Live Classroom

Module 3 – Correlation and Simple Linear Regression

Topics

- Announcements
- Scatterplots
- Correlation
- Simple Linear Regression
- Quantities from the F-distribution

Announcements

- Next week's live classroom will be at the usual time!
  - Wednesday July 27 from 7-9 PM EST.
  - We will discuss the final project.

- The week 5 live classroom will be a day later!
  - Thursday, August 4 from 6-8 PM EST.
  - We will allow you to submit quiz 5 and assignment 5 a day later (i.e., on Wednesday instead of Tuesday).
  - I will update the dates on the course website and post an announcement.

Introduction

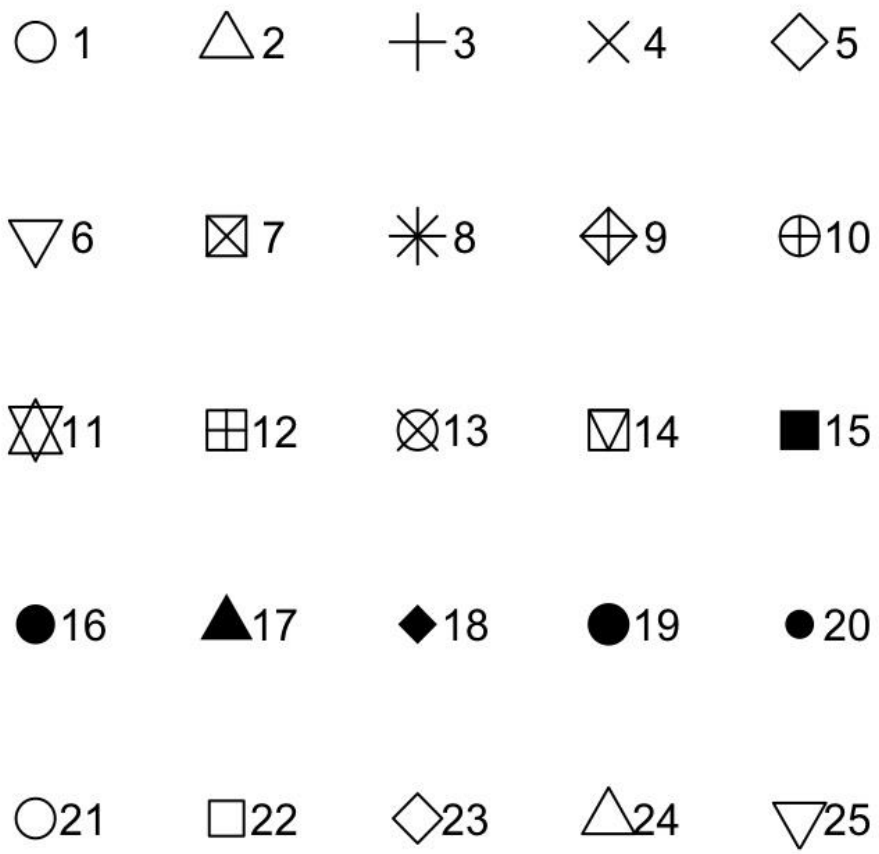
- This module is focused on understanding the relationship or association between two continuous or quantitative factors.
  - Weight versus height.
  - Family income and child's SAT score.
  - Minutes spent studying and test score.
- We will learn to
  - Visualize the relationship (via scatterplots).
  - Describe the relationship via numerical summaries (via correlation and simple linear regression).
  - Use information about a sample to make inferences about a population.

Scatterplots

- Use plot() function to visualize the relationship between two quantitative variables
  - plot(*data$explanatoryvariable*, *data$responsevariable*)
  - Use main, xlab, and ylab to label the plot appropriately
  - Use xlim and ylim to control x and y axes, respectively
  - Change the type of point using pch and/or the color of the points using col
  - Change the size of the points or the labels using cex, cex.axis, or cex.main
- Use scatterplot to describe the form, direction, and strength of the relationship. Can also be used to identify outliers.

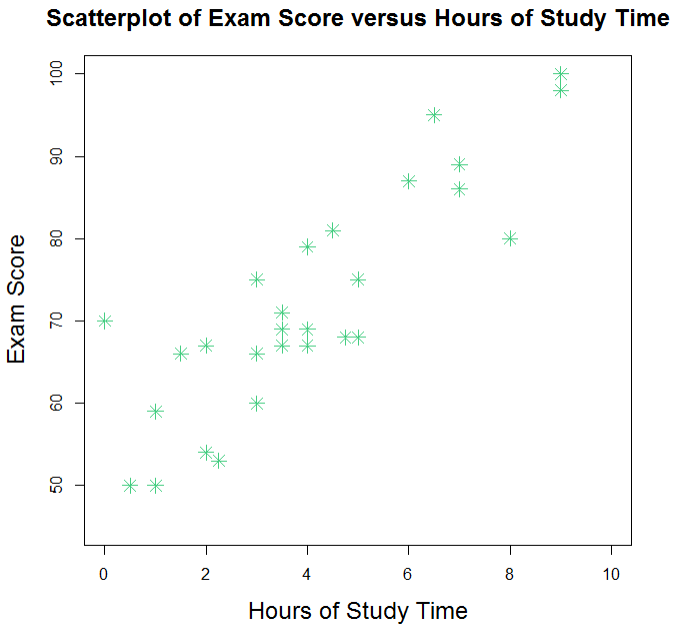
Plot options for pch

- pch=# in plot function call



Example- Study Time and Exam Score

- plot(examdata$Hours,examdata$Exam, main="Scatterplot of Exam Score versus Hours of Study Time", xlab = "Hours of Study Time", ylab="Exam Score", xlim=c(0, 10), ylim=c(45, 100), pch = 8, col="seagreen3", cex=1.5, cex.lab = 1.5, cex.main = 1.5)



Correlation

- Formula:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- Properties:
  - Between -1 and +1
  - Correlation between x and y = Correlation between y and x
  - Used for quantitative variables only
  - Measures strength of linear association
  - Independent of unit of measure
  - Outliers affect correlation.

Correlation - Inference

- Though the sample correlation coefficient is useful in helping describe the relationship in a sample, we are often interested in using the data collected from a sample to make conclusions about the correlation between the same parameters in the population.

- The sample correlation, *r*, is a point estimate for the population correlation coefficient, ρ (read "rho").

- In general, formal tests of hypotheses concerning ρ seek to determine whether or not there is a linear association between the variables in the population (we usually test whether $\rho = 0$).

Correlation

- Use cor() function to calculate sample correlation coefficient
  - cor(*data$explanatoryvariable*, *data$responsevariable*)
- Use cor.test() function to perform testing
  - cor.test(*data$explanatoryvariable*, *data$responsevariable*,
    alternative =[*alternative*],
    method = [*method*],
    conf.level = [*confidence level*])
  - *[alternative] = "less", "greater" or "**two.sided**"*
  - [*method*] = "**pearson**", "kendall" or "spearman"
  - FYI (optional material) : https://www.statisticssolutions.com/correlation-pearson-kendall-spearman/

Example- Study Time and Exam Score

- Calculate sample correlation
  - cor(examdata$Hours,examdata$Exam)
- Perform testing
  - cor.test(examdata$Hours,examdata$Exam)

Example- Study Time and Exam Score

- Is there a linear relationship between hours of study time and exam score for all students taking CS546?

(1) Set up the hypotheses and select the alpha level

$H_0: \rho = 0$ (there is no linear association between study time and exam score)

$H_1: \rho \neq 0$ (there is a linear association between these factors)

$\alpha = 0.05$

(2) Select the appropriate test-statistic

$t = r\sqrt{\frac{n-2}{1-r^2}}$ with df = n-2

Example- Study Time and Exam Score

(3) State the decision rule

Decision Rule: Reject H_0 if $p \leq \alpha$. Otherwise, do not reject H_0

Example- Study Time and Exam Score

(4) Compute the test statistic

cor.test(examdata$Hours,examdata$Exam) gives

$$t = 10.1574, \quad df = 29, \quad p - value = 4.625e - 11$$

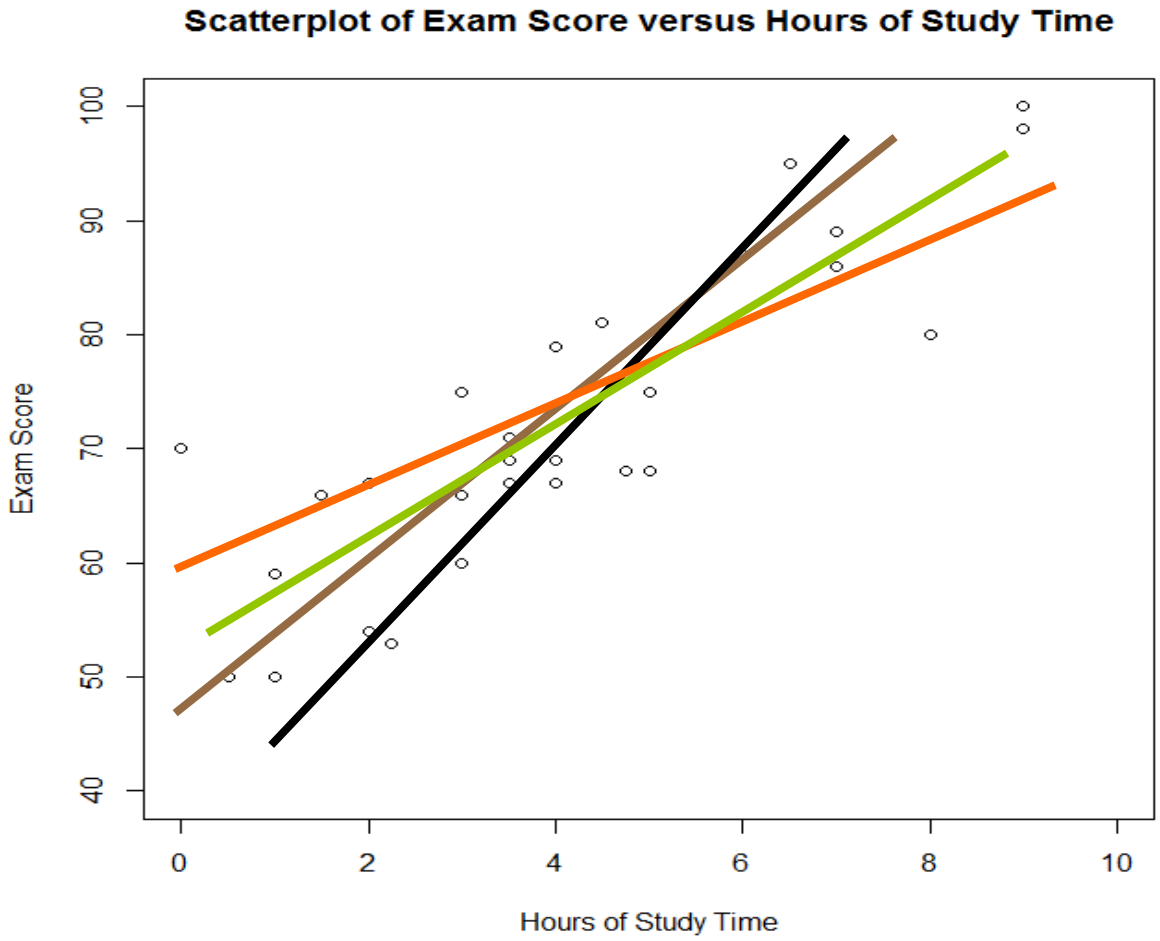
(5) Conclusion

Reject H_0 since $p \leq \alpha$. We have significant evidence at the $\alpha = 0.05$ level that $\rho \neq 0$. That is, there is evidence of a significant linear association between study time and exam score among students in CS546. The sample correlation coefficient is 0.8835 indicating a strong positive association between study time and exam score. The positive correlation between these factors indicates that as study time increases, exam scores increase.

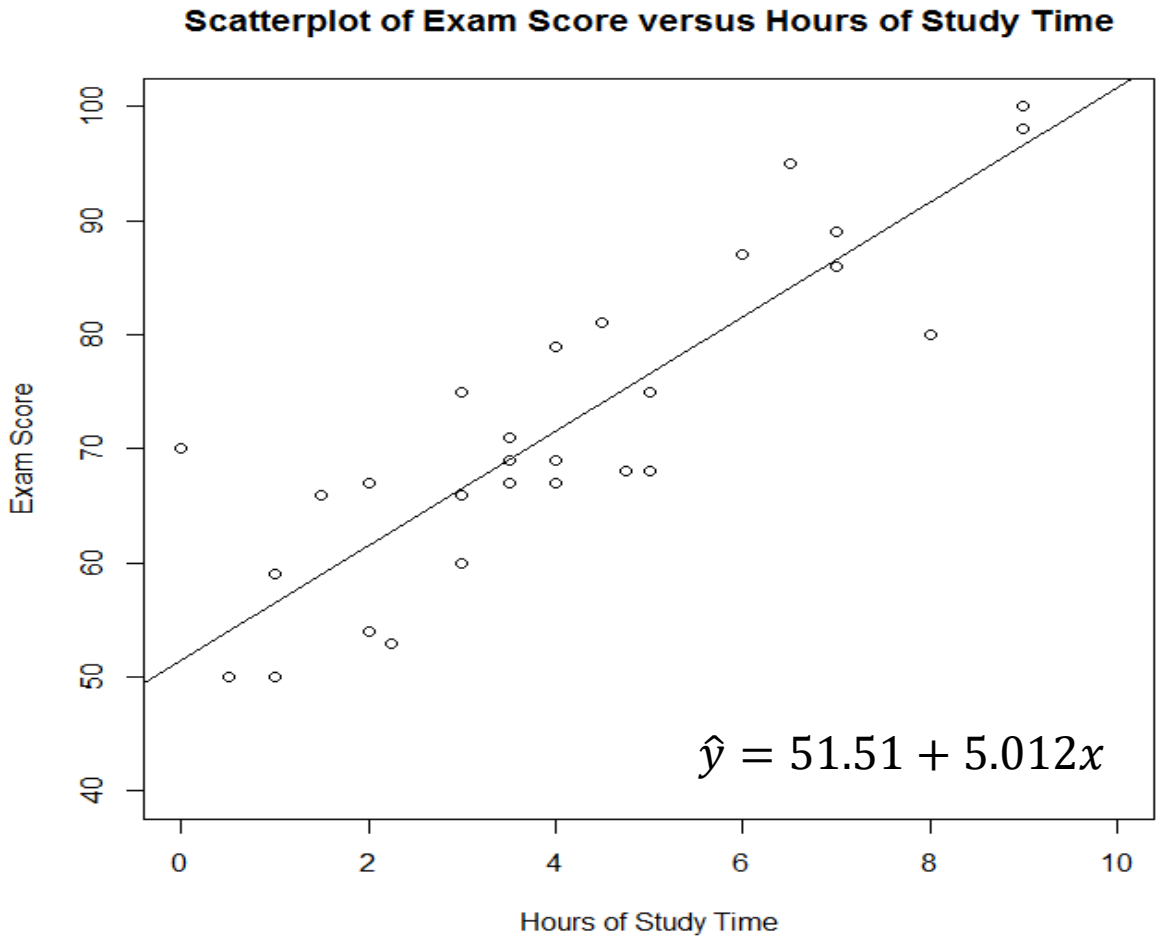
Simple Linear Regression

- Describes the nature of the relationship between two continuous variables
- Assert straight line on the scatterplot that best fits the pattern of the relationship
- Form: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  - $\hat{y}$ (read "y hat") is the expected (average) or predicted value of y for a given value of x
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least-squares estimates of β_0 (the intercept) and β_1 (slope), respectively
    - $\hat{\beta}_1$ = expected (average) or predicted change in response (y) for a one unit increase in explanatory variable (x)
    - $\hat{\beta}_0$ = expected (average) or predicted value of the response when the explanatory variable (x) equals 0

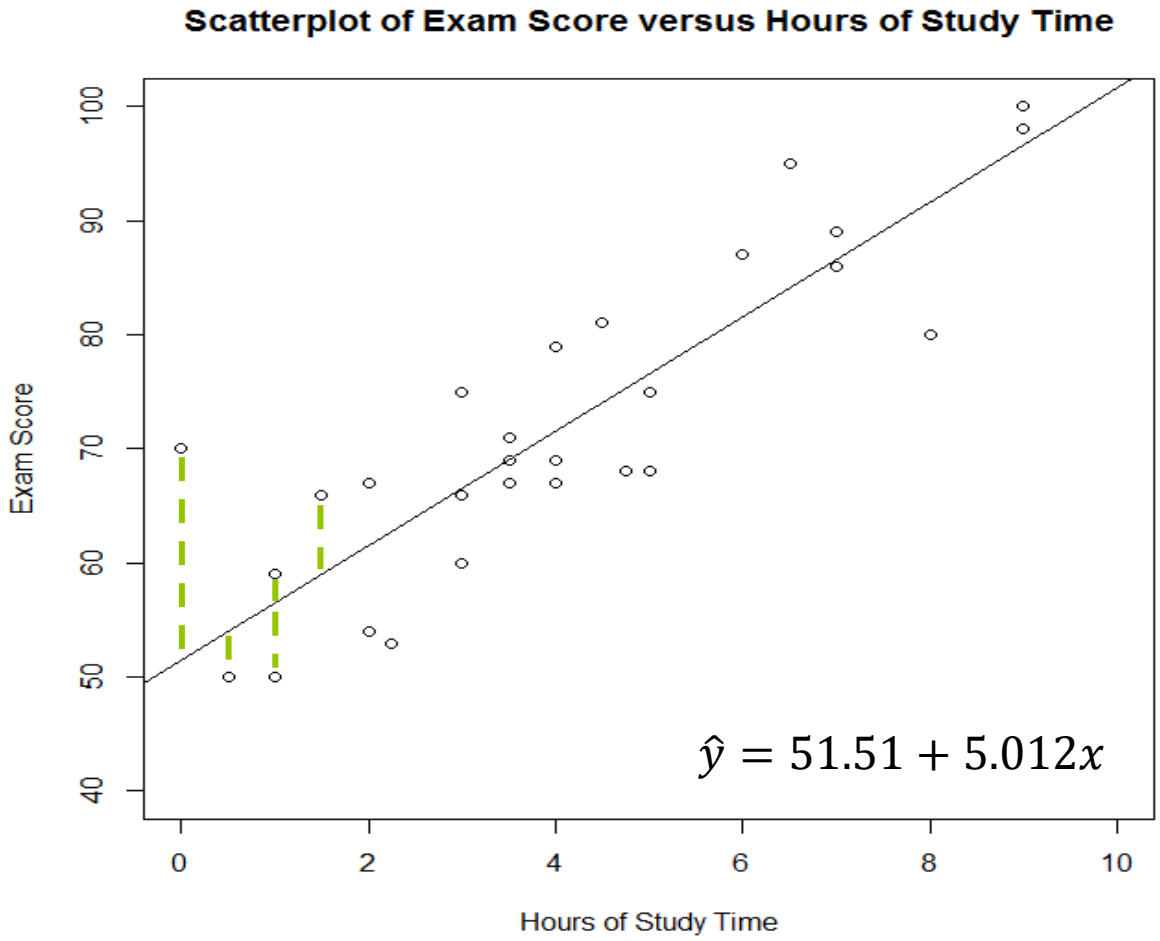
Example – What's the best fit line?



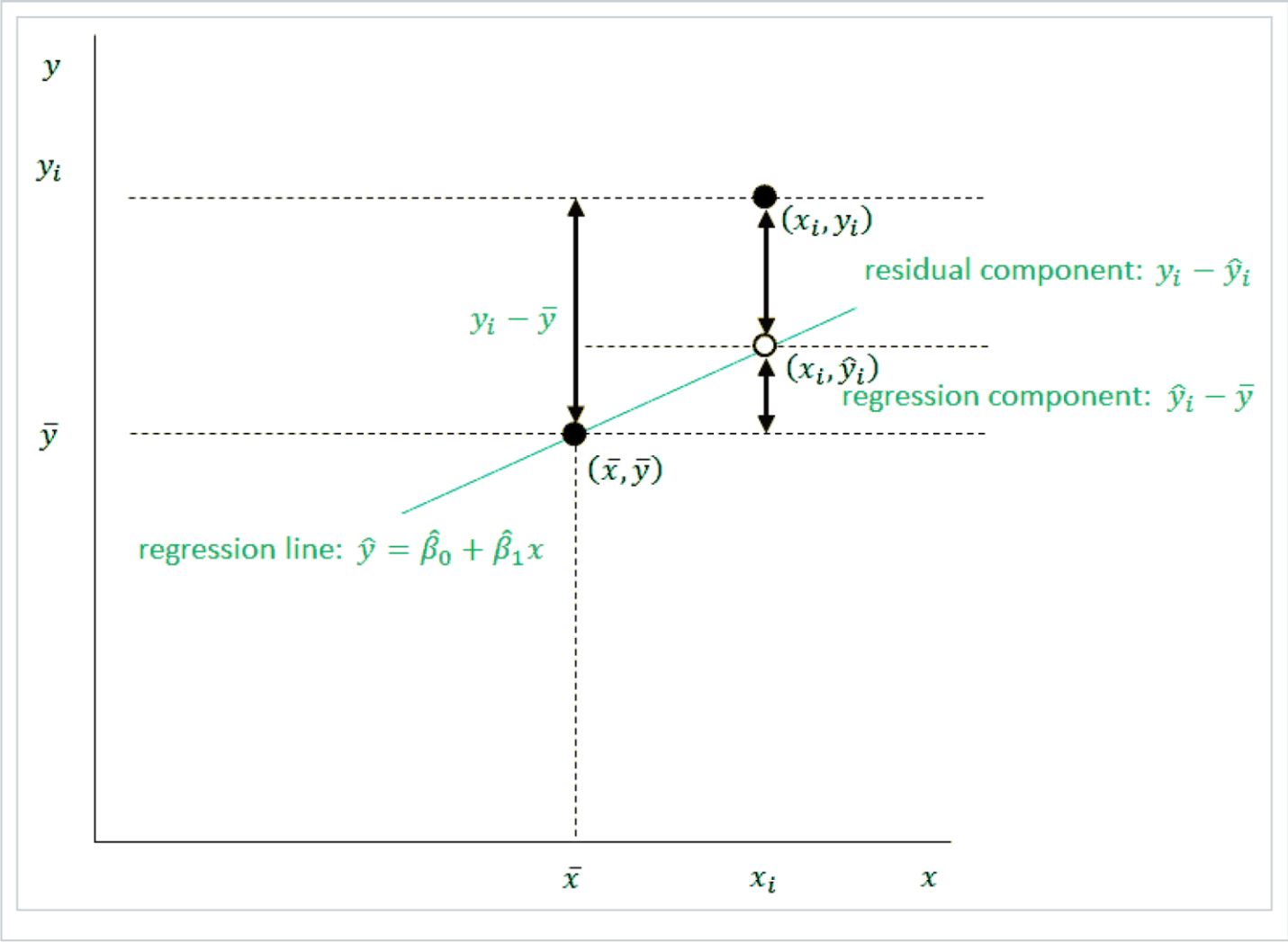
Example – What's the best fit line?



Example – What's the best fit line?



Simple Linear Regression



ANOVA table

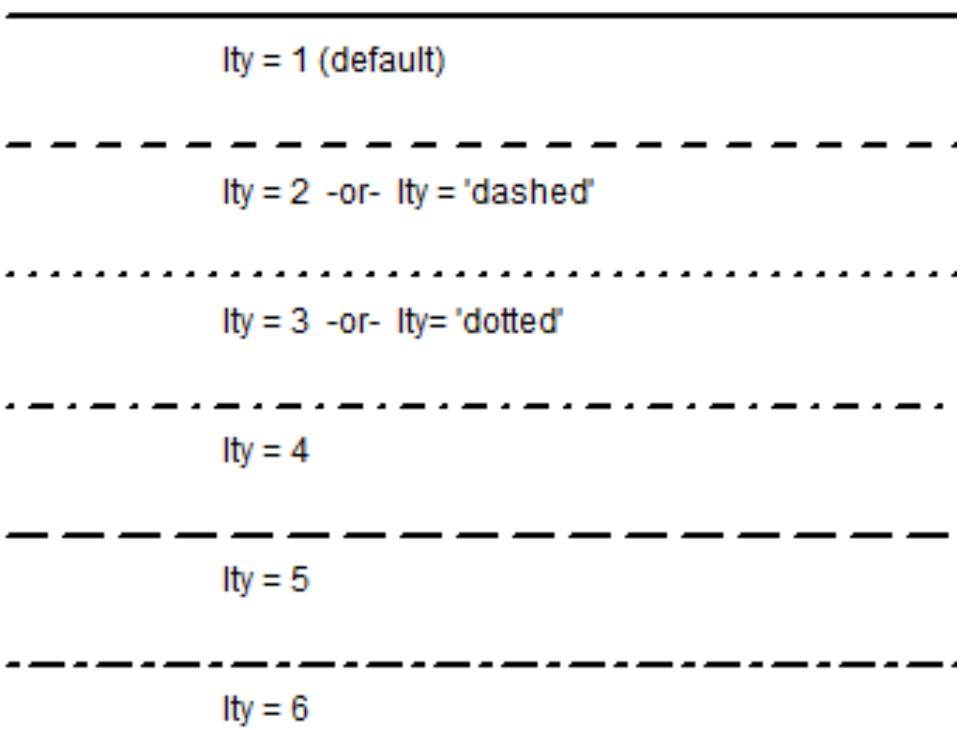
	SS (Sum of Squares)	df (degrees of freedom)	MS (Mean Square)	F-statistic	p-value
Regression	Reg SS	Reg df = k	Reg MS = Reg SS/Reg df	F=Reg MS/Res MS	$P(F_{\text{Reg df,Res df, }\alpha}>F)$
Residual	Res SS	Res df = n-k-1	Res MS = Res SS/Res df		
Total	Total SS = Reg SS + Res SS				

Simple Linear Regression

- Use lm() function to perform regression
  - lm(*data$response~data$explanatory*)
- Use abline() function on the lm() result to plot the regression line on the scatterplot
- Use anova() function on the lm() result to obtain the ANOVA table (F-test)
- Use summary() function on the lm() result to obtain the beta coefficients (t-test), R-squared value, F-statistic
- Use confint() on the lm() result to obtain confidence intervals for the regression coefficients

abline() function

- Control the line type by lty=#
- Control color using col = "*color*"



Quantities from the F-distribution

- Calculating probabilities from F-statistics
  - Use the pf() function to calculate the area to the left of a given *F*-statistic
  - pf(*[F statistic]*, df1 = *[degrees of freedom of the numerator]*, df2 = *[degrees of freedom of the denominator]*)
- Calculating F-statistics from probabilities
  - Use the qf() function to find the F-statistic with the specified area to the left
  - qf(*[area]*, df1 = *[degrees of freedom of the numerator]*, df2 = *[degrees of freedom of the denominator]*)

Table C. F-Distribution Critical Values

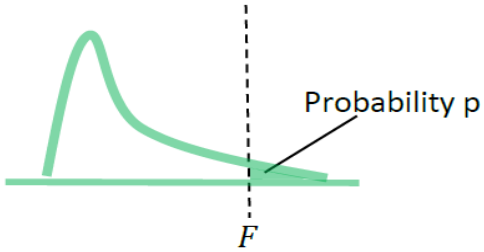


Table entry for p is the critical value F with probability p lying to its right

		Degrees of freedom in the numerator									
	p	1	2	3	4	5	6	7	8	9	10
	0.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
	0.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
1	0.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63
	0.010	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85
	0.001	405284	499999	540379	562500	576405	585937	592873	598144	602284	605621
	0.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
	0.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
2	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
	0.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
	0.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39	999.40

- pf(18.51, df1=1,df2=2)
  - R gives 0.95 (area to the left of F=18.51)
- qf(0.95, df1=1,df2=2)
  - R gives back 18.51

Example- Study Time and Exam Score

- m<-lm(examdata$Exam~examdata$Hours)
- plot(examdata$Hours,examdata$Exam)
- abline(m)
- anova(m)
- summary(m)
- confint(m, level = 0.95)

Example- Study Time and Exam Score – F-test

- Is there a linear relationship between hours of study time and exam score for all students taking CS546?

(1) Set up the hypotheses and select the alpha level

H_0: $\beta_1 = 0$ (there is no linear association between study time and exam score)

H_1: $\beta_1 \neq 0$ (there is a linear association between these factors)

$\alpha = 0.05$

(2) Select the appropriate test-statistic

$F = \frac{MS\ Reg}{MS\ Res}$ with 1 and n-2 degrees of freedom

Example- Study Time and Exam Score – F-test

(3) State the decision rule

Decision Rule: Reject H_0 if $p \leq \alpha$. Otherwise, do not reject H_0

Example- Study Time and Exam Score – F-test

(4) Compute the test statistic

anova(m) or summary(m) gives

$$\mathrm{F-statistic:}\ 103.2 \text{ on } 1 \text{ and } 29 \text{ DF}, \qquad \mathrm{p-value:}\ 4.625\mathrm{e}-11$$

(5) Conclusion

Reject H_0 since $p \leq \alpha$. We have significant evidence at the $\alpha = 0.05$ level that $\beta_1 \neq 0$. That is, there is evidence of a significant linear association between study time and exam score among students in CS546. The beta coefficient for the regression is 5.01 indicating an increase of approximately 5 percentage points on the exam for each additional hour of study time. The 95% confidence interval for the beta coefficient is approximately 4 to 6 indicating that we are 95% confident that the underlying increase in exam score (in percentage points) for each hour spent studying is between 4 and 6.

Example- Study Time and Exam Score – t-test

- Is there a linear relationship between hours of study time and exam score for all students taking CS546?

(1) Set up the hypotheses and select the alpha level

H_0: $\beta_1 = 0$ (there is no linear association between study time and exam score)

H_1: $\beta_1 \neq 0$ (there is a linear association between these factors)

$\alpha = 0.05$

(2) Select the appropriate test-statistic

$$t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2/n-2} \Big/ \sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}}$$ with df = n-2

Example- Study Time and Exam Score – t-test

(3) State the decision rule

Decision Rule: Reject H_0 if $p \leq \alpha$. Otherwise, do not reject H_0

Example- Study Time and Exam Score – t-test

(4) Compute the test statistic

summary(m)

$$t = 10.16, \qquad df = 29, \qquad p - value = 4.625e - 11$$

(5) Conclusion

Reject H_0 since $p \leq \alpha$. We have significant evidence at the $\alpha = 0.05$ level that $\beta_1 \neq 0$. That is, there is evidence of a significant linear association between study time and exam score among students in CS546. The beta coefficient for the regression is 5.01 indicating an increase of approximately 5 percentage points on the exam for each additional hour of study time. The 95% confidence interval for the beta coefficient is approximately 4 to 6 indicating that we are 95% confident that the underlying increase in exam score (in percentage points) for each hour spent studying is between 4 and 6.

Things to note

- Correlation between x and y is independent of order
- Regression and correlation will give same conclusion, but regression coefficient depends on which variable is specified as explanatory
- In regression, t-test and F-test give same result (in SLR $F = t^2$) – same p-value
- T-test from correlation and t-test from regression are equivalent and also give same result