

- **Normal error model:** $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where ϵ_i are iid $N(0, \sigma^2)$

- **Estimated regression (fitted) regression line:** $\hat{Y} = b_0 + b_1 X$

where $b_1 = \frac{SS_{xy}}{SS_x}$ and $b_0 = \bar{Y} - b_1 \bar{X}$

1. Distributions of b_1 and b_0 under normal error model:

$$b_1 \sim N(\beta_1, \sigma^2/SS_x) \quad \text{and} \quad b_0 \sim N(\beta_0, \sigma^2[\frac{1}{n} + \frac{\bar{X}^2}{SS_x}])$$

2. Variance estimators $s^2\{b_1\}$ and $s^2\{b_0\}$:

$$s^2\{b_1\} = \frac{MSE}{SS_x} \quad \text{and} \quad s^2\{b_0\} = MSE[\frac{1}{n} + \frac{\bar{X}^2}{SS_x}]$$

3. Distributions of $\frac{b_1 - \beta_1}{s\{b_1\}}$ and $\frac{b_0 - \beta_0}{s\{b_0\}}$: both have a $t(n - 2)$ distribution

4. Confidence interval for β_1 : $b_1 \pm t \cdot s\{b_1\}$

Confidence interval for β_0 : $b_0 \pm t \cdot s\{b_0\}$

5. Hypothesis test for β_1 (testing for a linear regression relationship):

$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_a: \beta_1 \neq 0$$

Decision rule: Reject H_0 if $t^* = \frac{b_1}{s\{b_1\}}$ is too large or too small.

(or Decision rule: Reject H_0 if the p -value \leq pre-specified level of significance α .)

* To find the power of the test for a specified noncentrality measure δ , use Table in Appendix B.

6. Confidence interval for the mean response $E(Y_h)$:

$$\hat{Y}_h \pm t_{n-2} s\{\hat{Y}_h\} \quad \text{where } s\{\hat{Y}_h\} = \sqrt{\text{MSE} \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\text{SS}_X} \right)}$$

7. Prediction interval for a new observation $Y_{h(\text{new})}$:

$$\hat{Y}_h \pm t_{n-2} s\{Y_{h(\text{new})}\} \quad \text{where } s\{Y_{h(\text{new})}\} = \sqrt{\text{MSE} \left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\text{SS}_X} \right)}$$

8. Prediction interval for the mean of m new observations $\bar{Y}_{h(\text{new})}$:

$$\hat{Y}_h \pm t_{n-2} s\{\bar{Y}_{h(\text{new})}\} \quad \text{where } s\{\bar{Y}_{h(\text{new})}\} = \sqrt{\text{MSE} \left(\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\text{SS}_X} \right)}$$

9. Partitioning the Total Sum of Squares SSTO:

$$\begin{aligned} \text{SSTO} &= \text{SSE} + \text{SSR} \\ \sum (Y_i - \bar{Y})^2 &= \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 \end{aligned}$$

SSTO is a measure of the total variation in the response variable Y.

SSE is a measure of the unexplained random variation in Y.

SSR is a measure of the variation in Y that can be explained by variation in X.

Degrees of freedom: SSTO has $n - 1$ df, SSE has $n - 2$ df, and SSR has 1 df.

10. Cochran's Theorem: If all n observations come from the same normal distribution

with mean μ and variance σ^2 , and SSTO is decomposed into k sums of squares SS_r , each with degrees of freedom df_r such that $\sum_{r=1}^k \text{df}_r = n - 1$, then each $\frac{\text{SS}_r}{\sigma^2}$ is an independent $\chi^2(\text{df}_r)$ random variable.

11. F test for β_1 $H_0: \beta_1=0$ vs. $H_a: \beta_1 \neq 0$

$$\text{Test statistic } F^* = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n - 2) \text{ under } H_0.$$

Reject H_0 if F^* exceeds the appropriate percentile of the F distribution.

12. Confidence Band for Regression Line $E(Y) = \beta_0 + \beta_1 X$

For level X_h , Working-Hotelling $1 - \alpha$ confidence band has boundary values given by

$$\hat{Y}_h \pm \sqrt{2F(1 - \alpha; 2, n - 2)} s\{\hat{Y}_h\}$$

$$\text{where } s\{\hat{Y}_h\} = \sqrt{\text{MSE} \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{SSX} \right)}$$

13. ANOVA Table (augmented with expected mean squares)

Source of Variation	SS	df	MS	E{MS}	F*	p-value
Regression	SSR	1	MSR=SSR/1	$\sigma^2 + \beta_1^2 \cdot SSX$	MSR/MSE	p-value
Error	SSE	$n - 2$	MSE	σ^2		
Total	SSTO	$n - 1$				

The p -value is used for testing $H_0: \beta_1 = 0$.

14. The general linear test approach using the reduced model (under H_0) and full model:

$$\text{test statistic } F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

15. The coefficient of determination: $r^2 = \frac{SSR}{SSTO}$,

which measures the proportion of the variability in Y that can be explained by its linear relationship to X .

Limitations of r^2

- i. high r^2 does not imply that useful predictions can be made
- ii. high r^2 does not imply the estimated regression line is a good fit for the data
- ii. a low r^2 does not imply that X and Y are not related

16. The correlation coefficient: $r = \frac{SSXY}{\sqrt{SS_x SS_y}} = \pm \sqrt{r^2}$

measures the linear association between Y and X , and $-1 \leq r \leq 1$.

Note that (from 5) $t^* = \frac{b_1}{s\{b_1\}} = \frac{SSXY/SSX}{\sqrt{MSE/SSX}} = r \sqrt{\frac{SSY}{MSE}} = r \sqrt{\frac{(n-2)SSTO}{(SSTO-SSR)}} = r \sqrt{\frac{n-2}{1-r^2}}$.

17. Considerations in applying regression analysis

1. Be careful when making inferences for the future.
2. Be careful in making inferences when X levels fall out of the observed range.
3. Don't automatically assume cause-and-effect relationship.

18. Case when X is random

Previous results hold provided that

- (i) $Y_i \mid X_i$ independent $N(\beta_0 + \beta_1 X_i, \sigma^2)$
- (ii) The X_i are independent and with probability distribution which does not involve the parameters β_0 , β_1 , and σ^2 .