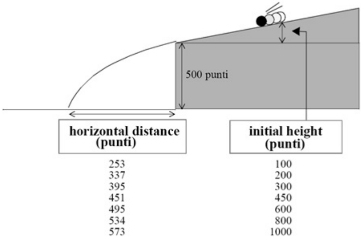
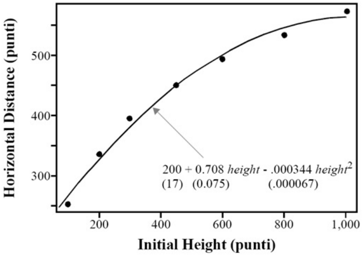
Stat GR5205 Lecture 7

Jingchen Liu

Department of Statistics
Columbia University

Galileo's experiment





$$distance = \beta_0 + \beta_1 height + \beta_2 height^2 + \varepsilon$$

Galileo's experiment

variable	coefficient	standard error	t-statistic	p-value
intercept	199.91	16.8	11.93	0.0003
height	0.71	0.075	9.5	0.0007
height^2	- 0.00034	0.000067	5.15	0.007

$R^2 = 0.99 \quad \hat{\sigma} = 13.6$

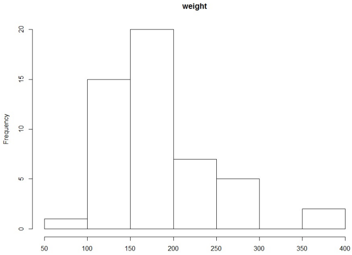
Galileo's experiment

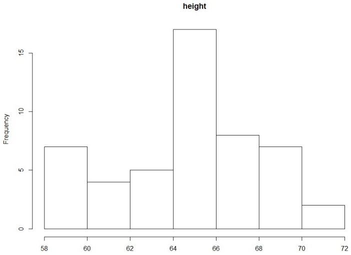
$$distance = \beta_0 + \beta_1 height + \beta_2 height^2 + \beta_3 height^3 + \varepsilon$$

Example

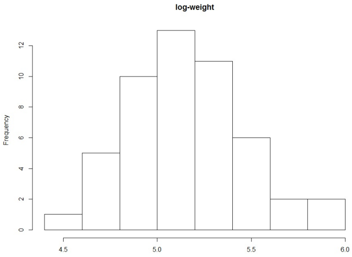
- 50 samples
- 5 Asian, 15 African American, 30 Whites
- Weight and height
- Coding of the design matrix

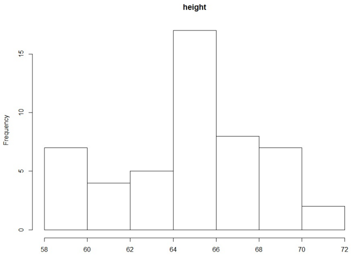
Example


weight
Frequency
0
5
10
15
20
50
100
150
200
250
300
350
400


height
Frequency
0
5
10
15
58
60
62
64
66
68
70
72

Example


log-weight
Frequency
0
2
4
6
8
10
12
4.5
5.0
5.5
6.0


height
Frequency
0
5
10
15
58
60
62
64
66
68
70
72

Example

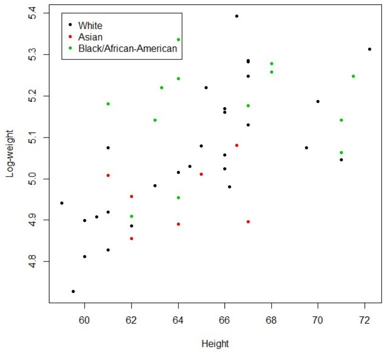


Figure: Height (inch) versus log-weight (log-lb)

Example

- $$\log(weight) = \beta_0 + \beta_1 height + \varepsilon$$
- $$\log(weight) = \beta_0 + \beta_{Asian} I_{Asian} + \beta_{Black} I_{Black} + \beta_1 height + \varepsilon$$
- $$\begin{aligned}\log(weight) &= \beta_0 + \beta_{Asian} I_{Asian} + \beta_{Black} I_{Black} + \beta_1 height \\ &\quad + \beta_{Asian,H} I_{Asian} height + \beta_{Black,H} I_{Black} height + \varepsilon\end{aligned}$$

Example

- $$\log(weight) = \beta_0 + \beta_1 height + \varepsilon$$
- $$\log(weight) = \beta_0 + \beta_{Asian} I_{Asian} + \beta_{Black} I_{Black} + \beta_1 height + \varepsilon$$
- $$\begin{aligned} \log(weight) &= \beta_0 + \beta_{Asian} I_{Asian} + \beta_{Black} I_{Black} + \beta_1 height \\ &\quad + \beta_{Asian,H} I_{Asian} height + \beta_{Black,H} I_{Black} height + \varepsilon \end{aligned}$$

Example

- $$\log(weight) = \beta_0 + \beta_1 height + \varepsilon$$

- $$\log(weight) = \beta_0 + \beta_{Asian} I_{Asian} + \beta_{Black} I_{Black} + \beta_1 height + \varepsilon$$

- $$\begin{aligned} \log(weight) &= \beta_0 + \beta_{Asian} I_{Asian} + \beta_{Black} I_{Black} + \beta_1 height \\ &\quad + \beta_{Asian,H} I_{Asian} height + \beta_{Black,H} I_{Black} height + \varepsilon \end{aligned}$$

The extra sum-of-squares F test

- Question: is there any difference among the three groups aside from that due to height difference
- The formulation

$$\log(weight) = \beta_0 + \beta_{Asian} + \beta_{Black} + \beta_1 height + \varepsilon$$

- The hypotheses

$$H_0 : \beta_{Asian} = \beta_{Black} = 0 \quad H_1 : \ otherwise$$

The extra sum-of-squares F test

- Question: is there any difference among the three groups aside from that due to height difference
- The formulation

$$\log(weight) = \beta_0 + \beta_{Asian} + \beta_{Black} + \beta_1 height + \varepsilon$$

- The hypotheses

$$H_0 : \beta_{Asian} = \beta_{Black} = 0 \quad H_1 : \text{ } otherwise$$

The extra sum-of-squares F test

- Question: is there any difference among the three groups aside from that due to height difference
- The formulation

$$\log(weight) = \beta_0 + \beta_{Asian} + \beta_{Black} + \beta_1 height + \varepsilon$$

- The hypotheses

$$H_0 : \beta_{Asian} = \beta_{Black} = 0 \quad H_1 : \ otherwise$$

The full model versus the reduced model

- Full model (H_1)

$$\log(weight) = \beta_0 + \beta_{Asian} + \beta_{Black} + \beta_1 height + \varepsilon$$

- Reduced model (H_0)

$$\log(weight) = \beta_0 + \beta_1 height + \varepsilon$$

- Comparing the full model against the reduce model

The full model versus the reduced model

- Full model (H_1)

$$\log(\mathit{weight}) = \beta_0 + \beta_{Asian} + \beta_{Black} + \beta_1 \mathit{height} + \varepsilon$$

- Reduced model (H_0)

$$\log(\mathit{weight}) = \beta_0 + \beta_1 \mathit{height} + \varepsilon$$

- Comparing the full model against the reduce model

The full model versus the reduced model

- Full model (H_1)

$$\log(weight) = \beta_0 + \beta_{Asian} + \beta_{Black} + \beta_1 height + \varepsilon$$

- Reduced model (H_0)

$$\log(weight) = \beta_0 + \beta_1 height + \varepsilon$$

- Comparing the full model against the reduce model

Analysis of variance

- ANOVA of the full model

$$SST = SSR_{full} + SSE_{full}$$

- ANOVA of the reduced model

$$SST = SSR_{reduced} + SSE_{reduced}$$

- Inequality

$$SSE_{extra} = SSE_{reduced} - SSE_{full} > 0$$

- Reject H_0 if SSE_{extra} is large
- Distribution of SSR_{extra}

Analysis of variance

- ANOVA of the full model

$$SST = SSR_{full} + SSE_{full}$$

- ANOVA of the reduced model

$$SST = SSR_{reduced} + SSE_{reduced}$$

- Inequality

$$SSE_{extra} = SSE_{reduced} - SSE_{full} > 0$$

- Reject H_0 if SSE_{extra} is large
- Distribution of SSR_{extra}

Analysis of variance

- ANOVA of the full model

$$SST = SSR_{full} + SSE_{full}$$

- ANOVA of the reduced model

$$SST = SSR_{reduced} + SSE_{reduced}$$

- Inequality

$$SSE_{extra} = SSE_{reduced} - SSE_{full} > 0$$

- Reject H_0 if SSE_{extra} is large
- Distribution of SSR_{extra}

Analysis of variance

- ANOVA of the full model

$$SST = SSR_{full} + SSE_{full}$$

- ANOVA of the reduced model

$$SST = SSR_{reduced} + SSE_{reduced}$$

- Inequality

$$SSE_{extra} = SSE_{reduced} - SSE_{full} > 0$$

- Reject H_0 if SSE_{extra} is large
- Distribution of SSR_{extra}

Analysis of variance

- ANOVA of the full model

$$SST = SSR_{full} + SSE_{full}$$

- ANOVA of the reduced model

$$SST = SSR_{reduced} + SSE_{reduced}$$

- Inequality

$$SSE_{extra} = SSE_{reduced} - SSE_{full} > 0$$

- Reject H_0 if SSE_{extra} is large
- Distribution of SSR_{extra}

Extra sums of squares test

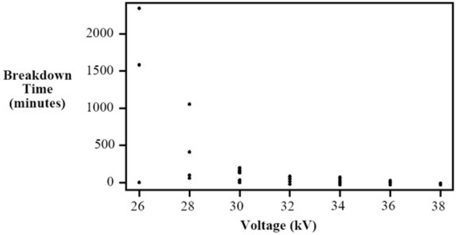
- Test statistic

$$F - statistic = \frac{SSE_{extra}/(p_{full} - p_{reduced})}{SSE_{full}/(n - p_{full})}$$

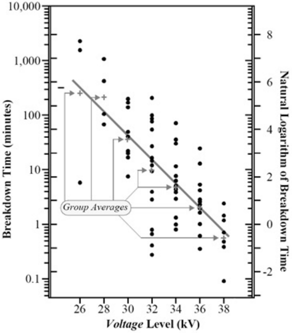
ANOVA

source	sum of sq	d.f.	mean sq	F-stat	p-value
regression	28.7	3	9.6	248	< 0.0001
height	21.4	1	21.4	553	< 0.0001
additional race	7.3	2	7.6	94	< 0.0001
residual	55.7	1441	0.039		
total	84.4	1444			

Lack-of-fit test



Lack-of-fit test



Lack-of-fit test

source	sum of sq	d.f.	mean sq	F-stat	p-value
regression	190	1	190	78	< 0.0001
residual	180	74	2.4		
total	370	75			

source	sum of sq	d.f.	mean sq	F-stat	p-value
between group	196	6	33	13	< 0.0001
residual	174	69	2.5		
total	370	75			

Lack-of-fit test

$$F - statisitc = \frac{(196 - 190)/5}{174/69} = 0.48$$

Some notation

- Reduced model

$$SST = SSR(X_1) + SSE(X_1)$$

- Full model

$$SST = SSR(X_1, X_2) + SSE(X_1, X_2)$$

- Extra sums of squares

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$

- The coefficients of partial determination

$$R^2_{X_2|X_1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

Some notation

- Reduced model

$$SST = SSR(X_1) + SSE(X_1)$$

- Full model

$$SST = SSR(X_1, X_2) + SSE(X_1, X_2)$$

- Extra sums of squares

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$

- The coefficients of partial determination

$$R^2_{X_2|X_1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

Some notation

- Reduced model

$$SST = SSR(X_1) + SSE(X_1)$$

- Full model

$$SST = SSR(X_1, X_2) + SSE(X_1, X_2)$$

- Extra sums of squares

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$

- The coefficients of partial determination

$$R^2_{X_2|X_1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

Some notation

- Reduced model

$$SST = SSR(X_1) + SSE(X_1)$$

- Full model

$$SST = SSR(X_1, X_2) + SSE(X_1, X_2)$$

- Extra sums of squares

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$

- The coefficients of partial determination

$$R^2_{X_2|X_1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

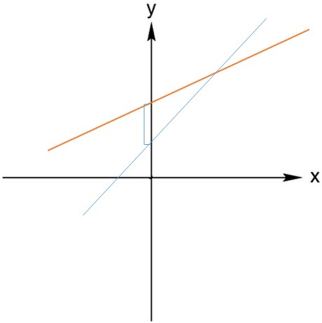
ANOVA table

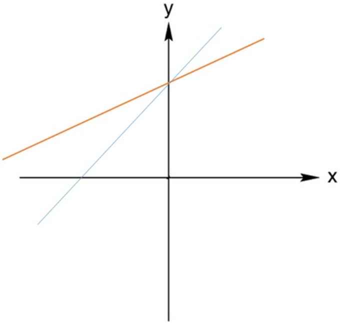
$$SST = SSR + SSE$$

source	sums of sq	d.f.	mean sum of sq	F-stat	p-value
Regression	SST	p-1	SST/(p-1)		
Residual	SSE	n-p	SSE/(n-p)		
Total	SST				

Standardization

Correlation between β_0 and β_1





Standardization

- Reducing the correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$

$$x_i^* = x_i - \bar{x}$$

- Rescaling the covariates

$$x_i^* = \frac{x_i - \bar{x}}{SD(y)}$$

- Rescaling the response

$$y_i^* = \frac{y_i - \bar{y}}{SD(y)}$$

Standardization

- Reducing the correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$

$$x_i^* = x_i - \bar{x}$$

- Rescaling the covariates

$$x_i^* = \frac{x_i - \bar{x}}{SD(y)}$$

- Rescaling the response

$$y_i^* = \frac{y_i - \bar{y}}{SD(y)}$$

Standardization

- Reducing the correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$

$$x_i^* = x_i - \bar{x}$$

- Rescaling the covariates

$$x_i^* = \frac{x_i - \bar{x}}{SD(y)}$$

- Rescaling the response

$$y_i^* = \frac{y_i - \bar{y}}{SD(y)}$$

Standardization of covariates

- The standardized regression coefficients

$$\hat{\beta}_0^* = 0 \qquad \hat{\beta}_1^* = \rho_{xy}$$

- Transform back to the original coefficients
- Standardization does not alter prediction

Standardization of covariates

- The standardized regression coefficients

$$\hat{\beta}_0^* = 0 \qquad \hat{\beta}_1^* = \rho_{xy}$$

- Transform back to the original coefficients
- Standardization does not alter prediction

Standardization of covariates

- The standardized regression coefficients

$$\hat{\beta}_0^* = 0 \qquad \hat{\beta}_1^* = \rho_{xy}$$

- Transform back to the original coefficients
- Standardization does not alter prediction

Multicollinearity – orthogonality

$$X = (x_{ij})_{n \times p}$$

- Standardized covariates

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, \quad SD(x_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) = 1$$

- Uncorrelated covariates,

$$\sum_{i=1}^{n} x_{ij_1} x_{ij_2} = 0$$

Multicollinearity – orthogonality

$$X = (x_{ij})_{n \times p}$$

- Standardized covariates

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, \quad SD(x_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) = 1$$

- Uncorrelated covariates,

$$\sum_{i=1}^{n} x_{ij_1} x_{ij_2} = 0$$

Multicollinearity – orthogonality

- Uncorrelated covariates

$$X^\top X = (n-1)I_{p\times p}$$

- The least-square estimate

$$\hat\beta = (X^\top X)^{-1}X^\top Y = X^\top Y/(n-1)$$

- Individual estimated coefficient

$$\hat\beta_j = \frac{1}{n-1}\sum_{i=1}^{n} x_{ij}y_i$$

Multicollinearity – orthogonality

- Uncorrelated covariates

$$X^\top X = (n-1) I_{p \times p}$$

- The least-square estimate

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y = X^\top Y / (n-1)$$

- Individual estimated coefficient

$$\hat{\beta}_j = \frac{1}{n-1} \sum_{i=1}^{n} x_{ij} y_i$$

Multicollinearity – orthogonality

- Uncorrelated covariates

$$X^\top X = (n-1) I_{p\times p}$$

- The least-square estimate

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y = X^\top Y/(n-1)$$

- Individual estimated coefficient

$$\hat{\beta}_j = \frac{1}{n-1}\sum_{i=1}^{n} x_{ij} y_i$$

Multicollinearity – orthogonality

- Variance

$$Var(\beta_j) = \frac{1}{(n-1)^2} \sum_{i=1}^{n} x_{ij}^2 = \frac{1}{n-1}$$

- Covariance

$$Cov(\beta_{j_1}, \beta_{j_2}) = \frac{1}{(n-1)^2} \sum_{i=1}^{n} x_{ij_1} x_{ij_2} = 0$$

Multicollinearity – orthogonality

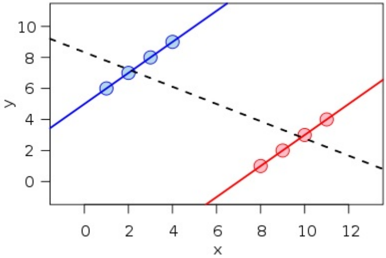
- Variance

$$Var(\beta_j) = \frac{1}{(n-1)^2} \sum_{i=1}^{n} x_{ij}^2 = \frac{1}{n-1}$$

- Covariance

$$Cov(\beta_{j_1}, \beta_{j_2}) = \frac{1}{(n-1)^2} \sum_{i=1}^{n} x_{ij_1} x_{ij_2} = 0$$

A paradox due to multicollinearity



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Quantification of multicollinearity

$$y, x_1, x_2, \ldots, x_p$$

- With all covariates standardized, we consider

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- $Var(\hat{\beta}_1) = \sigma^2/(n-1)$
- Consider

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + \varepsilon$$

- $Var(\tilde{\beta}) = \sigma^2 (X^\top X)^{-1}$

Quantification of multicollinearity

$$y, x_1, x_2, \ldots, x_p$$

- With all covariates standardized, we consider

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- $Var(\hat{\beta}_1) = \sigma^2/(n-1)$
- Consider

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + \varepsilon$$

- $Var(\tilde{\beta}) = \sigma^2 (X^\top X)^{-1}$

Quantification of multicollinearity

$$y, x_1, x_2, ..., x_p$$

- With all covariates standardized, we consider

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- $Var(\hat{\beta}_1) = \sigma^2/(n-1)$
- Consider

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_p x_p + \varepsilon$$

- $Var(\tilde{\beta}) = \sigma^2(X^\top X)^{-1}$

Quantification of multicollinearity

$$y, x_1, x_2, ..., x_p$$

- With all covariates standardized, we consider

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- $Var(\hat{\beta}_1) = \sigma^2/(n-1)$
- Consider

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_p x_p + \varepsilon$$

- $Var(\tilde{\beta}) = \sigma^2 (X^\top X)^{-1}$

Quantification of multicollinearity

- Variance inflation factor (VIF)

$$VIF(\beta_1) = \frac{Var(\tilde{\beta}_1)}{Var(\hat{\beta}_1)}$$

- A representation

$$VIF(\beta_1) = \frac{1}{1 - R_1^2}$$

where R_1^2 is the coefficient of determination of x_1 on x_2, $x_3, \ldots, x_p$.

Quantification of multicollinearity

- Variance inflation factor (VIF)

$$VIF(\beta_1) = \frac{Var(\tilde{\beta}_1)}{Var(\hat{\beta}_1)}$$

- A representation

$$VIF(\beta_1) = \frac{1}{1 - R_1^2}$$

where R_1^2 is the coefficient of determination of x_1 on x_2, x_3,...,x_p.