Registered Nurse Hourly Wage Analysis:

Using 2012 U.S. Census Data to Create Predictions on Future Hourly Wages for Registered Nurses

Introduction

Salaries are important. Most if not everyone wants to know how much money they can make from a job they are applying for. As children we may have dreams of having a specific profession and not concern ourselves over what the salary for said profession might be. This does not change the importance of how much is to be earned, unfortunately. Considering this, I decided to examine the hourly wages of registered nurses to create a model for predicting wages in live data. This report aims to predict average hourly wages for registered nurses based on age as the main independent/explanatory variable accounting for a variety of additional factors (explained below) to create the best prediction of future earnings.

Data

The data used for analysis was 2012 U.S. Census data compiled by Gabor Bekes which includes a cross-section of 149,316 individuals as observations amongst various professions. Registered Nurses, the profession selected for this analysis, has 3455 observations. The dataset includes the following key variables used for analysis:

- **hrwage** – Average dollars earned per hour for each individual. This variable was created using two other variables within the original dataset, `earnwke` (dollars earned per week on average) dividing it by `uhours` (hours worked per week on average). This variable was used as the **dependent variable** for prediction analysis.
- **grade92** – Level of education for each observation ranging from less than high school to PHDs
- **race** – The listed race of the individual (Black, Asian, Pacific Islander, etc.)
- **age** – The age for each individual at the time the data was collected. This variable was selected as the **independent/main explanatory variable** for analysis
- **agesq** – The `age` variable squared
- **sex** – The individual's identified gender, male or female only
- **marital** – Whether the individual has never married, is separated, divorced, etc. Data is denoted by numbers
- **ownchild** – How many children the individual has of their own, 0 if none
- **occ2012** – The occupation code per U.S. Census data for each individual. Registered Nurses was selected, occupation code 3255
- **unionmme** – Whether the individual is a member of a union within their given profession, listed as 'Yes' or 'No'. This variable was converted into a dummy variable with 'Yes'=1 and 'No'=0

Null variables were examined in the dataset. There were 3264 null values within the `ethnic` variable of the original data and 610 null values within the `unioncov` variable. Both variables were examined, the `ethnic` variable determined what type of Hispanic the individual identified as, if they were Hispanic, and was listed as null if no data was entered. A decision was made to denote that all null values were to be replaced as "Missing" and that a missing entry was to mean that the individual was not Hispanic. The `unioncov` variable denoted whether the union member was not employed. For the purposes of this analysis this variable was not used and therefore the null values were not deemed relevant.

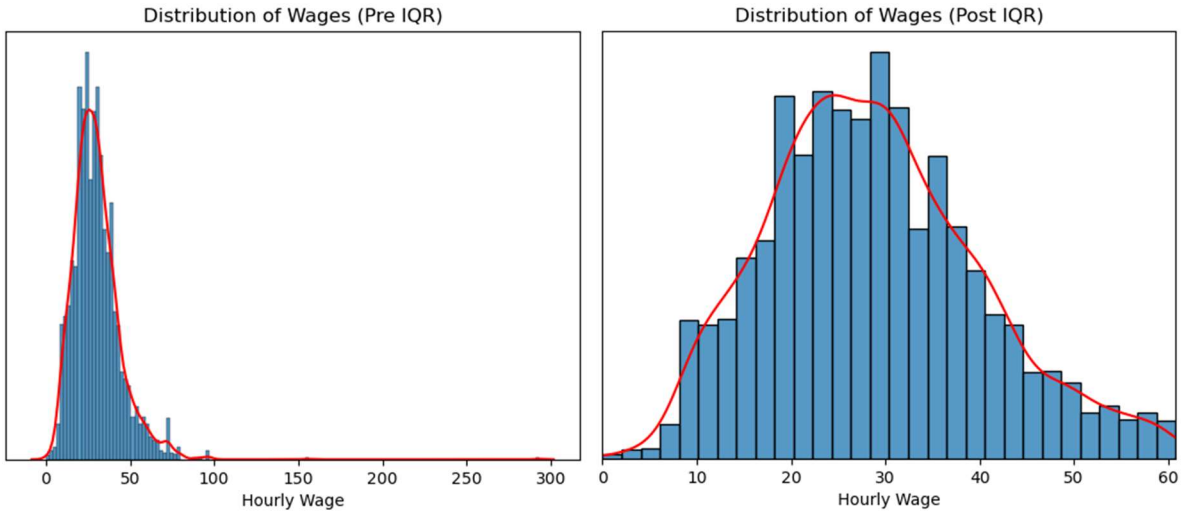
For further information as to the detail of each variable and how it is denoted within the original data, please refer to this link.

The following dummy variables were created from the dataset for further analysis:

- **black** – using `race`, equal to 1 if the individual is Black, 0 if not
- **asian** – using `race`, equal to 1 if the individual is Asian, 0 if not
- **hispanic** – using `ethnic`, equal to 1 if a value was present, 0 if it was "Missing"
- **female** – using `sex`, equal to 1 if the individual is female, 0 if not
- **one_child** – using `ownchild`, equal to 1 if household has 1 child, 0 if not
- **two_child** – using `ownchild`, equal to 1 if household has 2 children, 0 if not
- **three_plus_child** – using `ownchild`, equal to 1 if household has 3 or more children, 0 if not
- **now_single** – using `marital`, equal to 1 if individual is separated or divorced, 0 if not
- **never_married** – using `marital`, equal to 1 if individual was never married, 0 if not
- **associate** – using `grade92`, equal to 1 if an associate degree only was earned, 0 if not
- **higher_edu** – using `grade92`, equal to 1 if any degree above a bachelors was earned, 0 if not
- **no_degree** – using `grade92`, equal to 1 if no degree of any type was earned, 0 if not

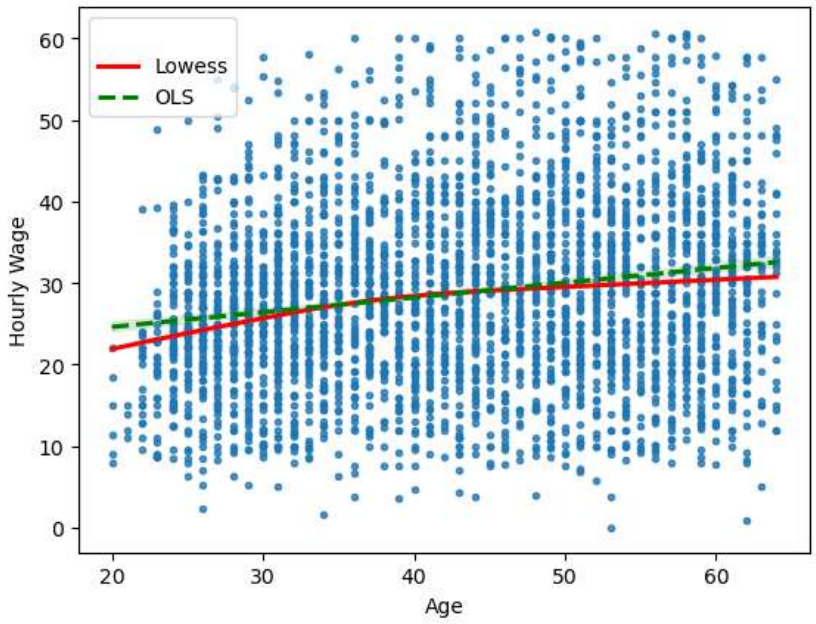
Data Review and Decisions

An examination of the data was conducted of the distribution of wages within the data to check for extreme values. A descriptive table was viewed showing the mean wage as ~$30/hr with the max value in the data was ~$293/hr. This is evidence of a fairly significant right skew in the data. The decision was made to use the 1.5 Interquartile Rage formula to identify outliers and exclude them from analysis. The data for analysis post IQR contained N=3341 observations. The following charts were created using the seaborn python package to display the distribution of wages pre and post IQR:



The distribution of the hourly wages post-IQR slightly skewed to the right and approximates a normal distribution. With this information we can use a level analysis of hourly wages without having to transform it to a log value for a normal distribution.

A robustness check was also done to check how a non-linear model compared to a linear model. Lowess and OLS regression models were compared again using seaborn:



From the review of the Lowess and OLS regression lines on the plot, the Lowess and OLS regression lines are mostly concurrent. The Lowess line deviates below roughly age 32 and above age 48 but only slightly. From this we can infer that a linear model, when appropriately fitted, can make for a good prediction model within the data.

Models

For this study I constructed multiple models for use in prediction in order from simplest to most complex. All models are constructed using a covariance type of HC1 for a heteroskedasticity and robustness check.

Model 1

$$hrwage = \beta_0 + \beta_1 age + \beta_2 agesq + \beta_3 female$$

Model 1 is a simple regression of `hrwage` on `age` with `female` as an explanatory variable and `agesq` to account for potential curvature in the data and for the potential influence of gender on hourly wage since gender pay gaps may exist in the data.

Model 2

$$hrwage = \beta_0 + \beta_1 age + \beta_2 agesq + \beta_3 female + \beta_4 nodegree + \beta_5 associate + \beta_6 higheredu$$

Model 2 adds the education dummy variables as explanatory variables to account for education potentially influencing the hourly wage. The level of education could affect the wage that a person is initially hired which may account for variance in hourly wages amongst people with different ages and levels of education.

Model 3

$$hrwage = \beta_0 + \beta_1 age + \beta_2 agesq + \beta_3 female + \beta_4 nodegree + \beta_5 associate + \beta_6 higheredu + \beta_7 black + \beta_8 asian + \beta_9 hispanic + \beta_{10} unionmme + \beta_{11} unionmme * age + \beta_{12} female * age$$

Model 3 adds the race dummy variables as explanatory variables in order to account for race potentially influencing the hourly wage as well as adding `unionmme` to account for union membership potentially affecting hourly wages. Also adding interaction terms between `female` and `age` as well as `unionmme` and `age`. The reasoning for this is that there may be an effect that gender has on the age viewed in the data, such as women maybe staying in the profession longer than men or visa versa. Similarly, union members may be older an average and using an interaction term between union members and age could account for this skew in the data.

Model 4

$$hrwage = \beta_0 + \beta_1 age + \beta_2 agesq + \beta_3 female + \beta_4 nodegree + \beta_5 associate + \beta_6 higheredu + \beta_7 black + \beta_8 asian + \beta_9 hispanic + \beta_{10} unionmme + \beta_{11} unionmme * age + \beta_{12} female * age + \beta_{13} one\_child + \beta_{14} two\_child + \beta_{15} three\_plus\_child + \beta_{16} never\_married + \beta_{17} now\_single + \beta_{18} never\_married * age + \beta_{19} now\_single * age$$

Model 4 adds all dummy variables accounting for children as well as marital status to account for families and/or children potentially influencing wages. The reasoning behind this is that the amount of children an individual has may potentially affect the quality of their work either positively or negatively, and could affect the amount of hours they can provide to a profession which could give license to an employer to hire them at a lower rate of pay. Similarly, if an individual was never married they potentially could have devoted more time to their career. Adding in these dummy variables can help account for the effects of these possibilities. The interaction terms between the marital status dummy variables can help account for the effect of age on whether a the individual in the data is married, currently single but previously married, or was never married, since it could be more likely that those who are divorced are older than those who are not, same for those are have never been married. These interaction terms help account for these scenarios.

Determining Best Predictive Model

A regression analysis was performed, and the results were displayed via Stargazer. To evaluate which model is the best suited to use for predictions of hourly wage, a Bayesian Information Criterion (BIC) test was performed on each model for comparison. The BIC test on each model gives an approximation of each model's fit given the number of parameters used for each fit, penalizing models that are potentially overfitted to the data. Along with this a Root Mean Squared Error (RMSE) test was done on the entire sample. This value always goes down the more complex the model, however it is instructive to compare in case of huge differences in RMSE between models. The stargazer results comparing each model is below:

	(1)	(2)	(3)	(4)
Observations	3341	3341	3341	3341
R^2	0.042	0.103	0.139	0.144
Adjusted R^2	0.041	0.102	0.136	0.139
Residual Std. Error	11.098 (df=3337)	10.743 (df=3334)	10.534 (df=3328)	10.517 (df=3321)
F Statistic	61.606*** (df=3; 3337)	69.130*** (df=6; 3334)	48.362*** (df=12; 3328)	31.795*** (df=19; 3321)
RMSE	11.091	10.732	10.513	10.485
BIC	25591.75	25395.88	25307.06	25346.01
Note:				*p<0.1; **p<0.05; ***p<0.01

From these results we can see that the model with the lowest BIC is model 3, our third most complex model. The first two models may not be complex enough and therefore are underfitted for the data but model 4 appears to be overfitted given that, despite being the most complex model, has a higher BIC score than model 3. Also, of note the RMSE scores for models 3 and 4, while continuing to decrease as mentioned above, are not too disparate, especially when compared to the difference in RMSE values between models 2 and 3. Along with this, a view of the Adjusted r^2 shows as well that model 4 captures the variance in the model .3% better than model 3, whereas model 3's Adjusted r^2 is 3.4% higher than model 2 (which has a greater 6.1% difference compared to model 1 which definitively is underfitting a line in the data). Our initial inference from these results is that model 3 is the best model to choose for use in predictions.

Cross-fold Validation

Cross-fold validation was performed on the data to better estimate the performance of each model on the new data. Five folds were selected to test each model. The results were as follows:

	Model1	Model2	Model3	Model4
Fold1	11.192307	10.828663	10.573649	10.546650
Fold2	11.009291	10.677538	10.483365	10.454148
Fold3	11.093651	10.727483	10.530777	10.504527
Fold4	11.042620	10.665740	10.434913	10.383836
Fold5	11.113534	10.749430	10.520046	10.494279
Average	11.090281	10.729771	10.508550	10.476688

The results show a very similar output to the RMSE from the main sample. The total average amongst all the folds is slightly higher for models 1 and 2 and lower for models 3 and 4, thereby increasing the gap between models 2 and 3. As such, it does not change our inference that model 3 is the best model to use for predictions.

Prediction Analysis

Having selected model 3 as the best model for use in predictions, we will now select values in order check the prediction of the model. Several predictions will be made. The prediction intervals for both 80% and 95%

Prediction 1

The following values will be selected for testing the first prediction with their results below:

- Bachelors: Yes
- Union member: Yes
- Age: 35
- Sex: Female
- Race: Black

PI	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
80%	28.869	0.861	27.765	29.972	15.324	42.413
95%	28.869	0.861	27.181	30.556	8.154	49.583

From the prediction results based on the values that were inputted, we can see that the average predicted hourly wage of a woman who is black, 35, and has a bachelor's degree (nothing higher) and who is a union member is 28.89/hr. Of note, the prediction intervals are quite disparate for both the 80% and 95% values, with the 80% PI being between 15.32/hr and 42.41/hr and the 95% between 8.15/hr and 49.58/hr.

Prediction 2

For the second prediction, the same values from prediction 1 were used except for changing union membership to 'No':

PI	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
80%	24.533	0.682	23.660	25.407	11.006	38.061
95%	24.533	0.682	23.197	25.869	3.845	45.222

Here we can see that by simply changing the union membership value to 'No' the average predicted hourly wage drops to 24.53. The prediction interval remains large. This implies that being a member of a union as a registered nurse can have a significant positive impact when predicting hourly wages.

Prediction 3

Another more prediction will be run this time using the following values:

- Age: 31
- Sex: Female
- Race: White
- Union Member: No
- Bacherlors: Yes

These values are being used in comparison due to the sample size being larger for these predictive values. Results below:

PI	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
80%	27.584	0.302	27.198	27.971	14.079	41.089
95%	27.584	0.302	26.993	28.175	6.930	48.238

With these new predictive values, we have a mean hourly wage of 27.58 while the prediction intervals remain large. This implies that even when using predictive values that are the most representative within the data, the model does not predict a significant difference in the average hourly wage of an individual.

Prediction 4

One more prediction will be run using the same values as prediction 1 except changing the race to “White”. Results below:

PI	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
80%	33.359	0.646	32.531	34.187	19.835	46.884
95%	33.359	0.646	32.093	34.626	12.675	54.044

Here we can see that simply by changing the race evaluated to “White” it has increased the predicted mean hourly wages to $33.36/hr implying a correlation between hourly wage and race.

Predictive Comparison Between Models

Despite having chosen model 3 to be the best fitting model for predictive purposes, it is still instructive to show a comparison of predictions between the models. Model 1 was chosen to show the difference between model 3 and an underfit model in model 1. The results are below using the values from prediction 3:

	Model1	Model3		Model1	Model3
Predicted	26.401	27.584	Predicted	26.401	27.584
PI_low(80%)	12.175	14.079	PI_low(95%)	4.645	6.930
PI_high(80%)	40.627	41.089	PI_high(95%)	48.158	48.238

From the comparison of models, we can see that there is a difference in the predicted values from the simplest model (model 1) to the more complex model (model 3). The prediction intervals are smaller, albeit slightly, and model 3 predicts a higher mean hourly wage average.

Conclusion

From the findings of the analysis, it can be shown that model 3, the model that regresses hourly wage on age and accounts for gender, education, race, and union membership and for agesq as a quadratic does the best job of predicting hourly wages for registered nurses. The overall fit using the data, however, does not account for much of the variance within it which is noted in the large prediction intervals that were seen. For example, while the mean average hourly wage predicted for prediction 3 is $27.58/hr, we can only be 80% sure that the true value lies within an approximately $27 range which is about as large as the mean wage itself. This unfortunately is not the most reliable prediction of the wages of registered nurses, despite being the best fit model based on the results. This could be due to the nature of the original data being Census data that, despite being polled amongst a wide range of population, the actual data points might contain errors from which we cannot do anything about, or the models created could be refined to better fit and reduce the variance in the prediction intervals. Regardless, the predictions that were run using the data can show some interesting notes about age and union membership, that union membership can have a positive impact on hourly wages and that, intuitively, the older you are the higher your average wage. Along with this, race seems to have an impact as well on hourly wages. The model showed an increase in hourly wages when evaluating only those individuals that identify as “white.” While the model might have large prediction intervals and may not give the most accurate prediction, it does still show instructive correlations on hourly wage.