Chapter 2

Simple Linear Regression

Summary

Properties of the least-squares estimators (pages 3-4)

Gauss-Markov theorem (page 5)

Some useful properties of the least-squares fit (page 6)

Estimation of σ^2 (pages 7-8)

- Example – Rocket propellant data (pages 9-11)

Hypotheses testing and confidence interval of the slope and intercept (pages 12-16)

- Example – Rocket propellant data (pages 17-19)

Analysis of variance (pages 20-26)

- Example – Rocket propellant data (page 27)

Equivalence between the t and F tests (page 28)

Properties of the Least-Squares Estimators

1. $\hat{\beta}_0$ and $\hat{\beta}_1$ are **linear combinations** of the observations y_i

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i$$

$$c_i = (x_i - \bar{x})/S_{xx} \text{ for } i = 1, 2, \ldots, n.$$

2. The least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are **unbiased estimators** of the model parameters β_0 and β_1.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\Rightarrow E(y_i) = \beta_0 + \beta_1 x_i + \underbrace{E(\varepsilon_i)}_{0}$$

$$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i E(y_i)$$

$$= \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

$$= \beta_1 \qquad \sum_{i=1}^{n} c_i = 0 \text{ and } \sum_{i=1}^{n} c_i x_i = 1$$

$$E(\hat{\beta}_0) = \beta_0$$

3. $$\mathrm{Var}\left(\hat{\beta}_1\right) = \mathrm{Var}\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2 \mathrm{Var}\left(y_i\right) = \sigma^2 \left[\sum_{i=1}^{n} c_i^2\right]$$ use definition of c_i

$\mathrm{Var}(y_i) = \sigma^2$

$$= \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{S_{xx}^2}$$

$$= \frac{\sigma^2}{S_{xx}}$$

$n \uparrow \quad S_{xx} \uparrow \quad \mathrm{var}(\hat{\beta}_1)$ and $\mathrm{var}(\hat{\beta}_0) \downarrow$

4. $$\mathrm{Var}\left(\hat{\beta}_0\right) = \mathrm{Var}\left(\bar{y} - \hat{\beta}_1 \bar{x}\right)$$

$$= \mathrm{Var}\left(\bar{y}\right) + \bar{x}^2 \mathrm{Var}\left(\hat{\beta}_1\right) - 2\bar{x}\mathrm{Cov}\left(\bar{y}, \hat{\beta}_1\right)$$

$\mathrm{Cov}(\bar{y}, \hat{\beta}_1) = 0$

$$= \mathrm{Var}\left(\bar{y}\right) + \bar{x}^2 \mathrm{Var}\left(\hat{\beta}_1\right)$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

Gauss-Markov theorem

$\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance when compared with all other unbiased estimators that are linear combinations of the y_i.

Least-squares estimators are **best linear unbiased estimators**

Some useful properties of the least-squares fit

1. $\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$ (residue) $\Rightarrow$ difference between y_n and $\hat{y}_n$

2. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$ ← results in

3. The least-squares regression line always passes through the **centroid** [the point $(\bar{x}, \bar{y})$] of the data. $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

4. $\sum_{i=1}^{n} x_i e_i = 0$

5. $\sum_{i=1}^{n} \hat{y}_i e_i = 0$

Estimation of σ^2

1. **Corrected sum of squares or total variation in y**

measure of the spread of y_i

$$SS_T \equiv \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

2. **Residual sum of squares**

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy}$$

SS_T

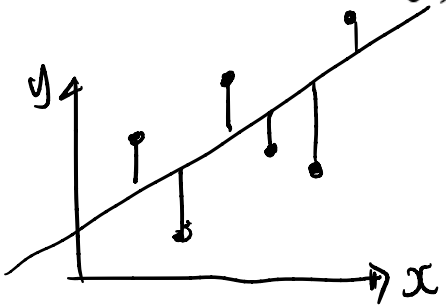
3. **Sum of squares due to the regression** $SS_R = \hat{\beta}_1 S_{xy}$

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}$$

$$SS_T = SS_{Res} + SS_R$$

4. $E(SS_{\text{Res}}) = (n-2)\sigma^2$ **(Linear model theory)**

Unbiased estimator of σ^2



σ^2

$$\hat{\sigma}^2 = \frac{SS_{\text{Res}}}{n-2} = MS_{\text{Res}}$$

MS_{Res} is called the **residual mean square.**

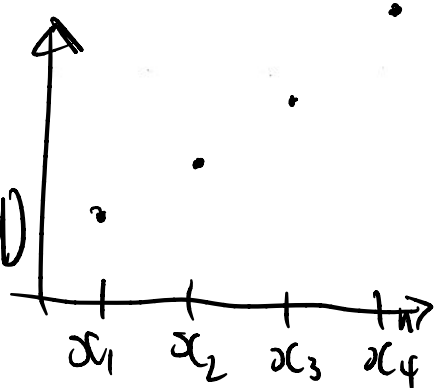
The square root of $\hat{\sigma}^2$ is called the **standard error of regression.**

change model then estimate will change

$\hat{\sigma}^2$ is a **model-dependent** estimate of σ^2.

line required to estimate residue (e_1, e_2, e_3, e_4 etc.)

each time only have 1 value (cannot calculate variation hence model/line required)



The square root of σ^2 is called the **standard error of regression**.

$\hat{\sigma}^2$ is a **model-dependent** estimate of σ^2.

$x_1, x_2, \ldots, x_n \sim \sigma^2$

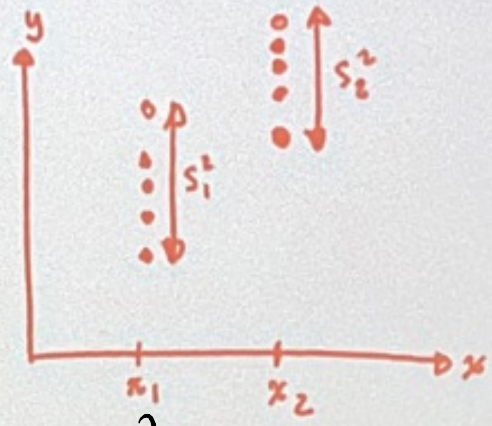
$S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$

x

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

$\text{Var}(\varepsilon) = \sigma^2$

(no model required if multiple data points present)



→ already have variations for us to calculate

$$\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

↑ hence model independent

Example The Rocket Propellant Data

Estimate σ^2 for the rocket propellant data

$$SS_{\mathrm{T}} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$$

$$SS_{\mathrm{Res}} = SS_{\mathrm{T}} - \hat{\beta}_1 S_{xy}$$

$$\hat{\sigma}^2 = \frac{SS_{\mathrm{Res}}}{n-2}$$

```
#estimate sigma^2
SST <- sum(y^2) - (sum(y))^2/length(y)
SST
SSxy <- sum(y*x) - sum(y)*sum(x)/length(y)
SSxy
SSRes <- SST - beta1*SSxy
SSRes
sigma2hat <- SSRes/(length(y)-2)
sigma2hat

#ANOVA table of fitted model
summary.aov(fitted_model)
```

```
> #estimate sigma^2
> SST <- sum(y^2) - (sum(y))^2/length(y)
> SST
[1] 1693738
> SSxy <- sum(y*x) - sum(y)*sum(x)/length(y)
> SSxy
[1] -41112.65
> SSRes <- SST - beta1*SSxy
> SSRes
[1] 166254.9
> sigma2hat <- SSRes/(length(y)-2)
> sigma2hat
[1] 9236.381
>
> #ANOVA table of fitted model
> summary.aov(fitted_model)
             Df  Sum Sq Mean Sq F value   Pr(>F)
x             1 1527483 1527483   165.4 1.64e-10 ***
Residuals    18  166255    9236
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

Hypotheses testing and confidence interval of the slope and intercept

1. The simple linear regression model for a response variable y and a regressor variable x based on observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ can be written as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, 2, ..., n,$$

where ϵ_i is a random variable such that $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, $i = 1, 2, ..., n$ and ϵ_i's are independent.

Assume further that ϵ_i's are normally distributed.

(fit least square line do not need this)

β_{10}

2. We want to test $H_0 : \beta_1 = \beta_{10}$ versus $H_1 : \beta_1 \neq \beta_{10}$

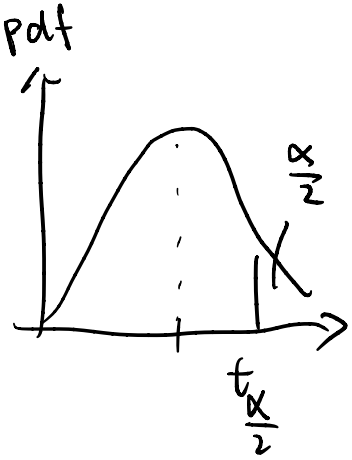
   If H_0 is true, $T = \dfrac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}}$ follows the t distribution with $n-2$ degrees of freedom according to linear model theory.

3. For testing $H_0 : \beta_1 = \beta_{10}$ versus $H_1 : \beta_1 \neq \beta_{10}$, we reject H_0 if

   $$\frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} < -t_{\alpha/2,n-2} \quad \text{or} \quad \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} > t_{\alpha/2,n-2}$$

   where α is the level of significance.

assume error is normally distributed



4. A $100(1-\alpha)\%$ confidence interval for β_1 is given as

   $$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{MS_{Res}/S_{xx}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{MS_{Res}/S_{xx}}$$

5. We want to test $H_0 : \beta_0 = \beta_{00}$ versus $H_1 : \beta_0 \neq \beta_{00}$

   If H_0 is true, $T = \dfrac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})}}$ follows the t distribution with $n - 2$ degrees of freedom according to linear model theory.

6. For testing $H_0 : \beta_0 = \beta_{00}$ versus $H_1 : \beta_0 \neq \beta_{00}$, we reject H_0 if

$$\frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})}} < -t_{\alpha/2,n-2} \quad \text{or} \quad \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})}} > t_{\alpha/2,n-2}$$

7. A $100(1 - \alpha)\%$ confidence interval for β_0 is given as

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})}$$

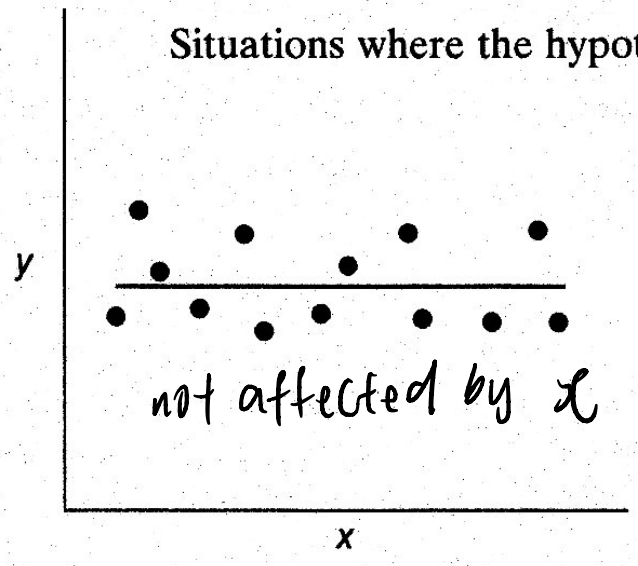
8. For example, to perform tests and construct confidence intervals on individual regression coefficients, the following R codes can be used for the model $y = \beta_0 + \beta_1 x + \epsilon$,

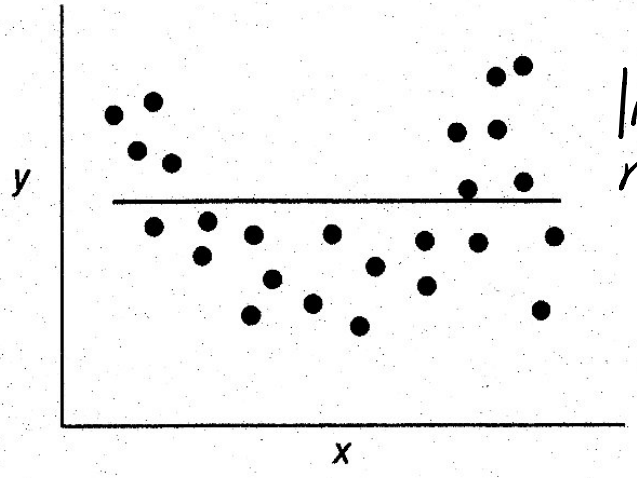
```
summary(lm(y~x))

confint(lm(y~x), level=0.95)
```

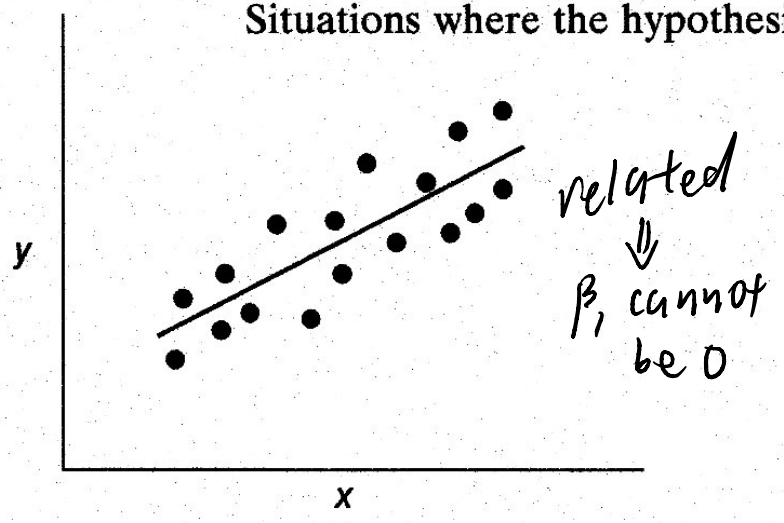
Failing to reject H_0: $\beta_1 = 0$ is equivalent to saying that there is **no linear relationship between *y* and *x*.**

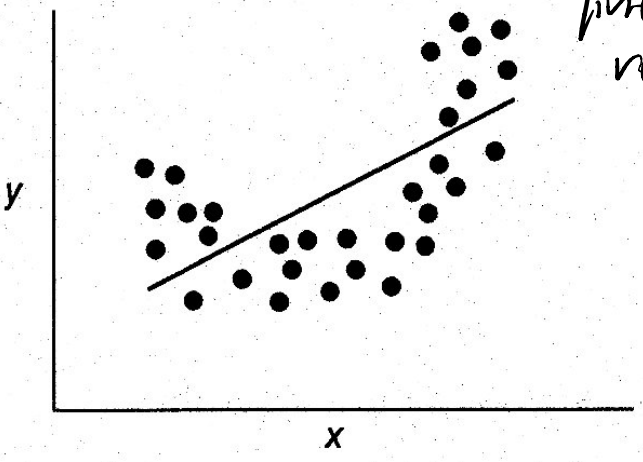
Situations where the hypothesis H_0: $\beta_1 = 0$ is not rejected.





Situations where the hypothesis H_0: $\beta_1 = 0$ is rejected.





Example - Rocket propellant data

```
74  #t test and confidence interval of intercept and slope
75  summary(fitted.model)
76  confint(fitted.model, level=0.95)
77
78  #alternative approach
79  #extract data from fitted model
80  beta0 <- summary(fitted.model)$coeff[1,1]
81  beta0
82  se.beta0 <- summary(fitted.model)$coeff[1,2]
83  se.beta0
84  beta1 <- summary(fitted.model)$coeff[2,1]
85  beta1
86  se.beta1 <- summary(fitted.model)$coeff[2,2]
87  se.beta1
88
89  #95% confidence interval of beta1
90  n <- length(y)
91  beta1 - qt(0.975,df=n-2)*se.beta1
92  beta1 + qt(0.975,df=n-2)*se.beta1
93
94
95  #95% confidence interval of beta0
96  beta0 - qt(0.975,df=n-2)*se.beta0
97  beta0 + qt(0.975,df=n-2)*se.beta0
98
```

```
> #t test and confidence interval of intercept and slope
> summary(fitted.model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-215.98  -50.68   28.74   66.61  106.76

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2627.822     44.184   59.48  < 2e-16 ***
x            -37.154      2.889  -12.86 1.64e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 96.11 on 18 degrees of freedom
Multiple R-squared:  0.9018,    Adjusted R-squared:  0.8964
F-statistic: 165.4 on 1 and 18 DF,  p-value: 1.643e-10

> confint(fitted.model, level=0.95)
                2.5 %    97.5 %
(Intercept) 2534.99540 2720.6493
x            -43.22338  -31.0838
```

p-values (small so null hypothesis rejected)

```
> #alternative approach
> #extract data from fitted model
> beta0 <- summary(fitted.model)$coeff[1,1]
> beta0
[1] 2627.822
> se.beta0 <- summary(fitted.model)$coeff[1,2]
> se.beta0
[1] 44.18391
> beta1 <- summary(fitted.model)$coeff[2,1]
> beta1
[1] -37.15359
> se.beta1 <- summary(fitted.model)$coeff[2,2]
> se.beta1
[1] 2.889107
>
> #95% confidence interval of beta1
> n <- length(y)
> beta1 - qt(0.975,df=n-2)*se.beta1
[1] -43.22338
> beta1 + qt(0.975,df=n-2)*se.beta1
[1] -31.0838
>
>
> #95% confidence interval of beta0
> beta0 - qt(0.975,df=n-2)*se.beta0
[1] 2534.995
> beta0 + qt(0.975,df=n-2)*se.beta0
[1] 2720.649
```

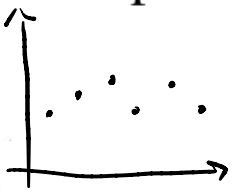
Analysis of Variance

1. The total variation in y, SS_T can be decomposed as

$$SS_T = SS_{Res} + SS_R$$

where SS_{Res} is the residual sum of squares and SS_R is the regression sum of squares. For a given data set, SS_T is a constant. How large SS_R is with respect to SS_{Res} depends on how close the points are to the regression model.

eg. if y does not depend on x ⇒ no variation in y



constant

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

SS_T SS_R SS_{Res}

(total variation in residue)

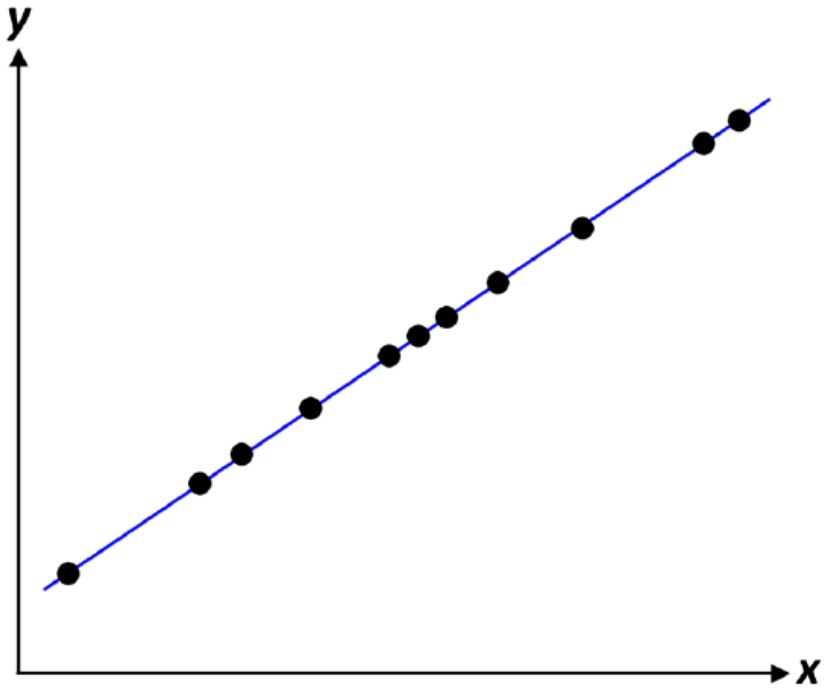
if all points fall on the line, $SS_{Res} = 0$

$$SS_T = SS_R + SS_{Res}$$

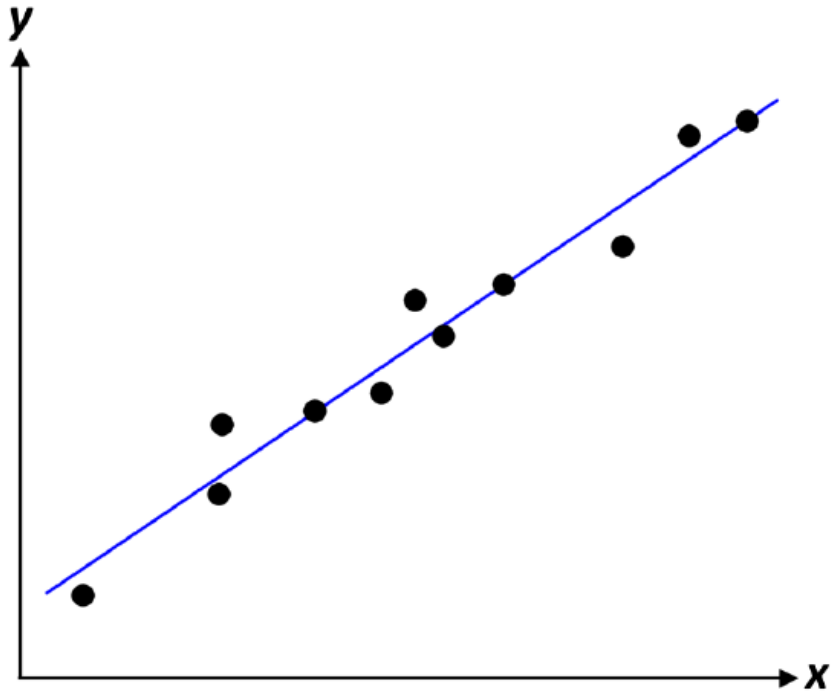
$$SS_R = \hat{\beta}_1 S_{xy}$$

variation either explained by regression line or by error

2. If all the points fall exactly on the model, SS_{Res} will be zero, and $SS_R = SS_T$. In other words, all the variation in y is contributed by the regression model. See the figure for an example of a simple linear regression model where all the points fall exactly on the model.



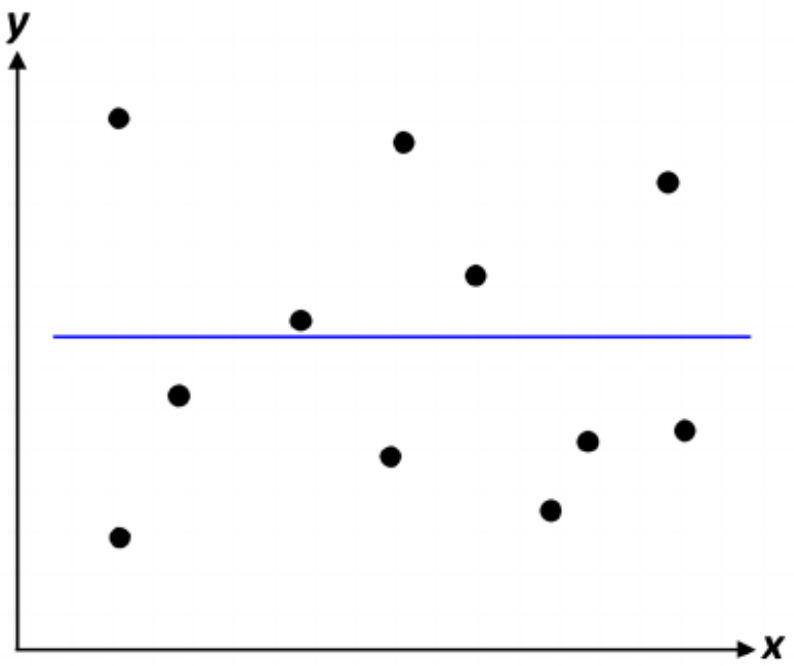
3. If all the points fall close to the regression model, SS_{Res} will be small, and SS_R will be large, hence SS_R/SS_{Res} is large. In other words, most of the variation in y is contributed by the regression model. See the figure for an example of a simple linear regression model where all the points fall close to the model.



4. If the points do not fall close to the model $y = \beta_0 + \beta_1 x + \epsilon$, this means y does not depend on the regressor variables x, consequently SS_{Res} will be large and SS_R will be small, hence SS_R/SS_{Res} is small. The figure shows an example based on simple linear regression where the response y does not depend on x, hence the points will not fall along the line $y = \beta_0 + \beta_1 x + \epsilon$, instead they fall randomly along the line $y = \beta_0 + \epsilon$. In other words, the regression line $y = \beta_0 + \beta_1 x + \epsilon$ contributes very little or none to the variation in y.

test whether β_1 is strong

$SS_{Res} = SS_T$



5. It makes sense to use the ratio SS_R/SS_{Res} to test the hypotheses

   $H_0 : \beta_1 = 0$

   $H_1 : \beta_1 \neq 0$

   If H_0 is true, the ratio will be small. On the other hand, if H_0 is not true, the ratio will be large, so we reject H_0 if the ratio is large.

6. Instead of SS_R/SS_{Res}, we will use $F = \dfrac{SS_R/1}{SS_{Res}/(n-2)} = \dfrac{MS_R}{MS_{Res}}$ because according to linear model theory (can be learned in a graduate course on regression analysis), F follows the F distribution with degrees of freedom 1 and $n-2$ when H_0 is true. For a given level of significance level α, we reject H_0 if $F > F_{\alpha,1,n-2}$.

7. The decomposition of variance and test of $H_0 : \beta_1 = 0$ can be summarized in the following table. This table is commonly known as an analysis of variance (ANOVA) table.

Analysis of variance table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SS_R	1	MS_R	MS_R/MS_{Res}
Residual	SS_{Res}	$n-2$	$MS_{Res} = \hat{\sigma}^2$	
Total	SS_T	$n-1$		

8. For example, to an analysis of variance, the following R codes can be used for the model $y = \beta_0 + \beta_1 x + \epsilon$,

```
summary.aov(lm(y~x))
```

```
anova(lm(y~x))
```

Example - Rocket propellant data

```
55 #ANOVA table of fitted model
56 summary.aov(fitted.model)
57 anova(fitted.model)
```

```
> #ANOVA table of fitted model
> summary.aov(fitted.model)
            Df  Sum Sq Mean Sq F value   Pr(>F)
x            1 1527483 1527483   165.4 1.64e-10 ***
Residuals   18  166255    9236
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(fitted.model)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 1527483 1527483  165.38 1.643e-10 ***
Residuals 18  166255    9236
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Equivakence between the t and F tests

$$T = \frac{\hat{\beta}_1}{\text{se}\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1}{\sqrt{MS_{\text{Res}}/S_{xx}}}$$

$$T^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MS_{\text{Res}}} = \frac{\hat{\beta}_1 S_{xy}}{MS_{\text{Res}}} = \frac{MS_{\text{R}}}{MS_{\text{Res}}} = F$$

The End