André Filipe Silva – 26005
João Seixo - 40510
Márcia Serra - 41221

**EXERCISE 1**

**1a)**

LPM models, even when estimated with robust standard-errors, suffer from some unavoidable problems. Mainly, unit increases in $X_i$ always change the estimated $P[y_i = 1|x_i]$ in the same amount, regardless of the initial level. This is a particularly big deal for modelling the probability of labor force participation, as it is probably safe to think that our independent variable does not evolve in a straightly linear fashion – for instance, 1000$ dollars change in the husband's income probably has more impact on women choosing to work or not to work if the husband's income was small in level prior to the change.

With that said, our variable is binomial and the probability of $y_i = 1$ should never be bigger than 1 or smaller than 0. But because of the mentioned issues with the coefficients, we may get probabilities outside the (0,1) interval for the dependent variable. That is why LPM is rarely a suitable way to model probabilistic dependent variables.

**1b)**

First, it should be said that the only thing we can interpret directly from the probit and logit estimated coefficients are its signs: they give us the positive or negative influence of the independent variables on the probability of a woman to be part of the labor force outside the home (i.e. working for a wage outside the home at some point during the year).

Looking only at the signs of the coefficients, we get exactly the same results on Probit and Logit – which shouldn't be surprising since Probit and Logit are two very similar models.

The husband's income (*nwifeinc*), the age of the woman (*age*), and number of children aged less than six years old (*kidslt6*) affect negatively the probability of a married woman working for a wage outside the home. The other variables, *educ*, *exper* and *kidsge6* have a positive impact on the probability of a married woman to work for a wage outside the home. However, both models' p-value for variable *kidsge6* seem to indicate that this variable is not statistically significant.

André Filipe Silva – 26005
João Seixo - 40510
Márcia Serra - 41221

PROBIT:

Interpreting the coefficients with regard to the average marginal effects:

*nwifeinc*: An increase of 1 thousand dollars in the husband's income leads to decrease of the probability of a married woman reporting working for a wage outside the home at some point during the year by 0.3532 p.p.

*educ*: An increase of 1 year of education increases the probability of a married woman reporting working for a wage outside the home at some point during the year by 4.08301 p.p.

*exper*: An increase of 1 year of labor market experience increases the probability of a married woman reporting working for a wage outside the home at some point during the year by 2.14447 p.p.

*age*: An increase of 1 year in the age of a married woman decreases the probability of a married woman reporting working for a wage outside the home at some point during the year by 1.69669 p.p.

*kidslt6*: Having one more kid aged less than six years old decreases the probability of a married woman reporting working for a wage outside the home at some point during the year by 26.70162 p.p.

*kidsge6*: Having one more kid aged between six and 18 years old increases the probability of a married woman reporting working for a wage outside the home at some point during the year by 1.05506 p.p. However, given the p-value shown for this variable, it is not statistically significant at a 5% significance level.

LOGIT:

Interpreting the coefficients with regard to the average marginal effects:

*nwifeinc*: An increase of 1 thousand dollars in the husband's income leads to decrease of the probability of a married woman reporting working for a wage outside the home at some point during the year by 0.36634 p.p.

*educ*: An increase of 1 year of education increases the probability of a married woman reporting working for a wage outside the home at some point during the year by 4.11306 p.p.

*exper*: An increase of 1 year of labor market experience increases the probability of a married woman reporting working for a wage outside the home at some point during the year by 2.16992 p.p.

*age*: An increase of 1 year in the age of a married woman decreases the probability of a married woman reporting working for a wage outside the home at some point during the year by 1.65062 p.p.

*kidslt6*: Having one more kid aged less than six years old decreases the probability of a married woman reporting working for a wage outside the home at some point during the year by 26.08333 p.p.

*kidsge6*: Having one more kid aged between six and 18 years old increases the probability of a married woman reporting working for a wage outside the home at some point during the year by 1.05417 p.p. However, given the p-value shown for this variable, it is not statistically significant at a 5% significance level.

As we can see, the results for both models are very similar. This should not come as a surprise, as the models follow similar cumulative distribution functions (CDF): Standard Normal for the Probit model, and Logistic Distribution for the Logit model.

**1c)**

First, it should be said that LPM and Probit/Logit coefficients are not directly comparable. That is why we need to compare LPM coefficients with Probit and Logit average marginal effect coefficients. This allows for a direct comparison.

The real problem with the LPM model is that, as mentioned above, there is no accounting for the level of the variables, which can lead to predicted probabilities outside the (0,1) interval and bias.

The sign of the coefficients is the same in the three different methods, which corroborates what we have already said about the LPM coefficients.

In what concerns to the magnitude, the LPM coefficients either underestimate or overestimate. We can't say they overestimate all coefficients or underestimate all

coefficients, as that is not what we observe. The Probit and Logit magnitudes are much closer to each other than they are to LPM.

**1d)**

**PROBIT:**

1. The probability of a women working outside home if she has no kids younger than 6 years old is 61,92%, with reference to the average value of the regressors.
2. The probability of a women working outside home if she has one kid younger than 6 years old is 39,48%, with reference to the average value of the regressors.
3. The probability of a women working outside home if she has two kids younger than 6 years old is 25,20%, with reference to the average value of the regressors.

**LPM:**

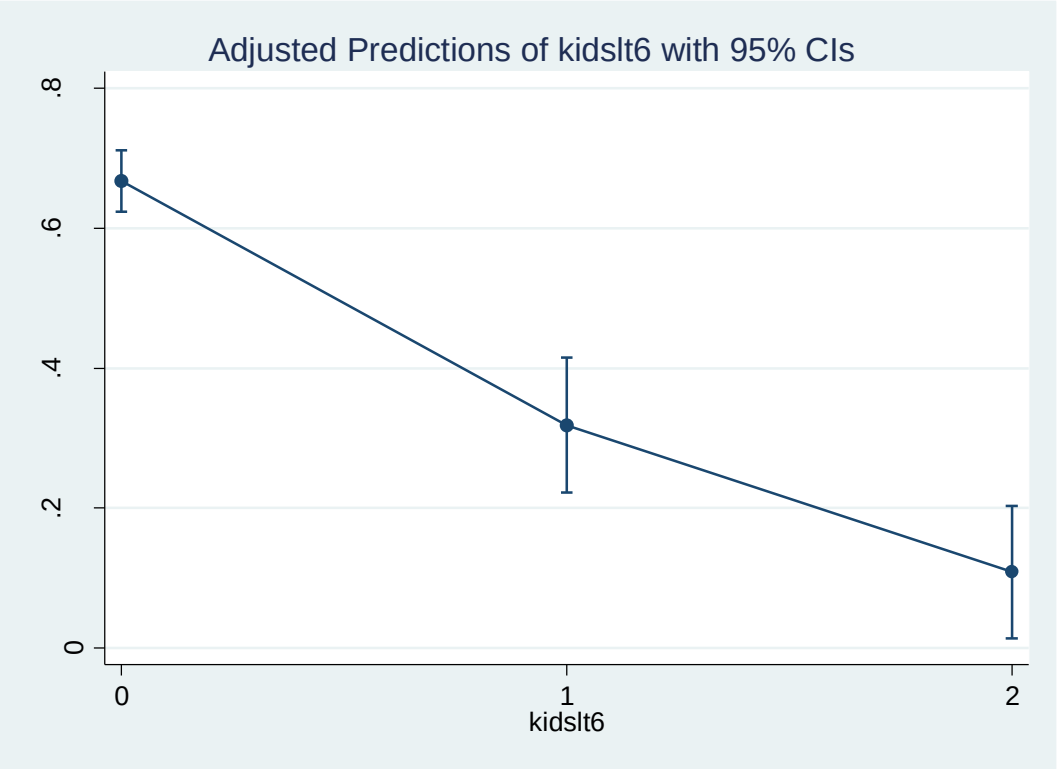
1. On average, the probability of a women working outside home if she has no kids younger than 6 years old is 61,68%, all else constant.
2. On average, the probability of a women working outside home if she has one kid younger than 6 years old is 40,55%, all else constant.
3. On average, the probability of a women working outside home if she has two kids younger than 6 years old is 26,74%, all else constant.

At the average value of the regressors, the probabilities are very close to one another. As the number of children with age lower than 6 increases, the probability of a woman getting into the labor force falls dramatically - but that probability drops faster for the Probit model (as should be expected because it takes into account the level). We can see the drop in probability on both LPM and Probit models. However we can see that for LPM the range (Min, Max) falls outside the (0,1) interval for all estimates – which is why we prefer probit/logit models.

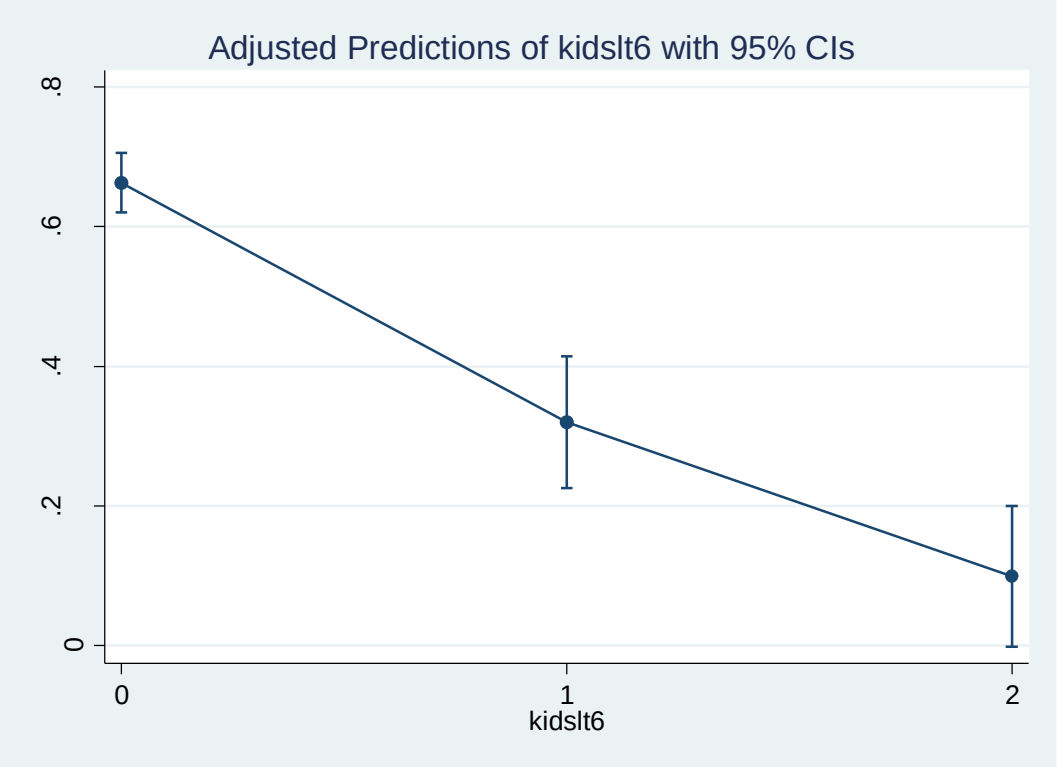
André Filipe Silva – 26005
João Seixo - 40510
Márcia Serra - 41221

**1e)**

PROBIT:




LOGIT:



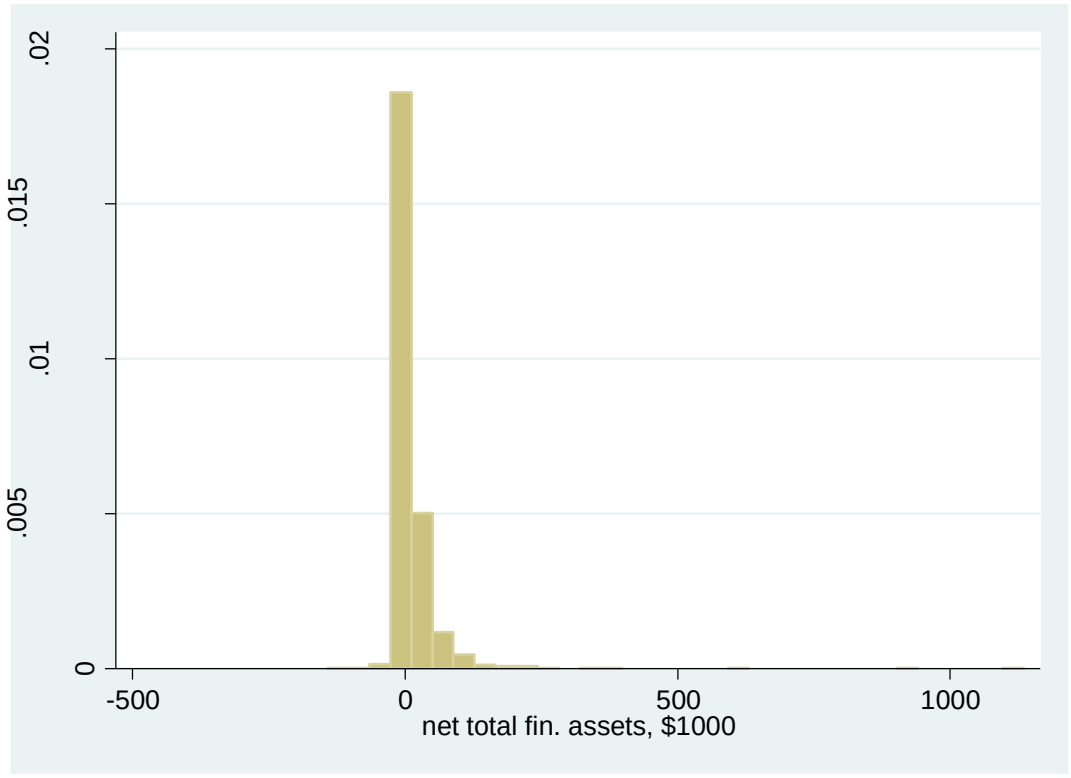
André Filipe Silva – 26005
João Seixo - 40510
Márcia Serra - 41221

**EXERCISE 2**

**2a)**

Focusing on the variable *nettfa*, it is very evident that most individuals do not have considerable financial assets. Even though the mean stands at $13 594,98 , the standard deviation is enormous, showing huge dispersion in the distribution of said financial wealth. Also, if we combine this information with the (min, max) range, the following is clear: most individuals actually have negative financial assets – they are in debt.

This is one example of a variable where the "mean" is a bad way to describe the distribution – because there is huge variation across individuals, and the heteroskedasticity is one of the most important parts – assuming it away would be to not address our dataset in a proper way to get proper conclusions.



André Filipe Silva – 26005
João Seixo - 40510
Márcia Serra - 41221

**2b)**

Given the model:

$$nettfa_i = \beta_0 + \beta_1 inc_i + \beta_2 age_i + \beta_3 agesq_i + \beta_4 e401k_i + \epsilon_i$$

$\beta_1$: An increase of \$1000 in annual income of an individual leads on average to an increase in the net total financial assets of \$782,59, all else constant.

$\beta_2$ and $\beta_3$: *age* and *agesq* must be interpreted jointly. As such, taking the derivative of *nettfa* in order to *age*, we get the following result: A variation of one year in age of the individual leads on average to a variation in the net total financial assets of \$[(-1,568 + 0,0568**age*)*1000], all else constant - depending thus on the age level.

$\beta_4$: An individual that is eligible for a 401(k) pension fund through her employer has on average more 683,66\$ in net total financial assets than an individual with the same characteristics but not eligible for the pension fund.

Age leads to an increase in net total financial assets from age 28 onwards (technically 27,6 – but we rounded it upwards because age is usually represented only as integer). The derivation to conclude this is the following:

$$\frac{\partial netffa}{\partial age} = 0 \Leftrightarrow -1{,}567659 + 2 * 0{,}0283926 * (\text{age}) = 0 \Leftrightarrow$$

$$\Leftrightarrow -1.568 + 0.0568 * (age) = 0 \Leftrightarrow 0.0568 * (age) = 1.568 \Leftrightarrow age = 27.6$$

**2c)**

We test for the presence of heteroskedasticity in the model using White's test.

The test regression:

$$(e_i)^2 = \beta_0 + \beta_1 inc + \beta_2 inc^2 + \beta_3 age + \beta_4 age^2 + \beta_5 age^4 + \beta_6 e401k + \beta_7 inc * age + \beta_8 inc * age^2 + \beta_9 inc * e401k + \beta_{10} age * age^2 + \beta_{11} age * e401k + \beta_{12} age^2 * e401k + \epsilon_i$$

André Filipe Silva – 26005
João Seixo - 40510
Márcia Serra - 41221

The test hypothesis:

$H_0$: $\beta_1 = \beta_2 = \cdots = \beta_{12} = 0$
$H_1$: $\exists \beta_j \neq 0, j = 1,2, \ldots 12$

The test-statistic: $n * R^2$

The decision rule: Reject $H_0$ if: $n * R^2 > \chi^2_{(12)}$ , with 12 being the number of regressors of White's auxiliar regression.

$\chi^2$ observed = 33.34 > $\chi^2_{(12)}$ , 5% significance

We reject the null hypothesis. We conclude for the presence of heteroskedasticity for a 5% significance level, due to the p-value of 0.0009 obtained through STATA.

**2d)**

First, let us state that since the Standard Errors are estimated resorting to Bootstrap calculations, they will most likely be different for each time anyone runs the code. However, the coefficients remain the same across estimations, and as we only need to interpret coefficients, this is not a big problem.

For $\tau = 10$, the 10$^{th}$ percentile, an increase in annual income by \$1000 leads to a decrease in net total financial assets of \$ -17,91.
For $\tau = 25$, the 25$^{th}$ percentile, an increase in annual income by \$1000 leads to an increase in net total financial assets of \$71,29.
For $\tau = 50$, the 50$^{th}$ percentile, an increase in annual income by \$1000 leads to an increase in net total financial assets of \$323,93.
For $\tau = 75$, the 75$^{th}$ percentile, an increase in annual income by \$1000 leads to an increase in net total financial assets of \$797,72.
For $\tau = 90$, the 90$^{th}$ percentile, an increase in annual income by \$1000 leads to an increase in net total financial assets of \$1291,106.

Apart from this, we can say that the coefficient for income increases with each increase of the percentile – the top quantiles increase their net total financial assets more than the bottom quantiles. Meaning that there is a very spread-out unequal distribution of *nettfa* – within-group inequality. There is more net total financial assets increases for the top quantiles that have more income than for the bottom quantiles with lower incomes.

The only comparable to OLS estimate is the one for the $50^{th}$ percentile. For OLS we got an increase of $782,59 on average, and with Quantile Regression we get an increase of $323,93. This gap between estimates shows the upper outliers in the data, and really show the difference between average and median. In OLS the average is way beyond the $50^{th}$ percentile value, because the upper percentile values drag the average up. In QR, we can get a fairer picture of the impact of income on net total financial assets across percentiles. All of this also points towards the heteroskedasticity in the OLS model that we determined previously – when you compare OLS with QR, it makes perfect sense. The variation (or squared S.E.) is higher for higher percentiles.

**2e**)

Endogeneity through omitted variable bias. Income affects both the probability of an individual being eligible for a (401 k) pension and its net total financial assets. As such, this regression is inconsistent, and the coefficients are biased.

André Filipe Silva – 26005
João Seixo - 40510
Márcia Serra - 41221

**EXERCISE 3**

**3a)**

$$\Pr[U_i = j] = \frac{exp(X_{ij} + z_i\gamma_j)}{\sum_{l=1}^{m} exp(X_{il}\beta + z_i\gamma_l)}$$

I = 1,2,3,4 and $\gamma_3 = 0$

1 = AIR

2= BUS

3= CAR

4= TRAIN

**3b)**

**Summary statistics: N mean sd min max by(mode_travel)**
**MODE_TRAVEL**

| | N | mean | sd | min | max |
|---|---|---|---|---|---|
| **AIR** | | | | | |
| gc | 210 | 102.648 | 30.575 | 56 | 197 |
| ttme | 210 | 61.01 | 15.719 | 5 | 99 |
| hinc | 210 | 34.548 | 19.711 | 2 | 72 |
| **BUS** | | | | | |
| gc | 210 | 115.257 | 44.934 | 45 | 222 |
| ttme | 210 | 41.657 | 12.077 | 5 | 60 |
| hinc | 210 | 34.548 | 19.711 | 2 | 72 |
| **CAR** | | | | | |
| gc | 210 | 95.414 | 46.827 | 30 | 238 |
| ttme | 210 | 0 | 0 | 0 | 0 |
| hinc | 210 | 34.548 | 19.711 | 2 | 72 |
| **TRAIN** | | | | | |
| gc | 210 | 130.2 | 58.235 | 42 | 269 |
| ttme | 210 | 35.69 | 12.279 | 1 | 99 |
| hinc | 210 | 34.548 | 19.711 | 2 | 72 |

**3c)**

According to the model, the coefficient associated with $hinc_i$ corresponds to $\gamma_h$. On our STATA output, this corresponds to the coefficient on the 2nd row (AIR), associated with *hinc*.

From the class materials, we can see that only the signs for "alternative-varying" regressors are directly interpretable. *hinc* is an "alternative-invariant regressor" i.e. it only varies by individual and not by alternative.

As such, I would say there is no direct interpretation possible for the coefficient $\gamma_h$.

Marginal effect calculation and interpretation is possible, but as question 3e) asks exactly that, we are assuming that there is no further explanation required here.

**3d)**

We can see that the relative frequencies of the choice of travel and the predicted probabilities are very close to each other, for all modes of travel. This is an indicator that we have estimated a good model.

**3e)**

As the household income variable is associated only with the AIR mode of travel, we believe the interpretation is only due on that coefficient. However, for completion sake, all coefficients will be interpreted.

When choice=AIR, the marginal effect of *hinc* is 0.004085.

This means that an increase in 1 unit ($1000) in income increases by 0.004085 the probability to travel by air rather than by bus, train or car, with regard to the mean values of the independent variables.

When choice=BUS, the marginal effect of *hinc* is -0.000952.

This means that an increase in 1 unit ($1000) in income decreases by -0.000952 the probability to travel by bus rather than by air, train or car, with regard to the mean values of the independent variables.

When choice=CAR, the marginal effect of *hinc* is 0.00686.

This means that an increase in 1 unit ($1000) in income increases by 0.00686 the probability to travel by car rather than by air, train or bus, with regard to the mean values of the independent variables.

When choice=TRAIN, the marginal effect of *hinc* is -0.009993.

This means that an increase in 1 unit ($1000) in income decreases by -0.009993 the probability to travel by train rather than by air, bus or car, with regard to the mean values of the independent variables.

**3f)**

First of all, by looking at the table, we see that only two of the three coefficient estimates of household income are statistically significant at the 5% level, but the results of individual testing may vary with the base outcome/omitted category. As such, we'll perform a Wald test to be more certain that household income has significance.

The result clearly shows us the significance of household income in explaining the travel mean choice.

Keeping in mind that our base category is travel by car, we have negative coefficients for all other means of travel. This lets us interpret that as household income increases, individuals are more likely to travel by car than by bus, train, or air. This, however, is not completely coherent with our previous interpretations on the Conditional Logit model. One might reason that the transformation of the model has made the Multinomial Logit not totally equivalent to the Conditional Logit previously estimated.

Transforming into relative risk-ratio for interpretation is also interesting.

With this transformation, we can see that with a one-unit ($1000) increase in household income, the relative odds of choosing to travel by air, bus, or train rather than car decline – because the coefficients are smaller than one. However, they are only marginally smaller than one, which points that a big household income variation is needed for an actual relevant change in the odds to occur.

**3g)**

It does not seem reasonable to think the IIA assumption holds. If the category AIR is omitted, one would guess that it is not reasonable to assume that individuals will evenly split between CAR, BUS, and TRAIN, as the choice probably changes with income levels

(other examples could be made). But let us do the formal testing through the Hausman-McFaden test.

$H_0$: Coefficients are equal regardless of the presence of other alternatives (IIA holds)
$H_1$: Coefficients are different depending on the presence of different alternatives (IIA does not hold)

The p-values (for a $\chi^2$ distribution) lead us not to reject the null hypothesis, for a 5% significance level.
Based on the Hausman test for the Independence from Irrelevant Alternatives (IIA), this assumption holds. Given the IIA assumption holds, the multinomial logit model is consistent and our coefficients are properly estimated and can be used.

We are a bit surprised to see that the assumption holds, and thus we even looked for a alternative ways to test the assumption (see our STATA commands), to make sure we were not doing anything wrong. In all of our tests, the assumption held.