Assignment 1 Submission

Aaryaman Aggarwal, 230020 Manya Dixit, 230

August 28, 2025

Question 1

a) A descriptive summary of the mean, median, standard deviation, maximum, and minimum are as follows:

Variable Metrics	IgG	Age
Mean	5.29	2.77
Median	5.00	2.58
Standard Deviation	2.28	1.66
Maximum	14.40	6.00
Minimum	0.90	0.50

The histogram is as follows:

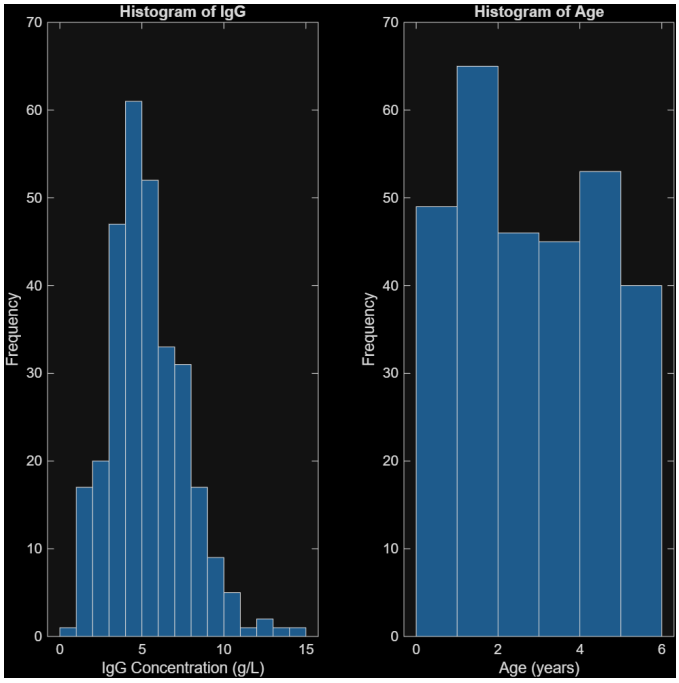

Figure 1: Histogram of IgG(Left) and Age(Right)

The IgG histogram appears to be left skewed, whereas the histogram for age doesn't appear

to be skewed. This should imply that Mean < Median for IgG but from the values we can see that isn't the case as few high values of IgG push the mean to be higher.

The histogram for age seems to be uniformly distributed and there appears to be no skewness.

b) The *Model* 1:

$$y_i = \beta_1 + \beta_2 x + u_i \tag{1}$$

The following are the estimates for the model using Ordinary Least Squares which was given as follows:

Coefficients	Estimates	Standard Errors	t-statistic
β_1 (Intercept)	3.3640	0.2214	15.194
β_2	0.6951	0.068469	10.152

$$R^2 = 0.258$$

From the coefficient estimates we can see that age has a positive effect of IgG, implying that as age increases, the IgG value(g/L) also increases. The low value of R^2 however indicates that the model doesn't fit very well and explains a very little amount of the variability of IgG.

c) The *Model* 2:

$$y_i \quad = \quad \beta_1 \quad + \quad \beta_2 x \quad + \quad \beta_3 x^2 \quad + \quad \varepsilon_i \tag{2}$$

The following are the estimates of the model using Ordinary Least Squares:

Coefficients	Estimates	Standard Errors	t-statistic
β_1 (Intercept)	3.0839	0.3833	8.0458
β_2	0.9674	0.3118	3.1029
β_3	-0.0454	0.05077	-0.8954

$$R^2 = 0.26$$

In the fit for Model 2 as well, we can see it is evident that there is a positive correlation between age and IgG but the coefficient for age^2 is negative with a very low t-statistic. Additionally, the R^2 is only slightly greater than Model 1 and explains around the same amount of variability as model 1.

d) In this it is said we can assume that

$$\varepsilon \sim N(0,1) \implies y_i - (\beta_1 + \beta_2 x_i + \beta_3 x_i^2) \sim N(0,1) \tag{3}$$

From this assumption we can write the conditional distribution of $y_i|x_i$ as:

$$y_i|x_i \sim N(\beta_1 + \beta_2 x_i + \beta_3 x_i^2, 1) \tag{4}$$

Therefore the PDF of y_i can be written as:

$$f(y_i|x_i,\beta) = \frac{1}{\sqrt{2\pi}}\ exp(\frac{-(y_i-(\beta_1+\beta_2 x_i+\beta_3 x_i^2))}{2}\ )$$

Therefore the log likelihood is:

$$ln\ L(\beta)\ = \sum_{i=1}^{n} ln\ f(y_i|x_i,\beta) \tag{5}$$

$$ln\ L(\beta)\ = \sum_{i=1}^{n} [-\frac{1}{2}ln(2\pi)\ -\ \frac{(y_i-\beta_1-\beta_2 x_i-\beta_3 x_i^2)^2}{2}] \tag{6}$$

$$ln\ L(\beta)\ =\ -\frac{n}{2}ln(2\pi)\ -\ \frac{1}{2}\sum_{i=1}^{n}\ \frac{(y_i-\beta_1-\beta_2 x_i-\beta_3 x_i^2)^2}{2}] \tag{7}$$

The model estimates using the Maximum Likelihood Method (Using Newton Rhapson) is as follows:

Maximum Likelihood Estimates

Coefficients	Estimates	Standard Errors	t-statistic
β_1 (Intercept)	3.0839	0.3833	8.0458
β_2	0.9674	0.3118	3.1029
β_3	-0.0454	0.05077	-0.8954

Question 2

a) For the response variable $q85$ coded as y_i

Variable	Frequency Count	Percentages
$y_i = 1$ (Yes, legal)	659	53.1452%
$y_i = 0$ (No, illegal)	581	46.8548%

b) For the variable 'age' coded as x_2 and 'hh1' (Number of members in the household) as x_3.

Variable	Mean	Std. Deviation
x_2 (Age)	50.5008	17.7579%
x_3 (hh1)	2.7194	1.4403%

c) For the variable 'past use' encoded as x_4 which is 1 if 'Yes' or else '0' if 'No'

Variable	Frequency Count	Percentages
$x_4 = 1$ (Yes)	587	47.3387%
$x_4 = 0$ (No)	653	52.6613%

d) For the variable 'sex' as $x_5 = 1$ if Male and $x_5 = 0$ if female. Also for the variable 'Parent' as $x_6 = 1$ if 'Yes' or $x_6 = 0$ if 'No'

Variable	Count	Percentage
$x_5 = 1$ (Male)	603	48.6290%
$x_5 = 0$ (Female)	637	51.3710%
$x_6 = 1$ (Yes for parents)	358	28.8710%
$x_6 = 0$ (No for parents)	882	71.1290 %

e) For the variable 'Marital Status'
Category 1('Single') = '$x_7 = 1$ if $x \in$'Never Been Married'
Category 2('Post Married') = '$x_8 = 1$ if $x \in$ 'Divorced', 'Separated', 'Widowed'
Category 3('Couple') = '*Base Category* = if $\in$'Married', 'Living with a partner'

Variable	Count	Percentage
$x_7 = 1$ (Single)	235	18.9516%
$x_7 = 0$ (Not Single)	1005	81.0484%
$x_8 = 1$ (Post Married)	285	22.9839%
$x_8 = 0$ (Not Post Married)	955	77.0161 %
Couple Category	720	58.0645

f) For the variable 'Income'. The categories are as follows

1. 'Poor' = People $\in$ Less than 10000, 10 to under 20000, 20 to under 30000, 30 to under 40000, 40 to under 50000. Encoded as $x_9 = 1$ if individual $\in$ 'Poor' and $x_9 = 0$ otherwise.

2. 'Middle' = People $\in$ 50 to under 75000, 75 to under 100000. Encoded as $x_{10} = 1$ if individual $\in$ 'Poor' and $x_{10} = 0$ otherwise.
3. 'Rich' = People $\in$ 100 to under 150000, 150000 or more.This is considered as the *Base Category* and not be included in the regression

Variable	Count	Percentage
$x_9 = 1$ (Poor)	519	41.8548%
$x_9 = 0$ (Not Poor)	721	58.1452%
$x_{10} = 1$ (Middle)	366	29.5161%
$x_{10} = 0$ (Not Middle)	874	70.4839 %
Rich Category	232	18.7097%

g) For the variable 'educ'. The categories are as follows:

1. 'HSandBelow' = People $\in$ Less than HS, HS Incomplete, HS. Encoded as $x_{11} = 1$ if individual $\in$ 'HSandBelow' and $x_{11} = 0$ otherwise.
2. 'lessThanBachelors' = People $\in$ Some College, Associate Degree. Encoded as $x_{12} = 1$ if individual $\in$ 'lessThenaBachelors' and $x_{12} = 0$ otherwise.
3. 'BachelorsandAbove' = People $\in$ Bachelors, Postgraduate Degree, Some Postgraduate. This is considered as the *Base Category* and not included in the regression.

Variable	Count	Percentage
$x_{11} = 1$ (HSandBelow)	414	33.3871%
$x_{11} = 0$ (Not HSandBelow)	826	66.6129%
$x_{12} = 1$ (lessThanBachelors)	381	30.7258%
$x_{12} = 0$ (not lessThanBachelors)	859	69.2742 %
BachelorsandAbove Category	445	35.8871%

h) For the variable 'race'. The categories are as follows:

1. 'white' = Encoded as $x_{13} = 1$ if individual $\in$ 'white' and $x_{13} = 0$ otherwise.
2. 'black' = Encoded as $x_{14} = 1$ if individual $\in$ 'black' and $x_{14} = 0$ otherwise.
3. 'allOther' = Considered as *Base Category* which will not be included in the regression.

Variable	Count	Percentage
$x_{13} = 1$ (White)	954	76.9355%
$x_{13} = 0$ (Not White)	286	23.0645%
$x_{14} = 1$ (Black)	146	11.7742%
$x_{14} = 0$ (not Black)	1094	88.2258%
allOthers Category	140	11.2903%

i) For the variable 'party'. The categories are as follows:

1. 'democrat' = Encoded as $x_{15} = 1$ if individual $\in$ 'Democrat' and $x_{15} = 0$ otherwise.
2. 'republican' = Encoded as $x_{16} = 1$ if individual $\in$ 'Republican' and $x_{16} = 0$ otherwise.
3. 'independentOthers' = *Base Category* which consists of individuals who affiliate with all other parties excluding Democratic and Republican. This will not be included in the regression.

Variable	Count	Percentage
$x_{15} = 1$ (Democrat)	422	34.0323%
$x_{15} = 0$ (Not Democrat)	818	65.9677%
$x_{16} = 1$ (Republican)	357	28.7903%
$x_{16} = 0$ (not Republican)	883	71.2097%
independentOthers Category	461	37.1774%

Question 3

a) To derive the probability of $y_i = 1$ and $y_i = 0$ using a binary probit model we can use an underlying 'Latent Variable' z_i such that:

$$f(x) = \begin{cases} 0 & \text{if} \quad z_i \leq 0 \\ 1 & \text{if} \quad z_i > 0 \end{cases} \tag{8}$$

Where z_i is defined as:

$$z_i \ = \ x_i'\beta \ + \ \varepsilon_i \tag{9}$$

β and x_i are vectors kx1 and by the definition of the binary probit model:

$$\varepsilon_i \sim N(0,1) \tag{10}$$

We can derive it as follows:

$$\Pr(y = 1) \ = \ \Pr(z_i > 0) \tag{11}$$

$$\Pr(z_i > 0) = \Pr(x_i'\beta \ + \ \varepsilon_i \ \ > 0) \tag{12}$$

$$\Pr(\varepsilon \ > \ -x_i'\beta) \ \ = \ \ \Pr(\varepsilon \ < \ x_i'\beta) \tag{13}$$

$$\implies Pr(\varepsilon \ < \ x_i'\beta) \ = \ \Phi(x_i'\beta) \tag{14}$$

$$\implies \Pr(y_i = 1) = \Phi(x_i'\beta) \tag{15}$$

We can do this because of Equation (10). ($\Phi(x)$ is the CDF of the Standard Normal Distribution).

Similarily, we can derive the same for $\Pr(y = 0)$

$$\Pr(y = 0) \ = \ \Pr(z_i < 0) \tag{16}$$

$$\Pr(z_i < 0) = \Pr(x_i'\beta \ + \ \varepsilon_i \ \ < 0) \tag{17}$$

$$\implies Pr(\varepsilon \ < \ -x_i'\beta) \ = \ \Phi(-x_i'\beta) \tag{18}$$

$$\implies Pr(\varepsilon \ < \ -x_i'\beta) \ = \ 1 - \ \Phi(x_i'\beta) \tag{19}$$

$$\implies \Pr(y_i = 0) = 1 - \Phi(x_i'\beta) \tag{20}$$

Due to the symmetry of the Normal Distribution

The likelihood function can be constructed as:

$$L(\beta) \ = \prod_{i=1}^{n} \ f(y_i|x_i, \beta) \tag{21}$$

$$L(\beta) \ = \prod_{i=1}^{n} \ [f(y_i = 1|x_i, \beta)]^{y_i} [f(y_i = 0|x_i, \beta)]^{1-y_i} \tag{22}$$

$$L(\beta) \ = \prod_{i=1}^{n} \ [\ \Phi(x_i'\beta) \ ]^{y_i} \ [\ 1 - \ \Phi(x_i\beta)]^{1-y_i} \tag{23}$$

b) The following table is of the coefficient estimates:

Binary Probit Model				
Variable	Coefficients	Estimates	Standard Errors	t-statistic
Intercept	β_1	0.4496	0.2598	1.7302
Age	β_2	-0.0084	0.0029	-2.8755
hh1	β_3	-0.0552	0.0332	-1.6625
Past Use	β_4	0.8035	0.07837	10.2531
Sex	β_5	0.1771	0.0786	2.2527
Parents	β_6	0.0526	0.1051	0.5007
Single	β_7	0.1242	0.1199	1.0363
Post Married	β_8	0.05555	0.1049	0.5294
Poor	β_9	-0.2148	0.0961	-2.2341
Middle	β_{10}	-0.1711	0.1008	-1.6983
HSandBelow	β_{11}	-0.1915	0.0949	-2.0179
lessThanBachelors	β_{12}	0.0798	0.0973	0.8205
White	β_{13}	-0.0346	0.1252	-0.2761
Black	β_{14}	-0.3382	0.1598	-2.1159
Democrat	β_{15}	0.2324	0.0939	2.4738
Republican	β_{16}	-0.4638	0.0950	-4.8786

c) The covariate effect on $\Pr(y_i = 1|x)$ if age is increased by 5 years is calculated as:

$$\textit{Covariate Effect}_{Age} = \Pr(y_i = 1|x_{i,-2},\ x_{i,2} = x_{i,2} + 5, \beta) - \Pr(y_i = 1|x_i, \beta) \tag{24}$$

All values of x_i are kept constant apart from x_2 which is the age variable. For each individual, the value of age is increased by 5, while everything else is kept constant. The reported value of the covariate effect of increasing age by 5 years is.

$$\textit{Covariate Effect}_{Age} = -0.0144 \tag{25}$$

This is the mean of the covariate effect over all of datapoints of x.

d) The covariate effect of parents on $\Pr(y_i = 1|\beta, x_i)$

$$\textit{Covariate Effect}_{Parents} = \Pr(y_i = 1|\beta,\ x_{i,-6},\ x_{i,6} = 1) - \Pr(y_i = 1|\beta,\ x_{i,-6},\ x_{i,6} = 0) \tag{26}$$

All of x_i is kept constant apart from $x_{i,6}$ which is fixed as $x_{i,6} = 1$ in the first term and $x_{i,6} = 0$ for the second term. The reported value of the covariate effect of being a Parent is:

$$\textit{Covariate Effect}_{Parents} = 0.0179 \tag{27}$$

This is the mean of all the covariate effects of the individual being a parent or not in the dataset.

Question 4

a) The probability $\Pr(y_i = 1)$ and $\Pr(y_i = 0)$ can be derived in a similar manner (i.e using a latent variable z_i)

$$f(x) = \begin{cases} 0 & \text{if} \quad z_i \leq 0 \\ 1 & \text{if} \quad z_i > 0 \end{cases} \tag{28}$$

Where z_i is defined as:

$$z_i \;=\; x_i'\beta \;+\; \varepsilon_i \tag{29}$$

β and x_i are vectors of dimension kx1 and by the definition of the binary logit model:

$$\varepsilon_i \sim \text{Logistic}(0, 1) \tag{30}$$

This is where the logit model differs from the probit model, as we can see in Equations (28) and (10). The errors in the logit model follow a standard normal distribution.

For a random variable X, if $X \sim$ Logisitic(0,1) then:

$$E(X) = 0 \tag{31}$$

$$Var(X) = \frac{\pi^2}{3} \tag{32}$$

The probability density function for the standard logistic distribution is:

$$f(x) = \frac{e^{-x}}{[1 + e^{-x}]^2} \tag{33}$$

The cummulative density function for the standard logisitc distribution is donated as:

$$\Pr(X < \alpha) = \Lambda(\alpha) \tag{34}$$

It is evident that due to the symmetricity of the logistic distribution, the following will hold true:

$$\Lambda(\alpha) = 1 - \Lambda(-\alpha) \tag{35}$$

As we can observe, since the variance of the Standard Logistic Distribution is greater than that of the Standard Normal Distribution we can infer that the logistic distribution has comparatively fatter tails.

Using these properties of ε and Logistic distribution, we can derive the probability $\Pr(y_i = 1)$ as follows:

$$\Pr(y = 1) \;=\; \Pr(z_i > 0) \tag{36}$$

$$\Pr(z_i > 0) = \Pr(x_i'\beta \;+\; \varepsilon_i \quad > 0) \tag{37}$$

$$\Pr(\varepsilon \;>\; -x_i'\beta) \quad = \quad \Pr(\varepsilon \;<\; x_i'\beta) \tag{38}$$

$$\implies \Pr(\varepsilon \;<\; x_i'\beta) \;=\; \Lambda(x_i'\beta) \tag{39}$$

$$\Pr(y_i = 1) = \Lambda(x_i'\beta) \tag{40}$$

Similarily, we can also derive the probability for $\Pr(y_i = 0)$ as:

$$\Pr(y = 0) \;=\; \Pr(z_i < 0) \tag{41}$$

$$\Pr(z_i < 0) = \Pr(x_i'\beta \;+\; \varepsilon_i \;\; < 0) \tag{42}$$

$$\implies \Pr(\varepsilon \;<\; -x_i'\beta) \;=\; \Lambda(-x_i'\beta) \tag{43}$$

$$\implies \Pr(\varepsilon \;<\; -x_i'\beta) \;= 1 -\; \Lambda(x_i'\beta) \tag{44}$$

$$\Pr(y_i = 0) = 1 - \Lambda(x_i'\beta) \tag{45}$$

Using these values of $\Pr(y_i = 1)$ and $\Pr(y_i = 0)$ we can construct the likelihood function:

$$L(\beta) \;= \prod_{i=1}^{n} \; f(y_i|x_i, \beta) \tag{46}$$

$$L(\beta) \;= \prod_{i=1}^{n} \; [f(y_i = 1|x_i, \beta)]^{y_i} [f(y_i = 0|x_i, \beta)]^{1-y_i} \tag{47}$$

$$L(\beta) \;= \prod_{i=1}^{n} \; [\Lambda(x_i'\beta) \;]^{y_i} \; [1 -\; \Lambda(x_i\beta)]^{1-y_i} \tag{48}$$

b) The following are the estimates, standard error and t-values of the coefficients using a binary logit model.

Binary Logit Model				
Variable	Coefficients	Estimates	Standard Errors	t-statistic
Intercept	β_1	0.838	0.43	1.948
Age	β_2	-0.015	0.005	-2.999
hh1	β_3	-0.096	0.055	-1.737
Past Use	β_4	1.308	0.13	10.052
Sex	β_5	0.267	0.13	2.058
Parents	β_6	0.074	0.173	0.426
Single	β_7	0.183	0.198	0.921
Post Married	β_8	0.08	0.174	0.46
Poor	β_9	-0.364	0.16	-2.28
Middle	β_{10}	-0.291	0.167	-1.743
HSandBelow	β_{11}	-0.327	0.158	-2.074
lessThanBachelors	β_{12}	0.119	0.161	0.738
White	β_{13}	-0.052	0.206	-0.25
Black	β_{14}	-0.523	0.264	-1.978
Democrat	β_{15}	0.36	0.155	2.316
Republican	β_{16}	-0.776	0.158	-4.905

c) To calculate the covariate effect of increasing age by 5 years on $\Pr(y_i = 1|\beta)$

$$\textit{Covariate Effect}_{Age} = \Pr(y_i = 1|\beta, x_{i,-2},\; x_{i,2} = x_{i,2} + 5) - \Pr(y_i = 1|\beta, x_i) \tag{49}$$

All values of x_i are kept constant apart from x_2 which is the age variable. For each individual, the value of age is increased by 5, while everything else is kept constant. The reported value of the covariate effect of increasing age by 5 years is.

$$\textit{Covariate Effect}_{Age} = -0.0151 \tag{50}$$

This is the mean of the covariate effect over all of datapoints of x.

d) To calculate the covariate effect of being a parent

$$\textit{Covariate Effect}_{Parent} = \Pr(y_i = 1|\beta, x_{i,-6},\ x_{i,6} = 1) - \Pr(y_i = 1|\beta, x_{i,-6}, x_{i,6} = 0) \tag{51}$$

While keeping all other values of x_i the same, and fixing the "Parents" variable. The value of the covariate effect is:

$$\textit{Covariate Effect}_{parents} = 0.0152 \tag{52}$$

This is the mean of the covariate effect of all the datapoints.

e) Yes there is a difference in the Covariate effect calculated in the probit and logit model. This is due to the difference in their variances.

$$Var_{logit} = \frac{\pi^2}{3} \tag{53}$$

$$Var_{probit} = 1 \tag{54}$$

Due to this difference in variance, the estimates can be compared as:

$$\beta_{logit} \approx 1.81\beta_{probit} \tag{55}$$

So hence, the values of the covariate effects may be slightly different, as the β-estimates for the logit model are greater. Even due to the different values, the meaning of the values can still be interpreted as the same. For example, from both models we can conclude that if age increases, the probability of the individual supporting the legalization of marijuana decreases.

Question 5

a) The ordinary probit model generalizes the binary model as y_i can have more than 2 outcomes which have some ordinal significance, i.e they can be ranked.
For this model, we also use the latent variable z_i.

$$z_i \;=\; x_i'\beta \;+\; \varepsilon_i \tag{56}$$

β and x_i are vectors kx1 and by the definition of the ordinal probit model:

$$\varepsilon_i \sim N(0,1) \tag{57}$$

And we use it in the following manner:

$$\gamma_{j-1} < z_i \leq \gamma j \implies y_i = j \tag{58}$$

Where

$$j = \text{Categories into which we have to classify the datapoint to} \tag{59}$$

$$j = 1, 2 \ldots J \tag{60}$$

$$\gamma_j = \text{Threshold values for category } j \tag{61}$$

Therefore, we can derive $\Pr(y_i = j)$ as follows:

$$\Pr(y_i = j) = \Pr(z_i \leq \gamma_j \;,\; \gamma_{j-1} < z_i) \tag{62}$$

$$\implies \Pr(y_i = j) = \Pr(z_i \leq \gamma_j) - \Pr(z_i \leq \gamma_{j-1}) \tag{63}$$

$$\implies \Pr(y_i = j) = \Pr(x_i'\beta + \varepsilon_i \leq \gamma_j) - \Pr(x_i'\beta + \varepsilon_i \leq \gamma_{j-1}) \tag{64}$$

$$\implies \Pr(y_i = j) = \Pr(\varepsilon_i \leq \gamma_j - x_i'\beta) - \Pr(\varepsilon_i \leq \gamma_{j-1} - x_i'\beta) \tag{65}$$

$$\Pr(y_i = j) = \Phi(\gamma_j - x_i'\beta) - \Phi(\gamma_{j-1} - x_i'\beta) \tag{66}$$

We define an indicator variable $\mathbb{Z}_{ij}$, which equals 1 if y_i = j and 0 otherwise. Then log-likelihood function can be defined as follows :

$$L(\beta) \;= \sum_{i=1}^{N} \sum_{j=0}^{m} Z_{ij}\; ln[\; \Phi(\gamma_j - x_i'\beta) \;-\; \Phi(\gamma_{j-1} - x_i'\beta)] \tag{67}$$

b) A descriptive summary of categorical variables with mean and standard deviation is as follows

Variables	Mean	Standard Deviation
educlevel	5.29	2.77
motherWorked	5.00	2.58
urban	2.28	1.66
south	14.40	6.00
educFather	0.90	0.50
educMother		
famIncome		
female		
black		
age15		
age16		
age17		

c)

d) For a one-unit change in covariate x_{jk} , the marginal effect on category j is given by :

$$\frac{\partial Pr(y_i = j|x_i)}{\partial x_{ik}} = \beta_k[\ \Phi(\gamma_j - x_i'\beta)\ -\ \Phi(\gamma_{j-1} - x_i'\beta)] \tag{68}$$

Now for a finite change of $\Delta\ x_{ik}$ we will modify the equation a little bit and say:

$$\Delta Pr(y_i = j) = (\ \Phi(\gamma_j - x'^T{}_i\beta)\ -\ \Phi(\gamma_{j-1} - x'^T{}_i\beta)) - [\ \Phi(\gamma_j - x_i'\beta)\ -\ \Phi(\gamma_{j-1} - x_i'\beta)] \tag{69}$$

where $x^T{}_i = x_i + \Delta\ x_{ik}e_{ik}$ and we average over i to obtain the ACE
Following are the average changes in Pr(educLevel=j) across the sample, rounded to four decimals, and they sum to approximately zero as expected.

Categories	Reported ACE
Pr(educLevel=1)	-0.0289 (decrease)
Pr(educLevel=2)	+0.0153 (increase)
Pr(educLevel=3)	+0.0031 (increase)
Pr(educLevel=4)	+0.0105 (increase