

回归分析报告

本文以勐腊县、勐仑县、勐海县和景洪县从 1961 年至 2000 年 12 个月的降雨量及气温数据为数据集。其中，

Mla=勐腊县；Mlun=勐仑县； Mhai=勐海县； Jhong=景洪县.

R=降雨量； AT=气温； 01, 02, ..., 12 表示十二个月份。

一. 分析

1. 一个地区某一个月的降雨量与气温之间的关系；
2. 一个地区的平均降雨量与平均气温之间的关系；
3. 一个地区的降雨量与相邻地区降雨量之间的关系；
4. 一个地区的气温与相邻地区气温之间的关系；

本文从线性回归的角度出发，考虑建立一元线性回归模型和多元线性回归模型，并利用最小二乘法进行估计，同时对回归系数和回归方程进行显著性检验。

二. 模型的建立与求解

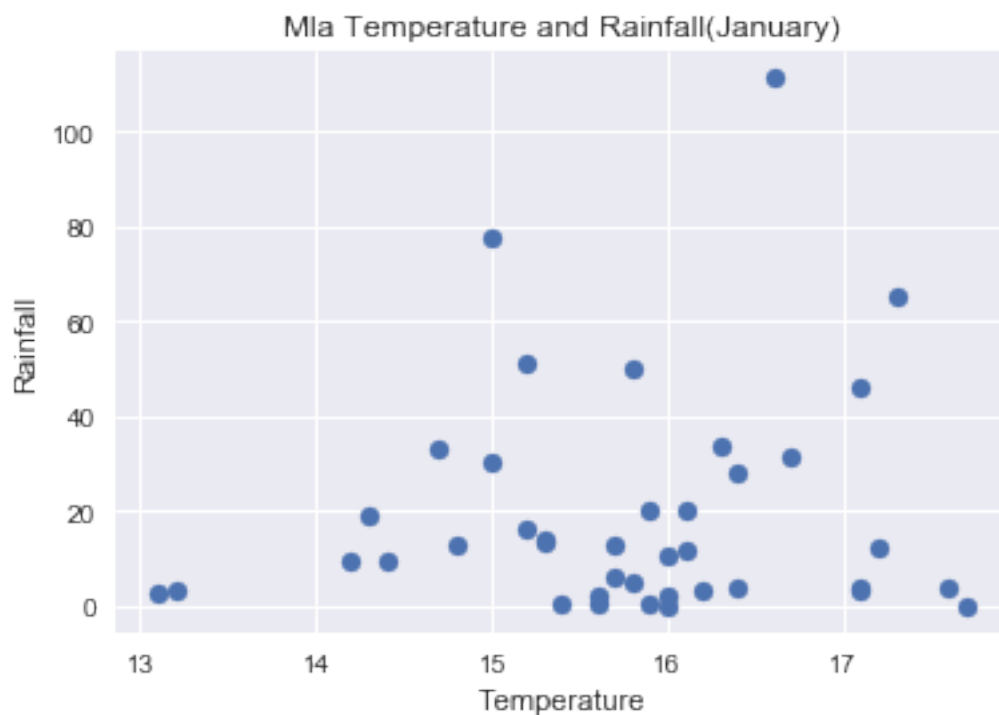
模型 1:

以勐腊县所有 1 月份的降雨量及气温数据为例，记气温为自变量 X ，记降雨量为因变量 Y ，假设基本模型为线性模型：

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

其中 $\varepsilon \sim N(0, \sigma^2)$ ， β_0, β_1 和 σ^2 是未知参数.

首先作 Y 对 X 的散点图：



OLS Regression Results

Dep. Variable:	Y	R-squared:	0.013
Model:	OLS	Adj. R-squared:	-0.013
Method:	Least Squares	F-statistic:	0.4916
Date:	Wed, 18 Oct 2017	Prob (F-statistic):	0.487
Time:	16:20:35	Log-Likelihood:	-183.16
No. Observations:	40	AIC:	370.3
Df Residuals:	38	BIC:	373.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-20.4015	57.126	-0.357	0.723	-136.047	95.244
X	2.5334	3.613	0.701	0.487	-4.781	9.848

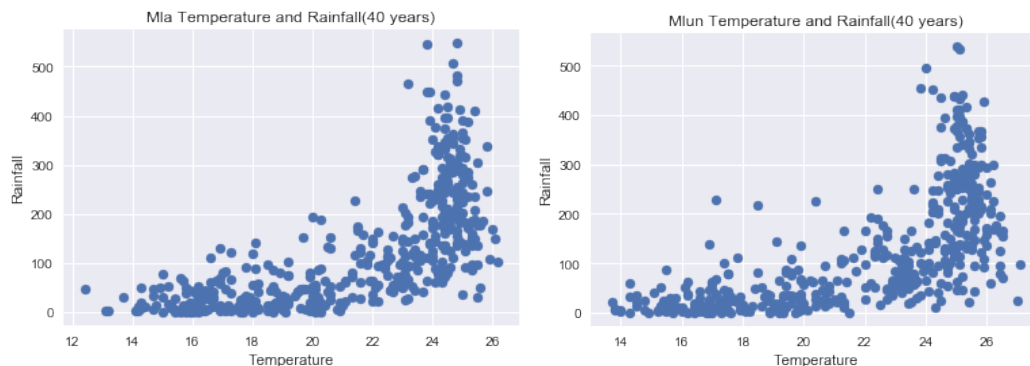
由最小二乘法的计算结果得到

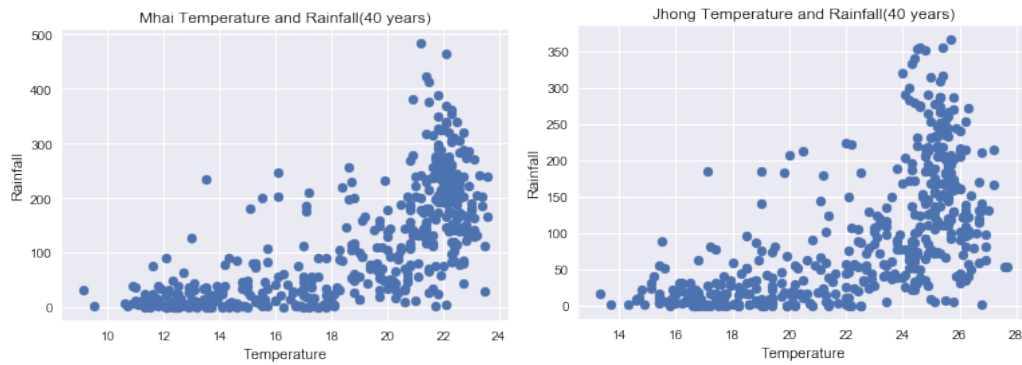
$$\hat{\beta}_0 = -20.4015, \hat{\beta}_1 = 2.5334, sd(\hat{\beta}_0) = 57.126, sd(\hat{\beta}_1) = 3.613$$

但对应于两个系数的 P 值均大于 0.1，说明该模型未能通过 t 检验和 F 检验。另外，从散点图上看，X 和 Y 之间并不构成线性关系，所以用一元线性模型模拟效果并不好。

模型 2:

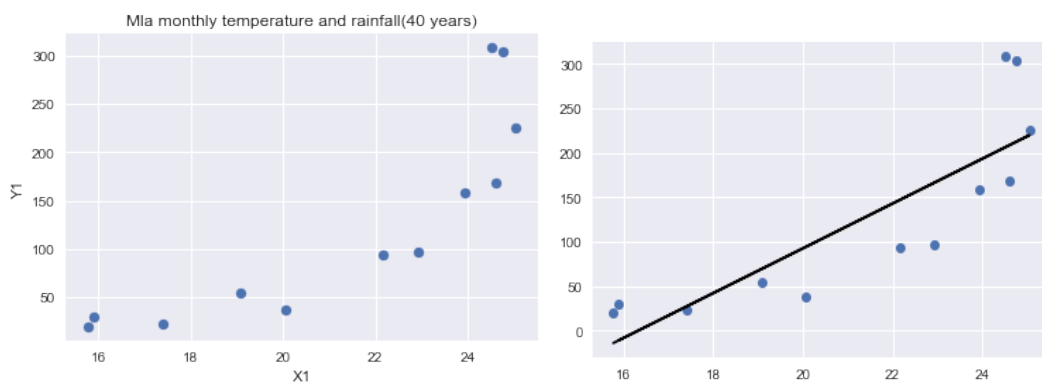
作四个地区 40 年的平均降雨量与平均气温的散点图，看出降雨量与气温之间并不构成线性关系，散点图如下：





模型 3:

考虑勐腊县这四十年的平均降雨量与气温的关系，作出散点图，并尝试用一条回归直线来拟合样本点：



从散点图上看，这些点大多数没有落在一条直线上，所以线性回归分析显然没有意义。可能用二次曲线拟合效果更好（待研究）。

既然一元线性关系都不满足，那么接下来不妨考虑一下多变量的情况：

模型 4:

以四个地区 1 月份的降雨量为例，记 Y 为勐腊县 1 月的降雨量， X_1 、 X_2 、 X_3 分别为勐仑县、勐海县和景洪县 1 月份的降雨量，做多元线性回归分析。

设变量 Y 与变量 X_1 ， X_2 ， X_3 间有线性关系

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon,$$

其中 $\varepsilon \sim N(0, \sigma^2)$ ， $\beta_0, \beta_1, \beta_2, \beta_3$ 和 σ^2 是未知参数。

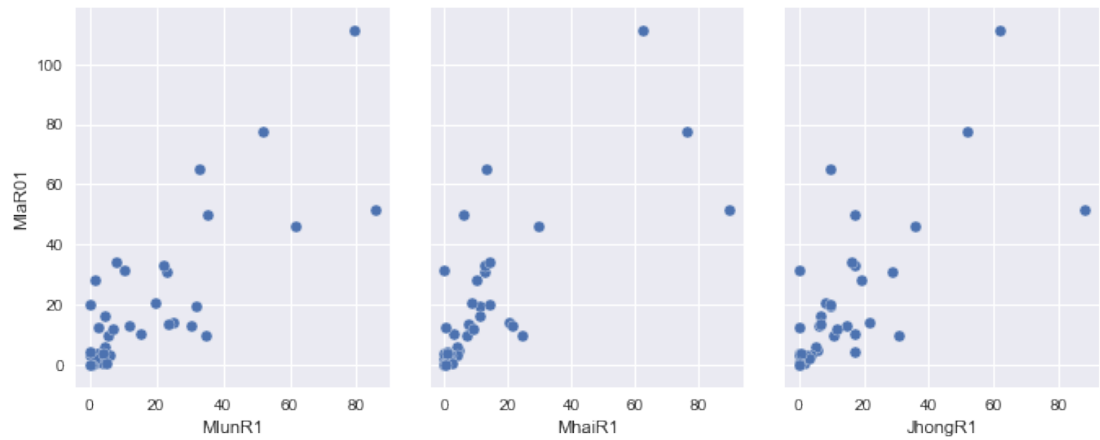
OLS Regression Results

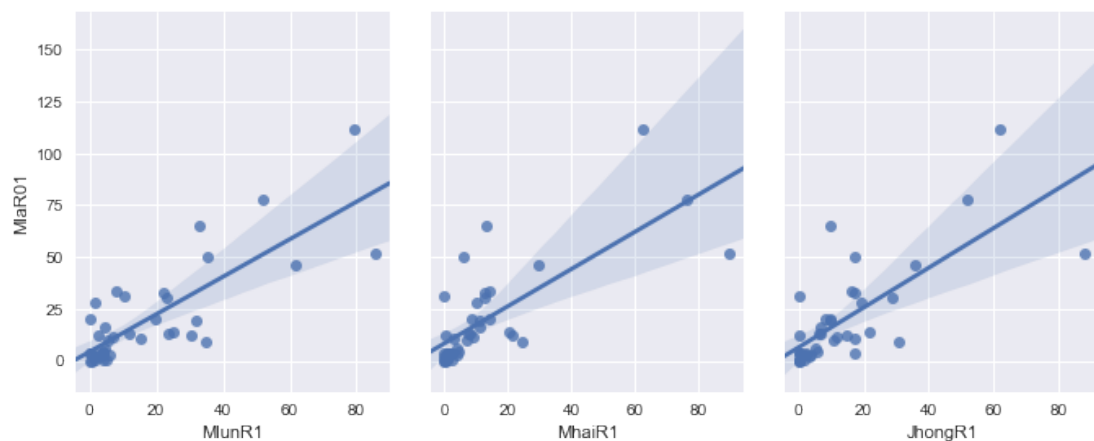
Dep. Variable:	MlaR01	R-squared:	0.810
Model:	OLS	Adj. R-squared:	0.780
Method:	Least Squares	F-statistic:	26.99
Date:	Wed, 18 Oct 2017	Prob (F-statistic):	4.63e-07
Time:	16:19:02	Log-Likelihood:	-90.163
No. Observations:	23	AIC:	188.3
Df Residuals:	19	BIC:	192.9
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	1.3148	3.922	0.335	0.741	-6.894	9.523
MlunR1	1.0535	0.304	3.466	0.003	0.417	1.690
MhaiR1	0.3423	0.362	0.944	0.357	-0.416	1.101
JhongR1	-0.0857	0.457	-0.187	0.853	-1.042	0.871

由最小二乘法的计算结果看出，在系数检验的环节，只有 X_1 的系数 β_1 通过了检验，其他系数都不显著；关于回归方程的检验，相关系数的平方 $R^2 = 0.810$ ，关于 F 分布的 P 值也非常显著。

以下是关于自变量与因变量的散点图，并添加了一条最佳拟合直线和 95% 的置信带：





结论:，该线性模型能够显著预测因变量的变化，但只有一个自变量能够显著预测因变量的变化，该模型还需进一步改进。

模型 5:

以四个地区 1 月份的气温为例，记 Y 为勐腊县 1 月的气温，X1、X2、X3 分别为勐仑县、勐海县和景洪县 1 月份的气温，做多元线性回归分析。

设变量 Y 与变量 X1，X2，X3 间有线性关系

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon,$$

其中 $\varepsilon \sim N(0, \sigma^2)$ ， $\beta_0, \beta_1, \beta_2, \beta_3$ 和 σ^2 是未知参数。

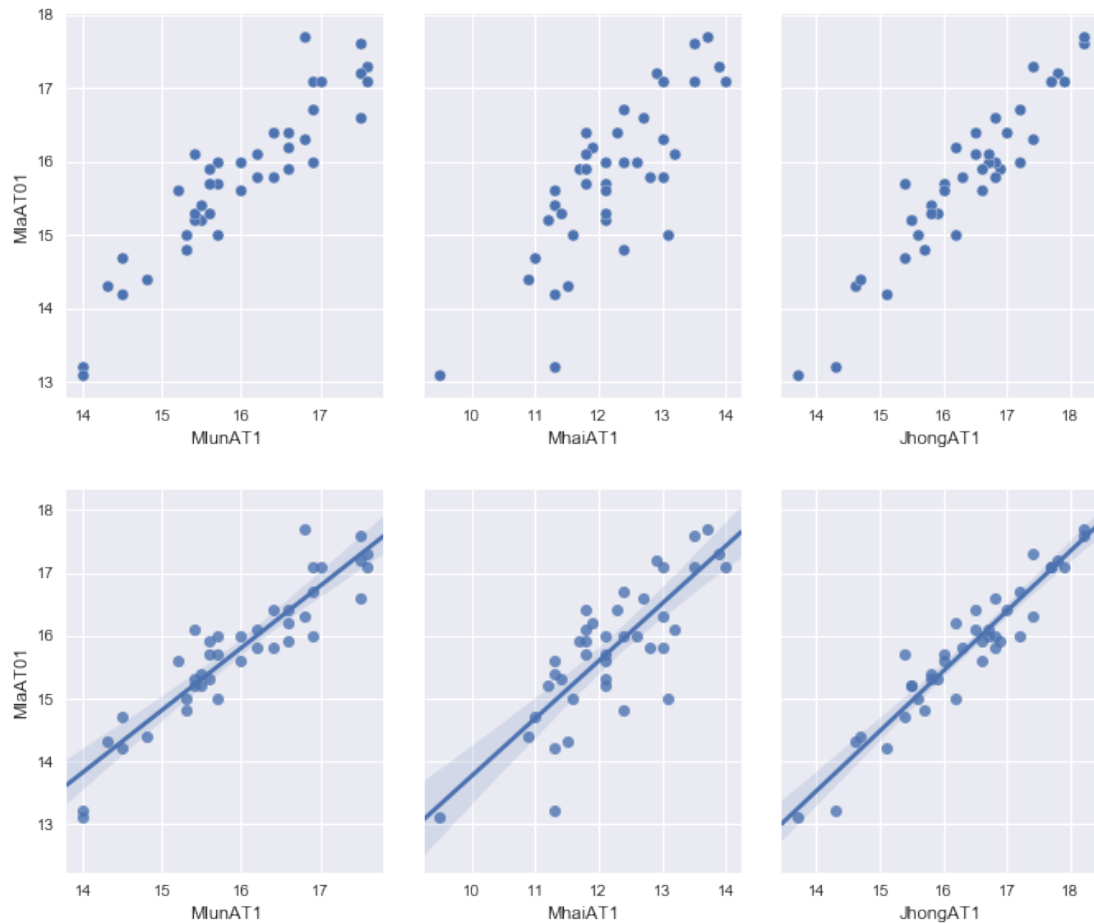
OLS Regression Results

Dep. Variable:	MlaAT01	R-squared:	0.909
Model:	OLS	Adj. R-squared:	0.895
Method:	Least Squares	F-statistic:	63.52
Date:	Wed, 18 Oct 2017	Prob (F-statistic):	4.32e-10
Time:	16:20:27	Log-Likelihood:	-6.1092
No. Observations:	23	AIC:	20.22
Df Residuals:	19	BIC:	24.76
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-0.8238	1.268	-0.650	0.524	-3.477	1.829
MIunAT1	0.4209	0.168	2.507	0.021	0.069	0.772
MhaiAT1	-0.0204	0.157	-0.131	0.898	-0.348	0.307
JhongAT1	0.6198	0.194	3.191	0.005	0.213	1.026

由最小二乘法的计算结果看出，在系数检验的环节，只有 β_1 和 β_3 显著，其他系数都不显著；在方程检验的环节，相关系数的平方 $R^2 = 0.909$ ，关于 F 分布的 P 值也非常显著。

以下是关于自变量与因变量的散点图，并添加了一条最佳拟合直线和 95% 的置信带：



结果表明：关于气温的多元线性回归方程比降雨量数据的拟合效果稍好，但该多元线性回归方程依然不是最适合的，需要考虑其他模型（如高斯过程模型，分层模型等）来拟合该数据集（待研究）。

附： [本文 python 代码](#)