Master Qualify Exam  
Applied Section  

9:30-11:30am, January 24, 2006

The number of points assigned to each question is indicated before the question. The full score for the exam is 100.

1. An agronomist studied the effects of moisture \((X_1, \text{ in inches})\) and temperature \((X_2, \text{ in centigrade})\) on the yield of a new hybrid tomato \((Y)\). The experimental data follow:

\[

tabular{|c|c|c|c|c|c|c|c|c|c|c|c|c|}
\hline
i: & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 \\
\hline
X_{1i}: & 6 & 6 & 6 & 6 & 8 & 8 & 8 & 8 & 8 & 10 & 10 & 10 & 10 \\
X_{2i}: & 20 & 21 & 22 & 23 & 24 & 20 & 21 & 22 & 23 & 24 & 20 & 21 & 22 \\
Y_i: & 49.2 & 48.1 & 48.0 & 49.6 & 47.0 & 51.5 & 51.7 & 50.4 & 51.2 & 48.4 & 51.1 & 51.5 & 50.3 \\
\hline
\hline
\hline
X_{1i}: & 10 & 10 & 12 & 12 & 12 & 12 & 14 & 14 & 14 & 14 & 14 & 14 \\
X_{2i}: & 23 & 24 & 20 & 21 & 22 & 23 & 24 & 20 & 21 & 22 & 23 & 24 \\
Y_i: & 48.9 & 48.7 & 48.6 & 47.0 & 48.0 & 46.4 & 46.2 & 43.2 & 42.6 & 42.1 & 43.9 & 40.5 \\
\hline
\end{tabular}

(a) (5) Consider a linear regression model \(Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \epsilon_i\), where \(\epsilon_i\) are assumed independent and \(\epsilon_i \sim N(0, \sigma^2)\). Find the least square estimate for the coefficients \(\beta_k, k = 0, 1, 2\). It is known that for

\[
X = \begin{bmatrix}
1 & X_{1,1} & X_{1,2} \\
1 & X_{2,1} & X_{2,2} \\
\vdots & \vdots & \vdots \\
1 & X_{25,1} & X_{25,2}
\end{bmatrix}, \quad Y = \begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_{25}
\end{bmatrix}, \quad X'Y = \begin{bmatrix}
1194.1 \\
1789.0 \\
\vdots \\
26244.0
\end{bmatrix}
\]

\[
X'X = \begin{bmatrix}
25 & 250 & 550 \\
250 & 2700 & 5500 \\
550 & 5500 & 12150
\end{bmatrix}, \quad (X'X)^{-1} = \begin{bmatrix}
10.22 & -0.05 & -0.44 \\
-0.05 & 0.0 & 0 \\
-0.44 & 0 & 0.02
\end{bmatrix}
\]

(b) (5) The error sum of squares for the above fitted linear model is 111.224. Find the Bonferroni joint confidence interval for \(\beta_1\) and \(\beta_2\) with confidence coefficient 95%.

(c) (5) The residual plots against \(X_1, X_2\) and \(X_1 \cdot X_2\) are provided in Figure 1. Comment on the appropriateness of the above linear model based on these plots.

(d) (5) Find the 95% confidence interval for a new observed tomato yield \(Y\) at \(X_1 = 7\) and \(X_2 = 22\).
(e) A second order model $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 \cdot X_2 + \epsilon_i$ is fitted to the above data by least squares. The estimated coefficients for $\beta_0 \sim \beta_5$ are $-27.90, 5.216, 5.622, -0.293, -0.139, -0.005$. The error sum of squares is 13.81. Test whether second order terms can be dropped from the regression model at the level of significance $\alpha = 0.05$. The alternatives are:

$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$

$H_a : \text{not all of } \beta_3, \beta_4, \beta_5 \text{ equal zero}$

Figure 1: Residual plots
2. **Regression.** The setup is as follows. We have data \((x_1, y_1), \ldots, (x_n, y_n)\). We wish to build a regression model with response variable \(y\) and predictor variable \(x\) such that \(E[Y|X = x]\) is one line to the left of breakpoint \(a\), and a second line to the right of \(a\). To be specific,

\[
E[Y|X = x] = \begin{cases} 
\alpha_0 + \alpha_1 x & x < a \\
\gamma_0 + \gamma_1 x & x \geq a
\end{cases}
\]

Define the new variable \(z = z(x) = (x - a)I\{x \geq a\}\). For fixed \(a\), the indicator function \(I\{x \geq a\}\) is defined to be 1 if \(x \geq a\) is true and 0 if it is false.

(a) (5) Sketch \(z\) as a function of \(x\). Argue why it is continuous at \(x = a\).

(b) (5) The regression model \(E[Y|X = x] = \beta_0 + \beta_1 x + \beta_2 z\) is linear in \(x\) on both sides of \(a\). Express the above parameters \(\alpha\) and \(\gamma\) in terms of the \(\beta\)’s.

(c) (5) For the model in part (b), is the regression model continuous at \(a\)? Why or why not?

(d) (5) Define the new variable \(w = w(x) = (x - a)^2I\{x \geq a\}\). Sketch \(w\) as a function of \(x\).

(e) (5) Using the variables \(x, z, w\), give a regression model such that the response function is a quadratic function of \(x\) on both sides of \(a\), where the two quadratics are constrained so that \(E[Y|X = x]\) is continuous and the derivative exists at \(x = a\).
3. Suppose we have the following model: \( Y_{ij} = \mu_i + \epsilon_{ij}, j = 1, ..., n \) and \( i = 1, ..., r \) in which \( \mu_i \) are iid \( n(\mu, \sigma^2_\mu) \), \( \epsilon_{ij} \) are iid \( n(0, \sigma^2) \), and the \( \mu \)'s and \( \epsilon \)'s are independent.

(a) (5) Describe an example of an experimental situation for which this model would be appropriate.

(b) (5) What is the \( \text{var}(Y_{ij}) \) ?

(c) (6) Find the values of the following two covariances and a correlation:

\[
\text{cov}(Y_{ij}, Y_{ik}) \text{ for } j \neq k , \\
\text{cov}(Y_{ij}, Y_{kl}) \text{ for } i \neq k , \\
\text{corr}(Y_{ij}, Y_{ik}) \text{ for } j \neq k .
\]

(d) (5) Construct a test of \( H_0 : \mu_1 = ... = \mu_r \).

(e) (4) Explain why \( \text{corr}(Y_{ij}, Y_{ik}) \) for \( j \neq k \) is of interest.
4. The Pennsylvania System of State Assessment (PSSA) science exam is given to 8th and 12th grade students. Research is continuing to explore ways to raise test scores on the PSSA. An experiment has been conducted to assess the effectiveness of three teaching methods: 1 = the standard approach, 2 = a new hands-on approach, and 3 = a multimedia approach. Each method is used in an 8th grade classroom and a 12th grade classroom; each classroom has 25 students.

(a) (8) The linear model \( y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \) was fit where \( \mu \) is the overall mean, \( \alpha_i \) is the effect of the \( i \)-th teaching method and \( \beta_j \) is the effect of the \( j \)-th grade. The following partial output was obtained.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td></td>
<td>15.4935</td>
<td></td>
<td></td>
</tr>
<tr>
<td>grade</td>
<td></td>
<td>0.0019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>method*grade</td>
<td></td>
<td>33.9823</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td>478.5799</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td></td>
<td>528.0576</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Calculate the df, mean squares and F tests in the table above. Based on these, does the impact of the teaching method on PSSA scores depend upon grade level? Support your statement with appropriate evidence.

(b) (6) The researchers would like to know whether, on average, the two new methods (2) and (3) result in higher PSSA scores than the standard teaching method (1) for each of the grade levels. Calculate the contrast between Method 1 and Methods 2 and 3 first for the 8th grade and again for the 12th grade. Draw an appropriate conclusion.

<table>
<thead>
<tr>
<th>method</th>
<th>grade</th>
<th>LS Means Estimate</th>
<th>Standard Error</th>
<th>DF</th>
<th>t Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>6.5662</td>
<td>0.3646</td>
<td>144</td>
<td>18.01</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>7.9128</td>
<td>0.3646</td>
<td>144</td>
<td>21.7</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>8.2496</td>
<td>0.3646</td>
<td>144</td>
<td>22.63</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>7.7038</td>
<td>0.3646</td>
<td>144</td>
<td>21.13</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>8.237</td>
<td>0.3646</td>
<td>144</td>
<td>22.59</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>7.4576</td>
<td>0.3646</td>
<td>144</td>
<td>20.45</td>
</tr>
</tbody>
</table>

(c) (5) Based on the results in (a) and (b) above, the researchers expanded the study so that each method is used at three different randomly selected schools; at each school, the method is used both in an 8th grade and in a 12th grade classroom of 25 students. Write the full model to use in this setting.

(d) (6) Outline an ANOVA table including the source, df, and F test for each term.