STATS 500 Homework 1

Changbai Liu

1. Chapter 1, problem 2 (page 12)
   The dataset uswages is drawn as a sample from the Current Population Survey in 1988. Make a numerical and graphical summary of the data, commenting on any features that you find interesting.

   Take a look at the data:

```
> uswages[1:10,]
      wage educ exper race smsa ne mw so we pt
6085  771.60   18    18    0    1  1  0  0  0  0
23701 617.28   15    20    0    1  0  0  0  1  0
16208 957.83   16     9    0    1  0  0  1  0  0
2720  617.28   12    24    0    1  1  0  0  0  0
9723  902.18   14    12    0    1  0  1  0  0  0
22239 299.15   12    33    0    1  0  0  0  1  0
14379 541.31   16    42    0    1  0  0  1  0  1
12878 148.39   16     0    0    1  0  1  0  0  1
23121 273.19   12    36    0    1  0  0  0  1  1
13086 666.67   12    37    0    0  0  1  0  0  0
dim(uswages)
[1] 2000   10
```

   Comment:
   There are two thousands of people in the data.

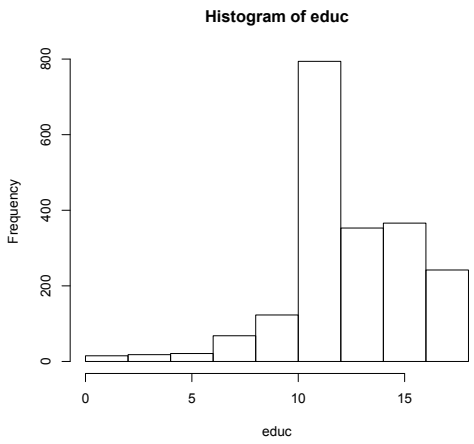
   Summary of data:

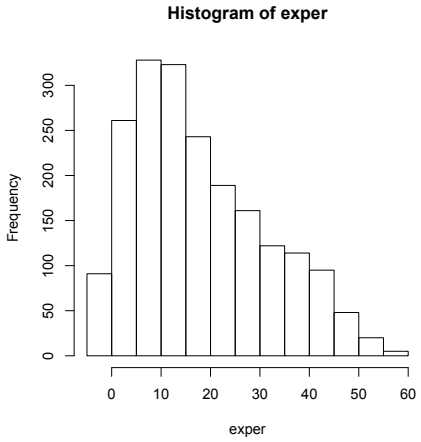
```
> summary(uswages)
     wage             educ           exper            race              smsa
 Min.   :  50.39   Min.   : 0.00   Min.   :-2.00   Min.   :0.000   Min.   :0.000
 1st Qu.: 308.64   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:0.000   1st Qu.:1.000
 Median : 522.32   Median :12.00   Median :15.00   Median :0.000   Median :1.000
 Mean   : 608.12   Mean   :13.11   Mean   :18.41   Mean   :0.078   Mean   :0.756
 3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00   3rd Qu.:0.000   3rd Qu.:1.000
 Max.   :7716.05   Max.   :18.00   Max.   :59.00   Max.   :1.000   Max.   :1.000
       ne              mw              so              we              pt
 Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00   Min.   :0.0000
 1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:0.0000
 Median :0.000   Median :0.0000   Median :0.0000   Median :0.00   Median :0.0000
 Mean   :0.229   Mean   :0.2485   Mean   :0.3125   Mean   :0.21   Mean   :0.0925
 3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.00   3rd Qu.:0.0000
 Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00   Max.   :1.0000
```

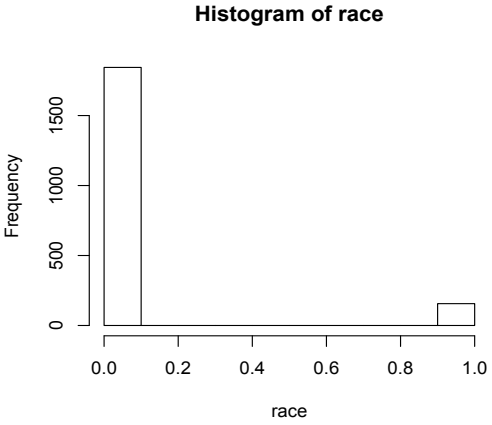
   Comment:
   The wages range is (50.4-7716.1),which is a pretty big range.
   The median and mean number of “educ” are both around 13, which means most of the people in this data do not have a college degree.
   Most of the people in data work more than 15 years and they are mostly back people, also live in Standard Metropolitan Statistical Area.
   Just 9.25% of the people in data take part-time job.

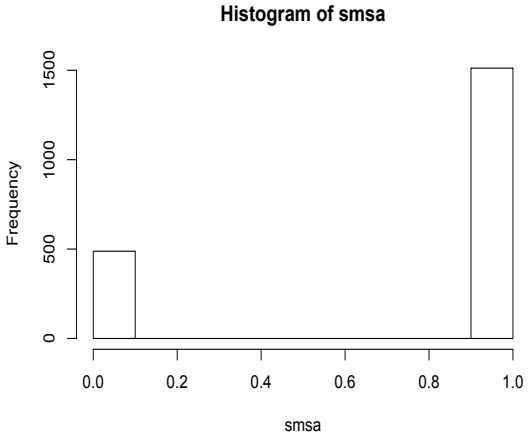
   Question not solved:
   How come a people get -2 years’ experience?

Histograms:

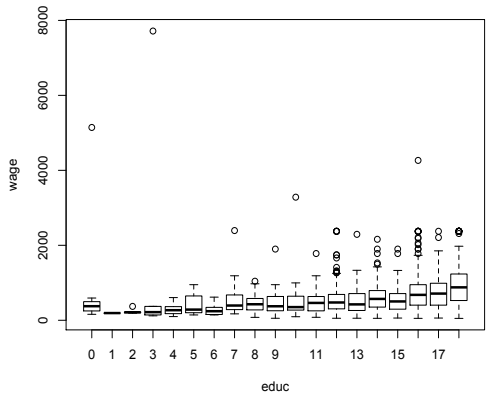

Histogram of educ
Frequency
0
200
400
600
800
0
5
10
15
educ

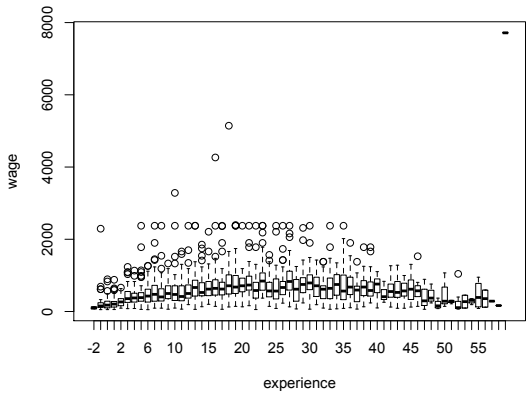

Histogram of exper
Frequency
0
50
100
150
200
250
300
0
10
20
30
40
50
60
exper

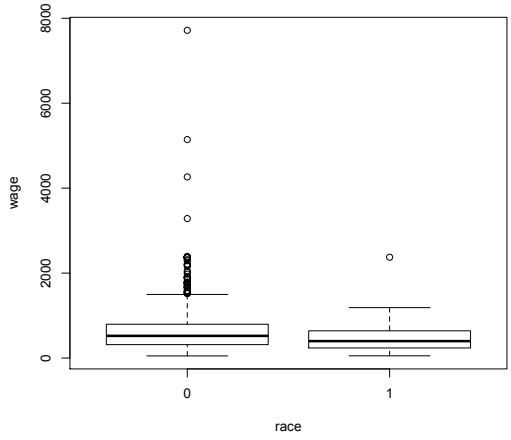

Histogram of race
Frequency
0
500
1000
1500
0.0
0.2
0.4
0.6
0.8
1.0
race

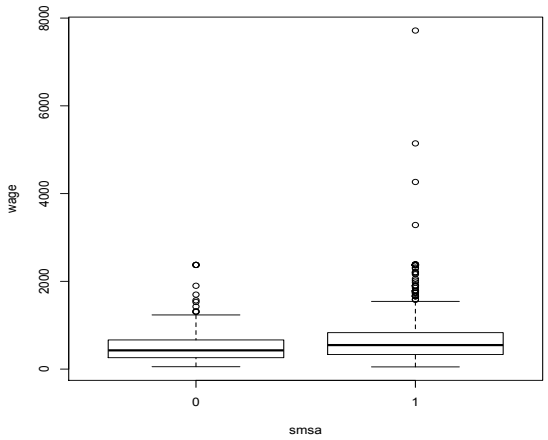

Histogram of smsa
Frequency
0
500
1000
1500
0.0
0.2
0.4
0.6
0.8
1.0
smsa

How factors impact on wage:





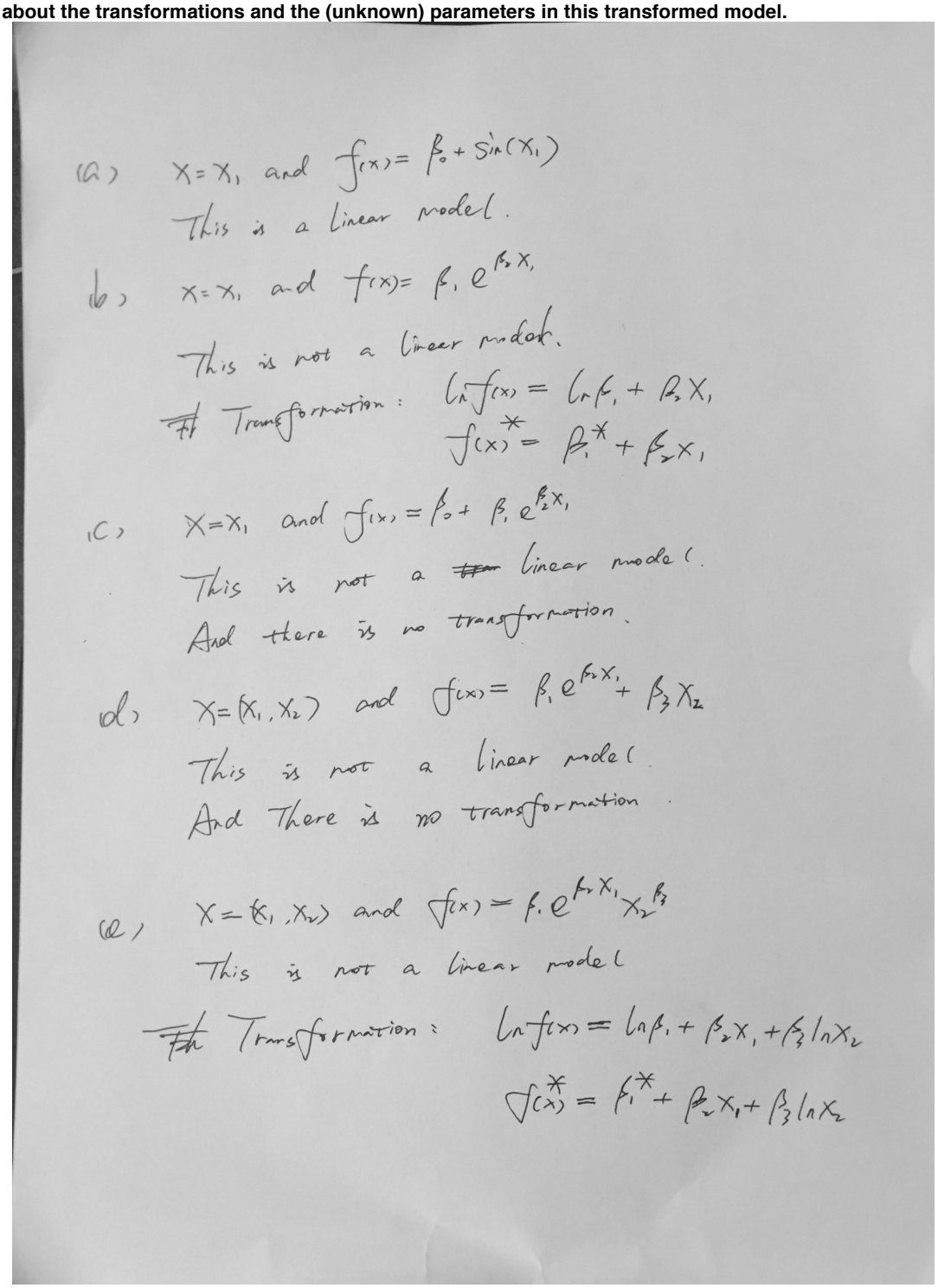




Comment:

1) First diagram shows that the more education people get, the more wage they get.
2) Second diagram shows that in the first 20 years in career, the more experience people have, the more wage they get.
   But it appears exactly opposite when people's experience is more than twenty years, especially when they have 50 years' working experience. It keep getting worse and worse.
3) It shows that white people have more wages than the black.
4) The people in Metropolitan Statistical Area get more wage than the people don't live in this area.

2. Determine which of the following models for the dependency of y on x = (x_1; x_2; : : : ; x_p) are linear models according to the denitions in class. If they are not linear models, determine if there are transformations that could be applied to make them linear models { if this is the case, be explicit about the transformations and the (unknown) parameters in this transformed model.



(a) $X = X_1$ and $f(x) = \beta_0 + \sin(X_1)$

This is a linear model.

(b) $X = X_1$ and $f(x) = \beta_1 e^{\beta_2 X_1}$

This is not a linear model.

Transformation: $\ln f(x) = \ln\beta_1 + \beta_2 X_1$

$f(x)^* = \beta_1^* + \beta_2 X_1$

(c) $X = X_1$ and $f(x) = \beta_0 + \beta_1 e^{\beta_2 X_1}$

This is not a linear model.

And there is no transformation.

(d) $X = (X_1, X_2)$ and $f(x) = \beta_1 e^{\beta_2 X_1} + \beta_3 X_2$

This is not a linear model.

And There is no transformation.

(e) $X = (X_1, X_2)$ and $f(x) = \beta_1 e^{\beta_2 X_1} X_2^{\beta_3}$

This is not a linear model

Transformation: $\ln f(x) = \ln\beta_1 + \beta_2 X_1 + \beta_3 \ln X_2$

$f(x)^* = \beta_1^* + \beta_2 X_1 + \beta_3 \ln X_2$