Types of Variables 
Variables 
Categorical 
Nominal 
Male, Female 
Red, Green, Blac 
so nominal and the word names they sound similar and therefore 
as with names you can order them nominal 
Data Science Training 
@ Kirill Eremenko 

Example - Names and colors - You cannot order them

Types of Variables 
Variables 
Categorical 
Nominal 
Male, Female 
Ordinal 
Small, Medium, Large 
Red, Green, Blac 
And moreover you can order them so you can put these groups of 
clothes in a logical order meaning that 
Data Science Training 
Kirill Eremenko 

Example - Clothes and grades- You can order them

Types of Variables 
Variables 
Categorical 
Numeric 
Nominal 
Male, Female 
Ordinal 
Small, Medium, Large 
Discrete 
Red, Green, Blac 
They are basically; numbers so how many units of something 
or they are or how much of something 
Data Science Training 
@ Kirill Eremenko 

It can't be a fraction, has to be a whole number. It can be an infinite number of observations.

Example 1 2 3 5 business or 543 or 583 people.

Types of Variables 
Variables 
Categorical 
Numeric 
Nominal 
Male, Female 
Red, Green, Black 
Data Science Training 
Ordinal 
Small, Medium, Large 
A, B, C (grades) 
Discrete 
1, 2, 3 businesses 
568 people 
But discrete they can be a fraction: 
Continuous 
@ Kirill Eremenko 

Example- age or height

---------------------------------------------------------------------

Regression –

Linear and Logistic – Four types -

Regressions 
In statistics, regression analysis is a statistical process 
for estimating the relationships among variables. 
The fOCUS is on the relationship between a dependent 
variable and one or more independent variables. 
relationships among 
Data Science Training 
-Wikipedia 
@ Kirill Eremenko 

Regressions 
Linear 
Simple 
Linear 
Regression 
Logistic 
Multiple 
Linear 
Regression 
Simple 
Logistic 
Regression 
Multiple 
Logistic 
Regression 
and the multiple Just vigorish And in terms of logistic regression. 
Data Science Training 
@ Kirill Eremenko 

 It is mostly Multiple Logistic Regression while dealing with Logistic Regression.

We will study that later in the other section.

.Regressions 
Constant 
Simple 
Linear 
Regression 
Coefficient 
Dependent variable (DV) 
Multiple 
Linear 
Regression 
Independent variable (IV) 
And B is the constants as a constant term and we'll talk about 
the constant just in a few seconds 
Data Science Training 
@ Kirill Eremenko 

Regressions 
Simple 
Linear 
Regression 
Multiple 
Linear 
Regression 
bl*Xl + b2*X2 + 
So bizzarro one constant and then many combinations or many 
pairs of B and X and it can go up to quite 
Data Science Training 
... + bn*xn 
@ Kirill Eremenko 

**Simple Linear Regression -**

This is the line that best fits the data.

And so the steeper the line the more you get more money you get per extra year of experience.

B1= slope of the line

Regressions 
Simple Linear Regression: 
Salary ($) 
Ex erience 
y = bo + bl*x 
Salary = bo + bl 
This is the line that best fits as Darren even looks like it right. 
Data Science Training 
*Experience 
@ Kirill Eremenko 

Regressions 
Simple Linear Regression: 
Salary ($) 
+10k 
30k 
y = bo + bl*x 
Salary 
that onto the salary access and you can see that here for one of 
your experience the person will get 
Data Science Training 
Experience 
@ Kirill Eremenko 

**Ordinary Least Squares –**

So it's the difference between the observed and the modelled values for a particular independent variable.

**Sum(yi – yi^ )**

**R-Squared -**

Sum of square of residuals - Ss res = Sum(Yi - Yi^)2 ---------> Yi hat

The total sum of squares - Ss tot = Sum(Yi - Yavg)2

R² = 1 – SS res/ SS tot

Normally R squared is between 0 and 1. The closer it is to 1 the better.

R Squared 
Simple Linear Regression: 
Salary ($) 
SUM (y - Y? min 
We talked about the simpie introgression been constructed through 
the ordinary least squares method 
Data Science Training 
@ Kirill Eremenko 

R Squared 
Simple Linear Regression: 
Salary ($) 
avg 
sst 
-SUM (y 
-SUM (y - 
Yavg 
ssres 
sstot 
And what are squared our R„sguared equals to 1 minus sum of 
squares of residua's divided by total 
Data Science Training 
@ Kirill Eremenko 

**Adjusted R-squared (R²)**

SS res 🡪 should reduce -- should be minimum

Adjusted R² = 1 – ( 1 - R² ) n-1/n – p -1

p- number of regressors

n = sample size

**Simple Linear Regression with GRETL**

Step 1 -

Dataset – SalaryData.csv

Launch Gretl 🡪File 🡪 Open Data 🡪SalaryData.csv 🡪 View 🡪 Summary Statistics (of data)

Salary or YearsofExp to the other side 🡪 Press Ok

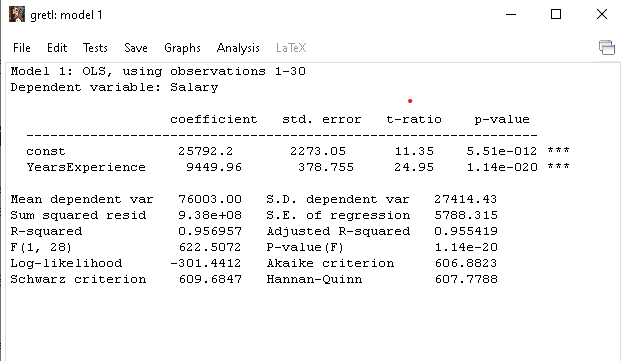
Right click on the column 🡪 Edit 🡪 Explore more

Step 2 -

Model 🡪 Ordinary least square ( it is SLR )

Salary to the Dependant Variable and Years of exp to the Independent Variable 🡪 Press Ok

Check the p-value 🡪 Lower the p-value, the better



Step 3 –

Plotting and analysing graphs

Go to Graphs 🡪 Fitted Actual Plot 🡪 Actual vs Fitted OR

Analysis 🡪 Forecast 🡪 Ok

Confidence Intervals – 95% of the times the data will remain in between the 2 parallel blue lines



Section 3. Multiple Linear Regression

Continuous output -

Dummy Variables –

Convert the categorical features to dummy variables to avoid the dummy variable trap

(One hot encoder or Label encoder)

Watch again – Using Adjusted R-Squared to create a Robust model

<https://www.udemy.com/course/datascience/learn/lecture/3495916#overview>

Interpreting the co-efficients of MLR -

<https://www.udemy.com/course/datascience/learn/lecture/3495918#overview>

A Caveat 
Assumptions of a Linear Regression: 
1. Linearity 
2. Homoscedasticity 
3. Multivariate normality 
4. Independence of errors 
5. Lack of multicollinearity 
Data Science Training 
@ Kirill Eremenko 

1

The first assumption is linearity.

We want to make sure that there is a linear relationship between our dependent variable and each independent variable.

2

The second assumption is homo scarcity.

It actually simply means equal variance, meaning that you don't want to see a cone type shape on your chart, whether an increasing or decreasing cone, which would mean that variance is dependent on the independent variable.

3

The third assumption is multivariate normality or normality of error distribution.

The best way to intuitively think about it is if you look along the line of the linear regression,

you want to see a normal distribution of your data points.

In the case o

4

The fourth assumption is independence of observations, and this includes the term no or a correlation.

And what that means is that we don't want to see any kind of pattern in our data.

A pattern in the data like we see here indicates that our rows are not independent, that some rows

are fixing other rows and other rows, etc..

A classic example of this would be the stock market, where previous prices affect future prices, which

affect future prices and so on.

5

The fifth assumption is lack of multi collinearity.

Basically, we want our independent variables or predictors not to be correlated with each other.

If they're not correlated, then we can build a linear regression.

If they are correlated, then if we do proceed and build a linear regression model, then the coefficient

estimates that we get in the model will become unreliable.

6

And the sixth point is the outlier check.

This is not an actual assumption, but rather an extra check.

That is important to keep in mind when building linear regression models.

If you look at the chart here on the right, you can see that the outlier is significantly affecting

the linear regression line that we get.

So something that we want to consider is should we remove the outliers before building a linear regression,

or do we want to build a linear regression with the outliers included?

\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\

Dummy Variables for categorical values

Dummy Vart— 
Dummy Variables 
New York California 
Profit 
192,261.83 
191,792.06 
191,050.39 
182,901.99 
166,187.94 
spend 
165,349.20 
162,597.70 
153,441.51 
144,372.41 
142,107.34 
Admin 
136,897.80 
151,377.59 
101,145.55 
118,671.85 
91,391.77 
Marketing 
471,784.10 
443,898.53 
407,934.54 
383,199.62 
366,168.42 
State 
New York 
California 
California 
New York 
California 
1 
1 
bl*Xl + b2*X2 + b3*X3 + 
An two new columns are called dummy and 
building your regression mode! from here 
Data Science Training 
1 
1 
1 
@ Kirill Erernenko 

Dummy Vart— 
Dummy Variables 
New York California 
Profit 
192,261.83 
191,792.06 
191,050.39 
182,901.99 
166,187.94 
spend 
165,349.20 
162,597.70 
153,441.51 
144,372.41 
142,107.34 
Admin 
136,897.80 
151,377.59 
101,145.55 
118,671.85 
91,391.77 
Marketing 
471,784.10 
443,898.53 
407,934.54 
383,199.62 
366,168.42 
State 
ew York 
ifornia 
Call rnia 
New k 
Californi 
1 
1 
+ b4*Dl 
1 
1 
bl*Xl + b2*X2 + b3*X3 
Bu now I would like to discuss so first of all the New 
York or of the dummy 
Data Science Training 
@ Kirill Eremenko 

Building A Model 
5 methods of building models: 
All¯in 
1. 
2. Backward Elimination 
3. Forward Selection 
4. Bidirectional Elimination 
5. Score Comparison 
And number five is score comparison. 
Data Science Training 
@ Kirill Eremenko 

Building A Model 
5 methods of building models: 
All¯in 
1. 
2. Backward Elimination 
Stepwise 
3. Forward Selection 
Regression 
4. Bidirectional Elimination 
5. Score Comparison 
But sometimes you will hear people or stepwise 
regression in reference to just not before 
Data Science Training 
@ Kirill Eremenko 

--------------------------------------------------------------------------------------------------------------------------------------

Buildin A Model 
Backward Elimination 
STEP 1: Select a significance level to stay in the model (e.g. SL 
STEP 2: Fit the full model with all possible predictors 
= 0.05) 
STEP 3: Consider the predictor with the highest 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* 
Data Science Training 
@ Kirill Eremenko 

-------------------

Buildin A Model 
Forward Selection 
STEP 1: Select a significance level to enter the model (e.g. SL = 0.05) 
STEP 2: Fit all simple regression models y xn 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 
Data Science Training 
@ Kirill Eremenko 

-------------------------

Buildin A 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 
Data Science Training 
@ Kirill Eremenko 

------------------------------

Building A Model 
STEP 1: 
STEP 2: 
STEP 3: 
Possible Models 
Select a criterion of goodness of fit (e.g. Akaike criterion) 
Construct All Possible Regression Models: 2N-1 total combinations 
Select the one with the best criterion 
FIN: Your Model Is Ready 
Data Science Training 
Example: 
10 columns means 
1,023 models 
@ Kirill Eremenko 

gretl: model 2 
Eile Edi lests Save Graphs LaTeX 
2: I—SO 
gretl 
file 
50 Sta 
gretl: rnodel 
Eile Edit Jests Save graphs Anawsis LaTeX 
Model OLS, I—SO 
Dependent variable: Profit 
8 
Std. 
error 
66S3. sq 
o.oszesS9 
p—value 
1.43e-og 
o. 1209 
o. 6229 
Dependent : Prcfit 
coe ent 
47721.8 
cong t 
at ion 
Ha: ketingSpend 
NewYork 
coefficient 
S0416.s 
o. 0236200 
o .02€3€92 
—1332.09 
Std. 
error 
301B .34 
o. 01 seoB6 
IS.EI 
1.809 
—0.49S2 
Mean dependent var 112012.6 
Sum squared resid 
ke t ingSpend 
dependent 
Sum 
R— squared 
46) 
Log— ke 
Schwarz 
o .02%947 
—2424.€1 
o. 950787 
296.23" 
-szs .3649 
10€€.378 
S. D. dependent 
S.E. regression 
us Ced R—zquared 
P-value (F) 
Akaike on 
Hannan—Quinn 
p —value 
3.22e—020 
0.0770 
9222.426 
0.947578 
loss. 730 
1061.642 
Log — 
Schwarz criterion 
1070.0 
S. D. dependent 
S.E. of regression 
Ad: R—squared 
P-vaiue (E) 
Hannan—Quinn 
40306. IS 
9309.026 
7. sse-29 
1060. soo 
10€4.140 
Zhe constant, p—vaiue 
was € (NevYcrk) 
Excluding the constant, p—value 
gretl: "'Odel 3 
Eile Edit rests Save Graphs Analysis 
was highest for variable 2 
LaTex 
Model 3: OCS, using observations I—SO 
Dependent 
file lets Save LaTeX 
Yodel 4: OLS, using observations 
std. 
error 
2629.93 
0.0413476 
0.01SS200 
Dependent 
enz 
49032.9 
o.8S4291 
I-so 
zzd. 
2537.90 
o .02930% 
const 
RDSpend 
g Spend 
Mean dependeaz 
Sum squared resid 
Schwarz criterion 
coefftcienz 
46975.9 
o. 796584 
0.0299079 
112012. € 
3.94e+09 
450.7713 
-sas.sssg 
1062 .807 
19.27 
1.927 
p—vaiue 
3.soe-022 
6.04e-024 
0.0600 
40306. z: 
9160. 966 
2.16e-31 
loss. ass 
19.32 
29. IS 
S. D. dependent 
S.E. of regression 
R —squared 
P-value (F) 
Rann an 
Yean dependent 
Sum squared 
R— squared 
log — ihOOd 
Schwarz 
112012.6 
4.2€e.09 
o. 946S3S 
-S27.436S 
. €97 
S. D. dependent var 
S.E. 
Ad: ugted A—squared 
Akaike criterion 
Hannan—Quinn 
p—vaiue 
2.7 
3. SOe—032 
40306.1e 
9416.349 
o. 945421 
lose.E73 
1060.329 

File-open data-user file-select file

State-Add-Dummy for discreet variable

Model-ordinary least squares-profit into dependent variable-all others one by one into regressors or independent variable.

Graphs-fitted actual plot-against marketing spend or New York and etc.

So R-squared here is basically a characteristic or a parameter of your model which tells you about the goodness of fit. So how will your model has been fitted and R squared can never be greater than one and you wanted to be as close to one as possible the closer to 1 it is the better your model is deemed to be fitted. However, we also discussed this part in the refresher R-Squared is biased and it's biased in a way that it is constructed and the way these models are run. So the ordinary least squared method it doesn't allow R-squared to ever decrease so the more variables you add to your model the greater R-squared will be. So basically what we're what this means is that as long as you keep adding variables R-squared will always grow and we can observe that here.

Adjusted r squared … So basically just like R squared would grow if you add more variables that are square would also grow but there is a penalization factor which makes it small which reduces adjusted R-squared as you add more variables… it's constructed on the other hand the penalization factor is penalizing you or penalizing adjusted R-squared and reducing it every time you add a variable.

Checking adjusted r squared to see if it is increasing or decreasing.

In the third window it increased like the 1st and 2nd but in the 4th it decreased so that means that the third window serves a better understanding of the situation even though we did not go as the backward elimination process and did not remove the p value of 0.06 which was higher than our significance level which is 0.05.

So the third model was better fitted to predict or explain the variance in profit.

Also check akaike criterion. Its better when its lower as in our third model.

 you can say that a dollar increase in R&D spend drives 79 cents dull 79 cents increase in profit and mark a dollar increase in marketing spend drives only 3 cents increase in profits so basically if you're the venture capitalist which companies are going to invest in well from this model you're going to decide that you should invest into companies that are spending more on R&D. Just remember about that per unit trick or tip I guess because if you forget about that you can you know make the wrong conclusions.

Section Recap 
In this section we learned: 
1. How to create dummies for categorical IVs 
2. How to avoid the dummy variable trap 
3. Backward, Forward, Bidirectional, All Possible 
4. We actually built a model. Step-By-Step!! 
5. How to use adjusted R-squared in modelling 
6. How to interpret coefficients of a MLR 
Data Science Training 
@ Kirill Eremenko 

**Logistic Regression**

For a binary outcome – 1 or 0 – Yes or No - Discreet

False Positives – We predicted the outcome will occur but it did not occur

Meaning – We said Yes but in reality it was a No.

Type 1 error is Less dangerous

False Negative – We predicted the outcome will not occur but it did occur

Meaning – We said No but in reality it was a Yes

Type 2 error is More dangerous

Confusion Matrix

1. Accuracy Rate = Correct / Total
2. Error Rate = Wrong / Total

**Section 5. Building a Robust Geodemographic Segmentation Model**

[**https://www.udemy.com/course/datascience/learn/lecture/3496662#overview**](https://www.udemy.com/course/datascience/learn/lecture/3496662#overview)

