R：Statistical Programming Methods
R：程式、機率與統計

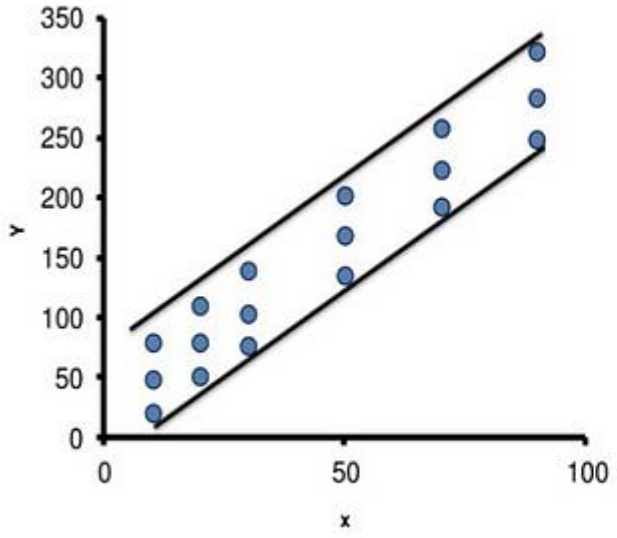
Interpretation of Linear Regression

Conducting Linear Regression Analysis

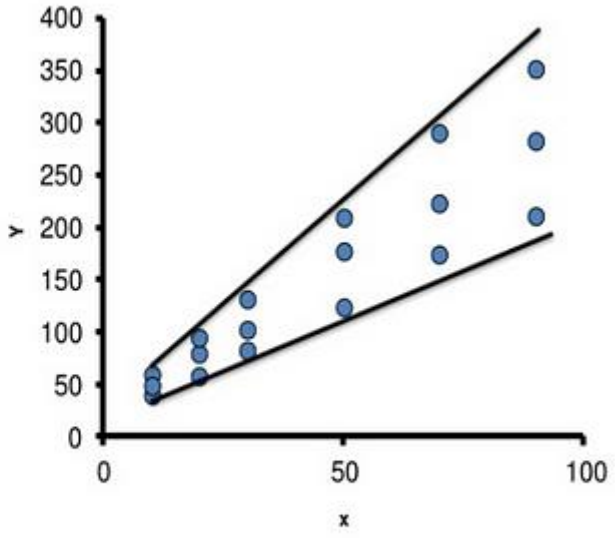
1. Checking multicollinearity (共線性) between independent variables
   - Correlation coefficients
   - Variance Inflation Factor (VIF) – above 5 shows potential high correlation between predictor variables
2. Normality Check
   - **residuals** are normally distributed
3. Checking homoscedasticity (同質性)
   - The residuals have equal variance (homoscedasticity) for every value of the fitted values and of the predictors

Is this model a good model?

Violation of Homoscedasticity



homoscedasticity



heteroscedasticity

- Source: Linear Regression: Assumptions, Violation of Assumptions & Rectification | by Akshatha Vijay | Jul, 2023 | Medium

Checking multicollinearity (共線性)

```
check <- cbind(df$Adult.Mortality, df$infant.deaths, df$Alcohol,
df$BMI, df$GDP, df$Schooling, df$Population)
cor(check)
```

```
##             [,1]        [,2]        [,3]        [,4]        [,5]        [,6]
## [1,]  1.00000000  0.04245024 -0.17553509 -0.35154248 -0.25503473 -0.42117052
## [2,]  0.04245024  1.00000000 -0.10621692 -0.23442515 -0.09809202 -0.21437190
## [3,] -0.17553509 -0.10621692  1.00000000  0.35339621  0.44343279  0.61697481
## [4,] -0.35154248 -0.23442515  0.35339621  1.00000000  0.26611397  0.55484390
## [5,] -0.25503473 -0.09809202  0.44343279  0.26611397  1.00000000  0.46794697
## [6,] -0.42117052 -0.21437190  0.61697481  0.55484390  0.46794697  1.00000000
## [7,] -0.01501184  0.67175831 -0.02888023 -0.08141598 -0.02036896 -0.04031242
##             [,7]
## [1,] -0.01501184
## [2,]  0.67175831
## [3,] -0.02888023
## [4,] -0.08141598
## [5,] -0.02036896
## [6,] -0.04031242
## [7,]  1.00000000
```

Normality Check

```
life_lm <- lm(Life.expectancy~
Adult.Mortality+infant.deaths+Alcohol+BMI+GDP+Schooling++log(Population)+Sta
tus, data=df)
summary(life_lm)

#2. normality check (residual)
library(olsrr)

hist(life_lm$residuals)

ols_test_normality(life_lm)

## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9334         0.0000
## Kolmogorov-Smirnov        0.0648         0.0000
## Cramer-von Mises        109.4532         0.0000
## Anderson-Darling         17.4865         0.0000
## -----------------------------------------------
```

Checking homoscedasticity (同質性)

```
#3. checking Homoscedasticity (homogeneity of variances )
ncvTest(life_lm)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.512834, Df = 1, p = 0.21871
```

$$life\ expectancy = \beta_0 + \beta_1 Adult\ Mortality + \beta_2 infant\ deaths + \beta_3 Alcohol + \beta_4 \mathrm{BMI} + \beta_5 \mathrm{GDP} + \beta_6 Schooling + \beta_7 \log(Population) + \beta_8 Status + e$$

```
Call:
lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
    Alcohol + BMI + GDP + Schooling + +log(Population) + Status,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-24.7372  -2.2314   0.3934   2.9431  13.3102

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.765e+01  1.041e+00  55.358  < 2e-16 ***
Adult.Mortality  -3.146e-02  1.017e-03 -30.921  < 2e-16 ***
infant.deaths    -1.292e-03  9.823e-04  -1.315  0.18873
Alcohol          -1.288e-01  3.975e-02  -3.241  0.00122 **
BMI               5.310e-02  6.917e-03   7.676 2.79e-14 ***
GDP               7.048e-05  1.163e-05   6.060 1.69e-09 ***
Schooling         1.375e+00  6.160e-02  22.319  < 2e-16 ***
log(Population)  -3.509e-02  4.215e-02  -0.833  0.40523
StatusDeveloping -1.201e+00  4.220e-01  -2.845  0.00449 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.5 on 1640 degrees of freedom
Multiple R-squared:  0.7396,    Adjusted R-squared:  0.7384
F-statistic: 582.4 on 8 and 1640 DF,  p-value: < 2.2e-16
```

Confidence Intervals for Parameters

- The interval is the set of values for which a hypothesis test to the level of 5% cannot be rejected.
- The interval has a probability of 95% to contain the true value of βi. So in 95% of all samples that could be drawn, the confidence interval will cover the true value of βi.

```
#confidence interval for parameter
confint(life_lm)
```

```
##                              2.5 %        97.5 %
## (Intercept)         5.560889e+01  5.969423e+01
## Adult.Mortality    -3.345107e-02 -2.946036e-02
## infant.deaths      -3.218237e-03  6.350482e-04
## Alcohol            -2.067683e-01 -5.085129e-02
## BMI                 3.953060e-02  6.666517e-02
## GDP                 4.766362e-05  9.328864e-05
## Schooling           1.254058e+00  1.495715e+00
## log(Population)    -1.177721e-01  4.758480e-02
## StatusDeveloping   -2.028706e+00 -3.731117e-01
```

Interpreting Dummy Variables

$$life\ expectancy = \beta_0 + \beta_1 Adult\ Mortality + \beta_2 infant\ deaths + \beta_3 Alcohol + \beta_4 \mathrm{BMI} + \beta_5 \mathrm{GDP} + \beta_6 Schooling + \beta_7 \log(Population) + \beta_8 Status + e$$

$Developing = 1$
$Developed = 0$

How to interpret when β_8= - 1.21?

Developing countries has the 1.21 lower life expectancy rate compared to developed countries.

Try using country variable to see what is the result.

Interaction and Main Effects

$$life\ expectancy = \beta_0 + \beta_1 Adult\ Mortality + \beta_2 infant\ deaths + \beta_3 Alcohol + \beta_4 \mathrm{BMI} + \beta_5 \mathrm{GDP} + \beta_6 Schooling + \beta_7 \log(Population) + \beta_8 Status + \beta_9 Status \times Schooling + e$$

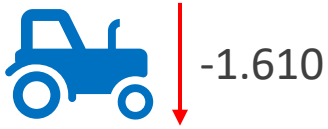
```
#interaction
life_lm2 <- lm(Life.expectancy~
Adult.Mortality+infant.deaths+Alcohol+BMI+GDP+Schooling+log(Populatio
n)+Status+Status*Schooling, data=df)
summary(life_lm2)
```

```
Call:
lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths + 
    Alcohol + BMI + GDP + Schooling + log(Population) + Status + 
    Status * Schooling, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-24.7480  -2.2765   0.2888   2.8701  13.8149 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                7.124e+01  2.674e+00  26.639  < 2e-16 ***
Adult.Mortality           -3.104e-02  1.011e-03 -30.694  < 2e-16 ***
infant.deaths             -1.105e-03  9.742e-04  -1.134   0.2568    
Alcohol                   -1.670e-01  4.000e-02  -4.175 3.14e-05 ***
BMI                        4.726e-02  6.938e-03   6.812 1.35e-11 ***
GDP                        7.585e-05  1.157e-05   6.556 7.39e-11 ***
Schooling                  5.311e-01  1.649e-01   3.221   0.0013 ** 
log(Population)           -2.668e-02  4.181e-02  -0.638   0.5235    
StatusDeveloping          -1.610e+01  2.737e+00  -5.882 4.90e-09 ***
Schooling:StatusDeveloping 9.685e-01  1.758e-01   5.508 4.20e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.46 on 1639 degrees of freedom
Multiple R-squared:  0.7444,    Adjusted R-squared:  0.743 
F-statistic: 530.3 on 9 and 1639 DF,  p-value: < 2.2e-16
```

Interaction Effect

Coefficients	
Schooling	0.531
Developing	-1.610
Schooling*Developing	0.968

When other variables being constant, one year more schooling and in developing countries



$$0.531 \times 1 - 1.610 + 0.531 \times (-0.610) = -1.934$$

Confidence Intervals v.s. Prediction Intervals

- Confidence intervals (信賴區間) express sampling uncertainty in quantities estimated from many data points. The more data, the less sampling uncertainty, and hence the thinner the interval.
  - The mean of the estimation
- Prediction interval (預測區間) is an estimated range of values that may contain the value of a single new observation, based on previous data.

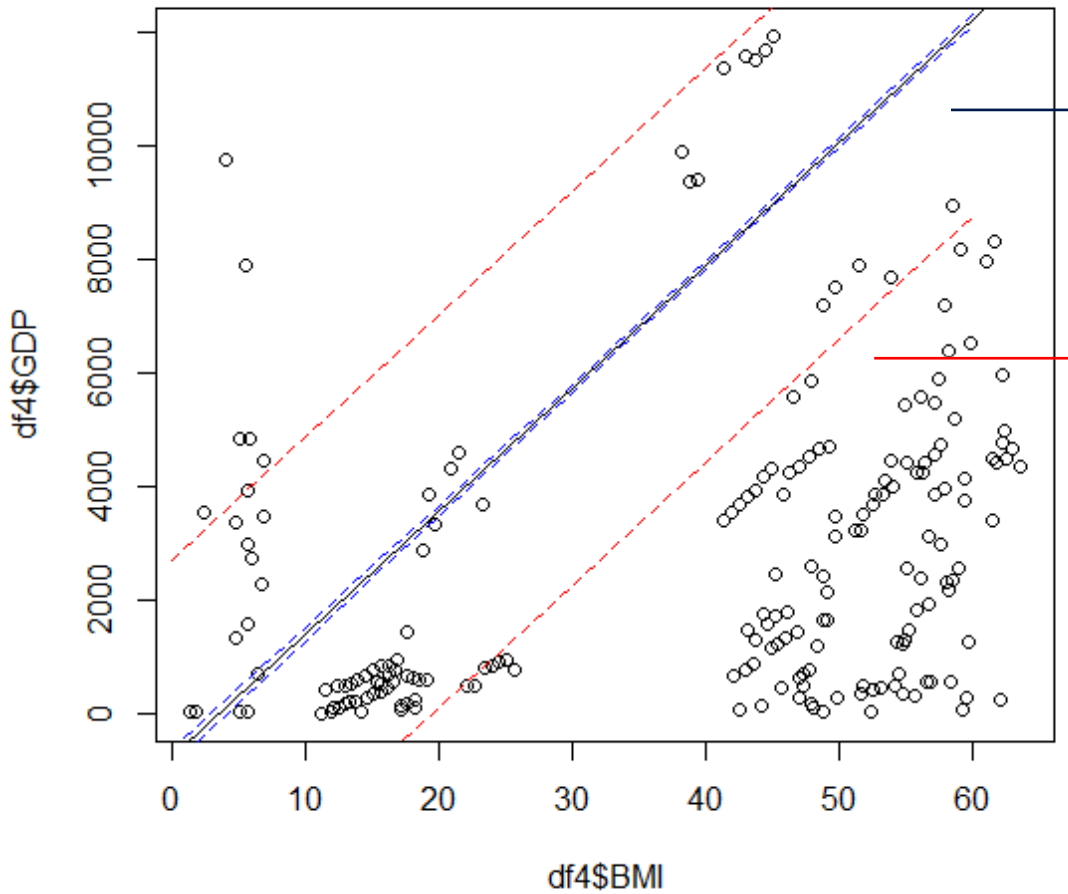
```
predict_data <- df[1:5, ]
life_confidence <- predict(life_lm2, predict_data, interval = "confidence");
life_confidence
```

```
##         fit      lwr      upr
## 1  62.53575 62.06322 63.00829
## 2  62.23746 61.83389 62.64104
## 3  62.03421 61.56415 62.50427
## 4  61.79341 61.39338 62.19344
## 5  61.18908 60.79357 61.58459
```

```
life_prediction <- predict(life_lm2, predict_data, interval = "prediction");
life_prediction
```

```
##         fit      lwr      upr
## 1  62.53575 53.77535 71.29616
## 2  62.23746 53.48051 70.99442
## 3  62.03421 53.27394 70.79448
## 4  61.79341 53.03662 70.55020
## 5  61.18908 52.43249 69.94567
```

```r
df3 <- subset(df, df$GDP<12000)
df4 <- df3[1:200,]
life_lm3 <- lm(GDP~BMI, data=df)
plot(df4$BMI, df4$GDP)
abline(life_lm3)


newx <- seq(0,60, by=5)

cont_interval <- predict(life_lm3, newdata=data.frame(BMI=newx), interval="confidence", level=0.2)
lines(newx, cont_interval[,2], col="blue", lty=2)
lines(newx, cont_interval[,3], col="blue", lty=2)

pre_interval <- predict(life_lm3, newdata=data.frame(BMI=newx), interval="prediction", level=0.2)
lines(newx, pre_interval[,2], col="red", lty=2)
lines(newx, pre_interval[,3], col="red", lty=2)
```



Confidence Interval

Prediction Interval