

# Introduction to Modeling: Takeaways

by Dataquest Labs, Inc. - All rights reserved © 2021

## Syntax

### VISUALIZING BIVARIATE RELATIONSHIPS

- Generating scatterplots to visualize bivariate relationships:

```
ggplot(data = uber_trips,  
       aes(x = distance, y = cost)) +  
  geom_point()
```

- Visualizing a linear regression line:

```
ggplot(data = uber_trips,  
       aes(x = distance, y = cost)) +  
  geom_point() +  
  geom_smooth(method = "lm", se = FALSE)
```

### ANALYZING THE RESIDUALS

- Calculating mean absolute error (MAE):

```
MAE <- mean(abs(df$residuals))
```

## Notation

### GENERAL FORM OF A PREDICTIVE MODEL:

$$Y = f(X) + \epsilon$$

- In this context,  $X$  represents a set of **inputs** and  $Y$  represents a set of **outputs**. The random error term  $\epsilon$  is independent of  $X$  and has a mean of approximately zero. In reality, the error term is unknown, so we can represent our estimate of  $Y$  as a function of  $X$  as:

$$\hat{Y} = \hat{f}(X)$$

- We can omit the error term because it averages to zero. The "hat" symbol indicates an estimate.

### BIVARIATE LINEAR MODEL:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- Here  $\hat{y}$  indicates a prediction of  $Y$  assuming  $X = x$ . Intercept ( $\hat{\beta}_0$ ) refers to the value on the y-axis where the value on the x-axis is equal to zero. Slope ( $\hat{\beta}_1$ ) is the change in  $\hat{y}$  for every single unit change in  $x$ .

---

## MEAN ABSOLUTE ERROR (MAE) :

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- $y_i$  = observed output value
- $\hat{y}_i$  = predicted output value
- $\sum_{i=1}^n$  = the sum of...
- $|y_i - \hat{y}_i|$  = the absolute value of each residual
- $\frac{1}{n}$  = divide the above by the total number of observations (to return the average value)

## Concepts

- An *input*, or input variable is also sometimes referred to as a predictor, independent variable, feature, attribute, descriptor, or simply variable\*.
- An *output* or *output variable* is also known as a *dependent variable*, *outcome*, *response variable*, *response*, *target*, or *class*.
- **Prediction:** If our primary purpose of building a model is to generate an accurate prediction, we aren't too concerned if the function of the input to predict the response is a "black box." Rather than understanding  $f$ , our primary concern is that our model gives us accurate output predictions for each input.
- **Inference:** When motivated by inference, we may or may not be interested in generating predictions for our response. Instead, we wish to understand  $f$  and how the response variable is affected by changes in the input variable.
- **Error:** The accuracy of our prediction for an output depends on two types of error: *reducible error*, and *irreducible error*. Even though the "E" in mean absolute error stands for error, it does not refer to the epsilon error described here.
- **Reducible error** can be minimized by selecting a statistical method that provides a good estimate of the response. In linear regression, one key to minimizing reducible error is to select the input variable that provides the most accurate estimate of the output variable.
- **Irreducible error** is out of our control. Unmeasurable variation contributes to error. Another term for this is random noise.
- **Residual:** The difference between the observed value and the model's prediction.
- **Summary measures:** Statisticians have developed various summary measurements (mean absolute error, for example) that can take the residuals from our model and transform them into a single value that represents the predictive ability of our model.
- **Overfitting:** When a model finds patterns in the training data that are not present in the unseen data.

# Resources

- [An Introduction to Statistical Learning with Applications in R by James et al.](#)
- [Applied Predictive Modeling by Max Kuhn and Kjell Johnson.](#)

Takeaways by Dataquest Labs, Inc. - All rights reserved © 2021