# STA501: Data-based Decision Making

## Problem Set 5

**Question 1.** For the population of men who grew up with disadvantaged backgrounds, let poverty be a dummy variable equal to one if a man in currently living below the poverty line, and zero otherwise. The variable *age* is an individual's age and *educ* is total years of schooling. Let *vocat* be an indicator equal to unity if a man's high school offered vocational training. Using a random sample of 850 men, you obtain

$$Pr(poverty = 1|educ, age, vocat) = \Lambda(.453 - .016age - .087educ - .049vocat)$$

where $\Lambda(z) = exp(z)/(1 + exp(z))$ is the logistic function.

(a) For a 40-year old man with 12 years of education, what is the estimated effect of having vocational training available in high school on the probability of currently living in poverty?

(b) A new welfare program for "out-of-poverty by education" is designed to foster longer support in education from elementary school (K-6) to high school (K-12). What will be the partial effect of the average individual affected by the welfare package for education?

(c) What about average partial effect of all individuals?

(d) Given (c), what's the form of average treatment effect?

(e) Are the effect constant over varying age, educ, vocat? What will be the sign of coefficient, if you add $age^2$ in the logistic regression? How can you justify the sign? What about $educ^2$?

**Question 2.** The following estimates were calculated using a sample of 7,634 women respondents from the General Household Survey 1995. The dependent variable takes the value 1 if the woman was in paid employment, and 0 otherwise. where high is 1 if the respondent has a higher educational qualification, 0 otherwise; noqual

| | OLS | Logit | Probit |
|---|---|---|---|
| *high* | 0.093 | 0.423 | 0.259 |
| | (0.015) | (0.071) | (0.043) |
| *noqual* | −0.210 | −0.898 | −0.554 |
| | (0.013) | (0.056) | (0.035) |
| *age* | 0.038 | 0.173 | 0.108 |
| | (0.003) | (0.124) | (0.008) |
| *age2* | −0.051 | −0.230 | −0.142 |
| | (0.003) | (0.069) | (0.009) |
| *mar* | 0.024 | 0.103 | 0.063 |
| | (0.009) | (0.057) | (0.035) |
| *Constant* | −0.068 | −2.587 | −1.593 |
| | (0.049) | (0.225) | (0.137) |

is 1 if the respondent has no qualification, 0 otherwise; age is the age in years; age2 is (age×age)/100; mar is 1 is married, 0 otherwise. Conventially calculated standard errors are in brackets for the ordinary least squares (OLS) results and asymptotic standard errors are in brackets elsewhere.

(a) Explain how the Probit estimates are obtained when the model has no intercept and there is only one explanatory variable.

(b) Using all three sets of estimates, test the null hypothesis that the coefficient of *mar* is zero. Which test statistic(s) would you consider more reliable? Explain

(c) Using the OLS and Probit estimates calculate the estimated probability of being in paid employment for a married woman, aged 40 with a higher education qualification. Comment on your results.

(d) Why is OLS (or Linear Probability Model) an inferior option to the other two models? Statisticians often claim that Logit and Probit are practically equivalent in real world applications. Given above, can you support the claim?