Total marks = 30

Notes: Short answers are preferable to long answers (answers in "point form" are acceptable).

1. The multiple regression model can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon$$

where ϵ represents the error term. What three assumptions are made about the errors in a standard regression model? For each assumption, identify one diagnostic plot that can be used to check that assumption. [3 marks]

2. Partial regression plots can be used to help decide if a particular regressor should be included in the model. Suppose that partial regression plots are constructed for a dataset that consists of a response, Y, and regressors X_1, X_2, ..., X_k. Consider the partial regression plot for X_1. Describe what this plot would look like for each of the following situations:

   (a) X_1 does not need to be included in the regression model.

   (b) X_1 should be included in the model and does not need to be transformed.

   (c) X_1 should be included in the model but it should either be transformed or a polynomial of X_1 should be used. [3 marks]

3. The influence of an observation can be investigated using leave-one-out diagnostics. For each of the following situations explain what effect leaving out the observation would have on the fitted model.

   (a) An observation that has an unusual covariance ratio.

   (b) An observation that has a large value of Cook's distance.

   (c) An observation that has a large value of DFFITS. [3 marks]

4. Multicollinearity is one problem that can occur in a dataset.

   (a) What is meant by multicollinearity?

   (b) What diagnostic is used to detect multicollinearity? How does this diagnostic indicate multicollinearity is present?

   (c) Name one problem that is caused by multicollinearity? [3 marks]

Questions 5 and 6 both refer to the following data. The selling prices (in British pounds) at auction of 32 antique grandfather clocks were recorded. The age of each clock (in years) and the number of people who participated in the bidding were also recorded.

Age	Bidders	Price	Age	Bidders	Price	Age	Bidders	Price
127	13	1235	115	12	1080	127	7	845
150	9	1522	156	6	1047	182	11	1979
156	12	1822	132	10	1253	137	9	1297
113	9	946	137	15	1713	117	11	1024
137	8	1147	153	6	1092	117	13	1152
126	10	1336	170	14	2131	182	8	1550
162	11	1884	184	10	2041	143	6	854
159	9	1483	108	14	1055	175	8	1545
108	6	729	179	9	1792	111	15	1175
187	8	1593	111	7	785	115	7	744
194	5	1356	168	7	1262			

5. Splus was used to fit a multiple regression model that relates the expected selling price (Price) to the age of the clock (Age) and the number of people bidding (Bidders).

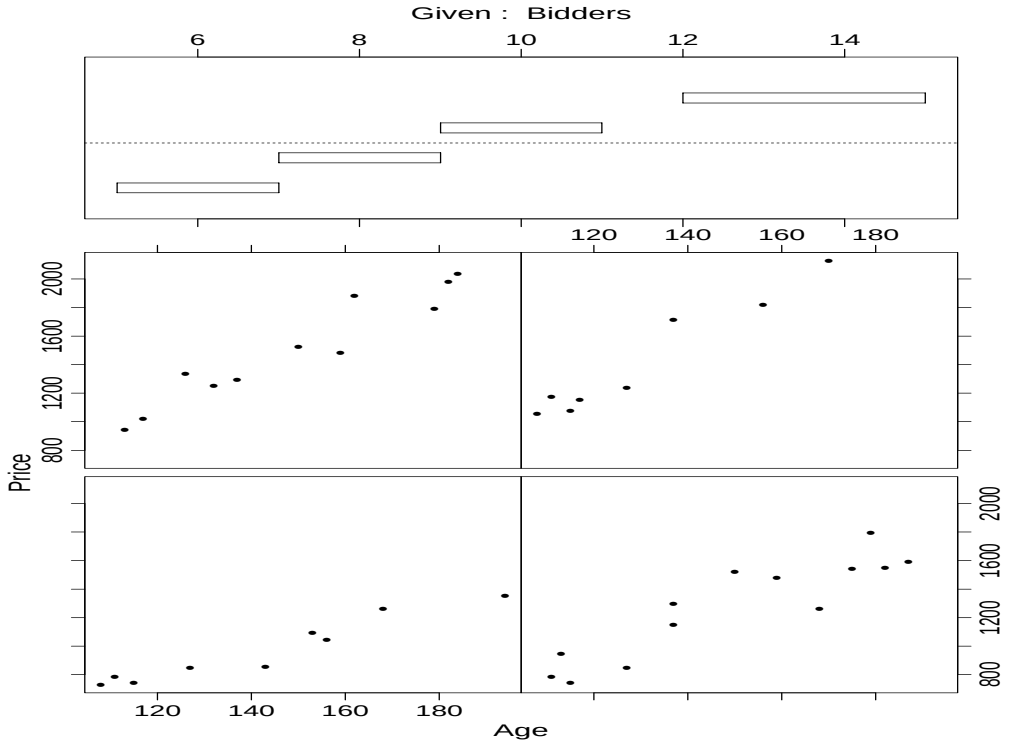
```
> clock.fit<-lm(Price~Age+Bidders,data=clock.df)
> summary(clock.fit)
                 Value Std. Error    t value   Pr(>|t|)
(Intercept) -1336.7221   173.3561    -7.7108     0.0000
        Age    12.7362     0.9024    14.1140     0.0000
    Bidders    85.8151     8.7058     9.8573     0.0000

Residual standard error: 133.1 on 29 degrees of freedom
Multiple R-Squared: 0.8927
F-statistic: 120.7 on 2 and 29 degrees of freedom,
                                the p-value is 8.771e-15
```

(a) Write down the fitted model. Use the fitted coefficients for Age and Bidders to explain what this model indicates about the relationship between Price and the explanatory variables (be precise in your explanation). [3 marks]

(b) The p-value for the overall F-test is very small. What hypothesis is this p-value testing? Explain what the result of this F-test indicates about the fitted regression model. [3 marks]

(c) Suppose an antique dealer claims that the value of an antique grandfather clock increases by 100 pounds for each additional bidder present at the auction. Assuming that the fitted model is reasonable, suggest a hypothesis test that could be used to check the validity of this statement. To answer this question you should state the hypothesis being tested, show how you would calculate the test statistic, and explain how you would obtain the p-value (note that you won't be able to actually calculate the p-value). [4 marks]

6. A trellis plot of Price versus Age conditioned on the level of Bidders was created using coplot.



This plot suggests that Age and Bidders interact. The output for the regression model containing the Age:Bidders interaction is:

```
                Value Std. Error   t value  Pr(>|t|)
(Intercept)  322.7544  293.3251     1.1003    0.2806
        Age    0.8733    2.0197     0.4324    0.6688
    Bidders  -93.4099   29.7077    -3.1443    0.0039
Age:Bidders    1.2979    0.2110     6.1504    0.0000

Residual standard error: 88.37 on 28 degrees of freedom
Multiple R-Squared: 0.9544
F-statistic: 195.2 on 3 and 28 degrees of freedom, the p-value is 0
```

(a) What is meant by "Age and Bidders interact". Explain what feature of the coplot indicates that there is an interaction between Age and Bidders (your explanation should clearly identify how the coplot would be different if the two variables did not interact). [4 marks]

(b) Write down the fitted model. Write a short paragraph that explains how E(Price) is related to Bidders using this fitted model. [4 marks]