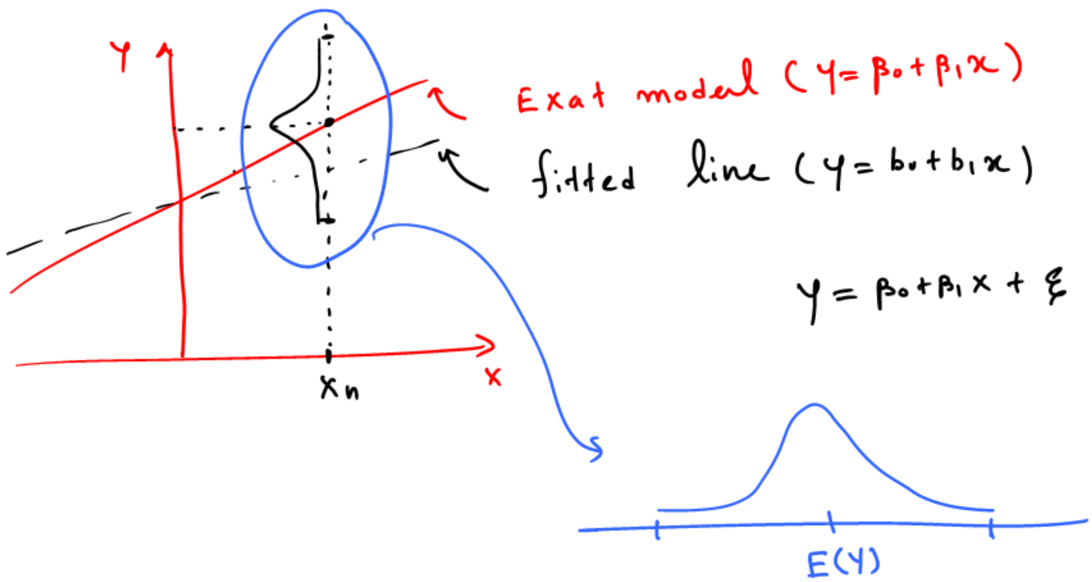
where $S\{\hat{Y}_h\} = \sqrt{MSE\left[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]}$.

HW - 2 : M (1-29) | Quiz - 1 : M (1-29) | HW - 3 : M (2-05) | Quiz - 2 : M (2-05) | Class-3

Prediction Interval for a new observation ($\hat{Y}_{h(new)}$)

consider a new obsenation $\hat{Y}_{h(new)}$ at $X = X_h$. Suppose $\hat{Y}_{h(new)}$ is independent of the observations on which the regression model is based.



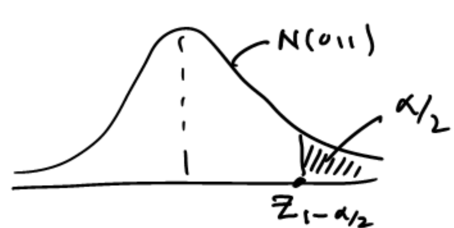
* when model parameters known

$$\hat{Y}_{h(new)} \sim N\left(E(Y), \sigma^2\right)$$

$\therefore$ $(1-\alpha)100\%$ Prediction interval for $\hat{Y}_{h(new)}$ is

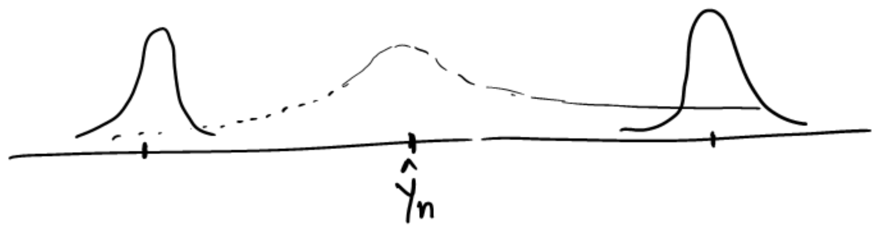
$$E(Y_h) \pm Z_{1-\alpha/2}\,\sigma,$$

where $P(Z > Z_{1-\alpha/2}) = \alpha/2$, and $Z \sim N(0,1)$.



* When model parameters unknown

Now we have to estimate the mean response $E(Y_n)$ too.



So the point estimator for $\hat{Y}_{n(new)}$ is $\hat{Y}_n$.

But,
There are two variations to consider

1. variation of $\hat{Y}_{n(new)}$
2. variation of $\hat{Y}_n$.

Suppose the total variation is $\sigma^2\{Pred\}$.

Then considering all of the above,

$$\frac{\hat{Y}_{n(new)} - \hat{Y}_n}{\sigma\{Pred\}} \sim N(0,1) \text{, when } \sigma^2 \text{ is know.}$$

* when σ^2 is unknown

not a constant

$$\frac{\hat{Y}_{n(new)} - \hat{Y}_n}{S\{pred\}} \sim t_{n-2}.$$

constant

$$\frac{b_1 - \beta_1}{\sigma\{b_1\}}$$

$$\sigma^2\{b_1\} = \sigma^2\{\underline{b_1} - \beta_1\}$$

linear combination of $Y_1, Y_2, \dots Y_n$

Now,
$$\sigma^2\{Pred\} = \sigma^2\{\hat{Y}_{n(new)} - \hat{Y}_n\} = \sigma^2\{\hat{Y}_{n(new)}\} + \sigma^2\{\hat{Y}_n\} \ (\because \text{independent})$$

$$= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(X_n - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]$$

$$= \sigma^2\left(1 + \frac{1}{n} + \frac{(X_n - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right).$$

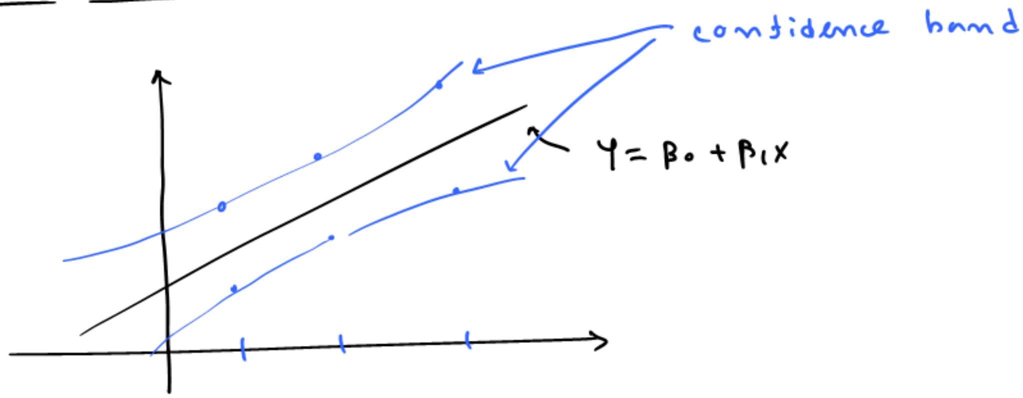
* when σ^2 is unknown,

$$S^2\{pred\} = MSE\left[1 + \frac{1}{n} + \frac{(X_n-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right].$$

$\therefore (1-\alpha)100\%$ prediction interval for $\hat{Y}_{n(new)}$ is

$$\hat{Y}_n \pm t_{1-\alpha/2}\, S\{pred\}.$$

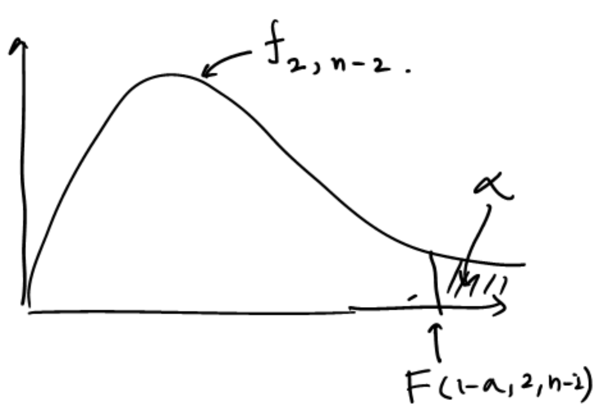
Confidence band for the regression line



Working Hotelling $(1-\alpha)100\%$ confidence level for the regression line for any level of X_n is given by

$$\hat{Y}_n \pm W\, S\{\hat{Y}_n\}, \quad \text{where} \quad W^2 = 2\underbrace{F(1-\alpha,\, 2,\, n-2)}_{}$$

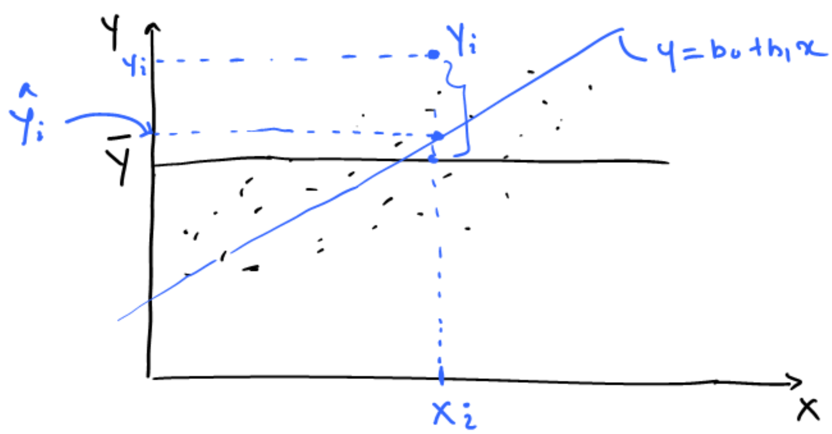
F distribution value with degrees of freedom 2 and $n-2$.



Note:

* $|W|$ value is larger than $t(1-\alpha/2, n-2)$. So the boundary points of the confidence band for X_h is wider than the corresponding confidence interval.

Analysis of variance approach



$Y = \mu + \varepsilon_i$ – center + error model

LSE of $\mu = \hat{\mu} = \bar{Y}$

when $\beta_1 = 0$.

Total variation of variable Y (without fitting a regression model) is

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$ – total Sum of Squares.

By fitting a regression model we reduce the variation. But still there is a variation due to the error term and it is given by

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$ – error Sum of Squares.

Variation reduced by the regression model is

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$ – regression Sum of Squares.

From the graph:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Note:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Proof: HW

ie $\quad SSTO = SSE + SSR$

* degrees of freedom:

df of $\quad SSTO = \sum (y_i - \bar{y})^2 = n-1.$

df of $\quad SSE = \sum (y_i - \underset{b_0 + b_1 x_i}{\hat{y}_i})^2 = n-2.$

df of $\quad SSR = \sum (\hat{y}_i - \bar{y})^2 = 1.$

ie $\quad df(SSTO) = df(SSE) + df(SSR).$

* Mean Squares

A sum square divided by it's degrees of freedom is called the mean square.

Sample variance

* Total mean square $= MSTO = \dfrac{\sum (y_i - \bar{y})^2}{n-1} = \dfrac{SSTO}{n-1}$

* Regression mean square $= MSR = \dfrac{SSR}{1} = \dfrac{\sum (\hat{y}_i - \bar{y})^2}{1}$

* Error mean square $= MSE = \dfrac{SSE}{n-2} = \dfrac{\sum (y_i - \hat{y}_i)^2}{n-2}$.

Note:

mean squares are not additive

ie $\quad MSTO \neq MSR + MSE.$

Analysis of Variance (ANOVA) table

All of the above values are summerized in a table called ANOVA table.

Source of variation	SS	df	MS	E[MS]
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$		

Note:

1. $E[MSE] = \sigma^2$

Proof:

$Y_i - \hat{Y}_i \sim$ distribution

$$MSE = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}.$$

Note: $\frac{SSE}{\sigma^2} \sim \chi^2_{n-2}$ — chi-square distribution with df $n-2$.

$$\Rightarrow E\left[\frac{SSE}{\sigma^2}\right] = n-2 \quad \left[\because E(\chi^2_\nu) = \nu\right]$$

$$\Rightarrow E\left[\frac{SSE}{n-2}\right] = \sigma^2$$

$$\Rightarrow E[MSE] = \sigma^2.$$

2. $E(MSR) = \sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$

Proof:

$$SSR = \sum(\hat{Y}_i - \bar{Y})^2 = \sum\left[\cancel{b_0} + b_1 X_i - (\cancel{b_0} + b_1\bar{X})\right]^2$$

$$= b_1^2 \sum(X_i - \bar{X})^2$$

$$E(SSR) = E\left[b_1^2 \underbrace{\sum (X_i - \bar{X})^2} \right]$$

$$= \sum (X_i - \bar{X})^2 \, E(b_1^2)$$

$$= \sum (X_i - \bar{X})^2 \left[Var(b_1) + (E(b_1))^2 \right]$$

$$= \sum (X_i - \bar{X})^2 \left(\frac{\sigma^2}{\sum (X_i - \bar{X})^2} + \beta_1^2 \right)$$

$$= \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2 \quad \square.$$

$Var(X) = E(X^2) - (E(X))^2$

$\Rightarrow E(X^2) = Var(X) + (E(X))^2$

F – test

Another hypothesis test for β_1.

Steps:

1) Hypotheses

$$H_0 : \underbrace{\beta_1 = 0} \qquad vs \qquad H_1 : \underbrace{\beta_1 \neq 0}$$

no relationship between X and Y

X and Y are related.

2) Test statistic:

$$F = \frac{MSR}{MSE} \sim f_{1, n-2}$$

– F - distribution with df 1 and n-2.

3) Calculating the critical value or the p-value.

critical value: $F(1-\alpha, 1, n-2)$.

p-value (Let f^* be the observed value of F)

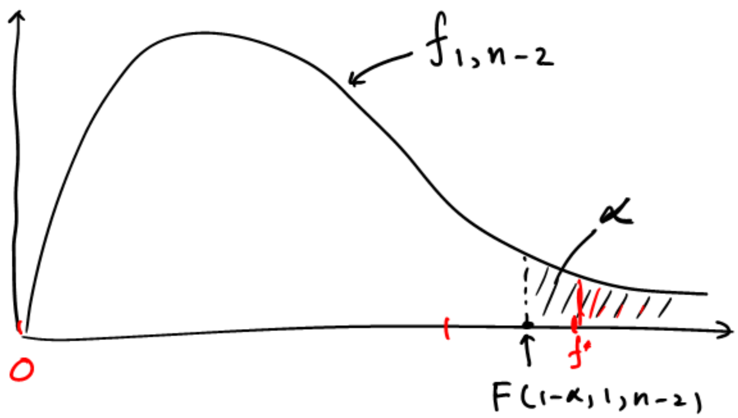
$$p\text{-value} = P(F > f^*)$$

4) <u>conclusion</u>

Reject H_0 if p-value $< \alpha$ (usually $\alpha = 0.05$)

OR

Reject H_0 if $f^* > F(1-\alpha, 1, n-2)$.



<u>Note:</u>

$$\frac{MSR}{MSE} = \frac{\frac{MSR}{\sigma^2}}{\frac{MSE}{\sigma^2}} = \frac{\left(1 \cdot \frac{MSR}{\sigma^2}\right)/1}{\left((n-2)\frac{MSE}{\sigma^2}\right)/(n-2)} = \frac{\chi^2_{(1)}/1}{\chi^2_{(n-2)}/(n-2)} \sim f_{1,n-2}.$$

General Linear Test approach

General Linear test is used to compare two linear models.

<u>Steps:</u>

1. Calculate SSE for the full model (unrestricted model)
2. Calculate SSE for the reduced model (restricted model)

Then,

$$F = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} \sim f_{df_R - df_F,\, df_F}$$

If we compare the SLR model with $Y_i = \mu + \varepsilon_i$, $i = 1, 2, \ldots n$, model:

$$\underbrace{\text{SLR}}_{\substack{\text{Full model} \\ Y = \beta_0 + \beta_1 X}} \qquad \underbrace{Y_i = \mu + \varepsilon_i}_{\text{reduced model}}$$

1) H_0: Restricted model is better (ie $\beta_1 = 0$)

H_1: The full model is better (ie $\beta_1 \neq 0$)

2)
$$F = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

$$= \frac{SSTO - SSE}{(n-1) - (n-2)} \div \frac{SSE}{n-2}$$

$$= \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} \sim F_{1, n-2}$$

* This is same as the test statistic of the F-test.

Note:

If the full model is SLR model, the general linear test and the F-test are both same. But these are different for muliple linear regression models.

Descriptive Measures of linear association between X and Y

1. Coefficient of Determination [R^2]

Coefficient of determination is a measure of reduction of variation as a proportion to the total variation.

$$R^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSR}{SSTO}$$

$$= 1 - \frac{SSE}{SSTO}.$$

Note:

* Since $0 \le SSE \le SSTO$, $0 \le R^2 \le 1$.
* when all the observations fall on the fitted line, $SSE = 0$ and then $R^2 = 1$.

* When the fitted regression line is horizontal, $b_1 = 0$ and then $\hat{Y}_i = \bar{Y}$ for all $i = 1, 2, \dots n$.

Then $SSE = SSTO$

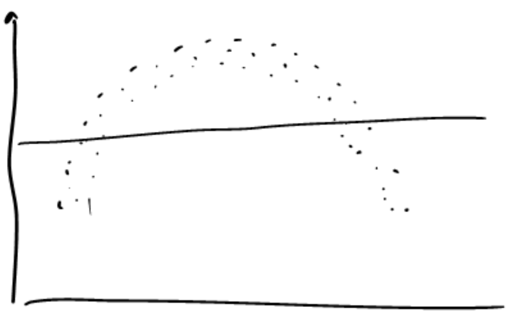
$\Rightarrow R^2 = 0$

ie regression does nothing.

(ie No linear relationship between X and Y).

* There may be another reulationship with $R^2 = 0$.

Eg:-



$R^2 = 0$

2) coeffient of correlution (r)

$$r = \pm \sqrt{R^2}$$

The sign of r is decided based on the sign of b_1.

Chapter-3: Diagnostic and Remidial Measures

Before using the fitted regression model for future predictions we need to check the quality of the fit for data. Here we need to check the assumptions made for the model.

SLR model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

constant.

Fitted model:

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

There are two types of tools to check the quality of a fit. There are graphical tools (ie graphs) and numerical tools (tests).

Residuals play the main rule of accessing the quality of the fit. The key point is the following.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$\downarrow \quad \downarrow \quad \downarrow$$
$$Y_i = b_0 + b_1 X_i + e_i$$

Residuals can be used to estimate the error terms. So residuals should show properties (ie assumptions) of terms ε_i if the fitted model is appropriate.

Properties of residuals:

residuals

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i) \quad , \quad i = 1, 2, \ldots n.$$

mean: $\bar{e} = \dfrac{\sum e_i}{n} = 0.$

Variance:

$$\mathrm{var}\{e\} = \frac{\Sigma(e_i-\bar{e})^2}{n-1} = \frac{\Sigma e_i^2}{n-2} = \frac{SSE}{n-2} = MSE.$$

* Nonindependence

Since all e_is are involved the fitted value $\hat{Y}_i$ which based on the same regression function $(b_0+b_1X_i)$, e_is are not independent.

But for large sample, dependency among e_is can be ignored.

* Semistudentized residuals.

$$e_i^* = \frac{e_i-\bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

* Things to check about the quality of the SLR fitted model

1. Nonlinearity
2. Non constancy of the error variance.
3. Presence of outliers
4. Non-independence of error variance
5. Non-normality of error variance
6. Omission of important predictor variables.