Problem Set 4 Answers

Question 1

The baseline model regresses log wages on education. However, this model would likely suffer from omitted variable bias, so we could include controls for experience, age, education, and ability. Other factors, such as region, might affect labour market outcomes so that they could be relevant controls.

Question 3

a)

Most but not all individuals had a non-negative value for the primary work hours variable, which is consistent with some unemployment or intentional non-responses. The number of negative values is much higher for the secondary work variable, which would be consistent with fewer people working two jobs. The table includes observations with zeros as positive.

name	not_missing	count
hrp1	0	1727
hrp1	1	5344
hrp2	0	5875
hrp2	1	1196

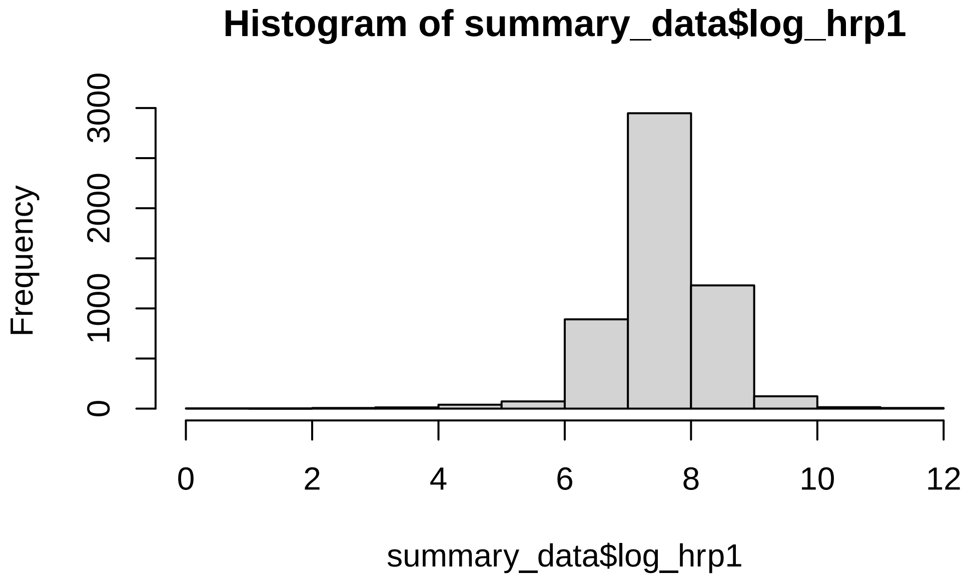
b)

Values are high because they are coded in integers, with the two leftmost digits representing cents—for example, 500 stands for 5 dollars.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0    1225    1910    2633    3000  110000 
```

c)

The distribution has thin tails and is symmetric. It is centred around 8.



Question 4

a)

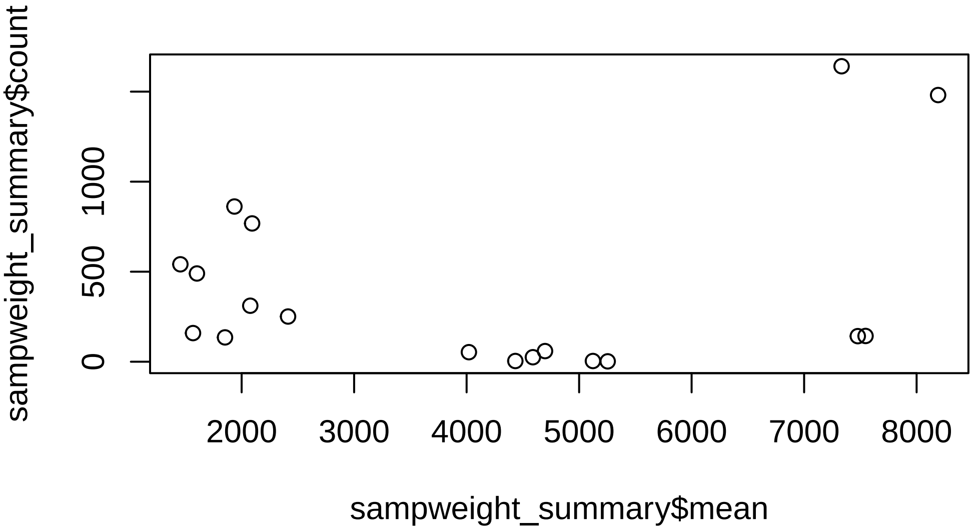
The weight should represent the number of individuals each respondent's answers represent, so the divided weight is how many 100s of individuals each response represents. The mean is the average number of one hundred people represented by an average person in the sample.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  56749  172806  412156  474608  733119 1603933 
```

b)

Since the group count and the mean sample seem unrelated, the survey does not represent the population.

```
[1] 4746.077
```



c)

Since each observation is a person, these are most likely the weights for each person in the sample.

Question 5

a)

For `q3_4`, the mode is to have completed high school, and the second most frequent is to have completed college. Few people drop out of university or high school.

q3_4	n
12	2878
16	974

b)

We dropped six observations for individuals who did not know or refused to answer.

```
[1] 6
```

Question 6

a)

People are between 49 and 59 years old, which could mean higher wages than the average population.

```
[1] 49 59
```

b)

We might control for age squared if we expect non-linear effects of wage, e.g., income decreases when near retirement after peaking.

c) and d)

The new generation is more educated, but a gender gap persists.

female	child	parent	difference
0	13.46236	11.81952	1.642847
1	13.71652	11.61324	2.103287

e)

We observe positive correlations with both mother and father, but the correlation with the father is slightly higher.

```
 mom_schl  pop_schl
0.3971862 0.4027837
```

f)

It might measure general academic skills rather than manual or technical ones.

g)

```
[1] 0.2495402
```

Question 7

a)

The coefficient is around 0.115, which means that each extra year increases mean education by 11.5%. This result is in line with the previous literature.

```
OLS estimation, Dep. Var.: log_hrp1
Observations: 5,339
Weights: corrected_weight
Standard-errors: Heteroskedasticity-robust
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 6.040299   0.080785 74.7697 < 2.2e-16 ***
yschl       0.115597   0.005966 19.3758 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 55.0   Adj. R2: 0.132209
```

b)

The coefficient is now around 10.5%. The additional controls probably removed some omitted variable bias, so it is expected. Coefficients on parental schooling are positive, while age is negative. The latter is likely true because the individuals in the sample are older. The coefficient on “female” is negative, which reflects the gender wage gap.

```
OLS estimation, Dep. Var.: log_hrp1
Observations: 4,487
Weights: corrected_weight
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error   t value  Pr(>|t|)
(Intercept) 16.490681   8.466290   1.94780  0.051501 .
yschl        0.104690   0.006566  15.94491 < 2.2e-16 ***
female      -0.355444   0.026377 -13.47578 < 2.2e-16 ***
age         -0.386452   0.317103  -1.21870  0.223023
age_sqd      0.003595   0.002969   1.21107  0.225934
mom_schl     0.010735   0.007272   1.47618  0.139966
pop_schl     0.010378   0.006314   1.64360  0.100329
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 53.6   Adj. R2: 0.180594
```

c)

The coefficient is even higher, at 8.3%. The coefficient on AFQT is positive and significant, which implies the existence of ability bias.

```
OLS estimation, Dep. Var.: log_hrp1
Observations: 5,339
Weights: corrected_weight
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error    t value  Pr(>|t|)
(Intercept) 11.128332   7.922059   1.404727   0.16016
yschl        0.082851   0.006470  12.804577 < 2.2e-16 ***
female      -0.333204   0.024982 -13.337942 < 2.2e-16 ***
age         -0.180596   0.296741  -0.608598   0.54282
age_sqd      0.001704   0.002779   0.613104   0.53983
afqt         0.005779   0.000479  12.056560 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 52.7   Adj. R2: 0.203444
```

d)

The coefficient is -0.38. Since schooling decisions might also affect later health outcomes, which, according to the variable definitions, affect labour market outcomes, they should be included to avoid omitted variable bias.

```
OLS estimation, Dep. Var.: log_hrp1
Observations: 5,339
Weights: corrected_weight
Standard-errors: Heteroskedasticity-robust
                 Estimate Std. Error    t value   Pr(>|t|)
(Intercept)     11.128854   7.809937   1.424961 1.5423e-01
yschl            0.080582   0.006405  12.581372  < 2.2e-16 ***
female          -0.317770   0.024974 -12.723911  < 2.2e-16 ***
age             -0.180887   0.292569  -0.618269 5.3642e-01
age_sqd          0.001737   0.002740   0.633818 5.2623e-01
afqt             0.005557   0.000472  11.782092  < 2.2e-16 ***
```

```
health_problems -0.383476   0.049531  -7.742200 1.1614e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 52.0   Adj. R2: 0.223952
```

e)

Each dummy is generally not statistically significant individually. Although the model is more flexible as it captures non-linear effects, the sample size might not be sufficiently large to estimate these values with precision.

term	estimate	std.error	statistic	p.value
(Intercept)	12.0868050	7.9069017	1.5286399	0.1264133
factor(yschl)2	0.1784713	0.5257353	0.3394699	0.7342692
factor(yschl)3	-1.3051897	0.6608667	-1.9749667	0.0483236
factor(yschl)4	-0.3198519	0.5266652	-0.6073154	0.5436676
factor(yschl)5	0.0786784	0.5702259	0.1379776	0.8902633
factor(yschl)6	-0.2420055	0.5313765	-0.4554312	0.6488176
factor(yschl)7	-2.1166419	1.3002962	-1.6278152	0.1036234
factor(yschl)8	-0.2481249	0.5348926	-0.4638780	0.6427542
factor(yschl)9	-0.5299321	0.5379222	-0.9851463	0.3245971
factor(yschl)10	-0.3291790	0.5272064	-0.6243835	0.5324026
factor(yschl)11	-0.3634864	0.5270186	-0.6897032	0.4904110
factor(yschl)12	-0.1251123	0.5237227	-0.2388903	0.8111999
factor(yschl)13	-0.0450625	0.5251420	-0.0858101	0.9316206
factor(yschl)14	0.0379512	0.5240443	0.0724198	0.9422706
factor(yschl)15	0.0574172	0.5286396	0.1086130	0.9135135
factor(yschl)16	0.2666939	0.5242965	0.5086699	0.6110047
factor(yschl)17	0.2301463	0.5283377	0.4356044	0.6631415
factor(yschl)18	0.3446913	0.5280915	0.6527113	0.5139706
factor(yschl)19	0.5306184	0.5339073	0.9938400	0.3203460
factor(yschl)20	0.5175680	0.5284925	0.9793289	0.3274621
female	-0.3345135	0.0250018	-13.3795657	0.0000000
age	-0.1738777	0.2953176	-0.5887820	0.5560325
age_sqd	0.0016433	0.0027656	0.5941959	0.5524064
afqt	0.0053086	0.0004772	11.1253335	0.0000000

The dummies increase yearly and are statistically significant after the 16th year. This result indicates that among the most educated group, the payoff is concentrated around the ones with the highest levels of education.

term	estimate	std.error	statistic	p.value
(Intercept)	12.1948651	7.8687153	1.5497911	0.1212517
factor(yschl)9	-0.2815914	0.1627736	-1.7299570	0.0836965
factor(yschl)10	-0.0811535	0.1232518	-0.6584366	0.5102863
factor(yschl)11	-0.1153987	0.1218363	-0.9471619	0.3435997
factor(yschl)12	0.1229819	0.1084765	1.1337192	0.2569639
factor(yschl)13	0.2028486	0.1173696	1.7282897	0.0839948
factor(yschl)14	0.2859370	0.1111068	2.5735321	0.0100937
factor(yschl)15	0.3053804	0.1315991	2.3205364	0.0203498
factor(yschl)16	0.5143708	0.1160962	4.4305581	0.0000096
factor(yschl)17	0.4779100	0.1324110	3.6092934	0.0003099
factor(yschl)18	0.5924731	0.1299773	4.5582818	0.0000053
factor(yschl)19	0.7780982	0.1535785	5.0664524	0.0000004
factor(yschl)20	0.7652448	0.1334248	5.7354018	0.0000000
female	-0.3349365	0.0250006	-13.3971360	0.0000000
age	-0.1867885	0.2953196	-0.6324963	0.5270900
age_sqd	0.0017599	0.0027657	0.6363409	0.5245819
afqt	0.0053186	0.0004775	11.1375178	0.0000000

f)

Of the variables I discuss, experience would likely be a better control than age here, given that the sample is restricted to a small range of ages. Additionally, occupation control could also be helpful. The survey team creates these based on answers to questions. As such, they could also be measured with error due to uninformative answers.