附件一：调查方法和技术分析方法

调查方法

据了解，关于居民收入的官方统计数据之所以发生较严重的失真，主要是由于一部分居民对调查敏感，不愿意透露其真实收入情况，因此常常发生严重低报收入或者拒绝调查的现象。官方调查可能过度依赖统计调查所要求的分层随机抽样方法和居民自身对家庭收支的记录，但缺乏相应的手段降低调查的敏感度，也没有相应的措施对收入的真实性进行检验。

为此，本调查采用定向调查的方法，以提高数据的真实性。定向调查方法在以下方面不同于常规的分层随机抽样调查。

第一，分层随机抽样方法所抽出的样本户是随机形成的，而我们的调查采用定向方法。首先在全国各地慎重选择信誉好、工作规范的专业调查公司作为合作对象，并对这些公司符合要求的调查员进行严格培训，按照一定的配额，通过他们的社会关系，对其熟悉并了解其基本家庭情况的人群进行调查。这包括调查人员的亲属、朋友、近邻、前同学和前同事等。由于调查者事先了解被调查者的家庭情况，包括基本经济情况，并与被调查者之间存在一定的私人信任关系，被调查者一般没有必要对调查者大量隐瞒收入。而且当被调查者明显隐瞒或虚报收入时，调查者常常能够对此做出判断。多次调查的经验证明，这些措施对减少瞒报、漏报收入的现象有非常大的帮助。

第二，采取了若干降低敏感度的措施，最大限度地减少被调查者对调查的心理防范。除了调查者和被调查者之间的信任关系，这些措施还包括：首先，问卷调查采取匿名的方式，并提供保密承诺；其次，在问卷设计中，严格遵循先询问敏感度较低的问题、后询问敏感度较高的问题的原则。因此，问卷设计先询问分项消费支出，后询问分项收入和各项资产变动情况，总和数据取自分项的加总而不是由被调查者提供。

第三，调查采取面访的方式，而不像官方调查让被调查者自己记账。如果被调查者无条件配合调查，后一种方式会更准确。但如果他们对提供某些信息敏感，则后一种方式也会给他们提供足够的时间考虑如何隐瞒某些数据，同时又避免各项收支数据互相不一致。而面访的调查方式则不给被调查者从容的时间这样做。

第四，设计了严格的质量检验程序对数据的真实性进行核查。某些被调查者刻意瞒报、虚报某些信息，通常会造成收支数据的不一致。通过数据分析和逻辑检查，可以发现这种情况，剔除不合格问卷。此外，收支检查和其他逻辑检查也有助于发现个别调查人员的问卷造假行为。

在2012年调查中，我们进一步改善了对调查人员的事先培训和调查过程中的监督指导，提高了调查质量；总共收回问卷5756份，经质量检查筛除了7.2%的不合格问卷，保留有效样本5344份。

上述这些措施当然不可能保证问卷信息百分之百准确，但在可能的条件下最大限度地减少了调查信息的系统性偏差，基本保证了调查数据真实可靠。

技术分析方法

由于抽样方法不同，我们不能直接使用样本数据来推断全国城镇居民的收入分配状况，而是基于比较真实可靠的样本数据进行计量模型分析，通过回归分析估算收入水平与一些主要消费特征参数之间的函数关系，并在此基础上对收入水平随影响变量而变动的趋势进行模拟，据此对统计数据进行校正。

其中，一个关键的消费特征参数是恩格尔系数（即居民家庭的食品消费支出占家庭消费支出总额的比例）。经济学界公认，恩格尔系数是一个与收入水平密切相关的变量。居民食品支出在消费支出中的比例会随着收入水平提高而递减，使恩格尔系数呈下降趋势。

根据这个原理，我们可以利用一个数据可靠并有代表性的调查样本，使用计量经济学方法，来建立居民家庭的恩格尔系数和人均可支配收入水平间的函数关系。依据所得到的函数关系，我们可以对任意一组居民收入统计数据进行检验。也就是说，只要我们能够得到某一组统计样本的相对可靠的恩格尔系数和其他相关参数，就可以用模拟的方法近似推算出该组居民的人均收入水平。据此，我们可以对官方公布的分组城镇住户的人均可支配收入数据进行检验，以发现这些统计数据是否存在系统性误差、误差有多大，并进行校正。

这样做的前提，是要求统计样本的恩格尔系数真实可信。曾经有人质疑，如果居民的收入统计数据有系统性误差，同一样本的恩格尔系数是否也会有系统性误差，使用这样的恩格尔系数进行推算是否可行？事实上，当居民收入数据存在偏差时，他们的消费和食品消费支出数据很可能也存在一定偏差。但前面已经指出，由于两者敏感度的不同，消费支出偏差通常会显著小于收入偏差。更重要的是，只要消费支出偏差和食品消费支出偏差同方向，并在统计意义上大体保持同比例，那么推算出的恩格尔系数仍然是基本准确的。最后，如果消费支出与食品消费支出的偏差不保持同比例，在计算恩格尔系数时，同方向的偏差仍可以在很大程度上互相抵消（两者是分母和分子的关系），使恩格尔系数的偏差远远小于收入水平的偏差。因此，统计样本的恩格尔系数仍然可以用来推算收入水平，只是推算结果的准确程度会下降。

研究发现，居民的恩格尔系数不仅受到其收入水平的影响，同时还会受到其他一些因素的影响。因此在进行计量分析时，需要把这些因素作为控制变量包括在内，并在计算恩格尔系数与收入水平的关系时考虑这些因素的影响。

除了收入，其他已知影响恩格尔系数的因素主要有：第一，不同地区居民的消费习惯差异。有些地区居民比其他地区居民有较高的饮食偏好。第二，不同规模城市的居民消费特征差异。这主要是不同规模城市某些消费品（特别是食品）价格水平的差别造成的。第三，家庭人口规模的差异。人口较多的家庭在食品支出方面可能具有规模效应，能够节约食品支出。第四，家庭成员的平均教育程度。教育程度较高的居民有多方面消费需求，包括精神层面需求，如通信联络、教育、文化娱乐、旅游等。第五，家庭成员的就业面（就业的家庭成员占全部家庭成员的比例）。一方面，家庭就业率较高，可能节约食品支出，因为从业者有可能在单位就餐。另一方面，较高的就业率又有可能导致较多的外出就餐，从而导致较高的食品支出。究竟哪种因素占上风，还需要通过检验来证明。

由于恩格尔系数与人均收入之间存在某种非线性关系，但与其他一些影响变量的关系可能是线性或接近线性的，作者通过数据分析选择了半对数函数的形式建立计量模型。该函数以恩格尔系数为被解释变量：

$$N=C+a_1\ln Y+a_2S+a_3H+a_4E+a_5M+a_6D_0+a_7D_1+a_8D_2 \qquad (1)$$

上式中，N是恩格尔系数，解释变量lnY是人均可支配收入的对数，S是城市规模，H是家庭规模，E是家庭成员教育程度，M是家庭就业面，D_0、D_1、D_2是区域消费差异虚拟变量，C是常数项，a_1到a_8分别为各变量的系数。在前期分析中发现就业面变量虽然在多数函数形式中具有负系数，但不具有统计显著性，故从函数中剔除。

回归结果显示，除了城市规模估计值在接近1%水平上统计显著，其余各解释变量，包括虚拟变量在内，均达到0.1%的统计显著程度，说明各解释变量均与被解释变量的关系紧密，模型有很好的解释力。

根据函数（1）的回归结果进行计算，就可以得出恩格尔系数随人均收入水平和其他控制变量的变化而变化的值。再从居民收入统计数据中查出各组居民的恩格尔系数，并代入各控制变量的全国赋值，就可以倒推出对应于不同恩格尔系数的人均收入水平。