HW9: John Andrus

Code

Hide

```
library(tidyverse)
library(magrittr)
library(lmtest)
library(sandwich)
```

0. Rename the variables that you are going to use to something sensible – variable names that have both periods and capital letters are not sensible. :fire: Better would be, for example changing Metrics.Sales to just sales.

Hide

```
vg <- read_csv("./video_games.csv") %>%
  rename(title = 'Title',
    handheld = 'Features.Handheld?',
    genre = 'Metadata.Genres',
    sales = 'Metrics.Sales',
    score = 'Metrics.Review Score',
    rating = 'Release.Rating',
    year = 'Release.Year',
    comptime = 'Length.Completionists.Average') %>%
    select(
      title, handheld, genre, sales, score, rating, year, comptime)
```

```
[36m──[39m [1m[1mColumn specification[1m[22m [36m────────────────────────────────────────────────────────
────────────────────────────────────────────────────────────[39m
cols(
  .default = col_double(),
  Title = [31mcol_character()[39m,
  `Features.Handheld?` = [33mcol_logical()[39m,
  `Features.Multiplatform?` = [33mcol_logical()[39m,
  `Features.Online?` = [33mcol_logical()[39m,
  Metadata.Genres = [31mcol_character()[39m,
  `Metadata.Licensed?` = [33mcol_logical()[39m,
  Metadata.Publishers = [31mcol_character()[39m,
  `Metadata.Sequel?` = [33mcol_logical()[39m,
  Release.Console = [31mcol_character()[39m,
  Release.Rating = [31mcol_character()[39m,
  `Release.Re-release?` = [33mcol_logical()[39m
)
[36mi[39m Use [38;5;235m[48;5;253m[38;5;235m[48;5;253m`spec()`[48;5;253m[38;5;235m[49
m[39m for the full column specifications.
```

Hide

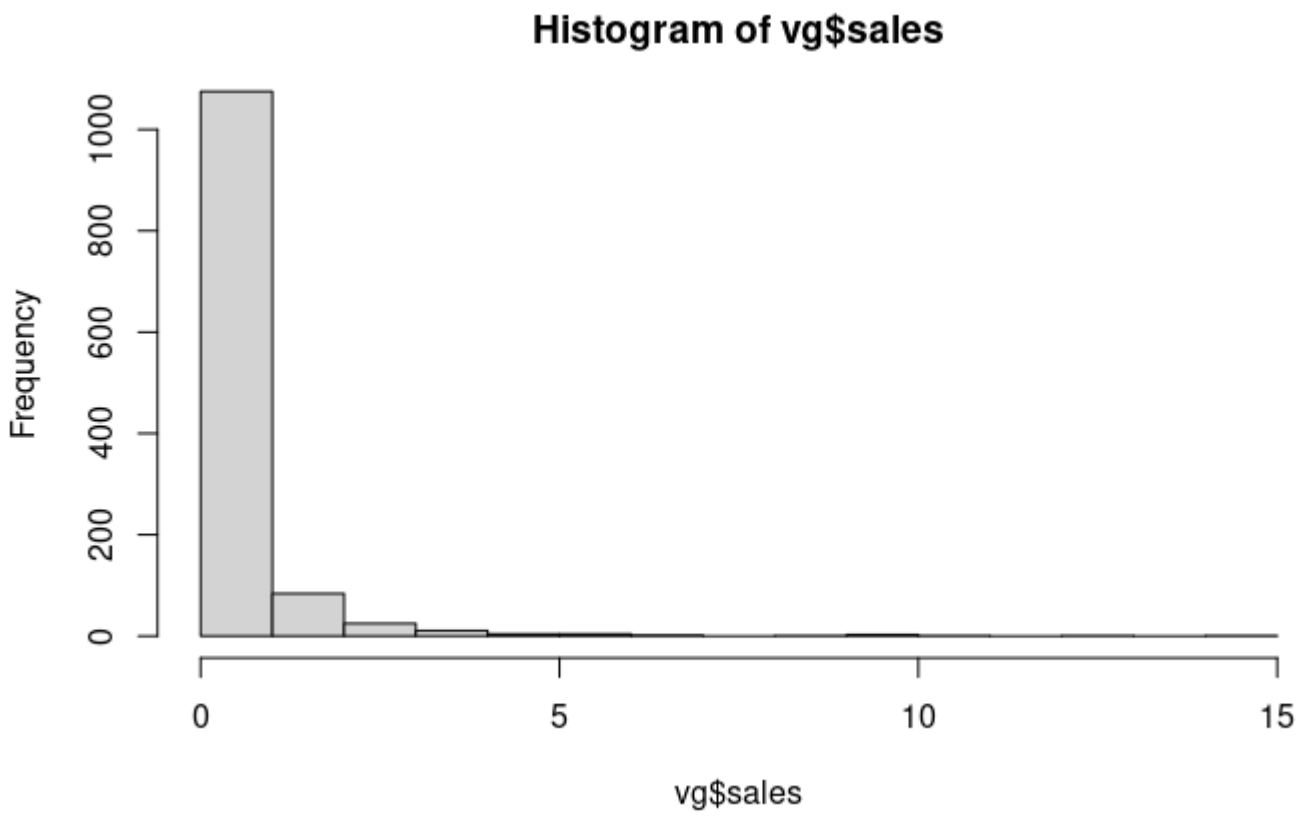
```
glimpse(vg)
```

```
Rows: 1,212
Columns: 8
$ title    [3m[90m<chr>[39m[23m "Super Mario 64 DS", "Lumines: Puzzle Fusion", "WarioWare To
uched!", "Hot Shots Golf: Open Tee", "S…
$ handheld [3m[90m<lgl>[39m[23m TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE,
TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRU…
$ genre    [3m[90m<chr>[39m[23m "Action", "Strategy", "Action,Racing / Driving,Sports", "Spo
rts", "Action", "Simulation", "Racing /…
$ sales    [3m[90m<dbl>[39m[23m 4.69, 0.56, 0.54, 0.49, 0.45, 0.41, 0.36, 0.34, 0.25, 0.22,
0.20, 0.16, 0.15, 0.14, 0.13, 0.12, 0.1…
$ score    [3m[90m<dbl>[39m[23m 85, 89, 81, 81, 61, 67, 88, 75, 68, 46, 62, 75, 63, 74, 51,
73, 60, 74, 72, 48, 66, 76, 91, 83, 77,…
$ rating   [3m[90m<chr>[39m[23m "E", "E", "E", "E", "E", "M", "E", "E", "T", "T", "E", "T",
"E", "E", "E", "E", "E", "T", "T", "E",…
$ year     [3m[90m<dbl>[39m[23m 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004,
2004, 2004, 2004, 2004, 2004, 2004, 2004, 200…
$ comptime [3m[90m<dbl>[39m[23m 29.766667, 0.000000, 10.000000, 0.000000, 72.566667, 30.0333
33, 1.250000, 80.000000, 0.000000, 12.0…
```

1. Examining the data, and using your background knowledge, evaluate the assumptions of the large-sample linear model.

Hide

```
hist(vg$sales)
```



The sample is quite large, but the data is very skewed which can cause problems when applying the Central Limit Theorem. More specifically, we may not trust the accuracy of our standard error calculation.

2. Whether you consider the large-sample linear model sufficiently valid or not, proceed to fit the linear model using lm().

Hide

```
lslm <- lm(sales ~ score + comptime, data = vg)
print(lslm)
```

```
Call:
lm(formula = sales ~ score + comptime, data = vg)

Coefficients:
(Intercept)        score     comptime
   -1.09197      0.02239      0.00273
```

3. Examine the coefficient for Metrics.Review.Score and give an interpretation of what it means.

Hide

```
coeftest(lslm, vcov = vcovHC)
```

```
t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) -1.0919732  0.1806027 -6.0463 1.972e-09 ***
score        0.0223899  0.0028686  7.8053 1.278e-14 ***
comptime     0.0027297  0.0011670  2.3391   0.01949 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This test shows that there is a positive correlation between a game’s scores and its sales and pratically no relationship between the game’s completion time and its sales. Additionally, the intercept of this line is negative which does not have practical meaning given that sales can not be less than zero. This is because all games take at least some amount of time to play ant the intercept coefficient attempts to estimate sales when the completion time of a game is equal to zero, resulting in a meaningless negative value.

4. Perform a hypothesis test to assess whether video game quality has a relationship with total sales. Please use vcovHC from the sandwich package with the default options (“HC3”) to compute robust standard errors. To conduct the test, use coeftest from the lmtest package.

Hide

```
coeftest(lslm, vcov = vcovHC)
```

```
t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) -1.0919732  0.1806027 -6.0463 1.972e-09 ***
score        0.0223899  0.0028686  7.8053 1.278e-14 ***
comptime     0.0027297  0.0011670  2.3391   0.01949 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In order to reject the null hypothesis that there is no relationship between quality and sales, we would need to see a P value less than 0.05. In this case the P values are much smaller than 0.05, so we can reject the null.

5. How many more copies of a game are sold when a game is one standard-deviation higher than the mean review, vs. when it is one standard-deviation lower than the mean review, holding all else equal? Answer this in two different ways:

A. Compute the standard deviation of the review score, and multiply the appropriate model coefficient by two-times this standard deviation.

Standard deviation was computed in part 4. It is predicted that an increase in one standard deviation of review will correspond to selling approximately 0.58 more games.

B. Use the predict function with the model that you have estimated. You can read the documentation for predict.lm which is the predict method for linear model objects (the type that you have fit here). Include a data frame (that has the same variable names as the data frame that you fitted the model against) in the newdata argument to predict. This data frame should have two rows and two columns. The column for the reviews should change from μ−σ to μ+σ; the column for the play time should be set to a constant, sensible level (perhaps the μ of this variable).

Hide

```
Warning messages:
1: In readChar(file, size, TRUE) : truncating string with embedded nuls
2: In readChar(file, size, TRUE) : truncating string with embedded nuls
3: In readChar(file, size, TRUE) : truncating string with embedded nuls
4: In readChar(file, size, TRUE) : truncating string with embedded nuls
5: In readChar(file, size, TRUE) : truncating string with embedded nuls
6: In readChar(file, size, TRUE) : truncating string with embedded nuls
7: In readChar(file, size, TRUE) : truncating string with embedded nuls
```

Hide

```
predict.lm(coeftest(lslm, vcov = vcovHC))
```

```
Error: $ operator is invalid for atomic vectors
```