Lecture 6-
Remedies
& Miscellaneous Topics

周在莹

清华大学统计与数据科学系

zyzhou@mail.tsinghua.edu.cn



本节揭晓：

- 你知道在 R 中如何对一个简单线性回归模型进行诊断吗?
- GLT的工作理念你还记得吗？我们将介绍如何用它来辅助模型诊断
- 如果模型诊断有问题，该如何补救?
- 如何从几何上理解相关系数?



学习目标（精简版）

- 掌握
  - ✓ 会读 R 中提供的回归诊断图
  - ✓ Box-Cox 变换的作用及如何使用
  - ✓ 精确解读决定系数
- 理解
  - ✓ 离群点、高杠杆点、高影响力点的区别和联系
  - ✓ GLT 在失拟检验中的应用（问题的提出、完整模型、简化模型、过程）
- 了解
  - ✓ 过原点的回归所存在的弊端
  - ✓ 反向预测的概念、和“正向”的关系



Topic 1:

Remedies



Outline

- 1. More Diagnostics for Residuals
  - 1.1 Review residual plots
  - 1.2 Built-in diagnostic plots in R
  - 1.3 *F*-test for lack of fit
- 2. Discuss Remedies
  - 2.1 Nonlinear relationship
  - 2.2 Non-constant variance
  - 2.3 Non-normal distribution
  - 2.4 The general method: transformations



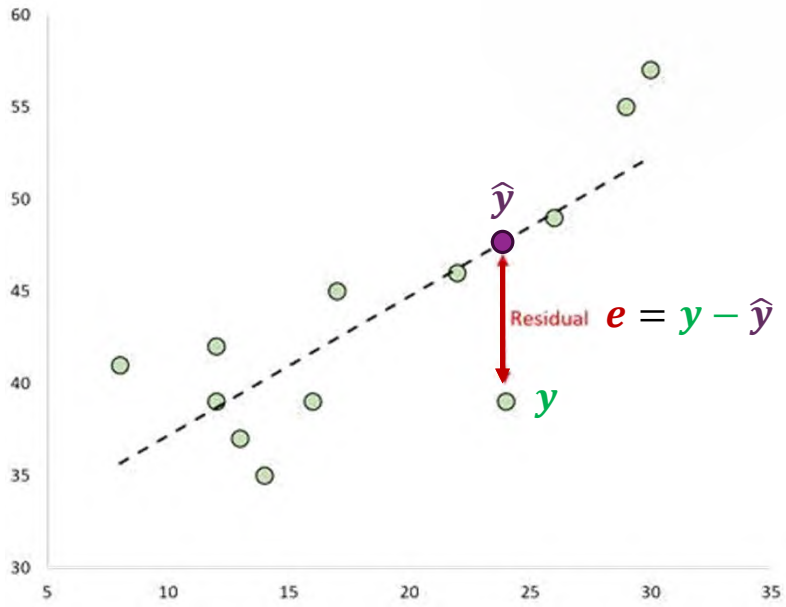
Diagnostics for Residuals

- Residuals are leftover of the outcome variable after fitting a model (predictors) to data and they could reveal unexplained patterns in the data by the fitted model

- Look at residuals to find serious/obvious violations of the model assumptions
  - nonlinear relationship
  - non-constant variance
  - non-Normal errors
  - dependence

  or/and

  - presence of outliers
  - a strongly skewed distribution



Recommendations for Checking Assumptions

- Plot Y vs X (Is it a linear relationship?)
- Use `scatter.smooth` to get a smoothed curve fit
- If reasonable, fit model to get residuals
- Look at distribution of residuals
- Plot residuals vs
  - time (run order)
  - X or any other potential explanatory variable
  - predicted value ($\hat{Y} = b_0 + b_1 X$)
- In all cases look for
  - nonrandom patterns
  - outliers (unusual observations)



填空题 15分

设置

8

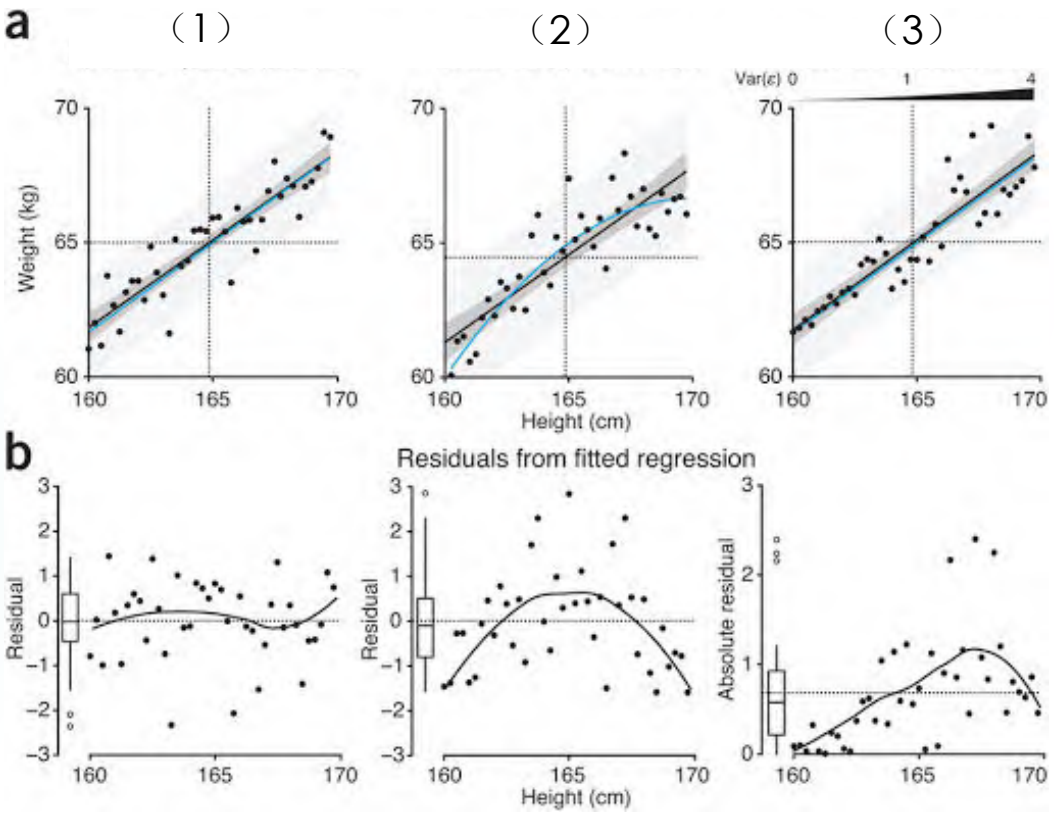
请对下面三组图做出诊断，填入相应的字母。如有违反假设的情况，选出你认为最严重的问题。

（1） [填空1]

（2） [填空2]

（3） [填空3]

- A. Nonlinearity
- B. Non-constant variance
- C. Non-Normality
- D. Dependency
- E. No obvious violation

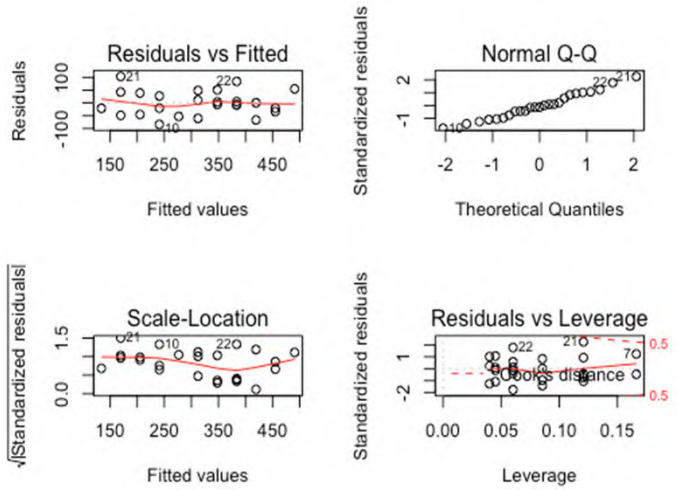


作答

Built-in Diagnostic Plots in R

```
> reg <- lm(hours ~ lotsize, data=toluca)
> plot(reg)
```

- Use a `plot` function to an `lm` object after running an analysis
- Then R will show four diagnostic plots one by one:
  - Residuals vs Fitted
  - Normal Q-Q
  - Scale-Location
  - Residuals vs Leverage

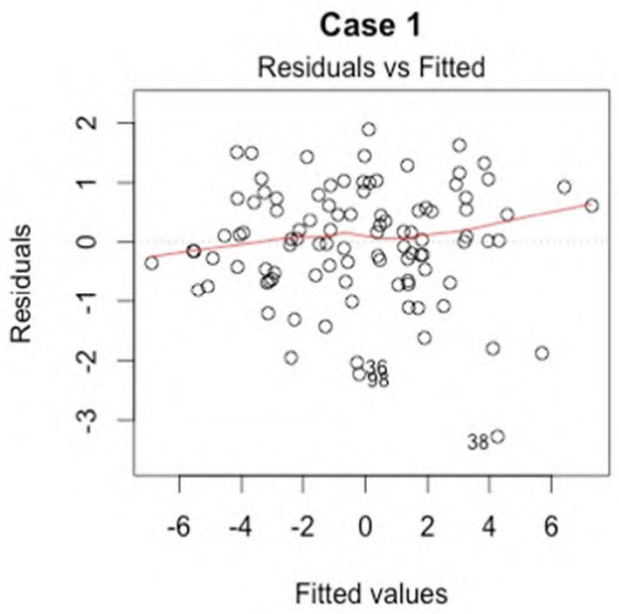


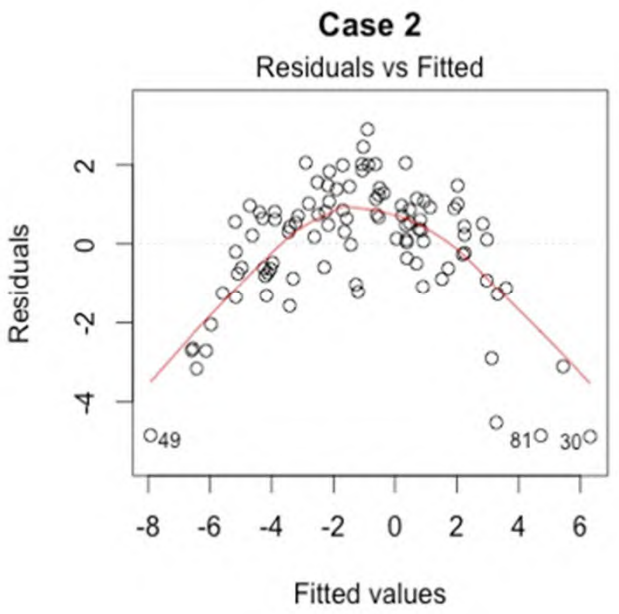
- Points with numbers next to them in each plot are extreme values



1. Residuals *vs* Fitted

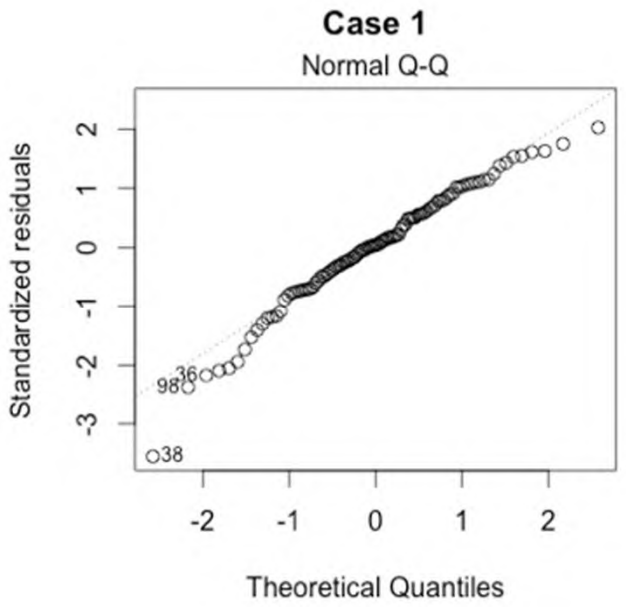
- This plot shows if residuals have non-linear patterns
- It's good if residuals equally spread around a horizontal line without distinct patterns

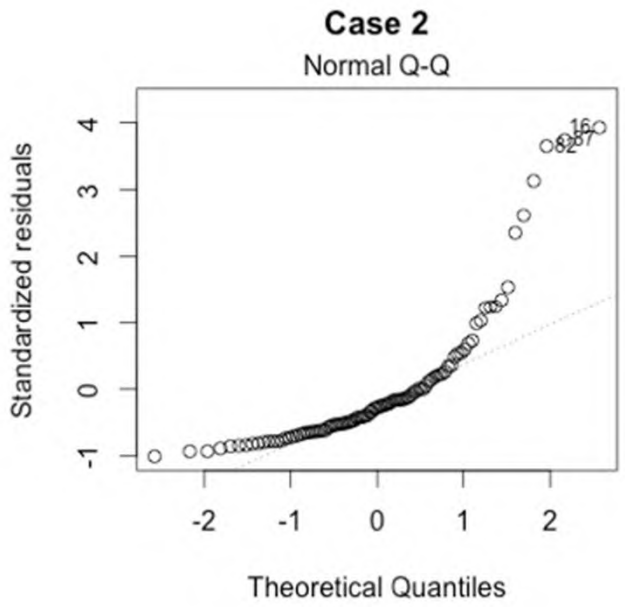




2. Normal Q-Q

- This plot shows if residuals are normally distributed
- It's good if residuals are lined well on the straight dashed line

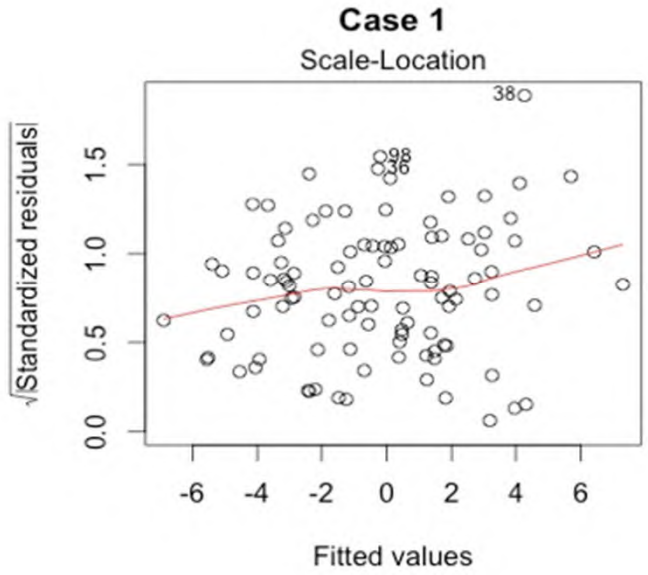


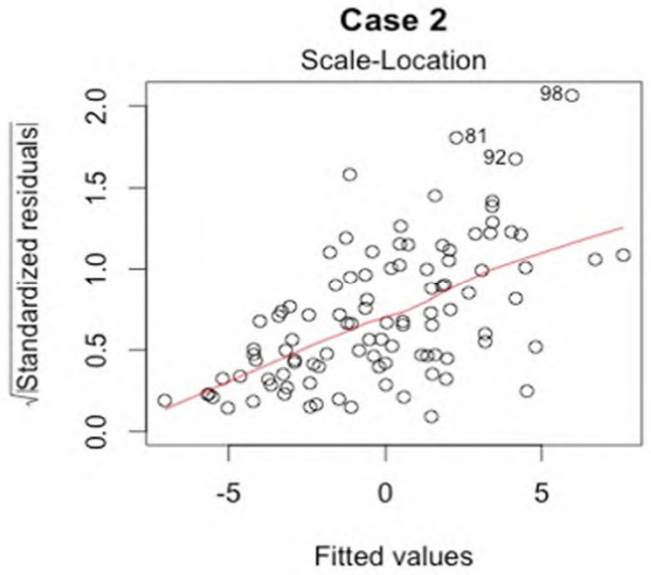


3. Scale-Location (or Spread-Location)

- It shows if residuals are spread equally along the ranges of predictors
- Check the assumption of equal variance (homoscedasticity)
- It's good if you see a horizontal line with equally (randomly) spread points

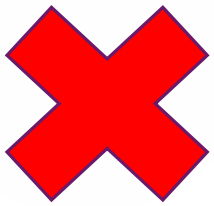
$$Var(e_i) = (1 - h_{ii})\sigma^2$$





4. Residuals *vs* Leverage

- This plot helps us to find influential cases if any
- What is a 'leverage'?
  - Extreme cases against a regression line which can alter the results if we exclude them from analysis
  - Another way to put it is that they don't get along with the trend in the majority of the cases



Leverage is a measure of how far away the independent variable values of an observation are from those of the other observations

- Also known as the **observation self-sensitivity** or **self-influence**

$$h_{ii} = \frac{\partial \hat{Y}_i}{\partial Y_i} \in [0,1]$$



Notes

- Prove that

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

- Proof:

$$h_{ii} = \frac{\partial \hat{Y}_i}{\partial Y_i} = \frac{\partial}{\partial Y_i}(b_0 + X_i b_1) = \frac{\partial}{\partial Y_i}[\bar{Y} + b_1(X_i - \bar{X})]$$

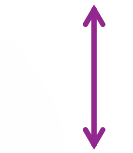
$$= \frac{1}{n} + \frac{\partial}{\partial Y_i}\left[(X_i - \bar{X})\frac{\sum_{j=1}^{n}(X_j - \bar{X})Y_j}{\sum_{j=1}^{n}(X_j - \bar{X})^2}\right]$$

$$= \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$



Outliers *vs* High Leverage Observations

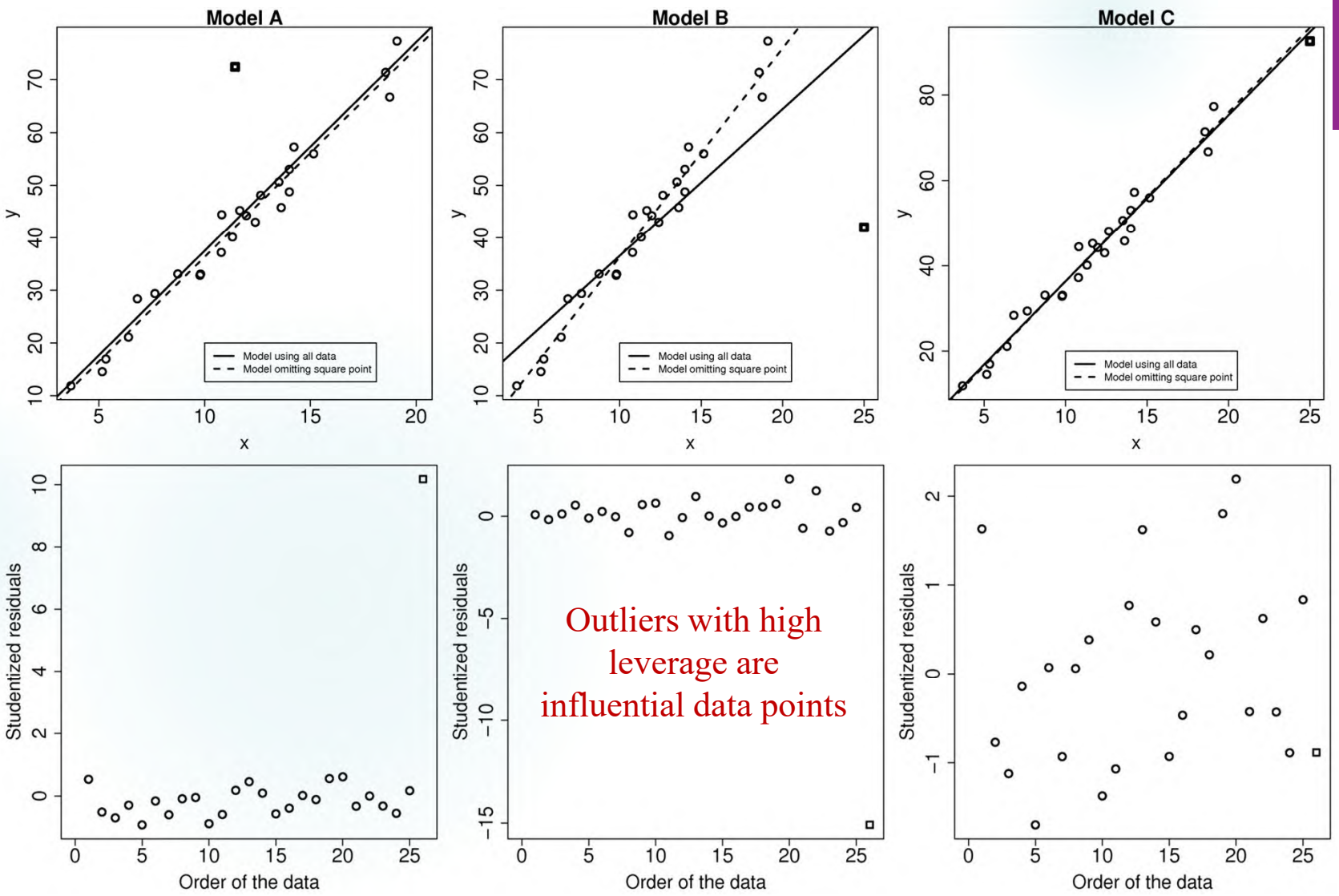
- Two types of **extreme values** (异常值)
  - An **outlier** (离群值) is a data point whose response ***y*** does not follow the general trend of the rest of the data (high **discrepancy**, with large residual)
  - A data point has **high leverage** (杠杆值) if it has "extreme" predictor ***x*** values. A small change in this point has the potential to be influential



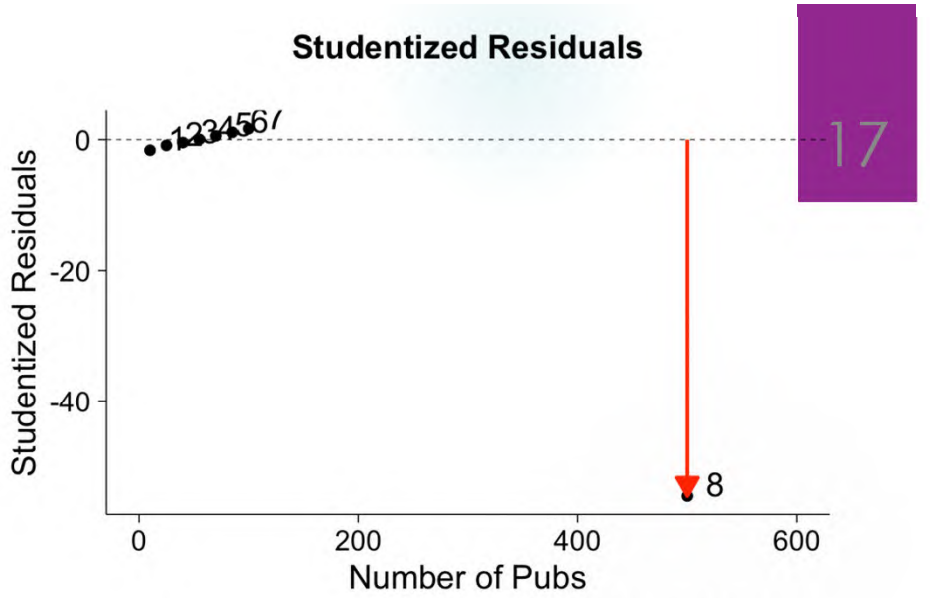


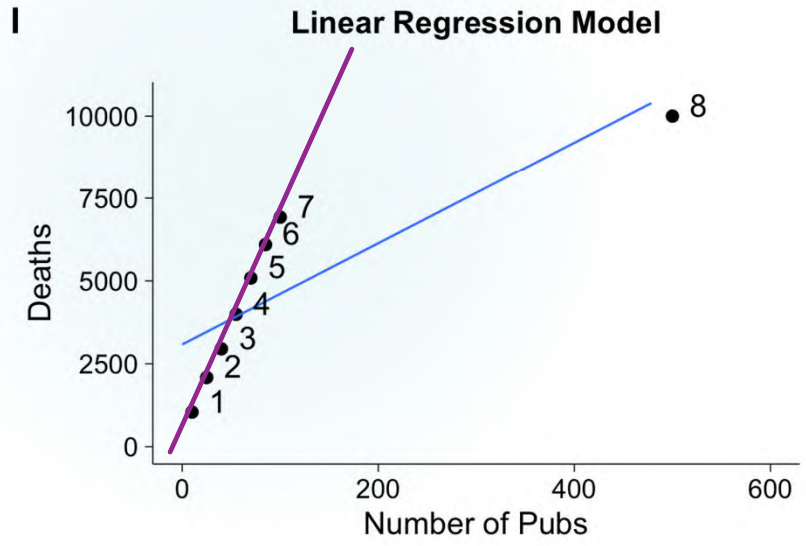
- A data point is **influential** (强影响点) if it unduly influences any part of a regression analysis
  - Removing the observation substantially changes the estimate of coefficients

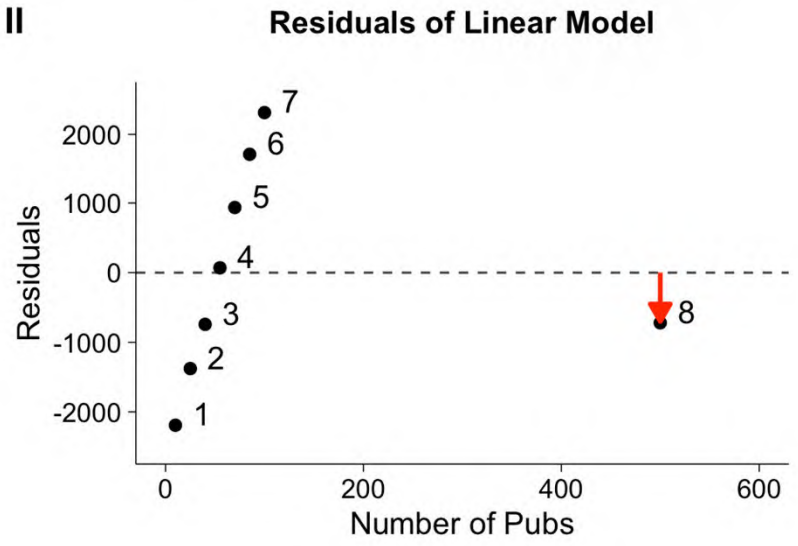



Model A
Model B
Model C
Model using all data
Model omitting square point
x
y
Studentized residuals
Order of the data
Outliers with high
leverage are
influential data points

From Residual to Studentized Residual: Motivation







Standardized Residual *vs* Studentized Residual

- Standardized residual: $\frac{e_i}{s(e_i)}$. Almost all should be between $\pm$ 2

$Var(e_i) = (1 - h_{ii})\sigma^2$

- Studentized residuals offer an alternative criterion for identifying outliers: Standardizing the deleted residuals
  - Deleted residuals: $d_i = Y_i - \hat{Y}_{i(-i)}$
    - $\hat{Y}_{i(-i)}$: the predicted response for the i^{th} observation based on the estimated model with the i^{th} observation deleted
  - Studentized residuals:

$$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{MSE_i(1 - h_{ii})}}$$

  - Semistudentized residual: $t_i = \frac{e_i}{\sqrt{MSE_i}}$ (p. 394 Section 10.2)

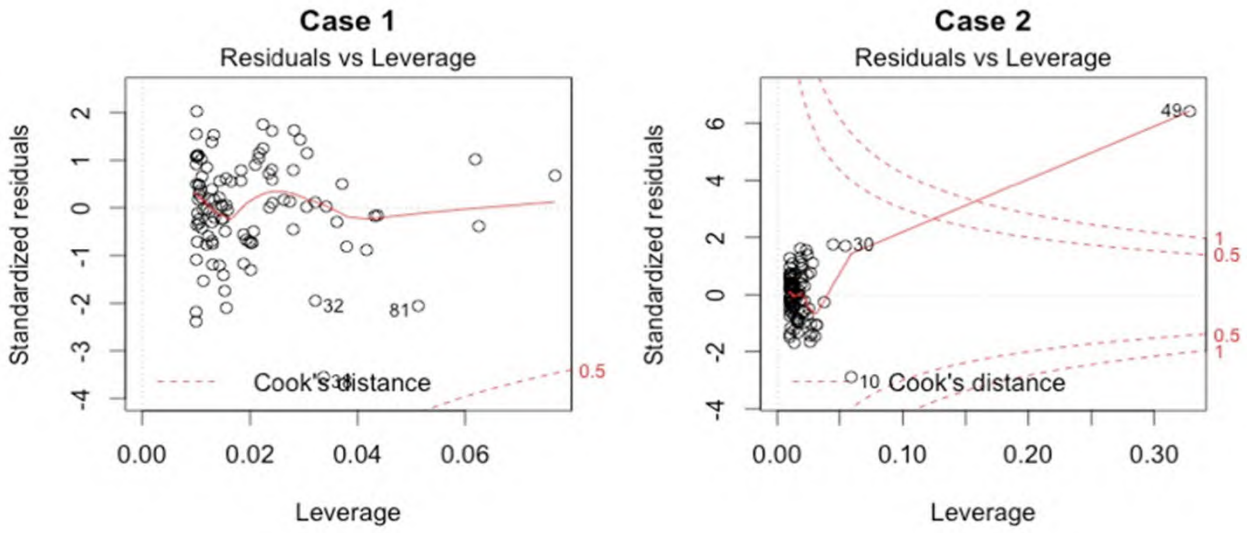


Back to 4. Residuals *vs* Leverage (Cont'd)

- Again, this plot helps us to find influential cases if any
- This time patterns are not relevant. We watch out for outlying values at the upper right corner or at the lower right corner
- Look for cases outside of a dashed line, Cook's distance
  - For $X_{n\times p}$, $s^2 = \frac{e^T e}{n-p}$,

$$D_i = \frac{\sum_j \left(\hat{Y}_j - \hat{Y}_{j(-i)}\right)^2}{ps^2}$$

  - $\hat{Y}_{j(-i)}$: fitted response value obtained when excluding the i^{th} case



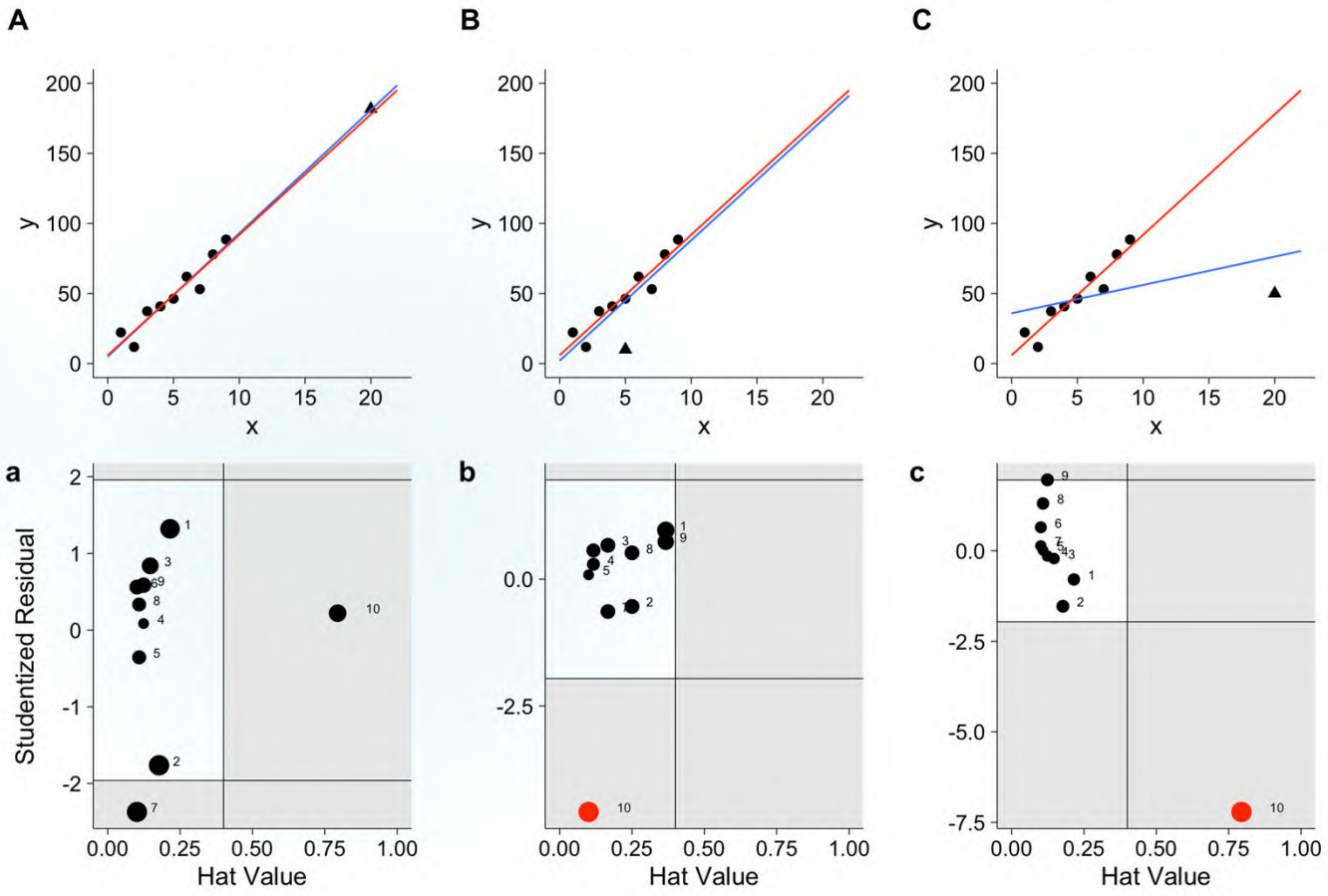
填空题 3分

设置

20

以下三个图中的三角形点，分别是（选最合适的一个描述）

[填空1] [填空2] [填空3]



（请填写1，2，或3，表示：

1. 离群值点；
2. 高杠杆值点；
3. 强影响力点）

作答

Recap of New (Difficult) Concepts

- **Standardized residual & Studentized residual**
- **Leverage**, $h_{ii} = \frac{\partial \hat{Y}_i}{\partial Y_i} \in [0,1]$ $\quad \hat{Y} = HY$
  - A point with zero leverage has no effect on the regression model
  - If a point has leverage equal to 1 the line must follow the point perfectly
  - Measures the "Distance" of X from center $\quad h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$
  - Helps determine outlying X values in multivariable setting… may be influential
- **Cook's D**, $D_i = \frac{\sum_j (\hat{Y}_j - \hat{Y}_{j(-i)})^2}{ps^2}$: Influence of i^{th} case on all predicted values

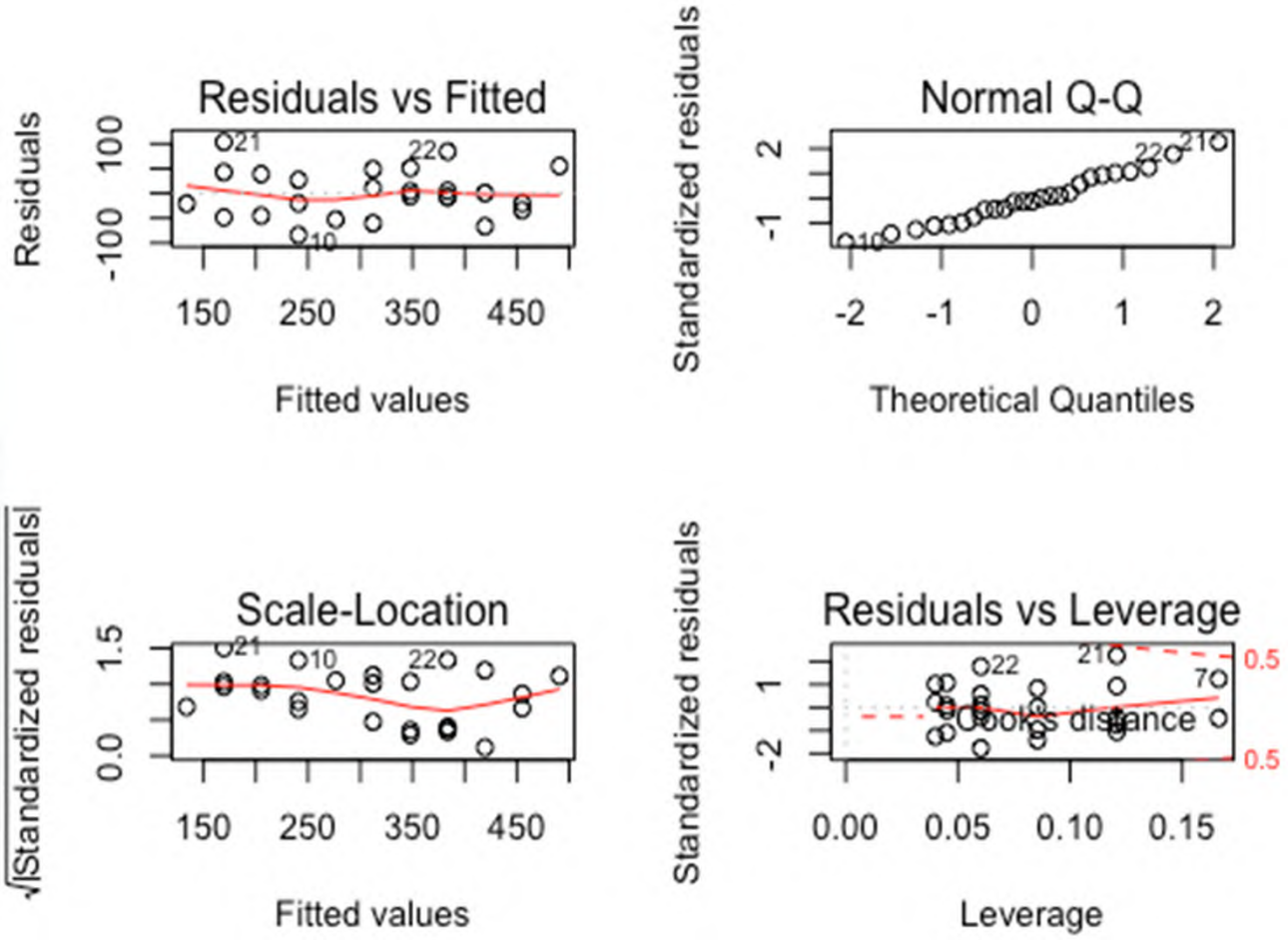
$$= \frac{e_i^2}{ps^2}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$



Toluca Example `> reg <- lm(hours ~ lotsize, data=toluca)`

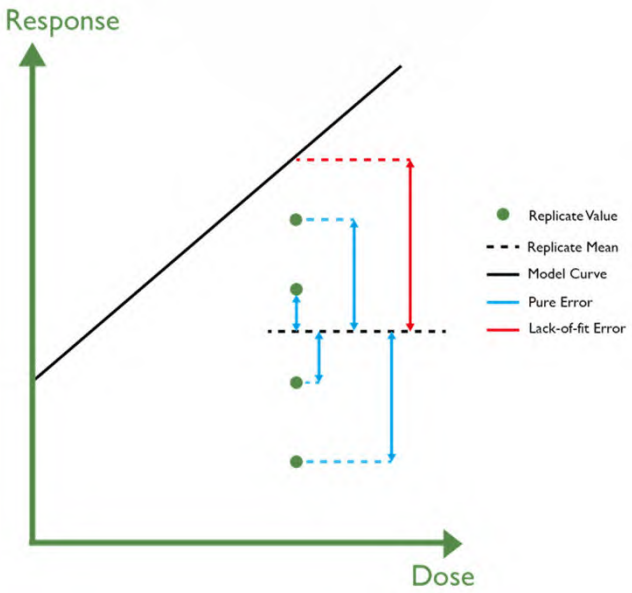
- `par(mfrow=c(2,2))`
  `plot(reg)`
  `par(mfrow=c(1,1))`

- Will discuss these diagnostics more with multiple regression
- Plots provide rule of thumb limits
- Questionable observation Ob21(30,273)



Lack of fit for Nonlinearity

- Goodness of fit, to test if the model is adequate
- Prerequisites: replicates!
  - When we have repeat observations at various values of X, the error term will be partitioned into pure error (error within replicates) and a lack of fit error



- H_0: There is no lack of fit in the regression model

  vs H_1: There is a lack of fit in the regression model
- Use GLT, basic idea is to compare two models
- Description in KNNL Section 3.7



填空题 2分

设置

24

回忆前面的 general linear test，$H_0: \beta_1 = 0$，我们比较

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (full model)

$Y_i = \beta_0 + \varepsilon_i$ (reduced model)

这里，失拟检验的

$$H_0: \mu_i = \beta_0 + \beta_1 X_i \quad vs \quad H_1: \mu_i \neq \beta_0 + \beta_1 X_i$$

要比较的 full model 是 [填空1]， reduced model 是 [填空2]。

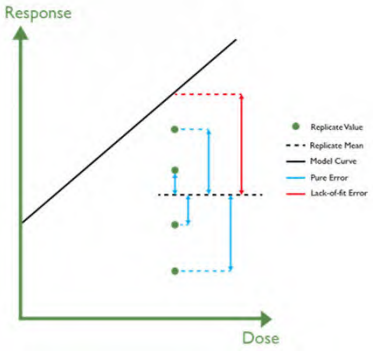
候选答案，填数字即可：

1. $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
2. $Y_i = \beta_0 + \varepsilon_i$

$i = 1, \dots, n$

3. $Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_{ij}$
4. $Y_{ij} = \mu_i + \varepsilon_{ij}$

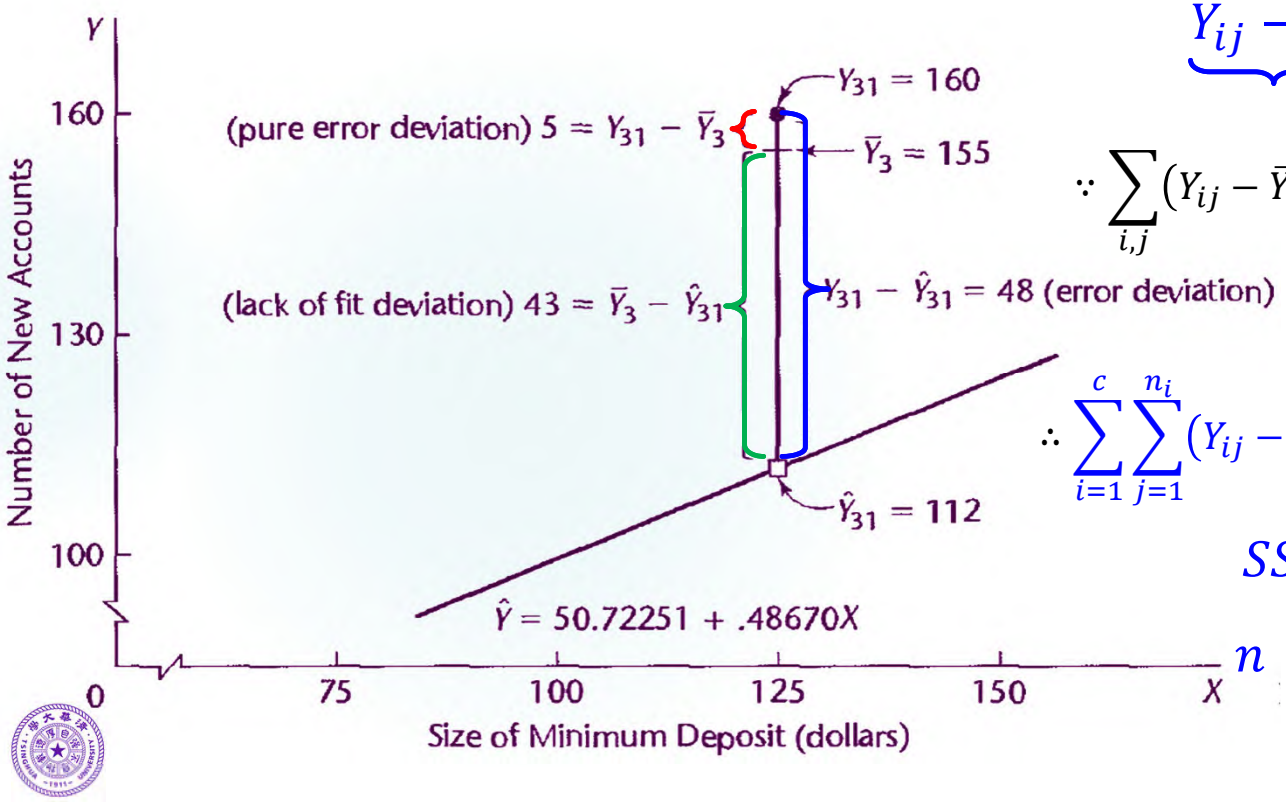
$i = 1, \dots, c; j = 1, \dots, n_i; \sum n_i = n$



作答

Decomposition of Error Deviation

$$Y_{ij} = \mu_i + \varepsilon_{ij} \overset{H_0}{\Rightarrow} Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_{ij}\ , \qquad i = 1, \dots, c; j = 1, \dots, n_i; \sum n_i = n$$



$$\underbrace{Y_{ij} - \hat{Y}_{ij}}_{} = \underbrace{(Y_{ij} - \bar{Y}_i)}_{} + \underbrace{(\bar{Y}_i - \hat{Y}_{ij})}_{}$$

$$\because \sum_{i,j} (Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \hat{Y}_{ij}) = \sum_i (\bar{Y}_i - \hat{Y}_i) \sum_j (Y_{ij} - \bar{Y}_i) = 0$$

$$\therefore \sum_{i=1}^{c} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2 = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\bar{Y}_i - \hat{Y}_{ij})^2$$

$$SSE = SSPE + SSLF$$

$$n - 2 = n - c + c - 2$$

Why replicates?

ANOVA Table

- For testing lack of fit of simple linear regression function

Source of Variation	SS	df	MS	
Regression	$SSR = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	
Error	$SSE = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	Reduced model $Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_{ij}$
Lack of fit	$SSLF = \sum\sum(\bar{Y}_i - \hat{Y}_{ij})^2$	$c-2$	$MSLF = \frac{SSLF}{c-2}$	
Pure error	$SSPE = \sum\sum(Y_{ij} - \bar{Y}_i)^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$	Full model $Y_{ij} = \mu_i + \varepsilon_{ij}$
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y})^2$	$n-1$		

$$E(MSLF) = \sigma^2 + \frac{\sum_i n_i(\mu_i - (\beta_0 + \beta_1 X_i))^2}{c-2}$$

$$E(MSPE) = \sigma^2$$

R Code and Output

$$H_0: \mu_i = \beta_0 + \beta_1 X_i \quad vs \quad H_1: \mu_i \neq \beta_0 + \beta_1 X_i$$

```
> Full <- lm(hours ~ 0 + as.factor(lotsize), data = toluca)
> anova (reg, Full)
```

Full model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Reduced model

$$Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_{ij}$$

Analysis of Variance Table

Model 1: hours ~ lotsize
Model 2: hours ~ 0 + as.factor(lotsize)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	23	54825				
2	14	37581	9	17245	0.7138	0.6893

$n = 25, \quad c = 11$

$$F = \frac{17245/9}{37581/14}$$

80	399	20	113
30	121	110	435
50	221	100	420
90	376	30	212
70	361	50	268
60	224	90	377
120	546	110	421
80	352	30	273
100	353	90	468
50	157	40	244
40	160	80	342
70	252	70	323
90	389		

投票 最多可选1项

设置

29 非线性

以下属于是回归模型吗?

$$Y = \beta_0 + \beta_1 \log(X) + \varepsilon \text{ (Log−linear)}$$

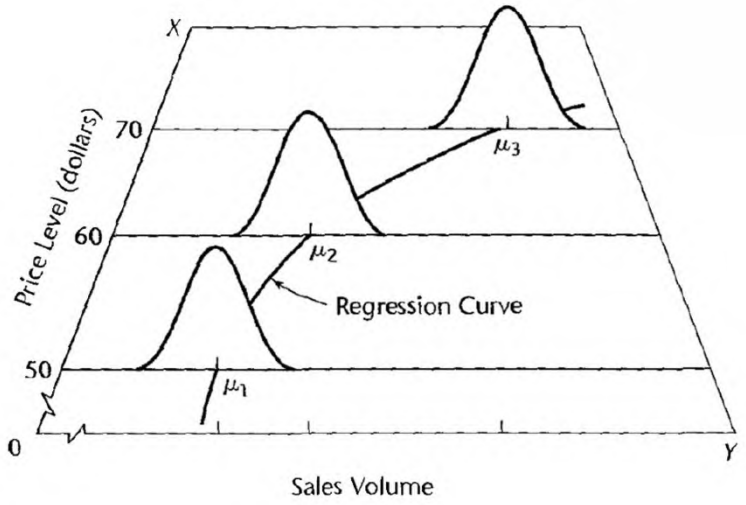
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \text{ (Quadratic)}$$

A 都是

B 是，不是

C 不是，是

D 都不是



提交

Nonlinear Relationships

- We can model many nonlinear relationships with linear models, some have several explanatory variables (i.e., need to move to multiple linear regression)

- Sometimes one can transform a nonlinear equation into a linear equation
  - Consider

$$Y = \beta_0 \exp(\beta_1 X + \varepsilon)$$

  - Can form linear model using log

$$\log(Y) = \log(\beta_0) + \beta_1 X + \varepsilon$$

- Many other times, such transformations may not exist. Then, nonlinear regression analysis is required



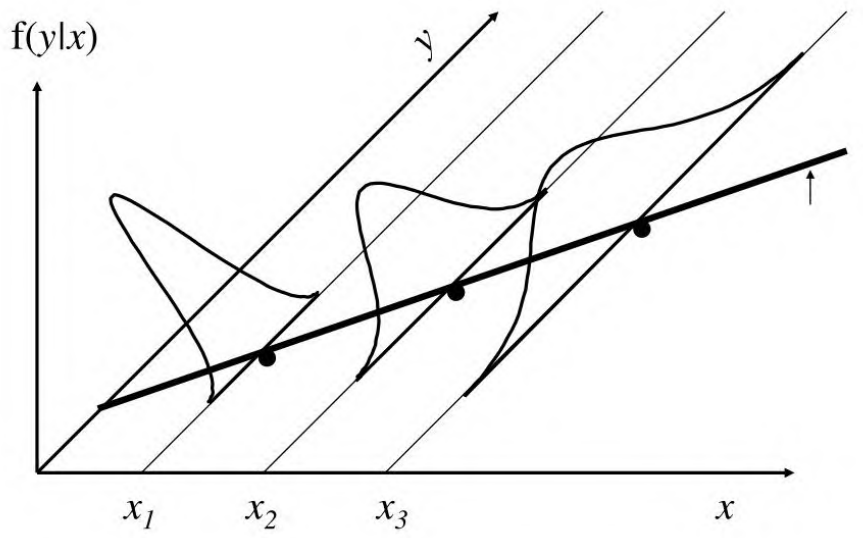
Nonlinear Relationships

- If we don't want to alter assumptions on errors, we can instead perform a nonlinear regression analysis
  - Means vary in nonlinear manner
  - Observations deviate about these means just as in linear regression
- KNNL Chapter 13
  - Introduction to Nonlinear Regression and Neural Networks
- R `nls` function (nonlinear squares)



Nonconstant Variance

- Sometimes we need to model the way in which the error variance changes
  - May be functionally related to *X*
  - In other words changes with the mean
- In this case, we use a weighted analysis
- This is discussed in KNNL 11.1 (WLS)
- Specify `weights` in the **`lm`** function





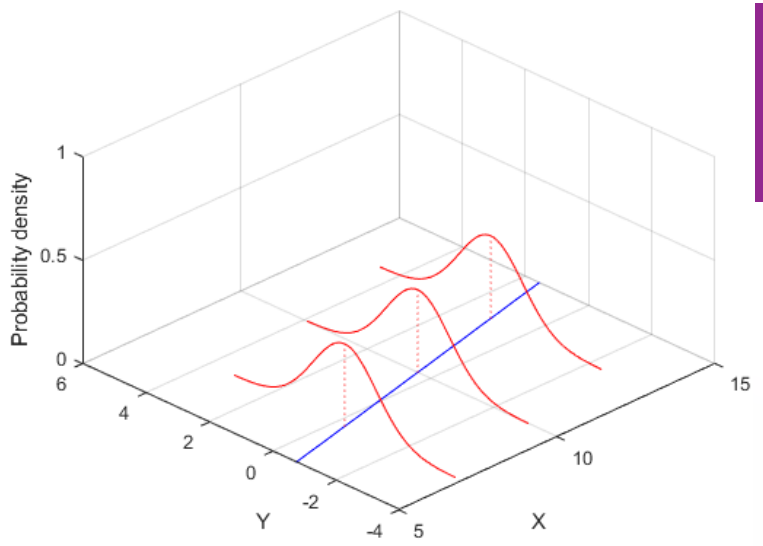
Non-Normal Errors

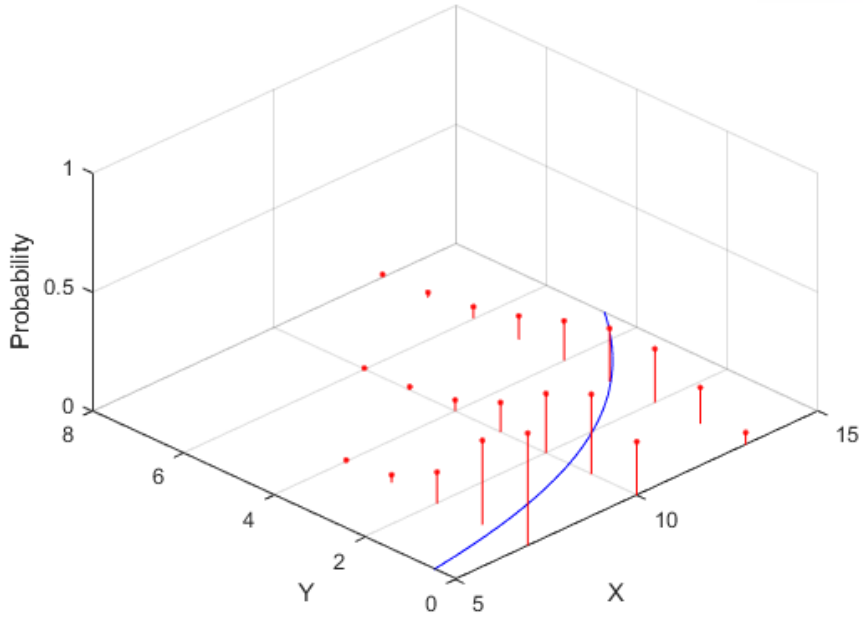
- Transformations of Y often allow you to still use linear regression
- If not, use a procedure that allows different distributions for the error term
  - R `glm`



Generalized Linear Model

- Possible distributions of Y:
  - Binomial (Yes/No or percentage data)
  - Poisson (Count data)
  - Gamma (exponential)
  - Inverse gaussian/Wald
  - Negative binomial
  - Multinomial
- $E(Y) = g(\beta_0 + \beta_1 X_1 + \cdots)$
  - g: a monotonic nonlinear transfor
  - g^{-1}: known as the "link" function





Transformations

- Large values are more variable than smaller values
- Transformations are typically performed to:
  - Stabilize variance - to achieve equal variance
  - Improve normality - often violated because it is easy to have an outlier when variance is large which can be 'reined in' with a transformation
  - Simplify the model
- Sometimes transformations will solve a couple of these problems



Variance Stabilizing Transformations

- $E(Y) = \mu_x$, and $Var(Y) = h(\mu_x)$: variance changes with μ_x
- Consider transformation $f(Y)$
  - $f(Y) \approx f(\mu_x) + f'(\mu_x)(Y - \mu_x)$
  - $Var(f(Y)) = (f'(\mu_x))^2 h(\mu_x)$
- Aim to have $(f'(\mu_x))^2 h(\mu_x) = \text{constant} \Longrightarrow f'(\mu) = \frac{c}{\sqrt{h(\mu)}} \Longrightarrow$

$$f(\mu) = \int \frac{c d\mu}{\sqrt{h(\mu)}}$$

- Examples:
  - When $h(\mu) = \mu^2,\ f(\mu) = \log(\mu)$
  - When $h(\mu) = \mu^{2\nu}, \nu \neq 1,\quad f(\mu) = c\mu^{1-\nu}$



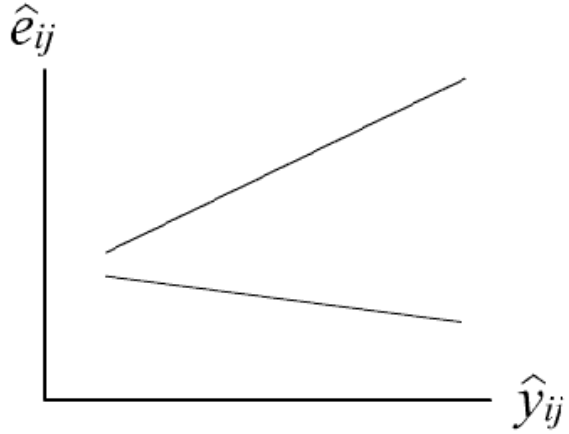
Box-Cox Transformations (1964)

- Also called power transformations
- Designed for strictly positive responses
- Adjust for non-Normality and nonconstant variance(*variance stabilization*)
- $Y^* = Y^\lambda$, or rescale it as $Y^* = \left(Y^\lambda - 1\right)/\lambda$
  - In the second form, the limit as λ approaches zero is the (natural) log
- Choose the transformation (λ) to find the the best fit to the data



Important Special Cases

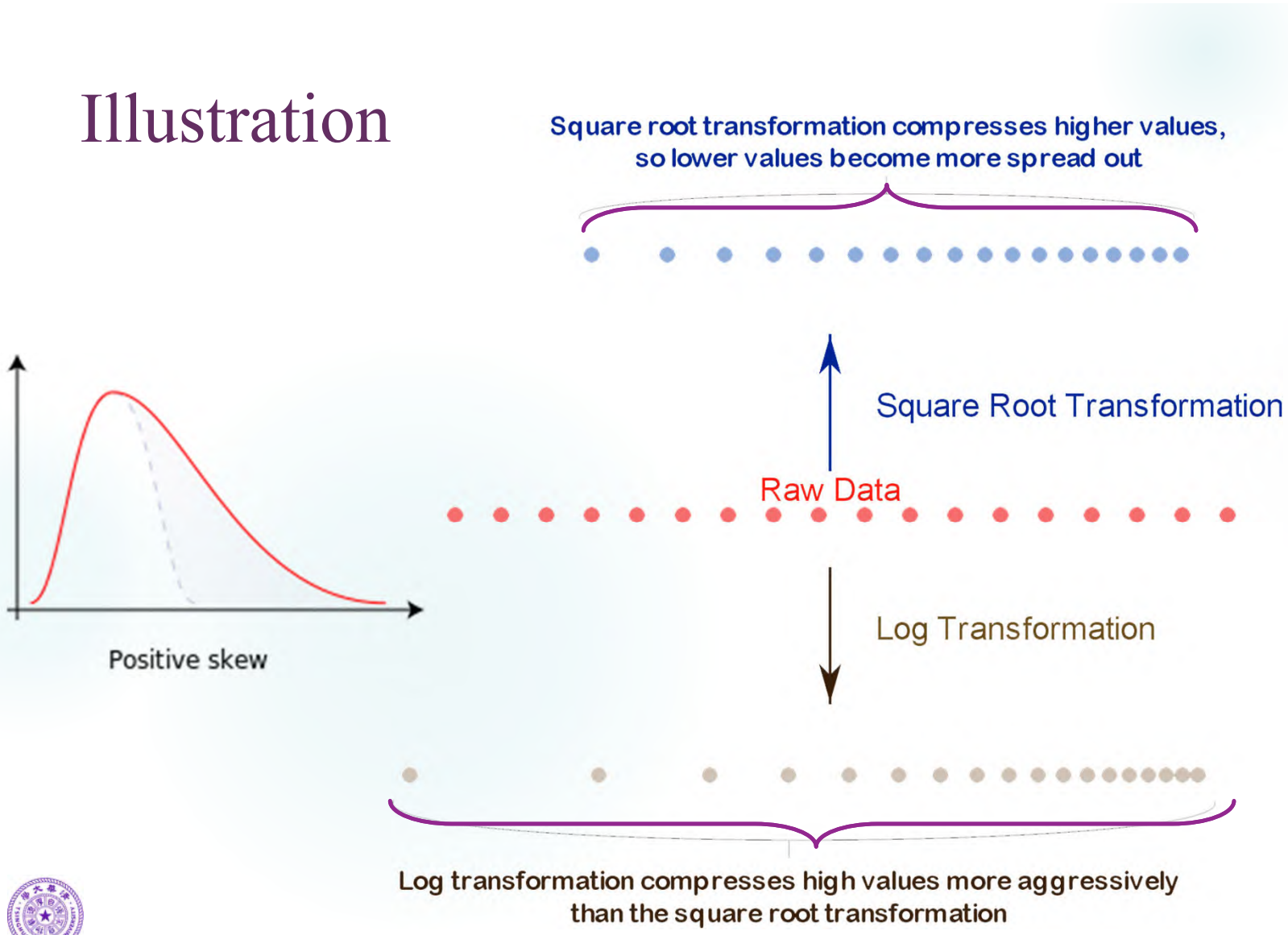
- If the variance is small at the low end and quite large on the right side, producing a fanning effect. Consider the family of transformations
  - $\lambda = 1, Y^* = Y$, no transformation
  - $\lambda = 0.5, Y^* = Y^{1/2}$, square root transformation
  - $\lambda = 0, Y^* =$ (natural) log of Y
  - $\lambda = -0.5, Y^* = Y^{-1/2}$, one over square root
  - $\lambda = -1, Y^* = Y^{-1} = 1/Y$, inverse, reciprocal



- Transformations towards the bottom of the list are stronger in how they shrink large values more than they shrink small values



Illustration



Box-Cox Details

- We can estimate λ by including it as a parameter in a non-linear model

$$\frac{Y_i^{\lambda} - 1}{\lambda} = \beta_0 + \beta_1 X_i + \varepsilon_i$$

and use the method of maximum likelihood to estimate it

- Details are in KNNL p 134-137



单选题 1分

设置

41

当我们决定要使用 Box-Cox 变换，首先要明确的目标是：

λ 和 β, σ^2的估计

λ 是多少，不关心 β, σ^2的估计

β, σ^2 是多少，不关心 λ 的估计

σ^2 是多少，不关心 β, λ 的估计



提交

- $L(\lambda,\beta,\sigma^2)=\frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}exp\left\{-\frac{\sum_i^n\left((Y_i^\lambda-1)/\lambda-\beta_0-\beta_1X_i\right)^2}{2\sigma^2}\right\}J(\lambda;y)$

$$J(\lambda;y)=\prod_{i=1}^{n}\left|\frac{d(Y_i^\lambda-1)/\lambda}{dY_i}\right|=\begin{cases}\prod_{i=1}^{n}Y_i^{\lambda-1}, \text{if } \lambda\neq 0\\ \prod_{i=1}^{n}1/Y_i, \text{if } \lambda=0\end{cases}$$

- Take $K_2=(\prod Y_i)^{1/n}$ (geometric mean), then $J(\lambda;y)=K_2^{n(\lambda-1)}$
- $log\left(L(\lambda,\beta,\sigma^2)\right)=-\frac{n}{2}log(2\pi\sigma^2)-\frac{\sum_i^n\left((Y_i^\lambda-1)/\lambda-\beta_0-\beta_1X_i\right)^2}{2\sigma^2}+n(\lambda-1)\log(K_2)$

$$=-\frac{n}{2}log\left(2\pi\sigma^2/K_2^{2(\lambda-1)}\right)-\frac{\sum_i^n\left((Y_i^\lambda-1)/\lambda-\beta_0-\beta_1X_i\right)^2/K_2^{2(\lambda-1)}}{2\sigma^2/K_2^{2(\lambda-1)}}$$

- Take $\tilde{\sigma}^2=\frac{\sigma^2}{K_2^{2(\lambda-1)}}$, $h(Y_i,\lambda)\triangleq\frac{Y_i^\lambda-1}{\lambda}\frac{1}{K_2^{(\lambda-1)}}$, $SSE(\lambda)\triangleq\sum_i^n\left(h(Y_i,\lambda)-\tilde{\beta}_0-\tilde{\beta}_1X_i\right)^2$

$$log\left(L(\lambda,\beta,\sigma^2)\right)=log\left(L(\lambda,\tilde{\beta},\tilde{\sigma}^2)\right)=-\frac{n}{2}log(2\pi\tilde{\sigma}^2)-\frac{SSE}{2\tilde{\sigma}^2}$$

$$h(Y_i,\lambda)=\begin{cases}K_1\left(Y_i^\lambda-1\right), \text{if } \lambda\neq 0\\ K_2\log(Y_i), \quad \text{if } \lambda=0\end{cases}$$

where $K_2=(\prod Y_i)^{1/n}$ and $K_1=1/\left(\lambda K_2^{\lambda-1}\right)$

- $\max_{\lambda,\tilde{\beta},\tilde{\sigma}^2} log\left(L(\lambda,\tilde{\beta},\tilde{\sigma}^2)\right)=\max_{\lambda,\beta,\sigma^2} log\left(L(\lambda,\beta,\sigma^2)\right)$

$$\Rightarrow \operatorname{argmax}_\lambda log\left(L(\lambda,\beta,\sigma^2)\right)=\operatorname{argmax}_\lambda log\left(L(\lambda,\tilde{\beta},\tilde{\sigma}^2)\right)=\operatorname{argmin}_\lambda SSE(\lambda)$$



Box-Cox Solution

$$SSE(\lambda) \triangleq \sum_i^n \left(h(Y_i,\lambda) - \tilde{\beta}_0 - \tilde{\beta}_1 X_i\right)^2$$

- Standardized transformed Y is
  - $h(Y_i,\lambda) = \begin{cases} K_1\left(Y_i^{\lambda} - 1\right), \text{if } \lambda \neq 0 \\ K_2 \log(Y_i), \quad \text{if } \lambda = 0 \end{cases}$

  where $K_2 = \left(\prod Y_i\right)^{1/n}$ (geometric mean) and $K_1 = 1/\left(\lambda K_2^{\lambda-1}\right)$
- For each λ, regress $h(Y_i,\lambda)$ against X_i and obtain $SSE(\lambda)$
- Best choice of λ minimizes $SSE(\lambda)$, denoted as $\hat{\lambda}(MLE)$
- One degree freedom <u>confidence interval</u>

$$(\lambda_l, \lambda_u) = \left\{\lambda: \log(L(\lambda)) \geq \log(L(\hat{\lambda})) - \frac{1}{2}\chi^2_{1,1-\alpha}\right\}$$

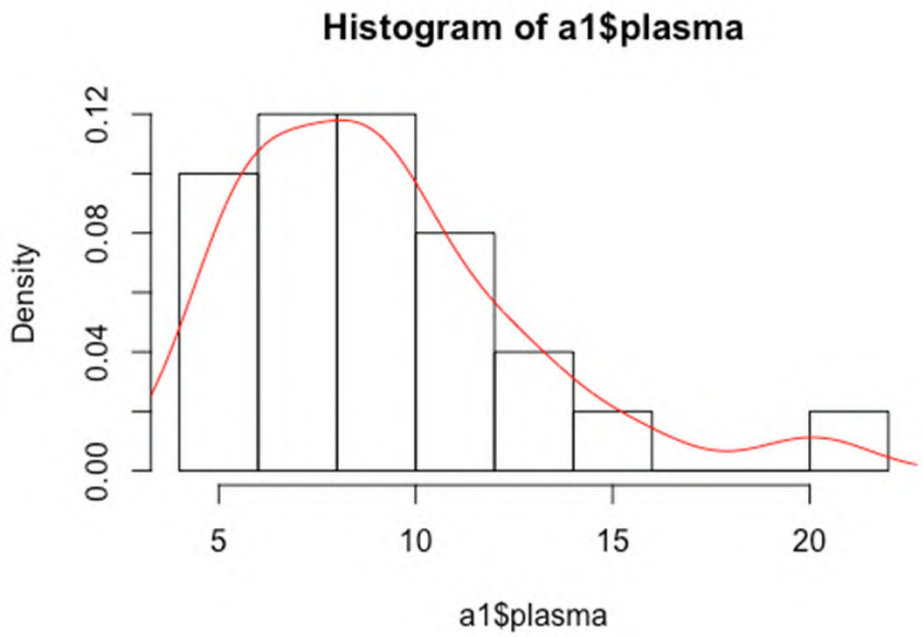
By Wilks' Theorem

$$-2\log\left(\frac{L(\lambda)}{L(\hat{\lambda})}\right) \leq \chi^2_{1,1-\alpha}$$



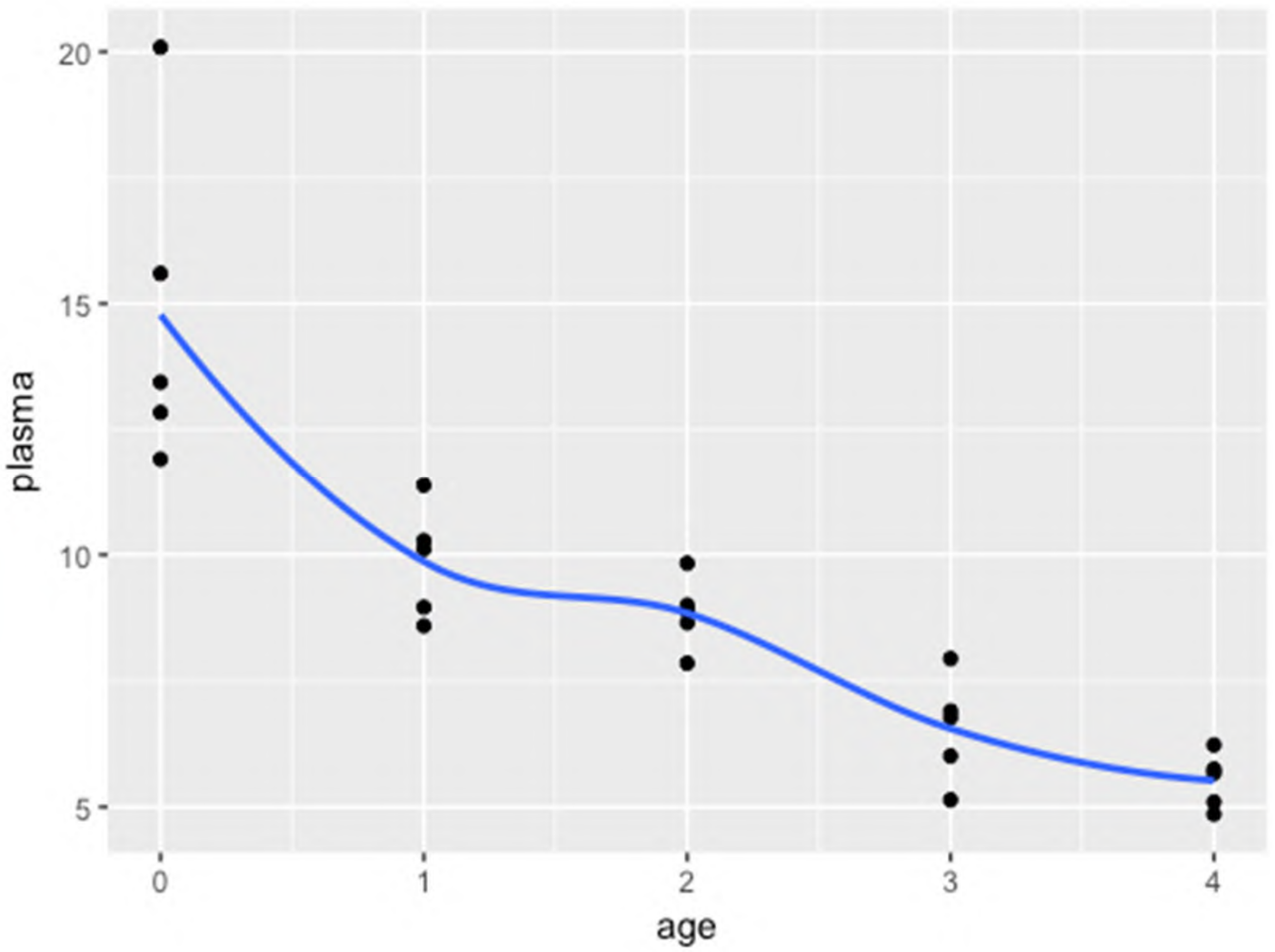
Example

CH03TA08.txt lplasma=$\log_{10}$plasma



	age	plasma	lplasma
1	0	13.44	1.1284
2	0	12.84	1.1086
...			
6	1	10.11	1.0048
7	1	11.38	1.0561
...			
11	2	9.83	0.9926
...			
16	3	7.94	0.8998
17	3	6.01	0.7789
...			
21	4	4.86	0.6866
...			
25	4	6.23	0.7945



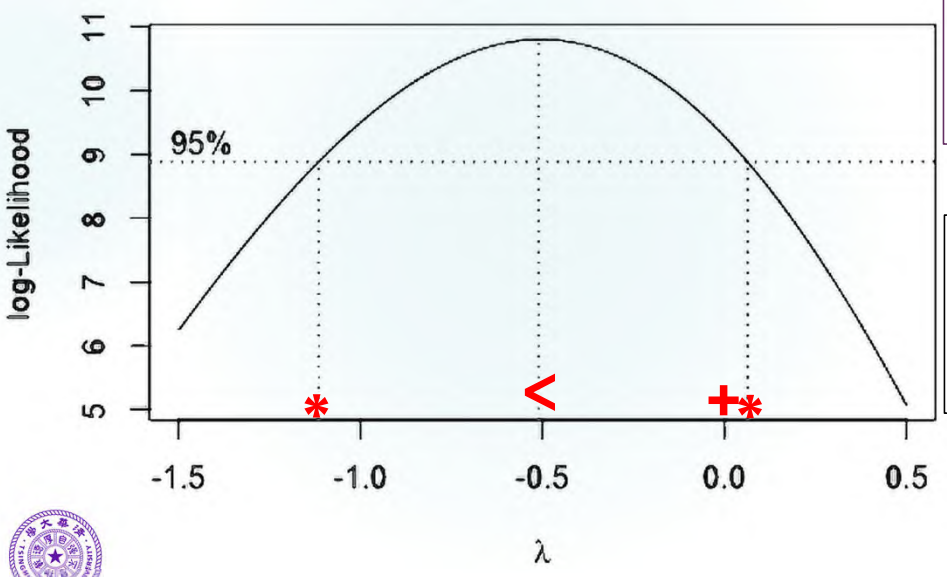


- Variability large initially and gets smaller
- Is it just different at age=0? Or is it more a function of age?

Box Cox Procedure

- Procedure that will automatically find the Box-Cox transformation

```
> library(MASS)
> boxcox(plasma ~ age, data = a1,
         lambda = seq(-1.5, 0.5, length = 10))
```




```
loglike <- boxcox(plasma ~ age, data = a1,
                  lambda = seq(-1.5, .5, length = 10))
loglike$x[which.max(loglike$y)]
```

< - Best Lambda

* - Confidence Interval

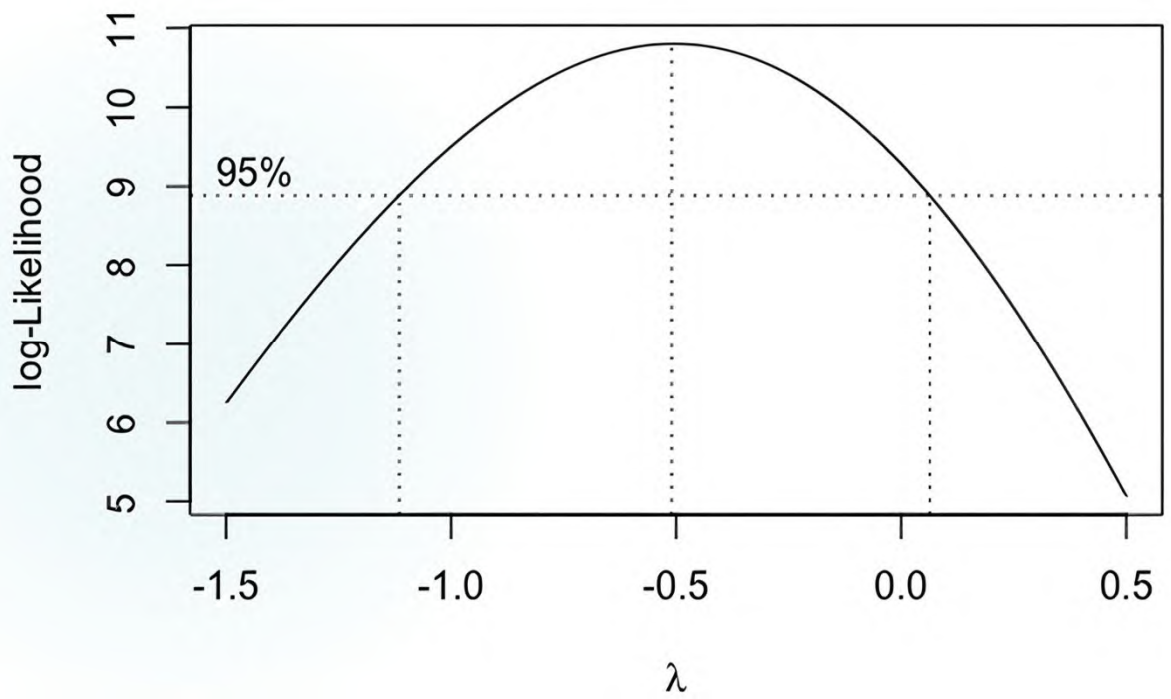
\+ - Convenient Lambda

- Based on the plot, either take inverse square root or log

Functions and datasets to support Venables and Ripley, ``Modern Applied Statistics with S" (4th edition, 2002)

Discussion

- 你猜，为什么我们关心 λ 的置信区间呢?





Doing Procedure by Hand

- First, get the geometric mean

```
> n <-  nrow(a1)
> k2 <- (prod(a1$plasma))^(1/n)
> lambda <- seq(-1, 1, by=0.1)
```

or:

```
> k2  <- exp(mean(log(a1$plasma)))
```

$$h(Y_i, \lambda) = \begin{cases} K_1(Y_i^{\lambda} - 1) & \text{if } \lambda \neq 0 \\ K_2 \log(Y_i) & \text{if } \lambda = 0 \end{cases}$$

$$K_2 = (\prod Y_i)^{1/n}$$

$$K_1 = 1/(\lambda K_2^{\lambda - 1})$$



- Get the transformed data set a2

$$h(Y_i, \lambda) = \begin{cases} K_1(Y_i^{\lambda} - 1) & \text{if } \lambda \neq 0 \\ K_2 \log(Y_i) & \text{if } \lambda = 0 \end{cases}$$

$$K_2 = (\prod Y_i)^{1/n}$$

$$K_1 = 1/(\lambda K_2^{\lambda - 1})$$

```
transformed <-  NULL

for (i in 1:length(lambda)){
  k1 <-  1/(lambda[i]*k2^(lambda[i]-1))

trans_y <- if (lambda[i]==0)
     {k2*(log(a1$plasma))} else
     {k1*((a1$plasma)^lambda[i]-1)}

  a2 <- cbind(a1, lambda = rep(lambda[i], n), trans_y)
  transformed <- rbind(transformed, a2)
}
```



```
# extract SSE's of linear regressions by lambda

sse <-  by(transformed, transformed[,"lambda"],
                  function(x) anova(lm(trans_y ~ age,
                                       data = x))[,2][2] )
plot(sse~lambda, type='l')
```

Recall that

```
> anova(lm(plasma ~ age, data = a1))
```

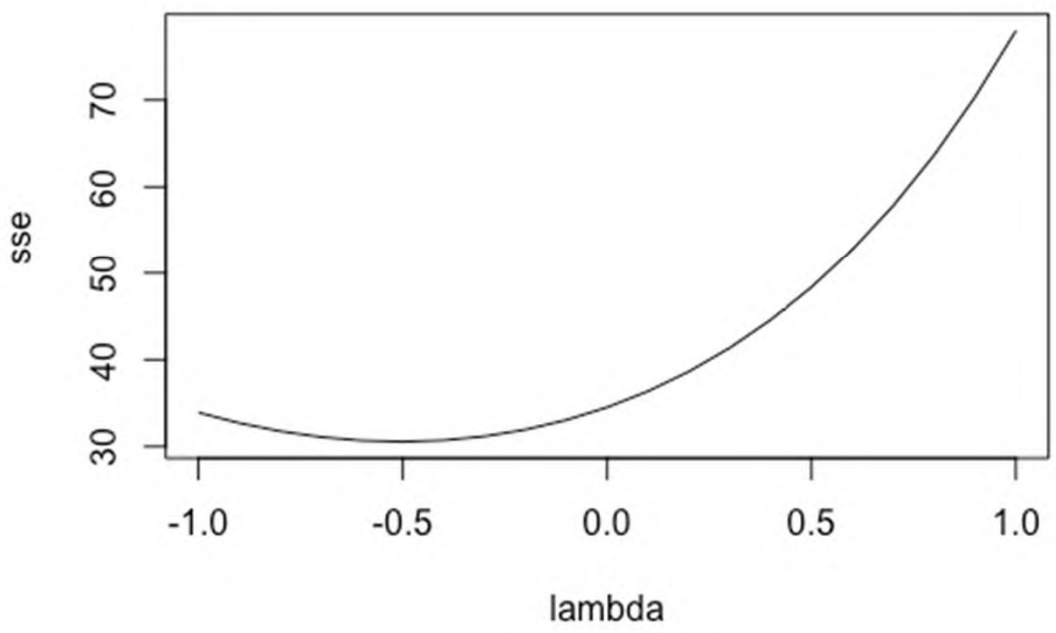
Analysis of Variance Table

Response: plasma

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	238.056	238.056	70.211	1.92e-08 ***
Residuals	23	77.983	3.391		



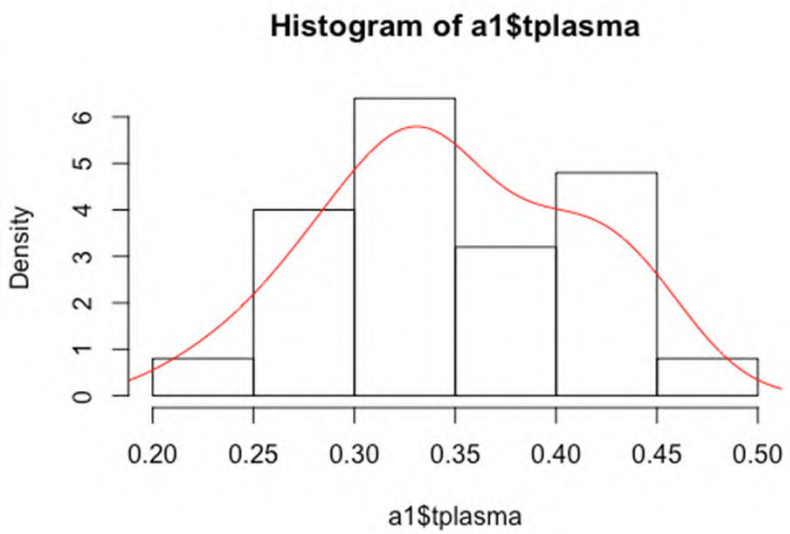
Obs	lambda	sse
1	-1.0	33.9089
2	-0.9	32.7044
3	-0.8	31.7645
4	-0.7	31.0907
5	-0.6	30.6868
6	-0.5	30.5596***
7	-0.4	30.7186
8	-0.3	31.1763
9	-0.2	31.9487
10	-0.1	33.0552

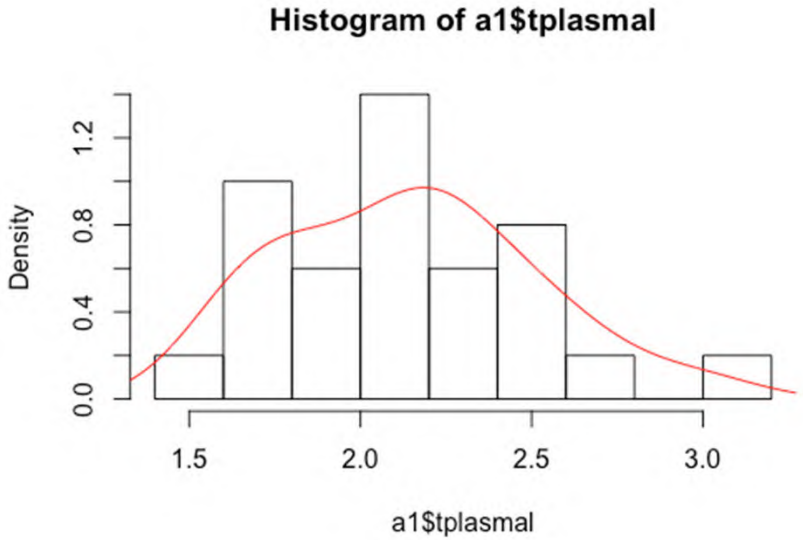


- Apply the transforms

```
a1$tplasma <- a1$plasma^(-0.5)
a1$tplasmal <- log(a1$plasma)

hist(a1$tplasma, prob=T)
lines(density(a1$tplasma), col='red')
```

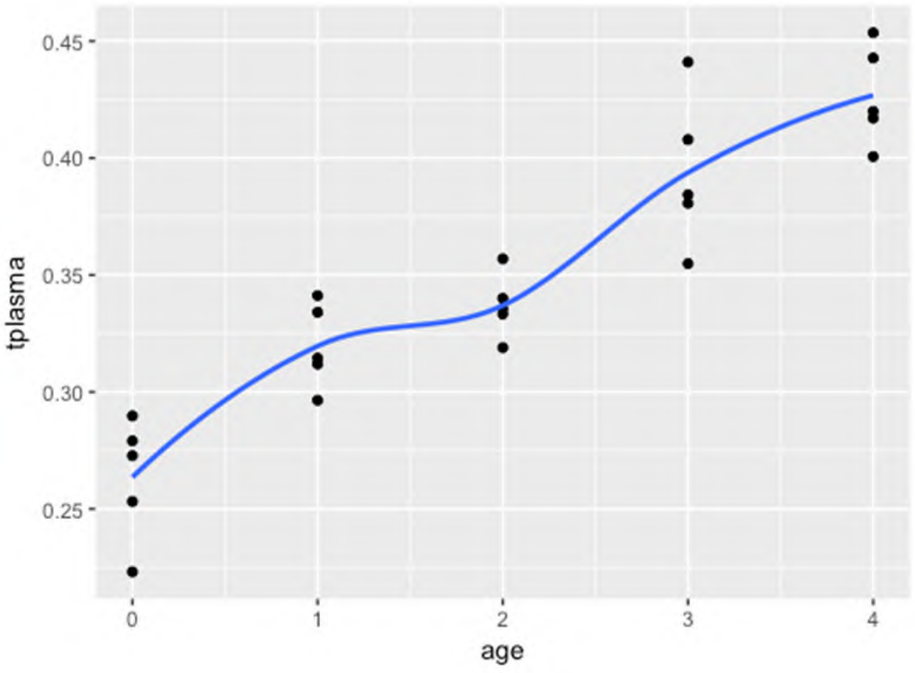


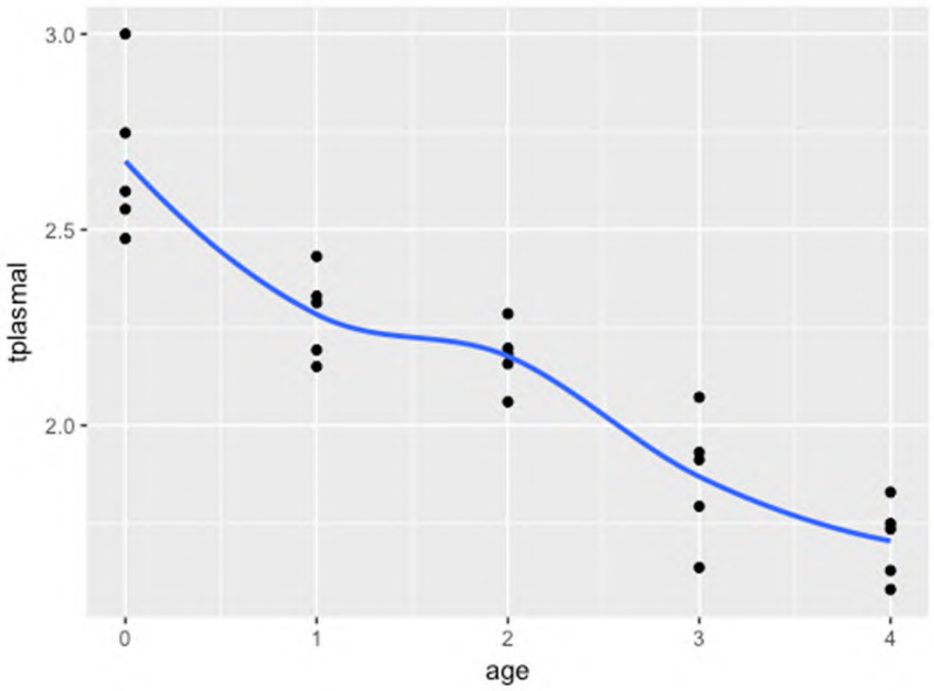




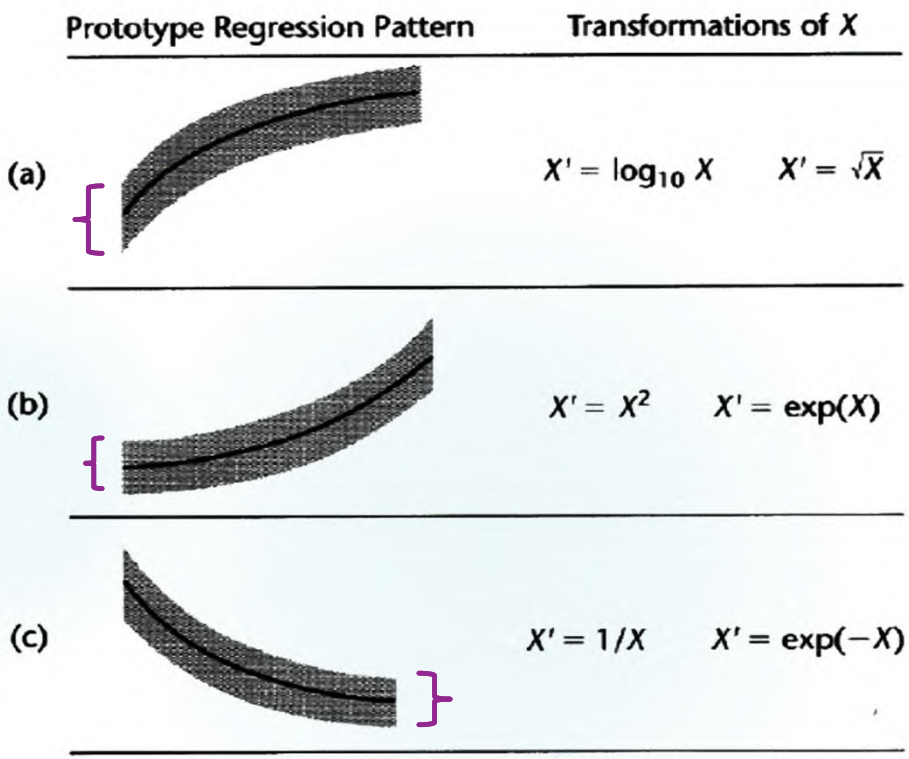
- Transformed data

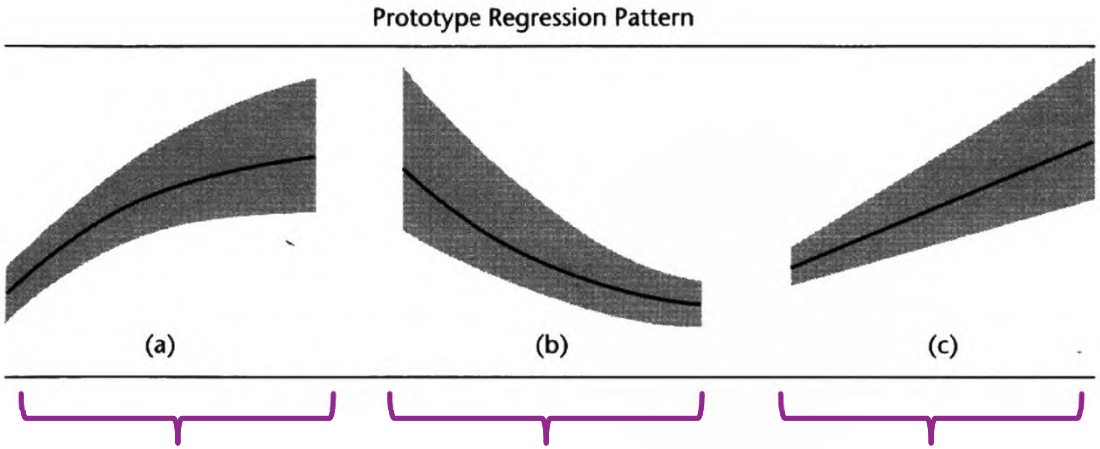
```
library(ggplot2)
ggplot(a1, aes(age, tplasma)) + geom_point() +
      geom_smooth(method="loess", se=FALSE)
```





Prototype Regression Patterns





- Note change in error distribution as function of input
- Can be combined with transformation on X

Transformations on Y

$Y' = \sqrt{Y}$

$Y' = \log_{10} Y$

$Y' = 1/Y$

- Note constancy error terms, for nonlinear relation only



Background Reading

- Sections 3.4 - 3.7 describe significance tests for assumptions

  (read it if you are interested)

- Read Chapter 4, especially Sections 4.1, 4.2, 4.4, 4.5, and 4.6



Topic 2:

Additional Topics



Outline

- 1. Regression Through the Origin
- 2. Inverse Predictions



Regression through the Origin

- $Y_i = \beta_1 X_i + \varepsilon_i$, that is, assume $\beta_0 = 0$
- `lm(Y ~ 0 + X )` in R
- Generally *not* a good idea because of model misspecification and chance variation
- Might be forcing model to behave certain way in area with no data
- Problems with residuals, R^2 and other statistics
- See cautions, KNNL p. 164

$$b = \frac{\sum X_i Y_i}{\sum X_i^2}$$

$$Cov(b, \bar{Y}) \neq 0$$
$$\sum e_i \neq 0$$

$$(Y_i - \bar{Y})^2 = (Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

$$SST \neq SSE + SSR,$$
$$n - 1 \neq n - 1 + 1$$

Inverse Predictions

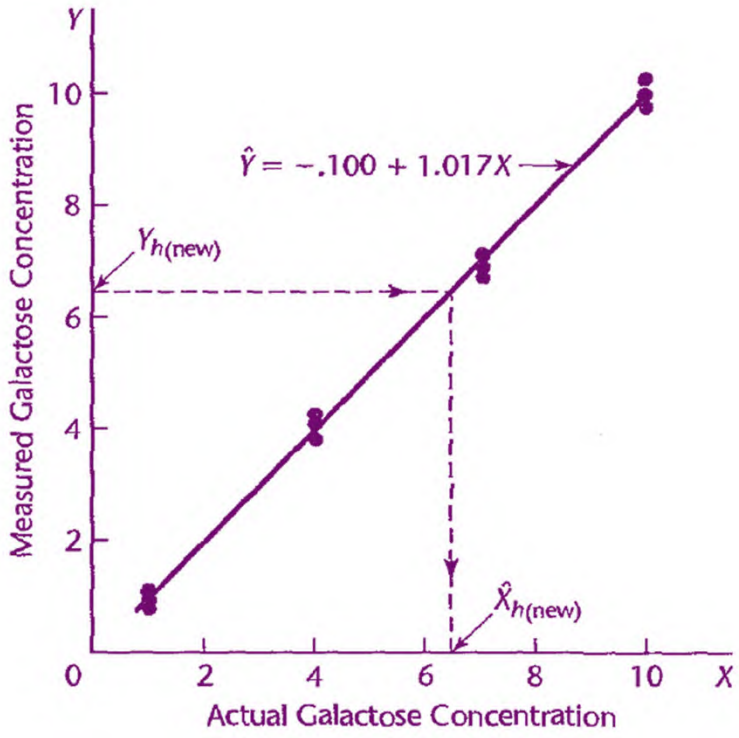
- The same model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- The fitted regression function:

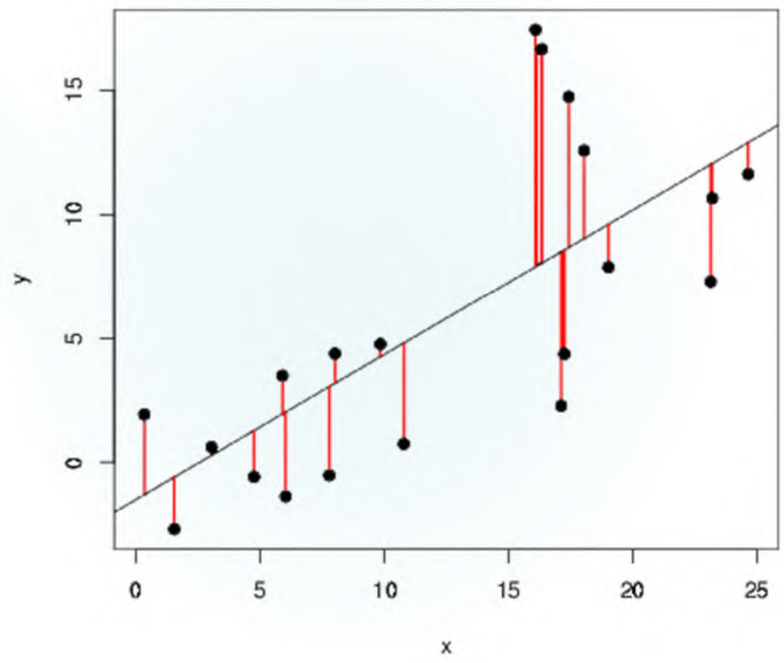
$$Y = b_0 + b_1 X$$

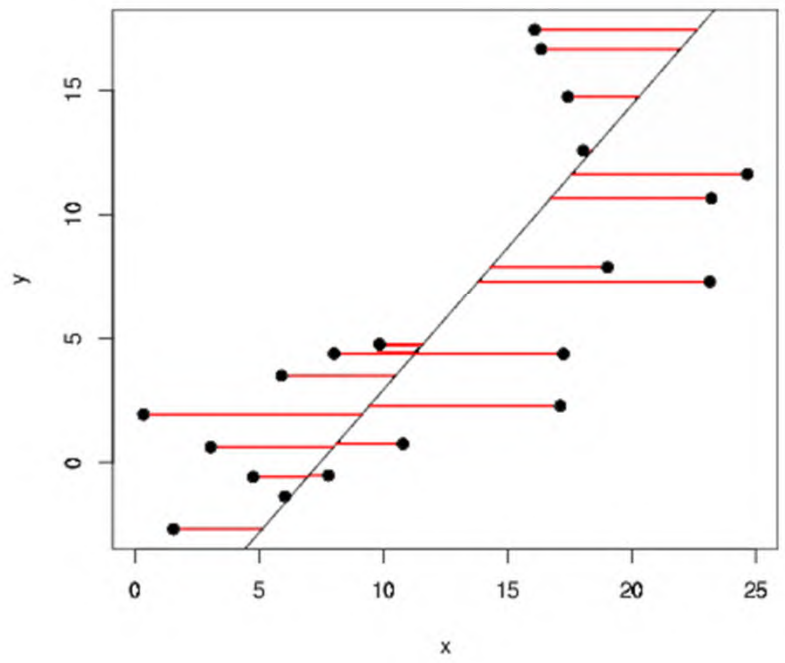
- Instead of predicting Y_h, predict the corresponding X_h, $\hat{X}_h$, given Y_h
  - Solve the fitted equation for X_h
  - $\hat{X}_h = \frac{Y_h - b_0}{b_1}$, where $b_1 \neq 0$
- Approximate CI can be obtained, see KNNL p. 169



Inverse Predictions/Calibration

- Technical applications
  - Validation of new instruments
  - Assessment of sample "unknowns" against a set of standard values





主观题 10分

设置

61

设想在同一个坐标系中同时画出 X 关于 Y 的回归线

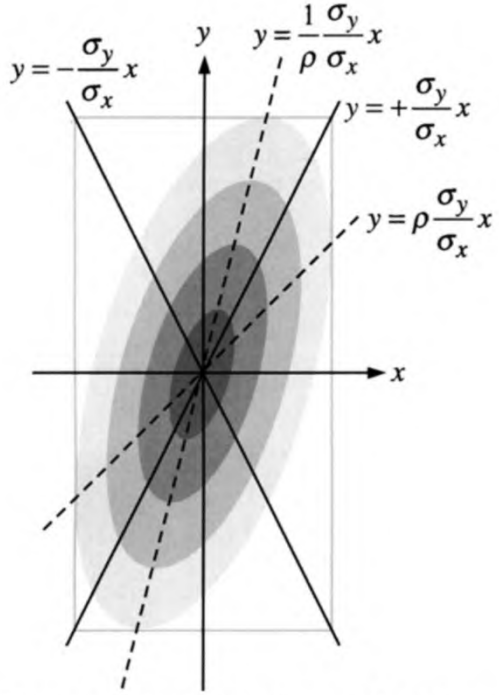
$$Y = b_0 + b_1 X$$

和 Y关于X 的回归线

$$X = a_0 + a_1 Y$$

(假设 $\sigma_x = \sigma_y$)

1. 假如 $Cor(X,Y) = 1$, 你会看到什么?
2. 假如 $Cor(X,Y) = -1$, 你会看到什么?
3. 假如 $Cor(X,Y) = 0$, 你会看到什么?



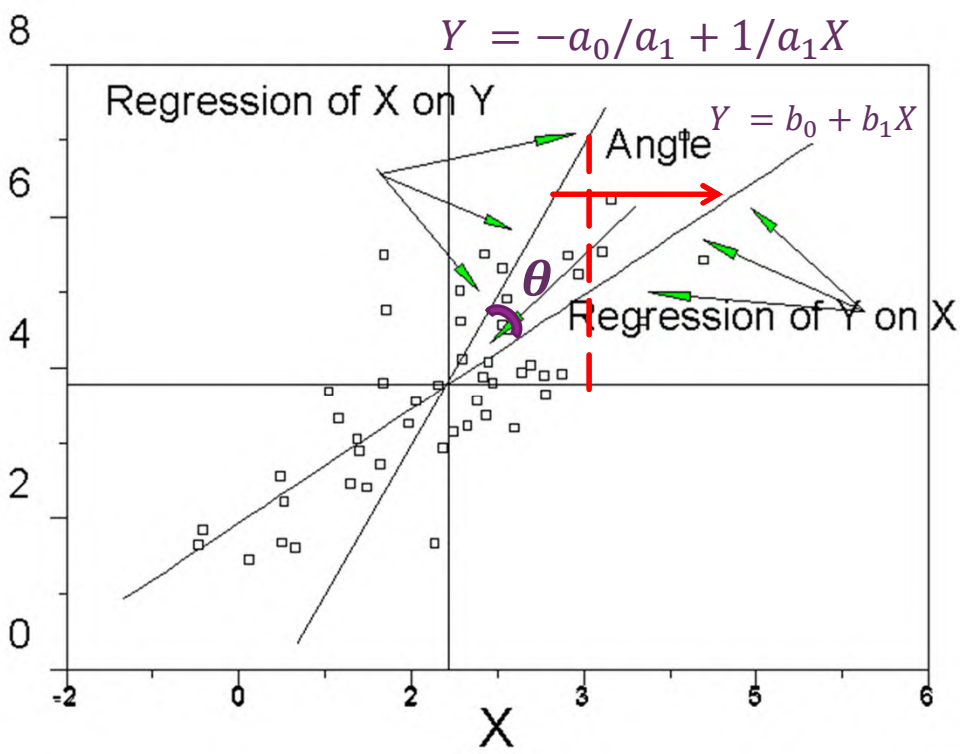
The Regression and SD Lines when $\mu_x = \mu_y = 0, \sigma_y = \sigma_x = 2$, and $\rho = +1/2$. The regression lines are dashed and the SD lines are solid.

作答

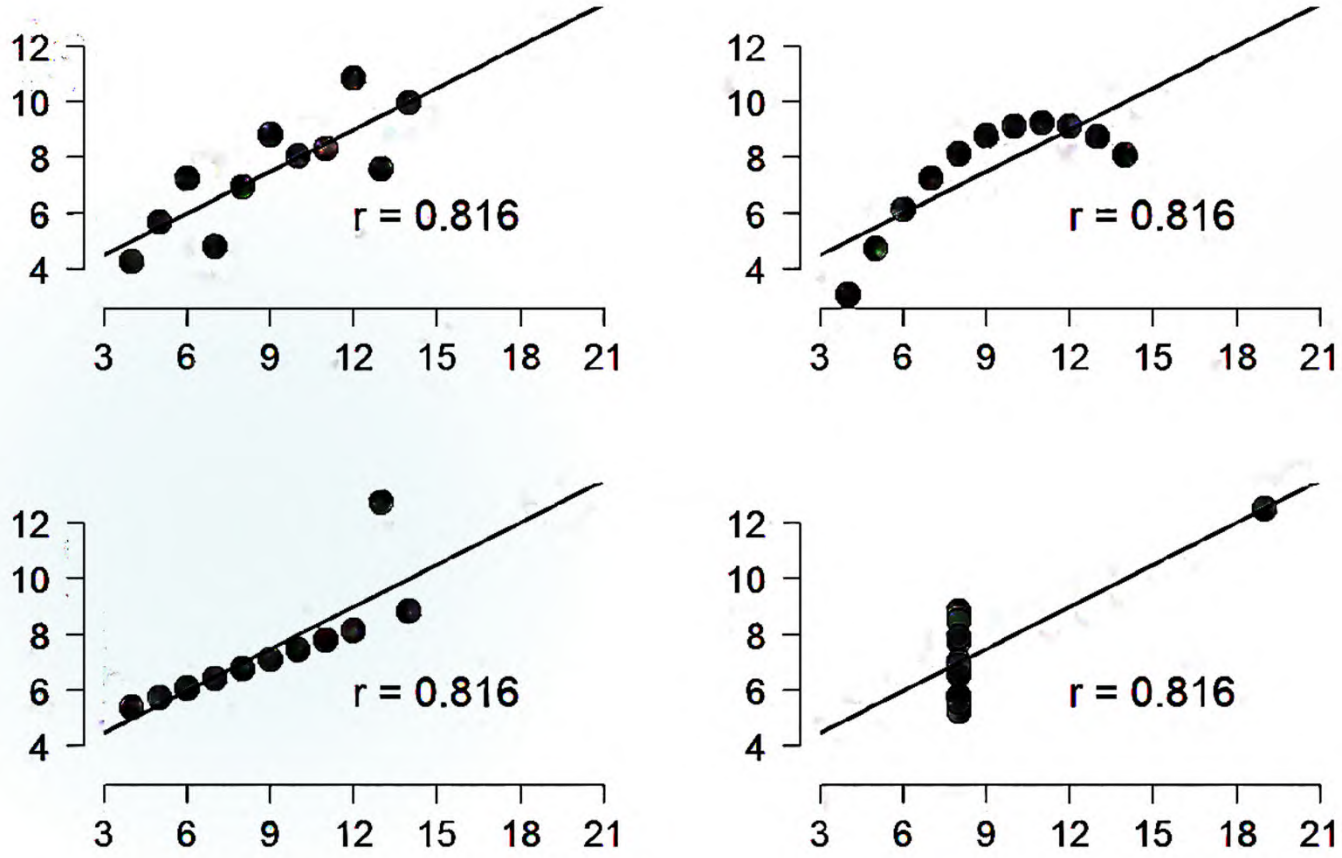
Angle between Two Regression Lines

- The geometric interpretation of r is the size of the angle θ between the regression of Y on X and of X on Y
  - The smaller θ, the larger the r
- $r = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2}\sqrt{\sum(Y_i-\bar{Y})^2}}$
- $Y = b_0 + b_1X;\ X = a_0 + a_1Y$
- $b_1 = r\frac{S_Y}{S_X};\ a_1 = r\frac{S_X}{S_Y};\Rightarrow a_1b_1 = r^2$
- $\tan(\theta) = \frac{\frac{1}{a_1}-b_1}{1+\frac{1}{a_1}b_1} = \frac{1-a_1b_1}{a_1+b_1} = \frac{1-r^2}{r\left(\frac{S_X}{S_Y}+\frac{S_Y}{S_X}\right)}$
  - When $S_X = S_Y, r = \frac{1}{1+\sqrt{1-cos\theta}}$

$0 \le a_1b_1 \le 1$
$\Rightarrow 0 < |b_1| \le 1/|a_1|$
Systematically away from $\bar{X}$

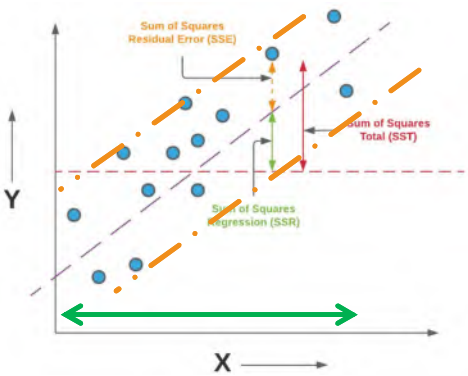


Always Plot Your Data!



Limitations of R^2

$$R^2 = \frac{SSR}{SST} = \frac{b_1^2 \sum(X_i - \bar{X})^2}{\sum(Y_i - \bar{Y})^2}$$



- `summary(mod1)$r.squared`
- R^2 does not measure **goodness of fit**
  - With a large σ^2, R^2 can be tiny even when the model is completely correct
  - R^2 can still be high, with a nonlinear term obviously missed
  - R^2 can be anywhere between 0 and 1 just by changing the range of X
- R^2 says nothing about **predictive ability**
  - MSE is a better measure of prediction error
- R^2 does not allow you to **compare** models using transformed responses
  - R^2 can easily go down when the model assumptions are better fulfilled
- R^2 does not measure how one variable **explains** another
  - If we regressed X on Y, we'd get exactly the same R^2

Adjusted R^2 does nothing to address any of these issues



第六讲 模型的补救以及补充议题

- 6.1 模型的补救措施
  - 6.1.1 模型的诊断补充
    - 6.1.1.1 复习残差图
    - 6.1.1.2 R中的诊断图
    - 6.1.1.3 失拟检验
  - 6.1.2 模型的补救措施
    - 6.1.2.1 非线性关系
    - 6.1.2.2 异方差
    - 6.1.2.3 非正态分布
    - 6.1.2.4 一般方法：做变换
- 6.2 补充议题
  - 6.2.1 过原点的回归
  - 6.2.2 反向预测



Objectives

- Diagnose a regression by the build-in plots in R.
- Understand leverage, and know how to detect outlying X values using leverages.
- Distinguish among outliers, high leverage observations and influential points.
- Distinguish between standardized residuals and studentized residuals.
- Know how to detect outliers by way of studentized residuals or studentized deleted residuals.
- Know that the (linear) lack of fit test only gives you evidence against linearity. If you reject the null, and conclude lack of linear fit, it doesn't tell you what (non-linear) regression function would work.
- Understand the "derivation" of the linear lack of fit test. That is, understand the decomposition of the error sum of squares, and how the expected mean squares tell us to use the ratio MSLF/MSPE to test for lack of linear fit.
- Know how to perform a Box-Cox transformation.
- Know the common transformations.
- Understand the limitations of a regression through the origin.
- Know the relationship of two regression lines.
- Understand the cautions necessary in using the R^2 value as a way of assessing the strength of the linear association.



Homework

- Prob-1. KNNL 3.9
- Prob-2. KNNL 3.15(Data Set: CH03PR15.txt)
- Prob-3. KNNL 3.16
- Prob-4. Consider the following data set (on the next slide) that describes the relationship between "velocity" of an enzymatic reaction (V) and the substrate concentration (C). A common model used to describe the relationship between "velocity" and concentration is the Michaelis-Menten model,

$$V = \frac{\theta_1 C}{\theta_2 + C}$$

where θ_1 is the maximum velocity of the reaction and θ_2 describes how quickly (in terms of increasing concentration) the reaction will reach maximum velocity. With this model, $1/V$ can be written as a linear model with explanatory variable $1/C$.

$$\frac{1}{V} = \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1}\frac{1}{C} = \beta_0 + \beta_1 \frac{1}{C}$$



Homework Cont'd

You are asked to investigate whether this linear transformation results in a good or poor fit by doing the following steps:

- a). Generate a scatterplot of V vs C. Comment on the shape.
- b). Define new variables for $1/V$ and $1/C$ and generate a scatterplot.
- Does the fit appear linear? Does there appear to be any violation of assumptions?
- c). Is the distribution of $1/C$ different than C? Are there any points that may be more influential in determining the fit?
- d). Determine the least squares regression line for $1/V$ vs $1/C$. Save the residuals and predicted values and generate a residual plot. Does the residual plot suggest any problems with assumptions?
- e). Generate a scatterplot of V versus C and add the predicted values of V from your linear regression model (i.e., transform back to the original units). Comment on the fit.

Concentration	Velocity
0.02	76
0.02	47
0.06	97
0.06	107
0.11	123
0.11	139
0.22	159
0.22	152
0.56	191
0.56	201
1.10	207
1.10	200



Background Reading

- 3.8 Overview of remedial measures
- 3.9 Transformations

- Next class we will do simple regression with vectors and matrices so that we can generalize to multiple regression
- Scan through KNNL 5.1 to 5.7 if this is unfamiliar to you



下节预告：

- 下一次课我们将进入崭新的*Multiple Linear Regression*征程！
- 如何用矩阵和向量的形式来描述回归模型？
  - 相应的参数估计和分布函数又该怎么写？
- 线性回归的几何直观是什么？
- 你知道调整的 R^2 吗？好好的为什么要“调整”？
- …
- 期待下次课与你相约 MLR ~~~

