Lecture 3-d
OLS – Goodness of Fit, and Introduction to MLE

Brooks (4th edition): Chapters 3 & 4

[1]

Review: OLS – Summary

• *Classical linear regression model* (CLM) - Assumptions:

(**A1**) DGP: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is correctly specified.

(**A2**) $E[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$

(**A3**) $Var[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2 \mathbf{I}_T$

(**A4**) $\mathbf{X}$ has full column rank –rank($\mathbf{X}$) = k, where $T \geq k$.

• Objective function: $S(x_i, \boldsymbol{\beta}) = \sum_i \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$

First order condition: $-2\,\mathbf{X}'\boldsymbol{y} + 2\,\mathbf{X}'\mathbf{X}\,\mathbf{b} = 0$

Solving for $\mathbf{b}$ $\Rightarrow \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ (kx1) vector

• Finite Properties for $\mathbf{b}$.

1) Unbiased: $E[\mathbf{b}|\mathbf{X}] = \boldsymbol{\beta}$

2) Efficiency (& BLUE): $Var[\mathbf{b}|\mathbf{X}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$

3) If (**A5**) $\boldsymbol{\varepsilon}|\mathbf{X} \sim$ *i.i.d.* $N(\mathbf{0}, \sigma^2\mathbf{I}_T)$ $\Rightarrow \mathbf{b}|\mathbf{X} \sim$ *i.i.d.* $N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$

Review: OLS – Summary

• Asymptotic properties for **b**.

4) Consistent: $\mathbf{b} \xrightarrow{p} \boldsymbol{\beta}$

5) Asymptotic Normality: $\mathbf{b} \xrightarrow{a} N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}' \mathbf{X})^{-1})$

We use these asymptotic properties when we introduce more realistic assumptions about the data (**X** is an RV) and (**A5**) does not apply.

Review: OLS – Testing One Parameter

• We are interested in testing a hypothesis about one parameter in our linear model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

1. Set H_0 and H_1 (about only one parameter): $H_0: \beta_k = \beta_k^0$
$H_1: \beta_k \neq \beta_k^0$

2. Appropriate T(*X*): *t-statistic*. Under H_0:

If (**A5**), $t_k = \frac{b_k - \beta_k^0}{s_{b,k}} \sim t_{T-k}$

Otherwise, $t_k \xrightarrow{d} N(0,1)$

3. Compute t_k, $\hat{t}$, using b_k, β_k^0, *s*, and $(\mathbf{X}'\mathbf{X})^{-1}$. Get *p-value*($\hat{t}$).

4. Rule: Set an α level. If *p-value*($\hat{t}$) $< \alpha$ $\Rightarrow$ Reject $H_0: \beta_k = \beta_k^0$
Alternatively, if $|\hat{t}| > t_{T-k,1-\alpha/2}$ $\Rightarrow$ Reject $H_0: \beta_k = \beta_k^0$.

Review: OLS – Testing One Parameter

• Special case: H_0: $\beta_k = 0$
H_1: $\beta_k \neq 0$.

Then,

$$t_k = \frac{b_k}{SE[b_k]} = \textit{t-value} \text{ or } \textit{t-ratio}.$$

• Usually, $\alpha = 5\%$, then if $|\ t_k\ | > 1.96 \approx 2$, we say the coefficient b_k is "*significant*."

Review: OLS Estimation – Testing the CAPM

Example: We test the CAPM for IBM using the time-series.

CAPM: $E[r_{i,t} - r_f] = \beta_i\ E[(r_{m,t} - r_f)]$.

According to the CAPM, equilibrium excess returns are only determined by excess market returns –i.e., the CAPM is a one factor model. There is no constant or extra factors besides the market.

Then, there are two ways to test the CAPM:

1) Check if a constant is significant

2) Check if other factors are significant.

In this example, we are going to check if a constant is significant.

Review: OLS Estimation – Testing the CAPM

Example (continuation):

A linear data generating process (DGP) consistent with the CAPM is:

$$r_{i,t} - r_f = \alpha_i + \beta_i\,(r_{m,t} - r_f) + \varepsilon_{i,t},\ i = 1, ..., N\ \&\ t = 1, \ldots, T$$

Thus, we test the CAPM by testing H_0 (CAPM holds): $\alpha_{i=IBM} = 0$

H_1 (CAPM rejected): $\alpha_{i=IBM} \neq 0$

```
SFX_da <-
read.csv("http://www.bauer.uh.edu/rsusmel/4397/Stocks_FX_1973.csv",head=TRUE,sep=",")
x_ibm <- SFX_da$IBM                    # Extract IBM price data
x_Mkt_RF <- SFX_da$Mkt_RF              # Extract Market excess returns (in %)
x_RF <- SFX_da$RF                      # Extract risk free rate (in %)
T <- length(x_ibm)                     # Sample size
lr_ibm <- log(x_ibm[-1]/x_ibm[-T])     #  Log returns for IBM (lost one observation)
Mkt_RF <- x_Mkt_RF[-1]/100             # Adjust size (take one observation out )
RF <- x_RF[-1]/100
```

Review: OLS Estimation – Testing the CAPM

Example (continuation):

```
ibm_x <- lr_ibm – RF                   # Define excess returns for IBM

fit_ibm_capm <- lm(ibm_x ~ Mkt_RF)     # OLS estimation with lm package in R
> summary(fit_ibm_capm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.005791   0.002487  -2.329   0.0202 *
xMkt_RF      0.895774   0.053867  16.629   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q: Is $\alpha_{i=IBM} = 0$? Compute the t-value of $\alpha_{i=IBM}$:

$$\hat{t}_\alpha = \frac{\alpha_{i=IBM}}{SE[\alpha_{i=IBM}]} = \frac{-0.005791}{0.002487} = -2.329$$

$\Rightarrow |\hat{t}_\alpha| > 1.96$ $\Rightarrow$ Reject H_0 (CAPM) at 5% level

Review: OLS Estimation – Testing the CAPM

Example (continuation):

$\Rightarrow |\hat{t}_\alpha| > 1.96$ $\Rightarrow$ Reject H_0 (CAPM) at 5% level

Conclusion: The CAPM is rejected for IBM at the 5% level.

Note: You can also reject H_0 by looking at the *p-value* of intercept

p-value: **0.0202**. $< \alpha = 5\% \Rightarrow$ Reject H_0 at 5% level

Interpretation: Given that the intercept is significant (& negative), IBM **underperformed** relative to what the CAPM expected:

- $r_{IBM,t} - r_f$: mean(ibm_x) = **-0.00073141**
- $r_{IBM,t} - r_f$ (CAPM): β_i * mean(Mkt_RF) = **0.895774** * 0.0056489
  = **0.0050601**
- Ex-post difference: **-0.00073141** - **0.0050601** = **-0.00579151** ($\approx \alpha_{IBM}$)

OLS Estimation – Testing the CAPM: Remark

• We tested (& rejected) the CAPM for one asset only, IBM. But, the CAPM should apply to all assets. Suppose we have N assets. Then, a test for the CAPM involves testing N α_i's:

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_N = 0$

H_0: at least one $\alpha_i \neq 0$.

• This test is a **joint** test. It requires a simultaneous estimation of N CAPM equations.

• There are different ways to approach this test. Two popular methods are the 2-step procedure of Fama-MacBeth (1973) and the Likelihood Ratio test.

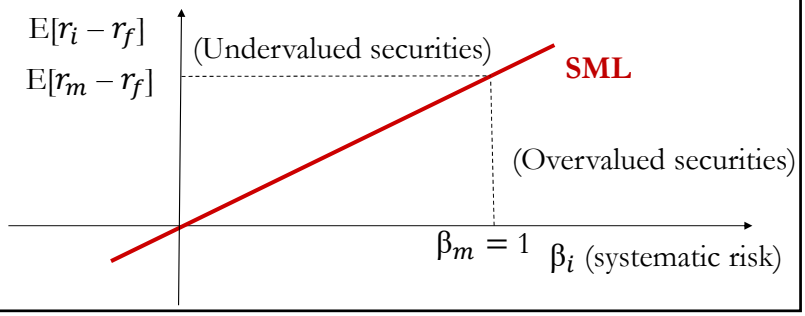
OLS Estimation – Testing the CAPM (SML)

The CAPM also tells a cross-section story for asset returns: Assets with higher β_i should get, on average, higher compensation.

CAPM (cross-section): $E[r_i - r_f] = \beta_i \lambda$

where λ, in equilibrium, is the market excess return (or factor return).

If we have β_i's for N assets, we can estimate the *security market line* (SML), where we show the effect of β_i on $E[r_i - r_f]$.



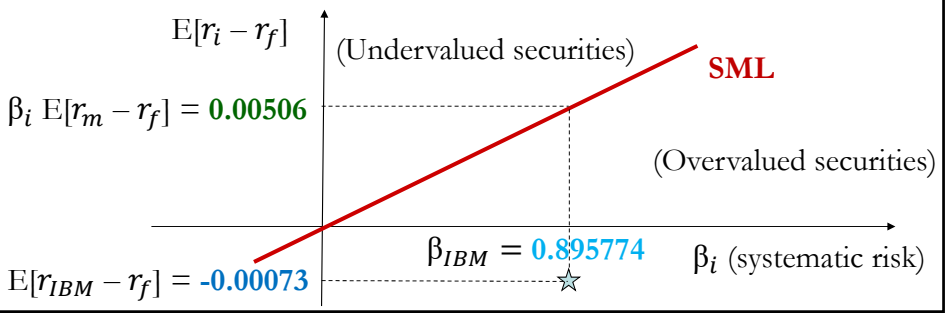
OLS Estimation – Testing the CAPM (SML)

Example (continuation):

IBM underperformed relative to what the CAPM expected by

$$\alpha_{i=IBM} = -0.005791$$

Then, according to the CAPM, IBM has been overvalued. The average, negative, performance (**-0.00073**) is the performance of a much safer asset, with a small, negative β!



OLS Estimation – Testing the CAPM (CS)

Q: Which assets pay a higher return? The SML answers this question: Assets with the higher exposure to market risk –i.e., higher β_i.

A linear DGP consistent with the CAPM is:

$$(r_i - r_f) = \alpha + \beta_i \lambda + \varepsilon_i, \qquad i = 1, ..., N$$

Testing implication of the SML for the cross-section of stock returns:

H_0 (CAPM holds in the CS): $\alpha_i = 0$ & $\lambda = E[r_{m,t} - r_f] > 0$

H_1 (CAPM rejected in the CS): $\alpha_i \neq 0$ and/or $\lambda \neq E[r_{m,t} - r_f] > 0$

• Again, we have a **joint** test. There are different ways to approach this simultaneous estimation, a common approach is a two-step estimation, popularly known as Fama-MacBeth (1973).

OLS Estimation – Testing the CAPM (CS)

• Fama-French (1992, 1993) adapted Fama-MacBeth to produce a well-known two-step approach to test the CAPM in the cross-section:

(1) Estimate β_i using the time series (T observations) for each asset i.

$r_{i,t} - r_{f,t} = \alpha_i + \beta_i (r_{M,t} - r_{f,t}) + \varepsilon_{i,t}, \ t = 1, \ldots, T \quad \Rightarrow$ Get N b_i's.

(2) Using the N b_i's as regressors, estimate

$$(\bar{r}_i - \bar{r}_f) = \alpha + b_i \lambda + \varepsilon_i, \qquad i = 1, ..., N$$

where $(\bar{r}_i - \bar{r}_f)$ is the average excess return of asset i in our sample.

The usual execution of almost all 2-step procedures involves:

1) Since returns are estimated with a lot of noise, portfolios are used.

2) The estimation takes into account the possible change over time of beta coefficients, by estimating the coefficients every 5 or 10 years.

OLS Estimation – Testing the CAPM (CS)

Example: We test the CAPM, in the cross-section, using the 2-step Fama-French method. We use returns of 25 Fama-French portfolios (sorted by Size (ME) and Book-to-Market), downloaded, along the 3-Fama-French factors from Ken French's website.

```
FF_p_da <- read.csv("https://www.bauer.uh.edu/rsusmel/4397/FF_25_portfolios.csv", head=TRUE, sep=",")
FF_f_da <- read.csv("https://www.bauer.uh.edu/rsusmel/4397/FF_3_factors.csv", head=TRUE, sep=",")

# Extract variables from imported data
Mkt_RF_fm <- FF_f_da$Mkt_RF        # extract Market excess returns (in %)
HML_fm <- FF_f_da$HML              # extract HML returns (in %)
SMB_fm <- FF_f_da$SMB              # extract HML returns (in %)
RF_fm <- FF_f_da$RF                # extract Risk-free rate (in %)
Y_p <- FF_p_da[,2:26] - RF_fm      # Compute excess returns of 25 portfolios

T <- length(HML_fm)                # Number of observations (1926:July on)
x0 <- matrix(1,T,1)                # Vector of ones, represents constant in X
```

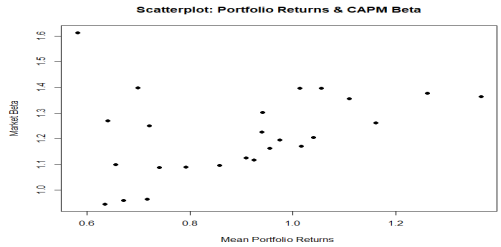
OLS Estimation – Testing the CAPM (CS)

Example (continuation):

```
x <- cbind(x0, Mkt_RF_fm)          # Matrix X (Tx2)
k <- ncol(Y_p)

## First Pass
Allbs = NULL                       # Initialize empty (a space to put betas)
for (i in seq(1,k,1)){
 y <- Y_p[,i]     # select Y (portfolio)
 b <- solve(t(x)%*% x)%*% t(x)%*%y # OLS regression = (X'X)^(-1) X'y
 Allbs  =cbind(Allbs,b)            # accumulate b as rows
}

beta_ret <- cbind(colMeans(Y_p),t(Allbs)) # Mean portfolio returns along alpha & beta estimates
cor(beta_ret[,1], beta_ret[,3])           # Correlation of mean portfolio return & beta

> cor(beta_ret[,1], beta_ret[,3])
[1] 0.3326008

plot(beta_ret[,1], beta_ret[,3], main="Scatterplot: Portfolio Returns & CAPM Beta",
    xlab="Mean Portfolio Returns ", ylab="Market Beta", pch=19)
```

OLS Estimation – Testing the CAPM (CS)

Example (continuation):




```
## Second Pass (CAPM)
fit_fm_capm_25 <- lm(beta_ret[,1] ~ beta_ret[,3])
> summary(fit_fm_capm_25)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3728	0.3113	1.198	0.243
beta_ret[, 3]	0.4289	0.2536	**1.691**	0.104

⇒ Not significant: Beta plays no role!

Conclusion: CAPM's beta does not seem to be useful to explain expected returns.

OLS Estimation – Testing the CAPM (CS)

- Fama and French (1992, 1993) estimated variations of the DGP with more factors. They found that β was weakly significant or not significant, even with the wrong sign, in explaining the C-S of stock returns, which created a big splash in the literature ("*Beta is dead*").

- Other researchers dispute the "Beta is dead" finding, criticizing the selection of estimation period, construction of portfolios, number of factors, statistical problems like measurement error and incorrect SE, etc.

- The debate about β & what (& how many) factors to include in the DGP continues.

OLS Estimation – The 3-Factor F-F Model

• The CAPM is routinely rejected. A popular alternative is the empirically derived 3-Factor Fama-French Model (1993), which adds two factors, related to firm's characteristics, to the CAPM's market excess return factor:

a) *Size* factor (**SMB**) measured as returns of small (size portfolio) minus returns of big (size portfolio) = long **S**mall & short **B**ig

b) *Value* factor or book-to-market factor (**HML**), measured as returns of high (B/M portfolio) minus returns of low (B/M portfolio) = long **H**igh & short **L**ow.

• The three factors are, in theory, "*factor mimicking portfolios*," that is, portfolios with unit exposure to the factor in question (market, size, or value), and no exposure to any other factor. If significant any factor beyond the market is considered a "CAPM *anomaly*."

OLS Estimation – The 3-Factor F-F Model

• Then, a linear DGP generating this model is:

$$r_{i,t} - r_f = \alpha_i + \beta_1 (r_{m,t} - r_f) + \beta_2 SMB_t + \beta_3 HML_t + \varepsilon_{i,t},$$

under this model, the main drivers of expected returns are sensitivity to the market, sensitivity to size, and sensitivity to value stocks, as measured by the book-to-market ratio.

• Interpretation of coefficients (also called "*factor loadings*"):
- β_1 has the same as the interpretation in the CAPM, it measures the relation between asset i risk and market risk.
- β_2 measures how tilted asset i is towards small stock (in general, $\beta_2 > 0$ means that returns of asset i behaves like small stocks.
- $\beta_3$measures how tilted asset i is towards value stock (in general, $\beta_3 > 0$ means that returns of asset i behave like high book-to-market stocks).

OLS Estimation – The 3-Factor F-F Model

• Like the CAPM, the 3-factor FF model produces expected excess returns:

$$E[r_{i,t} - r_f] = \beta_1 E[r_{m,t} - r_f] + \beta_2 E[SMB_t] + \beta_3 E[HML_t]$$

A significant constant would be evidence against this model: something is missing in the model.

• Questions:
- How where these factors determined to be drivers of stock returns? By looking at data characteristics, not theory. Data mining issues are likely present.

- Are these 3 factors the definitive number of factor?
No. There have been over 200 factors proposed! Big number, likely due to data mining. Feng, Giglio and Xiu (2020), who try to propose a method to select factors, call their paper "Taming the **Factor Zoo**."

OLS Estimation – The 3-Factor F-F Model

• In 2014, Fama and French added two additional factors to their 3-factor model: RMW & CMA.
- RMW measures the return of the portfolio of most profitable firms ("robust") minus the returns of a portfolio least profitable ("weak").

- CMA measures the return of a portfolio of firms that invest "conservatively" minus a portfolio of firms that invest "aggressively".

• Again, the 5-factor FF model produces expected excess returns:

$$E[r_{i,t} - r_f] = \beta_1 E[r_{m,t} - r_f] + \beta_2 E[SMB_t] + \beta_3 E[HML_t] + \\ + \beta_4 E[RMW_t] + \beta_5 E[CMA_t]$$

• There is debate regarding the validity of this extension, especially, outside the U.S. market.

OLS Estimation – The 3-Factor F-F Model

• Computation of factor portfolios in the Fama-French Models.

The portfolios are formed as follows:

Step 1. At the end of June of year t, sort the stock returns by attribute (size of Size, B/M or Operating Profitability).

Step 2. Split the sorted assets by attribute into 3 equal/value-weighted portfolios (3 *tercile* portfolios). Split can be thinner (quintile portfolios) or based on more complicated sorts, for example, using 6 portfolios constructed by intersecting 2 size portfolios & 3 value portfolios.

Step 3. At the end of each month (week or day), from July of year t to June of year t+1, based on the portfolios constructed in **Step 1**, compute the returns of each of the split portfolios.

Step 4 Form a "hedge portfolio": long the top portfolio (say, top tercile) and short the bottom portfolio (say, bottom tercile).

OLS Estimation – The 3-factor F-F Model: IBM

Example (continuation): Now, using the time-series, we test the significance of the factors in the Fama-French (1993) model for IBM returns with.

```
SFX_da <-
read.csv("http://www.bauer.uh.edu/rsusmel/4397/Stocks_FX_1973.csv",head=TRUE,
sep=",")

x_SMB <- SFX_da$SMB
x_HML <- SFX_da$HML
x_RF <- SFX_da$RF
SMB <- x_SMB[-1]/100
HML <- x_HML[-1]/100
y <- ibm_x                    # Define y (IBM excess returns)
x1 <- Mkt_RF                  # Regressor 1 (Mkt_RF)
x2 <- SMB                     # Regressor 2 (SMB)
x3 <- HML                     # Regressor 3 (HML)
```

OLS Estimation – The 3-factor F-F Model: IBM

Example (continuation):

```
T <- length(y)                    # New sample size (Original – 1 observation)
x0 <- matrix(1,T,1)               # Define vector of ones (the constant in X)
x <- cbind(x0,x1,x2,x3)           # Matrix X
k <- ncol(x)                      # Number of regressors (=rank(X)=k)
b <- solve(t(x)%*% x)%*% t(x)%*%y # b = (X'X)^-1 X' y  (OLS regression)
e <- y - x%*%b                    # regression residuals, e
k <- ncol(x)                      # number or regressors, k
RSS <- as.numeric(t(e)%*%e)       # RSS
Sigma2 <- as.numeric(RSS/(T-k))   # Estimated σ^2 = s^2
SE_reg <- sqrt(Sigma2)            # Estimated σ – Regression stand error
Var_b <- Sigma2*solve(t(x)%*% x)  # Estimated Var[b|X] = s^2 (X'X)^-1
SE_b <- sqrt(diag(Var_b))         # SE[b|X]
t_b <- b/SE_b                     # t-values
y_hat <- x%*%b                    # fitted values
```

OLS Estimation – The 3-factor F-F Model: IBM

Example (continuation):

```
> t(b)
              Mkt_RF      SMB       HML
[1,] -0.005088944 0.9082989 -0.2124596 -0.1715002
> t(SE_b)
              Mkt_RF      SMB       HML
[1,] 0.002487509 0.05672206 0.08411188 0.08468165
> t(t_b)
          Mkt_RF      SMB       HML
[1,] -2.045799 16.01315 -2.525917 -2.025235
```

⇒ Negative signs of β_2 & β_3.

⇒ all coefficients are *significant* ($|t|>2$).

Conclusion: Consistent with the 3-factor Fama-French model, Mkt_RF, SMB and HML are drivers of the expected returns for IBM. The signs of β_2 & β_3: IBM behaves like a large & low B/M firm.

Note: The constant is significant, that is, there is an "extra" component of expected returns not explained by the 3 F-F factors.

OLS Estimation – The 3-factor F-F Model: IBM

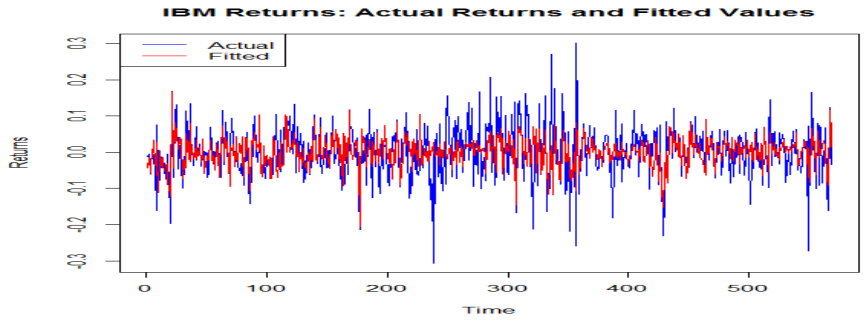
Example (continuation):

You should get the same coefficients and S.E.'s using *lm* (use summary(.) to print results) and extracting information from lm:

```
> summary(fit_ibm_ff3)                          # print lm results

Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept) -0.005089  0.002488   -2.046   0.0412 *
Mkt_RF       0.908299  0.056722   16.013   <2e-16 ***
SMB         -0.212460  0.084112   -2.526   0.0118 *      ⇒ significant at 5% level (against CAPM).
HML         -0.171500  0.084682   -2.025   0.0433 *      ⇒ significant at 5% level (against CAPM).
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05848 on 565 degrees of freedom
Multiple R-squared:  0.3389,    Adjusted R-squared:  0.3354
F-statistic: 96.55 on 3 and 565 DF,  p-value: < 2.2e-16
```

OLS Estimation – The 3-factor F-F Model: IBM

```
plot(y, type = "l", col = "blue",                          # Plot IBM returns
main = "IBM Returns: Actual Returns and Fitted Values", ylab = "Returns", xlab = "Time")
lines(y_hat, type = "l", col = "red")                      # Add fitted values to plot
legend("topleft",                                          # Add legend to plot
    legend = c("Actual", "Fitted"),     col = c("blue", "red"),     lty = 1)
```



⇒ Some periods with good fit –early & late periods- & some periods with poor fit –middle period.

OLS Estimation – Is IBM's Beta equal to 1?

Example: Using the 3-factor F-F model for IBM returns, we test if IBM's market $\beta = 1$, that is, if IBM bears the same market risk as the market. Using the previous estimation:

```
> t(b)
              Mkt_RF      SMB      HML
[1,] -0.005088944 0.9082989 -0.2124596 -0.1715002
 > t(SE_b)
              Mkt_RF      SMB      HML
[1,] 0.002487509 0.05672206 0.08411188 0.08468165
```

• Q: Is the market beta (β_1) equal to 1? That is,

H_0: $\beta_1 = 1$ vs.

H_1: $\beta_1 \neq 1$

Compute $t_k = \dfrac{b_k - \beta_k^0}{\text{Est. SE}[b_k]}$ $\Rightarrow \hat{t}_1 = \dfrac{0.9082989 - 1}{0.05672206} = -1.6167$

OLS Estimation – Is IBM's Beta equal to 1?

Example (continuation):

$\hat{t}_1 = -1.6167$

$\Rightarrow |\hat{t}_1| < 1.96$ $\Rightarrow$ Cannot reject H_0 at 5% level

Conclusions: IBM has a one-to-one risk relation with the market.

Note: You should get the same numbers using *lm* (use summary(.) to print results) and extracting information from lm:

```
fit_ibm_ff3 <- lm(ibm_x ~ Mkt_RF + SMB + HML)
b_ibm <- fit_ibm_ff3$coefficients        # Extract from lm function OLS coefficients
SE_ibm <- sqrt(diag(vcov(fit_ibm_ff3)))# SE from fit_ibm (also a kx1 vector)
t_beta1 <- (b_ibm[2] - 1)/SE_ibm[2]      # t-stat for H0: Beta1 - 1
> t_beta1
[1] -1.616674
p_val <- (1- pnorm(abs(t_beta1))) * 2     # pvalue for t_beta (adjusted b/c two sided test)
> p_val
[1] 0.1059487
```

OLS Estimation – Linear Algebra Interpretation

- Disturbances and Residuals

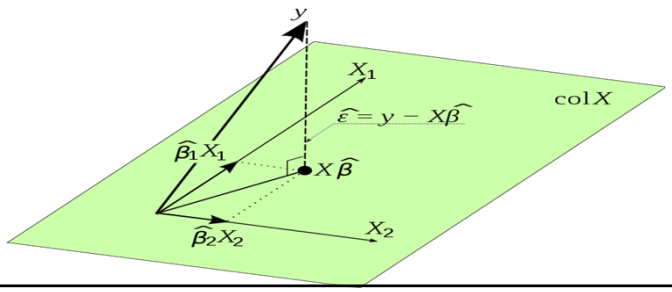
  In the population: $E[\mathbf{X}'\boldsymbol{\varepsilon}] = 0$.

  In the sample: $\mathbf{X}'\boldsymbol{e} = \mathbf{X}'(\boldsymbol{y} - \mathbf{X}\mathbf{b}) = \mathbf{X}'\boldsymbol{y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} = 0$.

- We have two ways to look at $\boldsymbol{y}$:

  $\boldsymbol{y} = E[\mathbf{y}|\mathbf{X}] + \boldsymbol{\varepsilon}$ = Conditional mean + disturbance

  $\boldsymbol{y} = \mathbf{X}\mathbf{b} + \boldsymbol{e}$ = Projection ("fitted values") + residual



Results when X Contains a Constant Term

- Let the first column of $\mathbf{X}$ be a column of ones ($\boldsymbol{x}_1 = \boldsymbol{\imath}$). That is,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{21} & \cdots & x_{k1} \\ 1 & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2T} & \cdots & x_{kT} \end{bmatrix}$$

- Recall $\boldsymbol{\imath}'\boldsymbol{z} = \sum_i^T z_i$, where $\boldsymbol{z}$ and $\boldsymbol{\imath}$ are Tx1 vectors. Then,

(1) Residuals sum to zero.

Since $\mathbf{X}'\boldsymbol{e} = \mathbf{0}$

$$= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{21} & x_{22} & \cdots & x_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kT} \end{bmatrix} * \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_T \end{bmatrix} = 0 \quad \Rightarrow \sum_i^T e_i = 0$$

$\Rightarrow \mathbf{x}_1'\boldsymbol{e} = \boldsymbol{\imath}'\boldsymbol{e} = \sum_i^T e_i = 0$ –the residuals sum to zero.

Results when X Contains a Constant Term

(2) Regression line passes through the means

Recall we can write $\boldsymbol{y}$ = fitted values + residuals

$$\boldsymbol{y} = \mathbf{Xb} + \boldsymbol{e}$$

Pre-multiply by $\boldsymbol{\imath}'$: $\boldsymbol{\imath}'\boldsymbol{y} = \boldsymbol{\imath}'\mathbf{Xb} + \boldsymbol{\imath}'\boldsymbol{e}$

$$\Rightarrow \sum_i^T y_i = \sum_i^T \{b_1 1 + b_2 x_{2i} + ... + b_k x_{ki}\} + \sum_i^T e_i$$
$$\Rightarrow \sum_i^T y_i = b_1 \sum_i^T 1 + b_2 \sum_i^T x_{2i} + ... + b_k \sum_i^T x_{ki}$$
$$\Rightarrow \sum_i^T y_i = b_1 T + b_2 \sum_i^T x_{2i} + ... + b_k \sum_i^T x_{ki}$$

Dividing both sides by T:

$$\sum_i^T y_i/T = b_1 + b_2 \sum_i^T x_{2i}/T + ... + b_k \sum_i^T x_{ki}/T$$
$$\bar{y} = b_1 + b_2 \bar{x}_2 + \cdots + b_k \bar{x}_k$$
$$\Rightarrow \bar{y} = \bar{\mathbf{x}}'\mathbf{b}$$

- That is, the regression line passes through the means.

Goodness of Fit of the Regression

- After estimating the model (**A1**), we would like to judge the adequacy of the model. There are two ways to do this:
  - Visual: Plots of fitted values and residuals, histograms of residuals.
  - Numerical measures: R^2, Adjusted R^2, AIC, BIC, etc.

- Numerical measures. In general, they are simple and easy to compute. We call them *goodness-of-fit* measures. Most popular: R^2.

- Definition: Variation

In the context of a model, we consider the *variation* of a variable as the movement of the variable, usually associated with movement of another variable.

Goodness of Fit of the Regression

• Total variation = Total sum of squares (TSS) = $\sum_i (y_i - \bar{y})^2$.

We want to decompose TSS in two parts: one explained by the regression and one unexplained by the regression.

• TSS $= \sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$
$= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$
$= \sum_i e_i^2 + \sum_i (\hat{y}_i - \bar{y})^2$

since $\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_i e_i\,(\hat{y}_i - \bar{y}) = 0$

Or TSS = RSS + SSR

RSS: Residual Sum of Squares (also called SSE: SS of errors)
SSR: Regression Sum of Squares (also called ESS: *explained* SS)

A Goodness of Fit Measure

• TSS = SSR + RSS

• We want to have a measure that describes the fit of a regression. Simplest measure: the standard error of the regression (SER)
SER = sqrt{RSS/(T − k)} ⇒ SER depends on units. Not good!

• R-squared (R^2)
1 = SSR/TSS + RSS/TSS
R^2 = SSR/TSS = Regression variation/Total variation
R^2 = 1 – RSS/TSS

As introduced here, R^2 lies between 0 and 1 (& it is independent of units of measurement!). It measures how much of total variation (TSS) is explained by regression (SSR): the higher R^2, the better.

A Goodness of Fit Measure

- $R^2 = SSR/TSS$

Interpretation: The percentage of total variation (TSS) explained by the variation of regressors.

Note: R^2 is bounded by zero and one only if:

(a) There is a constant term in **X**.

(b) The line is computed by OLS.

- Main problem with R^2: Adding regressors

It can be shown that R^2 never falls when regressors (say **z**) are added to the regression. This occurs because RSS decreases with more "information" (in the sense of more regressors).

Problem: Judging a model based on R^2 tends to over-fitting.

A Goodness of Fit Measure

- Comparing Regressions

- Make sure the denominator in R^2 is the same - i.e., same left hand side variable. For example, when modeling sales, it is common to use log(Sales). Cannot compare R^2 to the one with Sales. Loglinear will almost always appear to fit better, taking logs reduces variation.

- Linear Transformation of data does not change R^2.

- Based on $\mathbf{X}$, $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$.

Suppose we work with $\mathbf{X}^* = c\mathbf{X}$, instead ($c$ is a constant).

$$\hat{\boldsymbol{y}}^* = \mathbf{X}^* \mathbf{b}^* = c\mathbf{X} (c\mathbf{X}' c\mathbf{X})^{-1} c\mathbf{X}'\boldsymbol{y}$$
$$= c\mathbf{X} (c^2 \mathbf{X}'\mathbf{X})^{-1} c\mathbf{X}'\boldsymbol{y}$$
$$= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{y} = \mathbf{X}\mathbf{b} = \hat{\boldsymbol{y}}$$

$\Rightarrow$ same fit, same residuals, same R^2!

Adjusted R-squared

• R^2 is modified with a penalty for number of parameters: *Adjusted-*R^2

$$\bar{R}^2 = 1 - \frac{(T-1)}{(T-k)}(1-R^2) = 1 - \frac{(T-1)}{(T-k)}\frac{RSS}{TSS} = 1 - \frac{s^2}{TSS/(T-1)}$$

$$\Rightarrow \text{maximizing } \bar{R}^2 \iff \text{minimizing } [RSS/(T-k)] = s^2$$

• *Degrees of freedom* –i.e., $(T-k)$– adjustment assumes something about "unbiasedness."

• $\bar{R}^2$ includes a penalty for variables that do not add much fit. Can fall when a variable is added to the equation.

• It will rise when a variable, say **z**, is added to the regression if and only if the *t-ratio* on **z** is larger than one in absolute value.

Adjusted R-squared

• Theil (1957) shows that, under certain assumptions (an important one: the true model is being considered), if we consider several linear models:

M_1: $\boldsymbol{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \varepsilon_1$ - true model

M_2: $\boldsymbol{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \varepsilon_2$

M_3: $\boldsymbol{y} = \mathbf{X}_3\boldsymbol{\beta}_3 + \varepsilon_3$

& choose the model with smaller s^2 (or, larger Adjusted R^2), we select the true model, M_1, on average.

• In this sense, we say that "maximizing Adjusted R^2" is an *unbiased* model-selection criterion.

Other Goodness of Fit Measures

• There are other goodness-of-fit measures that also incorporate penalties for number of parameters (degrees of freedom). We minimize these measures.

• Information Criteria (IC)

- *Amemiya*: $[\boldsymbol{e'e}/(T-k)] * (1 + k/T) = s^2 * (1 + k/T)$

- *Akaike Information Criterion* (AIC)

$$\text{AIC} = -2/T(\ln L - k) \qquad L\text{: Likelihood}$$

$$\Rightarrow \text{if normality AIC} = \ln(\boldsymbol{e'e}/T) + (2/T)\, k \qquad (+\text{constants})$$

- *Bayes-Schwarz Information Criterion* (BIC)

$$\text{BIC} = -(2/T \ln L - [\ln(T)/T]\, k)$$

$$\Rightarrow \text{if normality AIC} = \ln(\boldsymbol{e'e}/T) + [\ln(T)/T]\, k \quad (+\text{constants})$$

Goodness of Fit Measures – Example

Example: 3 Factor F-F Model (continuation) for IBM returns:

```
b <- solve(t(x)%*% x)%*% t(x)%*%y          # b = (X'X)^-1 X' y  (OLS regression)
e <- y - x%*%b                             # regression residuals, e
k <- ncol(x)                               # Number of parameters estimated
RSS <- as.numeric(t(e)%*%e)                # RSS
R2 <- 1 - as.numeric(RSS)/as.numeric(t(y)%*%y)   # R-squared w/ TSS approximation
Adj_R2 <- 1 - (T-1)/(T-k)*(1-R2)           # Adjusted R-squared
AIC <- log(RSS/T) + 2*k/T                  # AIC under N(.,.) –i.e., under (A5)
```

```
> R2
[1] 0.338985
> Adj_R2
[1] 0.3354752
> AIC
[1] -5.671036
```

⇒ The 3 F-F factors explain **34%** of the variability of IBM returns.

Maximum Likelihood Estimation

• Idea: Assume a particular distribution with unknown parameters. Maximum likelihood (ML) estimation chooses the set of parameters that maximize the likelihood of drawing a particular sample.

• Consider a sample $(X_1, \dots, X_n)$ which is drawn from a pdf $f(\boldsymbol{X}|\theta)$ where θ are parameters. If the X_i's are *independent* with pdf $f(X_i|\theta)$ the joint probability of the whole sample is:

$$L(X|\theta) = f(X_1, \dots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

The function $L(X|\theta)$ –also written as $L(X;\theta)$– is called the *likelihood function*. It represents how likely it is to get a particular sample from the model. This function $L(X|\theta)$ can be maximized with respect to θ to produce maximum likelihood estimates: $\hat{\theta}_{MLE}$.

Maximum Likelihood Estimation

• It is often convenient to work with the *Log of the likelihood* function. That is,

$$\ln L(X|\theta) = \sum_{i=1}^{n} \ln f(X_i|\theta)$$

• ML estimation approach is general. We need a model (say, **A1**) and a pdf for the errors (say, normal) to apply ML. Now, if the model is not correctly specified, the estimates are sensitive to misspecification.

• A lot of applications in finance and economics: Time series, volatility (GARCH and stochastic volatility) models, factor models of the term structure, switching models, option pricing, logistic models (mergers and acquisitions, default, etc.), trading models, etc.

• In general, we rely on numerical optimization to get MLEs.

Ronald A. Fisher, England (1890 – 1962)

Maximum Likelihood Estimation: Properties

- ML estimators (MLE) have very appealing properties:

(1) *Efficiency*. Under general conditions, they achieve lowest possible variance for an estimator.

(2) *Consistency*. As the sample size increases, the MLE converges to the population parameter it is estimating:

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta$$

(3) *Asymptotic Normality:* As the sample size increases, the distribution of the MLE converges to the normal distribution.

$$\hat{\theta}_{MLE} \xrightarrow{a} N(\theta, [n\,\mathbf{I}(\theta|x_i)\,]^{-1}) = N(\theta, \mathbf{I}(\theta|X)^{-1})$$

where $\mathbf{I}(\theta|x_i)$ is the *Information matrix* for observation x_i:

$$E\left[\left(\frac{\partial \log f(\theta|x_i)}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log f(\theta|x_i)}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}}\right] = \mathbf{I}(\theta|x_i) \quad (k\text{x}k \text{ matrix})$$

Maximum Likelihood Estimation: Properties

and $$E\left[\left(\frac{\partial \log L}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log L}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}}\right] = \mathbf{I}(\theta|X)$$

is the information matrix for the whole sample.

(4) *Invariance*. The ML estimate is invariant under functional transformations. That is, if $\hat{\theta}_{MLE}$ is the MLE of θ and if $g(\theta)$ is a function of θ, then $g(\hat{\theta}_{MLE})$ is the MLE of $g(\theta)$.

Example: Suppose we estimated $\hat{\sigma}^2_{MLE}$ -i.e., the MLE of σ^2. Then, $\hat{\sigma}_{MLE}$ = sqrt($\hat{\sigma}^2_{MLE}$)

(5) *Sufficiency*. If a single sufficient statistic exists for θ, the MLE of θ must be a function of it. That is, $\hat{\theta}_{MLE}$ depends on the sample observations only through the value of a sufficient statistic.

ML Estimation: Example I

Let the sample be $\boldsymbol{X} = \{5, 6, 7, 8, 9, 10\}$ drawn from a Normal(μ, 1). The probability of each of these points based on the unknown mean, μ, can be written as:

$$f(5|\mu) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(5-\mu)^2}{2}\right]$$
$$f(6|\mu) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(6-\mu)^2}{2}\right]$$
$$\vdots$$
$$f(10|\mu) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(10-\mu)^2}{2}\right]$$

Assume that the sample is *independent*. Then, the joint pdf is given by:

$$L(X|\mu) = f(5|\mu) * f(6|\mu) * \ldots * f(10|\mu)$$

ML Estimation: Example I

Then, the joint pdf function can be written as:

$$L(X|\mu) = \frac{1}{(2\pi)^{6/2}}\exp\left[-\frac{(5-\mu)^2}{2} - \frac{(6-\mu)^2}{2} - \cdots - \frac{(10-\mu)^2}{2}\right]$$

The value of μ that maximizes the likelihood function of the sample can then be defined by $\max_{\mu} L(X|\mu)$.

It easier to maximize the *Log likelihood*, ln $L(X|\mu)$:

$$\max_{\mu} \ln\big(L(X|\mu)\big) = -{}^{6}/_{2}\ln(2\pi) + \left[-\frac{(5-\mu)^2}{2} - \frac{(6-\mu)^2}{2} - \cdots - \frac{(10-\mu)^2}{2}\right]$$

1st-derivative $\Rightarrow \frac{\partial}{\partial\mu}\left[K - \frac{(5-\mu)^2}{2} - \frac{(6-\mu)^2}{2} - \cdots - \frac{(10-\mu)^2}{2}\right]$

f.o.c. $\Rightarrow$ $(5-\hat{\mu}_{MLE}) + (6-\hat{\mu}_{MLE}) + \cdots + (10-\hat{\mu}_{MLE}) = 0$

ML Estimation: Example I

Then, the first order conditions:

$$(5 - \hat{\mu}_{MLE}) + (6 - \hat{\mu}_{MLE}) + \cdots + (10 - \hat{\mu}_{MLE}) = 0$$

Solving for $\hat{\mu}_{MLE}$:

$$\hat{\mu}_{MLE} = \frac{5 + 6 + 7 + 8 + 9 + 10}{6} = 7.5 = \bar{x}$$

That is, the MLE estimator $\hat{\mu}_{MLE}$ is equal to the sample mean. This is good for the sample mean: MLE has very good properties!