Statistics GIDP
Ph.D. Qualifying Exam
Methodology

Jan 7th, 2014, 9:00am-1:00pm

Instructions: Provide answers on the supplied pads of paper; write on only one side of each sheet. Complete exactly 3 of the first 4 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. True or false question - only circle "true " or "false".

   (a) True or false? F-statistic can be used for checking the equality of two population variances where the populations are assumed to have normal distributions.

   (b) True or false? In the simultaneous confidence interval calculation Bonferroni approach gives a wider interval than Scheffe method.

   (c) True or false? We use the type I sums of squares more often than the type III one because type I sums of squares do not depend upon the order in which effects are specified in the model.

   (d) True or false? When the number of treatments a=9, the number of blocks b=10, and the other parameters r =10 and k=9, it is a BIBD design.

   (e) True or false? If it is impossible to perform all of the runs in a 2^k factorial experiment under homogeneous conditions, we need to consider the design technique blocking.

   (f) True or false? In a split-plot design the easy-to-change factor is usually chosen as the whole-plot factor and the hard-to-change one as the subplot factor.

   (g) A Graeco-Latin square design is run with a = 5 treatments.
      i. true or false? Two nuisance factors are blocked.

      ii. true or false? The total number of runs is 125 (=5*5*5).

      iii. true or false? The degree of freedom for the error term is 12.

      iv. true or false? The degrees of freedom for the overall F-test in ANOVA table are 16 and 8.

   (h) True or false? In a single replicate of 2^4 design with factors A, B, C, and D in two blocks, the term ABCD should be confounded with the blocking factor.

   (i) True or false? The word length patterns for three IV designs 2^{7-2} are {4,4,4}, {4,4,6}, {4,5,5}. The design with the pattern {4,4,4} is the best choice among them.

2. To simplify production scheduling, an industrial engineer is studying the possibility of assigning one time standard to a particular class of jobs, believing that differences between jobs is negligible. To see if this simplification is possible, six jobs are randomly selected. Each job is given to a different group of three operators. Each operator completes the job twice at different times during the week, and the following results are obtained.

	Operator 1		Operator 2		Operator 3	
Job	1	2	1	2	1	2
1	158.3	159.4	159.2	159.6	158.9	157.8
2	154.6	154.9	157.7	156.8	154.8	156.3
3	162.5	162.6	161.0	158.9	160.5	159.5
4	160.0	158.7	157.5	158.9	161.1	158.5
5	156.3	158.1	158.3	156.9	157.7	156.9
6	163.7	161.0	162.3	160.3	162.6	161.8

Computer Output:

ANOVA: Time versus Job, Operator

Factor	Type	Levels	Values
Job	random	6	1 2 3 4 5 6
Operator(Job)	random	3	1 2 3

Analysis of Variance for Time

Source	DF	SS	MS	F	P
Job	__	_____	29.622	_____	_____
Operator(Job)	__	_____	1.721	_____	_____
Error	__	_____	1.092		
Total	35	188.430			

(a) What design/experiment is this?
(b) Write the statistic model and the corresponding assumptions
(c) Fill in the missing values for the output.
(d) Estimate the variability between jobs. Write the hypothesis in notation for testing the equality of the jobs.
(e) Estimate the variability between the operators.
(f) What are your conclusions about the use of a common time standard for all jobs in this class?

3. The shear strength of an adhesive is thought to be affected by the application pressure and temperature. A factorial experiment is performed in which both factors are assumed fixed.

	Temperature (°F)		
Pressure (lb/in^2)	250	260	270
120	9.60	11.28	9.00
130	9.69	10.10	9.57
140	8.43	11.01	9.03
150	9.98	10.44	9.80

The following SAS codes are run.

```
data adhesive;
      input pressure temp strength @@;
      datalines;
      120 250 9.6 120 260 11.28 120 270 9
      130 250 9.69 130 260 10.10 130 270 9.57
      140 250 8.43 140 260 11.01 140 270 9.03
      150 250 9.98 150 260 10.44 150 270 9.8
      ;

proc glm;
       class pressure temp;
       model strength = pressure temp;
       output out=one r=res p=pred;

data two;
set one;
q = pred*pred;

proc glm data=two;
class pressure temp;
model strength=pressure temp q/ss3;
run;
```

SAS output:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
pressure	3	0.55757954	0.18585985	0.56	0.6651
temp	2	0.55761607	0.27880803	0.84	0.4856
q	1	0.48947296	0.48947296	1.47	0.2794

(a) What is the degree of freedoms for the error term in the first "glm" procedure result output (not given)?

(b) What is the purpose of the second "glm" procedure?

(c) Write the corresponding statistical model and state the hypothesis test in terms of mathematical notation.

(d) What is the test name for the above hypothesis?

(e) What conclusion can you get for the hypothesis in c) from the SAS output (α=0.05)?

(f) Based on the SAS output, can we make a conclusion that neither temperature nor pressure affects the shear strength of an adhesive? Why or why not?

4. The yield of a chemical process is being studied. The two factors of interest are temperature and pressure. Three levels of each factor are selected; however, only 9 runs can be made in one day. The experimenter runs a complete replicate of the design on each day - the combination of the levels of pressure and temperature is chosen randomly. The data are shown in the following table. Analyze the data assuming that the days are blocks.

	Day 1 Pressure			Day 2 Pressure		
Temperature	250	260	270	250	260	270
Low	86.3	84.0	85.8	86.1	85.2	87.3
Medium	88.5	87.3	89.0	89.4	89.9	90.3
High	89.1	90.2	91.3	91.7	93.2	93.7

Part of SAS output:

```
Factor     Type       Levels Values
Day       fixed       2      1      2
temp      fixed       3     Low    Medium   High
pres      fixed       3     250    260      270

                Sum of             Mean
   Source       Squares      DF    Square   F-value   Prob > F
   Day          13.01         1    13.01
   temp         5.51          2    2.75     ______    _______
   pres         99.85         2    49.93    ______    _______
   temp*pres    4.45          4    1.11
   Residual     4.25          8    0.53
   Cor Total    127.07             17
```

(a) What design is this?

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

(c) Fill up the blanks in the ANOVA table below

(d) Why are the terms “day*temp” “day*pres” not included in the model?

(e) Can you calculate the F values and p-values for the terms “day”, “temp*pres”? If yes, calculate them. If not, explain why.

(f) Draw conclusions at α=0.05.

5. Consider the following data, which are frequencies of mutations seen in survivors of atomic radiation exposure (x is radiation exposure dose, Y is mutation rate):

x	Y
1.980	37.9213
2.540	39.8544
5.000	45.4828
4.670	45.4816
6.890	51.2148
8.120	54.7815
9.995	59.6586
11.578	63.0588
14.400	70.0128
15.001	71.3794
16.870	75.5131
19.560	81.4950

The data are found **in the file data1.csv**. Fit an appropriate model to these data (you may assume they are independent, normally distributed, with a common variance). Be sure to assess the quality of your fit to the data in a comprehensive manner. Is mutation rate significantly affected by radiation exposure (at a false positive rate $\alpha = 0.1$)?

6. A southeastern city's newspaper reported data from n = 46 nearby counties on each county's proportion of minority police officers, relative to its proportion of minority residents. The full data are available **in the file data2.csv**; they are

County Number	% minority officers	% minority population	County Number	% minority officers	% minority population
1	26	31	24	16	30
2	13	24	25	38	54
3	75	68	26	11	17
4	8	17	27	26	57
5	37	62	28	14	28
6	29	43	29	11	25
7	13	28	30	20	28
8	15	24	31	31	62
9	38	52	32	11	11
10	19	35	33	34	55
11	6	21	34	44	49
12	24	40	35	29	59
13	24	33	36	14	35
14	40	57	37	8	9
15	31	45	38	44	58
16	19	40	39	5	7
17	30	43	40	27	42
18	19	23	41	15	33
19	28	46	42	15	21
20	45	58	43	2	43
21	29	39	44	12	30
22	25	43	45	53	64
23	13	18	46	11	20

Assuming the data are normally distributed, test the assertion that minorities are underrepresented on police forces in these counties. Set your false positive rate $\alpha = 0.01$.

7. A set of 3 quantitative predictor variables was considered for a study of their effects on a laboratory animal's response to a drug in an upcoming study. The predictor variables were X_1=body weight (grams), X_2=age (months), and X_3=drug dose to be administered (mg). These were recorded as follows:

X_1	X_2	X_3	X_1	X_2	X_3
176	6.5	0.88	165	7.9	0.84
176	9.5	0.88	158	6.9	0.80
190	9.0	1.00	148	7.3	0.74
176	8.9	0.88	149	5.2	0.75
200	7.2	1.00	163	8.4	0.81
167	8.9	0.83	170	7.2	0.85
188	8.0	0.94	186	6.8	0.94
195	10.0	0.98	146	7.3	0.73
176	8.0	0.88	181	9.0	0.90
149	6.4	0.75			

The data are found **in the file data3.csv**. Assume a multiple linear regression model will be appropriate once the n=19 responses Y_i are observed. Are there any concerns with the values of these predictor variables? If so, is any solution possible? Explain.

PhD Qualifying exam – Methodology Jan 2014

Solutions

1. True or false question - only circle "true " or "false"

   (a) **True** or false? F-statistic can be used for checking the equality of two population variances where the populations are assumed to have normal distributions.

   (b) True or **false**? In the simultaneous confidence interval calculation Bonferroni approach gives a wider interval than Scheffe method.

   (c) True or **false**? We use the type I sums of squares more often than the type III one because type I sums of squares do not depend upon the order in which effects are specified in the model.

   (d) True or **false**? When the number of treatments a=9, the number of blocks b=10, and the other parameters r =10 and k=9, it is a BIBD design.

   (e) **True** or false? If it is impossible to perform all of the runs in a 2^k factorial experiment under homogeneous conditions, we need to consider the design technique blocking.

   (f) True or **false**? In a split-plot design the easy-to-change factor is usually chosen as the whole-plot factor and the hard-to-change one as the subplot factor.

   (g) A Graeco-Latin square design is run with a = 5 treatments.
   i. true or **false**? Two nuisance factors are blocked.

   ii. true or **false**? The total number of runs is 125 (=5*5*5).

   iii. true or **false**? The degree of freedom for the error term is 12.

   iv. **true** or false? The degrees of freedom for the overall F-test in ANOVA table are 16 and 8.

   (h) **true** or false? In a single replicate of 2^4 design with factors A, B, C, and D in two blocks, the term ABCD should be confounded with the blocking factor.

   (i) true or **false**? The word length patterns for three IV designs 2^{7-2} are {4,4,4}, {4,4,6}, {4,5,5}. The design with the pattern {4,4,4} is the best choice among them.

2. To simplify production scheduling, an industrial engineer is studying the possibility of assigning one time standard to a particular class of jobs, believing that differences between jobs is negligible. To see if this simplification is possible, six jobs are randomly selected. Each job is given to a different group of three operators. Each operator

completes the job twice at different times during the week, and the following results are obtained.

	Operator 1		Operator 2		Operator 3	
Job	1	2	1	2	1	2
1	158.3	159.4	159.2	159.6	158.9	157.8
2	154.6	154.9	157.7	156.8	154.8	156.3
3	162.5	162.6	161.0	158.9	160.5	159.5
4	160.0	158.7	157.5	158.9	161.1	158.5
5	156.3	158.1	158.3	156.9	157.7	156.9
6	163.7	161.0	162.3	160.3	162.6	161.8

Computer Output:

ANOVA: Time versus Job, Operator

Factor	Type	Levels	Values
Job	random	6	1 2 3 4 5 6
Operator(Job)	random	3	1 2 3

Analysis of Variance for Time

Source	DF	SS	MS	F	P
Job	__	_____	29.622	_____	_____
Operator(Job)	__	_____	1.721	_____	_____
Error	__	_____	1.092		
Total	35	188.430			

(a) (4 pts) What design/experiment is this?

Nested design (with 2 random factors).

(b) (6 pts) Write the statistic model and the corresponding assumptions

$$y_{ijk} = \mu + \tau_i + \beta_{j(i)} + \epsilon_{k(ij)}$$

where $\tau_i \sim N(0, \sigma_\tau^2)$ and $\beta_{j(i)} \sim N(0, \sigma_\beta^2)$.

(c) (5 pts) Fill in the missing values.

Analysis of Variance for Time

Source	DF	SS	MS	F	P
Job	**5**	**148.111**	29.622	**17.21**	**0.000**
Operator(Job)	**12**	**20.653**	1.721	**1.58**	**0.186**
Error	**18**	**19.665**	1.092		
Total	35	188.430			

(d) (4 pts) Estimate the variability between jobs. Write the hypothesis in notation for testing the equality of the jobs.

$$\hat{\sigma}_{\tau}^2 = (MS_{Job} - MS_{op(job)}) / (nb) = (29.622 - 1.721) / 6 = 4.65016$$

$$H_0 : \sigma_{\tau}^2 = 0$$

$$H_0 : \sigma_{\tau}^2 \neq 0$$

(e) (3 pts) Estimate the variability between the operators.

$$\hat{\sigma}_{\beta}^2 = (MS_{op(job)} - MS_E) / n = (1.721 - 1.092) / 2 = 0.3145$$

(f) (3 pts) What are your conclusions about the use of a common time standard for all jobs in this class?

The jobs differ significantly; the use of a common time standard would likely not be a good idea.

3. The shear strength of an adhesive is thought to be affected by the application pressure and temperature. A factorial experiment is performed in which both factors are assumed to be fixed.

	Temperature (°F)		
Pressure (lb/in^2)	250	260	270
120	9.60	11.28	9.00
130	9.69	10.10	9.57
140	8.43	11.01	9.03
150	9.98	10.44	9.80

The following SAS codes are run.

```
data adhesive;
      input pressure temp strength @@;
      datalines;
      120 250 9.6 120 260 11.28 120 270 9
      130 250 9.69 130 260 10.10 130 270 9.57
      140 250 8.43 140 260 11.01 140 270 9.03
      150 250 9.98 150 260 10.44 150 270 9.8
      ;
```

```
proc glm;
      class pressure temp;
      model strength = pressure temp;
      output out=one r=res p=pred;

data two;
set one;
q = pred*pred;

proc glm data=two;
class pressure temp;
model strength=pressure temp q/ss3;
run;
```

SAS output:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
pressure	3	0.55757954	0.18585985	0.56	0.6651
temp	2	0.55761607	0.27880803	0.84	0.4856
q	1	0.48947296	0.48947296	1.47	0.2794

(a) (3 pts) What is the degree of freedoms for the error term in the first "glm" procedure result output (not given)?

df=11-3-2=6

(b) (4 pts) What is the purpose of the second "glm" procedure?

check the non-additivity of the model.

(c) (6 pts) write the corresponding statistical model and state the hypothesis test in terms of mathematical notation.

$$y_{ij} = \mu + \tau_i + \beta_j + \gamma\tau_i\beta_j + \epsilon_{ij}$$

$$H_0 : \gamma = 0 \text{ vs } H_1 : \gamma \neq 0$$

(d) (3 pts) What is the test name for the above hypothesis?

Tukey's one degree of freedom test

(e) (4 pts) What conclusion can you get for the hypothesis in c) from the SAS output (α=0.05)?

The p-value =0.2794>0.05, so we have no evidence of non-additivity. Therefore, we can move on the analyze the experiment assuming a additive model, i.e.,

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

(f) (5 pts) Based on the SAS output, can we make a conclusion that neither temperature nor pressure affects the shear strength of an adhesive? Why?

No. The p-values here are not meaningful for testing the temperature and pressure.

4. The yield of a chemical process is being studied. The two factors of interest are temperature and pressure. Three levels of each factor are selected; however, only 9 runs can be made in one day. The experimenter runs a complete replicate of the design on each day - the combination of the levels of pressure and temperature is chosen randomly. The data are shown in the following table. Analyze the data assuming that the days are blocks.

	Day 1 Pressure			Day 2 Pressure		
Temperature	250	260	270	250	260	270
Low	86.3	84.0	85.8	86.1	85.2	87.3
Medium	88.5	87.3	89.0	89.4	89.9	90.3
High	89.1	90.2	91.3	91.7	93.2	93.7

Part of SAS output:

```
Factor      Type       Levels Values
Day        fixed       2      1      2
temp       fixed       3     Low    Medium   High
pres       fixed       3     250    260      270

                Sum of             Mean
    Source      Squares      DF    Square   F-value   Prob > F
    Day         13.01         1    13.01
    temp        5.51          2    2.75     ______    _______
    pres        99.85         2    49.93    ______    _______
    temp*pres   4.45          4    1.11
    Residual    4.25          8    0.53
    Cor Total   127.07             17
```

(a) (3 pts) What design is this?

Blocked factorial design

(b) (6 pts) State the statistical model and the corresponding assumptions.

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_k + \varepsilon_{ijk}, i = 1,.2,3; j = 1,.2,3, k = 1,2.$$
$$\sum \alpha_i = 0, \sum \beta_j = 0, \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$$
$$, \sum \delta_k = 0, \varepsilon_{ijk} \sim N(0, \sigma^2)$$

(c) (4 pts) Fill up the blanks in the ANOVA table below

Source	Sum of Squares	DF	Mean Square	F-Value	Prob > F
day	13.01	1	13.01		
temp	5.51	2	2.75	**5.18**	**0.0360**
pres	99.85	2	49.93	**93.98**	**< 0.0001**
temp*pres	4.45	4	1.11		
Residual	4.25	8	0.53		
Cor Total	127.07	17			

(d) (4 pts) Why are the terms "day*temp" "day*pres" not included in the model?

"day" is a blocking factor and it's assumed that has no interaction with the treatment factors.

(e) (4 pts) Can you calculate the F values and p-values for the terms "day", "temp*pres"? If yes, calculate them. If not, explain why.

Yes. F_{day}=13.01/0.53=25.84, pvalue< 0.0001
$F_{temp*press}$=1.11/0.53=2.1, p-value=0.1733

(f) (4 pts) Draw conclusions at α=0.05.

Both main effects, temperature and pressure, and the blocking factor are significant.

5. Consider the following data, which are frequencies of mutations seen in survivors of atomic radiation exposure (x is radiation exposure dose, Y is mutation rate):

x	Y
1.980	37.9213
2.540	39.8544
5.000	45.4828
4.670	45.4816
6.890	51.2148
8.120	54.7815
9.995	59.6586
11.578	63.0588
14.400	70.0128
15.001	71.3794
16.870	75.5131
19.560	81.4950

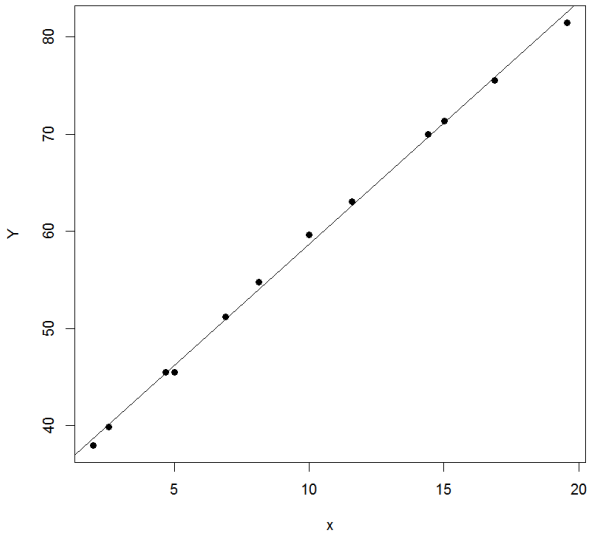
The data are found **in the file data1.csv**. Fit an appropriate model to these data (you may assume they are independent, normally distributed, with common variance). Be sure to

assess the quality of your fit to the data in a comprehensive manner. Is mutation rate significantly affected by radiation exposure (at a false positive rate of $\alpha = 10\%$)?

Answer:

Always plot the data! Sample R code:

```
q1.df = read.csv( file.choose() )
attach( q1.df )
x = exposure.dose
Y = mutation.rate
plot( Y~x, pch=19 ); abline( (lm(Y~x)) )
```

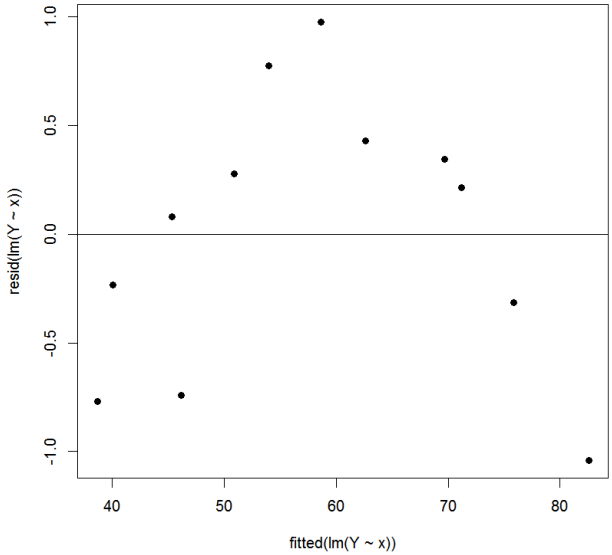


The plot indicates an increase over exposure; a fairly straightforward linear fit may be reasonable:

```
lm( Y~x )
```

producing $\hat{Y} = 33.753 + 2.494x$. However, a residual plot of $e_i = Y_i - \hat{Y}_i$ vs. x_i shows a clearly concave relationship:

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



(It is not unreasonable for mutation rate to increase at a decreasing rate as radiation exposure increases: after a while, the cells get about as damaged as they're going to be...) A better fit would include a quadratic term:

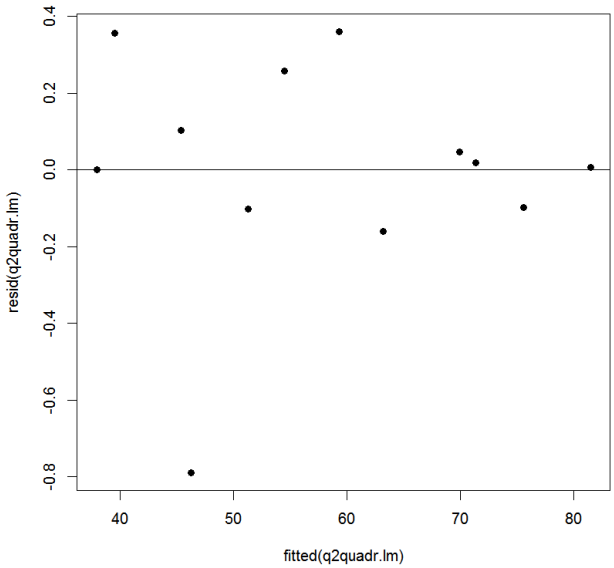
```
q1quadr.lm <- lm( Y~x+I(x^2) )
```

yielding $\hat{Y} = 32.24927 + 2.90355x - 0.01975x^2$.

Now, the residual plot from

```
plot( resid(q1quadr.lm)~fitted(q1quadr.lm), pch=19 ); abline( h=0 )
```

is somewhat more randomly scattered about $e = 0$:



(But, there's a possible outlier; see below.)

To test for a relationship between mutation and radiation, set H_o:$\beta_1=\beta_2=0$ vs. H_a:any departure. The test statistic is $F^* = 1.031\times10^4$ on 2 and 9 d.f. by examining the lower output from

```
summary( q1quadr.lm )
```

The P-value is $P = 7.564\times10^{-16}$, which is clearly less than $\alpha = 0.10$. Reject H_o and conclude a significant relationship exists between mutation rate and radiation exposure.

By the way, as seen above the residual plot from the quadratic polynomial fit indicates an extremely low residual at $\hat{Y}_3 = 46.27333$ ($x_3 = 5.0$). To assess this definitively, find the Studentized deleted residuals, t_i, from the fit and plot them against $\hat{Y}_i$. Mark a t_i as an outlier if $|t_i| > t(1–[\alpha/2n]; n–p–1) = t(1–[0.1/24]; 12–3–1) = t(0.995833; 8) = 3.4789$. Check this via

```
abs(rstudent( q1quadr.lm )) > qt(.99583333333,8)
```

from which we find

```
    1     2     3     4     5     6     7     8     9    10    11    12
FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

i.e., the observation at $x_3 = 5.0$ is the only violator of the outlier criterion:
$|t_3| = |–4.5776| > 3.4789$.
Further examination of this particular observation may be called for.

6. A southeastern city's newspaper reported data from n = 46 nearby counties on each county's proportion of minority police officers, relative to its proportion of minority residents. The full data are available **in the file data2.csv**; they are

County Number	% minority officers	% minority population	County Number	% minority officers	% minority population
1	26	31	24	16	30
2	13	24	25	38	54
3	75	68	26	11	17
4	8	17	27	26	57
5	37	62	28	14	28
6	29	43	29	11	25
7	13	28	30	20	28
8	15	24	31	31	62
9	38	52	32	11	11
10	19	35	33	34	55
11	6	21	34	44	49
12	24	40	35	29	59
13	24	33	36	14	35
14	40	57	37	8	9
15	31	45	38	44	58
16	19	40	39	5	7
17	30	43	40	27	42
18	19	23	41	15	33
19	28	46	42	15	21
20	45	58	43	2	43
21	29	39	44	12	30
22	25	43	45	53	64
23	13	18	46	11	20

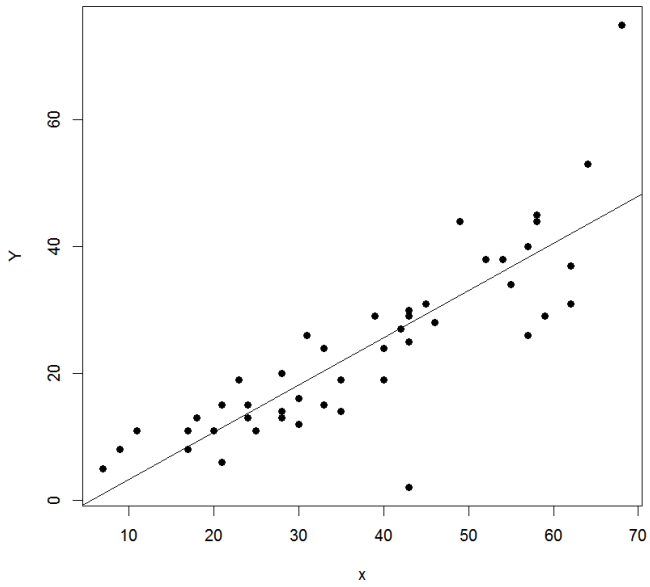
Assuming the data are normally distributed, test the assertion that minorities are underrepresented on police forces in these counties. Set your false positive rate to $\alpha = 0.01$.

Answer:

Always plot the data! Sample R code:

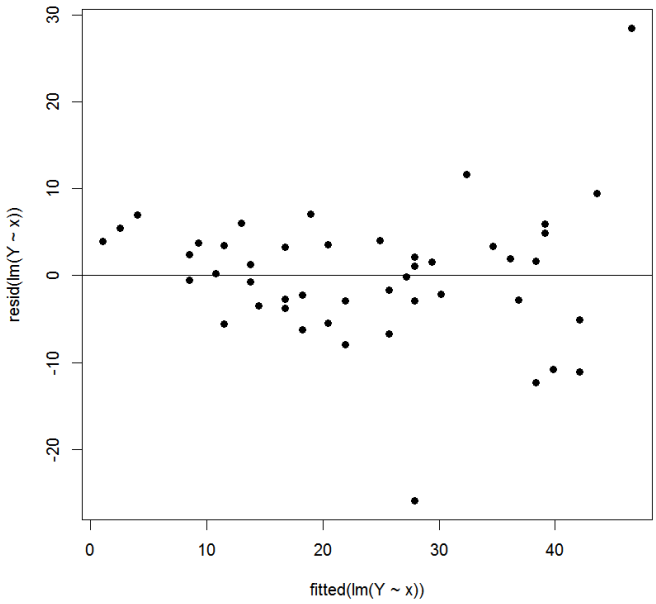
```
q2.df = read.csv( file.choose() )
attach( q2.df )
Y = minority.officers
x = minority.popln
plot( Y~x, pch=19 ); abline( (lm(Y~x)) )
```

Plot shows increase, but heterogeneous variation may be present.



Check this:

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



Seems like variance increases as $\hat{Y}$ grows. (There are a couple of potential outliers, too.) So, apply weighted least squares. A reasonable choice of weights assumes $Var[Y_i] \propto x_i$, so set the weights inversely proportional to x: $w_i = 1/x_i$. (One could reasonably take $Var[Y_i] \propto x_i^2$ instead here, but the resulting weighted MSE isn't as close to 1.0.)

```
summary( lm(Y~x, weights=1/x) )
```

Output (edited) is:

```
Call:
lm(formula = Y ~ x, weights = 1/x)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-3.8772 -0.6536  0.0208  0.5243  3.7822

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.76035    1.87713  -0.405    0.687
x            0.65546    0.05742  11.416 9.64e-15 ***

Residual standard error: 1.173 on 44 degrees of freedom
Multiple R-squared:  0.7476,    Adjusted R-squared:  0.7418
F-statistic: 130.3 on 1 and 44 DF,  p-value: 9.645e-15
```

The minority 'representation' would be fair if the slope were 1.0, so use the WLS analysis to test H_o:$\beta_1 = 1$ vs. H_o:$\beta_1 < 1$ (one sided alternative: check for 'underrepresented' status). The test statistic is

$$t^* = (b_{w1} - 1)/s[b_{w1}] = (0.65546 - 1)/0.05742 = -6.000627$$

i.e.,

```
(coef(lm(Y~x,weights=1/x))[2] - 1)/sqrt(vcov(lm(Y~x,weights=1/x))[2,2])
```

Reject H_o if $t^* = -6.000627 < t(.99; 44) = 2.414134$. This is clearly true (approx. *P*-value is $P[t(44) <= -6.000627] = 1.68\times10^{-7}$), so reject H_o and conclude that minority officers are significantly underrepresented in these 46 counties, relative to minority population.

A number of alternative operations could be applied here. For instance, one could employ White's robust standard error for b_1 instead of WLS.

Or, a quadratic polynomial fit could also be used (which would make the test for underrepresentation much more complex). The need for the higher-order predictor isn't indicated from the residual plot, however. In fact, if the datum at x = 68 is a true outlier, then removing it also removes much of the impetus to expand upon the simple linear predictor.

7. A set of 3 quantitative predictor variables was considered for a study of their effects on a laboratory animal's response to a drug in an upcoming study. The predictor variables were X_1=body weight (grams), X_2=age (months), and X_3=drug dose to be administered (mg). These were recorded as follows:

X_1	X_2	X_3	X_1	X_2	X_3
176	6.5	0.88	165	7.9	0.84
176	9.5	0.88	158	6.9	0.80
190	9.0	1.00	148	7.3	0.74
176	8.9	0.88	149	5.2	0.75
200	7.2	1.00	163	8.4	0.81
167	8.9	0.83	170	7.2	0.85
188	8.0	0.94	186	6.8	0.94
195	10.0	0.98	146	7.3	0.73
176	8.0	0.88	181	9.0	0.90
149	6.4	0.75			

The data are found **in the file data3.csv**. Assume a multiple linear regression model will be appropriate once the n=19 responses Y_i are observed. Are there any concerns with the values of these predictor variables? If so, is any solution possible? Explain.

Answer:

Check for (a) multicollinearity among the p–1=3 x-variables, and (b) for any high leverage points among the n=19 triplets.

(a) For multicollinearity, examine the correlations and the VIFs. Sample R code:

```
q3.df = read.csv( file.choose() )
attach( q3.df )
cor( cbind(X1,X2,X3) )
library ( car )
Ydummy = runif( length(X1) )  # create "sample" of n=19 Y values
vif( lm(Ydummy ~ X1+X2+X3))
mean( vif(lm(Ydummy ~ X1+X2+X3)) )
```

Output (edited) is:

```
           X1        X2        X3
X1 1.0000000 0.5000101 0.9902126
X2 0.5000101 1.0000000 0.4900711
X3 0.9902126 0.4900711 1.0000000
```

for the correlations, and

```
       X1        X2        X3
52.101917  1.335679 51.427154
```

for the VIFs and

```
[1] 34.95492
```

for $\overline{VIF}$. Since $\max\{VIF_k\} = 52.102$ clearly exceeds 10, and $\overline{VIF} = 34.95$ is far larger than 1.0, there is a clear problem with multicollinearity here. The correlation matrix suggests that X_1 and X_3 are the greater culprits. (Remove X_3 and the VIFs drop to near 1.33.)
Now, X_1=weight is not adjustable, but if the X_3=dose variable can be changed prior to beginning the study, repositioning it to decrease the multicollinearity may be prudent.

(b) For high leverage, examine the hat matrix diagonal elements, h_{ii}, to determine if any exceed the rule of-thumb cut-off of $2p/n = (2)(4)/19 = 0.421$:

```
hii = hatvalues( lm(Ydummy ~ X1+X2+X3) )
cutoff3 = 2*4/length(X1)
index = which( hii > cutoff3 )
c(X1[index], X2[index], X3[index])
```

which indicates that the single data triplet at $i = 3$,

```
[1] 190   9   1
```

is (the only) high-leverage point ($h_{33} = 0.851 > 0.421$). Again, if X_3=dose can be adjusted before collecting the response data, specific attention to the dose for animal #3 may be warranted. (But, be sure to recheck the hat values after adjusting X_3, in case other high leverage points might appear.)