Lecture 5-
Model Assumptions and Diagnostics

周在莹

清华大学统计与数据科学系

zyzhou@mail.tsinghua.edu.cn



本节揭晓：

- 简单线性回归模型的假设LINE，你还记得吗?
- 所以你明确回归诊断的入手点了吗?
  - 我们要查看什么?
  - 具体要怎么看?
  - 图形办法？正式检验?

A typical regression analysis involves the following steps:

1. Model formulation
2. Model estimation
3. Model evaluation
4. Model use



学习目标（精简版）

- 掌握
  - ✓ 回归需要诊断的方面
  - ✓ 偏度和峰度的概念
  - ✓ 做残差图
- 理解
  - ✓ 为什么关心 X 的“分布”
  - ✓ 残差图的作用
  - ✓ 离群值对回归结果的影响
- 了解
  - ✓ 同方差、正态性、相关性的常用检验方法



Outline

- 1. Model Assumptions
- 2. Diagnostics to Check Model Assumptions
  - 2.1 Diagnostics concerning X
  - 2.2 Diagnostics using the residuals

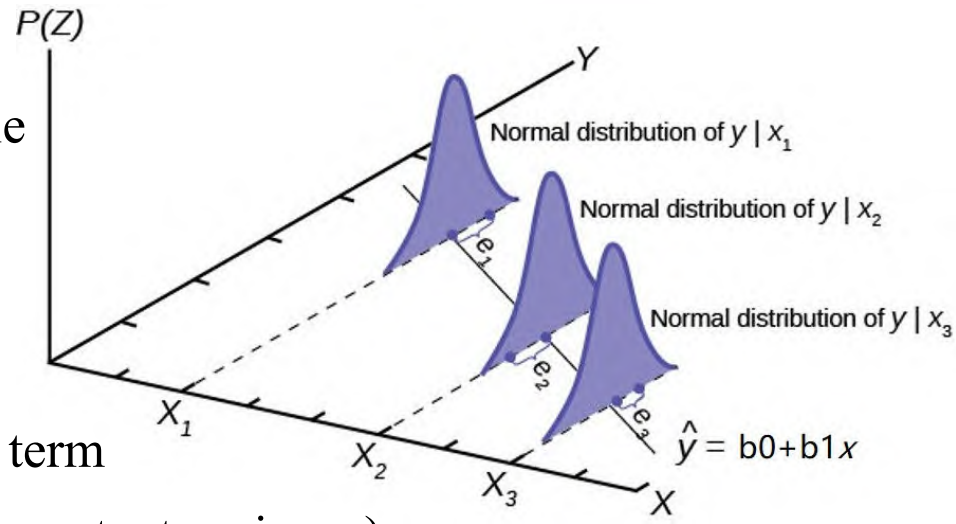


Simple Linear Normal Error Regression Model

- Observe cases
  - $Y_i = i^{th}$ response or outcome variable
  - $X_i = i^{th}$ explanatory or predictor variable
- Assume

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

  - β_0 is the intercept, β_1 is the slope
  - ε_i is a normally distributed random error term
  - $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2(\varepsilon_i) = \sigma^2$ (constant variance)
  - $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent





Fitted Regression Model and Residuals

- $\hat{Y}_i = b_0 + b_1 X_i$
  - b_0 is the estimated intercept
  - b_1 is the estimated slope
- e_i: residual for the i^{th} case
- $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$





Assumed Features of the Model (LINE)

- Mean of Y_i is a linear function of X_i
  - $E(Y_i) = \beta_0 + \beta_1 X_i$ (linearity)
- $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent (independence)
- $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are normal distributed
  - $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \sim N(0, \sigma^2)$ (normality)
- Variance of Y_i is constant, not dependent on X_i
  - $Var(Y_i) = \sigma^2$ (equal/constant variance)



Diagnostics and Remedies

- Diagnostics: Use visualization tools (graphs) and/or formal procedures to check whether the assumptions of the model are violated
- Consequences of violations
  - Inference results cannot be trusted
    - t tests, F tests, P-values not correct
    - Coverage of confidence intervals off
- Remedial Measures:
  - Use transformation and other methods to fix/adjust for the problems
  - Change to different models/analytic strategies to accommodate violations



单选题 1分

设置

9

线性回归模型的概率方面的假设都会体现在 Y 的概率分布上，那么我们是否应该查看 Y的概率分布，以核实其满足正态分布、同方差、独立?

当然

不是

不确定



提交

Look at the Data

- Before trying to describe the relationship between a response variable (Y) and an explanatory variable (X), we should look at the 'distributions' of these variables
- We should always look at X
- If Y depends on X, looking at Y alone may not be very informative – looking marginally at Y is a common mistake



Diagnostics for *X*

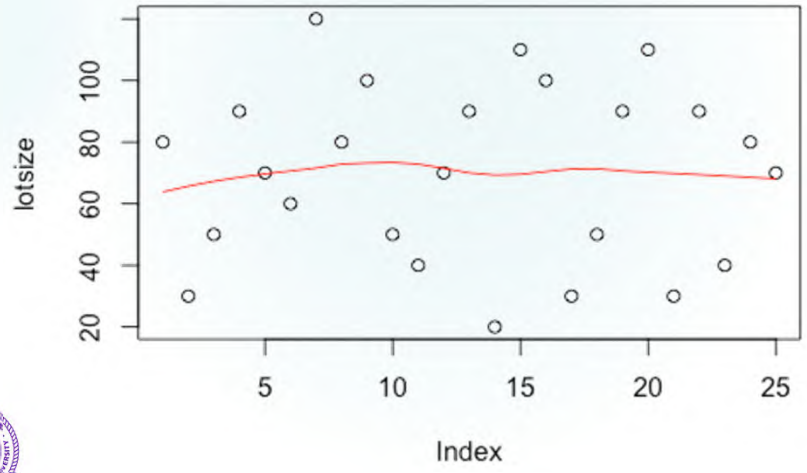
- Though no particular or explicit assumptions are imposed on *X*, the "distribution" or design of *X* is important
  - If *X* has many values, use '`summary`' to get numerical summaries (e.g., mean, median, quartiles)
  - If *X* has only a few values, use '`table`' to get counts
- Examine the "distribution" of *X*
  - Is it skewed?
  - Are there extreme values?
- Can inflate variance or influence trend

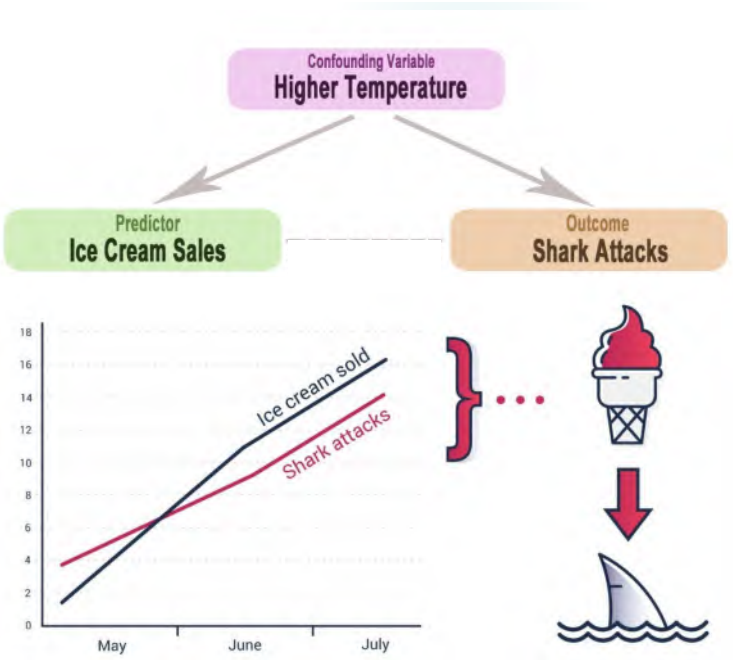
- Good design not only leads to more power in statistical inference, it also ensures the validity of the whole process
- Bad design does the opposite



Trend in Order

- Sequence plots can be useful for identifying potential problems with independence

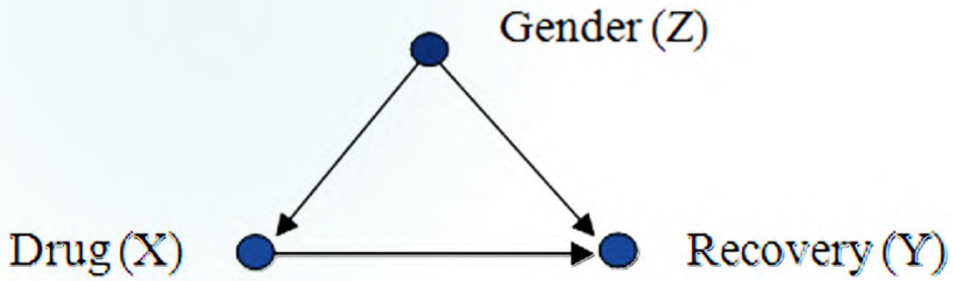





```
> plot(lotsize)
> lines(lowess(lotsize),
        col = 'red')
```

Notes on Confounding

- Do the values of *X* depend on time (i.e., the order in which they were collected) or other types of order ?
  - Suggests possible confounding factors
- Confounders may alter the relation between *X* and *Y*
- Were any important predictors excluded from the model?





Important Statistics

… is a measure of

- Mean - Location
- Standard deviation - Variability
- Skewness (偏度) - The lack of symmetry

$$g_1 = \frac{m_3}{{m_2}^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}},$$

- Kurtosis (峰度) - Heavy/light-tailed relative to a normal distribution
- Range - Spread

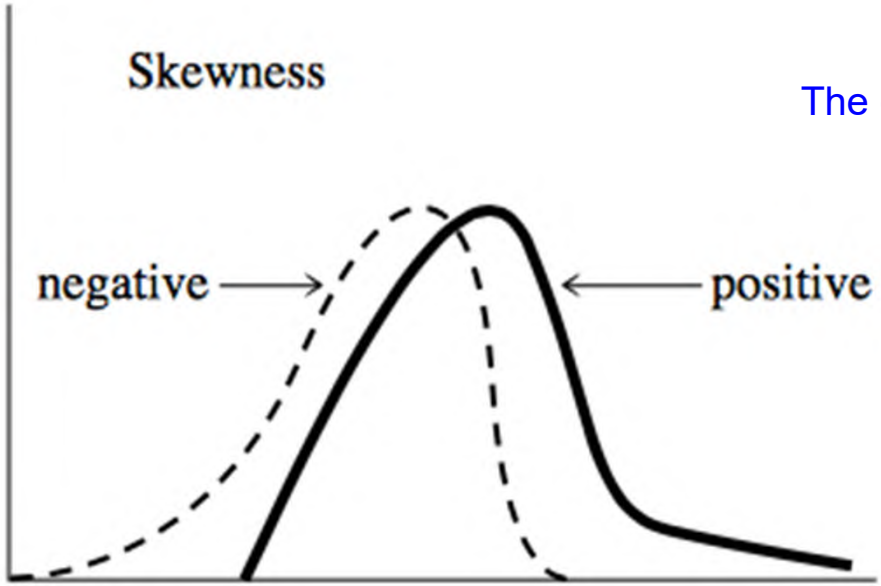
$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3$$

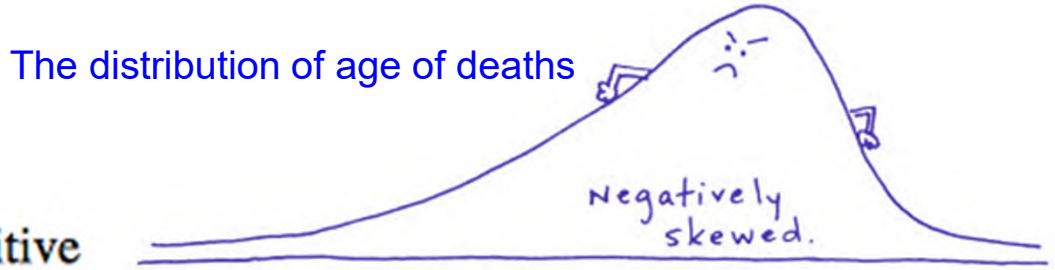


Skewness

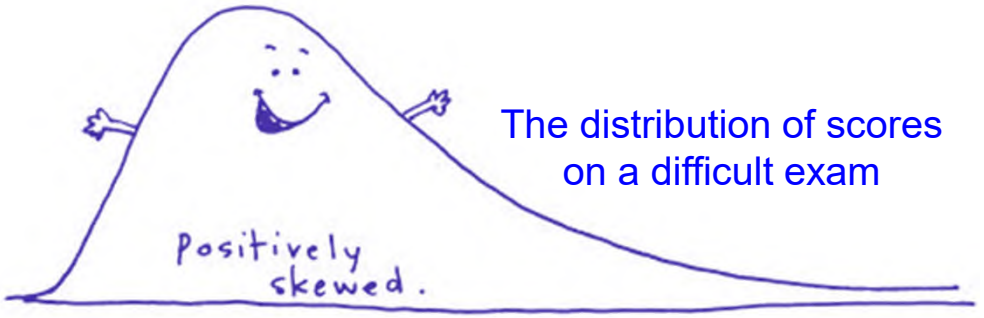
$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

- $g_1 < 0$： skewed left- the left tail is long relative to the right tail
- $g_1 > 0$： skewed right- the right tail is long relative to the left tail





The distribution of age of deaths



The distribution of scores on a difficult exam

单选题 1分

设置

16

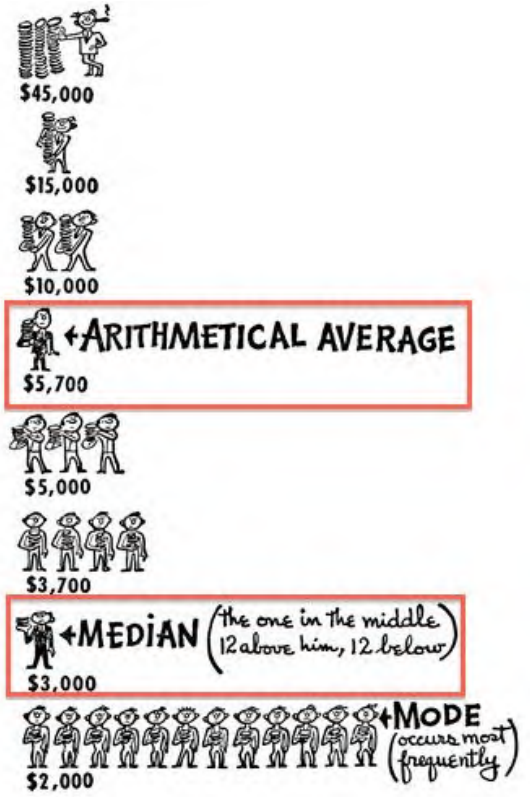
你觉得图中收入数据是左偏还是右偏的?

左偏

右偏

无法判断

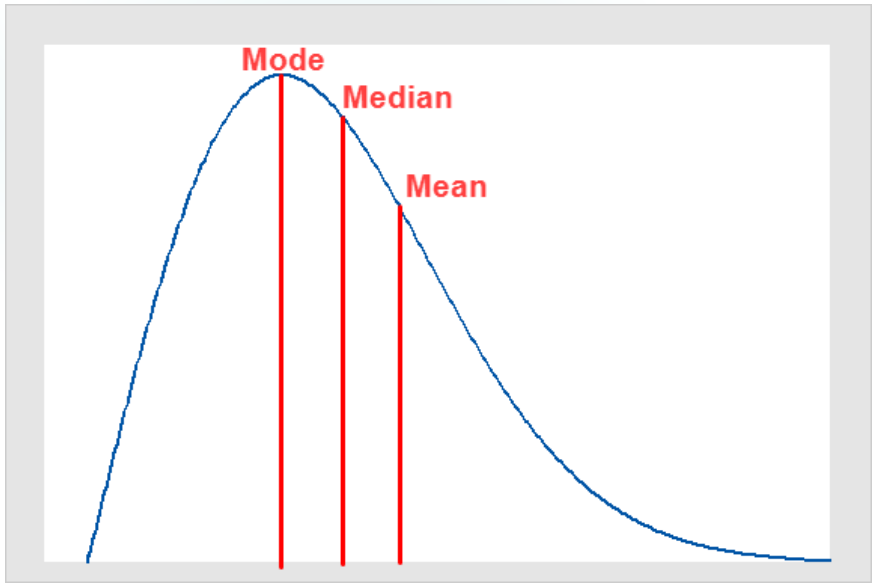
都有可能



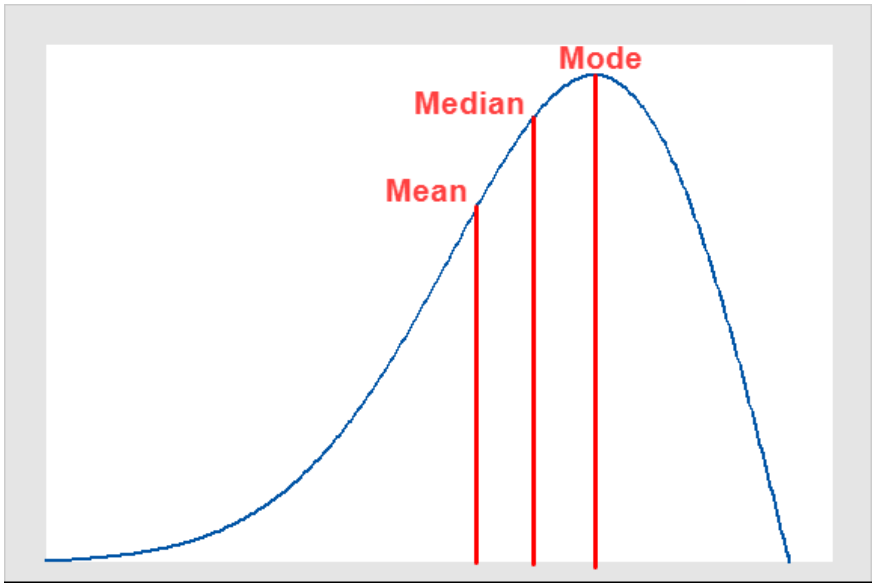
提交

Measures of Central Tendency

- The mean, median, and mode are all equal in the normal distribution and other symmetric distributions
- **Generally**:
  - **Right-skewed**



  - **Left-skewed**



Skewness: Key Takeaways

- Its sign reflects the direction of skewness
  - Distributions can exhibit right (positive) skewness or left (negative) skewness to varying degrees
- It compares the sample with a normal (symmetric) distribution
  - Skewness is the degree of distortion from the symmetrical bell curve
  - Values far from zero suggest a non-normal (skewed) population
- If the data are multi-modal, it may affect the sign of the skewness



Kurtosis

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

- How far rare measurements scatter from the mean
- $g_2 = 0$ for a normal distribution
- g_2 is sometimes referred to as the excess kurtosis, where the term kurtosis is defined as:

$$G_2 = \frac{m_4}{m_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

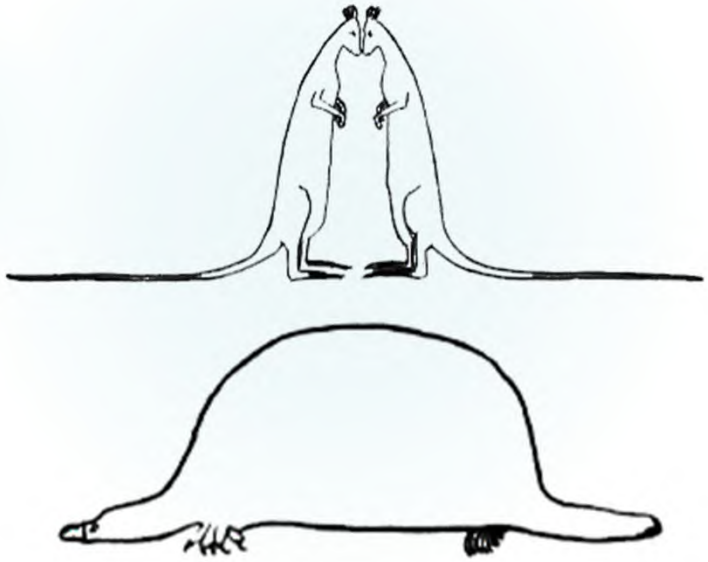
- When using software to compute the sample kurtosis, you need to be aware of which convention is being followed

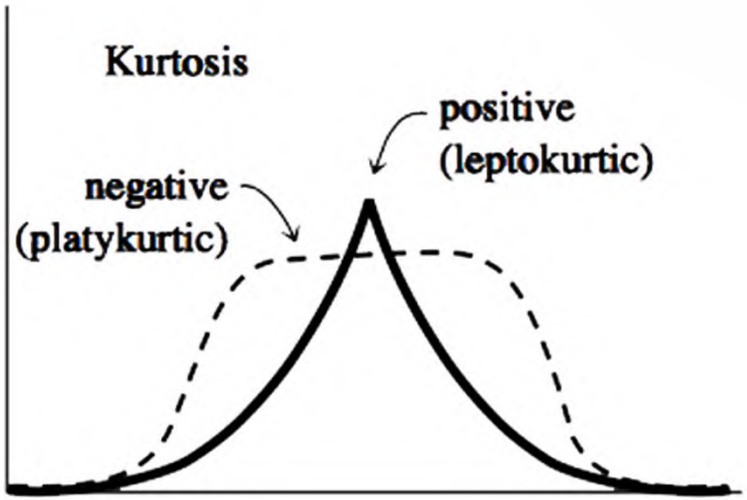


Kurtosis: Leptokurtic(尖峰态) or Platykurtic(低峰态)

- $g_2 > 0$: a heavy tail distribution, called a leptokurtic distribution, where “lepto-" means "slender”
- $g_2 < 0$: a light tail distribution, called a platykurtic distribution, where “platy-" means "broad”

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3$$





单选题 1分

设置

21

尖峰态的分布，其方差与低峰态的分布相比较，会

较小

较大

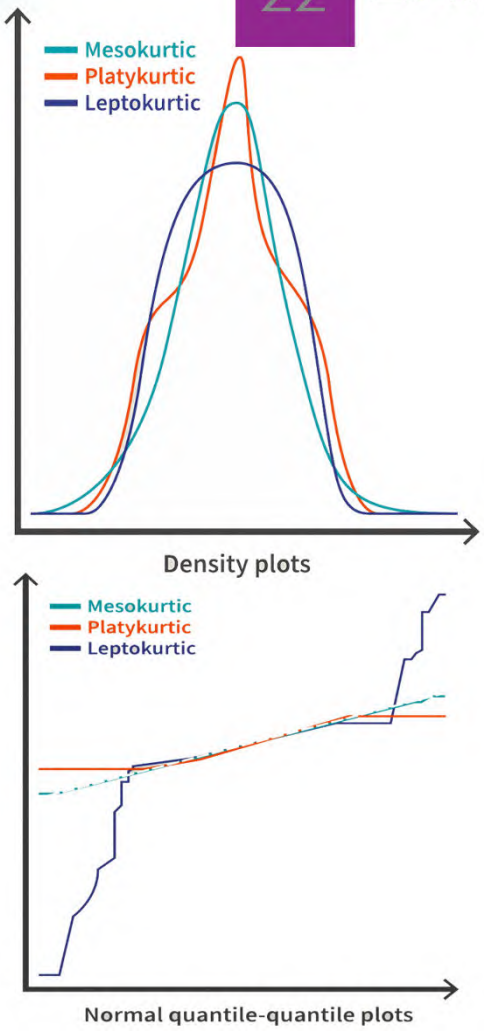
相等

无法判断

提交

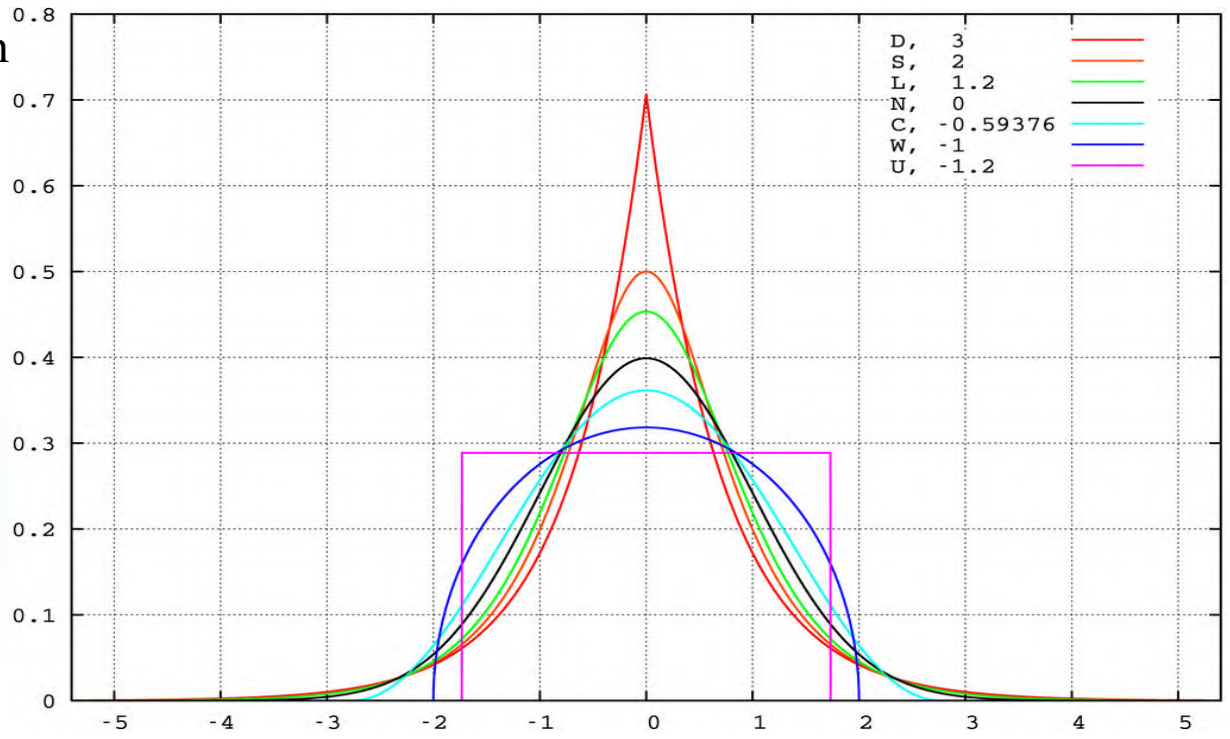
Kurtosis: Common Misconceptions

- 1. Kurtosis is defined solely in terms of peakedness, with no mention of the importance of the tails
  - **Tailedness** and peakedness are both components of kurtosis
- 2. The relation between the peak and tails of a distribution with excess kurtosis is described or illustrated incorrectly
  - Positive kurtosis indicates peakedness and **heavy** tails and negative kurtosis flatness and **light** tails
- 3. Descriptions and illustrations of kurtosis fail to distinguish between kurtosis and the variance
  - The kurtosis measure is scaled with respect to the variance, so it is not affected by it (it is **scale free**). Kurtosis reflects the shape of a distribution apart from the variance



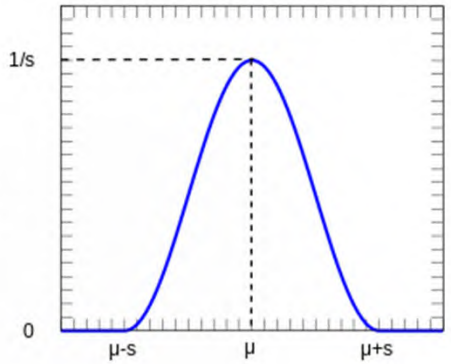
Kurtosis: Examples

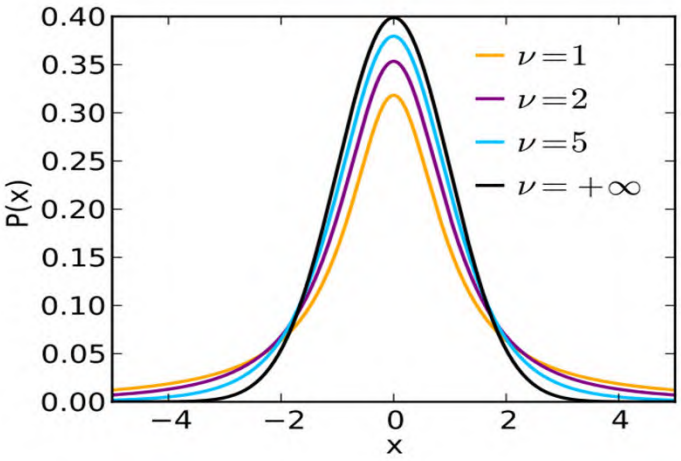
- D: Laplace distribution, also known as the double exponential distribution, red curve
- S: hyperbolic secant distribution, orange curve
- L: logistic distribution, green curve
- N: normal distribution, black curve
- C: raised cosine distribution, cyan curve
- W: Wigner semicircle distribution, blue curve
- U: uniform distribution, magenta curve



Kurtosis: Usefulness

- Super-Gaussian: leptokurtic distributions including the Student's t, Rayleigh, Laplace, exponential, Poisson and the logistic
- Sub-Gaussian: platykurtic distributions including the continuous or discrete uniform, and the raised cosine
  - The most platykurtic distribution: Bernoulli $(1/2)$ (a coin toss), $g_2 = -2$
- Application: Kurtosis risk. When analyzing historical returns, kurtosis can help an investor gauge an asset's level of risk





单选题 1分

设置

25

在查看金融投资的历史回报数据时，出于尽量规避风险的考虑，你倾向于选择低峰分布的还是尖峰分布的哪一个？

选低峰的

选尖峰的

都可以吧

不太明白

提交

Kurtosis: Key Takeaways

- Whereas skewness differentiates extreme values in one versus the other tail, kurtosis measures extreme values in either tail
- It compares the sample with the standard normal distribution
  - Distributions with large kurtosis (leptokurtic) exhibit tail data that are generally more extreme than the tails of the normal distribution
  - Distributions with low kurtosis (platykurtic) have a lower likelihood of extreme events as compared to a normal distribution



Notes: Why Compared with Normal Distributions

- Regression model does **NOT** assume that X comes from a single Normal distribution
  - X can follow any distribution
  - Or simply, we see X as fixed (not a r.v.)
- Regression model does not state that Y comes from a single Normal distribution
  - Assume $(Y|X_i)$ is Normal
- In some cases, however, X and/or Y may be Normal and it can be useful to know this



Example: Toluca lot size

```
> attach(toluca)
> summary(lotsize)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
20	50	70	70	90	120

- Quartiles- divide a rank-ordered data set into four equal parts
- The interquartile range (IQR) = Q3 − Q1



Moments

- `install.packages("moments")`

Moments			
N	25	**SumWeights**	25
Mean	70	**SumObservations**	1750
StdDeviation	28.7228	**Variance**	825
Skewness	-0.0969	**Kurtosis**	-1.106
UncorrectedSS	142300	**CorrectedSS**	19800



Location and Spread

Basic Statistical Measures			
Location			**Variability**
Mean	70.00000	**StdDeviation**	28.7228
Median	70.00000	**Variance**	825.00000
Mode	90.00000	**Range**	100.00000
		InterquartileRange	40.00000



Stem-and-Leaf Plots

```
> stem(lotsize)
```

The decimal point is 2 digit(s) to the right of the |

0 | 233344
0 | 55567778889999
1 | 00112

```
> stem(lotsize, scale = 3)
```

The decimal point is 1 digit(s) to the right of the |

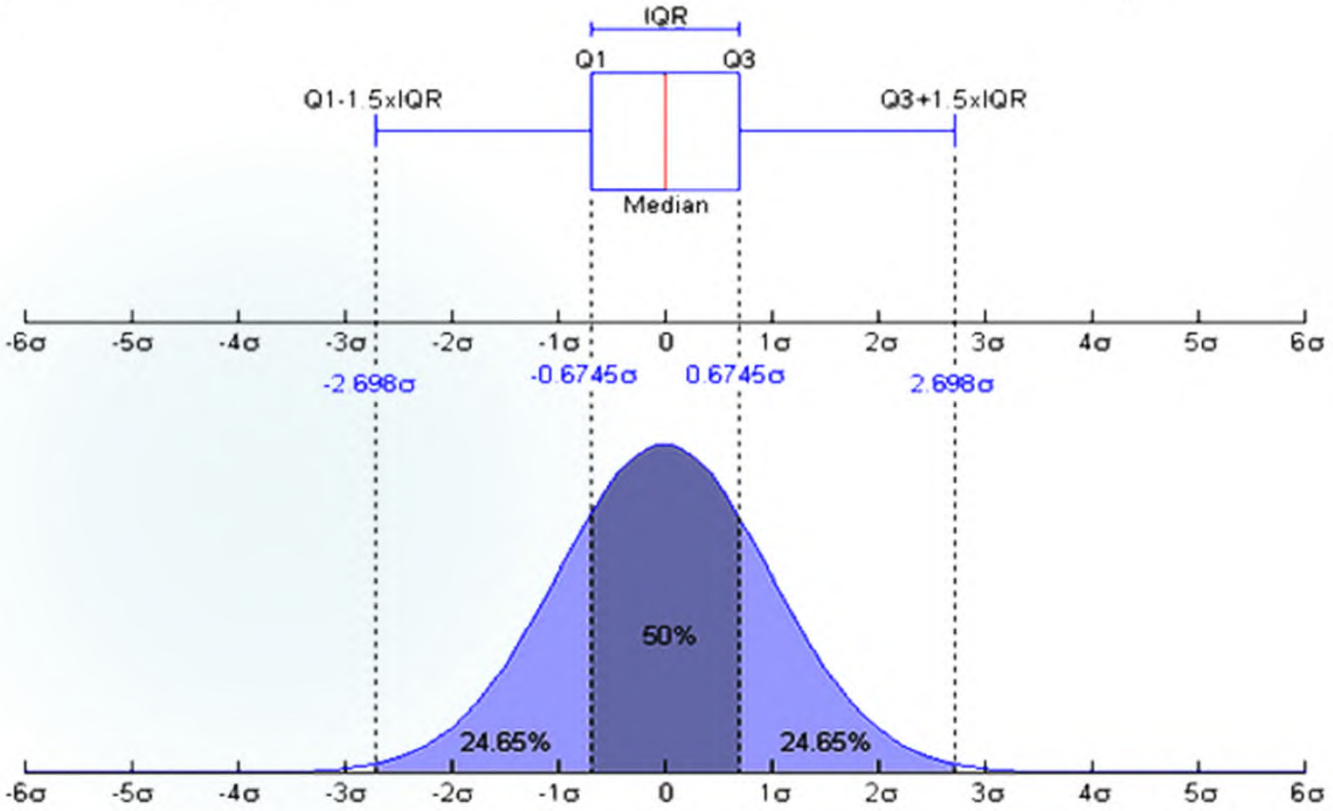
2 | 0
3 | 000
4 | 00
5 | 000
6 | 0
7 | 000
8 | 000
9 | 0000
10 | 00
11 | 00
12 | 0

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	



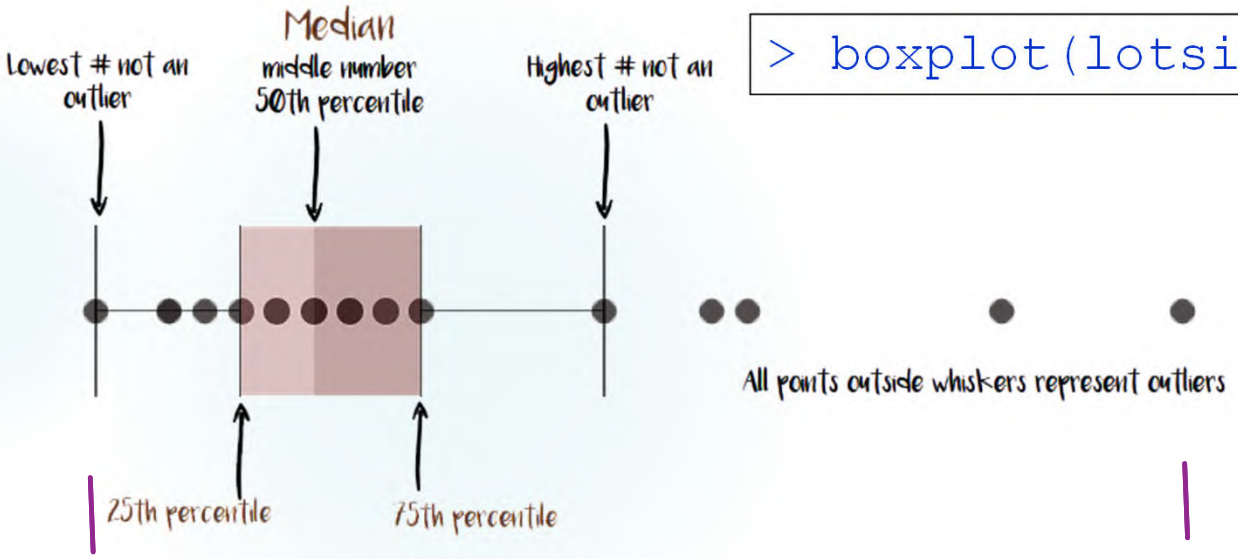
Box Plots

- The spacings between the different parts of the box help indicate the degree of dispersion (spread) and skewness in the data and identify outliers

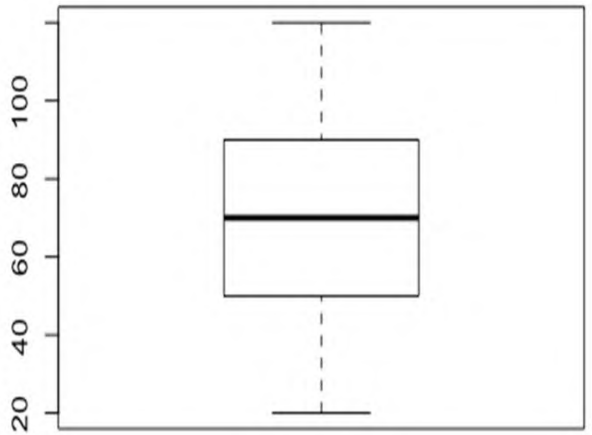


Box Plots

- The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box




```
> boxplot(lotsize)
```



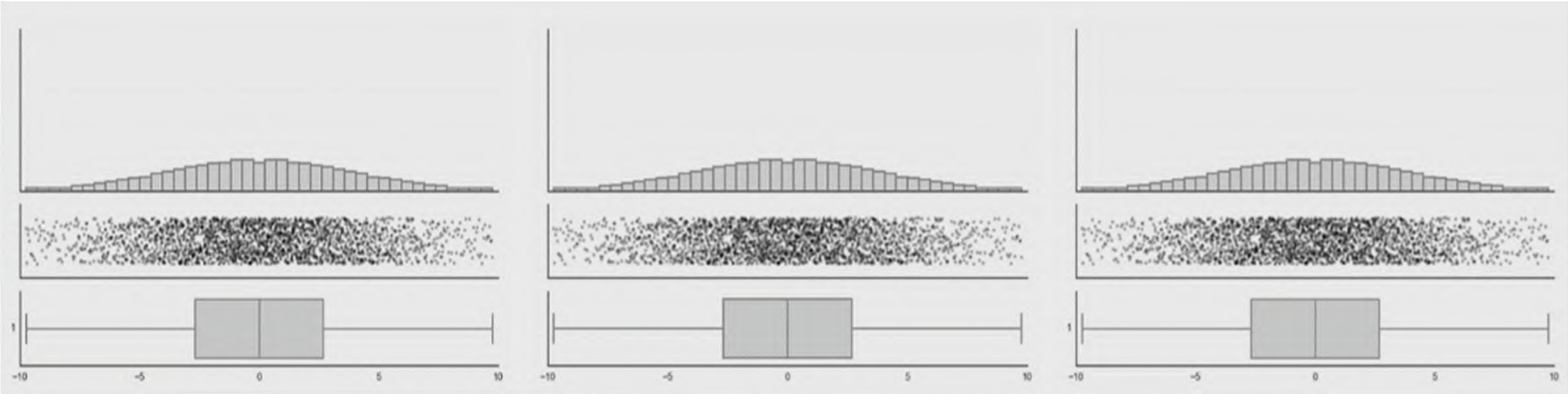

```
> boxplot(x,range=0)
```

主观题 10分

设置

34

请评论下面的图：



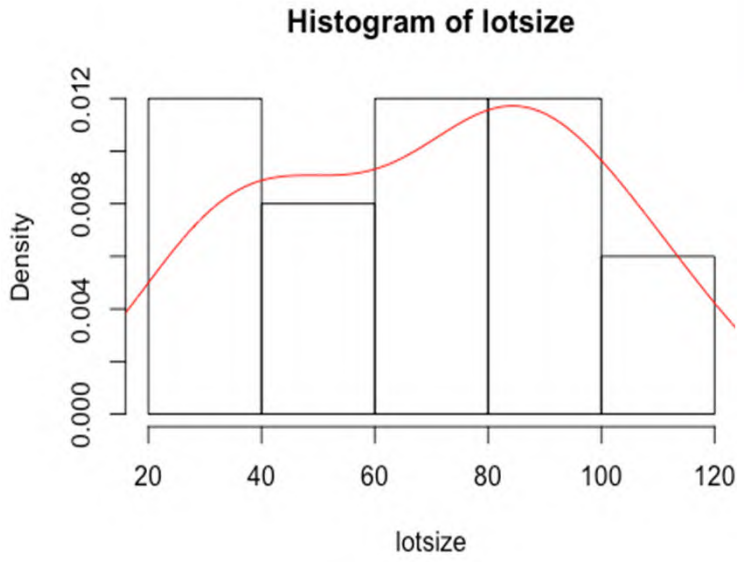
作答

Histogram Plots

```
> hist(lotsize)
```

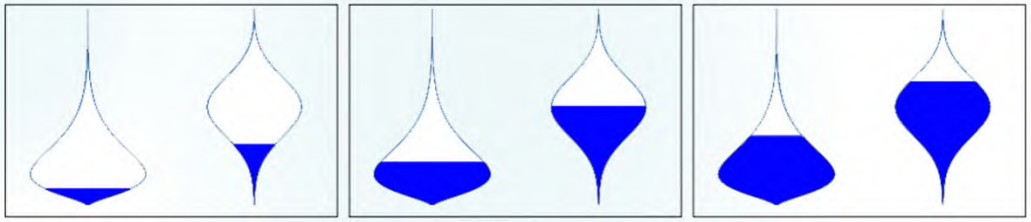



```
> hist(lotsize,freq=F)
> lines(density(lotsize),col='red')
```



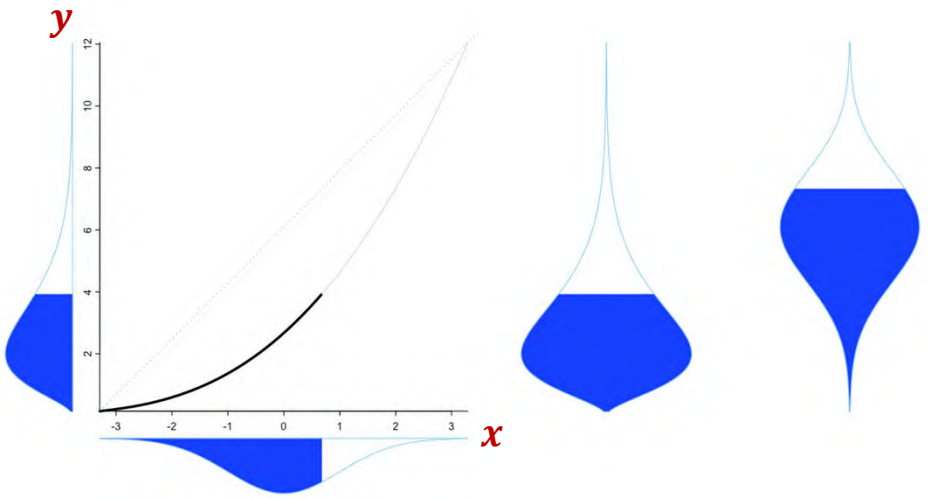
Quantile-Quantile Plots, Revisited

- General quantile-quantile (Q-Q) plot
  - F and G be the CDFs of two distributions
  - F^{-1} and G^{-1} the inverse CDF or quantile functions $F^{-1}(p) = \inf\{y: F(y) \geq p\}$
  - The Q-Q plot plots $F^{-1}(q)$ against $G^{-1}(q)$ for $0 < q < 1$



The pairs are 10% full, 50% full, 80% full

CDF = 0.1, 0.5, 0.8…



Normal Quantile Plots

- Normal quantile/probability plot is to plot sample quantiles (order statistics) against standard normal quantiles
  - Sample: $X_1, X_2, \ldots, X_n$.
  - Order statistics: The realized quantiles $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, corresponding to $p = 1/n, 2/n, \ldots, n/n$
  - Z scores: Quantiles of standard normal distribution, $Z_i = \Phi^{-1}(p_i)$, where $p_i = (i - 0.5)/n$ for $i = 1, 2, \ldots, n$
  - The normal probability plot: Plotting $X_{(i)}$ *versus* Z_i for $i = 1, 2, \ldots, n$
- If a sample can be considered drawn from a normal distribution, we expect to see the points roughly follow a linear line:

$$X_{(i)} = \mu + \sigma Z_i$$

  - Departure from a linear line indicates departure from normality

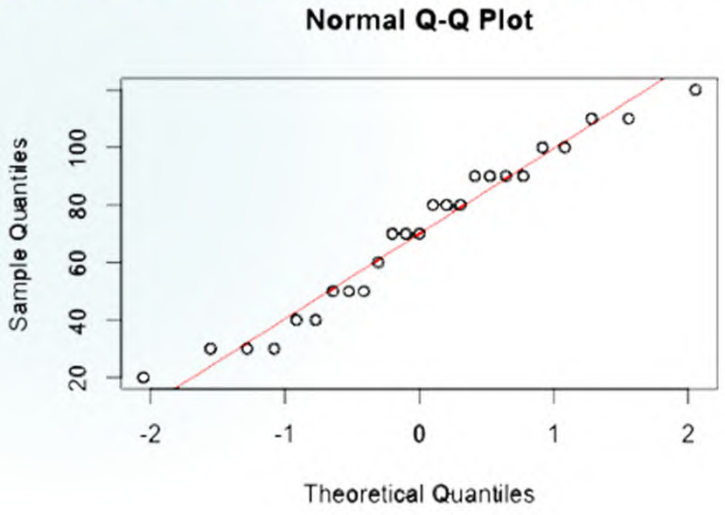


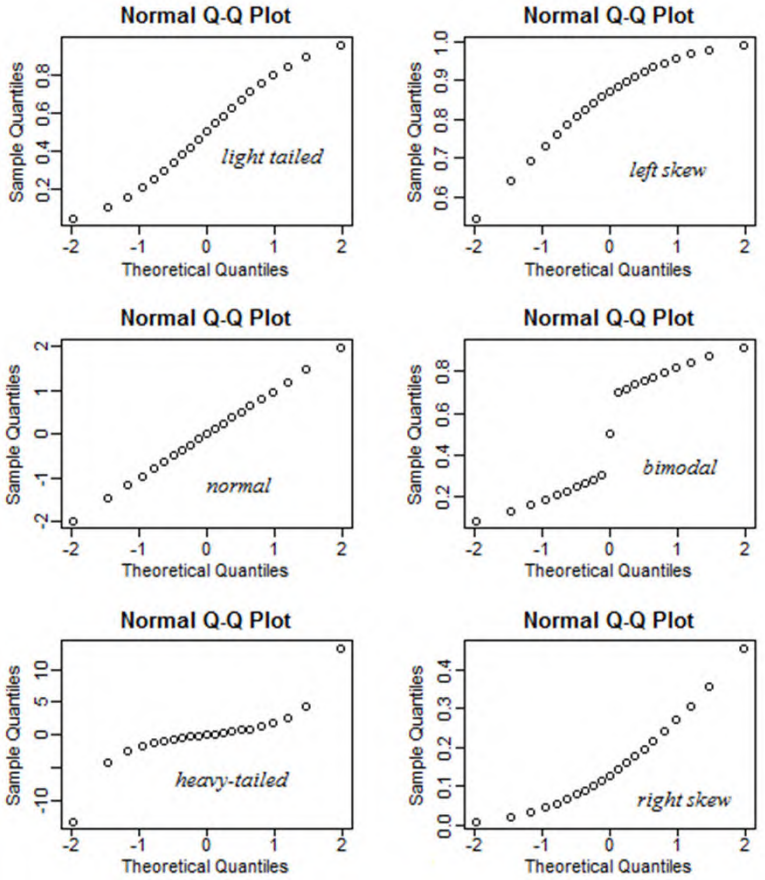
QQ Plots: Applications

$$X_{(i)} = \mu + \sigma Z_i$$

- Can also reveal the shape of the underlying distribution, such as long/short tails, skewness, etc
- KNNL plots $X_{(i)}$ against sZ_i, which does not alter the nature of the plot

```
qqnorm(lotsize)
qqline(lotsize, col = 'red')
```





Diagnostics Using Residuals

- Recall residuals $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$
- Show model fit and approximate error term ε_i
- Can be used for multiple diagnostic purposes:
  - Is the linear model/relationship appropriate?
  - Is the error variance constant?
  - Are the errors normal?
  - Are the errors independent?
  - Are there outliers?



Is the Relationship Linear?

- Plot Y_i vs X_i
- Plot e_i vs X_i or Y_i (residual plot)
- Residual plot better emphasizes deviations from linear pattern
- Recall Lecture 2 plots of scatterplot and residual plot



投票 最多可选1项

设置

41

如果一个回归的决定系数 $R^2 = 0.69$, 凭此你是否有把握认定该组数据的线性拟合较好?

A 可以

B 不行

C 不确定



提交

R Code: Synthetic #1

- Generates data set where $Y = X^2 - 10X + 30 + \varepsilon$
- Errors are Normally distributed with $\sigma = 25$

```
> x = seq(1, 30, by=1)
> n = length(x)
> y = x^2 - 10*x + 30 + 25*rnorm(n)
> a100 = data.frame(x, y)

> fit1 = lm(y ~ x, data=a100)
> anova(fit1)
> summary(fit1)
```



Output

$$Y = X^2 - 10X + 30 + \varepsilon$$

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	1016230	1016230	182.52	8.658e-14 ***
Residuals	28	155898	5568		

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-138.179	27.942	-4.945	3.22e-05 ***
x	21.264	1.574	13.510	8.66e-14 ***

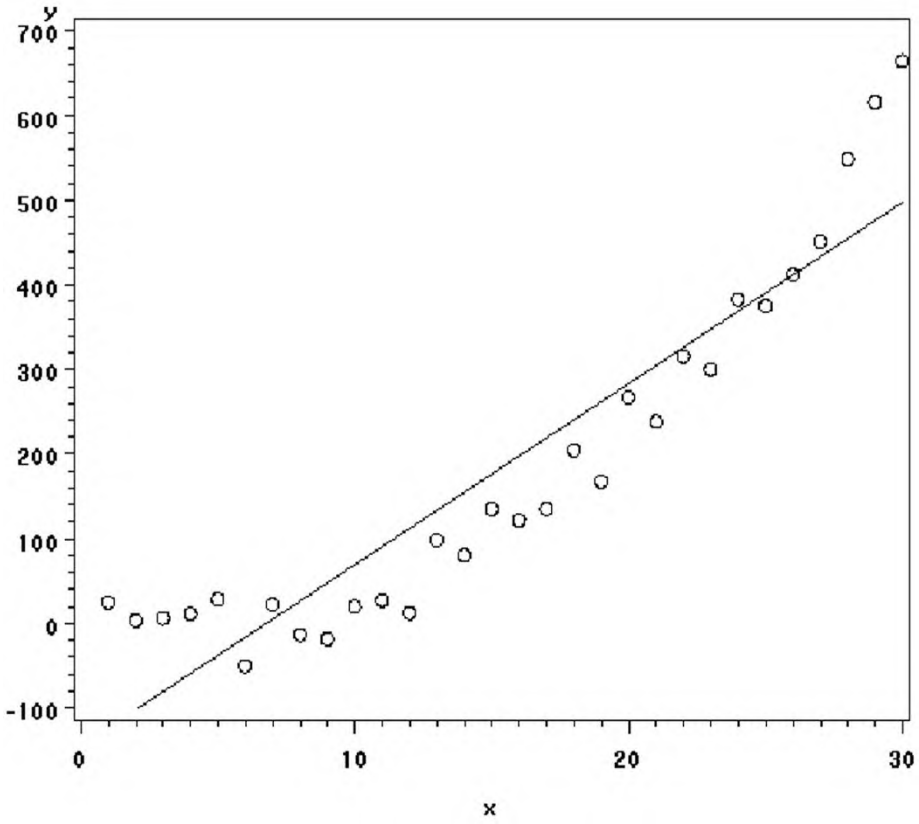
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



R Code: Visual Checks

```
> plot(y ~ x, data=a100)
> abline(fit1)
```

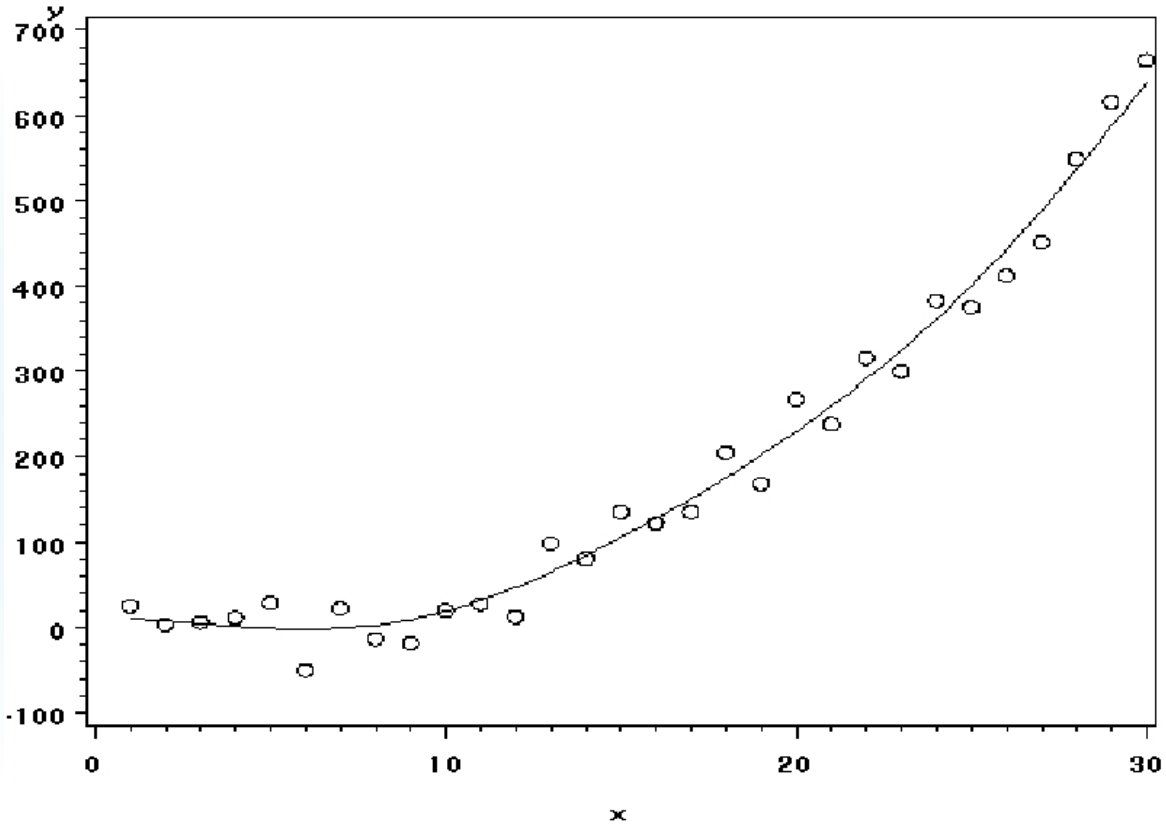
Scatterplot with regression line





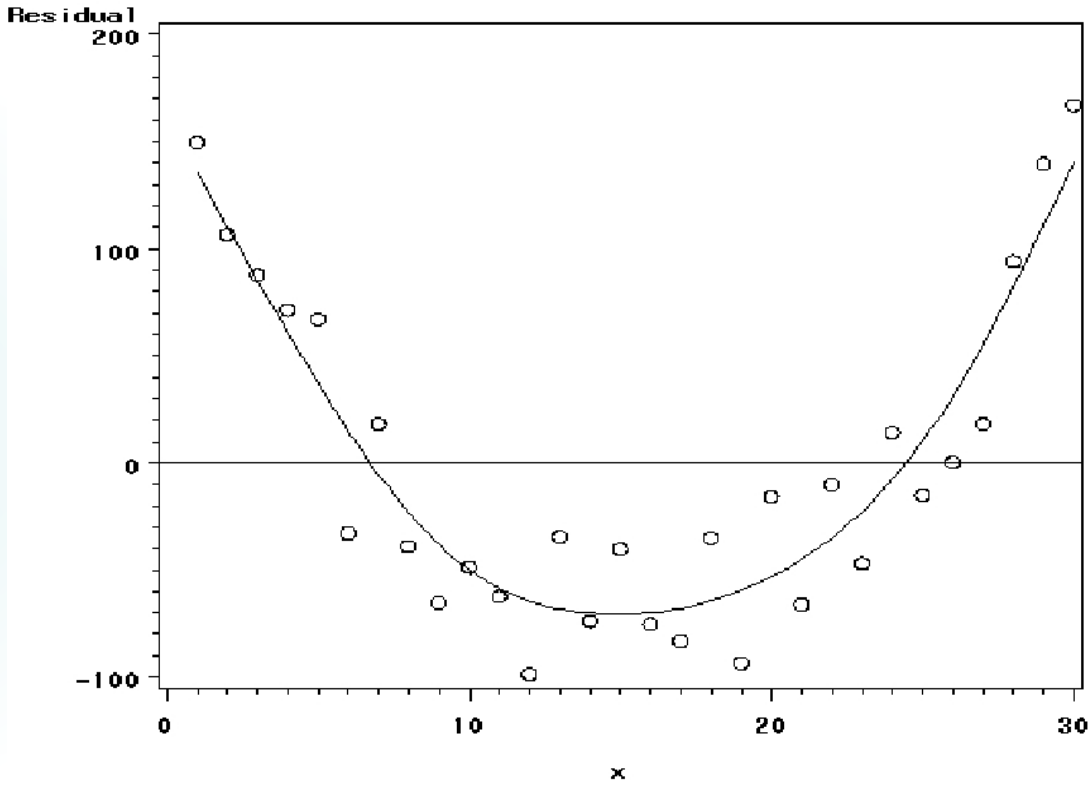
```
> scatter.smooth(x,y)
```

Scatterplot with smoothed curve



```
> scatter.smooth(x,fit1$residuals)
> abline(h=0)
```

Residual plot



Does the Variance Differ across X ?

- Plot Y_i vs X_i
- Plot e_i vs X_i or $\hat{Y}_i$
- Plot of e_i vs X_i will emphasize problems with the variance assumption



R Code: Synthetic #2

- Generates data set where $Y = 30 + 100X + 10X\varepsilon$
- Errors are Normally distributed with $\sigma = 10X$

```
>   x = seq(1, 100, by=1)
>   n = length(x)
>   y = 30 + 100*x + 10*x*rnorm(n)
> a100a = data.frame(x, y)

> fit2 = lm(y ~ x, data = a100a)
> anova(fit2)
> summary(fit2)
```



Output

$$Y = 30 + 100X + 10X\varepsilon$$

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	845726258	845726258	2431.1	< 2.2e-16 ***
Residuals	98	34091704	347875		

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	71.299	118.852	0.60	0.55
x	100.746	2.043	49.31	<2e-16 ***

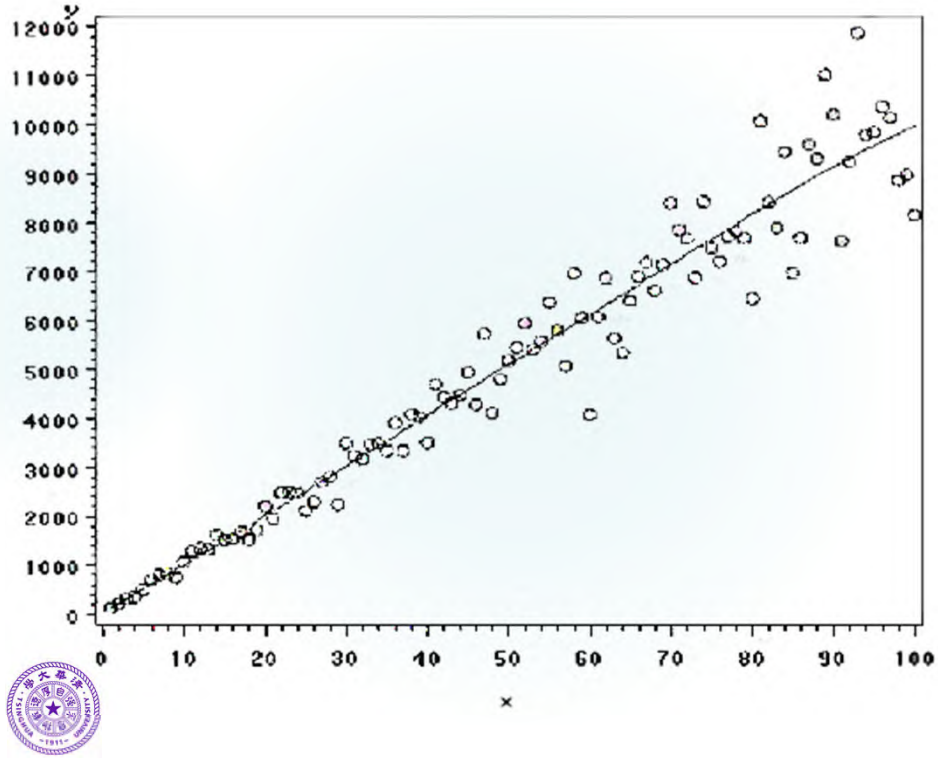
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



R Code: Visual Checks

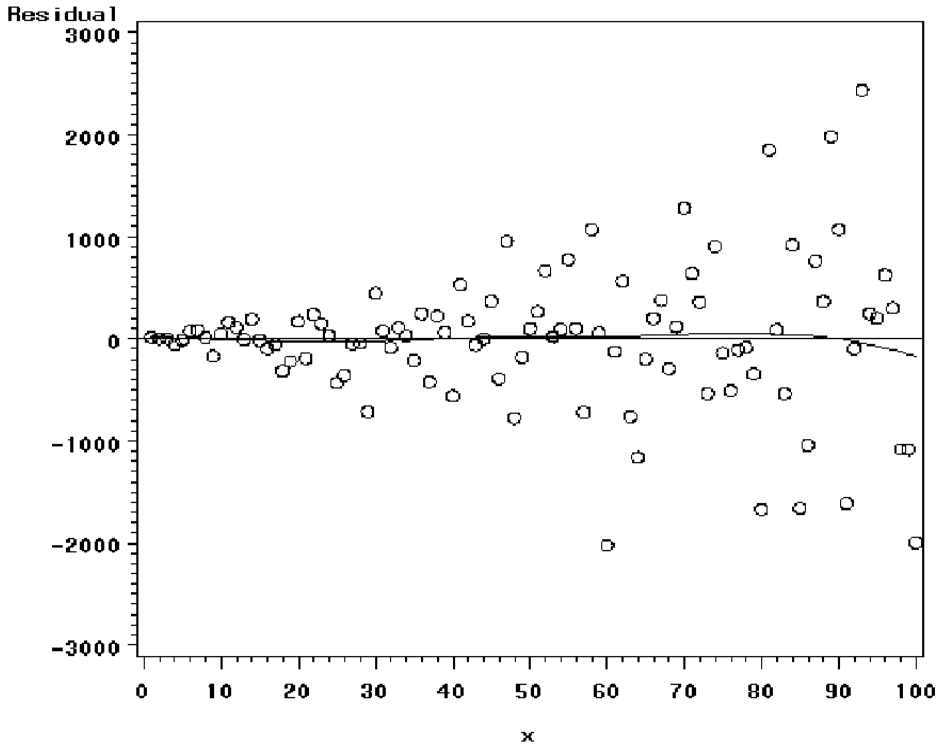
- Scatterplot with smoothed curve

```
> scatter.smooth(x,y)
```



- Residual plot

```
> scatter.smooth(x,fit2$residuals)
> abline(h=0)
```



单选题 1分

设置

51

就 BLUE 而言，你觉得异方差“会”或者“不会”产生什么问题呢?

无偏性保持，方差最小保持

无偏性不成立，方差最小不成立

无偏性保持，方差最小不成立

无偏性不成立，方差最小保持



提交

What's the Concern?

- Since
  - Trend appears linear
  - Estimates close to the truth
- Why is non-constant variance an issue here?

- Answer:
  - Inference including testing, CI, prediction can be off, wrong, or difficult to do
  - Estimation is not the most efficient



主观题 10分

设置

53

既然说无偏性没有受异方差影响，为什么

$$Y = 30 + 100X + 10X\varepsilon$$

的例子中，$b_1 \approx \beta_1 = 100$，而 b_0 和 $\beta_0 = 30$ 却差别非常大?

```
Coefficients:
             Estimate  Std. Error  t value Pr(>|t|)
(Intercept)   71.299    118.852     0.60     0.55
x            100.746      2.043    49.31    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

作答



Testing the Assumption of Equal Variances

- Goal: To test if k samples are from populations with equal variances
- Bartlett's test (likelihood ratio test, χ^2, sensitive to departures from normality)
- Levene's test ($Z_{ij} = |Y_{ij} - \bar{Y}_i|$)
  - Monotonous, sample with greater dispersion will have observations with larger absolute deviations from center, t or ANOVA
- Brown-Forsythe test ($Z_{ij} = |Y_{ij} - \tilde{Y}_i|$, robustness)
  - Levene with trimmed means better for symmetrical but heavy-tailed distributions, B-F (median) better for skewed

$Z_i^T \alpha = \alpha_0 + \alpha_1 Z_{i1} + \cdots + \alpha_p Z_{ip}$

- Breusch-Pagan (test $\sigma_i^2 = h(Z_i^T \alpha)$, $h(\cdot)$ shall possess first and second derivatives
  - $H_0: \alpha_1 = \cdots = \alpha_p = 0$. Under H_0, $\sigma_i^2 = h(Z_i^T \alpha) = h(\alpha_0) = \sigma^2$

Take Z as X, $h = exp, p = 1$

  - Test statistic: $LM = \frac{SSR^*}{2} \Big/ \left(\frac{SSE}{n}\right)^2 \approx \chi^2(p)$, SSR^*: regression e^2 upon Z)



Assessment of Normality

- Look at the distribution of residuals
  - Can look at them together because $E(\varepsilon) = 0$ (unlike Y's)
- Common to use eye-test
  - Scatter in residual plot
  - Box plot
  - Histogram
  - Normal quantile plot

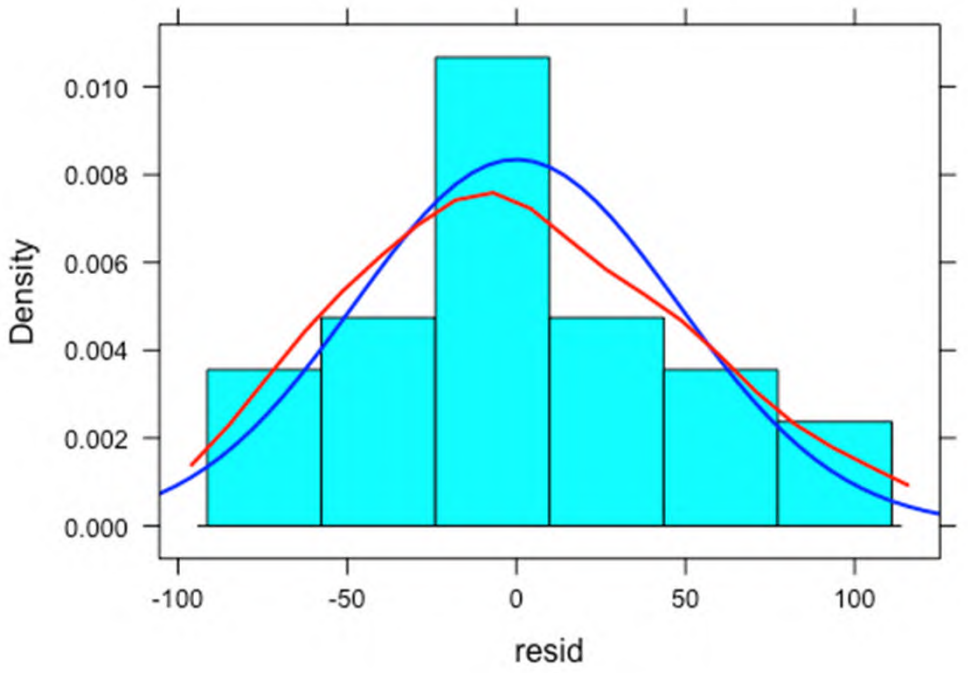


Histogram Example: Toluca


```
> reg = lm(hours~lotsize, data=toluca)
> resid = resid(reg)
```

- Bell-shaped?
- Symmetric?
- Can often fit Normal curve on top of histogram to assess fit

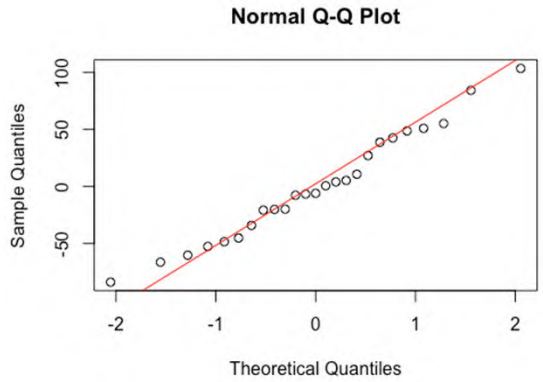
The assumed normal density *vs* The fitted density



Tests for Normality

- H_0: data are an i.i.d. sample from a Normal population
- H_1: data are *not* an i.i.d. sample from a Normal population
- KNNL (p 115) suggests a correlation test that requires a table look-up based on relationship between observations and normal scores

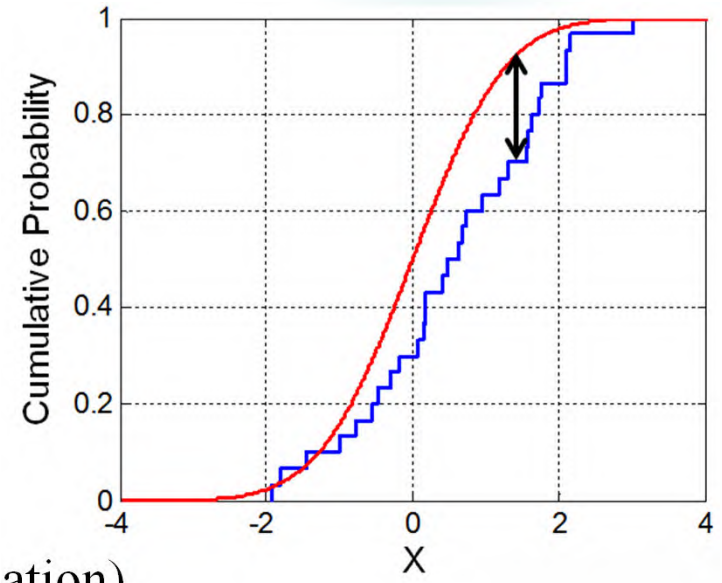
```
qqp <- qqnorm(resid)
cor(qqp$x,qqp$y)
```







Tests for Normality



- R has several choices for the formal significance testing procedure
- Four common procedures
  - Shapiro-Wilk (coefficient of correlation)
  - Kolmogorov-Smirnov
  - Cramér-von Mises
  - Anderson-Darling
- Shapiro-Wilk is most common choice

$$D_n = \sup_x |F_n(x) - F(x)|$$

$$T = n \int_{-\infty}^{\infty} \left(F_n(x) - F(x)\right)^2 dF(x)$$

$$A^2 = n \int_{-\infty}^{\infty} \left(F_n(x) - F(x)\right)^2 w(x)\, dF(x)$$

$$w(x) = \frac{1}{F(x)(1 - F(x))}$$



Shapiro-Wilk Test

- In statistical hypothesis testing, data are tested against the null hypothesis that it is normally distributed

```
> shapiro.test(resid)      # Shapiro-Wilk test
```

```
        Shapiro-Wilk normality test

data:  resid
W = 0.9789, p-value = 0.8626
```

- Has the best power
- Sample-size biased

```
No obvious deviations from normality as
P-values are greater than 0.05
```



Recap: Are the Errors Normal?

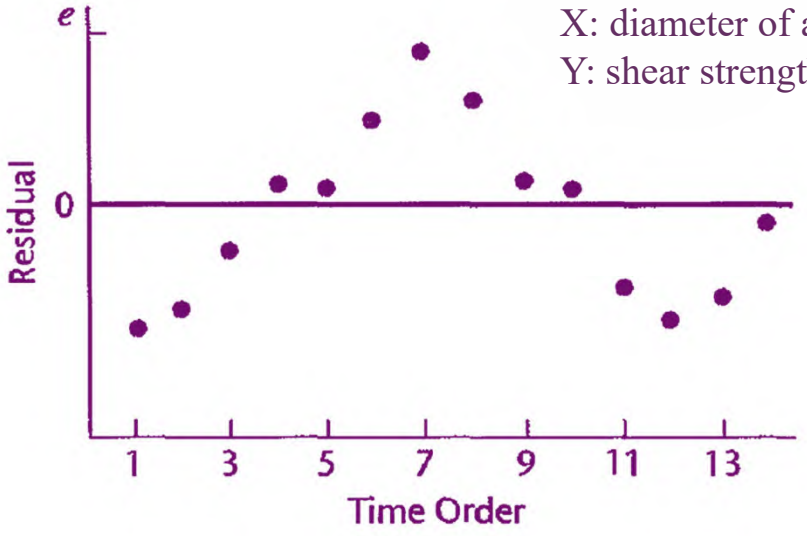
- The **real** question is whether the distribution of the errors is far enough away from Normal to invalidate our confidence intervals, significance tests, etc
- Look at the residuals' distribution
- Use a Normal quantile plot
- Be wary of tests of Normality



Are Errors Dependent?

- Usually we see this in a plot of residuals vs time order (KNNL)
- Pattern usually a linear or quadratic trend and/or cyclical
- Observations not independent





X: diameter of a weld
Y: shear strength of the weld

If you are interested read KNNL pgs108-110



Tests for Dependence

- Durbin-Watson test for serially correlated errors/randomness (KNNL p. 114, p. 487 Section 12.3)

$$\varepsilon_i = \rho\varepsilon_{i-1} + u_i \text{ where } -1 < \rho < 1, u_i \text{ i.i.d} \sim N(0, \sigma^2)$$

$$d = \frac{\sum_2^n (e_i - e_{i-1})^2}{\sum e_i^2} \approx 2(1 - \hat{\rho}) \in (0,4)$$

  - Require the use of tables. Rule of thumb: $d \in (1.5, 2.5)$ is relatively fine; $d < 1$ or $d > 3$ shall be a definite cause for concern

- Ljung-Box test for a time series tests the randomness of "overall" (a number of lags) instead of at each distinct lag

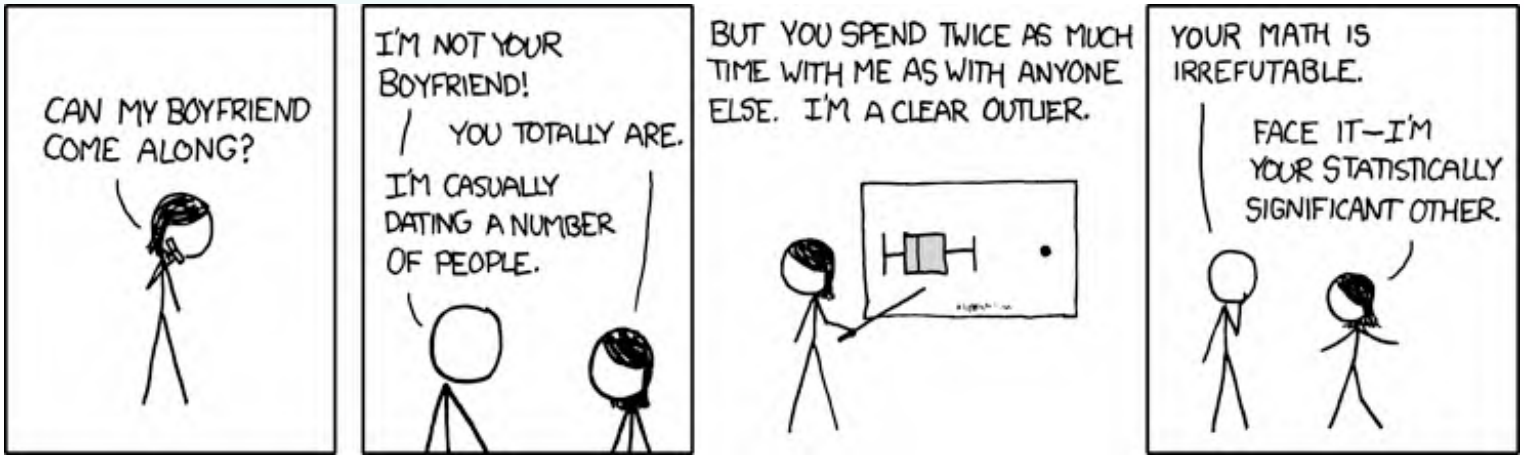
$$Q = n(n+2)\sum_{k=1}^{h} \frac{\hat{\rho}_k^2}{n-k} \overset{H_0}{\approx} \chi^2_{(h)}$$

  - $\hat{\rho}_k^2$: the sample autocorrelation at lag k
  - h: the number of lags being tested



Are There Outliers?

- Model estimates based on means and sums of squares
  - These numerical summaries are not robust to outliers
- Plot Y_i vs X_i
- Plot e_i vs X_i or Y_i
- Plot of e_i vs X_i should emphasize on outliers



R Code: Synthetic #3

- Generates data set with $Y = 30 + 50X + \varepsilon$
- Errors are Normally distributed with $\sigma = 200$

```
>  x = seq(1, 100, by=5)
>  n = length(x)
>  y = 30 + 50*x + 200*rnorm(n)

>  x = c(x, 50)
>  y = c(y, 30+50*50+10000)
# the last pair is the outlier

>  a100b1 = data.frame(x, y)
```

```
# With Outlier
>  fit3 = lm(y~x, data=a100b1)
>  anova(fit3)
>  summary(fit3)

# Without Outlier
>  fit4 = lm(y~x, data=a100b1[-(n+1),])
>  anova(fit4)
>  summary(fit4)
```



With Outlier

$Y = 30 + 50X + \varepsilon$

s = 2237

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
x	1	44475362	44475362	8.8859	0.007679	**
Residuals	19	95098399	5005179			

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	458.98	973.92	0.471	0.64281	
x	51.72	17.35	2.981	0.00768	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Without Outlier

$Y = 30 + 50X + \varepsilon$

s = 185.8

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	43010130	43010130	1245.7	< 2.2e-16 ***
Residuals	18	621474	34526		

Coefficients:

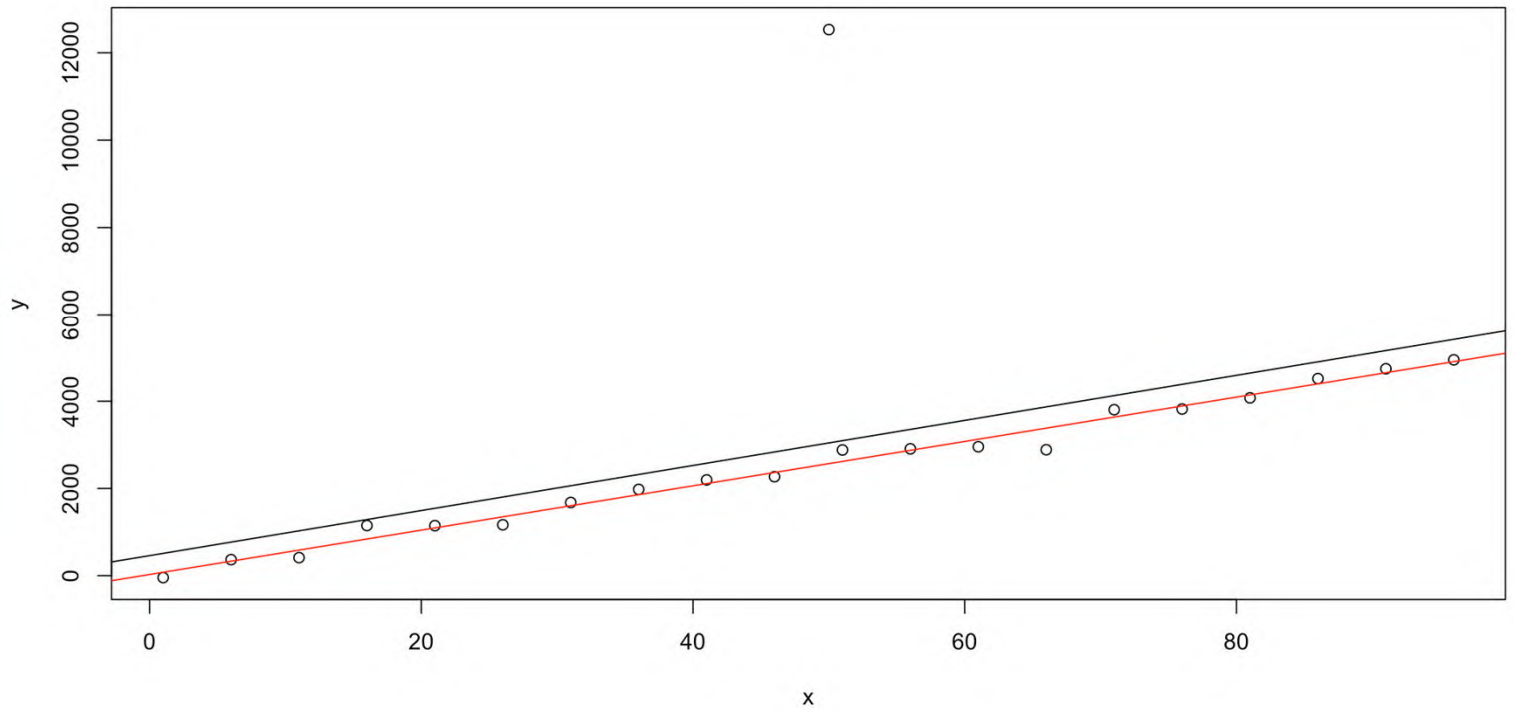
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	26.235	81.311	0.323	0.751
x	50.863	1.441	35.295	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



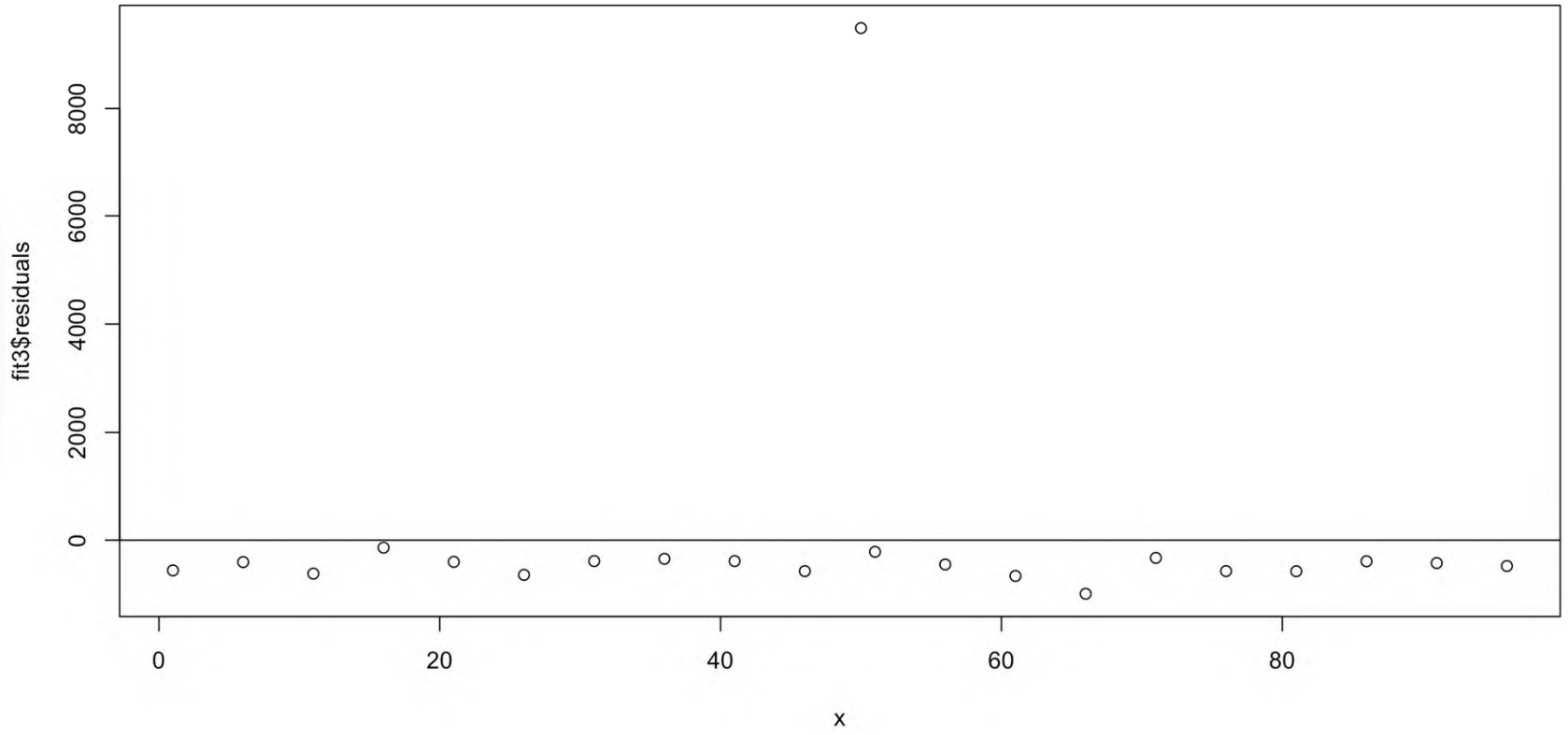
R Code: Visual Checks

```
> plot(y ~ x, data=a100b1)
> abline(fit3)
> abline(fit4, col='red')
```





```
> plot(fit3$residuals ~ x, data=a100b1)
> abline(h=0)
```





Different Kinds of Outliers

- The outlier in the last example *influenced* the intercept but not the slope
- It inflated all of our standard errors
- Here is an example of an outlier that *influences* the slope



R Code: Synthetic #4

- Generates data set with $Y = 30 + 50X + \varepsilon$
- Errors are Normally distributed with $\sigma = 200$

```
>  x = seq(1, 100, by=5)
>  n = length(x)
>  y = 30 + 50*x + 200*rnorm(n)

>  x = c(x, 100)
>  y = c(y, 30+50*100 -10000)
# the last pair is the outlier

>  a100c1 = data.frame(x, y)
```

```
# With Outlier
>  fit5 = lm(y~x, data=a100c1)
>  anova(fit5)
>  summary(fit5)

# Without Outlier
>  fit6 = lm(y~x, data=a100c1[-(n+1),])
>  anova(fit6)
>  summary(fit6)
```



With Outlier

$$Y = 30 + 50X + \varepsilon$$

s = 2111

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	12008430	12008430	2.6947	0.1171
Residuals	19	84668824	4456254		

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	861.83	903.49	0.954	0.352
x	25.04	15.25	1.642	0.117

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Without Outlier

$Y = 30 + 50X + \varepsilon$

s = 185.5

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	43010130	43010130	1245.7	< 2.2e-16 ***
Residuals	18	621474	34526		

Coefficients:

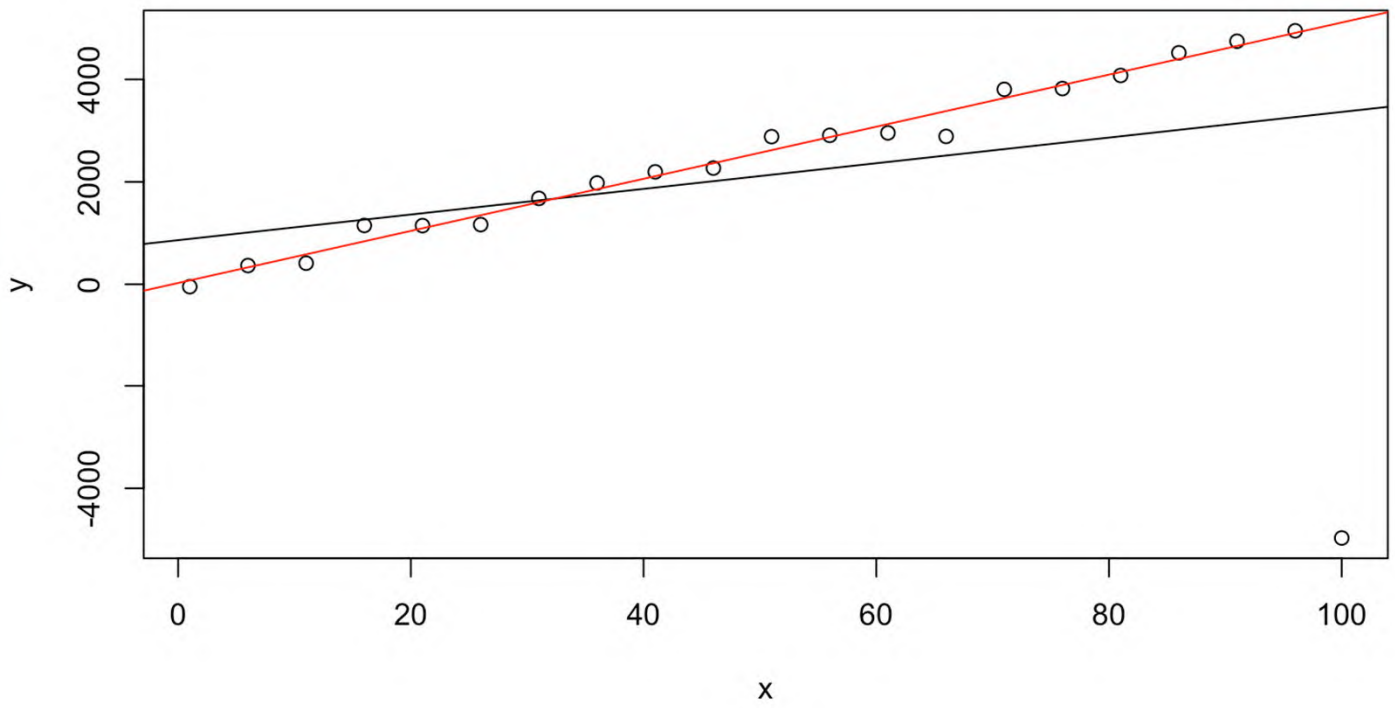
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	26.235	81.311	0.323	0.751
x	50.863	1.441	35.295	<2e-16***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



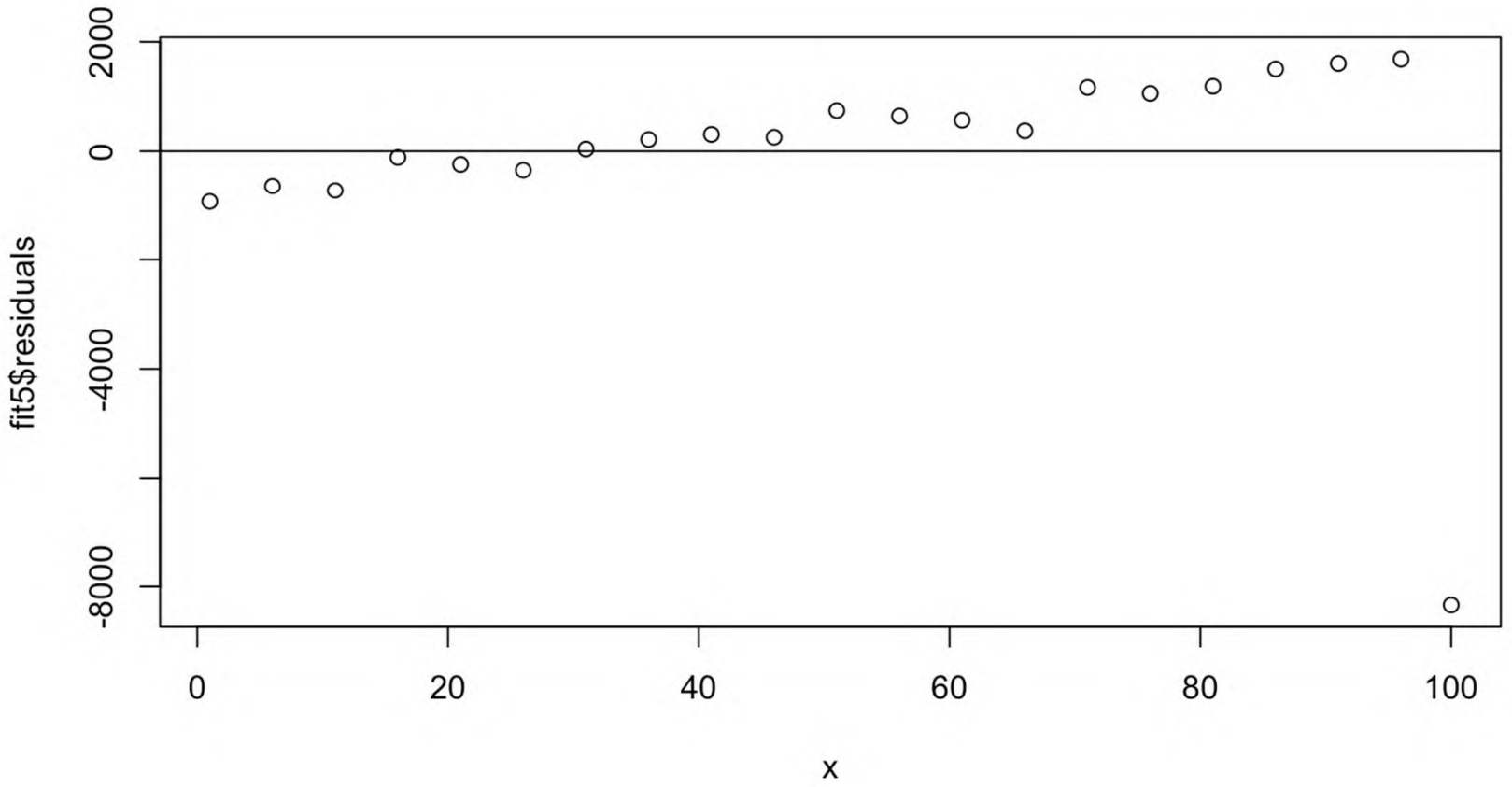
R Code: Visual Checks

```
> plot(y ~ x, data=a100c1)
> abline(fit5)
> abline(fit6 , col='red')
```





```
> plot(fit5$residuals ~ x, data=a100c1)
> abline(h=0)
```



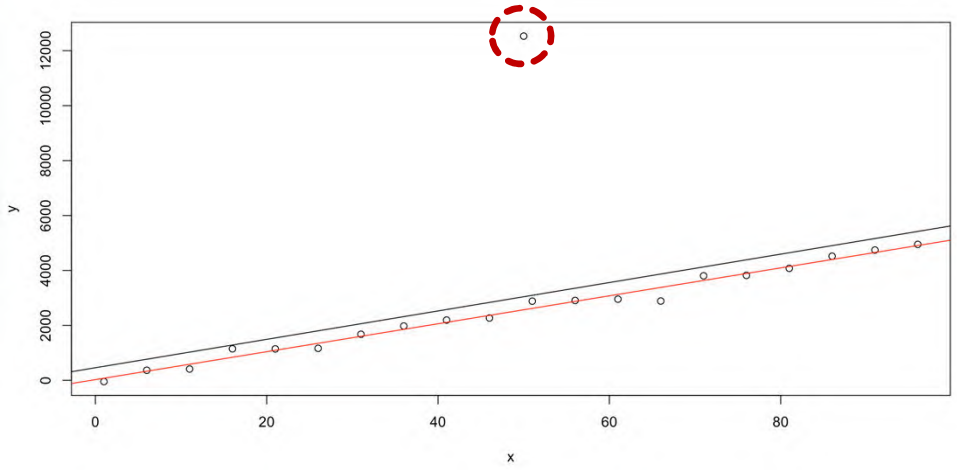
主观题 10分

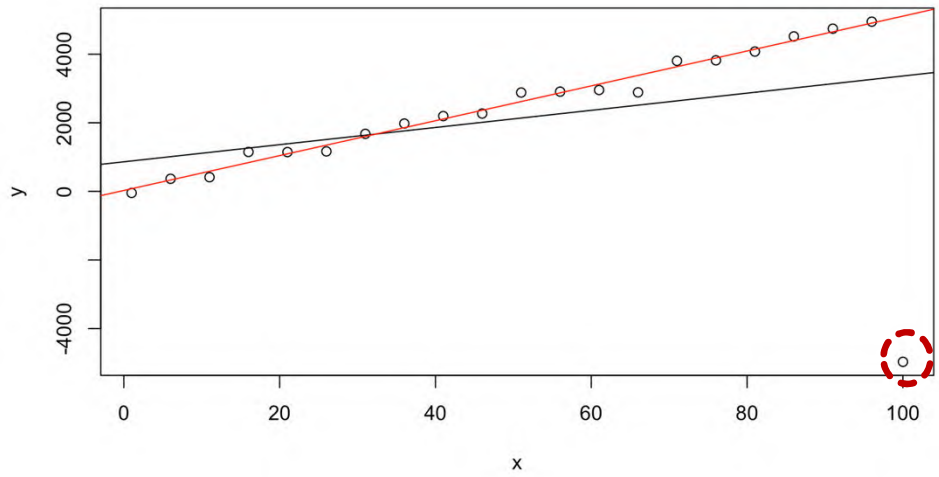
设置

75

你觉得什么情况下离群值会对斜率的估计产生较大影响？

什么情况下影响较小？





作答

第5讲 模型假设及其诊断

- 5.1 模型和假设
  - 5.1.1 模型
  - 5.1.2 假设
- 5.2 模型诊断
  - 5.2.1 关于X
    - 5.2.1.1 分布的四参数
    - 5.2.1.2 常用图
  - 5.2.2 关于残差
    - 5.2.2.1 检验线性关系
    - 5.2.2.2 检验方差齐性
    - 5.2.2.3 检验正态性
    - 5.2.2.4 检验相关性
    - 5.2.2.5 检验离群值



Objectives

- Understand why we need to check the assumptions of our model.
- Know the things that can go wrong with the linear regression model.
- Know why and what we check for X.
- Distinguish between left and right skewness.
- Distinguish between a leptokurtic distribution and a platykurtic distribution.
- Know what box plots, histogram plots, normal quantile plots are and how to use them.
- Know how we can detect various problems with the model using a residuals plot.
- Undertand the ideas of significance tests for equal variances.
- Undertand the ideas of significance tests for normality.
- Undertand the ideas of significance tests for dependence.
- Detect outliers and pay attention to their influence to regression results.



Homework

- Prob-1. KNNL 3.7 (For part *b*, instead of a dot plot required there, please draw a histogram.)
- Prob-2. Please comment on the title of the news.



Background Reading

- We have learnt through 3.1~3.6



下节预告：

- 你知道在 R 中如何对一个简单线性回归模型进行诊断吗?
- 我们如何使用GLT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?
- …
- 有你感兴趣的话题吗? 再接再厉，下次课不见不散!

