Introduction to Statistics

MSc Mathematical and Computational Finance

Week 1 Class

Wen Su

University of Oxford

wen.su@maths.ox.ac.uk

UNIVERSITY OF OXFORD

Mathematical Institute

Oxford Mathematics

Downloading Stock Prices from Yahoo Finance

- Download daily adjusted closing prices for Tesla (`TSLA`) and Goldman Sachs (`GS`), from 1 January 2013 to 14 August 2020.
- `pandas_datareader` is an older Yahoo Finance interface package, which is no longer reliable. We use `yfinance` instead.

```
import yfinance as yf

raw = yf.download(
    ['TSLA', 'GS'],
    start='2013-01-01', end='2020-08-15',
    auto_adjust=False,
    progress=False, threads=False
)
prices = raw['Adj Close'][['TSLA', 'GS']]
prices.head()
```

The end date is exclusive.

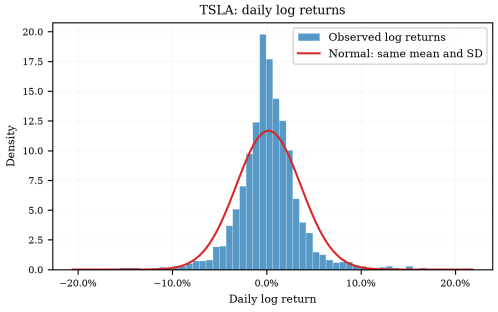
Simple and Log Returns

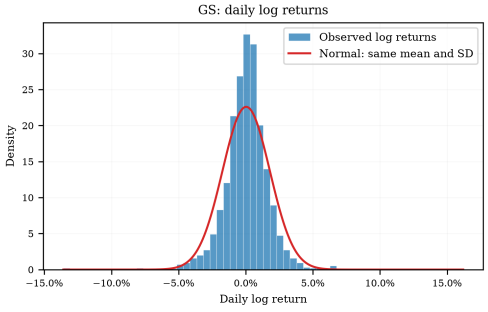
- **Simple return**: $R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$.
- **Log return**: $r_t = \log\left(\frac{P_t}{P_{t-1}}\right) = \log(1 + R_t)$.
- Simple returns compound across time, whereas log returns add:

$$\frac{P_T}{P_0} = \prod_{t=1}^{T}(1 + R_t), \qquad \log\left(\frac{P_T}{P_0}\right) = \sum_{t=1}^{T} r_t.$$

For small daily price changes, $r_t = \log(1 + R_t) \approx R_t$.

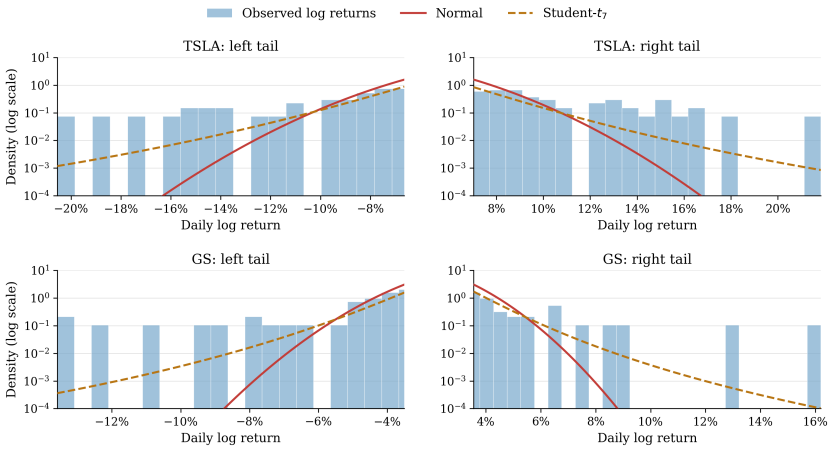
Histograms of Daily Log Returns





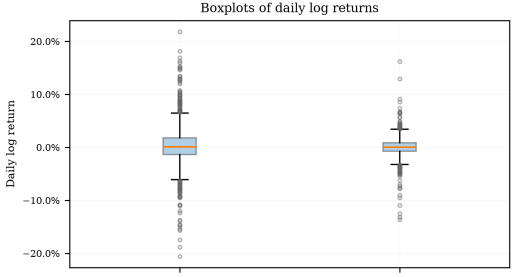
- The red curves are normal densities with the same sample mean and standard deviation as each stock's log returns.
- Both distributions are more peaked and have heavier tails than the corresponding normal distributions.

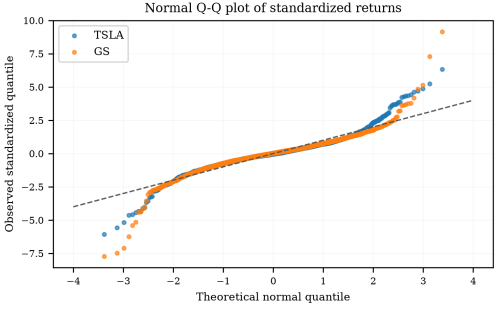
Normal and Student-t Tails



Tails beyond $\bar{r} \pm 2s$; log density scale. Both models use the full sample.

Boxplots and Q–Q Plots





- **Boxplots:** The box shows the middle 50% and the line marks the median. TSLA has greater dispersion than GS.
- **Normal Q–Q plots:** Both series fall below the reference line in the left tail and above it in the right tail, indicating heavier tails than normal.

Trading Strategy

Invest \$1 million in TSLA at its closing price S_0 on 14 August 2020. Monitor daily closing prices for the next 100 trading days.

Sell the entire position according to the following rules:

- **Stop loss:** sell when the closing price first falls below $0.9S_0$.
- **Take profit:** sell when the closing price first rises above $1.2S_0$.
- **Time limit:** if neither threshold is crossed, sell at the close of day 100.

Let τ denote the selling day. The realised profit and loss (P&L) is

$$\text{P\&L} = \frac{10^6}{S_0} S_\tau - 10^6 = 10^6 \left(\frac{S_\tau}{S_0} - 1 \right).$$

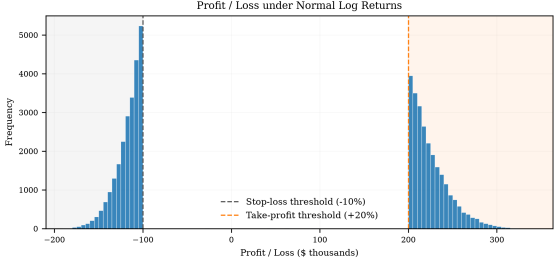
Goal: use simulation to estimate the probability of a loss and the average P&L.

Trading Profits under Normal Log Returns

Assume independent normal daily log returns:

$$r_t \overset{\text{iid}}{\sim} N(\hat{\mu}, \hat{\sigma}^2), \qquad S_t = S_{t-1} e^{r_t},$$

where $\hat{\mu} = 0.002004$ and $\hat{\sigma} = 0.034189$ are estimated from historical TSLA returns.



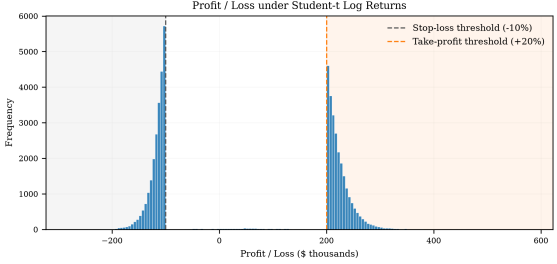
- Across 50,000 simulated paths: loss probability 48.25%; mean P&L \$59,458.

Trading Profits under Student-t Log Returns

Keep the same trading rules, but replace normal log returns with

$$r_t = \hat{\ell} + \hat{s} Z_t, \qquad Z_t \overset{\text{iid}}{\sim} t_7,$$

where the fitted location and scale are $\hat{\ell} = 0.001636$ and $\hat{s} = 0.025173$.



- Across 50,000 simulated paths: loss probability 47.59%; sample mean P&L \$60,181.

Normal versus Student-t

Same trading strategy, with 50,000 simulated paths under each model.

	Normal	Student-t_7
Loss probability	48.25%	47.59%
Sample mean P&L ($)	59,458	60,181
1st percentile ($)	−155,985	−162,543
99th percentile ($)	284,431	290,552
Minimum P&L ($)	−199,548	−313,199
Maximum P&L ($)	353,341	612,305

- Loss probabilities and sample means are similar.
- The Student-t simulation produces more extreme gains and losses.

Under Student-t log returns, $\mathbb{E}[e^{r_t}] = \infty$. The reported sample mean is therefore a finite-simulation summary, not an estimate of a finite expected P&L.

Exploring the Fuel Data

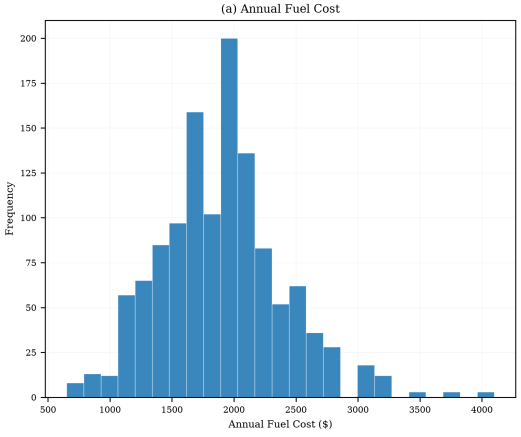
`FuelData.csv` contains 1,234 observations and 167 columns.

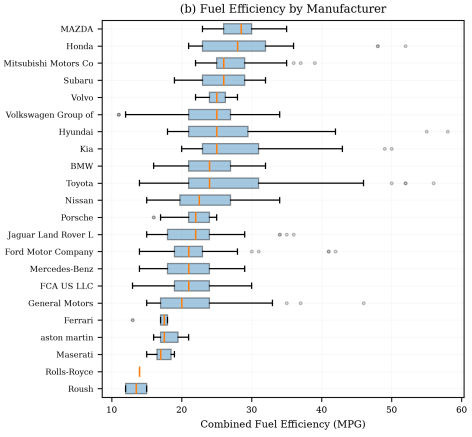
```
import pandas as pd

fuel = pd.read_csv('FuelData.csv')
fuel.head()
fuel.info()
fuel.describe()
```

- `head()`: displays the first five rows.
- `info()`: reports column names, data types and non-missing counts.
- `describe()`: summarises numerical columns with counts, means, standard deviations, quartiles, minima and maxima.

Fuel Cost and Fuel Efficiency





Most and Least Fuel-Efficient Cars

```
col = 'Comb FE (Guide) - Conventional Fuel'
cols = ['Mfr Name', 'Carline', col]

print('Most fuel-efficient cars:')
display(fuel.loc[fuel[col] == fuel[col].max(), cols])

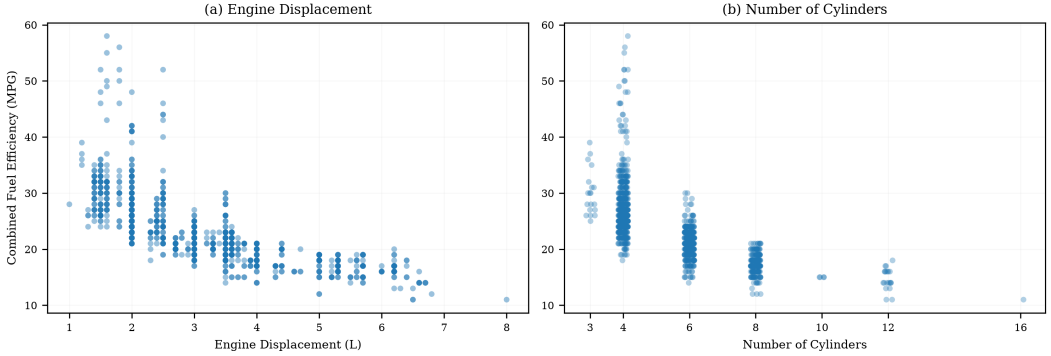
print('Least fuel-efficient cars:')
display(fuel.loc[fuel[col] == fuel[col].min(), cols])
```

- `max()` and `min()` return the highest and lowest combined MPG.
- `loc[condition, cols]` selects the matching rows and specified columns, retaining all ties.

Most efficient: Hyundai Ioniq Blue (58 MPG).

Least efficient: Chiron, Aventador Coupe and Aventador Roadster (11 MPG).

Fuel Efficiency and Engine Size



- Vehicles with larger engines and more cylinders generally have lower fuel efficiency.
- Correlation with MPG: -0.720 for engine displacement and -0.688 for the number of cylinders.

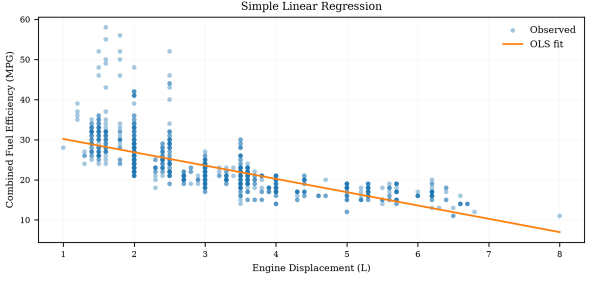
Simple Linear Regression

Let x_i be engine displacement (litres) and y_i be combined fuel efficiency (MPG).

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

The fitted regression is

$$\hat{y}_i = 33.503 - 3.317 x_i, \qquad R^2 = 0.519.$$



Regression Assumptions and Diagnostics

Let $e_i = y_i - \hat{y}_i$ denote a residual and e_i^* its standardised version.

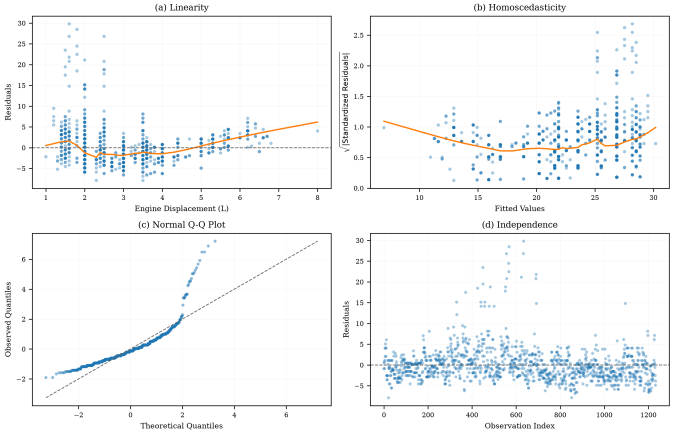
1. **Linearity:** $\mathbb{E}[y_i \mid x_i] = \beta_0 + \beta_1 x_i$.
   Plot e_i against x_i: the smooth trend should stay near zero.

2. **Constant variance:** $\mathrm{Var}(\varepsilon_i \mid x_i) = \sigma^2$.
   Plot $\sqrt{|e_i^*|}$ against $\hat{y}_i$: the smooth trend should be approximately flat.

3. **Normality:** errors are normally distributed.
   Normal Q–Q plot of residuals: points should lie approximately on a straight line.

4. **Independence:** errors are mutually independent.
   Check residual patterns in a meaningful time or group order.

Regression Diagnostics



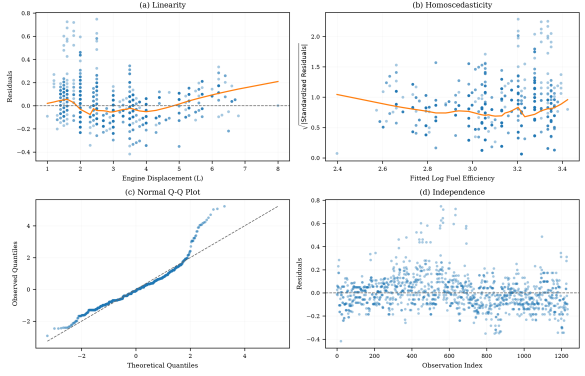
The first three plots suggest **nonlinearity**, **unequal error variance** and **departures from normality**, especially in the upper tail.

Log Fuel Efficiency

Mathematical Institute

Replace MPG by its logarithm and refit the linear model:

$$\widehat{\log y_i} = 3.5702 - 0.1466 x_i, \qquad R^2 = 0.638.$$



Extending the Regression Model

Let y_i denote MPG, x_i engine displacement, and

$$z_i = \begin{cases} 1, & \text{manual transmission,} \\ 0, & \text{automatic transmission.} \end{cases}$$

Consider four specifications:

$$M_1: \quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

$$M_2: \quad y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i,$$

$$M_3: \quad y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$$

$$M_4: \quad y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 z_i + \varepsilon_i.$$

- You can even also do nonlinear regression using polynomials:

$$y_i = f(x_i) + \varepsilon_i \approx \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_K x_i^K + \varepsilon_i.$$

Model Comparison

Adding regressors must increase the in-sample R^2 in OLS:

$$R^2 = 1 - \frac{RSS}{TSS}, \quad RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \;\; TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

We need to account for model complexity:

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} \qquad \text{Higher is better,}$$

$$\text{Akaike Information Criterion (AIC)} = n\log(RSS/n) + 2k \qquad \text{Lower is better,}$$

$$\text{Bayesian Information Criterion (BIC)} = n\log(RSS/n) + k\log n \qquad \text{Lower is better.}$$

Here k counts number of parameters.

The F-Test

Are the additional regressors helpful?

The F-statistics is

$$F = \frac{(RSS_R - RSS_U)/(k_U - k_R)}{RSS_U/(n - k_U)} \overset{H_0}{\sim} F_{k_U - k_R,\, n - k_U},$$

where k_R and k_U count number of coefficients in Model M_R and M_U.

Example: M_1 **vs** M_4:

$$M_1: \quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

$$M_4: \quad y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 z_i + \varepsilon_i.$$

Test whether the quadratic and transmission terms are jointly zero:

$$H_0: \beta_2 = \beta_3 = 0, \qquad k_R = 2, \quad k_U = 4.$$

Reject H_0 when the F statistic is sufficiently large.

Model Comparison Results

Model	Adjusted R^2	AIC	BIC
MPG			
M_1: Linear	0.518	7009.3	7019.5
M_2: + Transmission	0.520	7007.6	7022.9
M_3: + Quadratic	0.570	6870.3	6885.7
M_4: + Both	**0.574**	**6860.6**	**6881.1**
Log MPG			
M_1: Linear	0.638	−1294.0	−1283.7
M_2: + Transmission	0.638	−1295.6	−1280.2
M_3: + Quadratic	0.666	−1395.3	−1379.9
M_4: + Both	**0.669**	**−1403.2**	**−1382.7**

F-Test Results

	MPG		Log MPG	
Comparison	F	p-value	F	p-value
M_1 vs. M_2	3.706	0.0544	3.606	0.0578
M_1 vs. M_3	148.939	2.03×10^{-32}	107.458	3.38×10^{-24}
M_2 vs. M_4	157.783	3.89×10^{-34}	114.272	1.45×10^{-25}
M_3 vs. M_4	11.725	0.000637	9.969	0.00163

- The quadratic term is significant with or without the transmission variable.
- Transmission is significant at the 5% level only after including the quadratic term.

Comparing Cars Using Residuals

Using M_4 with MPG as the response, define

$$e_i = y_i - \hat{y}_i.$$

A positive residual means that the car is more fuel-efficient than predicted from its engine displacement and transmission.

Car	Observed MPG	Predicted MPG	Residual
Hyundai Ioniq Blue	58	30.01	+27.99
Chiron	11	19.06	−8.06

Thanks!