Regression Analysis

What Determines Health Insurance Charges?

Fall 2018 @Purdue University

Motivation

- In 2017 about **91.3%** of people in Unitied States have health insurance. (United States Census Bureau)

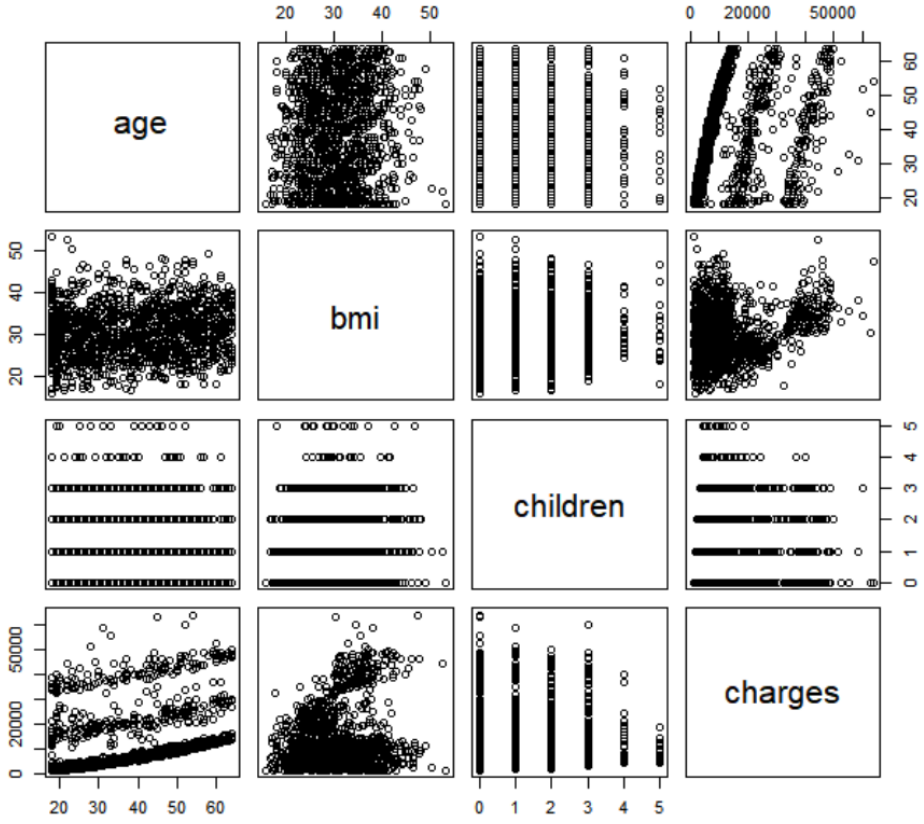
- In 2018 the average annual premium for employer-based family coverage is **$19,616**; for single coverage is **$6,896**. (National Conference of State Legislators)

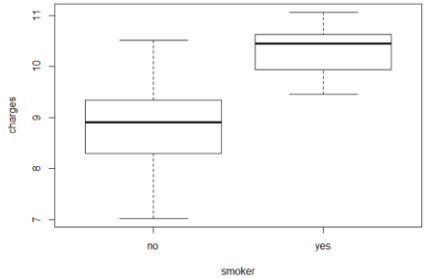
- Questions people want to know...
  - What factors influence the charges of health insurance in US?
  - What's the relationship between factors and charges?
  - What factors impact the charges the most?
  - Can I predict my charges?

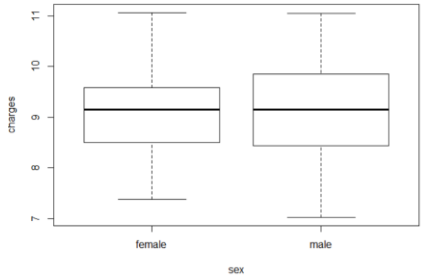
Dataset

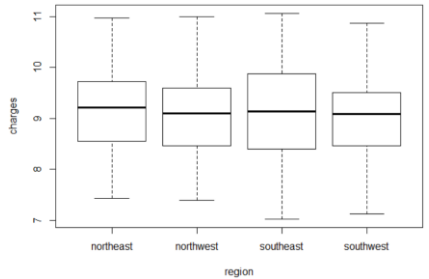
- Health insurance dataset from **Kaggle** (7 columns, 1338 rows)
- **Response variable**
  charges: individual medical costs billed by health insurance
- **Predictors:** **3** continuois variables + **3** categorical variables
  age: age of primary beneficiary
  bmi: body mass index
  children: number of children covered by health insurance
  sex: insurance contractor gender, may be female or male
  smoker: whether smoke or not, may be yes or no
  region: the beneficiary's residential area in the US, may be northeast, southeast, southwest, northwest

Preliminary Analysis









Model Building Process

1. Main Effect Model (m1)
2. Full Interaction Model (m2)
3. Half-way Model (m3)
4. Semi-Final Model (m4)
5. Final Model (m5)

	Name	Model	Adjusted R^2	AIC	p-value
	m5	lm(charges ~ age + bmi + children + smoker + region + bmiOver30:smoker)	0.8625	26313.04	<2.2e-16
Create new variable	m4	lm(charges ~ age + sex + bmi + children + smoker + region + bmiOver30:smoker + bmiOver30:region)	0.8629	26313.65	<2.2e-16
ANOVA	m3	lm(charges ~ age + sex + bmi + children + smoker + region + bmi:smoker + bmi:region)	0.8408	26512.74	<2.2e-16
Step-wise	m2	lm(charges ~ age * sex * bmi * children * smoker * region)	0.8408	26621.12	<2.2e-16
	m1	lm(charges ~ age + sex + bmi + children + smoker + region)	0.7494	27115.51	<2.2e-16

Model Checking

(i) Multicolinearity test ○
(ii) Normality test ○
(iii) Homoscedasticity test ○
(iv) Independence test of error ○

ANOVA Table

	Df	Sum of Square	Mean Square	F value	Pr(>F)
age	1	1.7530e+10	1.7530e+10	869.6224	< 2.2e-16 ***
bmi	1	5.4464e+09	5.4464e+09	270.1827	< 2.2e-16 ***
children	1	5.7152e+08	5.7152e+08	28.3514	1.187e-07 ***
smoker	1	1.2345e+11	1.2345e+11	6123.8808	< 2.2e-16 ***
region	3	2.3320e+08	7.7734e+07	3.8561	0.009211 **
smoker:bmiOver30	2	2.2075e+10	1.1037e+10	547.5361	< 2.2e-16 ***
Residuals	1328	2.6770e+10	2.0158e+07	-	-

Parameter Estimates

Parameter	Point Estimate	Standard Error	t-statistic	p-value	Confidence Interval	
					2.5 %	97.5 %
(Intercept)	-4967.297	954.272	-5.205	2.24e-07	-6839.34248	-3095.2515
age	263.663	8.812	29.920	< 2e-16	246.37553	280.9506
bmi	114.090	34.595	3.298	0.001000	46.22317	181.9578
children	516.796	102.056	5.064	4.69e-07	316.58841	717.0038
smokeryes	13383.160	444.301	30.122	< 2e-16	12511.55257	14254.7679
regionnorthwest	-264.050	352.801	-0.748	0.454328	-956.15797	428.0584
regionsoutheast	-823.426	355.196	-2.318	0.020588	-1520.23154	-126.6203
regionsouthwest	-1221.147	354.076	-3.449	0.000581	-1915.75495	-526.5382
smokerno:bmiOver30	-869.805	426.244	-2.041	0.041485	-1705.99018	-33.6197
smokeryes:bmiOver30	18874.901	643.239	29.344	< 2e-16	17613.02575	20136.7767