SM339 ▪ Applied Statistics ▪ Spring 2023 ▪ Uhan

Exam 1 — Part 2 — 2/17/2023

Instructions

This part is worth 40 points total. The exam (both parts) is worth 100 points total.

You have until the end of the class period to complete this part of the exam.

You may use your plebe-issue TI-36X Pro calculator.

You may refer to notes that *you have handwritten*, not to exceed *one side* of an 8.5" × 11" piece of paper.

You may *not* use any other materials.

No applications except for JupyterLab may be open on your laptop during the exam.

No collaboration allowed. All work must be your own.

Do not discuss the contents of this exam with any midshipmen until it is returned to you.

Type your answers **directly in this Jupyter notebook**, and submit this notebook (just the `ipynb` file) using the submission form on the course website.

Problem 1

a.

For $X \sim N(\mu = 3, \sigma^2 = 4)$, calculate $P(2.1 \leq X < 3.7)$.

> **Feedback.** Most of you had the right idea here. Note that the variance σ^2 of X is given, but the R functions `dnorm`, `pnorm`, and `qnorm` take the standard deviation `sd` as input. See Problem 4b from the Lesson 2 Exercises for a similar problem.

```
In [ ]:
```

b.

For $Y \sim t(df = 11)$, calculate the 25th quantile.

> **Feedback.** See Problem 1 from the Lesson 2 Exercises for similar problems.

```
In [ ]:
```

Problem 2

You've recently been hired by Jacobian Jewelers to analyze the diamond market. The data frame `Diamonds2` from the `Stat2Data` library contains several variables measured on 307 randomly selected diamonds, including $TotalPrice$, the price for each diamond.

Load the `Stat2Data` library and the `Diamonds2` data frame, and examine the first few rows by running the cell below.

In [1]:
```
library(Stat2Data)
data(Diamonds2)
head(Diamonds2)
```

A data.frame: 6 × 6

	Carat	Color	Clarity	Depth	PricePerCt	TotalPrice
	<dbl>	<fct>	<fct>	<dbl>	<dbl>	<dbl>
1	1.08	E	VS1	68.6	6693.3	7228.8
2	0.31	F	VVS1	61.9	3159.0	979.3
3	0.32	F	VVS1	60.8	3159.0	1010.9
4	0.33	D	IF	60.8	4758.8	1570.4
5	0.33	G	VVS1	61.5	2895.8	955.6
6	0.35	F	VS1	62.5	2457.0	860.0

a.

Compute the mean of the prices of the diamonds in the sample.

Feedback. Most of you had the right idea here. See Lesson 2 for details on how to compute the mean of a column.

In [ ]:

b.

Create a boxplot of the prices of the diamonds in the sample.

Feedback. Most of you had the right idea here. See Lesson 2 for details on how to create a boxplot.

In [ ]:

Problem 3

A faculty member has supplemental retirement account (SRA) to invest money for retirement. The data frame `Retirement` from the `Stat2Data` library contains two variables: the contribution (SRA) to that account in each $Year$. The data frame has 16 observations.

Load the `Retirement` data frame and examine the first few rows by running the cell below.

In [2]:
```
data(Retirement)
head(Retirement)
```

A data.frame: 6 × 2

	Year	SRA
	<int>	<dbl>
1	1997	787.08
2	1998	968.16
3	1999	1975.08
4	2000	3990.00
5	2001	5455.80
6	2002	6338.60

a.

Fit a simple linear regression model for predicting the annual contribution (SRA) using $Year$.

Provide **only** the summary output for this part.

Feedback. Most of you had the right idea here.

- Be careful with identifying the response variable and the explanatory variable! Note that you're asked in this problem to predict SRA.
- Note that if you use `data = ...` in `lm()`, you **don't** need to specify the data frame again in your regression formula. For example:

```
lm(Y ~ X, data = my.data.frame)
```

  is the same as

```
lm(my.data.frame$Y ~ my.data.frame$X)
```

In [ ]:

b.

Create a plot of the standardized residuals, leverage, and Cook's distance.

You will be asked to comment on this plot in the next part.

Feedback. Most of you had the right idea here. See Lesson 9 for details on how to create the relevant diagnostic plot.

In [ ]:

c.

Using the rules of thumb we covered in class, comment on how "unusual" point 15 is in terms of its standardized residual, leverage, and Cook's distance. For each of these three measures, briefly explain.

Feedback. See Example 2 in Lesson 9 for a similar problem.

Write your answer here. Double-click to edit.

d.

Write R code to display **only** the 7th and 15th rows of the data frame `Retirement`.

Feedback. See Lesson 9 for details on how to keep or delete certain rows from a data frame in R.

In [ ]:

```
```

Problem 4

Simplexville Showers wants to better understand the market for shower heads, and has hired you as their consultant.

In the same folder as this notebook, there is a CSV file `data/Shower.csv` containing the $Price$ (in dollars) and average $Rating$ for 74 shower heads, randomly sampled from the online retailer, Jungle.com.

Load the CSV file into a data frame called `Shower` by running the cell below.

In [3]:

```
Shower <- read.csv('data/Shower.csv')
```

a.

Fit a simple linear regression model for predicting $Price$ using $Rating$.

Provide **only** the summary output for this part.

Feedback. Most of you had the right idea here.

- Be careful with identifying the response variable and the explanatory variable! Note that you're asked in this problem to predict Price.
- Note that if you use `data = ...` in `lm()`, you **don't** need to specify the data frame again in your regression formula. For example:

```
            lm(Y ~ X, data = my.data.frame)
```

  is the same as

```
            lm(my.data.frame$Y ~ my.data.frame$X)
```

In [ ]:

```
```

b.

Is the **constant variance** condition for simple linear regression met?

Using the code cell below, create a **single** diagnostic plot that will help you answer this question.

In the Markdown cell below, comment on whether the constant variance condition is met, based on your diagnostic plot.

Feedback for parts b and c. Most of you created a plot of residuals vs. fitted values. Some notes on using this plot:

- Recall that in class, we said that an ideal residual versus fitted values plot looks like **white noise**, that is, there is no discernible pattern in the residuals. In particular:
    - The points span a constant vertical width, moving from left to right (e.g., no blatant fanning) — this indicates that **constant variance** is met.
    - The points are evenly distributed above and below the residuals = 0 (zero horizontal) line, moving from left to right, without any discernible pattern — this indicates that **linearity** is met.

In [ ]:

Write your answer here. Double-click to edit.

c.

Is the **linearity** condition for simple linear regression met?

Using the code cell below, create a **single** diagnostic plot that will help you answer this question.

In the Markdown cell below, comment on whether the linearity condition is met, based on your diagnostic plot.

In [ ]:

Write your answer here. Double-click to edit.

d.

Is the **normality** condition for simple linear regression met?

Using the code cell below, create a **single** diagnostic plot that will help you answer this question.

In the Markdown cell below, comment on whether the normality condition is met, based on your diagnostic plot.

Feedback. Most of you created a Normal Q-Q plot of the residuals. Recall that in class, this ideally should look like a straight line.

In [ ]:

Write your answer here. Double-click to edit.

Grading rubric

Problem	Weight
1a	0.2
1b	0.2
2a	0.2
2b	0.2
3a	0.4
3b	0.4
3c	0.4
3d	0.4
4a	0.4

Problem	Weight
4b	0.4
4c	0.4
4d	0.4
Max Score	**40**