Solution to HW8

Problem 5.28

(a) Here $\pi_1 = 0.2, \pi_2 = 0.3$. Since we are using 90% CI, so the correspoding significance level is $\alpha = 0.1$. The power is 80% so $\beta = 0.2$. Under equal group sample size assumption, the sample size is:

$$\begin{aligned} n_1 = n_2 &= (z_{0.1/2} + z_{0.2})^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]/(\pi_1 - \pi_2)^2 \\ &= (1.645 + 0.842)^2[0.2 \times 0.8 + 0.3 \times 0.7]/(0.2 - 0.3)^2 = 229. \end{aligned}$$

(b) (i) 90% CI with 90% power:

$$n_1 = n_2 = 317.$$

(ii) 95% CI with 80% power:

$$n_1 = n_2 = 291.$$

(iii) 95% CI with 90% power:

$$n_1 = n_2 = 389.$$

Problem 6.1

(a)

Problem

(a) The prediction equation for $\log(\widehat{\pi}_R/\widehat{\pi}_D)$ is

$$\log(\widehat{\pi}_R/\widehat{\pi}_D) = \log(\widehat{\pi}_R/\widehat{\pi}_I) - \log(\widehat{\pi}_D/\widehat{\pi}_I) = (1.0 - 3.3) + (0.3 - (-0.2))x = -2.3 + 0.5x.$$

Given that the preference in president is Republican or Democrate, with \$10,000 increase in annual income, the odds in favor of Republican is $e^{0.5} = 1.65$ times the odds for Democrate.

(b) $\widehat{\pi}_R > \widehat{\pi}_D$ if and only if $-2.3 + 0.5x > 0$, $\Leftrightarrow x > 4.6$ (\$46000).

(c) Prediction equation for $\widehat{\pi}_I$:

$$\widehat{\pi}_I = (1 + e^{1.0+0.3x} + e^{3.3-0.2x})^{-1}.$$

Problem 6.3

(a) Using F to mean fish, I to mean invertebrate, R to mean reptile, B to mean bird and O to mean other, and using other as the reference category, we fit baseline category logit model with the following SAS program:

```
data prob6_3;
  input lake $ length $ n1-n5;
  size=(length=">2.3");
  datalines;
    Hancock =2.3 23 4 2 2 8
    Hancock >2.3 7 0 1 3 5
    Oklawaha =2.3 5 11 1 0 3
    Oklawaha >2.3 13 8 6 1 0
    Trafford =2.3 5 11 2 1 5
    Trafford >2.3 8 7 6 3 5
    George =2.3 16 19 1 2 3
    George >2.3 17 1 0 1 3
    ;
data prob6_3; set prob6_3;
  array temp {5} n1-n5;
  do y=1 to 5;
    count=temp(y);
    output;
  end;
run;

proc logistic;
  class lake / param=ref;
  freq count;
  model y (ref="5") = lake size / link=glogit aggregate scale=none;
run;
```

Part of the ouput is:

Analysis of Maximum Likelihood Estimates

Parameter		y	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1	0.0564	0.5022	0.0126	0.9107
Intercept		2	1	1.0875	0.4700	5.3530	0.0207
Intercept		3	1	-0.6742	0.6506	1.0739	0.3001
Intercept		4	1	-1.5796	0.7926	3.9717	0.0463
lake	George	1	1	1.5164	0.6214	5.9541	0.0147
lake	George	2	1	0.3944	0.6263	0.3965	0.5289
lake	George	3	1	-1.4183	1.1890	1.4229	0.2329
lake	George	4	1	0.4286	0.9383	0.2087	0.6478
lake	Hancock	1	1	0.6902	0.5597	1.5207	0.2175
lake	Hancock	2	1	-2.0901	0.7184	8.4653	0.0036
lake	Hancock	3	1	-1.0023	0.8297	1.4593	0.2270
lake	Hancock	4	1	0.2975	0.8342	0.1272	0.7214
lake	Oklawaha	1	1	1.5107	0.7532	4.0229	0.0449
lake	Oklawaha	2	1	1.3260	0.7468	3.1527	0.0758
lake	Oklawaha	3	1	1.0343	0.8402	1.5154	0.2183
lake	Oklawaha	4	1	-0.2303	1.3005	0.0313	0.8595
size		1	1	0.3316	0.4483	0.5471	0.4595
size		2	1	-1.1267	0.5049	4.9790	0.0257
size		3	1	0.6828	0.6514	1.0988	0.2945
size		4	1	0.9622	0.7127	1.8227	0.1770

Based on the output, we have the following prediction equations:

$$\log(\widehat{\pi}_F/\widehat{\pi}_O) = 0.0564 + 1.5164L_G + 0.6902L_H + 1.5107L_O + 0.3316size$$

$$\log(\widehat{\pi}_I/\widehat{\pi}_O) = 1.0875 + 0.3944L_G - 2.0901L_H + 1.3260L_O - 1.1267size$$

$$\log(\widehat{\pi}_R/\widehat{\pi}_O) = -0.6742 - 1.4183L_G - 1.0023L_H + 1.0343L_O + 0.6828size$$
$$\log(\widehat{\pi}_B/\widehat{\pi}_O) = -1.5796 + 0.4286L_G + 0.2975L_H - 0.2303L_O + 0.9622size$$

where L_G, L_H, L_O are dummy variables for Lake George, Hancock and Oklawaha, and *size* is the indicator variable for lenght > 2.3 meters.

(b) For alligators in Lake Oklawaha whose length ≤ 2.3 meters, the above equation becomes

$$\log(\widehat{\pi}_F/\widehat{\pi}_O) = 0.0564 + 1.5107 = 1.5671$$
$$\log(\widehat{\pi}_I/\widehat{\pi}_O) = 1.0875 + 1.3260 = 2.4135$$
$$\log(\widehat{\pi}_R/\widehat{\pi}_O) = -0.6742 + 1.0343 = 0.3601$$
$$\log(\widehat{\pi}_B/\widehat{\pi}_O) = -1.5796 - 0.2303 = -1.8099$$

Since $\widehat{\pi}_F + \widehat{\pi}_I + \widehat{\pi}_R + \widehat{\pi}_B + \widehat{\pi}_O = 1$, we have

$$\widehat{\pi}_F = \frac{e^{1.5671}}{1 + e^{1.5671} + e^{2.4135} + e^{0.3601} + e^{-1.8099}} = 0.258.$$

For alligators in Lake Oklawaha whose length > 2.3 meters, the above equation becomes

$$\log(\widehat{\pi}_F/\widehat{\pi}_O) = 0.0564 + 1.5107 + 0.3316 = 1.8987$$
$$\log(\widehat{\pi}_I/\widehat{\pi}_O) = 1.0875 + 1.3260 - 1.1267 = 1.2868$$
$$\log(\widehat{\pi}_R/\widehat{\pi}_O) = -0.6742 + 1.0343 + 0.6828 = 1.0429$$
$$\log(\widehat{\pi}_B/\widehat{\pi}_O) = -1.5796 - 0.2303 + 0.9622 = -0.8477$$

Again, $\widehat{\pi}_F + \widehat{\pi}_I + \widehat{\pi}_R + \widehat{\pi}_B + \widehat{\pi}_O = 1$, we have

$$\widehat{\pi}_F = \frac{e^{1.8987}}{1 + e^{1.8987} + e^{1.2868} + e^{1.0429} + e^{-0.8477}} = 0.458.$$

From these estimated probabilities, we know that larger alligators in Lake Oklawaha liked fish as primary food much more than smaller alligators (the prefence probability is almost doubled).

Problem 6.8

(a) Assign scores 1, 2, 3, 4 to the four ordinal categories "Progressive Disease", "No Change", "Partial Remission" and "Complete Remission", use `trt` as the indicator variable for Sequential and `male` the dummy variable for Male. Then we consider the cumulative logit model

for cumulative probabilities $\tau_j = P[Y \geq j|trt, male]$ for $j = 4, 3, 2$ with main effect of `trt` and `male`:

$$\text{logit}(\tau_j) = \alpha_j + \beta_1 trt + \beta_2 male, \quad j = 4, 3, 2.$$

The SAS program and part of relevant output are:

```
data prob6_8;
  input trt male n1-n4;
  datalines;
  1 1 28 45 29 26
  1 0  4 12  5  2
  0 1 41 44 20 20
  0 0 12  7  3  1
;

data prob6_8; set prob6_8;
    array temp {4} n1-n4;

    do y=1 to 4;
      count=temp(y);
      output;
    end;
run;

proc logistic data=prob6_8 descending;
   freq count;
   model y = trt male / aggregate scale=none;
run;
```

```
*************************************************************************
            Score Test for the Proportional Odds Assumption

                  Chi-Square       DF     Pr > ChiSq

                      2.9280        4         0.5699


           Deviance and Pearson Goodness-of-Fit Statistics

    Criterion          Value       DF     Value/DF     Pr > ChiSq

    Deviance          5.5677        7       0.7954         0.5910
    Pearson           5.3527        7       0.7647         0.6170

                        The LOGISTIC Procedure

              Analysis of Maximum Likelihood Estimates

                                    Standard          Wald
Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept 4     1     -2.4221      0.3276       54.6609        <.0001
Intercept 3     1     -1.3713      0.3059       20.0903        <.0001
Intercept 2     1      0.1960      0.2947        0.4424        0.5060
trt             1      0.5807      0.2119        7.5131        0.0061
male            1      0.5414      0.2953        3.3619        0.0667
```

Regardless of gneder, the odds of having better outcomes for patients treated with Sequential therapy is $e^{0.5807} = 1.79$ times the odds for patients treated with Alternating therapy.

(b) Cumulative logit model with treatment and gender interaction:

$$\text{logit}(\tau_j) = \alpha_j + \beta_1 trt + \beta_2 male + \beta_3 trt \times male.$$

The effect of Sequential therapy relative to Alternating therapy in terms of log odds-ratio is:

$$\text{logit}(\tau_j(trt = 1)) - \text{logit}(\tau_j(trt = 0)) = \alpha_j + \beta_1 + \beta_2 male + \beta_3 male - (\alpha_j + \beta_2 male) = \beta_1 + \beta_3 male,$$

or equivalently, the odds-ratio is:

$$\frac{\tau_j(trt=1))/(1-\tau_j(trt=1))}{\tau_j(trt=0))/(1-\tau_j(trt=0))} = e^{\beta_1+\beta_3 male}$$

```
proc logistic data=prob6_8 descending;
   freq count;
   model y = trt male trt*male / aggregate scale=none;
run;
```

```
**********************************************************************
            Score Test for the Proportional Odds Assumption

                  Chi-Square       DF     Pr > ChiSq

                      3.8245        6         0.7004


           Deviance and Pearson Goodness-of-Fit Statistics

    Criterion          Value       DF     Value/DF     Pr > ChiSq

    Deviance          4.5209        6       0.7535         0.6066
    Pearson           4.4151        6       0.7359         0.6207

                        The LOGISTIC Procedure

              Analysis of Maximum Likelihood Estimates

                                    Standard          Wald
Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept 4     1     -2.6978      0.4260       40.1002        <.0001
Intercept 3     1     -1.6484      0.4102       16.1462        <.0001
Intercept 2     1     -0.0770      0.3986        0.0373        0.8468
trt             1      1.0786      0.5498        3.8490        0.0498
male            1      0.8646      0.4309        4.0268        0.0448
trt*male        1     -0.5906      0.5935        0.9901        0.3197
```

Based on the output, the estimated odds-ratio of have better outcomes between Sequential therapy relative to Alternating therapy is

$$\frac{\widehat{\tau}_j(trt=1))/(1-\widehat{\tau}_j(trt=1))}{\widehat{\tau}_j(trt=0))/(1-\widehat{\tau}_j(trt=0))} = e^{1.0786-0.5906 male}.$$

So for females, the estimated odds-ratio of have better outcomes between Sequential therapy relative to Alternating therapy is $e^{1.0786} = 2.94$.

For males, the estimated odds-ratio of have better outcomes between Sequential therapy relative to Alternating therapy is $e^{1.0786-0.5906} = 1.63$.

Therefore, the treatment effect is greater for female patients than male patients.

(c) The score tests for the goodness-of-fit of the model with interaction are $\chi^2 = 3.8245$ with $df = 6$ (P-value=0.7004) and $\chi^2 = 2.9280$ with $df = 4$ (P-value=0.5699) for the model without interaction. The P-values indicate that the interaction model does not provide a much better fit compared to the model without interaction. Similarly, the LRT comparing the models with and without interaction is $G^2 = 5.5677 - 4.5209 = 1.0468$ with $df = 1$. The

P-value $=P(\chi^2_1 \geq 1.0468) = 0.306$, indicating the model with interaction does not provide a much better fit.

Problem 6.11

(a) Denote $Y = 1, 2, 3, 4$ for the four ordinal catergories of job satisfaction. We consider the cumulative logit model for the cumulative probabilities $\tau_j = P[Y \geq j|income, gender]$ for $j = 4, 3, 2$:

$$\text{logit}(\tau_j) = \alpha_j + \beta_1 income + \beta_2 male, \quad j = 4, 3, 2,$$

where `income` is the income scores 3, 10, 20, 35 for 4 ornidal income categories and `male` is the dummy variable for gender male.

The SAS program and relevant output are:

```
data prob6_11;
  input gender$ income$ incscore n1-n4;
  male=(gender="Male");
  cards;
  Female <5000          3  1 3 11 2
  Female 5000~15,000    10 2 3 17 3
  Female 15,000~25,000 20 0 1  8 5
  Female >25,000        35 0 2  4 2
  Male   <5000          3  1 1  2 1
  Male   5000~15,000    10 0 3  5 1
  Male   15,000~25,000 20 0 0  7 3
  Male   >25,000        35 0 1  9 6
  ;

data prob6_11; set prob6_11;
  array temp {4} n1-n4;

  do y=1 to 4;
    count=temp(y);
    output;
  end;
run;

proc logistic data=prob6_11 descending;
  freq count;
  model y = incscore male / aggregate=(income male) scale=none;
run;
```

```
*************************************************************************

            Score Test for the Proportional Odds Assumption

                  Chi-Square       DF     Pr > ChiSq

                      1.3416        4         0.8543


           Deviance and Pearson Goodness-of-Fit Statistics

    Criterion          Value       DF     Value/DF     Pr > ChiSq

    Deviance         13.9519       19       0.7343         0.7865
    Pearson          14.3128       19       0.7533         0.7652

              Analysis of Maximum Likelihood Estimates

                                Standard          Wald
Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept 4     1     -2.0780      0.4206       24.4101        <.0001
Intercept 3     1      0.8940      0.3603        6.1569        0.0131
```

```
Intercept 2     1      2.5795      0.5618       21.0840        <.0001
incscore        1      0.0444      0.0185        5.7372        0.0166
male            1      0.0259      0.4274        0.0037        0.9516

                    Odds Ratio Estimates

                       Point          95% Wald
          Effect      Estimate      Confidence Limits

          incscore       1.045       1.008       1.084
          male           1.026       0.444       2.372
```

Since $\widehat{\beta}_1 = 0.0444 > 0$ is significant P-value < 0.5, and the associated odds-ratio estimate is 1.045, we conclude that income has a positive effect on the job satisfaction. Since the income score roughly represents the mid-income (in thousands) for each category, we estimated that with every \$ 10 increase in income, the odds of more job satisfaction increases by $e^{0.0444\times 10} - 1 = 1.56 - 1 = 57\%$ regardless of gender.

(b) We combine the first two categories of jobsatisfaction by defining $Y^* = Y$ for $Y = 3, 4$ and $Y^* = 2$ for $Y = 1, 2$ and consider the cumulative logit model for Y^*. If the cumulative logit model is true if Y, it will also be true for Y^* with the same α_4, α_3 and the same β_1, β_2.

The SAS program and the relevant output are:

```
data  prob6_11; set prob6_11;
  newy = y;
  if y=1 then newy=2;
run;

proc logistic data=prob6_11 descending;
  freq count;
  model newy = incscore male / aggregate=(income male) scale=none;
run;

*****************************************************************

            Score Test for the Proportional Odds Assumption

                  Chi-Square       DF     Pr > ChiSq

                      0.0557        2         0.9725

          Deviance and Pearson Goodness-of-Fit Statistics

    Criterion          Value       DF     Value/DF     Pr > ChiSq

    Deviance          9.1948       12       0.7662         0.6862
    Pearson           8.9175       12       0.7431         0.7100

              Analysis of Maximum Likelihood Estimates

                                 Standard          Wald
Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept 4     1     -2.0582      0.4197       24.0511        <.0001
Intercept 3     1      0.9154      0.3616        6.4091        0.0114
incscore        1      0.0435      0.0186        5.4991        0.0190
male            1      0.0247      0.4286        0.0033        0.9541
```

Indeed, the estimates of α_4, α_3 and the β_1, β_2 from the model for Y^* are basically the same as those obtained in (a) for the original Y. The means the covariate effects are invariant to

the choice of cut points used to form the categories for the response.

(c) Yes, we can drop gender from the model in (a) since the P-value 0.9516.