Linear Regression

The Descriptive Angle - I

Explanatory vs. Predictive

- Explain/describe population relationships
- Small sample, few variables
- Retrospective
- Find good fitting regression model
- Confidence intervals, hypothesis test, p-value

- Predict values of new records
- Large sample, many variables
- Prospective
- Regression with high predictive power
- Predictive power on holdout data

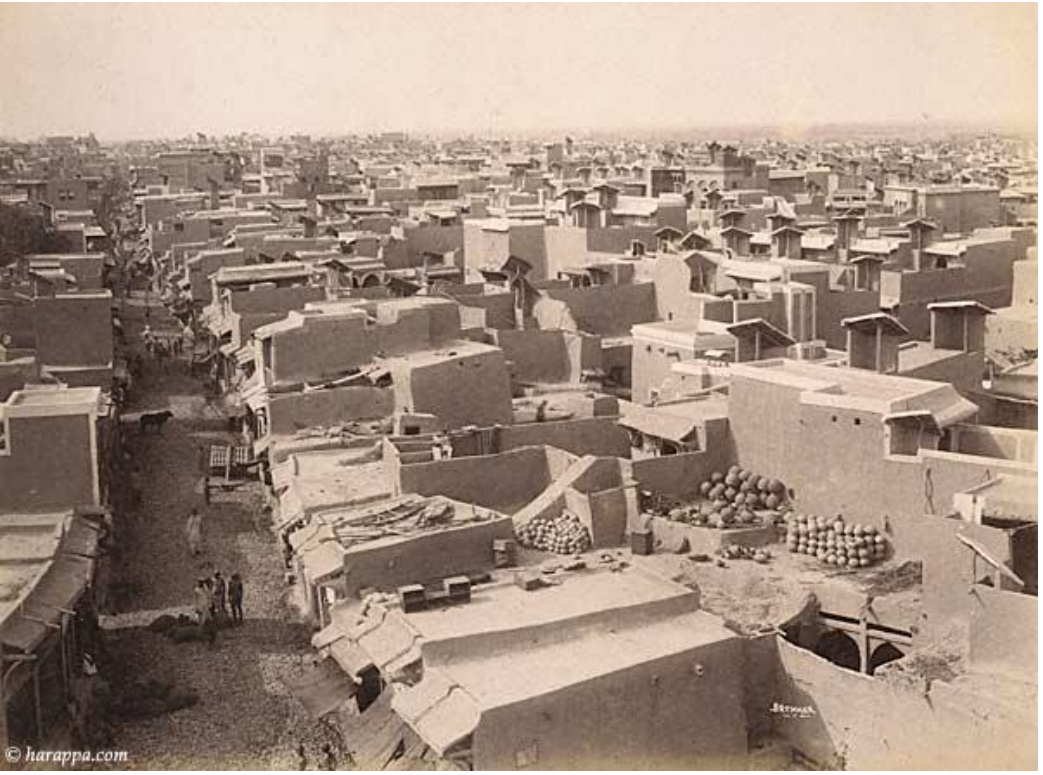
Basic courses in Statistics teach how to use linear regression for explanation/description

Reminder: Linear Regression as an explanatory/descriptive tool

- Running regressions
- Interpreting output
- Checking model validity
- Testing hypotheses

Example: Housing prices in MidCity

Housing Prices.xlsx :
128 recent sales of single-family houses in MidCity



Price: Final sale price

SqFt: Floor area in ft^2

Bedrooms: # bedrooms

Bathrooms: # bathrooms

Offers: # offers made on the house prior to sale

Brick: Brick construction? (yes/no)

Neighborhood: East/West/North

Explanatory Objective:

Estimate and interpret the pricing structure of houses in MidCity

Predictive Objective:

Predict the sale price of a house that is on the market

Data sample

Home	Price	SqFt	Bedrooms	Bathrooms	Offers	Brick	Neighborhood
1	114300	1790	2	2	2	No	East
2	114200	2030	4	2	3	No	East
3	114800	1740	3	2	1	No	East
4	94700	1980	3	2	3	No	East
5	119800	2130	3	3	3	No	East
6	114600	1780	3	2	2	No	North
7	151600	1830	3	3	3	Yes	West
8	150700	2160	4	2	2	No	West
9	119200	2110	4	2	3	No	East
10	104000	1730	3	3	3	No	East
11	132500	2030	3	2	3	Yes	East
12	123000	1870	2	2	2	Yes	East
13	102600	1910	3	2	4	No	North
14	126300	2150	3	3	5	Yes	North
15	176800	2590	4	3	4	No	West
16	145800	1780	4	2	1	No	West
17	147100	2190	3	3	4	Yes	East
18	83600	1990	3	3	4	No	North
19	111400	1700	2	2	1	Yes	East
20	167200	1920	3	3	2	Yes	West

Fitting a regression model for explanation

1. **Choose** X variables
2. **View** scatter plots (careful!)
3. **Fit** a simple/multiple regression model to data. Get estimated model
4. Check **validity** of model assumptions
5. Use estimated model to **test/infer** relationship in the population

CHOOSE VARIABLES

Based on theory or domain knowledge

Real-estate agent claims that the collected variables should all affect House Price

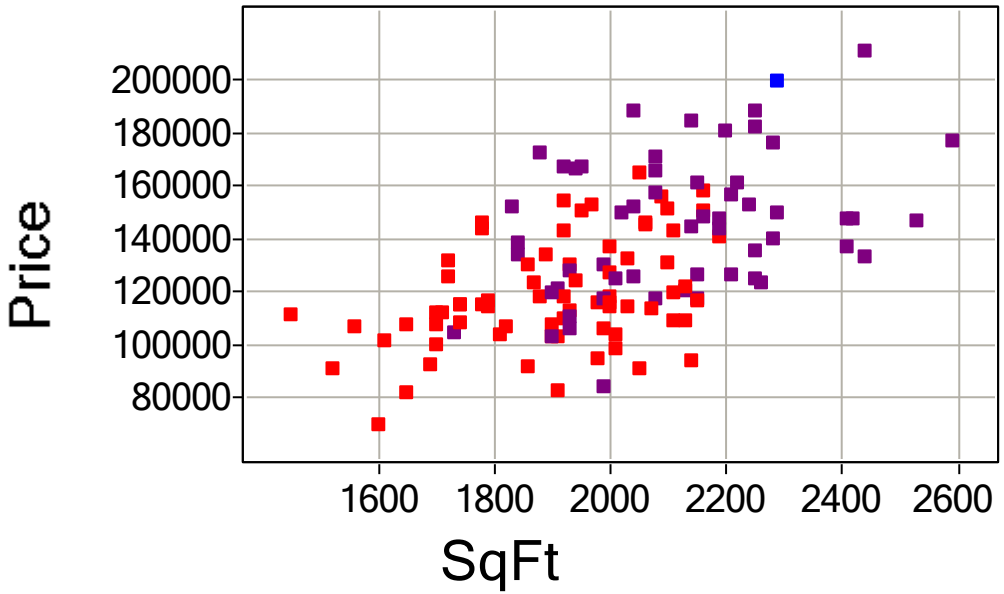
EXAMINE SCATTER PLOTS

Data exploration: TIBCO Spotfire

Interactive!

Right click > Properties

Edit> Copy Special > Visualization



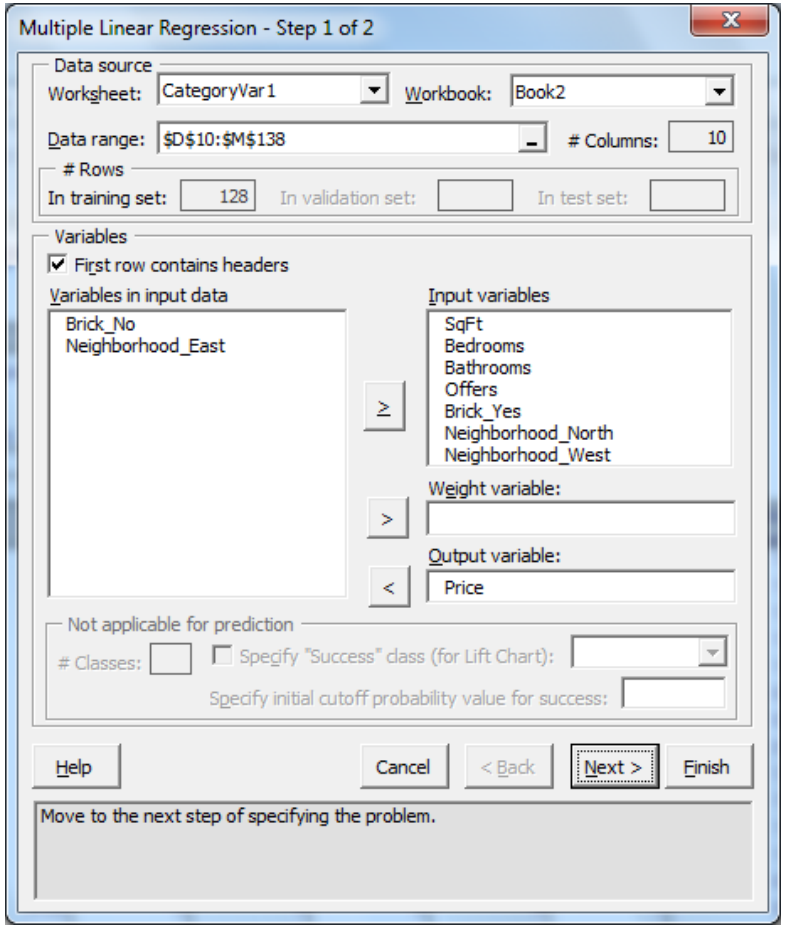
Data Preparation

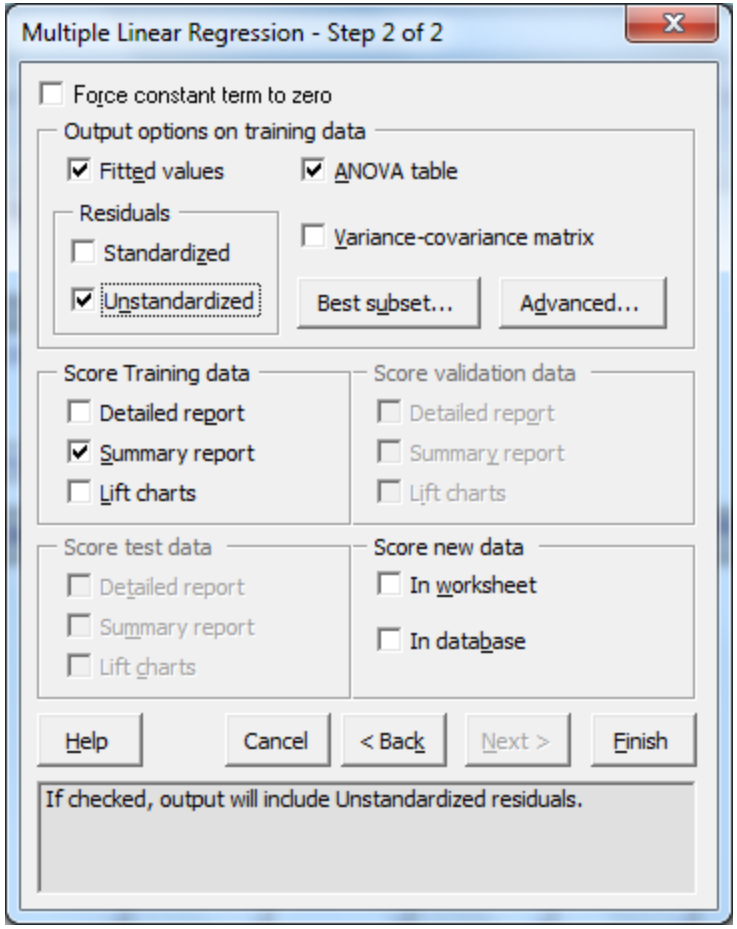
What types of variables are *Brick* & *Neighborhood*?

How to include them in a regression model?

XLMiner: *Data Utilities> Transform Categorical Data> Create Dummies*

(See sheet CategoryVar1)





ESTIMATE MODEL

Estimated Model (XLMiner, sheet MLR_output1)

Business Analytics Using Data Mining ISB

The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	598.9199219	9552.197266	0.95010996	2.17745E+12
SqFt	52.99374008	5.73424006	0	28036362240
Bedrooms	4246.793945	1597.910767	0.00893895	7992064000
Bathrooms	7883.27832	2117.0354	0.00030041	4272561664
Offers	-8267.48828	1084.776733	0	23712010240
Brick_Yes	17297.34961	1981.616333	0	7782121472
Neighborhood_North	1560.579224	2396.765381	0.51621467	97610600
Neighborhood_West	22241.61719	2531.758301	0	7746974720

Residual df	120
Multiple R-squared	0.868621033
Std. Dev. estimate	10018.94434
Residual SS	12045508608

ANOVA

Source	df	SS	MS	F-statistic	p-value
Regression	7	79639704936	11377100705	113.3411738	8.24658E-50
Error	120	12045508608	100379238.4		
Total	127	91685213544			

$$Y = \alpha + \beta_1 X_1 + \ldots \beta_k X_k + \varepsilon$$

RESIDUAL ANALYSIS

$$e_i = y_i - \hat{y}_i$$

CHECK MODEL ASSUMPTIONS

Linear Regression Assumptions

Model: $Y = \alpha + \beta_1 X_1 + \ldots \beta_k X_k + \varepsilon$

Error (ε) is

- **Normally** distributed with mean **0** and **constant variability**
- **Independent** across records

If any assumption is violated, interpretations based on the estimated model might be incorrect

Computing a residual (model-dependent!)

For record *i*: $e_i = y_i - \hat{y}_i$

Using our model:

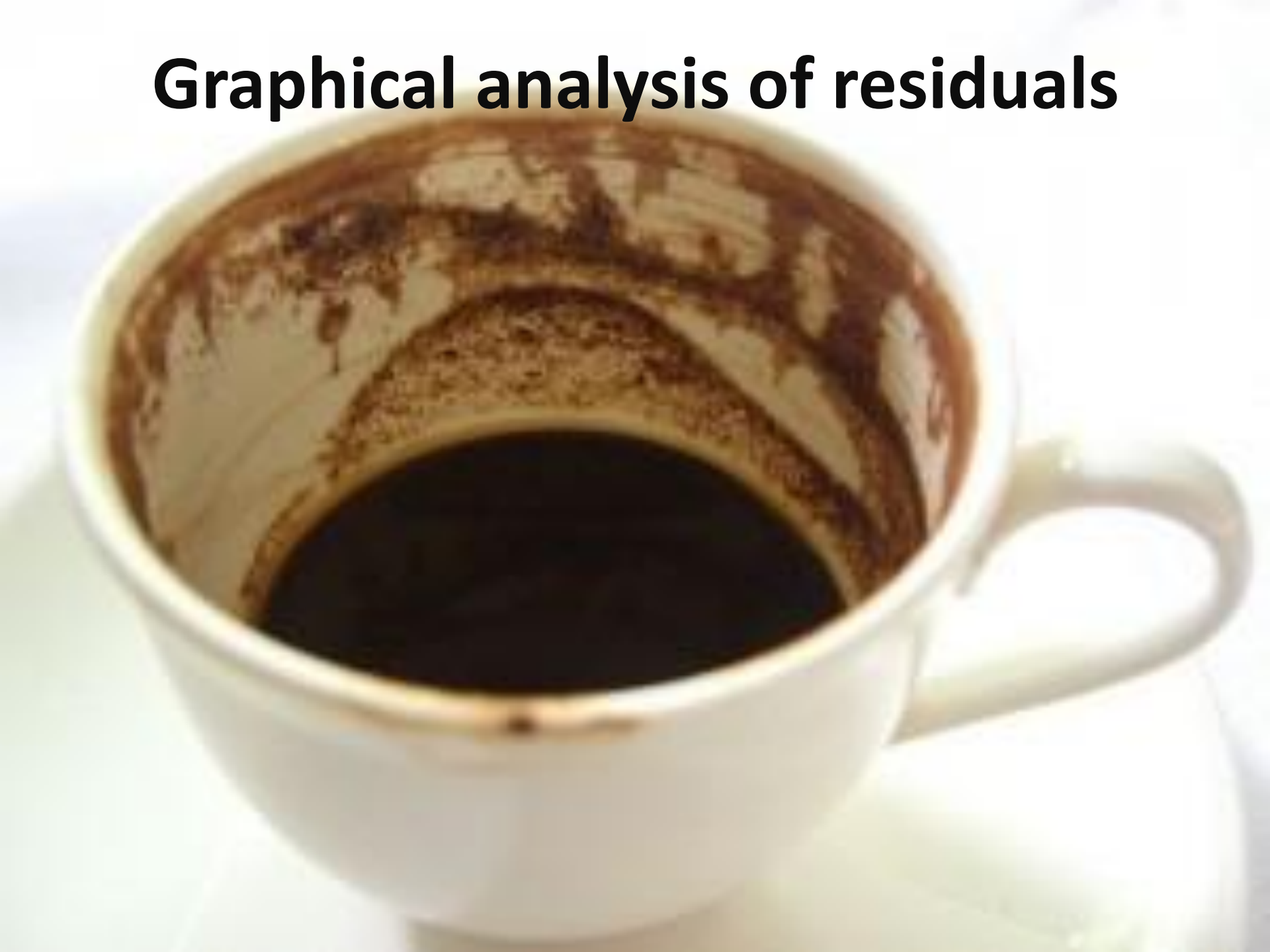
Y_1 = 114,300

$\hat{Y}_1$ = 103,182.88

e_1 = 11,117.12

XLMiner: check "*unstandardized residuals*"

Graphical analysis of residuals

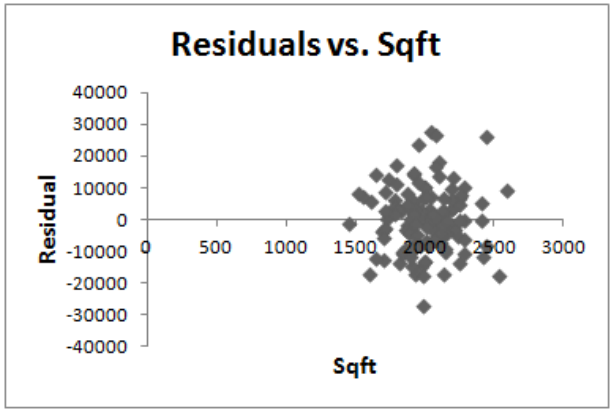


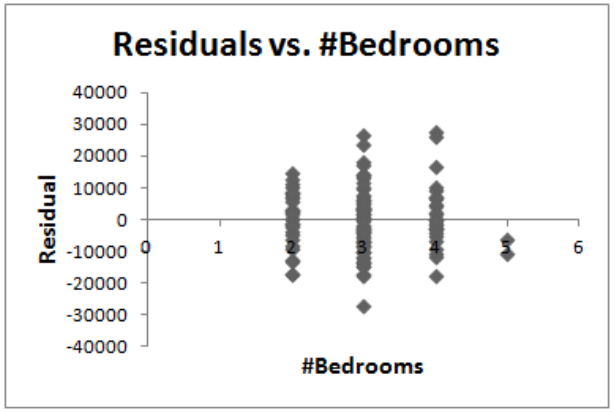
Residuals vs. each X

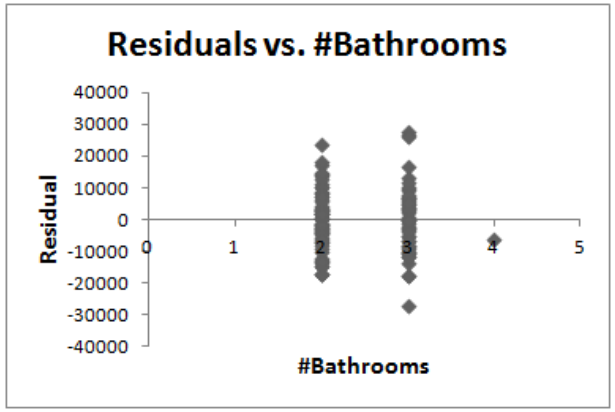
is variability constant regardless of the value of X?

Should see a random cloud

If not, transform X (for “fan shape” try logarithm)





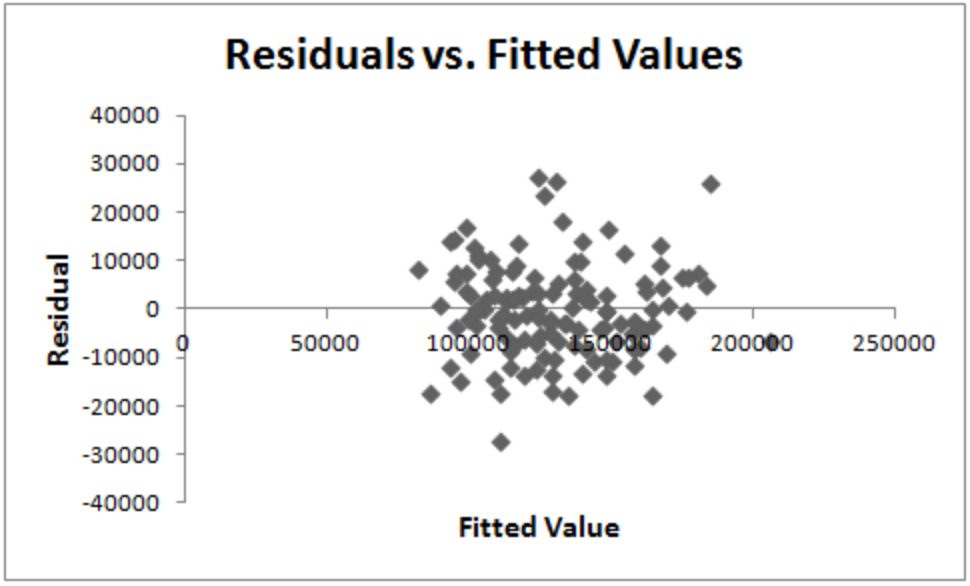


Residuals vs. fitted values

is linear assumption valid?

Should see a random cloud

If not, transform Y

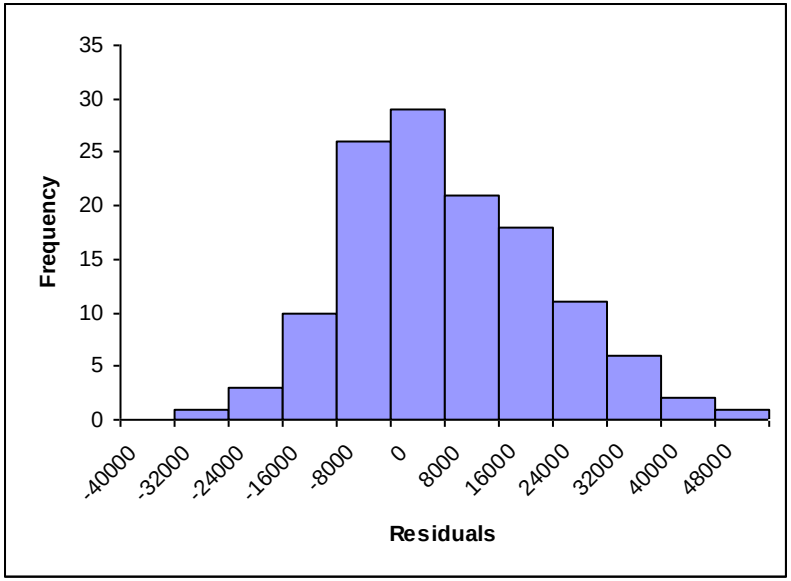


Histogram of residuals

Approx Normal distribution around 0?

Look at spread to each side

If not, try transforming Y (logarithm, inverse, etc.)



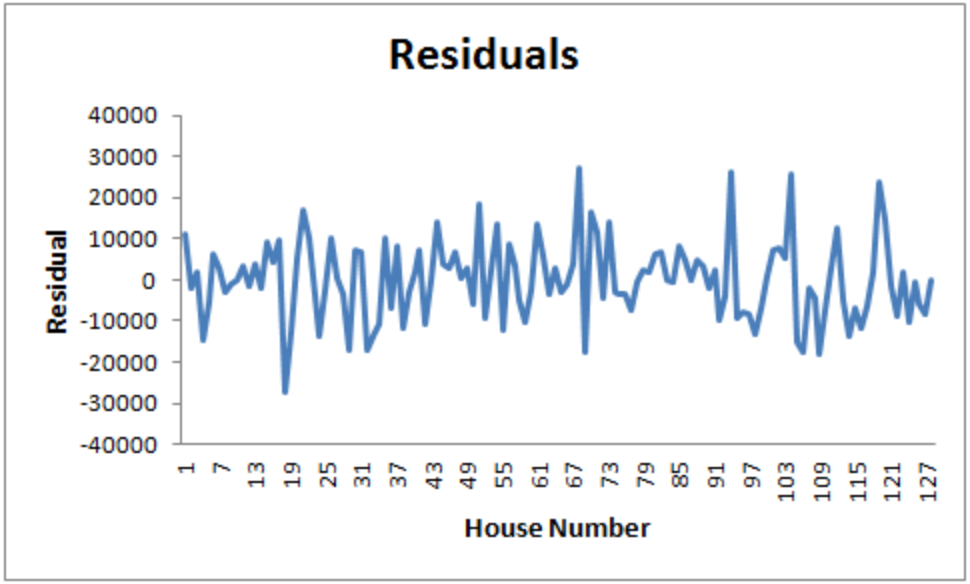
XLMiner: Charts

Time plot of residuals

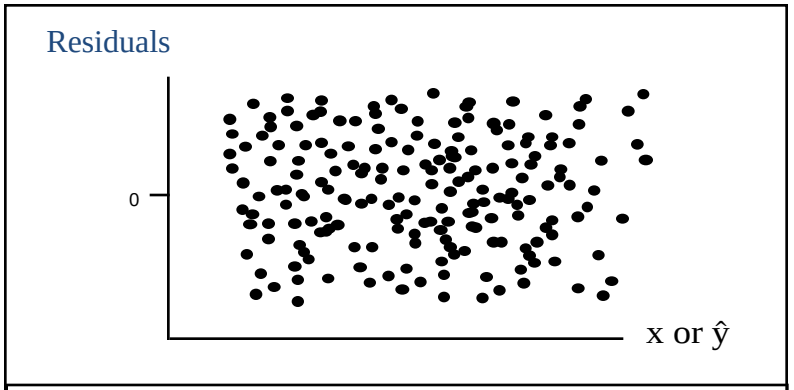
Are points independent?

Expect no pattern

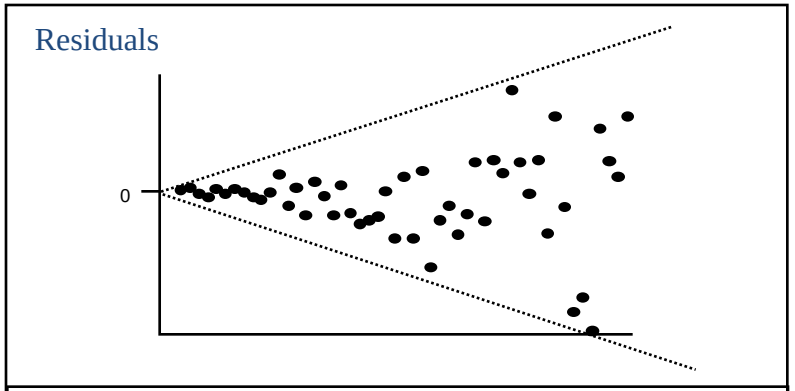
If pattern exists, need a time-series model



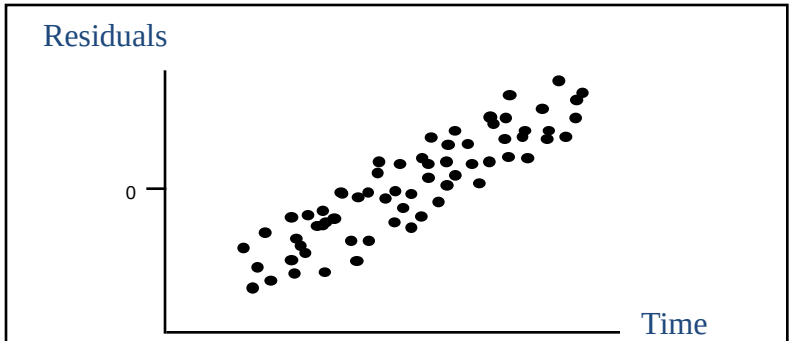
Common Patterns



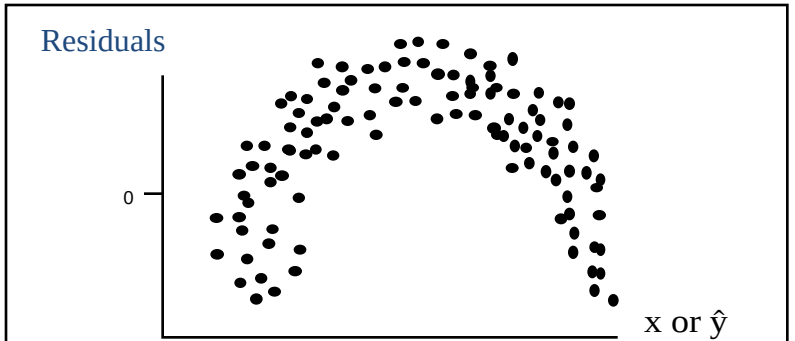
Homoscedasticity: Residuals appear completely random. No indication of model inadequacy.



Heteroscedasticity: Variance of residuals changes when x changes.



Residuals exhibit a linear trend with time.



Curved pattern in residuals resulting from underlying nonlinear relationship.

Source: Aczel, A. (1998), *Complete Business Statistics*, McGraw-Hill/Irwin, Mass., 4th ed. (CD-ROM)

What have we seen thus far?

1. **Choose** X variables
2. **View** scatter plots (careful!)
3. **Fit** a simple/multiple regression model to data. Get estimated model
4. Check **validity** of model assumptions
5. Use estimated model to **test/infer** relationship in the population [coming up next]