Data Analysis 3 - Assignment 1

Ali Sial

1/26/2022

Introduction

The aim of this document is to build models and **predict earnings per hour** using linear regression. The data used for this analysis is retrieved from ***cps-earnings dataset*** (please click for the link) and the profession selected is **Secretaries and Administrative Assistants** (Occupation Code: 5700). The data has a total of 3511 observations for this occupation, majority being females.

Data Cleaning and Mungging

The first step after loading the data was to filter observations for full-time employees (at least 40hrs per week) and also removed variables with missing values. After this our target variable was added to the data that is "Earnings per Hour". Upon completing the basic cleaning, I investigated each predictor to check if further feature engineering is required. A squared term was added for "Age" to incorporate all observations when running linear regression. I also dropped observation that had education level higher than Masters and regrouped the remaining in a new variable. Dummy variables were created for gender, race and own child. To avoid complexity the States were regrouped into four regions. The final observations that will be used for analysis are **2576**.

Selecting Interactions and Prediction Models

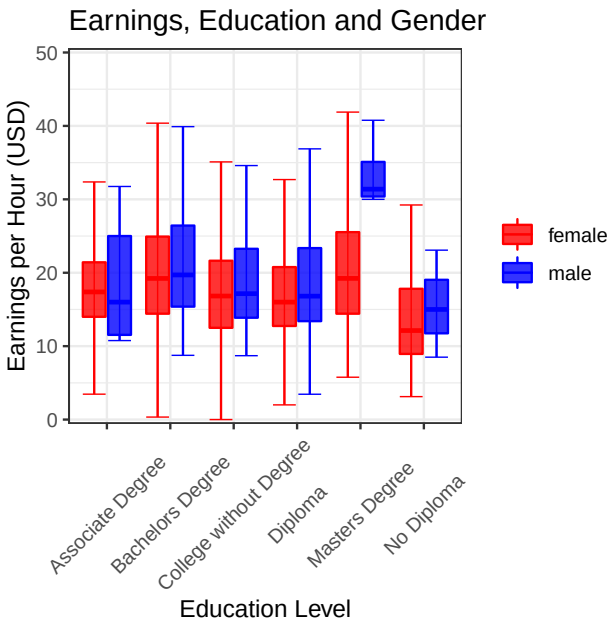
Understanding interactions within the variables was a challenging task, but in the end out of the 11 predictors used for modeling, gender, race and union membership had significant interaction terms. Total 11 interaction terms are used to build the models, majority for gender. The most complex interaction that uses 3 predictors is between gender, race and education level. Four models were built using predictors and interactions. The first model is the simplest with only one variable which is education level. In the second model i also included age, age square and gender. The third model includes all the variables and an interaction term for gender and education level. The last model is the most complex with all predictors and the interaction terms. Next step was to run regression and cross validate the performance of models.

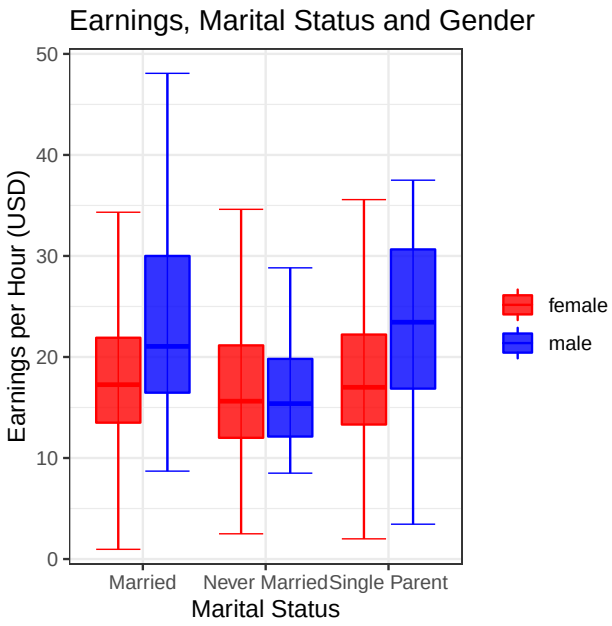
Interpreting the Results

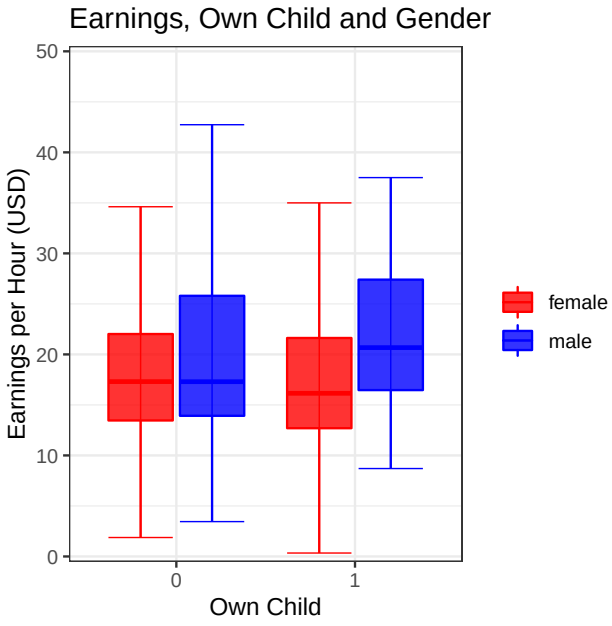
The performance of these prediction models was first compared using Root Mean Square Error (RMSE). After running the regressions model 4 had the lowest RMSE (the most complex model) i.e. 7.8509. Next, I cross validated RMSE using the K-Fold method by creating 4 training and test samples. The cross validated RMSE was lowest for model 3, but the difference between model 2 and model 3 was very minute. Based on this comparison, even though model 3 had the lowest RMSE, model 2 is less complex, thus I would prefer that over model 3. This can be further validated by using Bayesian Information Criterion (BIC), which adds a penalty for model complexity. When comparing the models using BIC, the model 2 had lowest BIC value. Therefore, we can conclude that model 2 would the most preferred model for predicting earnings per hour for this specific occupation.

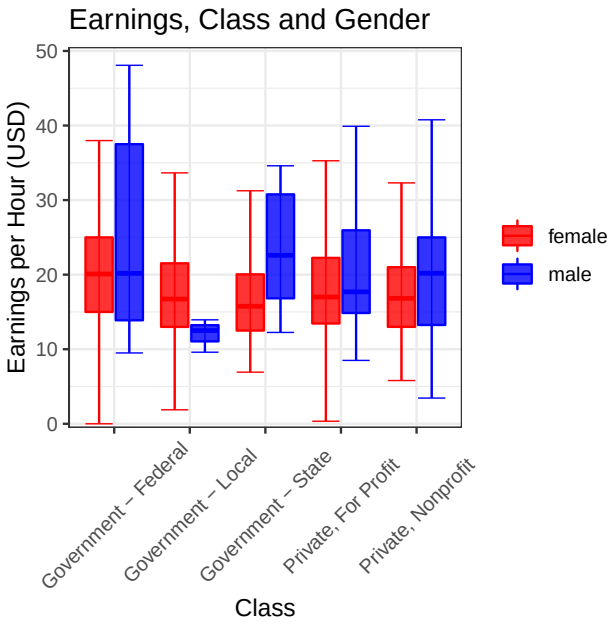
Appendix

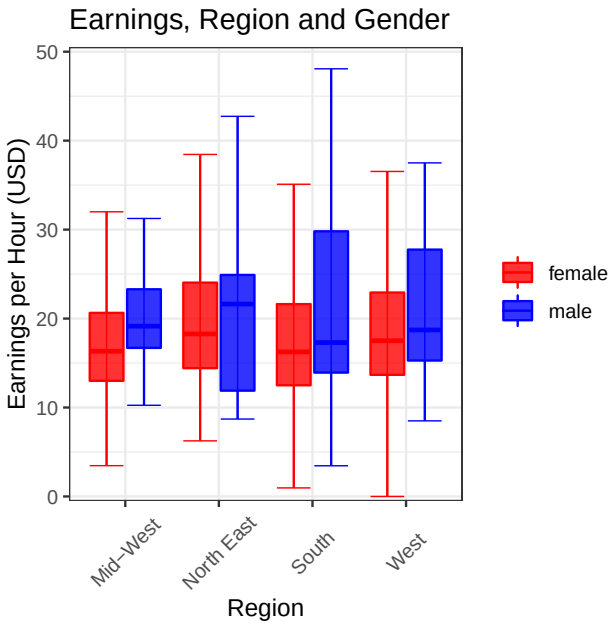
Interactions with Gender

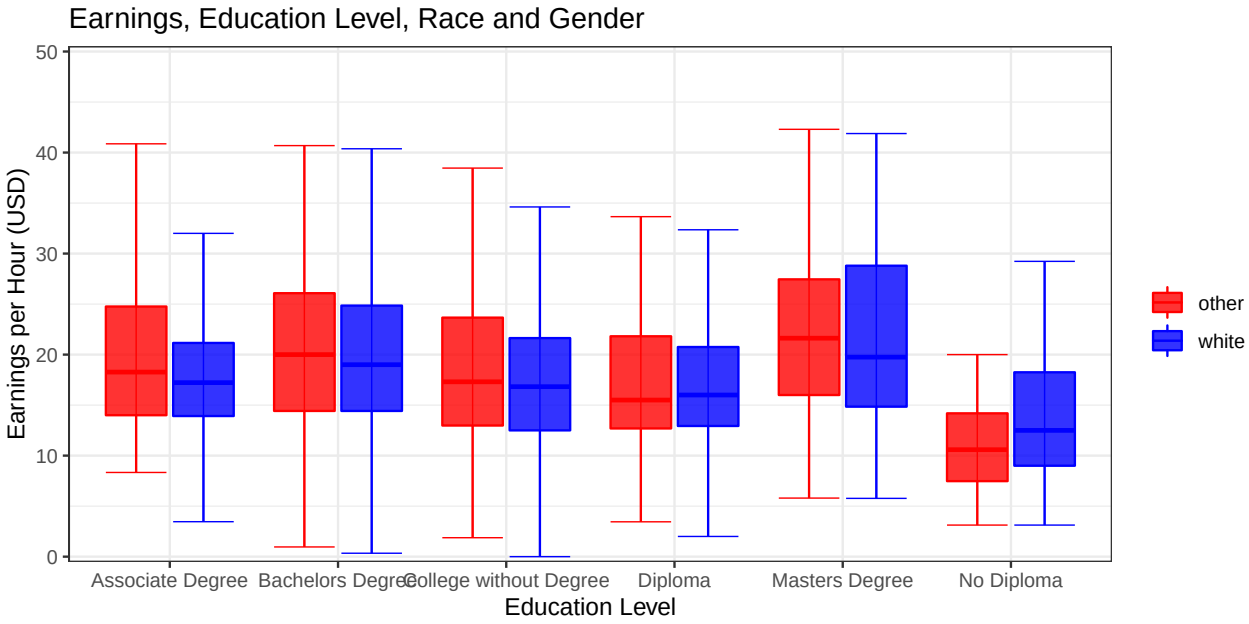




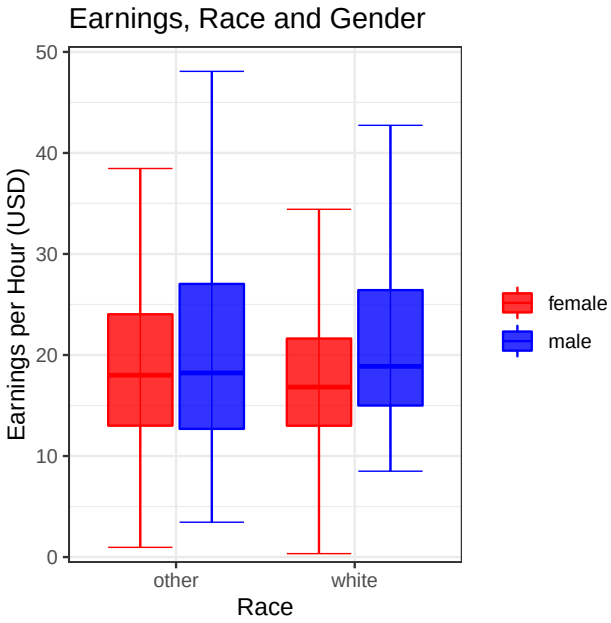


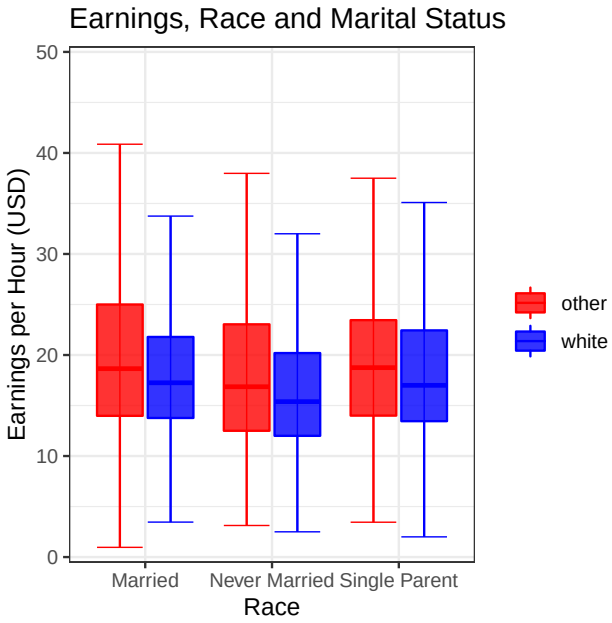

Earnings, Class and Gender
Earnings per Hour (USD)
0
10
20
30
40
50
Government – Federal
Government – Local
Government – State
Private, For Profit
Private, Nonprofit
Class
female
male


Earnings, Region and Gender
Earnings per Hour (USD)
0
10
20
30
40
50
Mid-West
North East
South
West
Region
female
male

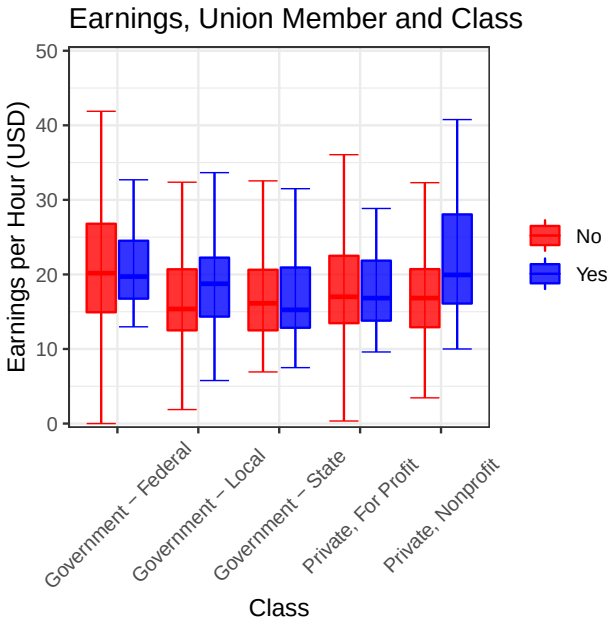

Earnings, Education Level, Race and Gender
Earnings per Hour (USD)
0
10
20
30
40
50
Associate Degree
Bachelors Degree
College without Degree
Diploma
Masters Degree
No Diploma
Education Level
other
white

Interactions with Race





Interactions with Union Membership



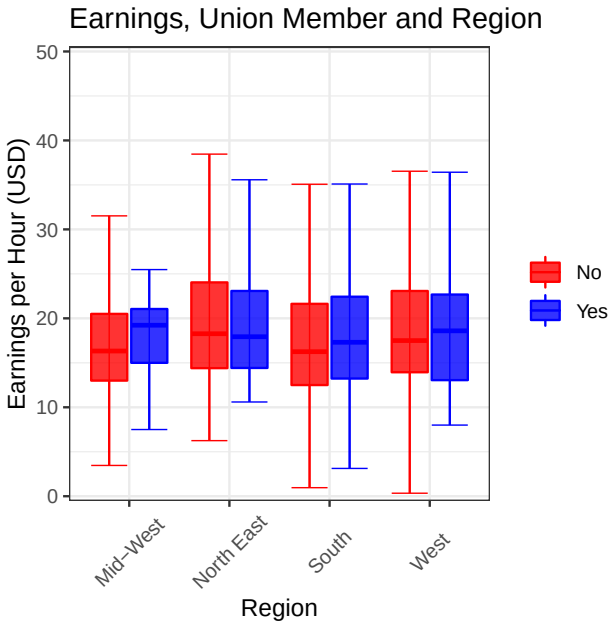


Table 1: Model evaluation based on full sample RMSE and BIC

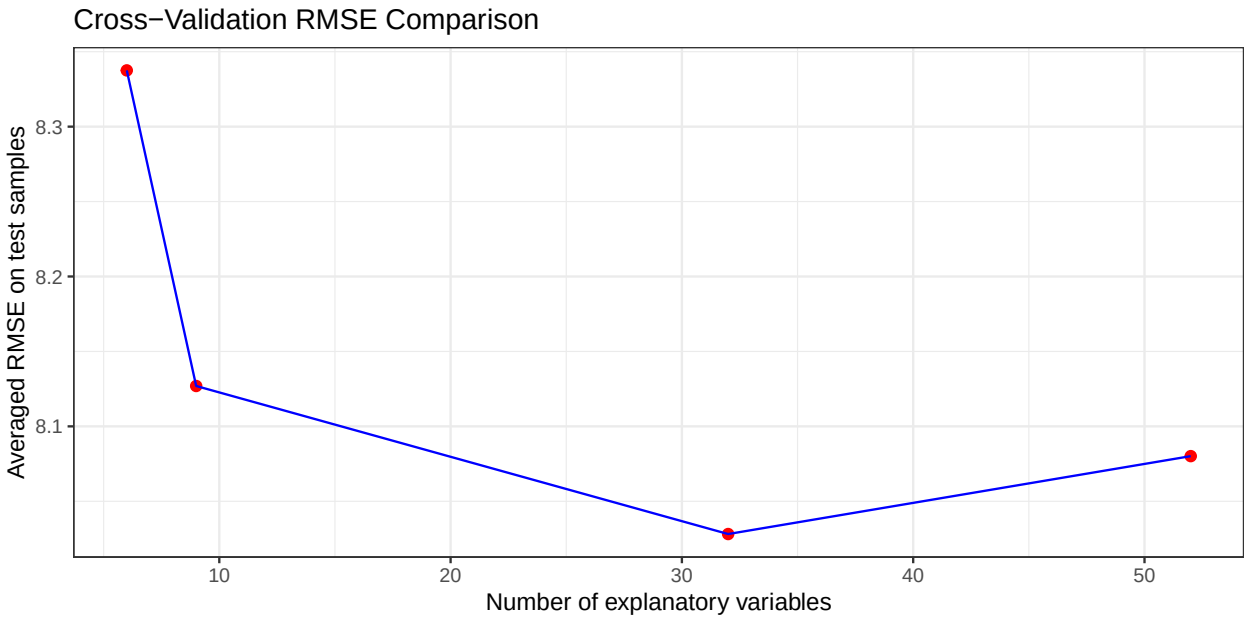
Model	BIC	RMSE	No. of coeff
Model 1	18,268.7	8.3133	5
Model 2	18,150.1	8.0871	8
Model 3	18,206.6	7.8945	31
Model 4	18,335.2	7.8509	51

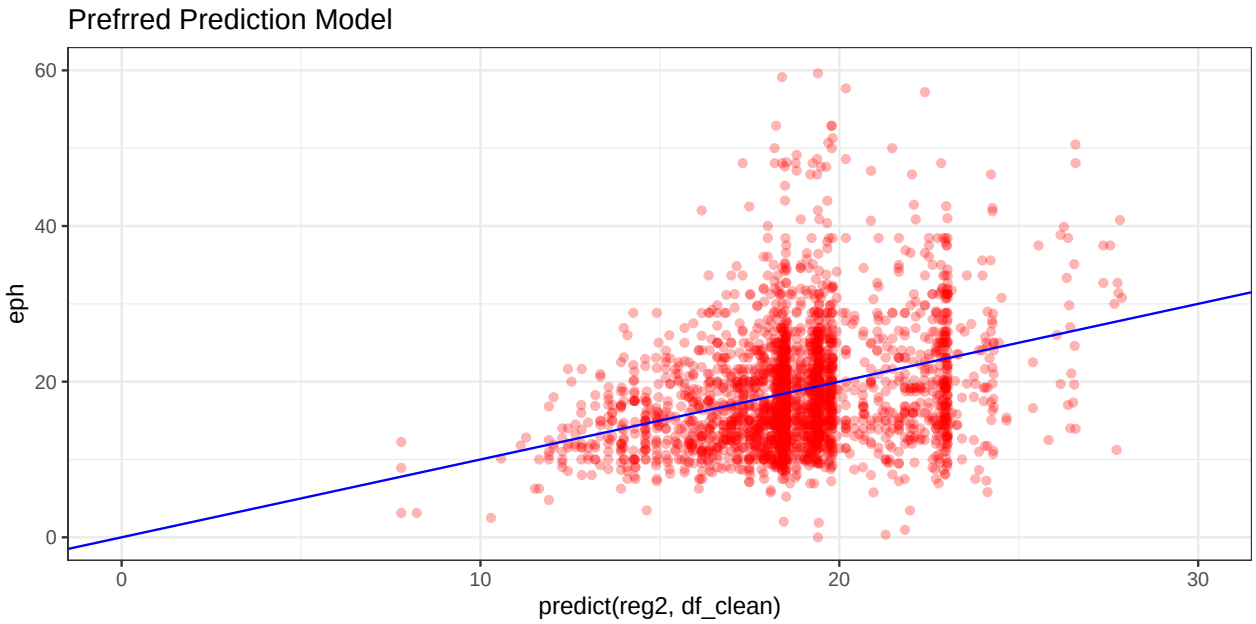
Table 2: Cross Validation Matrix

Resample	Model 1	Model 2	Model 3	Model 4
Fold1	7.685258	7.599731	7.503885	7.608538
Fold2	8.812244	8.581648	8.475649	8.535908
Fold3	8.864477	8.514606	8.306304	8.333844
Fold4	7.922091	7.764600	7.788760	7.807030
Average	8.337525	8.126933	8.028144	8.080134

Table 3: Model Complexity

model	complexity	RMSE
reg1	6	8.337525
reg2	9	8.126933
reg3	32	8.028144
reg4	52	8.080134




Prefrred Prediction Model
60
40
20
0
eph
0
10
20
30
predict(reg2, df_clean)

Regression Table

[H]

Table 4:

	reg1	reg2	reg3	reg4
Dependent Var.:	eph	eph	eph	eph
(Intercept)	18.61*** (0.3621)	2.158 (1.970)	5.377* (2.571)	5.229. (2.782)
eduBachelorsDegree	2.792*** (0.5633)	3.202*** (0.5551)	2.962*** (0.5556)	4.306* (1.788)
eduCollegewithoutDegree	-0.5336 (0.4631)	-0.3732 (0.4506)	-0.4553 (0.4497)	-1.242 (1.463)
eduDiploma	-0.9797* (0.4450)	-1.290** (0.4383)	-1.419** (0.4349)	-3.076* (1.360)
eduMastersDegree	5.014*** (1.421)	4.494** (1.399)	2.718. (1.451)	0.4595 (2.280)
eduNoDiploma	-4.415*** (1.232)	-3.360** (1.096)	-3.879*** (1.156)	-4.859 (3.542)
age		0.6224*** (0.0956)	0.5984*** (0.1123)	0.6020*** (0.1144)
age2		-0.0055*** (0.0011)	-0.0051*** (0.0013)	-0.0051*** (0.0013)
gendermale		3.567*** (0.9160)	7.279 (4.562)	2.789 (5.200)
classGovernment-Local			-3.634** (1.150)	-3.382** (1.200)
classGovernment-State			-3.944*** (1.181)	-3.299** (1.225)
classPrivate,ForProfit			-1.941. (1.064)	-1.662 (1.096)
classPrivate,Nonprofit			-3.921*** (1.118)	-3.989*** (1.147)
marital_statusNeverMarried			0.2959 (0.4950)	-0.3091 (1.086)
marital_statusSingleParent			-0.0451 (0.4018)	1.502 (1.531)
race_dummywhite			-1.803*** (0.5296)	-1.986 (1.239)
ownchild			0.3308 (0.4229)	0.3295 (0.4305)
unionmmeYes			1.236. (0.6573)	2.404 (3.512)
regionNorthEast			1.975*** (0.4687)	2.086*** (0.5029)
regionSouth			0.6200 (0.4150)	0.4969 (0.4229)
regionWest			1.277** (0.4119)	1.362** (0.4370)
eduBachelorsDegree x gendermale			5.155* (2.427)	10.21. (5.548)
eduCollegewithoutDegree x gendermale			2.728 (2.666)	6.384 (4.383)
eduDiploma x gendermale			3.766 (2.531)	10.65. (6.218)
eduMastersDegree x gendermale			10.80** (4.120)	9.186* (4.510)
eduNoDiploma x gendermale			1.557 (3.026)	0.8303 (3.100)
gendermale x classGovernment-Local			-20.20*** (4.355)	-20.27*** (4.447)
gendermale x classGovernment-State			-3.548 (4.474)	-2.377 (4.432)
gendermale x classPrivate,ForProfit			-6.334 (4.056)	-6.236 (4.102)
gendermale x classPrivate,Nonprofit			-3.485 (4.891)	-4.324 (4.916)
gendermale x marital_statusNeverMarried			-5.808** (2.021)	-5.375** (2.038)
gendermale x marital_statusSingleParent			-4.701. (2.467)	-4.485. (2.574)
gendermale x race_dummywhite				4.512 (3.849)
gendermale x unionmmeYes				-1.295 (3.207)
gendermale x ownchild				1.059 (2.218)
race_dummywhite x marital_statusNeverMarried				0.7914 (1.177)
race_dummywhite x marital_statusSingleParent				-1.831 (1.574)
unionmmeYes x classGovernment-Local				-0.7617 (3.614)
unionmmeYes x classGovernment-State				-3.001 (3.594)
unionmmeYes x classPrivate,ForProfit				-1.529 (3.742)
unionmmeYes x classPrivate,Nonprofit				3.345 (4.014)
unionmmeYes x regionNorthEast				-1.020 (1.408)
unionmmeYes x regionSouth				2.022 (2.415)
unionmmeYes x regionWest				-0.9953 (1.308)
eduBachelorsDegree x race_dummywhite				-1.685 (1.872)
eduCollegewithoutDegree x race_dummywhite				0.8377 (1.537)
eduDiploma x race_dummywhite				1.874 (1.428)
eduMastersDegree x race_dummywhite				3.069 (2.927)
eduNoDiploma x race_dummywhite				1.053 (3.759)
eduBachelorsDegree x gendermale x race_dummywhite				-5.706 (6.060)
eduCollegewithoutDegree x gendermale x race_dummywhite				-4.130 (5.317)
eduDiploma x gendermale x race_dummywhite				-7.886 (6.787)
S.E. type	Heteroskeda.-rob.	Heteroskedast.-rob.	Heteroskedast.-rob.	Heteroskedast.-rob.
AIC	18,233.6	18,097.4	18,019.3	18,030.8
BIC	18,268.7	18,150.1	18,206.6	18,335.2
RMSE	8.3133	8.0871	7.8945	7.8509
R2	0.04282	0.09420	0.13683	0.14633
Observations	2,576	2,576	2,576	2,576
No. Variables	5	8	31	51