2017—2018 第二学期
统计计算期末作业
北京市昌平区二手房数据分析

姓名：韩锡雅

学号：2016310815

班级：统计学 16

目录

1.数据提取

选取赶集网北京市昌平区二手房信息页面，利用 R 中 rvest 包的相关函数，写针对此网页的数据提取函数 DataScrap，用此函数抓取前 25 页二手房信息，包括标题、单价、面积、总价、户型、装修、楼层、朝向和地址，存为数据框。

2.数据描述

本次提取到赶集网北京市昌平区二手房信息 1100 条，每条信息包含二手房标题、单价、面积、总价、户型、装修、楼层、朝向和地址，对此数据进行进一步分析，有助于我们了解北京市昌平区二手房市场现状，为有意购买二手房的消费者提供昌平区二手房市场宏观信息，提高选房效率。同时，尝试对变量中的一个或多个进行分布猜想，建立合适的回归模型。

3.初步分析

对于购买者而言，单价、面积、户型、装修、楼层、朝向都是影响他们做出选择的因素，但往往单价（每平米价格）是他们关注的第一个要素。

对数据集中的单价变量进行探索，得到单价的箱线图和直方图（图 1）。

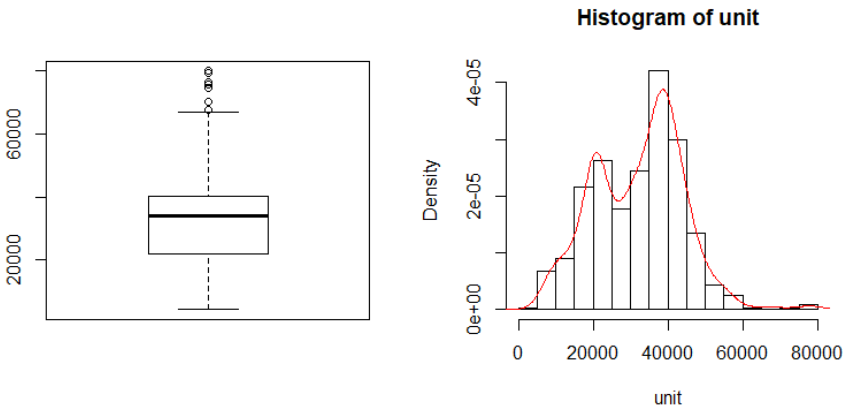

图 1：单价的箱线图和直方图

每平米价格均值为 32227.3 元，标准差为 12014.9，从直方图中可以看出单价的分布呈双峰，但整体仍大致有正态分布轮廓。

我们关心装修情况和朝向对二手房单价的影响，数据集中二手房中过半（692 户）为精装修房，但从不同装修的单价箱线图（图 2）对比中，发现不同装修状况对二手房单价的影响并不明显。

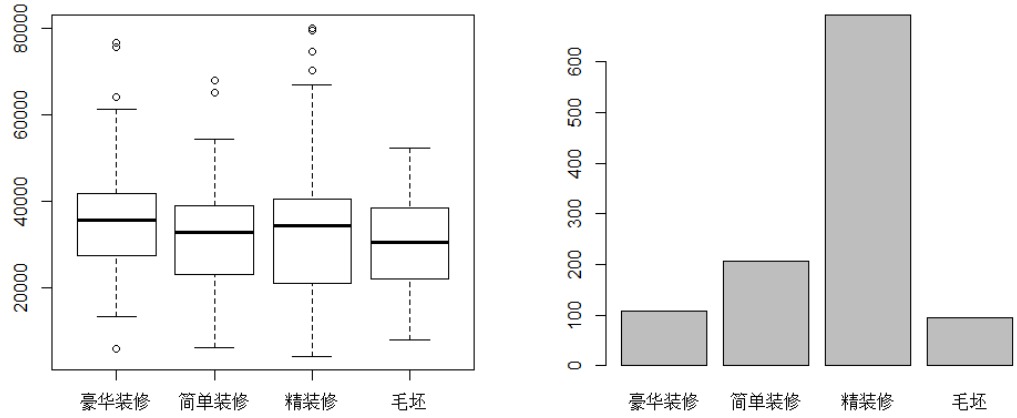

图 2：不同装修的单价箱线图和装修情况条形图

同时，数据集中二手房中朝向基本为南、北向，从不同朝向的单价箱线图（图 3）对比中，发现其他朝向二手房单价均值稍高于南北向二手房，但由于其他朝向二手房数量过少，所以不具有强参考价值。

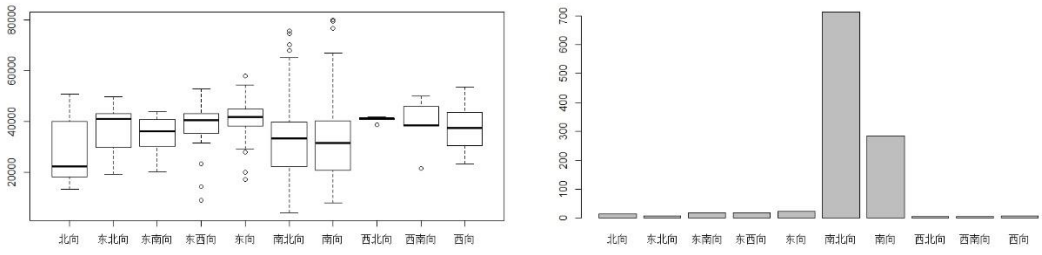

图 3：不同朝向的单价箱线图和朝向条形图

对类型较多的户型和楼层变量进行词频统计并画出词频图（图 4、5），可以看出二手房主要户型为 2 室 1 厅 1 卫（324 户）和 1 室 1 厅 1 卫（176 户），且楼层多为 6 层楼左右的低、中、高层。

图 4：户型词频图

图 5：楼层词频图

对于二手房面积，均值为 149.2 平方米，从图 6 可以看出呈右偏分布。

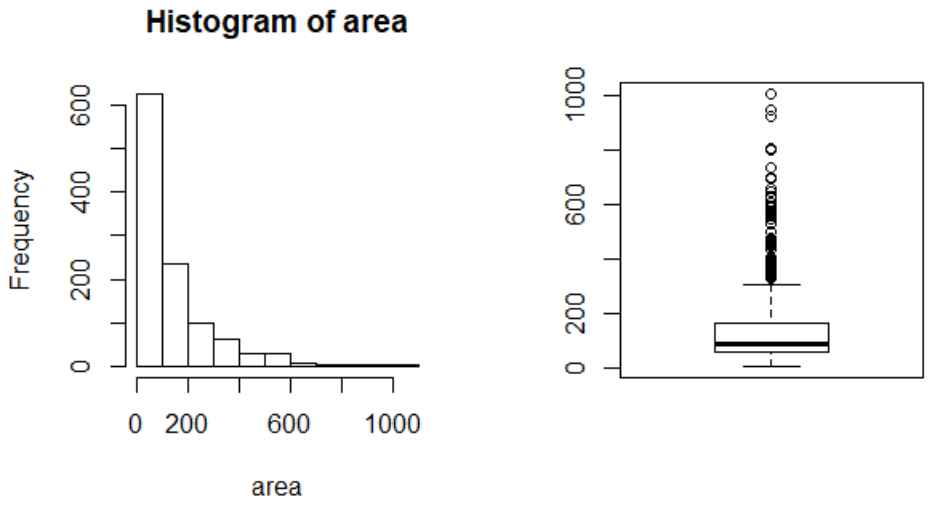

图 6：面积直方图和箱线图

对于二手房总价，其均值为 536.4 万元，从如图 7 左中看出总价呈明显的右偏分布，不服从正态分布。生成 1000 个服从均值为 536.4，方差为 0.006 的逆高斯分布的随机数，作出随机数的直方图，与总价直方图对比发现两者有相似性，所以猜想二手房总价这个变量服从逆高斯分布。

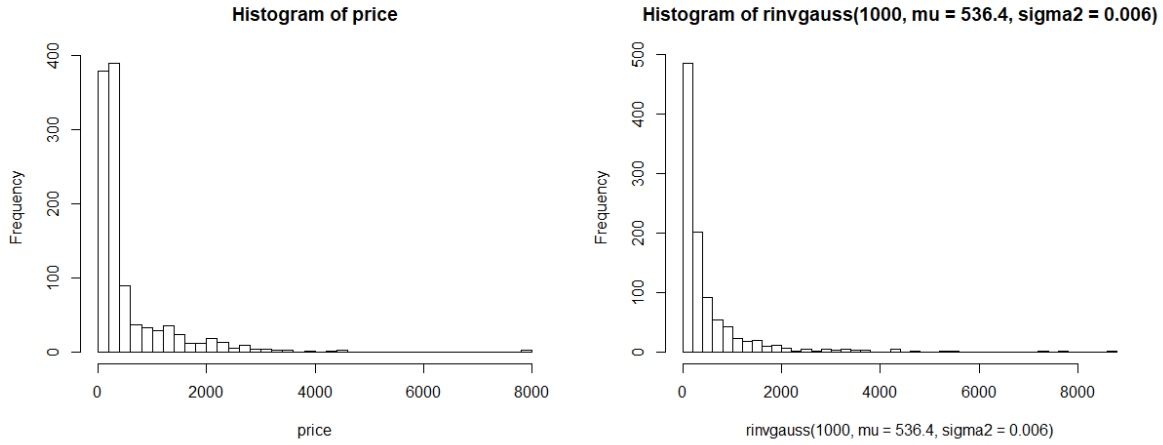

图 7：总价直方图和逆高斯分布直方图

4.逆高斯分布参数估计

写出逆高斯分布的对数似然函数l，进行μ和σ^2最佳值的搜寻。首先假定$\sigma^2=0.1$，给定μ范围，找到对数似然函数l极值点对应的$\hat{\mu}$；然后固定$\mu=\hat{\mu}$，给定σ^2范围，找到对数似然函数l极值点对应的$\hat{\sigma}^2$，则$\sigma^2=\hat{\sigma}^2$。

最终得到结果为：$\mu=536.42$，$\sigma^2=0.00326$。

最终得到结论，猜想是合理的，二手房单价变量服从一个$\mu=536.42$，$\sigma^2=0.00326$的逆高斯分布。

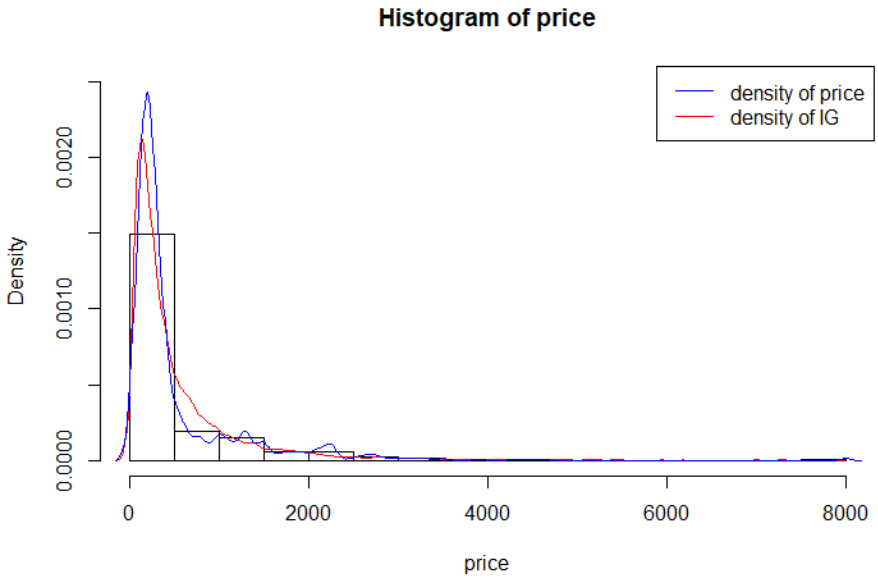

图 8：单价直方图、密度图与逆高斯密度图

5.逆高斯回归模型

以二手房总价 price 为因变量 y，分类变量户型 type 为自变量 x，进行回归模型的建立：

在 3 分析中，发现 y 不服从一般线性回归的正态假定，故不能使用一般线性回归模型。假定因变量 y 服从指数分布族中的逆高斯分布，建立广义线性回归模型—逆高斯回归模型。

借助指数族分布，对响应变量 Y 的描述将不再局限于正态分布，称观测 $y_{1,\ldots,}y_n$ 来自指数族分布，如果其概率密度函数可以表达为如下形式：

$$f(y_i|\theta_i,\varphi_i)=\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{\varphi_i}+c(y_i,\varphi_i)\right\}. \qquad (4.1)$$

其中，θ_i 是指数族的自然参数，φ_i 称为尺度参数；$b(\cdot)$ 以及 $c(\cdot)$ 是依据不同指数族而确定的函数且 $c(\cdot)$ 只由 y_i 和 φ 决定。

指数族分布的均值为：

$$E(y_i)=b'(\theta_i). \qquad (4.2)$$

指数族分布的方差为：

$$Var(y_i)=\varphi_i b''(\theta_i). \tag{4.3}$$

逆高斯分布是统计学中一种常见的分布，属于指数族分布，其密度函数为：

$$f(y;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi y^3\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2(\mu\sigma)^2 y}\right\},y>0,\mu>0,\sigma>0. \tag{4.4}$$

该分布含有两个参数μ和σ^2，当σ^2趋近于 0 时，逆高斯分布逐渐趋近于高斯分布，逆高斯分布有多项类似于高斯分布的特性。

式 4.4 可转化为逆高斯分布的指数分布族形式：

$$f(y;\mu,\sigma^2)=\exp\left\{\frac{-y/(2\mu)^2+1/\mu}{\sigma^2}-\frac{1}{2y\sigma^2}-\frac{1}{2}\ln(2\pi y^3\sigma^2)\right\},\quad y>0,\mu>0,\sigma>0. \tag{4.5}$$

将式 4.1 和式 4.5 对比，可得，$\varphi=\sigma^2$，$b(\theta)=-\frac{1}{\mu}$，$\theta=-\frac{1}{2\mu^2}$，故逆高斯分布的均值为μ，方差为$\sigma^2\mu^3$。

逆高斯分布假设下的对数似然函数为：

$$l=\sum_{i=1}^{n}\left\{\frac{-y_i/(2\mu^2)+1/\mu_i}{\mu^2}-\frac{1}{2y_i\sigma^2}-\frac{1}{2}\ln(2\pi y_i^3\sigma^2)\right\}. \tag{4.6}$$

逆高斯回归模型的残差偏差：

$$D=\sum_{i=1}^{n}\left\{\frac{(y_i-\mu_i)^2}{y_i\mu_i^2}\right\}. \tag{4.7}$$

利用迭代加权最小二乘估计[1]，对回归系数进行估计。

逆高斯回归模型通常使用对数联系函数：

$$g(\mu)=\ln(\mu). \tag{4.8}$$

迭代加权最小二算法中，

$$W=\mathrm{diag}\left[\frac{\omega_i}{\varphi v(\mu_i)[g'(\mu_i)]^2}\right]_{(n\times n)}=\mathrm{diag}\left[\frac{1}{\sigma^2\mu_i}\right]_{(n\times n)}. \tag{4.9}$$

$$z=[\eta_i+(y_i-\mu_i)g'(\mu_i)]_{(n\times 1)}=[\eta_i+(y_i-\mu_i)/\mu_i]_{(n\times 1)}. \tag{4.10}$$

在 R 中实现上述算法时，先设定残差偏差D、均值μ和η的初始值，开始循环。其中，$\beta=(X^TWX)^{-1}X^TWz$，$\eta=X\beta,\mu=\exp(\eta)$，当残差变化$\Delta D$足够小时停止循环，输出回归系数$\beta$。将此过程写为函数$IG_reg(y,X)$，参数为因变量$y$，自变量矩阵$X$，输出结果为回归系数、拟合值、残差平方和、AIC 值、皮尔逊卡方统计量$\chi^2=\Sigma\frac{(y-\mu)^2}{\mu^3}$。

将y与X代入函数$IG_reg(y,X)$运行，得到迭代加权最小二乘法下的回归系数估计如下：

```
(Intercept) 4.7338203
x           0.1046764
```

6.比较分析

在 R 中，可以建立广义线性模型的函数有 glm 和 gamlss。Glm 专门用于建立广义线性模型，gamlss 可以建立更加一般意义上的回归模型。

在 4 中建立的逆高斯广义回归模型可以直接用 glm 或 gamlss 函数实现。

Glm 函数结果如下：

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.733869   0.023012  205.72   <2e-16 ***
  x         0.104668   0.002997   34.93   <2e-16 ***
```

gamlss 函数结果如下：

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.734303   0.041596  113.81   <2e-16 ***
  x         0.104598   0.005994   17.45   <2e-16 ***
```

glm 显著性更高。

迭代加权最小二乘算法、glm、gamlss 三者回归系数结果对比如下：

```
                WLS       glm       gamlss
(Intercept) 4.7338203 4.7338694 4.7343027
x           0.1046764 0.1046685 0.1045985
```

可以看出，用迭代加权最小二乘算法得到的回归系数和用 R 中 glm 函

数、gamlss 函数得到的回归系数基本相同，所以三者的拟合值对观测值散点图没有差别。

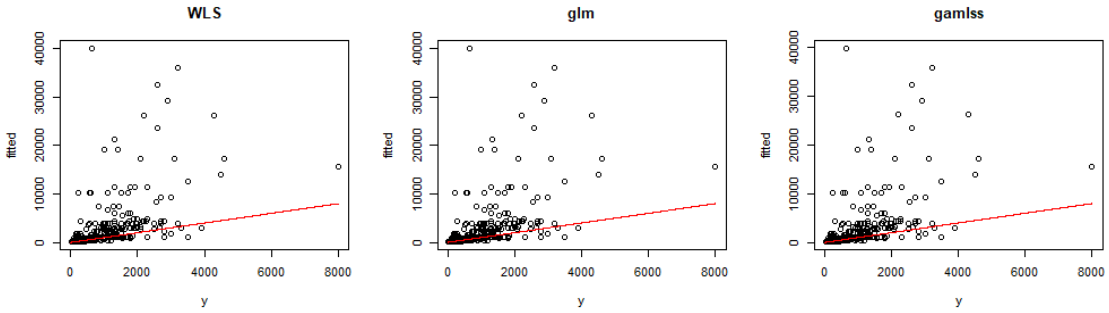

图 9：三个模型拟合值对观测值散点图

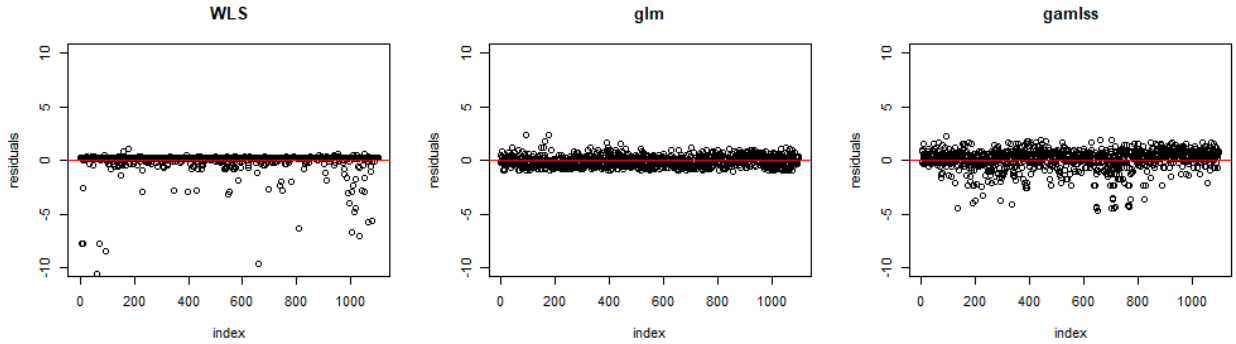

图 12：三个模型残差顺序图

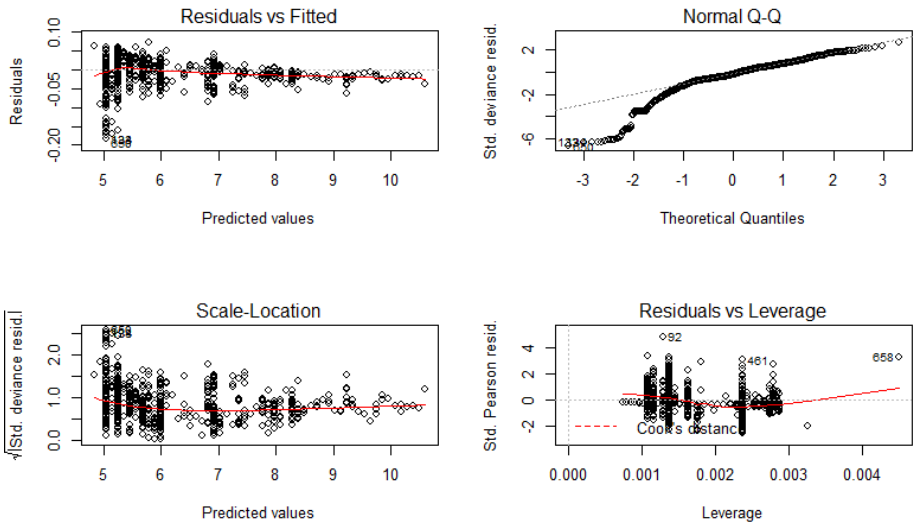

图 10：Glm 函数

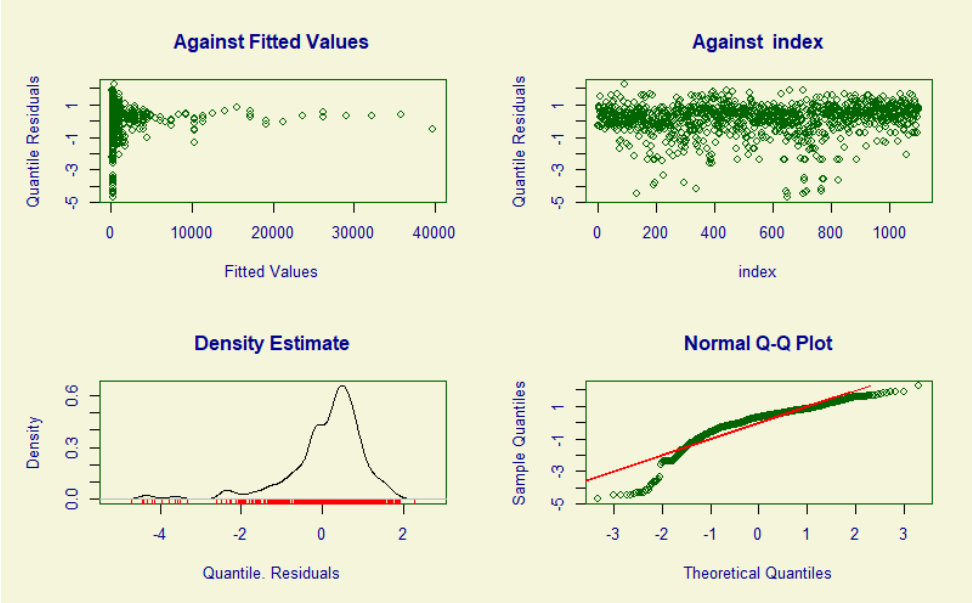

图 11：Gamlss 函数

经计算，三个模型的 AIC 值如下：

	WLS	glm	gamlss
AIC	20812.41	14851.63	14851.63

Glm 函数与 gamlss 函数的 AIC 值相同且明显小于迭代加权最小二乘算法，所以虽然三者的回归系数结果相差不大，但 R 中函数模型拟合优良性要高于用迭代加权最小二乘算法建立的模型。

7.结论

通过分析在赶集网提取的二手房数据了解到目前北京市昌平区二手房市场的大体情况，每平米价格集中在 2 万元到 4 万元之间，户型多为 2 室 1 厅 1 卫和 1 室 1 厅 1 卫这样的基本户型，精装修房比较多但装修状况对每平米价格的影响并不明显，房屋朝向基本都是南向或北向，较多为中低层。

通过探究二手房总价的分布，发现总价这个变量并不具有正态性，呈明显的右偏分布，通过与逆高斯分布对比，猜想总价服从一个逆高斯分布，对参数进行估计，最终验证了猜想是合理的。

在建立总价与户型的线性回归模型时，根据因变量的逆高斯分布建立广义线性回归模型——逆高斯回归模型，并用迭代加权最小二乘算法的到回归结果，与 R 自带函数对比后发现回归系数结果相同，但 AIC 值大，拟合优良性相对较差。

8.参考文献

[1]孟生旺. 回归模型[M]. 中国人民大学出版社, 2015.