Lab 7: Linear Regression Analysis

1. **Scatter Plots with Simple Regression Line**
2. **Linear Regression Output Analysis**
3. **Multiple Regression Analysis**

Based on the textbook Statistical Thinking through Media Examples by Anthony Donoghue.

As we saw in lab-6, scatter plots and correlation coefficients are useful ways of characterizing the relationship between two quantitative variables. In order to further summarize this relationship, we can fit a line to the scatter plot of data points using a statistical technique known as ***simple linear regression***.

The aim of simple linear regression is to choose a line to fit to our scatter plot that gets as close to all the data points as possible. We want to minimize the difference between the actual y-values, and the predicted y-values we calculate using the fitted line. For this reason, the line we choose is often called the *line of best fit* (the regression line).

> The equation for the regression line is: $\hat{y}_i = b_0 + b_1 x_1$ where b_0 is the intercept of the line and b_1 is the coefficient associated with the explanatory variable x_1, and also represents the slope of the line.

Using this equation, we can make predictions for the y-values based off of a given x-value. Our x-values are called *explanatory variables*.

Simple Linear Regression code

```
x_vals = df[“x_val_column"]
x_vals = sm.add_constant(x_vals.values)

y_vals = df["y_val_column"]

reg_model = OLS(y_vals, x_vals).fit()

display(reg_model.summary())
```

Multiple Linear Regression code

```
X = df[["x_val_column1", "x_val_column2"]]
X = sm.add_constant(X.values)

y = df["y_val_column"]

reg_model = OLS(y, X).fit()

display(reg_model.summary())
```

To access the first parameter of a model:

```
b0 = reg_model.params[0]
```

To access all other parameters, simply replace the number 0 with the number corresponding to the parameter you need.