Econ 483 presentation

Jingyi Cui

September 23, 2020

Research Question

- What are the effects of age, age, numbers of children, being a smoker or not, living region on medical expenses?
- Econometric Model:
  $\text{charges} = \alpha + \beta_0 \cdot \text{ age} + \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{bmi} + \beta_3 \cdot \text{children} + \beta_4 \cdot \text{smoker} + \beta_4 \cdot \text{region} + \epsilon$

Variables

	age	sex	bmi	children	smoker	region	charges
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31	female	25.740	0	no	southeast	3756.622
7	46	female	33.440	1	no	southeast	8240.590
8	37	female	27.740	3	no	northwest	7281.506
9	37	male	29.830	2	no	northeast	6406.411
10	60	female	25.840	0	no	northwest	28923.137
11	25	male	26.220	0	no	northeast	2721.321

Variables

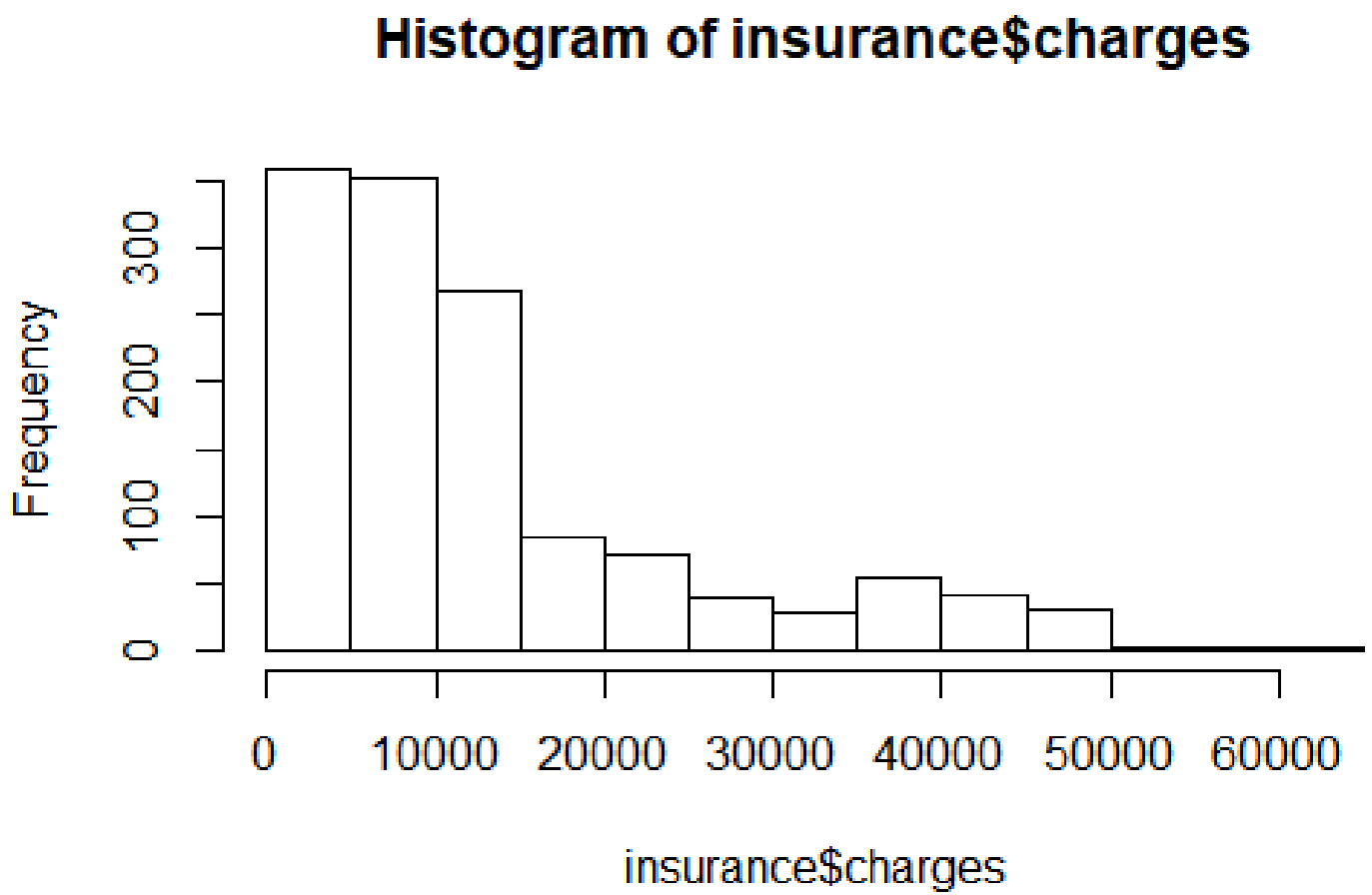
Dependent Variable:

- age: An integer number.
- sex: Female or Male.
- bmi: Body Mass Index. A high BMI can be an indicator of high body fatness. $\text{BMI} = \frac{\text{weight(kg)}}{\text{height(m)}^2}$
- chidren: number of children in the insurance plan. It is an integer.
- smoker: if smoking or not.
- region: A categorical variable and northeast, southeast, southwest and northwest in this dataset.

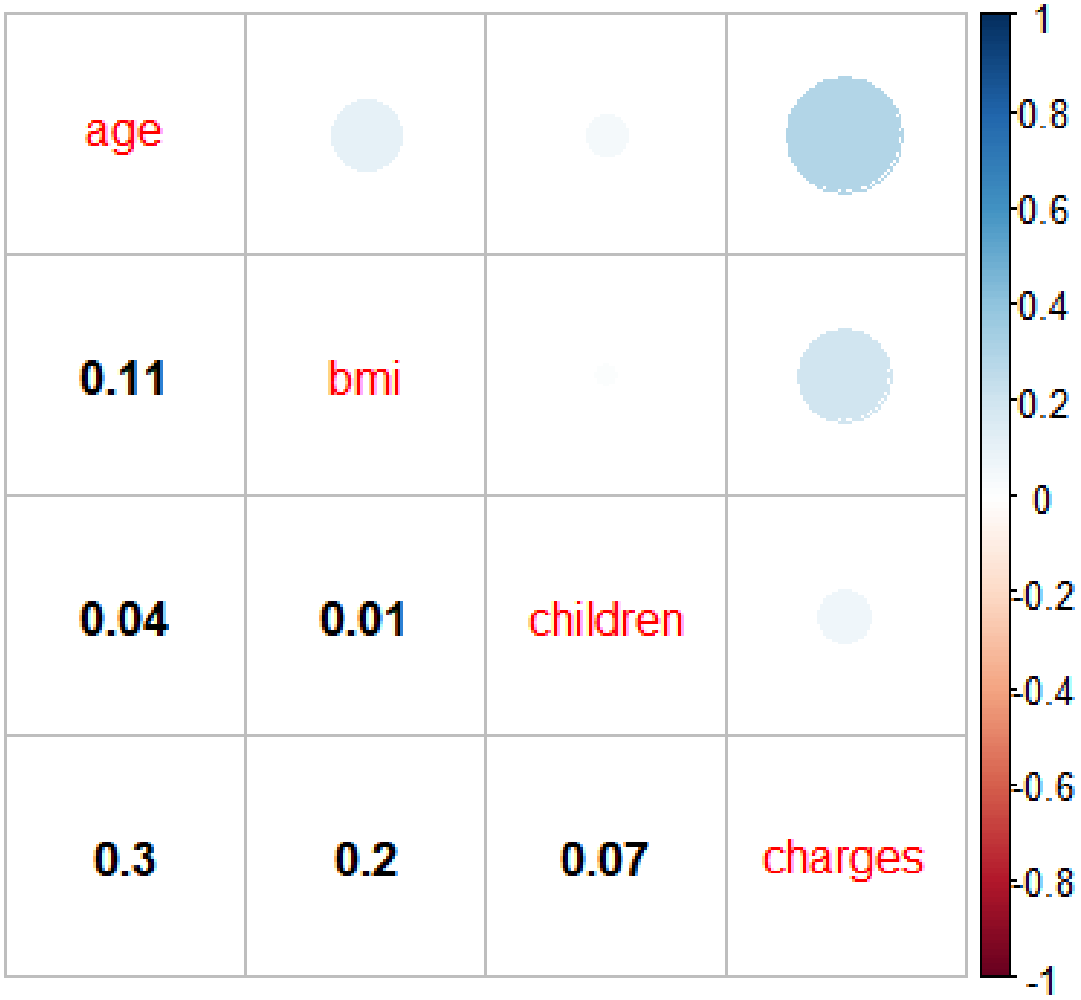
Independent Variable:

- charges: An numerical variables representing total spending in treatment.

Histogram of Charges



Correlation plot



Not very strong relationship among age, bmi, children and charges.

Econometrics Model

$$\text{fit} = \text{lm}(\text{charges} \sim ., \text{data} = \text{insurance})$$

```
> fit$coefficients
    (Intercept)             age          sexmale              bmi         children       smokeryes
    -11938.5386        256.8564       -131.3144        339.1935        475.5005      23848.5345
regionnorthwest regionsoutheast regionsouthwest
     -352.9639      -1035.0220       -960.0510
```

Comparing with female, charges for male decreases by 131.5 on average.

Charges for smokers increase 23848.5345 each year.

The coefficients show that smoking or age increasing may increase charges and having children in the family may result in the increment of charges, such as annual physical exams.

Econmetrics Model

```
> summary(fit)

Call:
lm(formula = charges ~ ., data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Econmetrics Model

fit1 = lm(charges age+children+bmi+sex+bmi30*smoker+region,data=insurance)

```
> summary(fit)$adj.r.squared
[1] 0.7494136
> summary(fit1)$adj.r.squared
[1] 0.8636438
> fit1$coefficients
    (Intercept)             age        children           bmi30
     -1978.9682        265.7278        511.6939        207.5195
regionnorthwest regionsoutheast regionsouthwest bmi30:smokeryes
      -275.0322       -597.1524      -1098.8956      19967.3301
```

Interaction between bmi and smoke or not has a big effect on the charges .

Follow-up Questions

- Focusing on improving the above model
- Nonlinear relationship, such as age and charges
- If omitting a few variables, will the model be improved?