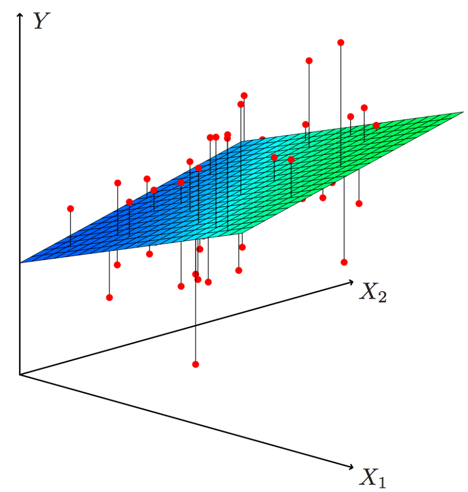
Linear Regression



Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Fitted Value

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Residual Sum of Squares

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

Coefficient Estimates

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Assessing Accuracy

Residual Sum of Squares

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

Not great…

Residual Standard Error

$$RSE = \sqrt{\frac{1}{n-p-1}RSS} = \sqrt{\frac{(y_i - \hat{y}_i)^2}{n-p-1}}$$

Better…can roughly think of as average amount that response will deviate from regression line

R-Squared, or "Proportion of Variance Explained"

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

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

☺ Nice interpretation
Independent of scale of y

Interpretation

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.933
Model:                            OLS   Adj. R-squared:                  0.928
Method:                 Least Squares   F-statistic:                     211.8
Date:                Mon, 03 Nov 2014   Prob (F-statistic):           6.30e-27
Time:                        14:45:06   Log-Likelihood:                -34.438
No. Observations:                  50   AIC:                             76.88
Df Residuals:                      46   BIC:                             84.52
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             0.4687      0.026     17.751      0.000         0.416     0.522
x2             0.4836      0.104      4.659      0.000         0.275     0.693
x3            -0.0174      0.002     -7.507      0.000        -0.022    -0.013
const          5.2058      0.171     30.405      0.000         4.861     5.550
==============================================================================
Omnibus:                        0.655   Durbin-Watson:                   2.896
Prob(Omnibus):                  0.721   Jarque-Bera (JB):                0.360
Skew:                           0.207   Prob(JB):                        0.835
Kurtosis:                       3.026   Cond. No.                         221.
==============================================================================
```

Proportion of Variance Explained by model is 93.3%

Measure of the significance of the fit …my model isn't utterly useless ☺

There is an approximately 95% chance that [0.275, 0.693] will contain the true value of β_2

Each coefficient is really significant. Can also think of this as a Partial F-test.

"The average effect on Y of a one unit increase in X_2, holding all other predictors (X_1 & X_3) fixed, is 0.4836"

- However, interpretations are generally pretty hazardous due to correlations among predictors.
- p-values for each coefficient ≈ 0, so might be okay here

Note: Magnitude of the Beta coefficients is NOT how to determine whether predictor contributes. Why?

Linear Regression - Woes of Interpretation

- Don't use the magnitude of coefficient to determine *how significant* the variable is.
  - You can get a sense of contribution in the context of other predictors, but that's about it.
  - Feet vs. Inches

- If p-value of the coefficient is not significant, don't interpret coefficient.

```
Residuals:
    Min      1Q  Median      3Q     Max
-149.95  -34.42  -14.74   11.58  560.38

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.3483    11.6908   4.563 1.17e-05 ***
Cr            1.8577     0.2324   7.994 6.66e-13 ***
Co            2.1808     1.7530   1.244    0.216
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 74.76 on 128 degrees of freedom
Multiple R-squared: 0.544,      Adjusted R-squared: 0.5369
F-statistic: 76.36 on 2 and 128 DF,  p-value: < 2.2e-16
```

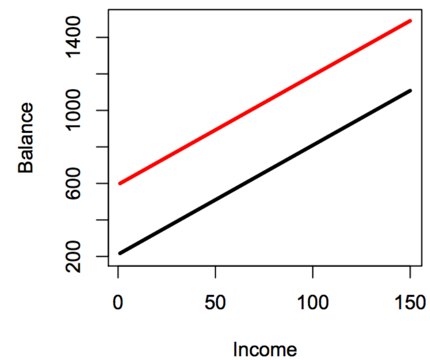
Interactions

Interacting ***student*** (qualitative) and ***income*** (quantitative)

No Interaction

$$balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i$$

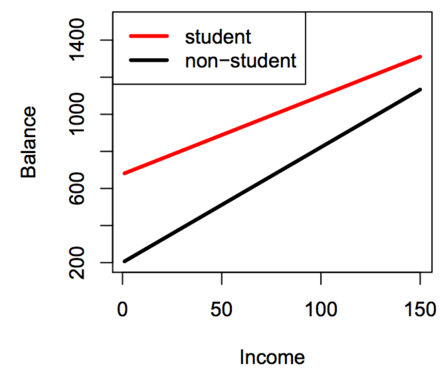
$$\begin{aligned} \texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\ &= \underline{\beta_1} \times \texttt{income}_i + \begin{cases} \underline{\beta_0 + \beta_2} & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases} \end{aligned}$$



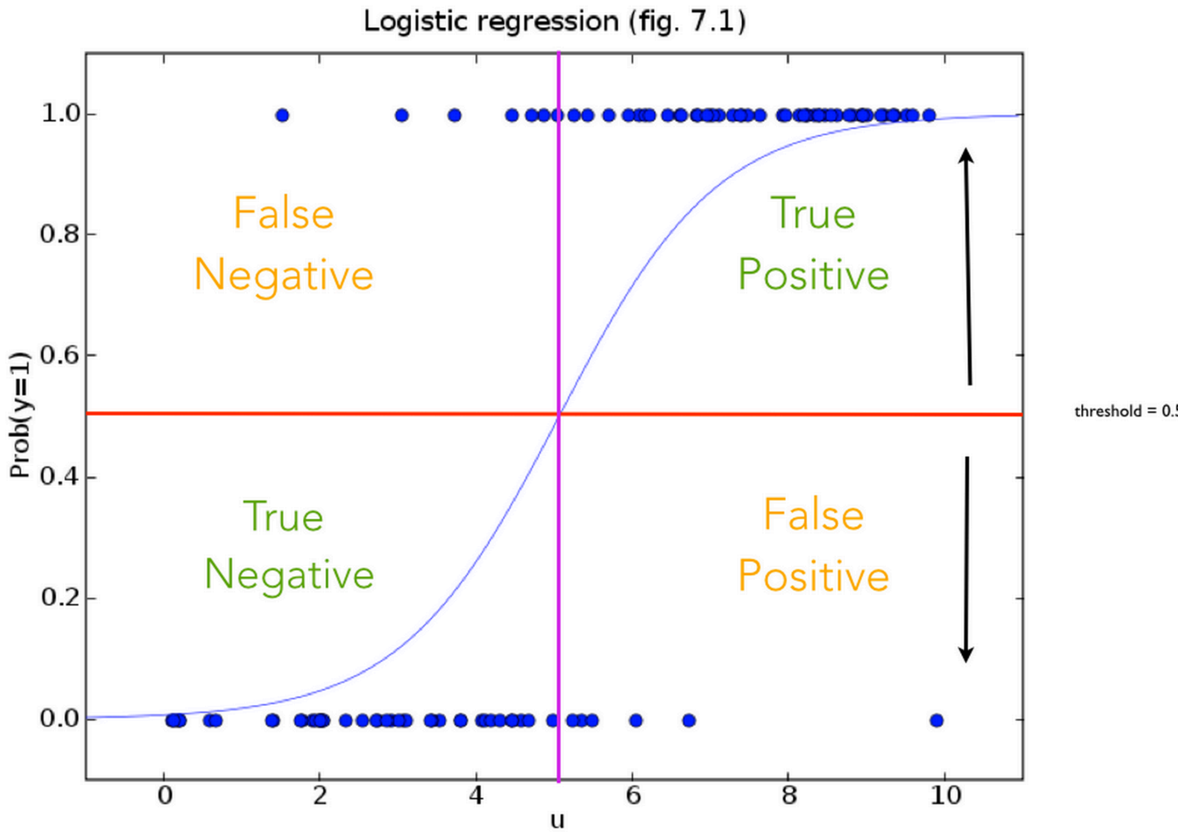
With Interaction

$$balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i + \beta_3 * income_i * student_i$$

$$\begin{aligned} \texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases} \\ &= \begin{cases} \underline{(\beta_0 + \beta_2)} + \underline{(\beta_1 + \beta_3)} \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases} \end{aligned}$$

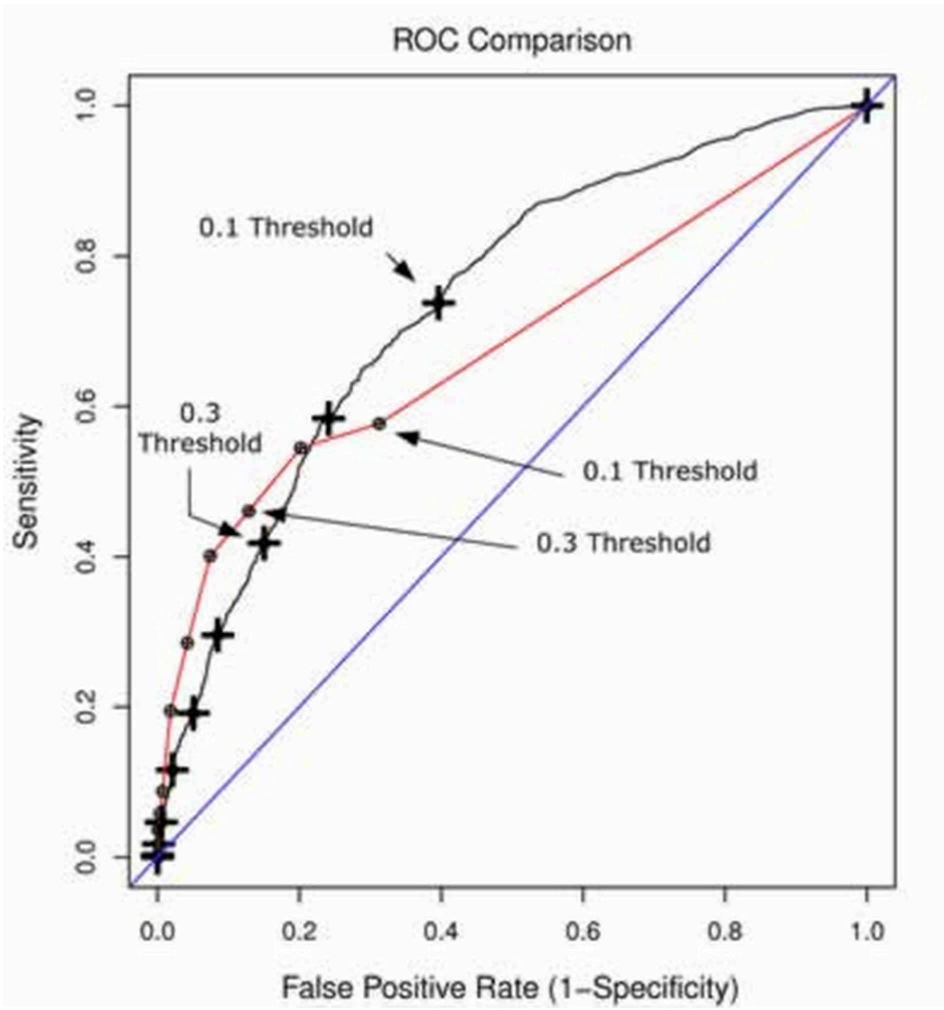


Logistic Regression



	Predicted Yes	Predicted No
Actual Yes	True positive	False negative
Actual No	False positive	True negative

Logistic Regression - Evaluation



- Area under the Curve (AUC)
- F1 Score
- Precision / Recall
- Sensitivity / Specificity

Logistic Regression - Evaluation

true positive (TP)

eqv. with hit

true negative (TN)

eqv. with correct rejection

false positive (FP)

eqv. with false alarm, Type I error

false negative (FN)

eqv. with miss, Type II error

accuracy (ACC)

$$ACC = (TP + TN)/(P + N)$$

F1 score

is the harmonic mean of precision and sensitivity

$$F1 = 2TP/(2TP + FP + FN)$$

sensitivity or true positive rate (TPR)

eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

specificity (SPC) or true negative rate (TNR)

$$SPC = TN/N = TN/(FP + TN)$$

precision or positive predictive value (PPV)

$$PPV = TP/(TP + FP)$$

negative predictive value (NPV)

$$NPV = TN/(TN + FN)$$

fall-out or false positive rate (FPR)

$$FPR = FP/N = FP/(FP + TN)$$

false discovery rate (FDR)

$$FDR = FP/(FP + TP) = 1 - PPV$$

false negative rate (FNR)

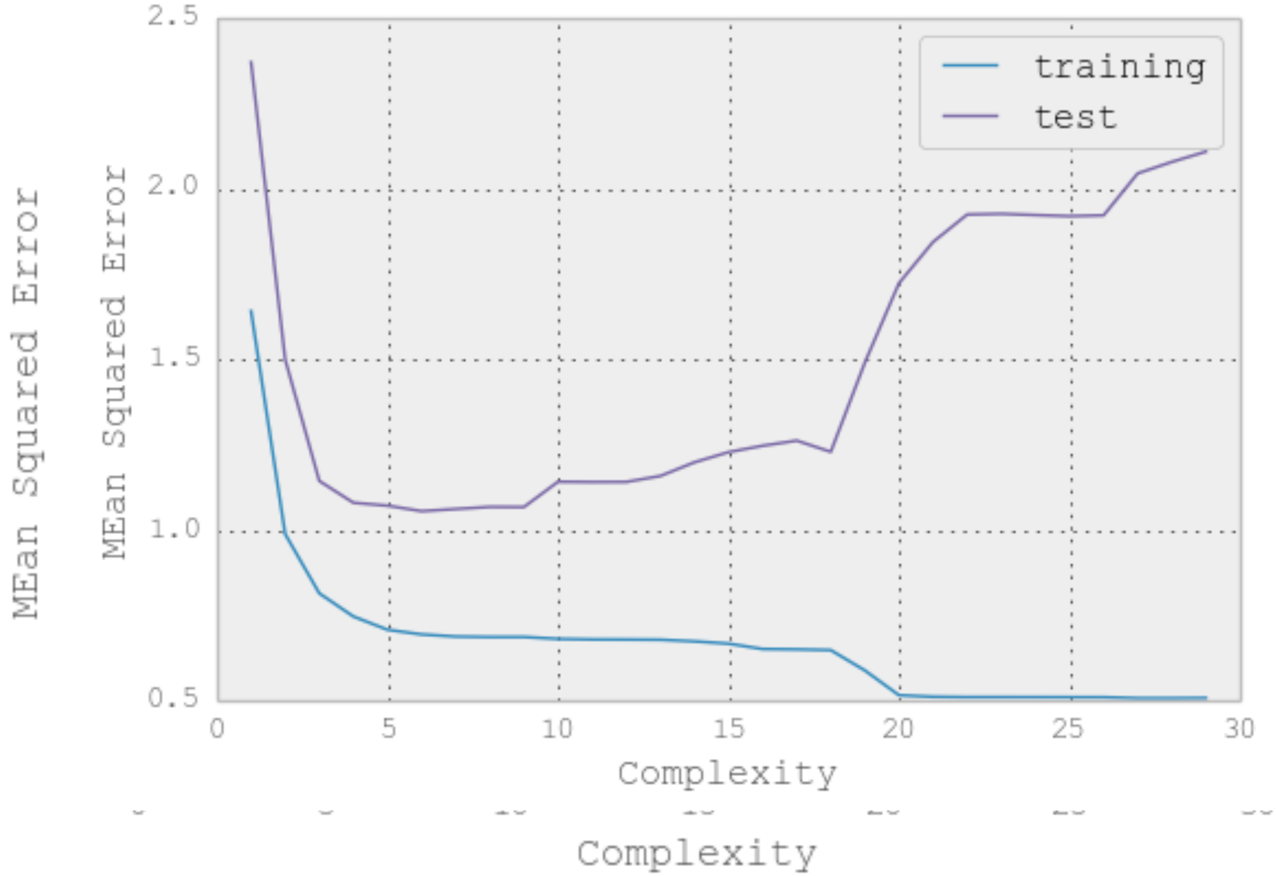
$$FNR = FN/(FN + TP) = 1 - TPR$$

Logistic Regression - Interpretation

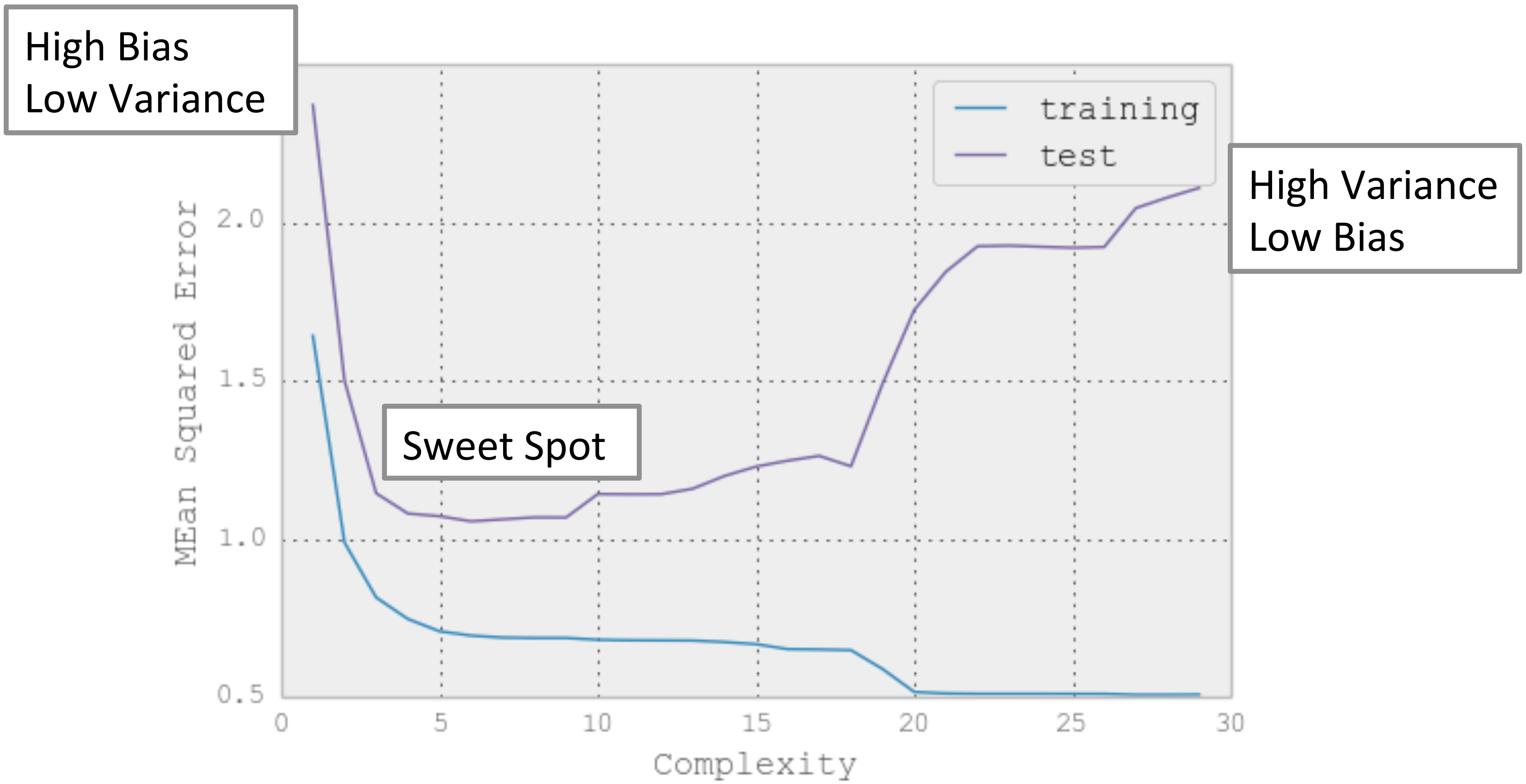
$$\frac{p}{1-p} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n} = e^{\beta_0} e^{\beta_1 X_1} e^{\beta_2 X_2} \ldots e^{\beta_n X_n}$$

It tells you how much 1-unit increase of a feature increases the odds of being classified in the positive class. In this way, the coefficients of the logistic regression can be interpreted similarly to that of linear regression

Model Framework - Evaluation

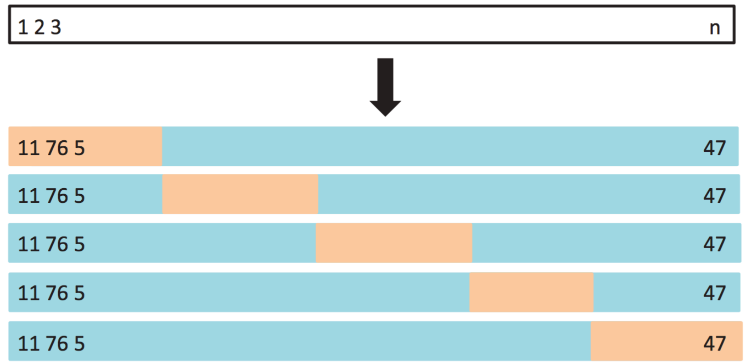


Model Framework - Evaluation



- Can break this complexity tradeoff into what we call “bias” and “variance”

K-Fold Cross-Validation



Randomly divide data into K=5 folds. Typically choose K=5 or 10.

Run K times

1. Fit model on training set, using (K-1) folds
2. Use fitted model in 1. to predict responses for validation set, 1 of the folds
3. Compute validation-set error
   - Quantitative Response: Typically MSE
   - Qualitative Response: Typically Misclassification Rate

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i$$

Model Framework – Cross-Validate & Test

<u>TRY STUFF</u>
- tune parameter 1
- tune parameter 2
- feature engineering
- feature selection
- scale factors 1 way
- scale factors another way
- etc. etc.

Test Set

Don’t touch until end for final evaluation