ST430 Exam 2 Solutions

Date: November 9, 2015

Name: ____________________

Guideline:

• You may use one-page (front and back of a standard A4 paper) of notes.
• No laptop or textbook are permitted but you may use a calculator.
• Giving or receiving assistance from other students is not allowed.
• Show work to receive partial credit! Partial credit will be given, but only for work written on the exam.
• The total points are 25.
• Good luck!
1. Consider the regression with response $Y =$ teacher’s performance rating and two predictors: 1) a continuous predictor: $X_1 =$ the teacher’s years of experience; 2) another continuous predictor: $X_2 =$ the teacher’s age. We consider three models:

- Model 1: $Y = \beta_0 + \beta_1 X_1 + \epsilon$
- Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- Model 3: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$

(a) (1 point) Explain how the interpretation of $\beta_1$ differs between Model 1 and Model 2.

Answer: In Model 2, unlike Model 1, $\beta_1$ is the effect of teacher’s years of experience while accounting for teacher’s age.

(b) (1 point) Give the change in the expected response if age ($X_2$) increases by one while years of experience ($X_1$) is held constant for Model 2.

Answer: $[\beta_0 + \beta_1 X_1 + \beta_2 (X_2 + 1)] - [\beta_0 + \beta_1 X_1 + \beta_2 X_2] = \beta_2$.

(c) (1 point) Give the change in the expected response if age ($X_2$) increases by one while years of experience ($X_1$) is held constant for Model 3.

Answer:

$$[\beta_0 + \beta_1 X_1 + \beta_2 (X_2 + 1) + \beta_3 X_1 (X_2 + 1)] - [\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2] = \beta_2 + \beta_3 X_1.$$ 

(d) (1 point) If age ($X_2$) has positive effect on performance rating for teachers in their first year ($X_1 = 0$) but no effect for teacher in their tenth year ($X_1 = 9$), should $\beta_3$ in Model 3 be positive, negative or zero?

Answer: The effect of age ($X_2$) is $\beta_2 + \beta_3 X_1$. A positive effect in the first year means that $\beta_2 + \beta_3 0 = \beta_2 > 0$. No effect in the tenth year means that $\beta_2 + \beta_3 9 = 0$. It follows that $\beta_3 = -\beta_2/9$ must be negative.
2. Consider the sale prices of residential properties (denoted by \( Y \) and in the unit of 1,000 dollars) in one neighborhood as a function of two predictors: land value (denoted by \( X_1 \) and in the unit of 1,000 dollars) and improvement value (denoted by \( X_2 \) and in the unit of 1,000 dollars). Consider the model

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.
\]

Below is the summary and anova table of the fitted model.

Call:
\texttt{lm(formula = Y ~ X1 * X2)}

Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-117.848</td>
<td>-45.208</td>
<td>-1.167</td>
<td>24.546</td>
<td>228.616</td>
</tr>
</tbody>
</table>

Coefficients:

|             | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------|----------|------------|---------|----------|
| (Intercept)  | 22.782754 | 45.196951  | 0.504   | 0.6163   |
| X1           | 0.108923  | 0.833835   | 0.131   | 0.8966   |
| X2           | 1.213412  | 0.191601   | 6.333   | 5.71e-08 *** |
| X1:X2        | 0.004930  | 0.002282   | 2.160   | 0.0354 *  |

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 62.64 on 52 degrees of freedom
Multiple R-squared: 0.9357, Adjusted R-squared: 0.932
F-statistic: 252.4 on 3 and 52 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: Y

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1</td>
<td>1814019</td>
<td>462.3601</td>
<td>&lt; 2e-16 ***</td>
</tr>
<tr>
<td>X2</td>
<td>1</td>
<td>1138831</td>
<td>290.2670</td>
<td>&lt; 2e-16 ***</td>
</tr>
<tr>
<td>X1:X2</td>
<td>1</td>
<td>18307</td>
<td>4.6662</td>
<td>0.03539 *</td>
</tr>
<tr>
<td>Residuals</td>
<td>52</td>
<td>204016</td>
<td>3923</td>
<td></td>
</tr>
</tbody>
</table>

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(a) (1 point) Give the estimated least squares regression equation.

\textbf{Answer:} \( E(Y) = 22.782754 + 0.108923X_1 + 1.213412X_2 + 0.004930X_1X_2. \)
(b) (2 points) Conduct a test if there is an interaction effect between land value and improvement value: specify the null and alternative, give the testing statistic, specify its null distribution, and make a conclusion.

**Answer:** Null $H_0 : \beta_3 = 0$; alternative $H_a : \beta_3 \neq 0$. The testing statistic is the $t$-testing statistic and its null distribution is a $t$-distribution with 52 degrees of freedom. The $t$-statistic has a value of 2.16. Since the p-value is 0.0354 < 0.05, we reject the null hypothesis at the 0.05 significance level.

An alternative testing statistic is the $F$-testing statistic and its null distribution is a $F$-distribution with 1 and 52 degrees of freedom. The $F$-statistic has a value of 4.6662. Since the p-value is 0.03539 < 0.05, we reject the null hypothesis at the 0.05 significance level.

(c) (1 points) Based on the model, what would you predict for the sale price of a residential property with land value 50,000 dollars and improvement value 100,000 dollars?

**Answer:** The unit of variables are 1000 dollars, hence $X_1 = 50$ and $X_2 = 100$. The predicted sale price is

$$22.782754 + 50 \times 0.108923 + 1.213412 \times 100 + 0.004930 \times 50 \times 100 = 174.2201,$$

which gives 174220.1 dollars.

3. Consider a two-factor regression model with the mean performance of a diesel engine as a function of fuel type (three types: $F_1$, $F_2$ and $F_3$) and engine brand (two brands: $B_1$ and $B_2$).

(a) (1 point) Give the main effects model without interactions. Identify and code all dummy variables.

**Answer:** Let $X_1$ be 1 if fuel type $F_2$ and 0 if not; $X_2$ be 1 if fuel type $F_3$ and 0 if not. Let $X_3$ be 1 if engine brand $B_2$ and 0 if engine brand $B_1$. Then the main effects model is

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

(b) (2 points) Interpret all of model (a)’s parameters (the $\beta_j$’s) in the context of the problem.

**Answer:** $\beta_0$ is the mean performance of engines of fuel type $F_1$ and brand $B_1$. $\beta_1$ is the difference in mean performance between engines of fuel types $F_2$ and $F_1$ when the brands are the same. $\beta_2$ is the difference in mean performance between engines of fuel types $F_3$ and $F_1$ when the brands are the same. $\beta_3$ is the difference in mean performance between engines of brands $B_2$ and $B_1$ when the fuel types are the same.

(c) (2 points) Give the two-factor model with interactions. Identify and code all dummy variables and specify the reference group.
Answer: With similar coding of dummy variables in (a), the model is

\[ E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3 + \beta_5 X_2 X_3 \]

and the reference group is engines of fuel type \( F_1 \) and brand \( B_1 \).

(d) \( (2 \text{ points}) \) Fill in the following table of expected responses \( E(Y) \) using your model from (e). Note: the entries in the table should be combinations of \( \beta_j \)'s.

<table>
<thead>
<tr>
<th>( E(Y) )</th>
<th>Engine brand ( B_1 )</th>
<th>Engine brand ( B_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuel type ( F_1 )</td>
<td>( \beta_0 )</td>
<td>( \beta_0 + \beta_3 )</td>
</tr>
<tr>
<td>Fuel type ( F_2 )</td>
<td>( \beta_0 + \beta_1 )</td>
<td>( \beta_0 + \beta_1 + \beta_3 + \beta_4 )</td>
</tr>
<tr>
<td>Fuel type ( F_3 )</td>
<td>( \beta_0 + \beta_2 )</td>
<td>( \beta_0 + \beta_2 + \beta_3 + \beta_5 )</td>
</tr>
</tbody>
</table>

(e) \( (4 \text{ points}) \) The analysis of variance table for model (a) is below:

Analysis of Variance Table

Response: PERFORM

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUEL</td>
<td>2</td>
<td>170.17</td>
<td>85.08</td>
<td>0.4501</td>
</tr>
<tr>
<td>BRAND</td>
<td>1</td>
<td>688.09</td>
<td>688.09</td>
<td>3.6397</td>
</tr>
<tr>
<td>Residuals</td>
<td>8</td>
<td>1512.41</td>
<td>189.05</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

And the analysis of variance table for model (c) is below:

Analysis of Variance Table

Response: PERFORM

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUEL</td>
<td>2</td>
<td>170.17</td>
<td>85.08</td>
<td>7.5443</td>
</tr>
<tr>
<td>BRAND</td>
<td>1</td>
<td>688.09</td>
<td>688.09</td>
<td>61.0130</td>
</tr>
<tr>
<td>FUEL:BRAND</td>
<td>2</td>
<td>1444.74</td>
<td>722.37</td>
<td>64.0526</td>
</tr>
<tr>
<td>Residuals</td>
<td>6</td>
<td>67.67</td>
<td>11.28</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Based on the above outputs and models (a) and (c), conduct a test of interaction effect between fuel type and engine brand: specify the null and alternative, calculate the testing statistic, specify the null distribution of the testing statistic, and make a conclusion. Note: at the 0.05 level, the cutoff value here is 5.14.

Answer: Based on model (c), the null is \( H_0 : \beta_4 = \beta_5 = 0 \) and the alternative is \( H_a : \text{ at least one of } \beta_4 \text{ and } \beta_5 \text{ is non-zero. } \) The testing statistic is the \( F \)-statistic, defined as

\[ F = \frac{(\text{SSE}_r - \text{SSE}_c)/2}{\text{SSE}_c/6} = \frac{(1512.41 - 67.67)/2}{67.67/6} = 64.05. \]
The null distribution is a $F$-distribution with 2 and 6 degrees of freedom. Since $64.05 > 5.14$, we reject the null hypothesis at the 0.05 significance level.

4. (a) (2 points) Suppose $Y$ is the response, $X_1, X_2$ are two continuous predictors, and $X_3$ is a categorical predictor with three levels: $A$, $B$ and $C$. Give the complete second order models using all predictors.

Answer: Let $X_3 = 1$ if the categorical predictor has level $B$ and 0 if not; $X_4 = 1$ if the categorical predictor has level $C$ and 0 if not. Then the complete second order model is

$$E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + \beta_4X_2^2 + \beta_5X_2^2 + \beta_6X_3 + \beta_7X_4$$
$$+ \beta_8X_1X_3 + \beta_9X_1X_4$$
$$+ \beta_{10}X_2X_3 + \beta_{11}X_2X_4$$
$$+ \beta_{12}X_1X_2X_3 + \beta_{13}X_1X_2X_4$$
$$+ \beta_{14}X_1^2X_3 + \beta_{15}X_1^2X_4$$
$$+ \beta_{16}X_2^2X_3 + \beta_{17}X_2^2X_4.$$  

(b) (1 point) We would like to test if there is an interaction effect between continuous predictors and categorical predictor in model (a) above. Specify the null and alternative.

Answer: $H_0 : \beta_8 = \beta_9 = \cdots = \beta_{17} = 0$; $H_a :$ at least one of $\beta_8, \ldots, \beta_{17}$ is not zero.

(c) (1 point) Give the formula for Akaike’s information criterion AIC (Explain any abbreviation or symbol).

Answer: The formula is

$$n \log(\hat{\sigma}^2) + 2(k + 1),$$

where $n$ is the sample size, $k$ is the number of predictors, and $\hat{\sigma}^2$ is the estimated error variance.

(d) (1 point) Explain why the AIC criteria rewards models that fit the data well and rewards models with a small number of predictors.

Answer: Models with small AIC are preferred. Models that fit the data well will have small $\hat{\sigma}^2$ and hence small AIC; models that have a small number of predictors will also have small AIC.

(e) (1 point) Explain why the stepwise variable selection procedure generally should not be used for building second order models.

Answer: Because when we build second order models, if a second order term is included in the model then all lower order terms have to be included. But the stepwise variable selection procedure does not necessarily follow the rule.