

DA3 Exercise1

Peter Kaiser

2022 01 23

```
## -- Attaching packages ----- tidyverse 1.3.1 --

## v ggplot2 3.3.5    v purrr  0.3.4
## v tibble  3.1.5    v dplyr  1.0.7
## v tidyr   1.1.4    v stringr 1.4.0
## v readr   2.0.2    v forcats 0.5.1

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##   lift

## New names:
## * ' ' -> ...1

## Rows: 149316 Columns: 23

## -- Column specification -----
## Delimiter: ","
## chr  (9): intmonth, stfips, prcitshp, state, ind02, class, unionmme, unionco...
## dbl (14): ...1, hhid, weight, earnwke, uhours, grade92, race, ethnic, age, s...

##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

- Target Variable (Y): Hourly wage
- Predictor variables (X): Education (High school graduate as base), Race (White as base), Age, Sex (Male as base), Marital Status (Married as base), Union Status (Not in union as base)

```

## Linear Regression
##
## 2980 samples
##    2 predictor
##
## No pre-processing
## Resampling: Cross-Validated (4 fold)
## Summary of sample sizes: 2235, 2235, 2236, 2234
## Resampling results:
##
##    RMSE      Rsquared    MAE
##    8.622476  0.01786559  6.10027
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

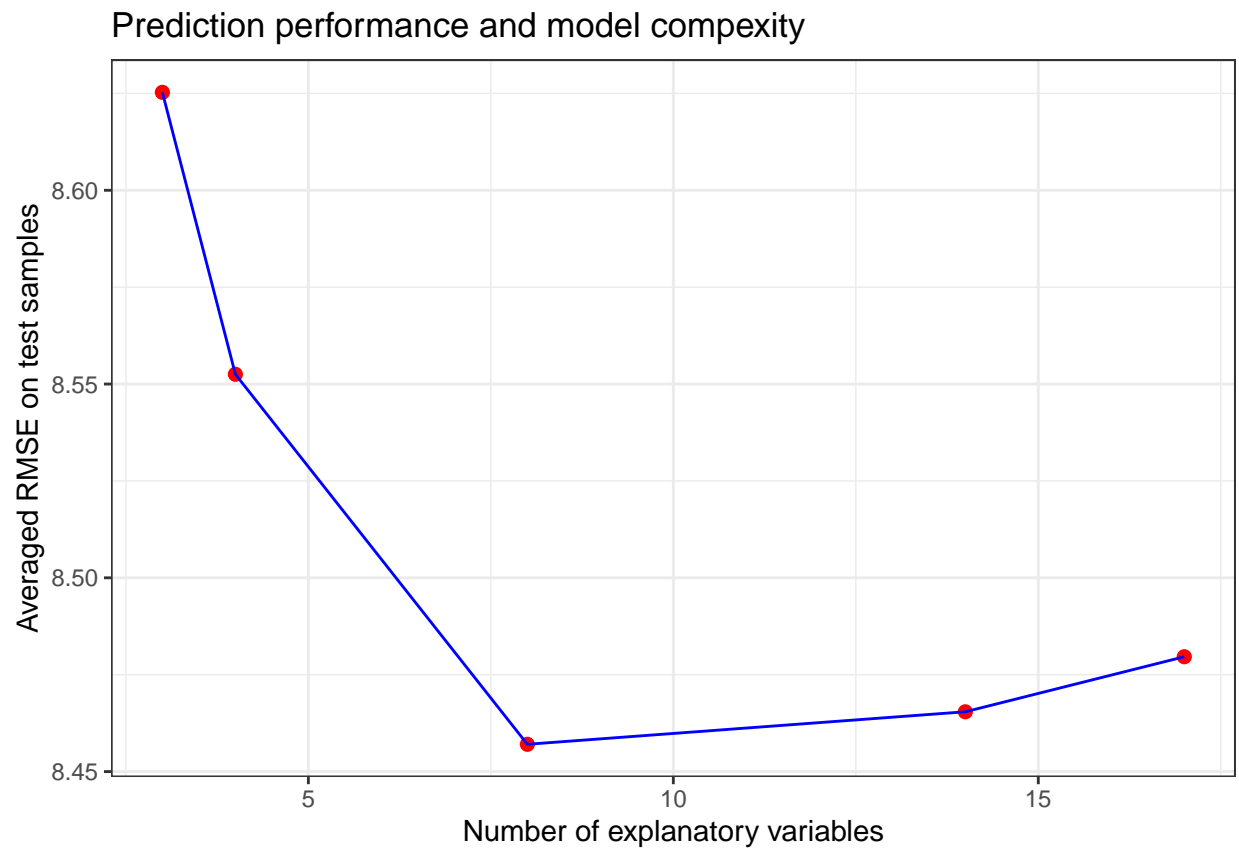
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.358  -5.487  -1.565   3.620  55.254
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.915498   2.599660   0.737   0.461
## age          0.687602   0.121813   5.645 1.81e-08 ***
## agesq       -0.006892   0.001368  -5.038 4.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.619 on 2977 degrees of freedom
## Multiple R-squared:  0.01795,    Adjusted R-squared:  0.01729
## F-statistic: 27.2 on 2 and 2977 DF,  p-value: 1.964e-12

##    intercept      RMSE    Rsquared      MAE    RMSESD    RsquaredSD    MAESD
## 1      TRUE 8.622476 0.01786559 6.10027 0.2538065 0.009511182 0.1596935

##          RMSE      Rsquared      MAE Resample
## 1 8.590490 0.017126070 6.237623   Fold1
## 2 8.926210 0.031429311 6.215915   Fold2
## 3 8.309046 0.009930978 5.895469   Fold3
## 4 8.664158 0.012975984 6.052072   Fold4

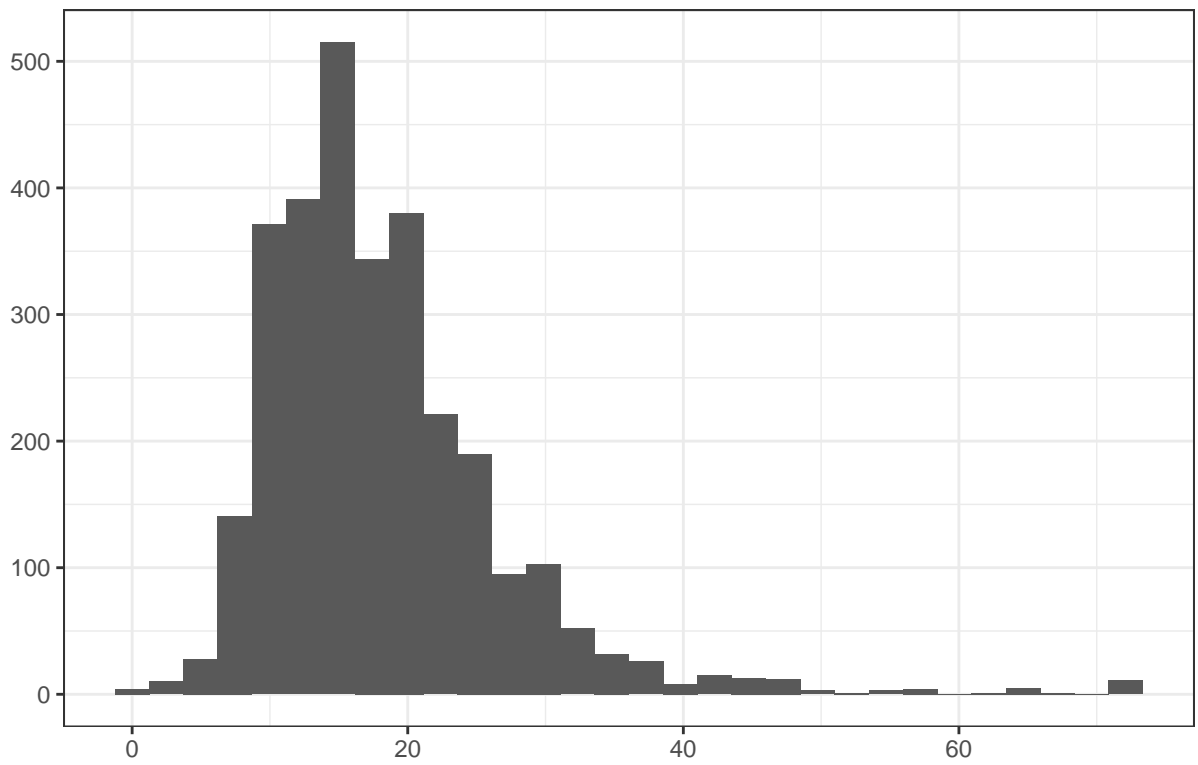
## Resample  Model1  Model2  Model3  Model4  Model5
## 1   Fold1 8.590490 8.474870 8.338192 8.293918 8.297041
## 2   Fold2 8.926210 8.814151 8.682432 8.744360 8.745257
## 3   Fold3 8.309046 8.223868 8.162299 8.162094 8.186240
## 4   Fold4 8.664158 8.685491 8.634375 8.647591 8.676497
## 5 Average 8.625277 8.552521 8.457033 8.465423 8.479631

```

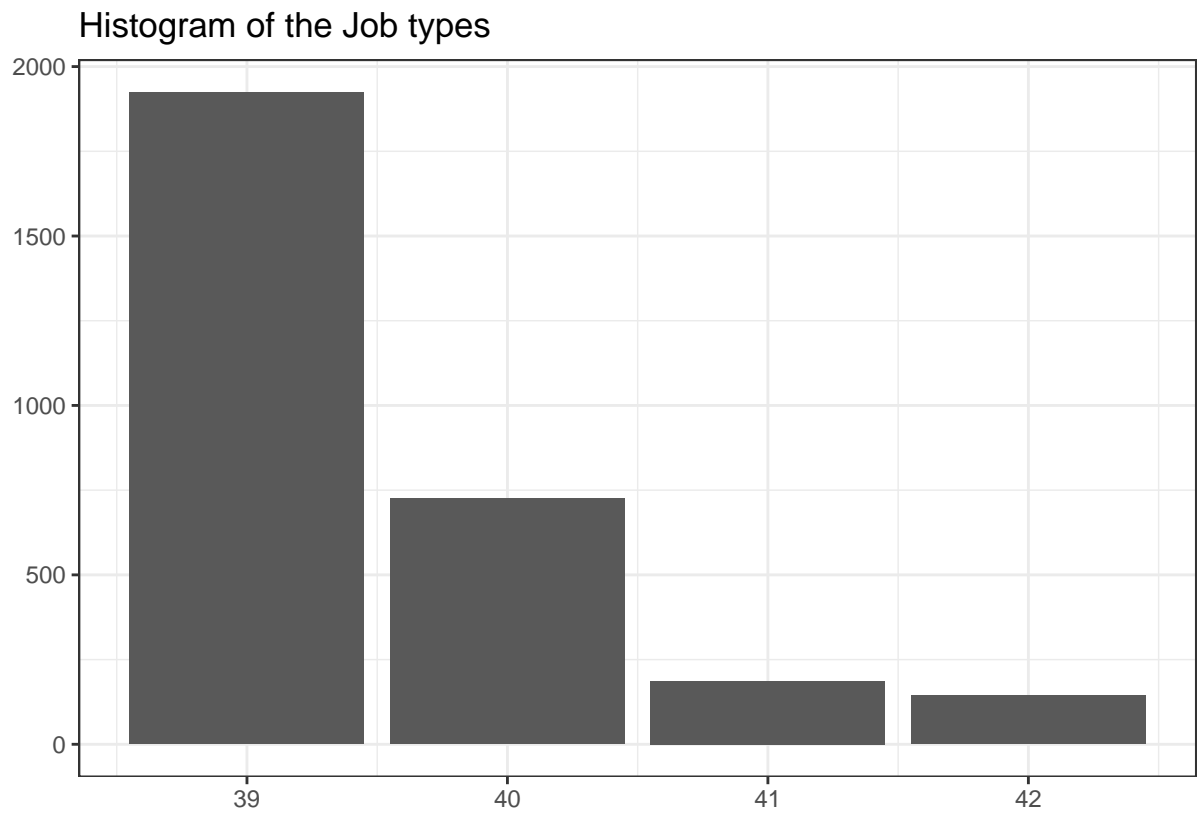


```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

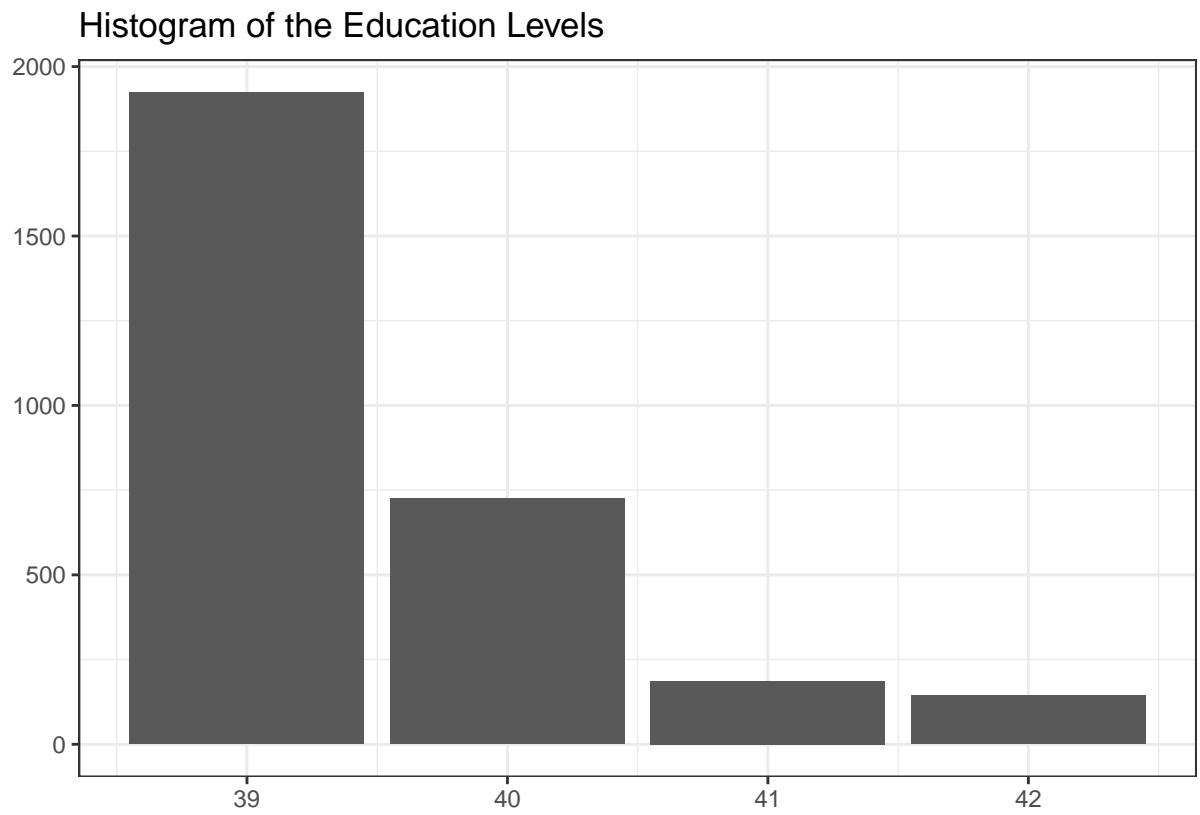
Histogram of Hourly Wage



Warning: Ignoring unknown parameters: binwidth, bins, pad

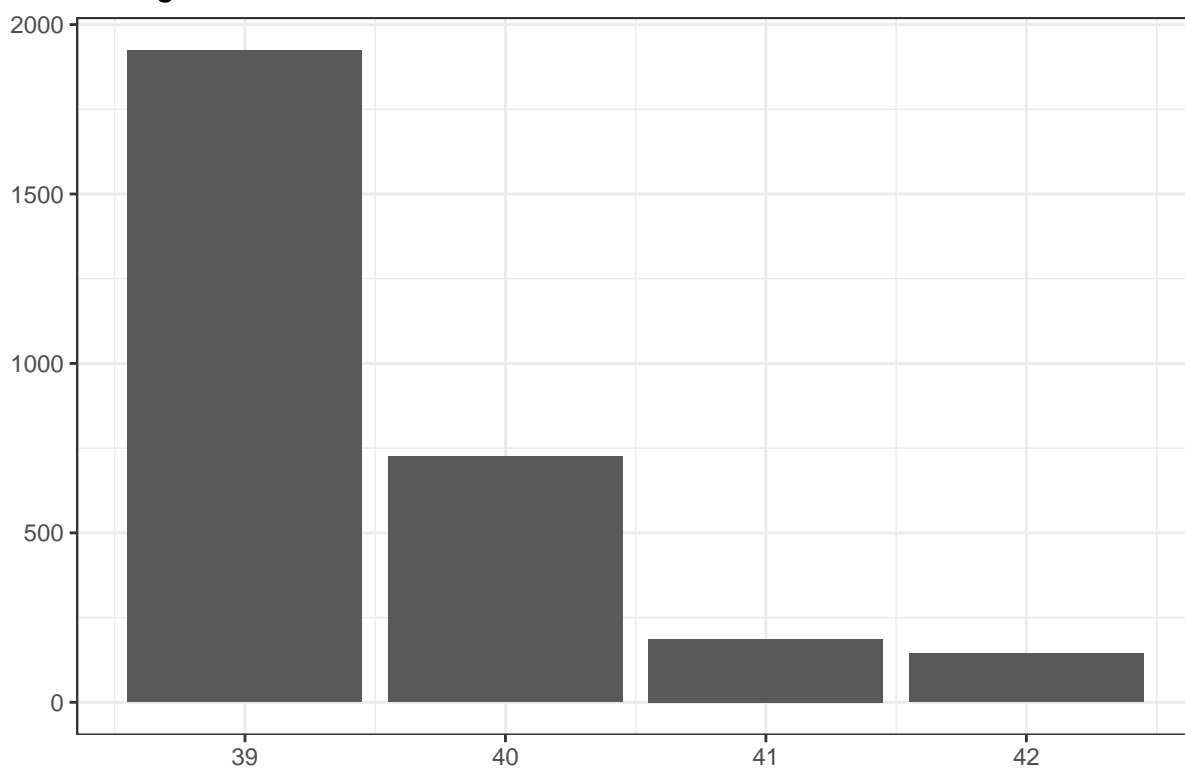


```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

Histogram of the Education Levels



Evaluation of the models using all the sample

	reg1	reg2	reg3	reg4	reg5
Dependent Var.:	earnho	earnho	earnho	earnho	earnho
(Intercept)	1.915 (2.392)	5.315* (2.384)	3.992 (2.504)	3.407 (2.563)	-2.449 (9.631)
age	0.6876*** (0.1157)	0.7182*** (0.1149)	0.6146*** (0.1172)	0.6460*** (0.1187)	1.095 (0.7073)
agesq	-0.0069*** (0.0013)	-0.0072*** (0.0013)	-0.0063*** (0.0013)	-0.0067*** (0.0013)	-0.0173 (0.0167)
sex		-3.639*** (0.4378)			
s_femalef			-4.000*** (0.4510)	-4.073*** (0.4575)	-4.134* (1.917)
unionf			3.532*** (0.4268)	3.543*** (0.4212)	0.4839 (1.890)
m_divorcedf			0.4361 (0.4939)	0.3766 (0.4937)	0.3872 (0.4928)
m_nevermarriedf			-0.9001* (0.4349)	-0.7635. (0.4357)	-0.7563. (0.4352)
m_otherf			-0.1950 (0.5845)	-0.1103 (0.5855)	-0.1091 (0.5852)
r_blackf				-0.8400* (0.3966)	-0.8445* (0.3972)
r_asianf				-1.219 (1.022)	-1.175 (1.023)
r_otherf				-1.991* (0.8207)	-2.073* (0.8215)
e_collegedrop				0.5206 (0.3849)	0.5151 (0.3854)
e_occupationdegree				-1.838*** (0.5385)	-1.849*** (0.5381)
e_academicdegree				2.000* (0.8355)	2.019* (0.8368)
agecu					7.87e-5 (0.0001)
age x					0.0012
s_femalef					(0.0428)
age x unionf					0.0645 (0.0396)
S.E. type	Heteroskedast.- rob.	Heteroskedast.- rob.	Heteroskedast.- rob.	Heteroskedast.- rob.	Heteroskedas.- rob.
AIC	21,297.8	21,243.4	21,173.7	21,157.6	21,160.7
BIC	21,315.8	21,267.4	21,221.7	21,241.6	21,262.7
RMSE	8.6150	8.5339	8.4234	8.3837	8.3797
R2	0.01795	0.03636	0.06115	0.06997	0.07085
Observations	2,980	2,980	2,980	2,980	2,980
No. Variables	2	3	7	13	16