2.2 – Modeling with Linear Functions

1 Introduction

We typically begin the modeling process with simple models and build progressively more complex models when appropriate. To begin the discussion of continuous modeling, we will explore the simplest of the continuous variable models covered in this course, the **linear** model [1]. Recall from our last reading, the slope-intercept form of a **linear** function in two dimensions is:

$$\hat{y} = mx + b$$

where x is the explanatory variable and $\hat{y}$ is the predicted value of the response variable; m is the rate of change of y with respect to x (slope), and b is the y-intercept.

We use the notation $\hat{y}$, read as y-hat, to indicate that this value is not an observed data point, but a prediction made by our model. This helps distinguish between actual outcomes (collected data depicted as y) and estimated or predicted values. We may also see models written as $\hat{f}(x) = mx + b$, where $\hat{f}(x)$ is the predicted value for a given input x.

In the equation of a line, the values of x and $\hat{y}$ can vary depending on the inputs and predictions made, but the parameters m and b are fixed for a given model. These parameter values are determined during the solve step of the modeling process, using either first principles or data fitting techniques [1].

2 Building a Linear Model using a First Principles Approach

If we understand the context for the problem we are trying to model, we may be able to estimate model parameters based on how we expect the system to behave. For a linear model we must ask ourselves the following questions:

- What is the value of the dependent variable when the independent variable is 0? (e.g. what is the population at time 0?) This gives us b.
- Do we expect the dependent variable to increase linearly ($m > 0$) or decrease linearly ($m < 0$), or stay the same ($m = 0$) as the independent variable changes?
- How quickly do we expect the dependent variable to change? This gives us the value of the slope, m.

Problem 2.2.1: First Principles Approach

You and your company mates are taking a taxi to the train station for a weekend trip to the city, but unsure of where you want to catch the train. There is a base fee of \$10 for hiring the taxi plus \$0.50 per mile. You want to model the total cost of the ride based on the number of miles you drive.

Solution: We expect a constant of change, for every mile we drive, the total cost increases by \$0.50, so our slope, $m = .5$. We also have a fixed starting cost, or initial value. This becomes our y-intercept, $b = 10$.

$$\hat{y} = 0.5x + 10$$

We arrived at this model simply by thinking our way through the context of the problem.

3 Building a Linear Model using an Empirical Approach

Given a data set, we need to model the data with the appropriate type of model. As a general rule, the empirical modeling process we will use to develop continuous models for a data set is:

- Plot the data and decide which family of functions best represents the trend of the data. Estimate the parameters and develop the model.
- Use the model to predict a value.
- Restate the predicted value in real-world terms and reflect on your solution.
- Evaluate the effectiveness of the model and make adjustments.

We saw data visualization techniques in descriptive analytics and this will be very useful for predictive analytics. After plotting the data and deciding to use a linear model, we need to estimate the slope and y-intercept. In general, the parameters of any function (linear, trigonometric, exponential, etc.) determine its shape and location. It is possible to draw a line through any two points. The slope of the line through points (x_1, y_1) and (x_2, y_2) is calculated by the following equation:

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

The sign of the slope will determine if the function is increasing or decreasing. The magnitude of the slope will determine how steep the line is, which indicates how fast the dependent variable is changing with respect to the independent variable.
Once we have determined the slope of the line through the two points we can calculate the y-intercept using slope-intercept form. Solving for parameter b we are left with:

$$b = -mx_1 + y_1$$

After determining the parameters we can now use the linear model to predict a value and interpret the value in real world terms.

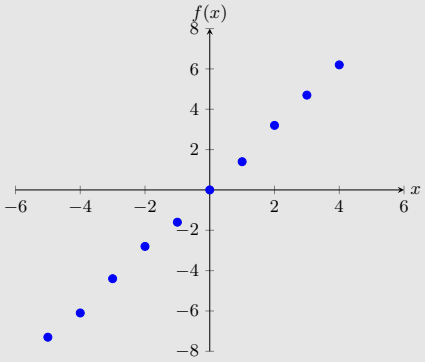
Problem 2.2.2: Empirical Approach

Develop a linear model for the data below. Make a prediction for $f(x)$ when $x = 2.5$.

x	-5	-4	-3	-2	-1	0	1	2	3	4
$f(x)$	-7.3	-6.1	-4.4	-2.8	-1.6	0	1.4	3.2	4.7	6.2

3.1 Transform the Problem

We have no context for this data, so we will start with the empirical approach. The first step is to plot the given data to see if it exhibits a pattern: transforming the data from tabular to graphical form. Which family of functions (linear, exponential, etc.) best models the data?



The plot of the data exhibits a pattern that appears linear. Thus, we choose to model this data with a linear function $y = mx + b$ and estimate the parameters m and b.

3.2 Solve

Notice that since the data contains the point $(0, 0)$, an initial estimate for the y-intercept is zero. Since b equals zero, there is one parameter left to find, the slope m.

One way to estimate the slope for a linear model is to use the average rate of change between any two data points that are representative of the general trend in the data. For this example we will use the points $(-4, -6.1)$ and $(-1, -1.6)$, but we could use any two points that are representative of the trend. The average rate of change between the two points selected is:

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{(-1.6) - (-6.1)}{(-1) - (-4)} = \frac{4.5}{3} = 1.5 \quad (1)$$

After estimating the slope ($m = 1.5$) and the y-intercept ($b = 0$) we arrive at the model

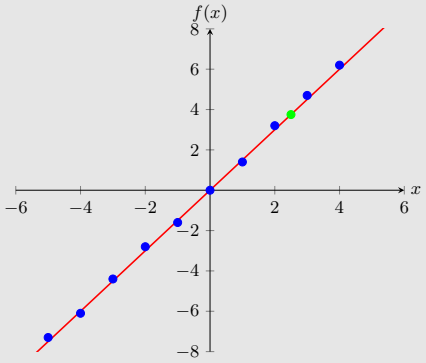
$$\hat{f}(x) = 1.5x + 0 \quad (2)$$

In order to predict the value of $f(x)$ when $x = 2.5$ we simply substitute $x = 2.5$ into the model.

$$\hat{f}(2.5) = 1.5(2.5) + 0 = 3.75 \quad (3)$$

3.3 Interpret

The model for the data with b = 0 and m = 1.5 is $f(x) = 1.5x$. To reflect, graph this line on the same set of axes along with the data. Our prediction of $f(2.5) = 3.75$, plotted in green below, follows the trend of the data and further validates our model. This prediction is an example of ***interpolation*** because we have estimated a value of the response variable *between* observed values of the explanatory variable. If we were to predict the value of $f(x)$ when $x = 5$, this would be an example of ***extrapolation*** because we would predict a value *outside* of the observed values of the explanatory variable. Consequently, we are less certain about the accuracy of our prediction [2].



3.4 Model Assessment

In the last reading, we learned about two tools for model assessment. Let's find SSE and R^2 for this model. When calculating SSE, it is often useful to make a table.

x	-5	-4	-3	-2	-1	0	1	2	3	4
$f(x)$	-7.3	-6.1	-4.4	-2.8	-1.6	0	1.4	3.2	4.7	6.2
Prediction $\hat{f}(x)$	-7.5	-6	-4.5	-3	-1.5	0	1.5	3	4.5	6
Error $\hat{f}(x) - f(x)$	-0.2	0.1	-0.1	-0.2	0.1	0	0.1	-0.2	-0.2	-0.2
Squared Error $(\hat{f}(x) - f(x))^2$	0.04	0.01	0.01	0.04	0.01	0	0.01	0.04	0.04	0.04
										Sum: 0.24

We will not manually calculate R^2 in this class, but we can find it by creating a scatterplot in Excel and adding a trend line. We find that $R^2 = 0.99$. Remember that the closer R^2 is to 1, the better the fit of the model to the data. This is generally true, but doesn't guarantee a good model. We will address the dangers of *overfitting* in upcoming lessons.

4 Model Assessment

In the previous reading, we explored a qualitative way to interpret R^2 as a tool for assessing model fit—observing how the value increases as the trendline better matches

the data. In this reading, we will build on that intuition by understanding where R^2 comes from and introducing a second tool for model assessment: the **Sum of Squared Error (SSE)**.

The **Coefficient of Determination,** R^2, is given by the following formula:

Definition 2.2.1 (**The Coefficient of Determination,** R^2)
is the percentage of the total observed variation in the response variable that is accounted for by changes in the explanatory variable [3].

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where y_i is a data point, $\hat{y}_i$ is the value the model predicted, and $\bar{y}$ is the sample mean.

R^2 is related to the second tool we will use, the **Sum of Squared Error (SSE)**, which is given by the following formula:

Definition 2.2.2 (**Sum of Squared Error**)
is the squared difference between the observed and predicted values.

$$SSE = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where y_i is a known data point and $\hat{y}_i$ is the value the model predicted.

Take a moment to walk yourself through the formula for SSE. You must first find the error in your model by finding the difference between the predicted and known value, then square the result, and finally sum up each individual error to find the total squared error. By squaring each error, we ensure that all error is positive and give more weight to large errors. Sum of squared error can help us measure the accuracy of our predictions and compare models [3].

Let's go back and look at the equation for R^2. You should see that the numerator of the fractional term is SSE again, but the denominator is slightly different. In the denominator we are calculating the total sum of squares, SST, by summing the squared difference between the observed values and the mean. Our formula now becomes:

$$R^2 = 1 - \frac{SSE}{SST}$$

This helps us with our conceptual understanding of R^2 as the percentage of the total observed variation in the response variable that is accounted for by changes in the explanatory variable [3]. Observe that as SSE decreases to zero, R^2 approaches one (a perfect model where all variation in the response variable is accounted for by the explanatory variable), and as SSE increases, R^2 will decrease to zero (a horizontal line where none of the variation in the response is accounted for by changes to the explanatory). Note: it is mathematically possible for R^2 to be negative. This indicates that the model is worse than simply taking the average of the data and should set off alarm bells. If we calculate a negative R^2, we should immediately reevaluate our model. In later readings we will discuss how a higher R^2 is usually an indication of better fit, but does not guarantee a good model.

As we continue modeling, we will use both R^2 and SSE to assess how well our functions fit the data, and to compare the effectiveness of different model types. These tools help guide our modeling decisions, but always require interpretation in the context of the real-world problem.

5 Ethical Checklist

As we discussed in our introductory reading, we must consider data and model validity, and clearly communicate our linear models. Below are some questions to think through as you consider whether your linear model is ethical.

- **Data validity**. As with descriptive statistics, you need look at where you data is coming from and how it was collected. Normally, when you are using data collected by someone else, there should be some report or description of how the data was collected. How much did you need to clean the data? If you had to clean a lot of it, what are the possible implications about this data set?
- **Model validity**. Examine your modeling decisions.
  - Did you make modeling decisions while cleaning or exploring your data? How did you identify your explanatory and response variables, are they the best variables to choose? Does your model answer the question you are trying to find?
  - Is your predictive model accurate? How did you evaluate your model? How does your model perform on new data? Does your choice of linear model make sense in the context of the problem?
- **Communication**. Is your visualization clear? Is it misleading, or does it convey an honest representation of the data? Did you thoroughly communicate each of your assumptions and modeling decisions? Did you communicate the limitations of your model?

References

[1] US Military Academy. *Modeling in a Real and Complex World*. West Point, New York: Department of Mathematical Sciences, 2022.

[2] James Stewart, Daniel Clegg, and Saleem Watson. *Calculus: Early Transcendentals*. Cengage, 2020.

[3] Nathan Tintle et al. *Introduction to Statistical Investigations*. Siley, 2021.