

ECO351: Homework 3

Question 1. Truncation and Censoring.

- (a) Substituting $\mu_i = \exp(x'_i\beta)$ to parametrize the pdf, we get the PDF conditional on x_i :

$$f(y_i|x_i) = \mu_i^{-1} \exp(-y_i/\mu_i)$$

$$\boxed{f(y_i|x_i) = \exp(-x'_i\beta) \exp(-y_i \exp(-x'_i\beta))} \quad y_i \geq 0$$

Similarly, the CDF can be written by substituting the same value:

$$F(y_i|x_i) = 1 - \exp(-y_i/\mu_i)$$

$$\boxed{F(y_i|x_i) = 1 - \exp(-y_i \exp(-x'_i\beta))}$$

- (b) The probability that a worker is employed for more than 36 months is given by,

$$\begin{aligned} Pr(y_i > 36|x_i) &= 1 - Pr(y_i \leq 36|x_i) \\ &= 1 - F(36|x_i) \\ &= 1 - [1 - \exp(-36 \exp(-x'_i\beta))] \end{aligned}$$

$$\boxed{Pr(y_i > 36|x_i) = \exp(-36 \exp(-x'_i\beta))}$$

- (c) The sample is **truncated** since we do not observe x_i and y_i for $y_i \leq 36$. This implies that we cannot utilise the information for individuals who have been employed for less than or equal to 36 months. It can be represented as:

$$z_i = x'_i\beta + \varepsilon, \varepsilon \sim \text{Exp}(\exp(x'_i\beta))$$

$$y_i = \begin{cases} z_i & \text{if } z_i > 36 \\ \text{unobserved} & \text{otherwise} \end{cases}$$

The **individual likelihood** is:

$$\begin{aligned} l(\beta) &= f(y_i|y_i > 36, x_i) \\ &= \frac{f(y_i|x_i)}{Pr(y_i > 36|x_i)} \\ &= \frac{f(y_i|x_i)}{1 - F(36|x_i)} \\ &= \frac{\exp(-x'_i\beta) \exp(-y_i \exp(-x'_i\beta))}{\exp(-36 \exp(-x'_i\beta))} \end{aligned}$$

$$\boxed{l(\beta) = \exp(-x'_i\beta - (y_i - 36) \exp(-x'_i\beta))}$$

For the **individual log-likelihood function**, we take log on both sides.

$$\boxed{\ln l(\beta) = -x'_i\beta - (y_i - 36) \exp(-x'_i\beta)}$$

The **sample log-likelihood function** can be derived by taking the logarithm of the sample likelihood function. Here, the number of individuals is \tilde{n} . Let the sample likelihood function be $L(\beta)$

$$L(\beta) = \prod_{i=1}^{\tilde{n}} \exp(-x'_i\beta - (y_i - 36)\exp(-x'_i\beta))$$

$$\ln L(\beta) = \ln\left(\prod_{i=1}^{\tilde{n}} \exp(-x'_i\beta - (y_i - 36)\exp(-x'_i\beta))\right)$$

$$\boxed{\ln L(\beta) = \sum_{i=1}^{\tilde{n}} \left[-x'_i\beta - (y_i - 36)\exp(-x'_i\beta) \right]}$$

(d) This model can be called **censored** since we do observe the x_i s for all workers in the company instead of just those who have been employed for more than 36 months.

The **individual contributions** are:

- For an observed (uncensored) worker i with observed $y_i > 36$:

$$l_i(\beta) = f(y_i|x_i)$$

$$= \exp(-x'_i\beta) \exp(-y_i \exp(-x'_i\beta))$$

- For a censored worker i for whom we only know $y_i \leq 36$:

$$l_i(\beta) = \Pr(y_i \leq 36|x_i)$$

$$= F(36|x_i)$$

$$= 1 - \exp(-36 \exp(-x'_i\beta))$$

The **individual log-likelihoods** are:

- For an observed (uncensored) worker i with observed $y_i > 36$:

$$\ln l_i(\beta) = \ln(\exp(-x'_i\beta) \exp(-y_i \exp(-x'_i\beta)))$$

$$= -x'_i\beta - y_i e^{-x'_i\beta}$$

- For a censored worker i for whom we only know $y_i \leq 36$:

$$\ln l_i(\beta) = \ln(1 - \exp(-36 \exp(-x'_i\beta)))$$

The sample likelihood is given as:

$$L(\beta) = \prod_{i:y_i > 36} f(y_i|x_i) \times \prod_{j:y_j \leq 36} F(36|x_j)$$

The **sample log-likelihood** is calculated as below:

$$\ln L(\beta) = \ln\left(\prod_{i:y_i > 36} f(y_i|x_i) \times \prod_{j:y_j \leq 36} F(36|x_j)\right)$$

$$= \sum_{i:y_i > 36} [-x'_i\beta - y_i e^{-x'_i\beta}] + \sum_{j:y_j \leq 36} \ln(1 - \exp(-36 \exp(-x'_j\beta)))$$

Question 2. Tobit Type I Model.

a) **Number and percentage of observations.** The number and percentage of observations in each program category are summarized below:

Program Type	Number of Observations	Percentage (%)
Academic	45	22.50
General	105	52.50
Vocational	50	25.00

TABLE 1. Distribution of program types among 200 observations.

b) **Descriptive Summary.** The descriptive summary for the three variables *read*, *math*, and *apt* is presented below. The table reports the mean, median, standard deviation, minimum, and maximum values for each variable.

Variable	Mean	Median	Std. Dev.	Min	Max
Read	52.23	50	10.25	28	76
Math	52.65	52	9.37	33	75
Apt	640.03	633	99.22	352	800

TABLE 2. Descriptive summary statistics for *read*, *math*, and *apt*.

c) **Histogram of apt.** The histogram shows the distribution of **aptitude scores** across the three program type.

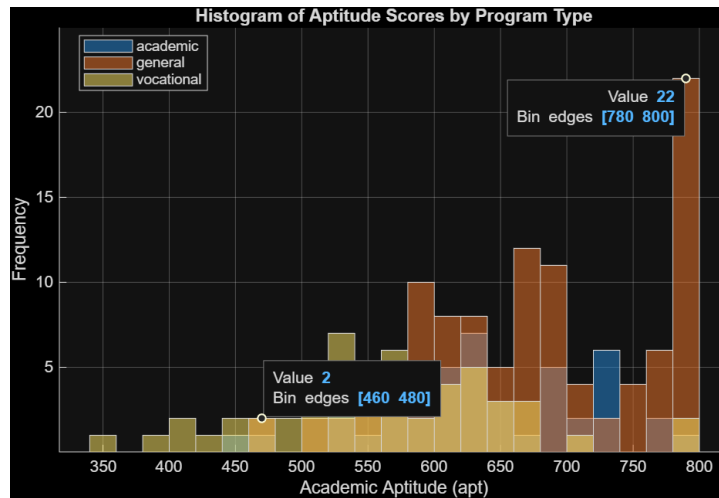


FIGURE 1. Histogram of Academic Aptitude (apt) Scores.

The histogram reveals distinct yet overlapping distributions of **Academic Aptitude (apt)** scores across the three program types: academic, general, and vocational.

- The **academic program** (blue) scores are predominantly concentrated at the **higher end**, with a clear mode between 750 and 800, suggesting students in this program tend to have the highest aptitude.
- The **vocational program** (olive/gold) scores are generally **lower**, with a greater frequency in the middle-to-lower range, spanning approximately 450 to 650. Its distribution appears somewhat left-skewed relative to the other two.

- The **general program** (brown) exhibits the **widest spread**, covering the entire range of scores (~ 400 to 800). It has a high frequency across the distribution but also has the largest absolute frequency in the 750 to 800 bin, indicating a large number of high-aptitude students are in this category.

Overall, there is considerable **overlap** in the middle aptitude range (~ 550 to 700) for all three programs, but the academic and vocational programs show a clear separation in their central tendencies towards the upper and lower halves, respectively.

d) Correlation table and Scatter plots. The **correlation matrix** for the continuous variables *read*, *math*, and *apt* is presented below.

Variable	Read	Math	Apt
Read	1.000	0.662	0.645
Math	0.662	1.000	0.733
Apt	0.645	0.733	1.000

TABLE 3. Correlation matrix for Read, Math, and Aptitude scores.

The correlation coefficients indicate a strong positive relationship among all three variables. The **scatter plots** between the pairs of continuous variables are shown below.

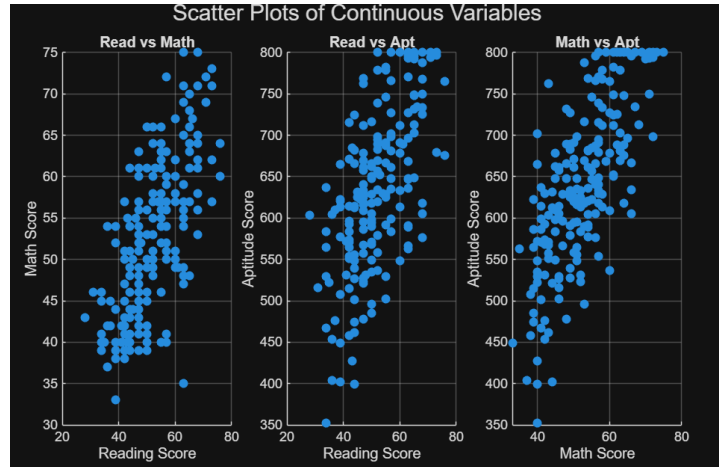


FIGURE 2. Scatter plots.

Observation: All three scatter plots indicate a positive relationship among the variables. Higher reading and math scores are associated with higher aptitude scores. In the "Read vs Apt" plot and the "Math vs Apt" plot, the data points for Aptitude Score (the y-axis) cluster densely at the very top, around 800. This suggests that 800 is the maximum possible score, and for individuals who performed well enough to score above 800, their recorded score is capped at 800. Therefore, **Aptitude score is censored from above at 800.**

e) Tobit Type I Model Estimation. The Tobit Type I model (also known as the Censored Regression Model) is used when the dependent variable is censored at a certain threshold but the underlying latent variable is assumed to follow a normal distribution. In this case, the dependent variable is *aptitude* (*apt*), which represents students' academic aptitude scores. The model assumes that there exists an unobserved latent variable apt_i^* related to the observed variables as:

$$apt_i^* = \beta_0 + \beta_1 read_i + \beta_2 math_i + \beta_3 general_i + \beta_4 vocational_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

and the observed variable follows:

$$apt_i = \begin{cases} apt_i^*, & \text{if } apt_i^* < 800, \\ 800, & \text{if } apt_i^* \geq 800. \end{cases}$$

The model parameters are estimated using Maximum Likelihood Estimation (MLE) via the `fminunc` optimizer in MATLAB. The academic program category serves as the base group.

Variable	Estimate	Std. Error	t-Statistic
Constant	209.57	32.774	6.3943
Read	2.6979	0.6188	4.3600
Math	5.9145	0.7099	8.3314
General	-12.715	12.406	-1.0249
Vocational	-46.144	13.724	-3.3622
log(σ)	4.1847	0.0530	78.901

TABLE 4. Tobit Type I Model Estimates (Right-Censored at 800)

Interpretation:

- The coefficients of both *read* and *math* are **positive and statistically significant** at the 5% level ($|t| > 1.96$), implying that higher reading and mathematics scores are associated with higher latent academic aptitude. Specifically, holding other factors constant, a one-unit increase in the reading score increases the latent aptitude by approximately 2.70 units, while a one-unit increase in the mathematics score increases it by about 5.91 units.
- The dummy variable for the general program is negative but statistically insignificant, indicating that students in the general program do not differ significantly in aptitude from those in the academic program.
- In contrast, the coefficient for the vocational program is negative and statistically significant, suggesting that, on average, students enrolled in the vocational program have lower aptitude scores (approximately 46 units less) than those in the academic program, after controlling for reading and mathematics performance.

Overall, the Tobit Type I results suggest that both reading and math proficiency are strong predictors of students' academic aptitude, while program type plays a significant role in explaining variation across groups.

f) Covariate Effects. The Tobit Type I latent variable specification is

$$apt_i^* = \beta_0 + \beta_1 read_i + \beta_2 math_i + \beta_3 general_i + \beta_4 vocational_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

and the observed outcome is right-censored at $c = 800$:

$$apt_i = \begin{cases} apt_i^*, & \text{if } apt_i^* < 800, \\ 800, & \text{if } apt_i^* \geq 800. \end{cases}$$

Define $z_i = \frac{c - x_i' \beta}{\sigma} = \frac{800 - x_i' \beta}{\sigma}$. The conditional expectation of the *observed* aptitude is

$$E[apt_i | x_i] = \Phi(z_i) x_i' \beta - \sigma \varphi(z_i) + c(1 - \Phi(z_i)),$$

where $\Phi(\cdot)$ and $\varphi(\cdot)$ are the standard normal CDF and PDF. Intuitively, the first two terms give the contribution from uncensored (latent) values and the third term accounts for the mass of observations censored at c .

The marginal effect of a continuous covariate x_{ij} on the expected *observed* aptitude equals

$$\frac{\partial E[apt_i | x_i]}{\partial x_{ij}} = \beta_j \Phi(z_i).$$

Thus the effect on the observed (censored) outcome equals the Tobit coefficient β_j scaled by the probability of being *uncensored* (i.e. $P(\text{apt}_i^* < 800 \mid x_i) = \Phi(z_i)$). The Average Marginal Effect (AME) is

$$AME_j = \frac{1}{N} \sum_{i=1}^N \beta_j \Phi(z_i).$$

For a binary covariate (e.g. the vocational dummy) it is conventional to report the average discrete change:

$$AME_{\text{voc}} = \frac{1}{N} \sum_{i=1}^N \left[E(\text{apt}_i \mid \text{prog} = \text{voc}, x_i) - E(\text{apt}_i \mid \text{prog} = \text{academic}, x_i) \right].$$

Using the estimates from part (e) ($\hat{\beta}_{\text{Read}} = 2.697$, $\hat{\beta}_{\text{Math}} = 5.914$, $\hat{\beta}_{\text{Voc}} = -46.144$, $\hat{\sigma} = \exp(4.205) \approx 66.9$):

- **Reading:** The average marginal effect on the *observed* aptitude for observation i is approximately 2.501 which means the expected change in *apt* for a one unit increase in *read* is 2.501.
- **Mathematics:** Similarly, the average marginal effect due to mathematics on observed aptitude is 5.482.
- **Vocational program (discrete change):** The average effect of switching from academic to vocational on the *observed* expected aptitude for an observation i is -43.838.

(g) **Scatter Plots of Residuals and Fitted Values.**

- The scatter plot shows a distinct cluster of points forming a negatively sloped boundary at the upper fitted values (around 700–800).
- These red points correspond to censored observations, where the latent aptitude would have been higher than 800 but is censored at the limit.
- The uncensored points (blue) appear more symmetrically scattered around zero for lower fitted values, indicating a roughly homoskedastic residual pattern for the uncensored portion.
- The sharp cutoff among red points reflects the censoring mechanism, a classic signature of an upper-censored Tobit model.
- Therefore, the residuals show a clear truncation boundary for censored observations at 800, confirming that censoring is correctly modeled. Uncensored residuals appear randomly distributed, indicating no major specification issues.

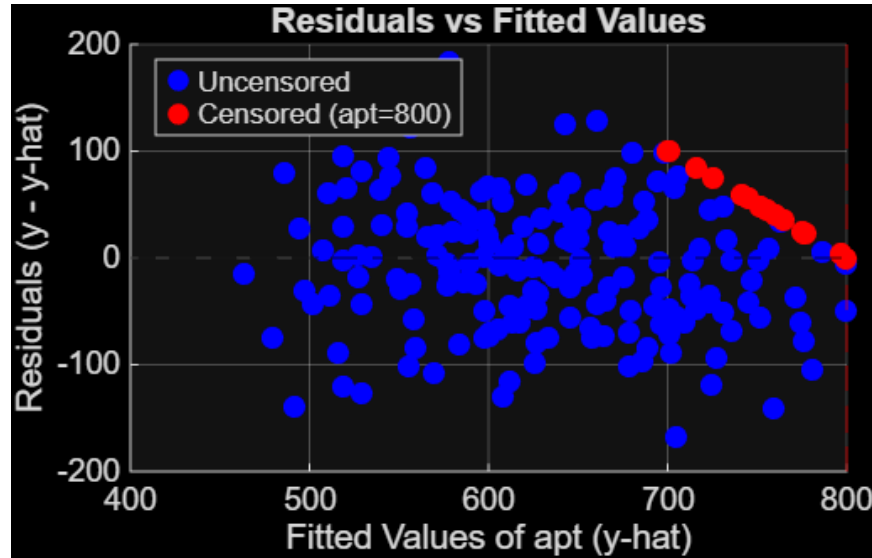


FIGURE 3. Residuals vs. Fitted Values of *apt*

- The observed–fitted plot shows that most points lie close to the 45° line for uncensored observations, suggesting a good overall model fit.
- However, the censored cases (in red) all pile up at the top ($y = 800$), deviating from the 45° line.

- This concentration at the upper limit again reflects right-censoring, as the true latent aptitudes for these individuals exceed the observable maximum.
- Therefore, the fitted values align closely with observed aptitude for uncensored cases, while the flat cluster at the top indicates censoring at 800. This pattern is consistent with a correctly specified Tobit model.

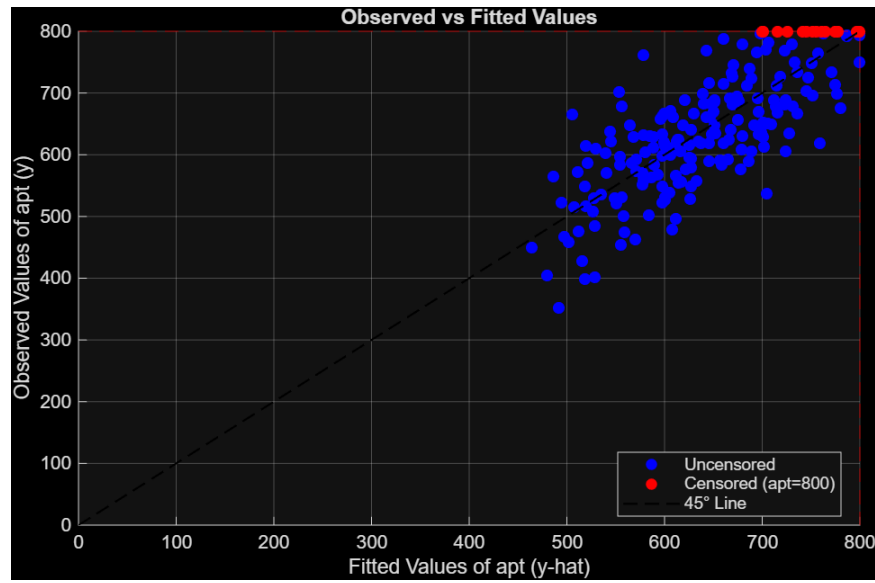


FIGURE 4. Observed *apt* vs. Fitted Values of *apt*

Question 3.

a) OLS. The regression model we are using with the the independent variables **educ**, **exper**, and **hhours** which are her education, her experience, and her husband's hours of work to explain the woman's hours of work. The model is represented as follows:

$$(1) \quad \text{hours} = \beta_1 + \beta_2 \text{educ} + \beta_3 \text{exper} + \beta_4 \text{hhours} + \varepsilon_i$$

The estimates for the above model using OLS are:

Variable	Coefficients	Estimates	Standard Errors	t-statistic
Intercept	β_1	-20.6750	190.7690	-0.108
Educ	β_2	31.1346	12.7894	2.434
Exper	β_3	42.8949	3.6208	11.847
Hhours	β_4	-0.0341	0.0491	-0.694

These OLS estimates are not consistent specifically because we include the data of Women who are not part of the labor force (i.e $lfp = 0$) which introduces a very large number of zeroes. These unusually large number of zero-working hours even when other variables are consistent with that of high working hours(for example education) impacts the coefficient estimate incorrectly. For example: Two individuals with the same education level but with one being part of the labor force and the other not, in this case the coefficient is impacted negatively and hence it will be inconsistent. Therefore, the inflated number of zeroes makes the coefficient estimates inconsistent.

b) OLS with selection. While estimating this model using OLS we only consider the women whose working hours > 0 . The estimates for the above model using OLS are:

Variable	Coefficients	Estimates	Standard Errors	t-statistic
Intercept	β_1	1143.9	245.24	4.6645
Educ	β_2	-21.021	15.779	-1.3322
Exper	β_3	28.883	4.4735	6.4565
Hhours	β_4	0.021734	0.0621033	0.34997

The OLS estimates in this case are also not consistent as there is a sampling selection bias which leads to the error term being correlated with the selection rule of $hours > 0$.

c) Estimating Probit model. The following probit model is used for estimating women's decision to be in the labour force i.e $LFP = 1$:

This is the selection model, the participation latent variable(z_{1i}) equation can be written as:

$$(2) \quad z_{1i} = \gamma_1 + \gamma_2 \text{exper} + \gamma_3 \text{kidsl6} + \gamma_4 \text{kids618} + \gamma_5 \text{MTR} + \gamma_6 \text{largacity} + \varepsilon_{1i}$$

$$(3) \quad LFP_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$(4) \quad \Pr(LFP = 1) = \Pr(z_i > 0) = \Phi(\gamma_1 + \gamma_2 \text{exper} + \gamma_3 \text{kidsl6} + \gamma_4 \text{kids618} + \gamma_5 \text{MTR} + \gamma_6 \text{largacity})$$

The following are the coefficient estimates, standard errors and t-stats for the above model:

Variable	Coefficients	Estimates	Standard Errors	t-statistic
Intercept	γ_1	1.39	0.44472	3.1255
Exper	γ_2	0.064384	0.00707	9.094
kidsl6	γ_3	-0.39382	0.09645	-4.0832
kids618	γ_4	0.14778	0.03977	3.7158
MTR	γ_5	-2.755	0.62034	-4.441
largacity	γ_6	-0.16179	0.10604	-1.5258

As we can see the coefficients of Experience, Kids below 6, Kids between 6 and 18, and MTR have significant t-statistics and hence play an important role in the woman's decision to participate in the labor force. The coefficient of experience is still significantly lower in value than the rest, hence we can also say that the other variable play a more crucial role in explaining the decision.

The variables MTR and kids below 6 have a negative effect on the decision while kids between 6 to 18 has a positive effect on the decision.

d). Using the estimates from the probit model we are to obtain:

$$(5) \quad \tilde{w} = \tilde{\gamma}_1 + \tilde{\gamma}_2 \text{ exper} + \tilde{\gamma}_3 \text{ kidsl6} + \tilde{\gamma}_4 \text{ kids618} + \tilde{\gamma}_5 \text{ MTR} + \tilde{\gamma}_6 \text{ largacity}$$

And using this, we create the inverse Mills ratio

$$(6) \quad \tilde{\lambda} = \frac{\varphi(\tilde{w})}{\Phi(\tilde{w})}$$

$$(7) \quad \text{Sample Mean of } \tilde{\lambda} = 0.716302$$

$$(8) \quad \text{Sample Variance of } \tilde{\lambda} = 0.1173$$

e) Estimating the hours model. The following model is to be estimated

$$(9) \quad \text{hours} = \beta_1 + \beta_2 \text{educ} + \beta_3 \text{exper} + \beta_4 \text{hhours} + \beta_\lambda \tilde{\lambda} + \varepsilon$$

And we use the observations for which hours > 0.

The following are the estimates for the above model to estimate hours:

Variable	Coefficients	Estimates	Standard Errors	t-statistic
Intercept	β_1	2094.6	260.41	8.0436
Educ	β_2	-25.807	11.908	-2.1671
Exper	β_3	5.4648	5.6407	0.96882
hhours	β_4	-0.029992	0.047783	-0.62766
Lambda	β_λ	-797.7	154.18	-5.1737

Including $\tilde{\lambda}$ corrects the bias from estimating hours only for women in the labor force as we did in part (b). Since the coefficient of $\tilde{\lambda}$ is statistically significant, there was a sample selection bias which is being corrected by including $\tilde{\lambda}$. Suppose, if it wasn't significant then the OLS estimation from part(b) would have been unbiased and consistent. That is clearly not the case.

The standard errors are still not correct because the error term is heteroscedastic conditional on selection, hence violating homoscedastic OLS assumptions.

f) Using robust Standard Errors. We estimate the above model using heteroscedasticity robust standard errors to obtain the heteroscedastically consistent estimates and standard errors.

$$\begin{aligned} \hat{\Omega}_{n*n} &= \text{diag}[\hat{\varepsilon}_1, \hat{\varepsilon}_2, \dots, \hat{\varepsilon}_n] \\ \hat{\beta}_{FGLS} &= (X' \hat{\Omega}^{-1} X)^{-1} (X' \hat{\Omega}^{-1} y) \\ \hat{V}(\hat{\beta}) &= (X' X)^{-1} (X' \hat{\Omega} X) (X' X)^{-1} \end{aligned}$$

The above estimates are as follows:

Variable	Coefficients	Estimates	Standard Errors	t-statistic
Constant	β_1	1352.1	349.04	3.8737
Educ	β_2	-23.579	15.067	-1.5649
Exper	β_3	23.515	7.306	3.2186
hhours	β_4	0.023549	0.064558	0.036476
Lambda	β_λ	-177.26	198.81	-0.89162

g) Bootstrap Errors. Following are the estimates with the Bootstrap standard errors:

Variable	Coefficients	Estimates	Bootstrap Standard Errors	t-statistic
Intercept	β_1	2094.6	412.54	5.0774
Educ	β_2	-25.807	16.276	-1.5855
Exper	β_3	5.4648	9.0439	0.60425
hhours	β_4	-0.029992	0.060781	-0.49344
Lambda	β_λ	-797.7	215.99	-3.1656

The bootstrap standard errors are larger than the standard errors in part (e) because the naive OLS Standard Errors contain lambda which is generated and hence violates homoscedasticity. The bootstrap SEs are larger than the estimates in (e) and (f) but they are still closer to the values in (f).

The bootstrap SEs affect the t-statistics, and it reduces the significance of variables with marginal significance in (e). For example, the significance of Education falls below 1.96. The inverse-mills ratio is still significant and hence indicates that sample selection bias is present. The bootstrap SE accounts for the estimation of the inverse mills ratio and the heteroskedasticity in the two step procedure.

h) Heckman's Sample Selection Model. Equation (1) for hours is the outcome equation and equation (4) is used as the selection criteria. We use the full information maximum likelihood to compute the estimates which are as follows:

Variable	Coefficients	Estimates	Standard Errors	t-statistic
Outcome Intercept	β_1	1145.4	286.31	4.0007
Educ	β_2	-5.0625	15.77	-0.32103
Outcome -Exper	β_3	22.051	5.7161	3.8578
hhours	β_4	0.046664	0.062431	0.74745
Selection Constant	γ_1	1.6664	0.44587	3.7374
Selection Exper	γ_2	0.0642	0.0071252	9.0102
Kidsl6	γ_3	-0.42455	0.095034	-4.4674
Kids618	γ_4	0.13877	0.039936	3.4747
MTR	γ_5	-3.1008	0.62052	-4.997
Largecity	γ_6	-0.19052	0.10383	-1.835
Log of s.d	$\ln \sigma$	6.6417	0.046454	142.97
Correlation	$atanh(\rho)$	-0.42026	0.17395	-2.416

The values of σ and ρ are as follows:

$$\hat{\sigma} = 766.40468$$

$$\hat{\rho} = -0.3972$$

The Heckman estimates indicate the same as the estimates in (e) in terms of the direction of the effect of the given variables. Some coefficients like that of experience are larger in value and significance due to the efficiency of the FIML, which maximizes the joint likelihood.

Some of the coefficient estimates in (e) are larger than those in FIML due to the outcome variance changing the coefficient scale.

The bias term which is λ in (e) and ρ in (h) are significant in both and negative as well which confirms the selection bias. Overall the two-step model provides consistent but less efficient estimates and the FIML produces a more reliable specification.

Question 4.

a) Fraction receiving job training : To determine the proportion of individuals who participated in job training, we used the indicator variable **train**, which takes the value 1 if a man received job training between 1975–1977, and 0 otherwise. The variable **train** is binary, so its mean represents the proportion of workers who underwent training:

$$\text{Fraction Trained} = \frac{\sum_{i=1}^N \mathbf{train}_i}{N}$$

From the computation, we obtain:

$$\text{Number trained} = 185, \quad \text{Total sample size} = 445$$

$$\Rightarrow \text{Fraction trained} = \frac{185}{445} \approx 0.416$$

\Rightarrow Approximately 41.6% of the men in the sample received job training during the 1975–1977 period. This fraction will serve as the treatment proportion in subsequent analysis of training effects on 1978 earnings

(b) Difference on receiving job training: The variable **re78** represents real earnings in 1978, measured in thousands of 1982 dollars. Comparing the average earnings of individuals who received training (**train** = 1) with those who did not (**train** = 0). The computed sample means are:

$$\bar{y}_{\text{trained}} = 6.349, \quad \bar{y}_{\text{untrained}} = 4.555, \quad \Rightarrow \Delta = 6.349 - 4.555 = 1.794$$

\Rightarrow On average, men who received job training earned approximately \$1,794 more in 1978 (in 1982 dollars) than those who did not receive training. Considering the relatively low-income nature of the sample, this difference is economically meaningful. It indicates that participation in job training programs was associated with noticeably higher post-training earnings, suggesting a potentially positive treatment effect.

(c) Effect of Job Training in Unemployment : The variable **unem78** is a binary indicator taking the value 1 if an individual was unemployed in 1978, and 0 otherwise. To compute the unemployment fractions, we separate the sample based on the job training indicator **train** and take the mean of **unem78** within each group.

$$\text{Fraction unemployed (trained)} = \frac{\sum_{i:T_i=1} \mathbf{unem78}_i}{N_{T=1}} = 0.243$$

$$\text{Fraction unemployed (untrained)} = \frac{\sum_{i:T_i=0} \mathbf{unem78}_i}{N_{T=0}} = 0.354$$

$$\text{Difference} = 0.243 - 0.354 = -0.111$$

Hence, about 24.3% of trained men were unemployed in 1978, compared to 35.4% among those who did not receive training. The difference of -0.111 indicates that job training is associated with an 11.1 percentage point reduction in unemployment, suggesting a potentially meaningful economic effect.

Appendix

MATLAB Code

```

1 clear;
2 clc;
3
4 %% Question 3 Assignment 3
5
6 data = readtable('mroz.xlsx','Sheet','Data');
7
8 hours = data.hours;
9 educ = data.educ;
10 exper = data.exper;
11 hhours = data.hhours;
12 lfp = data.lfp;
13 kidsl6 = data.kidsl6;
14 kids618 = data.kids618;
15 mtr = data.mtr;
16 largecity = data.largecity;
17
18 n = height(data);
19
20 X = [ones(n,1), educ, exper, hhours];
21 Z = [ones(n,1), exper, kidsl6, kids618, mtr, largecity];
22
23 y = hours;
24 d = lfp;
25
26 kx = size(X,2);
27 kz = size(Z,2);
28
29 % OLS estimation
30 b = (X' * X) \ (X' * y); % OLS formula
31 yhat = X * b; % fitted values
32 res = y - yhat; % residuals
33
34 k = size(X,2);
35 sigma2 = (res' * res) / (n - k);
36 Varb = sigma2 * inv(X' * X);
37 se = sqrt(diag(Varb));
38 tstat = b ./ se;
39 pval = 2 * (1 - tcdf(abs(tstat), n - k));
40
41 fprintf('\n===== Q3(a): OLS Estimation =====\n');
42 varNames = {'Const','Educ','Exper','Hhours'};
43 for j = 1:k
44     fprintf('%-8s: %10.4f    SE=%8.4f    t=%8.3f    p=%8.4f\n', ...
45         varNames{j}, b(j), se(j), tstat(j), pval(j));
46 end
47 fprintf('-----\n');
48

```

```

49
50 %% Part b)
51
52 idx = hours > 0;
53 y_b = hours(idx);
54 X_b = [ones(sum(idx),1), educ(idx), exper(idx), hhours(idx)];
55 varNames = {'const','educ','exper','hhours'};
56
57 [n_b, k] = size(X_b);
58 beta = (X_b' * X_b) \ (X_b' * y_b);
59 u = y_b - X_b*beta;
60
61 sigma2 = (u' * u) / (n_b - k);
62 V = sigma2 * inv(X_b' * X_b);
63 se = sqrt(diag(V));
64 tstat = beta ./ se;
65
66 % --- Display results ---
67 disp('OLS on hours > 0');
68 fprintf('N (hours>0): %d\n\n', n_b);
69
70 T_out = table(varNames', beta, se, tstat, ...
71     'VariableNames', {'Variable','Coefficient','StdError','tStat'});
72 disp(T_out);
73
74 %% Part C)
75
76 X_probit = [ones(size(lfp)), exper, kidsl6, kids618, mtr, largecity];
77 k_probit = size(X_probit, 2);
78 llfun = @(b) - sum(lfp .* log(normcdf(X_probit*b)) + (1 - lfp) .* log(1
    - normcdf(X_probit*b)));
79 opts = optimoptions('fminunc', 'Algorithm','quasi-newton', ...
80     'Display','iter', 'MaxFunctionEvaluations',1e5, 'MaxIterations',1e4
    );
81 b0 = zeros(k_probit,1);
82 [bhat, fval, exitflag, output, grad, hess] = fminunc(llfun, b0, opts);
83
84 varb_probit = inv(hess);
85 se_probit = sqrt(diag(varb_probit));
86
87 tstat_probit = bhat ./ se_probit;
88
89 varNames = {'const','exper','kidl6','kids618','MTR','largecity'};
90 disp(' ');
91 disp('Q3(c) PROBIT MODEL FOR LABOR FORCE PARTICIPATION');
92 T_out = table(varNames', bhat, se_probit, tstat_probit, ...
93     'VariableNames', {'Variable','Coefficient','StdError','tStat'});
94 disp(T_out);
95
96 %% Part D)
97

```

```

98 w_tilde = X_probit * bhat;
99 w_pdf = normpdf(w_tilde);
100 w_cdf = normcdf(w_tilde);
101
102 lambda = w_pdf ./ w_cdf;
103
104 lambda_mean = mean(lambda);
105 lambda_var = var(lambda);
106
107 fprintf('Sample mean of lambda = %.6f\n', lambda_mean);
108 fprintf('Sample var. of lambda = %.6f\n', lambda_var);
109
110 %% Part E)
111
112 X_ss = [ones(sum(idx), 1), educ(idx), exper(idx), hhours(idx), lambda(
    idx)];
113 [n_ss, k_ss] = size(X_ss);
114 y_ss = hours(idx);
115
116 beta_ss = (X_ss' * X_ss) \ (X_ss' * y_ss);
117 u_ss = y_ss - X_ss * beta_ss;
118
119 sigma2_ss = (u' * u) / (n - k);
120 var_ss = sigma2_ss * inv(X_ss' * X_ss);
121 se_ss = sqrt(diag(var_ss));
122 tstat_ss = beta_ss ./ se_ss;
123
124
125 varNames_ss = {'const', 'educ', 'exper', 'hhours', 'lambda'};
126 disp(' ');
127 disp('Q3(e): Heckman Two-Step Regression (Hours on and Covariates)');
128 T_out = table(varNames_ss', beta_ss, se_ss, tstat_ss, ...
129     'VariableNames', {'Variable', 'Coefficient', 'StdError', 'tStat'});
130 disp(T_out);
131
132 %% Part F
133
134 S = X_ss' * (diag(u_ss.^2)) * X_ss;
135 V_rob = inv(X_ss' * X_ss) * S * inv(X_ss' * X_ss);
136
137 beta_FGLS = (X_ss'*(diag(1./(u.^2 + eps)))*X_ss) \ (X_ss'*(diag(1./(u
    .^2 + eps)))*y_ss);
138
139 se_rob = sqrt(diag(V_rob));
140
141 t_fgls = beta_FGLS ./ se_rob;
142 % --- Display results for FGLS ---
143 varNames_fgls = {'const', 'educ', 'exper', 'hhours', 'lambda'};
144 T_out_fgls = table(varNames_fgls', beta_FGLS, se_rob, t_fgls, ...
145     'VariableNames', {'Variable', 'Coefficient', 'StdError', 'tStat'});
146 disp(' ');

```

```

147 disp('Q3(f): FGLS Regression Results');
148 disp(T_out_fgls);
149
150 %% Part g)
151
152 probit_fit = @(y, X) ...
153     fminunc(@(b) -sum( y.*log(max(normcdf(X*b), 1e-12)) + ...
154         (1-y).*log(max(1 - normcdf(X*b), 1e-12)) ), ...
155         zeros(size(X,2),1), ...
156         optimoptions('fminunc','Algorithm','quasi-newton','Display'
157             , 'off', ...
158                 'MaxFunctionEvaluations',1e5,'MaxIterations',1
159                 e4)));
160
161 Xsel = X_probit;
162
163 B = 400;
164 k = numel(beta_ss);
165 boot_betas = NaN(k, B);
166
167 for b = 1:B
168     try
169         ii = randi(n, n, 1);
170
171         lfp_b = lfp(ii);
172         exper_b = exper(ii);
173         kidsl6_b = kidsl6(ii);
174         kids618_b = kids618(ii);
175         mtr_b = mtr(ii);
176         largecity_b = largecity(ii);
177         hours_b = hours(ii);
178         educ_b = educ(ii);
179         hhours_b = hhours(ii);
180
181         %Selection
182         Xsel_b = [ones(n,1), exper_b, kidsl6_b, kids618_b, mtr_b,
183             largecity_b];
184         b_sel_b = probit_fit(lfp_b, Xsel_b);
185         w_b = Xsel_b * b_sel_b;
186
187         lambda_b = normpdf(w_b) ./ max(normcdf(w_b), 1e-12);
188
189         idx_b = hours_b > 0;
190         if (sum(idx_b) <= k)
191             continue;
192         end
193         y_boot = hours_b(idx_b);
194         X_boot = [ones(sum(idx_b), 1), educ_b(idx_b), exper_b(idx_b),
195             hhours_b(idx_b), lambda_b(idx_b)];

```

```

194         beta_boot = (X_boot'*X_boot) \ (X_boot' * y_boot);
195         boot_betas(:, b) = beta_boot;
196     catch
197         continue
198     end
199 end
200
201 se_boot = std(boot_betas, 0, 2);
202
203 t_boot = beta_ss ./ se_boot;
204
205 disp('Q3(g): Heckman 2-step with Bootstrap Standard Errors (B = 400)');
206 fprintf('Valid bootstrap replications: %d (of %d)\n\n', size(boot_betas
    ,2), B);
207
208 T_out = table(varNames_ss', beta_ss, se_boot, t_boot, ...
209     'VariableNames', {'Variable', 'Coef', 'SE_boot', 't_boot'});
210 disp(T_out);
211
212 %% Heckman model
213
214 sel1 = (lfp == 1);
215 sel0 = (lfp == 0);
216
217 beta0 = X(sel1,:) \ hours(sel1);
218 gamma0 = zeros(kz, 1);
219 lnsig0 = log(std(hours(sel1)));
220 arho0 = atanh(0.1);
221 theta0 = [beta0; gamma0; lnsig0; arho0];
222
223 function nll = heckman_nll(theta, X, Z, y, lfp, sel1, sel0)
224     kx = size(X,2);
225     kz = size(Z,2);
226     beta      = theta(1:kx);
227     gamma     = theta(kx+1:kx+kz);
228     ln_sigma  = theta(kx+kz+1);
229     atanh_rho = theta(kx+kz+2);
230
231     sigma = exp(ln_sigma);
232     rho   = tanh(atanh_rho);
233     one_m_r2 = max(1 - rho^2, 1e-12);
234     srt = sqrt(one_m_r2);
235
236     XB  = X*beta;
237     ZG  = Z*gamma;
238
239
240     y1   = y(sel1);
241     XB1  = XB(sel1);
242     ZG1  = ZG(sel1);
243     a1   = (y1 - XB1) / sigma;

```

```

244
245     comp1 = -log(sigma) + log(normpdf(a1)) + log( max( normcdf( (ZG1 +
rho*a1)/srt ), 1e-12) );
246
247     ZG0 = ZG(sel0);
248     comp0 = log( max( normcdf(-ZG0), 1e-12) );
249
250     ll = sum(comp1) + sum(comp0);
251     nll = -ll;
252 end
253
254 obj = @(th) heckman_nll(th, X, Z, hours, lfp, sel1, sel0);
255
256 opts = optimoptions('fminunc', 'Algorithm','quasi-newton', 'Display', '
iter', ...
257     'MaxFunctionEvaluations',2e5, 'MaxIterations', 2e4);
258
259 [theta_hat, fval, exitflag, output, grad, hess] = fminunc(obj, theta0,
opts);
260
261 beta_he = theta_hat(1:kx);
262 gamma_he = theta_hat(kx+1:kx+kz);
263 lnsigma_hat = theta_hat(kx+kz+1);
264 atanhrho_hat = theta_hat(kx+kz+2);
265 sigma_hat = exp(lnsigma_hat);
266 rho_hat = tanh(atanhrho_hat);
267
268 V_he = inv(hess);
269 se_he = sqrt(diag(V_he));
270
271 pnames = [ ...
272     strcat("beta_", ["const","educ","exper","hhours"]), ...
273     strcat("gamma_", ["const","exper","kid16","kids618","mtr","
largecity"]), ...
274     "ln_sigma","atanh_rho" ]';
275
276 est = [beta_he; gamma_he; lnsigma_hat; atanhrho_hat];
277 tstats = est ./ se_he;
278
279 disp('Q3(h): Heckman Selection (FIML) Estimates');
280 T_out = table(pnames, est, se_he, tstats, 'VariableNames', {'Parameter'
, 'Estimate', 'StdError', 'tStat'});
281 disp(T_out);

```