Multiple regression – the “models of all models”!

Part I (continuation):

model, properties of estimators and sensibility to assumptions

Part II:

Goodness of fit and model simplicity metrics, hypotheses testing, standardized slopes, model selection, examples and diagnostics

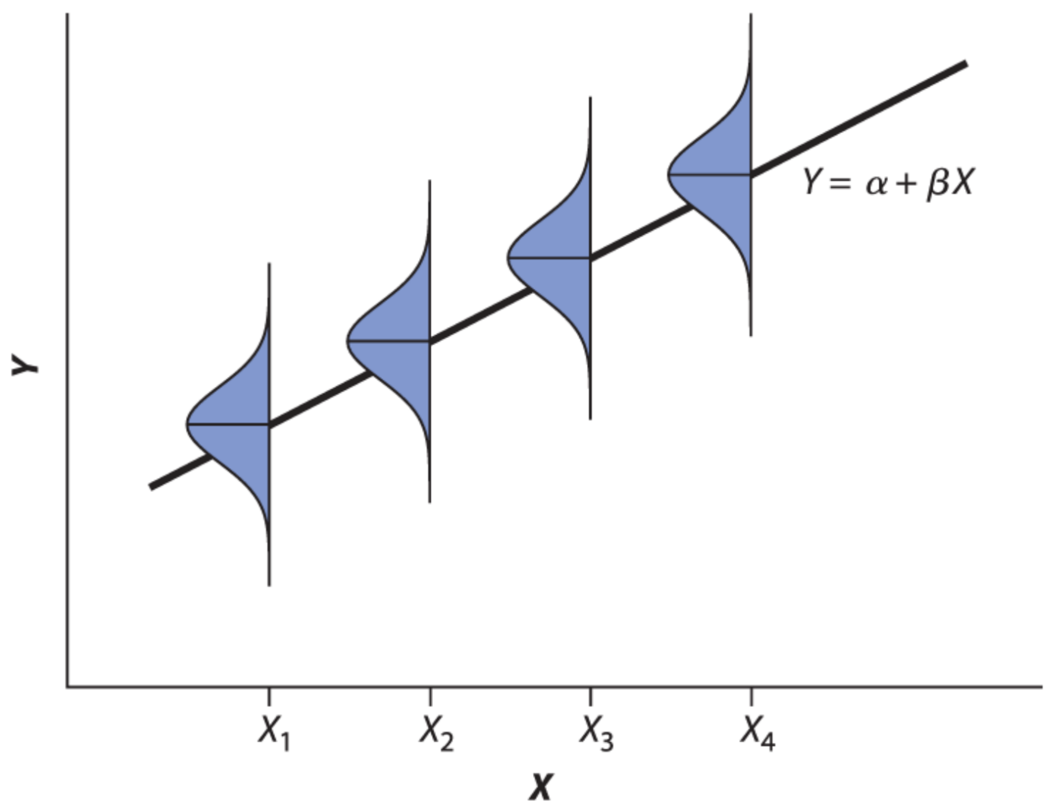
The properties of a regression model -

[1] Properties of errors in response Y and predictors X

Properties of errors

multiple regression assumes measurement errors in Y but not X

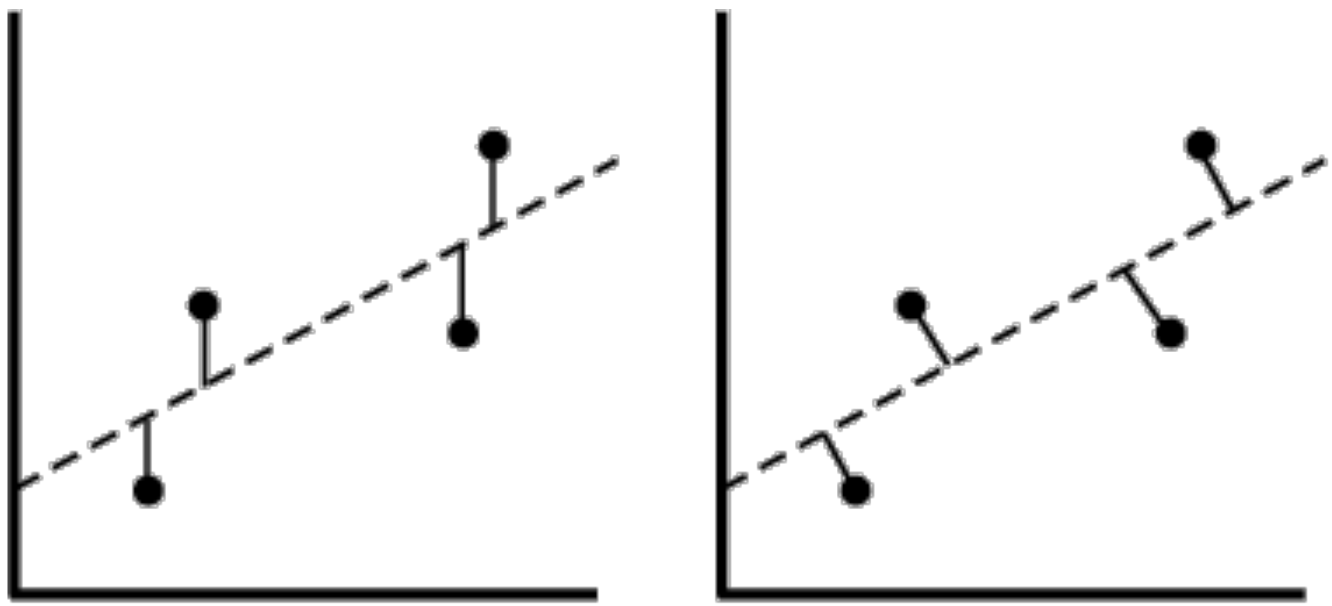
A regression model aims at predicting the average Y based on X, i.e., predict the average Y based on X.



Values of X (predictor) are measured without error (hard to assess, often assumed).

Properties of errors: Values of X (predictor) are measured without measurement error (hard to assess, often assumed)

multiple regression assumes vertical offsets (residuals)



vertical offsets

perpendicular offsets

Residuals for Type I regression
Error in Y but not in X

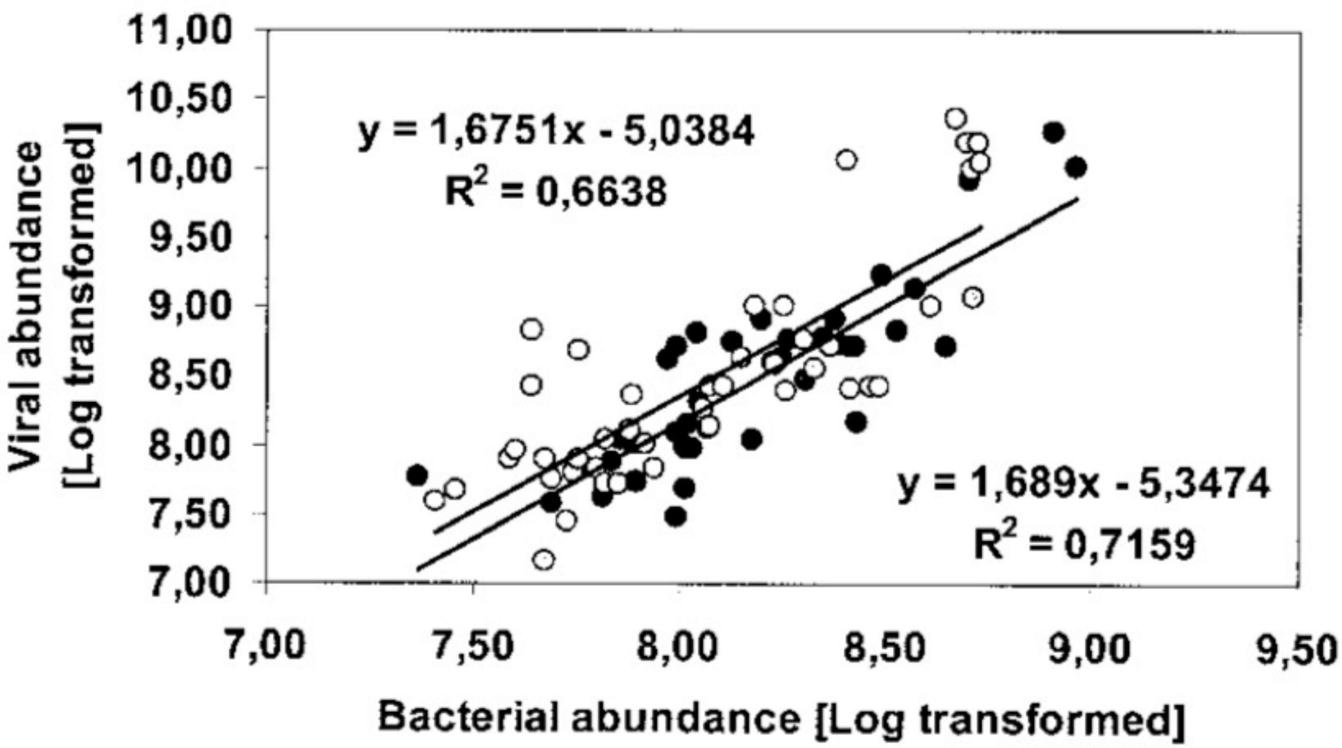
Residuals for Type II regression
Error in both Y and X

Type I and III sum-of-squares

Type II sum-of-squares

Properties of errors (assumption): values of X (predictor) is measured without measurement error

Viral abundance (log transformed)



Bacterial abundance (log transformed)

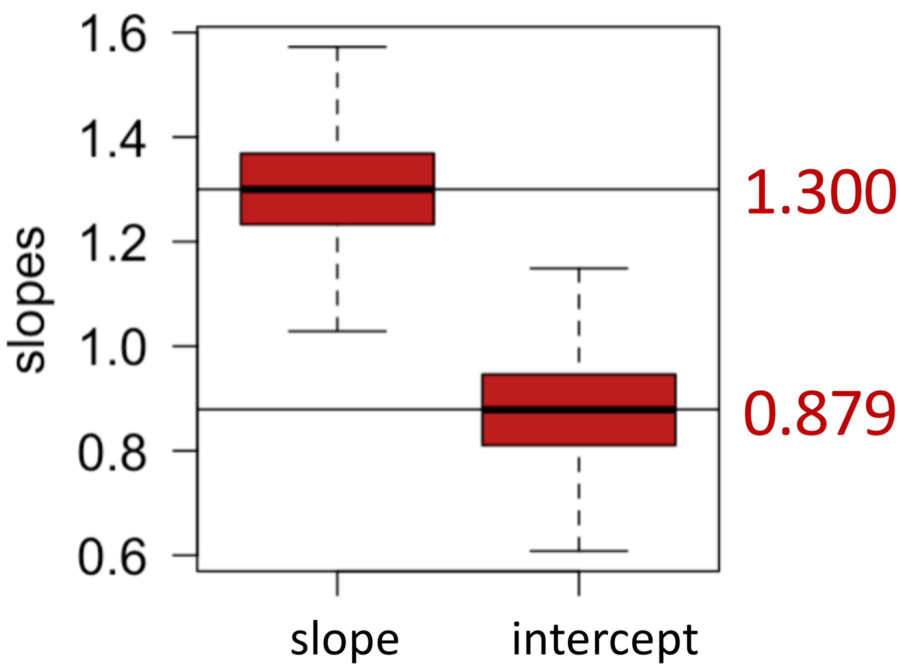
If we assume here that bacterial and viral abundance have the same measurement errors, then we can't use the regular regression model (the authors used a type II regression that is appropriate for this issue).

Corinaldesi et al. (2003); APPLIED AND ENVIRONMENTAL MICROBIOLOGY, May: 2664–2673.

Properties of errors (assumption): values of X (predictor) is measured without measurement

But first we need to revisit understand that the regression model based on samples are an unbiased estimate of the true intercepts and slopes. Let's assume the following population regression model:

$$Y = 0.879 + 1.300X$$

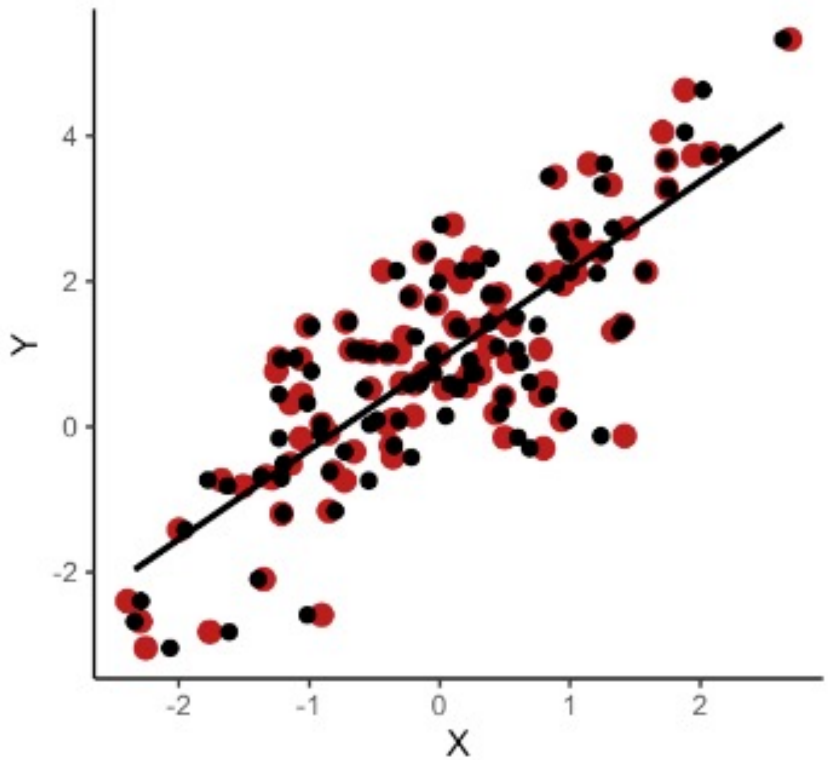


Sampling variation in estimates

```r
slopes <- c()
intercept <- c()
for (i in 1:10000){
  X <- rnorm(100)
  e <- rnorm(100)
  Y <- 0.879 + 1.3*X + e
  lm.fit <- lm(Y ~ X)
  slopes[i] <- lm.fit$coefficients["X"]
  intercept[i] <- lm.fit$coefficients["(Intercept)"]
}
boxplot(slopes,intercept,col="firebrick",outline = FALSE,
        ylab="slopes",las = 1,cex.axis=1.3,cex.lab=1.3)
```

Properties of errors (assumption): values of X (predictor) is measured without error (hard to assess, often assumed)

SMALL MEASUREMENT ERROR

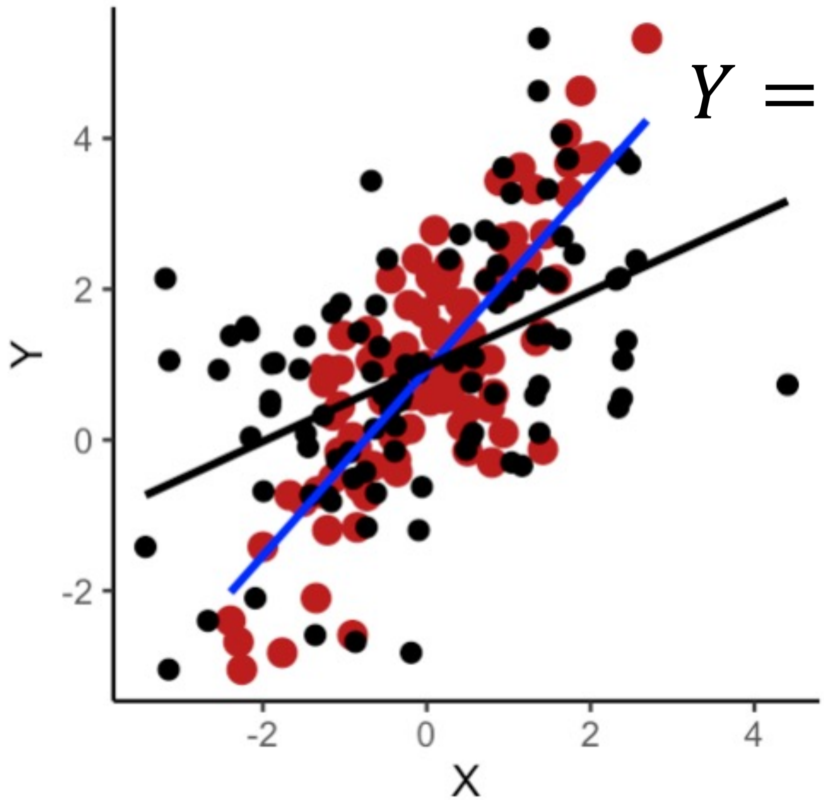



```
X <- rnorm(100)
e <- rnorm(100)
Y <- 0.879 + 1.3*X + e
X.error <- rnorm(100,X,sd=0.1)
```

Red dots are X values "measured" without error, whereas the smaller black dots are X values "measured" with error.

In this case there is little consequence because the error is small (0.1).

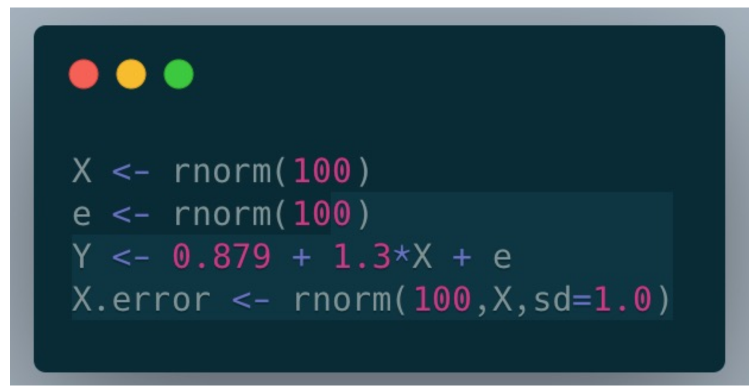
Properties of errors (assumption): values of X (predictor) is measured without error (hard to assess, often assumed)



$$Y = 0.929 + 1.23X$$ without error in X

$$Y = 0.977 + 0.498X$$ with error in X

LARGE MEASUREMENT ERROR



BLUE line = Regression model without error in X.

BLACK line = Regression model with error in X.

ERROR IN X REDUCES SLOPES.

Red dots are X values “measured” without error, whereas the smaller black does are X values “measured” with error.

The consequence here is much bigger for estimating the regression model because the error is large (1.0).

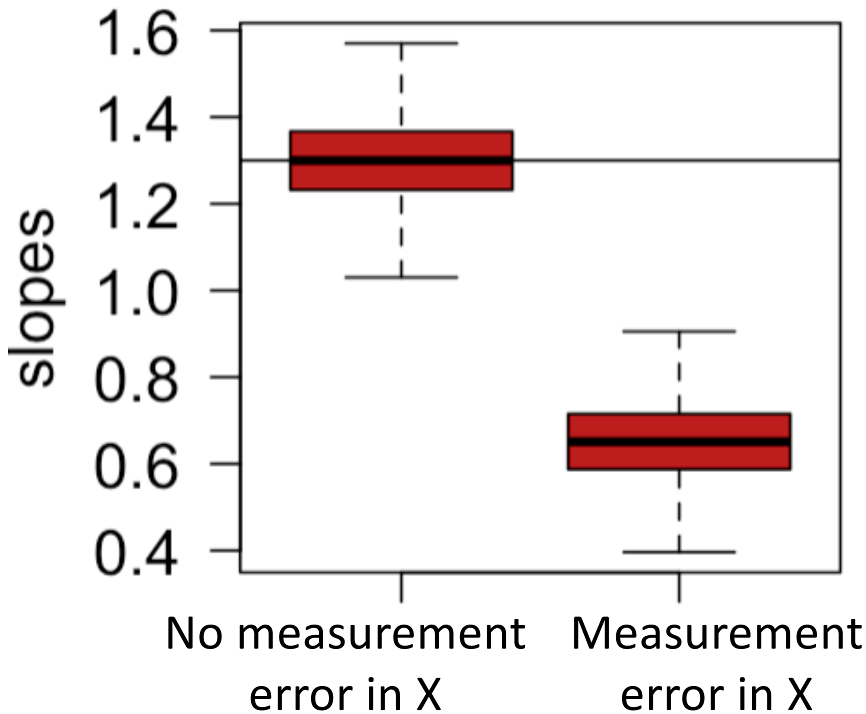
Properties of errors (assumption): values of X (predictor) is measured without error (hard to assess, often assumed)

$$Y = 0.879 + 1.300X$$ True population model

```
slopes <- c()
slopes.error <- c()
for (i in 1:10000){
  X <- rnorm(100)
  e <- rnorm(100)
  Y <- 0.879 + 1.3*X + e
  lm.fit <- lm(Y ~ X)
  slopes[i] <- lm.fit$coefficients["X"]
  X.error <- rnorm(100,X,sd=1)
  lm.fit <- lm(Y ~ X.error)
  slopes.error[i] <- lm.fit$coefficients["X.error"]
}
boxplot(slopes,slopes.error,col="firebrick",outline = FALSE,
        ylab="slopes",las = 1,cex.axis=1.3,cex.lab=1.3)
```

Properties of errors (assumption): values of X (predictor) is measured without error (hard to assess, often assumed)

$$Y = 0.879 + 1.300X$$ True population model




```r
slopes <- c()
slopes.error <- c()
for (i in 1:10000){
  X <- rnorm(100)
  e <- rnorm(100)
  Y <- 0.879 + 1.3*X + e
  lm.fit <- lm(Y ~ X)
  slopes[i] <- lm.fit$coefficients["X"]
  X.error <- rnorm(100,X,sd=1)
  lm.fit <- lm(Y ~ X.error)
  slopes.error[i] <- lm.fit$coefficients["X.error"]
}
boxplot(slopes,slopes.error,col="firebrick",outline = FALSE,
        ylab="slopes",las = 1,cex.axis=1.3,cex.lab=1.3)
```

The properties of a regression model -

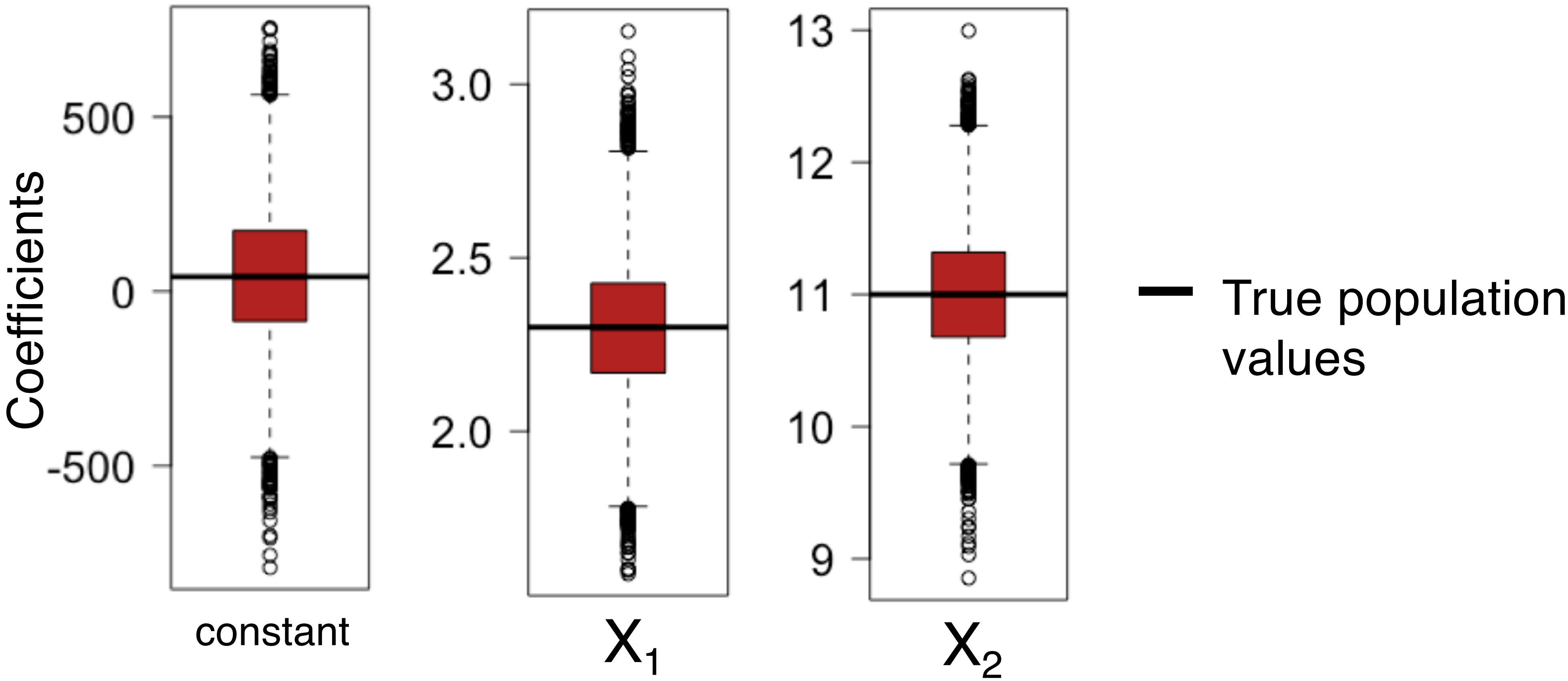
[2] Properties of estimators of coefficients and residual variance

Properties of estimators of coefficients

(sampling variation of coefficients; 10000 samples)

True population model:

$$Y = 42\text{cm} + \mathbf{2.3}X_1 + \mathbf{11}X_2 + e$$



Note that there is much more relative sampling error around constant than the slopes.

Properties of estimators of residual variance

mean of residuals is always zero

$$\sigma^2 = E(s^2) = \frac{\sum_{i=1}^{n}(e_i - 0)^2}{n - (k+1)}$$

number of parameters estimated (intercept + number of slopes)

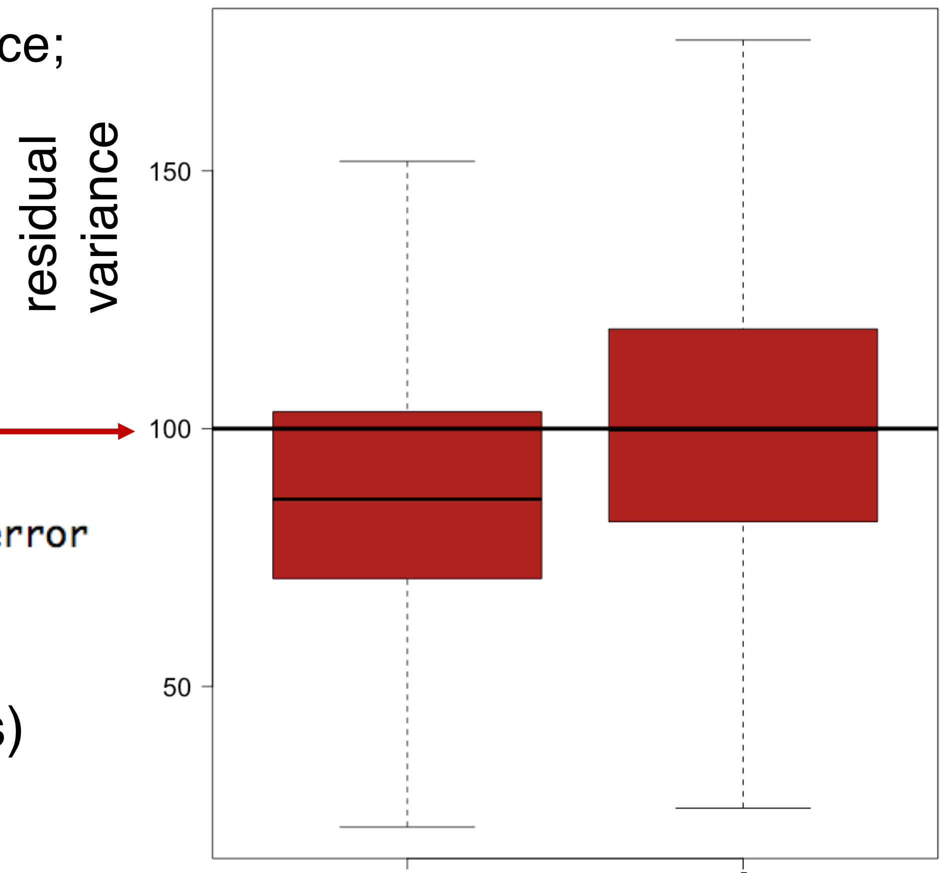
1 degree of freedom is lost because of the mean of residuals, which is always zero here

$$e_i = residual\ of\ the\ ith\ observation$$

Properties of estimators of residual variance and the roles of degrees of freedom

(sampling variation of residual variance; 10 000 samples)

```
n = 30
constant = 42
X1 = rnorm(n,1000,10)
X2 = rnorm(n,40,4)
error = rnorm(n,0,10)

Y = constant + 2.3*X1 + 11*X2 + error
```



k = 3 (intercept + 2 predictors)

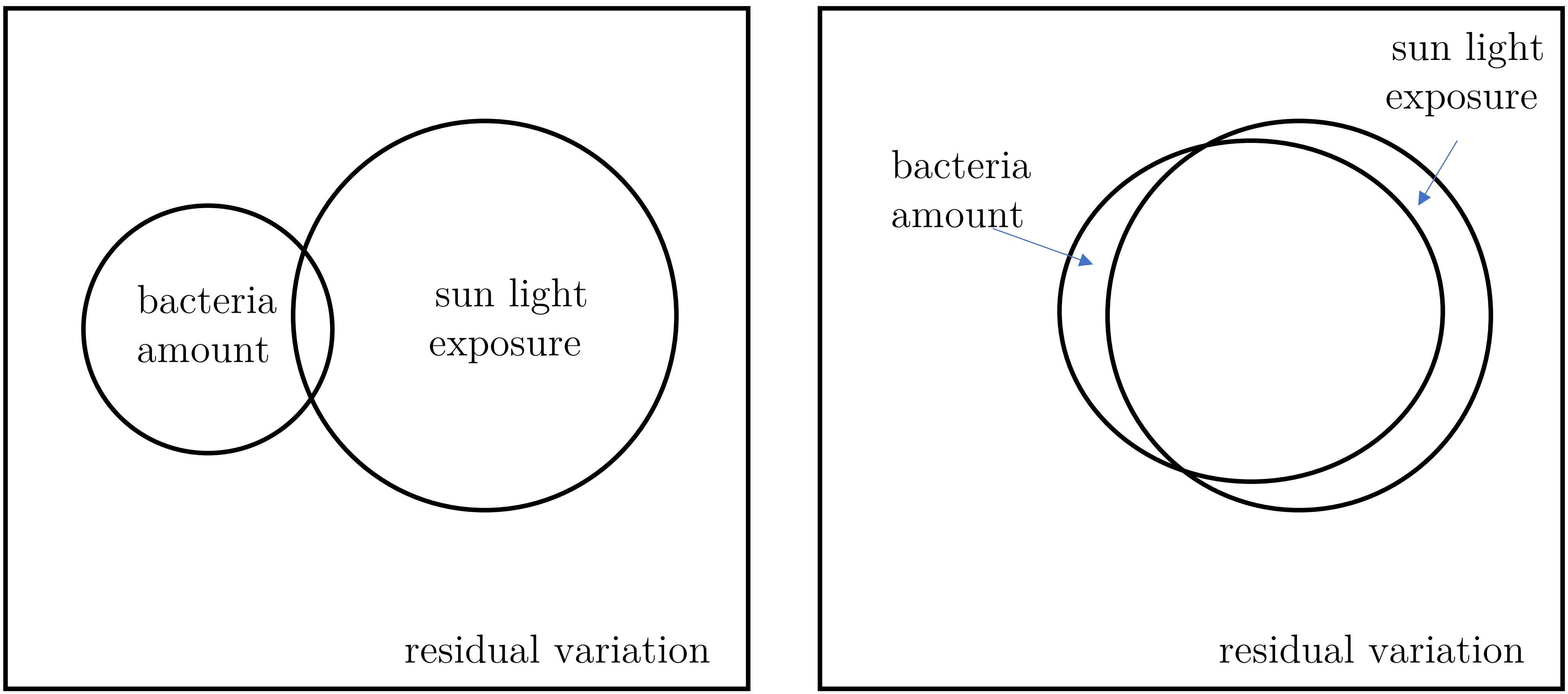
$$\sigma^2 = E(s^2) = \frac{\sum_{i=1}^{n}(Y_i - 0)^2}{n - (k+1)}$$

To be properly estimated, the variance of residuals needs to take into account the number of predictors in the model

$$\sigma^2 = E(s^2) = \frac{\sum_{i=1}^{n}(Y_i - 0)^2}{n}$$

The properties of a regression model

[3] The influence of missing predictors that correlate with measured predictors (e.g., measuring the effect of bacteria without sun light); e.g., extreme cases are called multicollinearity



versus

```r

n = 1000
constant = 42
X1 = rnorm(n,1000,10)
X2 = rnorm(n,40,4)
error = rnorm(n,0,10)
Y = constant + 2.3*X1 + 11*X2 + error

```

```
> cor(X1,X2)
[1] -0.009406406
```

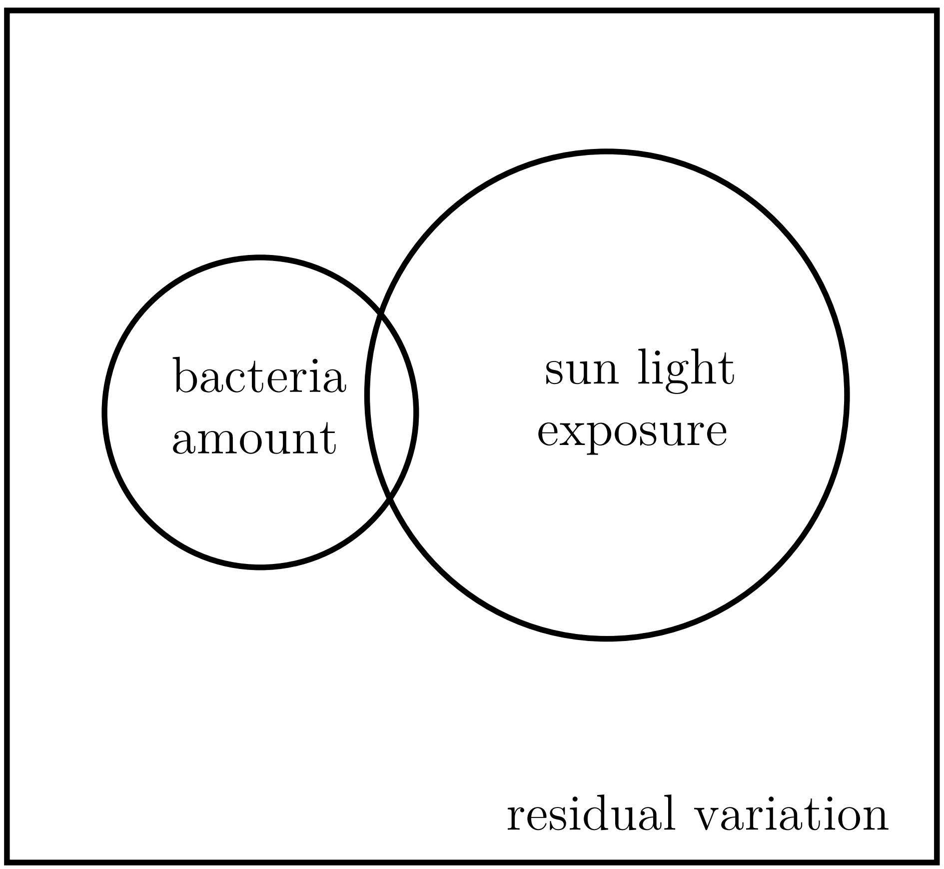
```
> lm(Y~X1)

Call:
lm(formula = Y ~ X1)

Coefficients:
(Intercept)           X1
     561.39         2.22
```

```
> lm(Y~X1+X2)

Call:
lm(formula = Y ~ X1 + X2)

Coefficients:
(Intercept)           X1           X2
     83.941        2.262       10.919
```

Compare the two models – both slopes for X1 are very similar

The properties of a regression model

Small influence of missing predictors that do not correlate strongly with measured predictors




```
> cor(X1,X2)
[1] -0.009406406
```

```
n = 1000
constant = 42
X1 = rnorm(n,1000,10)
X2 = X1+rnorm(n,40,4)
error = rnorm(n,0,10)
Y = constant + 2.3*X1 + 11*X2 + error
```

```
> cor(X1,X2)
[1] 0.9366205
```

```r

n = 1000
constant = 42
X1 = rnorm(n,1000,10)
X2 = X1+rnorm(n,40,4)
error = rnorm(n,0,10)
Y = constant + 2.3*X1 + 11*X2 + error

```

```
> cor(X1,X2)
[1] 0.9366205
```

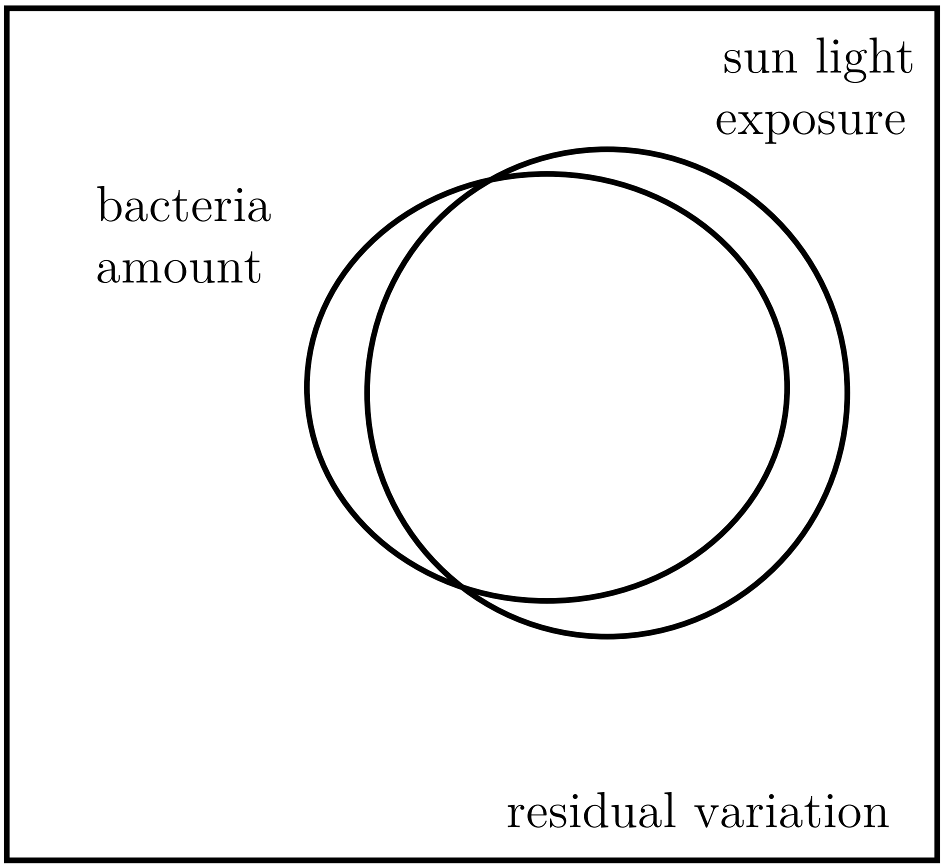
```
> lm(Y~X1)

Call:
lm(formula = Y ~ X1)

Coefficients:
(Intercept)           X1
     293.89        13.49
```

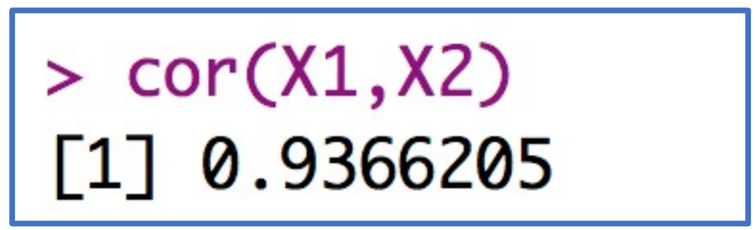
```
> lm(Y~X1+X2)

Call:
lm(formula = Y ~ X1 + X2)

Coefficients:
(Intercept)           X1           X2
      9.267        2.252       11.077
```

Compare the two models – slopes are now very different, i.e., the missing predictor X2 in the first model affected the true estimation of X1.

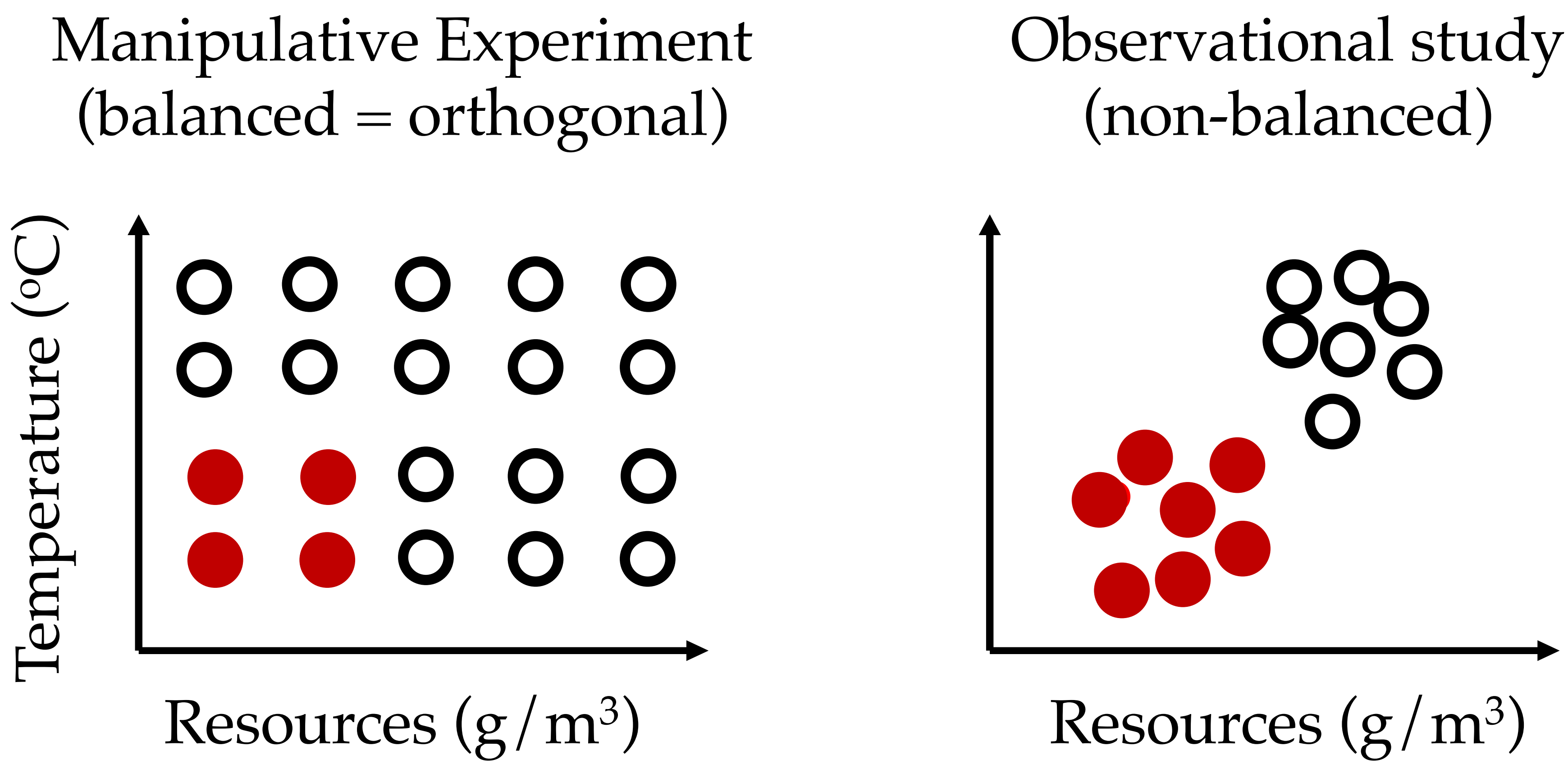
The properties of a regression model

Strong influence of missing predictors that correlate strongly with measured predictors





Experimental (likely close to orthogonal) versus observational (likely non-orthogonal) approaches.



Optimal combination of the two variables for fish growth.

The properties of a regression model
(now let's use a small simulation)

Properties of estimators
[4] sampling variation of coefficients

low *versus* high correlation among predictors

```
n = 1000
constant = 42
X1 = rnorm(n,1000,10)
X2 = X1+rnorm(n,40,4)
error = rnorm(n,0,10)
Y = constant + 2.3*X1 + 11*X2 + error
```

```
> cor(X1,X2)
[1] 0.9366205
```

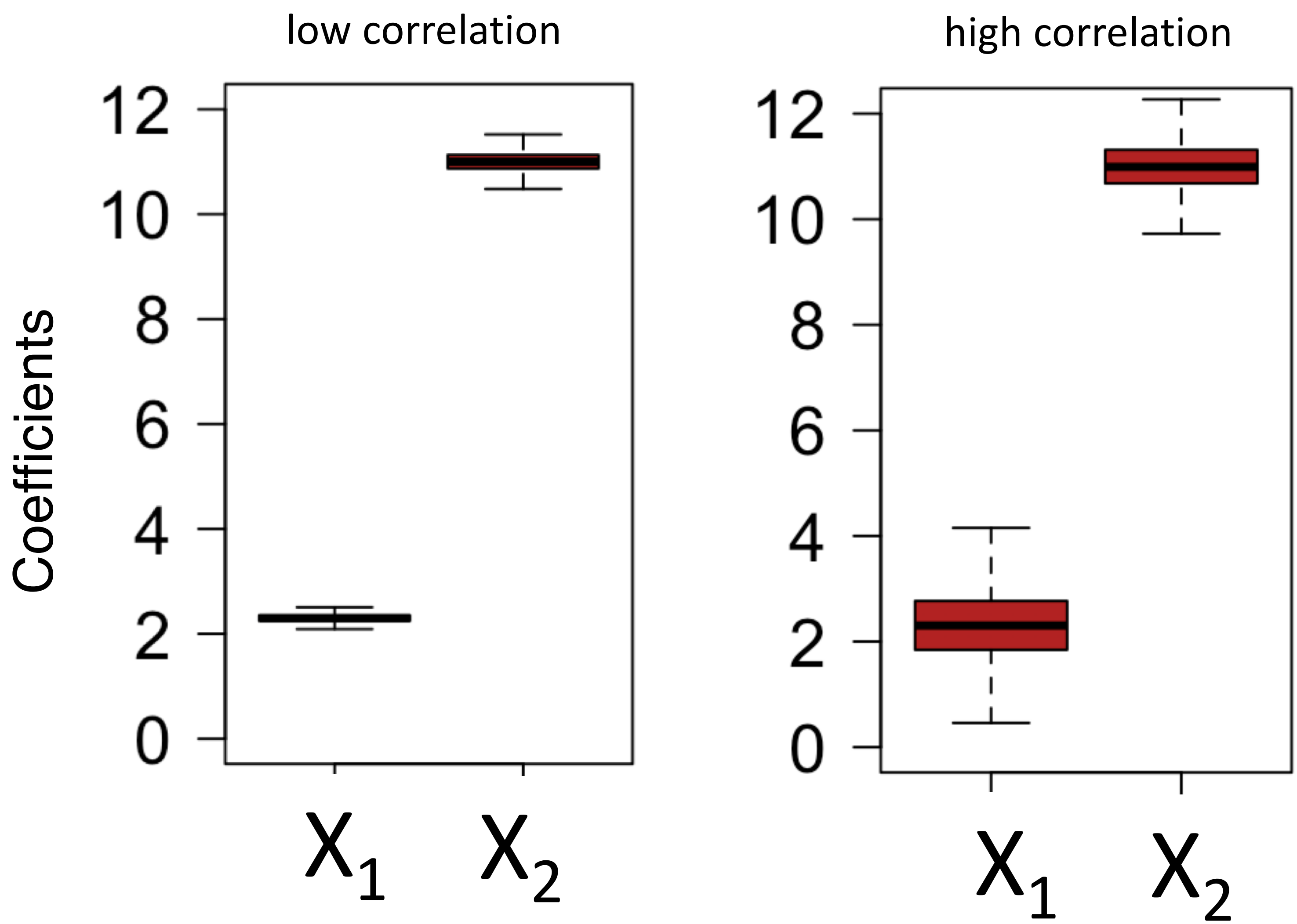
```
> lm(Y~X1)

Call:
lm(formula = Y ~ X1)

Coefficients:
(Intercept)           X1
     293.89        13.49
```

```
> lm(Y~X1+X2)

Call:
lm(formula = Y ~ X1 + X2)

Coefficients:
(Intercept)           X1           X2
      9.267        2.252       11.077
```

But even when we consider the « correct » predictors, the error estimation (sampling error) of slopes is affected when they are very correlated.

Level of correlation between predictors affects estimation accuracy (Variation Inflation) – we can trust less the slopes of predictors that are correlated

$$Y = 42\text{cm} + \mathbf{2.3}X_1 + \mathbf{11}X_2 + e$$



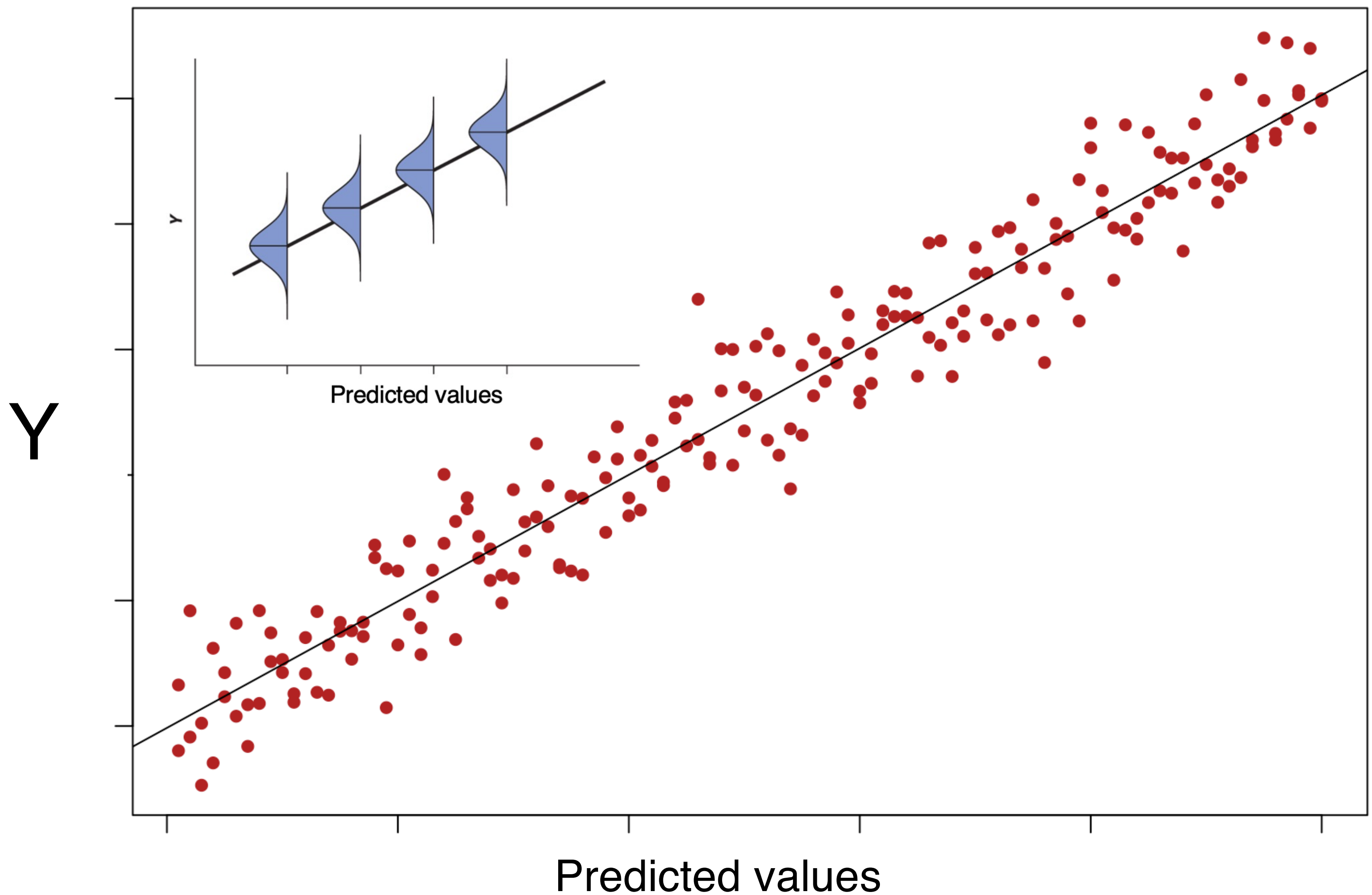
[5] Homoscedasticity of residuals

(the assumption of constant variance)

e residual error assumed to be $N(0, \sigma^2)$

$$Y = 42\text{cm} + \mathbf{2.3}\text{X}_1 + \mathbf{11}\text{X}_2 + e$$

e residual error assumed to be $N(0, \sigma^2)$
The assumption of constant residual variance
(homoscedasticity)

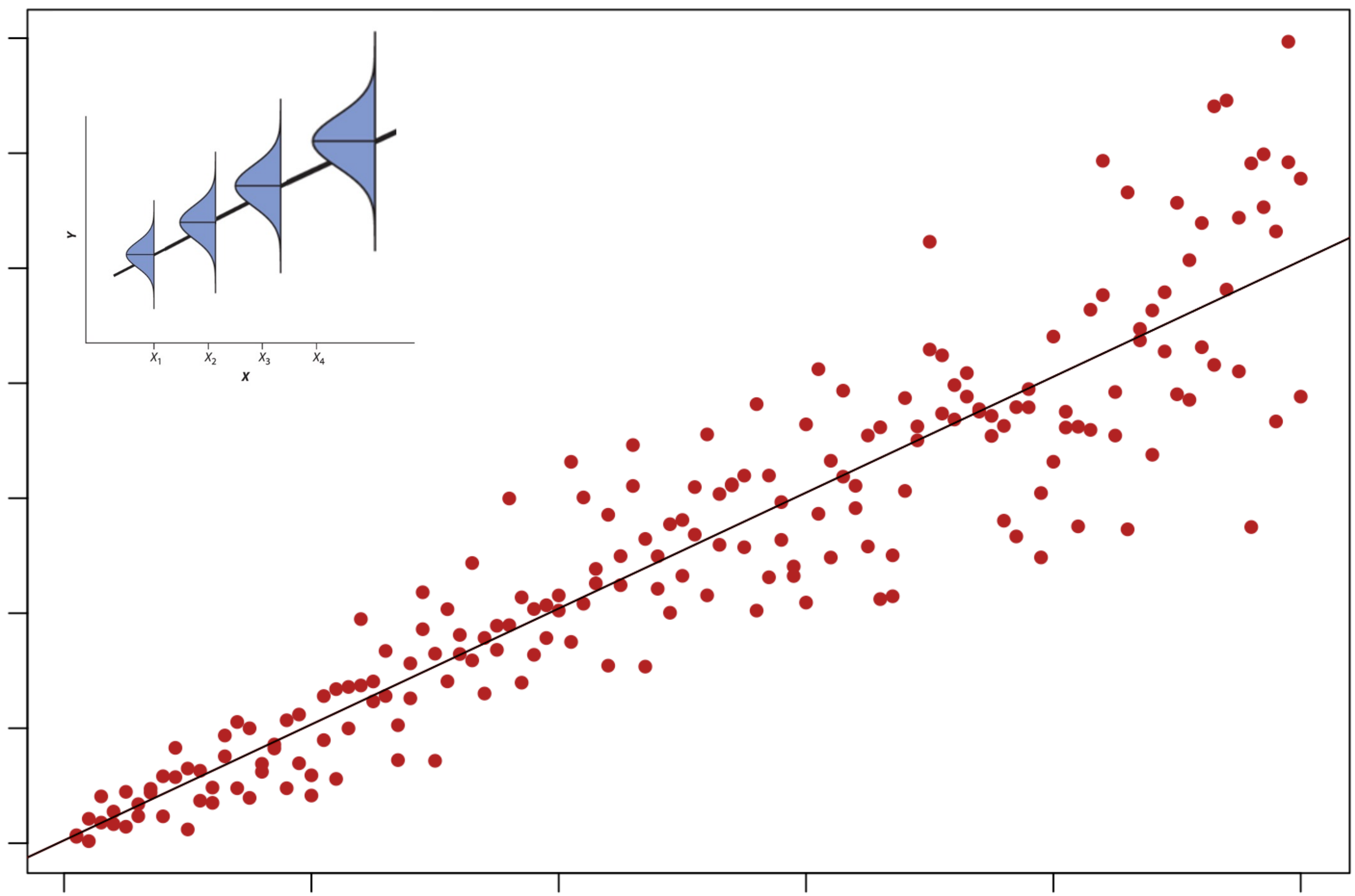


e residual error assumed to be $N(0, \sigma^2)$

The assumption of constant residual variance

(this one is not constant)

Y

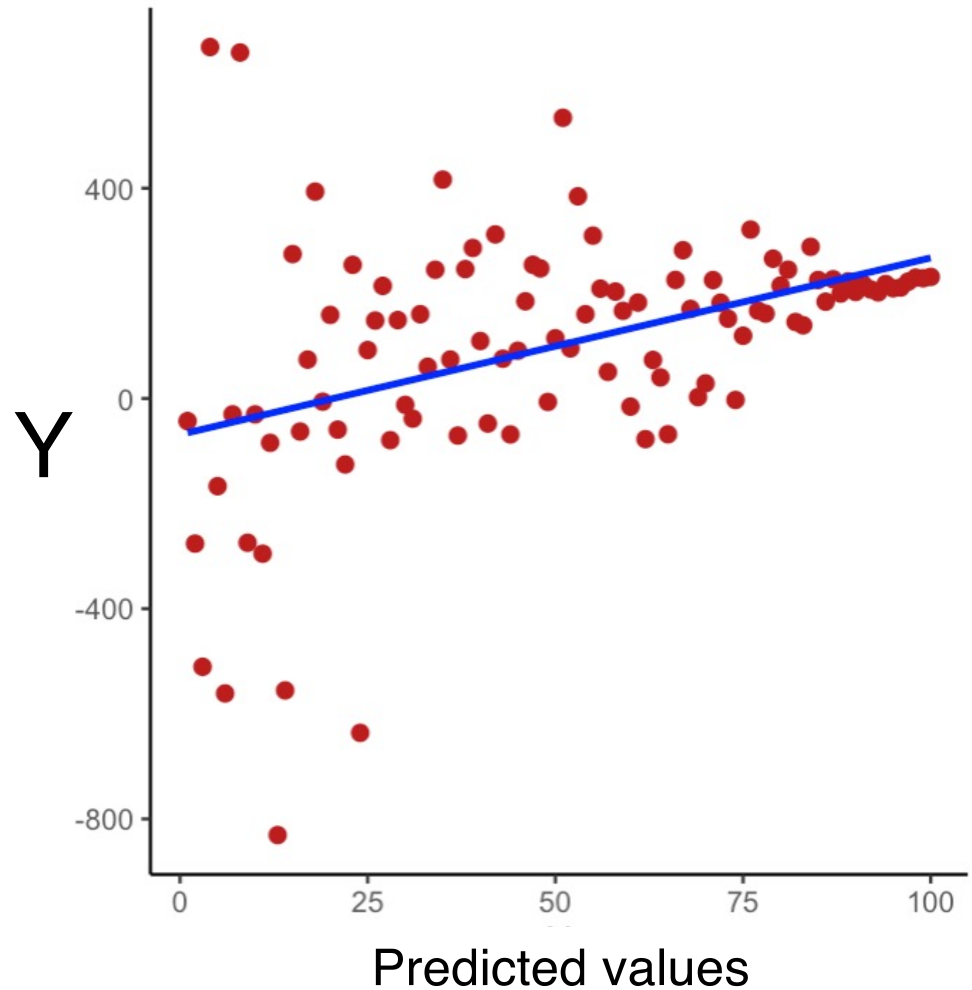


Predicted values

e residual error assumed to be $N(0, \sigma^2)$

The assumption of constant residual variance

(this one is not constant)

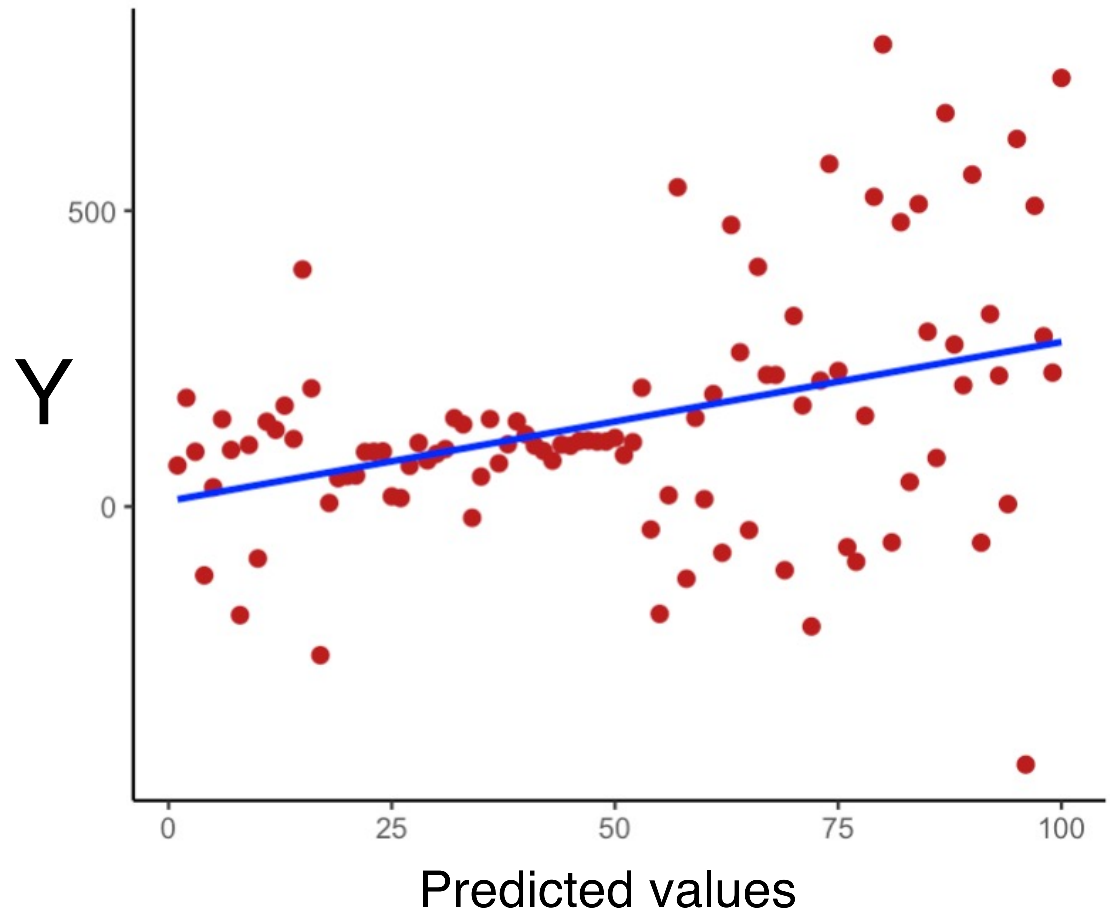


Another example of residual heteroscedasticity

e residual error assumed to be $N(0, \sigma^2)$

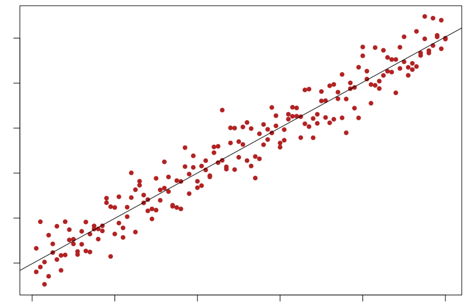
The assumption of constant residual variance

(this one is not constant)

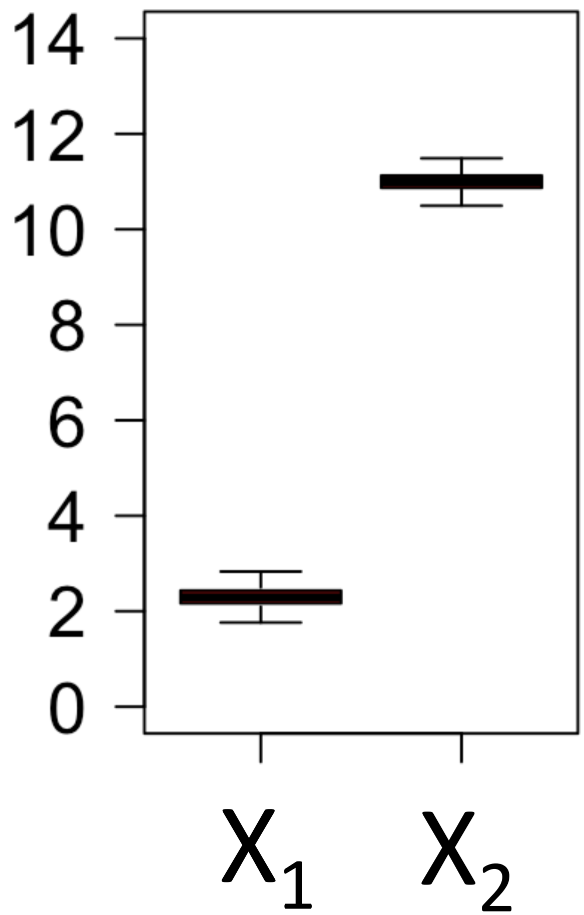


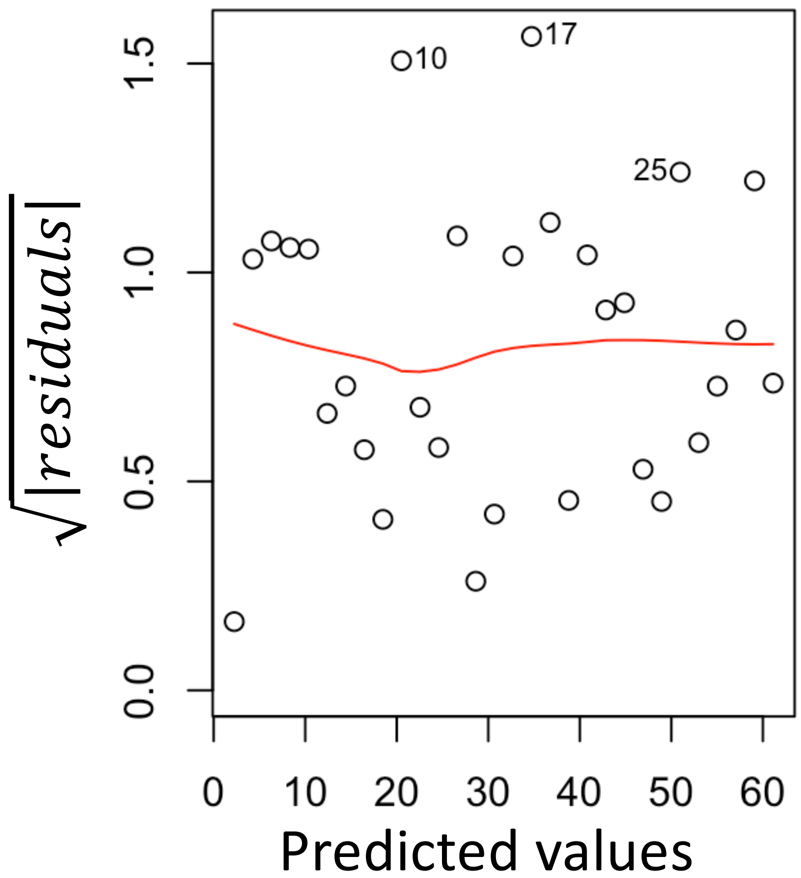
Another example of residual heteroscedasticity non constant residual variance affects estimation accuracy

$$Y = 42\text{cm} + \mathbf{2.3}X_1 + \mathbf{11}X_2 + e$$

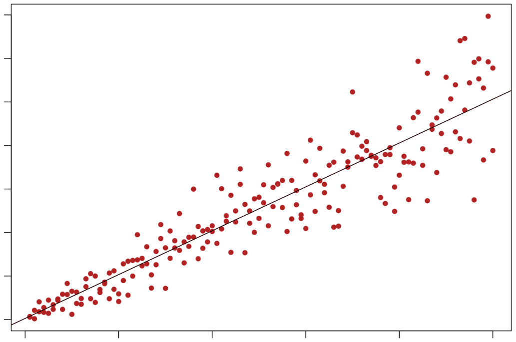


constant variance



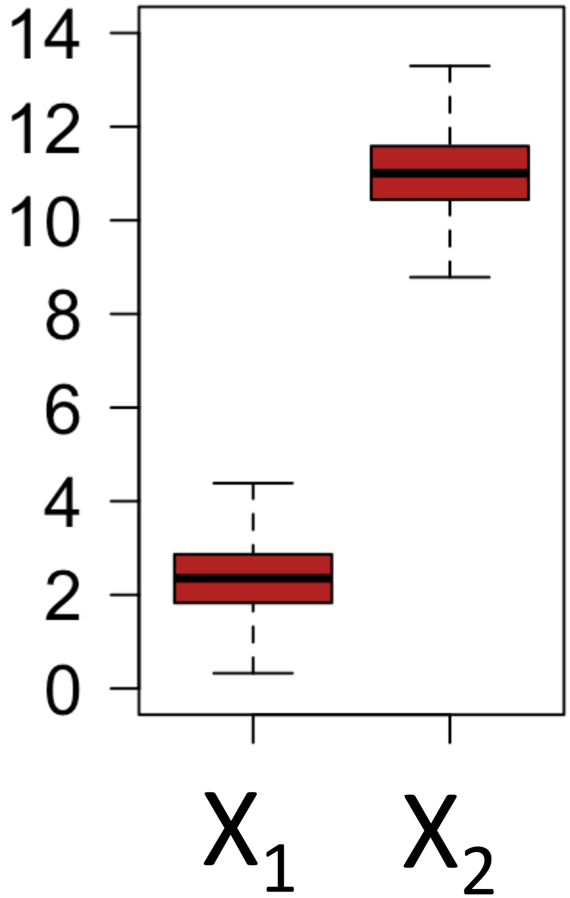


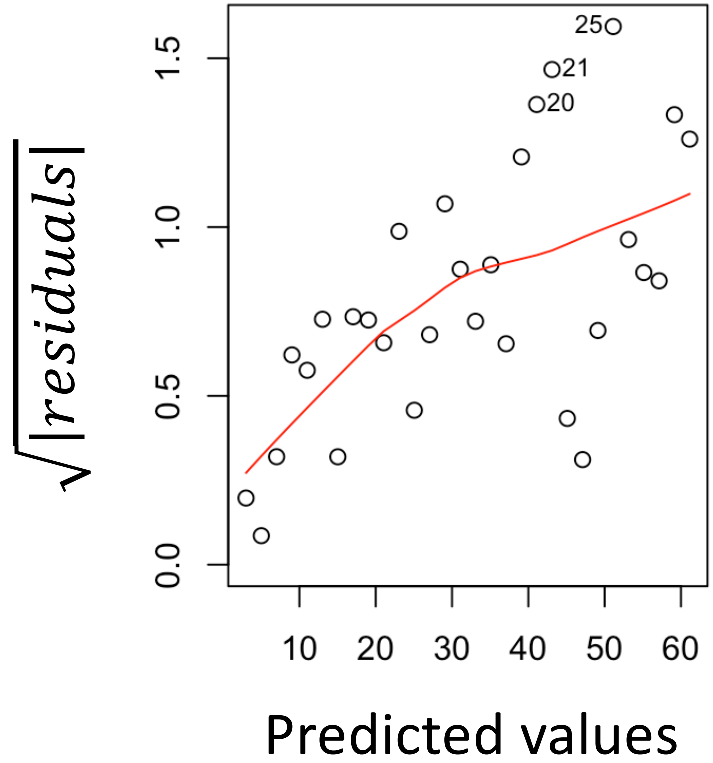
non constant residual variance affects estimation accuracy



$$Y = 42\text{cm} + \color{red}{\mathbf{2.3}}X_1 + \color{red}{\mathbf{11}}X_2 + e$$

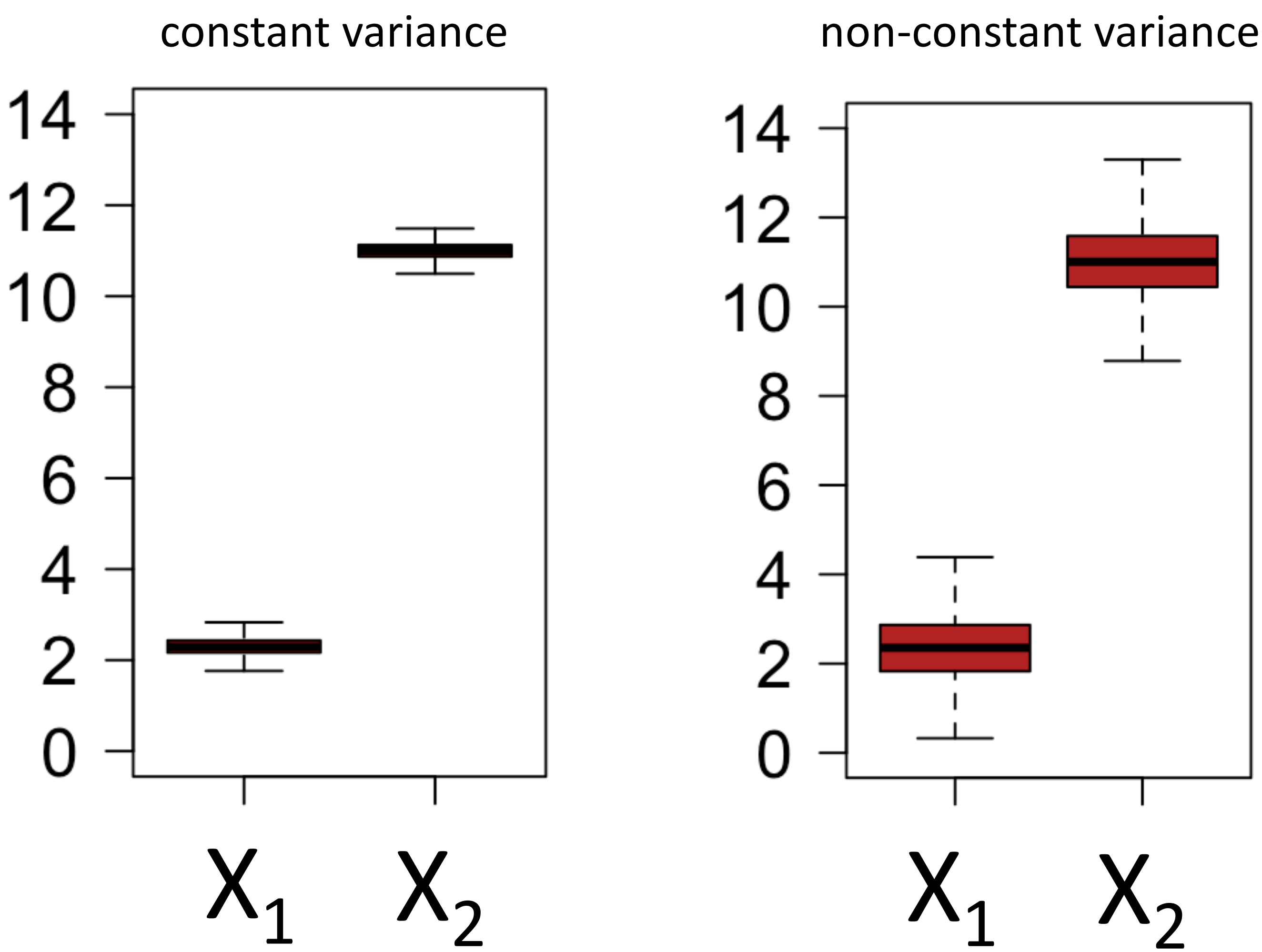
non-constant variance





non constant residual variance affects estimation precision BUT not accuracy

$$Y = 42\text{cm} + \mathbf{2.3}X_1 + \mathbf{11}X_2 + e$$



Multiple regression – the “model of all models”!

Part I:

model, properties of estimators and sensibility to assumptions

Part II:

Goodness of fit and model simplicity metrics, hypotheses testing, standardized slopes, model selection, examples and diagnostics