EDA

COGS 108 Fall 2025
Jason Chen
Week 6

xic007@ucsd.edu

OH: Tue 3-5 pm

Due dates

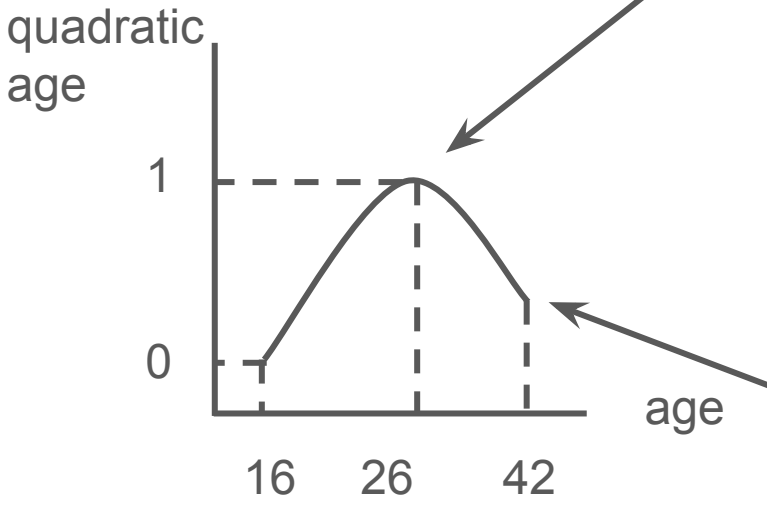
- A2: Wednesday (11/05)
- D5: Friday (11/07)

Part II : quadratic transformation

Choose quad_b so that
max(quad_a * (age - 26)^2 + quad_b) = 1

max(quad_a * (age - 26)^2 + quad_b)
= quad_a * (26 - 26)^2 + quad_b
= quad_b

quad_b = ?



Choose quad_a so that for the max value of (age - 26)^2,
quad_a * (max((age - 26)^2) = 1

max((age - 26)^2) = (42- 26)^2 = 256

quad_a = 1/max((age - 26)^2) = ?

Part III : statsmodels.regression.linear_model.OLS

What is **OLS** (Ordinary Least Squares)?

A fundamental method for estimating relationships between variables in **regression** analysis.

It fits a **line** (or plane) through the data that **minimizes the sum of squared differences** between the observed values and the model's predicted values.

Introduction :

A linear regression model establishes the relation between a dependent variable(**y**) and at least one independent variable(**x**) as :

$\hat{y} = b_1 x + b_0$

In *OLS* method, we have to choose the values of $b_1$and $b_0$such that, the total sum of squares of the difference between the calculated and observed values of y, is minimised.

Formula for OLS:

$$S = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_1 x_1 - b_0)^2 = \sum_{i=1}^{n}(\hat{\epsilon}_i)^2 = min$$

Where,

$\hat{y}_i$= predicted value for the ith observation

y_i= actual value for the ith observation

ϵ_i= error/residual for the ith observation

n = total number of observations

To get the values of $b_0$and $b_1$which minimise S, we can take a partial derivative for each coefficient and equate it to zero.

Part III : statsmodels.regression.linear_model.OLS

What OLS does

Estimates how **changes** in predictor variables affect the **outcome variable**.

Provides estimates for:

Coefficients (β) → strength and direction of relationships.

Intercept → expected value of the outcome when predictors are 0.

R² and F-statistic → overall model fit and significance.

p-values → how likely it is that each effect (or the model overall) occurred by chance.

Introduction :

A linear regression model establishes the relation between a dependent variable(**y**) and at least one independent variable(**x**) as :

$\hat{y} = b_1 x + b_0$

In *OLS* method, we have to choose the values of $b_1$and $b_0$such that, the total sum of squares of the difference between the calculated and observed values of y, is minimised.

Formula for OLS:

$$S = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_1 x_1 - b_0)^2 = \sum_{i=1}^{n}(\epsilon_i)^2 = min$$

Where,

$\hat{y}_i$= predicted value for the ith observation

y_i= actual value for the ith observation

ϵ_i= error/residual for the ith observation

n = total number of observations

To get the values of $b_0$and $b_1$which minimise S, we can take a partial derivative for each coefficient and equate it to zero.

Part III : statsmodels.regression.linear_model.OLS

What **R²** means

Proportion of variance in the outcome explained by the model (0–1).

Higher R² ⇒ better in-sample fit (for the same outcome and dataset).

OLS Regression Results

Dep. Variable:	value_eur	R-squared:	0.452
Model:	OLS	Adj. R-squared:	0.452
Method:	Least Squares	F-statistic:	4038.
Date:	Mon, 03 Nov 2025	Prob (F-statistic):	0.00
Time:	16:50:50	Log-Likelihood:	-82386.
No. Observations:	4902	AIC:	1.648e+05
Df Residuals:	4900	BIC:	1.648e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.674e+07	6.28e+05	-58.534	0.000	-3.8e+07	-3.55e+07
overall	5.991e+05	9428.092	63.542	0.000	5.81e+05	6.18e+05

Omnibus:	5334.671	Durbin-Watson:	0.117
Prob(Omnibus):	0.000	Jarque-Bera (JB):	543549.845
Skew:	5.433	Prob(JB):	0.00
Kurtosis:	53.429	Cond. No.	607.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Part III : statsmodels.regression.linear_model.OLS

What the **p-value (F-statistic)** means

Tests whether the model as a whole explains a significant amount of variance.

Null hypothesis: all regression coefficients = 0 (no predictive value).

A small p-value (e.g., < 0.05) ⇒ at least one predictor significantly improves model fit.

OLS Regression Results

Dep. Variable:	value_eur	R-squared:	0.452
Model:	OLS	Adj. R-squared:	0.452
Method:	Least Squares	F-statistic:	4038.
Date:	Mon, 03 Nov 2025	Prob (F-statistic):	0.00
Time:	16:50:50	Log-Likelihood:	-82386.
No. Observations:	4902	AIC:	1.648e+05
Df Residuals:	4900	BIC:	1.648e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.674e+07	6.28e+05	-58.534	0.000	-3.8e+07	-3.55e+07
overall	5.991e+05	9428.092	63.542	0.000	5.81e+05	6.18e+05

Omnibus:	5334.671	Durbin-Watson:	0.117
Prob(Omnibus):	0.000	Jarque-Bera (JB):	543549.845
Skew:	5.433	Prob(JB):	0.00
Kurtosis:	53.429	Cond. No.	607.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Part III : model comparison

How to **compare** two OLS models

1. Compare **R²** – higher means better overall fit.
2. Check **Prob (F-statistic)** – must be < 0.05 for the model to be statistically significant.

This model is the winner!

OLS Regression Results

Dep. Variable:	log_value_eur	R-squared:	0.890
Model:	OLS	Adj. R-squared:	0.890
Method:	Least Squares	F-statistic:	3.973e+04
Date:	Mon, 03 Nov 2025	Prob (F-statistic):	0.00
Time:	16:50:50	Log-Likelihood:	669.09
No. Observations:	4902	AIC:	-1334.
Df Residuals:	4900	BIC:	-1321.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5053	0.028	18.373	0.000	0.451	0.559
overall	0.0823	0.000	199.314	0.000	0.082	0.083

Omnibus:	6866.292	Durbin-Watson:	0.729
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6398339.823
Skew:	-7.659	Prob(JB):	0.00
Kurtosis:	179.328	Cond. No.	607.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS Regression Results

Dep. Variable:	value_eur	R-squared:	0.452
Model:	OLS	Adj. R-squared:	0.452
Method:	Least Squares	F-statistic:	4038.
Date:	Mon, 03 Nov 2025	Prob (F-statistic):	0.00
Time:	16:50:50	Log-Likelihood:	-82386.
No. Observations:	4902	AIC:	1.648e+05
Df Residuals:	4900	BIC:	1.648e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.674e+07	6.28e+05	-58.534	0.000	-3.8e+07	-3.55e+07
overall	5.991e+05	9428.092	63.542	0.000	5.81e+05	6.18e+05

Omnibus:	5334.671	Durbin-Watson:	0.117
Prob(Omnibus):	0.000	Jarque-Bera (JB):	543549.845
Skew:	5.433	Prob(JB):	0.00
Kurtosis:	53.429	Cond. No.	607.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS Regression Results

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              potential   R-squared:                       0.009
Model:                            OLS   Adj. R-squared:                  0.009
Method:                 Least Squares   F-statistic:                     32.28
Date:                Wed, 05 Nov 2025   Prob (F-statistic):           7.18e-33
Time:                        22:24:18   Log-Likelihood:                -55680.
No. Observations:               17244   AIC:                         1.114e+05
Df Residuals:                   17238   BIC:                         1.114e+05
Df Model:                           5
Covariance Type:            nonrobust
=============================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
Intercept                                72.7900      0.173    419.910      0.000      72.450      73.130
side[T.Left]                             -1.2958      0.216     -5.997      0.000      -1.719      -0.872
side[T.Right]                            -1.0691      0.307     -3.487      0.000      -1.670      -0.468
preferred_foot[T.Right]                  -1.0401      0.190     -5.476      0.000      -1.412      -0.668
side[T.Left]:preferred_foot[T.Right]      1.6812      0.263      6.392      0.000       1.166       2.197
side[T.Right]:preferred_foot[T.Right]     0.0619      0.329      0.188      0.850      -0.582       0.706
==============================================================================
Omnibus:                      179.576   Durbin-Watson:                   0.039
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              185.092
Skew:                           0.254   Prob(JB):                     6.43e-41
Kurtosis:                       2.981   Cond. No.                         15.5
==============================================================================
```

Section Materials

Section materials can be accessed at:
https://github.com/JasonC1217/COGS108_FA25_B07-B08

THANKS!

Questions on EdStem or office hours

Office hours: Tue, 3-5 PM