American University, Department of Government

Ph.D. Methodology Qualifying Exam, Summer 2020, due 5PM on August 20, answer 6 questions out of 8.

1. **LINEAR AND GENERALIZED LINEAR MODELS.** Suppose a researcher would like to estimate the relationship between some vector of covariates and a dichotomous outcome variable, Y.

   (a) If the researcher were to estimate this model using OLS, what would be the properties of the resulting estimators? Would OLS be a good or bad idea, and exactly why?

   (b) If the researcher were intent on using OLS, how might the researcher be able to fix some of the problems you discussed in the previous question? What problems would remain un-addressable?

   (c) Discuss, briefly, how logit and probit models address the shortcomings of using OLS when Y is dichotomous. Be sure to address the assumptions inherent in these models.

   (d) How would a researcher calculate predicted probabilities and marginal effects for these models, and why would a researcher want to do so?

   (e) How would a researcher report on (and calculate) the uncertainty surrounding the predicted probabilities and marginal effects? Why would a researcher want to do so?

   Answer:

   (a) The linear model does not work well in this situation. The resulting estimators are biased and predictions outside of $[0:1]$ are possible.

   (b) One can use truncated forms or the Constrained Linear-Probability Model. The unaddressed problem is in the interpretation of the estimated coefficients.

   (c) They should talk about the functional form of logit/probit and how this related to dichotomous outcomes.

   (d) They should discuss first differences for logit/probit and how the linear model can give misleading predictions.

   (e) Bootstrapping, simulation, etc. They should discuss the importance of standard errors.

2. **MAXIMUM LIKELIHOOD ESTIMATION.** Consider the following posited relationships:

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i \tag{1}$$

$$Var(\epsilon_i) = \exp(\mathbf{Z}_i\boldsymbol{\gamma})^2 \tag{2}$$

   Where $\mathbf{Y}_i$ is a dichotomous outcome variable, $\mathbf{X}$ is a vector of covariates, and $\mathbf{Z}$ is a vector of covariates. Consider the following estimator, built from these posited relationships:

$$\ell(\boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{n}\left[y_i \log\left(\Phi\left(\frac{\mathbf{X}\boldsymbol{\beta}}{e^{\mathbf{Z}\boldsymbol{\gamma}}}\right)\right) + (1-y_i)\log\left(1-\Phi\left(\frac{\mathbf{X}\boldsymbol{\beta}}{e^{\mathbf{Z}\boldsymbol{\gamma}}}\right)\right)\right]. \tag{3}$$

   Answer the following questions:

(a) Translate the relationships in (1) and (2) into everyday English, including an explanation of why you would use this specification. That is, how would you explain this set of posited relationships to an undergraduate?

(b) Describe one substantive political science application to which this set of relationships might pertain.

(c) What are the criteria that researchers should use in determining how "good" this new estimator is?

(d) What are the ways in which a researcher would go about determining how "good" this estimator is? Specify the analytic and computational methods that a researcher could use to evaluate the properties of this estimator (Note: you do not have to actually run these methods; just specify what they are).

Answer:

(a) This is the Harvey Heteroscedastic Probit model (they don't need to know that). They should know that the coefficients in the denominator are variance modeling. They should say something about probit in general.

(b) This is useful where there is heteroscedasticity and a dichotomous outcome.

(c) This is wide open and a gift. Look for something dumb here.

(d) They should discuss standard errors, AIC, predicted values, deviances, residual plots, etc.

3. **DATA ANALYSIS.** Using the following dataset construct a linear model (in `R`) with excellent fit properties where the outcome variable is `time`, and all three explanatory variables are included. You may employ any standard extensions or enhancements to the linear model, but not a GLM. Perform and report appropriate diagnostics, giving details. Submit your regression model, the diagnostics, and a one page explanation of the model results.

day	output	weight	time
1	55	5593	1738
1	20	2011	491
0	35	3574	999
1	45	4593	1370
0	40	4072	1150
1	25	2524	684
0	45	4576	1650
1	30	3034	876
1	60	6095	1910
1	45	4561	1380
0	35	3562	995
1	25	2516	660
1	45	4566	1390
1	35	3559	1025
1	30	3036	821

Answer: This model doesn't work unless there are interactions and transformations. Something like...

```
> lin.fit <- lm(time ~ day*output*log(weight),data=lin.dat)
> summary(lin.fit)

Call:
lm(formula = time ~ day * output * log(weight), data = lin.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-26.531  -3.048   0.000   1.724  28.218

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           82384.2    26215.5   3.143   0.0163
day                  -81411.5    26653.5  -3.054   0.0185
output                -2603.3     1011.6  -2.574   0.0368
log(weight)           -9769.3     4000.1  -2.442   0.0446
day:output             2700.7     1039.0   2.599   0.0355
day:log(weight)        9572.2     4077.3   2.348   0.0513
```

```
output:log(weight)        313.1      100.1   3.129   0.0166
day:output:log(weight)   -319.2      102.7  -3.107   0.0172

Residual standard error: 16.92 on 7 degrees of freedom
Multiple R-squared:  0.9992,Adjusted R-squared:  0.9984
F-statistic:  1231 on 7 and 7 DF,  p-value: 2.829e-10
```

The writeup should have the usual linear model diagnostics like Cook's D, residuals plots, etc. The explanation should correctly described the Wald statistics, R-square, F, and so on, without interpretation errors.

4. **CAUSAL INFERENCE.** The following results refer to the New Haven voter mobilization experiment, in which a random subset of the subject pool was assigned to be canvassed, but only some of those assigned to be canvassed were actually canvassed. The outcome is voter turnout. (8 points each)

Voter turnout by experimental group, New Haven voter mobilization experiment.

	Treatment Group	Control Group
Turnout rate among those contacted by canvassers	54.43 (395)	
Turnout rate among those not contacted by canvassers	36.48 (1,050)	37.54 (5,645)
Overall turnout rate	41.38 (1,445)	37.54 (5,645)

Note: Entries are percent voting, with number of observations in parentheses.
Sample restricted to households containing a single registered voter.

(a) Define a "Complier." Answer: A complier is a subject who takes treatment if and only if assigned to the treatment group. In this case, a complier is a subject who opens the door to a canvasser if and only if assigned to the treatment group.

(b) Estimate the proportion of Compliers in the subject pool. Answer: pr.compliers $\sim \frac{395}{1445} = 0.2734$. The proportion of compliers is 0.273.

(c) Show (with algebra) that under the assumptions of non-interference and excludability, the CACE (Complier-Average Causal Effect) is identified in this application. Answer: The CACE is defined as $E(Y_i(1) - Y_i(0)|D_i(1) = 1)$.

- Expected value of voting rate (Y) in the control group $=$ $E(Y_i(0)|D_i(1) = 1) * ITT_d + E(Y_i(0)|D_i(1) = 0) * (1 - ITT_d)$

- Expected value of voting rate in treatment group $=$
  $E(Y_i(1)|D_i(1)=1) * ITT_d + E(Y_i(0)|D_i(1)=0) * (1 - ITT_d)$

- Expected value of rate of successful canvassing $= E(D_i(1)) = ITT_d$

- Expected value of voting rate among treatment group minus voting rate of control group $=$
  $(E(Y_i(1)|D_i(1)=1) - E(Y_i(0)|D_i(1)=1)) * (ITT_d) = CACE * ITT_d$.

- $CACE = (E(Y_i(1)|D_i(1)=1) - E(Y_i(0)|D_i(1)=1))/(ITT_d)$.

(d) Are non-interference and excludability plausible in this example? Answer: Non-interference requires that the subjects respond only to their own treatment assignment, and not to the treatment assignment of others. This would be violated if, perhaps, neighbors called each other after being canvassed and talked about voting. The possibility of interference was explored experimentally by Sinclair, McConnell, and Green (2012) – Having treated neighbors does not appear to increase turnout. Excludability requires that nothing about assignment to canvassing itself affected potential outcomes, only the canvassing itself. This is plausible in this experiment.

(e) Estimate (by hand) the CACE. Provide a substantive interpretation of your estimate. Answer: itt $\sim .4138 - .3754$
ittd $\sim \frac{395}{1445}$
cace $\sim \frac{itt}{ittd}$
The $\widehat{CACE}$ is 0.14, meaning that compliers are 14 percentage points more likely to vote as a result of canvassing.

5. **RESEARCH DESIGN.** Imagine that you want to do a empirical study of sentencing procedure in the county courts of a state with 100 counties, 10 of which are urban. Your expectation is that the sentences should be less severe in urban counties. The problem is that gathering the data is quite laborious, and it requires that you spend a week pouring through the records in each of the counties. Hence, you decide that you can only afford to include 20 counties in the analysis. How would you do the study?

   Answer: There is a lot of lattitude on this question. Mark them down for saying something dumb or leaving out something obvious.

6. **STATISTICAL COMPUTING**

Consider the bivariate normal PDF:

$$f(x_1, x_2) = \left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\right)^{-1} \times \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right].$$

for $-\infty < \mu_1, mu_2 < \infty$, $\sigma_1, \sigma_2 > 0$, and $\rho \in [-1:1]$.

For $\mu_1 = 3, \mu_2 = 2, \sigma_1 = 0.5, \sigma_2 = 1.5, \rho = 0.75$, calculate a grid search using `R` for the mode of this bivariate distribution on $\mathbb{R}^2$. A grid search bins the sample space into equal space intervals on each axis and then systematically tests each resulting subspace. First setup a two dimensional coordinate system stored in a matrix covering 99% of the support of this bivariate density, then do a systematic search for the mode without using "for" loops. Hint: see the `R` help menu for the `apply` function. There are other approaches to not using for loops as well. Use the `contour` function to make a figure depicting bivariate contours lines at 0.5, 0.9, and 0.95 levels.

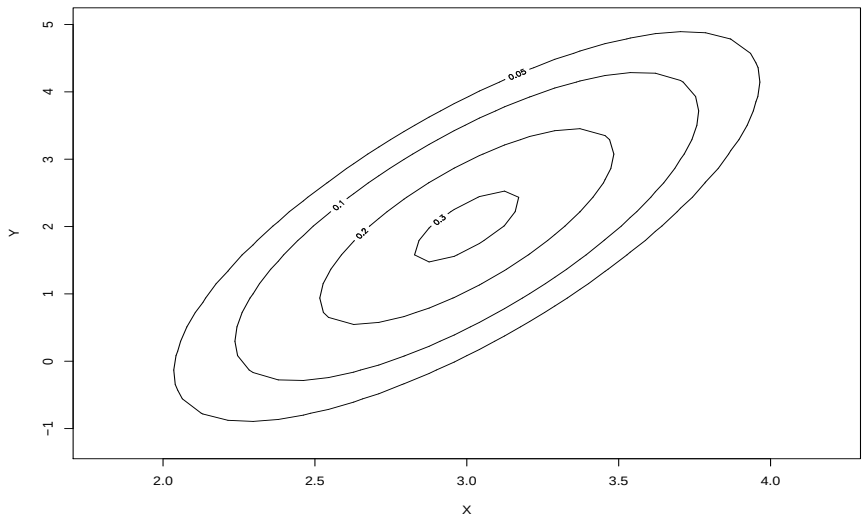
Answer:

```
dmultnorm <- function(x,y,mu1,mu2,sigma.mat)  {
      rho <- sigma.mat[1,2]/prod(sqrt(diag(sigma.mat)))
      nlizer <- 1/(2*pi*prod(sqrt(diag(sigma.mat)))*sqrt(1-rho^2))
      e.term1 <- (x - mu1)/sqrt(sigma.mat[1,1])
      e.term2 <- (y - mu2)/sqrt(sigma.mat[2,2])
      like <- exp( -1/(2*(1-rho^2)) *
                  (e.term1^2 + e.term2^2 - 2*rho*e.term1*e.term2) )
      nlizer*like
}

x.ruler <- seq(1.8,4.2,length=30); y.ruler <- seq(-1.2,5,length=30)
xy.cov.mat <- matrix(c(0.5^2,0.75*0.5*1.5,0.75*0.5*1.5,1.5^2),2,2)
xy.grid <- outer(x.ruler,y.ruler,dmultnorm,3,2,xy.cov.mat)
contours <- c(0.05,0.1,0.2,0.3)
contour(x.ruler,y.ruler,xy.grid,levels=contours,xlab="X",ylab="Y", cex=2)
```

7. **CODE.**

NBA playoff series are played in a "best of 7" format, where two teams play one another repeatedly until one team wins four games. Assume that two teams A and B have an equal probability to win each game in a best of 7 series. Write code to computationally estimate the probability that the series between A and B will go to 7 games. What happens if the two teams are not equal? Add lines to your code that will



estimate the probability that a series between A and B goes to 7 games if the probability that team A wins each game goes from 50% to 100% at intervals of 5 percentage points. Within each scenario, assume the probability of team A winning each game is the same as the probability that team A wins each other game.

```
strengths <- seq(.5,1,.05) p7 <- c()
for (strength in strengths) {
  total <- 100000
  sims <- rep(NA, total)

  for (ii in seq_len(total)) {
    a <- 0
    b <- 0
    games <- 0

    while (a < 4 && b < 4) {
      c <- runif(1,0,1)
      if (c > strength) {
        a <- a+1;
      } else {
        b <- b+1
      }
      games <- games + 1
```

```
    }
    sims[ii] <- games

  }
  prob <- sum(sims==7)/length(sims)
  p7 <- c(p7, prob)

}

names(p7) <- strengths
p7
```

8. **NON-LINEAR MODELING.** Consider the following R session:

```
X <- matrix(NA,32,3)
X[,1] <- c(2.66,2.89,3.28,2.92,4,2.86,2.76,2.87,3.03,3.92,2.63,
       3.32,3.57,3.26,3.53,2.74, 2.75,2.83,3.12,3.16,2.06,
       3.62,2.89,3.51,3.54,2.83,3.39,2.67,3.65,4,3.10,2.39)
X[,2] <- c(20,22,24,12,21,17,17,21,25,29,20,23,23,25,26,19,
        25,19,23,25,22,28,14,26,24,27,17,24,21,23,21,19)
X[1:18,3] <- 0
X[19:32,3] <- 1
#X <- cbind(rep(1,32),X)
Y <- c(0,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0,0,1,0,1,0,0,1,1,1,0,1,1,0,1)
dimnames(X)[[2]] <- list("GPA","TUCE","PSI")

apply(X,2,summary)

            Constant   GPA  TUCE    PSI
    Min.           1 2.060 12.00 0.0000
    1st Qu.        1 2.813 19.75 0.0000
    Median         1 3.065 22.50 0.0000
    Mean           1 3.117 21.94 0.4375
    3rd Qu.        1 3.515 25.00 1.0000
    Max.           1 4.000 29.00 1.0000


spector.logit <- glm(formula = Y ~ X, family=binomial(link=logit))
summary.glm(spector.logit)

    Deviance Residuals:
        Min       1Q   Median       3Q      Max
    -1.9551  -0.6453  -0.2570   0.5888   2.0966

    Coefficients:
                Estimate Std. Error z value Pr(>|z|)
    (Intercept) -13.02119    4.91635  -2.649  0.00808
    GPA           2.82609    1.26051   2.242  0.02496
    TUCE          0.09516    0.14135   0.673  0.50083
    PSI           2.37866    1.06242   2.239  0.02516
    ---

    Null deviance: 41.183  on 31  degrees of freedom
    Residual deviance: 25.779  on 28  degrees of freedom
    AIC: 33.779
```

Now answer the following questions. You do not need to re-run the model.

(a) Write the likelihood function for this model and give its value at the MLE.

(b) Briefly comment on the quality of this model using Wald tests.

(c) Calculate the first differences:

- for `GPA` across the interquartile range,
- for `TUCE` across the interquartile range,
- for `PSI`.

(d) Prove that using the linear probability model for these data violates the standard assumptions.

Answer:

(a) Something along the lines of...

$$\begin{aligned} L(\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}) &= \prod_{y_i=0} [1 - F(\mathbf{X}_i\boldsymbol{\beta})] \prod_{y_i=1} [F(\mathbf{X}_i\boldsymbol{\beta})] \\ &= \prod_{i=1}^{n} [1 - F(\mathbf{X}_i\boldsymbol{\beta})]^{1-y_i} [F(\mathbf{X}_i\boldsymbol{\beta})]^{y_i} \\ \ell(\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}) &= \sum_{i=1}^{n} [(1-y_i)\log(1 - F(\mathbf{X}_i\boldsymbol{\beta})) + y_i \log(F(\mathbf{X}_i\boldsymbol{\beta}))] \end{aligned}$$

with $F()_i = 1/[1 + \exp(\mathbf{X}_i\boldsymbol{\beta})]$.

(b) Only TUCE fails.

(c) Calculating (take the difference in values below)...

```
ilogit <- function(Xb)  1/(1+exp(-Xb))
spector.mean.vec <- apply(X,2,mean)
gpa.25 <- c(1,quantile(X[,1],0.25),spector.mean.vec[2:3])
gpa.75 <- c(1,quantile(X[,1],0.75),spector.mean.vec[2:3])
( pred.25 <- ilogit(gpa.25 %*% spector.logit$coefficients) )
[1,] 0.1251315
( pred.75 <- ilogit(gpa.75 %*% spector.logit$coefficients) )
[1,] 0.5101576
tuce.25 <- c(1,spector.mean.vec[1],quantile(X[,2],0.25),spector.mean.vec[3])
tuce.75 <- c(1,spector.mean.vec[1],quantile(X[,2],0.75),spector.mean.vec[3])
( pred.25 <- ilogit(tuce.25 %*% spector.logit$coefficients) )
[1,] 0.2155509
( pred.25 <- ilogit(tuce.75 %*% spector.logit$coefficients) )
[1,] 0.3116951
psi.0 <- c(1,spector.mean.vec[1:2],0)
psi.1 <- c(1,spector.mean.vec[1:2],1)
( pred.0 <- ilogit(psi.0 %*% spector.logit$coefficients) )
[1,] 0.1067571
```

```
( pred.1 <- ilogit(psi.1 %*% spector.logit$coefficients) )
[1,] 0.5632555
```

(d) Wrong distributional implications:

$$Y_i = \alpha + \beta x_i + \epsilon \;\Rightarrow\; \pi_i = \alpha + \beta x_i, \tag{4}$$

but since $Y_i \in \{0, 1\}$, then ϵ_i is dichotomous not normally distributed:

$$\epsilon_i = 1 - \mathbb{E}[Y_i] = 1 - (\alpha + \beta x_i) = 1 - \pi_i$$

$$\epsilon_i = 0 - \mathbb{E}[Y_i] = 0 - (\alpha + \beta x_i) = -\pi_i$$

The expectation is okay:

$$\mathbb{E}[\epsilon_i] = \mathbb{E}[Y_i - \alpha - \beta x_i] = \pi_i - \pi_i = 0, \tag{5}$$

but the variance is wrong:

$$\mathsf{Var}[\epsilon_i] = \mathbb{E}[\epsilon_i^2] - (\mathbb{E}[\epsilon_i])^2 = \mathbb{E}[\epsilon_i^2], \tag{6}$$

which turns out to be:

$$\begin{aligned}
\mathbb{E}[\epsilon_i^2] &= \sum_{i=0}^{1} \epsilon_i^2 p(\epsilon_i) \\
&= (1 - \pi_i)^2(\pi_i) + (-\pi_i)^2(1 - \pi_i) \\
&= (1 - \pi_i)[(1 - \pi_i)(\pi_i) + \pi_i^2] \\
&= (1 - \pi_i)\pi_i \\
&= \pi_i - \pi_i^2.
\end{aligned}$$

This is a quadratic form and is therefore heteroscedastic, especially near zero and one.