量化金融R语言第一讲——定价模型

（一）资产定价模型

一、资本资产定价模型

1. **资本资产定价模型**(Capital Asset Pricing Model, CAPM)：从经济学角度解释资产价格

2. CAPM假设：

- 个体投资者是价格的接受者
- 单阶段的投资界限
- 投资限定于可交易的金融资产
- 没有税收，也没有交易成本
- 信息免费，并且所有的投资者都可以获得
- 投资者是理性的，遵照均值方差最优化的原则进行决策
- 期望同质
  如果以上假设全部成立，所有的投资者都持有相同的风险资产组合。市场的风险溢价取决于所有市场参与者的平均风险厌恶程度。

3. 市场风险溢价与个体证券风险之间的线性关系：

$$E(r_i) - r_f = \beta_i[E(r_m) - f_f]$$

$E(r_i)$是某一确定证券的预期收益率，r_f是无风险收益率，$E(r_m)$是市场组合的预期收益率。

4. 贝塔β_i度量了CAPM的风险，它是这个证券与市场组合协方差与市场组合收益率方差的函数。

$$\beta_i = \frac{Cov_{i,m}}{Var_m}$$

$Cov_{i,m}$是给定证券的收益率和市场组合收益率之间的协方差，Var_m是市场组合收益率的方差。
一方面，贝塔显示了一个股票的收益率相当于市场组合收益率的敏感性。另一方面，某个证券的贝塔还显示了该证券向市场组合增加了多少风险。

5. 市场仅仅在系统风险更高的情况下对股票才会给出更高的收益率，因为非系统风险可以分散，所以不会为它支付风险溢价。

6. **证券市场线**(Security Market Line,SML)：

$$E(r_i) = r_f + \beta_i[E(r_m) - f_f]$$

当市场均衡时,每个证券都应该位于SML上。

二、贝塔估计

证券对因子的敏感性可以通过价格的历史运动进行估计，将从单因素模型中估计贝塔。
1. 获得数据

下载某只股票(GOOGLE)及市场指数(标普500)在一段时间的价格数据，下载一个月美元libor利率作为无风险收益率标的。

完整代码见R文件

```
library(quantmod)
# GOOGLE
getSymbols("GooG", from = '2009-06-01', to = '2013-06-01')
# 标普500
getSymbols("^GSPC", from = '2009-06-01', to = '2013-06-01')
# 一个月美元LIBOR利率
getSymbols("USD1MTD156N", from = '2009-06-01', to = '2013-06-01', src = 'FRED')
LIBOR = na.omit(USD1MTD156N)
```

2. 计算收益率
计算股票以及市场指数的对数收益率：

$$r_{it} = ln(\frac{s_t}{s_{t-1}})$$

通过减去无风险的日对数收益率r_{ft}来确定风险溢价：

$$r_{ft} = ln(1 + \frac{USDLIBOR}{36000} * ((t+1) - t))$$

$$R_{it} = r_{it} - r_{ft}$$

```
#获取指标共同日期
NG = merge(G, sp500, by = 'date')
new3 = merge(NG, LIBOR, by = 'date')
#求对数收益率
logreturn = function(x){
  log(tail(x, -1)/head(x, -1))
}
# 计算一个月美元LIBOR利率日对数收益率的计算方法
rft = log(1 + head(new3$USD1MTD156N, -1)/36000*diff(as.numeric(new3$date)))
```

3. 计算β

   使用风险溢价取代收益率，计算β。

```
cov((logreturn(new3$GOOG.Close) - rft), (logreturn(new3$GSPC.Adjusted) - rft))/var(logretur
n(new3$GSPC.Adjusted) - rft)
```

4. 基于线性回归估计β

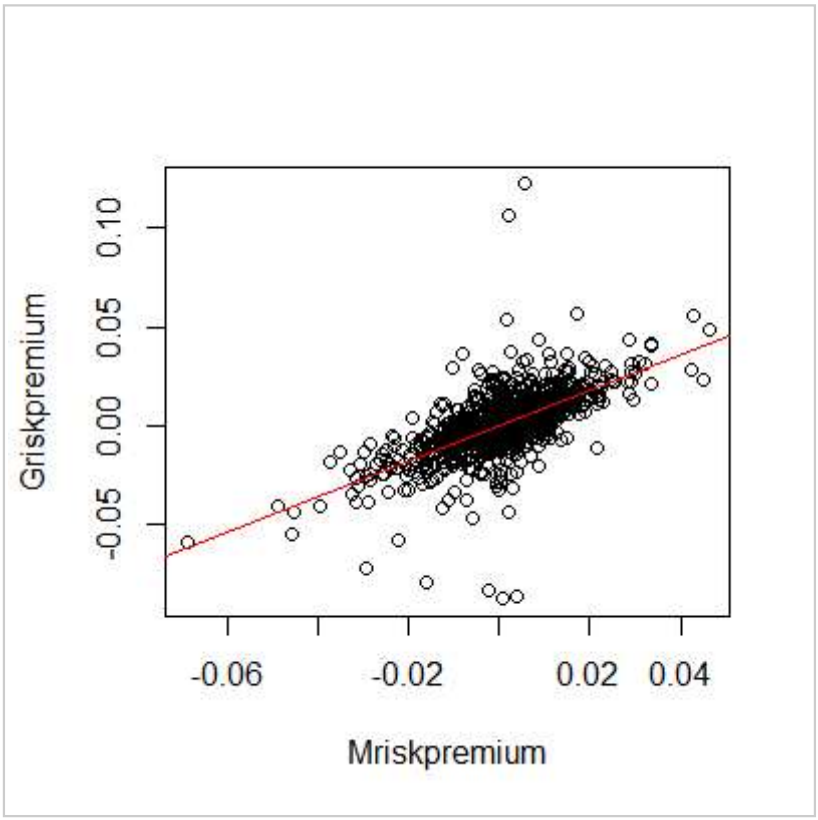
   线性回归的解释变量是市场风险溢价，因变量是证券的风险溢价。 回归方程的形式如下：

$$R_i = \alpha_i + \beta_i R_m + e_i$$

   我们将使用普通最小二乘估计来确定线性回归模型。特征线的截距是α，股票的收益率中不能被市场因素解释的部分。函数的斜率表示对市场因子的敏感性，通过贝塔来度量。

```
# 证券风险溢价
Griskpremium = logreturn(new3$GOOG.Close) - rft
#市场风险溢价
Mriskpremium = logreturn(new3$GSPC.Adjusted) - rft
fit <- lm(Griskpremium ~ Mriskpremium)
```

可以发现，两种方法估计的β一样。



根据CAPM，α等于0，假设α_i为0。

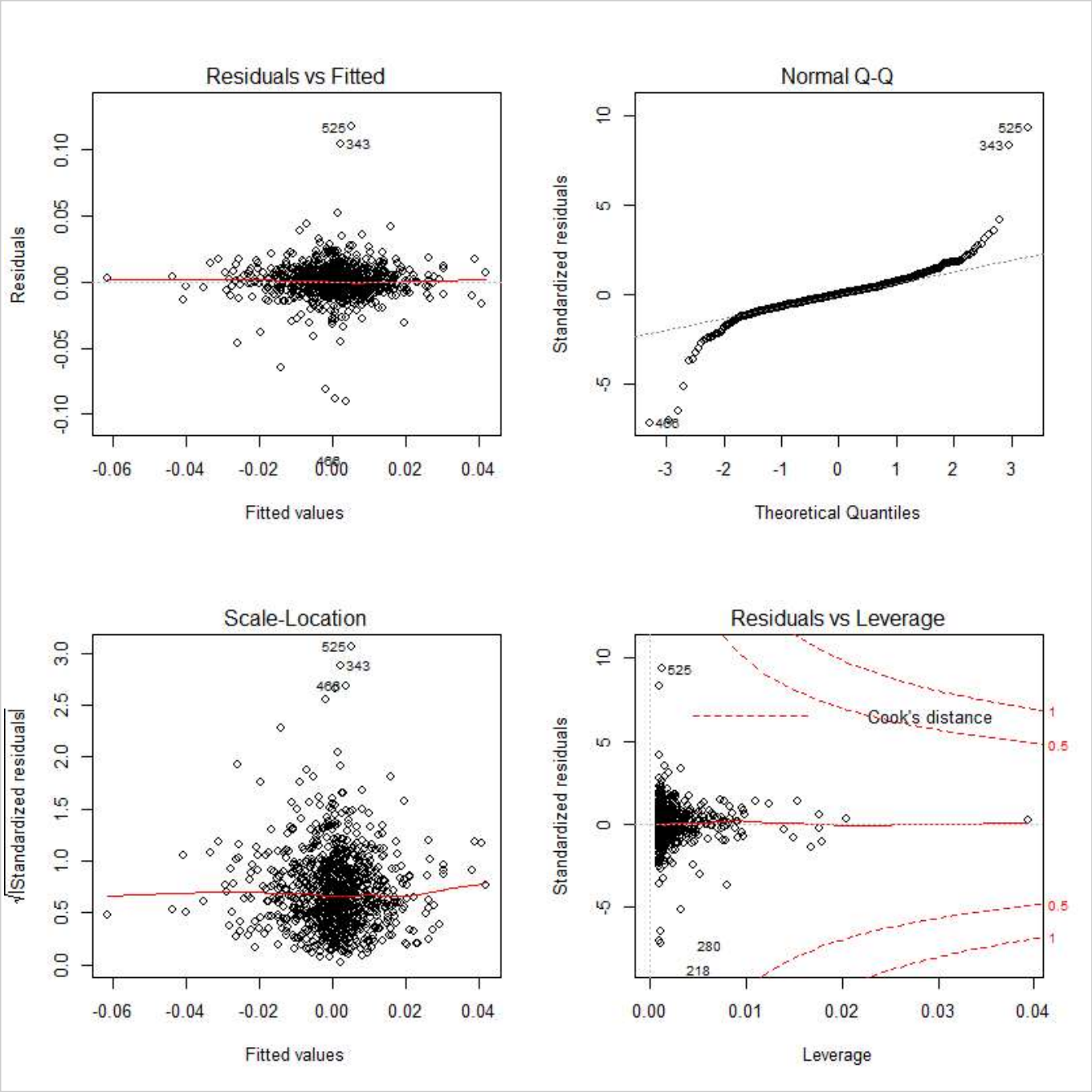
```
fit2 <- lm(Griskpremium ~ -1 + Mriskpremium)
```

```
Call:
lm(formula = Griskpremium ~ -1 + Mriskpremium)

Residuals:
      Min        1Q    Median        3Q       Max
-0.089774 -0.005531  0.000142  0.005444  0.117111

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
Mriskpremium  0.89707    0.03517    25.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0125 on 985 degrees of freedom
Multiple R-squared:  0.3977,    Adjusted R-squared:  0.3971
F-statistic: 650.5 on 1 and 985 DF,  p-value: < 2.2e-16
```

高的F-statistic值说明模型具备解释能力，贝塔证明是显著的。并且，贝塔为0的原假设被拒绝。

```
Call:
lm(formula = Griskpremium ~ Mriskpremium)

Residuals:
      Min         1Q     Median         3Q        Max
-0.089996 -0.005748 -0.000081  0.005229  0.116891

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0002253  0.0003987   0.565    0.572
Mriskpremium 0.8961131  0.0352266  25.439   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01251 on 984 degrees of freedom
Multiple R-squared:  0.3967,    Adjusted R-squared:  0.3961
F-statistic: 647.1 on 1 and 984 DF,  p-value: < 2.2e-16
```

如果我们通过放松α为0的假设来运行检验，可以看到截距与0没有显著的区别。



三、模型检验

检验股票风险溢价和β是否有显著关系。 关于贝塔-收益率关系的第一个检验使用了两阶段线性回归。第一个回归估计证券特征线以及单个证券的贝塔。在第二个回归中，证券风险溢价是因变量，而贝塔是解释变量。

1. 数据收集

   收集了69只股票2003-2007年的月度收益率。此处略去具体过程。

2. 对SCL建模

使用股票收益率的时间序列，我们能对每个证券计算贝塔。因此，能得到样本均值表示的风险溢价向量和一个包含贝塔的向量。

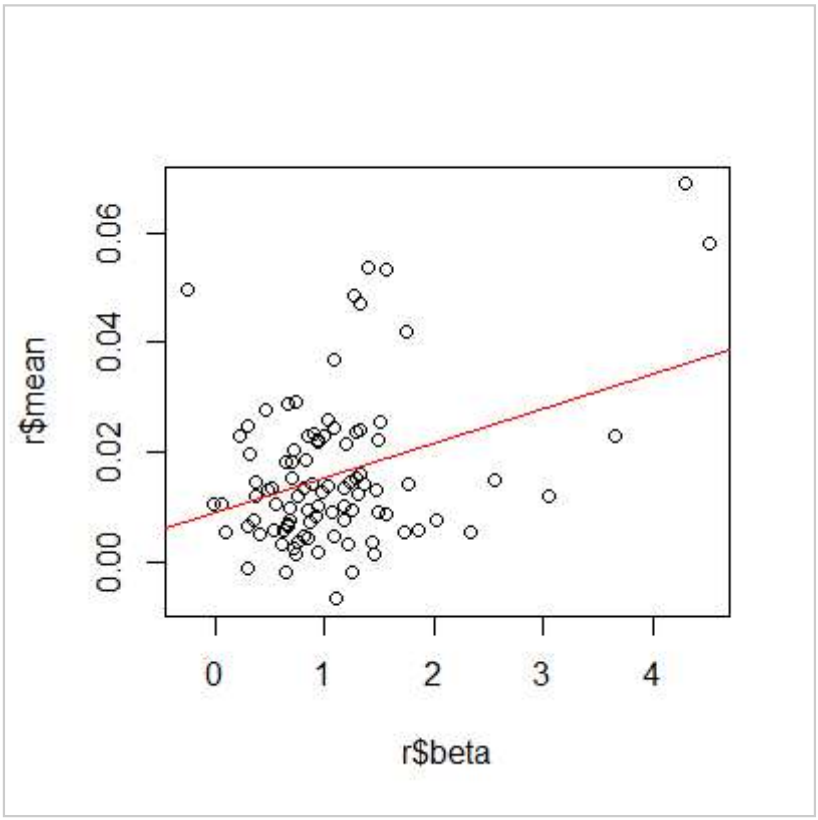
$$\overline{R}_l = \gamma_0 + \gamma_1 \beta_i$$

```
# 分别求69个股票超额收益率
resss = apply(ress, 2, logreturn)
riskpremium = resss - rft
# SP500超额收益
premium500 = new500_return - rft
# 计算beta和股票差额收益的均值mean
r = t(sapply(symbols, function(symbol)
c(beta = lm(riskpremium[, symbol] ~ premium500[, 1])$coefficients[[2]], mean = mean(resss[,
symbol]))))
```

因此，通过迭代所有的股票代码，我们来画出贝塔计算值的返回列表以及风险溢价的均值。

```
r = as.data.frame(r)
plot(r$beta, r$mean)
abline(lm(r$mean ~ r$beta), col = 'red')
summary(lm(r$mean ~ r$beta))
```

```
Call:
lm(formula = r$mean ~ r$beta)

Residuals:
      Min         1Q     Median         3Q        Max
-0.022635 -0.008878 -0.003006  0.006557  0.042208

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.009052   0.002126   4.258 4.69e-05 ***
r$beta      0.006322   0.001573   4.019 0.000114 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01268 on 99 degrees of freedom
Multiple R-squared:  0.1402,    Adjusted R-squared:  0.1316
F-statistic: 16.15 on 1 and 99 DF,  p-value: 0.0001142
```



根据检验，不能拒绝贝塔-收益率的关系。

3. 检验个体方差的解释能力
   检验可以进一步地深化，包括把非系统风险作为第二个解释变量的检验。证券的个体风险可以如下计算：

$$\sigma_{ei}^2 = \sigma_i^2 - \beta_i^2 \sigma_m^2$$

   因此，必须先计算方差的向量，接着得到单个方差的向量。回归方程如下估计：

$$\overline{R}_l = \gamma_0 + \gamma_1 \beta_i + \gamma_2 \sigma_{ei}^2$$

   更新循环。

```
k = t(sapply(symbols, function(symbol){
  stock <- riskpremium[, symbol]
  beta = lm(stock ~ premium500[, 1])$coefficients[[2]]
  c(beta = beta,
    mean = mean(stock),
    risk = var(stock)- beta^2*var(New500)
    )}))
k <- as.data.frame(k)
```

现在，模型如下图：

```
Call:
lm(formula = k$mean ~ k$beta + k$risk)

Residuals:
      Min         1Q     Median         3Q        Max
-0.020740 -0.008307 -0.002167  0.006941  0.038907

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.348e-02  3.192e-03   4.222 5.41e-05 ***
k$beta      -4.589e-03  4.250e-03  -1.080   0.2830
k$risk      -1.487e-07  5.407e-08  -2.750   0.0071 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01228 on 98 degrees of freedom
Multiple R-squared:  0.2018,    Adjusted R-squared:  0.1855
F-statistic: 12.39 on 2 and 98 DF,  p-value: 1.594e-05
```

新模型中β的回归系数是负值，在99.9%的显著性水平上，风险参数是显著的，不能拒绝β_2为0的原假设。

（二）因素模型

一、套利定价理论

1. **套利定价理论**(Arbitrage Pricing Theory,APT)：
   用于确定不同证券的收益率；
   均衡时不存在套利机会，并且一个资产的收益率是多个随机因素的线性组合。
   这些因素可以是多种宏观经济因素，也可以是市场指数，都有一个对应的贝塔系数。

$$r_i = E(r_i) + \sum_{j=1}^{n} \beta_{ij} F_j + e_i$$

$E(r_i)$是资产i的预期收益率，F_j表示第j个因素的非预期变动，β_{ij}是证券i对因素j的敏感程度，e_i表示非预期的公司特定事件引起的收益。因此，$\sum_{j=1}^{n}\beta_{ij}F_j$表示随机系统影响系统性风险可以分散化，$e_i$表示非系统影响，即总影响中无法被系统因素捕捉的那部分。$\sum_{j=1}^{n}\beta_{ij}F_j$和$e_i$都具有无条件的零均值。

2. APT的其他重要假设：
    - 市场中存在有限多个投资者，每个投资者为了下一期而做最优化组合选择。他们拥有相同的信息，并且都没有市场影响力。
    - 市场中存在一个无风险资产和无穷多个连续交易的风险资产。因此，公司特定风险可以通过分散化完全消除。一个公司特定风险为零的组合称为完全分散化的组合。
    - 投资者是理性的，这意味着市场中如果出现套利机会（金融资产相互之间发生错误定价），那么投资者会迅速买入低估证券卖出高估证券，并为了尽可能多获取无风险收益而持有无穷大的头寸。因此，任何错误定价会瞬间消失。
    - 因素组合存在，并可连续交易。

3. **因素组合**(factorportfolio)：一个因素组合是一个充分分散的投资组合，它仅仅对一个特定因素贝塔为1，对其他所有因素的贝塔都为0。
   假设存在因素组合，任何充分分散的投资组合的风险溢价，等于因素组合的风险溢价的加权和。 如果这对每一个充分分散的投资组合都成立，单个证券的预期收益率可以通过因素的风险溢价(RP_j)建立以及对这个因素的敏感性β_{ij}组合而成。

$$E[r_i] = r_f + \sum_{j=1}^{n}\beta_{ij}RP_j$$

4. 两因素模型:

$$E(r_i - r_f) = \beta_{i1}(r_1 - r_f) + \beta_{i2}(r_2 - r_f)$$

5. 线性回归方程式：

$$r_i - r_f = \alpha_i + \beta_{i1}(r_1 - r_f) + \beta_{i2}(r_2 - r_f) + \epsilon_i$$

6. **指数模型**：APT中仅有一个因素，市场组合的收益率

7. CAPM与APM的差异：
    - CAPM是一个均衡模型，建立于经济学的思考，而APT是一个统计模型，使用了套利观点。

- 在APT理论中，如果投资者有一个充分分散的投资组合，那么可以在实际中通过持有大量资产来建立这个组合，进而可以给出预期收益率——贝塔关系。而CAPM 则不能建立所谓的市场组合。
- CAPM说明，每个证券都成立预期收益率-贝塔关系，而APT说明，几乎每个证券都成立这种关系。
- 如果市场中存在错误定价，APT认为，只要有一部分投资者改变投资组合结构就可以，而CAPM认为，每个投资者都会这么做。

8. APT实现
   - （1）识别因素：因为APT本身不包含关于因素的任何内容，所以因素需要通过实证分析来识别。这些因素通常考虑宏观经济因素，如股票市场收益率、通货膨胀率、商业周期等。使用宏观经济因素的一大问题是各个因素相互不独立。因此常常需要使用因子分析来识别因素。但是，通过因子分析识别出的因素在经济学上不容易有好的解释。
   - （2）估计因素系数：为了估计多变量线性回归模型的系数，我们使用两因素模型的一个一般形式。（$r_i - r_f = \alpha_i + \beta_{i1}(r_1 - r_f) + \beta_{i2}(r_2 - r_f) + \epsilon_i$)
   - （3）估计因素溢价：因素溢价基于历史数据来估计，对因素组合溢价的历史时间序列数据取均值。
   - （4）给出APT定价方程：通过代入适合的变量，用两因素模型来计算任何资产的预期收益率。

9. Fama-French三因素模型

   Fama和French在1996年提出一个多因素模型。他们使用公司指标因素替代宏观因素，因为他们发现这些因素能够更好地描述资产的系统风险。他们向市场组合收益率中增加了公司规模和净值市值比作为收益率生成因素，扩展了指数模型。
   公司规模因素定义为小公司与大公司的收益率之差(r_{SMB});
   净值市值比因素定义为高净值市值比减去低净值市值比的公司收益率之差(r_{HML})。
   模型如下：

$$(r_i - r_f) = \alpha_i + \beta_{iM}(r_M - r_f) + \beta_{iHML} r_{HML} + \beta_{iSMB} r_{SMB} + e_i$$

二、在R中建模

1. 数据选择

   在这里，我们仅仅讲述股票价格时间序列和其他相关信息如何获取和如何用于因素模型估计。使用quantmod包来收集数据库。

```
library(quantmod)
stocks <- stockSymbols()
```

现在，我们得到一个数据框，这个数据框涵盖了在各个交易所（如AMEX、NASDAQ、NYSE）进行交易的大约6500只股票。为了查看数据集涵盖的变量，我们用str命令：

```
str(stocks)
```

```
> str(stocks)
'data.frame':   6941 obs. of  8 variables:
 $ Symbol   : chr  "AAMC" "AAU" "ACU" "ACY" ...
 $ Name     : chr  "Altisource Asset Management Corp" "Almaden Minerals, Ltd." "Acme United Corporation."
 "AeroCentury Corp." ...
 $ LastSale : num  11.5 0.59 22.58 7.27 34.28 ...
 $ MarketCap: chr  "$18.28M" "$65.92M" "$75.69M" "$11.24M" ...
 $ IPOyear  : int  NA 2015 1988 NA NA NA 2018 2014 NA NA ...
 $ Sector   : chr  "Finance" "Basic Industries" "Capital Goods" "Technology" ...
 $ Industry : chr  "Real Estate" "Precious Metals" "Industrial Machinery/Components" "Diversified Commerci
al Services" ...
 $ Exchange : chr  "AMEX" "AMEX" "AMEX" "AMEX" ...
```

Alt text

我们也会需要无风险收益率的时间序列。我们通过计算月度LIBOR市场上的美元利率确定这个序列：

```
library(Quandl)
LIBOR <- Quandl('FED/RILSPDEPM01_N_B',start_date = '2010-06-01', end_date = '2014-06-01')
```

我们可以删掉确实不需要的变量，也可以增加来自不同数据库的市场资本化信息和公司账面价值作为新变量，这些变量我们需要用来估计Fama-French模型：

```
stocks[1:5, c(1, 3:4, ncol(stocks))]
```

我们有一张表，记录了接近5000只股票价格在2010年6月1日到2014年6月1日之间的时间序列。其中第一列和最后几列如下所示：

```
d <- read.table("data.csv", header = TRUE, sep = ";", colClasses = c("Date", rep("numeric",
4014)))
d[1:7, c(1:5, (ncol(d) - 6):ncol(d))]
```

```
> d[1:7, c(1:5, (ncol(d) - 6):ncol(d))]
         Date   SP500  AAU   ACU   ACY   YPF
1 2010-06-01 1070.71 0.96 11.30 20.64 35.50
2 2010-06-02 1098.38 0.95 11.70 20.85 35.60
3 2010-06-03 1102.83 0.97 11.86 20.90 35.40
4 2010-06-04 1064.88 0.93 11.65 18.95 35.60
5 2010-06-07 1050.47 0.97 11.45 19.03 35.66
6 2010-06-08 1062.00 0.98 11.35 18.25 36.75
7 2010-06-09 1055.69 0.98 11.90 18.35 37.97
    YUM   YZC   ZEP   ZMH   ZNH  ZQK
1 40.49 21.91 17.20 54.17 21.55 4.45
2 41.41 22.55 17.21 55.10 21.79 4.65
3 41.92 22.00 17.37 55.23 21.63 4.63
4 40.86 20.95 17.11 53.18 20.88 4.73
5 39.77 20.24 16.79 52.66 20.24 4.18
6 40.89 20.57 16.85 52.99 20.96 3.96
7 41.21 20.62 16.86 53.22 20.45 4.02
```

现在，我们得到了所需数据：市场组合（标普500）、股票价格、无风险利率（月度LIBOR）。 为了清洗数据库，我们删除了有缺失值、0价格或者负价格的变量。最简单的实现如下：

```
d <- d[, colSums(is.na(d)) == 0]
d <- d[, c(T, colMins(d[, 2:ncol(d)]) > 0)]
```

为了使用colMins函数，需要应用matrixStats包。下面，我们可以开始处理数据。

2. 通过主成分分析估计APT
   由于识别影响证券收益率的宏观变量很困难，我们在实践中使用因子分析相当不易。通常，我们通过主成分分析来寻找驱动收益率变动的潜因子。 在最初下载的6500只股票中，我们可以使用4500只股票数据。其他部分由于缺失值或者0价格的缘故删除了。现在，由于这里用不到日期，还有我们将标普500本身视为一个独立的因子，在主成分分析（PCA）中不考虑，我们又删除了前两列数据。然后，计算对数收益率：

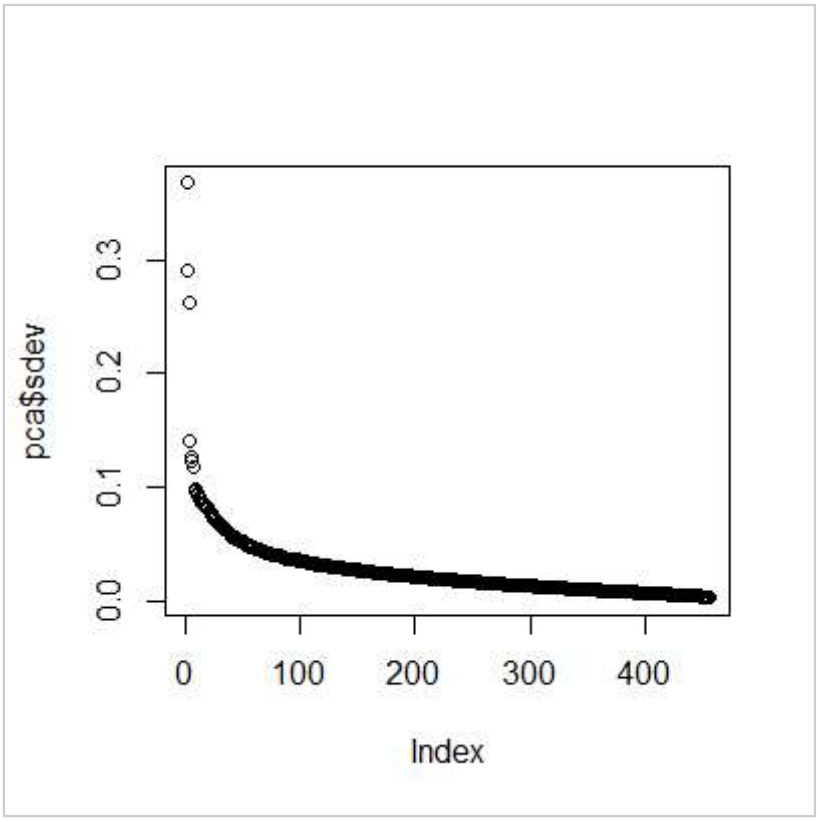
```
p <- d[, 3:ncol(d)]
r <- log(p[2:nrow(p), ] / p[1:(nrow(p) - 1), ])
```

因为我们的股票数目过于庞大，为了实施PCA，要么需要数据时间长度至少25年，要么需要减少股票数量。使因素模型在几十年间保持稳定，这是不可能的。因此，为了达到说明的目的，我们随机选择了百分之十的股票，并对这个样本计算PCA模型：

```
r <- r[, runif(nrow(r)) < 0.1]
pca <- princomp(r)
```

结果，我们得到一个princomp类对象。这个对象有8个属性，其中最重要的属性是加载矩阵和sdev属性（包括组成部分的标准差）。第一主成分是数据集方差最大的向量。 我们确认一下主成分的标准差：

```
plot(pca$sdev)
```



结果如图所示。我们可以看出，前5个成分是独立的。因此，我们应该选择5个因子。但是，其他因子的标准差同样显著。所以，不能通过几个因子解释整个市场。 我们可以通过调用factanal函数确认结果。这个函数估计了五因子模型：

```
factanal(r, 5)
```

因子分析与PCA相关，但从数学角度看稍复杂一些。结果，我们得到一个factanal类对象。它有很多属性。但是，此时我们仅对以下输出部分有兴趣：

```
                Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings      61.206  28.749  11.283   9.604   4.502
Proportion Var    0.158   0.074   0.029   0.025   0.012
Cumulative Var    0.158   0.232   0.261   0.286   0.297

Test of the hypothesis that 5 factors are sufficient.
The chi square statistic is 90629.74 on 73148 degrees of freedom.
The p-value is 0
```

结果显示，五因子模型适合数据。但是，可解释的方差仅仅接近30%。这意味着模型应该考虑扩展进其他因子。

3.Fama-French模型估计

我们有一个包含4015只股票5年价格的数据框架，和包含LIBOR时间序列的LIBOR数据框架。首先，我们需要计算收益率，再与LIBOR利率合并。 第一步，我们删掉数学计算不需要的数据。然后，对保留下来的每一列计算对数收益率：

```
d2 <- d[, 2:ncol(d)]
d2 <- log(tail(d2, -1)/head(d2, -1))
```

在计算了对数收益率之后，我们把日期放回到收益率中。然后，在最后一步合并两个数据集：

```
d <- cbind(d[2:nrow(d), 1], d2)
d <- merge(LIBOR, d, by = 1)
```

结果如下：

```
> print(d[1:5, 1:5])
        Date Value         SP500          AAU
1 2010-06-02   0.4  0.025514387 -0.01047130
2 2010-06-03   0.4  0.004043236  0.02083409
3 2010-06-04   0.4 -0.035017487 -0.04211149
4 2010-06-07   0.4 -0.013624434  0.04211149
5 2010-06-08   0.4  0.010916240  0.01025650
            ACU
1  0.034786116
2  0.013582552
3 -0.017865214
4 -0.017316450
5 -0.008771986
```

我们将LIBOR利率转换为日度收益率：

```
d$LIBOR <- d[,2] / 36000
```

由于LIBOR利率的报价基于货币市场基差——以及（实际天数/360）的天数计算约定——而且时间序列包含使用百分比表示的利率，我们把LIBOR除以36000。现在，我们需要计算Fama-French模型的3个变量。正如在数据选择部分中所讲，我们有股票的数据框：

```
> d[1:5, c(1,(ncol(d) - 3):ncol(d))]
        Date           ZMH           ZNH
1 2010-06-02  0.017022466  0.011075332
2 2010-06-03  0.002356568 -0.007369909
3 2010-06-04 -0.037823898 -0.035289477
4 2010-06-07 -0.009826232 -0.031130919
5 2010-06-08  0.006247062  0.034955015
            ZQK        LIBOR
1  0.043963123 1.111111e-05
2 -0.004310352 1.111111e-05
3  0.021368334 1.111111e-05
4 -0.123613956 1.111111e-05
5 -0.054067221 1.111111e-05
```

删掉那些没有价格数据的股票：

```
stocks = stocks[stocks$Symbol %in% colnames(d),]
```

我们将市场上限作为一个变量。我们仍需对每只股票计算账面市值比：

```
stocks$BookToMarketRatio <- stocks$BookValuePerShare / stocks$LastSale
str(stocks)
```

现在，我们需要计算SMB因素和HML因素。为了简化，我们将BIG公司定义为大于平均水平的公司，并定义SMB因素、HML因素：

```
#SMB
avg_size <- mean(stocks$MarketCap)
BIG   <- as.character(stocks$Symbol[stocks$MarketCap > avg_size])
SMALL <- as.character(stocks[stocks$MarketCap < avg_size,1])
d$SMB <- rowMeans(d[,colnames(d) %in% SMALL]) - rowMeans(d[,colnames(d) %in% BIG])
#HML
avg_btm <- mean(stocks$BookToMarketRatio)
HIGH <- as.character(stocks[stocks$BookToMarketRatio > avg_btm, 1])
LOW <- as.character(stocks[stocks$BookToMarketRatio < avg_btm, 1])
d$HML <- rowMeans(d[, colnames(d) %in% HIGH]) - rowMeans(d[, colnames(d) %in% LOW])
#第三个因素
d$Market <- d$SP500 - d$LIBOR
```

定义完3个因素，我们在花旗集团（Citi）股票和伊克塞利克斯（EXEL）股票上试一试：

```
d$C   <- d$C - d$LIBOR
model <- glm( formula = "C ~ Market + SMB + HML" , data = d)
```

GLM（general linear model，广义线性模型）函数将数据和公式作为参数读入。公式是一个形为"响应～条件"的字符串，其中响应是数据框中的一个变量名，条件指定了模型中的预测子，它包含在数据集中通过操作符"+"分隔开的变量名。这个函数也可以用于Logistic回归，只是缺省状态设定为线性。 模型输出如下：

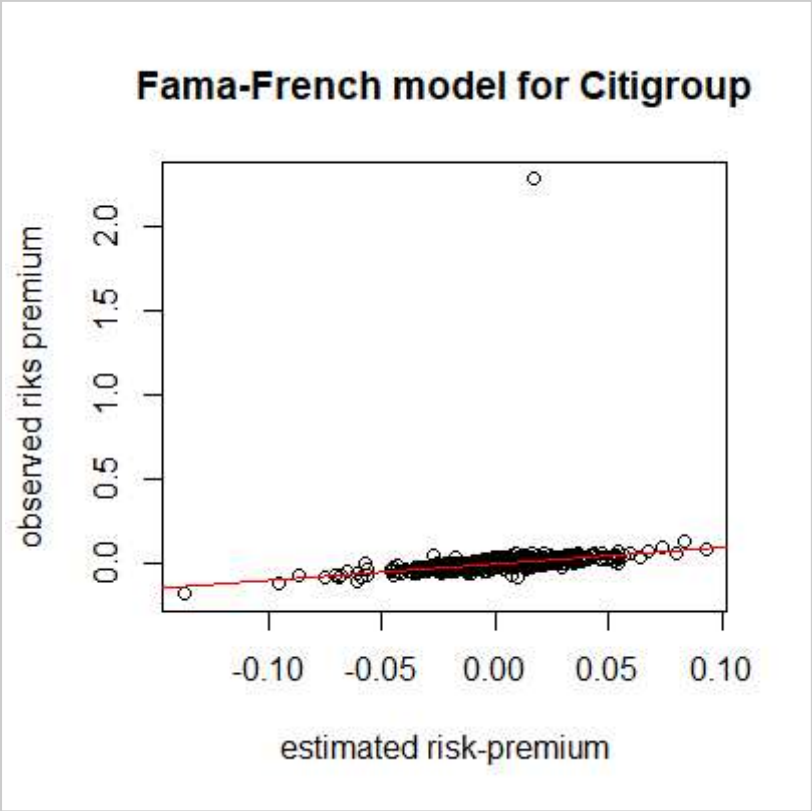
```
> model

Call:  glm(formula = "C ~ Market + SMB + HML", data = d)

Coefficients:
(Intercept)       Market          SMB
   0.002583     2.322486     0.336115
        HML
   2.912915

Degrees of Freedom: 1001 Total (i.e. Null);  998 Residual
Null Deviance:      5.74
Residual Deviance: 5.326        AIC: -2394
```

```
> summary(model)

Call:
glm(formula = "C ~ Market + SMB + HML", data = d)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.09573  -0.01051  -0.00266   0.00567   2.26257

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.002583   0.002350   1.099  0.27188
Market      2.322486   0.275166   8.440  < 2e-16 ***
SMB         0.336115   0.654617   0.513  0.60775
HML         2.912915   1.061474   2.744  0.00617 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.005337046)

    Null deviance: 5.7397  on 1001  degrees of freedom
Residual deviance: 5.3264  on  998  degrees of freedom
AIC: -2394

Number of Fisher Scoring iterations: 2
```

结果显示，唯一显著的因素是市场溢价。这表明花旗集团的股票收益率倾向于与整个市场本身共同变动。

使用以下命令可以绘出了对花旗集团的Fama-French模型估计的风险溢价：

```
estimation <- model$coefficients[1]+
              model$coefficients[2] * d$Market +
              model$coefficients[3]*d$SMB +
              model$coefficients[4]*d$HML
plot(estimation, d$C, xlab = "estimated risk-premium",ylab = "observed riks premium",main =
"Fama-French model for Citigroup")
lines(c(-1, 1), c(-1, 1), col = "red")
```



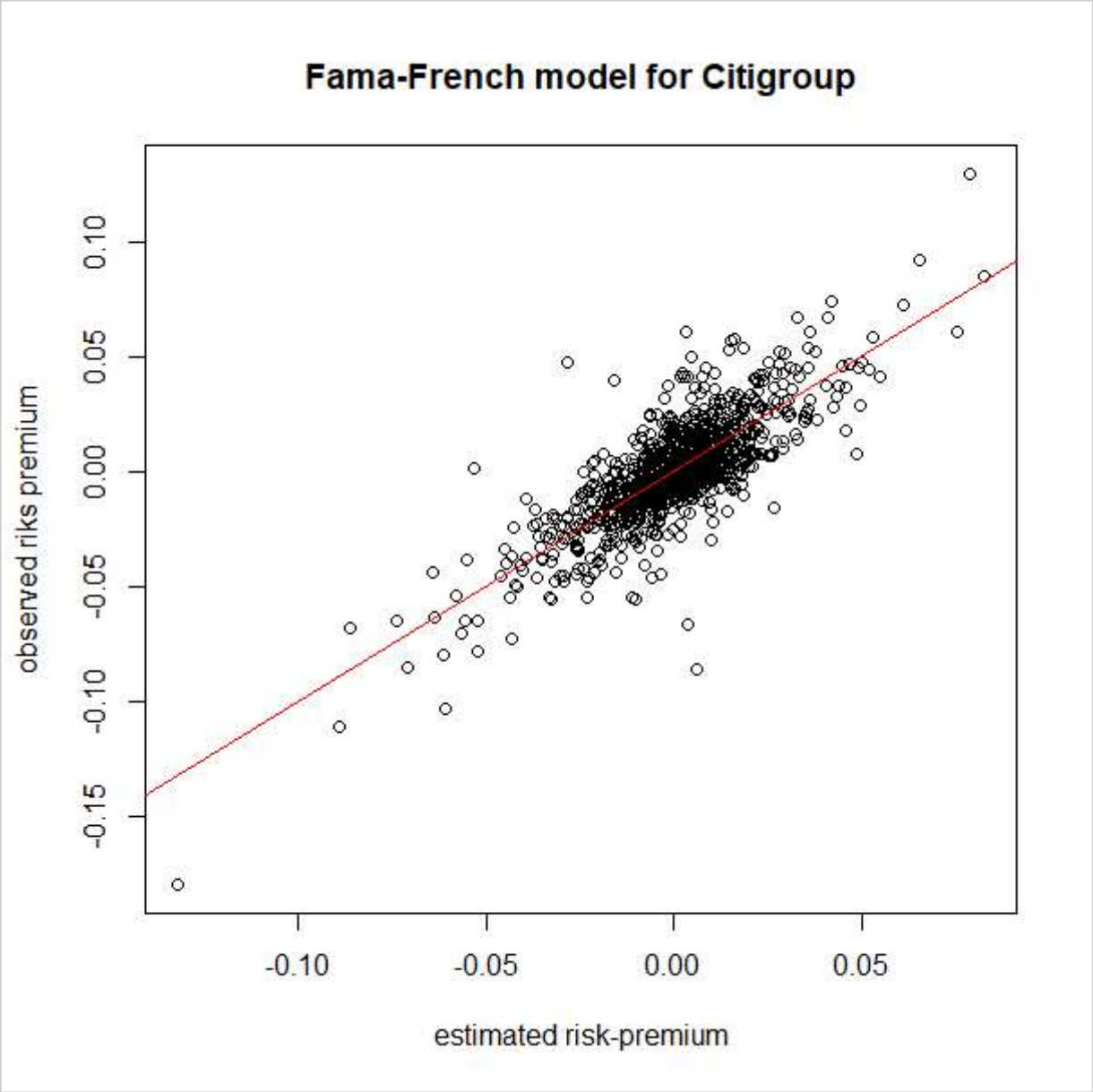
看图可以发现，收益率中存在一个异常值。我们将这个异常值设为0，看看不考虑它之后的结果：

```
outlier <- which.max(d$C)
d$C[outlier] <- 0
model_new <- glm( formula = "C ~ Market + SMB + HML" , data = d)
```

```
glm(formula = "C ~ Market + SMB + HML", data = d)

Deviance Residuals:
      Min          1Q      Median          3Q
-0.091733  -0.007827  -0.000633   0.007972
      Max
 0.075853

Coefficients:
              Estimate Std. Error t value
(Intercept) -0.0000864  0.0004498  -0.192
Market       2.0726607  0.0526659  39.355
SMB          0.4275055  0.1252917   3.412
HML          1.7601956  0.2031631   8.664
            Pr(>|t|)
(Intercept) 0.847703
Market       < 2e-16 ***
SMB         0.000671 ***
HML          < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.0001955113)

    Null deviance: 0.55073  on 1001  degrees of freedom
Residual deviance: 0.19512  on  998  degrees of freedom
AIC: -5707.4

Number of Fisher Scoring iterations: 2
```

根据以上结果，所有3个因素均显著。GLM函数不返回R_2。lm函数对线性回归同样可以得到精确值。我们可以从模型总结中读出r.squared = 0.6446。 结果显示，变量可以解释花旗集团风险溢价中超过64%的变动。

```
estimation_new <- model_new$coefficients[1]+
                  model_new$coefficients[2] * d$Market +
                  model_new$coefficients[3]*d$SMB +
                  model_new$coefficients[4]*d$HML
dev.new()
plot(estimation_new, d$C, xlab = "estimated risk-premium",ylab = "observed riks premium",ma
in = "Fama-French model for Citigroup")
lines(c(-1, 1), c(-1, 1), col = "red")
```

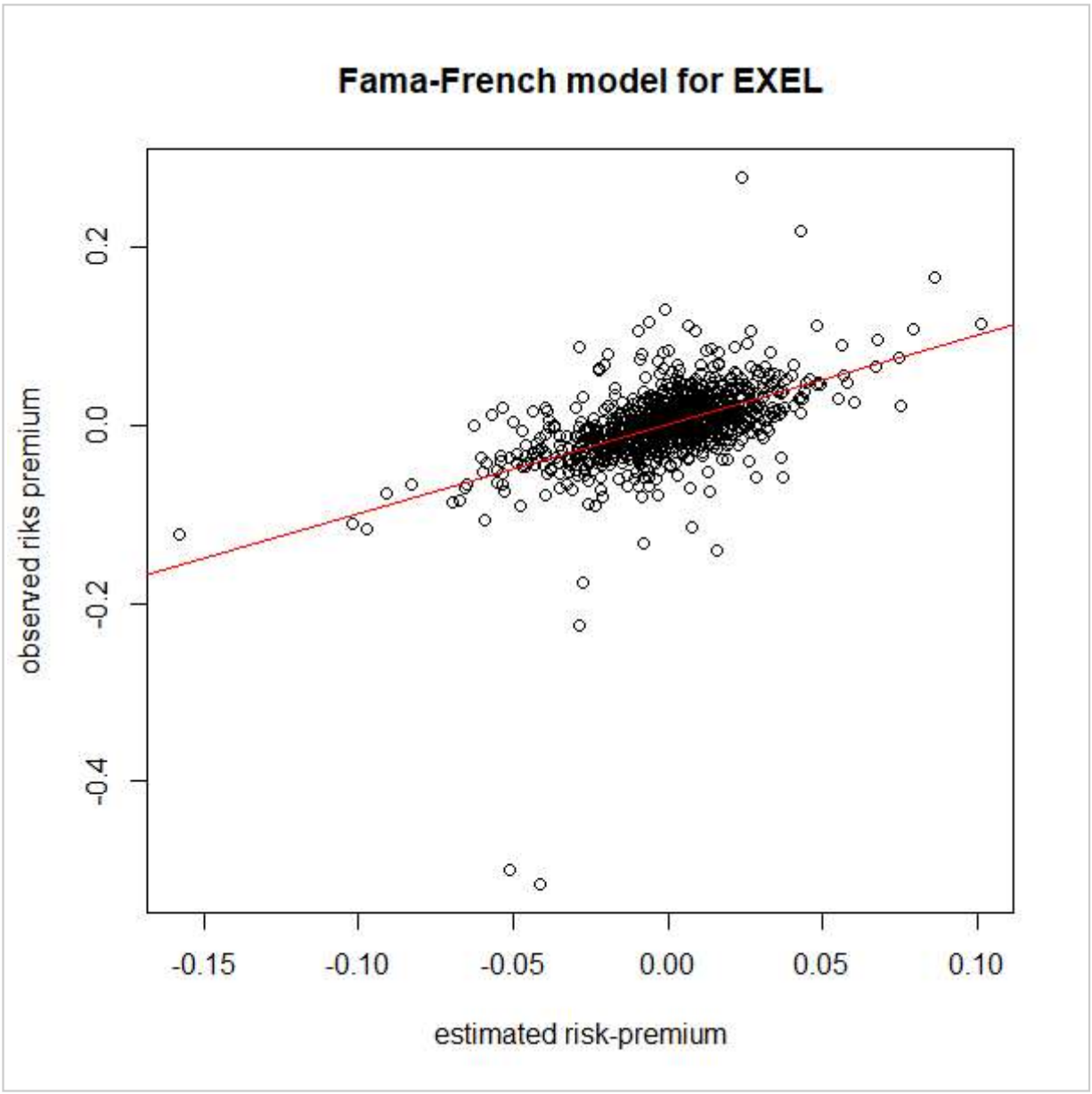


同理，我们再检验另一只股票EXEL，模型小结输出如下：

```
d$EXEL <- d$EXEL - d$LIBOR
model2 <- glm( formula = "EXEL~Market+SMB+HML" , data = d)
```

```
Call:
glm(formula = "EXEL~Market+SMB+HML", data = d)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.47367  -0.01480  -0.00088   0.01500   0.25348

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.001773   0.001185  -1.495  0.13515
Market       1.843306   0.138801  13.280  < 2e-16 ***
SMB          2.939550   0.330207   8.902  < 2e-16 ***
HML         -1.603046   0.535437  -2.994  0.00282 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.001357998)

    Null deviance: 1.8681  on 1001  degrees of freedom
Residual deviance: 1.3553  on  998  degrees of freedom
AIC: -3765.4

Number of Fisher Scoring iterations: 2
```

根据以上结果，所有3个因子均显著。同样地，用lm函数对线性模型得到精确的结果——r.squared = 0.2723，即我们认为变量可以解释EXEL风险溢价的超过27%的变动。并绘出EXEL的Fama-French模型。


Fama-French model for EXEL
observed riks premium
0.2
0.0
-0.2
-0.4
-0.15
-0.10
-0.05
0.00
0.05
0.10
estimated risk-premium