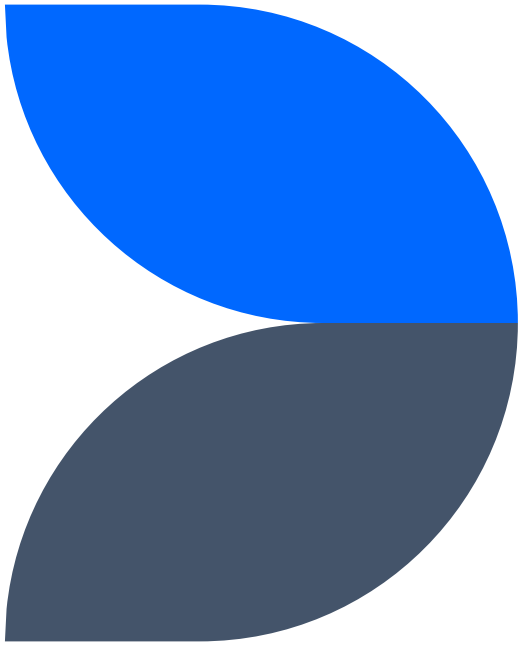
Modeling Regression Supervised ML

Melinda Higgins & Vicki Hertzberg

Goals

Supervised Machine Learning Models



“Supervised” Models

- Required:
  - Target must be “known” or identified

- Uses:
  - Accurate and (ideally) precise prediction of “target”
  - List of potential “predictors”
  - Evaluating association of predictors to target (implied mechanism)
  - (possibly) evaluate “importance” (tests) of predictors

Linear Regression

Target (Y)

- Continuous numeric
- Ideally ratio, integer and interval ok
- Normal (Gaussian) distribution

Linear Regression

Target (Y)

- Continuous numeric
- Ideally ratio, integer and interval ok
- Normal (Gaussian) distribution

Predictors (X's)

- Continuous numeric
- Ratio, integer and interval ok
- Ordinal – so-so, be careful interpreting
- Categorical ok but dummy variable coding needed for > 2 categories
- *Factor vs numeric class*

Linear Regression

Target (Y)

- Continuous numeric
- Ideally ratio, integer and interval ok
- Normal (Gaussian) distribution

Predictors (X's)

- Continuous numeric
- Ratio, integer and interval ok
- Ordinal – so-so, be careful interpreting
- Categorical ok but dummy variable coding needed for > 2 categories
- *Factor vs numeric class*

Evaluate

- Model fit: r2, adj r2
- Confidence (95%) or Prediction Intervals (for predicted Y's)
- Predictor coefficients (slopes)
- Assumptions
  - Independent X's
  - Linear slope

Linear Regression

Target (Y)

- Continuous numeric
- Ideally ratio, integer and interval ok
- Normal (Gaussian) distribution

R code

```
lm(y ~ x1 + x2, data)
```

Predictors (X's)

- Continuous numeric
- Ratio, integer and interval ok
- Ordinal – so-so, be careful interpreting
- Categorical ok but dummy variable coding needed for > 2 categories
- *Factor vs numeric class*

Evaluate

- Model fit: r2, adj r2
- Confidence (95%) or Prediction Intervals (for predicted Y's)
- Predictor coefficients (slopes)
- Assumptions
  - Independent X's
  - Linear slope

Logistic Regression

Target (Y)

- Binary
- Categorical
- Can be A/B; Yes/No; 0/1

Logistic Regression

Target (Y)

- Binary
- Categorical
- Can be A/B; Yes/No; 0/1

Predictors (X's)

- Continuous numeric
- Ratio, integer and interval ok
- Ordinal – so-so, be careful interpreting
- Categorical – reference category required

- *Factor vs numeric class*

Logistic Regression

Target (Y)

- Binary
- Categorical
- Can be A/B; Yes/No; 0/1

Predictors (X's)

- Continuous numeric
- Ratio, integer and interval ok
- Ordinal – so-so, be careful interpreting
- Categorical – reference category required
- *Factor vs numeric class*

Evaluate

- Model fit: AUC from ROC
- Contingency Table / Confusion Matrix (FP/FN tradeoffs)
- Predictor coefficients (odds ratios = exp(beta))
- Assumptions
  - Independent X's
  - Linear (in logit)

Logistic Regression

Target (Y)

- Binary
- Categorical
- Can be A/B; Yes/No; 0/1

R code

```
glm(y ~ x1 + x2, data,
    family = binomial)
```

Predictors (X's)

- Continuous numeric
- Ratio, integer and interval ok
- Ordinal – so-so, be careful interpreting
- Categorical – reference category required
- *Factor vs numeric class*

Evaluate

- Model fit: AUC from ROC
- Contingency Table / Confusion Matrix (FP/FN tradeoffs)
- Predictor coefficients (odds ratios = exp(beta))
- Assumptions
  - Independent X's
  - Linear (in logit)