0.1 Simple Linear Regresion - Lecture 3

$\text{col\_1}_i = \beta_0 + \beta_1 \, \text{col\_2}_i + \varepsilon_i$

β_0 is the population intercept, and β_1 is the population slope parameter measuring how col_1 changes with an associated change in col_2. For each observation i, ε_i represents the random error term.

Slope: a one-unit increase in X (col_2) is associated with an expected increase of β_1 units in Y (col_1).

Intercept: The average value of Y when $X = 0$ is β_0

```
SL_reg <- lm(col1 ~ col2, data = data_frame)
tidy_SL_reg <- tidy(SL_reg) #Get estimate, std.error, statistic,
    p.value
# Extract coefficients
beta_0_hat = SL_reg$coef[1]
beta_1_hat = SL_reg$coef[2]
```

0.2 Conclusions from above

Can be asked to check linearity of the model based on the slope of the coefficients

Slope Coefficient Hypothesis are as follows:-

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

From the regression output we got the following get:

Slope estimate: ¡tidy_SL_reg -¿ estimate column¿

p-value: ¡tidy_SL_reg -¿ p-value column¿

If the p-value is far smaller than significance value α (0.05 usually):

p-value $< 0.05 \quad \Rightarrow \quad$ Reject H_0

Has is extremely strong statistical evidence that the slope coefficient differs from zero. We therefore conclude that col1 is a significant predictor of col2 in the population.

0.3 The range problem

The linear model assumes that the relationship between X and $E[Y—X]$ is linear, which may or may not be true

Sometimes, there's a linear association only in part of the data range. The linear model could still be useful when restricted to that specific range; We need to exercise caution when using the model outside the range of the data, as the relationship between X and Y may differ significantly.

0.4 Multiple Linear Regresion - Lecture 4

$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$

$\varepsilon_i \sim N(0, \sigma^2) \quad$ for all i

$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad$ for $i \neq j$

The MLR model is no longer a line, but a hyperplane in a $(p+1)$-dimensional space. In three dimensions (two predictors), it is a plane.

```
ML_reg <- lm(col1 ~ col2 + col3 + col4, data = data_frame)
library(GGally)
ggpairs(data = dataFrame[, c("col1","col2","col3","col4",)])
```

0.5 "+" and "*" interaction in lm()

In an `lm()` formula, the `+` symbol adds a variable as a main effect only, meaning it includes the predictor in the model without forming any interaction terms. The `*` symbol expands to include both the main effects of the variables and their interaction term—for example, `x1 * x2` is equivalent to `x1 + x2 + x1:x2`. Thus, using `+` simply adds predictors, while using `*` fits a richer model that accounts for combined (interaction) effects between predictors.

```
# Examples in R:
lm(y ~ x1 + x2) # main effects only
lm(y ~ x1 * x2) # x1, x2, and interaction x1:x2
lm(y ~ x1:x2) # interaction only
format(3.904359e-02, scientific = FALSE) #convert exponents to dec
```

The bootstrap sampling distribution of the slope is approximately bell-shaped and symmetric, closely resembling a normal distribution. There are no major signs of skewness or heavy tails, indicating that the sampling distribution of the slope is well-approximated by the normal model commonly assumed in linear regression.