DEPARTMENT OF STATISTICS

STATS 330/762: Advanced Statistical Modelling/Special Topic in Regression

Term Test: 8.00am - 9:00am, Thursday September 15, 2011

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. Suppose we have a data set consisting of three continuous variables X, Y and Z. We want to fit a regression model using Y as the response and X and Z as the explanatory variables. Which of the following plots **WOULD NOT** be useful for examining the suitability of the standard regression model **BEFORE** any model was fitted?

   (1) A 3-d "spinning" plot.

   (2) A coplot of Y versus X conditioning on Z.

   (3) A plot produced by the R function `cloud`.

   (4) A normal plot of Y.

   (5) A coplot of Y versus Z conditioning on X.

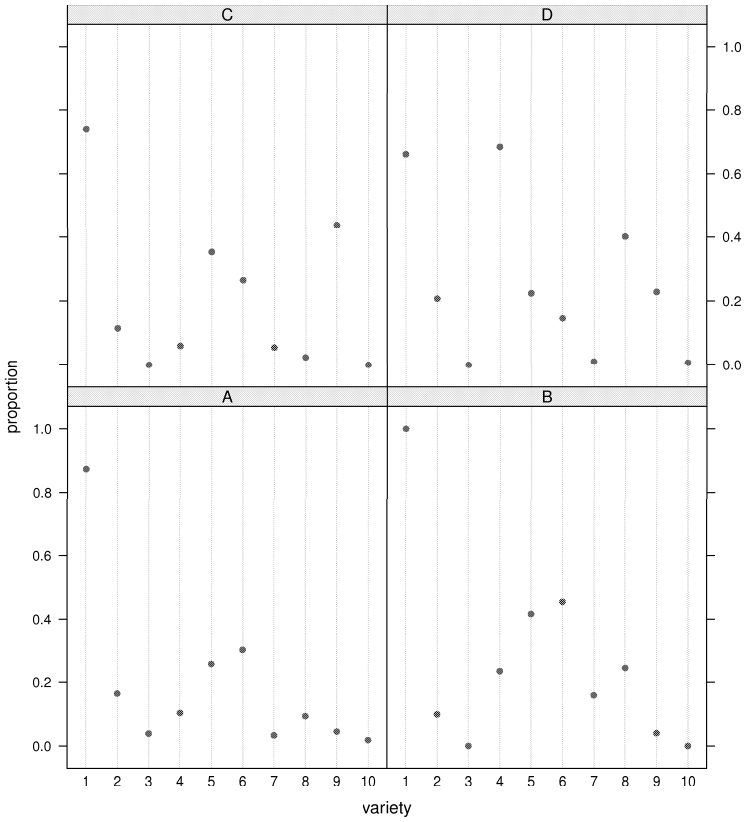

Figure 1: Trellis plot for Question 2.

2. In an experiment to examine the resistance of different varieties of sugar cane to a certain disease, specimen plants were deliberately infected with the disease and then planted out in one of four blocks. The proportion of plants dying was recorded, along with the variety and which block (one of A, B, C, D) the specimen was planted in. Figure 1 shows a trellis plot of the data, which consists of three variables, `proportion`, `variety` and `block`.

Which of the following is **FALSE**?

(1) Variety six does worst in block D.

(2) Varieties three and ten do not seem to be very vulnerable to the disease.

(3) With the exception of variety one, the varieties in block A have less variable proportions than the other blocks.

(4) Overall, variety one seems the most vulnerable to the disease.

(5) Some varieties do better in some blocks than others.

3. Which of the following bits of R code produced the graph in Figure 1?

(1) `xyplot(variety~proportion|block, xlab = "variety",ylab = "proportion")`

(2) `bwplot(proportion~variety|block, xlab = "variety",ylab = "proportion")`

(3) `dotplot(variety~block|proportion, xlab = "variety",ylab = "proportion")`

(4) `dotplot(proportion~variety|block, xlab = "variety",ylab = "proportion")`

(5) `dotplot(variety~proportion|block, xlab = "variety",ylab = "proportion")`

4. When fitting a linear regression model, which of the following is the **BEST** plot for detecting serial correlation in the errors?

(1) An acf plot of the responses.

(2) An acf plot of the residuals.

(3) A plot of residuals versus fitted values.

(4) A normal plot of the errors.

(5) A Box-Cox plot.

5. In a regression, a pair of explanatory variables $X1$ and $X2$ have a correlation of 0.95, and p-values considerably greater than 0.05. Which of the following is the **WORST** interpretation?

(1) It is possible that X1 is related to the response.

(2) One or more of the VIF's will be large.

(3) Both variables are unimportant in the regression.

(4) Either X1 or X2 could be deleted from the regression.

(5) It is possible that X2 is related to the response.

6. Which of the following statements is **correct**?

   (1) The adjusted R^2 is always less than the R^2.

   (2) Deleting a variable from the regression cannot increase the R^2.

   (3) In a regression the R^2 will be equal to 1 if all the estimated regression coefficients (apart from the constant term) are zero.

   (4) Deleting an observation from the regression always decreases the estimate of error variance.

   (5) Deleting an observation from the regression always decreases the R^2 .

Questions 7-10 are concerned with the following problem: Every year, the magazine US News and World Report publishes a list of rankings for US colleges and universities. The construction of the ranking involves predicting the graduation rate from other variables relating to the composition of the student body. A data set containing information on 173 US colleges is available for analysis.

The variables in the data set are

university: the name of the college,

grad.rate: the graduation rate (the percentage of the entering class of 2001 that graduated by 2006),

fresh.ret: the freshman retention rate (percentage of first year students that return),

per20: percentage of classes with fewer than 20 students,

per50: percentage of classes with more than 50 students,

SAT75: 75th percentile of the SAT scores of admitted students (the SAT is a standard college admission test),

top10: percentage of freshmen in top 10school class,

accept.rate: The percent of applicants accepted for admission,

alumni.giving: The percentage of alumni who contribute money.

7. A regression model was fitted to these data and the summary is shown below:

```
> college.lm = lm(grad.rate~.-University,data=college.df)
> summary(college.lm)

Call:
lm(formula = grad.rate ~ ., data = colleges.df[, -1])

Residuals:
      Min        1Q    Median        3Q       Max
-0.254498 -0.037653  0.009696  0.039694  0.142059
```

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.499e-01  1.264e-01  -6.726 2.73e-10 ***
fresh.ret      1.281e+00  9.581e-02  13.375  < 2e-16 ***
per20          6.221e-02  6.311e-02   0.986 0.325697
per50          3.744e-02  1.232e-01   0.304 0.761524
SAT75          2.231e-04  9.764e-05   2.285 0.023608 *
top10          4.501e-02  4.101e-02   1.098 0.274018
accept.rate    5.186e-02  4.092e-02   1.267 0.206885
alumni.giving  2.346e-01  6.723e-02   3.490 0.000619 ***
Residual standard error: 0.06582 on 165 degrees of freedom
Multiple R-Squared: 0.8679,     Adjusted R-squared: 0.8623
F-statistic: 154.8 on 7 and 165 DF,  p-value: < 2.2e-16
```

Which of the following is **NOT** indicated by this output?

(1) The graduation rate goes down 2.231e-04% if the average 75th percentile of the SAT scores goes up by one.

(2) Freshman retention helps predict the graduation rate

(3) It seems that the percentage of classes under 20 and the percentage of classes under 50 are not both required in the model.

(4) At least one of the explanatory variables is related to the response.

(5) Alumni giving helps predict the graduation rate.

8. Some diagnostic plots from the regression are shown in Figure 2.

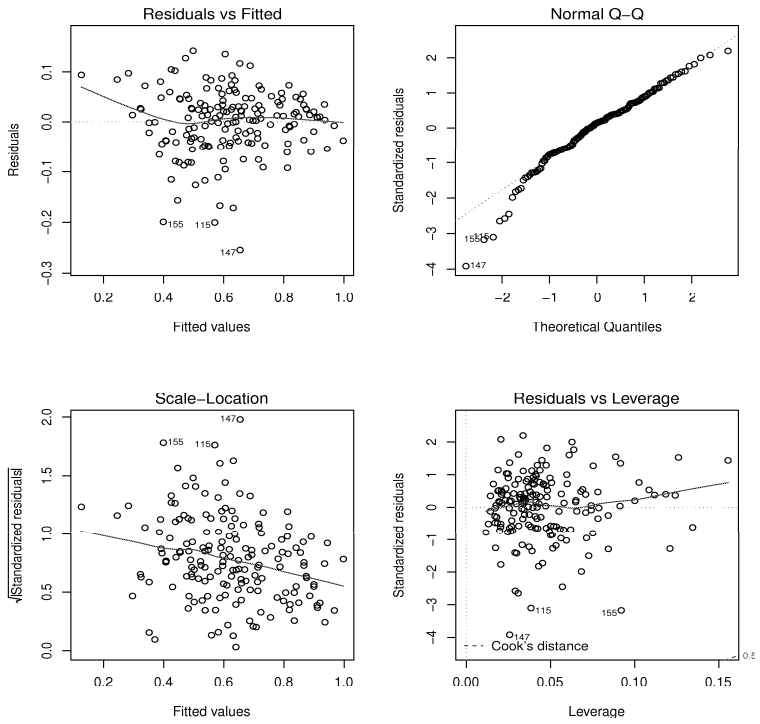

Figure 2: Diagnostic plots for Question 8.

Which of the following is **TRUE**?

(1) There is evidence the data are not planar.

(2) The errors seem to have two heavy tails.

(3) There is evidence that the variance increases with the mean.

(4) The graphs indicate that there are no serious problems with the regression.

(5) Some points have a high Cook's distance.

9. Here is some R output relating to influence measures for the college data. Note that the column on the left is the observation number: only points that are influential on the various criteria are shown.

	dfb.1_	dfb.frs.	dfb.pr20	dfb.pr50	dfb.SAT7	dfb.tp10	dfb.acc.	dfb.alm.	dffit	cov.r	cook.d	hat
2	-0.047	0.043	-0.034	-0.060	0.041	0.010	0.049	-0.189	-0.241	1.191	0.007	0.135
32	0.000	0.005	0.026	0.037	-0.014	0.024	0.002	-0.022	0.069	1.164	0.001	0.100
39	0.060	0.043	0.010	0.002	-0.107	0.137	0.010	-0.052	0.187	1.162	0.004	0.109
44	0.016	0.011	-0.012	0.020	-0.034	0.088	0.022	-0.042	0.135	1.174	0.002	0.112
71	-0.004	-0.006	-0.008	-0.005	0.010	0.001	0.004	0.001	0.014	1.155	0.000	0.091
80	0.036	0.008	-0.001	0.000	-0.064	0.115	0.043	-0.019	0.140	1.191	0.002	0.124
81	0.032	-0.032	-0.018	-0.016	-0.032	0.129	0.053	-0.023	0.148	1.182	0.003	0.119
115	-0.070	-0.176	0.265	0.450	-0.025	0.063	0.199	0.297	-0.637	0.675	0.048	0.038
118	0.097	-0.201	0.074	0.131	-0.153	0.424	0.327	0.210	0.618	1.124	0.047	0.156
123	-0.009	-0.152	-0.025	0.031	0.191	-0.067	-0.246	-0.236	-0.454	0.778	0.025	0.029
128	0.240	-0.086	0.062	0.012	-0.295	0.237	0.052	0.322	-0.481	0.765	0.028	0.031
147	-0.166	-0.262	0.256	0.054	0.328	-0.178	-0.070	-0.261	-0.666	0.493	0.051	0.026
155	0.028	0.259	-0.590	-0.026	-0.105	-0.011	-0.051	0.497	-1.039	0.699	0.128	0.092
164	-0.528	0.016	0.221	0.224	0.418	-0.172	0.326	-0.189	-0.613	0.827	0.045	0.057

Hint for question 9: Points are influential if (i) $|DFBETAS| > 1$, (ii) $|DFFITS| > 3\sqrt{p/(n-p)}$, (iii) $|COVRATIO - 1| > 3p/n$.

Which of the following is **FALSE**?

(1) Point 115 is having an effect on the standard errors.

(2) Point 118 is influential.

(3) Point 147 is having an effect on its fitted value.

(4) Point 118 has high leverage.

(5) Point 164 is having an effect on the standard errors.

10. Which of the following statements is **FALSE**, based on the all possible regressions output below?

```
all.poss.regs(grad.rate~.-University,data=college.df, dp=4)
    rssp sigma2 adjRsq      Cp      AIC      BIC     CV
1 0.9351 0.0055 0.8261 46.8830 219.8830 226.1896 0.0950
2 0.7844 0.0046 0.8533 14.0789 187.0789 196.5388 0.0804
3 0.7314 0.0043 0.8624  3.8528 176.8528 189.4659 0.0759
4 0.7252 0.0043 0.8628  4.4079 177.4079 193.1743 0.0761
5 0.7201 0.0043 0.8629  5.2503 178.2503 197.1701 0.0763
6 0.7151 0.0043 0.8630  6.0924 179.0924 201.1654 0.0771
7 0.7147 0.0043 0.8623  8.0000 181.0000 206.2263 0.0780
```

	fresh.ret	per20	per50	SAT75	top10	accept.rate	alumni.giving
1	1	0	0	0	0	0	0
2	1	0	0	1	0	0	0
3	1	0	0	1	0	0	1
4	1	0	0	1	1	0	1
5	1	0	0	1	1	1	1
6	1	1	0	1	1	1	1
7	1	1	1	1	1	1	1

(1) The model `grad.rate~fresh.ret+SAT75+alumni.giving` is indicated by the CV criterion.

(2) The model `grad.rate~fresh.ret+SAT75+alumni.giving` is indicated by the AIC criterion.

(3) The model `grad.rate~fresh.ret+per20+SAT75+top10 + accept.rate + alumni.giving` is indicated by the adjusted R^2 criterion.

(4) The model `grad.rate~fresh.ret+SAT75+alumni.giving` is indicated by the BIC criterion.

(5) The model `grad.rate~fresh.ret+SAT75+alumni.giving` is indicated by the `sigma2` criterion.

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

(1) The estimate of prediction error based on the residual sum of squares is usually optimistic.

(2) The R^2 of a model selected by stepwise regression will usually be too high.

(3) In the regression summary of a model selected by stepwise regression, the p-values can't be trusted.

(4) Overfitting doesn't matter because we have included all relevant variables.

(5) Underfitting leads to biased estimates.

12. In class, we studied remedies for regression when the variances of the errors are not equal. Which of the following statements relating to these remedies is **FALSE**?

(1) We should estimate the variances by smoothing the plot of residuals versus fitted values.

(2) If the variances are functions of the mean, we can detect this by examining a plot of the square root of the absolute value of the residuals versus the fitted values.

(3) We could use weighted least squares with weights inversely proportional to the variances.

(4) If the variances are functions of the mean, we could try transforming the response.

(5) If we take no action, the estimated standard errors of the regression coefficients will be incorrect.

13. Suppose we have a continuous response Y, a categorical explanatory variable A and a continuous explanatory variable X .Which of the following is **FALSE**?

    (1) The code `lm(Y∼X)` fits a simple linear regression model.

    (2) In the "non-parallel lines" model, the constant term corresponds to the slope of the baseline.

    (3) The code `lm(Y∼A)` fits a one-way analysis of variance model.

    (4) The code `lm(Y∼A + X)` fits a the parallel lines model.

    (5) The coefficient of X when fitting the parallel lines model represents the slope of the lines.

Questions 14 and 15 are based on the following: The data frame `cats.df` contains data on 97 male and female cats. There are three variables

Sex: A factor for the sex of the cat (levels are F and M).

Bwt: Body weight in kg.

Hwt: Heart weight in g.

14. We first fitted a model to the data using `Hwt` as the response and obtained the output below:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9813     1.8428   1.618 0.107960
SexM         -4.1654     2.0618  -2.020 0.045258 *
Bwt           2.6364     0.7759   3.398 0.000885 ***
SexM:Bwt      1.6763     0.8373   2.002 0.047225 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 1.442 on 140 degrees of freedom
Multiple R-squared: 0.6566,     Adjusted R-squared: 0.6493
F-statistic: 89.24 on 3 and 140 DF,  p-value: < 2.2e-16
```

Which of the following is **FALSE**?

(1) The fitted line for female cats has intercept 2.9813.

(2) The fitted line for male cats has intercept -1.1841.

(3) The fitted line for female cats has slope 2.6364.

(4) The fitted line for male cats has has slope 0.9601.

(5) The baseline corresponds to female cats.

15. Base your answer to Q15 on the output above and the additional output below.

```
> anova(cats.lm, cats1.lm)
Analysis of Variance Table

Model 1: Hwt ~ Sex + Bwt
Model 2: Hwt ~ Sex * Bwt
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    141 299.38
2    140 291.05  1    8.3317 4.0077 0.04722 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> anova(cats2.lm, cats1.lm)
Analysis of Variance Table

Model 1: Hwt ~ Bwt
Model 2: Hwt ~ Sex * Bwt
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    142 299.53
2    140 291.05  2    8.4865 2.0411 0.1337
```

Which of the following is **FALSE**?

(1) There is strong evidence of a relationship between body weight and heart weight.

(2) The "parallel lines" model has a residual sum of squares that is 8.3317 more than that of the "non-parallel lines" model.

(3) There is no real evidence that the lines for male and female cats are different.

(4) Under the "non-parallel" lines model, the estimated mean weight for male cats with a bodyweight of 3 kg is about 10.89 gm.

(5) Assuming that there are different lines for male and female cats, there is some evidence that the lines are not parallel.