PREDICTION INTERVALS FOR NEW OBSERVATIONS, Y

Goal:

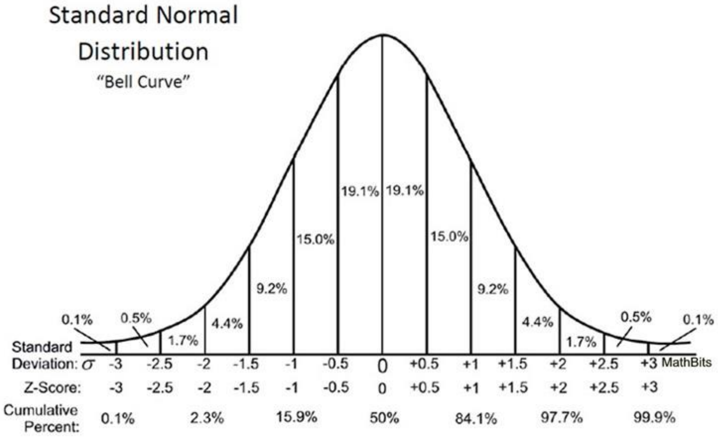
Therefore, we need a **prediction** ***interval*****.**

PREDICTION INTERVAL: PARAMETERS KNOWN

Assuming a Normal error regression model with known parameter values β_0, β_1, $E(Y_h)$, and σ^2,

the **lower and upper limits** of the

$(1-\alpha) \times 100\%$ **prediction interval** for $\boldsymbol{Y_{h(new)}}$ are:



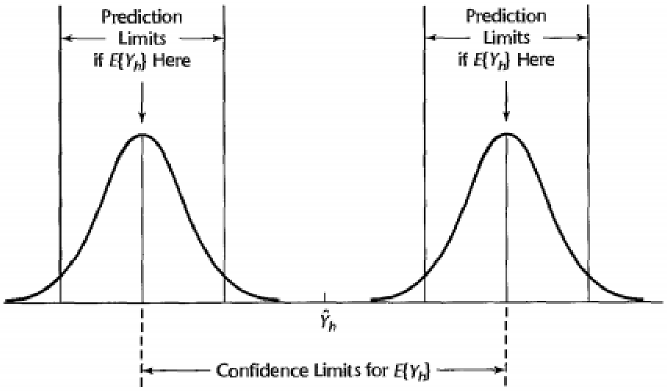
PREDICTION INTERVAL: PARAMETERS UNKNOWN

- The parameter values β_0, β_1, $E(Y_h)$, and σ^2 are unknown, so the mean and variance of the distribution of must be estimated.
    - **MEAN**:
    - **VARIANCE:**

Two major considerations:

1. Variation in the possible center of the distribution of Y (since the center, $E(Y_h)$, is estimated by a confidence interval based on $\widehat{Y_h}$).
2. Variation within the probability distribution of Y (determined by prediction limits based on the estimated spread of the distribution of Y).

FIGURE 2.5 Prediction of $Y_{h(new)}$ when Parameters Unknown.



So, assuming a Normal error regression model, the **prediction limits for $Y_{h(new)}$ are based on** :

- is a point estimate for the unknown mean E(Y_h)
- $pred =$ =prediction error:

  How far might the new observation vary from the estimated mean?

The lower and upper limits of the $(1-\alpha) \times 100\%$ **prediction interval** for $Y_{h(new)}$ are:

- The **variance o**f the **prediction error**, $pred = (\boldsymbol{Y_{h(new)}} - \widehat{\boldsymbol{Y_h}})$, is:

  $$\sigma^2(pred) =$$

- It is estimated, *unbiasedly*, by: $s^2(pred) =$

Example: Lean Body Mass (LBM) and Calorie Rate

Previously, our regression model predicted Calorie Rate (in calories per day) based on LBM (in kg). Suppose a new woman is added to the study whose lean body mass is$X_h = 50\ kg.$ Find a 96% prediction interval for the calorie rate for this new subject.

We need to know:

- From previous example: t **critical values** for **C=0.96** (from R), value of the estimated/predicted mean ($\boldsymbol{\hat{Y}_h}$), value of error mean square($\boldsymbol{MSE}$), and value of the estimated standard deviation $\mathbf{s}(\boldsymbol{\hat{Y}_h})$.
  - $\alpha = 1 - 0.96$=0.04 and $1 - \frac{\alpha}{2} = 0.98$

    $df = 12 - 2 = 10$→qt(**0.98**,10)→[1] 2.359315→ $\pm\, \boldsymbol{t(0.98; 10)} =$
  - $\boldsymbol{\hat{Y}_h}$= **201.1616+ 24.0260666(50)=**
  - $\boldsymbol{MSE} =$
  - $\boldsymbol{S(\hat{Y}_h)} = \sqrt{9040.352[\frac{\mathbf{1}}{\mathbf{12}} + \frac{48.53445}{518.926667}]}$=
- The value of $\mathbf{s}(\boldsymbol{pred})$.
  - $\boldsymbol{s(pred)} = \sqrt{\boldsymbol{MSE} + \boldsymbol{s^2(\hat{Y}_h)}} =$

  **Note: The sample size is still 12 since our prediction is based on the original regression model.

The **lower and upper limits** are:

Note: This **interval is wider than the interval found previously for the mean:

$$1308.125 \leq \boldsymbol{E(Y_h)} \leq 1496.805$$

since the prediction interval incorporates the variability of estimated mean calorie rate from sample to sample as well as the variability of calorie rates from one woman to the next (for women with LBM=50 kg).

Final Comments about Prediction Intervals for One Observation

- As we saw in the last example, the **prediction interval** for the same level of X and same confidence level will always be **wider than** the **confidence interval** since prediction intervals must consider the variability (spread) of the distribution of Y in addition to the sampling variability of $\widehat{Y}_h$.
- Unlike confidence limits, **prediction limits are sensitive to distinct non-normality** of the distribution of error terms.
- **Prediction limits** apply to **one prediction** based on the sample data.
- A **prediction interval** is **statement about** the **value taken on by a random variable**.

(This is not the same as a confidence interval which is intended to capture the true value of a parameter.)

PREDICTION OF M NEW OBSERVATIONS FOR GIVEN X

Goal:

The **lower** and **upper limits** of the $(1-\alpha)\times 100\%$ prediction interval are:

Where:

$$s^2(predmean) =$$

Example: Lean Body Mass (LBM) and Calorie Rate

Suppose a 4 new women are added to the study with lean body mass is $X_h = 50\ kg$. Find a 96% prediction interval for the mean calorie rate for these new subjects.

- The value of $\boldsymbol{s(predmean)}$.

$$\boldsymbol{s(predmean)} = \sqrt{\frac{\boldsymbol{MSE}}{\boldsymbol{m}} + \boldsymbol{s^2(\hat{Y}_h)}} =$$

The lower and upper limits are:

__Note:__** This **interval is wider** than the interval found previously for the mean:

$$1308.125 \leq \boldsymbol{E(Y_h)} \leq 1496.805$$

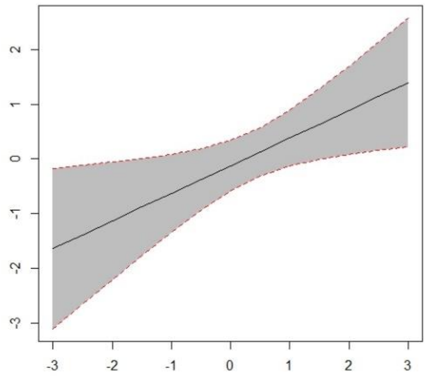
but **narrower** than the interval found previously for one new observation:

$$1159.109 \leq \boldsymbol{Y_{h(new)}} \leq 1645.821$$

CONFIDENCE BANDS FOR THE REGRESSION LINE

Goal:

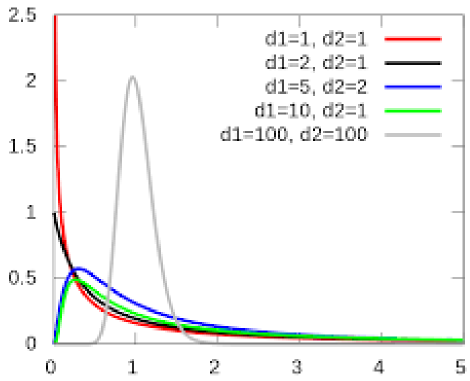
This region is called a **confidence band**.



Assuming the Normal error regression model, the Working-Hotelling $(1-\alpha) \times 100\%$ **confidence band** for $E(Y_h) = \beta_0 + \beta_1 X_h$ at any level of X_h has **boundary points**:

Where:

$\boldsymbol{W^2} =$ is used as the critical value rather than $t\left(1-\frac{\alpha}{2}; n-2\right)$.



Example: Lean Body Mass (LBM) and Calorie Rate-How accurately are we able to estimate the regression function? Find a 96% confidence band for the regression line.

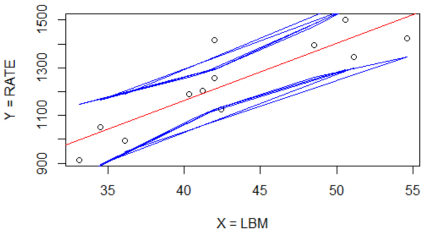
To show the calculation for $X_h = 50\ kg$ we need to know:

- $\boldsymbol{W}$ **critical values** for **C=0.96** (based on F and from R).
    - $\text{C} = 1 - \alpha = 0.96; df_2 = 12 - 2 = 10 \rightarrow$

The **boundary points** at level $\boldsymbol{X_h = 50\ kg}$ are:

****<u>Note:</u>** This **interval is wider** than the interval found previously for the mean:

$$1308.125 \leq \boldsymbol{E(Y_h)} \leq 1496.805$$



The whole confidence band from R.

<u>Final Comments about Confidence Bands</u>

- $\boldsymbol{W}$ **is larger than** $\boldsymbol{t}$ and thus the **boundary points are farther apart** than confidence interval limits.

- Boundary points will still be farther apart the further X_h is from the mean $\bar{X}$.

- $\boldsymbol{(1-\alpha) \times 100\%}$ represents

ANOVA APPROACH TO REGRESSION ANALYSIS

This is a different approach to testing the same hypothesis, $\boldsymbol{H_0}$:

- Based on partitioning the sums of squares and degrees of freedom associated with Y.

Partitioning Total Variation

$$\textbf{SST} \quad = \quad \textbf{SSR} \quad + \quad \textbf{SSE}$$

$$\boldsymbol{SST = \sum(Y_i - \overline{Y})^2}$$

$$\boldsymbol{SSR = \sum(\widehat{Y}_i - \overline{Y})^2}$$

$$\boldsymbol{= b_1^2 s_{XX}}$$

$$\boldsymbol{SSE = \sum(Y_i - \widehat{Y}_i)^2}$$

Example: Lean body mass (LBM) and calorie rate

x_i (LBM)	y_i (Rate)	$\hat{y}_i$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
36.1	995	1068.502604	57640.0069	27749.1393	5402.63283
54.6	1425	1512.984836	36068.3403	77229.24538	7741.33143
48.5	1396	1366.42583	25894.1736	17250.85146	874.631525
42	1418	1210.256397	33458.5069	616.3767578	43157.4045
50.6	1502	1416.88057	71244.5069	33050.23525	7245.31737
42	1256	1210.256397	437.506944	616.3767578	2092.4772
40.3	1189	1169.412084	2123.67361	4312.712992	383.686454
33.1	913	996.4244045	103737.674	56958.08433	6959.63126
42.4	1124	1219.866824	12339.5069	231.5421612	9190.44791
34.5	1052	1030.060898	33519.5069	42034.19911	481.32421
51.1	1347	1428.893603	12525.3403	37562.42073	6706.56225
41.2	1204	-1.833333333	966.173611	1940.207752	168.077121
	SUM	14820.99999	389954.917	299551.392	**90403.5241**

The Regression Sum of Squares:

$$\boldsymbol{SSR = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 =}$$

The Error Sum of Squares:

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

Total Sum of Squares:

$$\boldsymbol{SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 =}$$

Why does $\boldsymbol{SST = SSR + SSE}$ follow from $\boldsymbol{Y_i - \overline{Y} = \left(\widehat{Y}_i - \overline{Y}\right) + (Y_i - \widehat{Y}_i)}$?

Partitioning of Degrees of Freedom

SST $=$ SSR $+$ SSE

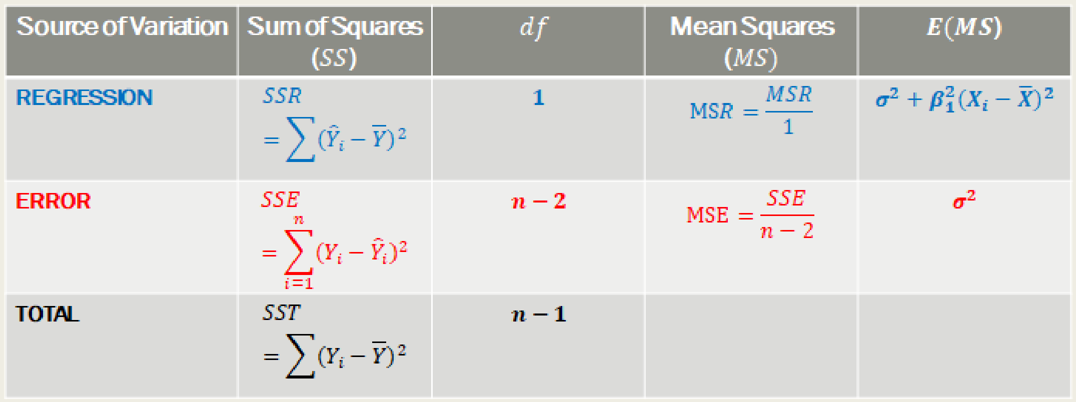
Mean Squares

Previously we have worked with $\boldsymbol{MSE}$, error mean square.

In general, a **mean square** is .

So, $\boldsymbol{MSR} =$ $= \boldsymbol{SSR}$ and $\boldsymbol{MSE} =$.

One-Way ANOVA Table for Regression



Source of Variation	Sum of Squares (SS)	df	Mean Squares (MS)	$E(MS)$
REGRESSION	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	$\mathbf{1}$	$\text{MSR} = \frac{MSR}{1}$	$\boldsymbol{\sigma^2 + \beta_1^2(X_i - \bar{X})^2}$
ERROR	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$\boldsymbol{n-2}$	$\text{MSE} = \frac{SSE}{n-2}$	$\boldsymbol{\sigma^2}$
TOTAL	$SST = \sum(Y_i - \bar{Y})^2$	$\boldsymbol{n-1}$		

Expected Value of Mean Squares

- $\boldsymbol{E(MSE)} =$
  - Center(mean) of the sampling distribution of MSE is $\boldsymbol{\sigma^2}$ even if $\beta_1 \neq 0$.
- $\mathbf{E}(\boldsymbol{MSR}) =$
  - Show that $\boldsymbol{MSR}$ is an unbiased estimator of $\boldsymbol{\sigma^2 + \beta_1^2 s_{XX}}$.
  - When $\beta_1 = 0$ $\mathbf{E}(\boldsymbol{MSR}) = \boldsymbol{\sigma^2}$ and when $\beta_1 \neq 0$ $\mathbf{E}(\boldsymbol{MSR}) > \boldsymbol{\sigma^2}$ and hence $\boldsymbol{MSR}$ will tend to be larger than $\boldsymbol{MSE}$.

F TEST for TESTING $\boldsymbol{H_0}$: $\boldsymbol{\beta_1 = 0}$ vs $\boldsymbol{H_A}$: $\boldsymbol{\beta_1 \neq 0}$.

- The **test statistic** is:

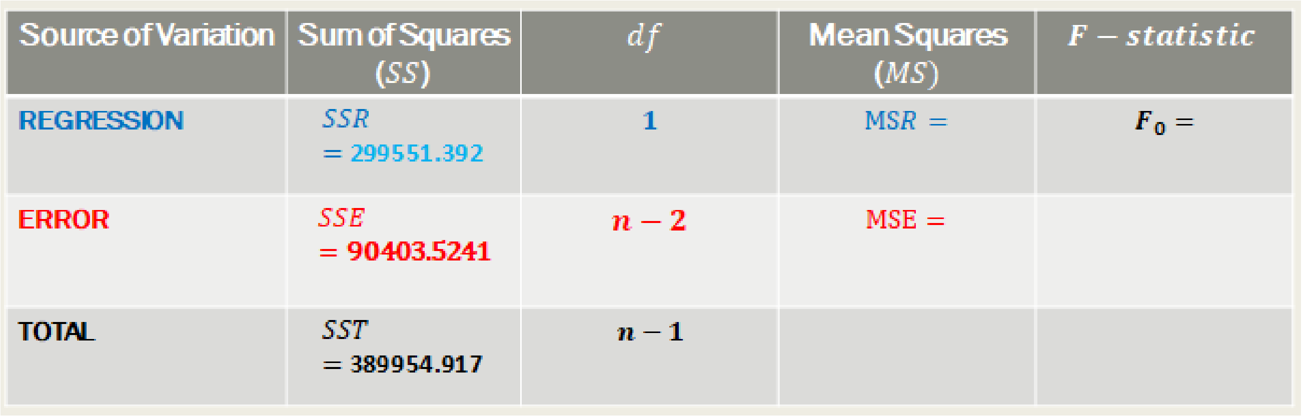
$$\boldsymbol{F_0} =$$

  - Large values (far from 1) support H_A, so the test is really an .
  - Sampling distribution: Assuming $\boldsymbol{H_0}$: $\boldsymbol{\beta_1 = 0}$, .

- **Decision Rule:**
  - Reject H_0 if .
  - Otherwise, fail to reject H_0.

Example: Lean body mass (LBM) and calorie rate

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



Source of Variation	Sum of Squares (SS)	df	Mean Squares (MS)	$F - statistic$
REGRESSION	$SSR = 299551.392$	$\mathbf{1}$	$MSR =$	$\boldsymbol{F_0} =$
ERROR	$SSE = 90403.5241$	$\boldsymbol{n-2}$	$MSE =$	
TOTAL	$SST = 389954.917$	$\boldsymbol{n-1}$		

(Note: The two-sided p-value for the previous part with the t test was $2 * P(t(10) = 2 *$ 9.178927e-05 $=$ 0.0001835785.)

Descriptive Measures of Linear Association

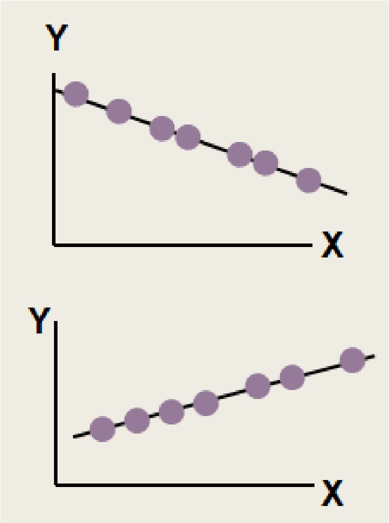
- No single descriptive statistic can fully explain whether a particular regression model is useful.
    - Two common statistics:
        - **Coefficient of Determination**, R^2 (or r^2)

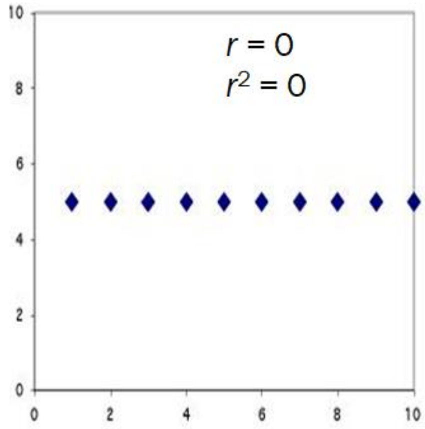
          $R^2 = (r)^2 =$

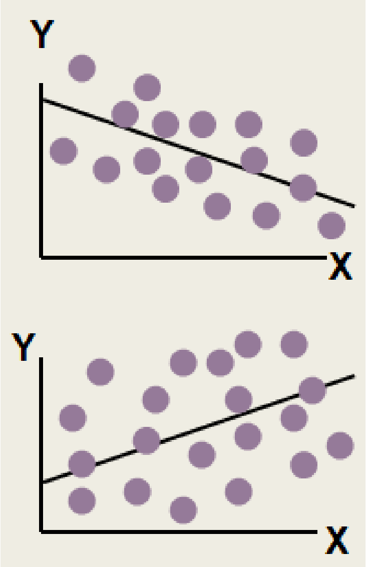
The **coefficient of determination is** the proportion of the total variation in the response variable (Y) that is explained by variation in the explanatory variable (X), or the proportion of variation in Y that is explained by the regression model.

$\boldsymbol{R^2}$**=1**

$\mathbf{0 < \boldsymbol{R^2} < 1}$







- **Correlation Coefficient**, r

  $\boldsymbol{r} =$

- *The closer to –1, the stronger the negative linear relationship*
- *The closer to 1, the stronger the positive linear relationship*
- *The closer to 0, the weaker the linear relationship*