

Political Science 150C/350C  
Spring 2007  
Problem Set 2 - Answer Guide

Again we've interleaved some R code. Again we stress that this is for your edification only; an article-quality write-up does not include R code or raw output.

1. Consider the data frame `nes96` in the R package `faraway`. This data frame consists of 944 responses to the 1996 National Election Study on selected variables. A seven point party identification measure is in the data frame as variable `PID` (an ordered factor). Model the ordinal responses to for `PID` as a function of income and levels of education. Fit a series of models to the data: a null "intercept-only" model, a model with income as a predictor, and a model with income and levels of education as a predictor. Use likelihood ratio tests and other goodness of fit criteria to compare the models. For your preferred model, generate a graph showing how the predicted probabilities of specific party identification categories change as a function of income, for one or two different values of the education variable. Provide a paragraph or two of supporting text detailing what strikes you as interesting/compelling in the data, as revealed by the model fitting and/or the predicted probability graphs etc.

Notes:

- The `income` variable is given as an ordered factor. Before using the variable in analysis, convert this to a continuous variable by assigning the mid-point of each income category (in thousands of dollars) to the corresponding level of the ordered factor. You can do this with the following R commands:

```
inca <- c(1.5,4,6,8,9.5,10.5,11.5,12.5,13.5,14.5,  
         16,18.5,21,23.5,  
         27.5,32.5,37.5,42.5,47.5,55,67.5,82.5,97.5,115)  
nes96$nincome <- inca[unclass(nes96$income)]
```

- The `educ` variable is also an ordered factor. Convert this to a regular factor in R with the command:

```
nes96$neduc <- factor(nes96$educ,ordered=FALSE)
```

- I will supply R code that will generate (1) `hitmiss` tables and (2) McFadden's pseudo- $r^2$  for objects of class `polr` (and classes `glm` and `multinom`, for that matter). Watch my web site and/or your e-mail.

**Answer:** Here's my (Nathan's) not-nearly-so-commented-as-Simon's code. Comments to follow.

R Code

```
1 > library(MASS)
2 > source("helper.R")
3 > data(nes96, package = "faraway")
4 > inc <- c(1.5, 4, 6, 8, 9.5, 10.5, 11.5, 12.5, 13.5, 14.5, 16,
5 + 18.5, 21, 23.5, 27.5, 32.5, 37.5, 42.5, 47.5, 55, 67.5, 82.5,
6 + 97.5, 115)
7 > nes96$nincome <- inc[unclass(nes96$income)]
8 > nes96$neduc <- factor(nes96$educ, ordered = FALSE)
9 > m0 <- PID ~ 1
10 > m1 <- PID ~ nincome
11 > m2 <- PID ~ nincome + neduc
12 > o0 <- polr(m0, data = nes96, Hess = TRUE, method = c("logistic"))
13 > summary(o0)
```

Call:  
polr(formula = m0, data = nes96, Hess = TRUE, method = c("logistic"))

No coefficients

Intercepts:

	Value	Std. Error	t value
strDem weakDem	-1.3137	0.0796	-16.4937
weakDem indDem	-0.3949	0.0664	-5.9500
indDem indind	0.0678	0.0651	1.0413
indind indRep	0.2255	0.0655	3.4427
indRep weakRep	0.6443	0.0685	9.4053
weakRep strRep	1.4803	0.0838	17.6745

Residual Deviance: 3500.693  
AIC: 3512.693

R Code

```
1 > o1 <- polr(m1, data = nes96, Hess = TRUE, method = c("logistic"))
2 > summary(o1)
```

Call:  
polr(formula = m1, data = nes96, Hess = TRUE, method = c("logistic"))

Coefficients:

	Value	Std. Error	t value
nincome	0.0127599	0.001870919	6.820126

Intercepts:

	Value	Std. Error	t value
strDem weakDem	-0.7618	0.1126	-6.7676
weakDem indDem	0.1924	0.1085	1.7730
indDem indind	0.6761	0.1109	6.0983
indind indRep	0.8394	0.1120	7.4978
indRep weakRep	1.2708	0.1158	10.9775
weakRep strRep	2.1273	0.1287	16.5356

Residual Deviance: 3453.106  
AIC: 3467.106

R Code

```
1 > o2 <- polr(m2, data = nes96, Hess = TRUE, method = c("logistic"))
2 > summary(o2)
```

Call:  
polr(formula = m2, data = nes96, Hess = TRUE, method = c("logistic"))

```

Coefficients:
              Value Std. Error t value
nincome      0.01218715 0.002023671 6.0223004
neduchSdrop  0.23397272 0.551651669 0.4241313
neduchS      0.63664195 0.507331243 1.2548842
neducColl    0.80827482 0.512167514 1.5781454
neducCdeg    0.76311017 0.529641702 1.4408045
neducBAdeg   0.79031546 0.513019984 1.5405159
neducMAdeg   0.51039905 0.526790959 0.9688835

```

```

Intercepts:
              Value Std. Error t value
strDem|weakDem -0.1248 0.4943 -0.2526
weakDem|indDem 0.8351 0.4950 1.6868
indDem|indind 1.3209 0.4962 2.6618
indind|indRep 1.4850 0.4966 2.9900
indRep|weakRep 1.9196 0.4980 3.8550
weakRep|strRep 2.7815 0.5016 5.5450

```

```

Residual Deviance: 3444.638
AIC: 3470.638

```

So there are the estimated models. Let's first look at how the goodnesses-of-fit compare. First, the AICs indicate that the income model is a discernible improvement over the null model, and whatever additional improvement we get from adding education is outweighed by the number of additional parameters (six).

```

_____ R Code _____
1 > anova(o0, o1)

```

Likelihood ratio tests of ordinal regression models

```

Response: PID
      Model Resid. df Resid. Dev  Test   Df LR stat.    Pr(Chi)
1         1      938   3500.693
2 nincome      937   3453.106 1 vs 2    1 47.58727 5.260903e-12

```

```

_____ R Code _____
1 > anova(o0, o2)

```

Likelihood ratio tests of ordinal regression models

```

Response: PID
      Model Resid. df Resid. Dev  Test   Df LR stat.    Pr(Chi)
1         1      938   3500.693
2 nincome + neduc      931   3444.638 1 vs 2    7 56.0557 9.20653e-10

```

```

_____ R Code _____
1 > anova(o1, o2)

```

Likelihood ratio tests of ordinal regression models

```

Response: PID
      Model Resid. df Resid. Dev  Test   Df LR stat.    Pr(Chi)
1         nincome      937   3453.106
2 nincome + neduc      931   3444.638 1 vs 2    6 8.468435 0.2057528

```

So both the income model and the income-and-education model are statistically distinguishable from the intercept-only model; the LRs are large and the p-values are teeny-tiny. The income-and-education model is at best a marginal improvement over the income-only model; the difference isn't statistically significant at conventional levels. (That's why the AIC declines from the income to income-and-education model.)

The PCP for the null model is about 21 percent; it improves to about 25 percent for the income model and declines very slightly from 25 percent for the income-and-education model — but don't take that decline very seriously, because PCP is a rough measure of GOF and the decline is small. The key thing is the improvement when you add income. Here are the hit-miss tables and pseudo- $R^2$  as well.

```

1 > hitmiss(o0)

```

Table of Actual (y) Against Predicted (p)  
Classification rule: outcome with highest probability.

	p=strDem	p=weakDem	p=indDem	p=indind	p=indRep	p=weakRep	p=strRep
y=strDem	200	0	0	0	0	0	0
y=weakDem	180	0	0	0	0	0	0
y=indDem	108	0	0	0	0	0	0
y=indind	37	0	0	0	0	0	0
y=indRep	94	0	0	0	0	0	0
y=weakRep	150	0	0	0	0	0	0
y=strRep	175	0	0	0	0	0	0
Row PCP							
y=strDem	100						
y=weakDem	0						
y=indDem	0						
y=indind	0						
y=indRep	0						
y=weakRep	0						
y=strRep	0						

Percent Correctly Predicted, Fitted Model: 21.19%  
Percent Correctly Predicted, Null Model : 21.19%

```

1 > hitmiss(o1)

```

Table of Actual (y) Against Predicted (p)  
Classification rule: outcome with highest probability.

	p=strDem	p=weakDem	p=indDem	p=indind	p=indRep	p=weakRep	p=strRep
y=strDem	156	0	0	0	0	0	44
y=weakDem	128	0	0	0	0	0	52
y=indDem	57	0	0	0	0	0	51
y=indind	19	0	0	0	0	0	18
y=indRep	47	0	0	0	0	0	47
y=weakRep	80	0	0	0	0	0	70
y=strRep	86	0	0	0	0	0	89
Row PCP							
y=strDem	78.00						
y=weakDem	0.00						
y=indDem	0.00						
y=indind	0.00						

```

y=indRep      0.00
y=weakRep     0.00
y=strRep      50.86

```

```

Percent Correctly Predicted, Fitted Model: 25.95%
Percent Correctly Predicted, Null Model : 21.19%

```

```

R Code
1 > hitmiss(o2)

```

Table of Actual (y) Against Predicted (p)

Classification rule: outcome with highest probability.

	p=strDem	p=weakDem	p=indDem	p=indind	p=indRep	p=weakRep	p=strRep
y=strDem	155	0	0	0	0	0	45
y=weakDem	128	0	0	0	0	0	52
y=indDem	59	0	0	0	0	0	49
y=indind	19	0	0	0	0	0	18
y=indRep	56	0	0	0	0	0	38
y=weakRep	80	0	0	0	0	0	70
y=strRep	90	0	0	0	0	0	85

```

Row PCP
y=strDem      77.50
y=weakDem     0.00
y=indDem      0.00
y=indind      0.00
y=indRep      0.00
y=weakRep     0.00
y=strRep      48.57

```

```

Percent Correctly Predicted, Fitted Model: 25.42%
Percent Correctly Predicted, Null Model : 21.19%

```

Note what the null model says: everybody's a strong Dem. The income model and income-and-education model say everybody's either a strong Dem or Republican. So basically the model's not picking up any variation in the strength of party ID, and it's getting some of the party IDs wrong. So, while I'm generally skeptical of PCP and hit-miss as a measure of GOF, I am more skeptical of this model: the model predicts that about a quarter of the strong Democrats are strong Republicans, which is pretty weird. Pseudo- $R^2$  tells the same story: the null model doesn't explain much of anything, and the other models are an improvement but aren't doing a great job by any means.

Finally, some plots. You should sorta expect that the biggest difference is between getting a college degree and not getting one, so I'll look at high school and bachelor's levels of education in model 2. (Though there isn't a lot of evidence that education makes much difference, let's see what the pictures look like.)

```

R Code
1 > p2.hs <- predict(o2, newdata = expand.grid(neduc = "HS", nincome = inc),
2 +   type = "probs")
3 > p2.ba <- predict(o2, newdata = expand.grid(neduc = "BAdeg", nincome = inc),
4 +   type = "probs")

```

```

R Code
1 > par(mfrow = c(2, 1), mgp = c(1.65, 0.65, 0), mar = c(2.5, 3,
2 + 3, 0.5), oma = rep(0, 4), cex = 1)
3 > matplot(inc, p2.hs, type = "l", col = 1:7, pch = 16, lwd = 2,
4 + cex = 0.5, bty = "n", lty = 1:7, ylim = c(0, 0.5))
5 > title("High School education")
6 > legend(legend = levels(nes96$PID), col = 1:7, lwd = 2, pch = 16,
7 + cex = 0.5, pt.cex = 0.35, x = "topleft", bty = "n")
8 > matplot(inc, p2.ba, type = "l", col = 1:7, pch = 16, lwd = 2,
9 + cex = 0.5, bty = "n", lty = 1:7, ylim = c(0, 0.5))
10 > title("College Degree")
11 > legend(legend = levels(nes96$PID), col = 1:7, lwd = 2, pch = 16,
12 + cex = 0.5, pt.cex = 0.35, x = "topleft", bty = "n")

```

What's apparent from the graphs (Figure 1) is that there isn't a big difference; it appears that a college education will make slightly more conservative for a given income level, but the differences aren't big (and aren't significant, anyway, if you look back to the coefficients). Unsurprisingly, increasing income makes you more likely to identify with the Republican party. Somewhat surprisingly, the model says that you become more likely to be a Republican than Democratic identifier at around \$40,000-60,000 income, which strikes me (at least) as kinda low. But maybe not.

2. [Download](#) the file `nomocc2.dta` from my web site. This file is in `stata` format, and you'll need the `foreign` library in R in order to be able to read it. The file contains 337 observations from the General Social Survey. The outcome of interest is occupational type, with (unordered) categories, "menial", "blue collar", "craft", "white collar" and "professional". The predictors are

- `white`, 1 if white, 0 otherwise
- `ed`, years of education
- `exper`, an estimate of the possible number of years that the respondent could have been in the workforce (age minus years of education minus 5).

Analyze these occupational attainment data via multinomial logit analysis, so as to address the following questions in a short write-up (say, two to three pages).

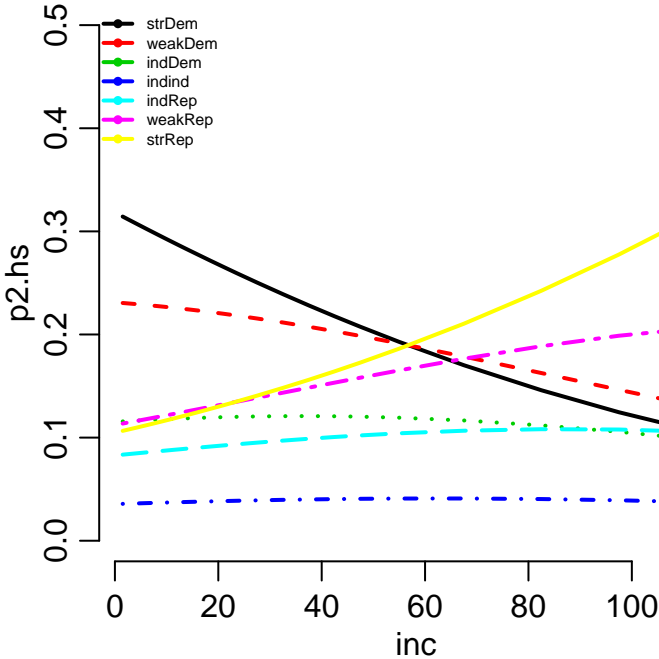
```

R Code
1 > library(foreign)
2 > dat <- read.dta("nomocc2.dta")
3 > dat$white <- factor(dat$white, levels = c(0, 1), labels = c("Non-White",
4 + "White"))
5 > library(nnet)

```

- Leaving aside other the predictors, does the distribution of occupational attainment vary by race?

### High School education



### College Degree

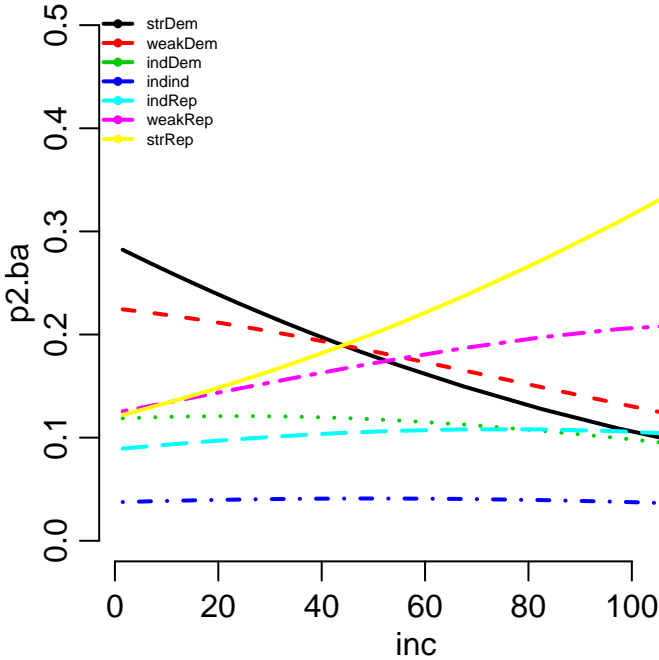


Figure 1: Predicted probabilities of party ID by income and education.

```

1 > library(lattice)
2 > print(histogram(~occ | white, xlab = "", data = dat))

```

**Answer:** See Figure 2. We also can run a simple test of association, using Fisher's exact test due to the small cell counts:

```

1 > tab <- table(dat$occ, dat$white)
2 > print(tab)

```

	Non-White	White
Menial	5	26
BlueCol	4	65
Craft	10	74
WhiteCol	2	39
Prof	7	105

```

1 > fisher.test(tab)

```

```

Fisher's Exact Test for Count Data

data: tab
p-value = 0.2401
alternative hypothesis: two.sided

```

There doesn't seem to be any statistically significant evidence of association here.

- After controlling for education and experience, does race plays a role in determining occupational type?

```

1 > q2 <- multinom(occ ~ white + ed + exper, data = dat)

```

```

# weights: 25 (16 variable)
initial value 542.380576
iter 10 value 472.154249
iter 20 value 426.809681
final value 426.800478
converged

```

```

1 > q2Restricted <- multinom(occ ~ ed + exper, data = dat)

```

```

# weights: 20 (12 variable)
initial value 542.380576
iter 10 value 449.361211
iter 20 value 430.848185
iter 20 value 430.848182
iter 20 value 430.848182
final value 430.848182
converged

```

```

1 > anova(q2Restricted, q2, test = "Chi")

```

Likelihood ratio tests of Multinomial Models

```

Response: occ

```

	Model	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
1	ed + exper	1336	861.6964				
2	white + ed + exper	1332	853.6010	1 vs 2	4	8.095408	0.08814514

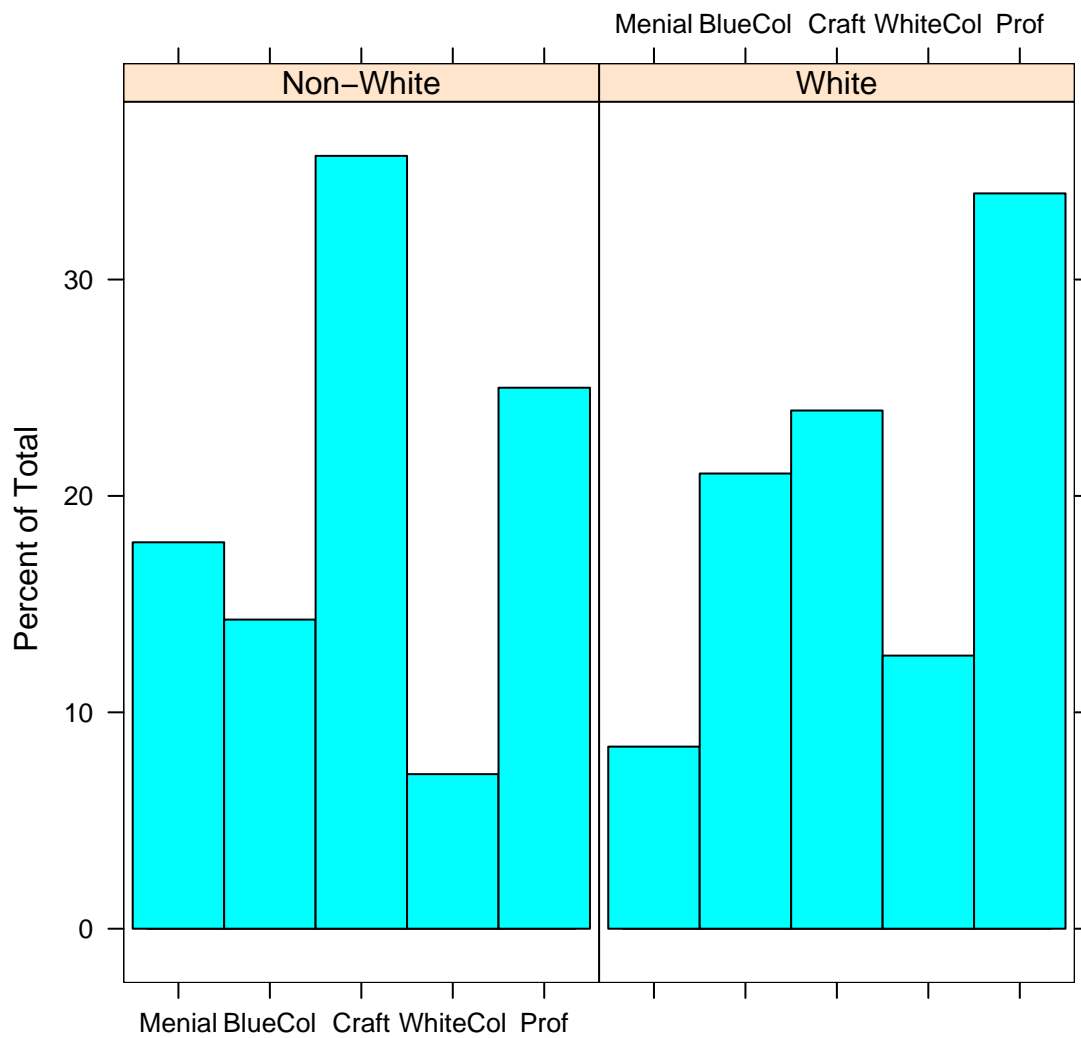


Figure 2: Histogram of Occupational Status by Race

Outcome	Intercept	White	Educ	Experience
BlueCol	0.741 (1.52)	1.24 (0.724)	-0.0994 (0.102)	0.00472 (0.0174)
Craft	-1.09 (1.45)	0.472 (0.604)	0.0938 (0.0976)	0.0277 (0.0167)
WhiteCol	-6.24 (1.9)	1.57 (0.903)	0.353 (0.117)	0.0346 (0.0188)
Prof	-11.5 (1.85)	1.77 (0.755)	0.779 (0.115)	0.0356 (0.018)

Table 1: Maximum Likelihood Parameter Estimate, Multinomial Logit Model of Occupation Status, standard errors in parentheses. Baseline category is Menial.

**Answer:** Table 1 presents the maximum likelihood estimates of a multinomial model for these data, using three predictors: a binary indicator for white respondents, years of education and years of experience. A likelihood ratio test suggests that after controlling for education and experience, race does not have a statistically significant impact on occupational classification. That is, we fail to reject the null hypothesis that the 4 multinomial logit coefficients on the dummy variable for white are jointly zero, at least at conventional  $p = .05$  levels of statistical significance ( $\chi_4^2 = 8.10, p = .09$ ). Note that the AIC from the unrestricted, “with race effects” model is 885.6, while the AIC from the restricted “no race effects” model is 885.7. Coupled with the  $p$ -value of .09, for the remainder of the solution key I will retain race as a predictor.

- What is the role of education in determining occupational type; that is, how does the predicted probability of different levels of occupational attainment vary by education, net of the effects of other predictors?

```

R Code
1 > par(mfrow = c(2, 1), mgp = c(1.65, 0.65, 0), mar = c(2.5, 3,
2 +   3, 0.5), oma = rep(0, 4), cex.lab = 0.75, cex.axis = 0.75,
3 +   cex.main = 0.75)
4 > xseq <- sort(unique(dat$ed))
5 > newdata <- expand.grid(list(ed = xseq, white = "White", exper = 17))
6 > phat <- predict(q2, newdata, type = "probs")
7 > par(las = 1)
8 > matplot(x = xseq, y = phat, xlab = "Education", ylab = "Probability",
9 +   type = "b", pch = 16, lwd = 2, cex = 0.5, bty = "n", lty = 1)
10 > title("Whites")
11 > legend(legend = levels(dat$occ), col = 1:5, lwd = 2, pch = 16,
12 +   cex = 0.5, pt.cex = 0.35, x = "topleft", bty = "n")
13 > phat <- predict(q2, newdata = expand.grid(list(ed = xseq, white = "Non-White",
14 +   exper = 17)), type = "probs")
15 > matplot(x = xseq, y = phat, xlab = "Education", ylab = "Probability",
16 +   type = "b", pch = 16, lwd = 2, cex = 0.5, bty = "n", lty = 1)
17 > title("Non-Whites")

```

```

18 > legend(legend = levels(dat$occ), col = 1:5, lwd = 2, pch = 16,
19 +       cex = 0.5, pt.cex = 0.5, x = "topleft", bty = "n")

```

**Answer:** See Figure 3. For whites, the two most likely occupational categories are blue collar and professional, with a break between the two becoming apparent at around 14 years of education. For non whites, the model predicts craft occupations for all 8 through 15 years of education, at which point professional occupations are the most likely predicted occupation choice.

- Are the effects of education are conditional on race? (i.e., does the same marginal change in levels of education translates into different levels of occupational attainment, depending on whether the subject is white or non-white?)

**Answer:** Again, see Figure 3 and the discussion above. We can also test this formally, by estimating an expanded version of the model with education effects specific to the two racial groups:

```

R Code
1 > q2More <- multinom(occ ~ white * ed + exper, data = dat)

# weights: 30 (20 variable)
initial value 542.380576
iter 10 value 489.103083
iter 20 value 425.941100
iter 30 value 425.800327
final value 425.795524
converged

R Code
1 > anova(q2, q2More, test = "Chi")

```

Likelihood ratio tests of Multinomial Models

```

Response: occ

```

	Model	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
1	white + ed + exper	1332	853.601				
2	white * ed + exper	1328	851.591	1 vs 2	4	2.009906	0.7339368

We fail to reject the restrictions in our working model; the effects of education on occupational attainment do not seem to be conditional on race.

- Based on these data and your modeling, does it seem that education and experience help equalize racial differences in occupational attainment?

**Answer:** The probability of attaining a professional occupation for non-whites trails the corresponding probability for white, at all levels of educational attainment. For the scenario graphed in Figure 3, a professional occupation becomes the most likely predicted outcome for a non-white worker at around 15 years of age, while this occurs at around 13 years of education for a white worker.

- How well do your models fit the data?

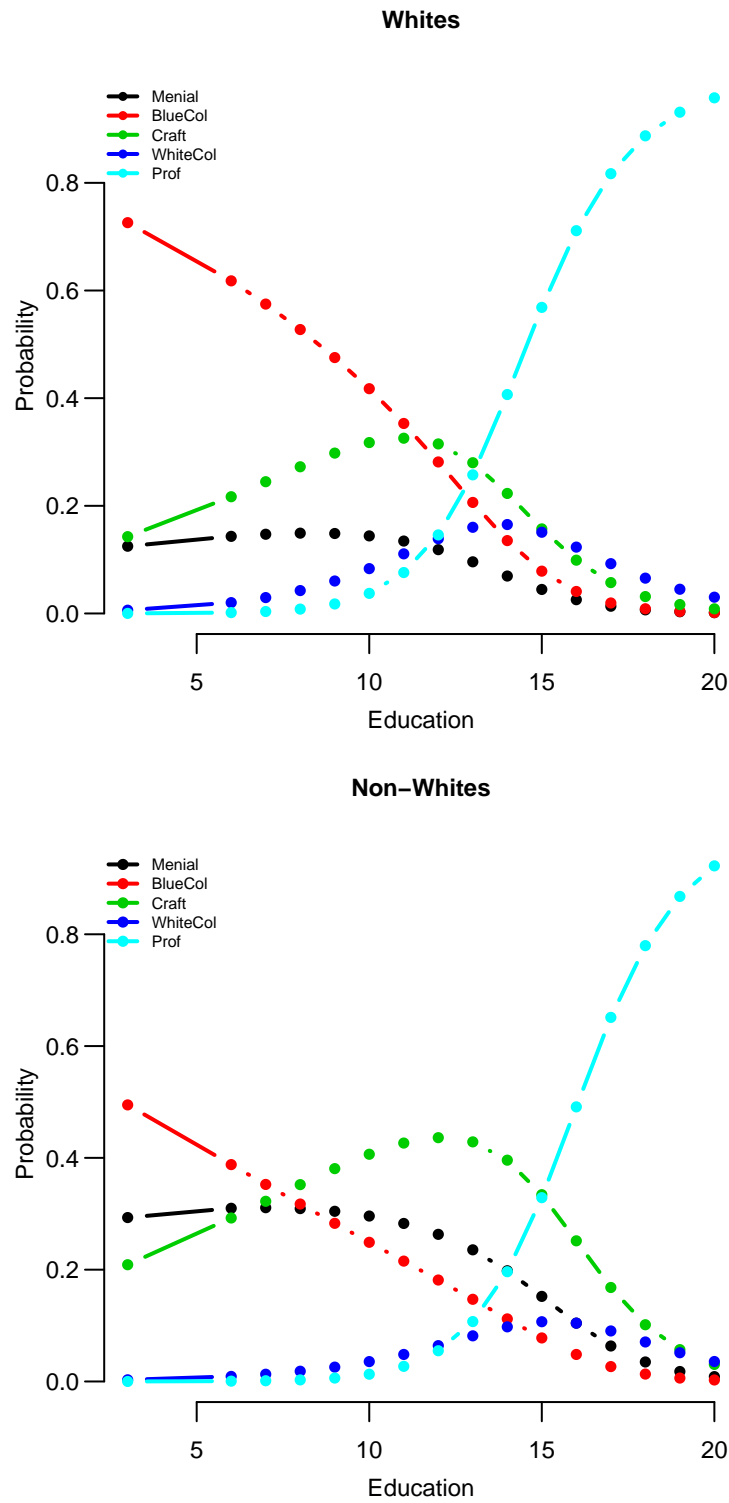


Figure 3: Predicted Probabilities of Occupational Attainment, by Race and Education, with 17 years of Experience

```

1 > source("helper.R")
2 > hitmiss(q2)

```

```

refitting the model to extract responses...
# weights: 25 (16 variable)
initial value 542.380576
iter 10 value 472.154249
iter 20 value 426.809681
final value 426.800478
converged

```

Table of Actual (y) Against Predicted (p)

Classification rule: outcome with highest probability.

	p=Menial	p=BlueCol	p=Craft	p=WhiteCol	p=Prof	Row	PCP
y=Menial	0	12	12	0	7		0.00
y=BlueCol	0	37	22	0	10		53.62
y=Craft	0	22	43	0	19		51.19
y=WhiteCol	0	12	14	0	15		0.00
y=Prof	0	7	16	0	89		79.46

Percent Correctly Predicted, Fitted Model: 50.15%

Percent Correctly Predicted, Null Model : 33.23%

**Answer:** The working model used here fits the data reasonably well, correctly classifying 50.1% of the data; a null model which assigned all respondents to the modal occupational status category would attain 33.3% correct classification. The model assigns no one to the two smallest categories, menial and white collar, but does a good job of predicting the professional outcome (79.5%) correctly predicted.

Include a table of coefficients and standard errors for your preferred model, along with any supporting tables demonstrating model comparison tests and/or goodness of fits. Your write up should also include some graphs demonstrating marginal effects, where appropriate.