CHAPTER 8

MULTIPLE REGRESSION

8.4 For each of two models give the following points

(a) Centering predictors: (5 points)
(b) Model interpretation: (10 points)
(c) Model assessment: (10 points)
(d) Plots: (10 points)

Also, give (10 points) for model comparison. Thus the total points on this problem are (80 points).

As an example of how the problem might be solved, two possible models (using centered predictors CPH and CDAYS) are:

- **Model 1 (ADDITIVE)** (Plots in Figures 1 and 2)
  - Fitted model:
    \[
    \hat{LEAD} = 15.97 - 5.02 \cdot CPH + 0.56 \cdot CDAYS.
    \]
  - \( R^2 = 0.5409, \quad R_a^2 = 0.5235, \quad MSE = 25.72. \)
  - The intercept, 15.97, is the predicted lead level at mean PH and DAYS.
  - The change in predicted lead per unit change in CPH is \(-5.02\), and the change in predicted lead per unit change in CDAYS is 0.56.
  - There is a slight curvature in the plot of residuals versus fitted values. Plots of studentized residuals look reasonable.

- **Model 2 (INTERACTION)** (Plots in Figures 3 and 4)
  - Fitted model:
    \[
    \hat{LEAD} = 15.97 - 5.02 \cdot CPH + 0.56 \cdot CDAYS - 0.50 \cdot CPH \cdot CDAYS.
    \]
  - \( R^2 = 0.6568, \quad R_a^2 = 0.6370, \quad MSE = 19.59. \)
  - The intercept, 15.97, is the predicted lead level at mean PH and DAYS.
  - The change in predicted lead per unit change in CPH is
    \[
    \frac{\partial \hat{LEAD}}{\partial CPH} = -5.02 - 0.50 \cdot CDAYS,
    \]
    and the change in predicted lead per unit change in CDAYS is
    \[
    \frac{\partial \hat{LEAD}}{\partial CDAYS} = 0.56 - 0.50 \cdot CPH.
    \]
  - Plots of studentized residuals look reasonable.

Model 2 is preferred because of \( R^2, R_a^2 \) and \( MSE \).

8.6 \( R^2 = 0.6568. \)

- (10 points) 65.68% of the variation in LEAD is explained by the model: \( R^2 = SSR/SSTO = 1949.8326/2968.7347 = 0.6568. \)

- (10 points) The model reduces the uncertainty in predicting LEAD by 65.68%: \( R^2 = 1 - SSE/SSTO = 1 - 1018.9021/2968.7347 = 0.6568. \)

8.8 (5 points) The VIFs all equal 1.0, so there is no multicollinearity at all. The previous conclusions are unchanged.

8.12 (5 points) The overall model is significant, yet none of the regressors is. This seems a paradox. The explanation is multicollinearity, as shown by the large VIFs. I would try centering the predictors.
LEAD = CPH CDAYS
Response Distribution: Normal
Link Function: Identity

Model Equation
LEAD = 15.9661 - 5.0180 CPH + 0.5603 CDAYS

Summary of Fit
Mean of Response  15.9661  R-Square  0.5409
Root MSE  5.0712  Adj R-Sq  0.5135

Analysis of Variance
Source | DF | Sum of Squares | Mean Square | F Stat | Prob > F
---|---|---|---|---|---
Model | 1 | 1605.7037 | 802.8519 | 31.2180 | 0.0001
Error | 53 | 1163.0010 | 22.7174 | | |
C Total | 55 | 2768.7047 | | | |

Type III Tests
Source | DF | Sum of Squares | Mean Square | F Stat | Prob > F
---|---|---|---|---|---
CPH | 1 | 647.4925 | 647.4925 | 23.1711 | 0.0001
CDAYS | 1 | 958.2112 | 958.2112 | 37.2590 | 0.0001

Parameter Estimates
<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Std Error</th>
<th>T Stat</th>
<th>Prob &gt;</th>
<th>Tolerance</th>
<th>Var Inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>1</td>
<td>15.9661</td>
<td>1.6777</td>
<td>9.5692</td>
<td>0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPH</td>
<td>1</td>
<td>-5.0180</td>
<td>1.0001</td>
<td>-5.0177</td>
<td>0.0001</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>CDAYS</td>
<td>1</td>
<td>0.5603</td>
<td>0.0918</td>
<td>6.1049</td>
<td>0.0001</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

![Scatter plot with regression line](attachment:image.png)

Figure 1: SAS/INSIGHT output, additive model, lead data.
Figure 2: Studentized residual plots, additive model, lead data.
**LEAD = CPH CDAYS CPH*CDAYS**

Response Distribution: Normal
Link Function: Identity

<table>
<thead>
<tr>
<th>Parameter Information</th>
<th>Parameter</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>INTERCEPT</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>CPH</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>CDAYS</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>CPH*CDAYS</td>
</tr>
</tbody>
</table>

**Model Equation**

LEAD = 15.9661 - 5.0180 CPH + 0.5603 CDAYS - 0.4955 P dag

**Summary of Fit**

Mean of Response 15.9661 R-Square 0.6588
Root MSE 4.4245 Adj R-Sq 0.6370

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Stat</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Error</td>
<td>51</td>
<td>1869.9396</td>
<td>36.7948</td>
<td></td>
<td>0.0001</td>
</tr>
<tr>
<td>C Total</td>
<td>55</td>
<td>2968.7347</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Type III Tests**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Stat</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPH</td>
<td>1</td>
<td>647.4825</td>
<td>647.4825</td>
<td>33.0450</td>
<td>0.0001</td>
</tr>
<tr>
<td>CDAYS</td>
<td>1</td>
<td>650.2112</td>
<td>650.2112</td>
<td>48.9026</td>
<td>0.0001</td>
</tr>
<tr>
<td>CPH*CDAYS</td>
<td>1</td>
<td>344.1289</td>
<td>344.1289</td>
<td>17.3827</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

**Parameter Estimates**

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Std Error</th>
<th>T Stat</th>
<th>Prob &gt;</th>
<th>Tolerance</th>
<th>Var Inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>1</td>
<td>15.9661</td>
<td>0.5935</td>
<td>26.9937</td>
<td>0.0001</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CPH</td>
<td>1</td>
<td>-5.0180</td>
<td>0.8719</td>
<td>-5.7454</td>
<td>0.0001</td>
<td>1</td>
<td>1.0000</td>
</tr>
<tr>
<td>CDAYS</td>
<td>1</td>
<td>0.5603</td>
<td>0.0681</td>
<td>8.2930</td>
<td>0.0001</td>
<td>1</td>
<td>1.0000</td>
</tr>
<tr>
<td>CPH*CDAYS</td>
<td>1</td>
<td>-0.4955</td>
<td>0.1182</td>
<td>-4.1904</td>
<td>0.0001</td>
<td>1</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

![Figure 3: SAS/INSIGHT output, interaction model, lead data.](image-url)
Figure 4: Studentized residual plots, interaction model, lead data.