dm25s1

Topic 07 : Regression1

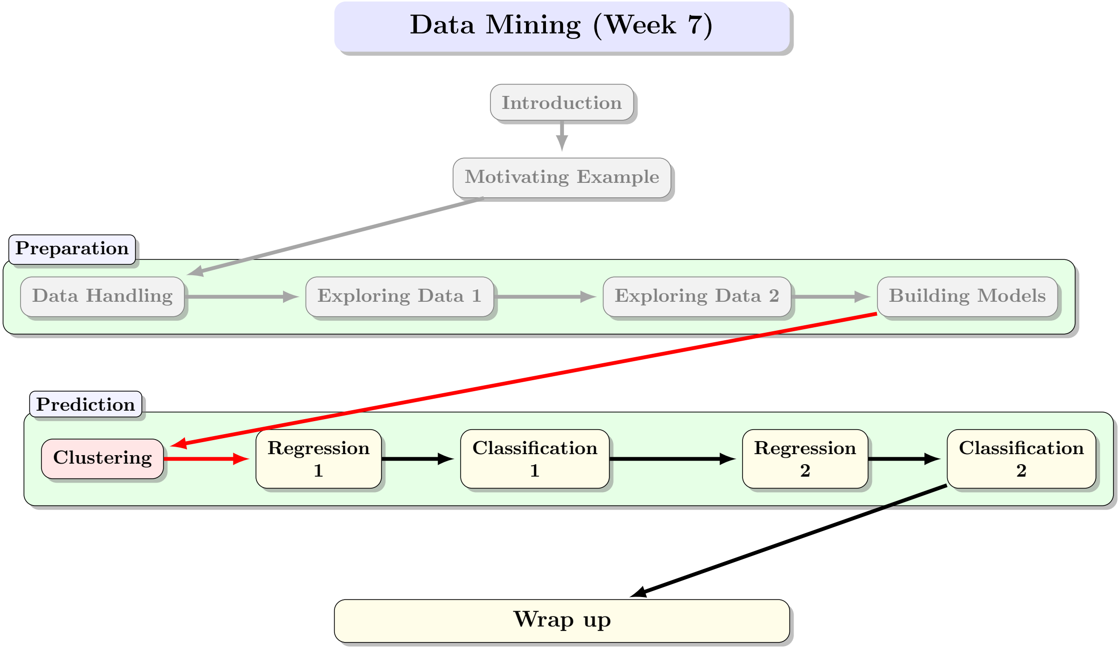
Part 01 : Regression - Overview

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

Autumn Semester, 2025

Outline

- Regression as a means of minimising sum of the squared errors
- Regression assumptions - what they mean, how they can be used for validation and model building
- Role of residuals


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

Regression - Overview — Summary

1. Introduction

2. Linear regression assumptions

3. Reviewing regression results

This Week's Aim

This week's aim is to give an overview of the linear regression: fitting linear models to data, to predict a numeric value.

This Week's Aim

This week's aim is to give an overview of the linear regression: fitting linear models to data, to predict a numeric value.

- High level view of regression: where it came from, what it attempts to do.

This Week's Aim

This week's aim is to give an overview of the linear regression: fitting linear models to data, to predict a numeric value.

- High level view of regression: where it came from, what it attempts to do.
- Examine some extensions to the simplest case of linear regression.

This Week's Aim

This week's aim is to give an overview of the linear regression: fitting linear models to data, to predict a numeric value.

- High level view of regression: where it came from, what it attempts to do.
- Examine some extensions to the simplest case of linear regression.
- Consider how to check that the regression was successful, and make some improvements if necessary

This Week's Aim

This week's aim is to give an overview of the linear regression: fitting linear models to data, to predict a numeric value.

- High level view of regression: where it came from, what it attempts to do.
- Examine some extensions to the simplest case of linear regression.
- Consider how to check that the regression was successful, and make some improvements if necessary
- In this topic, we will cover the basics of Regression - we will return later in the module

This Week's Aim

This week's aim is to give an overview of the linear regression: fitting linear models to data, to predict a numeric value.

- High level view of regression: where it came from, what it attempts to do.
- Examine some extensions to the simplest case of linear regression.
- Consider how to check that the regression was successful, and make some improvements if necessary
- In this topic, we will cover the basics of Regression - we will return later in the module
- To provide context we will use the following datasets over this and Regression 2:

This Week's Aim

This week's aim is to give an overview of the linear regression: fitting linear models to data, to predict a numeric value.

- High level view of regression: where it came from, what it attempts to do.
- Examine some extensions to the simplest case of linear regression.
- Consider how to check that the regression was successful, and make some improvements if necessary
- In this topic, we will cover the basics of Regression - we will return later in the module
- To provide context we will use the following datasets over this and Regression 2:
    - Diamond dataset: predicting diamond prices given their weights
    - Generated data (various)

This Week's Aim

This week's aim is to give an overview of the linear regression: fitting linear models to data, to predict a numeric value.

- High level view of regression: where it came from, what it attempts to do.
- Examine some extensions to the simplest case of linear regression.
- Consider how to check that the regression was successful, and make some improvements if necessary
- In this topic, we will cover the basics of Regression - we will return later in the module
- To provide context we will use the following datasets over this and Regression 2:
    - Diamond dataset: predicting diamond prices given their weights
    - Generated data (various)
    - Advertising dataset: predicting widgets sold based on spending in different advertising channels
    - Credit dataset: predicting credit balance using income, status, etc.

Simple Linear Regression: Background

- Linear regression was discovered by Gauss and others around 1800. The "name" came later!
- With small data sets, calculations can be done by hand, but they are tedious and error-prone.
- The goal is simple: Given a training set of (x, y) data where y is assumed to have a linear relationship with x
  - Find the line that is the "best fit" to that data
  - Use the specification of that line to *predict* y for the test x values
- Note that the "linear relationship" of y upon x is just one of the underlying assumptions
- In practice, the data does not have an exact linear relationship, but it should be "close enough"—but what does that mean?
- In terms of Week 3's ML models taxonomy: regression is geometric and parametric
- In terms of Week 3's Components of a Machine Learning Problem
  - Representation is based on (fitting) hyperplanes to point clouds
  - Evaluation usually based on MSE, with assumption checks to help identify the best model family
  - Optimization is one-step (no search needed) because we have a constraint on the errors we allow
- Hyperparameter tuning: polynomial degree, regularisation λ, weights, loss function, . . .

Review: Linear combinations and scalar products

Definition 1 (Scalar (dot) product of two vectors)

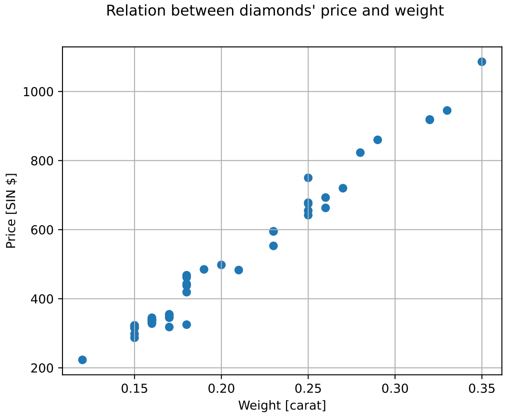
Given two vectors $\mathbf{a}$ and $\mathbf{b}$, each with n elements, the *scalar product* (c) of $\mathbf{a}$ and $\mathbf{b}$ is

$$c \equiv a_1 b_1 + a_2 b_2 + \ldots + a_n b_n = \sum_{i=1}^{n} a_i b_i \equiv |\mathbf{a}||\mathbf{b}| \cos(\mathbf{a}, \mathbf{b})$$

Remarks

- The scalar product of 2 vectors is a scalar, which can be seen as "mixing" two vectors.
- Matrix-vector multiplication $X\mathbf{a}$ can be seen as the scalar product of each row in the matrix X, which is $X(i,:)$ for row i, with the column vector $\mathbf{a}$.
- Alternatively, matrix-vector multiplication can be seen as the *linear combination* of the matrix columns, such as $X(:,j)$, with the column multipliers being the elements of $\mathbf{a}$.
- For linear regression, the matrix columns are the feature vectors $X(j)$ and the column multipliers are the regression parameters $\mathbf{a}$.
- Two nonzero vectors $\mathbf{a}$ and $\mathbf{b}$ can have a scalar product that is zero if $\cos(\mathbf{a}, \mathbf{b}) = 0$, i.e., the $\mathbf{a}$ and $\mathbf{b}$ vectors are perpendicular to each other.

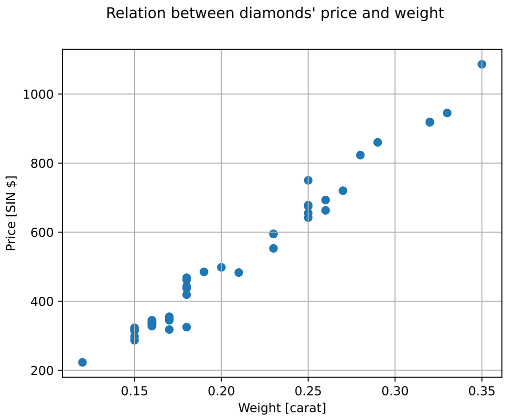
Motivating example: Diamond data



Diamond Prices by Weight

- Given the data on the left, can we use it to predict the price of a diamond that weighs 0.22 carat?
- NB - we have not seen a diamond with that weight before in the data
- Can you think of at least 3 other factors that might affect the price?

Motivating example: Diamond data

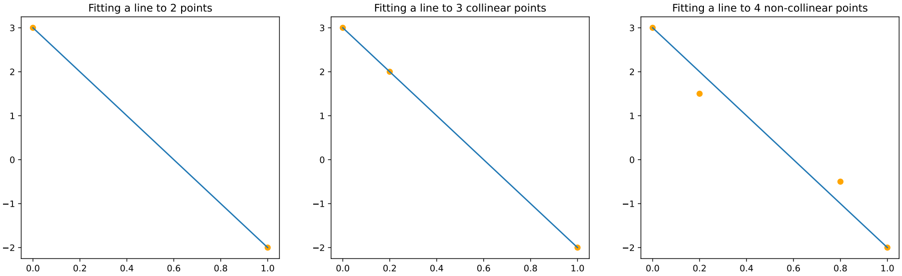


Diamond Prices by Weight

- Given the data on the left, can we use it to predict the price of a diamond that weighs 0.22 carat?
- NB - we have not seen a diamond with that weight before in the data
- Can you think of at least 3 other factors that might affect the price?
- Various(!) - some examples: clarity, cut, provenance, part of a set, . . .

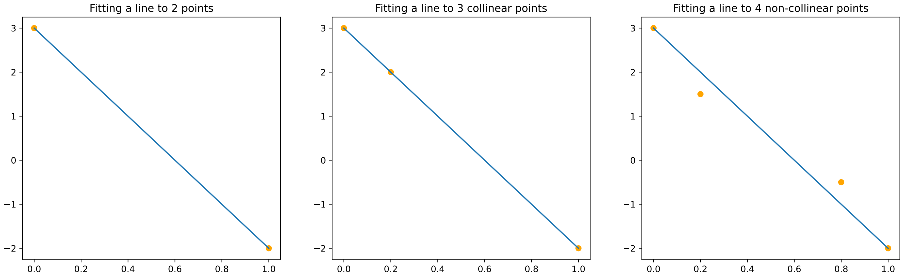
Simple Linear Regression: Geometric Intuition

- Given data $\{x_i, y_i\}$ where $i = 2, 3, \ldots, n$ and β_0, β_1 as the (unknown, but to be determined) *intercept* and *slope* of the regression line for this data.



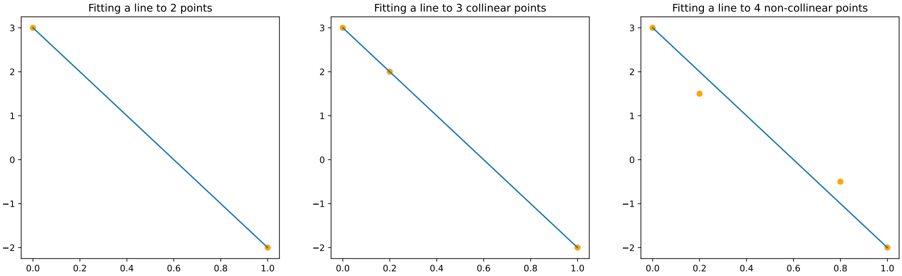
Simple Linear Regression: Geometric Intuition

- Given data $\{x_i, y_i\}$ where $i = 2, 3, \ldots, n$ and β_0, β_1 as the (unknown, but to be determined) *intercept* and *slope* of the regression line for this data.
- If $n = 1$, the problem is underdetermined: any line through the point will do - the solution is not unique.



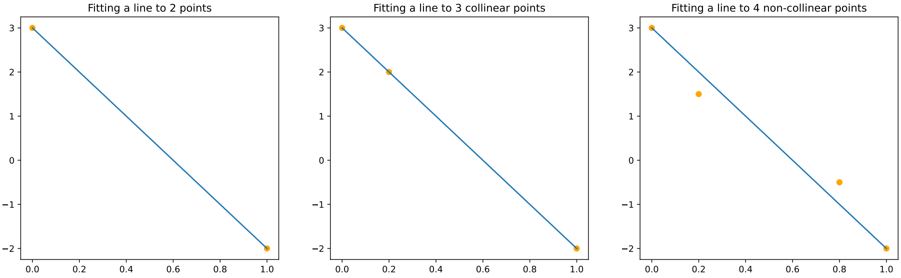
Simple Linear Regression: Geometric Intuition

- Given data $\{x_i, y_i\}$ where $i = 2, 3, \ldots, n$ and β_0, β_1 as the (unknown, but to be determined) *intercept* and *slope* of the regression line for this data.
- If $n = 1$, the problem is underdetermined: any line through the point will do - the solution is not unique.
- For $n = 2$ points with $x_2 \neq x_1$, this can be solved uniquely for β_0, β_1, using techniques you learnt for your Junior/Inter Cert.



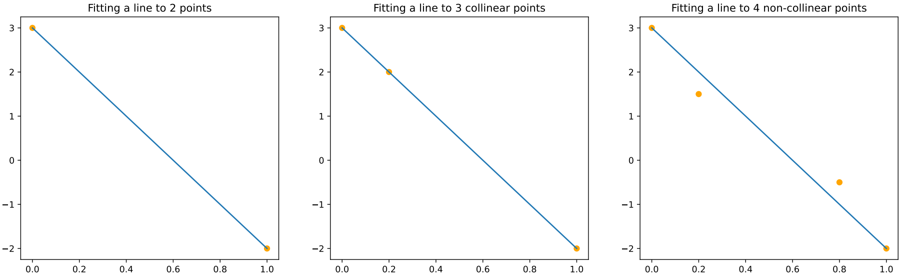
Simple Linear Regression: Geometric Intuition

- Given data $\{x_i, y_i\}$ where $i = 2, 3, \ldots, n$ and β_0, β_1 as the (unknown, but to be determined) *intercept* and *slope* of the regression line for this data.
- If $n = 1$, the problem is underdetermined: any line through the point will do - the solution is not unique.
- For $n = 2$ points with $x_2 \neq x_1$, this can be solved uniquely for β_0, β_1, using techniques you learnt for your Junior/Inter Cert.
- For $n > 2$ collinear points, just pick any two points and solve as before.



Simple Linear Regression: Geometric Intuition

- Given data $\{x_i, y_i\}$ where $i = 2, 3, \ldots, n$ and β_0, β_1 as the (unknown, but to be determined) *intercept* and *slope* of the regression line for this data.
- If $n = 1$, the problem is underdetermined: any line through the point will do - the solution is not unique.
- For $n = 2$ points with $x_2 \neq x_1$, this can be solved uniquely for β_0, β_1, using techniques you learnt for your Junior/Inter Cert.
- For $n > 2$ collinear points, just pick any two points and solve as before.
- Otherwise the problem is overdetermined so need a more general formulation to solve for β_0, β_1.



Simple Linear Regression: Formulation

Definition 2 (Matrix formulation)

- General equation is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \hat{y}_i + \epsilon_i$ (data = model + error), where $\hat{y}$ is the predicted y for these values of β_0, β_1.

Simple Linear Regression: Formulation

Definition 2 (Matrix formulation)

- General equation is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \hat{y}_i + \epsilon_i$ (data = model + error), where $\hat{y}$ is the predicted y for these values of β_0, β_1.
- Matrix form is $\mathbf{y} = X\beta$. Remember matrix-vector multiplication: inner product of i^{th} row of X times the vector β = $1 \times \beta_0 + x_i \times \beta_1 = \hat{y}_i$.

Simple Linear Regression: Formulation

Definition 2 (Matrix formulation)

- General equation is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \hat{y}_i + \epsilon_i$ (data = model + error), where $\hat{y}$ is the predicted y for these values of β_0, β_1.
- Matrix form is $\mathbf{y} = X\beta$. Remember matrix-vector multiplication: inner product of i^{th} row of X times the vector β = $1 \times \beta_0 + x_i \times \beta_1 = \hat{y}_i$.
- However, we don't know β yet, nor do we know $\hat{y}_i$, so we use y_i as an estimate of $\hat{y}_i$ and solve for all data in the training set.

Simple Linear Regression: Formulation

Definition 2 (Matrix formulation)

- General equation is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \hat{y}_i + \epsilon_i$ (data = model + error), where $\hat{y}$ is the predicted y for these values of β_0, β_1.
- Matrix form is $\mathbf{y} = X\beta$. Remember matrix-vector multiplication: inner product of i^{th} row of X times the vector β = $1 \times \beta_0 + x_i \times \beta_1 = \hat{y}_i$.
- However, we don't know β yet, nor do we know $\hat{y}_i$, so we use y_i as an estimate of $\hat{y}_i$ and solve for all data in the training set.
- So: our task is to solve the *overdetermined* (number of rows exceeds the number of columns) system of equations $\mathbf{y} = X\beta$ for β

Simple Linear Regression: Formulation

Definition 2 (Matrix formulation)

- General equation is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \hat{y}_i + \epsilon_i$ (data = model + error), where $\hat{y}$ is the predicted y for these values of β_0, β_1.
- Matrix form is $\mathbf{y} = X\beta$. Remember matrix-vector multiplication: inner product of i^{th} row of X times the vector β = $1 \times \beta_0 + x_i \times \beta_1 = \hat{y}_i$.
- However, we don't know β yet, nor do we know $\hat{y}_i$, so we use y_i as an estimate of $\hat{y}_i$ and solve for all data in the training set.
- So: our task is to solve the *overdetermined* (number of rows exceeds the number of columns) system of equations $\mathbf{y} = X\beta$ for β
- Our geometric intuition is that the errors should be "balanced": no benefit to changing intercept (sliding up or down) or slope (tilting the line).

Simple Linear Regression: Normal Equations

$$
\begin{aligned}
\mathbf{y} &\approx X\beta \\
\mathbf{y} &= X\beta + \epsilon \\
X^{\mathsf{T}}\mathbf{y} &= X^{\mathsf{T}}X\beta + X^{\mathsf{T}}\epsilon \\
X^{\mathsf{T}}\mathbf{y} &= X^{\mathsf{T}}X\beta
\end{aligned}
$$

Simple Linear Regression: Normal Equations

$$
\begin{aligned}
\mathbf{y} &\approx X\beta \\
\mathbf{y} &= X\beta + \epsilon \\
X^{\mathsf{T}}\mathbf{y} &= X^{\mathsf{T}}X\beta + X^{\mathsf{T}}\epsilon \\
X^{\mathsf{T}}\mathbf{y} &= X^{\mathsf{T}}X\beta
\end{aligned}
$$

because $X^{\mathsf{T}}\epsilon \equiv 0$ implies the fitted line gives balanced errors and so is 'best". Swapping sides, we have

$$
\begin{aligned}
(X^{\mathsf{T}}X)\beta &= X^{\mathsf{T}}\mathbf{y} \\
(X^{\mathsf{T}}X)^{-1}(X^{\mathsf{T}}X)\beta &= (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}
\end{aligned}
$$

Simple Linear Regression: Normal Equations

$$\begin{aligned}\mathbf{y} &\approx X\beta \\ \mathbf{y} &= X\beta + \epsilon \\ X^\mathsf{T}\mathbf{y} &= X^\mathsf{T}X\beta + X^\mathsf{T}\epsilon \\ X^\mathsf{T}\mathbf{y} &= X^\mathsf{T}X\beta\end{aligned}$$

because $X^\mathsf{T}\epsilon \equiv 0$ implies the fitted line gives balanced errors and so is ‘best”. Swapping sides, we have

$$\begin{aligned}(X^\mathsf{T}X)\beta &= X^\mathsf{T}\mathbf{y} \\ (X^\mathsf{T}X)^{-1}(X^\mathsf{T}X)\beta &= (X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y}\end{aligned}$$

which is equivalent to the *Normal equations*

$$\beta = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y} \tag{1}$$

Simple Linear Regression: Normal Equations

$$
\begin{aligned}
\mathbf{y} &\approx X\beta \\
\mathbf{y} &= X\beta + \epsilon \\
X^\mathsf{T}\mathbf{y} &= X^\mathsf{T}X\beta + X^\mathsf{T}\epsilon \\
X^\mathsf{T}\mathbf{y} &= X^\mathsf{T}X\beta
\end{aligned}
$$

because $X^\mathsf{T}\epsilon \equiv 0$ implies the fitted line gives balanced errors and so is ‘best”. Swapping sides, we have

$$
\begin{aligned}
(X^\mathsf{T}X)\beta &= X^\mathsf{T}\mathbf{y} \\
(X^\mathsf{T}X)^{-1}(X^\mathsf{T}X)\beta &= (X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y}
\end{aligned}
$$

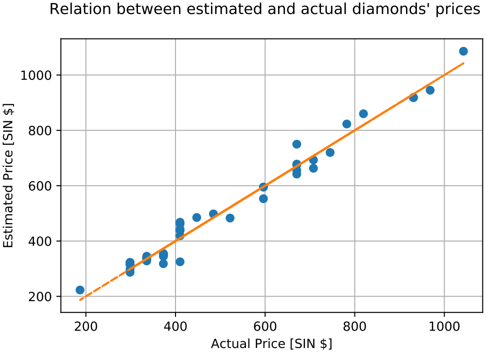
which is equivalent to the *Normal equations*

$$\beta = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y} \tag{1}$$

Note that everything on the right is a set of operations on the data.

For more info, and an alternative construction of the Normal equations, see `https://goo.gl/TbLru3`.

Simple Linear Regression: Balanced Errors



What makes this look like a good fit?
The fitted line passes through the data centroid and errors pass are balanced *- cf. see-saw*

- More generally: a weighted sum of the errors should be 0.
- Weights should depend on the features.
- The $X^\top \epsilon = 0$ criterion works well, so we apply it.
- If you imagine the centroid as being the fulcrum of the line, viewed as a lever, we wish to “balance” the errors around that point

Simple Linear Regression: Implementation

When implemented in software, the Normal equations are not used directly: faster and more numerically accurate algorithms are used instead, but the results are equivalent in exact arithmetic (remember: digital computers perform finite-precision arithmetic and so cannot be exact).

Simple Linear Regression: Implementation

When implemented in software, the Normal equations are not used directly: faster and more numerically accurate algorithms are used instead, but the results are equivalent in exact arithmetic (remember: digital computers perform finite-precision arithmetic and so cannot be exact).

One option is to use statsmodels: consistent with R (separate model specification), excellent diagnostics as standard

Another option is to use sklearn: consistent with other sklearn algorithms, more controls

Simple Linear Regression: Implementation

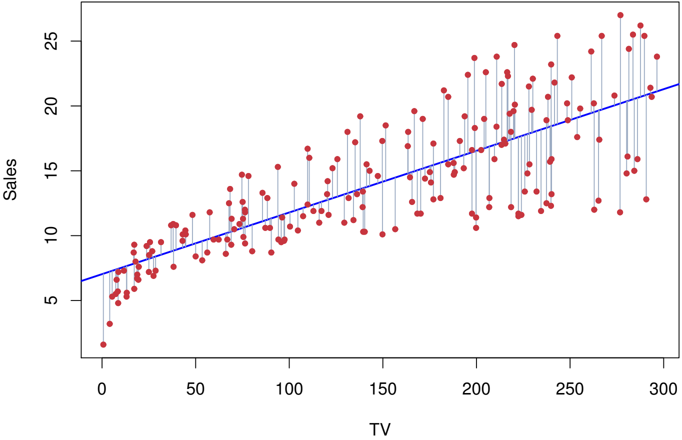
When implemented in software, the Normal equations are not used directly: faster and more numerically accurate algorithms are used instead, but the results are equivalent in exact arithmetic (remember: digital computers perform finite-precision arithmetic and so cannot be exact).

One option is to use statsmodels: consistent with R (separate model specification), excellent diagnostics as standard

Another option is to use sklearn: consistent with other sklearn algorithms, more controls

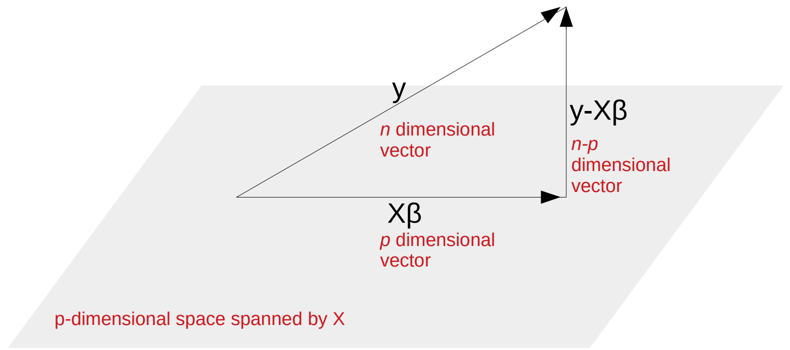
Remember: after *learning* the β parameters using the training data $\{\mathbf{x}_i, y_i\}$, with the model encoded in the feature matrix X, it is then possible to predict $\hat{y}_k$ for “new” (test) $\mathbf{x}_k$ values, using separate *prediction* function calls.

SLR: Residual Plot for the model



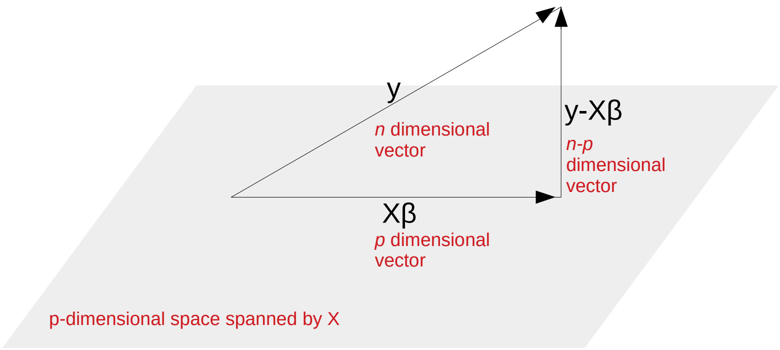
Source: ISLR, Fig 3.1: Advertising data with the model "Sales $\sim$ TV".
Note the vertical distance between the red dots (data points) $\mathbf{y}$ and the corresponding $\hat{\mathbf{y}}$ on the regression line, which is termed the *error* ϵ.

Geometrical interpretation of regression: n rows, p features, $n > p$



- Analogy: achieving photorealism with a limited palette of colours.
- Grey plane represents all the colours mixable from those colours.
- Point above plane: a colour that needs to be approximated.

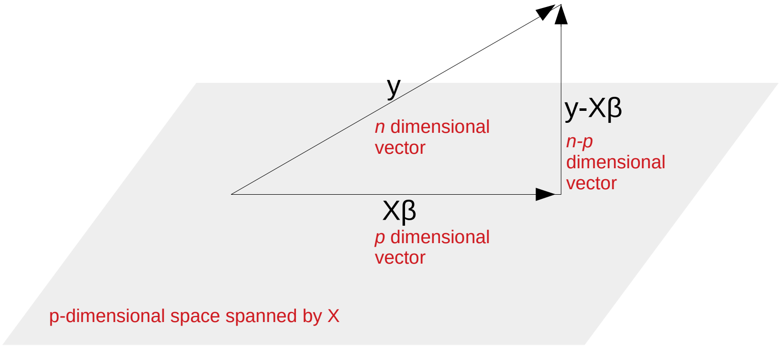
Geometrical interpretation of regression: n rows, p features, $n > p$



- Analogy: achieving photorealism with a limited palette of colours.
- Grey plane represents all the colours mixable from those colours.
- Point above plane: a colour that needs to be approximated.

- The X matrix spans the $p \times p$ space represented by the grey plane.
- But $\mathbf{y}$ has $n > p$ dimensions and so is represented by a point that lies outside the grey plane.
- When $\mathbf{y}$ is projected onto the nearest point in the X space,
  - The projected point is $\hat{\mathbf{y}}$.
  - The residuals (errors) ϵ are $\mathbf{y} - X\beta \equiv \mathbf{y} - \hat{\mathbf{y}}$.

Geometrical interpretation of regression: n rows, p features, $n > p$



- Analogy: achieving photorealism with a limited palette of colours.
- Grey plane represents all the colours mixable from those colours.
- Point above plane: a colour that needs to be approximated.

- The X matrix spans the $p \times p$ space represented by the grey plane.
- But $\mathbf{y}$ has $n > p$ dimensions and so is represented by a point that lies outside the grey plane.
- When $\mathbf{y}$ is projected onto the nearest point in the X space,
  - The projected point is $\hat{\mathbf{y}}$.
  - The residuals (errors) ϵ are $\mathbf{y} - X\beta \equiv \mathbf{y} - \hat{\mathbf{y}}$.

This decomposition of n data dimensions (observations) into p model parameters and n residuals with rank $n - p$ is helpful when interpreting regression diagnostics.

OLS and Linear Regression

Definition 3 (BLUE)

According to the Gauss-Markov theorem, *Ordinary Least Squares* (OLS), which uses the Normal equations to minimise the sum of the squares of the errors ($\|\epsilon\|_2 \equiv \sqrt{e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2}$), is the *Best, Linear, Unbiased, Estimator* of that model that can be derived from the training data, provided some reasonably loose assumptions hold.

OLS and Linear Regression

Definition 3 (BLUE)

According to the Gauss-Markov theorem, *Ordinary Least Squares* (OLS), which uses the Normal equations to minimise the sum of the squares of the errors ($\|\epsilon\|_2 \equiv \sqrt{e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2}$), is the *Best, Linear, Unbiased, Estimator* of that model that can be derived from the training data, provided some reasonably loose assumptions hold.

When we discuss Bias, Variance and Irreducible Error, it is clear that low bias is not enough. OLS might be BLUE but that does not guarantee low variance, because overfitting can still be a problem.

OLS and Linear Regression

Definition 3 (BLUE)

According to the Gauss-Markov theorem, *Ordinary Least Squares* (OLS), which uses the Normal equations to minimise the sum of the squares of the errors ($\|\epsilon\|_2 \equiv \sqrt{e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2}$), is the *Best, Linear, Unbiased, Estimator* of that model that can be derived from the training data, provided some reasonably loose assumptions hold.

When we discuss Bias, Variance and Irreducible Error, it is clear that low bias is not enough. OLS might be BLUE but that does not guarantee low variance, because overfitting can still be a problem.
In practice, the assumptions required for OLS to be appropriate can be stated in terms of properties of the residual vector ϵ.

OLS and Linear Regression

Definition 3 (BLUE)

According to the Gauss-Markov theorem, *Ordinary Least Squares* (OLS), which uses the Normal equations to minimise the sum of the squares of the errors ($\|\epsilon\|_2 \equiv \sqrt{e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2}$), is the *Best, Linear, Unbiased, Estimator* of that model that can be derived from the training data, provided some reasonably loose assumptions hold.

When we discuss Bias, Variance and Irreducible Error, it is clear that low bias is not enough. OLS might be BLUE but that does not guarantee low variance, because overfitting can still be a problem.
In practice, the assumptions required for OLS to be appropriate can be stated in terms of properties of the residual vector ϵ.
In the rest of this lecture, we will generalise from Simple to Multiple Linear Regression, where $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ and $2 \leq p \leq n$, so instead of fitting lines, we fit (hyper)planes to data.

Assumptions required for the linear model to be meaningful

Definition 4 (Linear Regression Assumptions)

1. The underlying relationship between the predictors and the response is linear in the regression parameters β.

Assumptions required for the linear model to be meaningful

Definition 4 (Linear Regression Assumptions)

1. The underlying relationship between the predictors and the response is linear in the regression parameters β.
2. The residual errors ϵ are drawn from a (multivariate) Normal distribution $N(\mu, \sigma^2)$ where $\mu = \mathbf{0}$.

Assumptions required for the linear model to be meaningful

Definition 4 (Linear Regression Assumptions)

1. The underlying relationship between the predictors and the response is linear in the regression parameters β.
2. The residual errors ϵ are drawn from a (multivariate) Normal distribution $N(\mu, \sigma^2)$ where $\mu = \mathbf{0}$.
3. The predictors are not pairwise collinear, i.e., each pair of predictors β_{j_1} and β_{j_2} (associated with columns $X(:, j_1)$ and $X(:, j_2)$) have low correlation (equivalently, the inner product of $X(:, j_1)$ and $X(:, j_2)$ is far from zero).

Assumptions required for the linear model to be meaningful

Definition 4 (Linear Regression Assumptions)

1. The underlying relationship between the predictors and the response is linear in the regression parameters β.
2. The residual errors ϵ are drawn from a (multivariate) Normal distribution $N(\mu, \sigma^2)$ where $\mu = \mathbf{0}$.
3. The predictors are not pairwise collinear, i.e., each pair of predictors β_{j_1} and β_{j_2} (associated with columns $X(:, j_1)$ and $X(:, j_2)$) have low correlation (equivalently, the inner product of $X(:, j_1)$ and $X(:, j_2)$ is far from zero).
4. There is no auto-correlation in $\mathbf{y}$: each observation is independent of its "neighbours".

Assumptions required for the linear model to be meaningful

Definition 4 (Linear Regression Assumptions)

1. The underlying relationship between the predictors and the response is linear in the regression parameters β.
2. The residual errors ϵ are drawn from a (multivariate) Normal distribution $N(\mu, \sigma^2)$ where $\mu = \mathbf{0}$.
3. The predictors are not pairwise collinear, i.e., each pair of predictors β_{j_1} and β_{j_2} (associated with columns $X(:, j_1)$ and $X(:, j_2)$) have low correlation (equivalently, the inner product of $X(:, j_1)$ and $X(:, j_2)$ is far from zero).
4. There is no auto-correlation in $\mathbf{y}$: each observation is independent of its "neighbours".
5. The errors are *homoscedastic* (i.e., $\mathrm{Var}(\epsilon)$ is constant over the range of $\mathbf{x}$ or $\mathbf{y}$).

Assumptions required for the linear model to be meaningful

Definition 4 (Linear Regression Assumptions)

1. The underlying relationship between the predictors and the response is linear in the regression parameters β.
2. The residual errors ϵ are drawn from a (multivariate) Normal distribution $N(\mu, \sigma^2)$ where $\mu = \mathbf{0}$.
3. The predictors are not pairwise collinear, i.e., each pair of predictors β_{j_1} and β_{j_2} (associated with columns $X(:, j_1)$ and $X(:, j_2)$) have low correlation (equivalently, the inner product of $X(:, j_1)$ and $X(:, j_2)$ is far from zero).
4. There is no auto-correlation in $\mathbf{y}$: each observation is independent of its “neighbours”.
5. The errors are *homoscedastic* (i.e., Var(ϵ) is constant over the range of $\mathbf{x}$ or $\mathbf{y}$).

Because these assumptions depend both on the data and on the model fitted to that data, it is meaningless to say that “Data set A does not satisfy the linear regression assumptions”, because this observation might not apply to all formulations of all models applied to that data.

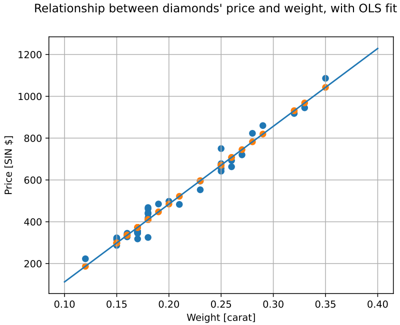
Assumptions required for the linear model to be meaningful

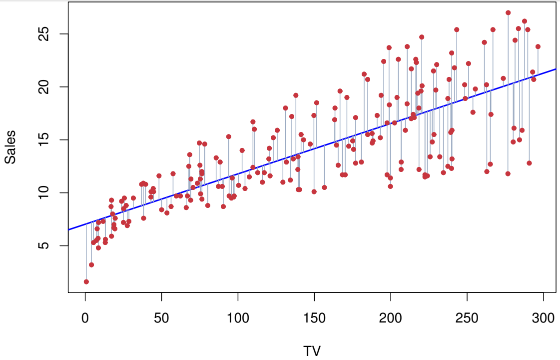
Definition 4 (Linear Regression Assumptions)

1. The underlying relationship between the predictors and the response is linear in the regression parameters β.
2. The residual errors ϵ are drawn from a (multivariate) Normal distribution $N(\mu, \sigma^2)$ where $\mu = \mathbf{0}$.
3. The predictors are not pairwise collinear, i.e., each pair of predictors β_{j_1} and β_{j_2} (associated with columns $X(:, j_1)$ and $X(:, j_2)$) have low correlation (equivalently, the inner product of $X(:, j_1)$ and $X(:, j_2)$ is far from zero).
4. There is no auto-correlation in $\mathbf{y}$: each observation is independent of its "neighbours".
5. The errors are *homoscedastic* (i.e., Var(ϵ) is constant over the range of $\mathbf{x}$ or $\mathbf{y}$).

Because these assumptions depend both on the data and on the model fitted to that data, it is meaningless to say that "Data set A does not satisfy the linear regression assumptions", because this observation might not apply to all formulations of all models applied to that data.
Consequently, these assumptions can be used constructively, when model building, or as checks, when validating models.

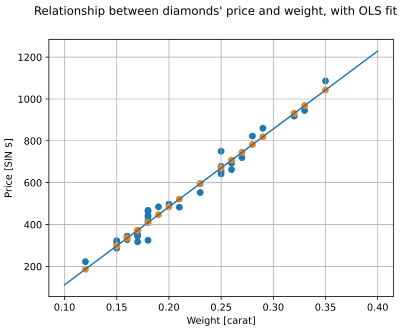
Linear relationship

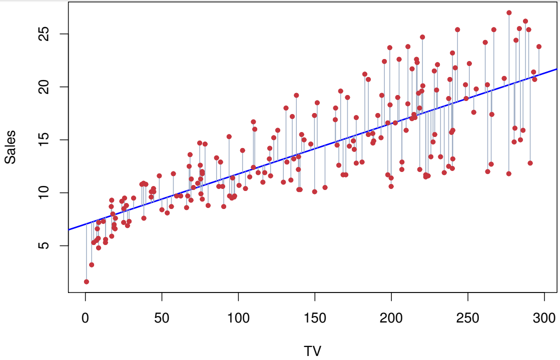




- In both cases, the relationship between predictor (feature) and target is approximately linear.

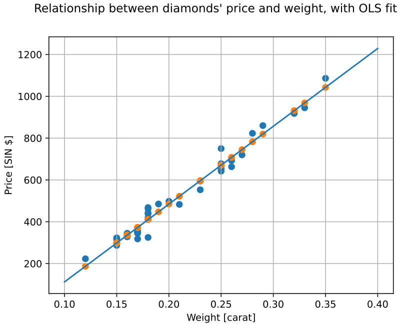
Linear relationship

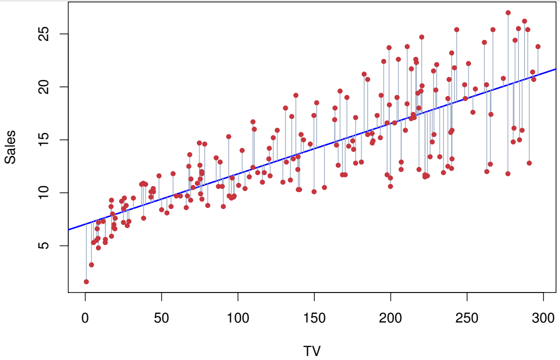




- In both cases, the relationship between predictor (feature) and target is approximately linear.
- Given a feature value, we can predict the target value using a simple linear formula.

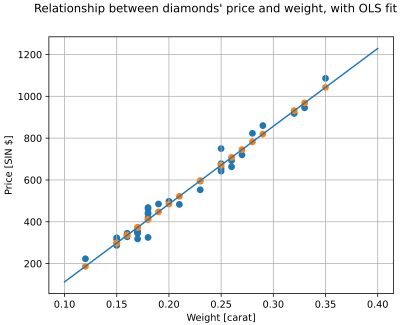
Linear relationship

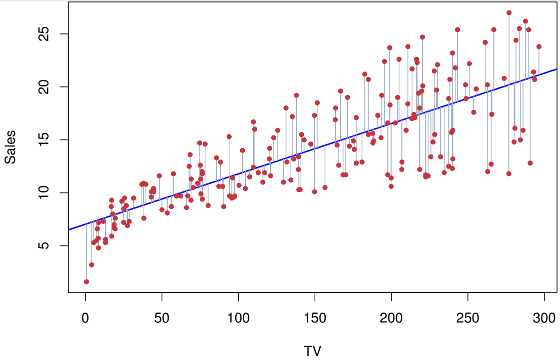




- In both cases, the relationship between predictor (feature) and target is approximately linear.
- Given a feature value, we can predict the target value using a simple linear formula.
- The predicted parameters are the *intercept* β_0 and *slope* β_1 of the line.

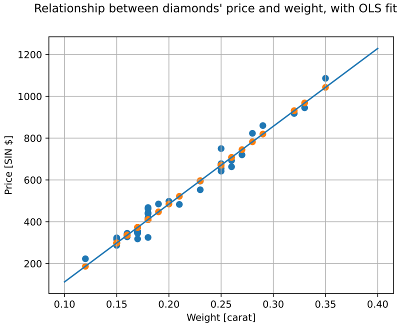
Linear relationship

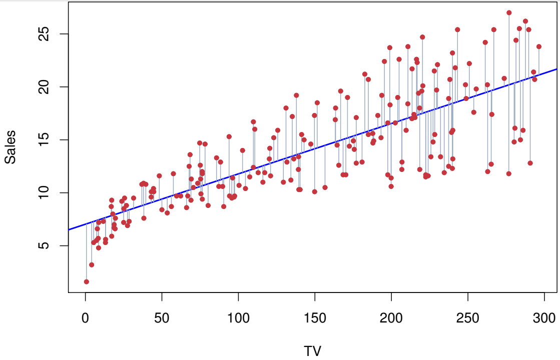




- In both cases, the relationship between predictor (feature) and target is approximately linear.
- Given a feature value, we can predict the target value using a simple linear formula.
- The predicted parameters are the *intercept* β_0 and *slope* β_1 of the line.
- Usually the vertical distance between a data point x_i and its predicted value $\hat{y}_i$ is $\epsilon_i \neq 0$.

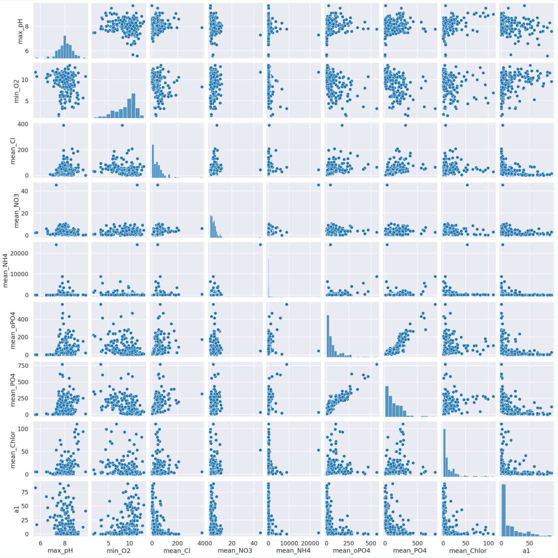
Linear relationship





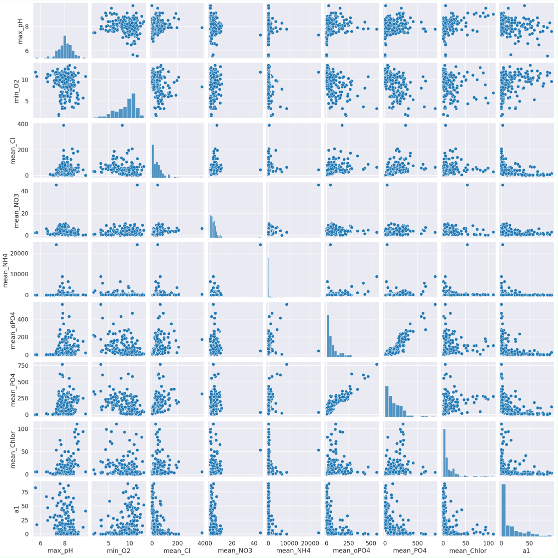
- In both cases, the relationship between predictor (feature) and target is approximately linear.
- Given a feature value, we can predict the target value using a simple linear formula.
- The predicted parameters are the *intercept* β_0 and *slope* β_1 of the line.
- Usually the vertical distance between a data point x_i and its predicted value $\hat{y}_i$ is $\epsilon_i \neq 0$.
- ϵ_i is the *residual error*. It quantifies data behaviour not included in our model.

Collinearity (high pairwise correlation) among the algae bloom predictors



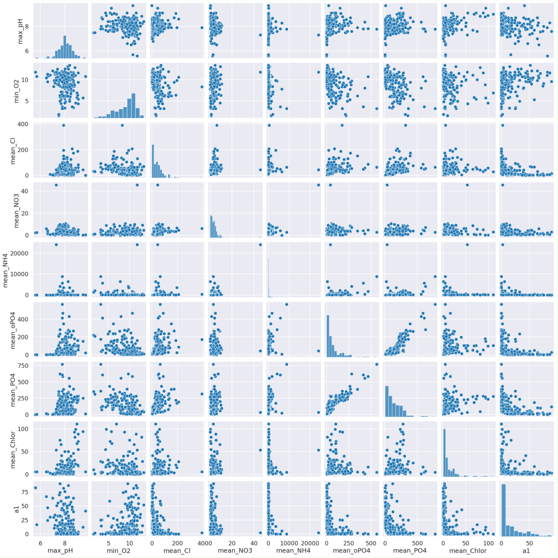
- The pairplot confirms what we saw in the corresponding correlation matrix: `mean_PO4` and `mean_OPO4` are highly correlated with each other (indeed, the relevant scatterplots indicate a strong linear relationship).

Collinearity (high pairwise correlation) among the algae bloom predictors



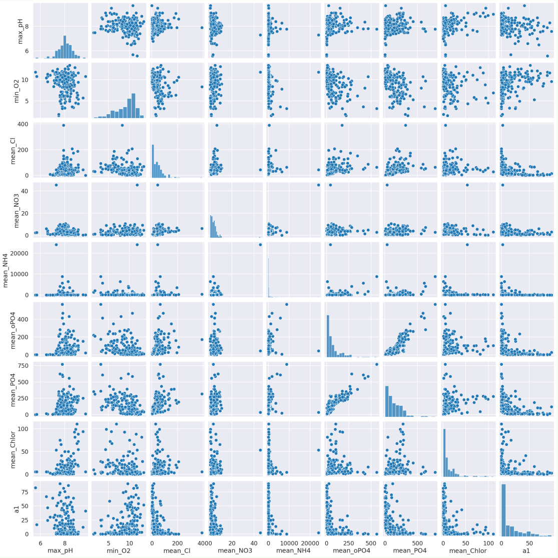
- The pairplot confirms what we saw in the corresponding correlation matrix: `mean_PO4` and `mean_OPO4` are highly correlated with each other (indeed, the relevant scatterplots indicate a strong linear relationship).
- Either `mean_PO4` or `mean_OPO4` can be included in the model, but not both of them.

Collinearity (high pairwise correlation) among the algae bloom predictors



- The pairplot confirms what we saw in the corresponding correlation matrix: `mean_PO4` and `mean_OPO4` are highly correlated with each other (indeed, the relevant scatterplots indicate a strong linear relationship).
- Either `mean_PO4` or `mean_OPO4` can be included in the model, but not both of them.
- Also, the individual predictors do not have a strong linear relationship with `a1` (look at the scatterplots in the last row and column) so, on their own, they are not likely to predict `a1` well with a linear model.

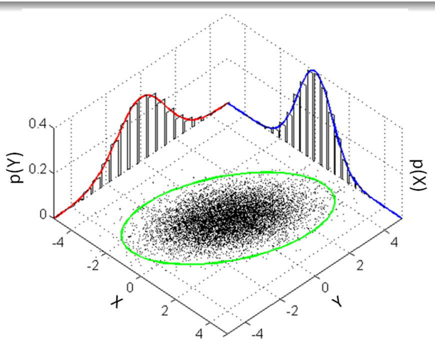
Collinearity (high pairwise correlation) among the algae bloom predictors

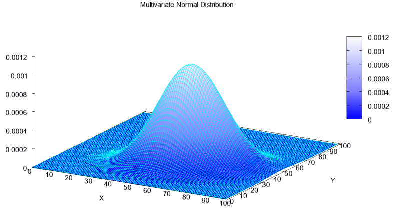


- The pairplot confirms what we saw in the corresponding correlation matrix: `mean_PO4` and `mean_OPO4` are highly correlated with each other (indeed, the relevant scatterplots indicate a strong linear relationship).
- Either `mean_PO4` or `mean_OPO4` can be included in the model, but not both of them.
- Also, the individual predictors do not have a strong linear relationship with `a1` (look at the scatterplots in the last row and column) so, on their own, they are not likely to predict `a1` well with a linear model.
- However, it is still possible that a combination of predictors might predict a1 well.

Errors are normally distributed

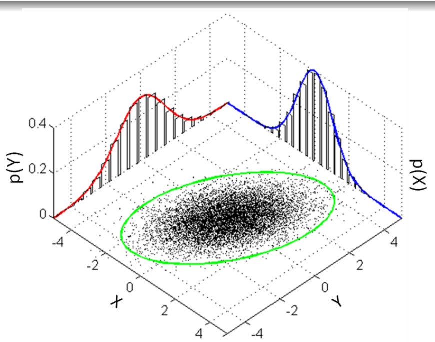
- Centred on zero so small errors are more common

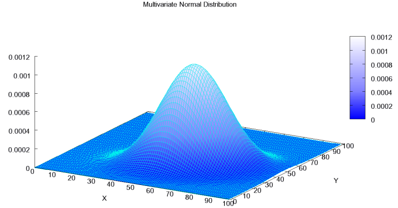




Errors are normally distributed

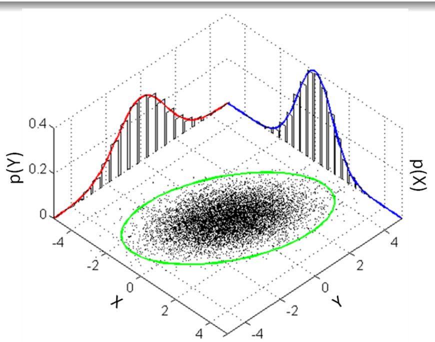
- Centred on zero so small errors are more common
- Symmetric so positive and negative balance out

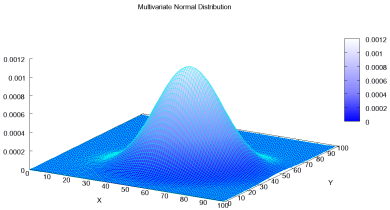




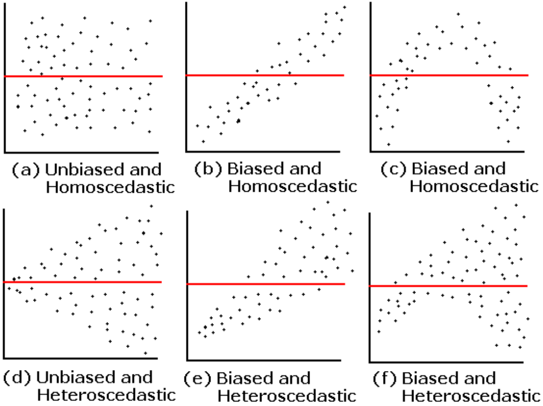
Errors are normally distributed

- Centred on zero so small errors are more common
- Symmetric so positive and negative balance out
- Normal distribution is a also called the Gaussian distribution and is “bell-shaped”.





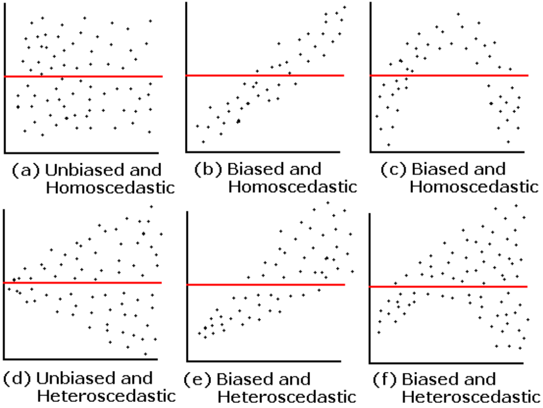
Bias and variance in regression



Source: https://bit.ly/3vC9zK7

- Bias is caused by underfitting.

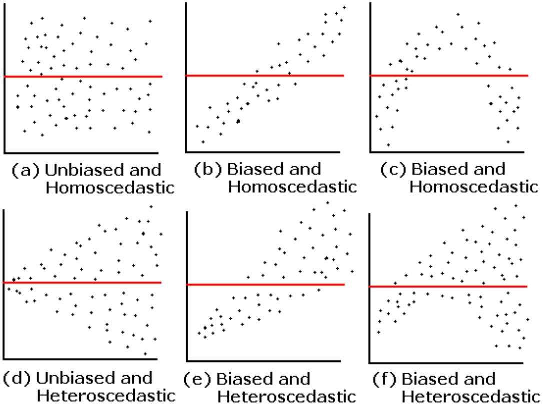
Bias and variance in regression



Source: `https://bit.ly/3vC9zK7`

- Bias is caused by underfitting.
- Fix bias by adding suitable predictors.

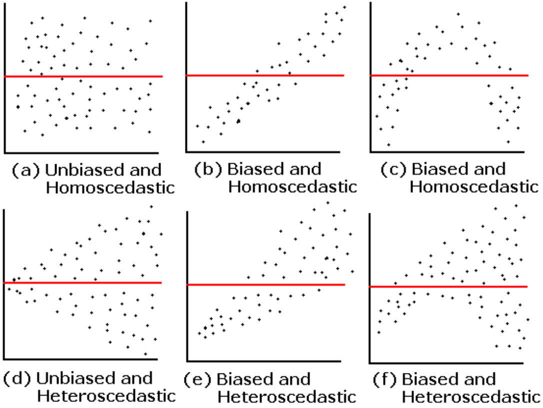
Bias and variance in regression



Source: `https://bit.ly/3vC9zK7`

- Bias is caused by underfitting.
- Fix bias by adding suitable predictors.
- Overfitting causes large variance.

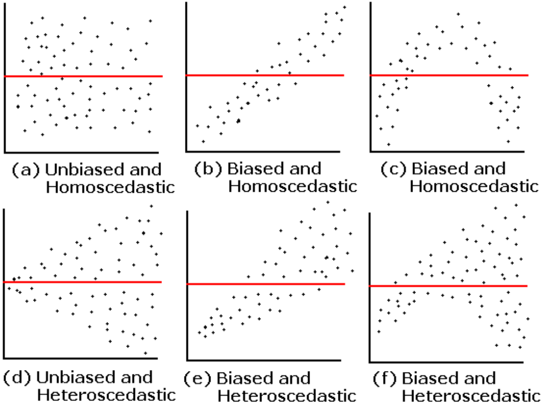
Bias and variance in regression



Source: `https://bit.ly/3vC9zK7`

- Bias is caused by underfitting.
- Fix bias by adding suitable predictors.
- Overfitting causes large variance.
- If variance changes over the range, some errors get undue attention.

Bias and variance in regression

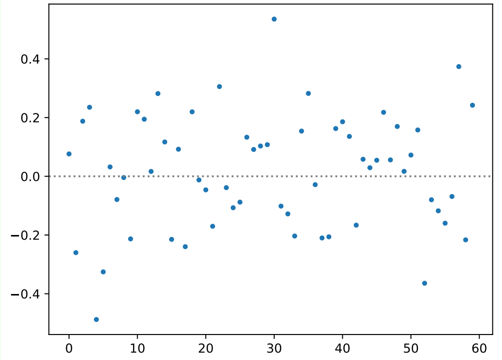


Source: `https://bit.ly/3vC9zK7`

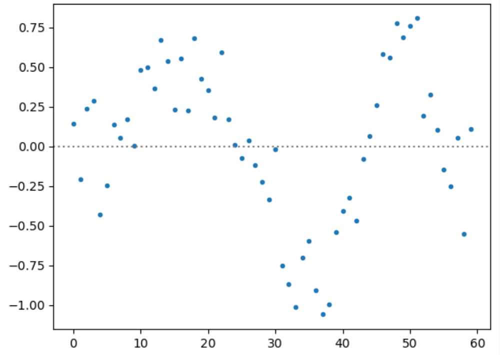
- Bias is caused by underfitting.
- Fix bias by adding suitable predictors.
- Overfitting causes large variance.
- If variance changes over the range, some errors get undue attention.
- Fix this by weighting the errors so they are balanced.

Errors should not be serially correlated

No serial correlation



Positive serial correlation

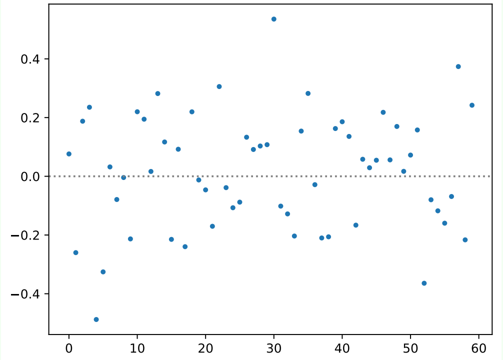


Apparent seasonal effects - can they be removed?

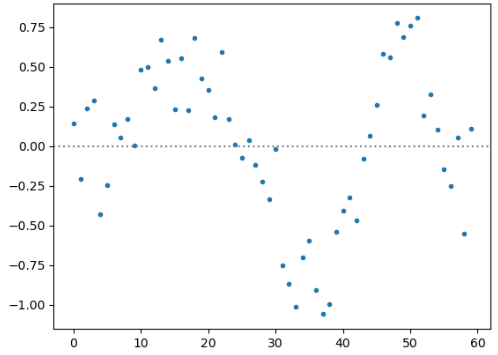
1. Add feature to the model

Errors should not be serially correlated

No serial correlation



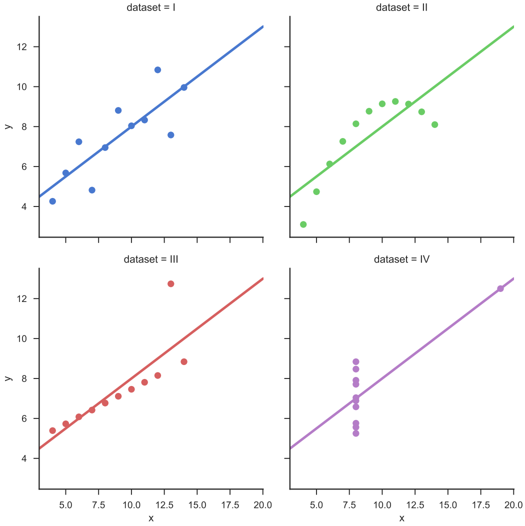
Positive serial correlation



Apparent seasonal effects - can they be removed?

1. Add feature to the model
2. Include autoregressive terms (but then it is no longer Ordinary Least Squares (OLS)!)

Anscombe's quartet (1973)



Francis Anscombe devised 4 data sets to show different forms of misalignment between data and models. Sets I,II,III share the same x values. All 4 sets share approximately the same descriptive statistics (mean and variance), but little else is common to all 4!
Only I appears suited as it stands. The other data sets require some work, particularly IV.
What do you think needs to be done for each data set?

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.

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 5 (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.

Choices of Vector norms

Definition 5 (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 6 (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.

Choices of Vector norms

Definition 5 (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 6 (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 7 (Minkowski $p-$norm)

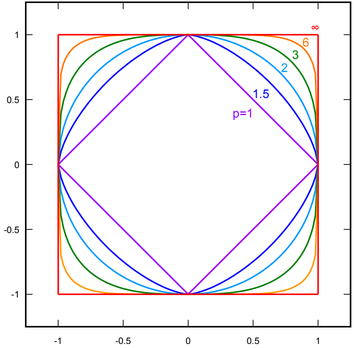
For a real number $1 \leq p < \infty$, the p-norm of $\mathbf{x}$ is defined by

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

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?

Huber loss