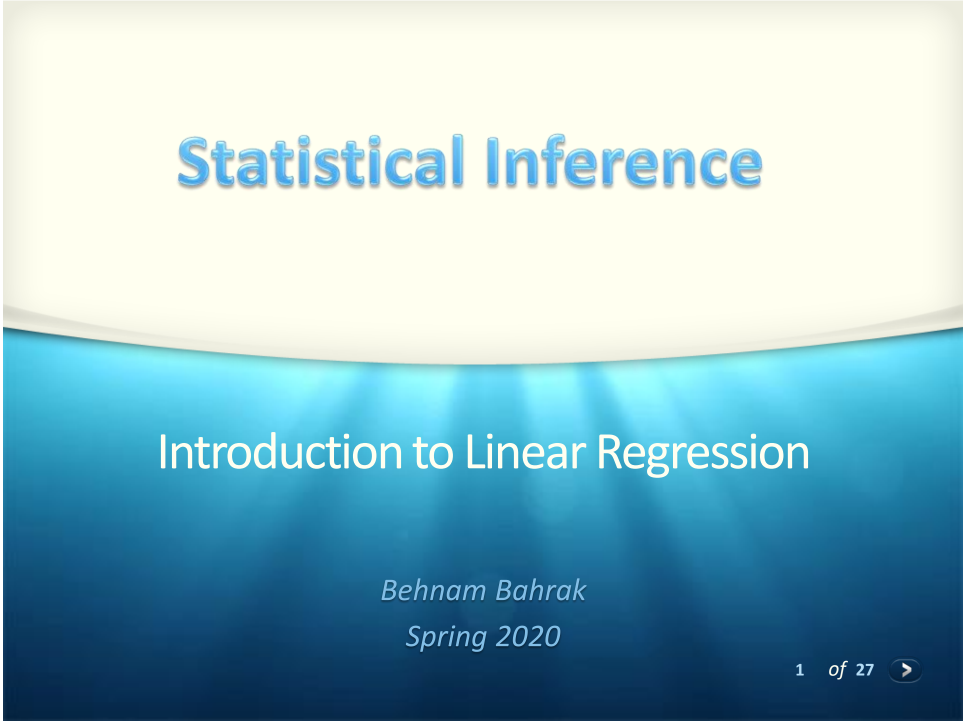
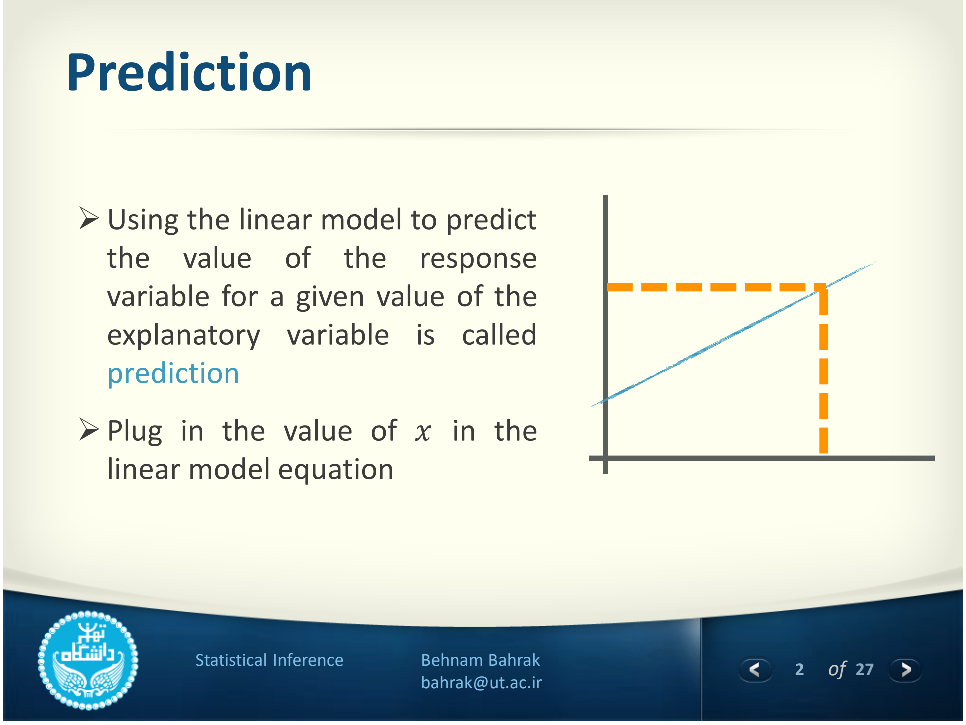


Statistical Inference

Introduction to Linear Regression

Behnam Bahrak

Spring 2020



Prediction

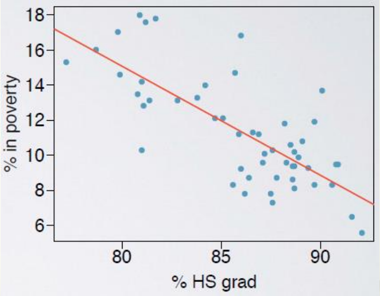
- Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called prediction
- Plug in the value of x in the linear model equation

Example

- According to the following linear model, what is the predicted % living in poverty in states where the HS graduation rate is 82%? $\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$

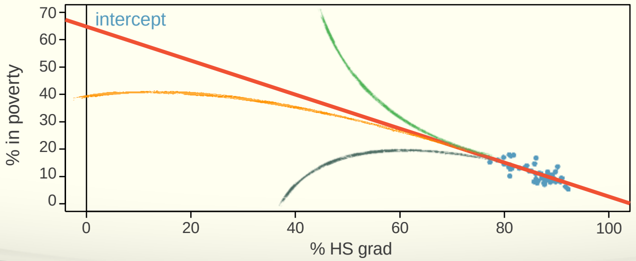
$$\widehat{\%\text{ in poverty}} = 64.68 - 0.62 \times 82$$

$$= 13.84\ \%$$



Extrapolation

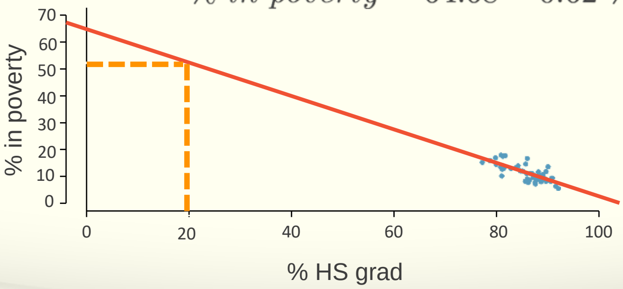
- Applying a model estimate to values outside of the realm of the original data is called extrapolation
- Sometimes the intercept might be an extrapolation

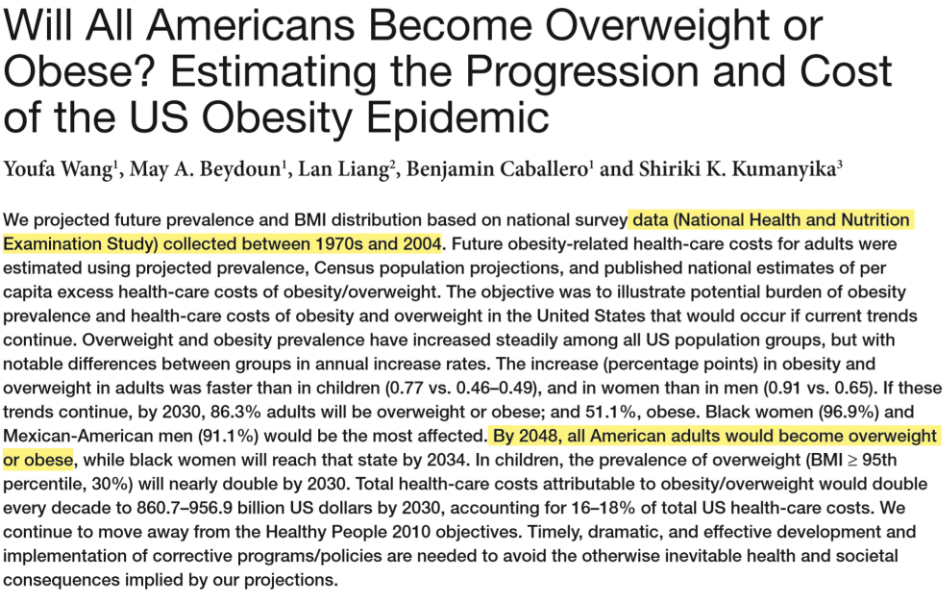


Example

- According to the following linear model, what is the predicted % living in poverty in states where the HS graduation rate is 20%?

$$\widehat{\% \ in \ poverty} = 64.68 - 0.62 \ \% \ HS \ grad$$

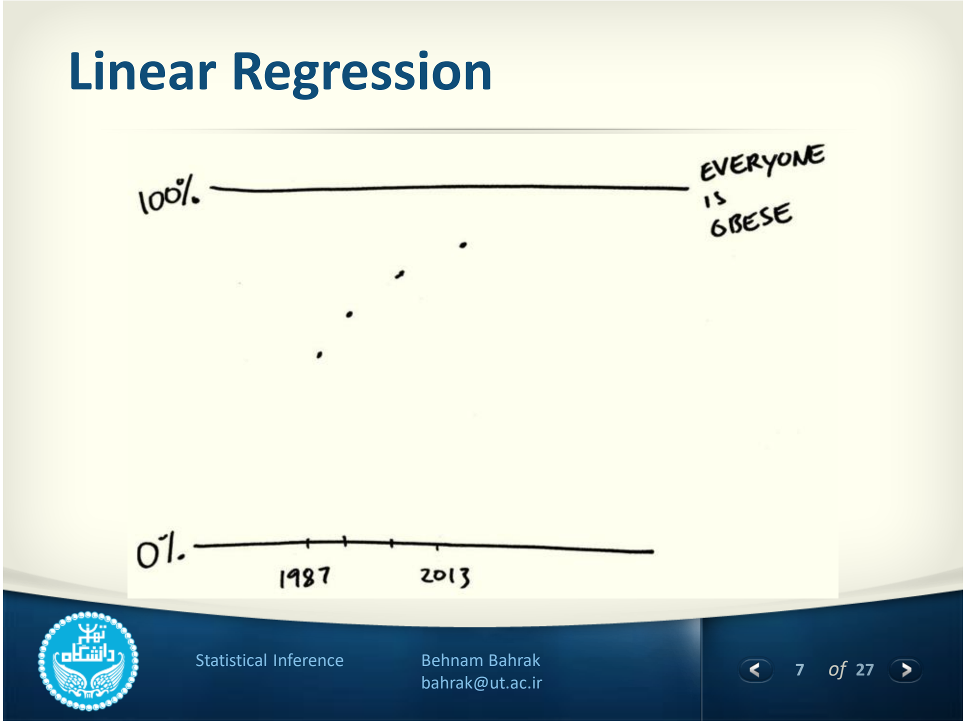




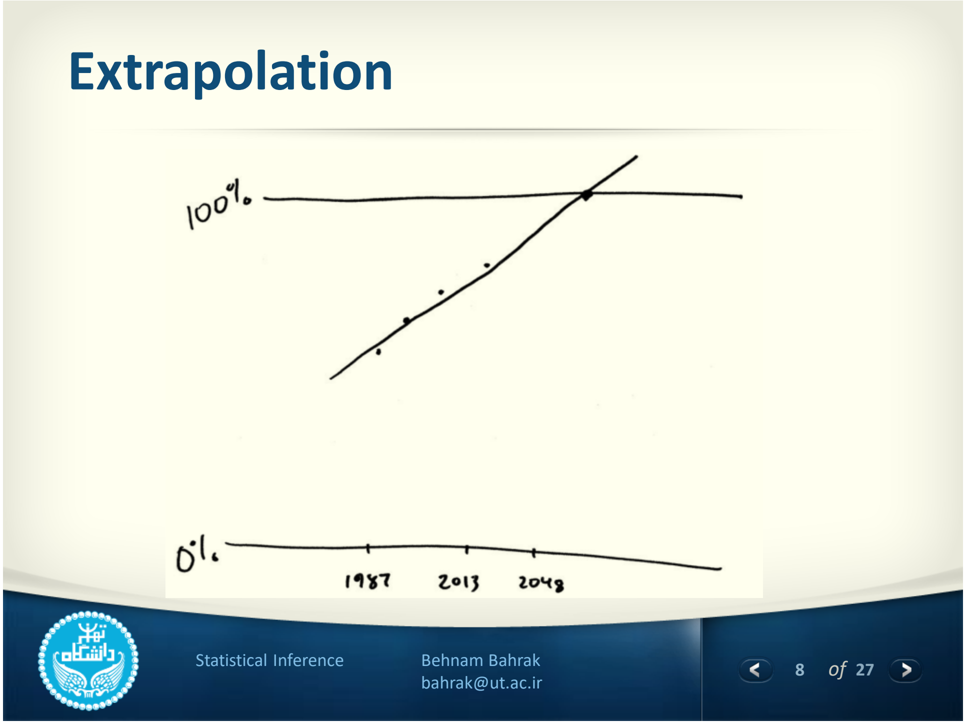
Will All Americans Become Overweight or Obese? Estimating the Progression and Cost of the US Obesity Epidemic

Youfa Wang[1], May A. Beydoun[1], Lan Liang[2], Benjamin Caballero[1] and Shiriki K. Kumanyika[3]

We projected future prevalence and BMI distribution based on national survey data (National Health and Nutrition Examination Study) collected between 1970s and 2004. Future obesity-related health-care costs for adults were estimated using projected prevalence, Census population projections, and published national estimates of per capita excess health-care costs of obesity/overweight. The objective was to illustrate potential burden of obesity prevalence and health-care costs of obesity and overweight in the United States that would occur if current trends continue. Overweight and obesity prevalence have increased steadily among all US population groups, but with notable differences between groups in annual increase rates. The increase (percentage points) in obesity and overweight in adults was faster than in children (0.77 vs. 0.46–0.49), and in women than in men (0.91 vs. 0.65). If these trends continue, by 2030, 86.3% adults will be overweight or obese; and 51.1%, obese. Black women (96.9%) and Mexican-American men (91.1%) would be the most affected. By 2048, all American adults would become overweight or obese, while black women will reach that state by 2034. In children, the prevalence of overweight (BMI ≥ 95th percentile, 30%) will nearly double by 2030. Total health-care costs attributable to obesity/overweight would double every decade to 860.7–956.9 billion US dollars by 2030, accounting for 16–18% of total US health-care costs. We continue to move away from the Healthy People 2010 objectives. Timely, dramatic, and effective development and implementation of corrective programs/policies are needed to avoid the otherwise inevitable health and societal consequences implied by our projections.

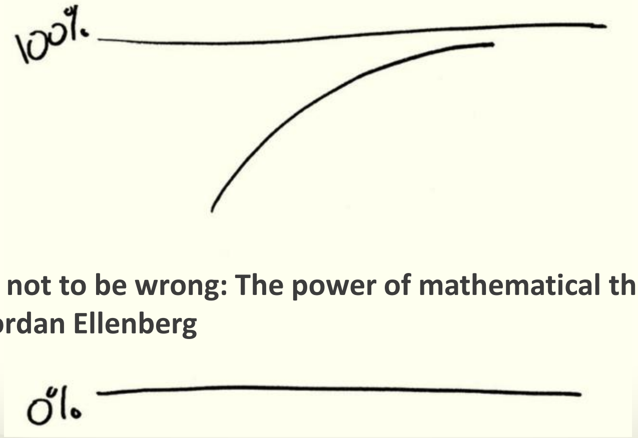


Linear Regression



Extrapolation

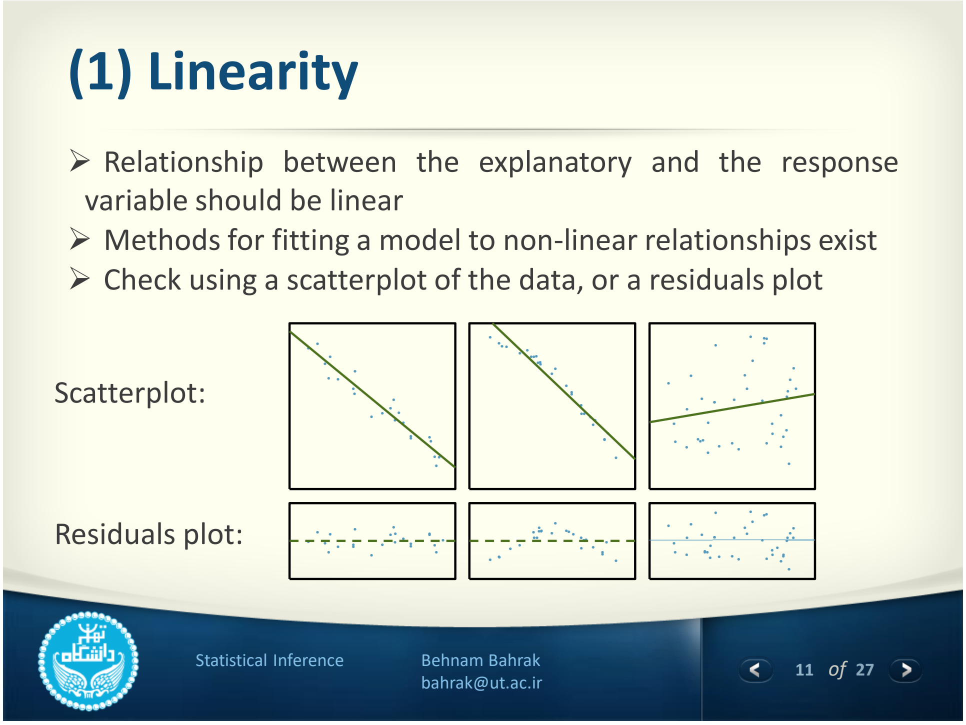
More Realistic Curve



How not to be wrong: The power of mathematical thinking
By **Jordan Ellenberg**

Conditions for linear regression

- **Linearity**
  - relationship between the explanatory and the response variable should be linear
- **Nearly normal residuals**
  - residuals should be nearly normally distributed
- **Constant variability**
  - variability of points around the least squares line should be roughly constant



(1) Linearity

- Relationship between the explanatory and the response variable should be linear
- Methods for fitting a model to non-linear relationships exist
- Check using a scatterplot of the data, or a residuals plot

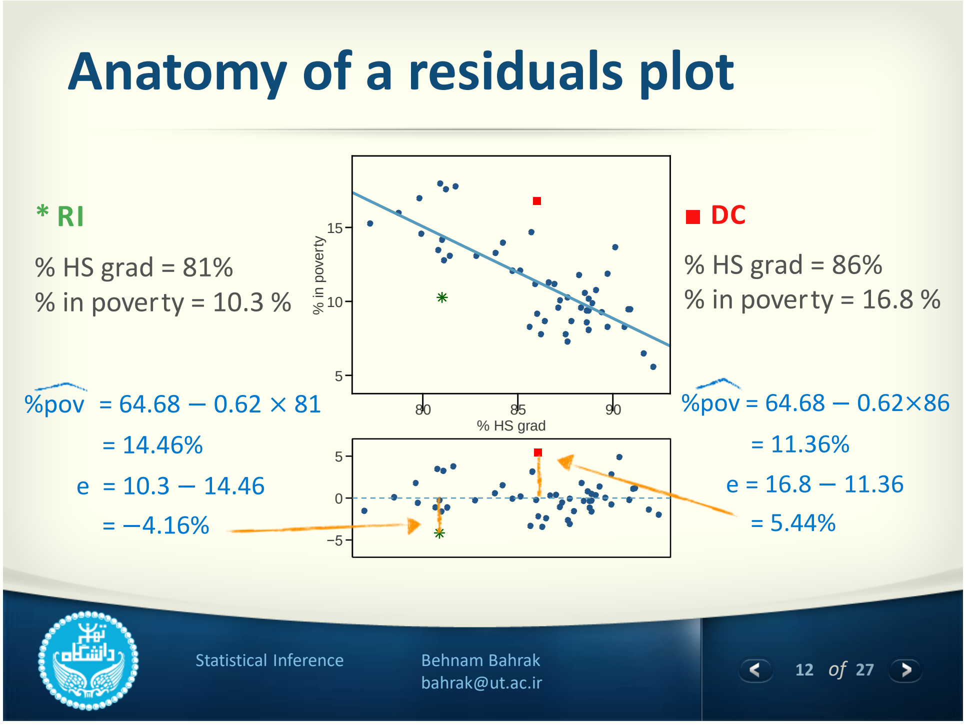
Scatterplot:

Residuals plot:

Statistical Inference

Behnam Bahrak
bahrak@ut.ac.ir

11 of 27



Anatomy of a residuals plot

* RI

% HS grad = 81%

% in poverty = 10.3 %

$\widehat{\%pov} = 64.68 - 0.62 \times 81$

$= 14.46\%$

$e = 10.3 - 14.46$

$= -4.16\%$

■ DC

% HS grad = 86%

% in poverty = 16.8 %

$\widehat{\%pov} = 64.68 - 0.62 \times 86$

$= 11.36\%$

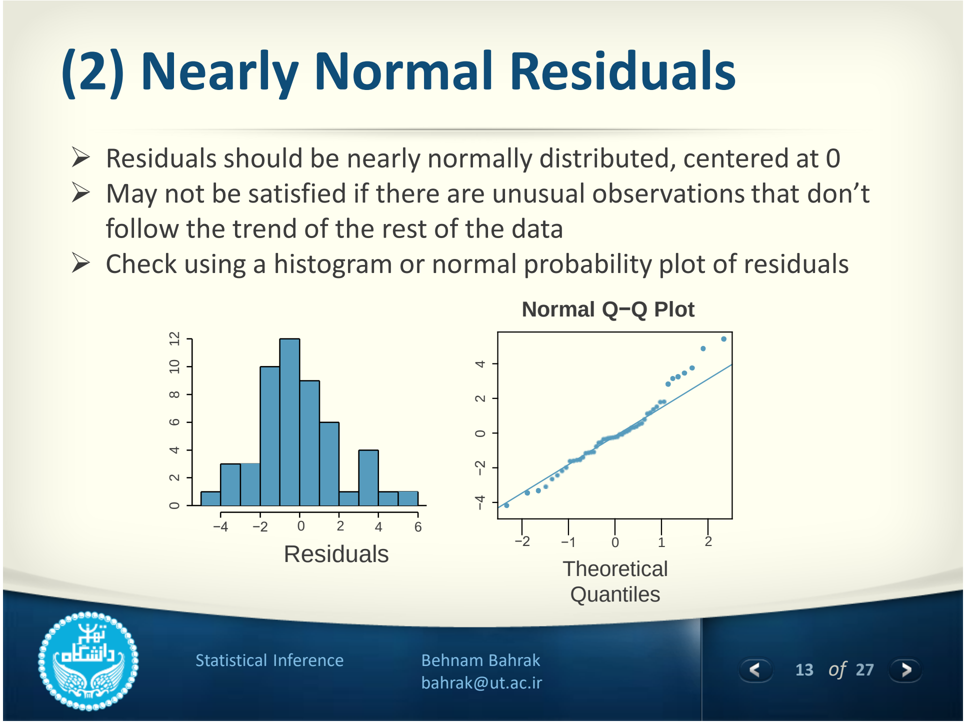
$e = 16.8 - 11.36$

$= 5.44\%$

Statistical Inference

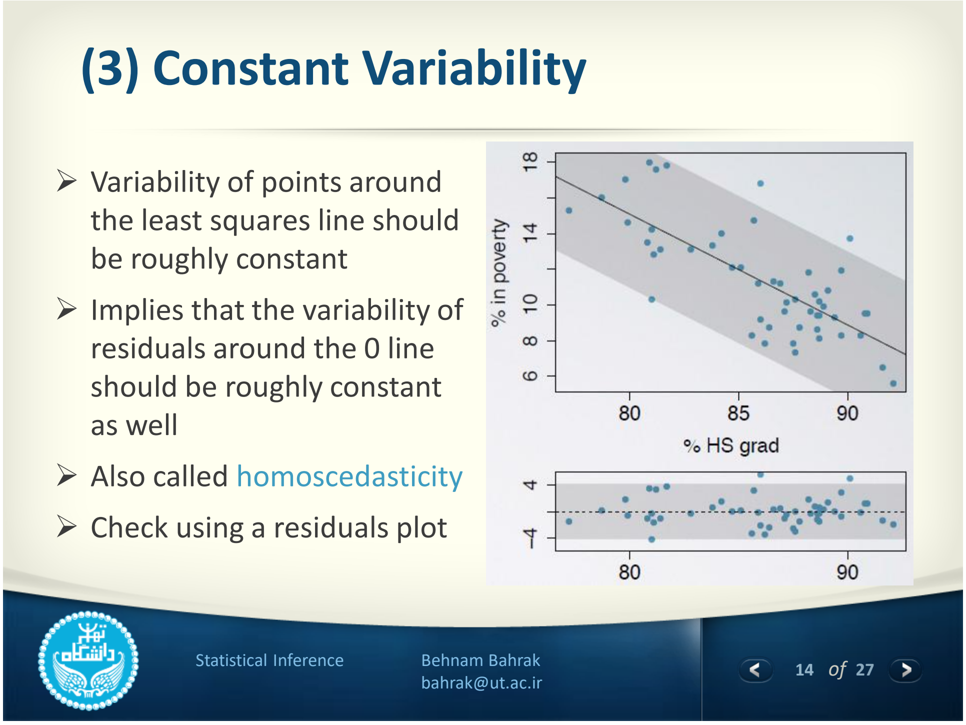
Behnam Bahrak
bahrak@ut.ac.ir

12 of 27



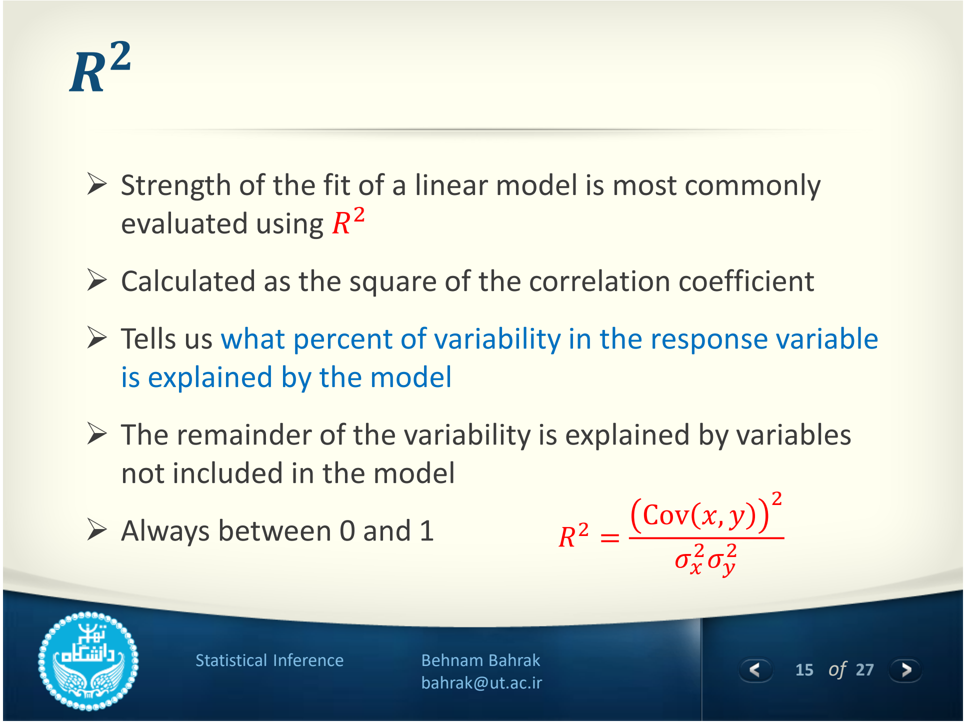
(2) Nearly Normal Residuals

- Residuals should be nearly normally distributed, centered at 0
- May not be satisfied if there are unusual observations that don’t follow the trend of the rest of the data
- Check using a histogram or normal probability plot of residuals



(3) Constant Variability

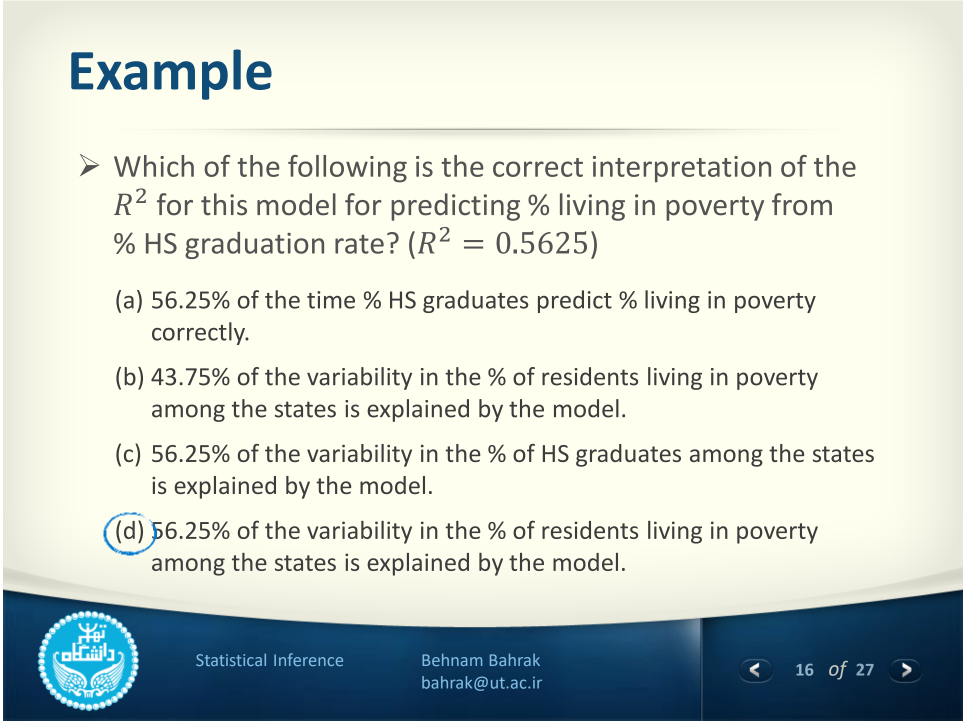
- Variability of points around the least squares line should be roughly constant
- Implies that the variability of residuals around the 0 line should be roughly constant as well
- Also called homoscedasticity
- Check using a residuals plot



R^2

- Strength of the fit of a linear model is most commonly evaluated using R^2
- Calculated as the square of the correlation coefficient
- Tells us what percent of variability in the response variable is explained by the model
- The remainder of the variability is explained by variables not included in the model
- Always between 0 and 1

$$R^2 = \frac{\left(\mathrm{Cov}(x,y)\right)^2}{\sigma_x^2 \sigma_y^2}$$



Example

- Which of the following is the correct interpretation of the R^2 for this model for predicting % living in poverty from % HS graduation rate? ($R^2 = 0.5625$)

(a) 56.25% of the time % HS graduates predict % living in poverty correctly.

(b) 43.75% of the variability in the % of residents living in poverty among the states is explained by the model.

(c) 56.25% of the variability in the % of HS graduates among the states is explained by the model.

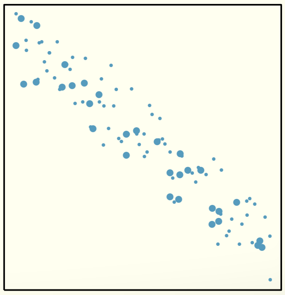
[✓ (d)] 56.25% of the variability in the % of residents living in poverty among the states is explained by the model.

Example

- The R^2 for the relationship displayed in the scatterplot is 92.16%. What is the correlation coefficient?

$$|R| = \sqrt{R^2} = \sqrt{0.9216} = 0.96$$

$$\Rightarrow \quad R = -0.96$$



Regression with categorical explanatory variable

Poverty vs. Region

explanatory variable: region
0: east
1: west

$$\widehat{poverty} = 11.17 + 0.38\ region : west$$

for eastern states
plug in **0** for x

for western states
plug in **1** for x

$$\widehat{poverty} = \underbrace{11.17} + 0.38 \times 0 = 11.17 \qquad \widehat{poverty} = 11.17 + 0.38 \times 1 = 11.55$$

reference level:

In regression models, with explanatory categorical variables, we always code one of the levels of that categorical variable to be what we call the reference level. This is the level that we plug in zero for.

Slope and Intercept

$$\widehat{poverty} = 11.17 + 0.38\ region : west$$

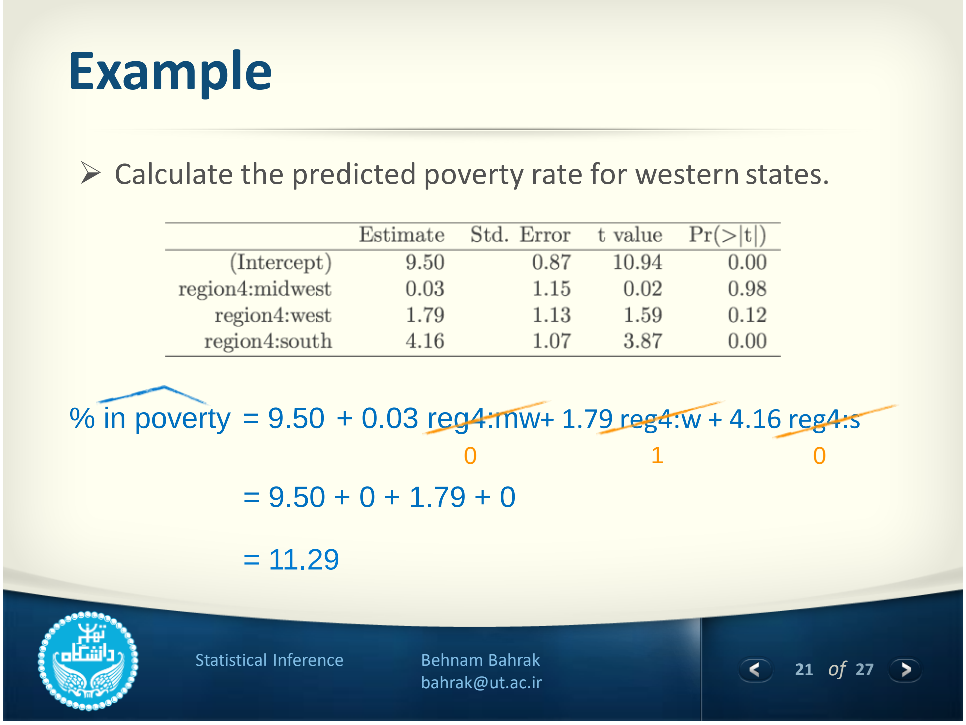
- Intercept: The model predicts an 11.17% average poverty percentage in eastern states.
- This is the value we get if we plug in **0** for the explanatory variable
- Slope: The model predicts that the average poverty percentage in western states is 0.38% higher than in the eastern states.

Example

- Next, we use a new region variable (`region4`) with four levels: *northeast*, *midwest*, *west*, *south*. Write the linear regression model based on the regression output below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4:midwest	0.03	1.15	0.02	0.98
region4:west	1.79	1.13	1.59	0.12
region4:south	4.16	1.07	3.87	0.00

$\widehat{\text{\% in poverty}}$ = 9.50 + 0.03 reg4:mw+ 1.79 reg4:w + 4.16 reg4:s



Example

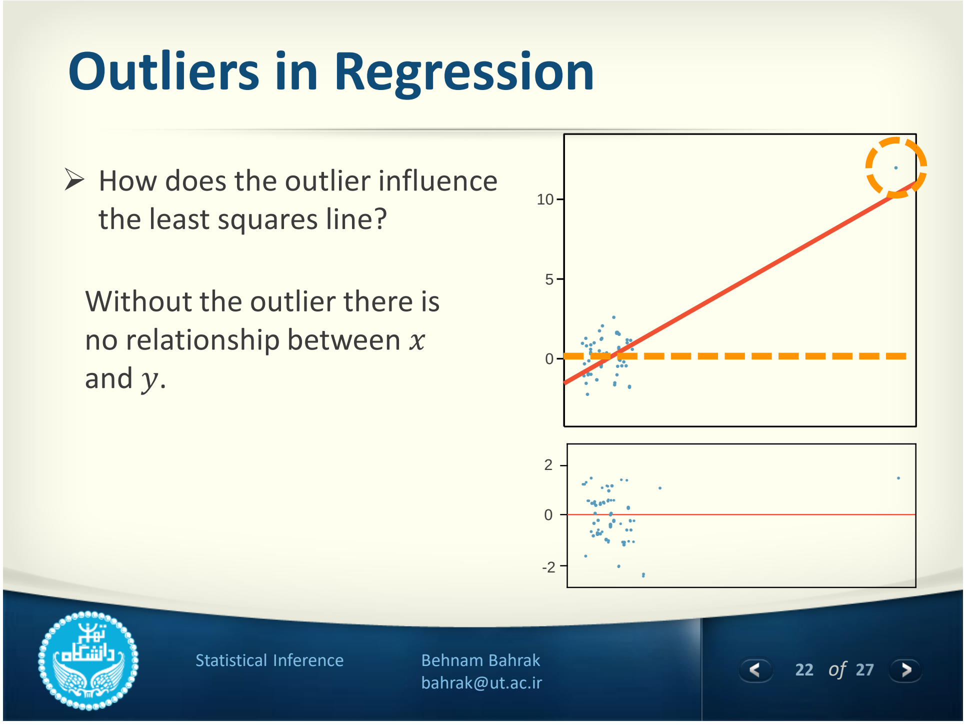
- Calculate the predicted poverty rate for western states.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.87	10.94	0.00
region4:midwest	0.03	1.15	0.02	0.98
region4:west	1.79	1.13	1.59	0.12
region4:south	4.16	1.07	3.87	0.00

$$\widehat{\%\text{ in poverty}} = 9.50 + 0.03\,\underset{0}{\cancel{\text{reg4:mw}}} + 1.79\,\underset{1}{\cancel{\text{reg4:w}}} + 4.16\,\underset{0}{\cancel{\text{reg4:s}}}$$

$$= 9.50 + 0 + 1.79 + 0$$

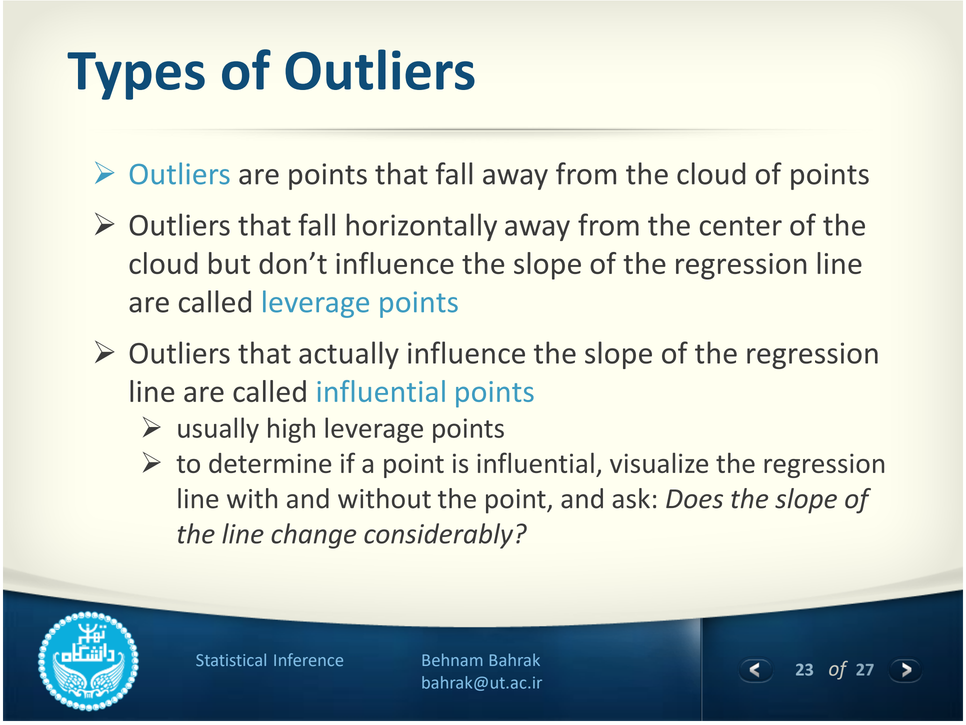
$$= 11.29$$



Outliers in Regression

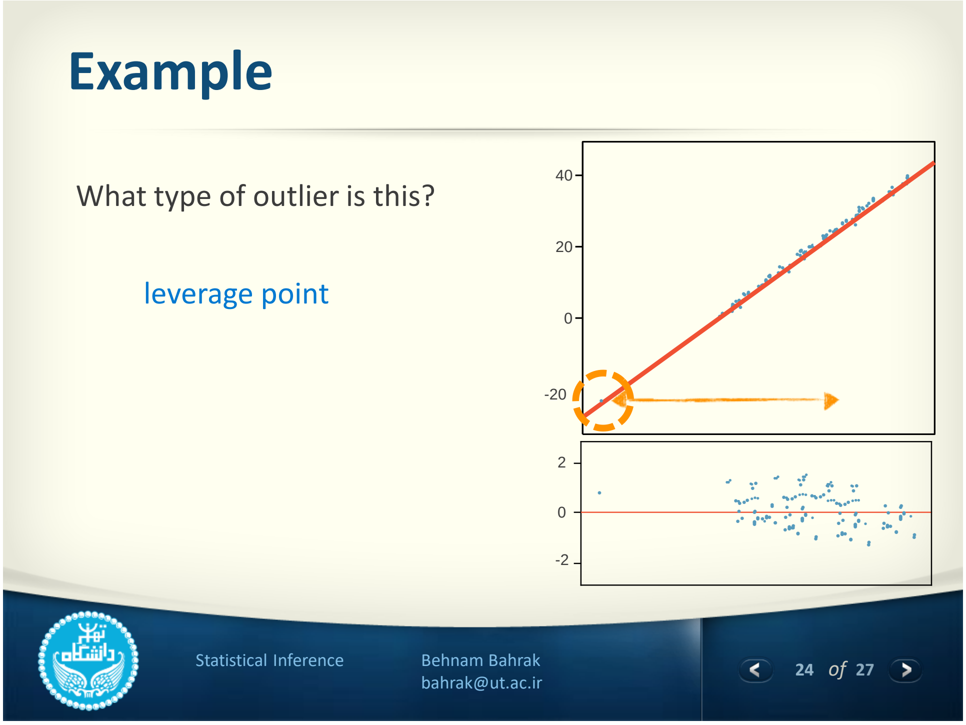
- How does the outlier influence the least squares line?

Without the outlier there is no relationship between x and y.



Types of Outliers

- Outliers are points that fall away from the cloud of points
- Outliers that fall horizontally away from the center of the cloud but don’t influence the slope of the regression line are called leverage points
- Outliers that actually influence the slope of the regression line are called influential points
  - usually high leverage points
  - to determine if a point is influential, visualize the regression line with and without the point, and ask: *Does the slope of the line change considerably?*



Example

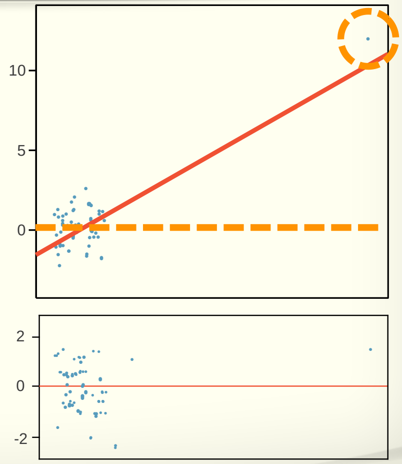
What type of outlier is this?

leverage point

Example

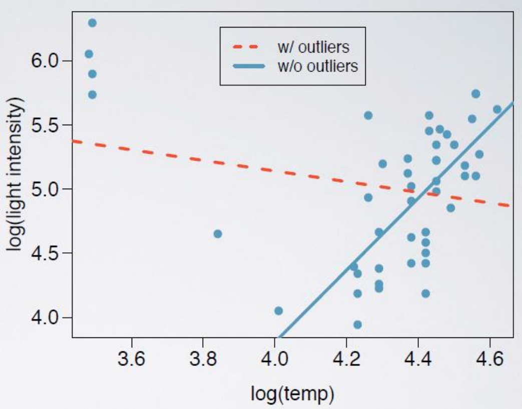
What type of outlier is this?

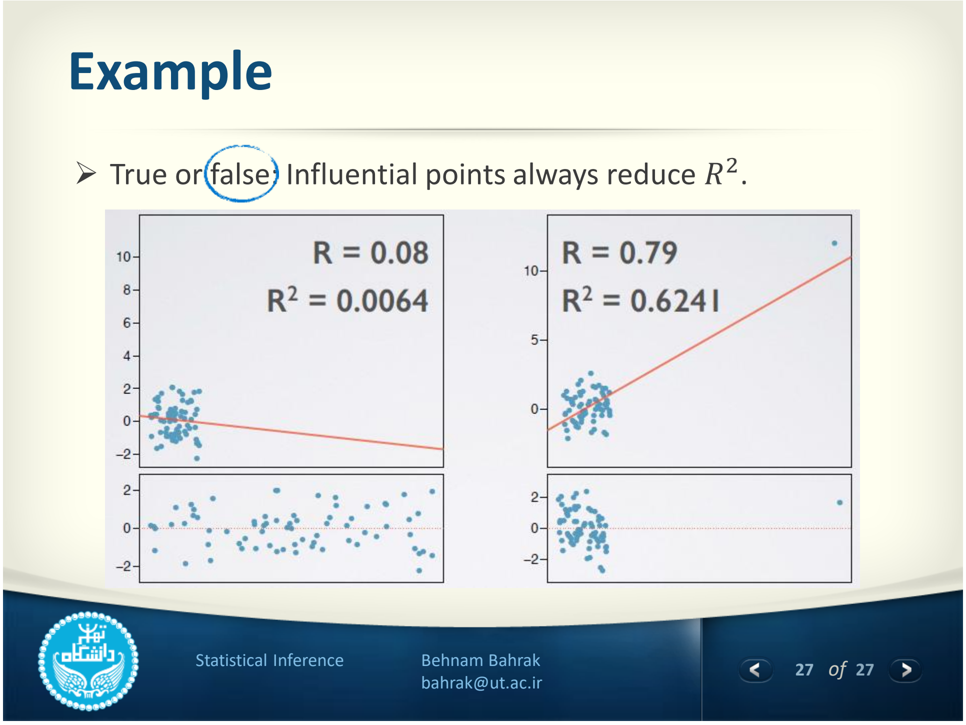
influential point



Influential Points

- Light intensity and surface temperature (logged) of 47 stars in the star cluster CYG OB1





Example

- True or [✓ false]: Influential points always reduce R^2.

Statistical Inference

Behnam Bahrak
bahrak@ut.ac.ir

27 of 27