中国各城市平均工资影响因素统计分析报告

吴若愚*，曲文琦*，谭昊*

上海财经大学 2020 届统计实验班

【摘　要】 本文以统计研究中的主流方法，对 2006 年中国各城市平均工资影响因素进行定量分析。首先采用描述性统计的方法，将人均工资和人均工资总产值等众多因子之间的联系可视化。之后采用回归分析手段构建了人均工资与自变量之间的回归模型，并建立在详实的数学理论基础上，阐述了假设检验、残差分析、多重共线性分析、选模型分析等检验拟合效果的方法的原理和流程；最后采用回归的结果对其他年份人均工资作出预测。

【关键词】 **平均工资，线性回归，残差诊断，多重共线性，选模型分析**

1　研究目的

平均工资是一项反映收入总体水平的指标，指企业、事业、机关单位的职工在一定时期内平均每人所得的货币工资额。本世纪以来，随着中国 GDP 的欣欣向荣，人均工资也水涨船高。但工资的分配具有不平衡、不充分的特点，如不同地区、不同职业、不同岗位的就业人员的工资待遇不平等，这导致了贫富差距、社会保障金过高等社会经济问题。

因此，在宏观经济的研究中，哪些因素对某个地区的人均工资具有显著的影响是一个非常重要的问题。由于考虑时间因素时，经济转型导致产业结构调整及收入再分配，导致不同年份的数据不易横向对比。所以我们考虑特定年份下不同地区的产业结构差异及综合的发达程度对当地平均工资影响，以此既可以微观上对企业管理者决定职员薪资时给出建议，也可以在宏观上给出某一地区的行业发展作长期规划。这也便于研究者在其他条件已知的情况下，用于预测未知的平均工资。

2　数据来源和相关说明

研究的理论基础在于如下的显著而深刻的论断：一个地区的总收入中总是有确定比例的收入被用于分发工资，因此一个地区的人均工资水平应当取决于当地的整体经济发展。计算 GDP 时，我们通常将产业分为第一、第二和第三产业，即农业、工业和服务业，不同产业的产值之和。而第三产业相较于一二产业又更加宽泛，涉及零售、金融、房地产、信息、物流、旅游、娱乐、教育、科研、卫生等领域。在刻画地区经济状况对该地区人均工资的影响时，我们认为地区在每个方面表现的好坏一起决定了经济的发展，因此我们从这些领域中各取一个指标作为自变量，而把平均工资作为因变量。注意到这些指标应该以“人均”为量纲，这是因为假设论断成立时，一个地区所有职工的收入总和应当和当地总的经济生产能力成正比，而总的经济生产能力在确定的经济水平下又与人口数成正比，因此平均工资应当和某种人均生产能力成正比。指标选取如下:

(1) 人均占有耕地面积：反映一个地区的农业水平，一个城市人均耕地越多，自然该城市偏向

于农业城市。注意平均工资未必和该指标正相关，因为耕地可能挤占了能用于开发附加值更高的产业的空间

(2) 人均工业总产值：将地方工业总产值除以人口数，反映一个地区的工业水平

(3) 人均固定资产投资额：固定资产是指企业为生产产品、提供劳务、出租或者经营管理而持有的、使用时间超过 12 个月的，价值达到一定标准的非货币性资产。其中很大一部分是房地产投资，房地产是我国的支柱产业之一

(4) 人均社会消费零售额：社会消费品零售总额划分为商品零售和餐饮收入两部分，能够反映商品贸易和餐饮零售的繁荣程度

(5) 人均学校数：我们将一个地区的学校总数除以地方人口总数得到人均学校数，这是一个二级指标，反映了该地区的教育水平

(6) 人均货运量：总货运量除以人口总数，能够综合地反映当地进出口业务的多少，交通航运能力等

(7) 失业率百分数：我们认为就业率和平均工资也存在一定的相关性，但正负性未知。这是因为一方面经济状况不景气时失业率较高，而薪水不会提高；但另一方面，根据微观经济学的理论，政府设定最低工资会导致失业率提高

(8) 人均机构存款量：存款是比较常规的银行业务，人均机构存款量反映金融业的发达程度和当地人的富裕程度

(9) 人均客运量：即年客运人次除以人口总数。一个地区的客运主要由外来务工人员和游客组成

(10) 沿海与否：与上述的连续性变量不同的是，这是一个 0-1 型离散变量。由于沿海城市开放程度高，利于旅游、外贸等发展，我们猜想沿海城市可能薪酬会高于内陆城市

由于相关政策，近年的相关数据不予公布，我们对较早的数据进行分析，但可以在有数据支持的情况下较容易地推广到最新的数据。我们分别选取了 2005 年、2006 年的各 286 个城市的有关数据，其中 2006 年的 286 个城市作为训练样本用于建立模型；而 2005 年由于和 2006 年紧邻，我们认为回归系数大致相同，将 2005 年 286 个城市作为用于检验预测精度。我们的数据来自于 eps 数据平台。

3 描述性分析

为了获得对数据的整体了解，我们对数据进行简单的描述性分析。描述性分析分为两个部分，其一是描述样本指标的数字特征，其二是分别建立指标之间的联系。

3.1 变量数字特征

我们用均值、中位数、最大值、最小值这四个数字特征描述样本变量的特点。

变量名	均值	最小值	中位数	最大值
人均工资	17717	7378	16317	41189
人均占有耕地面积	1.4484	0.0000	1.0300	11.9600
人均客运量	18.997	2.918	13.280	285.830
人均机构存款量	24698	3351	12764	484704
人均工业总产值	28589.7	698.4	13739.0	606035.7
人均固定资产投资额	9103	1180	6152	64709
人均社会消费零售额	6087.7	798.4	4221.0	84910.5
人均学校数	0.6180	0.3128	0.5982	1.4075
人均货运量	20.449	2.354	13.259	215.170
失业率百分数	0.6543	0.0000	0.4854	5.6038

图 1 变量数字特征

人均工资介于 7378 与 41189 之间，均值为 17717，中位数为 16317。这说明有超过一半的地区人均工资低于平均水平，表明我国目前贫富差距仍然较大。将城市按工资从高到低排序，发现北上广等大城市普遍较高，而平均工资较低的多为西部省市。此外，我们发现相邻城市的工资具有粘性，即大小接近、变化趋势相同。这启示我们进一步将地理位置相近的城市归并为一类比较。

值得注意的是，人均工业总产值的取值范围很大（从 698.4 到 606035.7），这表明不同地区的第二产业发展水平差距较大。从人均社会消费零售额的均值 6087.7 和较大的取值范围（从 798.4 到 84910.5）可以看出，不同地区的消费水平与消费能力差距较大。从人均学校数的均值（0.6180）和较小的取值范围可以看出，我国的教育普及程度较好，但仍有部分地区存在教育能力低下的问题。从失业率百分数的均值和中位数可以看出，我国失业率不高，但仍存在部分地区就业难赚钱难的问题。人均客运量和货运量的最大值远远高于均值，说明我国的客运和货运呈现高度中心化的特点。人均固定资产投资额的取值范围（从 1180 到

64709）也很大，说明房价在不同地区差异也很大。

将每个城市当作一个样本，画出人均工资的频数分布直方图。

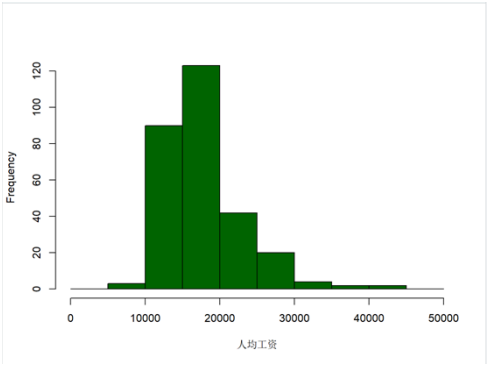

图 2 人均工资频数分布直方图

从图中可以判断出，人均工资的分布中间高两头低，但并不服从正态分布，左边厚尾右边长尾。多数地区的人平均工资集中在 10000 到 30000 之间。但对于经济学中的很多指标，我们经常考虑其对数分布。并且取对数能够将分布曲线左边的尾加长而右边的尾加厚，这可能使得我们的频数分布直方图更变得接近与正态。因此我们考虑将人均工资取对数并可视化。

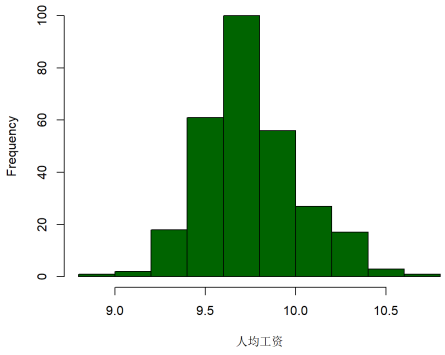

图 3 对数人均工资频数分布直方图

此时，由图 3 可知不同地区人均工资确实服从对数正态分布。

3.2 因变量和自变量的联系

当考虑所处地区导致的平均工资差异这样的属性数据时，用箱型图是直观且高效的。箱型图的“盒子”上沿表示 75% 分位数，下沿表示 25% 分位数。“盒子”中间的横线表示中位数。之所以采用中位数而不是均值，是因为中位数对异常值不敏感。此外，盒外上下的两根横线表示正态假设下取值的一个合理范围。若一个数据落在两根横线所夹的范围外，原因一可能是该数据是离群的异常值，也可能是因为数据不符合正态假设。按照地理位置，我们将中国城市划分为华东地区、华北地区、东北地区、西北地区、中南地区、西南地区六大地区，用箱型图反应平均工资的地区特点。

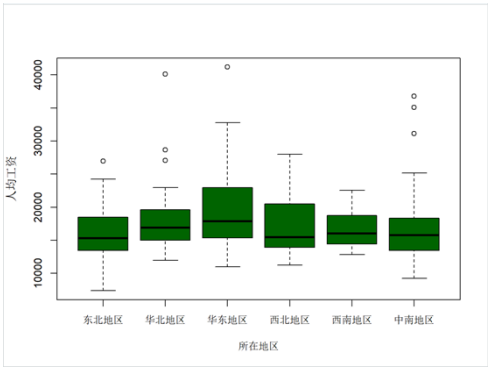

图 4 人均工资关于所在地区箱型图

根据图 4 结果，我们注意到华北、东北、西北、西南、中南五个地区平均工资水平较为接近，但华东地区显著高于这五个地区。这启发我们寻找华东地区的特殊性。于是我们猜想一个城市的人均工资水平和是否沿海具有一定关系。

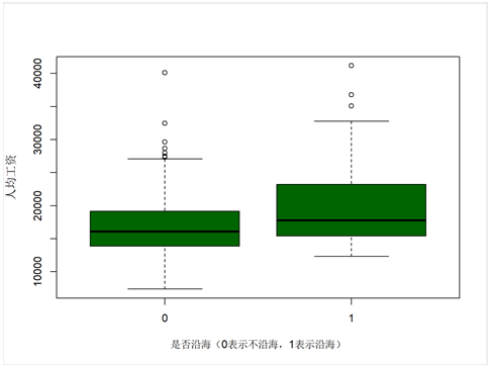

图 5 人均工资关于沿海与否箱型图

沿海地区的人均工资确实明显高于非沿海地区，这可能是由于靠近海边运输业发达，贸易交流更为频繁致使经济发达，人均工资高。

现在，我们绘制人均工资关于平均工业总产值等自变量的散点图。

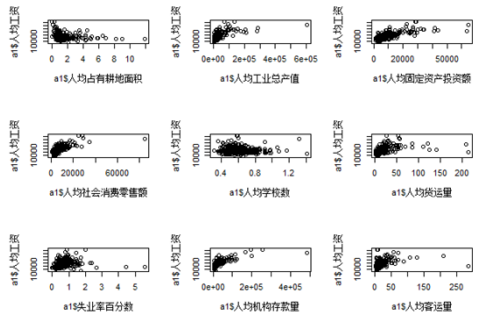

图 6 人均工资关于回归因变量散点图

由图 6，人均工资与人均社会消费零售额、人均货运量、人均机构存款量、人均固定投资额、人均工业总产值和人均客运量存在明显的近似线性相关性，且均为正相关。而人均工资与失业率百分数、人均学校数、人均占有耕地面积之间的也存在一定的线性相关性，但相关性比较弱，其中工资与学校数正相关，与失业率、耕地面积负相关。这些近似线性关系支持我们使用线性回归，而关联正负性符合我们在选择指标时的初步认识。

4 回归分析

4.1 模型建立

我们用矩阵形式表述线性模型：

$$y = X\beta + \epsilon \tag{1}$$

这里，y 是因变量的观测向量，X 是 n*p 的设计矩阵，其中第一列全是 1，β 是未知系数向量。ϵ 是随机扰动向量即残差项，满足 Gauss-Markov 假设，即：

$$E(\epsilon) = 0, \qquad Cov(\epsilon) = \sigma^2 I$$

绝大多数情况下，我们还假设残差项服从正态假设，即：

$$\epsilon \sim N(0, \sigma^2 I) \tag{2}$$

为了估计未知系数向量 β，线性模型参数估计的核心思想是最小二乘法，即使得误差向量 $\epsilon = y - X\beta$ 的度量取到最小，也就是最小化：

$$Q(\beta) = \|\epsilon\|^2 = (y - X\beta)'(y - X\beta)$$

给出 β 估计值 $\hat{\beta}$:

$$\hat{\beta} = (X'X)^{-1}X'y \tag{3}$$

进一步的：

$$\hat{\epsilon} = (I - P_X)y = (I - X(X'X)^{-1}X')y$$

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{n - p - 1}$$

这里 n 是观测的次数，即设计矩阵的行数；p 是自变量的个数，即设计矩阵的列数。这些估计都是无偏的。以下的 Gauss-Markov 定理表面最小二乘估计是好的估计。

定理 1 (Gauss-Markov 定理)**.** $\hat{\beta}$ 是 β 唯一的最佳线性无偏估计。此外，$\hat{\beta}$ 与 $\hat{\sigma}^2$ 相互独立。在正态假设时 $\hat{\beta}$ 还是唯一的最小方差无偏估计

在我们的问题中。因变量是人均工资，而自变量是其余九个因素。n 为城市的个数。利用 R 语言对线性模型进行回归。

表 1 原模型回归结果

变量名	系数估计值	标准差	t 检验 p 值
常数项	1.646e+04	9.500e+02	0.00
人均占有耕地面积	-2.891e+02	1.487e+02	0.05
人均工业总产值	-4.297e-02	9.914e-03	0.00
人均固定资产投资额	3.737e-01	4.840e-02	0.00
人均社会消费零售额	-1.236e-01	1.023e-01	0.23
人均学校数	-4.213e+03	1.385e+03	0.00
人均货运量	1.591e+01	1.053e+01	0.13
失业率百分数	1.615e+02	3.688e+02	0.66
人均机构存款量	8.864e-02	1.622e-02	0.00
人均客运量	8.169e+00	1.002e+01	0.42
残差项标准差 3095		F 检验 p 值 <0.0001	
R^2 0.6475		调整后的 R^2 0.6351	

我们希望的回归的拟合优度作定量的判断, 即判断因变量与自变量线性关系的强弱，于是引入判决系数 R^2:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y_i - \hat{y}_i))}{\sum(y_i - \bar{y})}$$

显然 R^2 是一个小于 1 大于 0 的数，且这个值越大证明拟合效果越好。但是问题在于自变量增多时 R^2 一定会变大，导致我们为了增加拟合优度引入众多次要的变量。因此需要增加一个惩罚因子，使得自变量在过于“次要”的时候适得其反，

这时引入调整后的 R^2：

$$R_d^2 = 1-\frac{(n-1)SSE}{(n-p-1)SST} = 1-\frac{(n-1)\sum(y_i-\hat{y_i}))}{(n-p-1)\sum(y_i-\bar{y})}$$

由上表，在本题中 R^2 为 0.6475，调整后的 R^2 为 0.6351，拟合优度良好，证明使用线性回归的策略切实可行。但我们注意到回归系数的大小不能反映自变量影响的强弱，这是因为我们没有经过标准中心化来统一量纲。

4.2 假设检验

尽管我们已经得到了模型的回归结果，但我们并不能确定这种拟合足够好，原因在于不能确定每种自变量影响的显著性。当我们计算得到一些自变量回归系数过小时，存在这样一种可能性，就是实际上，这些因素对平均工资实际上没有影响，真正的回归系数应当是 0，而导致计算出回归系数不为 0 的原因是随机噪音 ϵ。为了避免这种问题，我们进行假设检验：

$$H_0 : \beta_i = 0 \quad vs \quad H_1 : \beta_i \neq 0 \tag{4}$$

为了刻画两种情况的本质不同，我们先引入约束最小二乘估计，将上面的形式推广到一般的情况。这是指在满足式 (1) 的线性模型中，增加条件对 β 的限制条件：

$$H\beta = d$$

这里，H 是一个 k*p 的矩阵，每一行都描述了一个自变量间的线性关系条件。显然，“对于任意 j，$B_j = 0$” 是 $H\beta = 0$ 的特例，这时取 H 为单位阵即可。

未知系数向量的约束最小二乘估计 $\hat{\beta_H}$ 是：

$$\hat{\beta_H} = \hat{\beta} - (X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}(H\hat{B} - d)$$

这也是一个无偏估计。

定理 2. *若* $H\beta = 0, \quad F = \frac{(SSHE-SSE)/m}{SSE/(n-p)} \sim F_{m,n-p}$

基于这一定理，我们引入 F 检验，对给定的水平 α，当 $F > F_{m,n-p}(\alpha)$ 时拒绝原假设 H_0，反之接受之。但是 F 检验在处理特殊的问题时，比如式 (4) 的假设检验，相对不便捷。因此，我们根据下面的定理引入 t 检验。

定理 3. *正态假设成立时,* $t_i = \frac{\hat{\beta_i}}{\sqrt{c_{ii}}\hat{\sigma}} \sim t_{n-p}$

因此，在 $|t_i| > t_{n-p}(\alpha/2)$ 时，可以拒绝原假设 H_0。当然，对偶地，我们也可以使用检验的 p 值来衡量是否应该接受原假设。p 值很小说明发生了极端事件，那么应该拒绝假设。值得一提的是，t 检验在样本量比较小的时候表现较好，适合我们的课题。

本回归模型中，我们通过考察 t 检验的 p 值，刻画因变量与各自变量相关性的强弱。在 0.1 的置信水平下，我们断言人均工资和人均固定资产投资额、人均机构存款量有着比较显著的正相关关系，而与人均耕地面积、人均工业总产值和人均学校数有着相对显著的负相关关系。与其他自变量的关系较弱。我们认为原因可能在于以下两点：第一，第三产业的附加值高于第一产业和第二产业，比较发达的地区产业重心会倾向于第三产业，如金融业，这些行业就业者工资比较高，因此农业和工业城市在影响职工工资时处于不利位置：第二，实际上根据散点图平均工资和学校数是正相关的，但在变量的联合作用下学校数却有负系数，这可能是因为我国教育普及度高，不同城市人间学校数的分异弱于平均工资的差异，而平均工资差异又弱于机构存款量等的差异，学校数可能需要一个负系数去“中和”两种差异。

4.3 回归诊断

在我们之前的讨论中，其实默认了所有条件都是理想的从而忽略了很多隐患。一般来说，这种隐患存在于三种情况。第一，我们对模型关于误差项作了一些假设，如 Gauss-Markov 假设或者正态假设，但这种假设实际经不起推敲。第二，样本本身中本身存在一些强影响点或者离群异常值，其可能来自于统计数据时的人为失误，但是极大地干扰了正常的回归结果。第三，自变量之间具有多重共线性，即自变量的线性组合恒为 0，这表明自变量的选择重复。我们分别就这三个问题进行诊断。

4.3.1 残差诊断

在我们的研究中，对残差 ϵ 作出了正态假设，即式 (2)。如何验证假设的正确与否？我们有如下定理。

定理 4. *(1)* $E(\hat{\epsilon}) = 0$

(2) $Cov(\hat{\epsilon}) = \sigma^2(I - P_X)$

(3) $Cov(\hat{y}, \hat{\epsilon}) = 0$

(4) 若 $\epsilon \sim N(0, \sigma^2 I)$，则 $\hat{\epsilon} \sim N(0, \sigma^2(I - P_X))$

根据如上定理，我们在正态假设下定义学生化残差 r_i：

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - p_{ii}}}$$

r_i 的分布比较复杂，且彼此不独立，但容易证明 r_i 的渐进分布是标准正态分布，且近似“独立”。这一观点指使我们绘制残差图。当图示的结果不太好时，我们有时会采用 Box-Cox 变换对自变量预先加工。

以我们的课题为例，我们引入两类残差图——RL 图和 RF 图。首先利用 RL 图对我们的模型进行残差分析。

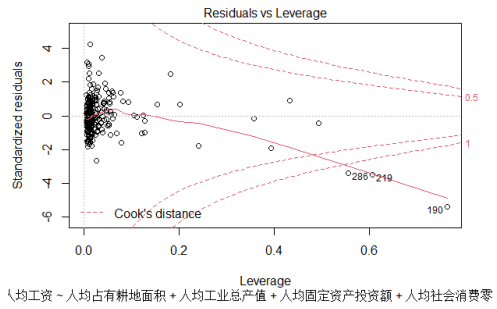

图 7 原模型 RL 图

由于在正态假设下，r_i 能够被视为来自标准正态分布的简单随机样本。根据正态分布的性质，应当有 95% 的 r_i 落在 [-2,2] 中，即宽度为 4 的无限长水平带中。其次，由于 r_i 和 $\hat{y}_i$ 无关，r_i 不应该表现出关于 $\hat{y}_i$ 变大或变小的趋势，这说明等方差性成立。模型的 RL 图确实能够满足这两个性质，检验成功通过。

RF 图本质与 RL 图有很大相似性，但读法有一定差异。

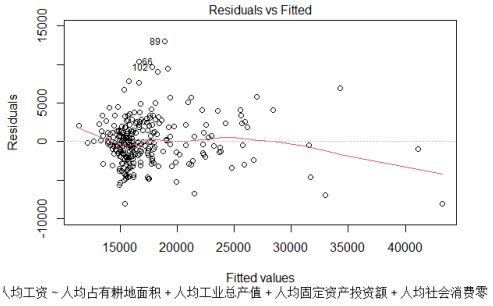

图 8 原模型 RF 图

我们注意到 RF 图和 RL 图尽管纵坐标都是残差，但后者使用的是学生化残差，而前者只需要一般的残差就可以。水平上来看，点的分布杂乱无章，与正态性相洽。从中间直线来看，假如所有样本残差刚好是服从正态分布的，那该直线应当是一条水平的直线。“直”是因为线性相关，假如是“微笑曲线”或“倒微笑曲线”，那么实际上因变量和自变量的相关性就不是线性的了，需要对自变量变换处理。水平是因为 Gauss-Markov 假设下的常方差性。从我们的图中可以看出，左边点很密集而右边点稀疏，直线在左边平坦，而在右边有下倾的趋势。这反应正态假设基本能成立。右边向下倾斜是因为点太少因而离群点的干扰过大，不必过多在意模型整体的稳健性，但是离群点值得我们格外注意。

位置尺度图与 RF 图大同小异，唯一的区别是纵坐标用学生化残差的平方根代替了残差。位置尺度图和后者的性质是几乎完全相同的，并且也需要保证绝大多数点学生化残差平方根不太大就可以。

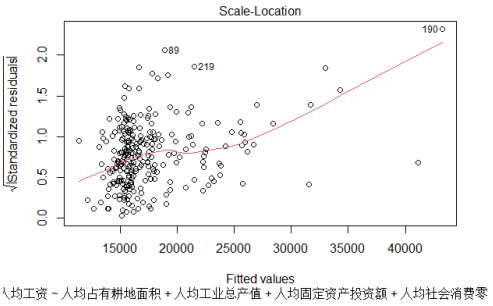

图 9 原模型位置尺度图

最后我们想要使用一种图是 QQ 图。这种方法和前面三图有本质上的差异：前面的方法本质上

是参数估计的方法，而 QQ 图是非参数的方法。因此，某种意义上来说，QQ 图适用的范围更加广泛。QQ 图的思想来源是很直观的：假如两个随机变量分布的分位数相同，那么我们自然可以认为该两个随机变量相同。特别的，假如一个随机变量和某个正态分布具有相同分位数，那么我们就可以证明它是正态分布。这也是正规 QQ 图的来历。我们更加严谨地阐释 QQ 图的原理。

在 $\epsilon_i \sim N(\mu, \sigma^2)$ 的时候，设 q_α 是 ϵ_i 的 α 分位数，即：

$$P(\epsilon_i < q_\alpha) = \alpha$$

则有:

$$q_\alpha = z_\alpha \sigma + \mu$$

其中 z_α 是标准正态分布 α 分位数。于是在正态假设下，ϵ_i 与 z_α 呈线性关系。我们对 $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ 从小到大排序得到次序统计量 $\epsilon_{(1)}, \epsilon_{(2)}, \dots, \epsilon_{(n)}$，并以 $\epsilon_{(1)}, \epsilon_{(2)}, \dots, \epsilon_{(n)}$ 为 X 轴，$z_{1/n}, z_{2/n}, \dots, z_{n/n}$ 为 y 轴作散点图。若散点图接近于直线，可以认为满足正态假设。

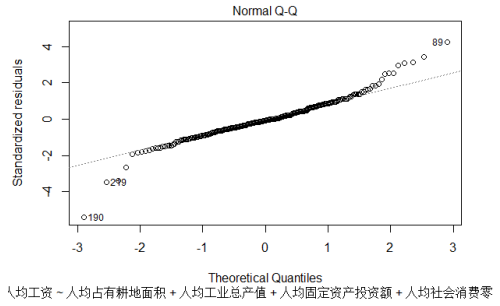

图 10 原模型正规 QQ 图

模型的 QQ 图接近直线，能够认为符合正态假设。

4.3.2 强影响点诊断

让我们回到样本本身。之前我们考虑的是样本的整体性质，即残差的正态假设，要么所有的样本都能够满足，要么所有的样本都不能。实际上，还有另外一种可能，就是一部分样本是好的，而另一部分不是。比如我们有一张表格记录了班级里所有学生的身高，想要据此对学生身高作统计分析。其中有一位同学被记录的身高和姚明相同。毫无疑问，这是一个极端的数字，且远远高于被记录的其他身高。这里有两种可能：其一，这个数字来源于记录时的错误；其二，确实有一位学生具有这个身高。然而无论在哪种情况下，这个数字都极大地影响了我们对样本全体的统计分析。此时的最优估计可能为了容纳这个极端样本而达到最优，但是和主流身高的分布有所偏离。因此这个身高对我们的统计分析产生了非常不利的影响，且这个影响相比于 1 米 8 这样常规偏高的身高来说要大得多。

我们用上面这个例子来解释强影响点。强影响点可能来自于离群值或者异常值。实际研究中我们最好将两者加以区分，通常需要反复核查并剔除或修改人为过失或条件失控导致的异常值，或者收集更多数据或者使用稳健统计的方法缩小离群值对估计的影响。但在像我们的课题这样样本量有限的黑箱模型下，这难以做到，我们只好退而求其次机械地删除强影响点。

现在，我们引入 Cook 距离的概念刻画强影响点。我们先引入一些记号。用 $y_{(i)}$，$X_{(i)}$ 和 $\epsilon_{(i)}$ 分别表示 y，X 和 ϵ 删除第 i 行后得到的矩阵或向量。此时剩余 n-1 组数据的线性回归模型为：

$$y_{(i)} = X_{(i)}\beta + \epsilon_{(i)}, \quad \epsilon \sim N(0, \sigma^2 I_{n-1})$$

此时 β 的最小二乘估计 $\hat{\beta_{(i)}} = (X'_{(i)}X_{(i)})^{-1}X'_{(i)}y_{(i)}$。显然，$\hat{\beta} - \hat{\beta_{(i)}}$ 反映了第 i 组数据对回归系数估计的影响。它是一个向量，我们考虑将其数量化。于是 Cook 统计量定义为：

$$D_i = \frac{(\hat{\beta} - \hat{\beta_{(i)}})' X' X (\hat{\beta} - \hat{\beta_{(i)}})}{p\hat{\sigma}^2}$$

统计量 D_i 的大小同样反映了第 i 组数据对回归系数估计的影响，其值越大，说明第 i 个样本影响力越强。但是这样的定义下的 D_i 计算比较麻烦。我们引入一个简便的形式。

定理 5. $D_i = \frac{1}{p}(\frac{p_{ii}}{1-p_{ii}})r_i^2$

这里 p_{ii} 是 P_X 的第 i 个主对角元，r_i 是学生化残差。一般来说，我们认为第 i 组数据使得 $D_i > 0.2$ 时，该数据是强影响点。于是，我们水到渠成地引入 Cook 距离图。横轴上均匀分布着所有点的编号，其纵坐标为对应点的 D_i。我们主要注意查看所有纵坐标大于 0.2 的点。

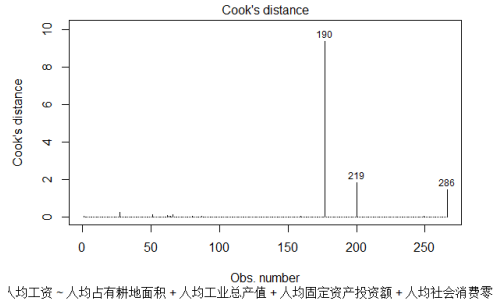

图 11 原模型 Cook 距离图

本题中，第 190,219,286 组样本的 Cook 距离远远大于 0.2，而其他样本组的 Cook 距离均小于 0.2。我们有足够的理由相信有且仅有这三组样本点是强影响点。删除后再次回归并计算其他点的 cook 距离，发现此时第 199 组样本成为了强影响点，进行删除。我们分析离群点的来源，分别是深圳、东莞、玉林、克拉玛依四座城市。我们猜想其成为离群点的原因可能在于：这几座城市产业结构比较特殊，我们选取的因变量指标不能涵盖当地人口的收入来源，使得这些城市在其他经济指标比较低的情况下人口收入较高。深圳处于发展的上升期，经济年增长率很高，整体已经处于全国领先位置，但人均工资的上升具有滞后性，不能跟上经济的发展；且使用了过多廉价劳动力，其工资不能与创造的价值达到平衡。

删除这四组样本后，我们利用剩下的城市，重新对人均工资进行线性回归并输出拟合结果。模型的判决系数更大了，证明删去强影响点后回归效果更好。如下图，再一次模型诊断。删除四个强影响点后，诊断图的趋势变好，残差图点的分布更加杂乱无章，点的大小更加可比；QQ 图更加接近一条直线；所有点的 Cook 距离不超过 0.2。残差诊断和强影响点诊断能够通过。我们用修正后的模型回归系数作为全模型最终的回归结果。

表 2 修正后的模型回归结果

变量名	系数估计值	标准差	t 检验 p 值
常数项	1.470e+04	9.258e+02	0.00
人均占有耕地面积	-1.373e+02	1.373e+02	0.32
人均工业总产值	5.739e-03	1.145e-02	0.62
人均固定资产投资额	1.821e-01	6.250e-02	0.00
人均社会消费零售额	-3.212e-02	1.284e-01	0.80
人均学校数	-2.314e+03	1.336e+03	0.08
人均货运量	2.292e+01	1.036e+01	0.03
失业率百分数	-3.278e+02	3.423e+02	0.34
人均机构存款量	9.622e-02	1.545e-02	0.00
人均客运量	3.713e+01	1.571e+01	0.02
残差项标准差 2786		F 检验 p 值 <0.0001	
R^2 0.6931		调整后的 R^2 0.6821	

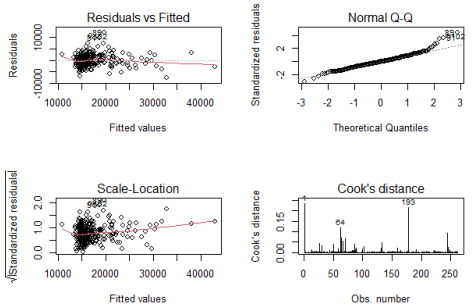

图 12 修正后的模型统计诊断图

4.3.3 多重共线性诊断

在线性回归的假设中，除了样本需要足够好之外，我们还应当正确地选取了回归的自变量。前者我们已经在前面的部分讨论了，现在我们来讨论一个问题：回归的自变量是不是越多越好。的确，乍一看自变量增多时 R^2 一定增加；那么其他方面我们是否会有损失。基于这个看法我们必须先引入全模型和选模型的概念。全模型是指我们使用所有的自变量做最小二乘估计，是我们上面写的通常形式 (1)。相对的，选模型是指我们去除了一部分自变量后使用剩余自变量对因变量做的最小二乘估计。我们从数学上考察全模型和选模型的差异。

记 $X = (X_q : X_t)$,$\beta' = (\beta'_q : \beta'_t)$，使用原设计矩阵前 q 列回归得到选模型：

$$y = X_q\beta_q + \epsilon_q, \quad \epsilon \sim N(0, \sigma^2 I)$$

选模型 β_q 的最小二乘估计为:

$$\tilde{\beta_q} = (X_q'X_q)^{-1}X_q'y$$

我们对全模型的估计 $\hat{\beta}$ 也做对应的分块 $\hat{\beta}' = (\hat{\beta_q}' : \hat{\beta_t}')$。我们用下面的定理，刻画全模型和选模型的关系:

定理 6. 当全模型成立时:

(1) $Cov(\hat{\beta_q}) - Cov(\tilde{\beta_q})$ 是正定阵

(2) 若 $Cov(\hat{\beta_t}) \geq \beta_t\beta_t'$，$MSEM(\hat{\beta_q}) - MSEM(\tilde{\beta_q})$ 是正定阵

(3) $R^2 \geq R_c^2$

(4) 定义全模型预测偏差 $z = x_0'\hat{\beta} - y_0$ 及选模型预测偏差 $z_q = x_{0q}'\hat{\beta_q} - y_0$，则 $Var(z) \geq Var(z_q)$

(5) 若 $Cov(\hat{\beta_t}) \geq \beta_t\beta_t'$, $MSEP(\hat{y_0}) - MSEP(\tilde{y_0}) \geq 0$

定理 6 反映了这样的事实: 即使全模型正确，且全模型的判决系数始终大于选模型，但选模型的方差始终要比全模型小。并且，当选模型去除的那一部分自变量对因变量的影响比较小时，选模型的精度高于全模型。因此存在这样一种可能性，使用全模型的拟合优度非常好，但是这让全模型的方差异常大。这就是多重共线性的后果之一，说明全模型中存在着本来不应该存在的自变量。

多重共线性在具有共同趋势的经济变量中是很常见的，它表明几个因变量之间存在近似线性关系。我们现在来说明为什么多重共线性会导致模型失真。首先定义估计量 $\hat{\theta}$ 的均方误差:

$$MSE(\hat{\theta}) = E|\hat{\theta} - \theta|^2$$

它度量了 $\hat{\theta}$ 偏离 θ 真值的程度，越小越好。进一步的，我们有:

定理 7. $MSE(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$

这里 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ 为 $X'X$ 所有的特征值。

这表明，假如 $X'X$ 有一个很小的特征值，估计的均方误差就会很大; 并且 $\hat{\beta}$ 的平均长度也比 β 真值的长度大得多，导致一些分量的绝对值更大。这时，模型的估计一定不是好的估计。与此同时，假设 ϕ 是 λ_p 对应的规范特征向量，则:

$$X'X\phi = \lambda_p\phi \approx 0$$

$$\phi'X'X\phi = \lambda_p\phi'\phi = \lambda \approx 0$$

$$X\phi \approx 0$$

这表明回归自变量近似线性相关，这也是多重共线性名字的由来。我们常用条件数 k 检验所有自变量整体多重共线性:

$$k = \frac{\lambda_1}{\lambda_p}$$

条件数是 $X'X$ 最大特征值和最小特征值的比值，反映了自变量整体多重共线性的严重程度。$k > 1000$ 时，认为存在严重的多重共线性。

进一步的，我们还想拷问，究竟哪些自变量是多重共线性的罪魁祸首。对于每个自变量，我们都能定义方差膨胀因子 VIF_i:

$$VIF_i = \frac{1}{1 - R_i^2}$$

其中，R_i^2 是把 x_i 当作因变量，其他因子当作自变量进行线性回归而得到的判决系数。VIF 很好地衡量了对应自变量的信息有多少已经被其他自变量涵盖。盲目加入 VIF 太大的自变量会严重降低模型的效果。一般来说，我们认为 VIF 大于 10 的自变量就明显存在多重共线性。

现在，计算人均工资有关自变量的方差膨胀因子。

表 3 自变量方差膨胀因子

人均占有耕地面积	人均工业总产值	人均固定资产投资额
1.31	3.80	6.48
人均社会消费零售额	人均学校数	人均货运量
9.87	1.38	1.82
失业率百分数	人均机构存款量	人均客运量
1.26	5.43	1.62

可以注意到所有的 VIF_i 均没有超过 10，不存在严重的多重共线性，没必要将任何因变量删除。尽管如此，“人均固定资产投资额”、“人均社会消费零售额”、“人均机构存款量” 三项的 VIF 值比较高，存在一定的多重共线性。这可能是因为这些

自变量与其他自变量存在先后因果的关系。这三项反映了居民的消费和投资，而不是和其他因变量一样是居民就业的产业，本身就和居民的创造价值的能力挂钩，存在一些多重共线性是合理的。

4.4 选模型分析

前面我们已经初步讨论了选模型和全模型的回归的差异，当多重共线性不严重时，我们使用全模型就可以有比较好的拟合效果。尽管如此，定理也表明了假如剔除所有自变量中的次要自变量，回归精度更高。我们已经使用 p 值找到了五个和人均工资相关显著性比较强的自变量，那么，一个最简单的想法就是取这些自变量作为选模型的回归自变量。然而，这种办法只是寻找了局部最优解，我们并不能断言它是整体最优的，原因可能在于其他自变量尽管与人均工资关系略弱，但是包含了这些自变量中未包含的信息；若把这样的自变量纳入回归中，预测能力一定是更好的。所以，我们必须考虑一种整体性的手段，即同时考虑自变量子集的联合影响，比较所有子集的作用并从中选取最优的。这种系统性的比较方法被称为自变量选择准则。常用的自变量选择准则有 RMS_q 准则，C_p 准则，AIC 准则，BIC 准则等。我们这里使用 AIC 准则和 BIC 准则这两种比较普适的方法分析选模型。

4.4.1 AIC 准则

日本统计学家 Akaike 在极大似然估计的基础上提出了 AIC 准则。我们用熵表示用估计分布代替原始分布时损失信息的多寡。对于两个分布函数 p 和 q，定义 KL 散度 (即相对熵) 为：

$$I(p||q) = E(log\frac{p}{q})$$

特别的，当 θ 是一个 k 维的未知量，θ_0 为其真值，并设 p 取 y 的真实分布 $g(y) = f(y|\theta_0)$，q 取条件概率分布 $f(y|\theta)$。

对于极大似然函数

$L(\theta|y) = f(y_1|\theta)f(y_2|\theta)\cdots f(y_n|\theta)$，极大似然估计 $\hat{\theta}$ 定义为使得 $L(\theta|y)$ 取到最大的 θ，由大数定理，这同时一定使得 $E(log(f(y|\theta)))$ 取到最大。与此同时，

$$I = E(log(g(y))) - E(log(f(y|\theta)))$$

显然，$E(log(g(y)))$ 是一个常数，因此极大似然估计等价于使得 KL 散度最小的估计。我们用 KL 散度的概念刻画了极大似然估计的本质，现在用它来推导 AIC 准则的定义。

我们不妨定义 $d(\theta) = -2E(log(f(y|\theta)))$。显然极大似然估计 $\hat{\theta}$ 使得 $d(\hat{\theta})$ 最小。我们用 $E(\hat{\theta})$ 作为 I 的估计量并使之最小化。利用 Taylor 展开可知，

$$E(-2log(f(y|\theta_0))) - E(-2log(f(y|\hat{\theta}))) = k + o(1)$$

$$E(d(\hat{\theta})) - E(-2log(f(y|\theta_0))) = k + o(1)$$

因此可知

$$E(d(\hat{\theta})) = E(-2log(f(y|\hat{\theta})))+2k = -2log(L(\hat{\theta}|y))+2k$$

于是我们把这个量定义为 AIC 统计量。AIC 统计量越小，证明估计的概率分布和真值的分布 KL 距离散度越小，也就是估计的效果越好。正态假设成立时，AIC 的计算可以继续化简。

定理 8. 正态假设成立时，$AIC = nln(SSE_q)+2q$

AIC 准则总结为：寻找使得 AIC 达到最小的自变量子集。当子集中的元素个数增加时 SSE_q 单调递减，而 $2q$ 单调递增。因此，当变量增加带来方差降低的收益超过变量增加带来的惩罚时，我们会选择加入变量；如果变量影响力不强而导致带来方差降低的效果不明显时，我们就停止这个过程。AIC 最小确实是整体的性质。

利用 AIC 对我们的回归自变量进行选择，并得到选模型的回归结果。

表 4 AIC 准则的选模型回归结果

变量名	系数估计值	标准差	t 检验 p 值
常数项	1.462e+04	8.427e+02	0.00
人均固定资产投资额	1.822e-01	4.393e-02	0.00
人均学校数	-2.910e+03	228e+03	0.02
人均货运量	1.977e+01	1.002e+01	0.05
人均机构存款量	9.501e-02	9.431e-03	0.00
人均客运量	4.227e+01	1.487e+01	0.00
残差项标准差 2782		F 检验 p 值 <0.0001	
R^2 0.6891		调整后的 R^2 0.6830	

回归后，对自变量进行 t 检验，p 值均很小，证明回归显著性良好。判决系数比较大，拟合优度较好。利用 AIC 准则得到的自变量是人均固定资产

投资额、人均学校数、人均货运量、人均机构存款量、人均货运量五个因素。其中，人均工资与人均学校数负相关，而与其他四个因素均正相关。

4.4.2 BIC 准则

BIC 准则考虑最大后验概率，将将 AIC 准则的方法加以改进，使得自变量的选取满足使 BIC 统计量最小。BIC 统计量定义为：

$$BIC = -2log(L(\hat{\theta}|y)) + log(n)k$$

当满足正态假设时，

$$BIC = nln(SSE_q) + log(n)q$$

可以发现相比于 AIC 准则，BIC 准则在自变量个数项前乘了样本个数的对数。也就是说，相比于用 AIC 准则做回归，使用 BIC 准则会使得增加变量个数受到的惩罚更苛刻，故使用后者得到的变量个数更少。AIC 的选变量策略更为保守，可能保留更多的自变量；而 BIC 的最优模型变量个数一般与真实模型更接近，而避免选用自变量过多导致过拟合的情况。

现在，利用 BIC 选择回归自变量。

表 5 BIC 准则的选模型回归结果

变量名	系数估计值	标准差	t 检验 p 值
常数项	1.277e+04	2.931e+02	0.00
人均固定资产投资额	2.369e-01	3.880e-02	0.00
人均机构存款量	9.153e-02	9.438e-03	0.00
人均客运量	4.601e+01	1.467e+01	0.00
残差项标准差 2811		F 检验 p 值 <0.0001	
R^2 0.6801		调整后的 R^2 0.6763	

只剩下人均固定资产投资额，人均机构存款量，人均客运量三个自变量。人均工资与这三个自变量均正相关。不难证明，用 BIC 准则获得的自变量一定是用 AIC 准则获得自变量的子集。

最后，值得一提的是，之所以我们要做选模型分析，是因为我们需要排除全模型中相关性不显著的自变量，从而改善拟合结果。但是这就意味着我们不需要做全模型分析或者说用 BIC 准则做回归一定要比 AIC 准则好吗？自然不是这样。理由在于有时我们就想要指定研究因变量和特定自变量之间的联系，但假如使用选模型致使这些自变量被舍弃了，我们就不能达成目的，这也违背了我们回归分析的初衷。

4.5 模型预测

现在，我们利用全模型和选模型分别在其他条件已知的情况下，对上一年的人均工资进行预测。预测出于两点目的，第一是证明模型具有应用的现实意义，第二是检验模型的的预测能力。我们之所以选用上一年 2005 年的数据做预测而不是下一年的数据，是因为人均工资这样的经济学指标往往呈现指数级成长的趋势，往前的数量波动要比往后的小，预测一般也更精确。预测方法也很简单，只需代入每座城市的自变量指标数值并乘以回归系数，求和即可。我们用 R 语言完成。

为了比较预测能力，我们引入均方误差 MSE 和平均绝对误差 MAE。

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

均方误差和平均绝对误差本质上都是以几何度量每个预测值偏离真值的程度作为损失函数并求损失函数的均值。当然我们也可以定义“平均 n 次误差”为 L^n 空间中距离的均值，这和均方误差也没有什么区别。均方误差和平均绝对误差越小，证明预测结果越好。

我们还引入一种更加直观的预测方法“直接预测法”，指的是用今年某个的人均工资直接作为上一年该城市人均工资的预测值。然后分别用全模型预测，AIC 准则下的选模型预测，BIC 准则下的选模型预测，并比较四种预测方式 MAE 值和 MSE 值的差异。

表 6 预测结果 MAE 值比较

模型	直接预测	全模型	AIC	BIC
平均预测误差	2215.841	2621.322	2620.622	2692.862

表 7 预测结果 MSE 值比较

模型	直接预测	全模型	AIC	BIC
均方误差	6e+06	1.6e+07	1.5e+07	1.7e+07

MAE 值和 MSE 值都很大，但是无需担心，

这是因为我们的数字没有进行标准中心化的预处理。全模型、AIC 法模型和 BIC 法模型三者误差都比较接近，其中 AIC 法的精确度略好一些，而全模型次之。在这种情况下，由于全模型保留了多得多的自变量而没有几乎损失太多回归精度，我们一般采用全模型即可。但是 AIC 和 BIC 准则下的选模型也提醒了我们哪些自变量是相对比较重要的，这给了我们在研究时的重点考察对象，大大提高了工作效率。

此外，我们还观察到我们用回归分析法得到的预测结果均方误差反而大于直接预测法的结果，也就是，我们历经千辛万苦的回归成果还不如直接套用今年数值做的直观预测。实际上，这一事实并没有完全否定我们的努力，而只是局部地体现了“用现在的回归套未来的”这种思想的错误。这也反映了经济学研究和统计学研究的本质不同，统计上正确的东西并不完全适用于现实中的经济现象。这也证明了我们之前的论断：人均工资完全取决于社会经济的发展——在现实中不成立。现实中，个人的工资要比整体的经济环境稳定。经济不景气时，企业一般采取裁员而不是降薪的方式降低成本。当公司大赚一笔的时候，员工也不可能一下子工资上涨很多。并且个人工资的变化是有滞后性的，社会生产力在一段时间后才会变成财富，社会创造的财富也需要一段时间让普通员工得到收益。直接用今年的预测去年的是不妥的，因为每一年我国的产业结构都不一样，产业的境况和就业者的薪资待遇也不一样，不应该假定有着相同的回归系数，而是要考虑时间要素。

5 模型总结、推广与不足

5.1 模型总结

我们采用了线性回归的方法把不同地区当作样本，研究了人均占有耕地面积、人均工业总产值、人均固定资产投资额等九个因素对人均工资的影响。根据线性回归操作简单、实现高效、方法完善、理论系统、具有良好解释性的特点，用于此处是非常合适的。从保守的角度看，人均固定资产投资额、人均学校数、人均货运量、人均机构存款量、人均货运量五个指标对于人均工资的影响最为显著，尤其是人均固定资产投资额、人均机构存款量、人均客运量三者，且它们均与人均工资正相关。

人均固定资产投资额反映了房地产等行业的繁荣程度。人均机构存款量反映了当地银行业的成熟度及居民的富裕程度。人均客运量体现了外来劳务收入和旅游业收入的多少。三者越高，经济活动越频繁，人均工资也越高，符合人们的常识。人均耕地占有面积、人均工业总产值、人均社会消费零售额三个自变量系数为负，人均工资呈现与它们相对不显著的负相关性。这可能是因为农业、工业、零售业这些传统行业的附加值都不够高，经济发达的城市里，这些行业的就业者逐渐转向信息业、金融业等附加值更高的新产业。

本文的另外一大优点是除了实际应用外，还阐述了每种方法的理论基础，以及在进行一次完整的统计分析时每一步的动机。线性回归结果的产生来源于最小二乘法。显著性检验的实质是一个约束最小二乘估计，可以将统计量变形为一个具有特定分布的量并以 F 检验或 t 检验完成。回归诊断分为对样本的诊断和自变量的诊断。样本的错误可能来自于样本残差不符合正态假设，或样本中混入了过多强影响点；必须在回归时删去错误样本。多重共线性是自变量错误选取的表现，可以用方差膨胀因子或条件数衡量。选模型帮助我们更好地选取自变量来最优化模型，变量选取常常服从 AIC 或 BIC 准则。

5.2 模型推广

模型只考虑了很多自变量对单一因变量人均工资的影响。我们可以把它推广到多元线性回归的形式，考察自变量同时对若干个因变量的影响，使得模型更具有一般性。

模型只取了一年的数据做回归，在忽略了年份差异可能存在的影响的情况下，不便于直接用于预测其他年份的人均工资，导致模型实用性略有欠缺。假如结合时间序列分析的方法建立模型则更佳。

5.3 模型不足

从模型预测一章的分析我们发现，由于现实经济问题的复杂性，对于人均工资和社会经济线性关系的假定不一定正确，实际上可能会是非线性关系。假如采用机器学习的方法对人均工资和自变量进行回归，可能回归的效果更好，模型使用面也更广。

模型把不同地区当作独立实验，且认为具有 Gauss-Markov 假设。但现实中相邻地方的经济往往会有粘性，样本与样本之间是存在联系的。而相比于较小城市来说，大城市的经济结构一般更加稳定，而小城市有着比较明显的支柱产业，也更倾向于被某个新开发的产业影响，故对于等方差性的假设也可能不圆满。

缺乏因果分析。比如人均机构存款量和人均工资的联系，单纯的线性回归可以指出其相关正负性，但是不能体现到底是谁影响了谁。很大一种可能是居民收入较高后，将衣食住行外的财产存在银行中，使得机构存款量提高，这就不能体现在回归分析的方法中。

参考文献

[1] 王汉生. 应用商务统计分析 [M]. 北京: 北京大学出版社,2008.

[2] 王松桂, 史建红, 尹素菊, 吴密霞. 线性模型引论 [M]. 北京: 科学出版社,2004.