2022.4.7

更正数据后的实验

这里将参考集与测试集扩展到了28期，并对基准预测进行了修正。

固定效应的惩罚项是10/50（如果按照之前设置为1则结果较差），还尝试了选择5个参考集上最优的模型进行组合（并不能保证处理共线性，但是处理共线性基本都要减少变量）。Pinball loss 结果如下：

分位数	0.01	0.025	0.165	0.25	0.5	0.75	0.835	0.975	0.99
quantGAM	0.0146	0.0368	0.2336	0.3378	0.5638	0.5554	0.4769	0.1592	0.0887
VZ	0.0143	0.0356	0.2275	0.3289	0.5374	0.5320	0.4551	0.1454	0.0785
WSS	0.0143	0.0356	0.2270	0.3322	0.5740	0.5672	0.4745	0.1520	0.0754
poisson_static	0.0409	0.0706	0.2692	0.3631	0.5448	0.5434	0.4786	0.2146	0.1491
poisson_damped	0.0243	0.0495	0.2347	0.3243	0.4974	0.4926	0.4279	0.1683	0.1084
poisson_undamped	0.0249	0.0497	0.2322	0.3201	0.4915	0.4906	0.4306	0.1880	0.1318
nb_static	0.0144	0.0358	0.2295	0.3382	0.5878	0.6205	0.5292	0.1498	0.0759
nb_damped	0.0150	0.0366	0.2206	0.3179	0.5172	0.5129	0.4428	0.1553	0.0839
nb_undamped	0.0144	0.0356	**0.2164**	0.3105	0.4958	0.4979	0.4359	0.1681	0.1043
固定效应组合9-10	0.0143	**0.0355**	0.2174	0.3122	0.5099	0.5088	0.4267	0.1356	0.0704
固定效应组合9-50	0.0143	0.0356	0.2168	0.3091	**0.4915**	**0.4897**	**0.4217**	**0.1351**	**0.0702**
固定效应组合5-10	**0.0143**	**0.0355**	0.2171	0.3125	0.5120	0.5102	0.4281	0.1369	0.0718
固定效应组合5-50	0.0143	0.0356	0.2181	**0.3090**	0.4932	0.4901	0.4231	0.1373	0.0717
简单平均-9	0.0177	0.0397	0.2217	0.3161	0.5091	0.5071	0.4327	0.1362	0.0710
倒数损失平均-9	0.0159	0.0381	0.2213	0.3158	0.5078	0.5052	0.4313	0.1360	0.0706

结论如下：

- 回归组合方法基本可以在各个分位数上战胜基准模型与简单/倒数损失平均的；
- 关于固定效应的惩罚项，惩罚系数设为10时有一些固定效应项是超过0.01的，但是惩罚系数设为50时最大固定效应仅为0.0001量级，即固定效应几乎不起作用；非固定效应的截距项基本都高于0.01。这说明，固定效应对回归组合的作用并不显著，而截距确实对组合起到一定作用。结合3.21报告的实验结果，可以认为：分位数回归组合的截距项是有用的，但是固定效应的作用不明显。

分位数回归组合——截距项作用

所参考的文献是 ***Combining forecast quantiles using quantile regression: Investigating the derived weights, estimator bias and imposing constraints*** (Taylor & Bunn, 1998) (Journal of Applied Statistics)。这篇文章基于 Granger (1989) 的分位数回归组合的实验(***INTERVAL FORECASTING An Analysis Based Upon ARCH-quantile Estimators***)，进一步探讨了分位数回归组合的问题，特别是截距项的问题——**截距项与分位数回归的无偏性有关。**（这一点和无截距的线性回归是对称的）

首先，分位数回归组合的求解即是优化下式（分位数回归代入可得）

$$\min_{\beta_1,\beta_A,\beta_B}\left[\sum_{t|y_t\geq \beta_1+\beta_A Q^A_{yt}(\theta)+\beta_B Q^B_{yt}(\theta)}\theta\left|y_t-\beta_1-\beta_A Q^A_{yt}(\theta)-\beta_B Q^B_{yt}(\theta)\right|\right.$$

$$\left.+\sum_{t|y_t<\beta_1+\beta_A Q^A_{yt}(\theta)+\beta_B Q^B_{yt}(\theta)}(1-\theta)\left|y_t-\beta_1-\beta_A Q^A_{yt}(\theta)-\beta_B Q^B_{yt}(\theta)\right|\right]$$

Granger认为分位数回归是一个一致估计，故渐近无偏，起到了修偏的作用（这与他在1984的线性组合回归的思路类似）；其还认为如果两个预测在样本上无偏，则可以将组合改为权重和为1的加权平均，以保证简单性与稳健性。——这是本篇论文要批判的点

在论文的论证中，先说明了独立同分布误差下分位数回归组合意义不大（不需要引入回归变量，常数即可解决问题），而在异方差情况下才需要引入变量。其对分位数组合的解释是，可以将分位数基准预测写成下式，即均值成分与误差的分位数之和：

$$Q^A_{yt}(\theta)\equiv m_{At}+Q^A_{et}(\theta)$$

则回归组合可以写成下式，即包含均值与误差分位数解释变量的受约束回归：

$$\hat{Q}_{yt}(\theta)=\beta_1+\beta_A m_{At}+\beta_A Q^A_{et}(\theta)+\beta_B m_{Bt}+\beta_B Q^B_{et}(\theta)$$

写成此式则可利用分位数回归的渐进理论。上式包含了均值的解释变量和误差分位数的解释变量，分位数回归将分别对均值和误差分位数进行建模。

关于截距项的讨论， Koenker and Bassett (1978) 在 ***Regression Quantiles*** 一文中证明，如果回归中有截距项，则以下划分不等式成立：

$$N(\mathbf{u}(\theta))\leqslant T\theta\leqslant T-P(\mathbf{u}(\theta))=N(\mathbf{u}(\theta))+Z(\mathbf{u}(\theta))$$

其中 N, P, Z 分别是误差项 $u=Y-X\hat{\beta}$ 小于0、大于0、等于0的个数，T 是样本数，θ 是指定分位数的概率。当解唯一时，以上等式严格成立。

这个式子的含义是，数据被分位数分割，**大约 θ 的估计值大于观测值，而剩下的估计值小于观测值。**而没有截距，此定理失效。为了进一步验证，其做数值模拟实验以证明这一点，最终结论是有截距的回归组合是可以满足上式，而没有截距的组合不满足。

要进一步检验无偏性，首先要确定无偏性定义：对于分位数预测，不能使用线性回归的误差定义偏差，因为在估计分位数时，无法解释的变化是估计的一部分（误差的部分进入了估计中，因是分位数）。故使用 Granger (1989) 的定义（***Invited Review Combining Forecasts-Twenty Years Later***）（qreg unbiasedness)：**随样本增大，观测值小于估计的分位数的比例趋于指定概率。**其证明了以下事实：

- qreg unbiasedness充要条件是：序列 $y-Q(\theta)$ 在没有解释变量的常数上的分位数回归会随着观察次数变大而导致常数为零。
- 满足划分不等式的回归满足“无偏性”，（不需要考虑基准的无偏性）；
- 无偏基准预测的单位组合不一定“无偏”，观察集相同时是无偏的，但是当两个基准预测依托的观察集不同时是不一定满足无偏的（不能直接不等式相加）。作者的实验也验证了这一点。

  ——暴论：Granger对“无偏”的定义是一个大样本性质，一致-->渐近无偏 等价于 渐近不无偏-->不一致，故不加截距的组合系数是不一致的？

之后计划

- 考虑Global的基于特征方法（比如基于深度学习得到的特征），以及如何在非回归的 Global 组合中引入截距项；
- 引入两个全局基准预测（比如 LSTM/LightGBM 等），一个针对分布，一个针对分位数，以接近M5的方法；
- 11周左右《智能优化算法》课程作业，可能会尝试用启发式算法（遗传算法、粒子群等）进行组合权重优化尝试。