PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

Linear Regression

"All models are wrong – but some are useful."

-- George Box

slido

Please download and install the Slido app on all computers you use

How are we feeling about R?

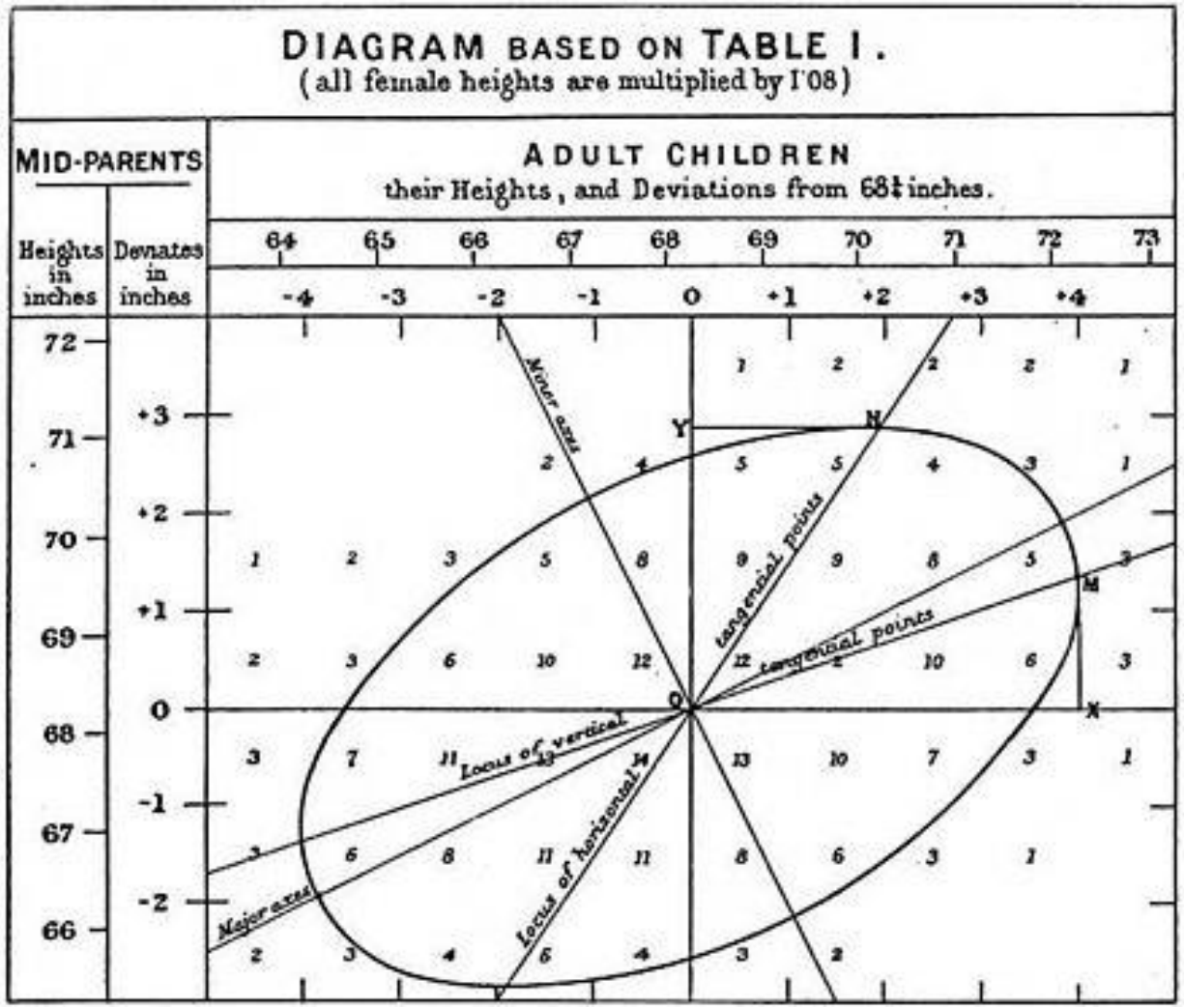
ⓘ Start presenting to display the poll results on this slide.

PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

Linear Regression

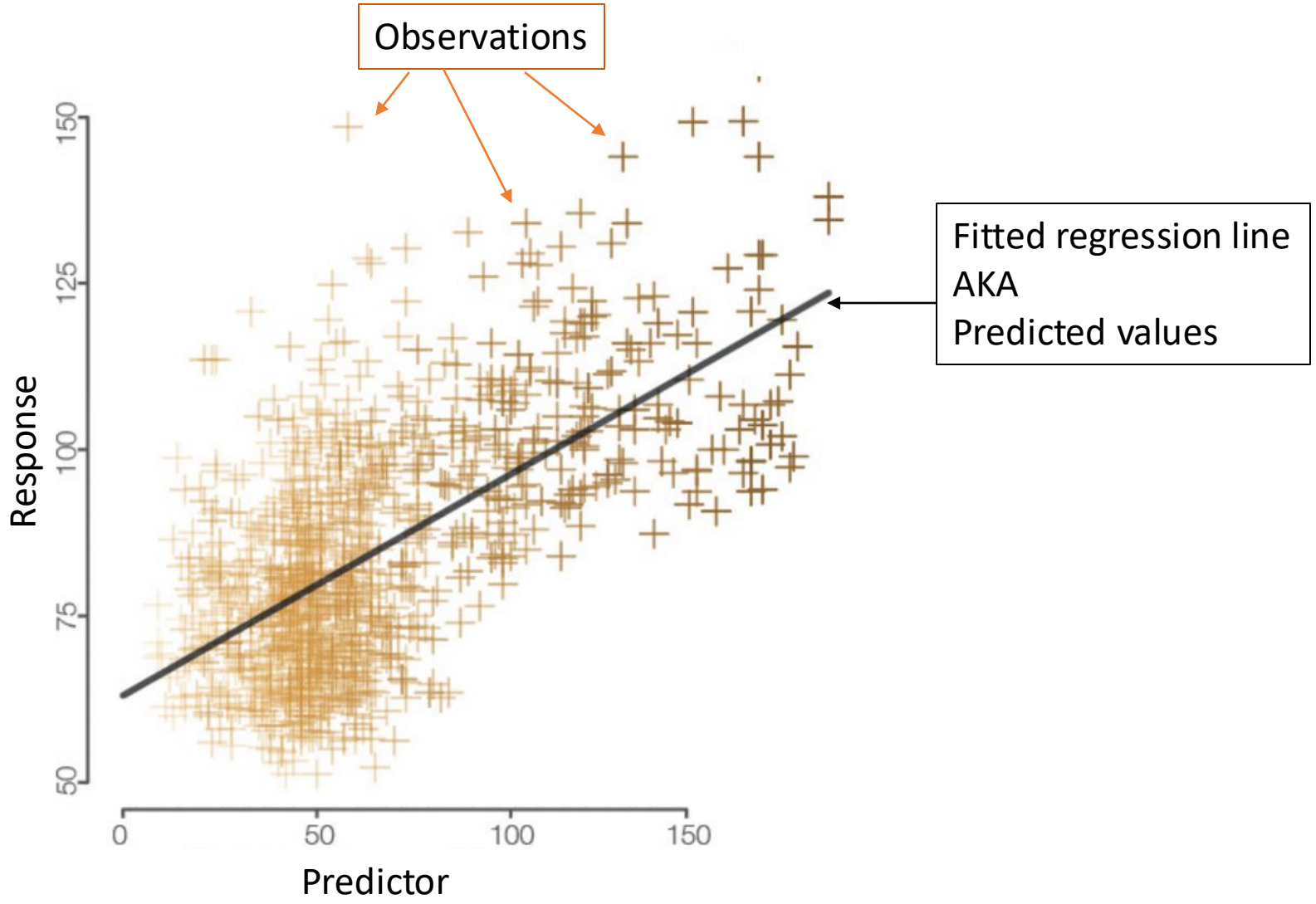
"All models are wrong – but some are useful."

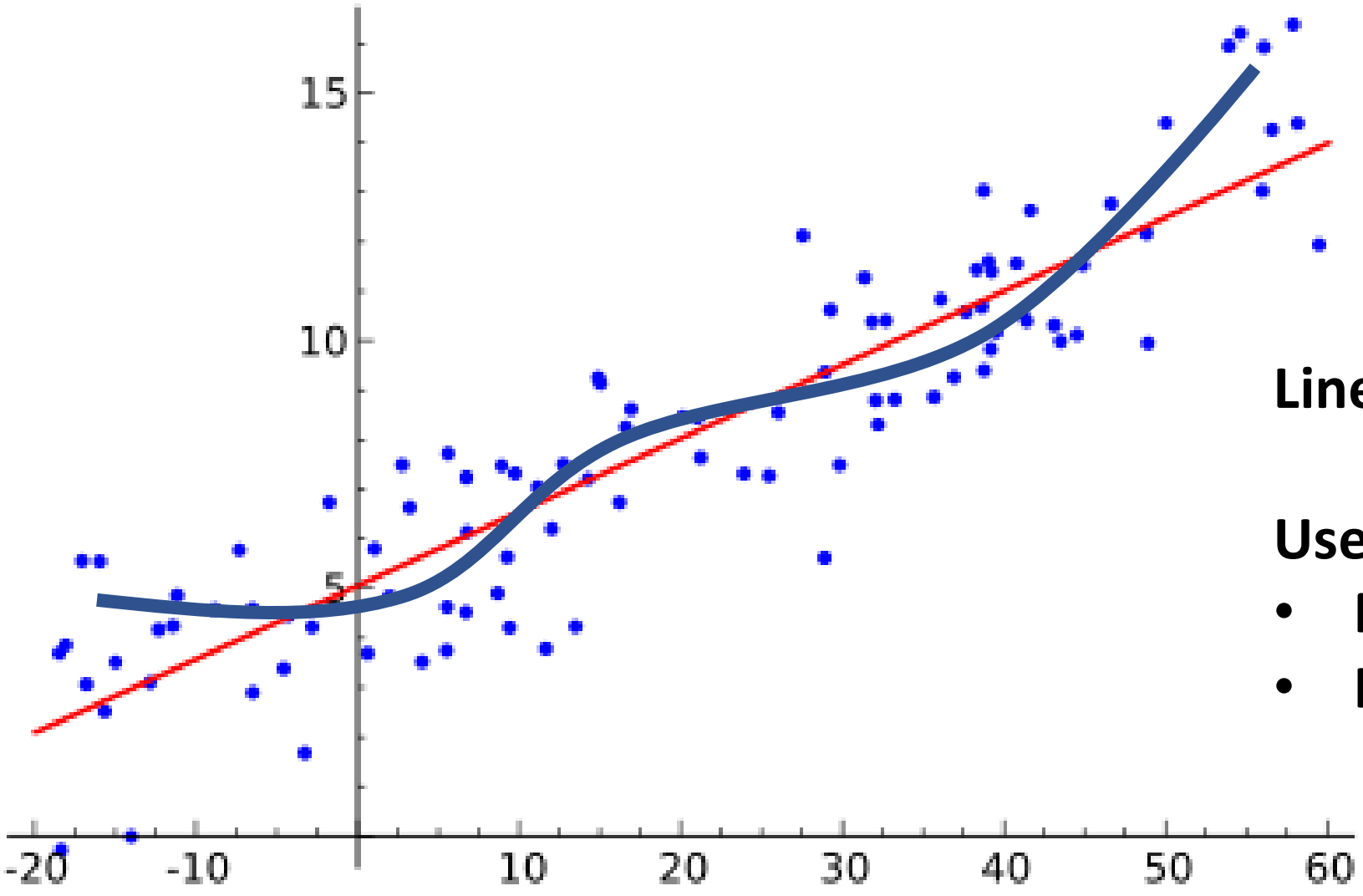
-- George Box



“Regression” described in 1886

Compared to the the height of their parents, heights of children was more similar to (i.e “**regressed**” to) the population mean


Observations
Fitted regression line
AKA
Predicted values
Response
Predictor
50
75
100
125
150
0
50
100
150



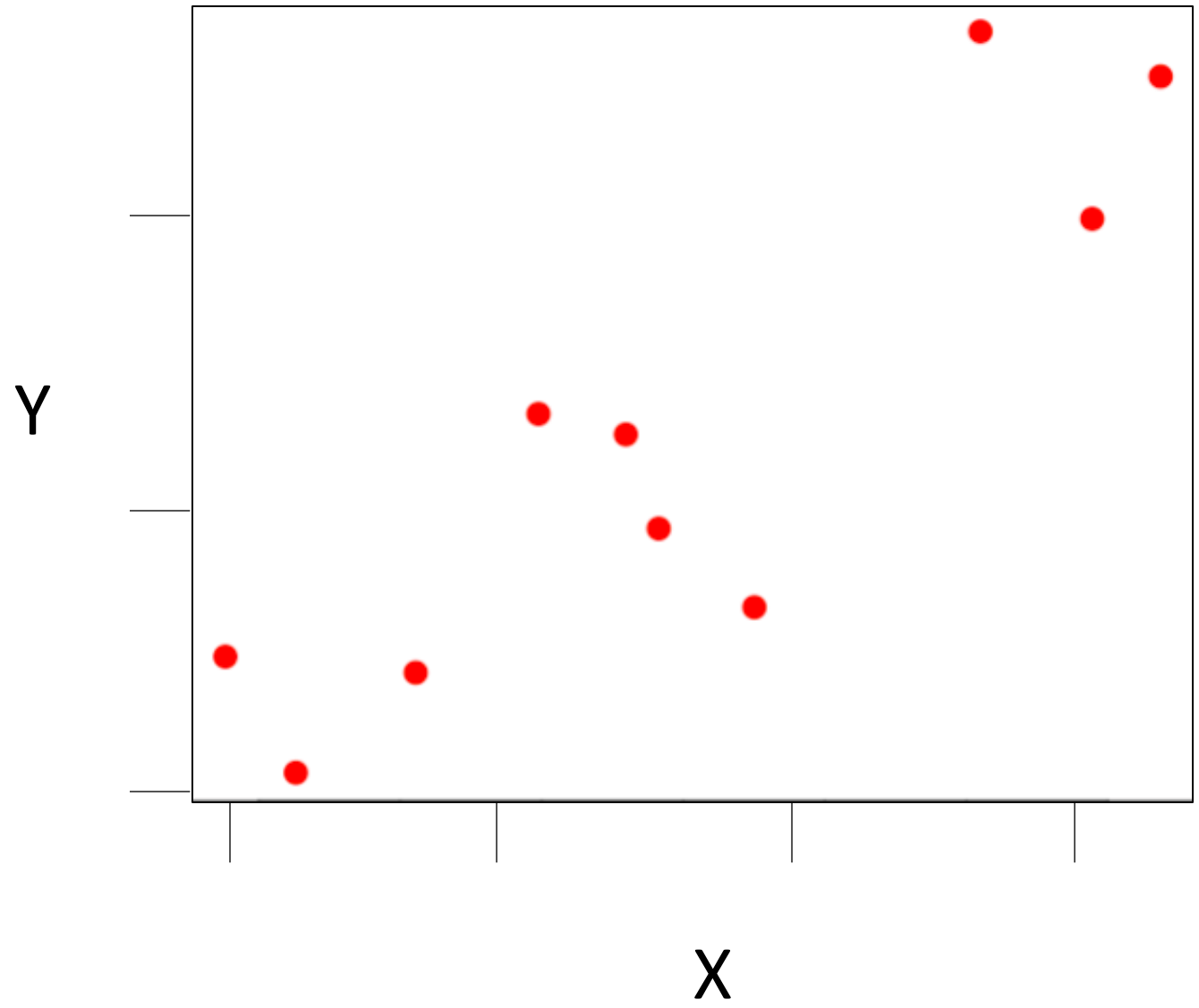
Linear regression is simple

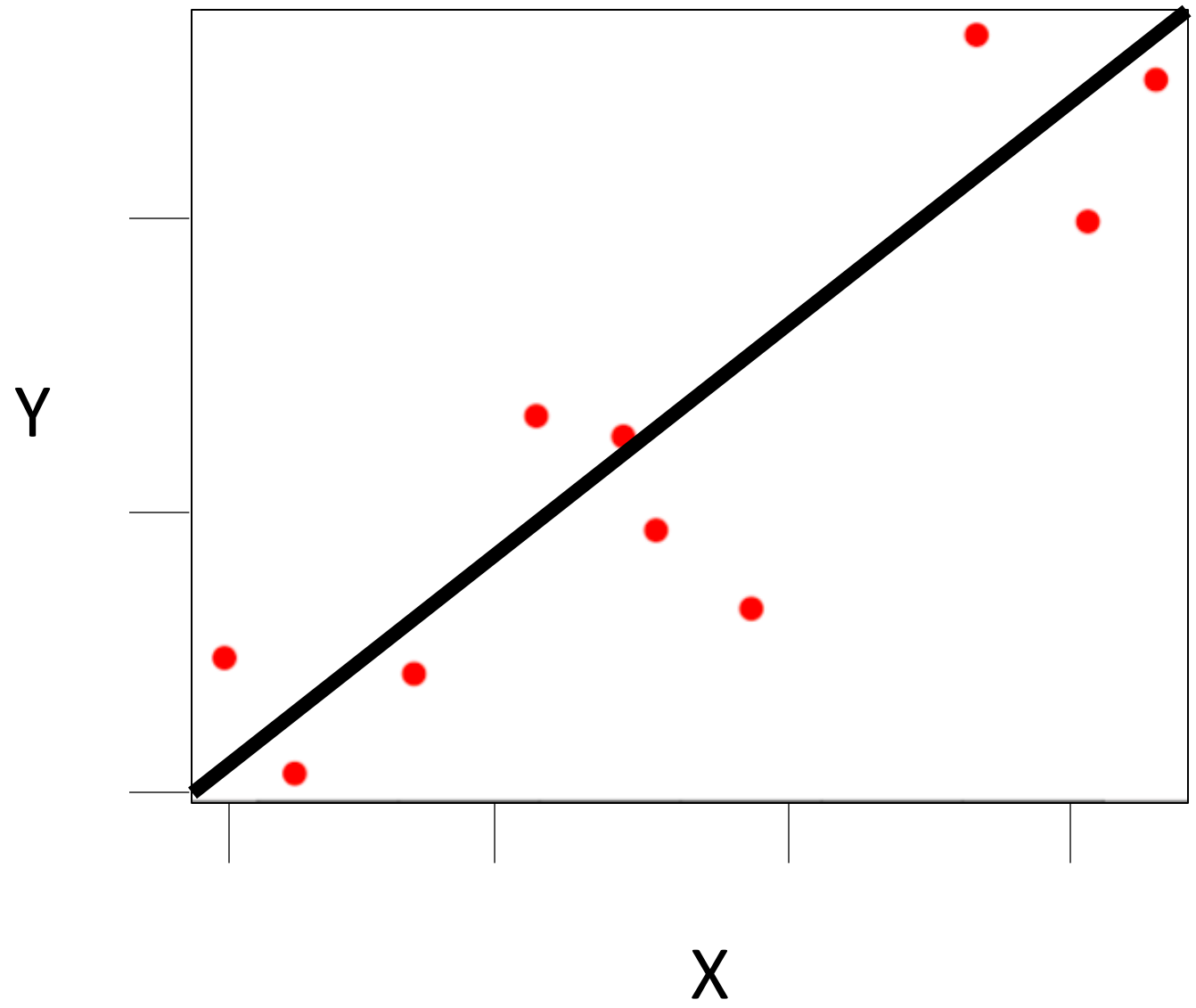
Used for

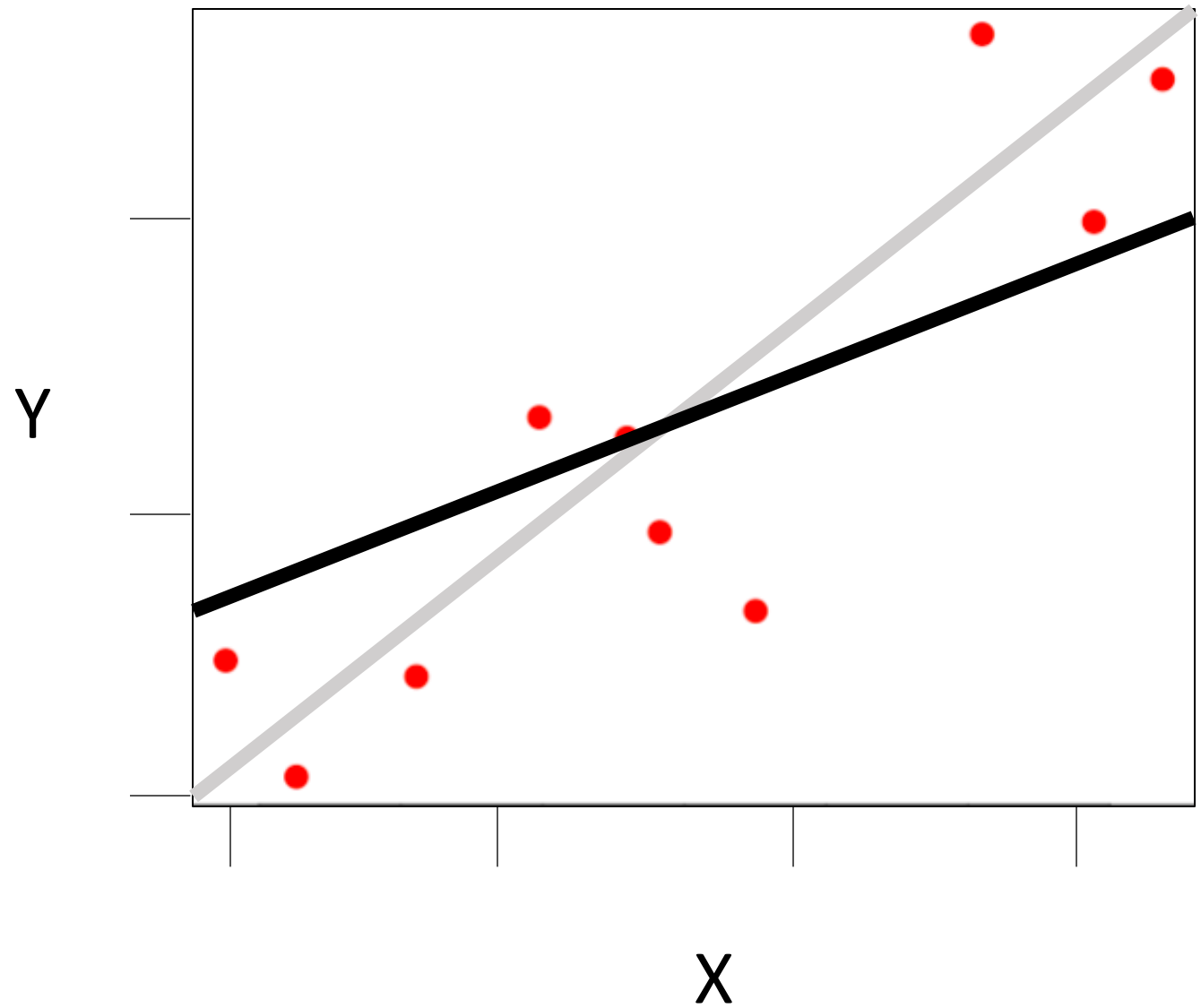
- **Making predictions**
- **Hypothesis testing**

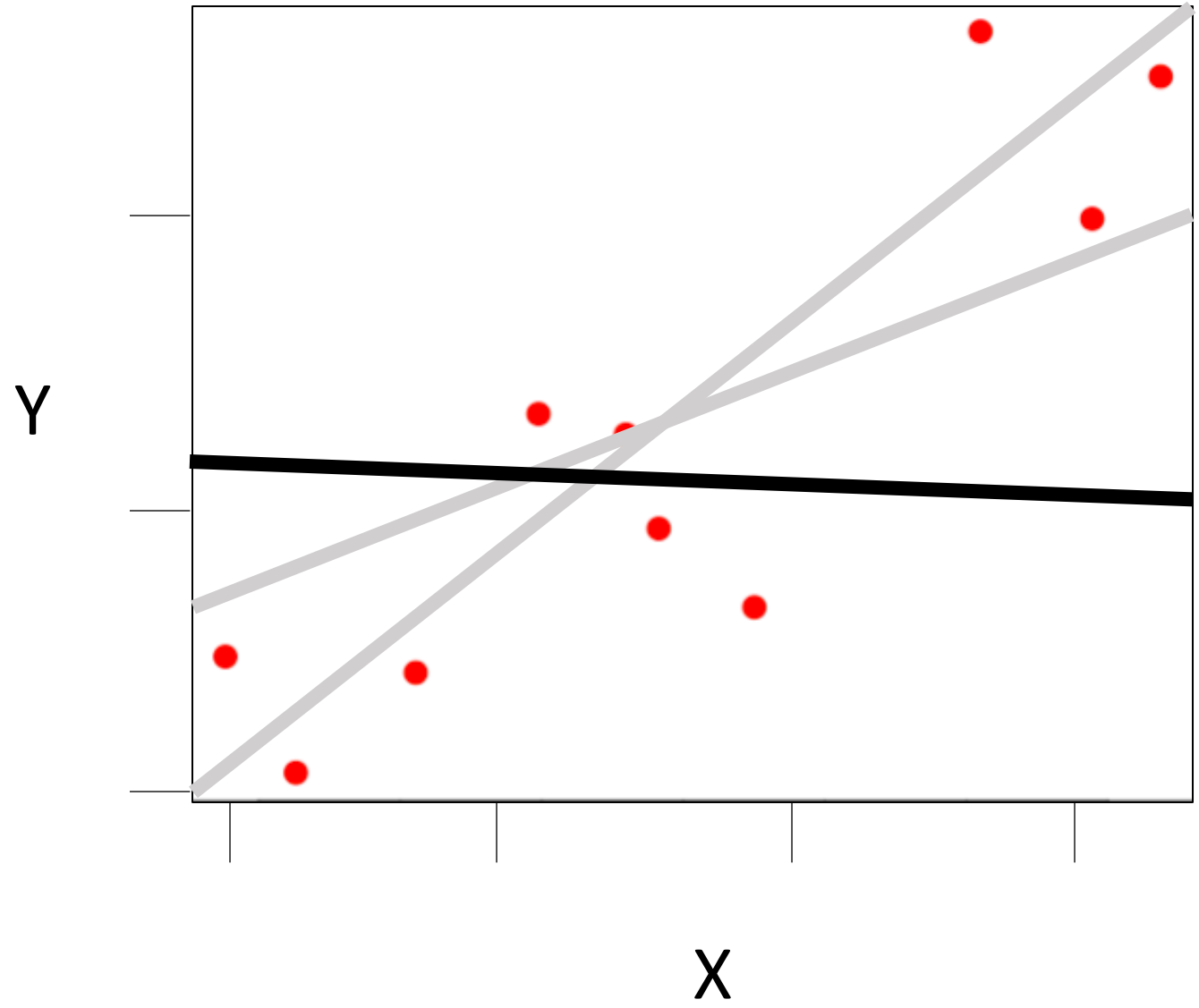
PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

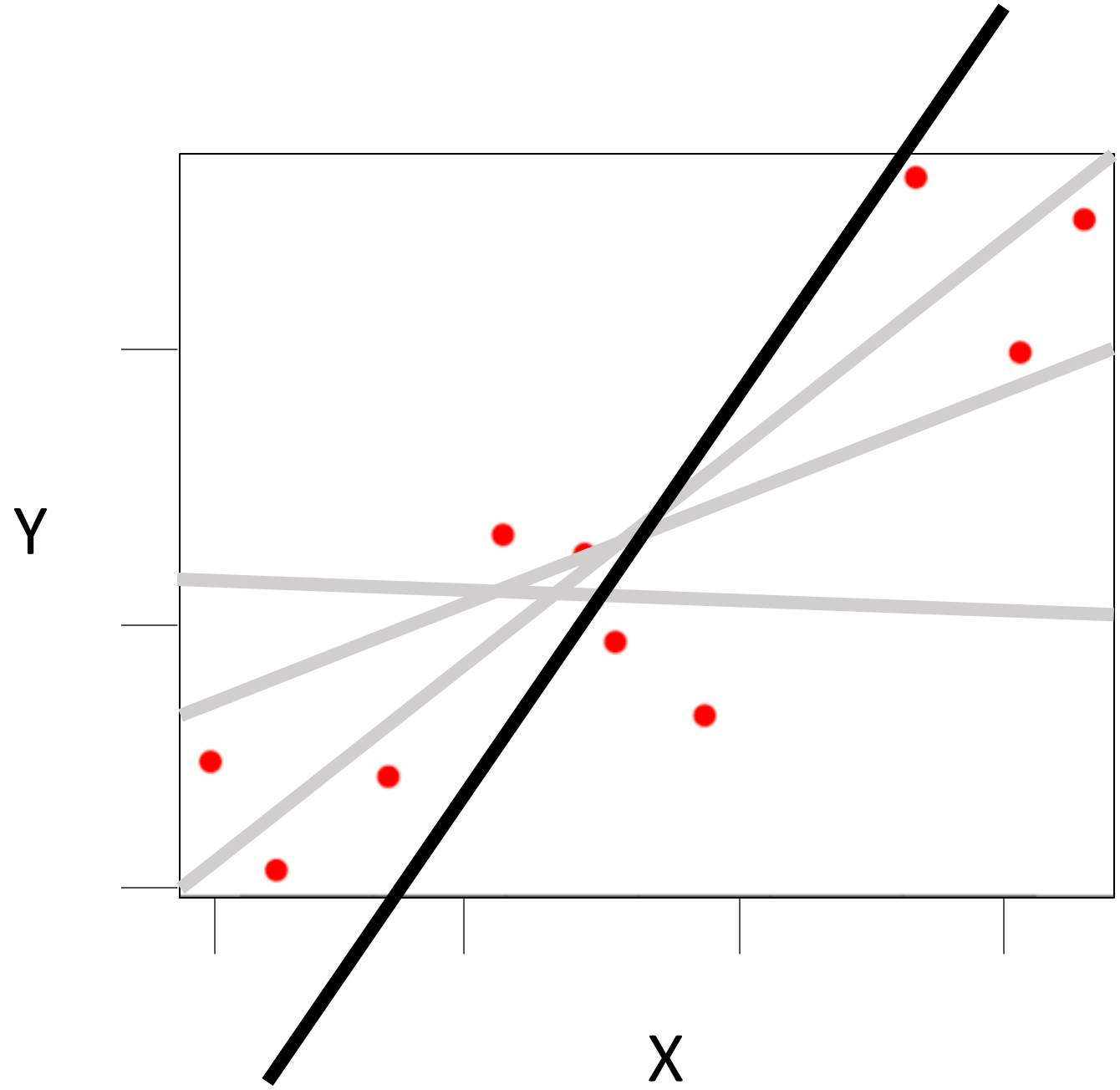
- What is linear regression?
- Residuals & residual sum of squares (RSS)
- What are the assumptions of linear regression?
- R^2 (proportion of variance explained)
- Hypothesis testing

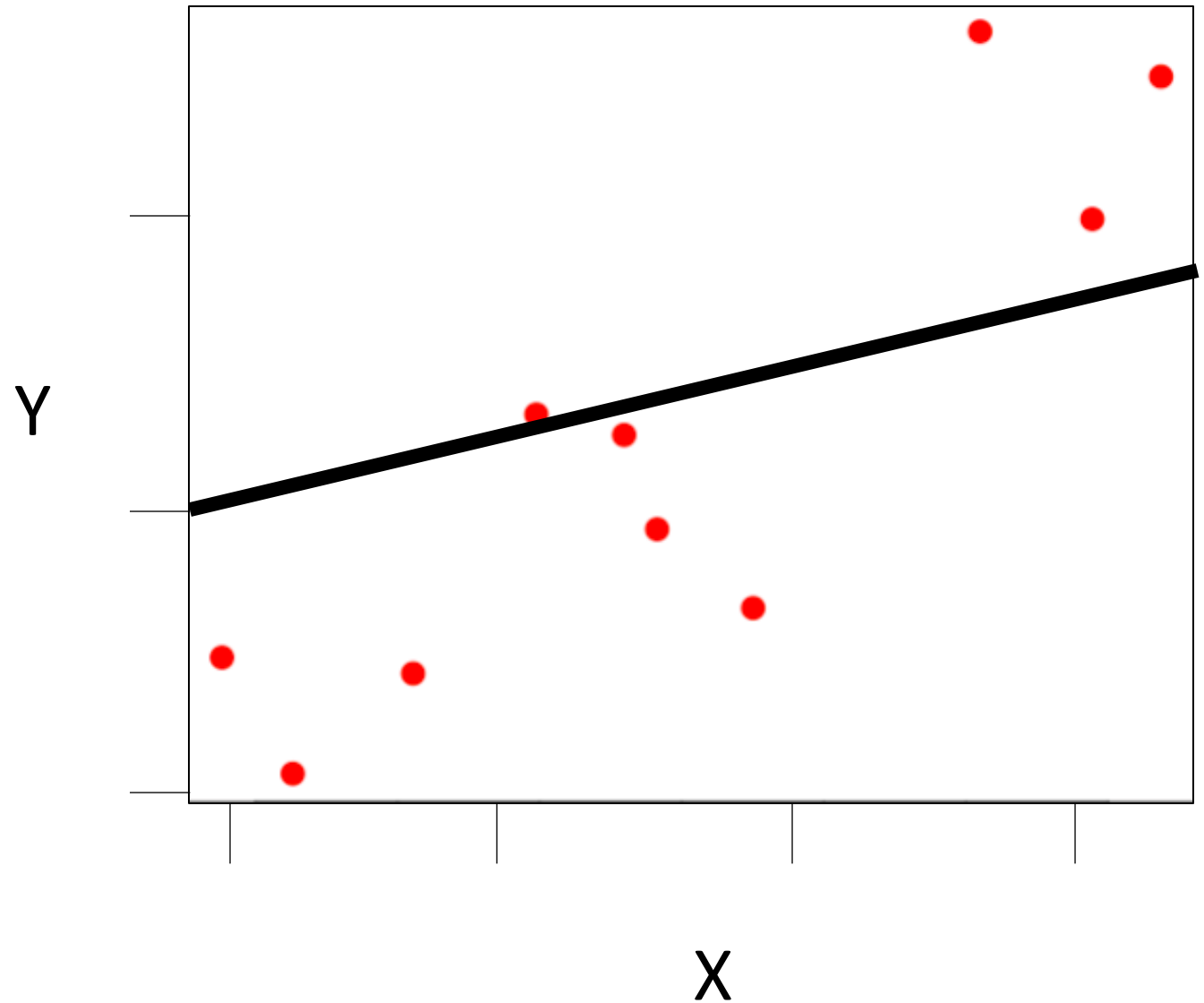

Y
X

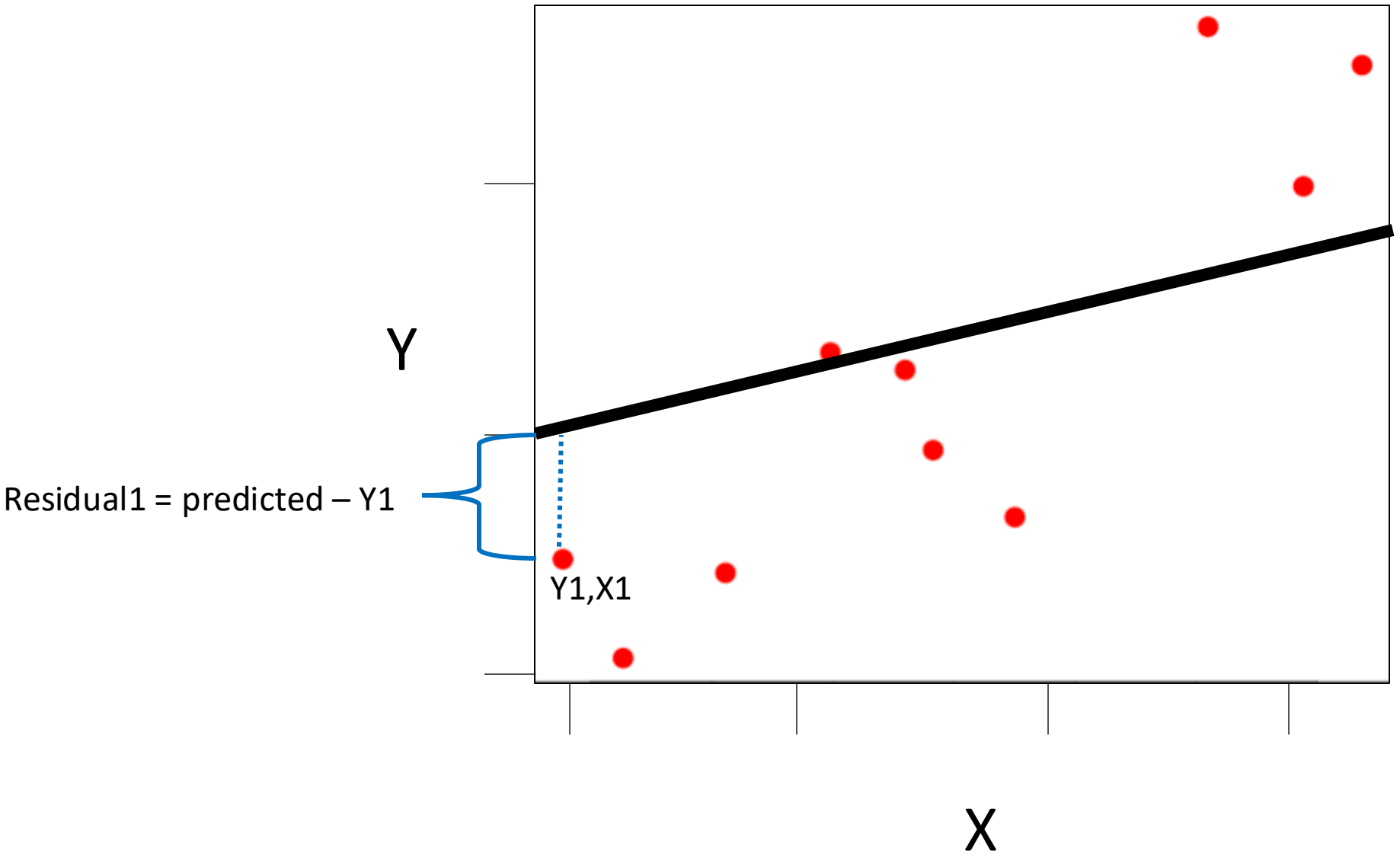

Y
X


Y
X

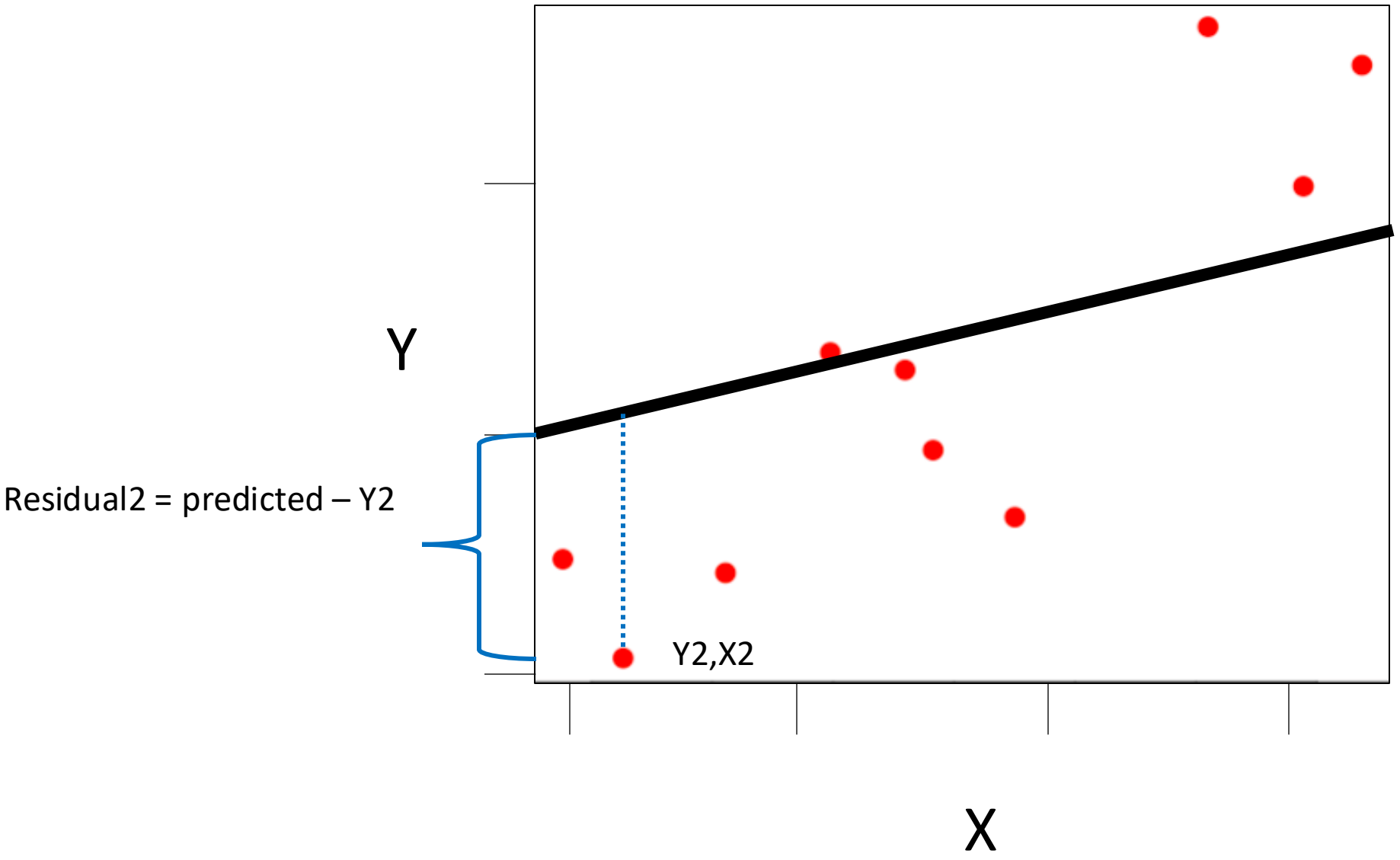

Y
X


Y
X

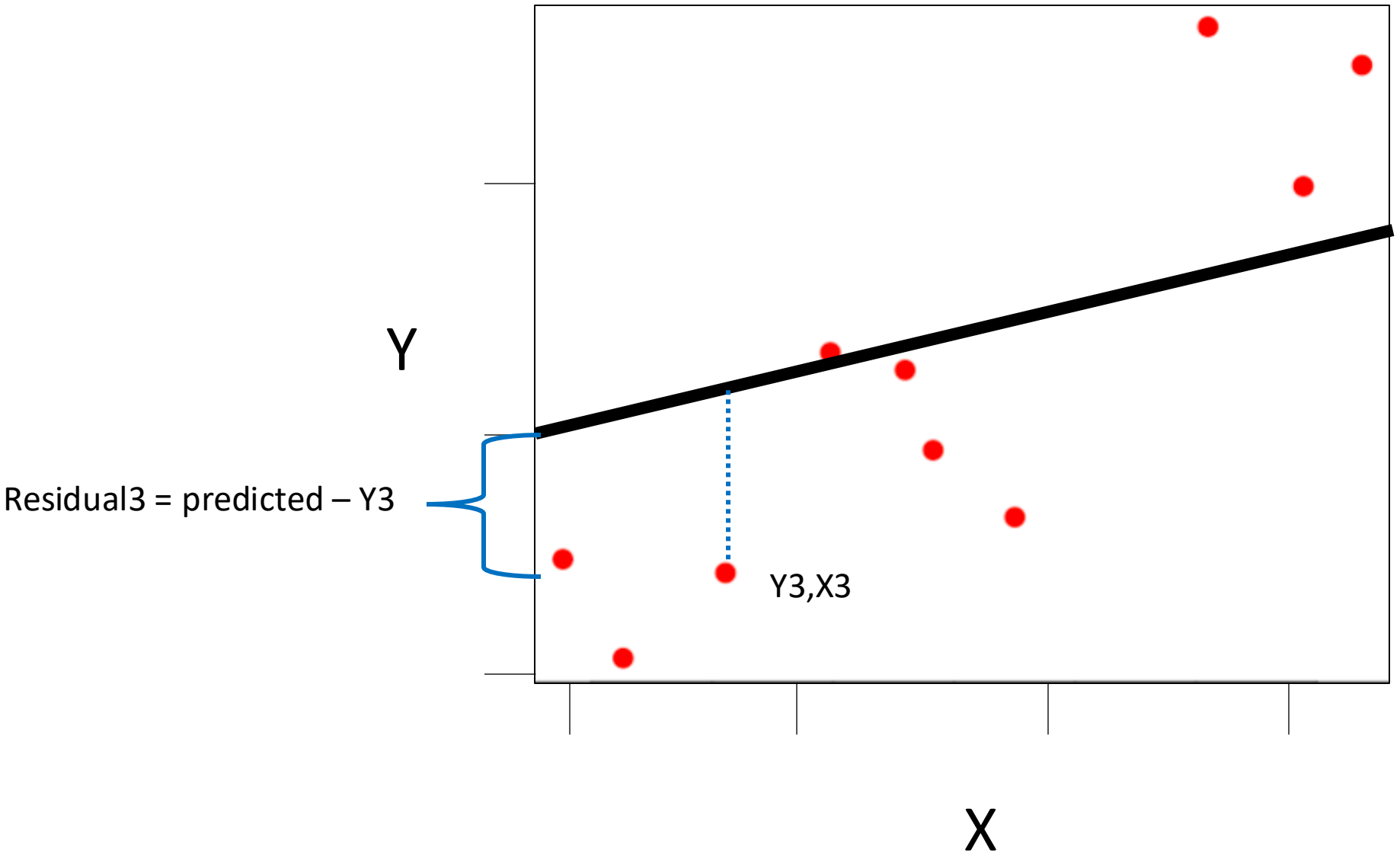

Y
X


Y
Residual1 = predicted – Y1
Y1,X1
X

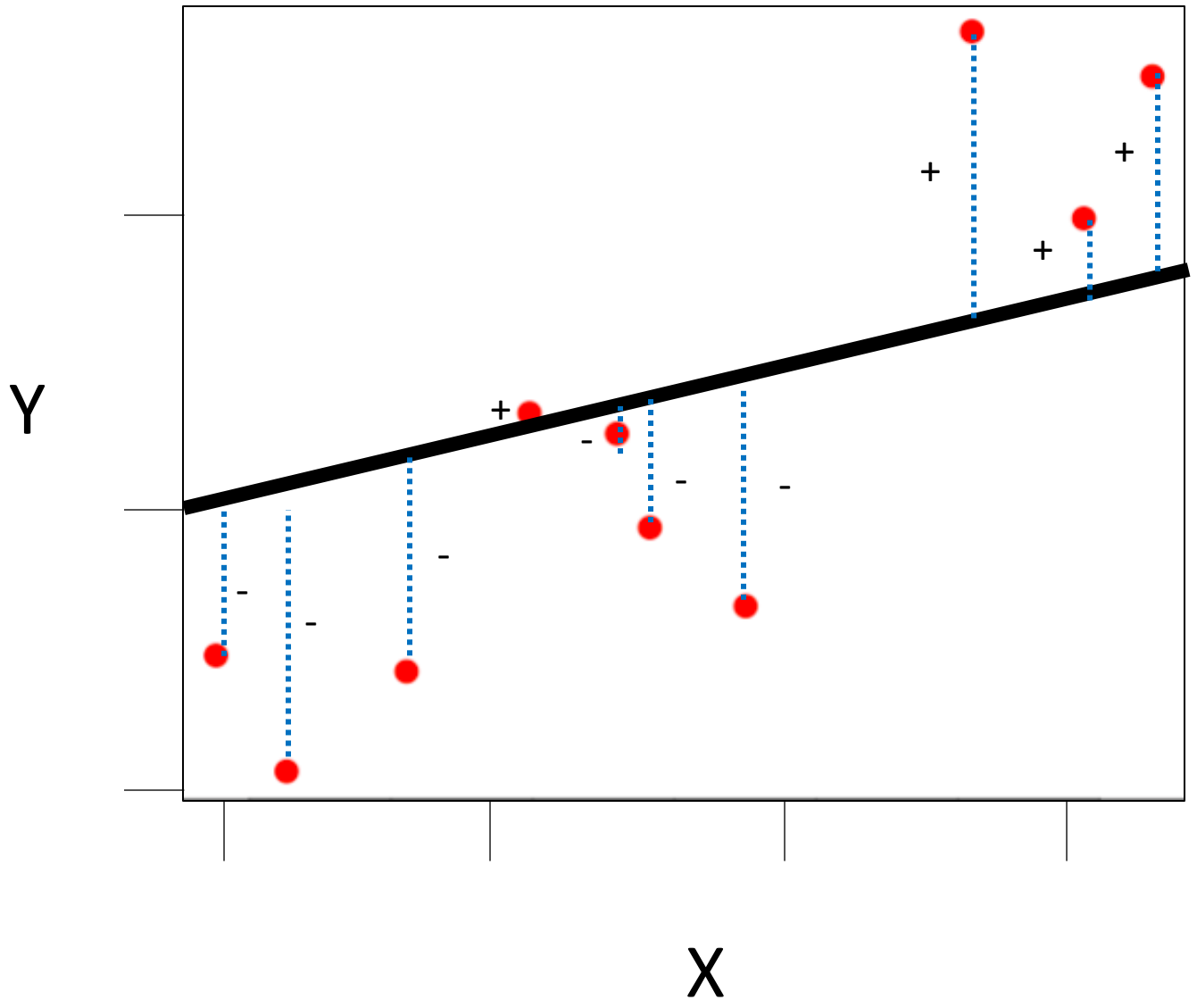
Residual1 + Residual2



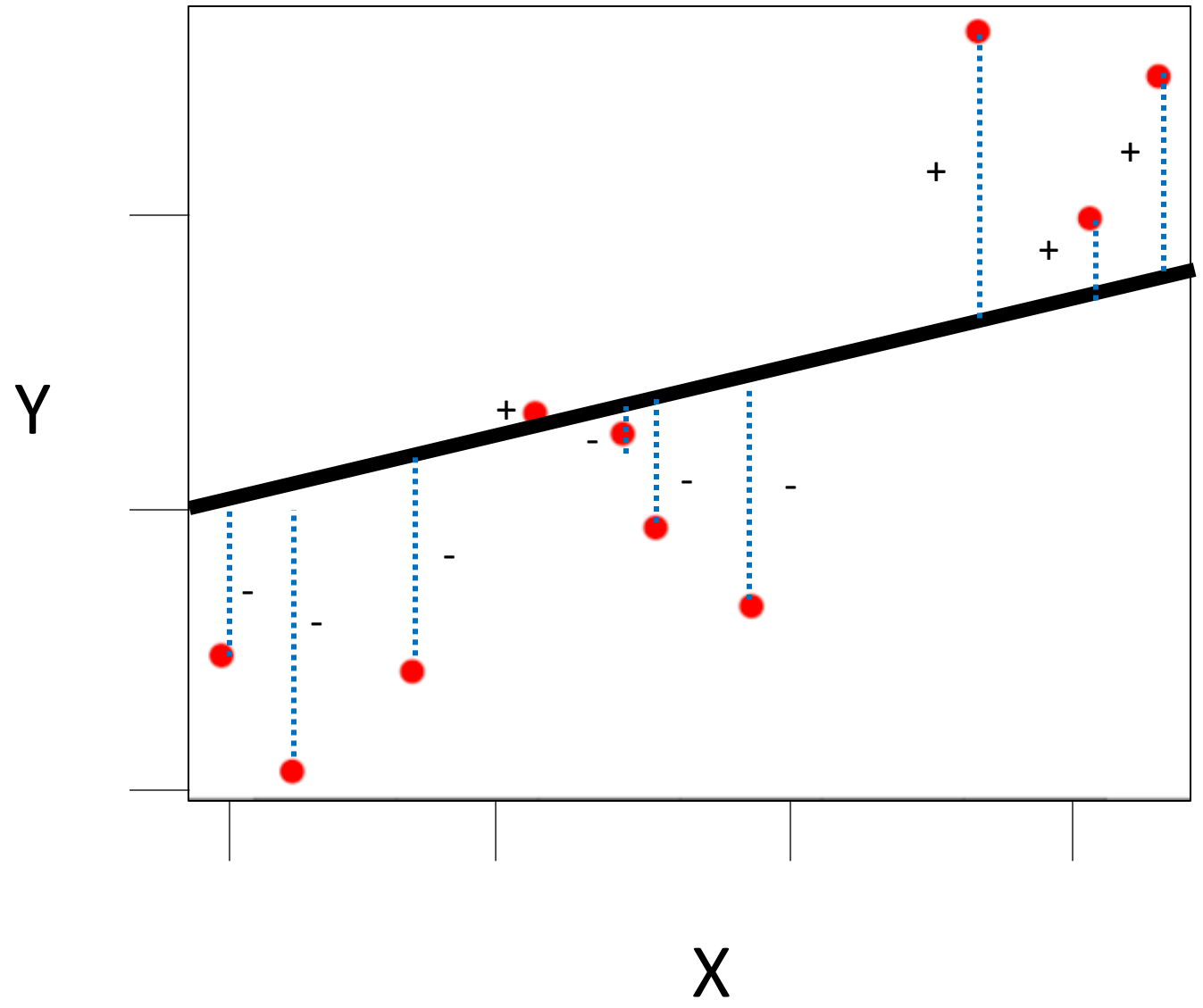
Residual1 + Residual2 + Residual3



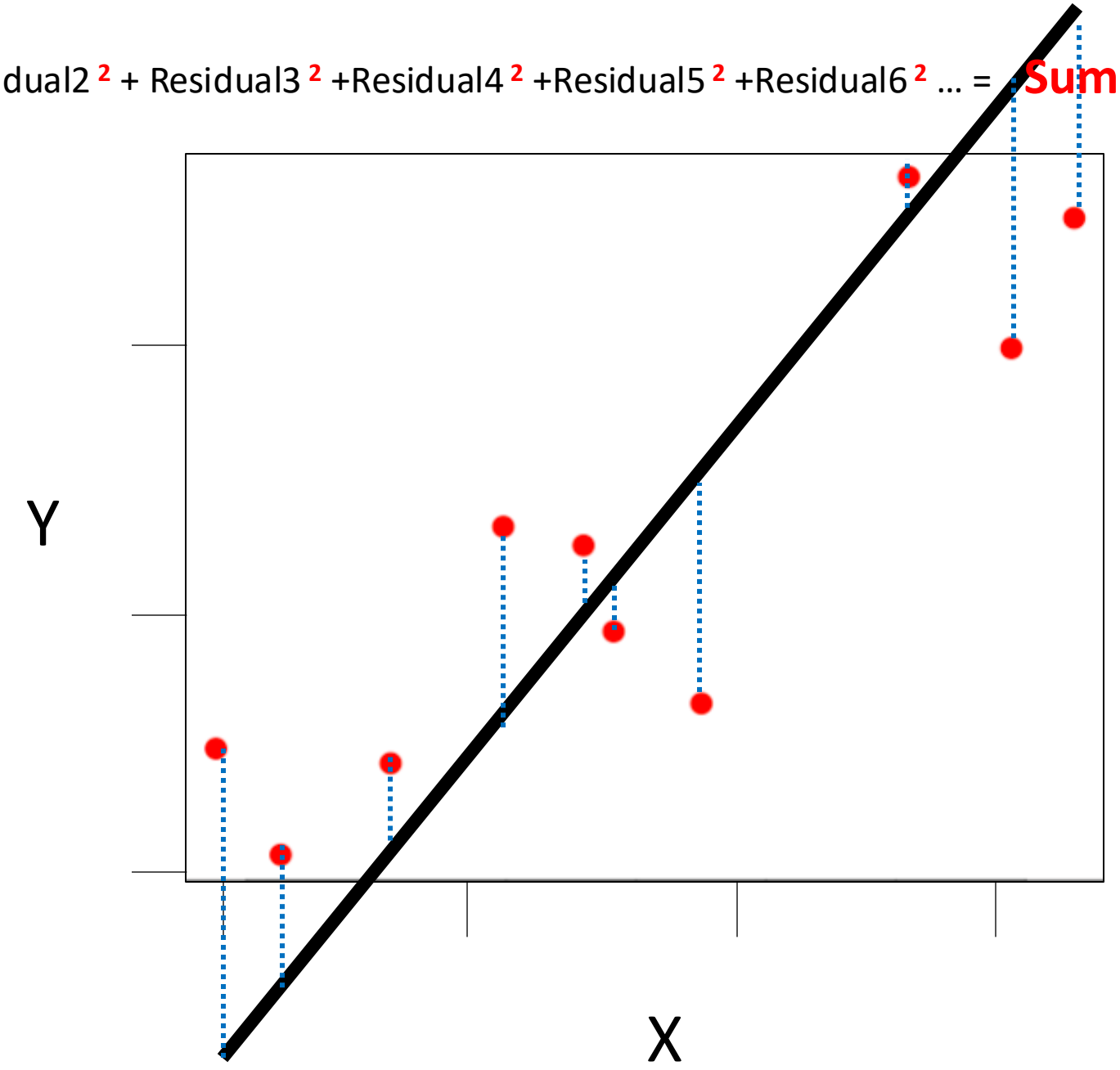
Residual1 + Residual2 + Residual3 +Residual4 +Residual5 +Residual6 …



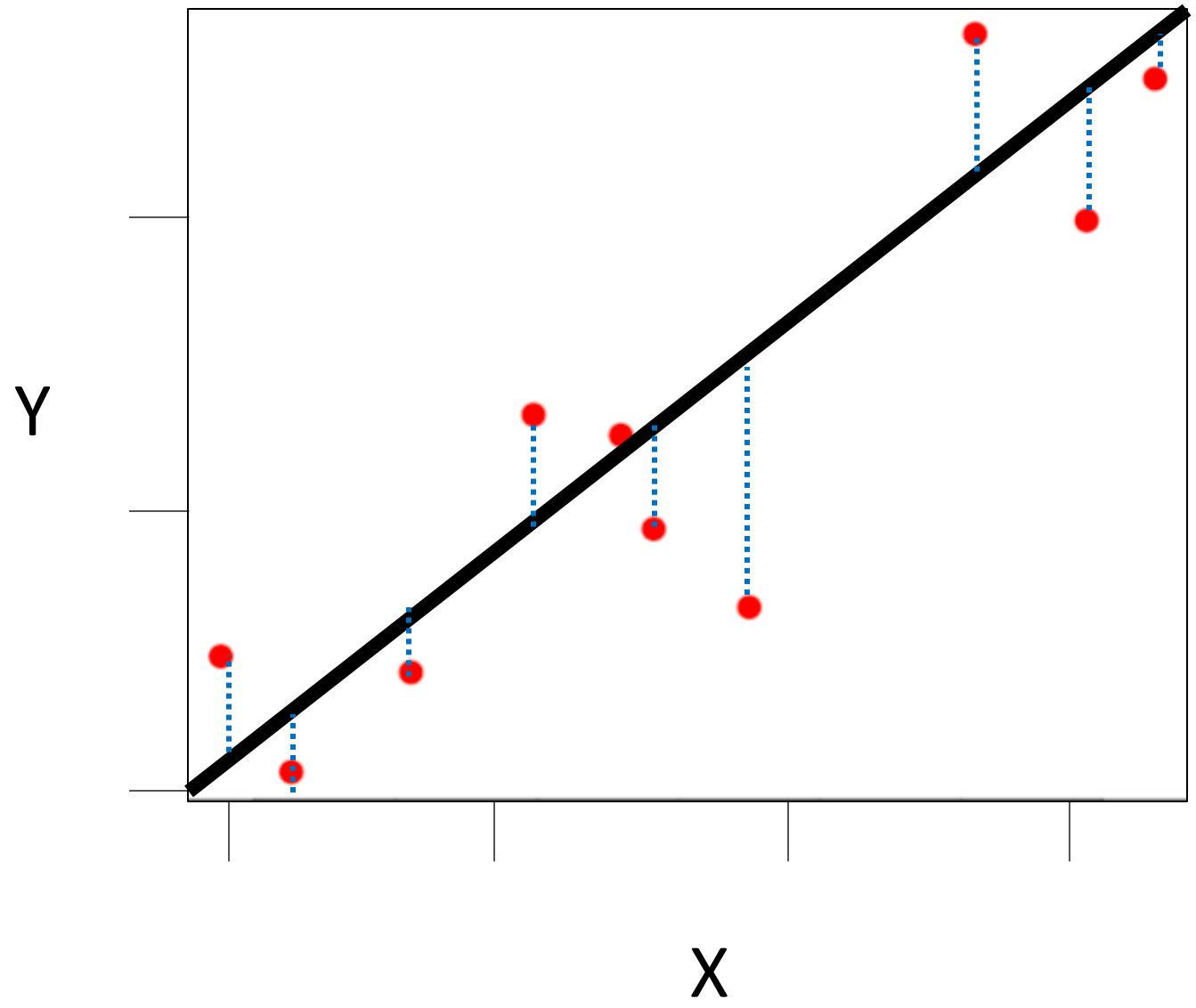
Residual1 ^{2} + Residual2 ^{2} + Residual3 ^{2} +Residual4 ^{2} +Residual5 ^{2} +Residual6 ^{2} … = **Sum of Squares**



$\text{Residual1}^2 + \text{Residual2}^2 + \text{Residual3}^2 + \text{Residual4}^2 + \text{Residual5}^2 + \text{Residual6}^2 \ldots =$ **Sum of Squares**



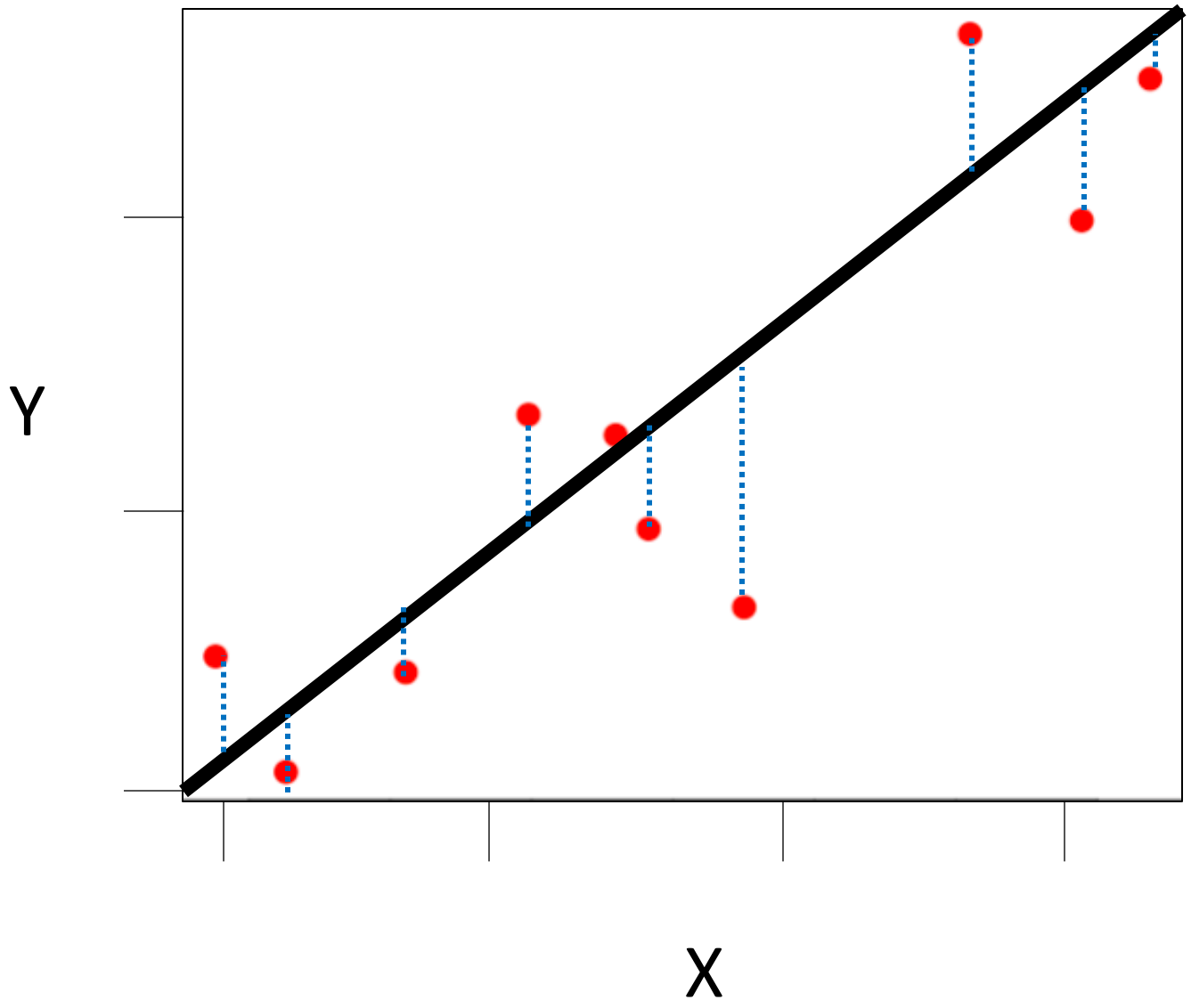
Residual1 ^{2} + Residual2 ^{2} + Residual3 ^{2} +Residual4 ^{2} +Residual5 ^{2} +Residual6 ^{2} … = **Sum of Squares**



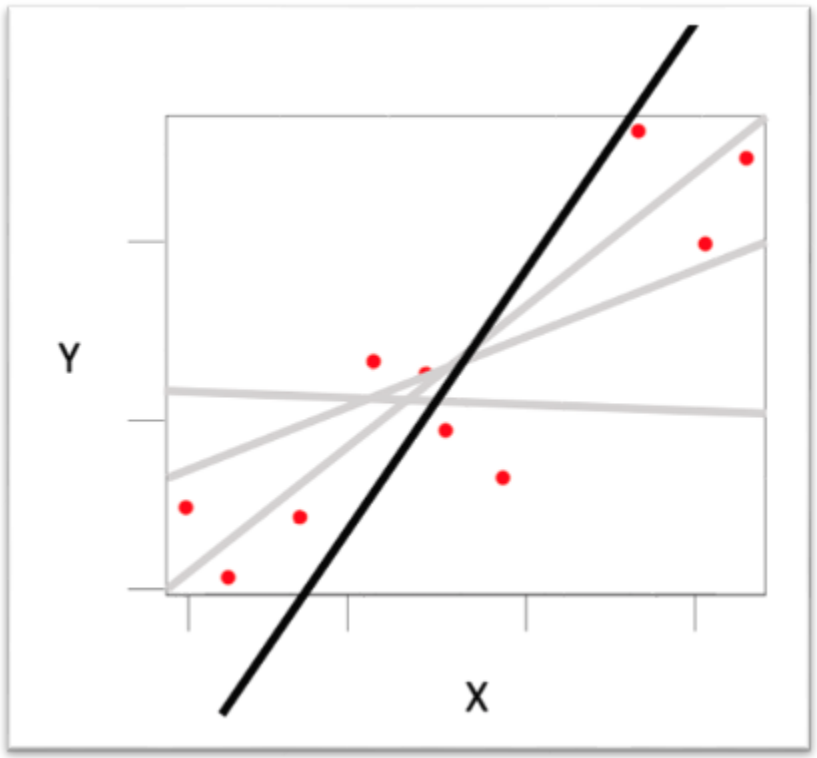
Regression is finding the line that produces the least sum of squares

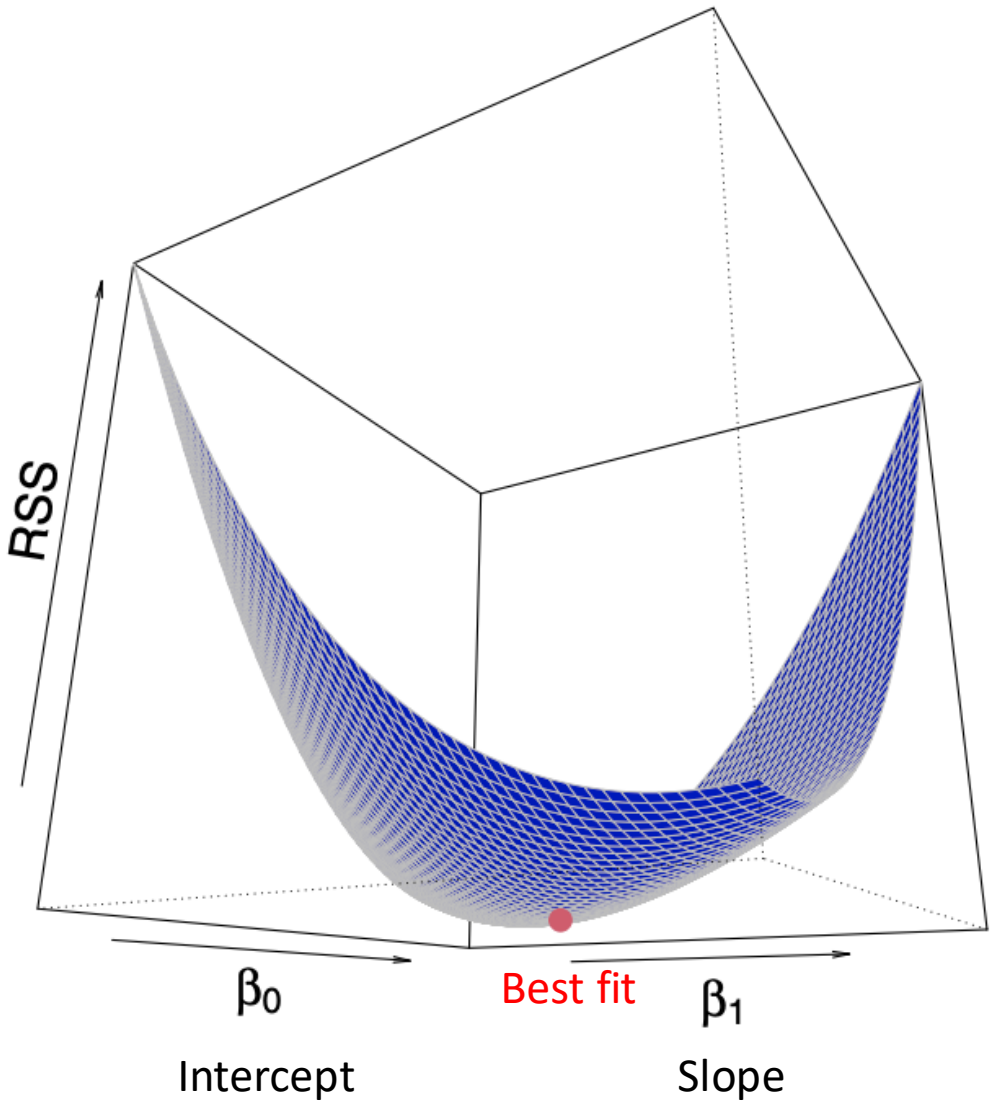
$\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \varepsilon_4^2 + \varepsilon_5^2 + \varepsilon_6^2 \ldots =$ **Residual Sum of Squares (RSS)**

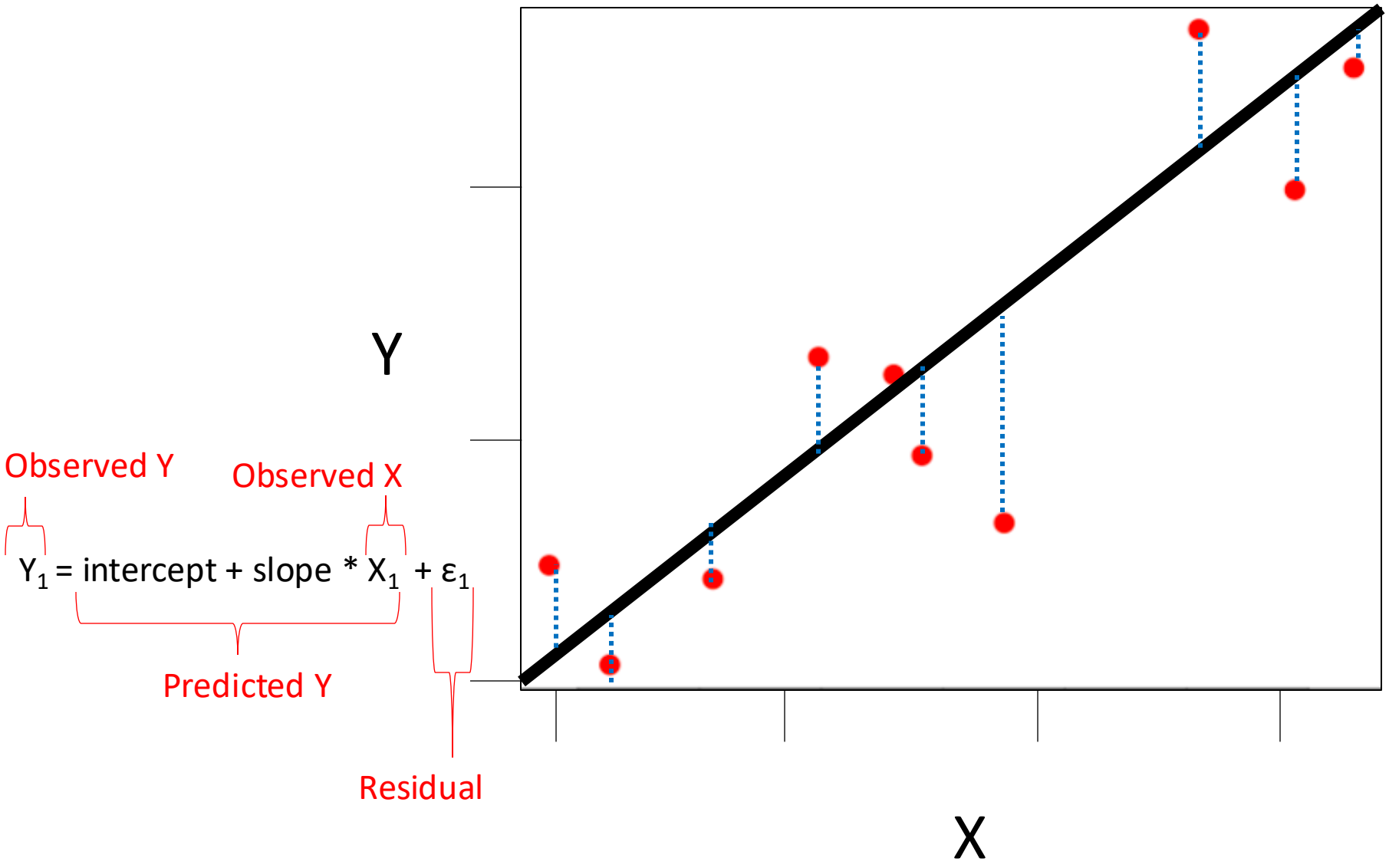
ε (residuals)



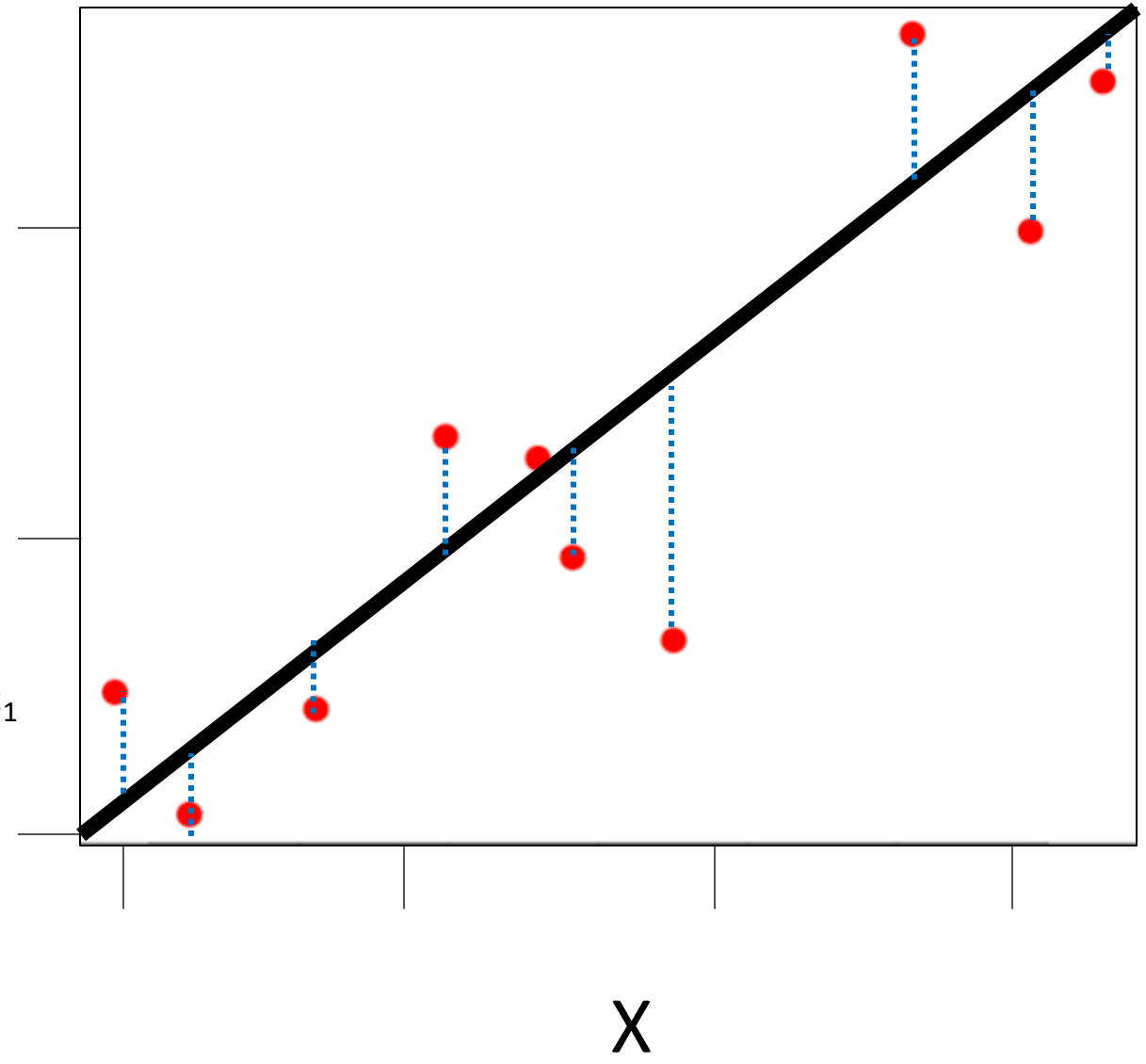
Regression is finding the line that produces the least sum of squares


Y
X


RSS
β0
Best fit
β1
Intercept
Slope

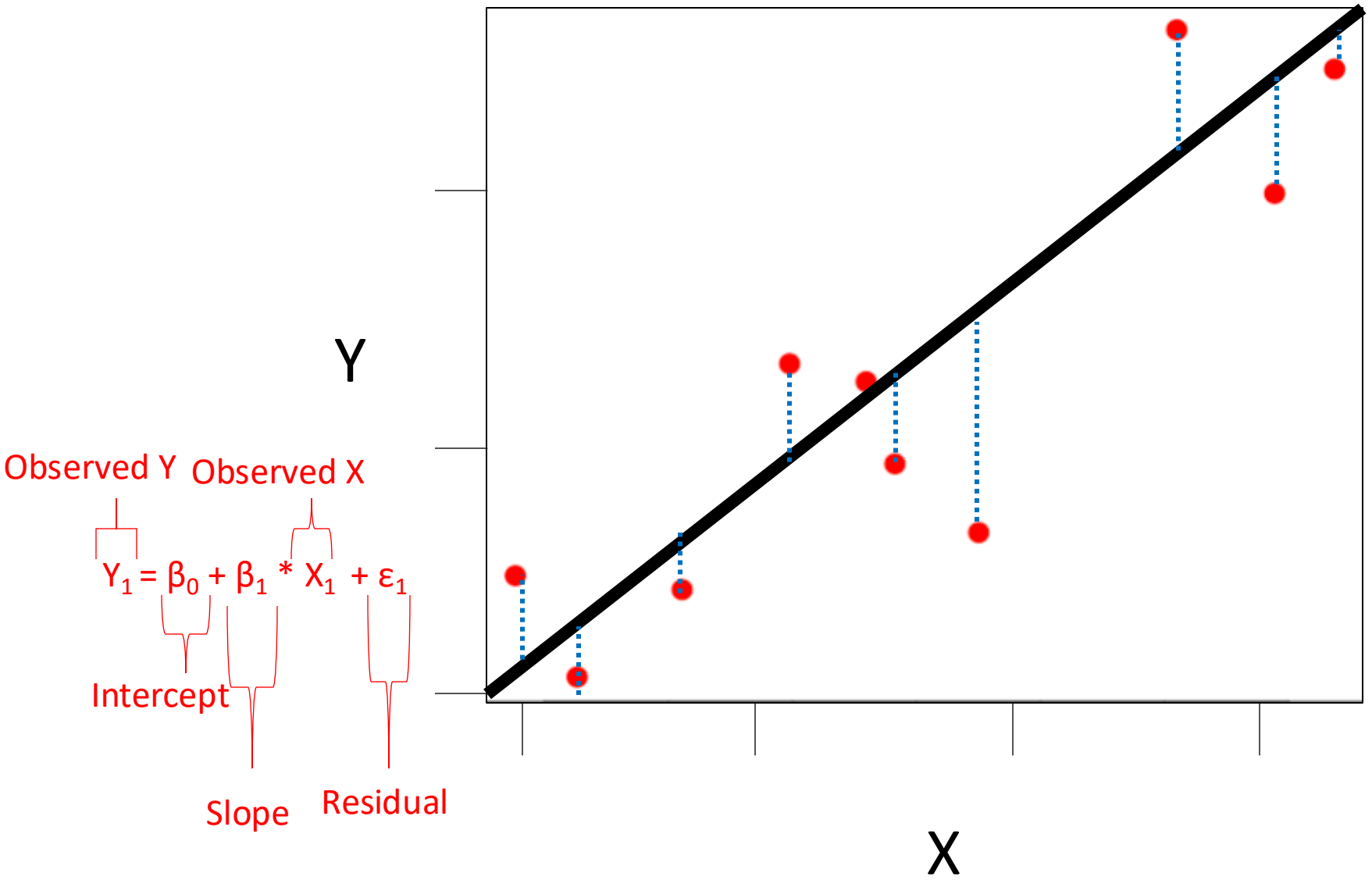


Regression is finding the line that produces the least sum of squares

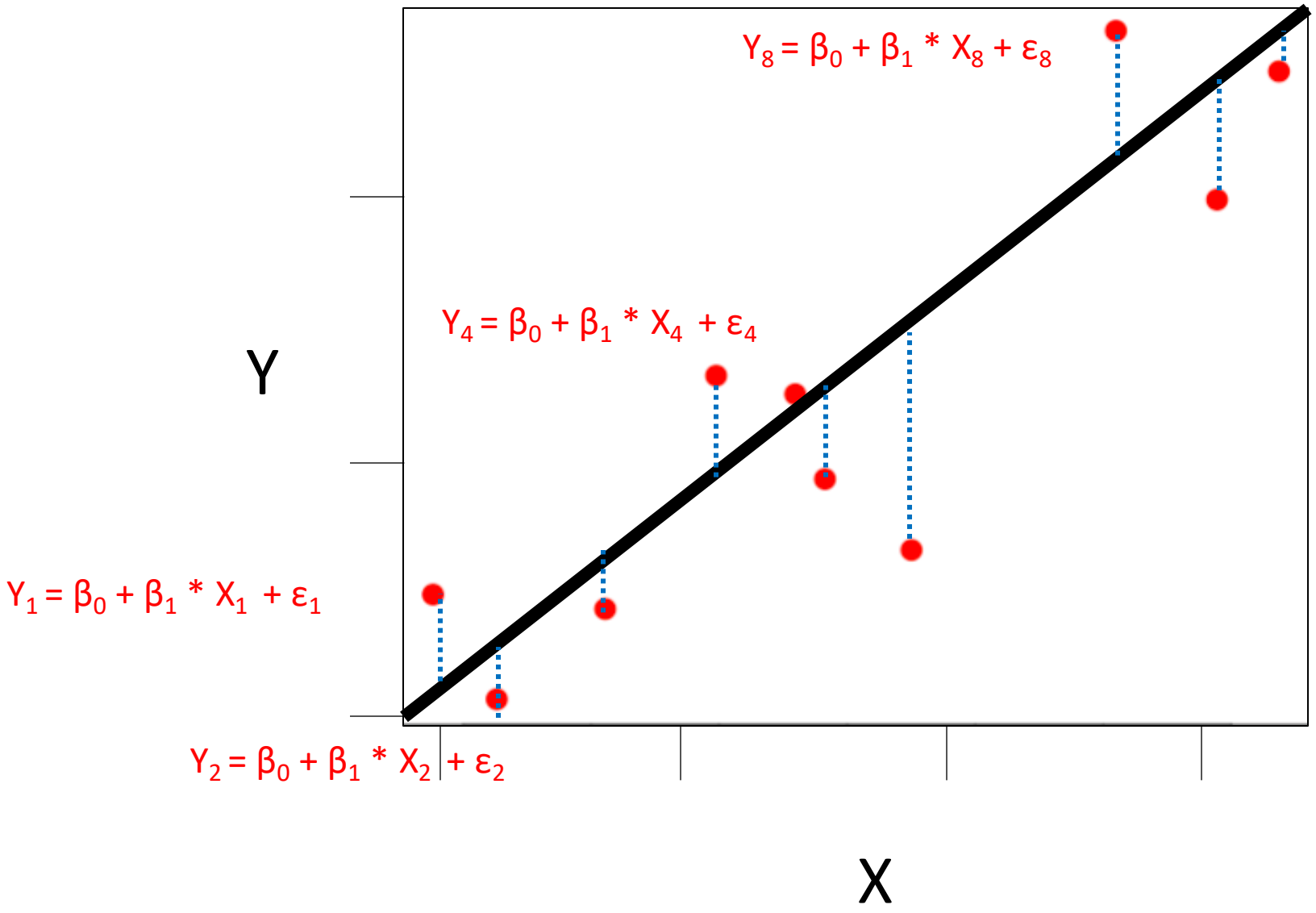


$Y_1 = \text{intercept} + \text{slope} * X_1 + \varepsilon_1$

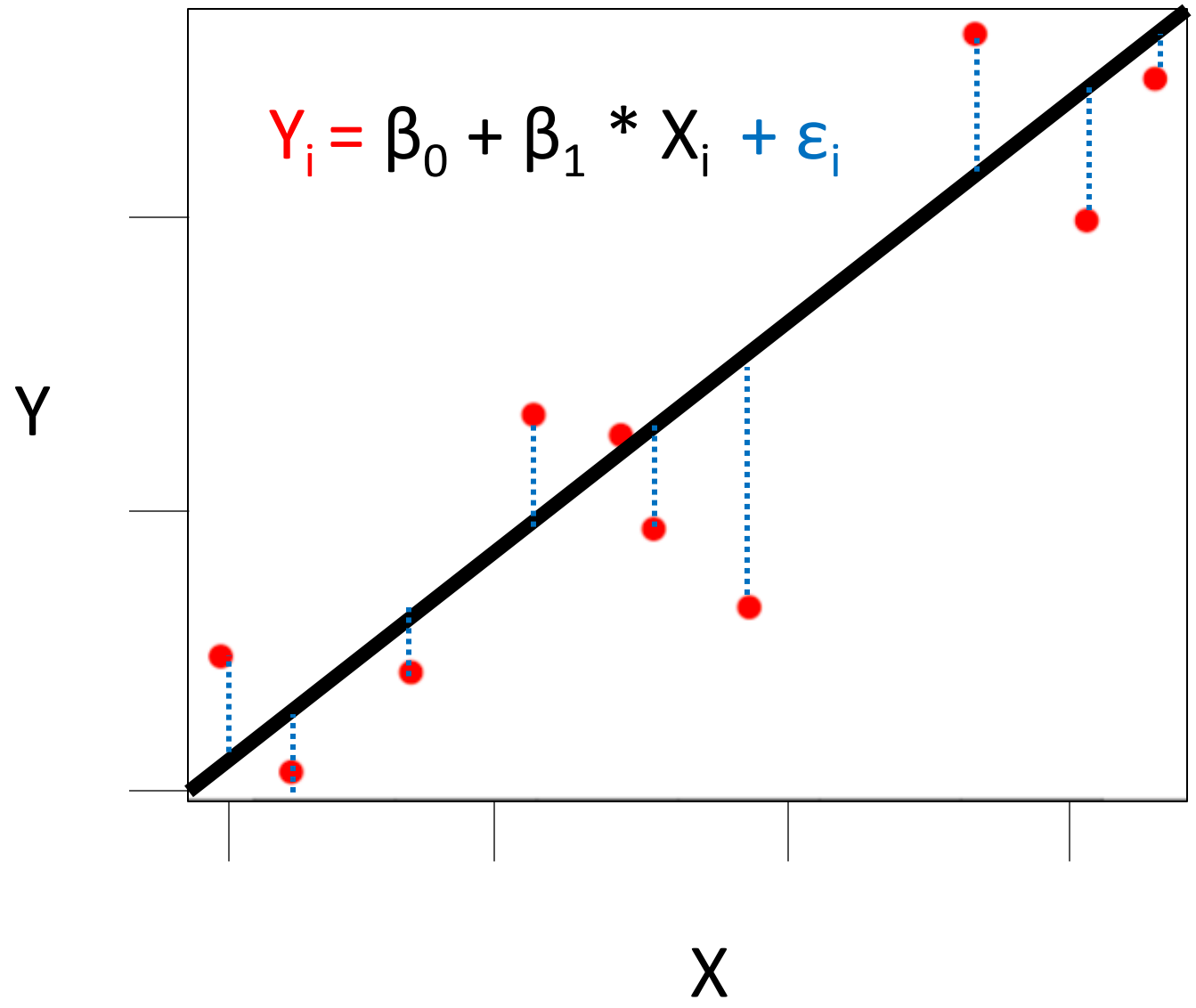
Regression is finding the line that produces the least sum of squares



Regression is finding the line that produces the least sum of squares

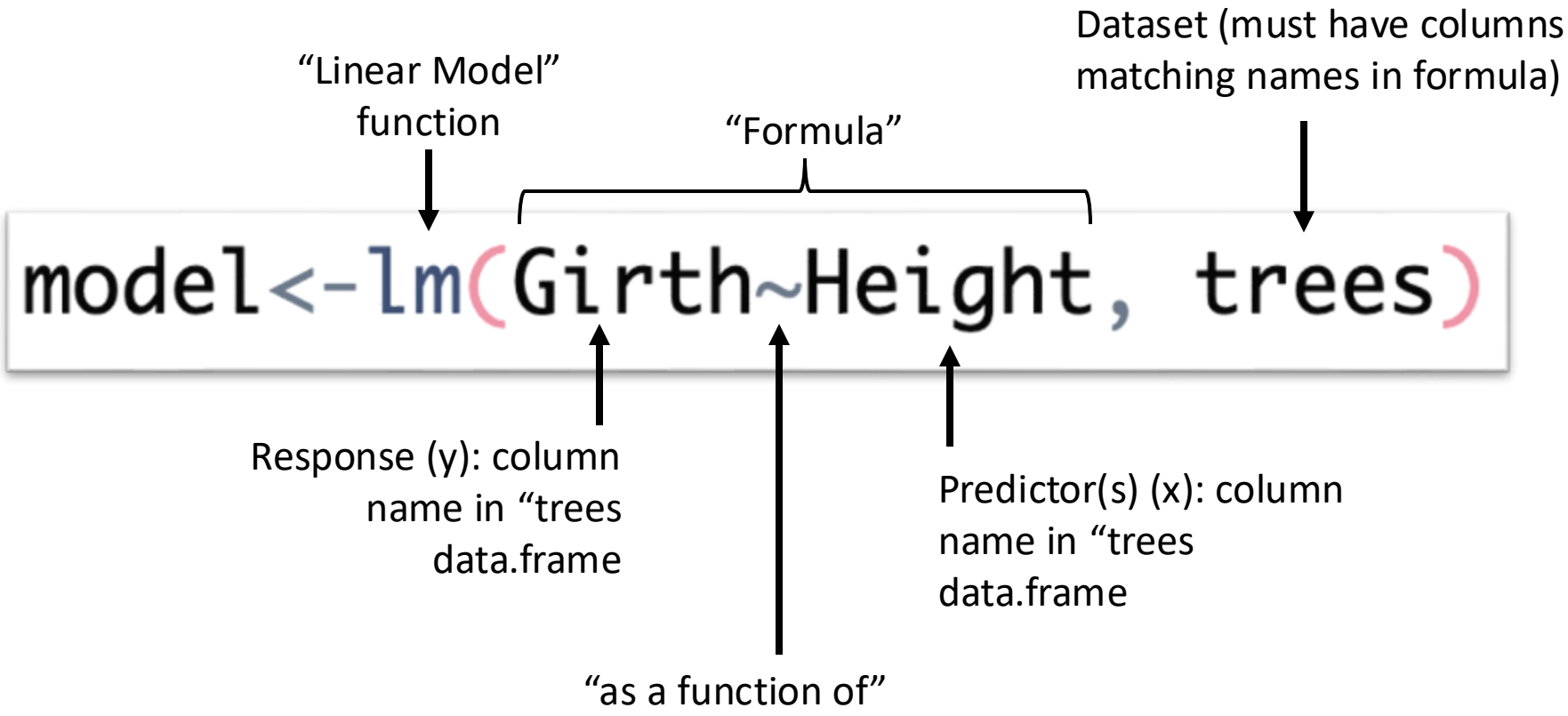


Regression is finding the line that produces the least sum of squares



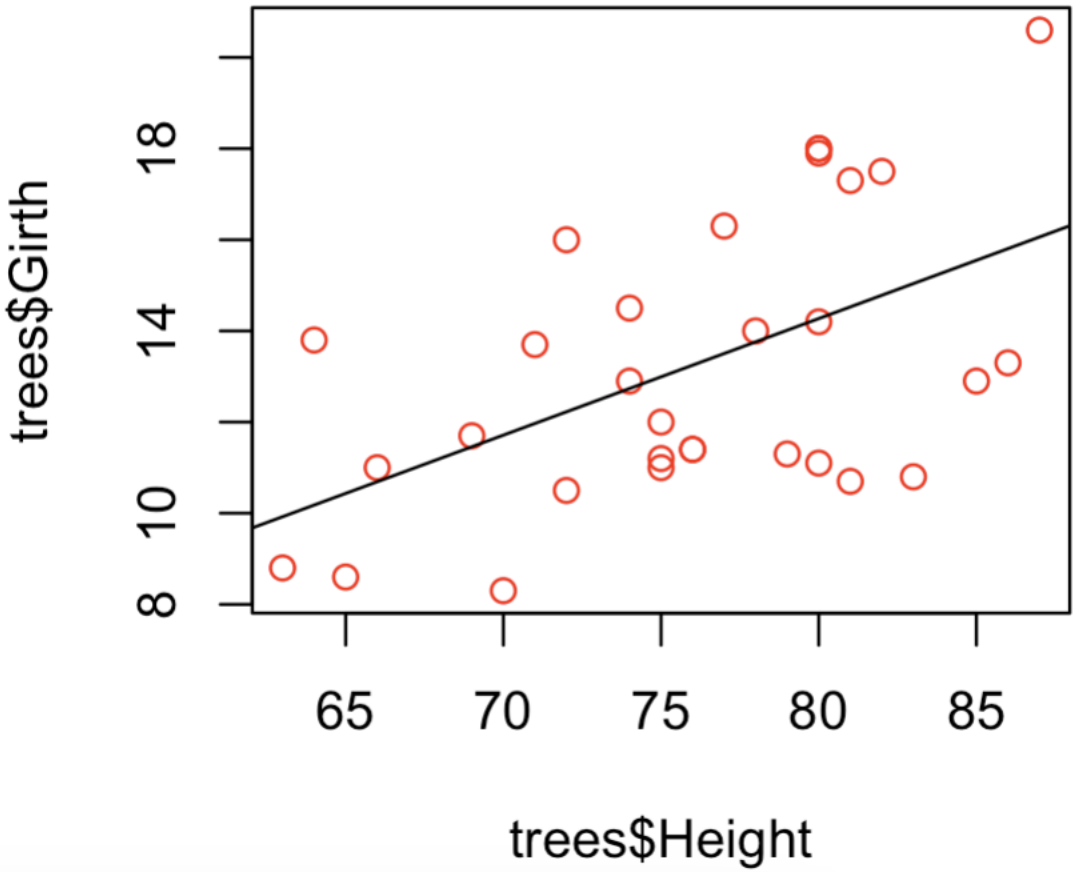
Regression is finding the line that produces the least sum of squares

Syntax of linear regression in R… **lm() function**

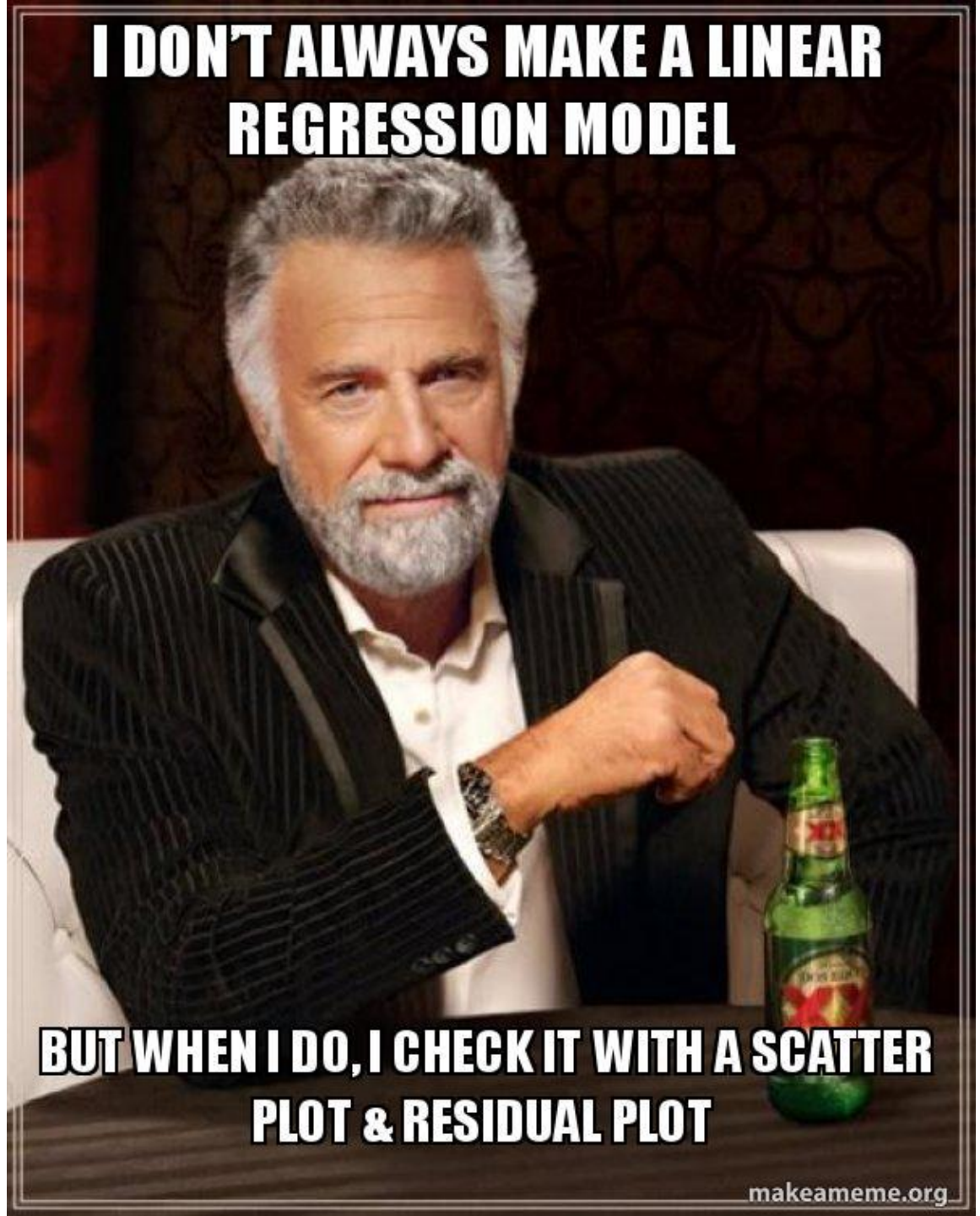


Linear regression in R. lm() function

```r
#plot a scatter plot between height and girth
plot(trees$Height, trees$Girth, col="red")

#create a linear (regression) model with lm()
model<-lm(Girth~Height, trees)
abline(model) #draw the regression line on the plot
```



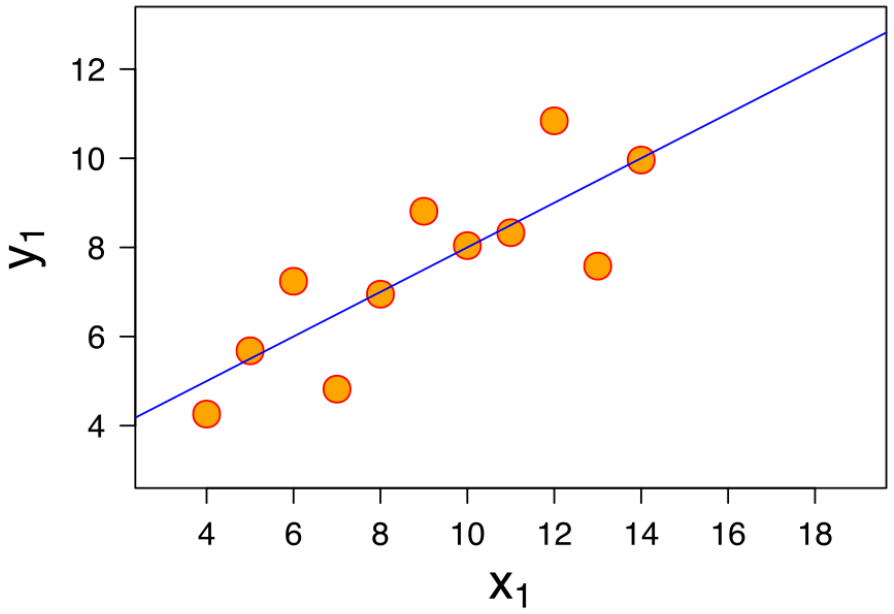

```
> summary(model)

Call:
lm(formula = Girth ~ Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2386 -1.9205 -0.0714  2.7450  4.5384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.18839    5.96020  -1.038  0.30772
Height       0.25575    0.07816   3.272  0.00276 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.728 on 29 degrees of freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:  0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

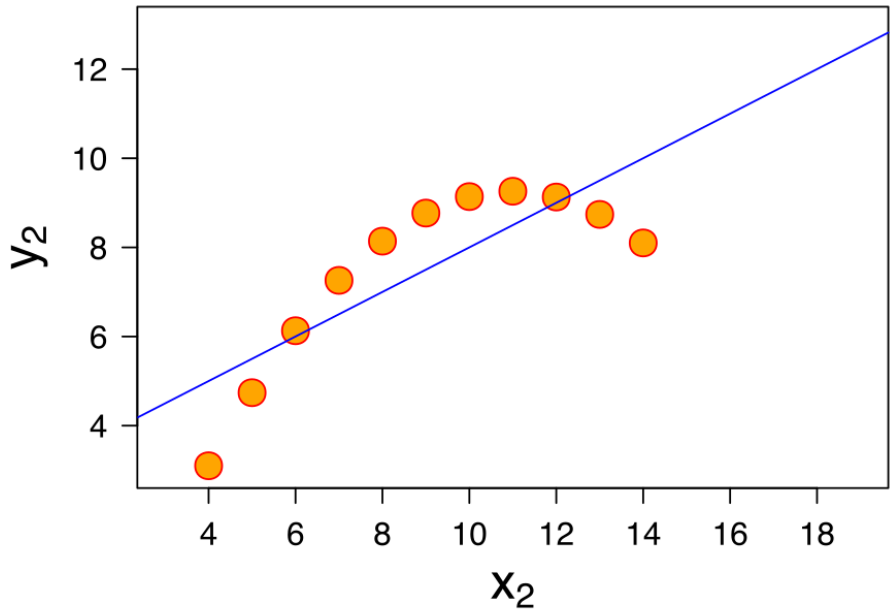

I DON'T ALWAYS MAKE A LINEAR REGRESSION MODEL
BUT WHEN I DO, I CHECK IT WITH A SCATTER PLOT & RESIDUAL PLOT
makeameme.org

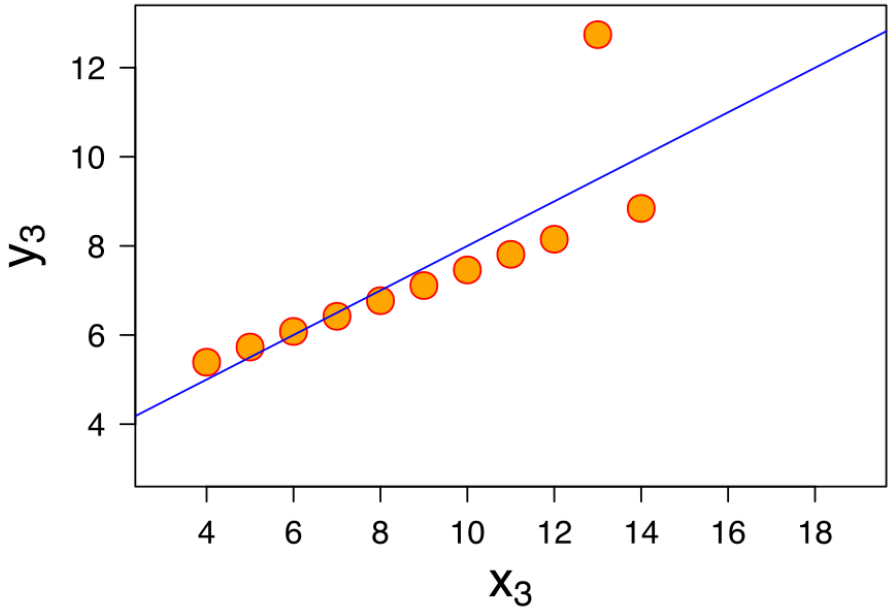
Assumptions of Linear Regression
(in practice these are rarely met "perfectly")

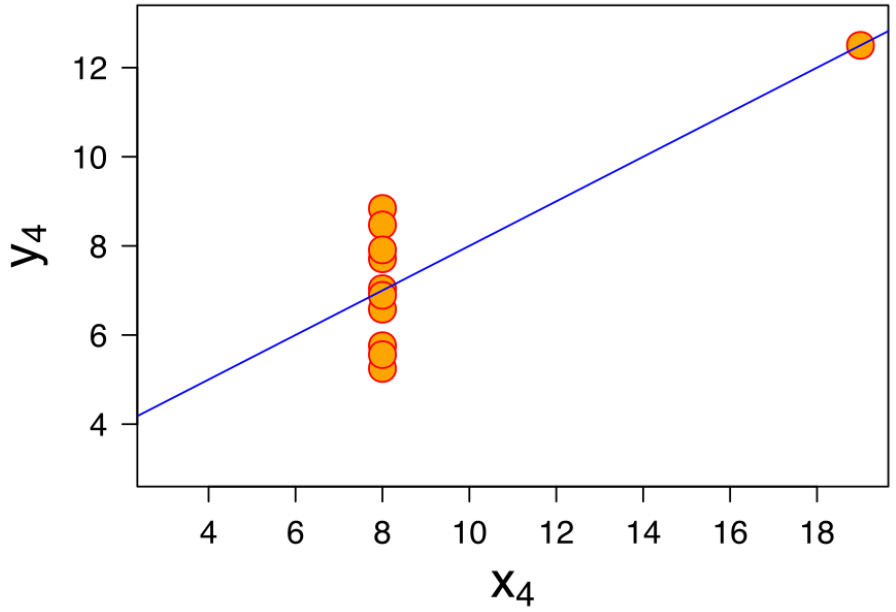
- **Weak exogeneity**: related to experimental design. How much error is there in the measurement? Linear regression assumes that there isn't much.
- **Linearity:** Assumes the relationship is really linear.
- **Constant variance**: variance in errors doesn't depend on predictor.
- **Lack of perfect multicollinearity:** predictors aren't highly correlated (will come back to this when we look at multiple linear regression)

Linearity


12
10
8
6
4
y_1
4 6 8 10 12 14 16 18
x_1

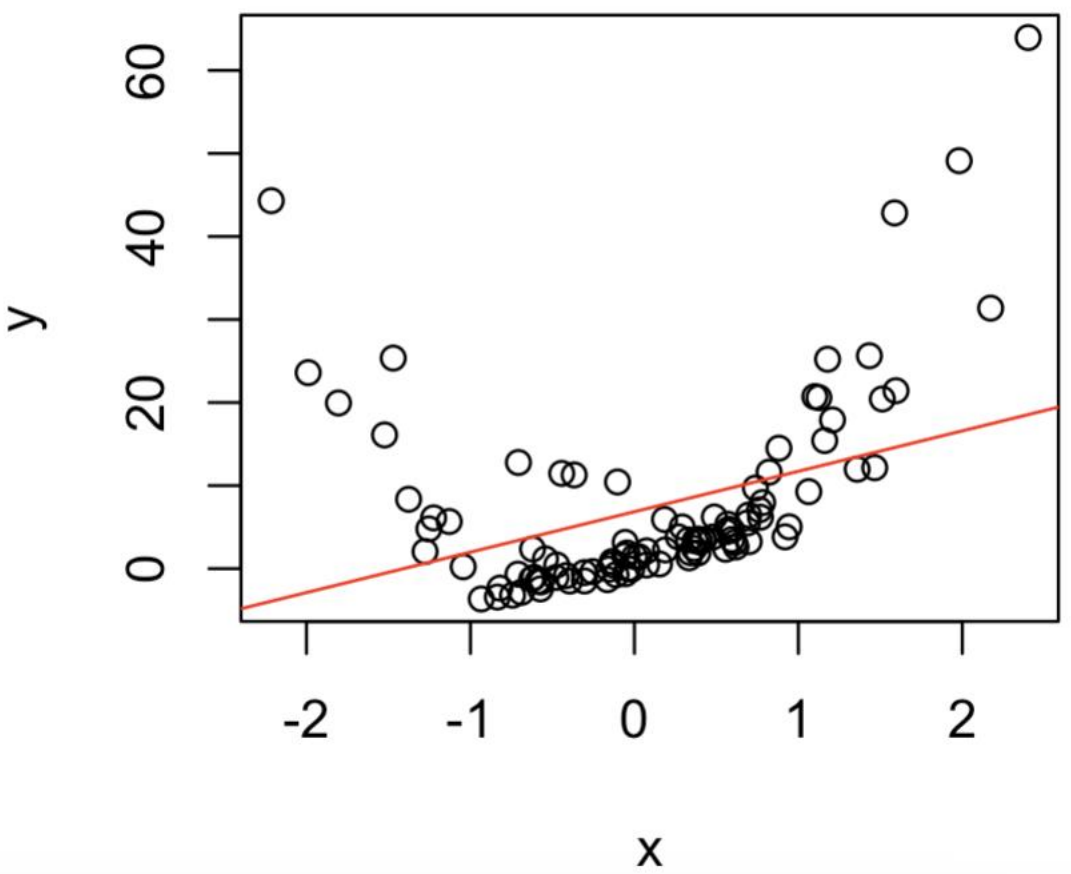

12
10
8
6
4
y_2
4 6 8 10 12 14 16 18
x_2


12
10
8
6
4
y_3
4 6 8 10 12 14 16 18
x_3

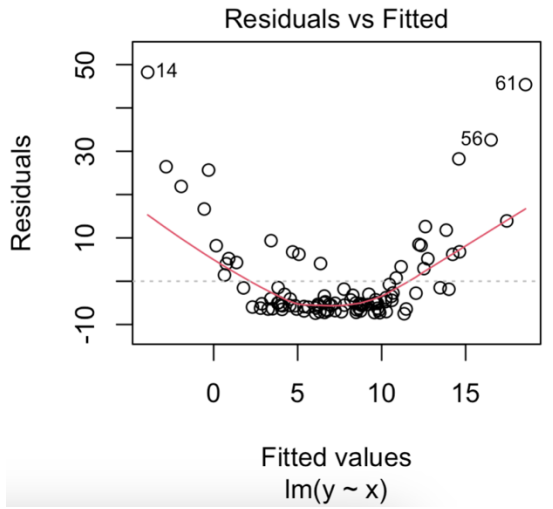

12
10
8
6
4
y_4
4 6 8 10 12 14 16 18
x_4

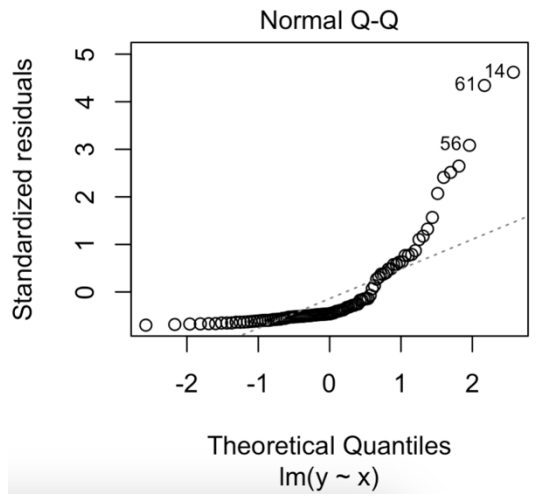
Violating assumptions (linearity)

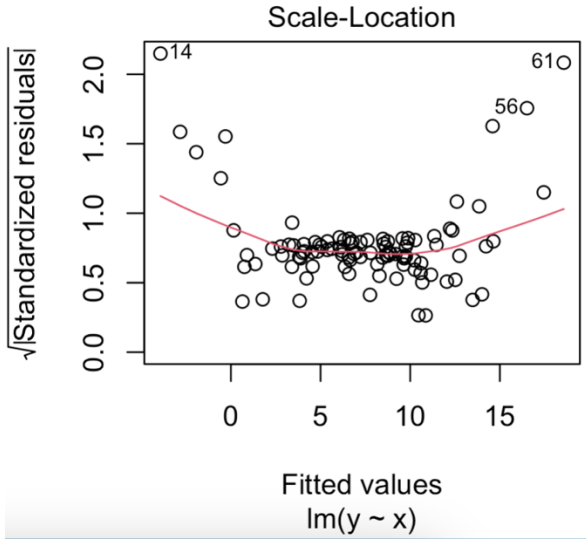
```
plot(x,y)
model<-lm(y~x)
abline(model, col="red")
```

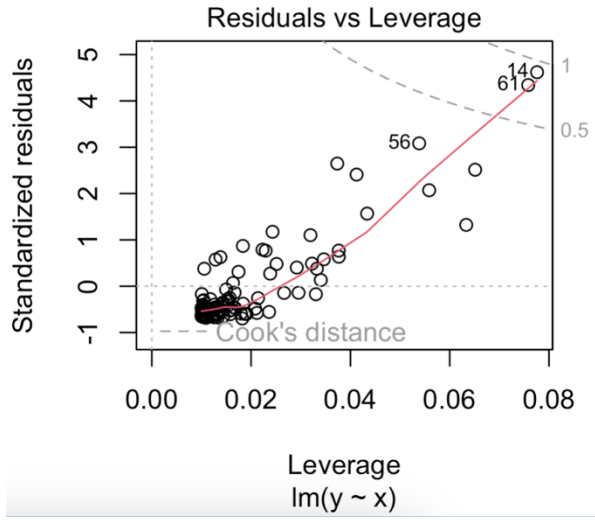



```
> plot(model)
Hit <Return> to see next plot:
```



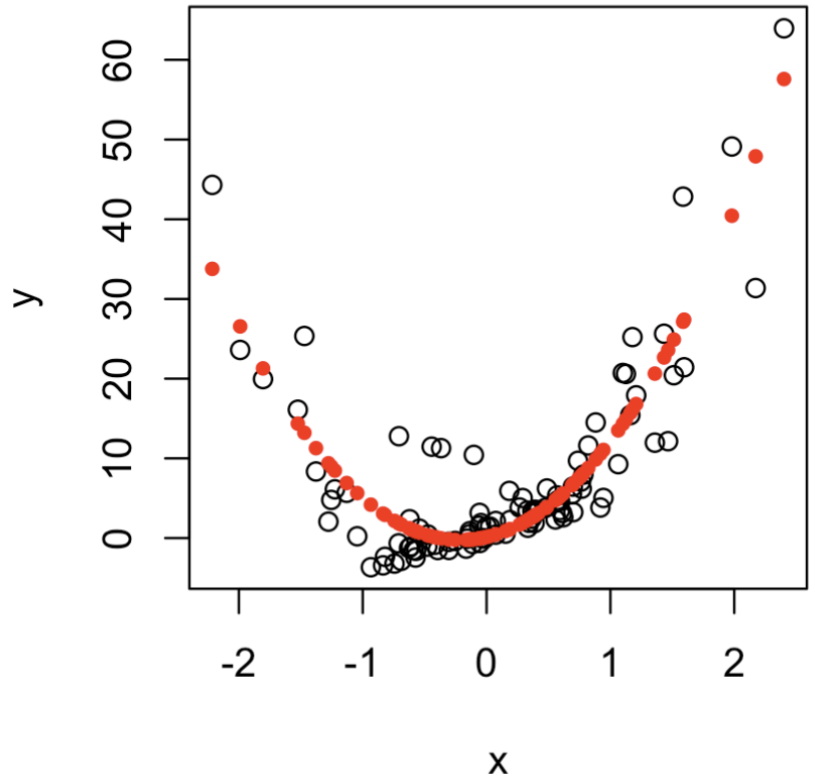






Non-linear model (will go into much more detail on non-linearity later this quarter)

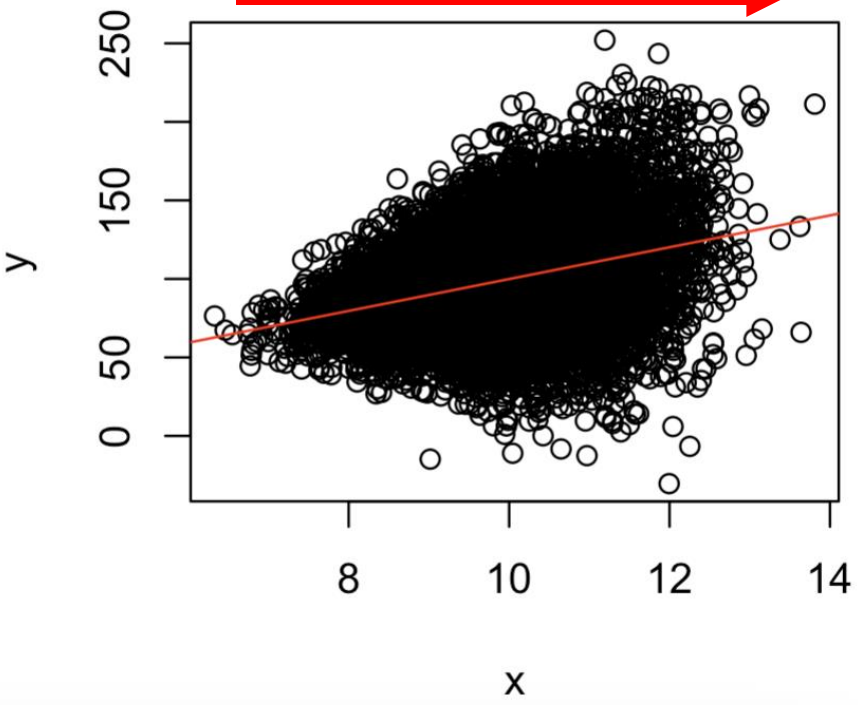
```r
# adding a non-linear transformation to the model
model2<-lm(y~x+I(x^2))
plot(x,y)
points(data.frame(x=x, y=predict(model2)), col="red", pch=20)
```



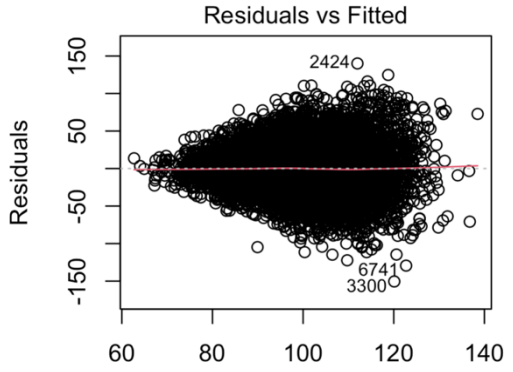
Violating assumptions (constant variance)

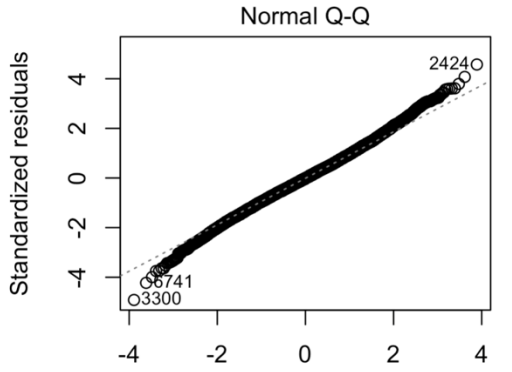
```
plot(x,y)
model<-lm(y~x)
abline(model, col="red")
```

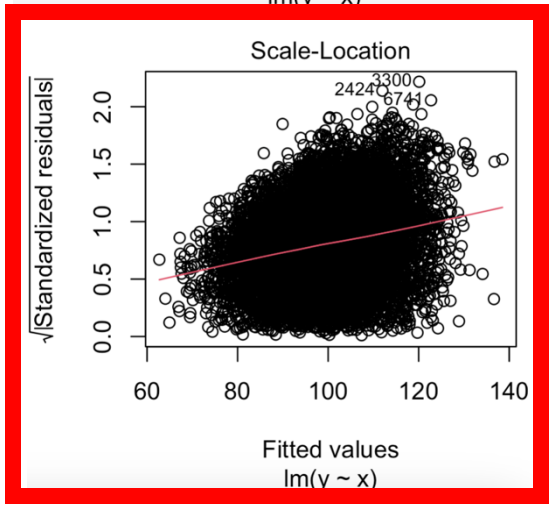
Low variance → High variance

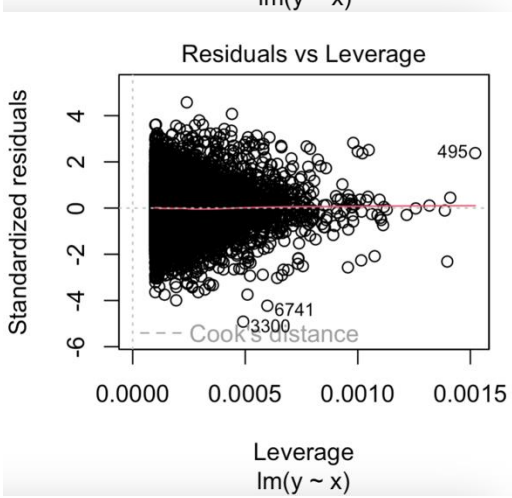



```
> plot(model)
Hit <Return> to see next plot:
```





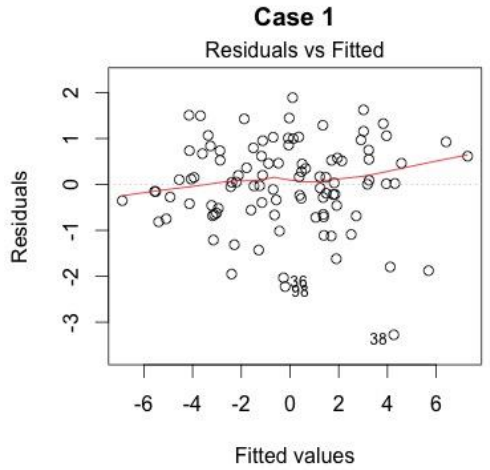


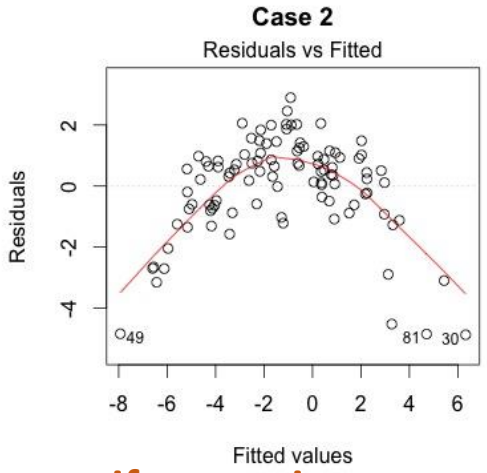


Correlation between fitted values and residuals

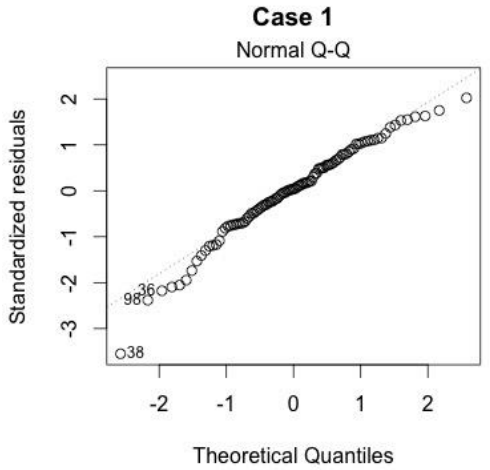
Checking model assumptions

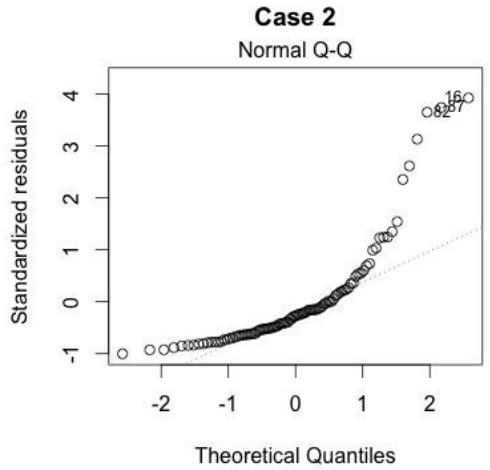
Non linear relationship – consider adding non-linear predictor



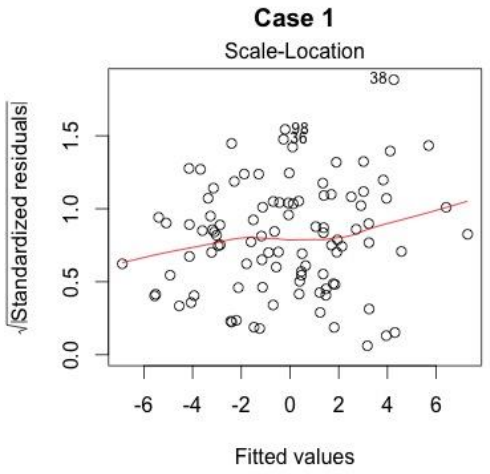


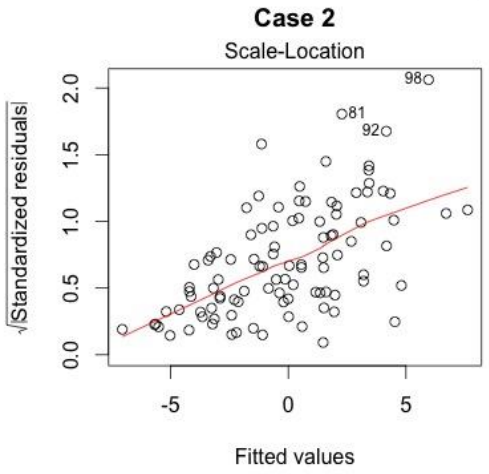
Non-normal residuals – consider transformation such as log()



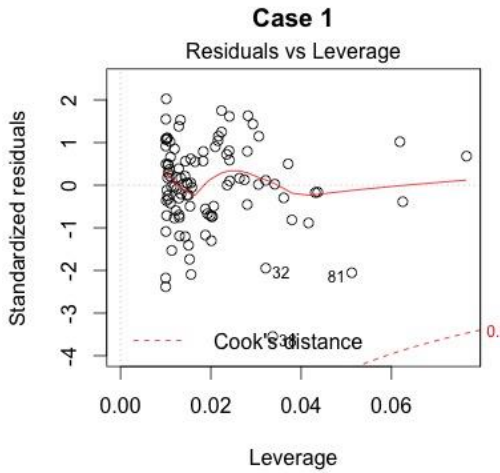


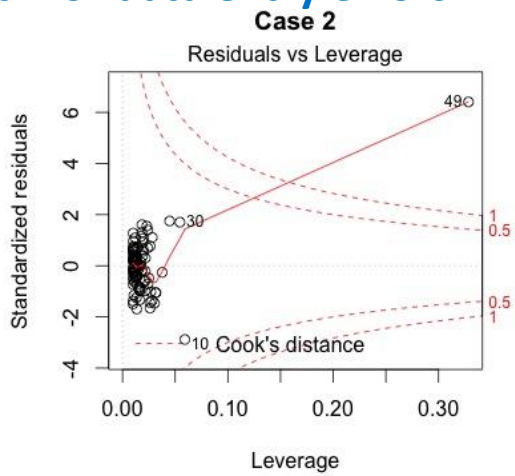
Non-uniform variance - consider transformation such as log()





Outliers – consider transformation and check for data entry errors





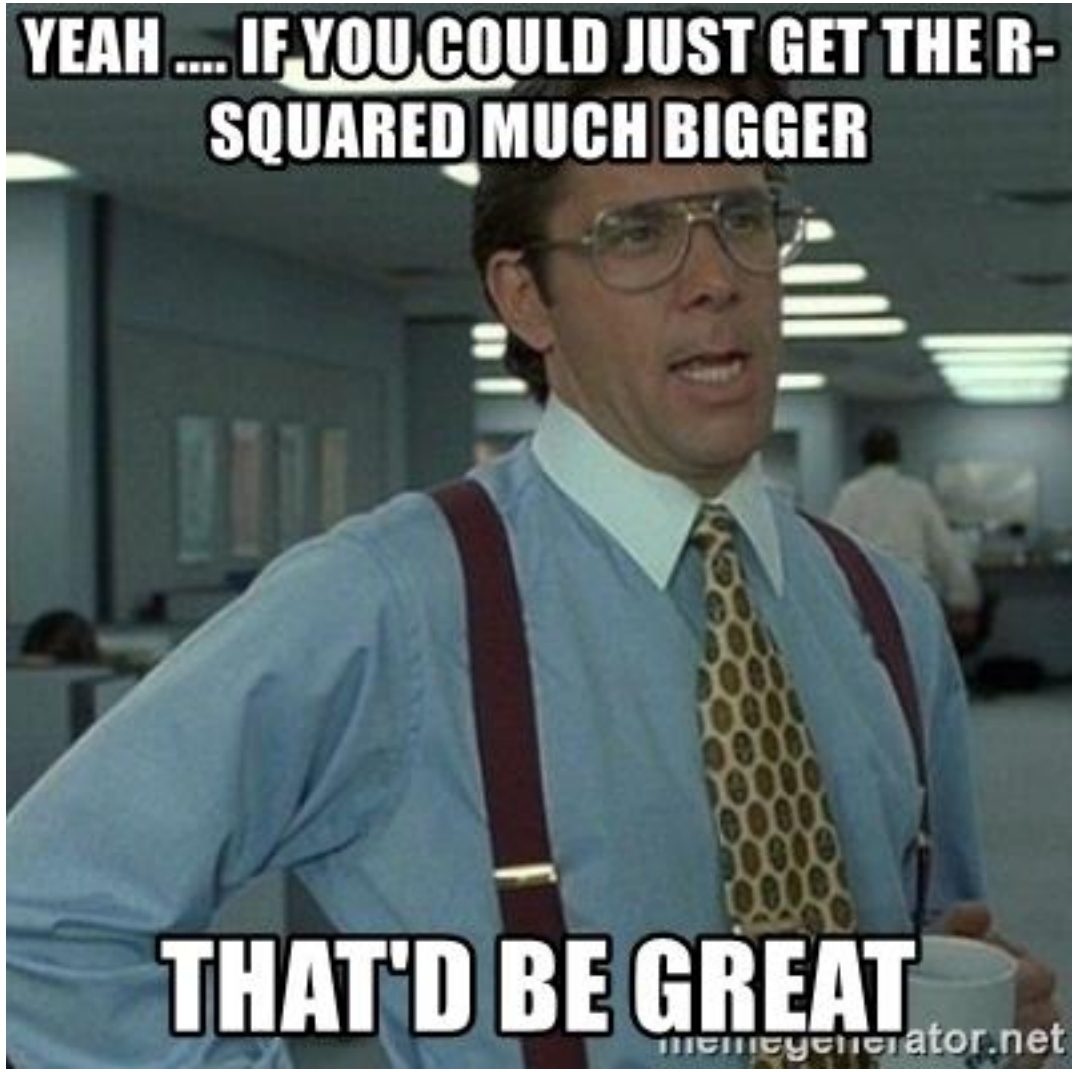
PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

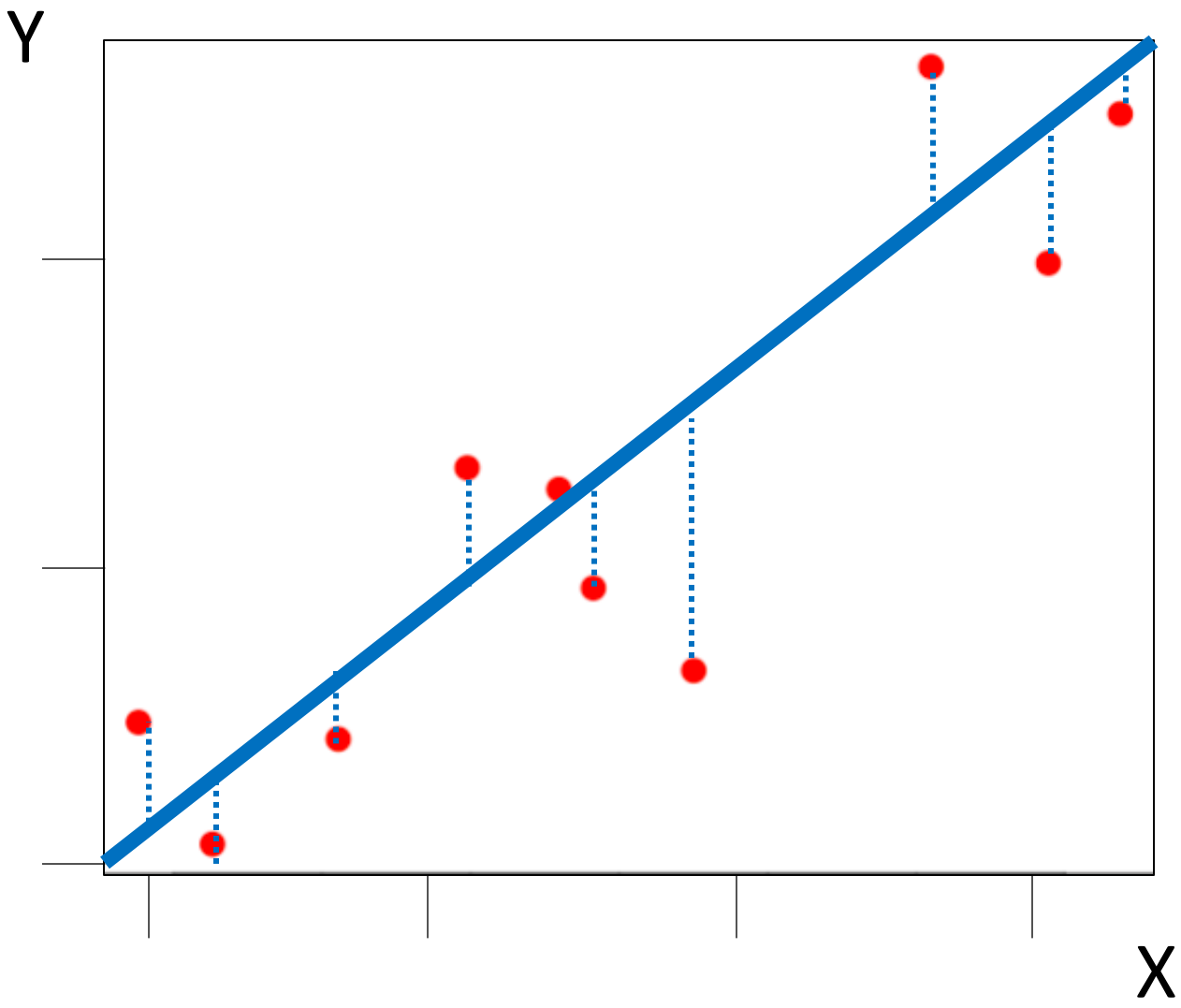
Linear Regression

- R^2
- Hypothesis testing

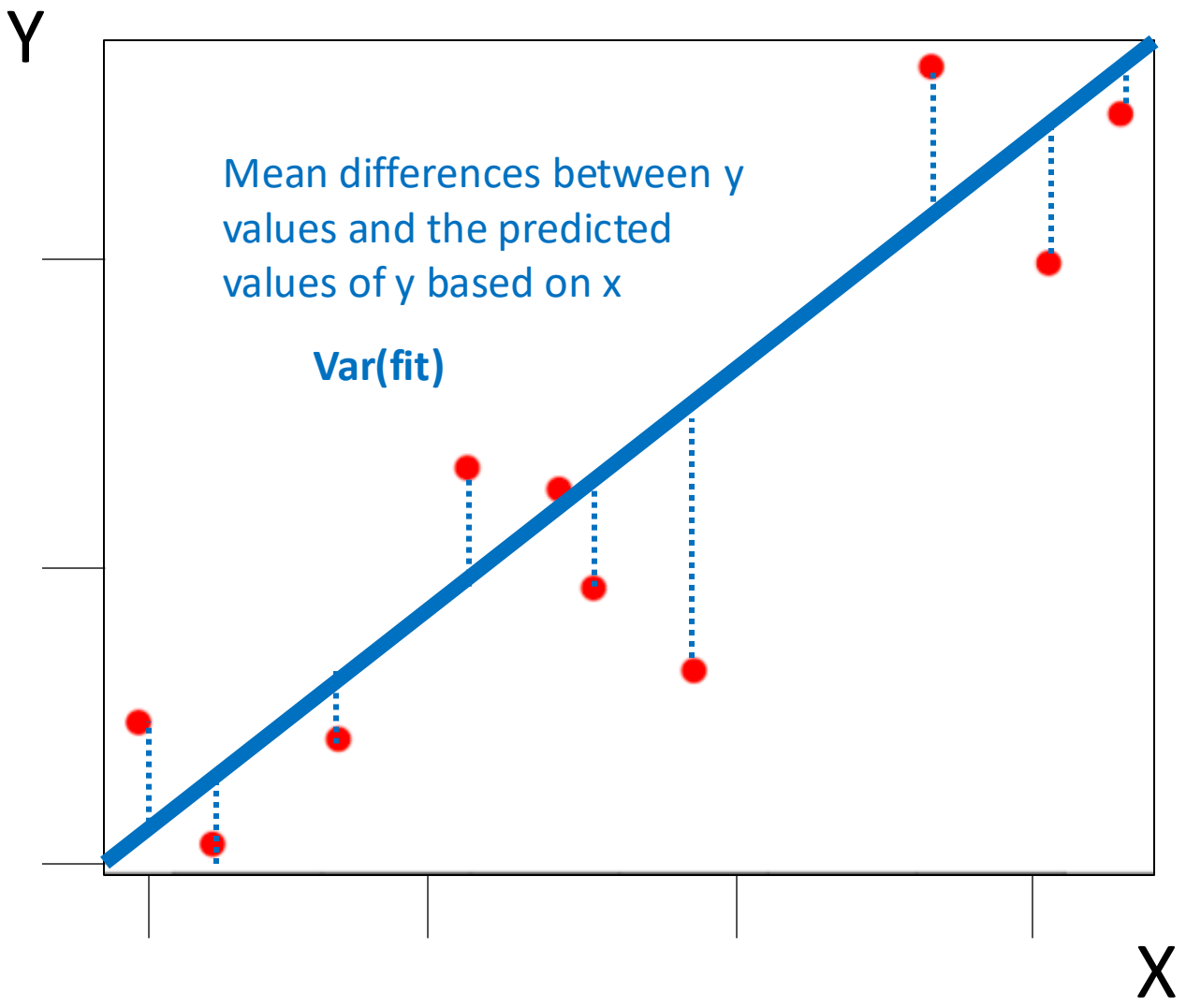
What is R^2?

Proportion of variance in Y (response variable) explained by variance in X (predictor variable(s))



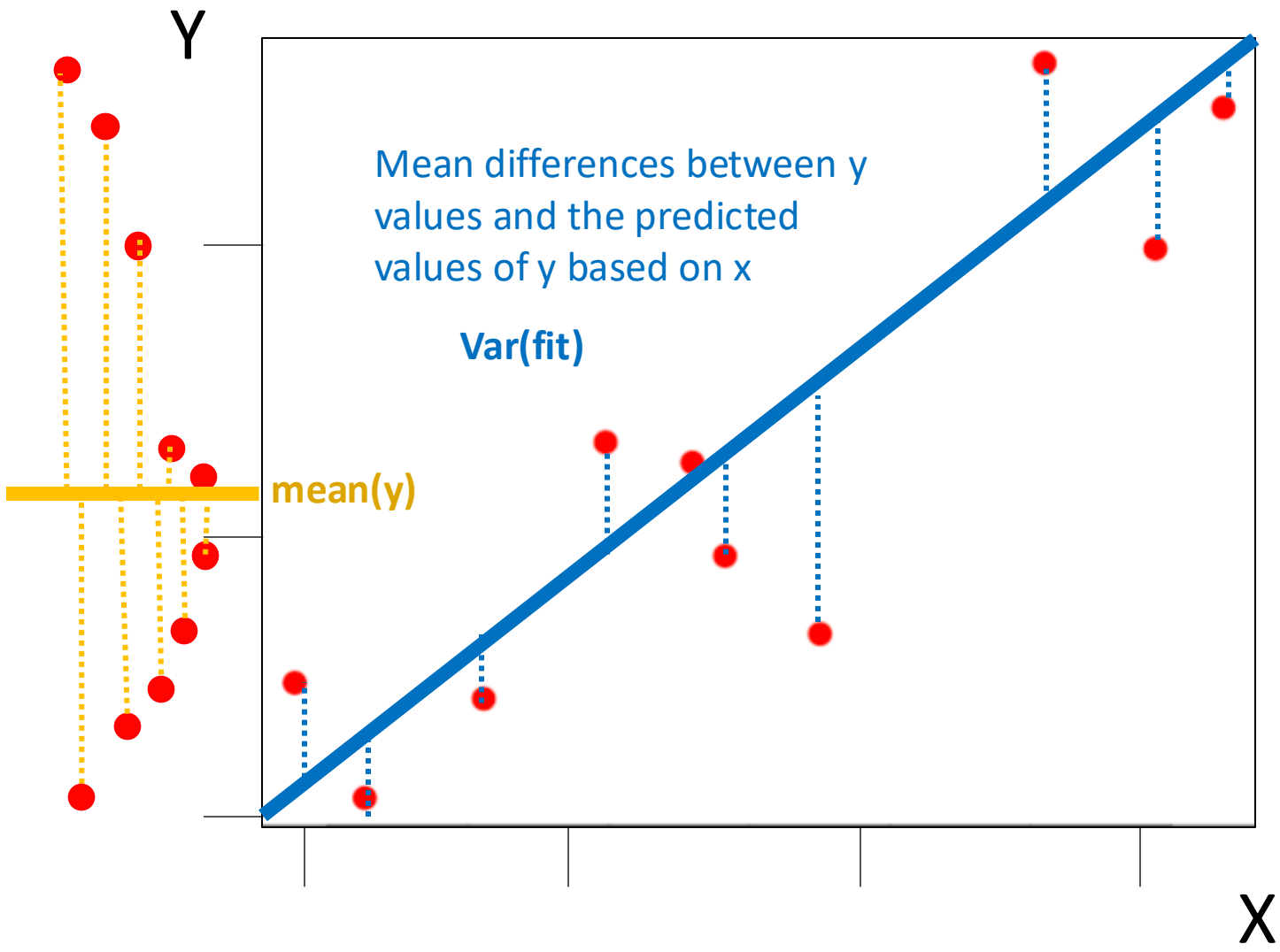

Y
X

Var = Sum of Squares / (n-1)



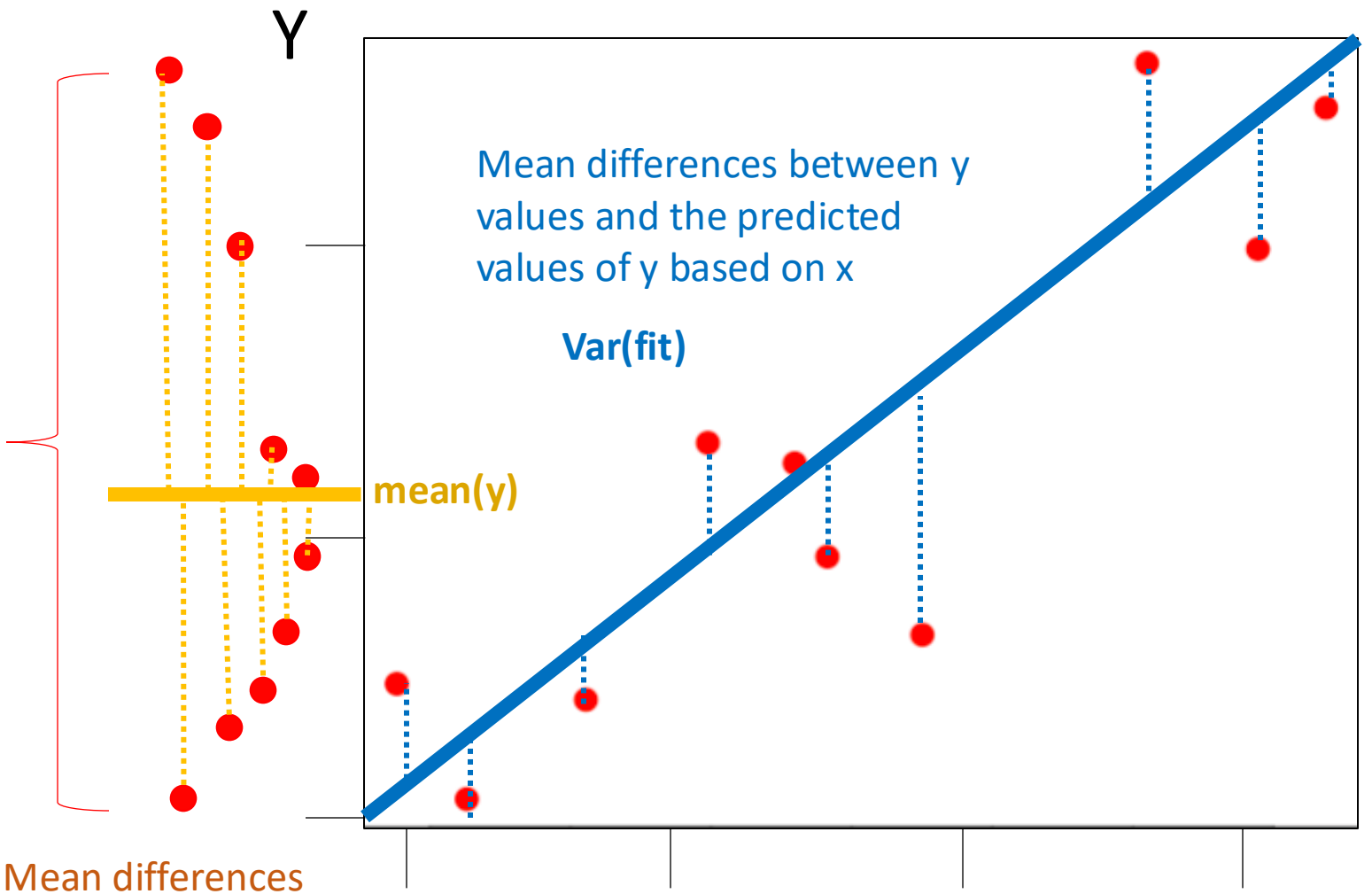
*__Fitted__ values is a synonym for __predicted__ values of y based on x

Var = Sum of Squares / (n-1)



*__Fitted__ values is a synonym for __predicted__ values of y based on x

Var = Sum of Squares / (n-1)

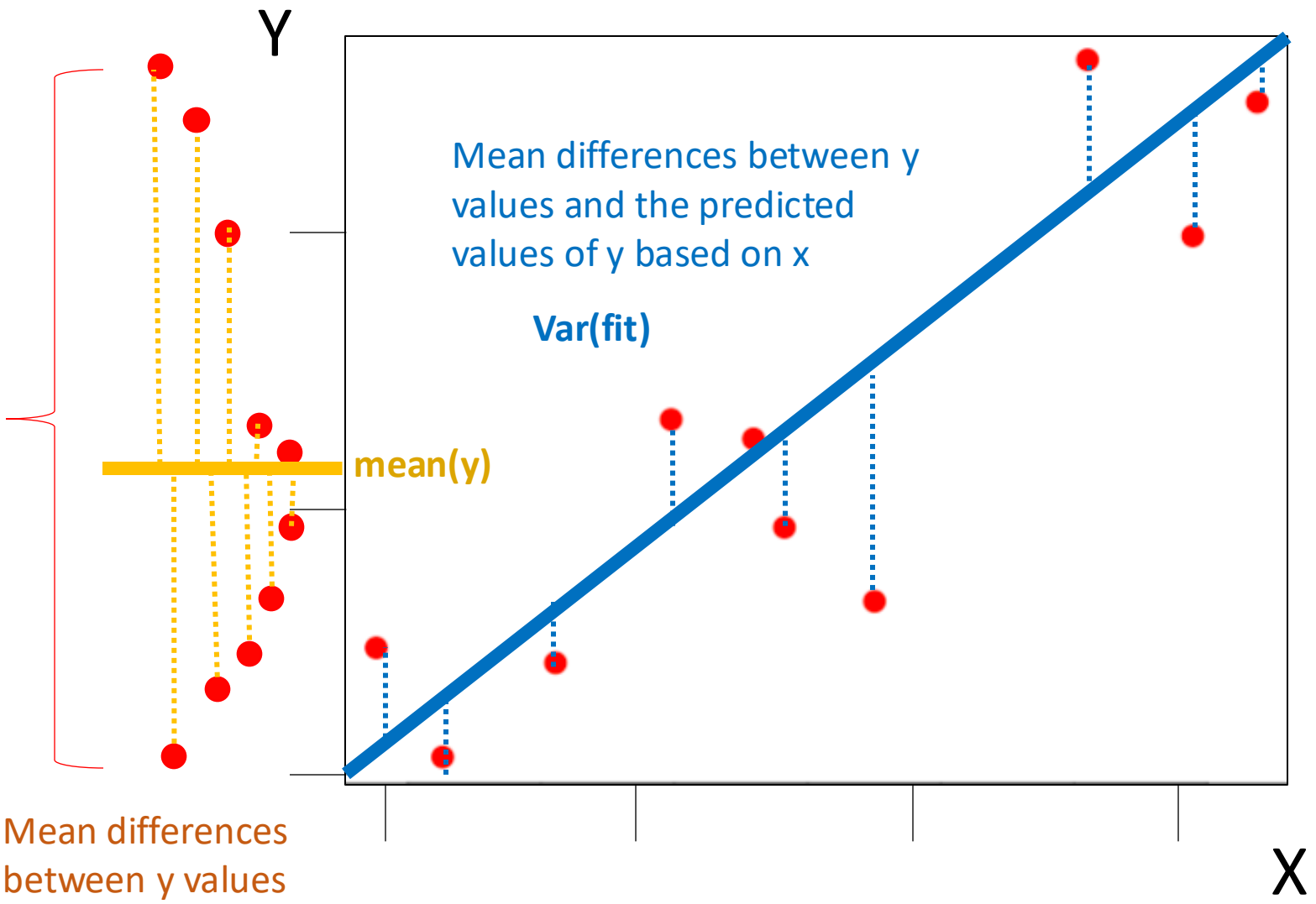


Mean differences between y values and the mean of y
Var(mean)

*__Fitted__ values is a synonym for __predicted__ values of y based on x

Var = Sum of Squares / (n-1)

$$R^2 = \frac{\text{var(mean)} - \text{var(fit)}}{\text{var(mean)}}$$



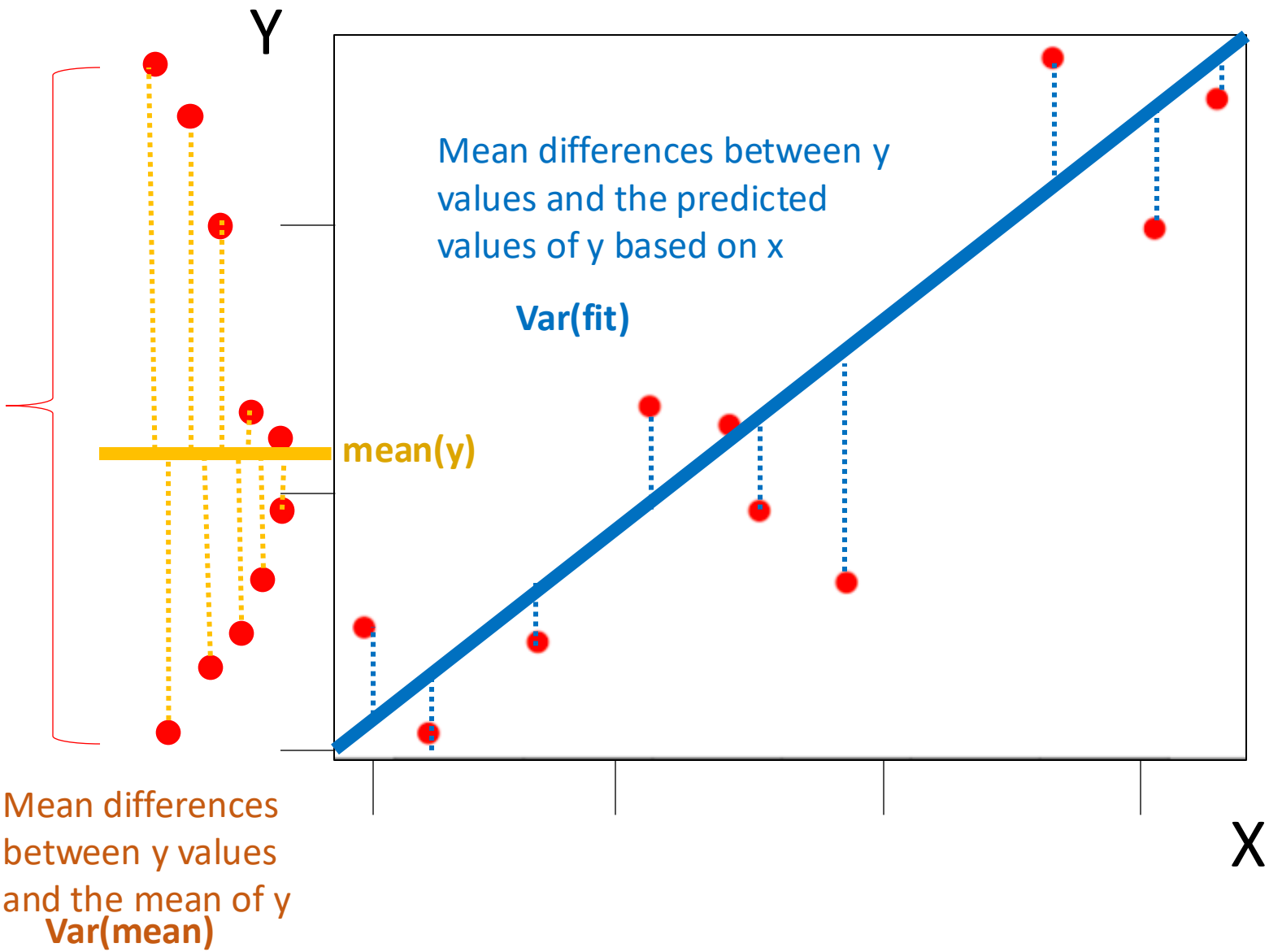
*__Fitted__ values is a synonym for __predicted__ values of y based on x

Var = Sum of Squares / (n-1)

R^2 = "proportion of variance in response variable explained by predictor variable(s)"

$$R^2 = \frac{\text{var(mean)} - \text{var(fit)}}{\text{var(mean)}}$$

If y is perfectly on the fitted line var(fit) = 0 and $\mathbf{R^2}$ =1



*__Fitted__ values is a synonym for __predicted__ values of y based on x

Hypothesis testing

- Null hypothesis: predictor(s) explain no more variance in y than expected by chance
- Alternative hypothesis: predictor(s) explain more variance in y than expected by chance
- F-test
- P-value depends approximately on R^2, sample size, and number of predictors (in multiple linear regression)

$$F = \frac{R^2 / k}{(1 - R^2) / (N - k - 1)}$$

where…

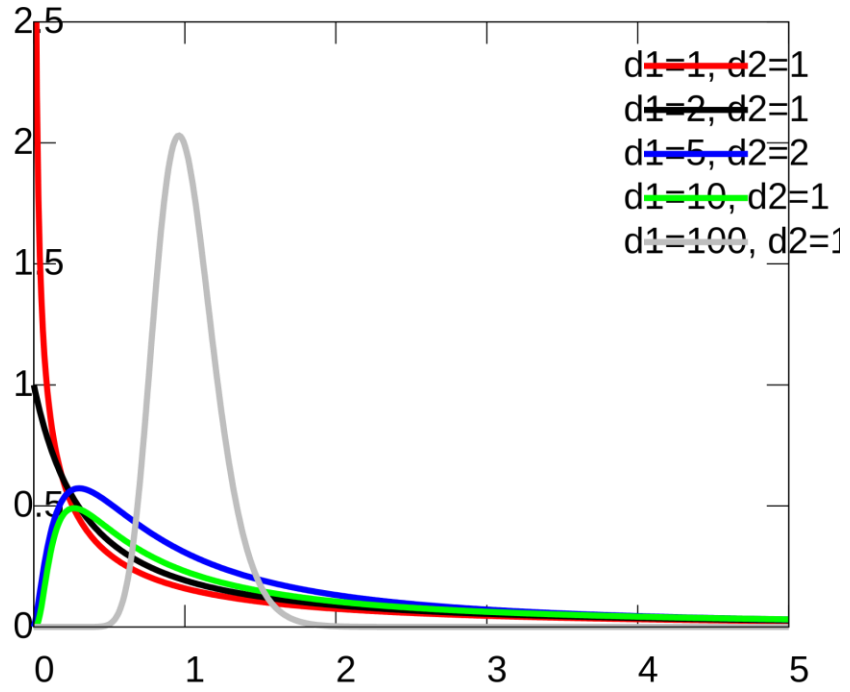
***R²** = proportion of variance explained*

***k** = number of predictors*

***N** = number of observations*

$$df1 = k$$

$$df2 = N\text{-}k\text{-}1$$



F-table of Critical Values of α = 0.05 for F(df1, df2)

	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79

$$F = \frac{R^2 / k}{(1 - R^2) / (N - k - 1)}$$

where…

***R²** = proportion of variance explained*

***k** = number of predictors*

***N** = number of observations*

$$df1 = k$$

$$df2 = N\text{-}k\text{-}1$$

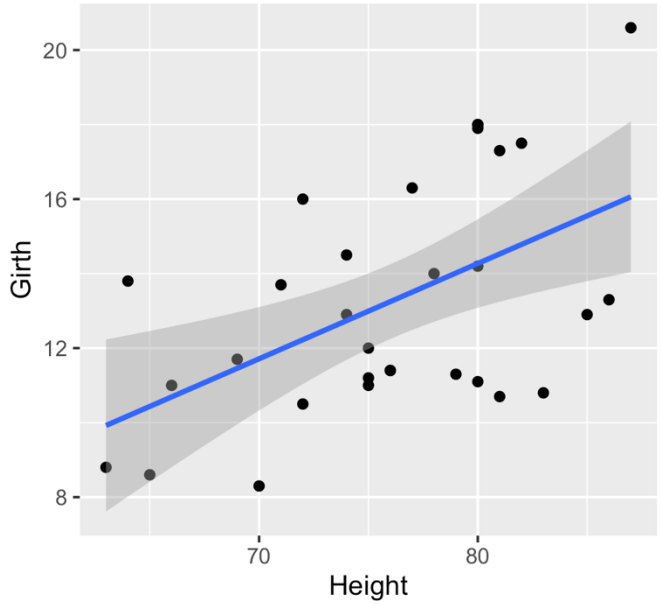
```
> model_summary

Call:                          k=1          N=nrow(trees)=31
lm(formula = Girth ~ Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2386 -1.9205 -0.0714  2.7450  4.5384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.18839    5.96020  -1.038  0.30772
Height       0.25575    0.07816   3.272  0.00276 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.728 on 29 degrees of freedom
Multiple R-squared:  0.2697,     Adjusted R-squared:  0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

```
> pf(f, df1, df2, lower.tail = F)
[1] 0.002757815
```

Linear regression T statistic (testing each β)

$$T = \frac{\beta_j}{SE(\beta_j)}$$

"Does including this predictor (compared to if the β = 0) explain more variance in response than expected by chance?"



Linear regression T statistic (testing each β)

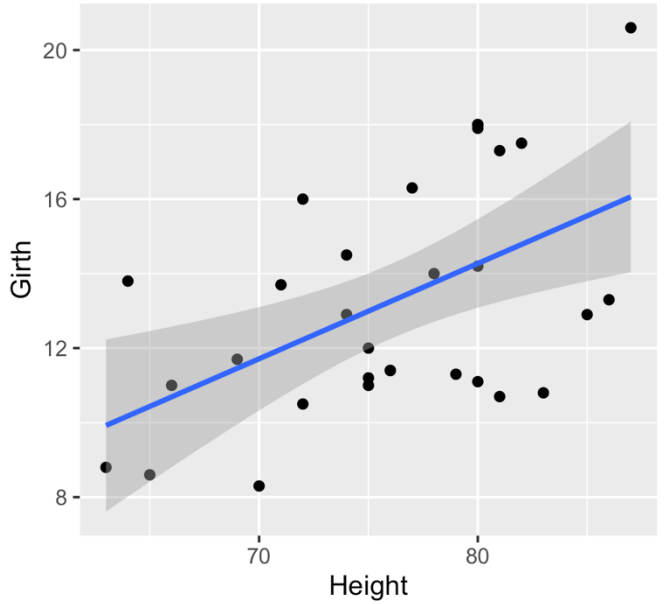
$$T = \frac{\beta_j}{SE(\beta_j)}$$

"Does including this predictor (compared to if the β = 0) explain more variance in response than expected by chance?"

```
> model_summary

Call:
lm(formula = Girth ~ Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2386 -1.9205 -0.0714  2.7450  4.5384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.18839    5.96020  -1.038  0.30772
Height       0.25575    0.07816   3.272  0.00276 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.728 on 29 degrees of freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:  0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

Visualizing SE in R

```
ggplot(trees, aes(x=Height, y=Girth))+
  geom_point()+
  geom_smooth(method="lm")
```



Intercept….

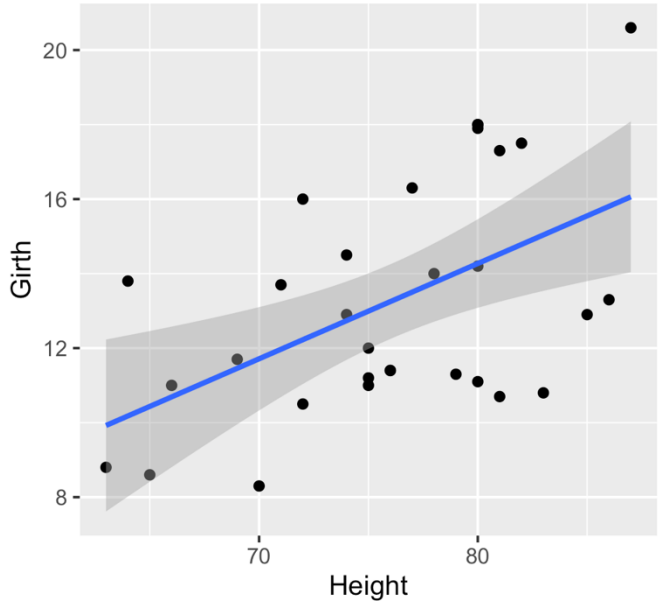
$$T = \frac{\beta_j}{SE(\beta_j)}$$

Intercept: significance is simply regarding whether it is different from 0, often not particularly important

```
> model_summary

Call:
lm(formula = Girth ~ Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2386 -1.9205 -0.0714  2.7450  4.5384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.18839    5.96020  -1.038  0.30772
Height       0.25575    0.07816   3.272  0.00276 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.728 on 29 degrees of freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:  0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

Visualizing SE in R

```
ggplot(trees, aes(x=Height, y=Girth))+
  geom_point()+
  geom_smooth(method="lm")
```

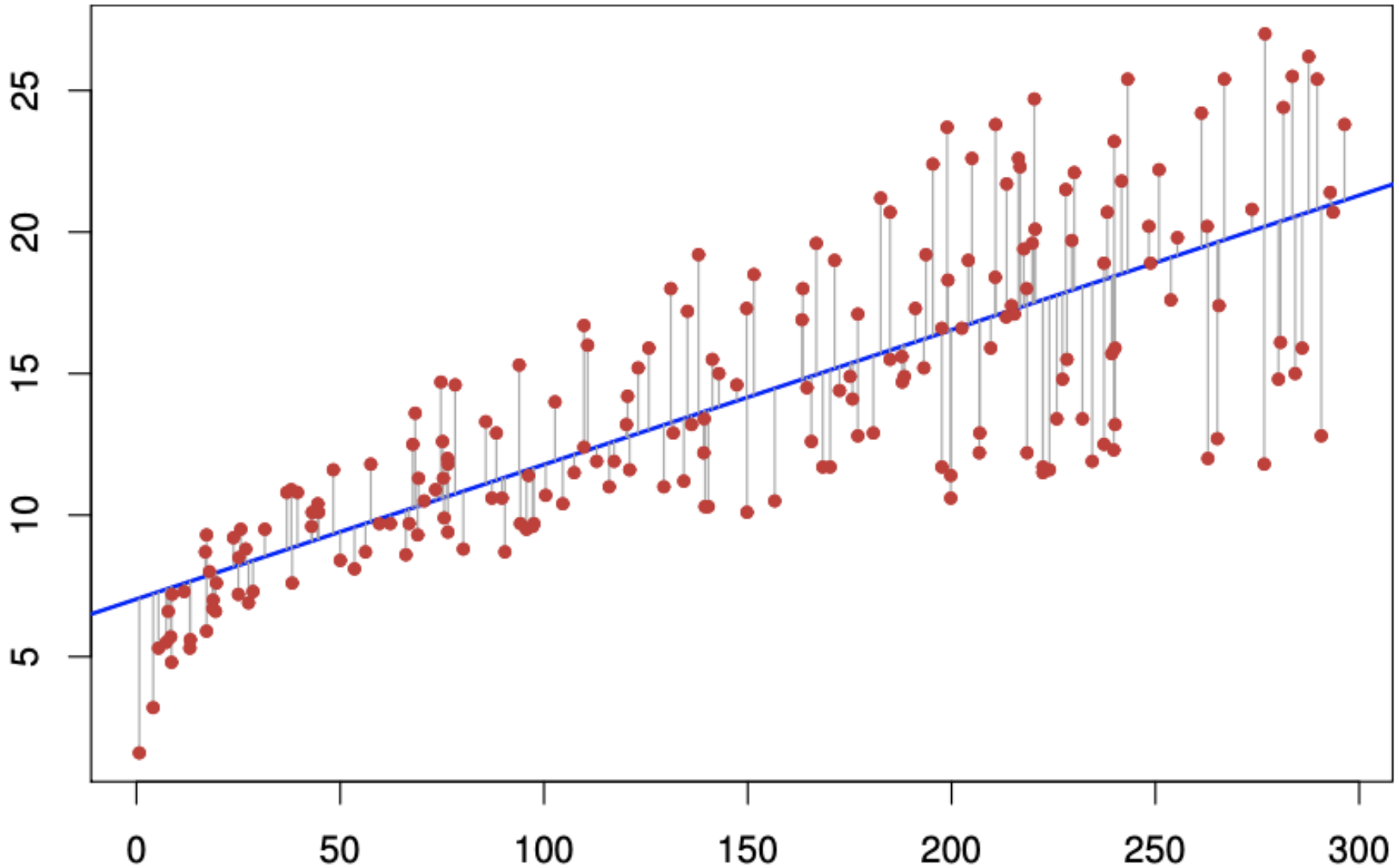


PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

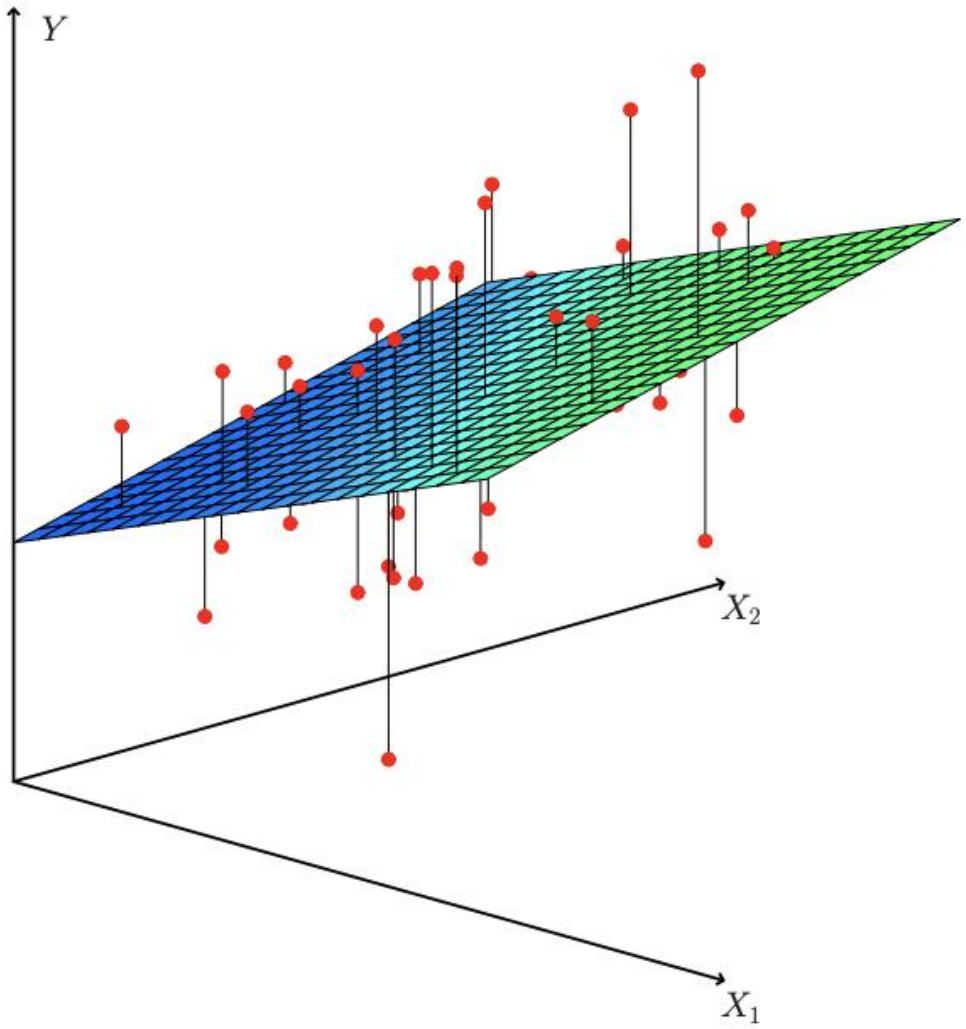
Multiple Linear Regression

- In R
- Hypothesis testing
- Assumptions (detecting multicollinearity)

Single Linear Regression



Multiple Linear Regression



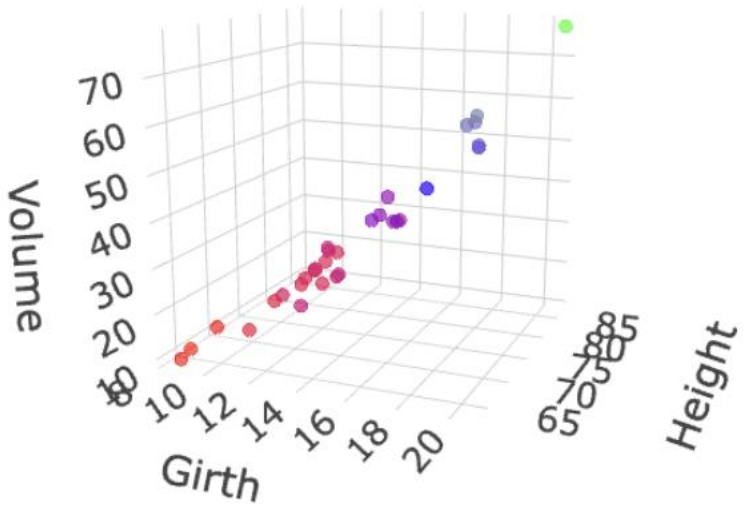
Instead of fitted line we have fitted (p-dimensional) plane

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

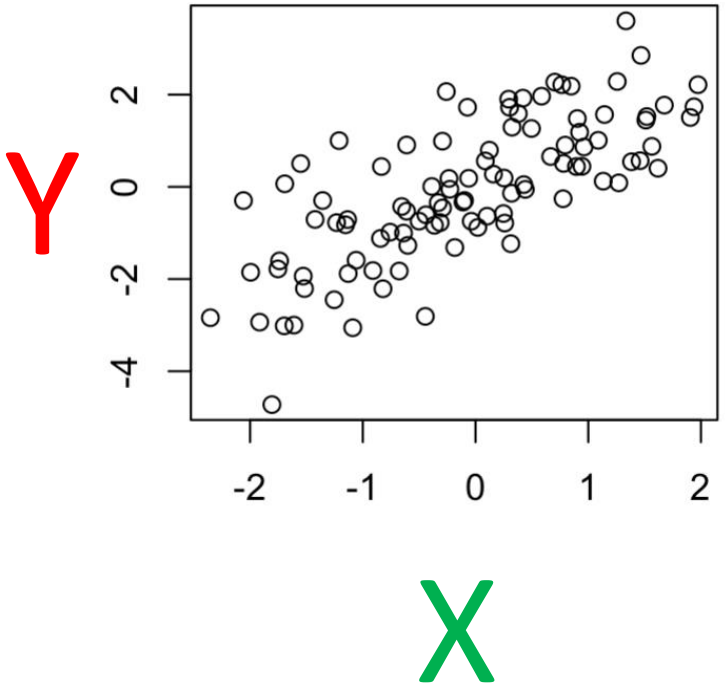
Same rules apply: multiple linear regression finds the slope and intercepts that minimize the RSS

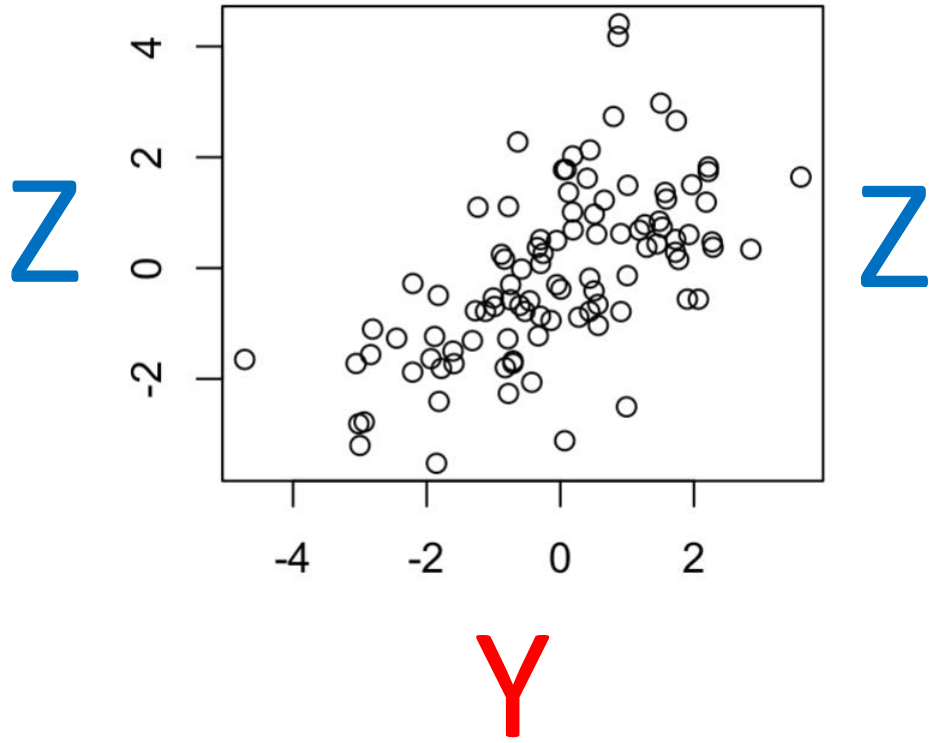
Multiple Linear Regression in R

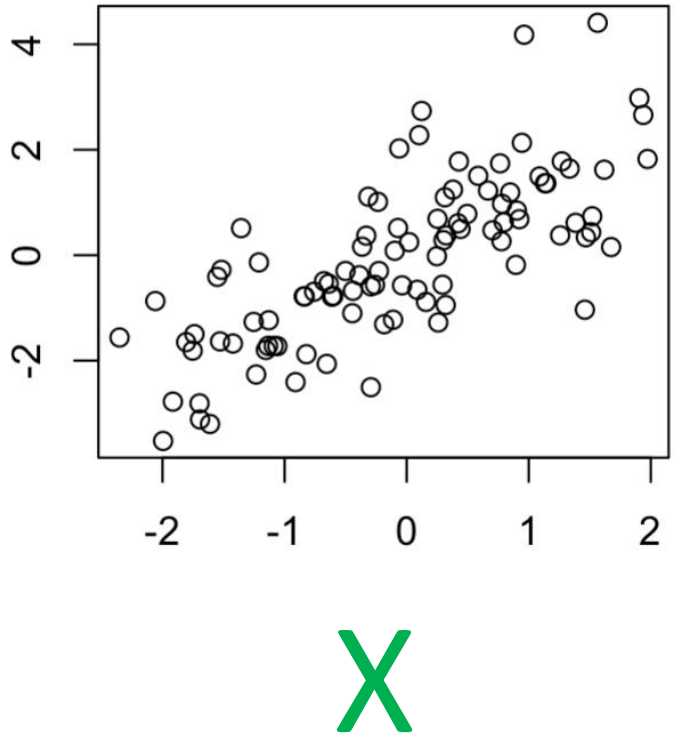
```
model<-lm(Volume~Girth+Height, trees)
```

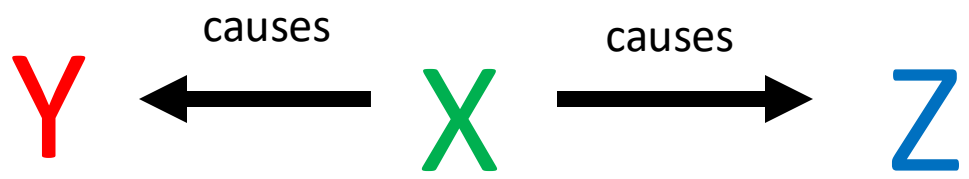


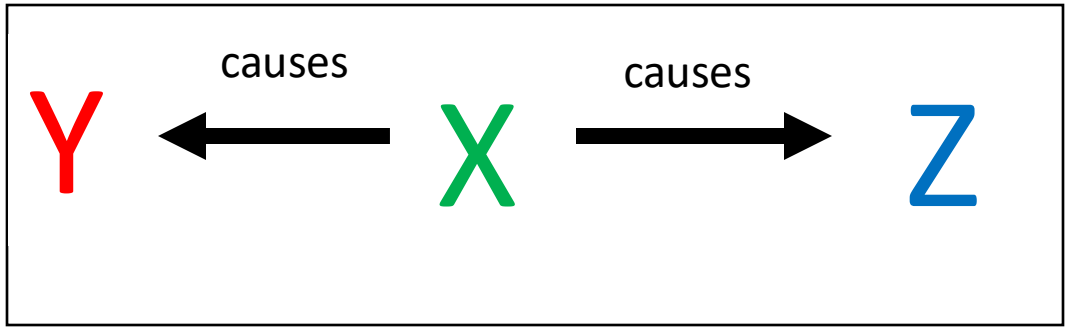

```
> summary(model)

Call:
lm(formula = Volume ~ Girth + Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,     Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

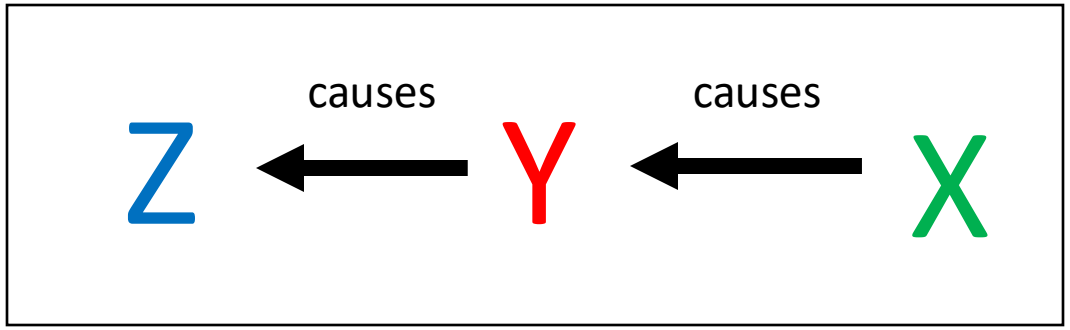

Y
2
0
-2
-4
-2
-1
0
1
2
X

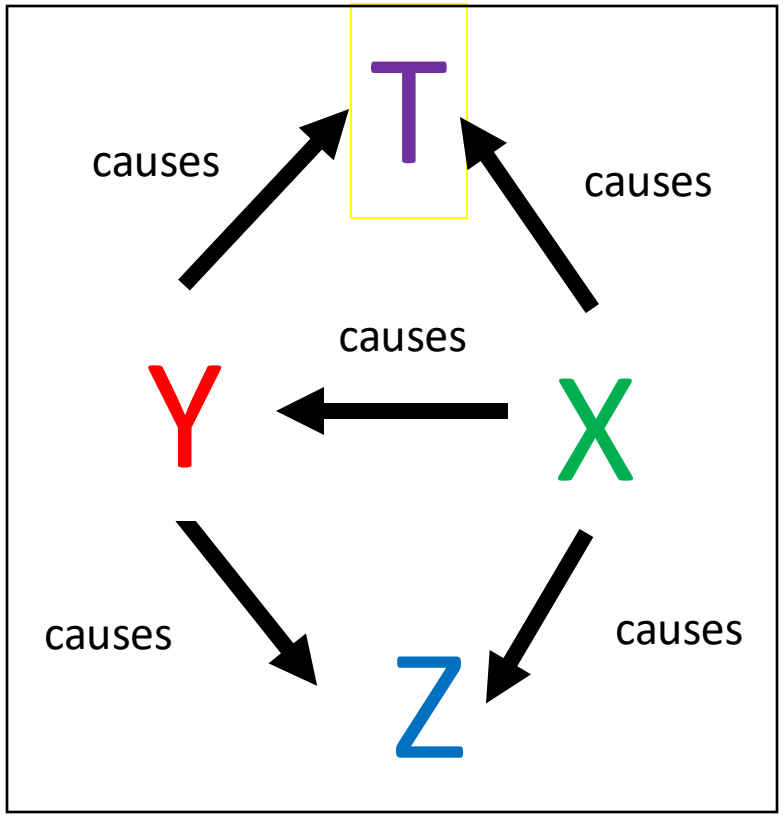

Z
4
2
0
-2
-4
-2
0
2
Y


Z
4
2
0
-2
-2
-1
0
1
2
X

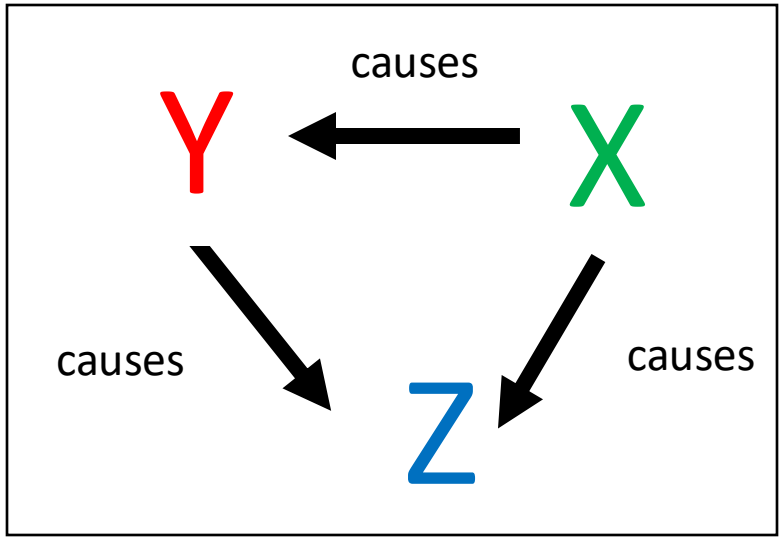

causes
causes
Y
X
Z

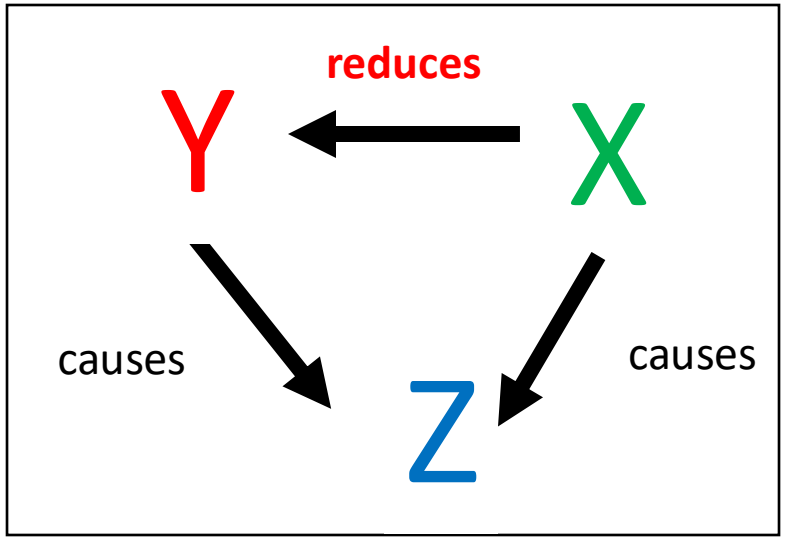

causes
causes
Y
X
Z


causes
causes
Z
Y
X


T
causes
causes
causes
Y
X
causes
causes
Z

?


causes
Y
X
causes
causes
Z


reduces
Y
X
causes
causes
Z


$V = \frac{1}{3}\pi r^2 . h$
h
r
sin
cos
tan
$y = ax^2 + bx + c$
$(x_1, x_2) = \frac{-b \pm \Delta}{2a}$

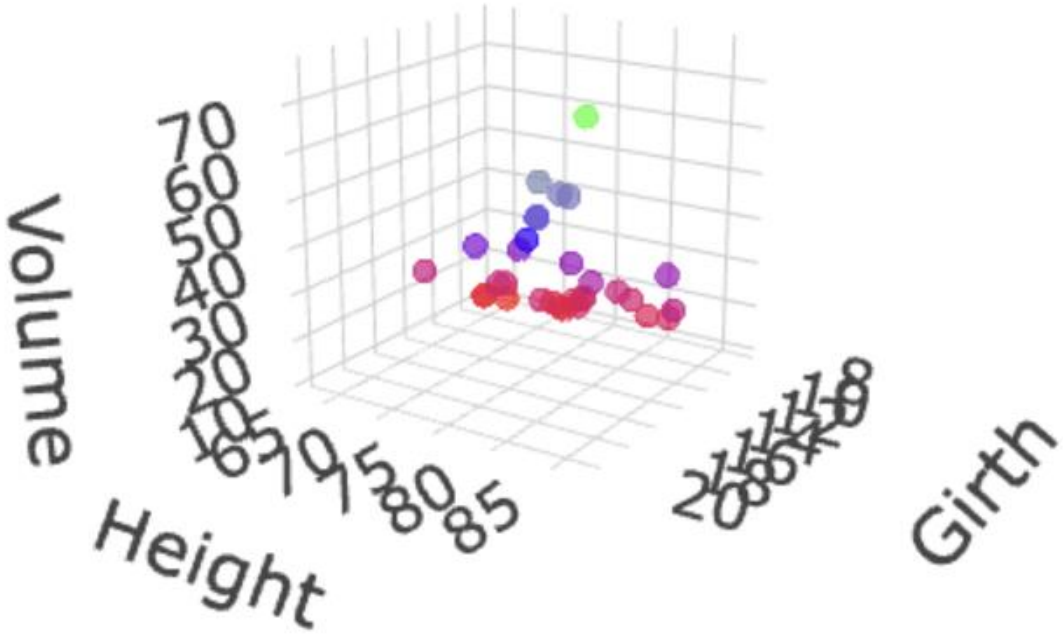
Assumptions of Linear Regression
(practically never met perfectly)

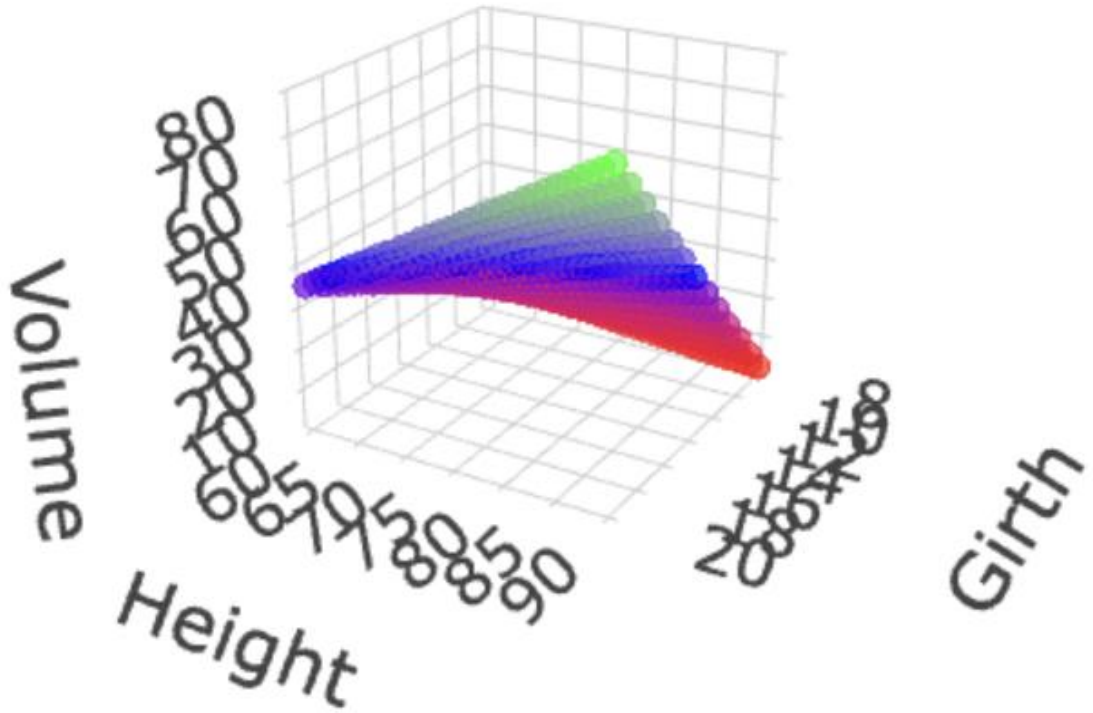
- **Weak exogeneity**: related to experimental design. How much error is there in the measurement? Linear regression assumes that there isn't much.
- **Linearity:** Assumes the relationship is really linear.
- **Constant variance**: variance in errors doesn't depend on predictor.
- **Lack of multicollinearity:** predictors aren't highly correlated

How to check if there is excess multicollinearity?

- **Insignificant regression coefficients in the multiple regression but significant F test for whole model.**
- **Insignificant coefficient of a particular predictor in multiple linear regression but single linear regression (correlation) is significant**
- **Variance Inflation Factor (VIF) > 10 (* somewhat “arbitrary” threshold)**

* Bonus * making 3d plots in R





$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Flowering	atitude	Longitude	bio1	bio2
118	47.41667	19.33333	10.6	8.9
118	49.00000	11.50000	8.0	8.7
121	49.00000	11.50000	8.0	8.7
123	51.91667	11.50000	9.0	8.1
104	34.35000	47.23333	12.9	15.8
102	33.23333	48.56667	15.3	15.7
105	33.23333	48.56667	15.3	15.7
113	39.25000	45.50000	12.0	11.8
98	40.93333	27.28333	13.0	9.7
117	38.16667	38.31667	9.6	11.6
116	37.75000	39.71667	14.4	11.4
113	39.25000	45.50000	12.0	11.8
115	45.00000	34.00000	10.9	9.5
109	39.25000	45.50000	12.0	11.8

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Flowering	atitude	Longitude	bio1	bio2
118	47.41667	19.33333	10.6	8.9
118	49.00000	11.50000	8.0	8.7
121	49.00000	11.50000	8.0	8.7
123	51.91667	11.50000	9.0	8.1
104	34.35000	47.23333	12.9	15.8
102	33.23333	48.56667	15.3	15.7
105	33.23333	48.56667	15.3	15.7
113	39.25000	45.50000	12.0	11.8
98	40.93333	27.28333	13.0	9.7
117	38.16667	38.31667	9.6	11.6
116	37.75000	39.71667	14.4	11.4
113	39.25000	45.50000	12.0	11.8
115	45.00000	34.00000	10.9	9.5
109	39.25000	45.50000	12.0	11.8

>summary(model)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 142.36492   20.07386   7.092 4.78e-09 ***
Latitude      0.09498    0.34348   0.277    0.783
Longitude     0.16722    0.16999   0.984    0.330
bio1         -2.82833    0.38874  -7.276 2.48e-09 ***
bio2         -0.57061    0.58483  -0.976    0.334
```

```
> model$residuals
          1
 2.95716100  -3.35
          9
-10.51200682  -1.62
         17
 7.27865394 -13.65
         25
 7.08219667 -11.63
```

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

Flowering	atitude	Longitude	bio1	bio2
118	47.41667	19.33333	10.6	8.9
118	49.00000	11.50000	8.0	8.7
121	49.00000	11.50000	8.0	8.7
123	51.91667	11.50000	9.0	8.1
104	34.35000	47.23333	12.9	15.8
102	33.23333	48.56667	15.3	15.7
105	33.23333	48.56667	15.3	15.7
113	39.25000	45.50000	12.0	11.8
98	40.93333	27.28333	13.0	9.7
117	38.16667	38.31667	9.6	11.6
116	37.75000	39.71667	14.4	11.4
113	39.25000	45.50000	12.0	11.8
115	45.00000	34.00000	10.9	9.5
109	39.25000	45.50000	12.0	11.8

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

>summary(model)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 142.36492   20.07386   7.092 4.78e-09 ***
Latitude      0.09498    0.34348   0.277    0.783
Longitude     0.16722    0.16999   0.984    0.330
bio1         -2.82833    0.38874  -7.276 2.48e-09 ***
bio2         -0.57061    0.58483  -0.976    0.334
```

```
> model$residuals
          1
 2.95716100  -3.35
          9
-10.51200682  -1.62
         17
 7.27865394 -13.65
         25
 7.08219667 -11.63
```

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

”Dummy” variable for intercept

$$X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

Flowering	ıtitude	Longitude	bio1	bio2
118	47.41667	19.33333	10.6	8.9
118	49.00000	11.50000	8.0	8.7
121	49.00000	11.50000	8.0	8.7
123	51.91667	11.50000	9.0	8.1
104	34.35000	47.23333	12.9	15.8
102	33.23333	48.56667	15.3	15.7
105	33.23333	48.56667	15.3	15.7
113	39.25000	45.50000	12.0	11.8
98	40.93333	27.28333	13.0	9.7
117	38.16667	38.31667	9.6	11.6
116	37.75000	39.71667	14.4	11.4
113	39.25000	45.50000	12.0	11.8
115	45.00000	34.00000	10.9	9.5
109	39.25000	45.50000	12.0	11.8

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

>summary(model)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 142.36492   20.07386   7.092 4.78e-09 ***
Latitude      0.09498    0.34348   0.277    0.783
Longitude     0.16722    0.16999   0.984    0.330
bio1         -2.82833    0.38874  -7.276 2.48e-09 ***
bio2         -0.57061    0.58483  -0.976    0.334
```

```
> model$residuals
          1
 2.95716100  -3.35
          9
-10.51200682  -1.62
         17
 7.27865394 -13.65
         25
 7.08219667 -11.63
```

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

”Dummy” variable for intercept

$$X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

Flowering	ıtitude	Longitude	bio1	bio2
118	47.41667	19.33333	10.6	8.9
118	49.00000	11.50000	8.0	8.7
121	49.00000	11.50000	8.0	8.7
123	51.91667	11.50000	9.0	8.1
104	34.35000	47.23333	12.9	15.8
102	33.23333	48.56667	15.3	15.7
105	33.23333	48.56667	15.3	15.7
113	39.25000	45.50000	12.0	11.8
98	40.93333	27.28333	13.0	9.7
117	38.16667	38.31667	9.6	11.6
116	37.75000	39.71667	14.4	11.4
113	39.25000	45.50000	12.0	11.8
115	45.00000	34.00000	10.9	9.5
109	39.25000	45.50000	12.0	11.8

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

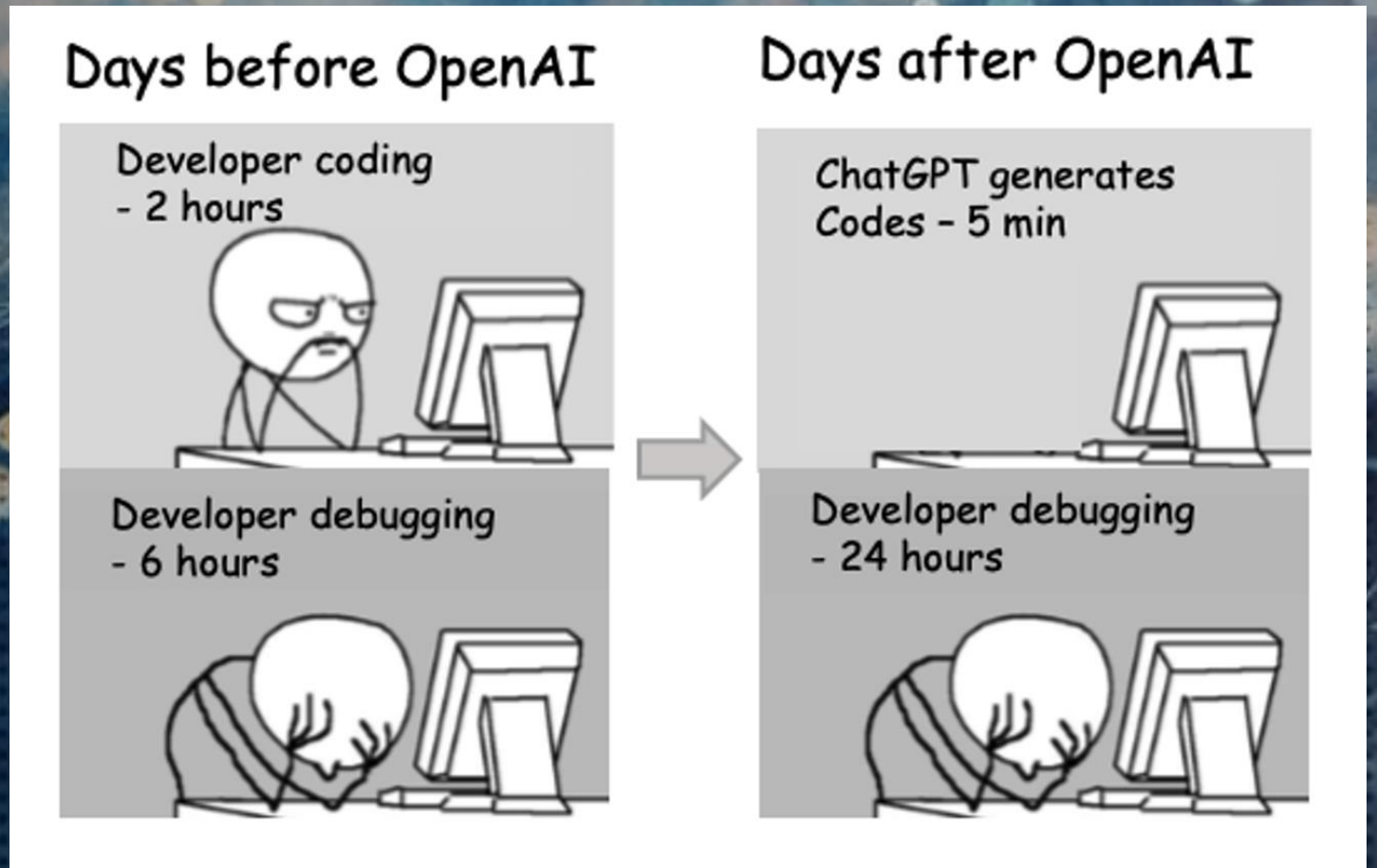
$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

>summary(model)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 142.36492   20.07386   7.092 4.78e-09 ***
Latitude      0.09498    0.34348   0.277    0.783
Longitude     0.16722    0.16999   0.984    0.330
bio1         -2.82833    0.38874  -7.276 2.48e-09 ***
bio2         -0.57061    0.58483  -0.976    0.334
```

```
> model$residuals
          1
 2.95716100  -3.35
          9
-10.51200682  -1.62
         17
 7.27865394 -13.65
         25
 7.08219667 -11.63
```

Friday: more R code and how to use ChatGPT responsibly



Next week: General Linear Models