

DEPARTMENT OF STATISTICS

Paper 475.330

Term Test

Date: Sept. 15

**Note 1:** Each part within question is worth equal marks.

**Note 2:** Short answers are preferable to long answers.

1. The following output is produced by Splus for a regression analysis.

Call:

```
lm(formula = dist ~ speed + I(speed^2), data = cars)
```

Residuals:

Min	1Q	Median	3Q	Max
-28.720	-9.184	-3.188	4.628	45.152

Coefficients:

	Estimate	Std.Error	t Value	Pr(> t )
(Intercept)	2.4701	14.8172	0.1667	0.8683
speed	0.9133	2.0342	0.4490	0.6555
I(speed^2)	0.1000	0.0660	1.5153	0.1364

Residual standard error: 15.18 on 47 degrees of freedom

Multiple R-Squared: 0.6673

F-statistic: 47.14 on 2 and 47 degrees of freedom,  
the p-value is 5.852e-12

- a) Write down the *theoretical* form of the model being fitted here (i.e. in the form of an equation containing  $\beta$ s).
- b) There are four  $p$ -values in the output. For each of these  $p$ -values, write down the null and alternative hypothesis being tested.
- c) Consider the statement:

*The coefficient of speed gives an estimate of the change in dist which will occur if speed increases by 1.*

Do you agree or disagree with this statement? Explain why or why not.

- d) None of the model coefficients appear to be significant and yet the overall  $F$ -statistic is highly significant. Explain why this is.

2. a) What is the principal problem you would expect to detect by examining a plot of residuals versus fitted values? How would you correct this kind of problem?

- b) Explain how (don't give Splus commands) to produce a partial regression plot. Under what circumstances would you expect a partial regression plot to look different from the corresponding plot of residuals against independent variable?
- c) What assumption would a plot showing the  $i$ th residual plotted against the  $i - 1$ st residual be checking? What feature in the plot would indicate that this assumption was not satisfied.
- d) What assumption is made about the distribution of the errors in a regression analysis? How could this assumption be checked?
3. a) Explain what it means for a point to have high leverage in a regression analysis. What diagnostic quantity gives a direct measure of leverage?
- b) What kind of problem is indicated by a significant *covariance ratio*? Approximately what value should a covariance ratio take on for an observation which is causing no problems in a regression analysis?
- c) Suppose a observation has a large negative DFFITS value. What effect would leaving that observation out of the analysis have on the results?
4. a) Why is it not useful to use  $R^2$  as a guide when selecting a regression subset? Name a more useful criterion.
- b) Suppose that you have selected a set of predictor variables by stepwise regression. Would you expect all the coefficients of the variables to be significant? Explain.