Regression

HARVARD UNIVERSITY

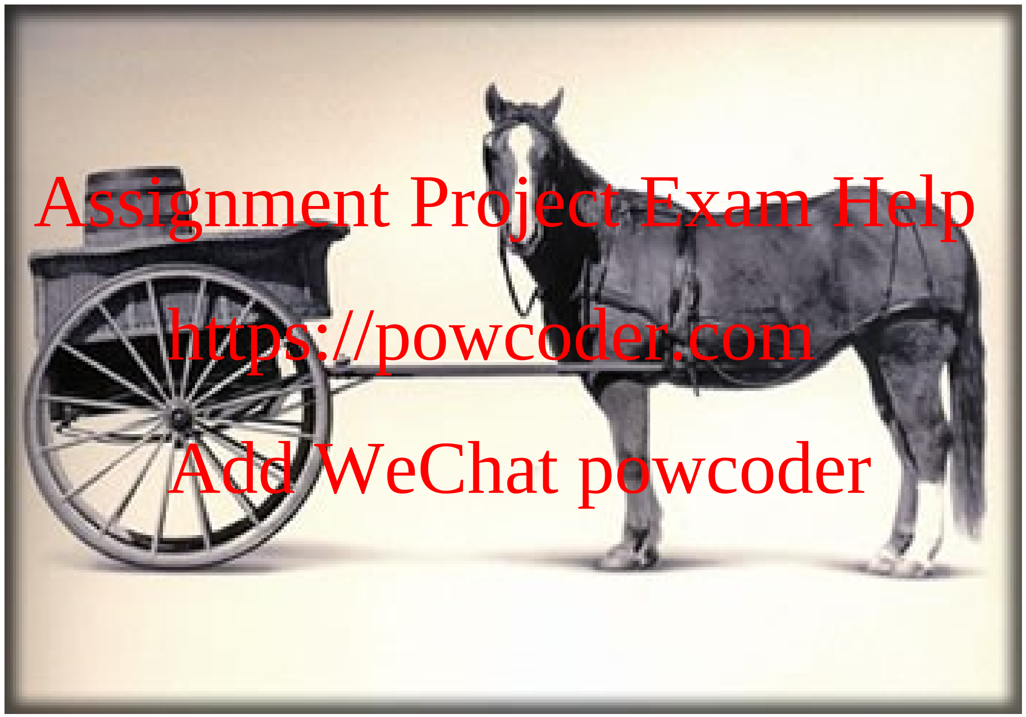
Agenda

Start	End	Item
		Regression in Action
		Partitioning Data
		Understanding Regression (simple univariate)
		Build & Evaluate a Multiple Linear Regression
		Housekeeping

HARVARD UNIVERSITY

Let's Practice

<u>Open A_Regression.R</u>



Predict Diamond Prices with linear regression.


COAL FOR CHRISTMAS?
WITH HEAT AND COMPRESSION
I CAN MAKE MY OWN DIAMONDS
DIYLOL.COM

Agenda

Start	End	Item
		Regression in Action
		Partitioning Data
		Understanding Regression (simple univariate)
		Build & Evaluate a Multiple Linear Regression
		Housekeeping

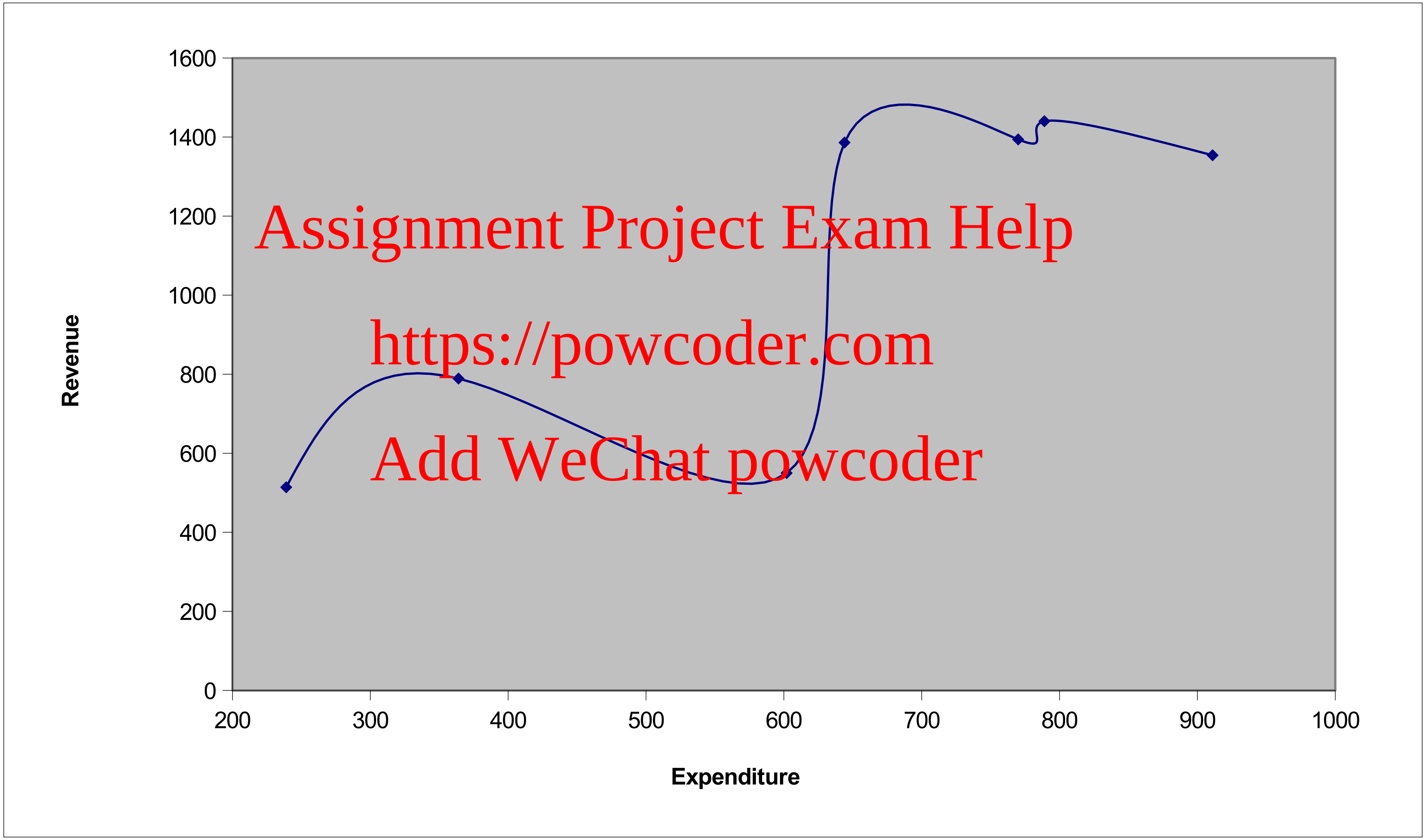
HARVARD UNIVERSITY

The Problem of Overfitting

- Statistical models can produce highly complex explanations of relationships between variables
- The "fit" may be excellent
- When used with new data, models of great complexity may do not do so well.

HARVARD UNIVERSITY

100% fit – not useful for new data



HARVARD UNIVERSITY

Another view of overfitting to a problem…

Overfitting, continued.

Causes:

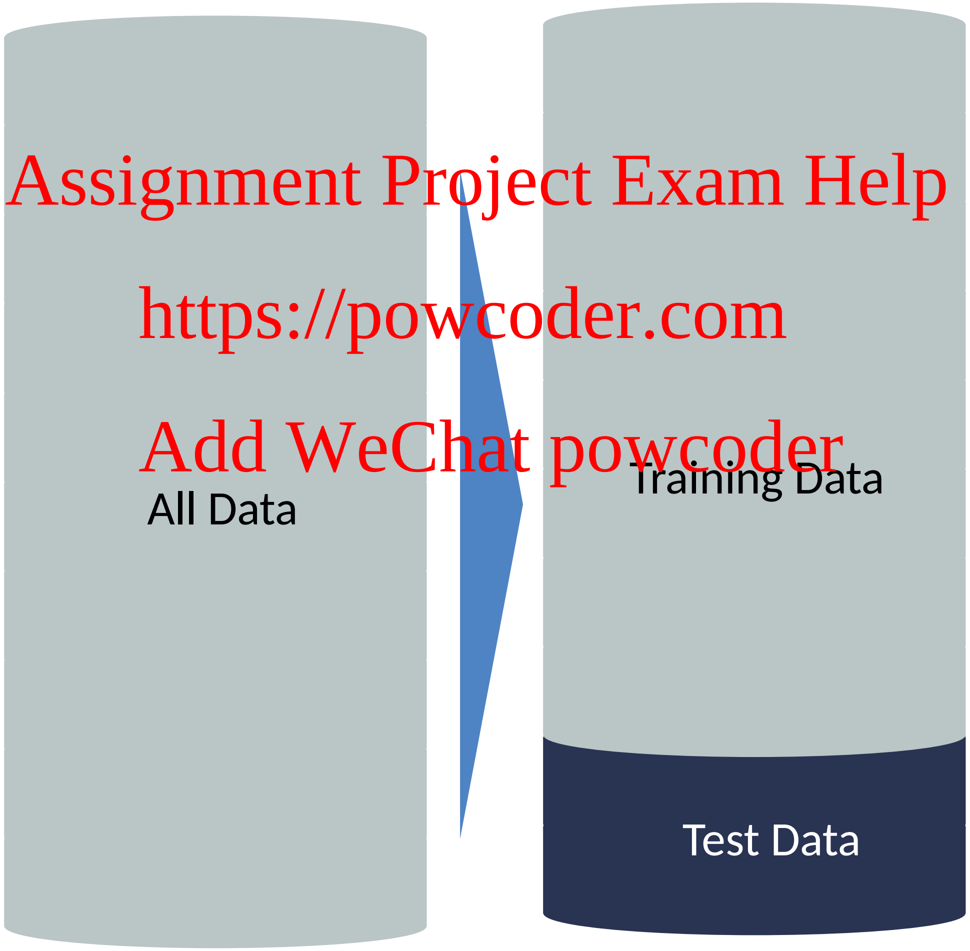
- Too many predictors start to inject noise not signal
- Not adhering to a priori partitioning – *up next*
- Lack of data knowledge & problem understanding

Consequence: Deployed model will not work as well as expected with completely new data.

Minimize Overfitting - Partitioning

Divide data into training portion and validation portion

Test model on the test portion

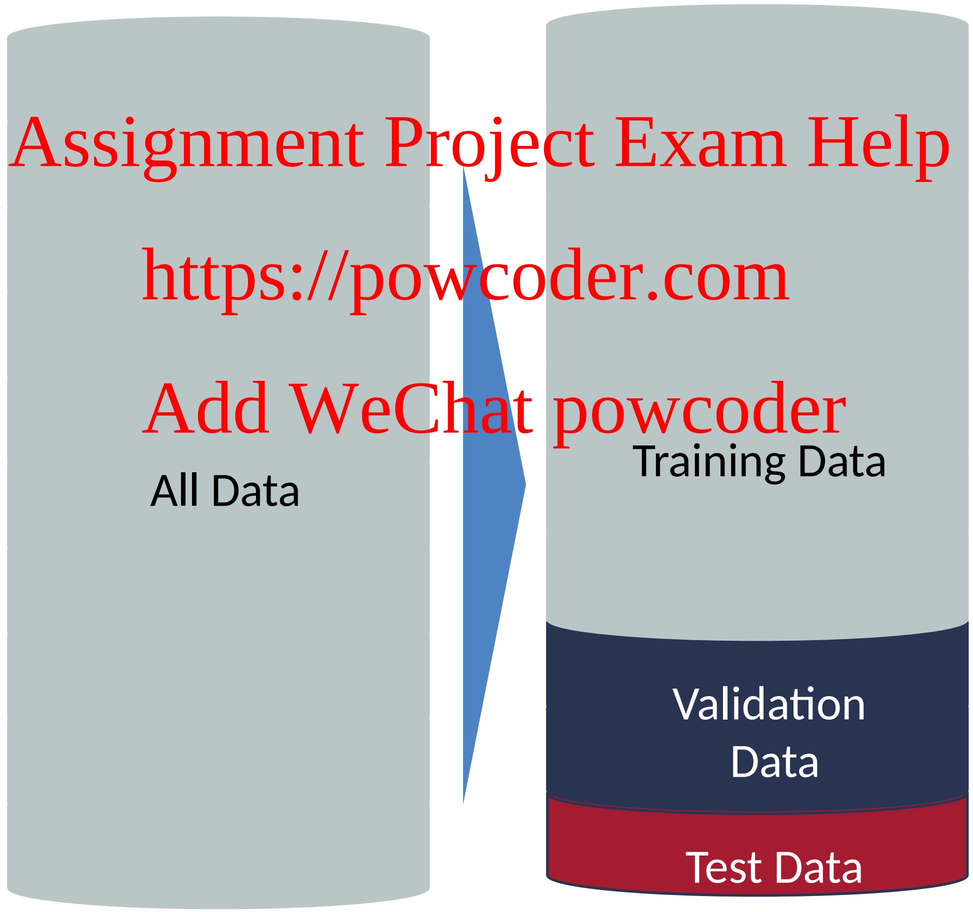


Minimize Overfitting - Partitioning

Divide data into training portion , validation & test portions

Tune a model and/or compare models with the validation portion

The “true” way a model will behave when launched on new data.

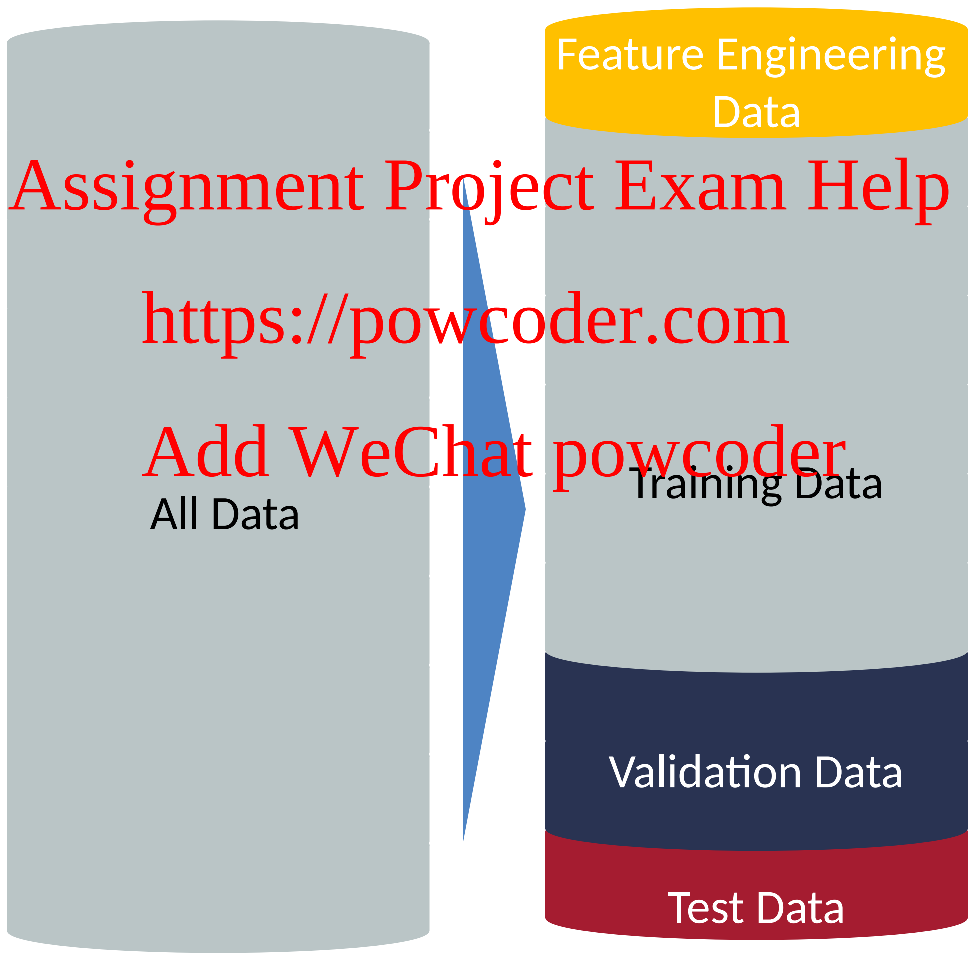


HARVARD UNIVERSITY

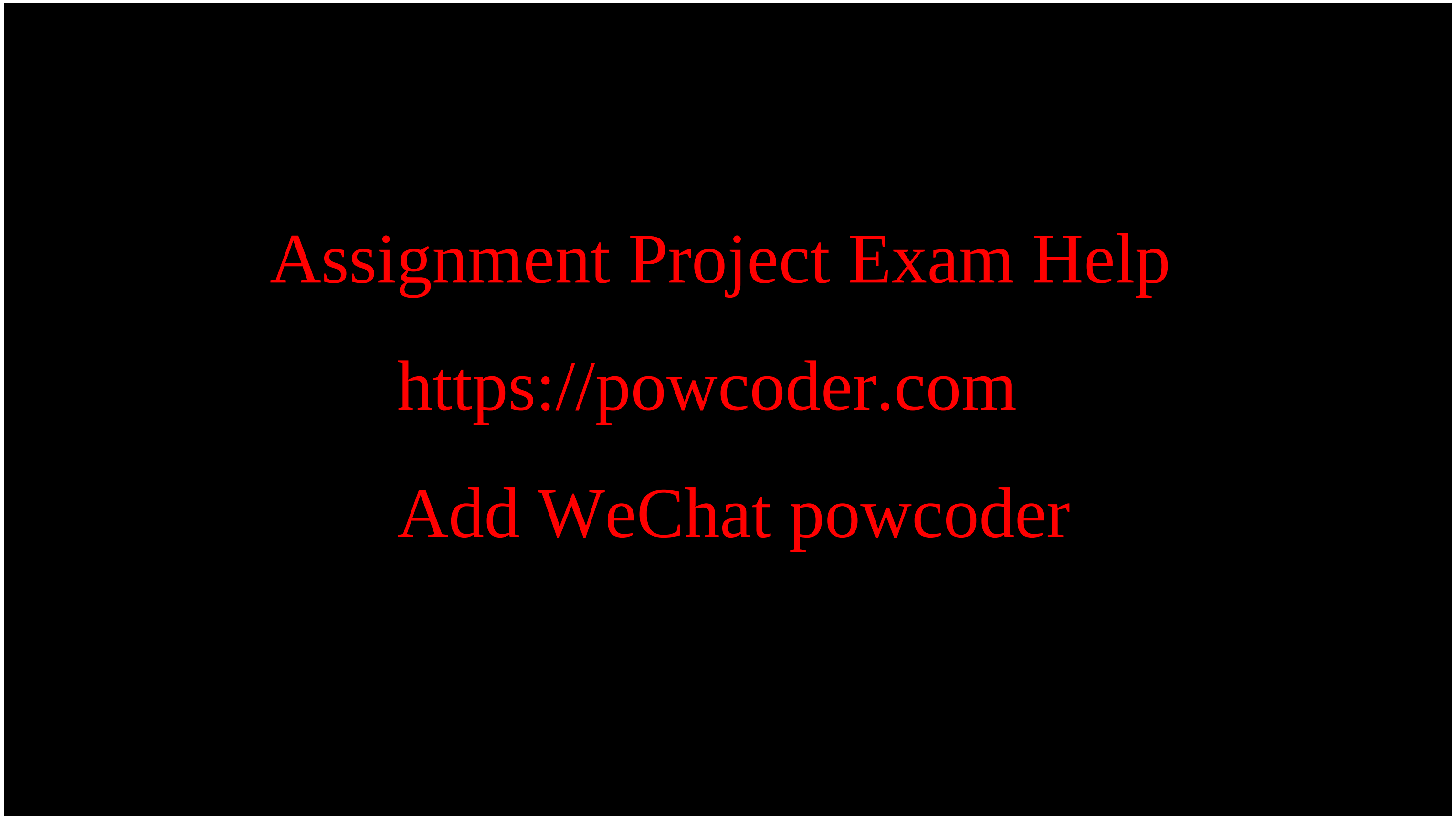
Best Practice

If you have enough data and the model impact is large, this is a good partitioning schema

However, this much effort is seldom undertaken.



Next Glass



Next Glass is not out of business but leveraged chemistry & modeling for it’s business model.

Let’s Practice

Open B_anotherGlass.R

Train/Test

Train/Validation/Test

Engineering/Train/
Validation/Test

Review

- Data Mining:
  - Supervised - Classification & Prediction
  - Unsupervised- Association Rules, Data Reduction, Data Exploration & Visualization
- Before algorithms can be applied, data must be explored then pre-processed (treated)
- To evaluate performance and to avoid overfitting, data partitioning is used
- Models are fit to the training partition and assessed on the validation and test partitions

Today's lesson explores partitioning and simple prediction.

HARVARD UNIVERSITY

Agenda

Start	End	Item
		Regression in Action
		Partitioning Data
		Understanding Regression (simple univariate)
		Build & Evaluate a Multiple Linear Regression
		Housekeeping

HARVARD UNIVERSITY

Supervised Learning

- Goal: Predict a single “target” or “outcome” variable
- Training data, where target value is known
- Score to data where value is not known
- Methods: Classification and Prediction

Supervised Learning

Inferring a function from labeled data.

"Learn from telling", "Look at my data and I will tell you what to predict"

Business Context

Marketing-Will a customer buy yes or no? How much will a customer spend?

Operations- Will an applicant default? When will a machine break?

Sports Analytics- How many points will the Bears' QB score? What is the Bears' probability of winning?

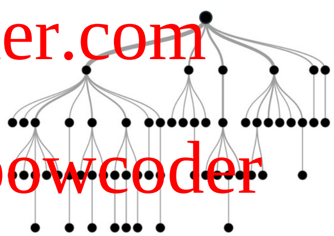
Requires expertise and stakeholder buy in

Data Setup

Flat "Excel" file. Each row is a record or observation. Each column is an attribute of the record.

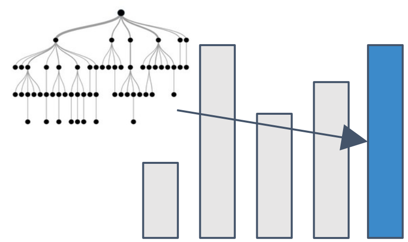
One column is the outcome, y or target attribute.

Algorithm



Modeling e.g. K-NN, linear regression, decision tree, random forest etc.

Application



Use the model to make predictions for the target label on the new data.

HARVARD UNIVERSITY

Supervised Learning Example

Inferring a function from labeled data.

“Learn from telling”, “Look at my data and I will tell you what to predict”

\#  $=f(\ldots)$

What impacts ice cream sales?

Linear Regression for continuous outcomes

 = + (*temperature) + (*day) + (*price) + error

Some linear combination of temperature values, day of the week dummy variables and price estimate the number of cones that will sell.

The linear combination equation captures information

outcome

coefficients

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

constant

error (noise)

predictors

HARVARD UNIVERSITY

The linear combination equation captures information

Outcome:
The "dependent", "y" or "target".
Number of Ice Cream Cones

$$\boxed{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

The linear combination equation captures information

$$Y = \boxed{\beta_0} + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

Constant:
The "intercept" or "beta-naught" has no predictor associated with it.
Avg. Number of Ice Cream Cones expected to sell if predictors were all 0.

HARVARD UNIVERSITY

The linear combination equation captures information

$$Y = \beta_0 + \beta_1 \boxed{x_1} + \beta_2 \boxed{x_2} + \cdots + \beta_p \boxed{x_p} + \epsilon,$$

Predictors:
The "informative features", "x" or "independent" variables.
Variables affecting sales in the data, temp, day, & price.

The linear combination equation captures information

$$Y = \beta_0 + \boxed{\beta_1} x_1 + \boxed{\beta_2} x_2 + \cdots + \boxed{\beta_p} x_p + \epsilon,$$

Coefficients:
The “weight”, “betas” or “coefficients” multiplied with the specific “x” variable value.

The linear combination equation captures information

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \boxed{\epsilon},$$

Error:
The “error” or “noise” represents the value the equation is wrong compared to the actual Y.

The linear combination equation captures information

outcome

coefficients

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

constant

error (noise)

predictors

The combinations of beta coefficients seeks to minimize the squared errors between the actual Y values and the equation. **This combination manifests as the “best fit line”**

 = + 0.*temperature + *satuday_dummy + -0.5*price

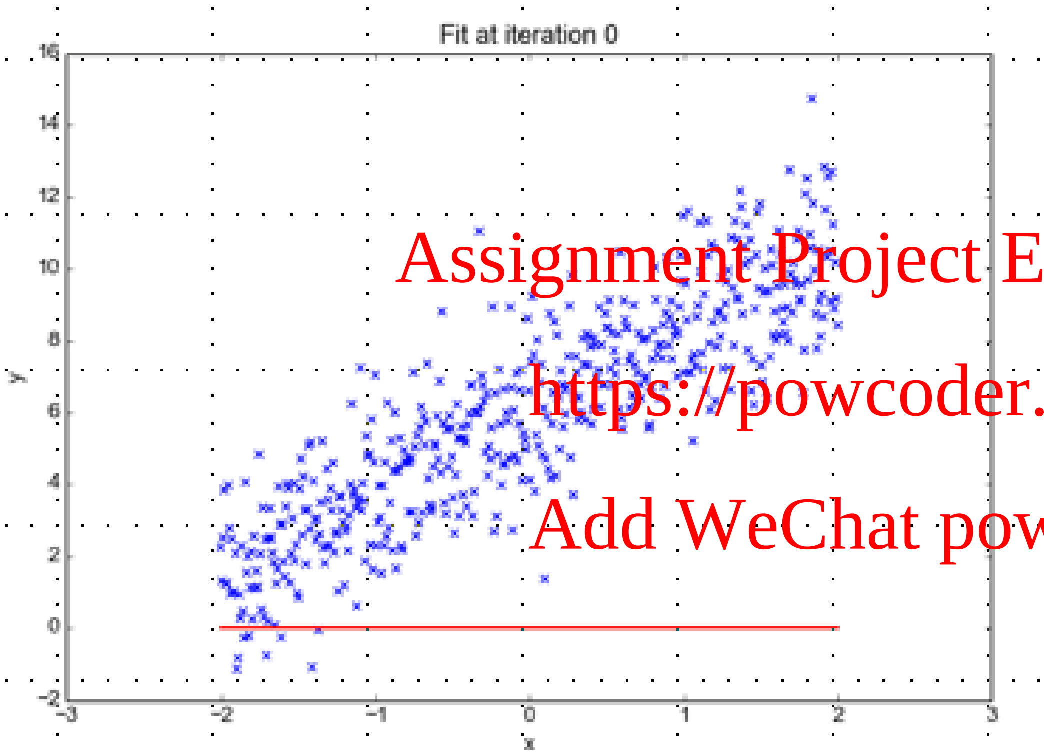
Beta-Naught	Temperature	Saturday_dummy	Price
6	0.25 * 80 degrees	3 * 1	-0.5 * 5
5	0.25 * 88	3 * 0	-0.5 * 2

 = + 0.*temperature + *satuday_dummy + -0.5*price

Beta-Naught	Temperature	Satuday_dummy	Price	Best Fit Prediction
6	0.25 * 80 degrees	3 * 1	-0.5 * 5	26.5
6	0.25 * 88	3 * 0	-0.5 * 2	27

HARVARD UNIVERSITY

Minimizing the Sum of Ordinary Least Squared Errors



Equation:
Y= 0 +(0*x)

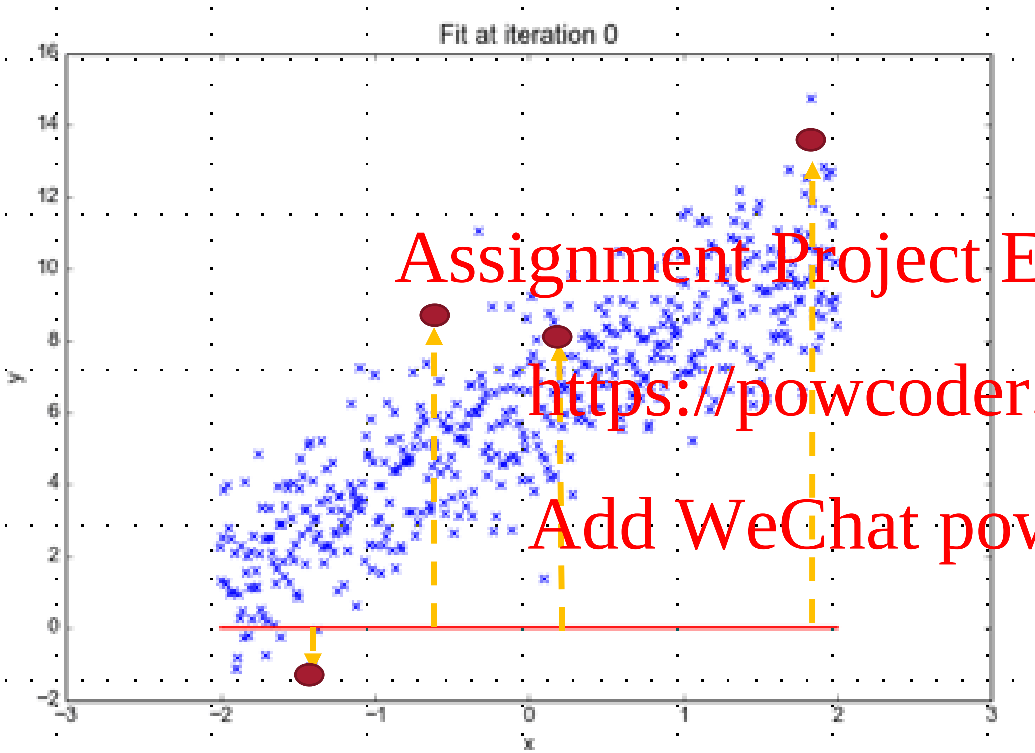
Beta "Naught" = 0
Intercept is 0

X beta coefficient = 0
No slope

Blue points Y Values represent actual outcome. | MINUS | Red line is the predicted outcome | Equals the Error

HARVARD UNIVERSITY

Big Errors



Equation:
Y= 0 +(0*x)

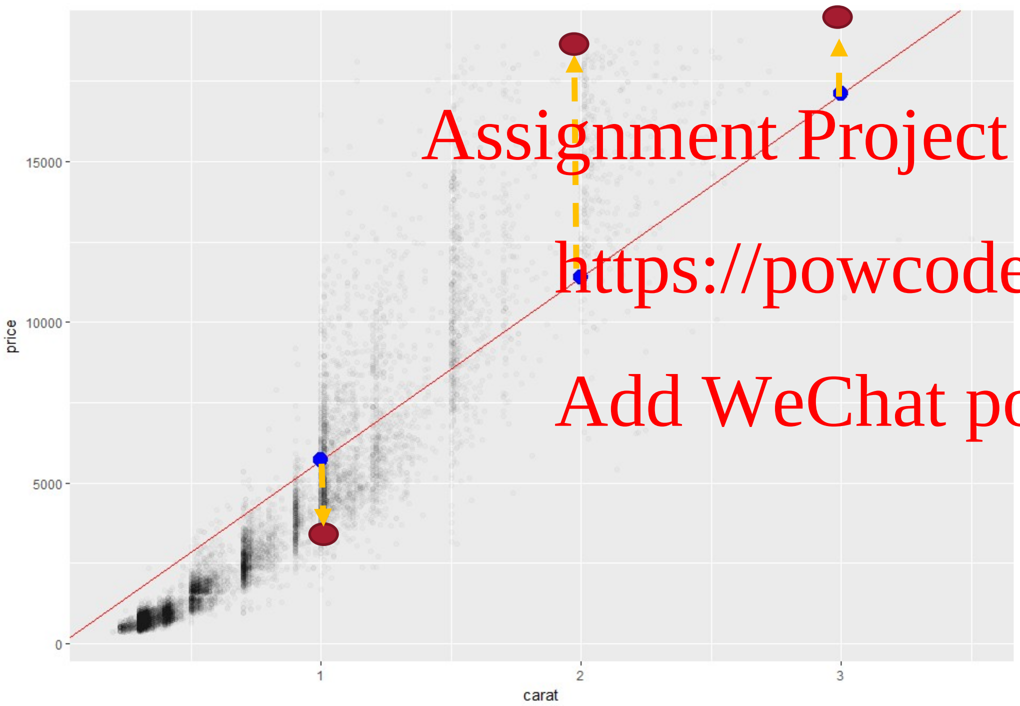
Beta "Naught" = 0
Intercept is 0

X beta coefficient = 0
No slope

Blue points Y Values represent actual outcome. | MINUS | Red line is the predicted outcome | Equals the Error

HARVARD UNIVERSITY

What’s really going on?



- Errors between a prediction and actual.
- Notice some are negative and some are positive

Why Squared Error?

Blue points Y Values represent actual outcome.

MINUS

Red line is the predicted outcome

Equals the Error

Cones	Prediction	Error
85	67	18
48	42	6
45	54	-9
27	95	-68
32	1	31
30	48	-18
69	51	18
80	95	-15
15	20	-5
61	22	39

$$(18 + 6 + -9 + -68 + 31 + \mathbf{-18 + 18} + -15 + -5 + 39) = \mathbf{-3}$$

Why Squared Error?

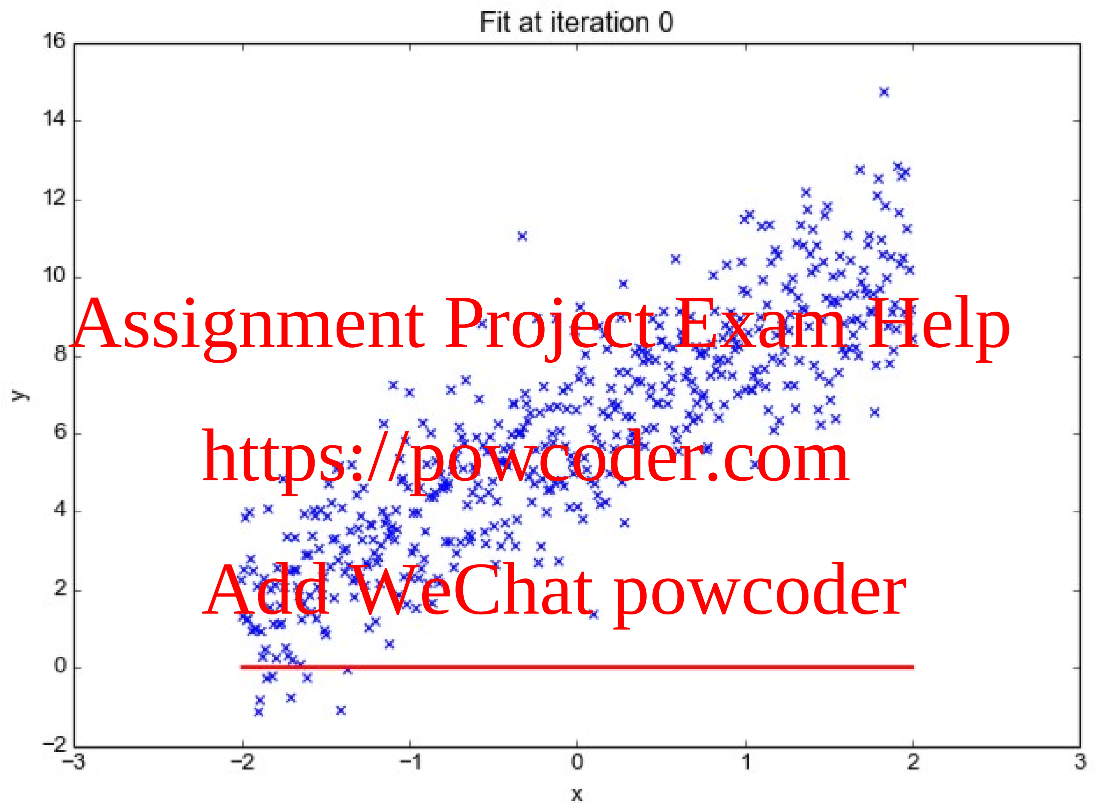
Without squaring the errors positive & negative prediction errors cancel each other out.

$(18^2 + 6^2 + -9^2 + -68^2 + 31^2 + \mathbf{-18^2 + 18^2} + -15^2 + -5^2 + 39^2)$

$324 + 36 + 81 + 4624 + 961 + \mathbf{324 + 324} + 225 + 25 + 1521 = \mathbf{8445}$

Squaring the error means all errors have the same impact on the optimization function.

So what is really going on?



The algorithm is optimizing the inputs and weights (beta's) to **minimize the sum of squared errors**.

This is called "ordinary least squares (OLS).

HARVARD UNIVERSITY

Let's Practice

Open C_Regression_v1.R

HARVARD UNIVERSITY

Topics

- Explanatory vs. predictive modeling with regression
- Example: prices of Toyota Corollas
- Fitting a predictive model
- Assessing predictive accuracy
- Selecting a subset of predictors

HARVARD UNIVERSITY

Explanatory Vs Predictive

Reviewing beta coefficients can explain relationships

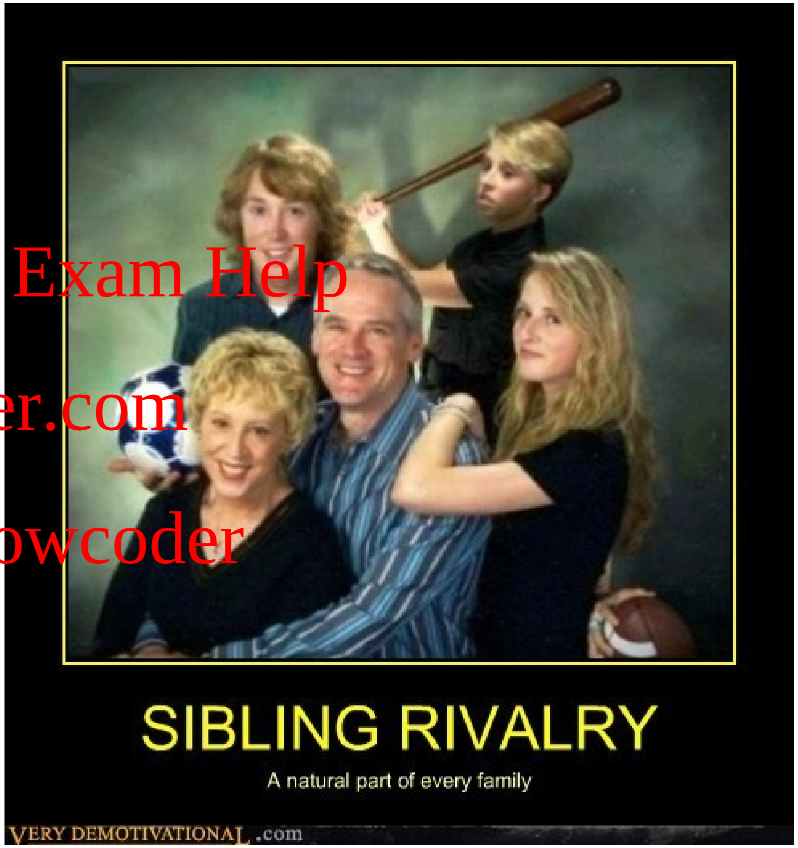
- There is a positive relationship between number of rooms and housing prices.

```
> fit2

Call:
lm(formula = MEDV ~ RM, data = trainSet)

Coefficients:
(Intercept)           RM
    -38.704        9.534
```

- *Holding all other inputs constant* the median price would increase 9.534 for each room.



Explanatory Modeling

Goal: Explain relationship between predictors (explanatory variables) and target

- Familiar use of regression in data analysis

- Model Goal: Fit the data well and understand the contribution of explanatory variables to the model

- “goodness-of-fit”: R^2, residual analysis, p-values

HARVARD UNIVERSITY

Predictive Modeling

Goal: predict target values in other data where we have predictor values, but not target values

- Classic data mining context
- Model Goal: **Optimize predictive accuracy**
- Train model on training data
- Assess performance on validation (hold-out) data
- Explaining role of predictors is not primary purpose (but useful)

Explanatory Vs Predictive Modeling

Make sure you understand the point of your project (explanatory or predictive)

- Do leaders want to understand a phenomena?
- Do leaders want to make accurate predictions about the future?

This impacts how you evaluate the model and even what variables you choose.

Sibling rivalry never ends...

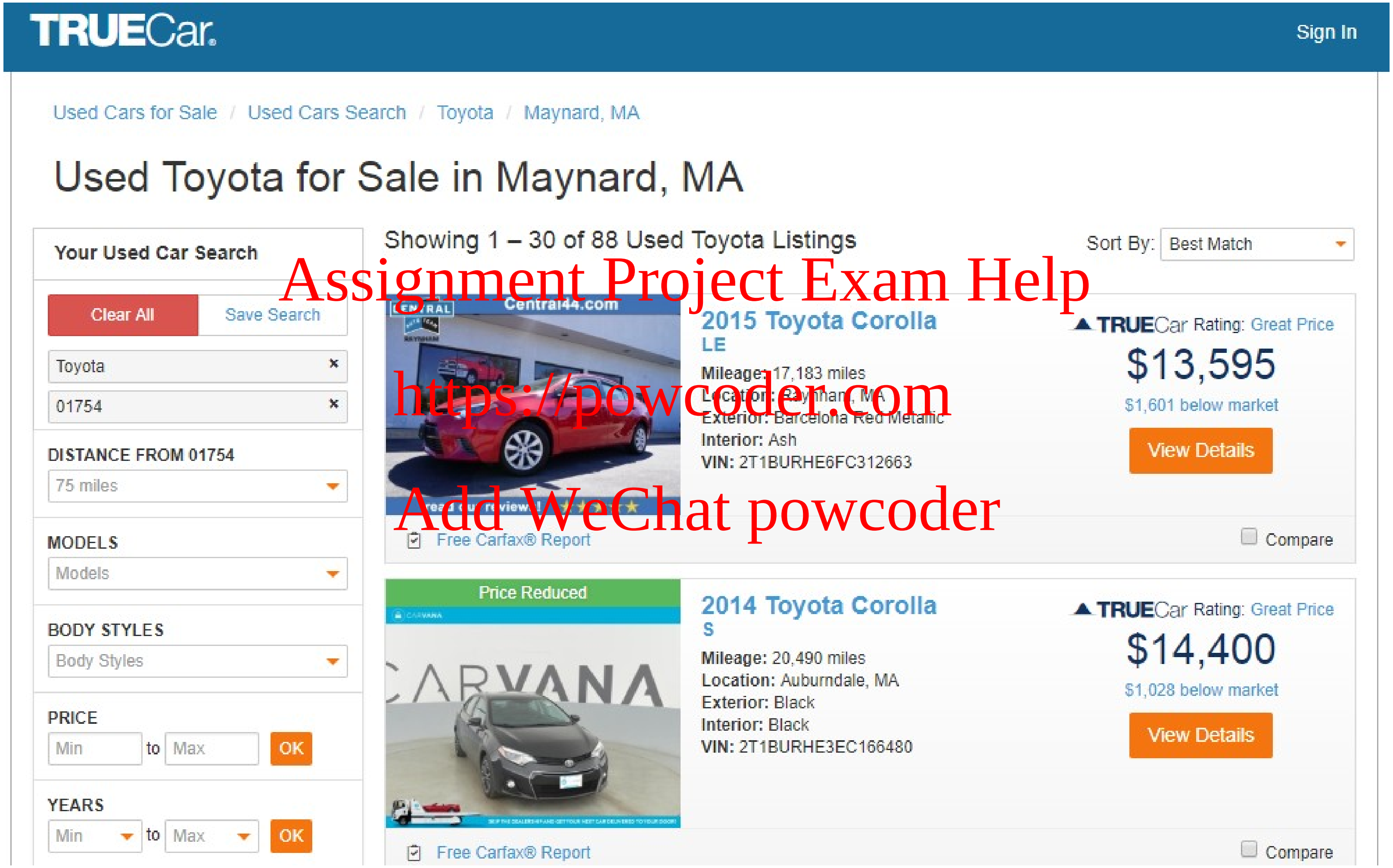
HARVARD UNIVERSITY

Agenda

Start	End	Item
		Regression in Action
		Partitioning Data
		Understanding Regression (simple univariate)
		Build & Evaluate a Multiple Linear Regression
		Housekeeping

HARVARD UNIVERSITY

How does truecar.com know the price is “great?”



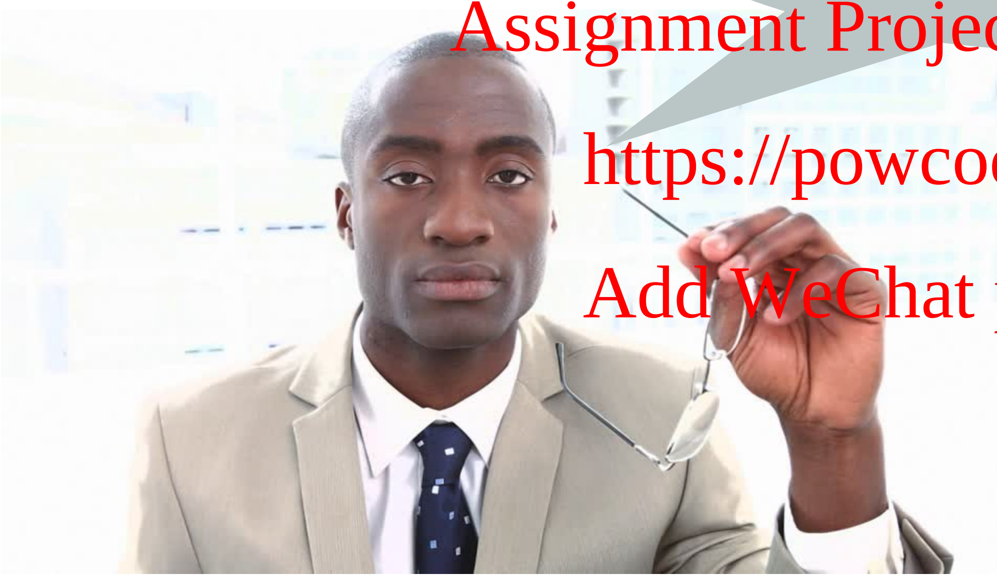

Yeah! I got a new job!

Welcome Dale to TrueCar's Competitor: OldCar

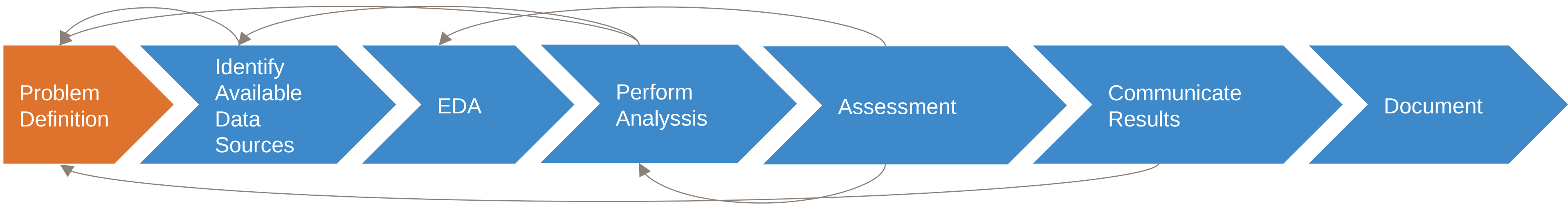
Dale, can you predict used car prices? Then we will know if a car is priced above/below expected value

Let’s help Dale again.

But of course. This sounds like a predictive modeling project.



Remember the Workflow?



1. **Problem Formulation >> Predict Toyota Prices.**
2. **Define data requirements >> Use ~1400 cars & car attributes**
3. **Explore the data >> in script**
4. **Perform Analysis & Create Project Artifacts >> fit a linear regression**
5. **Asses/Adjust the Project Artifacts >> Adj. R-Squared, P-Values etc.**
6. **Communicate Results >> examine the coefficients and readout Adj R^2**
7. **Document to make it repeatable >> Keep notes in script**

HARVARD UNIVERSITY

Let’s practice.

Open D_oldCar.R

Using the preprocessing and partitioning code in your toolbox, create a linear model of Toyota prices.

Price in Euros

Tires binary 0/1 are new tires on car

Age in months as of 8/04

KM (kilometers)

Fuel Type (diesel, petrol, CNG)

HP (horsepower)

Metallic color (1=yes, 0=no)

Automatic transmission (1=yes, 0=no)

CC (cylinder volume)

Doors

Quarterly_Tax (road tax)

Weight (in kg)

Regression is susceptible to **multi-colinearity**

Math

The presence of two or more predictor variables sharing the same linear relationship with the outcome variable.

In other words

Two+ informative features are measuring essentially the same thing.

Example

When predicting ice cream sales you include Fahrenheit and Celsius temperatures as two separate informative features.

Don't shock the cat by having multi-collinearity.

Regression is susceptible to **multi-colinearity**

What happens with multi-collinearity?

The effect is exaggerated e.g. double counted.

The good news

- The problem is so prevalent/impactful that R's linear regression function handles it.
- Other algorithms are not affected by double counting

Best Practice

Even though R removes it, do not rely on the function. Understanding the data you put in will avoid "garbage in, garbage out" scenarios.

Make happy cats by knowing your data

HARVARD UNIVERSITY

Back to R, script D!

The Summary of the Fit

```
> summary(fit)

Call:
lm(formula = Price ~ ., data = treatedTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-11141.3   -774.3    -19.9    763.0   6653.1 

Coefficients: (3 not defined because of singularities)
                          Estimate   Std. Error t value            Pr(>|t|)    
(Intercept)           -6476.090893  1363.877191  -4.748     0.0000023128693 ***
Weight_clean             18.866071     1.291646  14.606 < 0.0000000000000002 ***
Quarterly_Tax_clean      11.299234     1.895942   5.960     0.0000000336699 ***
Doors_clean             -69.703648    44.544760  -1.565               0.118    
CC_clean                 -0.037332     0.091886  -0.406               0.685    
Automatic_clean         185.378082   177.884776   1.042               0.298    
Met_Color_clean         106.772762    84.118619   1.269               0.205    
HP_clean                 26.793641     3.812888   7.027     0.0000000000036 ***
Fuel_Type_catP         2132.597511   358.370269   5.951     0.0000000354884 ***
Fuel_Type_catN            0.319050     0.211948   1.505               0.133    
Fuel_Type_catD                  NA           NA      NA                  NA    
KM_clean                 -0.017617     0.001488 -11.840 < 0.0000000000000002 ***
Age_08_04_clean        -123.703581     2.940716 -42.066 < 0.0000000000000002 ***
Fuel_Type_lev_x_Diesel          NA           NA      NA                  NA    
Fuel_Type_lev_x_Petrol          NA           NA      NA                  NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1328 on 1137 degrees of freedom
Multiple R-squared:  0.8701,    Adjusted R-squared:  0.8688 
F-statistic: 692.4 on 11 and 1137 DF,  p-value: < 0.00000000000000022
```


HARVARD UNIVERSITY

The Summary of the Fit

```
> summary(fit)

Call:
lm(formula = Price ~ ., data = treatedTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-11141.3   -774.3    -19.9    763.0   6653.1 

Coefficients: (3 not defined because of singularities)
                       Estimate   Std. Error t value            Pr(>|t|)    
(Intercept)        -6476.090893  1363.877191  -4.748     0.00000231286938 ***
Weight_clean          18.866071     1.291646  14.606 < 0.0000000000000002 ***
Quarterly_Tax_clean   11.299334     1.895942   5.960     0.00000000336699 ***
Doors_clean          -69.703648    44.544760  -1.565                0.118    
CC_clean              -0.037332     0.091886  -0.406                0.685    
Automatic_clean      185.378082   177.884776   1.042                0.298    
Met_Color_clean      106.772762    84.113619   1.269                0.205    
HP_clean              26.796611     3.812888   7.027     0.00000000000363 ***
Fuel_Type_catP      2132.597511   358.370269   5.951     0.00000000354884 ***
Fuel_Type_catN         0.319050     0.211948   1.505                0.133    
Fuel_Type_catD               NA           NA      NA                   NA    
KM_clean              -0.017617     0.001488 -11.840 < 0.0000000000000002 ***
Age_08_04_clean     -123.703581     2.940716 -42.066 < 0.0000000000000002 ***
Fuel_Type_lev_x_Diesel       NA           NA      NA                   NA    
Fuel_Type_lev_x_Petrol       NA           NA      NA                   NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1328 on 1137 degrees of freedom
Multiple R-squared:  0.8701,	Adjusted R-squared:  0.8688 
F-statistic: 692.4 on 11 and 1137 DF,  p-value: < 0.00000000000000022
```

Auto ID of multi colinearity

HARVARD UNIVERSITY

The Summary of the Fit

```
> summary(fit)

Call:
lm(formula = Price ~ ., data = treatedTrain)

Residuals:
     Min       1Q   Median       3Q      Max
-11141.3   -774.3    -19.9    763.0   6653.1

Coefficients: (3 not defined because of singularities)
                         Estimate   Std. Error t value               Pr(>|t|)
(Intercept)          -6476.090893  1363.877191  -4.748      0.00000231286938 ***
Weight_clean            18.866071     1.291646  14.606 < 0.0000000000000002 ***
Quarterly_Tax_clean     11.[illegible]09234     1.[illegible]95[illegible]4[illegible]   5.960      0.00000000336699 ***
Doors_clean            -69.703648    44.544760  -1.565                 0.118
CC_clean                -0.037332     0.091886  -0.406                 0.685
Automatic_clean        185.378082   177.884776   1.042                 0.298
Met_Color_clean        106.772762    84.113619   1.269                 0.205
HP_clean                26.793641     3.812888   7.02[illegible]      0.00000000000363 ***
Fuel_Type_catP        2132.597511   358.370269   5.951      0.00000000354884 ***
Fuel_Type_catN           0.319050     0.211948   1.505                 0.133
Fuel_Type_catD                 NA           NA      NA                    NA
KM_clean                -0.017617     0.001488 -11.840 < 0.0000000000000002 ***
Age_08_04_clean       -123.703581     2.940716 -42.066 < 0.0000000000000002 ***
Fuel_Type_lev_x_Diesel         NA           NA      NA                    NA
Fuel_Type_lev_x_Petrol         NA           NA      NA                    NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1328 on 1137 degrees of freedom
Multiple R-squared:  0.8701,	Adjusted R-squared:  0.8688
F-statistic: 692.4 on 11 and 1137 DF,  p-value: < 0.00000000000000022
```

Another name for errors.
Summary stats for the errors.

HARVARD UNIVERSITY

The Summary of the Fit

```
> summary(fit)

Call:
lm(formula = Price ~ ., data = treatedTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-11141.3   -774.3    -19.9    763.0   6653.1 

Coefficients: (3 not defined because of singularities)
                          Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)           -6476.090893 1363.877191  -4.748     0.00000231286938 ***
Weight_clean             18.866071    1.291646  14.606 < 0.0000000000000002 ***
Quarterly_Tax_clean      11.299234    1.895942   5.960     0.00000000336699 ***
Doors_clean             -69.703648   44.514760  -1.565                0.118    
CC_clean                 -0.037332    0.091886  -0.406                0.685    
Automatic_clean         185.378082  177.884776   1.042                0.298    
Met_Color_clean         106.772762   84.113619   1.269                0.205    
HP_clean                 26.793641    3.812888   7.027     0.00000000000363 ***
Fuel_Type_catP         2132.597511  358.370269   5.951     0.00000000354884 ***
Fuel_Type_catN            0.319050    0.211948   1.505                0.133    
Fuel_Type_catD                  NA          NA      NA                   NA    
KM_clean                 -0.017617    0.001488 -11.840 < 0.0000000000000002 ***
Age_08_04_clean        -123.703581    2.940716 -42.066 < 0.0000000000000002 ***
Fuel_Type_lev_x_Diesel          NA          NA      NA                   NA    
Fuel_Type_lev_x_Petrol          NA          NA      NA                   NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1328 on 1137 degrees of freedom
Multiple R-squared:  0.8701,    Adjusted R-squared:  0.8688 
F-statistic: 692.4 on 11 and 1137 DF,  p-value: < 0.00000000000000022
```

Treated Variable Names
i.e. informative features

HARVARD UNIVERSITY

The Summary of the Fit

```
> summary(fit)

Call:
lm(formula = Price ~ ., data = treatedTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-11141.3   -774.3    -19.9    763.0   6653.1 

Coefficients: (3 not defined because of singularities)
                        Estimate  Std. Error t value            Pr(>|t|)    
(Intercept)         -6476.090893 1363.877191  -4.748    0.00000231286938 ***
Weight_clean           18.866071    1.291646  14.606 < 0.0000000000000002 ***
Quarterly_Tax_clean    11.299234    1.895942   5.960    0.00000000336699 ***
Doors_clean           -69.703648   44.544760  -1.565               0.118    
CC_clean               -0.037332    0.091836  -0.406               0.685    
Automatic_clean       185.378082  177.884776   1.042               0.298    
Met_Color_clean       106.772762   84.113619   1.269               0.205    
HP_clean               26.793641    3.812888   7.027    0.00000000000363 ***
Fuel_Type_catP       2112.597511  358.170369   5.851    0.00000000354884 ***
Fuel_Type_catN          0.319050    0.211948   1.505               0.133    
Fuel_Type_catD                NA          NA      NA                  NA    
KM_clean               -0.017617    0.001488 -11.840 < 0.0000000000000002 ***
Age_08_04_clean      -123.703581    2.940716 -42.066 < 0.0000000000000002 ***
Fuel_Type_lev_x_Diesel        NA          NA      NA                  NA    
Fuel_Type_lev_x_Petrol        NA          NA      NA                  NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1328 on 1137 degrees of freedom
Multiple R-squared:  0.8701,	Adjusted R-squared:  0.8688 
F-statistic: 692.4 on 11 and 1137 DF,  p-value: < 0.00000000000000022
```

Coefficients or Beta values

HARVARD UNIVERSITY

The Summary of the Fit

```
> summary(fit)

Call:
lm(formula = Price ~ ., data = treatedTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-11141.3   -774.3    -19.9    763.0   6653.1 

Coefficients: (3 not defined because of singularities)
                         Estimate  Std. Error t value            Pr(>|t|)    
(Intercept)          -6476.090893 1363.877191  -4.748    0.00000231286938 ***
Weight_clean            18.866071    1.291646  14.606 < 0.0000000000000002 ***
Quarterly_Tax_clean     11.299234    1.895942   5.960    0.00000000336699 ***
Doors_clean            -69.703648   44.544760  -1.565               0.118    
CC_clean                -0.037332    0.091836  -0.406               0.685    
Automatic_clean        185.378082  177.884776   1.042               0.298    
Met_Color_clean        106.772762   84.113619   1.269               0.205    
HP_clean                26.793641    3.812888   7.027    0.00000000000363 ***
Fuel_Type_catP        2132.597511  358.370269   5.951    0.00000000354884 ***
Fuel_Type_catN           0.319050    0.211948   1.505               0.133    
Fuel_Type_catD                 NA          NA      NA                  NA    
KM_clean                -0.017617    0.001488 -11.840 < 0.0000000000000002 ***
Age_08_04_clean       -123.703581    2.940716 -42.066 < 0.0000000000000002 ***
Fuel_Type_lev_x_Diesel         NA          NA      NA                  NA    
Fuel_Type_lev_x_Petrol         NA          NA      NA                  NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1328 on 1137 degrees of freedom
Multiple R-squared:  0.8701,	Adjusted R-squared:  0.8688 
F-statistic: 692.4 on 11 and 1137 DF,  p-value: < 0.00000000000000022
```

The average amount that the coefficients vary from the actual average value of our response variable.

HARVARD UNIVERSITY

The Summary of the Fit

```
> summary(fit)

Call:
lm(formula = Price ~ ., data = treatedTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-11141.3   -774.3    -19.9    763.0   6653.1 

Coefficients: (3 not defined because of singularities)
                         Estimate   Std. Error t value            Pr(>|t|)    
(Intercept)          -6476.090893  1363.877191  -4.748    0.00000231286938 ***
Weight_clean            18.866071     1.291646  14.606 < 0.0000000000000002 ***
Quarterly_Tax_clean     11.299234     1.895942   5.960    0.00000000336699 ***
Doors_clean            -69.703648    44.544760  -1.565               0.118    
CC_clean                -0.037332     0.091886  -0.406               0.685    
Automatic_clean        185.378082   177.884776   1.042               0.298    
Met_Color_clean        106.772762    84.113619   1.269               0.205    
HP_clean                26.793641     3.812888   7.027    0.00000000000363 ***
Fuel_Type_catP        2112.597511   358.170269   5.851    0.00000000354884 ***
Fuel_Type_catN           0.319050     0.211948   1.505               0.133    
Fuel_Type_catD                 NA           NA      NA                  NA    
KM_clean                -0.017617     0.001488 -11.840 < 0.0000000000000002 ***
Age_08_04_clean       -123.703581     2.940716 -42.066 < 0.0000000000000002 ***
Fuel_Type_lev_x_Diesel         NA           NA      NA                  NA    
Fuel_Type_lev_x_Petrol         NA           NA      NA                  NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1328 on 1137 degrees of freedom
Multiple R-squared:  0.8701,	Adjusted R-squared:  0.8688 
F-statistic: 692.4 on 11 and 1137 DF,  p-value: < 0.00000000000000022
```

The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. In another way, values away from 0 indicate a real relationship

HARVARD UNIVERSITY

The Summary of the Fit

```
> summary(fit)

Call:
lm(formula = Price ~ ., data = treatedTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-11141.3   -774.3    -19.9    763.0   6653.1 

Coefficients: (3 not defined because of singularities)
                        Estimate   Std. Error t value         Pr(>|t|)    
(Intercept)         -6476.090893  1363.877191  -4.748    0.00000231286938 ***
Weight_clean           18.866071     1.291646  14.606 < 0.0000000000000002 ***
Quarterly_Tax_clean    11.299234     1.895942   5.960    0.00000000336699 ***
Doors_clean           -69.703648    44.544760  -1.565               0.118    
CC_clean               -0.037332     0.091886  -0.406               0.685    
Automatic_clean       185.378082   177.884776   1.042               0.298    
Met_Color_clean       106.772762    84.113619   1.269               0.205    
HP_clean               26.793641     3.812888   7.027    0.00000000000363 ***
Fuel_Type_catP       2112.597511   358.170369   5.851    0.00000000354884 ***
Fuel_Type_catN          0.319050     0.211948   1.505               0.133    
Fuel_Type_catD                NA           NA      NA                  NA    
KM_clean               -0.017617     0.001488 -11.840 < 0.0000000000000002 ***
Age_08_04_clean      -123.703581     2.940716 -42.066 < 0.0000000000000002 ***
Fuel_Type_lev_x_Diesel        NA           NA      NA                  NA    
Fuel_Type_lev_x_Petrol        NA           NA      NA                  NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1328 on 1137 degrees of freedom
Multiple R-squared:  0.8701,	Adjusted R-squared:  0.8688 
F-statistic: 692.4 on 11 and 1137 DF,  p-value: < 0.00000000000000022
```

P-Values
The probability of seeing a value larger than the t value.
Small p-values mean its less likely due to chance.

HARVARD UNIVERSITY

The Summary of the Fit

```
> summary(fit)

Call:
lm(formula = Price ~ ., data = treatedTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-11141.3   -774.3    -19.9    763.0   6653.1 

Coefficients: (3 not defined because of singularities)
                          Estimate   Std. Error t value           Pr(>|t|)    
(Intercept)           -6476.090893  1363.877191  -4.748   0.00000231286938 ***
Weight_clean             18.866071     1.291646  14.606 < 0.0000000000000002 ***
Quarterly_Tax_clean      11.299234     1.895942   5.960   0.00000000336699 ***
Doors_clean             -69.703648    44.544760  -1.565              0.118    
CC_clean                 -0.037132     0.091886  -0.406              0.685    
Automatic_clean         185.378082   177.884776   1.042              0.298    
Met_Color_clean         106.772762    84.113619   1.269              0.205    
HP_clean                 26.793641     3.812888   7.027   0.00000000000363 ***
Fuel_Type_catP         2112.597511   358.370369   5.851   0.00000000354884 ***
Fuel_Type_catN            0.319050     0.211948   1.505              0.133    
Fuel_Type_catD                  NA           NA      NA                 NA    
KM_clean                 -0.017617     0.001488 -11.840 < 0.0000000000000002 ***
Age_08_04_clean        -123.703581     2.940716 -42.066 < 0.0000000000000002 ***
Fuel_Type_lev_x_Diesel          NA           NA      NA                 NA    
Fuel_Type_lev_x_Petrol          NA           NA      NA                 NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1328 on 1137 degrees of freedom
Multiple R-squared:  0.8701,	Adjusted R-squared:  0.8688 
F-statistic: 692.4 on 11 and 1137 DF,  p-value: < 0.00000000000000022
```

P-Values
In stats p <0.05 is good but in business I have seen p<0.2.
It's a good idea to rebuild a model without variables that do not meet the cutoff.

HARVARD UNIVERSITY

The Summary of the Fit

```
> summary(fit)

Call:
lm(formula = Price ~ ., data = treatedTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-11141.3   -774.3    -19.9    763.0   6653.1 

Coefficients: (3 not defined because of singularities)
                          Estimate   Std. Error t value            Pr(>|t|)    
(Intercept)           -6476.090893  1363.877191  -4.748      0.000002312869 ***
Weight_clean             18.866071     1.291646  14.606 < 0.0000000000000002 ***
Quarterly_Tax_clean      11.299234     1.895942   5.960      0.000000003366 ***
Doors_clean             -69.703648    44.544760  -1.565               0.118    
CC_clean                 -0.037132     0.091886  -0.406               0.685    
Automatic_clean         185.378082   177.884776   1.042               0.298    
Met_Color_clean         106.772762    84.113619   1.269               0.205    
HP_clean                 26.793641     3.812888   7.027      0.000000000003 ***
Fuel_Type_catP         2112.597511   358.170269   5.851      0.000000006354 ***
Fuel_Type_catN            0.319050     0.211948   1.505               0.133    
Fuel_Type_catD                  NA           NA      NA                  NA    
KM_clean                 -0.017617     0.001488 -11.840 < 0.0000000000000002 ***
Age_08_04_clean        -123.703581     2.940716 -42.066 < 0.0000000000000002 ***
Fuel_Type_lev_x_Diesel          NA           NA      NA                  NA    
Fuel_Type_lev_x_Petrol          NA           NA      NA                  NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1328 on 1137 degrees of freedom
Multiple R-squared:  0.8701,	Adjusted R-squared:  0.8688 
F-statistic: 692.4 on 11 and 1137 DF,  p-value: < 0.00000000000000022
```


P-Values
We could drop ??? Based on p values?

What variables should we drop with a p-value >=0.2?

HARVARD UNIVERSITY

Back to the script D

Parsimonious Model

```
> summary(fit2)

Call:
lm(formula = Price ~ ., data = treatedTrainParsimony)

Residuals:
     Min       1Q   Median       3Q      Max 
-11231.1   -766.9    -24.4    769.7   6665.8 

Coefficients: (3 not defined because of singularities)
                          Estimate  Std. Error t value            Pr(>|t|)    
(Intercept)           -6652.426248 1337.407929  -4.974    0.00000075614045 ***
Weight_clean             19.086315    1.263358  15.108 < 0.0000000000000002 ***
Quarterly_Tax_clean      11.289780    1.895568   5.956    0.00000000344140 ***
Doors_clean             -71.134917   44.179384  -1.610               0.108    
HP_clean                 26.230472    3.716559   7.058    0.00000000000293 ***
Fuel_Type_catP         2169.457810  354.025248   6.128    0.00000000122404 ***
Fuel_Type_catN            0.282825    0.208225   1.358               0.175    
Fuel_Type_catD                  NA          NA      NA                  NA    
KM_clean                 -0.017808    0.001482 -12.014 < 0.0000000000000002 ***
Age_08_04_clean        -123.620949    2.908408 -42.470 < 0.0000000000000002 ***
Fuel_Type_lev_x_Diesel          NA          NA      NA                  NA    
Fuel_Type_lev_x_Petrol          NA          NA      NA                  NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1328 on 1140 degrees of freedom
Multiple R-squared:  0.8698,	Adjusted R-squared:  0.8689 
F-statistic:   952 on 8 and 1140 DF,  p-value: < 0.00000000000000022
```


Parsimony or compactness is desirable in models. The more features in a model, the more complexity we introduce, data integrity, data interactions, time to score and time to predict are all impacted.

HARVARD UNIVERSITY

Parsimonious Model

```
> summary(fit2)

Call:
lm(formula = Price ~ ., data = treatedTrainParsimony)

Residuals:
     Min       1Q   Median       3Q      Max 
-11231.1   -766.9    -24.4    769.7   6665.8 

Coefficients: (3 not defined because of singularities)
                          Estimate   Std. Error t value           Pr(>|t|)    
(Intercept)           -6652.426248  1337.407929  -4.974   0.00000075614045 ***
Weight_clean             19.086315     1.263358  15.108 < 0.0000000000000002 ***
Quarterly_Tax_clean      11.289780     1.895568   5.956    0.0000000034140 ***
Doors_clean             -71.134917    44.179384  -1.610              0.108    
HP_clean                 36.230472     8.716559   4.058    0.0000000000293 ***
Fuel_Type_catP         2169.457810   354.025248   6.128    0.0000000122404 ***
Fuel_Type_catN            0.282825     0.208225   1.358              0.175    
Fuel_Type_catD                  NA           NA      NA                 NA    
KM_clean                 -0.017808     0.001482 -12.014 < 0.0000000000000002 ***
Age_08_04_clean        -123.520949     2.908408 -42.470 < 0.0000000000000002 ***
Fuel_Type_lev_x_Diesel          NA           NA      NA                 NA    
Fuel_Type_lev_x_Petrol          NA           NA      NA                 NA    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1328 on 1140 degrees of freedom
Multiple R-squared:  0.8698,	Adjusted R-squared:  0.8689 
F-statistic:   952 on 8 and 1140 DF,  p-value: < 0.00000000000000022
```

R-Sq: how much of the variation are the model is fitting. R-Sq measures the linear relationship between Price & features It always lies between 0 and 1

Compare the two models

fit1

```
Residual standard error: 1328 on 1137 degrees of freedom
Multiple R-squared:  0.8701,    Adjusted R-squared:  0.8688
F-statistic: 692.4 on 11 and 1137 DF,  p-value: < 0.0000000000000022
```


fit2

```
Residual standard error: 1328 on 1140 degrees of freedom
Multiple R-squared:  0.8698,    Adjusted R-squared:  0.8689
F-statistic:   952 on 8 and 1140 DF,  p-value: < 0.0000000000000022
```

It can be said that both models explain ~87% of the variation in car prices. Dropping the variables improved accuracy and reinforces the fact that the variables didn't add value.

Evaluating a Prediction Model

RMSE- Root Mean Squared Error

MAPE- Mean Absolute Percentage Error

y	y-hat
Actual Values	Predicted/Forecasted
10	6
12	8
20	18
36	20

Besides P-Values which is a variable level KPI, and adjusted R-Sq there are two popular KPI for evaluating continuous model predictions.

RMSE

RMSE- Root Mean Squared Error

y	y-hat	ERROR	ERROR-SQ
Actual Values	**Predicted/Forecasted**	**Errors**	**Squared Errors**
10	16	-6	36
12	8	4	16
20	17	3	9
36	34	2	4

Mean

$$\frac{36+16+9+4}{4}$$

To manually calculate RMSE, work the acronym backwards.

HARVARD UNIVERSITY

RMSE

RMSE- Root Mean Squared Error

Actual Values	Predicted/Forecasted	Errors	Squared Errors
10	16	-6	36
12	8	4	16
20	17	3	9
36	34	2	4

Mean

$$\frac{36+16+9+4}{4}$$

Square Root

=4.03

In the same units being measured, tells you +/- the prediction error.

HARVARD UNIVERSITY

MAPE

MAPE- Mean Absolute Percentage Error

Actual Values	Predicted/Forecasted	Errors	Absolute	As % of Forecast
10	16	-6	6	=6/16 or **37%**
12	8	4	4	=4/8 or **50%**
20	17	3	3	=3/17 or **17%**
36	34	2	2	=2/34 or **5%**

Mean of Percentages

$$\frac{37\%+50\%+17\%+5\%}{4}$$

=27.7%

Instead of squaring error, take the absolute error. Then divide that by the **forecast** value. Lastly calculate a mean average of all the percentage errors.

Back to script D

How good is Dale's model?

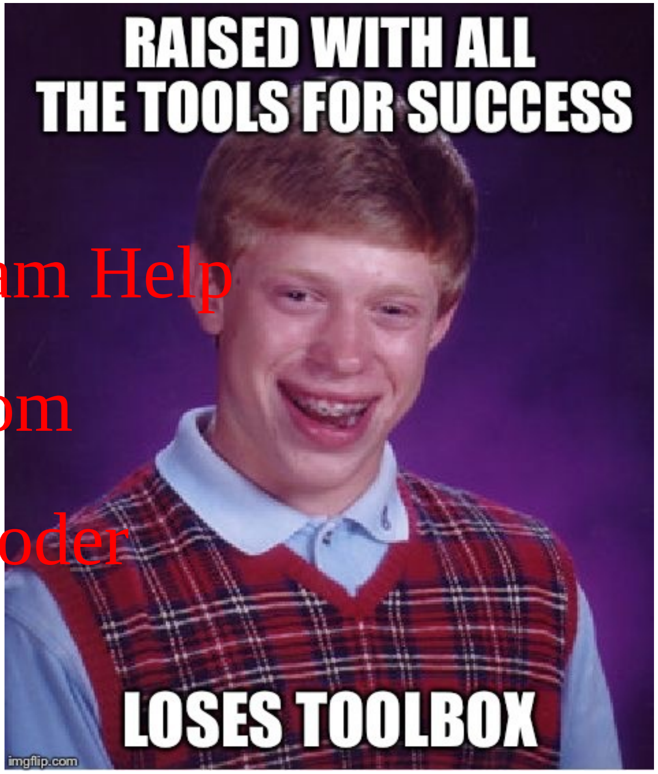
Your Data Mining Toolbox

<u>Previous Lessons</u>

- Some R Programming
- Knowledge of Data Preparation
- Exploratory Data Analysis
- Basic Visualization

<u>After today</u>

- You can predict continuous business outcomes simplistically



Regression is an initial starting algorithm. It puts you on a path to more complex machine learning but more importantly you can start to frame business problems in terms algorithms can understand.

Housekeeping , Reading & Homework

- Next Week is Logistic Regression, KNN

- Chapter 7
- Chapter 10

- Homework – check syllabus!

HARVARD UNIVERSITY