Linear Regression - Data Analysis and Visualization

Dr. Farshid Alizadeh-Shabdiz

Spring 2021

Simple Linear Regression (SLR)

The equation for the simple linear regression line is given by

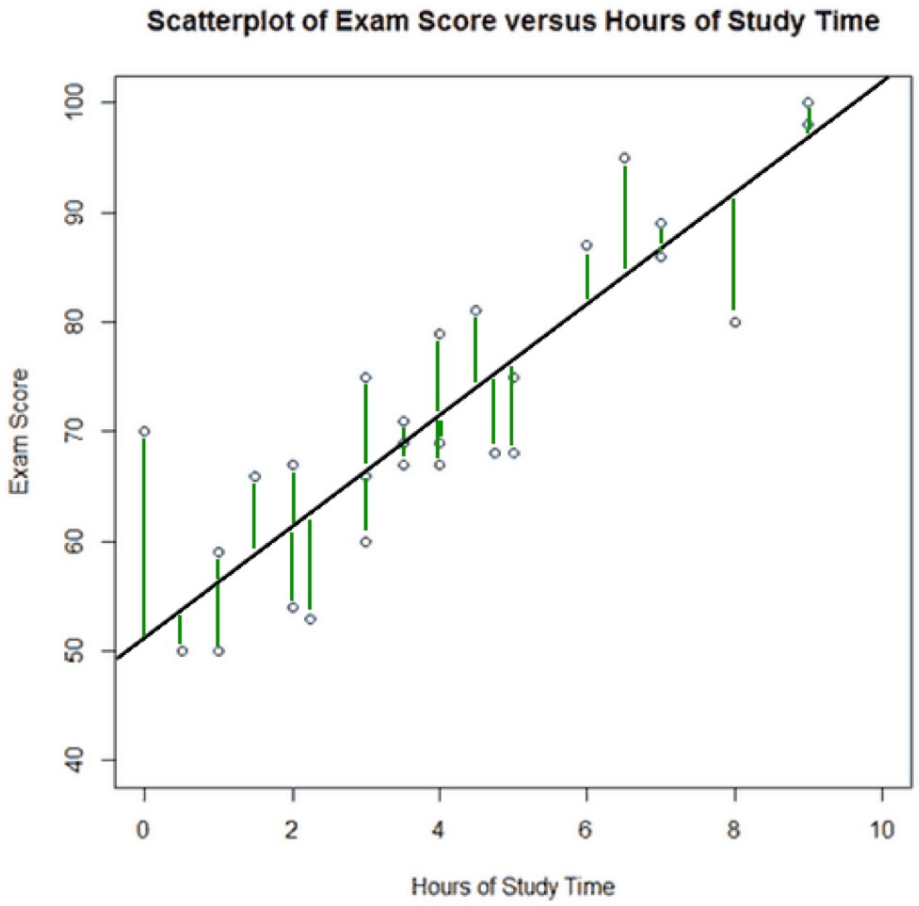
$$y = \beta_0 + \beta_1 x$$

- y is the response or dependent variable
- x is the explanatory or independent variable
- **β_0 is the intercept (the value of y when $x = 0$)**
- **β_1 is the slope (the expected change in y for each one-unit change in x)**

How to find the regression line that best fits the data

There are several ways to find parameters of the line. The most common way is to **minimize the sum of the squares** of the distances between the points and the regression line.

This approach is called the **least-squares method**.



Equation for the least-squares regression line

SLR equation is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = r\frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

r is correlation coefficient

S_x the sample standard deviation of x, S_y is SD of y

$\bar{x}$ sample mean of x, and $\bar{y}$ sample mean of y

An Example Study Hours vs. Exam Scores

```
> student <- read.csv("student.csv")
> attach(student)

> xbar <- mean(study.hours)
> sx <- sd(study.hours)
> ybar <- mean(score)
> sy <- sd(score)
> r <- cor(study.hours, score)

> beta1 <- r*sy/sx
> beta1

> beta0 <- ybar - beta1*xbar
> beta0


#Use lm() function lm(data$responsevariable ~ data$explanatory)
> m <- lm(score ~ study.hours)
> summary(m)
```

$$\hat{y} = 51.51 + 5.012x$$

Linear Regression – Two parameters example

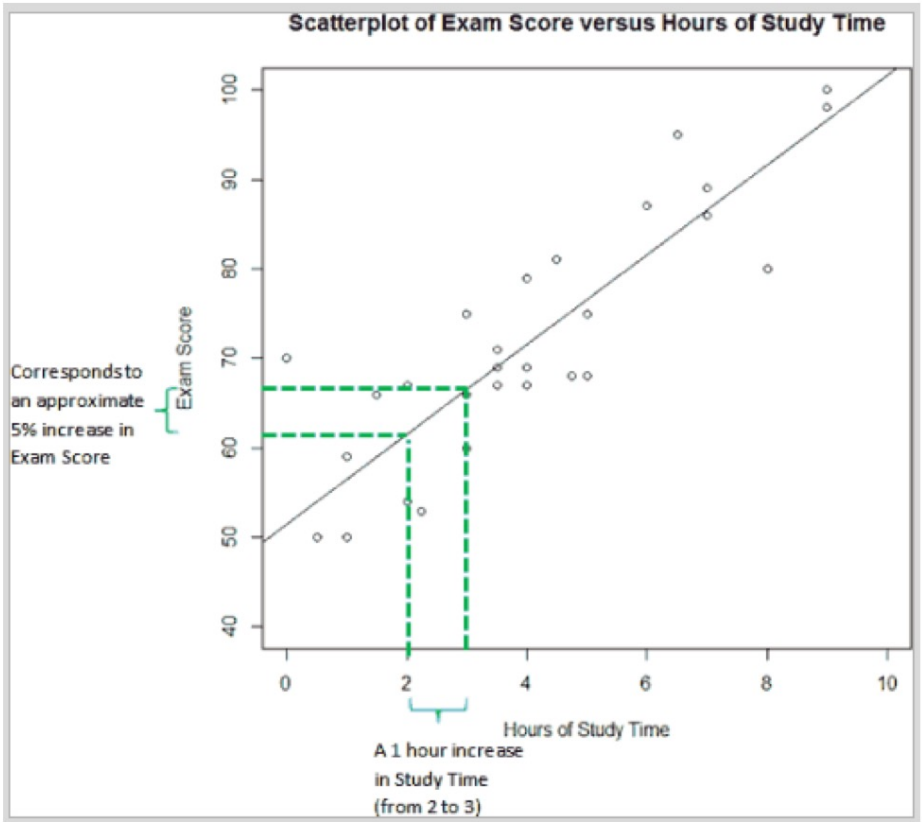
- Parameters of linear regression model will be added by "+" sign
  - m = lm(score ~ hours+student$id)
- All the parameters can be used by using "." sign. In that case, data has to be identified
  - m = lm(score ~ . , data=student)
- Notes
  - Shows how to run multi-parameter linear regression
  - The analysis shows importance of "id" in the analysis
  - Overfitting in high dimensionality space

Interpretation of results

Interpretation of results

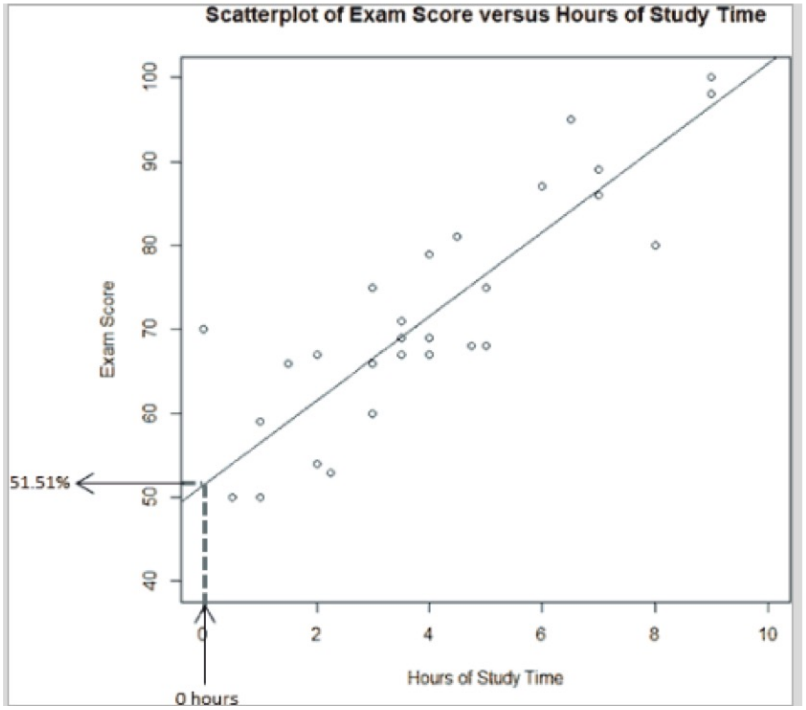
The estimate of the slope parameter ($\hat{\beta}_1$) gives the **expected or predicted change** in the response variable $\hat{y}$ **for a one-unit increase in the explanatory variable (x)**. Here, $\hat{\beta}_1 = 5.012$.

- Can be interpreted as the **increase in exam score for every one-hour increase in study time**.
- For **each additional hour** that students studied, their exam score **improved by around 5%** points on average.



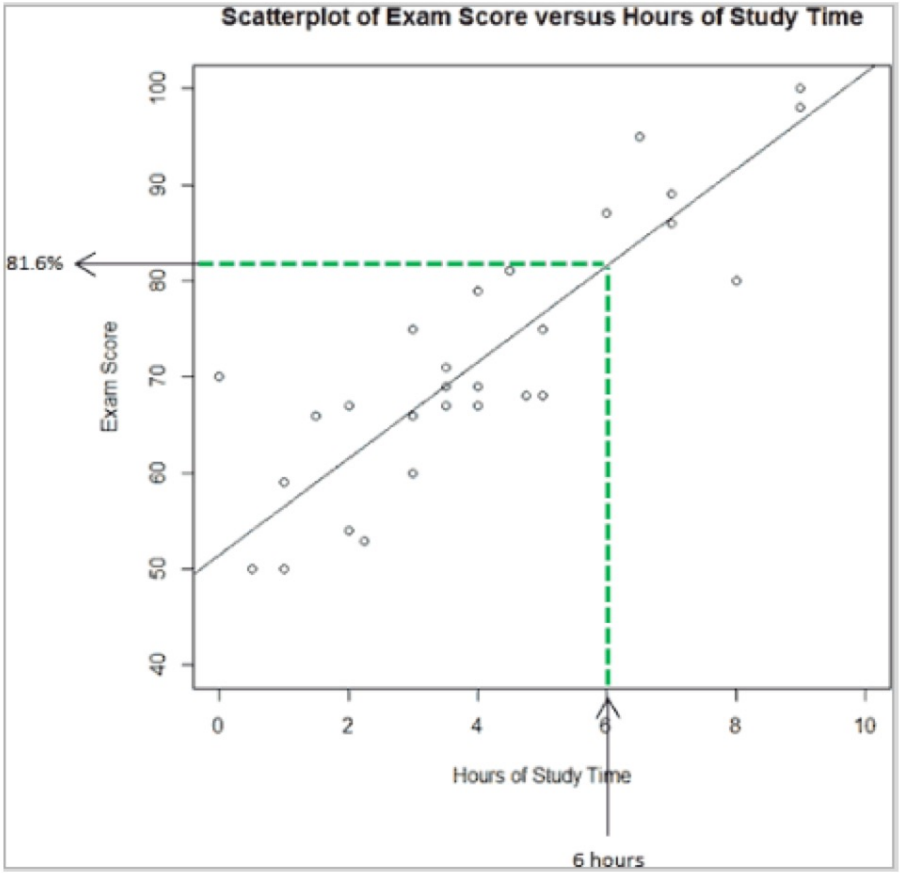
Interpretation of results

- The **Linear Nature** of the relationship and the equation implies that the increase in exam score is the same **for any 1 unit change**. **Increase from 3 to 2 hours is the same as increase from 9 hours versus 8 hours.**
- The estimate for the intercept (β_0) is meaningful in this case since values of the explanatory variable near 0 are possible. Here, $\beta_0 = 51.515$. **This can be interpreted as the average exam grade for those who did not study.**



Interpretation of Results - Example

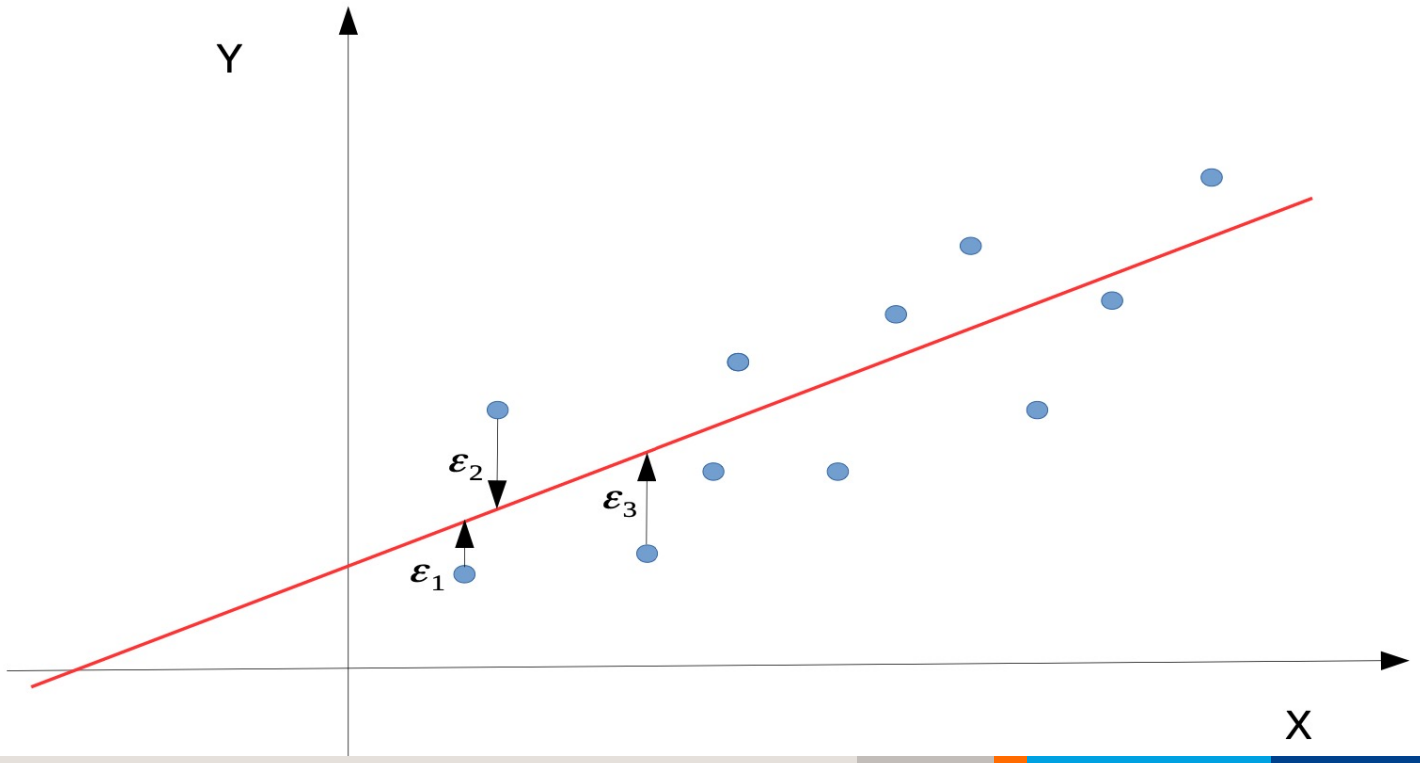
- The least-squares regression is $y = 51.51 + 5.012x$
- **If I were planning to study for 6 hours, what should I expect my score to be?**



- If I study for 6 hours by plugging in $x = 6$ into the regression equation $\hat{y} = 51.51 + 5.01x$
- That is, my average expected exam score is $y = 51.51 + 5.012(6) \approx 81.6$ if I study for 6 hours.

Random Error ϵ

- The **true value of the response variable will vary** from the value predicted by the regression.
- We assume that **the random error term, ϵ, is normally distributed** with a mean of 0 and a variance of σ^2.
- The **larger the random error, the more the individual data points are scatter around** the linear regression line.
- We need to assess the **goodness of fit of the regression line.**

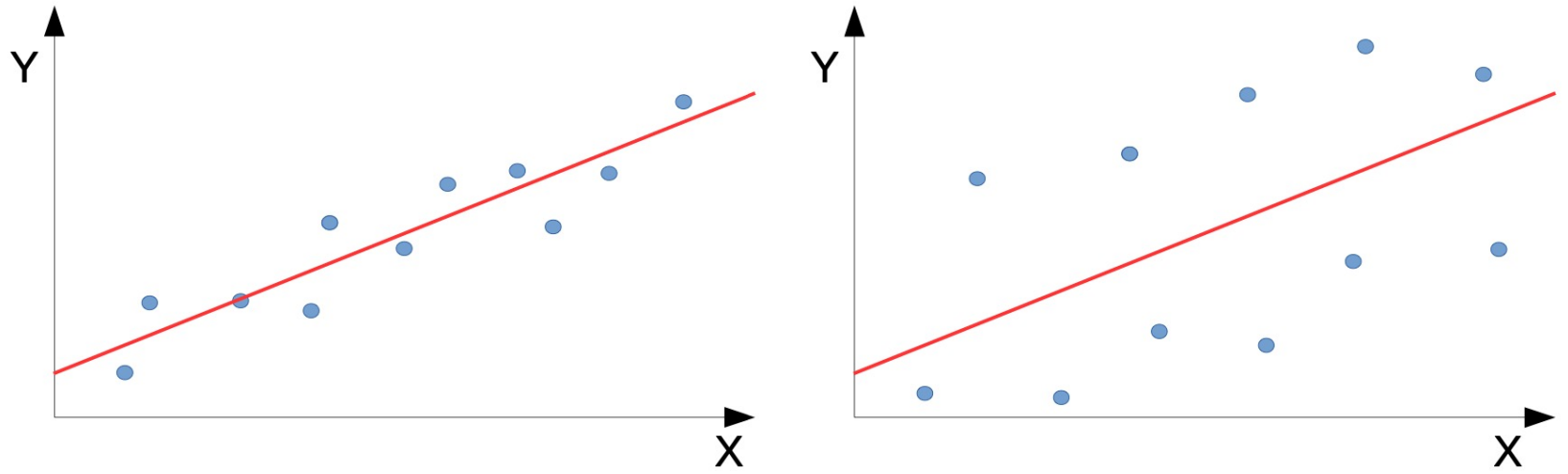


Assessing the Fit of Linear Regression Model

Question – How do you suggest assessing a linear regression model?

The coefficient of determination or R-squared

- The coefficient of determination, or R^2, is a number that indicates how well data fit a statistical model - in our case, the regression line.
- R^2 represents the proportion (percentage) of the variation in the response variable **explained by the regression model** (equation).

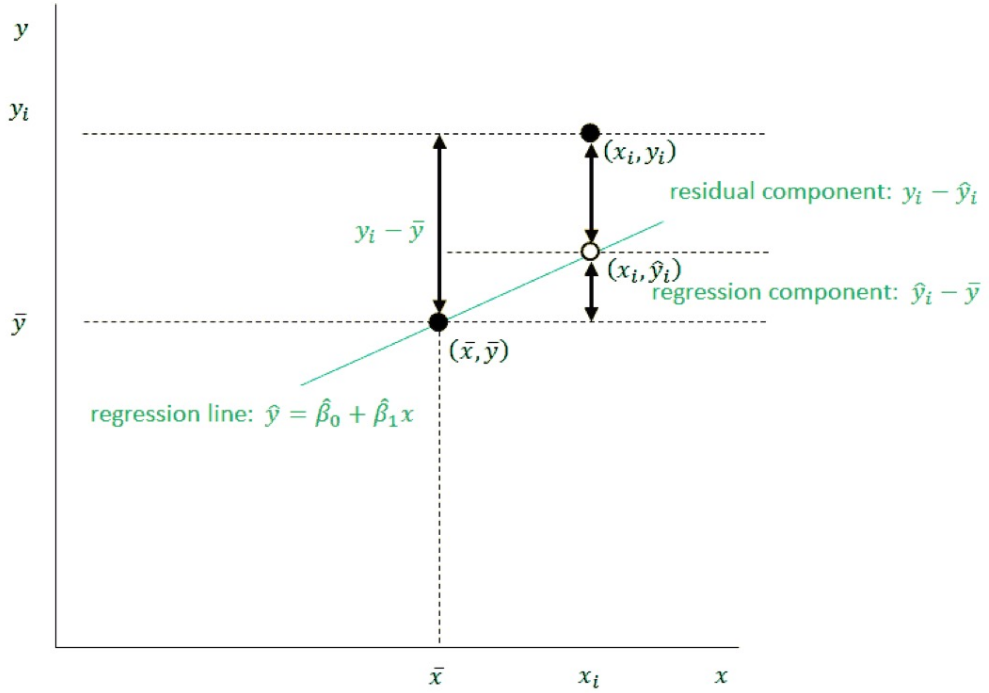


Assessing the Fit of the Regression Line

For any given data point, the difference between the mean response and the observed response value y i can be split into two parts:

(1) the residual component

(2) the regression component

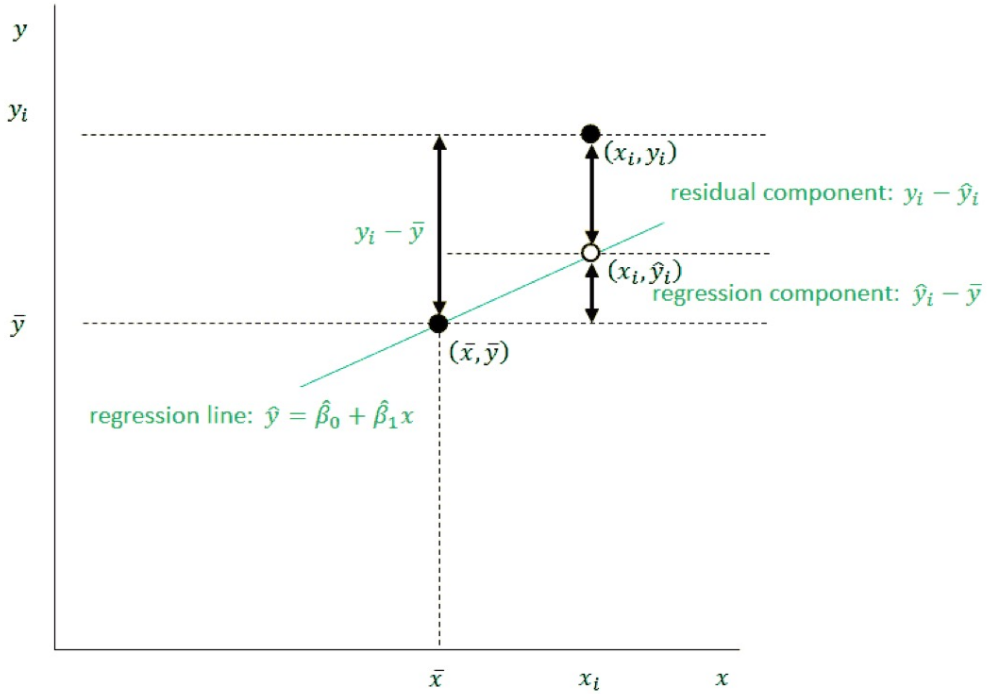


Assessing the Fit of the Regression Line

For any given data point, the difference between the mean response and the observed response value y i can be split into two parts:

(1) the residual component

(2) the regression component



For any sample point (x_i, y_i)

Residual component is the vertical distance between the **observed response, y_i**, and the regression **predicted response for x_i.**

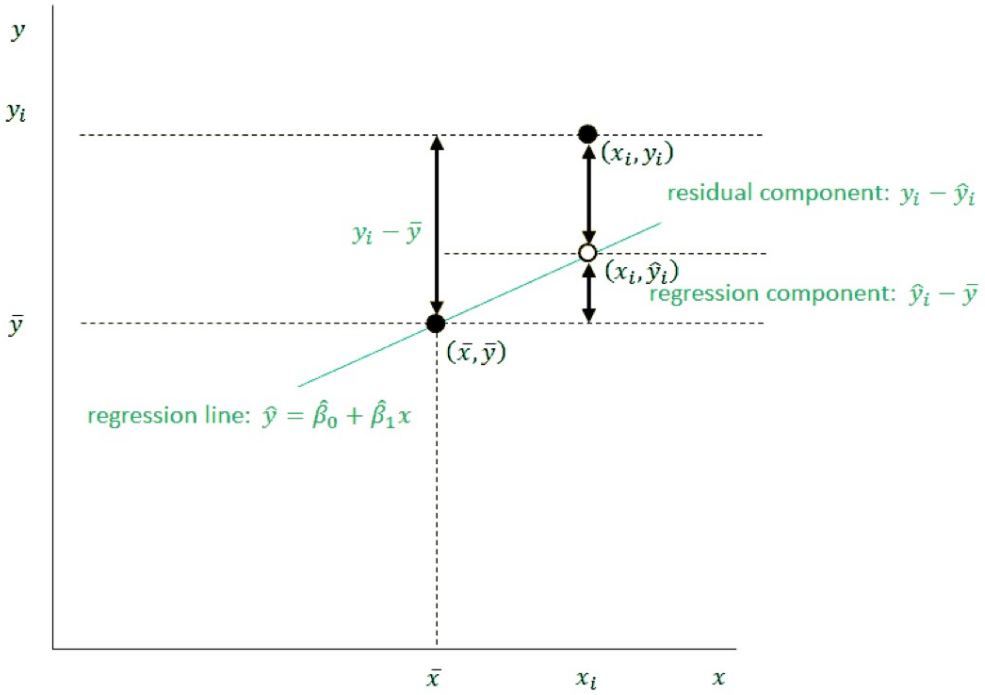
Regression Component is the **vertical distance** between the regression **predicted response** for the value of explanatory variable x_i and the **average value of the response variable**.

Assessing the Fit of the Regression Line

For any given data point, the difference between the mean response and the observed response value y i can be split into two parts:

(1) the residual component

(2) the regression component



For any sample point (x_i, y_i)

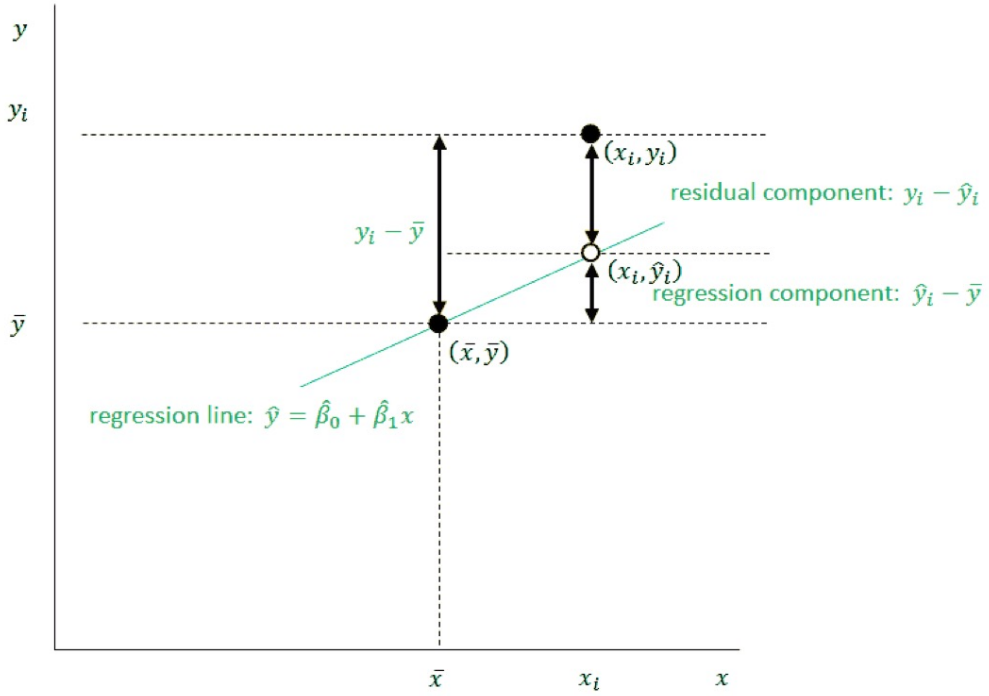
Residual component is $y_i - \hat{y} = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

Regression Component is $\hat{y}_i - \bar{y} = (\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}$

Assessing the Fit of the Regression Line

The **sum of the regression component and the residual component** gives us back the difference between the mean response $\bar{y}$ and the observed response value y_i

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



- If **all data points fell on or very close to the regression line**, then $y_i \approx \hat{y}_i$ and the residual component $y_i - \hat{y}_i$ will be 0 or very close to 0.
- **A well-fitted regression line will have regression components that are larger than the residual components across all data points.**

The coefficient of determination or R-squared

Sum of all squared deviations and break it into each of the component parts to **see what proportion represents the regression components versus the residual components.**

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

where

- $\sum_{i=1}^{n}(y_i - \bar{y})^2$ **(Total Sum of Squares)** represents the sum of squares of the deviations of the individual sample points from the sample mean
- $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ **(Residual Sum of Squares)** represents the sum of squares of the residual components
- $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ **(Regression Sum of Squares)** represents the sum of squares of the regression components

The Coefficient of Determination or R-squared

One of the measures that we use to **assess the fit of the data is the coefficient of variation** (R^2, read "R-squared")

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{Reg SS}}{\text{Total SS}} = r^2$$

- **Coefficient of determination R2** is the quantity that represents the proportion of the variation explained by the regression (or the "model").
- R^2 ranges between 0 and 1
- $R^2 = 1$ mean that model explains everything
- For SLR $R^2 = r^2$ Correlation Coefficient.

An Example: Linear Regression and R^2

- For age of five couples, predict age of a husband based on age of the wife.
- Calculate R^2 value

Couple	Age of Wife	Age of Husband
1	20	20
2	30	32
3	24	22
4	28	26
5	28	30
Sample mean	26	26
Sample standard deviation	4.0	5.1

An example: Calculate R2

The association between husbands and wives ages was calculated to be $\hat{y} = -4.94 + 1.19x$. Using this and the fact that $\bar{y} = 26$, calculate by hand the quantities from the ANOVA table.

Calculate R-squared and give its interpretation.

Couple	Age of Wife	Age of Husband
1	20	20
2	30	32
3	24	22
4	28	26
5	28	30
Sample mean	26	26
Sample standard deviation	4.0	5.1

An example: Calculate R2

The association between husbands and wives ages was calculated to be $\hat{y} = -4.94 + 1.19x$. Using this and the fact that $\bar{y} = 26$, calculate by hand the quantities from the ANOVA table.

Calculate R-squared and give its interpretation.

Reg $df = 1$ for SLR, Res $df = n - k - 1 = n - 2 = 5 - 2 = 3$.

Couple	Age of Wife	Age of Husband
1	20	20
2	30	32
3	24	22
4	28	26
5	28	30
Sample mean	26	26
Sample standard deviation	4.0	5.1

Couple	x_i	y_i	$\hat{y}_i$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	20	20	18.86	-7.14	50.98	1.14	1.30
2	30	32	30.76	4.76	22.66	1.24	1.54
3	24	22	23.62	-2.38	5.66	-1.62	2.62
4	28	26	28.38	2.38	5.66	-2.38	5.66
5	28	30	28.38	2.38	5.66	1.62	2.62
Sum					90.63		13.75

An example: Calculate R2

	SS (Sum of Squares)	df (degrees of freedom)	MS (Mean Square)
Regression	Reg SS $= 90.63$	Reg df $= k = 1$	Reg MS $= 90.63/1 = 90.63$
Residual	Res SS $= 13.75$	Res df $= n - k - 1 = 5 - 1 - 1 = 3$	Res MS $= 13.75/3 = 4.58$
Total	Total SS = Reg SS + Res SS $= 104.38$		

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{Reg SS}}{\text{Total SS}} = \frac{90.63}{104.38} = 86.8\%$$

86.8% of the variability in husbands ages can be explained by wives' ages.

Inference about Regression Coefficients

R^2 Issues

- Adding any new parameter always results to increase in R^2, even if the new parameter is not significant
- As number of parameters increase, R^2 approaches 1
- Overfitting would also make R^2 approach one (-> 1)
- Can not be used to assess multiple parameters

What is Degree of Freedom

- Degree of freedom (df) is number of values that can vary freely.
  - If n number of points are randomly selected from a normal distribution – the point can be any where in n-space dimension. Therefore, there are df=n
  - If we write the samples as mean and residual.
    - The mean can be any where: df=1
    - Residual will have (df=n-1) since mean is fix and by selecting (n-1) the last one will get dictated by the mean

Alternate solution – Using Hypothesis Testing and Confidence Interval

T-test & F-test

Inference about Regression Coefficients

- **$\hat{\beta}_0$ and $\hat{\beta}_1$ are statistics** calculated using sample data, not population parameters.
- If we **had a different sample**, we would get **different values of $\hat{\beta}_0$ and $\hat{\beta}_1$**
- Formal inference involves considering β_0, β_1 as unknown population parameters and determining what we can say about the unknown population parameters given the data we observed from our sample.

One Sample T-Test

- Going from population with mean μ to a sample with mean of $\bar{X}$

$$y = \beta_0 + \beta_1 x$$

  - H0: $\beta_1 = 0$
  - H1: $\beta_1 \neq 0$
- $t = \frac{y_i - \hat{y}}{s\sqrt{n}} \sim t_{n\_2}$ with S being standard deviation of the sample set

Five Step for Hypothesis Testing

- **Set up the hypotheses and select the alpha level**
- **Select the appropriate test statistic**
- **State the decision rule**
- **Compute the test statistic**
- **Conclusion**

Inference from regression - t-test

In linear regression, the sampling distribution of the coefficient estimates form a normal distribution, which is approximated by a ***t* distribution due to approximating σ by *s***.

We can calculate a confidence interval for each estimated coefficient or **perform a hypothesis test**:

$H_0 : \beta_1 = 0$ (there is no linear association)

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

Test Statistic **t-test**

$$t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$$

with $df = n - 2 = 31 - 2 = 29$ degrees of freedom

Normality Assumption - Standard Error of Regression Coefficients

How is the standard error of the estimator $\hat{\beta}$ calculated?

$$s_{\hat{\beta}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

No worries! We use R to compute this.

T–Distribution from Wikipedia

Definition [edit]

Probability density function [edit]

Student's ***t*-distribution** has the probability density function given by

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where ν is the number of *degrees of freedom* and Γ is the gamma function. This may also be written as

$$f(t) = \frac{1}{\sqrt{\nu}\,\mathrm{B}(\frac{1}{2},\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where B is the Beta function. In particular for integer valued degrees of freedom ν we have:

For $\nu > 1$ even,

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} = \frac{(\nu-1)(\nu-3)\cdots 5\cdot 3}{2\sqrt{\nu}(\nu-2)(\nu-4)\cdots 4\cdot 2}.$$

For $\nu > 1$ odd,

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} = \frac{(\nu-1)(\nu-3)\cdots 4\cdot 2}{\pi\sqrt{\nu}(\nu-2)(\nu-4)\cdots 5\cdot 3}.$$

T-test for SLR - Confidence Interval

The decision rule for a two-sided level α t-test is:

Reject $H_0 : \beta_1 = 0$ if $|t| \geq t_{n-2,\frac{\alpha}{2}}$ OR $p \leq \alpha$

Otherwise, do not reject $H_0 : \beta_1 = 0$

where $t_{n-2,\frac{\alpha}{2}}$ is the value from the t-distribution table with $n-2$ degrees of freedom and associated with a **right hand tail probability of $\alpha/2$**.

The two-sided $100\% \times (1-\alpha)$ confidence interval for β_1 is given by:

$$\hat{\beta}_1 \pm t_{n-2,\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_1}$$

Interpretation: We can say with 95% confidence (for standard $\alpha = 0.05$) that the true value of β_1 is between

$$\hat{\beta}_1 - t_{n-2,\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_1} \text{ and } \hat{\beta}_1 + t_{n-2,\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_1}$$

A Example: t-test for Simple Linear Regression

`Is there a linear relationship between hours of study time and exam score?`

Perform a t-test at the $\alpha = 0.05$ level, construct and interpret the 95% confidence interval for β_1.

1. Set up the hypotheses and select the alpha level

$H_0 : \beta_1 = 0$ (there is no linear association) $H_1 : \beta_1 \neq 0$ (there is a linear association) $\alpha = 0.05$

2. Select the appropriate test statistic

$$t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}} \tag{1}$$

with $df = n - 2 = 31 - 2 = 29$ degrees of freedom

3. State the decision rule Determine the appropriate value from the t-distribution table with 29 degrees of freedom and associated with a right hand tail probability of $\alpha/2 = 0.025$

A Example: t-test for Simple Linear Regression (Continued)

3. State the decision rule Determine t-value with $df = 29$ and associated with a right hand tail probability of $\alpha/2 = 0.025$

$t_{n2,\alpha/22} = t_{29,0.025} = 2.045$

Decision Rule: Reject H_0 if $t \geq 2.045$ or if $t \leq -2.045$

Otherwise, do not reject H_0

4.Compute the test statistic. Using R we can have

	Estimate	SE	t-statistic	p-value
Intercept	51.5147	2.3820	21.63	2e-16
Hours	5.0121	0.4934	10.16	4.63e-11

$$t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}} = \frac{5.0121}{0.4934} \approx 10.16$$

A Example: t-test for Simple Linear Regression (Continued)

Also we can calculate the confidence interval

$$\hat{\beta}_1 \pm t_{n-2,\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_1} = 5.0121 \pm 2.045 \cdot 0.4934$$
$$= (4.00, 6.02)$$

5. Conclusion

- Reject H_0 since $10.16 \geq 2.045$.
- 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 (here, $p < 0.001$ as calculated using software program).
- We are 95% confident that the true value of is between 4.00 and 6.02.

Example – Age vs Cholesterol

Age	Cholesterol	Age	Cholesterol
25	180	42	183
25	195	48	204
28	186	51	221
32	180	51	243
32	210	58	208
32	197	62	228
38	239	65	269

- Find relationship between age and cholestrol?
- And also R^2

Example cont.

- Line: Y=151.3537 + 1.3991x
- $R^2 = 0.515$

- Test for linear relationship?
- Find confidence interval of slope

- Standard error = sqrt[(4553.7)/14-2] = 19.48

$$standard\ error = \sqrt{\frac{4553.7}{14-2}} = 19.48$$

$$Sample\ standard\ error = \frac{\beta_1 - \beta_1}{\frac{Standard\ error}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} = \frac{1.3991}{19.48/\sqrt{2472.92}}$$

$$= 3.572\ (the\ p - value = 0.003)$$

Two side t-dis with 95% CI with (14-2)=12 degree of freedom = 2.179
So, we reject H0 and there is a linear regression relationship.

Slope confidence interval

- $Lower\ bound = 1.3991 \pm {}_{2}^{t_\alpha} \frac{Standard\ error}{\sqrt{\sum (x_i - \bar{x})^2}}$
- $= 1.3991 \pm 2.179 \frac{19.48}{\sqrt{2472.9}}$
- $= 1.3991 \pm 0.8536$

Inference about Regression Coefficients

In regression analysis, assessment of the fit of the model to the data is performed using the quantities in the **ANOVA (analysis of variance) 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				

Inference about Regression Coefficients

In regression analysis, assessment of the fit of the model to the data is performed using the quantities in the **ANOVA (analysis of variance) 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				

- **Reg *df* = *k***, the df of Reg SS, k = number of predictors in the model
- **Res *df* = *n* – *k* – 1** is the df of Res SS.
- **Reg MS = Reg SS/Reg df** (the regression mean square)
- **Res MS = Res SS/Res df** (the residual mean square)
- **F= Reg MS/Res MS** (F statistic value)
- **p-value** = the probability that the observed value of test statistic

Formal Tests of Hypotheses

In SLR, formal tests of hypotheses concern β_1.

They are generally of the form

$H_0 : \beta_1 = 0$ (there is no linear association)

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

- $\beta_1 = 0$ would mean that the regression line had a slope of 0 **(a horizontal line)**
- $\beta_1 = 0$ is equivalent to saying that the **response variable does not change at all when the explanatory variable changes**.

F-test for Simple Linear Regression

$\boldsymbol{H_0 : \beta_1 = 0}$ (there is no linear association)

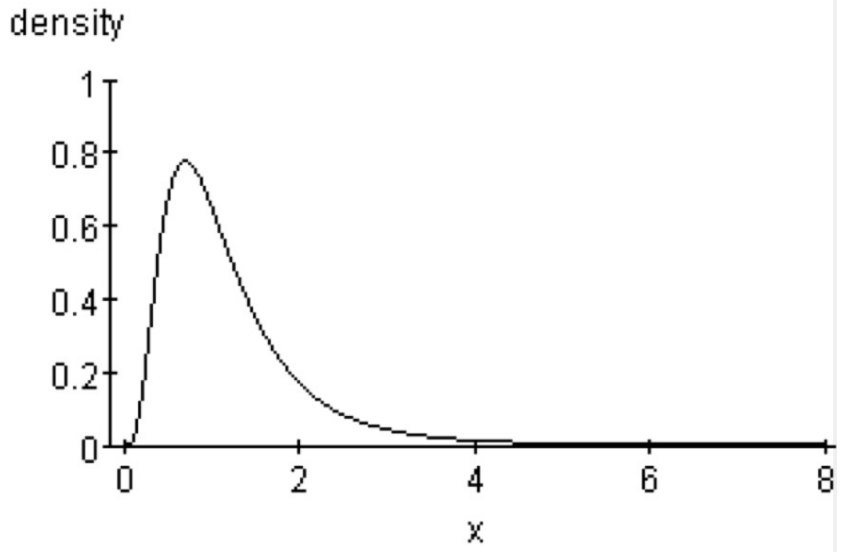
$\boldsymbol{H_1 : \beta_1 \neq 0}$ (there is a linear association)

$$F = \frac{\text{Reg MS}}{\text{Res MS}}$$

F-distribution with 1 and $n - 2$ degrees of freedom under H_0.

F-distribution

- The **F-distribution** is named after the famous statistician R. A. Fisher[1].
- **F is the ratio of two variances.**
- The F-distribution is most commonly used in **Analysis of Variance (ANOVA) and the F test** (to determine if two variances are equal).
- It has a **minimum of 0, but no maximum value** (all values are positive).
- The peak is not far from 0.
- When referencing the F-distribution the numerator degrees of freedom are always given first, and switching the degrees of freedom changes the distribution **(F(10,12) does not equal F(12,10))**.



[1] https://en.wikipedia.org/wiki/Ronald_Fisher

F-test for Simple Linear Regression

$H_0 : \beta_1 = 0$ (there is no linear association)

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

$$F = \frac{\text{Reg MS}}{\text{Res MS}}$$

F-distribution with 1 and $n - 2$ degrees of freedom under H_0.

The decision rule is:

Reject $H_0 : \beta_1 = 0$ if $F \geq F_{1,n-2,\alpha}$

Otherwise, do not reject $H_0 : \beta_1 = 0$

where **$F_{1,n-2,\alpha}$ is the value from the F-distribution** with 1 degree of freedom (numerator) and $n - 2$ degrees of freedom (denominator) and associated with a right-hand tail probability of α.

R Functions - Quantities from the F-distribution

```
# Calculating probability from F-statistics
# Use pf() function to calculate the area to the left of a given F-
    statistic
> pf([F statistic], df1=[degree of freedom of the numerator], df2=[
    degree of freedom of the denominator])

# Calculating F-statistics from probability
# Use qf() function to calculate F-statistic with the specifies area to
    the left
> qf([probability], df1=[degree of freedom of the numerator], df2=[
    degree of freedom of the denominator])
```

Quantities from the F-distribution

```
> pf(18.51, df1=1, df2=2)
[1] 0.9499929  # (the area to the left)
> pf(18.51, df1=1, df2=2, lower.tail = F) # (area to the right)
[1] 0.05000706
> qf(0.95, df1=1, df2=2)  # getting the p-value out of F-value
[1] 18.51282
```

Table C. F-Distribution Critical Values

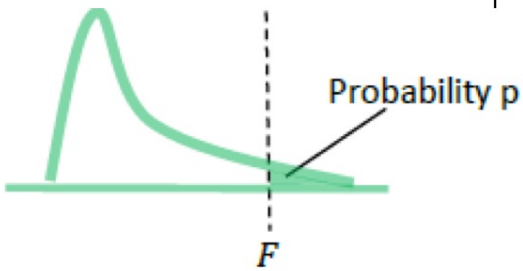


Table entry for p is the critical value F with probability p lying to its right

		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
	0.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23

A Example: F-test for Simple Linear Regression

Is there a linear relationship between hours of study time and exam score? Perform this test at the $\alpha = 0.05$ level.

1. Set up the hypotheses and select the alpha level

$H_0 : \beta_1 = 0$ (there is no linear association)

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

$\alpha = 0.05$

2. Select the appropriate test statistic

$$F = \frac{\text{MS Reg}}{\text{MS Res}}$$

with 1 and n-2 = 31-2 = 29 degrees of freedom

A Example: F-test for Simple Linear Regression (Continued)

3. State the decision rule

F-distribution with 1, 29 degrees of freedom and associated with $\alpha = 0.05$.

```
> qf(.95, df1=1, df2=29)
```

$F_{1,29,0.05} = 4.1830$

Decision Rule: Reject H_0 if $F \geq 4.1830$

Otherwise, do not reject H_0

A Example: F-test for Simple Linear Regression (Continued)

4. Compute the test statistic

Using R function anova(), we got the following ANOVA table:

	SS	df	MS	F-statistic	p-value
Regression	4973.5	1	4973.5	103.2	4.625e-11
Residual	1398.0	29	48.2		
Total					

$$F = \frac{\text{MS Reg}}{\text{MS Res}} = \frac{4973.5}{48.2} \approx 103.2 \text{ with } 1 \text{ and } 29 \text{ degrees of freedom.}$$

F-statistic can also be calculated using summary()

5. Conclusion

- Reject H_0 since $103.2 \geq 4.183$.
- We have significant evidence at the $\alpha = 0.05$ level that $\beta_1 \neq 0$.
- There is evidence of a significant linear association between study time and exam score (here, $p < 0.001$ as calculated using software program).

General F function

$$F = \frac{Reg\ SS / Reg\ df}{Res\ SS / Original\ model\ \ df}$$

Or

$$F = \frac{(Total\ SS\ -\ Res\ SS) / Reg\ df}{Res\ SS / Original\ model\ \ df}$$

Or comparing model 2 with base model 1

$$F = \frac{(Res\ SS1\ -\ Res\ SS2) / (df2 - df1)}{Res\ SS2 / (n - df2)}$$

- In words: error captured by model divided by remaining error

F – Test
Testing for Significance of Multiple Parameter

- In order to assess a multi variable linear regression, an equation is needed reflecting impact of multi-parameters
  - F test is the ratio of variance captured by regression to residual error
  - Numerator has been standardized by dividing by number of parameter
  - Denominator has been standardized by dividing by (number of samples-number of parameters-1)
  - (-1) in denominator because of one parameter in the alternative equation
- This is called F-distribution

F- test example1

- Predicting stock value as a function of PE ratio, earning per stock, and dividend.
- We have taken **350** companies.
- Total standard deviation: 130
- Standard deviation of residual after model: 30
- Test the significance of this solution with 95% CI

Answer to Example 1

- Hypothesis
  - H0: $\beta_1 = \beta_2 = \beta_3 = 0$
  - H1: at least one of the coefficients is not zero
- Number of parameters: fd1 = 3
- Number of samples: fd2 = 350

$$F_{test} = \frac{(130 - 30)/3}{30/(350 - 4)} = 346$$

- BTW: qF(0.95, df1=3,df2=346) = 2.63

F-Test Example 2

- Based on 50 samples
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, total standard deviation of residual after the model =100
  - $y = \beta_0 + \beta_1 x_1$, total standard deviation of 130

  Answer
  - H0: b2=b3=0
  - H1: b2 or b3 not equal to zero

F-Test Example 2 - Solution

$$y = \frac{(Square\ error\ residual\ H0\ - square\ error\ residual\ H1) \Big/ (\#\ of\ parameters\ H1)}{(Square\ \ error\ residual\ H1) \Big/ (\#sample - total\ parameters\ - 1)}$$

$$F_{test} = \frac{(130 - 100)/2}{100/(50 - 4)} = 6.9$$

R Functions - lm(), anova(), fitted(), resid()

```
# create linear model object
my.model <- lm( student$score ~ student$hours)

# report a summary of the model, reports regression coefficients,
    RSquared, F values, t values, p-values
summary(my.model)

# create Analysis of Variance Table
anova(my.model)

# fitted values of my model
fitted(my.model)

# residuals of my model
resid(my.model)
```

Adjusted R^2

R Squared is the estimate of the variability of the response variable y given a particular value of the explanatory variable x. It is a statistic and depends on two parameters, 1. number of samples, 2. number of variables in the model (in SLR =1).

We want to be more conservative and adjust (reduce) it to state claims that are more true.

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1} \right)$$

- n is the number of points in your data sample.
- k is the number of independent regressors, i.e., the number of variables in the model (in SLR k=1)
- Always $R^2_{adj} \leq R^2$
- In SLR for large n, $R^2_{adj} \approx R^2$

Closing Remarks - Some key points

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