	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
logwage	1546	31	1.7	0.7	−1.0	1.7	4.2
hgc	14	0	12.5	2.4	5.0	12.0	18.0
college	2	0	0.1	0.3	0.0	0.0	1.0
exper	1932	0	6.4	4.9	0.0	6.0	25.0
married	2	0	0.6	0.5	0.0	1.0	1.0
kids	2	0	0.4	0.5	0.0	0.0	1.0
union	2	0	0.2	0.4	0.0	0.0	1.0

Table 1: Summary statistics of the dataset

1 Regressions and Models

2 Missing Data

Seems to be missing at random there is no apparent explanantion for it being missing that can be ascertained from the data-set although from a social standpoint one could say that maybe people didnt want to let others know what their salary is.

	Estimate	Std. Error	t value	Pr(>—t—)	
(Intercept)	0.911787	0.135133	6.747	2.13e-11	***
hgc	0.059042	0.009035	6.535	8.62e-11	***
union.L	0.156733	0.061809	2.536	0.01132	*
college.L	-0.046061	0.074747	-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	**
Residual standard error: 0.676 on 1539 degrees of freedom (684 observations deleted due to missingness)					

	Estimate	Std. Error	t value	Pr(>—t—)	
(Intercept)	0.911787	0.135133	6.747	2.13e-11	***
hgc	0.059042	0.009035	6.535	8.62e-11	***
union.L	0.156733	0.061809	2.536	0.01132	*
college.L	-0.046061	0.074747	-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	**
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					

	Estimate	Std. Error	t value	Pr(>—t—)	
(Intercept)	0.911787	0.135133	6.747	2.13e-11	***
hgc	0.059042	0.009035	6.535	8.62e-11	***
union.L	0.156733	0.061809	2.536	0.01132	*
college.L	-0.046061	0.074747	-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	**
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					

	(1)	(2)	(3)
(Intercept)	0.912	0.912	1.120
	(0.135)	(0.135)	(0.091)
hgc	0.059	0.059	0.036
	(0.009)	(0.009)	(0.006)
union.L	0.157	0.157	0.048
	(0.062)	(0.062)	(0.033)
college.L	−0.046	−0.046	−0.089
	(0.075)	(0.075)	(0.034)
exper	0.050	0.050	0.021
	(0.013)	(0.013)	(0.007)
	−0.004	−0.004	−0.001
	(0.001)	(0.001)	(0.000)
Num.Obs.	1545	1545	2229
R2	0.038	0.038	0.020
R2 Adj.	0.035	0.035	0.018
AIC	3182.4	3182.4	3808.4
BIC	3219.8	3219.8	3848.4
Log.Lik.	−1584.189	−1584.189	−1897.193
F		12.106	9.207
RMSE	0.67	0.67	0.57

Table 2: Model summaries of regressions

The LaTeX code for the remaining text is:

Results

The summary of the data in Table 1 shows that logwage has a high percentage (31 percent) of missing values. This missingness is suspected to be MAR as there is no apparent reason why it would be MNAR. The approach taken to deal with the missing values was to use mean imputation, which is a common method used in such cases. The data was split into two sets, one containing only complete cases and the other containing the mean-imputed data. A linear regression model was fitted to both datasets using the lm() function in R. The model formula used was $logwage \sim hgc + union + college + exper + I(exper^2)$.