PS7: Logistic Regression

Pooja Sadarangani

2022-11-27

1 Titanic: What Happened During Her Last Hours? (40pt)

1.1 Titanic Data

1. (2pt) load file titanic.csv.bz2 Download titanic.csv.bz2, and do quick sanity checks

```
## [1] 1309   14
```

```
##  [1] "pclass"    "survived"  "name"      "sex"       "age"       "sibsp"
##  [7] "parch"     "ticket"    "fare"      "cabin"     "embarked"  "boat"
## [13] "body"      "home.dest"
```

```
##   pclass survived                                          name    sex
## 1      1        1                    Allen, Miss. Elisabeth Walton female
## 2      1        1                   Allison, Master. Hudson Trevor   male
## 3      1        0                     Allison, Miss. Helen Loraine female
## 4      1        0             Allison, Mr. Hudson Joshua Creighton   male
## 5      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
## 6      1        1                                 Anderson, Mr. Harry   male
##       age sibsp parch ticket     fare   cabin embarked boat body
## 1 29.0000     0     0  24160 211.3375      B5        S    2   NA
## 2  0.9167     1     2 113781 151.5500 C22 C26        S   11   NA
## 3  2.0000     1     2 113781 151.5500 C22 C26        S        NA
## 4 30.0000     1     2 113781 151.5500 C22 C26        S       135
## 5 25.0000     1     2 113781 151.5500 C22 C26        S        NA
## 6 48.0000     0     0  19952  26.5500     E12        S    3   NA
##                         home.dest
## 1                     St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
## 6                     New York, NY
```

2. (3pt) find the number of missings in the important variables. You are definitely going to use variables survived, pclass, sex, age, and you may use more (see below).

```
## There are missing values in columns Age, Fare, and Body. The total number of missing values in the da
```

3. (4pt) Are there implausible values that are technically not missing?

Ans. There are blank values in certain columns like boat,cabin, fare, home.dest, etc.

1.2 Logistic Regression

1. (4pt) Based on the survivors accounts, which variables do you think are the most important ones to describe survival? How should those be related to the survival? (should they increase or decrease chances of survival?

According to me, variables age, sex, and pclass are the most crucial variables for describing the survival.

pclass: First class passengers had a greater chance of survival as compared to second and third class passengers.

age: Children i.e Age $<$ 14 had a higher chance of survival as compared to adults.

sex: Females had a greater chance of survival as compared to males.

2. (2pt) Create a new variable child, that is 1 if the passenger was younger than 14 years old.

```
##  [1] "pclass"    "survived"  "name"      "sex"       "age"       "sibsp"
##  [7] "parch"     "ticket"    "fare"      "cabin"     "embarked"  "boat"
## [13] "body"      "home.dest" "data"      "child"
```

3. (4pt) Explain why do we have to treat pclass as categorical. Convert it to categorical using factor(pclass).

Ans: We need to consider pclass as categorical because they are discrete values {1,2,3} and not continuous.

```
## [1] "factor"
```

4. (4pt) Estimate a multiple logistic regression model where you explain survival by these variables. Show the results.

```
##
## Call:
## glm(formula = survived ~ pclass + sex + age, family = binomial(),
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6399  -0.6979  -0.4336   0.6688   2.3964
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.522074   0.326702  10.781  < 2e-16 ***
## pclass2     -1.280570   0.225538  -5.678 1.36e-08 ***
## pclass3     -2.289661   0.225802 -10.140  < 2e-16 ***
## sexmale     -2.497845   0.166037 -15.044  < 2e-16 ***
## age         -0.034393   0.006331  -5.433 5.56e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1414.62  on 1045  degrees of freedom
```

```
## Residual deviance:  982.45  on 1041  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 992.45
##
## Number of Fisher Scoring iterations: 4
```

5. (6pt) Interpret the results. Did men or women, old or young have larger chances of survival? What about different passenger classes? How big were the effects?

Ans.

Passenger Class - First class passenger, versus second class passenger, changes the log odds of survival by -1.280570. First class passenger, versus third class passenger, changes the log odds of survival by -2.289661. This suggests that more passengers from first class survived as compared to second and third class passengers. Least number of third class passengers survived.

Sex - Being a female, versus being a males, changes the log odds of survival by -2.497845. This means more number of females had survived.

Age - The effect -0.034393 shows that a negative correlation exists between age and survived variables. Thus, more number of children survived than adults.

6. (5pt) But what about young men? Were they able to force their way to the boats? Create a variable "young man" (e.g. males between 18 and 35, or anything else you see suitable) and see if they survived more likely than others.

```
##   pclass survived                                             name    sex
## 1      1        1                    Allen, Miss. Elisabeth Walton female
## 2      1        1                   Allison, Master. Hudson Trevor   male
## 3      1        0                     Allison, Miss. Helen Loraine female
## 4      1        0             Allison, Mr. Hudson Joshua Creighton   male
## 5      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
## 6      1        1                              Anderson, Mr. Harry   male
##       age sibsp parch ticket     fare   cabin embarked boat body
## 1 29.0000     0     0  24160 211.3375      B5        S    2   NA
## 2  0.9167     1     2 113781 151.5500 C22 C26        S   11   NA
## 3  2.0000     1     2 113781 151.5500 C22 C26        S        NA
## 4 30.0000     1     2 113781 151.5500 C22 C26        S       135
## 5 25.0000     1     2 113781 151.5500 C22 C26        S        NA
## 6 48.0000     0     0  19952  26.5500     E12        S    3   NA
##                         home.dest data.pclass data.survived
## 1                     St Louis, MO           1             1
## 2 Montreal, PQ / Chesterville, ON           1             1
## 3 Montreal, PQ / Chesterville, ON           1             0
## 4 Montreal, PQ / Chesterville, ON           1             0
## 5 Montreal, PQ / Chesterville, ON           1             0
## 6                     New York, NY           1             1
##                                     data.name data.sex data.age data.sibsp
## 1                    Allen, Miss. Elisabeth Walton   female  29.0000          0
## 2                   Allison, Master. Hudson Trevor     male   0.9167          1
## 3                     Allison, Miss. Helen Loraine   female   2.0000          1
## 4             Allison, Mr. Hudson Joshua Creighton     male  30.0000          1
## 5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels)   female  25.0000          1
## 6                              Anderson, Mr. Harry     male  48.0000          0
##   data.parch data.ticket data.fare data.cabin data.embarked data.boat data.body
## 1          0       24160  211.3375         B5             S         2        NA
```

```
## 2          2      113781  151.5500    C22 C26             S        11        NA
## 3          2      113781  151.5500    C22 C26             S                  NA
## 4          2      113781  151.5500    C22 C26             S                 135
## 5          2      113781  151.5500    C22 C26             S                  NA
## 6          0       19952   26.5500        E12             S         3        NA
##                   data.home.dest data.data.pclass data.data.survived
## 1                   St Louis, MO               1                   1
## 2 Montreal, PQ / Chesterville, ON               1                   1
## 3 Montreal, PQ / Chesterville, ON               1                   0
## 4 Montreal, PQ / Chesterville, ON               1                   0
## 5 Montreal, PQ / Chesterville, ON               1                   0
## 6                   New York, NY               1                   1
##                                 data.data.name data.data.sex data.data.age
## 1                   Allen, Miss. Elisabeth Walton        female       29.0000
## 2                  Allison, Master. Hudson Trevor          male        0.9167
## 3                    Allison, Miss. Helen Loraine        female        2.0000
## 4            Allison, Mr. Hudson Joshua Creighton          male       30.0000
## 5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels)        female       25.0000
## 6                           Anderson, Mr. Harry          male       48.0000
##   data.data.sibsp data.data.parch data.data.ticket data.data.fare
## 1               0               0           24160       211.3375
## 2               1               2          113781       151.5500
## 3               1               2          113781       151.5500
## 4               1               2          113781       151.5500
## 5               1               2          113781       151.5500
## 6               0               0           19952        26.5500
##   data.data.cabin data.data.embarked data.data.boat data.data.body
## 1              B5                  S              2             NA
## 2         C22 C26                  S             11             NA
## 3         C22 C26                  S                            NA
## 4         C22 C26                  S                           135
## 5         C22 C26                  S                            NA
## 6             E12                  S              3             NA
##               data.data.home.dest data.child child youngmen
## 1                   St Louis, MO          0     0        1
## 2 Montreal, PQ / Chesterville, ON          1     1        0
## 3 Montreal, PQ / Chesterville, ON          1     1        0
## 4 Montreal, PQ / Chesterville, ON          0     0        1
## 5 Montreal, PQ / Chesterville, ON          0     0        1
## 6                   New York, NY          0     0        0
```

```
##
## Call:
## glm(formula = survived ~ factor(youngmen), family = binomial(),
##     data = titanic_men)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7014  -0.7014  -0.6546  -0.6546   1.8142
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.2770     0.1352  -9.447   <2e-16 ***
## factor(youngmen)1  -0.1545     0.1932  -0.799    0.424
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 667.85  on 657  degrees of freedom
## Residual deviance: 667.21  on 656  degrees of freedom
##   (185 observations deleted due to missingness)
## AIC: 671.21
##
## Number of Fisher Scoring iterations: 4
```

Ans: No, Youngmen weren't able to force their way into the boat. This is clear from the effect of category 1 (category for age between 18 and 35) i.e -0.1545. Less number of young men survived.

7. (7pt) Based on the results above, explain what can you tell about the last hours on Titanic. Are the survivors' accounts broadly accurate? Did the order break down? Can you tell anything else interesting

Based on the results above, survivors' account is broadly accurate. More women survived as compared to men. More children survived as compared to adults. More number of people from first class survived as compared to those from second class and third class.

2 Predict AirBnB Price

1. (2pt) Load the data. Select only relevant variables you need below, otherwise the dataset is hard to comprehend. Do basic sanity checks.

```
##     price bedrooms       room_type accommodates
## 1 $158.00        2 Entire home/apt            5
## 2 $150.00       NA Entire home/apt            4
## 3  $85.00        1 Entire home/apt            2
## 4 $149.00        1 Entire home/apt            2
## 5 $150.00        1 Entire home/apt            4
## 6 $350.00        2 Entire home/apt            4
```

```
## [1] 4448    4
```

2. (4pt) Do the basic data cleaning:

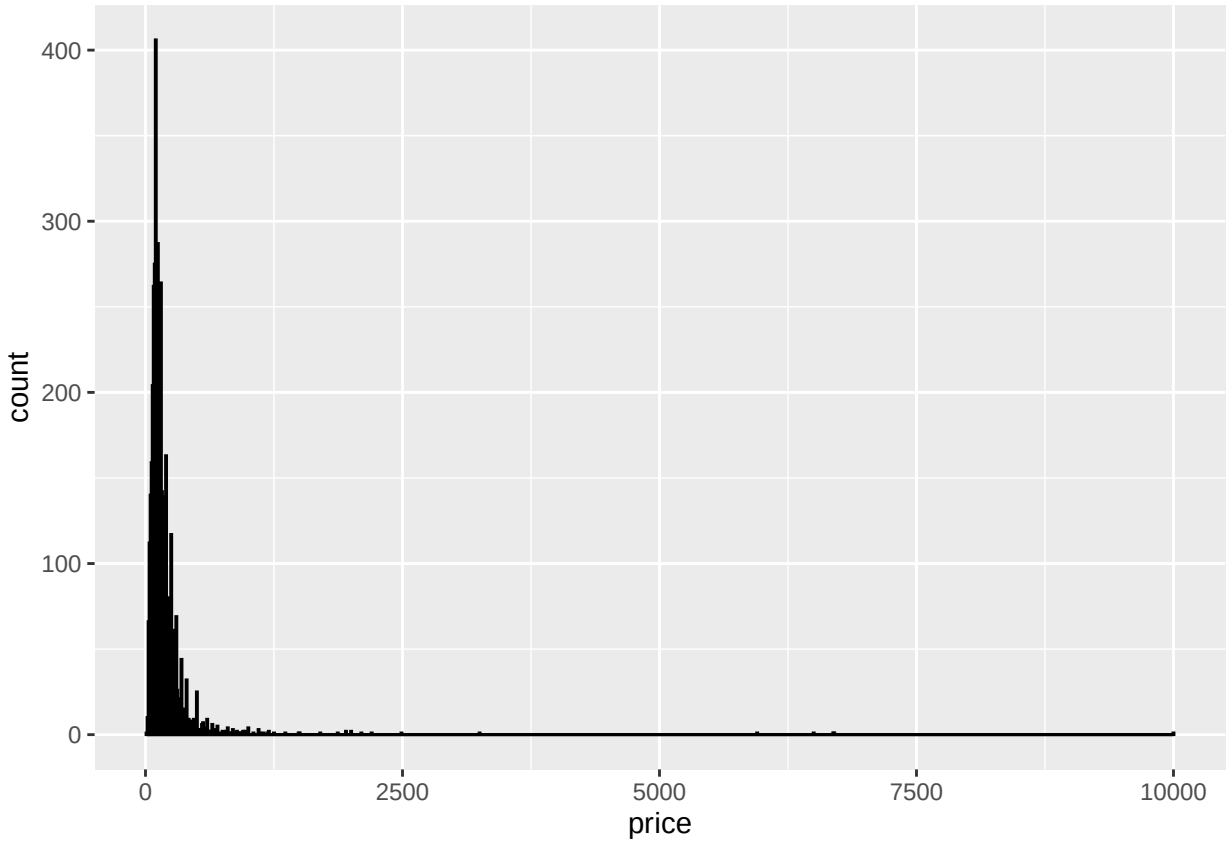
(a) convert price to numeric.

(b) remove entries with missing or invalid price, bedrooms, and other variables you need below.

Hint: there are many NA-s in bedrooms. Check out what are the listings with missing bedrooms, and fill in the values accordingly! Explain what/why are you doing!

Ans. I have replaced NAs in bedroom columns with 0 because from looking at the 'name' column of the listings, I could tell that most of the bedrooms were studios.

3. (4pt) Analyze the distribution of price. Does it look like normal? Does it look like something else? Does it suggest you should do a log-transformation?



Ans. The distribution of price is a normal distribution. The above graph suggests that we should do a log transformation because price has a lower bound which makes the graph right skewed.

4. Convert the number of bedrooms into another variable with a limited number of categories only, such as 0, 1, 2, 3+, and use these categories in the models below.

```
##
##    0    1    2   3+
##  312 2440 1150  546
```

5. (7pt) Now estimate a linear regression model where you explain log price with number of BR-s (the BR categories you did above). Interpret the results. Which model behaves better in the sense of R2?

Linear Regression model for outcome response variable = price

```
##
## Call:
## lm(formula = price ~ factor(bedrooms), data = airbnb)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -365.6  -63.7  -27.5   23.5 9872.5
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          142.75      15.71   9.087  < 2e-16 ***
## factor(bedrooms)1    -16.20      16.68  -0.971  0.33164
## factor(bedrooms)2     51.32      17.71   2.897  0.00378 **
## factor(bedrooms)3+   247.84      19.69  12.585  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 277.5 on 4444 degrees of freedom
## Multiple R-squared:  0.08508,    Adjusted R-squared:  0.08446
## F-statistic: 137.7 on 3 and 4444 DF,  p-value: < 2.2e-16
```

Linear Regression model for outcome response variable = log(price)

```
##
## Call:
## lm(formula = log(price) ~ factor(bedrooms), data = airbnb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4411 -0.3292 -0.0415  0.3315  4.5636
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.72903    0.03085 153.281   <2e-16 ***
## factor(bedrooms)1   -0.08235    0.03277  -2.513    0.012 *
## factor(bedrooms)2    0.42656    0.03479  12.262   <2e-16 ***
```

```
## factor(bedrooms)3+  0.93094    0.03868  24.071   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.545 on 4444 degrees of freedom
## Multiple R-squared:  0.2964, Adjusted R-squared:  0.2959
## F-statistic: 623.9 on 3 and 4444 DF,  p-value: < 2.2e-16
```

Ans: In the sense of R2, linear regression model for outcome response = log(price) works better as it has a greater R2 value. This indicates that the interdependency between the outcome variable (price) and dependent variable (bedrooms) is stronger.

6. (2pt) What kind of values do these two variables take? Show the counts!

For room_type

```
##
## Entire home/apt      Hotel room    Private room     Shared room
##            3582               4             854               8
```

For accommodates

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
##  288 1699  462 1108  247  412   61  101    8   27    5   14    3    7    1    5
```

7. (4pt) Convert the room type into 3 categories: Entire home/apt, Private room, Other; and recode accommodates into 3 categories: "1", "2", "3 or more".

```
## [1] "For room_type:"
```

```
##
## Entire home/apt           Other    Private room
##            3582              12             854
```

```
## [1] "For accommodates:"
```

```
##
##         1         2 3 or more
##       288      1699      2461
```

8. (6pt) Now amend your previous model with these two variables (the 3-category version you did above). Interpret and comment the more interesting/important results. Do not forget to explain what are the relevant reference categories and R2.

```
##
## Call:
## lm(formula = log(price) ~ factor(bedrooms) + factor(room_type) +
##     factor(accommodates), data = airbnb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4534 -0.3134 -0.0555  0.2572  4.8504
```

```
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.441962   0.043801 101.412  < 2e-16 ***
## factor(bedrooms)1              0.005536   0.030714   0.180    0.857
## factor(bedrooms)2              0.274670   0.035704   7.693 1.76e-14 ***
## factor(bedrooms)3+             0.766063   0.039203  19.541  < 2e-16 ***
## factor(room_type)Other        -0.118880   0.145918  -0.815    0.415
## factor(room_type)Private room -0.396423   0.023202 -17.085  < 2e-16 ***
## factor(accommodates)2          0.308814   0.034615   8.921  < 2e-16 ***
## factor(accommodates)3 or more  0.464217   0.038710  11.992  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4999 on 4440 degrees of freedom
## Multiple R-squared:  0.4085, Adjusted R-squared:  0.4076
## F-statistic: 438.1 on 7 and 4440 DF,  p-value: < 2.2e-16
```

The reference categories are: bedrooms = 0, room_type = "Entire home/apt", accomodates = 1. The R2 value i.e 0.4069 is greater than the R2 value of the previous model that only considered the variable bedrooms. This suggests that the updated model is better than the previous model as the interdependency between outcome variable and dependent variables is stronger. We can also see that apart from bedrooms = 1 and room_type = "Other", every dependent variable is statistically significant at 1%.

9. (4pt) You should see that type "Other" is not statistically significant. What does this mean? Why do you think this is the case?

Ans. This means that the price of the house does not depend much on other room_types i.e hotel rooms and shared rooms. We get this result because we don't have much data in the other category to be able to determine it's effect on the price of the house.

10. (3pt) Now use the model above to predict (log) price for each listing in your data.

```
##   price bedrooms       room_type accommodates predictedlogprice
## 1   158        2 Entire home/apt    3 or more          5.180849
## 2   150        0 Entire home/apt    3 or more          4.906179
## 3    85        1 Entire home/apt            2          4.756312
## 4   149        1 Entire home/apt            2          4.756312
## 5   150        1 Entire home/apt    3 or more          4.911715
## 6   350        2 Entire home/apt    3 or more          5.180849
```

11. (5pt) Compute root-mean-squared-error (RMSE) of your predictions. RMSE is explained in lecture notes, 4.1.5 "Model evaluation: MSE, RMSE, R2".

```
## [1] 0.4994184
```

12. (5pt) Now use your model to predict log price for a 2-bedroom apartment that accommodates 4 (i.e., a full 2BR apartment).

```
##        1
## 5.180849
```