STA108
Final Project

Team 1

Grant, Jinghong, Natalie, Selam

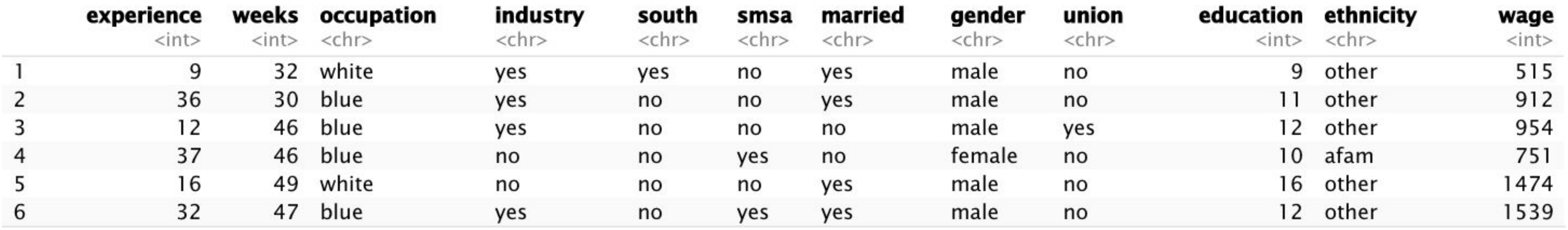
Instructor: Prof. Hao Chen

Fall 2020

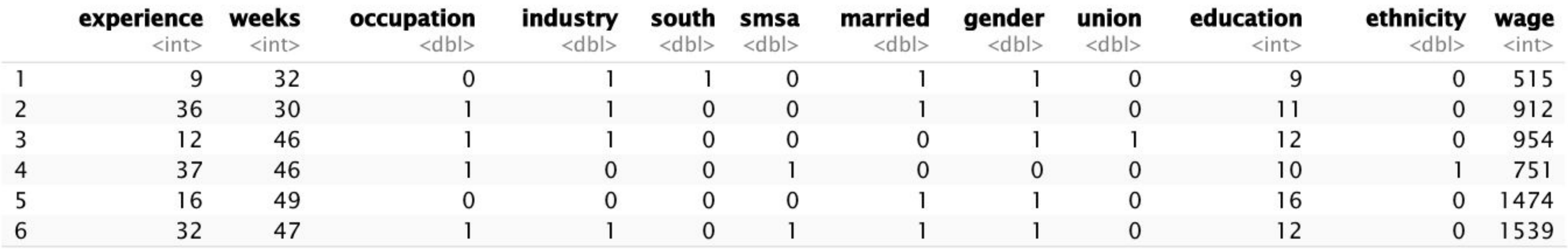
Overview on Earnings data

1. Data preparation
2. Data exploration
3. Data transformation
4. Significance testing
5. Model selection(s)
6. Model validation(s)

Data preparation (Earnings Data)

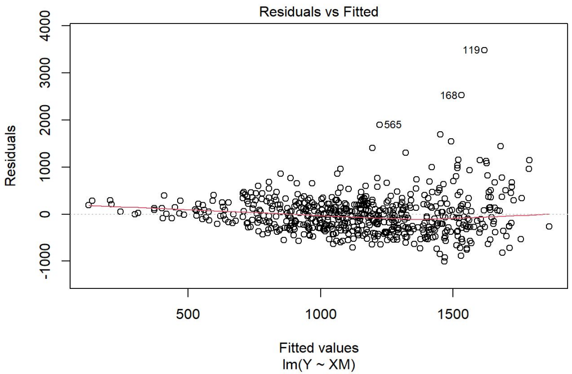


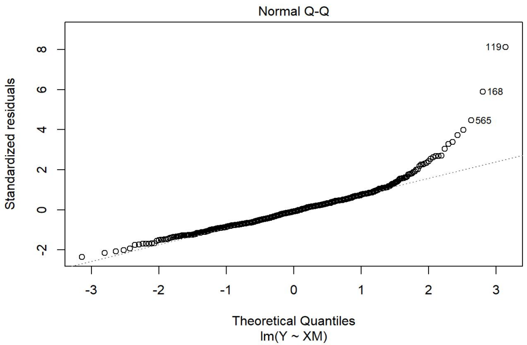
	experience	weeks	occupation	industry	south	smsa	married	gender	union	education	ethnicity	wage
	<int>	<int>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<int>	<chr>	<int>
1	9	32	white	yes	yes	no	yes	male	no	9	other	515
2	36	30	blue	yes	no	no	yes	male	no	11	other	912
3	12	46	blue	yes	no	no	no	male	yes	12	other	954
4	37	46	blue	no	no	yes	no	female	no	10	afam	751
5	16	49	white	no	no	no	yes	male	no	16	other	1474
6	32	47	blue	yes	no	yes	yes	male	no	12	other	1539

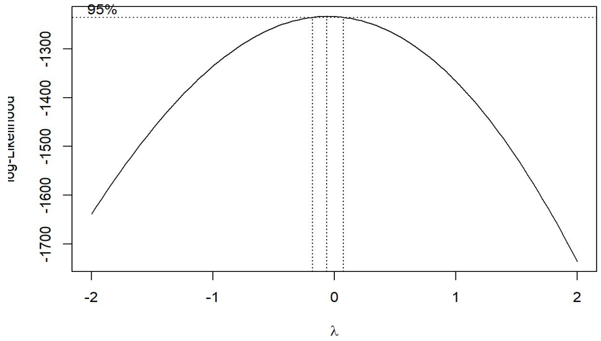


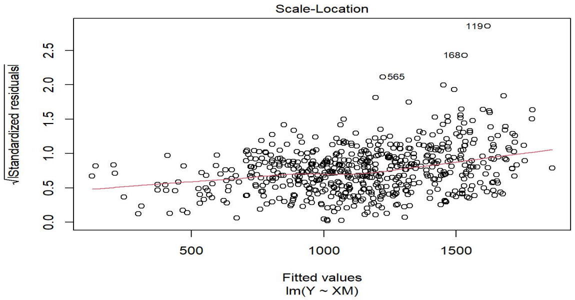
	experience	weeks	occupation	industry	south	smsa	married	gender	union	education	ethnicity	wage
	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<int>
1	9	32	0	1	1	0	1	1	0	9	0	515
2	36	30	1	1	0	0	1	1	0	11	0	912
3	12	46	1	1	0	0	0	1	1	12	0	954
4	37	46	1	0	0	1	0	0	0	10	1	751
5	16	49	0	0	0	0	1	1	0	16	0	1474
6	32	47	1	1	0	1	1	1	0	12	0	1539

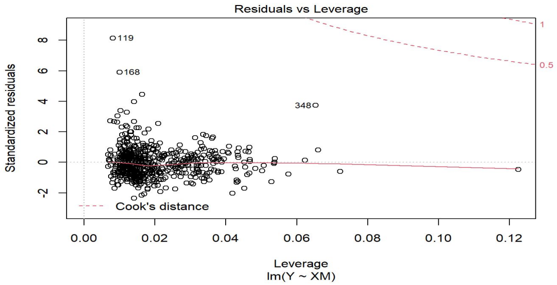
Data exploration


Residuals vs Fitted


Normal Q-Q

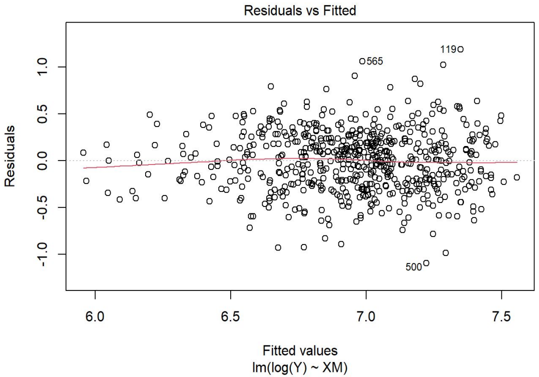


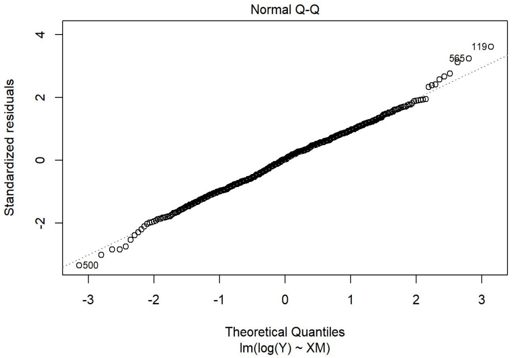

Scale-Location

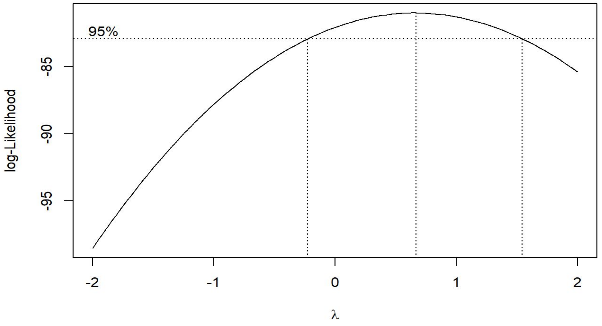

Residuals vs Leverage

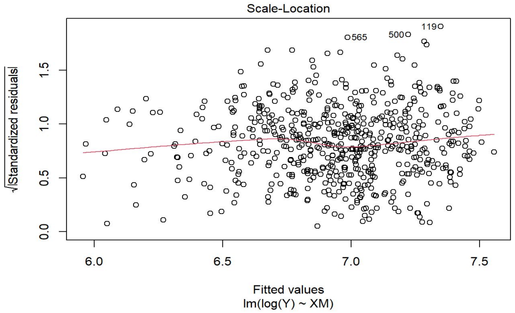
$R^2 = 0.35751$

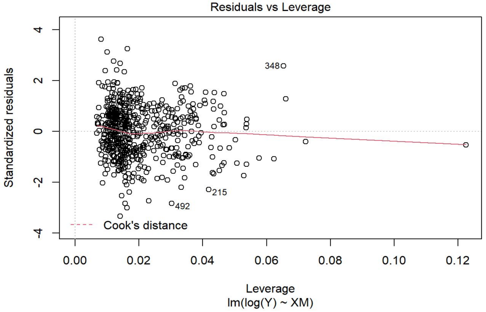
Data transformation











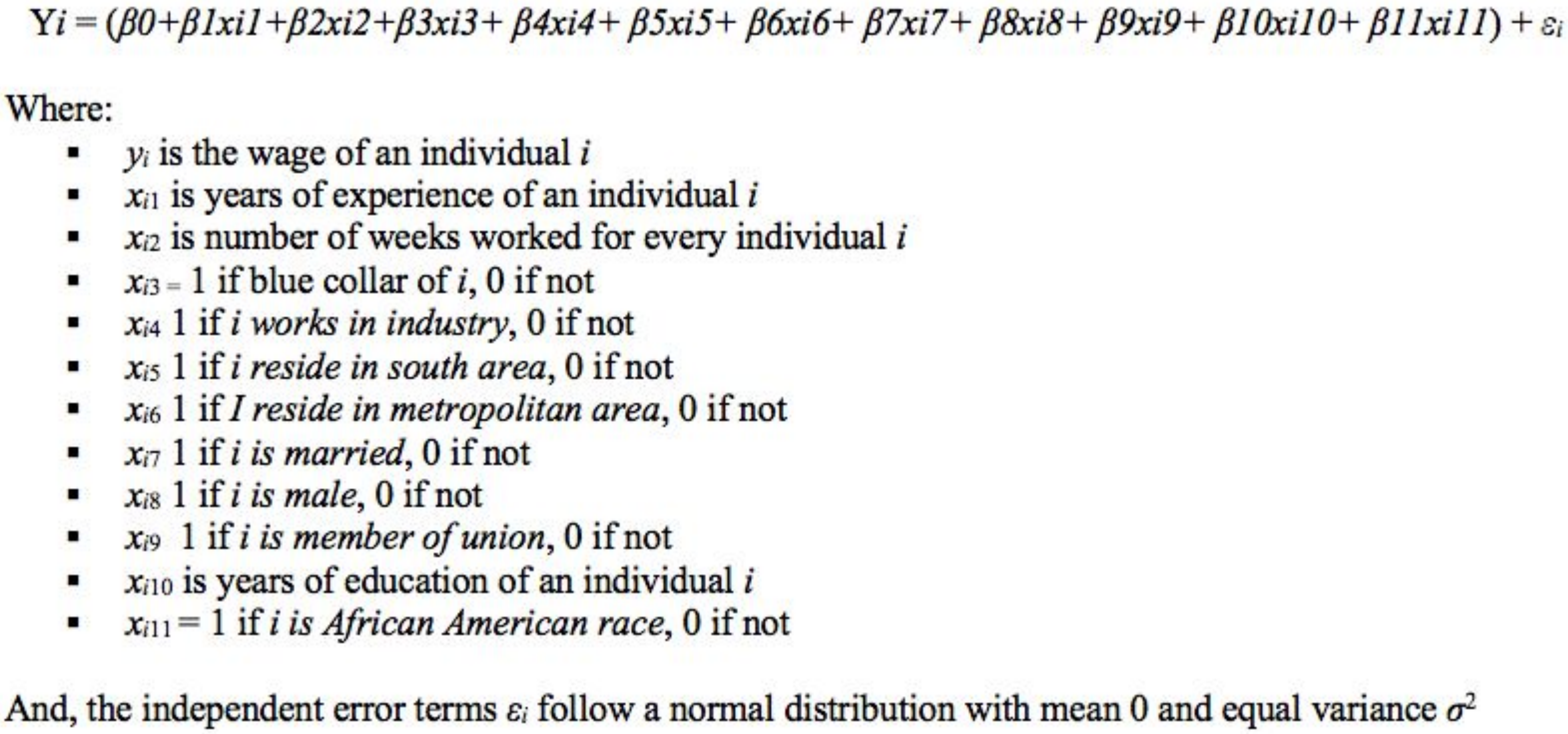
$$R^2 = 0.446054$$

Data splitting

```
set.seed(123)
index = sample(1:595,298)
training_data = mydata[index,]  #the first half
validate_data = mydata[-index,] #the second half
```

- Randomly splitting data into two portions to allow for independent model building and model validation
- Random sample removes potential bias in ordering of entries

Significance Testing



$$Yi = (\beta 0+\beta 1xi1+\beta 2xi2+\beta 3xi3+ \beta 4xi4+ \beta 5xi5+ \beta 6xi6+ \beta 7xi7+ \beta 8xi8+ \beta 9xi9+ \beta 10xi10+ \beta 11xi11) + \varepsilon_i$$

Where:

- y_i is the wage of an individual *i*
- x_{i1} is years of experience of an individual *i*
- x_{i2} is number of weeks worked for every individual *i*
- $x_{i3} =$ 1 if blue collar of *i*, 0 if not
- x_{i4} 1 if *i works in industry*, 0 if not
- x_{i5} 1 if *i reside in south area*, 0 if not
- x_{i6} 1 if *I reside in metropolitan area*, 0 if not
- x_{i7} 1 if *i is married*, 0 if not
- x_{i8} 1 if *i is male*, 0 if not
- x_{i9} 1 if *i is member of union*, 0 if not
- x_{i10} is years of education of an individual *i*
- $x_{i11} =$ 1 if *i is African American race*, 0 if not

And, the independent error terms ε_i follow a normal distribution with mean 0 and equal variance σ^2

Significance Testing (cont) : Hypothesis Testing

- Is the wage of an individual significantly related **one slope parameter** to the working experience of an individual?

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

$$\left\{ F^* = 13.403 \; F_q = 3.857 \right\}$$

Reject null hypothesis, in favor of Full Model

- Is the wage significantly related **subsets parameter** to individual working hours and designation?

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_a\text{: at least } \beta_1 \text{ or } \beta_2 \neq 0$$

$$\left\{ F^* = 3.436 \; F_q = 2.38 \right\}$$

Reject null hypothesis, in favor of Full Model

- Is the regression model containing **at least one predictor** useful in predicting the average wage of an individual?

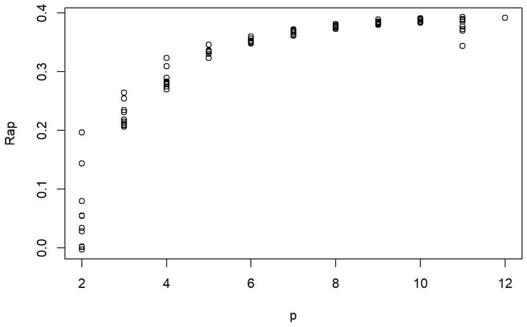
$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$$
$$H_a\text{: at least one } \beta_j \neq 0 \text{ (for } j = 1, 2, 3, 4, 5, 6, 7, 8, 10, 11\text{)}$$

$$\left\{ F^* = 42.68 \; F_q = 1.81 \right\}$$

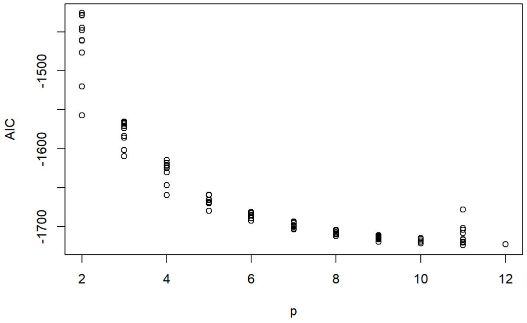
Reject null hypothesis, in favor of Full Model

Model selection **first-order model**

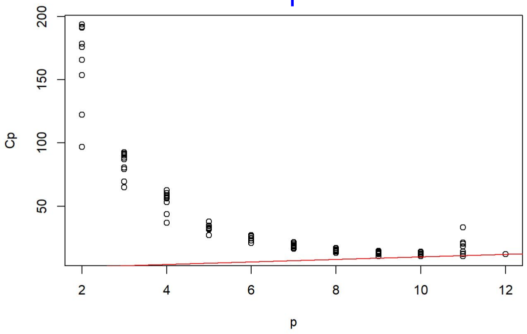
Adjusted R^2



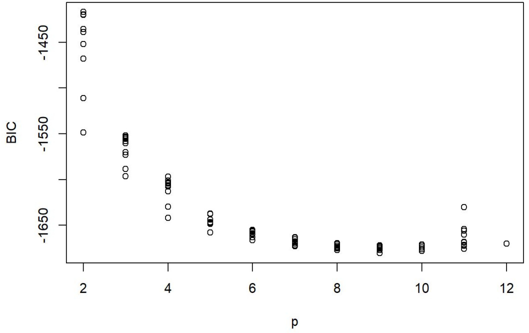
A.I.C



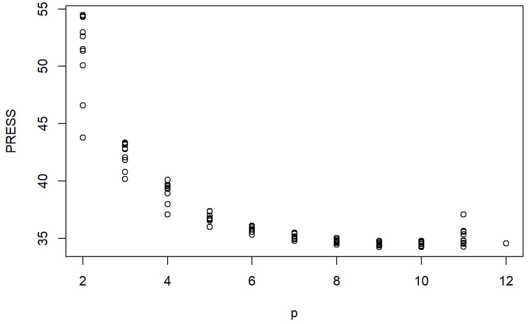
Cp

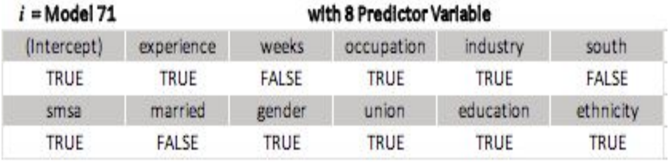


B.I.C



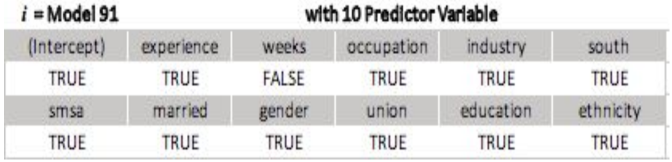
PRESS_p





i = Model 71 with 8 Predictor Variable

(Intercept)	experience	weeks	occupation	industry	south
TRUE	TRUE	FALSE	TRUE	TRUE	FALSE
smsa	married	gender	union	education	ethnicity
TRUE	FALSE	TRUE	TRUE	TRUE	TRUE



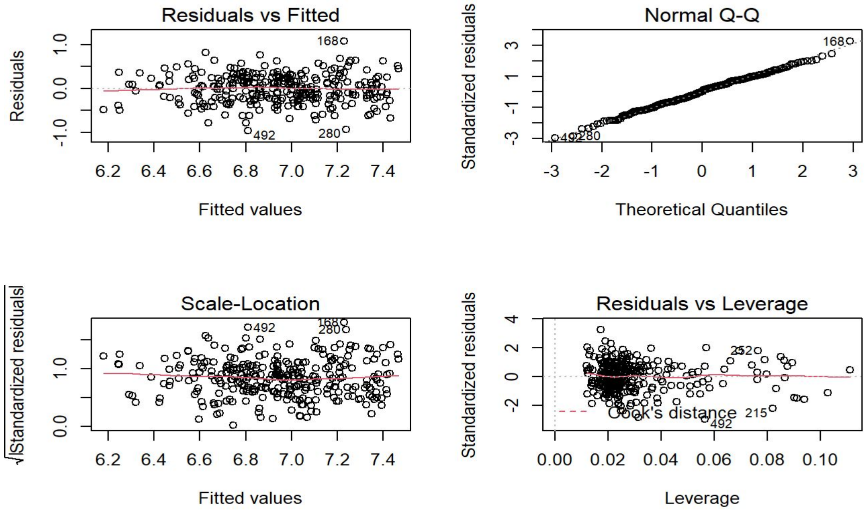
i = Model 91 with 10 Predictor Variable

(Intercept)	experience	weeks	occupation	industry	south
TRUE	TRUE	FALSE	TRUE	TRUE	TRUE
smsa	married	gender	union	education	ethnicity
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

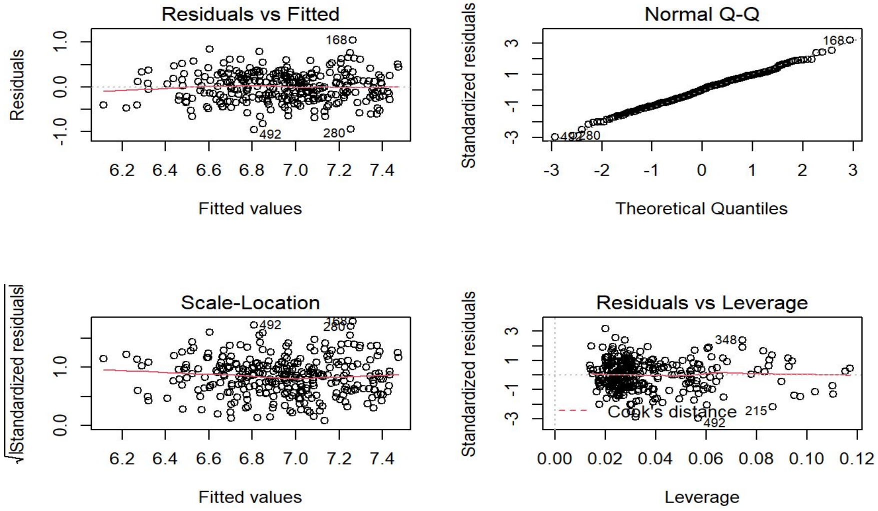
Model diagnostic

first-order model

Model 71



Model 91

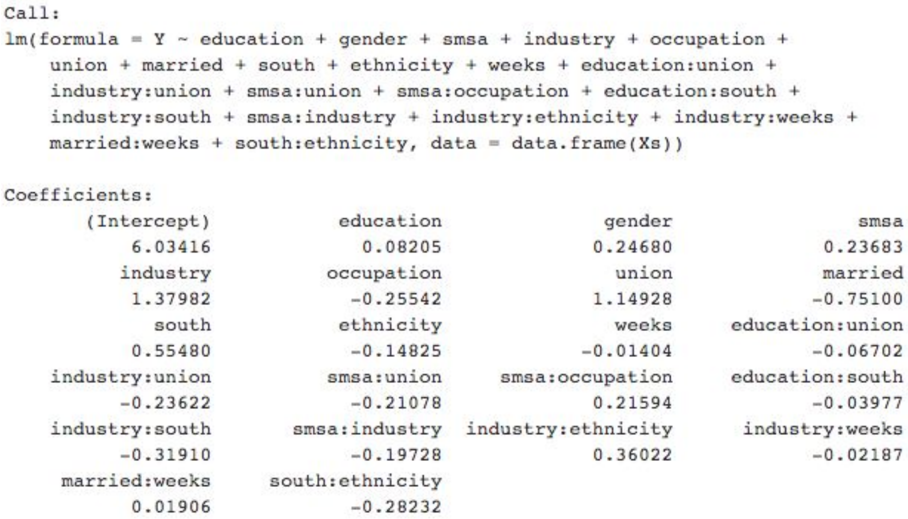


MSE = 0.109
MSPR = 0.106

MSE = 0.111
MSPR = 0.107

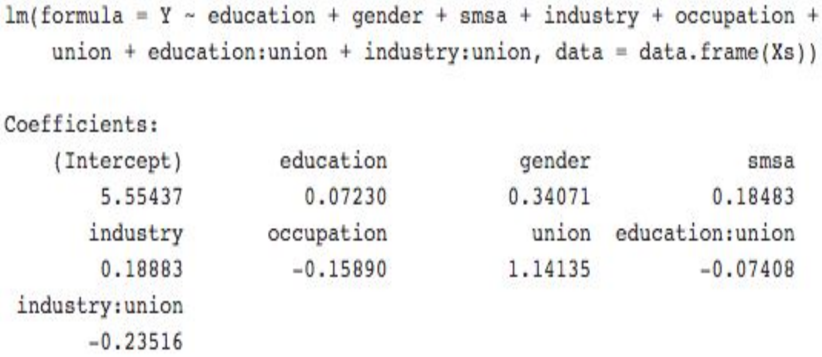
Model selection **first-order with two-way interaction “Stepwise selection”**

AIC criteria



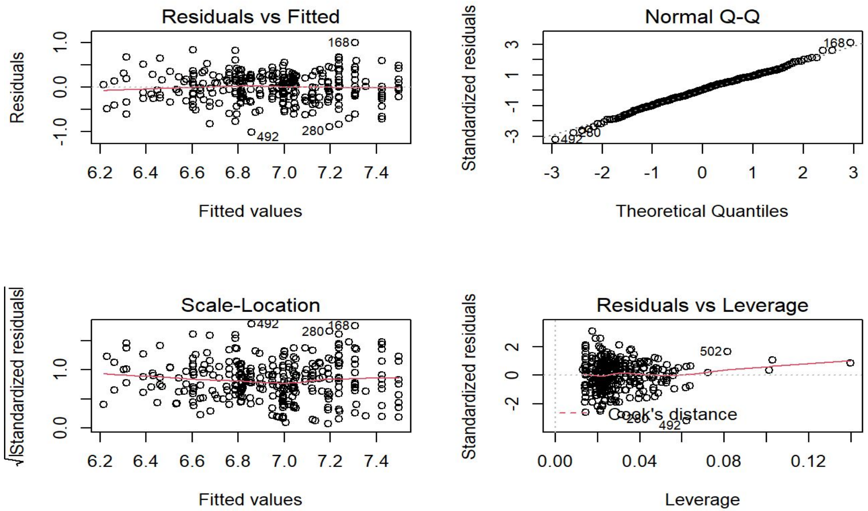
```
Call:
lm(formula = Y ~ education + gender + smsa + industry + occupation +
    union + married + south + ethnicity + weeks + education:union +
    industry:union + smsa:union + smsa:occupation + education:south +
    industry:south + smsa:industry + industry:ethnicity + industry:weeks +
    married:weeks + south:ethnicity, data = data.frame(Xs))

Coefficients:
      (Intercept)            education               gender                 smsa
          6.03416              0.08205              0.24680              0.23683
         industry           occupation                union              married
          1.37982             -0.25542              1.14928             -0.75100
            south            ethnicity                weeks      education:union
          0.55480             -0.14825             -0.01404             -0.06702
   industry:union           smsa:union      smsa:occupation      education:south
         -0.23622             -0.21078              0.21594             -0.03977
   industry:south        smsa:industry   industry:ethnicity       industry:weeks
         -0.31910             -0.19728              0.36022             -0.02187
    married:weeks      south:ethnicity
          0.01906             -0.28232
```

BIC criteria

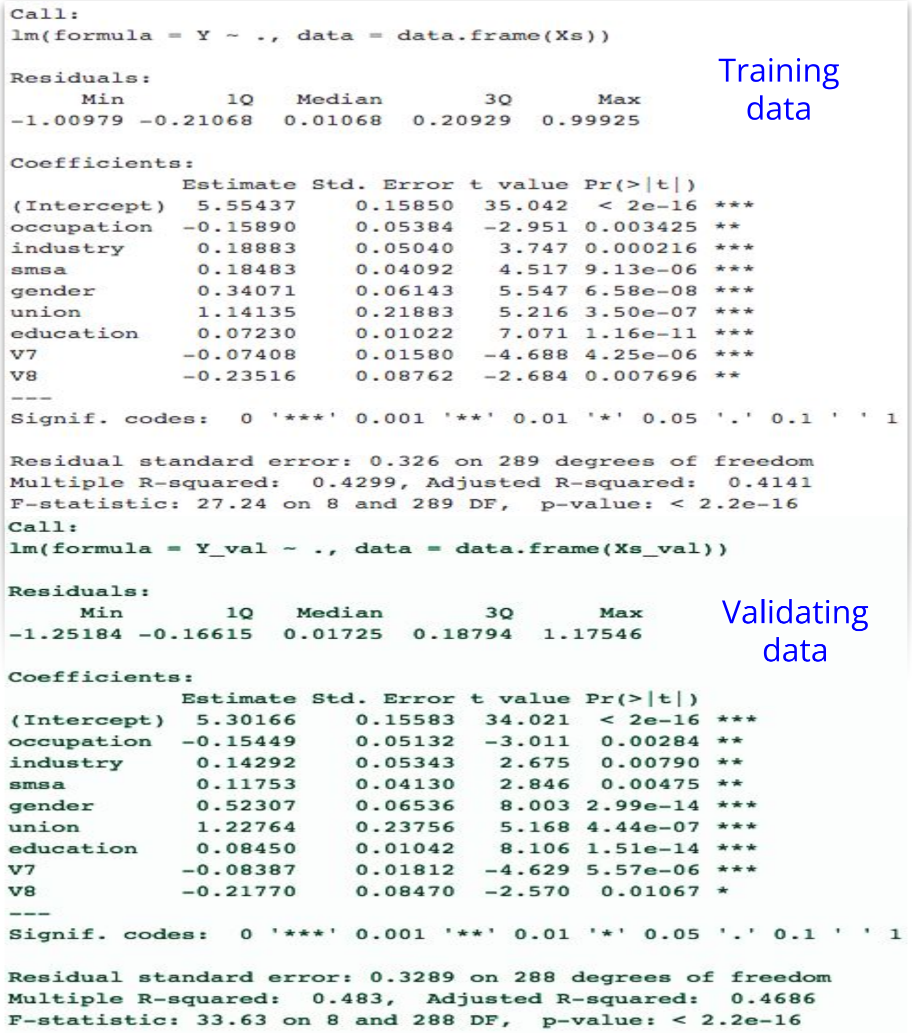


```
lm(formula = Y ~ education + gender + smsa + industry + occupation +
    union + education:union + industry:union, data = data.frame(Xs))

Coefficients:
    (Intercept)        education           gender             smsa
        5.55437          0.07230          0.34071          0.18483
       industry       occupation            union  education:union
        0.18883         -0.15890          1.14135         -0.07408
 industry:union
       -0.23516
```

Model validation **first-order with two-way interaction**



MSE = 0.103
MSPR = 3.202



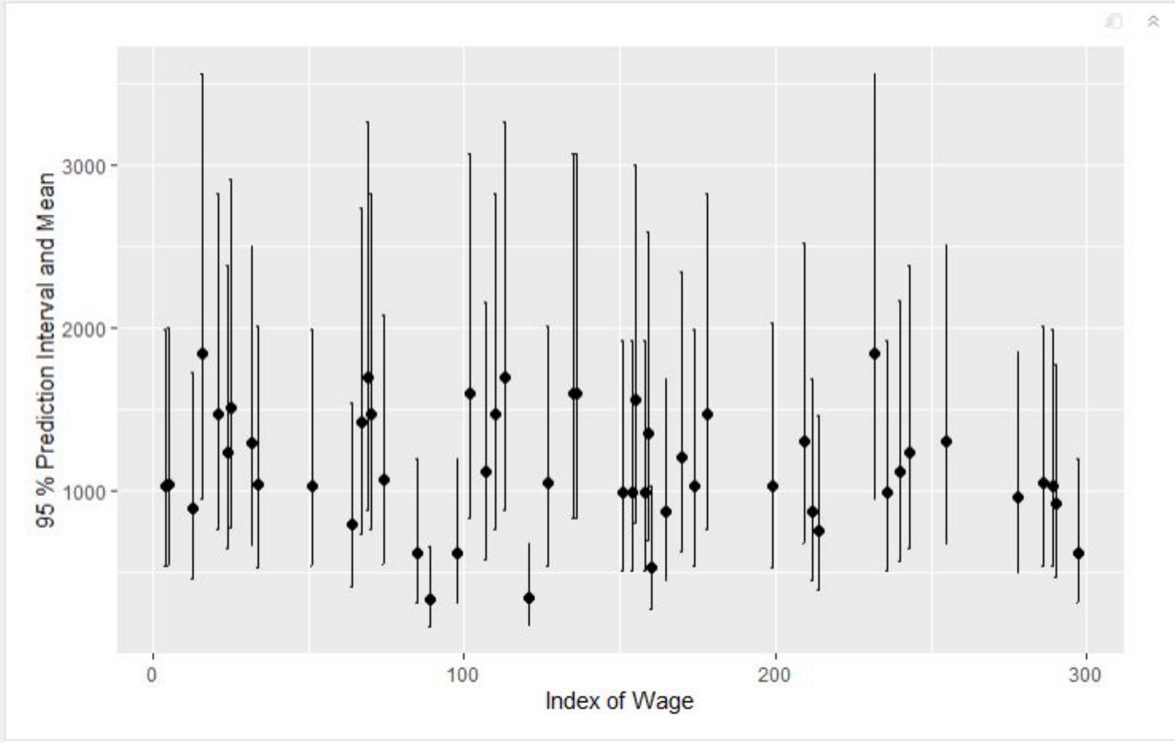
Training data

```
Call:
lm(formula = Y ~ ., data = data.frame(Xs))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.00979 -0.21068  0.01068  0.20929  0.99925 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.55437    0.15850  35.042  < 2e-16 ***
occupation  -0.15890    0.05384  -2.951 0.003425 ** 
industry     0.18883    0.05040   3.747 0.000216 ***
smsa         0.18483    0.04092   4.517 9.13e-06 ***
gender       0.34071    0.06143   5.547 6.58e-08 ***
union        1.14135    0.21883   5.216 3.50e-07 ***
education    0.07230    0.01022   7.071 1.16e-11 ***
V7          -0.07408    0.01580  -4.688 4.25e-06 ***
V8          -0.23516    0.08762  -2.684 0.007696 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.326 on 289 degrees of freedom
Multiple R-squared:  0.4299,	Adjusted R-squared:  0.4141 
F-statistic: 27.24 on 8 and 289 DF,  p-value: < 2.2e-16
Call:
lm(formula = Y_val ~ ., data = data.frame(Xs_val))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.25184 -0.16615  0.01725  0.18794  1.17546 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.30166    0.15583  34.021  < 2e-16 ***
occupation  -0.15449    0.05132  -3.011  0.00284 ** 
industry     0.14292    0.05343   2.675  0.00790 ** 
smsa         0.11753    0.04130   2.846  0.00475 ** 
gender       0.52307    0.06536   8.003 2.99e-14 ***
union        1.22764    0.23756   5.168 4.44e-07 ***
education    0.08450    0.01042   8.106 1.51e-14 ***
V7          -0.08387    0.01812  -4.629 5.57e-06 ***
V8          -0.21770    0.08470  -2.570  0.01067 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3289 on 288 degrees of freedom
Multiple R-squared:  0.483,	Adjusted R-squared:  0.4686 
F-statistic: 33.63 on 8 and 288 DF,  p-value: < 2.2e-16
```

Validating data

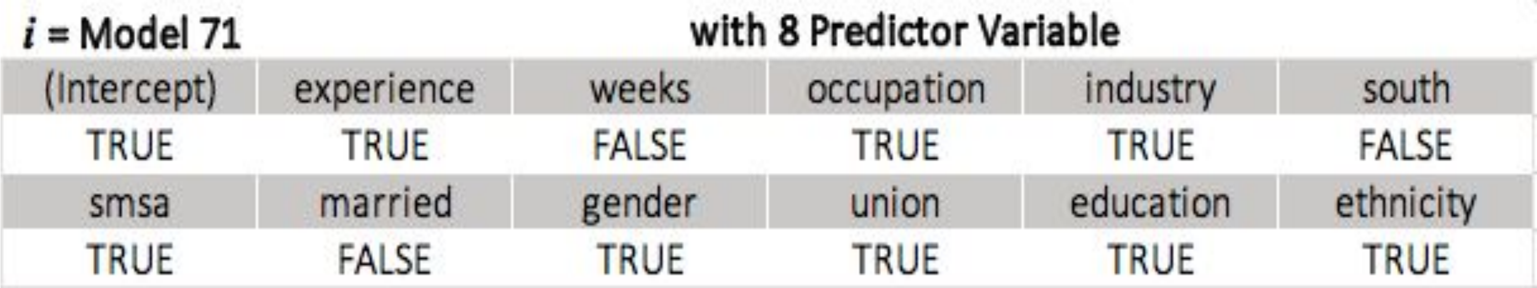
Prediction interval (on validation dataset)

- Prediction intervals much wider than confidence interval.
- Y_{h_new} means greater variance
- Mean of interval is not at the true middle (due to transformation)



Selected Model

We chose the following first-order model with these predictor variables.



i = Model 71 with 8 Predictor Variable

(Intercept)	experience	weeks	occupation	industry	south
TRUE	TRUE	FALSE	TRUE	TRUE	FALSE
smsa	married	gender	union	education	ethnicity
TRUE	FALSE	TRUE	TRUE	TRUE	TRUE

Acknowledgment:

We would like to give special thanks to Prof. Hao Chen, and Mr. Yi-Wei for sharing their knowledge and codes. Also, we would like to thank you all for giving your time to listen to our presentation.