Introduction to Machine Learning. Lec.4 Multiple Linear Regression

Aidos Sarsembayev, IITU, 2018

Regression is…

- a technique for determining the statistical **relationship between** two or more variables where a change in a dependent variable is associated with, and depends on, a change in one or more independent variables.

http://www.businessdictionary.com/definition/regression.html

Types of regression models

- Simple Linear Regression
- **Multiple Linear Regression**
- Polynomial Regression
- Support Vector Regression (SVR)
- Decision Tree Regression
- Random Forest Regression

SLR. Formula

Constant

Coefficient

$$y = b_0 + b_1 * x_1$$

Dependent
variable (DV)

Independent
variable (IV)

MLR. Formula

$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + b_3{*}x_3 + \ldots + b_n{*}x_n$$

MLR. Formula

Constant

Coefficients

$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + b_3{*}x_3 + \ldots + b_n{*}x_n$$

Dependent variable (DV)

Independent variables (IVs)

A caveat. Assumptions of LR

- Linearity
- Homoscedasticity
- Multivariate normality
- Independence of errors
- Lack of multicollinearity

An assumption that all the errors of ind.vars are similar to each other

Significance level

Dummy variable trap

A caveat. Assumptions of LR

- Linearity

Dummy variables

Profit	Administration	Marketing Spend	R&D Spend	State
192261.83	136897.8	471784.1	165349.2	New York
191792.06	151377.59	443898.53	162597.7	California
191050.39	101145.55	407934.54	153441.51	New York
182901.99	118671.85	383199.62	144372.41	New York
166187.94	91391.77	366168.42	142107.34	California
156991.12	99814.71	362861.36	131876.9	New York
156122.51	147198.87	127716.82	134615.46	California

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

Dummy variables

Profit	Administration	Marketing Spend	R&D Spend	State
192261.83	136897.8	471784.1	165349.2	New York
191792.06	151377.59	443898.53	162597.7	California
191050.39	101145.55	407934.54	153441.51	New York
182901.99	118671.85	383199.62	144372.41	New York
166187.94	91391.77	366168.42	142107.34	California
156991.12	99814.71	362861.36	131876.9	New York
156122.51	147198.87	127716.82	134615.46	California

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

Categorical variables cannot be fit in the equations

Dummy variables > Label encoding

Profit	Administration	Marketing Spend	R&D Spend	State
192261.83	136897.8	471784.1	165349.2	New York
191792.06	151377.59	443898.53	162597.7	California
191050.39	101145.55	407934.54	153441.51	New York
182901.99	118671.85	383199.62	144372.41	New York
166187.94	91391.77	366168.42	142107.34	California
156991.12	99814.71	362861.36	131876.9	New York
156122.51	147198.87	127716.82	134615.46	California

New York	California
1	0
0	1
1	0
1	0
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

Dummy variables > Label encoding

Profit	R&D Spend	Administration	R&D Spend	State
192261.83	165349.2	136897.8	165349.2	~~New York~~
191792.06	162597.7	151377.59	162597.7	~~California~~
191050.39	153441.51	101145.55	153441.51	~~New York~~
182901.99	144372.41	118671.85	144372.41	~~New York~~
166187.94	142107.34	91391.77	142107.34	~~California~~
156991.12	131876.9	99814.71	131876.9	~~New York~~
156122.51	134615.46	147198.87	134615.46	~~California~~

New York	California
1	0
0	1
1	0
1	0
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ???$$

Dummy variables > Label encoding

Profit	R&D Spend	Administration	R&D Spend	State
192261.83	165349.2	136897.8	165349.2	New York
191792.06	162597.7	151377.59	162597.7	California
191050.39	153441.51	101145.55	153441.51	New York
182901.99	144372.41	118671.85	144372.41	New York
166187.94	142107.34	91391.77	142107.34	California
156991.12	131876.9	99814.71	131876.9	New York
156122.51	134615.46	147198.87	134615.46	California

New York	California
1	0
0	1
1	0
1	0
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \ldots\ldots + b_4 * D1$$

D1 is a dummy variable for New York

Dummy variables > Label encoding

Profit	R&D Spend	Administration	R&D Spend	State
192261.83	165349.2	136897.8	165349.2	New York
191792.06	162597.7	151377.59	[illegible]	[illegible]
191050.39	153441.51	101145.55	[illegible]	[illegible]
182901.99	144372.41	118671.85	[illegible]	[illegible]
166187.94	142107.34	91391.77	[illegible]	[illegible]
156991.12	131876.9	99814.71	131876.9	New York
156122.51	134615.46	147198.87	134615.46	California

New York	California
1	0
0	1
1	0
1	0
0	1
1	0
0	1

$$y=b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + \dots \quad + b_4*D1$$

This is like a light switch

Dummy variables > Label encoding

Profit	R&D Spend	Administration	R&D Spend	~~State~~
192261.83	165349.2	136897.8	165349.2	~~New York~~
191792.06	162597.7	151377.59	162597.7	~~California~~
191050.39	153441.51	101145.55	153441.51	~~New York~~
182901.99	144372.41	118671.85	144372.41	~~New York~~
166187.94	142107.34	91391.77	142107.34	~~California~~
156991.12	131876.9	99814.71	131876.9	~~New York~~
156122.51	134615.46	147198.87	134615.46	~~California~~

New York	~~California~~
1	~~0~~
0	~~1~~
1	~~0~~
1	~~0~~
0	~~1~~
1	~~0~~
0	~~1~~

$$y = b_0 + \quad b_1 * x_1 \quad + \quad b_2 * x_2 \quad + b_3 * x_3 + \ \ldots\ldots \qquad + b_4 * D1$$

Aren't we suppressing the 'California' variable in this case?
No, because of b_0

Dummy variables > Label encoding

Profit	R&D Spend	Administration	R&D Spend	State
192261.83	165349.2	136897.8	165349.2	New York
191792.06	162597.7	151377.59	162597.7	California
191050.39	153441.51	101145.55	153441.51	New York
182901.99	144372.41	118671.85	144372.41	New York
166187.94	142107.34	91391.77	142107.34	California
156991.12	131876.9	99814.71	131876.9	New York
156122.51	134615.46	147198.87	134615.46	California

New York	California
1	0
0	1
1	0
1	0
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \ldots\ldots \quad + b_4 * D_1$$

Why should we bother about getting rid of one of the dummy variables?

Dummy variables > Label encoding

Profit	R&D Spend	Administration	R&D Spend	State
192261.83	165349.2	136897.8	165349.2	New York
191792.06	162597.7	151377.59	162597.7	California
191050.39	153441.51	101145.55	153441.51	New York
182901.99	144372.41	118671.85	144372.41	New York
166187.94	142107.34	91391.77	142107.34	California
156991.12	131876.9	99814.71	131876.9	New York
156122.51	134615.46	147198.87	134615.46	California

New York	California
1	0
0	1
1	0
1	0
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \ldots\ldots \quad + b_4 * D_1$$

Because of the so called 'dummy variable trap'

Dummy variables trap

Profit	R&D Spend	Administration	R&D Spend	~~State~~
192261.83	165349.2	136897.8	165349.2	~~New York~~
191792.06	162597.7	151377.59	162597.7	~~California~~
191050.39	153441.51	101145.55	153441.51	~~New York~~
182901.99	144372.41	118671.85	144372.41	~~New York~~
166187.94	142107.34	91391.77	142107.34	~~California~~
156991.12	131876.9	99814.71	131876.9	~~New York~~
156122.51	134615.46	147198.87	134615.46	~~California~~

New York	California
1	0
0	1
1	0
1	0
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \ldots\ldots + b_4 * D1 + b_5 * D2$$

We don’t need both, because they are simply duplicating each other ->>>

Dummy variables trap

Profit	R&D Spend	Administration	R&D Spend	State
192261.83	165349.2	136897.8	165349.2	~~New York~~
191792.06	162597.7	151377.59	162597.7	~~California~~
191050.39	153441.51	101145.55	153441.51	~~New York~~
182901.99	144372.41	118671.85	144372.41	~~New York~~
166187.94	142107.34	91391.77	142107.34	~~California~~
156991.12	131876.9	99814.71	131876.9	~~New York~~
156122.51	134615.46	147198.87	134615.46	~~California~~

New York	California
1	0
0	1
1	0
1	0
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \ldots\ldots + b_4 * D1 + b_5 * D2$$

D2 = 1 – D1 . The phenomenon when one or several independent variables in a linear regression predict each other is called - multicollinearity

Dummy variables trap

Profit	R&D Spend	Administration	R&D Spend	State
192261.83	165349.2	136897.8	165349.2	New York
191792.06	162597.7	151377.59	162597.7	California
191050.39	153441.51	101145.55	153441.51	New York
182901.99	144372.41	118671.85	144372.41	New York
166187.94	142107.34	91391.77	142107.34	California
156991.12	131876.9	99814.71	131876.9	New York
156122.51	134615.46	147198.87	134615.46	California

New York	California
1	0
0	1
1	0
1	0
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \ldots\ldots + b_4 * D1 + b_5 * D2$$

As the consequence of this phenomenon the model cannot distinguish between the effects of D1 from the effects of D2

Dummy variables trap

Profit	R&D Spend	Administration	R&D Spend	State
192261.83	165349.2	136897.8	165349.2	~~New York~~
19179[illegible]06	162597.7	151377.59	162597.7	~~California~~
19105[illegible]39	153441.51	101145.55	153441.51	~~New York~~
18290[illegible]99	144372.41	118671.85	144372.41	~~New York~~
16618[illegible]94	142107.34	91391.77	142107.34	~~California~~
15699[illegible]12	131876.9	99814.71	131876.9	~~New York~~
15612[illegible]51	134615.46	147198.87	134615.46	~~California~~

New York	California
1	0

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \ldots\ldots + b_4 * D1 + b_5 * D2$$

The reality is that you cannot have all of these three elements in your model

Dummy variables trap

Profit	R&D Spend	Administration	R&D Spend	State
192261.83	165349.2	136897.8	165349.2	~~New York~~
191792.06	162597.7	151377.59	162597.7	~~California~~
191050.39	153441.51	101145.55	153441.51	~~New York~~
182901.99	144372.41	118671.85	144372.41	~~New York~~
166187.94	142107.34	91391.77	142107.34	~~California~~
156991.12	131876.9	99814.71	131876.9	~~New York~~
156122.51	134615.46	147198.87	134615.46	~~California~~

New York	California
1	0
0	1
1	0
1	0
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \ldots\ldots + b_4 * D1 + \cancel{b_5 * D2}$$

You have to drop one of the dummy variables.
If you have 10 (or 95), leave only 9 (or 94) etc.

Dummy variables trap

Profit	R&D Spend	Administration	R&D Spend	~~State~~
192261.83	165349.2	136897.8	165349.2	~~New York~~
191792.06	162597.7	151377.59	162597.7	~~California~~
191050.39	153441.51	101145.55	153441.51	~~New York~~
182901.99	144372.41	118671.85	144372.41	~~New York~~
166187.94	142107.34	91391.77	142107.34	~~California~~
156991.12	131876.9	99814.71	131876.9	~~New York~~
156122.51	134615.46	147198.87	134615.46	~~California~~

New York	California
1	0
0	1
1	0
1	0
0	1
1	0
0	1

$$y = b_0 + b_1 {*} x_1 + b_2 {*} x_2 + b_3 {*} x_3 + \ldots\ldots \quad + b_4 {*} D1 + b_5 {*} D2$$

What if we would have not one, but two or more sets of dummy variables?

Dummy variables trap

Profit	R&D Spend	Administration	R&D Spend	State
192261.83	165349.2	136897.8	165349.2	~~New York~~
191792.06	162597.7	151377.59	162597.7	~~California~~
191050.39	153441.51	101145.55	153441.51	~~New York~~
182901.99	144372.41	118671.85	144372.41	~~New York~~
166187.94	142107.34	91391.77	142107.34	~~California~~
156991.12	131876.9	99814.71	131876.9	~~New York~~
156122.51	134615.46	147198.87	134615.46	~~California~~

New York	California
1	0
0	1
1	0
1	0
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \ldots\ldots + b_4 * D1 + b_5 * D2$$

We should repeat the same for the other sets of dummy variables as well

Model optimization

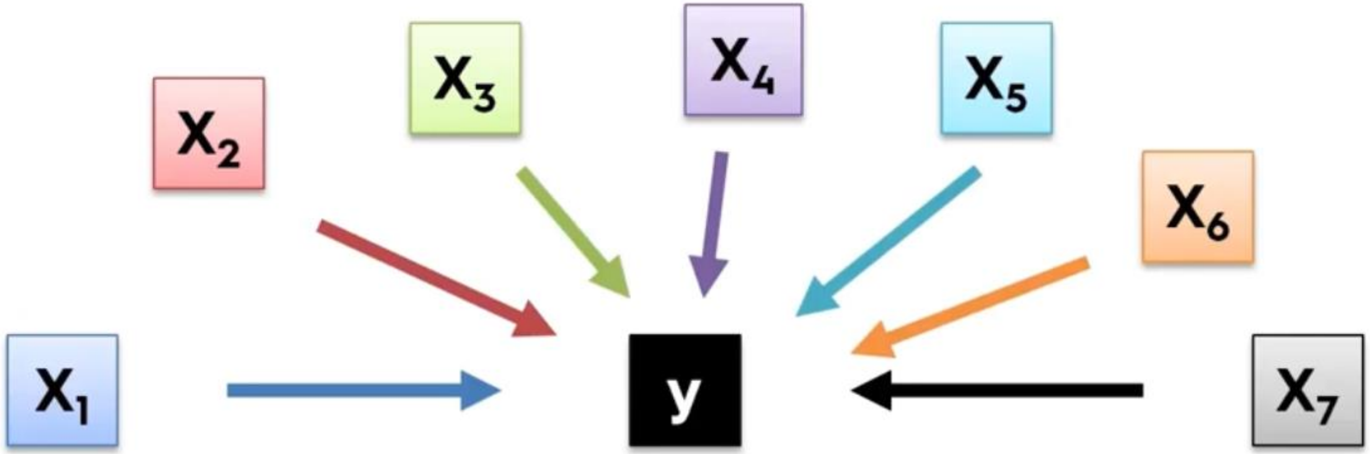
Among non-categorical variables not every independent variable is necessary for the optimal model building

We have to optimize the model by getting rid of the insignificant variables

Model optimization

Among non-categorical variables not every independent variable is necessary for the optimal model building

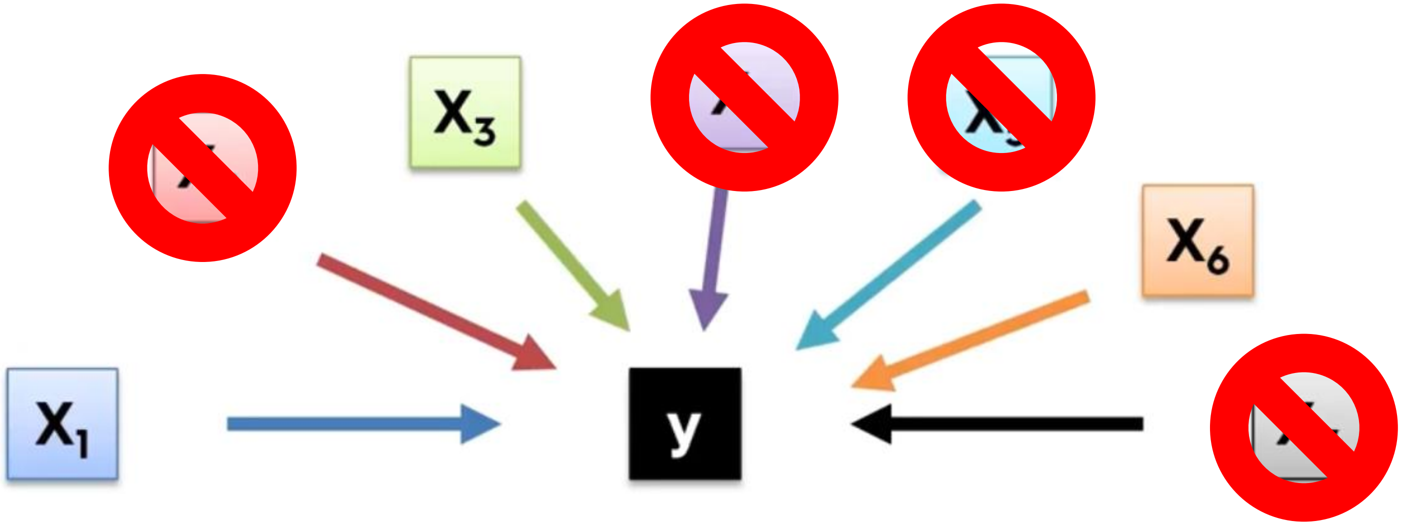
We have to optimize the model by getting rid of the insignificant variables



Model optimization

Among non-categorical variables not every independent variable is necessary for the optimal model building

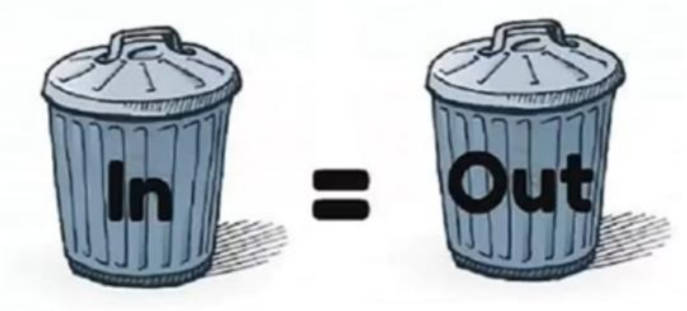
We have to optimize the model by getting rid of the insignificant variables



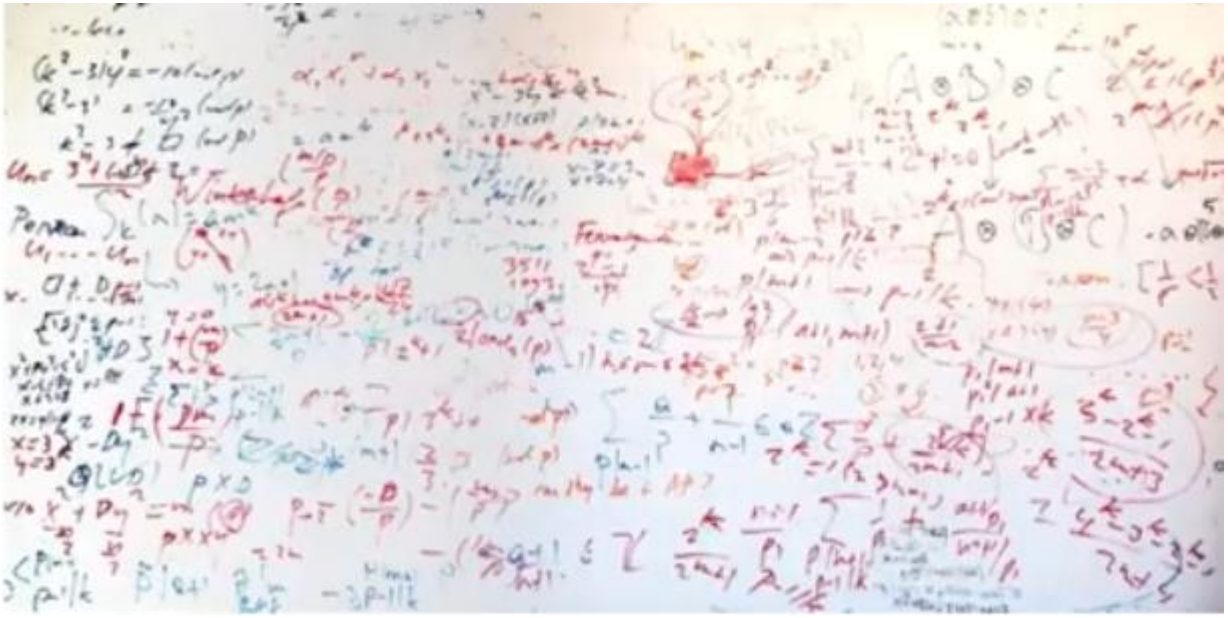
WHY???

Why?

1)



2)



Building a model

There are 5 ways of building the models:

- All – in
- Backward elimination
- Forward elimination
- Bidirectional elimination
- Score Comparison

Building a model

There are 5 ways of building the models:

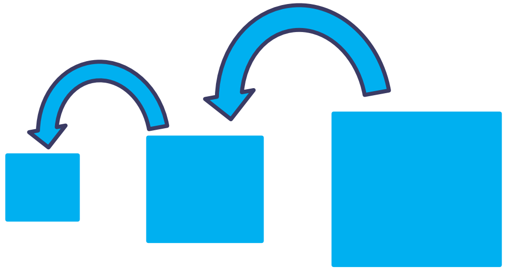
- All – in
- Backward elimination
- Forward elimination
- Bidirectional elimination
- Score Comparison

Stepwise Regression

All-in cases

- Prior knowledge; OR
- You have to; OR
- Preparing for Backward Elimination

Backward elimination



STEP 1: Select a significance level to stay in the model (e.g. SL = 0.05)

SL – significance level

STEP 2: Fit the full model with all possible predictors

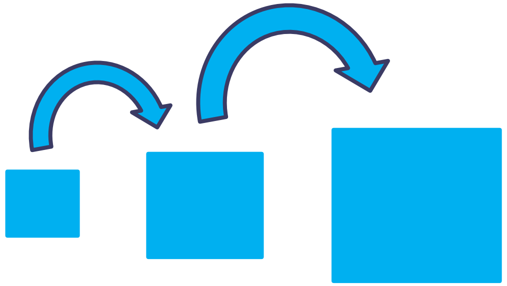
STEP 3: Consider the predictor with the <u>highest</u> P-value. If P > SL, go to STEP 4, otherwise go to FIN

STEP 4: Remove the predictor

STEP 5: Fit model without this variable*

FIN: Your Model Is Ready

Forward elimination



STEP 1: Select a significance level to enter the model (e.g. SL = 0.05)

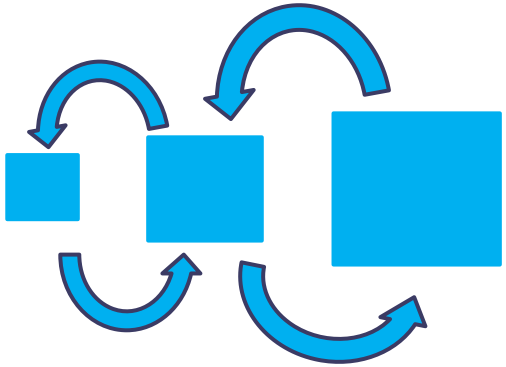
STEP 2: Fit all simple regression models $\mathbf{y} \sim \mathbf{x}_n$ Select the one with the lowest P-value

STEP 3: Keep this variable and fit all possible models with one extra predictor added to the one(s) you already have

STEP 4: Consider the predictor with the lowest P-value. If P < SL, go to STEP 3, otherwise go to FIN

FIN: Keep the previous model

Bidirectional elimination



STEP 1: Select a significance level to enter and to stay in the model
e.g.: SLENTER = 0.05, SLSTAY = 0.05

STEP 2: Perform the next step of Forward Selection (new variables must have: P < SLENTER to enter)

STEP 3: Perform ALL steps of Backward Elimination (old variables must have P < SLSTAY to stay)

STEP 4: No new variables can enter and no old variables can exit

FIN: Your Model Is Ready

Score Comparison / All possible models

STEP 1: Select a criterion of goodness of fit (e.g. Akaike criterion)

⇓

STEP 2: Construct All Possible Regression Models: 2^N-1 total combinations

⇓

STEP 3: Select the one with the best criterion

⇓

FIN: Your Model Is Ready

Example:
10 columns means
1,023 models

Significance level. P-value

- How do we get the significance level?
- And what is the P-value?

Significance level. P-value

- How do we get the significance level?
- And what is the P-value?

- You can find the answer here (1.5x speed, first ~4 minutes): https://www.youtube.com/watch?v=128yz0OCG-I
  P-value is a **probability value** which indicates how likely it is that the result happened **by chance** alone

Significance level. P-value

- If the result happened not by chance, then the P-value will be low, meaning that the result has been caused by some clear factors.
- The significance level (or Sig.) is a value determined by a user. Usually it's 0.01 or 0.05
- If:
  p<Sig. – the test is significant
  p>Sig. – the test is NOT significant

Test's significance summary

Dep. Variable:	y	R-squared:	0.951
Model:	OLS	Adj. R-squared:	0.945
Method:	Least Squares	F-statistic:	169.9
Date:	Tue, 25 Sep 2018	Prob (F-statistic):	1.34e-27
Time:	13:23:04	Log-Likelihood:	-525.38
No. Observations:	50	AIC:	1063.
Df Residuals:	44	BIC:	1074.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.013e+04	6884.820	7.281	0.000	3.62e+04	6.4e+04
x1	198.7888	3371.007	0.059	0.953	-6595.030	6992.607
x2	-41.8870	3256.039	-0.013	0.990	-6604.003	6520.229
x3	0.8060	0.046	17.369	0.000	0.712	0.900
x4	-0.0270	0.052	-0.517	0.608	-0.132	0.078
x5	0.0270	0.017	1.574	0.123	-0.008	0.062

Omnibus:	14.782	Durbin-Watson:	1.283
Prob(Omnibus):	0.001	Jarque-Bera (JB):	21.266
Skew:	-0.948	Prob(JB):	2.41e-05
Kurtosis:	5.572	Cond. No.	1.45e+06

Test's significance summary

Dep. Variable:	y	R-squared:	0.951
Model:	OLS	Adj. R-squared:	0.945
Method:	Least Squares	F-statistic:	169.9
Date:	Tue, 25 Sep 2018	Prob (F-statistic):	1.34e-27
Time:	13:23:04	Log-Likelihood:	-525.38
No. Observations:	50	AIC:	1063.
Df Residuals:	44	BIC:	1074.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.013e+04	6884.820	7.281	0.000	3.62e+04	6.4e+04
x1	198.7888	3371.007	0.059	0.953	6595.030	6992.607
x2	-41.8870	3256.039	-0.013	0.990	6604.003	6520.229
x3	0.8060	0.046	17.369	0.000	0.712	0.900
x4	-0.0270	0.052	-0.517	0.608	-0.132	0.078
x5	0.0270	0.017	1.574	0.123	-0.008	0.062

Omnibus:	14.782	Durbin-Watson:	1.283
Prob(Omnibus):	0.001	Jarque-Bera (JB):	21.266
Skew:	-0.948	Prob(JB):	2.41e-05
Kurtosis:	5.572	Cond. No.	1.45e+06

Test's significance summary

OLS Regression Results

Dep. Variable:	y	R-squared:	0.950
Model:	OLS	Adj. R-squared:	0.948
Method:	Least Squares	F-statistic:	450.8
Date:	Tue, 25 Sep 2018	Prob (F-statistic):	2.16e-31
Time:	13:23:36	Log-Likelihood:	-525.54
No. Observations:	50	AIC:	1057.
Df Residuals:	47	BIC:	1063.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.698e+04	2689.933	17.464	0.000	4.16e+04	5.24e+04
x1	0.7966	0.041	19.266	0.000	0.713	0.880
x2	0.0299	0.016	1.927	0.060	-0.001	0.061

Omnibus:	14.677	Durbin-Watson:	1.257
Prob(Omnibus):	0.001	Jarque-Bera (JB):	21.161
Skew:	-0.939	Prob(JB):	2.54e-05
Kurtosis:	5.575	Cond. No.	5.32e+05

Coefficients

- How to interpret coefficients?
- If the sign is positive – then it means that your independent variable is correlated with the depended variable.
  Means that if you increase this value, the output values will also increase
- The negative has opposite effect respectively

Coefficients

- Magnitude – measures the impact as well.
- However, magnitude states that a variable A per unit has a bigger impact on the output than a variable B per unit has.
- The variables can be measured in different units ($, hours, kg…)

Higher magnitude

Lower magnitude

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     4.698e+04  2.690e+03  17.464   <2e-16 ***
R.D.Spend       7.966e-01  4.135e-02  19.266   <2e-16 ***
Marketing.Spend 2.991e-02  1.552e-02   1.927     0.06 .  
```

Test's significance summary

Dep. Variable:	y	R-squared:	0.951
Model:	OLS	Adj. R-squared:	0.945
Method:	Least Squares	F-statistic:	169.9
Date:	Tue, 25 Sep 2018	Prob (F-statistic):	1.34e-27
Time:	13:23:04	Log-Likelihood:	-525.38
No. Observations:	50	AIC:	1063.
Df Residuals:	44	BIC:	1074.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.013e+04	6884.820	7.281	0.000	3.62e+04	6.4e+04
x1	198.7888	3371.007	0.059	0.953	-6595.030	6992.607
x2	-41.8870	3256.039	-0.013	0.990	-6604.003	6520.229
x3	0.8060	0.046	17.369	0.000	0.712	0.900
x4	-0.0270	0.052	-0.517	0.608	-0.132	0.078
x5	0.0270	0.017	1.574	0.123	-0.008	0.062

Omnibus:	14.782	Durbin-Watson:	1.283
Prob(Omnibus):	0.001	Jarque-Bera (JB):	21.266
Skew:	-0.948	Prob(JB):	2.41e-05
Kurtosis:	5.572	Cond. No.	1.45e+06

Datasets sources

- http://archive.ics.uci.edu/ml/datasets.html?task=reg
- http://people.sc.fsu.edu/~jburkardt/datasets/regression/regression.html
- https://www.quora.com/Where-can-I-find-data-sets-for-regression
- https://vincentarelbundock.github.io/Rdatasets/datasets.html