Question 6

```
\begin{tabular}{lrrlrrl}
\multicolumn{1}{c}{logwage} & \multicolumn{1}{c}{hgc} & \multicolumn{1}{c}{college} & \multi
Min.    & -0.9561   & Min.    & 5.00  & 0:1996 & Min.    & 0.0000   & 0:1700 \\
1st Qu. & 1.2012    & 1st Qu. & 11.00 & 1: 233 & 1st Qu. & 0.0000   & 1: 529 \\
Median  & 1.6897    & Median  & 12.00 &        & Median  & 0.0000   &        \\
Mean    & 1.6518    & Mean    & 12.45 &        & Mean    & 0.4289   &        \\
3rd Qu. & 2.1200    & 3rd Qu. & 14.00 &        & 3rd Qu. & 1.0000   &        \\
Max.    & 4.1660    & Max.    & 18.00 &        & Max.    & 1.0000   &        \\
NA's    & 684       &         &       &        &         &          &        \\
\end{tabular}
```

- **logwage**: The median log wage is approximately 1.69, with a range from approximately -0.96 to 4.17. A negative log wage might seem unusual because log transformation is typically applied to variables that are strictly positive.

- **hgc (Years of schooling)**: The median years of schooling is approximately 12 years, with a range from 5 to 18 years. This seems reasonable for a workforce dataset.

- **college (College education indicator)**: The data shows that 233 individuals have attended college, while 1996 have not. This distribution might need further examination based on the context of your analysis.

- **exper (Years of experience)**: The median years of experience is approximately 5.97, with a range from 0 to 25 years. These values are plausible.

- **married (Marital status indicator)**: There are 1415 married individuals and 814 unmarried individuals. This distribution seems reasonable.

- **kids (Number of children)**: The median number of children is 0, which seems low.

- **union (Union membership indicator)**: There are 529 individuals who are union members and 1700 who are not. This distribution could be valid.

At what rate are log wages missing? In about 30% of the cases, log wages are missing.

Do you think the logwage variable is most likely to be MCAR, MAR, or MNAR? MCAR - the missingness of log wages is unrelated to the observed or unobserved values in the dataset.

Question 7

(only complete cases)

```
Call:
lm(formula = logwage ~ hgc + union + college + exper + I(exper^2),
    data = complete_cases_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.32511 -0.43303  0.00805  0.44808  2.52985

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.161958   0.090400  12.854  < 2e-16 ***
hgc          0.034514   0.007210   4.787 1.80e-06 ***
union1       0.103406   0.054856   1.885   0.0596 .
college1    -0.114841   0.056058  -2.049   0.0406 *
exper        0.035362   0.008022   4.408 1.09e-05 ***
I(exper^2)  -0.002703   0.000504  -5.363 9.02e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6596 on 2223 degrees of freedom
Multiple R-squared:  0.02195, Adjusted R-squared:  0.01975
F-statistic: 9.978 on 5 and 2223 DF,  p-value: 1.864e-09
```

(mean imputation)

```
Call:
lm(formula = logwage_imputed ~ hgc + union + college + exper +
    I(exper^2), data = complete_cases_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.14385 -0.43986  0.02331  0.45580  2.55898

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.833530   0.113032   7.374 2.69e-13 ***
hgc          0.059042   0.009035   6.535 8.62e-11 ***
union1       0.221654   0.087410   2.536  0.01132 *
college1    -0.065139   0.105709  -0.616  0.53784
exper        0.050359   0.012646   3.982 7.15e-05 ***
I(exper^2)  -0.003691   0.001176  -3.137  0.00174 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.676 on 1539 degrees of freedom
Multiple R-squared:  0.03784, Adjusted R-squared:  0.03472
F-statistic: 12.11 on 5 and 1539 DF,  p-value: 1.596e-11
```

Tobit 2 model (sample selection model)

```
2-step Heckman / heckit estimation
2229 observations (684 censored and 1545 observed)
16 free parameters (df = 2214)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.55276    1.11124  18.495  < 2e-16 ***
hgc         -1.10366    0.06627 -16.655  < 2e-16 ***
union1      -1.11334    0.21334  -5.219 1.97e-07 ***
college1    -0.56499    0.22736  -2.485    0.013 *
exper       -0.50551    0.03011 -16.788  < 2e-16 ***
married1    -2.27529    0.16220 -14.027  < 2e-16 ***
kids         0.49540    0.11443   4.329 1.56e-05 ***
Outcome equation:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.446456   0.121902   3.662 0.000256 ***
hgc          0.091461   0.009789   9.344  < 2e-16 ***
union1       0.185728   0.084203   2.206 0.027507 *
college1     0.091996   0.100138   0.919 0.358357
exper        0.054162   0.012051   4.494 7.34e-06 ***
I(exper^2)  -0.001802   0.001094  -1.646 0.099828 .
Multiple R-Squared:0.0919, Adjusted R-Squared:0.0883
Error terms:
              Estimate Std. Error t value Pr(>|t|)
invMillsRatio -0.69455    0.06036  -11.51   <2e-16 ***
sigma          0.69571         NA      NA       NA
rho           -0.99833         NA      NA       NA
```

Comment on the differences of $\hat{\beta}_1$ across the models:

- Complete Cases Model: $\hat{\beta}_1 \approx 0.0345$
- Mean Imputation Model: $\hat{\beta}_1 \approx 0.059$
- Tobit 2 Model (Heckman selection model):
  - Outcome Equation: $\hat{\beta}_1 \approx 0.091$

Heckman selection model was far closer than the other two models.
What can you conclude about the veracity of the various imputation methods? I would trust the Heckman selection model much more than the others.

Question 8

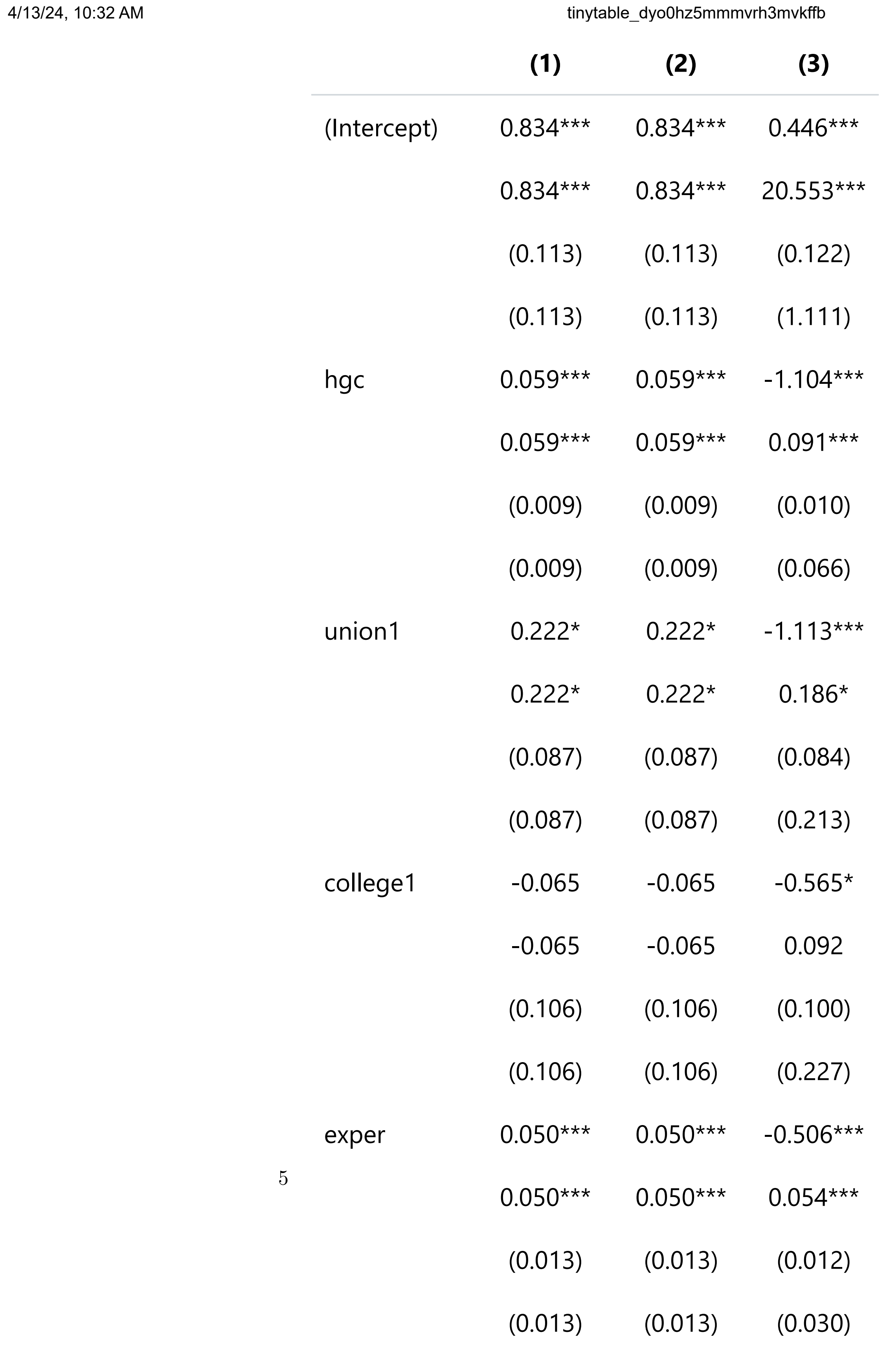
```
Call:
glm(formula = union ~ hgc + college + exper + married + kids,
    family = binomial(link = "probit"), data = wages_data)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -6.74260    0.80372  -8.389  < 2e-16 ***
hgc         -1.00903    0.09761 -10.337  < 2e-16 ***
college1     0.39722    0.42662   0.931  0.35181    
exper        1.84899    0.15594  11.857  < 2e-16 ***
married1     0.58780    0.20554   2.860  0.00424 ** 
kids         0.79927    0.20208   3.955 7.65e-05 ***
```

Question 9

```
Original   Counterfactual
0.2373394      0.2373394
```

This could be plausible, as being married and having kids are not likely to affect the chance of having union jobs. The union/employer is not discriminating.



	(1)	(2)	(3)
(Intercept)	0.834***	0.834***	0.446***
	0.834***	0.834***	20.553***
	(0.113)	(0.113)	(0.122)
	(0.113)	(0.113)	(1.111)
hgc	0.059***	0.059***	-1.104***
	0.059***	0.059***	0.091***
	(0.009)	(0.009)	(0.010)
	(0.009)	(0.009)	(0.066)
union1	0.222*	0.222*	-1.113***
	0.222*	0.222*	0.186*
	(0.087)	(0.087)	(0.084)
	(0.087)	(0.087)	(0.213)
college1	-0.065	-0.065	-0.565*
	-0.065	-0.065	0.092
	(0.106)	(0.106)	(0.100)
	(0.106)	(0.106)	(0.227)
exper	0.050***	0.050***	-0.506***
	0.050***	0.050***	0.054***
	(0.013)	(0.013)	(0.012)
	(0.013)	(0.013)	(0.030)