IDS 702

Prediction modeling metrics

How is modeling different for predictive problems?

Compared to inference problems…

- Less concerned about model interpretation
- More flexibility with model selection
- Different model metrics/evaluation

Model diagnostics
transforming variables

Don't base model selection on p-values

Mean Squared Error (MSE)

- One particularly useful metric for measuring model fit (especially when the goal is prediction) is the mean squared error

observed y_i

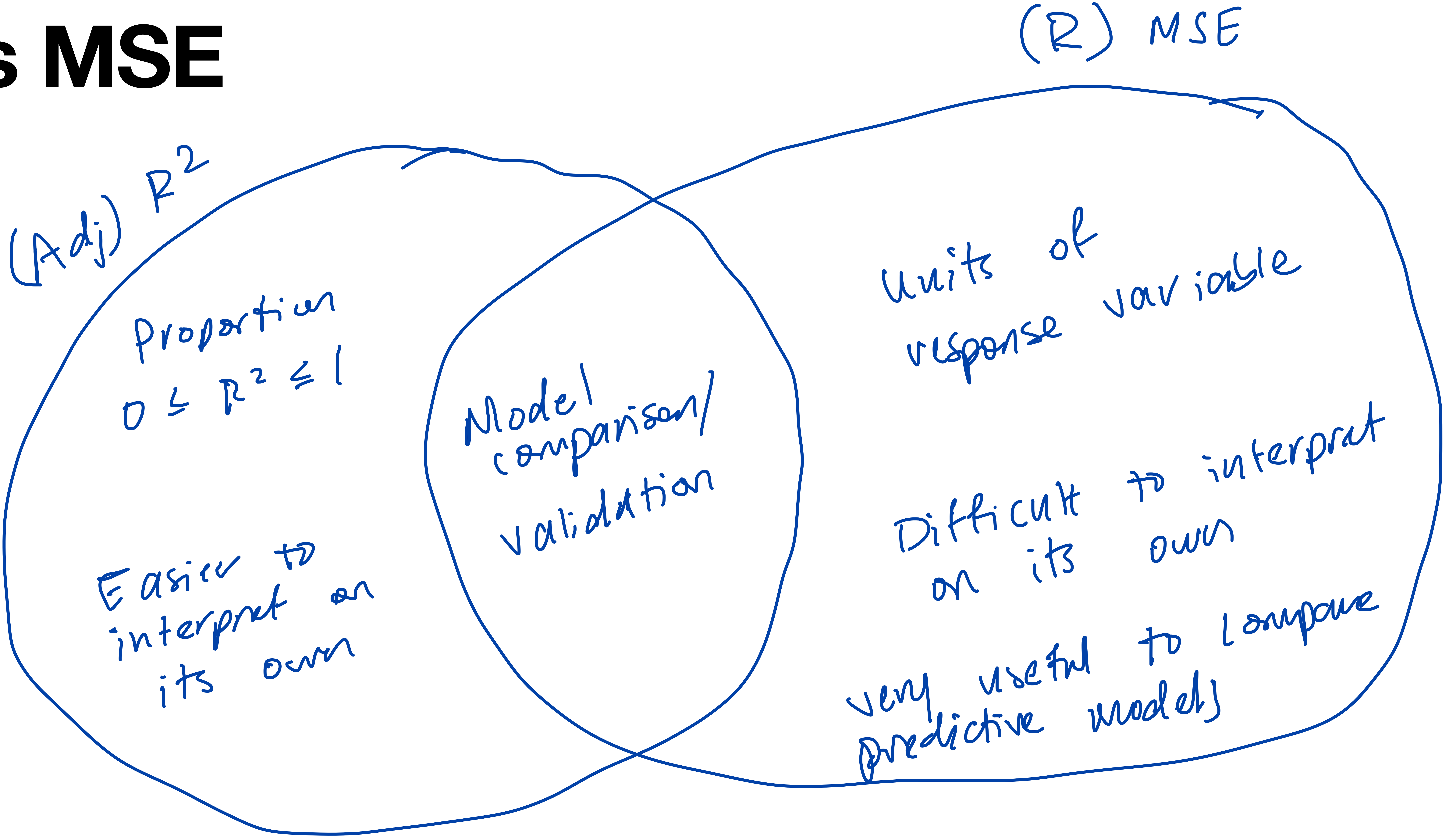
- $$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^{\boxed{2}}$$

predicted outcome

- Some also report Root Mean Squared Error (RMSE) which is measured in the same units as the response variable

- The smaller the MSE/RMSE, the better

R^2 vs MSE


(Adj) R^2
(R) MSE
Proportion
$0 \leq R^2 \leq 1$
Easier to interpret on its own
Model comparison/validation
Units of response variable
Difficult to interpret on its own
very useful to compare predictive models

MSE

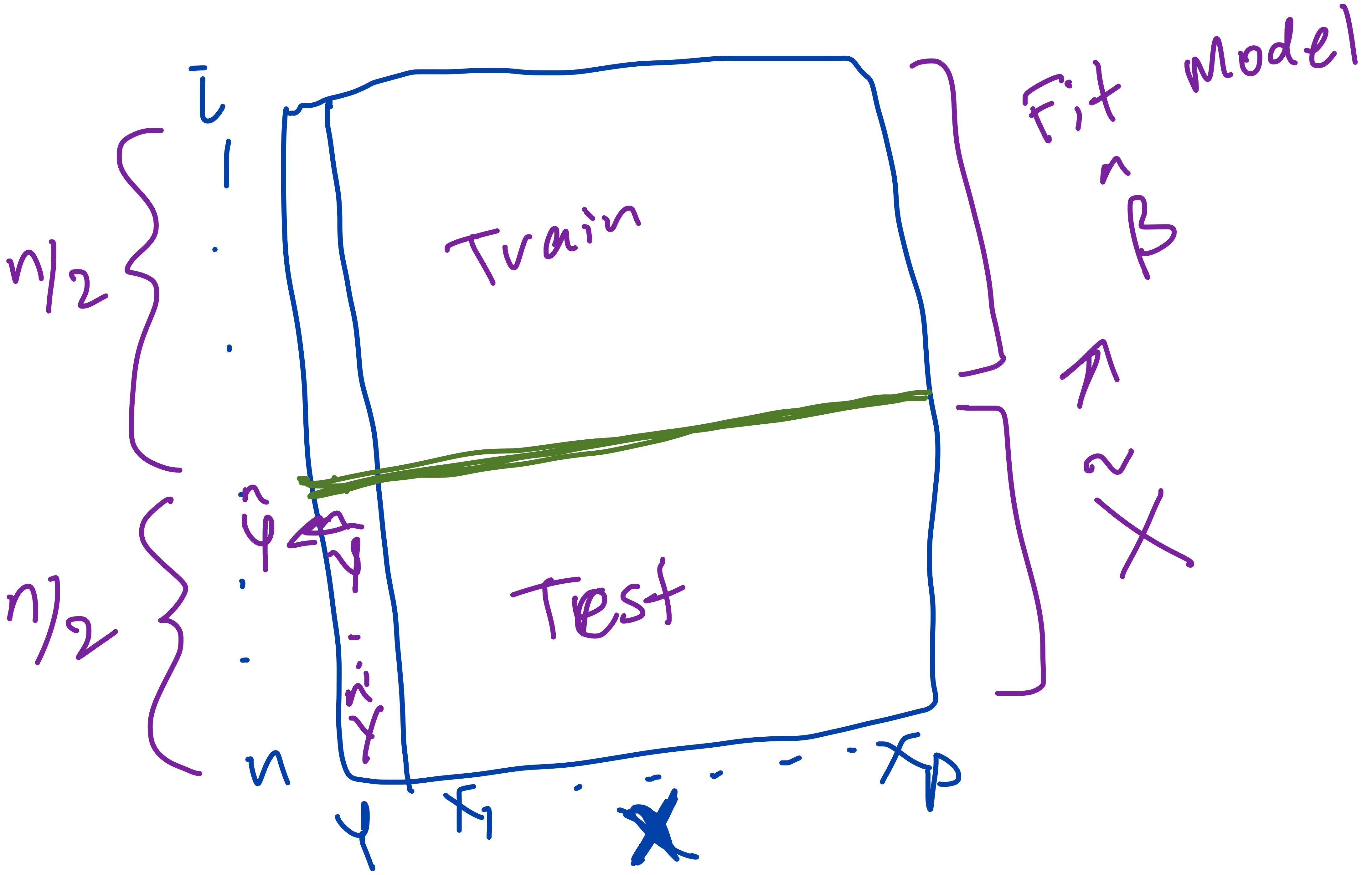
- It is often useful to calculate out-of-sample MSE using a different dataset
- What does our model tell us about what might happen in the future?
- To do this, we can split our sample into **training** and **test** datasets

Fit the Model

Evaluate the model

$\tilde{x}$

Training and test data



- Training data: (x_{0i}, y_{0i})
- Test data: (x_{1i}, y_{1i})
- Test MSE or out-of-sample MSE is then given by

test data

$$MSE_{test} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_{1i} - \hat{y}_{1i})^2$$ where $\hat{y}_{1i}$ is the predicted response for an observation in the test data using the model fitted with the training data

$\tilde{X}$

$\hat{\beta}$

Overfitting

- Using test data is often important because of the problem of overfitting
- Model fits very well to the training data but is not generalizable
- Train MSE or Test MSE: which will generally be larger?

minimize RSS

$y_i - \hat{y}_i$

to fit the model i.e. obtain $\hat{\beta}$