MATH253 cheatsheet

MATH253 cheatsheet

Chapter 1 基础分布

Chi-squared distribution

Definition: If $Z_1, \ldots, Z_n$ are independent, identically distributed (i.i.d.) standard normal random variables, that is $Z_i \sim N(0,1)$ for all i, then the distribution of $Y = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ is called the χ^2-distribution with n degrees of freedom.

- Notation: $Y \sim \chi_n^2$

t distribution

Definition: If $Z \sim N(0,1)$ and $Y \sim \chi_n^2$ with Z and Y independent, then the distribution of $T = \frac{Z}{\sqrt{Y/n}}$ is called the t-distribution with n degrees of freedom.

- Notation: $T \sim t_n$

F distribution

Definition: If $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ with X and Y independent, then the distribution of $F = \frac{X/m}{Y/n}$ is called the F-distribution with (m, n) degrees of freedom.

- Notation: $F \sim Fm, n$

Chapter 2

Unbiased estimators

Definition: Suppose $\hat{\theta}$ is an estimator of a parameter θ. Then $\hat{\theta}$ is called unbiased if $E[\hat{\theta}] = \theta$. Otherwise, we say $\hat{\theta}$ is biased.
The value of $E[\hat{\theta}] - \theta$ is called the bias of $\hat{\theta}$.

Chapter 3 置信区间CI

CI for μ when σ^2 is known

When σ^2 is known, we have that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and so then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

For the 95% confidence interval what we want is an interval $[\bar{X} - \epsilon, \bar{X} + \epsilon]$ containing μ where the probability of producing this interval is 95% and where ϵ is the margin of error.

$$\begin{aligned} \Pr(\mu - \epsilon < \bar{X} < \mu + \epsilon) &= 0.95 \\ \Rightarrow \Pr\left(\frac{-\epsilon}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{\epsilon}{\sigma/\sqrt{n}}\right) &= 0.95 \\ \Rightarrow \Pr(-k < Z < k) &= 0.95 \end{aligned}$$

where $Z \sim N(0, 1)$ and $k = \frac{\epsilon}{\sigma/\sqrt{n}}$
The 95% confidence interval (CI) is

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

以 $\bar{X}$ 为中心的估计区间

CI对应的 (1-α)%的critical value可以通过查询tables得到，也可以通过R得到

CI for μ when σ^2 is unknown

We know that the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ is unbiased estimator of σ^2.
It can be shown that

- $\bar{X} \sim N\left(\mu, \sigma^2/n\right)$ - this was shown before,
- $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$ - this will be shown later,
- $\bar{X}$ and s^2 are independent random variables

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

$100(1-\alpha)\%$ confidence interval is now

$$\left[\bar{X} - t_{n-1}\left(\frac{\alpha}{2}\right)\frac{s}{\sqrt{n}}, \bar{X} + t_{n-1}\left(\frac{\alpha}{2}\right)\frac{s}{\sqrt{n}}\right]$$

where $t_{n-1}\left(\frac{\alpha}{2}\right)$ represents the point with $\Pr\left(T > t_{n-1}\left(\frac{\alpha}{2}\right)\right) = \frac{\alpha}{2}$ for $T \sim t_{n-1}$

Chapter 4 Hypothesis Test

Type I and Tpye II error

In reaching a decision, we may make two types of errors:

- Type I error (false positive): We reject H_0 when H_0 is in fact true.
  The probability of Type I error is denoted by α : $\quad \alpha = \Pr(\text{ Type I error })$
- Type II error (false negative): We fail to reject H_0 when H_0 is in fact false.
  The probability of Type II error is denoted by β : $\quad \beta = \Pr$ (Type II error)

The probability of Type I error, α, is the significance level.
The probability of correctly rejecting H_0 is called power and is calculated as
Power $= 1 - \beta$ = Pr(reject $H_0 \mid H_1$).

此处H_1是测试增加/减少几个单位的 μ，题目会给定数值。需要先算出reject H_0时 $\bar{X}$ 满足的条件再代入。

power 与 β 运算的区别是>和<的转换，所以结果满足 power $= 1 - \beta$。

power的值与confidence level密不可分，需要依据 α 得到。

Chapter 5 检测μ是否正确

The underlying **assumption** for these tests is that the population is **normally distributed**.

Z test when σ^2 is known

Use: To test whether the population mean, μ, has a specified value μ_0, when the population variance σ^2 **is known**.
To carry out a Z-test for a population mean we use the following steps:

1. State hypotheses, the null hypothesis $H_0: \mu = \mu_0$ and the alternative hypothesis $H_1: \mu < \mu_0$ for the lower tail, $H_1: \mu > \mu_0$ for the upper tail, or $H_1: \mu \neq \mu_0$ for the two-tailed alternative.
2. Calculate the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

3. Compare the value of the test statistic to the critical value in the standard normal distribution tables for the relevant significance level, remembering to half the significance level for a two-tailed test. Alternatively, compute the p-value.
4. Decide whether to reject or not to reject the null hypothesis at α.
   Remember that $Z \sim N(0,1)$ and so we use critical values $z(\alpha)$ or $z(\alpha/2)$ as following:

- For the alternative $H_1: \mu < \mu_0$, we reject H_0 at α if $Z < -z(\alpha)$.
- For the alternative $H_1: \mu > \mu_0$, we reject H_0 at α if $Z > z(\alpha)$.
- For the alternative $H_1: \mu \neq \mu_0$, we reject H_0 at α if $Z < -z\left(\frac{\alpha}{2}\right)$ or $Z > z\left(\frac{\alpha}{2}\right)$, in other words, if $|Z| > z\left(\frac{\alpha}{2}\right)$.
  If we calculate the exact p-value, then, of course, we use general rules: That is if $p \leq \alpha$, then we reject H_0 at α, otherwise we fail to reject H_0 at α.

5. Interpret your conclusion practically in the context of the question.

```
library(BSDA) #z-test不在R里面，在BSDA里面需要单独调用
z.test(x ,y=NULL,alternative=c('t'),mu=5.1,sigma.x=0.22,sigma.y=NULL,conf.level=0.95)
```

T test when σ^2 is unknown

The t-test is robust to non-normality. ($n \geq 30$近似正态分布).
Use: To test whether the population mean, μ, has a specified value μ_0, when the population variance σ^2 **is unknown**.
To carry out a t-test for a population mean we use the following steps:

1. State hypotheses 同上
2. Calculate the test statistic

$$T = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

where s is the sample standard deviation, that is $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$. （方差未知情况的代替）

3. 与表格比较或计算p值
4. Decide whether to reject or not to reject the null hypothesis at α.
   As shown before $T \sim t_{n-1}$ and so we use critical values $t_{n-1}(\alpha)$ or $t_{n-1}(\alpha/2)$ as following:

准则完全类似 z test

5. Interpret your conclusion practically in the context of the question.

```
t.test(x ,y=NULL, alternative=c('t'),mu=5.1,paired=FALSE,var.equal=FALSE,conf.level=0.95)

#不是two sample的情况，所以y的位置是NULL
```

Chapter 6

Confidence Intervals and Hypothesis Testing

General rule:
The $1-\alpha$ confidence interval consists of all values μ_0 such that $H_0 : \mu = \mu_0$ is not rejected at α level in a two-sided test. And so, when the value μ_0 is outside the $1-\alpha$ confidence interval, then $H_0 : \mu = \mu_0$ is rejected at α level in a two-sided test.

Chapter 7 测试两组样本 μ 是否相同

Two sample tests for means are used when we want to compare two samples to assess whether their population means differ.

Paired t-test when σ^2 is know

Assumption: difference 是 normally distributed

$T = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{n-1}$

To carry out a paired t-test we use the following steps:

1. State hypotheses, the null $H_0 : \mu_d = 0$, where μ_d is the mean of the population from which the differences are sampled, and the alternative $H_1 : \mu_d < 0, H_1 : \mu_d > 0$, or $H_1 : \mu_d \neq 0$
2. Calculate the differences between each pair of observations, along with the mean and standard deviation of these differences, $\bar{d}$ and s_d
3. Calculate the test statistic

$$T = \frac{\bar{d}}{s_d/\sqrt{n}}$$

4. Compare the value of the test statistic to the critical value in the t-distribution tables for the relevant significance level, remembering to half the significance level for a two-tailed test. Alternatively, compute the p-value.
5. Decide whether to reject or not to reject the null hypothesis at α.
   As shown before $T \sim t_{n-1}$ and so we use critical values $t_{n-1}(\alpha)$ or $t_{n-1}(\alpha/2)$ as following:

- For the alternative $H_1 : \mu_d < 0$, we reject H_0 at α if $T < -t_{n-1}(\alpha)$.
- For the alternative $H_1 : \mu_d > 0$, we reject H_0 at α if $T > t_{n-1}(\alpha)$.

- For the alternative $H_1 : \mu_d \neq 0$, we reject H_0 at α if $|T| > t_{n-1}\left(\frac{\alpha}{2}\right)$.
  If we calculate the exact p-value, then, of course, we use general rules: That is if $p \leq \alpha$, then we reject H_0 at α, otherwise we fail to reject H_0 at α.

6. Interpret your conclusion practically in the context of the question.

Confidence interval for difference of means
Adapting the formula for CI for the population mean (with unknown σ^2), we get the formula for the $100(1-\alpha)\%$ confidence interval for μ_d :

$$\left[\bar{d}-t_{n-1}\left(\frac{\alpha}{2}\right)\frac{s_d}{\sqrt{n}}, \bar{d}+t_{n-1}\left(\frac{\alpha}{2}\right)\frac{s_d}{\sqrt{n}}\right]$$

> t.test(A, y=B, alternative=c('g'), paired=TRUE, var.equal=FALSE,conf.level=0.95)
>
> #paried t test 需要设定 paired=TRUE, var.equal=FALSE
>
> #R 自动设定difference为x-y 需要注意题目要求左右关系的小于和大于号，要和R的习惯进行对照转换
>
> #例如看if μ_2 increases，需要设定 alternative=c('l')

Independent two sample tests

Known common standard deviation – independent two-sample Z-test

Assumption: 1.The samples are independent,

2.each sample comes from a normal distribution

3.with the **same variance**

注意：没有假设difference也满足normal distribution

A $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$ is

$$\left(\bar{X}_1-\bar{X}_2\right) \pm z\left(\frac{\alpha}{2}\right)\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

that is, $\left(\bar{X}_1-\bar{X}_2\right) \pm z\left(\frac{\alpha}{2}\right) \times$ s.e. $\left(\bar{X}_1-\bar{X}_2\right)$
To test $H_0 : \mu_1=\mu_2$ versus a one-sided or two-sided alternative, we use the two-sample Z-statistic

$$Z=\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Note that under H_0, this statistic has standard normal distribution $N(0,1)$.

Unknown common standard deviation – independent two-sample t-test

假设同上

A $100(1-\alpha)\%$ CI for $\mu_1-\mu_2$ is

$$\left(\bar{X}_1-\bar{X}_2\right) \pm t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

that is, $\left(\bar{X}_1-\bar{X}_2\right) \pm t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) \times$ e.s.e. $\left(\bar{X}_1-\bar{X}_2\right)$
To test $H_0: \mu_1=\mu_2$ versus an alternative, we use the two-sample t-statistic

$$T=\frac{\bar{X}_1-\bar{X}_2}{sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Under H_0, this statistic has t-distribution with (n_1+n_2-2) degrees of freedom.

$$\begin{aligned} s_p^2 &= \frac{\sum_{i=1}^{n_1}\left(X_{1i}-\bar{X}_1\right)^2+\sum_{i=1}^{n}\left(X_{2i}-\bar{X}_2\right)^2}{n_1+n_2-2} \\ &= \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \end{aligned}$$

```
t.test(A, B, alternative=c('t'), paired=FALSE, var.equal=TRUE,conf.level=0.95)
```

Chapter 8 推断σ^2，比较两个σ^2

Sampling distribution of sample variance 推断σ^2

Suppose $X_1, \ldots, Xn$ is a random sample from $N\left(\mu, \sigma^2\right)$, that is $X_1, \ldots, Xn$ are independent and $X_i \sim N\left(\mu, \sigma^2\right)$ for all i. We want to show that $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$

So, the $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left[\frac{(n-1)s^2}{\chi_{n-1}^2(\alpha/2)}, \frac{(n-1)s^2}{\chi_{n-1}^2(1-\alpha/2)}\right]$$

The $100(1-\alpha)\%$ confidence interval for σ is

$$\left[\sqrt{\frac{(n-1)s^2}{\chi_{n-1}^2(\alpha/2)}}, \sqrt{\frac{(n-1)s^2}{\chi_{n-1}^2(1-\alpha/2)}}\right]$$

此分布不对称，所以两头要分开计算

```
#找90%CI for the population variance sigma^2 for A


```

```
library(TeachingDemos)
sigma.test(A, sigma=1, alternative=c('t'), conf.level=0.95)
sigma.test(A, alternative=c('t'), conf.level=0.90)
```

sigma=多少不重要，我们只是想找CI for sigma，所以参数默认sigma=1，可省略

Comparison of two population variances 比较两个σ^2

Confidence interval for the ratio of variances

For independent two samples t-test, we assumed both populations had the same **(unknown) variance**. In this section we will discuss how we can check this.

We use the notation introduced previously:

- Sample 1: $X_{11}, \ldots, X_{1n_1}$
- Sample 2: $X_{21}, \ldots, X_{2n_2}$
- With $X_{1i} \sim N\left(\mu_1, \sigma_1^2\right)$ and $X_{2i} \sim N\left(\mu_2, \sigma_2^2\right)$ independently
  We want to test whether two population variances are the same, that is, we test

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

Equivalently, we can write

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad \text{versus} \quad H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

For each sample, we know (from one-sample theory in previous section)

$$\frac{(n_1-1)s_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1} \quad \frac{(n_2-1)s_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$$

independently.
Using the definition of F-distribution, we have $\frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$

So, the $100(1-\alpha)\%$ CI for σ_1^2/σ_2^2 is therefore

$$\left[\frac{s_1^2}{s_2^2}\frac{1}{F_{n_1-1,n_2-1}(\alpha/2)}, \frac{s_1^2}{s_2^2}F_{n_2-1,n_1-1^{(\alpha/2)}}\right]$$

如果1在区间内说明很大可能性$\sigma_1^2 = \sigma_2^2$

R 中计算方差用var()，分母自动为n-1，填补空值加上命令na.rm=TRUE

```
a<-var(ex$Instant, na.rm = TRUE)
b<-var(ex$`Drip-brew`)
c<-qf(0.05, 6,8,lower.tail = FALSE)
d<-qf(0.05, 8,6,lower.tail = FALSE)
```

```
a/b/c
a/b*d
```

F test for comparison of two population variances (鲁棒性差)

To carry out a F-test for comparison of two population variances we use the following steps:

1. State hypotheses, the null $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$, and the alternative $H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1, H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$, or $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$.
2. For each sample calculate the sample variances s_1^2 and s_2^2.
3. Calculate the test statistic

$$F = \frac{s_1^2}{s_2^2}$$

4. Compare the value of the test statistic to the critical value in the $F_n - 1, n_2 - 1$ distribution tables for the relevant significance level, remembering to half the significance level for a two-tailed test. Alternatively, compute the p-value.
5. Decide whether to reject or not to reject the null hypothesis at α.
   As shown before $F \sim F_{n_1} - 1, n_2 - 1$ and so we use critical values from the F_{n_1-1,n_2-1} distribution as following:

- For the alternative $H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1$, we reject H_0 at α if $F < F_{n_1-1,n_2-1}(1-\alpha)$
- For the alternative $H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$, we reject H_0 at α if $F > F_{n_1-1,n_2-1}(\alpha)$
- For the alternative $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$, we reject H_0 at α if $F < F^{(1-\alpha/2)}_{n_1-1,n_2-1}$ or $F > F_{n_1-1,n_2-1}(\alpha/2)$
  If we calculate the exact p-value, then, of course, we use general rules: That is if $p \leq \alpha$, then we reject H_0 at α, otherwise we fail to reject H_0 at α.

6. Interpret your conclusion practically in the context of the question.

```
var.test(A,B, ratio=1, alternative=c('t'),conf.level=0.95)
```

Chapter 9 检测是否正态分布

用专业方法之前可以借助histgram人肉做一个粗糙的判断

```
difference <- leaf$Treated - leaf$Untreated
hist(difference, main = paste("Histogram of differences"), breaks = 10,
  xlab = "Differences", col="purple", xlim = range(-0.5,1.5))
```

Normal probability plot

```
qqnorm(x,col='blue')
qqline(x,col='red')
```

点分布越与直线重合，normal的可能性更大

Anderson-Darling test for normality

```
library(nortest)
ad.test(x)
```

之前诸多test的assumption都要求sample或difference满足normal distribution，所以测试显得尤为重要。

p值越大，正态分布可能性大(H_0不能被拒绝)

“The p-value for the AndersonDarling test is given as p = 0.7573, meaning that we cannot reject the null hypothesis (data normal) even at the 75% level. That is, there is no evidence that data do not come from a normal distribution. The assumption underlying the paired t-test (normality of the differences) appears to be justified.”

Chapter 10 ANOVA 又名Analysis of Variance

a statistical methodology for comparing means of several populations

ANOVA model

The ANOVA model is derived in a similar way. (data=fit+residual.)

- We have k independent samples, from different normally distributed populations.
- Observations or response values are denoted in the following way:

  Group 1: $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$
  Group 2: $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$
  Group k : $Y_{k1}, \ldots \ldots, Y_{kn_k}$
- We have n_i observations from group i for $i = 1, 2, \ldots, k$.
- Note that Y_{ij} is response j from group i.
- We assume that

$$Y_{ij} \sim N\left(\mu_i, \sigma^2\right)$$

for parameters $\mu_1, \ldots, \mu_k$ and σ^2

- Note that we assume a common variance σ^2 for all groups - this is called the **homoscedastic assumption**. 组间同方差假设
- The population mean for group i is denoted by μ_i. Our goal is to test whether these population means differ.

Estimating parameters - Least Squares estimation

To sum up, the **Least Squares parameter estimates for the ANOVA model** are

$$\widehat{\mu_i} = \overline{y_i\cdot} \quad \text{where} \quad \overline{y_i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

group内的均值

For next discussion, we also define **overall sample mean**

$$\overline{y\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$$

所有样本的均值

where $N = n_1 + \cdots + n_k$ is the total sample size.

ANOVA 参数、公式准备

- Now we define the quantities which measure variability and average variability between group sample means while taking into account the group sizes:
  - Between Groups Sum of Squares

    $$SSG = \sum_{i=1}^{k} n_i(\overline{y_i\cdot} - \overline{y\cdot})^2$$

  - Between Groups Mean Square

    $$MSG = \frac{SSG}{k-1}$$

- estimators for σ^2

  We generalise this for k independent groups to use the estimate of σ^2 given by

  $$\widehat{\sigma^2} = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{N-k}$$

  where $N = n_1 + n_2 + \cdots + n_k$

  $\widehat{\sigma^2}$ is an unbiased estimator of σ^2.

- Next we define the quantities which measure the variation and average variation of data within the groups:
  - Error Sum of Squares 单个数据与均值的计算

    $$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2$$

  - Error Mean Square

    $$MSE = \frac{SSE}{N-k}$$

实际上 MSE is actually equal to $\widehat{\sigma^2}$

These are also known as the 'Within Groups' SS and MS.

- Similarly, we define the Total Sum of Squares as

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y\cdot})^2 = (N-1)s_T^2$$

where $s_T^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y..})^2$.
It can be shown that (not required to show this in this module)

$$SST = SSE + SSG$$

- distributions of MSE and MSG.
  - $\frac{SSE}{\sigma^2} \sim \chi^2_{N-k}$
  - if H_0 is true, then $\frac{SST}{\sigma^2} \sim \chi^2_{N-1}$ and $\frac{SSG}{\sigma^2} \sim \chi^2_{k-1}$
- Distribution of MSG/MSE
  So far we know:

$$\frac{SSG}{\sigma^2} \sim \chi^2_{k-1} \text{ and } \frac{SSE}{\sigma^2} \sim \chi^2_{N-k} \quad \text{independently}$$

  Using definition of the F-distribution,

$$\begin{aligned} &\frac{\frac{SSG}{\sigma^2}/(k-1)}{\frac{SSE}{\sigma^2}/(N-k)} \sim F_{k-1,N-k} \\ \Rightarrow &\frac{MSG}{MSE} \sim F_{k-1,N-k} \end{aligned}$$

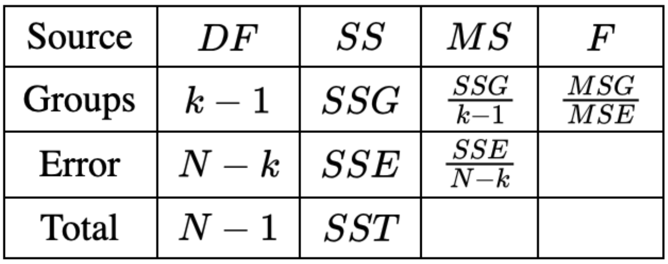
  Conclusion: If the ratio MSG/MSE is big compared to $F_{k-1,N-k}$ critical values, that is evidence of real groups.
  So, MSG/MSE is the **test statistic** for ANOVA test.

ANOVA的关键是比较组间population mean是否一致。所以H_0就是population mean for different groups相等

If H_0 is not true, MSE still estimates σ^2, but the differences between groups make MSG bigger, so we expect $MSG/MSE > 1$. So, if MSG is (significantly) large compared to MSE, this would indicate evidence of differences between groups.
To determine this, we have to find the distribution of MSG/MSE which is the final step in deriving ANOVA test.

ANOVA table 非常有用！

Since there are many calculations before we arrive to the test statistic MSG/MSE, it is convenient to construct the ANOVA table which helps us to organise all necessary calculations.
A general layout of the ANOVA table:



Source	DF	SS	MS	F
Groups	$k-1$	SSG	$\frac{SSG}{k-1}$	$\frac{MSG}{MSE}$
Error	$N-k$	SSE	$\frac{SSE}{N-k}$	
Total	$N-1$	SST		

where $SST = SSE + SSG, DF$ denotes degrees of freedom, SS denotes sum of squares, MS denotes mean squares, k is the number of groups, $N = n_1 + \cdots + n_k$ is the total sample size, SSG, SSE, SST are calculated using formulas introduced earlier.

ANOVA hypothesis test

Use: To test whether several population means are all equal or whether some population means are different.

ANOVA assumptions

1. Samples are independent.
2. Responses are normally distributed, $Y_{ij} \sim N\left(\mu_i, \sigma^2\right)$.
3. Each group has same variance.

（The groups are independent, responses of each group come from normally distributed populations and all groups have the common variance.） 假设非常重要，有一些方法可以检验假设是否成立，详见mobius。

To carry out the ANOVA test we use the following steps:

1. State hypotheses, the null $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ versus the alternative $H_1 : \mu_i$ not all equal.
2. Construct the ANOVA table, including the test statistic $F = MSG/MSE$.
3. Compare the value of the test statistic to the critical value in the $F_{k-1,N-k}$ distribution tables for the relevant significance level. Note that the ANOVA test is a one-sided test. Alternatively, compute the p-value.
4. Decide whether to reject or not to reject the null hypothesis at α.
   As shown before $F \sim F_{k-1,N-k}$ and so we use critical values $F_{k-1,N-k}(\alpha)$ as following: we reject H_0 at α if $F > F_{k-1,N-k}(\alpha)$, otherwise we cannot reject H_0 at α
   If we calculate the exact p-value, then, of course, we use general rules: That is if $p \leq \alpha$, then we reject H_0 at α, otherwise we fail to reject H_0 at α
5. Interpret your conclusion practically in the context of the question.

注意事项

- If you are not given α, find the lowest level at which we can reject H_0 and the highest level at which we cannot reject H_0, following the rules in step 4 for using critical values.
- The ANOVA test is a **one-sided** test because in this test we actually compare two estimators of variance - MSG and MSE. We actually test whether MSG is significantly bigger than MSE, and that is why the ANOVA test is one-sided. 目标是看test statistics即 $F = MSG/MSE$ 是否大于1.

```
qf(0.01, 3, 18, lower.tail = FALSE) #(m,n)为自由度
qf(0.05, 3, 18, lower.tail = FALSE)
qf(0.1, 3, 18, lower.tail = FALSE)
```

这里使用F分布，是one-sided，故百分比不用乘2。

R实现 ANOVA F-test

```
A <- salt$A
B <- salt$B
C <- salt$C
D <- salt$D
E <- salt$E

combined <- data.frame(cbind(A, B, C, D, E)) #确保了组的关系
stacked <- stack(combined) #变成两列的data，一列
anovaresults <- aov(values ~ ind, data = stacked) #～左右两边不要搞错
summary(anovaresults)
```

#normality test for residuals

```
res <- residuals(anovaresults)
```

ANOVA Assumptions and How to Check Them

Assumption of equality of variances

There is an **informal test** (rule of thumb) to check this assumption: provided the largest **standard deviation** is less than twice the smallest **standard deviation**, then methods based upon the assumption of equal variances/standard deviations will be OK to use.

There are also more **formal tests** to check the assumption of equal variances: Bartlett’s test or Levene’s test. These can be carried out by statistical software.

- **Barlett’s test** assumes that the data are normally distributed.
- **Levene’s test** does not require normality.

```
#tests for equal variances
bartlett.test(values ~ ind, data = stacked)
library(car)
leveneTest(values ~ ind, data = stacked)

#boxplot
means <- c(mean(A), mean(B), mean(C), mean(D), mean(E))
boxplot(A, B, C, D, E, col = "lightblue", names=c("A", "B", "C", "D", "E"))
points(means)
#box长度都差不多说明equal variance
```

Assumption of normality

```
#normality test - Anderson-Darling test
library(nortest)
ad.test(res)

#normal probability plot
qqnorm(res, col="blue")
qqline(res, col="red")
```

Robustness of the ANOVA F-test

If assumptions of normality or equal variances do not hold, the ANOVA F-test still gives approximately the correct answer.

- The ANOVA F-test is very robust against non-normality.
  - Large sample size helps - Central Limit Theorem then implies sample means are approximately normal.
  - Most robust for a balanced design (equal group sizes).
- The ANOVA F-test is very robust against heterogeneity of variances (variances not all equal).
  - Most robust for a balanced design.

Chapter 11 Post Hoc Tests

ANOVA test只告诉我们组均值是否相同，或是否有一些均值不同。然而，如果平均值之间存在显著差异，它并不能确切地告诉我们这些差异出现在哪里。

如果我们想知道组中哪个平均值不同，我们必须进行进一步的测试。

我们可以看个别的方法和boxplot来做非正式的评估。然而，这并不是一个正式的组间差异的统计测试!

组间差异的正式统计检验(在方差分析检验显示平均值之间的显著差异后进行)称为Post Hoc Tests（事后检验）。

Fisher’s Least Significant Differences (LSD)

Fisher's least significant difference (LSD) procedure is a two-step testing procedure for pairwise comparisons of several treatment groups. In the first step of the procedure, a global test is performed for the null hypothesis that the expected means of all treatment groups under study are equal. If this global null hypothesis can be rejected at the pre-specified level of significance, then in the second step of the procedure, one is permitted in principle to perform all pairwise comparisons at the same level of significance.

（通过ANOVA test判定H_0被否定，说明some population means for groups不同，要找出这些不同，两两组间实施independent two-sample t-test。）

MSE is based on all available observations, whereas sp would use only the observations from two groups at a time. We know that MSE follows χ^2-distribution on $N-k$ degrees of freedom.
So, we use the t-statistic

$$T = \frac{\overline{y_i\cdot} - \overline{y_{j\cdot}}}{\sqrt{MSE}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

which follows t-distribution with $N - k$ degrees of freedom.

T可以看绝对值

```
library(PMCMRplus)
summary(lsdTest(anovaresults))
```

Multiple Test

One should be careful when performing multiple tests as performing many tests increases overall significance level.
If we perform K tests on independent samples, each test at level α, then the probability of failing to reject H_0 for all tests is

$$(1-\alpha)^K$$

And the probability of rejecting H_0 for at least one of these tests is (i.e. making Type I error) is

$$\alpha_K = 1 - (1-\alpha)^K$$

a_K随着K增大而增大

Note that if there are k groups in ANOVA, then the number of pairs of means to compare is

$$K = \binom{k}{2} = \frac{k!}{2!(k-2)!} = k(k-1)/2$$

To ensure that the overall test level is at the required significance level, the significance level of individual tests is usually adjusted. 个体因总体目标而校准设定

There are many different methods of “correcting” the significance level.

The most common ones are

- Bonferroni correction,
- Tukey’s HSD (Honest Significance Difference) test,
- Scheffe’s test.

我们在此只学习前两个。

Bonferroni correction

The significance level for each individual test is set as

$$\alpha^* = \frac{\alpha}{K}$$

举例：

degree of freedom is 40；10 groups; 目标：是否different at significant level 1%

test statistic与$t_{40}(0.0005)$进行比较

0.0005从何而来?

$\alpha^* = \frac{\alpha}{K} = \frac{0.01}{2*10} = 0.0005$

10是组数(Bonferroni correction)，2是two-sided test

Tukey’s HSD （优先选择的方法）

Tukey's Honestly Significant Difference (HSD) is the **default** method used in pair-wise comparisons in **R**. This method deals with the problem of inflating the overall significance level effectively within the design of the method.
R outputs the p-values for all pair-wise comparisons which are already adjusted.

这个方法不需要掌握原理，但是要会运用（R）

```
TukeyHSD(anovaresults)
library(PMCMRplus)
summary(lsdTest(anovaresults))
```

Chapter 12 Simple Linear Regression

Simple Linear Regression Model and Estimating Parameters

If the points lie reasonably close to a straight line, we can fit linear regression model.
We know that a straight line relationship is $y = \beta_0 + \beta_1 x$, where β_1 is the slope and β_0 the intercept.
But clearly the points aren't exactly on a straight line, so one needs to allow for random variation.

- Since Y is continuous, the simplest thing is to assume that for each value of x, the individual responses of Y are normally distributed with mean $\beta_0 + \beta_1 x$. In other words, we assume that the means of the responses Y all lie on a line $\mu_Y = E(Y) = \beta_0 + \beta_1 x$. This line is called the **population regression line**.
- For each observation of an explanatory variable x_i, we can write the mean of response Y_i as $E\left(Y_i\right) = \beta_0 + \beta_1 x_i$. Of course, not every observed value of the response will equal the mean $E\left(Y_i\right)$. There will be some variation, error (denoted by ϵ_i). So, we can write $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ which is the **simple linear regression model**.

Simple Linear Regression model

The Simple Linear Regression model is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \text{ for } i = 1, 2, \ldots, n$$

where

- $\epsilon_1, \ldots, \epsilon_n$ are independent with $\epsilon_i \sim N\left(0, \sigma^2\right)$. The variables ϵ_i are called the random errors.
- β_0, β_1 are unknown parameters to be estimated
- β_0 is the intercept, β_1 is the slope.
- Y is a response variable (random), with observed values $y_1, \ldots, y_n$.

- $x_1, \ldots, x_n$ are observed values of an explanatory variable (not random).
  Note that $Y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$.

To predict, we use $\widehat{Y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$.
$\widehat{Y_0}$ is referred to as the **fitted value** at x_0.

Least Squares estimation of parameters β_0, β_1

We define the Sum of Squared Deviations as

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

and our aim is to find the estimates $\hat{\beta}_0, \hat{\beta}_1$ to make S as small as possible.

通过求导实现

To sum up, the Least Squares parameter estimates for the Simple Linear Regression are

$$\widehat{\beta_1} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} \bar{x}$$

where $\bar{x} = \frac{1}{n} \sum x_i$ and $\bar{y} = \frac{1}{n} \sum y_i$
Note: The estimate of β_0 is obtained simply by observing that the sample mean point $(\bar{x}, \bar{y})$ lies on the fitted line, that is, $\bar{y} = \widehat{\beta_0} + \widehat{\beta_1} \bar{x}$

Residual

The difference is called the residual 预测值和真实值的差距

residual $e_i = y_i - \widehat{y_i}$ for each $i = 1, 2, \ldots, 8$.

Distributions of $\widehat{\beta_0}, \widehat{\beta_1}$

In a similar way to ANOVA, we define the Error Sum of Squares

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)^2$$

It can be shown that

$$\frac{SSE}{\sigma^2} \sim \chi_{n-2}^2$$

We define the Mean Square Error as

$$MSE = \frac{SSE}{n-2}.$$

Expectations of $\widehat{\beta_0}, \widehat{\beta_1}$

$$E\left[\widehat{\beta_1}\right] = \beta_1$$

The estimator $\widehat{\beta_1}$ is unbiased for β_1.

$$E\left[\widehat{\beta_0}\right] = \beta_0$$

The estimator $\widehat{\beta_0}$ is unbiased for β_0.

因此, $\begin{aligned} E\left[\widehat{Y_0}\right] &= E\left[\widehat{\beta_0} + \widehat{\beta_1} x_0\right] = E\left[\widehat{\beta_0}\right] + x_0 E\left[\widehat{\beta_1}\right] \\ &= \beta_0 + \beta_1 x_0 = E\left[Y_0\right]. \end{aligned}$

Distribution of $\widehat{\beta_1}$

Variance: $\operatorname{Var}\left[\widehat{\beta_1}\right] = \frac{\sigma^2}{\sum_i x_i^2 - n\bar{x}^2}$

结合expectation, 得到分布为 $\widehat{\beta_1} \sim N\left(\beta_1, \frac{\sigma^2}{\sum_i x_i^2 - n\bar{x}^2}\right)$

Distribution of $\widehat{\beta_0}$

$\operatorname{Var}\left[\widehat{\beta_0}\right] = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i x_i^2 - n\bar{x}^2}\right)$

$\widehat{\beta_0} \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i x_i^2 - n\bar{x}^2}\right)\right)$.

Confidence Intervals for β_0, β_1

Inference in linear regression

Recalling that $\widehat{\sigma} = \sqrt{MSE}$, then the estimated standard error of $\widehat{\beta_1}$ is $e.s.e.\left(\widehat{\beta_1}\right) = \frac{\widehat{\sigma}}{\sqrt{\sum x_i^2 - n\bar{x}^2}}$

Going through the same calculations for $\widehat{\beta_0}$, we get that

$$\frac{\widehat{\beta_0} - \beta_0}{e.s.e.\left(\widehat{\beta_0}\right)} \sim t_{n-2}$$

with

$$e.s.e.\left(\widehat{\beta_0}\right) = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2 - n\bar{x}^2}}$$

Confidence intervals for β_0, β_1

The $100(1-\alpha)\%$ confidence interval for β_1 is

$$\widehat{\beta_1} \pm t_{n-2}(\alpha/2) \times e.s.e.\left(\widehat{\beta_1}\right).$$

The $100(1-\alpha)\%$ confidence interval for β_0 is

$$\widehat{\beta_0} \pm t_{n-2}(\alpha/2) \times e.s.e. \left(\widehat{\beta_0}\right)$$

Confidence Intervals for a Mean Response and Prediction Intervals

Confidence Intervals for a Mean Response

Recall that

- we have $Y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0$ with $E[\epsilon_0] = 0$,
- the mean response at x_0 is $E[Y_0] = \beta_0 + \beta_1 x_0$,
- the predicted response is $\widehat{Y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$.
  The mean response $\beta_0 + \beta_1 x_0$ is not a random variable, it is an **unknown constant**, so we can work out a confidence interval for it.
  The **actual** response Y_0 is a **random variable**.

$$\operatorname{Var}\left[\widehat{Y_0}\right] = \sigma^2 h_{00}$$

$$\text{where} \quad h_{00} = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}.$$

We need to estimate σ, using $\widehat{\sigma} = \sqrt{MSE}$ as before.
Then a $100(1-\alpha)\%$ CI for the mean response at x_0 is

$$\left[\widehat{Y_0} - t_{n-2}\left(\frac{\alpha}{2}\right)\sqrt{MSE \times h_{00}}, \widehat{Y_0} + t_{n-2}\left(\frac{\alpha}{2}\right)\sqrt{MSE \times h_{00}}\right]$$

Prediction intervals

Suppose we want to predict the response Y_0 corresponding to x_0.
We have

$$Y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0$$

So far, we have worked out a confidence interval for the **mean response** $\beta_0 + \beta_1 x_0$.
If we want to work out an interval for the **particular response** Y_0, we need to add in the extra uncertainty from ϵ_0.
That is, the mean is unchanged (since $E[\epsilon_0] = 0$), but we add an extra σ^2 to the variance term (since $\operatorname{Var}[\epsilon_0] = \sigma^2$).
The relevant variance becomes $\sigma^2(1 + h_{00})$.
Estimating σ by $\sqrt{MSE}$, then a $100(1-\alpha)\%$ prediction interval (PI) for a response at x_0 is

$$\left[\widehat{Y_0} - t_{n-2}\left(\frac{\alpha}{2}\right)\sqrt{MSE \times (1 + h_{00})}, \widehat{Y_0} + t_{n-2}\left(\frac{\alpha}{2}\right)\sqrt{MSE \times (1 + h_{00})}\right]$$

加上一个噪声的范围

Notes about confidence and prediction intervals

- The PI for an individual observation is wider than the CI for a mean response.
- Averaging reduces uncertainty / variability.
- The width of the CI and PI depends on x_0 only through the value of h_{00}.
- Recall $h_{00} = \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum x_i^2 - n\bar{x}^2}$ and notice that the value x_0 only appears in the term $(x_0 - \bar{x})^2$.
- Hence the further away from $\bar{x}$ we go, the bigger h_{00} gets, and the wider the interval (CI or PI) is.
- Our prediction becomes more uncertain as we move away from the mean $\bar{x}$ of our original observed data.

区分题目中所言 mean response 和 response 的区别

Hypothesis Testing in Linear Regression

Test for the slope

To test $H_0 : \beta_1 = c$ versus an alternative $H_1 : \beta_1 < c, H_1 : \beta_1 > c$, or $H_1 : \beta_1 \neq c$, for any given constant c, we compute the test statistic

$$T = \frac{\widehat{\beta_1} - c}{e.s.e.\left(\widehat{\beta_1}\right)}$$

and compare it with t_{n-2} tables. The rejection rules are the same as for a one-sample t-test, only with the test statistic stated above.

- For the alternative $H_1 : \beta_1 < c_1$ we reject H_0 at α if $T < -t_{n-2}(\alpha)$.
- For the alternative $H_1 : \beta_1 > c$, we reject H_0 at α if $T > t_{n-2}(\alpha)$.
- For the alternative $H_1 : \beta_1 \neq c$, we reject H_0 at α if $T < -t_{n-2}\left(\frac{\alpha}{2}\right)$ or $T > t_{n-2}\left(\frac{\alpha}{2}\right)$, in other words, if $|T| > t_{n-2}\left(\frac{\alpha}{2}\right)$. Remember to interpret your conclusion practically in the context of the question.

Test for the intercept 看清是测试斜率还是截距，是one-sided还是two-sided

To test $H_0 : \beta_0 = c$ versus an alternative $H_1 : \beta_0 < c_1 H_1 : \beta_0 > c$, or $H_1 : \beta_0 \neq c$ for any given constant c, we compute the test statistic

$$T = \frac{\widehat{\beta_0} - c}{e.s.e.\left(\widehat{\beta_0}\right)}$$

and compare with t_{n-2} tables.

- For the alternative $H_1 : \beta_0 < c_1$ we reject H_0 at α if $T < -t_{n-2}(\alpha)$.
- For the alternative $H_1 : \beta_0 > c$, we reject H_0 at α if $T > t_{n-2}(\alpha)$.
- For the alternative $H_1 : \beta_0 \neq c$, we reject H_0 at α if $T < -t_{n-2}\left(\frac{\alpha}{2}\right)$ or $T > t_{n-2}\left(\frac{\alpha}{2}\right)$, in other words, if $|T| > t_{n-2}\left(\frac{\alpha}{2}\right)$. Remember to interpret your conclusion practically in the context of the question.

Anova for regression

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

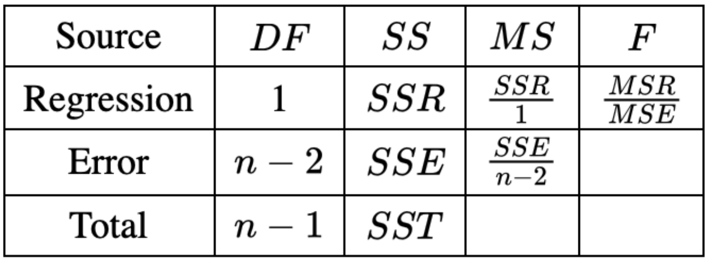
$$SSE = \sum_{i=1}^{n} \left(y_i - \widehat{y_i}\right)^2 \quad MSE = \frac{SSE}{n-2}$$

$$SSR = \sum_{i=1}^{n} \left(\widehat{y}_i - \bar{y}\right)^2 \quad MSR = \frac{SSR}{1}$$

因此， $SST = SSR + SSE$

利用ANOVA table实现

We can draw up an ANOVA table and compute an F-statistic (similar to the standard ANOVA). This is used to test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$



Source	DF	SS	MS	F
Regression	1	SSR	$\frac{SSR}{1}$	$\frac{MSR}{MSE}$
Error	$n-2$	SSE	$\frac{SSE}{n-2}$	
Total	$n-1$	SST		

with $SST = SSE + SSR$.
We compare $F = \frac{MSR}{MSE}$ with $F_{1,n-2}$ tables since $\frac{MSR}{MSE} \sim F_{1,n-2}$.
We reject $H_0 : \beta_1 = 0$ if $F > F_{1,n-2}(\alpha)$
That is, we reject H_0 if the variation explained by the regression relationship is large, compared to the variation explained by pure randomness.
In other words, we reject H_0 if the regression relationship does usefully explain some of the variation in the data.
So, the ANOVA F-test for regression tests whether there really is a linear relationship between x and y, which is equivalent to testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

Notes about the ANOVA F-test for regression:

- The ANOVA F-test for regression is a one-sided test for the same reason as the standard ANOVA test. We actually compare two estimators of variance - MSR and MSE and we test whether MSR is significantly bigger than MSE.
- The ANOVA F-test for regression can be only used to test the alternative $H_1 : \beta_1 \neq 0$ and it is equivalent to the two-sided t test for the slope. These two tests are equivalent because $F_{1,n-2} \sim \left(t_{n-2}\right)^2$.
- If you want to test the alternative $H_1 : \beta_1 > 0$ or $H_1 : \beta_1 < 0$, then you need to use a one-sided t-test for the slope which was discussed earlier.

Model assessment

We might want to assess how useful the model is in explaining the data. This can be assessed (informally) by R^2, formally tested using a t -test for the slope or ANOVA F-test for regression, and also assessed by other tools discussed in this section.
Recall:

- SSR measures how much of the variation in the y values is explained by the regression,
- SST measures the total amount of variation in the y values.
  To measure the proportion of total variation explained by the regression, we can use the ratio

$$R^2 = \frac{SSR}{SST}$$

Note that $0 \leq R^2 \leq 1$ since $0 \leq SSR \leq SST \Rightarrow 0 \leq \frac{SSR}{SST} \leq \frac{SST}{SST} \Rightarrow 0 \leq R^2 \leq 1$.

R 接近0模型不可靠 接近1可靠

Residuals

Once the model has been fitted, the residuals are defined by

$$e_i = y_i - \widehat{y_i} = y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i$$

and the Error Sum of Squares is

$$SSE = \sum_{i=1}^{n} e_i^2$$

The Least Squares estimation method chooses $\widehat{\beta_0}, \widehat{\beta_1}$ to make SSE as small as possible.
The e_i values are estimates of the ϵ_i.
The standard assumption about the errors is that $\epsilon_1, \ldots, \epsilon_n$ are independent with $\epsilon_i \sim N\left(0, \sigma^2\right)$.
To check this, we examine the residuals e_i :

- Plot a histogram of the residuals to see if it looks like a normal distribution.
- More formally, we do a normal probability plot and find the p-value for the normality test.
  To see if the model is appropriate, plot the residuals against the fitted values (or plot the residuals against the explanatory variable).
- If the linear regression model is appropriate, there should be no pattern in the plots.

Residual 本身满足正态分布， 但是 residuals against the fitted values 的图像应该是没有pattern的（区分）

Key steps in regression

1. Draw a scatterplot of the data, see if it looks like a straight line that could usefully describe the relationship.

2. If so, compute the Least Squares parameter estimates $\widehat{\beta_0}, \widehat{\beta_1}$.

3. Assess whether the fitted model is adequate by
   - computing R^2 value,
   - formally testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$,
   - examining residuals to see if normally distributed,
   - plot residuals against fitted values or explanatory variables to see if there are any patterns.
4. We may then go on to
   - work out confidence intervals for β_0, β_1,
   - test other hypotheses, e.g. $H_0 : \beta_0 = 0$ against $H_1 : \beta_0 \neq 0$,
   - compute prediction intervals for future response values Y_0.

Multiple Linear Regression and Model Transformations

Usually, a response Y depends on many things, not just on a single explanatory variable x.
Example: Amount of fuel (tons of coal) required to heat a particular building for a week may depend upon average temperature during the week, average wind velocity, and cloud cover.

$$\begin{aligned} Y_i &= \text{Coal required (tons) during week i} \\ x_{1i} &= \text{Average temperature (degrees Celsius)} \\ x_{2i} &= \text{Average wind velocity (mph)} \\ x_{3i} &= \text{Cloud cover (percent)} \end{aligned}$$

A linear model here would be

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

and one may test whether this provides a good fit to the data.

transformation是一些变形，为了满足模型标准式子的一次形式。例如我们可以将polynomial中的高次令为新的一次变量，类似$y = x^2$

附录 R操作

- 创建dataframe

```
fuelDF <- data.frame('temp'=c(28,28,32.5,39,45.9,57.8,58.1,62.5),
                             'fuel'=c(12.4,11.7,12.4,10.8,9.4,9.5,8.0,7.5))
```

c()代表生成list或者vector，上下行个数不匹配可以插入NA，计算时记得加上na.rm=TRUE

- 导入excel表格

```
library(readxl)
fuel <- read_excel("Desktop/fuel.xlsx")
```

- Scatterplot

```
plot(fuel$temp,fuel$fuel, main='scatterplot', xlab='temp', ylab='fuel')
```

main代表title

- p求cdf概率

```
pnorm(1.56, mean=1.2, sd=sqrt(8))
```

- q求概率对应的critical value

```
qnorm(0.05,mean=0,sd=1,lower.tail=FALSE) #lower.tail取左半边还是右半边
qnorm(0.95,mean=0,sd=1,lower.tail=TRUE)
```

```
#t分布
qt(0.95, 10, lower.tail = TRUE) #10是df:degree of freedom
qt(0.05, 10, lower.tail = FALSE) #0.05在表中对应P=5，即a/2=0.05（P需要除以100为所求真实值）
```

```
#F分布
qf(0.05, 5, 7, lower.tail = FALSE) #(m,n)为自由度
qf(0.05, 7, 5, lower.tail = FALSE)
```

```
#chi-squared
qchisq(0.01, 9, lower.tail = FALSE) #第二位为自由度
qchisq(0.95, 9, lower.tail = FALSE)
```

- 一张图上画两条曲线 d是pdf的值

```
#standard normal and t_10 in one picture
y4<-seq(from=-5,to=+5, length.out=100000)
plot(y4,dt(y4, 10), main="t-distribution", type="l", col ="purple")
points(y4,dnorm(y4), main="standard normal", type="l", col ="red") #points函数将点画于已经存的坐标系上
```

type是连接形式，l是用line连接，p是用point，b是既line又point

dnorm(x,mean=-2,sd=5) 所有pdf的值，cdf是pdf的积分

- histgram

```
hist(density$Density, main = paste("Histogram of Density"), xlab = "Density", col="purple")
#paste()可以将任意数量的参数组合到一起，此处只有一个字符串，可以省略
```

- boxplot箱线图(额外画出mean)

```
A <- boilers$`Type A

B <- boilers$`Type B
means <- c(mean(A), mean(B,na.rm=TRUE)) # na.rm是删除空值的意思
boxplot(A, B, col="purple", main = paste("Boxplots"), names=c("Type A", "Type B"),
    boxwex=0.4)
points(means, pch=9)
```

计算sd，var，mean

```
sd(x)

var(x) #分母自动设置为n-1
```

> mean(x)