David Forteguerre
Professor A. Basu
MAR653
10/8/2018

Individual Assignment

1) Linear regression #1: *regression equations and price elasticities*

1.(a) Regression equations for each brand and price elasticities.
Equations:

Brand	Estimated Regression model
Amstel	$\hat{Y} = \beta_0 + \beta_1 + \beta_4 x + \beta_5 x$ $\hat{Y} = (\beta_0 + \beta_1) + (\beta_4 + \beta_5)x$
Bud	$\hat{Y} = (\beta_0 + \beta_2) + (\beta_4 + \beta_6)x$
Michelob	$\hat{Y} = (\beta_0 + \beta_3) + (\beta_4 + \beta_7)x$
Miller	$\hat{Y} = \beta_0 + \beta_4 x$

If we run a linear regression with natural log of move as dependent variable and the natural log of price as the independent variable for a single brand, the coefficient of natural log of price (i.e. the coefficient of X) is the own price elasticity of demand of the brand. Let's express those price elasticities by brand.

Price elasticities of demand:

Brand	Price elasticity
Amstel	$\beta_4 + \beta_5$
Bud	$\beta_4 + \beta_6$
Michelob	$\beta_4 + \beta_7$
Miller	β_4

1.(b) We will now express each of the following hypotheses in terms of the parameters of the regression model.

1.(b)(i) "The price elasticity of demand is the same for all four brands" can be restated as:
$\boldsymbol{\beta_5 = \beta_6 = \beta_7 = 0}$

1.(b)(ii) "The price elasticity of demand is the same for Amstel and Bud" can be restated as:
$\beta_4 + \beta_5 = \beta_4 + \beta_6$
$\boldsymbol{\beta_5 = \beta_6}$

1.(b)(iii) "The regression line is the same (that is, intercept and slope of X are both the same) for Michelob and Miller" can be restated as:
intercept: $\beta_3 = 0$ and slope: $\beta_7 = 0$
$\boldsymbol{\beta_3 = \beta_7 = 0}$

2) Linear regression #2: *hypothesis testing*

Let us first define the following:

- $n = 65$ (sample size)
- $m = 4$ (number of β's other than β_0 in the full model)

(a) H_0: $\beta_3 = 0$ (k = 1)

Full model (X_1, X_2, X_3, X_4): $R^2_{full} = 0.70$
Restricted model (X_1, X_2, X_4): $R^2_{res} = 0.66$

$$F = \frac{0.70-0.66}{1-0.70} \times \frac{65-4-1}{1} = 8$$

Decision rule: at a 99% level of confidence, reject H_0 if $F > F_{.01}(1,60) = 7.08$
(see F Distribution table with $\alpha = 0.1$)
Conclusion: Since $F = 8 > 7.08$, we reject the null hypothesis at a 99% level of confidence and conclude that β_3 is not zero.

(b) H_0: $\beta_1 = \beta_2 = 0$ (k = 2)

Full model (X_1, X_2, X_3, X_4): $R^2_{full} = 0.70$
Restricted model (X_3, X_4): $R^2_{res} = 0.32$

$$F = \frac{0.70-0.32}{1-0.70} \times \frac{65-4-1}{2} = 38$$

Decision rule: at a 99% level of confidence, reject H_0 if $F > F_{.01}(2,60) = 4.98$
(see F Distribution table with $\alpha = 0.1$)
Conclusion: Since $F = 38 > 4.98$, we reject the null hypothesis at a 99% level of confidence and conclude that at least one of β_1 and β_2 is not zero.

(c) H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

Note that in this case:

- The restricted model is the naïve model, and as a result, $R^2_{res} = 0$
- $k = m = 4$

Full model (X_1, X_2, X_3, X_4): $R^2_{full} = 0.70$
Restricted model: $R^2_{res} = 0$

$$F = \frac{0.70-0}{1-0.70} \times \frac{65-4-1}{4} = 35$$

Decision rule: at a 99% level of confidence, reject H_0 if $F > F_{.01}(4,60) = 3.65$
(see F Distribution table with $\alpha = 0.1$)
Conclusion: Since $F = 35 > 3.65$, we reject the null hypothesis at a 99% level of confidence and conclude that at least one of β_1, β_2, β_3 and β_4 is not zero.

3) Logistic regression #1: *regression equations and hypothesis testing*

(a) Let's restate the following null hypothesis, "if an account holder person does not have a graduate degree but owns a home, then the probability of taking a loan does not depend on income", in terms of parameters.

In order to do that, we need to first rewrite I for this specific scenario, considering that "does not have a graduate degree" means that $D_1 = 0$, and that "owns a home" means that $D_2 = 1$.

$$\begin{aligned} I &= \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 X + \beta_4 D_1 X + \beta_5 D_2 X \\ &= \beta_0 + \beta_2 + \beta_3 X + \beta_5 X \\ &= \boldsymbol{(\beta_0 + \beta_2) + (\beta_3 + \beta_5)X} \end{aligned}$$

Now, the fact that the probability of taking a loan does not depend on income means that the coefficient of X must be 0, since X represents the annual income of the person. As a result, the null hypothesis for the scenario above is:

H_0: $\boldsymbol{\beta_3 + \beta_5 = 0}$

(b) We will now test the following null hypotheses at a 99% level of confidence:

Reminders:

χ^2 = (residual deviance from restricted model) – (residual deviance from full model)
$\quad$ = (-2 ln L_{res}) – (-2 ln L_{full})

Decision rule: at a 99% level of confidence, reject H_0 if $\chi^2 > \chi^2_{.01}$ at degrees of freedom k.

(i) H_0: $\boldsymbol{\beta_1 = \beta_4 = 0}$ (k = 2)

Full model (D_1, D_2, X, D_1X, D_2X): -2 ln L_{full} = 480
Restricted model (D_2, X, D_2X): -2 ln L_{res} = 484

χ^2 = (-2 ln L_{res}) – (-2 ln L_{full}) = 484 – 480 = 4

Decision rule: at a 99% level of confidence, reject H_0 if $\chi^2 > \chi^2_{.01}$ (at df = 2) = 9.21
(see Chi-Square (χ^2) Distribution table with α = 0.1)
Conclusion: Since χ^2 = 4 < 9.21, we cannot reject the null hypothesis at a 99% level of confidence.

(ii) H_0: $\boldsymbol{\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0}$ (k = 5)

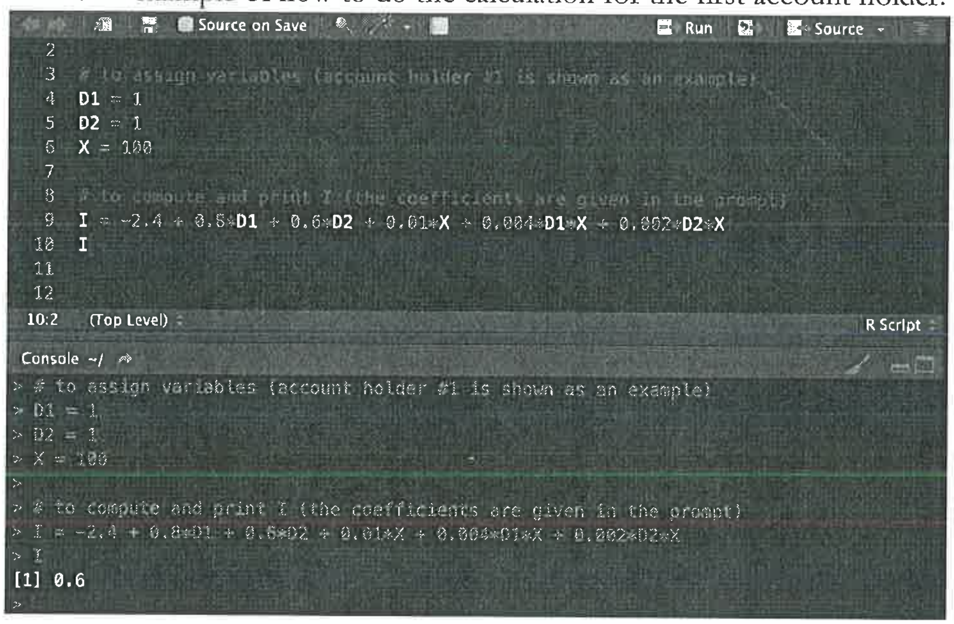
Full model (D_1, D_2, X, D_1X, D_2X): -2 ln L_{full} = 480
Restricted model: -2 ln L_{res} = 570 (note that the restricted model here is the naïve model)

χ^2 = (-2 ln L_{res}) – (-2 ln L_{full}) = 570 – 480 = **90**

Decision rule: at a 99% level of confidence, reject H_0 if $\chi^2 > \chi^2_{.01}$ (at df = 5) = 15.09
(see Chi-Square (χ^2) Distribution table with α = 0.1)
Conclusion: Since χ^2 = 90 > 15.09, we reject the null hypothesis at a 99% level of confidence.

(c) Let's compute the probability that all four of the account holders discussed took a loan in the last five years, using the coefficients provided for I.

First, we need to compute I for each account holder. Let's enter the full equation in R using the three variables D1, D2, and X for simplicity. Then, we will just adapt the variable assignments to quickly compute the result for each account holder.
Here's an example of how to do the calculation for the first account holder:




```
# R CODE:
# to assign variables (account holder #1 is shown as an example)
D1 = 0
D2 = 1
X = 75

# to compute and print I (the coefficients are given in the prompt)
I = -2.4 + 0.8*D1 + 0.6*D2 + 0.01*X + 0.004*D1*X + 0.002*D2*X
I
```

Below are the final results:

Account Holder	D1	D2	X	I
1	1	1	100	**0.6**
2	0	0	80	**-1.6**
3	1	0	120	**0.08**
4	0	1	75	**-0.9**

Second, we need to compute the probability that a loan was taken in the last five years for each account holder, using the following formula:

$$P(Y = 1 \mid I) = \frac{1}{1+e^{-I}}$$

Again, we can do that easily in R. Here's the first example:

```
# to compute and print P (account holder #1 is shown as an example)
P = 1/(1+exp(-(0.6)))
P = round(P, 3) # to round using 3 decimal places
P

11:1    (Top Level)                                   R Script

Console ~/
> # to compute and print P (account holder #1 is shown as an example)
> P = 1/(1+exp(-(0.6)))
> P = round(P, 3) # to round using 3 decimal places
> P
[1] 0.646
>
```

```
# R CODE:
# to compute and print P (account holder #1 is shown as an example)
P = 1/(1+exp(-(0.6)))
P = round(P, 3) # to round using 3 decimal places
P
```

Below are the final results:

Account Holder	D1	D2	X	I	P(Y=1|I)
1	1	1	100	0.6	**0.646**
2	0	0	80	-1.6	**0.168**
3	1	0	120	0.08	**0.52**
4	0	1	75	-0.9	**0.289**

Finally, we just need to multiply those four probabilities to obtain the probability that ***all four*** of the account holders took a loan in the last five years.

P = 0.646 * 0.168 * 0.52 * 0.289 = **0.0163** (rounded at 4 decimal places)

To conclude, the probability that all four account holders took a loan in the last five years is 1.63%.

4) Logistic regression #2: *classifying a new case*

In this exercise, we will determine if, based on the criterion of minimizing expected loss, the banker should approve the loan for the person described in the scenario.

Let's first compute I for the person who has applied for a loan from the banker:

$I = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 D$
$= \mathbf{-2.1 - 0.002}*X_1 \mathbf{- 0.02}*X_2 \mathbf{+ 1.6}*X_3 \mathbf{- 0.6}*D$ *(replacing the coefficients by their known values)*
$= -2.1 - 0.002*\mathbf{120} - 0.02*\mathbf{15} + 1.6*\mathbf{(1/4)}$ *(replacing by the known variables for this person)*
$= \underline{-2.24}$

We can now summarize the costs from incorrect classification in the following table:

Assign to:	ACTUAL Group 0 (Y = 0) *(person **doesn't default** to pay off loan)*	 Group 1 (Y = 1) *(person **defaults** to pay off loan)*
Group 0 *(banker **approves** loan)*	\$0 (loss)	C(0 | 1) = \$100,000 (loss)
Group 1 *(banker does **not approve** loan)*	C(1 | 0) = \$9,000 (loss)	\$0 (loss)

Let's now compute the probabilities that P(Y=1) (i.e. person defaults to pay), that P(Y=0) (i.e. person doesn't default to pay), and the odds ratio.

P(Y=1) $= \frac{1}{1+e^{-I}} = \frac{1}{1+e^{-(-2.24)}} =$ **0.096** (rounded at 3 decimal places)

P(Y=0) = 1 - P(Y=1) $= \frac{1}{1+e^{I}} = \frac{1}{1+e^{-2.24}} =$ **0.904** (rounded at 3 decimal places)

Odds ratio $= \frac{P(Y=1)}{P(Y=0)} = \frac{0.096}{0.904} =$ **0.106** (rounded at 3 decimal places)

However, note that for the case of the logit model, $\frac{P(Y=1)}{P(Y=0)} = e^{I} = \mathbf{0.106}$

Decision rule: assign the new case to Group 1 *(banker does **not approve** loan)* if and only if:

$I \geq \ln(\frac{C(1|0)}{C(0|1)})$
$I = -2.24 \geq \ln(\frac{9000}{100000})$
$I = -2.24 \geq \ln(\frac{9000}{100000})$
$I = -2.24 \geq \ln(0.9)$
$I = -2.24 \geq -2.408$

Conclusion: Since $I > \ln(\frac{C(1|0)}{C(0|1)})$, we assign the new case to Group 1. As a result, the banker should ***not*** approve the loan.

```
# R CODE:
# to compute I for the given scenario
I = -2.1 -0.002*120 -0.02*15 + 1.6*(1/4)
I

# to compute P(1) and print it
P1 = 1/(1+exp(-(-2.24)))
P1 = round(P1, 3) # to round using 3 decimal places
P1

# to compute P(0) and print it
P0 = 1/(1+exp(-2.24))
P0 = round(P0, 3) # to round using 3 decimal places
P0
# or alternatively:
## P0 = 1 - P1

# to compute odds ratio
oddsratio = P1/P0
oddsratio = round(oddsratio, 3) # to round using 3 decimal places
oddsratio

# to check that odds ratio = exp(I) for the logit model
value1 = oddsratio
value2 = exp(I)
value2 = round(value2, 3) # to be consistant with decimal places
value1 == value2 # returns TRUE, which confirms the statement above

# to compute log of costs
finallog = log(9000/100000)
finallog

# to check if I is greater or equal to final log of cost
I>=finallog # returns TRUE
```

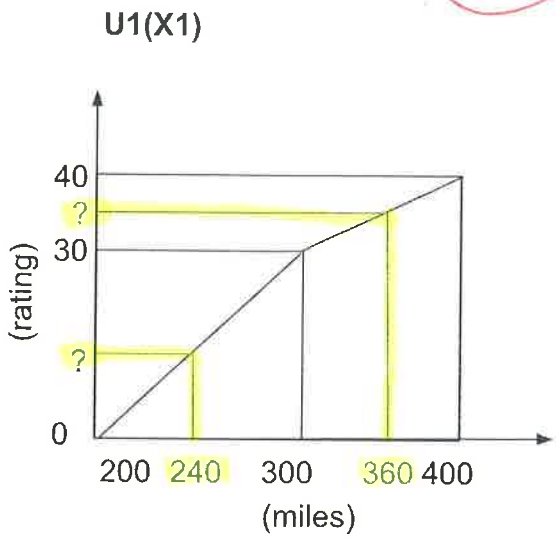
5) Conjoint analysis: *electric cars*

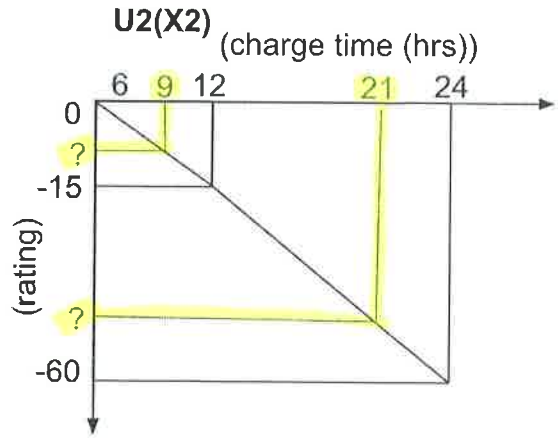
Reminders:

- There are two attributes in this analysis:
    - X_1: The number of highway miles one can drive the car after a full charge. Range: 200-400 miles.
    - X_2: The time it takes to recharge the car after a full discharge. Range: 6-24 hours.
- A person's rating of an electric car on a 0-100 scale (from very poor to excellent) can be expressed as: $\mathbf{U(X_1, X_2) = U_0 + U_1(X_1) + U_2(X_2)}$
    - $U_1(200) = 0$ and $U_2(6) = 0$
    - A higher score means the person likes the product better.
- We will assume that the hypothetical cars differ on only X_1 and X_2, but that they are identical in every other aspect.

Which of the following two electric cars will the person prefer? Let's compute X_1 and X_2 for each car concept. But first, let's visualize our data using graphs:

40/40





a) Car #1 (X_1 = 360, X_2 = 21):

$U_1(360) = U_1(300) + (\frac{360-300}{400-400})(U_1(400)-U_1(300)) = 30 + (\frac{6}{10})(40-30) = 30 + (\frac{6}{10})(10) = \underline{\mathbf{36}}$

$U_2(21) = U_2(12) + (\frac{21-12}{24-12})(U_2(24)-U_2(12)) = -15 + (\frac{9}{12})(-60-(-15)) = -15 + (\frac{9}{12})(-45) = \underline{\mathbf{-48.75}}$

b) Car #2 (X_1 = 240, X_2 = 9):

$U_1(240) = U_1(200) + (\frac{240-200}{300-200})(U_1(300)-U_1(200)) = 0 + (\frac{4}{10})(30-0) = \underline{\mathbf{12}}$

$U_2(9) = U_2(6) + (\frac{9-6}{12-6})(U_2(12)-U_2(6)) = 0 + (\frac{3}{6})(-15-0) = \underline{\mathbf{-7.5}}$

Conclusion:

Car #1 concept score: $U(360, 21) = U_0 + U_1(360) + U_2(21) = 60 + 36 - 48.75 = \underline{\mathbf{47.25}}$ /100
Car #2 concept score: $U(240, 9) = U_0 + U_1(240) + U_2(9) = 60 + 12 - 7.5 = \underline{\mathbf{64.5}}$ /100

Given that 64.5 > 47.25, the respondent prefers the car #2 concept and cares more about charge time than distance range (in miles) overall. Indeed, this person would rather choose a car that charges fast even though its distance range is smaller than a car that has a higher distance range in miles but does not charge fast.