Homework #7

1. *From Kutner et al., Applied Linear Regression Models, p. 609.*

Refer to the Pregnancy Duration Data (p. 609), repeat the analysis on p. 617 (the response variable is treated as Ordinal categorical and a proportional odds model is used) using R or other statistical software. Compare your results with the ones in the text (from Minitab). Are they the same? If not, what is the cause? Interpret the parameters in the context of the problem.

```
df1$OrderedRes <- ordered(df1$preg, c("1", "2", "3"))
polr_reg <- polr(OrderedRes ~ .-preg-preg1-preg2-preg3, data=df1 , Hess=T)
summary(polr_reg)

## Call:
## polr(formula = OrderedRes ~ . - preg - preg1 - preg2 - preg3,
##     data = df1, Hess = T)
##
## Coefficients:
##            Value Std. Error t value
## nutri    0.04887    0.01182   4.133
## age1    -1.97601    0.57616  -3.430
## age3    -1.36348    0.54648  -2.495
## alcohol -1.66987    0.47537  -3.513
## smoking -1.59154    0.45165  -3.524
##
## Intercepts:
##     Value   Std. Error t value
## 1|2  2.9301  1.4929     1.9627
## 2|3  5.0249  1.5445     3.2535
##
## Residual Deviance: 173.5122
## AIC: 187.5122
```

The signs of the coefficients are not the same because of different parameterization of two softwares. Keeping other predictors constant, as nutrition status increases 1 unit, the odds ratio that the mother is in a lower pregnancy category vs. a higher pregnancy category is $e^{-0.049} = 0.952$. After adjusted for other predictors, when a mother's age change from age category 2 to age category 1, the odds of the mother is in a lower pregnancy category vs. a higher pregnancy category will change by a factor $e^{1.97601} = 7.214$. Similarly, interpretation for other parameters can be drawn according to the output.

2. *From Dobson & Barnett, An Introduction to Generalized Linear Models, p. 163 Exercises 8.2 (c, d)*

c. Do you think an ordinal model would be appropriate for associations between the levels of satisfaction and the other variables? Justify your answer. If you consider such a model to be appropriate, fit a suitable one and compare the results with those from (b).

- In part (c), use a proportional odds model without interaction. Then,

```
attach(df2)
df2$OrderedRes <- ordered(df2$satisfaction2, c("1", "2", "3"))
polr_reg_sat <- polr(OrderedRes ~ contact + type, weights = frequency, data = df2)
summary(polr_reg_sat)

## Call:
## polr(formula = OrderedRes ~ contact + type, data = df2, weights = frequency)
##
## Coefficients:
##                   Value Std. Error t value
## contactlow     -0.2524    0.09306  -2.713
## typeHouse      -0.2353    0.10521  -2.236
## typeTowerBlock  0.5010    0.11675   4.291
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -0.7488  0.0818    -9.1570
## 2|3  0.3637  0.0801     4.5393
##
## Residual Deviance: 3610.286
## AIC: 3620.286
```

  i. Conduct a Pearson goodness of fit test.

```
df2 %>%
  arrange(type, contact) -> df2
observed<-matrix(df2$frequency, byrow=T, ncol=3)
observed
##      [,1] [,2] [,3]
## [1,]  141  116  191
## [2,]  130   76  111
## [3,]  130  105  104
## [4,]   67   48   62
## [5,]   34   47  100
## [6,]   65   54  100
df3 <- df2 %>%
  group_by(type, contact) %>%
  summarise(n = sum(frequency))
df3 <- df3[rep(row.names(df3), each = 3),]
yhat<-predict(polr_reg_sat, type="probs")*df3$n
yhat
##            1         2         3
## 1  120.71972 116.16444 211.11584
```

```
## 2   99.78654 108.86302 239.35044
## 3  120.71972 116.16444 211.11584
## 4   70.60789  77.03031 169.36181
## 5   85.41998  82.19671 149.38331
## 6   70.60789  77.03031 169.36181
## 7  128.27413  91.85073 118.87514
## 8  108.84520  91.14175 139.01305
## 9  128.27413  91.85073 118.87514
## 10  56.83068  47.58728  72.58204
## 11  66.97499  47.95746  62.06755
## 12  56.83068  47.58728  72.58204
## 13  78.75208  48.10424  54.14368
## 14  67.76020  49.06122  64.17858
## 15  78.75208  48.10424  54.14368
## 16  81.98610  59.36137  77.65253
## 17  95.28567  58.20348  65.51086
## 18  81.98610  59.36137  77.65253
```

```
expected<-yhat[c(1:6)*3, ]
expected
```

```
##            1         2         3
## 3  120.71972 116.16444 211.11584
## 6   70.60789  77.03031 169.36181
## 9  128.27413  91.85073 118.87514
## 12  56.83068  47.58728  72.58204
## 15  78.75208  48.10424  54.14368
## 18  81.98610  59.36137  77.65253
```

```
rsp<-(observed-expected)/sqrt(expected)  # Standardized Pearson Residuals
c("PearsonChiSq"=sum(rsp^2), "df" = 12-5, "p-value"= 1-pchisq(sum(rsp^2), 12-5
```

```
## PearsonChiSq           df      p-value
##     157.2687        7.0000         0.0000
```

The p-value of the Pearson goodness of fit test is close to 0, so the model does not fit the data well.

ii. Use Likelihood Ratio Test to test whether adding interaction improves the model.

```
full <- polr(OrderedRes ~ contact * type, weights = frequency, data = df2)
DevChi<-polr_reg_sat$dev - full$dev
c("DevChiSq"=DevChi, "p-value"=1-pchisq(DevChi, 2))  # df=2
```

```
##   DevChiSq    p-value
## 6.19554642 0.04514963
```

H_0: $log(\frac{P(y\leq j)}{1-P(y\leq j)}) = \alpha_j - (\beta_1 X_1 + beta_2 X_2)$

H_A: $log(\frac{P(y\leq j)}{1-P(y\leq j)}) = \alpha_j - (\beta_1 X_1 + beta_2 X_2 + beta_3 X_1 X_2)$

Where $j = 1, 2, \cdots, (J-1)$ The p-value for this test is $0.045 < 0.05$, so we reject the null hypothesis and conclude that adding interaction improves the model.

- Note that in this case, the proportional odds model with interaction is NOT the saturated model.

c. From the best model you obtained in (c), calculate the standardized residuals and use them to find where the largest discrepancies are between the observed frequencies and expected frequencies estimated from the model.

```
rsp
```

```
##             1           2         3
## 3   1.8458007 -0.01525710 -1.384451
## 6   7.0680913 -0.11739107 -4.484572
## 9   0.1523839  1.37201983 -1.364318
## 12  1.3489644  0.05982826 -1.242095
## 15 -5.0429215 -0.15921094  6.231970
## 18 -1.8759599 -0.69586299  2.536007
```

```
df2
```

```
##          type satisfaction contact satisfaction2 frequency OrderedRes
## 1   Apartment          low    high             1       141          1
## 2   Apartment       medium    high             2       116          2
## 3   Apartment         high    high             3       191          3
## 4   Apartment          low     low             1       130          1
## 5   Apartment       medium     low             2        76          2
## 6   Apartment         high     low             3       111          3
## 7       House          low    high             1       130          1
## 8       House       medium    high             2       105          2
## 9       House         high    high             3       104          3
## 10      House          low     low             1        67          1
## 11      House       medium     low             2        48          2
## 12      House         high     low             3        62          3
## 13 TowerBlock          low    high             1        34          1
## 14 TowerBlock       medium    high             2        47          2
## 15 TowerBlock         high    high             3       100          3
## 16 TowerBlock          low     low             1        65          1
## 17 TowerBlock       medium     low             2        54          2
## 18 TowerBlock         high     low             3       100          3
```

The prediction for the satisfaction of the house that is an apartment and has a low type contact has the largest discrepancy.