On Regression and Model Building

Six Sigma and Data Science. Differences and Synergies.

Ernesto L. Garcia C., PhD

General Objectives of this webinar

- Regression modeling and forecasting are briefly analyzed considering:

1. Parsimonious Modeling, and
2. Data Science Model Building.

Forecasting and model adequacy are compared using real examples with R and Python.

Suggestions and implications on the use of Data Science Regression Modeling for Six Sigma projects are discussed.

Why Regression?

- The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon.
- The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean). [see Ref.]
- Today, I'll use regression in a general sense: the modeling of a process using historical or "real time" data with the purpose of gaining insight and/or forecast its behavior.

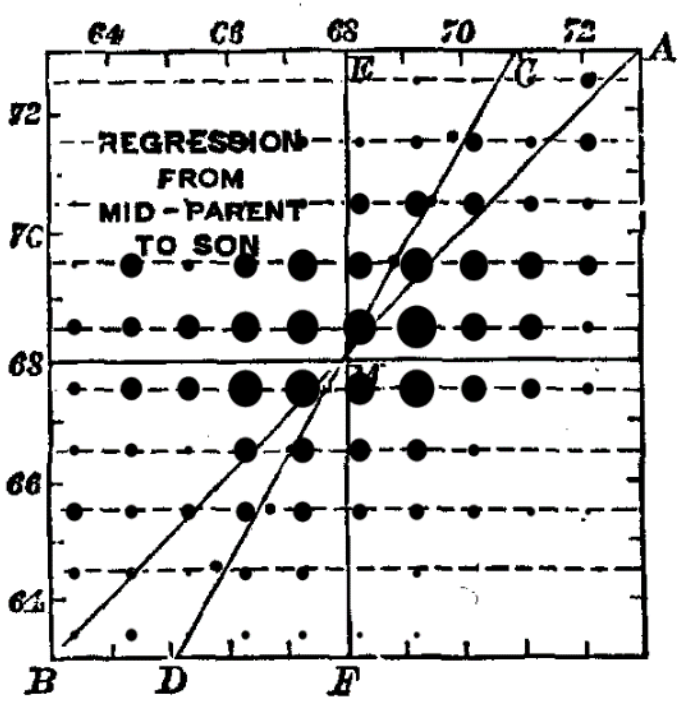


Fig: Galton's height regression.

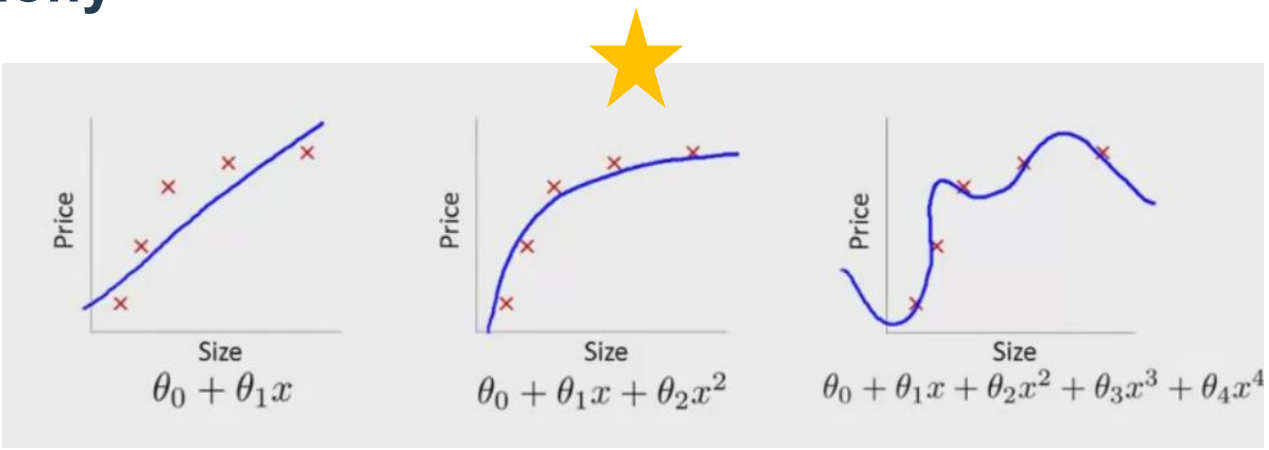
Figure source and credits:
https://commons.wikimedia.org/wiki/File:Galton-height-regress.png
Credit to Madeleine Price Ball.

Ref: https://stats.stackexchange.com/questions/11087/why-are-regression-problems-called-regression-problems.

Occam's razor or the principle of parsimony

- Applied to statistics, a model that has few parameters but achieves a satisfactory level of goodness of fit should be preferred over a model that has many parameters and achieves only a marginally higher level of goodness of fit.

1. **Parsimonious models are easier to interpret and appreciate.** Models with fewer parameters are easier to understand and explain.
2. **Parsimonious models tend to have better predictive ability:** Models with fewer parameters usually perform better when applied to new data.



High Bias (Underfit) | "Just Right" | High Variance (Overfit)

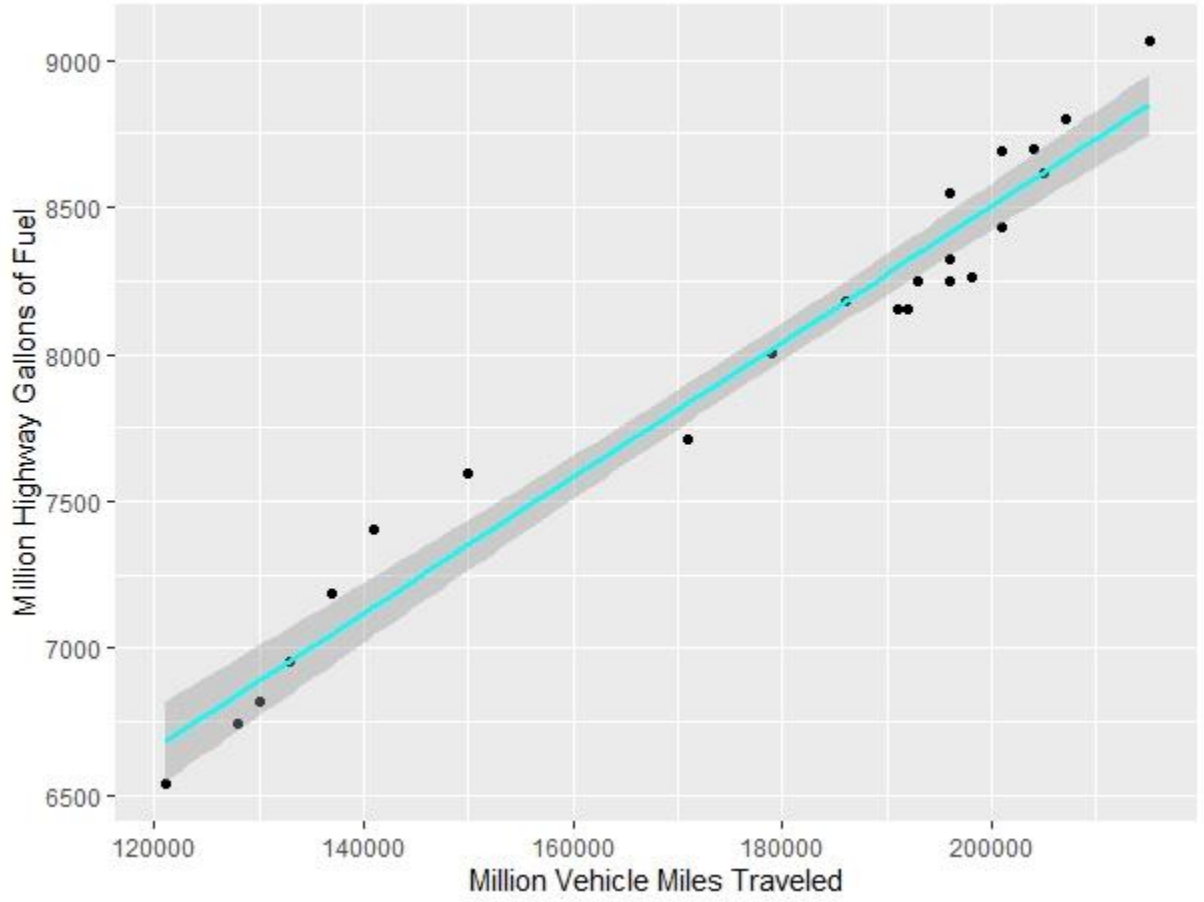
Bias vs Variance in Machine Learning

Source: https://www.statology.org/parsimonious-model/

Fig. source: Machine Learning by Andrew Ng. Coursera

FDOT example

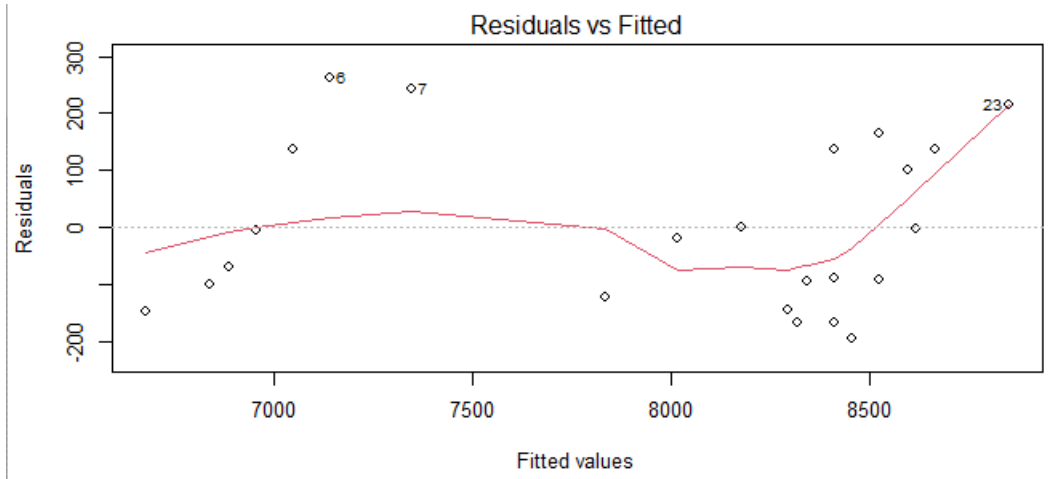
Relation between Vehicle Miles Traveled, and Highway Fuel Gallons consumed (in Millions)

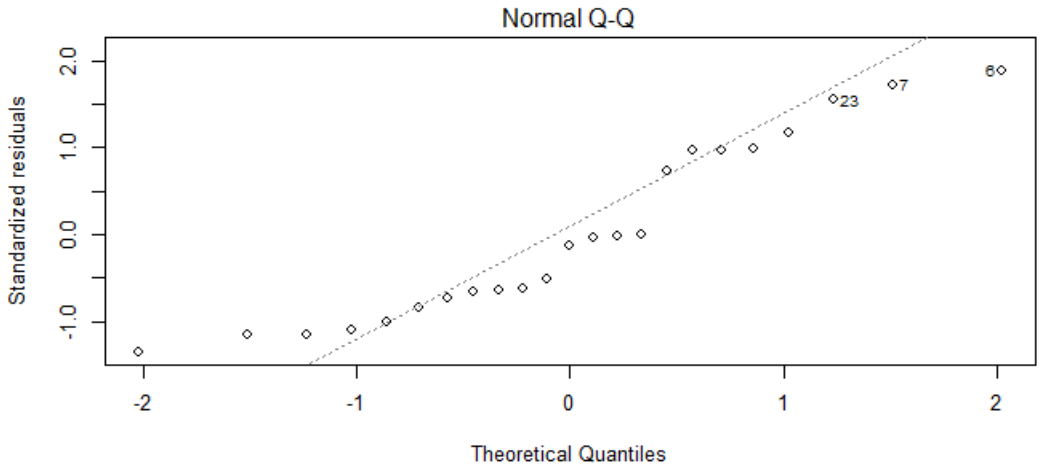


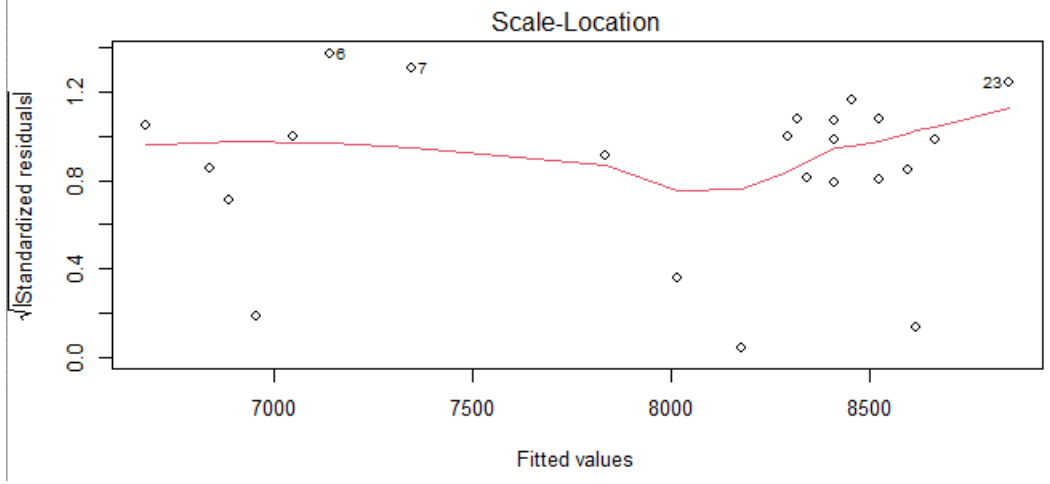

```
Call:
lm(formula = THGF ~ TVMT, data = data <- dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-193.86 -109.42  -18.33  137.15  263.69

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.890e+03  1.847e+02   21.06 1.33e-15 ***
TVMT        2.307e-02  1.029e-03   22.42 3.78e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 147.8 on 21 degrees of freedom
Multiple R-squared:  0.9599,    Adjusted R-squared:  0.958
F-statistic: 502.7 on 1 and 21 DF,  p-value: 3.776e-16
```

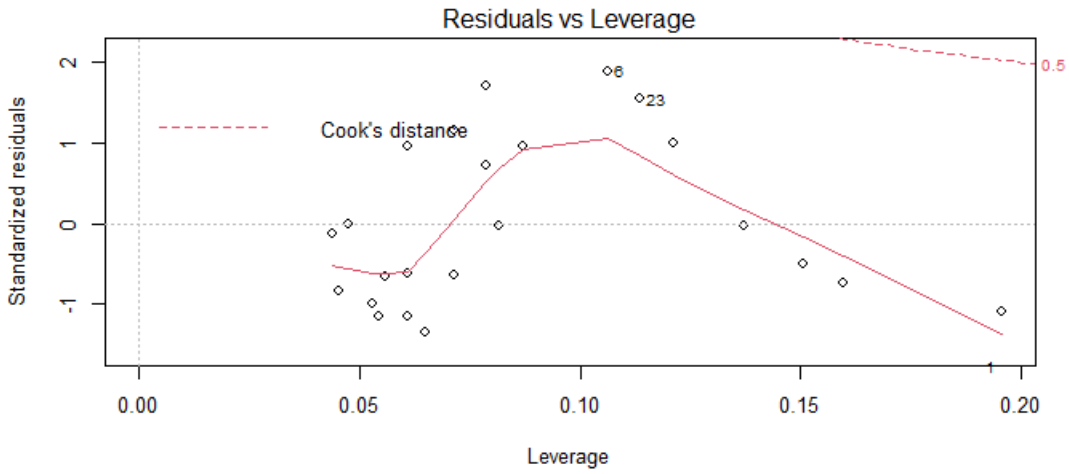
Q: Can I work with this model?

Residual plots for FDOT example.





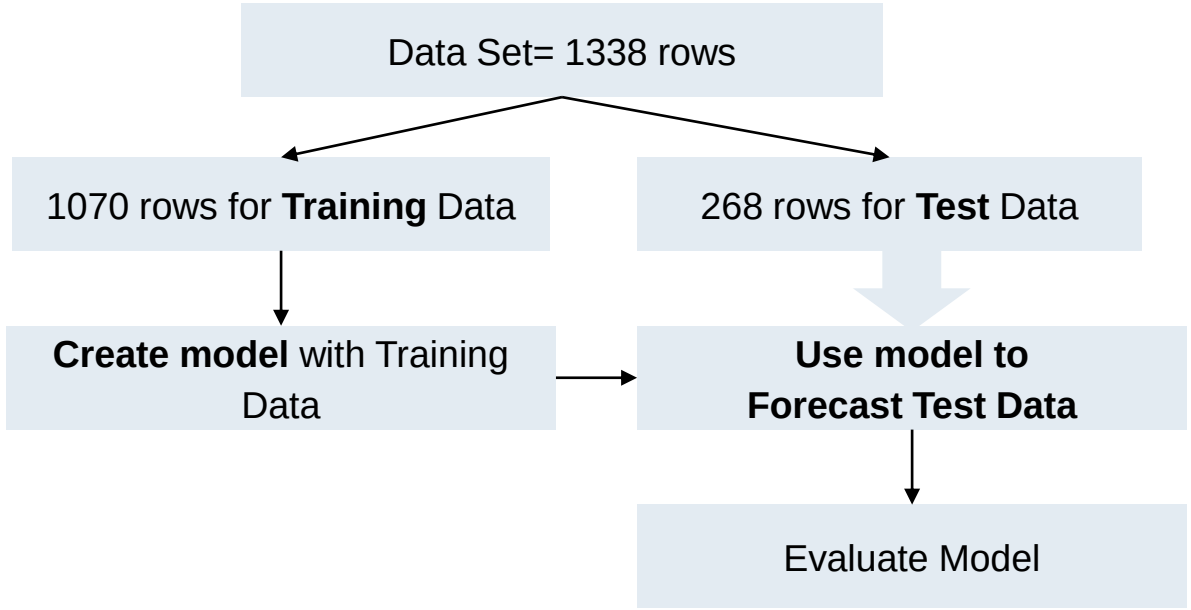




Example: Insurance Charges vs Age, Sex, BMI, Children, Smoking, and Region

Coded Data Set

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	3	16884.92400
1	18	1	33.770	1	0	2	1725.55230
2	28	1	33.000	3	0	2	4449.46200
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.880	0	0	1	3866.85520



Scikit-Learn in Python:

Creating Training and Test Data Sets

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test=train_test_split(X,y, test_size=0.2, random_state=44)
```

Training Multiple Linear Regression on Training set

```
from sklearn.linear_model import LinearRegression #Calling LinearRegression Class
regressor=LinearRegression() #Creating an object of LinearRegression
regressor.fit(X_train, y_train) #Method of the linear regression class
regressor.score(X_train,y_train)
# Printing R^2
```

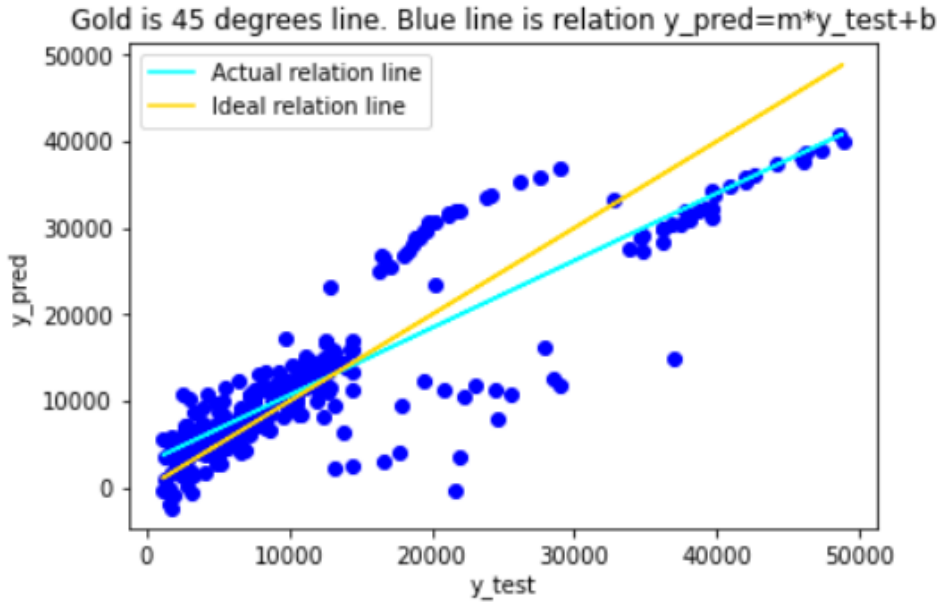
```
']: 0.7497846167934157
```

Predicting Test set results

```
y_pred=regressor.predict(X_test) # Predicting values on test data
#print(y_pred)
```


Source: Kaggle dataset for insurance cost prediction

ML approach: plot *y_pred* vs *y_test*




```
Slope: m= 0.78
Intercept: b= 2924.65
```

y_test= Real response value for test data
y_pred= Predicted response value for test data
Correlation(***y_test***, ***y_pred***)~0.87
Aspect Ratio=(1-Correlation (***y_test***, ***y_pred***))**(0.5-1)=2.76

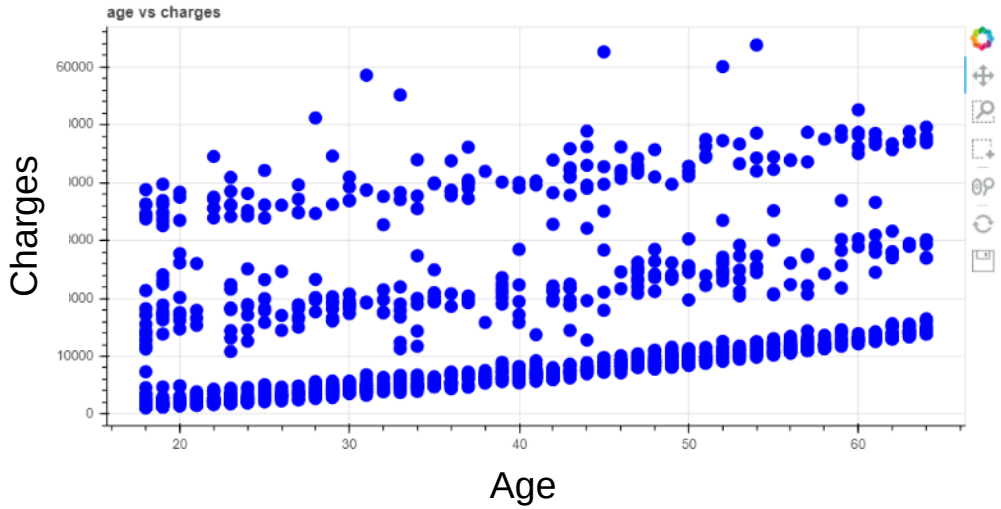
A perfect model will have a scatter plot of ***y_pred*** vs ***y_test*** along the Gold line.

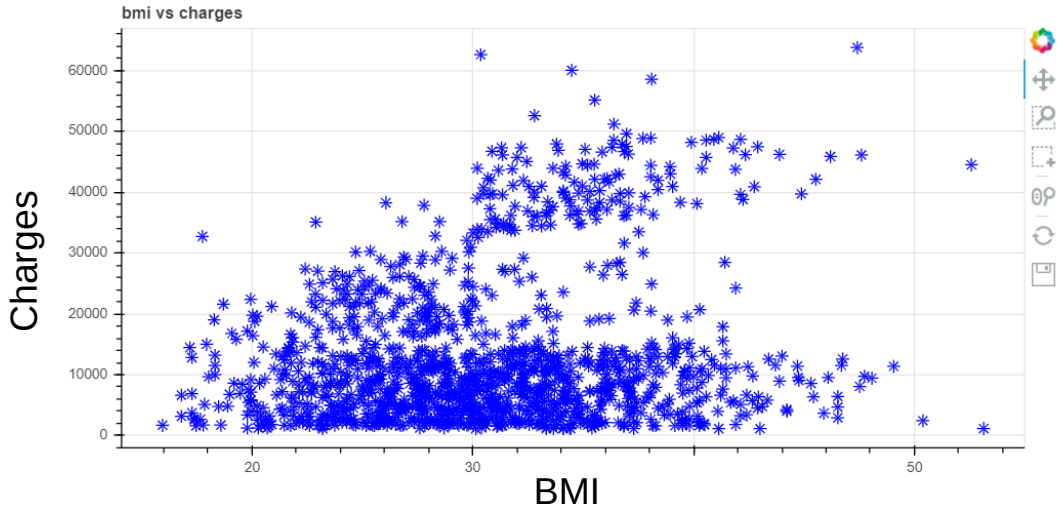
Perhaps we can use other “engines” available:

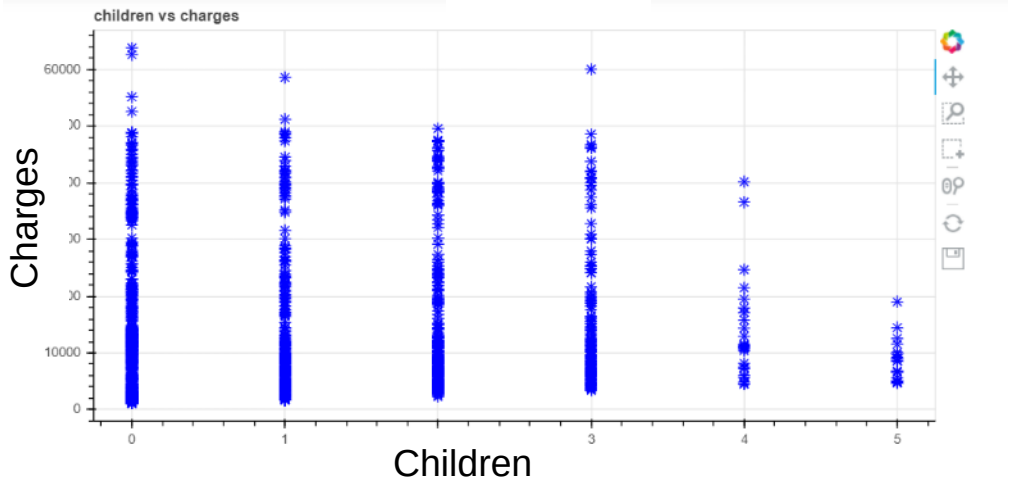
- Deep Neural Nets
- Regression Trees
- Ensemble models
- Ridge regression
- Etc.

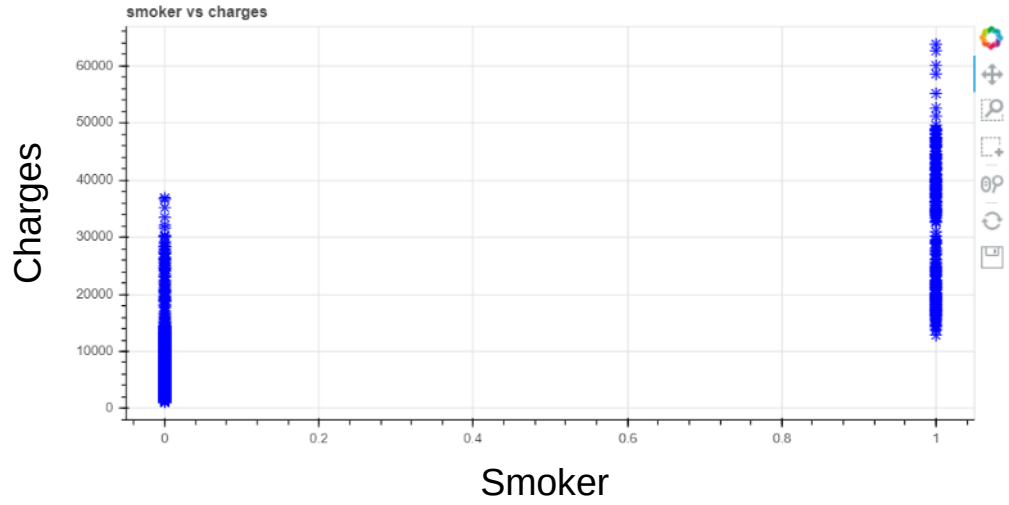
But before doing that…

Look at the data: EDA

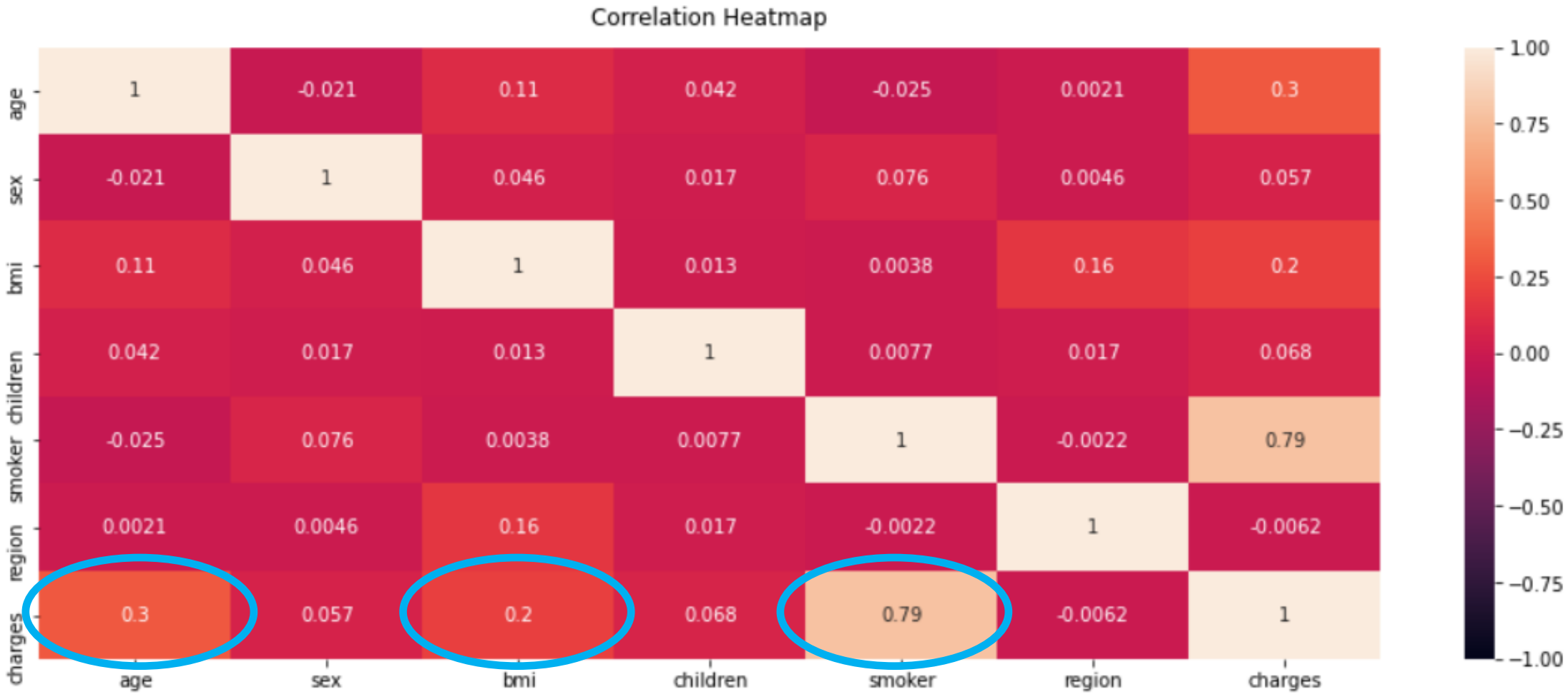








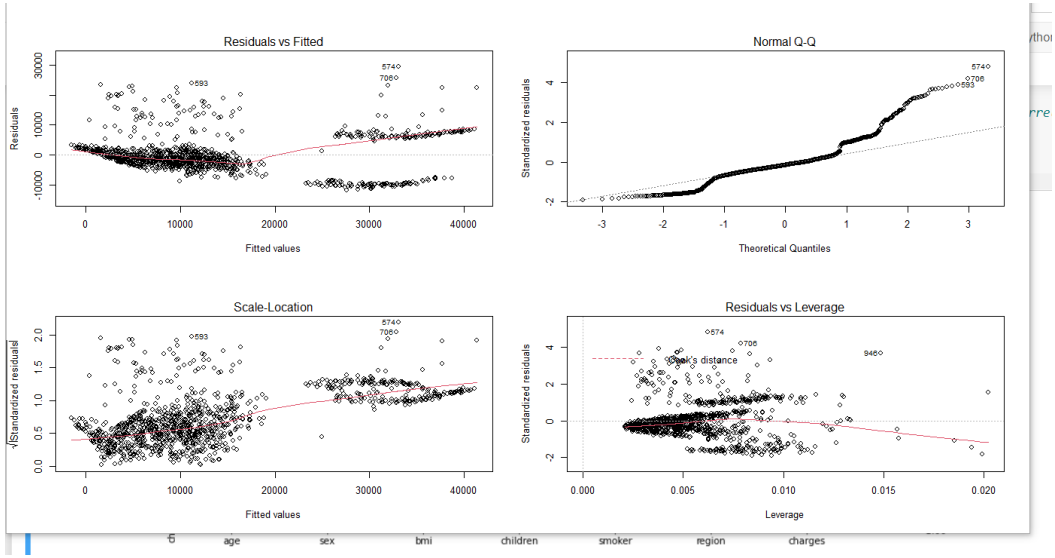
EDA


Correlation Heatmap
age
sex
bmi
children
smoker
region
charges
1
-0.021
0.11
0.042
-0.025
0.0021
0.3
-0.021
1
0.046
0.017
0.076
0.0046
0.057
0.11
0.046
1
0.013
0.0038
0.16
0.2
0.042
0.017
0.013
1
0.0077
0.017
0.068
-0.025
0.076
0.0038
0.0077
1
-0.0022
0.79
0.0021
0.0046
0.16
0.017
-0.0022
1
-0.0062
0.3
0.057
0.2
0.068
0.79
-0.0062
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Smoker

Regression and Residual plots for Insurance cost example.

```
Call:
lm(formula = charges ~ bmi + children + smoker + Age + Sex, data = data <- dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-11758.8  -2977.1   -952.9   1471.9  29517.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12322.66    1089.19 -11.314  < 2e-16 ***
bmi            336.00      31.78  10.574  < 2e-16 ***
children       495.82     157.11   3.156  0.00164 **
smoker       23978.47     466.63  51.386  < 2e-16 ***
Age            254.25      13.42  18.940  < 2e-16 ***
Sex           -222.53     378.09  -0.589  0.55627
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6160 on 1064 degrees of freedom
Multiple R-squared:  0.7488,    Adjusted R-squared:  0.7477
F-statistic: 634.5 on 5 and 1064 DF,  p-value: < 2.2e-16
```



ANOVA of regression: Only Sum of Squares.

Response: charges			
Feature	Df	Sum Sq	"Practical" contribution
bmi	1	5.90E+09	3.7%
children	1	6.41E+08	0.4%
smoker	1	1.00E+11	62.3%
Age	1	1.36E+10	8.5%
Sex	1	1.31E+07	0.0%
Residuals	1064	4.04E+10	25.1%
Total		1.60748E+11	

Can we use ***smoker*** as a first initial solution and develop a better model with more investigation on other variables, possibly, not included?

Can we live with it?

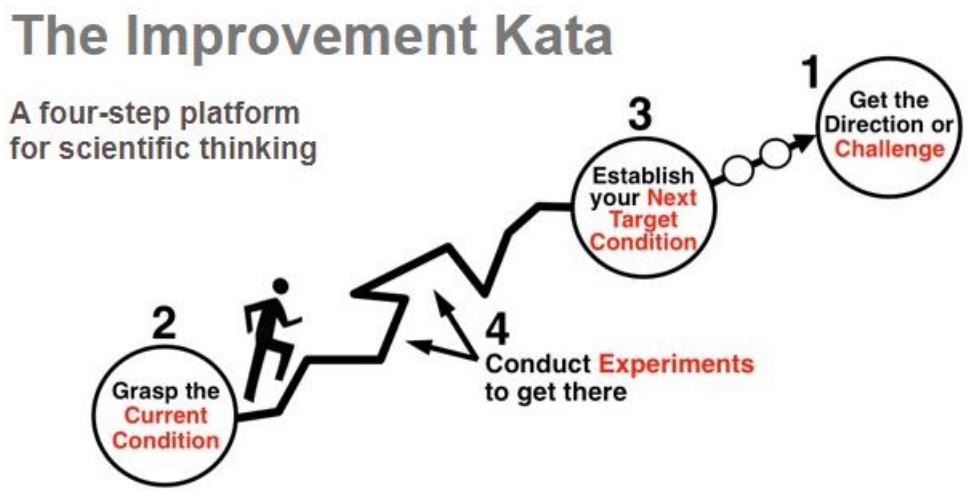
Since all models are wrong, the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.

George P. Box

Source: *Box, G. E. P. (1976), "Science and statistics", Journal of the American Statistical Association, **71** (356): 791–799, p:792.*

Some thoughts

- Know your processes. If you are not an expert, work with one. Better: work with a team. They can provide more insight.
- Do your homework: EDA .
- Can you answer the question with an elegant model? This depends on the customer.
- Faster and available algorithms and modules are tools used with your Improvement Kata: What is what the customer want? What is the challenge?



Sources: http://www-personal.umich.edu/~mrother/The_Improvement_Kata.html

Thank you for your time

Appendix: Linear Regression Assumptions

There are four assumptions associated with a linear regression model:

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residuals is the same for any value of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed (for significance testing).

Appendix: Residuals Plots explanation

- **Scale-Location plot**: It's also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It's good if you see a horizontal line with equally (randomly) spread points.
- **Residuals vs Fitted plot:** This plot shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and the response. The pattern could show up in this plot if the model doesn't capture the non-linear relationship. If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships.
- **Normal Q-Q plot**: This plot shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely? It's good if residuals are lined well on the straight dashed line.
- **Residuals vs Leverage plot**: This plot helps us to find influential cases if any. Not all outliers are influential in linear regression analysis. Even though data have extreme values, they might not be influential to determine a regression line. That means, the results wouldn't be much different if we either include or exclude them from analysis. On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis. Unlike the other plots, this time patterns are not relevant. We watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line. Look for cases outside of a dashed line, Cook's distance. When cases are outside of the Cook's distance (meaning they have high Cook's distance scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases.

Source: https://data.library.virginia.edu/diagnostic-plots/