

# STAT 401A - Statistical Methods for Research Workers

## Regression diagnostics

Jarad Niemi (Dr. J)

Iowa State University

last updated: October 20, 2014

# All models are wrong!

George Box (Empirical Model-Building and Response Surfaces, 1987):

*All models are wrong, but some are useful.*

http:

[//stats.stackexchange.com/questions/57407/what-is-the-meaning-of-all-models-are-wrong-but-some-are-useful](http://stats.stackexchange.com/questions/57407/what-is-the-meaning-of-all-models-are-wrong-but-some-are-useful)

*“All models are wrong” that is, every model is wrong because it is a simplification of reality. Some models, especially in the “hard” sciences, are only a little wrong. They ignore things like friction or the gravitational effect of tiny bodies. Other models are a lot wrong - they ignore bigger things.*

*“But some are useful” - simplifications of reality can be quite useful. They can help us explain, predict and understand the universe and all its various components.*

*This isn't just true in statistics! Maps are a type of model; they are wrong. But good maps are very useful.*

# Regression

The simpler linear regression model is

$$Y_i \stackrel{\text{ind}}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

this can be rewritten as

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad e_i \stackrel{\text{ind}}{\sim} N(0, \sigma^2)$$

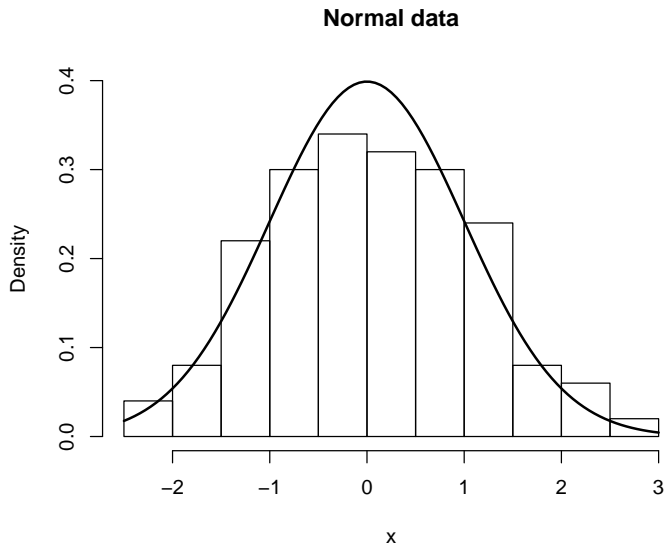
where we estimate the errors via the residuals

$$r_i = \hat{e}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$

Key assumptions are:

- Normality of the errors
- Constant variance of the errors
- Independence between observations
- Linearity between mean response and explanatory variable

# Histograms with best fitting bell curves



# Normal QQ-plot

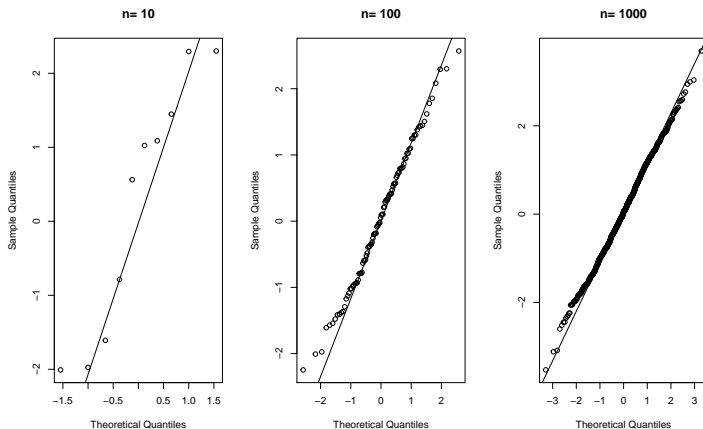
## Definition

The quantile-quantile or qq-plot is an exploratory graphical device used to check the validity of a distributional assumption for a data set.

A normal qq-plot graphs the theoretical quantiles from a normal distribution versus the observed quantiles.

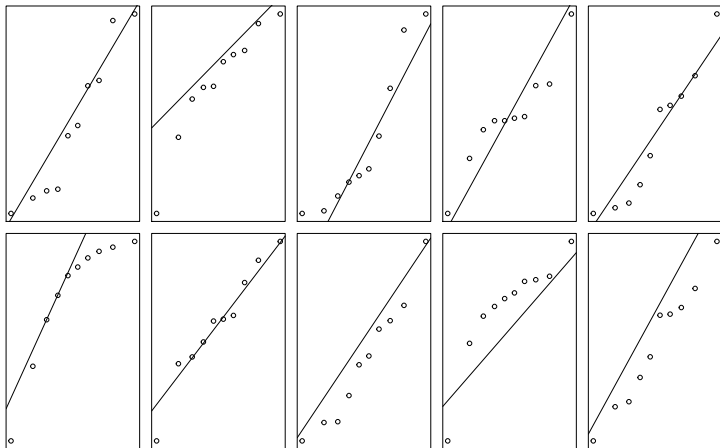
**Remark** The bottom line is that, if the distribution assumption is satisfied, the points should fall roughly along the  $y=x$  line.

# Normal



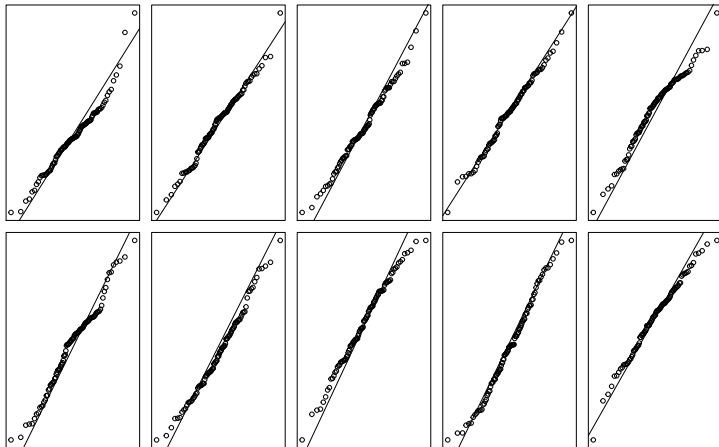
SAS swaps the x and y axes

# Normal ( $n=10$ )



SAS swaps the x and y axes

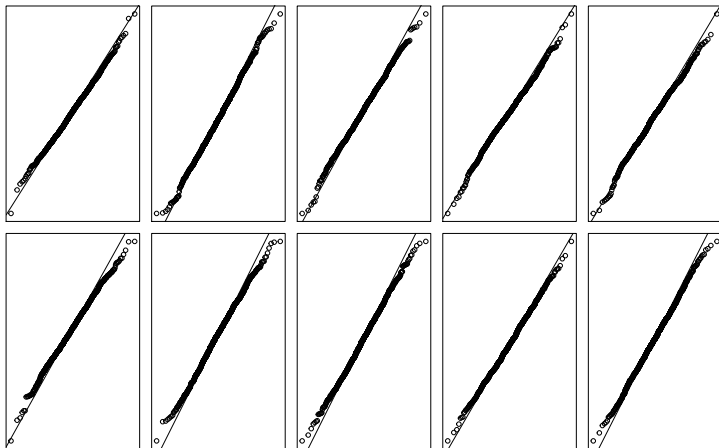
# Normal( $n=100$ )



SAS swaps the x and y axes



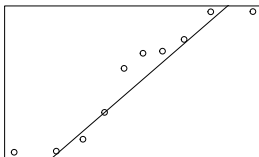
# Normal ( $n=1000$ )



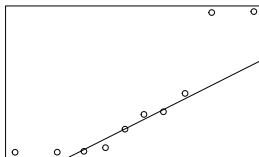
SAS swaps the x and y axes

# Not normal ( $n=10$ )

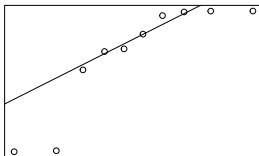
normal



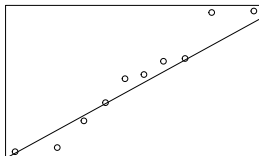
right-skewed



left-skewed

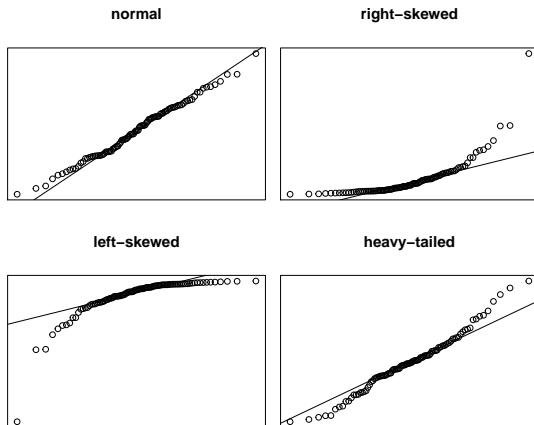


heavy-tailed



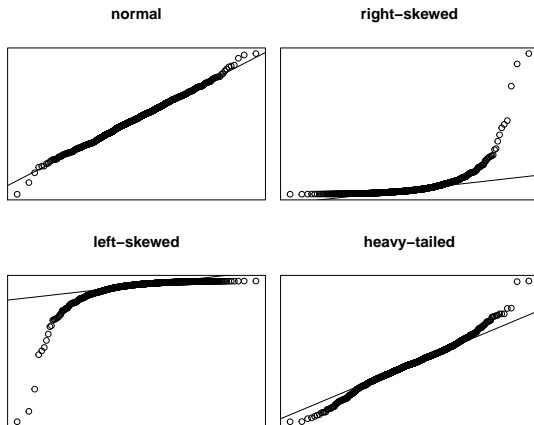
SAS swaps the x and y axes

# Not normal ( $n=100$ )



SAS swaps the x and y axes

# Not normal ( $n=1000$ )



SAS swaps the x and y axes

# Constant variance

Recall the model

$$Y_i = \beta_0 + \beta_x X_i + e_i \quad e_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

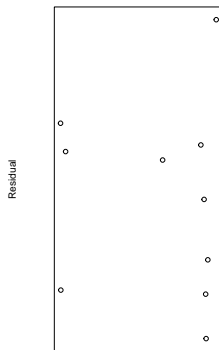
so the variance for the  $e_i$  is constant.

To assess this assumption, we look at plots of residuals vs anything and look for patterns that show different “spreads”, e.g.

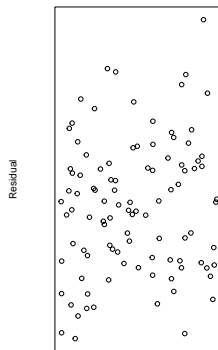
- funnels
- football shapes

The most common way this assumption is violated is by having increasing variance with increasing mean, thus we often look at a residuals vs predicted (fitted) mean plot.

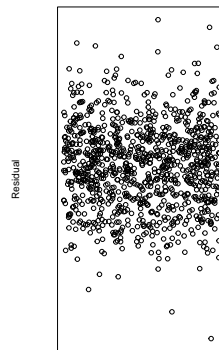
# Constant variance



Predicted mean

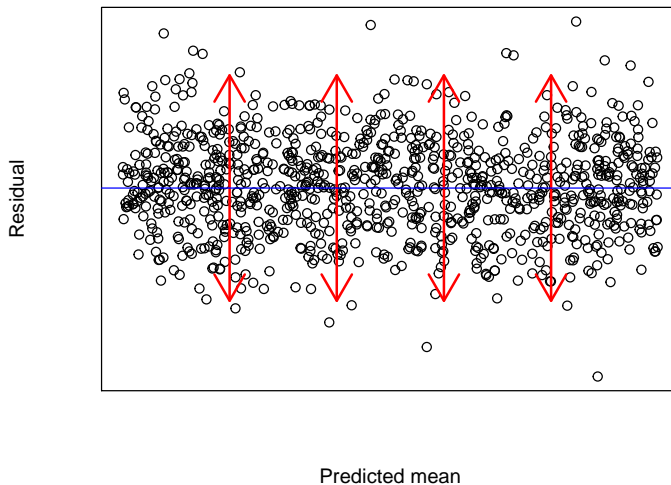


Predicted mean

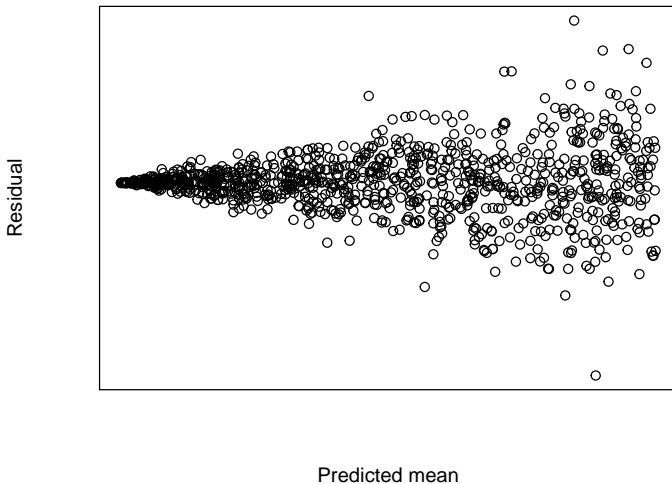


Predicted mean

# Constant variance

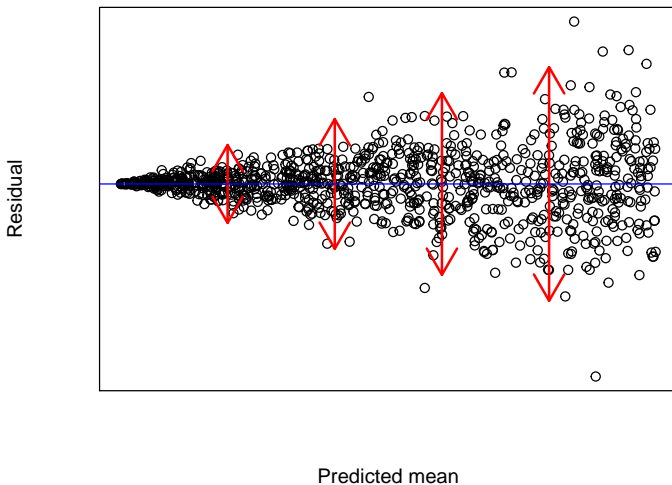


# Extreme Non-Constant variance (funnel)

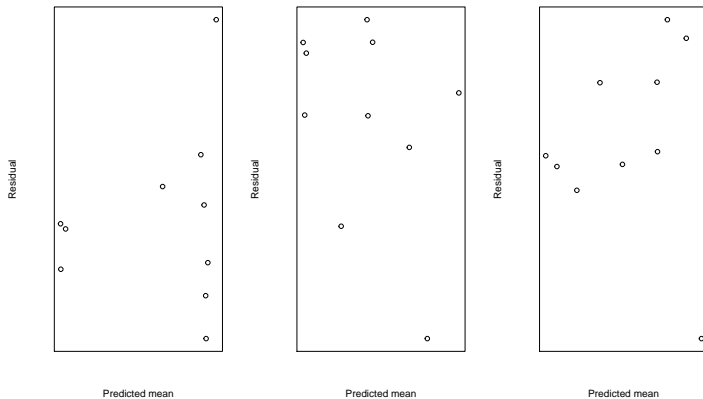




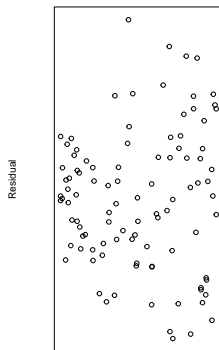
# Extreme non-Constant variance (funnel)



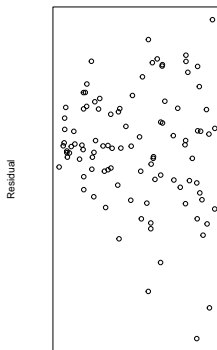
# Non-constant variance ( $n=10$ , $\sigma_2/\sigma_1 = 4$ )



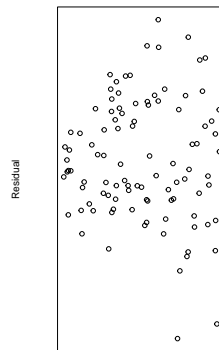
# Non-constant variance ( $n=100$ , $\sigma_2/\sigma_1 = 4$ )



Predicted mean

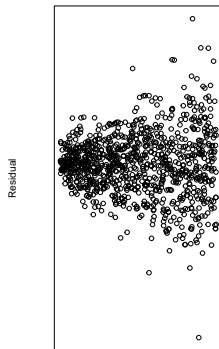


Predicted mean

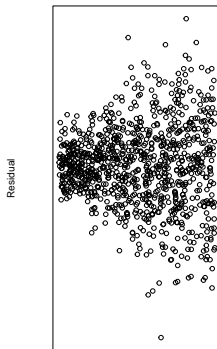


Predicted mean

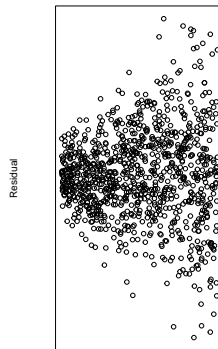
# Non-constant variance ( $n=1000$ , $\sigma_2/\sigma_1 = 4$ )



Predicted mean

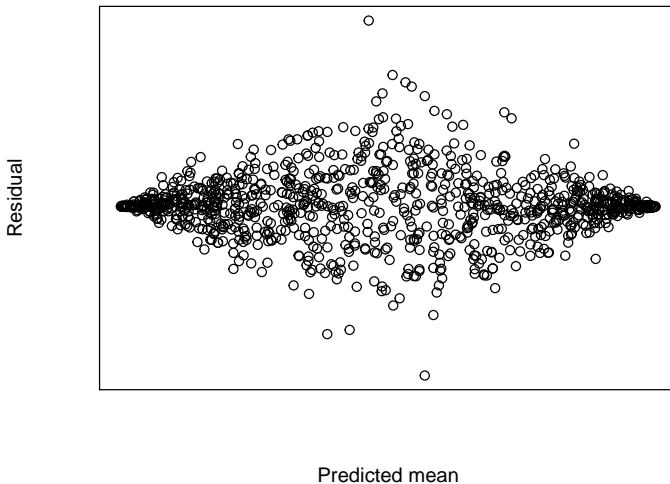


Predicted mean



Predicted mean

# Extreme non-Constant variance (football)



# Independence

Lack of independence includes

- Cluster effect
- Serial correlation
- Spatial association

Make plots of residuals vs relevant explanatory variables and look for patterns, e.g.

- Residuals vs groups (prefer blocking)
- Residuals vs time (or observation number)
- Residuals vs spatial variable

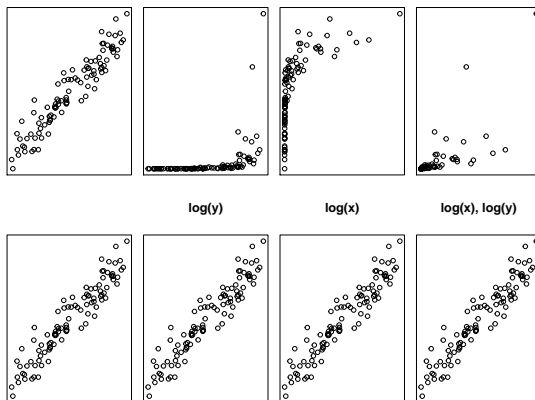
# Summary

Often the best strategy is graphical exploration of the data, here are some relevant graphs:

- transformed response vs transformed explanatory
- transformed response vs transformed explanatory
- qqplot of residuals
- residual vs fitted value
- residual vs explanatory
- residual vs observation number
- residual vs any other variable

# Linearity

Assess using scatterplots of (transformed) response vs (transformed) explanatory variable:





# Testing Composite hypotheses

Comparing two models

- $H_0$  : (reduced)
- $H_1$  : (full)

Do the following

1. Calculate extra sum of squares.
2. Calculate extra degrees of freedom
3. Calculate

$$\text{F-statistic} = \frac{\text{Extra sum of squares} / \text{Extra degrees of freedom}}{\hat{\sigma}_{full}^2}$$

4. Compare this to an F-distribution with

- numerator degrees of freedom = extra degrees of freedom
- denominator degrees of freedom = degrees of freedom in estimating  $\hat{\sigma}_{full}^2$

# Lack-of-fit F-test

Let  $Y_{ij}$  be the  $i^{th}$  observation from the  $j^{th}$  group where the group is defined by those observations having the same explanatory variable value ( $X_j$ ).

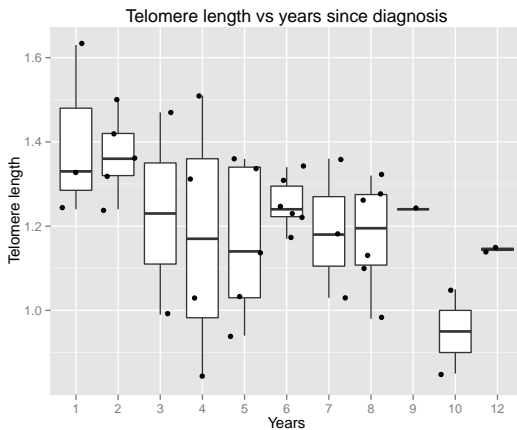
Two models:

$$\text{ANOVA: } Y_{ij} \stackrel{\text{ind}}{\sim} N(\mu_j, \sigma^2) \quad (\text{full})$$

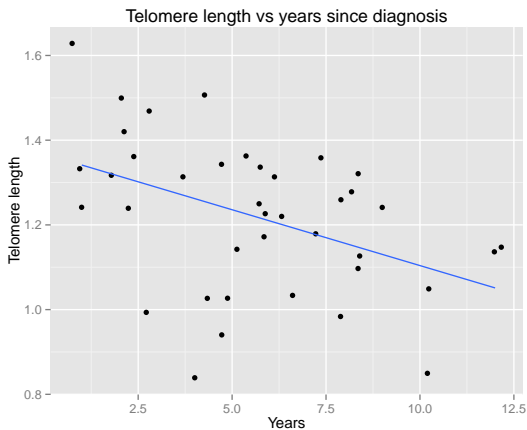
$$\text{Regression: } Y_{ij} \stackrel{\text{ind}}{\sim} N(\beta_0 + \beta_1 X_j, \sigma^2) \quad (\text{reduced})$$

- Regression model is reduced:
  - ANOVA has  $J$  parameters for the mean
  - Regression has 2 parameters for the mean
  - Set  $\mu_j = \beta_0 + \beta_1 X_j$ .
- Small pvalues indicate a lack-of-fit, i.e. the reduced model is not adequate.
- Lack-of-fit F-test requires multiple observations at a few  $X_j$  values!

# Telomere length



# Telomere length



# SAS code

```
DATA t;
  INFILE 'telomeres.csv' DSD FIRSTOBS=2;
  INPUT years length;

PROC REG DATA=t;
  MODEL length = years / CLB LACKFIT;
  RUN;
```

The REG Procedure  
 Model: MODEL1  
 Dependent Variable: length

Number of Observations Read	39
Number of Observations Used	39

## Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.22777	0.22777	8.42	0.0062
Error	37	1.00033	0.02704		
Lack of Fit	9	0.18223	0.02025	0.69	0.7093
Pure Error	28	0.81810	0.02922		
Corrected Total	38	1.22810			

Indicates no evidence for a lack of fit, i.e. regression seems adequate.

```
# Use as.factor to turn a continuous variable into a categorical variable
m_anova = lm(telomere.length ~ as.factor(years), Telomeres)
m_reg   = lm(telomere.length ~ years, Telomeres)
anova(m_reg, m_anova)
```

Analysis of Variance Table

Model 1: telomere.length ~ years

Model 2: telomere.length ~ as.factor(years)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	37	1.000				
2	28	0.818	9	0.182	0.69	0.71

No evidence of a lack of fit.

# Lack-of-fit F-test summary

- Lack-of-fit F-test tests the assumption of linearity
- Needs multiple observations at various explanatory variable values
- Small pvalue indicates a lack-of-fit, i.e. means are not linear
  - Transform response, e.g. log
  - Transform explanatory variable
  - Add other explanatory variables