BIOSTAT 705 - Homework 3

Austin Allen

March 29, 2024

```
## Warning: package 'tidyverse' was built under R version 4.2.3
## Warning: package 'ggplot2' was built under R version 4.2.3
## Warning: package 'tibble' was built under R version 4.2.3
## Warning: package 'tidyr' was built under R version 4.2.3
## Warning: package 'readr' was built under R version 4.2.3
## Warning: package 'purrr' was built under R version 4.2.3
## Warning: package 'dplyr' was built under R version 4.2.3
## Warning: package 'stringr' was built under R version 4.2.3
## Warning: package 'forcats' was built under R version 4.2.3
## Warning: package 'lubridate' was built under R version 4.2.3
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Problem 1.

An investigator is interested in whether Bayer aspirin (treatment "A"), Tylenol (acetaminophen) (treatment "B") or Aleve (naproxen) (treatment "C") works more quickly to relieve the pain of a common headache. She recruits n individuals with frequent headaches, randomly assigns them to one of the three pain killers, asks them to take the medication upon first signs of the headache, and to record the time until the pain is gone (Y).

Part (a): Using indicator (dummy) variables in regression and set up the model for the experiment above. Assume $n_i = 2; i = 1, 2, 3$ subjects are randomized to group i, where $n = n_1 + n_2 + n_3$.

Let $X_1 = \begin{cases} 1 = \text{Treatment A} \\ 0 = \text{Otherwise} \end{cases}$, $X_2 = \begin{cases} 1 = \text{Treatment B} \\ 0 = \text{Otherwise} \end{cases}$, and $Y = \text{Time to Relief}$. The model for this experiment can be written as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Part (b): Write the model in part a) in a matrix form as $Y = X\beta + \epsilon$. Identify the design-matrix X as well as the Y , β and ϵ vectors.

$$\overbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{bmatrix}}^{Y} = \overbrace{\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}}^{\text{X, Design Matrix}} \overbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}}^{\beta} + \overbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}}^{\epsilon}$$

Part (c): State the null hypothesis of no difference among the means of the three groups.

$$H_0 : \beta_1 = \beta_2 = 0$$

Problem 2.

For the experiment in the problem above, express the regression model as "cell-mean" model, that is $Y_{ij} = \mu_i + \epsilon_{ij}$ where μ_i is the mean of the ith treatment group.

Part (a): Write the "cell-mean" model above in a matrix form as $Y = X\mu + \epsilon$. Identify the design-matrix X and the parameters vector μ for the "cell-mean" model.

$$\overbrace{\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix}}^{Y} = \overbrace{\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}}^{\text{X, Design Matrix}} \overbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}}^{\mu} + \overbrace{\begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{bmatrix}}^{\epsilon}$$

Part (b): State the null hypothesis of no difference among the means of the three groups.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

Problem 3.

A rehabilitation center researcher was interested in examining the relationship between physical fitness level prior to surgery of persons undergoing corrective knee surgery and time required in physical therapy until successful rehabilitation (in days). Patient records from the rehabilitation center were examined, and 24 male subjects ranging in age from 18 to 30 years who had undergone similar corrective knee surgery during the past year were selected for the study. The number of days required for successful recovery and fitness level (1=below average, 2=average and 3=above average) were recorded. The patient's age at the time of surgery was also recorded because previous studies have shown that younger patients tend to recover more quickly. The dataset can be found in HW3-prob3.dat.

Part (a): Fit the ANCOVA model. Based on this model fit, is there evidence that time to recovery differs across fitness levels after accounting for a patient's age? Please provide the hypotheses, test statistic, and p-value of the test used to address this question.

Part (b): Using the model fit in part (a), provide the estimated regression model for each fitness level group.

Part (c): Estimate the overall slope between days to recovery and patient age by removing fitness level from the model fit in part (a). How does this slope estimate compare to the slope estimates reported in part (b)?

Part (d): Compute the unadjusted mean days to recovery for each fitness level group. Using this numerical output, what factor level differences appear to be driving the global signal from fitness level if one exists?

Part (e): Compute the mean days to recovery for each fitness level adjusted for average patient age. Using this numerical output, what factor level difference appear to be driving the global signal from fitness level if one exists? Does your conclusion differ from the conclusion you reached in part (d)? If so, why does the discrepancy exist?

Part (f): Was it beneficial to include patient age in this analysis when the primary goal was to assess the relationship between days to recovery and fitness level? Explain your reasoning.

Problem 4.

A 2year experiment was conducted to study the effectiveness of stannous fuoride (SF) and acid-phosphate fluoride (APF) in dental caries reduction. A total of 69 female children completed the study, 22 of them were treated with SF, 27 were treated with APF, and the remaining 20 were a control group treated with distilled water (W). At the beginning and the end of the study, the number of decayed, missing or filled teeth (DMFT) was measured for each child. Researchers were interested in examining the effect of treatment on the difference in DMFT before and after the study period. The data for this study can be found in HW3-prob4.csv.

Data Dictionary for FLUORIDE:

	Variable Name	Description
(1)	SUBJNO	Subject identifier
(2)	INST	Institution where the subject was treated (1; 2; 3)
(3)	AGE	Age of patient (years)
(4)	BEFORE	Number of DMFT at beginning of study period
(5)	AFTER	Number of DMFT at end of study period
(6)	TRT	Treatment group (SF vs. APF vs. W)
(7)	DIFF	Difference between after and before DMFT values
(8)	X1	Binary indicator for TRT = SF
(9)	X2	Binary indicator for TRT = APF
(10)	X3	Binary indicator for TRT = W

Part (a): Check the ANCOVA model assumptions before fitting a 1-way ANCOVA model: `diff = trt + age`. Plot age (covariate) vs. DIFF by TRT to Visualize the data for any interaction or confounding.

Part (b): Based on your answers in part (a), which ANCOVA assumption(s) have been violated (if any)?

Problem 5.

Suppose researchers performed an observational study to examine the impact of two treatments (A vs. B) on LDL levels. Because the study was observational, the researchers did not randomly assign patients to treatment group meaning that the treatment comparison could be confounded between the two groups. The data for this study can be found in HW3-prob5.csv.

Data Dictionary for LDL:

	Variable Name	Description
(1)	RISK	Baseline continuous covariate
(2)	LDL	LDL value { Outcome of interest
(3)	TRT	Treatment Group (A or B)
(4)	X1	Indicator variable for Treatment Group A

Part (a): Check the ANCOVA model assumptions before fitting a 1-way ANCOVA model: `ldl = trt + risk`. Plot risk (covariate) vs. ldl by TRT to Visualize the data for any interaction or confounding.

Part (b): Based on your answers in part (a), which ANCOVA assumption(s) been violated (if any)?