Comparing Several Means: ANOVA

Blue Lake snap beans were grown in 12 open-top chambers, which are subject to 4 treatments, 3 each, with O_3 and SO_2 present/absent. The total yield was measured for each chamber.

	Sulfur Dioxide	
Ozone	Absent	Present
Absent	1.52	1.49
	1.85	1.55
	1.39	1.21
Present	1.15	0.65
	1.30	0.76
	1.57	0.69

To compare the means of several, say I, groups (populations), one often uses an **analysis of variance** model, or ANOVA.

For the I populations, we use μ_1, μ_2, ..., μ_I and σ_1, σ_2, ..., σ_I to denote their respective means and standard deviations. Similarly, the sample mean, sample standard deviation, and sample size of the ith population are denoted by $\bar{x}_{i\cdot}$, s_i, and J_i.

Of most interest are the comparisons between the μ_i's.

Group Means and Grand Mean

For the bean growth data,

trt	J_i	$\sum_j x_{ij}$	$\bar{x}_{i\cdot}$
1	3	4.76	1.5867
2	3	4.25	1.4167
3	3	4.02	1.3400
4	3	2.10	0.7000

The grand total of $n = 12$ observations is $\sum_i \sum_j x_{ij} = 15.13$, so the grand mean is

$$\bar{x}_{\cdot\cdot} = \frac{15.13}{12} = 1.2608.$$

The J_i's here are all equal so $\bar{x}_{\cdot\cdot}$ is the mean of $\bar{x}_{i\cdot}$'s. This would not be the case for J_i's unequal.

For J_i's large, by CLT,

$$\bar{X}_{i\cdot} \sim N(\mu_i, \tfrac{\sigma_i^2}{J_i}),$$

and s_i^2 are reliable estimates of σ_i^2.

For J_i's small, one *assumes* normality and $\sigma_1^2 = \cdots = \sigma_I^2 = \sigma^2$.

The **individual sample means** are

$$\bar{x}_{i\cdot} = \tfrac{1}{J_i} \textstyle\sum_{j=1}^{J_i} x_{ij},$$

where x_{ij} is the jth observation in the ith group. The **grand mean** is

$$\bar{x}_{\cdot\cdot} = \tfrac{1}{n} \textstyle\sum_{i=1}^{I} \sum_{j=1}^{J_i} x_{ij},$$

where $n = \sum_{i=1}^{I} J_i$ is the total number of observations in the I groups.

Variation Within Groups

For the bean growth data,

trt	$\sum_j (x_{ij} - \bar{x}_{i\cdot})^2$	s_i^2
1	.112467	.056233
2	.065867	.032933
3	.090600	.045300
4	.006200	.003100

SSE is

$$\sum_i \sum_j (x_{ij} - \bar{x}_{i\cdot})^2 = .275134,$$

and MSE is

$$s_p^2 = \frac{.275133}{12-4} = .034392.$$

For J_i's all equal, $s_p^2 = \sum_i s_i^2 / I$. In general, s_p^2 is a weighted mean of s_i^2 with weights $\propto (J_i - 1)$.

Under the assumption

$$\sigma_1^2 = \cdots = \sigma_I^2 = \sigma^2,$$

one would like to estimate the common variance σ^2 using all available information. Such information is contained in the **sum of squared errors**,

$$\text{SSE} = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (x_{ij} - \bar{x}_{i\cdot})^2$$
$$= \sum_{i=1}^{I} (J_i - 1) s_i^2.$$

The pooled variance estimate is given by

$$s_p^2 = \text{MSE} = \frac{\text{SSE}}{n - I},$$

where $n - I = \sum_{i=1}^{I} (J_i - 1)$.

Variation Between Groups

For the bean growth data, SSTr is given by

$$\sum_i 3(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 = 1.353758,$$

and SST is given by

$$\sum_i \sum_j (x_{ij} - \bar{x}_{\cdot\cdot})^2 = 1.628892.$$

It is easy to verify that

$$\text{SST} = \text{SSTr} + \text{SSE}$$

• If one ignores the grouping, then the sample variance of the n observations is

$$s^2 = \frac{1}{n-1}\text{SST}.$$

To measure the variability between groups, one calculates the **sum of squares for treatments**,

$$\begin{aligned}\text{SSTr} &= \sum_{i=1}^{I} \sum_{j=1}^{J_i} (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 \\ &= \sum_{i=1}^{I} J_i (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2.\end{aligned}$$

It can be shown that

$$\begin{aligned}\sum_i \sum_j (x_{ij} - \bar{x}_{\cdot\cdot})^2 &= \sum_i \sum_j (x_{ij} - \bar{x}_{i\cdot})^2 \\ &\quad + \sum_i \sum_j (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2,\end{aligned}$$

where $\text{SST} = \sum_i \sum_j (x_{ij} - \bar{x}_{\cdot\cdot})^2$.

• For $I = 2$, it can be shown that

$$\text{SSTr} = \frac{(\bar{x}_{1\cdot} - \bar{x}_{2\cdot})^2}{\frac{1}{J_1} + \frac{1}{J_2}}.$$

ANOVA Table, F-Test

Associated with SSE and SST are degrees of freedom $n-I$ and $n-1$. Similarly, SSTr has df $I-1$. Note that

$$n-1=(n-I)+(I-1).$$

Dividing SS by the corresponding df, one gets a **mean square** (MS). An **ANOVA table** summarizes all the information.

Src	SS	df	MS
Trt	SSTr	$I-1$	$\frac{\text{SSTr}}{I-1}$
Error	SSE	$n-I$	$\frac{\text{SSE}}{n-I}$
Total	SST	$n-1$	

MSE is an unbiased estimate of σ^2.

For μ_i's all equal, MSTr is also an unbiased estimate of σ^2. When μ_i's are not all equal, MSTr tends to be larger.

To test the hypotheses

$$H_0: \mu_1=\cdots=\mu_I \quad \text{vs.} \quad H_a: \text{o.w.},$$

Calculate

$$f=\frac{\text{MSTr}}{\text{MSE}},$$

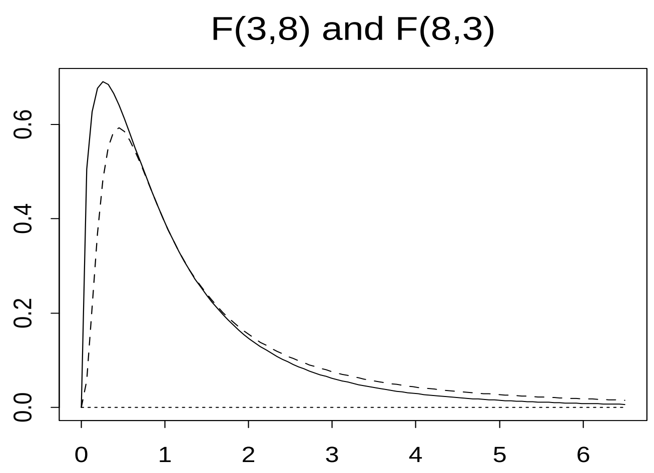
and reject H_0 when $f>F_{\alpha,\nu_1,\nu_2}$, where $\nu_1=I-1$ and $\nu_2=n-I$.

F-Distribution

Let $Y_i \sim N(0,1)$, $i = 1, \ldots, m$, and $Z_j \sim N(0,1)$, $j = 1, \ldots, n$, independent. The distribution of

$$\frac{\sum_{i=1}^m Y_i^2/m}{\sum_{j=1}^n Z_j^2/n}$$

is a **F-distribution** with degrees of freedom $\nu_n = m$, $\nu_d = n$.



For the bean growth data, the ANOVA table is given by

Src	SS	df	MS
Trt	1.3538	3	.4513
Error	0.2751	8	.0344
Total	1.6289	11	

It is easy to calculate

$$f = \frac{.4513}{.0344} = 13.12,$$

which is larger than $F_{.05,3,8} = 4.07$, so we reject H_0 at the 5% significance level.

To obtain $F_{.05,3,8}$ in R, use `qf(.95,3,8)`.

F- and t-tests, Computing Formulas

For $I = 2$, one has

$$f = \frac{\text{MSTr}}{\text{MSE}} = \frac{(\bar{x}_{1\cdot} - \bar{x}_{2\cdot})^2}{s_p^2(\frac{1}{J_1} + \frac{1}{J_2})}.$$

Reject H_0 when $f > F_{\alpha,1,n-2}$.

Compare this with the t-test for $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 \neq \mu_2$,

$$t = \frac{\bar{x}_{1\cdot} - \bar{x}_{2\cdot}}{s_p\sqrt{\frac{1}{J_1} + \frac{1}{J_2}}},$$

with a rejection region $|t| > t_{\alpha/2,n-2}$. We notice that $f = t^2$. Actually, one also has $F_{\alpha,1,\nu} = t^2_{\alpha/2,\nu}$, so the F-test is equivalent to the t-test we learned earlier.

Since SST = SSTr + SSE, one only needs to calculate two of the three terms.

$$\text{SST} = \textstyle\sum_i \sum_j (x_{ij} - \bar{x}_{\cdot\cdot})^2$$

$$= \textstyle\sum_i \sum_j x_{ij}^2 - \frac{(\sum_i \sum_j x_{ij})^2}{n},$$

$$\text{SSTr} = \textstyle\sum_i \sum_j (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$$

$$= \textstyle\sum_i \frac{(\sum_j x_{ij})^2}{J_i} - \frac{(\sum_i \sum_j x_{ij})^2}{n},$$

$$\text{SSE} = \textstyle\sum_i \sum_j (x_{ij} - \bar{x}_{i\cdot})^2$$

$$= \textstyle\sum_i \sum_j x_{ij}^2 - \sum_i \frac{(\sum_j x_{ij})^2}{J_i}.$$

Computing ANOVA: Example

Consider the following data

	Sample		
	1	2	3
	12	8	6
	10	5	2
		3	4
		4	
J_i	2	4	3
$\sum_j x_{ij}$	22	20	12
$\sum_j x_{ij}^2$	244	114	56
$\bar{x}_{i\cdot}$	11	5	4

$n = 9$, $\sum_i \sum_j x_{ij} = 54$, $\bar{x}_{\cdot\cdot} = 6$.

Using the computing formulas,

$$\begin{aligned} \text{SSE} &= 244 + 114 + 56 \\ &\quad - \left(\frac{22^2}{2} + \frac{20^2}{4} + \frac{12^2}{3}\right) \\ &= 24, \\ \text{SSTr} &= \left(\frac{22^2}{2} + \frac{20^2}{4} + \frac{12^2}{3}\right) - \frac{54^2}{9} \\ &= 66. \end{aligned}$$

Since $f = \frac{66/2}{24/6} = 8.25$ and $F_{.05,2,6} = 5.14$, we reject

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

at the 5% significance level.

Parameter Estimation and Testing

For the bean growth data,

$$\bar{x}_{1\cdot} = 1.5867, \quad \bar{x}_{2\cdot} = 1.4167,$$

$$s_p^2 = .0344 = .1855^2,$$

$$J_1 = J_2 = 3, \quad \nu = 8.$$

A 95% CI for μ_1 is

$$1.5867 \pm 2.306\sqrt{\tfrac{.0344}{3}},$$

or $(1.340, 1.834)$, where $t_{.025,8} = 2.306$.

A 95% CI for $\mu_1 - \mu_2$ is

$$.17 \pm 2.306(.1855)\sqrt{\tfrac{2}{3}},$$

or $(-.179, .519)$. One would accept $H_0: \mu_1 = \mu_2$ at the 5% level.

The inferences concerning means are derived from the fact that

$$\bar{X}_{i\cdot} \sim N(\mu_i, \frac{\sigma^2}{J_i}).$$

A $(1-\alpha)100\%$ CI for μ_i is

$$\bar{x}_{i\cdot} \pm t_{\alpha/2,\nu}\sqrt{\frac{s_p^2}{J_i}},$$

where $\nu = n - I$.

A $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ is

$$(\bar{x}_{1\cdot} - \bar{x}_{2\cdot}) \pm t_{\alpha/2,\nu}\sqrt{s_p^2(\frac{1}{J_1} + \frac{1}{J_2})}$$

Tests for hypotheses concerning these parameters can be similarly constructed.

Estimating and Testing Contrasts

For the bean growth data, a contrast of interest is

$$\theta = (\mu_1 - \mu_2) - (\mu_3 - \mu_4).$$

$\theta = 0$ implies no interaction between O_3 and SO_2.

The estimate is given by $\hat{\theta} = \bar{x}_{1\cdot} - \bar{x}_{2\cdot} - \bar{x}_{3\cdot} + \bar{x}_{4\cdot} = -.47$, with a standard error

$$\hat{\sigma}_{\hat{\theta}} = .1855\sqrt{4/3} = .2142.$$

A 95% CI for θ is

$$-.47 \pm 2.306(.2142),$$

or $(-.964, .024)$. One would conclude $\theta = 0$ at the 5% level.

A **linear combination** of means,

$$\theta = c_1\mu_1 + \cdots + c_I\mu_I,$$

is to be estimated by

$$\hat{\theta} = c_1\bar{x}_{1\cdot} + \cdots + c_k\bar{x}_{I\cdot},$$

with a standard error

$$\hat{\sigma}_{\hat{\theta}} = s_p\sqrt{\frac{c_1^2}{J_1} + \cdots + \frac{c_I^2}{J_I}}.$$

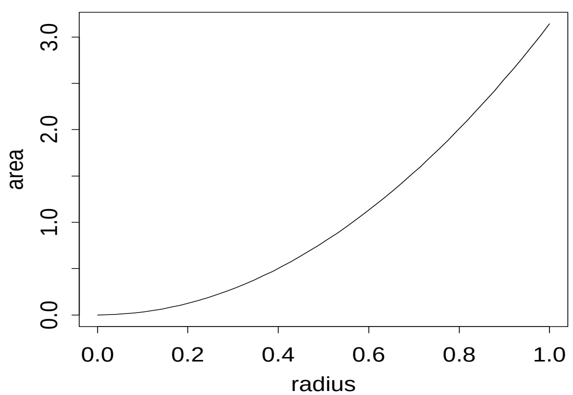
When $c_1, \ldots, c_I$ add to zero, $\sum_i c_i = 0$, such a θ is called a **contrast**. For example, $\mu_1 - \mu_2$ is a contrast.

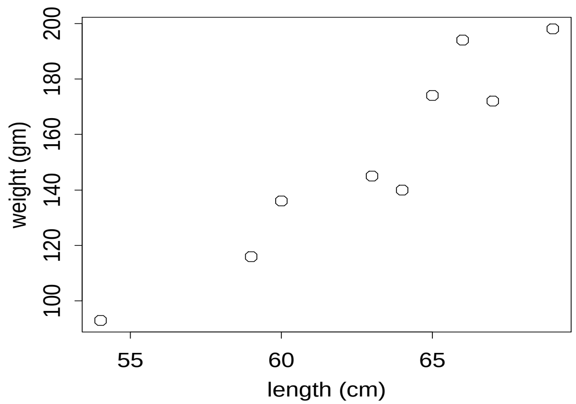
In applications, contrasts are often of the most interest.

Relations Between Variables

Functional relations: $y = f(x)$ deterministic, such as (i) $A = \pi r^2$ for the area A and radius r of a circle; or (ii) $y = \frac{5}{9}(x - 32)$ for thermometer readings $x^o F$ and $y^o C$.

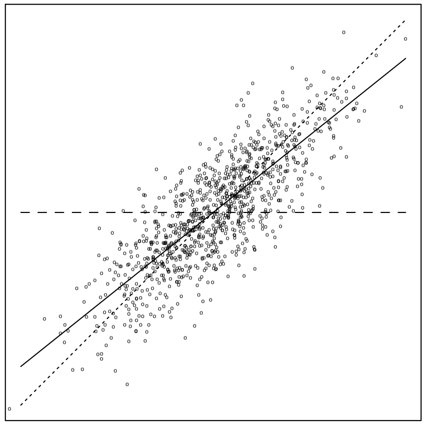
Statistical relations: Variables *tend to* vary together, but there is no deterministic coupling. Among examples are (i) ages of married couples; and (ii) lengths and weights of snakes.





Simple Linear Regression

When studying the heights of father-son pairs, Galton found, in late 19th century, that for fathers taller than average, the average height of their sons is between their height and the average. Ditto for fathers shorter than average.



A simple linear regression is of the form

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

Y – **response** or **dependent var.**

x – **predictor** or **indep. var.**

ϵ – **noise** or **random error**

- Y varies randomly given x. The distribution of Y varies systematically with x through the **regression function** $\mu_{Y \cdot x} = \beta_0 + \beta_1 x$.
- The model has a **systematic part**, $\beta_0 + \beta_1 x$, and a **random part**, ϵ.
- A causal structure is usually implied.

Model Assumptions in SLR

Data come in as pairs (x_i, y_i), and the model is written as

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

It is usually assumed that $\epsilon_i \sim N(0, \sigma^2)$.

Consider

$$Y = 12 + 8x + \epsilon,$$

where $\epsilon \sim N(0, 9)$. Since

$$Y|x=1 \sim N(20, 9),$$

one has

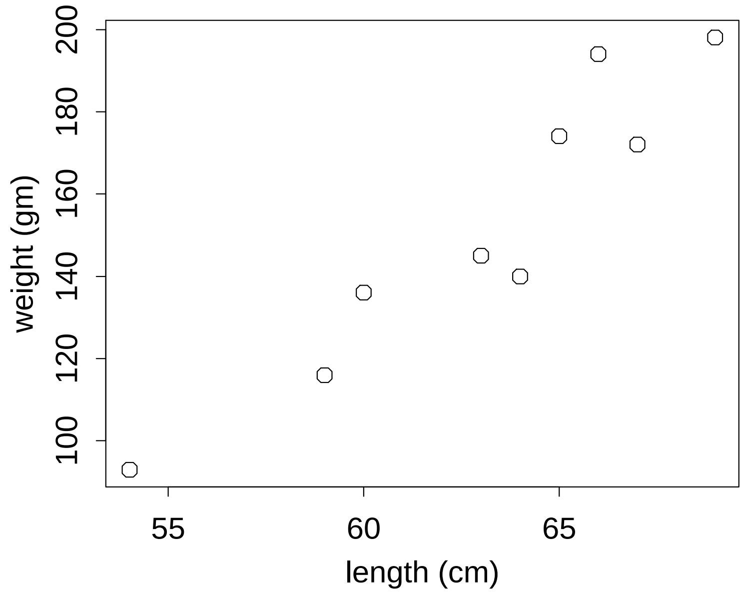
$$P(Y<17|x=1)$$
$$= P(Z< \frac{17-20}{3}) = .1587$$

- In practice, one observes pairs (x_i, y_i), and estimates model parameters β_0, β_1, and σ^2.
- $\mu_{Y \cdot x} = \beta_0 + \beta_1 x$ is a strong assumption.
- The normality assumption can sometimes be weakened to $\mu_{\epsilon_i} = 0$ and $\sigma^2_{\epsilon_i} = \sigma^2$.

Example: Length and Weight of Snakes

Length	Weight
60	136
69	198
66	194
64	140
54	93
67	172
59	116
65	174
63	145

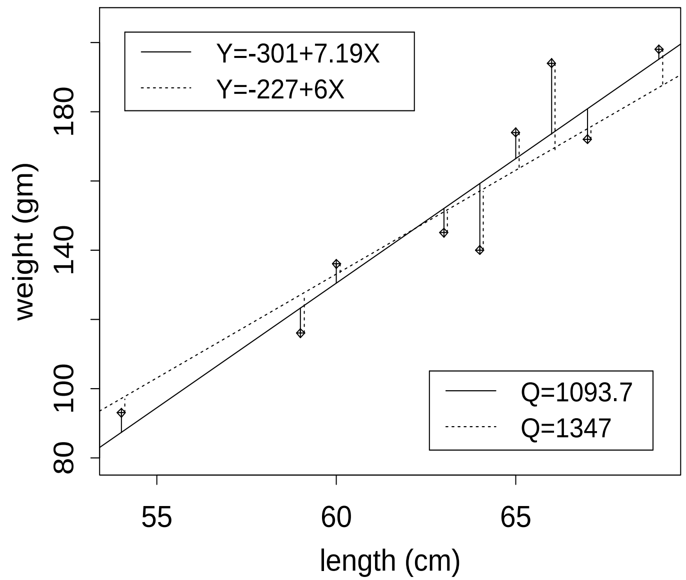
Nine adult females of the snake *Vipera berus* were caught and measured. The lengths and weights are listed on the left and plotted below.



Least Squares Estimates of β_0, β_1

The lengths and weights of female snakes.

The LS estimate of regression function is $Y = -301 + 7.19X$.



Minimizing w.r.t. β_0, β_1

$$Q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2,$$

one obtains the **least squares (LS) estimates** of (β_0, β_1),

$$b_1 = \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

$$b_0 = \hat{\beta}_0 = \bar{y} - b_1 \bar{x}.$$

where

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}),$$

$$S_{xx} = \sum_i (x_i - \bar{x})^2.$$

Fitted Values and Residuals

The lengths and weights of female snakes.

x	y	$\hat{y}$	e
60	136	130.4	5.6
69	198	195.2	2.8
66	194	173.6	20.4
64	140	159.2	-19.2
54	93	87.3	5.7
67	172	180.8	-8.8
59	116	123.2	-7.2
65	174	166.4	7.6
63	145	152.0	-7.0

The mean response $\mu_{Y\cdot x}$ at x is (unbiasedly) estimated by the fitted regression function

$$\hat{\mu}_{Y\cdot x} = \hat{Y} = b_0 + b_1 x.$$

At the data points, one has the **fitted values** (y-hat)

$$\hat{y}_i = b_0 + b_1 x_i,$$

and the **residuals**

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i).$$

The fitted values and residuals satisfy

$$\textstyle\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i,$$

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

Estimation of σ^2

Consider a model

$$Y_i = \mu + \epsilon_i,$$

where $\mu_{\epsilon_i} = 0$ and $\sigma^2_{\epsilon_i} = \sigma^2$. The estimate

$$\hat{y}_i = \hat{\mu} = \bar{y}$$

actually minimizes

$$Q = \textstyle\sum_{i=1}^n (y_i - \mu)^2.$$

An unbiased estimate of σ^2 is

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-1} = \frac{\sum_{i=1}^n e_i^2}{n-1},$$

where $\hat{y}_i$ contains one parameter.

To estimate σ^2, calculate the **residual sum of squares**

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2,$$

and use

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}.$$

- Unbiasedness: $\mu_{s^2} = \sigma^2$.
- To calculate s^2, use

$$\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}},$$

where

$$S_{yy} = \textstyle\sum_i (y_i - \bar{y})^2.$$

Details of Calculation

We use the lengths and weights of snakes to illustrate. Note that

$$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}, \quad S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}.$$

- First summarize the data.

$$\sum x_i = 567 \qquad \sum x_i^2 = 35893$$
$$\sum y_i = 1368 \qquad \sum y_i^2 = 217926$$
$$\sum x_i y_i = 87421$$

- Then calculate

$$\bar{x} = \tfrac{567}{9} = 63, \quad \bar{y} = \tfrac{1368}{9} = 152,$$
$$S_{xx} = 35893 - \tfrac{567^2}{9} = 172,$$
$$S_{yy} = 217926 - \tfrac{1368^2}{9} = 9990,$$
$$S_{xy} = 87421 - \tfrac{567(1368)}{9} = 1237.$$

- Now we have

$$b_1 = \tfrac{1237}{172} = 7.19$$
$$b_0 = 152 - 7.19(63)$$
$$= -301$$

- SSE is given by

$$9990 - \tfrac{1237^2}{172} = 1093.7,$$

so σ^2 is estimated by

$$s^2 = \tfrac{1093.7}{9-2} = 156.24.$$

Inferences Concerning β_1

Lengths and weights of snakes.

We have $b_1 = 7.19$ and

$$s_{b_1} = \sqrt{\tfrac{156.24}{172}} = .953.$$

A 95% CI for β_1 is given by

$$7.19 \pm 2.365(.953),$$

where $t_{.025,7} = 2.365$.

To test the hypotheses

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

we calculate

$$t = \tfrac{7.19-0}{.953} = 7.545,$$

and reject H_0 even at the 1%-level, as $|t| > 3.499 = t_{.005,7}$.

Assume $\epsilon_i \sim N(0, \sigma^2)$.

$$b_1 \sim N(\beta_1, \sigma^2_{b_1}),$$

where $\sigma^2_{b_1} = \sigma^2 / S_{xx}$ is to be estimated by

$$s^2_{b_1} = \frac{s^2}{S_{xx}}.$$

The inferences are based on

$$\frac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}.$$

For example, a $(1-\alpha)100\%$ CI for β_1 is given by

$$b_1 \pm t_{\alpha/2,n-2} s_{b_1}.$$

Analysis of Variance

The lengths and weights of female snakes.

Source	SS	df	MS	F
Model	8896.3	1	8896.3	56.94
Resid	1093.7	7	156.24	
Total	9990.0	8		

Decompose the deviation of y_i from $\bar{y}$,

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

where $(\hat{y}_i - \bar{y})$ is "systematic" and $(y_i - \hat{y}_i)$ is "random". It can be shown that

$$\textstyle\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

$$\text{SST} : (n-1) = \text{SSR} : 1 + \text{SSE} : (n-2)$$

The **ANOVA table** summarizes related information.

Source	SS	df	MS	f
Model	SSR	1	$\frac{\text{SSR}}{1}$	$\frac{\text{MSR}}{\text{MSE}}$
Resid	SSE	$n-2$	$s^2 = \frac{\text{SSE}}{n-2}$	
Total	SST	$n-1$		

F-Test for $\beta_1 = 0$

The lengths and weights of female snakes.

Since

$$f = \frac{8896.3}{156.24} = 56.94,$$
$$F_{.01,1,7} = 12.246,$$

we reject $H_0 : \beta_1 = 0$ at the 1% level.

This is equivalent to the t-test on Slide 19. Note that

$$f = 56.94 = 7.55^2 = t^2,$$
$$F_{.01,1,7} = 12.25 = 3.5^2 = t^2_{.005,7}.$$

It can be shown that

$$\mu_{\text{MSR}} = \sigma^2 + \beta_1^2 S_{xx},$$
$$\mu_{\text{MSE}} = \sigma^2.$$

When $\beta_1 = 0$, one has

$$f = \frac{\text{MSR}}{\text{MSE}} \sim F_{1,n-2}.$$

These lead to the F-test for

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

which rejects H_0 when $F_s > F_{\alpha,1,n-2}$.

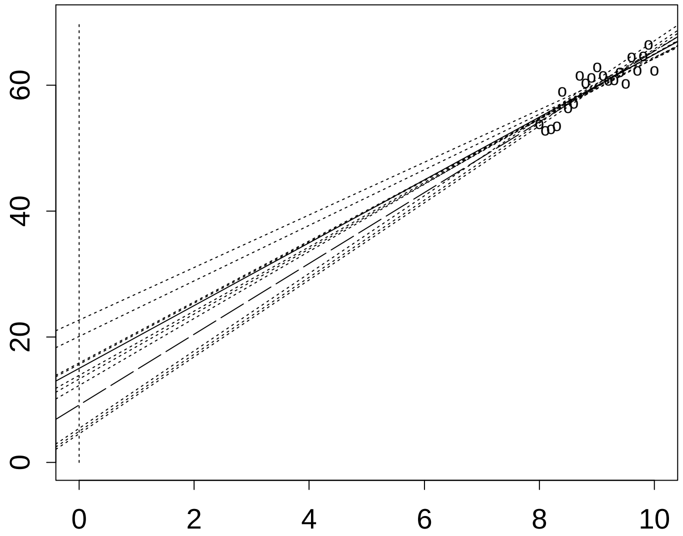
The F- and t-tests are equivalent:

$$\tfrac{\text{MSR}}{\text{MSE}} = f = t^2 = (\tfrac{b_1}{s_{b_1}})^2,$$
$$F_{\alpha,1,n-2} = t^2_{\alpha/2,n-2}.$$

Inferences Concerning β_0

For the lengths and weights of snakes, β_0 has no meaning.

Consider $Y = 15 + 5X + \epsilon$, where $\epsilon \sim N(0, 4)$. Given $x_i = 8(.1)10$, simulate Y_i and estimate the regression function.



Assume $\epsilon_i \sim N(0, \sigma^2)$.

$$b_0 \sim N(\beta_0, \sigma^2_{b_0}),$$

where

$$\sigma^2_{b_0} = \sigma^2\{\tfrac{1}{n} + \tfrac{\bar{x}^2}{S_{xx}}\}$$

is to be estimated by

$$s^2_{b_0} = s^2\{\tfrac{1}{n} + \tfrac{\bar{x}^2}{S_{xx}}\}$$

The inferences are based on

$$\frac{b_0 - \beta_0}{s_{b_0}} \sim t_{n-2}.$$

For $|\bar{x}|$ large, β_0 is hard to estimate, or to interpret.

Inferences Concerning $\mu_{Y\cdot x} = \beta_0 + \beta_1 x$

The lengths and weights of female snakes.

We are to estimate the average weight of snakes of length 60 cm.

$$\hat{Y} = -301 + 7.19(60) = 130.4,$$

$$s^2_{\hat{Y}} = 156.24\{\tfrac{1}{9} + \tfrac{(60-63)^2}{172}\} = 25.535 = 5.053^2,$$

so a 95% CI for $\beta_0 + \beta_1 60$ is

$$130.4 \pm 2.365(5.053),$$

or $(118.45, 142.35)$.

Assume $\epsilon_i \sim N(0, \sigma^2)$.

$$\hat{Y} \sim N(\beta_0 + \beta_1 x, \sigma^2_{\hat{Y}}),$$

where $\hat{Y} = b_0 + b_1 X$, and

$$\sigma^2_{\hat{Y}} = \sigma^2\{\tfrac{1}{n} + \tfrac{(x-\bar{x})^2}{S_{xx}}\}$$

is to be estimated by

$$s^2_{\hat{Y}} = s^2\{\tfrac{1}{n} + \tfrac{(x-\bar{x})^2}{S_{xx}}\}.$$

The inferences are based on

$$\frac{\hat{Y} - (\beta_0 + \beta_1 x)}{s_{\hat{Y}}} \sim t_{n-2}.$$

For $|x - \bar{x}|$ large, $\beta_0 + \beta_1 x$ is hard to estimate.

Prediction of New Observation

The lengths and weights of female snakes.

We are to predict the weight of a snake of length 60 cm.

$$\hat{Y} = 130.4,$$

$$s^2 = 156.24,$$

$$s^2_{\hat{Y}} = 25.535$$

so a 95% PI for Y at $X = 60$ is

$$130.4 \pm 2.365\sqrt{156.24 + 25.535},$$

or $(98.51, 162.29)$. This is wider than the CI for $\beta_0 + \beta_1 60$.

To predict a new response at x,

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

one has to allow for the variability of ϵ.

With β_0, β_1, and σ^2 *known*, the **prediction interval**

$$(\beta_0 + \beta_1 x) \pm z_{\alpha/2}\sigma$$

"covers" Y with probability $1 - \alpha$.

With $\beta_0 + \beta_1 x$ estimated by $\hat{Y} = b_0 + b_1 x$, we use

$$\hat{Y} \pm t_{\alpha/2,n-2}\sqrt{s^2 + s^2_{\hat{Y}}},$$

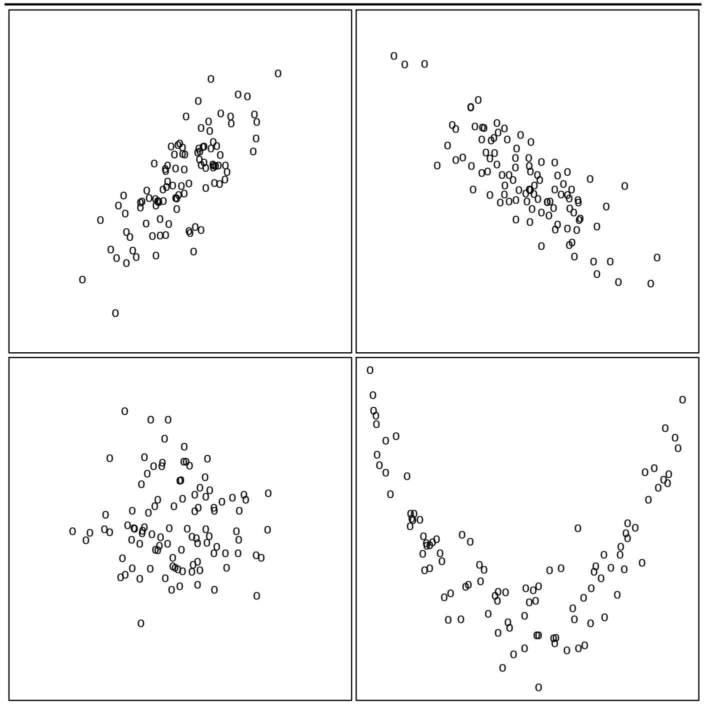
where the variances of $\hat{Y}$ and ϵ are estimated by $s^2_{\hat{Y}}$ and s^2.

R^2, Correlation

Lengths and weights of snakes.

$$R^2 = \frac{8896.3}{9990} = .891$$

$$r = \frac{1237}{\sqrt{172(9990)}} = .944$$



The **coefficient of determination**, or R^2,

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}},$$

measures the amount of variation explained by the model.

The **coefficient of correlation**,

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

measures the *linear* association between X and Y.

- $0 \le R^2 \le 1$. $-1 \le r \le 1$. $R^2 = r^2$.