MATH 530/630

Integrative Lab 2 - Linear Regression

Contents

Overview 1

Logistics 2

The Data 2

Your Mission 3

Exploratory Data Analysis 3

Regression Modeling 4

Residual Analysis 5

Outlier Analysis 5

Sums of Squares 5

The NULL Model 6

Replicate a Plot 6

Report your process 7

Grading 7

Overview

The goal of this lab is to carefully, thoroughly, and thoughtfully conduct a linear regression analysis. You are also asked to communicate clearly about the steps in your analysis process with others, by sharing your R code, output, and narrative. As such, your code cannot "stand alone"- it is meant to complement / enhance / support your narrative. This lab will be due in two stages:

1. A complete knitted `html` file submitted on Sakai.
2. A follow-up self assessment.

Using the key, your self-assessment should include even **more** narrative; where you made mistakes, you must discuss and analyze where you went wrong, and correct them without copying/pasting directly from the key (this typically means that you need to include more narrative than we provide in the key). A good self-assessment will include:

- Assessment of the accuracy and completeness of your "initial solutions"
- Correct worked solutions with some discussion and analysis of why your initial solution was incorrect, and reflection on the source of your confusion (if you got an answer correct, this is not necessary)
- Attributions as appropriate to other students who helped you, or other sources such as lecture notes, readings, online resources, etc. that helped you

Logistics

You will use R Markdown to construct your analysis report. You'll submit your work as an html file knit from your `.Rmd` file (please leave the default code chunk options for `eval = TRUE` and `echo = TRUE`). Your lab should serve as your own personal cheatsheet in the future for regression analyses. Give yourself the cheatsheet you deserve!

For all things, code and narrative, if you're dissatisfied with a result, discuss the problem, what you've tried and move on (remember my 30-minute rule). You'll need this loaded at the top:

```r
library(tidyverse) # all the good stuff
library(readxl) # for reading in xlsx files
library(janitor) # for clean_names
library(knitr) # for kable
library(moderndive) # for getting tables
library(corrr) # for correlation matrix
library(skimr) # for skim
library(GGally) # for ggpairs
library(broom) # for pulling out model results
```

The Data

You will work with an open access dataset from a publication in PLOS ONE titled: *Vitamin D Status among Thai School Children and the Association with 1,25-Dihydroxyvitamin D and Parathyroid Hormone Levels*. The data is available as an excel file on Data Dryad, where you can download the `.xlsx` file.

```r
library(readxl)
path_to_xlsx <- here::here("data","Thai_vitamin D dataset.xlsx")
vitd <- read_xlsx(path_to_xlsx, sheet = 1)
codebook <- read_xlsx(path_to_xlsx, sheet = 2) #The 2nd sheet has the codebook
```

I recommend you use the `janitor::clean_names` function, because some of these variable names start with an underscore and therefore will always need to be referenced surrounded by backticks.

```r
library(janitor)
vitd <- vitd %>%
  clean_names()
glimpse(vitd)
```

```
Observations: 537
Variables: 19
$ id          <dbl> 101, 102, 104, 105, 106, 107, 108, 110, 112, 113, ...
$ sex         <chr> "F", "F", "F", "F", "M", "M", "M", "F", "F", "F", ...
$ ageyears    <dbl> 7.749487, 7.040383, 7.289528, 7.251198, 6.845995, ...
$ height      <dbl> 121.80, 120.40, 117.00, 115.60, 113.30, 121.60, 12...
$ weight      <dbl> 20.0, 18.1, 19.6, 19.3, 17.0, 23.8, 24.1, 23.1, 22...
$ bmi         <dbl> 13.48141, 12.48606, 14.31807, 14.44248, 13.24308, ...
$ zwfa        <dbl> -1.29, -1.47, -1.09, -1.17, -2.18, -0.28, -0.07, -...
$ underweight <chr> "not underweight", "not underweight", "not underwe...
$ zbfa        <dbl> -1.45, -2.24, -0.75, -0.65, -1.88, 0.28, -0.48, -1...
$ wasted      <chr> "not wasted", "wasted", "not wasted", "not wasted"...
$ zhfa        <dbl> -0.58, -0.12, -0.98, -1.20, -1.45, -0.77, 0.34, -0...
$ stunted     <chr> "not stunted", "not stunted", "not stunted", "not ...
$ period      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
$ ifyeswhen   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
$ school      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
$ baseline_ca <dbl> NA, NA, NA, 57.6876, NA, NA, 290.4152, 131.2554, N...
$ pth         <dbl> 66.7, 55.1, 29.5, 48.9, 20.7, 23.9, 23.4, 24.7, 33...
$ x25d        <dbl> 72.2, 73.4, 64.6, 73.6, 88.5, 85.8, NA, 75.2, 82.9...
$ x125d       <dbl> 265, 144, 320, 245, 251, 206, 200, 205, 254, 223, ...
```

That is better! You are ready to start exploring the `vitd` data.

Your Mission

We'll conduct a linear regression analysis to examine the impact of age, and its interaction with sex, along with height-for-age Z score on serum 25(OH)D concentrations (a measure of vitamin D levels).

The relevant variables you will need are:

```r
lm_vars <- c("sex", "_zhfa", "ageyears", "_25D")
codebook %>%
  filter(Variable %in% lm_vars) %>%
  knitr::kable()
```

Variable

Description

Unit or coding where applicable

sex

Sex of child participant

F=female, M=male

ageyears

Age of child participant

YEARS

_zhfa

Height-for-age z-score

Standard deviation (SD)

_25D

Serum 25-hydroxyvitamin D concentrations

nmol/L

Please note that the variable names in the codebook won't exactly match those in your data if you use `janitor::clean_names()`! All `_` will be replaced by `x` so that you don't have to type backticks in your later R code.

Exploratory Data Analysis

Conduct a thorough EDA of the four variables defined above in the vitamin D dataset. Recall that a new exploratory data analysis involves three things:

- Looking at the raw values.

  - dplyr::glimpse()
- Computing summary statistics of the variables of interest.
  - skimr::skim()
  - corrr::correlate()
- Creating informative visualizations.
  - ggplot2::ggplot()
    - geom_histogram() or geom_density() for numeric continuous variables
    - geom_bar() or geom_col() for categorical variables
  - GGally::ggpairs()
    - Note that you can add transparency to points/density plots in the aes call, for example: aes(colour = sex, alpha = 0.7)

You may wish to have a level 1 header (#) for your EDA, then use level 2 sub-headers (##) to make sure you cover all three EDA bases. **At a minimum** you should answer these questions:

- How many variables/columns?
- How many rows/observations?
- Which variables are numbers?
- Which are categorical variables (numeric or character variables with variables that have a fixed and known set of possible values; aka factor variables)?
- Complete this sentence: "There is one row/observation per…"
- What are the correlations between variables? Does each scatterplot support a linear relationship between variables? Do any of the correlations appear to be conditional on the value of a categorical variable (like sex)?

At this stage, you may also find you want to use filter, mutate, arrange, select, or count. Let your questions lead you!

Regression Modeling

Fit a multiple regression model to predict serum 25(OH)D concentrations (x25d if you used janitor::clean_names(); _25D if you did not) and get the regression table. Your model should include:

- An intercept term
- A coefficient for age in years (ageyears)
- A coefficient for sex (sex)
- A coefficient for the interaction between age and sex (ageyears:sex)
- A coefficient for height-for-age z-scores (zhfa if you used janitor::clean_names(), _zhfa if not)

Interpret the output from the regression table (in complete sentences, but you may use bullet points to organize). You may wish to enhance the interpretability of your results by mean centering numerical predictor variables.

Some examples:

- Parallel slopes example here
- Interaction model here

The authors state: "Specifically, serum 25(OH)D concentrations were 19% higher in males at the mean age (9.9 years)." They also state: "females experienced a…4% decline in serum 25(OH)D levels for each increasing year of age; no decline was seen in male participants with increasing age". In your narrative, walk through how you would use the numbers in the regression table to arrive at these numbers- are they accurate based on your regression model output?

Residual Analysis

Examine the model residuals following using `get_regression_points(my_model)`. Perform a (raw) residual analysis first with a histogram, faceting by `sex`. Also look at the residuals as compared to the three predictor variables:

- x_1: numerical explanatory/predictor variable of `age`
- x_2: categorical explanatory/predictor variable of `sex`
- x_3: numerical explanatory/predictor variable of height-for-age z-scores (`zhfa` if you used `janitor::clean_names()`, `_zhfa` if not)

Explain (a) what you are looking for in the plots, and (b) what you see in the context of assessing how "well" the linear model fits the data.

Some examples:

- Parallel slopes example here
- Interaction model here

Outlier Analysis

Examine points with high leverage and discrepancy (use the externally studentized residuals as the index for discrepancy). Remember that these statistics are not in the output of `moderndive::get_regression_points()` function- you'll want to use `broom::augment()` and `rstudent` instead.

- How many observations would you expect to have high discrepancy in this sample?
  - You may want to account for how many females versus males you would expect
- Do you see any observations that have *both* high leverage and discrepancy?
- Do any points that are either high leverage or high discrepancy **also** have high influence on the regression estimates, as measured by Cook's distance?
- Would you exclude any observations from your model? Justify your answer either way.
- In general, would you say the overall model fit is better for females or males? Why or why not?

Sums of Squares

Fill in the blanks in the following code block to calculate the Residual, Model, and Total Sums of Squares:

```
vitd_ss <- d_diag %>%
  summarise(total_ss = sum((___ - mean(___))^2),
            resid_ss = sum((___ - ___)^2),
            model_ss = sum((___ - mean(___))^2))
vitd_ss
```

Using `dplyr`, show that:

- The total sums of squares is equal to the residual plus the model sums of squares

$$total\_ss = resid\_ss + model\_ss$$

- The total sums of squares divided by $(n-1)$ is equal to the variance of the `y` outcome variable (*hint:* you may need to look at how many observations actually contributed to the model- not the same as the original n due to missing values!)

$$var_y = \frac{total\_ss}{n-1}$$

- The R^2 value in your model output is the model sums of squares divided by the total sums of squares (*hint:* `broom::glance(my_model)`).

$$R^2 = \frac{model\_ss}{total\_ss}$$

The NULL Model

In class, I asserted that `lm` is by default comparing the model you specify in your `lm` call to a null model defined by using a line with an intercept, but slope = 0, which estimates the mean of y. Let's build an intercept-only linear regression model to prove this.

```
vitd_complete <- vitd %>%
  drop_na(x25d, sex, ageyears, zhfa)
int_mod <- lm(x25d ~ 1, data = vitd_complete)
get_regression_table(int_mod)
```

```
# A tibble: 1 x 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept     74.3     0.669      111.       0     73.0     75.6
```

What is the "estimate" equal to here (**hint:** look back at your `skim` output)?

```
anova(int_mod, dmod) # insert your model name from above instead of "dmod"
```

Use the code above to use the `anova` function to compare the intercept-only model to your linear regresssion model. Look very carefully at this output and answer these questions:

- What is the `RSS` for line 1 (corresponding to the intercept-only model, Model 1) equal to that you calculated above?
- What is the `RSS` for line 2 (corresponding to Model 2) equal to that you calculated above?
- What is the `Sum of Sq` equal to that you calculated above?
- In < 3 sentences, explain what it means to use `lm(y ~ x + z)` versus `lm(y ~ 1)`, and what happens "under the hood" here that you now see in the `anova` output.

Replicate a Plot

In the original paper, they presented this figure:

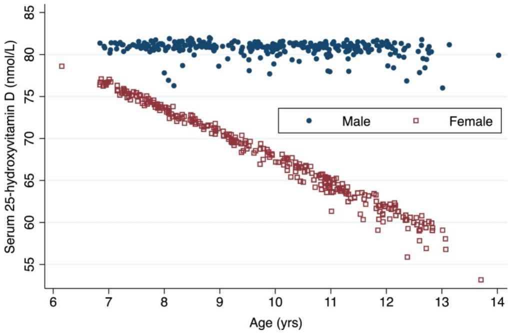

Figure 1 Relation between age in years and serum 25-hydroxyvitamin D (nmol/L) concentrations stratified by gender. Look at the published figure carefully:

- What appears to be the minimum value of serum 25D levels?
- What appears to be the max?
- Do those match your EDA?
- Recreate this figure in `ggplot2` using the observed data (ignore height-for-age z-scores at this point).
- Make the same plot where the y-axis should now be mapped to the fitted value for the outcome variable (using the full model with height-for-age z-scores in it). Discuss the differences you see between this plot and the previous. Were you able to recreate the published figure? Is it helpful in understanding the researchers' findings? Was it misleading at all?

Report your process

You're encouraged to reflect on what was hard/easy, problems you solved, helpful tutorials you read, etc. Give credit to your sources, whether it's a blog post, a fellow student, an online tutorial, etc.

Grading

This lab is worth 20 points total, scored as follows:

8 points for your initial submission being "in-good-faith":

- 8 (Strong attempt): narrative and code reflects strong independent problem solving, with clearly thought out attempts to approach the problems and a diligent and honest effort to find solutions.
- 4 (Adequate attempt): narrative and code reflects some attempt to approach the problems, but approach appears to be superficial and lacks depth of analysis.
- 0 (No attempt or incomplete): No submission, or didn't interpret anything but left it all to the "reader". Or more than one technical problem that is relatively easy to fix.

12 points for the quality of the final self-assessment:

- 12 (Exceptional): narrative is thorough, concise, and clearly demonstrates ability to analyze and interpret statistics as well as theoretical understanding of statistical concepts.
- 8 (Adequate): narrative addresses the questions with moderate inaccuracies in analysis and/or interpretation, or offers correct but incomplete solutions.
- 4 (Inadequate): narrative attempts to address questions with substantial inaccuracies in analysis and/or interpretation.
- 0 (Insufficient): narrative does not attempt to address questions or self-assessment is insufficient to grade