Answers to Homework 3

1. a. Download the wine data from my website under Data for Homework 3. The data come from thousands of wine auctions on prices of red Bordeaux wines from the 1952-1980 vintages. The order of the variables for the input statement is:

\[ \text{lnprice \, \text{wintrain \, avetemp \, harvrain \, age}} \]

Estimate the log of price (lnprice) as a function of winter rain (wintrain), average temperature (avetemp), the level of harvest rain (harvrain), and age of the vintage in years (age).


<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>8.66444</td>
<td>2.16611</td>
<td>26.39</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>22</td>
<td>1.80583</td>
<td>0.08208</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>26</td>
<td>10.47026</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE          0.28650  
R-Square          0.8275  
Dependent Mean    -1.45177  
Adj R-Sq           0.7962  
Coef Var          -19.73471

| Variable     | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|--------------|----|--------------------|----------------|---------|------|
| Intercept    | 1  | -12.312265466678   | 1.67721165009918 | -7.34   | <.0001 |
| wintrain     | 1  | 0.00116678229872   | 0.00048203680002 | 2.42    | 0.0242 |
| avetemp      | 1  | 0.616392441265     | 0.09517551015229 | 6.48    | <.0001 |
| harvrain     | 1  | -0.0038605535657   | 0.00080753497572 | -4.78   | <.0001 |
| age          | 1  | 0.0238474128805    | 0.00716671958521 | 3.33    | 0.0031 |
1. (cont.)

b. Interpret the coefficient estimates on wintrain, avetemp, harvrain and age. Express your answers in terms of the correct units for each variable.

\[ b_{\text{wintrain}} = 0.0012 \] implies that if winter rainfall increases by 1 millimeter, the logarithm of the wine price index is expected to increase by 0.0012, *ceteris paribus*.

\[ b_{\text{avetemp}} = 0.6164 \] implies that if average temperature increases by 1 degree Celsius, the log of the wine price index is expected to increase by 0.6164, *ceteris paribus*.

\[ b_{\text{harvrain}} = -0.0039 \] implies that if harvest rainfall increases by 1 millimeter, the log of the wine price index is expected to decrease by 0.0039, *ceteris paribus*.

\[ b_{\text{age}} = 0.0238 \] implies that if age of the vintage increases by 1 year, the log of the wine price index is expected to increase by 0.0238, *ceteris paribus*.

c. The *New York Times* abstract above ‘explains’ how to predict wine price. Does your estimated regression correspond to his explanation of how the equation predicts wine prices? Why or why not?

*Abstract (Document Summary)*

Calculate the winter rain and the harvest rain (in millimeters). Add summer heat in the vineyard (in degrees centigrade). Subtract 12.145. And what do you have? A very, very passionate argument over wine.

The abstract indicates the following predicting equation:

Wine Price = -12.145 + Winter Rain + Harvest Rain + Temperature

<also acceptable: lnprice = wintrain + harvrain + avetemp - 12.145 >

Our estimated regression is:

lnprice = -12.312 + 0.0012 wintrain + 0.6164 avetemp - 0.0039 harvrain + 0.0238 age

No, the coefficients differ.

d. Which of the parameter estimates are significantly different from zero at the 5 percent level of significance? Use the prob-value (p-value) method.

All parameter estimates are significantly different from zero at the 5 percent level of significance.

e. Explain in words the meaning of the estimated $R^2$.

The estimated value of $R^2 = 0.8275$, which indicates that 82.75% of the variation in lnprice is explained by the model.
1. (cont.)

f. Are the parameters of the model jointly significant by an F-test? Use $\alpha = 0.05$ and show your work.

$F = 26.39$ in the regression. The numerator degrees of freedom $= K - 1 = 4$, and the denominator degrees of freedom $= T - K = 27 - 5 = 22$. The corresponding critical value of $F$ is $2.82$. $F > F_c$ indicating that the parameters of the model are jointly significant.

g. Using a prob-F value test, are the parameters jointly significant at 5%? Yes. The prob-F value, 0.0001, is less than 0.05.

h. Does the adjusted $R^2$ indicate any problems? No. The adjusted $R^2 = 0.7962$ which is close to $R^2 = 0.8275$. It appears that irrelevant variables are not a big problem.

i. How many observations are in the data set? How many observations are used to estimate the regression? Why aren’t the number of observations the same? The number of observations in the data set is 39. The number of observations used to estimate the regression is 27. The numbers differ because 12 observations have missing values for some of the data.

2. a. Check for multicollinearity by obtaining correlation coefficients for all regressors.

Pearson Correlation Coefficients

<table>
<thead>
<tr>
<th></th>
<th>wintrain</th>
<th>avetemp</th>
<th>harvrain</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>wintrain</td>
<td>1.00000</td>
<td>-0.00187</td>
<td>-0.27334</td>
<td>-0.16286</td>
</tr>
<tr>
<td></td>
<td>0.9912</td>
<td>0.0968</td>
<td>0.3286</td>
<td></td>
</tr>
<tr>
<td>avetemp</td>
<td>-0.00187</td>
<td>1.00000</td>
<td>-0.11916</td>
<td>-0.14513</td>
</tr>
<tr>
<td></td>
<td>0.9912</td>
<td>0.4824</td>
<td>0.3914</td>
<td></td>
</tr>
<tr>
<td>harvrain</td>
<td>-0.27334</td>
<td>-0.11916</td>
<td>1.00000</td>
<td>0.23904</td>
</tr>
<tr>
<td></td>
<td>0.0968</td>
<td>0.4824</td>
<td>0.1483</td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>-0.16286</td>
<td>-0.14513</td>
<td>0.23904</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.3286</td>
<td>0.3914</td>
<td>0.1483</td>
<td></td>
</tr>
</tbody>
</table>

b. Does it look like multicollinearity might be a problem? Why or why not? No. All of the correlation coefficients (in absolute value) are far less than 0.80.
3. 
   a. Define a variable \(\text{YEAR} = 1980 - \text{age}\)

   Re-run the model above in part (1), obtain residuals (estimated error terms) and save them in an output data set as follows:

   ```
   PROC REG;
   MODEL lnprice = wintrain avetemp harvrain age;
   OUTPUT OUT=NEW R=RES;  \text{NEW} \text{ is the new data set containing the residuals, and } \text{RES} \text{ is the name of the estimated residuals.}
   PROC PLOT;
   PLOT RES*YEAR;  \text{Plot the residuals on the y-axis against year.}
   ```

   Plot of \(\text{res*year}\). Legend: \(A = 1 \text{ obs}, B = 2 \text{ obs, etc.}\)

   NOTE: 12 \text{ obs had missing values.}

3.  
   b. Does it look like there might be autocorrelation?

   No. There does not appear to be correlation between successive data points.
3. c. Re-estimate the model in question (1) but add the variable YEAR as a regressor. Do you have any estimation problems? Why?

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>8.66444</td>
<td>2.16611</td>
<td>26.39</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>22</td>
<td>1.80583</td>
<td>0.08208</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>26</td>
<td>10.47026</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE        0.28650    R-Square     0.8275
Dependent Mean  -1.45177    Adj R-Sq     0.7962
Coeff Var       -19.73471

NOTE: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.
NOTE: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

year = 1980 * Intercept - age

The SAS System 11:49 Tuesday, May 10, 2005 6

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Parameter Estimates

| Variable   | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|------------|----|--------------------|----------------|---------|------|----|
| Intercept  | B  | -12.312265466678   | 1.67721165009918 | -7.34   | <.0001 |
| wintrain   | 1  | 0.00116678229872   | 0.00048203680002 | 2.42    | 0.0242|
| avetemp    | 1  | 0.616392441265     | 0.09517551015229 | 6.48    | <.0001|
| harvrain   | 1  | -0.003860553567    | 0.00080753497572 | -4.78   | <.0001|
| age        | B  | 0.0238474128805    | 0.00716671958521 | 3.33    | 0.0031|
| year       | 0  | 0                  |                |         |      |    |

Do you have any estimation problems? Why?

Yes, the model can’t be estimated. There is perfect multicollinearity between year and age, i.e., year is linearly related to age.
4.  a. Suppose that a new soil additive is invented in 1960 which increases the quality of wine. Old school has it that the additive actually reduces wine quality but young vintners give it a try. Create a dummy variable

\[
\text{if year} > 1959 \text{ then } D = 1; \text{ else } D = 0;
\]

Using the model in question (1), test the hypothesis that wine quality on average (i.e., lnprice) changed in 1960. Who is right – old school, new school, or neither?

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>8.69894</td>
<td>1.73979</td>
<td>20.63</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>21</td>
<td>1.77133</td>
<td>0.08435</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>26</td>
<td>10.47026</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.29043  R-Square 0.8308
Dependent Mean -1.45177  Adj R-Sq 0.7905
Coeff Var -20.00522

Parameter Estimates

| Variable    | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|-------------|----|--------------------|----------------|---------|------|---|
| Intercept   | 1  | -12.570826464583   | 1.74760496095179 | -7.19   | <.0001 |
| wintrain    | 1  | 0.00126874493503   | 0.00051399330751 | 2.47    | 0.0222 |
| avetemp     | 1  | 0.61239876198827   | 0.09668198187058 | 6.33    | <.0001 |
| harvrain    | 1  | -0.0036779055558   | 0.000868698805485 | -4.24   | 0.0004 |
| age         | 1  | 0.03143194355274   | 0.01390720600992 | 2.26    | 0.0346 |
| d           | 1  | 0.14735682789458   | 0.23039961179066 | 0.64    | 0.5294 |

Neither is right. The coefficient on the dummy variable is not significantly different from zero (prob-value = 0.5294).

b. Suppose that winter rain has a greater effect on wine quality with the soil additive than without the additive. Create a dummy interaction variable

\[
D_{-\text{WRAIN}} = D*\text{WINTRAIN}
\]

Add this variable to the model in question 1 and estimate it. (Remember that after a proc statement you have to define a new data set before creating new variables.)
The REG Procedure  
Model: MODEL1  
Dependent Variable: lnprice

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>8.70113</td>
<td>1.74023</td>
<td>20.66</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>21</td>
<td>1.76913</td>
<td>0.08424</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>26</td>
<td>10.47026</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.29025  
R-Square 0.8310  
Dependent Mean -1.45177  
Adj R-Sq 0.7908  
Coeff Var -19.99283

Parameter Estimates

| Variable  | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|-----------|----|--------------------|----------------|---------|------|---|
| Intercept | 1  | -12.652264440173   | 1.77552501339161 | -7.13  | < .0001 |
| wintrain  | 1  | 0.00118494291207   | 0.0004891162438 | 2.42   | 0.0245 |
| avetemp   | 1  | 0.62062249462842   | 0.09663314596662 | 6.42   | < .0001 |
| harvrain  | 1  | -0.0036400196421   | 0.00088370539915 | -4.12  | 0.0005 |
| age       | 1  | 0.03114978152074   | 0.01323374041154 | 2.35   | 0.0284 |
| d_wrain   | 1  | 0.000239990905547  | 0.00036350040435 | 0.66   | 0.5164 |

c. What is the impact of d_wrain on lnprice? (Include the magnitude of the coefficient, not just the sign.)

In the period after 1959, a 1 millimeter increase in winter rain would have increased lnprice by 0.0002.

d. Is the parameter estimate on d_wrain significantly different from zero by a two-sided test?
No. The prob-value is 0.5164.

e. What does the result in (d) imply about the hypothesis that winter rain affects wine quality more when the additive is used?
The hypothesis (which is the alternative) is not accepted.

f. Since this is a hypothetical question, and there was no soil additive innovation in 1960, are the results regarding D and D_WRAIN reasonable?
Yes.
options pageno=1 nocenter ps=40 ls = 80;

NOTE: PROCEDURE REG used:
  real time           3:26.45
  cpu time            0.04 seconds

data wine;
infile 'wine.txt';
input lnprice wintrain avetemp harvrain age;
year=1980-age;
run;

NOTE: The infile 'wine.txt' is:
  File Name=C:\Program Files\SAS Institute\SAS\V8\wine.txt,
  RECFM=V,LRECL=256

NOTE: 39 records were read from the infile 'wine.txt'.
The minimum record length was 50.
The maximum record length was 50.
NOTE: Missing values were generated as a result of performing an operation on missing values.
  Each place is given by: (Number of times) at (Line):(Column).
  1 at 78:10
NOTE: The data set WORK.WINE has 39 observations and 6 variables.
NOTE: DATA statement used:
  real time           0.03 seconds
  cpu time            0.03 seconds

proc reg;
model lnprice = wintrain avetemp harvrain age;
run;

NOTE: 39 observations read.
NOTE: 12 observations have missing values.
NOTE: 27 observations used in computations.

proc corr;
var wintrain avetemp harvrain age;
run;

NOTE: PROCEDURE CORR used:
  real time           0.00 seconds
  cpu time            0.00 seconds

proc reg;
model lnprice = wintrain avetemp harvrain age;
output out = new r = res;
run;

NOTE: 39 observations read.
NOTE: 12 observations have missing values.
NOTE: 27 observations used in computations.

NOTE: The data set WORK.NEW has 39 observations and 7 variables.
NOTE: PROCEDURE REG used:
   real time 0.09 seconds
cpu time 0.09 seconds

proc plot;
plot res*year;
run;

NOTE: There were 39 observations read from the data set WORK.NEW.
NOTE: PROCEDURE PLOT used:
   real time 0.01 seconds
cpu time 0.01 seconds

proc reg;
model lnprice = wintrain avetemp harvrain age year;
run;

NOTE: 39 observations read.
NOTE: 12 observations have missing values.
NOTE: 27 observations used in computations.

NOTE: PROCEDURE REG used:
   real time 0.04 seconds
cpu time 0.04 seconds

data soil;
set wine;
if year>1959 then d = 1; else d = 0;
d_wrain = d*wintrain;

NOTE: Missing values were generated as a result of performing an operation on missing values.
   Each place is given by: (Number of times) at (Line):(Column).
   1 at 102:12
NOTE: There were 39 observations read from the data set WORK.WINE.
NOTE: The data set WORK.SOIL has 39 observations and 8 variables.
NOTE: DATA statement used:
   real time 0.03 seconds
cpu time 0.03 seconds

proc reg;
model lnprice = wintrain avetemp harvrain age d;
run;
NOTE: 39 observations read.
NOTE: 12 observations have missing values.
NOTE: 27 observations used in computations.

NOTE: PROCEDURE REG used:
   real time          0.06 seconds
   cpu time           0.06 seconds

proc reg;
model lnprice = wintrain avetemp harvrain age d_wrain;
run;

NOTE: 39 observations read.
NOTE: 12 observations have missing values.
NOTE: 27 observations used in computations.