Exam 2 — Part 2 — 4/4/2024

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 0

For this exam, we will use the `MLBStandings2016` dataset from the `Stat2Data` package. This dataset contains the standings and team statistics for all Major League Baseball teams for the 2016 season.

Run the cell below to load and preview the data.

In [1]:
```
options(repr.matrix.max.cols = 21)  # show all 21 columns
library(Stat2Data)
data(MLBStandings2016)
head(MLBStandings2016)
```

A data.frame: 6 × 21

	Team	League	Wins	Losses	WinPct	BattingAverage	Runs	Hits	HR	Doubles	Triples	RBI	SB	OB
	<fct>	<fct>	<int>	<int>	<dbl>	<dbl>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<db
1	Arizona Diamondbacks	NL	69	93	0.426	0.261	752	1479	190	285	56	709	137	0.3
2	Atlanta Braves	NL	68	93	0.422	0.255	649	1404	122	295	27	615	75	0.3
3	Baltimore Orioles	AL	89	73	0.549	0.256	744	1413	253	265	6	710	19	0.3
4	Boston Red Sox	AL	93	69	0.574	0.282	878	1598	208	343	25	836	83	0.3
5	Chicago Cubs	NL	103	58	0.640	0.256	808	1409	199	293	30	767	66	0.3
6	Chicago White Sox	AL	78	84	0.481	0.257	686	1428	168	277	33	656	77	0.3

Problem 1

For this problem, we will focus on the following variables in `MLBStandings2016`:

Variable	Description
WinPct	Proportion of games won
ERA	Earned run average (earned runs allowed per 9 innings)
Runs	Number of runs scored
RBI	Number of runs batted in
WHIP	Number of walks and hits per inning pitched
HitsAllowed	Number of hits against the team

a.

Consider the following model, which we will call Model A:

$$(\text{Model A}) \qquad WinPct = \beta_0 + \beta_1 ERA + \beta_2 Runs + \varepsilon \qquad \varepsilon \sim \text{iid } N(0, \sigma_\varepsilon^2)$$

Fit Model A. Provide **only** the summary output for this part.

In [ ]:

b.

Is the overall model effective?

- Answer yes or no.
- Report (1) the name of the hypothesis test, (2) the test statistic and (3) the p-value you used to make your decision. Use a significance level of 0.05.

Write your answer here. Double-click to edit.

Feedback. See Example 4 in Lesson 16 Part 1 for a similar example.

c.

Consider the following model, which we call Model B:

$$(\text{Model B}) \qquad WinPct = \beta_0 + \beta_1 ERA + \beta_2 Runs + \beta_3 RBI + \beta_4 WHIP + \beta_5 HitsAllowed + \varepsilon$$
$$\varepsilon \sim \text{iid } N(0, \sigma_\varepsilon^2)$$

Fit Model B. Provide **only** the summary output for this part.

In [ ]:

d.

Report the adjusted R^2 for Model A and Model B.

Without any formal hypothesis testing, which model do you think is "better"? Briefly explain. Your explanation should be in complete sentences, with correct spelling and grammar.

Write your answer here. Double-click to edit.

Feedback. See Example 2 in Lesson 17 for a similar example.

e.

Conduct a nested F-test to compare Model A and Model B.

In particular, in the cell below, compute the test statistic and p-value for this test. You are encouraged to use any appropriate "shortcut" functions in R.

You will be asked to interpret the results of the test in part f.

In [ ]:

Feedback. See Example 1 in Lesson 22 for a similar example. Make sure to look at part g of that example, which provides an R "shortcut" function for performing a nested F-test.

f.

Based on the nested F-test you performed in part e, which model is "better"? Report (1) the test statistic and (2) the p-value you used to make your decision. Use a significance level of 0.05.

Write your answer here. Double-click to edit.

Feedback. See Example 1 in Lesson 22 for a similar example.

Problem 2

For this problem, we will focus on the following variables in `MLBStandings2016`:

Variable	Description
WinPct	Proportion of games won
BattingAverage	Team batting average
Runs	Number of runs scored
Hits	Number of hits

Consider the following model:

$$WinPct = \beta_0 + \beta_1 BattingAverage + \beta_2 Runs + \beta_3 Hits + \varepsilon \qquad \varepsilon \sim \text{iid } N(0, \sigma_\varepsilon^2)$$

a.

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

In [ ]:

b.

Construct and interpret a 90% interval that predicts the winning percentage of a future team with a 0.275 batting average, 700 runs, and 1380 hits over the course of one season.

In [ ]:

Write your answer here. Double-click to edit.

> **Feedback.** See Example 4 in Lesson 17 for a similar example.

c.

Do you see any of the red flags for multicollinearity we discussed in class? Briefly explain. Your explanation should be in complete sentences, with correct spelling and grammar.

Write your answer here. Double-click to edit.

> **Feedback.** See Lesson 21 (and Examples 1 and 2 in that lesson) for the two red flags for multicollinearity that we discussed in class.

d.

Compute the VIFs of all the predictors in this model. Based on the VIFs, should we be worried about multicollinearity? Briefy explain. State any rules of thumb you use.

In [ ]:

Write your answer here. Double-click to edit.

> **Feedback.** See Lesson 21 (and Example 3 in that lesson) for how to compute the VIFs for each predictor in a multiple linear regression model and the rule of thumb for using them to detect multicollinearity.

Grading rubric

Problem	Weight
1a	0.4
1b	0.4
1c	0.4
1d	0.4
1e	0.4
1f	0.4
2a	0.4
2b	0.4
2c	0.4
2d	0.4
Max Score	**40**