Solutions to Method Exam May 2014

1. For the production of printed figures, four computer systems (A, B, C, D) were tested. Four comparable sets of rough sketches (I, II, III, IV) were used by four operators (1, 2, 3, 4). The number, y, of figures completed per hour was recorded. The purpose of the study was to compare the four systems in terms of the average y values. The design was the following:

	I	II	III	IV
1	D	C	A	B
2	A	B	D	C
3	B	D	C	A
4	C	A	B	D

For the four systems, $\bar{y}_A = 1.8$, $\bar{y}_B = 2.6$, $\bar{y}_C = 2.1$, $\bar{y}_D = 1.9$. Part of the ANOVA table was given:

	SS	df
operators	2.16	df_1
sets	0.24	df_2
systems	1.52	df_3
residuals	0.90	df_4
total	ss_1	df_5

(a) Name the treatment variables and the block variables.

The treatment variable is the computer system.
The block variables are operator and set.

(b) What experimental design was employed?

Latin square design.

(c) Give the df1, df2, df3, df4, df5, ss1 values.

df1=df2=df3=3, df4=6, df5=15, ss1=4.82

(d) Without computing the p-value, can you say there is a significant difference among the four systems? Why?

The largest difference between any two levels of systems is 2.6-1.8=0.8 and the MSE=0.90/6=0.15. So there is a very small probability that all these means can be covered by the same t distribution with scale factor of 0.194(=sqrt(MSE/4)).

(e) If the sum of squares for systems indicates a significant difference at level 0.01, does it imply that each of the 6 pairs of systems are significantly different at level 0.01?

Explain.

No, it implies at least one pair of the systems are significantly different at level 0.01.

(f) Suppose there is only one operator instead of four. Treat the four rows 1,2,3,4 as from the same operator. What experimental design would it refer to?

RCBD – randomized complete blocking design.

(g) Even further, suppose there is only one set of rough sketches instead of four based on (f). Treat the four column I, II, III< IV as from the same set. What experimental design would it refer to? How many replicates are there?

CRD- completely randomized design. 4

2. A nickel-titanium alloy is used to make components for jet turbine aircraft engines. Cracking is a potentially serious problem in the final part, as it can lead to non-recoverable failure. A test is run at the parts producer to determine the effects of four factors on cracks. The four factors are pouring temperature (*A*), titanium content (*B*), heat treatment method (*C*), and the amount of grain refiner used (*D*). Suppose that only 16 runs could be made on a single day, so each replicate was treated as block. The length of crack (in μm) induced in a sample coupon subjected to a standard test is measured. The data are shown below (also stored in a USB):

A	*B*	*C*	*D*	Treatment Combination	Replicate I	Replicate II
-	-	-	-	(1)	7.037	6.376
+	-	-	-	*a*	14.707	15.219
-	+	-	-	*b*	11.635	12.089
+	+	-	-	*ab*	17.273	17.815
-	-	+	-	*c*	10.403	10.151
+	-	+	-	*ac*	4.368	4.098
-	+	+	-	*bc*	9.360	9.253
+	+	+	-	*abc*	13.440	12.923
-	-	-	+	*d*	8.561	8.951
+	-	-	+	*ad*	16.867	17.052
-	+	-	+	*bd*	13.876	13.658
+	+	-	+	*abd*	19.824	19.639
-	-	+	+	*cd*	11.846	12.337
+	-	+	+	*acd*	6.125	5.904
-	+	+	+	*bcd*	11.190	10.935
+	+	+	+	*abcd*	15.653	15.053

(a) Estimate the factor effects. Which factors appear to be large?

Term	Effect
A	3.01888
B	3.97588
C	-3.59625

D	1.95775
AB	1.93412
AC	-4.00775
AD	0.0765
BC	0.096
BD	0.04725
CD	-0.076875
ABC	3.1375
ABD	0.098
ACD	0.019125
BCD	0.035625
ABCD	0.014125

In this case A, B, C, D, AB, AC, ABC appear to be large.

(b) Conduct an analysis of variance. Do any of the factors affect cracking? Use α=0.05.

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F
day	0.016	1	0.016		
Model	570.95	15	38.06	445.11	< 0.0001
A	*72.91*	*1*	*72.91*	*852.59*	*< 0.0001*
B	*126.46*	*1*	*126.46*	*1478.83*	*< 0.0001*
C	*103.46*	*1*	*103.46*	*1209.91*	*< 0.0001*
D	*30.66*	*1*	*30.66*	*358.56*	*< 0.0001*
AB	*29.93*	*1*	*29.93*	*349.96*	*< 0.0001*
AC	*128.50*	*1*	*128.50*	*1502.63*	*< 0.0001*
AD	*0.047*	*1*	*0.047*	*0.55*	*0.4708*
BC	*0.074*	*1*	*0.074*	*0.86*	*0.3678*
BD	*0.018*	*1*	*0.018*	*0.21*	*0.6542*
CD	*0.047*	*1*	*0.047*	*0.55*	*0.4686*
ABC	*78.75*	*1*	*78.75*	*920.92*	*< 0.0001*
ABD	*0.077*	*1*	*0.077*	*0.90*	*0.3582*
ACD	*2.926E-003*	*1*	*2.926E-003*	*0.034*	*0.8557*
BCD	*0.010*	*1*	*0.010*	*0.12*	*0.7352*
ABCD	*1.596E-003*	*1*	*1.596E-003*	*0.019*	*0.8931*
Residual	1.28	15	0.086		
Cor Total	572.25	31			

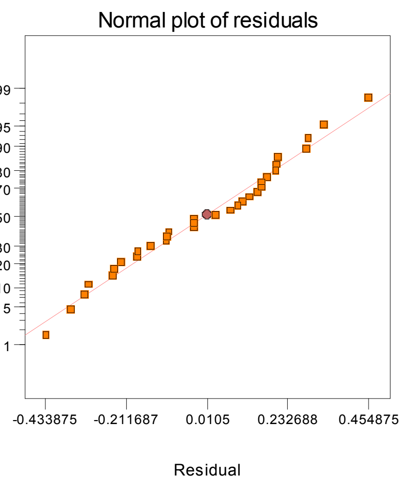
A, B, C, D, AB, AC, and ABC are significant effects.

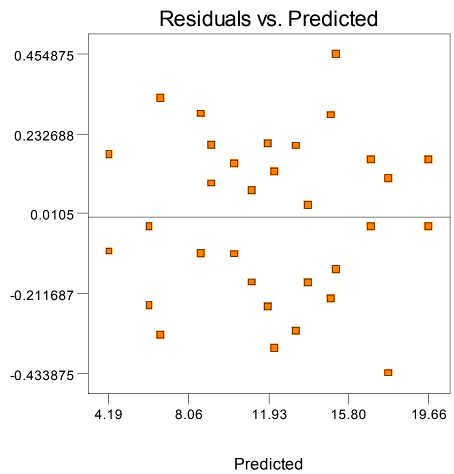
(c) Write down a regression model that can be used to predict crack length as a function of the significant main effects and interactions you have identified in part (b).

```
Crack Length=11.99

        +1.51  *A
        +1.99  *B
         -1.80  *C
        +0.98  *D
        +0.97  *A*B
         -2.00  *A*C
        +1.57  * A * B * C
```

(d) Analyze the residuals from this experiment.





There is nothing unusual about the residuals.

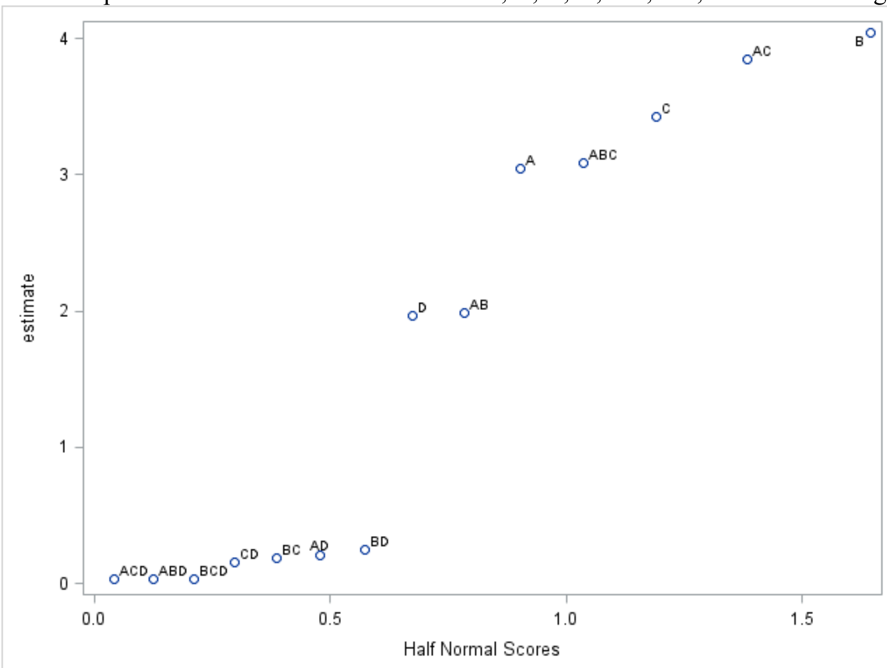
(e) Consider the data from the first replicate. Suppose that two operators run these 16 observations. Set up a design to run these observations in two blocks (i.e., operators) with 8 observations each. Which effect should be confounded? Analyze the data.

ABCD should be confounded.

A	*B*	*C*	*D*	*ABCD*	operator
-	-	-	-	+	1
+	-	-	-	-	2
-	+	-	-	-	2
+	+	-	-	+	1
-	-	+	-	-	2
+	-	+	-	+	1
-	+	+	-	+	1
+	+	+	-	-	2
-	-	-	+	-	2
+	-	-	+	+	1
-	+	-	+	+	1
+	+	-	+	-	2
-	-	+	+	+	1

+	-	+	+	-	2
-	+	+	+	-	2
+	+	+	+	+	1

From the half normal plot it is obvious that the effects of A, B, C, D, AB, AC, and ABC are significant.



(f) Attach your SAS/R code

```
data Q3;
  input A B C D day y;
  datalines;
-1  -1    -1    -1    -1    7.037
1   -1    -1    -1    -1    14.707
-1  1     -1    -1    -1    11.635
1   1     -1    -1    -1    17.273
-1  -1    1     -1    -1    10.403
1   -1    1     -1    -1    4.368
-1  1     1     -1    -1    9.36
1   1     1     -1    -1    13.44
-1  -1    -1    1     -1    8.561
1   -1    -1    1     -1    16.867
-1  1     -1    1     -1    13.876
1   1     -1    1     -1    19.824
-1  -1    1     1     -1    11.846
1   -1    1     1     -1    6.125
-1  1     1     1     -1    11.19
1   1     1     1     -1    15.653
-1  -1    -1    -1    1     6.376
1   -1    -1    -1    1     15.219
-1  1     -1    -1    1     12.089
1   1     -1    -1    1     17.815
```

```
-1  -1    1     -1    1     10.151
1   -1    1     -1    1     4.098
-1  1     1     -1    1     9.253
1   1     1     -1    1     12.923
-1  -1    -1    1     1     8.951
1   -1    -1    1     1     17.052
-1  1     -1    1     1     13.658
1   1     -1    1     1     19.639
-1  -1    1     1     1     12.337
1   -1    1     1     1     5.904
-1  1     1     1     1     10.935
1   1     1     1     1     15.053
;

data inter;                               /* Define Interaction Terms
*/
 set Q3;
 AB=A*B; AC=A*C; AD=A*D; BC=B*C; BD=B*D; CD=C*D; ABC=AB*C; ABD=AB*D;
 ACD=AC*D; BCD=BC*D; ABCD=ABC*D;

proc glm data=inter;                    /* GLM Proc to Obtain Effects
*/
 class A B C D AB AC AD BC BD CD ABC ABD ACD BCD ABCD day;
 model y=A B C D AB AC AD BC BD CD ABC ABD ACD BCD ABCD day;
 estimate 'A' A -1 1; estimate 'B' B -1 1;
  estimate 'C' C -1 1; estimate 'D' D -1 1;
 run;

proc reg outest=effects data=inter;    /* REG Proc to Obtain Effects
*/
 model y=A B C D AB AC ABC;
 run;

data part;
set Q3;
if day=1 then delete;
run;
proc print data=part;
run;

data part2;
set part;
 AB=A*B; AC=A*C; AD=A*D; BC=B*C; BD=B*D; CD=C*D; ABC=AB*C; ABD=AB*D;
 ACD=AC*D; BCD=BC*D; operator=ABC*D;
proc print data=part2;
 run;

proc reg outest=effects data=part2;
model y=A B C D AB AC AD BC BD CD ABC ABD ACD BCD operator;
run;

proc print data=effects;
run;
```

```
data effect2;
set effects;
drop _MODEL_ _TYPE_ _DEPVAR_ intercept y _RMSE_;
run;
proc print data=effect2;
run;


proc transpose data=effect2 out=effect3;
run;

data effect4;
set effect3;
effect=col1*2;
run;

data effect5;
set effect4;
where _NAME_ ^='operator';
run;

data effect6;
set effect5;
estimate=abs(effect);
run;

proc sort data=effect5;
by effect;
run;
proc print data=effect6;
run;



proc rank data=effect6 out=hnplots;
var estimate;
ranks rnk;
run;
proc print data=hnplots;
run;

/* the number 15 below is for 2^4 design, you need to change it for
your design: 7 for 2^3 design, 31 for 2^5 design ... */
data hnplots;
set hnplots;
zscore=probit((((rnk-0.5)/15)+1)/2);
run;
proc print data=hnplots;
run;


proc sgplot data=hnplots;
scatter x=zscore y=estimate/datalabel=_NAME_;
```

```
xaxis label='Half Normal Scores';
run;
proc reg outest=effects data=part2;
model y=A B C D AB AC ABC;
run;
```

3. A rocket propellant manufacturer is studying the burning rate of propellant from three production processes. Four batches of propellant are randomly selected from the output of each process and three determinations of burning rate are made on each batch. The results follow (data is provided in a USB).

	Process 1				Process 2				Process 3			
Batch	1	2	3	4	1	2	3	4	1	2	3	4
	25	19	15	15	19	23	18	35	14	35	38	25
	30	28	17	16	17	24	21	27	15	21	54	29
	26	20	14	13	14	21	17	25	20	24	50	33

(a) What design is this?

Nested design.

(b) Write the statistical model with assumptions.

$Y_{ijk}=u+\tau_i+\alpha_{j(i)}+\varepsilon_{k(ij)}$

τ represents the process effect, which is a fixed effect, $\sum\tau_i=0$,

α represents the batch effect, which is a random effect, $\alpha_{j(i)}\sim N(0, \sigma_\alpha^2)$.

$\varepsilon_{k(ij)}$ iid $\sim N(0, \sigma^2)$.

(c) Conduct an analysis of variance. Do any of the factors affect burning rate? Use α=0.05.

Analysis of Variance for Burn Rat

Source	DF	SS	MS	F	P
Process	2	676.06	338.03	1.46	0.281
Batch(Process)	9	2077.58	230.84	12.20	<0.0001
Error	24	454.00	18.92		
Total	35	3207.64			

There is a significant batch effect.

(d) What is the hypothesis for testing batch effect?

H_0: $\sigma_\alpha^2 = 0$

H_1: $\sigma_\alpha^2 > 0$

(e) What is the hypothesis for testing process effect?

H_0: $\tau_1=\tau_2=\tau_3=0$

H_1: at least one $\tau_i \neq 0$

(f) Estimate the variation for the batch factor and construct 95% confidence interval for it.

$\sigma_\alpha^2=(MS_{batch(process)}-MSE)/n=(230.84-18.92)/3=70.64$.

95% confidence interval is: [31.6664, 272.26]

(g) Attach your SAS/R code

```
input Process  Batch  Rate;
datalines;
1    1    25
1    1    30
1    1    26
2    1    19
2    1    17
2    1    14
3    1    14
3    1    15
3    1    20
1    2    19
1    2    28
1    2    20
2    2    23
2    2    24
2    2    21
3    2    35
3    2    21
3    2    24
1    3    15
1    3    17
1    3    14
2    3    18
2    3    21
2    3    17
3    3    38
3    3    54
3    3    50
1    4    15
1    4    16
1    4    13
2    4    35
2    4    27
2    4    25
3    4    25
3    4    29
3    4    33
;
proc mixed data=rocket CL;
class Process Batch;
model Rate=Process;
random Batch(Process);
run;
```

4. A study of the effects of a drug on reducing cell damage was conducted. The data were:
   x = treatment duration (months)
   Y = chromosome aberrations rate (per 100 cells)
   and are as follows (also see the file chromosome.csv):

x	Y	x	Y	x	Y	x	Y	x	Y
0	2.12379	1	1.28013	3	1.52619	6	2.14625	12	1.28915
0	1.95616	1	1.64591	3	1.59758	6	1.18598	12	0.87130
0	2.08874	1	2.26501	3	2.32749	6	0.69639	12	1.08913
0	2.26963	1	1.89014	3	1.25688	6	1.34690	12	1.22714
0	1.22742	1	0.96558	3	1.59787	6	1.61262	12	0.94220
0	1.70102	1	1.79096	3	1.20665	6	1.49434	12	1.23307
0	1.74595	1	1.55711	3	0.94922	6	1.78314	12	0.94560
0	1.76509	1	1.96382					12	0.86425
0	2.34254	1	1.85106					12	1.30250
0	2.28392	1	1.83534					12	1.11911
0	2.09915	1	1.99476					12	1.14881
0	1.61760	1	1.81533					12	1.43601
0	1.87913	1	1.23379					12	1.09781
0	1.78825	1	1.51438					12	1.14958
0	1.76288	1	2.03380					12	1.16802
0	2.18477	1	0.90620					12	1.28875
0	2.01024	1	1.63041					12	1.31049
0	2.51135							12	1.26559
0	2.56572							12	1.00396
0	2.59723								
0	2.43926								
0	2.65349								
0	1.88811								
0	1.94380								
0	1.94836								
0	2.52894								
0	1.58083								
0	1.83397								
0	2.03775								
0	1.96894								
0	2.00361								

(Chromosome aberrations are a form of genetic damage.) The authors reported that "...frequency of chromosome [aberrations] dropped significantly with time elapsed.... This dependency may be described via the equation $Y = 1.9 - 0.07x$ ($R^2 = 0.46$)."

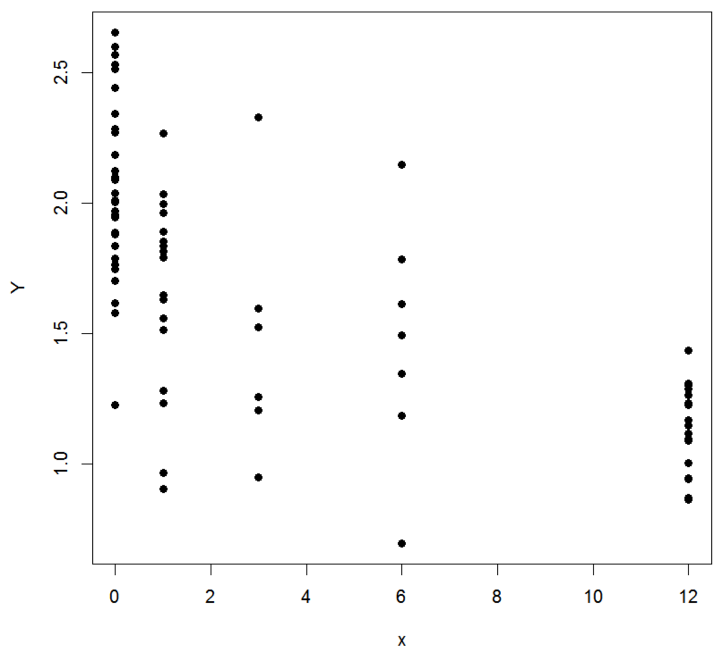
(a) Assume that the data are normally distributed. Comment on the authors' assertion. Is their analysis of these data reasonable?

(b) Provide a further analysis of these data to adjust for any problems you noted in part (a). Is it still reasonable to argue that frequency of chromosome aberrations drops significantly with time elapsed (at a false positive rate of 5%)?

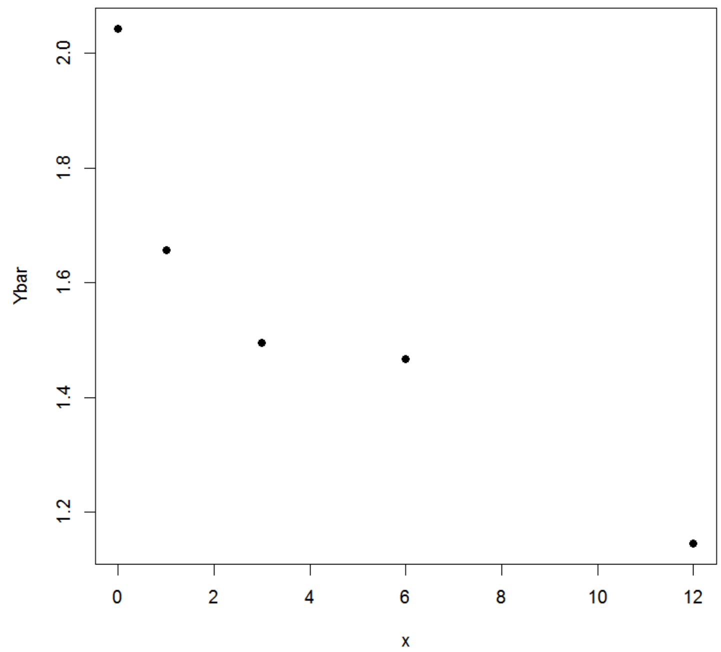
A4. (a) A plot of the data show that Y does clearly decrease with x, and the stated regression statistics are correctly calculated ($P \leq 0.0001$). A closer look at the plot (in particular, a plot of $\bar{Y}$ vs. x) reveals, however, that the trend may deviate from strictly linearity. A residual plot suggests clear heterogeneous variance, as well.

Sample R code:

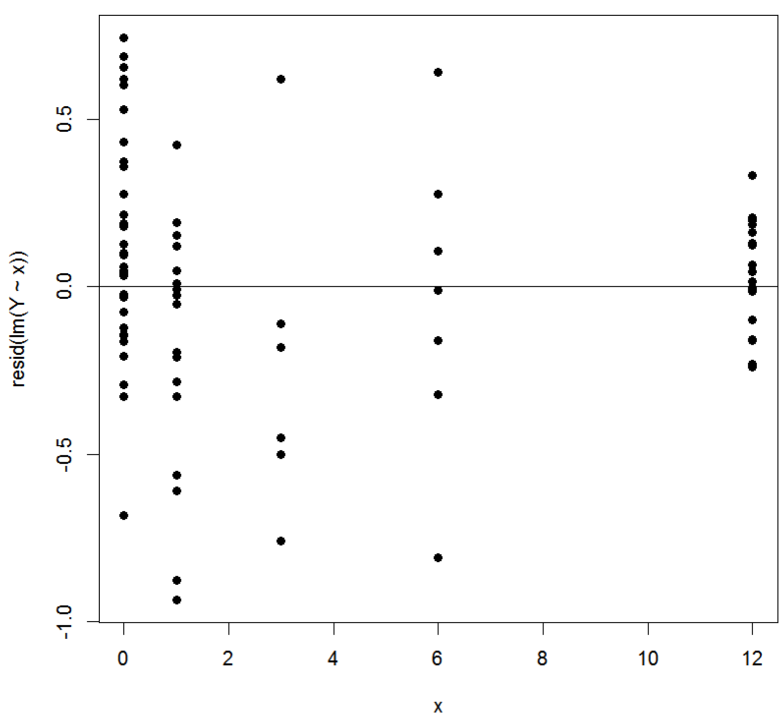
```
chromosome.df = read.csv( file.choose() )
attach( chromosome.df )
plot( Y~x, pch=19 )
```




```
plot( by( data=chromosome.df$Y, INDICES=x, FUN=mean )~unique(x),
pch=19,
                                                  xlab='x', ylab='Ybar')
```



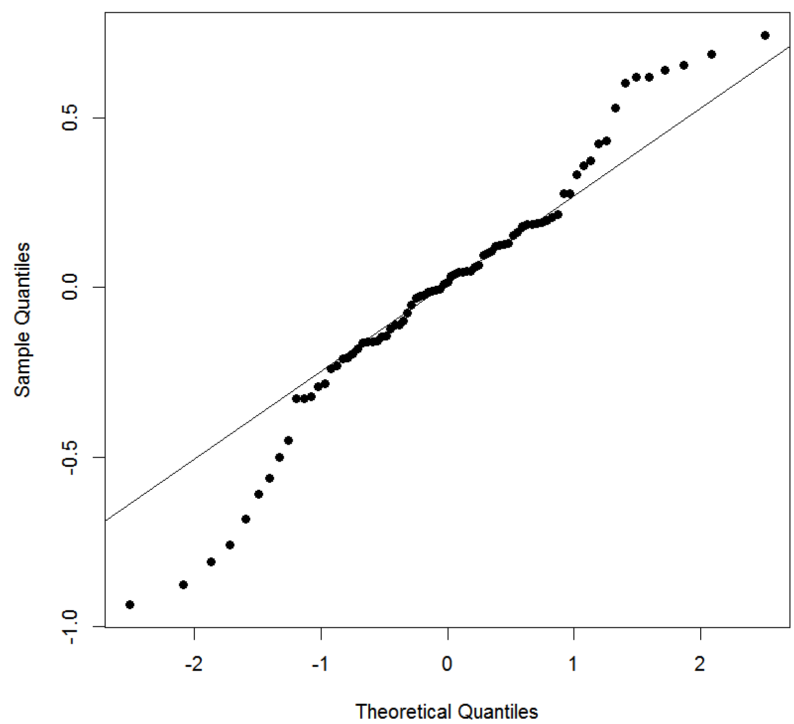

```
plot( resid(lm(Y~x))~x, pch=19 ); abline( h=0 )
```



Plotting residuals against fitted values would give a similar conclusion, although because the trend is negative, the pattern would be reversed. [Note that a Brown-Forsythe test on these residuals would have very low power, since only the last collection of residuals at the single highest *x*-value appears to be heterogeneous. The graphical diagnostic gives a stronger indication of the violation here. If a formal inference is desired, the older Bartlett test here gives a *P*-value of 0.0048 against constant variance, using, e.g., `bartlett.test(Y~x)`.]

A normal quantile plot of the residuals also gives one pause:

```
qqnorm( resid(lm(Y~x)), pch=19, main='' )
qqline( resid(lm(Y~x)))
```

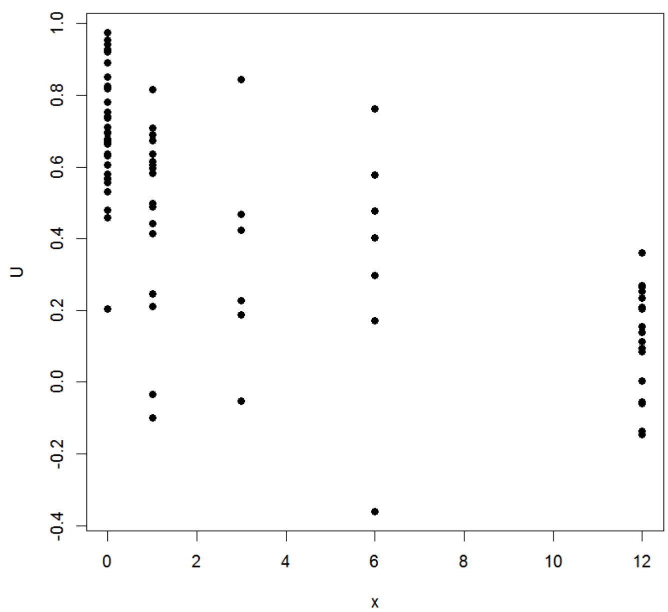


The quantile plot suggests nontrivial deviation from normality, with a potential skew to the right.

(b) A more-complete analysis would note the nonlinearity and the variance heterogeneity. In the former case, scientifically it would seem reasonable that the *rate* of chromosome damage would ameliorate over time, suggesting perhaps an exponential decay. A nonlinear regression fit using $E[Y] = \exp(\beta_0 + \beta_1 x)$ would be appropriate. To approximate this with linear regression methods, transform to $U = \log(Y)$ (which will also help account for the possible right skew in the data) and model $E[U] = \beta_0 + \beta_1 x$. To account for variance heterogeneity, use the replicated information at each x_i: find the per-time sample variances s_i^2 of the U_is and take as weights $w_i = 1/s_i^2$.

Start with a plot of the transformed data:

```
U = log(Y)
plot( U~x, pch=19 )
```

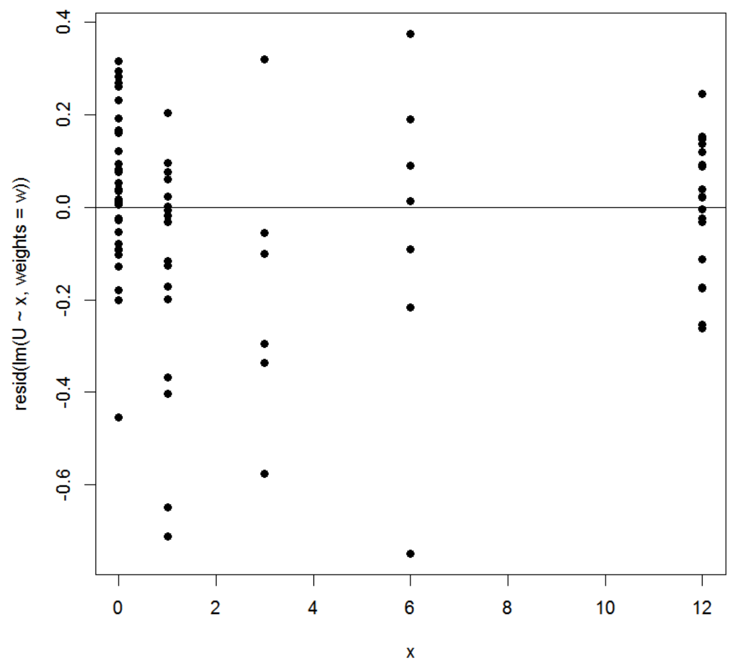


The decreasing trend remains, but so too does the variance heterogeneity. Now find the weights:

```
s2 = as.numeric( by( data=U, INDICES=x, FUN=var ) )
ni = as.numeric( by( data=U, INDICES=x, FUN=length ) )
w = 1/c( rep(s2[1],ni[1]), rep(s2[2],ni[2]), rep(s2[3],ni[3]),
              rep(s2[4],ni[4]), rep(s2[5],ni[5]) )
```

A weighted least squares/simple linear regression of U on x gives the following residual plot:

```
plot( resid(lm(U~x,weights=w))~x, pch=19 ); abline( h=0 )
```



The variance heterogeneity remains (no surprise, but now the weighting adjusts for this). A test for decreasing response over time tests $H_o:\beta_1 = 0$ vs. $H_a:\beta_1 < 0$. The call to

```
newfit.lm = lm( U~x, weights=w )
summary( newfit.lm )
```

shows the summary statistics as (edited):

```
Call:
lm(formula = U ~ x, weights = w)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.659544   0.028357   23.26   <2e-16
x           -0.045243   0.003795  -11.92   <2e-16

Residual standard error: 1.038 on 79 degrees of freedom
Multiple R-squared:  0.6428,    Adjusted R-squared:  0.6383
F-statistic: 142.2 on 1 and 79 DF,  p-value: < 2.2e-16
```

but for the one-sided test we need

```
tcalc = coef(newfit.lm)[2]/sqrt(vcov(newfit.lm )[2,2])
pt( tcalc, df=newfit.lm$df.residual,lower=T )
```

which gives

```
1.214723e-19
```

Since this *P*-value is far less than $\alpha = 0.5$, we see the trend is significant (!).

5. A study explored changes in body mass index (BMI) of North American models from the 1950s to the late 2000s. (BMI is a standardized measure that combines a person's weight in inches and height in pounds: BMI = 703×weight/height^2.) While most Western populations have seen increases in BMI over that time span, these models show a different pattern. The data comprise n = 609 data pairs, one from each independent model, and are available in the file bmi.csv; a sample follows:

Date:	Dec. 1953	Mar. 1954	Nov. 1954	⋯	Dec. 2008	Jan. 2009
BMI:	19.63408	19.04362	20.48249	⋯	17.48378	18.94921

(a) Plot Y = BMI against x = month. What pattern appears?

(b) Given the questions on the pattern of response, calculate a robust, linear, loess fit with soothing parameter set to q = 0.5. Overlay the loess fit on the scatterplot. Does this improve visualization of the pattern?

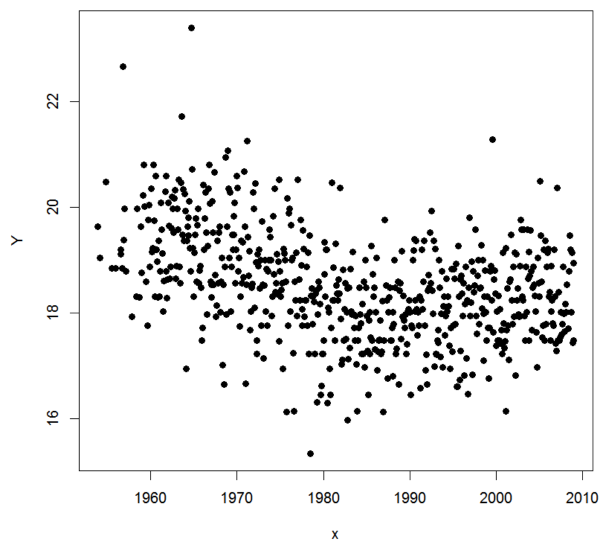
(c) From your loess fit, predict the (mean) BMI for x = Nov. 1989.

(d) Plot the residuals from the loess fit against x = month. Do any important patterns appear?

A5. (a) Sample R code for scatterplot:

```
q3.df = read.csv( file.choose() )
attach( q3.df )
month.digit = match( Month,month.abb )
date = paste( Year,month.digit,sep='-' )
library( zoo )  # load external 'zoo' package for as.yearmon function
month = as.yearmon(date)
x = as.numeric(month);
Y = BMI
plot( Y~x, pch=19 )
```

The pattern is generally decreasing (over the same time span, most Western cultures show an increasing BMI!), but obviously not in a linear fashion:



(b) Loess fit is conducted via

```
BMI1r.loess = loess( Y~x, span = 0.5, degree = 1, family='symmetric' )
```

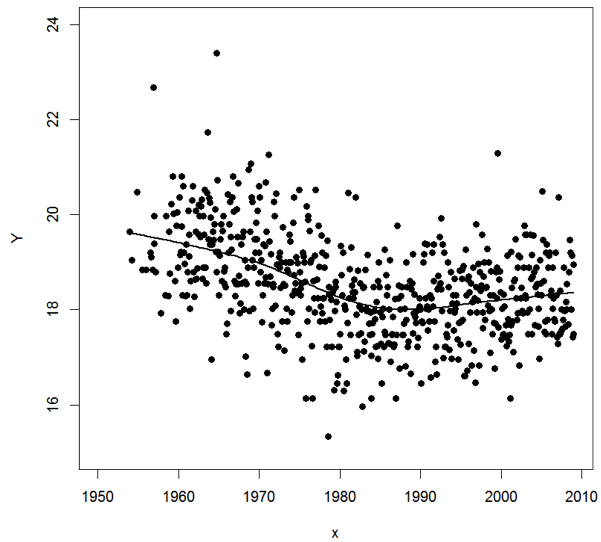
Smoothed predictions are found via

```
Ysmooth1r = predict( BMI1r.loess, data.frame(x = seq(1953,2009,.25) ))
```

Overlay plot via

```
plot( Y~x, pch=19, xlim=c(1950,2009), ylim=c(15,24) ); par( new=T )
plot( Ysmooth1r~seq(1953,2009,.25), type='l', lwd=2 , xaxt='n',
        yaxt='n' , xlab='', ylab='', xlim=c(1950,2009), ylim=c(15,24) )
```

The result visualizes better the decreasing pattern, and also highlights the nontrivial curvilinearity:



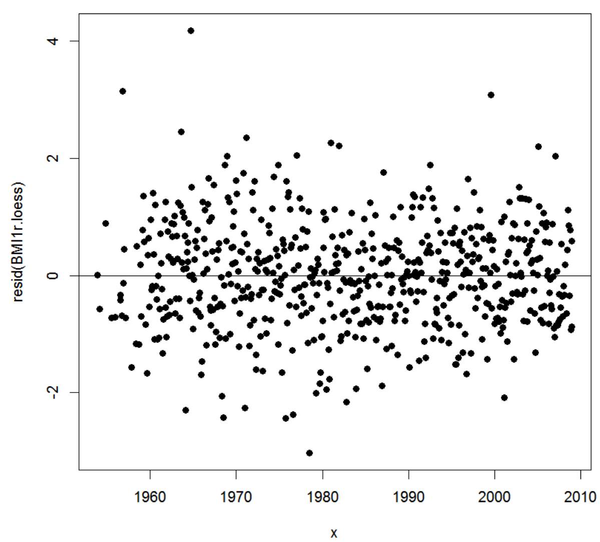
(c) Set x = 1989 + (10/12) = 1989.8303 for Nov. 1989. (Numeric months set ‘January’ to 0/12 and ‘December’ to 11/12 to keep them in the same ‘year’.) Find predicted value via

```
predict( BMI1r.loess, data.frame( x = 1989+(10/12) ))
```

which gives a value of $\hat{Y} = 18.00705$.

(d) Residual plot, using

```
plot( resid(BMI1r.loess)~x, pch=19 ); abline( h=0 )
```



appears randomly spread about $e = 0$, with perhaps a possible outlier near $e \approx 4.1$ in the early portion of the series.

6. From the simple linear regression (SLR) model with $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ($i = 1,...,n$) and $\varepsilon_i \sim$ i.i.d. $N(0,\sigma^2)$, let the ordinary least squares estimators be b_0 and b_1, with their usual properties. When the goal is prediction of a new observation $Y_{h(new)} = \beta_0 + \beta_1 x_h + \varepsilon_{h(new)}$ at some x_h, the predicted value is $\hat{Y}_h = b_0 + b_1 x_h$. Assume that $\varepsilon_{h(new)} \sim N(0,\sigma^2)$ is independent of the original Y_is. Prove that the ratio

$$t^* = \frac{Y_{h(new)} - \hat{Y}_h}{\sqrt{MSE \times \xi_h}}$$

is distributed as t with n–2 degrees of freedom, where

$$\xi_h = 1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}$$

with $S_{xx} = \sum(x_i - \bar{x})^2$ and where $MSE = S_{xx} = \sum(Y_i - \hat{Y}_i)^2/(n-2)$ is the usual mean square error of the regression fit using $\hat{Y}_i = b_0 + b_1 x_i$.

A6. This is Equation (2.35) from Kutner et al. (2005, ALRM, 4th ed.). Start with the numerator of t^*: $Y_{h(new)} - \hat{Y}_h = \beta_0 + \beta_1 x_h + \varepsilon_{h(new)} - b_0 - b_1 x_h = (\beta_0 - b_0) + (\beta_1 - b_1)x_h + \varepsilon_{h(new)}$. Each component here is itself normally distributed:

$(\beta_0 - b_0) \sim N(0,V_0)$ with $V_0 = \text{Var}\{b_0\} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$

$(\beta_1 - b_1) \sim N(0,V_1)$ with $V_1 = \text{Var}\{b_1\} = \sigma^2/S_{xx}$, and

$\varepsilon_{h(new)} \sim N(0,\sigma^2)$.

Thus the sum of the three is also normally distributed, with expected value

$$E[Y_{h(new)} - \hat{Y}_h] = E[\beta_0 - b_0] + x_h E[\beta_1 - b_1] + E[\varepsilon_{h(new)}] = 0 + (0)(x_h) + 0 = 0$$

and variance

$$\text{Var}\{Y_{h(new)} - \hat{Y}_h\} = \text{Var}\{(\beta_0 - b_0) + (\beta_1 - b_1)x_h + \varepsilon_{h(new)}\}.$$

For the latter term, recognize that $\varepsilon_{h(new)}$ is independent of b_0 and b_1, so

$$\text{Var}\{Y_{h(new)} - \hat{Y}_h\} = \text{Var}\{(\beta_0 - b_0) + (\beta_1 - b_1)x_h + \varepsilon_{h(new)}\}$$
$$= \text{Var}\{(\beta_0 - b_0) + (\beta_1 - b_1)x_h\} + \text{Var}\{\varepsilon_{h(new)}\}.$$

But now,

$$\text{Var}\{(\beta_0 - b_0) + (\beta_1 - b_1)x_h\} = \text{Var}\{\beta_0 - b_0\} + \text{Var}\{(\beta_1 - b_1)x_h\} + 2\text{Cov}\{\beta_0 - b_0, (\beta_1 - b_1)x_h\}$$

where the covariance is

$$\text{Cov}\{\beta_0 - b_0, (\beta_1 - b_1)x_h\} = x_h\text{Cov}\{\beta_0 - b_0, \beta_1 - b_1\} = x_h\text{Cov}\{-b_0, -b_1\} = x_h\text{Cov}\{b_0, b_1\}.$$

Since $\text{Cov}\{b_0, b_1\} = -\sigma^2\bar{x}/S_{xx}$, we have

$$\text{Cov}\{\beta_0 - b_0, (\beta_1 - b_1)x_h\} = -\sigma^2 x_h \bar{x}/S_{xx},$$

and thus, since

$$\text{Var}\{\beta_0 - b_0\} = (-1)^2\text{Var}\{b_0\} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

and

$$\text{Var}\{(\beta_1 - b_1)x_h\} = (-1)^2 x_h^2 \text{Var}\{b_1\} = x_h^2\sigma^2/S_{xx},$$

we see

$$\text{Var}\{(\beta_0 - b_0) + (\beta_1 - b_1)x_h\} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) + x_h^2\sigma^2/S_{xx} - 2\sigma^2 x_h\bar{x}/S_{xx}$$
$$= \sigma^2\left\{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} + \frac{x_h^2}{S_{xx}} - 2\frac{x_h^2\bar{x}^2}{S_{xx}}\right\} = \sigma^2\left\{\frac{1}{n} + \frac{\bar{x}^2 + x_h^2 - 2x_h^2\bar{x}^2}{S_{xx}}\right\} = \sigma^2\left\{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right\}$$

which is simply $\sigma^2(\xi_h - 1)$.

From this, we find

$$\text{Var}\{Y_{h(new)} - \hat{Y}_h\} = \text{Var}\{(\beta_0 - b_0) + (\beta_1 - b_1)x_h\} + \text{Var}\{\varepsilon_{h(new)}\} = \sigma^2(\xi_h - 1) + \sigma^2 = \sigma^2\xi_h.$$

Thus the standardized variate

$$Z = \frac{(Y_{h(new)} - \hat{Y}_h) - 0}{\sqrt{\text{Var}\{Y_{h(new)} - \hat{Y}_h\}}} = \frac{Y_{h(new)} - \hat{Y}_h}{\sigma\sqrt{\xi_h}}$$

is standard normal: $Z \sim N(0,1)$.

Now, recognize that since the MSE is known to be independent of b_0 and b_1 for this SLR model, and it is also assumed independent of $\varepsilon_{h(new)}$, then the MSE will be independent of Z, above (since the only random quantities in Z are these three variates). For that matter,

$$C^2 = \frac{(n-2)MSE}{\sigma^2} \sim \chi^2(n-2)$$

will also be independent of Z. But then from the definition of the *t* random variable, dividing Z by $\sqrt{C^2/(n-2)}$ produces a $t(n-2)$ variate. This is

$$t = \frac{Z}{\sqrt{\dfrac{(n-2)MSE}{(n-2)\sigma^2}}} = \frac{\left(\dfrac{Y_{h(new)} - \hat{Y}_h}{\sigma\sqrt{\xi_h}}\right)}{\sqrt{\dfrac{MSE}{\sigma^2}}} = \frac{(Y_{h(new)} - \hat{Y}_h)\sigma}{\sigma\sqrt{\xi_h}\sqrt{MSE}} = \frac{Y_{h(new)} - \hat{Y}_h}{\sqrt{\xi_h\,MSE}} = t^*$$

as desired.