

## Long Problem Set 5 – Answer Key

1. A plot of the data (not shown) suggests that a straight-line is an appropriate model for this data. To evaluate the two possible models (with and without an intercept), we create two linear models and examine their summaries.

```
> intno=lm(intensity~-1+conc) # the model  $y = \beta_1x$ 
> summary(intno)

Call:
lm(formula = intensity ~ -1 + conc)

Residuals:
    1     2     3     4     5     6 
4.00000 0.58545 3.37091 -0.04364 -4.45818 2.12727

Coefficients:
      Estimate Std. Error  t value  Pr(>|t|)
conc  2.06145   0.04353   47.36   7.93e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.228 on 5 degrees of freedom
Multiple R-Squared: 0.9978, Adjusted R-squared: 0.9973
F-statistic: 2243 on 1 and 5 DF, p-value: 7.93e-08

> intyes=lm(intensity~conc) # the model  $y = \beta_0 + \beta_1x$ 
> summary(intyes)

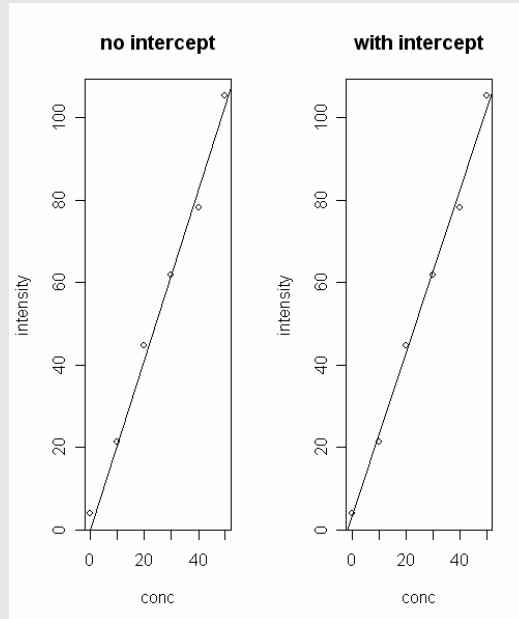
Call:
lm(formula = intensity ~ conc)

Residuals:
    1     2     3     4     5     6 
1.0762 -1.5410 2.0419 -0.5752 -4.1924 3.1905

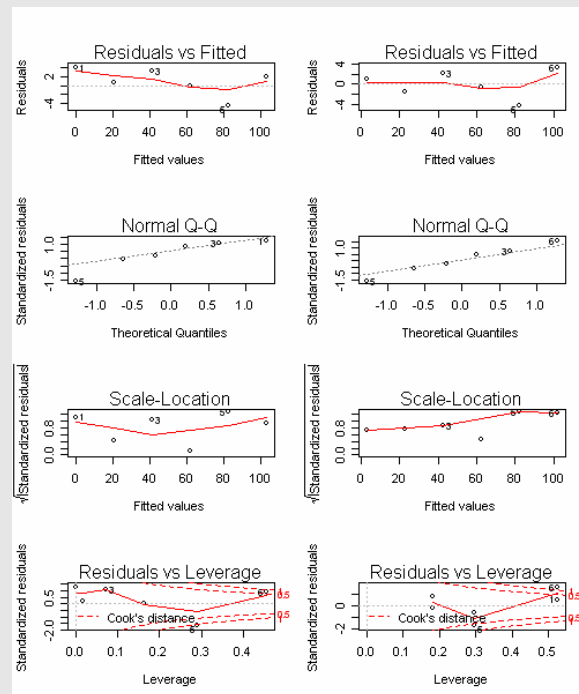
Coefficients:
      Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  2.9238     2.1648   1.351   0.248
conc         1.9817     0.0715  27.715  1.01e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.991 on 4 degrees of freedom
Multiple R-Squared: 0.9948, Adjusted R-squared: 0.9935
F-statistic: 768.1 on 1 and 4 DF, p-value: 1.008e-05
```

The slope for the two models are slightly different (2.06 for the model without an intercept and 1.98 for the model with an intercept). When included, the intercept of 2.92 is not found to be significantly different than zero ( $p = 0.248$ ), which suggests that it may not be needed. The model with the intercept has a slightly smaller residual standard error (2.991 vs. 3.228). A plot of the two models



doesn't suggest any particular advantage of one model over the other. Examination of residual plots, Q-Q plots and leverage plots for the two models



suggests that the model without the intercept (left) has the same number of leverage points, but a small trend in the residuals and a less normal Q-Q plot. In general, even if an intercept should, in theory, be zero (as is the case here), it is best to include an intercept in the model to correct for small determinate errors in the blank. Showing that there is a determinate error in the blank often is difficult due to the small number of degrees of freedom for calibration curves prepared with a limited number of standards. Sometimes intuition must take precedence over statistical tests!

2. The data in this problem comes from a classic data set created by Anscombe. Each data set produces a nearly identical statistical summary when fit to the model

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

```
> lm1=lm(y1~x1)
> summary(lm1)
```

```
Call:
lm(formula = y1 ~ x1)
```

```
Residuals:
    Min     1Q   Median     3Q      Max
-1.92127 -0.45577 -0.04136  0.70941  1.83882
```

```
Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)  3.0001     1.1247   2.667  0.02573 *
x1           0.5001     0.1179   4.241  0.00217 **
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-Squared: 0.6665, Adjusted R-squared: 0.6295
F-statistic: 17.99 on 1 and 9 DF, p-value: 0.002170
```

```
> lm2=lm(y2~x2)
> summary(lm2)
```

```
Call:
lm(formula = y2 ~ x2)
```

```
Residuals:
    Min     1Q   Median     3Q      Max
-1.9009 -0.7609  0.1291  0.9491  1.2691
```

```

Coefficients:
      Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  3.001      1.125    2.667   0.02576 *
x2           0.500      0.118    4.239   0.00218 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-Squared: 0.6662,    Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF, p-value: 0.002179

> lm3=lm(y3~x3)
> summary(lm3)

Call:
lm(formula = y3 ~ x3)

Residuals:
   Min     1Q   Median     3Q    Max
-1.1586 -0.6146 -0.2303  0.1540  3.2411

Coefficients:
      Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  3.0025      1.1245    2.670   0.02562 *
x3           0.4997      0.1179    4.239   0.00218 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

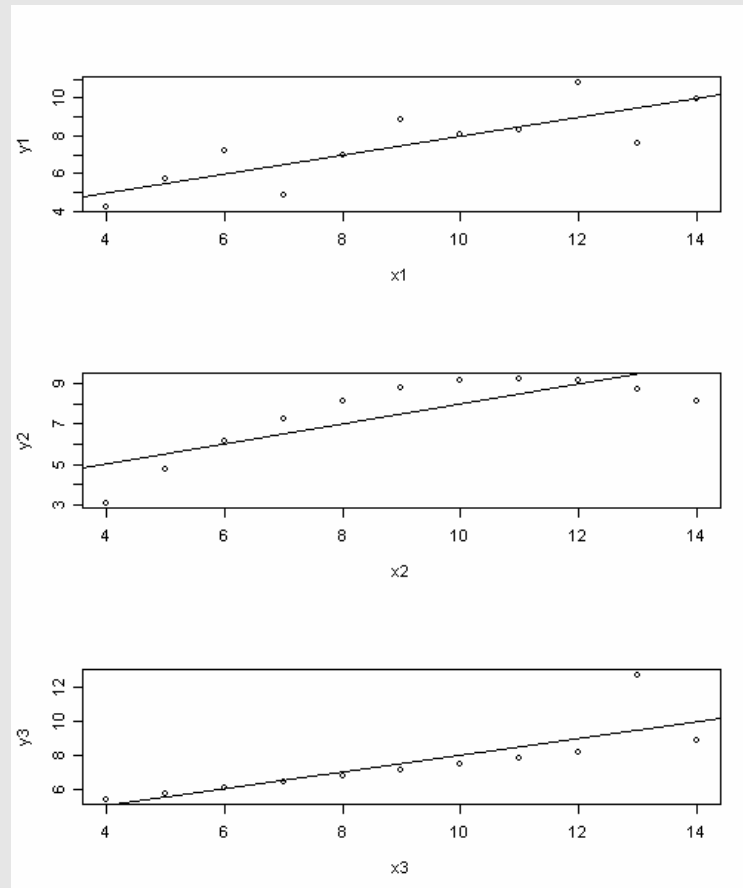
Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-Squared: 0.6663,    Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF, p-value: 0.002176

```

Note that each data set yields nearly identical values for the slope, the intercept, the residual standard error, the correlation coefficient and the F-statistic. Together, these values suggest that the model is equally good or equally poor for all three data sets. The only hint that the models are not identical is the information about residuals (given here in terms of quartiles), which show differences between the three models, although the information is hard to interpret.

Although residual plots, Q-Q plots and leverage plots are useful tools, it is always a good idea to first examine the actual data and the regression line; **there is no substitute in data analysis for looking at the data itself!** These plots, which are on the next page, clearly suggest that the first data set is the only one of the three data sets for which the straight-line model is suitable (although the data clearly are subject to much uncertainty). The second data set clearly shows curvature and is probably

best fit to a second-order polynomial and the third data set shows evidence of data that follows a straight-line with a single outlier.



Fitting the model

$$\psi = \beta_0 + \beta_1x + \beta_2x^2$$

to the second data set, gives the following summary information.

```
> lm4=lm(y2~x2+I(x2^2))
> summary(lm4)
```

```
Call:
lm(formula = y2 ~ x2 + I(x2^2))
```

Residuals:

Min	1Q	Median	3Q	Max
-0.0013287	-0.0011888	-0.0006294	0.0008741	0.0023776

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-5.9957343	0.0043299	-1385	<2e-16 ***
x2	2.7808392	0.0010401	2674	<2e-16 ***
I(x2^2)	-0.1267133	0.0000571	-2219	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

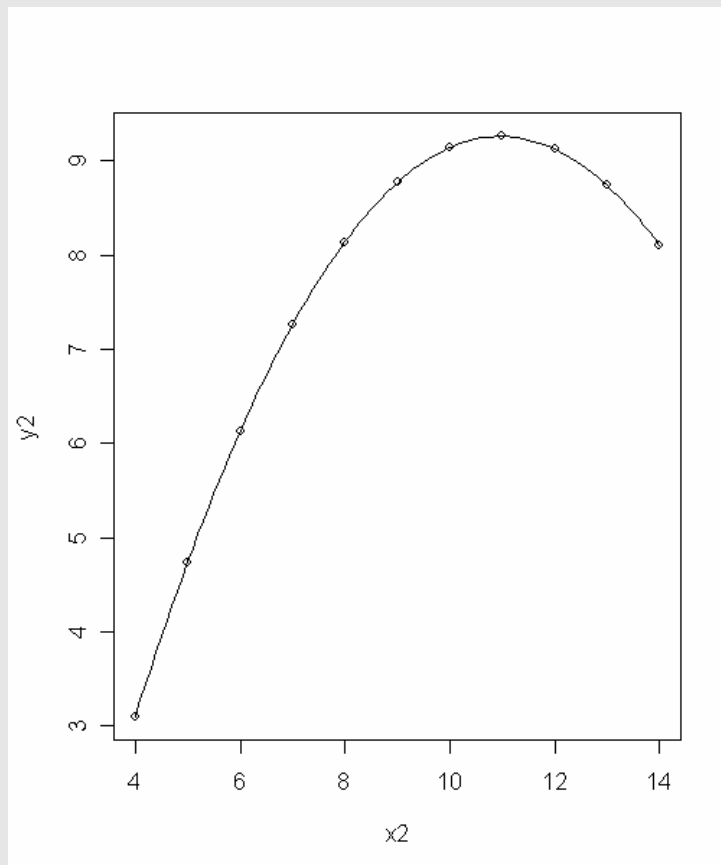
Residual standard error: 0.001672 on 8 degrees of freedom

Multiple R-Squared: 1, Adjusted R-squared: 1

F-statistic: 7.378e+06 on 2 and 8 DF, p-value: < 2.2e-16

Note that the residual errors and the residual standard error are substantially smaller for this model, and the F-statistic is much more significant. A plot of the data and the model

```
> newx=seq(4,14,0.1)
> newy=-5.9957343+2.7808392*newx-0.1267133*newx^2
> plot(x2,y2); lines(newx,newy)
```



shows that the model provides an excellent fit to the data. Removing the apparent outlier ( $x_3 = 13, y_3 = 12.74$ ) from the third data set (the third value in each object)

```
> x3new=x3[-3]
> y3new=y3[-3]
```

and completing the analysis

```
> lm5=lm(y3new~x3new)
> summary(lm5)
```

Call:

```
lm(formula = y3new ~ x3new)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.156e-03	-2.224e-03	6.494e-05	1.818e-03	5.065e-03

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.0056494	0.0029242	1370	<2e-16 ***
x3new	0.3453896	0.0003206	1077	<2e-16 ***

---

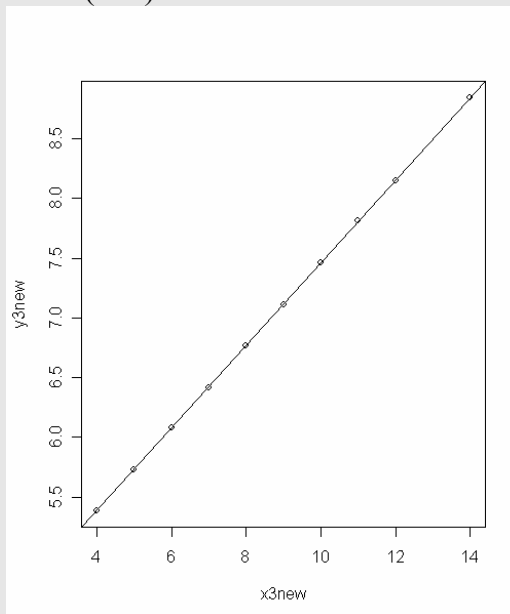
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.003082 on 8 degrees of freedom

Multiple R-Squared: 1, Adjusted R-squared: 1

F-statistic: 1.161e+06 on 1 and 8 DF, p-value: < 2.2e-16

```
> plot(x3new,y3new); abline(lm5)
```



shows that a straight-line model is appropriate.

3. Taking the gravimetric analysis as the independent variable and the ion-selective electrode method as the dependent variable (this is arbitrary; you can switch the designations without changing the conclusion), we arrive at the following linear model:

```
> lm.r=lm(ISE~Grav)
> summary(lm.r)

Call:
lm(formula = ISE ~ Grav)

Residuals:
    Min     1Q   Median     3Q     Max
-19.738  -6.937  -2.778   1.435  38.705

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.48367    8.69393   0.516   0.62
Grav         0.96294    0.08571  11.235 3.54e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.73 on 8 degrees of freedom
Multiple R-Squared:  0.9404,    Adjusted R-squared:  0.9329
F-statistic: 126.2 on 1 and 8 DF,  p-value: 3.537e-06
```

gives the slope as 0.96294 with a standard deviation of 0.08571. A t-test of

$$H_0: \beta_1 = 1.000 \quad H_A: \beta_1 \neq 1.000$$

shows

```
> tb1=abs((1.00-0.96294)/0.08571);tb1
[1] 0.4323883

> pt(tb1,8,lower.tail=FALSE)*2           # multiply by 2 to make it two-tailed
[1] 0.6768782
```

that there is insufficient evidence to suggest that slope is not 1.000; therefore, we conclude that there is no evidence for a difference in the methods.

A more traditional significance test for paired data (yes, the data are paired: each sample is analyzed by each method and the values for the samples themselves vary quite a bit, both of which are hallmarks of paired data):

```
> t.test(Grav,ISE,paired=TRUE)
```

Paired t-test

data: Grav and ISE

t = -0.2973, df = 9, p-value = 0.773

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

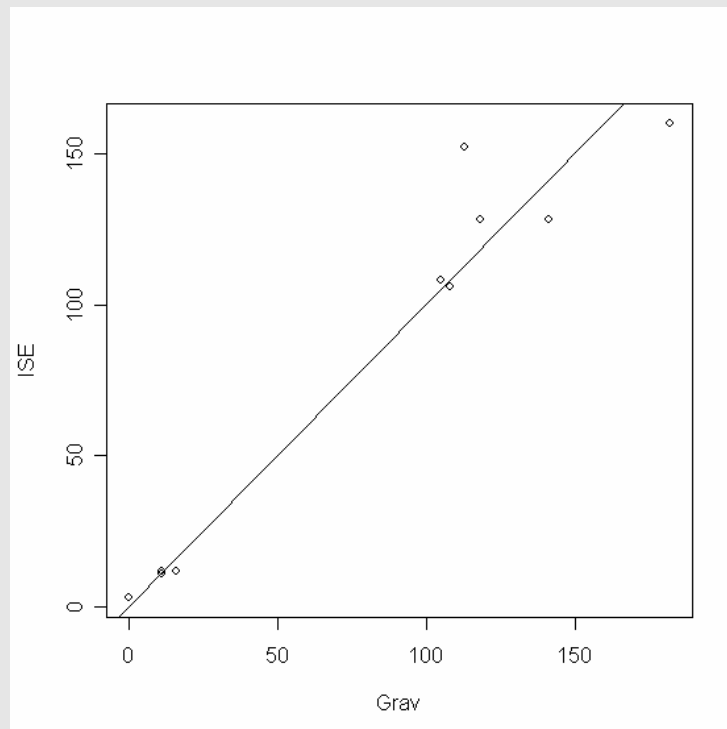
-12.912129 9.912129

sample estimates:

mean of the differences

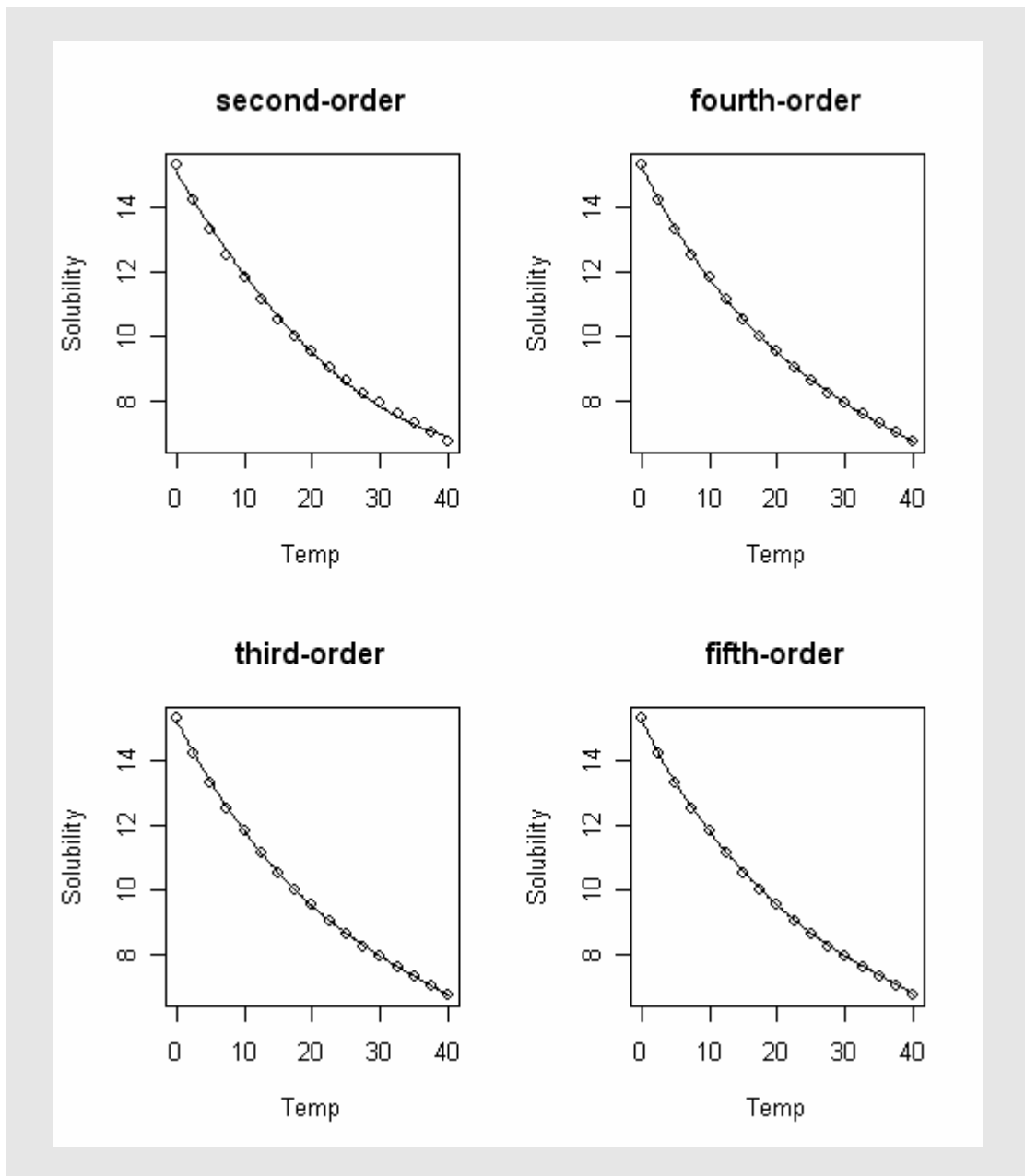
-1.5

gives the same result. As always, examining the data visually is a good idea (the line is not the regression line, but the ideal trend of an intercept of 0 and a slope of 1).



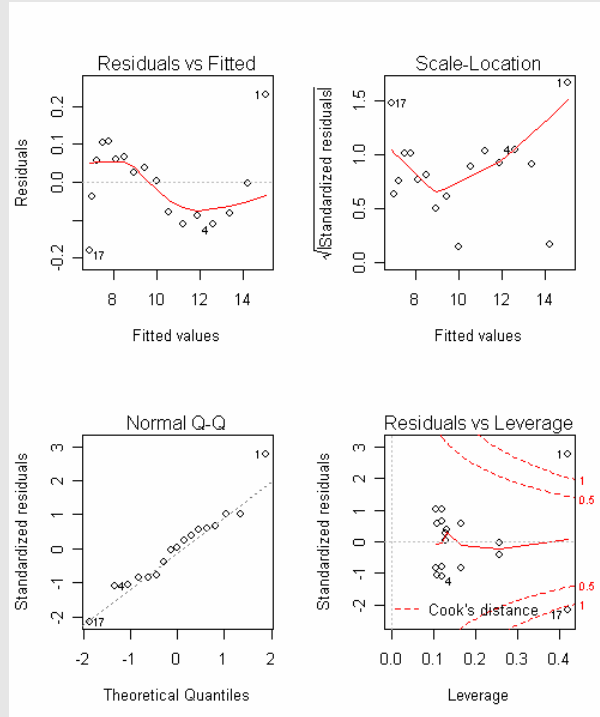
There is some scatter in the data, but most results show agreement between the two methods. Note that the absolute difference between the two methods becomes greater for larger concentrations of analyte, but the values remain randomly scattered on either side of the line.

4. A plot of the data (not shown) shows that the data are curved. To start, we will create models using polynomials of order 2, 3, 4 and 5 and examine the fits by looking at the data and predictions.

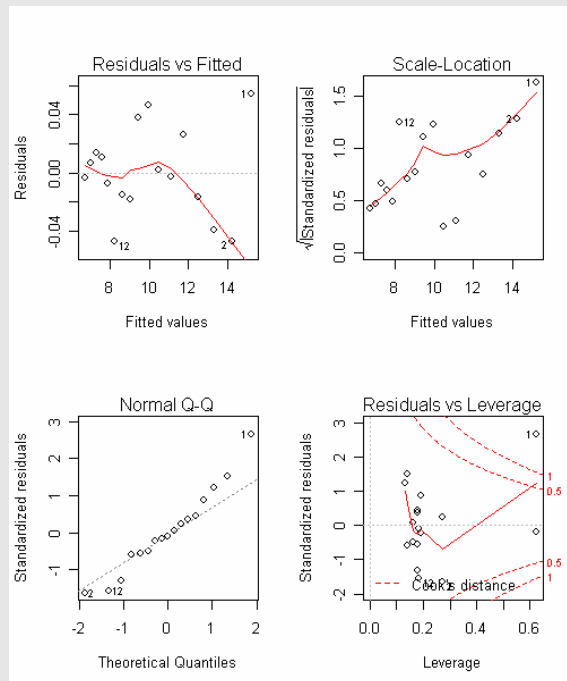


The second-order polynomial seems to fit most of the data but shows small deviations at the beginning and the end. In fact, if you look closely you will see that there are subtle differences between the data and the second-order polynomial model. A third-order polynomial fits all of the data fairly well and the fit does not appear to improve upon moving to higher-order polynomials.

As a first guess we might expect a third-order polynomial to be the best choice. To examine more closely, we might choose to look at additional plots, such as residual plots, Q-Q plots and leverage plots; these are shown on the next page for the second-order polynomial



and for the third-order polynomial



The residual errors for the second-order polynomial are consistent with our earlier observation that the model does not fit the extremes particularly well, but the other plots are more encouraging. The residual error plot for the third-order polynomial is less useful as the residual errors are generally small; there is evidence, however, that some data points have particular leverage on the regression.

All in all, the third-order polynomial provides the best fit to the data.