- ex. Find the average metabolic rate of a women who weighs 59 kg.
    - $\widehat{metabolicrate} = 811.23 + 7.06 \cdot bodywt$
    - $\widehat{metabolicrate} = 811.23 + 7.06 \cdot (59) = 1227.77 \text{ kg/24hr}$
- The average metabolic rate of a women who weighs 59 kg is 1227.77 kcal/24hr.

Note: Estimating outside of the rate of the data is called **extrapolating**.

Caution: The results are not necessarily reliable because we are inferring that the model that we have built is valid outside of the range of our data which may not be a reasonable assumption. In general, one should avoid extrapolating.

Properties of the Fitted Line: $\hat{Y}_i = b_0 + b_1 X_i$

1. $\sum_{i=1}^{n} e_i = 0$
2. $\sum_{i=1}^{n} {e_i}^2$ is a minimum.
3. $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$ (Consequence of (1), $e_i = Y_i - \hat{Y}_i$.) => $\bar{Y} = \bar{\hat{Y}}$
4. $\sum_{i=1}^{n} X_i e_i = 0$
5. $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$
6. The regression line always runs through the point $(\bar{X}, \bar{Y})$.

Properties of the OLS estimators

1. $E(b_0) = \beta_0$
2. $E(b_1) = \beta_1$

Properties (1) and (2) => that b_0 and b_1 are unbiased estimates of β_0 and β_1, respectively.

Proof:

(i) We need to show $E(b_1) = \beta_1$.

$$b_1 = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{\sum(X_i-\bar{X})Y_i - (X_i-\bar{X})\bar{Y}}{\sum(X_i-\bar{X})^2} = \frac{\sum(X_i-\bar{X})Y_i - \bar{Y}\overset{0}{\overset{\shortparallel}{\sum_1^n(X_i-\bar{X})}}}{\sum(X_i-\bar{X})^2}$$

$$= \frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2} = \sum_{i=1}^{n} \frac{(X_i-\bar{X})}{\sum(X_i-\bar{X})^2}\cdot Y_i$$

Now $E(b_1) = E\left(\sum_1^n \frac{(X_i-\bar{X})}{\sum(X_i-\bar{X})^2}\cdot Y_i\right) = \sum_1^n E\left(\frac{(X_i-\bar{X})}{\sum(X_i-\bar{X})^2}\cdot Y_i\right) = \sum_1^n \frac{(X_i-\bar{X})}{\sum(X_i-\bar{X})^2}\cdot E(Y_i)$

$$= \sum\frac{(X_i-\bar{X})}{\sum(X_i-\bar{X})^2}(\beta_0+\beta_1 X_i) = \beta_0\frac{\overset{0}{\overset{\shortparallel}{\sum_1^n(X_i-\bar{X})}}}{\sum(X_i-\bar{X})^2} + \beta_1\frac{\sum(X_i-\bar{X})X_i}{\sum(X_i-\bar{X})^2} \overset{*}{=} \beta_1\frac{\sum(X_i-\bar{X})^2}{\sum(X_i-\bar{X})^2} = \beta_1$$

$*$ $\sum(X_i-\bar{X})X_i = \sum(X_i-\bar{X})X_i - \bar{X}(X_i-\bar{X}) + \bar{X}(X_i-\bar{X}) = \sum(X_i-\bar{X})(X_i-\bar{X}) + \bar{X}\underset{0}{\underset{\shortparallel}{\sum(X_i-\bar{X})}}$

(ii) Show $E(b_0) = \beta_0$ on Homework #2

3. **Gauss-Markov Theorem**: Under assumptions 1-4 of the simple linear regression model, the ordinary least squares (OLS) estimators b_0 and b_1 have *minimum variance* among all *linear unbiased estimators*.

Note: The OLS estimators b_0 and b_1 are said to be the **Best Linear Unbiased Estimators (BLUE)** of β_0 and β_1.

Proof:

We need to show the estimators b_0, b_1 are linear, unbiased and have minimum variance amongst all other linear estimators of β_0 and β_1 respectively

(i) Let's consider the estimator b_1 first.

Linear: $b_1 = \dfrac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \sum \dfrac{(X_i-\bar{X})}{\sum(X_i-\bar{X})^2} \cdot Y_i = \sum k_i Y_i$ where

$k_i = \dfrac{X_i-\bar{X}}{\sum(X_i-\bar{X})^2}$. (We derived this above.)

Since $b_1 = \sum k_i Y_i$ is a linear combination of $Y_i \Rightarrow b_1$ is a linear estimator.

Unbiased: $E(b_1) = \beta_1$ (we've shown this above)

Has minimum variance amongst all unbiased linear estimators of β_1

Let $\tilde{b}_1$ be any other unbiased linear estimator of β_1

$\Rightarrow \tilde{b}_1 = \sum_{i=1}^{n} c_i Y_i$ for some set of constants c_i

Since $\tilde{b}_1$ is and unbiased estimator $E(\tilde{b}_1) = E\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_1^n c_i E(Y_i)$

$= \sum_1^n c_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum_1^n c_i + \beta_1 \sum_1^n c_i X_i = \beta_1$

$\Rightarrow \sum_1^n c_i = 0$ and $\sum_1^n c_i X_i = 1$

$Var(\tilde{b}_1) = Var\left(\sum_1^n c_i Y_i\right) = \sum_{i=1}^{n} c_i^2 Var(Y_i)$ since Y_i, Y_j are uncorrelated

$= \sum_1^n c_i^2 \cdot \sigma^2 = \sigma^2 \sum_1^n c_i^2$ since $Var(Y_i) = \sigma^2$ for all i

Let $c_i = k_i + d_i$ where k_i are the coefficients from b_1

$\Rightarrow Var(\tilde{b}_1) = \sigma^2 \cdot \sum(k_i + d_i)^2 = \sigma^2\left(\sum k_i^2 + 2\sum k_i d_i + \sum d_i^2\right)$

Since $b_1 = \sum_1^n k_i Y_i$ $\quad Var(b_1) = Var(\sum_1^n k_i Y_i) = \sum_1^n k_i^2 Var(Y_i)$ since

Y_i, Y_j are uncorrelated

$Var(b_1) = \sum_1^n k_i^2 \sigma^2$ $\quad$ ($Var(Y_i) = \sigma^2$ for all i)

Notice:

$$\sum_1^n k_i d_i = \sum_1^n k_i (c_i - k_i) = \sum_1^n c_i k_i - \sum k_i^2 = \sum_1^n c_i \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} - \sum_1^n \frac{(x_i - \bar{x})^2}{\{\sum (x_i - \bar{x})^2\}^2}$$

$c_i = k_i + d_i \Rightarrow d_i = c_i - k_i$

$$= \frac{\overset{1}{\overbrace{\sum_1^n c_i X_i}} - \bar{X} \overset{0}{\overbrace{\sum_1^n c_i}}}{\sum (X_i - \bar{X})^2} - \frac{\sum (X_i - \bar{X})^2}{\sum (X_i - \bar{X})^4} = \frac{1}{\sum (X_i - \bar{X})^2} - \frac{1}{\sum (X_i - \bar{X})^2} = 0$$

$$\Rightarrow Var(\tilde{b}_1) = \sigma^2 \sum k_i^2 + 2\sigma^2 \sum k_i d_i + \sigma^2 \sum d_i^2$$

$$= Var(b_1) + 0 + \sigma^2 \sum d_i^2$$

Since $\sum d_i^2 > 0$ for any set of d_i's that aren't all 0

$$\Rightarrow Var(\tilde{b}_1) > Var(b_1)$$

$\Rightarrow b_1$ is the linear, unbiased estimate of β_1 that minimizes the variance.

(ii) We can use a similar argument for b_0 using

$$b_0 = \sum_1^n m_i Y_i \quad \text{where } m_i = \frac{1}{n} + \frac{\bar{x}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

(I'd encourage you to see if you can derive that formula from $b_0 = \bar{Y} - b_1 \bar{X}$