Data Analysis 3: Assignment 1 Report
Corinne Williams

Introduction:
The goal of this assignment was to create models with predictor variables to predict earnings per hour for a certain occupation. Instead of actually predicting data, measure of fit values were calculated in order to see how models differ from one another, depending on the technique used to procure the values (BIC in the full sample, K-Fold cross-validated RMSE, and RMSE in the full sample). The significance of this is to show the trade offs that one would encounter and decide upon to choose which model would be best to use when predicting live data.

Exploratory Data Analysis:
I went through many occupations before settling on those with legal professions (namely lawyers, judges, magistrates, and other judicial positions) as it had a large number of observations (1,666 rows in the dataset). The outcome variable was to be earnings per hour, so I created a new variable, 'earnperhr' out of two existing variables, dividing weekly earnings by usual hours worked. I then chose predictor variables to be used in the models. I wanted to use the most important variables, as established by Gabor Bekes (age, sex, usual hours worked, and highest education level completed). However, I chose some additional variables that I believed might also provide an interesting insight into why legal employees are paid what they are. The additional variables were: state, race, U.S. citizenship, marriage status, and number of children.

Data Work:
For missing values in usual hours of work and earnings per week, I calculated how many missing values were in each variable. As there were not too many null observations, I decided not to drop them and rather to impute. I created histograms to track the distribution of each variable in order to choose the best measure of central tendency for imputation. The distribution for usual hours worked had long tails on either side and therefore was not normally distributed. I chose to impute the missing values using the median from the other non-null observations. The distribution for earnings per week was normally distributed for the most part, despite the last bin containing over 300 observations resulting in a small bias in that direction. Because it was mostly normal, I decided to impute using the mean of the other non-null observations. I then cleaned the other predictor variables as needed and created dummy variables for use in the regressions. The state variable became split into nine categories based on region, and then made into dummy variables. The race variable was split into ten categories, and then converted into dummies. The marital variable was split into observations of 'married' and 'single' and then converted into dummies. From the sex variable, a 'female' variable dummy was created. From the 'grade 92' variable (highest level of education completed), I split the variable into high and low education based on education levels found in the sample, and then converted them into dummies. The U.S. citizenship variable was also converted into a dummy variable stating whether someone had U.S.

citizenship or not. Lastly, the variable for number of children was split into separate variables for each additional child and also converted into dummies.

Model Creation:
For model creation, I decided to split my nine main variables (nine without dummy variable add-ons) into four parts for each model, with increasing model complexity.
Model 1 predictor variables: age and sex.
Model 2 predictor variables: age, sex, usual hours worked, highest level of education completed.
Model 3 predictor variables: age, sex, usual hours worked, highest level of education completed, state, race, U.S. citizenship.
Model 4 predictor variables: age, sex, usual hours worked, highest level of education completed, state, race, U.S. citizenship, marital status, number of children.

BIC in the full sample:
I created OLS linear regressions for each model. Each also utilized an HC1 robust standard error estimator to adjust for heteroscedasticity. I printed the summaries of each model along with the BIC values of each to measure goodness of fit while also penalizing for model complexity. Model 2 had the lowest, and therefore best, BIC value with approximately 9,558. However, looking at the R-squared values for each model, it can be assumed that the models created were not very good in general, and maybe why the overall BIC values seem to be very high. For example, the R-squared value of Model 2 is only 0.076, meaning that only 7.6% of the variance in earnings per hour is explained by the model. This means that perhaps there were better predictor variables that could have been used, others that could have been left out, interactions that were not explored, and/or predictor variables that are highly correlated with one another.

K-Fold Cross-Validated RMSE:
I split the dataset into two sets of training data and test data in order to imitate and manufacture original and live data. 80% of the observations belonged to the training set while 20% belonged to the test set. I created these sets five times (5 folds) and calculated the RMSE (root mean squared error) for each fold for each model, along with the average for each model. The K-Fold Cross-Validation technique performed well and the model with the lowest RMSE was the most complex, model 4. The average RMSE for model 4 was 16.5 USD, meaning that on average, model 4's predictions deviate from the actual values of earnings per hour by about 16.5 USD. While not great seeing as how 75% of all observations have employees making under 51 USD an hour, the RMSE calculated here is still much better than the RMSE taken from the full sample.

RMSE in the full sample:
Using the same regressions/models, I calculated RMSE for each model from the full sample of original data. All of the RMSE values were fairly similar, ranging from 28.46 at the highest and 26.14 at the lowest. Similarly to the RMSE calculated with K-Fold Cross-Validation, Model 4

performed the best out of the others (RMSE of 26.14). In comparison to the average K-Fold cross-validated RMSE however, this is much worse. The RMSE in the full sample says that the predicted values deviate on average from the actual values of earnings per hour by 26.1 USD.

Summary:
Lastly, I created a table comparing each measure of fit (BIC in the full sample, average K-Fold Cross-Validated RMSE, and RMSE in the full sample) for each model. From the comparison, it was seen that when using BIC in the full sample (penalizing for complexity) Models 1 and 2 performed better than 3 and 4 due to the lower BIC scores. However, as I stated before, their low R-squared measures likely mean that the models do not fit the data well in general. From the RMSE values in general (not penalizing for complexity), Model 4 outperforms the others in both cross-validated RMSE and RMSE taken from the full sample. However, from this experience, in the future rather than calculating RMSE from the full sample as a way to choose the best model, I would skip that completely and simply use cross-validated RMSE.
There is a clear trade off between model complexity and predictive accuracy seen in the comparisons between BIC and RMSE. BIC is more concerned about penalizing model complexity- therefore, providing a model with good fit and simplicity. It therefore also protects against overfitting the original data. RMSE with cross-validation also helps protect against the possibility of over- or underfitting the data; however overall, RMSE is not concerned with penalizing models based on their complexity and instead focuses on predicting accurate targets in the live data. This trade off is important to consider when deciding which model to use in order to predict the outcome variable in the future live data.

	Model 1	Model 2	Model 3	Model 4
Full Sample BIC	9594.08	9558.98	9644.23	9672.22
Avg Cross-Validated RMSE	17.43	17.05	16.71	16.59
Full Sample RMSE	28.46	26.40	26.20	26.14

Comparison table showing each measure of fit for each model