Model Building

DASC 512

Model Building

Model building can be an iterative process of trying different models and comparing them

- For simple scenarios with few variables, scatterplots can provide a lot of info
- For more complicated scenarios, we have a few methods to iteratively select variables for inclusion in the model

We'll talk about:

- Stepwise regression
- Ridge regression
- Lasso regression

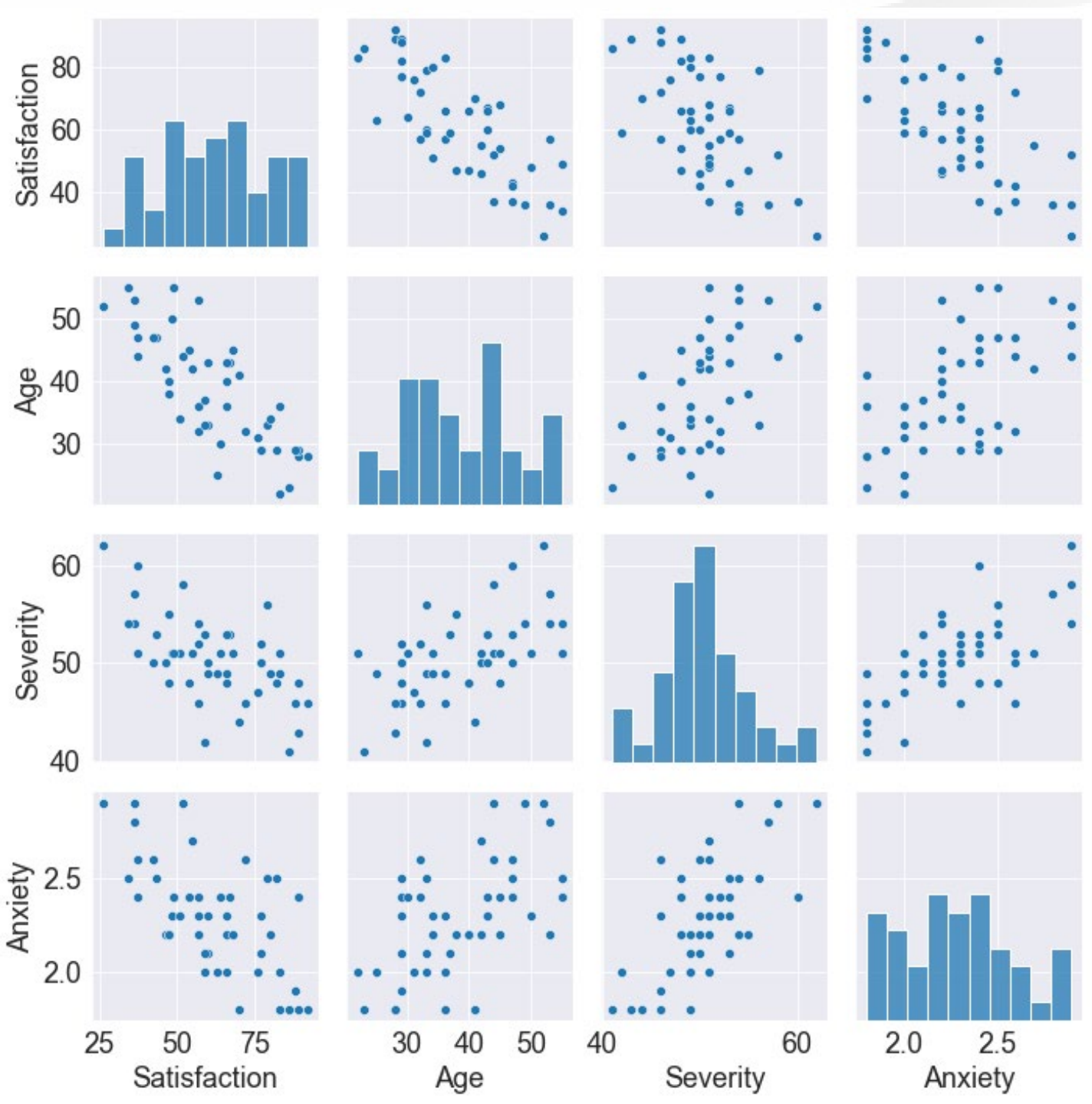
Model Building

It is best practice to split your data into three parts:

- Training set: the data you'll use to build a model
- Validation set: candidate models use this to compare accuracy
- Test set: used to assess accuracy of **only** the final model

In the final project, I'll hold back the test set results until evaluation.

Visual Method



Stepwise Regression

The most rudimentary method of iterative variable selection

First step is to hypothesize all potential predictor variables, including first order, higher order, and interaction terms

Operates using forward selection, backwards elimination, or some combo

- Forward Selection: At each step, look for the “best” parameter to include and add it to the model. Terms are never removed.
- Backward Elimination: Start with the full model. At each step, remove the “worst” parameter. Terms are never added.
- Mixed: Some mix of the two.

Backwards Elimination

Let’s run through this using the patient satisfaction data from Lesson 1

Forwards Selection

To do forwards selection, at each step we have to examine all possible models adding one variable and pick the best

Stepwise Regression

Stepwise regression is not implemented in Scipy, Statsmodels, or Sklearn

Why? It’s a good educational tool, so people know it, but it isn’t the best method for any real-world scenario

Ridge Regression (L2 Regularization)

Recall that our sum of squared errors is

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

Ridge regression adds a shrinkage penalty to penalize large parameters. This is the loss function that it will minimize.

$$Loss = SSE + Penalty = SSE + \alpha \sum_{j=1}^{k} \hat{\beta}_j^2$$

Ridge Regression

$$Loss = SSE + \alpha \sum_{j=1}^{k} \hat{\beta}_j^2$$

$\alpha > 0$ here is **not significance**. It is instead a parameter you can use to tune how much additional parameters are penalized.

If $\alpha = 0$, this is equivalent to OLS.

Increasing α artificially underestimates slopes $\hat{\beta}_1, \ldots, \hat{\beta}_k$

If $\alpha = \infty$, this will result in the model $\hat{y} = \hat{\beta}_0 = \bar{y}$

LASSO Regression (L1 Regularization)

LASSO (Least Absolute Shrinkage and Selection Operator) regression adds a different shrinkage penalty to penalize large parameters. This is the loss function that it will minimize.

$$Loss = SSE + Penalty = SSE + \alpha \sum_{j=1}^{k} |\beta_j|$$

A high value of α will result in a sparse model. A low value will result in a full model. This is much easier to interpret for model selection.

Next time…

Model Adequacy