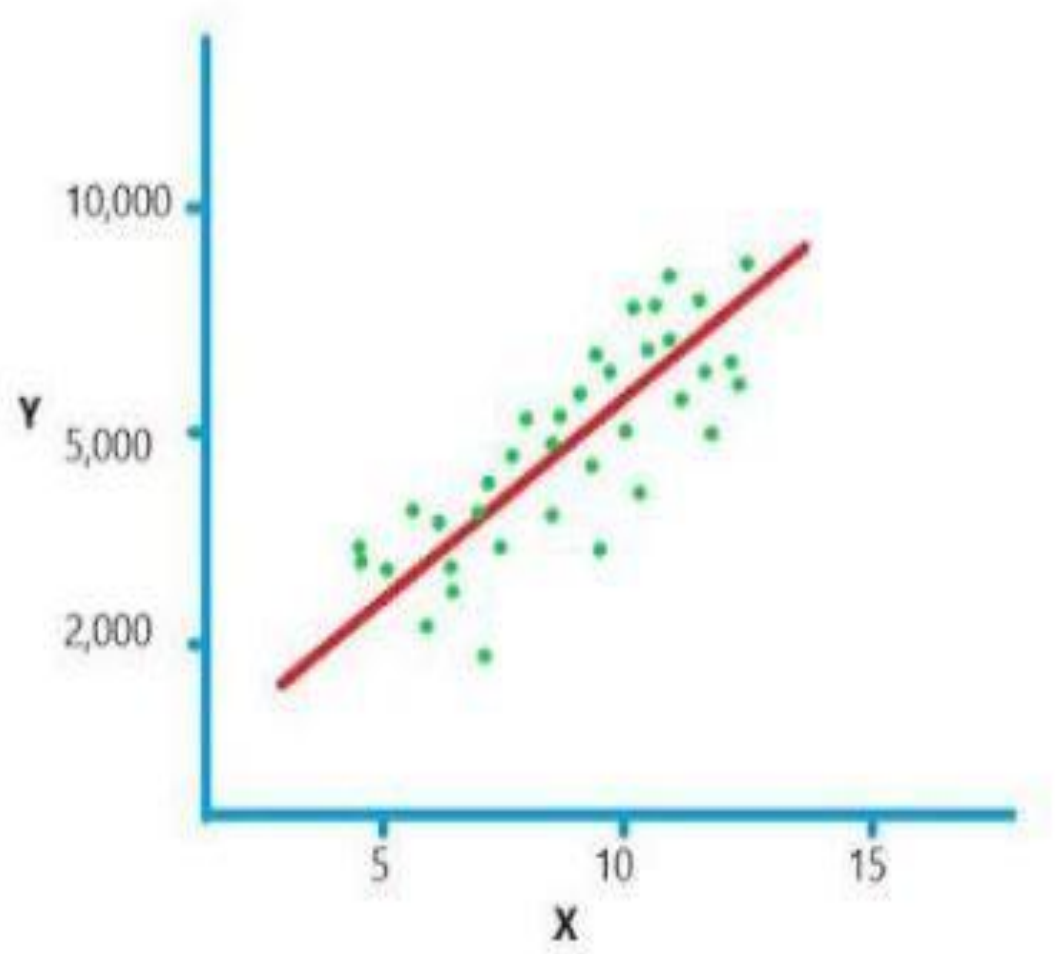
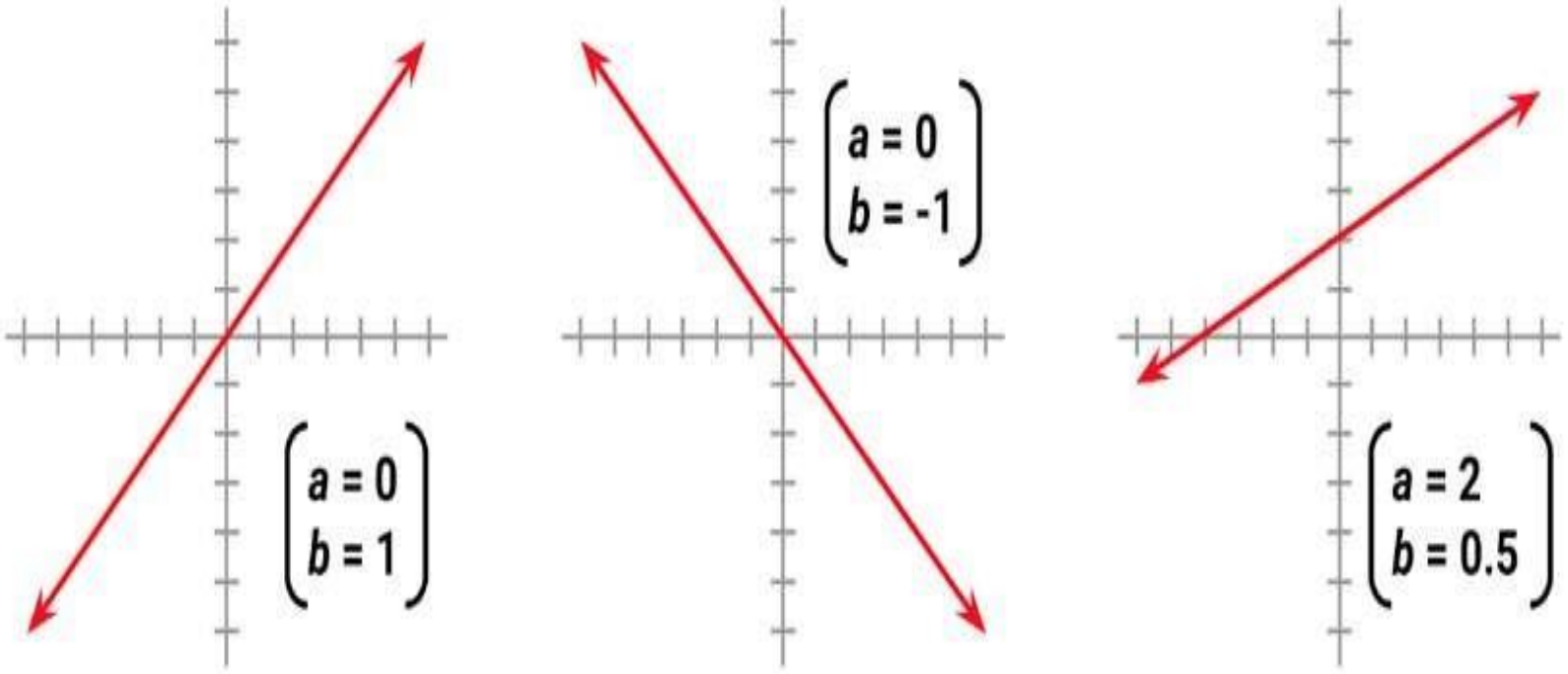
الفصل الثالث الانحدار الخطي
Linear Regression

يعتبر من أبسط خوارزميّات تعلّم الآلة (Machine Learning)ضمن فئة التعلّم تحت الإشراف (Supervised Learning) الذي يقوم بنمذجة مفهوم الانحدار .
تستخدم في تفسير متغيّر Yعبر متغيّر آخر X أو مجموعة من المتغييّرات (x1,x2,...xp) وفق دالّة خطيّة .
يسمى المتغيّر Y بالتّابع .
وتسمى المتغيّرات X بالمتغيّرات المستقلّة أو المفسّرة بمعنى أنّها تفسّر تغيّرات المتغيّر التّابع.Y



- قد يكون خط الانحدارذو ميل Slope موجب او سالب حسب العلاقة بين المتغير المستقل و المتغير او المتغيرات التابعة و اذا لم يمر من نقطة تقاطع المحاور يكون له قاطع Intercept



كيف يعمل الانحدار

- في الانحدار ، نريد كتابة الاستهلاك y، والذي يسمى المتغير التابع ، كدالة للدخل x، والذي يسمى المتغير المستقل. افترض أن الناتج هو مجموع دالة الإدخال $f\ (x)$ ومقدار الخطأالعشوائي كما هو موضح :

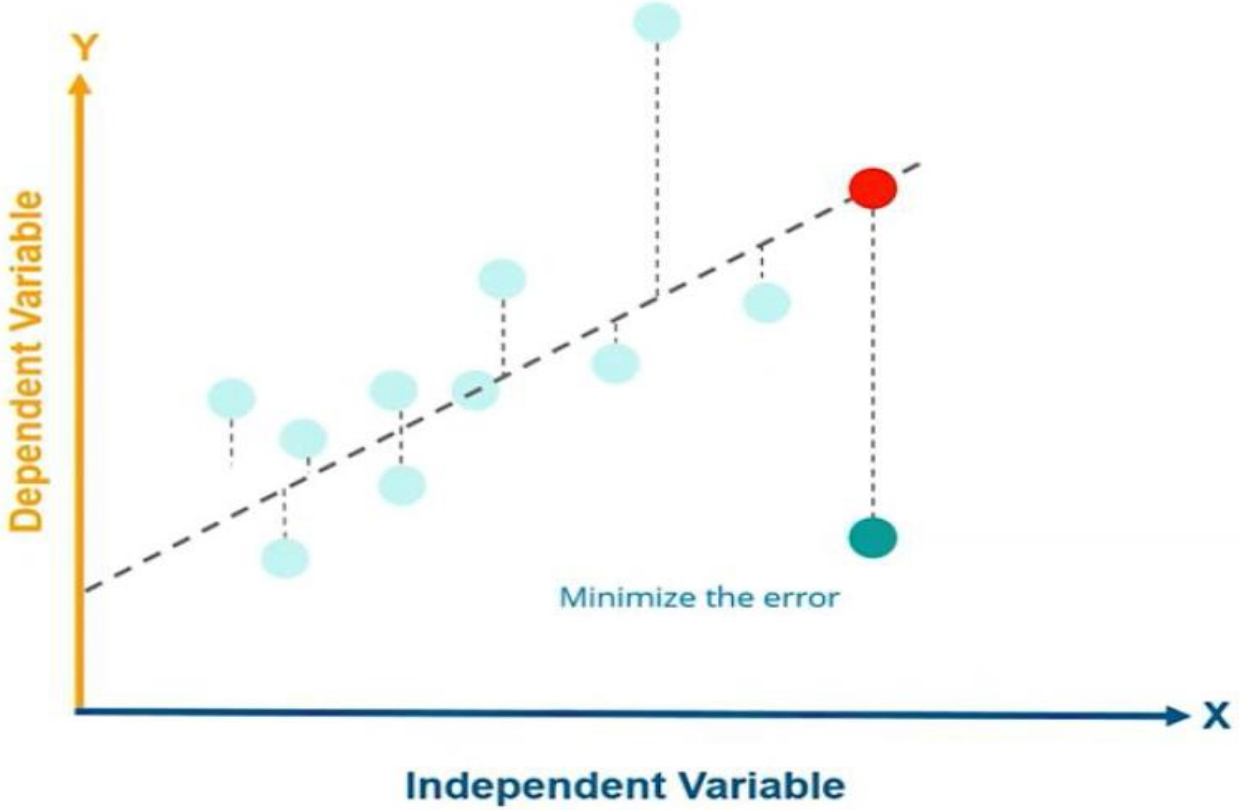
$$y = f(x) + \epsilon$$

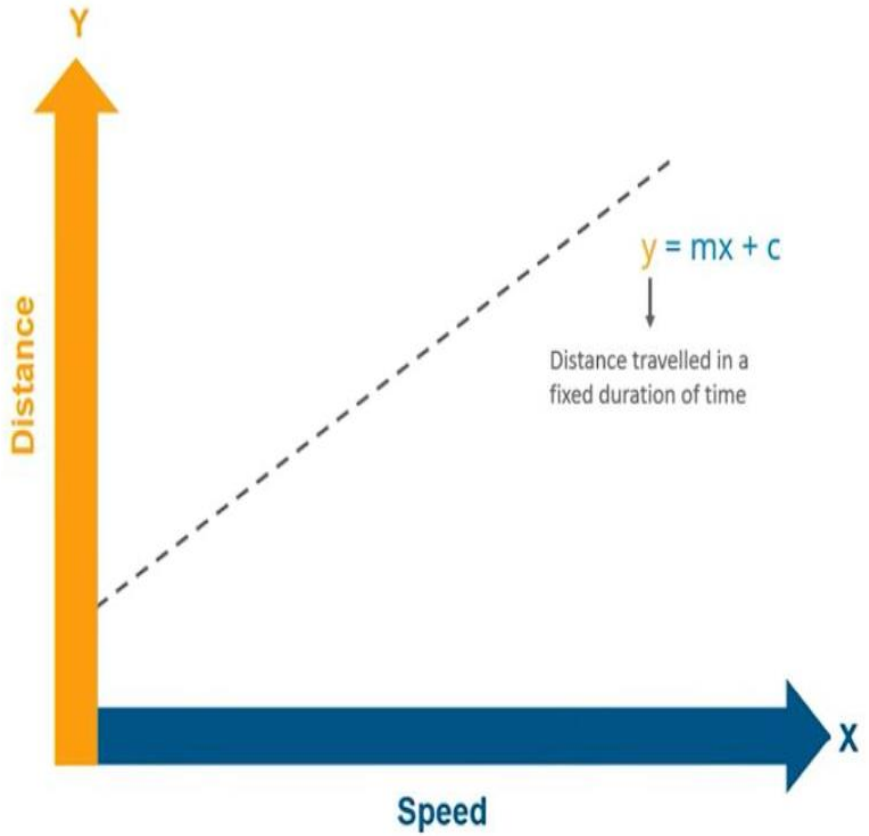
- هنا الدالة $f\ (x)$ غير معروفة ونريد تقريبها بالمقدر $g(x;\ \beta)$ الذي يحتوي على مجموعة من المعاملات β. افترض أن الخطأ العشوائي يتبع التوزيع الطبيعي بمتوسط 0. إذا كانت $x_1, \ldots, x_n$ هي عينة عشوائية لمشاهدات متغير المدخلات x و $y_1, \ldots, y_n$ القيم المرصودة مرتبطة بمتغير المخرجات y .
- بافتراض أن الخطأ يتبع التوزيع الطبيعي ، يمكننا استخدام طريقة تقدير الاحتمالية القصوى لتقدير قيم المعامل β. يمكن إظهار أن القيم β التي تعظم دالة الاحتمال هي القيم التي تقلل مجموع المربعات التالية:

$$E(\beta) = (y_1 - g(x_1, \beta))^2 + \cdots + (y_n - g(x_n, \beta))^2$$

-
- تُعرف طريقة إيجاد قيمة β كقيمة التي تقلل $E(\beta)$ باسم طريقة مربعات صغرى عاديةOLS.

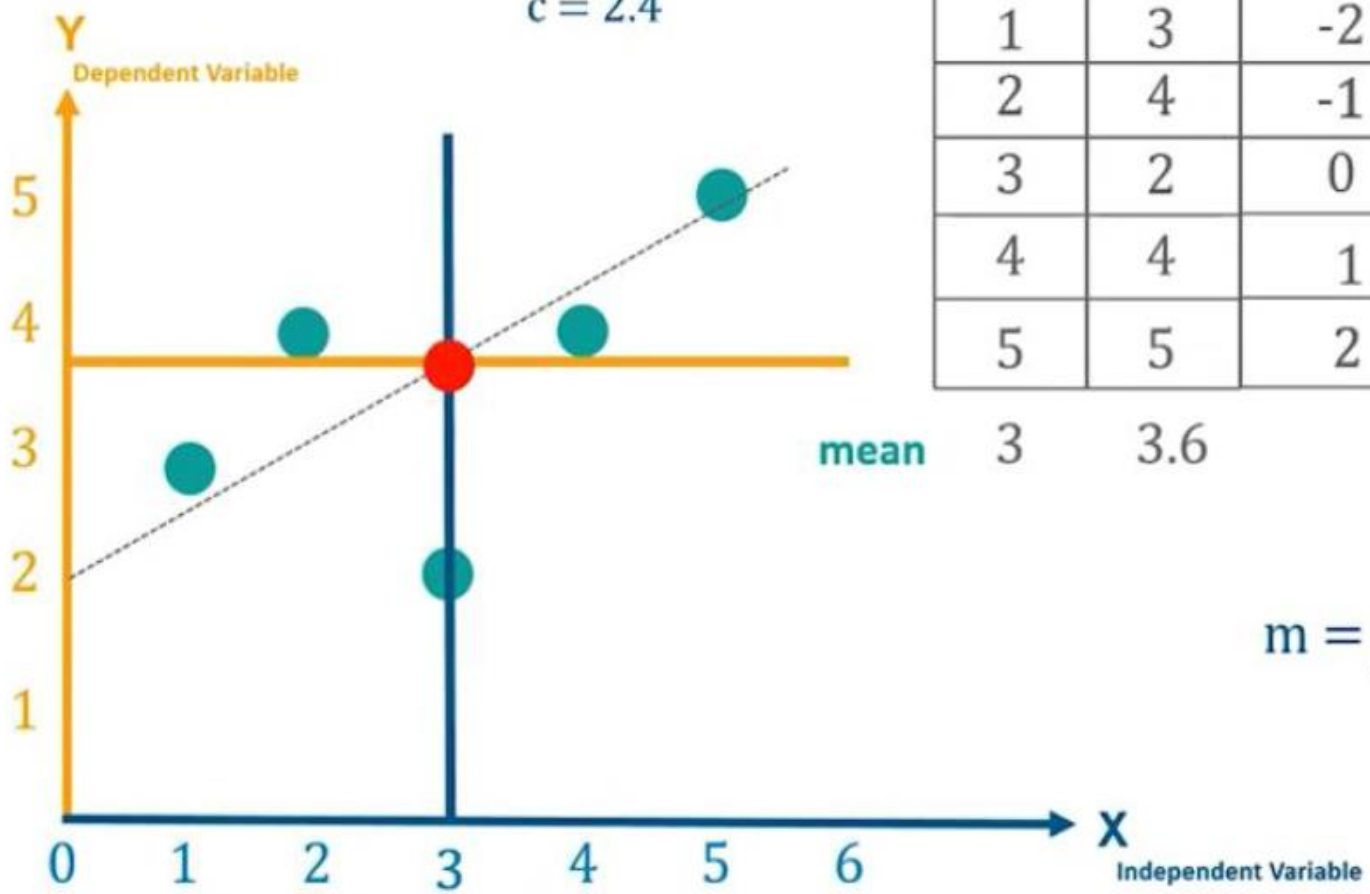
- **طريقة مربعات صغرى عادية Ordinary Least Squares**
- في طريقة مربعات صغرى عادية، يتم تحديد قيم نقاط التقاطع مع محور y والمنحدر لتقليل مجموع مربعات الأخطاء ؛ أي مجموع مربعات المسافة العمودية بين قيمة y المتوقعة وقيمة y الفعلية)





$y = mx+c$

$c = 2.4$



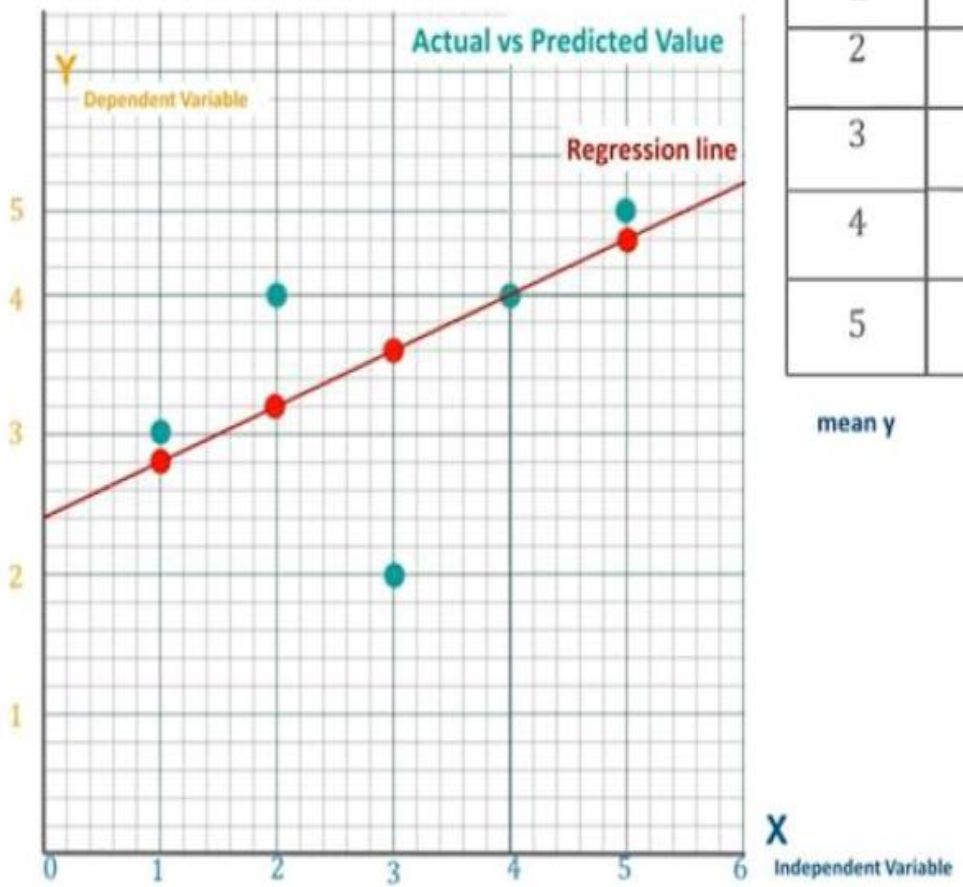
x	y	$x - \dot{x}$	$y - \dot{y}$	$(x - \dot{x})^2$	$(x - \dot{x})(y - \dot{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8
mean	3	3.6		$\Sigma = 10$	$\Sigma = 4$

$$m = \sum \frac{(x - \dot{x})(y - \dot{y})}{(x - \dot{x})^2} = \frac{4}{10}$$

$m = 0.4$

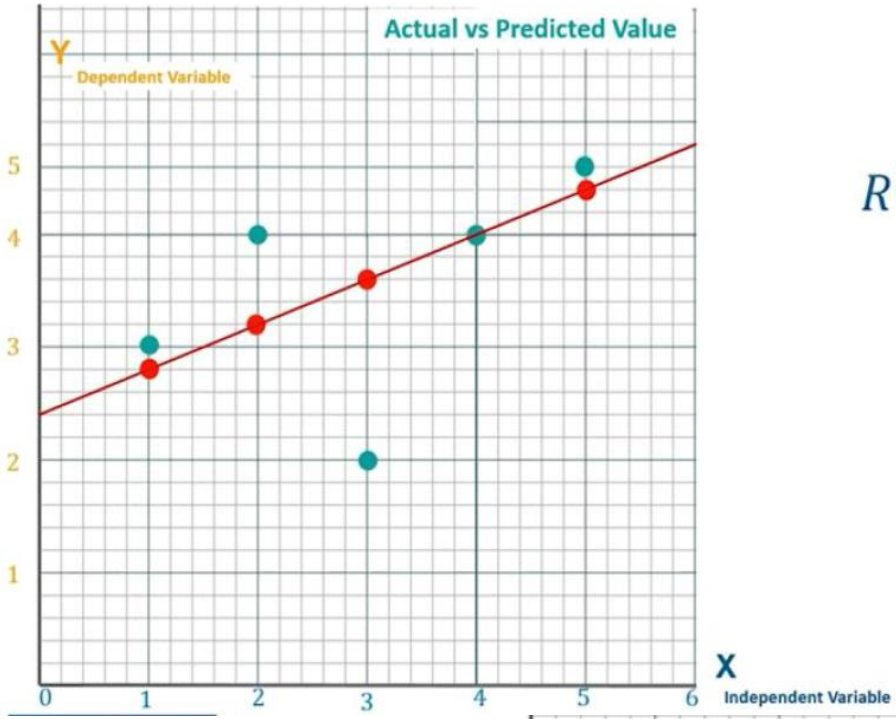
$c = 2.4$

$y = 0.4x + 2.4$

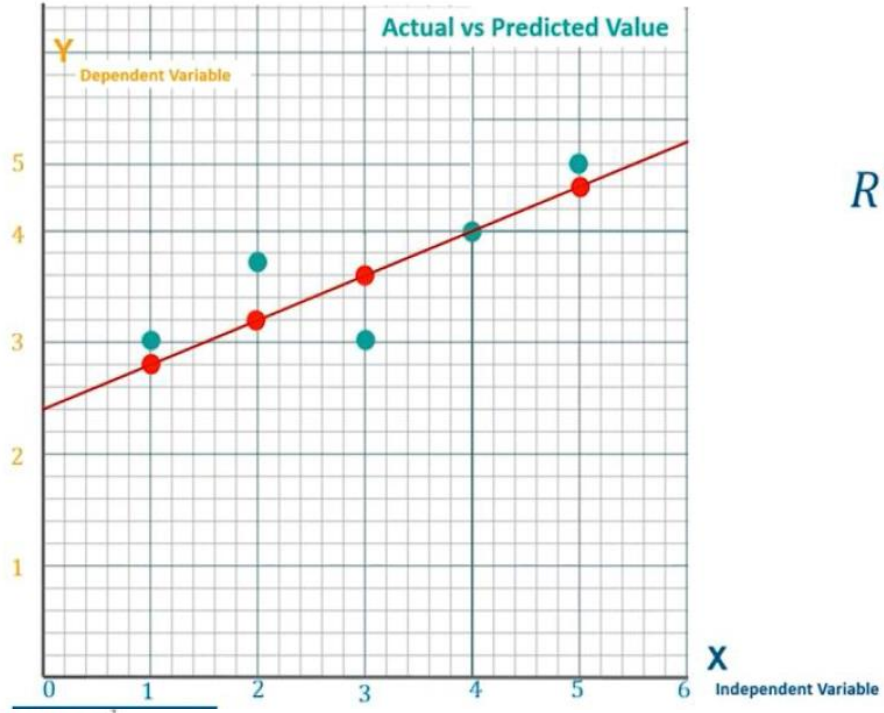


x	y	$y - \dot{y}$	$(y - \dot{y})^2$	y_p	$(y_p - \dot{y})$	$(y_p - \dot{y})^2$
1	3	– 0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	–1.6	2.56	3.6	0	0
4	4	0.4	0.16	4.0	0.4	0.16
5	5	1.4	1.96	4.4	0.8	0.64
mean y	3.6		$\sum$ 5.2			$\sum$ 1.6

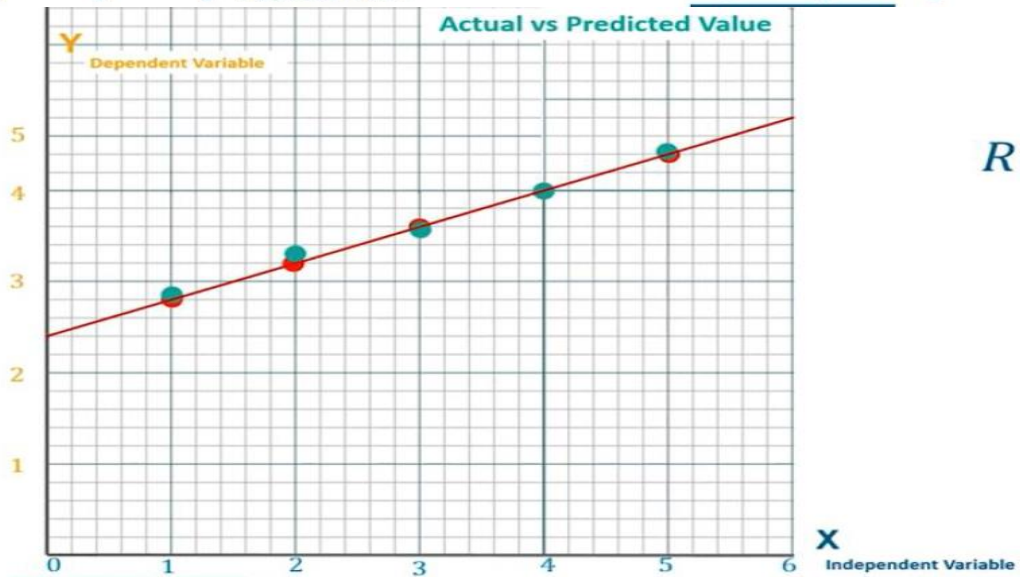
$$R^2 = \frac{1.6}{5.2} = \frac{\sum \left((y_p - \bar{y}\right)^2}{\sum (y - \dot{y})^2}$$



$R^2 \approx 0.3$



$R^2 \approx 0.7$

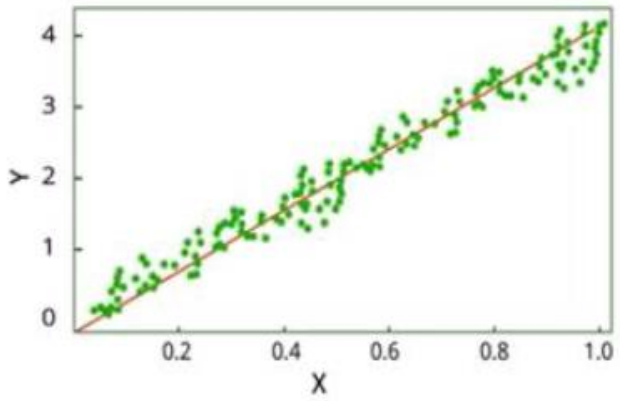


$R^2 \approx 0.9$

انواع نماذج الانحدار

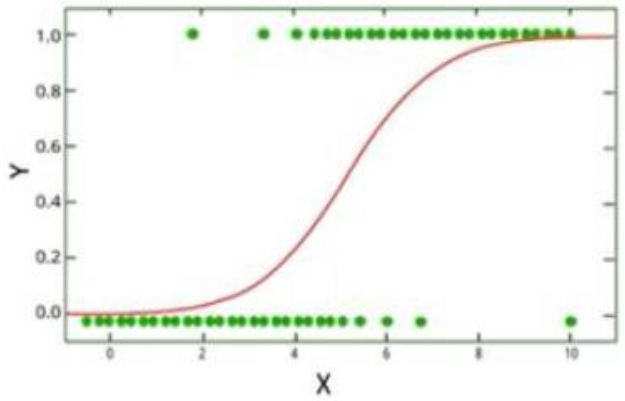
Linear Regression

- When there is a linear relationship between independent and dependent variables.



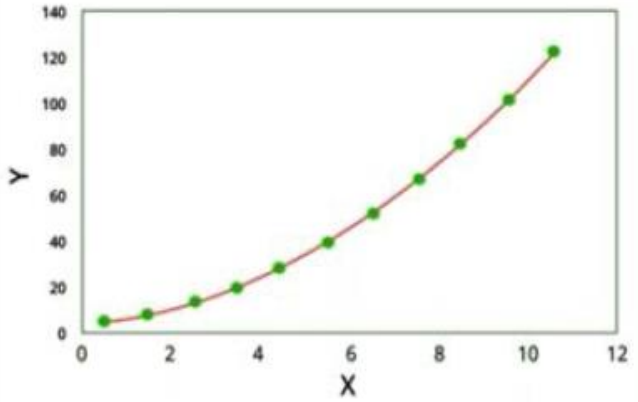
Logistic Regression

- When the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature.



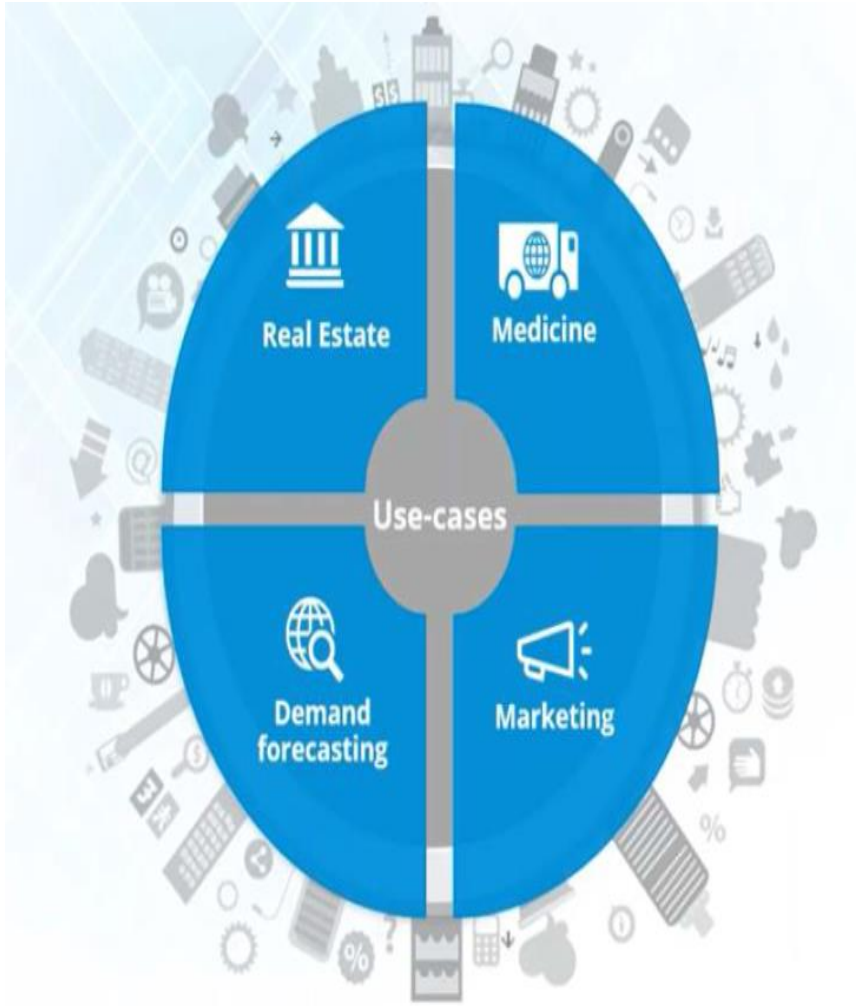
Polynomial Regression

- When the power of independent variable is more than 1.



استخدامات الانحدار الخطي

1-تقدير حالات واقعية: مثلا نريد تقدير سعر العقار فهو متعلق بمساحة العقار و الحي و طبيعة السكان و دخلهم و المدينة....



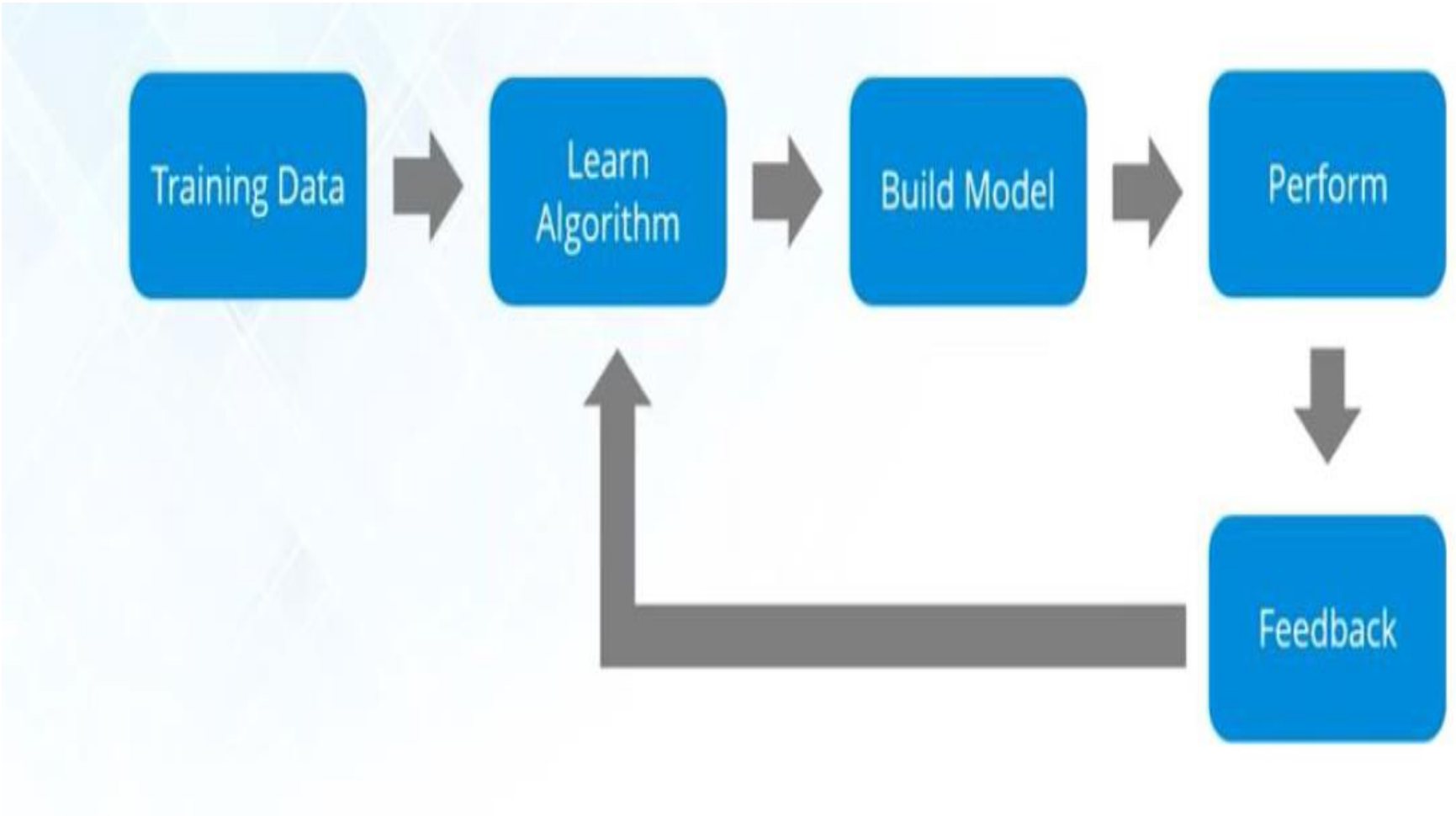
2-الطب: يستخدم الانحدارلتقدير فعالية استخدام الاشعة على مرضى السرطان بالاضافة الى العمر و الجنس...

3- التسويق و الاشهار: حيث يتم تقدير اثر الاشهار على ارتفاع المبيعات

4- التنبؤ: يستخدم الانحدار للتنبؤ بتطور الطلب على سلعة ما مثلا

تعلم الالة وفق حوازمية الانحدار الخطي البسيط

يمكن لجهاز الاعلام الالي ان يتعلم دون الحاجة الى برمجة كاملة



model <- lm(weight ~ height, data = mydata)

model <- lm(طول ~ وزن, data = قاعدة البيانات)

- model stores the created linear regression model.
- weight is the dependent variable (what we're trying to predict).
- height is the independent variable (what we're using to predict weight).
- data frame containing both weight and height variables.

-
-
- modelيخزن نموذج الانحدار الخطي الذي تم إنشاؤه.
- وزن هو المتغير التابع (ما نحاول التنبؤ به)Y.
- طول هو المتغير المستقل (ما نستخدمه للتنبؤ بالوزن)X.
- إطار البيانات والذي يحتوي على كل من المتغيرات الوزن والطول mydata .

#simple regression model انحدار خطي بسيط

```
lrmodel<-lm(medv~lstat,data =Boston)
summary(lrmodel)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-15.168   -3.990   -1.318    2.034   24.500

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,     Adjusted R-squared:  0.5432
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

>summary(model)

.فهم نتائج النموذج:

بعد تشغيل دالة lm ، يمكنك استخدام دالة summary لعرض نتائج النموذج:

سيوفر الملخص تفاصيل مختلفة مثل:

المعاملات: قاطع وميل خط الانحدار.

الانحراف المعياري : يقيس تباين المعاملات.

القيم الاحتمالية :(p-values) تشير إلى معنوية كل معامل.

:R-squared نسبة التباين التي يفسرها النموذج.

.تفسير النتائج:

يمثل القاطع القيمة المتوقعة للمتغير التابع عندما يكون المتغير المستقل يساوي صفر

يشير الانحدار إلى التغيير في المتغير التابع مقابل زيادة بمقدار واحد في المتغير المستقل.

تشير القيم الاحتمالية (p-values) الأقل من مستوى الأهمية (على سبيل المثال ، 0.05) إلى أن المعامل المقابل ذو دلالة إحصائية.

تشير قيمة R-squared القريبة من 1 إلى ملاءمة جيدة، بينما تشير القيم الأقرب إلى 0 إلى ملاءمة ضعيفة.

```
new_data <- data.frame(height = 70)     # Example data point
predicted_weight <- predict(model, new_data)
cat("Predicted weight for height 70:", predicted_weight)
```

.إجراء التنبؤات:

استخدم دالة predict للتنبؤ بقيم المتغير التابع لنقاط بيانات جديدة:

ينشئ مقطع الكود هذا إطار بيانات جديدًا بطول 70 ويستخدم النموذج للتنبؤ بالوزن المقابل.

ملاحظة:

يفترض الانحدار الخطي وجود علاقة خطية بين المتغيرات.

قم برسم بياناتك باستخدام مخطط تشتت للتحقق من الخطية.

تأكد من أن بياناتك مناسبة للانحدار الخطي.

يمكن للقيم المتطرفة والتوزيع غير الطبيعي للأخطاء أن يؤثر على دقة النموذج.

```
> predict2<-predict(model,testing_data)
>predict2
        4         10         11         15         18         23         24
29.405335 19.632603 20.174332 19.120567 16.947674 16.143685 13.823632
       27         29         30         42         44         50         51
```

Linear Regression in Machine Learning
#We will split the data into training and testing sets

```
> set.seed(2)
> split<-sample.split(Boston$medv,SplitRatio = 0.7)
➢Split
```

```
  [1]  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
TRUE [13]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
FALSE FALSE……
```

- .1**تحديد البذرة:**

السطر الأول **set.seed(2)** يضبط قيمة بذرة لمولد الأرقام العشوائية. يضمن هذا أنه إذا قمت بتشغيل الكود عدة مرات، فستحصل على نفس تقسيم البيانات إلى مجموعتي تدريب واختبار. وهذا يساعد على قابلية تكرار تحليلك.

.2تقسيم البيانات:

السطر الثاني **split <- sample.split(Boston$medv, SplitRatio = 0.7)**

يقسم البيانات في المتغير Boston$medv بافتراض أن إطار البيانات يحتوي على متغير تابع يسمى 'medv' إلى مجموعتي تدريب واختبار.

sample.split هي دالة من حزمة قاعدة R لتقسيم البيانات.

Boston عبارة عن مجموعة بيانات مضمنة في R تحتوي على معلومات حول أسعار المساكن.

medvهو المتغير المحدد داخل مجموعة بيانات Boston الذي نريد تقسيمه. و هو يمثل القيمة المتوسطة للوحدات السكنية المملوكة من قبل أصحابها بالآلاف من الدولارات.

SplitRatio = 0.7 يحدد نسبة البيانات التي سيتم تضمينها في مجموعة التدريب. في هذه الحالة، سيتم وضع 70٪ من البيانات في مجموعة التدريب وسيتم وضع الـ 30٪ المتبقية في مجموعة الاختبار.

عن طريق تقسيم البيانات إلى مجموعتي تدريب واختبار، يمكنك تدريب نموذج على بيانات التدريب ومن ثم تقييم أدائه على بيانات الاختبار. وهذا يساعد على منع الإفراط في التقدير Overfiting، والذي يمكن أن يحدث عندما يتم تقريب النموذج بشكل وثيق للغاية مع بيانات التدريب ولا يعمل بشكل جيد على البيانات الاخرى .

```
# dividing the data with the ratio 0,7
training_data<-subset(Boston,split=="TRUE")
testing_data<-subset(Boston,split=="FALSE")
```

- يعتمد مقطع الكود على حزمة `dplyr` في R لتقسيم مجموعة بيانات إلى مجموعتي تدريب واختبار. إليك شرح لما يفعله الكود:
- السطر `training_data <- filter(data, row_number() <= 0.7 * nrow(data))`
- يصفِّي إطار البيانات المسمى `data` لإنشاء مجموعة التدريب.
- `filter` هي دالة من حزمة `dplyr` تُستخدم لتصفية إطارات البيانات.
- `row_number()` تخصص رقم صف فريد لكل صف في إطار البيانات.
- يعبر الشرط `row_number() <= 0.7 * nrow(data)` عن تصفية إطار البيانات لإبقاء الصفوف التي يكون فيها رقم الصف أقل من أو يساوي 70٪ من إجمالي عدد الصفوف في إطار البيانات. (`nrow(data)`) وهذا يختار 70٪ من البيانات لمجموعة التدريب.

- إنشاء مجموعة الاختبار.
- السطر `testing_data <- filter(data, row_number() >0.7 *`
- `nrow(data))`
- ينشئ مجموعة الاختبار من البيانات المتبقية.
- على غرار السطر السابق، يقوم بتصفية إطار البيانات، ولكن هذه المرة يحتفظ فقط بالصفوف التي يكون فيها رقم الصف أكبر من 70٪ من إجمالي عدد الصفوف. وهذا يختار نسبة الـ 30٪ المتبقية من البيانات لمجموعة الاختبار.
- عن طريق تقسيم البيانات إلى مجموعتي تدريب واختبار، يمكنك تدريب نموذج على بيانات التدريب ومن ثم تقييم أدائه على بيانات الاختبار. وهذا يساعد على منع الإفراط في التقدير.
- ملاحظة:
- تفترض هذه الطريقة أن إطار البيانات `data` لا يحتوي على عمود يسمى `row_number`. إذا كان الأمر كذلك، فقد تحتاج إلى تعديل الكود لتجنب التعارضات.

Thank you