MOTIVATION FOR REGRESSION: Revisited

Example with **One Predictor Variable**

- **Questions of Interest:** Can we predict the number of votes received by Reform party candidate Pat Buchanan based on the total number of votes in that county?
- Was Buchanan's vote in Palm Beach County unusually high? (According to news accounts, many voters in PB county may have inadvertently voted for Buchanan, when in fact they intended to vote for Democratic candidate Al Gore.)
- **Response (dependent) variable: y**=
- **Predictor (independent) variable: x**=
- **Explore the relationship:**
  - Can we use a ***line*** to describe the relationship?

Example with **Multiple Predictor Variables**

- **Questions of Interest:** Can the quality of a French wine be predicted based on some various weather-related factors?
- **Response (dependent) variable: y=**
- **Predictors (independent) variables:** $x_1 =$, $x_2 =$, $x_3 =$
- **Explore the relationship:** We have **added dimensions**
  - With two predictors (X_1, X_2)we use a ***surface*** to describe the relationship**.**

Pitfalls of Regression Analysis

- Lacking an awareness of the assumptions underlying least-squares regression.
- Not knowing how to evaluate the assumptions.
- Not knowing the alternatives to least-squares regression if a particular assumption is violated.
- Using a regression model without knowledge of the subject matter.
- Extrapolating outside the relevant range.
- Assuming that responses depend causally on predictor variables.

Types of Data for Regression: Observational Study vs. Experiment

When a statistical study is designed there are two types of studies to choose from.

-An observational study simply measures variables of interest on a set of individuals.

-An experiment treats groups of individuals differently in order to observe responses in the response (dependent) variable.

**Completely Randomized Design: treatments are assigned randomly, or in such a way that each individual has an equal chance of receiving any one treatment.

- Certain research questions are often answered more effectively by selecting one type of study.
- Studies that are looking for an independent/dependent variable relationship tend to work better with an experimental design.

Uses of Regression Analysis

- Description
- Control
- Prediction

SIMPLE LINEAR REGRESSION MODEL

$$\mathbf{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

The **random error term** $\boldsymbol{\varepsilon_i}$ has**:**

- **Mean:** $\boldsymbol{E(\varepsilon_i) =}$
- **Variance:** $\boldsymbol{\sigma^2(\varepsilon_i) =}$
- **Covariance:** $\boldsymbol{\sigma(\varepsilon_i, \varepsilon_j) =}$ $\quad \forall\ \boldsymbol{i, j\ such\ that\ i \neq j}$

Review of Covariance and Correlation

COVARIANCE

$$\boldsymbol{\sigma(X,Y) = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big] = E(XY) - \big(E(X)\big)\big(E(Y)\big)}$$

- Measures the joint variability of two random variables.
- The sign (positive or negative) shows the tendency of the linear relationship between $X \; and \; Y$.

CORRELATION COEFFICIENT

$$\boldsymbol{\rho(X,Y) = \frac{\sigma(X,Y)}{\sigma(X)\sigma(Y)}}$$

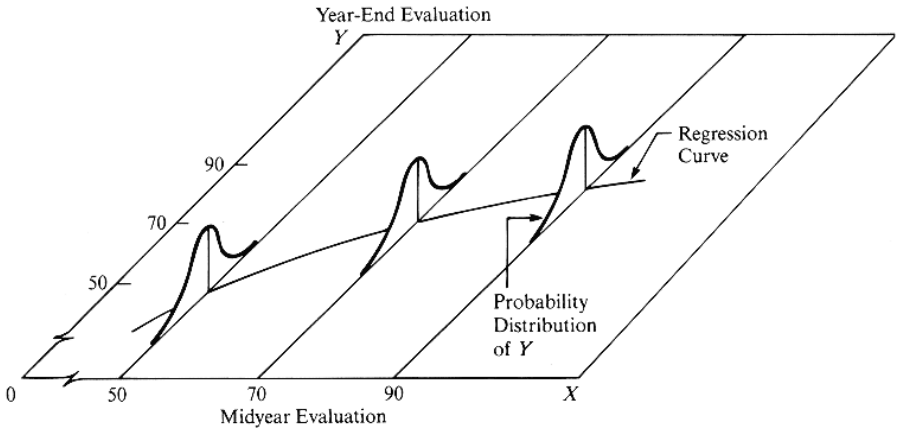
- $-1 \leq \rho(X,Y) \leq 1$
  - Magnitude related to the strength of the linear relationship between $X \; and \; Y$
  - $\rho(X,Y)$=0→ $X \; and \; Y$ are *uncorrelated*

Simple Linear Regression Model: $\boldsymbol{Y_i}$

- Y_i is a **random variable**
  - There is a probability distribution of Y for each level (i) of X where means of the distributions vary in a systematic fashion with X.
  - **Mean:** $E(Y_i) = \beta_0 + \beta_1 X_i$
  - **Variance:** $\sigma^2(Y_i) = \sigma^2$
  - **Covariance:** $\sigma\big(Y_i, Y_j\big) = 0 \; \forall \; i, j \; such \; that \; i \neq j$

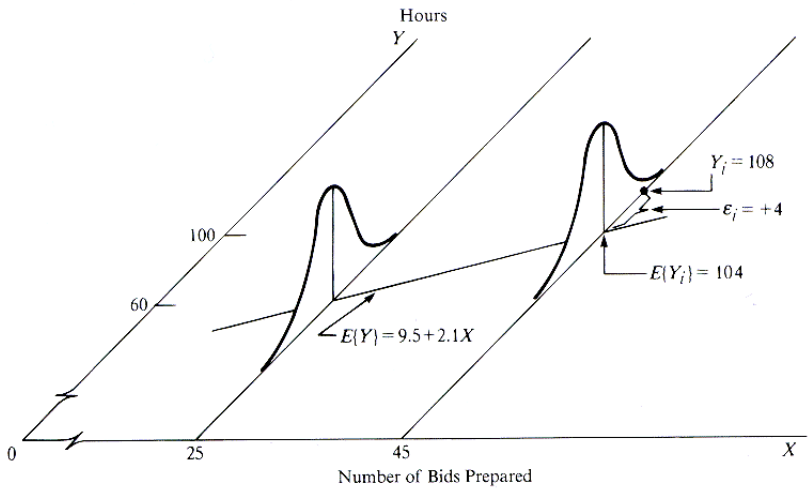
https://nielsen.sites.oasis.unc.edu/soci709/m1/m1004.gif

FIGURE 1.4 Pictorial Representation of Regression Model.

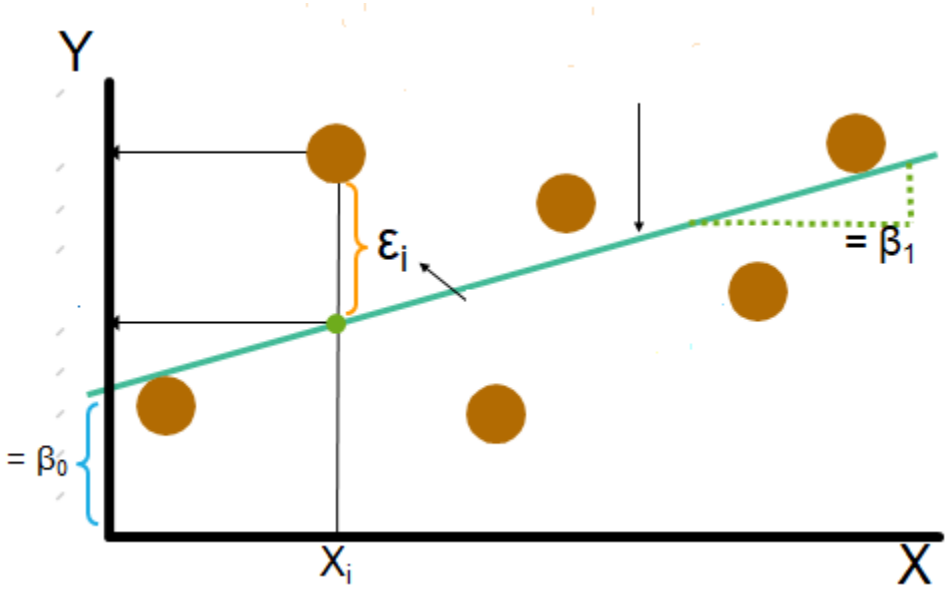


Some distributions based on the model below.

FIGURE 1.6 Illustration of Simple Linear Regression Model (1.1).



http://www.nielsen.sites.oasis.unc.edu/soci708/m15/m1005.gif



Regression Coefficients & The Error Term

Components	Interpretation
β_0	
β_1	
ε_i	

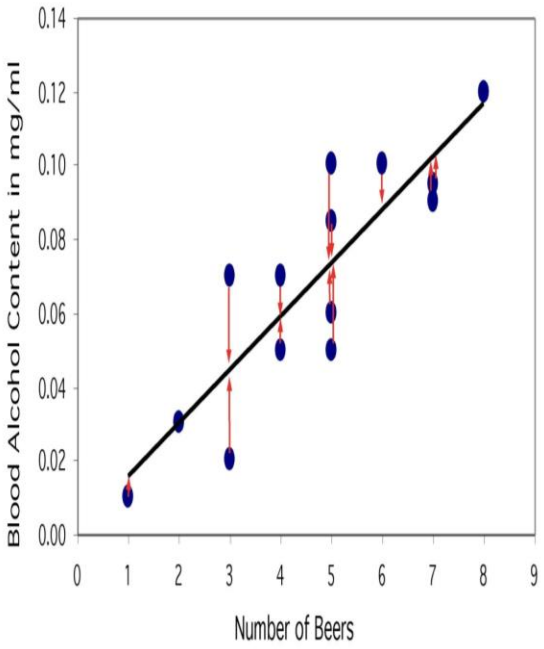
ESTIMATING THE SIMPLE LINEAR REGRESSION MODEL: METHOD OF LEAST-SQUARES

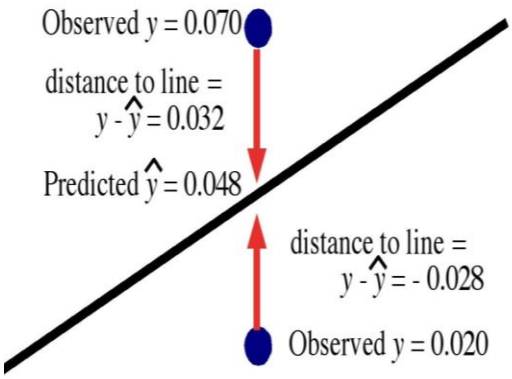
Least-squares regression equation

$$\hat{Y}_i = b_0 + b_1 X_i$$

Components	Interpretation
b_0	
b_1	

How is the LSRL fitted to the data?





The Least Squares Method

$\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ are obtained by finding the values that minimize the sum of the squared differences between Y and $\hat{Y}$:

$$min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = min \sum_{i=1}^{n} (Y_i - (\boldsymbol{b_0} + \boldsymbol{b_1} X_i))^2$$

Review: Special Sums

Sum of Squares for x:

$$S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2 = n\sum_{i=1}^{n} (X_i)^2 - (\sum_{i=1}^{n} X_i)^2$$

Sum of Cross Products:

$$S_{xy} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = n \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i$$

The Least-Squares Estimators

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

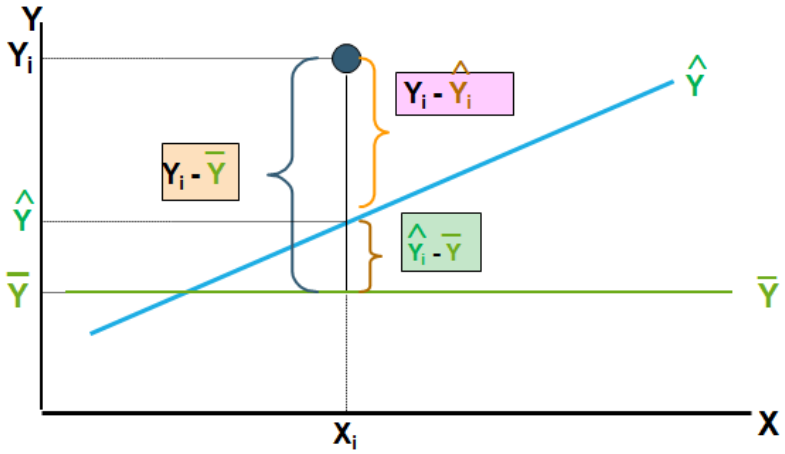
$$b_0 = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i\right) = \bar{Y} - b_1 \bar{X}$$

*** Later we will derive these values of the estimators that minimize the sum of the squared differences between Y and $\hat{Y}$.

Measures of Variation

- Total variation is made up of two parts:

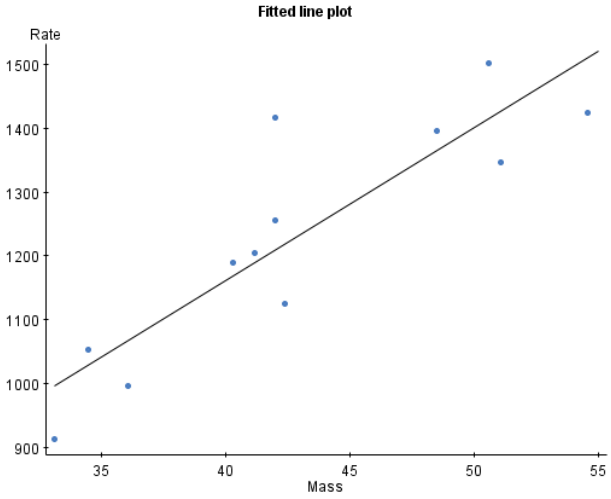
$$SST = SSR + SSE$$



Example **Research Question:** Do heavier people burn more energy?

Researchers decided to look at lean body mass (kg), includes nonfat parts of the body, and rate at which the body consumes energy (calories/24hour day), for 19 people.

a. **Fitted line plot** and **summary statistics** for metabolic rate vs. lean body mass for the females:



Summary statistics for **Mass**(kg):

Group by: Sex

Sex	n	Mean	Variance
F	12	43.033333	47.175152

Summary statistics for **Rate**(cal/24hr):

Group by: Sex

Sex	n	Mean	Variance
F	12	1235.0833	35450.447

b. Compute deviations from the mean, squared deviations from the mean and calculate the values of the least-squares estimators for β_1 and β_0 .

x_i (LBM)	y_i (Rate)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
36.1	995	-6.933333333	-240.0833333	48.0711111	57640.0069	1664.577778
54.6	1425	11.56666667	189.9166667	133.787778	36068.3403	2196.702778
48.5	1396	5.466666667	160.9166667	29.8844444	25894.1736	879.6777778
42	1418	-1.033333333	182.9166667	1.06777778	33458.5069	-189.0138889
50.6	1502	7.566666667	266.9166667	57.2544444	71244.5069	2019.669444
42	1256	-1.033333333	20.91666667	1.06777778	437.506944	-21.61388889
40.3	1189	-2.733333333	-46.08333333	7.47111111	2123.67361	125.9611111
33.1	913	-9.933333333	-322.0833333	98.6711111	103737.674	3199.361111
42.4	1124	-0.633333333	-111.0833333	0.40111111	12339.5069	70.35277778
34.5	1052	-8.533333333	-183.0833333	72.8177778	33519.5069	1562.311111
51.1	1347	8.066666667	111.9166667	65.0711111	12525.3403	902.7944444
41.2	1204	-1.833333333	-31.08333333	3.36111111	966.173611	56.98611111
	SUM	**2.84217E-14**	**9.09495E-13**	**518.926667**	**389954.917**	**12467.76667**

c. The equation of the LSRL is:

Gauss Markov Theorem

Under the given conditions for $\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$, the least squares estimators b_0 and b_1 in

$\boldsymbol{\hat{Y}_i = b_0 + b_1 X_i}$ are ***unbiased*** and have ***minimum variance*** among all unbiased estimators.

- Unbiased estimators

$E(\boldsymbol{b_0}) = \boldsymbol{\beta_0}$ and $E(\boldsymbol{b_1}) = \boldsymbol{\beta_1}$

- Minimum variance
  - The sampling distributions of $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ are less variable (more precise) than those for any other unbiased estimators.

*More on this in the next chapter!

Example: 2000 Elections

a) Treating **Total** votes as the explanatory variable (**X**) and **Buchanan** vote as the response variable (**Y**) find the values of any important descriptive statistics.

Buchanan (Y)

Mean : 260.9 Variance: 202948.5

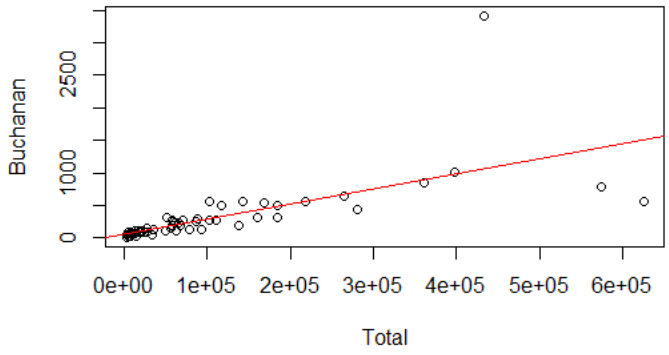
Total (X)

Mean : 88965 Variance: 17407451456

b) Find the **equation** of the LSRL and create a **fitted line plot**.

Coefficients:

(Intercept)	Total
54.229453	0.002323



c) Estimate the mean number (predicted number) of Buchanan votes for Palm Beach County.

```
predict(reg,data.frame(Total=433186))
      1
1060.456
```

d) Interpret the value of the **slope**.

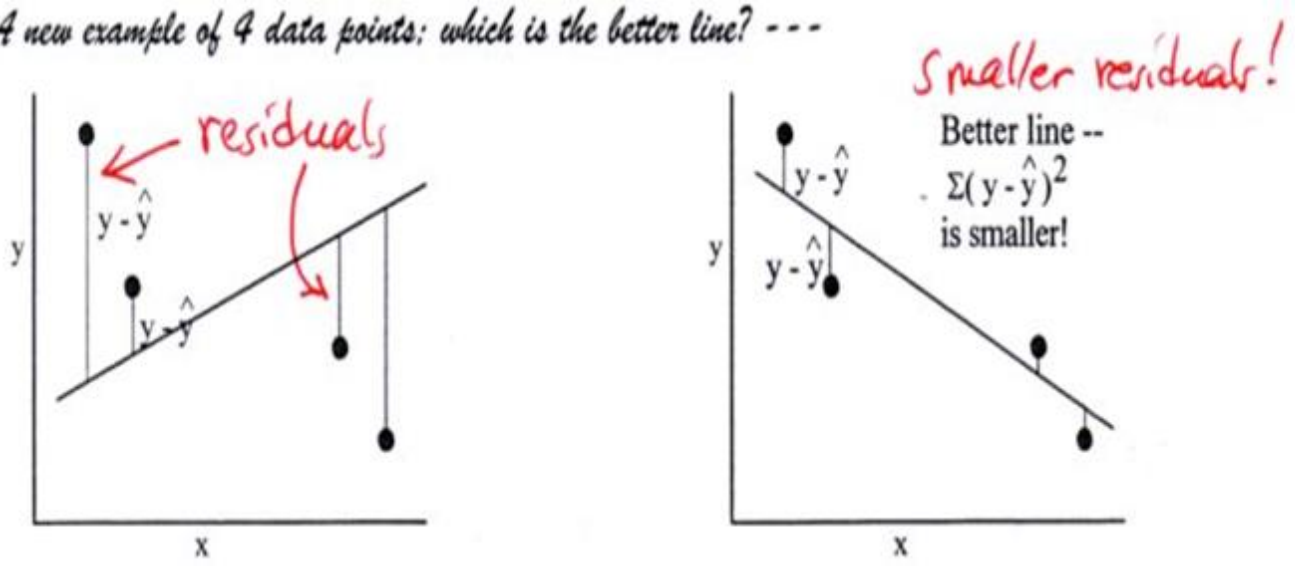
e) Does the **y-intercept** make sense practically?

RESIDUALS

$$e_i = Y_i - \hat{Y}_i$$

- The residual for observation i, e_i, is the difference between its observed and predicted Y value.
- There is a residual for each data point in the sample.
- Graphically a residual measures the vertical distance between a data point and the regression line. The LSRL minimizes the squared residuals overall!

A new example of 4 data points: which is the better line? ---



Properties of the Least-Squares Regression Line (LSRL)

1. The sum of the residuals is zero.

$$\sum_{i=1}^{n} e_i = 0$$

x_i (LBM)	y_i (Rate)	$\hat{y}$	$y_i - \hat{y}$	$(y_i - \hat{y})^2$
36.1	995	1068.502604	-73.50260426	5402.63283
54.6	1425	1512.984836	-87.98483636	7741.33143
48.5	1396	1366.42583	29.5741699	874.631525
42	1418	1210.256397	207.7436028	43157.4045
50.6	1502	1416.88057	85.11943004	7245.31737
42	1256	1210.256397	45.7436028	2092.4772
40.3	1189	1169.412084	19.58791602	383.686454
33.1	913	996.4244045	-83.42440446	6959.63126
42.4	1124	1219.866824	-95.86682384	9190.44791
34.5	1052	1030.060898	21.9391023	481.32421
51.1	1347	1428.893603	-81.89360326	6706.56225
41.2	1204	-1.833333333	12.96445608	168.077121
SUM	14821	14820.99999	7.76E-06	90403.5241

2. The sum of the squared residuals is a minimum.
   $\sum_{i=1}^{n} {e_i}^2$ is minimized by the LSRL

3. The sum of the observed values, Y_i, equals the sum of the predicted/fitted values, $\hat{Y}_i$.
$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$$

4. When the residuals are weighted by the level of the predictor variable in the i^{th} trial, the sum is zero.
$$\sum_{i=1}^{n} X_i e_i = 0$$

5. When the residuals are weighted by the predicted/ fitted value in the i^{th} trial, the sum is zero.
$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0$$

6. The LSRL always passes through the point $(\bar{X}, \bar{Y})$.

Example: In the 200 Elections example
The means are:
mean(Buchanan)= 260.8806
mean(Total)=88964.48

ESTIMATION OF ERROR TERM VARIANCE $\sigma^2(\varepsilon_i) = \sigma^2$

- Needed to indicate variability of probability distributions of Y.
- Also, important for inferences concerning regression function and predicted Y (coming up in next chapters).
- In a single population, we know that σ^2 is estimated *unbiasedly* by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \text{ (shown in Assignment 2).}$$

  - s^2 is often called a mean square because a sum of squared deviations is divided by $df = n - 1$.

SSE (Error Sum of Squares or residual sum of squares)

- Recall that $\sigma^2(\varepsilon_i) = \sigma^2$ and $\sigma^2(Y_i) = \sigma^2$.
- Each Y_i comes from a different probability distribution with a ***different mean*** (but a **constant variance** σ^2).
  - Therefore, we are calculating the squared variations of each Y_i from its estimated mean, or predicted value$\widehat{Y_i}$.
  - So, we are looking at the sum of the squared residuals, which has $df = n - 2$.

$$\boldsymbol{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 =$$

MSE (**E**rror **M**ean **S**quare or residual mean square)

- So, $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} {e_i}^2$ has $df = n - 2$ and we can write the appropriate mean square as $s^2 = \mathbf{MSE} =$
- MSE is an *unbiased* estimator of σ^2 for the regression model $\mathrm{Y_i} = \beta_0 + \beta_1 \mathrm{X_i} + \varepsilon_\mathrm{i}$.
  - Type equation here.

Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$\boldsymbol{s = RMSE =}$$

Example: Back to the Calorie data

x_i (LBM)	y_i (Rate)	$\hat{y}$	$y_i - \hat{y}$	$(y_i - \hat{y})^2$
36.1	995	1068.502604	-73.50260426	5402.63283
54.6	1425	1512.984836	-87.98483636	7741.33143
48.5	1396	1366.42583	29.5741699	874.631525
42	1418	1210.256397	207.7436028	43157.4045
50.6	1502	1416.88057	85.11943004	7245.31737
42	1256	1210.256397	45.7436028	2092.4772
40.3	1189	1169.412084	19.58791602	383.686454
33.1	913	996.4244045	-83.42440446	6959.63126
42.4	1124	1219.866824	-95.86682384	9190.44791
34.5	1052	1030.060898	21.9391023	481.32421
51.1	1347	1428.893603	-81.89360326	6706.56225
41.2	1204	-1.833333333	12.96445608	168.077121
SUM	14821	14820.99999	7.76E-06	**90403.5241**

The Error Sum of Squares:

$\boldsymbol{SSE =}$

The Error Mean Square:

$s^2 = \boldsymbol{MSE=}$

The Standard Error of Estimate:

$s = \boldsymbol{RMSE =}$

DERIVING THE LEAST-SQUARES ESTIMATORS

The Sum of Squared Errors (or the sum of squared vertical distances) for a regression is given by:

$$\textbf{SSE} = \sum_{i=1}^{n} (\boldsymbol{Y_i - \hat{Y}_i})^2$$

We can also write this as:

$$\boldsymbol{Q}(b_0, b_1) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

The Method of Least Squares seeks to find values of b_0 and b_1 to *minimize* $\boldsymbol{Q}(b_0, b_1)$.

What values of b_0 and b_1 minimize SSE = $Q(b_0, b_1)$?*

Some things to think about first.

1. **Derivatives and Partial Derivatives**
   - If $f(x) = 6x^3 + 3$, find $\frac{df}{dx}$.
   - If $f(x, y) = 6x^3 + y$, find $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$.
2. **Basic Sum Rules**
3. **Special Sums**

   $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 =$

   $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) =$
4. **The Least Squares Estimates of b_0 and b_1 for $\hat{Y}$.**
   - $b_0 =$
   - $b_1 =$

Back to our question: **What values of b_0 and b_1 minimize SSE = $Q(b_0, b_1)$?**

$$Q(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$