Statistics and Financial Data Analysis

MSc Mathematical and Computational Finance

Problem Sheet 1

Wen Su
University of Oxford
wen.su@maths.ox.ac.uk

UNIVERSITY OF OXFORD
Mathematical Institute

Oxford Mathematics

Q1(a): From Prices to Excess Returns

Question. Construct daily excess returns for 2019–2023.

Simple return

$$r_{i,t} = \frac{P_{i,t}}{P_{i,t-1}} - 1.$$

The relative change in the stock price, where $P_{i,t}$ is the price of stock i on day t.

Excess return

$$R_{i,t} = r_{i,t} - r_{f,t}.$$

The return relative to the risk-free benchmark, where $r_{f,t}$ is the daily risk-free return.

Similarly, the market excess return is $R_{m,t} = r_{m,t} - r_{f,t}$.

Q1(a): Market, Size, and Value Factors

CAPM (Sharpe, 1964). Expected excess returns depend on exposure to market risk:

$$\mathbb{E}[R_{i,t}] = \beta_i \mathbb{E}[R_{m,t}], \qquad R_{m,t} = r_{m,t} - r_{f,t}.$$

To assess this relation, we estimate

$$R_{i,t} = \alpha_i + \beta_i R_{m,t} + \varepsilon_{i,t}.$$

The CAPM predicts $\alpha_i = 0$. We therefore test whether $\widehat{\alpha}_i$ is significantly different from zero.

Fama–French (1993). The three-factor model adds two portfolio return spreads:

- **SMB**: small minus large market-capitalisation stocks.
- **HML**: high minus low book-to-market stocks (value minus growth).

Q1(b): HAC Standard Errors

Question. Estimate CAPM and the Fama–French model for JPM.

What is a HAC standard error?

Consider the usual t-test of $H_0 : \mu = \mu_0$. For i.i.d. observations $X_1, \ldots, X_n$, let $\bar{X} = n^{-1} \sum_{t=1}^{n} X_t$. Then

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X_1)}{n}, \qquad t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}},$$

where $s^2 = (n-1)^{-1} \sum_{t=1}^{n} (X_t - \bar{X})^2$ is the sample variance.

With unequal variances or serial correlation, we instead have

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2} \sum_{t=1}^{n} \mathrm{Var}(X_t) + \frac{2}{n^2} \sum_{1 \le s < t \le n} \mathrm{Cov}(X_s, X_t).$$

A **heteroskedasticity and autocorrelation consistent (HAC)** standard error accounts for both the variance and covariance terms.

Q1(b): The Newey–West Adjustment

For a weakly stationary series, define the lag-ℓ autocovariance and its sample estimate by

$$\gamma_\ell = \mathrm{Cov}(X_t, X_{t-\ell}), \qquad \widehat{\gamma}_\ell = \frac{1}{n} \sum_{t=\ell+1}^{n} (X_t - \bar{X})(X_{t-\ell} - \bar{X}).$$

In particular, $\gamma_0 = \mathrm{Var}(X_t)$.

Newey–West adjustment. Retain autocovariances up to lag L and apply Bartlett weights:

$$\widehat{\Omega} = \widehat{\gamma}_0 + 2 \sum_{\ell=1}^{L} \underbrace{\left(1 - \frac{\ell}{L+1}\right)}_{\text{Bartlett weight}} \widehat{\gamma}_\ell.$$

The resulting standard error and test statistic are

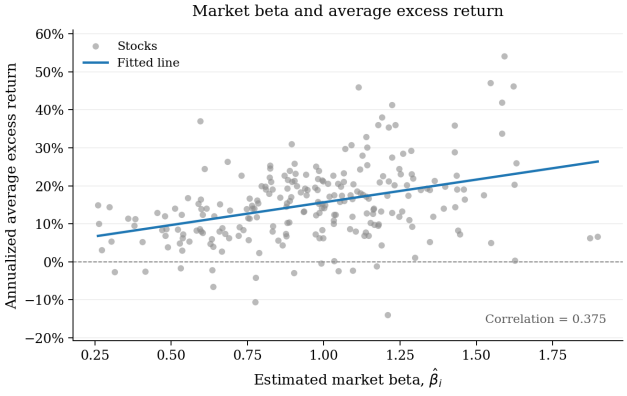
$$\mathrm{SE}_{\mathrm{HAC}}(\bar{X}) = \sqrt{\frac{\widehat{\Omega}}{n}}, \qquad t_{\mathrm{HAC}} = \frac{\bar{X} - \mu_0}{\mathrm{SE}_{\mathrm{HAC}}(\bar{X})}.$$

Q1(b): JPM Regression Results

Model	α	β_M	β_S	β_H	R^2	Adj. R^2	N
CAPM	0.000075	1.0596***	—	—	0.5274	0.5270	1,258
	(0.000371)	(0.0614)					
Fama–French	0.000065	1.1110***	−0.1463***	0.8594***	0.7617	0.7611	1,258
	(0.000272)	(0.0268)	(0.0560)	(0.0490)			

Q1(c): Market Beta and Average Excess Return

Question. Compare market betas with average excess returns across stocks.



Q1(d): Comparing Alphas

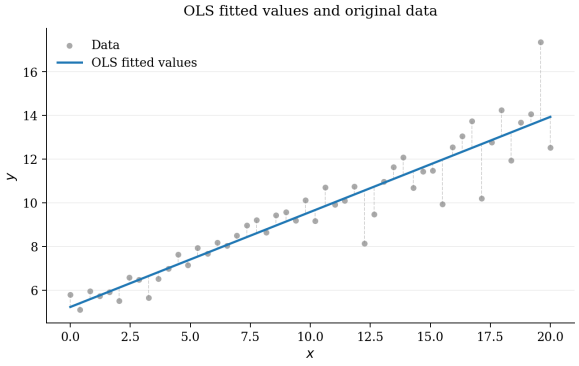
Question. Compare the estimated alphas under CAPM and Fama–French.

Model	Mean	Median	SD	Mean $\lvert\widehat{\alpha}\rvert$	Significant
CAPM	0.000041	0.000041	0.000374	0.000282	5 (2.12%)
Fama–French	0.000048	0.000046	0.000367	0.000277	6 (2.54%)

- Both distributions are centred near zero and have similar dispersion.
- Positive and negative alphas can cancel. Mean absolute alpha decreases only slightly after adding SMB and HML.
- More testing methods for α can be found in "A five-factor asset pricing model" (Fama and French, 2015, Journal of Financial Economics).

Q2(a): Ordinary Least Squares

Question. Regress `y_data` on `x` using OLS.



The residuals are much more dispersed for larger values of x, particularly among the last 20 observations. This suggests heteroskedasticity rather than constant error variance.

Q2(b): Weighted Least Squares

Question. Fit WLS and compare the estimates and standard errors with OLS.

Let w_i denote `df_WLS['weights']`.

The question states that $w_i^2 \propto \mathrm{Var}(\varepsilon_i \mid X)$. We assume

$$\mathrm{Var}(\varepsilon_i \mid X) = \sigma^2 w_i^2.$$

Therefore, WLS minimises

$$\sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\mathrm{Var}(\varepsilon_i \mid X)} \;\propto\; \sum_{i=1}^{n} w_i^{-2}(y_i - \beta_0 - \beta_1 x_i)^2.$$

In `statsmodels`, use `weights=1 / w**2`:

```
w = df_WLS['weights']
wls_model = sm.WLS(y, X, weights=1 / w**2).fit()
```

Q2(b): OLS and WLS Results

	$\widehat{\beta}_0$	SE	$\widehat{\beta}_1$	SE
OLS (conventional)	5.2426	0.2707	0.4349	0.0233
OLS (HC3)	5.2426	0.2070	0.4349	0.0300
WLS	5.2469	0.1426	0.4466	0.0180

- The fitted relationships are similar.
- WLS reports smaller standard errors for both coefficients, under the assumed variance model.
- Conventional OLS standard errors assume constant error variance. HC3 allows for heteroskedasticity without changing the estimates.

Q2(c): Prediction Intervals

Question. Construct pointwise 95% prediction intervals using WLS.

Assume independent normal errors with $\mathrm{Var}(\varepsilon_i \mid X) = \sigma^2 w_i^2$.

For a new observation at $\mathbf{x}_i = (1, x_i)^\top$, with relative error standard deviation w_i, the prediction interval is

$$\widehat{y}_i \pm t_{n-2,\,0.975}\,\widehat{\sigma}\sqrt{w_i^2 + \mathbf{x}_i^\top (X^\top W X)^{-1}\mathbf{x}_i},$$

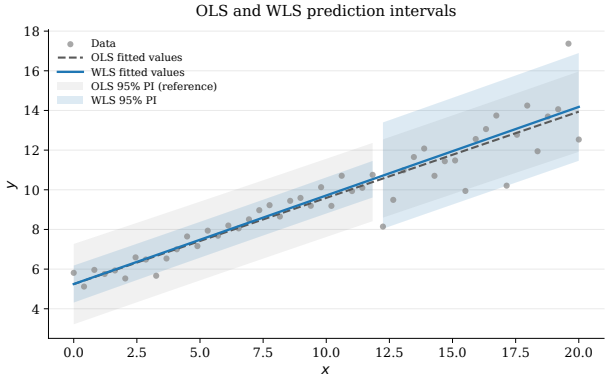
where

$$W = \mathrm{diag}(w_1^{-2}, \ldots, w_n^{-2}), \qquad \widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \frac{\widehat{\varepsilon}_i^2}{w_i^2},$$

and $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$ are the WLS residuals.

The term w_i^2 accounts for noise in the new observation; the quadratic term accounts for uncertainty in the fitted mean.

Q2(c): OLS and WLS Prediction Intervals



Q3(a): Bias of the Variance Estimator

Question. Find the bias of θ as an estimator of σ^2.

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ, variance $\sigma^2 > 0$, and $n > 1$. Consider

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad \theta = \alpha(n-1)S^2.$$

Since $\mathbb{E}[S^2] = \sigma^2$,

$$\mathbb{E}[\theta] = \alpha(n-1)\sigma^2.$$

The bias is therefore

$$\text{Bias}(\theta) = \mathbb{E}[\theta] - \sigma^2 = [\alpha(n-1) - 1]\sigma^2.$$

In particular, θ is unbiased when $\alpha = 1/(n-1)$.

Q3(a): Variance and Mean Squared Error

Question. Find the variance and mean squared error of θ.

Let $\gamma = \mathbb{E}[(X_i - \mu)^4]/\sigma^4 - 3$ be the finite excess kurtosis.

Using the variance formula given in the question,

$$\begin{aligned}\mathrm{Var}(\theta) &= \alpha^2(n-1)^2\,\mathrm{Var}(S^2) \\ &= \alpha^2(n-1)^2\left(\frac{\gamma}{n} + \frac{2}{n-1}\right)\sigma^4.\end{aligned}$$

Combining the variance and squared bias gives

$$\begin{aligned}\mathrm{MSE}(\theta) &= \mathbb{E}[(\theta - \sigma^2)^2] \\ &= \mathrm{Var}(\theta) + \mathrm{Bias}(\theta)^2 \\ &= \sigma^4\left[\alpha^2(n-1)^2\left(\frac{\gamma}{n} + \frac{2}{n-1}\right) + [\alpha(n-1) - 1]^2\right].\end{aligned}$$

Q3(b): Minimising the Mean Squared Error

Question. Find the value of α that minimises MSE.

Differentiate with respect to α:

$$\frac{\partial \,\mathrm{MSE}(\theta)}{\partial \alpha} = 2\sigma^4(n-1)\left[\alpha(n-1)\left(1+\frac{\gamma}{n}+\frac{2}{n-1}\right)-1\right].$$

Setting this equal to zero gives

$$\boxed{\alpha^* = \frac{1}{(n-1)(1+\gamma/n)+2} = \frac{1}{n+1+\frac{n-1}{n}\gamma}.}$$

This is the unique minimum because

$$\frac{\partial^2 \,\mathrm{MSE}(\theta)}{\partial \alpha^2} = 2\sigma^4(n-1)^2\left(1+\frac{\gamma}{n}+\frac{2}{n-1}\right) > 0.$$

The optimal α depends on n and γ, but not on σ^2.

Q3(c): The Normal Case

Question. Compare the minimum-MSE estimator with S^2 under normality.

For normal observations, $\gamma = 0$, so $\alpha^* = 1/(n+1)$. Hence,

$$\theta^* = \frac{n-1}{n+1} S^2, \qquad \text{Bias}(\theta^*) = -\frac{2\sigma^2}{n+1}.$$

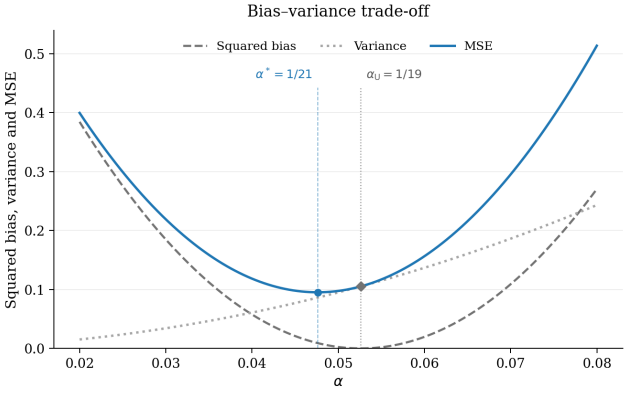
The mean squared errors are $\text{MSE}(\theta^*) = \frac{2\sigma^4}{n+1}, \text{MSE}(S^2) = \frac{2\sigma^4}{n-1}$.
The reductions relative to S^2 are

$$\text{Absolute reduction} = \text{MSE}(S^2) - \text{MSE}(\theta^*) = \frac{4\sigma^4}{n^2-1} > 0,$$

$$\text{Percentage reduction} = \frac{\text{MSE}(S^2) - \text{MSE}(\theta^*)}{\text{MSE}(S^2)} \times 100\% = \frac{2}{n+1} \times 100\%.$$

So, the use of MSE tends to be forgiving of underestimation.

Q3(d): The Bias–Variance Trade-off



Reducing α from the unbiased choice $1/19$ to $1/21$ increases squared bias but reduces variance by more. MSE falls from 0.1053 to 0.0952, a 9.52% reduction.