NORMAL ERROR 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:** $\boldsymbol{\sigma^2(\varepsilon_i\,) =}$

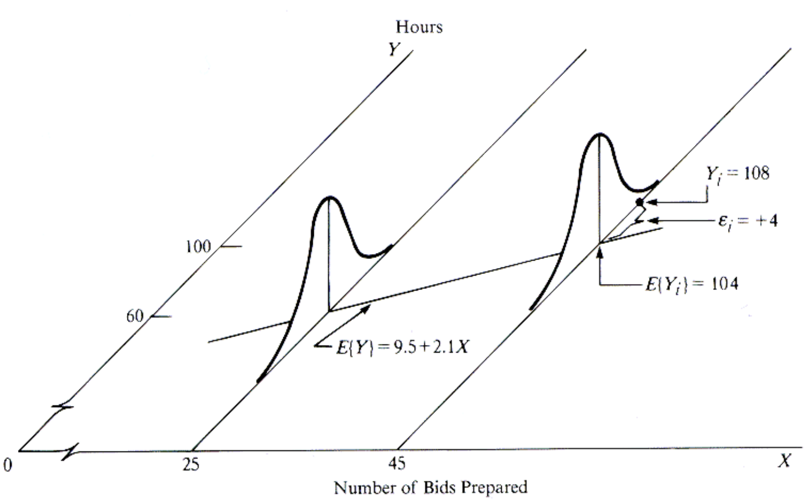
AND $\boldsymbol{\varepsilon_i}$

-
- **Are independent** $\forall\ \boldsymbol{i, j\ such\ that\ i \neq j}$

(The size of the error term for each trial has no effect on the size for any other.)

-
    - The model implies that Y_i are also independent normal random variables (as shown below).

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



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

Note: Both distributions shown are **Normally** distributed with:

- **Mean:** $E(Y_i) =$
- $\boldsymbol{E(Y|X = 25) =}$
- $\boldsymbol{E(Y|X = 45) =}$
- **Variance:** $\boldsymbol{\sigma^2}(Y_i) =$
- Spread is equal in both

**These are two distributions we showed in last week's notes!

Note: **Unless specified**, this **model is assumed** in the remainder of Chapters 2-5!

INFERENCE & SAMPLING DISTRIBUTIONS

As reviewed in the first two weeks, methods of inference are based on sampling distributions of estimators.

For each sampling distribution discussed, we will look at:

- **Shape:**
- **Center:**
- **Spread:**

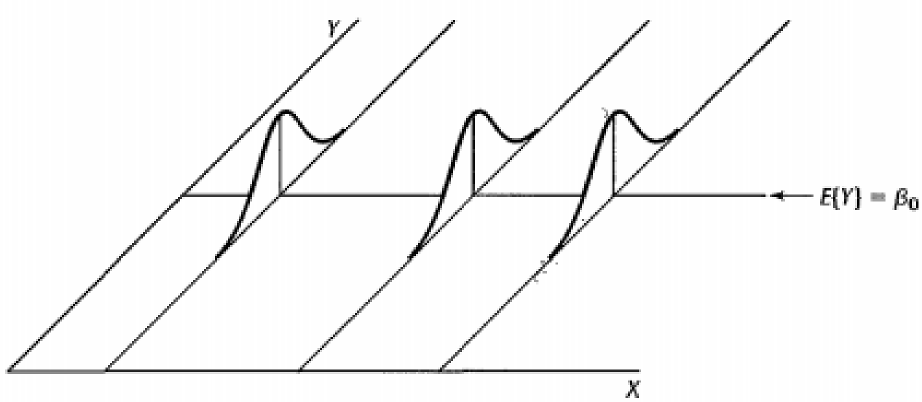
INFERENCE ABOUT THE SLOPE, β_1

Component	Interpretation
(slope) β_1	Change in the expected (average) value of Y per unit change in X

Implications of $\beta_1 = 0$

-
- With the Normal Error Regression Model (shown below)
    -
    -

FIGURE 2.1 Regression Model (2.1) when $\beta_1 = 0$.



Sampling Distribution of $\boldsymbol{b_1}$

- $\boldsymbol{b_1} =$ is the ***point estimator*** of β_1
- The sampling distribution of b_1 means that we are looking at all the different values of b_1 in repeated samples, *while* ***holding the level of the predictor variable constant*** *from sample to sample*.
- Assuming a Normal Error Regression Model, we have:
  - **SHAPE:**
  - **CENTER (MEAN):** $\boldsymbol{E(b_1)} =$
  - **SPREAD (VARIANCE):** $\boldsymbol{\sigma^2(b_1)} =$

SHAPE: WHY **NORMAL**? (b_1 is a linear estimator of β_1)

- A linear combination of the Y_i means that Y_i is not multiplied by itself or another variable, but may be multiplied by constants, and combined by addition or subtraction.

If $Y_1, Y_2, .., Y_n$ are independent normally distributed random variables, the linear combination $a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$ is normally distributed.

-

- Let's define $\boldsymbol{k_i} = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{X_i - \bar{X}}{s_{xx}}$.
- Then we can write the estimated slope as $b_1 =$.
- Some properties of $\boldsymbol{k_i}$:
    1. $\sum_{i=1}^{n} k_i =$
    2. $\sum_{i=1}^{n} k_i X_i =$
    3. $\sum_{i=1}^{n} {k_i}^2 =$

CENTER (MEAN): WHY IS $\boldsymbol{E(b_1) = \beta_1}$? (b_1 is an unbiased estimator of β_1)

SPREAD (VARIANCE): WHY IS $\boldsymbol{\sigma^2(b_1) = \frac{\sigma^2}{s_{xx}}}$?

Recall: Y_i are independent random variables with constant variance $\sigma^2(Y_i) = \sigma^2$.

ESTIMATED VARIANCE of $\boldsymbol{b_1}$

Replacing σ^2 with MSE we get:

$$\boldsymbol{s^2(b_1) =}$$

- $\boldsymbol{s^2(b_1)}$ is an unbiased estimator of $\boldsymbol{\sigma^2(b_1)}$.
- The estimated standard deviation is $\boldsymbol{s(b_1) =}$ and is an unbiased estimator of $\boldsymbol{\sigma(b_1)}$.
- Recall:

$$\text{MSE} = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} {e_i}^2}{n-2}$$

<u>Sampling Distribution of</u> $\frac{(\boldsymbol{b_1}-\boldsymbol{\beta_1})}{s(\boldsymbol{b_1})}$

- We know that the distribution of $\boldsymbol{b_1}$ is Normal. So, then we have that the standardized statistic

- As usual, we need to estimate the **unknown** value of $\boldsymbol{\sigma}(\boldsymbol{b_1})$ using $\mathbf{s}(\boldsymbol{b_1})$. This introduces more uncertainty and so,

  - $\frac{(\boldsymbol{b_1}-\boldsymbol{\beta_1})}{s(\boldsymbol{b_1})}$follows a ***t-distribution*** with $\boldsymbol{df} = \boldsymbol{n} - \mathbf{2}$ (we are estimating **two** parameters, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$, in our regression model)

CONFIDENCE INTERVAL FOR $\boldsymbol{\beta_1}$

lower and upper limits of the $(1-\alpha)\%$ confidence interval for $\boldsymbol{\beta_1}$ are:

How is this based on the sampling distribution of $\frac{(\boldsymbol{b_1}-\boldsymbol{\beta_1})}{s(\boldsymbol{b_1})}$?

Example: Lean Body Mass (LBM) and Calorie Rate

In our regression model predicting a Calorie Rate (in calories per day) based on LBM (in kg), find a 98% confidence interval for the slope, $\boldsymbol{\beta_1}$.

We need to know:

- t **critical values** for **C=0.98** (from R)
    - $\alpha =$ and $1 - \frac{\alpha}{2} =$ $; df =$

- Values of the estimated slope ($\boldsymbol{b_1}$), error mean square($\boldsymbol{MSE}$), sum of squares for X ($\boldsymbol{S_{xx}}$), and estimated standard deviation of slope $\boldsymbol{s(b_1)}$.

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

$$\boldsymbol{b_1} = \frac{S_{xy}}{S_{xx}} =$$

$$\boldsymbol{S_{xx}} =$$

$$\boldsymbol{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i{}^2 = \mathbf{90403.5241} \qquad s^2 = \boldsymbol{MSE} =$$

$$\boldsymbol{s(b_1)} = \sqrt{\frac{\boldsymbol{MSE}}{\boldsymbol{s_{xx}}}} =$$

The **lower** and **upper limits** are:

<u>T TESTS FOR $\boldsymbol{\beta_1}$</u>

We test the null hypothesis $\boldsymbol{H_0}$: .

- Represents ***no linear association*** between X and Y.

The alternative hypothesis, as usual, can be two-sided or one-sided.

- **Two-sided**

$$\boldsymbol{H_A}:$$

  - Represents ***a linear association*** between X and Y.

- **One-sided (lower-tail)**

$$\boldsymbol{H_A}:$$

  - Represents a ***negative*** slope (and thus negative association).

- **One-sided (upper-tail)**

$$\boldsymbol{H_A}:$$

  - Represents a ***positive*** slope (and thus positive association).

- The test statistic is:

  $\boldsymbol{t_0} =$

- <u>Note:</u> This follows the usual formula of $\frac{Estimator - Hypothesized\ Parameter\ Value}{SD\ of\ Estimator}$.

Example: Lean body mass (LBM) and Calorie Rate

Is there a **linear association** between LBM and Calorie Rate? Use $\alpha = 0.02$.

$$\boldsymbol{H_0}:$$

$$\boldsymbol{H_A}:$$

From our earlier example, we had a 98% confidence interval:

Is there a **positive association** between LBM and Calorie Rate? Use $\alpha = 0.02$.

$\boldsymbol{H_0}$: $\boldsymbol{H_A}$:

The test statistic is:

$\boldsymbol{t_0} =$

The p-value is:

$\boldsymbol{p - value} =$

INFERENCE ABOUT THE INTERCEPT, $\boldsymbol{\beta_0}$

Component	Interpretation
(Y-intercept) β_0	Mean value of Y when X=0 **(meaningful when X=0 is within the range of the model)**

SAMPLING DISTRIBUTION OF $\boldsymbol{b_0}$

- $\boldsymbol{b_0} =$ is the ***point estimator*** of β_0
- The sampling distribution of b_0 means that we are looking at all the different values of b_0 in repeated samples, *while **holding the level of the predictor variable constant** from sample to sample*.
- Assuming a Normal Error Regression Model, we have:
    - **SHAPE:**
    - **CENTER (MEAN):** $\boldsymbol{E(b_0)} =$
    - **SPREAD (VARIANCE):** $\boldsymbol{\sigma^2(b_0)} =$
    -
    - **ESTIMATED VARIANCE:** $\boldsymbol{S^2(b_0)} =$
    - **ESTIMATED STANDARD DEVIATION:** $\boldsymbol{S(b_0)} =$

Sampling Distribution of $\frac{(b_0-\beta_0)}{s(b_0)}$

- We know that the distribution of $\boldsymbol{b_0}$ is Normal.
- As usual, we need to estimate the **unknown** value of $\boldsymbol{\sigma(b_0)}$ using $\mathbf{s}(\boldsymbol{b_0})$.
    - $\frac{(b_0-\beta_0)}{s(b_0)}$ also follows a ***t-distribution*** with $\boldsymbol{df = n-2}$

Confidence Interval for $\boldsymbol{\beta_0}$

The **lower** and **upper limits** of the $(\mathbf{1}-\boldsymbol{\alpha})\%$ confidence interval for $\boldsymbol{\beta_0}$ are**:**

Considerations for Inference About $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$

- We are working with t distributions, so must remember that:
    - If the distributions of $Y_i$are not Normal:
        - Sampling distributions of $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$will be when there is not clear non-Normality in distributions of Y_i.
        - Sampling distributions of $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ *approach* Normal as sample size increases.
- **In summary:**

INTERVAL ESTIMATION OF PREDICTED VALUES, $E(Y_h)$

Inference About $\mathrm{E}(Y_h)$

- **GOAL:**
- Let X_h=

(may be an observed value of X occurring in the data or any value within the domain/scope of the model)

- $\mathbf{E}(Y_h) =$

Sampling Distribution Of $\widehat{Y}_h$

- $\widehat{Y}_h =$ is the ***point estimator*** of $E(Y_h)$=
- The sampling distribution of $\widehat{Y}_h$ means that we are looking at all the different values of $\widehat{Y}_h$ in repeated samples, *while **holding the level of the predictor variable constant** from sample to sample*.
- Assuming a Normal Error Regression Model, we have:
    - **SHAPE:**

(follows from fact that $\widehat{Y}_h$ is a linear combination of Y_i)

- **CENTER (MEAN):** $E(\widehat{Y}_h) =$

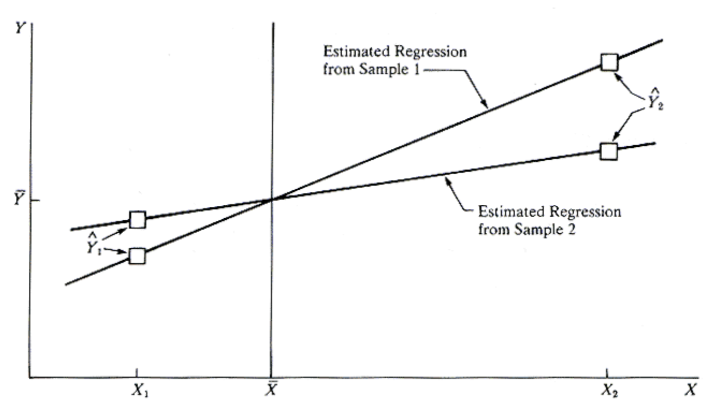
($\widehat{Y}_h$is an unbiased estimator of $E(Y_h)$)

- **SPREAD (VARIANCE):** $\sigma^2(\widehat{Y}_h) =$

Variation in $\widehat{Y}_h$ values will be **greater** from sample to sample when X_h is **far from** the **mean** and **smaller**

when X_h is **closer to** the **mean**.

FIGURE 2.3 Effect on $\widehat{Y}_h$ of Variation in b_1 from Sample to Sample in Two Samples with Same Means $\bar{Y}$ and $\bar{X}$.



- Assuming a Normal Error Regression Model, we have:
    - **ESTIMATED VARIANCE:** $\boldsymbol{S^2(\hat{Y}_h)} =$
    -
    - **ESTIMATED STANDARD DEVIATION:** $\boldsymbol{S(\hat{Y}_h)} =$

<u>**Sampling Distribution of**</u> $\frac{(\hat{Y}_h - E(Y_h))}{s(\hat{Y}_h)}$

$\frac{(\hat{Y}_h - E(Y_h))}{s(\hat{Y}_h)}$ also follows a ***t-distribution*** with $\boldsymbol{df = n - 2}$

<u>**Confidence Interval For**</u> $\boldsymbol{\hat{Y}_h}$

The **lower** and **upper limits** of the $(1-\alpha)\%$ confidence interval for $\boldsymbol{E(Y_h)}$are:

Example: Lean Body Mass (LBM) and Calorie Rate

In our regression model predicting a Calorie Rate (in calories per day) based on LBM (in kg), find a 96% confidence interval for the mean number of calories burned per day, $\boldsymbol{E(Y_h)}$, for a woman with LBM of

$X_h = 50\ kg.$

We need to know:

- t **critical values** for **C=0.96** (from R)
  - $\alpha =$ and $1 - \frac{\alpha}{2} =$; $df =$

- Values of the estimated/predicted mean ($\boldsymbol{\widehat{Y}_h}$), error mean square($\boldsymbol{MSE}$), sum of squares for X ($\boldsymbol{S_{xx}}$), deviation from the mean of X ($(X_h - \bar{X})^2$), and estimated standard deviation $\boldsymbol{S(\widehat{Y}_h)}$.
- The equation of the LSRL is:

$\hat{Y} =$ **201.1616**+ **24.0260666**$\boldsymbol{X_i}$

  - $\boldsymbol{\widehat{Y_h}}$=

$\boldsymbol{MSE} = 9040.352$ $\boldsymbol{S_{xx}}$=518.926667

$\boldsymbol{\bar{X}} =$

$(\boldsymbol{X_h} - \boldsymbol{\bar{X}})^2 =$

$$\boldsymbol{S(\widehat{Y}_h)} = \sqrt{MSE[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{s_{xx}}]} =$$

The lower and upper limits are: