MEDSTA 2: Regression models in medical research

29 April 2014

Øystein Ariansen Haaland, PHD

Department of Global Public Health and Primary Care, University of Bergen

Breaking assumptions

Multiple linear regression

- $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \epsilon_i$
- y_i is the observed value of subject i
- x_{ki} is the k'th observation of subject i
  - There are K observations per subject
  - E.g., age, sex, height, weight
- β's are regression coefficients
- ϵ_i is error

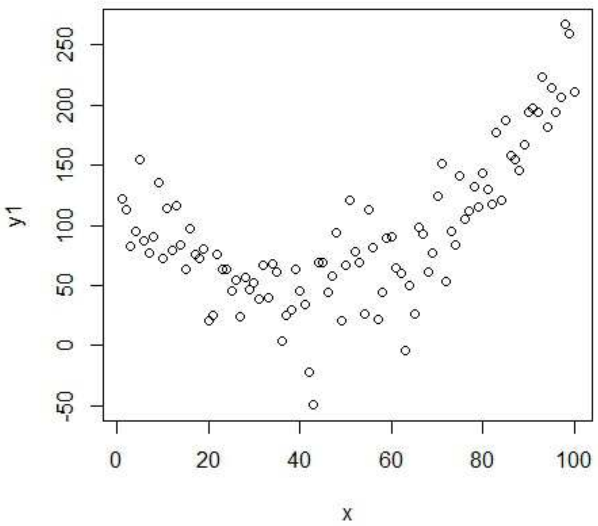
Breaking assumptions

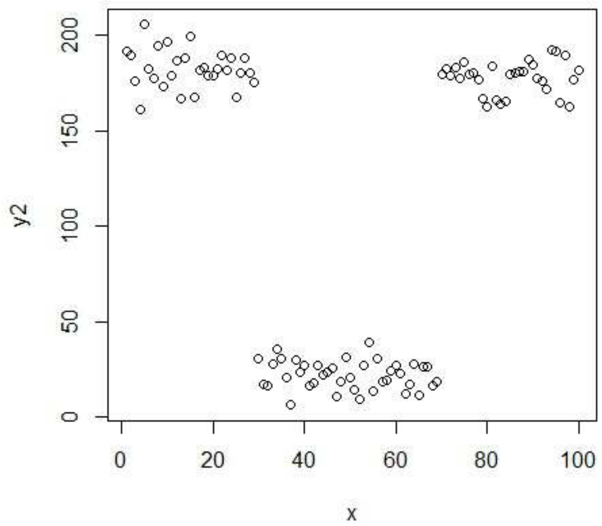
Multiple linear regression

- β's are unknown
- ϵ's are unknown
- Estimated line
  - $\hat{y}_i = b_0 + b_1 x_{1i} + \cdots + b_{Ki} x_K$

Breaking assumptions

What if relationship between y_i and x_{ki} is not linear?

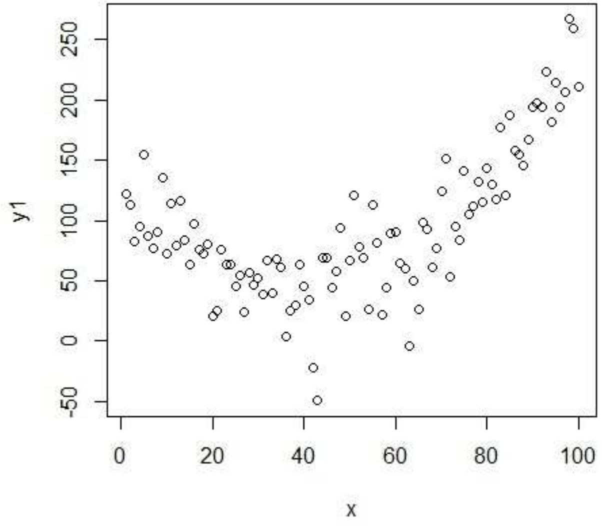




Breaking assumptions

What if relationship between y_i and x_{ki} is not linear?

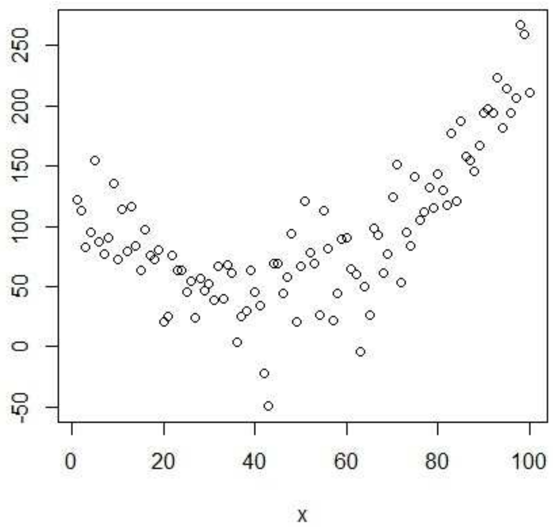
- Add quadratic term: $\hat{y}_i = b_0 + b_1 x + b_2 x^2$
  - x changes 1 unit, y changes approximately $b_1 + 2b_2 x$ units
  - Negative b_2 ⇒sad graph
  - Positive b_2 ⇒happy graph



Breaking assumptions

What if relationship between y_i and x_{ki} is not linear?

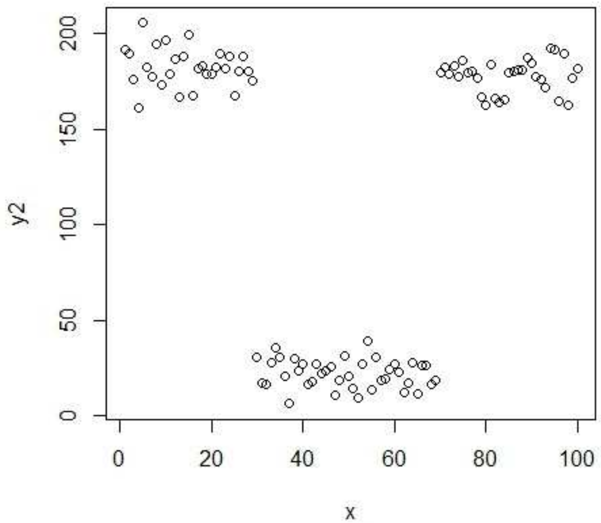
- Add quadratic term: $\hat{y}_i = b_0 + b_1 x + b_2 x^2$
- Stata:
  - gen x2=x^2
    - Creates new variable
  - regress y x x2
    - Normal regression



Breaking assumptions

What if relationship between y_i and x_{ki} is not linear?

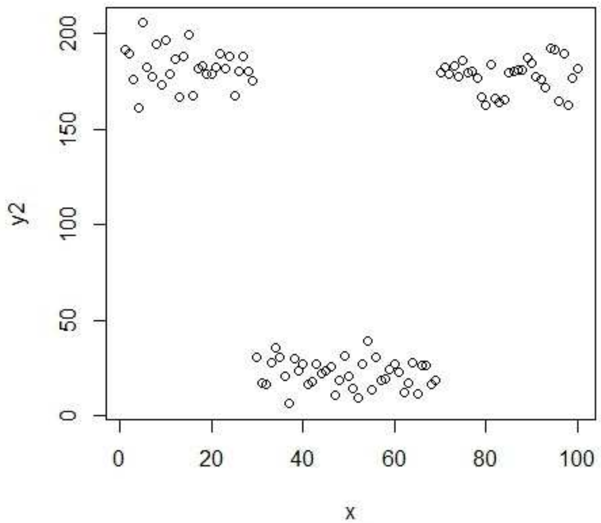
- Categorize x: $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2}$
    - x_{i0} is reference
    - b_1 is effect of x_{i1} relative to x_{i0}
    - b_2 is effect of x_{i2} relative to x_{i0}



Breaking assumptions

What if relationship between y_i and x_{ki} is not linear?

- Categorize x: $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2}$
- Stata:
  - gen x_cat=0
  - replace x_cat=1 if x>30
  - replace x_cat=2 if x>70
  - regress y i.x



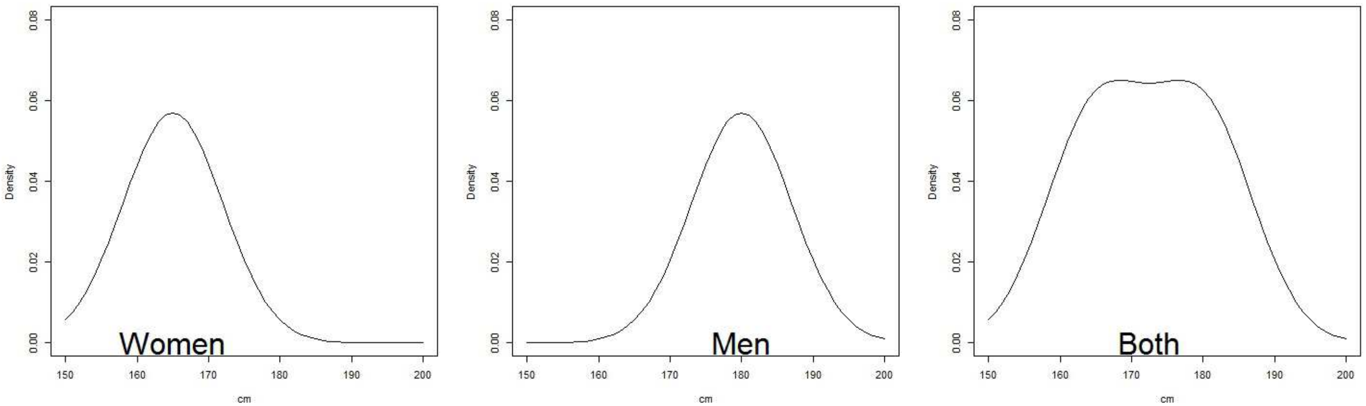
Breaking assumptions

What if y is not normal?

Breaking assumptions

What if y is not normal? NOT important!

- In a group of men and women, height is not normal
- Adjusting for height makes the error normal



Breaking assumptions

What if the error is not normal?

Breaking assumptions

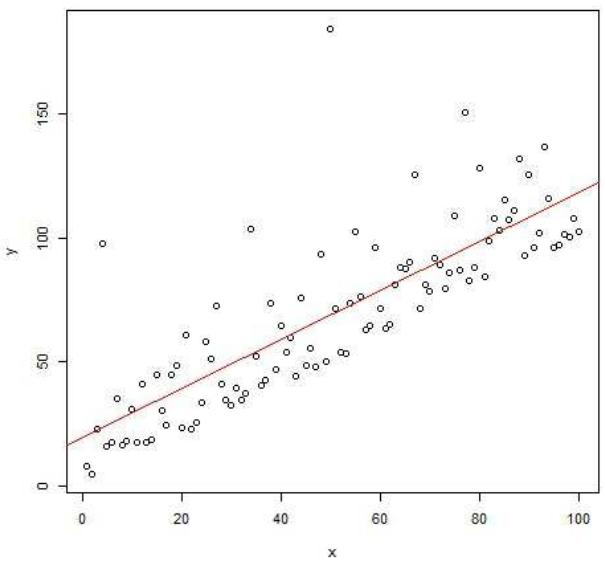
What if the error is not normal? ALSO not very important!

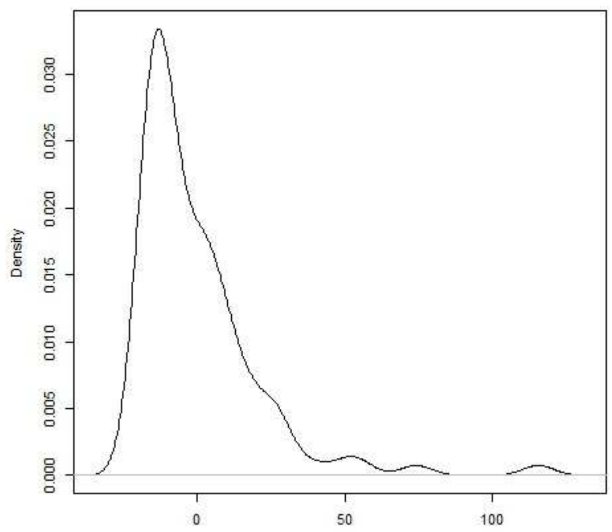
Coefficients will be normal if n is large because of Central limit theorem.

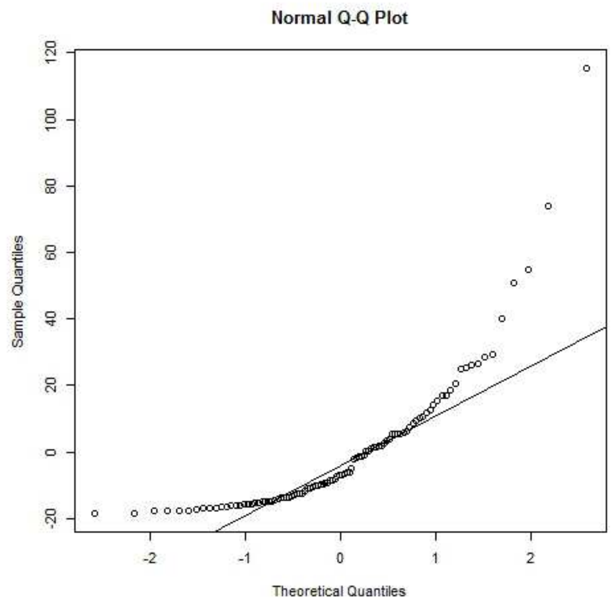
Breaking assumptions

What if the error is not normal? ALSO not very important!

- $\beta_0 = 20, \beta_1 = 1$



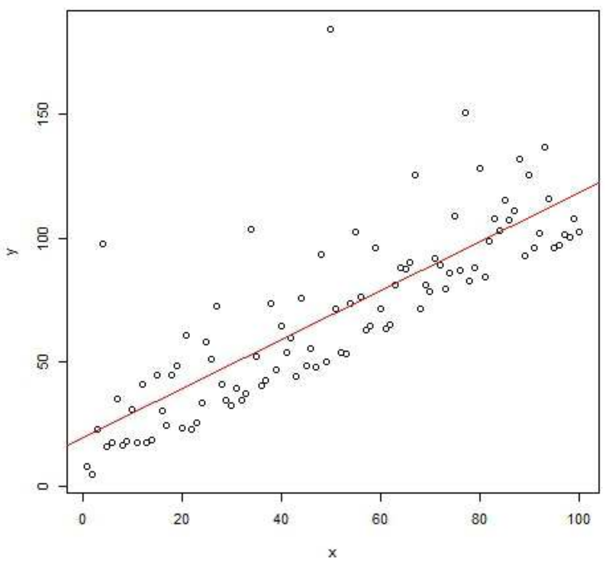


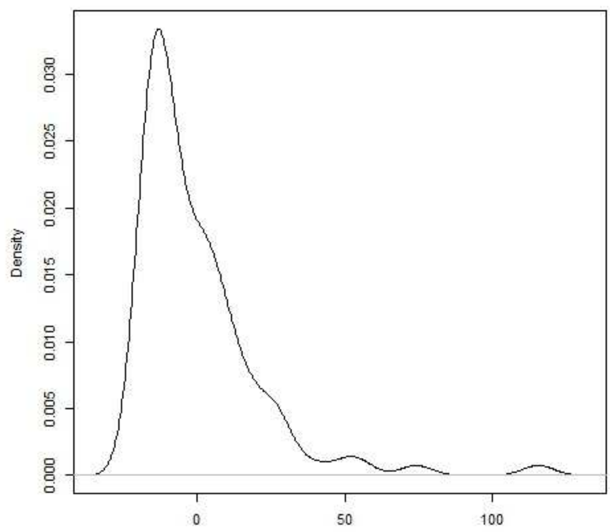


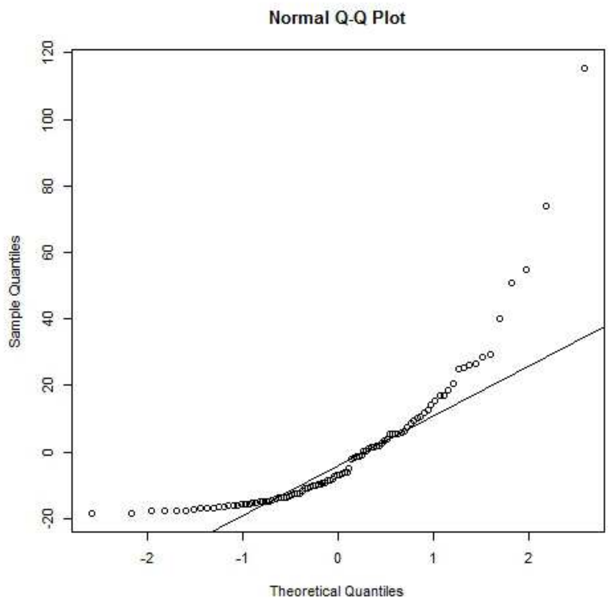
Breaking assumptions

What if the error is not normal? ALSO not very important!

- $b_0 = 19.5, b_1 = 0.99, \text{SE}(b_1) \approx 0.06$



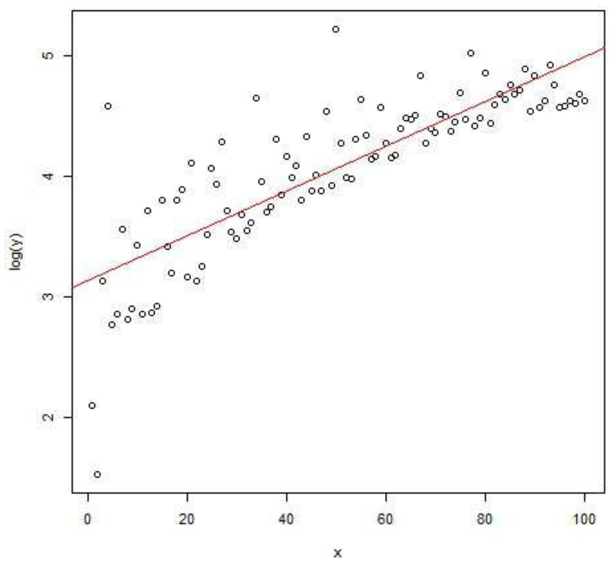


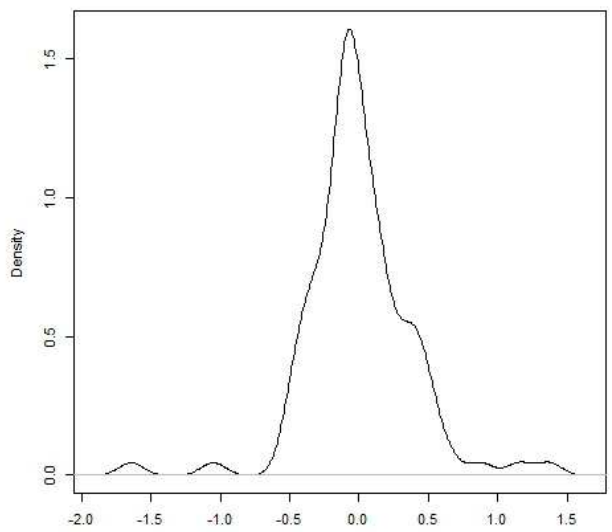


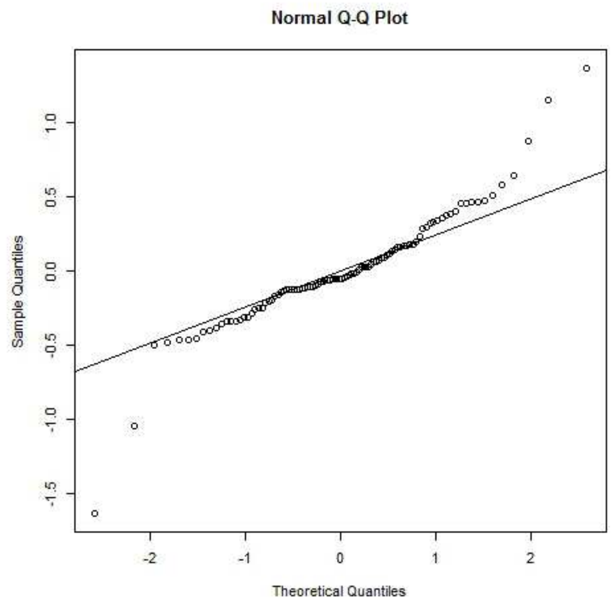
Breaking assumptions

What if the error is not normal?

- Log-transform y-variable
- Errors looks more symmetric



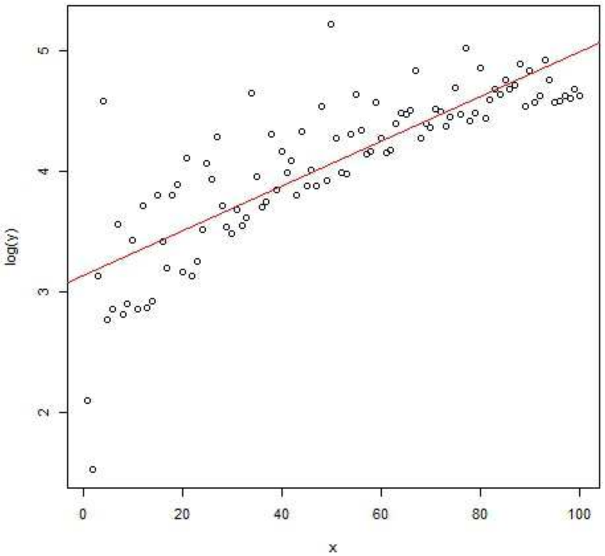


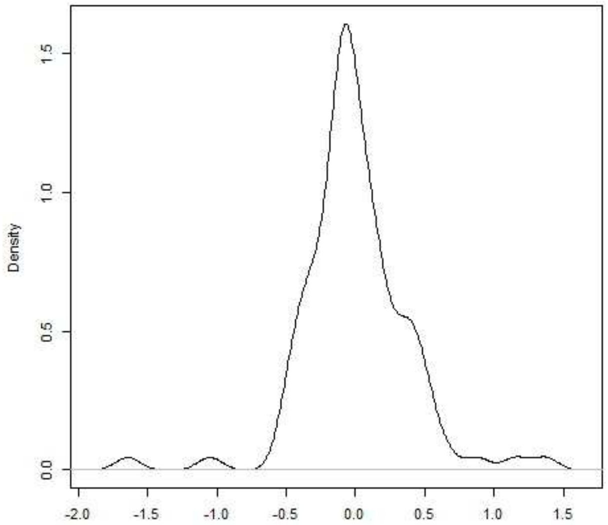


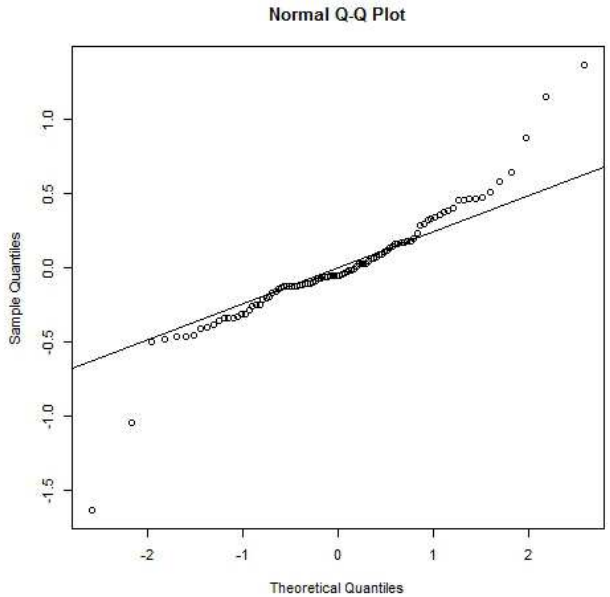
Breaking assumptions

What if the error is not normal?

- Log-transform y-variable
- Not straightforward to interpret coefficients



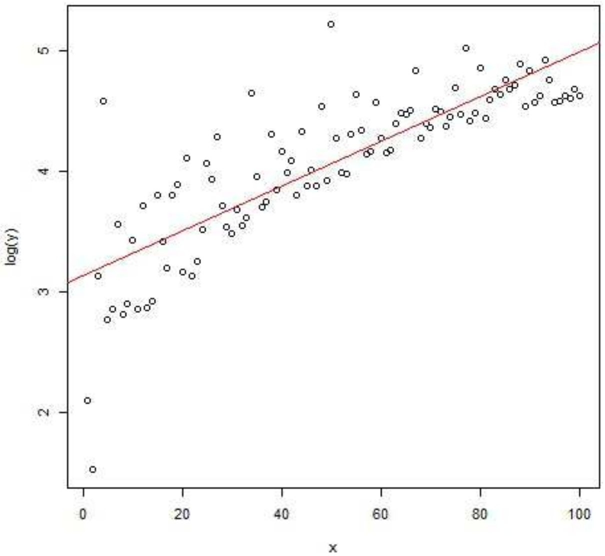


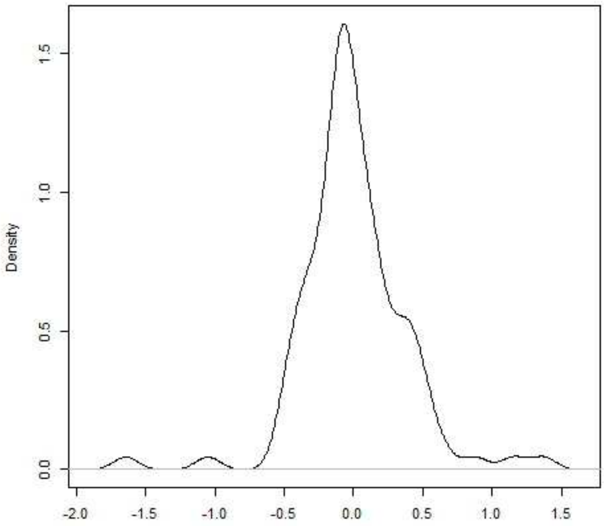


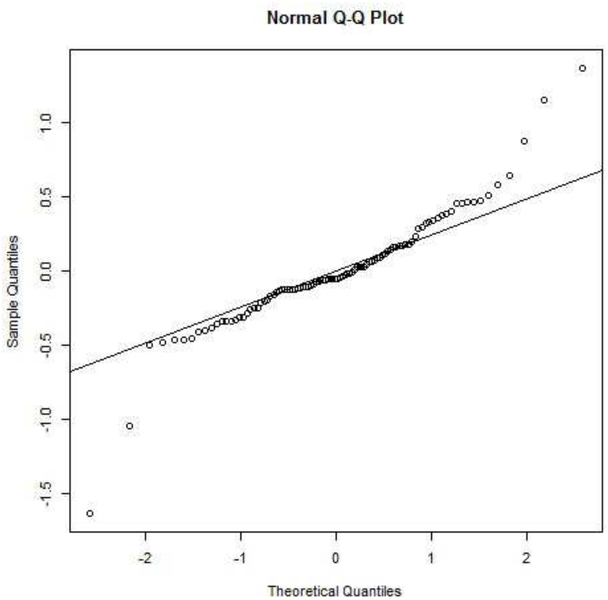
Breaking assumptions

What if the error is not normal?

- Log-transform y-variable
- β_1 is relative y-change with 1 unit change of x



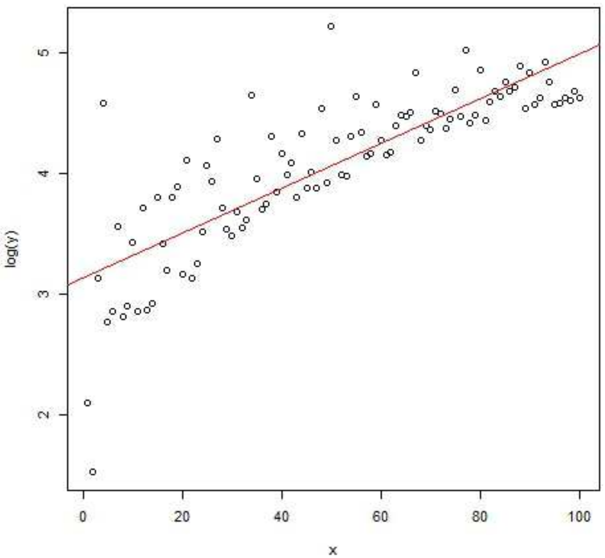


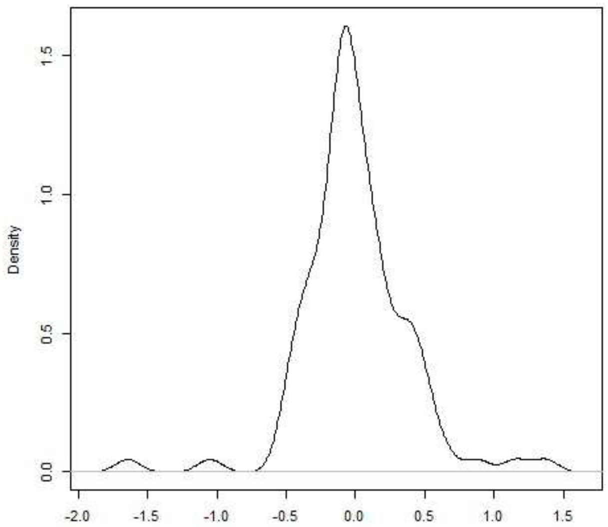


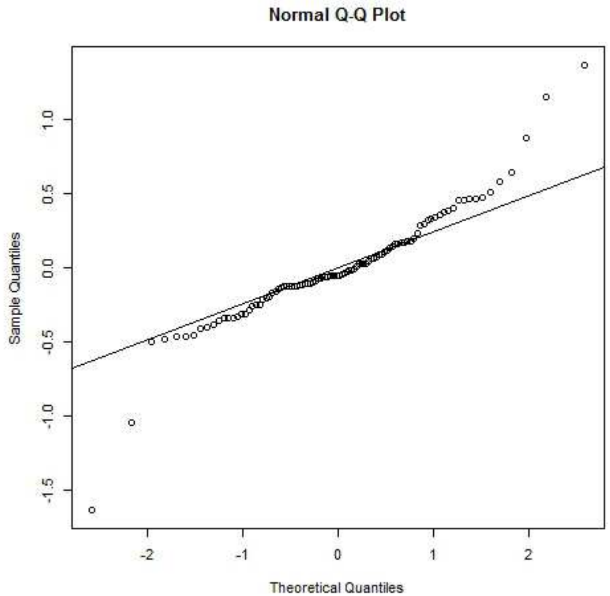
Breaking assumptions

What if the error is not normal?

- Log-transform y-variable
- $\beta_1 = y(x+1)/y(x)$



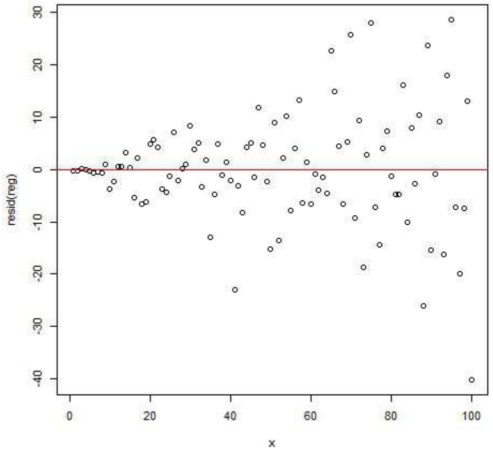


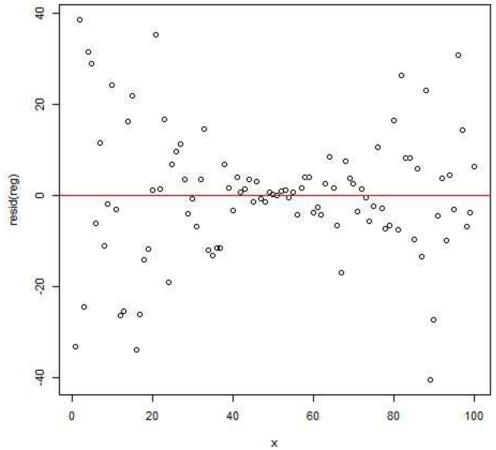


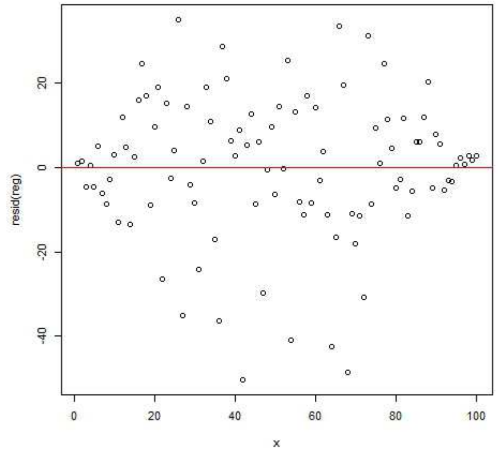
Breaking assumptions

What if the standard deviation varies with x?

- Heteroscedasticity
- Coefficients remain unbiased
- Standard errors (and p-values) are wrong



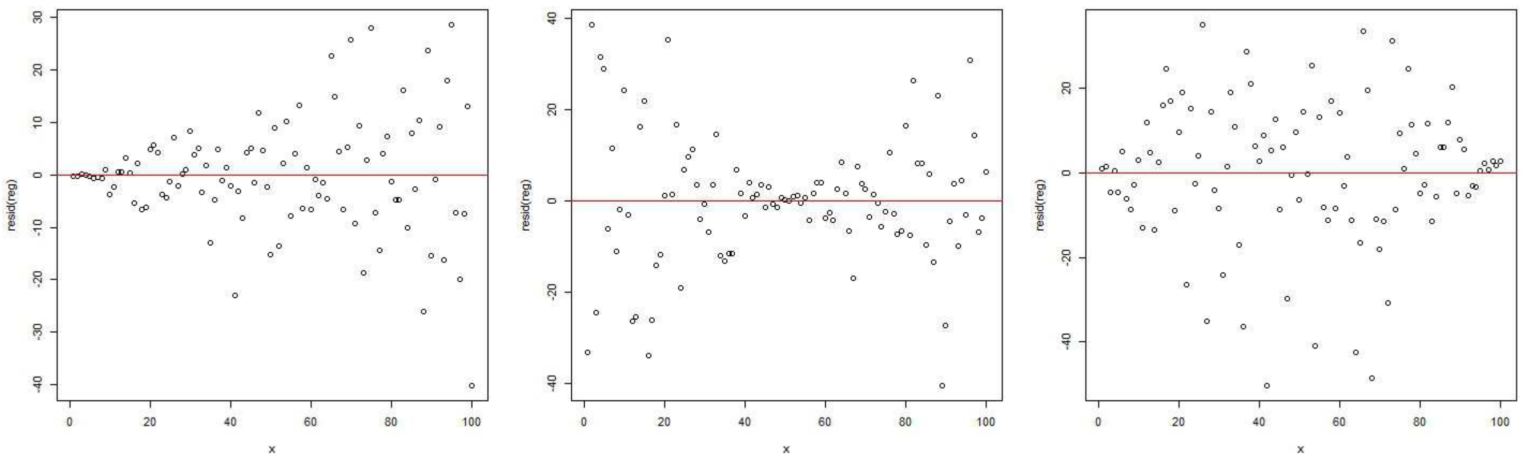




Breaking assumptions

What if the standard deviation varies with x?

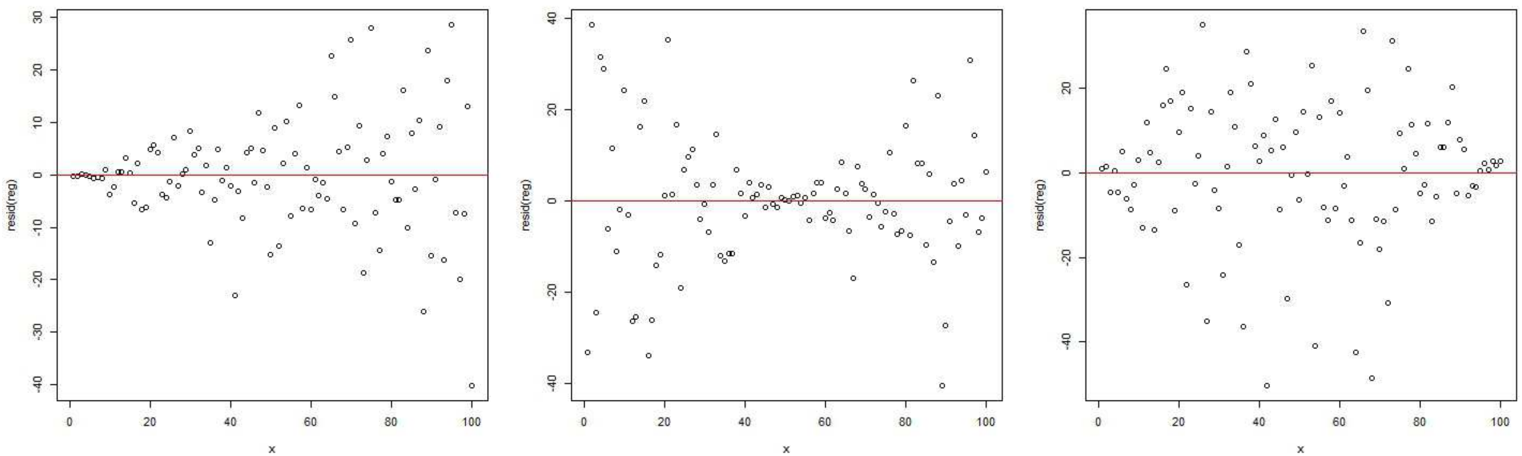
- More emphasis is put on areas where the standard deviation is large
- Normally NOT very important



Breaking assumptions

Causes of heteroscedasticity

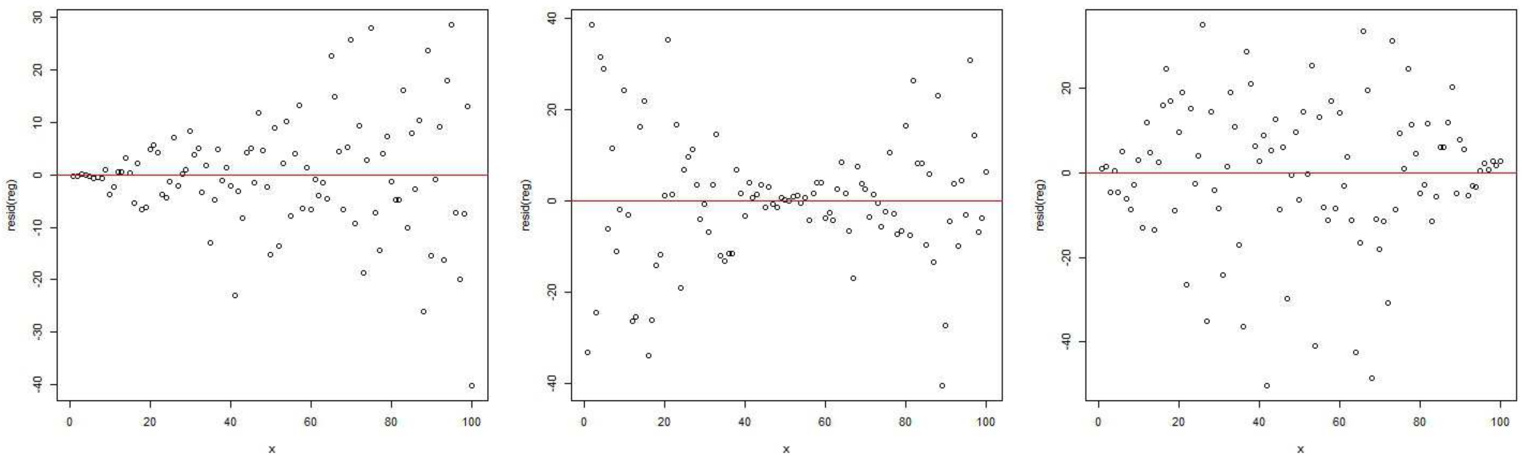
- Large X necessary for large Y (e.g., spending)
- Measurement errors (uncertainty about X)
- Subpopulation differences



Breaking assumptions

Causes of heteroscedasticity

- Bad model specification!



Breaking assumptions

Cures for heteroscedasticity

- Re-specify model
  - Add/remove variables
  - Transform variables
  - Categorize variables
- Use robust standard errors
  - Stata: regress y x, vce(robust)

Breaking assumptions

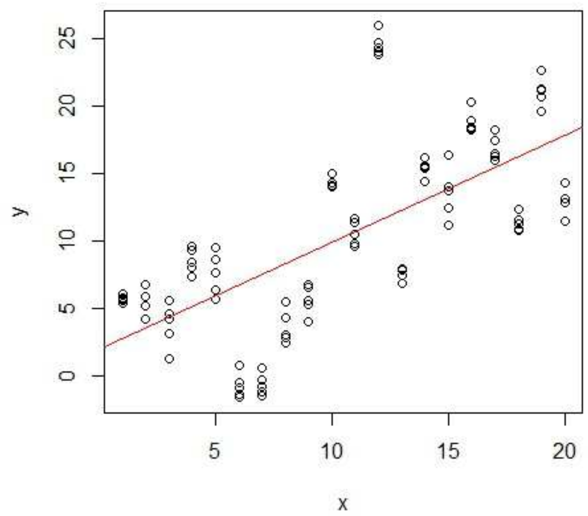
Cures for heteroscedasticity

- Logistic regression will always have heteorscedasticity
  - Outcome is 0 or 1
  - Predicted outcome is a probability (between 0 and 1)

Breaking assumptions

What if errors are not independent?

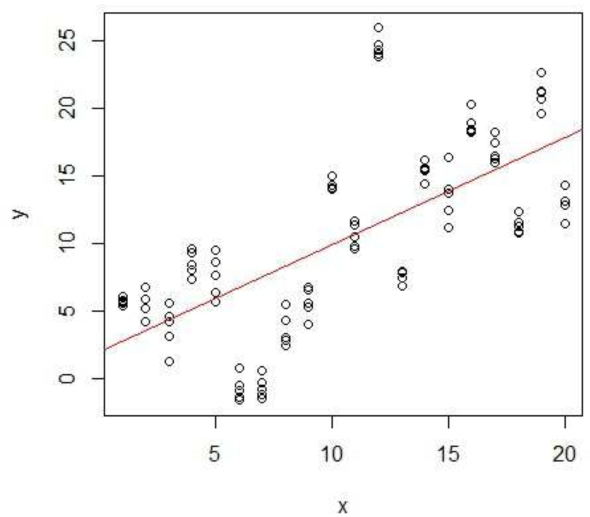
- E.g., we have $\epsilon_1 > 0$ if $\epsilon_2 > 0$
- Several observations on the same cluster
  - Same individual
  - Same family
  - Same ethnicity
  - Same gender



Breaking assumptions

What if errors are not independent?

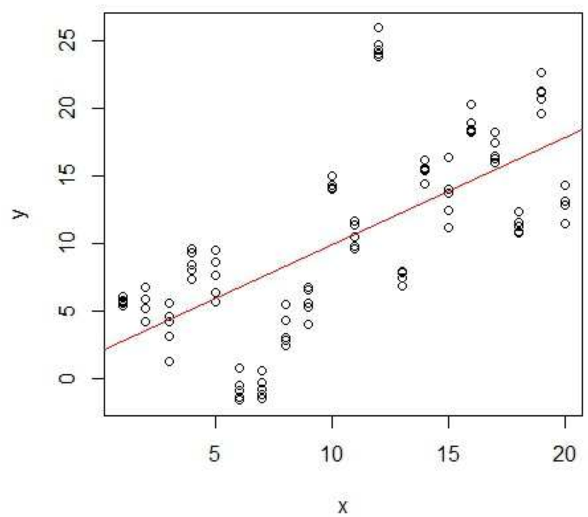
- E.g., we have $\epsilon_1 > 0$ if $\epsilon_2 > 0$
- Coefficients remain unbiased
- Too small standard errors
- Too low p-values



Breaking assumptions

What to do if errors are not independent?

- Account for the clusters
  - Stata: regress y x, vce(cluster id)
    - Mother with several children
- Random intercept
- Attend MEDSTA3/MEDLONG



Breaking assumptions

What if we have multicollinearity?

- High correlation between x-variables
  - E.g., gestational age and birth weight, or height and weight
- Coefficients will remain unbiased
- Standard errors will be too large
- p-values will be too high

Breaking assumptions

How do we detect multicollinearity?

- Calculate Variation inflation factor (VIF)
- Stata: regress y x1 x2 x3

  estat vif
- VIF>5 should prompt caution

Breaking assumptions

What to do in case of multicollinearity?

- NOT a problem…
  - …unless it involves study variable
  - …if it only involves variables that are functions of each other (e.g., x and x^2, or x, z and $x \cdot z$)
  - …if it only involves categorical varaiables with at least three categories

Breaking assumptions

What to do in case of multicollinearity?

- NOT a problem…
  - …unless it involves study variable

If the study variable has a low VIF, the standard error is not affected by a high VIF at other variables.

Breaking assumptions

What to do in case of multicollinearity?

- Multicollinearity is NOT a problem…
    - …if it only involves variables that are functions of each other (e.g., x and x^2, or x, z and $x \cdot z$)

Such variables are expected to be correlated, and p-values will not be affected. VIF can be reduced by subtracting the mean from each variable (before multiplication).

Breaking assumptions

What to do in case of multicollinearity?

- Multicollinearity is NOT a problem…
  - …if it only involves categorical varaiables with at least three categories

If the reference category is small, the VIF will be high. Choosing a reference with a larger fraction of the observations will reduce the VIF. A high VIF does not affect an overall test that all indicators have coefficients of zero.

Breaking assumptions

What to do in case of multicollinearity?

- If multicollinearity IS a problem
- Drop variables with high VIFs
    - E.g., use only one of GA and birth weight, or one of height and weight
- Change variables with high VIFs
    - E.g., use «Small for GA», not GA

Evaluation

- Home exam
  - Due on May 18
  - Focus on day 5 and day 6
  - You pass or fail (no grades)
- Oral presentation
  - May 12
  - Work in groups

Evaluation

- Will put a poll on My space (Mi side)
- Please evaluate the course
- First time in its current form
- Please don’t write:
  - «The book was retarded!»
  - «There was no use going to the lectures because he didn’t say anything that was not in the handouts.»