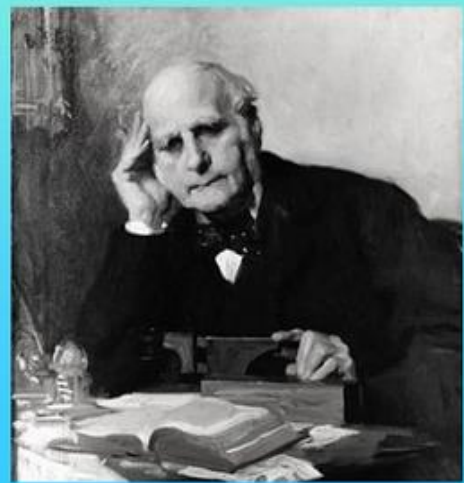
Simple Linear Regression (SLR)

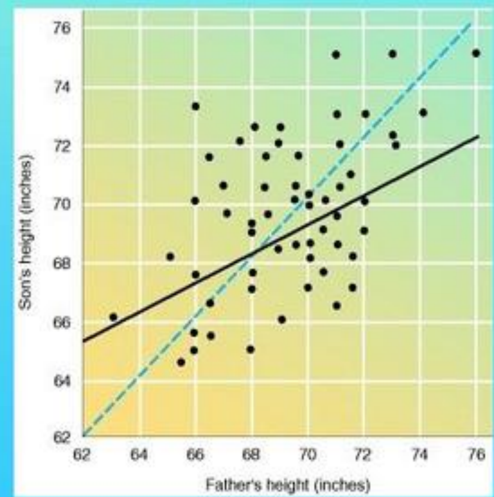
Simple Linear Regression

The purpose of regression?

- Describe functional relationships between variables
- Control
- Prediction of outcomes

Developed by Sir Francis Galton (1822-1911) in his article “Regression towards mediocrity in hereditary structure”





Simple Linear Regression

The basic concepts of regression

- Describe statistical relationships between variables
- The statistical relation has two essential ingredients
  - A tendency of the response variable Y to vary with the predictor variable X
    - There is a probability distribution of Y for each level of X.
  - A scattering of points around the curve of statistical relationship.
    - The means of these probability distributions vary in some systematic fashion with X

Example (diamonds.csv)

Variables:

Response Variable: Price in Singapore dollars (Y)

Explanatory Variable: Weight of diamond in carats (X)

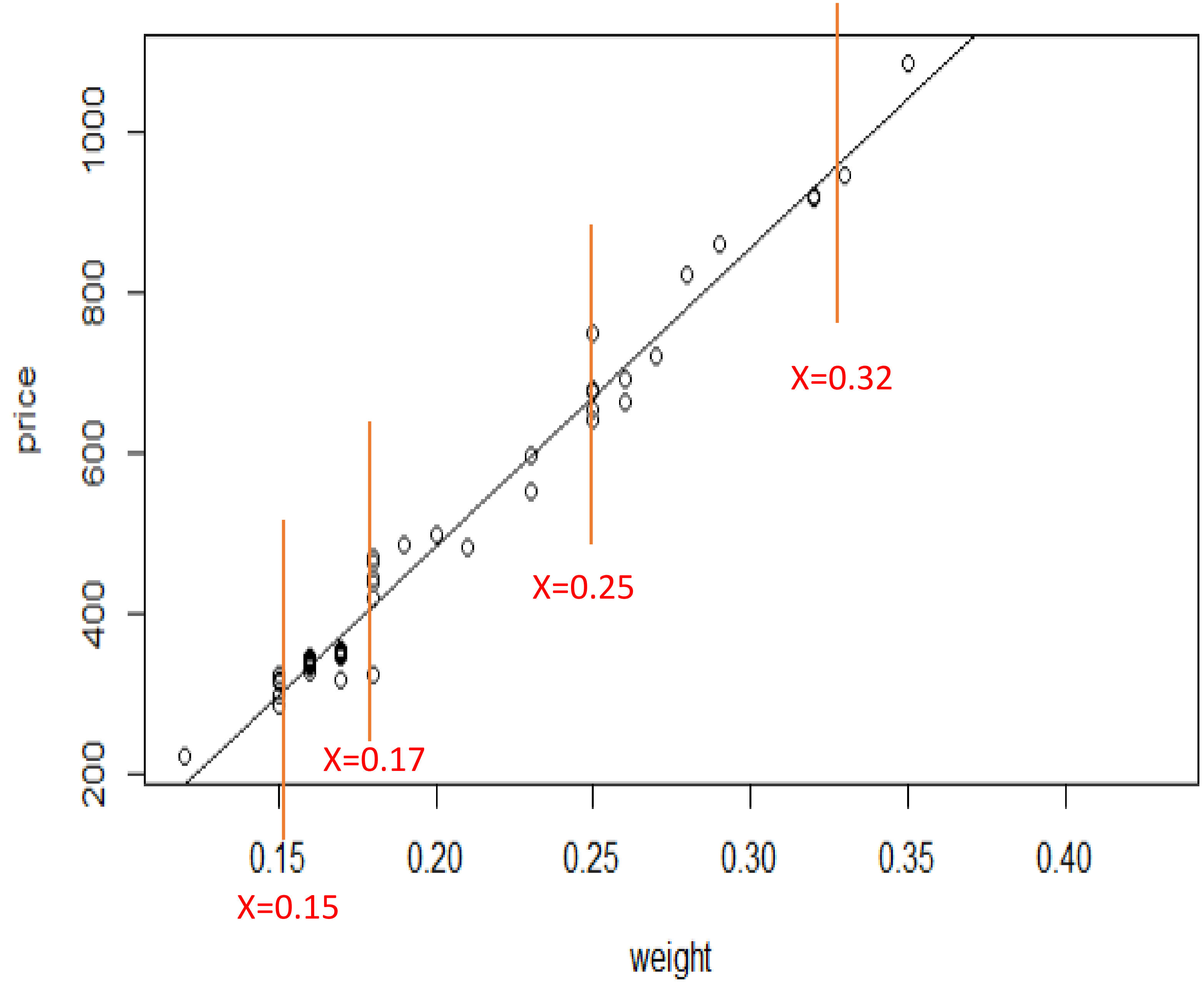
Goal:

Predict the price of a sale for a 0.43 carat diamond ring

What are the two ingredients in understudying statistical relationship between price and weight ?

Scatter plot

Mean price = intercept+ slope (weight)



- The means of the price distributions increase linearly with the weight
- For any given weight, the distribution of price varies, and we can see later that the distribution is Normal (the bell-shape distribution).

Notation for Simple Linear Regression (SLR)

- Observe a pair of variables (explanatory and response) on each of $i = 1, 2, \ldots, n$ samples
- Each pair often called a ***case*** or a ***data point*** (X_i, Y_i)
- Y_i is the value of the response for the i-th case
- X_i is the value of the explanatory variable for the i-th case

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \text{for} \quad i = 1, 2, \ldots, n$$

Simple Linear Regression Model Parameters

- β_0 is the intercept.
- β_1 is the slope.
- ε_i are independent, normally distributed random errors with mean 0 and variance σ^2,

$$\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Features of Simple Linear Regression Model

- Individual observations: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- Since ε_i are random, Y_i are also random and

$$\begin{aligned} \mathrm{E}(Y_i) &= \beta_0 + \beta_1 X_i + \mathrm{E}(\varepsilon_i) = \beta_0 + \beta_1 X_i \\ \mathrm{Var}(Y_i) &= 0 + 0 + \mathrm{Var}(\varepsilon_i) = \sigma^2. \end{aligned}$$

Since ε_i is Normally distributed, $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$

Fitted Regression Equation and Residuals

The parameters $\beta_0, \beta_1,$ and σ^2 are unknown and must be estimated from the data.

The "hat" symbol $\widehat{\ }$ is "point estimation" → $\hat{Y} = b_0 + b_1 X$

- b_0 estimates β_0 (intercept) $\boldsymbol{\hat{\beta}_0 = b_0}$ **or** $\boldsymbol{\hat{\beta}_1 = b_1}$
- b_1 estimates β_1 (slope)
- $\hat{Y}_i = b_0 + b_1 X_i$ gives the estimated mean of Y when the predictor is X_i.
- The *residual* for the i-th case is $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$
- $s^2 = \mathrm{Var}(e_i)$ estimates the error variance σ^2 $\boldsymbol{\widehat{\sigma^2} = s^2}$

The residual e_i (in one sample) is NOT the same as the error ε_i (in population) !

$\hat{\varepsilon} = e$

Estimating the parameters with Least Squares (LS) Solution

- We want to find the "best" estimates, b_0 and b_1.

- Minimize the sum of the squared residuals, $\sum_{i=1}^{n} e_i^2$, i.e., find

$$\underset{(b_0, b_1)}{\arg\min} = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

- How? Calculus!

  1. Take derivatives with respect to b_0 and with respect to b_1.
  2. Set equations equal to zero and solve for both b_0 and b.

Estimating the parameters with Least Squares (LS) Solution

- The best estimates of β_1 and β_0 given the data (X, Y) are:

$$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} = \frac{SS_{XY}}{SS_X}$$

SS is "sum of squares"

- $b_0 = \bar{Y} - b_1\bar{X}$

- This estimate is the "best" because it

1. is *unbiased* (its expected value is equal to the true value)
2. has *minimum variance*

Estimate the parameters with Maximum Likelihood Estimation (MLE)

Our model says that $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$.

Given X_i, the probability of data point i is,

$$f_i = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2}$$

β_0 and β_1 are unknown, but the *likelihood* of the proposed values $(\beta_0^* \ \beta_1^*)$ given the data is,

$$L(\beta_0^*, \beta_1^* | X, Y) = f_1 \times f_2 \times ... \times f_n = \Pi_{i=1}^{n} \mathrm{f_i}$$

L is maximized when $\beta_0^* = b_0$ and $\beta_1^* = b_1$. Thus, the LS estimates, b_0 and b_1, are also the estimated parameter values that are most (probabilistically) consistent with the data!

Estimation of stochastic variance, σ^2

We estimate σ^2 as the sum of the squared residuals, SSE, divided by the degrees of freedom:

$$s^2 = \frac{\Sigma\left(Y_i - \hat{Y}_i\right)^2}{\mathrm{n} - 2} = \frac{\Sigma \mathrm{e}_\mathrm{i}^2}{\mathrm{n} - 2} = \frac{SSE}{n-2} = \frac{SSE}{DFE} = \mathrm{MSE}$$

SSE stands for “sum of squares error”

DFE stands for “degree of freedom of error”

MSE stands for “mean squared error”

$\mathrm{E}\{MSE\} = \sigma^2$ → MSE is an unbiased estimator of σ^2

$s = \sqrt{MSE}$ → This is the **residual standard error, which estimates the residual standard deviation (σ)**

Estimation of stochastic variance, σ^2

We estimate σ^2 as the sum of the squared residuals, SSE, divided by the degrees of freedom:

$$s^2 = \frac{\Sigma\left(Y_i - \hat{Y}_i\right)^2}{\mathrm{n}-2} = \frac{\Sigma \mathrm{e}_\mathrm{i}^2}{\mathrm{n}-2} = \frac{SSE}{n-2} = \frac{SSE}{DFE} = \mathrm{MSE}$$

SSE stands for "sum of squares error"

DFE stands for "degree of freedom of error"

MSE stands for "mean squared error"

$\mathrm{E}\{MSE\} = \sigma^2$ → MSE is an unbiased estimator of σ^2

$$s = \sqrt{MSE}$$

MSE measures variability around the fitted regression line,
A_______(A. smaller/B larger) MSE is preferred and often used as a criterion for model selection

A comment on the notation

We will also estimate variances for other quantities.

These will also be denoted s^2, but will have a subscript to identify them, e.g. $s^2_{\{b_1\}}$.

Without any subscript, s^2 refers to the the estimated variance of the residuals.
And S refers to the standard error of the residuals.

Identifying statistics and estimates in the R output

```
diamond.mod<-lm(price~weight, diamond)
summary(diamond.mod)
anova(diamond.mod)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
weight       3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.84 on 46 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9783,    Adjusted R-squared:  0.9778
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

$Y = \beta_0 + \beta_1 X + \epsilon$

```
Analysis of Variance Table

Response: price
          Df  Sum Sq Mean Sq F value    Pr(>F)
weight     1 2098596 2098596    2070 < 2.2e-16 ***
Residuals 46   46636    1014
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$MSE =$

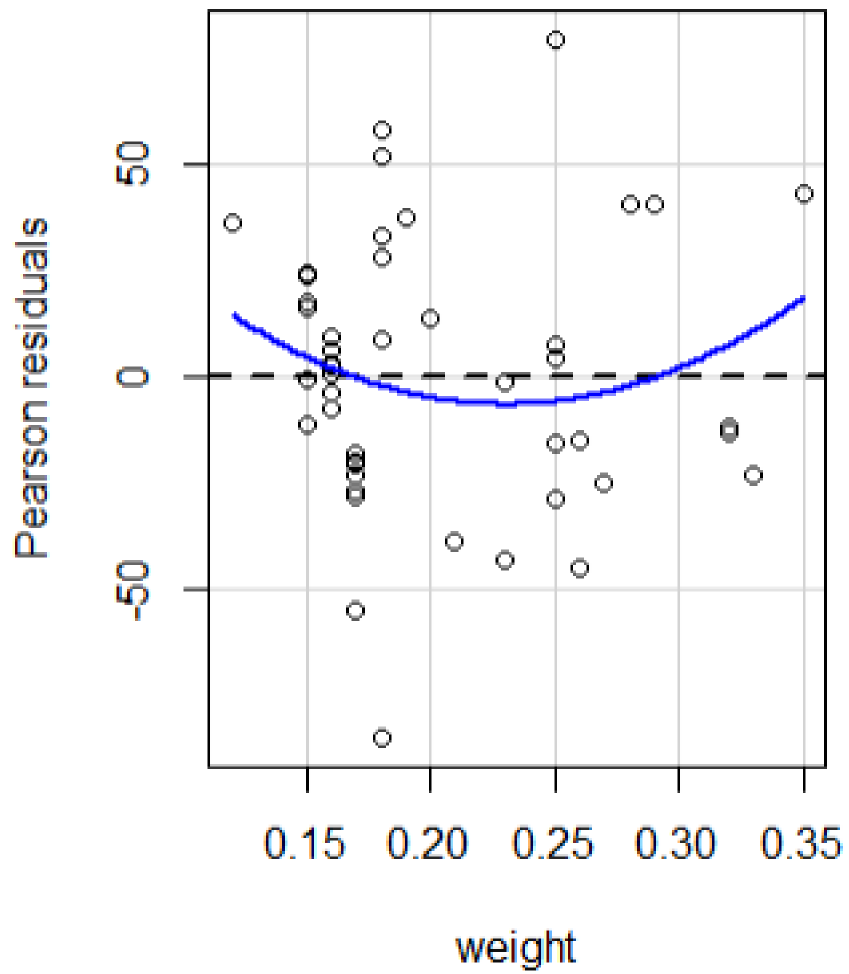
$s = \sqrt{MSE} =$

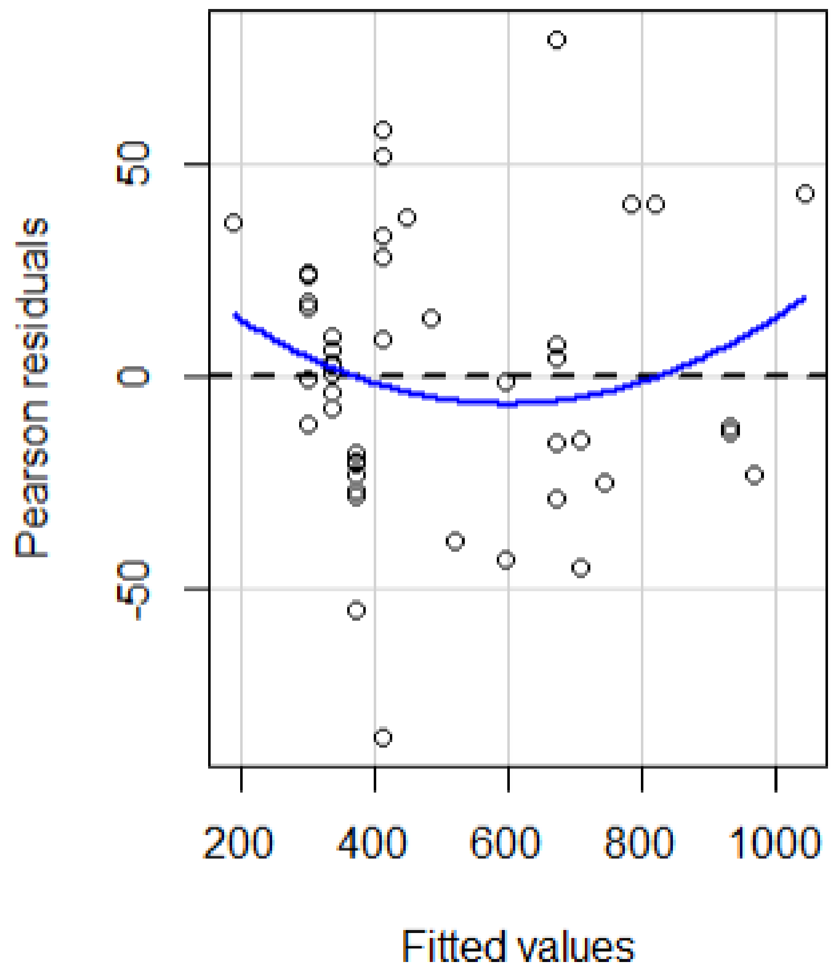
$DF = n - 2 =$

after remove 1 observation

Residual plots

```
residualPlots(diamond.mod)
```





Residuals show a random pattern.

Properties of the LS Line

- The least-squares line always passes through the point $(\bar{X}, \bar{Y})$.

- The residuals always sum to zero:

$$\begin{aligned}\Sigma e_i &= \Sigma [Y_i - (b_0 + b_1 X_i)] \\ &= \Sigma Y_i - b_0 - b_1 X_i \\ &= n\bar{Y} - nb_0 - nb_1\bar{X} \\ &= n[(\bar{Y} - b_1\bar{X}) - b_0] \\ &= 0\end{aligned}$$

- $\sum Y_i = \sum \hat{Y}_i$
- $\sum X_i e_i = 0$
- $\sum \hat{Y}_i e_i = 0$