

Data Mining

PATRICK HALL AND LISA SONG

DEPARTMENT OF DECISION SCIENCE

Contemporary Regression Models

- Linear Methods for Regression
- Bias-Variance Tradeoff
- Penalized Regression: Elastic Net
- Logistic Methods for Regression
- ROC and LIFT

Linear Regression



Carl Friedrich Gauss
(1777–1855)

Linear Regression

A linear regression model assumes that the regression function $E(Y|X)$ is linear in the input variables X_1, X_2, \dots, X_p . Linear models are:

- ▶ Simple and often provide an interpretable model
- ▶ Sometimes outperform nonlinear models with low signal-to-noise or sparse data
- ▶ Can be applied to transformations of the inputs - basis-function methods
- ▶ Many nonlinear models are direct generalizations of linear methods

Linear Regression

Let $X^T = (X_1, X_2, \dots, X_p)$ be the input vectors for which we want to predict output Y . The linear regression model has the form:

$$f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j \quad (1)$$

where β'_j s are unknown parameter coefficients and X_j are input vectors.

Linear Regression

Input vectors, X_j , can be:

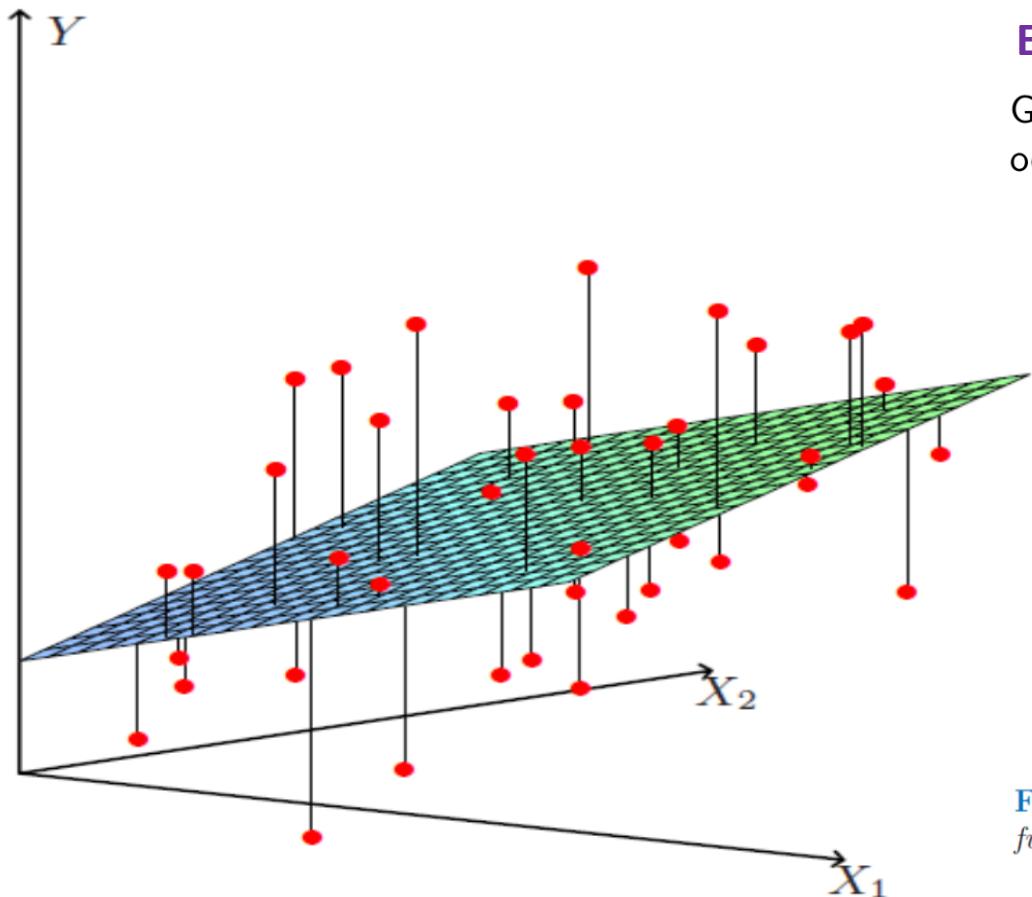
- ▶ Quantitative or transformations of quantitative inputs
- ▶ Basis expansion that leads to a polynomial representation
- ▶ Numeric or dummy coding: level-dependent constants
- ▶ Interactions between the input variables

Linear Regression

Regression methods estimates the model parameters β 's with the training data to minimize the residual sum of squares, $RSS(\beta)$.

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^N (y_i - f(x_i))^2 \\ &= \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2 \end{aligned} \tag{2}$$

Regression: Least-Squared Method



Elements of Statistical Learning (pg.45)

Geometry of least-square fitting in the \mathbb{R}^{p+1} -dimensional space occupied by the pairs (X, Y) .

FIGURE 3.1. Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y .

Interpreting Linear Regression

AVE_ave_provider_charge ~ AVE_ave_medicare_payment + AVE_num_service

| Analysis of Variance | | | | | |
|----------------------|------|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 2 | 3.85E+11 | 1.92E+11 | 1148.9 | <.0001 |
| Error | 3334 | 5.58E+11 | 167376011 | | |
| Corrected Total | 3336 | 9.43E+11 | | | |

| | | | |
|----------------|----------|----------|--------|
| Root MSE | 12937 | R-Square | 0.408 |
| Dependent Mean | 24721 | Adj R-Sq | 0.4076 |
| Coeff Var | 52.33355 | | |

| Parameter Estimates | | | | | | | |
|--------------------------|--------------------------|----|--------------------|----------------|---------|---------|--------------------|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > t | Variance Inflation |
| Intercept | Intercept | 1 | -1219.43 | 598.38 | -2.04 | 0.0416 | 0 |
| AVE_ave_medicare_payment | Average Medicare Payment | 1 | 3.83 | 0.08 | 47.88 | <.0001 | 1.02 |
| AVE_num_service | Number of Services | 1 | -5.84 | 1.17 | -4.96 | <.0001 | 1.02 |

AVE_ave_provider_charge ~ AVE_ave_medicare_payment + AVE_num_service

$$SSM = \sum(y\$ - \bar{y})^2$$

$$SSE = \sum(y\$ - \hat{y}\$)^2$$

$$SST = SSM + SSE$$

| Analysis of Variance | | | | | |
|----------------------|------|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 2 | 3.85E+11 | 1.92E+11 | 1148.9 | <.0001 |
| Error | 3331 | 5.58E+11 | 167376011 | | |
| Corrected Total | 3336 | 9.43E+11 | | | |

$F = MSM/MSE$, scaled ratio of the model variance to the error/residual variance.

Interpreted here as “rejecting the null hypothesis that all regression parameters equal 0,” i.e. the regression model is valid.

| | | | |
|----------------|----------|----------|--------|
| Root MSE | 12937 | R-Square | 0.408 |
| Dependent Mean | 24721 | Adj R-Sq | 0.4076 |
| Coeff Var | 52.33355 | | |

$R^2 = SSM/SST$, always interpreted as “the proportion of variance in the response variable explained by the model.”

R^2 adjusted for more than one variable.

| Parameter Estimates | | | | | | | |
|--------------------------|--------------------------|----|--------------------|----------------|---------|---------|--------------------|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > t | Variance Inflation |
| Intercept | Intercept | 1 | -1219.43 | 598.38 | -2.04 | 0.0416 | 0 |
| AVE_ave_medicare_payment | Average Medicare Payment | 1 | 3.83 | 0.08 | 47.88 | <.0001 | 1.02 |
| AVE_num_service | Number of Services | 1 | -5.84 | 1.17 | -4.96 | <.0001 | 1.02 |

VIF > 10 is considered an indicator of possible multicollinearity problems.

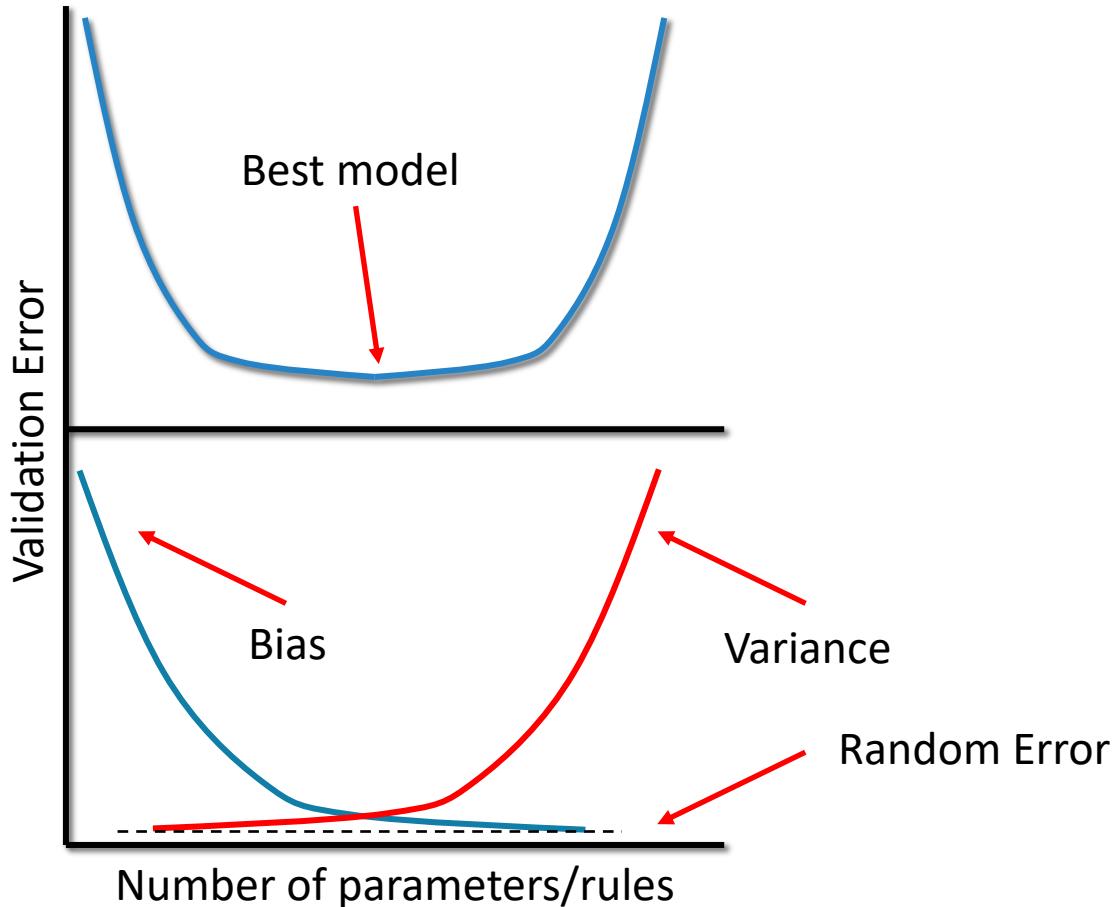
Estimated parameter for the input, here interpreted as “holding all other inputs constant, for a one unit increase in average Medicare payment, average provider charge will increase by 3.83 units on average.”

Standard error of the coefficient
- should be much smaller than the coefficient. (Std. deviation for the coefficient.)

t-test for the coefficient, here interpreted as “rejecting the null hypothesis that this coefficient is equal to 0,” i.e. this variable is “significant.”

Cross-Validation & Parameter Tuning

The Bias / Variance Trade-off



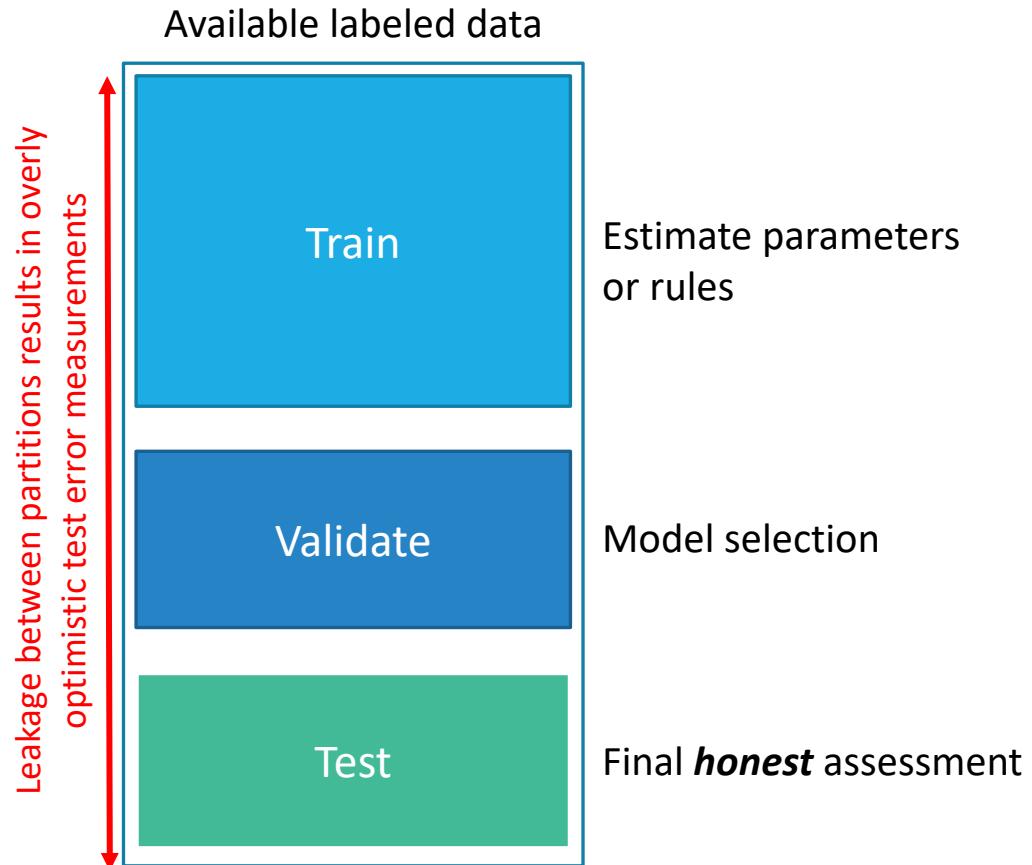
Total Error = Bias + Variance + Random

$$\text{Error} = (\hat{f}(x) - f(x))^2$$

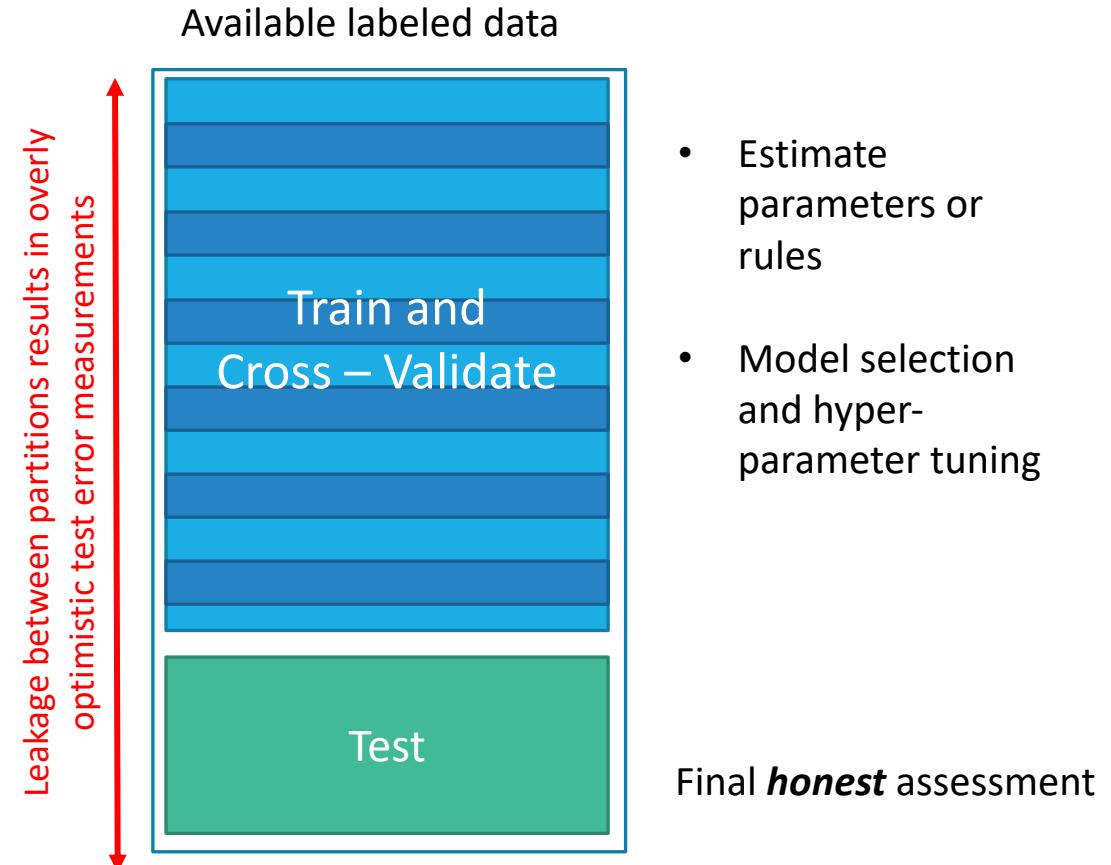
Bias = $E[\hat{f}(x)] - f(x)$ or the error that arises from a model's inability to replicate the fundamental phenomena represented by a data set.

Variance = $(\hat{f}(x) - E[\hat{f}(x)])^2$ or the error that arises from a model's ability to produce differing predictions from the values in a new data set.

Bias/Variance Trade-off in Practice: Honest assessment



Best suited for big data or linear models using traditional forward, backward, or stepwise selection.



Nearly always a more generalizable approach, but computationally intensive. Best suited for complex models with many hyper-parameters and small to medium sized data.

OLS Linear Regression Assumptions

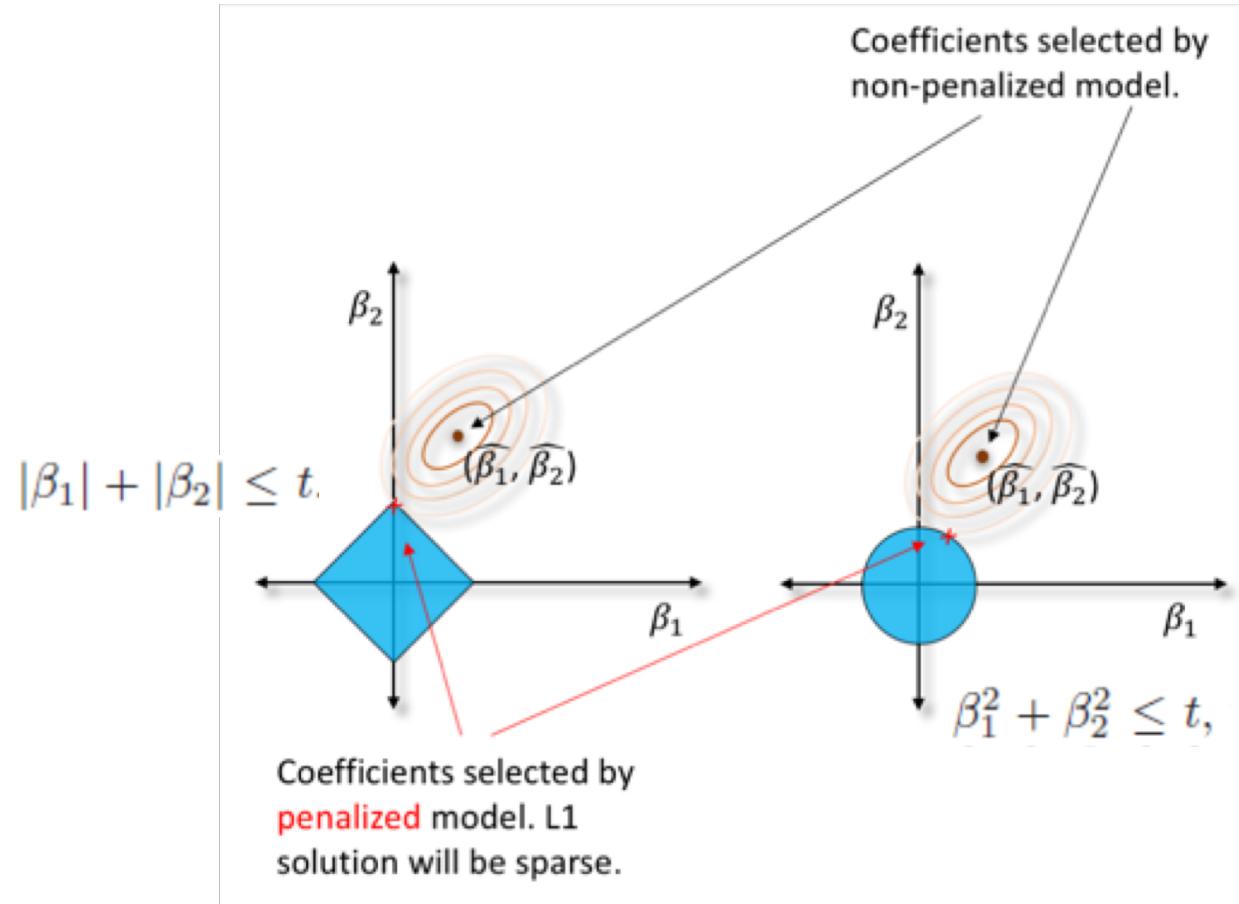
| Requirements | If broken ... |
|---|---|
| Linear relationship between inputs and targets; normality of y and errors | Inappropriate application/unreliable results; use a machine learning technique; use GLM |
| $N > p$ | Underspecified/unreliable results; use LASSO or Elastic Net penalized regression |
| No strong multicollinearity | Ill-conditioned/unstable/unreliable results; Use Ridge(L2/Tikhonov)/Elastic Net penalized regression |
| No influential outliers | Biased predictions, parameters, and statistical tests; use robust methods, i.e. IRLS, Huber loss, investigate/remove outliers |
| Constant variance/no heteroskedasticity | Lessened predictive accuracy, invalidates statistical tests; use GLM in some cases |
| Limited correlation between input rows (no autocorrelation) | Invalidates statistical tests; use time-series methods or machine learning technique |

Elastic Net - Modern Approach



Hui Zou and Trevor Hastie
Regularization and variable selection via the elastic net,
Journal of the Royal Statistical Society, 2005

Shrinkage/Regularization Method



Anatomy of Elastic Net: L1 & L2 Penalty

$$\tilde{\beta} = \min_{\beta} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} * \beta_j \right)^2 + \lambda \sum_{j=1}^p (\alpha * \beta_j^2 + (1 - \alpha) * |\beta_j|) \right\}$$

λ - Controls magnitude of penalties. Variable selection conducted by refitting model many times while varying λ . Decreasing λ allows more variables in the model.

L1/LASSO penalty – for variable selection.

L2/Ridge/Tikhonov Penalty – helps address multicollinearity.

α - tunes balance between L1 and L2 penalties.

Least squares minimization – finds β 's for linear relationship.

Elastic Net – Iteratively Reweighted Least Squares

Iteratively Reweighted Least Square complements fitting methods in the presence of the outliers by:

- Initially giving all observations equal weight then...
 - Train the model to estimate the β 's and find a linear relationship/equation

$$\tilde{\beta} = \min_{\beta} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} * \beta_j \right)^2 \right\}$$

← “Inner Loop”

- Calculate the residuals given these β 's/ linear equation
- Re-weight observations that cause high residuals to have a lower impact in the train model
- Re-train to find new β 's/linear equation
- Continue calculating residuals, re-weighting observations, and re-training until β 's become stable and weighted residuals are small...

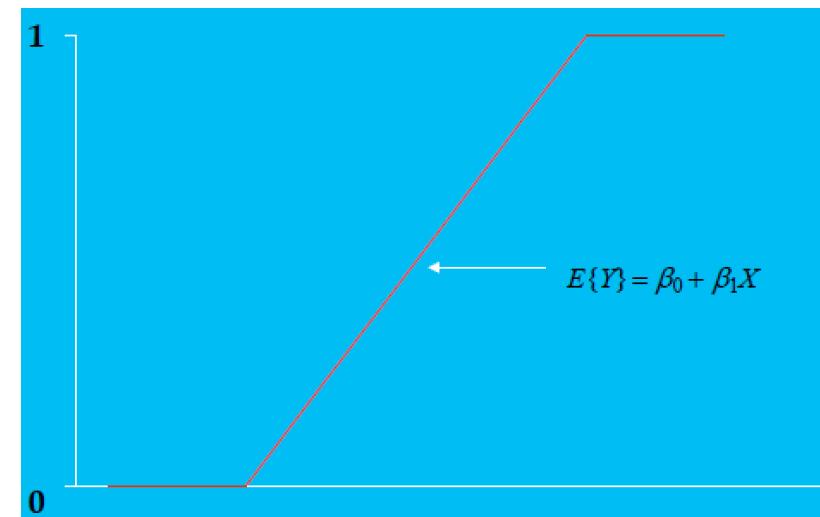
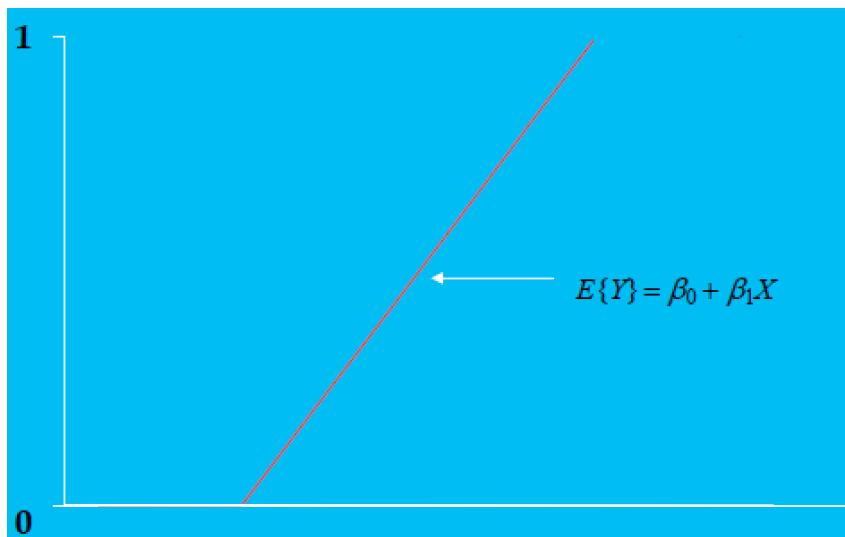
“Outer Loop”

Logistic Regression

Issues with Linear Regression

Best-Fit Linear Regression or Truncated Linear Regression:

- Boundary Problems
- Non-normal Error Terms
- Non-constant error variance



Non-Linear Logistic Model

- Solve the boundary constraint problems and also permit more plausible model that would model the relationship between the target and model parameters
- The model is constructed based on:
 - Ability to predict the probability of target occurrence based on various model parameter inputs
 - determine if there are any model parameters which are particularly salient to predict the model target

Linear Regression Prediction Formula

$$\hat{y} = \hat{w}_0 + \hat{w}_1 \cdot x_1 + \hat{w}_2 \cdot x_2$$

input measurement
intercept parameter estimate *estimate prediction estimate*

Choose intercept and parameter estimates to *minimize*.

squared error function

$$\sum_{\text{training data}} (y_i - \hat{y}_i)^2$$

Logistic Regression: Log of Odds

Copyright © 2014, SAS Institute Inc. All rights reserved.



Logistic Regression Prediction Formula

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2 \quad \text{logit scores}$$

RECALL: Odds=probability/(1-probability) AND Probability=Odds/(1+Odds)

Logit Link Function

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 \cdot x_1 + \hat{w}_2 \cdot x_2 = \text{logit}(\hat{p})$$

$$\hat{p} = \frac{1}{1 + e^{-\text{logit}(\hat{p})}}$$

To obtain prediction estimates, the logit equation is solved for \hat{p} .

Interpreting Logistic Regression

Probability to Log Odds:

For categorical

- p = event rate for that level
- odds = $p/(1-p)$
- odds ratio = $\text{odds}_{\text{level}} / \text{odds}_{\text{reference level}}$
- log odds ratio = $\ln(\text{odds ratio})$

Log odds ratio against reference level = 1.2

Odds ratio against reference level = $e^{1.2} = 3.32$

Probability/event rate in training data = $3.32/(1 + 3.32) = 0.76$

"Holding all other variables constant, a person being male changes the odds of the event occurring by a factor of 3.32 over the reference level on average."

$$\hat{y} = \text{"log odds"} = \log(p/(1 - p)) = 1.7 - 0.54 * \text{age} + 1.2 * \text{male}$$

For interval:

- p = change in event rate for one unit increase; this is **not** constant
- odds = $\text{odds}_{\text{level}} - \text{odds}_{\text{level}+1}$, this **is** constant
- Log odds = $\ln(\text{odds})$

Log odds = -0.54

Odds = $e^{-0.54} = 0.58$

"Holding all other variables constant, for a one unit increase in age, the odds of the event occurring change by a factor of 0.58 on average."

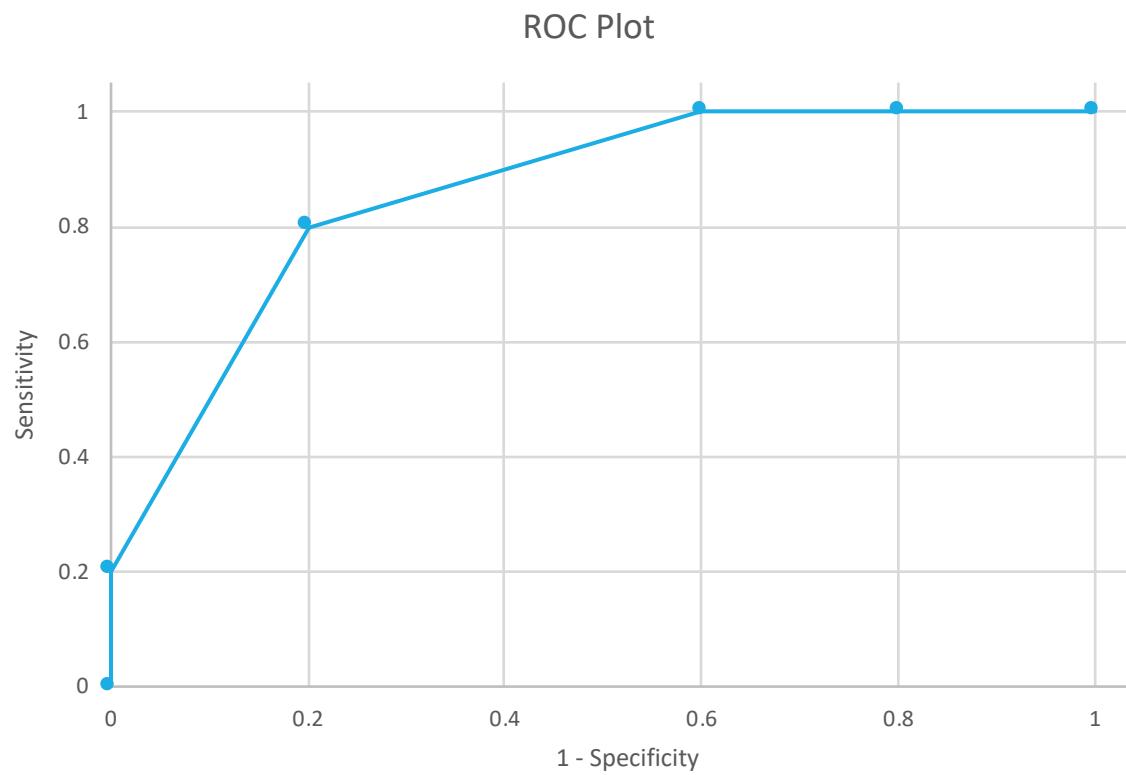
Confusion Matrix

| | | True condition | |
|---------------------|------------------------------|----------------------------------|---------------------------------|
| | | Condition positive | Condition negative |
| Total population | | | |
| Predicted condition | Predicted condition positive | True positive, Power | False positive, Type I error |
| | Predicted condition negative | False negative, Type II error | True negative |

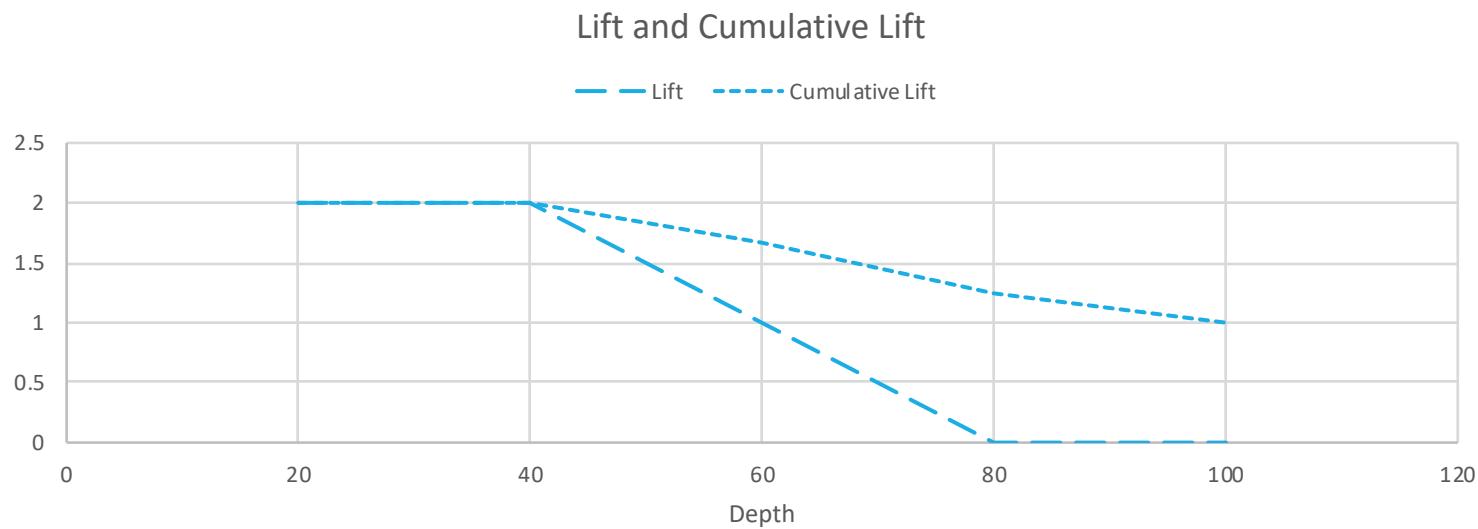
Confusion Matrix & Classification Table

| Actual | Predicted | Cutoff | TP | FP | FN | TN | Sens. | 1 - Spec. |
|--------|-----------|--------|----|----|----|----|-------|-----------|
| 1 | 0.85 | 0 | 5 | 5 | 0 | 0 | 1 | 1 |
| 1 | 0.75 | 0.2 | 5 | 4 | 0 | 1 | 0.8 | 1 |
| 1 | 0.7 | 0.4 | 5 | 3 | 0 | 2 | 0.6 | 1 |
| 1 | 0.65 | 0.6 | 4 | 1 | 1 | 4 | 0.2 | 0.8 |
| 0 | 0.65 | 0.8 | 1 | 0 | 4 | 5 | 0 | 0.2 |
| 1 | 0.55 | 1 | 0 | 0 | 5 | 5 | 0 | 0 |
| 0 | 0.55 | | | | | | | |
| 0 | 0.45 | | | | | | | |
| 0 | 0.3 | | | | | | | |
| 0 | 0.1 | | | | | | | |

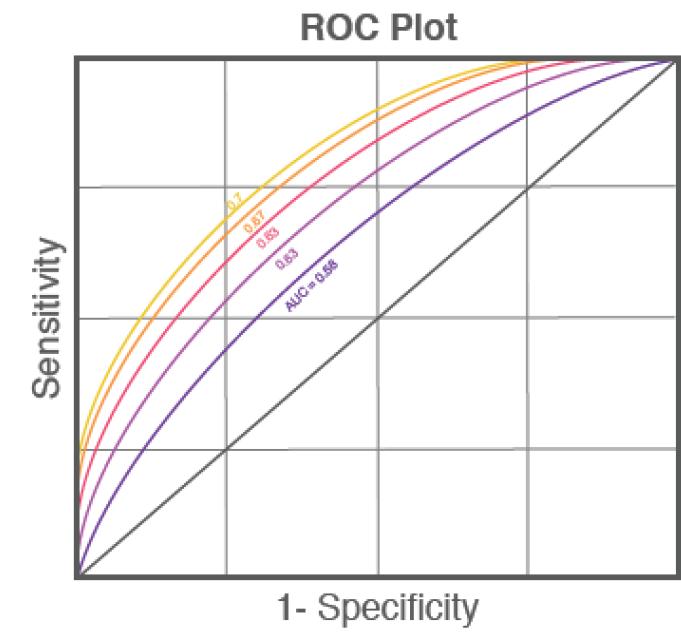
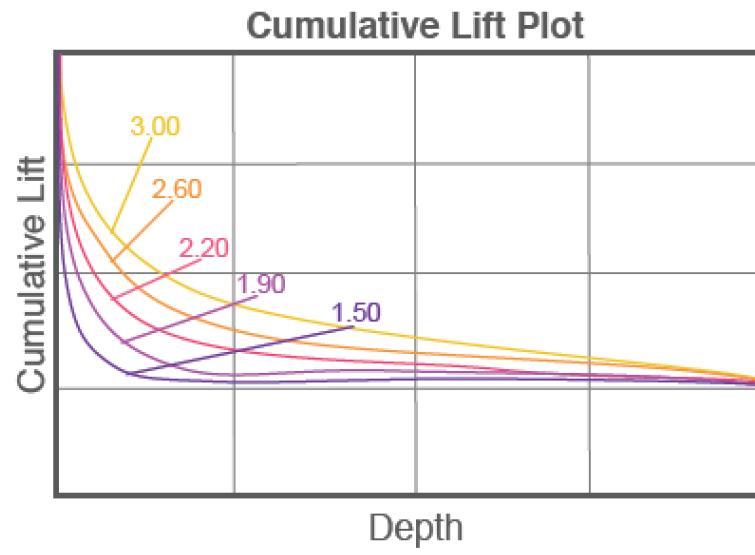
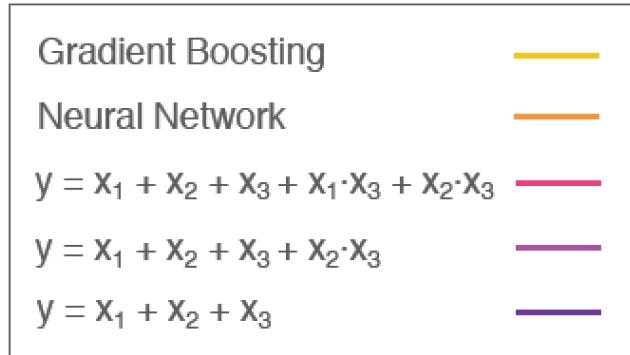
ROC Curve Example



LIFT PLOT

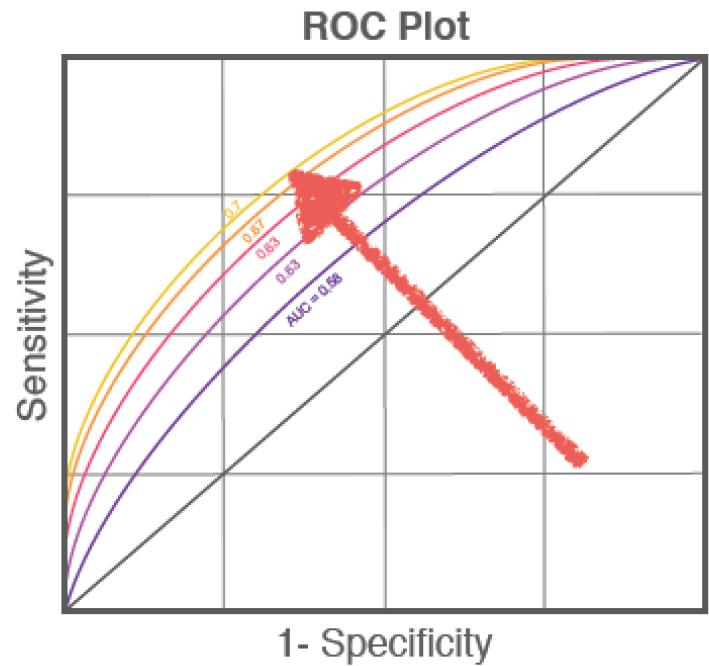
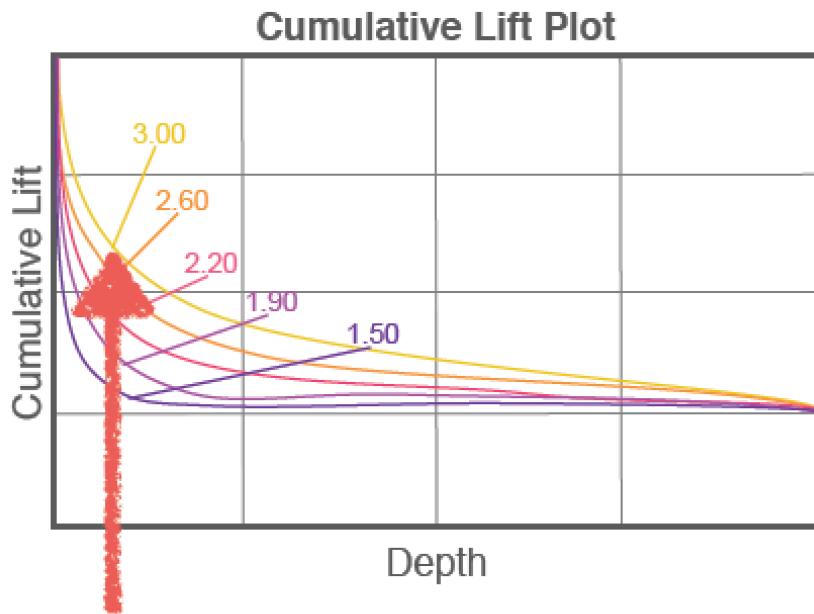


Interpreting Assessment Measures



- Area under the ROC curve (AUC) is bounded between 0 and 1. Values below and including 0.5 indicate serious problems with the model. Values above 0.5, as they approach 1, indicate a better model.
- AUC is interpreted as “The expectation that a uniformly drawn random positive is ranked before a uniformly drawn random negative.”

| | |
|---|--|
| Gradient Boosting | |
| Neural Network | |
| $y = X_1 + X_2 + X_3 + X_1 \cdot X_3 + X_2 \cdot X_3$ | |
| $y = X_1 + X_2 + X_3 + X_2 \cdot X_3$ | |
| $y = X_1 + X_2 + X_3$ | |



- Lift is typically measured at a certain percentile, say the 10th
- Higher lift indicates a better model
- Lift is interpreted as: “In the 10th percentile of highest predicted probabilities, this model predicted 3 times more events correctly than a random selection of 10% of the data.”