When Your Sample is Small But Your Predictors Are Many: A Hands-On Guide to Machine Learning & Ridge Regression

+ Lasso & Elastic Net Regression because they're real similar

M. Blake Rafferty
Dept. of Communication Disorders

Overview

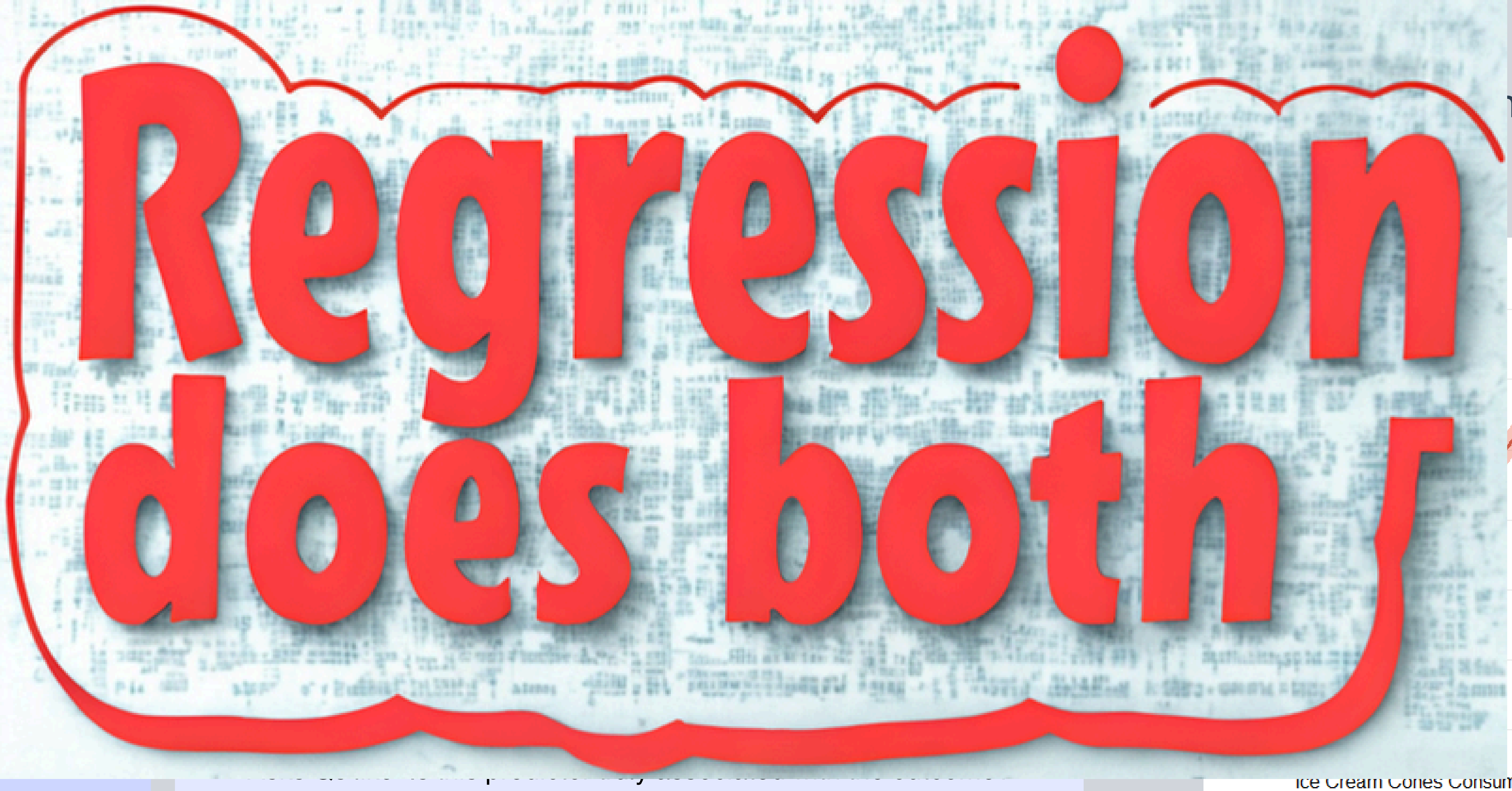
1. What is regression and why is it cool?
2. Least Squares Regression (aka the Classic Regression You Probably Know) and Its Limitations
3. Taming the beast: Meet Ridge
4. What about lambda?: Choosing parameters for stable predictions
5. How to handle more complicated models
6. Hands-on in R: glmnet, extracting coefficients, predicting test data
7. Practical guidance: When to use this method and how to interpret & report results

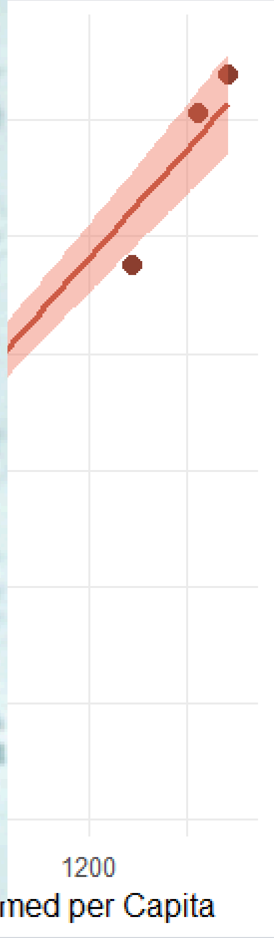
What is regression & why is it cool?

Regressi nd
one or m

Mod

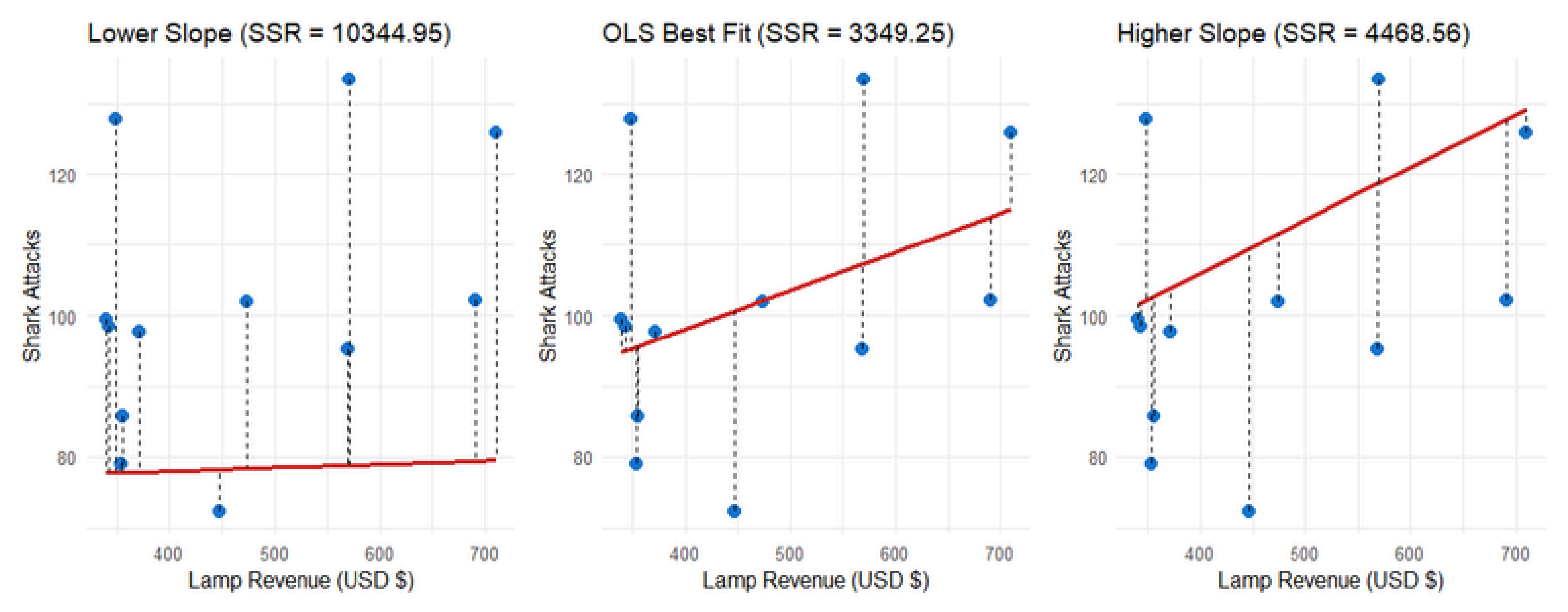
Infer





Ordinary Least Squares Regression and Its Limitations

OLS finds the line that minimizes the sum of squared differences (residuals) between the observed data points and the line's predicted values. The best fit line makes the predictions as close as possible to the actual data.



Works really well when:

Predictors (if > 1) are uncorrelated

# of observations > # of predictors

Doesn't work when:

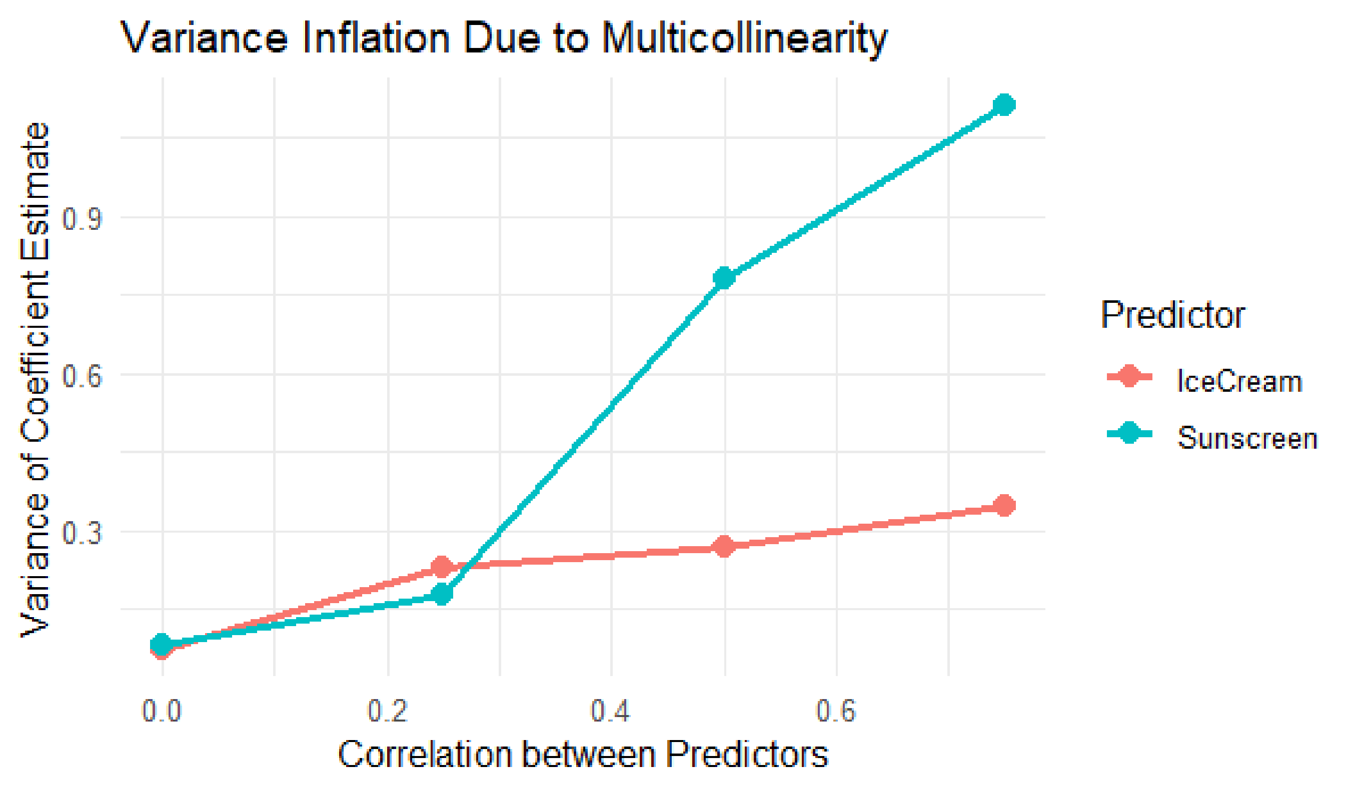
Predictors (if > 1) are collinear

# of observations < or = # of predictors

What’s the big deal?

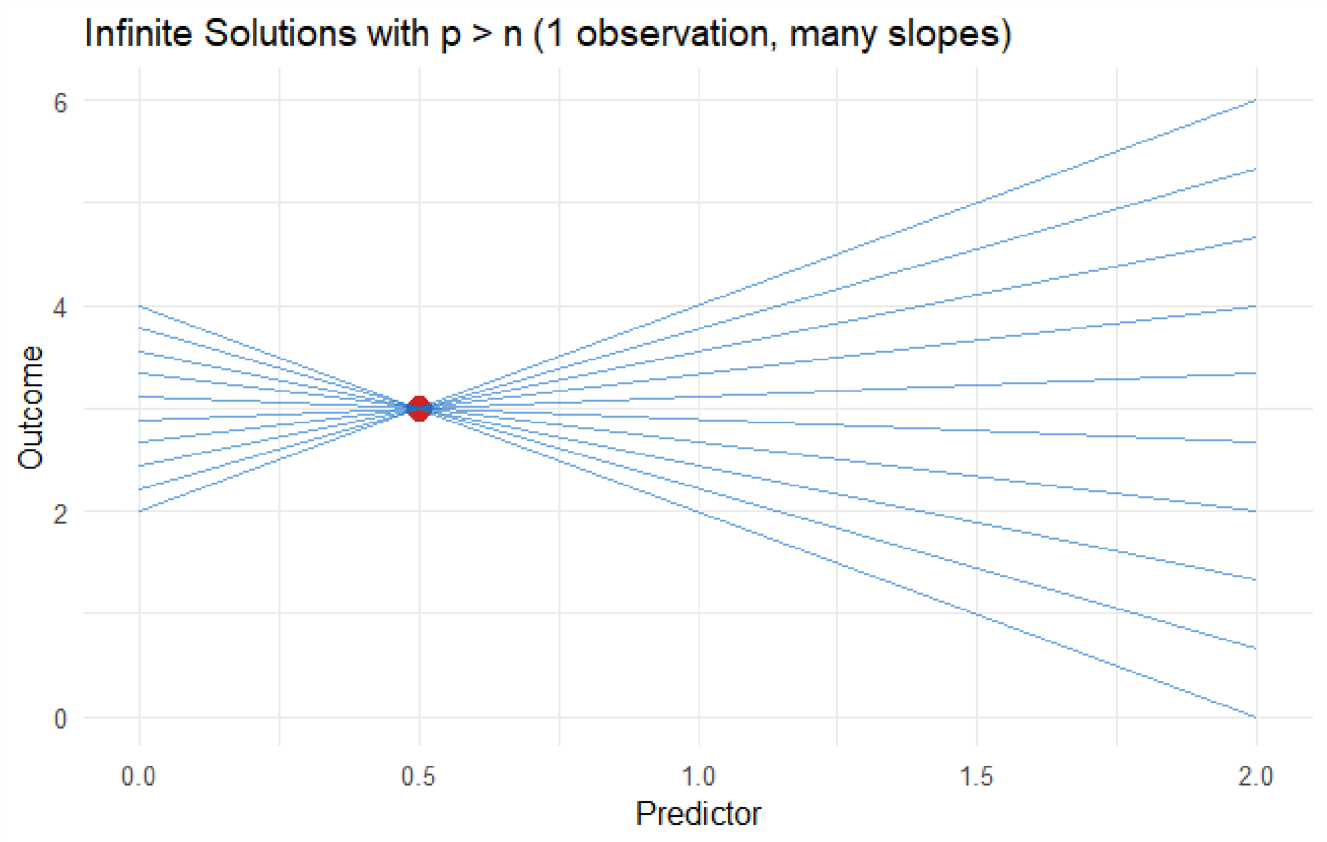
01

Multicollinearity results in unstable models that over estimate effects and don’t generalize to new data (over fit)



02

Models with less observations than parameters have an infinite number of possible correct solutions (all lines have the same SSR: 0)




Ridge Regression

Key Concepts Behind the Method

Ridge regression works like normal OLS regression, but incorporates a **penalty term** (lambda -> λ).

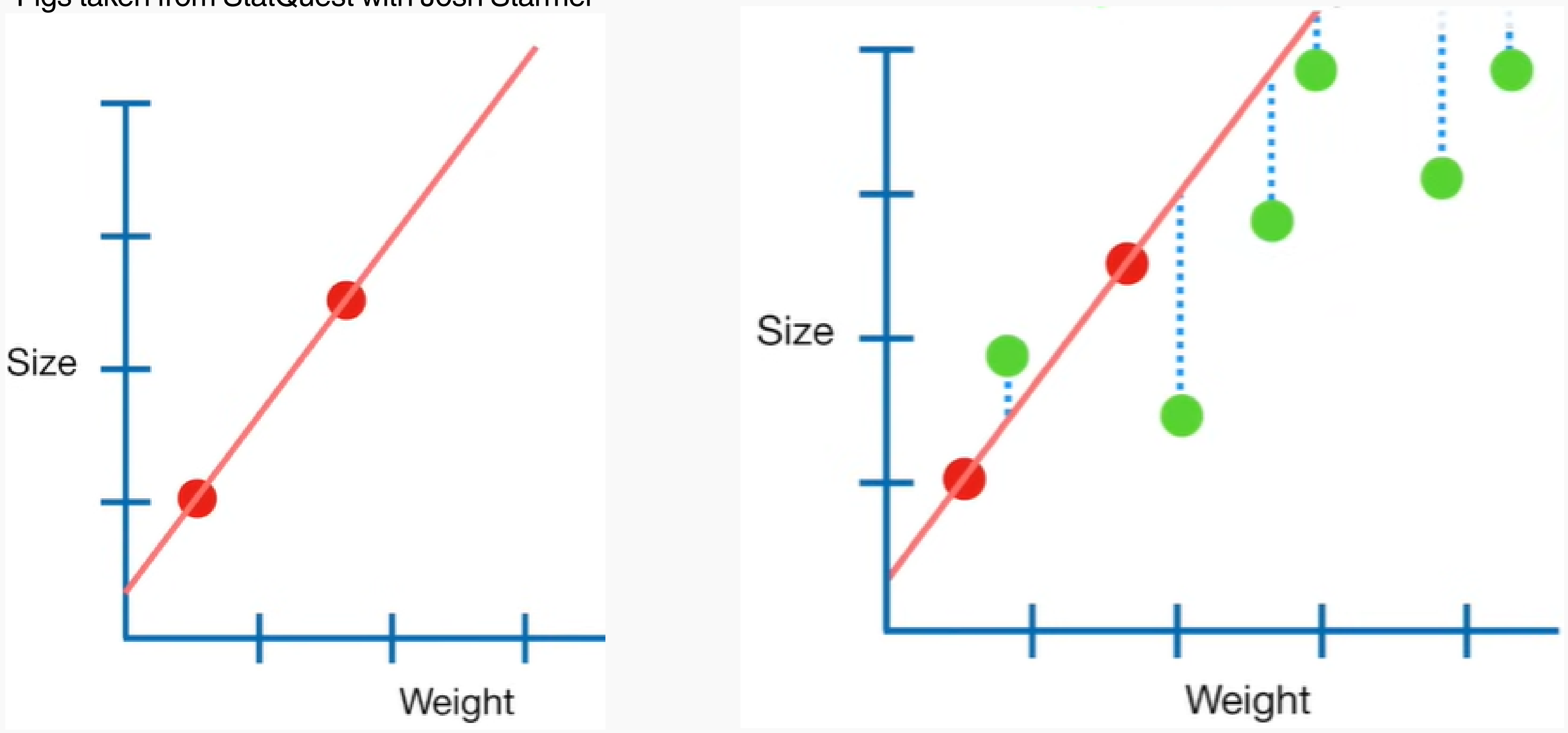
This helps prevent *overfitting* by shrinking the size of the coefficients,
enhancing model stability in the presence of multicollinearity.

Line Formula: Intercept + slope * x

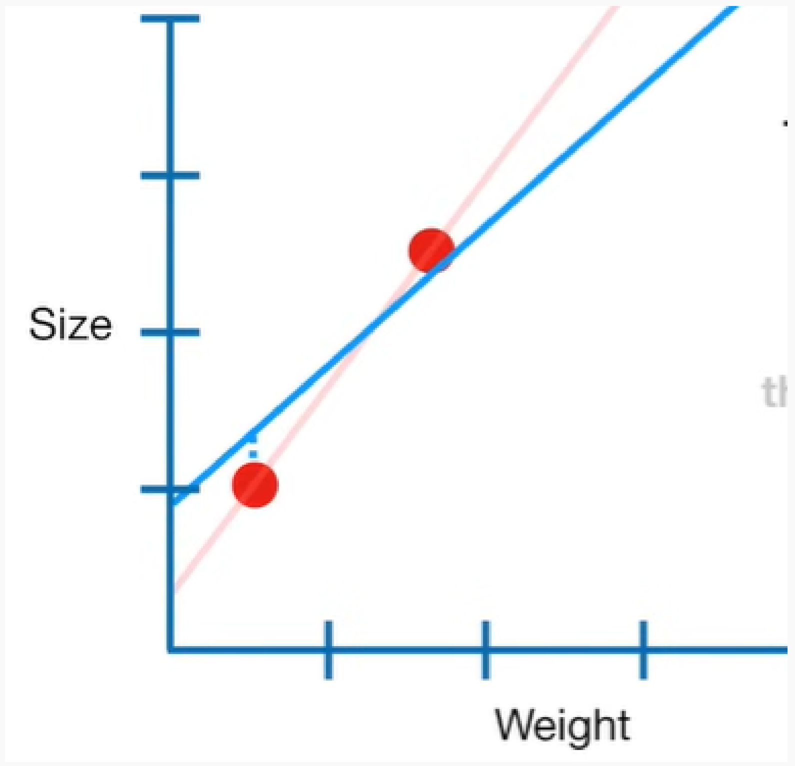
Goal of OLS is to find the line that minimizes SSR

Ridge finds the line that minimizes SSR + $\lambda * \text{slope}^2$

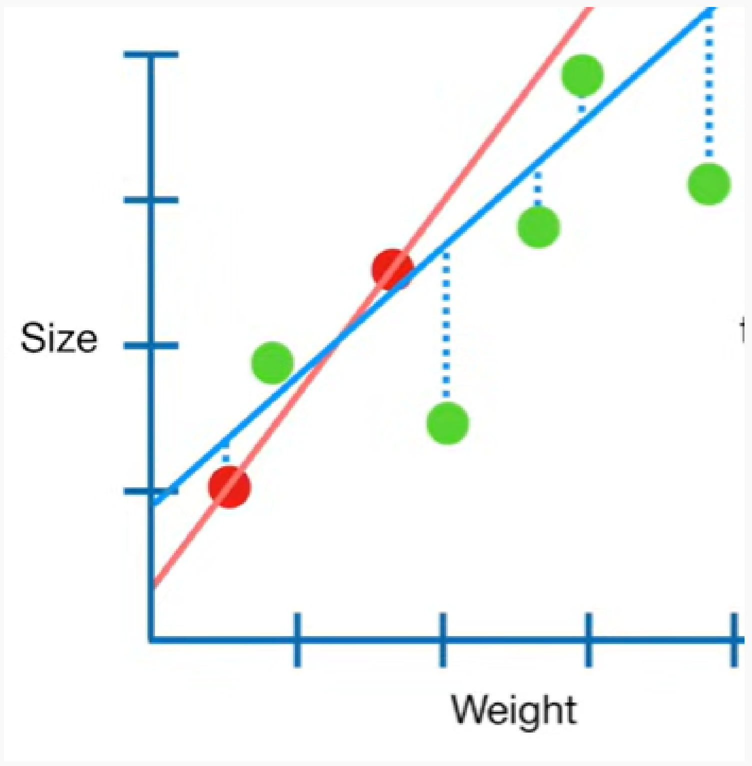
Figs taken from StatQuest with Josh Starmer



OLS: Size = 0.4 + 1.3 x Weight
If λ = 1
$0 + \lambda * 1.3^2 = 1.69$



Ridge: Size = 0.9 + 0.8 x Weight
If λ = 1
$0.3^2 + 0.1^2 + \lambda * 0.8^2 = 0.74$



Desensitizing the regression formula improves model fit to unseen data

What about lambda?

Can be any number between 0 and ∞.

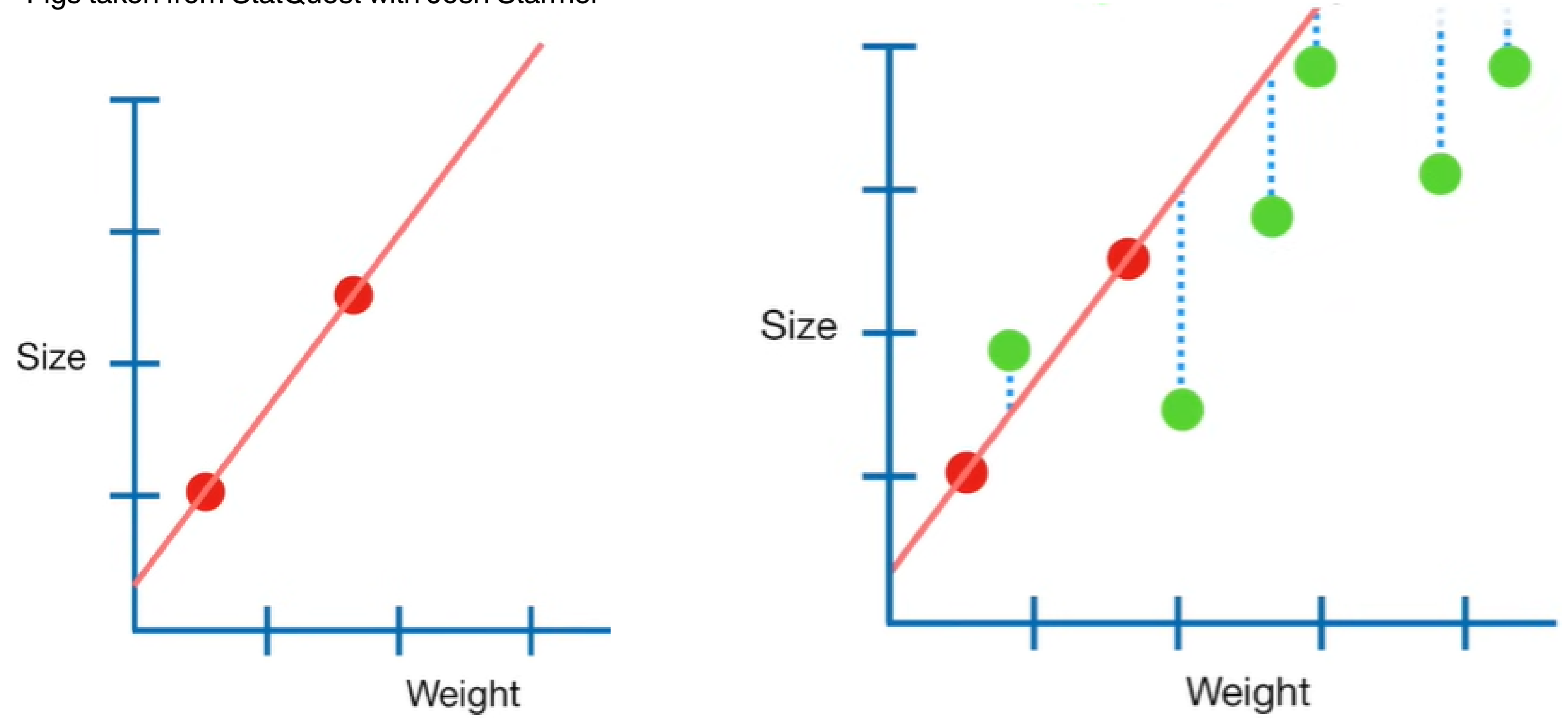
The larger we make λ, the closer the slope gets to zero.

How do we decide?

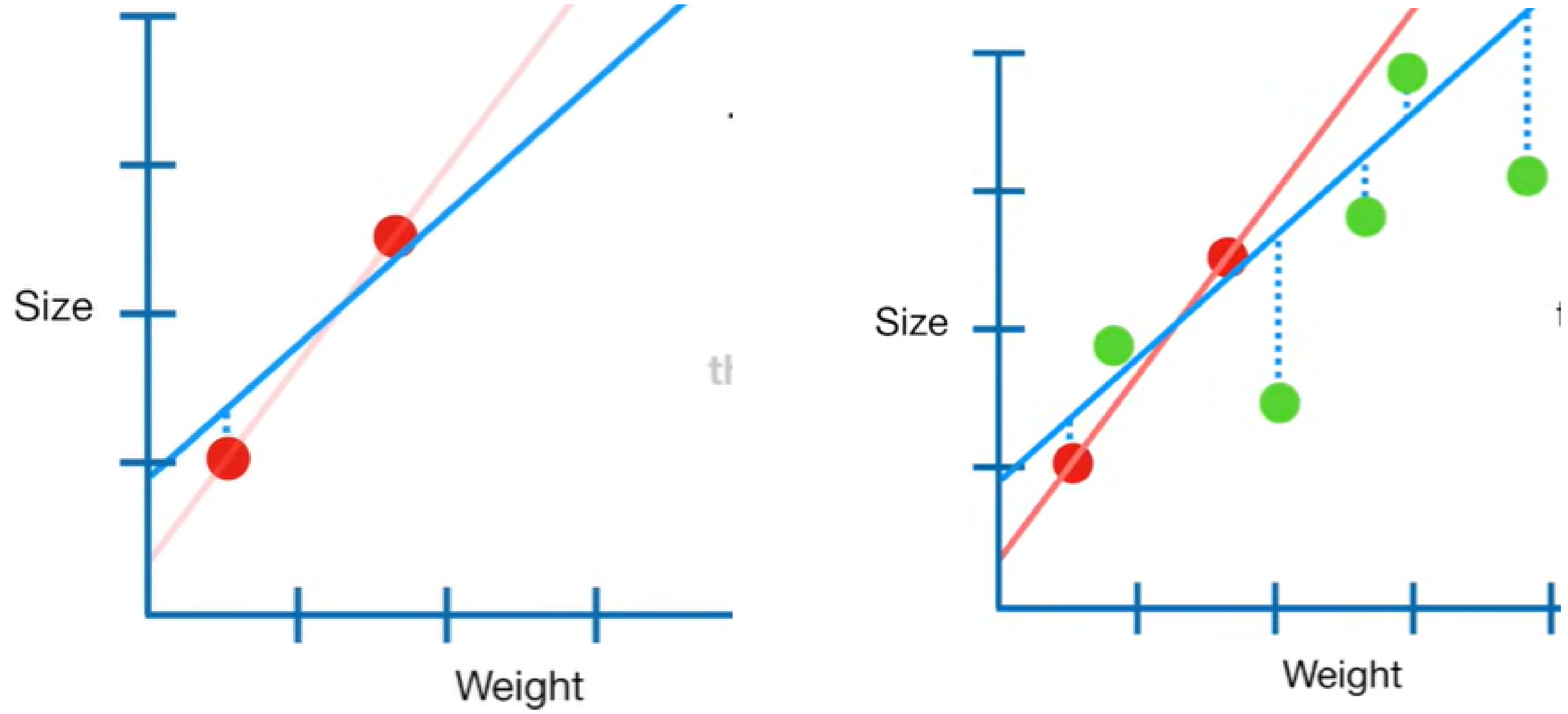
We just try a bunch of stuff and use a process called **cross-validation** to determine which values results in the lowest *variance*-->best fit to unseen data.

Typically 10-fold cross validation:
-Divide the data into 10 blocks.
-Leave one block out (test data).
-Train the model on remaining data (training data) with a particular lambda.
--->Find the ridge regression line on the training data.
-Evaluate how well the line predicts the held out data.
-Repeat for all blocks.
-Go with the lambda the produces the lowest variance and interpret the model

Figs taken from StatQuest with Josh Starmer



OLS: Size = 0.4 + 1.3 x Weight
If λ = 1
$0 + \lambda * 1.3^2 = 1.69$



Ridge: Size = 0.9 + 0.8 x Weight
If λ = 1
$0.3^2+0.1^2 + \lambda * 0.8^2 = 0.74$

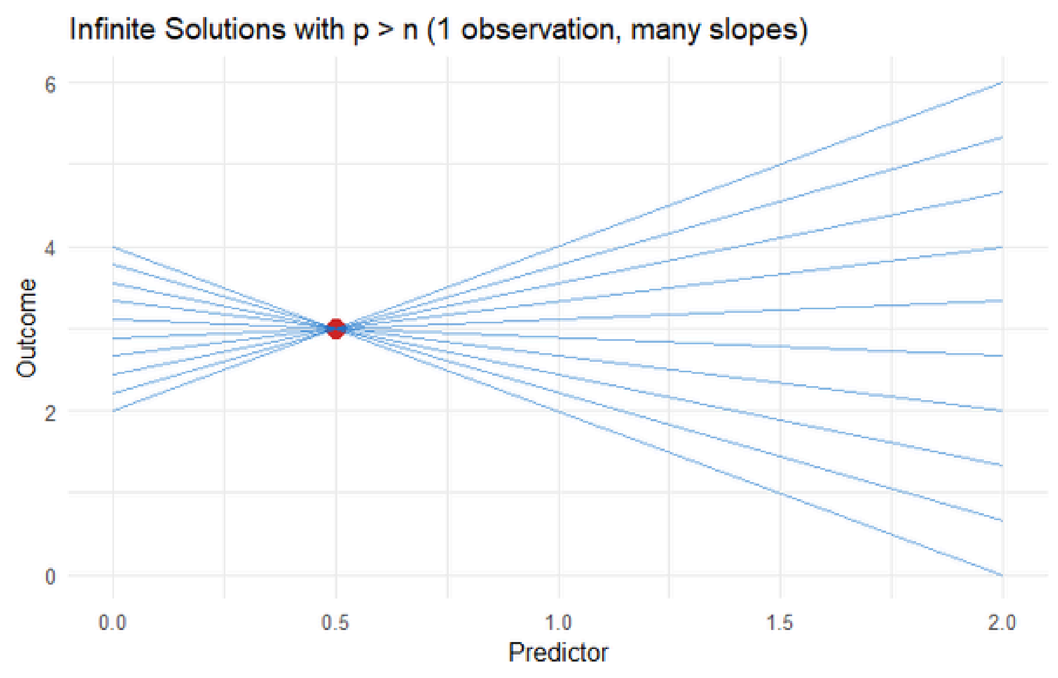
Desensitizing the regression formula improves model fit to unseen data

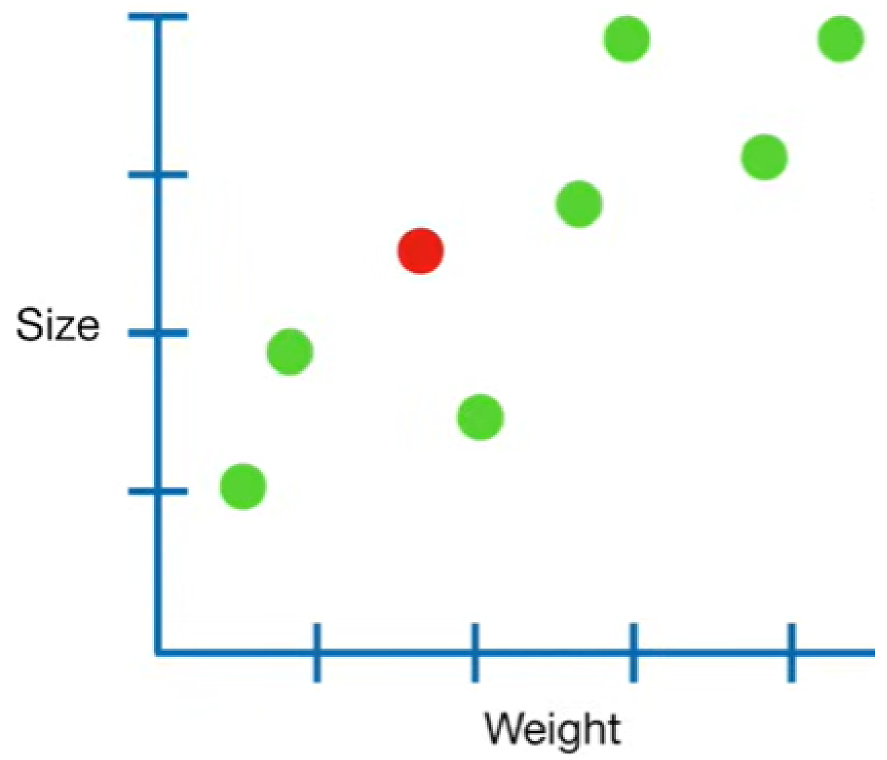
What about more complicated models?

Penalize all parameters!

Shark Attacks = SSR + λ * (ice cream cones^2 + lamp sales revenue^2 + white-collar crimes^2 + cat videos watched per household^2 + daily pizza deliveries^2)

What if we only have three cities (i.e. less observations than parameters)?





Practical Use Cases

01

OLS

Ordinary Least Squares is ideal for datasets with fewer predictors and when the relationship is linear, providing straightforward interpretation and reliable estimates under OLS assumptions.

02

Ridge

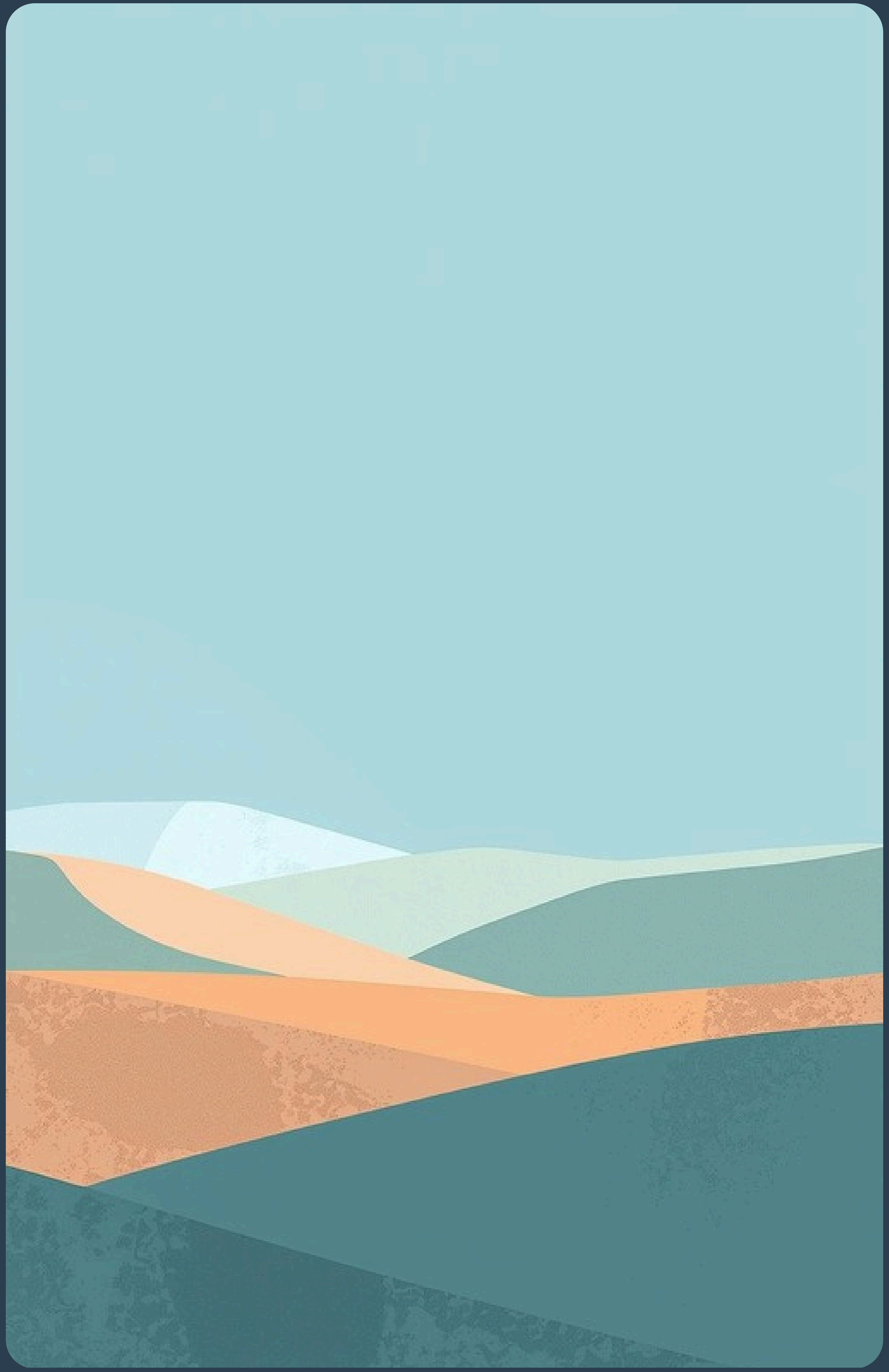
Ridge regression excels when dealing with multicollinearity or many predictors, effectively stabilizing coefficient estimates by adding a penalty term, thus reducing variance without greatly increasing bias.

Hands-On in R: glmnet Usage

Understanding how to use **glmnet** is essential for implementing ridge, lasso, and elastic net regression. This tool simplifies fitting models, plotting coefficient paths, and evaluating predictions effectively in R.

How to report?

To address multicollinearity among predictors, we fit a ridge regression model with the regularization parameter selected via 10-fold cross-validation. Coefficients reported below correspond to the model evaluated at the optimal penalty (λ_min). As expected under ridge regularization, all predictors retained non-zero coefficients, with effect sizes shrunk toward zero relative to ordinary least squares. The largest coefficients were observed for ice cream sales, sunscreen sales, and ice-cream-truck operating costs—three highly collinear predictors—indicating that ridge regression distributed explanatory weight across correlated variables rather than attributing disproportionate influence to any single predictor. Several predictors with negligible true effects (e.g., electricity usage, white-collar crime rates) exhibited coefficients near zero, suggesting limited contribution to the outcome once regularization was applied.



Summary