Regression and Calibration

BIM 283

Advanced Design of Experiments for Biomedical Engineers

Quantitative Prediction

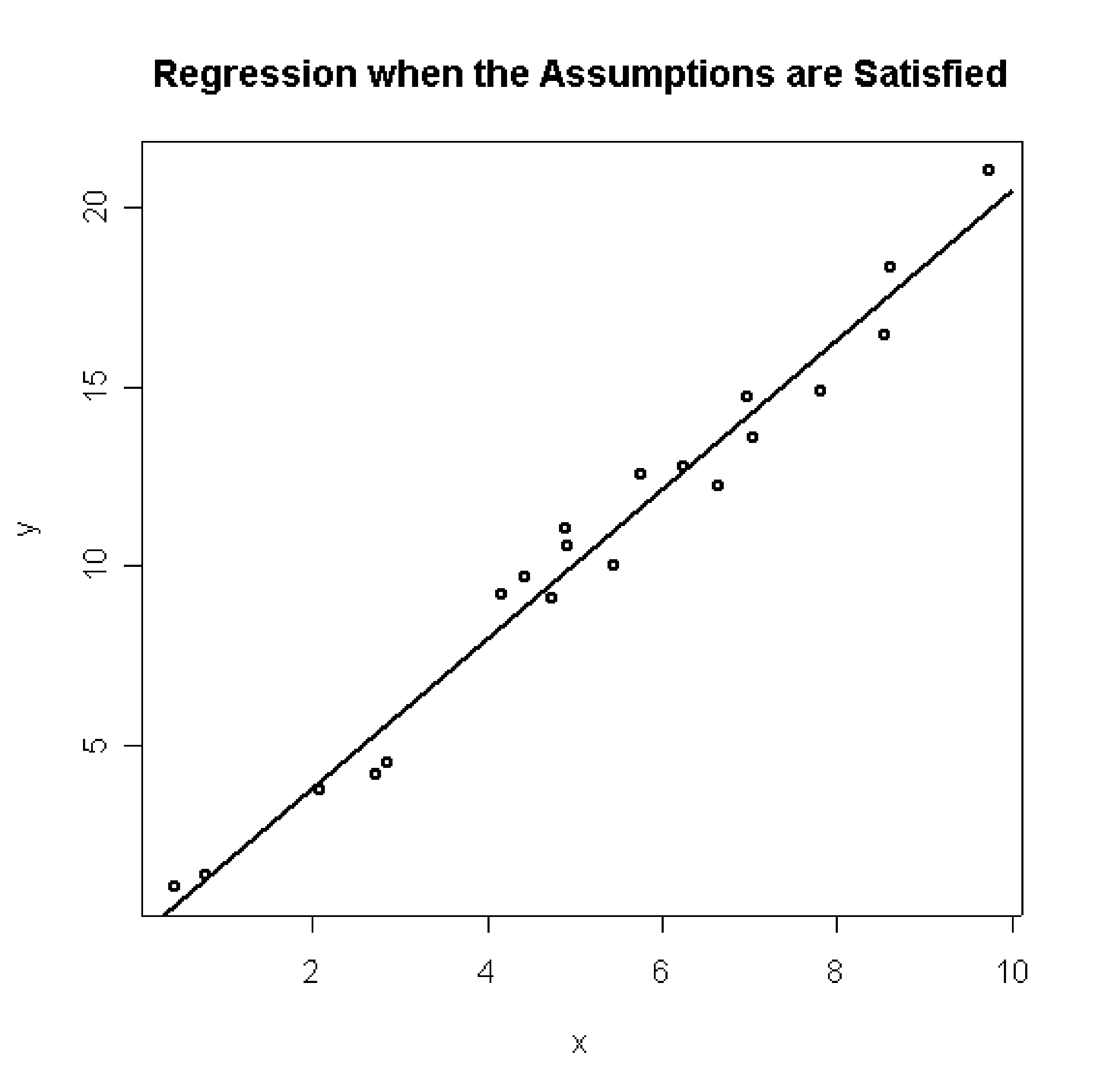
- Regression analysis is the statistical name for the prediction of one quantitative variable (fasting blood glucose level) from another (body mass index)
- Items of interest include whether there is in fact a relationship and what the expected change is in one variable when the other changes

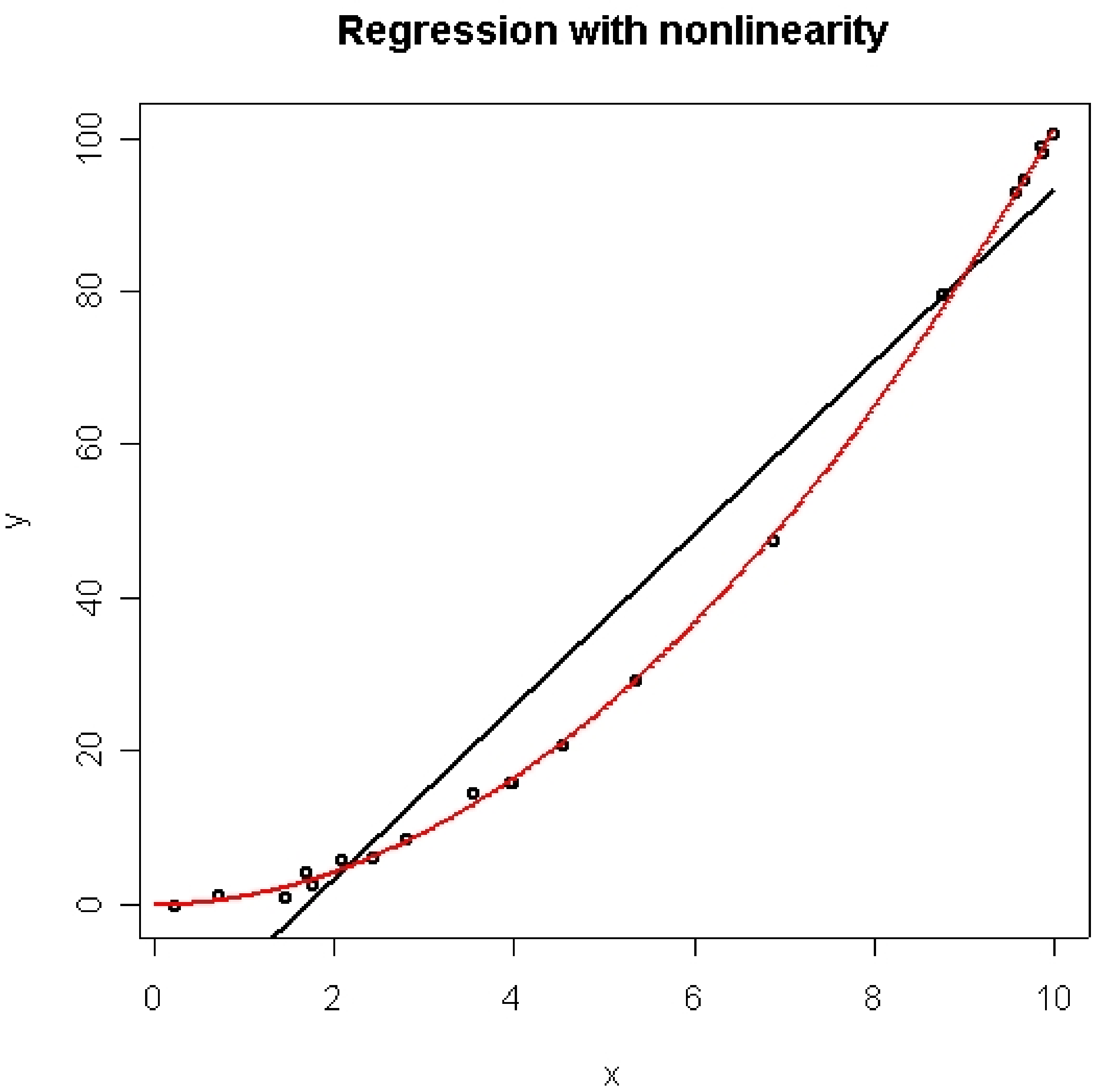
Assumptions

- Inference about whether there is a real relationship or not is dependent on several assumptions, many of which can be checked
- When these assumptions are substantially incorrect, alterations in method can often rescue the analysis
- No assumption is ever exactly correct

Linearity

- This is the most important assumption
- If *x* is the predictor, and *y* is the response, then we assume that the average response for a given value of *x* is a linear function of *x*
- $E(y) = a + bx$
- $y = a + bx + \varepsilon$
- ε is the *error* or variability
- Sometimes linearity is satisfied on the log scale, not on the original scale, especially when variables (like concentration) might range over orders of magnitude

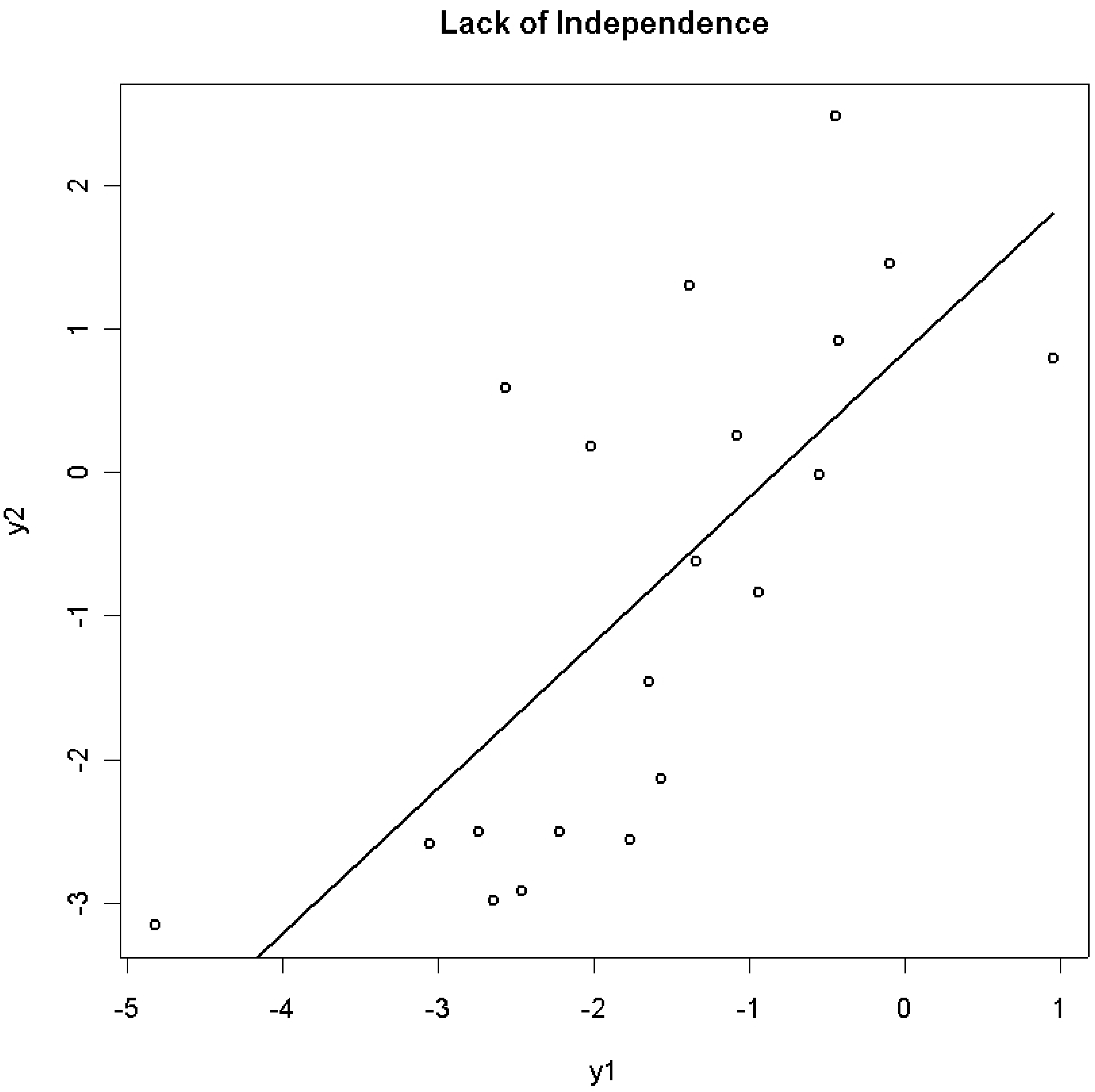

Regression when the Assumptions are Satisfied
y
5
10
15
20
x
2
4
6
8
10

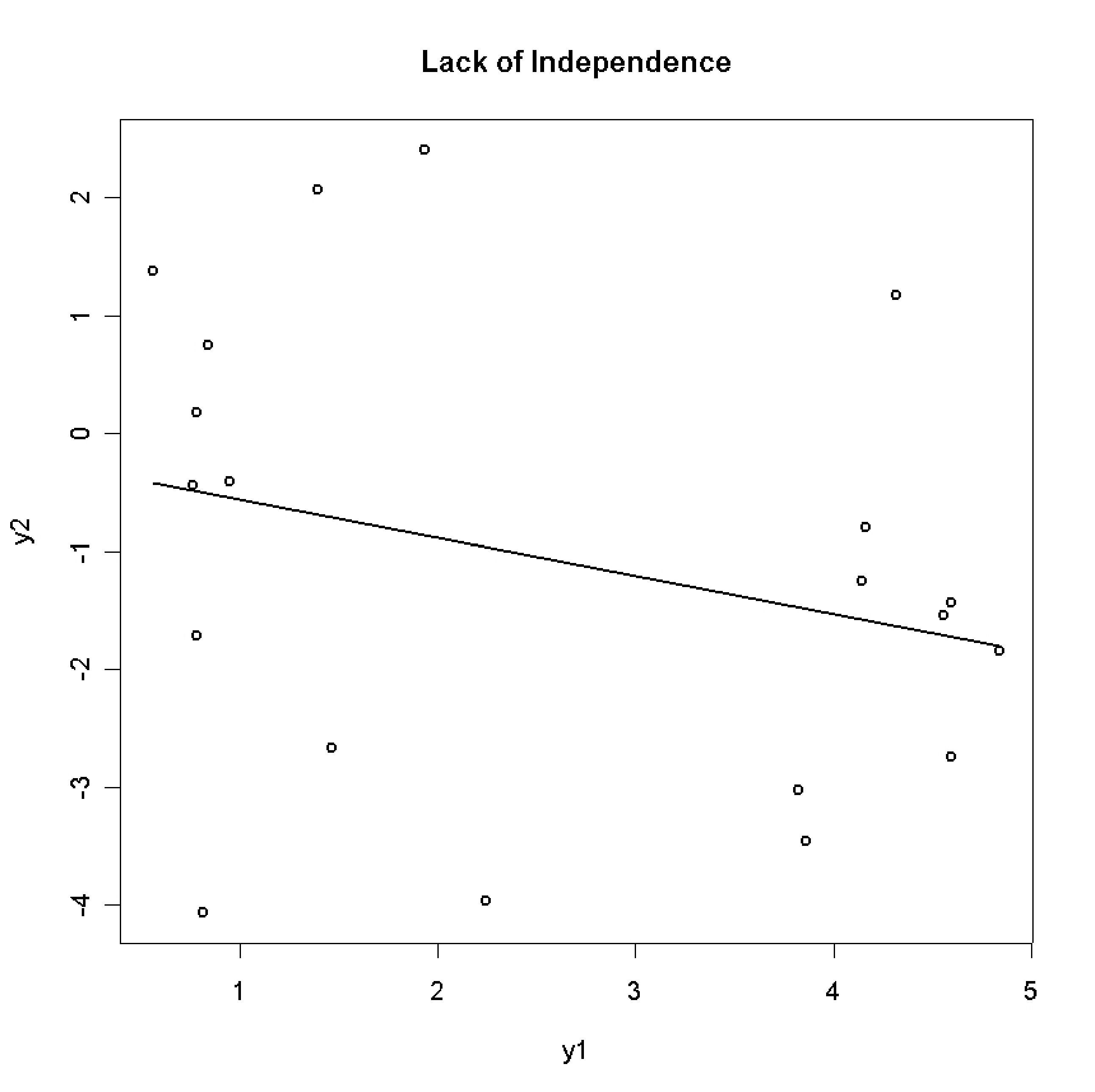

Regression with nonlinearity
y
0
20
40
60
80
100
x
0
2
4
6
8
10

- In general, it is important to get the model right, and the most important of these issues is that the mean function looks like it is specified
- If a linear function does not fit, various types of curves can be used, but what is used should fit the data
- Sometimes, a linear function will fit after a data transformation such as the log or square root
- Otherwise, predictions are biased

Independence

- It is assumed that different observations are statistically independent
- If this is not the case inference and prediction can be completely wrong
- There may appear to be a relationship even though there is not
- Randomization and then controlling the treatment assignment prevents this in general


Lack of Independence
y2
y1
-3
-2
-1
0
1
2
-5
-4
-3
-2
-1
0
1


Lack of Independence
y2
2
1
0
-1
-2
-3
-4
1
2
3
4
5
y1

- Note: no relationship between x and y
- These data were generated as follows:

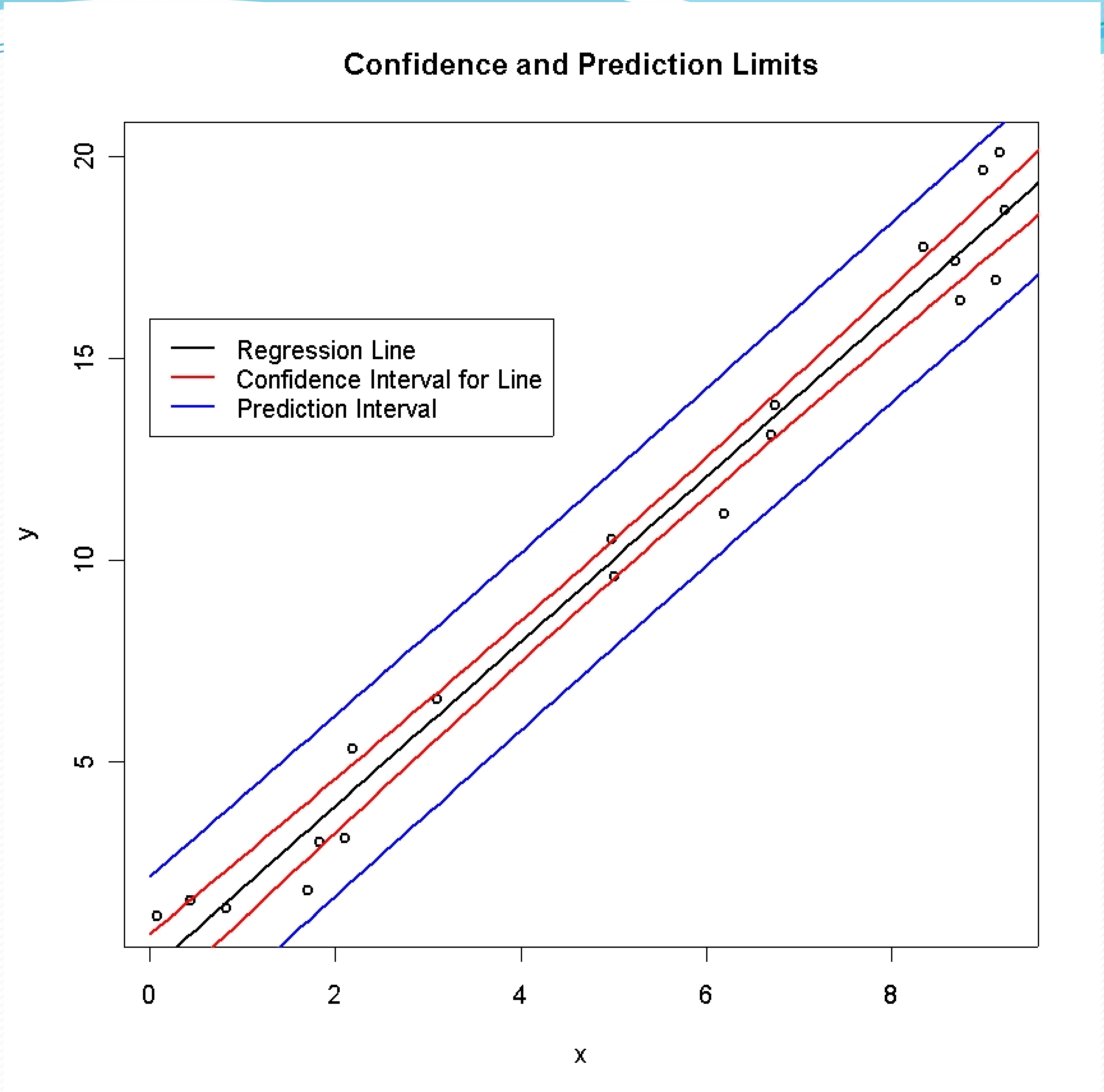
$$x_1 = y_1 = 0$$

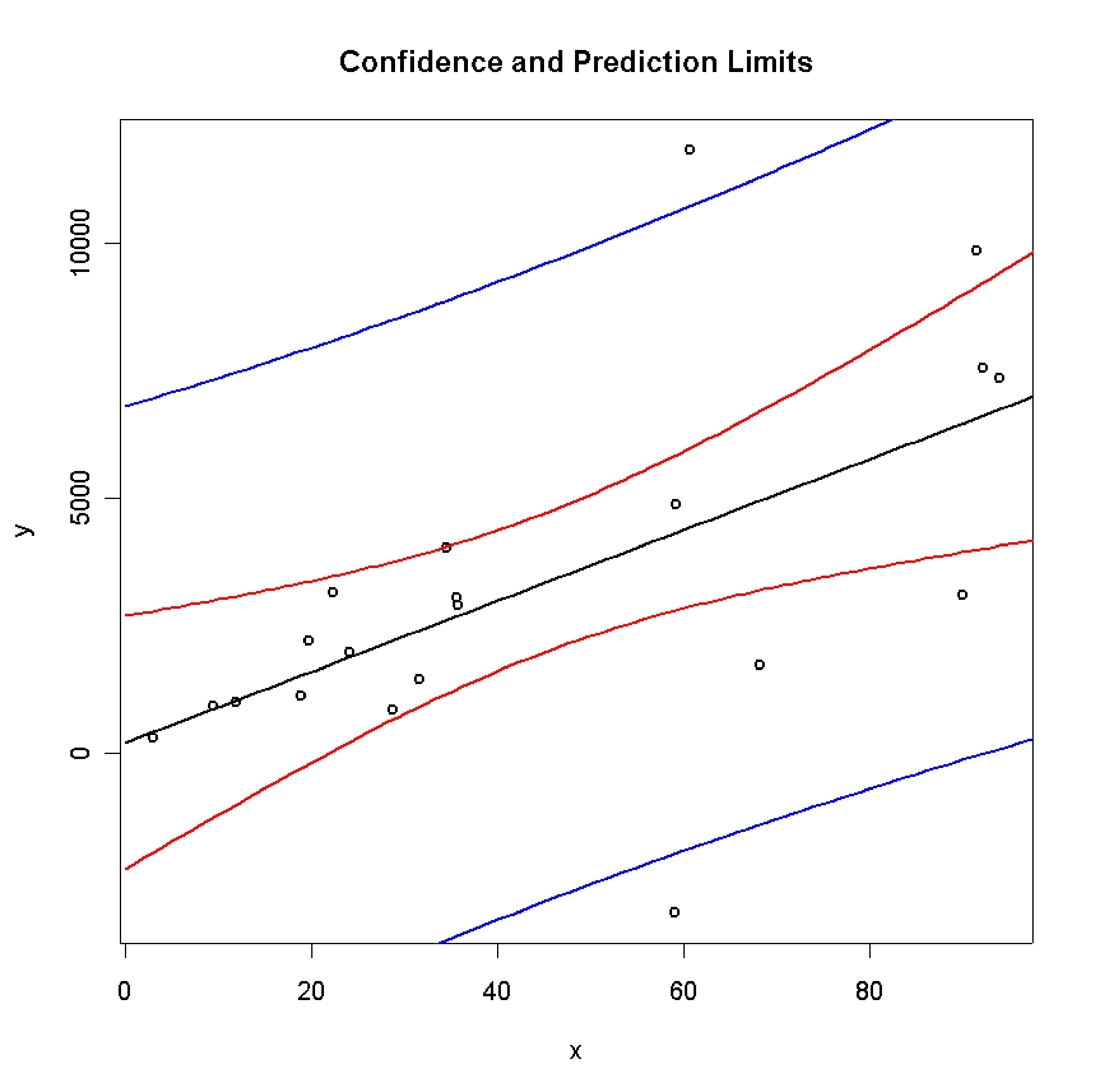
$$x_{i+1} = 0.95x_i + \varepsilon_i$$

$$y_{i+1} = 0.95y_i + \eta_i$$

Constant Variance

- Constant variance, or homoscedacticity, means that the variability is the same in all parts of the prediction function
- If this is not the case, the predictions may be on the average correct, but the uncertainties associated with the predictions will be wrong
- Heteroscedacticity is non-constant variance


Confidence and Prediction Limits
Regression Line
Confidence Interval for Line
Prediction Interval
y
20
15
10
5
x
0
2
4
6
8

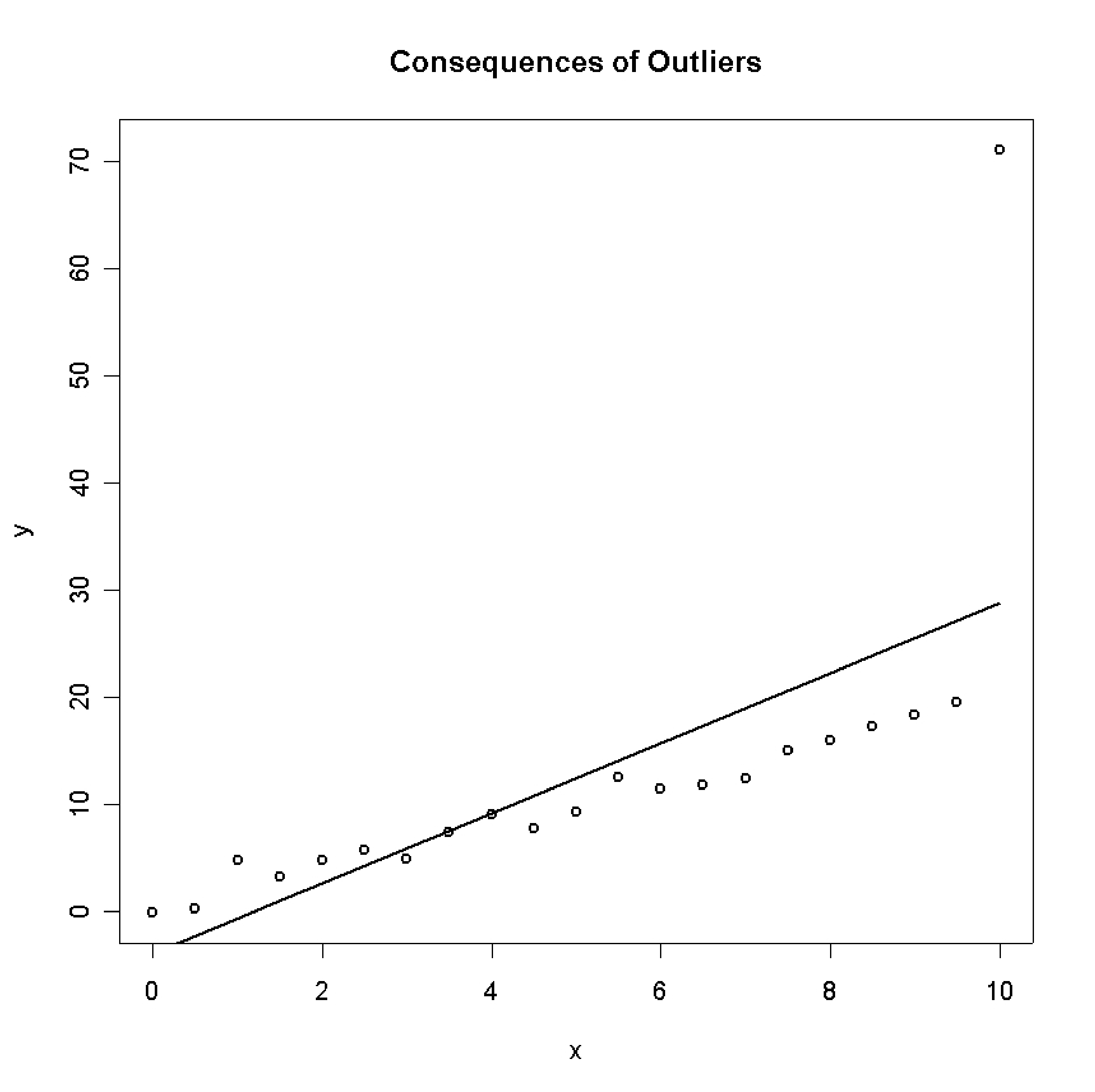

Confidence and Prediction Limits
y
10000
5000
0
x
0
20
40
60
80

Consequences of Heteroscedacticity

- Predictions may be unbiased (correct on the average)
- Prediction uncertainties are not correct; too small sometimes, too large others
- Inferences are incorrect (is there any relationship or is it random?)

Normality of Errors

- Mostly this is not particularly important
- Very large outliers can be problematic
- Graphing data often helps
- If in a gene expression array experiment, we do 40,000 regressions, graphical analysis is not possible
- Significant relationships should be examined in detail


Consequences of Outliers
y
0
10
20
30
40
50
60
70
x
0
2
4
6
8
10

Statistical Lab Books

- You should keep track of what things you try
- The eventual analysis is best recorded in a file of commands so it can later be replicated
- Plots should also be produced this way, at least in final form, and not done on the fly
- Otherwise, when the paper comes back for review, you may not even be able to reproduce your own analysis

Fluorescein Example

- Standard aqueous solutions of fluorescein (in pg/ml) are examined in a fluorescence spectrometer and the intensity (arbitrary units) is recorded
- What is the relationship of intensity to concentration
- Use later to infer concentration of labeled analyte

Concentration (pg/ml)	0	2	4	6	8	10	12
Intensity	2.1	5.0	9.0	12.6	17.3	21.0	24.7

```
> fluor.lm <- lm(intensity ~ concentration,data=fluor)
> summary(fluor.lm)

Call:
lm(formula = intensity ~ concentration)

Residuals:
       1        2        3        4        5        6        7
 0.58214 -0.37857 -0.23929 -0.50000  0.33929  0.17857  0.01786

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.5179     0.2949   5.146  0.00363 **
concentration   1.9304     0.0409  47.197 8.07e-08 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.4328 on 5 degrees of freedom
Multiple R-Squared: 0.9978,     Adjusted R-squared: 0.9973
F-statistic:  2228 on 1 and 5 DF,  p-value: 8.066e-08
```

Use of the calibration curve

$\hat{y} = 1.52 + 1.93x$

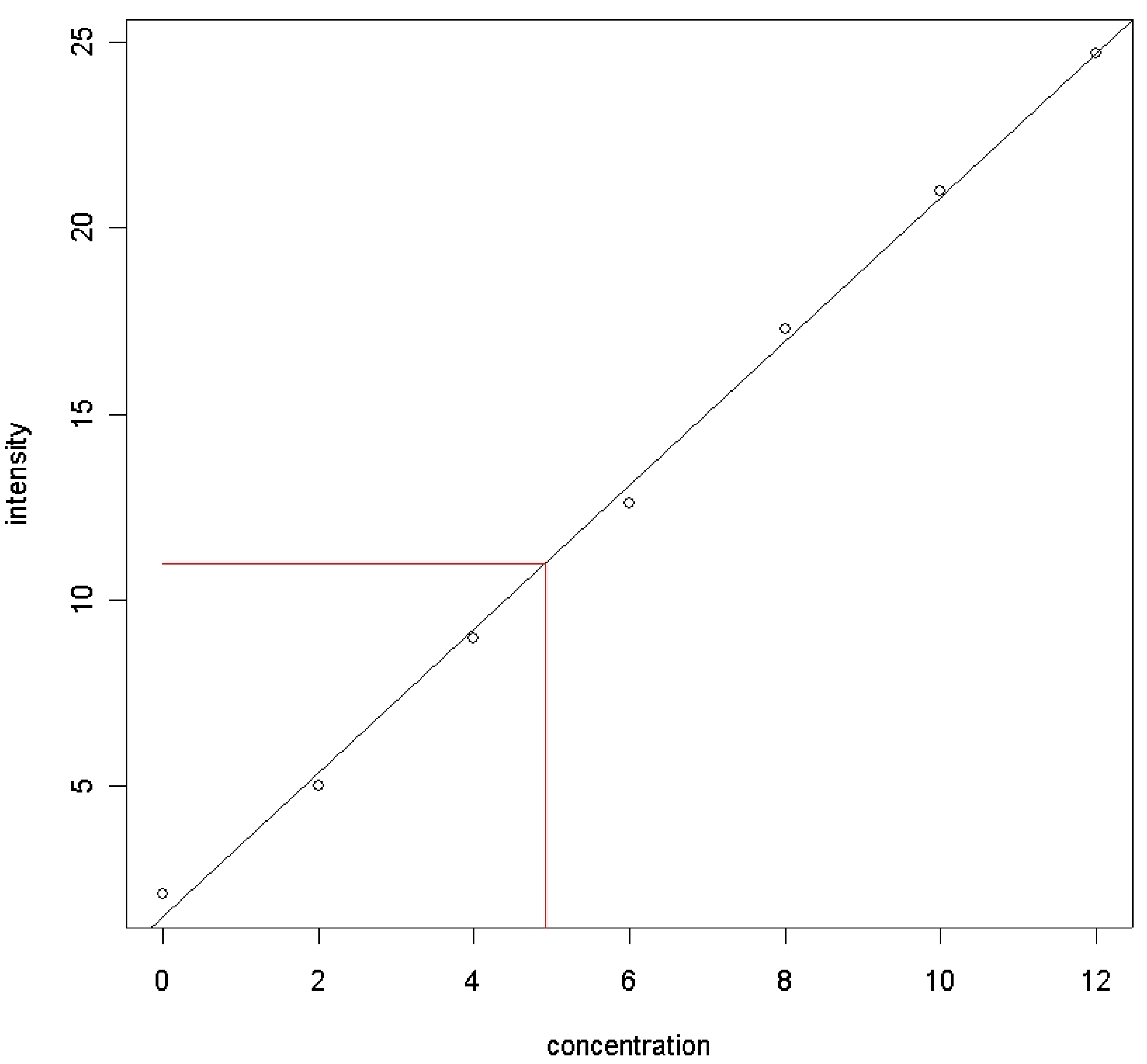
$\hat{y}$ is the predicted average intensity

x is the true concentration

$$\hat{x} = \frac{y - 1.52}{1.93}$$

y is the observed intensity

$\hat{x}$ is the estimated concentration


intensity
25
20
15
10
5
0
2
4
6
8
10
12
concentration

Measurement and Calibration

- Essentially all things we measure are indirect
- The thing we wish to measure produces an observed transduced value that is related to the quantity of interest but is not itself directly the quantity of interest
- Calibration takes known quantities, observes the transduced values, and uses the inferred relationship to quantitate unknowns

Measurement Examples

- Weight is observed via deflection of a spring (calibrated)
- Concentration of an analyte in mass spec is observed through the electrical current integrated over a peak (possibly calibrated)
- Gene expression is observed via fluorescence of a spot to which the analyte has bound (usually not calibrated) or by counting RNA fragments that map to a given gene (also not usually calibrated)

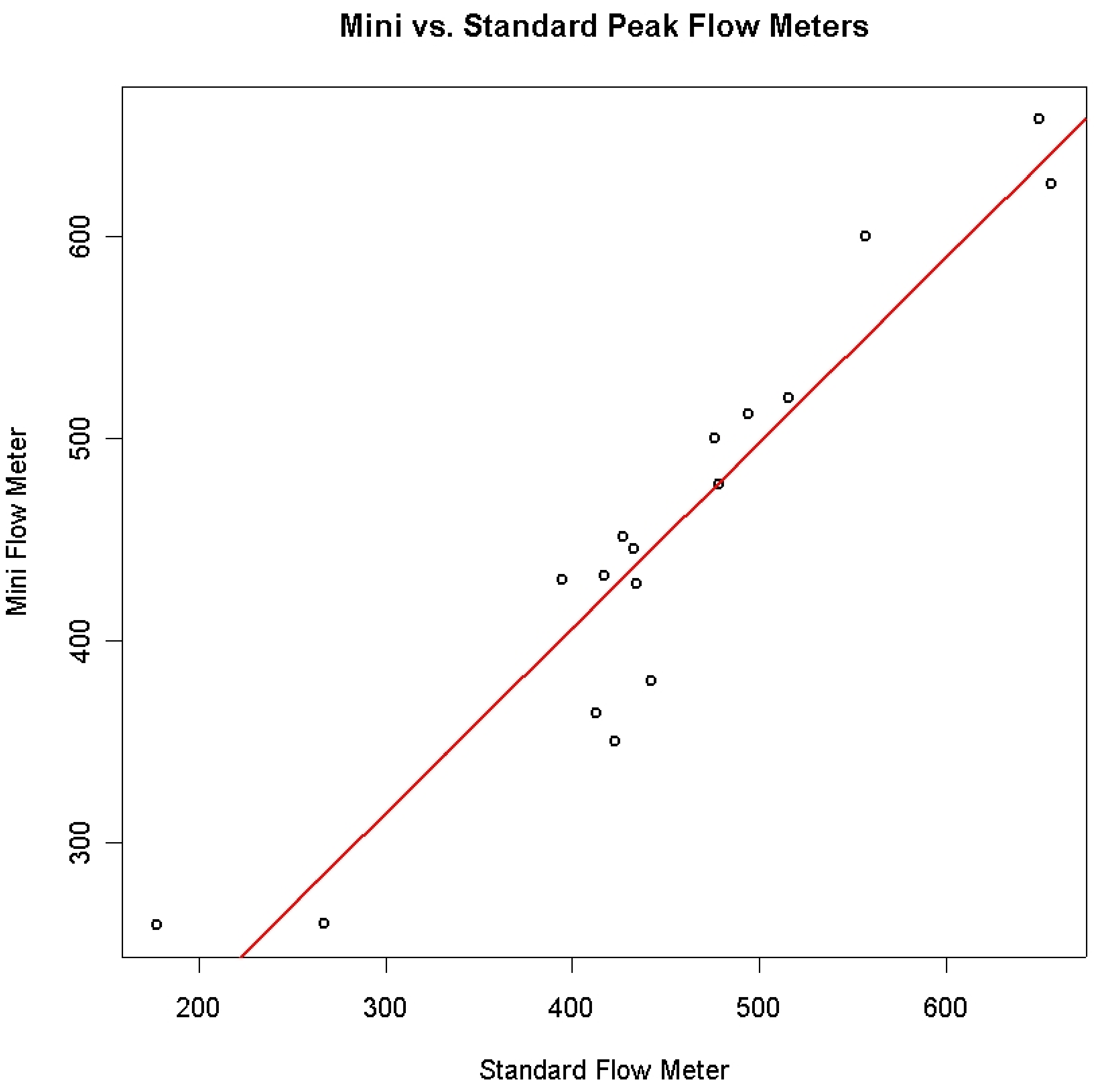
Correlation

- Wright peak-flow data set has two measures of peak expiratory flow rate for each of 17 patients in l/min.
- ISwR library, data(wright)
- Both are subject to measurement error
- In ordinary regression, we assume the predictor is known
- For two measures of the same thing with no error-free gold standard, one can use correlation to measure agreement

```
> source(“wright.r”)
> cor(wright)
            std.wright mini.wright
std.wright   1.0000000   0.9432794
mini.wright  0.9432794   1.0000000

> wplot1()
-----------------------------------------------------
```

File wright.r:

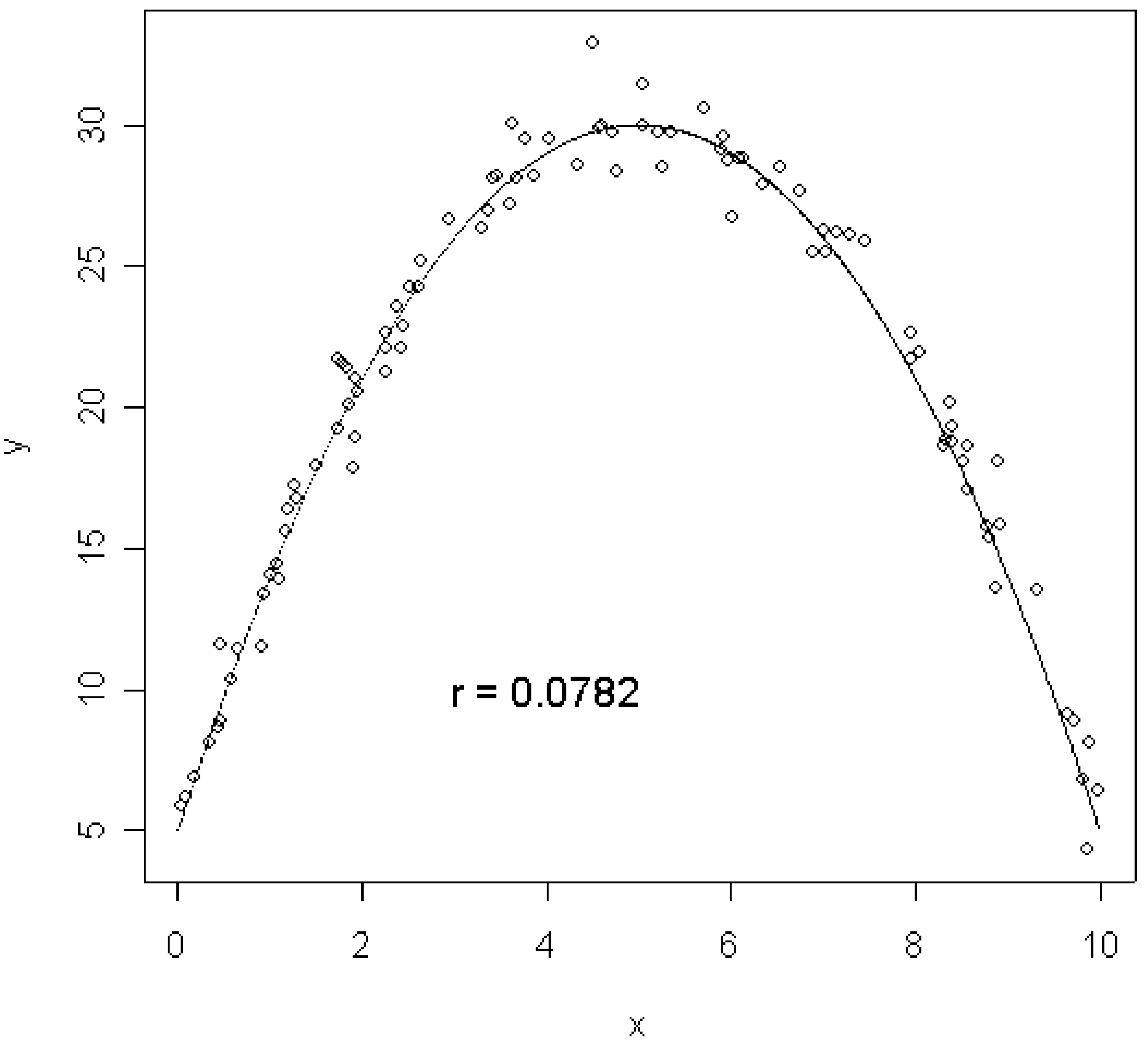
```
library(ISwR)
data(wright)

wplot1 <- function()
{
  with(wright,plot(std.wright,mini.wright,
     xlab="Standard Flow Meter",ylab="Mini Flow Meter",lwd=2))
  title("Mini vs. Standard Peak Flow Meters")
  wright.lm <- lm(mini.wright ~ std.wright,data=wright)
  abline(coef(wright.lm),col="red",lwd=2)
}
```


Mini vs. Standard Peak Flow Meters
Mini Flow Meter
300
400
500
600
200
300
400
500
600
Standard Flow Meter

Issues with Correlation

- For any given relationship between two measurement devices, the correlation will depend on the range over which the devices are compared. If we restrict the Wright data to the range 300-550, the correlation falls from 0.94 to 0.77.
- Correlation only measures linear agreement

A strong nonlinear relationship with low correlation



Measurement with no Gold Standard

$y_{1j} = a + b\xi_j + \epsilon_j$

$y_{2j} = \xi_j$

ξ_j is the true concentration

Method 2 is the gold standard, measured without error

We can estimate all the unknowns, including σ_ϵ^2

$y_{1j} = a + b\xi_j + \epsilon_j$

$y_{2j} = \xi_j + \eta_j$

ξ_j is the true unknown concentration

We have one more unknown (σ_η^2) but no additional data

Cannot be solved without information/assumptions