Solution to HW3

Problem 2.33

(a) Let X be defendant's race (W/B), Y be verdict (death penalty or no), Z be victom's rate (W/B). Then 3 table is

$Z = W$:

		Y: D	$\bar{D}$
X	W	19	132
	B	11	52

$Z = B$:

		Y: D	$\bar{D}$
X	W	0	9
	B	6	97

(b) The conditional odds-ratio estimates are:

$$\widehat{\theta}_{XY(1)} = \frac{19 \times 52}{11 \times 132} = 0.68, \quad \widehat{\theta}_{XY(2)} = \frac{0.5 \times 97.5}{6.5 \times 9.5} = 0.79.$$

Given the victom's race is white, the odds of getting a death penalty for a white defendant is 0.68 times the odds of getting a death penalty for a black defendant. Given the victom's race is black, the odds of getting a death penalty for a white defendant is 0.79 times the odds of getting a death penalty for a black defendant. Therefore, given the victim's race (regardless whether it is white or black) a white defendant is less likely to receive a death penalty than a black defendant.

(c) The marginal XY table is:

		Y: D	$\bar{D}$
X	W	19	141
	B	17	149

The sample marginal odds ratio between defendant's race and a death penalty is

$$\widehat{\theta}_{XY} = \frac{19 \times 149}{17 \times 141} = 1.18.$$

Marginally, the odds of a white defendant receiving a death penalty is 1.18 times the odds of black defendant receiving a death penalty. Therefore, marginally, a white defendant is more likely to receive a death penalty than a black defendant.

The data exhibit Simpson's parodox since the marginal XY association and the conditional XY association given Z are in different directions. The reason is that X and Z are related, and Z and Y are also related.

Problem 2.35

One possible reason is that the age distributions between South Carolina and Maine are different. Specifically, there are more younger people in South Carolina than in Maine (Older peopler prefer to live in Maine?). We know that older age groups tend to have higher death rates. Therefore, relatively more young people in South Carolina will pull down the overall death rate in South Carolina, making its death rate lower than that of Maine.

Problem 2.39

a. True; b. True; c. False; d. True; e. False.

Problem 3.4

(a) The SAS code and the revalent output are:

```
data prob3_4;
  input alcohol malform count @@;
  datalines;
  0     1 48   0   0 17066
  0.5   1 38   0.5 0 14464
  1.5   1 5    1.5 0 788
  4     1 1    4   0 126
  7     1 0    7   0 37
;

data y1; set prob3_4;
  y=count;
  if malform=1;
run;

data y0; set prob3_4;
  y0=count;
  if malform=0;
run;

data new; merge y1 y0;
  n=y+y0;
run;

title "Problem 3.4(a)";
proc genmod;
  model y/n = alcohol / link=identity lrci;
run;
*****************************************************************************
              Analysis Of Maximum Likelihood Parameter Estimates

                                        Likelihood Ratio
                             Standard    95% Confidence          Wald
 Parameter  DF  Estimate        Error        Limits          Chi-Square  Pr > ChiSq

 Intercept   1    0.0026       0.0003    0.0020     0.0034        58.11      <.0001
 alcohol     1    0.0007       0.0007   -0.0004     0.0023         0.81      0.3677
 Scale       0    1.0000       0.0000    1.0000     1.0000
```

So the result is sensitive to this single malformation observation. For example, $\widehat{\beta}_1$ became 0.0007 from 0.0011. Based on the output, the estimated probability of malformation at alcohol level 0 is 0.0026 (compared to 0.00255); the estimated probability of malformation at alcohol level 7 is 0.0075 (compared to 0.010).

(b) Using the new score (0, 1, 2, 3, 4), we refit the linear probability model and the revalent output is

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.0026	0.0004	0.0020	0.0034	52.33	<.0001
score	1	0.0005	0.0005	-0.0003	0.0015	1.16	0.2822
Scale	0	1.0000	0.0000	1.0000	1.0000		

Base on this model, the estimated malformation probability at the lowest alcohol consumption level is 0.0026; the estimated malformation probability at the highest alcohol consumption level is $0.0026 + 4 \times 0.0005 = 0.0046$. Even though the estimated malformation probablities at the lowest alcohol level are basically the same, the estimated malformation probablities at the highest alcohol level are dramatically different. Therefore, the fit is also sentive to the choice of the score.

(c) The revalent SAS output is

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.9605	0.1154	-6.1930	-5.7397	2666.41	<.0001
alcohol	1	0.3166	0.1254	0.0187	0.5236	6.37	0.0116
Scale	0	1.0000	0.0000	1.0000	1.0000		

So the prediction equation is:

$$\log \frac{P(\text{malformation}|\text{alcohol})}{1 - P(\text{malformation}|\text{alcohol})} = -5.96 + 0.32\text{alcohol}.$$

With one unit increase in the alcohol consumption score, the odds of having malformation increases by $e^{0.32} - 1 = 0.37 = 37\%$.

Problem 3.5

We fit the linear probablity with those 3 sets of scores and obtained the following output:

Snoring and heart disease data using s1 with identity link

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square
Intercept	1	0.0176	0.0035	0.0108	0.0244	25.52
s1	1	0.0181	0.0026	0.0130	0.0232	48.82
Scale	0	1.0000	0.0000	1.0000	1.0000	

****

Snoring and heart disease data using s2 with identity link

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square
Intercept	1	0.0176	0.0035	0.0108	0.0244	25.52
s2	1	0.0362	0.0052	0.0261	0.0464	48.82
Scale	0	1.0000	0.0000	1.0000	1.0000	

****

Snoring and heart disease data using s3 with identity link

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square
Intercept	1	-0.0186	0.0073	-0.0329	-0.0044	6.57
s3	1	0.0362	0.0052	0.0261	0.0464	48.82
Scale	0	1.0000	0.0000	1.0000	1.0000	

The fitted values will be the same. For example, for individuals who occasionally snored, the estimated probabities of heart disease are: (1) $0.0176 + 2 \times 0.0181 = 0.0538$; (2) $0.0176 + 1 \times 0.0362 = 0.0538$; (3) $-0.0186 + 2 \times 0.0362 = 0.0538$.

Suppose the intercept and slope estimates in a GLM with x are $\widehat{\beta}_0$ and $\widehat{\beta}_1$. When we do a linear transformation of x and use $z = a + bx (b \neq 0)$ in the same GLM, then the intercept and slope estimates with z will be $\widehat{\beta}_0 - a\widehat{\beta}_1/b$, and $\widehat{\beta}_1/b$.

Problem 3.9

(a) The prediction equation is:

$$\text{logit}\{P(\text{have one travel card}|\text{income})\} = -3.5561 + 0.0532\text{income}.$$

(b) With 1 millions of lira ($\approx$ 500 Euros) increase in income, the odds of having a travel card increases by $e^{0.0532} - 1 = 5.5\%$. You may interpret $\widehat{\beta}$ using 2 millions of lira ($\approx$ 1,000 Euros) increase in income. Then the odds increases by 11.2%.

(c) When $\widehat{\pi} = 0.5$, the estimated logit is

$$\text{logit}(\widehat{\pi}) = \log\{\widehat{\pi}/(1 - \widehat{\pi})\} = \log\{0.5/(1 - 0.5)\} = 0.$$

Solving

$$\text{logit}\{P(\text{have one travel card}|\text{income})\} - 3.5561 + 0.0532\text{income} = 0$$

gives income = 3.5561/0.0532=66.84 million lira.