Economics 217 **Name:**_____
Exam #1

Instructions. Closed book and notes, 95 minutes. Please directly answer on the exam paper. Partial credit will be granted for brief, relevant remarks and for partial results, but not unrelated equations and text from memory. There are 5 questions, 10 points each.

Problem 1

The binomial distribution is the basis for the logit and probit models that evaluate dichotomous variables. The PDF of the binomial distribution is written as:

$$f(y;p) = \binom{n}{y} p^y (1-p)^{n-y}$$

where y is the number of times the event has occurred after n trials, and p is the parameter of interest. Using (1), please show that the binomial distribution is a member of the canonical exponential family. Please report the functions $b(p)$, $c(p)$, and $d(y)$, and use these functions to compute the mean and variance of the Binomial distribution.

Answer:

p is the parameter of interest. Rearranging, we get:

$$f(y;p) = \exp\left[y\log(p) + \log(1-p)(n-y) + \log\binom{n}{y}\right]$$

Collecting y's

$$f(y;p) = \exp\left[y\log\left(\frac{p}{1-p}\right) + n\log(1-p) + \log\binom{n}{y}\right]$$

Thus,

$$\begin{aligned} b(p) &= \log\left(\frac{p}{1-p}\right) \\ c(p) &= n\log(1-p) \\ d(y) &= \log\binom{n}{y} \end{aligned}$$

Using the formulas from the lecture, the expected value of the binomial distribution is simplified as:

$$
\begin{aligned}
E(y) &= -\frac{c'(p)}{b'(p)} \\
&= -\frac{-\frac{n}{1-p}}{\frac{1}{p}+\frac{1}{1-p}} \\
&= -\frac{-\frac{n}{1-p}}{\frac{1-p}{p(1-p)}+\frac{p}{p(1-p)}} \\
&= np
\end{aligned}
$$

For variance, we first need to compute second derivatives:

$$
\begin{aligned}
b''(p) &= \frac{d}{dp}\frac{1}{(p(1-p))} \\
&= -\frac{1-2p}{(p(1-p))^2}
\end{aligned}
$$

$$
\begin{aligned}
c''(p) &= \frac{d}{dp}\left(-\frac{n}{1-p}\right) \\
&= -\frac{n}{(1-p)^2}
\end{aligned}
$$

$$
\begin{aligned}
V(y) &= -\frac{b''(p)E(y)+c''(p)}{b'(p)^2} \\
&= -\frac{-\frac{1-2p}{(p(1-p))^2}np-\frac{n}{(1-p)^2}}{\left(\frac{1}{(p(1-p))}\right)^2} \\
&= \frac{\frac{1-2p}{(p(1-p))^2}np+\frac{n}{(1-p)^2}}{\left(\frac{1}{(p(1-p))}\right)^2} \\
&= (1-2p)\,np+np^2 \\
&= np-2np^2+np^2 \\
&= np-np^2 \\
&= np(1-p)
\end{aligned}
$$

Problem 2

In this problem, we have N individuals, indexed by i, each with an observation for some variable y_i. Further, suppose that these y_i's originate from a Binomial distribution as written in question one. Further, we adopt an intercept model, where $g(\mu_i) = \beta_0 + \beta_1 x_i$. Assuming a *logit link*, please (1) write down the log-likelihood function, and (2) write the scoring functions for β_0 and β_1.
(Note: it is helpful to write $n = 1$, where $y \in \{0, 1\}$)

Answer To begin, the likelihood function is written as:

$$L = \prod_{i=1}^{N} f\left(y_i; p\right)$$

The Log-likelihood function, $l = \log(L)$, is

$$l = \sum_{i=1}^{N} \log\left(f\left(y_i; p\right)\right)$$

Within the canonical exponential family,

$$l = \sum_{i=1}^{N} \left[y_i b\left(p_i\right) + c\left(p_i\right) + d\left(y_i\right)\right]$$

The step that makes this question tolerable is to note that, after substituting that $n = 1$, we have:

$$b(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i \quad (1)$$

$$c(p_i) = \log(1-p_i) = -\log\left(1 + \exp\left(\beta_0 + \beta_1 x_i\right)\right) \quad (2)$$

Substituting into the log-likelihood function, we have:

$$l = \sum_{i=1}^{N} \left[y_i\left(\beta_0 + \beta_1 x_i\right) - \log\left(1 + \exp\left(\beta_0 + \beta_1 x_i\right)\right) + d\left(y_i\right)\right]$$

Differentiating with respect to β_0, we have:

$$U_0 \equiv \frac{dl}{d\beta_0} = \sum_{i=1}^{N} \left(y_i - \frac{\exp\left(\beta_0 + \beta_1 x_i\right)}{1 + \exp\left(\beta_0 + \beta_1 x_i\right)}\right)$$

Students may leave the answer here, or noting that $p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$, I will also accept:

$$\boxed{U_0 = \sum_{i=1}^{N} \left(y_i - p_i\right)}$$

Next, with respect to β_1, we have:

$$U_1 \equiv \frac{dl}{d\beta_1} = \sum_{i=1}^{N} \left(y_i x_i - \frac{\exp\left(\beta_0 + \beta_1 x_i\right)}{1 + \exp\left(\beta_0 + \beta_1 x_i\right)} x_i \right)$$

Again, one may leave the answer here. Or simplifying further, we have:

$$\boxed{U_1 \equiv \frac{dl}{d\beta_1} = \sum_{i=1}^{N} \left(y_i - p_i\right) x_i}$$

Problem 3

Using the Org dataset, we have run a number of regressions with *hourslw*, the hours worked, as reported by the respondent. For those in the labor force, we are interested in evaluating the relationship between educational attainment and hours worked, conditional on a flexible function of age and gender. To accomplish this, we run the following code:

```
> library(foreign)
> d<-read.dta("org_example.dta")
> ds<-subset(d,year==2013&nilf==0)
> ds$hourslw<-ifelse(is.na(ds$hourslw),0,ds$hourslw)
>
> glm1<-glm(hourslw~educ+female+age+I(age^2),ds,family=gaussian)
> summary(glm1)

Call:
glm(formula = hourslw ~ educ + female + age + I(age^2), family = gaussian,
    data = ds)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-42.573   -5.763    2.914    8.469  103.721

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.3630924  0.4450047   7.557 4.15e-14 ***
educHS            5.0476257  0.2139980  23.587  < 2e-16 ***
educSome college  5.6799652  0.2135217  26.601  < 2e-16 ***
educCollege       8.0570313  0.2220187  36.290  < 2e-16 ***
educAdvanced      8.2903120  0.2443321  33.931  < 2e-16 ***
female           -5.0581829  0.1043999 -48.450  < 2e-16 ***
age               1.3429915  0.0208131  64.526  < 2e-16 ***
I(age^2)         -0.0145830  0.0002341 -62.285  < 2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for gaussian family taken to be 271.1558)

    Null deviance: 29680410  on 100358  degrees of freedom
Residual deviance: 27210755  on 100351  degrees of freedom
AIC: 847097

Number of Fisher Scoring iterations: 2
```

a.) There are six lines of code here. Please describe briefly what each line accomplishes.

```
library(foreign)
```

This loads the foreign library, which is used to load stata files

```
d<-read.dta("org_example.dta")
```

This loads the stata file and assigns it to object 'd'

```
ds<-subset(d,year==2013&nilf==0)
```

This creates a subset of the data frame for year 2013 and only individuals who are in the labor force.

```
ds$hourslw<-ifelse(is.na(ds$hourslw),0,ds$hourslw)
```

This replaces missing values of hours worked with a zero.

```
glm1<-glm(hourslw~educ+female+age+I(age^2),ds,family=gaussian)
```

This runs a normal regression model with identity link.

```
summary(glm1)
```

This summarizes the output.

b.) Please interpret the coefficient educAdvanced

Holding age and gender constant, a person with an advanced degree works 8.29 hours per week more than somebody with less than a high school education.

Problem 4

Using the subsetted data from problem 3, we run a different model:

```
> glm2<-glm(hourslw~educ+female+age+I(age^2),ds,family=poisson(link="log"))
> summary(glm2)

Call:
glm(formula = hourslw ~ educ + female + age + I(age^2), family = poisson(link = "log"),
    data = ds)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-9.327  -1.054   0.494   1.445  13.064

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.522e+00  5.136e-03  491.07   <2e-16 ***
educHS            1.725e-01  2.425e-03   71.15   <2e-16 ***
educSome college  1.915e-01  2.421e-03   79.09   <2e-16 ***
educCollege       2.577e-01  2.475e-03  104.11   <2e-16 ***
educAdvanced      2.632e-01  2.669e-03   98.61   <2e-16 ***
female           -1.472e-01  1.083e-03 -135.90   <2e-16 ***
age               4.293e-02  2.347e-04  182.94   <2e-16 ***
I(age^2)         -4.667e-04  2.645e-06 -176.45   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1213941  on 100358  degrees of freedom
Residual deviance: 1139149  on 100351  degrees of freedom
AIC: 1625805

Number of Fisher Scoring iterations: 5
```

a.) Please re-interpret the coefficient on Advanced.

Holding age and gender constant, a person with an advanced degree works 26.3% more hours per week than somebody with less than a high school education.

b.) I now wish to test the model in problem 4 against an alternative without education dummy factor variables. Please write precisely the code you would use to execute such a test. What does each line do, mechanically (eg. code does X) and intuitively (eg. We are testing for Y because...)

There are two ways to answer this question. The easiest is:

```
library(lmtest)
lrtest(glm2,"educ")
```

The first line loads the lmtest library, which facilitates the likelihood ratio test. The second line calls the original regression, but instructs it to test agains the restricted model in which all education variables are removed. Intuitively, we are imposing restrictions and seeing how the log-likelihood changes. If the likelihood falls a sufficient amount, as determined by a chi-square distribution, we reject the restrictions in favor of the larger model.

The second way to answer this is the following.

```
glm3<-glm(hourslw~female+age+I(age^2),ds,family=poisson(link="log"))
LR<-(glm3$deviance-glm2$deviance)
pchisq(LR, 4, lower.tail = FALSE)
```

The first line runs a model with the restrictions imposed, the second line calculates the likelihood ratios statistic, and the third line calculates the p-value for inferrence.

Problem 5

Suppose that we wish to measure the impact of cracker prices on the choice between different cracker brands: Keebler, Nabisco, Sunshine, and Private label. To do so, we use the multinomial logit (MNL), based on the following specificiation:

$$\log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = \beta_j + \alpha \log(p_j) \text{ for each } j \ , \quad s.t. \quad \sum_j \pi_{ij} = 1$$

where π_{ij} is the probability that consumer i buys product j, π_{i1} is the probability that consumer i buys some reference category, 1, β_j is a product specific constant, and α is a common coefficient on price of brand j, $\log(p_j)$. To estimate this MNL, we use the "Cracker" data from class and run the following:

```
Call:
mlogit(formula = choice ~ price, data = data_cracker, method = "nr",
    print.level = 0)

Frequencies of alternatives:
 keebler  nabisco  private sunshine
0.068714 0.543934 0.314685 0.072666

nr method
5 iterations, 0h:0m:0s
g'(-H)^-1g = 2.32E-05
successive function values within tolerance limits

Coefficients :
                      Estimate Std. Error  t-value  Pr(>|t|)
nabisco:(intercept)   1.971844   0.071107  27.7307 < 2.2e-16 ***
private:(intercept)  -0.057197   0.116611  -0.4905    0.6238
sunshine:(intercept) -0.516202   0.100238  -5.1498 2.608e-07 ***
log(price)           -3.046660   0.172840 -17.6271 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -3354.6
McFadden R^2:  0.046404
Likelihood ratio test : chisq = 326.48 (p.value = < 2.22e-16)
```

Please use a simple **derivation** to **interpret** the coefficient on log price.

Differentiating the log-odds ratio with respect to p_j

$$\frac{d\pi_{ij}}{\pi_{ij}} - \frac{d\pi_{i1}}{\pi_{i1}} = \alpha \frac{dp_j}{p_j}$$

Thus:

$$\alpha = \frac{\frac{d\pi_{ij}}{\pi_{ij}} - \frac{d\pi_{i1}}{\pi_{i1}}}{\frac{dp_j}{p_j}}$$

The interpretation is that the percent change in probability for some variety j relative to the reference group changes by α with a percent change in the price of j. Put differently, the price elasticity of probability j relative to the reference group is α, or -3.05.