Fama-Macbeth回归讨论

邵一淼 191098180

一、引言——从CAPM到Fama-Macbeth

Fama-Macbeth回归是1973年Fama和Macbeth为验证CAPM模型而提出的一种因子统计方法，该模型现如今被广泛用于计量经济学的面板数据分析，而在金融领域在用于多因子模型的回归检验，用于估计各类模型中的因子暴露和因子收益（风险溢价）。

Fama-Macbeth回归是实证资产定价中最常用的方法之一。它的主要用途是验证因子对资产收益率是否产生系统性影响。与投资组合分析不同的是，Fama-Macbeth回归可以在同时控制多个因子对资产收益率的影响下，考察特定因子对资产收益率产生系统性影响，具体体现在因子是否存在显著的风险溢价。

CAPM：$r_i = r_f + (r_m - r_f) * \beta$

这个公式有三个含义：

1、风险与收益的关系是线性的

2、β是对系统性风险的完全度量

3、$r_m - r_f > 0$，在一个风险规避的世界，更高的风险要有更高的收益

要验证CAPM只需要看满不满足上面的三个条件，因此，设定要拟合的模型为：

$r_i = \gamma_0 + \gamma_1 * \beta + \gamma_2 * \beta^2 + \gamma_3 * s + \epsilon$

s是系统性风险，ϵ为残差项

条件1成立有：$\gamma_2 = 0$

条件2成立有：$\gamma_3 = 0$，非beta风险不具有系统性影响

条件3成立有：$\gamma_1 = r_m - r_f > 0$

sharpe-lintner capm假定：$\gamma_0 = r_f$

详细的步骤为（以论文为例）：

1、用四年1926-1929的月收益率，对个股进行时序回归，计算出beta，排序分组为20组

2、用之后五年共60个月1930-1934年的月收益率，重新时间序列回归计算出个股beta和个股残差标准差$s(\epsilon)$，在计算出beta（个股beta的直接简单平均）和组合残差标准差$s_p(\epsilon)$（个股残差标准查直接简单平均）

3、之后四年1935-1938，每一个月都进行一次截面回归，那么四年回归48次。每一次截面回归的因子都是用上期获得的因子（不是上个月的因子）。具体来说，他是个这样的结构化数据：

	ri	beta	beta**2	s(e)
1935-01	group1			
	group2			
	…			
	group20			
1935-02	group1			
	group2			
	…			
	group20			

…使用过去60个月的数据进行时间序列回归，得到因子（第二步）后进行截面回归。

然后进行滚动回归，每次都是用过去的60个月跟新beta与s(e)因子。

所以每个时点的截面回归也就20个样本。

到1935-02，重新滚动分组，从第一步开始，用1930-02~1935-01这60个月的数据进行时间序列回归后得到每组的beta与s(e)

（这里，关于什么是滚动分组：beta是滚动计算的，如果分组里的股票一直不变显然是不太合理哈。因为分组是按照beta分组，是重要保证高beta组的股票的beta始终高。所以分组不能固定，一个股票在上期可能是第一组，下一期可能是第二组。也就是1935-02的分组依据是用四年的月收益率，对个股进行时序回归，计算出beta然后分组）

我们将我们设定的模型加上下标t表示时间，加上p表示组合，因为最后的截面回归就是组合之间的：

$$r_{pt} = \gamma_{0t} + \gamma_{1t}\beta_{p,t-1} + \gamma_{2t}\beta^2_{p,t-1} + \gamma_{3t}s(\epsilon)_{p,t-1} + u$$

p=1,2,…20

u表示残差，ϵ是时间序列回归得到的残差。

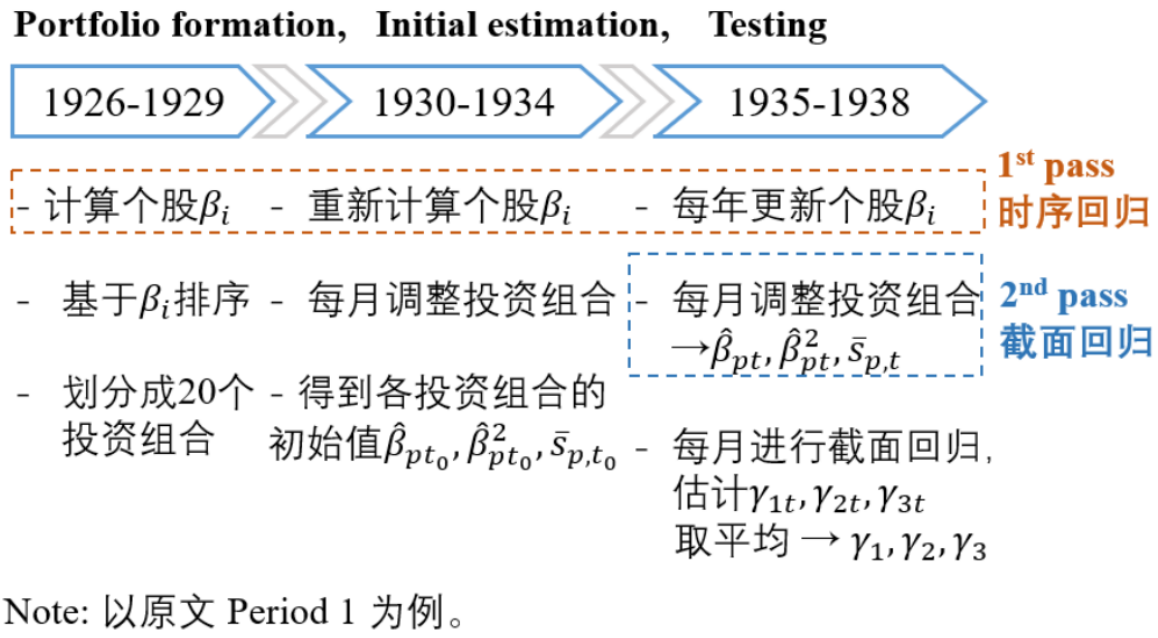
4、对所有截面回归得到的参数求均值，得到我们对参数的最终估计。第三步界面回归完成后，我们得到了这样的结构化数据：

	gamma0	gamma1	…
1935-01			
1935-02			
1935-03			
…			

参数$\gamma_i = \frac{1}{T}\sum \gamma_{it}, i = 0,1,2,3$，标准差就是直接求标准差，那么t统计量也有了：

$$t = \frac{\bar{\gamma}}{\sqrt{\frac{s_\gamma}{n}}}$$

然后就可以进行假设检验了



Note: 以原文 Period 1 为例。

二、方法概述

Fama-Macbeth与传统的截面回归类似，本质上也与是一个两阶段回归，不同的是它用巧妙的方法解决了截面相关性的问题，从而得出更加无偏，相合的估计。

时间序列回归

Fama-Macbeth模型与传统截面回归相同，第一步都是做时间序列回归。在因子分析框架中，时间序列回归是为了获得个股在因子上的暴露。如果模型中的因子是 portfolio returns（即使用投资组合收益率作为因子，例如Fama-French三因子模型中的SMB，HML和市场因子），那么可以通过时间序列回归（time-series regression）来分析$E[R_i]$和β_i在截面上的关系。

令f_t为因子组合在t期的收益率，R_{it}为个股i在t期的收益率，用f_t对每只股票的R_{it}回归，即可得到每只股票的全样本因子暴露β_i。

$$R_{it} = \alpha_i + \beta_i f_t + \epsilon_{it}, t = 1, 2, \ldots, T, \forall i \quad (1)$$

也可滚动计算某个时间段的因子暴露β_{it}，体现个股随市场的变化设置时间段长度为period

$$R_{ik} = \alpha_i + \beta_{it} f_k + \epsilon_{ik}, k = t - period, 2, \ldots, t, \forall i \quad (2)$$

截面回归

传统截面回归的第一步是通过时间序列回归得到个股暴露，这一步与Fama-Macbeth回归相同，而第二步回归体现了传统截面回归和Fama-Macbeth的最大不同。

传统截面回归：

在时序回归中回归式在时间序列上取均值，在$E[\epsilon] = 0$的假设下可以得出：

$$E[r_i] = \alpha_i + \beta_i E[f] \quad (3)$$

上式正是个股的期望收益与因子暴露在截面上的关系，截距α_i为个股的错误定价。

那么便可通过截面回归找到因子的期望收益率$E[f]$，方法是最小化个股定价错误α_i的平方和。对个股的的收益在时序上取均值得到个股期望收益$E[R_i]$，用全样本的个股因子暴露对个股期望收益做无截距回归。

$$E[r_i] = \beta_i \lambda + \alpha_i \quad (4)$$

回归残差α_i为个股的错误定价，λ为因子的期望收益率。

截面回归最大的缺陷在于忽略了截面上的残差相关性，使得OLS给出的标准误存在巨大的低估。

Fama-Macbeth回归

与截面回归相同，Fama-Macbeth回归第一步是通过时间序列回归得到因子暴露值，不同的是，第二步中，Fama-Macbeth在每个t上都做了一次无截距截面回归：

$$R_{it} = \beta_i \lambda_t + \alpha_{it}, i = 1, 2, \ldots, N, \forall t \text{ (5)}$$

上式中的β_i为全样本β，当然若使用滚动回归数据，也可以在不同截面的回归上使用对应时期的$\beta_{i,t}$。

Fama-Macbeth回归相当于在每个t上做一次独立的截面回归，这T次回归的参数取均值作为回归的估计值：

$$\hat{\lambda} = \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}_t$$

$$\hat{\alpha}_i = \frac{1}{T} \sum_{t=1}^{T} \hat{\alpha}_{it}$$

上述方法的巧妙之处在于它把 T 期的回归结果当作 T 个独立的样本。参数的 standard errors 刻画的是样本统计量在不同样本间是如何变化的。在传统的截面回归中，我们只进行一次回归，得到λ和α_i的一个样本估计。而在Fama-Macbeth截面回归中，把T期样本点独立处理，得到T个λ和α_i的样本估计。

若使用全样本因子暴露β_i进行估计,截面回归和Fama-Macbeth的估计结果相同，当使用滚动窗口进行估计时（Fama and MacBeth (1973)中作者使用了滚动窗口），截面回归和Fama-Macbeth回归会得到完全不同的估计结果。

Fama-Macbeth回归很好的解决了截面相关性的问题，但对于时间序列上的相关性仍然无力。

（用一些更通俗的话来说，Fama-Macbeth回归的两步是：

1、估计资产承担风险大小（beta值）。通过对资产收益率的时间序列分析，得到资产承担的风险水平。

2、估计风险溢价时间序列以及统计检验。通过在每个时点的资产收益率对得到的beta值进行截面回归，得到因子在每个时刻的风险溢价。对每个时刻的风险溢价进行平均，并检验均值是否显著异于0.）

三、模型设定

还是从最简单的CAPM模型开始：

$$E(r_i) - r_f = \beta_i [E(r_{Mkt}) - r_f], i = 1, 2, \ldots N$$

其中，$E[r_i]$、$E[r_{Mkt}]$分别指资产预期收益率和市场组合预期收益率，r_f是无风险利率，β_i是资产i对市场风险的因子暴露，$E(r_{Mkt}) - r_f$是市场组合风险溢价。上式说明了任何资产 的超额收益都由其对系统风险的暴露决定，即刻画了单个资产预期超额收益率在截面上和市场贝塔的线性关系。

用R_i表示资产i的超额收益，即$R_i = r_i - r_f$，用λ表示市场组合风险溢价，即$\lambda = E(r_{Mkt}) - r_f$，那么CAPM可以写为下式：

$$E(R_i) = \beta_i' \lambda, i = 1, 2, \ldots, N$$

更一般的，如果有k个因子， λ就是k维因子溢价向量， β_i是资产i在k个因子上的k维因子暴露向量。我们希望通过估计λ来检验模型中资产预期超额收益和因子暴露向量β的线性关系是否稳健，并确定定价误差的大小。

对于上述模型，FM方法分两步进行估计：

1、先对各个资产i进行时间序列回归，估计β_i：

$$R_i = \alpha_i + \beta_i' f_t + \epsilon_t^i, t = 1, 2, \ldots, N$$

2、用第一部得到的$\hat{\beta_i}$作为自变量，在各个时期t分别做截面回归：

$$R_i = \hat{\beta_i}' \lambda_t + \alpha_{it}, i = 1, 2, \dots N$$

最后，对每一期的估计结果$\hat{\lambda_t}$、$\hat{\alpha_{it}}$在时序上取平均值得到λ和α_i的估计：

$$\hat{\lambda} = \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda_t}$$

$$\hat{\alpha_i} = \frac{1}{T} \sum_{t=1}^{T} \hat{\alpha_{it}}$$

$\hat{\lambda}$和$\hat{\alpha_i}$的标准误可以由各个时期截面回归$\hat{\lambda_t}$、$\hat{\alpha_{it}}$的标准差得到：

$$\sigma^2(\hat{\lambda}) = \frac{1}{T} \sum_{t=1}^{T} (\hat{\lambda_t} - \hat{\lambda})^2$$

$$\sigma^2(\hat{\alpha_i}) = \frac{1}{T} \sum_{t=1}^{T} (\hat{\alpha_{it}} - \hat{\alpha_i})^2$$

FM 方法的优势是得到了异方差稳健的标准误，且修正了面板数据在截面上的相关性。此外，由于第二步是每期做一次截面回归，所以允许使用时变的β 做自变量。然而，这种以第一阶段估计量作为第二阶段自变量的方法引入了变量误差问题 (EIV 问题)，且 FM 标准误对残差序列时序相关是不一致的，这就需要对标准误进行修正。

EIV问题的修正

由于以第一阶段估计量作为第二阶段自变量引入了 EIV 问题，FM 标准误是不一致的。Shanken (1992) 给出了对 EIV 的修正项。如果残差项ϵ在时间上独立同分布且与因子收益率独立，均方误估计可以由下式给出 (Cochrane，2005；Shanken，1992)：

$$\sigma^2(\hat{\lambda_{OLS}}) = \frac{1}{T}[(\beta'\beta)^{-1} \sum \beta(\beta'\beta)^{-1}(1 + \lambda' \sum_f \lambda) + \sum_f]$$

其中，$\sum_f$是第一阶段回归因子收益率协方差矩阵$cov(f_t, f_t')$，$\sum$是第一阶段回归残差协方差矩阵$cov(\epsilon_t, \epsilon_t')$

时序相关问题和修正

累了，有空再更

四、Newey-West 调整

Fama-macbeth每一期使用当期因子暴露和个股下一期的收益率进行截面回归，得到因子的收益率；在全部期进行截面回归后，便可得到每个因子收益率的时间序列。将因子收益率在时序上取均值就得到每个因子的预期收益率，而我们关心的是该因子预期收益率是否显著不为零。

对于任何因子，其收益率序列在时序上很可能存在异方差和自相关性，因此在计算其均值标准误的时候需要进行 Newey-West 调整。然而，这和上面的多因子时序回归很不相同。如何进行 Newey-West 调整呢?

简单说明Newey West的原理：

考虑一个线性模型:

$$y = X\beta + \epsilon$$

$$E[\epsilon|X] = 0$$

$$E[\epsilon\epsilon'] = \sigma^2\Omega$$

$$\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\epsilon$$

$$\begin{aligned} Var[\beta] &= E[(\beta - \hat{\beta})(\beta - \hat{\beta})'|X] \\ &= \frac{1}{T}(\frac{1}{T}X'X)^{-1}(\frac{1}{T}X'\sigma^2\Omega X)(\frac{1}{T}(X'X)^{-1}) \end{aligned}$$

当残差不存在异方差和自相关性时，残差协方差阵为单位阵的倍数，回归系数的协方差估计是一致估计量，当残差存在异方差或自相关性时，协方差阵估计有问题，可以通过Newey West调整解决，具体来说是估计上式中的

$$Q = \frac{1}{T}X'\sigma^2\Omega X$$

Newey West调整即对Q进行估计，最终给出的估计量具有一致性，表达式如下，用S表示

$$S = \frac{1}{T}(\sum_{i=1}^{T} e_t^2 x_T x_t' + \sum_{l=1}^{L}\sum_{t=l+1}^{T} w_l e_t e_{t-1}(x_t x_{t-l}' + x_{t-l}x_t'))$$

$$where \quad w_l = 1 - \frac{l}{1+L}$$

上式中，括号中第一项为仅有异方差时的调整，后面一项为针对自相关的调整，其中，e为样本残差，L为计算自相关性影响的最大滞后阶数，w_l是滞后期l的系数，从公式来看，随着滞后期数的增加，影响减小。将S带入系数协方差阵的估计可以得到协方差的Newey West估计量

$$\hat{Var[\beta_{NW}]} = T(X'X)^{-1}S(X'X)^{-1}$$

以上是对于OLS的Newey West调整，对于Fama Macbeth回归，是对已经回归出来的一堆beta系数序列的方差进行调整，跟回归有一定差别，可以做一个转换：用回归出来的所有beta做因变量，1做自变量，做一个回归，这样回归出来的系数是所有beta的均值，残差也捕捉了beta中的异方差性和自相关性，对这个回归方程做newey west即可

Turan Bali、Robert Engle、Scott Murray 三位所著的经典教材 Empirical Asset Pricing, the cross section of stock returns（Bali et al. 2016）提出对于单个因子的收益率序列，将其用 1 作为 regressor 回归得到残差 —— 这相当于用因子收益率减去它在时序上的均值。然后把这个残差和 **X = 1** 代入到 Newey-West 调整中即可。

在这个简化版的Newey-West 调整中，Q的估计S简化为：

$$S = \frac{1}{T}\sum_{t=1}^{T} e_t^2 + 2\sum_{t=1}^{L}\sum_{t=l+1}^{T} w_l e_t e_{t-1}$$

$$where \quad e_t = f_t - E_t[f_t] \quad w_l = 1 - \frac{l}{1+L}$$

其中 f_t代表被检验因子的收益率时间序列，$E_t[f_t]$ 是它在时序上的均值。由于我们仅仅有一个 regressor，因此上述S其实是一个标量。将它代入到 **V**_OLS 的表达式中，在对其开方，就得到 $E_t[f_t]$ 的标准误:

$$s.e.(E_t[f_t]) = \sqrt{S/T}$$

对每个因子依次使用上述修正，获得其各自收益率均值的 standard error，然后就可以计算 t统计量以及 p-value 并检验它们的显著性

五、遗漏变量偏差与mimicking portfolio

六、Stata实现

为简单说明Fama-Macbeth两阶段回归的主要步骤，以下用投资组合数据估计一个简单的 CAPM 模型。数据主要使用了[25 Portfolios Formed on Size and Book-to-Market] 中的 25 个投资组合 1926.7-2020.10 期间的月度收益率(RP.csv)，和[Fama/French 3 Factors] 中的无风险收益、市场超额收益数据(Mkt-RF.csv)。

数据说明：仓库中RP.csv中存储的是25 个投资组合 1926.7-2020.10 期间的月度收益率，每行代表一个月份，每列代表一个投资组合；Mkt-RF.csv存储的是1926.7-2020.10 期间的无风险收益、市场超额收益数据，每行代表一个月份，Mkt-RF和RF列代表市场超额收益率和无风险收益。

数据预处理：

变量	含义
port_num	投资组合编号，1~25
t	时期，如1936m7格式
rpe	超额收益，投资组合收益-无风险收益

第一阶段：

pass1 1930.1-1938.11：25*48次时序回归 （1930.1-1934.12->1933.12-1938.11）

估计$beta_{it}, i = 1, 2...25$，窗口为五年，每次向后移动一个月

```
bys port_num: asreg rp mktrf if (t>=ym(1930,1) & t<=ym(1938,12)) , wind(t 60)
rmse se newey(4)
```

```
. list in 46/50
```

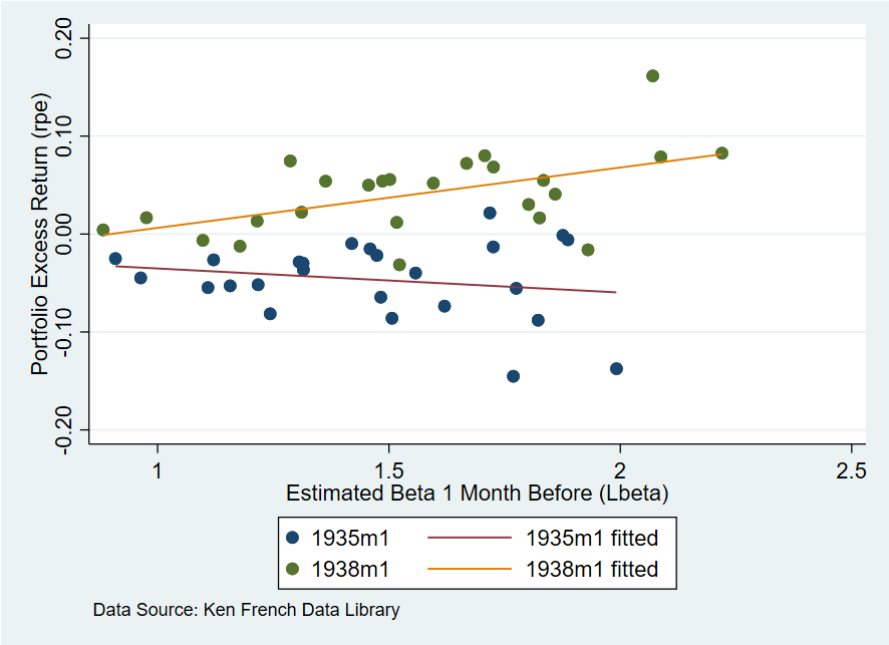
	port_num	t	mktrf	rpe	_rmse	_Nobs	_R2	_adjR2	_b_mktrf	_b_cons	_se_mk~f	_se_cons
46.	1	1938m9	0.01	-0.14	0.16	60	0.42	0.41	1.90	-0.00	0.33	0.01
47.	1	1938m10	0.08	0.12	0.16	60	0.42	0.41	1.91	-0.00	0.33	0.01
48.	1	1938m11	-0.02	-0.04	0.16	60	0.44	0.43	1.97	-0.00	0.38	0.01
49.	1	1938m12	0.04	0.04	0.15	60	0.47	0.46	1.97	0.00	0.38	0.02
50.	2	1934m12	0.00	-0.01	0.22	60	0.48	0.47	1.62	0.03	0.30	0.03

port_num	_b_mktrf	_se_mk~f	_R2	_rmse
1	1.89	0.32	0.48	0.21
2	1.82	0.30	0.52	0.19
3	1.87	0.20	0.72	0.12
4	1.75	0.17	0.70	0.12
5	1.95	0.22	0.65	0.15
6	1.39	0.15	0.66	0.11
7	1.51	0.13	0.77	0.09
8	1.55	0.16	0.79	0.09
9	1.62	0.15	0.80	0.09
10	1.74	0.15	0.75	0.10
11	1.25	0.08	0.81	0.06
12	1.20	0.06	0.90	0.04
13	1.36	0.09	0.92	0.04
14	1.47	0.09	0.86	0.06
15	1.79	0.10	0.89	0.07
16	0.96	0.05	0.92	0.03
17	1.13	0.07	0.94	0.03
18	1.32	0.06	0.93	0.04
19	1.50	0.09	0.91	0.05
20	1.98	0.15	0.86	0.08
21	0.90	0.03	0.97	0.02
22	1.09	0.03	0.98	0.02
23	1.26	0.06	0.95	0.03
24	1.50	0.06	0.92	0.05
25	1.69	0.18	0.78	0.10
Total	1.50	0.13	0.81	0.08

(_b_mkrtf就是beta)

为了截面回归更方便，直接将自变量取滞后项(beta滞后一个月)

在做截面回归之前，先看一下rpe和beta估计值的关系



该图画出了 1935m1 和 1938m1 两个时间节点上投资组合超额收益率 rpe 和上一月 估计值 **Lbeta** 的关系，横轴是 Lbeta，纵轴是 rpe。

接下来使用xtfmb进行第二阶段估计，也可以用asreg fmb，还可以用statsby

```
.   global regvar "rpe Lbeta"

.          *xtfmb
.          xtfmb $regvar
```

```
Fama-MacBeth (1973) Two-Step procedure          Number of obs      =      1200
                                                Num. time periods  =        48
                                                F(  1,    47)      =      0.91
                                                Prob > F           =    0.3437
                                                avg. R-squared     =    0.2567
```

rpe	Fama-MacBeth Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Lbeta	.0122536	.0128116	0.96	0.344	-.01352	.0380272
_cons	-.0029518	.0131503	-0.22	0.823	-.0294067	.0235032

[基于机器学习方法的宏观因子模拟投资组合构建 - 知乎 (zhihu.com)]