Data Analysis 3 Assignment 1
Balint Thaler
GitHub repository: https://github.com/balint-thaler/DA3-Assignment-1

Assignment 1

The data

The data used for the current exercise is the CPS dataset. The dataset contains information on individuals, their work and their earnings. For the current analysis, a singular occupation will be used which is civil engineers. The dataset contains 396 observations for this occupation with most variables having no missing values. For target variable the hourly wages are calculated for all observations using the weekly hours worked and weekly earnings. The hourly wages are then not converted further so as to be able to predict actual wage values instead of percentage changes.

Exploratory data analysis (EDA)

The distribution of hourly wages is almost symmetric (see chart 1 below), with a mean of 35.8 and a median of 32.5. There is an extremely high value of 180, however after observing its features, there is no obvious reason to assume that it is an error, therefore it is not deleted.

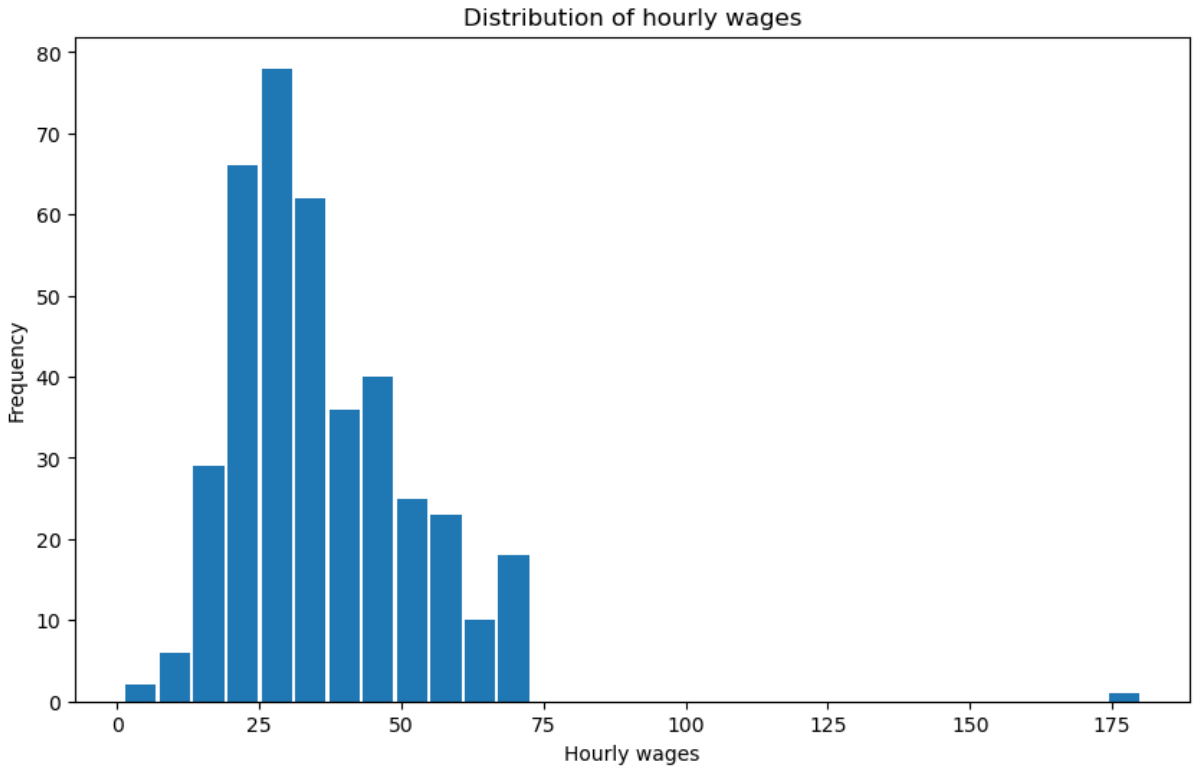

Chart 1: Distribution of hourly wages

The age of the civil engineers has a mean of 42.2 years and range from 20 to 64. Less than 20% of the dataset consists of women, and the average hourly wages of the genders in the dataset differ as well with men earning 36,3 and women earning 33.6 USD. The average hourly wages are higher for older, union members, non-whites, people with higher education levels.

The predictive models

There were four models created each with increasing complexity. Model 1 contains only age as a predictor variable, additionally to this Model 2 includes age squared and cubed to account for a potentially non-linear association and gender. Model 3 includes education on top of Model 2's variables, and finally Model 4 includes union membership and race as well on top of Model 3's variables. The covariates and their respective coefficients are summarised in Table 1.

	Dependent variable: wage			
	Model 1	Model 2	Model 3	Model 4
Age	0.36***	7.36***	5.55**	5.60**
	(0.06)	(2.42)	(2.30)	(2.31)
Age squared		-0.15***	-0.11**	-0.11**
		(0.06)	(0.06)	(0.06)
Age cubed		0.00**	0.00*	0.00*
		(0.00)	(0.00)	(0.00)
Gender		2.02	1.83	1.85
		(1.98)	(1.87)	(1.86)
BA education			7.86***	8.01***
			(2.91)	(2.89)
MA education			10.49***	10.71***
			(3.15)	(3.18)
PhD education			36.08**	36.40**
			(14.93)	(14.41)
Race				-0.00
				(0.89)
Union membership				1.44
				(2.15)
Intercept	20.83***	-80.05**	-63.86**	-64.52**
	(2.38)	(31.86)	(29.67)	(29.63)
Observations	396	396	396	396
R^2	0.07	0.09	0.16	0.16
Adjusted R^2	0.07	0.08	0.15	0.15
Residual Std. Error	16.03 (df=394)	15.88 (df=391)	15.30 (df=388)	15.34 (df=386)
F Statistic	39.65*** (df =1; 394)	18.24*** (df=4; 391)	11.99*** (df=7; 388)	9.82*** (df=9; 386)
BIC	3331.1	3338.59	3324.15	3335.8
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 1: summary of Models with heteroscedasticity robust standard errors in parenthesis

Evaluating the models

To assess the predictive power of the models, their BIC scores, their root mean squared errors (RMSE) and their cross-validated RMSEs will be compared.

Based on the BIC scores, Model 3 performs the best with a score of 3324.16. However, based on RMSE of the total dataset, the most complex Model 4 performs the best with a value of 15.141. This is however only marginally better than Model 3 with 15.147. Due to the simpler nature of the latter one, it is the preferred model.

To get the most accurate evaluation however we need to use the cross-validated RMSE scores. After iterating through the dataset with a 4-fold cross validation. These values show a similar ordering of the models to that of the full-sample RMSE: Model 4 performs the best with an RMSE of 15.00 however it is very close to Model 3's value of 15.03. Again, given that Model 3 is a simpler model, it is the preferred choice once again despite Model 4 being slightly less underfitted.

To get an insight into the prediction power of the models, their respective prediction intervals are calculated for an imaginary civil engineer who is a white 45-year-old male with a masters degree and is also a union member. The predicted hourly wage for this person (as predicted by the preferred choice, Model 3) is 40.3 USD. The prediction intervals are however rather wide even with a confidence of 80%, and even wider with 95% - as seen in the table below.

	80% confidence	95% confidence
Predicted wage (USD/hour)	40.3	40.3
PI lower bound	20.5	10.1
PI upper bound	60.0	70.4

Table 2: Prediction intervals of Model 3

We also see in Table 3 that model 4's prediction is rather close to model 3's, which again confirms that despite the fact that the RMSE scores suggest that model 4 is less underfitted, the additional complexity of the model offsets the marginal gain on accuracy.

	Model 3	Model 4
Predicted wage (USD/hour)	40.3	41.5
PI lower bound (80%)	20.5	21.6
PI upper bound (80%)	60.0	61.5

Table 3: predictions of Model 3 and 4

Conclusion

Overall, we can conclude that Model 3 would be the preferred choice from the presented models. We have also seen that even the "best" model contains high level of uncertainty around the predicted value. The RMSE scores of around 15 are rather high especially considering that the mean value of the hourly wages was 35.8. This is likely due to the fact that the models are rather simple and do not capture the right interactions or the right variables.