Assignment 2: Linear regression with interaction effect

Andrew Abisha Hu, Elaine Su, Eric Xiong, Tianye Wang

9/5/2018

Descriptions of variables:

```
Listing: ID for each house for sale
ListPrice: the list price of the house ($)
Sold: whether the house is sold (="TRUE") or not (="FALSE")
Age: the age of the house (year)
SquareFeet: the size of the house (square feet)
Beds: the number of bedrooms in the house
Baths: the number of bathrooms in the house
County: the county in which the house is located
Style: the style of the house
Construction: the construction type of the house
Garage: the type of garage with the house
Roof: the type of roof of the house
```

1. Regress ListPrice on Age, SquareFeet, Beds, Baths, and County. You add County in the model because you suspect that location matters. Note that County is a categorical variable. Use "Passaic" as the reference base for County. (provide R codes, 0.5%)

Hint: you can generate a new categorical variable x2 for an old variable x1 by usingrelevel (x1, ref="Y") to assign Y as the reference base for x2.

```
library(car)
```

```
## Loading required package: carData
```

```
setwd("/Users/andrewhu/Documents/GitHub/Machine-Learning/Class Practice/Assignment 2")
prop <- read.csv("Property.csv")
adv <- read.csv("Advertising.csv")
birth <- read.csv("Birth.csv")



relev <- relevel(prop$County, ref="Passaic")
faccounty <- factor(relev)
fit = lm(ListPrice ~Age + SquareFeet + Beds + Baths + faccounty, prop)
```

2. What is the estimation regression equation of the model in question (1)? (0.5%)

```
summary(fit)
```

```
##
## Call:
## lm(formula = ListPrice ~ Age + SquareFeet + Beds + Baths + faccounty,
##     data = prop)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29200  -8702   -994   8309  37241
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     6102.655   3558.547   1.715   0.0874 .
## Age              -33.284     62.497  -0.533   0.5947
## SquareFeet        97.131      1.601  60.686   <2e-16 ***
## Beds              56.819    993.503   0.057   0.9544
## Baths           -379.335    805.464  -0.471   0.6380
## faccountyBergen 1953.140   3100.190   0.630   0.5292
## faccountyEssex  4325.520   2967.326   1.458   0.1460
## faccountyHudson 5562.369   4020.032   1.384   0.1675
## faccountyMercer 5515.456   3438.907   1.604   0.1098
## faccountyMorris 5316.308   3154.615   1.685   0.0930 .
## faccountySussex 7389.402   3953.197   1.869   0.0626 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12730 on 292 degrees of freedom
## Multiple R-squared:  0.976,  Adjusted R-squared:  0.9752
## F-statistic:  1190 on 10 and 292 DF,  p-value: < 2.2e-16
```

3. How do you interpret the coefficient corresponding to a particular county, say Sussex? (0.5%)

```
Overall, the list price is 7389.4 higher in County Sussex than in County Passaic
```

4. What is the residual sum of square (RSS) of this model? (0.5%)

```
47,301,793,877
```

5. You want to know whether adding County can better explain ListPrice. You run another model to regress ListPrice on the same independent variables except County. What is the residual sum of square of this new model? (provide R codes, 0.5%)

```
fit2= lm(ListPrice ~ Age + SquareFeet + Beds + Baths, prop)
```

```
48,187,849,578
```

6. What is the difference in the degrees of freedom between the two models (with vs. withoutCounty)? (0.5%)

```
With County: 292 Without County:298 Difference:-6
```

7. What is the degree of freedom associated with the residuals of the model with County? (0.5%)

```
292
```

8. You are going to use an F test to test whether overall County has a significant effect on ListPrice. What are the null and alternative hypotheses? (0.5%)

```
H0: County doesn't have a significant effect on listprice
Ha: County doesn't have a significant effect on listprice
```

9. Use your answers to questions (4)-(7) to compute the value of the F-statistic associated with the test of the overall effect of County. (0.5%)

```
(48187849578-47301793877)/ 6 = 147,675,950.17
47301793877/292 = 161,992,444.78
147,675,950.17 / 161,992,444.78 = 0.9116
```

10. Use drop1() to find the p-value associated with the F-test. (provide R codes, 0.5%)

```
drop1(fit, test="F")
```

```
## Single term deletions
##
```

```
## Model:
## ListPrice ~ Age + SquareFeet + Beds + Baths + faccounty
##            Df  Sum of Sq        RSS    AIC   F value Pr(>F)
## <none>                   4.7302e+10 5738.4
## Age         1 4.5948e+07 4.7348e+10 5736.7    0.2836 0.5947
## SquareFeet  1 5.9659e+11 6.4389e+11 6527.6 3682.8308 <2e-16 ***
## Beds        1 5.2984e+05 4.7302e+10 5736.4    0.0033 0.9544
## Baths       1 3.5929e+07 4.7338e+10 5736.7    0.2218 0.6380
## faccounty   6 8.8606e+08 4.8188e+10 5732.0    0.9116 0.4868
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value=0.4868

11. What is your conclusion about the overall effect of County based on the p-value at the 0.05 significance level. (0.5%)

p-value>0.05, indicating that County may not be a variable that is significant in the overall model. Cou

Download the Advertising.csv data on Canvas and answer the following questions. Here is the description of variables:

TV: TV advertising budget
radio: radio advertising budget
newspaper: newspaper advertising budget
sales: product sales

12. sales is a nonnegative variable, so it does not follow a normal distribution. Typically, marketing researchers use log(sales) to deal with the issue. Regress log(sales) on TV,radio, and newspaper. (provide codes, 0.5%)

```
logsales <- log(adv$sales)
fit3 = lm(logsales~ TV + radio + newspaper, adv)
```

13. Some researchers argue that (a) the increase in sales will become smaller when using more advertising and (b) advertising should never have a negative marginal effect on sales. The model in question (12) can only satisfy (b) if the corresponding coefficients are positive. To satisfy both (a) and (b), you need to transform the three independent variables. Will you take square roots of each independent variable or adding square terms of each independent variable? Why? Create the regression model with the appropriate transformation. (provide R codes, 0.5%)

```
sqrtTV <- sqrt(adv$TV)
sqrtRadio <- sqrt(adv$radio)
sqrtNP <- sqrt(adv$newspaper)
fit4 = lm(logsales~ sqrtTV + sqrtRadio + sqrtNP, adv)
```

We will take square roots of independent variables because it captures the diminishing effect of advertising on sales. As advertising increases, its marginal effect on sales should be smaller and smaller. advertising will never have a negative marginal effect. However, advertising will never have a negative marginal effect because all the coefficients for advertising are positive.

14. Other researchers argue that while argument (a) in question (13) is true, argument

(b) is questionable. They argue the sales may decrease if consumers are tired of too much advertising (i.e., supersaturation). To test their hypothesis, will you take square roots of each independent variable or adding square terms of each independent variable? Why? Create the regression model with the appropriate transformation. (provide R codes, 0.5%)

```
sqTV = adv$TV^2
sqRadio = adv$radio^2
sqNP= adv$newspaper^2
fit5 = lm(logsales~ TV+ sqTV +radio +sqRadio + newspaper + sqNP, adv)
```

```
We choose to add square terms in addition to the original independent variables because the coefficients
The plot may look similar to this: (y = x - x^2) → Inverted U-Shape
```

15. Based on your findings from the model in question (14), supersaturation may describe the effect of advertising through which media(s)? Why? (0.5%)

```
Supersaturation may describe the effect of advertising through all the media (TV,Radio
and Newspaper), because the coefficients for all of the square terms are negative,
indicating that sales will decrease after a certain point.
```

16. Compute the VIF value for the model in question (14). Given your goal is to test the theory about supersaturation, how should you react to the VIF value? (Provide R codes, 0.5%)

```
vif(fit5)
```

```
##        TV      sqTV     radio   sqRadio newspaper      sqNP
## 16.101705 16.192011 15.328373 15.401475  8.699666  8.725552
```

```
Overall, all the variables have high VIF values, especially for TV, sqTV, radio and
sqRadio. Their VIF (>10) indicates serious multicollinearity problems. However, since
we are testing the theory about supersaturation, and choose to add in their square terms, it
is obvious that square terms will have high correlation with independent variables and
thus we may take VIF simply as a reference and it's not a big concern here.
```

17. Some IMC researchers focus on the synergy of advertising via different communication platforms. Use the original scale of independent variables and create a model that can test the synergy between any two of all media. (provide R codes, 0.5%)

```
fit6 = lm(logsales~TV +radio+ newspaper + TV:radio+ TV:newspaper + radio:newspaper, adv)
```

18. Expand the model in question (17) by further considering the specification to test supersaturation. (provide R codes, 0.5%)

```
fit66 = lm(logsales~ TV+ radio + newspaper + sqTV + sqRadio + sqNP + TV:radio + TV:newspaper + radio:ne
```

19. A multiplicative model implicitly assumes the synergy between all independent variables because the dependent variable can be viewed as an outcome of the multiplication of independent variables. Create a multiplicative model to explain sales by the three advertising variables. (provide R codes, 0.5%)

Hint: use log(x+1) if the minimum of x is 0.

```
logTV <- log(adv$TV)
logRadio <- log(adv$radio+1)
logNP <- log(adv$newspaper)
fit7 = lm(logsales~ logTV + logRadio + logNP)
```

```
Sales = TV*Radio*Newspaper
log(Sales) = log(TV*Radio*Newspaper) = log(TV) + log(Radio) + log(Newspaper)
```

20.Based on the results of the multiplicative model, when TV increases by one percent, what is the percentage change of sales? (0.5%)

```
0.351882 percentage change.
```

21.Which of the following properties are also considered in a multiplicative model (0.5%)?

(a) the increase in sales will become smaller when using more advertising
(b) advertising should never have a negative effect on sales

22. Based on the result of the multiplicative model, what is the mean predicted sales if TV =20, radio = 0, and newspaper = 1. (0.5%)

```
exp(predict(fit7,data.frame(TV = 20,radio = 0, newspaper = 1))+anova(fit7)[4,3]/2)
```

```
## Warning: 'newdata' had 1 row but variables found have 200 rows
```

```
##         1         2         3         4         5         6         7
## 20.248570 11.388242  8.447760 17.753019 14.583533  6.736673 11.934403
##         8         9        10        11        12        13        14
## 13.876232  3.630082 11.733082  9.058370 17.479125  8.975736 10.738066
##        15        16        17        18        19        20        21
## 18.797212 19.954966 13.165664 21.879810 11.587471 15.580691 18.644140
##        22        23        24        25        26        27        28
## 13.890780  6.233053 17.064889 10.176730 13.507539 15.962752 17.303211
##        29        30        31        32        33        34        35
## 19.242708 11.232392 20.707758 13.455571  8.491123 17.637125  8.238213
##        36        37        38        39        40        41        42
## 14.213822 21.295407 14.301339 10.405794 19.990346 17.296256 17.894842
##        43        44        45        46        47        48        49
## 19.880049 14.467211  8.560290 16.461564 11.149205 20.608183 16.946368
##        50        51        52        53        54        55        56
## 10.376232 12.116302 11.227187 20.072053 19.385396 19.764233 20.250221
##        57        58        59        60        61        62        63
##  5.637756 14.503606 20.572225 18.258795  7.112859 21.631950 17.073945
##        64        65        66        67        68        69        70
## 14.171618 16.851222  9.623535  8.876640 13.775353 18.818800 20.201158
##        71        72        73        74        75        76        77
## 18.332952 12.805414  9.112813 11.474059 17.772519  8.339209  5.464770
##        78        79        80        81        82        83        84
## 14.972973  5.043535 11.600015 12.660127 13.514683 11.995382 13.540295
##        85        86        87        88        89        90        91
## 20.061373 16.534146 12.677877 15.858022 13.389207 16.292984 11.168417
##        92        93        94        95        96        97        98
##  5.527107 19.353957 20.741626 12.497151 17.269631 12.045526 16.487281
##        99       100       101       102       103       104       105
## 22.372533 17.039799 13.311138 22.056947 16.607115 15.916818 19.507996
##       106       107       108       109       110       111       112
## 17.577854  7.236460  7.227437  3.722383 19.069577 14.987951 20.359750
##       113       114       115       116       117       118       119
## 14.864402 17.021870 14.330854 13.405822 13.885357  7.238129 16.316263
##       120       121       122       123       124       125       126
##  7.079052 15.865411  7.495613 12.030442 15.646872 19.628709 11.361392
##       127       128       129       130       131       132       133
##  6.166674  6.499110 20.246452 10.028746  2.595113 13.284708  5.682470
##       134       135       136       137       138       139       140
## 19.358088 10.671140 11.907985  9.188140 20.380367 10.270426 18.487116
##       141       142       143       144       145       146       147
## 11.364713 18.832509 19.305876 10.658238 12.332887  9.816620 14.672014
##       148       149       150       151       152       153       154
## 21.622484 10.657815 10.404249 17.753864 12.065287 17.132347 18.297854
##       155       156       157       158       159       160       161
## 16.429736  3.793985 15.131506  9.697153  7.029245 14.339718 15.696332
```

```
##       162       163       164       165       166       167       168
## 14.094993 16.156339 17.395609 12.899671 13.141561  8.123134 13.246138
##       169       170       171       172       173       174       175
## 17.998737 16.604223  9.274702 15.951786  7.399111 12.945233 12.618159
##       176       177       178       179       180       181       182
## 22.608095 19.616373 13.372378 12.944384 13.744364 10.650851 13.647767
##       183       184       185       186       187       188       189
##  8.550624 22.477366 18.549465 19.836714 10.057839 17.686639 17.392889
##       190       191       192       193       194       195       196
##  6.632203 10.754673 10.441662  5.337612 17.830224 16.712842  6.884234
##       197       198       199       200
##  9.842065 13.719351 22.241560 14.932205
```

23. Suppose you do not have any theoretical support to select a model. What you want is a model that can fit the data the best (i.e., explain the variation of the dependent variable the most), considering the number of variables used in the model. In this case, which statistic measure reported by summary() is suitable for model selection? (0.3%)

Adjusted R-square

24. Based on the statistic measure you selected in question (23), which model should you pick? (0.2%)

The multiplicative model →Adjusted R-square of 0.9431, highest among all models

Download the Birth.csv data on Canvas and answer the following questions. Here is the description of variables:

Age: age group of the respondent (<25, 25-29, 30-39, 40-49)
Education: education level of the respondent (Lower or Upper)
Desire: desire to have more children (Yes or No)
Use: Contraceptive use (Yes or No)
Total: number of respondents in the category

25. You want to study what affects the use of contraception. You think Desire must be one key reason, Education could also matter. You are not sure about Age, but you would like to try it. And you think Desire may depend on Age and thus consider the interaction term between the two factors. Run a logistic regression and specify the model as discussed above. (provide R codes, 0.5%)

```
fit8= glm(Use~ Desire*Age + Education, family=binomial,data=birth, weight=Total)
```

26. What is the estimated regression function of the logistic regression in question (25)? (0.5%) Hint: the left-hand side of the function is logit(Use).

logit(Use) = -(1.75336)-DesireYes*(0.06634)+Age25-29*(0.65489)+Age30-39*(1.66644)
+Age40-49*(1.95177)+EducationUpper*(0.36010)-DesireYes:Age25-29*(0.25881)-
DesireYes:Age30-39*(1.15124)-DesireYes:Age40-49*(1.36186)

27. On average, what is the difference in logit(Use) between people in their forties who still want more kids and those who do not want more kids? (0.3%)

-1.36186-0.06634 = -1.4282

28. Is the difference in question (27) statistically different from zero at a 0.01 significance level? (0.2%)

Yes, because p-value for the interaction effect between DesireYes and Age 40-49 is smaller than 0.01.

29. Based on the model, what is the predicted percentage of using contraception for people in their thirties with upper education and no desire to have more children? (0.5%)

0.2731804

30. You want to combine people less than 30 years old in one group, but you are not sure whether people below 25 and people among 25-29 show significantly different behavior. Therefore, you decided to run a likelihood ratio test. You run a new model as what you did for question (25) except that there are only three Age groups this time (<29, 30-39, 40-49). What is the value of -2log-likehood of this new model? (provide R codes, 0.5%)

```
levels(birth$Age) <- c("<29","<29", "30-39", "40-49")
cont2 = glm(Use~Desire + Age + Education + Desire*Age, birth, family=binomial, weight= Total)

-2*logLik(cont2)
```

```
## 'log Lik.' 1862.861 (df=7)
```

31. To perform a likelihood ratio test, you still need (a) the -2log-likehood for the original model in question (25) and (b) the difference in the degree of freedom of the two models. Collect the information you need. What is the p-value of the likelihood ratio test? (provide R codes, 0.3%)

```
-2*logLik(cont2)
```

```
## 'log Lik.' 1862.861 (df=7)
```

```
1-pchisq(1862.861 - 1855.454,2)
```

```
## [1] 0.02463715
```

```
Original model: 1855.454 (df=9)
p-value = 0.02463715
```

32. Give a 0.05 significance level, should you combine the two Age groups <25 and 25-29 as one group? (0.2%)

```
The p value (< .05) indicates that we can reject 0 and conclude that at least one predictor is
different from 0. Therefore, adding two different age groups is highly significant and we
shouldn't combine the two age groups as one group.
```