ISOM 2600 Business Analytics

TOPIC 5: INTRODUCTION TO LINEAR REGRESSION MODEL

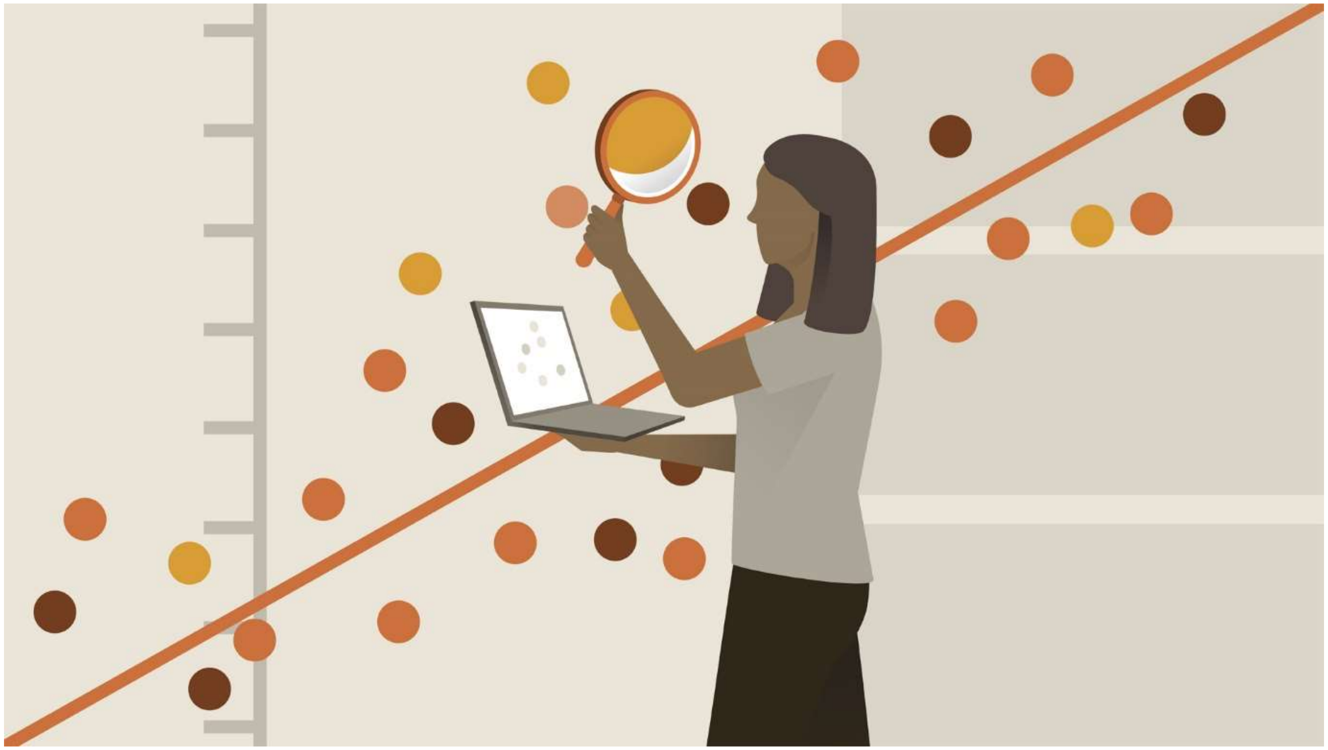
XUHU WAN

HKUST

JANUARY 15, 2022

Goals for this topic

- Review the basic concepts in simple linear regression and introduce multiple linear regression.



Why Regression?

- The motivation for using the technique:
  - Explanation: Explain the impact of changes in an predictor ($\boldsymbol{x}_i$) on the response ($\boldsymbol{y}$)
  - Prediction: Predict the value of a response ($\boldsymbol{y}$) based on the value of predictors ($\boldsymbol{x}_1$, $\boldsymbol{x}_i$,…$\boldsymbol{x}_k$)

Note: Response also called: dependent variable / outcome variable / target variable

Predictor also called: independent variable/ explanatory variable / covariate / risk factor / attribute

Simple Linear Regression

Simple Regression Model (SRM)

Observed values of the response y are linearly related to values of the explanatory variable x by equation

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where $\mathrm{y} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},\ x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$

n is sample size,

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2), i = 1,2, \dots, n$$

Finance Application: SIM

One of the most important applications of linear regression is the **Single Index Model (SIM)**.

It is assumed that risk adjusted return on an asset (R) is linearly related to the risk adjusted of return on the overall market (R_m).

$$R = \alpha + \beta R_m + \varepsilon$$

Example: Tesla

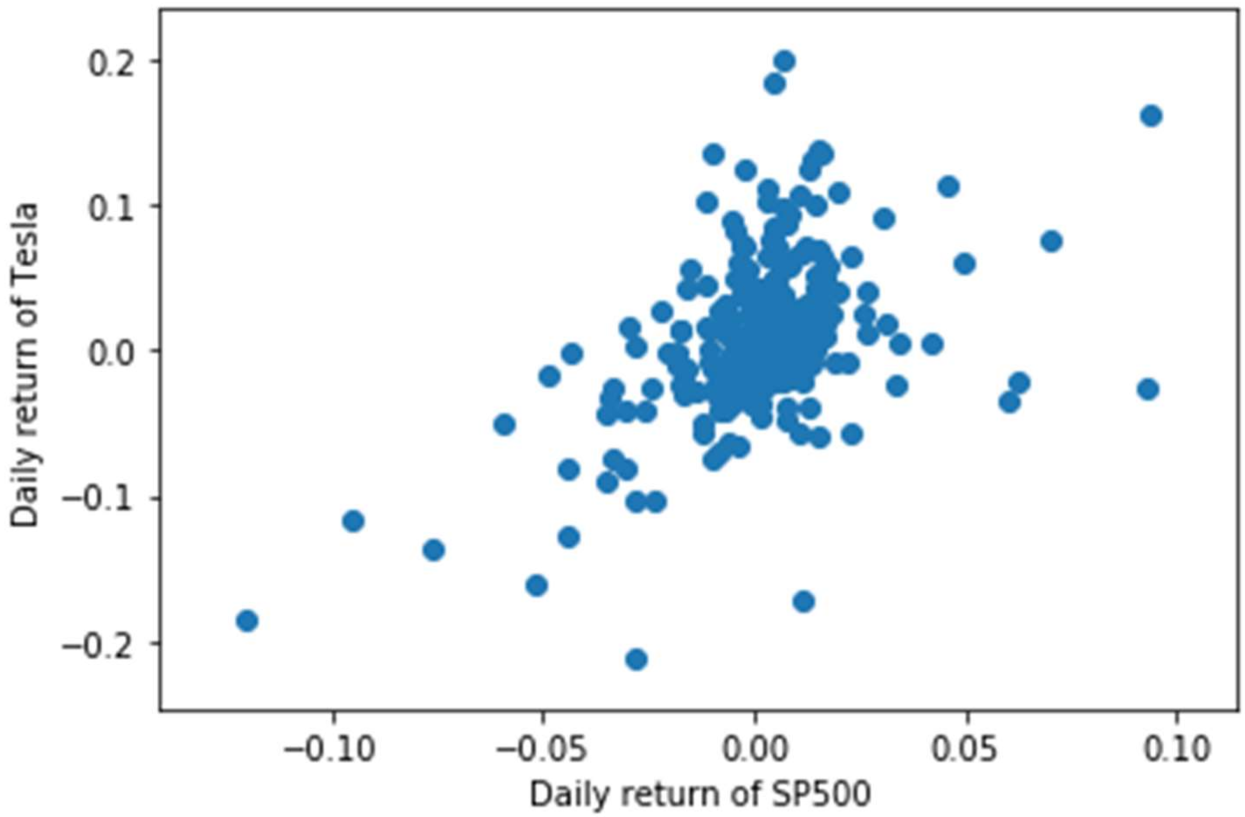
Tesla CEO, Elon Musk, has become the richest person in the world once again since the start of 2021.

In the last year alone, Tesla's share price has rocketed upward more than 700%.

We can use Single-Index Model (SIM) to analyze whether Tesla beat the market or whether it is aggressive or defensive significantly.

Example: Tesla (n=253)

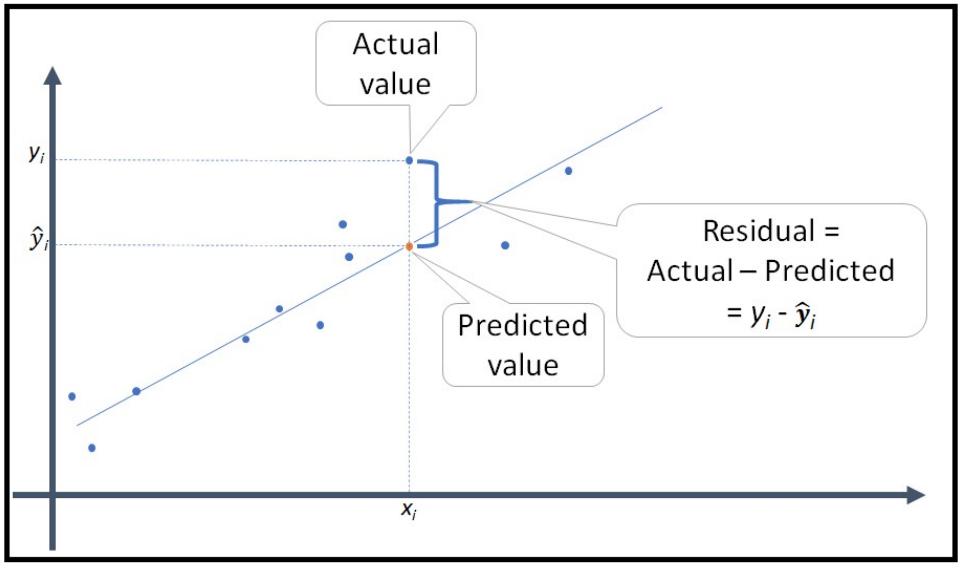
The scatterplot summarizes the relationship between the daily (simple) return of Tesla and the daily (simple) return of SP500 in 2020.



Determine Regression Equation

One goal of regression is to draw the 'best' line through the data points (least-squares regression line)

$e_i = y_i - \hat{y}_i$ is the residual, which describes the error of prediction



Least-Squares: Choose b_0 and b_1 to minimize the sum of squared residuals: $\mathrm{SSE} = \sum_{i=1}^{n} e_i^2$.

The fitted value $\hat{y} = b_0 + b_1 x$

Formula: $b_1 = r\frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1\bar{x}$, where r is the correlation between x,y, s_x, s_y *are sample standard deviation of* x, y

RMSE

The standard deviation of residuals measures how much the residuals vary around the fitted value, called root mean squared error (RMSE, or s_e)

$$\text{RMSE} = \sqrt{\frac{e_1^2 + e_2^2 + \cdots + e_n^2}{n-2}}$$

Assumptions

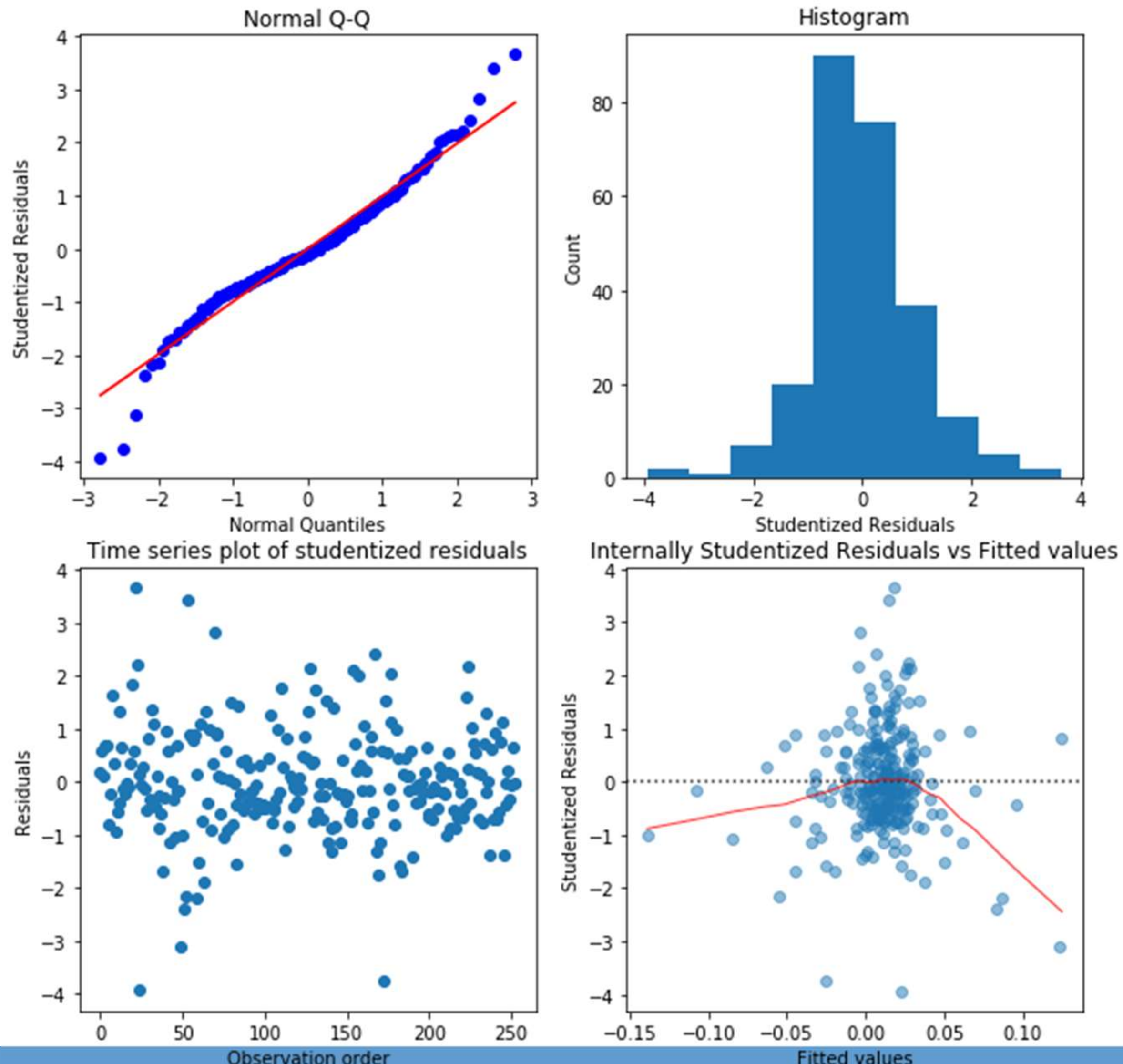
The random errors (ε)

- 1. have mean equal to zero
- 2. have equal variance σ_{ε}^{2}
- 3. are normally distributed
- 4. are independent of each another

Visual test for regression: Residual plots

Regression Diagnosis

Example – Tesla Data (Continued)




```
def four_in_one(dataframe,model):
    fitted_y = model.fittedvalues
    studentized_residuals = model.get_influen
    plt.figure(figsize=(10,10))
    ax1 = plt.subplot(221)
    stats.probplot(studentized_residuals, dis
    ax1.set_title('Normal Q-Q')
    ax1.set_xlabel('Normal Quantiles')
    ax1.set_ylabel('Studentized Residuals');
```

```
four_in_one(returns, model)
```

R-squared

Used to measure how useful is the regression model.

A regression line splits the response into two parts, a fitted value and a residual, $y = \hat{y} + e$

As a summary of the fitted line, it is common to report how much of the variation of y is explained by x in the regression model, the R-squared.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \ e_i = y_i - \widehat{y}_i$$

Regression output (python)

Example – Tesla Data (Continued)

```python
model = sm.OLS(returns['Tesla'], sm.add_constant(returns['SP500'])).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Tesla   R-squared:                       0.226
Model:                            OLS   Adj. R-squared:                  0.223
Method:                 Least Squares   F-statistic:                     73.23
Date:                Mon, 08 Feb 2021   Prob (F-statistic):           1.17e-15
Time:                        10:38:45   Log-Likelihood:                 401.99
No. Observations:                 253   AIC:                            -800.0
Df Residuals:                     251   BIC:                            -792.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0090      0.003      2.891      0.004       0.003       0.015
SP500          1.2326      0.144      8.557      0.000       0.949       1.516
==============================================================================
Omnibus:                       17.714   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               55.635
Skew:                           0.054   Prob(JB):                     8.30e-13
Kurtosis:                       5.295   Cond. No.                         46.2
==============================================================================
```

Inference

Three parameters identify the population described by the simple regression model. The least-squares regression provides the estimates: b_0 estimates β_0, b_1 estimates β_1, and s_e estimates σ_ε

The sampling distribution of b_0 and b_1 (normal distribution)

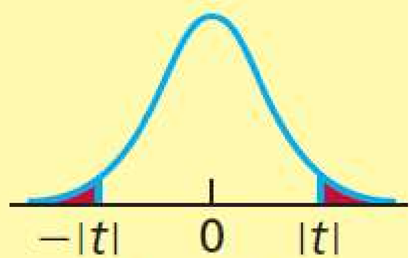
The confidence interval and hypothesis test of β_0 and β_1

In SIM, people are usually interested in testing whether $\beta_0 = 0$ and $\beta_1 = 1$.

Regression table

Hypothesis test: $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$

Test Statistic $t = \dfrac{b_1}{s_{b_1}}$



p-value = twice the area to the right of $|t|$

Here the $p - value$ is based on $n - 2$ degrees of freedom.

$p - value < \alpha$ => reject $H_0: \beta_1 = 0$ at the given significance level α,

=> we conclude that the predictor x is useful to predict y.

95% CI for β_1: $b_1 \pm 1.96\, s_{b_1}$

- For simplicity, can just round it to $b_1 \pm 2 s_{b_1}$

where $s_{b_1} = \dfrac{s_e}{\sqrt{n-1}} \times \dfrac{1}{s_x}$

- Example – CEO Data (Continued)

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.8653	0.401	4.654	0.000	1.075	2.656
Log10 Net Sales	0.5028	0.043	11.608	0.000	0.417	0.588

Prediction Interval

The 95% prediction interval for the response y_{new} under the Simple Regression Model equals

$$\hat{y}_{new} \pm t_{0.025,n-2}\, se(\hat{y}_{new}),$$

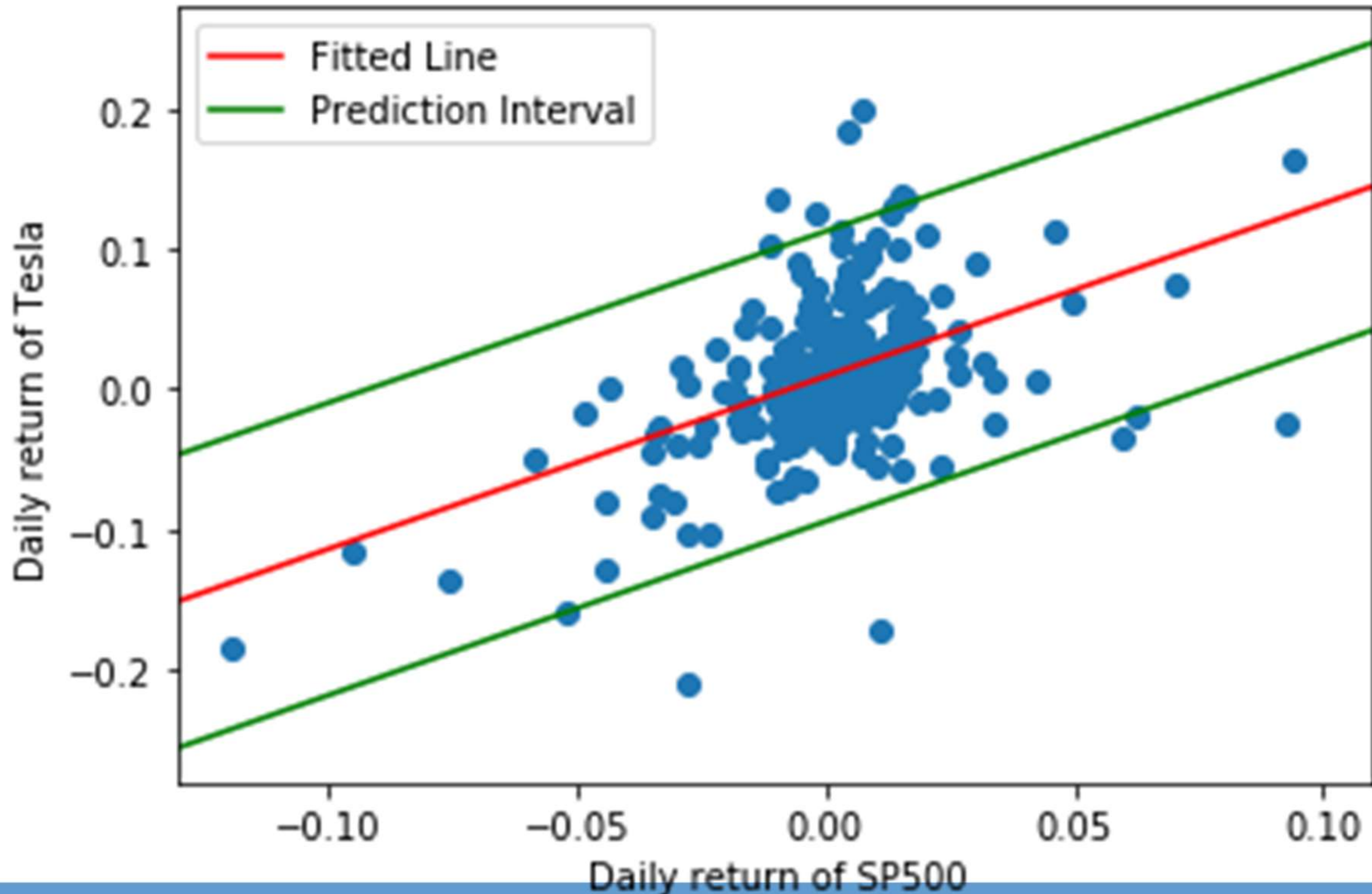
where $\hat{y}_{new} = b_0 + b_1 x_{new}$ and

$$\text{se}(\hat{y}_{new}) = \text{RMSE}\sqrt{1 + \frac{1}{n} + \frac{(x_{new}-\bar{x})^2}{(n-1)s_x^2}}$$

Example – Tesla Data (Continued)

PI illustration

```
plt.scatter(returns['SP500'], returns['Tesla'])
plt.plot([x_min, x_max], model.predict([[1, x_min], [1, x_max]]), color = 'red', label = 'Fitted Line')
plt.plot([x_min, x_max], model.get_prediction([[1, x_min], [1, x_max]]).summary_frame()['obs_ci_lower'].values, co
plt.plot([x_min, x_max], model.get_prediction([[1, x_min], [1, x_max]]).summary_frame()['obs_ci_upper'].values, co
```



Introduction to Multiple Linear Regression Part I

Case study: Retail profits

A chain of pharmacies is looking to expand into a new community.

It has data for 110 cities on the following variables:

- Annual profits of the pharmacies (in dollars)
- Income (median annual salary of the city)
- Disposable income (median income net of taxes),
- Birth rate (per 1,000 people in the local population)
- Social security recipients (per 1,000 people in the local population)
- Cardiovascular deaths (per 1,000 people in the local population)
- Percentage of the local population 65 years old or above.

Objectives

Management would like to

(a) know whether and how these variables are related to profits,

(b) provide a means to choose new communities for expansion

(c) predict sales at existing locations to identify underperforming sites.

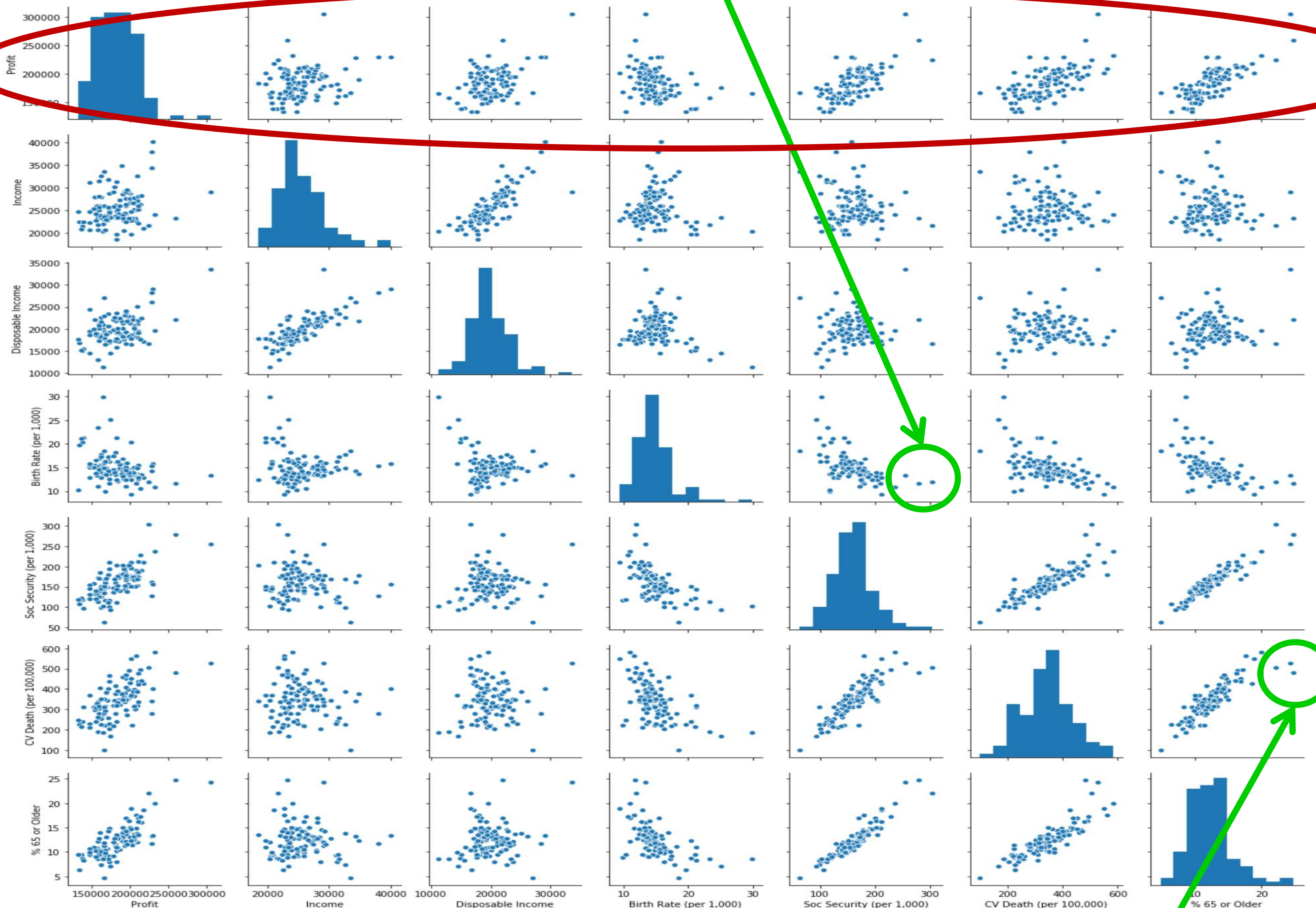
Step 1: Examine the correlation and scatterplot matrices.

- Which variables are related to profits?

```
corr = df_p.corr()
corr.style.background_gradient(cmap='coolwarm').set_precision(2)
```

	Profit	Income	Disposable Income	Birth Rate (per 1,000)	Soc Security (per 1,000)	CV Death (per 100,000)	% 65 or Older
Profit	1	0.26	0.47	-0.35	0.67	0.61	0.77
Income	0.26	1	0.78	-0.091	-0.13	-0.05	-0.056
Disposable Income	0.47	0.78	1	-0.26	0.063	0.056	0.16
Birth Rate (per 1,000)	-0.35	-0.091	-0.26	1	-0.58	-0.55	-0.55
Soc Security (per 1,000)	0.67	-0.13	0.063	-0.58	1	0.85	0.94
CV Death (per 100,000)	0.61	-0.05	0.056	-0.55	0.85	1	0.87
% 65 or Older	0.77	-0.056	0.16	-0.55	0.94	0.87	1

Lower birth rate (Texas and Utah)



Older communities (Florida)

```python
import seaborn as sns
sns_plot = sns.pairplot(df_p)
```

Method

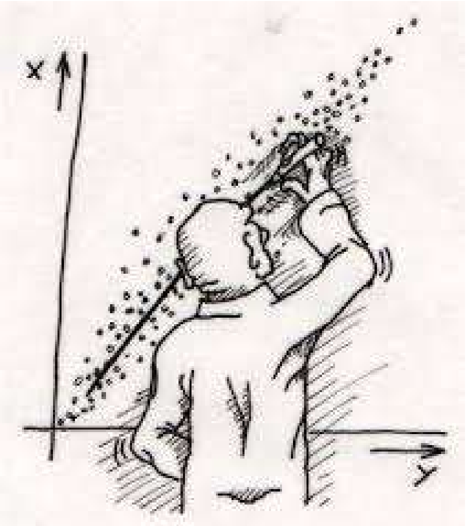
Use multiple regression (MR) with profit as the response variable.

Model

Model checking

Inference in MR

Prediction



Multiple Regression Model (MRM)

Under the MRM, the data are assumed to be a realization of

Predictors (/independent /explanatory variable /covariate /risk factor /attribute)

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i,$$

$$i = 1, \dots, n$$

Response (/dependent /outcome /target) Variable

$$\varepsilon_1, \dots\dots, \varepsilon_n \overset{i.i.d.}{\sim} N(0, \sigma_\varepsilon^2)$$

such that the conditional mean

$$E(y_i | x_{i1}, \dots, x_{ik}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

Parameters in MRM

In MRM, β_0, β_1 ..... β_k (the partial slopes), and σ_ε (the standard deviation of random errors) are (usually) unknown

An objective of regression is to estimate them

Case study: Retail profits Continued

```
# Specify the dependent variable and independent variables
X = df_p.drop(columns="Profit")
Y = df_p['Profit']
```

Least-squares regression

- In order to **estimate the "true" regression:**

$$E(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_k) = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_k \mathbf{x}_k$$

- We use the **least square (LS) regression:**

$$\hat{\mathbf{y}} \quad = b_0 + b_1 \mathbf{x}_1 + \cdots + b_k \mathbf{x}_k,$$

  which minimizes the sum of squared deviation from data to estimated signal $\min_{b_0 \ldots b_k} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

- The values b_0, b_1, …, b_k are called the **least squares (LS) estimates** of β_0, β_1 ….. β_k
- The values b_0, b_1, …, b_k are calculated by computer programs such as SAS, R, Python, Minitab and Excel.

Case study: Retail profits Continued

```python
model_fit1 = sm.OLS(Y,sm.add_constant(X)).fit()
print(model_fit1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Profit   R-squared:                       0.756
Model:                            OLS   Adj. R-squared:                  0.742
Method:                 Least Squares   F-statistic:                     53.12
Date:                Tue, 28 Jan 2020   Prob (F-statistic):           2.41e-29
Time:                        10:33:06   Log-Likelihood:                -1199.0
No. Observations:                 110   AIC:                             2412.
Df Residuals:                     103   BIC:                             2431.
Df Model:                           6
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                    1.316e+04   1.91e+04      0.689      0.493   -2.47e+04    5.11e+04
Income                      0.5986      0.589      1.017      0.312      -0.569       1.766
Disposable Income           2.5350      0.732      3.464      0.001       1.084       3.986
Birth Rate (per 1,000)   1703.8657    563.673      3.023      0.003     585.954    2821.777
Soc Security (per 1,000)  -47.5162    110.213     -0.431      0.667    -266.097     171.065
CV Death (per 100,000)    -22.6821     31.464     -0.721      0.473     -85.084      39.720
% 65 or Older            7713.8505   1316.210      5.861      0.000    5103.458    1.03e+04
==============================================================================
Omnibus:                        0.673   Durbin-Watson:                   1.546
Prob(Omnibus):                  0.714   Jarque-Bera (JB):                0.789
Skew:                          -0.097   Prob(JB):                        0.674
Kurtosis:                       2.634   Cond. No.                     4.85e+05
==============================================================================
```

Fitted values and residuals

As in simple regression, the LS regression line again serves to decompose the data into the fitted values and the residuals

$$y_i = \hat{y}_i + e_i,$$

- where $\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}$ are called fitted values
- and $e_i = y_i - \hat{y}_i$, the residuals.

Thus, LS regression decomposes the observed data into ‘signal’ plus ‘noise’.

RMSE in MRM

When the MRM holds, $\boldsymbol{\sigma_\varepsilon}$ is estimated by RMSE

As in a simple linear regression, RMSE is called the standard deviation of the residuals and measures the dispersion of the residuals about the LS regression line

- In MRM, RMSE again measures the predictive accuracy of the model used to forecast values for new cases

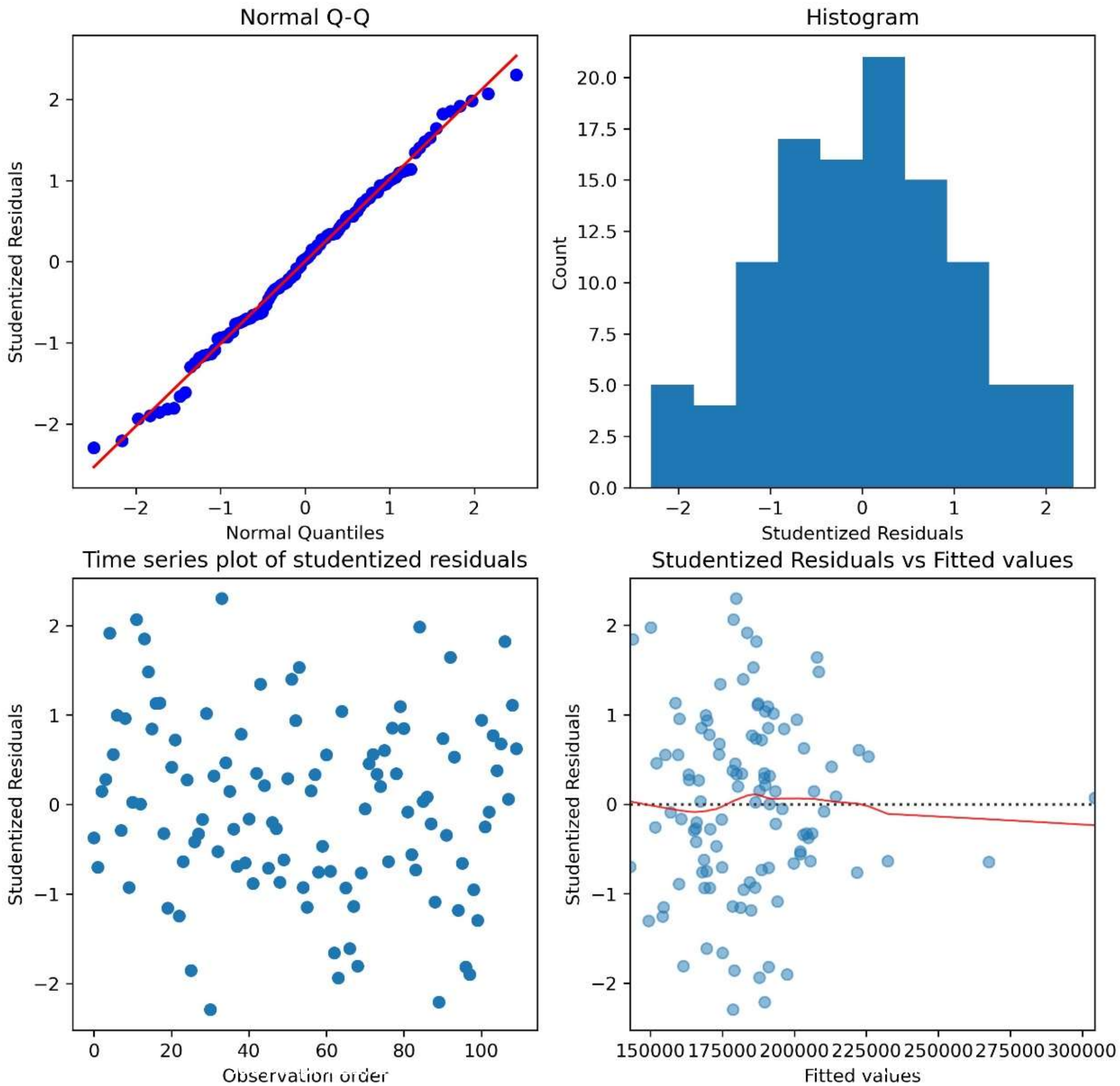
The formula is

$$RMSE = \sqrt{\sum_{i=1}^{n} (y_i - b_0 - b_i x_{i1} - \cdots - b_k x_{ik})^2 / (n - 1 - k)}$$

Case study: Retail profits Continued

Step 2: Examine Residuals

Q: Can we trust the result from multiple linear regression?




```
four_in_one(df_p,model_fit)
```

Take away from Topic 5

Review simple linear regression

- Check model assumption – residual plots
- R^2
- t-test and CI for beta
- Prediction

Multiple linear regression model

- Scatter plot matrix
- Least square regression
- Diagnose of model assumption

Appendix 1

Statistical Methods and Concepts Reported in Python Computer Output

Omnibus test is a statistical test of residual normality, the smaller the test statistic value the better, or equivalently, the larger P(Omnibus) the better.

Skewness, the standardized 3rd central moment, is a measure of symmetry of residuals – the closer to zero the better.

Kurtosis, the standardized 4th central moment, is a measure of heaviness of residual tails – the closer to 3 the better.

Jarque-Bera test is another test of residual normality based on skewness and kurtosis. The smaller the JB test statistic value the better, or equivalently, the larger P(JB) the better.

Durbin-Watson test is a statistical test of residual independence based on lag one autocorrelation. The closer DW statistic to 2 the better.

Condition number is the ratio of maximum to minimum eigenvalue of Gramian matrix X^TX, where X has n rows (correspond to n observation) and p columns (corresponds to p x-variables). A large conditional number indicates collinearity among independent variables.