Sergiu Buciumas

Problem 1

Use the data in table B11 (the quality of Pinot Noir wine data) to build regression models for

this problem. The response variable is quality.

a) Begin with all predictors except region. Use forward, backward and stepwise regression

to find the “best” model. For each method, indicate what variables are selected to be in

each model and what your criteria for entry and removal were.

**Stepwise Regression**

**Regression Analysis: Quality versus Aroma, Body, Flavor, Oakiness, Clarity**

Stepwise Selection of Terms

Candidate terms: Aroma, Body, Flavor, Oakiness, Clarity

-----Step 1---- -----Step 2---- -----Step 3----

Coef P Coef P Coef P

Constant 4.00 4.99 6.47

Aroma 0.483 0.086 0.530 0.054 0.580 0.034

Body 0.273 0.418

Flavor 1.168 0.001 1.264 0.000 1.200 0.000

Oakiness -0.684 0.017 -0.659 0.019 -0.602 0.029

Clarity 2.34 0.187 1.79 0.269

S 1.16254 1.15680 1.16131

R-sq 72.06% 71.47% 70.38%

R-sq(adj) 67.69% 68.01% 67.76%

R-sq(pred) 58.70% 61.02% 63.79%

Mallows’ Cp 6.00 4.67 3.93

α to enter = 0.05, α to remove = 0.1

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 3 108.935 36.312 26.92 0.000

Aroma 1 6.603 6.603 4.90 0.034

Flavor 1 25.689 25.689 19.05 0.000

Oakiness 1 6.999 6.999 5.19 0.029

Error 34 45.853 1.349

Total 37 154.788

Model Summary

S R-sq R-sq(adj) R-sq(pred)

1.16131 70.38% 67.76% 63.79%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 6.47 1.33 4.85 0.000

Aroma 0.580 0.262 2.21 0.034 2.21

Flavor 1.200 0.275 4.36 0.000 2.19

Oakiness -0.602 0.264 -2.28 0.029 1.04

Regression Equation

Quality = 6.47 +†0.580†Aroma +†1.200†Flavor -†0.602†Oakiness

In stepwise regression, body and clarity were removed in the first and second steps, respectively. The final model included Aroma, Flavor and Oakiness.

**Backward Regression**

**Regression Analysis: Quality versus Aroma, Body, Flavor, Oakiness, Clarity**

Backward Elimination of Terms

Candidate terms: Aroma, Body, Flavor, Oakiness, Clarity

-----Step 1---- -----Step 2---- -----Step 3----

Coef P Coef P Coef P

Constant 4.00 4.99 6.47

Aroma 0.483 0.086 0.530 0.054 0.580 0.034

Body 0.273 0.418

Flavor 1.168 0.001 1.264 0.000 1.200 0.000

Oakiness -0.684 0.017 -0.659 0.019 -0.602 0.029

Clarity 2.34 0.187 1.79 0.269

S 1.16254 1.15680 1.16131

R-sq 72.06% 71.47% 70.38%

R-sq(adj) 67.69% 68.01% 67.76%

R-sq(pred) 58.70% 61.02% 63.79%

Mallows’ Cp 6.00 4.67 3.93

α to remove = 0.1

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 3 108.935 36.312 26.92 0.000

Aroma 1 6.603 6.603 4.90 0.034

Flavor 1 25.689 25.689 19.05 0.000

Oakiness 1 6.999 6.999 5.19 0.029

Error 34 45.853 1.349

Total 37 154.788

Model Summary

S R-sq R-sq(adj) R-sq(pred)

1.16131 70.38% 67.76% 63.79%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 6.47 1.33 4.85 0.000

Aroma 0.580 0.262 2.21 0.034 2.21

Flavor 1.200 0.275 4.36 0.000 2.19

Oakiness -0.602 0.264 -2.28 0.029 1.04

Regression Equation

Quality = 6.47 +†0.580†Aroma +†1.200†Flavor -†0.602†Oakiness

The backward regression removed body and clarity in the first and second steps respectively. This model has Aroma, Flavor and Oakiness, like the Stepwise Regression.

**Forward Regression**

**Regression Analysis: Quality versus Aroma, Body, Flavor, Oakiness, Clarity**

Forward Selection of Terms

Candidate terms: Aroma, Body, Flavor, Oakiness, Clarity

----Step 1---- -----Step 2---- -----Step 3----

Coef P Coef P Coef P

Constant 4.941 6.91 6.47

Flavor 1.572 0.000 1.642 0.000 1.200 0.000

Oakiness -0.541 0.059 -0.602 0.029

Aroma 0.580 0.034

S 1.27119 1.22423 1.16131

R-sq 62.42% 66.11% 70.38%

R-sq(adj) 61.37% 64.17% 67.76%

R-sq(pred) 58.68% 60.58% 63.79%

Mallows’ Cp 9.04 6.81 3.93

α to enter = 0.1

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 3 108.935 36.312 26.92 0.000

Aroma 1 6.603 6.603 4.90 0.034

Flavor 1 25.689 25.689 19.05 0.000

Oakiness 1 6.999 6.999 5.19 0.029

Error 34 45.853 1.349

Total 37 154.788

Model Summary

S R-sq R-sq(adj) R-sq(pred)

1.16131 70.38% 67.76% 63.79%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 6.47 1.33 4.85 0.000

Aroma 0.580 0.262 2.21 0.034 2.21

Flavor 1.200 0.275 4.36 0.000 2.19

Oakiness -0.602 0.264 -2.28 0.029 1.04

Regression Equation

Quality = 6.47 +†0.580†Aroma +†1.200†Flavor -†0.602†Oakiness

The forward regression removed body and clarity in the first and second steps respectively. This model has Aroma, Flavor and Oakiness, same like the Stepwise Regression.

b) Use all possible subsets approach and choose the two models with best values of

Mallows Cp statistic.

- What variables are included in each of the two best models chosen using the Cp

criteria? 4 pts

- Do the models meet Mallows criteria? Hockings? 2 pts.

- Fit both of these models and give the equation for each. Comment on what you see as

far as significance of the variables in each. 6 pts.

- Is there much difference in their prediction ability (PRESS)? 2 pts

**Best Subsets Regression: Quality versus Clarity, Aroma, ...**

Response is Quality

O

C a

l F k

a A l i

r r B a n

i o o v e

R-Sq R-Sq Mallows t m d o s

Vars R-Sq (adj) (pred) Cp S y a y r s

1 62.4 61.4 58.7 9.0 1.2712 X

1 50.0 48.6 45.3 23.2 1.4658 X

2 66.1 64.2 60.6 6.8 1.2242 X X

2 65.9 63.9 60.8 7.1 1.2288 X X

3 70.4 67.8 63.8 3.9 1.1613 X X X

3 68.0 65.2 59.3 6.6 1.2068 X X X

4 71.5 68.0 61.0 4.7 1.1568 X X X X

4 70.5 66.9 61.0 5.8 1.1769 X X X X

5 72.1 67.7 58.7 6.0 1.1625 X X X X X

By using the Mallow's Cp criteria, we would take the two models with lowest Mallow's Cp values. This is the first of the two 3 variable models (cp=3.9 with Aroma, Flavor and Oakiness) and the first of the two 4 variable models (Cp=4.7 with clarity, aroma, flavor and Oakiness.

Mallow's criteria: Cp < p. In model 1: Cp = 3.9 and p = 4 (including the constant) so these criteria is met. In model 2: Cp=4.7 and p=5 (including constant) so again the criteria is met.

Hocking's criteria: Cp < 2p-pfull +1. In model 1: Cp is 3.9 and 2p-pfull +1 = 2\*4 - 6 +1 = 3 so these criteria is not met in model 1. In model 2: Cp is 4.7 and 2p-pfull +1 = 2\*5 -6 + 1 = 5 so these criteria is met in model 2.

Model 1: (Cp=3.9)

**Regression Analysis: Quality versus Aroma, Flavor, Oakiness**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 3 108.935 36.312 26.92 0.000

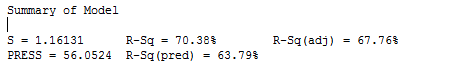
Aroma 1 6.603 6.603 4.90 0.034

Flavor 1 25.689 25.689 19.05 0.000

Oakiness 1 6.999 6.999 5.19 0.029

Error 34 45.853 1.349

Total 37 154.788



Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 6.47 1.33 4.85 0.000

Aroma 0.580 0.262 2.21 0.034 2.21

Flavor 1.200 0.275 4.36 0.000 2.19

Oakiness -0.602 0.264 -2.28 0.029 1.04

Regression Equation

Quality = 6.47 +†0.580†Aroma +†1.200†Flavor -†0.602†Oakiness

The equation is Quality = 6.47 + .580 Aroma + 1.200 Flavor - .602 Oakiness.

All the variables are significant in our model

Model2: (Cp=4.7)

**Regression Analysis: Quality versus Aroma, Flavor, Oakiness, Clarity**

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 4 110.629 27.657 20.67 0.000

Aroma 1 5.355 5.355 4.00 0.054

Flavor 1 27.328 27.328 20.42 0.000

Oakiness 1 8.081 8.081 6.04 0.019

Clarity 1 1.694 1.694 1.27 0.269

Error 33 44.160 1.338

Total 37 154.788

Model Summary



Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 4.99 1.87 2.67 0.012

Aroma 0.530 0.265 2.00 0.054 2.27

Flavor 1.264 0.280 4.52 0.000 2.29

Oakiness -0.659 0.268 -2.46 0.019 1.08

Clarity 1.79 1.59 1.12 0.269 1.08



The regression Equation is: Quality = 4.99 + 0.530 Aroma + 1.264 Flavor - 0.659 Oakiness + 1.79 Clarity.

Variable clarity is not significant in this model. Variable aroma has a p-value slightly higher than .05 now. Clarity being included in the model increased the p-value of aroma, making it insignificant if .05 is used as a strict cutoff.

There is not a big difference in the press values, but a lower press statistic in the first model would be preferable (56.05 vs. 60.33).

c) Incorporate indicator variables for region and find the model selected as best using the

Cp criteria? Fit this model. Compare it with the models found in part b). (Adjusted Rsquared

value, Cp, Press, etc.) Is there an indication that including the region information substantially improves the regression?

Region 1- r2 and r3=0. Region 2- r2=1 and r3=0. Region 3- r2=0 and r3=1.

**Best Subsets Regression: Quality versus Clarity, Aroma, ...**

Response is Quality

O

C a

l F k

a A l i

r r B a n

i o o v e

R-Sq R-Sq Mallows t m d o s r r

Vars R-Sq (adj) PRESS (pred) Cp S y a y r s 1 2

1 62.4 61.4 64.0 58.7 35.4 1.2712 X

1 50.0 48.6 84.7 45.3 58.3 1.4658 X

2 77.6 76.3 40.4 73.9 9.4 0.99552 X X

2 66.1 64.2 61.0 60.6 30.6 1.2242 X X

3 82.4 80.9 34.0 78.0 2.5 0.89464 X X X

3 79.9 78.2 38.3 75.3 7.1 0.95574 X X X

4 83.6 81.6 33.1 78.6 2.2 0.87632 X X X X

4 82.6 80.4 38.7 75.0 4.2 0.90442 X X X X

5 83.7 81.2 34.5 77.7 4.1 0.88788 X X X X X

5 83.7 81.1 35.7 76.9 4.1 0.88820 X X X X X

6 83.8 80.6 37.4 75.8 6.0 0.90054 X X X X X X

6 83.7 80.6 40.1 74.1 6.1 0.90187 X X X X X X

7 83.8 80.0 43.1 72.1 8.0 0.91543 X X X X X X X

According to Cp, the best model is the first of the two 4 variable models (Cp=2.2). This model includes Flavor, Oakiness., r1 and r2

**Regression Analysis: Quality versus Flavor, Oakiness, r1, r2**

Method

Categorical predictor coding (1, 0)

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 4 129.447 32.3617 42.14 0.000

Flavor 1 34.753 34.7532 45.26 0.000

Oakiness 1 1.871 1.8713 2.44 0.128

r1 1 5.715 5.7151 7.44 0.010

r2 1 25.777 25.7767 33.57 0.000

Error 33 25.342 0.7679

Total 37 154.788

Model Summary

S R-sq R-sq(adj) R-sq(pred)

0.876318 83.63% 81.64% 78.63%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 9.21 1.14 8.05 0.000

Flavor 1.192 0.177 6.73 0.000 1.60

Oakiness -0.318 0.204 -1.56 0.128 1.09

r1

1 -1.094 0.401 -2.73 0.010 1.97

r2

1 -2.609 0.450 -5.79 0.000 1.81

Regression Equation

r1 r2

0 0 Quality = 9.21 +†1.192†Flavor -†0.318†Oakiness

0 1 Quality = 6.61 +†1.192†Flavor -†0.318†Oakiness

1 0 Quality = 8.12 +†1.192†Flavor -†0.318†Oakiness

1 1 Quality = 5.51 +†1.192†Flavor -†0.318†Oakiness

Single Regression Equation

Quality = 9.21 +†1.192†Flavor -†0.318†Oakiness +†0.0†r1\_0 -†1.094†r1\_1 +†0.0†r2\_0

-†2.609†r2\_1

The model equation is shown in the screenshot above. Flavor and Oakiness are included in the model just like in the 2 models in part b, but Oakiness is not significant in this one. The adjusted R2 in this model is 81.64% vs. around 68% in the 2 models in part b, resulting that this model is definitely better in terms of R2. The press statistic is almost half what it was in the 2 models in part b (33 in this model vs. 56 and 60 in the 2 models in part b), so this model is better for prediction. The Cp value is also 2.2 (vs. 3.9 and 4.7 in part b), so that indicates that this model is better. The standard deviation is also slightly less in this model (.88 vs. around 1.15 in the other 2 models), so this addition of the region indicator variables definitely improves the regression.

2) The Binary Logistic Regression equation is = 1/(1+e-(-7.05+ 0.988Age + \*0.0738income)).

**Binary Logistic Regression: Buy versus Age, Income**

Method

Link function Logit

Rows used 20

Response Information

Variable Value Count

Buy 1 10 (Event)

0 10

Total 20

Deviance at Each Iterative Step

Step Deviance

1 21.407760

2 21.090016

3 21.081526

4 21.081518

5 21.081518

Deviance Table

Source DF Adj Dev Adj Mean Chi-Square P-Value

Regression 2 6.644 3.322 6.64 0.036

Age 1 5.909 5.909 5.91 0.015

Income 1 1.482 1.482 1.48 0.223

Error 17 21.082 1.240

Total 19 27.726

Model Summary

Deviance Deviance

R-Sq R-Sq(adj) AIC

23.96% 16.75% 27.08

Coefficients

Term Coef SE Coef VIF

Constant -7.05 4.67

Age 0.988 0.527 2.72

Income 0.0738 0.0637 2.72

Odds Ratios for Continuous Predictors

Odds Ratio 95% CI

Age 2.6856 (0.9553, 7.5497)

Income 1.0766 (0.9502, 1.2198)

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -7.05 +†0.988†Age +†0.0738†Income

Goodness-of-Fit Tests

Test DF Chi-Square P-Value

Deviance 17 21.08 0.223

Pearson 17 17.45 0.425

Hosmer-Lemeshow 8 4.06 0.852

Observed and Expected Frequencies for Hosmer-Lemeshow Test

Event

Probability Buy = 1 Buy = 0

Group Range Observed Expected Observed Expected

1 (0.000, 0.148) 0 0.2 2 1.8

2 (0.148, 0.201) 1 0.4 1 1.6

3 (0.201, 0.267) 0 0.5 2 1.5

4 (0.267, 0.419) 1 0.7 1 1.3

5 (0.419, 0.464) 2 1.4 1 1.6

6 (0.464, 0.599) 1 1.2 1 0.8

7 (0.599, 0.651) 1 1.3 1 0.7

8 (0.651, 0.830) 1 1.5 1 0.5

9 (0.830, 0.921) 2 1.8 0 0.2

10 (0.921, 0.993) 1 1.0 0 0.0

Measures of Association

Pairs Number Percent Summary Measures Value

Concordant 77 77.0 Somers’ D 0.55

Discordant 22 22.0 Goodman-Kruskal Gamma 0.56

Ties 1 1.0 Kendall’s Tau-a 0.29

Total 100 100.0

b) The p-value for deviance is 0.223, suggesting that the logistic regression is a good fit. We don't have evidence to suggest that the logistic regression is not a good fit.

c) A $1000 increase in income leads to a 0.0738 (coefficient of income) increase in the log-odds that the family buys the car in the next 6 months. A 1 year increase in the age of their oldest vehicle leads to a 0.988 (coefficient of age) increase in the log-odds that the family buys the car in the next 6 months.

A $1000 increase in income leads to a ~7.7% increase (odds ratio for income is 1.0766) in the likelihood that the family buys the car 6 months later. A 1 year increase in the age of their oldest vehicle leads to a ~169% increase (odds ratio for age is 2.6856) in the likelihood that the family buys the car 6 months later.

d. The odds that a family, who has a salary of $45,000 and has had a car that is 5 years old, buys a car after 6 months is .771.

**Prediction for Buy**

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -7.05 +†0.988†Age +†0.0738†Income

Variable Setting

Age 5

Income 45

Fitted

Probability SE Fit 95% CI

0.771028 0.152372 (0.382846, 0.948129)

3) A logistic regression model is fit to predict whether or not a bank note is counterfeit. For 200

notes, the value of the response y is 1 if the note is a counterfeit and 0 if it is not a counterfeit.

Although there were several possible predictors in the original data set, we focus on only two:

top, which is the width of the note’s top margin in mm; and right which is the note’s right edge

width in mm. The output contains results on two logistic model fits.

(M1): log-odds = β0 + β1\*right + β2\*top

(M2): log-odds = β0 + β1\*right+ β2\*top + β3\*right\*top

a) Use the output to write down the fitted model for the log odds that a bank note is

counterfeit if top and right are the only predictors and the interaction is not included.

1 pt.

**log[ p / (1-p) ] = -553.1 + 3.9153\*Right + 2.2682\*Top**

b) Is there any indication of a lack of fit (inadequate model) for M1? Briefly explain. 2 pts.

Regardless of your answer to part b, assume for the remainder of question 6 that the model

M1 was OK.

Goodness-of-Fit Test for both M1 models are ok.

H0: model has good fit

H1: model has bad fit

We see that p-values > 0.05 so fail to reject that it has good fit.

The HL = 5.3675 with degree of freedom DF = 8, the P value is 0.7177 and the ration HL/DF is very close to unity, does not exceed unity (~0.671), so there is no apparent reason to doubt the adequacy of the fit.

Also we quickly ascertain that this model passes both the Deviance and the Pearson goodness of fit tests. You do not want a small p-value for these tests. The Deviance value is 107.0975. It is supposed to have a chi-square distribution with 150 df. Its p-value is .9968. So we have no reason to doubt the good fit of this model.

c) Interpret in the context of this problem the meaning of β1 in M1. 2 pts.

For every one-unit increase in the value of Right the log-odd of counterfeit increases by 3.9153(assuming no other changes in model).

There is also a confidence interval given for this value. We are 95% confident that the odds ratio ranges from 12.066 to 208.568

d) Which model should be preferred: M1 or M2? Justify your choice as convincingly as you

can. For full points, you will need to use the most straightforward justification. (Note:

the choice can be convincingly justified without writing a lot.) 1 pt.

M1 model would be preferred because in Model M2 every variable is not significant (p-values > 0.05 ()), by adding interaction coefficient variables became non-significant. In model M1 every variable included are significant (p-values < 0.05 ())

e) Using your preferred model, give an estimate of the probability that a bank note is

counterfeit if the right edge width is 130 mm and the top margin is 10 mm. 2 pts.

= 1/ (1+e-(-553.1 + 3.9153\*130 + 2.2682\*10)).

= 4.9e-10

f) Suppose a bank note has estimated odds of 0.5 (under Model 1) of being counterfeit

(based on its values for right and top). If the right edge measurement had gone up by 1

mm but the top measurement is unchanged, what would the new estimated odds

(under model 1) of the note being counterfeit now be? 1 pt

log[ 0.5 ] = -553.1 + 3.9153\*Right + 2.2682\*Top =

Log odds increases by 3.9153 if Right edge measurement had gone up by 1mm. Exponentiation the e-( 3.9153)

=0.01993, which is the odds ratio.

odds ratio = p/ (1-0.5) = 0.5

P = 0.5(1-p)

P= 0.5 – 0.5p

1.5p = 0.5

p = 0.5/1.5 = 0.33

The probability of hitting the target is .33

the probability for missing is 1-.33 = .67

log (0.5) = -0.3

+ 3.9153 = 3.614

log (3.614) = 0.558

new log odds are 0.558 for Right.

g) Suppose that we are using Model 2, and the estimated odds of a particular note being

counterfeit is 0.5 (based on its values for right and top). We can’t tell what the new

estimated odds would be if the right edge measurement had gone up by 1 mm (and the

top measurement was unchanged) without also knowing an additional piece of

information (this is a piece of information we didn’t need for part f.) What is the

information we need, and why do we need it for model 2 but not for model 1?

**log[ 0.5 ] =** -214.3 **+** 1.4635 **\*Right -** 27.9578 **\*Top** + 0.2325 \*right\*top

Because there is interaction we do need to know value for Top, is no longer linear.

Imagine 2xy (let x = 2, y= 2), 2xy = 8, if x increases to 3, 2xy = 12, an increase of 4. Now imagine x = 2, y = 4, 2xy = 16, if x increases to 3, then 2xy = 24, an increase of 8 from previous value. Depending on what y is, an 1 unit increase in x affect the quantity of 2xy.