Biostatistics 200

Review & Short Topics

What have you accomplished this semester? (1)

- Understanding Data Types
  - Nominal/Categorical
  - Ordinal/Rank
  - Continuous/Integer
  - Continuous/Decimals
- Descriptive Statistics
  - Summarizing Data
  - Tabulating Data
  - Graphing Data

What have you accomplished this semester? (2)

- Probability
  - Union/Intersection
  - Independence
  - Mutually Exclusive
  - Conditional (Bayes)
- Distributions
  - Uniform/Binomial
  - Normal
  - Exponential/Poisson

What have you accomplished this semester? (3)

- Estimation & Confidence Bounds
  - Standardized values
- Hypothesis tests
  - Normal z test
  - Student’s t-test
- Sample Size & Power
- Non-parametric tests
- Chi-squared test (McNemar’s test)
- ANOVA & Multiple Comparisons

What have you accomplished this semester? (4)

- Correlation
  - Linear relationship
  - Pearson or Spearman?
- Simple Linear Regression
  - Continuous Dependent variable
  - Least Squares fit
  - Slope & Intercept Coefficients
  - F-test and R^2
- Logistic Regression
  - Categorical Dependent variable
  - Logit transformation
  - Odds ratios coefficients

What have you accomplished this semester? (5)

- Heteroscedasticity
- Multicollinearity
- Multiple Regression
  - Continuous Dependent variable
  - Least Squares fit
  - Continuous, ordinal, or binary independent variables
  - F-test and R^2
- Logistic Regression
  - Categorical Dependent variable
  - Logit transformation
  - Continuous, ordinal or binary independent variables
  - Odds ratios coefficients

Statistical hypothesis tests

Data and comparison type	Alternative hypotheses	Parametric test Stata command	Non-parametric test Stata command
Numerical; One mean	H_a: $\mu \neq \mu_a$ (two-sided) H_a: $\mu > \mu_a$ or $\mu < \mu_a$ (one-sided)	Z or t-test •ttest var1=hypoth val.*	
Numerical; Two means, paired data	H_a: $\mu_1 \neq \mu_2$ (two-sided) H_a: $\mu_1 > \mu_2$ or $\mu < \mu_a$ (one-sided)	Paired t-test •ttest var1=var2*	Sign test signtest var1=var2 •Wilcoxon Signed-Rank signrank var1=var2)
Numerical; Two means, independent data	H_a: $\mu_1 \neq \mu_2$ (two-sided) H_a: $\mu_1 > \mu_2$ or $\mu < \mu_a$ (one-sided)	T-test (equal or unequal variance) •ttest var1, by(byvar) unequal	Wilcoxon rank-sum test •ranksum var1, by(byvar)
Numerical, Two or more means, independent data	H_a: $\mu_1 \neq \mu_2$ or $\mu_1 \neq \mu_3$ or $\mu_2 \neq \mu_3$ etc.	ANOVA oneway var1 byvar	Kruskal Wallis test •kwallis var1, by(byvar)
Dichotomous; One proportion	H_a: $p \neq p_a$ (two-sided) H_a: $p > p_a$ or $p < p_a$ (one-sided)	Proportion test • prtest var1=hypoth value* • bitest var1=hypoth value	
Dichotomous; two proportions	H_a: $p_1 \neq p_2$ (two-sided) H_a: $p_1 > p_2$ (one-sided)	Proportion test (z-test) • prtest var1, by(byvar)	Chi-square test tab var1 var2, chi exact McNemar's for paired data: mcc var1 var2
Categorical by categorical (nxk)	H_a : The rows not independent of the columns		Chi-square test •tab var1 var2, chi exact

Further methods

Dependent variable (y)	**Independent variable(s) type(s) (x)**	**Methods**	**Non-parametric methods**
Numerical/continuous	One numerical/continuous variable	Pearson correlation; simple linear regression	Spearman rank correlation
Numerical/continuous	One categorical variable	Linear regression, t-test, ANOVA	Sign test, Signed rank test, Wilcoxon rank sum, Kruskal-Wallis
Numerical/continuous	Multiple variables/types of variables	Linear regression	
Dichotomous	One categorical variable	Logistic regression (or proportion test for 2x2)	Chi-square test or Fisher exact test
Dichotomous	One numerical/continuous variable	Logistic regression	
Dichotomous	Multiple variables/types of variables	Logistic regression	

Four things to do and remember

- Download the slides/examples/do files etc for your future research – some of the techniques may come in handy and you can repurpose them
- You can always find me for a short recap or question – I'm probably going to be around UCSF for awhile so even if you have a committee for your certificate or degree or your next clinical trial – feel free to contact me
- My email is: Isabel.allen@ucsf.edu and I'm on the 2nd floor
- If this course is part of your TICR Master's degree and you now wish you were a biostatistician, you can request to be my TA next fall.

Short Topics

Short Topic 1: Transformations!

- Getting Normal
- Using the Linear Model
- Do we need to?

Reasons to Transform

- Variables do not fit a normal distribution and parametric tests are desired
- A relationship between two variables is non-linear but transformation would allow the use of linear regression
- Errors in measurement are multiplicative rather than additive, e.g. ± 2% rather than ± 2mm
- Measurements are products rather than sums of other measurements, e.g. area, volume
- Counts follow binomial, Poisson, or negative binomial distributions which are often asymmetrical unless sample sizes are large

Solutions

- Use non-parametric methods that do not depend on the normality of the data (increasingly easy to do)
- Use data transformations that shift the distribution to one that is normal
- The transformation must preserve the order of the original data – we only change the spacing between data points
- Right skewed data with many zeros cannot be transformed effectively since nothing can stretch out observations that have the same value

Example: What do I need? Homoscedasticity

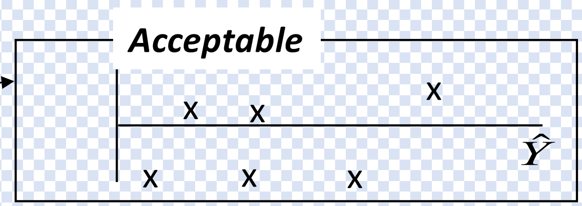
How do I know I have a problem?

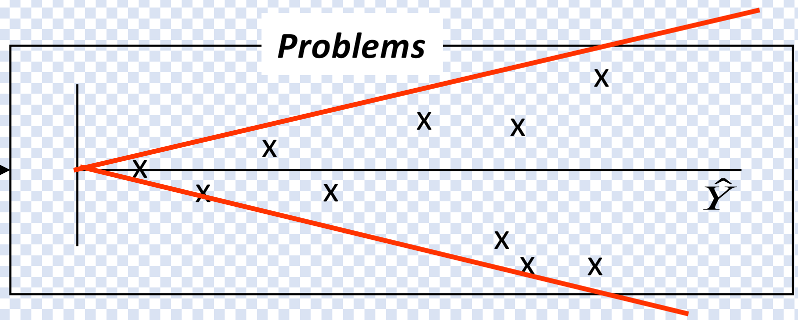
$$\hat{y}_i \ \text{vs.}\ \hat{\varepsilon}_i$$

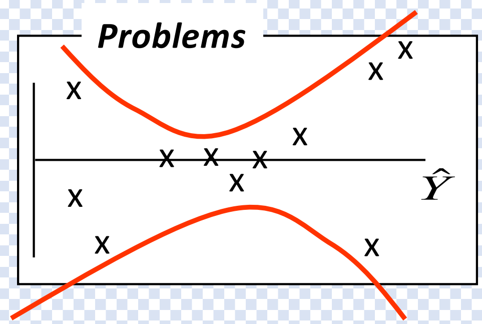
Plot predicted (fitted) values versus residuals.

What is the pattern of the spread in the residuals as the predicted values increase?

- Spread constant.
- Spread increases.
- Spread decreases then increases.





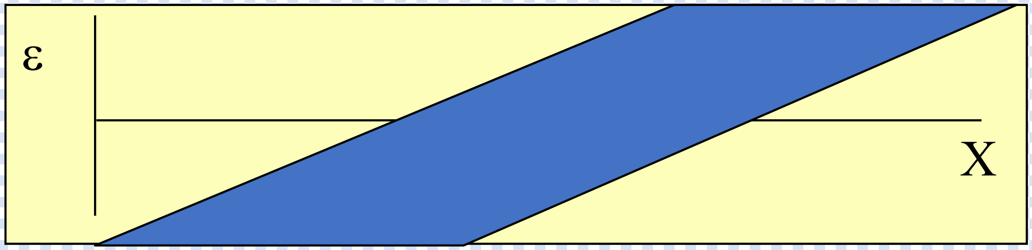


Example: What do I need? Homogeneity

What to do?

- Attempt a transformation.
- Weighted regression.
- Incorporate additional covariates.
- Non-linear regression.

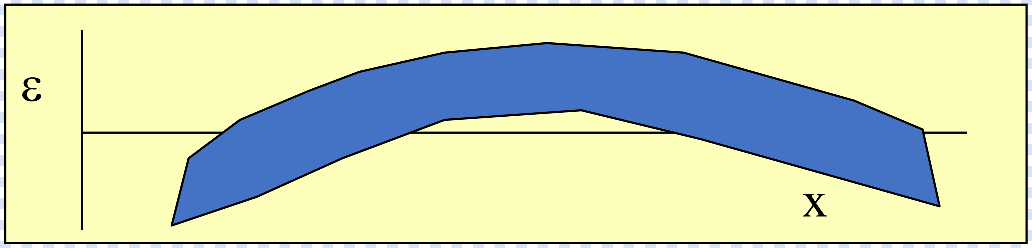
What to do if the spread of the residuals plotted versus **X** looks like this?



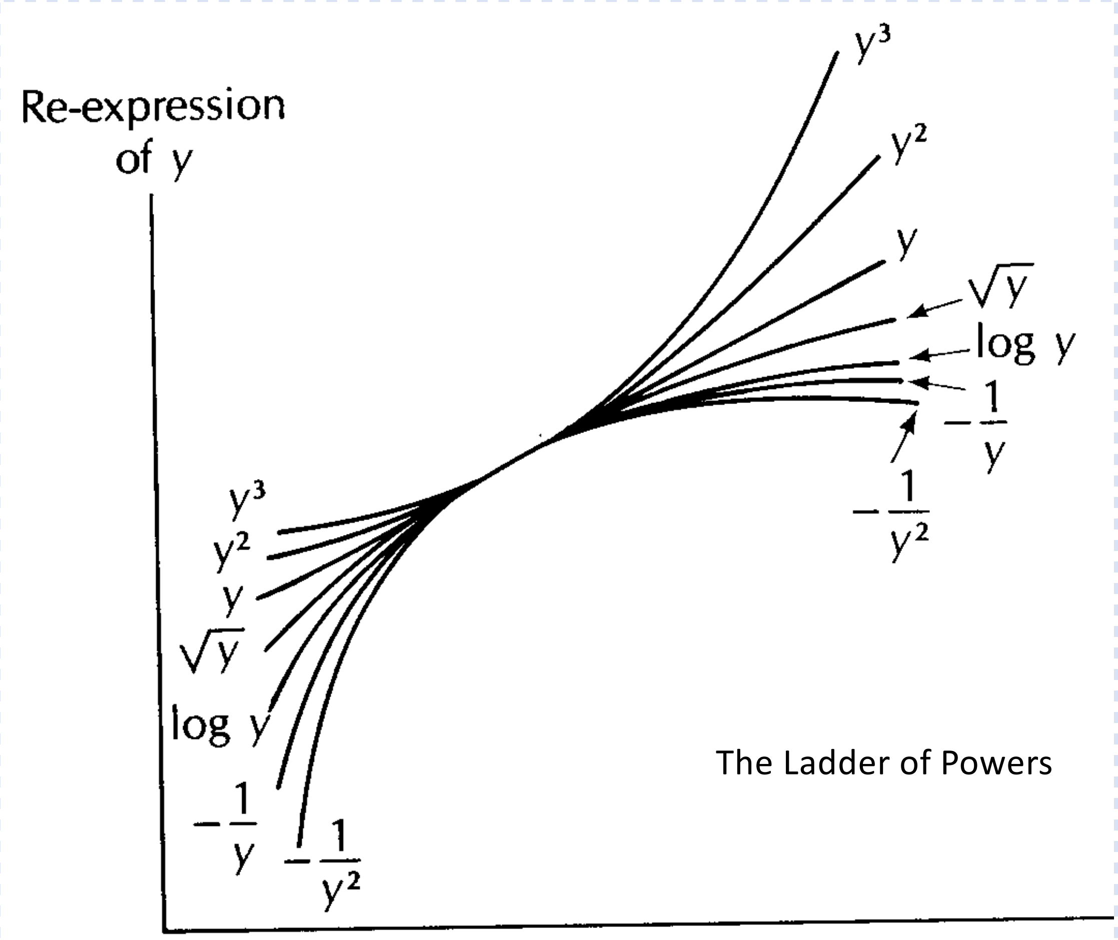
Need another x variable.

or this?

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$



Transforming the Response to achieve Linearity

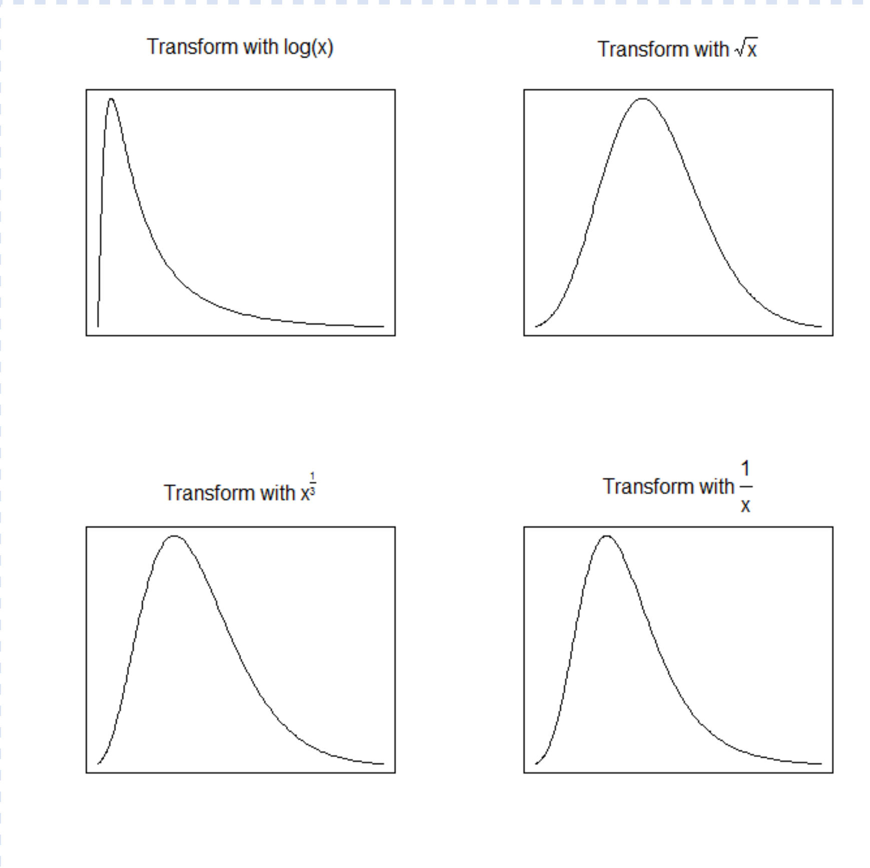


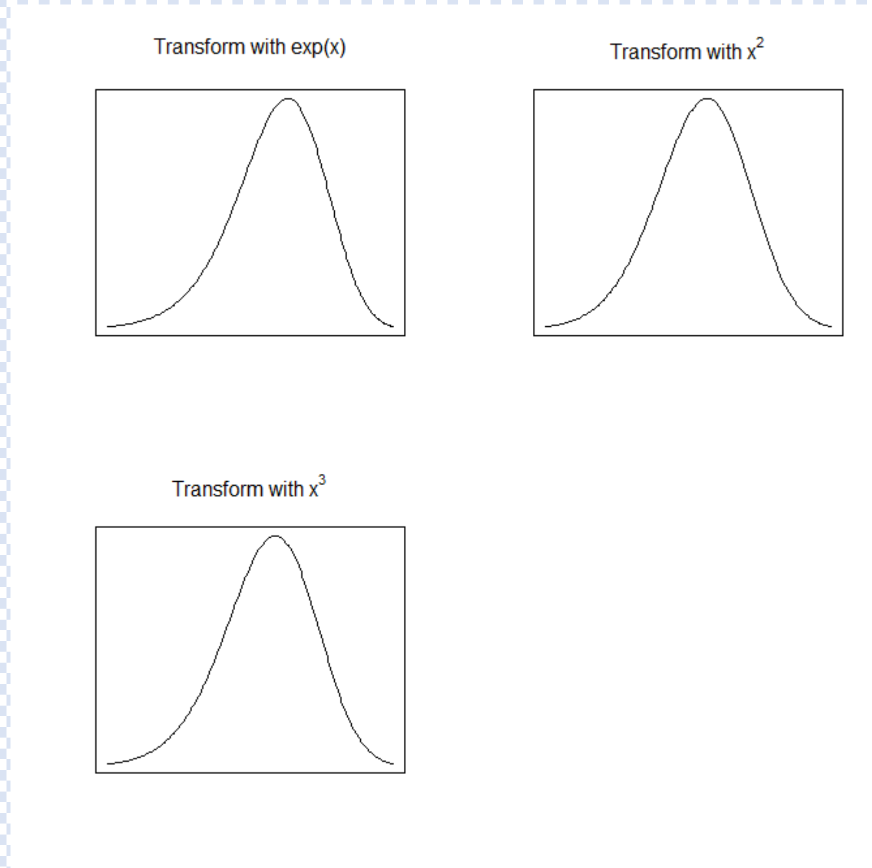
If a scatterplot of y versus x curves upward, proceed down on the scale to choose a transformation.

Common Transformations

- Tail to the right
  - Natural or common (base 10) logarithm – no zero values
  - Square root, cube root, etc – zeros ok
  - Inverse, $-1/x$, $-1/x^2$, etc – no zero values
- Tail to the left
  - Exponential e^x,10^x (low values)
  - Square, cube, etc

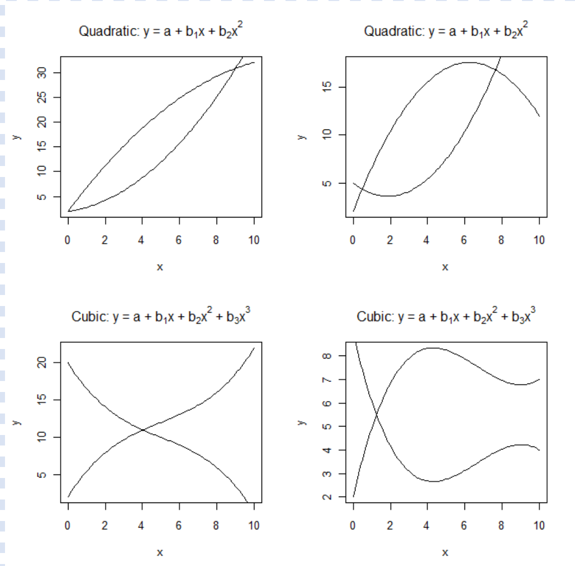
Common Transformations





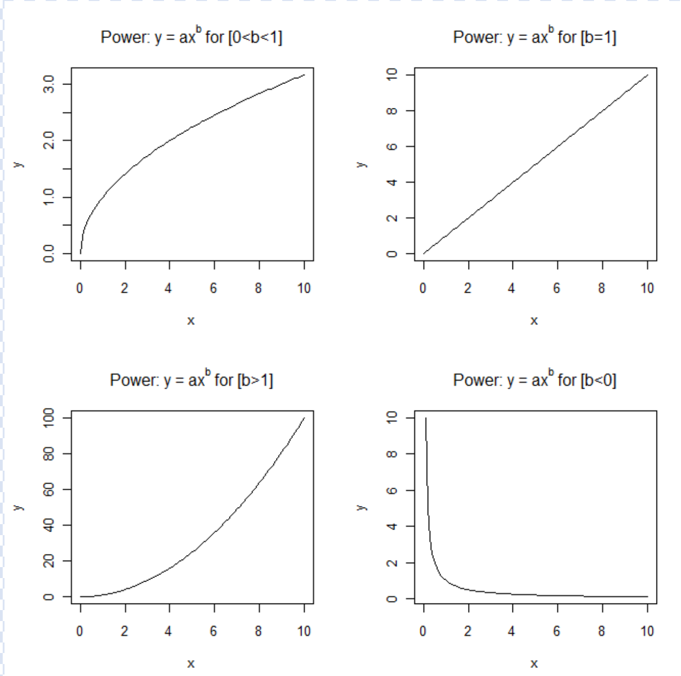
Transforming to Linear

- By transforming variables before using linear regression we can fit nonlinear equations
- In some cases we can express the fitted equation in terms of the original untransformed variables
- Polynomial: Begin with linear and then work up to quadratic, cubic, and so on until the new terms are not significant



Transforming to Linear

- Power Function
- Log-log transformation
- Use log() to transform dependent and independent variables



Transforming to Linear

- Exponential Function:
  - Semi-log transformation
  - Use log() to transform dependent variable, y > 0
- Inverse Function
  - Reciprocal transformation – 1/x where x ≠ 0
  - Used for distance models – marriage, trade, social interaction declines with distance