DEPARTMENT OF STATISTICS

Course STATS 330/762: Advanced Statistical Modeling/Special Topic in Regression

Term Test: 8.00am - 9:00am, Tuesday September 16, 2008

INSTRUCTIONS

- Answer **ALL 15** questions on the answer sheet provided.
- All questions have a single correct answer and carry the same mark value.
- If you give more than one answer to any question you will receive zero marks for that question.
- A correct answer is worth one point, an incorrect answer zero points.

1. A data set consists of a continuous variable Y, and two explanatory variables X and A (where X is continuous and A is a factor). We want to see if the relationship between X and Y changes with the different values of A. Which of the following pieces of R code would produce the most informative plot?

   (zz) `xyplot(Y~X|A)`

   (1) `xyplot(Y~A|X)`

   (1) `bwplot(Y~X|A)`

   (1) `pairs(data.frame(X,Y,A))`

   (1) `plot(lm(Y~X+A))`

2. Which of the following plots would be **least** helpful in deciding if a 3-dimensional data cloud was planar?

   (zz) A static 3-d plot.

   (1) A coplot.

   (1) A plot of residuals versus fitted values.

   (1) A gam plot.

   (1) A dynamic 3-d plot.

   The following data set will be used for several questions in this test. Criminologists are interested in the effect of punishment regimes and socio-economic factors on crime rates. This has been studied using aggregate data on 47 states of the USA. The data set contains the following variables:

   M: percentage of males aged 14-24

   So: indicator variable for a southern state 1=southern, 0=northern

   Ed: mean years of schooling

   Po1: police expenditure (percapita) in current year

   Po2: police expenditure (percapita) in previous year

   LF: labour force participation rate

   M.F: number of males per 1000 females

   Pop: state population

   NW: number of nonwhites per 1000 people

   U1: unemployment rate of urban males 14-24

   U2: unemployment rate of urban males 35-39

   GDP: gross domestic product per head

   Ineq: income inequality

   Prob: probability of imprisonment

   Time: average time served in state prisons

   y: The crime rate (Larceny per 100,000 population)

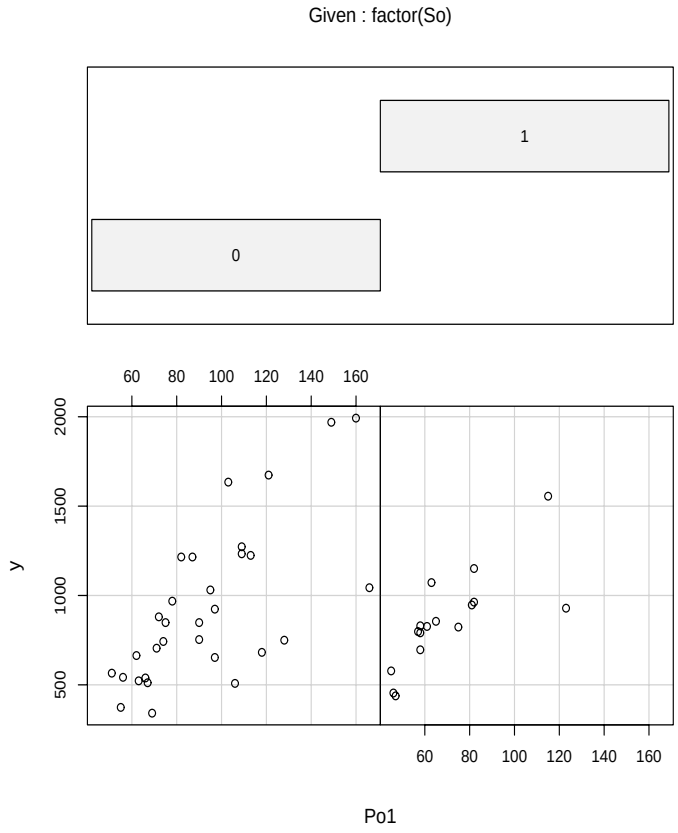

Figure 1: Trellis plot for Question 3.

3. A coplot of the variables `y`, `Po1` and `So` is shown in Figure 1. Which of the following is the best interpretation of this graph?

   (zz) The relationship between the crime rate and expenditure on police in the current year is about the same for both northern and southern states.

   (1) The relationship between the crime rate and being a southern state doesn't depend on expenditure on police.

   (1) It appears that expenditure on police is the same in northern and southern states.

   (1) It appears that expenditure on police does not depend on the crime rate.

   (1) As the crime rate goes up, expenditure on police goes down.

4. In the linear regression model, which is the most important assumption?

   (zz) The mean is a linear function of the explanatory variables.

   (1) The variances are equal.

   (1) The observations are independent.

   (1) The responses are normally distributed.

   (1) There are no high leverage points.

5. The correlation matrix of the crime data is

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	GDP	Ineq	Prob	Time	y
M	1.00	0.58	-0.53	-0.51	-0.51	-0.16	-0.03	-0.28	0.59	-0.22	-0.24	-0.67	0.64	0.36	0.11	-0.09
So	0.58	1.00	-0.70	-0.37	-0.38	-0.51	-0.31	-0.05	0.77	-0.17	0.07	-0.64	0.74	0.53	0.07	-0.09
Ed	-0.53	-0.70	1.00	0.48	0.50	0.56	0.44	-0.02	-0.66	0.02	-0.22	0.74	-0.77	-0.39	-0.25	0.32
Po1	-0.51	-0.37	0.48	1.00	0.99	0.12	0.03	0.53	-0.21	-0.04	0.19	0.79	-0.63	-0.47	0.10	0.69
Po2	-0.51	-0.38	0.50	0.99	1.00	0.11	0.02	0.51	-0.22	-0.05	0.17	0.79	-0.65	-0.47	0.08	0.67
LF	-0.16	-0.51	0.56	0.12	0.11	1.00	0.51	-0.12	-0.34	-0.23	-0.42	0.29	-0.27	-0.25	-0.12	0.19
M.F	-0.03	-0.31	0.44	0.03	0.02	0.51	1.00	-0.41	-0.33	0.35	-0.02	0.18	-0.17	-0.05	-0.43	0.21
Pop	-0.28	-0.05	-0.02	0.53	0.51	-0.12	-0.41	1.00	0.10	-0.04	0.27	0.31	-0.13	-0.35	0.46	0.34
NW	0.59	0.77	-0.66	-0.21	-0.22	-0.34	-0.33	0.10	1.00	-0.16	0.08	-0.59	0.68	0.43	0.23	0.03
U1	-0.22	-0.17	0.02	-0.04	-0.05	-0.23	0.35	-0.04	-0.16	1.00	0.75	0.04	-0.06	-0.01	-0.17	-0.05
U2	-0.24	0.07	-0.22	0.19	0.17	-0.42	-0.02	0.27	0.08	0.75	1.00	0.09	0.02	-0.06	0.10	0.18
GDP	-0.67	-0.64	0.74	0.79	0.79	0.29	0.18	0.31	-0.59	0.04	0.09	1.00	-0.88	-0.56	0.00	0.44
Ineq	0.64	0.74	-0.77	-0.63	-0.65	-0.27	-0.17	-0.13	0.68	-0.06	0.02	-0.88	1.00	0.47	0.10	-0.18
Prob	0.36	0.53	-0.39	-0.47	-0.47	-0.25	-0.05	-0.35	0.43	-0.01	-0.06	-0.56	0.47	1.00	-0.44	-0.43
Time	0.11	0.07	-0.25	0.10	0.08	-0.12	-0.43	0.46	0.23	-0.17	0.10	0.00	0.10	-0.44	1.00	0.15
y	-0.09	-0.09	0.32	0.69	0.67	0.19	0.21	0.34	0.03	-0.05	0.18	0.44	-0.18	-0.43	0.15	1.00

For the crime data, which variables do you think would have high VIF's?

(zz) `Po1` and `Po2`.

(1) `y`.

(1) `GDP` and `y`.

(1) `Time` and `Prob`.

(1) `LF`.

6. In a regression, which of the following is **TRUE**?

(zz) In a regression, an R^2 of 0 means that all the estimated regression coefficients except the constant term are zero.

(1) An R^2 of 0.8 means that no transformation of the response is necessary.

(1) An R^2 of 0.8 means the points are planar.

(1) R^2 is ratio of the residual sum of squares to the total sum of squares.

(1) Unless the R^2 is more that 0.9 the regression is worthless.

7. Suppose we fit a regression to the crime data, using the crime rate (`y`) as the response. The regression summary is shown below.

```
Call:
lm(formula = y ~ ., data = UScrime)

Residuals:
     Min       1Q   Median       3Q      Max
-395.738  -98.088   -6.695  112.989  512.671

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5984.2876  1628.3184  -3.675 0.000893 ***
M               8.7830     4.1714   2.106 0.043443 *
So             -3.8035   148.7551  -0.026 0.979765
Ed             18.8324     6.2088   3.033 0.004861 **
Po1            19.2804    10.6110   1.817 0.078892 .
Po2           -10.9422    11.7478  -0.931 0.358830
LF             -0.6638     1.4697  -0.452 0.654654
M.F             1.7407     2.0354   0.855 0.398995
Pop            -0.7330     1.2896  -0.568 0.573845
NW              0.4204     0.6481   0.649 0.521279
U1             -5.8271     4.2103  -1.384 0.176238
U2             16.7800     8.2336   2.038 0.050161 .
GDP             0.9617     1.0367   0.928 0.360754
Ineq            7.0672     2.2717   3.111 0.003983 **
Prob        -4855.2658  2272.3746  -2.137 0.040627 *
Time           -3.4790     7.1653  -0.486 0.630708
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 209.1 on 31 degrees of freedom
Multiple R-squared: 0.8031,     Adjusted R-squared: 0.7078
F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```

Which one of the following statements is **TRUE**?

(zz) Assuming the other variables are held constant, the crime rate in a southern state is about 3.8 per 100,000 less than in a non-southern state.

(1) Since the regression coefficient of U2 is quite large, this variable should be retained in the regression.

(1) There seems to be something wrong with the coefficient of Prob; it is too big.

(1) With a unit increase in `Ineq`, the crime rate goes down by about 7 per 100,000.

(1) The estimate of the variance of the errors is 209.1.

8. Some diagnostic plots for the regression fitted in Question 7 are shown in Figure 2.

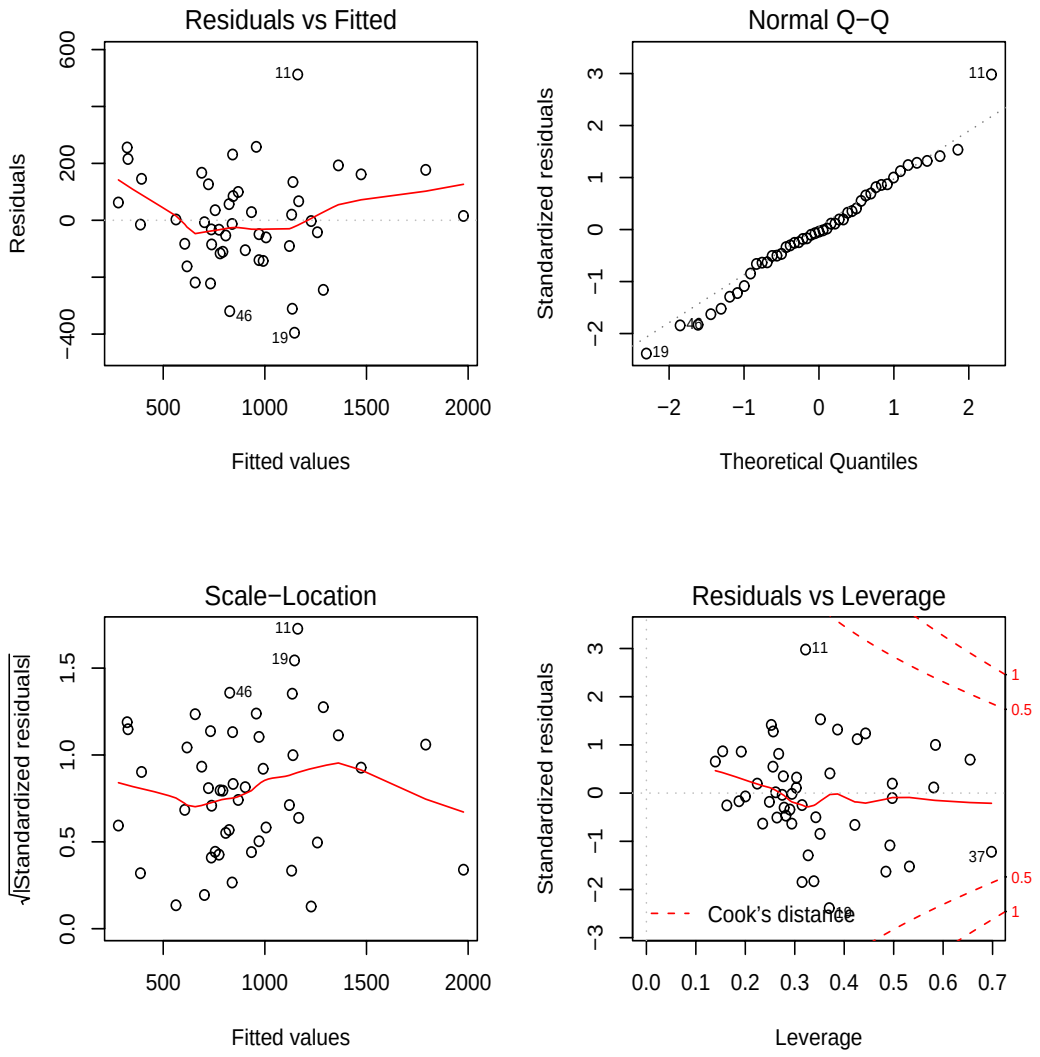

Figure 2: Diagnostic plots for Question 7.

Which of the following is the correct interpretation?

(zz) No point has a large Cook's distance.

(1) The errors seem to be right-skewed.

(1) There are really large outliers in the data.

(1) The variance of the errors seems to be increasing with the mean.

(1) There is evidence of serial correlation.

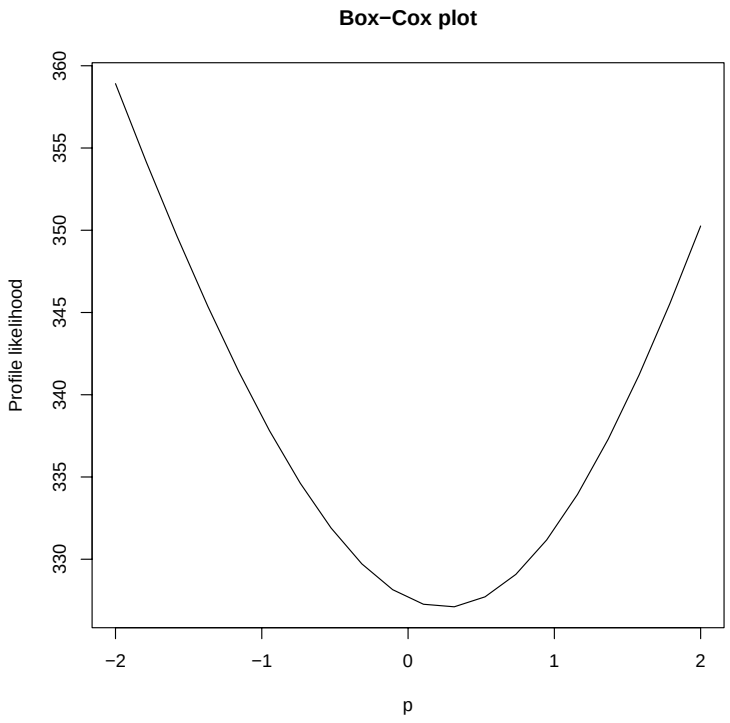

Figure 3: Box-Cox plot for Question 9.

9. A Box-Cox plot is shown in Figure 3. Which is the **BEST** interpretation of this plot?

   (zz) The response `y` should be transformed with power 1/4.

   (1) No transformation of the response is indicated.

   (1) No transformation of the explanatory variables is indicated.

   (1) The response `y` should be transformed with a reciprocal transformation.

   (1) The response should be transformed as a quadratic.

10. Suppose we wanted to predict the crime rate in one of the three states not included in the data set. The data for this state is in the data frame `newdata`. We get the following output:

```
> crime.lm = lm(y~., data=UScrime)
> predict(crime.lm, newdata = newdata, se=TRUE)
$fit
     48
991.763

$se.fit
[1] 123.8652
```

```
$df
[1] 31

$residual.scale
[1] 209.0644
> qt(0.975, 31)
[1] 2.039513
```

Which of the following is **TRUE**?

(zz) A 95% prediction interval for the crime rate of this state is is (496.1548, 1487.371).

(1) A 95% prediction interval for the crime rate of this state is is (739.1381, 1244.388).

(1) A 95% confidence interval for the mean crime rate is (496.1548, 1487.371).

(1) A 95% confidence interval for the mean crime rate is (523.4449, 1476.0171).

(1) The estimate of σ used is 14.459.

11. Some influence plots for the model fitted in Question 7 are shown in Figure 4. Which of the following statements is **NOT** a correct interpretation of these plots?

(zz) Observation 11 has a very big studentized residual.

(1) Several points are having an influence on the standard errors.

(1) No points have high leverage.

(1) Observation 11 is having an effect on the fitted value for point 11.

(1) No points are having an excessive influence on the regression coefficients.

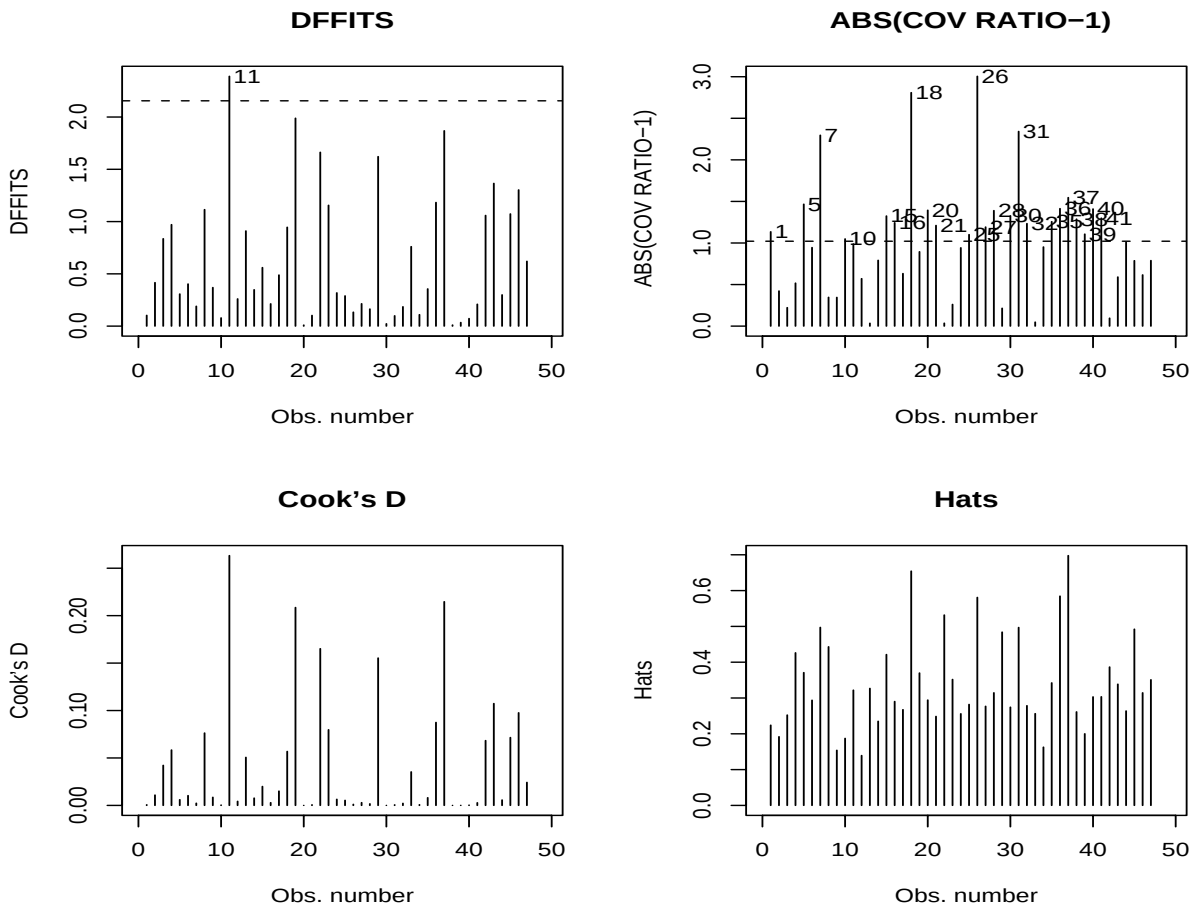

Figure 4: Influence plots for Question 10

12. In a regression with data collected sequentially in time, we suspect serial correlation is present. We calculate a Durbin-Watson test statistic of 1.50. The values of d_l and d_u are 1.44 and 1.54 respectively. What do we conclude?

(zz) The test is inconclusive, but there is certainly no evidence of negative serial correlation.

(1) There is nothing to indicate the data are serially correlated.

(1) There is evidence of positive serial correlation.

(1) The test is inconclusive, but there is certainly no evidence of positive serial correlation.

(1) There is evidence of negative serial correlation.

13. An all possible regressions was run on the model fitted in Q7 with the following results:

```
> all.poss.regs(y~., data=UScrime)
      rssp   sigma2 adjRsq     Cp    AIC    BIC       CV
1  3627626 80613.91  0.461 39.997 86.997 90.697 364163.7
2  2887807 65631.98  0.561 25.071 72.071 77.621 308070.7
3  2300757 53505.99  0.642 13.639 60.639 68.040 255880.1
4  2061353 49079.83  0.672 10.162 57.162 66.413 236224.1
5  1803290 43982.69  0.706  6.258 53.258 64.359 216904.7
6  1611057 40276.42  0.731  3.860 50.860 63.811 202861.4
7  1551147 39773.00  0.734  4.489 51.489 66.290 209867.6
8  1453068 38238.62  0.744  4.245 51.245 67.896 202094.9
9  1426575 38556.07  0.742  5.639 52.639 71.140 210280.5
10 1404229 39006.36  0.739  7.128 54.128 74.479 218575.2
11 1387523 39643.51  0.735  8.745 55.745 77.947 229517.1
12 1375848 40466.12  0.729 10.478 57.478 81.530 241861.8
13 1365315 41373.18  0.723 12.237 59.237 85.139 261522.9
14 1354974 42342.95  0.717 14.001 61.001 88.753 290788.7
15 1354946 43707.93  0.708 16.000 63.000 92.602 314255.4

   M So Ed Po1 Po2 LF M.F Pop NW U1 U2 GDP Ineq Prob Time
1  0  0  0   1   0  0   0   0  0  0  0   0    0    0    0
2  0  0  0   1   0  0   0   0  0  0  0   0    1    0    0
3  0  0  1   1   0  0   0   0  0  0  0   0    1    0    0
4  1  0  1   1   0  0   0   0  0  0  0   0    1    0    0
5  1  0  1   1   0  0   0   0  0  0  0   0    1    1    0
6  1  0  1   1   0  0   0   0  0  0  1   0    1    1    0
7  1  0  1   1   0  0   0   0  0  0  1   1    1    1    0
8  1  0  1   1   0  0   1   0  0  1  1   0    1    1    0
9  1  0  1   1   0  0   1   0  0  1  1   1    1    1    0
10 1  0  1   1   0  0   1   1  0  1  1   1    1    1    0
11 1  0  1   1   1  0   1   1  0  1  1   1    1    1    0
12 1  0  1   1   1  0   1   1  1  1  1   1    1    1    0
13 1  0  1   1   1  1   1   1  1  1  1   1    1    1    0
14 1  0  1   1   1  1   1   1  1  1  1   1    1    1    1
15 1  1  1   1   1  1   1   1  1  1  1   1    1    1    1
```

Which is the **wrong** interpretation?

(zz) Model 15 is the best model.

(1) Model 6 is a good model.

(1) Model 8 is a good model.

(1) Model 10 is a worse model than model 8.

(1) All possible regressions is a better method than stepwise regression.

14. Suppose we have a set of possible explanatory variables, and want to select a subset that gives a good model. Which of the following statements is **TRUE**?

    (zz) Choosing the model that has the biggest adjusted R^2 and choosing the model with the smallest residual standard error results in the same model.

    (1) There is no problem if we put too many variables in the model.

    (1) Stepwise regression always gives a smaller model than all possible regressions.

    (1) AIC tends to select smaller models (i.e. having fewer variables) than BIC.

    (1) We should pick the model with the biggest R^2.

15. Which of the following is **FALSE**?

    (zz) The hat matrix diagonal measures the extent to which a point is an outlier.

    (1) The hat matrix diagonals always lie between 0 and 1.

    (1) The hat matrix diagonals tell us if any explanatory variable has an extreme value.

    (1) "Externally Studentized" means the raw residual is divided by its standard deviation.

    (1) The studentized residuals do not depend on the units of measurement of the response variable.