# STAT 401A - Statistical Methods for Research Workers

## Regression diagnostics

Jarad Niemi (Dr. J)

Iowa State University

last updated: October 24, 2014

# All models are wrong!

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

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

# 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`

*"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.*

# 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`

*"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.*

# 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`

> *"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

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

# Regression

The simpler linear regression model is

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

this can be rewritten as

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

# Regression

The simpler linear regression model is

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

this can be rewritten as

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

where we estimate the errors via the residuals

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

# Regression

The simpler linear regression model is

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

this can be rewritten as

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

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

# Regression

The simpler linear regression model is

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

this can be rewritten as

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

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

# Regression

The simpler linear regression model is

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

this can be rewritten as

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

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 of the errors

# Regression

The simpler linear regression model is

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

this can be rewritten as

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

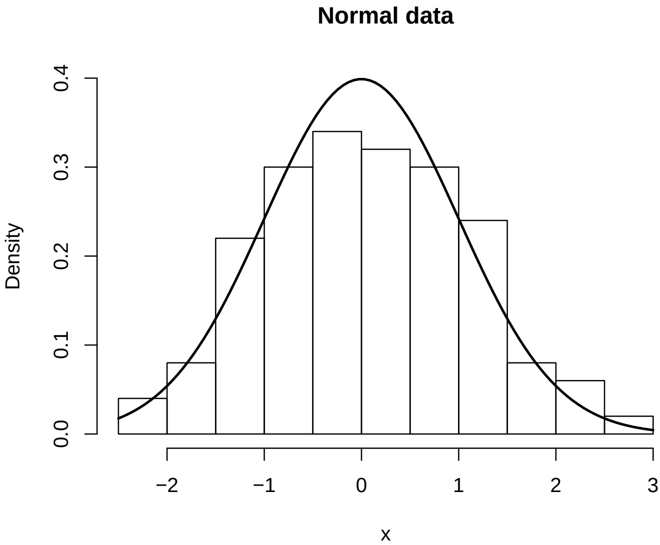
where we estimate the errors via the residuals

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

Key assumptions are:

- Normality of the errors
- Constant variance of the errors
- Independence of the errors
- 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.

# 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.

# 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. With a line that indicates perfect normality.

# 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. With a line that indicates perfect normality.

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

# Normal QQ-plot

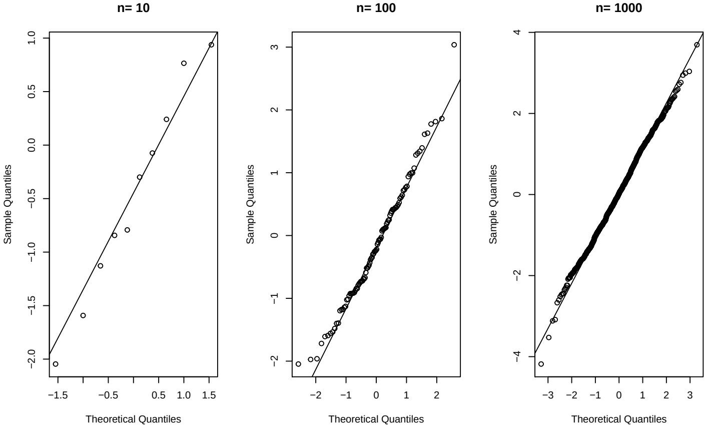
### Definition

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

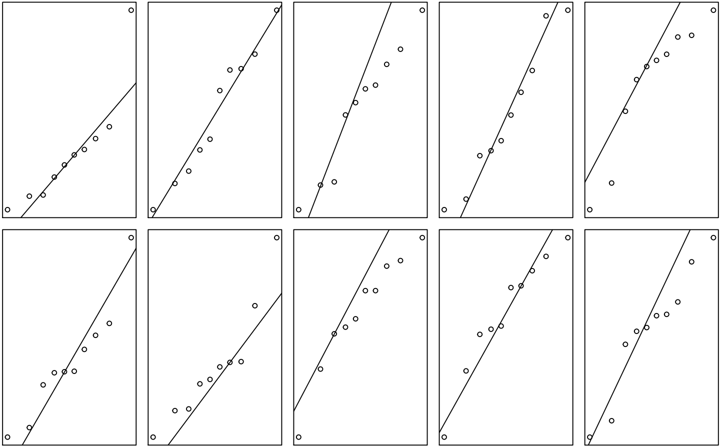
A normal qq-plot graphs the theoretical quantiles from a normal distribution versus the observed quantiles. With a line that indicates perfect normality.

**Remark** The bottom line is that, if the distribution assumption is satisfied, the points should fall roughly along the line. Systematic variation from this line indicates skewness,

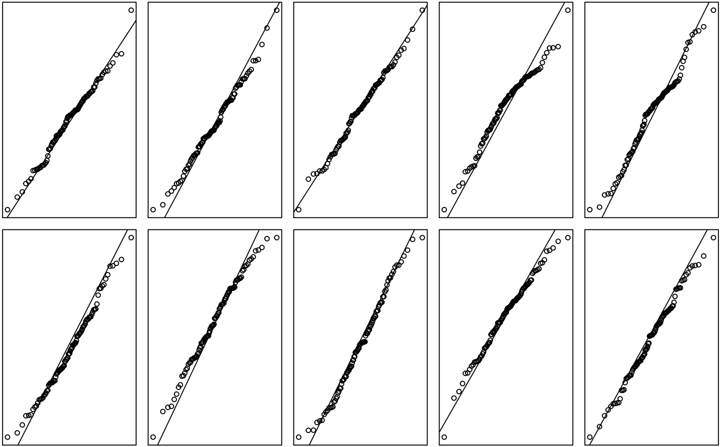
# Normal


n= 10
n= 100
n= 1000
Sample Quantiles
Theoretical Quantiles
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1.5
−0.5
0.5
1.0
1.5
3
2
1
0
−1
−2
−2
−1
0
1
2
4
2
0
−2
−4
−3
−2
−1
0
1
2
3

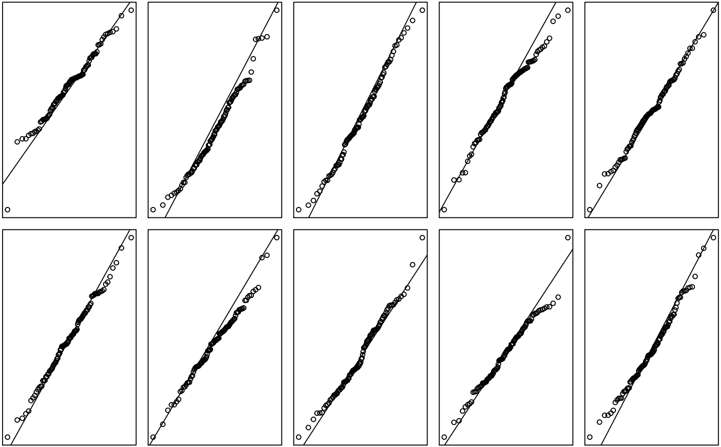
# Normal (n=10)



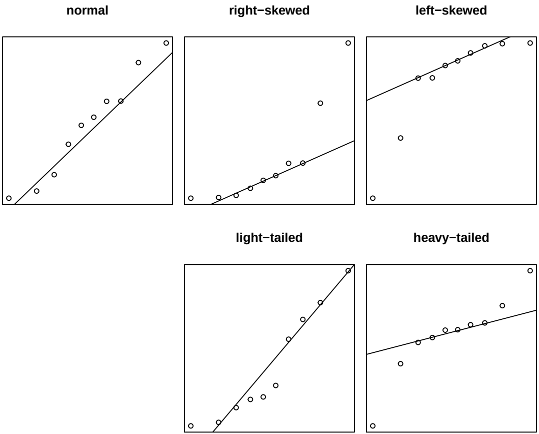
# Normal(n=100)



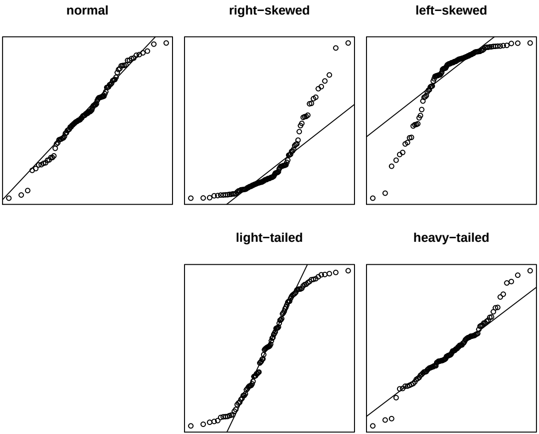
# Normal (n=1000)



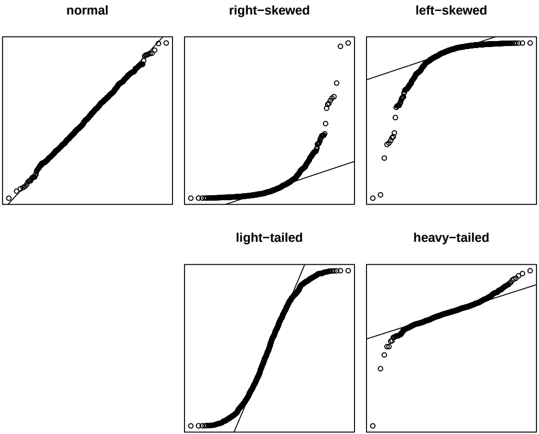
# Not normal (n=10)



# Not normal (n=100)



# Not normal (n=1000)



# Summary

For normal qq-plots with (standardized) residuals (y-axis) vs theoretical quantiles (x-axis), the following interpretations apply

# Summary

For normal qq-plots with (standardized) residuals (y-axis) vs theoretical quantiles (x-axis), the following interpretations apply

- If the residuals fall roughly along the line, then normality is reasonable.

# Summary

For normal qq-plots with (standardized) residuals (y-axis) vs theoretical quantiles (x-axis), the following interpretations apply

- If the residuals fall roughly along the line, then normality is reasonable.
- If the residuals have a U pattern, then there is right-skewness.

# Summary

For normal qq-plots with (standardized) residuals (y-axis) vs theoretical quantiles (x-axis), the following interpretations apply

- If the residuals fall roughly along the line, then normality is reasonable.
- If the residuals have a U pattern, then there is right-skewness.
- If the residuals have an upside down U pattern, then there is left-skewness.

# Summary

For normal qq-plots with (standardized) residuals (y-axis) vs theoretical quantiles (x-axis), the following interpretations apply

- If the residuals fall roughly along the line, then normality is reasonable.
- If the residuals have a U pattern, then there is right-skewness.
- If the residuals have an upside down U pattern, then there is left-skewness.
- If the residuals have an S pattern, then there are light tails (we are not too concerned about this situation because our inferences will be conservative).

# Summary

For normal qq-plots with (standardized) residuals (y-axis) vs theoretical quantiles (x-axis), the following interpretations apply

- If the residuals fall roughly along the line, then normality is reasonable.
- If the residuals have a U pattern, then there is right-skewness.
- If the residuals have an upside down U pattern, then there is left-skewness.
- If the residuals have an S pattern, then there are light tails (we are not too concerned about this situation because our inferences will be conservative).
- If the residuals have a rotated Z pattern, then there are heavy tails.

# Summary

For normal qq-plots with (standardized) residuals (y-axis) vs theoretical quantiles (x-axis), the following interpretations apply

- If the residuals fall roughly along the line, then normality is reasonable.
- If the residuals have a U pattern, then there is right-skewness.
- If the residuals have an upside down U pattern, then there is left-skewness.
- If the residuals have an S pattern, then there are light tails (we are not too concerned about this situation because our inferences will be conservative).
- If the residuals have a rotated Z pattern, then there are heavy tails.

# Summary

For normal qq-plots with (standardized) residuals (y-axis) vs theoretical quantiles (x-axis), the following interpretations apply

- If the residuals fall roughly along the line, then normality is reasonable.
- If the residuals have a U pattern, then there is right-skewness.
- If the residuals have an upside down U pattern, then there is left-skewness.
- If the residuals have an S pattern, then there are light tails (we are not too concerned about this situation because our inferences will be conservative).
- If the residuals have a rotated Z pattern, then there are heavy tails.

Other patterns are certainly possible, but these are the most common.

# Constant variance

Recall the model

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

so the variance for the $e_i$ is constant.

# Constant variance

Recall the model

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

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"

# Constant variance

Recall the model

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

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

# Constant variance

Recall the model

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

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

# Constant variance

Recall the model

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

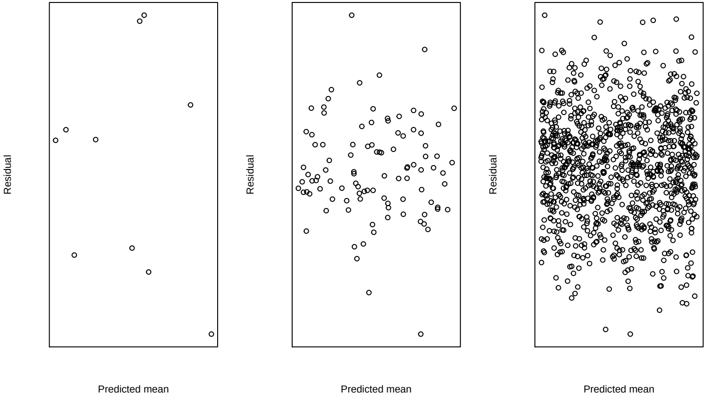
so the variance for the $e_i$ is constant.

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

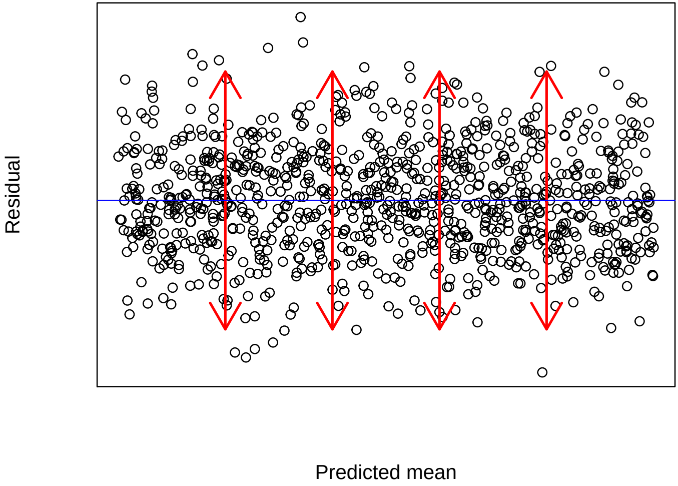
- funnels
- football shapes

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.

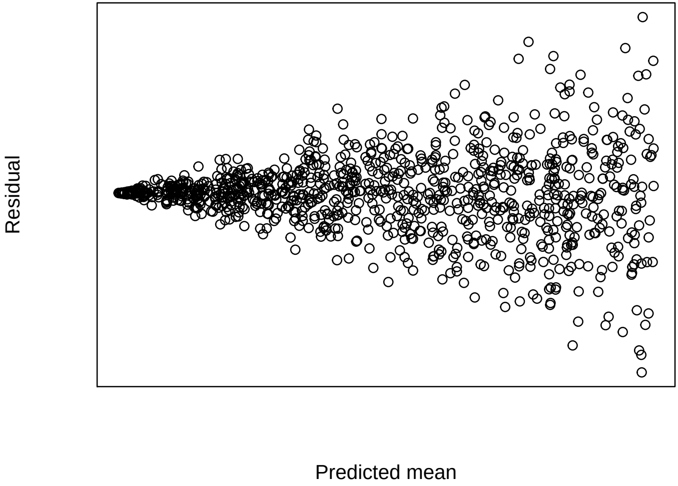
# Constant variance



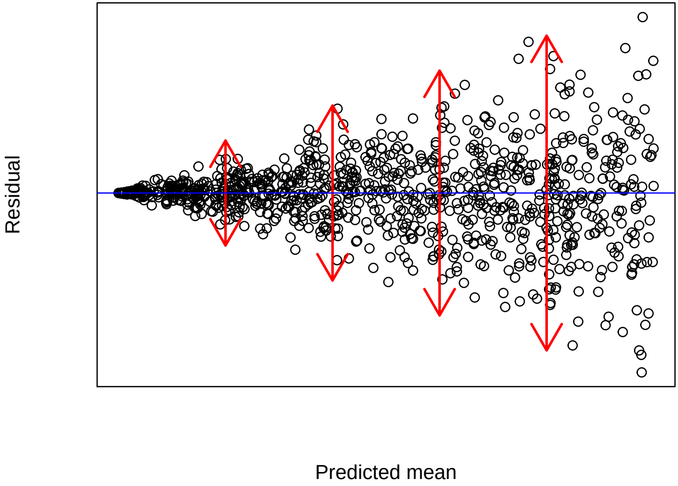
# Constant variance



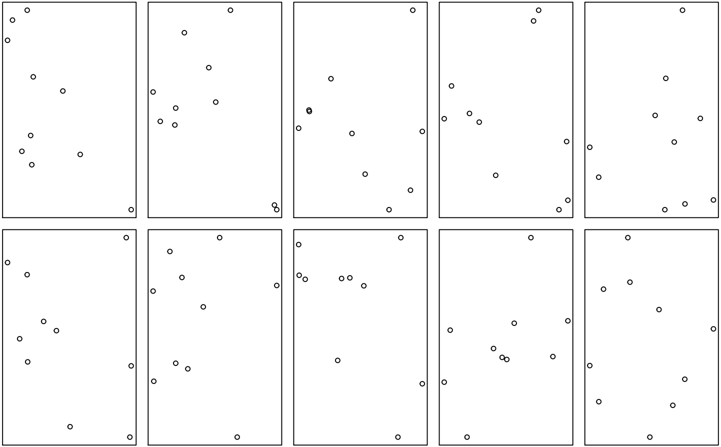
# Extreme non-constant variance (funnel)



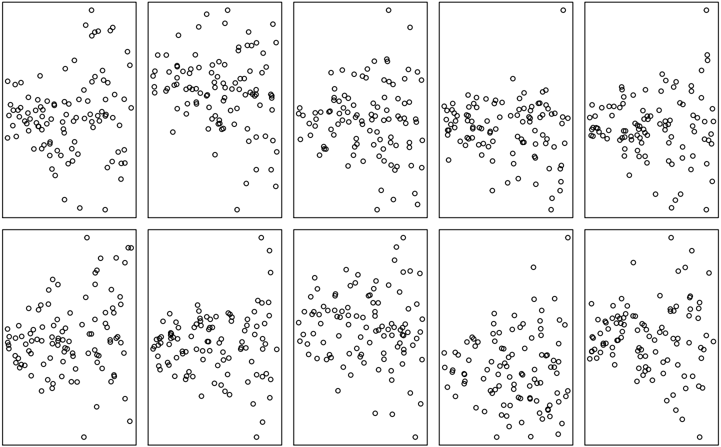
# Extreme non-constant variance (funnel)



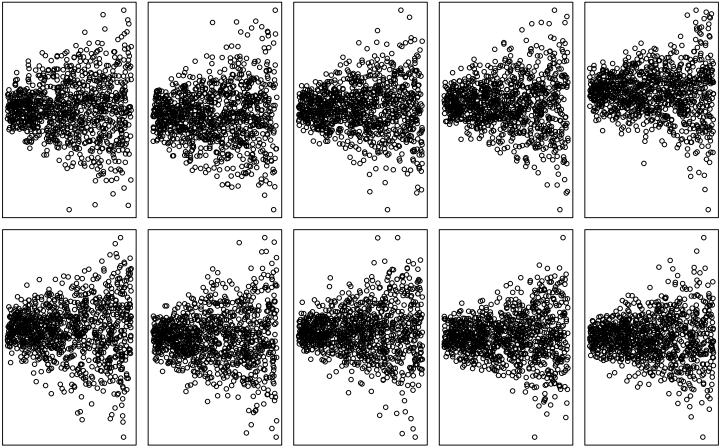
# Non-constant variance (n=10, $\sigma_2/\sigma_1 = 4$)



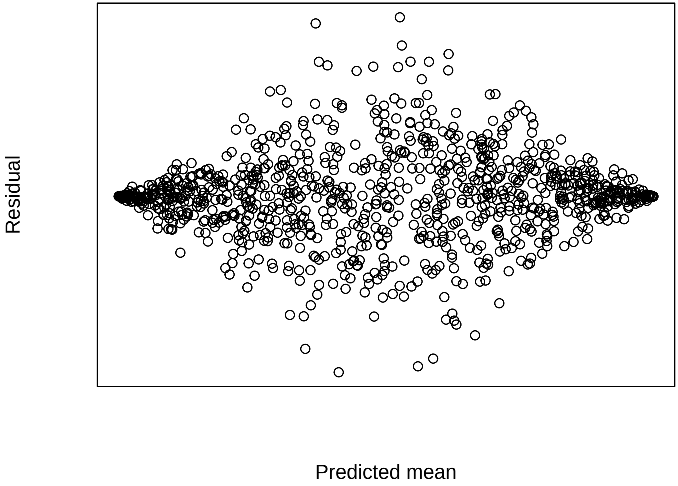
# Non-constant variance (n=100, $\sigma_2/\sigma_1 = 4$)



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



# Extreme non-constant variance (football)



# Independence

Lack of independence includes

- Cluster effect
- Serial correlation
- Spatial association

# Independence

Lack of independence includes

- Cluster effect
- Serial correlation
- Spatial association

Make plots of residuals vs relevant explanatory variable(s) and look for patterns,

# Independence

Lack of independence includes

- Cluster effect
- Serial correlation
- Spatial association

Make plots of residuals vs relevant explanatory variable(s) and look for patterns, e.g.

- Residuals vs groups

# Independence

Lack of independence includes

- Cluster effect
- Serial correlation
- Spatial association

Make plots of residuals vs relevant explanatory variable(s) and look for patterns, e.g.

- Residuals vs groups
- Residuals vs time (or observation number)

# Independence

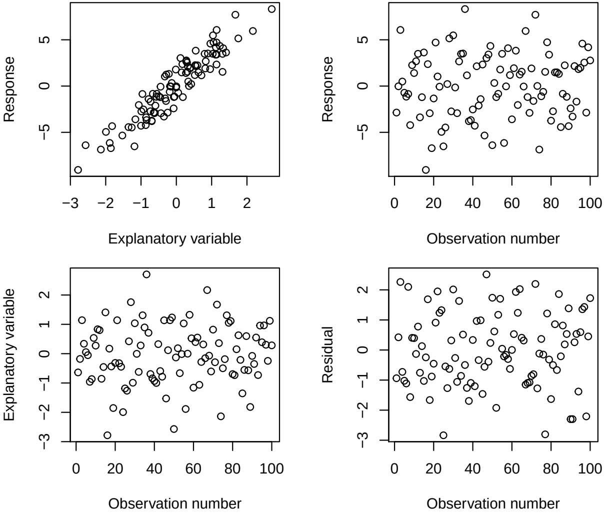
Lack of independence includes

- Cluster effect
- Serial correlation
- Spatial association

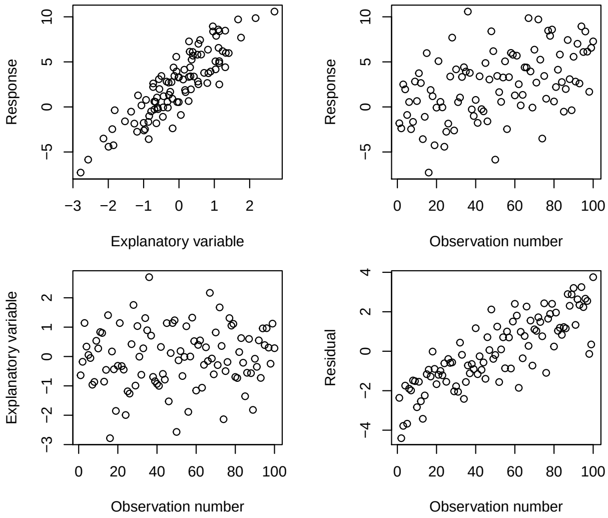
Make plots of residuals vs relevant explanatory variable(s) and look for patterns, e.g.

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

# No evidence for lack of independence

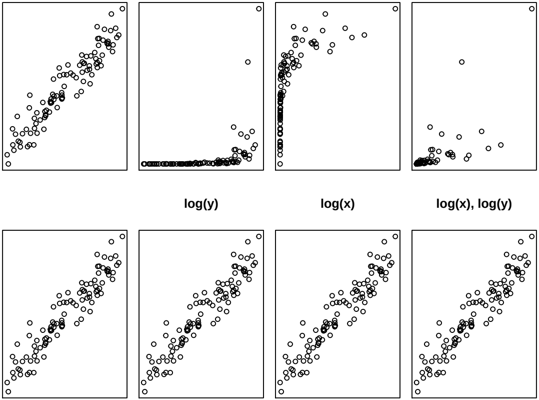


# Evidence for lack of independence



# Linearity

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



# Testing Composite hypotheses

Comparing two models

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

# 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

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

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 $\hat{\sigma}^2_{full}$

# 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 ($X_j$).

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

# 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:

| | | |
|---|---|---|
| ANOVA: | $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ | (full) |
| Regression: | $Y_{ij} \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_j, \sigma^2)$ | (reduced) |

# 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:

| | | |
|---|---|---|
| ANOVA: | $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ | (full) |
| Regression: | $Y_{ij} \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_j, \sigma^2)$ | (reduced) |

# 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:

| | | |
|---|---|---|
| ANOVA: | $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ | (full) |
| Regression: | $Y_{ij} \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_j, \sigma^2)$ | (reduced) |

- Regression model is reduced:

# 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:

| | | |
|---|---|---|
| ANOVA: | $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ | (full) |
| Regression: | $Y_{ij} \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_j, \sigma^2)$ | (reduced) |

- Regression model is reduced:
  - ANOVA has $J$ parameters for the mean

# 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:

| | | |
|---|---|---|
| ANOVA: | $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ | (full) |
| Regression: | $Y_{ij} \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_j, \sigma^2)$ | (reduced) |

- Regression model is reduced:
    - ANOVA has $J$ parameters for the mean
    - Regression has 2 parameters for the mean

# 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:

| | | |
|---|---|---|
| ANOVA: | $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ | (full) |
| Regression: | $Y_{ij} \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_j, \sigma^2)$ | (reduced) |

- Regression model is reduced:
    - ANOVA has $J$ parameters for the mean
    - Regression has 2 parameters for the mean
    - Set $\mu_j = \beta_0 + \beta_1 X_j$.

# 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:

| | | |
|---|---|---|
| ANOVA: | $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ | (full) |
| Regression: | $Y_{ij} \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_j, \sigma^2)$ | (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

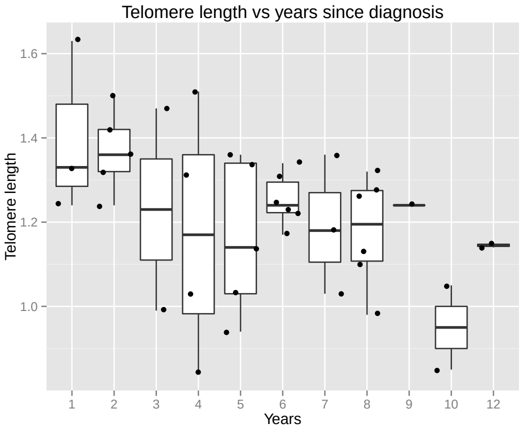
Let $Y_{ij}$ be the $i^{th}$ observation from the $j^{th}$ group where the group is defined by those observations having the same explanatory variable value ($X_j$).

Two models:

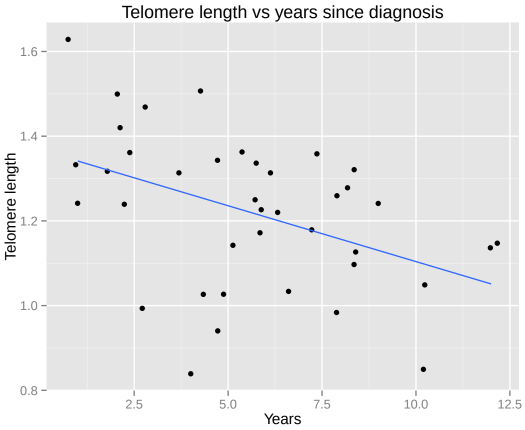
| | | |
|---|---|---|
| ANOVA: | $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ | (full) |
| Regression: | $Y_{ij} \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_j, \sigma^2)$ | (reduced) |

- 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

                                     Sum of           Mean
        Source               DF     Squares         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
```

```
# 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
```

```r
# 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

# Lack-of-fit F-test summary

- Lack-of-fit F-test tests the assumption of linearity
- Needs multiple observations at various explanatory variable values

# 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

# 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

# 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

# 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 variable(s)

# Summary of diagnostics

- Normality

# Summary of diagnostics

- Normality
  - Normal qq-plots should have points falling on the line

# Summary of diagnostics

- Normality
  - Normal qq-plots should have points falling on the line
- Constant variance

# Summary of diagnostics

- Normality
  - Normal qq-plots should have points falling on the line
- Constant variance
  - Residuals vs predicted values should have random scatter on y-axis

# Summary of diagnostics

- Normality
  - Normal qq-plots should have points falling on the line
- Constant variance
  - Residuals vs predicted values should have random scatter on y-axis
  - Residuals vs explanatory variable(s) should have random scatter on y-axis

# Summary of diagnostics

- Normality
  - Normal qq-plots should have points falling on the line
- Constant variance
  - Residuals vs predicted values should have random scatter on y-axis
  - Residuals vs explanatory variable(s) should have random scatter on y-axis
- Independence

# Summary of diagnostics

- Normality
  - Normal qq-plots should have points falling on the line
- Constant variance
  - Residuals vs predicted values should have random scatter on y-axis
  - Residuals vs explanatory variable(s) should have random scatter on y-axis
- Independence
  - Residuals vs anything should not have a pattern

# Summary of diagnostics

- Normality
  - Normal qq-plots should have points falling on the line
- Constant variance
  - Residuals vs predicted values should have random scatter on y-axis
  - Residuals vs explanatory variable(s) should have random scatter on y-axis
- Independence
  - Residuals vs anything should not have a pattern
- Linearity

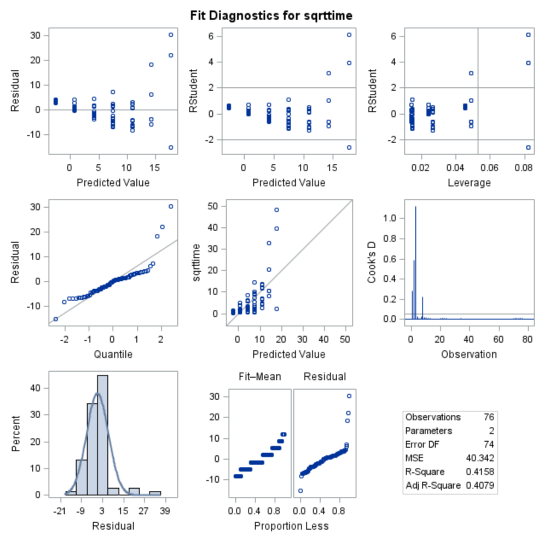
# Summary of diagnostics

- Normality
  - Normal qq-plots should have points falling on the line
- Constant variance
  - Residuals vs predicted values should have random scatter on y-axis
  - Residuals vs explanatory variable(s) should have random scatter on y-axis
- Independence
  - Residuals vs anything should not have a pattern
- Linearity
  - Response vs explanatory variable should be linear

# Summary of diagnostics

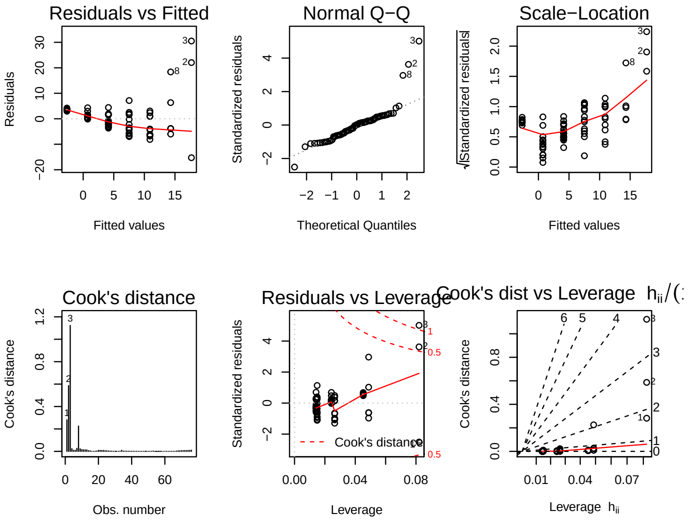
- Normality
  - Normal qq-plots should have points falling on the line
- Constant variance
  - Residuals vs predicted values should have random scatter on y-axis
  - Residuals vs explanatory variable(s) should have random scatter on y-axis
- Independence
  - Residuals vs anything should not have a pattern
- Linearity
  - Response vs explanatory variable should be linear
  - Lack-of-fit F-test should not be significant

# Default diagnostics in SAS




| | |
|---|---|
| Observations | 76 |
| Parameters | 2 |
| Error DF | 74 |
| MSE | 40.342 |
| R-Square | 0.4158 |
| Adj R-Square | 0.4079 |

# Default diagnostics in R



# Interpretations using logs

The most common transformation of either the response or explanatory variable(s) is to take logarithms

# Interpretations using logs

The most common transformation of either the response or explanatory variable(s) is to take logarithms because

- linearity will often then be approximately true,

# Interpretations using logs

The most common transformation of either the response or explanatory variable(s) is to take logarithms because

- linearity will often then be approximately true,
- the variance will likely be approximately constant, and

# Interpretations using logs

The most common transformation of either the response or explanatory variable(s) is to take logarithms because

- linearity will often then be approximately true,
- the variance will likely be approximately constant, and
- there is a (relatively) convenient interpretation.

# Interpretations using logs

The most common transformation of either the response or explanatory variable(s) is to take logarithms because

- linearity will often then be approximately true,
- the variance will likely be approximately constant, and
- there is a (relatively) convenient interpretation.

# Interpretations using logs

The most common transformation of either the response or explanatory variable(s) is to take logarithms because

- linearity will often then be approximately true,
- the variance will likely be approximately constant, and
- there is a (relatively) convenient interpretation.

We will talk about interpretation of $\beta_0$ and $\beta_1$ when

- only the response is logged,

# Interpretations using logs

The most common transformation of either the response or explanatory variable(s) is to take logarithms because

- linearity will often then be approximately true,
- the variance will likely be approximately constant, and
- there is a (relatively) convenient interpretation.

We will talk about interpretation of $\beta_0$ and $\beta_1$ when

- only the response is logged,
- only the explanatory variable is logged, and

# Interpretations using logs

The most common transformation of either the response or explanatory variable(s) is to take logarithms because

- linearity will often then be approximately true,
- the variance will likely be approximately constant, and
- there is a (relatively) convenient interpretation.

We will talk about interpretation of $\beta_0$ and $\beta_1$ when

- only the response is logged,
- only the explanatory variable is logged, and
- when both are logged.

# Neither response nor explanatory variable are logged

If

$$E[Y|X] = \beta_0 + \beta_1 X,$$

# Neither response nor explanatory variable are logged

If

$$E[Y|X] = \beta_0 + \beta_1 X,$$

then

- $\beta_0$ is the expected response when $X$ is zero and

# Neither response nor explanatory variable are logged

If

$$E[Y|X] = \beta_0 + \beta_1 X,$$

then

- $\beta_0$ is the expected response when $X$ is zero and
- $\beta_1$ is the expected change in the response for a one unit increase in the explanatory variable.

# Neither response nor explanatory variable are logged

If

$$E[Y|X] = \beta_0 + \beta_1 X,$$

then

- $\beta_0$ is the expected response when $X$ is zero and
- $\beta_1$ is the expected change in the response for a one unit increase in the explanatory variable.

# Neither response nor explanatory variable are logged

If

$$E[Y|X] = \beta_0 + \beta_1 X,$$

then

- $\beta_0$ is the expected response when $X$ is zero and
- $\beta_1$ is the expected change in the response for a one unit increase in the explanatory variable.

For the following discussion,

- $Y$ is always going to be the original response and

# Neither response nor explanatory variable are logged

If

$$E[Y|X] = \beta_0 + \beta_1 X,$$

then

- $\beta_0$ is the expected response when $X$ is zero and
- $\beta_1$ is the expected change in the response for a one unit increase in the explanatory variable.

For the following discussion,

- $Y$ is always going to be the original response and
- $X$ is always going to be the original explanatory variable.

# Example

Suppose

- $Y$ is corn yield per acre
- $X$ is fertilizer level in lbs/acre

# Example

Suppose

- $Y$ is corn yield per acre
- $X$ is fertilizer level in lbs/acre

Then, if

$$E[Y|X] = \beta_0 + \beta_1 X$$

- $\beta_0$ is the expected corn yield per acre when fertilizer level is zero and

# Example

Suppose

- $Y$ is corn yield per acre
- $X$ is fertilizer level in lbs/acre

Then, if

$$E[Y|X] = \beta_0 + \beta_1 X$$

- $\beta_0$ is the expected corn yield per acre when fertilizer level is zero and
- $\beta_1$ is the expected change in corn yield per acre when fertilizer is increase by 1 lbs/acre.

# Response is logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 X \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 X},$$

# Response is logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 X \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 X},$$

then

- $\beta_0$ is the expected $\log(Y)$ when $X$ is zero and

# Response is logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 X \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 X},$$

then

- $\beta_0$ is the expected $\log(Y)$ when $X$ is zero and
- $\beta_1$ is the expected change in $\log(Y)$ for a one unit increase in the explanatory variable.

# Response is logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 X \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 X},$$

then

- $\beta_0$ is the expected $\log(Y)$ when $X$ is zero and
- $\beta_1$ is the expected change in $\log(Y)$ for a one unit increase in the explanatory variable.

# Response is logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 X \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 X},$$

then

- $\beta_0$ is the expected $\log(Y)$ when $X$ is zero and
- $\beta_1$ is the expected change in $\log(Y)$ for a one unit increase in the explanatory variable.

Alternatively,

- $e^{\beta_0}$ is the median of $Y$ when $X$ is zero and

# Response is logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 X \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 X},$$

then

- $\beta_0$ is the expected $\log(Y)$ when $X$ is zero and
- $\beta_1$ is the expected change in $\log(Y)$ for a one unit increase in the explanatory variable.

Alternatively,

- $e^{\beta_0}$ is the median of $Y$ when $X$ is zero and
- $e^{\beta_1}$ is the multiplicative effect on the median of $Y$ for a one unit increase in the explanatory variable.

# Response is logged

Suppose

- $Y$ is corn yield per acre
- $X$ is fertilizer level in lbs/acre

# Response is logged

Suppose

- $Y$ is corn yield per acre
- $X$ is fertilizer level in lbs/acre

Then, if

$$E[\log(Y)|X] = \beta_0 + \beta_1 X \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 X}$$

- $e^{\beta_0}$ is the median corn yield per acre when fertilizer level is 0 and

# Response is logged

Suppose

- $Y$ is corn yield per acre
- $X$ is fertilizer level in lbs/acre

Then, if

$$E[\log(Y)|X] = \beta_0 + \beta_1 X \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 X}$$

- $e^{\beta_0}$ is the median corn yield per acre when fertilizer level is 0 and
- $e^{\beta_1}$ is the multiplicative effect in median corn yield per acre when fertilizer is increase by 1 lbs/acre.

# Explanatory variable is logged

If

$$E[Y|X] = \beta_0 + \beta_1 \log(X),$$

# Explanatory variable is logged

If

$$E[Y|X] = \beta_0 + \beta_1 \log(X),$$

then

- $\beta_0$ is the expected response when $\log(X)$ is zero and

# Explanatory variable is logged

If

$$E[Y|X] = \beta_0 + \beta_1 \log(X),$$

then

- $\beta_0$ is the expected response when $\log(X)$ is zero and
- $\beta_1$ is the expected change in the response for a one unit increase in $\log(X)$.

# Explanatory variable is logged

If

$$E[Y|X] = \beta_0 + \beta_1 \log(X),$$

then

- $\beta_0$ is the expected response when $\log(X)$ is zero and
- $\beta_1$ is the expected change in the response for a one unit increase in $\log(X)$.

# Explanatory variable is logged

If

$$E[Y|X] = \beta_0 + \beta_1 \log(X),$$

then

- $\beta_0$ is the expected response when $\log(X)$ is zero and
- $\beta_1$ is the expected change in the response for a one unit increase in $\log(X)$.

Alternatively,

- $\beta_0$ is the expected response when $X$ is 1 and

# Explanatory variable is logged

If

$$E[Y|X] = \beta_0 + \beta_1 \log(X),$$

then

- $\beta_0$ is the expected response when $\log(X)$ is zero and
- $\beta_1$ is the expected change in the response for a one unit increase in $\log(X)$.

Alternatively,

- $\beta_0$ is the expected response when $X$ is 1 and
- $\beta_1 \log(d)$ is the expected change in the response when $X$ increase multiplicatively by $d$, e.g.

# Explanatory variable is logged

If

$$E[Y|X] = \beta_0 + \beta_1 \log(X),$$

then

- $\beta_0$ is the expected response when $\log(X)$ is zero and
- $\beta_1$ is the expected change in the response for a one unit increase in $\log(X)$.

Alternatively,

- $\beta_0$ is the expected response when $X$ is 1 and
- $\beta_1 \log(d)$ is the expected change in the response when $X$ increase multiplicatively by $d$, e.g.
    - $\beta_1 \log(2)$ is the expected change in the response for each doubling of $X$ or

# Explanatory variable is logged

If

$$E[Y|X] = \beta_0 + \beta_1 \log(X),$$

then

- $\beta_0$ is the expected response when $\log(X)$ is zero and
- $\beta_1$ is the expected change in the response for a one unit increase in $\log(X)$.

Alternatively,

- $\beta_0$ is the expected response when $X$ is 1 and
- $\beta_1 \log(d)$ is the expected change in the response when $X$ increase multiplicatively by $d$, e.g.
  - $\beta_1 \log(2)$ is the expected change in the response for each doubling of $X$ or
  - $\beta_1 \log(10)$ is the expected change in the response for each ten-fold increase in $X$.

# Explanatory variable is logged

Suppose

- $Y$ is corn yield per acre
- $X$ is fertilizer level in lbs/acre

# Explanatory variable is logged

Suppose

- $Y$ is corn yield per acre
- $X$ is fertilizer level in lbs/acre

Then, if

$$E[Y|X] = \beta_0 + \beta_1 \log(X)$$

- $\beta_0$ is the expected corn yield per acre when fertilizer level is 1 lb/acre and

# Explanatory variable is logged

Suppose

- $Y$ is corn yield per acre
- $X$ is fertilizer level in lbs/acre

Then, if

$$E[Y|X] = \beta_0 + \beta_1 \log(X)$$

- $\beta_0$ is the expected corn yield per acre when fertilizer level is 1 lb/acre and
- $\beta_1 \log(2)$ is the expected change in corn yield when fertilizer level is doubled.

# Both response and explanatory variable are logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X) \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 \log(X)} = e^{\beta_0} X^{\beta_1},$$

# Both response and explanatory variable are logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X) \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 \log(X)} = e^{\beta_0} X^{\beta_1},$$

then

- $\beta_0$ is the expected $\log(Y)$ when $\log(X)$ is zero and

# Both response and explanatory variable are logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X) \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 \log(X)} = e^{\beta_0} X^{\beta_1},$$

then

- $\beta_0$ is the expected $\log(Y)$ when $\log(X)$ is zero and
- $\beta_1$ is the expected change in $\log(Y)$ for a one unit increase in $\log(X)$.

# Both response and explanatory variable are logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X) \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 \log(X)} = e^{\beta_0} X^{\beta_1},$$

then

- $\beta_0$ is the expected $\log(Y)$ when $\log(X)$ is zero and
- $\beta_1$ is the expected change in $\log(Y)$ for a one unit increase in $\log(X)$.

# Both response and explanatory variable are logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X) \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 \log(X)} = e^{\beta_0} X^{\beta_1},$$

then

- $\beta_0$ is the expected $\log(Y)$ when $\log(X)$ is zero and
- $\beta_1$ is the expected change in $\log(Y)$ for a one unit increase in $\log(X)$.

Alternatively,

- $e^{\beta_0}$ is the median of $Y$ when $X$ is 1 and

# Both response and explanatory variable are logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X) \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 \log(X)} = e^{\beta_0} X^{\beta_1},$$

then

- $\beta_0$ is the expected $\log(Y)$ when $\log(X)$ is zero and
- $\beta_1$ is the expected change in $\log(Y)$ for a one unit increase in $\log(X)$.

Alternatively,

- $e^{\beta_0}$ is the median of $Y$ when $X$ is 1 and
- $d^{\beta_1}$ is the multiplicative change in the median of the response when $X$ increase multiplicatively by $d$, e.g.

# Both response and explanatory variable are logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X) \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 \log(X)} = e^{\beta_0} X^{\beta_1},$$

then

- $\beta_0$ is the expected $\log(Y)$ when $\log(X)$ is zero and
- $\beta_1$ is the expected change in $\log(Y)$ for a one unit increase in $\log(X)$.

Alternatively,

- $e^{\beta_0}$ is the median of $Y$ when $X$ is 1 and
- $d^{\beta_1}$ is the multiplicative change in the median of the response when $X$ increase multiplicatively by $d$, e.g.
  - $2^{\beta_1}$ is the multiplicative effect on the median of the response for each doubling of $X$ or

# Both response and explanatory variable are logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X) \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 \log(X)} = e^{\beta_0} X^{\beta_1},$$

then

- $\beta_0$ is the expected $\log(Y)$ when $\log(X)$ is zero and
- $\beta_1$ is the expected change in $\log(Y)$ for a one unit increase in $\log(X)$.

Alternatively,

- $e^{\beta_0}$ is the median of $Y$ when $X$ is 1 and
- $d^{\beta_1}$ is the multiplicative change in the median of the response when $X$ increase multiplicatively by $d$, e.g.
  - $2^{\beta_1}$ is the multiplicative effect on the median of the response for each doubling of $X$ or
  - $10^{\beta_1}$ is the multiplicative effect on the median of the response for each ten-fold increase in $X$.

# Both response and explanatory variable are logged

Suppose

- $Y$ is corn yield per acre
- $X$ is fertilizer level in lbs/acre

# Both response and explanatory variable are logged

Suppose

- $Y$ is corn yield per acre
- $X$ is fertilizer level in lbs/acre

Then, if

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X) \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 \log(X)} = e^{\beta_0} X^{\beta_1}$$

- $e^{\beta_0}$ is the median corn yield per acre when fertilizer level is 1 lb/acre and

# Both response and explanatory variable are logged

Suppose

- $Y$ is corn yield per acre
- $X$ is fertilizer level in lbs/acre

Then, if

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X) \quad \text{or} \quad Median\{Y|X\} = e^{\beta_0} e^{\beta_1 \log(X)} = e^{\beta_0} X^{\beta_1}$$

- $e^{\beta_0}$ is the median corn yield per acre when fertilizer level is 1 lb/acre and
- $2^{\beta_1}$ is the multiplicative effect on median corn yield per acre when fertilizer level doubles.

# Summary of interpretations when using logarithms

- When using the log of the response,

# Summary of interpretations when using logarithms

- When using the log of the response,
  - $\beta_0$ will affect the median response

# Summary of interpretations when using logarithms

- When using the log of the response,
  - $\beta_0$ will affect the median response
  - $\beta_1$ will affect the multiplicative change in the median response

# Summary of interpretations when using logarithms

- When using the log of the response,
    - $\beta_0$ will affect the median response
    - $\beta_1$ will affect the multiplicative change in the median response
- When using the log of the explanatory variable ($X$),

# Summary of interpretations when using logarithms

- When using the log of the response,
  - $\beta_0$ will affect the median response
  - $\beta_1$ will affect the multiplicative change in the median response
- When using the log of the explanatory variable ($X$),
  - $\beta_0$ will affect the response when $X = 1$

# Summary of interpretations when using logarithms

- When using the log of the response,
  - $\beta_0$ will affect the median response
  - $\beta_1$ will affect the multiplicative change in the median response
- When using the log of the explanatory variable ($X$),
  - $\beta_0$ will affect the response when $X = 1$
  - $\beta_1$ will affect the change in the response when there is a multiplicative change in $X$

# Summary of interpretations when using logarithms

- When using the log of the response,
  - $\beta_0$ will affect the median response
  - $\beta_1$ will affect the multiplicative change in the median response
- When using the log of the explanatory variable ($X$),
  - $\beta_0$ will affect the response when $X = 1$
  - $\beta_1$ will affect the change in the response when there is a multiplicative change in $X$

# Summary of interpretations when using logarithms

- When using the log of the response,
  - $\beta_0$ will affect the median response
  - $\beta_1$ will affect the multiplicative change in the median response
- When using the log of the explanatory variable ($X$),
  - $\beta_0$ will affect the response when $X = 1$
  - $\beta_1$ will affect the change in the response when there is a multiplicative change in $X$

To construct confidence intervals for $e^\beta$, find a confidence interval for $\beta$ and exponentiate the endpoints, i.e. if $(L, U)$ is a confidence interval for $\beta$, then $(e^L, e^U)$ is a confidence interval for $e^\beta$.