CS555B1 Data Analysis and Visualization

Lecture 6

Simple Linear Regression and Assessing the Fit

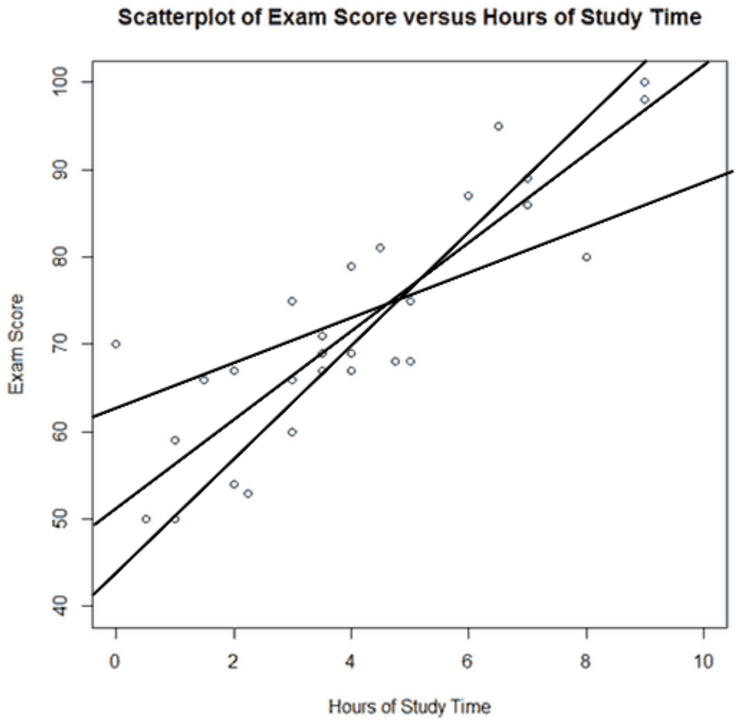
Kia Teymourian

Simple linear regression

The equation for the simple linear regression line is given by

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

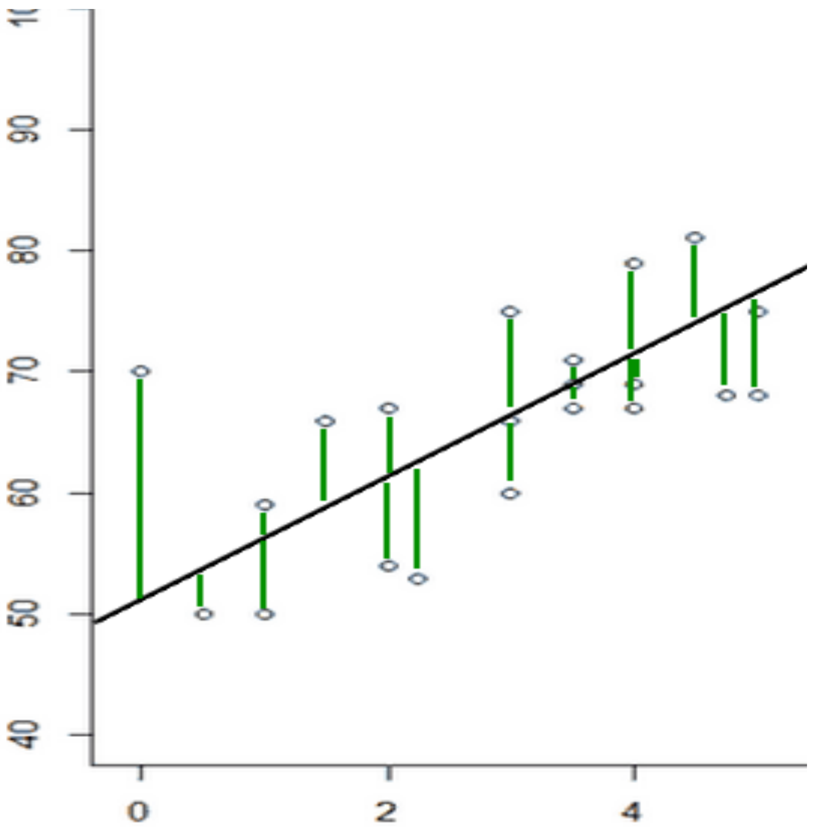
y is the response or **dependent** variable
x is the explanatory or **independent** variable
beta_0 is the intercept (the value of y when x=0)
beta_1 is the slope (the expected change in y for each one-unit change in x)



How to find the regression line that best fits the data

- We want to **minimize** the vertical distance between each of the points and the regression line.
- The most widely used method, **least-squares method**, aims to **minimize** the sum of the squares of the distances between the points and the regression line.
- Using the least-squares method, you can calculate the equation using just the correlation between the variables and each variable's mean and standard deviation.

Simple linear regression fits a straight line through the set of data points in such a way that **makes the sum of squared residuals of the model** (vertical distances between the points of the data set and the fitted line) **as small as possible**.



Equation for the least-squares regression line

The equation for the simple linear regression line is given by

$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}$x

$\hat{y}$ is the expected or predicted value of y for a given value of x
x is the explanatory or independent variable
$\widehat{\beta_0}$ is the least-squares estimates of β_0 (the intercept)
$\widehat{\beta_1}$ is the least-squares estimates of β_1 (the slope)

In the least-squares regression, the estimates of β_0 and β_1 are:
$\widehat{\beta_1}$=r$\frac{S_y}{S_x}$ and $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$

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

r = correlation coefficient, S_x = the sample standard deviation of x, S_y = the sample standard deviation of y, $\bar{x}$ = sample mean of x, $\bar{y}$ = sample mean of y

The equation for $\widehat{\beta_0}$ ensures that the least-squares regression line always passes through the "center of mass" point ($\bar{x}$, $\bar{y}$)

An example – study hours vs. exam scores

The least-squares regression line that describes the relationship between hours of study time and exam score is given by

$$\hat{y} = r\frac{S_y}{S_x} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

```
> 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
```

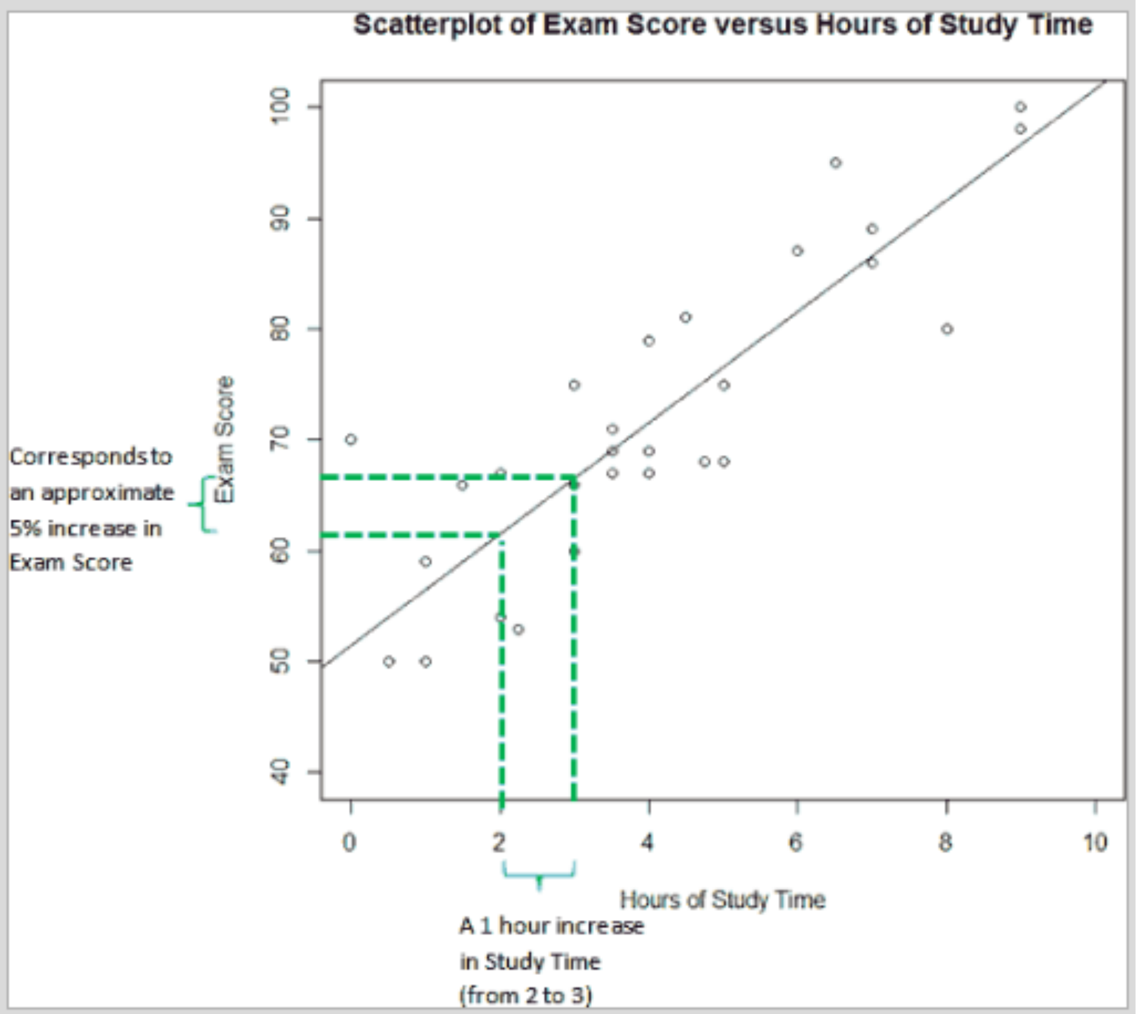
ŷ =51.51 + 5.012 x

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

Interpretation of Results

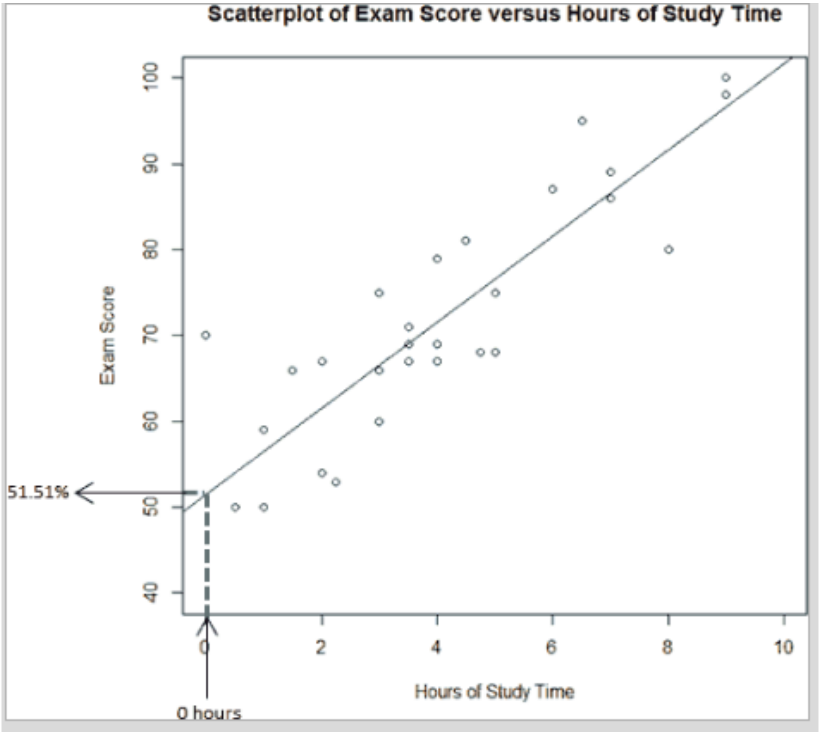
- The estimate of the **slope parameter ($\beta 1$)** gives the expected or predicted change in **the response variable ($\hat{y}$) for a one-unit** increase in the explanatory variable (x). Here, $\beta 1$^=5.012.

- This can be interpreted as the increase in exam score for every one-hour increase in study time. That is, for **each additional hour** that students studied, their exam score improved by around **five percentage 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 11 unit **change**. This means that the average increase in exam scores of students that studied 33 hours versus 22 hours **is the same as the average** increase in exam scores of students who studied 99 hours versus 88 hours.



- The estimate for the **intercept (β_0^) is meaningful** in this case since values of the explanatory **variable near 0 are possible** (that is, students could have theoretically not studied for the exam and in fact one student reported 0 hours of study time). Here, **β_0=51**. This can be interpreted as the average exam grade for those who did not study (spent 0 hours studying).

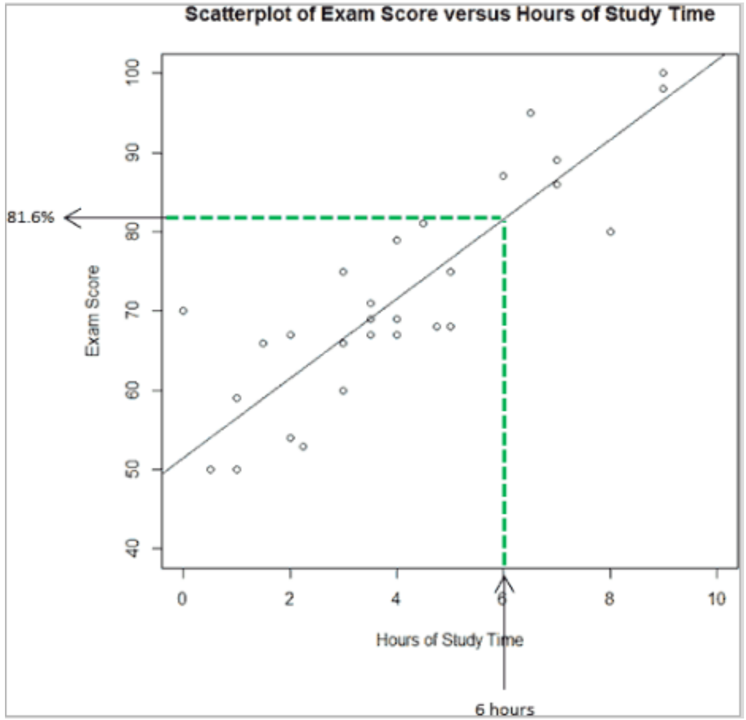
Interpretation of Results - Example

- The least-squares regression is $\hat{y}$ =51.51+5.012x , **If I were planning to study for 6 hours, what should I expect my score to be?**
- Using the equation of the least-squares regression line we can predict what my average exam score might be if I study for 66 hours by plugging in x=6 into the regression equation

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

  That is, my average expected exam score is

  $\hat{y}$ =51.51+5.012(6)≈81.6 if I study for 66 hours.



- The predicted value from the regression equation can be interpreted as follows:

  A student who studies x hours will have an exam score with a mean of

  $\hat{y}$ =51.51 + 5.012 x
- More specifically for this particular example where x=6 hours, the interpretation of the calculation above is as follows: students who study for 6 hours will have an exam score of 81.6 on average.

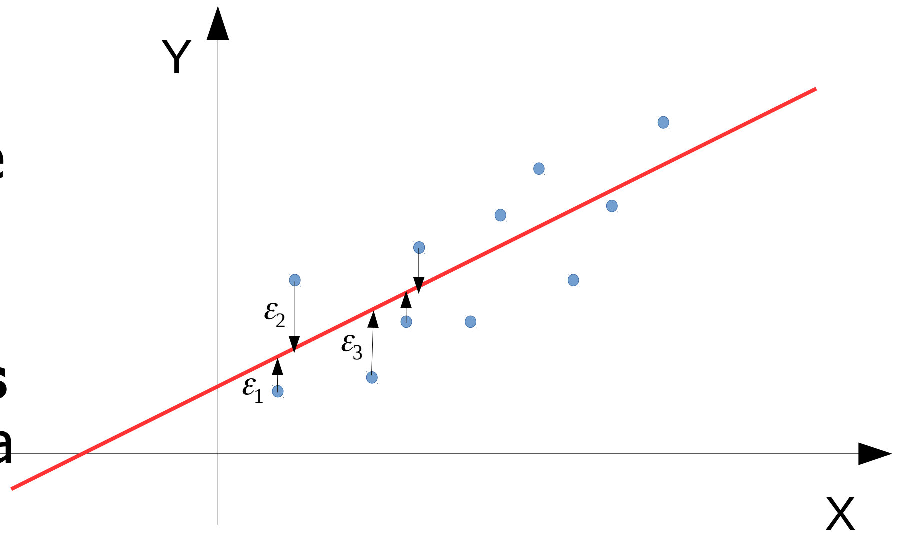
Random Error

- The **full linear regression** model is given by

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

where **ϵ** is the random error

- We recognize that the **true value** of the response variable will **vary** somewhat from the **value predicted** by the regression due to the variability.
- 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.
- After estimating the regression equation, the variability of the data about the regression line helps us to assess the **goodness of fit of the regression line**.

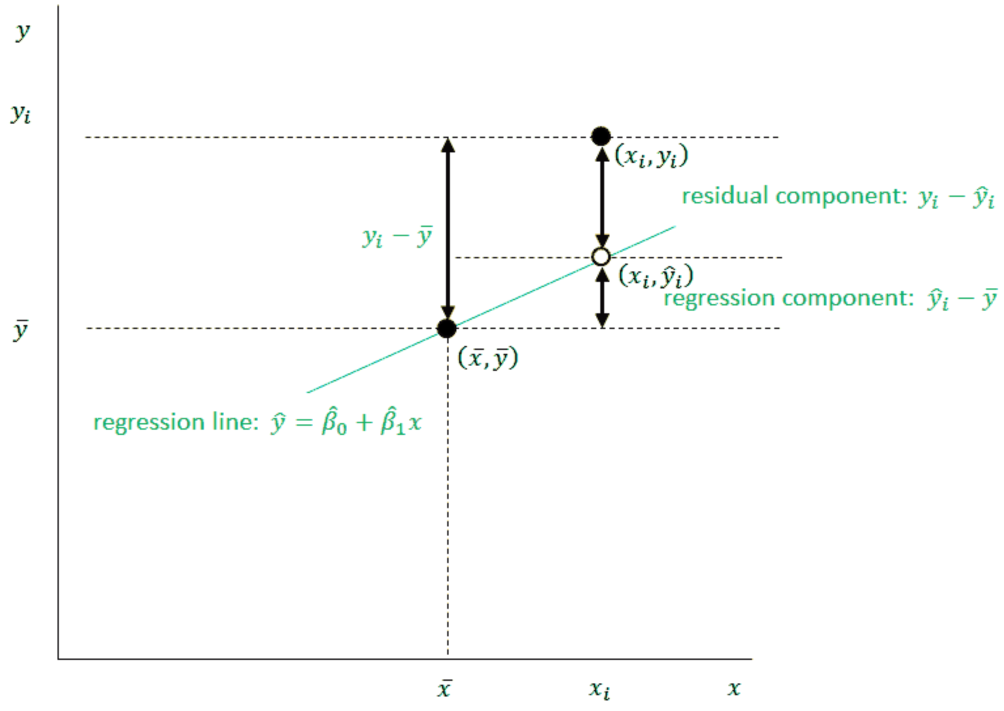


The coefficient of determination or R-squared

- 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.
- It is the **square of the sample correlation coefficient** (that is, $R^2=r^2$) and represents the proportion (percentage) of the variation in the response variable explained by the regression model (equation).
- 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 regression component** and the

  (2) **residual component**.

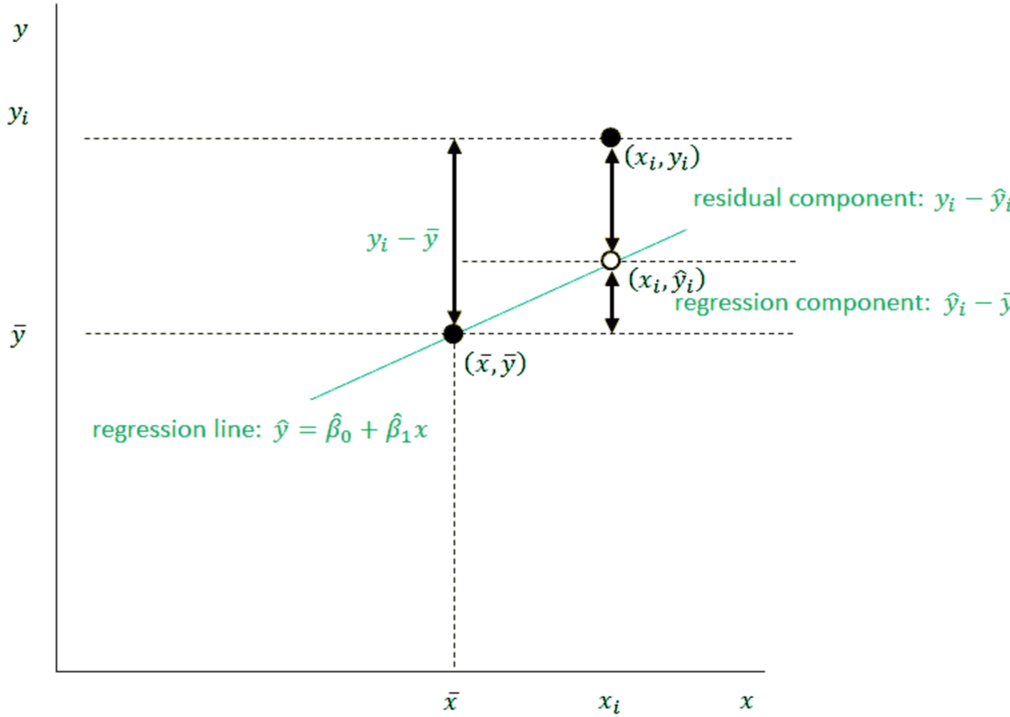


Assessing the Fit of the Regression Line

- For any sample point **(x_i, y_i)**, the **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.

  The **regression component** is equal to

  $$\left(\hat{\beta}_0+\hat{\beta}_1 \hat{x}_i\right)-\bar{y}=\hat{y}_i-\bar{y}$$



- For any sample point **(x_i,y_i)**, the residual or the **residual component** is the **vertical distance** between the **observed** response, **y_i**, and the regression **predicted** response for the value of explanatory variable **x_i**.

  The **residual component** is equal to $y_i-\left(\hat{\beta}_0+\hat{\beta}_1 x_i\right)=y_i-\hat{y}_i$

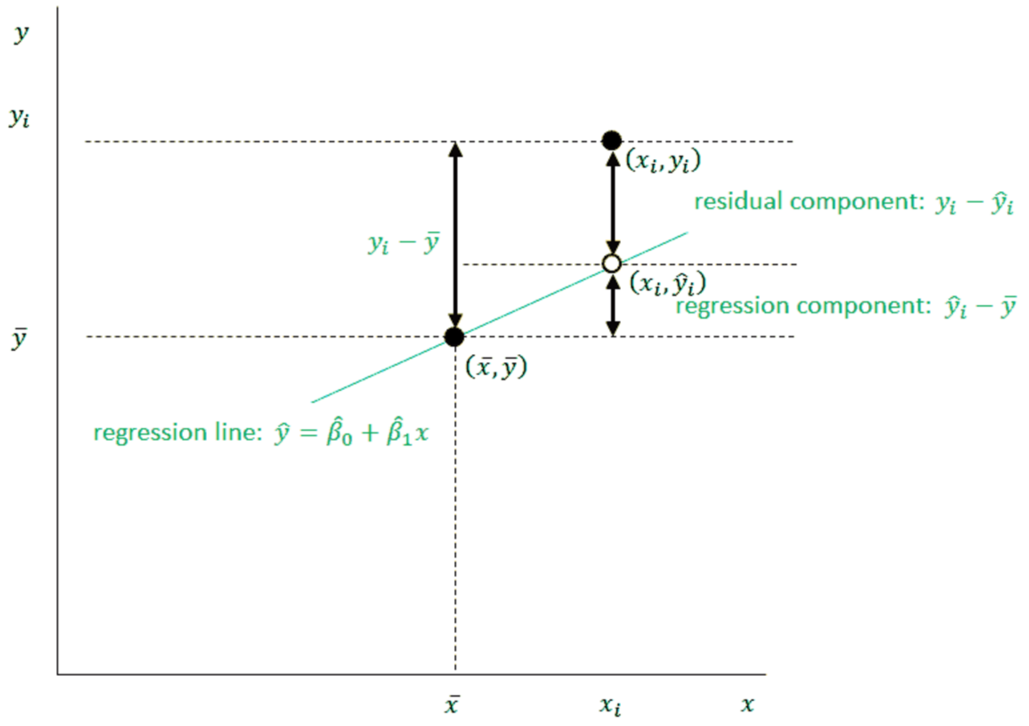
Assessing the Fit of the Regression Line

- 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.

- The regression **lines that fit the data well** will have **regression components** that are **larger** in size than the **residual components across all data points**.
- Regression lines lacking good fit will have residual components that are much larger in size than the regression components across all data points.



The coefficient of determination or R-squared

To quantify this is to take the sum of all squared deviations of the individual data points from the sample mean and break it into each of the component parts to see what proportion represents the regression components versus the residual components.

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

Where
$\sum_{i=1}^{n}(y_i - \bar{y})^2$ (the total sum of squares or Total SS) represents the sum of squares of the deviations of the individual sample points from the sample mean

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (the residual sum of squares or Res SS) represents the sum of squares of the residual components

$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ (the regression sum of squares or Reg SS) represents the sum of squares of the regression components

Assess the fit - The Coefficient of determination or R^2

- One of the measures that we use to **assess the fit of the data is the coefficient of variation** (R2, read "R-squared")

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

- **Coefficient of determination** R^2 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 simple linear regression ***R^2 = r^2 Correlation Coefficient***.

Inference

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are **statistics** and 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 $\hat{\beta}_0$ and $\hat{\beta}_1$ as **unknown** population parameters and determining what we can or can't say about the **unknowns** given the data we observed from our sample.
- 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 (continued)

- **SS is (Sum of Squares)**
- **Reg df = k**, the degrees of freedom of **Reg SS**. It equals to the number of predictors in the model (that is, the number of parameters being estimated besides the intercept).

- **Res df =n−k−1**= the degrees of freedom of **Res SS**. It equals to the number of data points **minus the number of predictors** in the model (that is, the number of parameters being estimated besides the intercept) minus 1.

- **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** (the statistic which is the ratio of the regression mean square to the residual mean square)

- **p-value =** the probability that the observed value of test statistic or a more extreme value could have been observed by chance

	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	$P(F_{\text{Reg df},\text{Res df},\alpha} > F)$
Residual	Res SS	Res df$=n-k-1$	Res MS = Res SS/Res df		
Total	Total SS = Reg SS + Res SS				

An example: calculate R^2

- 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 R^2 (continued)

	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		

- We can use the ANOVA table to calculate R^2: $R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{Reg SS}}{\text{Total SS}} = \frac{90.63}{104.38} = 86.8\%$
- 86.8% of the variability in husband's ages can be explained by wives' ages.
- In SLR, formal tests of hypotheses concern β_1.
- They are generally of the form $\beta_1=0$ (H_0: there is no linear association) versus $\beta_1 \neq 0$ (H_1: there is a linear association).
- $H_0: \beta_1=0$ is rejected if $\widehat{\beta_1}$ is sufficiently far from 0. We reject the claim that the population parameter β_1 is equal to 0, if $\widehat{\beta_1}$, the sample statistic is far from 0.
- There are two tests that can be used to assess these hypotheses: the F-test and the t-test.

F distribution

The **F-distribution** is named after the famous statistician R. A. Fisher.

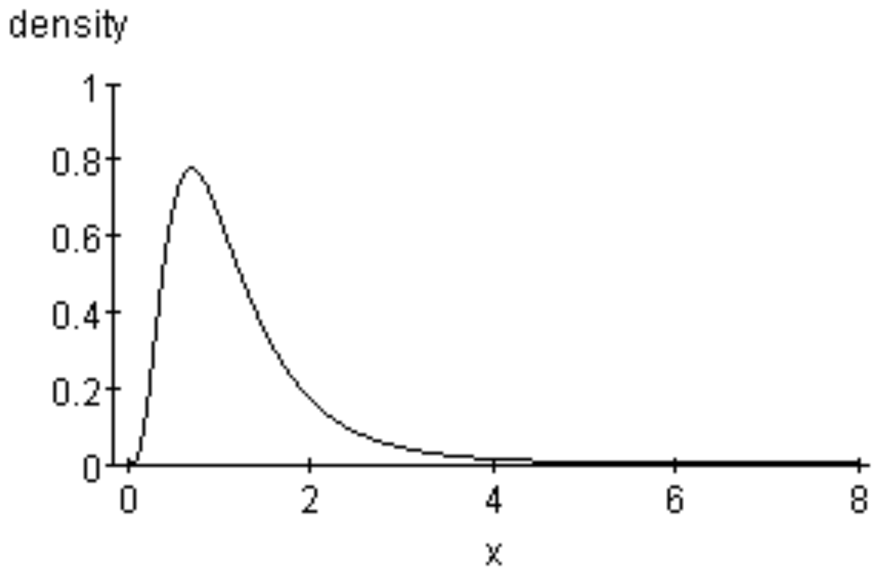
F is the ratio of two variances.

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 of the distribution 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)).**



F-test for Simple Linear Regression (SLR)

- In this test we use an **F statistic**:

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

  which follows an F-distribution with **1 and n−2 degrees of freedom under H_0**.

- The decision rule for a **two-sided level α test** 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** table with **1 degree of freedom** (**numerator**) and **n−2 degrees of freedom** (**denominator**) and associated with a right hand tail probability of α.

Quantities from the F-distribution - R Function

- **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)

> qf(0.95, df1=1, df2=2)
[1] 18.51282
```

Table C. F-Distribution Critical Values

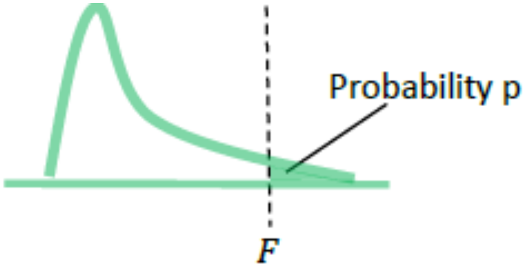


Table entry for p is the critical value F with probability p lying to its right

		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 α=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
$F = \frac{Reg\ MS}{Res\ MS}$ with 1 and n−2=31-2=29 degrees of freedom

3. State the decision rule
F-distribution with 1, 29 degrees of freedom and associated with α=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 SLR (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 $103.2 \geq 4.1830$. 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).

Inference from regression - Using t-test

- In linear regression, the sampling distribution of the **coefficient** estimates from a normal distribution, which is approximated by a **t distribution** due to approximating sigma (population sd) by s (sample sd).
- We can calculate a confidence interval for each estimated coefficient or perform a hypothesis test using t-test.

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

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

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

follows a t-distribution with n−2 degrees of freedom under H_0

T-test for SLR

The decision rule for a two-sided level α test is:
Reject $H_0: \beta_1 = 0$ if $|t| \geq t_{n-2,\alpha/2}$ OR $p \leq \alpha$
Otherwise, do not reject $H_0: \beta_1 = 0$

where $t_{n-2,\alpha/2}$ 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 β1 is given by:

$$\widehat{\beta_1} \pm t_{n-2,\frac{\alpha}{2}} SE_{\beta_1}$$

Interpretation: We can say with 95% confidence that the true value of β_1 is between

$\widehat{\beta_1}$-$t_{n-2,\frac{\alpha}{2}} SE_{\beta_1}$ and $\widehat{\beta_1} + t_{n-2,\frac{\alpha}{2}} SE_{\beta_1}$

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 α=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)
α=0.05

2. Select the appropriate test statistic
$t = \frac{\widehat{\beta_1}}{SE_{\beta_1}}$ 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 α/2=0.025

```
> qt(.975, df=29)
```

$t_{n-2,\ alpha/2}$=2.045

Decision Rule: **Reject H_0 if t≥2.045 or t≤-2.045** **O**therwise, do not reject H_0

A example: F-test for SLR (continued)

4. Compute the test statistic
Using R function summary(m), we get the table:

	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

$t = \frac{\widehat{\beta_1}}{SE_{\beta_1}} = \frac{5.0121}{0.4934} = 10.6$ with df=29
10.16

> confint(m, level=0.95)

$\widehat{\beta_1} \pm t_{n-2,\frac{\alpha}{2}} SE_{\beta_1}$=5.0121 ± 2.045*0.4934 = (4.00, 6.02)

5. Conclusion
Reject H_0 since 10.6≥2.045. We have significant evidence at the α=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). We are 95% confident that the true value of β_1 is between 4.00 and 6.02.

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, $\mathbf{F=t^2}$

- t-test from **correlation** and t-test from **regression** are equivalent and also give same result