Introduction to Linear Regression

```
key words: dependent and independent variables, regression line, sse, ssr,
sst, R^2, regression coefficient estimation
```

1. Linear Regression: Introduction

- Simple Linear Regression

    - definition: degree of variable **differs from mean** value(variation) in a dependent variable in term of variation in a independent variable.
        - Dependent variable: explained/ endogenous/ predicted variable
        - Independent variable: explanatory/ exogenous/ predicting variable
        - variation != variance, $variation\ in\ Y\ = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

- Least Squares and Regression Coefficient Estimation

    - Linear Regression model: $Y_i = b_0 + b_1 X_i + \epsilon_i,\ i = 1, ..., n$
        - b_0: regression intercept
        - b_1: regression slope coefficient
        - ϵ_i: residual for ith observation
    - Regression line: $\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i,\ i = 1, ..., n$
        - $\hat{Y}_i$: estimated value of Y_i given X_i
        - $\hat{b}_0, \hat{b}_1$: estimated intercept and slope
    - Sum of Squared Errors(SSE)
        - Intuition: we need to **minimize** the error between actual value Y_i and predicted one $\hat{Y}_i$
        - $SSE = e_i^2 = \sum(\hat{y}_i - y_i)^2$: unexplained variation
            - $SSR = \sum(\hat{y}_i - \bar{y})^2$

- - - variation in dependent variable that can be explained by independent variable
    - $SST = \sum(y_i - \bar{y})^2$
      - total sum of squares, total variation of dependent variable
      - differences between **actual Y and mean of Y**
    - $SST = SSE + SSR$, total variation=explained variation + unexplained variation
    - $R^2 = SSR/SST$
  - Estimated Slope Coefficient($\hat{b_1}$)
    - the change in Y for a one-unit change in X
    - $\hat{b_1} = \frac{Cov_{X,Y}}{\sigma_X^2}$
    - Applied in Portfolio Management: in stock return, this estimator is treated as stock β, systematic risk
  - Estimated Intercept($\hat{b_0}$)
    - $\hat{b_0} = \bar{Y} - \hat{b_1}\bar{X}$
    - estimate of the dependent variable when the independent variable is zero.

- Residuals with Violated Assumptions

  - Assumptions:
    1. **linear relationship** between independent and dependent.
    2. variance of residuals is **constant**(same) for all observation(homoskedasticity)
    3. residual term **iid**. One observation's residual not correlated with that of another.
    4. residual term **normally distributed**.
  - Violates:
    - Non-linear:
      - the sign of Residuals' over independent variable can check the the linearity or not.
    - Heteroskedasticity: residual for all observations without same variance
      - check by plot residual and time scatter plot, or fitted and observed graph.
    - Dependent:
      - if X and Y are not independent, residuals are not independent as well.
      - Both of variance and coefficient estimations are incorrect.
    - Non-Normality:

        - Outliers influence the distribution of residuals
        - if residual normally distributed, hypothesis can be conducted. Even through it is not normal distributed, with large enough samples, CLT applies, out parameter estimates are valid.

2. Goodness of Fit and Hypothesis Tests

- Analysis of Variance(ANOVA):
    - Summary of the variation of dependent variable
    - in this table, we only show the simple linear regression
        - k: number of slope parameters estimated
        - n: observation amount

Source of Variation	Degree of Freedom	Sum of Squares	Mean Sum of Squares
Regression(explained)	1	SSR	MSR=SSR/k=SSR/1=SSR
Error(unexplained)	n-k-1=n-2	SSE	MSE=SSE/n-k-1
Total	n-1	SST	

- Standard Error of Estimation(SEE)
    - $SSE = \sqrt[2]{MSE}$
    - the lower, the better
- Coefficient of Determination
    - *percentage* of total variation of dependent explained by independent variable
    - $R^2 = \frac{SSR}{SST}$
        - $R^2 = r^2$, for regression with one independent variable, where r^2 is the correlation coefficient
- F-Statistic
    - how well a set of independent variables *explains* the variation in dependent variable
    - $F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$
    - always one-tail test

- Hypothesis Testing: $H_0 : b_1 = 0, H_1 : b_1 = 0$
  - determine whether b_1 is statistically significant using F-test
  - compare the F-statistic we calculated and critical F_c value with degree of freedom

- Regression Coefficient Test

  - T-test:
    - $t_{b_1} = \frac{\hat{b_1} - b_1}{s_{\hat{b_1}}}$
      - degree of freedom: $n - 2$
    - compare the t statistic we calculated with the critical value t_c with degree of freedom and significant level
  - Hypothesis: to test whether the true slope coefficient=hypothesized value
    - Eg: $H_0 : b_1 = 0, H_1 : b_1 = 0$, in this case, the hypothesis value is 0.

3. Predicting Dependent Variables and Functional Forms

- Predicted Values

  - Definition: independent variable values based on estimated regression coefficients and a prediction value about independent variable.
  - Formula for simple regression: $\bar{Y} = \bar{b}_0 + \bar{b}_1 X_p$
    - $\bar{Y}$ and X_p: predicted and forecasted value of dependent and independent value respectively.
    - given the value of $\bar{b}_0$, $\bar{b}_1$ and X_p
  - Calculate Confidence Interval for Predicted Values
    - Range: $[\hat{Y} - (t_c * s_f), \hat{Y} + (t_c * s_f)]$
      - t_c: critical value(2-tailed) with given α and $df = n - 2$
      - s_f:
        - definition: standard error of forecast
        - function: $s_f^2 = SEE^2[1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{(n-1)s_x^2}]$
          - SEE^2: variance of the residuals

- Difference Form of Simple Linear Regression

  - log-In model:

        - dependent variable is logarithmic, the independent variable is linear.
        - Function: $lnY_i = b_0 + b_1X_i + \epsilon_i$
    - ln-log model:
        - dependent variable is linear, while independent variable is logarithmic.
        - Function: $Y_i = b_0 + b_1ln(X_i) + \epsilon_i$
    - log-log model:
        - both are logarithmic.
        - Function: $lnY_i = b_0 + b_1ln(X_i) + \epsilon_i$
    - Selection function form based on
        - natural of variables
        - goodness of fit measure($R^2, SSE, F - statistics$)