Regression HW

Introduction

In this homework, you'll get a chance to practice the linear regression skills you learned in class on some real-world data! This data was taken from an online data science competition called SLICED! SLICED is like the cooking competition CHOPPED, but for data science. The dataset in question contains metadata about Super Store sales of different products as well as the profit made. Here is a data dictionary:

Variable	Description
id	unique id per row
ship_mode	what mode was used to ship the item: 'First Class', 'Same Day', 'Second Class', 'Standard Class'
segment	whether the recipient is corporate or consumer
country	always United States
city	where the item was shipped (a city in the United States)
state	where the item was shipped (a state in the United States)
postal_code	ZIP Code in the United States
region	region in the United States: 'Central', 'East', 'South', 'West'
category	type of item: 'Furniture', 'Office Supplies', 'Technology'
sub_category	sub-type of item: 'Accessories', 'Appliances', 'Art', 'Binders', 'Bookcases', 'Chairs', 'Copiers', 'Envelopes', 'Fasteners', 'Furnishings', 'Labels', 'Machines', 'Paper', 'Phones', 'Storage', 'Supplies', 'Tables'
sales	sales made for the order (USD)
quantity	quantity sold of the item
discount	discount applied to the order(from 0 to 1)

Variable	Description
profit	profit made on the order (USD)

Getting started

Here are the steps for getting started:

- Start with the assignment link that creates a repo on GitHub with starter documents. I have sent this to you through email.
- Clone this repo in RStudio
- Make any changes needed as outlined by the tasks you need to complete for the assignment
- Periodically commit changes (the more often the better, for example, once per each new task)
  - Remember, git will yell at you when you try to commit before running the following lines in the terminal
    * `git config --global user.name "Your Name Here"`
    * `git config --global user.email "Your Email Here"`
- Push all your changes back to your GitHub repo

and voila, you're done! Once you push your changes back you do not need to do anything else to "submit" your work. And you can of course push multiple times throughout the assignment. At the time of the deadline I will take whatever is in your repo and consider it your final submission, and grade the state of your work at that time (which means even if you made mistakes before then, you wouldn't be penalized for them as long as the final state of your work is correct).

Assignment

The first thing that we'll need to do is setup our R/RStudio session so that we can do our data analysis and perform linear regression. This involves loading the packages we'll need for the project. I've loaded the correct packages below. Remember to install the packages with `install.packages()` before you load them (you only have to install once, but load when starting a new session). I've added an option to the chunk below called `message: FALSE` to turn off all the messages given when loading the two packages.

```
library("tidyverse");theme_set(theme_bw())
library("tidymodels");theme_set(theme_bw())
```

Now we need to load the data into R! There are two datasets in the project space, `super_stores_sales_train.csv` and `super_stores_sales_test.csv`. We don't fully know about the concept of training and testing datasets yet, but I want you to load them into R with the `read_csv()` function and call them `train` and `test` respectively. Do that below now!

```
# load train
train <- read_csv("super_store_sales_train.csv")
# load test
test <- read_csv("super_store_sales_test.csv")
```

IMPORTANT: From now on, assume you are working with the `train` dataset unless explicitly told to use `test`.

The next step in your data analysis should always be to make sure that your data was parsed correctly by the loading function. This can mean making sure that numbers are parsed as numbers, words as characters, TRUE's and FALSE's as logicals. For this assignment, you can trust that everything works out right when loading the data.

Before we begin to perform linear regression, we need to understand the goal of the analysis and do some exploratory data analysis (EDA) to understand how the data is behaving. In the SLICED competition, the goal was to predict profit based on any of the metadata on the sales. For this assignment, we will focus on both prediction and inference with profit being the response variable and everything else as possible predictors.

When engaging in EDA your goal is to extract information that will inform your modeling choices down the road. For linear regression, and really any other "learning" method, we are looking for predictors that are correlated to the response. That should be what we focus on finding out during this EDA.

EDA

It is important to understand how the response behaves and so I want you to make a histogram of `profit`. By default, it will show you one very large bar at 0 because of some outliers. Zoom in on the picture by manually setting the x-axis limits to -100 to 500 using the `xlim()` function. Comment on the plot and if anything surprises you.

```
# plot
train
```

```
# A tibble: 7,173 x 14
      id ship_mode    segment country city  state posta~1 region categ~2 sub_c~3
   <dbl> <chr>        <chr>   <chr>   <chr> <chr>   <dbl> <chr>  <chr>   <chr>
```

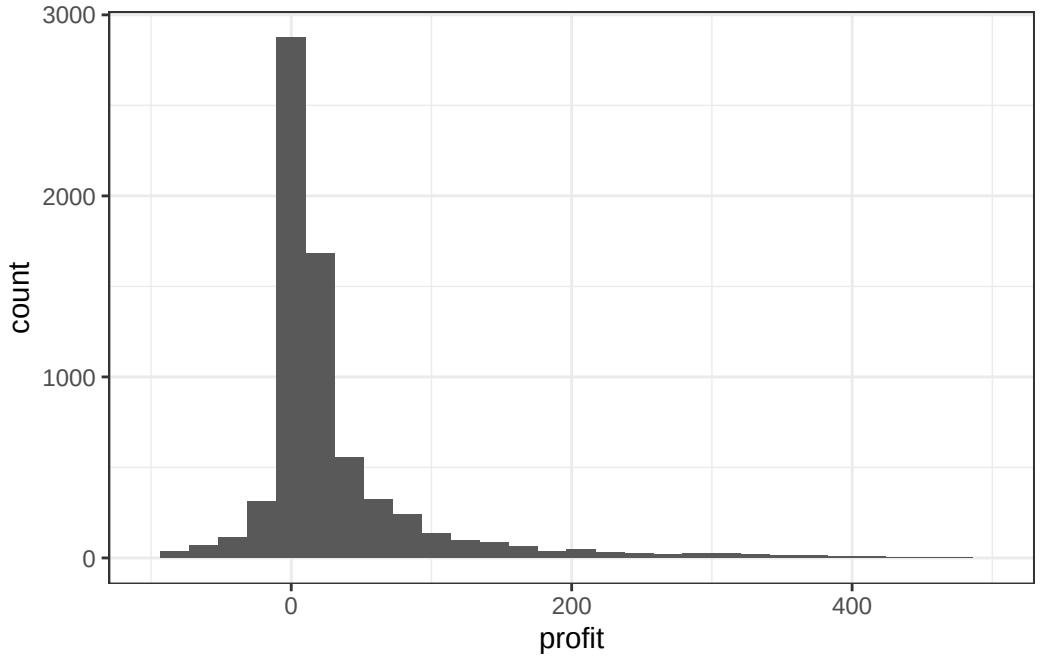
```
 1  8522 Second Class Consum~ United~ Ever~ Mass~    2149 East   Office~ Art
 2  7864 Standard Cl~ Corpor~ United~ Los ~ Cali~   90036 West   Furnit~ Chairs
 3  3522 Standard Cl~ Consum~ United~ New ~ New ~   10035 East   Office~ Storage
 4  8694 Standard Cl~ Consum~ United~ Los ~ Cali~   90045 West   Office~ Art
 5  2306 Second Class Home O~ United~ San ~ Cali~   94110 West   Furnit~ Bookca~
 6  8742 Standard Cl~ Corpor~ United~ Arli~ Texas   76017 Centr~ Furnit~ Furnis~
 7  2292 Standard Cl~ Consum~ United~ Yonk~ New ~   10701 East   Techno~ Phones
 8   713 Standard Cl~ Corpor~ United~ Spri~ Virg~   22153 South  Office~ Binders
 9  5904 First Class  Consum~ United~ Des ~ Illi~   60016 Centr~ Furnit~ Furnis~
10  6023 Second Class Consum~ United~ Prov~ Rhod~    2908 East   Office~ Binders
# ... with 7,163 more rows, 4 more variables: sales <dbl>, quantity <dbl>,
#   discount <dbl>, profit <dbl>, and abbreviated variable names
#   1: postal_code, 2: category, 3: sub_category
```

```
ggplot(data = train, aes(x=profit)) + geom_histogram() + xlim(-100, 500)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 275 rows containing non-finite values (stat_bin).
```

```
Warning: Removed 2 rows containing missing values (geom_bar).
```



Comments: The majority of profits are well below $200, which makes sense if we consider a "super store" to be something similar to a walmart which runs on volume and low costs. It is somewhat surprising how many transactions result in a loss, because the

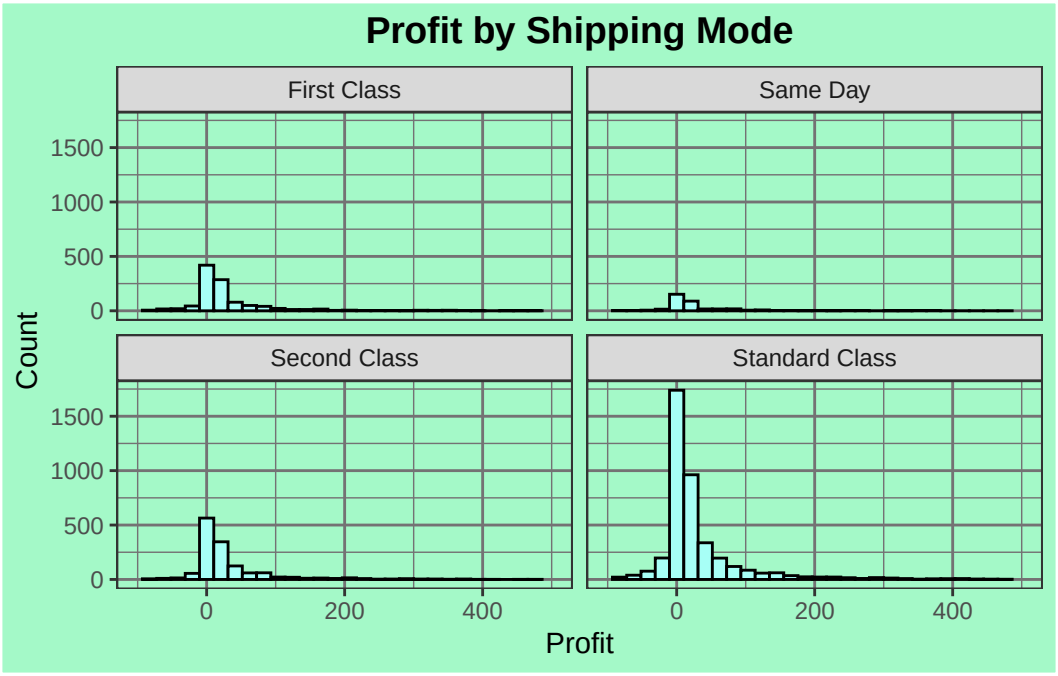
Now we are interested in seeing if we can determine if any variables or combination of variables are correlated with the response. Let's start with some categorical variables `ship_mode` and `segment`. First, I want you to find the mean profit amount for each mode of shipping using `group_by()` and `summarize()`. Next, I want you to make histograms of `profit` that are faceted by `ship_mode`. In addition to regular ggplotting functions, you'll want to use `facet_wrap()` to do the faceting. Also, remember to use `vars()` instead of `aes()` inside `facet_wrap()`. You'll probably want to limit the x-axis scale with `xlim()` like you did previously. Comment on any possible correlation between the variables.

```
# mean profit for every ship mode
meanProfitData <- train %>%
    group_by(ship_mode) %>%
    summarise(meanProfit = mean(profit))

#plot
ggplot(data = train, aes(x = profit)) + geom_histogram(fill="#A7FFF6", color="black") + fa
  theme(panel.background = element_rect(fill="#A4F9C8"), panel.grid.major= element_line(co
        plot.title = element_text(face="bold", size=14, hjust=.5),
        plot.subtitle = element_text(hjust=.5))
```

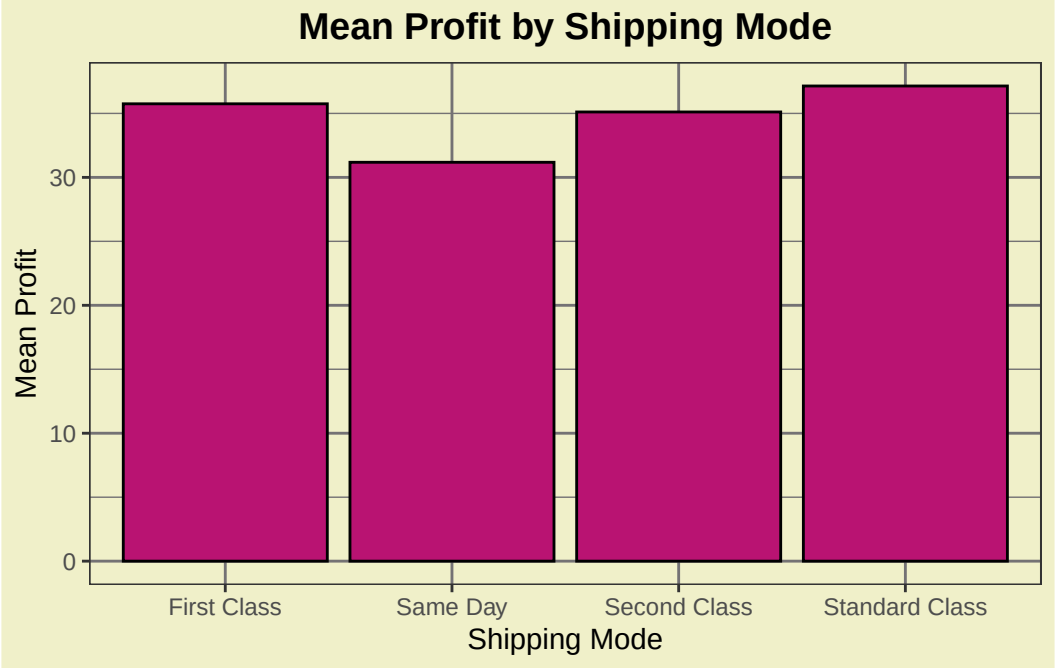
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 275 rows containing non-finite values (stat_bin).
```

```
Warning: Removed 8 rows containing missing values (geom_bar).
```




```
##Wasn't sure if this implied we should graph mean profit as well, so included in case

ggplot(data=meanProfitData, aes(x=ship_mode, y=meanProfit)) +
  geom_bar(stat="identity",fill="#B91372", color="black") + labs(title="Mean Profit by Shi
  theme(panel.background = element_rect(fill="#F0F0C9"), panel.grid.major= element_line(co
        plot.title = element_text(face="bold", size=14, hjust=.5),
        plot.subtitle = element_text(hjust=.5))
```



Comments: It's interesting to note that regardless of the profit distribution, the vast majority of classes are sent standard. Also interesting, the profit distribution of same day delivery appears to be less than that of first or second class.

What about `ship_mode` and `segment` together? Find the mean profit for each `ship_mode` and `segment` pair. You'll need to use `group_by()` and `summarize()` again. Any correlation?

```
# mean profit faceted on both variables
meanProfitDataTwo <- train %>%
    group_by(ship_mode, segment) %>%
    summarise(meanProfit = mean(profit))
```

```
`summarise()` has grouped output by 'ship_mode'. You can override using the
`.groups` argument.
```

```
meanProfitDataTwo
```

```
# A tibble: 12 x 3
# Groups:   ship_mode [4]
   ship_mode      segment     meanProfit
```

```
   <chr>          <chr>           <dbl>
 1 First Class    Consumer         29.5
 2 First Class    Corporate        27.2
 3 First Class    Home Office      69.2
 4 Same Day       Consumer         34.0
 5 Same Day       Corporate        20.1
 6 Same Day       Home Office      33.7
 7 Second Class   Consumer         29.5
 8 Second Class   Corporate        36.6
 9 Second Class   Home Office      48.8
10 Standard Class Consumer         32.4
11 Standard Class Corporate        42.8
12 Standard Class Home Office      41.8
```

Comments: Home office appears to profit well regardless of how it is shipped. On the other hand, profit for corporate goods appears to correlate quite a bit with how they are shipped.

There are a number of different categorical variables that correspond to location in some sense (`country`, `state`, `city`, `region`, and `zip code`). Certainly they all convey similar information. The question becomes how fine a mesh do we need to capture the variation in over location without overfitting. An important fact to know is that the holdout samples (the data in the test dataset) are certain states. Discuss why it would be unwise to use `state`, or anything finer than `state`, as a predictor in the model. Also, discuss why `id` and `country` are bad predictors as well.

Comments: Different states have different costs of living, so this does play a role: however, as a main predictor it doesn't tell us much at all about what the consumer is buying and even with this difference in cost of living it's not much of a difference. Country would be bad because pretty much the entire dataset is the US and it's too broad to derive much information from. City/region/zip code are too narrow and the variation in cost between these doesn't make up for the fact that it partitions the data into very small subsets

Now we'll explore the `category` and `sub_category` variables! Again use `group_by()` and `summarize()` to show the mean profit and sample size (which you can find using the `n()` function) for each combination of category and sub-category. Then comment on your findings.

```
# summary for combinations of category and sub_category
categories <- train %>%
    group_by(category, sub_category) %>%
    summarise(meanProfit = mean(profit), sampleSize = n())
```

```
`summarise()` has grouped output by 'category'. You can override using the
`.groups` argument.
```

```
categories
```

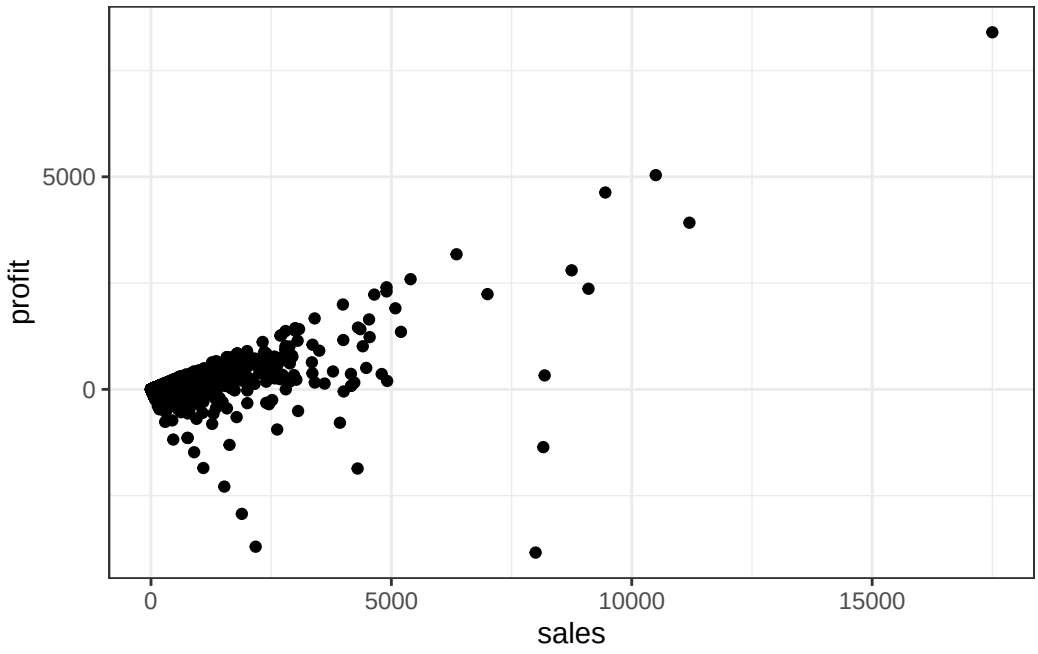
```
# A tibble: 17 x 4
# Groups:   category [3]
   category        sub_category meanProfit sampleSize
   <chr>           <chr>             <dbl>      <int>
 1 Furniture       Bookcases         13.2         172
 2 Furniture       Chairs            52.3         442
 3 Furniture       Furnishings       13.3         664
 4 Furniture       Tables           -56.9         230
 5 Office Supplies Appliances        40.1         342
 6 Office Supplies Art                8.94        574
 7 Office Supplies Binders           22.4        1058
 8 Office Supplies Envelopes         29.5         189
 9 Office Supplies Fasteners          4.49        152
10 Office Supplies Labels            14.2         264
11 Office Supplies Paper             26.1        1023
12 Office Supplies Storage           32.5         611
13 Office Supplies Supplies           5.24        140
14 Technology      Accessories       56.6         550
15 Technology      Copiers          897.           47
16 Technology      Machines         268.           78
17 Technology      Phones            70.1         637
```

Comments: Tables lose a lot of money, maybe because their size makes distribution difficult, and to some extnt a smaller sample size than some categories. Copiers seem to make a ton of money, which despite their cost seems like a larger profit margin that I wouldve expected

Finally lets look at the quantitative predictors! I would like you to find and comment on the following:

- Scatterplot of `sales` vs. `profit`
- Correlation between `sales` and `profit` using `cor()`
- Correlation between `quantity` and `profit`
- Correlation between `discount` and `profit`
- Scatterplot of `sales / quantity` and `profit`

```
ggplot(train, aes(x=sales, y=profit)) + geom_point()
```




```
cor(train$sales, train$profit)
```

```
[1] 0.6218926
```

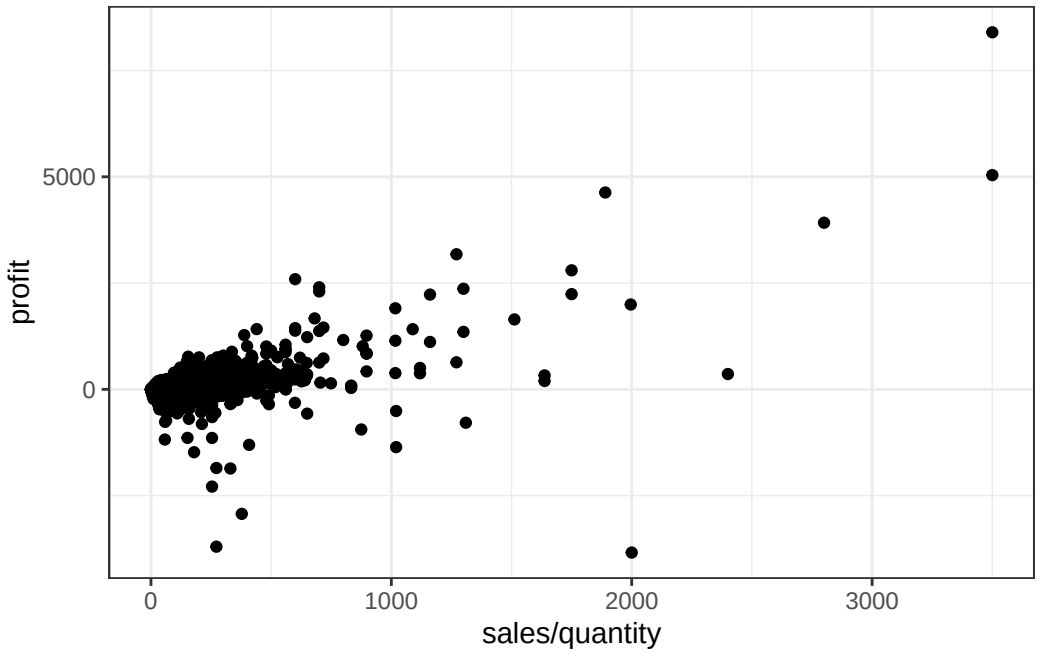
```
cor(train$quantity, train$profit)
```

```
[1] 0.0815635
```

```
cor(train$discount, train$profit)
```

```
[1] -0.2054159
```

```
ggplot(train, aes(x=sales/quantity, y=profit)) + geom_point()
```



Comments: 1) There is greater variance between data points at the far end of sales and profit. 2) There is a moderate correlation between sales and profit 3) There is no meaningful relationship between quantity and profit 4) There is a negligible negative correlation between discount and profit 5) There does appear to be correlation in this plot, even as sales/quantity approaches high/outlier values

Modeling

Let's start with just a simple model using `sales` as the predictor and `profit` as the response. Make the model, look at the summary, and comment on the usefulness of the model. Make sure to check the assumptions of simple linear regression!

```
fit <- lm(profit ~ sales, data = train)
fit
```

```
Call:
lm(formula = profit ~ sales, data = train)

Coefficients:
(Intercept)        sales
   -18.1576       0.2327
```

```r
summary(fit)
```

```
Call:
lm(formula = profit ~ sales, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-5683.6     5.2    19.6    24.9  4345.5

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.157582   2.209246  -8.219 2.42e-16 ***
sales         0.232721   0.003461  67.249  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 174.1 on 7171 degrees of freedom
Multiple R-squared:  0.3868,    Adjusted R-squared:  0.3867
F-statistic:  4522 on 1 and 7171 DF,  p-value: < 2.2e-16
```

Comments: The model estimates that about $0.23 of profit is made for every $1 of sales. The P value of 2.2e-16 indicates there is a strong relationship between these variables, but an r-squared of .39 indicates we should see significant variance along the way.

Now let's try `sales / quantity` instead of sales. To do this, you are going to have to use a special syntax in the formula. Instead of writing `profit ~ sales / quantity` you need to write `profit ~ I(sales / quantity)`. The `I()` function tells the formula to treat whatever is inside it as literal. The division sign has a special meaning inside the formula syntax without the `I()`. Comment on the quality of fit and compare it to the previous one. Make sure to also look at the assumptions when comparing the two.

```r
fit2 <- lm(profit ~ I(sales/quantity), data = train)
fit2
```

```
Call:
lm(formula = profit ~ I(sales/quantity), data = train)

Coefficients:
```

```
      (Intercept)  I(sales/quantity)
         -20.4585             0.9165
```

```
summary(fit2)
```

```

Call:
lm(formula = profit ~ I(sales/quantity), data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-5652.6    -2.7    19.2    25.4  5212.6

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -20.45853    2.33170  -8.774   <2e-16 ***
I(sales/quantity)   0.91654    0.01514  60.553   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 180.9 on 7171 degrees of freedom
Multiple R-squared:  0.3383,    Adjusted R-squared:  0.3382
F-statistic:  3667 on 1 and 7171 DF,  p-value: < 2.2e-16
```

Comments: The model estimates that about 92 cents of profit are gained for every 1 point move of sales/quantity. With an r-squared of .34, there is significant variance along the trendline, but the p-value of 2.2e-16 shows us there is a correlation here.

Let's try to improve on our fit by adding in more variables that we saw were important from our EDA. I want you to include `sales`, `quantity`, `discount`, the interaction between `category` and `subcategory`, and the interaction between `ship_mode` and `segment`. To convey interactions between variables, you use the colon operator. For example, `ship_mode:segment` conveys the interaction between variables. Comment on the fit and compare to the other fits

```
fit3 <- lm(profit ~ sales + quantity + discount + category:sub_category + ship_mode:segmen
fit3
```

```

Call:
lm(formula = profit ~ sales + quantity + discount + category:sub_category +
```

```
    ship_mode:segment, data = train)

Coefficients:
                                        (Intercept)
                                          -127.8206
                                              sales
                                             0.2406
                                           quantity
                                            -4.7890
                                           discount
                                           -219.4559
      categoryFurniture:sub_categoryAccessories
                                                 NA
categoryOffice Supplies:sub_categoryAccessories
                                                 NA
     categoryTechnology:sub_categoryAccessories
                                           165.4858
       categoryFurniture:sub_categoryAppliances
                                                 NA
 categoryOffice Supplies:sub_categoryAppliances
                                           168.0154
      categoryTechnology:sub_categoryAppliances
                                                 NA
              categoryFurniture:sub_categoryArt
                                                 NA
        categoryOffice Supplies:sub_categoryArt
                                           160.3569
             categoryTechnology:sub_categoryArt
                                                 NA
          categoryFurniture:sub_categoryBinders
                                                 NA
    categoryOffice Supplies:sub_categoryBinders
                                           209.6386
         categoryTechnology:sub_categoryBinders
                                                 NA
        categoryFurniture:sub_categoryBookcases
                                            67.0892
  categoryOffice Supplies:sub_categoryBookcases
                                                 NA
       categoryTechnology:sub_categoryBookcases
                                                 NA
           categoryFurniture:sub_categoryChairs
                                           107.4662
```

```
      categoryOffice Supplies:sub_categoryChairs
                                                NA
           categoryTechnology:sub_categoryChairs
                                                NA
           categoryFurniture:sub_categoryCopiers
                                                NA
     categoryOffice Supplies:sub_categoryCopiers
                                                NA
          categoryTechnology:sub_categoryCopiers
                                          529.2610
        categoryFurniture:sub_categoryEnvelopes
                                                NA
   categoryOffice Supplies:sub_categoryEnvelopes
                                          172.0297
       categoryTechnology:sub_categoryEnvelopes
                                                NA
        categoryFurniture:sub_categoryFasteners
                                                NA
   categoryOffice Supplies:sub_categoryFasteners
                                          163.8926
       categoryTechnology:sub_categoryFasteners
                                                NA
      categoryFurniture:sub_categoryFurnishings
                                          168.2187
 categoryOffice Supplies:sub_categoryFurnishings
                                                NA
     categoryTechnology:sub_categoryFurnishings
                                                NA
           categoryFurniture:sub_categoryLabels
                                                NA
     categoryOffice Supplies:sub_categoryLabels
                                          164.9791
          categoryTechnology:sub_categoryLabels
                                                NA
         categoryFurniture:sub_categoryMachines
                                                NA
    categoryOffice Supplies:sub_categoryMachines
                                                NA
        categoryTechnology:sub_categoryMachines
                                           45.6333
            categoryFurniture:sub_categoryPaper
                                                NA
       categoryOffice Supplies:sub_categoryPaper
```

```
                                      172.1167
       categoryTechnology:sub_categoryPaper
                                            NA
       categoryFurniture:sub_categoryPhones
                                            NA
 categoryOffice Supplies:sub_categoryPhones
                                            NA
      categoryTechnology:sub_categoryPhones
                                      148.1339
      categoryFurniture:sub_categoryStorage
                                            NA
 categoryOffice Supplies:sub_categoryStorage
                                      128.3388
     categoryTechnology:sub_categoryStorage
                                            NA
     categoryFurniture:sub_categorySupplies
                                            NA
categoryOffice Supplies:sub_categorySupplies
                                      104.2587
    categoryTechnology:sub_categorySupplies
                                            NA
       categoryFurniture:sub_categoryTables
                                            NA
 categoryOffice Supplies:sub_categoryTables
                                            NA
      categoryTechnology:sub_categoryTables
                                            NA
       ship_modeFirst Class:segmentConsumer
                                      -2.9132
          ship_modeSame Day:segmentConsumer
                                       5.5695
      ship_modeSecond Class:segmentConsumer
                                      -8.5393
    ship_modeStandard Class:segmentConsumer
                                      -1.8032
      ship_modeFirst Class:segmentCorporate
                                       0.2883
         ship_modeSame Day:segmentCorporate
                                     -72.5859
     ship_modeSecond Class:segmentCorporate
                                      -6.4441
   ship_modeStandard Class:segmentCorporate
                                       3.2000
```

```
        ship_modeFirst Class:segmentHome Office
                                        4.3741
           ship_modeSame Day:segmentHome Office
                                        1.0123
         ship_modeSecond Class:segmentHome Office
                                        2.9697
      ship_modeStandard Class:segmentHome Office
                                            NA
```

```
summary(fit3)
```

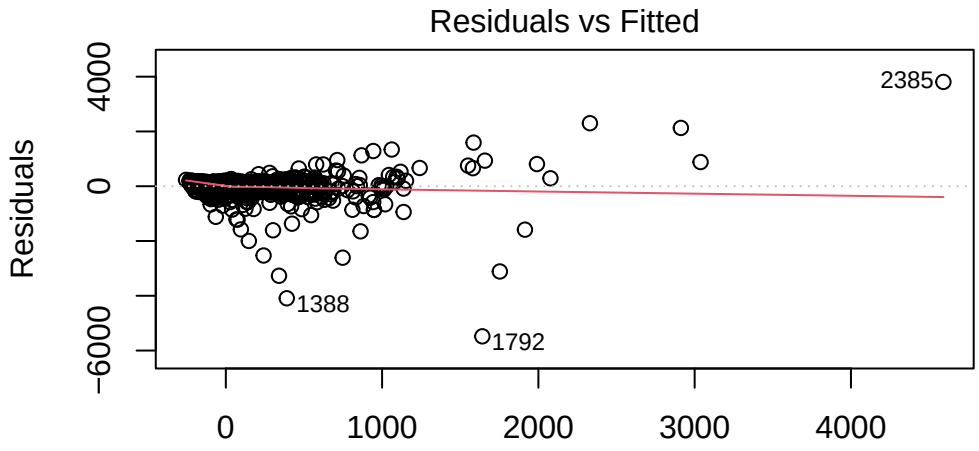
```
Call:
lm(formula = profit ~ sales + quantity + discount + category:sub_category +
    ship_mode:segment, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-5481.3   -23.3     0.3    30.4  3808.4

Coefficients: (36 not defined because of singularities)
                                                Estimate Std. Error t value
(Intercept)                                   -1.278e+02  1.287e+01  -9.931
sales                                          2.406e-01  3.767e-03  63.870
quantity                                      -4.789e+00  8.747e-01  -5.475
discount                                      -2.195e+02  1.077e+01 -20.374
categoryFurniture:sub_categoryAccessories             NA         NA      NA
categoryOffice Supplies:sub_categoryAccessories       NA         NA      NA
categoryTechnology:sub_categoryAccessories     1.655e+02  1.292e+01  12.808
categoryFurniture:sub_categoryAppliances              NA         NA      NA
categoryOffice Supplies:sub_categoryAppliances 1.680e+02  1.382e+01  12.155
categoryTechnology:sub_categoryAppliances             NA         NA      NA
categoryFurniture:sub_categoryArt                     NA         NA      NA
categoryOffice Supplies:sub_categoryArt        1.604e+02  1.296e+01  12.370
categoryTechnology:sub_categoryArt                    NA         NA      NA
categoryFurniture:sub_categoryBinders                 NA         NA      NA
categoryOffice Supplies:sub_categoryBinders    2.096e+02  1.185e+01  17.691
categoryTechnology:sub_categoryBinders                NA         NA      NA
categoryFurniture:sub_categoryBookcases        6.709e+01  1.624e+01   4.132
categoryOffice Supplies:sub_categoryBookcases         NA         NA      NA
categoryTechnology:sub_categoryBookcases              NA         NA      NA
```

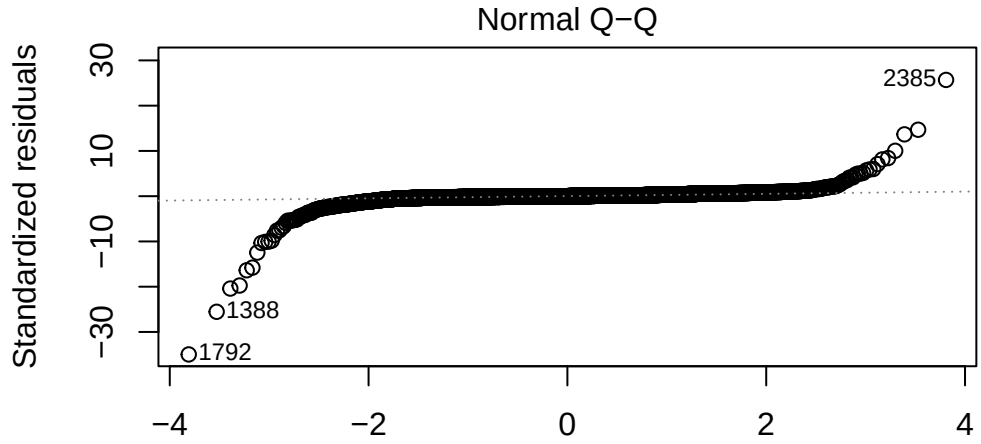
categoryFurniture:sub_categoryChairs	1.075e+02	1.311e+01	8.195
categoryOffice Supplies:sub_categoryChairs	NA	NA	NA
categoryTechnology:sub_categoryChairs	NA	NA	NA
categoryFurniture:sub_categoryCopiers	NA	NA	NA
categoryOffice Supplies:sub_categoryCopiers	NA	NA	NA
categoryTechnology:sub_categoryCopiers	5.293e+02	2.648e+01	19.990
categoryFurniture:sub_categoryEnvelopes	NA	NA	NA
categoryOffice Supplies:sub_categoryEnvelopes	1.720e+02	1.609e+01	10.694
categoryTechnology:sub_categoryEnvelopes	NA	NA	NA
categoryFurniture:sub_categoryFasteners	NA	NA	NA
categoryOffice Supplies:sub_categoryFasteners	1.639e+02	1.712e+01	9.573
categoryTechnology:sub_categoryFasteners	NA	NA	NA
categoryFurniture:sub_categoryFurnishings	1.682e+02	1.254e+01	13.413
categoryOffice Supplies:sub_categoryFurnishings	NA	NA	NA
categoryTechnology:sub_categoryFurnishings	NA	NA	NA
categoryFurniture:sub_categoryLabels	NA	NA	NA
categoryOffice Supplies:sub_categoryLabels	1.650e+02	1.488e+01	11.085
categoryTechnology:sub_categoryLabels	NA	NA	NA
categoryFurniture:sub_categoryMachines	NA	NA	NA
categoryOffice Supplies:sub_categoryMachines	NA	NA	NA
categoryTechnology:sub_categoryMachines	4.563e+01	2.147e+01	2.125
categoryFurniture:sub_categoryPaper	NA	NA	NA
categoryOffice Supplies:sub_categoryPaper	1.721e+02	1.215e+01	14.166
categoryTechnology:sub_categoryPaper	NA	NA	NA
categoryFurniture:sub_categoryPhones	NA	NA	NA
categoryOffice Supplies:sub_categoryPhones	NA	NA	NA
categoryTechnology:sub_categoryPhones	1.481e+02	1.250e+01	11.848
categoryFurniture:sub_categoryStorage	NA	NA	NA
categoryOffice Supplies:sub_categoryStorage	1.283e+02	1.271e+01	10.095
categoryTechnology:sub_categoryStorage	NA	NA	NA
categoryFurniture:sub_categorySupplies	NA	NA	NA
categoryOffice Supplies:sub_categorySupplies	1.043e+02	1.744e+01	5.979
categoryTechnology:sub_categorySupplies	NA	NA	NA
categoryFurniture:sub_categoryTables	NA	NA	NA
categoryOffice Supplies:sub_categoryTables	NA	NA	NA
categoryTechnology:sub_categoryTables	NA	NA	NA
ship_modeFirst Class:segmentConsumer	-2.913e+00	9.060e+00	-0.322
ship_modeSame Day:segmentConsumer	5.570e+00	1.257e+01	0.443
ship_modeSecond Class:segmentConsumer	-8.539e+00	8.421e+00	-1.014
ship_modeStandard Class:segmentConsumer	-1.803e+00	6.814e+00	-0.265
ship_modeFirst Class:segmentCorporate	2.883e-01	1.041e+01	0.028
ship_modeSame Day:segmentCorporate	-7.259e+01	1.963e+01	-3.698
ship_modeSecond Class:segmentCorporate	-6.444e+00	9.566e+00	-0.674

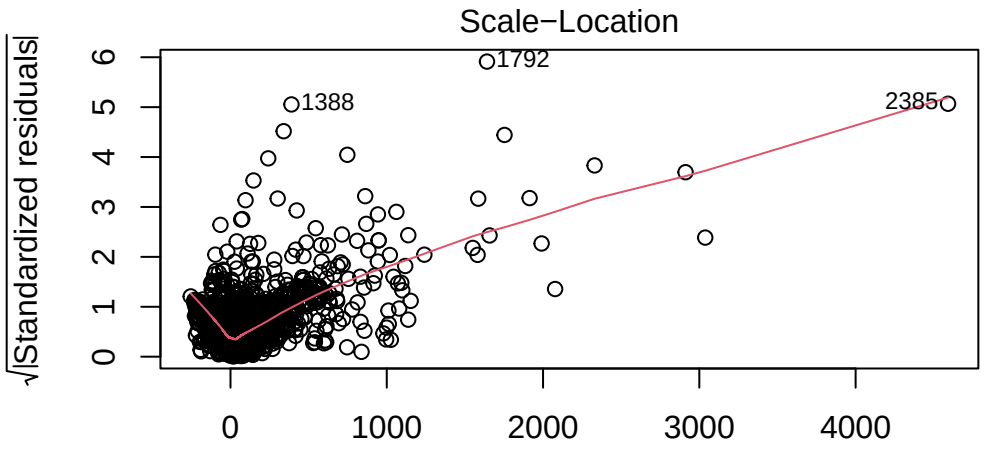
```
ship_modeStandard Class:segmentCorporate         3.200e+00  7.438e+00   0.430
ship_modeFirst Class:segmentHome Office          4.374e+00  1.301e+01   0.336
ship_modeSame Day:segmentHome Office             1.012e+00  1.822e+01   0.056
ship_modeSecond Class:segmentHome Office         2.970e+00  1.186e+01   0.250
ship_modeStandard Class:segmentHome Office              NA         NA      NA
                                                Pr(>|t|)
(Intercept)                                      < 2e-16 ***
sales                                            < 2e-16 ***
quantity                                        4.52e-08 ***
discount                                         < 2e-16 ***
categoryFurniture:sub_categoryAccessories             NA
categoryOffice Supplies:sub_categoryAccessories       NA
categoryTechnology:sub_categoryAccessories       < 2e-16 ***
categoryFurniture:sub_categoryAppliances              NA
categoryOffice Supplies:sub_categoryAppliances   < 2e-16 ***
categoryTechnology:sub_categoryAppliances             NA
categoryFurniture:sub_categoryArt                     NA
categoryOffice Supplies:sub_categoryArt          < 2e-16 ***
categoryTechnology:sub_categoryArt                    NA
categoryFurniture:sub_categoryBinders                 NA
categoryOffice Supplies:sub_categoryBinders      < 2e-16 ***
categoryTechnology:sub_categoryBinders                NA
categoryFurniture:sub_categoryBookcases         3.64e-05 ***
categoryOffice Supplies:sub_categoryBookcases         NA
categoryTechnology:sub_categoryBookcases              NA
categoryFurniture:sub_categoryChairs            2.95e-16 ***
categoryOffice Supplies:sub_categoryChairs            NA
categoryTechnology:sub_categoryChairs                 NA
categoryFurniture:sub_categoryCopiers                 NA
categoryOffice Supplies:sub_categoryCopiers           NA
categoryTechnology:sub_categoryCopiers           < 2e-16 ***
categoryFurniture:sub_categoryEnvelopes               NA
categoryOffice Supplies:sub_categoryEnvelopes    < 2e-16 ***
categoryTechnology:sub_categoryEnvelopes              NA
categoryFurniture:sub_categoryFasteners               NA
categoryOffice Supplies:sub_categoryFasteners    < 2e-16 ***
categoryTechnology:sub_categoryFasteners              NA
categoryFurniture:sub_categoryFurnishings        < 2e-16 ***
categoryOffice Supplies:sub_categoryFurnishings       NA
categoryTechnology:sub_categoryFurnishings            NA
categoryFurniture:sub_categoryLabels                  NA
categoryOffice Supplies:sub_categoryLabels       < 2e-16 ***
categoryTechnology:sub_categoryLabels                 NA
```

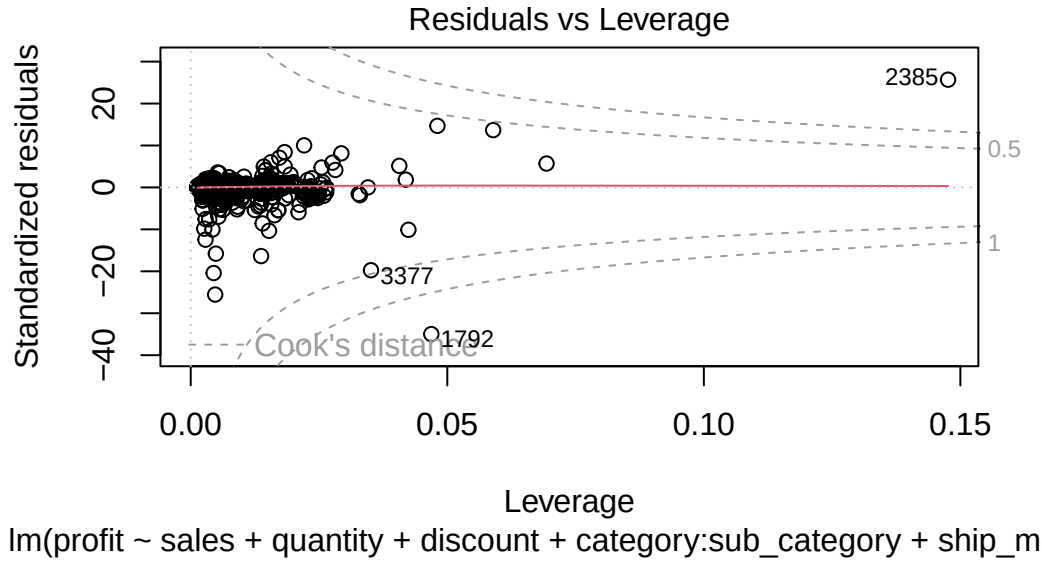
```
categoryFurniture:sub_categoryMachines               NA
categoryOffice Supplies:sub_categoryMachines         NA
categoryTechnology:sub_categoryMachines         0.033623 *
categoryFurniture:sub_categoryPaper                  NA
categoryOffice Supplies:sub_categoryPaper        < 2e-16 ***
categoryTechnology:sub_categoryPaper                 NA
categoryFurniture:sub_categoryPhones                 NA
categoryOffice Supplies:sub_categoryPhones           NA
categoryTechnology:sub_categoryPhones            < 2e-16 ***
categoryFurniture:sub_categoryStorage                NA
categoryOffice Supplies:sub_categoryStorage      < 2e-16 ***
categoryTechnology:sub_categoryStorage               NA
categoryFurniture:sub_categorySupplies               NA
categoryOffice Supplies:sub_categorySupplies    2.35e-09 ***
categoryTechnology:sub_categorySupplies              NA
categoryFurniture:sub_categoryTables                 NA
categoryOffice Supplies:sub_categoryTables           NA
categoryTechnology:sub_categoryTables                NA
ship_modeFirst Class:segmentConsumer            0.747799
ship_modeSame Day:segmentConsumer               0.657798
ship_modeSecond Class:segmentConsumer           0.310589
ship_modeStandard Class:segmentConsumer         0.791294
ship_modeFirst Class:segmentCorporate           0.977913
ship_modeSame Day:segmentCorporate              0.000219 ***
ship_modeSecond Class:segmentCorporate          0.500566
ship_modeStandard Class:segmentCorporate        0.667037
ship_modeFirst Class:segmentHome Office         0.736739
ship_modeSame Day:segmentHome Office            0.955705
ship_modeSecond Class:segmentHome Office        0.802300
ship_modeStandard Class:segmentHome Office           NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 160.6 on 7142 degrees of freedom
Multiple R-squared:  0.4807,    Adjusted R-squared:  0.4785
F-statistic: 220.3 on 30 and 7142 DF,  p-value: < 2.2e-16
```

```r
plot(fit3)
```


Residuals vs Fitted
Residuals
4000
0
−6000
2385
1388
1792
0
1000
2000
3000
4000
Fitted values
lm(profit ~ sales + quantity + discount + category:sub_category + ship_mo


Normal Q−Q
Standardized residuals
30
10
−10
−30
2385
1388
1792
−4
−2
0
2
4
Theoretical Quantiles
lm(profit ~ sales + quantity + discount + category:sub_category + ship_mo





Comments: In the interaction between ship_mode and segment, we see only one statistically significant grouping. These are the only outliers, as the rest appear to correlate strongly. The multiple r-squared of .48 isn't too great here, especially because we are working with multiple variables which should push the number up. It's also interesting that quantity negatively impacts profit so much.

Now we are going to use a simpler model to predict on the test set. Use a model that employs `sales`, `quantity`, and `discount` as the predictors. Fit the model and then use the `predict()`

function to predict new values on the test set. Unfortunately, we don't know the truth, so we can't compare our predictions.

```
fit4 <- lm(profit ~ sales + quantity + discount, data = train)
fit4
```

```
Call:
lm(formula = profit ~ sales + quantity + discount, data = train)

Coefficients:
(Intercept)        sales     quantity     discount
    25.1048       0.2339      -4.2897    -205.2987
```

```
summary(fit4)
```

```
Call:
lm(formula = profit ~ sales + quantity + discount, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-5616.5   -14.6     1.5    29.1  4303.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.510e+01  4.136e+00   6.070 1.34e-09 ***
sales        2.339e-01  3.431e-03  68.163  < 2e-16 ***
quantity    -4.290e+00  9.112e-01  -4.708 2.55e-06 ***
discount    -2.053e+02  9.941e+00 -20.652  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 168.9 on 7169 degrees of freedom
Multiple R-squared:  0.4233,    Adjusted R-squared:  0.423
F-statistic:  1754 on 3 and 7169 DF,  p-value: < 2.2e-16
```

```
predict(fit4, data.frame(sales = 1000, quantity = 10, discount = .2))
```

```
       1
175.0454
```

```
predict(fit4, data.frame(sales = 500, quantity = 1, discount = .1))
```

```
      1
117.234
```