I have neither given nor received unauthorized assistance on this exam.

KEY

Name Signed  Date

Name Printed
This exam concerns analysis of some data collected by the Ship Hydromechanics Laboratory of the Maritime and Transport Technology Department, Technical University of Delft (available on the UCI Machine Learning Repository). The object of data collection was prediction of hydodynamic performance of sailing yachts from dimensions and velocity of the yachts.

The original response variable
\[ y = \text{Residuary resistance per unit weight of displacement} \]
was measured for 308 different configurations of the predictors
- \( x_1 \) = Longitudinal position of the center of buoyancy
- \( x_2 \) = Prismatic coefficient
- \( x_3 \) = Length-displacement ratio
- \( x_4 \) = Beam-draught ratio
- \( x_5 \) = Length-beam ratio
- \( x_6 \) = Froude number (the ratio of a characteristic velocity to a gravitational wave velocity)

The analysis here is for the response variable \( y' = \log(y) \).

Consider first Simple Linear Regression of \( y' \) on \( x_6 \). Below is plot of \( y' \) versus \( x_6 \).

(a) What about this plot suggests that a SLR analysis be applied only to cases where \( x_6 \geq 0.175 \)?

The biggest issue is that variance for small \( x_6 \) seems to be larger than for large \( x_6 \). Dropping the two smallest values of \( x_6 \) from consideration looks like it would largely fix that problem.
Below is JMP report based on the cases from the study with \( x_6 \geq 0.175 \).

\textbf{b)} What fraction of raw variability in \( y' \) is explained using \( x_6 \) as a predictor?

\[ R^2 = 0.99 \]
c) What equation for the original response $y$ as a function of the original predictor $x_6$ corresponds to the fitted linear equation for $y' = \log(y)$ in terms of $x_6$?

$$\log y \approx -3.64 + 16.72 x_6$$

$$y \approx \exp(-3.64 + 16.72 x_6)$$

$$= 0.0263 \exp(16.72 x_6)$$

---

d) The $n = 264$ data pairs $(x_6, y')$ used here have $\bar{x}_6 = .3125$ and (standard deviation of the values of $x_6$) $s_{x_6} = .0864652$. Use these and the information on page 3 to produce 2-sided confidence limits for the mean of $y'$ when $x_6 = .3$. Plug in completely, but you need not simplify.

$$\hat{y} \pm t_{0.025, 262 df} \sqrt{\frac{1}{n} + \frac{(x-x)^2}{2(s-x)^2}}$$

i.e.,

$$(3.64 + 16.72(0.3)) \pm 1.96 (14.23) \sqrt{\frac{1}{264} + \frac{(0.0125)^2}{263(0.0864652)^2}}$$

---

e) A $4^{th}$ order polynomial equation for $y'$ in terms of $x_6$ (based on the cases from the study with $x_6 \geq .175$) has $R^2 = .99262$. What are the value of an appropriate $F$ statistic and its degrees of freedom for assessing whether there is a statistically significant increase in explanatory power for $y' = \log(y)$ provided by the higher order polynomial (as opposed to the linear equation) in $x_6$?

$$F = \frac{(SSR(\text{full}) - SSR(\text{reduced})/(k-p))}{MSE(\text{full})}$$

$SSR(\text{full}) = .99262, SStot = .99262(554.83851) = 550.7388$

$SSE(\text{full}) = 554.83851 - 550.7388 = 4.0947$

$$F = \frac{(550.7388 - 549.53349)/(4-1)}{4.0947/(264-4-1)}$$

$$F = 25.46 \quad d.f. = 3, 259$$
Across the range of values of predictors included in the study, the Froude number ($x_6$) is easily the most important predictor of response ($y$ or $y' = \log(y)$). But it makes sense to consider for a fixed value of $x_6$ how the other predictors impact $y'$. The balance of this exam treats only those cases in the original data set with $x_6 = .3$. Below is a JMP report for an initial analysis of these.

**f)** What are the value of an appropriate $F$ statistic and its degrees of freedom for assessing whether together $x_1, x_2, \ldots, x_5$ provide detectable explanatory power regarding $y' = \log(y)$ (when $x_6 = .3$)?

$$F = 14.3232 \quad d.f. = 5, 16$$
g) From the JMP report on page 5, what initially appear to be the most and the least important predictors for describing \( y' \) (at this Froude number)? (Pick 2 of \( x_1, x_2, \ldots, x_5 \))

Looking at the sizes of \( |t| \)

| least important predictor | \( x_1 \) | most important predictor | \( x_2 \) |

h) What is the value of \( \hat{y}' \) corresponding to \( x_1 = -2.3, x_2 = .568, x_3 = 4.78, x_4 = 3.99, \) and \( x_5 = 3.17 \)? (Plug in completely, but you do not need to simplify.)

\[
\hat{y}' = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5
\]

This is

\[
\hat{y}' = -2.48 + .005(-2.3) + 5.74(0.568) + .720(4.78) - .159(3.99) - .694(3.17)
\]

i) As it turns out, the value of \( \hat{y}' \) from part h) has corresponding standard error of .0171. Give 95% two-sided prediction limits for the next value of \( y' \) for the yacht hull configuration described by these values of \( x_1, x_2, \ldots, x_5 \) (at Froude number .3). (Plug in completely, but you do not need to simplify.)

\[
\text{Use } \hat{y} \pm t\sqrt{(SE_y)^2 + s^2}
\]

From above

\[
\hat{y} \pm 2.120 \sqrt{(.0171)^2 + (.071)^2}
\]

\[
\text{Upper 2.5% pt of } t_{16}
\]

j) Below are the correlations for the predictors in this small part of the original data set. What about them suggests that it may be problematic to assess "individual contributions" of the predictors to "explaining" \( y' \) (for \( x_6 = .3 \))?  

<table>
<thead>
<tr>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
</tr>
<tr>
<td>1.0000</td>
</tr>
<tr>
<td>-0.0086</td>
</tr>
<tr>
<td>-0.0027</td>
</tr>
<tr>
<td>0.0029</td>
</tr>
<tr>
<td>-0.0034</td>
</tr>
</tbody>
</table>

The fact that these entries are not all 0 (and some are sizable) indicates the presence of multicollinearity.
**k)** Below is a table taken from a JMP report for "All Possible Models" from a JMP Fit Model regression of $y'$ on $x_1, x_2, \ldots, x_5$. What set of predictors does it suggest that we consider using? Give 3 good reasons for your choice.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number</th>
<th>R Square</th>
<th>RMSE</th>
<th>Cp</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_2$</td>
<td>1</td>
<td>0.6103</td>
<td>0.0925</td>
<td>16.1464</td>
</tr>
<tr>
<td>$x_4$</td>
<td>1</td>
<td>0.4264</td>
<td>0.1123</td>
<td>32.2535</td>
</tr>
<tr>
<td>$x_5$</td>
<td>1</td>
<td>0.0381</td>
<td>0.1454</td>
<td>66.2796</td>
</tr>
<tr>
<td>$x_2, x_4$</td>
<td>2</td>
<td>0.7803</td>
<td>0.0713</td>
<td>3.2514</td>
</tr>
<tr>
<td>$x_2, x_3$</td>
<td>2</td>
<td>0.6604</td>
<td>0.0866</td>
<td>13.7568</td>
</tr>
<tr>
<td>$x_2, x_5$</td>
<td>2</td>
<td>0.6266</td>
<td>0.0929</td>
<td>16.7133</td>
</tr>
<tr>
<td>$x_2, x_3, x_5$</td>
<td>3</td>
<td>0.8037</td>
<td>0.0892</td>
<td>3.1989</td>
</tr>
<tr>
<td>$x_2, x_3, x_4$</td>
<td>3</td>
<td>0.7835</td>
<td>0.0727</td>
<td>4.9690</td>
</tr>
<tr>
<td>$x_1, x_2, x_4$</td>
<td>3</td>
<td>0.7830</td>
<td>0.0728</td>
<td>5.0165</td>
</tr>
<tr>
<td>$x_2, x_3, x_4, x_5$</td>
<td>4</td>
<td>0.8142</td>
<td>0.0683</td>
<td>4.2833</td>
</tr>
<tr>
<td>$x_1, x_2, x_3, x_5$</td>
<td>4</td>
<td>0.8066</td>
<td>0.0707</td>
<td>4.9436</td>
</tr>
<tr>
<td>$x_1, x_2, x_3, x_4, x_5$</td>
<td>5</td>
<td>0.8174</td>
<td>0.0708</td>
<td>6.0000</td>
</tr>
</tbody>
</table>

Predictors in model: $x_2, x_4$

Reasons: $R^2$ close to maximum (.7803 compared to .8174)
RMSE close to minimum (.0713 close to .0692)
$Cp$ not much bigger than 3
and $p$ is small (model is relatively simple)

**l)** Below are two graphics based on standardized residuals from a particular model for $y'$ (using some of $x_1, x_2, \ldots, x_5$). What do they suggest about the appropriateness of that model?

- **Diagnostic Plot**: Normal probability plot of standardized residuals.
  - Normality doesn't look implausible.

- **Studentized Resid logy vs. x1 & 4 more**: Studentized residuals vs. $x_1$.
  - There may be a need for a model offering curvature in $x_1$, $x_3$ and $x_5$.

- **Studentized Resid logy vs. x3**: Studentized residuals vs. $x_3$.
  - There may also be some hint of nonconstant response variance as $x_1, x_3$, and $x_5$ change.
m) The (unspecified) model referred to in part l) has $SSE = .0966$ and 2 predictors. Give 95% two-sided confidence limits for the standard deviation of $y'$ that accompanies any fixed choice of the 2 predictors under that MLR model. Plug in completely, but you need not simplify.

$$MSE = \frac{SSE}{n-k-1} = \frac{.0966}{22-2-1} = .0051$$

$$s = \sqrt{MSE} = .0713$$

$$USe \leq \sqrt{\frac{n-k-1}{\chi^2_{.025}}} \quad \text{and} \quad LSe = \sqrt{\frac{n-k-1}{\chi^2_{.975}}}$$

$$USe = .0713 \sqrt{\frac{12}{32.852}} \quad \text{and} \quad LSe = .0713 \sqrt{\frac{12}{8.967}}$$

n) Recall from the JMP report on page 5 that these data have $SSTot = .4396$ and $n = 22$. What is the value of $R^2$ for the model referred in in parts l) and m)? (Compute this based on this information and what is given in part m.)

$$R^2 = \frac{SSR}{SSTot} = \frac{SSTot - SSE}{SSTot} = \frac{.4396 - .0966}{.4396} = .7803$$

o) Running Fit Model on these data for the model $y' = \beta_0 + \beta_2 x_2 + \beta_4 x_4 + \epsilon$ produced $b_2 = 3.844$ and $SE_{b_2} = .695$. Give the value of a $t$ statistic and its degrees of freedom for testing whether after accounting for $x_4$, the variable $x_2$ provides important increased ability to predict $y'$.

$$t = \frac{b_2 - 0}{SE_{b_2}} \quad \text{which here is} \quad \frac{3.844}{.695} = 5.53$$

$$d.f. = n-k-1 = 22-2-1 = 19$$

value of the $t$ statistic ______ 5.53 degrees of freedom ______ 19