Lesson: Frameworks and goals of statistical modeling

Module: Introduction to statistical models

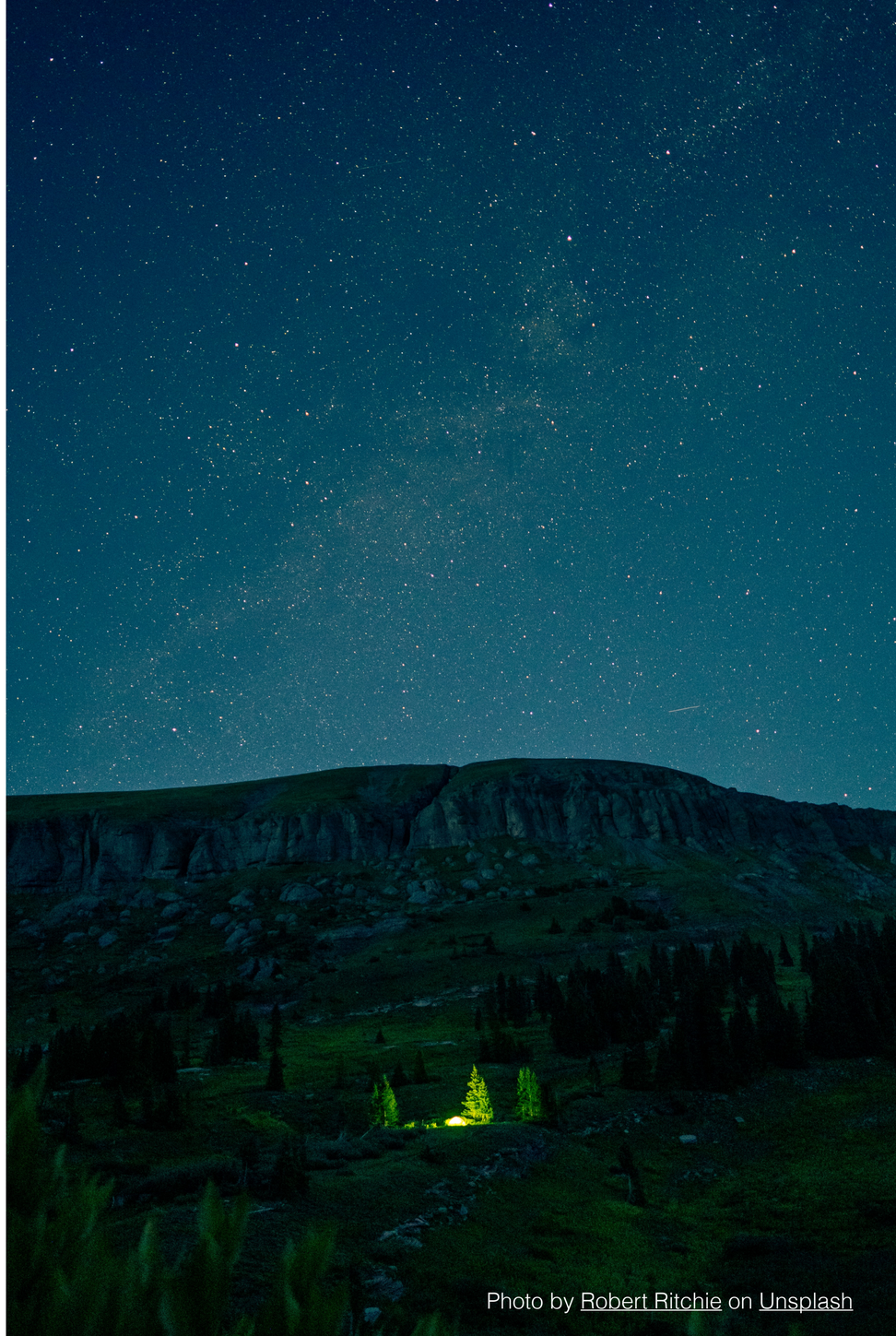
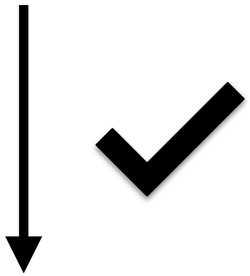


Photo by Robert Ritchie on Unsplash

switches + dials



Audio signal

Photo by Steve Harvey on Unsplash

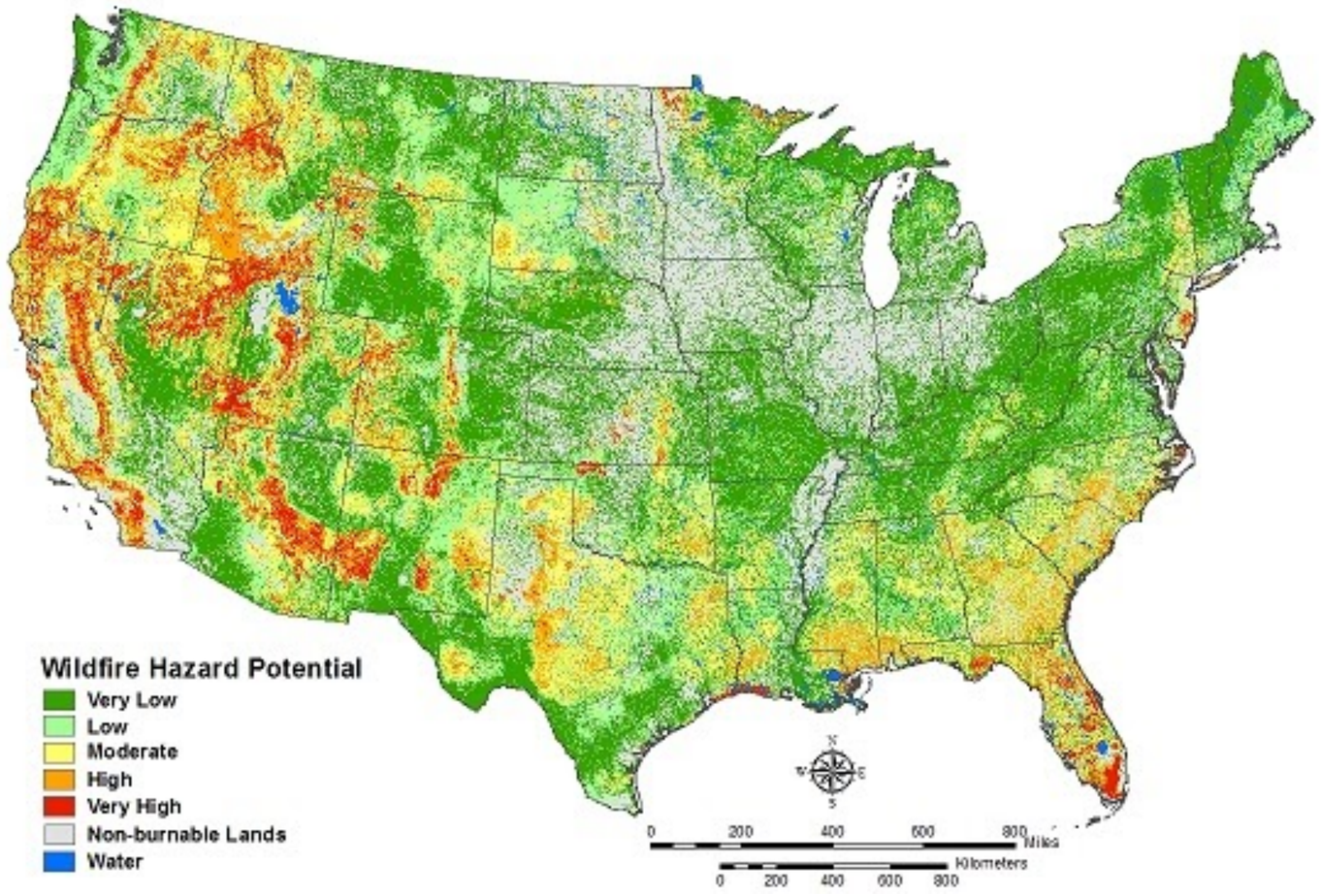
$$Y = f(\mathbf{x}) + \varepsilon$$

output

Systematic component

random component

Photo by Steve Harvey on Unsplash



temperature, humidity + fuel density + elevation, lightening storms + number of nearby campsites + …

↓ ?

Likelihood of large wildfire

Image from firelab.org: https://bit.ly/37q9EGA

Definition: A *statistical unit* is one member of the set of entities being studied.

Definition: A *population* is a collection of units about which research questions are asked.

Definition: A *sample* is a subset of the population. Typically, samples should be *representative.*

Photo by Max Adulyanukosol on Unsplash

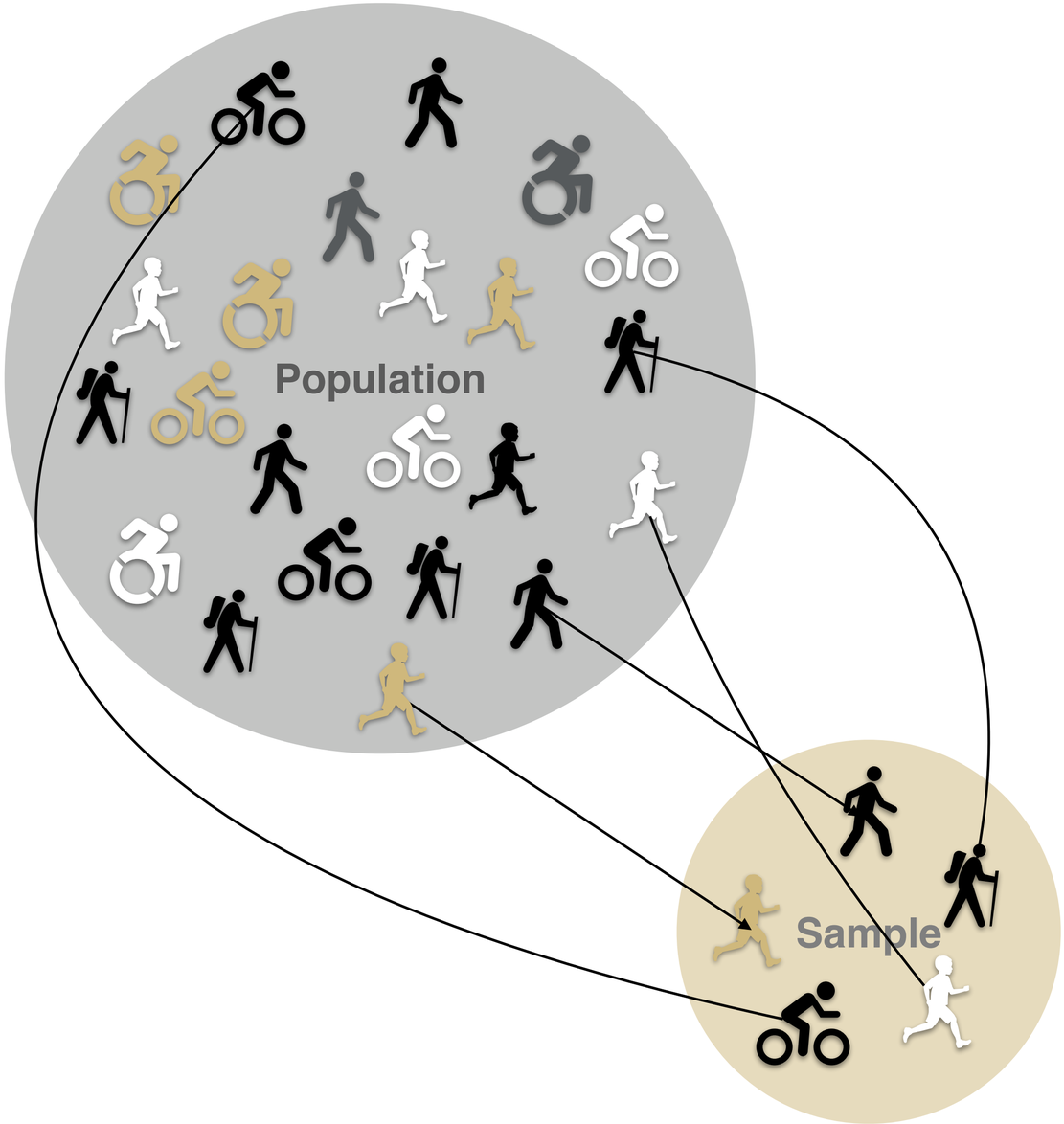


Photo by ismail mohamed - SoviLe on Unsplash

Definition: *Inferential statistics and data science* is the process of learning about relationships in a sample in a way that is reliable enough to *generalize* from the sample to a population of interest.

Photo by Danny Nee on Unsplash

Lesson: The assumption of *concept validity*

Module: Introduction to statistical models

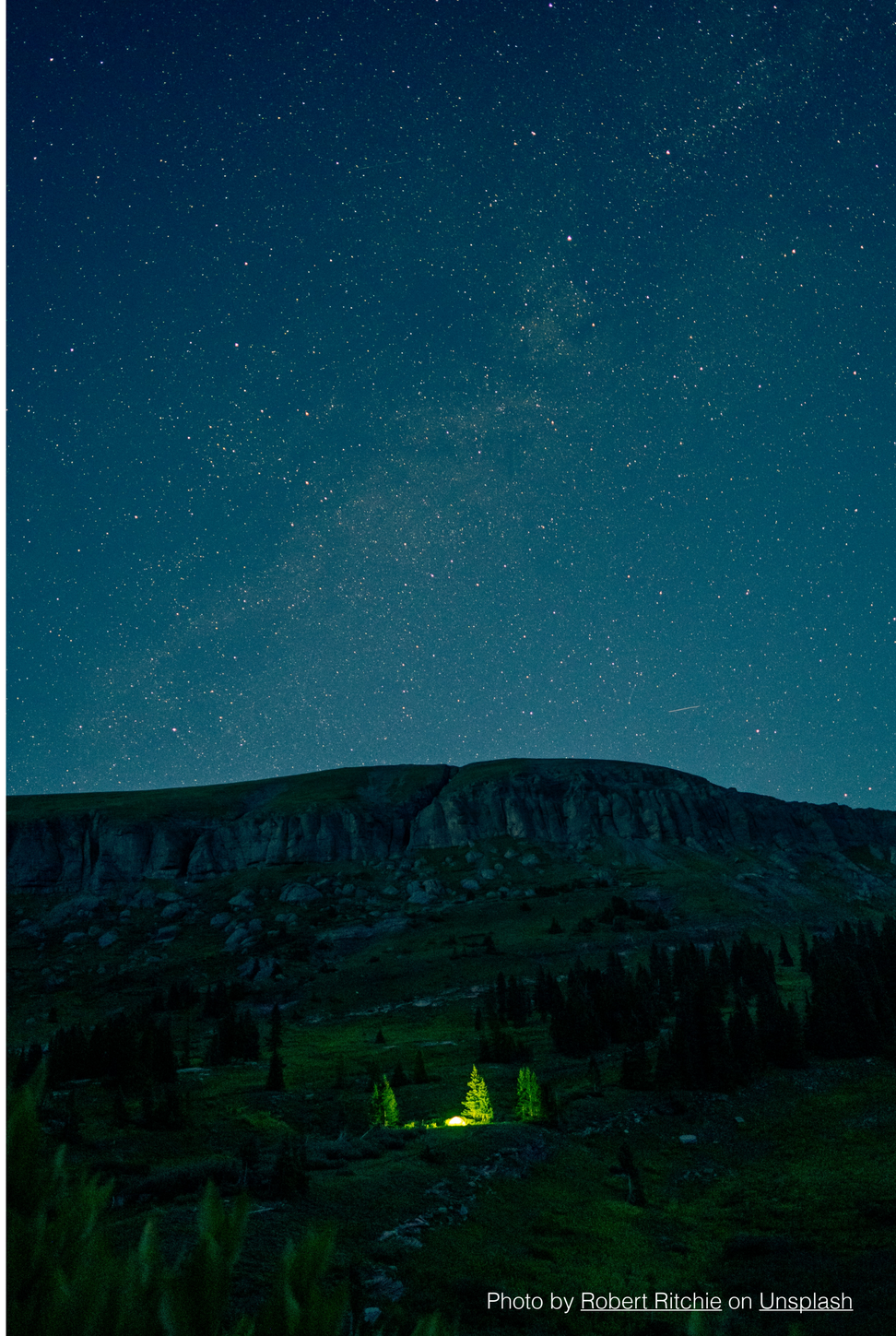


Photo by Robert Ritchie on Unsplash

Is cycling any less safe in the winter than in the summer?

Photo by Max Adulyanukosol on Unsplash

Definition: To *operationalize* a concept means to derive a set of steps to measure the concept.

Definition: The *validity* of a dataset or measurement tool is the extent to which the dataset or measurement tool measures what it claims to measure.

Photo by Max Adulyanukosol on Unsplash

Study on cycling safety:

http://bit.ly/3oZNd0x

Photo by Max Adulyanukosol on Unsplash

Lesson: The linear regression model

Module: Introduction to statistical models

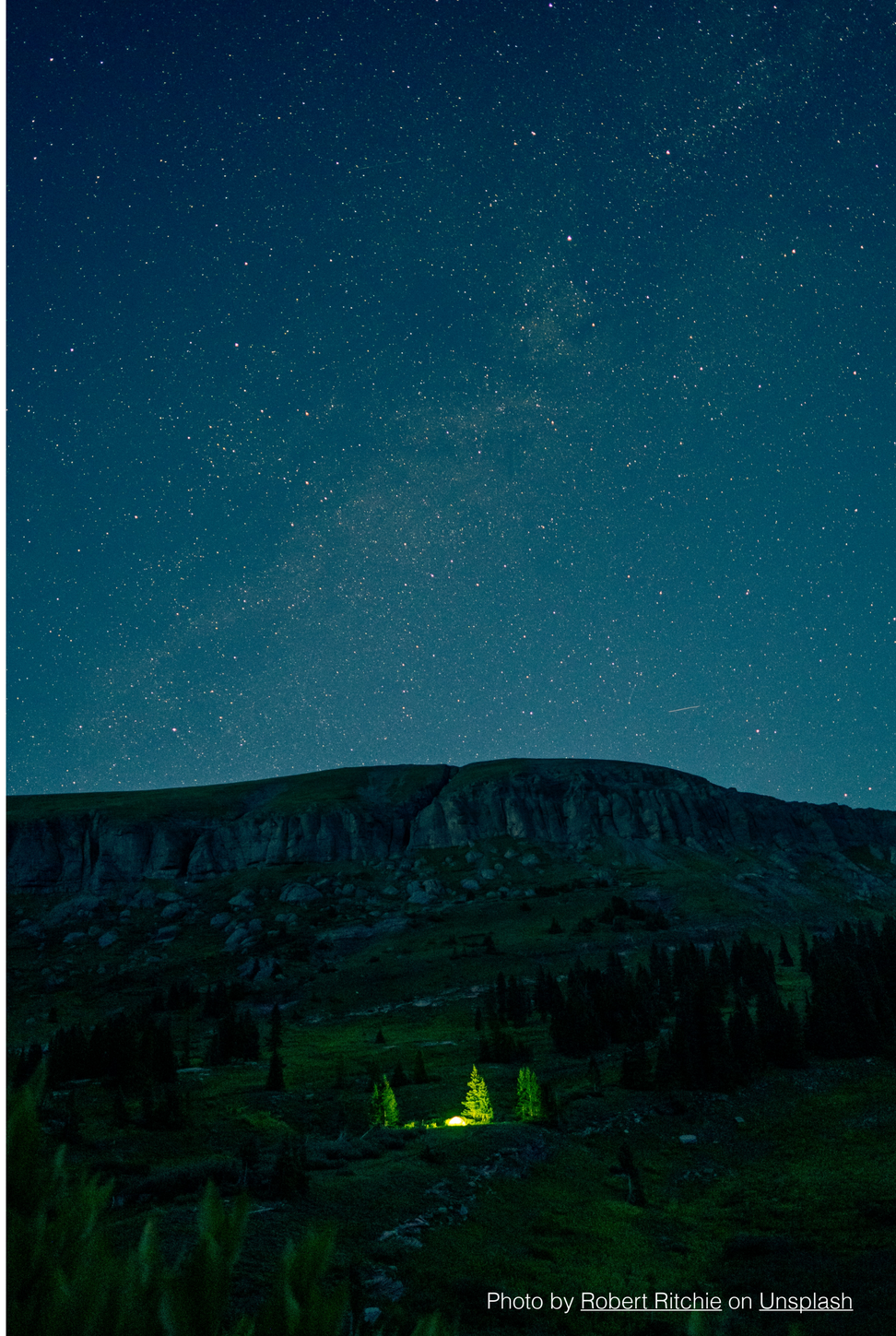


Photo by Robert Ritchie on Unsplash

Motivation and Context

Basic statistical inference: estimate or test hypotheses about data generated from a distribution with a constant mean. For example:

$$Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

μ: est. the mean

But what if the mean depended on another variable? For example:

$$Y_1, \ldots, Y_n \overset{iid}{\sim} N\left(\mu(X_1, X_2, \ldots, X_p), \sigma^2\right)$$

$f(x) = y$

The Context of Linear Regression

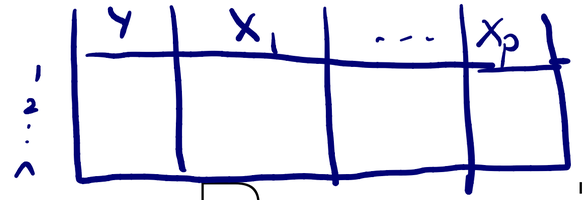
Linear regression is used to explain or model the relationship between a single variable Y, and one or more variables $x_1, \ldots, x_p$.

Definition: Y is called the *response, outcome, output,* or *dependent* variable.

Definition: $x_1, \ldots, x_p$ are called *predictors*, *inputs*, *independent variables*, *explanatory variables,* or *features*. In some contexts, they are also called *covariates*.

We will assume, for now, that all variables are continuous (but will soon extend our methods to allow for discrete variables!).

Photo by Jr Korpa on Unsplash



The Context of Linear Regression

The simplest relationship between these variables is a linear relationship: For $i = 1, \ldots, n$

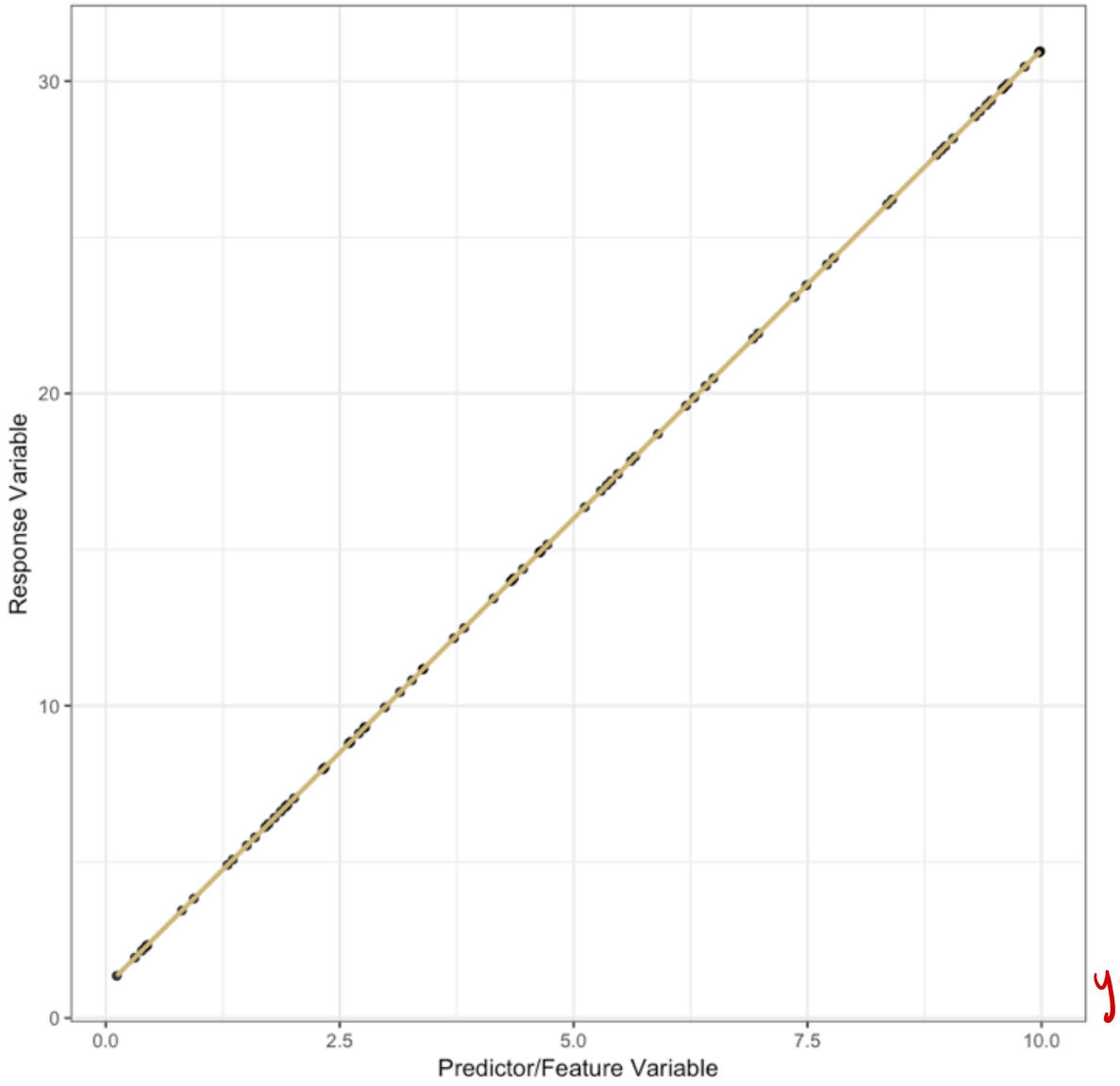
$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

But note that, when we actually measure data, measurements aren't perfect—there is error. So, we might use the following to model our data:

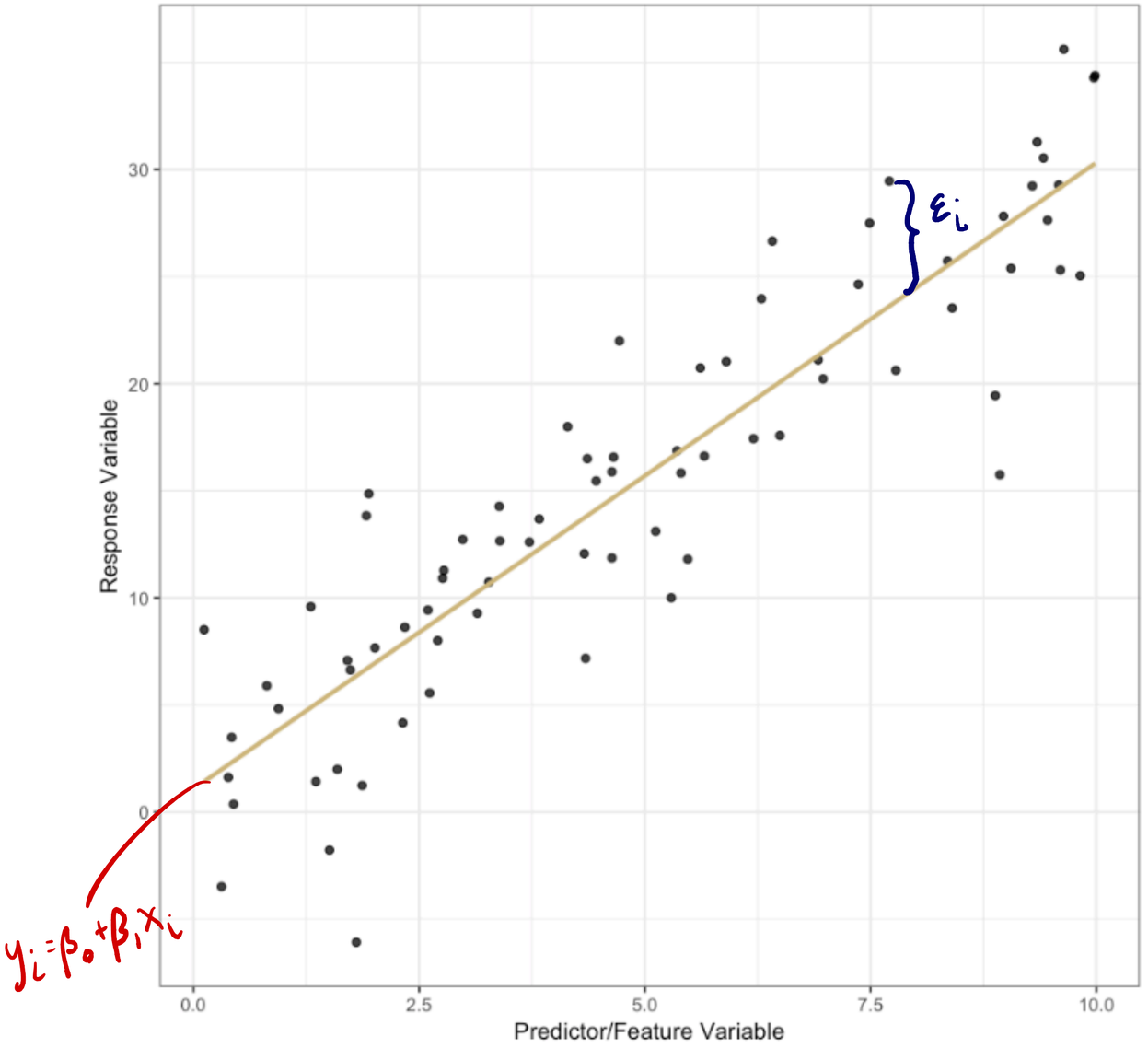
$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

ε_i: random variable

Exact Linear Relationship



Noisy Linear Relationship



The Context of Linear Regression

Regression analysis has two main objectives:

1. *Prediction:* Predict an unmeasured/unseen Y using observed $x_1, \ldots, x_p$.

2. *Explanation*: To assess the effect of, or explain the relationship between, Y and $x_1, \ldots, x_p$.

Can we infer causality?

Lesson: Matrix representation of the linear regression model

Module: Introduction to statistical models

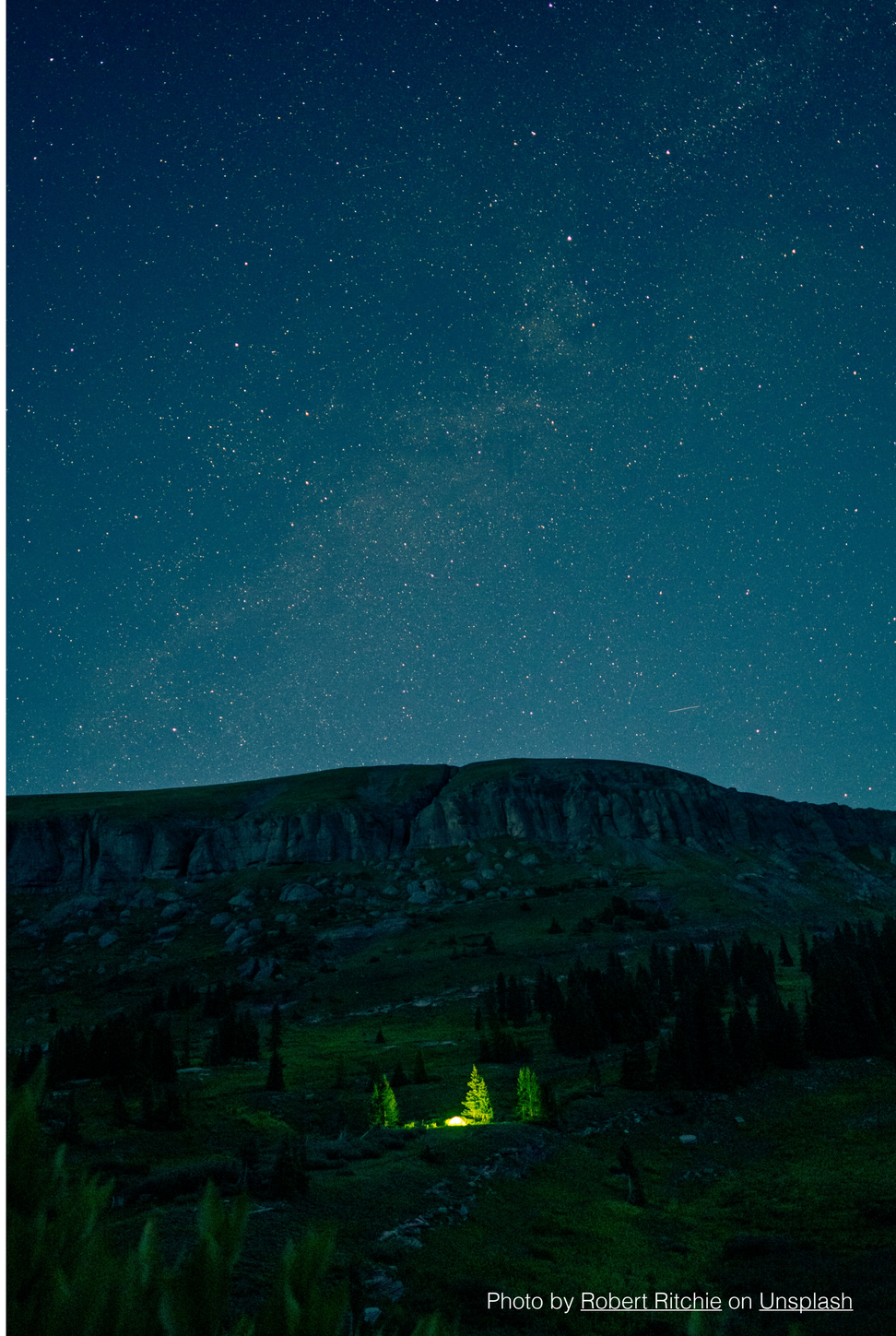


Photo by Robert Ritchie on Unsplash

$\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & & x_{np} \end{pmatrix}$

p = # of predictors

X r.v.
x realization

The Linear Regression Model

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be the response variable and $\mathbf{x_1} = \begin{pmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{n,1} \end{pmatrix}$, $\mathbf{x_2} = \begin{pmatrix} x_{1,2} \\ x_{2,2} \\ \vdots \\ x_{n,2} \end{pmatrix}$, $\ldots$, $\mathbf{x_p} = \begin{pmatrix} x_{1,p} \\ x_{2,p} \\ \vdots \\ x_{n,p} \end{pmatrix}$

$\underline{x}_1$

be predictors; we will collect the predictors in a matrix: $X = \begin{pmatrix} \mathbf{1} & \mathbf{x_1} & \mathbf{x_2} & \ldots & \mathbf{x_p} \end{pmatrix}$, where $\mathbf{1} = (1,1,...,1)^T$. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ be a vector of parameters. Finally, let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ be a vector of error terms. Then we can write our model as:

$n \times (p+1)$

$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$

$n \times 1$

$n \times 1$ $\quad$ $n \times (p+1)$ $\quad$ $(p+1) \times 1$

Matrix/vector representation

For $i = 1, \ldots, n$

$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$

Single equation representation

The Linear Regression Model

Examples of linear models (i.e., models that linear methods can handle):

like least squares

1. $Y_i = \beta_0 + \beta_1 \underbrace{\log(x_i)}_{z_i} + \varepsilon_i$

2. $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i$

Examples of nonlinear models:

1. $Y_i = \beta_1 [\cos(x_i)]^{\beta_2} + \varepsilon_i$

2. $Y_i = \beta_0 + \beta_1 \sin(\beta_2 x_i) + \varepsilon_i$

Photo by Jason Leung on Unsplash

Lesson: Assumptions of the linear regression model

Module: Introduction to statistical models

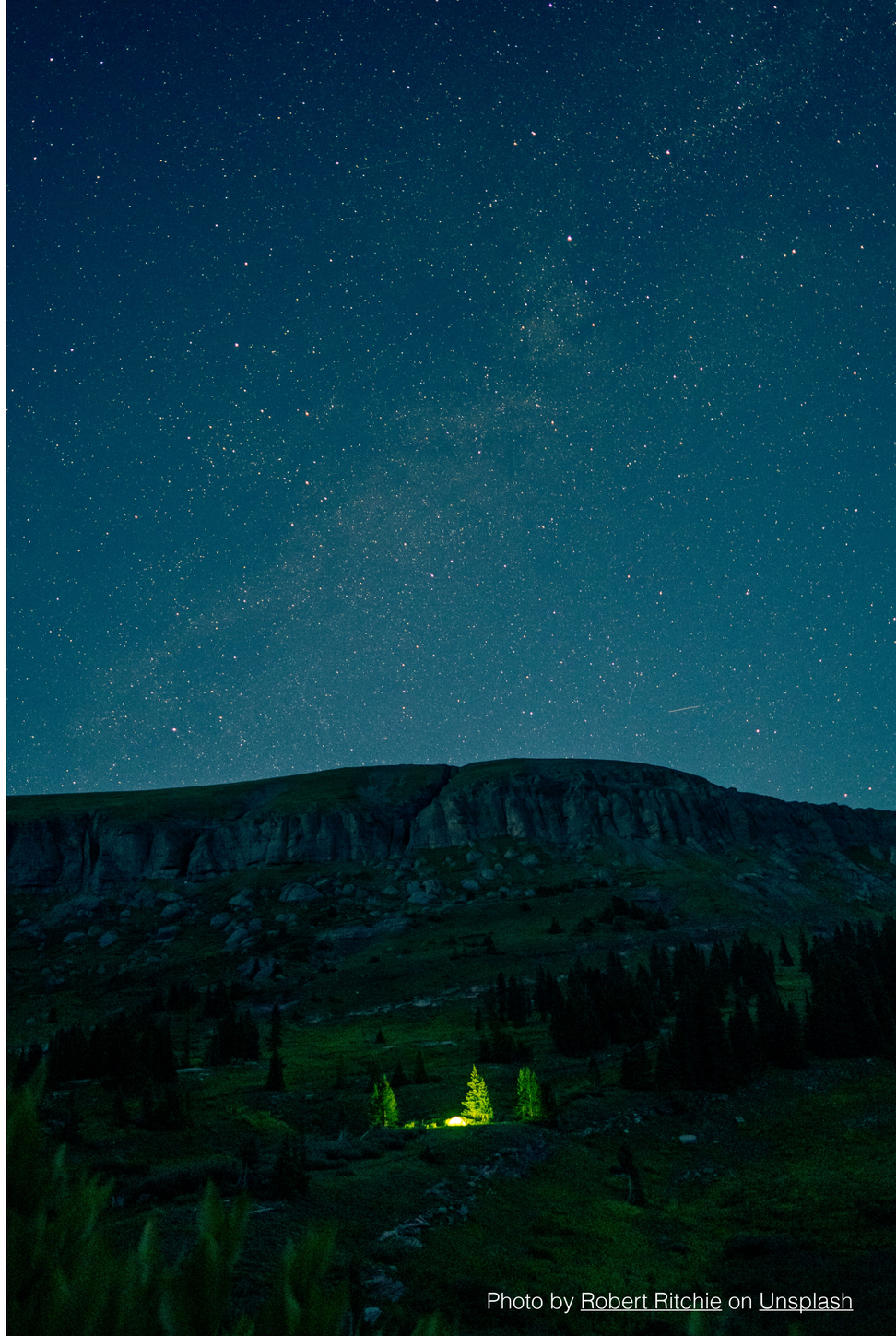


Photo by Robert Ritchie on Unsplash

$\underline{Y} \sim N(X\beta, \sigma^2 I_n)$, $I_n = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}_{n \times n}$

The Linear Regression Model

(Ind $\Rightarrow$ uncorrelated. uncorrelated $\not\Rightarrow$ Ind.)

<u>Definition/Assumptions</u> of the linear regression model:

1. *Linearity*: Linear relationship between $\underline{Y}$ and β
2. *Independence*: Y_i is independent from Y_j
   ($Cov(Y_i, Y_j) = 0 \quad i \neq j$)
3. *Homoskedasticity (constant variance)*:
   $Var(Y_i) = \sigma^2$ for all $i = 1, \ldots, n$
4. *Normality*:
   $Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \sigma^2)$

Lesson: Interpreting the linear regression model

Module: Introduction to statistical models

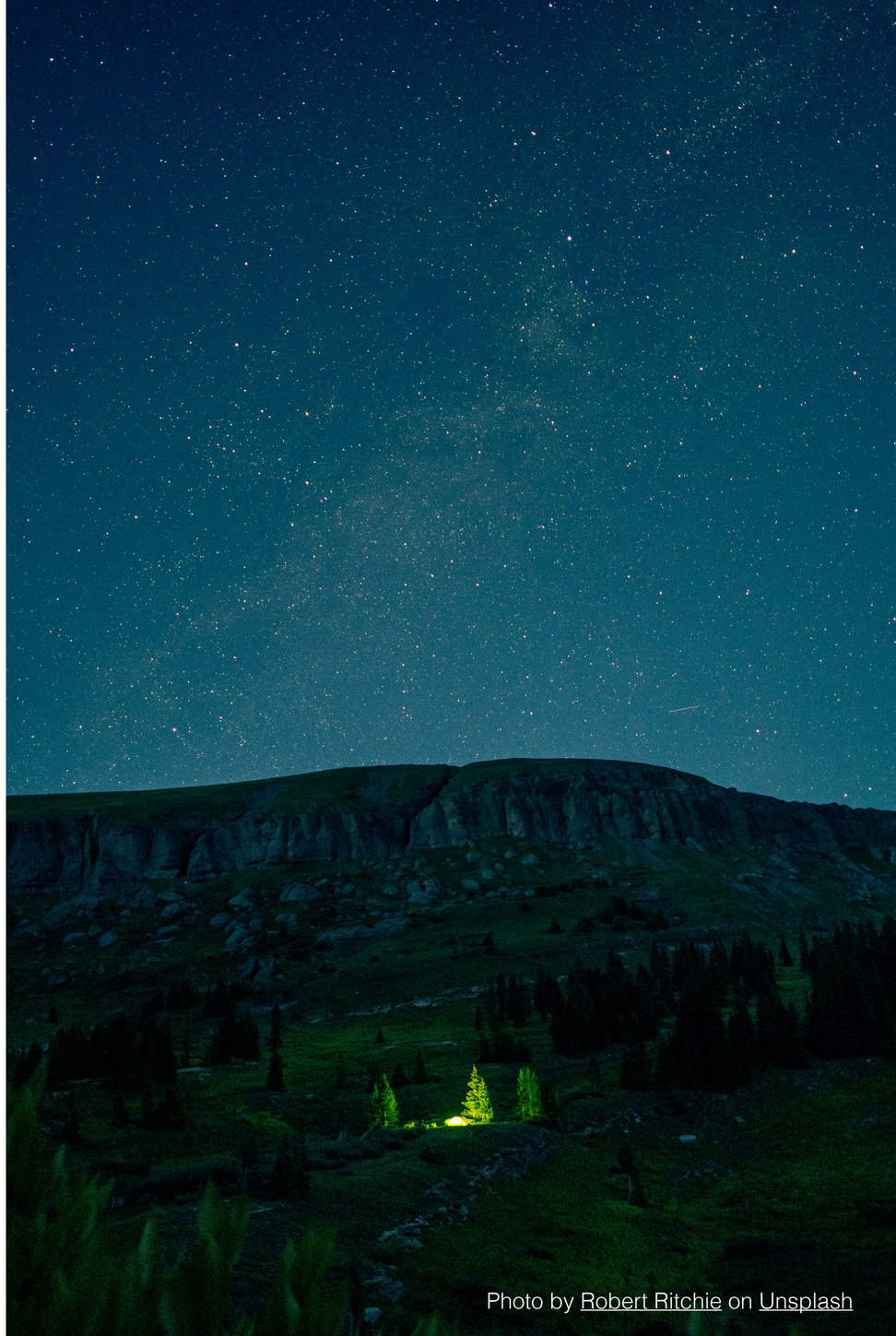


Photo by Robert Ritchie on Unsplash

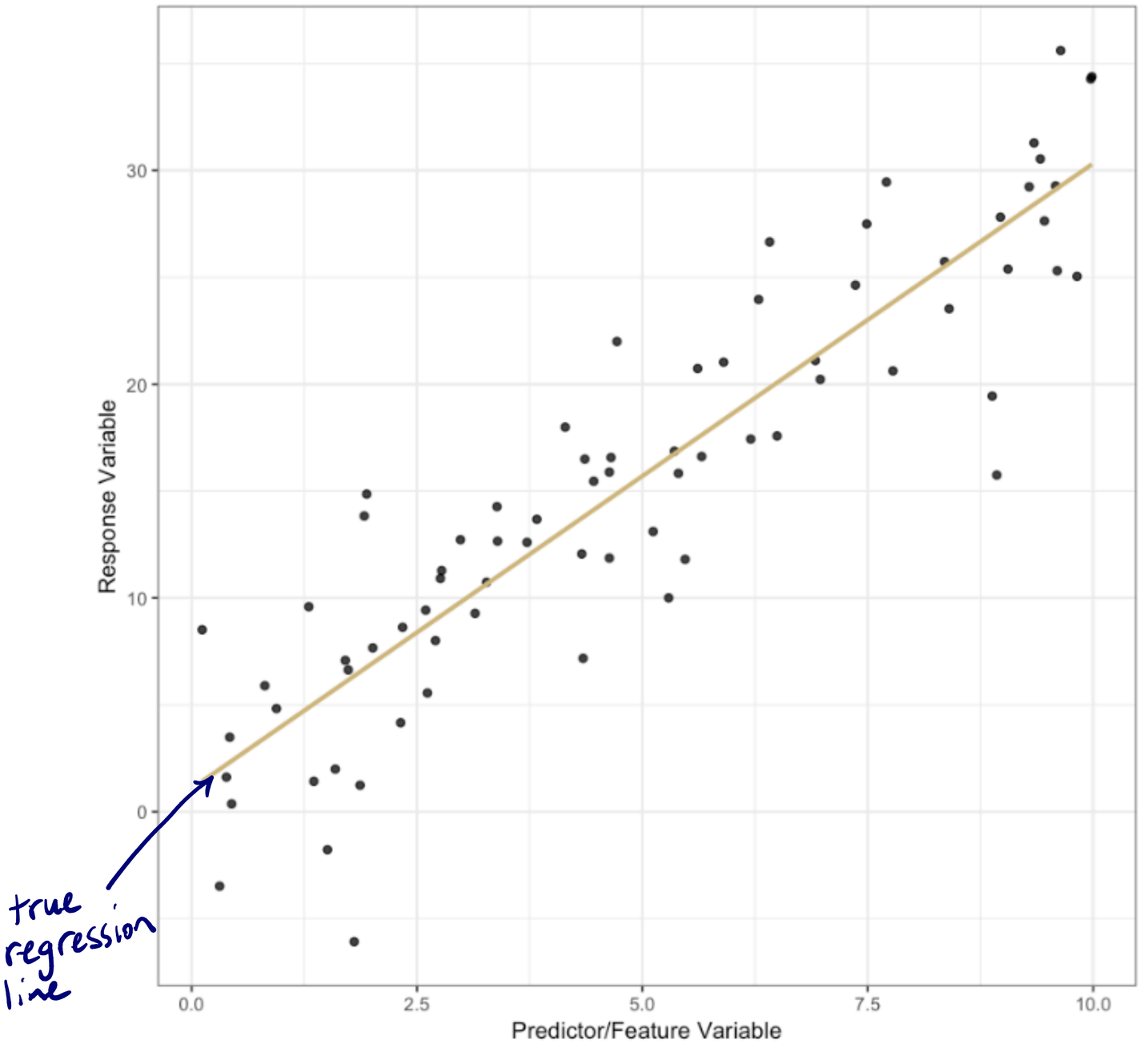


Photo by Jason Leung on Unsplash

Photo by Jr Korpa on Unsplash

Interpreting the Regression Parameters

Interpreting *simple* linear regression parameters:

1. β_0: the intercept of the true regression line
2. β_1: the slope of the true regression line.

Photo by Jr Korpa on Unsplash

Interpreting the Regression Parameters

Interpreting *simple* linear regression parameters:

$Y = \beta_0 + \beta_1 X + \varepsilon$, $E(\varepsilon) = 0 \Rightarrow E(Y) = \beta_0 + \beta_1 X$

1. β_0 *is the average value of* Y *when x is zero.* Usually this is called the "baseline average".

2. β_1: *is the average change in* Y *associated with a 1-unit increase in the value of* x.

Interpreting the Regression Parameters

Interpreting *multiple* linear regression parameters:

1. β_0: the intercept of the true regression line. We interpret this as *the average value of* Y *when all of the* x*'s are zero.*

2. β_j: the slope of the true regression line, $j = 1,...,p$.

Interpreting the Regression Parameters

Interpreting *multiple* linear regression parameters:

1. β_0 *is the average value of* Y *when all of the* x*'s are zero.*

2. β_j *is the average change in* Y *associated with a 1-unit increase in the value of* x_j *assuming all other predictors are held constant,* $j = 1,...,p$.

Thus, these "slope" parameters are called *partial* or *adjusted* regression parameters/coefficients.

Lesson: The appropriateness of linear regression

Module: Introduction to statistical models

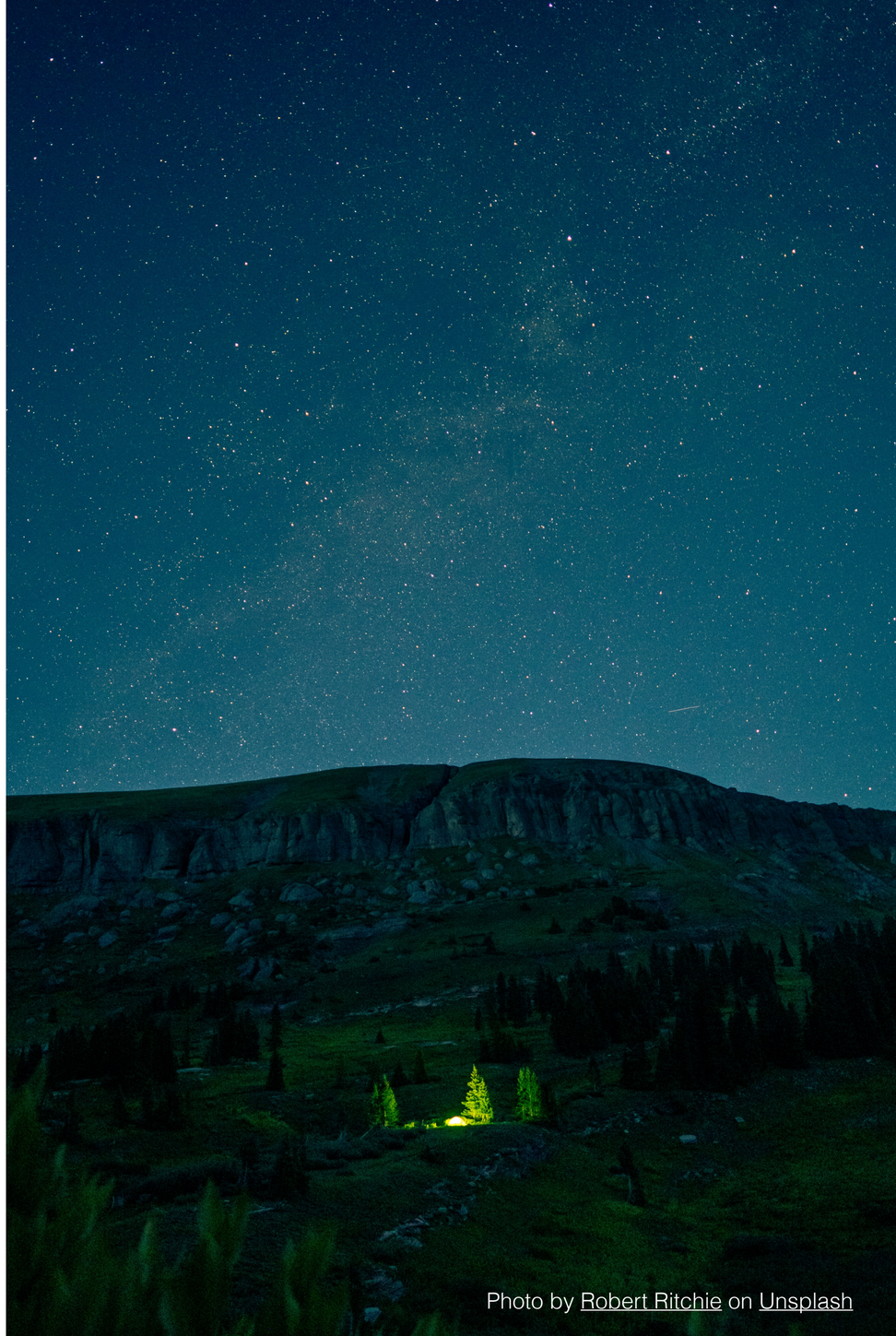


Photo by Robert Ritchie on Unsplash

How do we know linear regression is appropriate?

Photo by Jason Leung on Unsplash

Does a theory or “law” tell us how the variables relate?

Photo by Jason Leung on Unsplash

Photo by James Allen on Unsplash

Does a theory or “law” tell us how the variables relate?

$$F = kx$$

Force

Spring Constant

Displacement

What past studies related to this research question have been conducted?

Photo by Jason Leung on Unsplash

Can we explore the data to learn whether there are linear relationships between variables?

Photo by Jason Leung on Unsplash

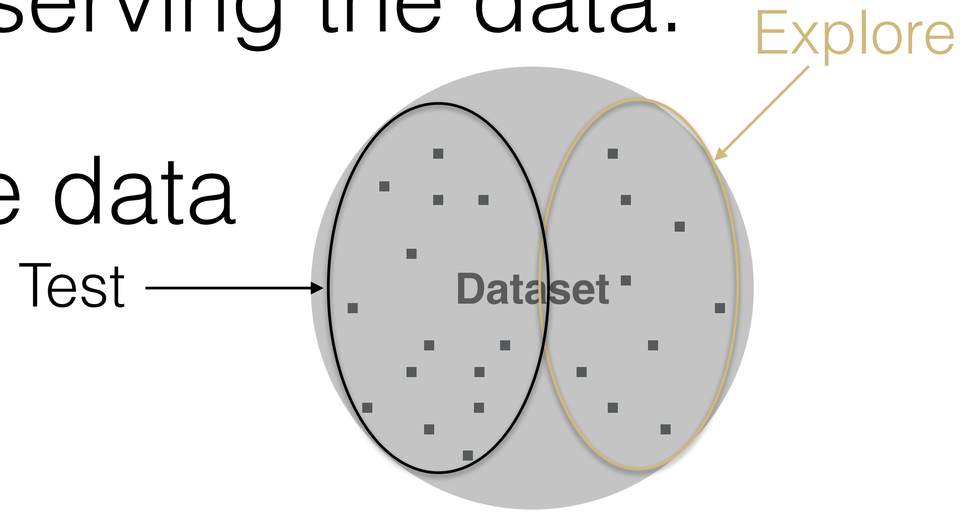
Definition: A *circular analysis* or *double dipping* is the process of exploring a dataset in an attempt to discover what relationships exist, and then test hypotheses related to that exploration on the same dataset.

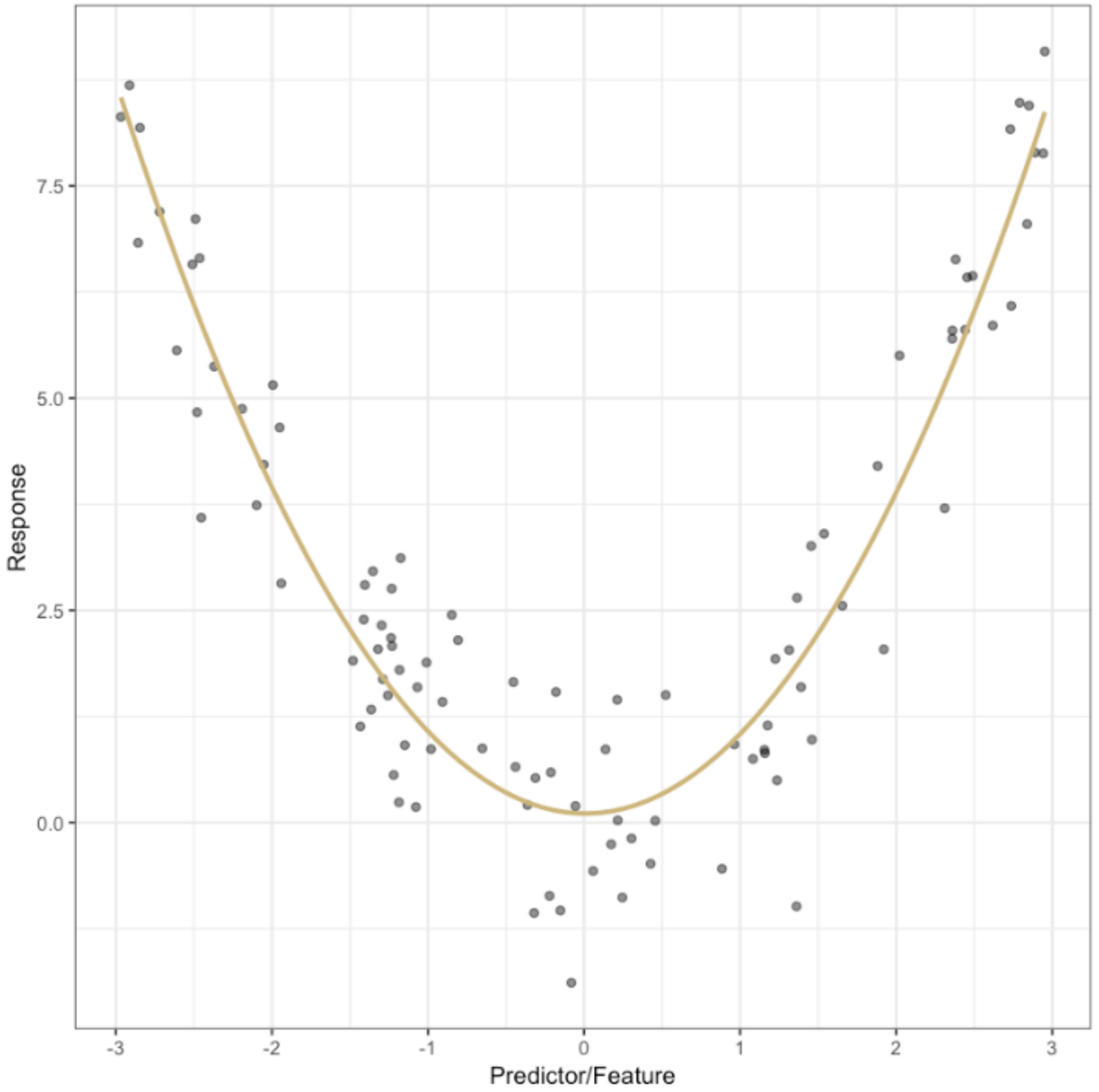
Photo by Jason Leung on Unsplash

Photo by Jason Leung on Unsplash

Ways to avoid circular analyses:

1. Design the analysis and pre-specify research hypotheses before observing the data.
2. Subset the data





Nonlinear relationships between the predictors and the response?