# The Randomization Test Procedure: Testing for a Non-Zero Quadratic Effect

Ling Huang, Sacramento City College
Paul Johnson, Biostatistical Software Development

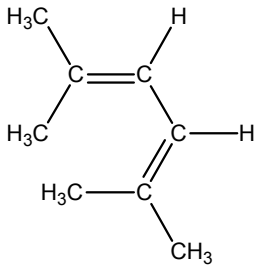
## ABSTRACT

A randomization test is performed for data from the ultraviolet light spectrum of 2,5-dimethyl-2,4-hexadiene (in methanol). This compound is a useful chemical intermediate for agricultural products. A quadratic model is fit to the data. Absorbance (the dependent variable) is a measure of the extent to which a compound absorbs radiation of a particular wavelength (the independent variable). The optimum value of absorbance is obtained and a two-tailed test for a non-zero quadratic effect is conducted. The SAS® macro is available for download from http://pages.prodigy.net/johnsonp12/rtest.html. The macro requires base SAS (SAS, 1990), SAS/GRAPH (1991) and SAS/STAT (1993) software to run.

## INTRODUCTION

The ultraviolet light spectrum of 2,5-dimethyl-2,4-hexadiene plots absorbance (y) against wavelength in nm (x) [see Browne and Foote, 2002].

The structure of 2,5-dimethyl-2,4-hexadiene is:




See http://www.eastman.com/Product_Information/ProductHome.asp?Product=906 for more background information on 2,5-dimethyl-2,4-hexadiene. This chemical is used only for illustrative purposes.

The (x,y) data pairs are listed in Appendix A. A regression model is fit to the data. The optimum value of absorbance is obtained. A two-tailed test for a non-zero quadratic effect is carried out.

## PROCEDURE

The regression model fit to the data is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

This is a quadratic model with a random error component PROC REG of SAS/STAT (SAS, 1993) is used to fit the model. The parameter estimates $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ are obtained. This is model 1.

The program performs a randomization test procedure for quadratic regression. The 2-tailed test for a non-zero quadratic effect is conducted and 'n_random' randomizations are carried out to compute an overall significance level (Manly, 1997). The y values are reallocated to the x values to carry out the randomizations.
The fitted model is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 \qquad \longrightarrow \text{(Model 1).}$$

A second model is fit considering wavelengths closer to the optimum absorbance. This is $\longrightarrow$ (Model 2).

### OPTIMUM OBTAINED

The optimum abundance ( $\hat{y}_{opt}$ ) is found by differentiation and setting $d\hat{y}/dx = 0$.

$d\hat{y}/dx = \hat{\beta}_1 + 2\hat{\beta}_2 x$. Setting to zero gives $\hat{\beta}_1 + 2\hat{\beta}_2 x = 0 \Rightarrow x \equiv x_{opt} = \dfrac{-\hat{\beta}_1}{2\hat{\beta}_2}$.

$$\hat{y}_{opt} = \hat{\beta}_0 - \left(\frac{\hat{\beta}_1^2}{4\hat{\beta}_2}\right). \longrightarrow \text{(eqn. 1)}$$ [see Appendix B].

The second derivative is $d^2\hat{y}/dx^2 = 2\hat{\beta}_2$.

The second derivative must be negative valued for an optimum (maximum). Hence for an optimum the condition $\hat{\beta}_2 < 0$ must be satisfied.

**HYPOTHESES**

The null and alternative hypotheses to test for a non-zero quadratic effect are:

$H_0 : \beta_2 = 0$ against the alternative $H_A : \beta_2 \neq 0$.

**TEST STATISTIC**

$$t = \hat{\beta}_2 / s.e.\left(\hat{\beta}_2\right) \longrightarrow \text{(eqn. 2)}.$$

It is appropriate to use a two-sided test to determine significance since patterns can make either high or low values of t likely. Let $n_2$ be the number of randomizations where the $|t|$ is greater or equal to the original $|t|$.

Then the significance level = $2(n_2 + 1)/(n\_random + 1) \longrightarrow$ (eqn. 3).

Manly (1997) suggests that 1000 (i.e., n_random = 999) randomizations is a reasonable minimum for a test at the 5% level of significance. The user may wish to change this to suit the user's needs. Manly (1997) for example indicates that 5,000 (i.e., n_random = 4,999) randomizations is a reasonable minimum to test at the 1% level of significance.

Finally a nonparametric regression program is used to fit a nonlinear model relating the dependent variable y and the independent variable x (see Neter, Wasserman and Kutner, 1985). This is model 3. A bandwidth of B% of the dynamic range of the x values is used. The predicted response curve using this method is constructed using the predicted y values. The bandwidth (B) is input and the bandwidth of B% of the range of the x values is calculated and is then used to produce the predicted response curve.

Model 3

The model fit is:

$$y_k = a + bx_k + e_k \longrightarrow \text{(Model 3)}.$$

with weights $w_k x_i$, where:

$$w_k[x_i] = \begin{cases} \left(1 - \left(\frac{x_k - x_i}{B}\right)^3\right)^3 & \text{if } |x_k - x_i| < B \\ 0 & \text{if } |x_k - x_i| \geq B \end{cases}.$$

A weighted least squares line is fitted for all values within the bandwidth and this line is used to predict the y values.

The second part of the fit involves the use of weights to 'down-weight' any outliers present.

The residuals are calculated. Robustness weights are computed that discount observations with large residuals ($e_i = y_i - \hat{y}_i$):

$$\delta_i = \omega_b(e_i / 6M).$$

M is the median of the absolute residuals $|e_i|$, and $\omega_b$ is the bisquare weight function:

$$\omega_b = \begin{cases} \left(1 - \left((x_k - x_i)/B\right)^2\right)^2 & \text{for } |(x_k - x_i)/B| < 1 \\ 0 & \text{for } |(x_k - x_i)/B| \geq 1 \end{cases}.$$

Compound weights, $\delta_k \; w_k x_i$ are used as above to provide for estimates of the predicted values.

## RESULTS

The results follow:

### MODEL 1

$$\hat{y}^{(1)} = -39.3115 + 0.3375x - 0.0007x^2 \qquad \left\{\hat{\beta}_0 = -39.311554; \hat{\beta}_1 = 0.337538; \hat{\beta}_2 = -0.000712\right\}.$$

The optimum abundance is found from eqn. 1 and equals 0.69 for a wavelength of 241.07 nm. This appears to be negatively biased. However on examination of the plot (see Figure 1) we see that the points for x values of 205 and 270 (the extreme points) appear to move the estimate of the optima and the curve itself to the left of the data. 95% confidence bounds for this point were found. Other confidence bounds and predicted values are obtained. Plots were produced using SAS/GRAPH (SAS, 1991).

### RANDOMIZATION TEST

$\hat{\beta}_2 = -0.000712,\ \text{s.e.}\left(\hat{\beta}_2\right) = 0.0008043,$

and t = $-8.848$ (eqn. 2). $n_2$ = 0 which gives a significance level = $2 * (0+1)/(999+1)$ = 0.002 (eqn. 3).

Figure 1 shows a plot of the predicted and observed values. Figure 2 shows a plot of the predicted values and a 95% confidence band for $205 \le x \le 270.$

**Figure 1.** Plot of Predicted and Observed Values ( $205 \le x \le 270$ ) [Model 1]

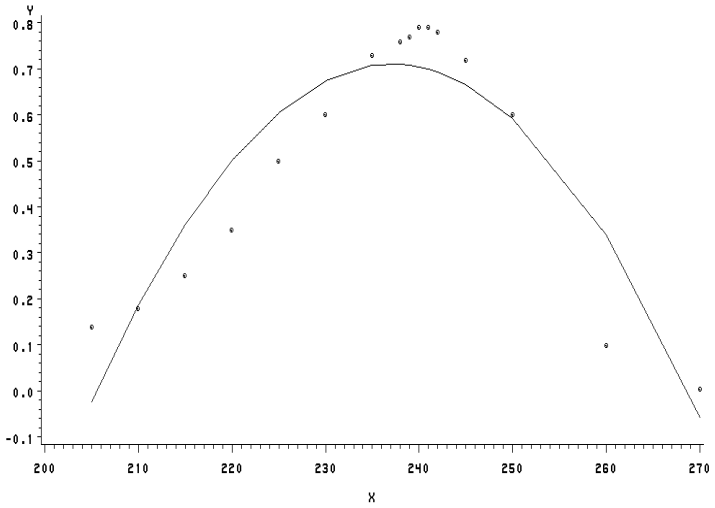
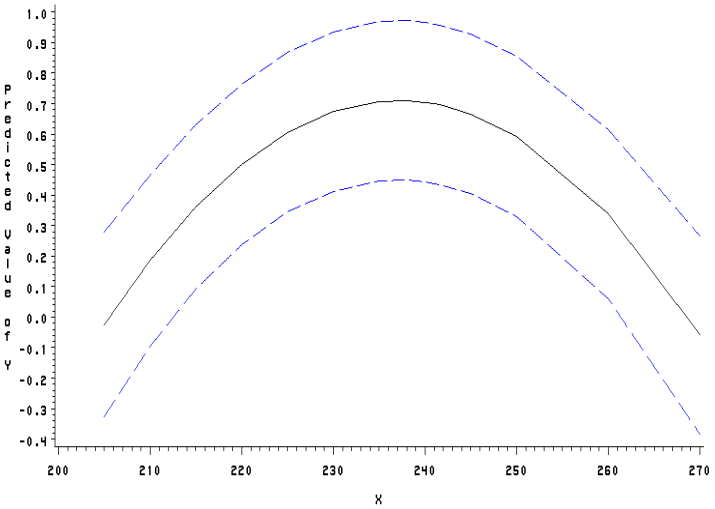



**Figure 2.** Plot of Predicted and 95% Confidence Band ( $205 \le x \le 270$ ) [Model 1]



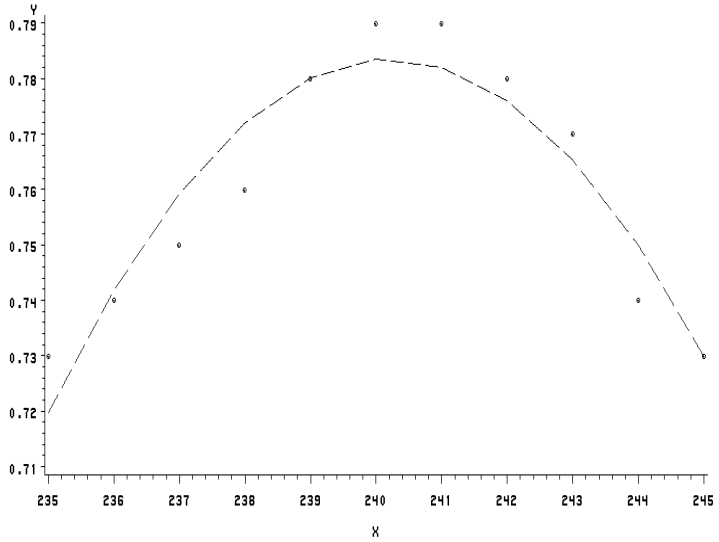
A secondary plot is produced for wavelengths between 235 nm and 245 nm showing predicted and observed values around the optimum (see Figure 3). A secondary model was fit and a 'new and improved' optimum found. This second model fit considering wavelengths between 235 nm and 245 nm is:

**MODEL 2**

$$\hat{y}^{(2)} = -134.3936 + 1.1255x - 0.0023x^2 .$$

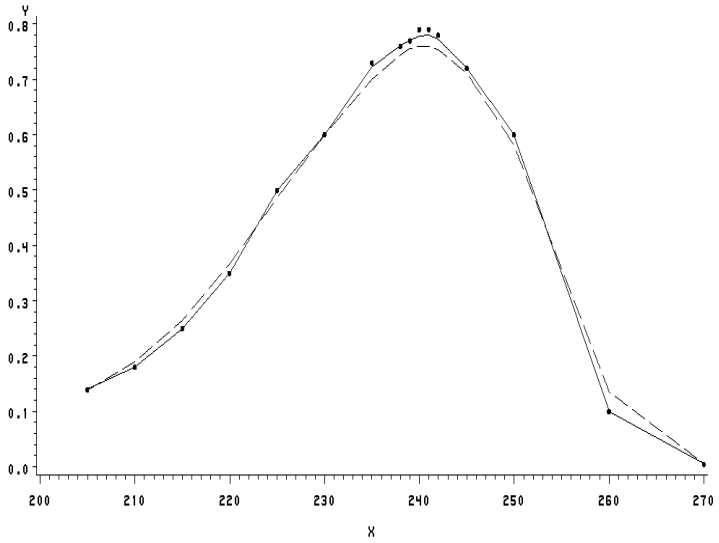
The optimum abundance from this model is found to be valued at 0.76 for a wavelength of 240.18 nm. The negative bias of the estimate of the maxima from model 1 has been removed. Model 2 provides for a more accurate refinement and improvement.

**Figure 3.** Plot of Predicted and Observed Values ( $235 \le x \le 245$ ) [Model 2]




A nonparametric regression program was used to fit a nonlinear model relating y and x. The nonparametric regression program can be obtained by e-mailing the second author. Bandwidths equal to 10% and 25% of the dynamic range of the x values were chosen. The predicted response curves are constructed using the predicted y values. Compound weights, $\delta_k \; w_k x_i$ are used as above to provide for estimates of the predicted values [yp for a 25% bandwidth and yp2 for a 10% bandwidth]. Down weighting the outlier produces a more robust fit (see Figure 4). The smoothing includes the down weighting of any outliers present.

**Figure 4.** Plot of Predicted Values and Observed Values (for chosen bandwidths of 10% and 25%) [Model 3]




Y - Actual observations ( • ), YP – Smoothing (---) with B = 25%, and YP2 - Smoothing (—) with B = 10%.

## CONCLUSIONS

The program performs a randomization test procedure for quadratic regression. The 2-tailed test for a non-zero quadratic effect is conducted. 'n_random' randomizations are carried out. The y values are reallocated to the x values to carry out these randomizations (see http://pages.prodigy.net/johnsonp12/rtest.html). A 95% confidence band around the predicted values was obtained and an optimal absorbance level was found. This optimum was further refined and improved upon by the fitting of a quadratic model. The wavelengths are chosen close to the values of the original optimum. Predicted values and observed values are plotted.

The significance level for testing for a non-zero quadratic effect is 0.002. At $\alpha = 0.05$, the significance level is less than 0.05, which indicates we have evidence to reject the null hypothesis in favor of the alternative. The randomization test thus indicates that there is evidence of a significant quadratic effect.

The optimum abundance from model 1 is 0.69 for a wavelength of 241.07 nm. The optimum abundance from model 2 is 0.76 for a wavelength of 240.18 nm. The optimum abundance from nonparametric model 3 is found to be valued at 0.78 for a wavelength of 240.48 nm. Model 3 provides for a more accurate refinement and improvement.

## REFERENCES


Brown, W. H., and Foote, C. S. (2002). *Organic Chemistry, 3rd Edition*, Singapore: Brooks/Cole Thomson Learning

Manly, B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, New York: Chapman & Hall.

Neter, J., Wasserman, W., and Kutner, M.H. (1985). *Applied Linear Statistical Models, Second Edition*, Homewood, Illinois: Richard D. Irwin, Inc.

SAS Institute Inc. (1990). *SAS Procedures Guide: Basics, Version 6, 3rd Edition*, Cary, NC: SAS Institute Inc.

SAS Institute. Inc. (1991). *SAS/GRAPH User's Guide, Version 6*, *3rd Edition*, Cary, NC: SAS Institute. Inc.

SAS Institute. Inc. (1993). *SAS/STAT User's Guide, Version 6, 4th Edition, Volume 1 and Volume 2,* Cary, NC: SAS Institute Inc.


## CODE

Program Name: RTEST
Randomization Procedure to Test for a Non-Zero Quadratic Effect

```
OPTIONS NODATE NONUMBER PAGESIZE = 60 LINESIZE =80 NONOTES;
goptions cback=white colors=(black cyan yellow green blue magenta);

data ctrl;input ind x y @@;cards;

1 205 0.14 2 210 0.18 3 215 0.25 4 220 0.35 5 230 0.60 6 235 0.73 7 245 0.72 8 250 0.60
9 260 0.10 10 270 0.005 11 225 0.50 12 240 0.79 13 241 0.79 14 239 0.77 15 238 0.76
16 242 0.78 17 201 . 18 211 . 19 221 . 20 231 . 21 251 . 22 261 . 23 271 .
         ;

data ctrl;set ctrl;x2=x*x;
proc reg data = ctrl outest = ma covout;
model y = x x2; output out = b pred=pred l95=l95 u95=u95;
TITLE1 'Randomization-Test Results for the Quadratic Regression Model';

proc sort data = b;by x;
data b2;set b;if y = . then delete;

proc print data = b2;
TITLE1 'Predicted and 95% Confidence Bound for the Individual Prediction';

data b3;set b;if y = .;
proc print data = b3;
TITLE1 'Predicted and 95% Confidence Bound for the Individual Prediction';
TITLE2 'Additional Points added with missing y values';
data mc;set ma;if _type_ = 'PARMS';keep x2 indx;indx = _N_;
       data mc;set mc;beta = x2;drop x2;

                proc transpose data = ma out = mb;var x2;

                       data mb;set mb;indx=_N_;var = x2;drop x2;
```

```
              proc sort data = mb;by indx; proc sort data = mc;by indx;
         data md;merge mb mc;by indx; data md;set md;t = beta/sqrt(var);

   proc print data=md;var beta var t;
   TITLE1 'Parameter Estimates for Fitting Least Squares Model to the Data';
   data ctrl;set ctrl;ind=_N_;

   %macro random;

     %let n_random= 999;
       %do i=1 %to &n_random;

       data cb&i;set ctrl;drop ind x;
          seed=floor(ind+&i+1000000000*(sqrt(time())-floor(sqrt(time()))));
                  k=500*(ranuni(seed));

                       proc sort data = cb&i;by k; data cb&i;set cb&i; ind=_N_;
                            proc sort data = ctrl;by ind; proc sort data = cb&i;by ind;
                                    data acb&i; merge ctrl cb&i;by ind; x2=x*x;

                                                proc reg data = acb&i outest = ma&i covout
noprint;model y = x x2;

                                                data mc&i;set ma&i;if _type_ = 'PARMS'; keep x2 indx;indx
= _N_;
                                  data mc&i;set mc&i;beta = x2;drop x2;

                              proc transpose data = ma&i out = mb&i;var x2;
                     data mb&i;set mb&i;indx=_N_;var = x2; drop x2;
                  proc sort data = mb&i;by indx; proc sort data = mc&i;by indx;

     data md&i;merge mb&i mc&i;by indx; data md&i;set md&i;t = beta/sqrt(var);

     %end;

     %do j = 2 %to &n_random;
     proc append base = md1 data = md&j;
     %end;

   %mend random;

   %random;
   data md;set md;dummy = 'a'; t2 = t;keep t2 dummy;
   data md1;set md1;dummy = 'a';keep t dummy;

   proc sort data = md;by dummy; proc sort data = md1;by dummy;

   data mf;merge md md1;by dummy;
       if abs(t)>=abs(t2) then ind_t =1;else ind_t = 0;
             proc means data = mf noprint;var ind_t;
                  output out = mdmc sum = sumlevel n = n_random;

   proc sort data = mf;by t;
   proc print data = mf;var t;
   TITLE2 'Randomization Results for Testing for a Non-Zero Quadratic Effect';

   data mdmc;set mdmc;siglevel =2*((sumlevel+1)/(n_random+1));n_2 = sumlevel;

   proc print data = mdmc;var siglevel n_random n_2;
   TITLE1 'Significance Level and Number of Randomizations Carried Out';
   TITLE2 'n_2 indicates the number of randomizations where the |t| value';
   TITLE3 'is greater than or equal to the original |t| (2-tailed alternative)';
   proc gplot data = b2;
       plot pred*x=1 u95*x=2 l95*x=3/overlay caxis=black ;
            symbol1 interpol = join color = black line=1 ;
               symbol2 interpol=join color = black line = 2 ;
                   symbol3 interpol=join color = black line=2 ;
                         TITLE1 'Plot of Predicted and 95% Confidence Bound';

   proc gplot data = b2;
        plot y*x pred*x=2 /overlay caxis=black ;
             symbol1 interpol=none color = black v=- h=.5;
                  symbol2 interpol=join color = black line = 2 v='';
                        TITLE1 'Plot of Predicted and Observed Values';
```

## ACKNOWLEDGMENTS

The authors would like to thank Michael Wilson, MWSUG Section chair, for his insightful comments to improve the quality of this article.

## CONTACT INFORMATION

Ling Huang,
Sacramento City College, Science and Allied Health,
3835 Freeport Blvd., Sacramento CA 95822-1386
E-Mail: HuangL@scc.losrios.edu

Paul Johnson,
P.O. Box 4146, Davis CA 95617-4146
E-Mail: PJohnson@biostatsoftware.com
Fax: (518) 723-1665

## APPENDIX

### APPENDIX A: DATA

| x | y |
|---|---|
| 205 | 0.140 |
| 210 | 0.180 |
| 215 | 0.250 |
| 220 | 0.350 |
| 230 | 0.600 |
| 235 | 0.730 |
| 245 | 0.720 |
| 250 | 0.600 |
| 260 | 0.100 |
| 270 | 0.005 |
| 225 | 0.500 |
| 240 | 0.790 |
| 241 | 0.790 |
| 239 | 0.770 |
| 238 | 0.760 |
| 242 | 0.780 |

y is the absorbance and x is the wavelength (nm).

### APPENDIX B: EQN 1

$x_{opt} = \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$ and $\hat{y}_{opt} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$.

Hence $\hat{y}_{opt} = \hat{\beta}_0 - \hat{\beta}_1\left(\frac{\hat{\beta}_1}{2\hat{\beta}_2}\right) + \hat{\beta}_2\left(\frac{\hat{\beta}_1}{2\hat{\beta}_2}\right)^2 = \hat{\beta}_0 - \left(\frac{\hat{\beta}_1^2}{2\hat{\beta}_2}\right) + \left(\frac{\hat{\beta}_{1\,1}^2}{4\hat{\beta}_2}\right) = \hat{\beta}_0 - \left(\frac{\hat{\beta}_1^2}{4\hat{\beta}_2}\right)$.