Statistics GIDP
Ph.D. Qualifying Exam
Methodology

Jan 9th, 2015, 9:00am-1:00pm

Instructions: Provide answers on the supplied pads of paper; write on only one side of each sheet. Complete exactly 2 of the first 3 problems, and 2 of the last 3 problems. Turn in only those sheets you wish to have graded. You may use the computer and/or a calculator; any statistical tables that you may need are also provided. Stay calm and do your best; good luck.

1. An engineer is studying the mileage performance characteristics of five types of gasoline additives. In the road test he wishes to use cars as blocks; however, because of a time constraint, he just can run a design as follow.

	Car				
Additive	1	2	3	4	5
1		17	14	13	12
2	14	14		13	10
3	12		13	12	9
4	13	11	11	12	
5	11	12	10		8

(a) What design is this?

(b) State the statistical model and the corresponding assumptions.

(c) Are Type I and III sum of squares equal in the SAS output for the model y= additive + car + ε? Why?

(d) If you're given $\sum \hat{\tau}_i^2 = 9.5289$, what is $SS_{additive(adjusted)}$?

(e) Fill up the blanks in the ANOVA table below and draw conclusions at α=0.05.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Additive	_____	31.7000	_____	_____	_____	_____
Car	_____	35.2333	35.2333	8.8083	9.67	0.001
Error	_____	10.0167	10.0167	0.9106		
Total	_____	76.9500				

2. The surface finish of metal parts made on four machines is being studies. An experiment is conducted in which each machine is run by two different operators and two specimens from each operator are collected and tested. Because of the location of the machines, different operators are used on each machine, and the operators are chosen at random. The results follow (**dataset "finish.csv**" is provided).

	Machine 1		Machine 2		Machine 3		Machine 4	
Operator	1	2	1	2	1	2	1	2
	79	94	92	85	88	53	36	40
	62	74	99	79	75	56	53	56

(a) What design is this?
(b) Write the statistical model with assumptions.
(c) Conduct an analysis of variance. Do any of the factors affect finish? Use α=0.05.
(d) What is the hypothesis (in term of mathematical notation) for testing operator effect?
(e) What is the hypothesis (in term of mathematical notation) for testing machine effect?
(f) Estimate the variation for the operator factor and construct 95% confidence interval for it.
(g) Attach your SAS/R code

3. A soft drink bottler is interested in obtaining more uniform fill heights in the bottles produced by his manufacturing process. Three variables are checked, the percent carbonation (A), the operating pressure in the filler (B), and the bottles produced per minute or the line speed (C). The data are shown below (also provided as "**deviation.csv**").

	Coded Factors			Fill Height Deviation	
Run	*A*	*B*	*C*	Replicate 1	Replicate 2
1	-	-	-	-3	-1
2	+	-	-	0	1
3	-	+	-	-1	0
4	+	+	-	2	3
5	-	-	+	-1	0
6	+	-	+	2	1
7	-	+	+	1	1
8	+	+	+	6	5

	Factor Levels	
	Low (-1)	High (+1)
A (%)	10	12
B (psi)	25	30
C (b/m)	200	250

(a) What design is this?
(b) Analyze the data from this experiment. Which factors significantly affect fill height deviation
(c) Analyze the residuals from the model in (b). Are there any indications of model inadequacy
(d) Assume that two replicates are done by two operators, respectively. Re-analyze the data.
(e) Attach your SAS/R code

4. Consider the variables
G = 2011 GDP per capita (in 2000 dollars, inflation-adjusted), and
Y = 2011 Life expectancy at birth
among n = 149 different nations worldwide. Data are

Nation	GDP	life.expect
Algeria	2255.225	73.131
Angola	629.955	51.093
Argentina	11601.63	75.901
Armenia	1384.085	74.241
⋮	⋮	⋮
Venezuela	5671.912	74.402
Vietnam	757.401	75.181
Zimbabwe	347.746	51.384

(The full data are available in the file **gapminder.csv**.)

a) Per capita values such as GDP are notoriously skewed. Verify this by plotting a histogram for G. (Indicate which binning rule you use for your histogram bins.) If available, overlay a simple kernel density estimator.
b) Plot Y against X = log(GDP). What pattern appears?
c) Given the questions on the pattern of response, work with X = log(GDP) and calculate a robust, linear, loess fit of Y vs. X with smoothing parameter set to q = 0.7. Overlay the loess fit on the scatterplot. Does this improve visualization of the pattern?
d) From your loess fit, predict the (mean) Life Expectancy at a GDP of 15000.
e) Plot the residuals from the loess fit against X = GDP. Do any important patterns appear?

5. Show that for the simple linear model $Y_i \sim$ indep. $N(\beta_0 + \beta_1 X_i, \sigma^2)$, i = 1,...,n, the least squares estimator of β_0,

$$b_0 = \bar{Y} + \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2}\bar{X},$$

is unbiased.

6. The following data on arable land (km^2), birth rate (per 1000 popl'n), and outbound migration (per 1000 popl'n) were recorded among 14 Middle East nations in 1994 (also see the file **arable.csv**):

Country	Arable Land	Birth Rate	Migration
Bahrain	2	26.6	+6.8
Egypt	3	28.7	–0.4
Jordan	4	37.8	+0.5
Iran	8	42.4	0
Iraq	12	44.1	+0.4
Israel	17	20.5	+8.0
Kuwait	0	29.4	+25.4
Lebanon	20	27.9	–1.5
Oman	1	40.4	0
Qatar	0	18.8	+10.1
Saudi Arabia	1	38.3	0
Syria	28	43.7	0
U.A.E.	0	27.7	+23.3
Yemen	6	50.7	–2.4

Assuming the Migration data are normally distributed, conduct a multiple linear regression on Y = Migration with predictors Arable Land, Birth Rate, and their interaction. Assess whether and how these variable may affect the Migration outcome. Be as complete as possible. For any inferences, set $\alpha = 0.10$.

Solutions to Method Exam – 2015 Jan

1. An engineer is studying the mileage performance characteristics of five types of gasoline additives. In the road test he wishes to use cars as blocks; however, because of a time constraint, he just can run a design as follow.

	Car				
Additive	1	2	3	4	5
1		17	14	13	12
2	14	14		13	10
3	12		13	12	9
4	13	11	11	12	
5	11	12	10		8

(a) What design is this?

BIBD (balanced incomplete block design)

(b) State the statistical model and the corresponding assumptions.

$$y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ij}, i = 1,...,5; j = 1,...,5$$
$$\sum \tau_i = 0, \sum \beta_j = 0, \varepsilon_{ij} \sim N(0, \sigma^2)$$

(c) Are Type I and III sum of squares equal in the SAS output for the model y= additive + car + ε? Why?

No, as the orthogonality does not hold.

(d) If you're given $\sum \hat{\tau}_i^2 = 9.5289$, what is $SS_{additive(adjusted)}$?

$$SS_{additive(adjusted)} = \frac{\lambda a}{k} \sum \hat{\tau}_i^2 = 3*5*9.5289/4 = 35.73338$$

(e) Fill up the blanks in the ANOVA table below and draw conclusions at α=0.05.

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Additive	4	31.7000	35.733	8.9333	9.8104	0.0012
Car	4	35.2333	35.2333	8.8083	9.67	0.001
Error	11	10.0167	10.0167	0.9106		
Total	19	76.9500				

Both car and additive are significant at α=0.05

2. The surface finish of metal parts made on four machines is being studies. An experiment is conducted in which each machine is run by two different operators and two specimens from each operator are collected and tested. Because of the location of the machines,

different operators are used on each machine, and the operators are chosen at random. The results follow (dataset "finish.csv" is provided).

	Machine 1		Machine 2		Machine 3		Machine 4	
Operator	1	2	1	2	1	2	1	2
	79	94	92	85	88	53	36	40
	62	74	99	79	75	56	53	56

(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 machine effect, which is a fixed effect, $\sum\tau_i=0$,
α represents the operator 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 finish? Use α=0.05.

Type 1 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Error Term	Error DF	F Value	Pr > F
Machine	3	4046.187500	1348.729167	MS(Operator(Machine))	4	8.10	0.0356
Operator(Machine)	4	665.750000	166.437500	MS(Residual)	8	1.93	0.1982
Residual	8	688.500000	86.062500	.	.	.	.

There is a significant effect for machine, but not operator.

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

H_0: $\sigma_\alpha^2 = 0$
H_1: $\sigma_\alpha^2 > 0$

(e) What is the hypothesis for testing machine 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 operator factor and construct 95% confidence interval for it.

Covariance Parameter Estimates

Cov Parm	Estimate	Alpha	Lower	Upper
Operator(Machine)	40.1875	0.05	-82.6133	162.99

(g) Attach your SAS/R code

```
data new;
input Finish      Machine     Operator;
datalines;
79     1      1
62     1      1
94     1      2
74     1      2
92     2      1
99     2      1
85     2      2
79     2      2
68     3      1
75     3      1
53     3      2
56     3      2
36     4      1
53     4      1
40     4      2
56     4      2
;

proc mixed data=new method=type1 CL;
class Machine Operator;
model Finish=Machine;
random Operator(Machine);
run;
```

3. A soft drink bottler is interested in obtaining more uniform fill heights in the bottles produced by his manufacturing process. Three variables are checked, the percent carbonation (A), the operating pressure in the filler (B), and the bottles produced per minute or the line speed (C). The data are shown below (also provided as "deviation.csv").

	Coded Factors			Fill Height Deviation	
Run	*A*	*B*	*C*	Replicate 1	Replicate 2
1	-	-	-	-3	-1
2	+	-	-	0	1
3	-	+	-	-1	0
4	+	+	-	2	3
5	-	-	+	-1	0
6	+	-	+	2	1
7	-	+	+	1	1
8	+	+	+	6	5

	Factor Levels	
	Low (-1)	High (+1)
A (%)	10	12
B (psi)	25	30
C (b/m)	200	250

(a) What design is this?

Factorial design (or 2^3 factorial design).

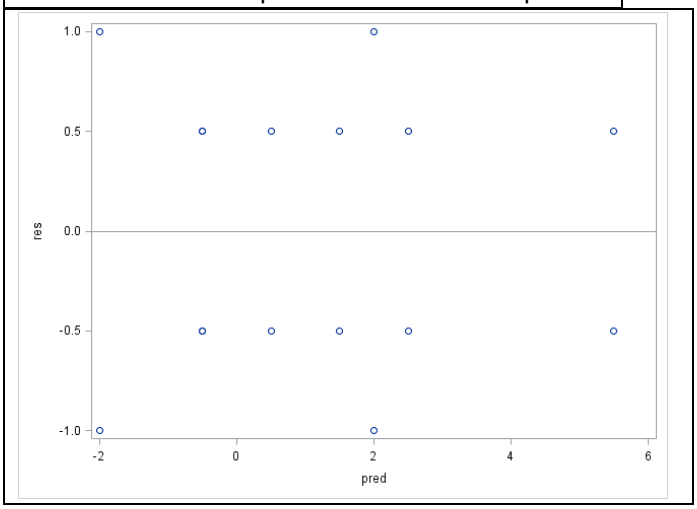
(b) Analyze the data from this experiment. Which factors significantly affect fill height deviation?

The analysis of variance in the Design Expert output below shows that factors *A*, *B*, and *C* are significant.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	1	30.25000000	30.25000000	34.57	0.0004
B	1	25.00000000	25.00000000	28.57	0.0007
A*B	1	1.00000000	1.00000000	1.14	0.3162
C	1	16.00000000	16.00000000	18.29	0.0027
A*C	1	0.00000000	0.00000000	0.00	1.0000
B*C	1	2.25000000	2.25000000	2.57	0.1475
A*B*C	1	0.25000000	0.25000000	0.29	0.6075

(c) Analyze the residuals from the model in (b). Are there any indications of model inadequacy?

There is no unusual pattern in the residual plot.



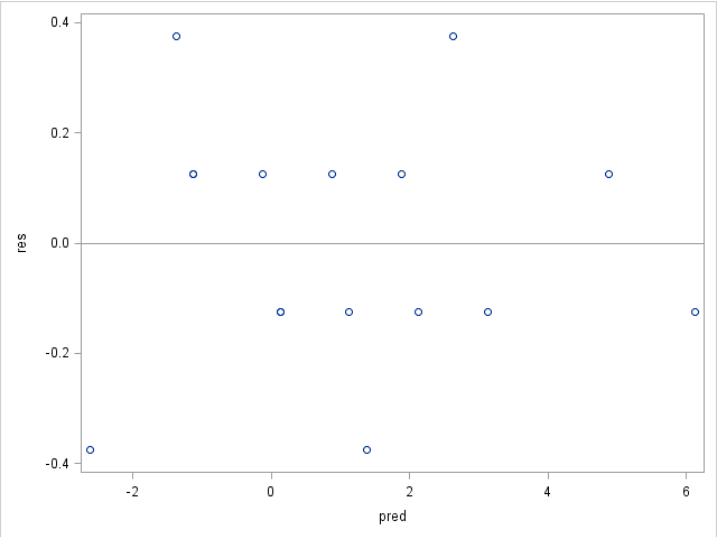
(You need to generate the QQ plot also and provide it here and make comments!).

(d) Assume that two replicates are done by two operators respectively, re-analyze the data.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
rep	1	6.25000000	6.25000000	58.33	0.0001
A	1	30.25000000	30.25000000	282.33	<.0001
B	1	25.00000000	25.00000000	233.33	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A*B	1	1.00000000	1.00000000	9.33	0.0185
C	1	16.00000000	16.00000000	149.33	<.0001
A*C	1	0.00000000	0.00000000	0.00	1.0000
B*C	1	2.25000000	2.25000000	21.00	0.0025
A*B*C	1	0.25000000	0.25000000	2.33	0.1705

The analysis of variance shows that factors *A*, *B*, and *C* and operator are significant, as well as the interactions AB and BC. The residual plot shows no special pattern.



also provide QQ plot here.

e) Attach your SAS/R code

```
data one;
input A B C rep deviation @@;
datalines;
-1    -1    -1    1     -3
-1    -1    -1    2      -1
+1    -1    -1    1     0
+1    -1    -1    2     1
-1    +1    -1    1     -1
-1    +1    -1    2     0
+1    +1    -1    1     2
+1    +1    -1    2     3
-1    -1    +1    1     -1
-1    -1    +1    2     0
+1    -1    +1    1     1
+1    -1    +1    2     2
-1    +1    +1    1     1
```

```
-1    +1    +1    2     3
+1    +1    +1    1     5
+1    +1    +1    2     6
;

/* part b)*/
proc glm data=one;
class A B C;
model deviation=A|B|C;
output out=onenew r=res p=pred;
run;
/*part c)*/

proc sgplot data=onenew;
scatter x=pred y=res; refline 0;
run;

/* part d) */

proc glm data=one;
class A B C rep;
model deviation=rep A|B|C;
output out=onenew2 r=res p=pred;
run;

proc sgplot data=onenew2;
scatter x=pred y=res; refline 0;
run;
```

4. Consider the variables
G = 2011 GDP per capita (in 2000 dollars, inflation-adjusted), and
Y = 2011 Life expectancy at birth
among n = 149 different nations worldwide. Data are

Nation	GDP	life.expect
Algeria	2255.225	73.131
Angola	629.955	51.093
Argentina	11601.63	75.901
Armenia	1384.085	74.241
⋮	⋮	⋮
Venezuela	5671.912	74.402
Vietnam	757.401	75.181
Zimbabwe	347.746	51.384

(The full data are available in the file **gapminder.csv**.)

a) Per capita values such as GDP are notoriously skewed. Verify this by plotting a histogram for G. (Indicate which binning rule you use for your histogram bins.) If available, overlay a simple kernel density estimator.
b) Plot Y against X = log(GDP). What pattern appears?

c) Given the questions on the pattern of response, work with X = log(GDP) and calculate a robust, linear, loess fit of Y vs. X with smoothing parameter set to q = 0.7. Overlay the loess fit on the scatterplot. Does this improve visualization of the pattern?

d) From your loess fit, predict the (mean) Life Expectancy at a GDP of 15000.

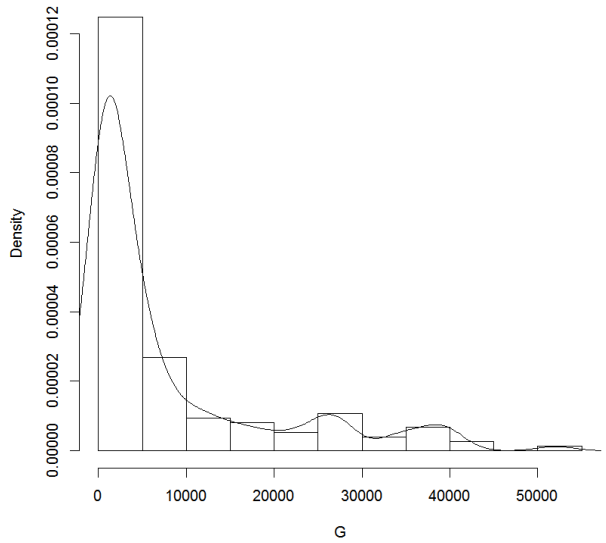
e) Plot the residuals from the loess fit against X = GDP. Do any important patterns appear?

4.answer.

(a) Sample R code for data retrieval and then for plotting a histogram with kernel density overlaid:

```
gapminder.df = read.csv( file.choose() )
G = GDP; Y = life.expect
hist( G, prob=T, main='' )
lines ( density(G) )
```

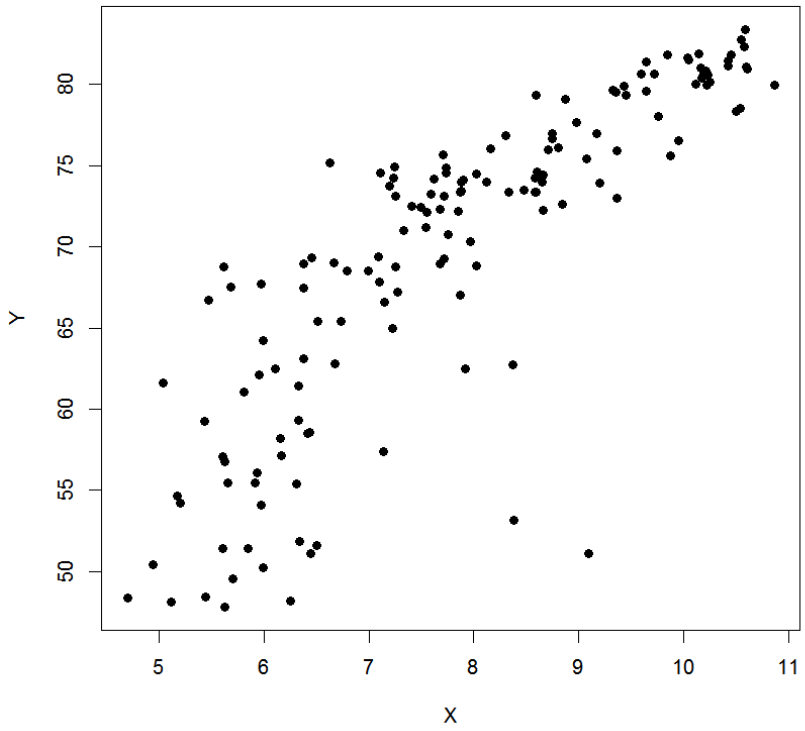
The R default for the bins is Sturges' Rule, which is used here:



As expected, the plot indicates a heavy right skew.

(b) Sample R code for plot:

```
X = log(GDP)
plot( Y ~ X, pch=19 )
```



The plot indicates a general increase in Y over X = log(GDP), with a hint of curvilinearity (and possibly a few outliers...).

(c) Sample R code for loess fit:

```
gapminder.loess = loess( Y~X, span=0.7, degree=1, family='symmetric' )
```

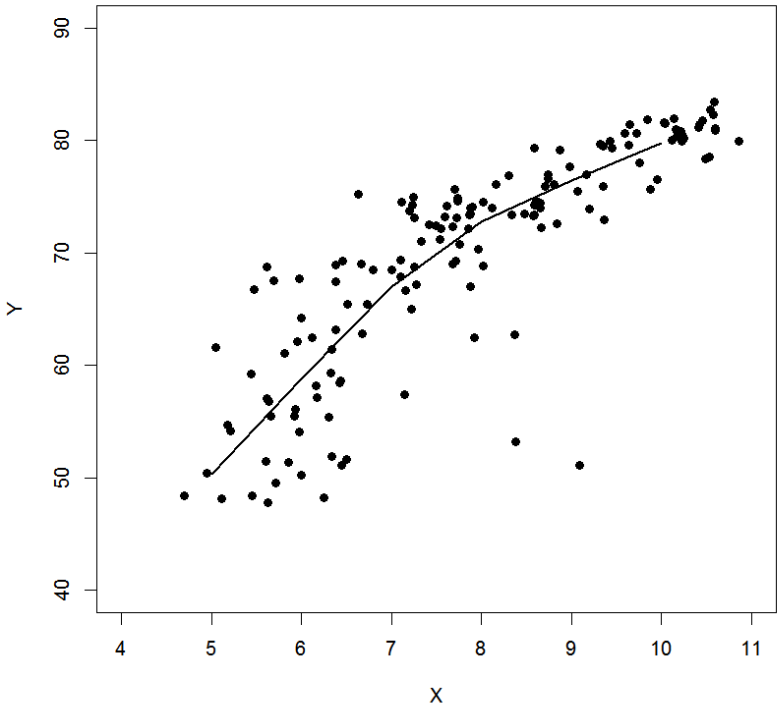
Smoothed predictions are found via

```
Ysmooth1r = predict( gapminder.loess, data.frame(X=seq(4,11)) )
```

Overlay plot via

```
plot( Y~X, pch=19, xlim=c(4,11), ylim=c(40,90) ); par( new=T )
plot( Ysmooth1r~seq(4,11), type='l', lwd=2 , xaxt='n',
        yaxt='n' , xlab='', ylab='', xlim=c(4,11), ylim=c(40,90) )
```

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



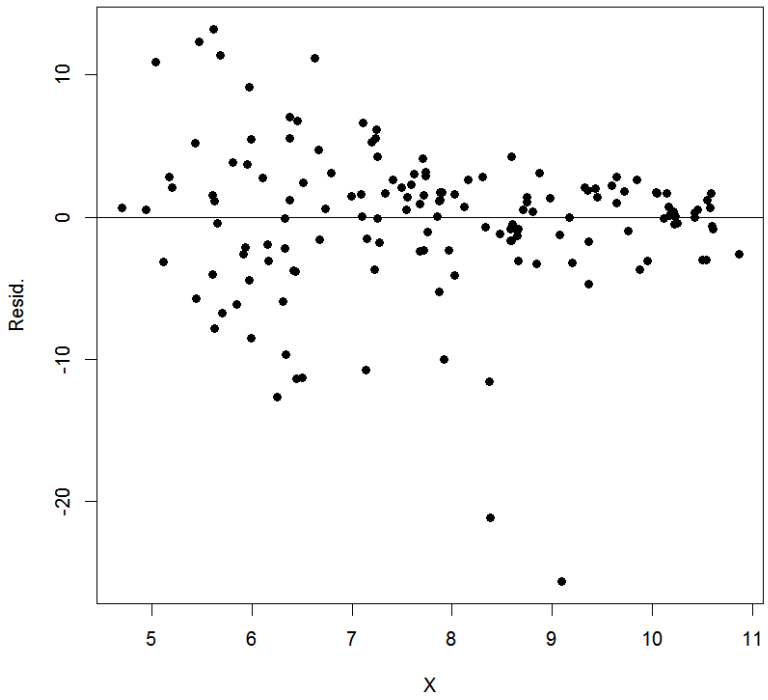
(d) Sample R command:

```
predict( gapminder.loess, data.frame(X=log(15000)) )
```

which gives 78.488 (yrs.).

(e) Residual plot, using

```
plot( resid(gapminder.loess)~X, pch=19, ylab='Resid.' ); abline( h=0 )
```



appears to show decreasing variance with increasing X = log(GDP) , and also highlights the two outliers near X = 8.2 and X = 9.1. The fit here requires more careful investigation.

5. Show that for the simple linear model $Y_i \sim$ indep. $N(\beta_0 + \beta_1 X_i, \sigma^2)$, $i = 1,...,n$, the least squares estimator of β_0,

$$b_0 = \bar{Y} + \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2}\bar{X},$$

is unbiased.

Answer:

Recognize that b_0 has the form $b_0 = \sum \kappa_i Y_i$ for

$$\kappa_i = \frac{1}{n} - \frac{\bar{X}(X_i - \bar{X})}{\sum(X_i - \bar{X})^2}.$$

Thus $E[b_0] = E[\sum \kappa_i Y_i] = \sum \kappa_i E[Y_i] = \sum \kappa_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum \kappa_i + \beta_1 \sum \kappa_i X_i$. But now,

$$\sum \kappa_i = \sum\left(n^{-1} - \bar{X}(X_i - \bar{X})/\sum(X_i - \bar{X})^2\right) = 1 - \bar{X}\sum(X_i - \bar{X})/\sum(X_i - \bar{X})^2.$$

But it is well known that $\sum(X_i - \bar{X}) = 0$, so $\sum \kappa_i = 1 - 0 = 1$. Also,

$$\sum \kappa_i X_i = \sum\left(X_i/n - X_i\bar{X}(X_i - \bar{X})/\sum(X_i - \bar{X})^2\right) = \bar{X} - \bar{X}\sum X_i(X_i - \bar{X})/\sum(X_i - \bar{X})^2.$$

It is straightforward to show that $\sum X_i(X_i - \bar{X}) = \sum(X_i - \bar{X})^2$, so that $\sum \kappa_i X_i = \bar{X} - \bar{X}(1) = 0$. Therefore $E[b_0] = \beta_0 \sum \kappa_i + \beta_1 \sum \kappa_i X_i = \beta_0(1) + \beta_1(0) = \beta_0$ and hence b_0 is unbiased for β_0.

6. The following data on arable land (km^2), birth rate (per 1000 popl'n), and outbound migration (per 1000 popl'n) were recorded among 14 Middle East nations in 1994 (also see the file **arable.csv**):

Country	Arable Land	Birth Rate	Migration
Bahrain	2	26.6	+6.8
Egypt	3	28.7	–0.4
Jordan	4	37.8	+0.5
Iran	8	42.4	0
Iraq	12	44.1	+0.4
Israel	17	20.5	+8.0
Kuwait	0	29.4	+25.4
Lebanon	20	27.9	–1.5
Oman	1	40.4	0
Qatar	0	18.8	+10.1
Saudi Arabia	1	38.3	0
Syria	28	43.7	0
U.A.E.	0	27.7	+23.3
Yemen	6	50.7	–2.4

Assuming the Migration data are normally distributed, conduct a multiple linear regression on Y = Migration with predictors Arable Land, Birth Rate, and their interaction. Assess whether and how these variable may affect the Migration outcome. Be as complete as possible. For any inferences, set $\alpha = 0.10$.

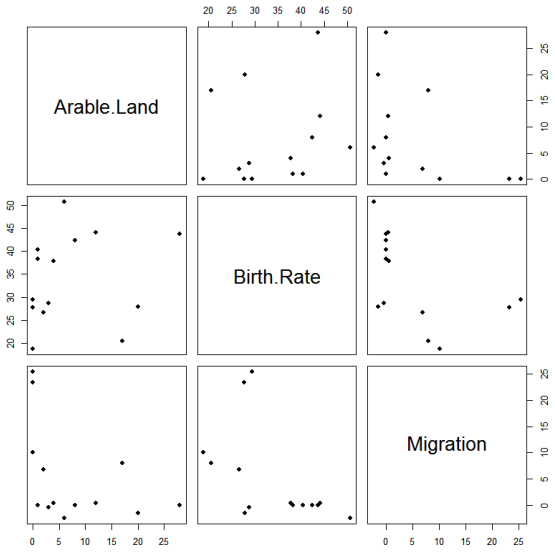
Answer:

Sample R code:

```
arable.df = read.csv( file.choose() )
attach( arable.df )
```

ALWAYS PLOT THE DATA! Start with a scatterplot matrix to examine possible relationships:

```
pairs( arable.df[,2:4], pch=19 )
```



ANOVA from the full-model MLR shows no signif. interaction, nor a signif. affect due to Arable Land, at (pointwise) $\alpha = .10$.

```
arable.lm = lm( Migration ~ Birth.Rate*Arable.Land )
anova( arable.lm )

Analysis of Variance Table
Response: Migration
                      Df Sum Sq Mean Sq F value Pr(>F)
Birth.Rate             1 299.97 299.972  5.0450 0.0485
Arable.Land            1  94.30  94.298  1.5859 0.2365
Birth.Rate:Arable.Land 1  68.02  68.016  1.1439 0.3100
Residuals             10 594.59  59.459
```

So, reduce the model to a SLR on x = Birth Rate

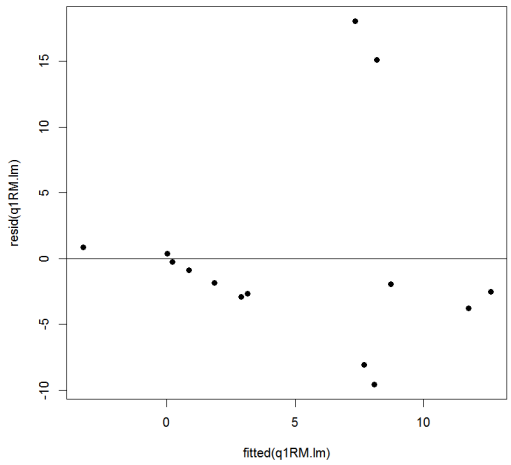
```
arableRM.lm = lm( Migration ~ Birth.Rate )
summary( arableRM.lm )

Call:
lm(formula = Migration ~ Birth.Rate)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.9537     8.0524   2.726   0.0184
Birth.Rate   -0.4972     0.2280  -2.181   0.0498

Residual standard error: 7.942 on 12 degrees of freedom
Multiple R-squared:  0.2838,    Adjusted R-squared:  0.2241
F-statistic: 4.756 on 1 and 12 DF,  p-value: 0.04982
```

Next check the residuals. A resid. plot shows clear variance heterogeneity with increasing response (i.e., with decreasing Birth Rate, since the regression has negative slope).

```
plot( resid(arableRM.lm)~fitted(arableRM.lm), pch=19 ); abline( h=0 )
```



Moving to a transformation in Migration, say, U = log{Migration + 3}, or adding a quadratic term in Birth Rate, does not assuage the variance heterogeneity. So, consider a weighted least squares (WLS) fit with, say, $w_i \propto 1/x_i$:

```
w = 1/Birth.Rate
arableWLS.lm = lm( Migration ~ Birth.Rate, weight=w )
summary( arableWLS.lm )

Call:
lm(formula = Migration ~ Birth.Rate, weights = w)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.620      7.878   2.617   0.0225
Birth.Rate    -0.458      0.241  -1.900   0.0817

Residual standard error: 1.487 on 12 degrees of freedom
Multiple R-squared:  0.2313,    Adjusted R-squared:  0.1672
F-statistic: 3.611 on 1 and 12 DF,  p-value: 0.08169
```

Now the regression is marginally signif. at $\alpha = .10$ ($P = 0.082$).

Maybe the two extreme inflow migration points (Kuwait and U.A.E.) are affecting the results. For an outlier analysis, find the Studentized deleted residuals, t_i. View any t_i as a potential outlier if $|t_i|$ exceeds the t-critical point $t_{\alpha/(2n)}(n–p–1) = t_{0.05/(28)}(14–2–1) = t_{0.0017857}(11) = 3.68867$:

```
ti = rstudent( arableWLS.lm )
which( abs(ti) > qt(.05/28,11,low=F) )
```

which gives `integer(0)`. The exceedance level is not reached (e.g., $\max\{|t_i| = 3.082$ at Kuwait) for any Country's residual, so there is no statistical motivation for removing any of the points from the analysis.

The relationship between Migration and Birth Rate here is apparently only marginally significant, although further study would be warranted to better understand the unusual features of these data. For instance, the small R-squared values suggest that the unexplained variability here is substantial; perhaps alternative, latent predictor variables could be identified that have a significant effect on the observed Migration patterns.