STATISTICS 3340/MATH 3340 Final Exam, Sat. Dec. 19, 2015 Solutions

1. A study was carried out to predict the *height* of larch trees using the mineral content of dried needles fallen from the trees. The predictor variables are the percent content of nitrogen *nitro*, the percent content of phosphorus *phos*, the percent content of potassium *potas*, and the percent content of residual ash *ash*. Output follows for the model linear in all predictor variables. Each predictor was centered by subtracting the mean.

```
> larch.out=lm(height~nitro+phos+potas+ash,data=larch2)
> summary(larch.out)

Call:
lm(formula = height ~ nitro + phos + potas + ash, data = larch2)

Residuals:
   Min     1Q Median     3Q    Max
-61.56 -29.11  10.28  24.72  80.29

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  196.577      7.428  26.466  < 2e-16 ***
nitro         97.764     24.572   3.979 0.000684 ***
phos         256.975    169.905   1.512 0.145321
potas        126.573     46.429   2.726 0.012653 *
ash           40.277     36.615   1.100 0.283773
---

Residual standard error: 37.87 on 21 degrees of freedom
Multiple R-squared:  0.8679
F-statistic: 34.48 on 4 and 21 DF,  p-value: 5.967e-09

> anova(larch.out)
Analysis of Variance Table

Response: height
          Df Sum Sq Mean Sq F value    Pr(>F)
nitro      1 152591  152591 106.381 1.124e-09 ***
phos       1  28274   28274  19.711  0.000227 ***
potas      1  15232   15232  10.620  0.003754 **
ash        1   1736    1736   1.210  0.283773
Residuals 21  30122    1434
---
```

(2) (a) What is the total sum of squares?

```
> SST=152591+28274+15232+1736+30122; SST
[1] 227955
```

(2) (b) What is the adjusted R^2 for this model?

```
> SSE=30122
> n=(21+5) #(n-p)+p
> R2adj=1-(SSE/21)/(SST/(n-1))
> R2adj
[1] 0.8426903
```

(2) (c) Explain why *phos* has a large P value in the *summary* output but a small P in the *anova* output.

phos is not an important predictor when nitro, potas and ash are already in the model, but it is an important predictor when only nitro is in the model

(2) (d) Explain how you would test whether the coefficients of *nitro* and *potas* are equal.

Would use the general linear model with $T = (0, 1, 0, -1, 0)$ and $c = 0$.

A second model was fitted which included the interaction between *nitro* and *phos*.

```
> larch2.out=lm(height~nitro+phos+potas+ash+nitro:phos,data=larch2)
> summary(larch2.out)

Call:
lm(formula = height ~ nitro + phos + potas + ash + nitro:phos,
    data = larch2)

Residuals:
    Min      1Q  Median      3Q     Max
-48.540 -26.313   6.115  16.557  67.602

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      185.20       9.52  19.454 1.83e-14 ***
nitro             99.40      23.40   4.247 0.000395 ***
phos             229.46     162.44   1.413 0.173167
potas            128.84      44.21   2.914 0.008574 **
ash               23.51      36.09   0.651 0.522186
nitro:phos       661.50     370.78   1.784 0.089595 .
---


Residual standard error: 36.05 on 20 degrees of freedom
Multiple R-squared:  0.886,     Adjusted R-squared:  0.8575
F-statistic: 31.09 on 5 and 20 DF,  p-value: 8.924e-09

> anova(larch2.out)
Analysis of Variance Table

Response: height
               Df Sum Sq Mean Sq  F value    Pr(>F)
nitro           1 152591  152591 117.4393 8.054e-10 ***
phos            1  28274   28274  21.7603 0.0001491 ***
potas           1  15232   15232  11.7233 0.0026887 **
ash             1   1736    1736   1.3358 0.2613923
nitro:phos      1   4136    4136   3.1829 0.0895948 .
Residuals      20  25986    1299
---
```

(2) (e) Explain why it is a good idea to center the predictors by subtracting their means.
This reduces colinearity in the X matrix. Resulting estimates are not so sensitive to small errors in the predictor variables.

(f) Assess the null hypothesis that *ash* and the interaction *nitro* : *phos* are not needed in the model.

(2) i. State the hypotheses.
$H_0 : \beta_4 = \beta_5 = 0$.
H_A : at least one of β_4, β_5 is non-zero.

(4) ii. Calculate the test statistic.
$F = \frac{(1736+4136)/2}{1299} = 2.26$

(2) iii. Bound the P value as accurately as possible. State the degrees of freedom and give comparison values from the table.
2 numerator and 20 denominator degrees of freedom.

```
> 1-pf(2.26,2,20)
[1] 0.1303453
```

(2) iv. Give a conclusion in the context of the problem. Do not use the word ‘reject’ or ‘accept’.
Very weak evidence that "ash" or the interaction between "nitro" and "phos" are significant.
This is based on a commonly used assessment of p-values as:

- if *pvalue* $< .01$ evidence is very strong,

- if $.01 \leq p - value < .05$ evindence is strong,
- if $.05 \leq p - value < .10$ evindence is weak,
- if $p - value > .10$ evindence is very weak.

(4) (g) Construct a 95% confidence interval for the interaction coefficient.

```
> 661.5 + c(-1,1)*qt(.025,20)*370.78
[1] 1434.9335 -111.9335
```

(2) (h) Which is not an appropriate interpretation for the confidence interval? Circle the Roman numeral.

i. We are 95% confident that the true value falls in this interval.
ii. 95% of intervals constructed in this way will contain the true value of the coefficient.
iii. **The probability is .95 that the true value of the coefficient falls in this interval.**

(8) 2. Match the terms in the list with the corresponding statements below, by writing the letter of the statement after the term

Term	Statement
multicollinearity	k
extrapolation	h
R^2 adjusted	d
quadratic regression	a
interaction	e
residual plots	c
fitted equation	j
indicator variables	l
multiple regression model	i
R^2	f
residual	g
influential points	b

Statements:

(a) Used when a numerical predictor has a curvilinear relationship with the response.

(b) The predictors are a long distance from the means of the predictors.

(c) Used to check the assumptions of the regression model.

(d) Used when trying to decide between two models with different numbers of predictors.

(e) Used when the effect of a predictor on the response depends on other predictors.

(f) Proportion of the variability in y explained by the regression model.

(g) Is the observed value of y minus the predicted value of y for the observed x.

(h) Can give bad predictions if the conditions do not hold outside the observed range of x's.

(i) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1} + \epsilon$.

(j) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_{p-1} x_{p-1}$.

(k) Problem that can occur when the information provided by several predictors overlaps.

(l) Used in a regression model to represent categorical variables.

(3) 3. In a regression problem, the deleted residual at case i is $e_{(i)} = 2.00$, and the leverage value for that case is $h_{ii} = .3$. What is the raw (undeleted) residual for this case?

$e_{(i)}(1 - h_{ii}) = 2(1 - .3) = 1.4$

(3) 4. In a different regression problem, the value for $s^2_{(i)} = MS_{Res,(i)} = 2.25$ is obtained when the ith case is deleted. If the raw residual for case i is 3.75, and $h_{ii} = .3$, what is the value of the externally studentized/standardized residual at case i?

```
> 3.75/(sqrt(2.25)*sqrt(1-.3))

[1] 2.988072
```

5. A regression of $\boldsymbol{y}$ on two sets of predictors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ with $n = 25$ is carried out in stages. The first set of predictors $\boldsymbol{X}_1$ consists of three predictors and the intercept term, and the second set contains a single predictor. First, Y is regressed on $\boldsymbol{X}_1$, giving the residuals $\boldsymbol{e}_1$. The total sum of squares is 120 and the residual sum of squares is 50 for this fit. Secondly, $\boldsymbol{X}_2$ is regressed on $\boldsymbol{X}_1$ giving the residuals $\boldsymbol{e}_2$. Third, $\boldsymbol{e}_1$ is regressed on $\boldsymbol{e}_2$ giving $\hat{e}_1 = -.7e_2$. The regression sum of squares for this fit is 20.

(2) (a) What is the estimated coefficient of X_2 in the regression of Y on both sets of predictors?

-.7

(5) (b) Write the extended ANOVA table showing sums of squares and degrees of freedom.

Source	SS	df	MS	F
$\boldsymbol{X}_1$	70	3	70/3	(70/3)/1.5
$\boldsymbol{X}_2 \vert \boldsymbol{X}_1$	20	1	20	20/1.5
Error	30	20	1.5	
Total	120	24		

(3) (c) In a multiple regression, the variance inflation factor for predictor X_j is $VIF_j = 45$. What proportion of the variation in X_j is explained by the other predictors?

(1-1/45) $\approx$ .98

(2) (d) Give one consequence of extreme multicollinearity.

large estimate variance of $\hat{\beta}_j$'s.

(e) A linear regression model was fitted to $n = 23$ cases. The fitted equation is

$$y = 4.60 + 1.50x_1 - 7.9x_2.$$

and $MS_{Res} = 25$. The $\boldsymbol{X}^T\boldsymbol{X}$ matrix is

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} 23 & 0 & 0 \\ 0 & 100 & 50 \\ 0 & 50 & 150 \end{pmatrix}$$

and its inverse is

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} 1/23 & 0 & 0 \\ 0 & 3/250 & -1/250 \\ 0 & -1/250 & 2/250 \end{pmatrix}$$

(3) i. What is the estimated standard error of $\hat{\beta}_1$?

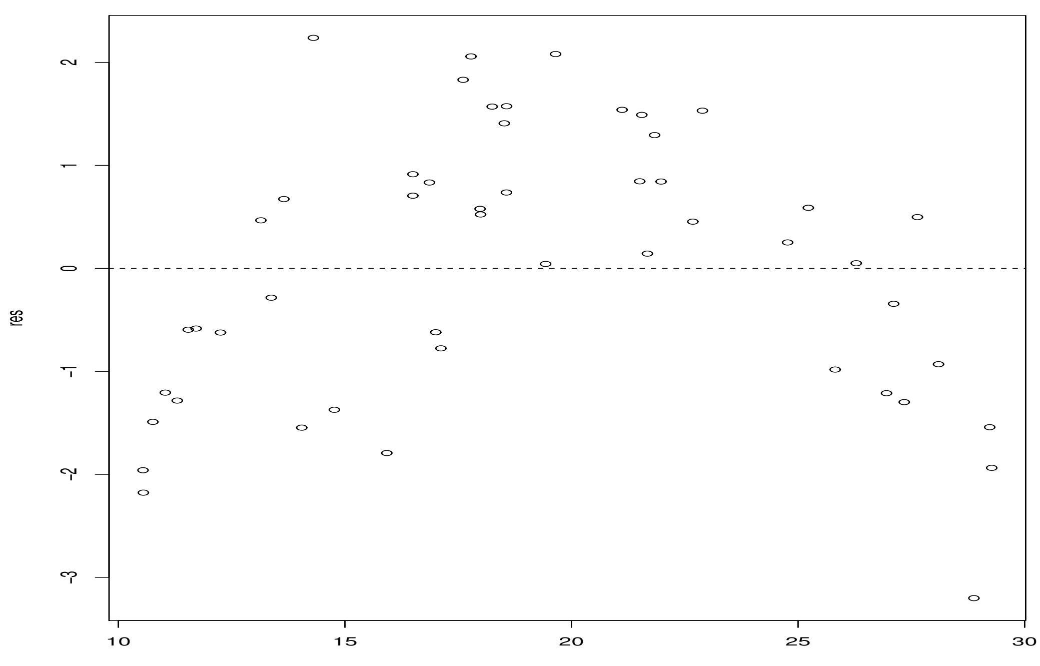
```
> sqrt(25)* sqrt(3/250)
[1] 0.5477226
```

(5) ii. Is the point (2.0,-8.0) in the joint 95% confidence region for β_1 and β_2? (done as in assignment 4, question 1d)

(5) iii. What is the standard error the estimate of the mean response when $x_1 = 5$ and $x_2 = 4$?
Let $x_0 = (1, 5, 4)$. The standard error is

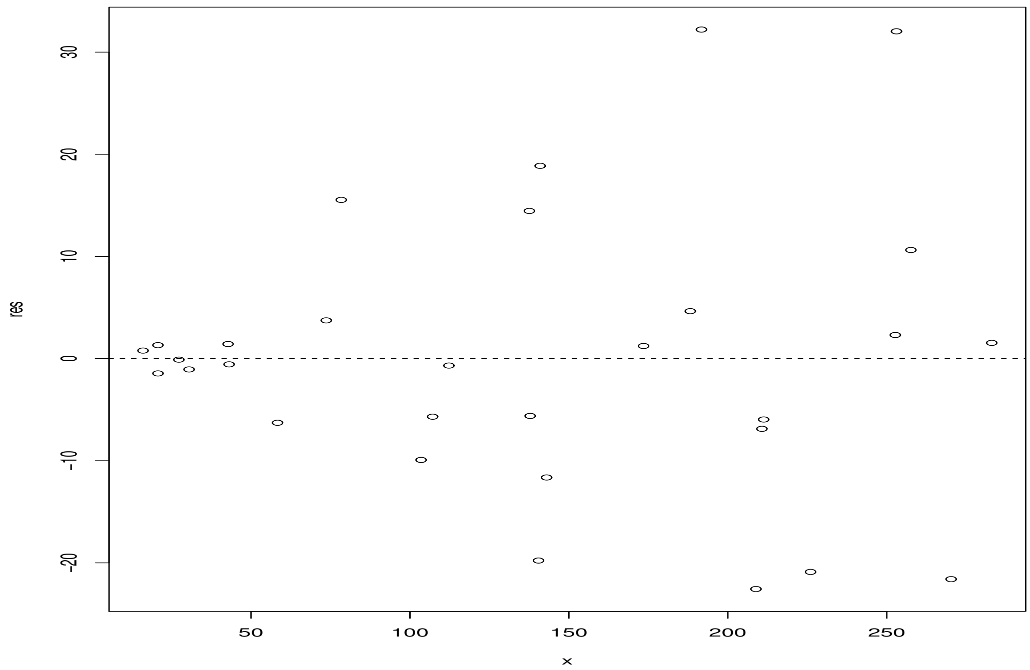
$$\hat{\sigma}\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}$$

(f) A plot of regression residuals versus X_1 follows.



(2) Is there a problem with these residuals? If so, explain how you would change the model.
The residuals have different mean as the predicted value changes. This pattern suggest that a quadratic term in x_1 should be included.

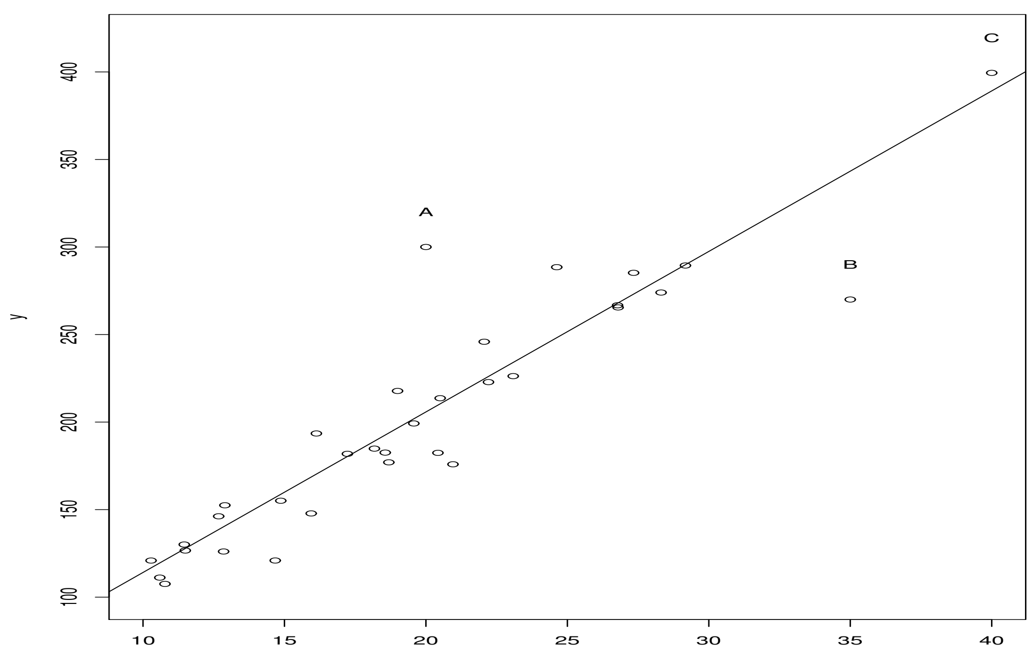
(g) In a multiple regression analysis, a plot of regression residuals versus X_1 follows.



(2) Is there a problem with these residuals? If so, explain how you would change the model.

Yes, variance of residuals increases with predicted value. This suggests a variance stabilizing transformation of y, perhaps a square root or logarithmic transform.

(h) The plot below shows the response y and predictor x to which a simple linear regression model is to be fitted. Three of the cases are labelled, with the label appearing about 20 above the y value.



Which of the labelled points A, B or C

(2) i. Has the highest leverage value?

C is the highest leverage point, because for C, x is furthest from $\bar{x}$.

(2) ii. Has the largest residual (in magnitude)?

A.

(2) iii. Has the largest Cook's distance?

Can't tell from the picture. perhaps B, perhaps A.