Regression

A skin-deep dive (oxymoron intended)
by @abulyomon

Agenda

Simple Linear Regression

Exercise

Multiple Linear Regression

Exercise

Logistic Regression

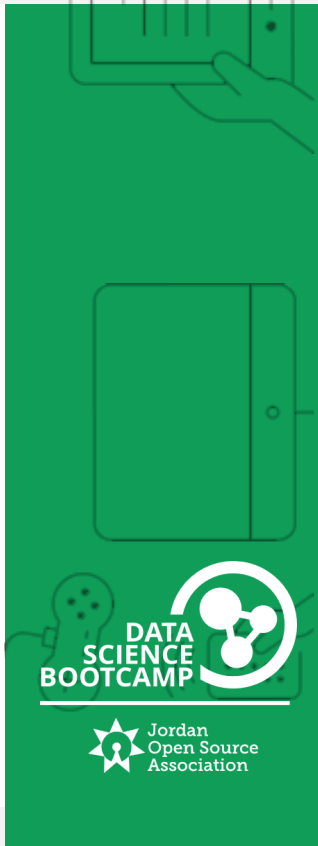
Exercise

Where am I?

By now, you should be able to:

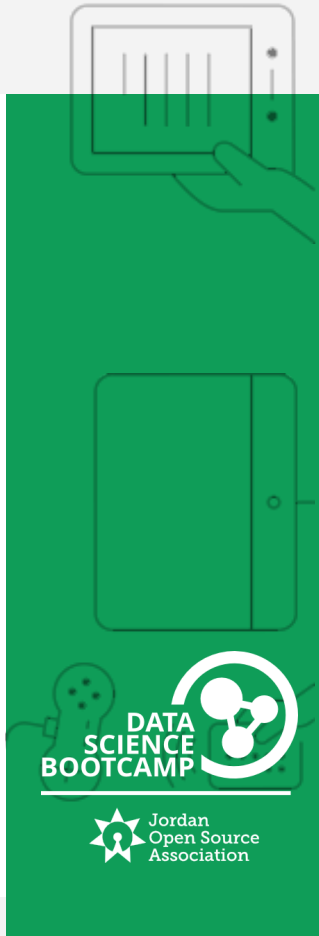
- Obtain and clean data.
- Conduct Exploratory Data Analysis as well as some visualization.

This session is your first step into explaining some cross-sectional data.



Disclaimer

- This session does not teach Python.
- Although it is necessary to understand the math behind modeling, we will not go through details. We will give recipes :(

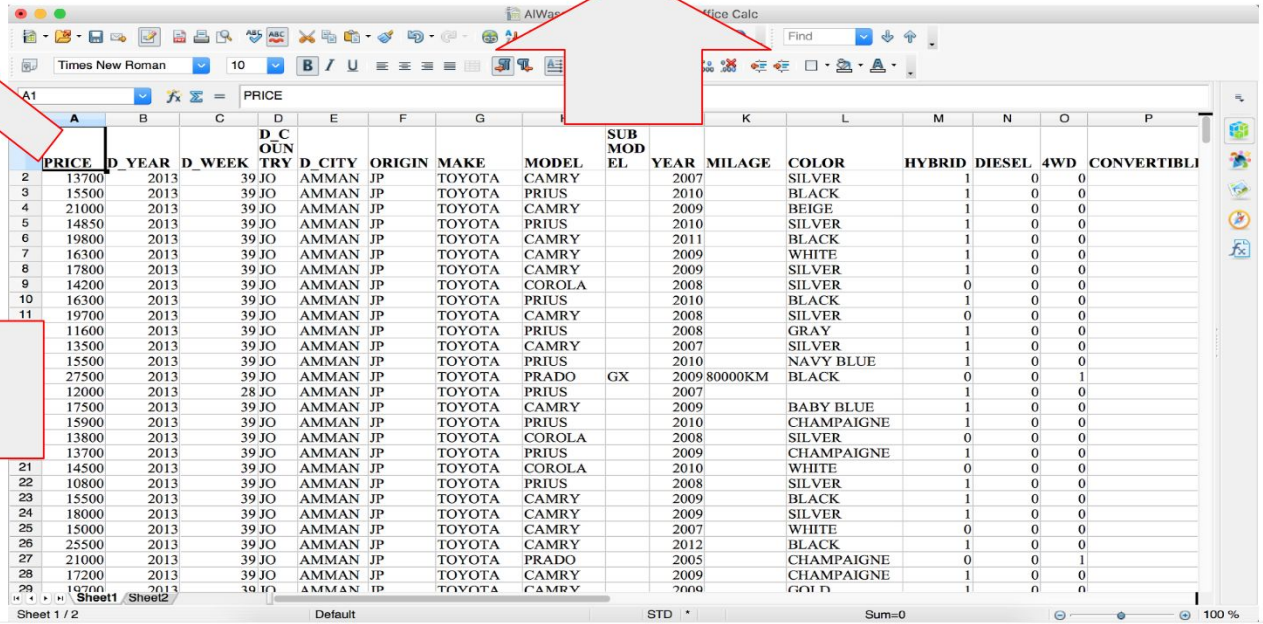


Context & Terminology

Y
Dependant
Explained
Predicted
Regressed
Response

X_i, (Independent), Explanatory, Predictor, Regressor, Control, Covariate, Carrier

Observations



	PRICE	D_YEAR	D_WEEK	D_COUNTRY	D_CITY	ORIGIN	MAKE	MODEL	SUBMODEL	YEAR	MILAGE	COLOR	HYBRID	DIESEL	4WD	CONVERTIBLI
2	13700	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2007		SILVER	1	0	0	
3	15500	2013	39	JO	AMMAN	JP	TOYOTA	PRIUS		2010		BLACK	1	0	0	
4	21000	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2009		BEIGE	1	0	0	
5	14850	2013	39	JO	AMMAN	JP	TOYOTA	PRIUS		2010		SILVER	1	0	0	
6	19800	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2011		BLACK	1	0	0	
7	16300	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2009		WHITE	1	0	0	
8	17800	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2009		SILVER	1	0	0	
9	14200	2013	39	JO	AMMAN	JP	TOYOTA	COROLA		2008		SILVER	0	0	0	
10	16300	2013	39	JO	AMMAN	JP	TOYOTA	PRIUS		2010		BLACK	1	0	0	
11	19700	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2008		SILVER	0	0	0	
	11600	2013	39	JO	AMMAN	JP	TOYOTA	PRIUS		2008		GRAY	1	0	0	
	13500	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2007		SILVER	1	0	0	
	15500	2013	39	JO	AMMAN	JP	TOYOTA	PRIUS		2010		NAVY BLUE	1	0	0	
	27500	2013	39	JO	AMMAN	JP	TOYOTA	PRADO	GX	2009	80000KM	BLACK	0	0	1	
	12000	2013	28	JO	AMMAN	JP	TOYOTA	PRIUS		2007			1	0	0	
	17500	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2009		BABY BLUE	1	0	0	
	15900	2013	39	JO	AMMAN	JP	TOYOTA	PRIUS		2010		CHAMPAIGNE	1	0	0	
	13800	2013	39	JO	AMMAN	JP	TOYOTA	COROLA		2008		SILVER	0	0	0	
	13700	2013	39	JO	AMMAN	JP	TOYOTA	PRIUS		2009		CHAMPAIGNE	1	0	0	
21	14500	2013	39	JO	AMMAN	JP	TOYOTA	COROLA		2010		WHITE	0	0	0	
22	10800	2013	39	JO	AMMAN	JP	TOYOTA	PRIUS		2008		SILVER	1	0	0	
23	15500	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2009		BLACK	1	0	0	
24	18000	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2009		SILVER	1	0	0	
25	15000	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2007		WHITE	0	0	0	
26	25500	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2012		BLACK	1	0	0	
27	21000	2013	39	JO	AMMAN	JP	TOYOTA	PRADO		2005		CHAMPAIGNE	0	0	1	
28	17200	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2009		CHAMPAIGNE	1	0	0	
29	19700	2013	39	JO	AMMAN	JP	TOYOTA	CAMRY		2009		GOLD	1	0	0	

DATA SCIENCE BOOTCAMP

Jordan Open Source Association

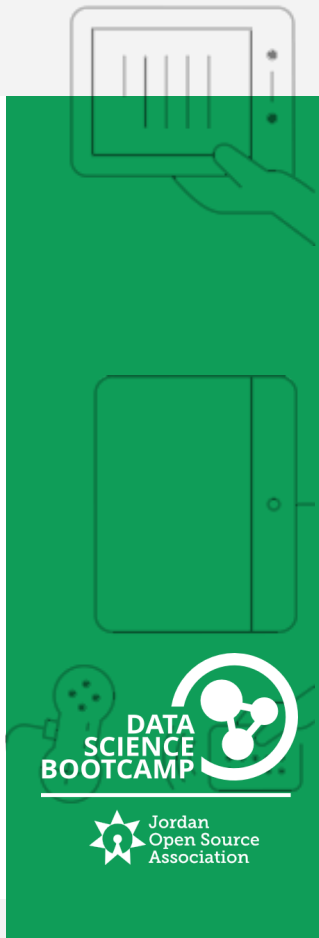
Linear Regression

A regression model approximates the relation between the dependant and independant variables:

$$Y = f(X_1, X_2, ..., X_i) + \varepsilon$$

when Y is continuous quantitative and the relation is linearizable:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i + \varepsilon$$



Linearity

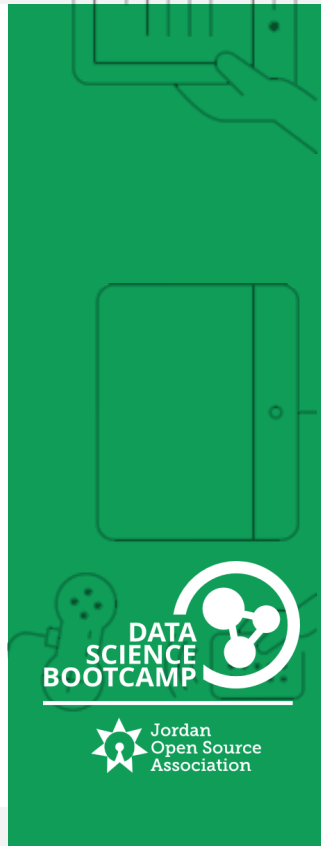
- Linear

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

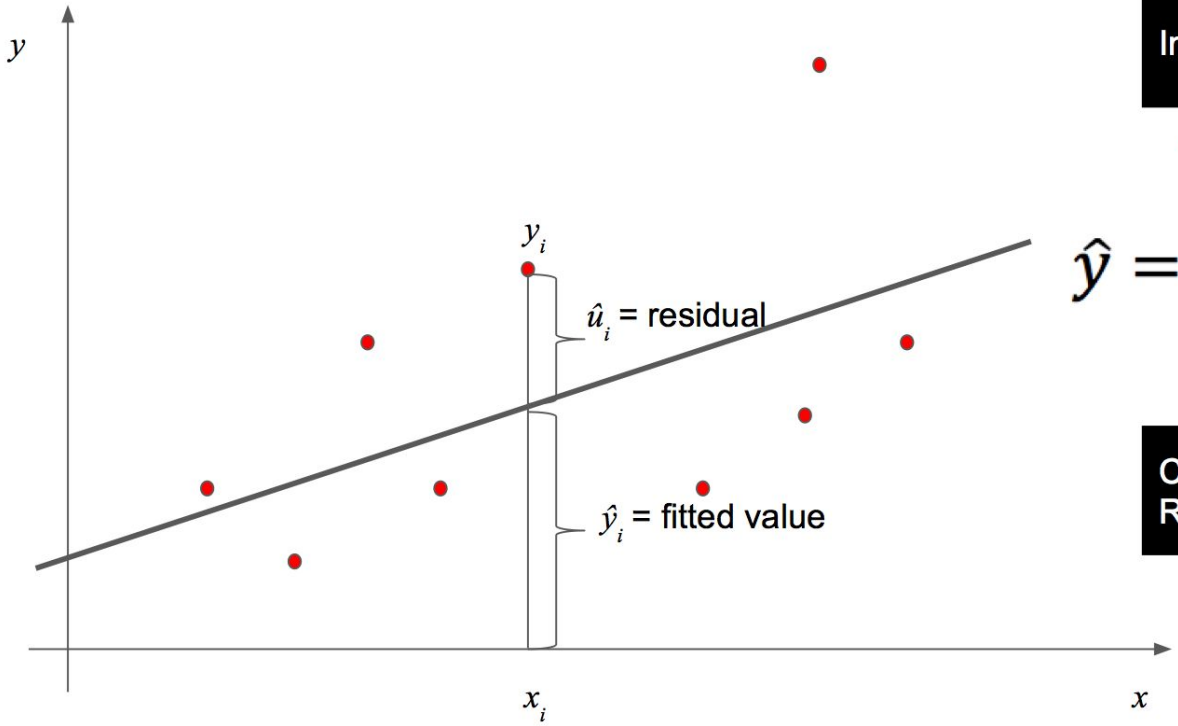
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Not linear

$$Y = \beta_0 + e^{\beta_1 X_1} + \varepsilon$$



Ordinary Least Squares Estimation



Intercept

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$

Coefficient
Regression Parameter

DATA SCIENCE BOOTCAMP

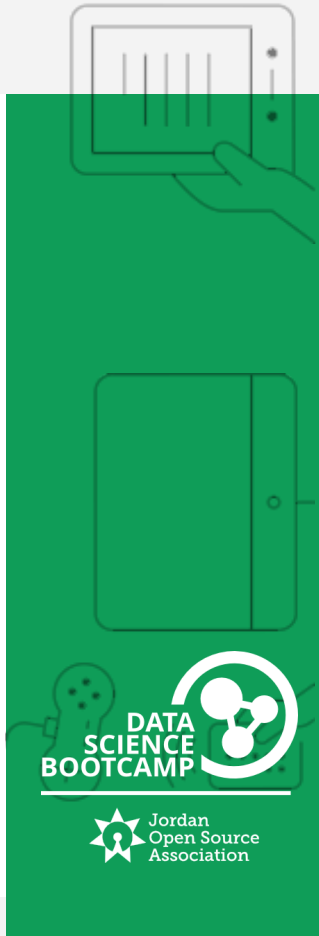
Jordan Open Source Association

Notes on OLS

OLS Assumes:

1. Linearity
2. Random sampling
3. Error terms are normally distributed
4. Error terms have constant variance (homoskedasticity)
5. Error terms are independent (autocorrelation?)

There is another estimation method called Maximum Likelihood Estimation. In the case of Linear Regression, MLE yields OLS findings!



Practice

File: AlWaseet.csv

Problem definition: What is the relation between car mileage and car price?

Model Interpretation + Inference

Coefficients

R^2: Predictive power based on correlation

= Explained Variation/ Total Variation

***p*-value**: Statistical significance

OLS Regression Results

Dep. Variable:	PRICE	**R-squared:**	0.010
Model:	OLS	**Adj. R-squared:**	-0.004
Method:	Least Squares	**F-statistic:**	0.7127
Date:	Sat, 19 Dec 2015	**Prob (F-statistic):**	0.401
Time:	19:37:36	**Log-Likelihood:**	-803.09
No. Observations:	72	**AIC:**	1610.
Df Residuals:	70	**BIC:**	1615.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	2.498e+04	3892.261	6.417	0.000	1.72e+04 3.27e+04
MILEAGE	-0.0476	0.056	-0.844	0.401	-0.160 0.065

10’

BREAK TIME

DATA SCIENCE BOOTCAMP

Jordan Open Source Association

Multiple Linear Regression

Where is the nice graph?

Problem of collinearity

R^2: Curse of Dimensionality

Practice

File: AlWaseet.csv

Can we build a better model for price with more predictors?

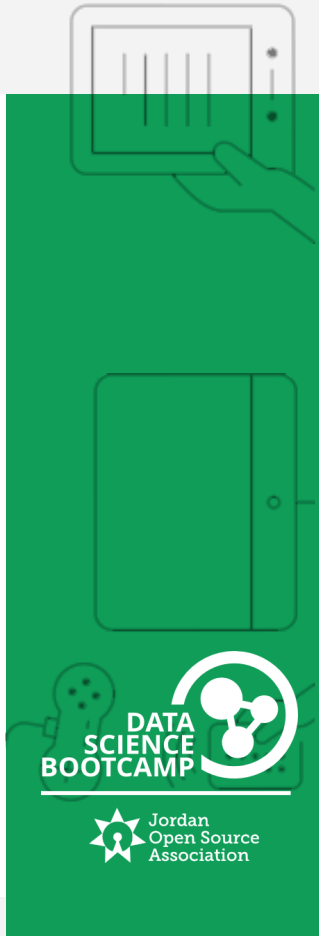
More inference + Model Comparison

R_a^2: Adjusted

F-test: tests the null hypothesis that coefficients are zero!

AIC: Information Criteria

BIC: Information Criteria with severe penalty to (i)



Goodness-of-fit: Visual Checks

Normal probability plot of residuals ->

Scatter plot of residuals vs predictor variables

Scatter plot of residuals vs fitted values

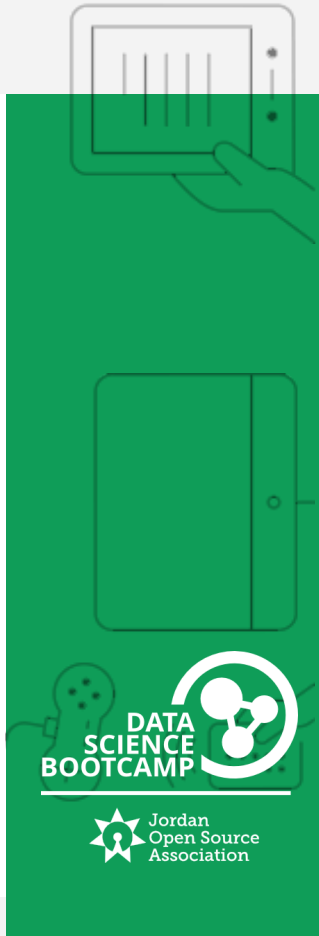
```
import statsmodels.graphics.gofplots as gof

import scipy.stats as stats

gof.qqplot(fitted_model.resid, stats.t, fit=True,
line='45')

plt.show()

###

plt.hist(fitted_model.resid_pearson)

plt.show()
```

Your HW, should you choose to accept it

Is the origin of the car significant in determining its price?

Are Korean made cars particularly different when it comes to price?

How do we handle categorical variables?



10’

BREAK TIME

DATA SCIENCE BOOTCAMP

Jordan Open Source Association

Logistic Regression

We have expressed earlier that a regression model approximates the relation between the dependant and independant variables:

$$Y = f(X_1, X_2, ..., X_i) + \varepsilon$$

when Y is binary:

$$ln(\pi/1-\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i$$

where $\pi = P(Y=1|X_1=x_1, ..., X_i=x_i)$ and $(\pi/1-\pi)$ is referred is to as the "odds."

Let’s demystify that cryptic slide!

Suppose Y, the dependant variable was “CHASSIS” which signifies if a car was damaged, denoted by: “1” if damaged, and “0” if not. One way to model this may be:

$$ln(o) = \beta_0 + \beta_1 PRICE + \beta_2 MILEAGE + \beta_3 AGE$$

The odds of a car of price p, mileage m, and age a being damaged would be:

$$o = e^{\beta_0 + \beta_1 p + \beta_2 m + \beta_3 a}$$

Practice

File: AlWaseet.csv

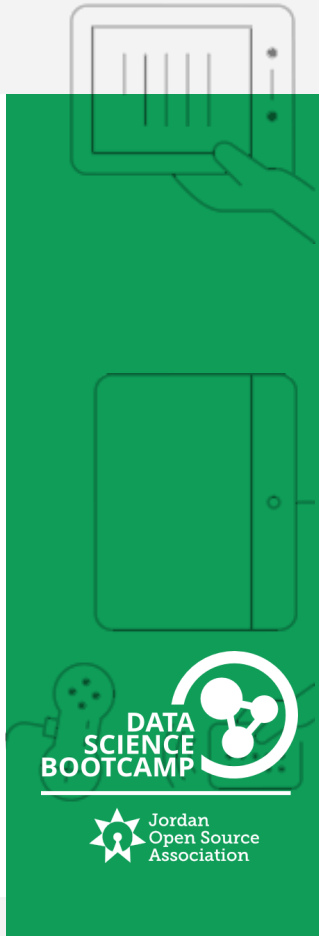
Can we guess the smashed car?

Predictive Accuracy

in-sample -> goodness-of-fit

Guard against **overfitting**

out-of-sample -> cross-validation



Next?

Further reading

- Regression Analysis by Example, Chatterjee & Hadi
- Logistic Regression using SAS, Allison

Contact me:

- @abulyomon
- abulyomon@gmail.com