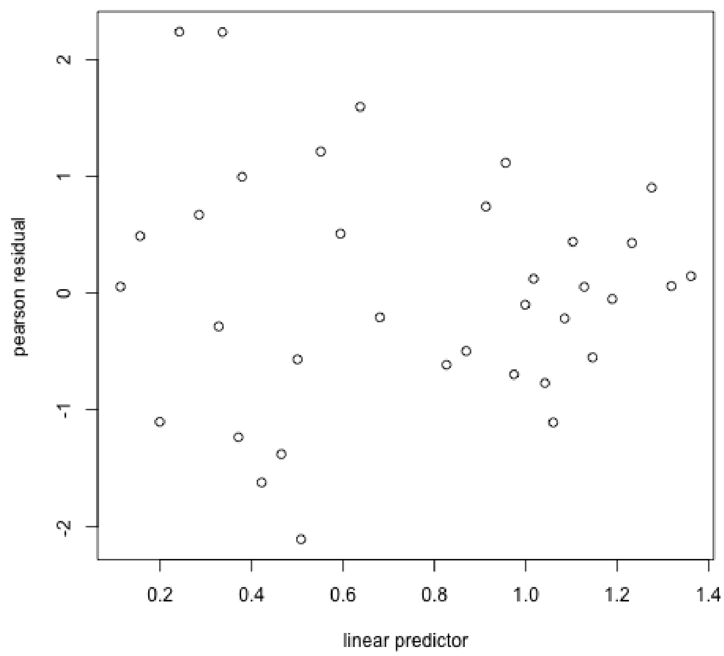
(b) $\phi_j = \frac{\pi_{1j}(1-\pi_{2j})}{\pi_{2j}(1-\pi_{1j})} = \exp(\alpha_1 + \beta_1 x_j)/\exp((\alpha_2 + \beta_2 x_j)) = \exp(\alpha_1 - \alpha_2)$
So $\log(\phi)$ is constant across tables.

3.

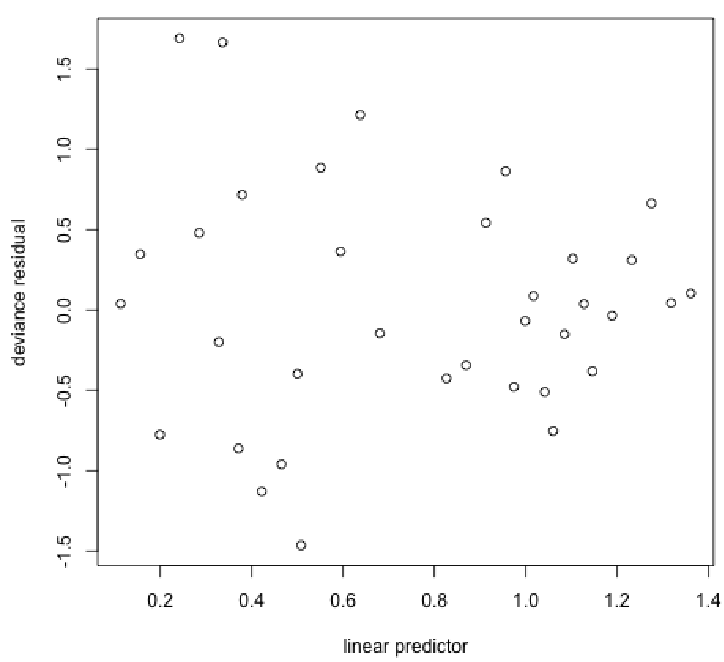
(a) $\hat{\beta}_0 = -0.66096$, estimated log odds of survival for students graduated from the science dept in 1900.
$\hat{\beta}_1 = 0.04302$, estimated difference in log adds ratio of survival per year increase in graduation, adjusting for other covariates.
$\hat{\beta}_2 = -0.86054$, estimated difference in log odds ratio of survival between ART and SCI students, adjusting for the year of graduation.

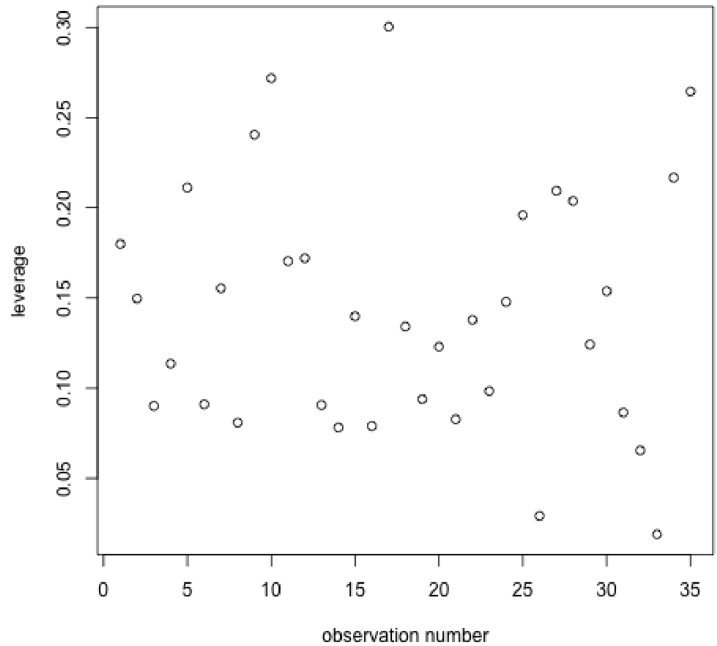
(b) i. residual:

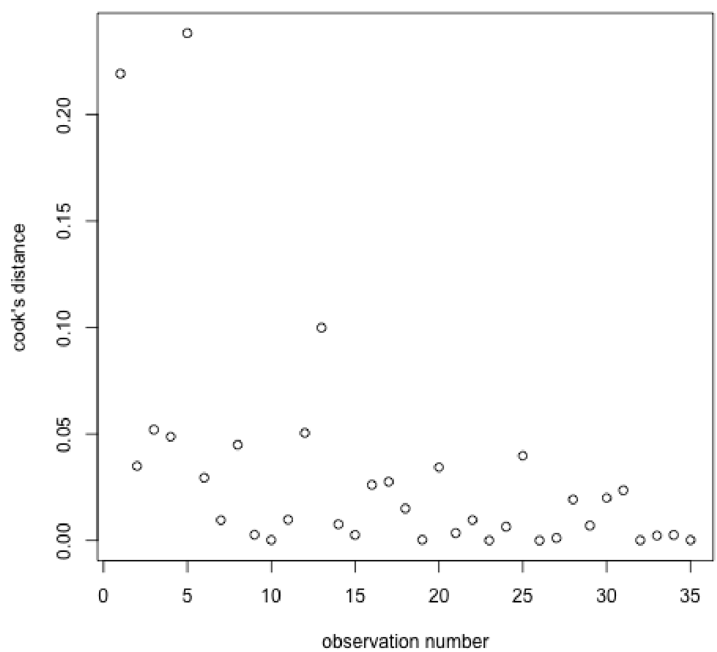
pearson residual



deviance residual







Residual plots show that although some observations have residuals $+/-2$ away from zero, there is no observation clearly separated from others. There are two observations with relatively large cooks distance, but their values are smaller than one. One observation (h=0.3003) has leverage $> 5/35*2 = 0.29$.

ii. VIF: VIF can be calculated using proc reg with a weight from proc genmod. $VIF_{year} = 1.04542$,
$VIF_{art} = 1.56467$,
$VIF_{med} = 1.58212$,

$VIF_{eng} = 1.36085$,
All VIFs are small or moderate, so we can conclude that there is no serious multi-collinearity problem.

iii. Pseudo R^2 (Cox & Snell): 0.0315
Max adjusted R^2: 0.1504

iv. HL Test
Null Model: The logistic regression model fits data well
Test statistic: 16.5094, DF:8, P-value: 0.0356

At level $\alpha = 0.05$, we can reject the null hypothesis. We can conclude that the logistic regression model does not fit the data well.

4.

a).

Score statistic : $U = \sum_{i=1}^{2} x_i (y_i - \mu_i)$

$i=1, \quad x_1 = (1, 0)^T, \quad \mu_1 = n_0 \cdot \exp(\alpha) / (1+\exp(\alpha))$

$i=2, \quad x_2 = (1, 1)^T, \quad \mu_2 = n_1 \cdot \exp(\alpha+\beta) / (1+\exp(\alpha+\beta))$

$$\therefore \quad 0 = U = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \left(n_{01} - n_0 \frac{\exp(\alpha)}{1+\exp(\alpha)} \right) + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \left(n_{11} - n_1 \frac{\exp(\alpha+\beta)}{1+\exp(\alpha+\beta)} \right)$$

$$\Leftrightarrow \begin{cases} 0 = n_{01} + n_{11} - n_0 \dfrac{\exp(\hat{\alpha})}{1+\exp(\hat{\alpha})} - n_1 \dfrac{\exp(\hat{\alpha}+\hat{\beta})}{1+\exp(\hat{\alpha}+\hat{\beta})} \\ 0 = n_{11} - n_1 \dfrac{\exp(\hat{\alpha}+\hat{\beta})}{1+\exp(\hat{\alpha}+\hat{\beta})} \end{cases}$$

$$\Leftrightarrow \begin{cases} \dfrac{\exp(\hat{\alpha}+\hat{\beta})}{1+\exp(\hat{\alpha}+\hat{\beta})} = \dfrac{n_{11}}{n_1} \Rightarrow \exp(\hat{\alpha}+\hat{\beta}) = \dfrac{n_{11}}{n_{10}} \\ \dfrac{\exp(\hat{\alpha})}{1+\exp(\hat{\alpha})} = \dfrac{n_{01}}{n_0} \Rightarrow \exp(\hat{\alpha}) = \dfrac{n_{01}}{n_{00}} \end{cases}$$

$$\therefore \quad \exp(\hat{\beta}) = \frac{\exp(\hat{\alpha}+\hat{\beta})}{\exp(\hat{\alpha})} = \frac{n_{00}\, n_{11}}{n_{10}\, n_{01}}$$

b)

$$\hat{\mathcal{I}} = \sum_{i=1}^{2} X_i X_i^T \, V(\hat{\mu}_i)$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} n_0 \frac{\exp(\hat{\alpha})}{(1+\exp(\hat{\alpha}))^2} + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} n_1 \frac{\exp(\hat{\alpha}+\hat{\beta})}{(1+\exp(\hat{\alpha}+\hat{\beta}))^2}$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cancel{n_0} \cdot \frac{n_{01}}{\cancel{n_0}} \cdot \frac{n_{00}}{n_0} + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \cancel{n_1} \frac{n_{10}}{\cancel{n_1}} \frac{n_{11}}{n_1}$$

$$= \begin{pmatrix} n_{01} n_{00}/n_0 + n_{10} \cdot n_{11}/n_1 & n_{10} n_{11}/n_1 \\ n_{10} n_{11}/n_1 & n_{10} n_{11}/n_1 \end{pmatrix}$$

The (2, 2) element of $\hat{J}^{-1}$ is

$$\frac{1}{n_{01} n_{00} n_{10} n_{11} / (n_0 n_1)} \cdot \left(\frac{n_{01} n_{00}}{n_0} + \frac{n_{10} n_{11}}{n_1} \right)$$

$$= \frac{n_1 (n_{01} n_{00}) + n_0 (n_{10} n_{11})}{n_{01} n_{00} n_{10} n_{11}}$$

$$= \frac{(n_{10} + n_{11}) n_{01} n_{00} + (n_{00} + n_{01}) n_{10} n_{11}}{n_{01} n_{00} n_{10} n_{11}}$$

$$= \frac{1}{n_{01}} + \frac{1}{n_{00}} + \frac{1}{n_{10}} + \frac{1}{n_{11}}$$

$$\therefore \hat{Var}(\hat{\beta}) = \frac{1}{n_{00}} + \frac{1}{n_{11}} + \frac{1}{n_{01}} + \frac{1}{n_{10}}$$

```
data copd;
input s p cases total;
datalines;
0 0 20 50
0 1 10 50
1 0 30 50
1 1 32 50
;
run;

proc genmod data=copd;
   model cases/total = s p s*p/dist=bin link=logit;
   contrast "Test" p 1, s*p 1;
run;

proc genmod data=copd;
   model cases/total = s /dist=bin link=logit;
run;

data HW33;
 infile "~/BIOSTAT651/Adelaide1.txt"  ;
 input YEAR DEPT $ SURVIVORS TOTAL;
 ART=(DEPT="ART");
 MED=(DEPT="MED");
 ENG=(DEPT="ENG");
 YEAR_1900 = YEAR - 1900;
run;

proc genmod data=HW33 plots=(RESCHI(XBETA) RESDEV(XBETA) LEVERAGE DOBS);
model SURVIVORS / TOTAL = YEAR_1900 ART MED ENG/ dist = bin link = logit;
    output out=Diagnostic1 XBETA=eta hesswgt=W Leverage=LEVERAGE RESCHI=RESCHI
    RESDEV=RESDEV STDRESCHI=STDRESCHI STDRESDEV=STDRESDEV COOKSD=COOKSD;
run;


proc logistic data=HW33;
model SURVIVORS / TOTAL = YEAR_1900 ART MED ENG/ Lackfit influence RSQ;
run;

/* calculate vif */
proc reg data=Diagnostic1;
        weight W;
        model SURVIVORS = YEAR_1900 ART MED ENG/ vif;
run;
```