HW 05: County Health

Model Comparison/Evaluation + Outliers

> **!** Important
>
> Due: Friday, October 4th, 11:59pm

In this homework, you'll use simple and multiple linear regression to analyze the relationship between the number of doctors in a county, the number of beds, and the number of hospitals.

Learning goals

By the end of the homework you will be able to...

- compare models using R^2 and RMSE
- transform observations to improve model fit
- talk about outliers
- fit a model with two predictors

Getting started

- Go to RStudio and login with your College of Idaho Email and Password.
- Make a subfolder in your `hw` directory to store this homework.
- Log into Canvas, navigate to Homework 5 and upload the `hw-05.qmd` file into the folder your just made.

Packages

We'll use the following packages in this homework.

```
library(tidyverse)
library(broom)
library(yardstick)
library(ggformula)
library(knitr)
library(patchwork)
library(Stat2Data)
library(GGally)
# add more packages as needed
```

Data: County Health

The data set for this homework is from the `Stat2Data` R package which is the companion package for this course's textbook. It is the same data set that we used in AE-08. The data was originally generated by the American Medical Association and concerns the availability of health care in counties in the United States. You can find information here by searching for the County Health Resources dataset.

```
data("CountyHealth") # Loads the data from the package
```

It is relatively easy to count the number of hospitals a county has, whereas counting the number of doctors is much more difficult. We'd like to build a linear model to predict the number of doctors, contained in the variable `MDs`, from the number of hospitals, `Hospitals` and the number of beds, `Beds`.

Exercise 1

Describe was an observational unit represents for this data set. How many are there?

Exercise 2

In Example 1.7 of Stat2, they consider a simple linear model to predict the number of doctors (`MDs`) from the number of hospitals (`Hospitals`) in a metropolitan area. In that example, they found that a square root transformation on the response variable produced a more linear relationship. Create a new variable in the `CountyHealth` data frame called `sqrtMDs`. Hint: use the `sqrt` function inside the `mutate` function.

Exercise 3

Use the function `ggpairs` from the package `GGally` to generate a grid scatter plots and correlations. Note that you will need to select the variables you want to use. Which explanatory variable (`Hospitals` or `Beds`) has the highest correlation with `SqrtMDs`? Is this consistent with your visual assessment?

Exercise 4

Fit a simple linear model using `SqrtMDs` as the response variable and `Hospitals` as the predictor. You may use `sqrt(MDs)` in your `lm` call instead of `SqrtMDs` if you like. How much of the variability in the `SqrtMDs` values is explained by `Hospitals`? How much of the variability in `MDs` is explained by the model you just fit. To figure this out:

1. Augment your model.
2. Convert the fitted and observed response variables back to number of MDs rather than square-root of the number of MDs.
3. Compute the R^2.

Why are these two numbers different?

Exercise 5

Do you think taking the square root of `Beds` would improve this model? Support your argument with plots and/or numbers.

Exercise 6

Repeat exercise 4 above with `Beds` as the predictor instead of `Hospitals`.

Exercise 7

For the model you just fit, are there any high-influence outliers? Justify your answer using something from the lecture on outliers. There appear to be at least two high-leverage points. Which observations are these?

Exercise 8

Fit a multiple linear model using `SqrtMDs` are the response variable and both `Hospitals` and `Beds` as the predictors. Interpret both slopes and the intercept in the context of the problem.

Exercise 9

How much of the variation in `MDs` is explained by the model you just fit? Which model would you say is the "best", given what we've learned through the first lecture on multiple linear regression.

Exercise 10

Using the "best" model, predict the average number of doctors in a metro area with 1,000 beds and 4 hospitals. Report a 95% prediction interval and interpret your results in context.

Grading

Total points available: 50 points.

Component	Points
Ex 1	3
Ex 2	3
Ex 3	5
Ex 4	8
Ex 5	4
Ex 6	3
Ex 7	5
Ex 8	6
Ex 9	4
Ex 10	4
Workflow & formatting	5[1]

[1]The "Workflow & formatting" grade is to assess the reproducible workflow, clarity, and professionalism.