Simple linear regression analysis

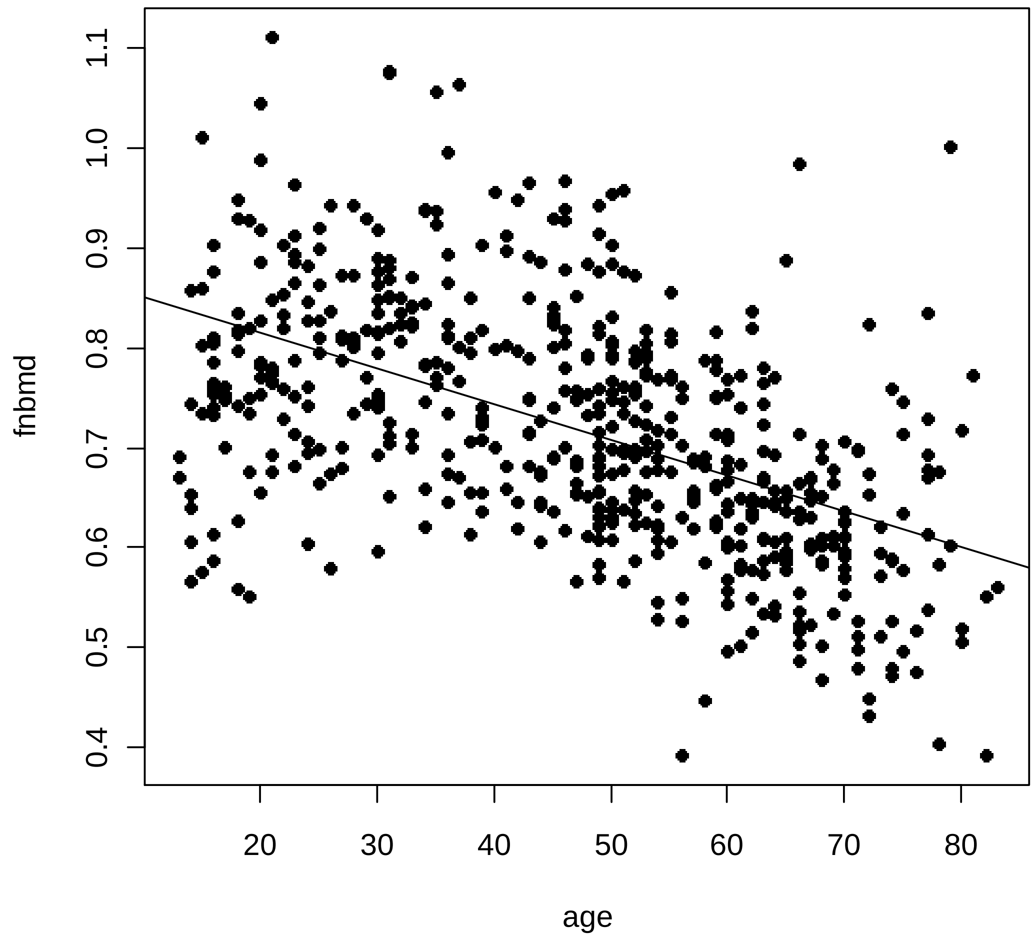
Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

What we are going to learn …

- Examples
- Purposes of linear regression analysis
- Questions of interest
- Model parameters
- R analysis
- Interpretation

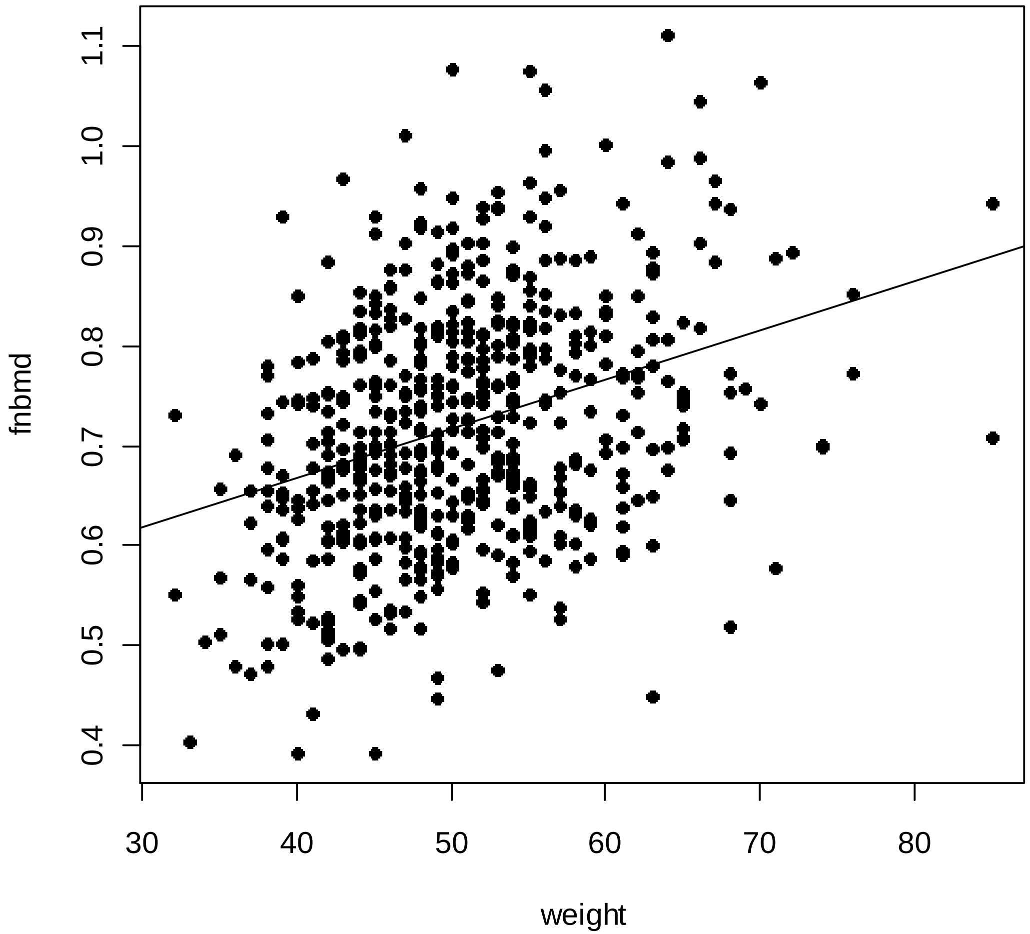
Femoral neck bone density and age

```
women = subset(vd, sex==2)
plot(fnbmd ~ age, pch=16)
abline(lm(fnbmd ~ age))
```



Weight and femoral neck bone density

```
plot(fnbmd ~ weight, pch=16)
abline(lm(fnbmd ~ weight))
```



Correlation analysis

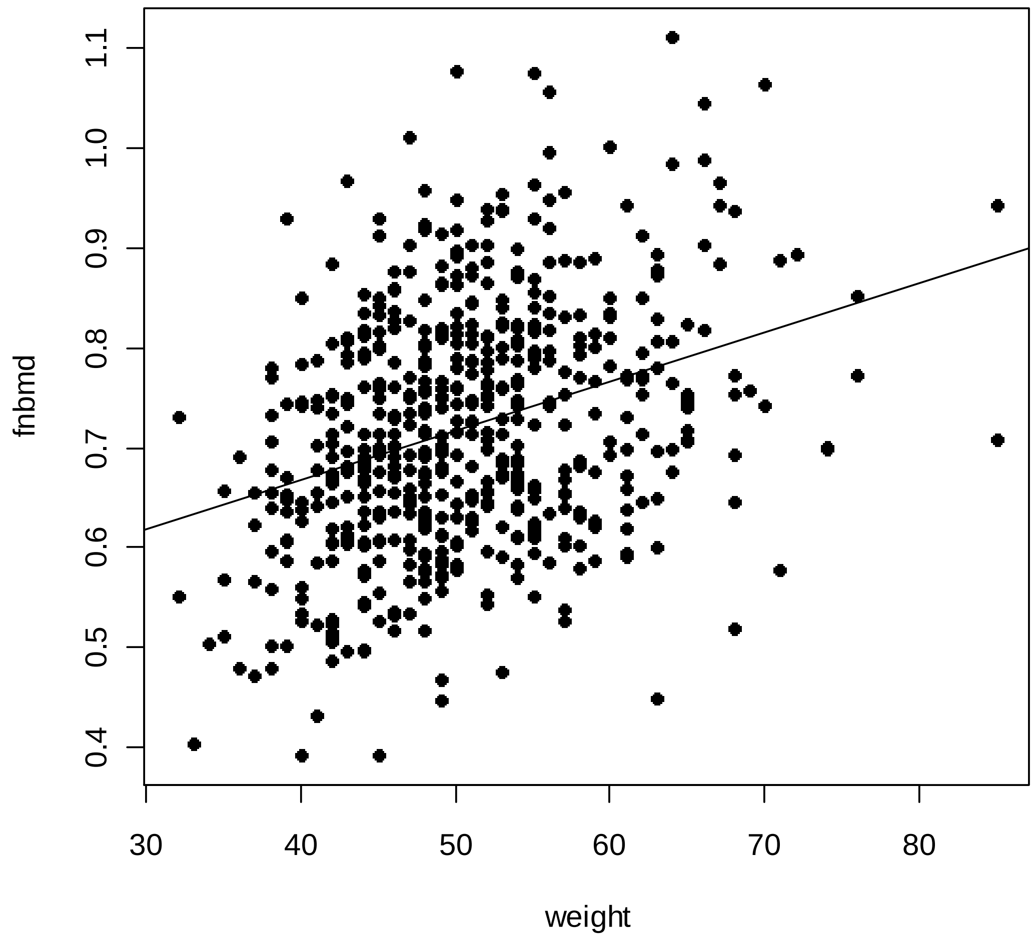
- Assessment of a relationship
- The coefficient of correlation: a measure of the relationship
- We want to know more …
  - The magnitude of effect of a *predictor* variable on the *outcome*
  - Prediction of outcome by using the predictor variable(s)

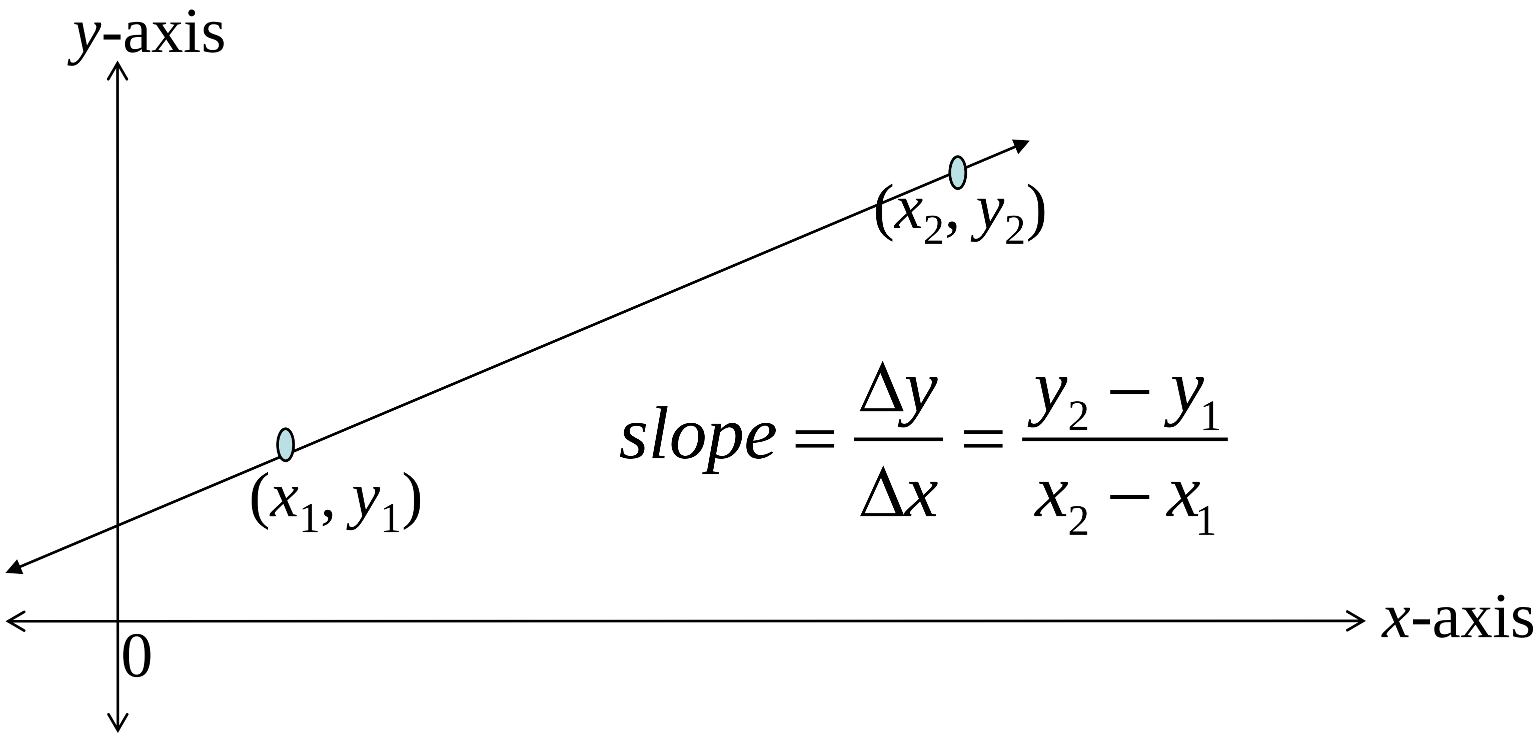
Our interests

- Finding a statistical model that decribes the relationship between age, weight, and BMD
- Adjustment of effect
- Prediction

Linear regression model

Weight and femoral neck bone density





$$slope = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

We can also describe the line in terms of a slope and an intercept

- **The slope the change in the *y*-value for a unit change in the *x*-value. In this simple situation we can think of this as the change in the height of the line as we progress along the *x*-axis**
- **The intercept is the height of the line when *x* = 0**

Linear regression: model

- *Y* : random variable representing a **response**
- *X* : random variable representing a **predictor** variable (predictor, risk factor)
  - Both *Y* and *X* can be a categorical variable (e.g., yes / no) or a continuous variable (e.g., age).
  - If *Y* is categorical, the model is a **logistic regression** model; if *Y* is continuous, a **simple linear regression** model.

- **Model**

$$Y = \alpha + \beta X + \varepsilon$$

α : intercept

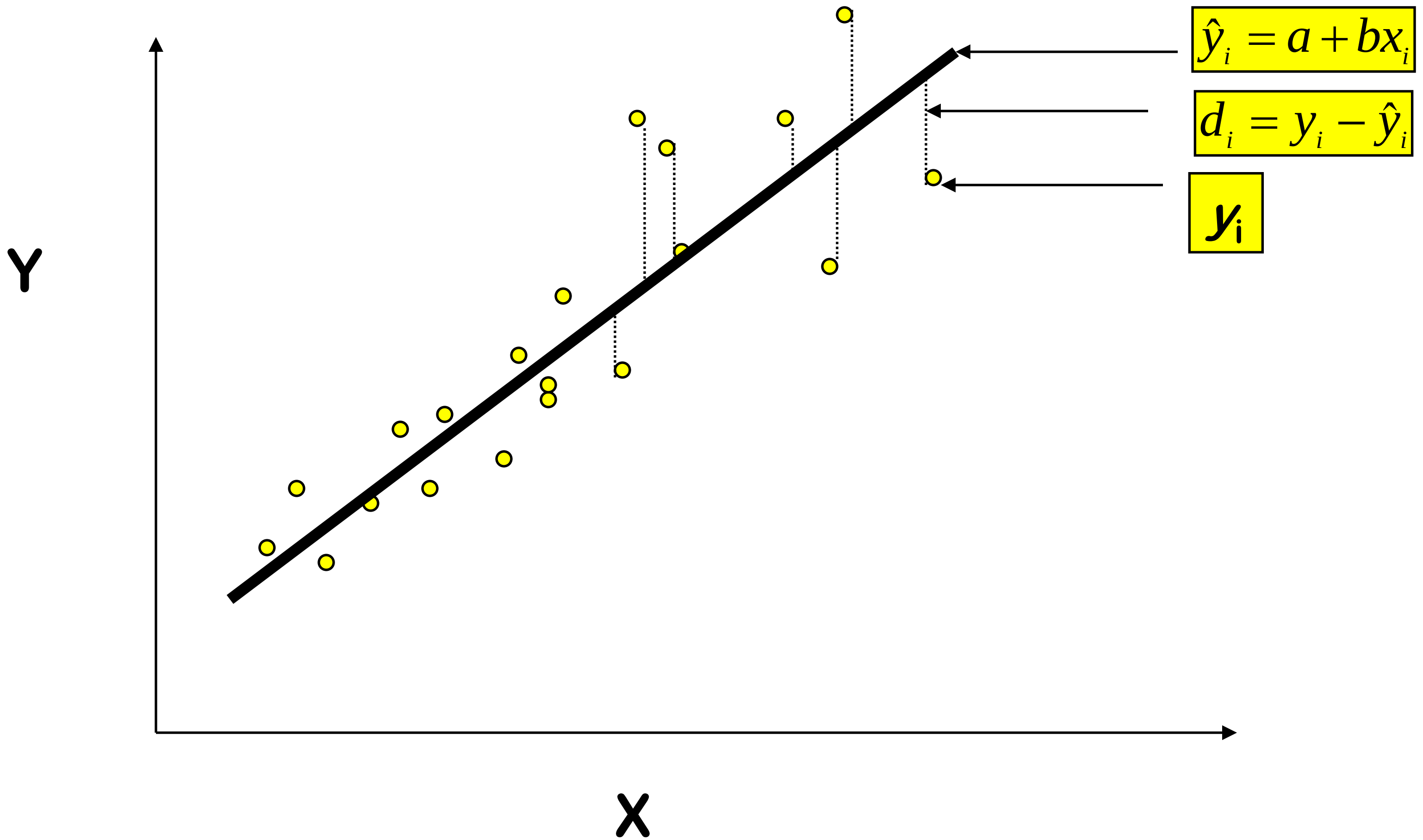
β : slope / gradient

ε : random error (variation between subjects in y even if *x* is constant, e.g., variation in cholesterol for patients of the same age.)

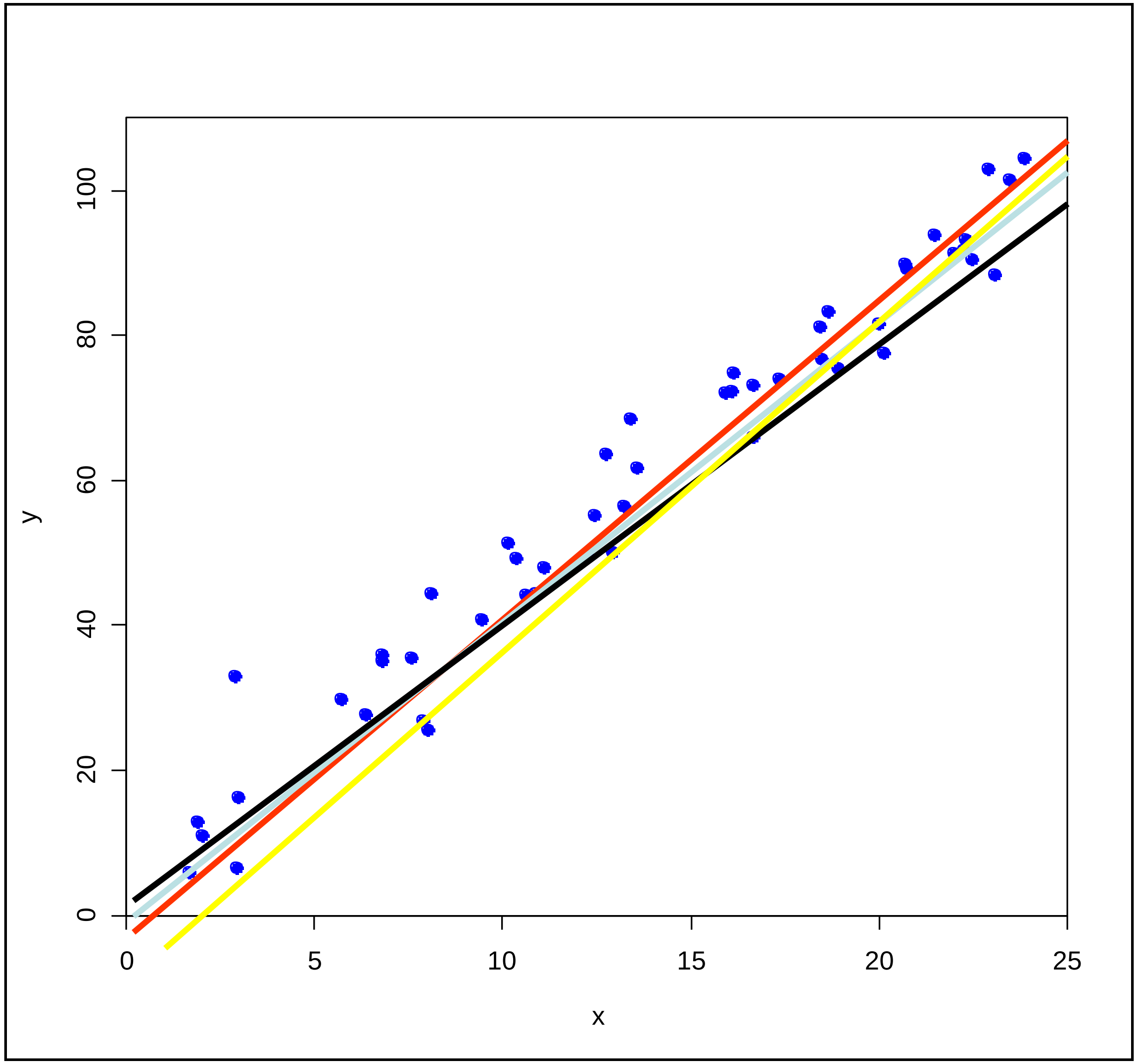
Linear regression: assumptions

- The relationship is linear *in terms of the parameter*;
- *X* is measured without error;
- The values of *Y* are independently from each other (e.g., Y_1 is not correlated with Y_2) ;
- The random error term (ε) is **normally** distributed with mean 0 and **constant** variance.

Criteria of estimation



The goal of least square estimator (LSE) is to find a and b such that the sum of d^2 is minimal.



- We could try fitting a line “by eye”
- But everyone’s best guess would probably be different
- We want consistency

Estimating parameters by R

- **Our interest: relationship between BMD and weight**
- **Model:**

$$BMD = a + b*weight + e$$

- **We want to estimate *a* and *b***
- **R language**

```
lm(bmd ~ weight)
```

R analysis

```
> m1 = lm(fnbmd ~ weight)
> summary(m1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.4699822  0.0310144   15.15  < 2e-16 ***
weight      0.0049416  0.0006041    8.18 1.95e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1152 on 556 degrees of freedom
Multiple R-squared: 0.1074,     Adjusted R-squared: 0.1058
F-statistic:  66.9 on 1 and 556 DF,  p-value: 1.945e-15
```

Interpretation of outputs

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.4699822  0.0310144   15.15  < 2e-16 ***
weight      0.0049416  0.0006041    8.18 1.95e-15 ***
```

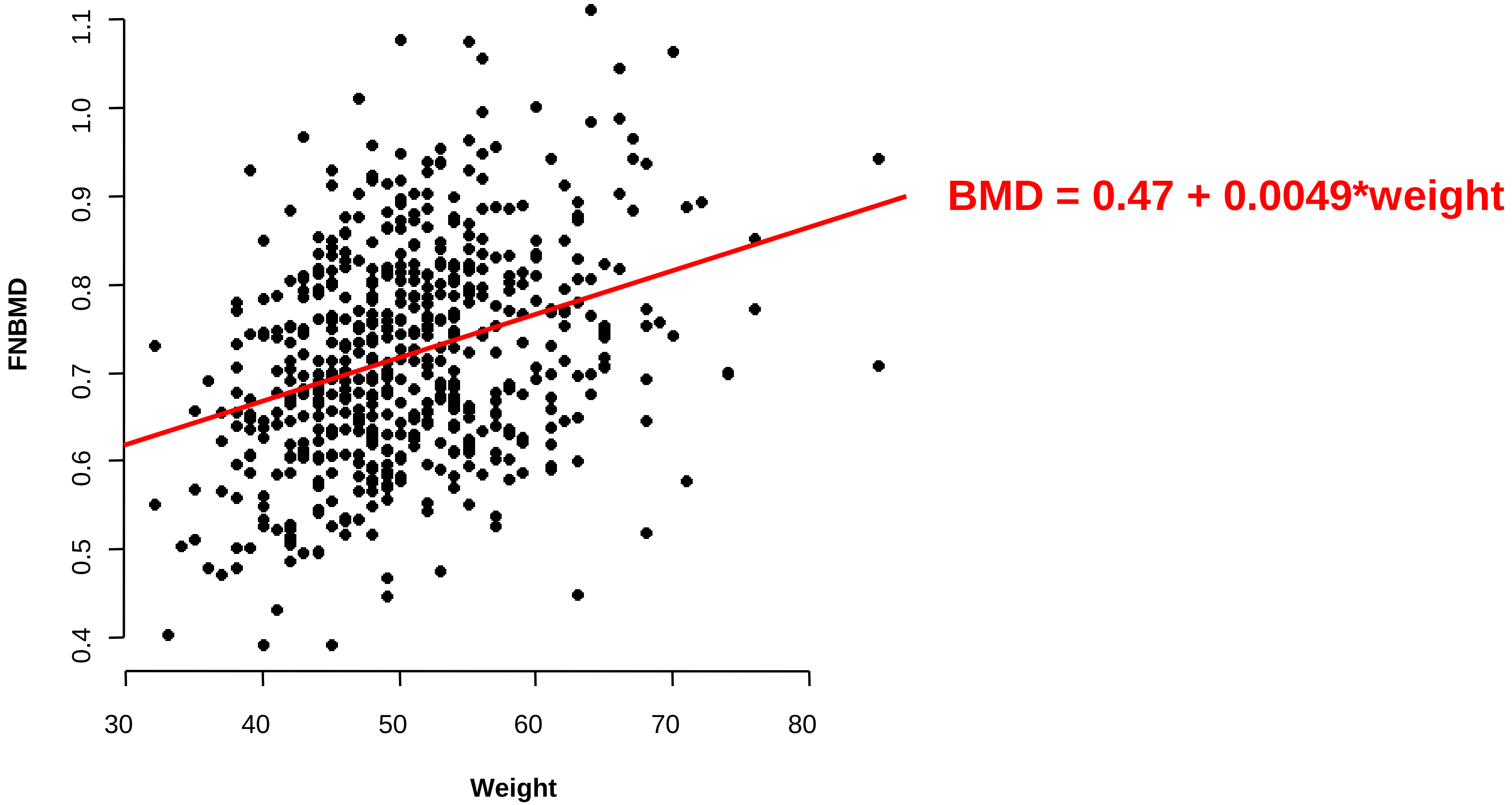
- Remember our model:

$$BMD = a + b*weight$$

- Our equation:

$$BMD = 0.47 + 0.0049*weight$$

- Interpretation: **1 kg increase in weight was associated with a 0.0049 g/cm^2 increase in BMD. The association is statistically significant (P < 0.0001)**


BMD = 0.47 + 0.0049*weight
FNBMD
Weight
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
30
40
50
60
70
80

Analysis of variance

- BMD = a + b*weight + e
- Observed variation = model + random

  "Variation" = sum of squares

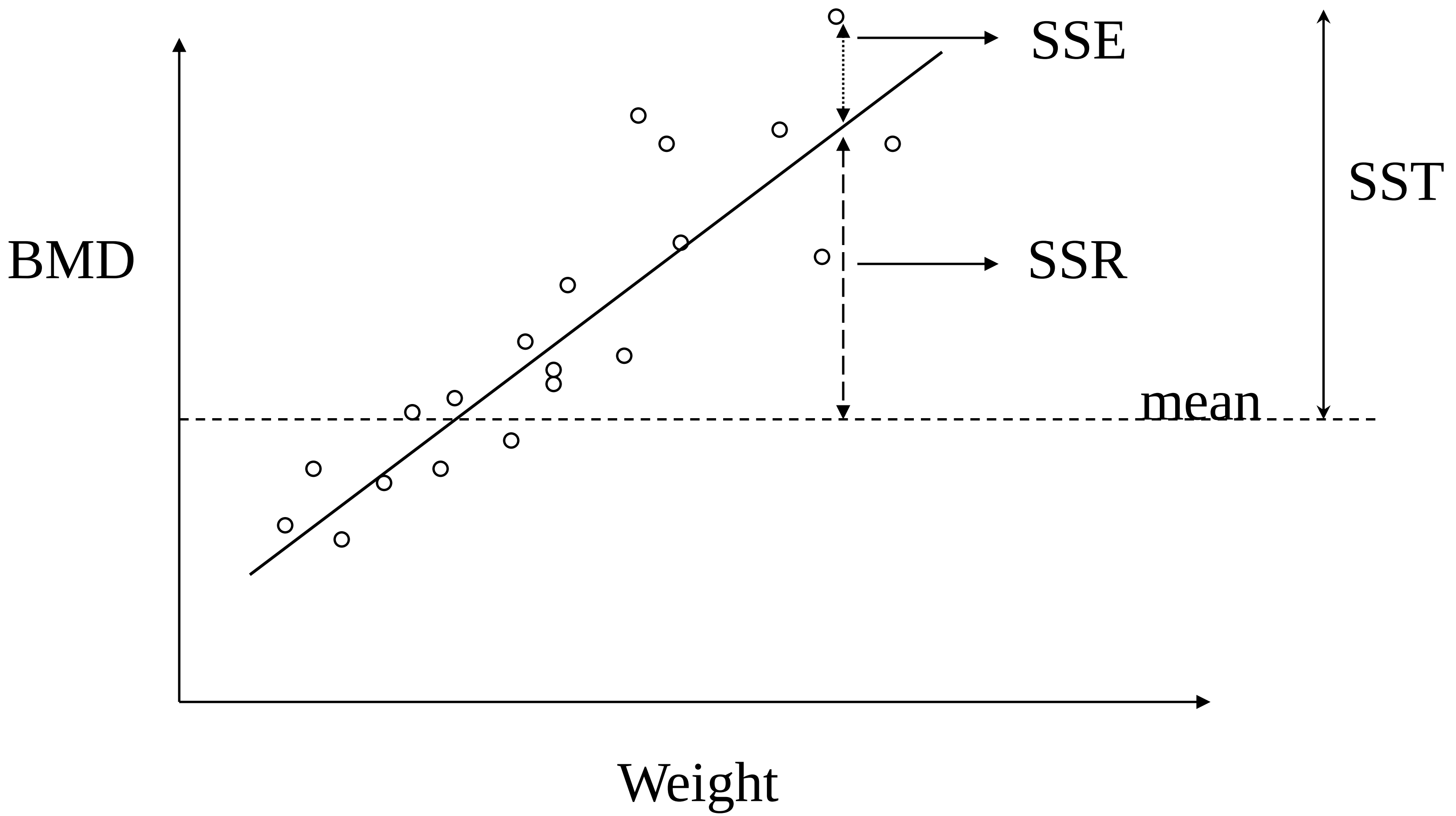
- SST = total sum of squares

  SSR = sum of squares due to the regresson model

  SSE = sum of squares due to random component

- SST = SSR + SSE
- R^2 = SSR / SST

Partitioning of variations: geometry



SST = SSR + SSE

Partitioning of variation by R

```
> m1 = lm(fnbmd ~ weight)
> anova(m1)
```

```
Analysis of Variance Table
Response: fnbmd
           Df Sum Sq Mean Sq F value    Pr(>F)
weight      1 0.8883 0.88829  66.905 1.945e-15 ***
Residuals 556 7.3819 0.01328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Total SS = 0.8883 + 7.3819 = 8.2702
- R2 = 0.8883 / 8.2702 = 0.107

Interpretation of outputs

```
Residual standard error: 0.1152 on 556 degrees of freedom
Multiple R-squared: 0.1074,     Adjusted R-squared: 0.1058
F-statistic:  66.9 on 1 and 556 DF,  p-value: 1.945e-15
```

- $R^2 = 0.107$
- Interpretation: **Approximately 11% of BMD variance could be accounted for by body weight**

Variance of BMD after adjusting for weight

```
> m1 = lm(fnbmd ~ weight)
> anove(m1)
```

```
Analysis of Variance Table
Response: fnbmd
           Df Sum Sq Mean Sq F value    Pr(>F)
weight      1 0.8883 0.88829  66.905 1.945e-15 ***
Residuals 556 7.3819 0.01328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Mean square (MS) = sum of squares / (degrees of freedom)
- MS(residuals) = 7.3819 / 556 = 0.01328

⇒ Variance of BMD after adjusting for weight is 0.01328

(variance of BMD before the adjustment: 0.01485

Prediction of BMD by weight

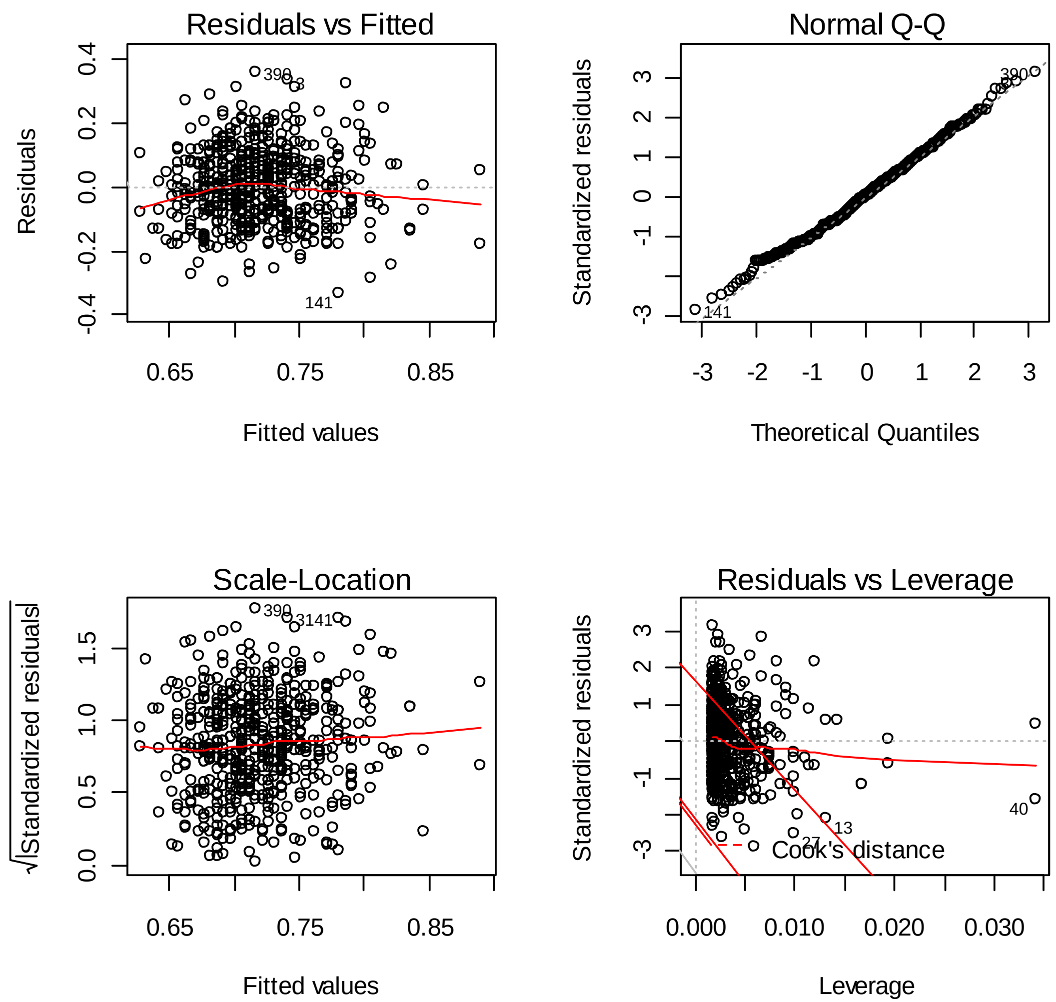
- The model: BMD = 0.47 + 0.0049*weight
- Without the knowledge of weight, the mean BMD is 0.72 g/cm^2
- With knowledge of weight, we know that BMD is dependent on weight
- Weight = 50 kg, BMD = 0.47 + 0.0049*50 = 0.72 g/cm^2

  Weight = 40 kg, BMD = 0.47 + 0.0049*40 = 0.67 g/cm^2

  Weight = 60 kg, BMD = 0.47 + 0.0049*60 = 0.76 g/cm^2

Checking model assumptions

```
par(mfrow=c(2,2))
plot(m1)
```



Be careful! Anscrombe's data

Frank Anscombe devised 4 sets of X-Y pairs

x	y1	y2	y3	x4	y4
10	8.04	9.14	7.46	8	6.58
8	6.95	8.14	6.77	8	5.76
13	7.58	8.74	12.74	8	7.71
9	8.81	8.77	7.11	8	8.84
11	8.33	9.26	7.81	8	8.47
14	9.96	8.10	8.84	8	7.04
6	7.24	6.13	6.08	8	5.25
4	4.26	3.10	5.39	19	12.50
12	10.84	9.13	8.15	8	5.56
7	4.82	7.26	6.42	8	7.91
5	5.68	4.74	5.73	8	6.89

Mean and SD of Anscombe's data

Data Set	N	X		Y	
		Mean	SD	Mean	SD
1	11	7.50	2.03	9.00	3.32
2	11	7.50	2.03	9.00	3.32
3	11	7.50	2.03	9.00	3.32
4	11	7.50	2.03	9.00	3.32

Correlation between X and Y: Anscombe's data

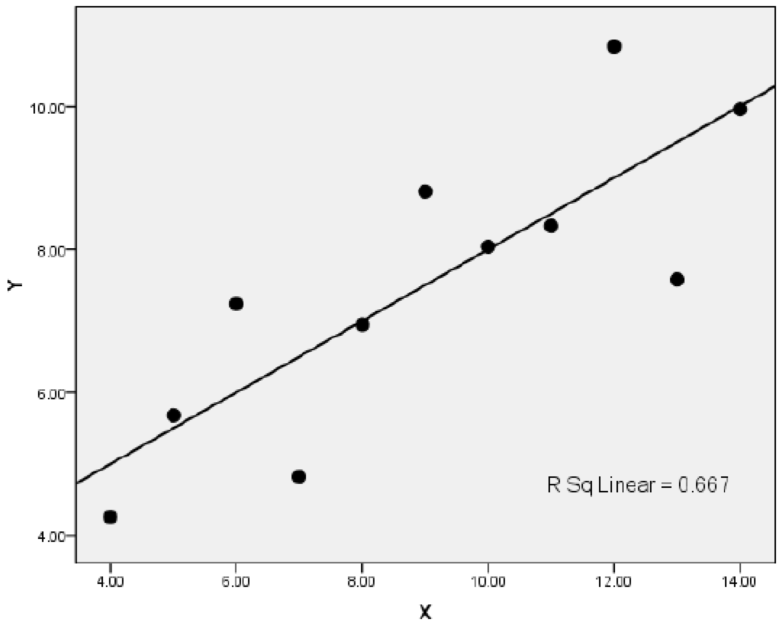
Data Set	Pearson r	R-squared	Adj. R-sq	SE
1	0.82	0.67	0.63	1.24
2	0.82	0.67	0.63	1.24
3	0.82	0.67	0.63	1.24
4	0.82	0.67	0.63	1.24

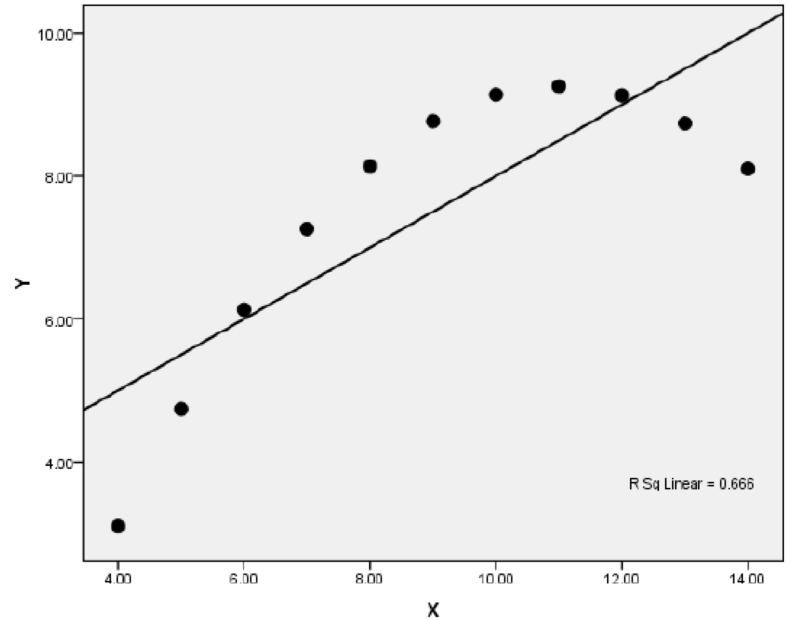
Regression analysis: Anscombe's data

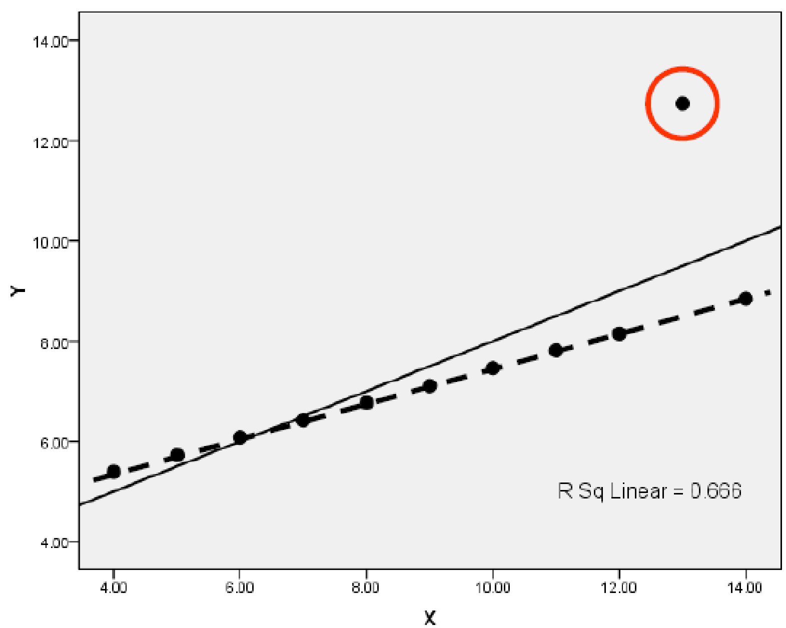
Data Set		B	SE	t	p	95% CI Lower	95% CI Upper
1	Constant	3.00	1.124	2.67	0.026	0.459	5.544
	X	0.50	0.118	4.24	0.002	0.233	0.766
2	Constant	3.00	1.124	2.67	0.026	0.459	5.546
	X	0.50	0.118	4.24	0.002	0.233	0.766
3	Constant	3.00	1.125	2.67	0.026	0.455	5.547
	X	0.50	0.118	4.24	0.002	0.233	0.767
4	Constant	3.00	1.125	2.67	0.026	0.456	5.544
	X	0.50	0.118	4.24	0.002	0.233	0.767

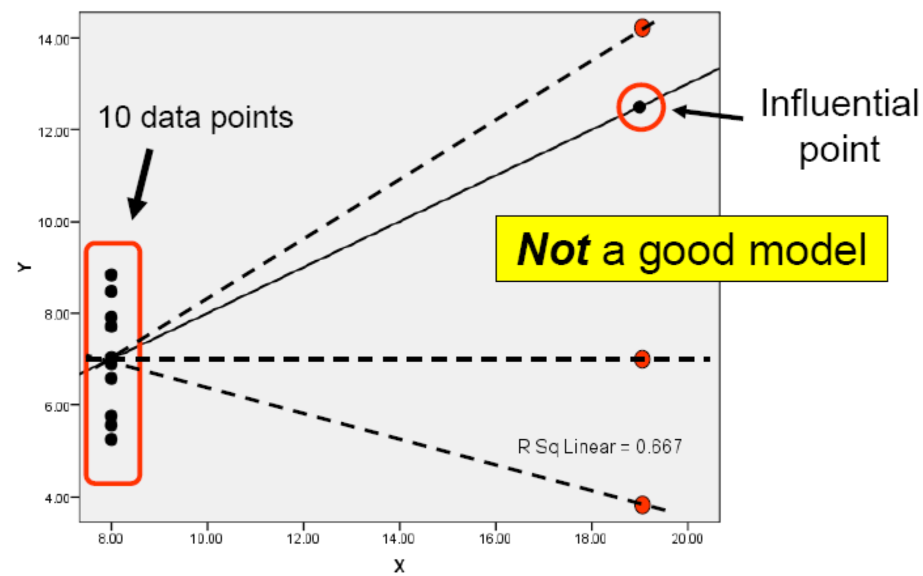
For all 4 models, $Y' = 0.5(X) + 3$

But …


R Sq Linear = 0.667
Y
X


R Sq Linear = 0.666
Y
X


R Sq Linear = 0.666
Y
X


10 data points
Influential point
Not a good model
R Sq Linear = 0.667
Y
X

Summary

- Simple linear regression model is used for
  - Understanding the effect of a risk factor or determinant on an outcome variable
  - Predicting an outcome variable
- It’s appropriate when the functional relationship is linear
- Always check assumptions!