Statistics GIDP
Ph.D. Qualifying Exam
Methodology

Jan 12, 2016, 9:00am-1:00pm

Instructions: Provide answers on the supplied pads of paper and/or use a Microsoft word document or equivalent to report your software code and outputs. Write on only one side of each sheet if you use paper. Complete exactly 5 of 6 problems; turn in only those sheets you wish to have graded. You may use the computer and/or a calculator. Stay calm and do your best. Good luck!

1. The yield of a chemical process is being studied. The two most important variables are thought to be the pressure and the temperature. Three levels of each factor are selected. There are **2** replications at each combination of pressure and temperature. The observation averages within each combination are given in the following table.

Temperature	Pressure			
	200	**215**	**230**	Average
150	90.3	90.6	90.3	90.4
160	90.1	90.5	90.0	90.2
170	90.5	90.7	90.3	90.5
Average	90.3	90.6	90.2	90.37

(a) Consider the mean model $y_{!"#} = !_{!"} + !_{!"#}$, where $!_{!"#}$! ! !0! ! ² !. The MSE is 0.!"#. Compute estimates of the $!_{i!}$, along with their standard errors.

(b) Propose an effect model with the interaction effect included; compute estimates of parameters.

(c) Compute standard errors of the parameter estimates for the main effects in part (b).

(d) Complete the following ANOVA table.

source	DF	SS	MS	F-values
Temperature				
Pressure				
Interaction		0.069		
Error			0.018	
Total				

(e) Compute R-square for this model fit. Also test significance of the interaction effect.

2. Steel is normalized by heating above a critical temperature, soaking, and then air cooling. This process increases the strength of the steel, refines the grain, and homogenizes the structure. An experiment is conducted to determine the effect of temperature and heat treatment time on the

strength of normalized steel. Two temperatures and three times are selected. The experiment is performed by heating the oven to a temperature (randomly choose from the two temperatures) and inserting three specimens. After 10 minutes one specimen (randomly chosen) is removed, after 20 minutes a second random specimen is removed, and after 30 minutes the final specimen is removed. Then, the temperature is changed to the other level and the process is repeated. Four shifts are required to collect the data, which are shown below.

		Temperature (F)	
Shift	Time(minutes)	1500	1600
1	10	63	89
	20	54	91
	30	61	62
2	10	50	80
	20	52	72
	30	59	69
3	10	48	73
	20	74	81
	30	71	69
4	10	54	88
	20	48	92
	30	59	64

Part of computer output:

Analysis of Variance for Strength

Source	DF	Sum Square	Mean Square	F-value	Pr>F
Shift	__	_____	_____		
Temp	__	2340.38	_____	_____	_____
Shift*Temp	___	240.46	_____		
Time	__	159.25	_____	_____	_____
Shift*Time	__	478.42	79.74		
Temp*Time	__	795.25	397.63	_____	_____
Shift*Temp*Time	__	244.42	40.74		
Residual	.	.	.	.	.
Total	23	4403.63			

(a) What design is this?

(b) Clearly specify which factor corresponds to what type of the treatment factors.

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

(d) Fill in the blanks in the ANOVA table below and draw conclusions at $\alpha=0.05$.

(e) Can you calculate the F values and p-values for the terms "Shift", "Shift*Temp", and "Shift*Time"? If yes, calculate them. If not, explain why.

(f) Draw conclusions at $\alpha=0.05$.

3. Sugar cane is very sensitive to climate and crop management practices. A sugar yield experiment involved 4 management practices (MANAGE) at each of 10 locations (LOCATION) in Louisiana. There were 2 fields assigned to each management practice at each location. The yield of extracted sugar (tons per acre) was calculated for each field. Use the partial SAS output below to answer the following questions.

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	___	_____	_____	_____	_____
Error	___	_____	_____		
Corrected	79	82.426			

Total

Source	DF	Sum Square	Mean Square	F Value	Pr > F
MANAGE	___	_____	_____	_____	_____
LOCATION	___	51.012	_____	_____	_____
MANAGE*LOCATION	___	15.865	_____	_____	_____

MANAGE	yield LSMEAN
1	4.666
2	5.053
3	4.841
4	4.267

(a) State whether you would consider each factor to be fixed or random and explain your reasoning.

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

(c) Complete the ANOVA table using the information above and summarize the results of the F tests.

(d) The researcher plans to compare each alternative management practice (MANAGE=2, 3, and 4) with the standard practice (MANAGE=1) and also compare the average of the alternative practices versus the standard. What's the standard error for each of these comparisons?

4. A cadre of modern recording artists had their Twitter activity examined to determine if the data could relate to first week sales of a new album. p = 3 predictor variables were taken: X_1 = {Number of Twitter followers (thousands)}, X_2 = {Average tweets per day}, and X_3 = log{Previous album's first week sales}. The response was Y = log{New album's first week sales}. The data are found in the file **albums.csv**, and also appear below

Artist	Y	X_1	X_2	X_3
Asher Roth	8.7160	118.5	2.1	11.0880
Ciara	11.3022	202.3	10.0	12.7321
Fabolous	11.6440	228.0	12.1	11.9767
Jordin Sparks	10.7579	350.8	17.4	11.6869
Maxwell	12.6635	70.0	1.1	12.5994
Trey Songz	11.9250	352.0	14.4	11.1982

(a) Fit an MLR model to these data. Test if the overall three-predictor model is significant. Operate at a false positive rate of 10%.

(b) Use partial t-tests to assess whether each individual predictor variable significantly affects mean (log-)new-album sales. Adjust for multiplicity at a false-positive FWE of 10%.

(c) What concerns might exist about the quality of the model fit with this data set?

5. Assume a simple linear regression model $Y_i \sim$ indep. $N(\beta_0 + \beta_1 X_i, \sigma^2)$, i = 1, ..., *n*. State the least squares estimator of the mean response and show that it has the form $\sum_{m=1}^{n} k_m Y_m$, where the constants k_m do not depend on the Y_ms. Be sure to fully explicate the form of these constants.

6. A study of website developers considered the following variables:
   X = number of months operating, and
   B = number of backlogged website orders at end of month
   for a series of different developer teams. The data are

months	# backlogged
3	12
6	18
⋮	⋮
17	37
20	26

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

   a) Count data are often notoriously non-normal. A useful stabilizing transformation for counts is the *Freeman-Tukey transform* $Y = \sqrt{B} + \sqrt{B+1}$. Calculate Y and plot it against X = months. What pattern appears?

   b) Calculate a robust, quadratic, loess fit of Y vs. X over a range of smoothing parameters: q = 0.25, 0.5, 0.75. Overlay the loess fits on the scatterplot. Comment on the differences among the different loess smooths.

   c) From the loess fit at q = 0.50, predict the (mean) number of backlogs when X = 12 months.

   d) Plot the residuals from the loess fit at q = 0.50 against X. Do any important patterns appear?

$$\widehat{\alpha\beta}_{21} = 90.1 - 90.2 - 90.3 + 90.37 = -0.03$$
$$\widehat{\alpha\beta}_{22} = 90.5 - 90.2 - 90.6 + 90.37 = 0.07$$
$$\widehat{\alpha\beta}_{23} = 90.0 - 90.2 - 90.2 + 90.37 = -0.03$$
$$\widehat{\alpha\beta}_{31} = 90.5 - 90.5 - 90.3 + 90.37 = 0.07$$
$$\widehat{\alpha\beta}_{32} = 90.7 - 90.5 - 90.6 + 90.37 = -0.03$$
$$\widehat{\alpha\beta}_{33} = 90.3 - 90.2 - 90.5 + 90.37 = -0.03$$

3) Compute standard errors of the parameter estimates for the main effects in part 2).

$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}$
$= \bar{y}_{i..} - \left(\bar{y}_{1..} + \bar{y}_{2..} + \bar{y}_{3..}\right)/3$

$V(\hat{\alpha}_1) = \frac{2}{3} * \bar{y}_{1..} + \frac{1}{3} * \bar{y}_{2..} + \frac{1}{3} * \bar{y}_{3..}$
$= \left(\frac{1}{n} + \frac{1}{n} + \frac{1}{n}\right) * V(\bar{y}_{1..})$
$= \frac{2}{3} * \frac{\sigma^2}{3 * 2}$
$= \frac{\sigma^2}{n}$

Since a=b=3, s.e of each parameter estimate=sqrt(MSE/9)=sqrt(0.002)=0.0447

4) Complete the following ANOVA table.

source	DF	SS	MS	F-values
Temperature				
Pressure				
Interaction		0.069		
Error			0.018	
Total				

source	DF	SS	MS	F-values
Temperature	2	0.2802	0.1401	7.783
Pressure	2	0.5202	0.2601	14.45
Interaction	4	0.069	0.01725	0.9583
Error	9	0.162	0.018	
Total	17	1.0314		

$$SS_{temperature} = bn\sum_i (\bar{y}_{i..} - \bar{y}_{...})^2 = 3 * 2 * 0.0467 = 0.2802$$

$$SS_{pressure} = an\sum_{i}(\bar{y}_{i..} - \bar{y}_{...})^2 = 3 * 2 * 0.0867 = 0.5202$$

5) Compute R-square. Also test significance of the interaction effect

R-squre=SS_{model}/SS_{total} = 0.8694/1.0314=0.843. We fail to reject the hypothesis that the interaction effect is 0 since p-value=0.47, which is greater than 0.05.

2. Steel in normalized by heating above the critical temperature, soaking, and then air cooling. This process increases the strength of the steel, refines the grain, and homogenizes the structure. An experiment is performed to determine the effect of temperature and heat treatment time on the strength of normalized steel. Two temperatures and three times are selected. The experiment is performed by heating the oven to a temperature (randomly choose from the two temperatures) and inserting three specimens. After 10 minutes one specimen (randomly choose) is removed, after 20 minutes the second specimen is removed, and after 30 minutes the final specimen is removed. Then the temperature is changed to the other level and the process is repeated. Four shifts are required to collect the data, which are shown below.

		Temperature (F)	
Shift	Time(minutes)	1500	1600
1	10	63	89
	20	54	91
	30	61	62
2	10	50	80
	20	52	72
	30	59	69
3	10	48	73
	20	74	81
	30	71	69
4	10	54	88
	20	48	92
	30	59	64

Part of computer output:

Analysis of Variance for Strength

Source	DF	Sum Square	Mean Square	F-value	Pr>F
Shift	_	____	_____		
Temp	_	2340.38	_____	_____	_____
Shift*Temp	__	240.46	_____		
Time	_	159.25	_____	_____	_____
Shift*Time	_	478.42	79.74		
Temp*Time	_	795.25	397.63	_____	_____
Shift*Temp*Time	_	244.42	40.74		
Residual	.	.	.	.	.
Total	23	4403.63			

(a) What design is this?

Split-plot design

(b) Clearly specify which factor is corresponding to what type of the treatment factors.

Temperature is the whole-plot factor; time is the sub-plot factor.

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

$$y_{ijk} = \mu + \gamma_i + \alpha_j + (\gamma\alpha)_{ij} + \beta_k + (\gamma\beta)_{ik} + (\alpha\beta)_{jk} + (\gamma\alpha\beta)_{ijk} + \varepsilon_{ijk}, i = 1,...,4; j = 1,.2, k = 1,2,3$$

$$\gamma_j \sim N(0,\sigma_r^2), \sum \alpha_i = 0, \sum \beta_k = 0, (\gamma\alpha)_{i,j} \sim N(0,\sigma_{r\alpha}^2), ((\gamma\beta)_{ik} \sim N(0,\sigma_{r\beta}^2), \gamma(\alpha\beta)_{ijk} \sim N(0,\sigma_{r\alpha\beta}^2),$$

$$\sum_j (\alpha\beta)_{jk} = \sum_{kj} (\alpha\beta)_{jk} = 0, \varepsilon_{ijk} \sim N(0,\sigma^2)$$

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

Source	DF	SumSquare	Mean Square	F-value	Pr>F
Shift	3	145.46	48.49		
Temp	1	2340.38	2340.38	**29.20**	**0.012**
Shift*Temp	3	240.46	80.15		
Time	2	159.25	79.63	**1.00**	**0.422**
Shift*Time	6	478.42	79.74		
Temp*Time	2	795.25	397.63	**9.76**	**0.013**
Shift*Temp*Time	6	244.42	40.74		

(e) Can you calculate the F values and p-values for the terms "Shift", "Shift*Temp", and "Shift*Time"? If yes, calculate them. If not, explain why.

No, as there is no degree of freedom for the residual/error term.

(f) Draw conclusions at α=0.05.

The temperature is significant, as well as the interaction between the temperature and the time.

3. Sugar cane is very sensitive to climate and crop management practices. A sugar yield experiment involved 4 management practices (MANAGE) at each of 10 locations (LOCATION) in Louisiana. There were 2 fields assigned to each management practice at each location. The yield of extracted sugar (tons per acre) was calculated for each field. Use the partial SAS output to answer the following questions.

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	___	_____	_____	_____	_____
Error	___	_____	_____		
Corrected Total	79	82.426			

Source	DF	Type III SS	Mean Square	F Value	Pr > F

MANAGE	___	_____	_____	_____	_____
LOCATION	___	51.012	_____	_____	_____
MANAGE*LOCATION	___	15.865	_____	_____	_____

MANAGE	yield LSMEAN
1	4.666
2	5.053
3	4.841
4	4.267

(a) State whether you would consider each factor to be fixed or random and explain your reasoning.

Manage is a fixed factor and location is a random factor, as manage has 4 levels and is the main treatment, and 10 locations are selected to represent the whole Louisiana.

(b) State the statistical model with assumptions.

$Y=\mu+ \tau_i+ \alpha_j + (\tau\alpha)_{ij}+ \varepsilon_{ijk}$

$\Sigma\tau_i=0$, $\alpha_j \sim N(0, \sigma_\alpha^2)$, $(\tau\alpha)_{ij} \sim N(0, \sigma_{\tau\alpha}^2)$, ε_{ijk} iid $N(0, \sigma^2)$

(c) Complete the ANOVA table using the information above and summarize the results of the F tests.

Grand mean=(4.666+5.053+4.841+4.267)/4=4.7068

SS_{MANAGE}=10*2*{(4.666-4.7068)^2+(5.053-4.7068)^2+(4.841-4.7068)^2+(4.267-4.7068)^2}=6.6591

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	39	73.5361	1.8855	8.4856	<0.0001
Error	40	8.8899	0.2222		
Corrected Total	79	82.426			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
MANAGE	3	6.6591	2.2197	9.9896	0.0001
LOCATION	9	51.012	5.668	25.5086	<0.0001
MANAGE*LOCATION	27	15.865	0.5876	2.6445	0.0026

(d) The researcher plans to compare each alternative management practice (MANAGE=2, 3, and 4) with the standard practice (MANAGE=1) as well as the average of the alternative practices versus the standard. What's the standard error for each of these comparisons?

The first three comparisons have the same standard error:

$Sqrt(MSE*\Sigma(c_i^2/n_i))=sqrt(0.2222*1/10)=0.1491$

The last one has:

$Sqrt(MSE*\Sigma(c_i^2/n_i))=sqrt(0.2222*((1/9+1/9+1/9+1)/20))=0.1217$

4. A cadre of modern recording artists had their Twitter activity examined to determine if the data could relate to first week sales of a new album. p = 3 predictor variables were taken: X_1 = {Number of Twitter followers (thousands)}, X_2 = {Average tweets per day}, and X_3 = log{Previous album's first week sales}. The response was Y = log{New album's first week sales}. The data are found in the file **albums.csv**, and also appear below

Artist	Y	X_1	X_2	X_3
Asher Roth	8.7160	118.5	2.1	11.0880
Ciara	11.3022	202.3	10.0	12.7321
Fabolous	11.6440	228.0	12.1	11.9767
Jordin Sparks	10.7579	350.8	17.4	11.6869
Maxwell	12.6635	70.0	1.1	12.5994
Trey Songz	11.9250	352.0	14.4	11.1982

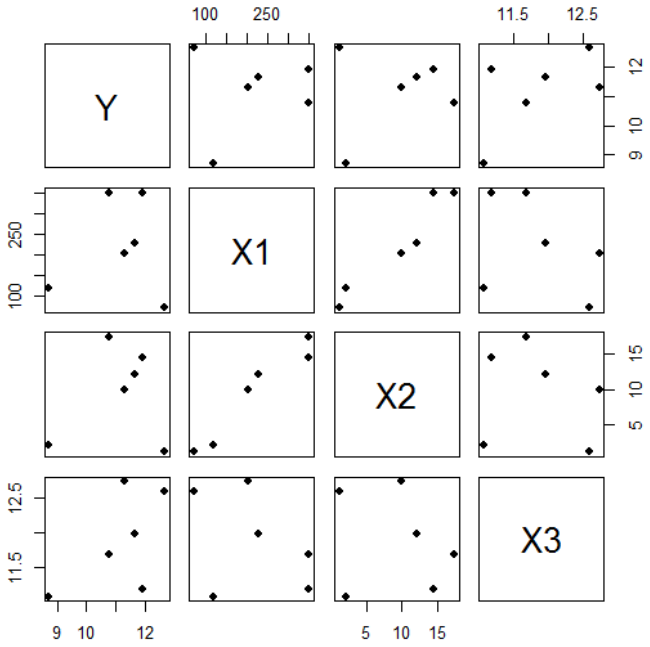
(a) Fit an MLR model to these data. Test if the overall three-predictor model is significant. Operate at a false positive rate of 10%.

(b) Use partial t-tests to assess whether each individual predictor variable significantly affects mean (log-)new-album sales. Adjust for multiplicity at a false positive FWE of 10%.

(c) What concerns might exist about the quality of the model fit with this data set?

Answer: Always plot the data! Sample R code:

```
albums.df = read.csv( file.choose() )
attach( albums.df )
pairs( Y~X1+X2+X3, pch=19 )
```



We see X_1 and X_2 may be multicollinear. (This can be verified by finding their VIFs: $VIF_1 = 33.623562$, $VIF_2 = 29.564078$, $VIF_3 = 2.588695$, so $\max\{VIF_k\} > 10$.) In any case, proceed with the MLR:

```
albums.lm = lm( Y~X1+X2+X3 )
summary( albums.lm )

Call:
lm(formula = Y ~ X1 + X2 + X3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.22718   19.28930  -0.893    0.466
X1            0.02597    0.03149   0.825    0.496
X2           -0.36205    0.51953  -0.697    0.558
X3            2.19868    1.47312   1.493    0.274

Residual standard error: 1.413 on 2 degrees of freedom
Multiple R-squared: 0.5675,     Adjusted R-squared: -0.08132
F-statistic: 0.8747 on 3 and 2 DF,  p-value: 0.5725
```

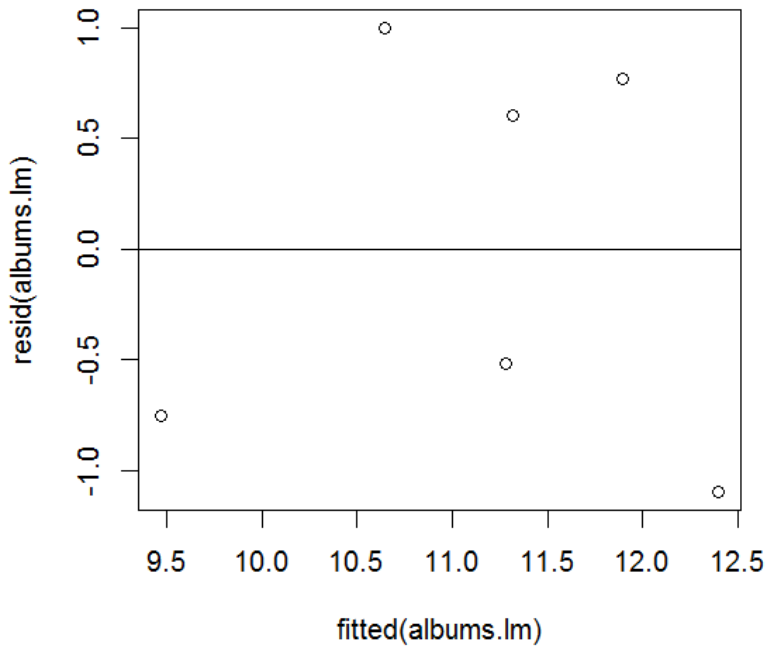
The (3,2) d.f. F-test P-value is reported as $P = 0.5726 > 0.10 = \alpha$. Hence, fail to reject H_o and conclude that as a group the three variables do not contribute significantly to the model.

(b) The partial t-test P-values appear in the final column of the **summary** output. Multiplying each by 3 corrects for multiplicity via a Bonferroni adjustment; notice, however, that since all three P-values were already very large, the correction does not affect the qualitative outcome: none of the three variables is seen to be significant at FWER of $\alpha = 0.10$.

(c) Even though tens of thousands of Twitter followers were studied here, the data really come down to only n = 6 artists. Thus, using p = 3 predictors leaves us with only n – p – 1 = 2 d.f. for the MSE. Thus creates very underpowered statistical analysis. By the way, a residual plot via

```
plot( resid(albums.lm) ~ fitted(albums.lm) )
abline( h=0 )
```

isn't that informative; again, the n = 6 observations don't give us much with which to work:



5. Assume a simple linear regression model $Y_i \sim$ indep. $N(\beta_0 + \beta_1 X_i, \sigma^2)$, $i = 1, \ldots, n$. State the least squares estimator of the mean response and show that it has the form $\sum_{m=1}^{n} k_m Y_m$, where the constants k_m do not depend on the Y_ms. Be sure to fully explicate the form of these constants.

Answer: The LS estimator for $E[Y_i]$ is $\hat{Y}_i = b_0 + b_1 X_i = (\bar{Y} - b_1 \bar{X}) + b_1 X_i = \bar{Y} + b_1(X_i - \bar{X})$. In effect, we are asked to show that this has the form $\sum_{m=1}^{n} k_m Y_m$. For simplicity, let $SSX = \sum_{m=1}^{n} (X_m - \bar{X})^2$ and notice that this is independent of any Y_i. Then, write $\hat{Y}_i = \bar{Y} + b_1(X_i - \bar{X})$ as

$$\hat{Y}_i = \bar{Y} + b_1(X_i - \bar{X}) = \left(\frac{1}{n}\right)\sum_{m=1}^{n} Y_m - \frac{\sum_{m=1}^{n}(X_m - \bar{X})Y_m}{SSX}(X_i - \bar{X})$$

$$= \left(\frac{1}{n}\right)\sum_{m=1}^{n} Y_m - \sum_{m=1}^{n} \frac{(X_m - \bar{X})(X_i - \bar{X})}{SSX} Y_m = \sum_{m=1}^{n} \left\{\frac{1}{n} + \frac{(X_m - \bar{X})(X_i - \bar{X})}{SSX}\right\} Y_m$$

so we see the k_m constants are

$$k_m = \frac{1}{n} + \frac{(X_m - \bar{X})(X_i - \bar{X})}{SSX} = \frac{1}{n} + \frac{(X_m - \bar{X})(X_i - \bar{X})}{\sum_{k=1}^{n}(X_k - \bar{X})^2}$$

which are also dependent upon X_i, as would be expected for estimating the mean response $E[Y_i]$.

6. A study of website developers considered the following variables:
   X = number of months operating, and
   B = number of backlogged website orders at end of month
   for a series of different developer teams. The data are

months	# backlogged
3	12
6	18
⋮	⋮
17	37
20	26

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

a. Count data are often notoriously non-normal. A useful stabilizing transformation for counts is the Freeman-Tukey transform $Y = \sqrt{B} + \sqrt{B+1}$. Calculate Y and plot it against X = months. What pattern appears?

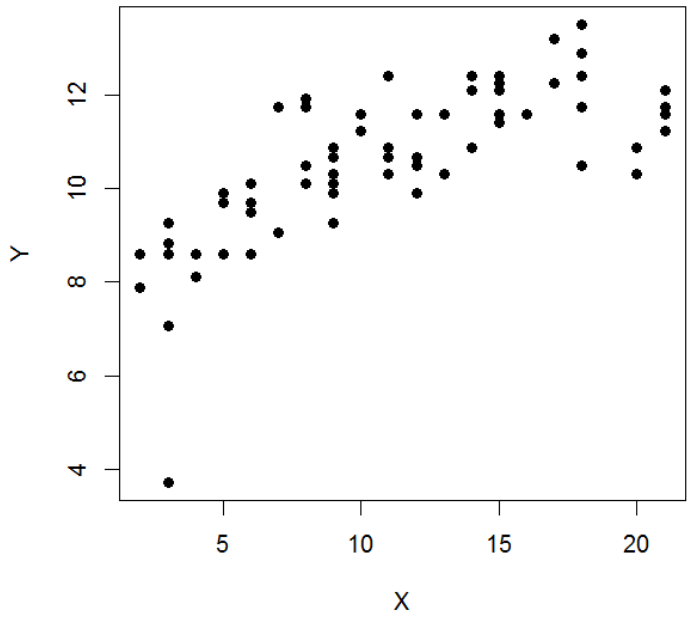
b. Calculate a robust, quadratic, loess fit of Y vs. X over a range of smoothing parameters: $q = 0.25, 0.5, 0.75$. Overlay the loess fits on the scatterplot. Comment on the differences among the different loess smooths.

c. From the loess fit at $q = 0.50$, predict the (mean) number of backlogs when X = 12 months.

d. Plot the residuals from the loess fit at $q = 0.50$ against X. Do any important patterns appear?

Answer. (a) Sample R code for data retrieval, transform, and scatterplot:

```
website.df = read.csv( file.choose() )
X = months; B = backlog;
Y = sqrt(B) + sqrt(B+1) #FT transform
plot( Y ~ X, pch=19 )
```



The plot indicates a general increase in Y over X, with an uncertain amount of curvilinearity (and possibly an outlier...).

(b) Sample R code for robust/quadratic loess fits:

```
website25.loess = loess( Y~X, span=0.25, degree=2,
family='symmetric' )
website50.loess = loess( Y~X, span=0.5, degree=2,
family='symmetric' )
website75.loess = loess( Y~X, span=0.75, degree=2,
family='symmetric' )
```

Smoothed predictions are found via

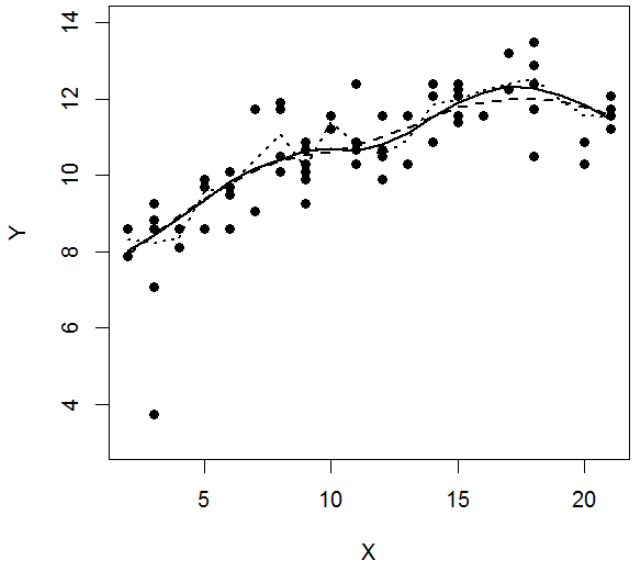
```
Ysmooth25 = predict( website25.loess, data.frame(X=seq(2,21)) )
```

```
Ysmooth50 = predict( website50.loess, data.frame(X=seq(2,21)) )
Ysmooth75 = predict( website75.loess, data.frame(X=seq(2,21)) )
```

Overlay plot via

```
plot( Y~X, pch=19, xlim=c(2,21), ylim=c(3,14) )
par( new=T )
plot( Ysmooth25~seq(2,21), type='l', lwd=2, lty=3, xaxt='n',
        yaxt='n' , xlab='', ylab='', xlim=c(2,21), ylim=c(3,14))
par( new=T )
plot( Ysmooth50~seq(2,21), type='l', lwd=2, lty=1, xaxt='n',
        yaxt='n' , xlab='', ylab='', xlim=c(2,21), ylim=c(3,14))
par( new=T )
plot( Ysmooth75~seq(2,21), type='l', lwd=2, lty=2, xaxt='n',
        yaxt='n' , xlab='', ylab='', xlim=c(2,21), ylim=c(3,14))
```

All three loess smooths help visualize the increasing pattern, and also highlight the nonlinearity. At q=0.25 (dotted curve) the loess fit is unnecessarily jagged, sugestinng undersmoothing. The loess smooths at q=0.5 (solid curve) and q=0.75 (dashed curve) are more stable.



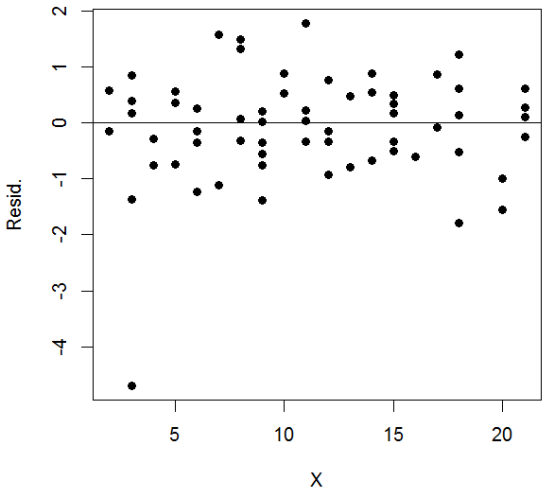
(c) Sample R commands, including reverse transform ($B = \frac{1}{4}\{Y - Y^{-1}\}^2$) to the original scale:

```
Yhat = predict( website50.loess, data.frame(X=12) )
Bhat = 0.25*(Yhat - 1/Yhat)^2
```

which produces 28.75234 (backlogs).

(d) Residual plot, using

```
plot( resid(website50.loess)~X, pch=19, ylab='Resid.' ); abline(
h=0 )
```



appears generally reasonable, except for a potential (large negative) outlier at X = 3 (identified, e.g, via `X[resid(website50.loess) < -4]`). That particular point may require more careful investigation.