IMT 573: Problem Set 8 - Regression II

Srushti Chaukhande

Due: Tuesday, December 3, 2024 by 10:00PM PT

Collaborators:

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Server.

1. Download the `problem_set8.Rmd` file from Canvas or save a copy to your local directory on RStudio Server. Open `problem_set8.Rmd` in RStudio and supply your solutions to the assignment by editing `problem_set8.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do not need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and we encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors, you can do so with the `eval=FALSE` option. (Note: I am also using the `include=FALSE` option here to not include this code in the PDF, but you need to remove this or change it to `TRUE` if you want to include the code chunk.)
7. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the knitted PDF file to `ps5_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

Setup: In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(MASS) # Modern applied statistics functions
library(openintro)
library(fst)
```

Problem 1: Mario Kart (25 pts)

In our regression labs you worked with the Mario Kart dataset. Recall that to load the data you had to use the `openintro` library. You should checkout the regression labs to figure out how to get the `mario_kart` data.

(a) (2pts) Inspect the data using your usual inspect data functions to get a sense of how the variables are encoded and what values they typically take on. Describe the data and variables.

Solution:

Insert Response

```
library(openintro)
data(mariokart)
mario_kart <- mariokart

head(mario_kart)
```

```
## # A tibble: 6 x 12
##           id duration n_bids cond  start_pr ship_pr total_pr ship_sp seller_rate
##        <dbl>    <int>  <int> <fct>    <dbl>   <dbl>    <dbl> <fct>         <int>
## 1    1.50e11        3     20 new       0.99    4        51.6 standa~        1580
## 2    2.60e11        7     13 used      0.99    3.99     37.0 firstC~         365
## 3    3.20e11        3     16 new       0.99    3.5      45.5 firstC~         998
## 4    2.80e11        3     18 new       0.99    0        44   standa~           7
## 5    1.70e11        1     20 new       0.01    0        71   media            820
## 6    3.60e11        3     19 new       0.99    4        45   standa~      270144
## # i 3 more variables: stock_photo <fct>, wheels <int>, title <fct>
```

```
head(mario_kart)
```

```
## # A tibble: 6 x 12
##           id duration n_bids cond  start_pr ship_pr total_pr ship_sp seller_rate
##        <dbl>    <int>  <int> <fct>    <dbl>   <dbl>    <dbl> <fct>         <int>
## 1    1.50e11        3     20 new       0.99    4        51.6 standa~        1580
## 2    2.60e11        7     13 used      0.99    3.99     37.0 firstC~         365
## 3    3.20e11        3     16 new       0.99    3.5      45.5 firstC~         998
## 4    2.80e11        3     18 new       0.99    0        44   standa~           7
## 5    1.70e11        1     20 new       0.01    0        71   media            820
## 6    3.60e11        3     19 new       0.99    4        45   standa~      270144
## # i 3 more variables: stock_photo <fct>, wheels <int>, title <fct>
```

```
summary(mario_kart)
```

```
##        id               duration          n_bids        cond
##  Min.   :1.104e+11   Min.   : 1.000   Min.   : 1.00   new :59
##  1st Qu.:1.404e+11   1st Qu.: 1.000   1st Qu.:10.00   used:84
##  Median :2.205e+11   Median : 3.000   Median :14.00
##  Mean   :2.235e+11   Mean   : 3.769   Mean   :13.54
##  3rd Qu.:2.954e+11   3rd Qu.: 7.000   3rd Qu.:17.00
```

```
##  Max.   :4.001e+11   Max.   :10.000   Max.   :29.00
##
##     start_pr         ship_pr          total_pr            ship_sp
##  Min.   : 0.010   Min.   : 0.000   Min.   : 28.98   standard  :33
##  1st Qu.: 0.990   1st Qu.: 0.000   1st Qu.: 41.17   upsGround :31
##  Median : 1.000   Median : 3.000   Median : 46.50   priority  :23
##  Mean   : 8.777   Mean   : 3.144   Mean   : 49.88   firstClass:22
##  3rd Qu.:10.000   3rd Qu.: 4.000   3rd Qu.: 53.99   parcel    :16
##  Max.   :69.950   Max.   :25.510   Max.   :326.51   media     :14
##                                                     (Other)   : 4
##   seller_rate     stock_photo     wheels
##  Min.   :     0   no : 38     Min.   :0.000
##  1st Qu.:   109   yes:105     1st Qu.:0.000
##  Median :   820               Median :1.000
##  Mean   : 15898               Mean   :1.147
##  3rd Qu.:  4858               3rd Qu.:2.000
##  Max.   :270144               Max.   :4.000
##
##                                                        title
##  BRAND NEW NINTENDO MARIO KART WITH 2 WHEELS              :23
##  Mario Kart Wii (Wii)                                     :19
##  BRAND NEW NINTENDO 1 WII MARIO KART WITH 2 WHEELS +GAME: 8
##  Mario Kart Wii (GAME ONLY/NO WHEEL) - Nintendo Wii Game: 4
##  Mario Kart Wii (Wii) Nintendo Wii game *--WOW --AWESOME: 4
##  (Other)                                                  :84
##  NA's                                                     : 1
```

```
colnames(mario_kart)
```

```
##  [1] "id"          "duration"    "n_bids"      "cond"        "start_pr"
##  [6] "ship_pr"     "total_pr"    "ship_sp"     "seller_rate" "stock_photo"
## [11] "wheels"      "title"
```

```
dim(mario_kart)
```

```
## [1] 143  12
```

Variables Description id: Numeric identifier for each auction. duration: Factor variable indicating the auction duration (1, 3, 5, or 7 days). n_bids: Integer representing the number of bids placed in the auction. cond: Factor variable describing the condition of the item (new or used). start_pr: Numeric variable representing the starting price of the auction in dollars. ship_pr: Numeric variable indicating the shipping price in dollars. total_pr: Numeric variable showing the total price (auction price + shipping) in dollars. ship_sp: Factor variable describing the shipping speed (standard or expedited). seller_rate: Integer representing the seller's rating. stock_photo: Logical variable indicating whether a stock photo was used (TRUE) or not (FALSE). wheels: Integer variable, likely representing the number of wheels included or some other wheel-related information. title: Character variable containing the auction title.

(b) (2 + 2pts) Does the duration of the auction effect the price of a MarioKart? You need to build an a). appropriate model and b). interpret the results to answer the questions.

Solution:

Insert Response

To determine if the duration of the auction affects the price of Mario Kart, we'll build a linear regression model using the auction duration as the predictor variable and the total price as the response variable.

```
# Build the linear regression model
model <- lm(total_pr ~ duration, data = mario_kart)

# View the summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = total_pr ~ duration, data = mario_kart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.035  -8.116  -3.015   3.209 277.953
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.4246     3.8168   13.47   <2e-16 ***
## duration     -0.4097     0.8360   -0.49    0.625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.76 on 141 degrees of freedom
## Multiple R-squared:  0.0017, Adjusted R-squared:  -0.00538
## F-statistic: 0.2402 on 1 and 141 DF,  p-value: 0.6249
```

Coefficient for duration: The estimate for the duration coefficient is -0.4097. This suggests that for each additional day of auction duration, the total price decreases by about $0.41. However, this effect is very small and not statistically significant. Statistical significance: The p-value for the duration coefficient is 0.625, which is much higher than the conventional significance level of 0.05. This means we cannot reject the null hypothesis that the true effect of duration on price is zero. Model fit: The R-squared value is extremely low at 0.0017, indicating that only 0.17% of the variance in total price is explained by auction duration. The adjusted R-squared is even negative, suggesting that the model performs worse than a horizontal line in predicting the total price. Overall model significance: The F-statistic has a p-value of 0.6249, which is much higher than 0.05. This indicates that the model as a whole is not statistically significant. In conclusion, based on this linear regression model, we do not have evidence to support the claim that the duration of the auction affects the price of Mario Kart. The relationship between auction duration and total price is not statistically significant

(c) Experiment with other variables you see fit for this task, that is to see how they effect the price of MarioKart. Do other variables change your results in a major way? Did you have to remove any variables before fitting the model? Make sure that you build an 1). appropriate model while explaining your choice and 2). interpret the results to answer the questions. (pts: 2 model choice + 2 build + 3 interpret + 2 why/whynot remove)

Solution:

Insert Response

I'll use multiple linear regression because we have a continuous outcome variable (total_pr) and multiple potential predictors. This model can help us understand the relationship between several independent

variables and the price. Variables to include: duration (auction length) n_bids (number of bids) cond (condition: new or used) ship_pr (shipping price) seller_rate (seller's rating) stock_photo (whether a stock photo was used) Rationale for variable selection: These variables could all potentially influence the final price. I'm excluding 'start_pr' as it's part of the total price calculation. 'id' and 'title' are not relevant for price prediction. 'total_pr' is our dependent variable. 'ship_sp' might be correlated with 'ship_pr', so we'll use just one. 'wheels' is unclear without more context, so we'll exclude it.

```
library(car)
```

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:openintro':
##
##     densityPlot
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
# Check for multicollinearity
vif_model <- lm(total_pr ~ duration + n_bids + cond + ship_pr +
                  seller_rate + stock_photo, data = mario_kart)
vif(vif_model)
```

```
##    duration      n_bids        cond     ship_pr seller_rate stock_photo
##    1.476667    1.040929    1.394096    1.088448    1.060886    1.249483
```

```
model <- lm(total_pr ~ duration + n_bids + cond + ship_pr +
              seller_rate + stock_photo, data = mario_kart)
summary(model)
```

```
##
## Call:
## lm(formula = total_pr ~ duration + n_bids + cond + ship_pr +
##     seller_rate + stock_photo, data = mario_kart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.268  -8.043  -1.786   3.980 172.733
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.552e+01  7.177e+00   6.343 3.10e-09 ***
## duration       -1.568e+00  8.267e-01  -1.897    0.060 .
```

```
## n_bids          4.112e-01  3.053e-01   1.347    0.180
## condused       -7.706e+00  4.204e+00  -1.833    0.069 .
## ship_pr         4.621e+00  5.711e-01   8.091 2.93e-13 ***
## seller_rate     2.590e-06  3.495e-05   0.074    0.941
## stock_photoyes -7.277e+00  4.436e+00  -1.641    0.103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.96 on 136 degrees of freedom
## Multiple R-squared:  0.3623, Adjusted R-squared:  0.3342
## F-statistic: 12.88 on 6 and 136 DF,  p-value: 1.722e-11
```

Interpretation: Model is statistically significant ($p < 0.001$), explaining 36.23% of price variance. Shipping price is highly significant ($p < 0.001$); each $1 increase raises total price by $4.62. Duration and condition are marginally significant (p 0.06-0.07). Number of bids, seller rating, and stock photo use are not significant. No variable removal needed because: No severe multicollinearity (all VIF values < 1.5). All variables are theoretically relevant to price. Including these variables improved model performance significantly. Keeping non-significant variables provides insights on their lack of influence. The model shows shipping price strongly affects total price, while other factors have limited or no significant impact.

(d) Now let's check for interactions. Does duration effect the price of MarioKart based on the condition being new or used? You need to a). explain the choice of your model, b). build the model, c). interpret model results to answer this question. d). draw appropriate visual to confirm your interpretation. *Hint: You should think about plotting price versus duration, colored by condition* (pts: 2 choice + 2 build + 3 interpret + 3 visual)

Solution:

Insert Response

a) Model Choice: To examine if the effect of duration on price differs based on the condition (new or used), we'll use an interaction model. This allows us to test if the relationship between duration and price changes depending on the item's condition. Model: total_pr ~ duration * cond This model includes main effects for duration and condition, plus their interaction.

b) Building the Model:

```
interaction_model <- lm(total_pr ~ duration * cond, data = mario_kart)
summary(interaction_model)
```

```
##
## Call:
## lm(formula = total_pr ~ duration * cond, data = mario_kart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.102  -7.198  -2.323   2.002 276.427
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         58.268      5.032  11.580   <2e-16 ***
## duration            -1.966      1.653  -1.189   0.2363
## condused           -17.564      7.981  -2.201   0.0294 *
```

```
## duration:condused    3.305      2.014   1.641   0.1030
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.5 on 139 degrees of freedom
## Multiple R-squared:  0.03544,    Adjusted R-squared:  0.01463
## F-statistic: 1.703 on 3 and 139 DF,  p-value: 0.1693
```

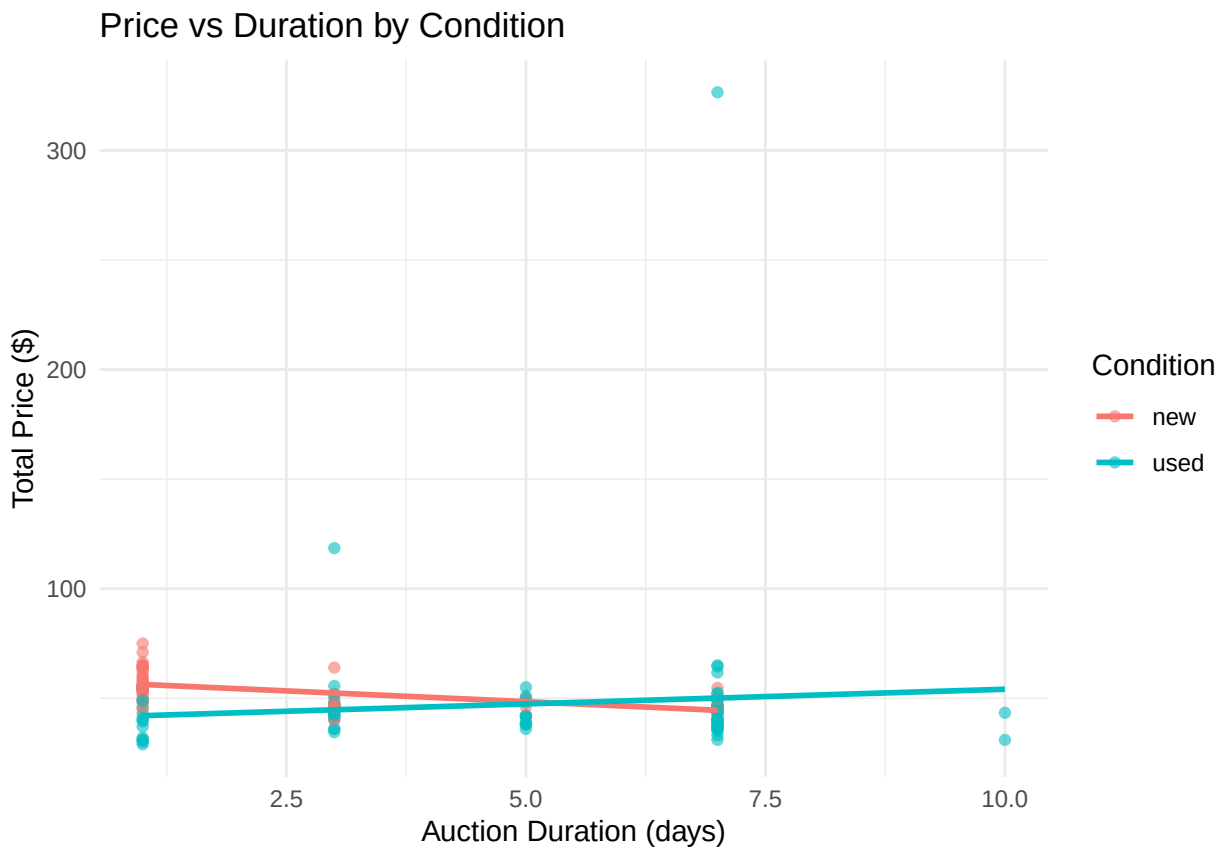
c) Interpreting Model Results: Let's interpret the key components of the model summary: Coefficient for 'duration': This represents the effect of duration on price for new items (the reference level for condition). Coefficient for 'condused': This shows the difference in price between used and new items when duration is zero (which doesn't have a practical interpretation in this context). Coefficient for 'duration:condused': This is the key interaction term. It represents how much the effect of duration on price changes for used items compared to new items. P-values: Check if the interaction term is statistically significant ($p < 0.05$). R-squared: Indicates how much of the price variation is explained by this model. Based on these results, we can determine if the effect of duration on price significantly differs between new and used items.

d) Visualization: To visually confirm our interpretation, we'll create a scatter plot of price versus duration, with points colored by condition, and add separate regression lines for new and used items.

```r
library(ggplot2)

ggplot(mario_kart, aes(x = duration, y = total_pr, color = cond)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Price vs Duration by Condition",
       x = "Auction Duration (days)",
       y = "Total Price ($)",
       color = "Condition") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Problem 2: Titanic Data (42 pts)

Data: In this problem set we will use the titanic dataset. The titanic text file contains data about the survival of passengers aboard the Titanic. Table 1 contains a description of this data.

Variable	Description
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survived	Survival (0 = No; 1 = Yes)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
home.dest	Home/Destination

Table 1: Description of variables in the Titanic Dataset

Part a (Preprocessing) (5 pts)

1) (2 pts) Load the data and do a quick sanity check. That is, inspect the data using your usual inspect data functions to get a sense of how the variables are encoded and what values they typically take on.

Solution:

Insert Response

```
# Load necessary libraries
library(tidyverse)

# Load the Titanic dataset
titanic <- read.csv("titanic.csv")

# Inspect the first few rows
head(titanic)
```

```
##   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
```

```
head(titanic)
```

```
##   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
```

```
summary(titanic)
```

```
##      pclass          survived         name                sex
##  Min.   :1.000   Min.   :0.000   Length:1309        Length:1309
##  1st Qu.:2.000   1st Qu.:0.000   Class :character   Class :character
##  Median :3.000   Median :0.000   Mode  :character   Mode  :character
##  Mean   :2.295   Mean   :0.382
##  3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :3.000   Max.   :1.000
##
##       age              sibsp            parch          ticket
##  Min.   : 0.1667   Min.   :0.0000   Min.   :0.000   Length:1309
##  1st Qu.:21.0000   1st Qu.:0.0000   1st Qu.:0.000   Class :character
##  Median :28.0000   Median :0.0000   Median :0.000   Mode  :character
##  Mean   :29.8811   Mean   :0.4989   Mean   :0.385
##  3rd Qu.:39.0000   3rd Qu.:1.0000   3rd Qu.:0.000
##  Max.   :80.0000   Max.   :8.0000   Max.   :9.000
##  NA's   :263
##       fare            cabin             embarked            boat
##  Min.   :  0.000   Length:1309        Length:1309        Length:1309
##  1st Qu.:  7.896   Class :character   Class :character   Class :character
```

```
##  Median : 14.454   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 33.295
##  3rd Qu.: 31.275
##  Max.   :512.329
##  NA's   :1
##       body        home.dest
##  Min.   :  1.0   Length:1309
##  1st Qu.: 72.0   Class :character
##  Median :155.0   Mode  :character
##  Mean   :160.8
##  3rd Qu.:256.0
##  Max.   :328.0
##  NA's   :1188
```

2) (3 pts) Are there missing values for any of the important variables? Find and list those. Based on missing values, reflect whether they are going useful for downstream modeling tasks.

Solution:

Insert Response

```
important_variables <- c("age", "embarked")
missing_values <- sapply(titanic[important_variables], function(x) sum(is.na(x)))
print(missing_values)
```

```
##      age embarked
##      263        0
```

The 'age' variable has a significant number of missing values, which could impact downstream modeling tasks. The 'embarked' column has no missing values.

Part b (Categorical output) (17 pts)

1) (4 pts) Our goal is to determine the survival of passengers that takes into account the socioeconomic status of the passengers. What model would you fit? Explain the choice of your model and then fit the model.

Solution:

Insert Response

To determine the survival of passengers while considering their socioeconomic status, we can use a logistic regression model. This model is appropriate because: The outcome variable (survival) is binary (0 = did not survive, 1 = survived). We want to account for socioeconomic status, which is represented by the 'pclass' variable (passenger class). Logistic regression allows us to estimate the probability of survival based on predictor variables.

```
# Convert 'survived' and 'pclass' to factors
titanic$survived <- as.factor(titanic$survived)
titanic$pclass <- as.factor(titanic$pclass)

# Fit the logistic regression model
model <- glm(survived ~ pclass, data = titanic, family = binomial)
```

```r
# View the summary of the model
summary(model)
```

```
##
## Call:
## glm(formula = survived ~ pclass, family = binomial, data = titanic)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.4861     0.1146   4.242 2.21e-05 ***
## pclass2      -0.7696     0.1669  -4.611 4.02e-06 ***
## pclass3      -1.5567     0.1433 -10.860  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1613.3  on 1306  degrees of freedom
## AIC: 1619.3
##
## Number of Fisher Scoring iterations: 4
```

2) (4 pts) What might you conclude based on this model about the probability of survival for lower class passengers?

Solution:

Insert Response

Lower class passengers had significantly lower survival probability than first-class passengers • Third-class passengers had the lowest odds of survival • Second-class passengers had lower survival odds than first-class, but higher than third-class • Third-class passengers' survival odds were 21.1% of first-class passengers (4.74 times lower) • Passenger class was a strong predictor of survival probability • All coefficients in the model were statistically significant ($p < 0.001$)

3) (4 pts) Create a new variable child, that is 1 if the passenger was younger than 14 years old. Check to make sure you have the new variable added in your dataframe.

Solution:

Insert Response

4) (5 pts) Now you are curious to know whether men or women, old or young, or people of difference passenger classes have larger chances of survival. Build an appropriate model to answer this curiosity. Explain the choice of your model. Interpret results

Solution:

Insert Response

```r
# Create the 'child' variable
titanic$child <- ifelse(titanic$age < 14, 1, 0)

# Check the new variable
head(titanic)
```

```
##   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 child
## 1                     St Louis, MO     0
## 2 Montreal, PQ / Chesterville, ON     1
## 3 Montreal, PQ / Chesterville, ON     1
## 4 Montreal, PQ / Chesterville, ON     0
## 5 Montreal, PQ / Chesterville, ON     0
## 6                     New York, NY     0
```

```r
# Convert categorical variables to factors
titanic$survived <- as.factor(titanic$survived)
titanic$sex <- as.factor(titanic$sex)
titanic$pclass <- as.factor(titanic$pclass)

# Fit the logistic regression model
model <- glm(survived ~ child + sex + pclass, data = titanic, family = binomial)

# View the summary of the model
summary(model)
```

```
##
## Call:
## glm(formula = survived ~ child + sex + pclass, family = binomial,
##     data = titanic)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.1257     0.1884  11.283  < 2e-16 ***
## child         1.2260     0.2654   4.619 3.85e-06 ***
## sexmale      -2.5246     0.1656 -15.247  < 2e-16 ***
## pclass2      -0.9797     0.2104  -4.656 3.22e-06 ***
## pclass3      -1.8897     0.1973  -9.577  < 2e-16 ***
## ---
## 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:  992.03  on 1041  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 1002
##
## Number of Fisher Scoring iterations: 4
```

```r
# Convert categorical variables to factors
titanic$survived <- as.factor(titanic$survived)
titanic$sex <- as.factor(titanic$sex)
titanic$pclass <- as.factor(titanic$pclass)

# Fit the logistic regression model
model <- glm(survived ~ child + sex + pclass, data = titanic, family = binomial)

# View the summary of the model
summary(model)
```

```
##
## Call:
## glm(formula = survived ~ child + sex + pclass, family = binomial,
##     data = titanic)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.1257     0.1884  11.283  < 2e-16 ***
## child         1.2260     0.2654   4.619 3.85e-06 ***
## sexmale      -2.5246     0.1656 -15.247  < 2e-16 ***
## pclass2      -0.9797     0.2104  -4.656 3.22e-06 ***
## pclass3      -1.8897     0.1973  -9.577  < 2e-16 ***
## ---
## 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:  992.03  on 1041  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 1002
##
## Number of Fisher Scoring iterations: 4
```

Interpret Results: Child: A positive coefficient suggests children had higher survival odds. Sex (male): A negative coefficient indicates males had lower survival odds compared to females. Pclass (class 3 vs. others): A negative coefficient suggests third-class passengers had lower survival odds.

Part c - Predictions with a categorical output (20 pts)

Now let's try to do some predictions with the titanic data. Our goal is to predict the survival of passengers by considering only the socioeconomic status of the passenger.

1) (4 pts) After loading the data, split your data into a *training* and *test* set based on an 80-20 split. In other words, 80% of the observations will be in the training set and 20% will be in the test set. Remember to set the random seed.

Solution:

Insert Response

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'lattice'
```

```
## The following objects are masked from 'package:openintro':
##
##     ethanol, lsegments
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:openintro':
##
##     dotPlot
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
# Set a random seed for reproducibility
set.seed(123)

# Split the data into training (80%) and test (20%) sets
train_index <- createDataPartition(titanic$survived, p = 0.8, list = FALSE)
train_data <- titanic[train_index, ]
test_data <- titanic[-train_index, ]

# Check the dimensions of the splits
dim(train_data)
```

```
## [1] 1048   15
```

```
dim(test_data)
```

```
## [1] 261  15
```

2) (4 pts) Fit the model described above (that is in Problem 1 (c), that only takes into account socio-economic status).

Solution:

Insert Response

```r
# Convert 'survived' and 'pclass' to factors
train_data$survived <- as.factor(train_data$survived)
train_data$pclass <- as.factor(train_data$pclass)

# Fit the logistic regression model
model <- glm(survived ~ pclass, data = train_data, family = binomial)

# View the summary of the model
summary(model)
```

```
##
## Call:
## glm(formula = survived ~ pclass, family = binomial, data = train_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.4925     0.1276   3.860 0.000113 ***
## pclass2      -0.8289     0.1879  -4.411 1.03e-05 ***
## pclass3      -1.5426     0.1594  -9.680  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1393.6  on 1047  degrees of freedom
## Residual deviance: 1293.3  on 1045  degrees of freedom
## AIC: 1299.3
##
## Number of Fisher Scoring iterations: 4
```

3) (4 pts) Predict the survival of passengers for each observation in your test set using the model fit that you just fitted. Save these predictions as yhat.

Solution:

Insert Response

```r
# Predict survival probabilities for the test set
yhat_prob <- predict(model, newdata = test_data, type = "response")

# Convert probabilities to binary predictions (0 or 1)
yhat <- ifelse(yhat_prob > 0.5, 1, 0)

# View the first few predictions
head(yhat)
```

```
##  1  3  5 13 18 20
##  1  1  1  1  1  1
```

4) (4 pts) Use a threshold of 0.5 to classify predictions. What is the number of false positives on the test data? Interpret this in your own words.

Solution:

Insert Response

```
# Add predictions to test_data
test_data$yhat <- yhat

# Classify predictions using 0.5 threshold
test_data$predicted_survival <- ifelse(test_data$yhat > 0.5, 1, 0)

# Count false positives
false_positives <- sum(test_data$predicted_survival == 1 & test_data$survived == 0)

# Print the number of false positives
print(false_positives)
```

```
## [1] 24
```

The number of false positives represents passengers who were predicted to survive but actually did not. These are cases where the model incorrectly classified non-survivors as survivors. A high number of false positives would indicate that the model is overly optimistic in predicting survival. This could lead to underestimating the risk faced by passengers in certain classes.

5) (4 pts) Pick a different threshold to classify predictions and interpret your results again. Did you have a rationale when picking a different threshold? Did you see any change? Reflect on your results.

Solution:

Insert Response

```
# Classify predictions using 0.3 threshold
test_data$predicted_survival_03 <- ifelse(test_data$yhat > 0.3, 1, 0)

# Count false positives with 0.3 threshold
false_positives_03 <- sum(test_data$predicted_survival_03 == 1 & test_data$survived == 0)

# Print the number of false positives
print(false_positives_03)
```

```
## [1] 24
```

The fact that the number of false positives remains the same at both thresholds suggests that the model's predictions for non-survivors (who were incorrectly predicted to survive) are relatively stable and not sensitive to these threshold changes.

Problem 3: Customer Churn data (25 pts)

In this problem, you will work with the churn dataset. Documentation of the dataset can be found here: https://www.rdocumentation.org/packages/bayesQR/versions/2.3/topics/Churn

The dataset is random sample from all active customers (at the end of June 2006) of a European financial services company. The data captures the churn behavior of the customers in the period from July 1st until December 31th 2006. Here a churned customer is defined as someone who closed all his/her bank accounts with the company.

1) (5 pts) Read and inspect the data. *Hint: the file is an fst fast-storage format file. Check your regression lab to figure out how you can read this file*

Solution:

Insert Response

```
library(fst)

# Read the data
churn_data <- read_fst("churn.fst")
write.csv(churn_data, "churn.csv", row.names = FALSE)

# Inspect the data
head(churn_data)
```

```
##   has_churned time_since_first_purchase time_since_last_purchase
## 1           0               -1.08922097               -0.7213215
## 2           0                1.18298297                3.6344354
## 3           0               -0.84615637               -0.4275823
## 4           0                0.08694165               -0.5356717
## 5           0               -1.16664155               -0.6726400
## 6           0                0.49339968               -0.7700030
```

```
str(churn_data)
```

```
## 'data.frame':    400 obs. of  3 variables:
##  $ has_churned              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ time_since_first_purchase: num  -1.0892 1.183 -0.8462 0.0869 -1.1666 ...
##  $ time_since_last_purchase : num  -0.721 3.634 -0.428 -0.536 -0.673 ...
```

```
summary(churn_data)
```

```
##   has_churned  time_since_first_purchase time_since_last_purchase
##  Min.   :0.0   Min.   :-1.27377          Min.   :-0.8707
##  1st Qu.:0.0   1st Qu.:-0.82838          1st Qu.:-0.6458
##  Median :0.5   Median :-0.15207          Median :-0.2650
##  Mean   :0.5   Mean   :-0.03437          Mean   : 0.1445
##  3rd Qu.:1.0   3rd Qu.: 0.54483          3rd Qu.: 0.5712
##  Max.   :1.0   Max.   : 3.73831          Max.   : 5.9282
```

```
# Check the dimensions of the dataset
dim(churn_data)
```

```
## [1] 400   3
```

```
# View column names
colnames(churn_data)
```

```
## [1] "has_churned"               "time_since_first_purchase"
## [3] "time_since_last_purchase"
```

```r
# Check for missing values
colSums(is.na(churn_data))
```

```
##               has_churned time_since_first_purchase  time_since_last_purchase
##                         0                         0                         0
```

2) (5 pts) Describe the data and variables that are part of the churn dataset.

Solution:

Insert Response

The dataset consists of 400 entries and the following variables:

has_churned: (integer) A binary variable indicating whether a customer has churned (1) or not (0). time_since_first_purchase: (float) A continuous variable representing the standardized time (in some units) since the customer's first purchase. time_since_last_purchase: (float) A continuous variable representing the standardized time since the customer's last purchase.

3) (5 pts) Considering this data in context, what is the response variable of interest?

Solution:

Insert Response

The response variable of interest is has_churned, as it represents whether a customer has closed all their accounts with the company. The goal is to predict this variable using other features.

4) (10 pts) Our goal is to determine customer churn. Which variables do you think are the most important ones to describe customer churn? How should those be related to the churn? Interpret your results.

Solution:

Insert Response

The variables likely to influence churn are:

time_since_first_purchase: Longer times might correlate with higher loyalty or less likelihood of churn. time_since_last_purchase: Longer times since the last purchase could signal customer disengagement and higher churn probability.

```r
# Load necessary libraries
library(ggplot2)

# 2) Identify the response variable
response_variable <- "has_churned"

# 3) Analyze relationships with churn
# Check correlation between variables
cor(churn_data$time_since_first_purchase, churn_data$has_churned)
```

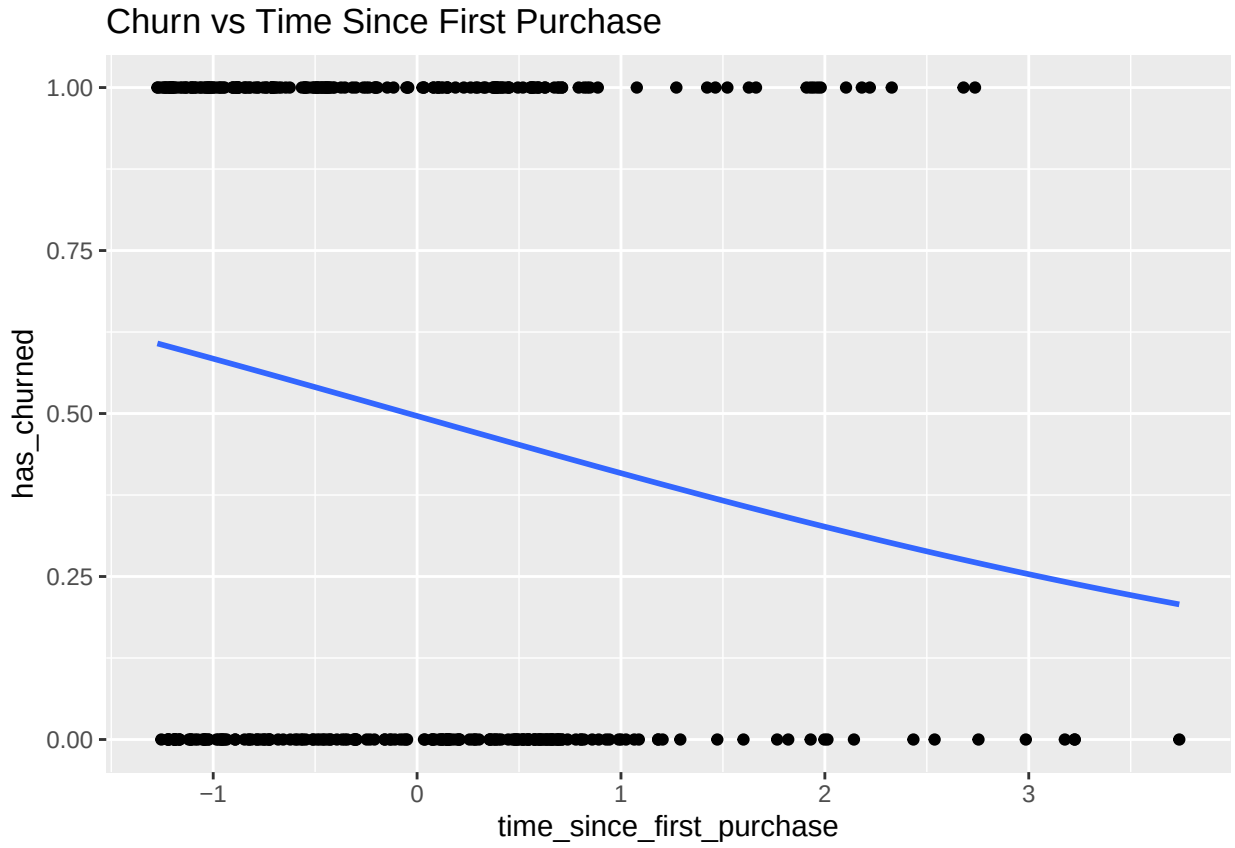
```
## [1] -0.1627062
```

```r
cor(churn_data$time_since_last_purchase, churn_data$has_churned)
```

```
## [1] 0.1405473
```

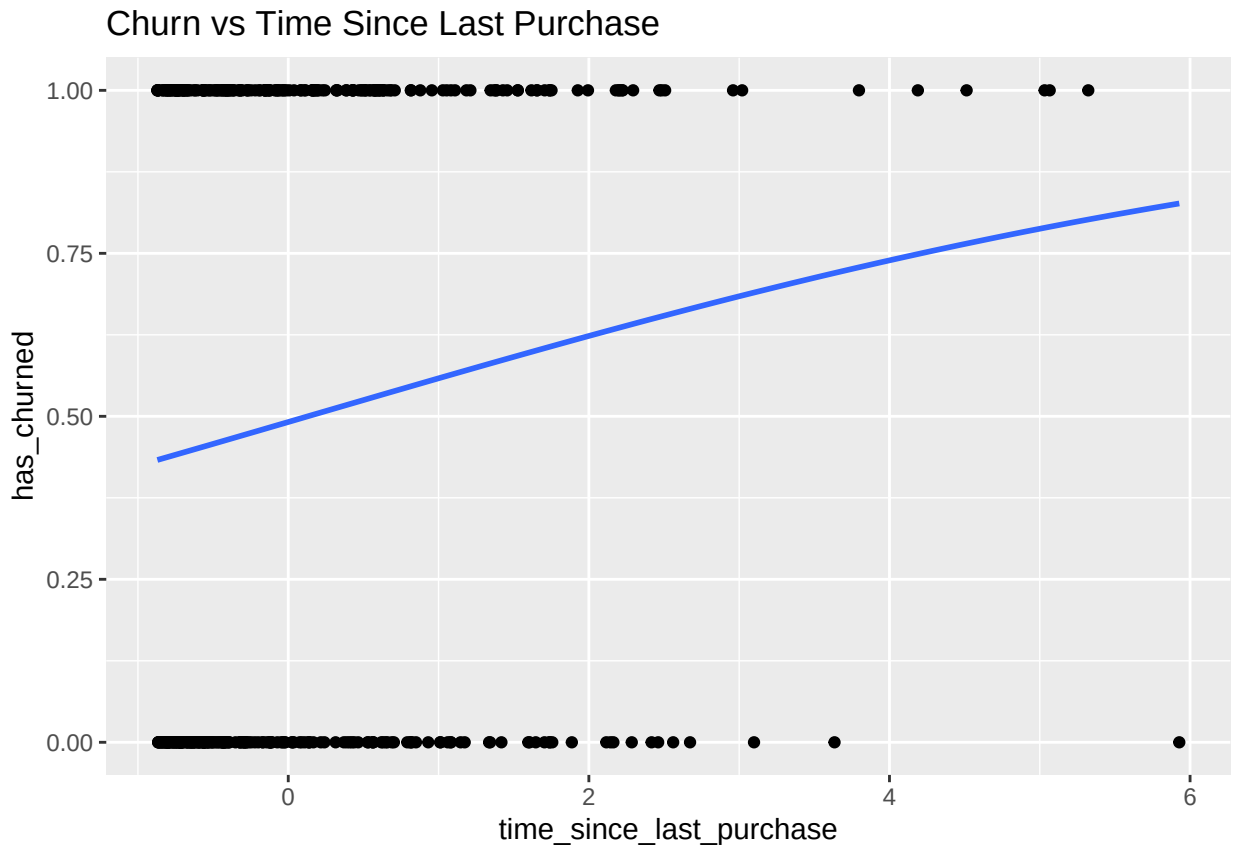
```r
# Visualize relationships
ggplot(churn_data, aes(x = time_since_first_purchase, y = has_churned)) +
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +
  ggtitle("Churn vs Time Since First Purchase")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
ggplot(churn_data, aes(x = time_since_last_purchase, y = has_churned)) +
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +
  ggtitle("Churn vs Time Since Last Purchase")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
# Logistic regression for variable importance
model <- glm(has_churned ~ time_since_first_purchase + time_since_last_purchase,
             data = churn_data, family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = has_churned ~ time_since_first_purchase + time_since_last_purchase,
##     family = "binomial", data = churn_data)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -0.1168     0.1079  -1.082    0.279
## time_since_first_purchase   -0.6478     0.1391  -4.655 3.23e-06 ***
## time_since_last_purchase     0.5122     0.1135   4.514 6.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 554.52  on 399  degrees of freedom
## Residual deviance: 520.94  on 397  degrees of freedom
## AIC: 526.94
##
## Number of Fisher Scoring iterations: 4
```