

Class30-32

November 2, 2020

1 Simple Linear Regression (SLR)

Suppose you're asked whether or not a patient's age increases their chances of a myocardial infarction (heart attack). An ecologist asks you how tree density (trees per square mile) are associated with the number of deer in a county. A stock broker, looking to capitalize on their investment, asks you to relate historical stock price data to predict future prices.

All of these examples relate one random variable to a set of others. Questions like this can begin to be answered with **regression**

1.0.1 Probabilistic and model form

Simple Linear Regression supposes a conditional probability between one random variable (denoted Y) and another (denoted X) as

$$p(Y = y|X = x) \sim N(\beta_0 + x * \beta_1, \sigma^2)$$

The conditional probability of Y is linearly related to X with two parameters: an intercept (β_0) and a slope (β_1).

When we write a regression model in terms of a single, or in more complex cases many, probability distributions, it is called **probabilistic form**. Probabilistic form highlights the distribution of our variable of interest (Y).

Another common way to write out this relationship is

$$y = \beta_0 + x * \beta_1 + \epsilon \tag{1}$$

$$\epsilon \sim N(0, \sigma^2) \tag{2}$$

This is called **model form** for SLR. Model form highlights the relationship between Y and X , focusing less on the distribution of Y .

1.0.2 LINE assumptions and plotting

Lets assume we collected data on our two random variables X and Y , and arranged them in a dataset

$$\mathcal{D} = \begin{bmatrix} X & Y \\ x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ \vdots & \vdots \\ x_N & y_N \end{bmatrix}$$

SLR makes a set of assumptions usually called the “LINE” assumptions:

- L - Our response (Y) is linearly related to X .
- I - The observations (y_i, x_i) are independent from one another.
- N - The conditional probability of our response y is normally distributed.
- E - The same σ applies to all values of X , i.e. and distribution of errors have equal variance.

With a scatter plot we can investigate linearity and see how the normal dist and “same sigma” assumptions fit with the sample we collected.

1.0.3 Data and scatterplot

The California Department of Public Health, Center for Healthcare Quality collected COVID-19 cases data on the number of residents and healthcare workers at skilled nursing facilities (SNFs), and made this data available to the public for analysis. The dataset is on the facility level, recording the total number of health care workers who were infected with COVID-19 and the total number of residents at SNFs who were infected with COVID-19.

The data is collected over time and we can look at data as of 2020-10-31. Below we will build a scatter plot of the total number of cases among health workers versus the total number of cases among residents.

```
[21]: import pandas as pd
d = pd.read_csv("https://data.chhs.ca.gov/dataset/
→7759311f-1aa8-4ff6-bfbb-ba8f64290ae2/resource/
→d4d68f74-9176-4969-9f07-1546d81db5a7/download/covid19datanursinghome.csv")

dataAsOf20201031 = d.loc[d.as_of_date=="2020-10-31",:]

plt.style.use("fivethirtyeight")
fig,axs = plt.subplots(1,2)

ax=axs[0]
sns.scatterplot( x="total_health_care_worker_cases"
                 ,y="total_resident_cases"
                 ,data=dataAsOf20201031
                 ,ax=ax)
ax.set_xlabel("Total Healthcare worker cases", fontsize=12)
ax.set_ylabel("Total resident cases", fontsize=12)

ax.set_xlim(-1,100)
```

```

ax.set_ylim(-1,250)

ax=axes[1]
sns.regplot( x="total_health_care_worker_cases"
             ,y="total_resident_cases"
             ,data=dataAsOf20201031
             ,ax=ax
             ,scatter_kws={"alpha":0.25}
             ,line_kws={"lw":2}
             )
ax.set_xlabel("Total Healthcare worker cases", fontsize=12)
ax.set_ylabel("Total resident cases", fontsize=12)

ax.set_xlim(-1,100)
ax.set_ylim(-1,250)

fig.set_size_inches(12,6)
fig.set_tight_layout(True)

```



1.0.4 Least Squares and finding optimal parameters

1.0.5 Interpreting SLR

1.0.6 Inference (tests and CIS)