620-270  Applied Statistics

Problem Set 11

Questions 11.1 to 11.5 refer to the following car price data, mentioned on page 18 of the lecture notes. (Here car model has been called “make”, to avoid confusing car model with statistical model.)

A student wanted to buy a second hand car, and was choosing between three makes — Barina, Corolla and Laser. He collected some data on prices from newspaper advertisements. Here are some of the data, with makes 1, 2 and 3 corresponding to Barina, Corolla and Laser respectively. The price of the car is in dollars, and the age in years.

<table>
<thead>
<tr>
<th>Row</th>
<th>Make</th>
<th>Price</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>14000</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>9995</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>18800</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>16950</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>3</td>
<td>23000</td>
<td>1</td>
</tr>
<tr>
<td>41</td>
<td>3</td>
<td>7200</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We are interested in comparing the prices of the three makes, and assessing whether age should be taken into account (and if so, how). Some models were fitted in Minitab, and the output is given below. The make of car (1, 2 or 3) is in c1, price is in c2, and age is in c3. Use the Minitab output to answer the following questions.

11.1  (a) Does there appear to be a price difference between the makes, if we ignore age? (Examine the relevant means and CIs.) What formal test can we use to check this? Use a suitable part of the Minitab output to perform the test.

   (b) Does there appear to be an age difference between the makes? Use the Minitab output to perform the formal test.

11.2 There are at least five statistical models that can be fitted to the data set that are basically sets of straight lines when price is plotted against age. For the most general of these models, draw a diagram to represent it and write down the mathematical expression for the model.

11.3  (a) Test whether the model allowing different slopes for each make of car is significantly better than the model with parallel lines.

   (b) Test whether the model with parallel lines is significantly better than the model with a single line for all makes. Write the mathematical expression of the model with parallel lines, and state the hypothesis being tested here. What is the meaning of the hypothesis in terms of comparing the prices of cars?

   (c) Test whether the model with parallel lines (non-specified slope) is significantly better than the model with parallel horizontal lines. State the hypothesis being tested.
11.4 Consider the model with parallel lines. (This is conventionally called the analysis of covariance model.)

(a) Write down the fitted line for each of the three car makes. Give an interpretation to the slope.

(b) Calculate the predicted price of a Barina at the mean age, and check your answer against the relevant “least squares” (i.e. adjusted) mean in the output.

(c) Suppose the student has a choice between buying an 8 year old Corolla and a 6 year old Laser. Which would he expect to be cheaper?

11.5 Consider now the model with individual lines for each car make.

(a) Write down the fitted line for Barina cars. Give an interpretation to the slope.

(b) Calculate the predicted price for the first car in the data set (see above). Hence find the residual for this observation under the fitted model.

```
MTB > oneway c3 levels in c1

Analysis of Variance on Age

Source   DF  SS     MS    F     p
Make     2  167.8  83.9   5.33  0.008
Error    56  881.1 15.7
Total    58 1048.8

Individual 95% CIs For Mean
Based on Pooled StDev

Level   N  Mean  StDev   ----------------------------------------
        1   19  5.684  2.888   (---------------------)
        2   20  9.800  4.938   (---------------------)
        3   20  8.250  3.754   (---------------------)

Pooled StDev = 3.967

5.0  7.5  10.0  12.5

MTB > oneway c2 levels in c1

Analysis of Variance on Price

Source   DF  SS      MS     F     p
Make     2 1295627 647813  0.04  0.965
Error    56 1.027E+09 18342650
Total    58 1.028E+09

Individual 95% CIs For Mean
Based on Pooled StDev

Level   N  Mean  StDev   ----------------------------------------
        1   19  9347  2904   (---------------------)
        2   20  9701  4931   (---------------------)
        3   20  9604  4665   (---------------------)

Pooled StDev = 4283

8400  9600  10800
```

MTB > brief 3
MTB > glm c2 = c1;
SUBC> covariate c3;
SUBC> mean c1.
**Analysis of Variance for Price**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Seq SS</th>
<th>Adj SS</th>
<th>Adj MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1</td>
<td>623868480</td>
<td>769819008</td>
<td>769819008</td>
<td>164.51</td>
<td>0.000</td>
</tr>
<tr>
<td>Make</td>
<td>2</td>
<td>147246192</td>
<td>147246192</td>
<td>73623096</td>
<td>15.73</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>55</td>
<td>257369312</td>
<td>257369312</td>
<td>4679442</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>58</td>
<td>1028483968</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Term** | **Coeff** | **Stdev** | **t-value** | **P**
---|---|---|---|---
Constant | 16945.5 | 641.7 | 26.41 | 0.000
Age | -934.74 | 72.88 | -12.83 | 0.000
Make | 1 | -2285.5 | 433.4 | -5.27 | 0.000
| 2 | 1915.7 | 419.9 | 4.56 | 0.000

**Means for Covariates**

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Mean</th>
<th>Stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>7.949</td>
<td>4.252</td>
</tr>
</tbody>
</table>

**Least Squares Means for Price**

<table>
<thead>
<tr>
<th>Make</th>
<th>Mean</th>
<th>Stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7230</td>
<td>523.0</td>
</tr>
<tr>
<td>2</td>
<td>11431</td>
<td>502.2</td>
</tr>
<tr>
<td>3</td>
<td>9885</td>
<td>484.2</td>
</tr>
</tbody>
</table>

MTB > glm c2 = c1|c3;
SUBC> covariate c3;
SUBC> mean c1.

**Analysis of Variance for Price**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Seq SS</th>
<th>Adj SS</th>
<th>Adj MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make</td>
<td>2</td>
<td>1295627</td>
<td>36566444</td>
<td>18278222</td>
<td>3.78</td>
<td>0.029</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>769819008</td>
<td>625928192</td>
<td>625928192</td>
<td>129.41</td>
<td>0.000</td>
</tr>
<tr>
<td>Make*Age</td>
<td>2</td>
<td>1029165</td>
<td>1029165</td>
<td>514582</td>
<td>0.11</td>
<td>0.899</td>
</tr>
<tr>
<td>Error</td>
<td>53</td>
<td>256340144</td>
<td>256340144</td>
<td>4836606</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>58</td>
<td>1028483968</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Term** | **Coeff** | **Stdev** | **t-value** | **P**
---|---|---|---|---
Constant | 16961.8 | 667.6 | 25.41 | 0.000
Age | -3290.4 | 936.8 | -2.66 | 0.010
Age*Make | 1 | 1715.0 | 927.7 | 1.85 | 0.070
| 2 | -934.46 | 82.14 | -11.38 | 0.000 |
| 1  | 32.9 | 132.2 | 0.25 | 0.805
| 2 | 18.5 | 101.1 | 0.18 | 0.855 |

**Means for Covariates**

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Mean</th>
<th>Stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>7.949</td>
<td>4.252</td>
</tr>
</tbody>
</table>
Least Squares Means for Price

<table>
<thead>
<tr>
<th>Make</th>
<th>Mean</th>
<th>Stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7305</td>
<td>648.0</td>
</tr>
<tr>
<td>2</td>
<td>11396</td>
<td>526.9</td>
</tr>
<tr>
<td>3</td>
<td>9900</td>
<td>493.4</td>
</tr>
</tbody>
</table>

11.6 For Assignment 4

A government economist wants to estimate the relationship between a worker’s productivity on a particular job, and (i) the worker’s sex; and (ii) their score on an aptitude test. The results are as follows (sex = 1 for male, 2 for female):

<table>
<thead>
<tr>
<th>worker</th>
<th>prod</th>
<th>score</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>52</td>
<td>70</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>63</td>
<td>75</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>74</td>
<td>92</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>48</td>
<td>63</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>52</td>
<td>75</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>61</td>
<td>73</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>55</td>
<td>80</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>54</td>
<td>69</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>52</td>
<td>73</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>39</td>
<td>65</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>42</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>45</td>
<td>61</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>39</td>
<td>62</td>
<td>1</td>
</tr>
</tbody>
</table>

(a) Test whether a model with parallel lines is appropriate for the data.

(b) Assuming the parallel line model is suitable, test whether the mean productivity for males and females is different, after allowance is made for any difference in aptitude test scores.

(c) Find a 95% confidence interval for the difference in productivity between males and females, having corrected for differences in aptitude.

(d) Find the predicted productivity for
   i. worker 3;
   ii. a male worker with an average aptitude score;
   iii. an “average male worker” (i.e. a male with an average male aptitude score).

(e) Estimate the error standard deviation.

(f) Plot productivity against aptitude score, distinguishing between male and female workers. Is any caution needed in interpreting the results?

(g) What are your main conclusions from these data?