Wages Prediction Modelling For The USA Computer Technology Occupations

Abstract

This paper investigates wage prediction models for computer technology occupations in the United States, employing a diverse set of predictors including age, gender, number of children in the household, employment class, education level, marital status, and citizenship status. The primary objective is to compare the performance between different models, from simple to complex, in predicting hourly wages for professionals in this sector. Four distinct models are considered, progressively incorporating additional predictors: a quadratic age model, a model adding gender and number of children, a model further incorporating employment class and education level and a model further accounting for marital and citizenship status. Through comparative analysis, I assess the predictive results of these models, seeking to identify the most effective model for wage estimation in computer technology occupations.

---

1. Data

Data Preparation

The data in the paper is from the cps-earnings dataset at https://osf.io/g8p9j/. I then applied some filtering and transformations to end up with the working dataset. The filtering and transformations are as follows:

1. Get the raw dataset (N = 149,316)
2. Apply filters (N = 4,304):
    - Computer tech occupation codes from 1005 to 1107
    - Employee with a high school diploma and above
    - Minimum 20 working hours per week
    - Hourly wages of more than $1/hr
3. Transform some columns (N = 4,304):
    - Get hourly wages by dividing weekly wages by the number of working hours
    - Create a boolean variable for gender
    - Transform citizenship status text to numerical values
    - Transform employment class (public, private, self-employed,...) text to numerical values
    - Create dummy variables for citizenship status, employment class, marital status and education level
    - Add the square of age for non-linear regression

The models in this paper use hourly wages as the target variable. Other variables are used as predictors including:

- Age
- Square of Age
- Gender
- Number of children in the household
- Employment class
- Education level
- Marital status
- Citizenship status

Data Exploration

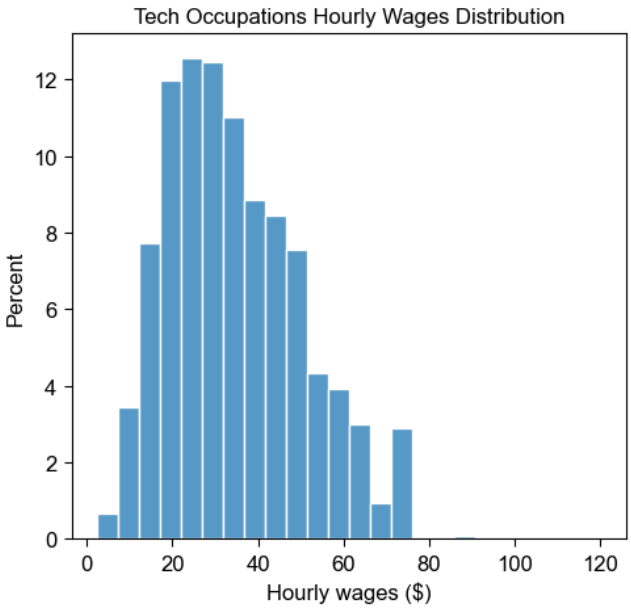

Fig. 1: Tech Occupations Hourly Wages Distribution

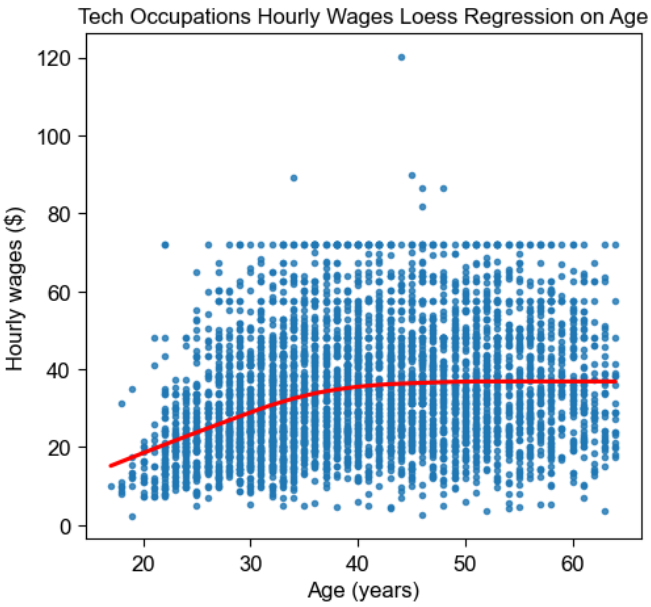

Fig. 2: Tech Occupations Hourly Wages Loess Regression on Age

From Fig. 1, it is clear that the target variable hourly wages is left-skewed. This means that the mean value of hourly wages is not the typical value. Within this paper, the models will assume that the mean is typical and will predict the mean value. It is worth noting that normally this is not the case and there might be impacts on the prediction results made by the models.

In Fig. 2, I construct the loess regression of hourly wages on age as the primary independent variable. The loess line shows a positive trend from age 15 to 35, then gradually flatten from 35 onward. There is a non-linear relationship between hourly wages and age. Thus, a square term is included in the subsequent model used for hourly wage prediction to represent this non-linearity.

---

2. Models

For this paper, I constructed four predictive models, from simple to complex, to try to predict the hourly wages for a professional in the computer technology sector. Each model is a linear regression with the OLS method, where the target variable is hourly wages and the predictors as listed in the **Data** section above.

Model 1 - Hourly wages on age and square of age

$$Hourly\ wages = \beta_0 + \beta_1 \times age + \beta_2 \times age^2 + \epsilon$$

Model 1 only uses age and the square of age as predictors for hourly wages. From intuition, employees of higher age usually have more experience and expertise than younger employees. Hence, age often correlates with professional experience and expertise, which can influence wages. The inclusion of age squared allows for capturing potential non-linear relationships, acknowledging that the impact of age on wages is non-linear as in the loess regression above.

Model 2 - Extension of model 1 with gender and number of children

$$Hourly\ wages = \beta_0 + \beta_1 \times age + \beta_2 \times age^2 + \beta_3 \times isFemale + \beta_4 \times numberChildren + \epsilon$$

Model 2 extends with the inclusion of gender and the number of children in the household as predictors. Gender has often been a focal point in discussions to pay equity. By incorporating gender as a predictor, model 2 aims to cover the wage disparities between male and female professionals. Additionally, the inclusion of the number of children variable represents any potential differences in hourly wages for professionals with children. Employees having more children might demand more wages to fulfill their responsibilities raising their offspring.

Model 3 - Extension of model 2 with employment class and education level

$$Hourly\ wages = \beta_0 + \beta_1 \times age + \beta_2 \times age^2 + \beta_3 \times isFemale + \beta_4 \times numberChildren$$
$$+\beta_i \times C(employmentClass) + \beta_k \times C(educationLevel) + \epsilon$$

Model 3 extends from model 2 with employment class and education level variables. Professionals who are employed in the public sector might receive different wages than those working in the private sector or self-employed. Furthermore, it is intuitive that professionals with higher education levels like a Master's degree or PhD degree might receive higher wages with their advanced knowledge or training. The inclusion of these variables acknowledges the potential variation in wages with the employment class and level of education.

Model 4 - Extension of model 3 with marital status and citizenship status

$$Hourly\ wages = \beta_0 + \beta_1 \times age + \beta_2 \times age^2 + \beta_3 \times isFemale + \beta_4 \times numberChildren$$
$$+\beta_i \times C(employmentClass) + \beta_k \times C(educationLevel) + \beta_j \times C(maritalStatus) + \beta_l \times C(citizenshipStatus) + \epsilon$$

Finally, model 4 further extends from model 3 with the inclusion of citizenship status and marital status variables. The USA attracts a large number of immigrant professionals to join the US workforce. Hence, the addition of the citizenship status in this model is necessary to represent the potential wage gap among immigration statuses. With martial status, professionals with different family obligations might have some relationship with their received hourly wage. Model 4 includes marital status variables to suggest this variation in hourly wage.

Out[10]:

	Dependent variable: hourly_wage			
	(1)	(2)	(3)	(4)
Intercept	-19.220***	-14.241***	-14.528***	-11.694***
	(2.755)	(2.901)	(2.868)	(3.135)
Age	2.387***	2.147***	1.813***	1.701***
	(0.143)	(0.153)	(0.147)	(0.153)
Square of Age	-0.024***	-0.021***	-0.017***	-0.016***
	(0.002)	(0.002)	(0.002)	(0.002)
is Female		-5.430***	-5.114***	-5.028***
		(0.519)	(0.492)	(0.495)
Number of children		0.900***	0.804***	0.565**
		(0.242)	(0.230)	(0.249)
Employment class (4 variables)			Yes	Yes
Education level (7 variables)			Yes	Yes
Marital status (6 variables)				Yes
Citizenship status (4 variables)				Yes
Observations	4304	4304	4304	4304
R^2	0.096	0.122	0.226	0.232
Adjusted R^2	0.096	0.122	0.223	0.227
Residual Std. Error	14.872 (df=4301)	14.657 (df=4299)	13.784 (df=4288)	13.748 (df=4278)
F Statistic	310.949*** (df=2; 4301)	196.409*** (df=4; 4299)	108.346*** (df=15; 4288)	68.085*** (df=25; 4278)
BIC	35473.68	35363.1	34915.51	34966.64
RMSE	14.867	14.649	13.759	13.707
CV RMSE	14.865	14.646	13.752	13.697
Note:				*p<0.1; **p<0.05; ***p<0.01

Fig. 3: Summary of the regression table

3. Model Performances

From Fig. 3, some analysis of the performance across all models can be made. The R-squared increases consistently from model 1 to model 4 (0.096 - 0.227). This suggests the additional variables have some relationships with the hourly wages and help explain the variation in hourly wage. The RMSE in the full sample across the four models is also decreasing, meaning that more predictors improve the model's generalization of the data and accuracy of the prediction. The same trend is observed with the RMSE from the 5-fold cross-validation. It is worth noting that the RMSE from the cross-validation of each model is smaller than its RMSE in the full sample. The cross-validated RMSE should be slightly larger than the full sample RMSE since each prediction is assessed on unseen data in the test split to mitigate the risk of overfitting to the training data. Unexpectedly, this is not the case with the four models in the paper.

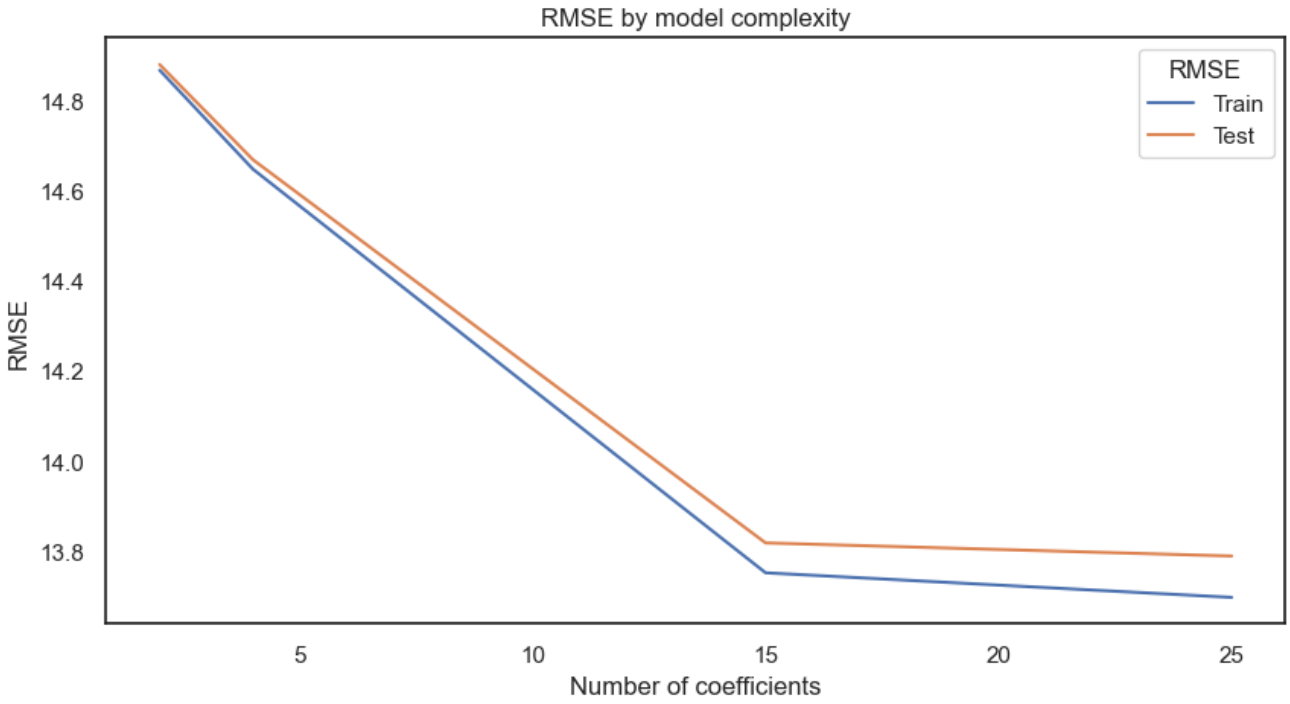

Fig. 4: RMSE by model complexity

So far, it seems that model 4 has the best performance with the lowest RMSE and cross-validated RMSE: 13.707 and 13.697 respectively. However, the previous model's performances are closely behind since the differences in the RMSE between each model are not great. On the other hand, from Fig. 4, the cross-validated RMSE difference between the train and test split of each model grows. While including more predictors can improve prediction accuracy, the more complex models might experience overfitting, especially with model 4. Looking at Fig. 3's BIC, it is decreasing from model 1 to model 3, then increasing between model 3 and 4. It seems like model 4 is indeed experiencing some overfitting. Hence, while having a slightly larger RMSE than model 4, model 3 has less complexity and is preferred to be the better model to mitigate overfitting.

---

4. Limitation

The models in this paper have some noticeable limitations worth mentioning. Firstly, the models assume that the mean value of hourly wages is the typical value, although this is in fact not the case from Fig. 1. Secondly, models 1 through 4 do not consider any interaction between the predictors, which might present in the dataset. Finally, the models are built with related predictors grouped together via intuition, which might include some unimportant predictors that unnecessarily complicate the models. With all of these limitations, the models might suffer losses in prediction accuracy and generalization performance.

Appendices

Appendix A

Full regression table of all models

Out[13]:

				Dependent variable: hourly_wage
	(1)	(2)	(3)	(4)
Intercept	-19.220***	-14.241***	-14.528***	-11.694***
	(2.755)	(2.901)	(2.868)	(3.135)
age	2.387***	2.147***	1.813***	1.701***
	(0.143)	(0.153)	(0.147)	(0.153)
agesq	-0.024***	-0.021***	-0.017***	-0.016***
	(0.002)	(0.002)	(0.002)	(0.002)
citizen_2				2.364***
				(0.778)
citizen_3				2.506***
				(0.757)
citizen_4				4.544**
				(2.011)
citizen_5				0.988
				(1.332)
class_2			-3.778***	-3.633***
			(0.878)	(0.874)
class_3			0.451	0.812
			(0.973)	(0.986)
class_4			-7.140***	-6.814***
			(0.785)	(0.789)
class_5			-6.260***	-6.152***
			(1.066)	(1.058)
edu_40			0.683	0.613
			(0.881)	(0.878)
edu_41			-0.475	-0.453
			(1.162)	(1.158)
edu_42			0.514	0.474
			(0.964)	(0.964)
edu_43			8.751***	8.386***
			(0.772)	(0.774)
edu_44			11.585***	10.540***
			(0.869)	(0.901)
edu_45			15.218***	14.387***
			(2.708)	(2.663)
edu_46			17.901***	16.352***
			(2.347)	(2.352)
female		-5.430***	-5.114***	-5.028***
		(0.519)	(0.492)	(0.495)
marital_2				1.748
				(2.919)
marital_3				-0.675

				(2.026)
marital_4				-0.374
				(2.589)
marital_5				-1.091
				(0.859)
marital_6				-0.893
				(1.614)
marital_7				-1.216**
				(0.604)
ownchild		0.900***	0.804***	0.565**
		(0.242)	(0.230)	(0.249)
Observations	4304	4304	4304	4304
R^2	0.096	0.122	0.226	0.232
Adjusted R^2	0.096	0.122	0.223	0.227
Residual Std. Error	14.872 (df=4301)	14.657 (df=4299)	13.784 (df=4288)	13.748 (df=4278)
F Statistic	310.949*** (df=2; 4301)	196.409*** (df=4; 4299)	108.346*** (df=15; 4288)	68.085*** (df=25; 4278)
BIC	35473.68	35363.1	34915.51	34966.64
RMSE	14.867	14.649	13.759	13.707
CV RMSE	14.865	14.646	13.752	13.697
Note:				*p<0.1; **p<0.05; ***p<0.01

Appendix B

80% Prediction Interval of all models

Out[14]:

	Model1	Model2	Model3	Model4
Predicted	29.450556	26.447708	29.429267	28.715455
PI_low(80%)	10.387172	7.646103	11.706046	10.929622
PI_high(80%)	48.513939	45.249313	47.152489	46.501288