INSTRUCTIONS TO CANDIDATES

1. This examination paper contains **FOUR (4)** questions and comprises **SEVEN (7)** printed pages.

2. Answer **ALL** the questions. The total mark is 100 points.

3. All **NOTATION** used here is the same as used in the lecture notes.

4. Write your answers **NEATLY** and label them **CLEARLY** with the associated question number in **your answer book**.

5. This is a **Closed** book examination but non-programmable calculators are allowed.

6. Candidates may bring in **TWO** A4 size (210 × 297 mm) help sheets.

⇒ Page 2
1. (10 pts) For each of the following statements, in your answer book, write down the one of the four choices that is most appropriate.

(a). (2 pts) To detect if there are outliers in the predictor variables of a sample, which of the following quantities is most appropriate?

   a) Ordinary residuals   b) Studentized residuals
   c) Leverages           d) Cook’s distances

(b). (2 pts) Which of the following simple linear regression models’ ordinary residuals may not sum up to zero?

   a) The general simple linear regression model   b) The intercept model
   c) The slope model                             d) The trivial model

(c). (2 pts) When the coefficient of determination of a multiple linear regression model is 0, which of the following must be true?

   a) All the regression coefficients are 0
   b) All the regression coefficients are not 0
   c) All the regression coefficients of the predictor variables are 0
   d) All the regression coefficients of the predictor variables are not 0

(d). (2 pts) Which of the following model selection procedures may require more computation effort than the others?

   a) Best subset       b) Backward elimination
   c) Forward selection d) Stepwise

(e). (2 pts) Let Cor(Y, \hat{Y}) be the correlation between the vector of the response values and that of the fitted values based on a simple linear regression fit of Y on X. If Cor(Y, \hat{Y}) > 0, which of the following must be false:

   a) Cor(X, Y) > 0   b) Cor(X, Y) < 0
   c) Cor(X, Y) = 0    d) Cor(X, Y) = Cor(Y, \hat{Y})
2. (20 pts)

(a) (6 pts) Let $Y$ and $X_1$ have a strong nonlinear relationship, $Y$ and $X_2$ have a strong linear relationship, and $X_1$ and $X_2$ be negatively linearly correlated. Let $\hat{e}$ be the studentized residuals when a standard multiple linear regression model $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon$ is fitted to the data. Draw a matrix plot to reflect the relationships between $\hat{e}, X_1,$ and $X_2$. What linear regression assumptions are violated for this data set?

(b) (6 pts) Let $(X_i, Y_i), i = 1, 2, \cdots, 15$ be a sample generated from the following linear regression model: $Y = 1 + 3X + \epsilon, \epsilon \sim N(0, 1)$. A least squares fit for $Y = \beta_0 + \beta_1 X + \epsilon$ led to $\hat{\beta}_0 = 1.1$ and $\hat{\beta}_1 = 2.9$. Roughly draw a plot for the data, together with the true regression line and the fitted line. Locate the prediction of $Y$ at $X = 2$.

(c) (4 pts) Roughly draw an index plot of the leverages of a data set with observations 5 and 10 being high leverage points. The data set has 20 observations, 1 response variable, and 3 predictor variables. Precisely indicate the cut-off line.

(d) (4 pts) Roughly draw an index plot of the studentized residuals to indicate a data set of size 20 with observations 5 and 7 being the outliers if the cut-off value is 3, and observations 5, 7, 10, 15 and 19 being the outliers if the cut-off value is 2.
3. (35 pts) A national insurance organization wanted to study the consumption pattern of cigarettes in all the 50 states of the USA and the District of Columbia. Let

\[ Y = \text{Sales: Number of packs of cigarettes sold in a state on a per capita basis} \]
\[ X_1 = \text{Age: Median age of a person living in a state} \]
\[ X_2 = \text{HS: Percentage of people over 25 years of age in a state who had completed high school} \]
\[ X_3 = \text{Income: Per Capita personal income for a state (income in dollars)} \]
\[ X_4 = \text{Black: Percentage of blacks living in a state} \]
\[ X_5 = \text{Female: Percentage of females living in a state} \]
\[ X_6 = \text{Price: Weighted average price (in cents) of a pack of cigarettes in a state} \]

Before doing a further regression analysis, we first wanted to find out which observations are unusual. Using Minitab, we fitted \( Y \) against all the predictor variables using the model \( Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon \). The table below lists the observations for 9 selected states, together with their leverage values and studentized residuals:

<table>
<thead>
<tr>
<th>State</th>
<th>AK</th>
<th>DC</th>
<th>FL</th>
<th>HI</th>
<th>NB</th>
<th>NV</th>
<th>NH</th>
<th>NC</th>
<th>UT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>121.3</td>
<td>200.4</td>
<td>123.6</td>
<td>82.1</td>
<td>108.1</td>
<td>189.5</td>
<td>265.7</td>
<td>172.4</td>
<td>65.5</td>
</tr>
<tr>
<td>Age</td>
<td>22.9</td>
<td>28.4</td>
<td>32.3</td>
<td>25.0</td>
<td>28.6</td>
<td>27.8</td>
<td>28.0</td>
<td>26.5</td>
<td>23.1</td>
</tr>
<tr>
<td>HS</td>
<td>66.7</td>
<td>55.2</td>
<td>52.6</td>
<td>61.9</td>
<td>59.3</td>
<td>65.2</td>
<td>57.6</td>
<td>38.5</td>
<td>67.3</td>
</tr>
<tr>
<td>Income</td>
<td>4644</td>
<td>5079</td>
<td>3738</td>
<td>4623</td>
<td>3789</td>
<td>4563</td>
<td>3737</td>
<td>3252</td>
<td>3227</td>
</tr>
<tr>
<td>Black</td>
<td>3.0</td>
<td>71.1</td>
<td>15.3</td>
<td>1.0</td>
<td>2.7</td>
<td>5.7</td>
<td>.3</td>
<td>22.2</td>
<td>.6</td>
</tr>
<tr>
<td>Female</td>
<td>45.7</td>
<td>53.5</td>
<td>51.8</td>
<td>48.0</td>
<td>51.2</td>
<td>49.3</td>
<td>51.1</td>
<td>51.0</td>
<td>50.6</td>
</tr>
<tr>
<td>Price</td>
<td>41.8</td>
<td>32.6</td>
<td>43.8</td>
<td>36.7</td>
<td>34.7</td>
<td>44.0</td>
<td>34.1</td>
<td>29.4</td>
<td>36.6</td>
</tr>
<tr>
<td>HI</td>
<td>.580</td>
<td>.719</td>
<td>.298</td>
<td>.216</td>
<td>.071</td>
<td>.227</td>
<td>.066</td>
<td>.189</td>
<td>.311</td>
</tr>
<tr>
<td>TRES</td>
<td>.744</td>
<td>.860</td>
<td>-.071</td>
<td>-.201</td>
<td>-.98</td>
<td>3.02</td>
<td>7.16</td>
<td>1.24</td>
<td>-1.16</td>
</tr>
</tbody>
</table>

\( \Rightarrow \) Page 5
3. (Continued)

(a). (6 pts) Assume we have collected all the high leverage points in the table. Identify all these high leverage points. Using one of the high leverage points, explain briefly why it is so unusual.

(b). (6 pts) Assume we have collected all the outliers in the table. Identify all these outliers. Use 2 as the cutoff value. Using one of the outliers, explain briefly why it is so unusual.

(c). (13 pts) Compute the Welsch and Kuh measures (DFITS) for the observations listed in the table. Identify all the influential observations.

(d). (10 pts) Based on the parts (a), (b) and (c) above, classify the type of the observations listed in the table. Based on these results, what conclusions can you draw about the relationships among the influential observations, the high leverage points and the outliers? What linear regression assumption is violated due to the existence of these unusual observations?
4. (35 pts) Information about domestic immigration in the USA is important to state and local governments. To find out which are the main factors that influence domestic immigration, a data set for 48 contiguous states has been created. The following are the response variable and predictor variables under consideration:

\[
Y = \text{NDIR: Net domestic immigration rate over the period 1990-1994}
\]
\[
X_1 = \text{Unemp: Unemployment rate in the civilian labor force in 1994}
\]
\[
X_2 = \text{Wage: Average hourly earnings of production workers in manufacturing in 1994}
\]
\[
X_3 = \text{Crime: Violent crime rate per 100,000 people in 1993}
\]
\[
X_4 = \text{Poor: Percentage of population who fall below the poverty level in 1994}
\]
\[
X_5 = \text{Taxes: Total state and local taxes per capita in 1993}
\]
\[
X_6 = \text{Educ: Percentage of population with a high school degree or higher in 1990}
\]

To find the best regression equation, a forward selection procedure was used. For each step, the regression coefficients, the associated P-values (except for the intercepts) and the $R^2$ were recorded as follows:

<table>
<thead>
<tr>
<th>Step</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>71.62</td>
<td>-44.49</td>
<td>-92.86</td>
<td>-34.06</td>
<td>-31.51</td>
<td>-37.68</td>
</tr>
<tr>
<td>P-value</td>
<td>.006</td>
<td>.001</td>
<td>.008</td>
<td>.000</td>
<td>.001</td>
<td>.001</td>
</tr>
<tr>
<td>Taxes</td>
<td>-0.0285</td>
<td>-0.0389</td>
<td>-0.0400</td>
<td>-0.0464</td>
<td>-0.0440</td>
<td>-0.0453</td>
</tr>
<tr>
<td>P-value</td>
<td>.010</td>
<td>.001</td>
<td>.000</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
</tr>
<tr>
<td>Educ</td>
<td>1.82</td>
<td>2.42</td>
<td>1.92</td>
<td>2.09</td>
<td>2.18</td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td>.031</td>
<td>.008</td>
<td>.069</td>
<td>.060</td>
<td>.061</td>
<td></td>
</tr>
<tr>
<td>Crime</td>
<td>.024</td>
<td>.031</td>
<td>.031</td>
<td>.031</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td>.092</td>
<td>.058</td>
<td>.060</td>
<td>.066</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>-1.4</td>
<td>-1.3</td>
<td>-1.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td>.357</td>
<td>.417</td>
<td>.390</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wage</td>
<td>-1.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td>.575</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>.1557</td>
<td>.2391</td>
<td>.2871</td>
<td>.3012</td>
<td>.3064</td>
<td>.3081</td>
</tr>
</tbody>
</table>
4. (Continued)
   (a). (4 pts) Specify an $\alpha_{in}$ so that the forward selection procedure will stop at Step 5.
       Indicate how you use $\alpha_{in}$ in a step.

   (b). (6 pts) Specify an $F_{in}$ so that the forward selection procedure will stop at Step 5.
       Indicate how you use $F_{in}$ in a step.

   (c). (6 pts) Specify a $T_{in}$ so that the forward selection procedure will stop at Step 5.
       Indicate how you use $T_{in}$ in a step.

   (d). (10 pts) Find out the best regression model for the data set. Test if this model is
       adequate compared with the model specified in Step 6. Use $\alpha = 5\%$.

   (e). (7 pts) Based on the best model found in (d), test if the effect of the predictor
       variable “Crime” is positive. Find the T-value and the associated P-value. Use
       $\alpha = 5\%$.

   (f). (2 pts) Predict the NDIR for a state whose Unemp=6.4,Wage=11.17, Crime=715,
       Income=31293, Metrop=84.7, Poor=15.9, Taxes=2122, Educ=78.7, BusFail=.51,
       Temp=61.09 based on the best model found in (d).

   ——End of the Paper——