dm25s1

Topic 07 : Regression1

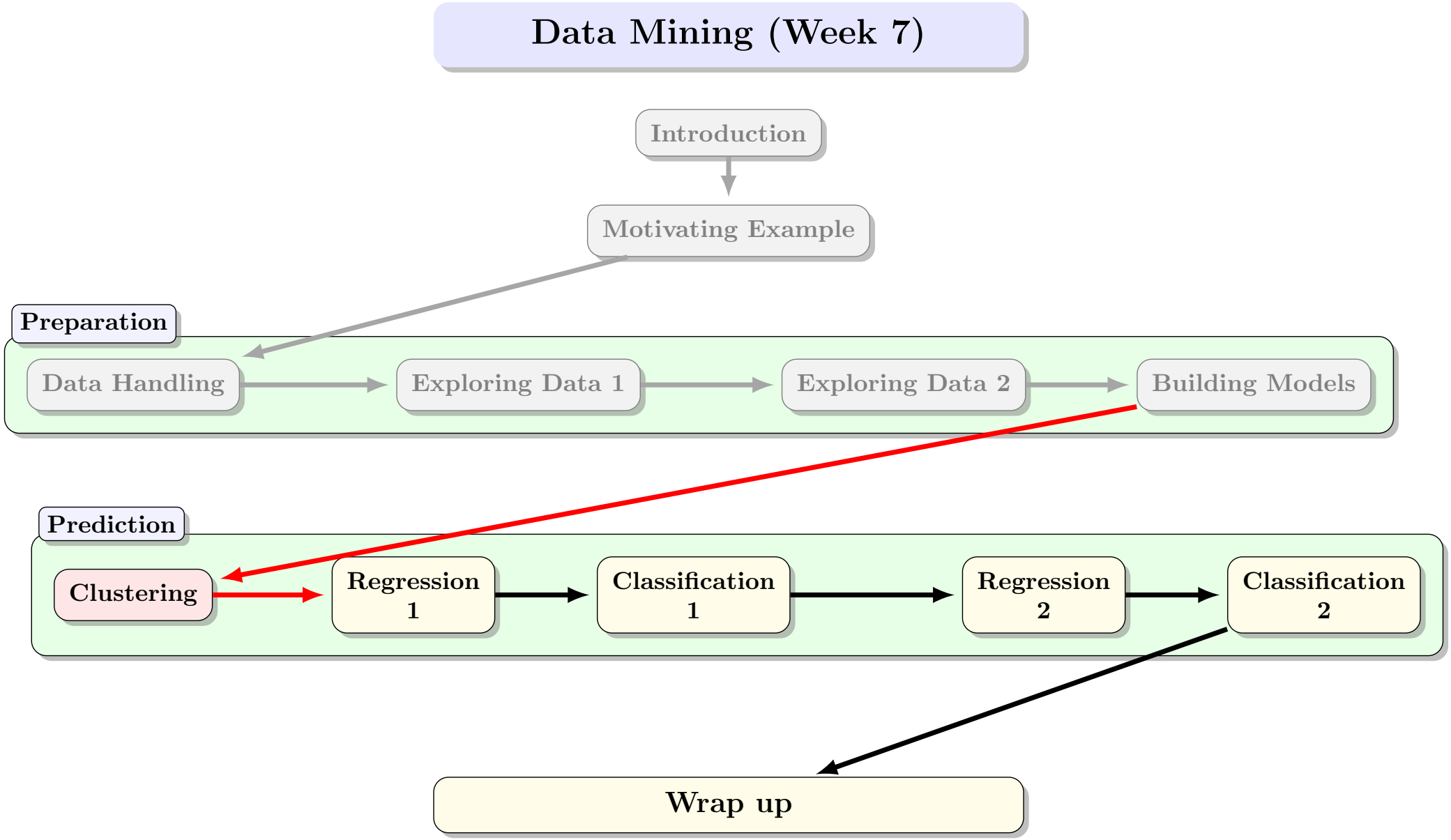
Part 02 : Loss Functions

Dr Bernard Butler
Department of Computing and Mathematics, WIT.
(bernard.butler@setu.ie)

Autumn Semester, 2025

Outline

- Regression loss functions - not just MSE!
- How regression can deal with observations that are outliers
- Regression summary and diagnostic metrics
- Diamond sales case study - interpretation of results


Data Mining (Week 7)
Introduction
Motivating Example
Preparation
Data Handling
Exploring Data 1
Exploring Data 2
Building Models
Prediction
Clustering
Regression 1
Classification 1
Regression 2
Classification 2
Wrap up

Loss Functions — Summary

Common Cost Functions in Regression Models

Remember: we are trying to minimise a loss function based on the error, which we approximate with the residuals of the training set.

Measure	Definition	Purpose
Mean square error (MSE)	$\frac{(p_1-a_1)^2+\cdots+(p_m-a_m)^2}{m}$	Mathematically tractable but places greater emphasise on observations with large error
Root mean square error (RMSE)	$\sqrt{\frac{(p_1-a_1)^2+\cdots+(p_m-a_m)^2}{m}}$	Has same units as data
Mean absolute error (RMAE)	$\frac{\lvert p_1-a_1\rvert+\cdots+\lvert p_m-a_m\rvert}{m}$	Does not overemphasise observations with large error (like MSE does)
Relative square error (RSE)	$\frac{(p_1-a_1)^2+\cdots+(p_m-a_m)^2}{(p_1-\bar{a})^2+\cdots+(p_m-\bar{a})^2}$	Relative metric compares the error in the predictions with errors in the simplest model possible (a model just always predicting the average value of y)
Root Relative square error (RRSE)	$\sqrt{\frac{(p_1-a_1)^2+\cdots+(p_m-a_m)^2}{(p_1-\bar{a})^2+\cdots+(p_m-\bar{a})^2}}$	
Relative absolute error (RAE)	$\frac{\lvert p_1-a_1\rvert+\cdots+\lvert p_m-a_m\rvert}{\lvert p_1-\bar{a}\rvert+\cdots+\lvert p_m-\bar{a}\rvert}$	

where a_j is the actual value, p_j is the predicted value, m is the number of observations, and $\bar{a}$ represents the mean of the a_j.

Choices of Vector norms

Definition 1 (Manhattan norm)

$\ell_1(\ldots) = \|\ldots\|_1$ is the *Manhattan* norm (length) of a vector. Let $\mathbf{x} = (x_1, x_2, \ldots, x_m)$. Then $\ell_1(\ldots) = \|\ldots\|_1 = |x_1| + |x_2| + \ldots + |x_m|$ is the *Manhattan* distance of $\mathbf{x}$ from the origin. Think of having to *walk* from one junction in Manhattan to another, the distance is the difference in the Street numbers plus the difference in the Avenue numbers.

Definition 2 (Euclidean norm)

$\ell_2(\ldots) = \|\ldots\|_2$ is the *Euclidean* norm (length) of a vector. Let $\mathbf{x} = (x_1, x_2, \ldots, x_m)$. Then $\ell_2(\ldots) = \|\ldots\|_2 = \sqrt{x_1^2 + x_2^2 + \ldots + x_m^2}$ is the *Euclidean* distance of $\mathbf{x}$ from the origin. Think of being able to *fly* over all the buildings using the shortest route (think: Pythagoras theorem!) from one junction in Manhattan to another.

The Euclidean norm is very common, but the Manhattan norm is gaining popularity, because it is robust to outliers and computers are becoming powerful enough. However we generally use Euclidean norm in this module.

Sidebar: Distance Measures for numeric data

Definition 3 (Minkowski $p-$norm)

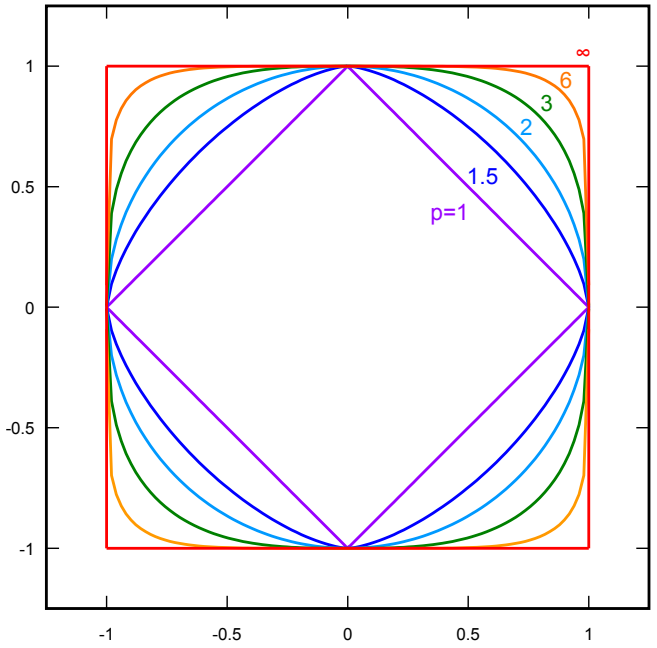
For a real number $1 \leq p < \infty$, the p-norm of $\mathbf{x}$ is defined by

$$\|\mathbf{x}\|_p \equiv \left(|x_1|^p + |x_2|^p + \ldots + |x_n|^p\right)^{\frac{1}{p}}.$$

The limiting case of $p = \infty$ is defined as

$$\|\mathbf{x}\|_\infty \equiv \max\{|x_1|, |x_2|, \ldots, |x_n|\}.$$

See the visualisation of the "unit balls" alongside, for $p = 1, 1.5, 2, 3, 6, \infty$.

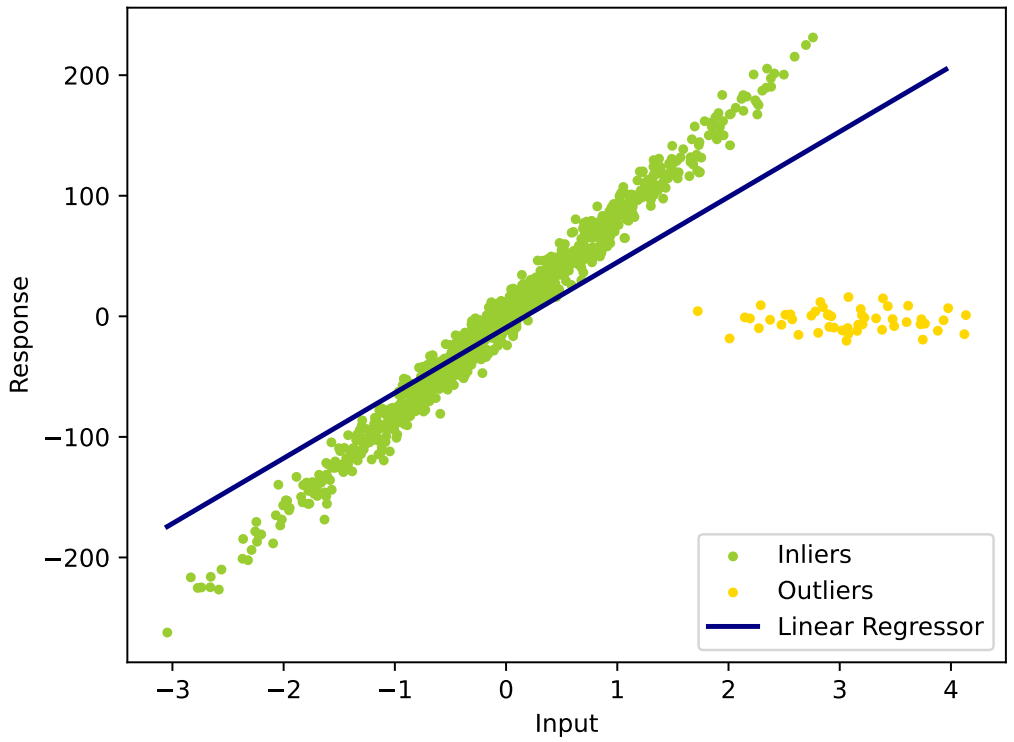


Source: wikipedia

The most common norms are when $p = 1, 2, \text{or}, \infty$. Choice of p depends on the application scenario. Can you think of when you would use each?

Dealing with outliers

The standard loss function (MSE) for linear regression squares the errors

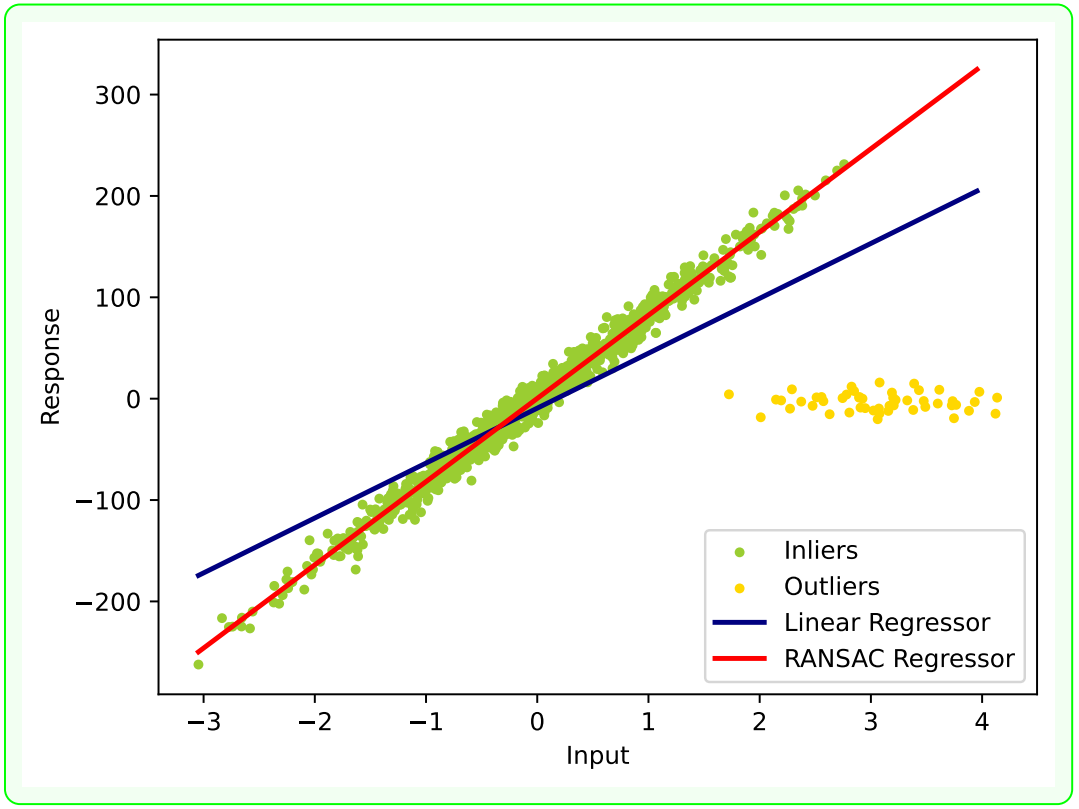


- The generated data has 950 inliers (green) and 50 outliers (yellow)
- The navy line indicates the model fitted using linear regression, with a mean-squared-error loss function.
- Since errors are squared, large errors introduced by the outliers are dominant.
- So the solver tries to reduce their effect when minimising the loss function.
- This means sacrificing the fit to the inliers so that the errors at the outliers are reduced. . .

Approach 1: Ignoring the Outliers

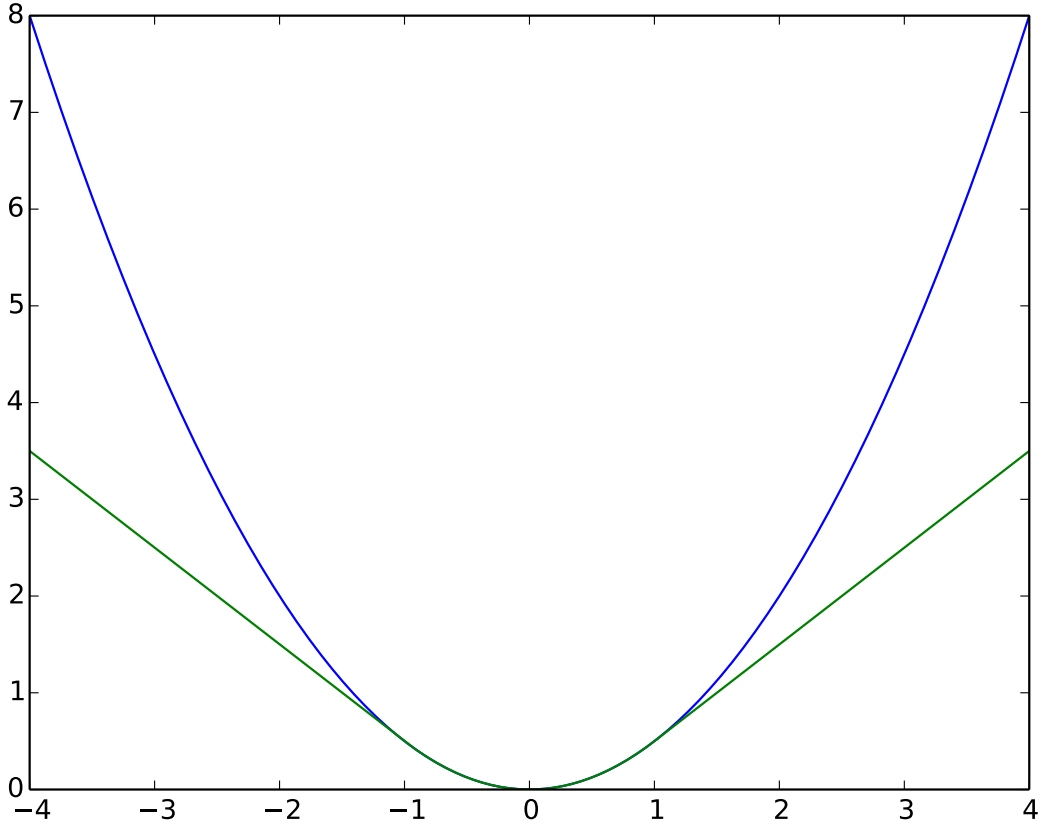
RANSAC applies linear regression to the inliers only

- RANSAC (RANdom SAmple Consensus) searches for a homogeneous set of inliers
- The search is nondeterministic, because it depends on the random seed value
- The goal is to find a representative subset of the inliers, excluding the outliers
- Linear regression is used to fit this subset, using the mean-squared-error loss function
- As can be seen, RANSAC fits the inliers very well by ignoring the outliers



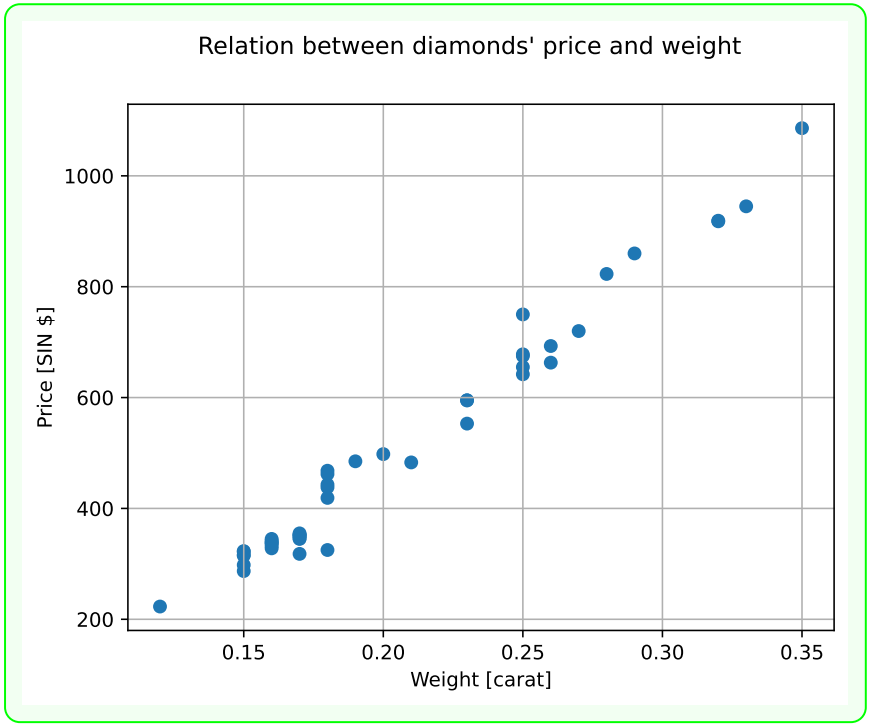
Approach 2: Adjusting the loss function

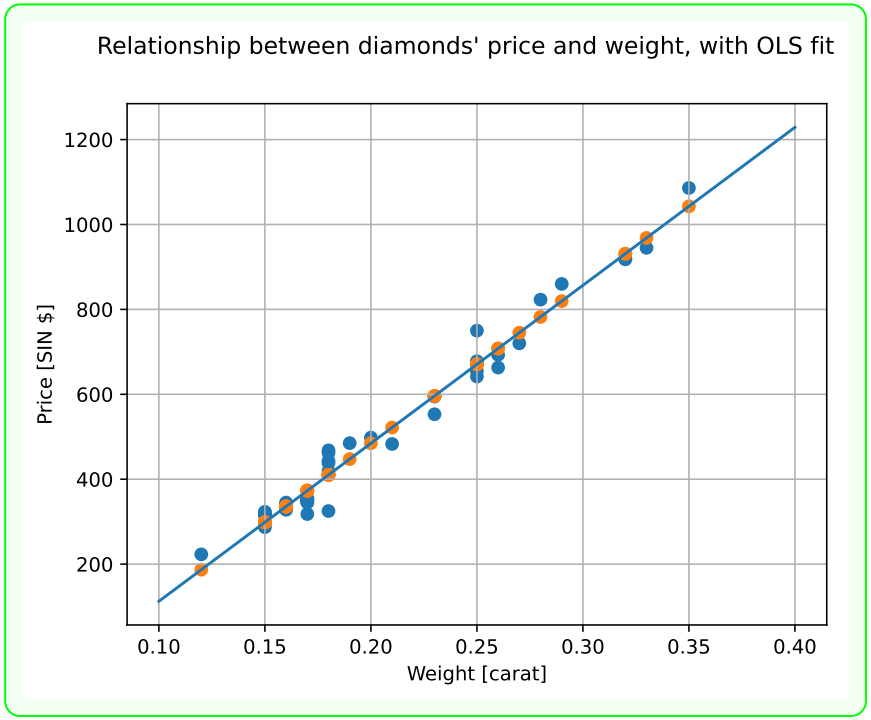
The Huber loss function varies quadratically with small errors and linearly with large errors



- Square-Error loss (Blue) and Huber loss (green) are computed from errors ranging from -4 to 4.
- Here, the Huber parameter $\alpha = 1$, where quadratic loss for $|e| < \alpha$ becomes linear loss for $|e| > \alpha$.
- The `HuberRegressor` uses α to control its sensitivity to outliers
- If α is relatively small, the Huber loss looks very like mean-absolute-error, and hence the RANSAC fit.
- Otherwise, the Huber loss looks more like mean-squared-error and the Linear Regression fit.

Case Study 2: Diamonds - Check relationship



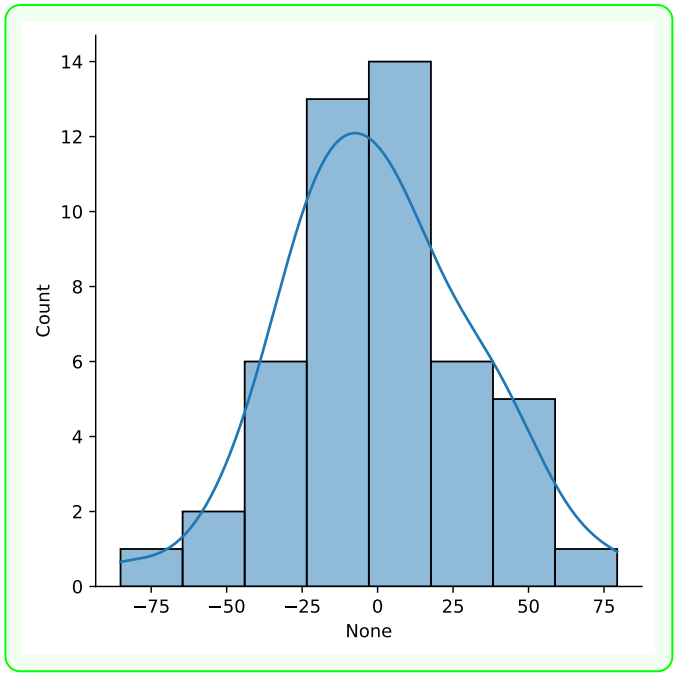


Clearly there is a linear relationship between a diamond's weight (in carats) and its price (in Singapore dollars, as here). So that is one assumption satisfied!

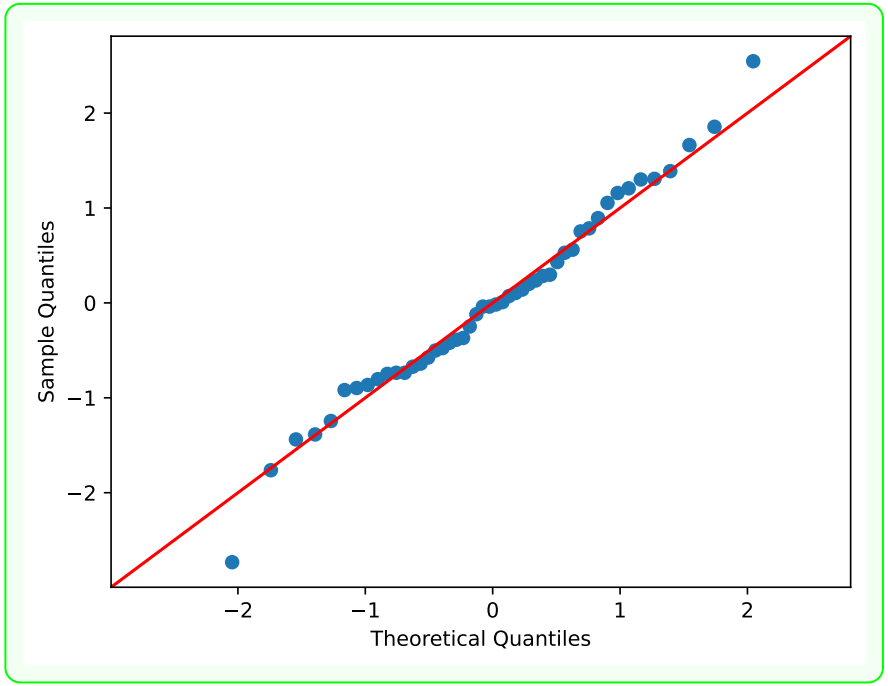
Sometimes the dependent variable has a linear dependence on a (computed) feature that is a function of one or more feature columns in the data. Example functions include log, exp, sqrt, polynomial, etc. Even if the function is nonlinear in the feature x (e.g., x^2, $\sqrt{x}$ or $\log(x)$), that does not matter, as long as the model is linear in the regression parameters β.

Case Study 2: Diamonds - Check residual distribution

```
import seaborn as sns
plot = sns.displot(x = residuals, kde=True)
#resFig = "res/residHist.pdf"
#plot.savefig(resFig)
plt.show()
```




```
# Q-Q plot to verify the residuals distribution
plot = sm.qqplot(residuals, fit=True, line = '45')
#resFig = "res/residualsqq.pdf"
#fig.savefig(resFig)
plt.show()
```



Both diagnostic plots indicate the residuals are reasonably close to Normal distribution centred on 0. Looking good so far!

Is the standardised residual distribution heavy-tailed or light-tailed relative to the Normal distribution? Any other features?

Case Study 2: Diamonds - model summary

Dep. Variable:	price	**R-squared:**	0.978
Model:	OLS	**Adj. R-squared:**	0.978
Method:	Least Squares	**F-statistic:**	2070.
Date:		**Prob (F-statistic):**	6.75e-40
Time:		**Log-Likelihood:**	-233.20
No. Observations:	48	**AIC:**	470.4
Df Residuals:	46	**BIC:**	474.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-259.6259	17.319	-14.991	0.000	-294.487	-224.765
carats	3721.0249	81.786	45.497	0.000	3556.398	3885.651

Omnibus:	0.739	**Durbin-Watson:**	1.994
Prob(Omnibus):	0.691	**Jarque-Bera (JB):**	0.181
Skew:	0.056	**Prob(JB):**	0.913
Kurtosis:	3.280	**Cond. No.**	18.5

Dep. Variable:	price	**R-squared:**	0.978
Model:	OLS	**Adj. R-squared:**	0.978
Method:	Least Squares	**F-statistic:**	2070.
Date:		**Prob (F-statistic):**	6.75e-40
Time:		**Log-Likelihood:**	-233.20
No. Observations:	48	**AIC:**	470.4
Df Residuals:	46	**BIC:**	474.1
Df Model:	1		

Model summary interpretation - 1

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.978
Model:	OLS	**Adj. R-squared:**	0.978
Method:	Least Squares	**F-statistic:**	2070.
Date:	Tue, 21 Oct 2025	**Prob (F-statistic):**	6.75e-40
Time:	17:26:45	**Log-Likelihood:**	-233.20
No. Observations:	48	**AIC:**	470.4
Df Residuals:	46	**BIC:**	474.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-259.6259	17.319	-14.991	0.000	-294.487	-224.765
carats	3721.0249	81.786	45.497	0.000	3556.398	3885.651

Omnibus:	0.739	**Durbin-Watson:**	1.994
Prob(Omnibus):	0.691	**Jarque-Bera (JB):**	0.181
Skew:	0.056	**Prob(JB):**	0.913
Kurtosis:	3.280	**Cond. No.**	18.5

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

Definition 4 (Dep. variable)

This is synonymous with the *target*, which is `price` (in this dataset).

Definition 5 (Model)

statsmodels uses it here in the sense of *problem formulation*. We wish to solve an Ordinary Least Squares problem (assumes all the regression assumptions are met, so no special treatment was applied).

Definition 6 (No. Observations)

This is the number of rows (also known as instances or cases) in the training set.

Model summary interpretation - 2

Definition 7 (Df Model)

The model has one named feature (carats (weight of the diamond)) and one unnamed feature (constant, independent of carats). df, the number of degrees of freedom counts the named features.

Definition 8 (Df Residuals)

The number of degrees of freedom in the residuals is the number of residuals minus the number of features. A higher value tends to go with smaller model variance.

Definition 9 (Covariance Type)

If residuals have the same variance (homoscedastic), nonrobust covariance (the default) can be used.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.978
Model:	OLS	**Adj. R-squared:**	0.978
Method:	Least Squares	**F-statistic:**	2070.
Date:	Tue, 21 Oct 2025	**Prob (F-statistic):**	6.75e-40
Time:	17:26:45	**Log-Likelihood:**	-233.20
No. Observations:	48	**AIC:**	470.4
Df Residuals:	46	**BIC:**	474.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-259.6259	17.319	-14.991	0.000	-294.487	-224.765
carats	3721.0249	81.786	45.497	0.000	3556.398	3885.651

Omnibus:	0.739	**Durbin-Watson:**	1.994
Prob(Omnibus):	0.691	**Jarque-Bera (JB):**	0.181
Skew:	0.056	**Prob(JB):**	0.913
Kurtosis:	3.280	**Cond. No.**	18.5

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

Model summary interpretation - 3

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.978
Model:	OLS	**Adj. R-squared:**	0.978
Method:	Least Squares	**F-statistic:**	2070.
Date:	Tue, 21 Oct 2025	**Prob (F-statistic):**	6.75e-40
Time:	17:26:45	**Log-Likelihood:**	-233.20
No. Observations:	48	**AIC:**	470.4
Df Residuals:	46	**BIC:**	474.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-259.6259	17.319	-14.991	0.000	-294.487	-224.765
carats	3721.0249	81.786	45.497	0.000	3556.398	3885.651

Omnibus:	0.739	**Durbin-Watson:**	1.994
Prob(Omnibus):	0.691	**Jarque-Bera (JB):**	0.181
Skew:	0.056	**Prob(JB):**	0.913
Kurtosis:	3.280	**Cond. No.**	18.5

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

Definition 10 (R-squared)

This is the ratio of the data variance explained by the model, to the variance of the data. It ranges from zero (model explains none of the data variance) to one (model explains all the data variance). A higher value is better, but be careful of overfitting the training set!

Definition 11 (Adj. R-squared)

Similar to R-squared, but it takes account of the number of features. Adding a feature generally increases R-squared, but if the feature did not help as much as its peers, adjusted R-squared shows a smaller increase than “normal” R-squared.

Model summary interpretation - 4

Definition 12 (F-statistic)

Ratio of the variance of a model with just the constant (intercept) feature to the variance of this model. Generally, large values of F are preferred.

Definition 13 (Prob (F statistic))

The value is assumed to follow the F distribution for given *dof*, so can lookup its probability. Small probability indicates that it is highly *unlikely* that the model is doing well purely by chance.

Definition 14 (log likelihood)

OLS is a special case of *maximum likelihood estimation*. Larger likelihood model fits the training data better.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.978
Model:	OLS	**Adj. R-squared:**	0.978
Method:	Least Squares	**F-statistic:**	2070.
Date:	Tue, 21 Oct 2025	**Prob (F-statistic):**	6.75e-40
Time:	17:26:45	**Log-Likelihood:**	-233.20
No. Observations:	48	**AIC:**	470.4
Df Residuals:	46	**BIC:**	474.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-259.6259	17.319	-14.991	0.000	-294.487	-224.765
carats	3721.0249	81.786	45.497	0.000	3556.398	3885.651

Omnibus:	0.739	**Durbin-Watson:**	1.994
Prob(Omnibus):	0.691	**Jarque-Bera (JB):**	0.181
Skew:	0.056	**Prob(JB):**	0.913
Kurtosis:	3.280	**Cond. No.**	18.5

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

Model summary interpretation - 5

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.978
Model:	OLS	**Adj. R-squared:**	0.978
Method:	Least Squares	**F-statistic:**	2070.
Date:	Tue, 21 Oct 2025	**Prob (F-statistic):**	6.75e-40
Time:	17:26:45	**Log-Likelihood:**	-233.20
No. Observations:	48	**AIC:**	470.4
Df Residuals:	46	**BIC:**	474.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-259.6259	17.319	-14.991	0.000	-294.487	-224.765
carats	3721.0249	81.786	45.497	0.000	3556.398	3885.651

Omnibus:	0.739	**Durbin-Watson:**	1.994
Prob(Omnibus):	0.691	**Jarque-Bera (JB):**	0.181
Skew:	0.056	**Prob(JB):**	0.913
Kurtosis:	3.280	**Cond. No.**	18.5

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

Definition 15 (AIC and BIC)

Akaike and Bayesian Information Criterion. These are calculated from the residuals and are derived from *information theory*. They allow for the number of features. Lower values are better.

Definition 16 (Features table: const,carats in this example)

coef is the parameter value for that feature, e.g., const=-259.6 here. $P > |t| = 0$ so it is highly unlikely the coef is zero, given the training data. We also have the 2.5% and 97.5% quantiles, giving the expected range of coef.

Model summary interpretation - 6

Definition 17 (Skew, Kurtosis)

Measures of asymmetry and of peak shape of the residual distribution. Ideal values are 0 (skew) and 3 (kurtosis).

Definition 18 (Durbin-Watson)

Measures the serial correlation of the residuals. Ideal value is 2 (no serial correlation).

Definition 19 (Cond. no)

OLS implementation solves a linear system of equations. Condition number measures column (hence feature independence). Large values mean the features are not independent (they are correlated), making the system more difficult to solve.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.978
Model:	OLS	**Adj. R-squared:**	0.978
Method:	Least Squares	**F-statistic:**	2070.
Date:	Tue, 21 Oct 2025	**Prob (F-statistic):**	6.75e-40
Time:	17:26:45	**Log-Likelihood:**	-233.20
No. Observations:	48	**AIC:**	470.4
Df Residuals:	46	**BIC:**	474.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-259.6259	17.319	-14.991	0.000	-294.487	-224.765
carats	3721.0249	81.786	45.497	0.000	3556.398	3885.651

Omnibus:	0.739	**Durbin-Watson:**	1.994
Prob(Omnibus):	0.691	**Jarque-Bera (JB):**	0.181
Skew:	0.056	**Prob(JB):**	0.913
Kurtosis:	3.280	**Cond. No.**	18.5

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

What does all this mean?

- Clearly, Linear Regression provides many diagnostics to assess how well the model works on the *training set*.
- We can test the model assumptions and many related desirable properties, when applied to the training set.
- However, in machine learning, our goal is to minimise prediction error on unseen data.
- As described last week, we need to perform a train-(validation)-test split and evaluate the errors on the test set.
- *The best model is the one that minimises the error when predicting the target in the test set!*
- The Regression Model diagnostics can help
  1. If a model does not do well, they can help to diagnose the problem(s).
  2. They can also be used constructively, to help identify promising candidate models.

Summary

- We described linear models
- We gave several ways to view what regression is: geometry, linear algebra, optimisation
- We described regression assumptions
- We looked at a simple example
- We consider ways of assessing the success of a regression model, even before checking its performance against the test set
- These quality metrics are not available for other problem formulations, but can help regression model building
- Next time - we return to classification...