Homework #4

Samuel Jones

4/1/2018

Instructions

Replicate the results of Gerfin (1996). Begin by reading section 3. The data can be obtained here:
library(AER) data("SwissLabor")

Write a short report that replicates the top part of Table I for Switzerland (page 327). The probit regression can be executed using the following form:

$$glm(y\ x, family = binomial(link = probit))$$

Here's a list of functions you might need:
* hist()
* table()
* gam() (mgcv package)
* npreg (np package)
* sims (arm package)

Make sure your report includes reproducible R code, either as an appendix, as a footnote through Harvard's Dataverse (or some other site of your choice), or through CMU's dropbox.

Problem 1a.

Conduct the probit regression reported in the paper. (1 point

```
#### Problem 1a - Table I ####
table1 <- glm(participation ~ age + I(age^2) + education + youngkids +
    oldkids + income + foreign, data = SwissLabor, family = binomial(link = "probit"))
```

Table 1: Probit estimation results (asymptotic standard errors in parentheses)

	Switzerland
(Intercept)	3.75^{**}
	(1.41)
age	2.08^{***}
	(0.41)
I(agesq)	-0.29^{***}
	(0.05)
education	0.02
	(0.02)
youngkids	-0.71^{***}
	(0.10)
oldkids	-0.15^{**}
	(0.05)
income	-0.67^{***}
	(0.13)
foreign(yes)	0.71^{***}
	(0.12)
AIC	1033.15
BIC	1071.32
Log Likelihood	-508.58
Deviance	1017.15
Num. obs.	872

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

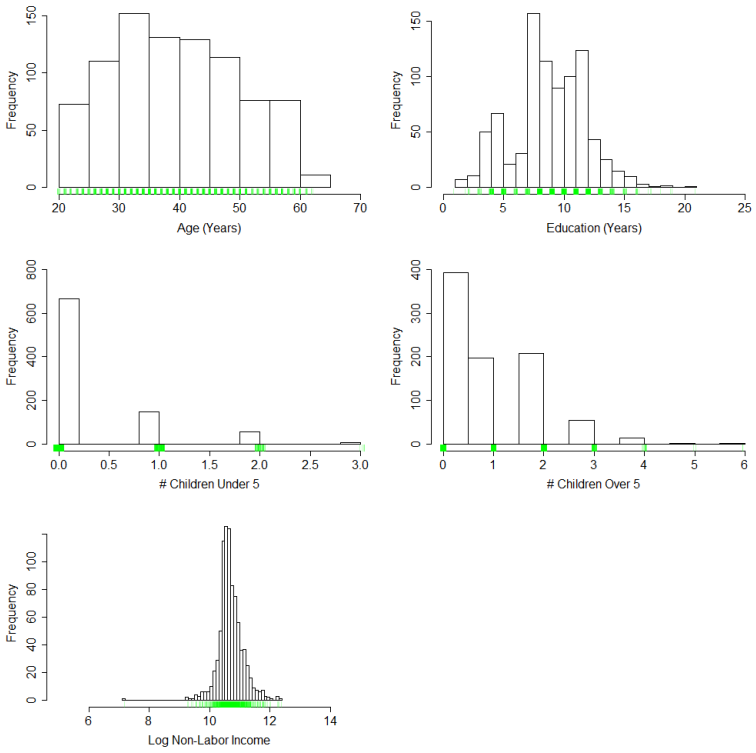
After attempting to reproduce the table, write a short report telling the five stories of these data. Use the probit model in Table I as a starting point, but use logistic regression or other tools (i.e., not probit). Here's what I expect to be included in the extension part of the report:

Problem 1b.

Create histograms of age, education, number of young children, log of yearly non-labor income, number of older kids, and tables of whether the woman participates in the labor force and whether she is a permanent foreign resident, summarizing what you see. (1 point)

```
#### Problem 1b - Histograms #### Based on Homework 3 solution
#### (Davis, 2018)
SwissLabor$agenorm <- SwissLabor$age * 10
png(filename = "HW4P1B Histograms.png", width = 1000, height = 1000,
    res = 100)
par(cex = 1.3, mar = c(5, 5, 2, 1), mfrow = c(3, 2), cex.lab = 1.5,
    cex.axis = 1.5)
hist(SwissLabor$agenorm, breaks = "FD", xlab = "Age (Years)", main = "",
```

```
    xlim = c(20, 70))
rug(jitter(SwissLabor$agenorm), col = rgb(0, 1, 0, 0.25))
hist(SwissLabor$education, breaks = "FD", xlab = "Education (Years)",
    main = "", xlim = c(0, 25))
rug(jitter(SwissLabor$education), col = rgb(0, 1, 0, 0.25))
hist(SwissLabor$youngkids, xlab = "# Children Under 5", main = "",
    ylim = c(0, 800))
rug(jitter(SwissLabor$youngkids, amount = 0.05), col = rgb(0, 1, 0,
    0.25))
hist(SwissLabor$oldkids, xlab = "# Children Over 5", main = "", ylim = c(0,
    400))
rug(jitter(SwissLabor$oldkids, amount = 0.05), col = rgb(0, 1, 0,
    0.25))
hist(SwissLabor$income, breaks = "FD", xlab = "Log Non-Labor Income",
    main = "", xlim = c(5, 15))
rug(jitter(SwissLabor$income), col = rgb(0, 1, 0, 0.25))
dev.off()
```



Histograms.bb

- We see that the ages of the women in this sample are fairly uniform over the plotted range of 20 to 60 years. The women's formal education is not uniform and there are high concentrations with local maxima 5, 8, and 12 years. 8 and 12 years probably correspond to middle and high school. The majority of women, ~75%, don't have children under the age of 5. A little over half of the women have children over the age of 5. The log income of the sample appears to be centered around 10.75 and is fairly normally distributed around that mean. There is at least one outlier around 7 ~ $e^7 = 1,096$ which is $\frac{1}{42}$ of the mean. This person is probably in extreme poverty.

```
kable(summary(SwissLabor$participation))
```

	x
no	471
yes	401

```
kable(summary(SwissLabor$foreign))
```

	x
no	656
yes	216

- From these two summary tables, we can see that about half of the women in the sample participated in the labor force which is great for our model. Most (~75%) of the women are not foreign residents.

Problem 2.

Write out the logit (log odds) regression equivalent of the model proposed in Table I (i.e., write the regression model in terms of a function that is linear in the log odds). Next, write the equivalent model in terms of the Bernoulli likelihood function using the logit link function. Explain how the logit function links the linear predictors to the conditional mean (probability). Explain whether it would be better or worse to just take the logit transform of the dependent variable directly. (1 point)

- Logit regression equivalent of previous model:

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1(age) + \beta_2(age^2) + \beta_3(education) + \beta_4(kids_{young}) + \beta_5(kids_{old}) + \beta_6(income) + \beta_7(foreign) + u$$

- Bernoulli likelihood function using logit link function

$$\eta(x) = log\left(\frac{p(x)}{1-p(x)}\right)$$

$$\eta(x) = \beta_0 + \beta_1(age) + \beta_2(age^2) + \beta_3(education) + \beta_4(kids_{young}) + \beta_5(kids_{old}) + \beta_6(income) + \beta_7(foreign) + u$$

- Then

$$L(p(x)|y_i) = \Pi_{i=1}^{n} p(x)^{y_i}(1-p(x))^{1-y_i} = \Pi_{i=1}^{n}\left(\frac{1}{1+e^{-\eta(x)}}\right)^{y_i}(1-frac11+e^{-\eta(x)})^{1-y_i}$$

- The logit link connect the unobserved conditional mean to the linear predictor. [Davis, 2018] It doesn't make sense to use the logit transform on the dependent variable because we would just end up with $\pm\infty$ for all observations.

Problem 3.

Conduct the logistic regression version of the model proposed in Table I. Interpret the logistic regression coefficients in terms of odds ratios. Note: Don't worry about interpreting the age coefficients. Plot the predicted probability of labor force participation against the woman's age, holding the other variables at their mean. (1 point)

```
#### Problem 3 - Logistic Regression ####
table2 <- glm(participation ~ agenorm + I(agenorm^2) + education +
    youngkids + oldkids + income + foreign, data = SwissLabor, family = binomial(link =
```

Table 4: Logistic Regression

	Switzerland
(Intercept)	6.20^{**}
	(2.38)
agenorm	0.34^{***}
	(0.07)
I(agenormSq)	-0.00^{***}
	(0.00)
education	0.03
	(0.03)
youngkids	-1.19^{***}
	(0.17)
oldkids	-0.24^{**}
	(0.08)
income	-1.10^{***}
	(0.23)
foreignyes	1.17^{***}
	(0.20)
AIC	1033.57
BIC	1071.74
Log Likelihood	-508.79
Deviance	1017.57
Num. obs.	872

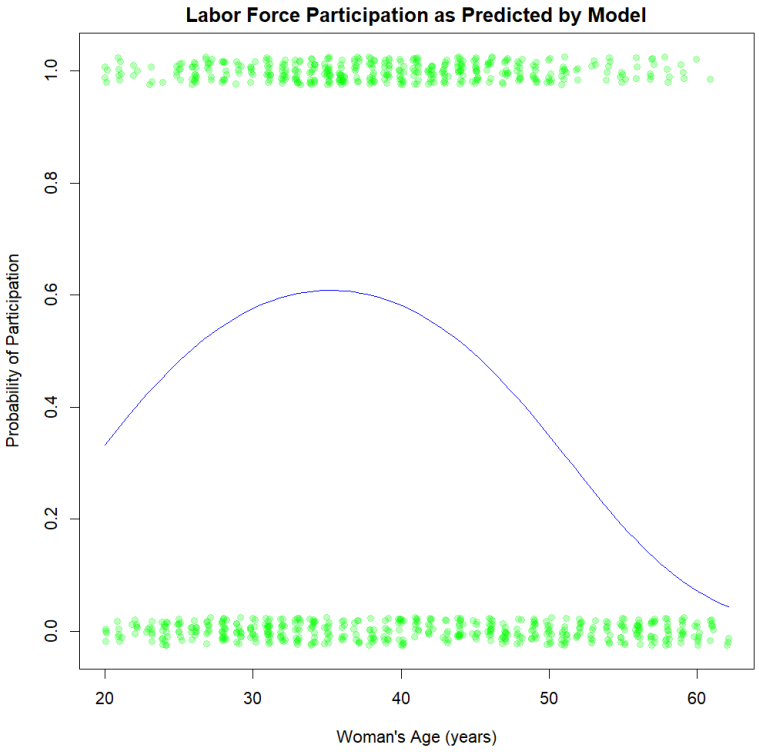
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

- Women with one additional year of education have an odds of participation that is $e^{0.032} = 1.03$ times greater than women without that additional year of education. Women with young children have an odds of $e^{-1.19} \approx 0.30$ times less likely to participate in the job market than women without young children. Women with older children have an odds of $e^{-0.24} \approx 0.78$ times less likely to participate in the job market than women without older children. Women with one additional log point of income have an odds of $e^{-1.1} \approx 0.33$ times less likely to participate in the job market than women without the additional log point of income. Leslty, foreign women have an odds of $e^{1.17} = 3.22$ times greater participation than non-foreign residents.

```
# Plot age convert non-numeric variables to binomial dummy
# variables
SwissLabor$biforn <- ifelse(SwissLabor$foreign == "yes", 1, 0)
SwissLabor$bipart <- ifelse(SwissLabor$participation == "yes", 1,
    0)
# Re-run the regression with the dummy variable; no change from
# table2
table3 <- glm(bipart ~ agenorm + I(agenorm^2) + education + youngkids +
    oldkids + income + biforn, data = SwissLabor, family = binomial(link = "logit"))
texreg(table3)
```

```r
# Create variables for mean levels of variables
meanedu <- mean(SwissLabor$education)
meanyoung <- mean(SwissLabor$youngkids)
meanold <- mean(SwissLabor$oldkids)
meanincome <- mean(SwissLabor$income)
meanforn <- mean(SwissLabor$biforn)
# Create an array with ages using the same range as the data
age <- seq(20, 60, by = 0.5)

# Create a plot of the coefficients vs. age based on (Davis, 2018)
png(filename = "HW4P3 Age Plot.png", width = 1000, height = 1000,
    res = 100)
par(cex = 1.3, mar = c(5, 4, 2, 1))
plot(jitter(SwissLabor$agenorm), jitter(SwissLabor$bipart, amount = 0.025),
    pch = 19, col = rgb(0, 1, 0, 0.25), main = "Labor Force Participation as Predicted b
    ylab = "Probability of Participation", xlab = "Woman's Age (years)")
curve(1/(1 + exp(-(coef(table3)[1] + coef(table3)[2] * x + coef(table3)[3] *
    x^2 + coef(table3)[4] * meanedu + coef(table3)[5] * meanyoung +
    coef(table3)[6] * meanold + coef(table3)[7] * meanincome + coef(table3)[8] *
    meanforn))), add = TRUE, col = rgb(0, 0, 1, 1))
dev.off()
```



Age Plot.bb

- As the plot shows, the likelihood of participation in the labor force as predicted by the model initially increases until peaking around age 35 at which point it declines.

Problem 4.

Create and examine a calibration plot and calibration table for the model proposed in Table 1. Does there seem to be model misspecification? Why or why not? (1 point)

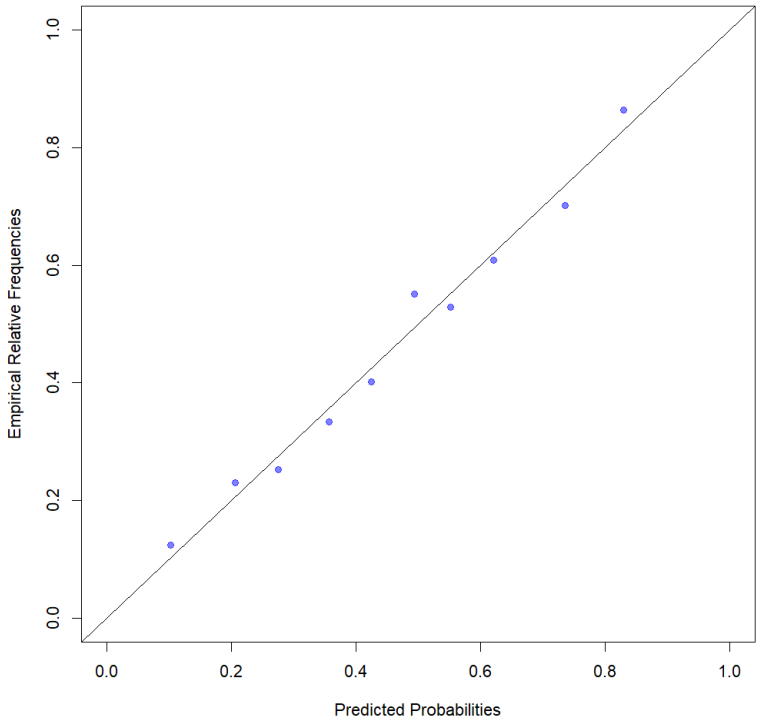
```
#### Problem 4 Calibration Table #### Based on Week10.r (Davis, 2018)
#### get the raw fitted probabilities from the first table
SwissLabor$probs <- predict(table1, type = "response")
# Get the deciles of the fitted probabilities
decile.cutpoints <- quantile(SwissLabor$probs, probs = seq(0, 1, 0.1))
# Create a new variable that identifies the quantile that each
# fitted probibility falls in
SwissLabor$decileID <- cut(SwissLabor$probs, breaks = decile.cutpoints,
    labels = 1:10, include.lowest = TRUE)
# Calculate the number that switched in each decile You can see
# the counts by using table:
tab <- table(SwissLabor$decileID, SwissLabor$participation)
# To turn the table into a data.frame, use as.data.frame.matrix
observed <- as.data.frame.matrix(tab)
# To calculate the expected values for each decile, we need to sum
```

```r
# the fitted probabilities for each decile We can do this by using
# tapply to compute the sum of wells$prob within each level of
# decileID:
expected1 <- tapply(SwissLabor$probs, SwissLabor$decileID, FUN = sum)
# Create a summary calibration table Create cutpoints that
# represent the 9 intervals that exist for deciles
interval.cutpoints <- round(quantile(SwissLabor$probs, probs = seq(0.1,
    1, 0.1)), 2)
# Create a dataframe with these cutpoints:
cal <- data.frame(interval.cutpoints)
# Add a column of observed switches
cal$observed1 <- observed[, 2]
# Add a column of expected switches
cal$expected1 <- round(expected1, 0)
# Add columns for observed and expected non-switches:
cal$observed0 <- observed[, 1]
cal$expected0 <- round(cal$observed1 + cal$observed0 - expected1,
    0)
# Add a column for the total # of observations in each decile
cal$total <- table(SwissLabor$decileID)
```

```
kable(cal)
```

	interval.cutpoints	observed1	expected1	observed0	expected0	total
10%	0.16	11	9	77	79	88
20%	0.24	20	18	67	69	87
30%	0.32	22	24	65	63	87
40%	0.39	29	31	58	56	87
50%	0.47	35	37	52	50	87
60%	0.52	48	43	39	44	87
70%	0.59	46	48	41	39	87
80%	0.67	53	54	34	33	87
90%	0.78	61	64	26	23	87
100%	0.94	76	73	12	15	88

```
# Calibration Plot Get the observed relative frequencies
freqs <- as.numeric(cal$observed1/cal$total)
# Get the predicted relative frequencies
probs <- as.numeric(cal$expected1/cal$total)
png(filename = "HW4P4 CalibrationPlot.png", width = 1000, height = 1000,
    res = 100)
par(cex = 1.3, mar = c(5, 4, 2, 1))
plot(probs, freqs, pch = 19, col = rgb(0, 0, 1, 0.5), ylab = "Empirical Relative Frequer
    xlab = "Predicted Probabilities", xlim = c(0, 1), ylim = c(0,
        1))
abline(0, 1)
dev.off()
```



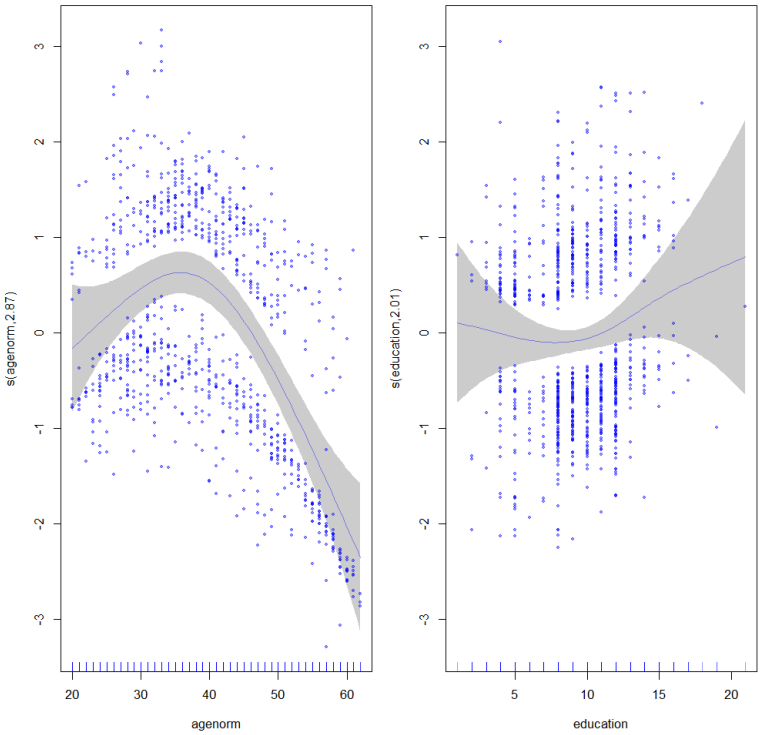
CalibrationPlot.bb

- Based on the table and plot it doesn't appear that there are any systematic misspecification issues. The predicted and observed are very close. Sometimes they are underpredicted and sometimes overpredicted, but all very close.

Problem 5.

Compare logistic regression and a generalized additive model with smoothing on age and education for the model proposed in Table I. Look at the partial residual plots from the GAM. Do any transformations seem necessary to the age or education variables? Why or why not? (1 point)

```
#### Problem 5 GAM ##### Based on Week67.r (Davis, 2018)
gam1 <- gam(participation ~ s(agenorm) + s(education) + youngkids +
    oldkids + income + foreign, data = SwissLabor, family = binomial(link = "logit"))
png(filename = "HW4P5 GAM Graph.png", width = 1000, height = 1000,
    res = 100)
par(cex = 1.3, mar = c(5, 4, 2, 1), mfrow = c(1, 2))
plot(gam1, residuals = TRUE, shade = TRUE, pch = 19, col = rgb(0,
    0, 1, 0.4), cex = 0.5)
dev.off()
```



GAM Graph.bb

- It is clear from these plots that Age is not monotonic. Our quadratic transformation seems very appropriate. Education isn't as clear cut. It may be necessary, but it also looks like it may be possible to use it without any transformation.

Problem 6.

Use Stukel's test for the logistic regression and previously suggested generalized additive model. Does the logit link function seem appropriate? Why or why not?

```
### Problem 6 Stukel's Test #### Based on Week 10.r and Section
### 9.5.3 (Davis, 2018)
eta.hat <- predict(table2)
positive <- ifelse(eta.hat >= 0, 1, 0)
negative <- ifelse(eta.hat < 0, 1, 0)
eta.hat.sq <- eta.hat^2

stukel1 <- glm(participation ~ agenorm + I(agenorm^2) + education +
    youngkids + oldkids + income + foreign + eta.hat.sq:positive +
    eta.hat.sq:negative, data = SwissLabor, family = binomial(link = "logit"))

logit <- glm(participation ~ agenorm + I(agenorm^2) + education +
    youngkids + oldkids + income + foreign, data = SwissLabor, family = binomial(link =

probit <- glm(participation ~ agenorm + I(agenorm^2) + education +
```

Table 6: Stukel's Test with Logit, Probit and Complementary Log-Log

	Stukel's Test	Logit	Probit	C-Log-Log
(Intercept)	4.54	6.20^{**}	3.75^{**}	4.06^{*}
	(2.78)	(2.38)	(1.41)	(1.68)
agenorm	0.26^{*}	0.34^{***}	0.21^{***}	0.25^{***}
	(0.11)	(0.07)	(0.04)	(0.05)
I(agenormSq)	-0.00^{**}	-0.00^{***}	-0.00^{***}	-0.00^{***}
	(0.00)	(0.00)	(0.00)	(0.00)
education	0.02	0.03	0.02	0.02
	(0.03)	(0.03)	(0.02)	(0.02)
youngkids	-0.90^{**}	-1.19^{***}	-0.71^{***}	-0.93^{***}
	(0.32)	(0.17)	(0.10)	(0.13)
oldkids	-0.18	-0.24^{**}	-0.15^{**}	-0.19^{**}
	(0.10)	(0.08)	(0.05)	(0.06)
income	-0.83^{*}	-1.10^{***}	-0.67^{***}	-0.77^{***}
	(0.34)	(0.23)	(0.13)	(0.15)
foreignyes	0.78^{*}	1.17^{***}	0.71^{***}	0.77^{***}
	(0.36)	(0.20)	(0.12)	(0.13)
eta.hat.sq:positive	0.35			
	(0.24)			
eta.hat.sq:negative	-0.09			
	(0.15)			
AIC	1034.99	1033.57	1033.15	1030.67
BIC	1082.70	1071.74	1071.32	1068.83
Log Likelihood	-507.49	-508.79	-508.58	-507.33
Deviance	1014.99	1017.57	1017.15	1014.67
Num. obs.	872	872	872	872

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 7: Statistical models

```
    youngkids + oldkids + income + foreign, data = SwissLabor, family = binomial(link =

cloglog <- glm(participation ~ agenorm + I(agenorm^2) + education +
    youngkids + oldkids + income + foreign, data = SwissLabor, family = binomial(link =

l <- list(stukel1, logit, probit, cloglog)
texreg(l)
```

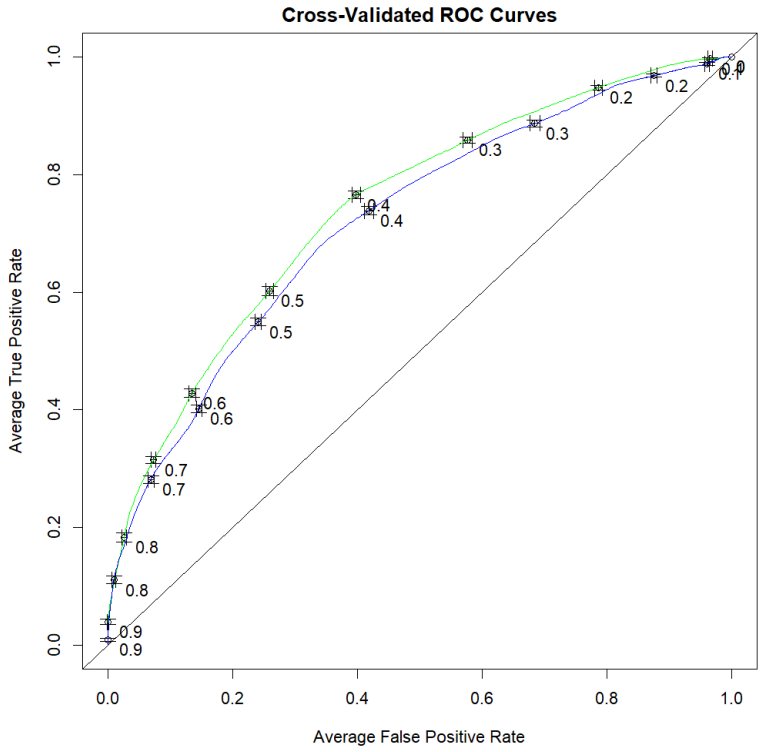
- The values are very similar for each of the three models. Based on the deviance, we might be slightly better off using hte complementary log-log model than the Stukel model. The logit model and probit models are not as good according to the deviance with logit being slightly worse than probit.

Problem 7.

Create a cross-validated ROC curve for the logistic regression with and without any transformations or interactions you think are necessary based on your previous work. Which model performs better in terms of cross-validation and which model do you prefer? Explain your reasoning. (1 point)

```
### Problem 7 ROC #### Based on Week 10.r (Davis, 2018) generate
### empty prediction and outcome vectors
predictions1 <- c()
predictions2 <- c()
labels <- c()
# logistic regression CV function
logit.cv <- function(data1 = SwissLabor, k = 5) {
    # select number of folds
    folds <- k
    # generate fold sequence
    fold.num <- rep(1:folds, length.out = nrow(data1))
    # randomize fold sequence
    fold.samp <- sample(fold.num)
    for (i in 1:k) {
        # Training data
        train <- data1[fold.samp != i, ]
        # Test data takes the remaining rows
        test <- data1[fold.samp == i, ]
        # Run glm on training data
        glm1 <- glm(participation ~ agenorm + I(agenorm^2) + education +
            youngkids + oldkids + income + foreign, data = train,
            family = binomial(link = "cloglog"))
        glm2 <- glm(participation ~ agenorm + education + youngkids +
            oldkids + income + foreign, data = train, family = binomial(link = "logit"))
        # Make probability predictions for test data
        glmpred1 <- predict(glm1, test, type = "response")
        glmpred2 <- predict(glm2, test, type = "response")
        # Add the predictions for this iteration to the data frame
        predictions1 <- append(predictions1, glmpred1)
        predictions2 <- append(predictions2, glmpred2)
        # Add the actual outcomes for this iteration to the data frame
        labels <- append(labels, test$participation)
    }
    return(list(predictions1, predictions2, labels))
}
cvdata1 <- replicate(100, logit.cv())
preds1 <- sapply(cvdata1[1, ], cbind)
preds2 <- sapply(cvdata1[2, ], cbind)
labs1 <- sapply(cvdata1[3, ], cbind)
```

```r
library(ROCR)
# Run the ROCR prediction and performance measures
glmerr1 <- prediction(preds1, labs1)
glmperf1 <- performance(glmerr1, measure = "tpr", x.measure = "fpr")
glmerr2 <- prediction(preds2, labs1)
glmperf2 <- performance(glmerr2, measure = "tpr", x.measure = "fpr")
# This gives a vector of AUCs
glmauc1 <- performance(glmerr1, measure = "auc")
glmauc2 <- performance(glmerr2, measure = "auc")
# Unlist the AUCs
glmauc1 <- unlist(glmauc1@y.values)
glmauc2 <- unlist(glmauc2@y.values)
# Take the average
glmauc1 <- mean(glmauc1)
glmauc2 <- mean(glmauc2)
# Plot the ROC curves:
png(filename = "HW4P7 ROC Plot.png", width = 1000, height = 1000,
    res = 100)
par(cex = 1.3, mar = c(5, 4, 2, 1))
# ROC curve for more complex model
plot(glmperf1, col = "green", main = "Cross-Validated ROC Curves",
    avg = "threshold", spread.estimate = "stddev", print.cutoffs.at = seq(0,
        0.9, by = 0.1), text.adj = c(-0.5, 1.2), xlab = "Average False Positive Rate",
    ylab = "Average True Positive Rate")
abline(0, 1)
# ROC curve for simpler model
plot(glmperf2, col = "blue", avg = "threshold", spread.estimate = "stddev",
    print.cutoffs.at = seq(0, 0.9, by = 0.1), text.adj = c(-0.5, 1.2),
    add = TRUE)
dev.off()
```



ROC Plot.bb

Figure 1: ROC Plot

- As expected the model with the quadratic performed better than the other. The more complicated model's ROC curve is always above the simpler model. Also the Area Under the Curve numbers are better for the more complicated model by about .023 it. (0.743 and 0.720) So, the more complicate model is better at forecasting and has a better explanatory ability.

Problem 8.

Briefly comment on the statistical inference and causality stories. Use simulations or confidence intervals if you desire. (1 point)

- The key assumption or factor with regard to the inference and causality stories is based upon the method used to collect the data. If the collection used anything less than a random sampling method, we can't really claim causality. So, after reviewing Gerfin's article, I was unable to find sufficient information to make any kind of informed diagnosis as to the true nater of this collection. [Gerfin, 1996] Therefore, I would have to abstain from making any causality claims.

References

Alex Davis. 19-703/4: Applied Data Analysis. 2018. 00000.

Michael Gerfin. Parametric and semi-parametric estimation of the binary response model of labour market participation. Journal of Applied Econometrics, 11(3):321□339, 1996. URL http://www.jstor.org/stable/2285067.