代码建议 三 （截止到章节 2.1.4.3 多元线性回归之前）

1. 对章节名称的建议

   2.1.4.1 预备的知识——反函数

   可以改成“2.1.4.1 预备知识”

   3）微分（与回归）

   可以改成“3）回归与微分”

   因为下面的章节，显示提到了一个回归例子，然后引出了微分

   但是此部分对微分的介绍太过简单，建议进行适当补充

函数的微分（英语：Differential of a function）是指对函数的局部变化的一种线性描述。微分可以近似地描述当函数自变量的取值作足够小的改变时，函数的值是怎样改变的。微分是指函数在某一点处（趋近于无穷小）的变化量，是一种变化的量。

微分本质是一个微小的线性变化量，是用一个线性函数作为原函数变化的逼近（或者叫近似）

微分在数学中的定义：由 y 是 x 的函数(y=f(x))。从简单的 x-y 坐标系来看，自变数 x 有微小的变化量时(d/dx)，应变数 y 也会跟着变动，但 x 跟 y 的变化量都是极小的。当 x 有极小的变化量时，我们称对 x 微分。微分主要用于线性函数的改变量，这是微积分的基本概念之一

微分法则

- $d(au+bv)=dau+dbv=adu+bdv$
- $d(uv)=udv+vdu$
- $d\left(\frac{u}{v}\right)=\frac{vdu-udv}{v^2}$
- 若函数$y(u)$可导，那么$d[y(u)]=y'(u)du$

- 常用公式求微分：

1. $y=x$，关于x进行微分，$\frac{dy}{dx}=1$
2. $y=x^2$，关于x进行微分，$\frac{dy}{dx}=2x$
3. $y=\frac{1}{x}$，关于x进行微分，$\frac{dy}{dx}=-x^{-2}$
4. $y=\frac{1}{x^2}$，关于x进行微分，$\frac{dy}{dx}=-2x^{-3}$
5. $y=(5x-7)^2$，关于x进行微分，$\frac{dy}{dx}=2(5x-7)\times 5$
6. $y=(ax+b)^n$，关于x进行微分，$\frac{dy}{dx}=n(ax+b)^{n-1}\times a$
7. $y=e^x$，关于x进行微分，$\frac{dy}{dx}=e^x$
8. $y=logx$，关于x进行微分，$\frac{dy}{dx}=\frac{1}{x}$
9. $y=log(ax+b)$，关于x进行微分，$\frac{dy}{dx}=\frac{1}{ax+b}\times a$
10. $y=log(1+ea^{x+b})$，关于x进行微分，$\frac{dy}{dx}=\frac{1}{1+e^{ax+b}}\times ae^{ax+b}$

“常用公式求微分：”

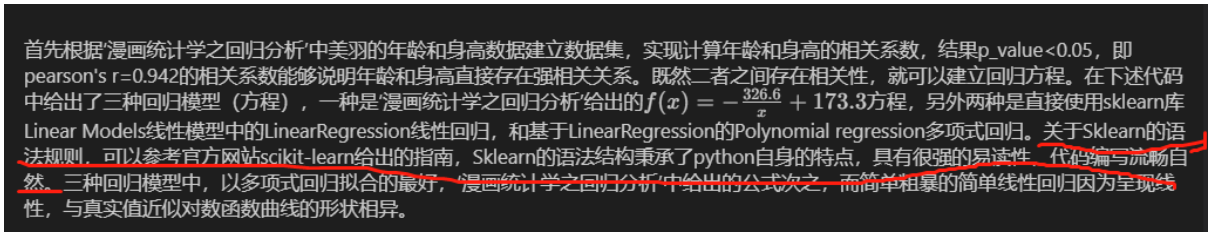
可以改为，**常见函数求微分举例**

2. 3）微分（与回归）中部分叙述的修订

首先根据‘漫画统计学之回归分析’中美羽的年龄和身高数据建立数据集，实现计算年龄和身高的相关系数，结果 p_value<0.05，即 pearson's r=0.942 的相关系数能够说明年龄和身高直接存在强相关关系。

改为

首先根据‘漫画统计学之回归分析’中美羽的年龄和身高数据建立数据集，实现计算年龄和身高的相关系数，结果 p 值小于 0.05，即值为 0.942 皮尔森关系系数能够说明年龄和身高直接存在强相关关系。



此句（下面给出以及上图红线标出）可以作为标注或者提示，位于本段之后

“关于 Sklearn 的语法规则，可以参考官方网站 scikit-learn 给出的指南，Sklearn 的语法结构秉承了 python 自身的特点，具有很强的易读性，代码编写流畅自然。”

3. 2.1.4.2 简单线性回归 修改

原文	修改
在统计学中，线性回归（linear regression）是利用称为线性回归方程的最小平方函数对一个或多个自变量和因变量之间关系进行建模的一种回归分析。这种函数是一个或多个称为回归系数的模型参数的线性组合。只有一个自变量的情况称为简单（线性）回归（simple linear regression），大于一个自变量情况的叫多元回归（multivariable linear regression）。 * 回归分析的流程： 1. 为了讨论是否具有求解回归方程的意义，画出自变量和因变量的散点图（求解相关系数）； 2. 求解回归方程； 3. 确认回归方程的精度； 4. 进行回归系数的检验； 5. 总体回归 Ax+b 的估计； 6. 进行预测	在统计学中，线性回归（linear regression）是利用称为线性回归方程的最小平方函数对一个或多个自变量和因变量之间关系进行建模的一种回归分析。这种函数是一个或多个称为回归系数的模型参数的线性组合。只有一个自变量的情况称为简单线性回归（simple linear regression），大于一个自变量情况的叫多元回归（multivariable linear regression）。 * 回归分析的流程： 1. 数据可视化：为了讨论是否具有求解回归方程的意义，画出自变量和因变量的散点图，并求解相关系数； 2. 求解回归方程； 3. 确认回归方程的精度； 4. 进行回归系数的检验； 5. 总体回归 Ax+b 的估计； 6. 进行预测

此部分的内容和代码也应该调整下
回归分析流程中，一中提到，先绘制散点图了，而现在文中，构造数据集之后，直接提到了求解回归方程，在其后的代码中，也包含了散点图的绘制，因此建议修改结构

2）求解回归方程

求解回归方程使用了两种方法，一种是逐步计算的方式；另一种是直接使用sklearn库的LinearRegression模型。逐步计算的方式可以更深入的理解回归模型，熟悉基本计算过程之后，直接应用sklearn机器学习库中的模型也会对各种参数的配置有个比较清晰的了解。首先计算温度与销量之间的相关系数，确定二者之间存在关联，其p_value=7.661412804450245e-06，小于0.05的显著性水平，确定pearson's r=0.90能够表明二者之间是强相关性。

求解回归方程即是使所有真实值与预测值之差的和为最小，求出a和b，就是所有变量残差residual的平方s_residual的和S_residual为最小。因为温度与销量为线性相关，因此使用一元一次方程式：$y = ax + b$，x为自变量温度，y为因变量销量，a和b为回归系数（参数），分别称为斜率（slop）和截距(intercept)，求解a和b的过程，可以使用最小二乘法（least squares method），又称最小平方法，通过最小化误差的平方（残差平方和）寻找数据的最佳函数匹配。为残差平方和：$(-34a - b + 93)^2 + (-33a - b + 91)^2 + (-32a - b + 80)^2 + (-31a - b + 73)^2 + (-31a - b + 75)^2 + (-31a - b + 84)^2 + (-30a - b + 84)^2 + (-29a - b + 64)^2 + (-29a - b + 77)^2 + (-28a - b + 62)^2 + (-26a - b + 65)^2 + (-25a - b + 51)^2 + (-25a - b + 59)^2 + (-24a - b + 58)^2$，

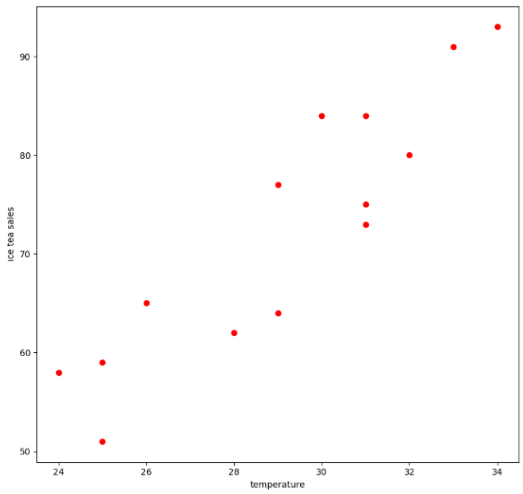
可以将以下部分的代码和图，放到“2）求解回归方程”之前，此部分文字和解释也在下面做了修改。
为了查看方便，其实可以把三个图分别绘制，
这样，横纵坐标也会更加清晰，查看方便，纸质书初版时，排版也会方便

此部分，文字较长，代码也很长，其实是不易于读着理解的，建议用类似的方法，拆分大段文字内容和代码，读者对于相对应的文字解释和代码，生成的图，会更加容易接受。

```
import math
import sympy
from sympy import diff,Eq,solveset,solve,simplify
import matplotlib.pyplot as plt
from matplotlib import cm
import numpy as np

r_=stats.pearsonr(iceTea_df.temperature,iceTea_df.iceTeaSales)
print("_"*50)
print(
    "pearson's r:",r_[0],"\n",
    "p_value:",r_[1]
    )
print("_"*50)

#原始数据散点图
fig, axs=plt.subplots(1,3,figsize=(25,8))
axs[0].plot(iceTea_df.temperature,iceTea_df.iceTeaSales,'o',label='ground truth',color='r')
axs[0].set(xlabel='temperature',ylabel='ice tea sales')
```



观察散点图可知，温度和销售具有一定的线性关系。进一步计算温度与销量之间的相关系数，得到 p value=7.661，小于 0.05 的显著性水平，且 pearson's r=0.90，可以确定二者之间是强相关性，适合采用线性拟合模型求解二者之间的关联公式。

此处我们可以假设两个变量之间是线性相关关系，因此采用一元一次函数$y=ax+b$，来拟合两者之间的关系，其中x为自变量温度，y为因变量销量，a和b为回归系数（参数），分别称为斜率（slop）和截距(intercept)。求解回归方程即是使所有真实值与预测值之差的和为最小，在此问题中，就是求出一组 a 和 b 的值，使所有变量残差`residual`的平方`s_residual`的和`S_residual`为最小。，求解 a 和 b 的过程，可以使用最小二乘法（least squares method），又称最小平方法，通过最小化误差的平方（残差平方和）寻找数据的最佳函数匹配。

手动计算的残差平方和为：

$(-34a-b+93)^{2} +(-33a-b+91)^{2}+(-32a-b+80)^{2}+(-31a-b+73)^{2}+(-31a-b+75)^{2}+(-31a-b+84)^{2}+(-30a-b+84)^{2}+(-29a-b+64)^{2}+(-29a-b+77)^{2}+(-28a-b+62)^{2}+(-26a-b+65)^{2}+(-25a-b+51)^{2}+(-25a-b+59)^{2}+(-24a-b+58)^{2}$,

4. 代码中注释不准确，需要修改

```
#4 - 对残差平方和S_residual关于a和b求微分，并使其为0
diff_S_residual_a=diff(S_residual,a)
diff_S_residual_b=diff(S_residual,b)
print("diff_S_residual_a=",)
pprint(diff_S_residual_a)
print("\n")
print("diff_S_residual_b=",)
pprint(diff_S_residual_b)

Eq_residual_a=Eq(diff_S_residual_a,0) #设所求a微分为0
Eq_residual_b=Eq(diff_S_residual_b,0) #设所求b微分为0
slop_intercept=solve((Eq_residual_a,Eq_residual_b),(a,b)) #计算二元一次方程组
print("_"*50)
print("slop and intercept:\n")
pprint(slop_intercept)
slop=slop_intercept[a]
intercept=slop_intercept[b]
```

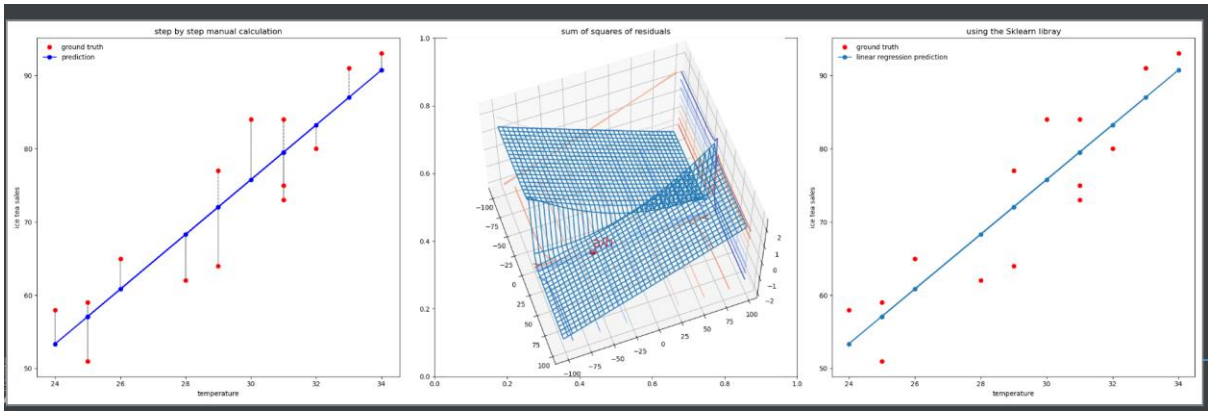
diff_S_residual_a=24040·a + 816·b - 60188

diff_S_residual_b=816·a + 28·b - 2032

```
# 对 a 和 b 的微分公式构造二元一次方程组，
# 即 diff_S_residual_a = 0，  diff_S_residual_a = 0
Eq_residual_a=Eq(diff_S_residual_a,0)
Eq_residual_b=Eq(diff_S_residual_b,0)
# 设置二元一次方程组中的变量为 a 和 b，并求解
slop_intercept=solve((Eq_residual_a,Eq_residual_b),(a,b))
```

5. 图片尺寸和分辨率问题

现有 notebook 中生成的图片，大多尺寸较大，且分辨率低，如下图



减低降低图片尺寸，设置图片分辨率

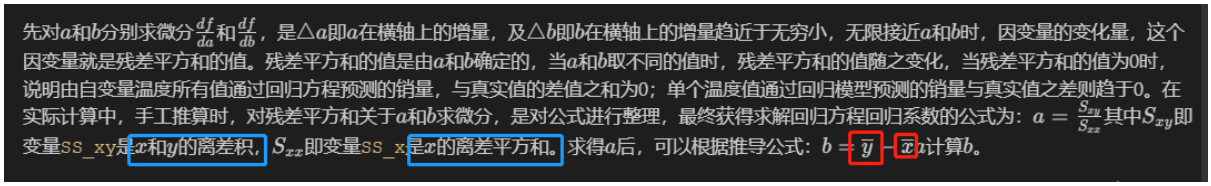
解决方法，设置矢量图

```
import matplotlib.pyplot as plt
%matplotlib inline
%config InlineBackend.figure_format = 'svg' #  'svg'改为'retina'也可以
```

解决方法，设置图片 dpi 分辨率

```
fig = plt.figure(figsize=(10,6), dpi=100)
```

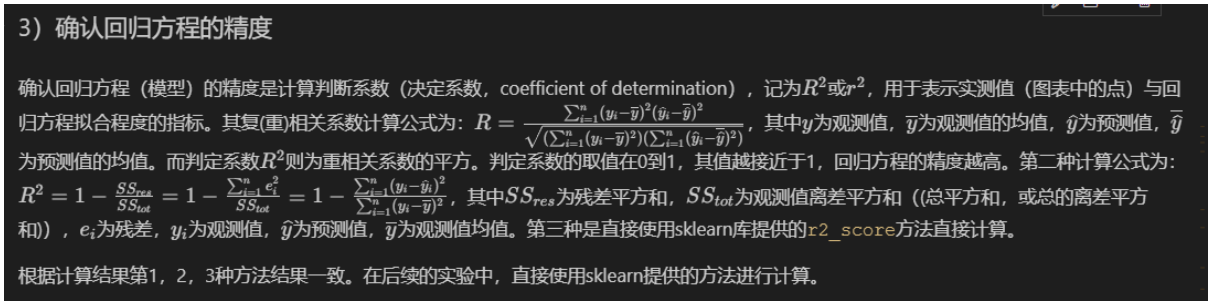
另外，figsize，dpi 会同时影响图片大小和字体相对大小，绘制时，需要根据图片内容密集程度来调整



此部分，建议添加，对$\bar{x}$和$\bar{y}$的解释，即，x 和 y 的平均值（即，iceTea_df.iceTeaSales.mean()和 iceTea_df.temperature.mean()）

更进一步，可以说明，此处的 a 和 b 对应代码中的变量 slop_和 intercept_

6. 3）确认回归方程的精度



此部分可以做出如下修改

原文	修改
确认回归方程（模型）的精度是计算判断系数（决定系数，coefficient of determination），记为$R^{2} $或$r^{2} $，用于表示实测值（图表中的点）与回归方程拟合程度的指标。其复(重)相关系数计算公式为	回归方程（模型）的精度可以通过计算判断系数确认。判断系数（也称决定系数，coefficient of determination），记为$R^{2} $或$r^{2} $，用于度量因变量的变异中可由自变量解释部分所占的比例，以此来判断回归模型的解释力，来表示实测值（图表中的点）与回归方程拟合的程度。对于简单线性回归而言，判断系数为样本相关系数 R 的平方。其复(重)相关系数计算公式为

7. 回归系数的检验，F 分布

4）回归系数的检验（回归显著性检验）| F分布与方差分析

F-分布（F-distribution）是一种连续概率分布，广泛应用于似然比率检验，特别是方差分析（Analysis of variance, ANOVA，或变异数分析）中。对于F-分布的阐释使用scipy.stats.f的官方案例。函数方法基本同正态分布和t分布。

```
from scipy.stats import f
import matplotlib.pyplot as plt
fig, ax=plt.subplots(1, 1)

dfn, dfd=29, 18
mean, var, skew, kurt=f.stats(dfn, dfd, moments='mvsk')
print("mean=%f, var=%f, skew=%f, kurt=%f"%(mean, var, skew, kurt))
```

此部分中，对 F 分布较为简单，建议稍加补充

F 分布是两个服从卡方分布的独立随机变量各除以其自由度后的比值的抽样分布，是一种非对称分布，且位置不可互换.

如果随机变量 X 有参数为 d_1 和 d_2 的 F-分布，我们写作 $X \sim F(d_1, d_2)$。那么对于实数 $x \geq 0$，X 的概率密度函数 (pdf)是

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1\,x)^{d_1}\;\;d_2^{d_2}}{(d_1\,x+d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

$$= \frac{1}{\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1+\frac{d_1}{d_2}\,x\right)^{-\frac{d_1+d_2}{2}}$$

一个F-分布的随机变量是两个卡方分布变量除以自由度的比率：

$$\frac{U_1/d_1}{U_2/d_2} = \frac{U_1/U_2}{d_1/d_2}$$

其中：

- U_1和U_2呈卡方分布，它们的自由度（degree of freedom）分别是d_1和d_2。
- U_1和U_2是相互独立的。

F-分布 - 维基百科，自由的百科全书
https://zh.wikipedia.org/wiki/F-%E5%88%86%E5%B8%83

且对于代码中的 “dfn, dfd=29, 18”
需要添加注释

```
# 分别是第一自由度即分子中卡方分布的自由度，和第二自由度，即分母中卡方分布的自由度
```

8. 对总平方和的解释的修改

此以下部分，建议做出如下修改，增加可读性

- 总平方和=回归平方和+残差平方和

公式为：$SS_{tot}=\sum_{i=1}^{n}(y_i-\overline{y})^2=SS_{reg}+SS_{res}=\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2$，式中$SS_{reg}$回归平方和，其它同上。回归平方和是预测值（回归值）与观测值（真实值、实测值）均值之差的平方和，该统计量反映了自变量$x_1,x_2,\ldots,x_m$,的变化引起的$y(y_k(k=1,2,\ldots,n))$的波动，其自由度为$df_{reg}=m$，其中m为自变量的个数，温度与销量求解的一元一次线性方程只有一个自变量，因此其自由度为1，即只有这一个因素可以自由变化；残差平方和是观测值与预测值之差的平方和，残差的存在是由实验误差及其它因素引起的，其自由度为$df_{res}=n-m-1$，其中n为样本数量，即对应的y的取值数量。总的离差平方和SS_{tot}的自由度为$n-1$。

修改后，效果如下

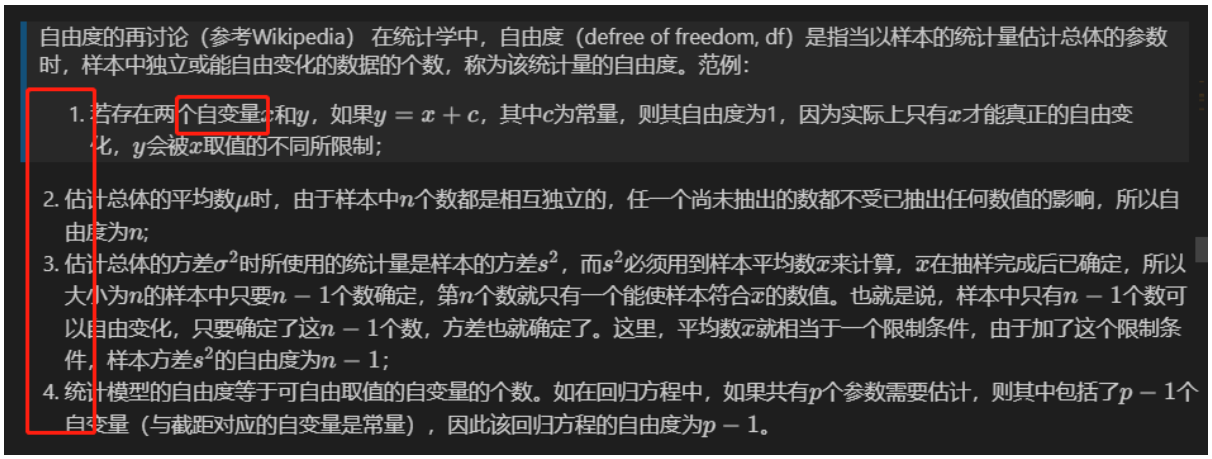
- $SS_{tot}=SS_{reg}+SS_{res}$，即总平方和=回归平方和+残差平方和

公式为：$SS_{tot}=\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2$，式中$SS_{reg}$回归平方和，其它同上。回归平方和$SS_{reg}=\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2$，是预测值（回归值）与观测值（真实值、实测值）均值之差的平方和，该统计量反映了自变量$x_1,x_2,\ldots,x_m$,的变化引起的$y(y_k(k=1,2,\ldots,n))$的波动，其自由度为$df_{reg}=m$，其中m为自变量的个数，温度与销量求解的一元一次线性方程只有一个自变量，因此其自由度为1，即只有这一个因素可以自由变化；残差平方和$SS_{res}=\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2$，是观测值与预测值之差的平方和，残差的存在是由实验误差及其它因素引起的，其自由度为$df_{res}=n-m-1$，其中n为样本数量，即对应的y的取值数量。总的离差平方和SS_{tot}的自由度为$n-1$。

观测值（样本）通常是给定的，因此总的离差平方和是固定的，构成总的离差平方和的因素为回归平方和和残差平方和，分布代表所求得的回归方程，或实验误差和其它因素引起y值的变化，当残差平方和越小（就是实验误差和其它因素影响小），则回归平方和越大，则说明所求得的回归方程的预测值越准确。

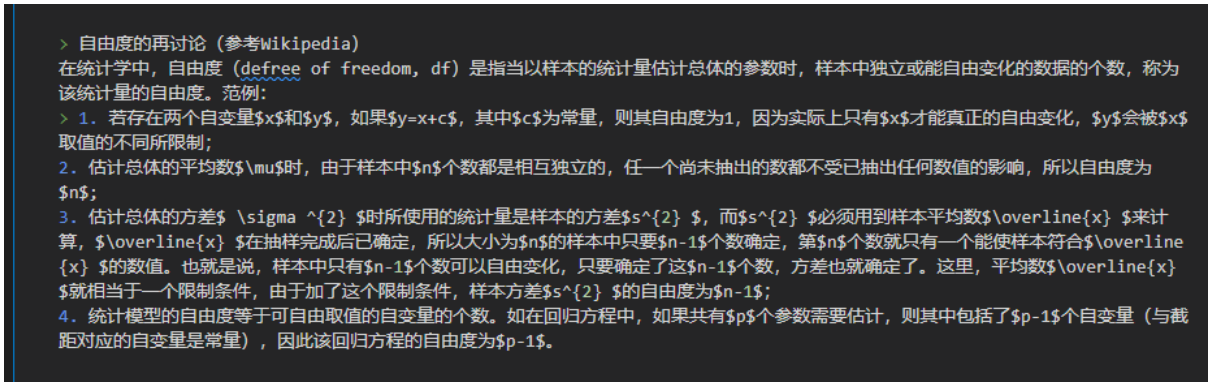
原始	修改
* 总平方和=回归平方和+残差平方和 公式为： $SS_{tot}=\sum_{i=1}^n (y_{i} - \overline{y})^{2}=SS_{reg}+SS_{res}= \sum_{i=1}^n (\widehat{y} _{i} - \overline{y})^{2} + \sum_{i=1}^n (y_{i} - \widehat{y} _{i})^{2} $，式中$SS_{reg}$回归平方和，其它同上。回归平方和是预测值（回归值）与观测值（真实值、实测值）均值之差的平方和，该统计量反映了自变量$x_{1}, x_{2}, \ldots ,x_{m}, $的变化引起的$y$($y_{k} (k=1,2, \ldots ,n)$)的波动，其自由度为$df_{reg}=m $，其中$m$为自变量的个数，温度与销量求解的一元一次线性方程只有一个自变量，因此其自由度为 1，即只有这一个因素可以自由变化；残差平方和是观测值与预测值之差的平方和，残差的存在是由实验误差及其它因素引起的，其自由度为$df_{res}=n-m-1 $，其中$n$为样本数量，即对应的$y$的取值数量。 总的离差平方和$SS_{tot}$的自由度为$n-1$。	* $SS_{tot}=SS_{reg}+SS_{res} $，即总平方和=回归平方和+残差平方和 公式为：$SS_{tot}=\sum_{i=1}^n (y_{i} - \overline{y})^{2}= \sum_{i=1}^n (\widehat{y} _{i} - \overline{y})^{2} + \sum_{i=1}^n (y_{i} - \widehat{y} _{i})^{2} $，式中$SS_{reg}$回归平方和，其它同上。回归平方和$SS_{reg}=\sum_{i=1}^n (\widehat{y} _{i} - \overline{y})^{2} $，是预测值（回归值）与观测值（真实值、实测值）均值之差的平方和，该统计量反映了自变量$x_{1}, x_{2}, \ldots ,x_{m}, $的变化引起的$y$($y_{k} (k=1,2, \ldots ,n)$)的波动，其自由度为$df_{reg}=m $，其中$m$为自变量的个数，温度与销量求解的一元一次线性方程只有一个自变量，因此其自由度为 1，即只有这一个因素可以自由变化；残差平方和$SS_{res} =\sum_{i=1}^n (y_{i} - \widehat{y} _{i})^{2}$，是观测值与预测值之差的平方和，残差的存在是由实验误差及其它因素引起的，其自由度为$df_{res}=n-m-1 $，其中$n$为样本数量，即对应的$y$的取值数量。总的离差平方和$SS_{tot}$的自由度为$n-1$。

9. 自由度的在讨论部分的修改建议



自由度的再讨论（参考Wikipedia） 在统计学中，自由度（defree of freedom, df）是指当以样本的统计量估计总体的参数时，样本中独立或能自由变化的数据的个数，称为该统计量的自由度。范例：

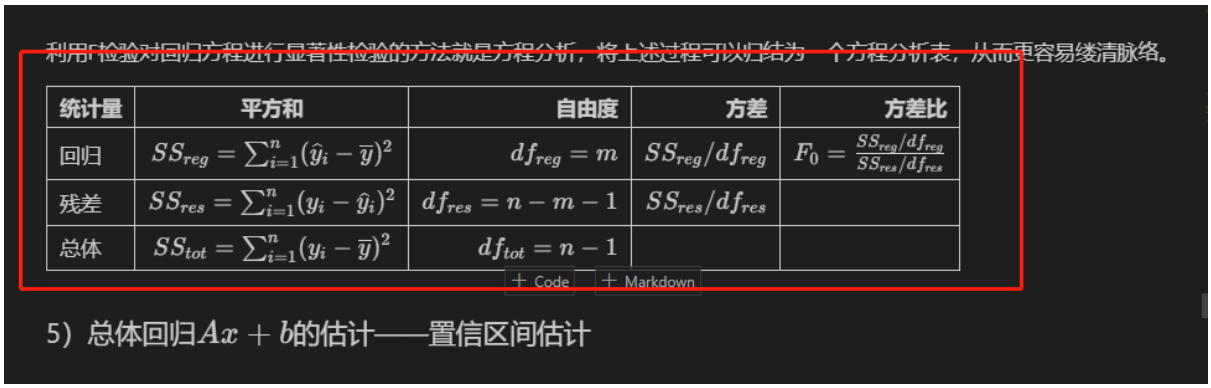
1. 若存在两个自变量x和y，如果$y = x + c$，其中c为常量，则其自由度为1，因为实际上只有x才能真正的自由变化，y会被x取值的不同所限制；
2. 估计总体的平均数μ时，由于样本中n个数都是相互独立的，任一个尚未抽出的数都不受已抽出任何数值的影响，所以自由度为n；
3. 估计总体的方差σ^2时所使用的统计量是样本的方差s^2，而s^2必须用到样本平均数$\bar{x}$来计算，$\bar{x}$在抽样完成后已确定，所以大小为n的样本中只要$n-1$个数确定，第n个数就只有一个能使样本符合$\bar{x}$的数值。也就是说，样本中只有$n-1$个数可以自由变化，只要确定了这$n-1$个数，方差也就确定了。这里，平均数$\bar{x}$就相当于一个限制条件，由于加了这个限制条件，样本方差s^2的自由度为$n-1$；
4. 统计模型的自由度等于可自由取值的自变量的个数。如在回归方程中，如果共有p个参数需要估计，则其中包括了$p-1$个自变量（与截距对应的自变量是常量），因此该回归方程的自由度为$p-1$。



> 自由度的再讨论（参考Wikipedia）
在统计学中，自由度（defree of freedom，df）是指当以样本的统计量估计总体的参数时，样本中独立或能自由变化的数据的个数，称为该统计量的自由度。范例：
> 1. 若存在两个自变量x和y，如果$y=x+c$，其中c为常量，则其自由度为1，因为实际上只有x才能真正的自由变化，y会被x取值的不同所限制；
2. 估计总体的平均数μ时，由于样本中n个数都是相互独立的，任一个尚未抽出的数都不受已抽出任何数值的影响，所以自由度为n；
3. 估计总体的方差$ \sigma ^{2} $时所使用的统计量是样本的方差$s^{2} $，而$s^{2} $必须用到样本平均数$\overline{x} $来计算，$\overline{x} $在抽样完成后已确定，所以大小为$n$的样本中只要$n-1$个数确定，第$n$个数就只有一个能使样本符合$\overline{x} $的数值。也就是说，样本中只有$n-1$个数可以自由变化，只要确定了这$n-1$个数，方差也就确定了。这里，平均数$\overline{x} $就相当于一个限制条件，由于加了这个限制条件，样本方差$s^{2} $的自由度为$n-1$；
4. 统计模型的自由度等于可自由取值的自变量的个数。如在回归方程中，如果共有p个参数需要估计，则其中包括了$p-1$个自变量（与截距对应的自变量是常量），因此该回归方程的自由度为$p-1$。

此处格式有误，数字前面应该全部添加“>”
且 1 中提及的“自变量”，应该改为“变量”。
后面公式可以看出，y 由 x 和 c 的值确定，此处，不适宜用自变量，

10. F 检验最后总结的方程分析表，可添加，对 m，n 的定义

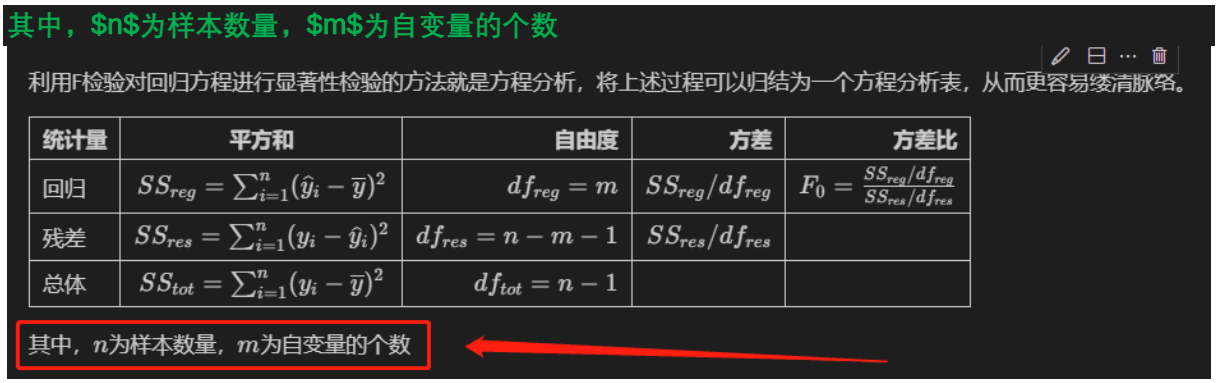


利用F检验对回归方程进行显著性检验的方法就是方程分析，将上述过程可以归结为一个方程分析表，从而更容易缕清脉络。

统计量	平方和	自由度	方差	方差比
回归	$SS_{reg}=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$	$df_{reg}=m$	SS_{reg}/df_{reg}	$F_0=\frac{SS_{reg}/df_{reg}}{SS_{res}/df_{res}}$
残差	$SS_{res}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$df_{res}=n-m-1$	SS_{res}/df_{res}	
总体	$SS_{tot}=\sum_{i=1}^{n}(y_i-\overline{y})^2$	$df_{tot}=n-1$		

+ Code + Markdown

5）总体回归$Ax+b$的估计——置信区间估计



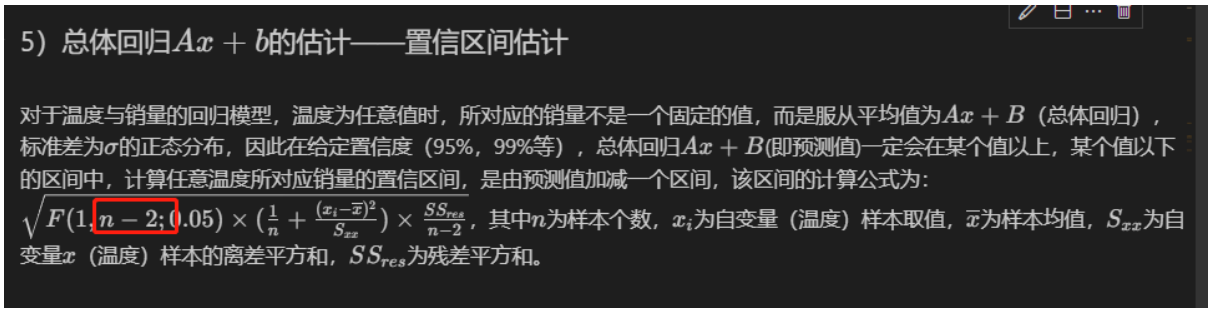
其中，n为样本数量，m为自变量的个数

利用F检验对回归方程进行显著性检验的方法就是方程分析，将上述过程可以归结为一个方程分析表，从而更容易缕清脉络。

统计量	平方和	自由度	方差	方差比
回归	$SS_{reg}=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$	$df_{reg}=m$	SS_{reg}/df_{reg}	$F_0=\frac{SS_{reg}/df_{reg}}{SS_{res}/df_{res}}$
残差	$SS_{res}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$df_{res}=n-m-1$	SS_{res}/df_{res}	
总体	$SS_{tot}=\sum_{i=1}^{n}(y_i-\overline{y})^2$	$df_{tot}=n-1$		

其中，n为样本数量，m为自变量的个数

11. 5）总体回归 Ax+b 的估计——置信区间估计
此部分出现的置信区间公式建议进行修改

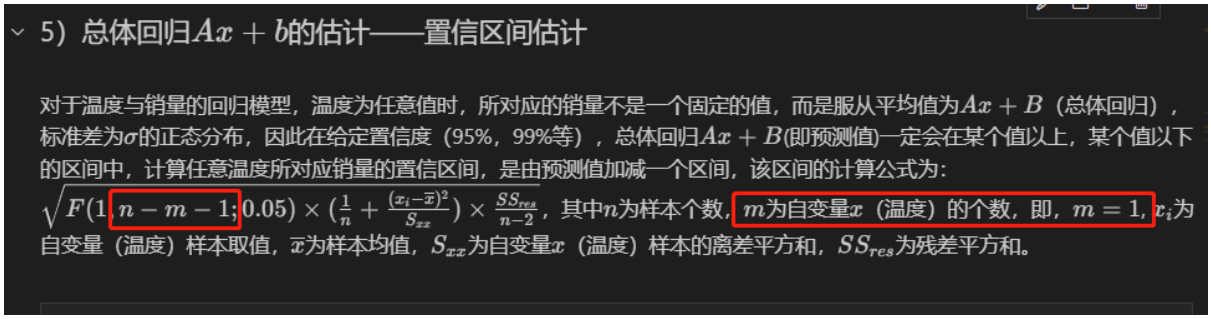


5）总体回归$Ax+b$的估计——置信区间估计

对于温度与销量的回归模型，温度为任意值时，所对应的销量不是一个固定的值，而是服从平均值为$Ax+B$（总体回归），标准差为σ的正态分布，因此在给定置信度（95%，99%等），总体回归$Ax+B$(即预测值)一定会在某个值以上，某个值以下的区间中，计算任意温度所对应销量的置信区间，是由预测值加减一个区间，该区间的计算公式为：

$\sqrt{F(1,n-2;0.05)\times(\frac{1}{n}+\frac{(x_i-\overline{x})^2}{S_{xx}})\times\frac{SS_{res}}{n-2}}$，其中$n$为样本个数，$x_i$为自变量（温度）样本取值，$\overline{x}$为样本均值，$S_{xx}$为自变量$x$（温度）样本的离差平方和，$SS_{res}$为残差平方和。

“n-2”应改为“n-m-1”，并添加对 m 的解释，即，“**m为自变量x（温度）的个数，即，$m=1$**”，



5）总体回归$Ax+b$的估计——置信区间估计

对于温度与销量的回归模型，温度为任意值时，所对应的销量不是一个固定的值，而是服从平均值为$Ax+B$（总体回归），标准差为σ的正态分布，因此在给定置信度（95%，99%等），总体回归$Ax+B$(即预测值)一定会在某个值以上，某个值以下的区间中，计算任意温度所对应销量的置信区间，是由预测值加减一个区间，该区间的计算公式为：

$\sqrt{F(1,n-m-1;0.05)\times(\frac{1}{n}+\frac{(x_i-\overline{x})^2}{S_{xx}})\times\frac{SS_{res}}{n-2}}$，其中$n$为样本个数，$m$为自变量$x$（温度）的个数，即，$m=1$，$x_i$为自变量（温度）样本取值，$\overline{x}$为样本均值，$S_{xx}$为自变量$x$（温度）样本的离差平方和，$SS_{res}$为残差平方和。

12.