

Lecture 13 – Correlation & Linear Regression

- Outline for today
 - Recall statistical methods of hypothesis testing 假设检验方法
 - Correlation between numerical variables 数值变量间的相关性 (Ch16)
 - Simple Linear Regression 简单线性回归 (Ch17)
 - Summary
 - R Lab & Discussion

生物统计学

李勤

生态与环境科学学院

一些提示

- 期末考试时间：暂定1月9日 10:30-12:30
- 注：请认真审题，按题目要求回答问题；题目有要求时需提供计算过程或者R代码。本试题考试形式为课堂开卷，允许学生查阅自己准备的包括电子文档在内的所有资料，但不允许使用互联网。本试题所涉数据皆为虚构，无任何生物学意义，仅供考查《生物统计学》的学习成效。

1. 回顾——假设检验

- 选择合适的检验方法（回答四个问题）
 - 检验涉及一个变量，还是检验两个或更多变量之间的关系？
 - 变量是分类的还是数值的？
 - 数据是否以配对形式出现？
 - 检验的假设是什么，数据是否符合这些假设？
- 参考：INTERLEAF 7 （Page 958）

表 1. 一个变量的检验方法 (一组数据)

Data type 数据类型	Goal	目标	Test	检验方法
Categorical 类型变量 (分类变量)	Use frequency data to test whether a population proportion equals a null hypothesized value	使用频数数据检测总体中的比例是否等于零假设中的值	Binomial test (7) χ^2 Goodness-of-fit test with two categories (use if sample size is too large for the binomial test) (8)	二项检验 (7) χ^2 拟合度检验, 应用于变量分为两类 (如果样本量过大, 无法进行二项检验, 则使用该检验) (8)
	Use frequency data to test the fit of a specific population model	使用频数数据检测特定总体模型的拟合程度	χ^2 Goodness-of-fit test (8)	χ^2 拟合度检验 (8)
Numerical 数值变量	Test whether the mean equals a null hypothesized value when data are approximately normal (possibly only after a transformation) (13)	当数据近似正态分布 (或经过转换后符合) 时, 检测平均值是否等于零假设中的值 (13)	One-sample <i>t</i> -test (11)	单样本 <i>t</i> 检验 (11)
	Test whether the median equals a null hypothesized value when data are not normal (even after transformation)	当数据不符合正态分布 (即使经过转换), 检测中位数是否等于零假设中的值	Sign test (13)	符号检验 (13)
	Use frequency data to test the fit of a discrete probability distribution	使用频率数据测试离散概率分布的拟合程度	χ^2 Goodness-of-fit test (8)	χ^2 拟合度检验 (8)
	Use data to test the fit of the normal distribution	检测数据是否符合正态分布	Shapiro-Wilk test (13)	Shapiro-Wilk 检验 (13)

表 2. 两个变量相关性的检验方法

		Type of explanatory variable 解释变量			
		Categorical 类型变量	类型变量	Numerical 数值	数值变量
Type of response variable 响应变量	Categorical 类型变量	Contingency analysis (9)	独立性检验 (9)	Logistic regression (17)	逻辑斯蒂回归 (17)
	Numerical 数值变量	<i>t</i> -tests, ANOVA, Mann-Whitney <i>U</i> -test, etc. [See Table 3 for more details.]	<i>t</i> 检验 方差分析 <i>U</i> 检验等 [更多细节见表 3]	Linear and nonlinear regression (17) Linear correlation (16) Spearman's rank correlation (when data are not bivariate normal) (16)	线性和非线性回归 (17) 线性相关 (16) Spearman 秩相关 (当数据不 是二元正态分布时) (16)

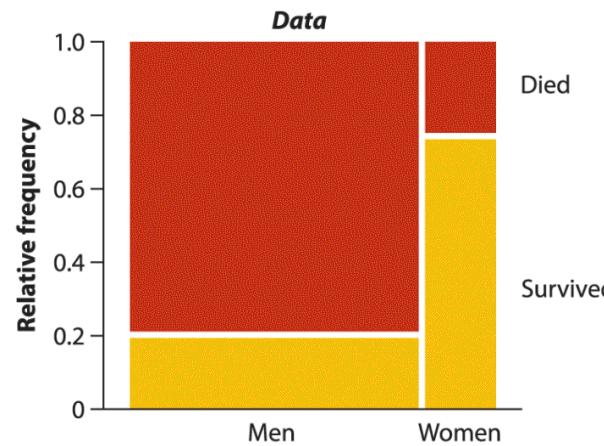
表 3. 两个变量相关性的检验方法及其前提假设 (或检验多组数据的差异)

Number of treatments 处理组数量	Tests assuming normal distribution 假设数据符合正态分布的检验	Tests not assuming normal distributions 假设数据不符合正态分布的检验	
Two treatments (independent samples) 两独立样本	<i>Welch's t</i> -test (12) Two-sample <i>t</i> -test (use when variance is equal in the two groups) (12)	<i>Welch's t</i> 检验 (12) 双样本 <i>t</i> 检验 (两组方差相 等时使用) (12)	Mann-Whitney <i>U</i> -test (Wilcoxon rank-sum test) (13)
Two treatments (paired data) 两配对样本	Paired <i>t</i> -test (12)	配对 <i>t</i> 检验 (12)	Sign test (13)
More than two treatments 超过两组设置	ANOVA (15)	方差分析 (15)	Kruskal-Wallis test (15)

1.2 The association between two variables 变量间的相关性



类型变量 ~ 类型变量

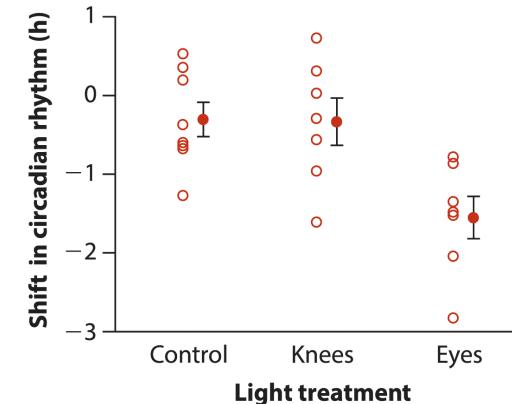
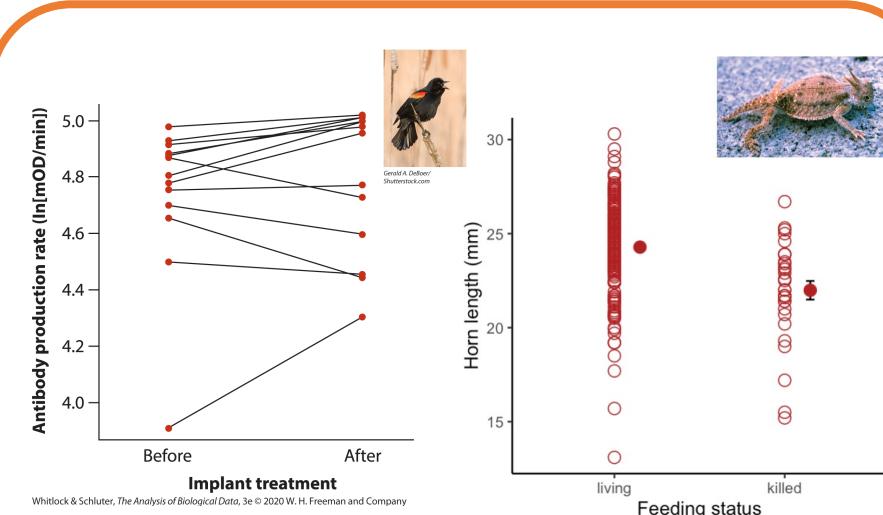


处理

	Treatment group (1)	Control group (2)
undesired outcome	<i>a</i>	<i>b</i>
(desired) outcome	<i>c</i>	<i>d</i>

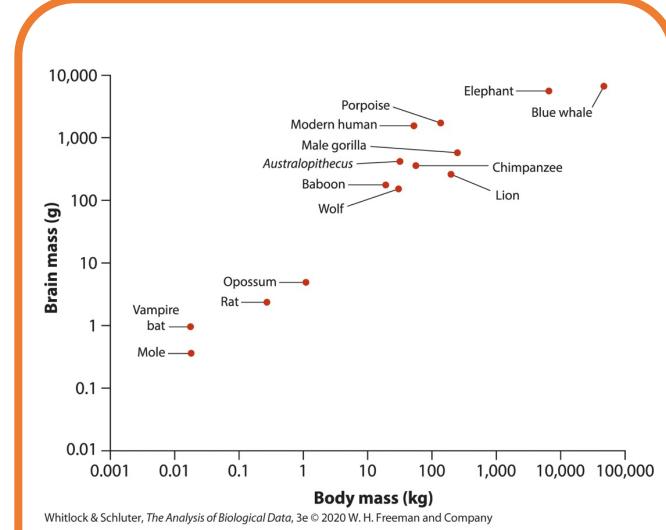
结果

数值变量 ~ 类型变量

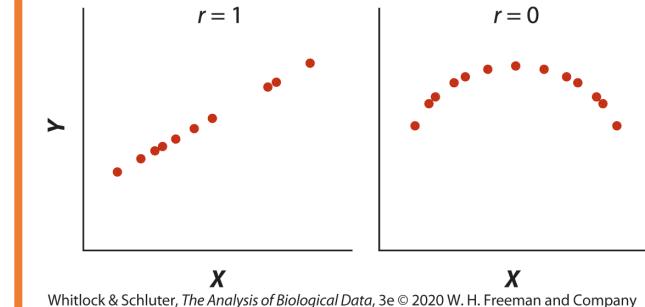


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

数值变量 ~ 数值变量



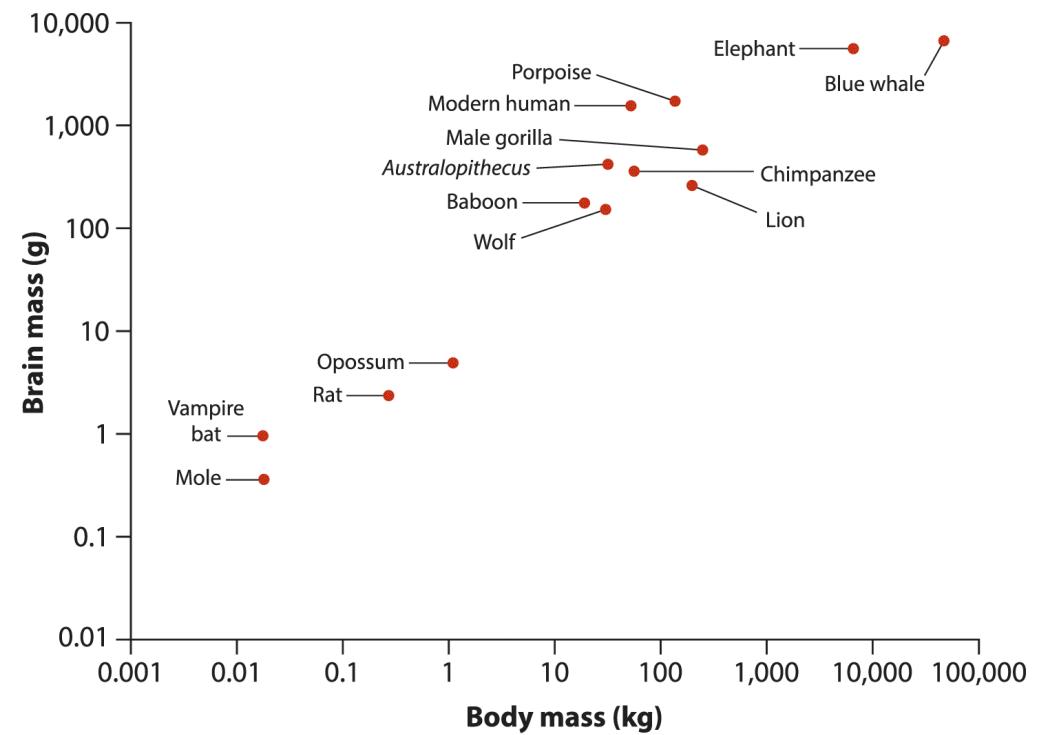
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2. 数值变量间的相关性 Correlation

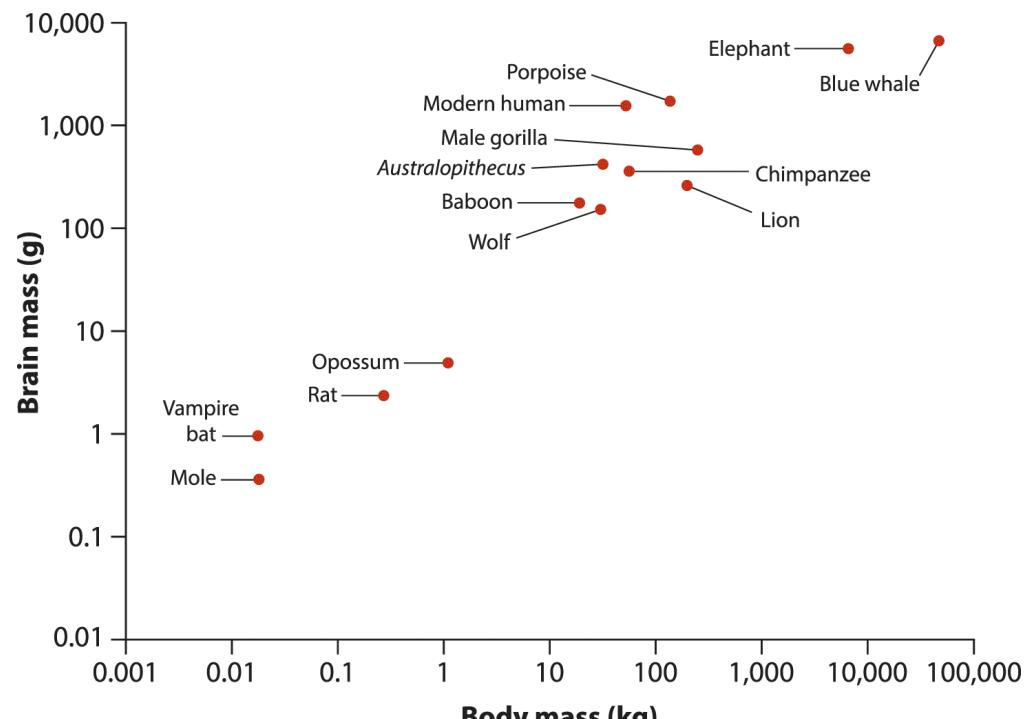
- Correlation 相关性
 - when two numerical variables are **associated**, we say that they are **correlated**.
 - 如果两个数值变量是关联的（非独立的），那么我们就说它们是相关的。



Whitlock & Schlüter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.1 相关性的统计学含义

- 相关性的定义
 - the amount of “scatter” in a scatter plot of the two variables
 - 两个变量变化的线性相关程度
- 相关性的量化
 - 相关性的强度 the strength
 - X 和 Y 的关联程度
 - 相关性的方向 the direction
 - X 和 Y 变化是否一致?



Whitlock & Schlüter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.2 相关性的量化

- 相关性的量化: Pearson相关系数 (Pearson's correlation coefficient)

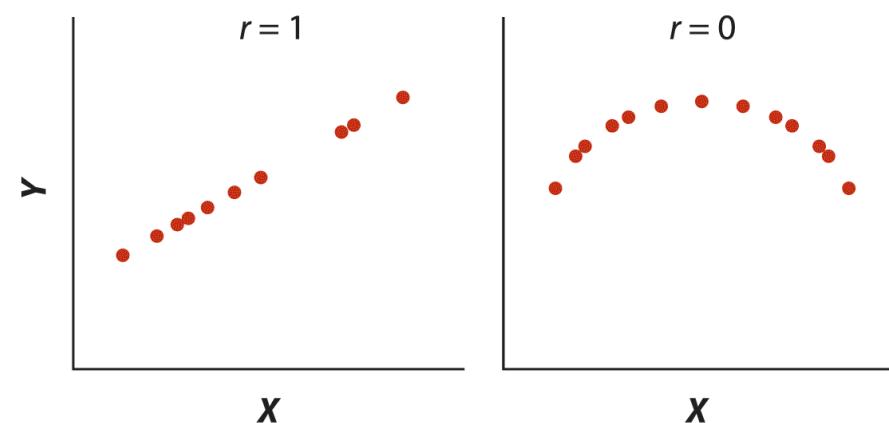
总体: ρ ← 样本: $r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{\text{Cov}(X, Y)}{s_X s_Y}$

- 其中, $\text{Cov}(X, Y)$ 为协方差 (covariance)

- $\text{Cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$
 - the product of two deviations



- 如果 X 和 Y 不相关, $\text{Cov}(X, Y) = ?$



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.2 相关性的量化

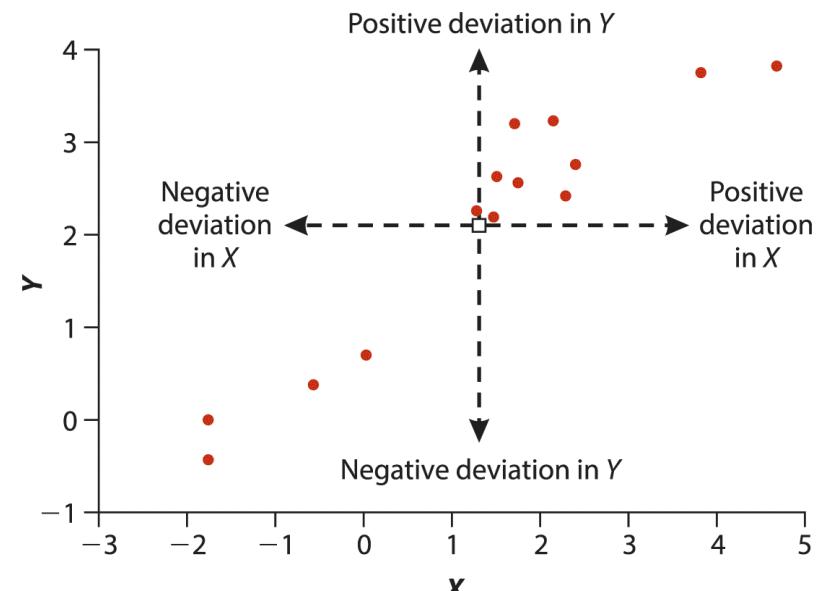
- 相关性的量化: Pearson相关系数 (Pearson's correlation coefficient)

总体: ρ ← 样本: $r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{\text{Cov}(X,Y)}{s_X s_Y}$

- 其中, $\text{Cov}(X,Y)$ 为协方差 (covariance)

- $\text{Cov}(X,Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$

- 相相关性的方向: X 和 Y 变化是否一致?
 - 正相关 (+)
 - 负相关 (-)



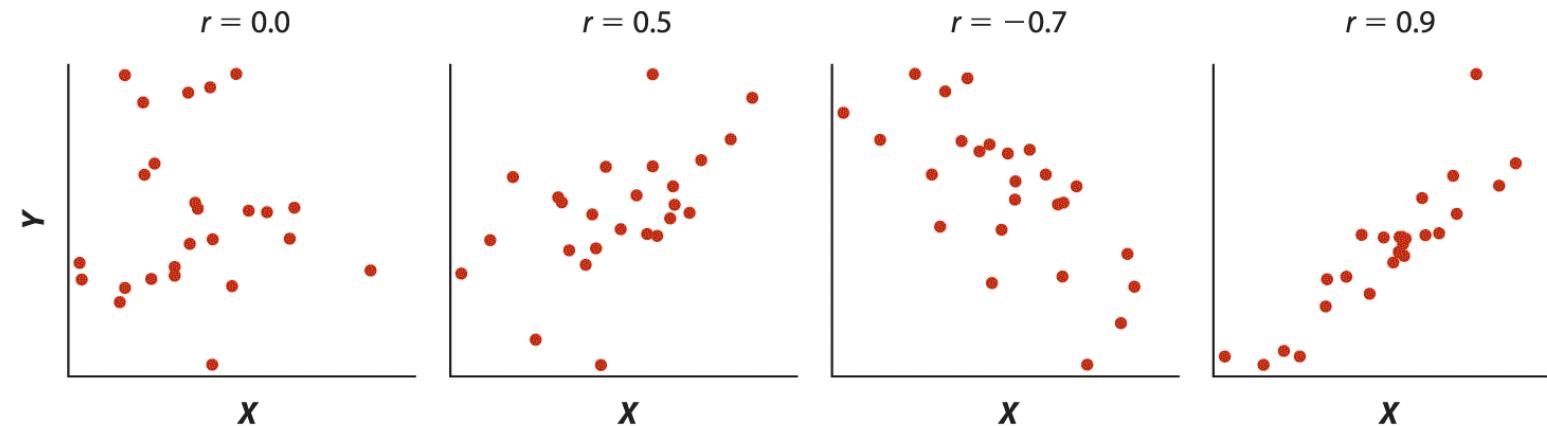
Whitlock & Schlüter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.2 相关性的量化

- 相关性的量化: Pearson相关系数 (Pearson's correlation coefficient)

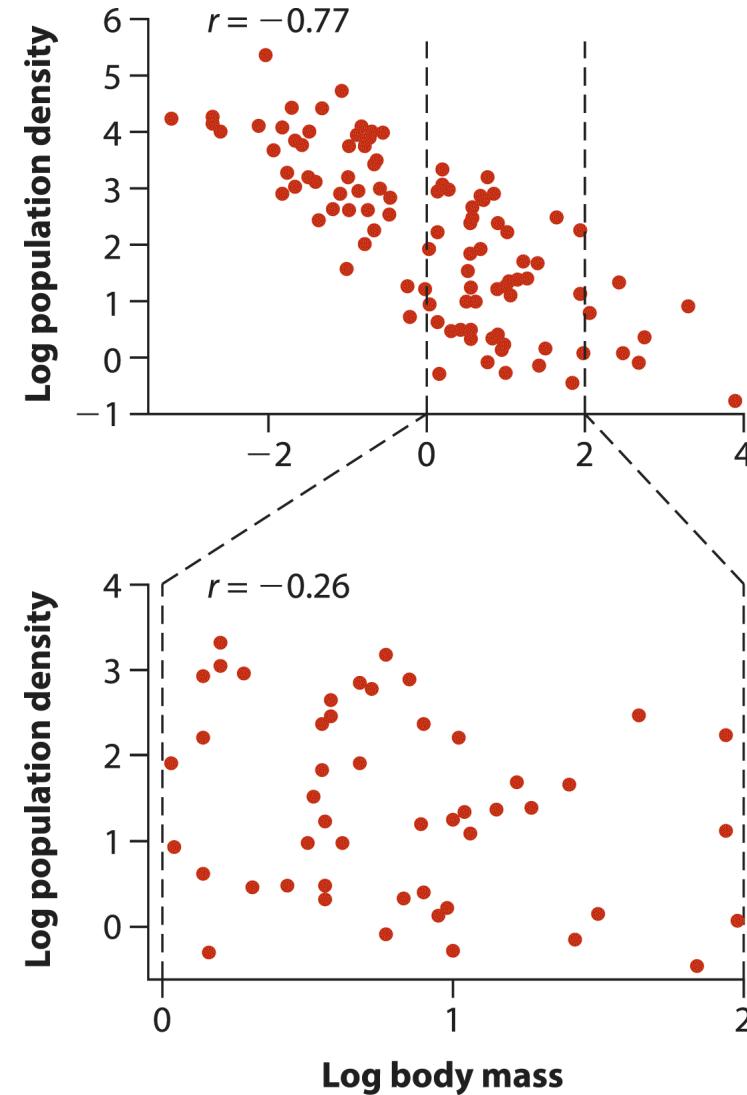
总体: ρ ← 样本: $r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{\text{Cov}(X, Y)}{s_X s_Y}$

- 相关的强度和方向 (无单位, no unit)



2.2 相关性的量化

- 相关性的量化的范围性
 - 两个变量 X 和 Y 之间的相关性取决于包含的数值范围;
- 例如
 - 河流中不同种类的无脊椎动物的种群密度 (Y) 与它们的体重 (X) 的相关性;



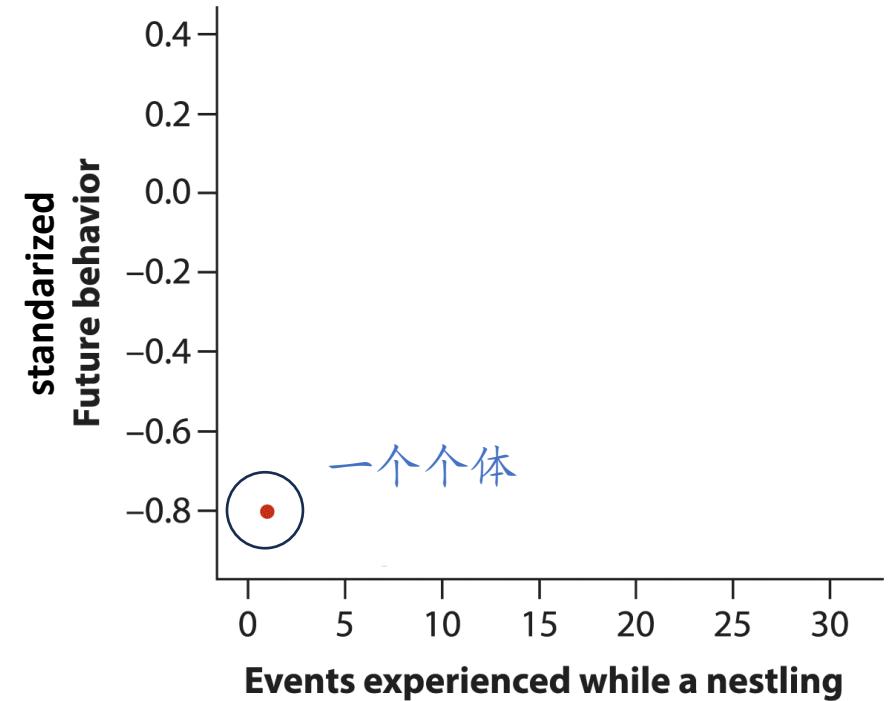
Whitlock & Schlüter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.3 相关性的估计和假设检验

- 从样本中计算相关系数，进行假设检验，并估计其置信区间
- 例子：鲣鸟之“原生家庭的影响” EXAMPLE 16.1: Flipping the bird (Nazca booby)
 - 问题：年幼时收到攻击的程度对未来的（攻击）行为有多大影响？
 - 攻击行为 (aggressive behavior)



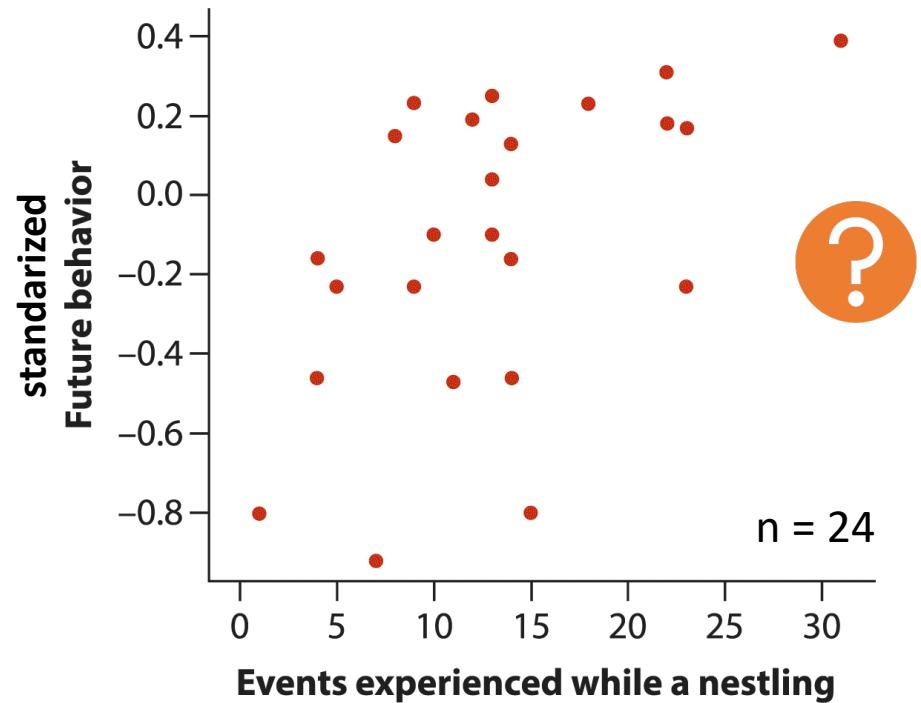
Adult on [España Island, Galapagos](#)



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.3 相关性的估计和假设检验

- 从样本中计算相关系数，进行假设检验，并估计其置信区间
- 例子：鲣鸟之“原生家庭的影响” EXAMPLE 16.1: Flipping the bird (Nazca booby)
 - 问题：年幼时收到攻击的程度对未来的（攻击）行为有多大影响？
 - 攻击行为 (aggressive behavior)
 - 测量值 X
 - 幼鸟阶段被非亲本鸟类攻击的次数；
 - 测量值 Y
 - 成鸟阶段去攻击非子代的幼鸟次数；



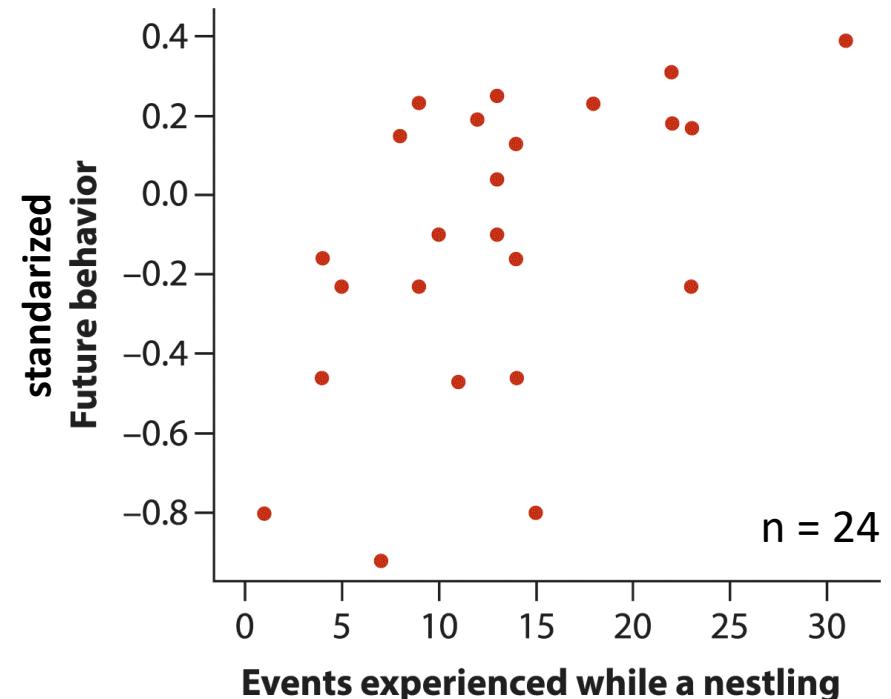
2.3 相关性的估计和假设检验

- 从样本中计算相关系数，进行假设检验，并估计其置信区间
- 例子：鲣鸟之“原生家庭的影响” EXAMPLE 16.1: Flipping the bird (Nazca booby)
 - 量化 X 和 Y 之间的相关性？

$$\begin{aligned} \bullet \quad r &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \\ &= \frac{33.086}{\sqrt{1194.625} \sqrt{3.217}} = 0.5337 \end{aligned}$$

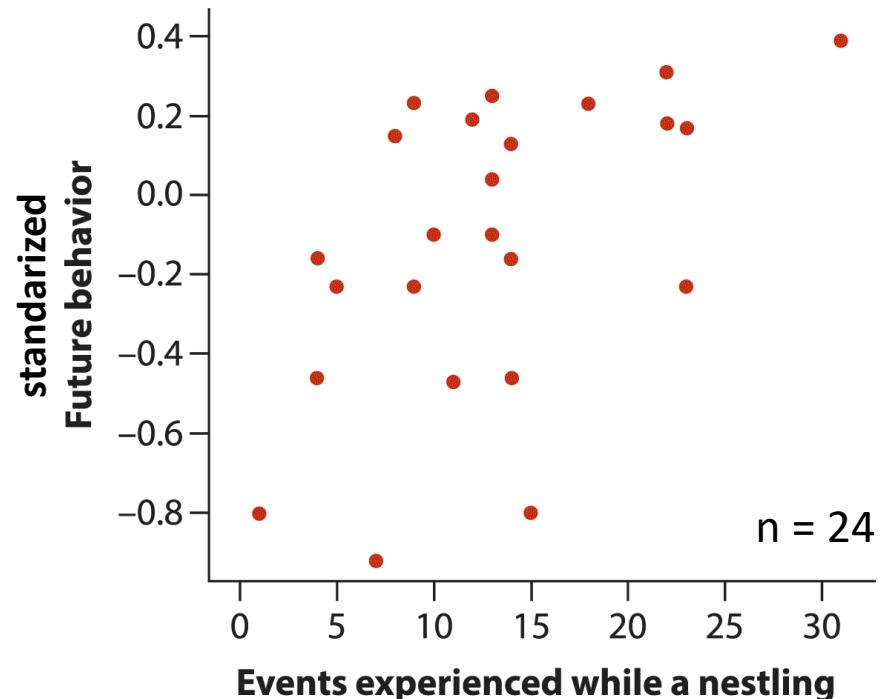
- 标准误 Standard Error

$$\bullet \quad SE_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0.5337^2}{24-2}} = 0.1803$$



2.3 相关性的估计和假设检验

- 从样本中计算相关系数，进行假设检验，并估计其置信区间
- 例子：鲣鸟之“原生家庭的影响” EXAMPLE 16.1: Flipping the bird (Nazca booby)
 - 假设检验：幼鸟阶段受到的攻击是否显著影响成鸟阶段的行为？
 - $H_0: \rho = 0$; 相关性和0没有显著差异;
 - $H_A: \rho \neq 0$; 相关性与0有显著差异;
 - 统计检验量 (t -test)
 - $t = \frac{r - \rho}{SE_r} = \frac{0.5337 - 0}{0.1803} = 2.960$
 - P值
 - $t > t_{critical} = t_{0.05(2), df=22} = 2.075$
 - $P = 0.0072$
 - 结论：拒绝零假设；两个变量显著相关；



2.3 相关性的估计和假设检验

- 相关系数 ρ 的置信区间的近似估计
 - 大样本时的近似值 (discovered by Fisher)
 - 数据转换 (\rightarrow 近似标准正态分布)

$$\bullet z = 0.5 \ln \left(\frac{1+r}{1-r} \right) = 0.5 \ln \left(\frac{1+0.5337}{1-0.5337} \right) = 0.0595$$

- 总体的 ρ 在 z -转换后的符号为 ζ
- 置信区间为:

$$\bullet z - 1.96 \sigma_z < \zeta < z + 1.96 \sigma_z, \text{ 其中 } \sigma_z = \sqrt{\frac{1}{n-3}} \text{ (近似的标准误计算)}$$

$$\bullet 0.168 < \zeta < 1.023 \text{ (其中 } \sigma_z = \sqrt{1/(24-3)} = 0.218 \text{)}$$

- 转换回原本的尺度:

$$\bullet r = \frac{e^{2z}-1}{e^{2z}+1} \text{ 或 } \rho = \frac{e^{2\zeta}-1}{e^{2\zeta}+1} \rightarrow 0.17 < \rho < 0.77 \text{ (很宽泛, 但不包括0)}$$



Adult on Española Island, Galapagos

(t-test)

$$\bar{Y} - t_{0.05(2), df} \text{SE}_{\bar{Y}} < \mu < \bar{Y} + t_{0.05(2), df} \text{SE}_{\bar{Y}}$$

2.3 相关性的估计和假设检验

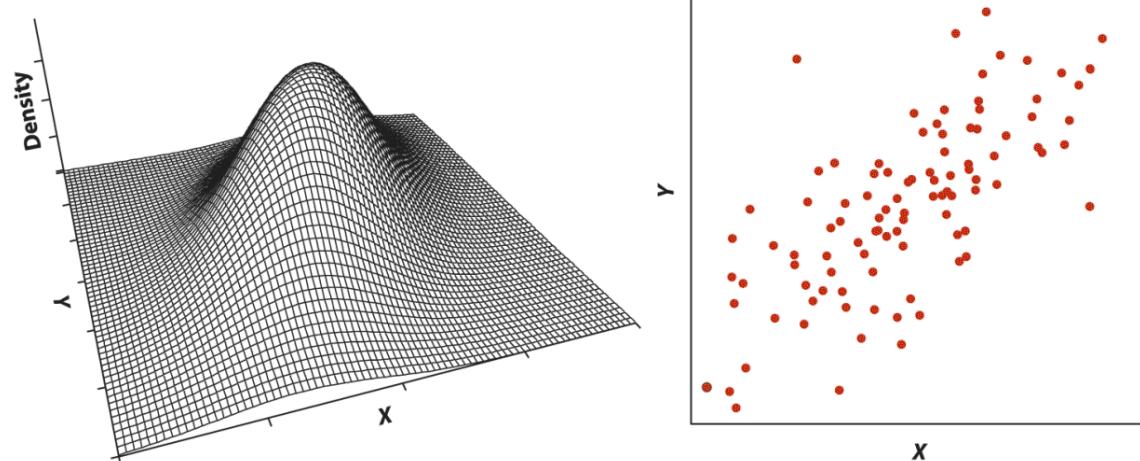
- 例子：鲣鸟之“原生家庭的影响” EXAMPLE 16.1: Flipping the bird (Nazca booby)
 - 相关系数 ρ 的估计、假设检验和置信区间的近似估计
 - $r = 0.5337$
 - $P = 0.0072 (< 0.05)$
 - $0.17 < \rho < 0.77$
 - R代码

```
boobyCor <- cor.test(~ futureBehavior + nVisitsNestling, data = booby)
boobyCor
```

```
##
## Pearson's product-moment correlation
##
## data: futureBehavior and nVisitsNestling
## t = 2.9603, df = 22, p-value = 0.007229
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval.
## 0.1660840 0.7710999
## sample estimates:
## cor
## 0.5337225
```

2.4 相关性分析的前提假设

- 前提假设 Assumption
 - 数据来自从总体中抽样而来的随机样本;
 - 数据服从二元正态分布 bivariate normal distribution
 - X和Y之间的关系是线性的;
 - 在X和Y的散点图中点的分布呈圆形或椭圆形;
 - X和Y的频数分布分别均是正态的;

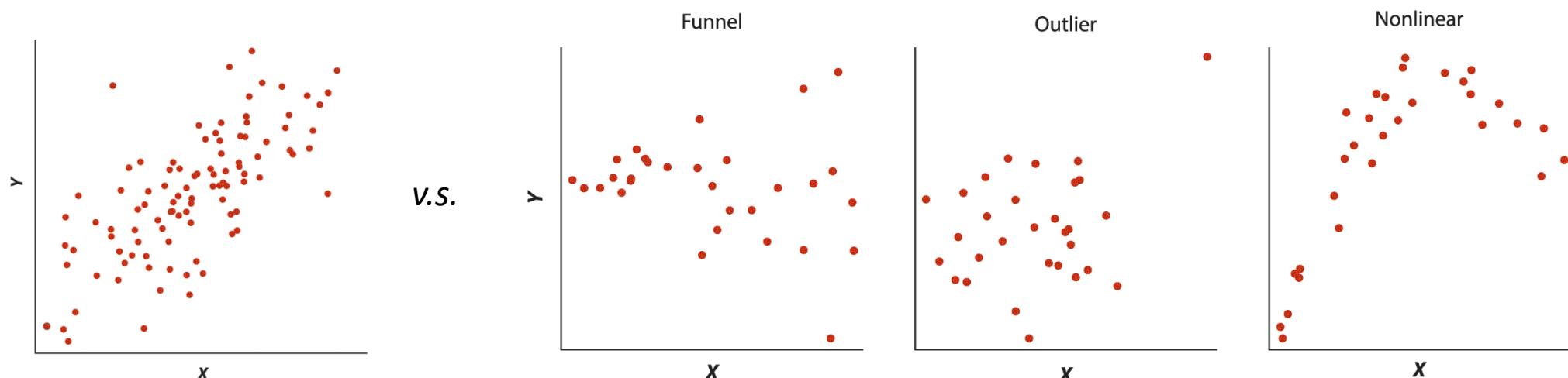


Whitlock & Schlüter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.4 相关性分析的前提假设

- 前提假设 Assumption

- 数据来自从总体中抽样而来的随机样本;
- 数据服从二元正态分布 bivariate normal distribution
 - X和Y之间的关系是线性的;
 - 在X和Y的散点图中点的分布呈圆形或椭圆形;
 - X和Y分别的频率分布是正态的;

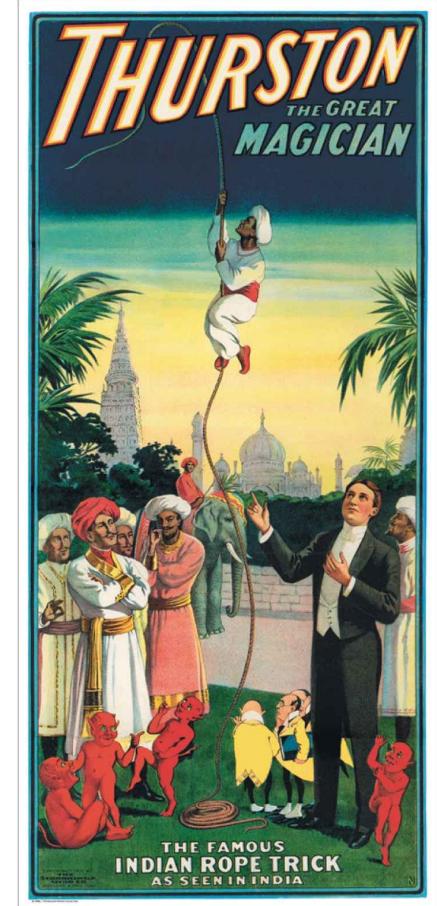
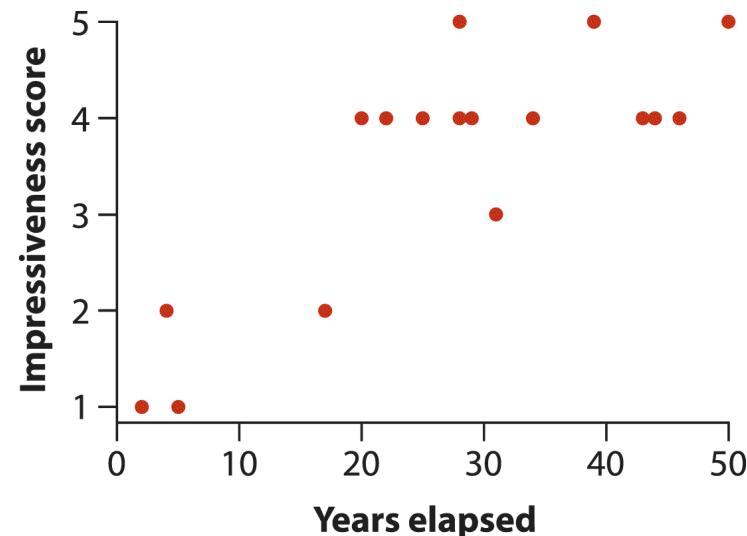


2.4 相关性分析的前提假设

- 不满足前提假设时
 - 数据转换
 - 从而满足二元正态分布
 - 非参方法： Spearman秩相关 (Spearman's rank correlation)
 - 衡量了两个变量的秩之间的线性关系的强度和方向；
 - 不对变量的分布做出假设，但仍然假设个体是从总体中随机抽取的；
 - 同时假设两个变量的关系为单调的（monotonic）
 - 其中 X 增大时， Y 增大或不变；
 - 或， X 减小时， Y 减小或不变；
 - 即两者之间的关系为线性的；

2.5 相关性分析的非参方法：Spearman秩相关

- 衡量了两个变量的秩之间的线性关系的强度和方向；
 - 例子：有关奇迹的回忆 EXAMPLE 16.5: The miracles of memory
 - 数据X: 目击到魔术的日期和记下这个记忆的日期之间经过的年数；
 - 数据Y: 对印度绳子魔术的第一手书面描述的印象评分；
 - 1分：“男孩爬上绳子，然后又爬下来”
 - 5分：“男孩爬上绳子，在顶端小时，然后重新出现在篮子中”



Billy Rose Theatre Division, The New York Public Library.

2.5 相关性分析的非参方法：Spearman秩相关

- 数据X和Y各自进行排序——秩 rank
 - 如果遇到相同的原始值，怎么矫正秩？



$$\bullet (3+4)/2 = 3.5$$

$$\bullet (1+2+3)/3 = 2$$

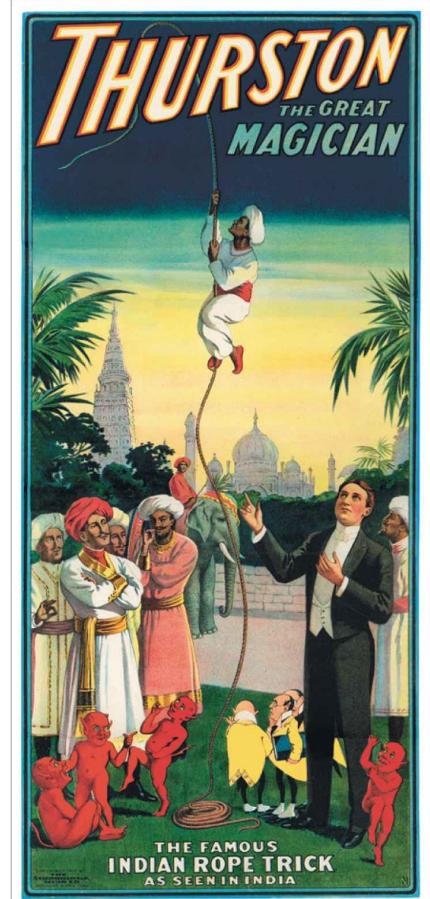
	years	year_rank	impressivenessScore	score_rank
1	2	1	1	1
2	5	3	1	2
3	5	4	1	3
4	4	2	2	4
5	17	5	2	5
6	17	6	2	6
7	31	13	3	7
8	20	7	4	8
9	22	8	4	9
10	25	9	4	10
11	28	10	4	11
12	29	12	4	12
13	34	14	4	13
14	34	15	4	14
15	43	17	4	15
16	44	18	4	16
17	46	19	4	17
18	28	11	5	18
19	39	16	5	19
20	50	20	5	20
21	50	21	5	21

2.5 相关性分析的非参方法：Spearman秩相关

- 衡量了两个变量的秩之间的线性关系的强度和方向；
 - 例子：有关奇迹的回忆 EXAMPLE 16.5: The miracles of memory
 - 数据X和Y各自进行排序，计算Spearman相关系数

$$r_s = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2} \sqrt{\sum_{i=1}^n (S_i - \bar{S})^2}} = \frac{566}{\sqrt{767.5} \sqrt{678.5}} = 0.784$$

- 其中， R_i 表示年份间隔的秩， S_i 表示印象评分的秩；
- 检验统计量： $t = \frac{r_s}{SE_{r_s}}$ ， 其中 $SE_{r_s} = \sqrt{\frac{1-r_s^2}{n-2}}$



Billy Rose Theatre Division, The New York Public Library.

2.5 相关性分析的非参方法：Spearman秩相关

- 衡量了两个变量的秩之间的线性关系的强度和方向；
 - 例子：有关奇迹的回忆 EXAMPLE 16.5: The miracles of memory
 - 数据X和Y各自进行排序，计算Spearman相关系数
 - $r_s = 0.784$

```
cor.test(~ years + impressivenessScore, data = trick, method = "spearman")
```

```
## Warning in cor.test.default(x = c(2L, 5L, 5L, 4L, 17L, 17L, 31L, 20L,
## 22L, : Cannot compute exact p-value with ties
```

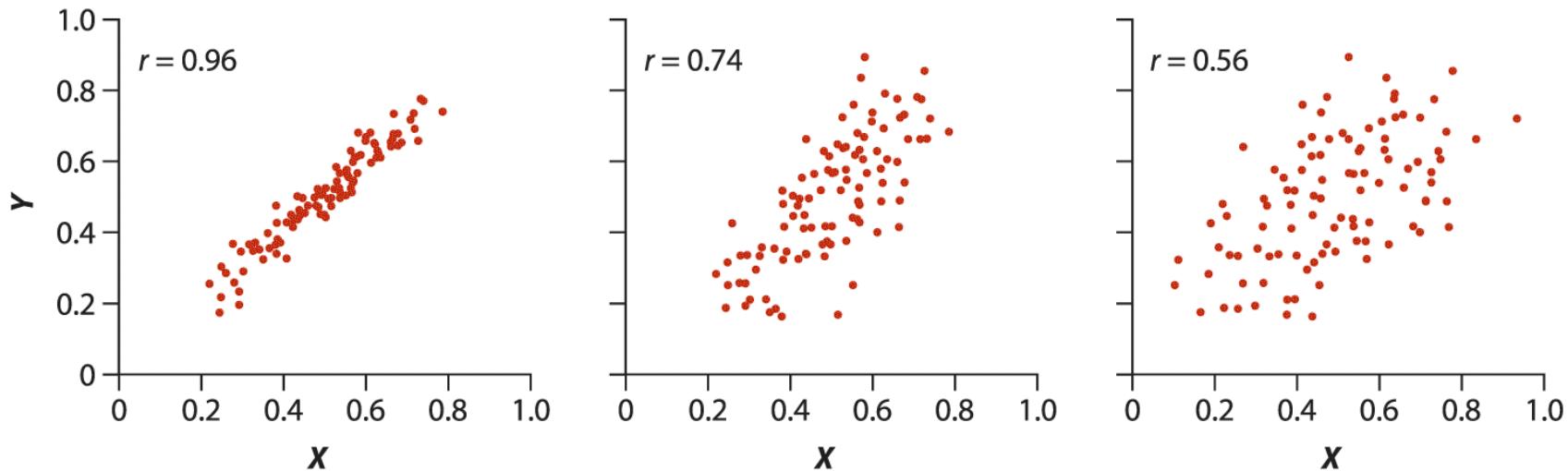
```
##
## Spearman's rank correlation rho
##
## data: years and impressivenessScore
## S = 332.12, p-value = 2.571e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##           rho
## 0.7843363
```

2.6 测量误差对相关性分析的影响

- 测量误差 measurement error
 - 个体的变量的真实值与其测量值之间的差异;
 - 当一个变量的测量不完美时，我们就说存在测量误差;
 - 测量误差可能是数据变异的一个重要组成部分;
 - 见随机效应方差分析 (random-effects ANOVA)
 - 例如，行为特征以低可重复性而著称:
 - 一个个体的行为测量可能在这一次和下一次测量时有很大的不同;

2.6 测量误差对相关性分析的影响

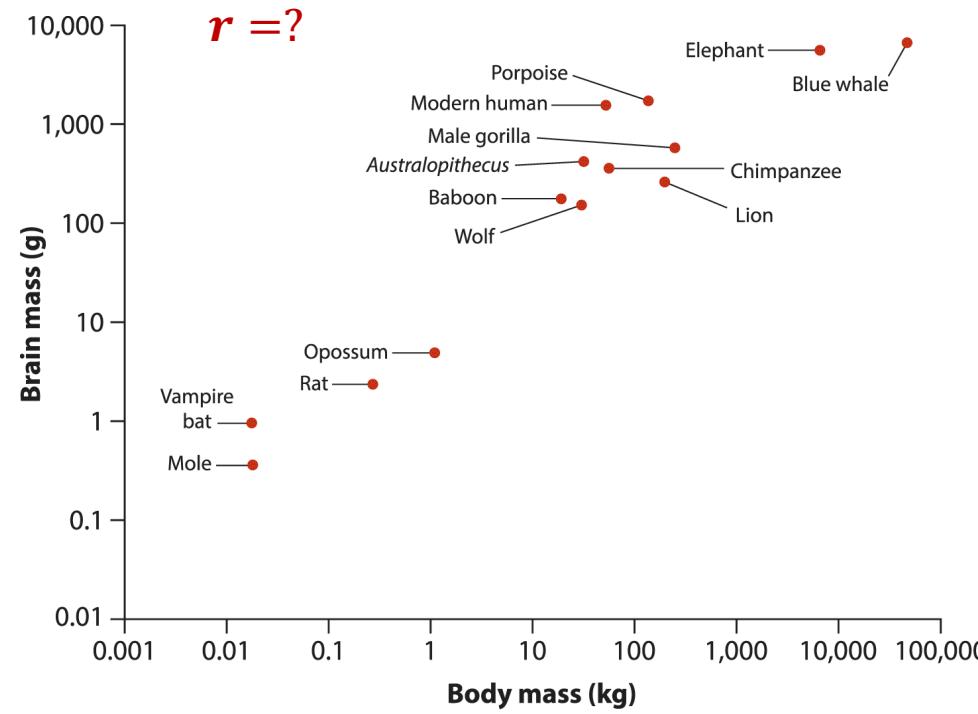
- 测量误差 measurement error
 - X 或 Y 中的测量误差往往会影响变量之间的观察到的相关性；
 - 因此， r 倾向于低估 ρ 的大小，这种偏差称为衰减 (attenuation)；
 - 精确测量有助于降低测量误差，如重复测量一个个体的值有利于矫正 r ；
 - (Section 16.8: Quick Formula Summary)



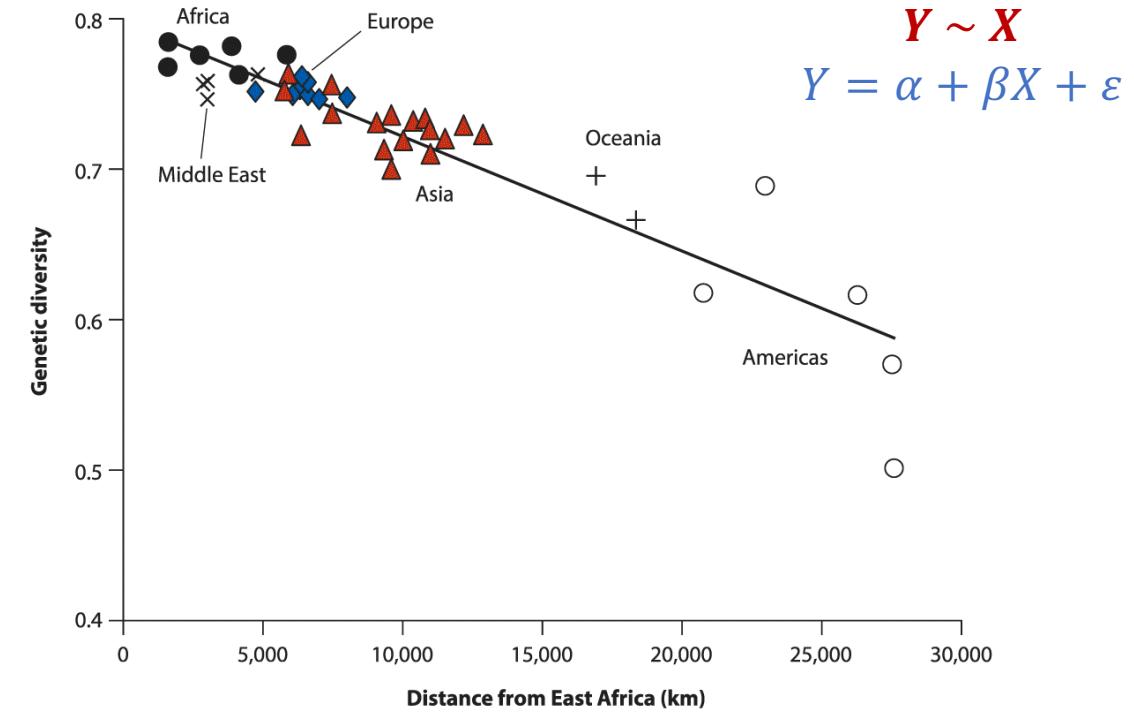


3. 简单线性回归 Simple Linear Regression

- 相关性分析和线性回归的异同？



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

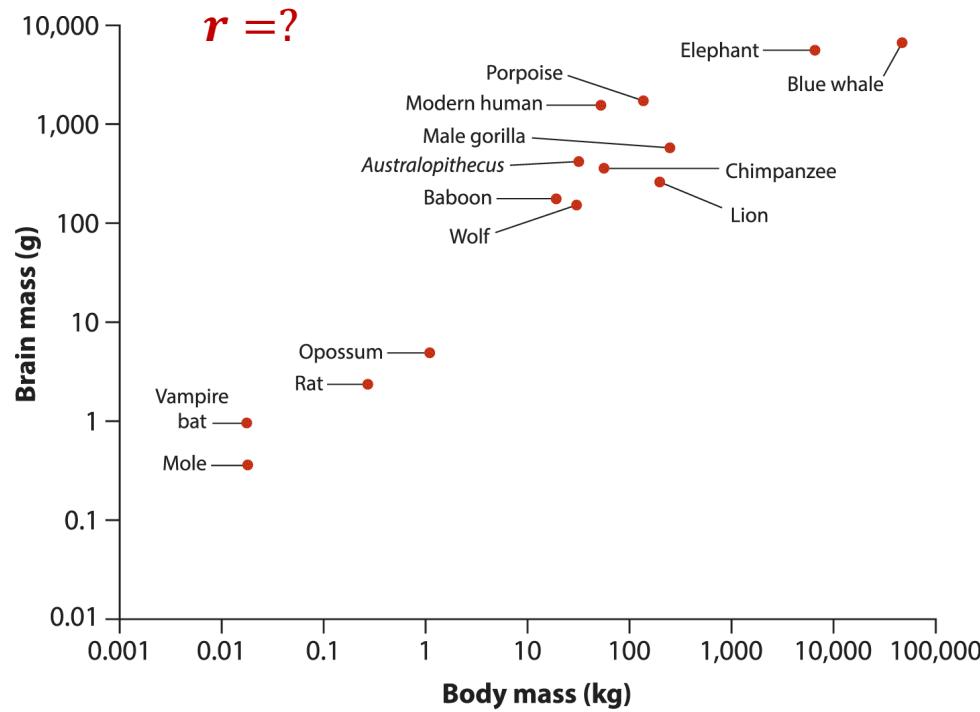


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

3.1 简单线性回归 Simple Linear Regression

- 相关性 correlation

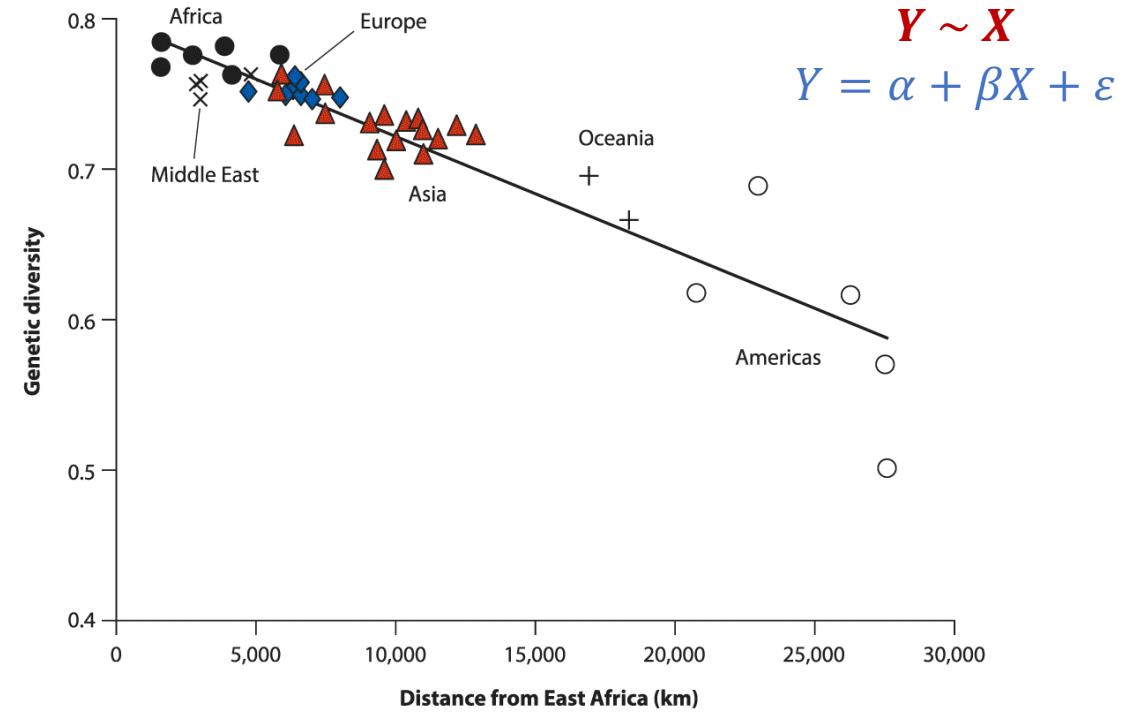
- 量化两个数值变量间的关联程度；
- 反应了数据的分散程度；



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

- 回归 regression

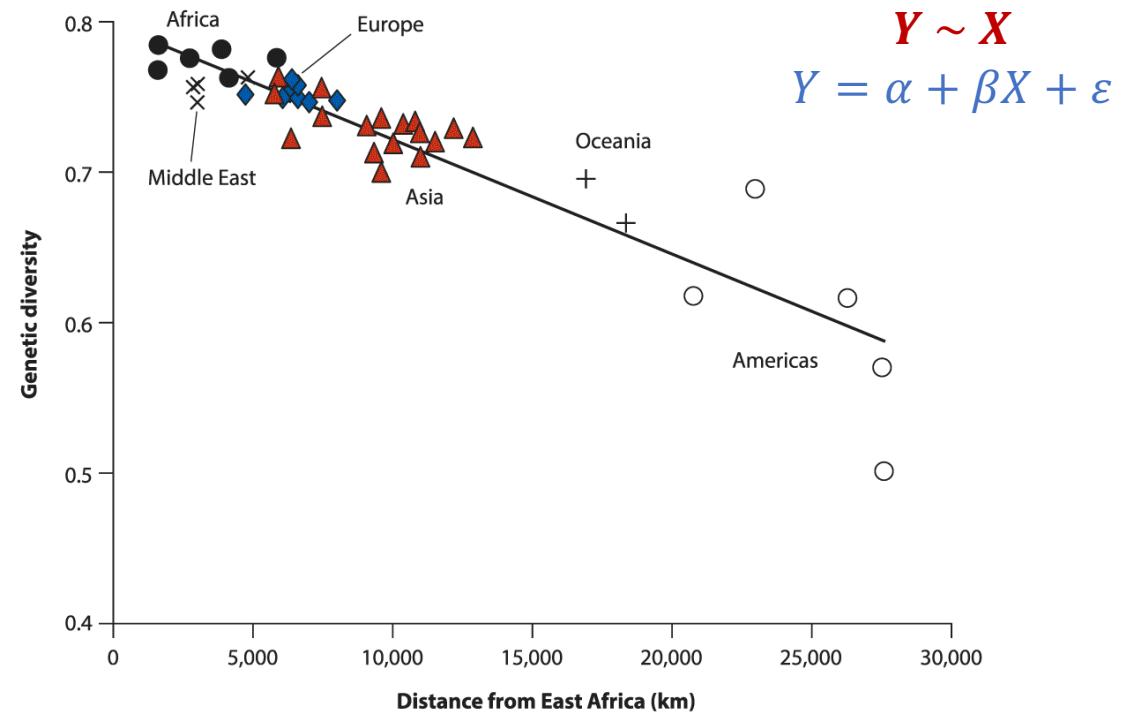
- 拟合回归线，可用以做预测；
- 斜率反应了 Y 随 X 变化的变化率；



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

3.1 简单线性回归 Simple Linear Regression

- 回归的目的
 - 是一种从一个数值变量的值预测另一个数值变量值的方法;
 - 线性回归拟合一条回归直线;
 - 其中回归线的斜率反应了 Y 随 X 变化的变化率;
 - the rate of change = slope: β
 - E.g., humans lose about 0.076 units of genetic diversity with every 10,000-km distance from East Africa.

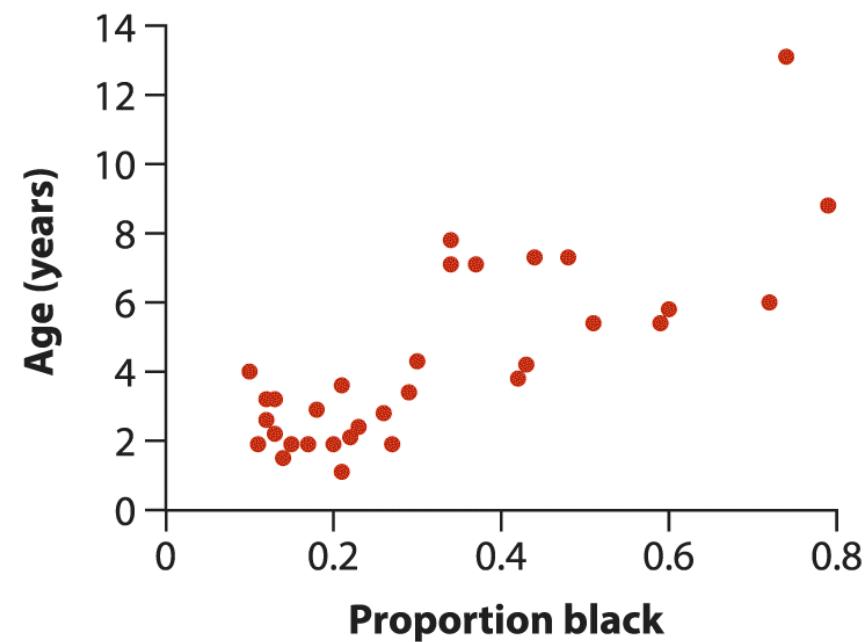


3.2 简单线性回归的拟合

- 线性回归的拟合 fit the linear regression
 - $Y = \alpha + \beta X + \varepsilon$
- 例子: EXAMPLE 17.1: The lion's nose
 - 维持狮群可行性
 - 管理非洲狮的狩猎活动
 - 通过鼻子来判断年龄 (移除老龄雄狮)
 - 样本: 坦桑尼亚32只雄性狮子
 - 数据:
 - 解释变量: 鼻子上黑色色素的比例
 - 响应变量: 年龄 (已知)
 - 目标: 构建两者之间的关系
 - 用鼻子黑色比例来预测年龄;



Deborah Kolb/Shutterstock.com

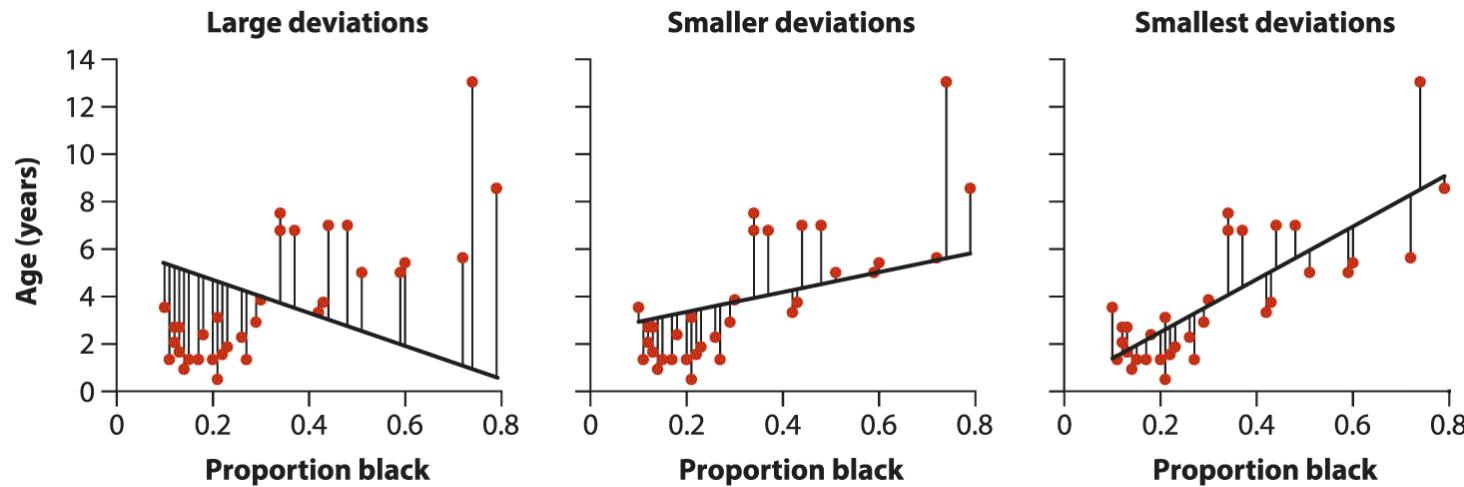


Whitlock & Schluter, *The Analysis of Biological Data*, 3e
© 2020 W. H. Freeman and Company

3.2 简单线性回归的拟合

- 拟合方法：最小二乘法 (the method of least squares)
 - 总体: $Y = \alpha + \beta X + \varepsilon, \varepsilon \sim N(0, \sigma^2)$ ← 样本: $Y = a + bX$
- 目标: 找到一条线回归线, 使得X预测Y的精度最高;
 - Y的偏差: 垂直方向上数据点 (Y_i) 和回归线 (预测值 \hat{Y}) 之间的线;
 - 最小二乘回归线是使得Y的所有偏差平方和最小的线;

$$\sum \varepsilon = \sum_{i=1}^n (Y_i - \hat{Y})^2$$



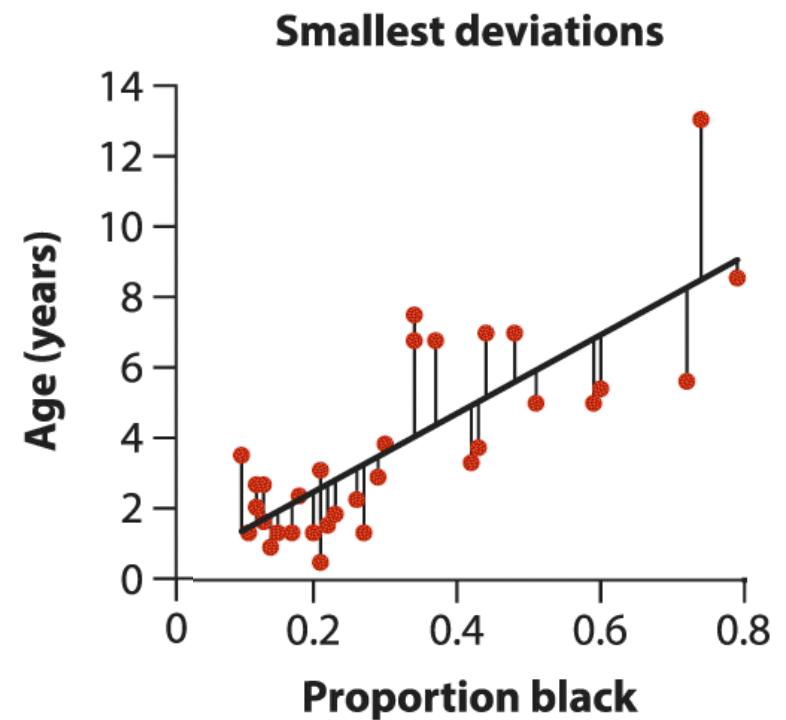
3.2 简单线性回归的拟合

- 最小二乘法 (the method of least squares)

$$\bullet Y = \alpha + \beta X + \varepsilon \leftarrow Y = a + bX$$

- 目标:

- X预测Y的精度最高;
- 即使得Y的所有偏差平方和最小;
- 对偏差进行平方处理的原因
 - 与计算普通方差时对偏差进行平方处理相同;
 - 因为在直接相加时, 为正值的偏差 (在回归线上方的点) 和为负值的偏差 (在回归线下方的点) 会互相抵消;



3.2 简单线性回归的拟合

- 最小二乘法 (the method of least squares)

- $$Y = \alpha + \beta X + \varepsilon \leftarrow Y = a + bX$$

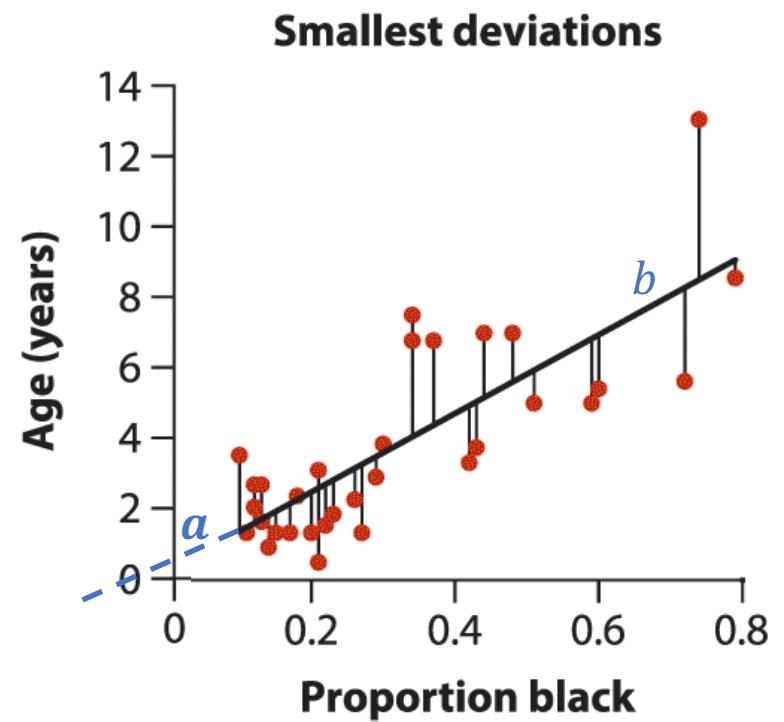
- 参数:

- α : 截距 intercept

- α : 截距 intercept
 - 当 X 为零时 a 的值就是 Y 的值;
 - 即回归线与 y 轴 “相交” 处的 Y 值;
 - 其单位与 Y 一致;

- b : 回归系数 regression coefficient

- b : 回归系数 regression coefficient
 - 即回归线的斜率 slope;
 - 衡量 Y 在 X 单位变化时的变化量;
 - 其单位是 Y 和 X 单位的比;



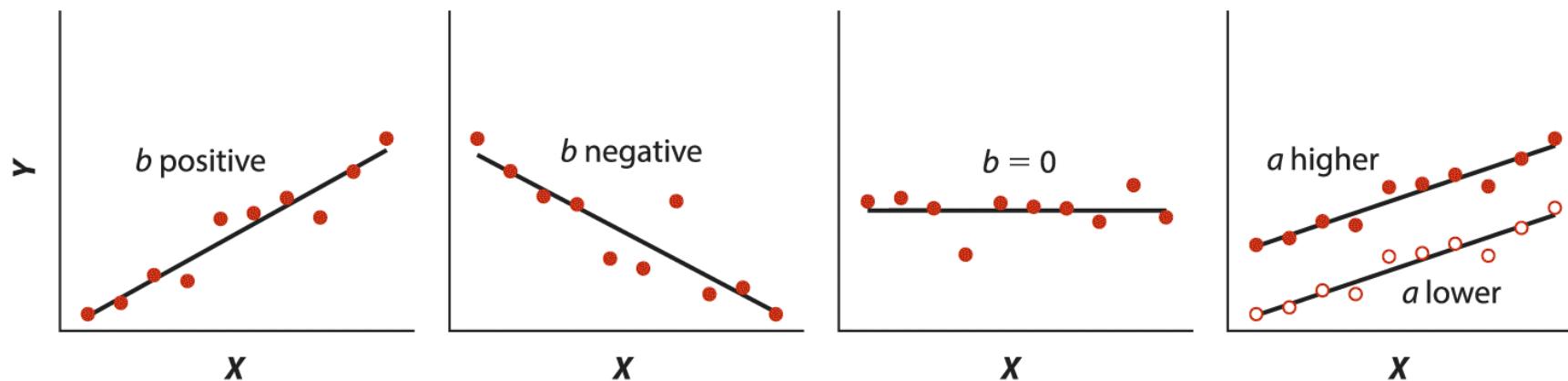
3.2 简单线性回归的拟合

- 最小二乘法 (the method of least squares)

- $Y = \alpha + \beta X + \varepsilon \leftarrow Y = a + bX$

- 参数:

- a : 截距
- b : 回归系数



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

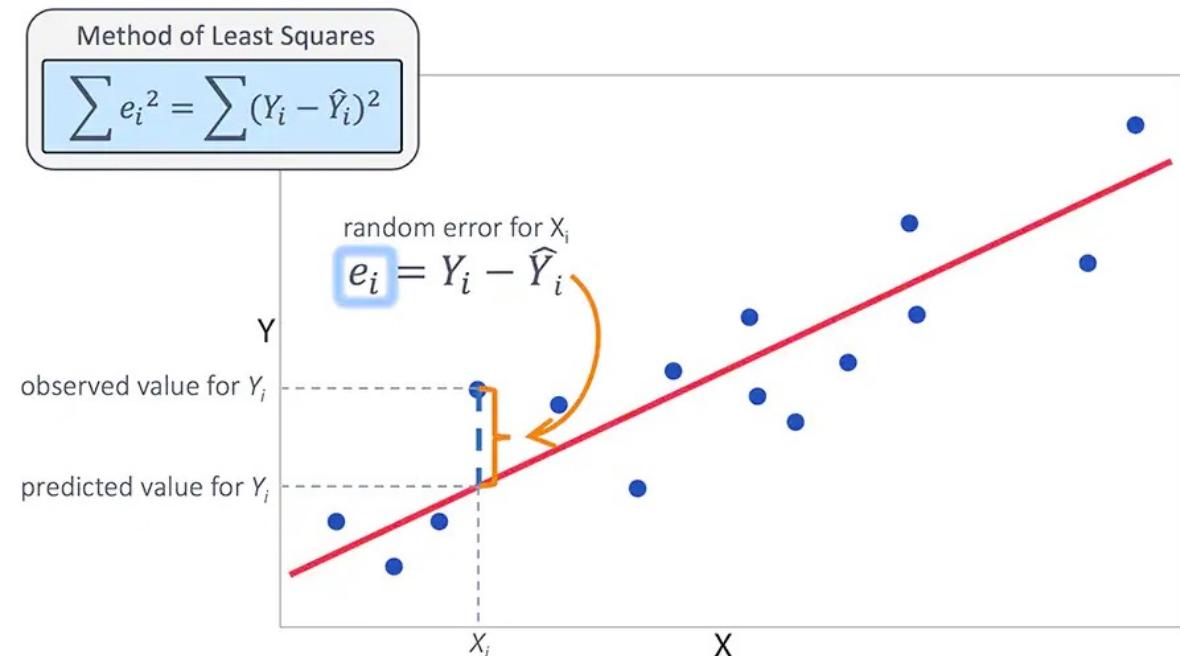
3.2 简单线性回归的拟合

- 最小二乘法: $Y = a + bX$
 - the ordinary least squares (OLS)
 - 目标: Y 的所有偏差平方和最小;
 - 即 $\sum \varepsilon^2$ 最小;
 - $\min \sum \varepsilon_i^2$

$$\text{观测值: } Y_i = a + bX_i + \varepsilon_i$$

$$\text{预测值: } \hat{Y}_i = a + bX_i$$

$$\text{偏差: } \varepsilon_i = Y_i - \hat{Y}_i$$



3.2 简单线性回归的拟合

- 最小二乘法: $Y = a + bX$
 - the ordinary least squares (OLS)
 - 目标: Y 的所有偏差平方和最小;
 - 即 $\sum \varepsilon^2$ 最小;
 - $\min \sum \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
 - 即 squared error sum 最小
 - $S(a, \beta) = \sum \varepsilon_i^2 = \sum (Y_i - a + bX_i)^2$

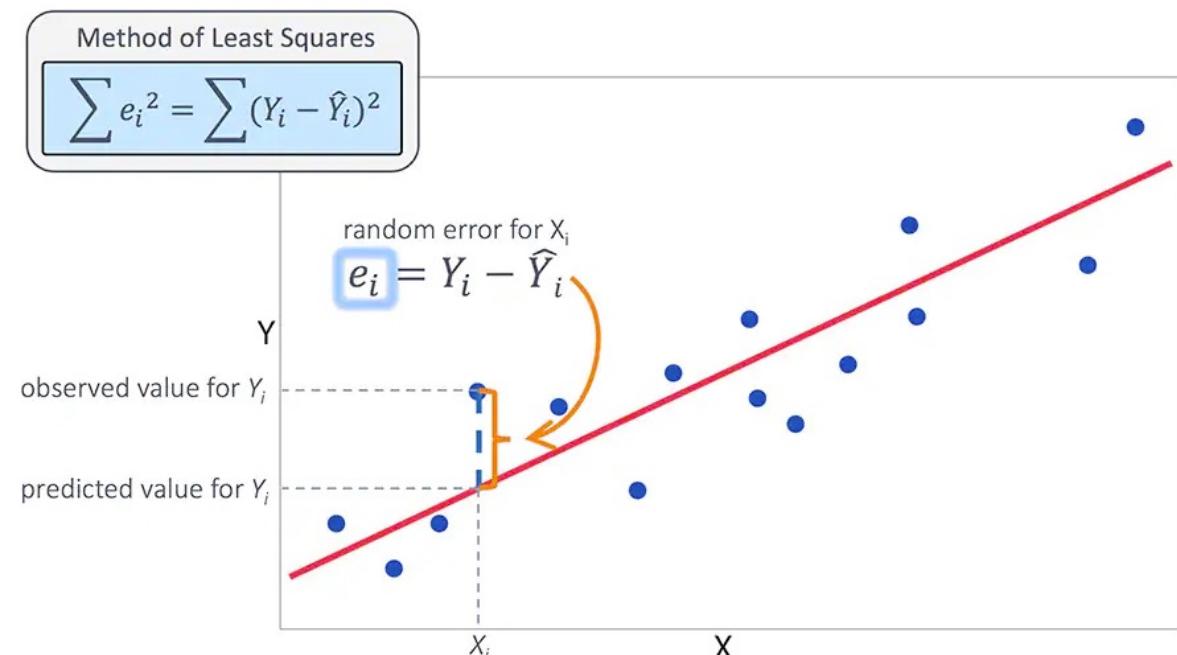
$$\hat{\beta}: \frac{\partial S(\hat{\alpha}, \hat{\beta})}{\partial \hat{\beta}} = 0$$

$$\hat{\alpha}: \frac{\partial S(\hat{\alpha}, \hat{\beta})}{\partial \hat{\alpha}} = 0$$

$$\text{观测值: } Y_i = a + bX_i + \varepsilon_i$$

$$\text{预测值: } \hat{Y}_i = a + bX_i$$

$$\text{偏差: } \varepsilon_i = Y_i - \hat{Y}_i$$



3.2 简单线性回归的拟合

- 最小二乘法: $Y = a + bX$

- the ordinary least squares (OLS)
- 目标: Y 的所有偏差平方和最小;
- 即 $\sum \varepsilon^2$ 最小;
- $\min \sum \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
 - 即 squared error sum 最小
 - $S(a, \beta) = \sum \varepsilon_i^2 = \sum (Y_i - a + bX_i)^2$

$$\hat{\beta}: \frac{\partial S(\hat{\alpha}, \hat{\beta})}{\partial \hat{\beta}} = 0$$

$$\hat{\alpha}: \frac{\partial S(\hat{\alpha}, \hat{\beta})}{\partial \hat{\alpha}} = 0$$

$$\begin{aligned}& \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} * x_i) x_i = 0 \\& \Rightarrow \sum_{i=1}^n y_i x_i - \hat{\alpha} x_i - \hat{\beta} x_i^2 = 0 \\& \Rightarrow \sum_{i=1}^n y_i x_i - (\bar{y} - \hat{\beta} \bar{x}) x_i - \hat{\beta} x_i^2 = 0 \\& \Rightarrow \sum_{i=1}^n y_i x_i - \bar{y} x_i + \hat{\beta} \bar{x} x_i - \hat{\beta} x_i^2 = 0 \Rightarrow \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta} \bar{x} - \hat{\beta} x_i) x_i = 0 \\& \Rightarrow \sum_{i=1}^n y_i - \bar{y} + \hat{\beta} (\bar{x} - x_i) = 0 \Rightarrow \sum_{i=1}^n (y_i - \bar{y}) = -\hat{\beta} \sum_{i=1}^n (\bar{x} - x_i) \\& \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})} \\& \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}$$

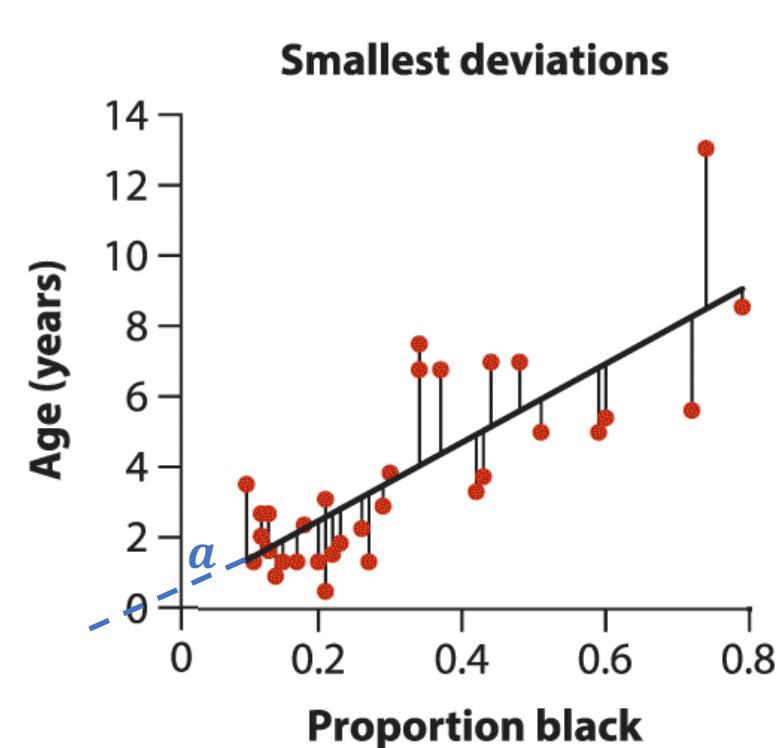
3.2 简单线性回归的拟合

- 最小二乘法: $Y = a + bX$
- 参数:

- 回归系数: $b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$
- 截距: $a = \bar{Y} - b\bar{X}$

- 例子: EXAMPLE 17.1: The lion's nose
 - $b = \frac{13.0123}{1.2221} = 10.647$
 - $a = 4.3094 - 10.647(0.3222) = 0.879$

proportionBlack	ageInYears
0.21	1.1
0.14	1.5
0.11	1.9
0.13	2.2
0.12	2.6
0.13	3.2
...	...



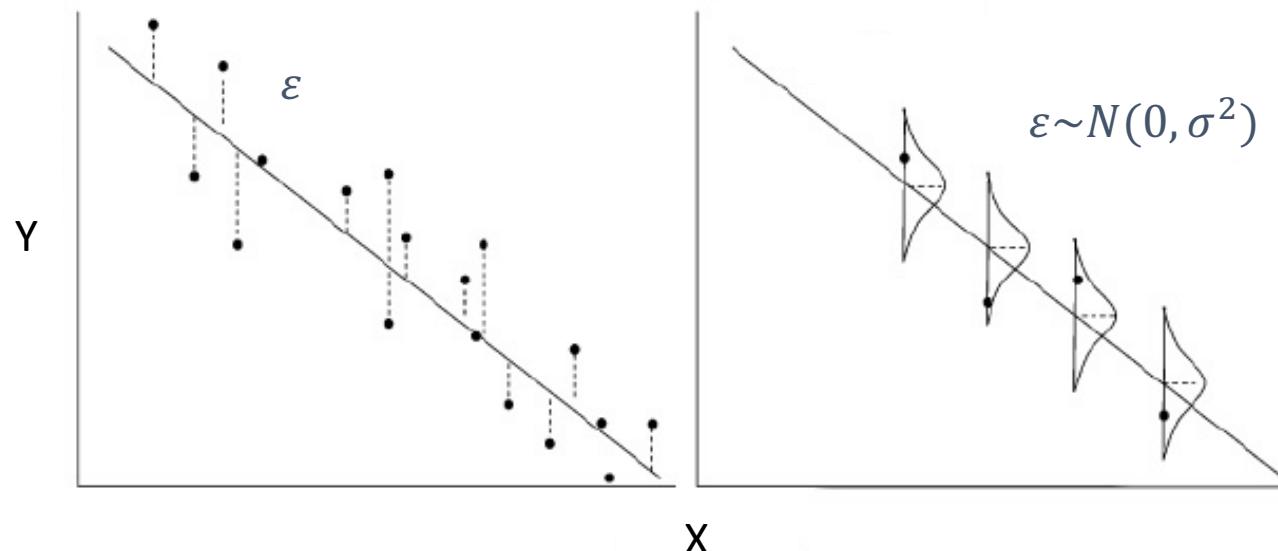
3.2 简单线性回归的拟合

- 最小二乘法

- 总体: $Y = \alpha + \beta X + \varepsilon, \varepsilon \sim N(0, \sigma^2)$ ← 样本: $Y = a + bX$

- 预测值: $\hat{Y}_i = a + bX_i$

- 可确定在指定的X值处所对应的回归线上的Y值;
- 该预测值估计了具有该X值的整个总体的Y的均值;



每个观测值 Y_i 落在了 X 为特定值 X_i 时的一个正态分布中;

3.2 简单线性回归的拟合

- 最小二乘法

- 总体: $Y = \alpha + \beta X + \varepsilon, \varepsilon \sim N(0, \sigma^2)$ ← 样本: $Y = a + bX$

- 预测值: $\hat{Y}_i = a + bX_i$

- 回归线的预测值估计了具有给定X值的所有个体的Y的均值;

$$Y = a + bX \\ = 0.879 + 10.647 X$$

- 例如:

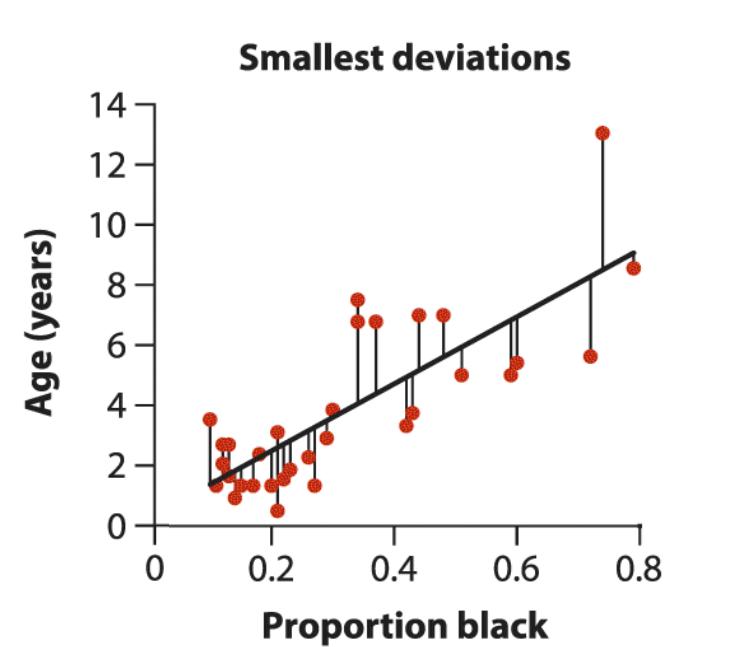
- 当 $X = 0.5$ (原始数据中没有这个值)

- $Y = a + bX = 0.879 + 10.647 \times 0.5 = 6.2$

- 即回归线预测:

- 鼻子黑色比例为0.5的狮子平均年龄为6.2岁;

- (即使我们以前从未见过完全相同的狮子)



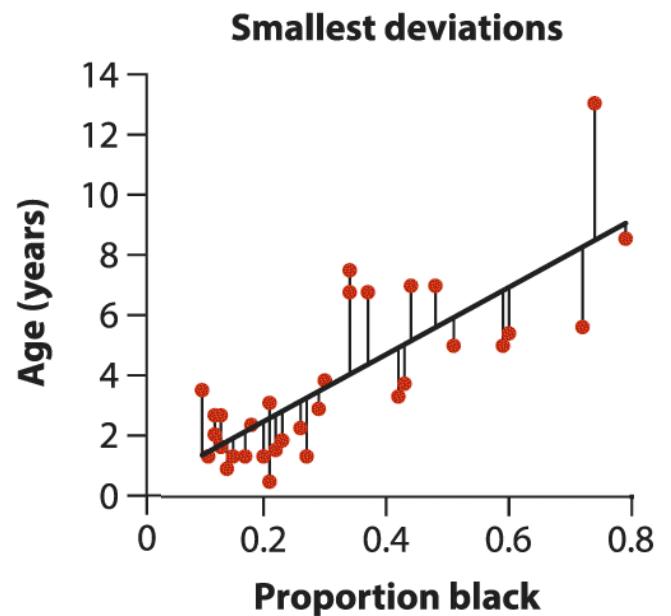
3.2 简单线性回归的拟合

- 最小二乘法 $Y = a + bX$
 - 预测值: $\hat{Y}_i = a + bX_i$
 - 残差 (偏差 ε): $residual_i = Y_i - \hat{Y}_i$
 - 即观测值与预测值的差异;
 - 残差均方: $MS_{residual} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum(Y_i - \bar{Y})^2 - b \sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-2}$
 - 回归系数 b (斜率) 的标准误: $SE_b = \sqrt{\frac{MS_{residual}}{\sum(X_i - \bar{X})^2}}$
 - 回归系数 b (斜率) 的95%置信区间 ($\alpha = 0.05$):
 - $b - t_{0.05(2), df} SE_b < \beta < b + t_{0.05(2), df} SE_b$
 - b 的抽样分布服从均值为 β 且标准差由斜率估计的标准误 SE_b 估计的正态分布;

3.2 简单线性回归的拟合

- 最小二乘法 $Y = a + bX$
 - 预测值: $\hat{Y}_i = a + bX_i$
 - 残差 (偏差 ε): $residual_i = Y_i - \hat{Y}_i$
- 例子: EXAMPLE 17.1: The lion's nose
 - 残差均方: $MS_{residual} = 2.785$
 - 回归系数 b (斜率) 的标准误: $SE_b = \sqrt{\frac{2.785}{1.221}} = 1.510$
 - 回归系数 b (斜率) 的 95% 置信区间 ($\alpha = 0.05$):
 - $df = n - 2 = 30$
 - $10.647 - 2.042 \times 1.150 < \beta < 0.647 + 2.042 \times 1.150$
 - $7.56 < \beta < 13.73$

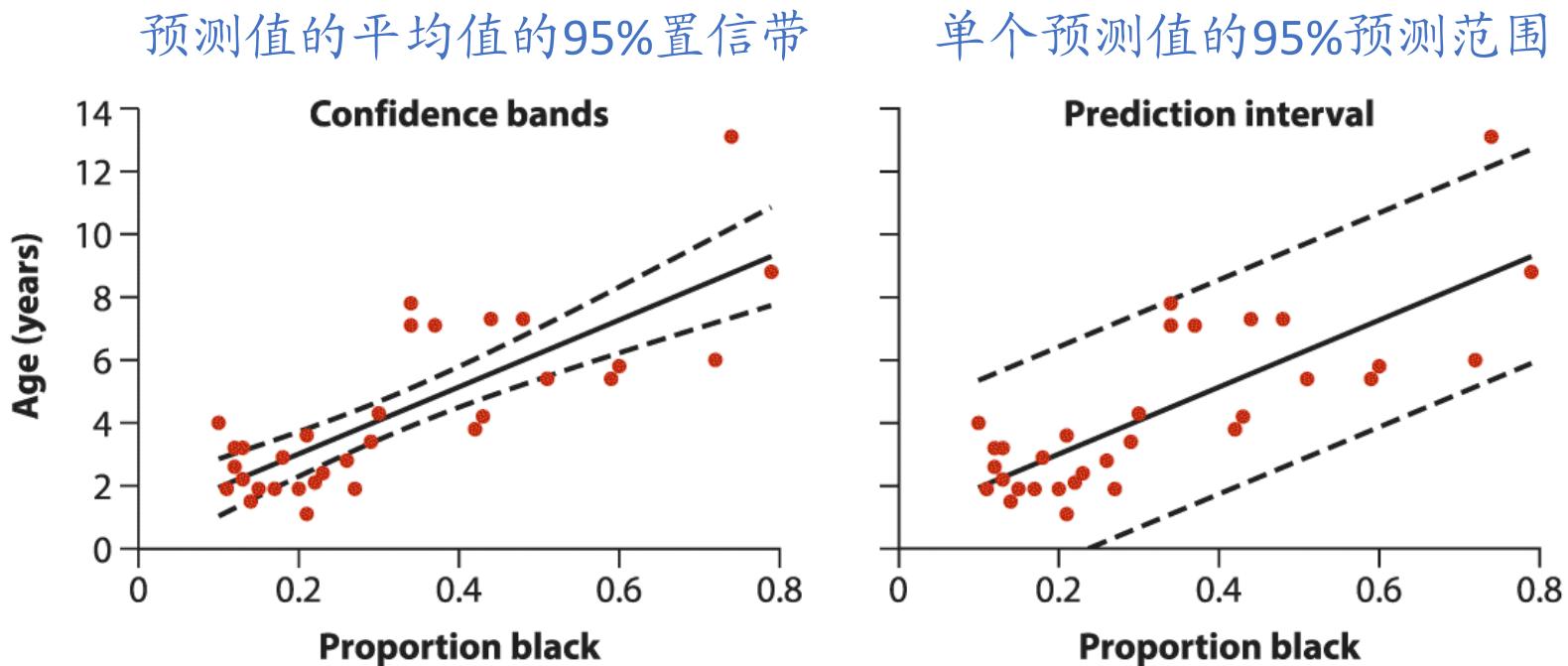
$$Y = a + bX \\ = 0.879 + 10.647 X$$



3.2 简单线性回归的拟合

- 预测

- 从鼻子的颜色预测一个狮子个体的年龄（右图）比预测所有具有相同鼻子黑色比例的狮子的平均年龄（左图）更不确定；

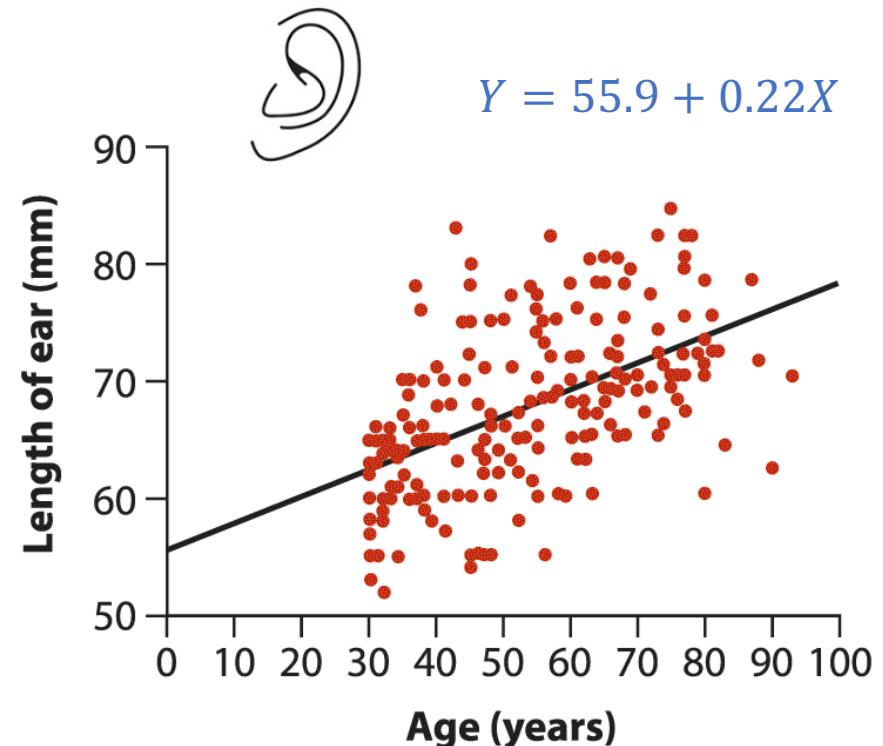


3.2 简单线性回归的拟合

- 预测
 - 预测个体值比预测均值更不确定（更宽泛的置信区间）；
 - (17.11 Quick formula summary)
 - 预测值的平均值的95%置信带 confidence bands
 - $\hat{Y} - t_{\alpha(2), df=n-2} \text{SE}_{\hat{Y}} < \text{predicted } Y < \hat{Y} + t_{\alpha(2), df=n-2} \text{SE}_{\hat{Y}}$
 - $\text{SE}_{\hat{Y}} = \sqrt{MS_{residual} \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right)}$
 - 单个预测值的95%预测范围 prediction interval
 - $\hat{Y} - t_{\alpha(2), df=n-2} \text{SE}_{1[\hat{Y}]} < \text{predicted } Y < \hat{Y} + t_{\alpha(2), df=n-2} \text{SE}_{1[\hat{Y}]}$
 - $\text{SE}_{1[\hat{Y}]} = \sqrt{MS_{residual} \left(1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right)}$

3.2 简单线性回归的拟合

- 外推 Extrapolation
 - 对数据范围之外的X值进行Y的预测被称为外推；
 - 当X值远远超出数据范围时，不能使用回归来预测响应变量的值；
 - 因为无法确保X和Y之间的关系在数据范围之外仍然是线性的；
 - E.g., age=0 → ear = 55.9mm

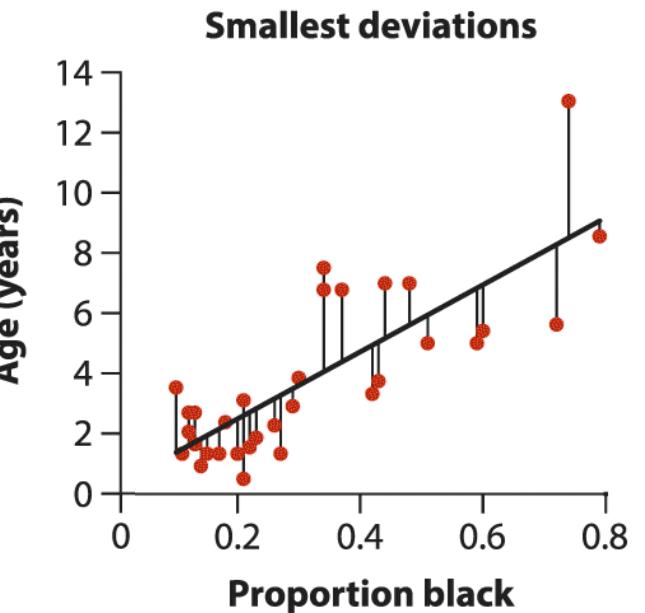


Whitlock & Schluter, *The Analysis of Biological Data*, 3e ©
2020 W. H. Freeman and Company

3.3 简单线性回归的假设检验

- 假设检验: $Y = a + bX \rightarrow$ 总体: $Y = \alpha + \beta X + \varepsilon$
 - 针对回归系数/斜率: 衡量 Y 在 X 单位变化时的变化量;
 - $H_0: \beta = \beta_0$; 回归线的斜率等于特定的值 β_0 ;
 - $H_A: \beta \neq \beta_0$; 回归线的斜率不等于特定的值 β_0 ;
 - 通常 $\beta_0 = 0$, 即 Y 不随 X 的变化而变化;
 - 统计检验量 (t -test)
 - $t = \frac{b - \beta_0}{SE_b} = \frac{10.647 - 0}{1.510} = 7.051$
 - P值
 - $t > t_{critical} = t_{0.05(2), df=22} = 2.075$
 - $P = 7.677e-08$
 - 结论: 拒绝零假设; 回归线斜率显著不等于 0;

$$Y = a + bX \\ = 0.879 + 10.647 X$$



3.3 简单线性回归的假设检验

- $H_0: \beta = \beta_0$; 回归线的斜率等于特定的值 β_0 ;
- $H_A: \beta \neq \beta_0$; 回归线的斜率不等于特定的值 β_0 ;

```
lionRegression <- lm(ageInYears ~ proportionBlack, data = lion)
```

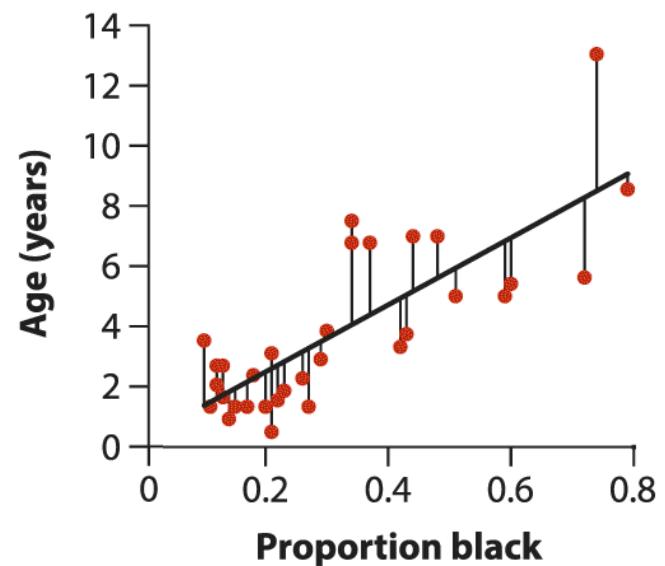
```
summary(lionRegression)
```

```
##  
## Call:  
## lm(formula = ageInYears ~ proportionBlack, data = lion)  
##  
## Residuals:  
##     Min      1Q  Median      3Q     Max  
## -2.5449 -1.1117 -0.5285  0.9635  4.3421  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  0.8790    0.5688   1.545   0.133  
## proportionBlack 10.6471   1.5095   7.053 7.68e-08 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 1.669 on 30 degrees of freedom  
## Multiple R-squared:  0.6238, Adjusted R-squared:  0.6113  
## F-statistic: 49.75 on 1 and 30 DF,  p-value: 7.677e-08
```

a → *b* →

$$Y = a + bX$$
$$= 0.879 + 10.647 X$$

Smallest deviations



3.3 简单线性回归的假设检验

- 方差分析的方法 the ANOVA approach
 - 回归均方：不同 X 值的数据点的 Y 值之间的变异量（组间）；
 - 残差均方：相同 X 值的数据点的 Y 值之间的变异量（组内）；
- 假设检验（当 $\beta_0 = 0$ ）
 - H_0 : 回归均方等于残差均方；
 - H_A : 回归均方不等于残差均方；

```
> anova(lionRegression)
Analysis of Variance Table

Response: ageInYears
              Df  Sum Sq Mean Sq F value    Pr(>F)
proportionBlack  1 138.544 138.544 49.751 7.677e-08 ***
Residuals      30  83.543   2.785
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} ? = 1$$

3.3 简单线性回归的假设检验

$$\bullet R^2 = \frac{SS_{\text{regression}}}{SS_{\text{total}}}$$

- 衡量了回归线对数据的拟合程度；即X对Y的数据变异的解释比例；

```
summary(lionRegression)
```

```
##  
## Call:  
## lm(formula = ageInYears ~ proportionBlack, data = lion)  
##
```

```
> anova(lionRegression)  
Analysis of Variance Table
```

Response: ageInYears

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
proportionBlack	1	138.544	138.544	49.751	7.677e-08 ***
Residuals	30	83.543	2.785		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
##
```

Residual standard error: 1.669 on 30 degrees of freedom

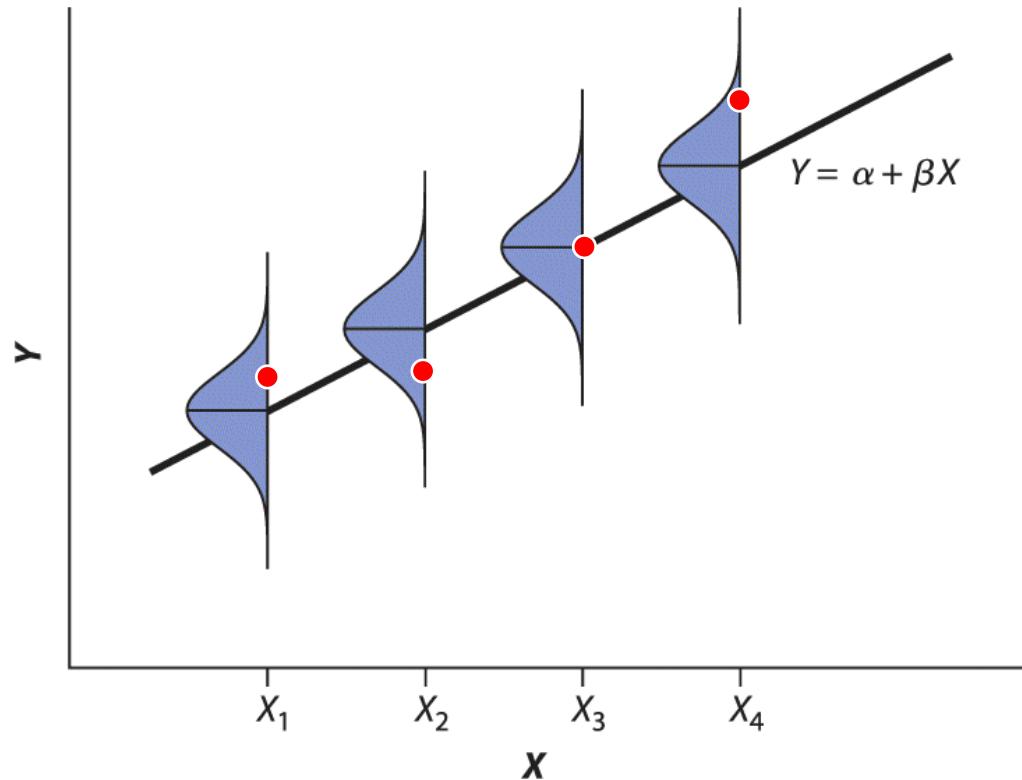
Multiple R-squared: 0.6238, Adjusted R-squared: 0.6113

F-statistic: 49.75 on 1 and 30 DF, p-value: 7.677e-08

3.4 简单线性回归的前提假设

- 前提假设 Assumptions

- 每个X值对应的所有可能的Y值遵从正态分布，其均值位于真实回归线上；
- 不同Y值的方差都相同；
- 在每个X值处的Y的测量值/观察值代表了可能的Y值总体的一个随机样本；
- 没有对X的正态性的要求；
- 也不要求数X是随机抽样而来；



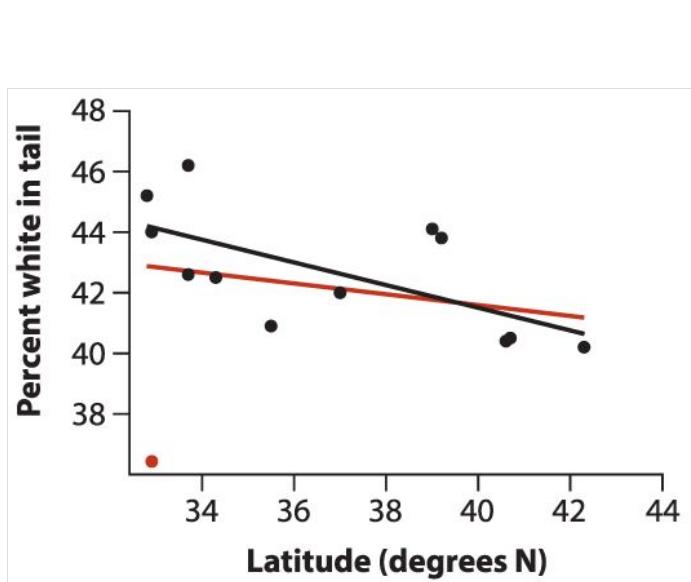
Whitlock & Schluter, *The Analysis of Biological Data*, 3e ©
2020 W. H. Freeman and Company

3.4 简单线性回归的前提假设

- 不满足前提假设时
 - (1) 离群值/异常值 Outliers (特别容易影响回归模型)
 - 尝试数据转换或其他方法 (robust regression methods & permutation tests)

$(\beta = 0?)$

- 不包含异常值
 - $b = -0.37$
 - $t = -2.66, P=0.024$
- 包含异常值
 - $b = -0.18$
 - $t = -0.81, P=0.43$



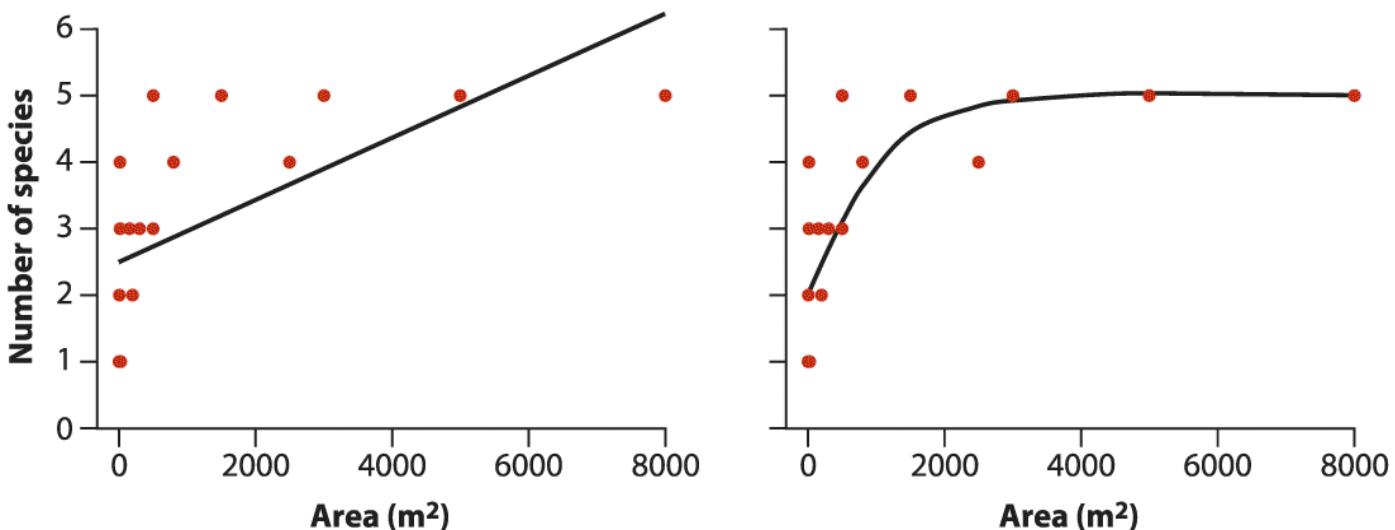
dark-eyed junco



Whitlock & Schlüter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company
Photo: Robert L Kohenbeutel/Shutterstock.com

3.4 简单线性回归的前提假设

- 不满足前提假设时
 - (2) 非线性 nonlinearity
 - 尝试数据转换



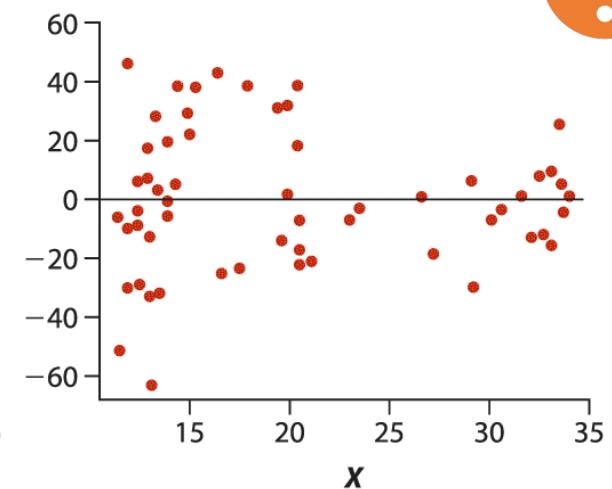
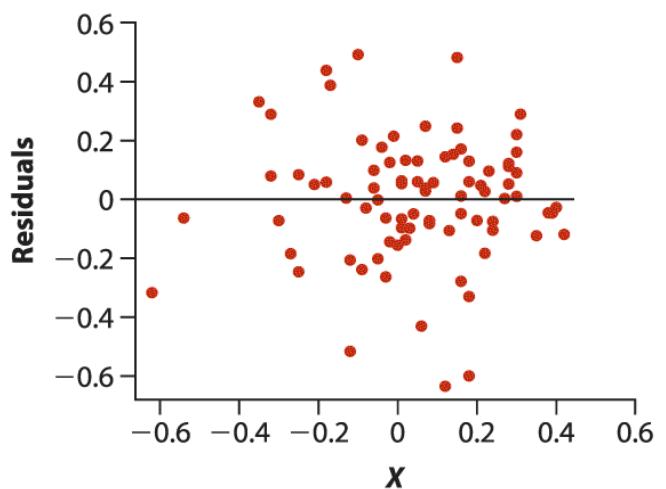
Whitlock & Schlüter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

3.4 简单线性回归的前提假设

- 不满足前提假设时
 - (3) 非正态性和不等方差 non-normality and unequal variance
 - 残差图residual plot 有助于检查残差是否符合正态分布和残差方差是否相等;
 - $\text{residual} \sim X \rightarrow (Y_i - \hat{Y}_i) \sim X_i$

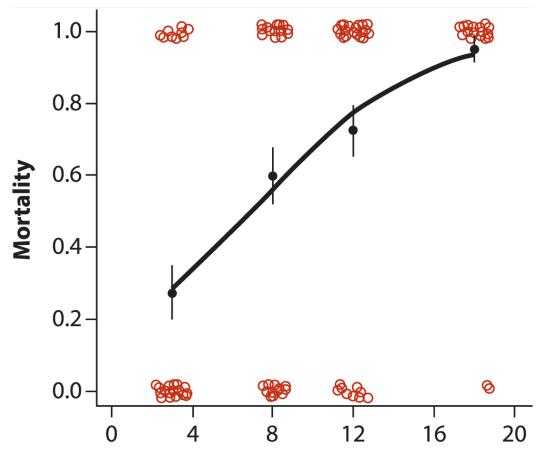
如果满足前提假设，则

- 散点在零水平线上方和下方大致对称分布;
- 在X的所有值上下方点的方差大致相等;
- 靠近该线的点比远离该线的点密集;
- 沿X轴没有明显的曲线;

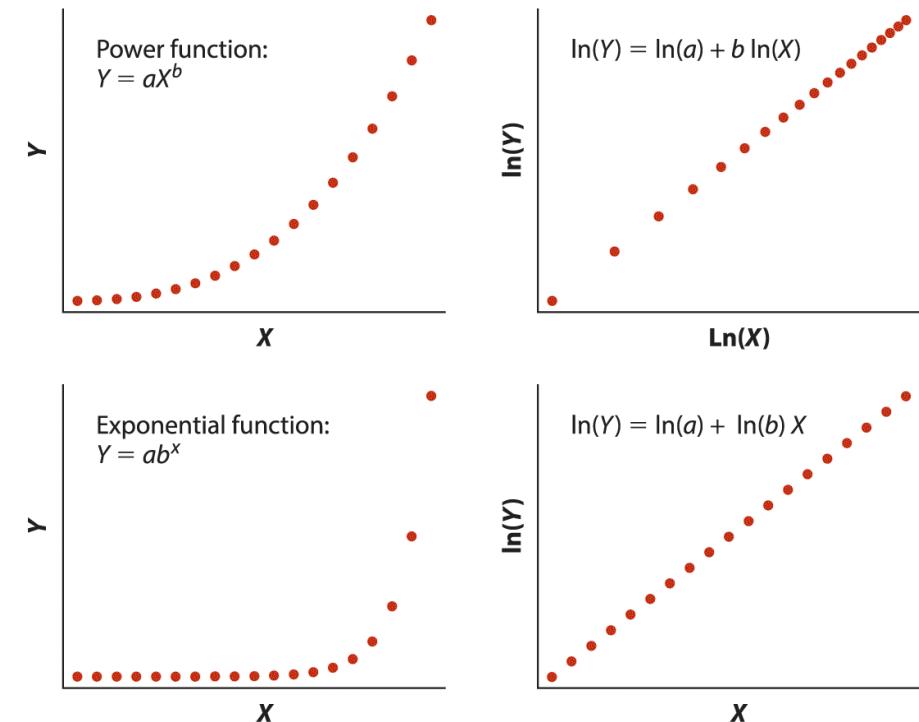


3.5 线性回归的变形——next week (& other methods)

- Transformations
- Regression with nonlinear relationships
 - Quadratic curves
- Logistic regression
 - fitting a binary response variable



Whitlock & Schlüter, *The Analysis of Biological Data*, 3e © 2020 W.H. Freeman and Company



Whitlock & Schlüter, *The Analysis of Biological Data*, 3e © 2020 W.H. Freeman and Company

4. 小结

- 相关系数 (r) 度量了两个数值变量之间关联的强度和方向。
 - 相关性不是因果关系；当在一组个体上测量两个变量时，无论这两个变量之间是否存在因果关系，都可以使用相关性；
 - 相关系数的取值范围从-1（最大负相关）到0（无相关）到1（最大正相关）；
- 使用相关性进行分析的前提包括
 - 这两个数值变量具二元正态分布，并且这些个体是随机抽样的；
 - X 和 Y 之间的关系是线性的；在 X 和 Y 的散点图中，点的分布呈圆形或椭圆形；
- 当不满足上述前提时，Spearman秩相关可以度量两个变量排名之间的线性相关性，其中每个变量都是从低到高排名的；
- 测量误差会使相关系数的估计趋近零；

4. 小结

- 回归是一种从解释变量X预测响应变量Y的方法（均为数值变量）；
- 线性回归通过一系列点拟合一条回归直线：
 - 总体回归线为 $Y = \alpha + \beta X + \varepsilon$, 其中 $\varepsilon \sim N(0, \sigma^2)$ ；
 - 样本估计为 $Y = a + bX$, 其中 b 是直线的斜率, a 是截距；
 - 最小二乘法通过最小化Y的观察值与预测值之间的偏差平方和来确定参数；
- 线性回归的前提包括
 - 每个X值对应的所有可能的Y值遵从正态分布，其均值位于真实回归线上；不同Y值的方差都相同；每个Y的观察值代表了该X处可能的Y值总体的一个随机样本；
- 线性回归的零假设通常设定总体回归线的斜率 β 为零；
 - 检验统计量 $t = b/\text{SE}_b$ 服从自由度为 $n-2$ 的t分布
 - ANOVA表和F检验也可用于检验总体斜率 $\beta=0$ 的零假设；

5. 课堂练习

- Ch17-1 Does face shape predict aggression?
 - 男性的面部宽高比 (width-to-height ratio) 平均高于女性——这反映了在青春期期间雄性激素表达的差异。
 - 已知雄性激素能预测攻击性行为；那么，面部形状是否能预测攻击性呢？
 - 为此， Carré 和 McCormick (2008) 比较了 21 名大学曲棍球 (hockey) 运动员的面部宽高比与每场比赛中因激烈违规行为（如打架或横扫）而获得的平均罚分。数据见 "chap17q01FacesAndPenalties.csv"。
- 请尝试
 - 计算面部宽高比与平均罚分的相关系数？
 - 通过面部宽高比来预测罚分的回归线模型是？