Further topics in regression

Krzysztof Podgórski
Department of Mathematics and Statistics
University of Limerick

November 12, 2009

Weighted regression lines

Weighted regression lines

- *Homoscedasticity* - the standard deviations of y-observations from the straight line are the same independently of the underlying x-observations.

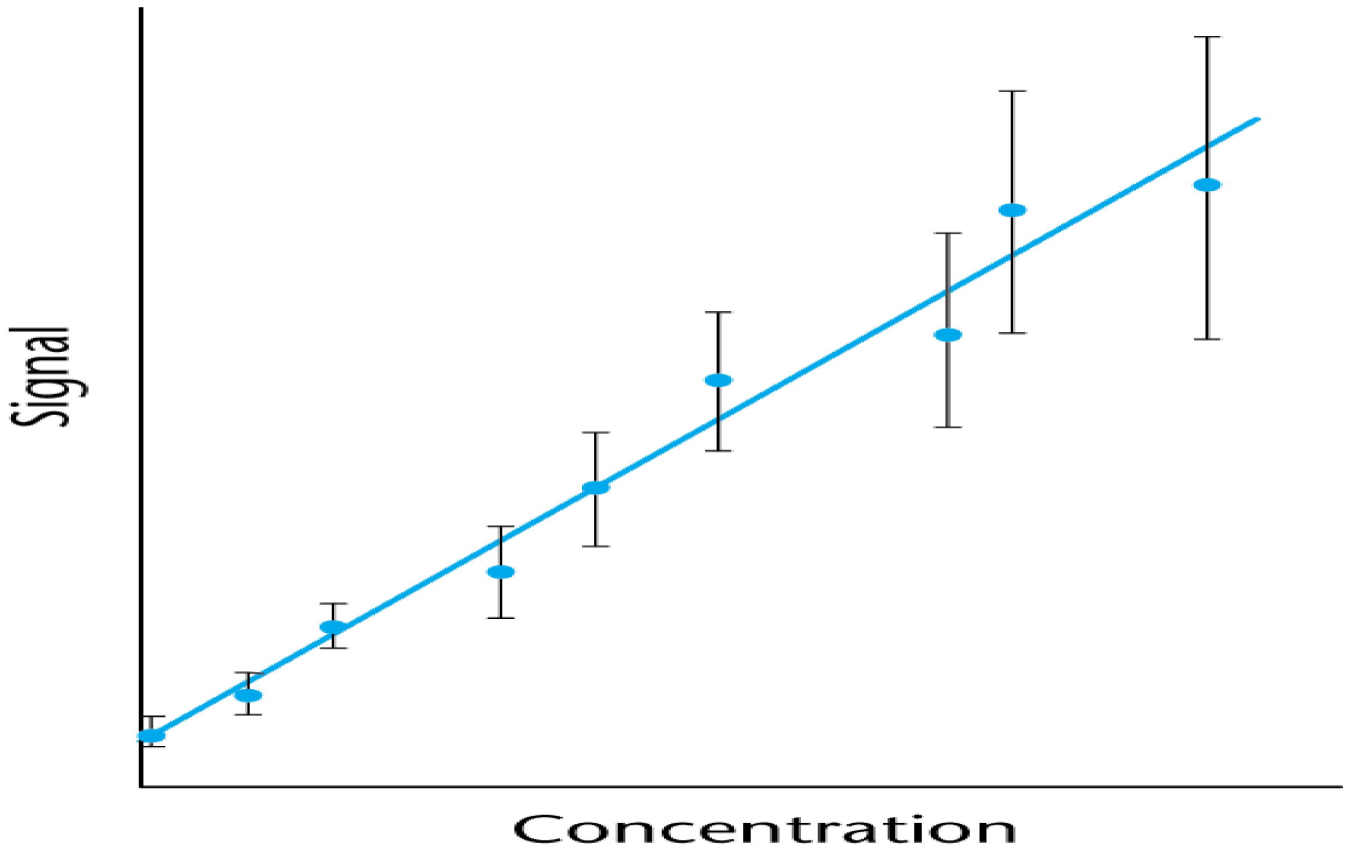
Weighted regression lines

- *Homoscedasticity* - the standard deviations of y-observations from the straight line are the same independently of the underlying x-observations.
- *Heteroscedasticity* - the standard deviations of y-observations depend on the underlying x-observations.

Weighted regression lines

- *Homoscedasticity* - the standard deviations of y-observations from the straight line are the same independently of the underlying x-observations.
- *Heteroscedasticity* - the standard deviations of y-observations depend on the underlying x-observations.
- In the first case, standard regression analysis should be performed, while in the second the weighted regression is more suitable.

Graphical representation of heteroscedasticity



Weights

The analysis has to be adjusted through weights that gives more emphasize to the values with small deviations and less to the ones with larger ones.

Weights

The analysis has to be adjusted through weights that gives more emphasize to the values with small deviations and less to the ones with larger ones.

- For this standard deviations s_i has to be given together with (x_i, y_i).

Weights

The analysis has to be adjusted through weights that gives more emphasize to the values with small deviations and less to the ones with larger ones.

- For this standard deviations s_i has to be given together with (x_i, y_i).
- They can be obtained through multiple measurements at a given x_i.

Weights

The analysis has to be adjusted through weights that gives more emphasize to the values with small deviations and less to the ones with larger ones.

- For this standard deviations s_i has to be given together with (x_i, y_i).
- They can be obtained through multiple measurements at a given x_i.
- Weights are defined as

$$w_i = \frac{s_i^{-2}}{\sum_{k=1}^{n} s_k^{-2}/n}.$$

Weights

The analysis has to be adjusted through weights that gives more emphasize to the values with small deviations and less to the ones with larger ones.

- For this standard deviations s_i has to be given together with (x_i, y_i).
- They can be obtained through multiple measurements at a given x_i.
- Weights are defined as

$$w_i = \frac{s_i^{-2}}{\sum_{k=1}^{n} s_k^{-2}/n}.$$

- Weights are inverse proportional to the variance.

Weighted regression fit

slope and the intercept of the regression line are then given by:

$$\text{Weighted slope:} \qquad b_w = \frac{\sum_i w_i x_i y_i - n \bar{x}_w \bar{y}_w}{\sum_i w_i x_i^2 - n \bar{x}_w^2} \tag{5.15}$$

and

$$\text{Weighted intercept:} \qquad a_w = \bar{y}_w - b \bar{x}_w \tag{5.16}$$

In equation (5.16) $\bar{y}_w$ and $\bar{x}_w$ represent the coordinates of the *weighted centroid* through which the weighted regression line must pass. These coordinates are given as expected by $\bar{x}_w = \sum_i w_i x_i / n$ and $\bar{y}_w = \sum_i w_i y_i / n$.

Example 5.10.1

Example 5.10.1

Calculate the unweighted and weighted regression lines for the following calibration data. For each line calculate also the concentrations of test samples with absorbances of 0.100 and 0.600.

Concentration, μg ml^{-1}	0	2	4	6	8	10
Standard deviation Absorbance	0.001 0.009	0.004 0.158	0.010 0.301	0.013 0.472	0.017 0.577	0.022 0.739

Example 5.10.1, cont.

Application of equations (5.4) and (5.5) shows that the slope and intercept of the *unweighted* regression line are respectively 0.0725 and 0.0133. The concentrations corresponding to absorbances of 0.100 and 0.600 are then found to be 1.20 and 8.09 μg ml^{-1} respectively.

The *weighted* regression line is a little harder to calculate: in the absence of a suitable computer program it is usual to set up a table as follows.

x_i	y_i	s_i	$1/s_i^2$	w_i	$w_i x_i$	$w_i y_i$	$w_i x_i y_i$	$w_i x_i^2$
0	0.009	0.001	10^6	5.535	0	0.0498	0	0
2	0.158	0.004	62500	0.346	0.692	0.0547	0.1093	1.384
4	0.301	0.010	10000	0.055	0.220	0.0166	0.0662	0.880
6	0.472	0.013	5917	0.033	0.198	0.0156	0.0935	1.188
8	0.577	0.017	3460	0.019	0.152	0.0110	0.0877	1.216
10	0.739	0.022	2066	0.011	0.110	0.0081	0.0813	1.100
Sums			1083943	5.999	1.372	0.1558	0.4380	5.768

These figures give $\bar{y}_w = 0.1558/6 = 0.0260$, and $\bar{x}_w = 1.372/6 = 0.229$. By equation (5.15), b_w is calculated from

$$b_w = \frac{0.438 - (6 \times 0.229 \times 0.026)}{5.768 - [6 \times (0.229)^2]} = 0.0738$$

so a_w is given by $0.0260 - (0.0738 \times 0.229) = 0.0091$.

These values for a_w and b_w can be used to show that absorbance values of 0.100 and 0.600 correspond to concentrations of 1.23 and 8.01 μg ml^{-1} respectively.

Regression vs. weighted regression

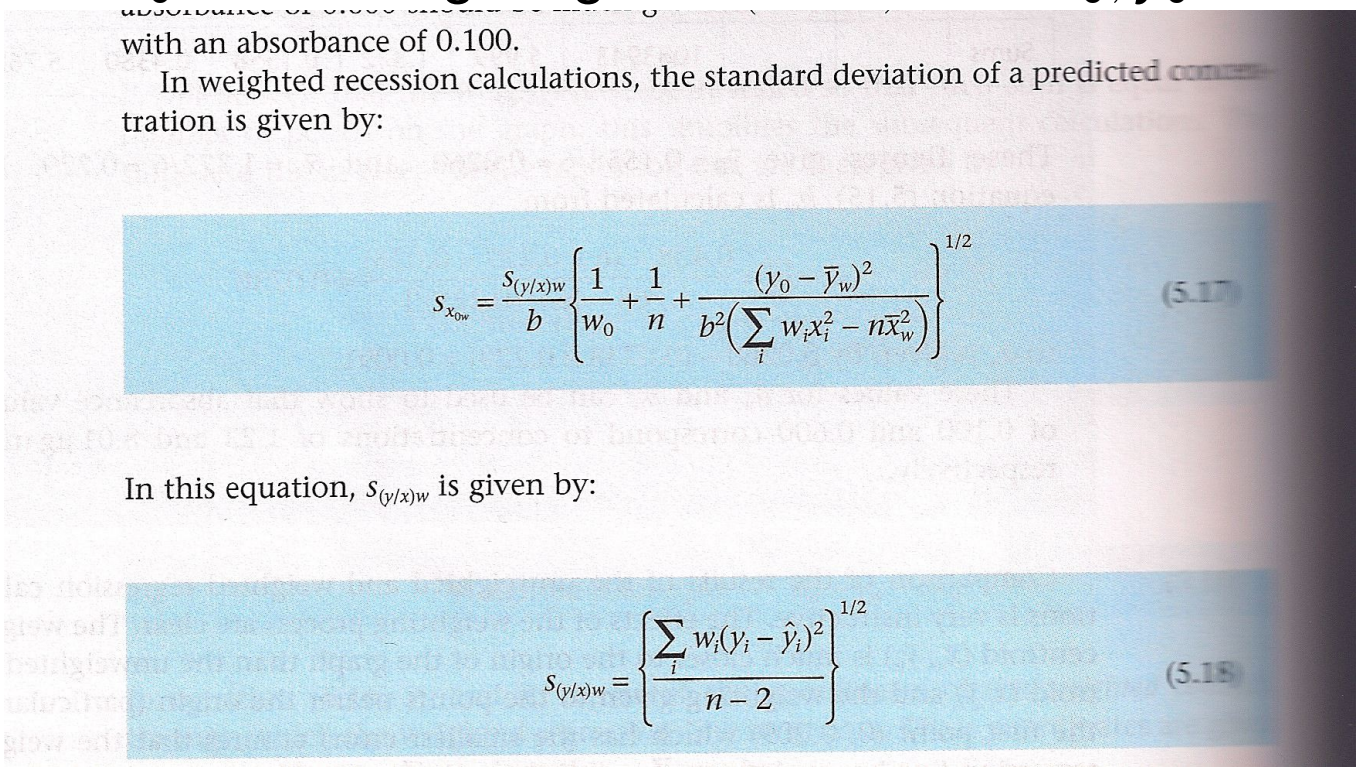
- Both approaches gives similar linear fits as expressed by slopes and and intercepts.

Regression vs. weighted regression

- Both approaches gives similar linear fits as expressed by slopes and and intercepts.
- They differ in error estimation.

Predicted concentration errors

Let w_0 is the weighting associated with x_0, y_0.



with an absorbance of 0.100.

In weighted recession calculations, the standard deviation of a predicted concentration is given by:

$$s_{x_{0w}} = \frac{s_{(y/x)w}}{b}\left\{\frac{1}{w_0} + \frac{1}{n} + \frac{(y_0 - \bar{y}_w)^2}{b^2\left(\sum_i w_i x_i^2 - n\bar{x}_w^2\right)}\right\}^{1/2} \quad (5.17)$$

In this equation, $s_{(y/x)w}$ is given by:

$$s_{(y/x)w} = \left\{\frac{\sum_i w_i(y_i - \hat{y}_i)^2}{n-2}\right\}^{1/2} \quad (5.18)$$

Computations in *R*

```
Conc=c(0,2,4,6,8,10)
StDev=c(0.001,0.004,0.010,0.013,0.017,0.022)
Abs=c(0.009,0.158,0.301,0.472,0.577,0.739)
n=length(Conc)
weights=StDev^(-2)/mean(StDev^(-2))
wreg=lm(Abs~Conc,weights=weights)
reg=lm(Abs~Conc)
summary(wreg)
```

Anova and regression calculations

- It is often convienent to express the regression analysis using ANOVA table. The following equation is the basis for such representation

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Anova and regression calculations

- It is often convienent to express the regression analysis using ANOVA table. The following equation is the basis for such representation

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- It is often shortened to

$$SS_T = SS_{LR} + SS_R,$$

where SS_T is refered to as the total sum of squares, SS_{LR} is the sum of squares due to linear regression (within regression), SS_R is the sum of squares due to residuals (outside regression).

R^2 – a measure of variation explained by regression

The following coefficient has a natural interpretation as amount of variability in the data that is explained by the regression fit:

$$R^2 = SS_{LR}/SS_T = 1 - SS_R/SS_T.$$

A similar interpretation is given to the adjusted coefficient R'^2 which is given by

$$R'^2 = 1 - MS_R/MS_T,$$

where MS_R is the mean squared error due to residuals, and MS_T is the total mean squared error. The adjusted coefficient is accounting for the degrees of freedom used for each source of variation and is often a more reliable indicator of variability than R^2. R'^2 is always smaller than R^2.

Example

Example 5.13.1

Investigate the linear calibration range of the following fluorescence experiment.

Fluorescence intensity	0.1	8.0	15.7	24.2	31.5	33.0
Concentration, $\mu g\ ml^{-1}$	0	2	4	6	8	10

Inspection of the data shows that the part of the graph near the origin corresponds rather closely to a straight line with a near-zero intercept and a slope of about 4. The fluorescence of the 10 $\mu g\ ml^{-1}$ standard solution is clearly lower than would be expected on this basis, and there is some possibility that the departure from linearity has also affected the fluorescence of the 8 $\mu g\ ml^{-1}$ standard. We first apply (unweighted) linear regression calculations to all the data. Application of the methods of Sections 5.3 and 5.4 gives the results $a = 1.357$, $b = 3.479$ and $r = 0.9878$. Again we recall that the high value for r may be deceptive, though it may be used in a comparative sense (see below). The y-residuals are found to be –1.257, –0.314, +0.429, +1.971, +2.314, and –3.143, with the sum of squares of the residuals equal to 20.981. The trend in the values of the residuals suggests that the last value in the table is probably outside the linear range.

We confirm this suspicion by applying the linear regression equations to the

```
Int=c(0.1,8.0,15.7,24.2,31.5,33.0)
Conc=c(0,2,4,6,8,10)
res=lm(Int~Conc)
res
```

Different outputs in *R* on regression

```
res

Call:
lm(formula = Int ~ Conc)

Coefficients:
(Intercept)         Conc
      1.357        3.479

anova(res)
Analysis of Variance Table

Response: Int
          Df Sum Sq Mean Sq F value    Pr(>F)
Conc       1 847.03  847.03  161.47 0.0002209 ***
Residuals  4  20.98    5.25
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Summary of regression

```
 summary(regr)
Call:
lm(formula = Int ~ Conc)

Residuals:
      1       2       3       4       5       6
-1.2571 -0.3143  0.4286  1.9714  2.3143 -3.1429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3571     1.6576   0.819 0.458916
Conc          3.4786     0.2737  12.707 0.000221 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.29 on 4 degrees of freedom
Multiple R-squared: 0.9758,     Adjusted R-squared: 0.9698
F-statistic: 161.5 on 1 and 4 DF,  p-value: 0.0002209
```

Curve fitting

Example 5.14.1

In an instrumental analysis the following data were obtained (arbitrary units).

Concentration	0	1	2	3	4	5	6	7	8	9	10
Signal	0.2	3.6	7.5	11.5	15.0	17.0	20.4	22.7	25.9	27.6	30.2

Fit a suitable polynomial to these results, and use it to estimate the concentrations corresponding to signal of 5, 16 and 27 units.

Even a casual examination of the data suggests that the calibration plot should be a curve, but it is instructive nonetheless to calculate the least-squares straight line through the points using the method described in Section 5.4. This line turns out to have the equation $y = 2.991x + 1.555$. The ANOVA table for the data has the following form:

Source of variation	Sum of squares	d.f.	Mean square
Regression	984.009	1	984.009
Residual	9.500	9	1.056
Total	993.509	10	99.351

As already noted, the number of degrees of freedom (d.f.) for the variation due to regression is equal to the number of terms (k) in the regression equation containing x, x^2, etc. For a straight line, k is 1. There is only one constraint in the calculation (viz. that the sum of the residuals is zero, see above), so the total number of degrees of freedom is $(n - 1)$. Thus the number of degrees of freedom assigned to the residuals is $(n - k - 1) = (n - 2)$ in this case. From the ANOVA table R^2 is given by $984.009/993.509 = 0.99044$, i.e. 99.044%. An equation which explains

Curve fitting, cont.

over 99% of the relationship between x and y seems quite satisfactory but, just as is the case with the correlation coefficient, r, we must use great caution in interpreting absolute values of R^2: it will soon become apparent that a quadratic curve provides a much better fit for the data. We can also calculate the R'^2 value from equation (5.27): it is given by $[1 - (1.056/99.351)] = 0.98937$, i.e. 98.937%.

As always, an examination of the residuals usually provides valuable information on the success of a calibration equation. In this case the residuals are as follows:

x	y_i	$\hat{y}_i$	y-residual
0	0.2	1.0	−1.4
1	3.6	4.5	−0.9
2	7.5	7.5	0
3	11.5	10.5	1.0
4	15.0	13.5	1.5
5	17.0	16.5	0.5
6	20.4	19.5	0.9
7	22.7	22.5	0.2
8	25.9	25.5	0.4
9	27.6	28.5	−0.9
10	30.2	31.5	−1.3

In this table, the numbers in the two right-hand columns have been rounded to one decimal place for simplicity. The trend in the signs and magnitudes of the residuals, which are negative at low x-values, rise to a positive maximum, and then return to negative values, is a sure sign that a straight line is not a suitable fit for the data.

When the data are fitted by a curve of quadratic form the equation turns

Curve fitting, cont.

When the data are fitted by a curve of quadratic form the equation turns out to be $y = 0.086 + 3.970x - 0.098x^2$, and the ANOVA table takes the form:

Source of variation	Sum of squares	d.f.	Mean square
Regression	992.233	2	494.116
Residual	1.276	8	0.160
Total	993.509	10	99.351

Note that the number of degrees of freedom for the regression and residual sources of variation have now changed in accordance with the rules described above, but that the total variation is naturally the same as in the first ANOVA table. Here R^2 is 992.233/993.509 = 0.99872, i.e. 99.872%. This figure is noticeably higher than the value of 99.044% obtained from the linear plot, and the R'^2 value is also higher at $[1 - (0.160/99.3511)] = 0.99839$, i.e. 99.839%. When the y-residuals are calculated, their signs (in increasing order of x-values) are + − − + + − + − + − +. There is no obvious trend here, so on all grounds we must prefer the quadratic over the linear fit.

Curve fitting, cont.

Lastly we repeat the calculation for a cubic fit. Here, the best-fit equation is $y = -0.040 + 4.170x - 0.150x^2 + 0.0035x^3$. The cubic coefficient is very small indeed, so it is questionable whether this equation is a significantly better fit than the quadratic one. The R^2 value is, inevitably, slightly higher than that for the quadratic curve (99.879% compared with 99.872%), but the value of R'^2 is slightly *lower* than the quadratic value at 99.827%. The order of the signs of the residuals is the same as in the quadratic fit. As there is no value in including unnecessary terms, we can be confident that a quadratic fit is satisfactory in this case.

When the above equations are used to estimate the concentrations corresponding to instrument signals of 5, 16 and 27 units, the results (x-values in arbitrary units) are:

	Linear	Quadratic	Cubic
$y = 5$	1.15	1.28	1.27
$y = 16$	4.83	4.51	4.50
$y = 27$	8.51	8.61	8.62

As expected, the differences between the concentrations calculated from the quadratic and cubic equations are insignificant, so the quadratic equation is used for simplicity.

The same analysis in *R*

```
Conc=c(0,1,2,3,4,5,6,7,8,9,10)
Sig=c(0.2,3.6,7.5,11.5,15.0,17.0,20.4,22.7,25.9,27.
linear=lm(Sig~Conc)
ConcQ=Conc^2
ConcC=Conc^3
quadratic=lm(Sig~Conc+ConcQ)
cubic=lm(Sig~Conc+ConcQ+ConcC)
```

Results *R*

```
Call:
lm(formula = Sig ~ Conc)
Coefficients:
(Intercept)         Conc
      1.555        2.991
Call:
lm(formula = Sig ~ Conc + ConcQ)
Coefficients:
(Intercept)         Conc        ConcQ
    0.08601      3.96993     -0.09790
Call:
lm(formula = Sig ~ Conc + ConcQ + ConcC)
Coefficients:
(Intercept)         Conc        ConcQ        ConcC
  -0.039860     4.169930    -0.150350     0.003497
```

Anova results *R*

```
Analysis of Variance Tables
Response: Sig
          Df Sum Sq Mean Sq F value    Pr(>F)
Conc       1 984.01  984.01  932.22 2.123e-10 ***
Residuals  9   9.50    1.06
---
Response: Sig
          Df Sum Sq Mean Sq  F value    Pr(>F)
Conc       1 984.01  984.01 6168.254 7.701e-13 ***
ConcQ      1   8.22    8.22   51.551 9.429e-05 ***
Residuals  8   1.28    0.16
---
Response: Sig
          Df Sum Sq Mean Sq   F value    Pr(>F)
Conc       1 984.01  984.01 5736.7100  1.84e-11 ***
ConcQ      1   8.22    8.22   47.9441 0.0002264 ***
ConcC      1   0.08    0.08    0.4403 0.5282168
Residuals  7   1.20    0.17
---
```

Summaries *R*: Linear vs. Quadratic

```
Call:
lm(formula = Sig ~ Conc)
Residuals:
    Min      1Q  Median      3Q     Max
-1.3545 -0.9091  0.2091  0.6955  1.4818
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.55455    0.57953   2.682   0.0251 *
Conc         2.99091    0.09796  30.532 2.12e-10 ***
---
Residual standard error: 1.027 on 9 degrees of freedom
Multiple R-squared: 0.9904,     Adjusted R-squared: 0.9894
F-statistic: 932.2 on 1 and 9 DF,  p-value: 2.123e-10

Call:
lm(formula = Sig ~ Conc + ConcQ)
Residuals:
     Min       1Q   Median       3Q      Max
-0.48811 -0.32168  0.01888  0.26259  0.60070
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.08601    0.30429   0.283    0.785
Conc         3.96993    0.14157  28.041 2.82e-09 ***
ConcQ       -0.09790    0.01364  -7.180 9.43e-05 ***
---
Residual standard error: 0.3994 on 8 degrees of freedom
Multiple R-squared: 0.9987,     Adjusted R-squared: 0.9984
F-statistic:  3110 on 2 and 8 DF,  p-value: 2.723e-12
```

Summaries *R*: Quadratic vs. Cubic

```
Call:
lm(formula = Sig ~ Conc + ConcQ)
Residuals:
     Min       1Q   Median       3Q      Max
-0.48811 -0.32168  0.01888  0.26259  0.60070
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.08601    0.30429   0.283    0.785
Conc         3.96993    0.14157  28.041 2.82e-09 ***
ConcQ       -0.09790    0.01364  -7.180 9.43e-05 ***
---
Residual standard error: 0.3994 on 8 degrees of freedom
Multiple R-squared: 0.9987,     Adjusted R-squared: 0.9984
F-statistic:  3110 on 2 and 8 DF,  p-value: 2.723e-12

Call:
lm(formula = Sig ~ Conc + ConcQ + ConcC)
Residuals:
     Min       1Q   Median       3Q      Max
-0.48811 -0.27098  0.07762  0.26434  0.54196
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.039860   0.368162  -0.108    0.917
Conc         4.169930   0.335257  12.438    5e-06 ***
ConcQ       -0.150350   0.080295  -1.872    0.103
ConcC        0.003497   0.005269   0.664    0.528
---
Residual standard error: 0.4142 on 7 degrees of freedom
Multiple R-squared: 0.9988,     Adjusted R-squared: 0.9983
F-statistic:  1928 on 3 and 7 DF,  p-value: 1.428e-10
```

Outliers in regression - examininig residuals