Linear Regression

Elsayed Hemayed

The original slides are from EMC Data Analytics Course and from Udmey course by SuperDataScience Team

Overview of Linear Regression

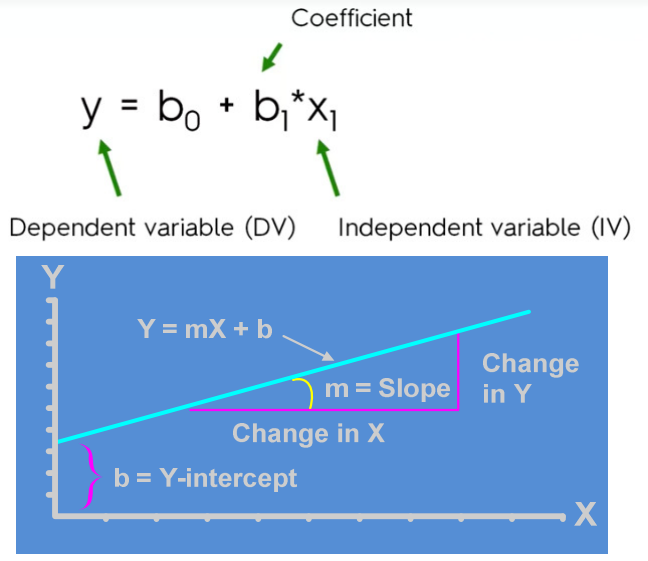
- General description of regression models
- Technical description
- Handling Categorical Values
- Building A Model
- Diagnostics for validating the linear regression model
- Overfitting Problem and Regularization
- The Reasons to Choose (+) and Cautions (-) of the linear regression model

2

Regression

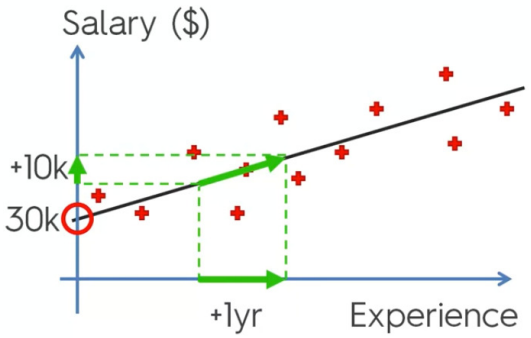
- Regression focuses on the relationship between an outcome and its input variables.
  - In other words, we don't just predict the outcome, we also have a sense of how changes in individual drivers affect the outcome.
- Examples:
  - Income as a function of years of education, age, gender
  - House price as function of median home price in neighborhood, square footage, number of bedrooms/bathrooms
  - Neighborhood house sales in the past year based on unemployment, stock price etc.

Simple Linear Regression



4

Simple Linear Regression



$y = b_0 + b_1 * x$

Salary = b_0 + b_1 *Experience

5

Linear Regression

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

Dependent variable (DV) Independent variables (IVs)

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

Constant Coefficients

6

Linear Regression

- Used to estimate a continuous value as a linear (additive) function of other variables

$$y = b_0 + b_1 x_1 + b_2 x_2 + ...$$

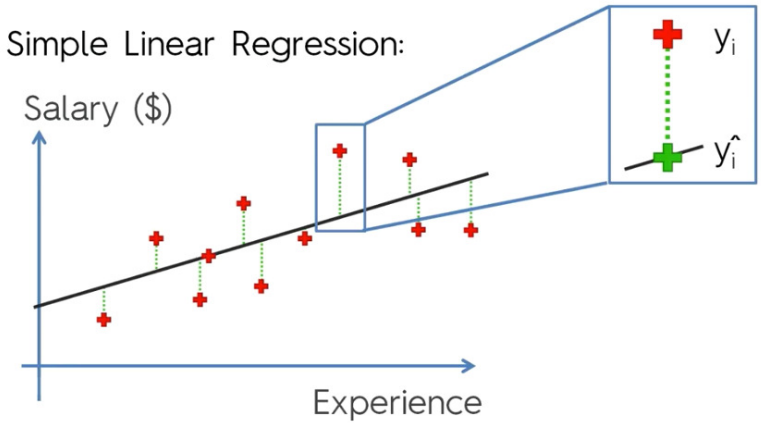
- The preferred method for almost any problem where we are predicting a continuous outcome
  - Try this first; if it fails, then try something more complicated

Technical Description

$$y = b_0 + b_1 x_1 + b_2 x_2 + ...$$

- Solve for the bi
  - Ordinary Least Squares
- Categorical variables are expanded to a set of indicator variables, one for each possible value.

Ordinary Least Squares



OLS: Minimize Sum$(y_i - \hat{y_i})^2$

9

Representing Categorical Variable

$$\text{income} = b_0 + b_1 \text{age} + b_2 \text{yearsOfEducation} + b_3 \text{gender} + b_4 \text{state}$$

- *State* is a categorical variable: 50 possible values.
- Expand it to 49 indicator (0/1) variables:
  - The remaining level is the "default level"
  - This is done automatically by standard packages
- *Gender* is categorical, too, but binary
  - so one variable: *genderMale*, which is 0 for females

Regression Example

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

11

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + ???$$

12

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State	Dummy Variables: New York	Dummy Variables: California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*D_1$$

13

Dummy Variable Trap

Profit	R&D Spend	Admin	Marketing	State	Dummy Variables: New York	Dummy Variables: California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*D_1 + b_5*D_2$$

14

Continuous Vs. Categorical Variables

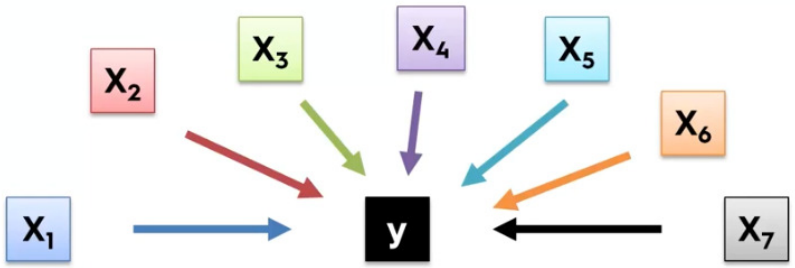
- General linear regression model:
$$y = b_0 + b_1 x_1 + b_2 x_2 + ...$$
- Independent variables (X's):
  - Continuous: age, income, height → use numerical value
  - Categorical: gender, city → use dummies.
- Dependent variable (y):
  - Continuous: consumption, time spent → use numerical values
  - Categorical: yes/no → use dummies.

15

What do the Coefficients b_i *M*ean?

- Change in *y* as a function of unit change in x_i
  - all other things being equal
- Example: income in units of \$10K, years in age, b_{age}= 2
  - For the same gender, years of education, and state of residence, a person's income increases by 2 units (20K)for every year older
- Standard packages also report the significance of the b_i: probability that, in reality, b_i = 0
  - b_i "significant" if P(b_i = 0) is small

Building A Model



Which X to use?

17

Building A Model

1. All-in
2. Backward Elimination
3. Forward Selection
4. Bidirectional Elimination
5. Score Comparison

Stepwise Regression

18

Backward Elimination

1. Select a significance level to stay in the model (e.g., SL=0.05)
2. Fit the full model with all possible predictors
3. Consider the predictor with the highest p-value. If p>SL then go to step 4, otherwise goto step 5
4. remove the selected predictor, and fit model without it.
5. The model is ready

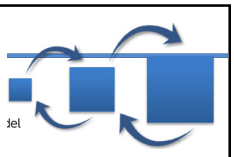
19

Forward Selection

1. Select a significance level to enter the model (e.g., SL=0.05)
2. Fit the full model with all possible predictors, select the one with the lowest p-value
3. Keep this variable and fit all possible model with one extra predictor added to the one(s) you already have.
4. Consider the predictor with the lowest p-value. If p<SL, go to step 3 otherwise go to step 5.
5. The model is ready

20

Bidirectional Elimination



1. Select a significance level to enter and to stay in the model (e.g., SLENTER=0.05, SLSTAY=0.05).
2. Preform the next step of Forward Selection (new Variables must have p<SLENTER to enter)
3. Perform all steps of Backward Elimination (old variables must have P<SSLSTAY to stay)
4. Repeat step 2 and 3 till no new variables can enter and no new variables can exit. Then the model is ready.

21

All Possible Models

1. Select a criteria of goodness of fit
2. Construct all possible regression models (2^N -1 total combinations for N predictors)
3. Select the one with the best criterion

22

Simple Linear Regression Assumptions

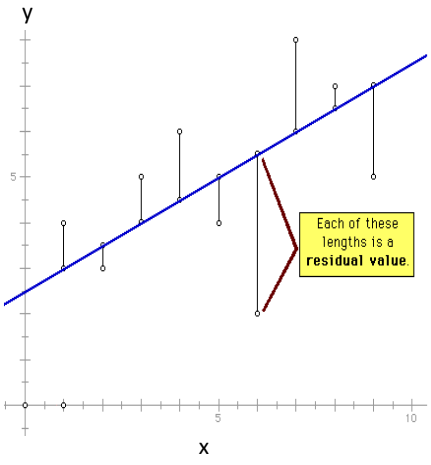
- Linearity
- Observations are independent
  - Based on how data is collected.
  - Check by plotting residuals in the order of which the data was collected.
- Constant variance
  - Check using a residual plot (plot residuals vs. $\hat{y}$).

23

The Residual

The residual values provide us some measure of how well the line fits the data, that is, the **goodness of fit**.

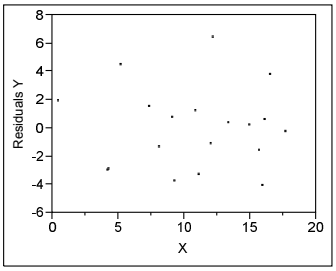
It is calculated as the difference between the actual y value and the predicted y value.



24

Diagnostics: Residual Plot

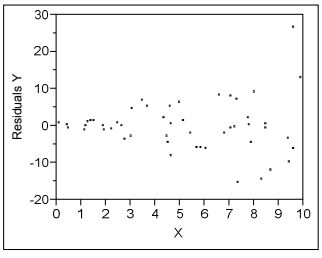
- A residual plot is used to check the assumption of constant variance and to check model fit (is a line a good fit).
- Good residual plot: no pattern

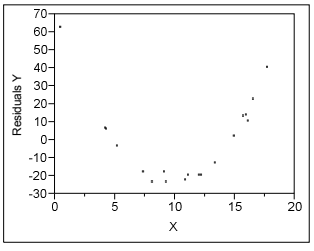


25

Diagnostics

- Left: Residuals show non-constant variance.
- Right: Residuals show non-linear pattern.





26

Test for Parameters (F-test)

- Test whether the true y-intercept is different from 0.
  - $H_0: \beta_0 = 0$
    $H_A: \beta_0 \neq 0$
- Test whether the true slope is different from 0.
  - $H_0: \beta_1 = 0$
    $H_A: \beta_1 \neq 0$
  - Note: For simple linear regression this test is equivalent to the overall F-test (p-value).

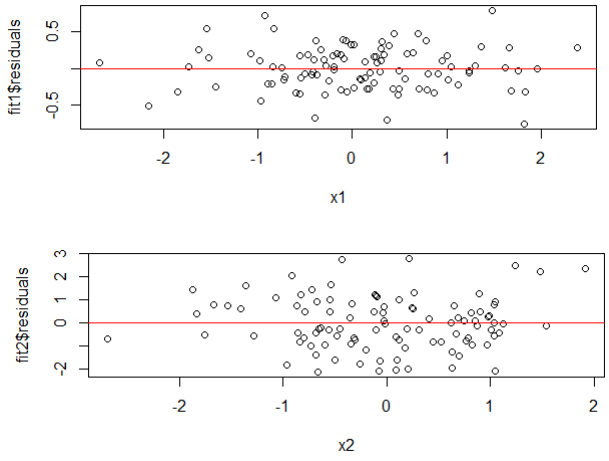
27

Linear Regression Example

```
n <- 100; nosim <- 1000
x1 <- rnorm(n);  x2<-rnorm(n);
y <- x1 + rnorm(n, sd = .3)
fit1<-lm(y~x1)
summary(fit1)
plot(x1, fit1$residuals); abline(a=0,b=0,col="red")

#trial
fit2<-lm(y~x2)
plot(x2,fit2$residuals); abline(a=0,b=0,col="red")
```

28

Residual Plots



29

Model Summary

Call: lm(formula = y ~ x1)

Residuals: Min 1Q Median 3Q Max

-0.82764 -0.19950 0.03874 0.19408 **0.62315**

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	**-0.0009003**	0.0295453	-0.03	0.976	
x1	**1.0125675**	0.0309821	32.68	<2e-16	***

--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: **0.2952** on 98 degrees of freedom

Multiple R-squared: 0.916, **Adjusted R-squared: 0.9151**

F-statistic: 1068 on 1 and 98 DF, **p-value: < 2.2e-16**

Try fit1<-lm(y~x1-1)

30

Multiple Variables General Rules

- Omitting variables results in bias in the coefficients of interest unless their regressors are uncorrelated with the omitted ones.
- Including variables that we shouldn't have increases standard errors of the regression variables.
- The model must tend toward perfect fit as the number of non redundant regressors approaches n (the correct regressors).
- R2 increases monotonically as more regressors are included.
- The SSE decreases monotonically as more regressors are included.

31

Multiple Variables Example

```
n <- 100; nosim <- 1000
x1 <- rnorm(n);
x2 <- rnorm(n);
x3 <- rnorm(n);
y <- x1 + x2+ rnorm(n, sd = .3)

fit1<-lm(y~x1)
fit2<-lm(y~x2)
fit3<-lm(y~x3)
fit12<-lm(y~x1+x2)
fit13<-lm(y~x1+x3)
fit123<-lm(y~x1+x2+x3)
```

```
# Data exploration
pairs(y~x1+x2+x3)

# Model Evaluation
summary(fit1)
summary(fit12)
summary(fit123)
anova(fit1,fit12,fit123)
```

32

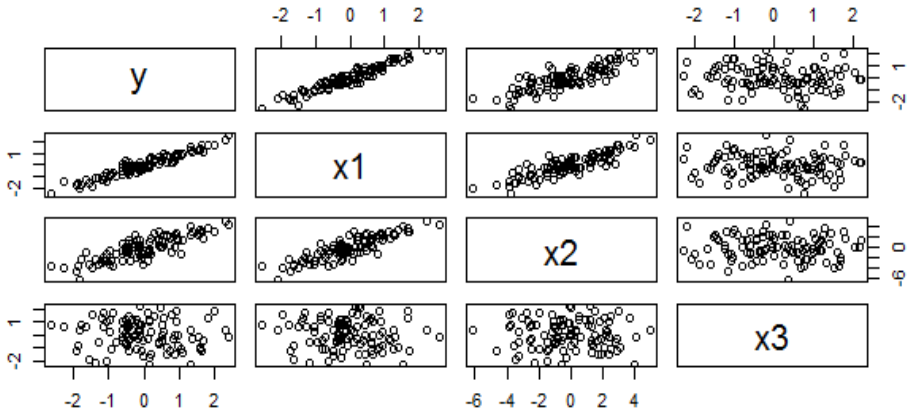
Dependent Variables Example

```
n <- 100; nosim <- 1000
x1 <- rnorm(n);
x2 <- 2*x1+rnorm(n);
x3 <- rnorm(n);
y <- x1 + x2+ rnorm(n, sd = .3)

fit1<-lm(y~x1)
fit2<-lm(y~x2)
fit3<-lm(y~x3)
fit12<-lm(y~x1+x2)
fit13<-lm(y~x1+x3)
fit123<-lm(y~x1+x2+x3)
```

```
# Data exploration
pairs(y~x1+x2+x3)

# Model Evaluation
summary(fit1)
summary(fit12)
summary(fit123)
anova(fit1,fit12,fit123)
```

33

Dependent Variables Example



34

Diagnostics

- Hold-out data
  - Does the model predict well on data it hasn't seen?
- N-fold cross-validation
  - Partition the data into N groups.
  - Fit N models, holding out each group, and calculate the residuals on the group.
  - Estimated prediction error is the average over all the residuals.
- R^2 : The fraction of the variance in the output variable that the model can explain.
  - It is also the square of the correlation between the true output and the predicted output. You want it close to 1.

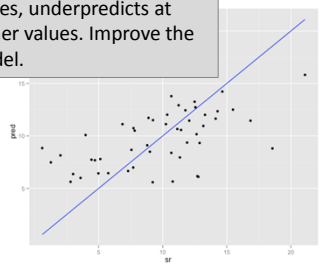
Diagnostics (Continued)

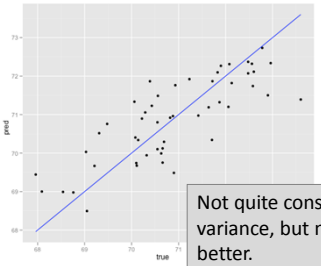
- Sanity check the coefficients
  - Do the signs make sense? Are the coefficients excessively large?
    - Wrong sign is an indication of correlated inputs, but doesn't necessarily affect predictive power.
    - Excessively large coefficient magnitudes may indicate strongly correlated inputs; you may want to consider eliminating some variables, or using regularized regression techniques.
      - Ridge, Lasso
    - Infinite magnitude coefficients could indicate a variable that strongly predicts a subset of the output (and doesn't predict well on the rest).
      - Plot output vs. this input, and see if you should segment the data before regressing.

Diagnostics (Continued)

- **Plot it!**
  - Prediction vs. true outcome
- Look for:
  - Systematic over/under prediction
  - Non-consistent variance
    - The data cloud should be symmetric about the line of true prediction
  - Glaring outliers
- You will see other diagnostic plots in the lab

overpredicts for low true values, underpredicts at higher values. Improve the model.



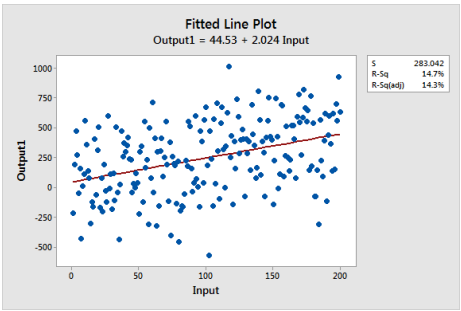


Not quite consistent variance, but much better.

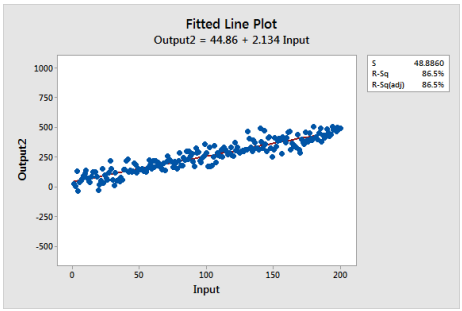
Comparing R-Squared

Regression equations: Output = 44 + 2 * Input
Input is significant with P < 0.001 for both models

R-Sq = 14.7%



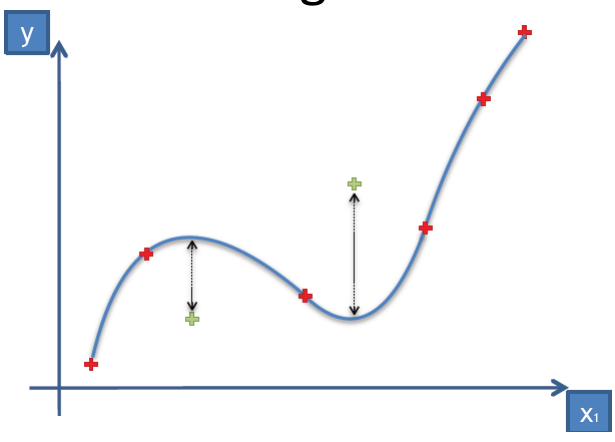
R-Sq = 86.5%



Coefficients, P-Value, and R-Squared

- The coefficients estimate the trends while R-squared represents the scatter around the regression line.
- The interpretations of the significant variables are the same for both high and low R-squared models.
- Low R-squared values are problematic when you need precise predictions.

Overfitting Problem



The fitting curve fits the data perfectly well,
but if we look at new observations, we can get large errors

40

Examples of Regularization
(Ridge – Lasso – Elastic Net)

$$\text{Minimize} \sum_{i=1}^{n} \left(y^i - (b_0 + b_1 x_1^i + \ldots + b_m x_m^i)\right)^2$$

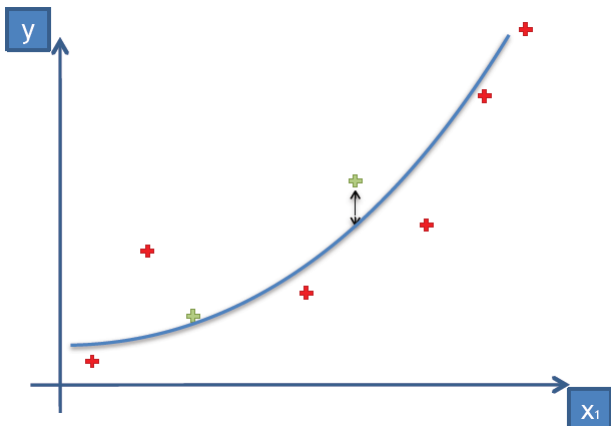
$$\text{Minimize} \sum_{i=1}^{n} \left(y^i - (b_0 + b_1 x_1^i + \ldots + b_m x_m^i)\right)^2 + \lambda(b_1^2 + \ldots + b_m^2)$$

$$\text{Minimize} \sum_{i=1}^{n} \left(y^i - (b_0 + b_1 x_1^i + \ldots + b_m x_m^i)\right)^2 + \lambda(|b_1| + \ldots + |b_m|)$$

$$\text{Minimize} \sum_{i=1}^{n} \left(y^i - (b_0 + b_1 x_1^i + \ldots + b_m x_m^i)\right)^2 + \lambda_1(|b_1| + \ldots + |b_m|) + \lambda_2(b_1^2 + \ldots + b_m^2)$$

41

Regularization Reduces Overfitting



42

Linear Regression - Reasons to Choose (+) and Cautions (-)

Reasons to Choose (+)	Cautions (-)
Concise representation (the coefficients)	Does not handle missing values well
Robust to redundant variables, correlated variables Lose some explanatory value	Assumes that each variable affects the outcome linearly and additively Variable transformations and modeling variable interactions can alleviate this A good idea to take the log of monetary amounts or any variable with a wide dynamic range
Explanatory value Relative impact of each variable on the outcome	Can't handle variables that affect the outcome in a discontinuous way Step functions
Easy to score data	Doesn't work well with discrete drivers that have a lot of distinct values For example, ZIP code