Universidade de Brasília

2023

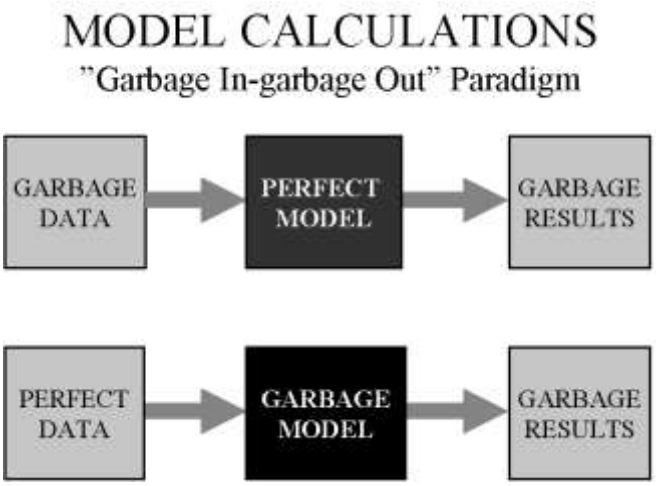
Experimentação 4 – Protocolo para Exploração de dados

Concepta McManus, Sonia Nair Báo, José Jivago Rolo, Andrea Queiroz Maranhão, Daniel Pimentel, Felipe Pimentel

1/1/2023

Sumário

Protocolo para exploração de dados



Exploração de dados

- Até 50% do tempo que você gasta com a análise em si
- Modelos devem ser decido a *priori*
- Usa gráficos

Quais são as pressuposições?

1. Ausência de observações atípicas;
2. Independência dos resíduos;
3. Aditividade dos efeitos do modelo;
4. Homogeneidade de variância dos resíduos para os tratamentos;
5. Normalidade dos resíduos;
6. Todas as amostras são retiradas de populações normalmente distribuídas
7. Todas as amostras são retiradas independentemente umas das outras
8. Dentro de cada amostra, as observações são amostradas aleatoriamente e independentemente umas das outras

Ordem de Importância

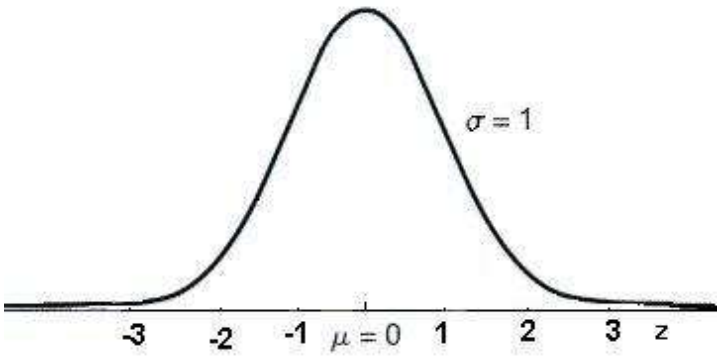
ANOVA

- Homoscedasticity
- Normality
- Additivity
- Independence

Regressão

- Additivity
- Homoscedasticity
- Normality
- Independence

Curva Normal



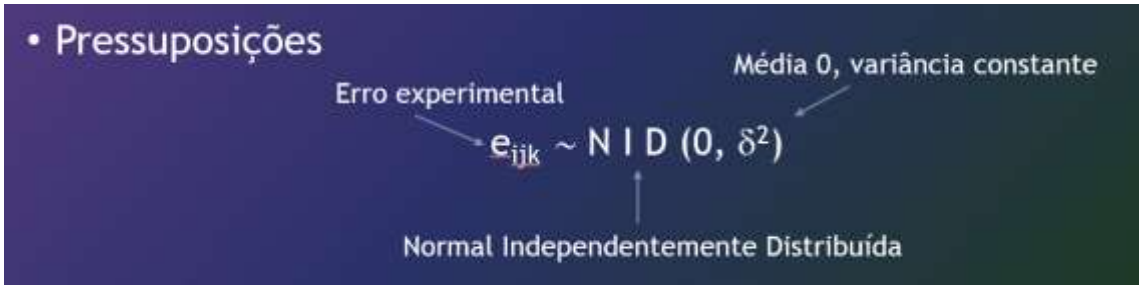
Características da Curva Normal:

- A distribuição normal é determinada por dois parâmetros:
  - Média da população
  - Desvio padrão da população
- A distribuição é simétrica em relação à média.
- Os valores de média, moda e mediana são iguais.

- A área total sob a curva é igual a 1 (100%)
- exatos 50% distribuídos à esquerda da média e 50% à sua direita

ANOVA

- Todas as amostras são retiradas de populações normalmente distribuídas
- Todas as populações têm uma variação comum
- Todas as amostras são retiradas independentemente umas das outras
- Dentro de cada amostra, as observações são amostradas aleatoriamente e independentemente umas das outras
- Os efeitos do fator são aditivos

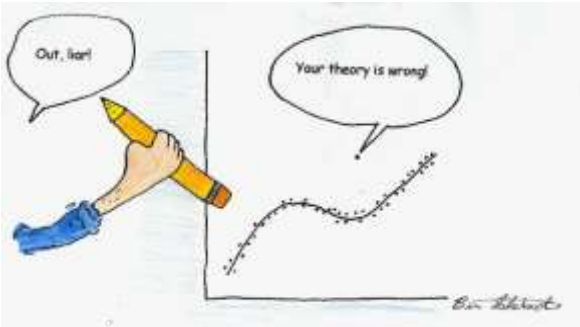


Problemas e Soluções

- Outliers Y & X
  - Boxplot e Cleveland dotplot
- Homogeneity Y
  - Conditional boxplot
- Normality Y
  - Histogram ou Qqplot
- Zero trouble Y
  - Frequency plot ou correlogram
- Collinearity X
  - VIF & scatterplots
  - Correlations & PCA
- Relationships Y & X
  - (multi panel) scatterplots

    - Conditional boxplots
- Interactions
    - Complots
- Independence Y
    - ACF & variogram
    - Plot Y vs tempo/espaço

Etapa 1. Tem outliers – observações atípicas?



Check raw data for errors

- Outlier – observation that has a relatively large/small value compared to the majority
- p.ex. Bill Gates ganha US $ 500 milhões por ano. Ele está em uma sala com 9 professores, 4 dos quais ganham $ 40k, 3 ganham $ 45k e 2 ganham $ 55k por ano. Qual é o salário médio de todos na sala? Qual seria o salário médio se Gates não fosse incluído?
    - Mean <u>With</u> Gates: $50,040,500
    - Mean <u>Without</u> Gates: $45,000

- Most likely explanation
    - Measurement (observer) errors
    - Drop the observation
- Assess whether the values are reasonable
    - Do these values affect the analyses?
- Transformation?

Métodos para identificação

- 2 SD Method: $x \pm 2$ SD
- 3 SD Method: $x \pm 3$ SD,

- where the mean is the sample mean and SD is the sample standard deviation.
- Z score
- Z-scores > 3 in absolute value - outliers

$$Z_i = \frac{x_i - \bar{x}}{sd},$$

Para encontrar quaisquer outliers em um conjunto de dados, precisamos encontrar o

Resumo de 5 números dos dados.

Encontre a média e a mediana do seguinte conjunto de números

3 12 7 40 9 14 18 15 17

Mean is 15

Median is 14

Etapa 1: classificar os números do menor para o maior

3 7 9 12 14 15 17 18 40

Etapa 2 Identificar a Mediana

3 7 9 12 14 15 17 18 40

Etapa 3: Identificar os números maiores e menores

3 7 9 12 14 15 17 18 40

Etapa 4: Identifique a mediana entre o menor número e a mediana para todo o conjunto de dados, e entre essa mediana e o maior número no conjunto

3 7 9 12 14 15 17 18 40

Estes são os cinco números no Resumo dos 5 números

3 - Menor número do conjunto

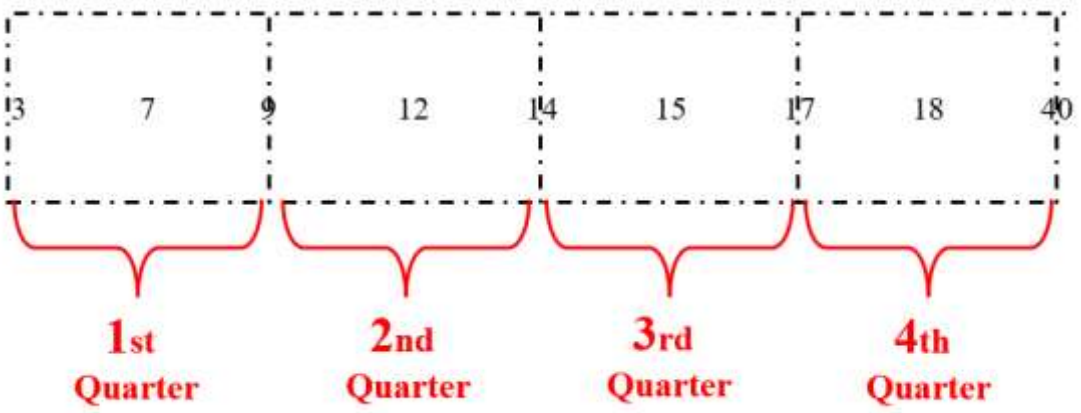
9 - Mediana entre o menor número e a mediana

14 - Mediana de todo o conjunto

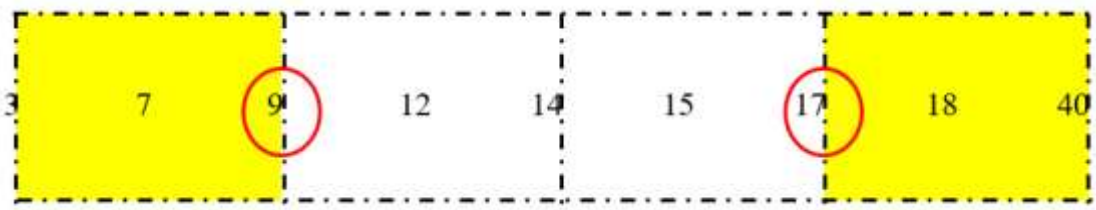
17 - Mediana entre o maior número e a mediana

40 - Maior número do conjunto

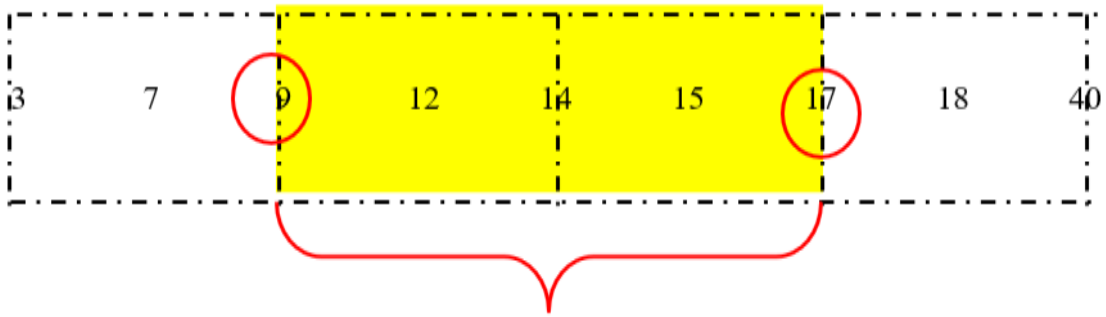
Um resumo de 5 números divide seus dados em quatro trimestres



- O quartil inferior (Q1) é o segundo número no resumo de 5 números
- 25% de todos os números do conjunto são menores do que Q1

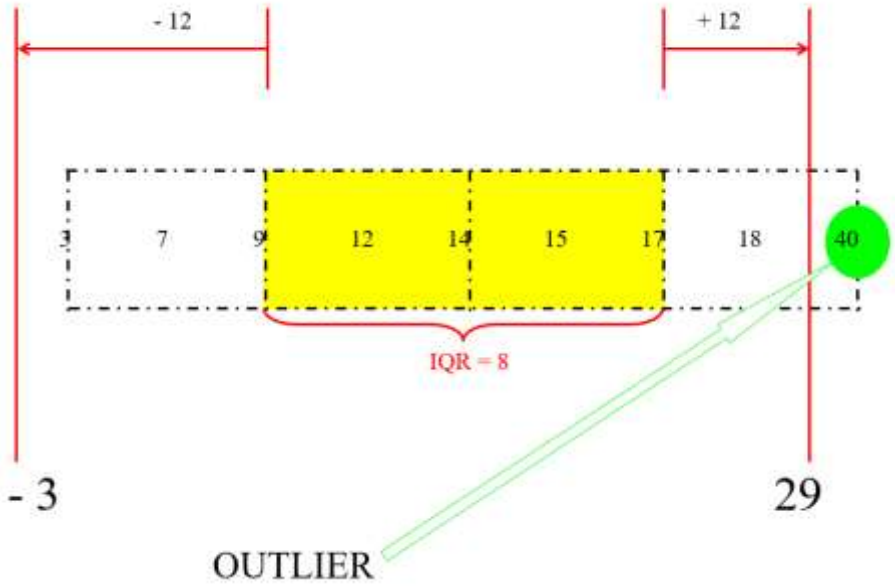


- O Quartil superior (Q3) é o quarto número no Resumo de 5 números
- 25% de todos os números do conjunto são maiores do que Q3
- 50% dos números estão entre Q1 e Q3



Isso é chamado de intervalo interquartil (IQR)

- O tamanho do IQR é a distância entre Q1 e Q3 = 17-9=8
- Para determinar se um número é um outlier multiplicar IQR * 1,5
- 8*1,5 = 12
- Um outlier é qualquer número 12 menor que Q1 ou 12 maior que Q3



Encontre a média e a mediana do seguinte conjunto de números

3 12 7 ~~40~~ 9 14 18 15 17

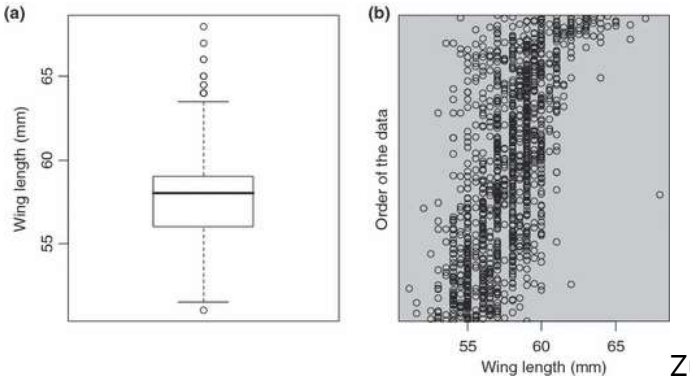
Antes **Depois**

Média é 15 **Média é 11,875**

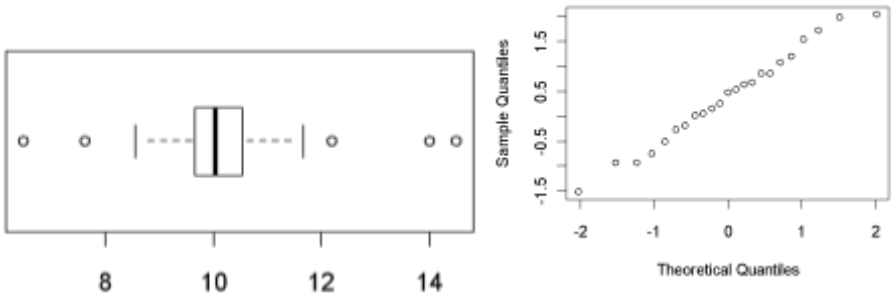
Mediana é 14 **Mediana é 13**

Outros Métodos

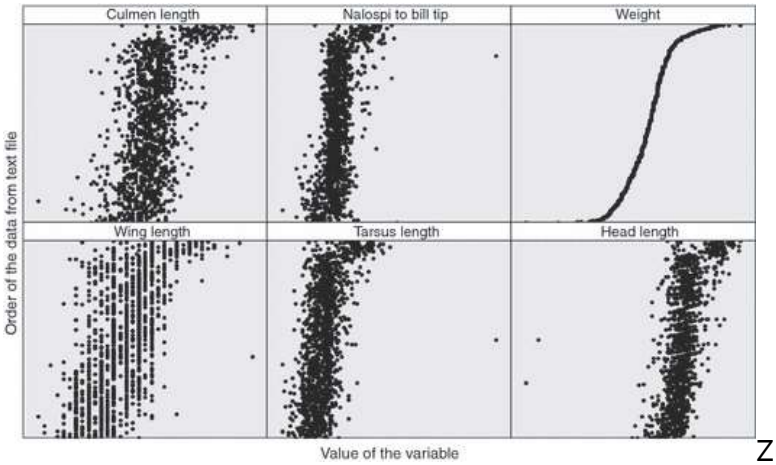
- Boxplot, Normal plot, Resíduos versus Preditos, and Cleveland plot



Zuur et al. (2008)

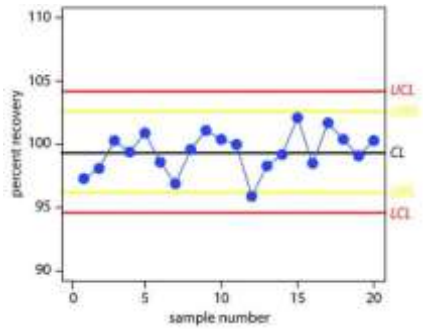


Multi-panel Cleveland plot



Zuur et al. (200*)

Control chart



- Os dados podem ser normalizados se houver discrepâncias moderada
- winsorização, "trim", corte ou transformação.
- Geralmente, outliers graves devem ser excluídos dos dados para atingir a normalidade.
- Esses procedimentos são "legais", desde que:
    - (1) eles são exercitados criteriosamente
    - (2) nunca usado para ajustar um valor P

Cuidado:

- Nunca exclua mais de 5% dos seus dados.
- Valores discrepantes graves podem estar legitimamente dentro dessa faixa.

- No entanto, se houver mais de 5% de discrepâncias graves, geralmente algo mais está acontecendo.

Winsorização

- Normalmente, mas não necessariamente, executado de forma simétrica.
    - Classifique os dados e dê aos extremos o mesmo valor que a classificação adjacente.
    - Recompute estatísticas e teste de normalidade.
- Exemplo:
    - 1, 2, 3, 4, 5, 7, 18
    - Média = 5,7
    - Normalidade M-I: rejeitar
    - 2, 2, 3, 4, 5, 7, 7
    - Média = 4,3
    - Normalidade M-I: aceitar
- mais apropriado quando os tamanhos das amostras são pequenos e você precisa proteger sua energia.

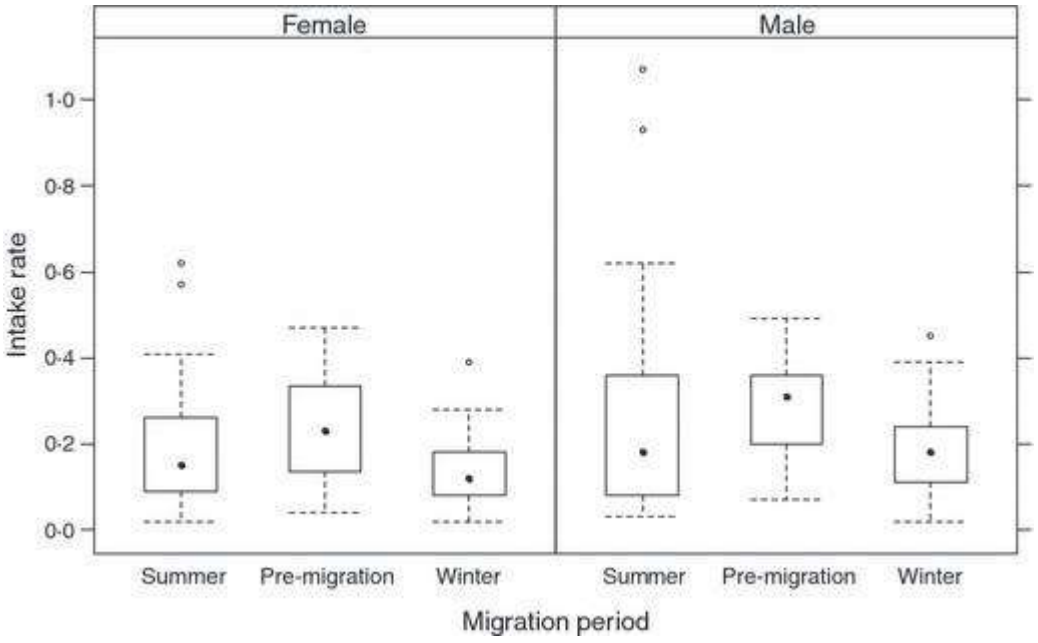
Trim

- Como alternativa, os dados podem ser cortados das caudas.
    - Normalmente, solte Xmin e Xmax
    - Isso reduz N e pode afetar a potência.
- Exemplo:
    - 1, 2, 3, 4, 5, 7, 18
    - Média = 5,7
    - Normalidade M-I: rejeitar
    - 2, 3, 4, 5, 7
    - Média = 4,2
    - Normalidade M-I: aceitar

Step 2 – Homogeneity of variance?

- Important in
    - ANOVA
    - Regression models
    - Discriminant analysis
- Use residuals
- Solution
- Transformation to stabilize the variance of response variable
- Use statistical tests that don´t require homogeneity (GLM)

Conditional Box plots



Plot residual vs fitted values

Hartley-Fisher´s Fmax

Levene's test

Cochran

Fligner Killeen test

Bartlett's test

- tende a *mascarar diferenças que existem* quando a curtose é negativa
- *achar diferenças que não existem* quando a curtose é positiva

Homogeneidade de variâncias dos resíduos

Formulação das hipóteses

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_I^2$$

$$H_a : pelo\ menos\ dois\ \sigma_i^2 's\ diferem\ entre\ si\ (i = 1,..., I).$$

Levene:

Resíduos ordinários (e_{ij})

$$e_{ij} = y_{ij} - \bar{y}_i$$

- Faça uma análise de variância com um critério de classificação (one way layout) dos quadrados desses resíduos.

$$(e_{ij})^2 = (y_{ij} - \bar{y}_i)^2$$

- Lógica
  - Quanto maiores são os quadrados dos resíduos, maiores são as variâncias.

- Se as variâncias são homogêneas, o resultado do teste *F* para comparar as médias dos quadrados dos resíduos será não significante.

Dados

Grupo					
A	B	C	D	E	Controle
25	10	18	23	11	8
17	-2	8	29	23	-6
27	12	4	25	5	6
21	4	14	35	17	0
15	16	6	33	9	2
21	8	10	29	13	2

Quadrados dos Resíduos

Grupo					
A	B	C	D	E	Controle
16	4	64	36	4	36
16	100	4	0	100	64
36	16	36	16	64	16
0	16	16	36	16	4
36	64	16	16	16	0

Resíduos padronizados (d_{ij})

$$d_{ij} = \frac{e_{ij}}{\sqrt{QMErro}}$$

Teste de Hartley-Fisher ou F_{max}

QUADRO 3.2.1 – Números de pulgões coletados 36 horas após a pulverização.

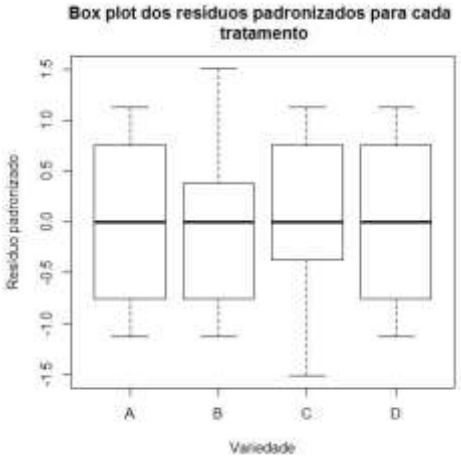
TRATAMENTOS	REPETIÇÕES						s^2
	1	2	3	4	5	6	
A	2.370	1.687	2.592	2.283	2.910	3.020	233.750
B	1.282	1.527	871	1.025	825	920	75.559
C	562	321	636	317	485	842	40.126
D	173	127	132	150	129	227	1.502
E	193	71	82	62	96	44	2.792

Pelo quadro, vemos que as variâncias parecem não ser homogêneas e, para verificar a hipóte
da homogeneidade de variâncias, aplicamos o teste de Hartley:

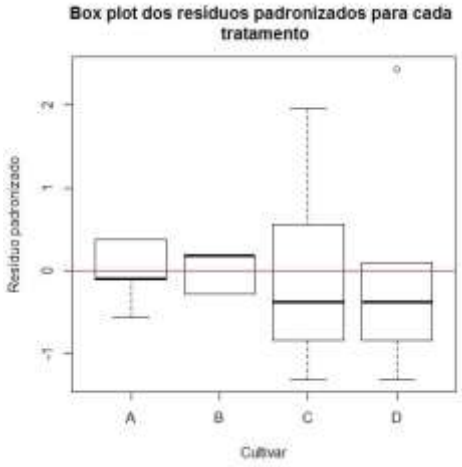
$$H_c = \frac{s^2_{max}}{s^2_{min}} = \frac{233.750}{1.502} = 155{,}63^{**}$$

Box plot dos resíduos padronizados;

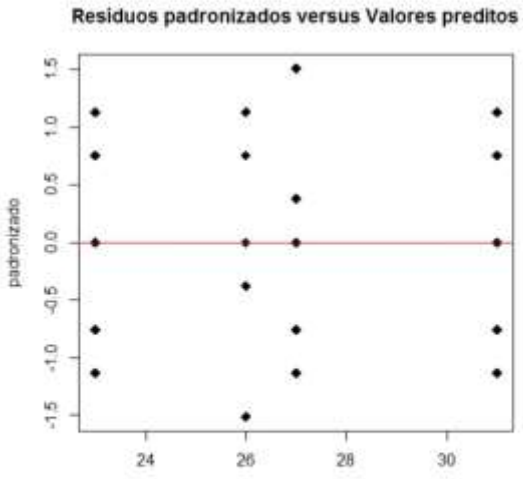
Resíduos padronizados versus valores preditos ($\hat{y}_i$).



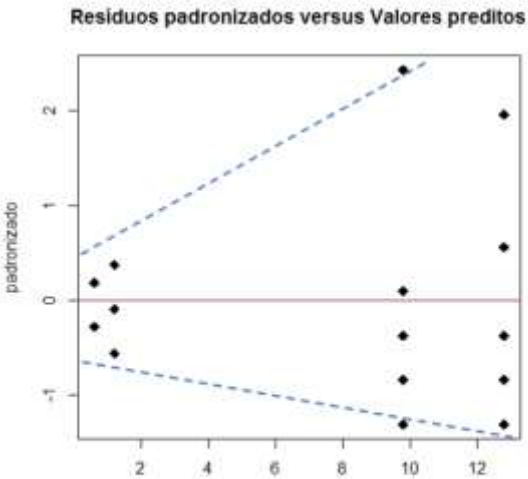
Variâncias **homogêneas**
(amplitudes semelhantes)



Variâncias **heterogêneas**
(amplitudes desiguais)

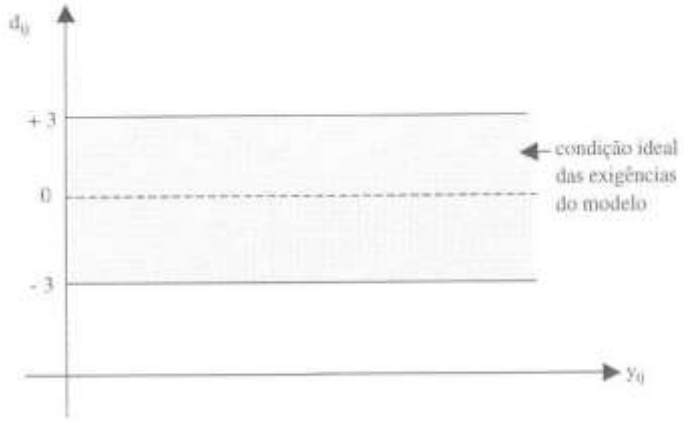


Variâncias homogêneas
(aleatório em torno do zero)

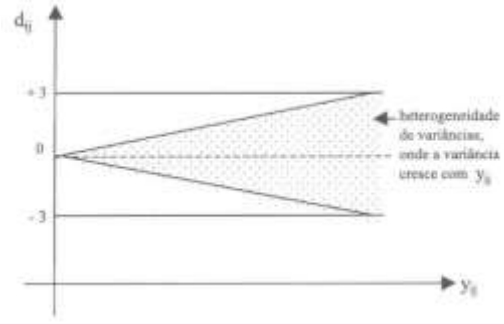


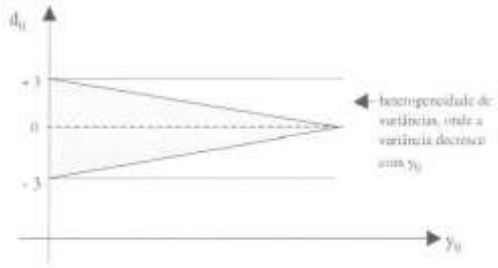
Variâncias **heterogêneas**
(variabilidade aumenta com os preditos)

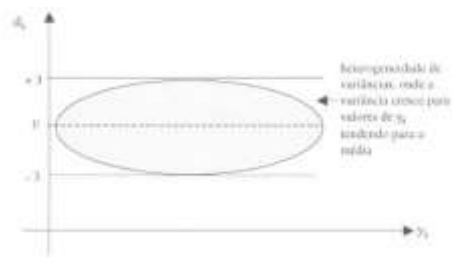
Padrão que indica homogeneidade

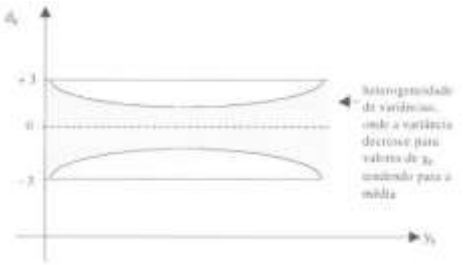


Padrões que indicam heterogeneidade









O que fazer?

- Não-paramétrico:
  - Kruskal-Wallis
- Testes Modificados:
  - Brown-Forsythe and Welch's ANOVA test
- Transformações (em baixo)

Homoscedasticity

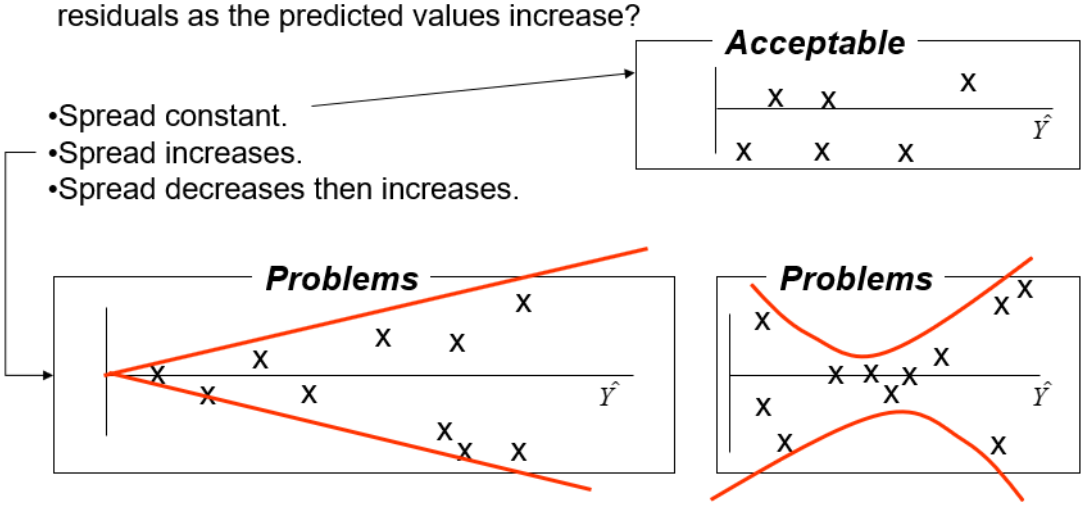
How do I know I have a problem?

$\hat{y}_i$ 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.



Step 3 - Are the data normally distributed?

- **Graphical methods**
  - Histogram (data and residuals)
- **Back-of-the-envelope test**
  - z-score
- **Frequentist tests**
  - D'Agostino's K-squared test,
  - Jarque–Bera test,
  - Anderson–Darling test,
  - Cramér–von Mises criterion,
  - Lilliefors test for normality (itself an adaptation of the Kolmogorov–Smirnov test),
  - Shapiro–Wilk test
  - Pearson's chi-squared test.
  - Skewness
  - Kurtosis

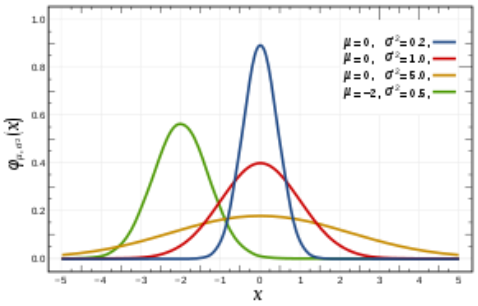
Normalidade

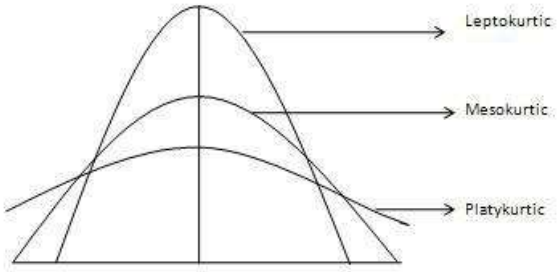
- Não é problema a menos que o número de amostras é grande
- Independência
  - implica que não existe relação entre o tamanho do erro e o agrupamento
- Parcelas adjacentes são mais relacionados de que parcelas aleatórios
  - O melhor segurança é casualização

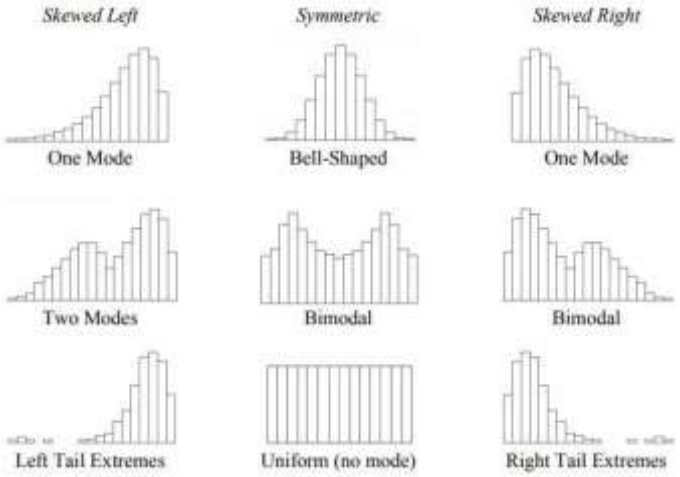
Teste de normalidade e simetria

- Revise a distribuição graficamente (por meio de histogramas, box plots, gráficos QQ)
- Análise a assimetria e curtose
    - Devem estar perto de 1
- Empregue testes estatísticos (especialmente Qui-quadrado, Kolmogorov-Smironov, Shapiro-Wilk)

Forma da Curva







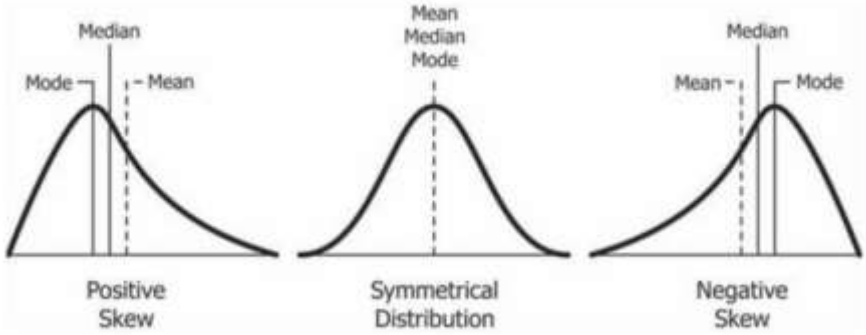
- A assimetria
    - medida da distribuição de probabilidade assumindo uma distribuição unimodal
    - terceiro momento padronizado.

- indica o quanto nossa distribuição se desvia da distribuição normal
- a distribuição normal tem assimetria o.

$$skewness = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{(N-1)s^3}$$

where:

- σ is the standard deviation
- $\bar{x}$ is the mean of the distribution
- N is the number of observations of the sample



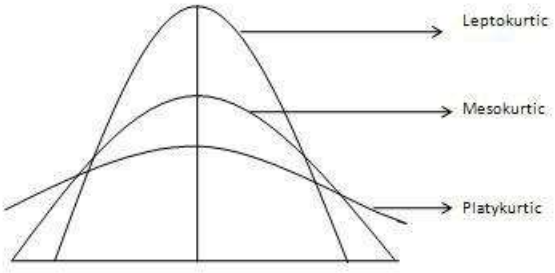
- **Simétrico**
  - assimetria é próxima de o e a **média** é quase igual à **mediana**
- **Inclinação negativa**
  - a cauda esquerda do histograma da distribuição é mais longa
  - a maioria das observações está concentrada na cauda direita.
  - "enviesado para a esquerda" ou "caudal para a esquerda" e a mediana é maior que a média.
- **Inclinação positiva**
  - quando a cauda direita do histograma da distribuição é mais longa
  - maioria das observações está concentrada na cauda esquerda

- "enviesado à direita" ou "caudal à direita" e a mediana é menor que a média.

CURTOSE

- Descrever a "cauda" da distribuição
  - média dos dados padronizados elevados à quarta potência
  - descreve a forma
  - medida do "pico" da distribuição
  - curtose alta tem um pico mais nítido e caudas mais longas e mais grossas,
  - curtose baixa tem um amendoim mais arredondado e caudas mais curtas e finas
- Valores <1
  - dados dentro de um desvio padrão da média, onde o "pico" seria
  - contribuem com praticamente nada para a curtose
    - aumentar um número menor que 1 à quarta potência o torna mais perto de zero
  - Únicos valores de dados que contribuem para a curtose de forma significativa
    - aqueles fora da região do pico

- outliers.
- **Portanto, a curtose mede apenas valores discrepantes; não mede nada sobre o "pico"** .



- **Mesocúrtico**
  - Esta é a distribuição normal
- **Leptocúrtico**
  - distribuição tem caudas mais **grossas** e um pico mais nítido.
  - curtose é "positiva" com um valor maior que **3**
- **Platicúrtico**
  - distribuição tem um pico mais baixo e mais largo e caudas mais finas.
  - curtose é "negativa" com um valor inferior a **3**

$$kurtosis = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)s^4}$$

where:

- σ is the standard deviation
- $\bar{x}$ is the mean of the distribution
- N is the number of observations of the sample

Interpretação

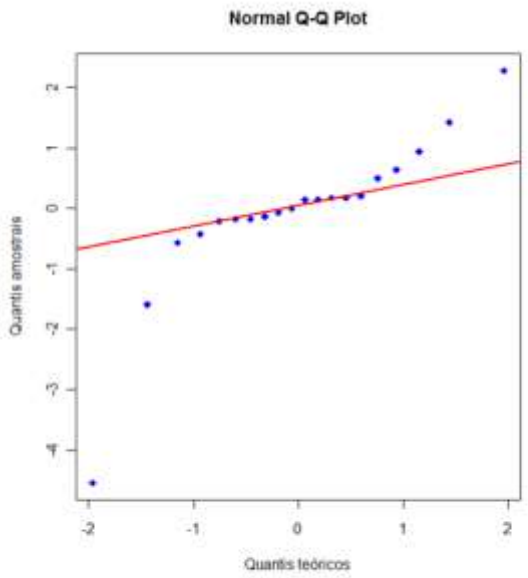
- **Simétrico**

    - Valores entre -0,5 a 0,5
- **Dados distorcidos moderados**
    - Valores entre **-1 e -0,5** ou entre **0,5 e 1**
- **Dados altamente distorcidos**
    - Valores **menores que -1** ou **maiores que 1**
- Sabemos que a distribuição **normal** é simétrica.

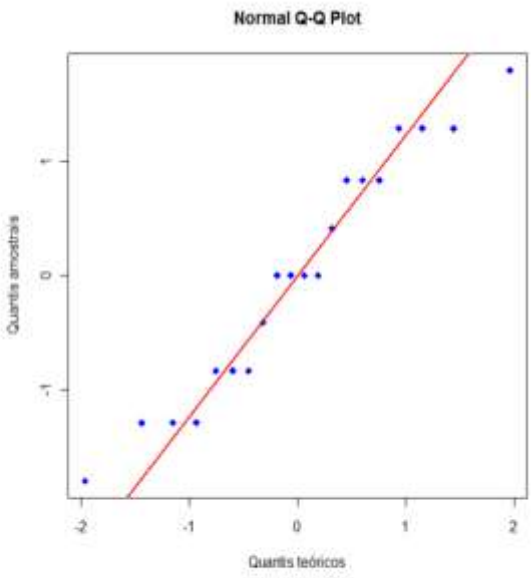
Testes

- Chi-square Test
- Kolmogorov-Smirnov (KS) Test
- Lilliefors Test
- Shapiro-Wilk (SW) Test
    - correlation between the raw data and the values that would be expected if the observations followed a normal distribution

Normalidade dos resíduos



Não Normal
(afastamento da reta)



Normal
(proximidade da reta)

Formulação das hipóteses

$$H_0 : os\ resíduos\ seguem\ uma\ distribuição\ Normal$$
$$H_a : os\ resíduos\ não\ seguem\ uma\ distribuição\ Normal$$

- O teste de Shapiro-Wilk é baseado na estatística W ($0 < W \leq 1$)
- Valores pequenos da estatística W levam a rejeitar a hipótese H_0
- Nos softwares R e SAS avaliamos o valor da probabilidade (valor p)
- Se o valor da probabilidade for menor que o nível de significância (α) rejeitamos a hipótese H_0

Tests for Normality

Test		--Statistic---		-----p Value------
Shapiro-Wilk	W	0.939623	Pr < W	0.2359
Kolmogorov-Smirnov	D	0.14496	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.07135	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.464289	Pr > A-Sq	0.2343

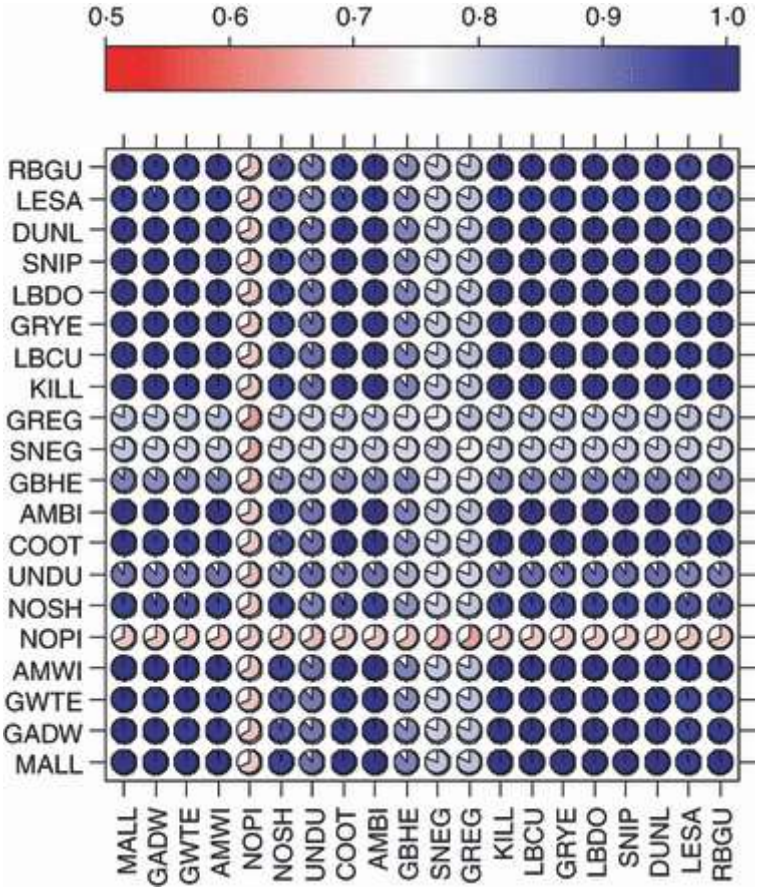
Violações de Normalidade

- A estatística F é relativamente robusta para violações da normalidade fornecida:
  - As populações são simétricas e unimodais.
  - Os tamanhos das amostras para os grupos são iguais e maiores que 10
  - Em geral, desde que os tamanhos das amostras sejam iguais (chamado de modelo balanceado) e suficientemente grandes, a suposição de normalidade pode ser violada, desde que as amostras sejam simétricas

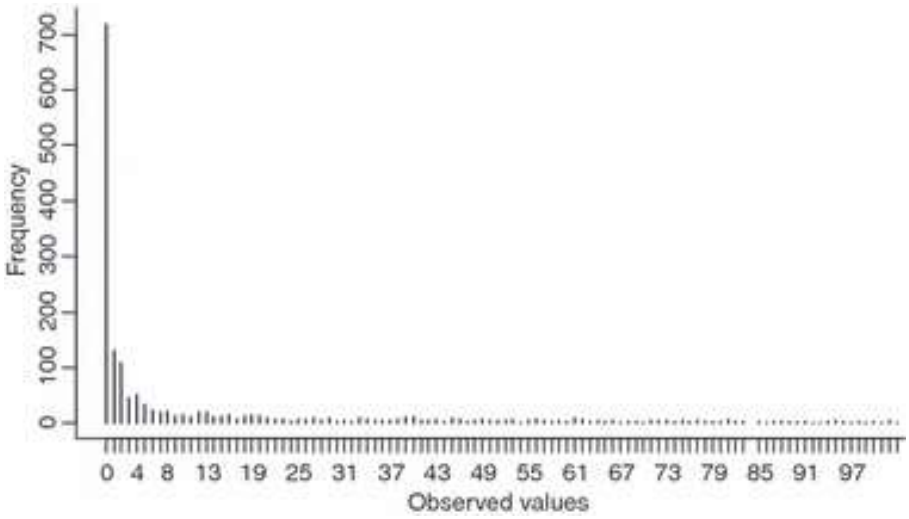
ou pelo menos semelhantes em forma (por exemplo, todas são negativamente inclinadas).

- A estatística F
    - não tão robusto a violações de homogeneidade de variâncias.
    - se a razão da maior variância para a menor variância for menor que 3 ou 4, o teste F será válido.
    - Se os tamanhos das amostras forem desiguais, diferenças menores nas variâncias podem invalidar o teste F.
    - Muito mais atenção deve ser dada às variâncias desiguais do que à não normalidade dos dados.

Etapa 4 - há muitos zeros nos dados?



Um corragram mostrando a frequência com que os pares de espécies de aves aquáticas têm abundância zero. A cor e a quantidade de preenchimento de um círculo correspondem à proporção de observações com zeros duplos. A diagonal que vai do canto inferior esquerdo ao canto superior direito representa a porcentagem de observações de uma variável igual a zero. As siglas de quatro letras representam diferentes espécies de aves aquáticas. A barra superior relaciona as cores no gráfico à proporção de zeros.



- Zero inflated GLM
- Logit regression
- Corragram
- Legendre & Legendre 1998
- Zuur et al. (2007)

Etapa 5: há colinearidade entre as covariáveis?

Table 1. ***P*-values of the *t*-statistic for three linear regression models and variance inflation factor (VIF) values for the full model. In the full model, the number of banded sparrows, which is a measure of how many birds were present, is modelled as a function of the covariates listed in the first column. In the second and third columns, the *P*-values and VIF values for the full model are presented (note that no variables have been removed yet). In the fourth column *P*-values are presented for the model after collinearity has been removed by sequentially deleting each variable for which the VIF value was highest until all remaining VIFs were below 3. In the last column, only variables with significant *P*-values remain, giving the most parsimonious explanation for the number of sparrows in a plot**

Covariate	*P*-value (full model)	VIF	*P*-value (collinearity removed)	*P*-value (reduced model)
%*Juncus gerardii*	0·0203	44·9953	0·0001	0·00004
% Shrub	0·9600	2·7818	0·0568	0·0727
Height of thatch	0·9989	1·6712	0·8263	
%*Spartina patens*	0·0640	159·3506	0·3312	
%*Distichlis spicata*	0·0527	53·7545	0·2538	
% Bare ground	0·0666	12·0586	0·8908	
% Other vegetation	0·0730	5·8170	0·9462	
%*Phragmites australis*	0·0715	3·7490	0·2734	
% Tall sedge	0·2160	4·4093	0·4313	
% Water	0·0568	17·0677	0·6942	
%*Spartina alterniflora* (short)	0·0549	121·4637	0·2949	
%*Spartina alterniflora* (tall)	0·0960	159·3828		
Maximum vegetation height	0·2432	6·1200		
Vegetation stem density	0·7219	3·2064		

Problemas

- Regressão linear múltipla
- ANOVA, modelos de efeitos mistos, análise de redundância (RDA), análise de correspondência canônica (CCA), modelos lineares gerais (GLM), GAM (modelos aditivos gerais)

Verifique se há inflação de variância

- Gráficos de dispersão
- Correlações
- PCA
- Funções de autocorrelação (ACF) - proc Arima
- variogramas
- Quanto custa medir as covariáveis?

Independence

When is this important?

Measurements over time on the same individual.
- Time series data (rainfall, temperature, etc).
- Repeated measures - split plots in time.
- Growth curves.

Measurements near each other in space.
- Split plot designs.
- Spatial data.

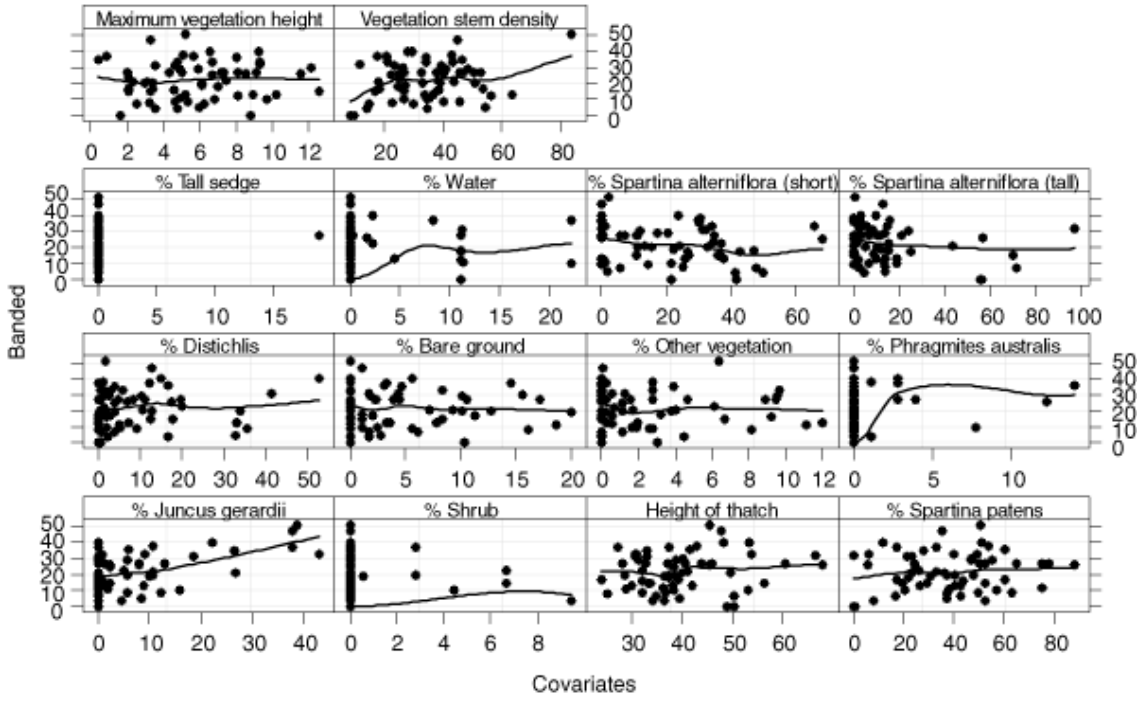
How do I know it's a problem?

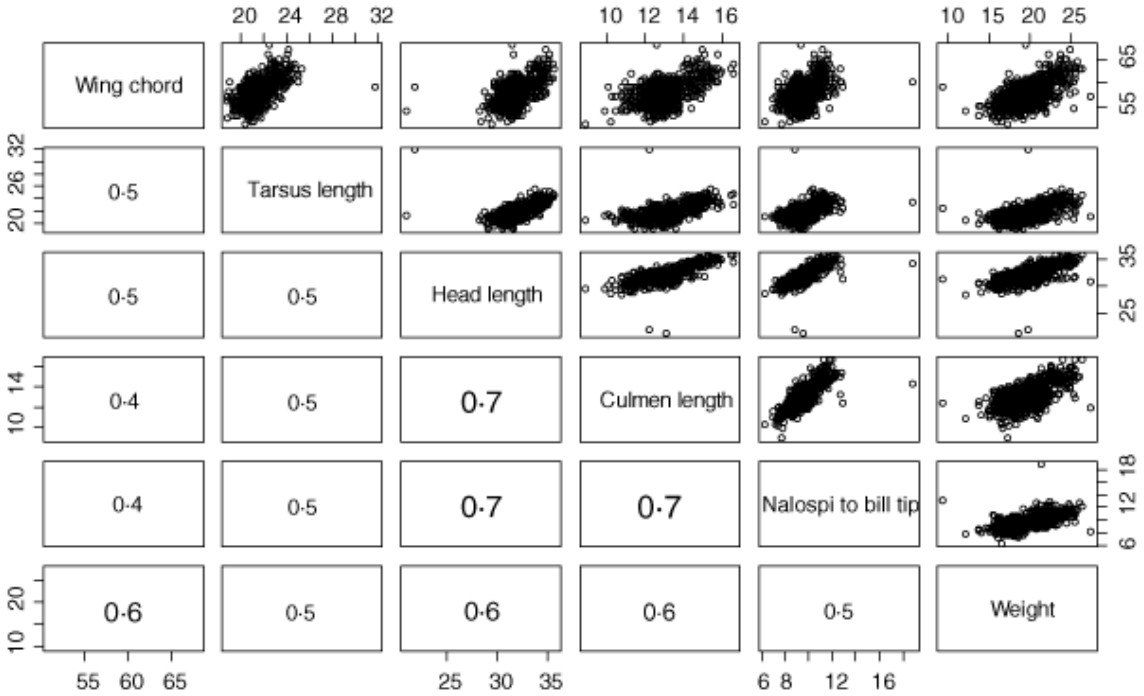
By design - how the data were collected.
Temporal/spatial autocorrelation analysis.

Rectifying a dependence problem.

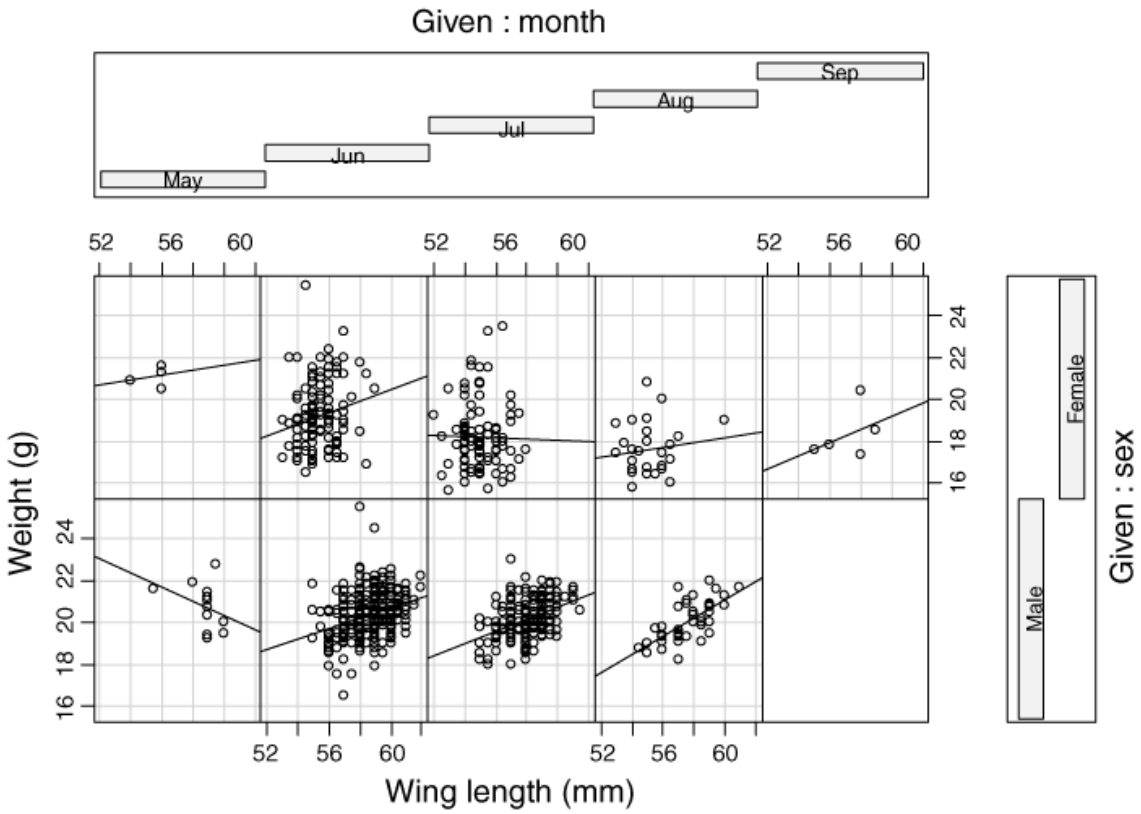
Modify the type of model to be fitted to the data.

Etapa 6: Quais são as relações entre as variáveis Y e X?



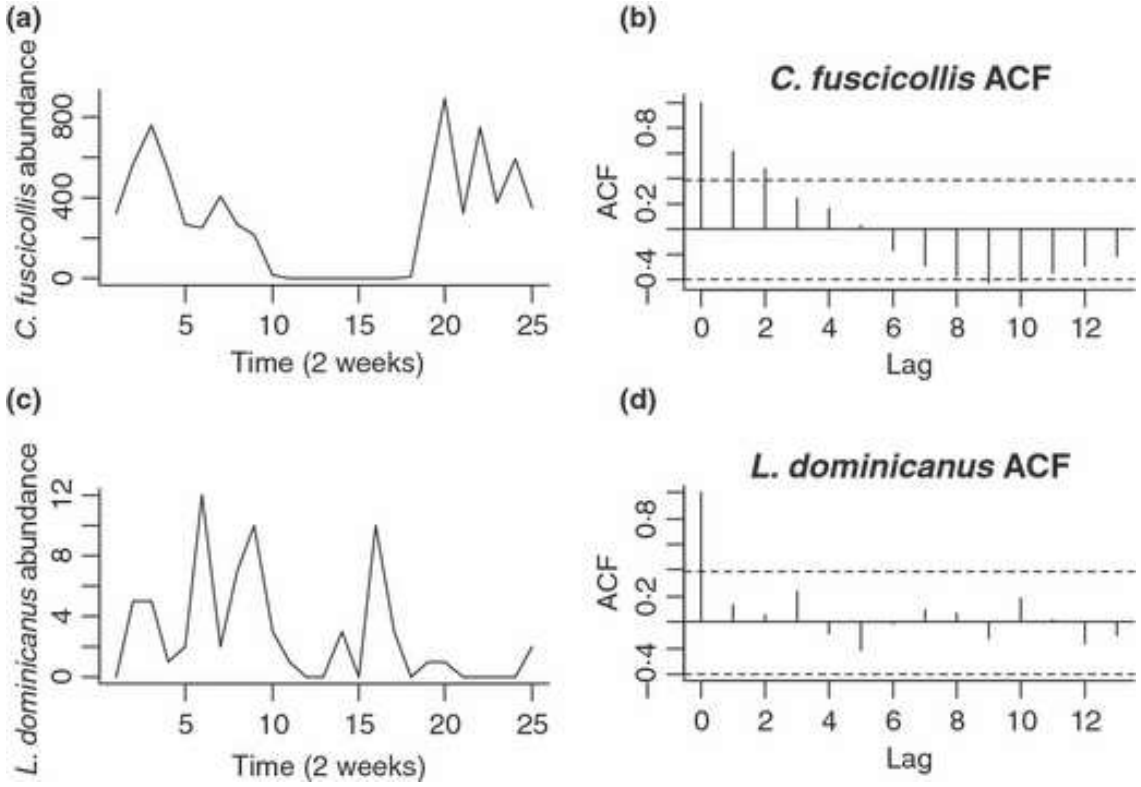

Wing chord
Tarsus length
Head length
Culmen length
Nalospi to bill tip
Weight
0·5
0·4
0·7
0·6

Etapa 7: Devemos considerar as interações?



Etapa 8: as observações da variável de resposta são independentes?

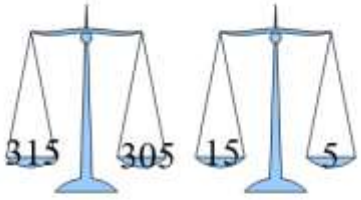
Espaço e Tempo



Independência de Médias e Variâncias

- A relação entre médias e variâncias é a fonte mais comum de heterogeneidade de variância
  - Acha mais freqüentemente onde existem muitas médias
  - Amostras com médias altas têm variâncias altas, as com médias baixas têm variâncias baixas

Lógica dita que existe uma diferença aqui

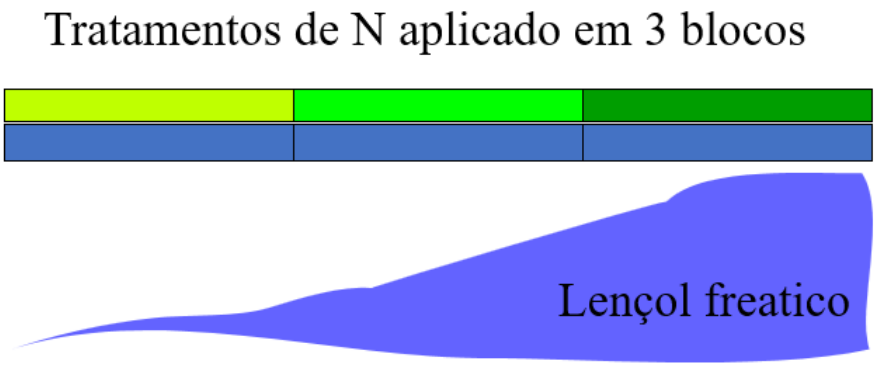


ANOVA dita "não" porque a diferença entre ambos é 10

Aonde está o problema...

Aditividade

- Cada delineamento experimental tem um modelo matemático chamada modelo linear aditivo
- $Y_i = Y + t_i + e_i$ DIC
- $Y_{ij} = Y + t_i + b_j + e_i$ DBC
- *Implica que o efeito do tratamento é o mesmo para todos os blocos e que o efeito de bloco é o mesmo para todos os tratamentos*
- Quando a presunção não é correta...



- Quando existe uma interação entre blocos e tratamentos - o modelo não é aditivo, mas multiplicativo

Independência dos Resíduos

- Garantida pela Casualização
- Princípio Básico da Experimentação;
  - Mesma unidade experimental é utilizada várias vezes para avaliar uma mesma característica;
  - Diferentes parcelas em contato físico direto;
  - Observações feitas por uma mesma pessoa durante um determinado intervalo de tempo;

Aditividade dos efeitos do modelo

Aditividade dos efeitos de tratamentos com os efeitos das variáveis de blocagem (DBC e DQL)

DIC: $y_{ji} = \mu + t_i + e_{ij}$

DBC: $y_{ji} = \mu + t_i + b_j + e_{ij}$

DQL: $y_{jik} = \mu + t_i + l_j + c_k + e_{ijk}$

O que pode ser feito?

Média do subgrupo

- A média dos subgrupos (tamanho recomendado maior que 4) geralmente produz uma distribuição normal
- Isso geralmente é feito com gráficos de controle
- Trabalha com o teorema do limite central
- Quanto mais distorcidos os dados, mais amostras são necessárias

Segmentação dos dados

- Os conjuntos de dados muitas vezes podem ser segmentados em grupos menores por estratificação de dados
- Esses grupos podem então ser examinados quanto à normalidade
- Depois de segmentados, os conjuntos de dados não normais muitas vezes se tornam grupos de conjuntos de dados normais

Transformando Dados

- Transformações Box-Cox de dados
  - $Y = X^a$
  - *"a"* vem dos dados
  - *Raiz quadrado se a=1/2*
  - *Recíproco se a= -1*
  - *Logarítmico se a = 0*
- Transformação Logit para dados Sim ou Não
- Dados exponenciais – classificação (rank)
- Distribuições truncadas para limites rígidos
- Aplicação de transformações que tem sentido
- As transformações nem sempre funcionam

Type of Distribution	Relationship of Mean & Variance	Type of Transformation
Poisson	Variance = Mean	Square Root
Binomial Proportions	Mean = p Variance = p(1–p)/n	Arcsine of Square Root
Exponential	SD = Mean	Log and Rank

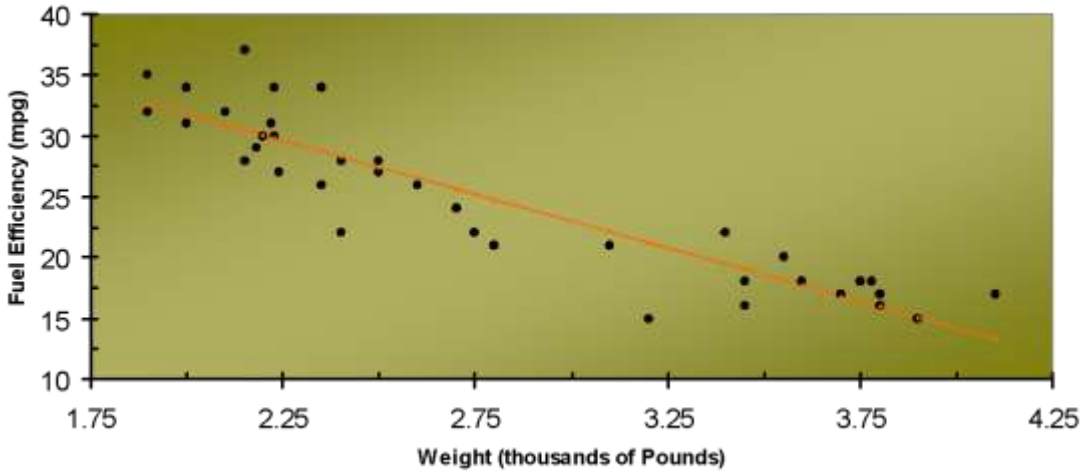
Exemplos de Distribuições (ver outro documento)

Distribution	Type Data	Examples
Normal	Continuous	Useful when it is equally likely the readings will fall above or below the average
Lognormal	Continuous	Cycle or lead time data
Weibull	Continuous	Mean time-to-failure data, time to repair and material strength
Exponential	Continuous	Constant failure rate conditions of products
Poisson	Discrete	Number of events in a specific time period (defect counts per interval such as arrivals, failures or defects)
Binomial	Discrete	Proportion or number of defectives

- Extreme value
- Logistic
- Log logistic

Estreitando relacionamentos

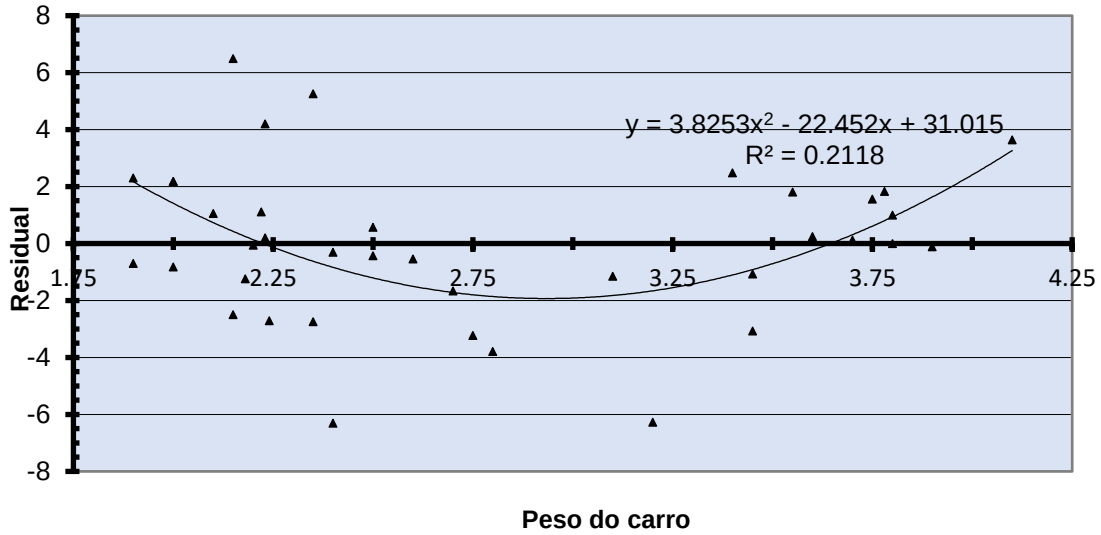
- A suposição fundamental para trabalhar com um modelo linear é que o relacionamento que você está modelando é, na verdade, linear.
- Frequentemente, é difícil dizer a partir do gráfico de dispersão que você está vendo antes de ajustar o modelo de regressão.
- Às vezes, você não pode ver uma curva no relacionamento até plotar os resíduos.
- Aqui estão os pesos e a eficiência de combustível de 38 carros. O que você vê?



- Obviamente, carros mais pesados tendem a consumir mais combustível para se mover.
- O gráfico de dispersão mostra uma forte associação linear negativa entre o peso de um carro e sua eficiência de combustível (r = 0,911).
- A equação de regressão linear é

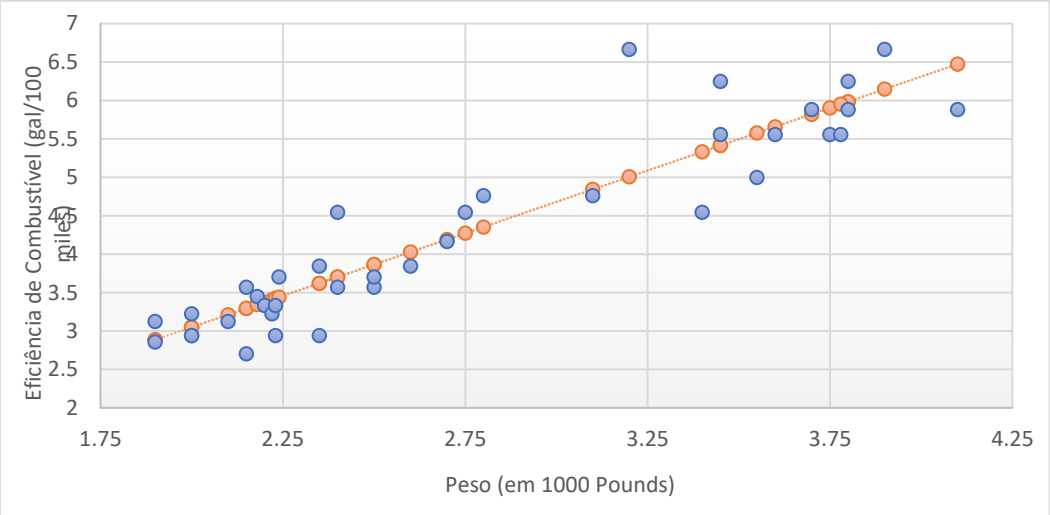
$$\widehat{mpg} = 49.4 - 8.79(weight)$$

O que você vê nos residuais?

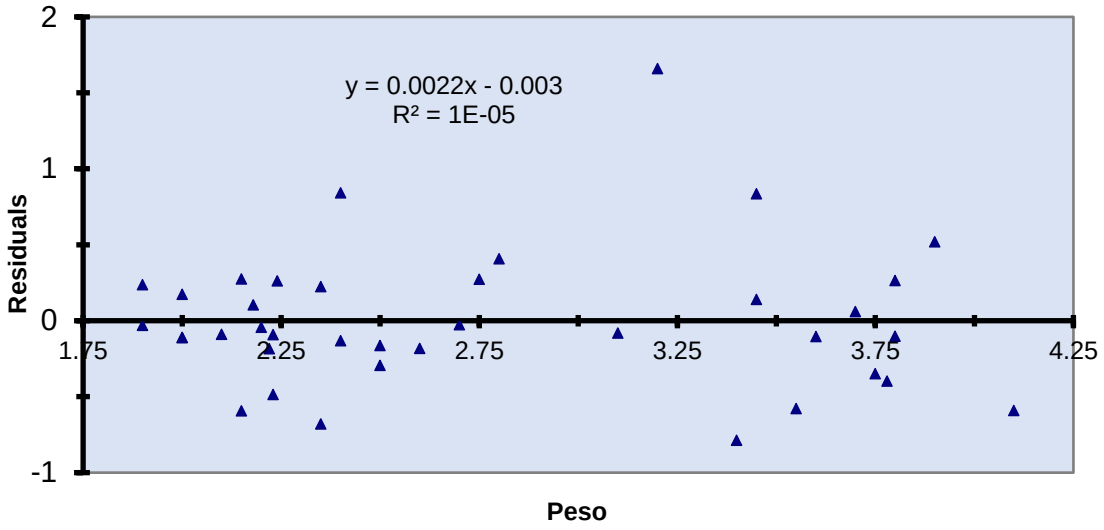


- O gráfico de dispersão dos resíduos versus os pesos não deve ter um padrão.

- O que vemos é uma curva
    - começando alto à esquerda, caindo no meio do gráfico e subindo novamente à direita.
- Os gráficos de resíduos podem revelar padrões como esse que são difíceis de ver no gráfico de dispersão original.
- Como podemos tentar corrigir essa deficiência em nosso modelo?
    - Em vez de olhar para milhas por galão
    - poderíamos fazer como a maioria do resto do mundo e considerar o número de galões por milha,
    - ou o número de galões por 100 milhas
    - Aqui, novamente, estão os pesos e a eficiência do combustível medidos em galões por 160 quilômetros de 38 carros. O que você vê agora?



Você vê alguma surpresa neste gráfico de residuais?



Outro exemplo

- Como os caibros de água branca sabem, um rio flui mais devagar perto de suas margens devido ao atrito.
- Para estudar a natureza da relação entre a velocidade da água e a distância da margem, os dados foram coletados sobre a velocidade (em centímetros por segundo) de um rio a diferentes distâncias (em metros) da costa.
- Represente graficamente esses dados e descreva o gráfico de dispersão.
- Faça uma regressão linear e examine o gráfico residual. Algo distinto?

Distance	Velocity
0.5	22.00
1.5	23.18
2.5	25.48
3.5	25.25
4.5	27.15
5.5	27.83
6.5	28.49
7.5	28.18
8.5	28.50
9.5	28.63

Vamos re-expressar os valores de x substituindo cada x por sua raiz quadrada

Distance	Sq Rt Distance	Velocity
0.5	0.707106781	22.00
1.5	1.224744871	23.18
2.5	1.58113883	25.48
3.5	1.870828693	25.25
4.5	2.121320344	27.15
5.5	2.34520788	27.83
6.5	2.549509757	28.49
7.5	2.738612788	28.18
8.5	2.915475947	28.50
9.5	3.082207001	28.63

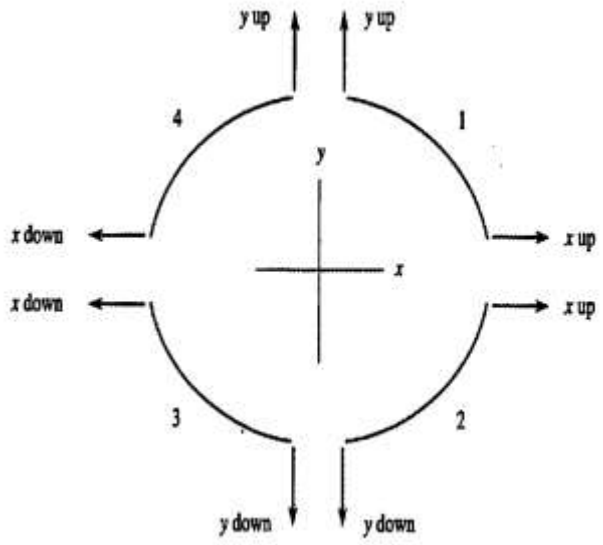
- Represente graficamente a raiz quadrada da distância do banco em relação à velocidade.
- Descreva o gráfico de dispersão.
- Faça uma regressão linear e escreva a nova equação de regressão.
- Examine o gráfico residual.
- Algo distinto?

Escada de poderes de Tukey (regra de protuberância)

- Portanto, transformar os valores de x usando a função de raiz quadrada funcionou bem.
- Como podemos escolher uma transformação que resultará em um padrão linear?
- A "escada de poderes" forneceu algumas orientações aos estatísticos.
- A tabela a seguir exibe as transformações de energia mais comumente usadas.
- O poder 1 corresponde a nenhuma transformação.
- Usar a potência o transformaria todos os valores em 1, o que certamente não é muito informativo, portanto, os estatísticos usam a transformação logarítmica em seu lugar na escada de transformações.

Power(p)	Transformation	Name
2	Y^2	Square
1	Y (No transformation)	Original Data
½	√ Y	Square root
"0"	log Y or $\log_{10}$ (Y)	Logarithm
-½	-1 / √ Y	Reciprocal Root
-1	-1 / Y	Reciprocal
-2	-1 / Y^2	Reciprocal Square

- Onde na escada devemos ir para encontrar uma transformação apropriada.
- As quatro curvas, rotuladas 1, 2, 3 e 4, representam formas de gráficos de dispersão curvos que são comumente encontrados.



Suponha que um gráfico de dispersão tenha uma dobra como a rotulada 1.

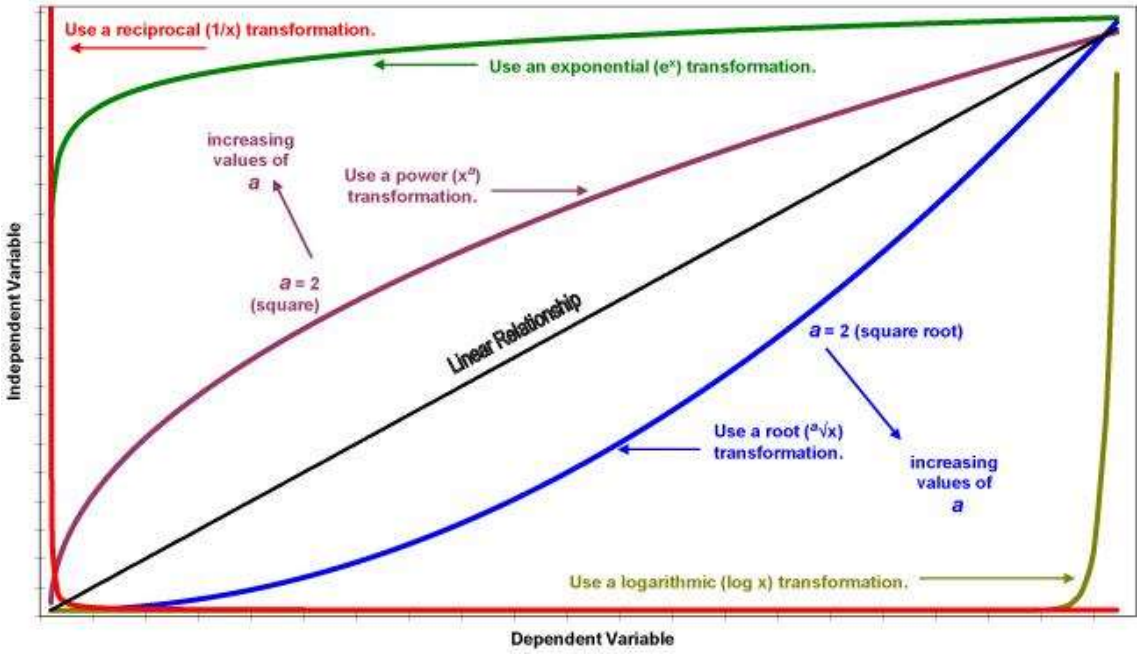
Para endireitar o gráfico, usaríamos uma potência de x que está subindo a escada da linha sem transformação (x^2 ou x^3) e / ou uma potência de y que também está subindo a escada de 1.

Assim, podemos ser levados a elevar x ao quadrado, talvez cubar cada y, e traçar os pares ordenados transformados.

Se a curvatura na dispersão original se parece com o segmento curvo 2,

uma potência para cima na escada de nenhuma transformação para x (x^2 ou x^3) e ou uma potência para baixo na escada para y (por exemplo, $\sqrt{y}$ ou log y) deve ser usada.

If a data relationship looks like one of these curves,
try using a transformation of the independent variable
to make the relationship linear.

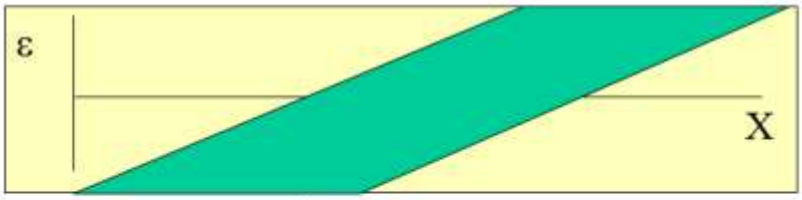


Lack of Homogeneity in Regression

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?

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

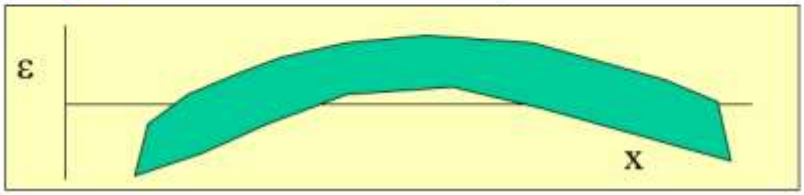
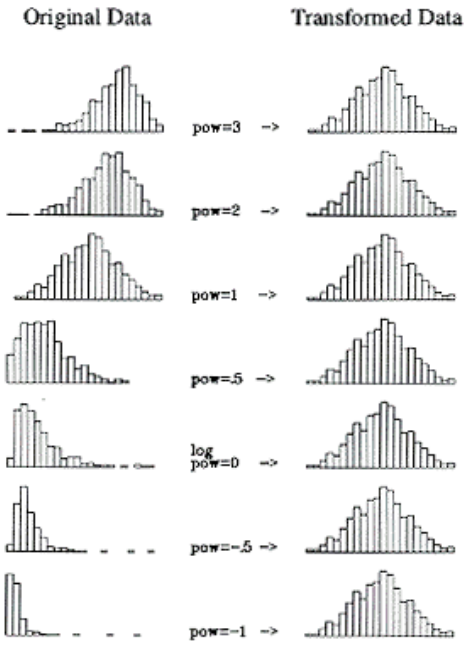
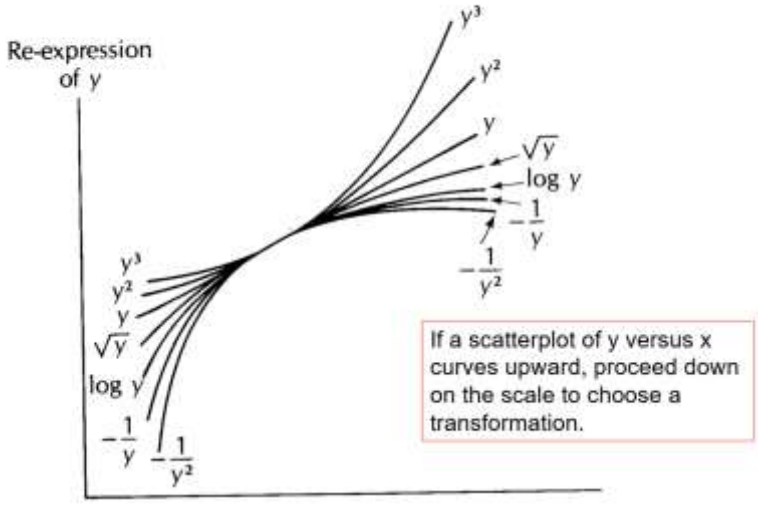


Figure 22.3
Ladder of Power Transformations

Power P	SYSTAT BASIC	Name	Notes
P	Y^P	power	DOWN: shorten upper tail.
⋮	⋮	⋮	⋮
3	Y^3	Cube	Not commonly used.
2	Y^2	Square	The highest commonly used power.
—> 1	Y^1	Original data	No transformation.
1/2	Y^(1/2)	Square root	Commonly used for counts.
"0"	LOG(Y)	Logarithm	Commonly used for financial data.
-1/2	-1/Y^(1/2)	Reciprocal root	The minus sign preserves order.
-1	-1/Y	Reciprocal	Lowest commonly used power.
-2	-1/Y^2	Reciprocal square	
⋮	⋮	⋮	⋮
- P	-1/Y^P	Reciprocal power	UP: shorten lower tail.

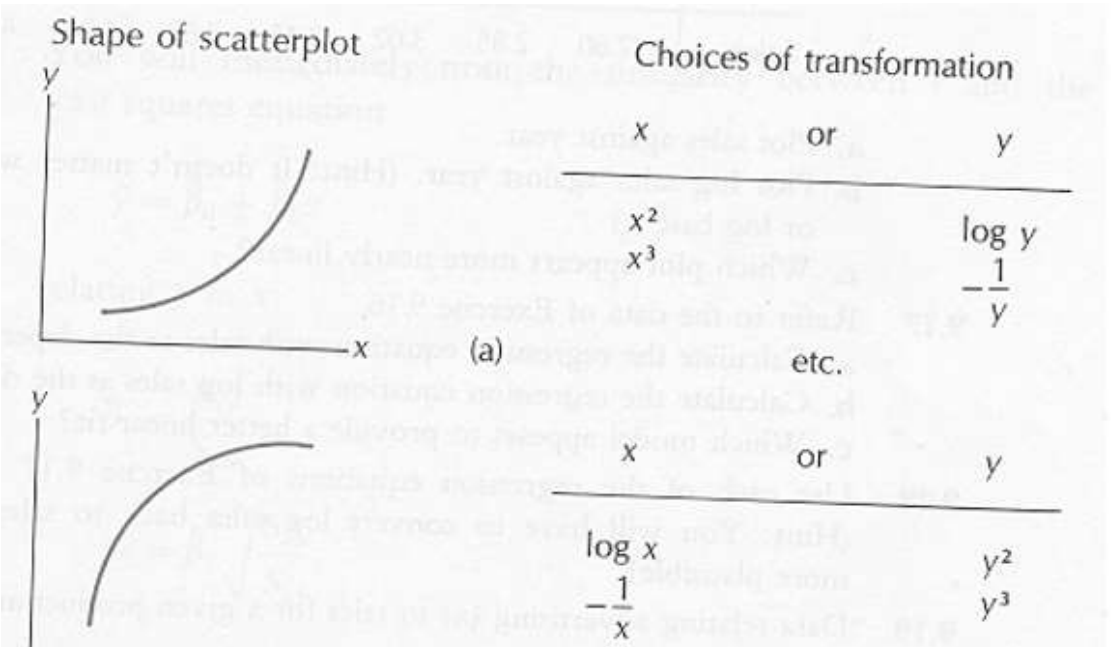


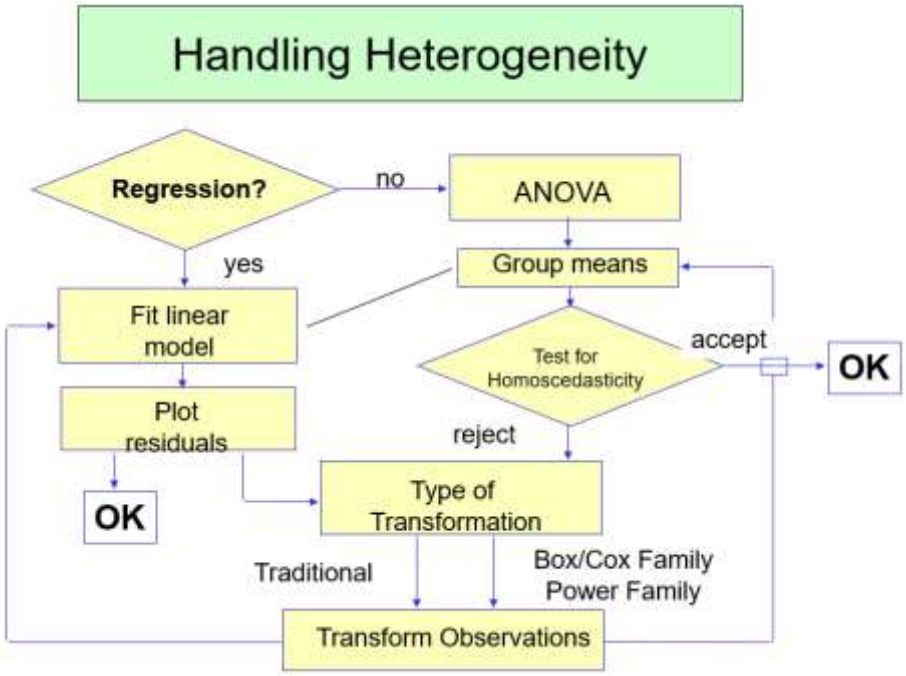
Transforming the Response to achieve Linearity



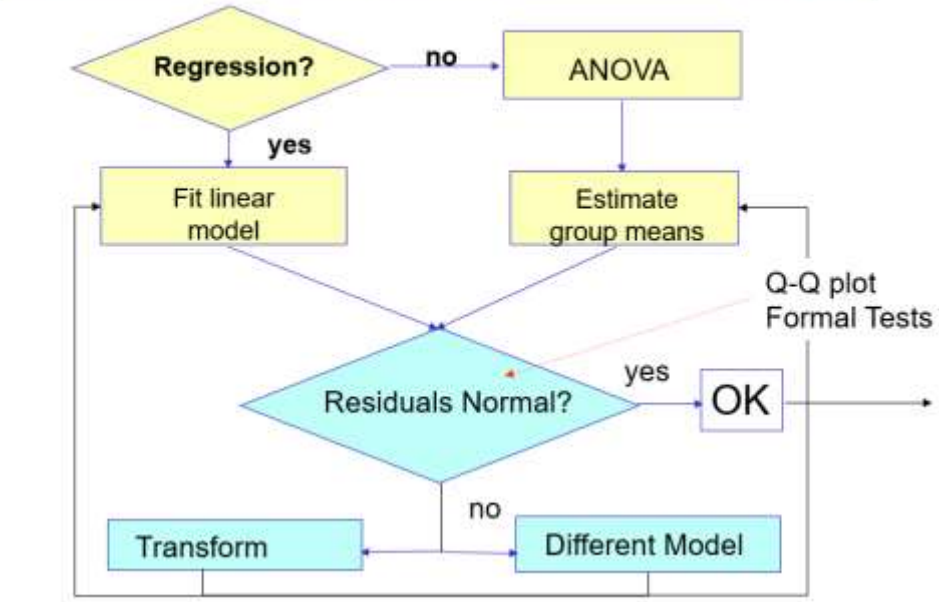
Examples of Transformations Used in Data Analysis

Type of Transformation		Example of Use	Number of Variables	Information Added	Result	How It Helps
Sample Adjustments	Deletion	Trimming tails, deleting outliers	One	No	Changed variable	Reduce variation
	Replacement by Fiat	Changing misleading values	One	No		
	Replacement with Constant	Filling in for censored or missing data	One	No		Keeps sample & variable in analysis
	Replacement from Same Variable	Averaging replicates	One	No		
	Replacement from Different Variable	Regression estimates	Any	Yes		
Dependent Variable Transformations		Simple	One	No	New variable	Achieve Normality
		Box-Cox				
Independent Variable Transformations	Variable Adjustments	Differencing	One	No	New variable	Explore data patterns
		Data Shifts				
		Smoothing				Reduce variation
		Standardization		Yes		
	Variable Rescaling	Convert Units: Simple	One	No	Changed variable	Convenience
		Convert Units: Add Information	One	Yes	New variable	Explore data patterns
		Computational Efficiency	One	No	Changed variable	Convenience
		Rescale to fewer divisions	One	No	New variable	Simplify Analysis
		Rescale to more divisions	One	Yes	New variable	Explore data patterns
	Variable Linearization	Reciprocals, roots, powers, logs	One	No	New variable	Explore data patterns
	Variable Combinations	Simple Arithmetic	Any	Yes	New variable	Explore data patterns
		More Complex Mathematics				
Supplemental Variables	Created by Stratification	Groupings from demographic data	Any	Yes	New variable	Explore data patterns
	Created by Concatenation	Sample ID, plot IDs	Any	Yes	New variable	Convenience
	Created from Metadata	Sampler, sampling device	Any	Yes	New variable	Explore data patterns
	Created from References	Government data	Any	Yes	New variable	Explore data patterns




Handling Heterogeneity
Regression?
no
ANOVA
yes
Group means
Fit linear model
Test for Homoscedasticity
accept
OK
Plot residuals
reject
OK
Type of Transformation
Traditional
Box/Cox Family
Power Family
Transform Observations

Transformations to Achieve Normality



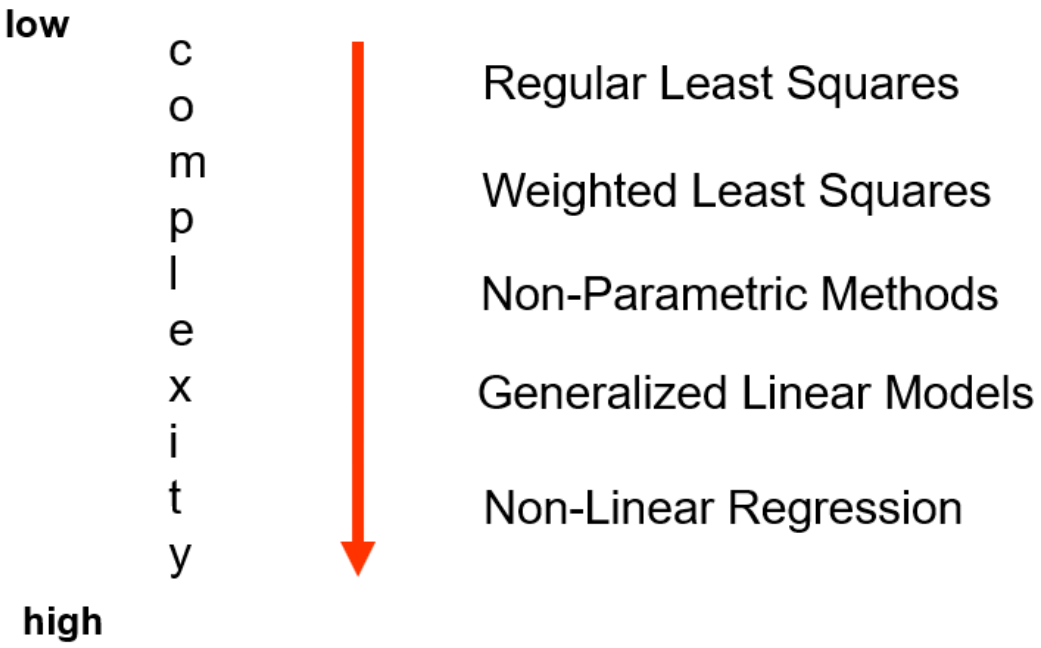
Non-normal! So what?

- Only very skewed distributions will have a marked effect on the significance level of the F-test for overall model or model effects
- Often the same transformations which are used to achieve homoscedasticity will produce more normal-looking observations (residuals).

Transformations to Achieve Model Simplicity

- GOAL: To provide as simple as possible a mathematical form for the relationship among response and explanatory variables.
- It May require transforming both response and explanatory variables

Alternative Models

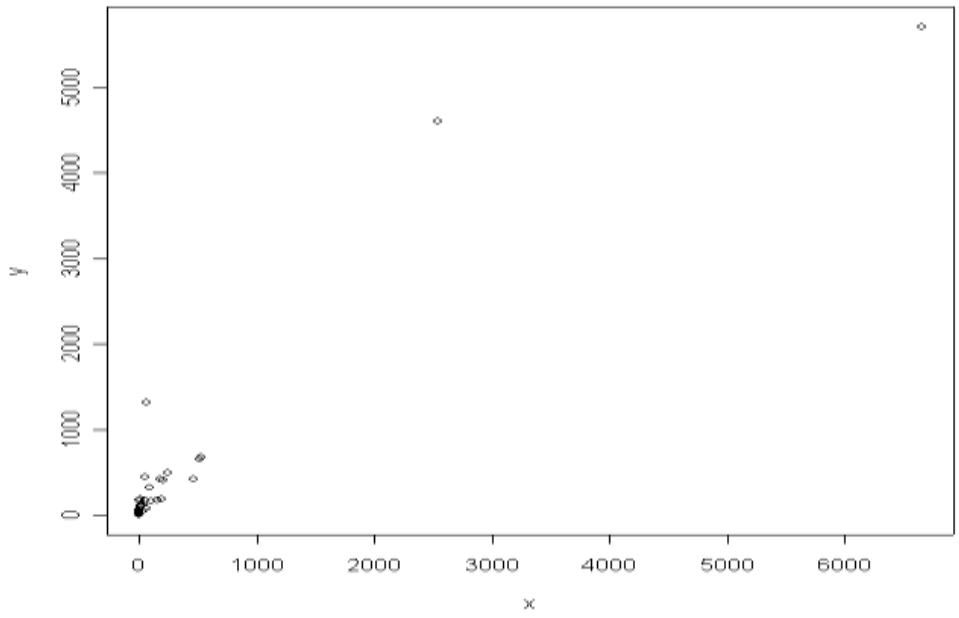


Example: Predicting brain weight from body weight in mammals via SLR

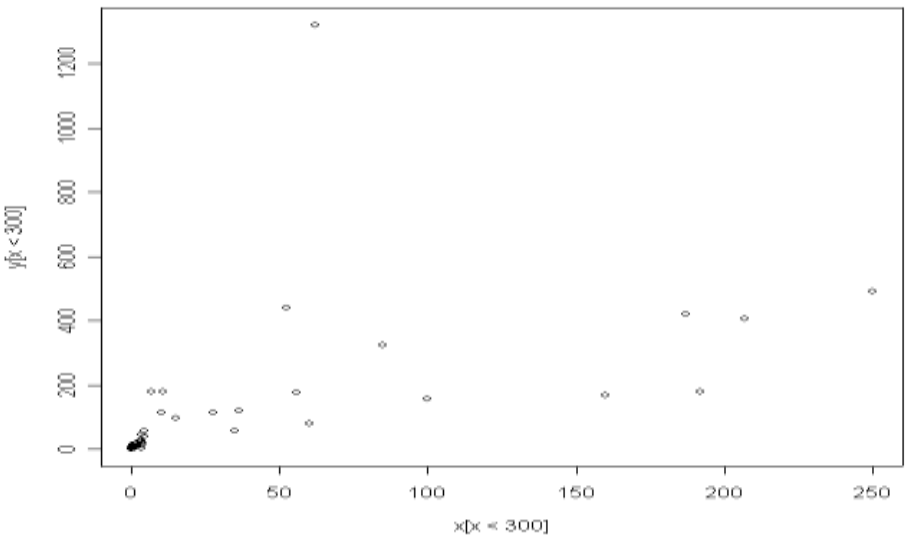
Data are average brain (Y, g) and body (X, kg) weights for 62 species of mammals (2 omitted). Source: Allison & Chicchetti (1976), *Science*.

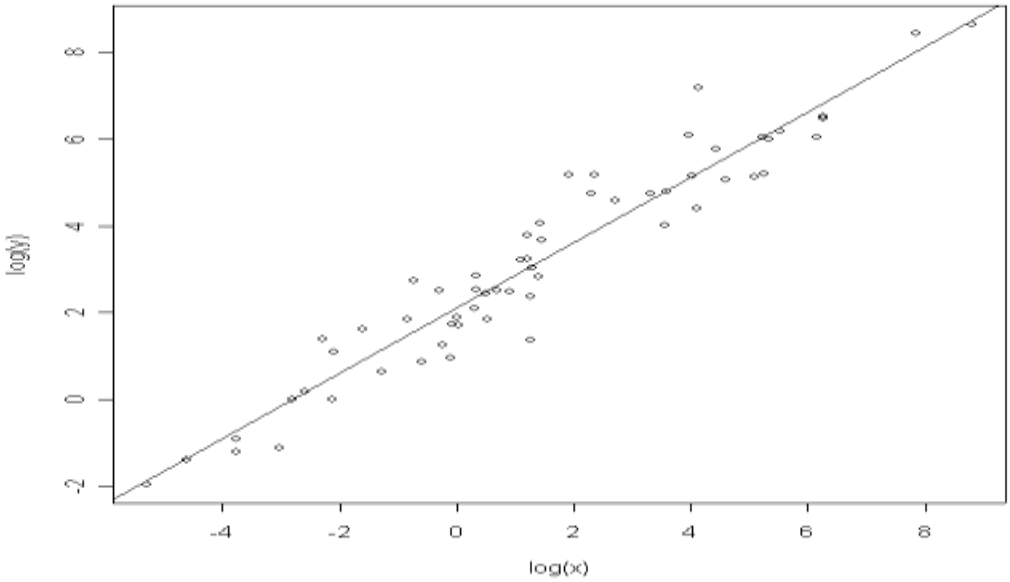
Species (common name)	body weight	brain weight
Arctic fox	3.385	44.500
Owl monkey	0.480	15.499
Horse	521.000	655.000
Kangaroo	35.000	56.000
Human	62.000	1320.000
African elephant	6654.000	5712.000
Asian elephant	2547.000	4603.000
...		
Chimpanzee	52.160	440.000
Tree shrew	0.104	2.500
Red fox	4.235	50.400

Scatterplot of data is non-informative. Most species have small weights compared to the elephants.



Viewing only those mammals with body weight below 300kgs suggests transforming to a log scale to linearize the relationship

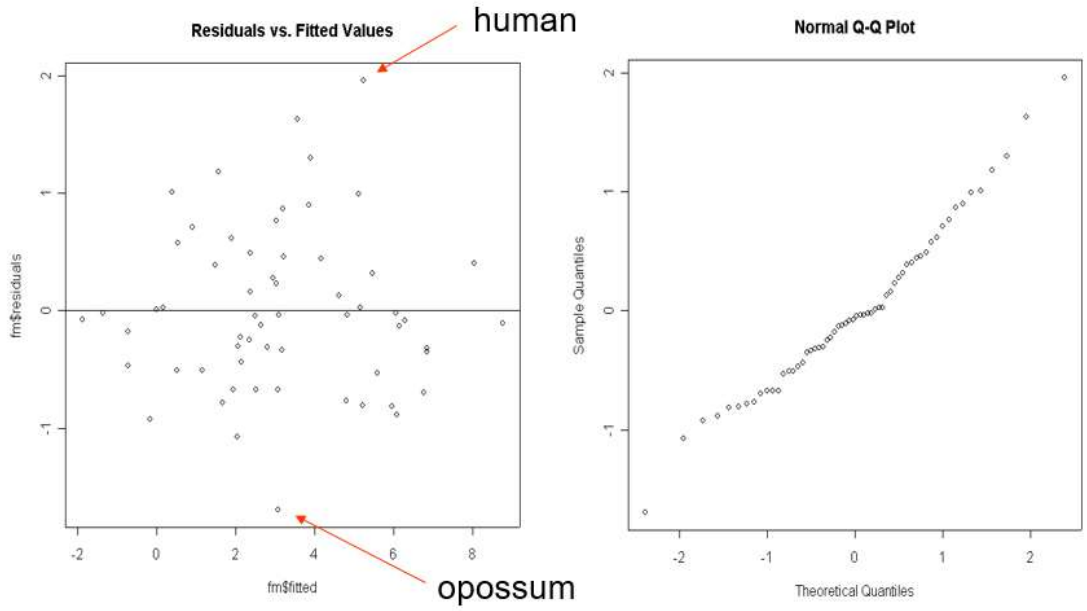




Scatterplot looks linear. Fitted regression equation is:

$$\log(y) = 2.111 + 0.755\log(x)$$

Body weight is a very significant predictor of brain weight (p-value<0.0001). Also, R^2=0.922.



Residual plot shows no obvious violations of the zero mean and constant variance assumption.

QQ-Plot demonstrates that the normality assumption for the residuals is plausible.

- **Detecting Unusual and Influential Data**
    - scatterplots of the dependent variables versus the independent variable
    - looking at the largest values of the studentized residuals, leverage, Cook's D, DFFITS and DFBETAs

Testes Não paramétricos

- Às vezes, a transformação leva à variação do valor p.
- Então, o que podemos fazer se não podemos transformar os dados?
- Geralmente é mais seguro usar métodos que não exijam tais suposições
- Isso inclui os métodos não paramétricos.

Quando usar?

- A variável de resposta (residual) não pode ser presumido ser distribuido normalmente
- Precisa disso p usar ANOVA
- Dados limitados
- Contagens e ordem são mais relevantes que médias
- Used for statistical tests when data is not normal
- Tests using medians rather than means
- Most often used when sample sizes of groups being compared are less than 100, but just as valid for larger sample sizes

Testes Paramétricos vs. Não Paramétricos

- Os testes paramétricos são mais "conservadores"
    - (ou seja, menos probabilidade de cometer um erro Tipo I).
- Testes não paramétricos são mais "liberais"
    - (ou seja, é mais provável cometer um erro Tipo I).
    - Quando os dados não são normalmente distribuídos e as medições, na melhor das hipóteses, contêm informações de ordem de classificação, calcular as estatísticas descritivas padrão (por exemplo, média, desvio padrão) às vezes não é a maneira mais informativa de resumir os dados
    - menos rigorosas dos dados (resistente a outliers, forma de distribuição)
    - às vezes pode ser usado para obter uma resposta rápida com poucos cálculos

- fornecer um ar de objetividade quando não houver uma escala subjacente confiável (universalmente reconhecida) para os dados originais
- Diferentes testes não paramétricos podem produzir resultados diferentes
- Aconselhável executar diferentes testes não paramétricos
- Os métodos não paramétricos são mais apropriados quando os tamanhos das amostras são pequenos
- Quando o conjunto de dados é grande, muitas vezes faz pouco sentido usar estatísticas não paramétricas

- Quando o conjunto de dados é grande (por exemplo, n> 100), muitas vezes faz pouco sentido usar estatísticas não paramétricas.

O que acontece quando você usa um teste não paramétrico com dados de uma distribuição normal?

- Maior incidência de erro tipo II
- Os testes não paramétricos carecem de poder estatístico com pequenas amostras
- Os testes não paramétricos não podem dar resultados muito significativos para amostras muito pequenas, pois todas as somas de classificação possíveis são bastante prováveis
- Eles não fornecem, sem a adição de suposições extras, intervalos de confiança para as médias ou medianas das distribuições subjacentes
- Suponha que os dados possam ser solicitados - o poder do teste diminui se houver muitos empates

Os testes paramétricos são frequentemente preferidos porque:

- Eles são robustos
- Eles têm maior eficiência energética (maior poder em relação ao tamanho da amostra)

- Eles fornecem informações exclusivas (por exemplo, a interação em um projeto fatorial)
- Testes paramétricos e não paramétricos geralmente abordam dois tipos diferentes de questões

- Portanto, na maioria das aplicações biológicas, deve-se sempre tentar usar um teste paramétrico primeiro.
- Os métodos não paramétricos são mais apropriados quando os tamanhos das amostras são pequenos.

Normal theory-based test	**Corresponding nonparametric test**	**Purpose of test**
t test for independent samples	Mann-Whitney U; Wilcoxon rank-sum	Compares two independent samples
Paired *t* test	Wilcoxon matched pairs signed-rank	Examines a set of differences
Pearson correlation coefficient	Spearman rank correlation coefficient	Assesses the linear association between two variables
One way analysis of variance (*F* test)	Kruskal-Wallis analysis of variance by ranks	Compares three or more groups
Two-way analysis of variance	Friedman two-way analysis of variance	Compares groups classified by two different factors

		Outcome variable					
		Nominal	Categorical (>2 Categories)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
Input Variable	Nominal	X^2 or Fisher's	X^2	X^2-trend or Mann - Whitney	Mann-Whitney	Mann-Whitney or log-rank[a]	Student's *t* test
	Categorical (2>categories)	X^2	X^2	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Analysis of variance[c]
	Ordinal (Ordered categories)	X^2-trend or Mann - Whitney	*	Spearman rank	Spearman rank	Spearman rank	Spearman rank or linear regression[d]
	Quantitative Discrete	Logistic regression	*	*	Spearman rank	Spearman rank	Spearman rank or linear regression[d]
	Quantitative non-Normal	Logistic regression	*	*	*	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and linear regression
	Quantitative Normal	Logistic regression	*	*	*	Linear regression[d]	Pearson and linear regression

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Sample		K Sample (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical or Nominal	X^2 or bi-nomi	X^2	Macnarmar's X^2	X^2	Cochran's Q	
Rank or Ordinal		Mann Whitney U	Wilcoxin Matched Pairs Signed Ranks	Kruskal Wallis H	Friendman's ANOVA	Spearman's rho
Parametric (Interval & Ratio)	z test or t test	t test between groups	t test within groups	1 way ANOVA between groups	1 way ANOVA (within or repeated measure)	Pearson's r
		Factorial (2 way) ANOVA				