Applied Regression Analysis

3

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Conditional v. Marginal Distribution **Marginal Distribution**: full range of x  
**Conditional Distribution**: distribution of conditional on range of

1. If has no forecasting power marginal distribution = conditional distribution
2. If has forecasting power, marginal distribution conditional distribution and standard deviation of will be less in the conditional distribution

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Least Squares is the prediction line or fitted values  
The residual is the discrepancy between the fitted and observed **Least squares chooses and x to minimize:**

**because it treats positiva and negative errors equally** -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Properties

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Correlation and Covariance   
**Correlation = 0 does no mean the variables are unrelated** -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Simple Linear Regression Model

Error is independent of x with **mean** 0 (not median) and fixed but unknown variance , constant over   
  
  
 least squares of estimate of the intercept,  least squares of estimate of the slope,   **They can both change when data changes**  
\*  **representes anything left, not captured in the linear function**

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Sampling distribution -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex For -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex For -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Joint distribution of and and can be dependent the estimation error in the slop is correlated with the estimation error in the intercept  
  
Usually they are negative correlated, and the correlation decreases with more x-spread

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Estimation of error variance or   
 is the number of regression coefficients

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Testing and   
**t-statistic** for this test is   
If is true then   
**Confidence intervals**: since

1)Center is your estimate , 2) length is how sure you are about your estimate, 3) if you increase the interval you decrease type 1 error but increase type 2 error -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex P-value Is the we need to set to just change your answer about a hypothesis  
-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Forecasting and Prediction Intervals   
The variance of our prediction error is: **Confidence/prediction interval for**   
  
 -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Polynomial Regression   
To check if you need a quadratic term, check for . **Watch out for overfitting**  
-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Log-log model where -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex MLR model

is the **average** change in per unit of change -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Inference for coefficients is a multivariate normal:   
 -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Dummy/Categorical variables $\mathbb{E}[Y \mid X]=\beta\_{0}+\beta\_{1} \mathbbm{1}\_{[X=2]}+\beta\_{2} \mathbbm{1}\_{[X=3]}+\cdots+\beta\_{R-1} \mathbbm{1}\_{[X=R]}$ -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Variable Interaction   
If you run all the variables together, the result is different than if you run each variable separately. first one you assume is the same for all in the second one each has its own -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Fit Measure   
If SSR = SST perfect fit. SSR = variation in explained by the regression. SSE = variation in that is left unexplained.  
If   
   
 Adjusted -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Multicollinearity Strong **linear** dependence between some of the covariates in a multiple regression  
+ change in one variable leads to change in others  
+ large uncertainty about the , you may fail to reject   
+ bigger intervals limit your ability to predict -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex F-test Joint hypothesis test. Tries to formalize what is a "big"   
  
If there is info about in we reject but we don’t know in which the info is. You can use a "partial" F-test. -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Non Constant Variance Plotting vs is your number 1 tool for finding fit problems. One possible solution is to stabilize the variance by transforming the model. or or something else.  
**You cannot compare values of transformations** -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Clustering Each observation is allowed to have unkown correlation with a **small** number others in a known pattern.   
Only standard errors change, the slope is the same slope for everyone. There are no assumptions for and

3

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Causality Best way to understand causality is to use randomized experiments. there must be no systematic relationship between units and treatments ( ) is independent

Where . Also is usually binary.

**Estimation:**

* For each subgroup you should do a random experiment to obtain ATE. If we can reject , is not effective to increase
* If you add an interaction your ATE changes to . **Clarification**:
* We assume the function is linear

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Causality without randomization We want the change in caused by moving independently all other influences. We need to be randomy assigned given X (no systematic relationship between and ).

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Binary Outcomes We need a model that gives mean/probability between 0 and 1. . And we need to assume that the average of will be linear.

**Def 1** (**Generalized Linear Model**).

Two main functions:

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Logistic Regression

$$\begin{gathered}
\begin{gathered}
\mathbb{P}\left( Y = 1 \mid X\_1 , X\_2 , X\_{d} \right) = S\left( x'\beta \right) = \frac{\exp\left( \beta\_0 + \beta\_1 X\_1 + \ldots +\beta\_{d} X\_{d }\right) } {1 + \exp\left( \beta\_0 + \beta\_1 X\_1 \ldots + \beta\_{d}X\_{d} \right) }
\shortintertext{to get linear odds ratio:}
\log\left( \frac{\mathbb{P}\left( Y = 1 \mid X\_1 \ldots X\_{d} \right)}{\mathbb{P}\left( Y = 0 \mid X\_1 \ldots X\_{d} \right)} \right) = \beta\_0 + \beta\_1 X\_1 \ldots + \beta\_{d} X\_{d}\\
\shortintertext{\textbf{Odds Ratio:}}
O R(x)=\frac{ \mathbb{P}[y=1 \mid x]}{ \mathbb{P} [y=0 \mid x]}=\frac{p}{1-p} \Rightarrow O R(x+1)=e^{\beta\_{j}} O R(x)
\end{gathered}
\end{gathered}$$

**Def 2**. *: change in the odds for each unit increase in holding everything else constant.*

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Classification Now the classification error is

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Confusion Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Prediction outcome** | |  |
|  | **1** | **0** |  |
| **1** |  |  |  |
| **0** |  |  |  |

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex ROC Curve

* Sensitivity:
* Specificity:

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Precision-Recall Curve

* Precision:
* Recall: = Sensitivity

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Model Building -1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Model Selection

1. Split data into testing/training samples.
2. Fit model on the training data.
3. Predict on the testing data.
4. Compute MSE/MAE: We want it to be low We only use testing data and training coefficients

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex [short]Variable Selection More variables always means higher . **We want to maximize fit but minimize complexity.**

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Penalized Regression

**Def 3**. ***Information Criteria*** *How well our model would have predicted the data, regardless of what you have estimated for*

* **BIC**: Bayes Information Criterion.
* where is the number of observations, is the number of parameters.
* **AIC**: Akaike Information Criterion.

AIC tends to prefer more complicated models because usually .  
We can interpret BIC as probabilities with

**Def 4**. ***LASSO***

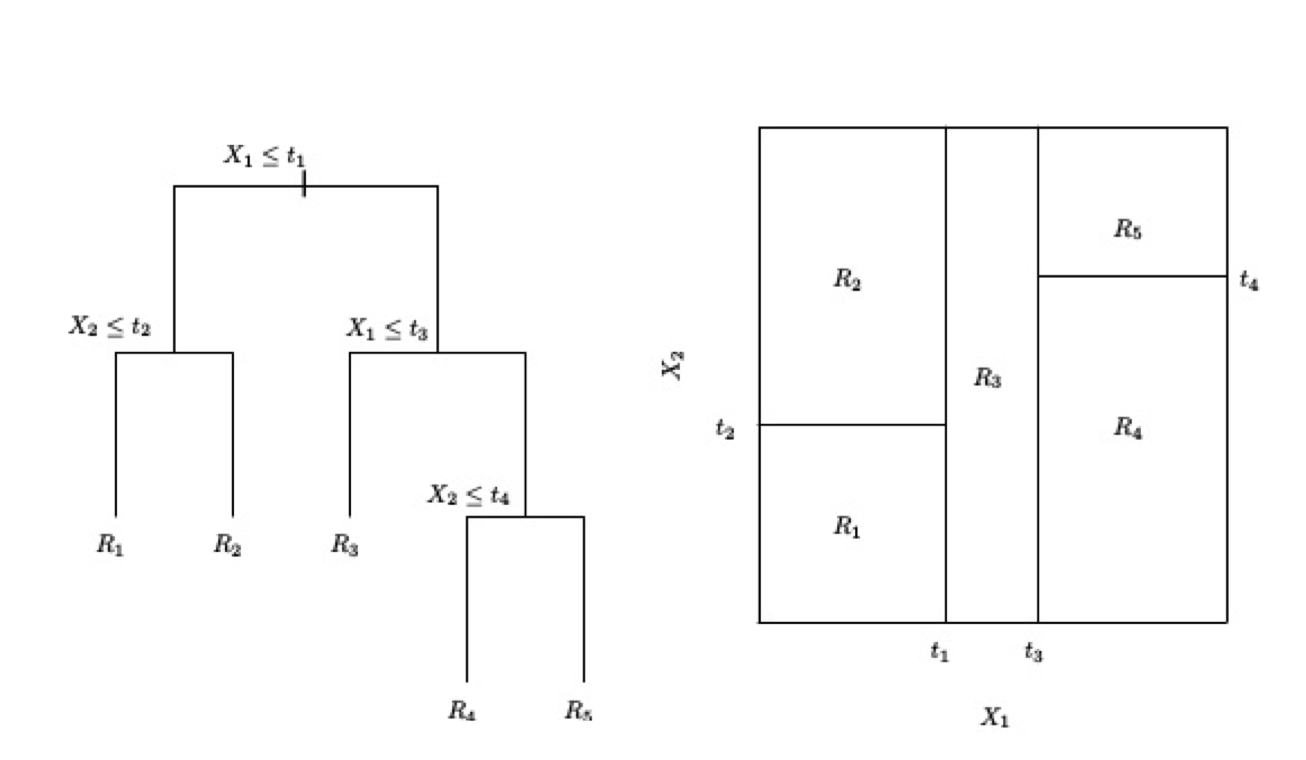
2

measures the models complexity. determines the penalty. we chose it with cross-validation.  
**Stepwise is a specific path through these models, but LASSO "searches" all combinations at once.**

Remember to refit!

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Trees Fit nonlinearities and selection automatically. We don’t need to specify transformations ahead of time ( )

* Initialize
* For each one at a time, find cutoff that most improves the predictions
* Whatever single is the best, keep that one Split
* Repeat, each time trying to improve one the prior iteration



-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex Time Series Analysis

**Def 5**. ***Autoregressive Model***  
*Previous lag values do not help predict if you already know .*

Correlation between and is called auto-correlation.

If , the series explodes. If , we have a random walk. If , the values are mean reverting. The random walk has some special properties ... , and is called the "drift parameter". The random walk without drift is a common model for simple processes , etc. - the expectation of what will happen is always what happened most recently.

-1ex plus -.5ex minus -.2ex 0.5ex plus .2ex AR(p) Model

-1ex plus -.5ex minus -.2ex 1ex plus .2ex Trending Series Just put “time” in the model -1ex plus -.5ex minus -.2ex 1ex plus .2ex Periodic Series Period model :