DA Group 6: Change List

**1. Introduction**

* Reworded opening paragraph
* From the source “Machine Learning with R”

“In order for a health insurance company to make money, it needs to collect more

in yearly premiums than it spends on medical care to its beneficiaries. As a result,

insurers invest a great deal of time and money in developing models that accurately

forecast medical expenses for the insured population.

Medical expenses are difficult to estimate because the most costly conditions are

rare and seemingly random. Still, some conditions are more prevalent for certain

segments of the population. For instance, lung cancer is more likely among smokers

than non-smokers, and heart disease may be more likely among the obese.

The goal of this analysis is to use patient data to estimate the average medical

care expenses for such population segments. These estimates can be used to create

actuarial tables that set the price of yearly premiums higher or lower, depending on

the expected treatment costs”

* State the range of the data

Age is between 18 and 24

Bmi between 19.8 and 53.13

**2. Exploratory Data Analysis**

* Include boxplot for smokers
* Histogram on diagonal of pairs plot



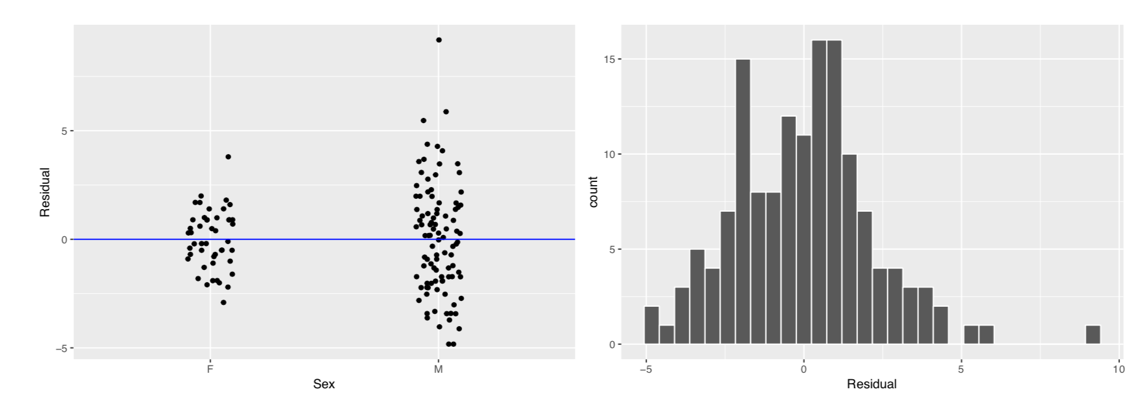
Consider changing the pairs plot to have histogram on diagonal

ggpairs(data=insur, mapping=ggplot2::aes(colour = smoker),

lower=list(combo=wrap("facethist",binwidth=1)))

**3. Formal Data Analysis**

* Combine Figures 4 and 5 such as:



* The fact that the parameter estimate for the categorical variable smoker is -19408.
* 629 is concerning. If someone smoke’s their premium decreases? Maybe we need to do an analysis on just the mean difference between smokers?

**4. Conclusion**

* Quadratic term for BMI
* More data at extremes
* Log transform charges (variation of chargers is an order of magnitude larger than the explanatory variables)
* Higher variance in charges for smokers indicating further complications or covariates

**5. Concerns**

* Need to answer a specific research question, not just ‘predict’ insurance premiums.

*Ex. What variable effects insurance premiums the most?*