Regression, Calibration and Analytical Error

BST 226
Statistical Methods for Bioinformatics
David M. Rocke

Quantitative Prediction

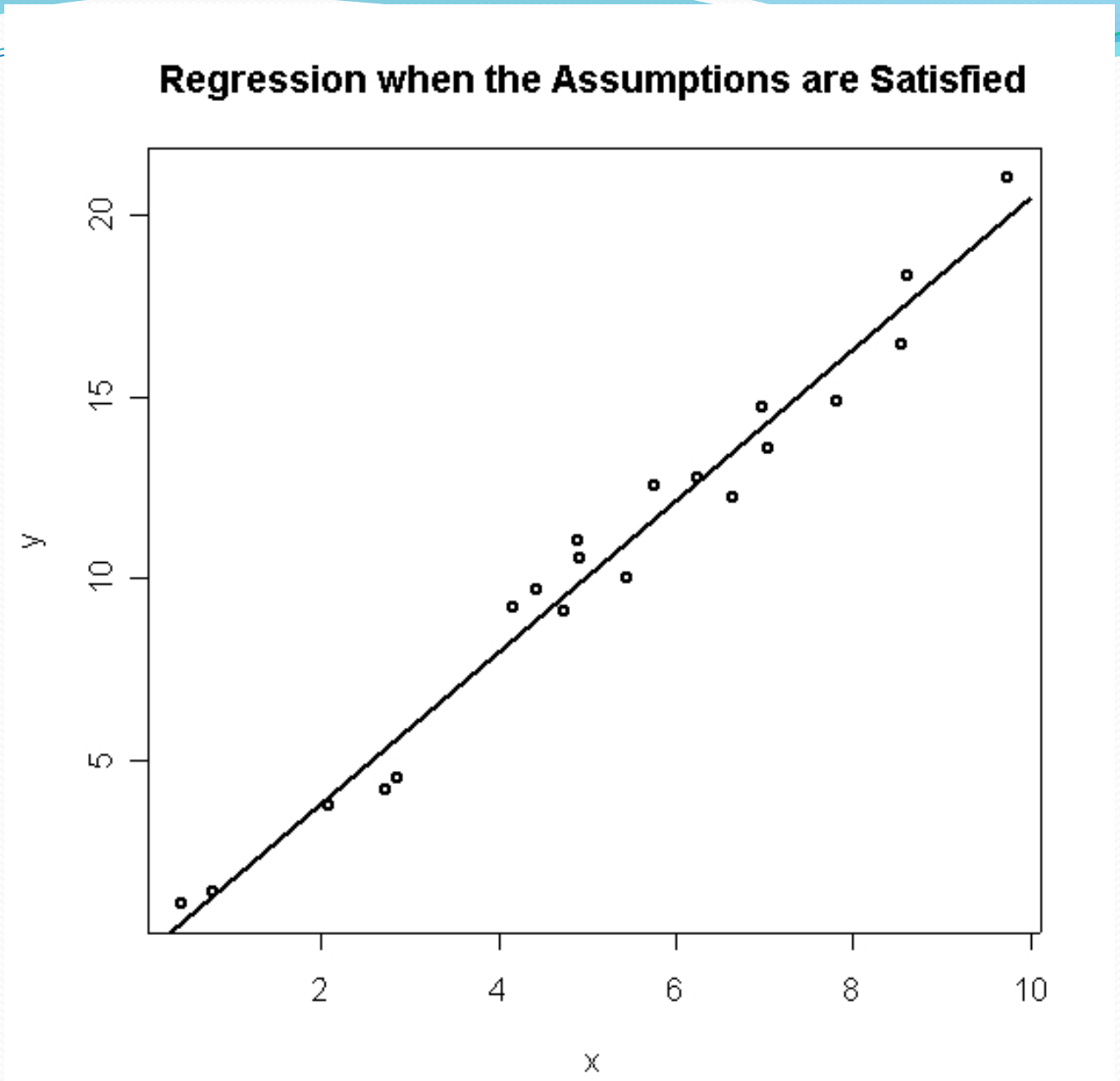
- Regression analysis is the statistical name for the prediction of one quantitative variable (fasting blood glucose level) from another (body mass index)
- Items of interest include whether there is in fact a relationship and what the expected change is in one variable when the other changes

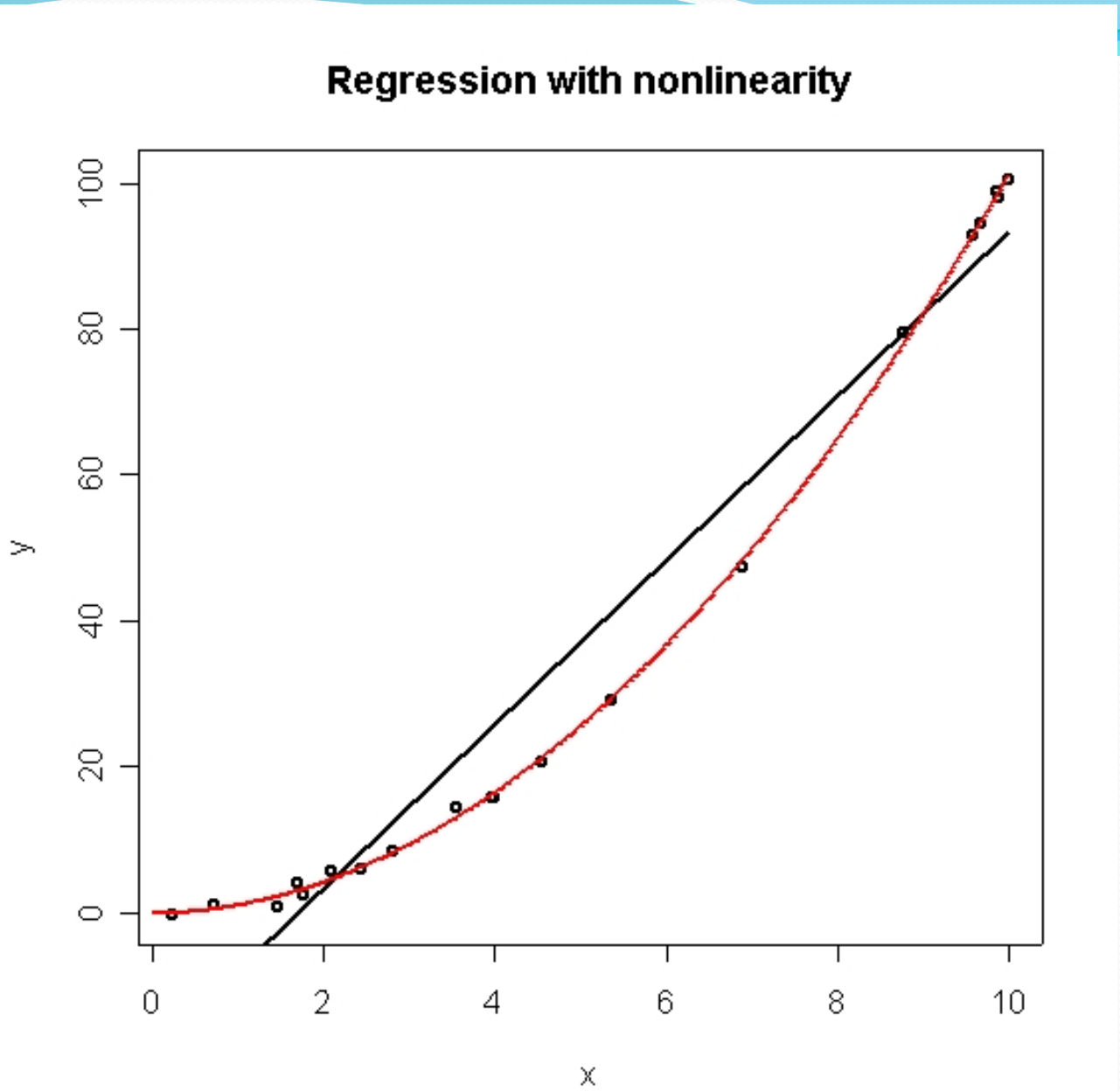
Assumptions

- Inference about whether there is a real relationship or not is dependent on a number of assumptions, many of which can be checked
- When these assumptions are substantially incorrect, alterations in method can 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

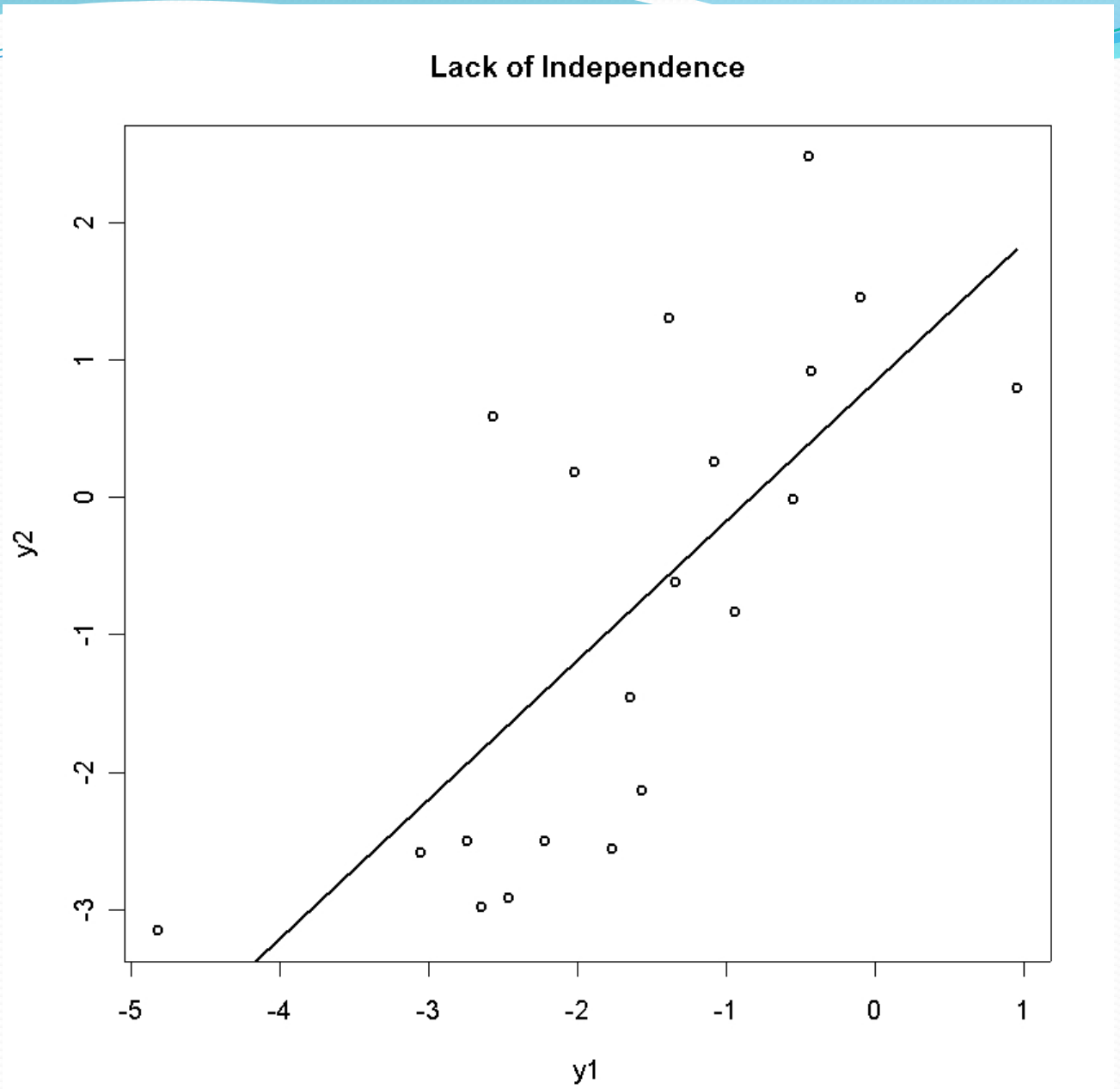

Regression when the Assumptions are Satisfied
y
x
5
10
15
20
2
4
6
8
10

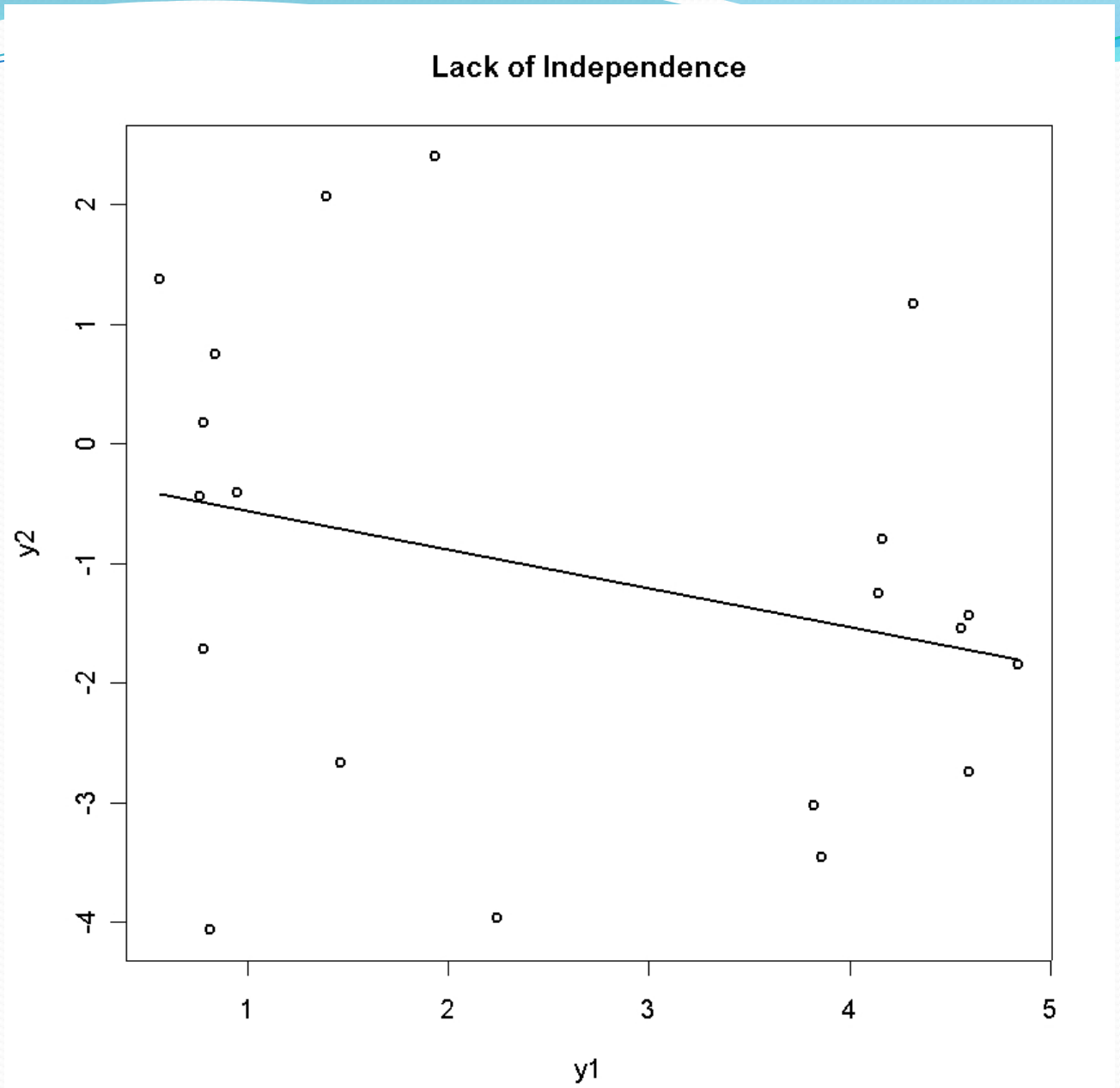

Regression with nonlinearity
y
x
0
20
40
60
80
100
0
2
4
6
8
10

- 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
- 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
2
1
0
-1
-2
-3
-5
-4
-3
-2
-1
0
1
y1


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

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

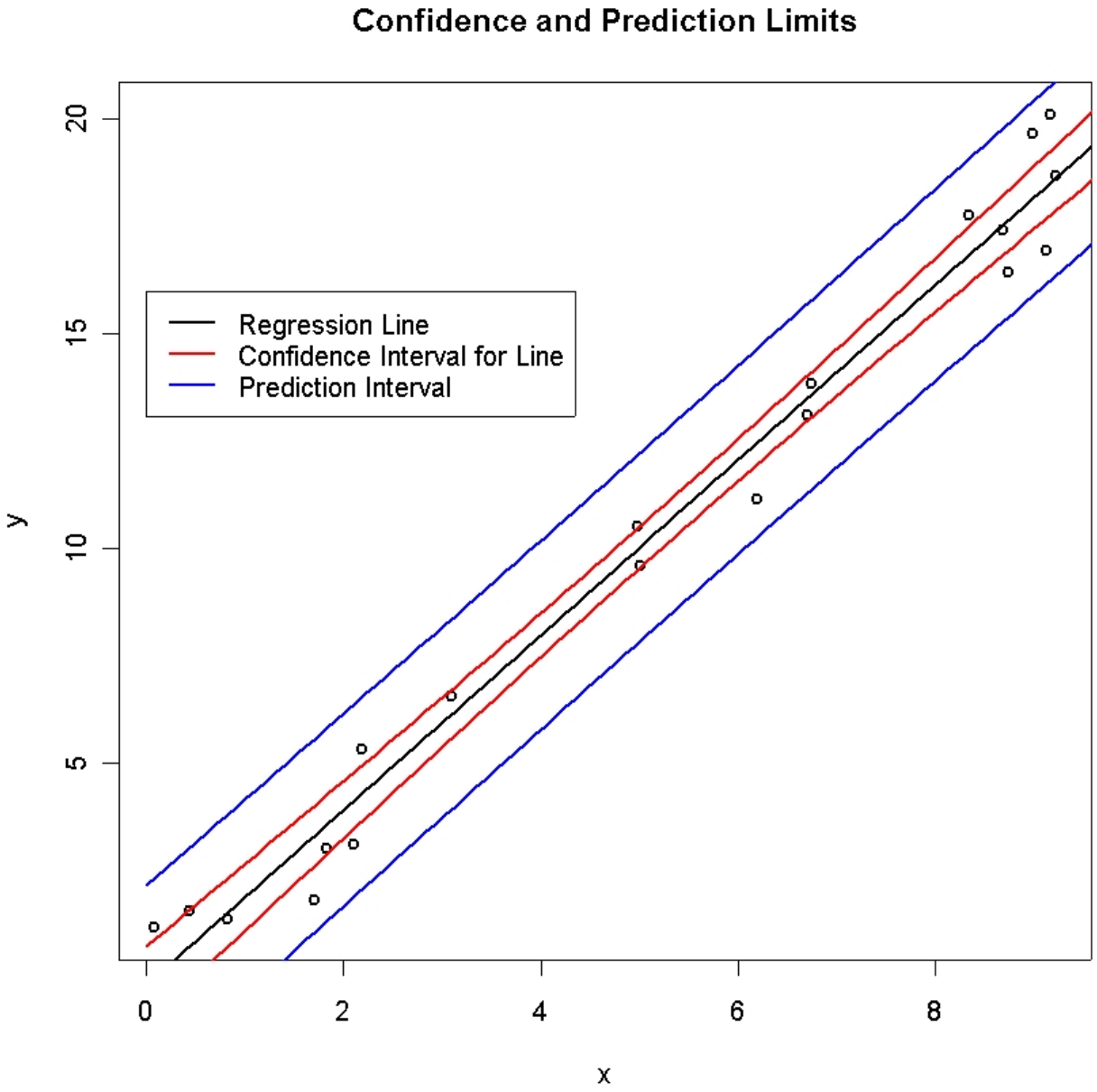
$$x_1 = y_1 = 0$$

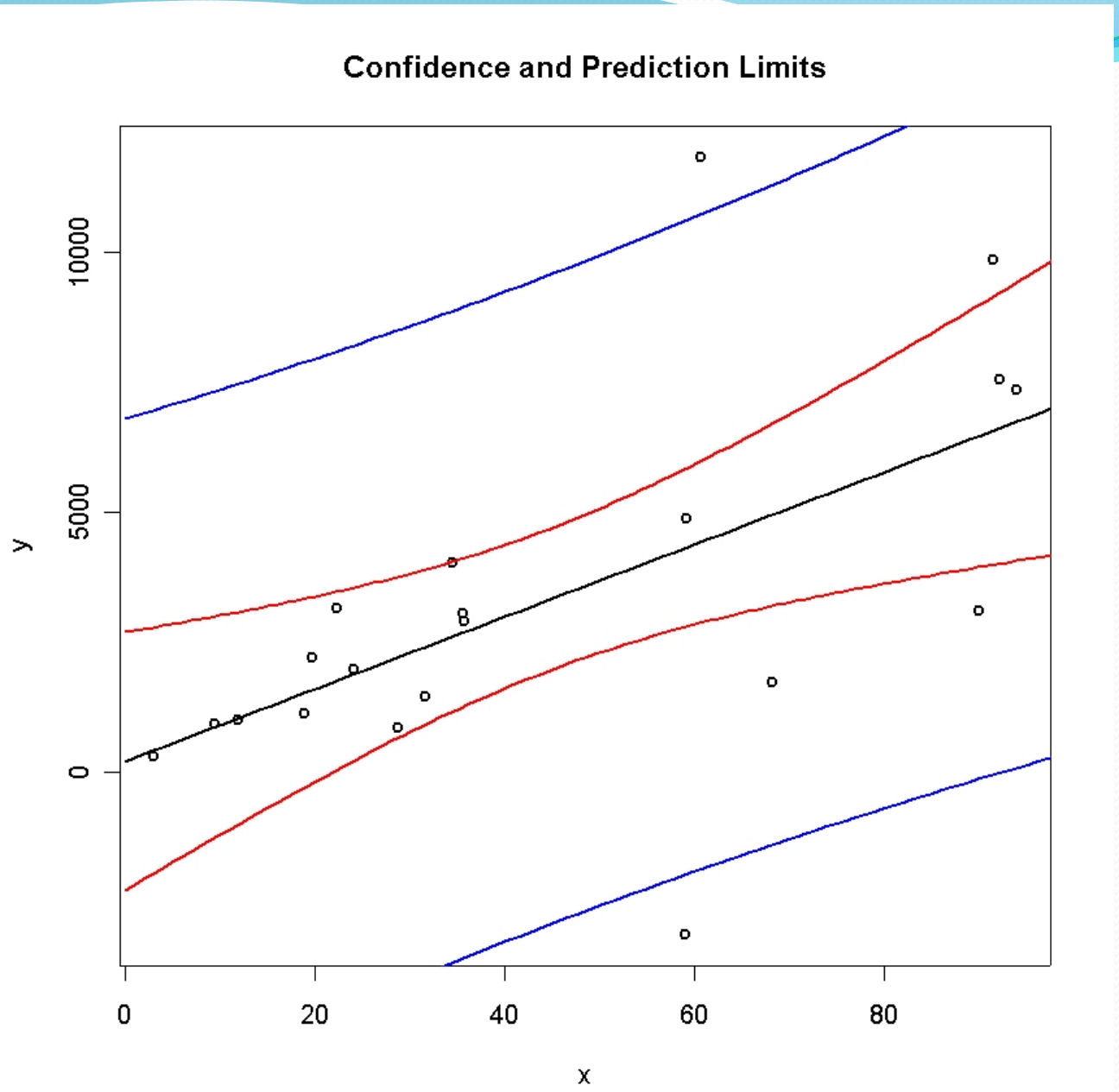
$$x_{i+1} = 0.95x_i + \varepsilon_i$$

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

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
x
0
2
4
6
8
5
10
15
20

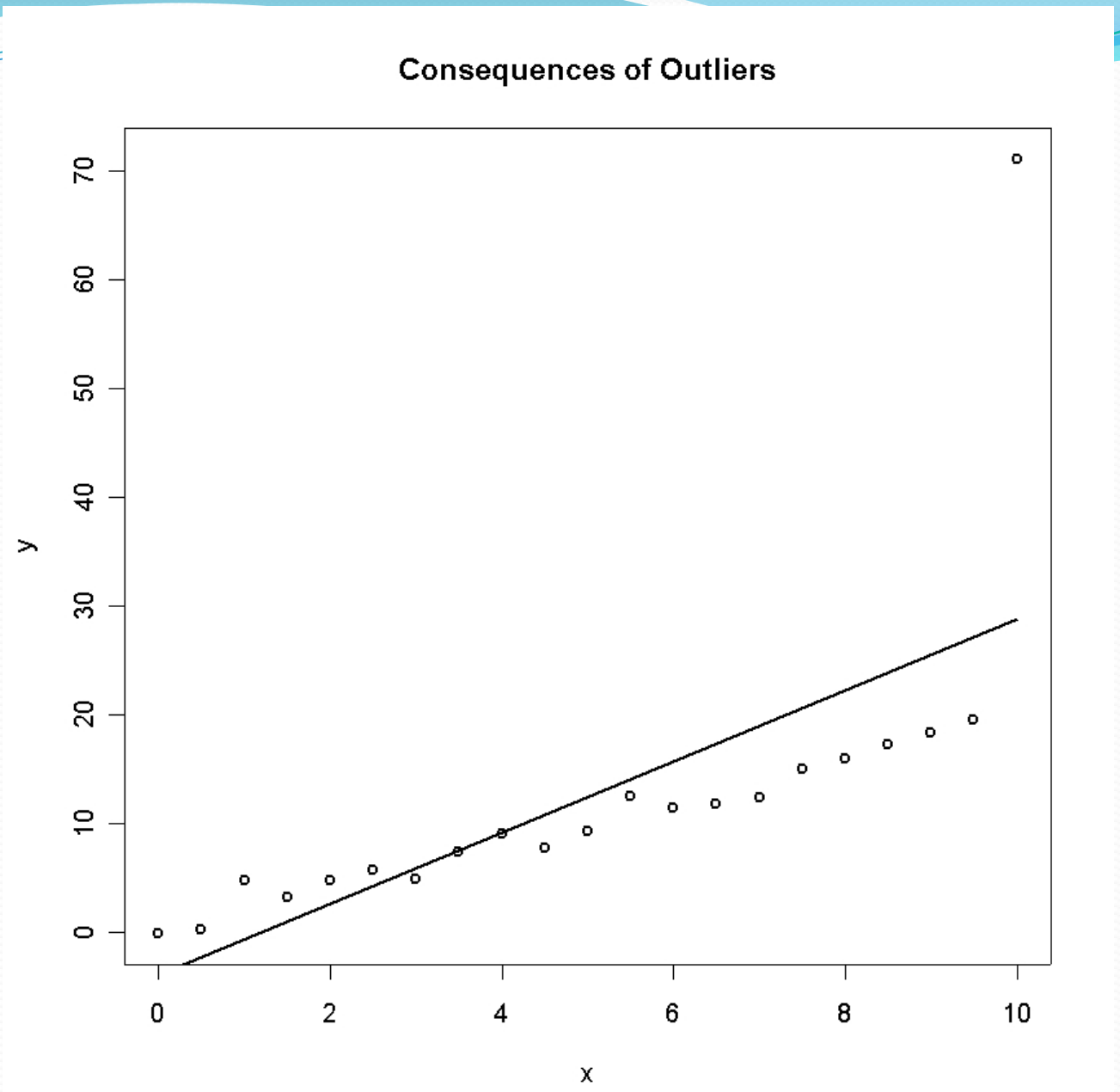

Confidence and Prediction Limits
y
10000
5000
0
0
20
40
60
80
x

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
x
0
10
20
30
40
50
60
70
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)
> 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

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

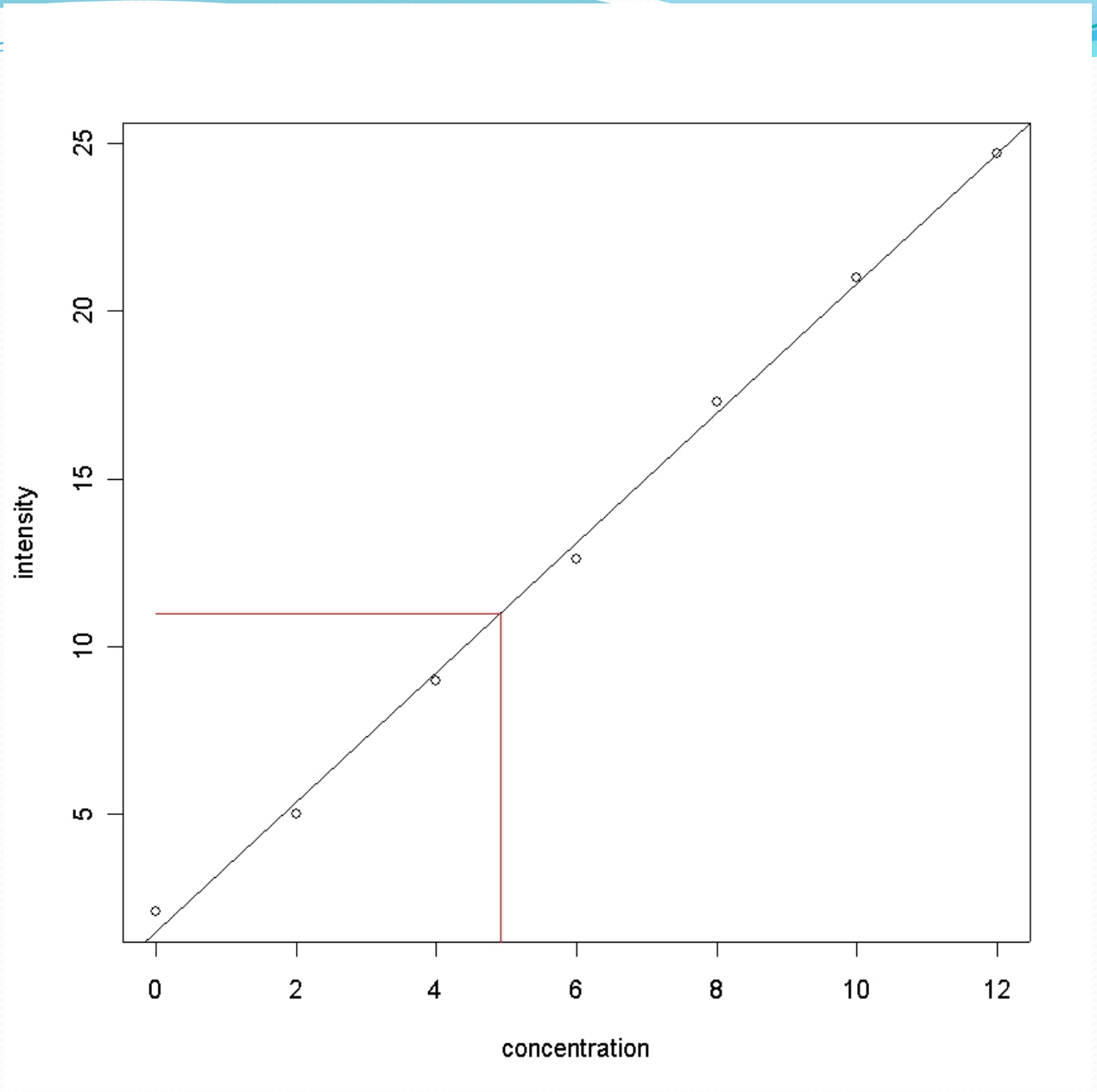
$\hat{y}$ is the predicted average intensity

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
5
10
15
20
25
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)

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

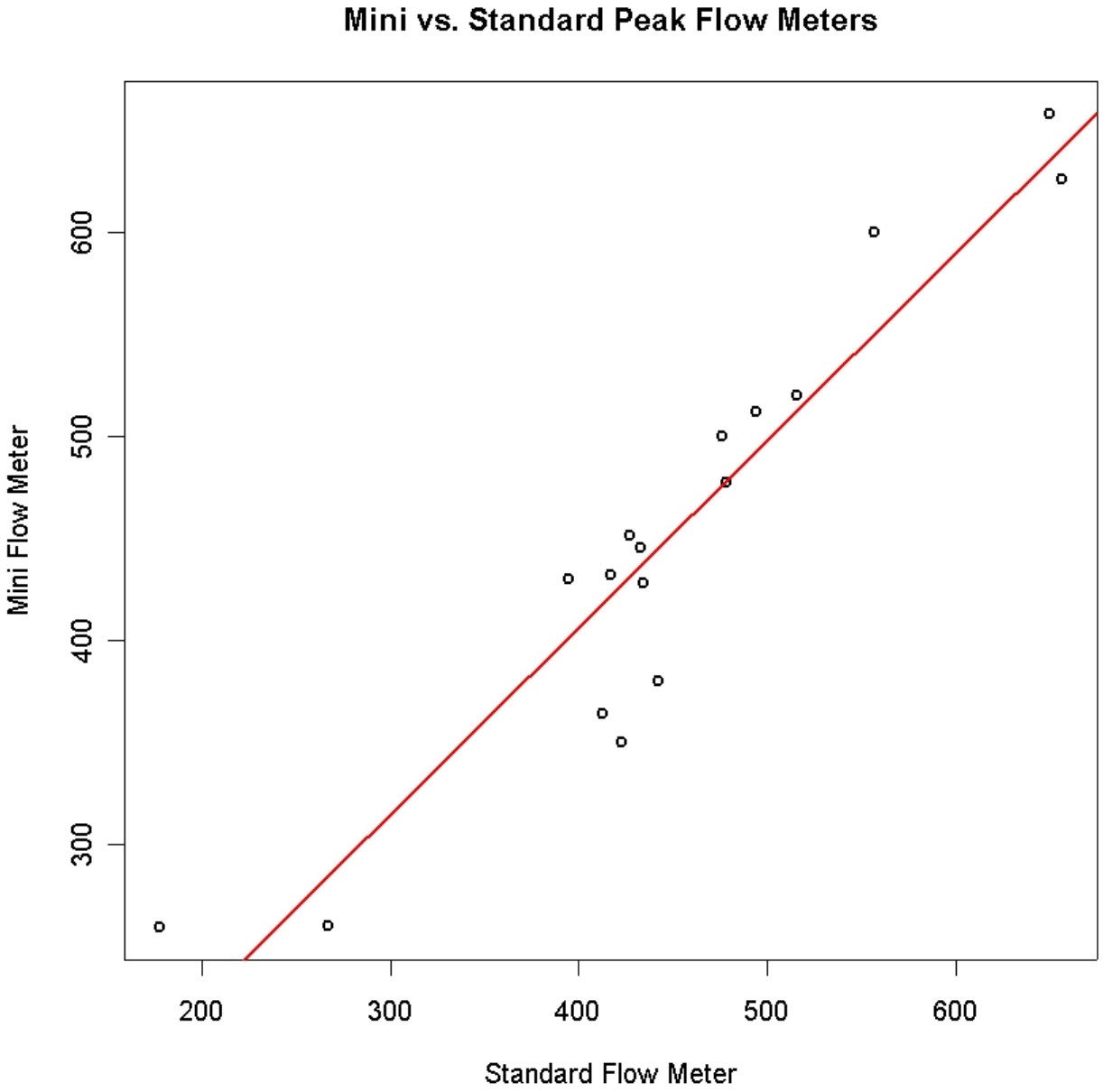
```
> cor(wright)
             std.wright mini.wright
std.wright   1.0000000   0.9432794
mini.wright  0.9432794   1.0000000

> wplot1()
```

File wright.r:

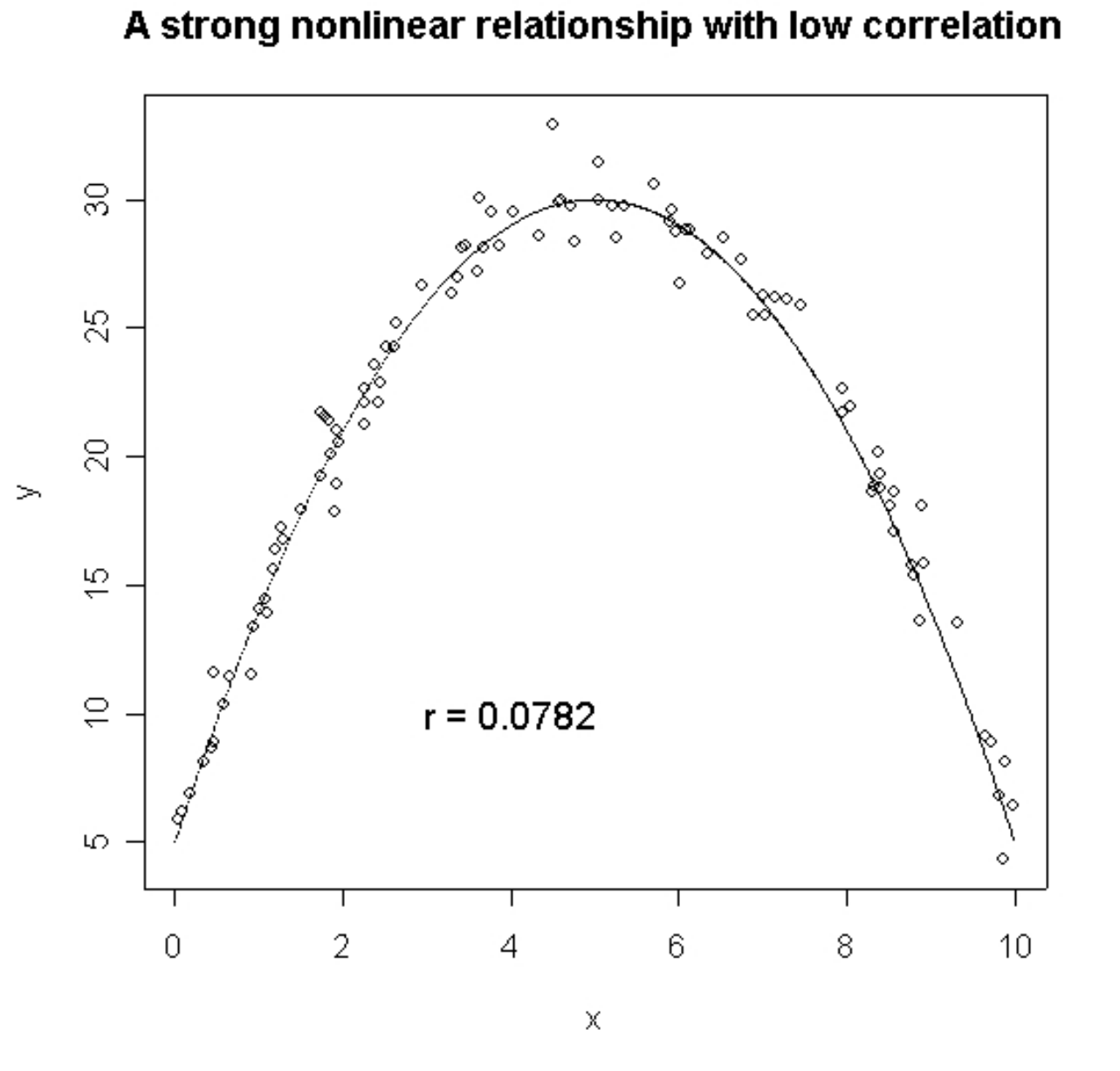
```
library(ISwR)
data(wright)
attach(wright)

wplot1 <- function()
{
  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)
  abline(coef(wright.lm),col="red",lwd=2)
}
detach(wright)
```


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
r = 0.0782
y
x
5
10
15
20
25
30
0
2
4
6
8
10

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

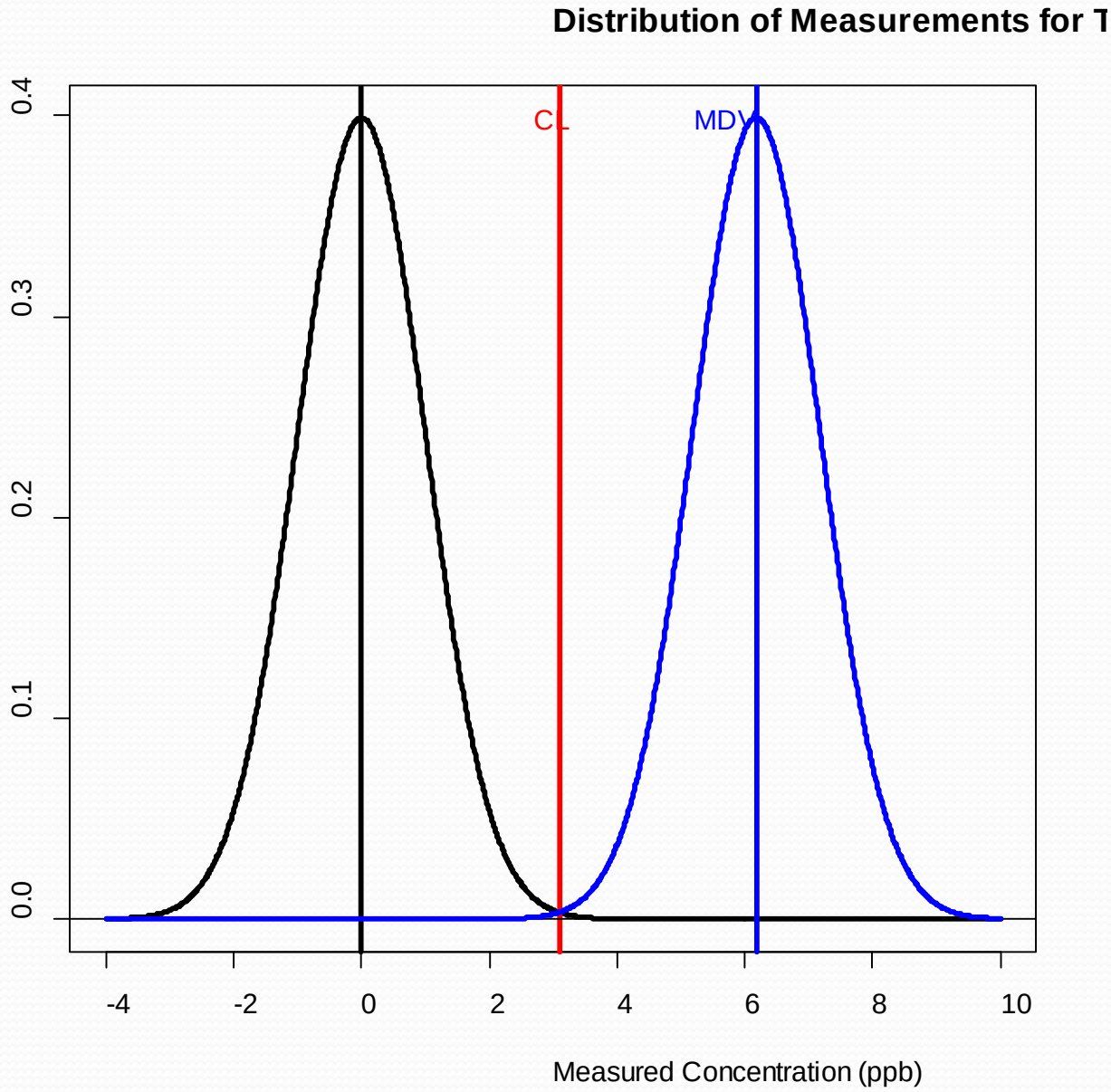
Cannot be solved without information/assumptions

Calibrated Measurements

- We produce a calibration curve of the form $y = f(x)$, where x is the concentration of the analyte and y is the transduced value, such as peak height or peak area in mass spectrometry.
- Often the curve is linear.
- It is estimated from measurements at a series of known concentrations and their responses.
- A new measurement y produces an estimated concentration using $x = f^{-1}(y)$.

Limits of Detection

- The term "limit of detection" is actually ambiguous and can mean various things that are often not distinguished from each other
- We will instead define three concepts that are all used in this context called the *critical level*, the *minimum detectable value*, and the *limit of quantitation*.
- The *critical level* is the measurement that is not consistent with the analyte being absent
- The *minimum detectable value* is the concentration that will almost always have a measurement above the critical level


Distribution of Measurements for T
CL
MDV
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
6
8
10
Measured Concentration (ppb)

$y = x + \epsilon$

$\epsilon \sim N(0, \sigma_\epsilon^2)$

If the concentration is zero, then

$y \sim N(0, \sigma_\epsilon^2)$

Negative measured values are possible and informative.

If $\sigma_\epsilon = 1$ppb, and if we want to use the 99.9% point of the normal, which is 3.090232, then the critical level is

$0 + (1)(3.090232) = 3.090232$

which is a measured level. If the reading is CL = 3.090232ppb or greater, then we are confident that there is more than 0 of the analyte.

If the true concentration is 3.090232, then about half the time the measured concentration is less than the CL.

If the true concentration is MDV = $3.090232 + 3.090232 = 6.180464$ppb, then 99.9% of the time the measured value will exceed the CL.

$$y = x + \epsilon$$

$$z_\alpha : \Pr(z > z_\alpha) = \alpha$$

$$z \sim N(0,1)$$

$$CL(\alpha) : 0 + z_\alpha \sigma_\epsilon$$

$$MDV(\alpha, \beta) : (0 + z_\alpha \sigma_\epsilon) + z_\beta \sigma_\epsilon$$

CI for x:

$$y \pm z_{\alpha/2} \sigma_\epsilon$$

Assuming σ_ϵ is known and constant.

Examples

- Serum calcium usually lies in the range 8.5–10.5 mg/dl or 2.2–2.7 mmol/L.
- Suppose the standard deviation of repeat measurements is 0.15 mmol/L.
- Using $\alpha = 0.01$, $z_\alpha = 2.326$, so the critical level is $(2.326)(0.15) = 0.35$ mmol/L.
- The MDV is 0.70 mg/L, well out of the physiological range.

- A test for toluene exposure uses GC/MS to test serum samples.
- The standard deviation of repeat measurements of the same serum at low levels of toluene is 0.03 µg/L.
- The critical level at $\alpha = 0.01$ is $(0.03)(2.326) = 0.070$ µg/L.
- The MDV is 0.140 µg/L.
- Unexposed non-smokers 0.4 µg/L
- Unexposed smokers 0.6 µg/L
- Chemical workers 2.8 µg/L
- One EPA standard is < 1 mg/L blood concentration.
- Toluene abusers may have levels of 0.3–30 mg/L and fatalities have been observed at 10–48 mg/L

- The EPA has determined that there is no safe level of dioxin (2,3,7,8-TCDD (tetrachlorodibenzodioxin)), so the Maximum Contaminant Level Goal (MCLG) is 0.
- The Maximum Contaminant Level (MCL) is based on the best existing analytical technology and was set at 30 ppq.
- EPA Method 1613 uses high-resolution GC/MS and has a standard deviation at low levels of 1.2 ppq.
- The critical level at 1% is (2.326)(1.2ppq) = 2.8 ppq and the MDV, called the Method Detection Limit by EPA, is 5.6 ppq.
- 1 ppq = 1pg/L = 1gm in a square lake 1 meter deep and 10 km on a side.
- The reason why the MCL is set at 30 ppq instead of 2.8 ppq will be addressed later.

Error Behavior at High Levels

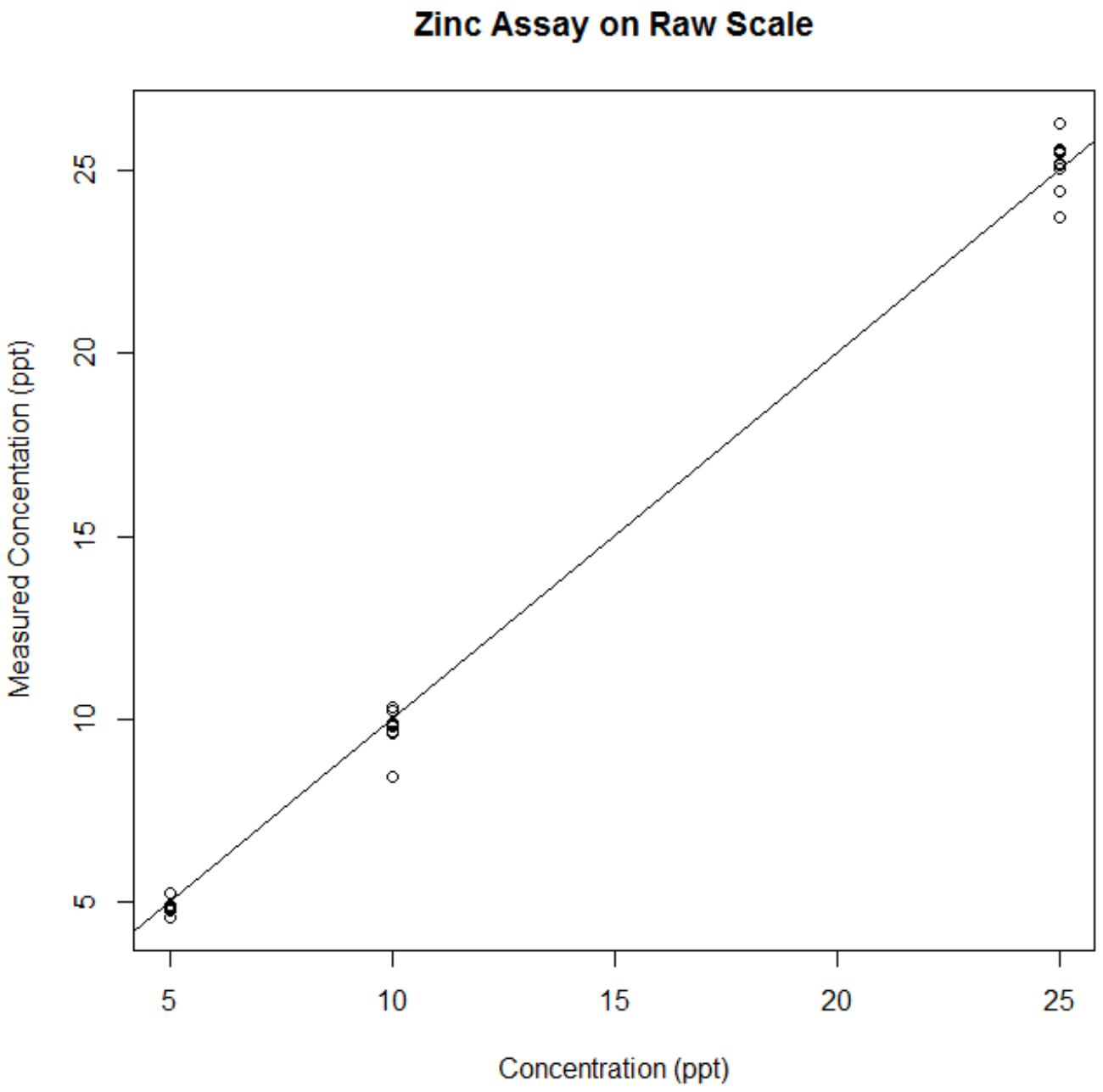
- For most analytical methods, when the measurements are well above the CL, the standard deviation is a constant multiple of the concentration
- The ratio of the standard deviation to the mean is called the coefficient of variation (CV), and is often expressed in percents.
- For example, an analytical method may have a CV of 10%, so when the mean is 100 mg/L, the standard deviation is 10 mg/L.
- When a measurement has constant CV, the log of the measurement has approximately constant standard deviation.
- If we use the natural log, then SD on the log scale is approximately CV on the raw scale

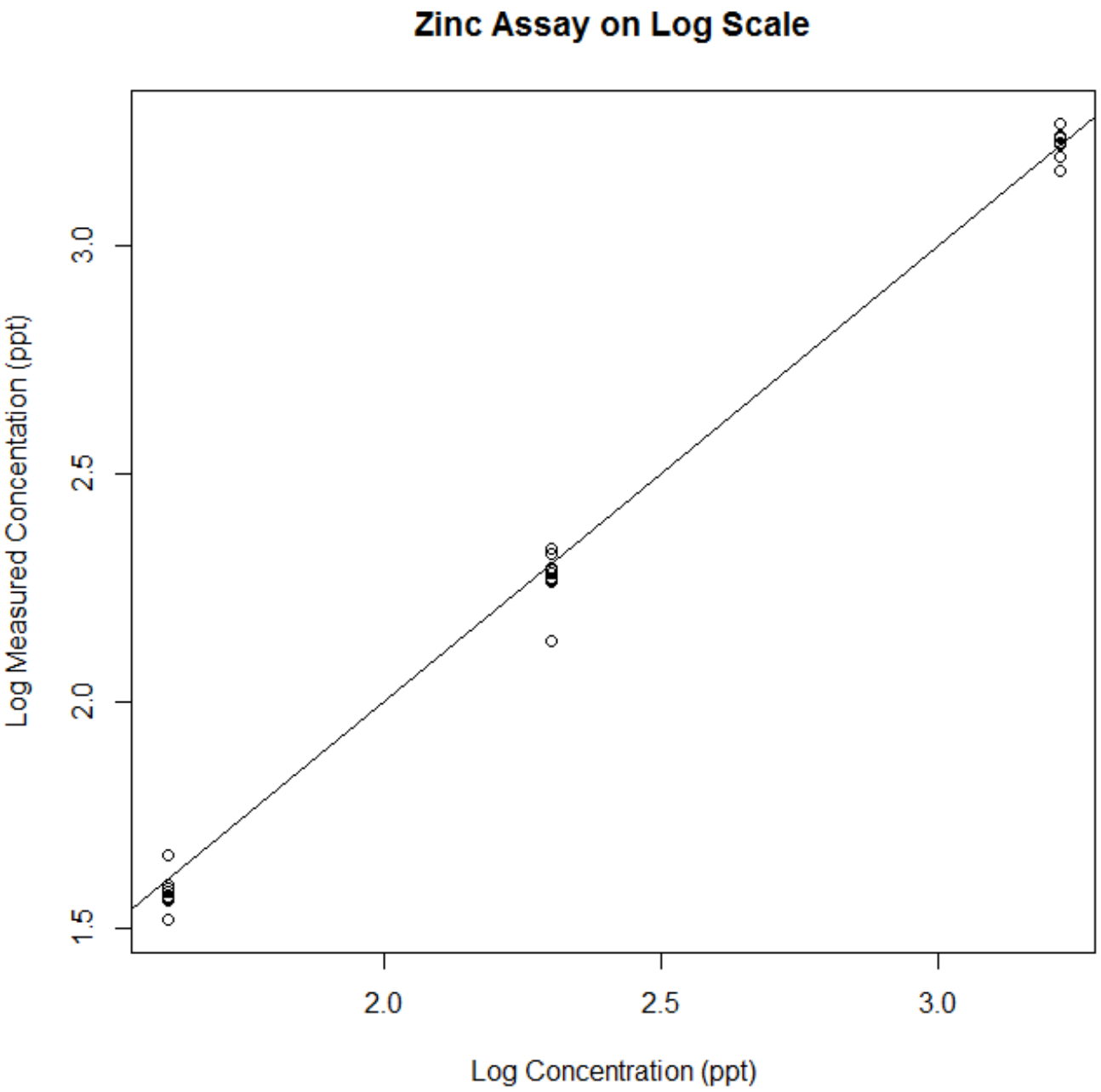
Zinc Concentration

- Spikes at 5, 10, and 25 ppb
- 9 or 10 replicates at each concentration
- Mean measured values, SD, and CV are below

Raw	5	10	25
Mean	4.85	9.73	25.14
SD	0.189	0.511	0.739
CV	0.039	0.053	0.029

Log	1.61	2.30	3.22
Mean	1.58	2.27	3.22
SD	0.038	0.055	0.030


Zinc Assay on Raw Scale
Measured Concentation (ppt)
Concentration (ppt)
5
10
15
20
25


Zinc Assay on Log Scale
Log Measured Concentation (ppt)
1.5
2.0
2.5
3.0
2.0
2.5
3.0
Log Concentration (ppt)

Summary

- At low levels, assays tend to have roughly constant variance not depending on the mean. This may hold up to the MDV or somewhat higher. For low level data, analyze the raw data.
- At high levels, assays tend to have roughly constant CV, so that the variance is roughly constant on the log scale. For high level data, analyze the logs.
- We run into trouble with data sets where the analyte concentrations vary from quite high to very low
- This is a characteristic of many gene expression, proteomics, and metabolomics data sets.

The two-component model

- The two-component model treats assay data as having two sources of error, an additive error that represents machine noise and the like, and a multiplicative error.
- When the concentration is low, the additive error dominates.
- When the concentration is high, the multiplicative error dominates.
- There are transformations similar to the log that can be used here.

$$y = xe^{\eta} + \epsilon$$

$$V(y) = x^2 V(e^{\eta}) + \sigma_{\epsilon}^2$$

$$= x^2 e^{\sigma_{\eta}^2}(e^{\sigma_{\eta}^2} - 1) + \sigma_{\epsilon}^2$$

$$\sim x^2 \sigma_{\eta}^2 + \sigma_{\epsilon}^2$$

$$e^{\sigma^2} = 1 + \sigma^2 + \frac{1}{2}\sigma^4 + \frac{1}{6}\sigma^6 + \cdots$$

$$\sim 1 + \sigma^2$$

$$e^{\sigma^2}(e^{\sigma^2} - 1) \sim (1 + \sigma^2)\sigma^2 \sim \sigma^2$$

$$\sigma = 0.1$$

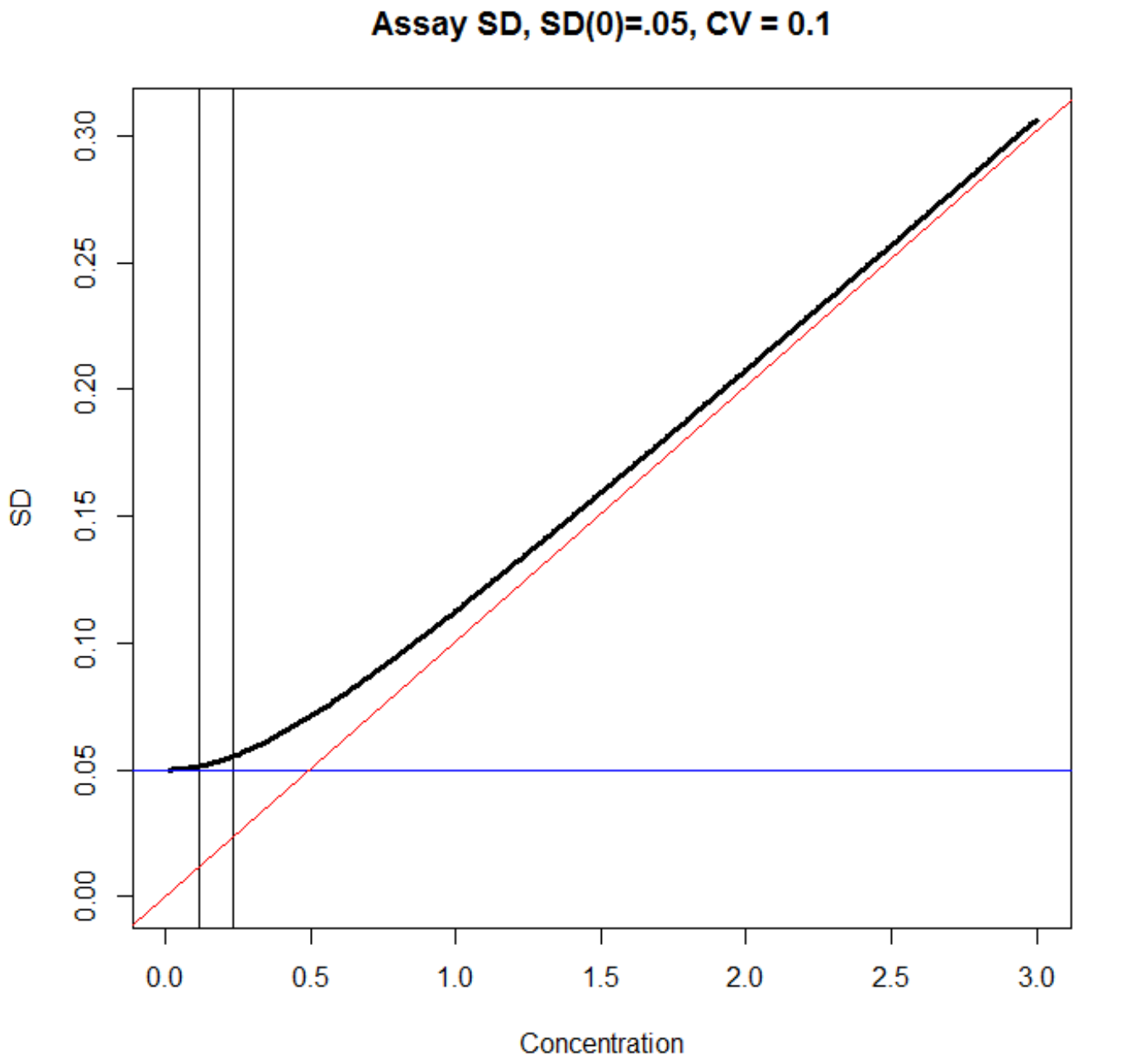
$$\sigma^2 = 0.01$$

$$\sigma^4 = 0.0001$$

$$\sigma = 0.2$$

$$\sigma^2 = 0.04$$

$$\sigma^4 = 0.0016$$


Assay SD, SD(0)=.05, CV = 0.1
SD
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Concentration

Detection Limits for Calibrated Assays

$$y = a + bxe^{\eta} + \epsilon$$

$$\hat{x} = \frac{y - a}{b}$$

$$= xe^{\eta} + \epsilon / b$$

- CL is in units of the response originally, can be translated to units of concentration.
- MDC is in units of concentration.

So-Called Limit of Quantitation

- Consider an assay with a variability near 0 of 29 ppt and a CV at high levels of 3.9%.
- Where is this assay most accurate?
  - Near zero where the SD is smallest?
  - At high levels where the CV is smallest?
- LOQ is where the CV falls to 20% from infinite at zero to 3.9% at large levels.
- This happens at 148ppt
- Some use 10*sd(0) = 290 ppt instead
- CL is at 67 ppt and MDV is at 135 ppt
- Some say that measurements between 67 and 148 show that there is detection, but it cannot be quantified.
- This is clearly wrong.

Conc	Mean	SD	CV
0	22	28	—
10	29	2	0.07
20	81	4	0.05
100	164	17	0.10
200	289	5	0.02
500	555	12	0.02
1,000	1,038	32	0.03
2,000	1,981	28	0.01
5,000	4,851	188	0.04
10,000	9,734	511	0.05
25,000	25,146	739	0.03

Confidence Limits

- Ignoring uncertainty in the calibration line.
- Assume variance is well enough estimated to be known
- Use $SD^2(x) = (28.9)^2 + (0.039x)^2$
- A measured value of 0 has SD(0) = 28.9, so the 95% CI is $0 \pm (1.960)(28.9) = 0 \pm 57$ or [0, 57]
- A measured value of 10 has SD(10) = 28.9 so the CI is [0, 67]
- A measured value of -10 has a CI of [0, 47]
- For high levels, make the CI on the log scale

Zinc Calibration

- If we take the spiked concentration and the peak area and predict the peak area from the concentration using linear regression, we get
  Peak Area = 104.5 + 7.2080 Concentration.
- The predicted concentration is then
  (Peak Area – 104.5)/7.2080
- The peak area measurements at 0 true concentration are 115, 631, 508, 317, 220, 93, 99, 135
- The predicted concentrations in ppt are then 1.45, 73.04, 55.97, 29.48, 16.02, –1.60, –0.77, 4.23
- Note that two of them are negative. These should not be reported as < 0 rather than the actual number.

Weighted Least Squares

$y = \beta_0 + \beta_1 x + \epsilon$

$V(\epsilon) = \sigma_\epsilon^2 + x^2 \sigma_\eta^2$

A simple approach is as follows:

Step 1: For each concentration level x_i, find the sample variance of the peak areas s_i^2

Step 2: Regress the s_i^2 on the x_i^2 to produce predicted variances $\hat{\sigma}_i^2$ for each x_i

Step 3: Regress y on x with weights $1/\hat{\sigma}_i^2$

Compare with unweighted regression.

More complex methods are in Wilson, Rocke, Durbin, and Kahn (2004).

These methods can fall into trouble if the estimated variance at 0 is negative or 0.

Exercise 1

- The standard deviation of measurements at low level for a method for detecting benzene in blood is 52 ng/L.
  - What is the Critical Level if we use a 1% probability criterion?
  - What is the Minimum Detectable Value?
  - If we can use 52 ng/L as the standard deviation, what is a 95% confidence interval for the true concentration if the measured concentration is 175 ng/L?
  - If the CV at high levels is 12%, about what is the standard deviation at high levels for the natural log measured concentration? Find a 95% confidence interval for the concentration if the measured concentration is 1850 ng/L?

Exercise 2

- Download data on measurement of zinc in water by ICP/MS ("Zinc.csv"). Use read.csv() to load.
- Conduct a regression analysis in which you predict peak area from concentration
- Which of the usual regression assumptions appears to be satisfied and which do not?
- What would the estimated concentration be if the peak area of a new sample was 1850?
- From the blanks part of the data, how big should a result be to indicate the presence of zinc with some degree of certainty?
- Try using weighted least squares for a better estimate of the calibration curve. Does it seem to make a difference?

References

- Lloyd Currie (1995) "Nomenclature in Evaluation of Analytical Methods Including Detection and Quantification Capabilities," *Pure & Applied Chemistry*, **67**, 1699–1723.
- David M. Rocke and Stefan Lorenzato (1995) "A Two-Component Model For Measurement Error In Analytical Chemistry," *Technometrics*, **37**, 176–184.
- Machelle Wilson, David M. Rocke, Blythe Durbin, and Henry Kahn (2004) "Detection Limits And Goodness-of-Fit Measures For The Two-component Model Of Chemical Analytical Error," *Analytica Chimica Acta*, **509**, 197–208.