

# CM 2.1-2 - Simple linear regression

Math 530/630

## Contents

<b>Overview</b>	<b>1</b>
<b>The data</b>	<b>1</b>
<b>Basics</b>	<b>2</b>
<b>EDA</b>	<b>3</b>
Look at the data . . . . .	3
Look at summary statistics . . . . .	3
Visualize the data . . . . .	3
<b>Simple linear regression: (see here)</b>	<b>4</b>
<b>Observed/fitted values: (see here)</b>	<b>4</b>
<b>Residual analysis: (see here)</b>	<b>5</b>

## Overview

- A complete knitted `html` file is due on Sakai by the beginning of the next class.
- This lab is based on Chapter 5: Basic Regression in ModernDive. Please open it and follow closely!
- You'll need to load these packages to do the lab (make sure they are installed first, not in your `.Rmd` file!):

```
library(moderndive)
library(tidyverse)
library(skimr)
```

## The data

Source: John Clay (1856). “On the Relation Between Crime, Popular Instruction, Attendance on Religious Worship, and Beer-Houses”, *Journal of the Statistical Society of London*, Vol. 20 #1, pp 22-32.

In 1856, the Reverend John Clay felt that it was high time to figure out what societal factors were playing a role in the incidence of criminal behavior in Britain. He stated that:

“It is a mere truism to say that the progress of popular education, and the formation of religious habits, are fatally opposed by the temptations to animal pleasures, which abound wherever BEER-HOUSES and low ALE-HOUSES abound.”



Clearly, the reverend considered public houses in Britain to be a scourge on society, namely that they “promote drunkenness and its consequent evil” (i.e., crime). Let’s investigate how well we can predict criminals (per 100k population) from the number of public houses (ale/beer houses per 100k population) using simple linear regression.

You’ll first need to download the data from the Reference Plus (scroll down to Data Sets) page on the course website into a “data” directory in your RStudio project. Assuming that, here is how to read in the data into the crime object:

```
crimenames <- c("county", "region_name", "region_code",
               "criminals", "public_houses", "school_attendance",
               "worship_attendance")

# The below assumes that you've downloaded the `beerhall.dat` file to your RStudio project, into a dire
crime <- read_table(here::here("data", "beerhall.dat"),
                   col_names = crimenames)
```

## Basics

Note: you don’t need to use R to answer these questions, but please create a section header using markdown format (# Basics) and type your answers there.

- What is the dependent variable?
- What is the independent variable?
- Copy and paste the provided equation that starts/ends with  $Y = \beta_0 + \beta_1 X$  into your narrative (not an R code chunk), and replace  $y$  and  $x$  in this formula with meaningful variable names (you may wish to reference

the `crimenames` object we made above):  $\hat{y} = b_0 + b_1x$

$$\hat{y} = b_0 + b_1x$$

- The “best-fitting” regression line is “best” in that it minimizes what?
- Why is this method called “simple linear regression” (as opposed to the method in Chapter 6)?

## EDA

Conduct a new exploratory data analysis, which involves three things:

- Looking at the raw values.
- Computing summary statistics of the variables of interest.
- Creating informative visualizations.

### Look at the data

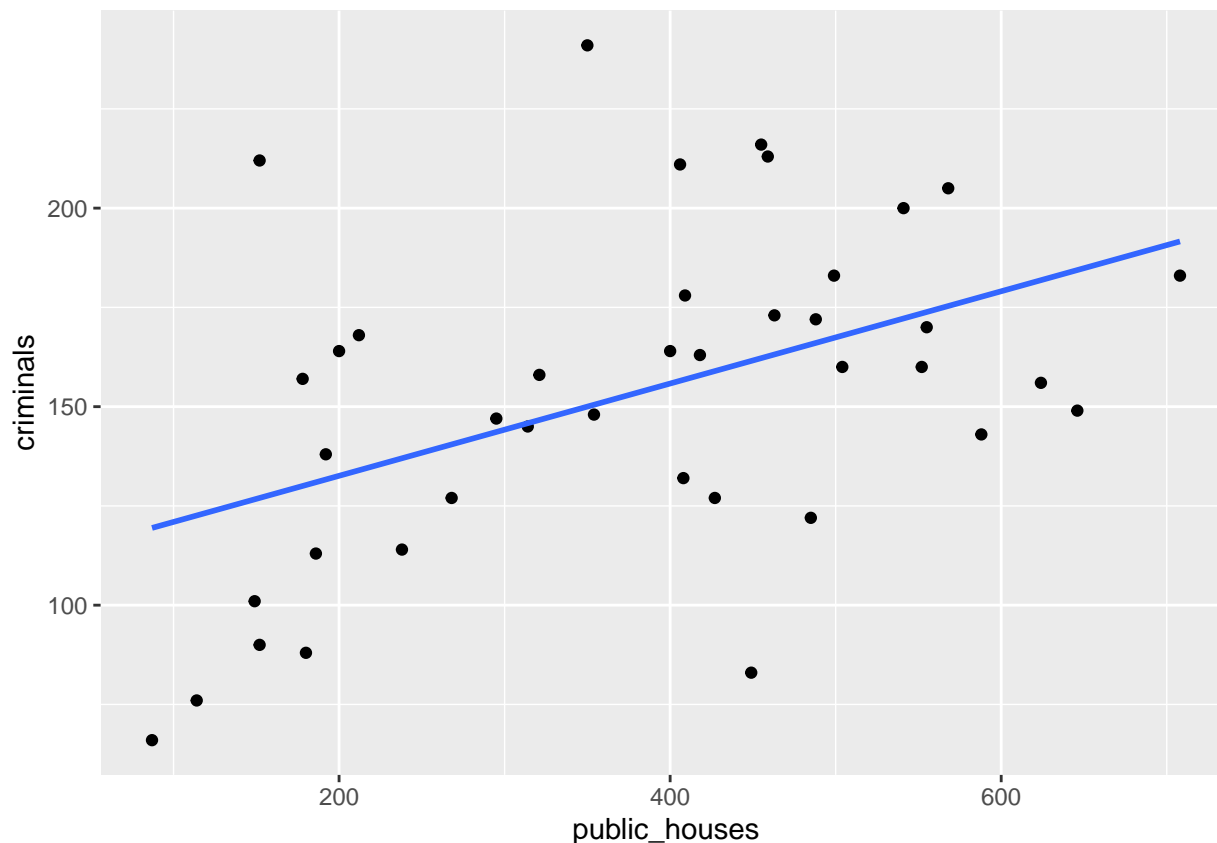
Use `dplyr` to figure out how many counties are in this dataset, and which variable names map onto the independent and dependent variables you identified above.

### Look at summary statistics

Use `select` to select only the independent and dependent variables you identified above, then pipe those variables to the `skim` function from the `skimr` package (you should have loaded this package at the top) to see summary statistics for each. Use `dplyr::summarize` to calculate the correlation coefficient.

### Visualize the data

Recreate the scatterplot below of ale/beer houses per 100K on the x-axis and criminals per 100K population on the y-axis. What can you say about the relationship between public houses and criminals based on this exploration?



## Simple linear regression: (see here)

**Part 1:** First “fit” the linear regression model to the data using the `lm()` function, then apply the `get_regression_table()` function from the `moderndive` R package to the model object. Use the output to fill in this formula with  $y$ ,  $x$ , and the intercept and slope coefficients: (copy and paste into your narrative:  $\hat{y} = b_0 + b_1x$ )

$$\hat{y} = b_0 + b_1x$$

**Part 2:** Interpret the intercept coefficient and the slope coefficient. How do the regression results match up with the results from your exploratory data analysis above?

## Observed/fitted values: (see here)

**Part 1:** What are the observed and fitted values for the Cornwall (ID = 20) and Monmouth regions (ID = 23)? Which region do you think the reverend called “the happiest example of the infrequency of crime”?

**Part 2:** In fact, we could argue with the reverend about the happiest example of the infrequency of crime. There are two ways you could define this. The first way is that the county had the lowest criminals overall. The second is that the county had the lowest criminals *given their number of public houses*. This would mean that the region has the lowest observed criminals, compared to the fitted value based on predicting criminals from public houses.

Filter the regression output using `filter` and the `|` operator (think: `or`) to extract the two alternative definitions above. You'll need to match the `ID` column as a row in the original data (you can just do this visually by comparing the tibbles and typing your answers in your narrative- you don't have to do a join here).

## Residual analysis: (see here)

**Part 1:** Perform a residual analysis and look for any systematic patterns in the residuals. Ideally, there should be little to no pattern- why? Does it seem like there is no systematic pattern to the residuals?

**Part 2:** Recreate this density plot of the residuals. Recall that we would like the residuals to be normally distributed with mean 0 (hint: `?geom_vline()`). Use `dplyr` to calculate the mean of the residuals- is it (pretty close to) 0? Do you think you have more positive residuals than negative, or vice versa?

