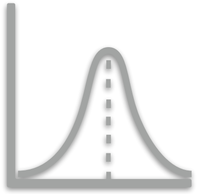
Lesson: Introduction to least squares estimation

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash

realization from a (normal) random variable



$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

measured response

Unknown, to be estimated

measured predictors

$$\varepsilon_i \sim N(0, \sigma^2)$$

Estimating Model Parameters

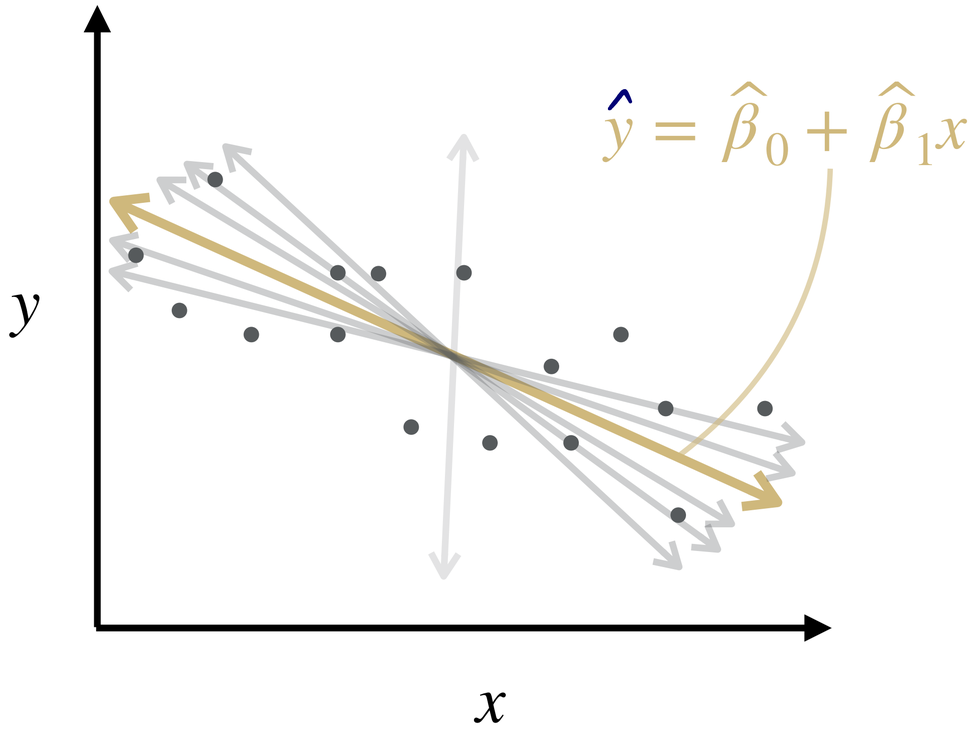
The line of best fit to the data is the line that minimizes the sum of the squared vertical distances between the line and the observed points:

$$\arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \left(y_i - [\beta_0 + \beta_1 x_i] \right)^2$$

measured response

Unknown, to be estimated

measured response



Estimating Model Parameters

The regression model partitions the response into a systematic component and a random component.

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times(p+1)}{X}\ \underset{(p+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

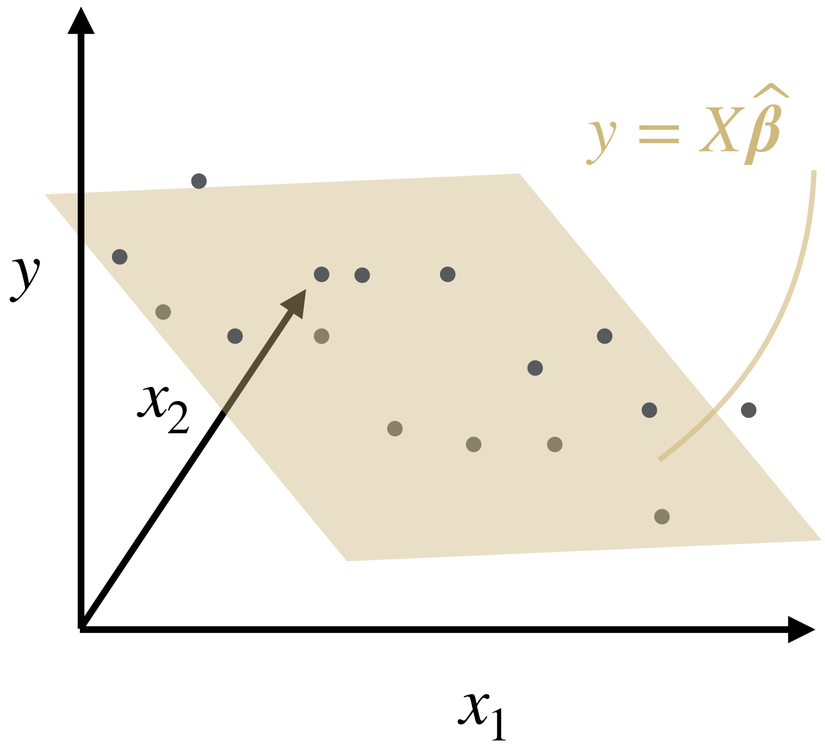
measured response

measured predictors

Unknown, to be estimated

The problem is the find a $\boldsymbol{\beta}$ so that $X\boldsymbol{\beta}$ is as close as possible to y.

Estimating Model Parameters



The *surface* of best fit to the data is the surface that minimizes the sum of the squared vertical distances between the surface and the observed points:

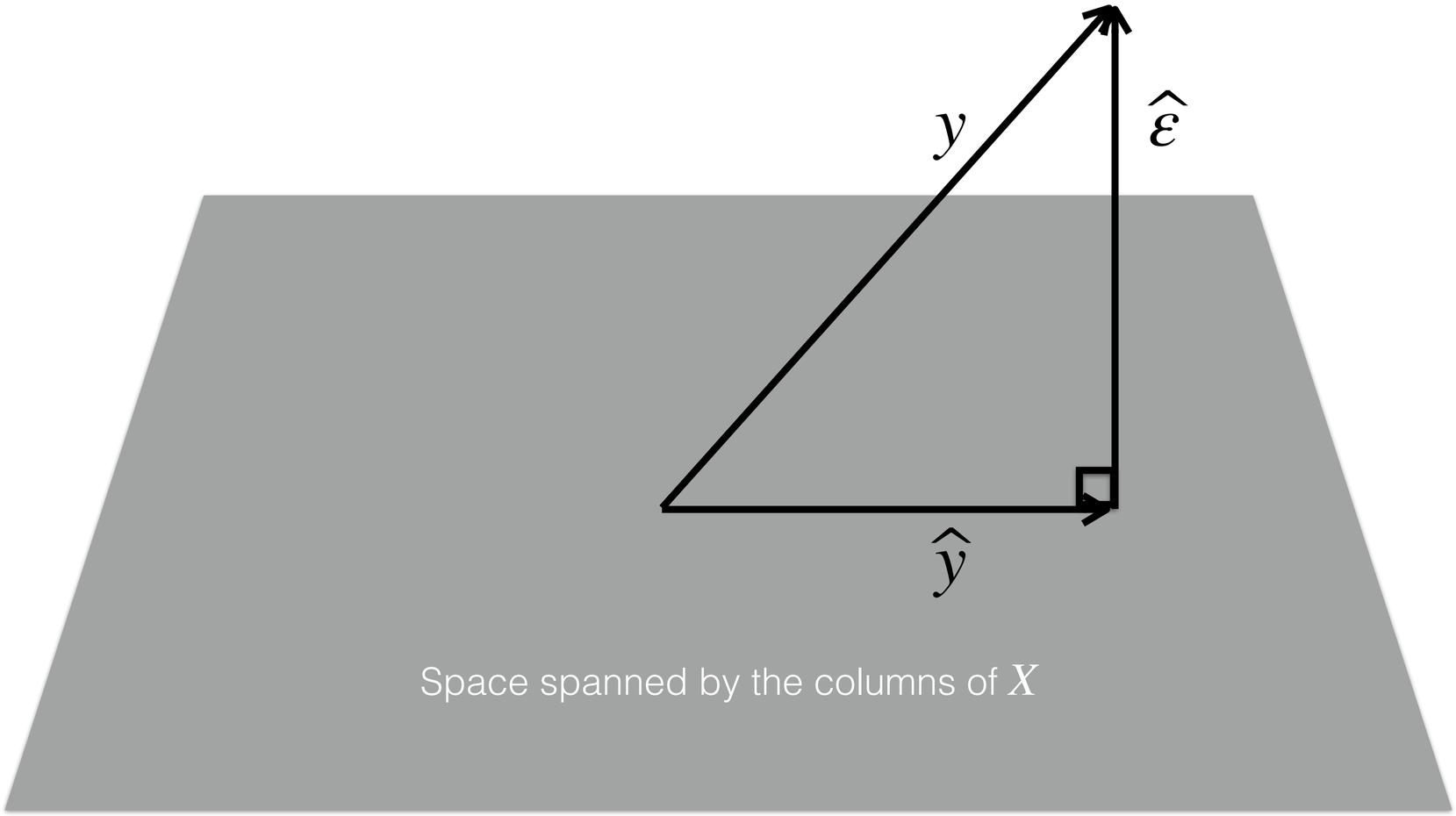
$$\arg\min_{\beta} \|\mathbf{Y} - X\boldsymbol{\beta}\|^2 = \underbrace{\left(\mathbf{Y} - X\boldsymbol{\beta}\right)^T}\left(\mathbf{Y} - X\boldsymbol{\beta}\right)$$

$$\underline{Y}^T - \underline{\beta}^T X^T)(\underline{Y} - X\underline{\beta})$$

$$\ldots = \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip})\right)^2$$

The problem is the find a $\boldsymbol{\beta}$ so that $X\boldsymbol{\beta}$ is as close as possible to y.

Estimating Model Parameters



Lesson: Linear algebra for least squares

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash

Let X be an $m \times n$ matrix, $\mathbf{v}$ be $n \times 1$, and $\mathbf{y}$ be $m \times 1$. Then:

$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & & \ddots & \\ x_{m1} & \cdots & & x_{mn} \end{pmatrix}$

1. *Lemma 1*: Then $X^T X$ is symmetric, i.e., $\left(X^T X\right)^T = X^T X$.

$n \times m \quad m \times n$

$(X^T X)^T = X^T (X^T)^T = X^T X$

2. *Lemma 2*: Let $\mathbf{y} = X\mathbf{v}$. Then $\dfrac{\partial y}{\partial \mathbf{v}} = X$ and $\dfrac{\partial y^T}{\partial \mathbf{v}} = X^T$

$y_i = x_{i1} v_1 + x_{i2} v_2 + \cdots + x_{in} v_n$

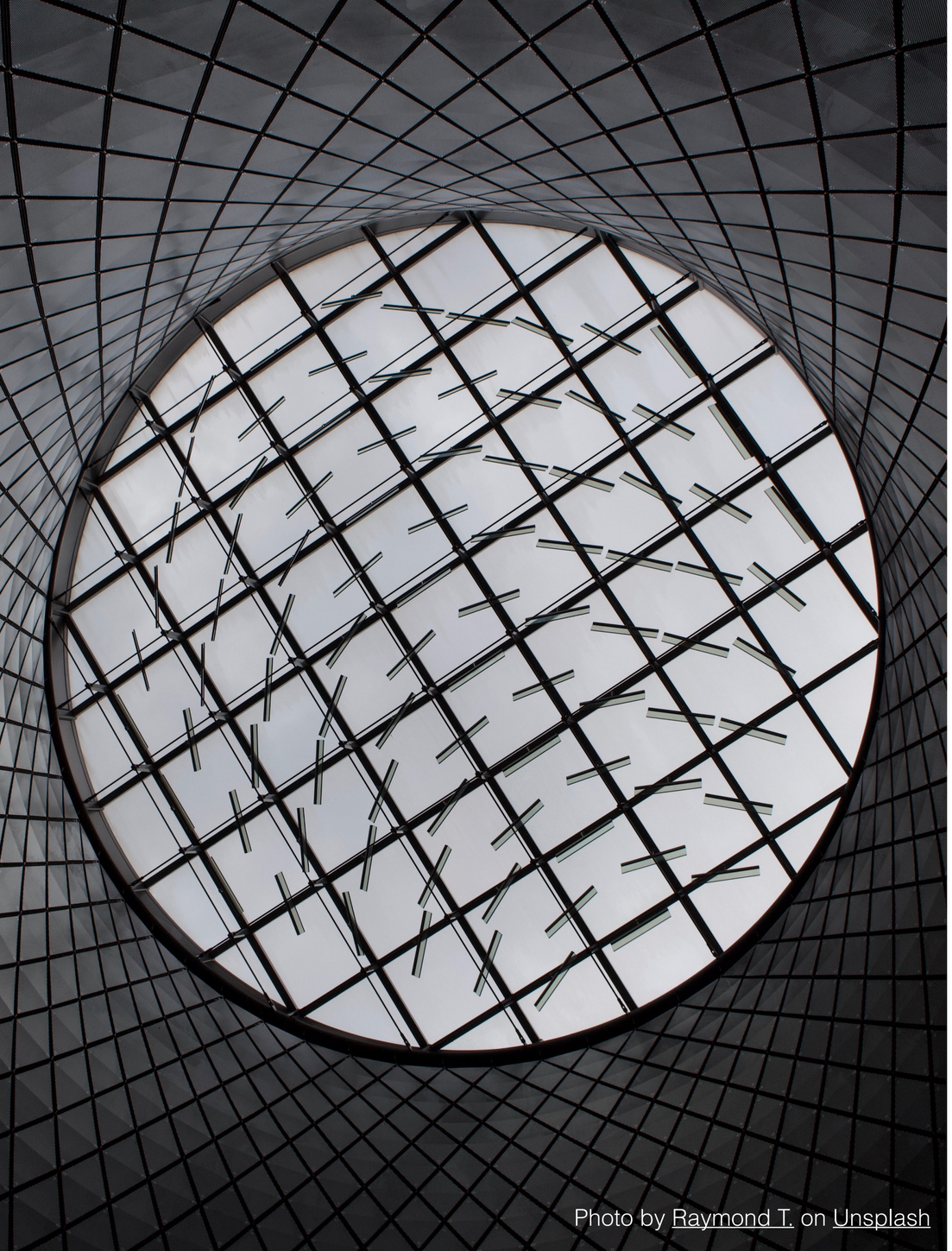
3. *Lemma 3*: Let $c = \mathbf{v}^T \left(X^T X\right) \mathbf{v}$. Then $\dfrac{\partial c}{\partial \mathbf{v}} = 2X^T X \mathbf{v}$

$\left(f(x) = 5x^2 \quad f'(x) = 2 \cdot 5x \right)$

Lesson: Deriving the least squares solution

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash



Definition: The *residuals* are defined as:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})$$

$$\hat{\underline{\varepsilon}} = (\underline{y} - \hat{\underline{y}})$$

Definition: The *fitted values* are defined as:

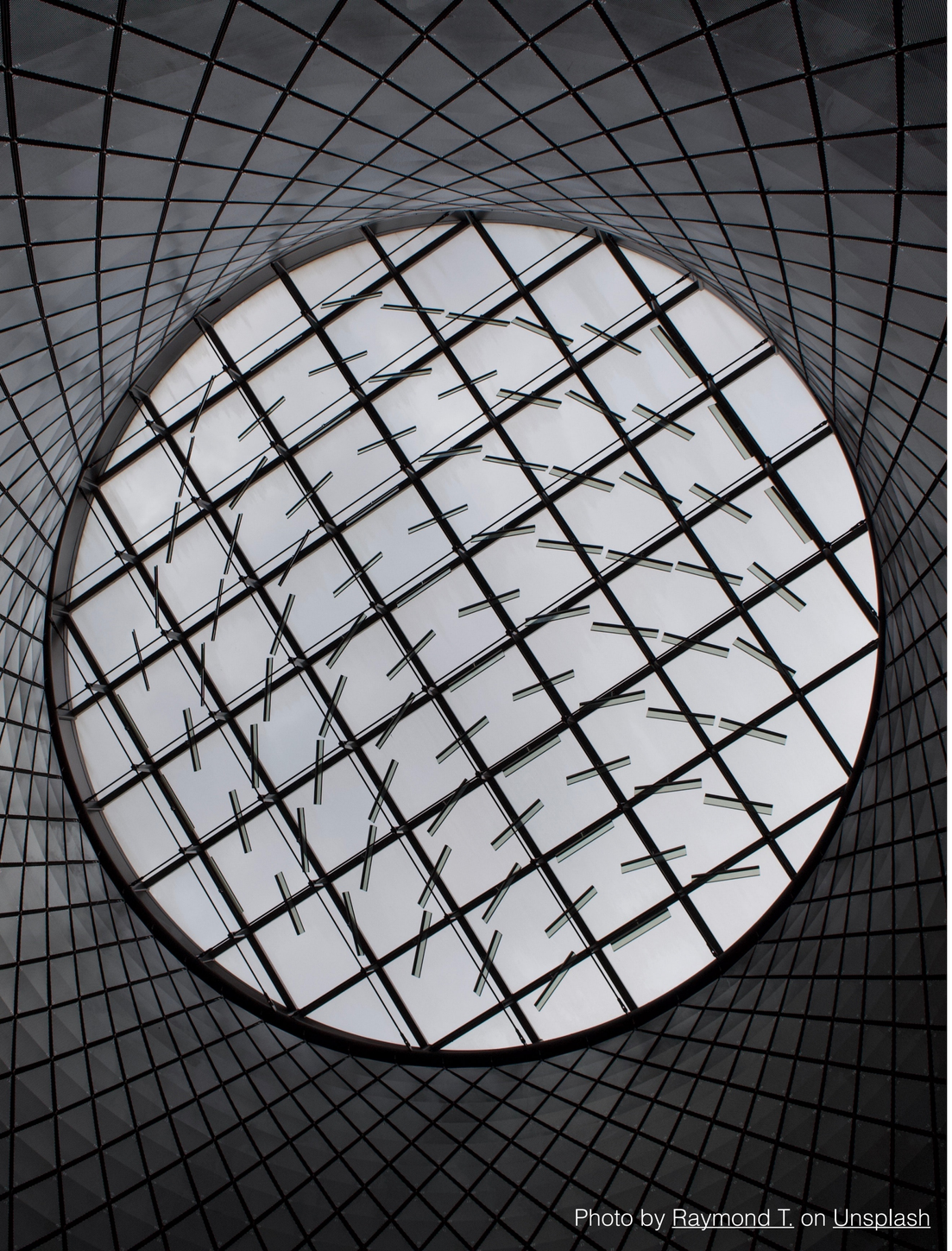
$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

as a function of β_j

Definition: The *hat matrix*, *H*, is defined as:

$$H = X(X^T X)^{-1} X^T$$

The hat matrix is useful in theoretical calculations related to least squares



Least Squares Estimation: We define the best estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ as the one that minimized the sum of the squared residuals: $RSS = (\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta})$

$= (\underline{y}^T - \underline{\beta}^T X^T)(\underline{y} - X\underline{\beta}) = \underline{y}^T\underline{y} - \underline{y}^T X\underline{\beta} - \underline{\beta}^T X^T\underline{y} + \underline{\beta}^T X^T X\underline{\beta}$

$= \underline{y}^T\underline{y} - 2\underline{\beta}^T X^T\underline{y} + \underline{\beta}^T X^T X\underline{\beta}$

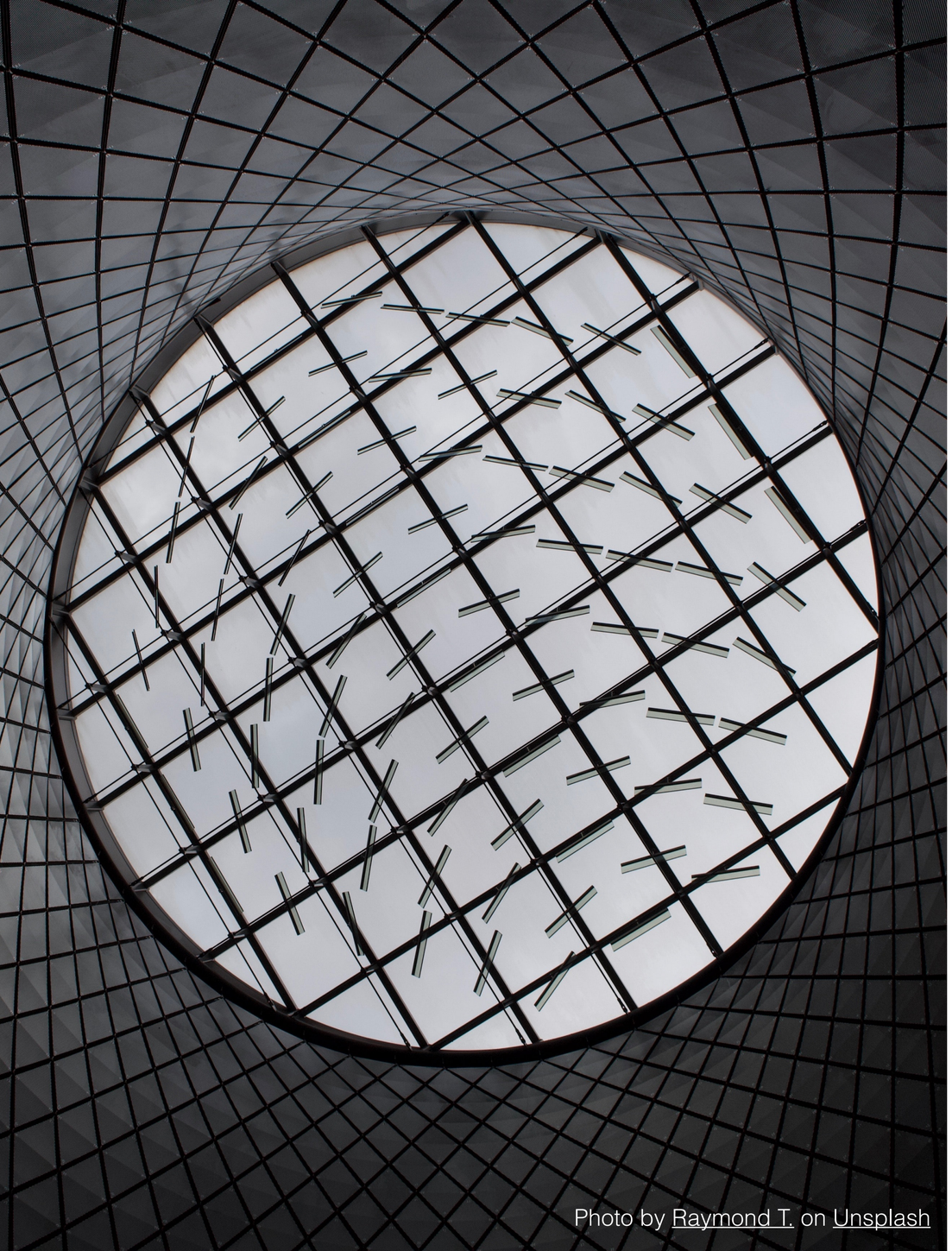
Differentiating with respect to $\boldsymbol{\beta}$, we get:

$$\frac{2RSS}{2\underline{\beta}} = 0 - 2X^T\underline{y} + 2X^TX\underline{\beta} \overset{set}{=} 0$$

$$\Rightarrow X^TX\underline{\beta} = X^T\underline{y}$$

$$\Rightarrow \hat{\underline{\beta}} = (X^TX)^{-1}X^T\underline{y} \quad \text{when } (X^TX)^{-1} \text{ exists}$$

$$(\underline{y}^T X\underline{\beta})^T = \underline{\beta}^T X^T\underline{y}$$



In order to use least squares, we assume that:

1. $E\left(\varepsilon_i\right) = 0$ for all $i = 1,\ldots,n$.

2. $E\left(Y_i\right) = \mathbf{x_i}^T\boldsymbol{\beta}$ for all $i = 1,\ldots,n$.

$\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$

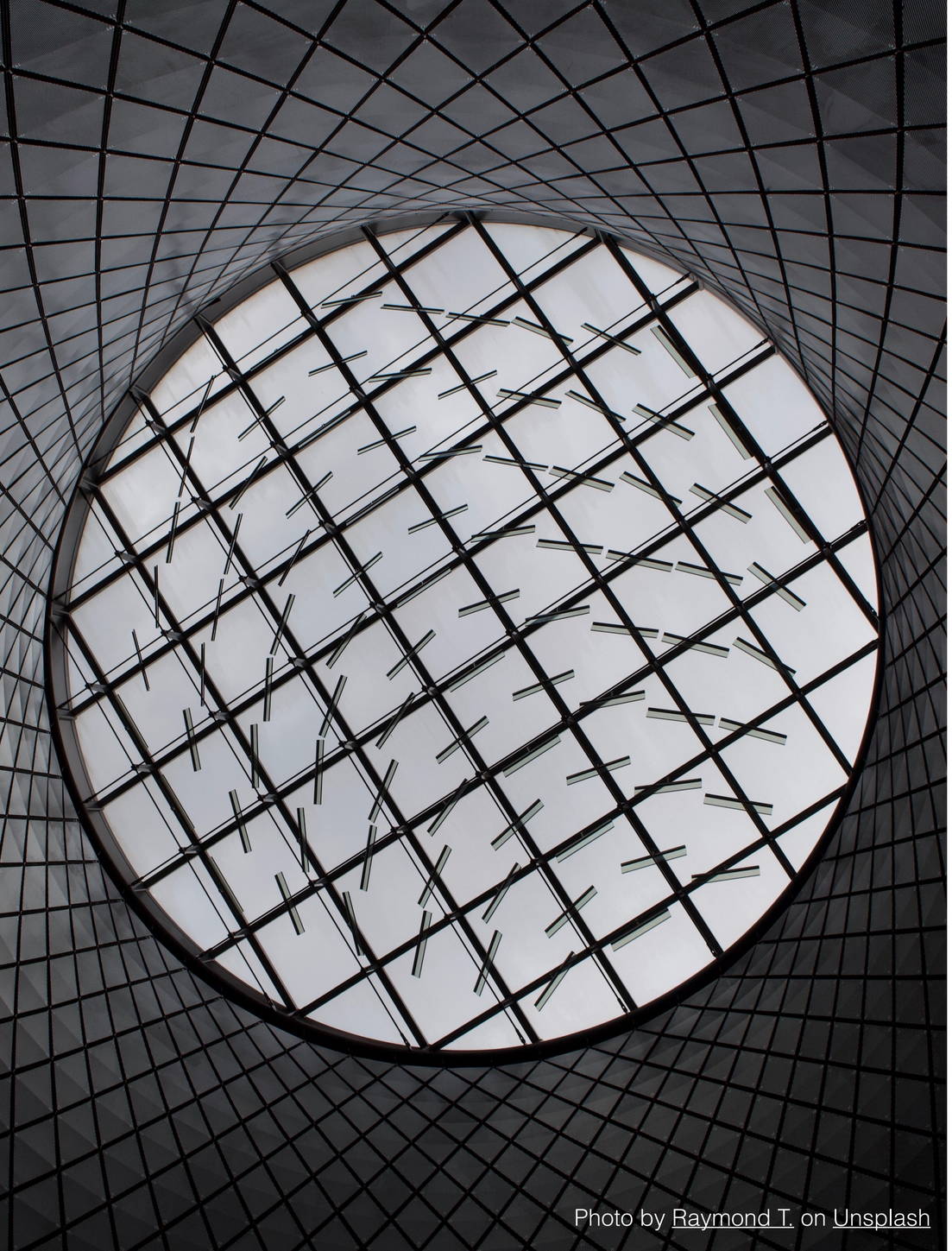
3. $Cov\left(\varepsilon_i, \varepsilon_j\right) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$

4. $\left(X^T X\right)^{-1}$ exists.

Lesson: Justifying least squares: the Gauss-Markov theorem and maximum likelihood estimation

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash



The Gauss-Markov Theorem: Suppose that:

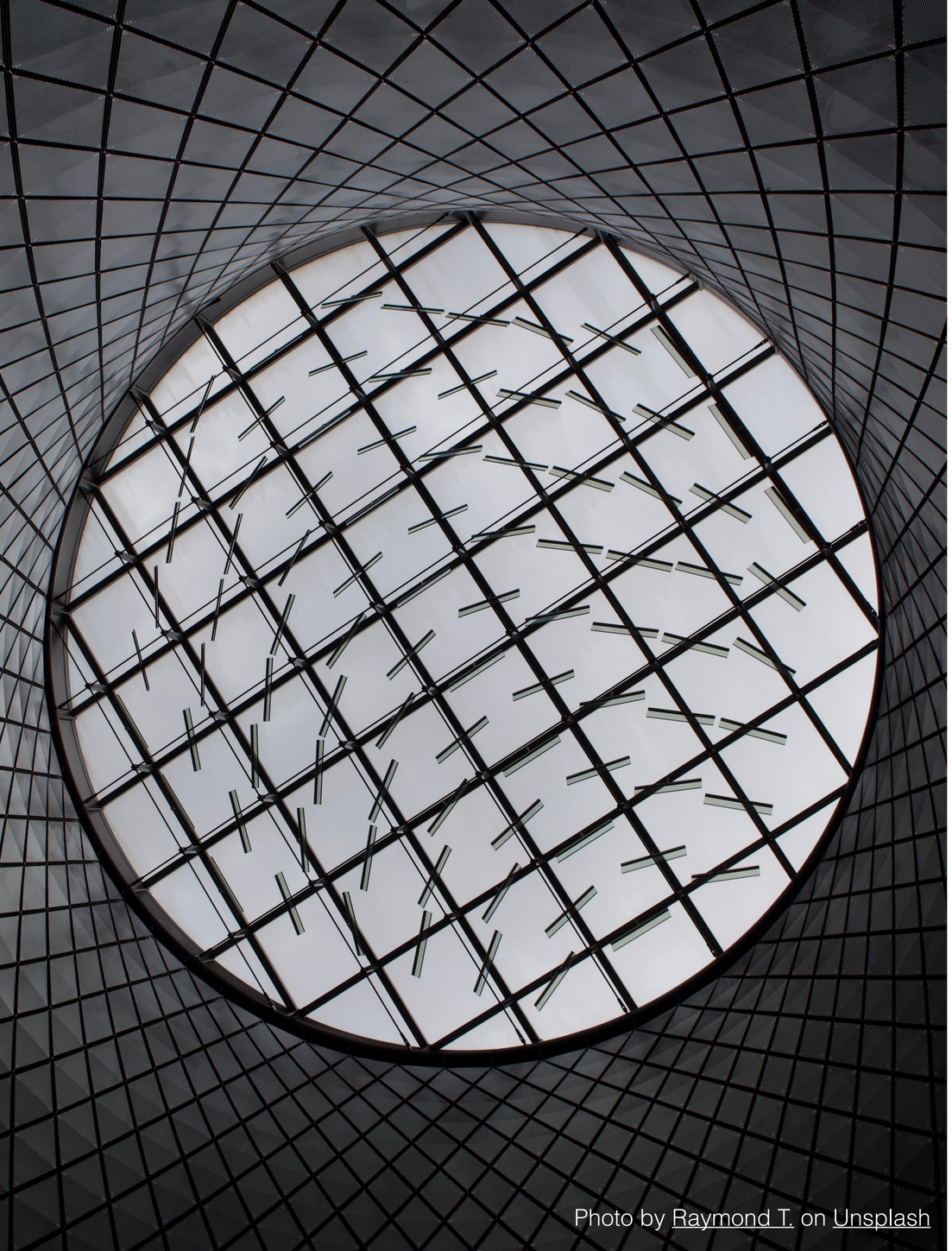
1. $E\left(\varepsilon_i\right) = 0$ for all $i = 1,...,n$.

2. $E\left(Y_i\right) = \mathbf{x_i}^T\boldsymbol{\beta}$ for all $i = 1,...,n$.

3. $Cov\left(\varepsilon_i, \varepsilon_j\right) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$

4. $\left(X^TX\right)^{-1}$ exists.

Then $\widehat{\boldsymbol{\beta}}$ is the "best" *unbiased* estimator of $\boldsymbol{\beta}$.



The maximum likelihood estimator.

Suppose that $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$. Then:

$\mu_y = \beta_0 + \beta_1 x_{i1} + \beta_p x_{ip}$

1. marginal pdf:

$$f(y_i;\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu_y)^2\right\}$$

2. joint pdf:

$$f(y;\beta) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_y)^2\right\}$$

3. log-likelihood:

$$\ell(\beta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i-\mu_y)^2$$

RSS

Lesson: Sums of squares and estimating the error variance

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash

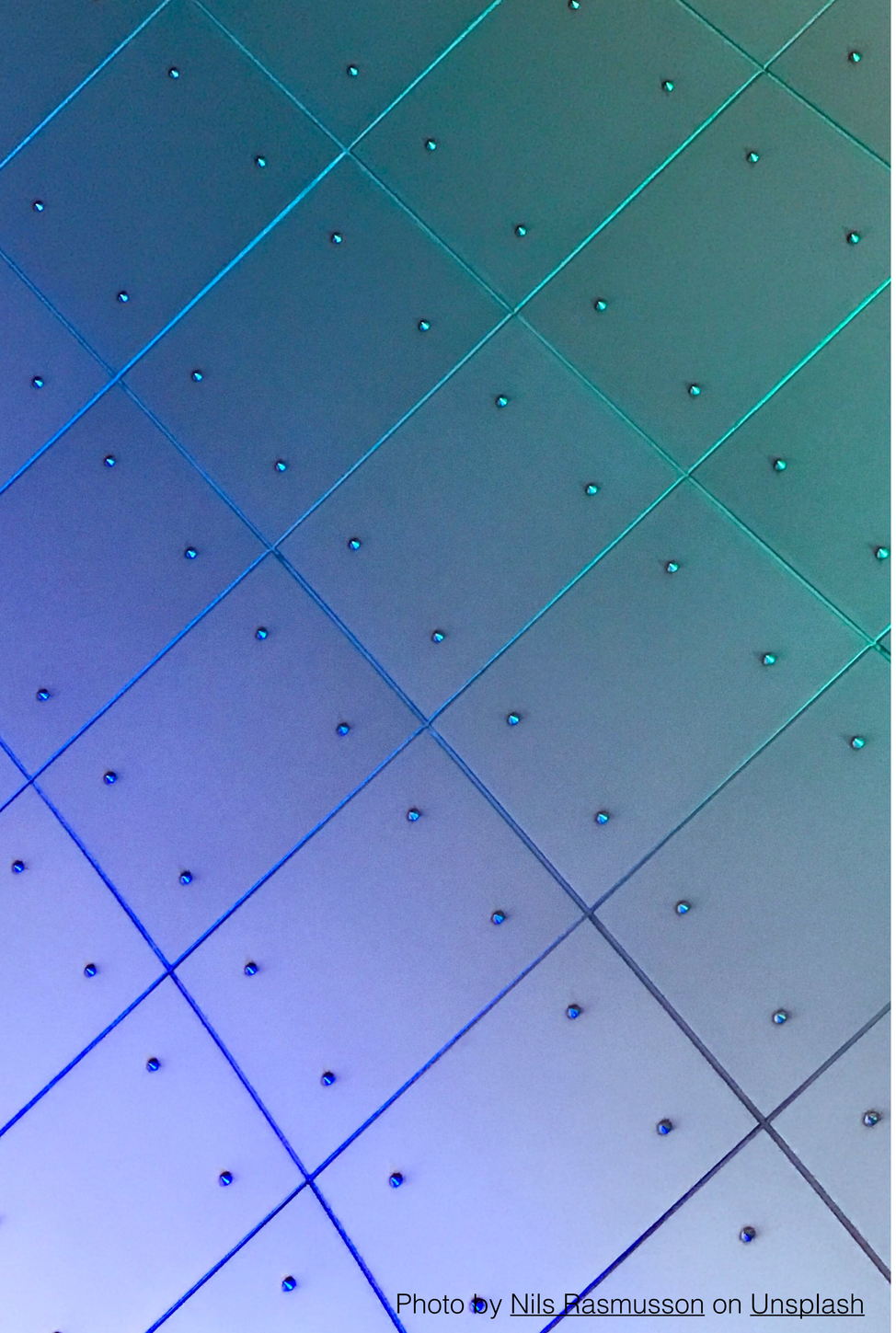
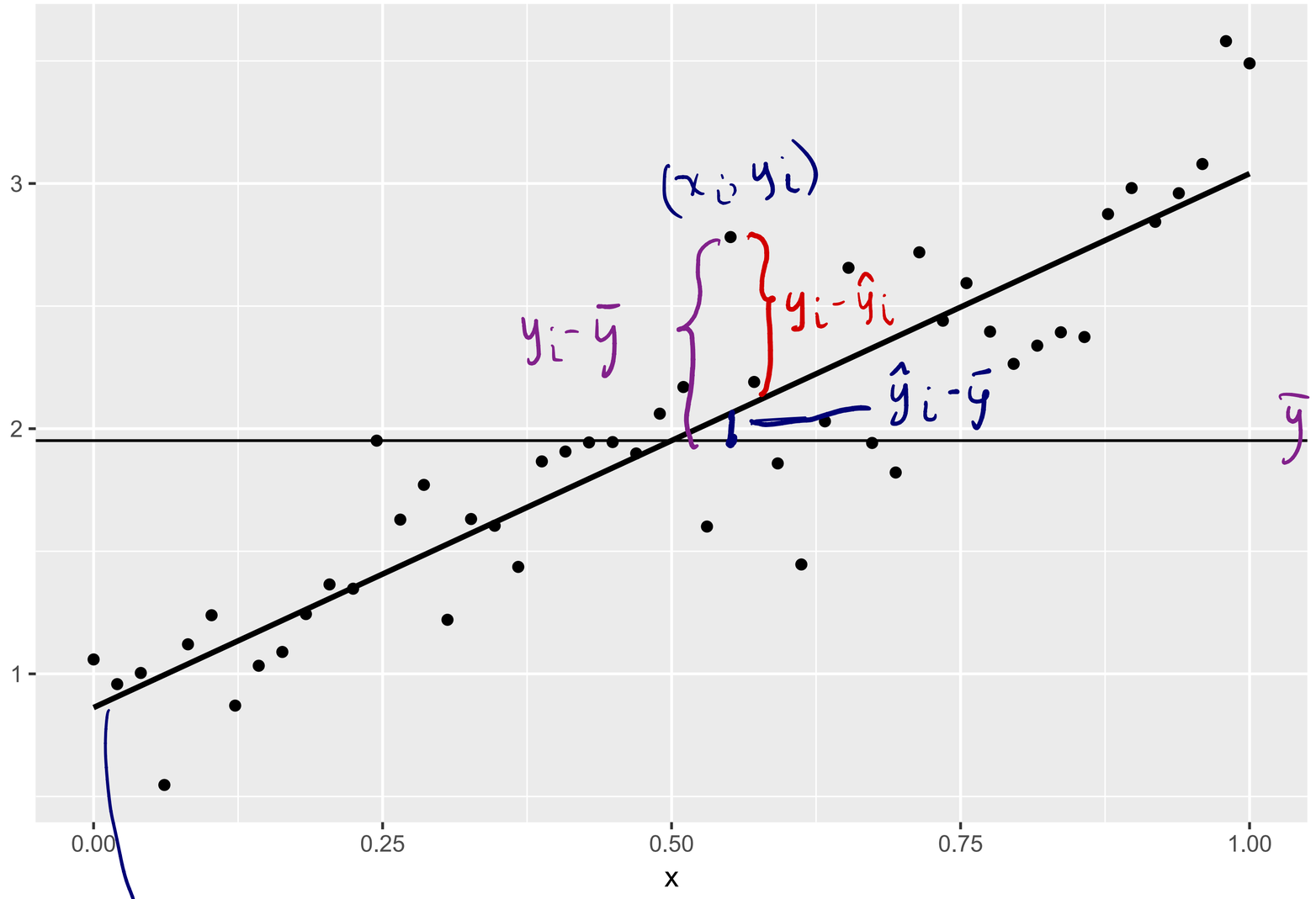


Photo by Nils Rasmusson on Unsplash



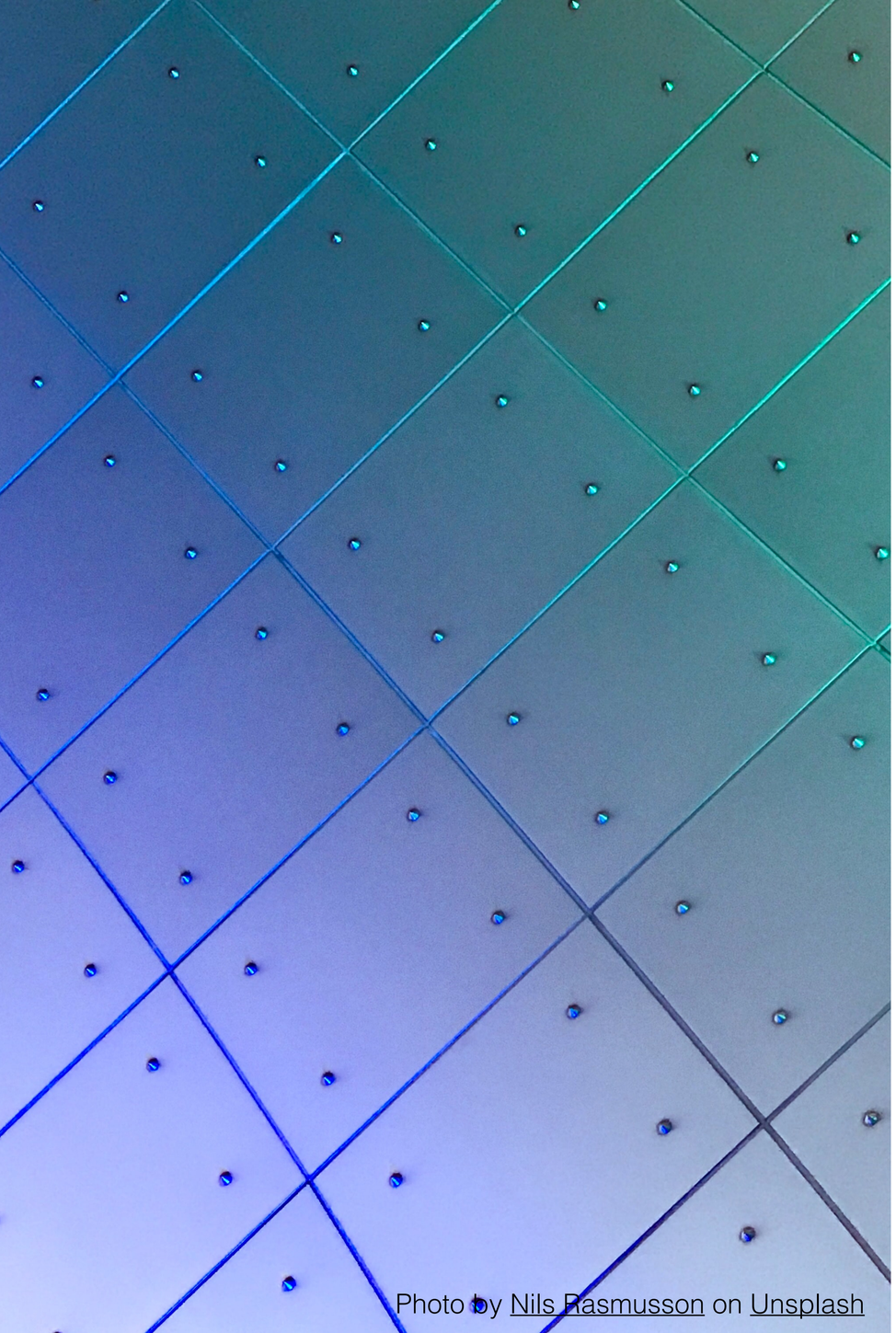


Photo by Nils Rasmusson on Unsplash

Sums of squares:

1. RSS: Residual sum of squares:

$$RSS = \sum_{i=1}^{n} \underbrace{\left(y_i - \hat{y}_i\right)}_{\hat{\varepsilon}_i}^2$$

2. ESS: *Explained* (or regression) sum of squares:

$$ESS = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$$

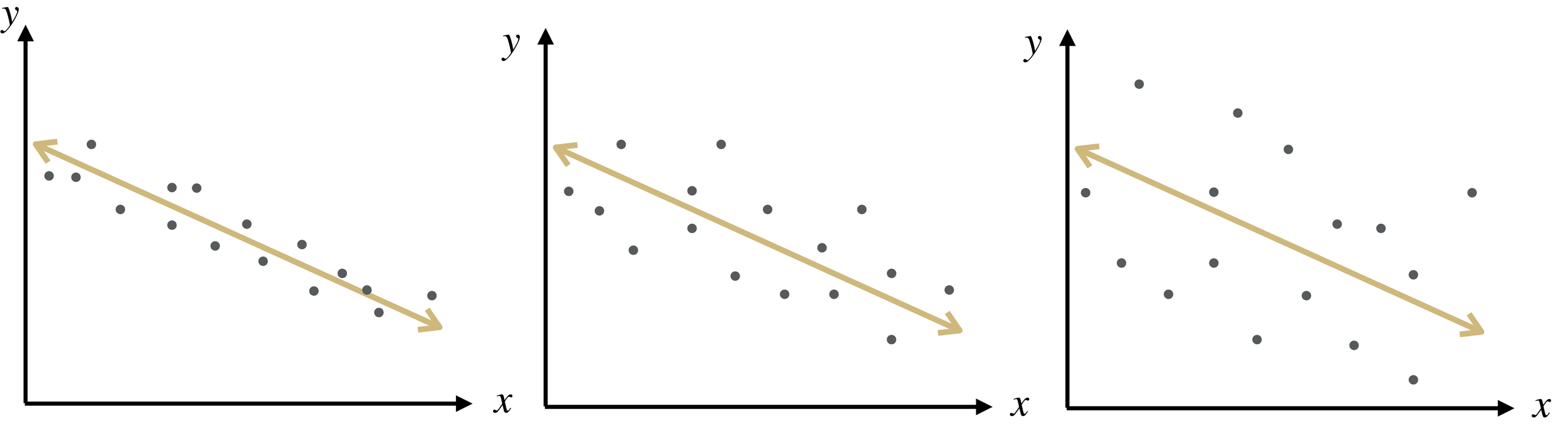
3. TSS: *Total* sum of squares:

$$TSS = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$$

$$\boxed{TSS = ESS + RSS}$$

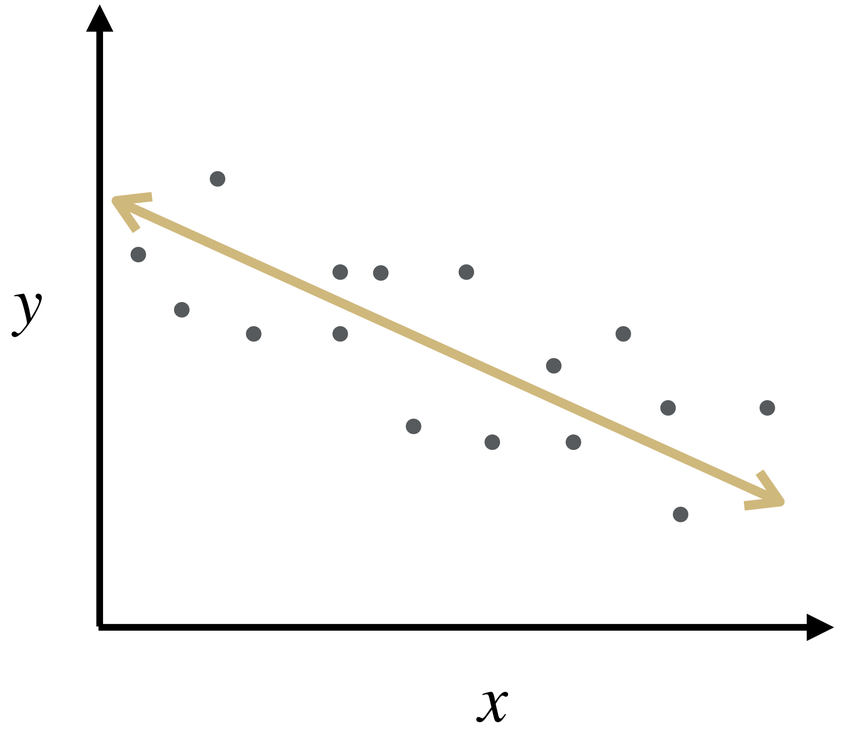
Sums of Squares

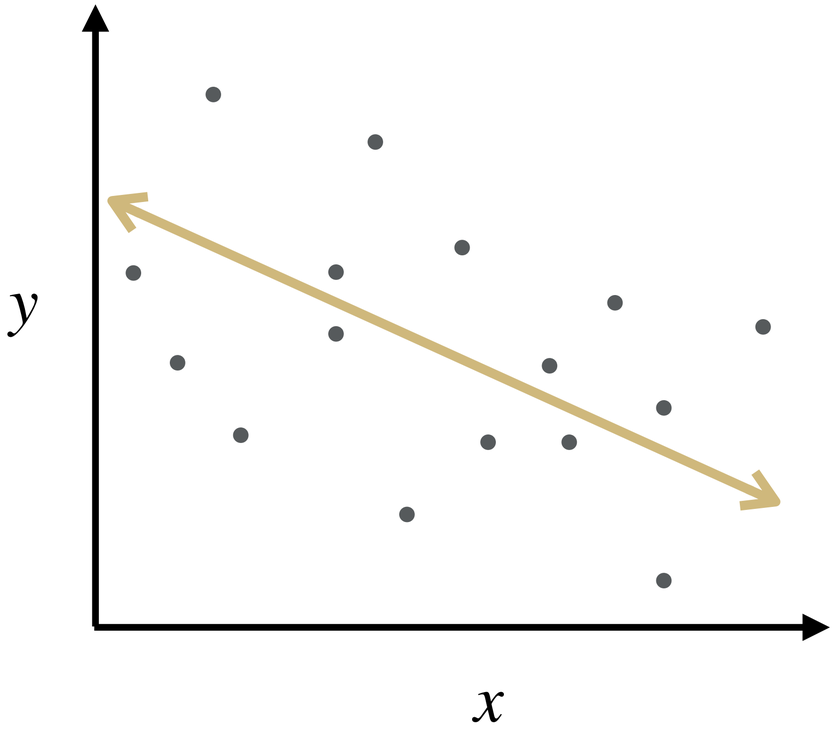
The residual sum of squares RSS can be interpreted as a measure of how much variation in y is left *unexplained by the model*—that is, how much cannot be attributed to a linear relationship.



Estimating σ^2

The parameter σ^2 determines the amount of spread about the true regression line. Two separate examples:





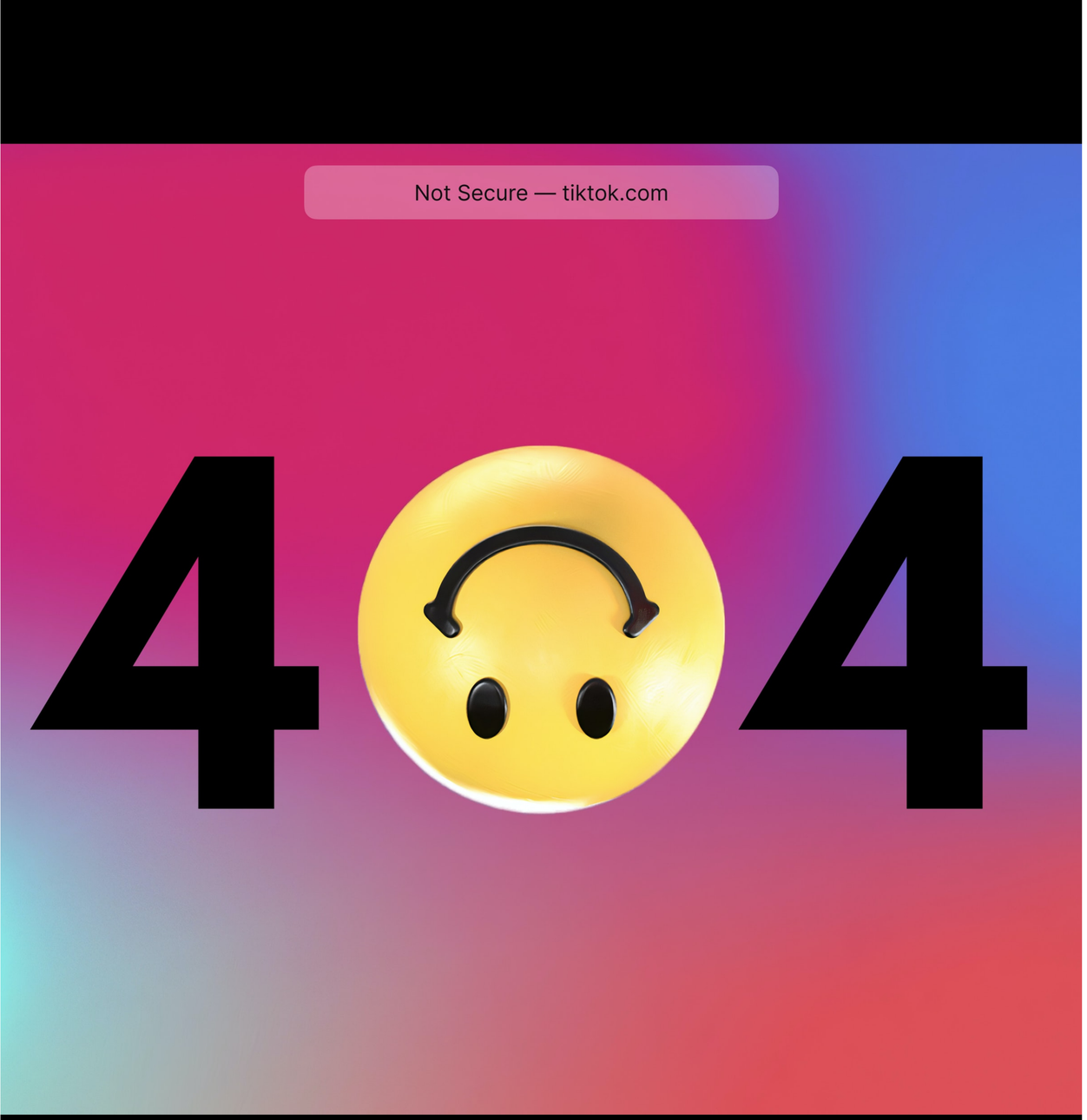


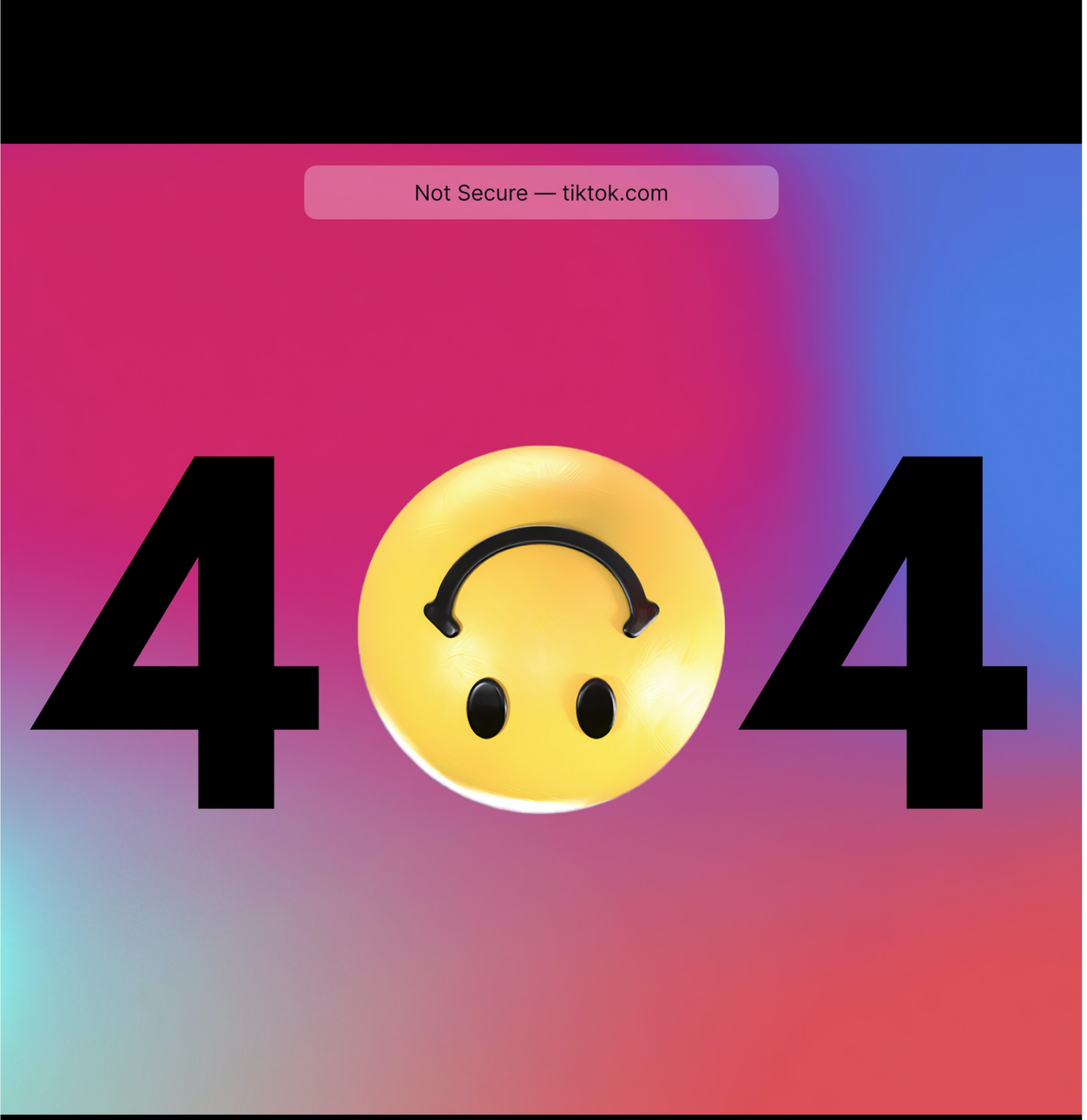
Photo by visuals on Unsplash

An estimate of σ^2 will be used in statistical inference (e.g., confidence interval formulas and hypothesis testing), presented in the next two sections.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Given our interpretation of the RSS, it seems plausible that contribute to an estimator of σ^2:

$$\hat{\sigma}^2 = \frac{RSS}{n-(p+1)}$$



Note:

1. The divisor $n-(p+1)$ in is the number of degrees of freedom (df) associated with RSS and $\hat{\sigma}^2$.

2. The RSS has $n-(p+1)$ df because $p+1$ parameters must first be estimated to compute it, which results in a loss of $p+1$ df.

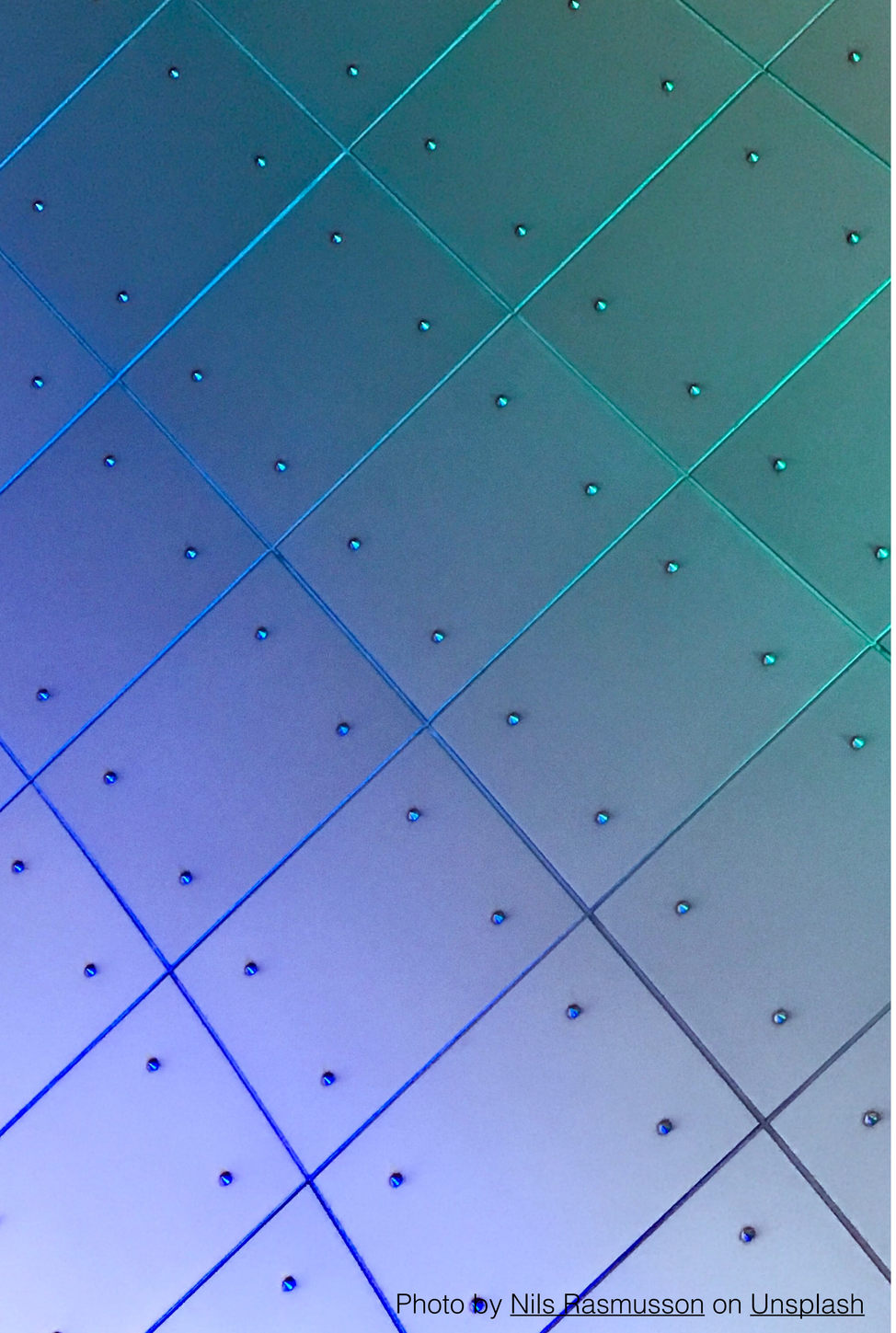
3. Replacing each y_i in the formula for $\hat{\sigma}^2$ by the r.v. Y_i gives a random variable.

4. It can be shown that the r.v. $\hat{\sigma}^2$ is an *unbiased estimator* for $\hat{\sigma}^2$.

Lesson: The coefficient of determination, "R-squared"

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash



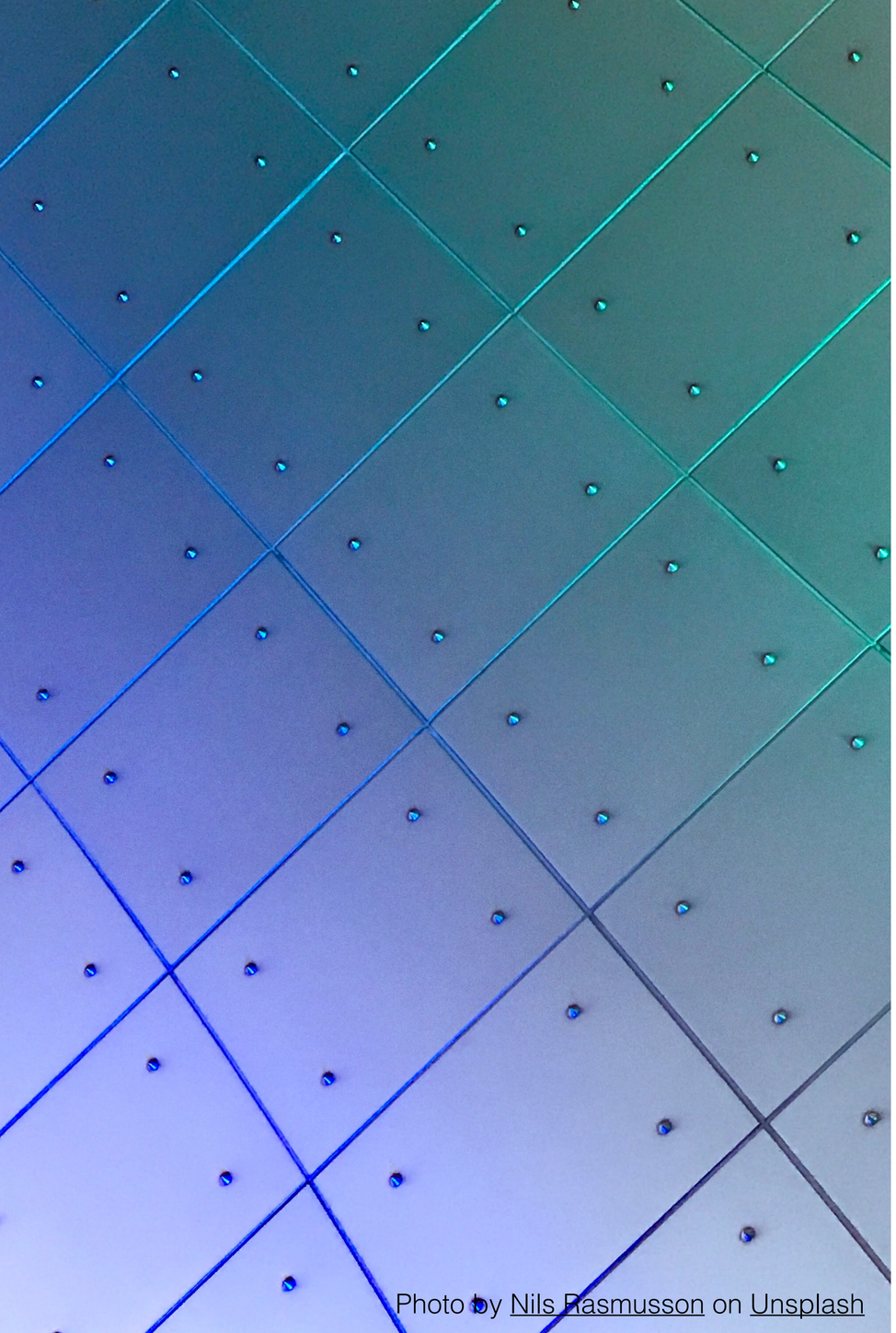
$TSS = RSS + ESS$

The *coefficient of determination*, R^2, is defined as:

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$

Note:

- $0 \leq R^2 \leq 1$
- Assuming that the model is correct, R^2 is interpreted as the proportion of observed variation in y explained by the model.



Warnings about R^2:

1. R^2 can be close to 1 but the model is the wrong fit for the data.
2. R^2 can be close to 0 even when the model is the correct fit for the data.
3. R^2 should **not** be used to compare models with a different number of predictors.
4. R^2 says nothing about the causal relationship between the predictors and the response.

Lesson: The problem of non-identifiability

Module: Linear regression parameter estimation

Photo by Ryan Hoffman on Unsplash

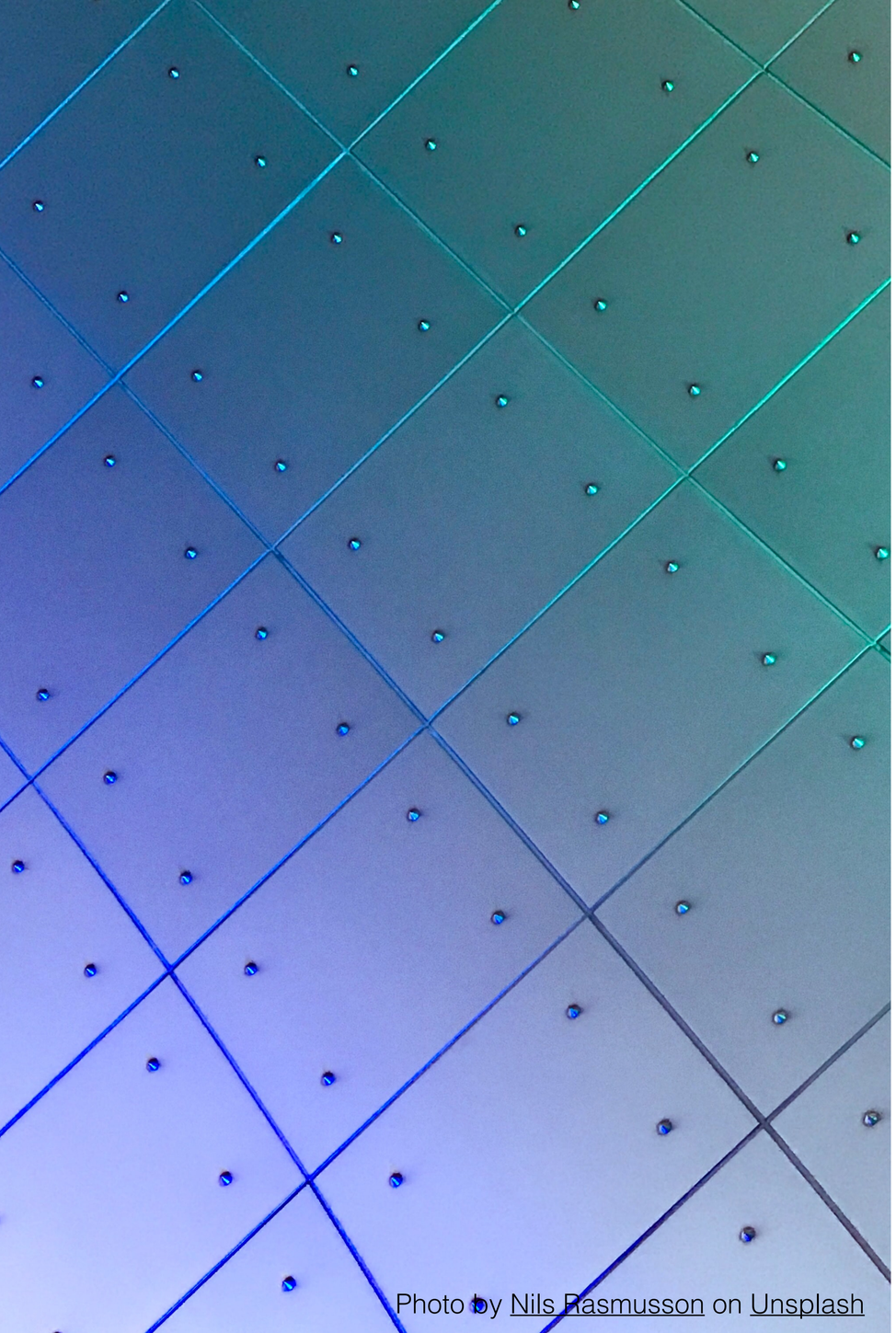


Photo by Nils Rasmusson on Unsplash

The least squares estimate is the solution to the *normal* equations:

$$X^T X \boldsymbol{\beta} = X^T \mathbf{Y}.$$

1. When $\left(X^T X\right)^{-1}$ exists, there is a unique solution, $\widehat{\boldsymbol{\beta}}$.

2. When $\left(X^T X\right)^{-1}$ does not exist, there will be infinitely many solutions.

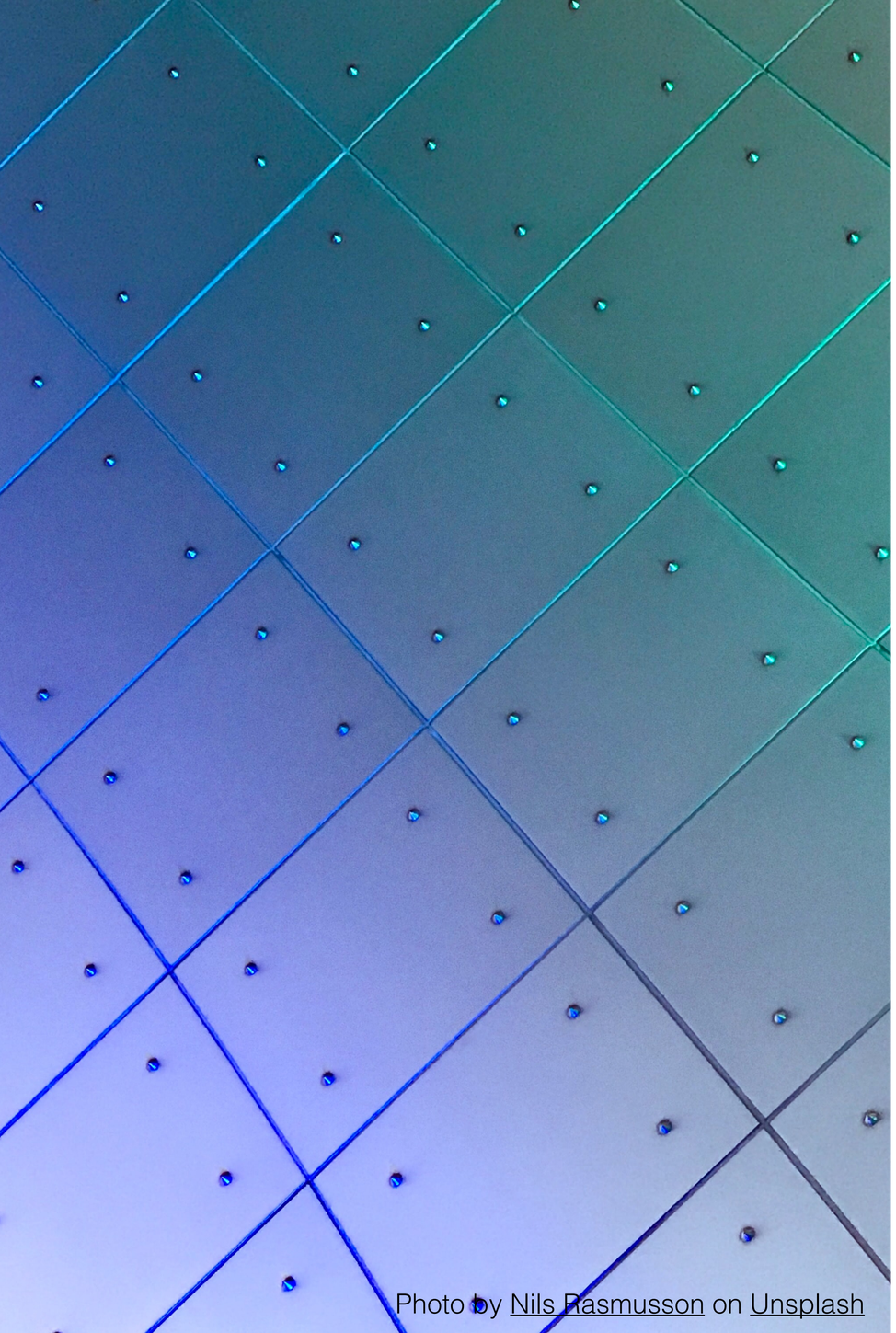


Photo by Nils Rasmusson on Unsplash

Definition: When $\left(X^T X\right)^{-1}$ does not exist, the regression model is said to be *non-identifiable* (or, *unidentifiable*).

When does this happen?

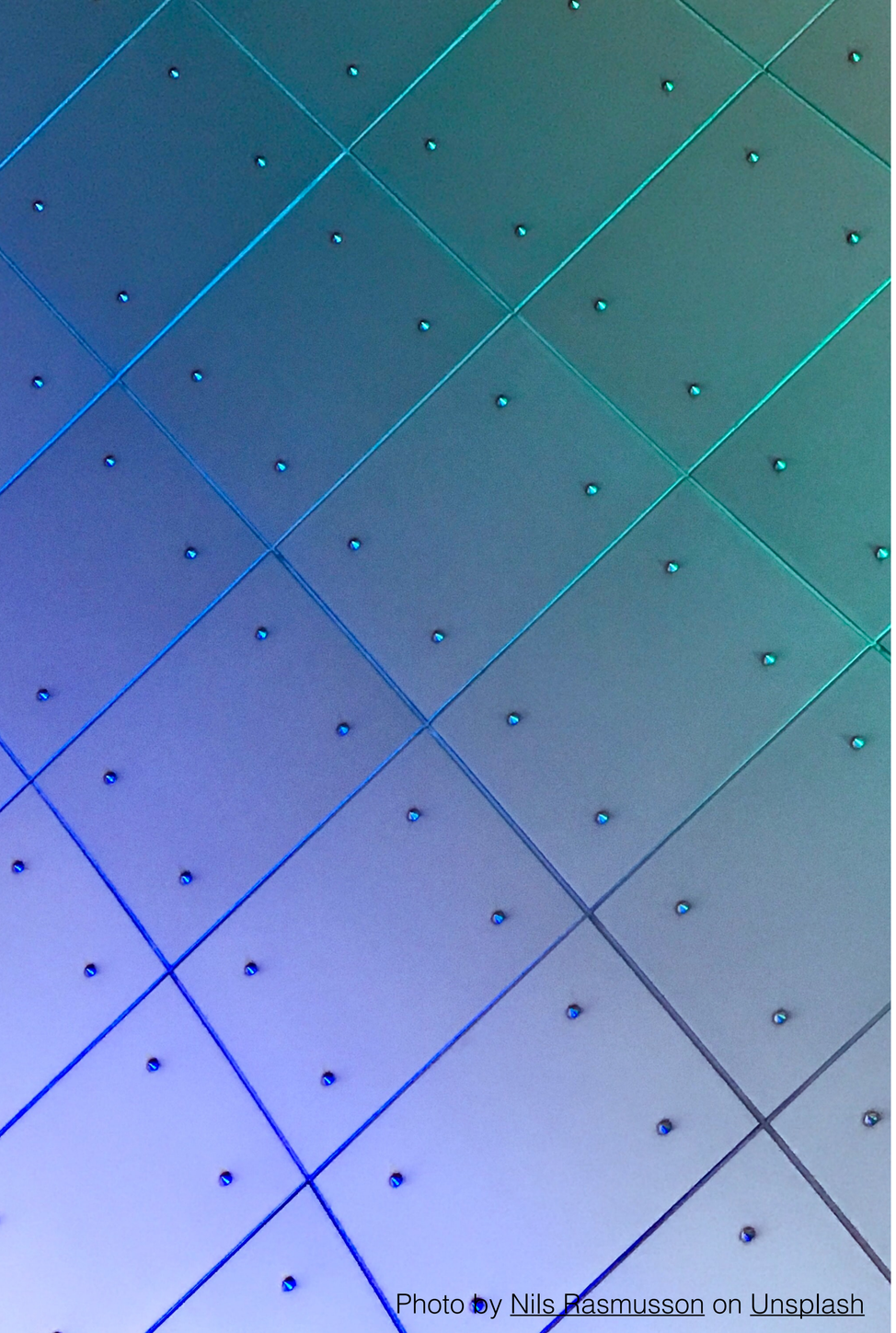


Photo by Nils Rasmusson on Unsplash

Why might we have non-identifiability?

1. One variable is just a multiple of another.

2. One variable is a *linear combination* of several others. $\underline{x}_1 = a\underline{x}_2 + b\underline{x}_3 + \ldots + p\underline{x}_p$

3. There are more variables than members in the sample.

Note: *Near non-identifiability* is trickier than exact non-identifiability.