Introductory Statistics
12: Correlations

Richard Veale

Graduate School of Medicine
Kyoto University

https://youtu.be/oq9lHbXXnpc

Lecture Video at above link

Today's goals

- To understand how to describe and test associations of two metric variables with correlations and linear regression.

- To learn about the F-distribution, its relation to the χ^2-distribution and the normal distribution, and its usage in regression models.

- To understand that simple linear regression (variables: x, y) can be extended to multiple variables (variables x_1, x_2, ..., x_i, y) -> multiple linear regression.

Today's topics

1) Covariance
2) Correlation
3) Significance test for a correlation
4) Linear regression
5) Multiple linear regression

Categorical vs Metric

Categorical data ~ Categorical data:

	Polio	No Polio
Vaccine	33	200712
Placebo	110	201119

Or we compared metric data between two groups:

Categorical data ~ Metric data:

Smoking Mums

$\bar{x}_1 = 3223$

$s_1 = 594$

Nonsmoking Mums

$\bar{x}_2 = 3051$

$s_1 = 673$

Metric-Metric?!

But what if we were to look at the relationship between one metric variable with another?

For example:

Amount of daily salt consumption [g/day]

~

Blood Pressure [mm Hg]

-> Correlation

Variance

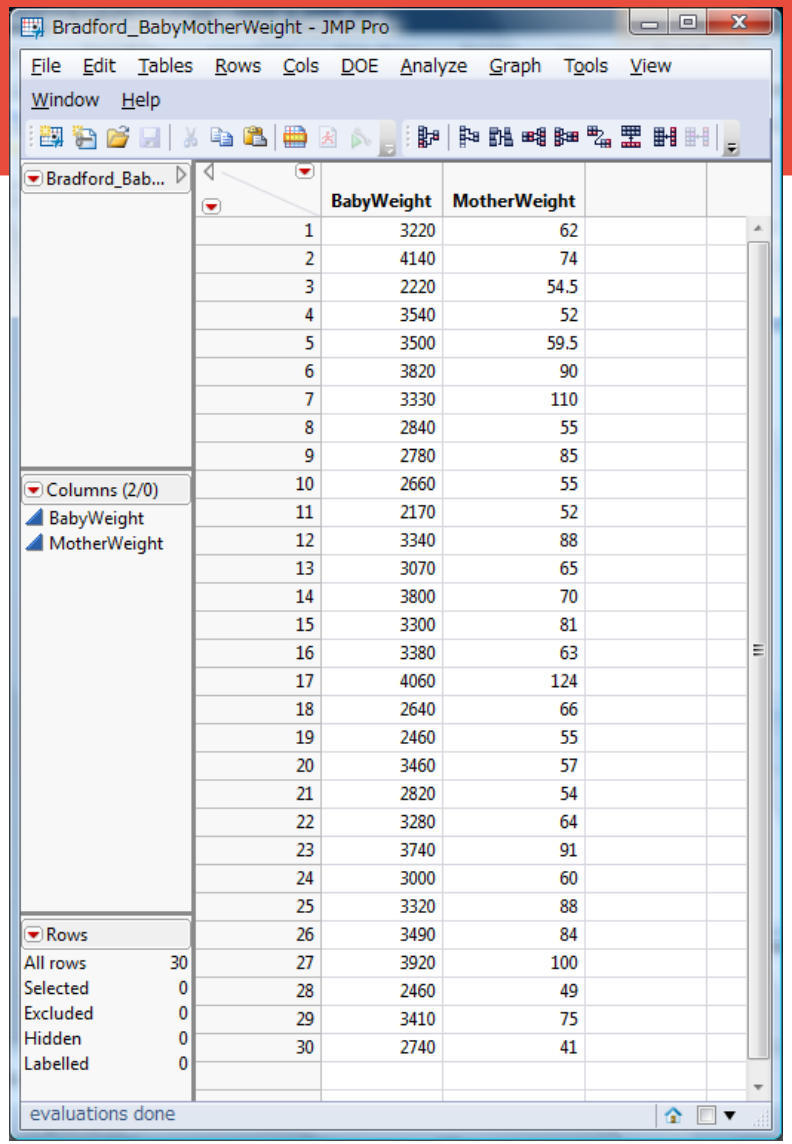
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

s^2: *sample variance*

$\bar{x}$: *arithmetic sample mean*

n: *sample size*

x_i: *ith sample*



	BabyWeight	MotherWeight
1	3220	62
2	4140	74
3	2220	54.5
4	3540	52
5	3500	59.5
6	3820	90
7	3330	110
8	2840	55
9	2780	85
10	2660	55
11	2170	52
12	3340	88
13	3070	65
14	3800	70
15	3300	81
16	3380	63
17	4060	124
18	2640	66
19	2460	55
20	3460	57
21	2820	54
22	3280	64
23	3740	91
24	3000	60
25	3320	88
26	3490	84
27	3920	100
28	2460	49
29	3410	75
30	2740	41

For example, Bradford Babies:

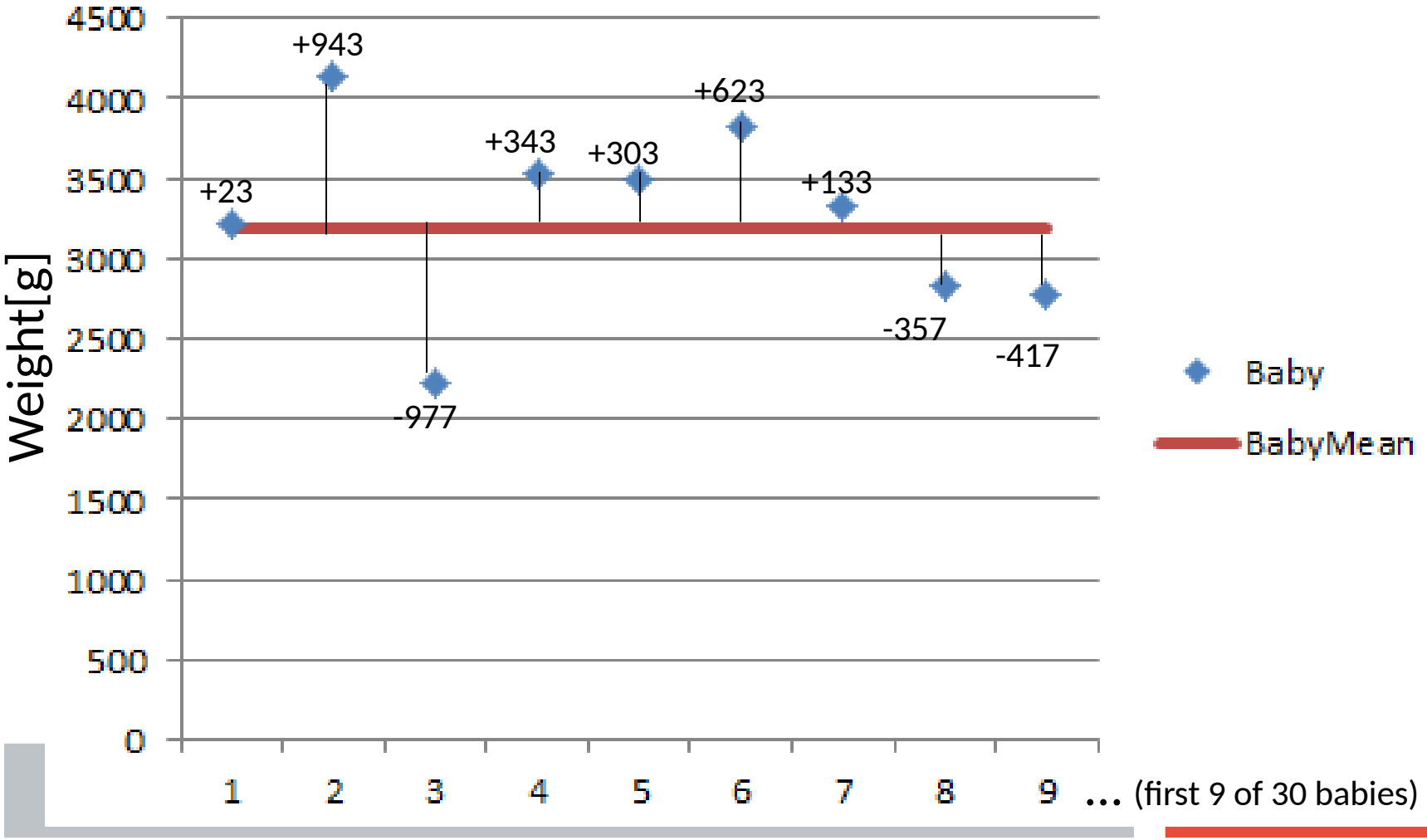
$$s^2 = \frac{1}{30-1}\sum_{i=1}^{n}(x_i - 3197)^2 =$$

$$\frac{1}{29}\left[(3220-3197)^2 + \cdots (2740-3197)^2\right]$$

$$s \approx 529.13$$

Variance of Metric Data

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_1 - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})$$

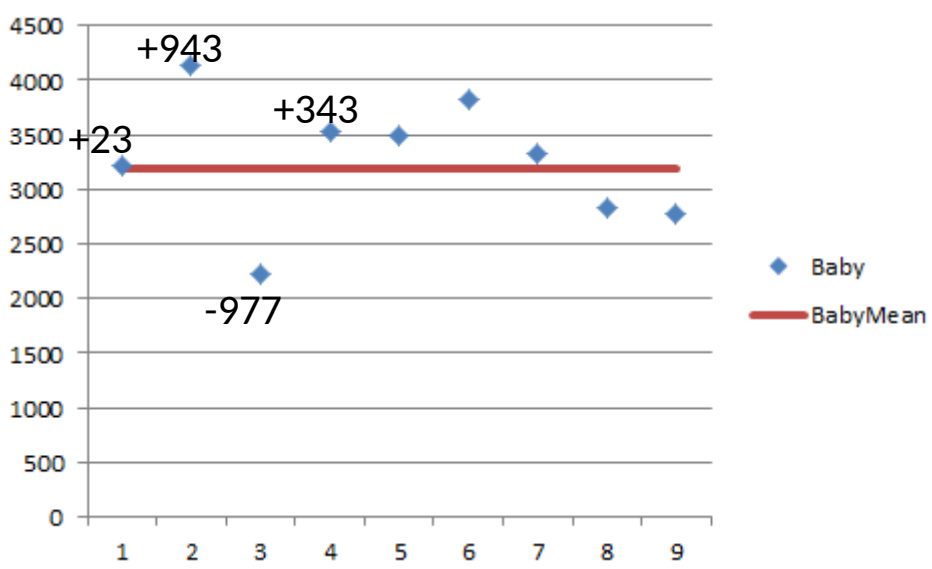


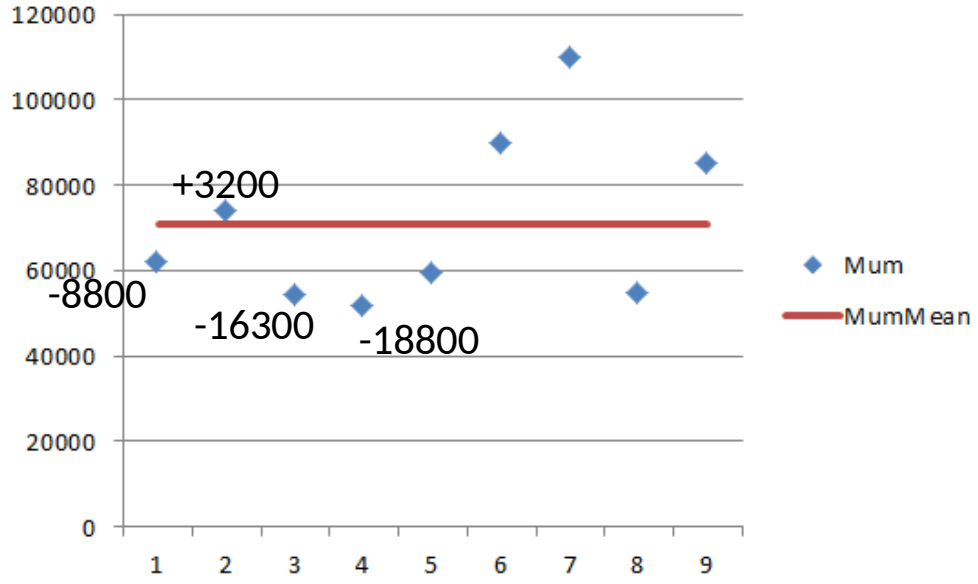
... (first 9 of 30 babies)

Covariance

$$cov(x,y)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

Weight[g]



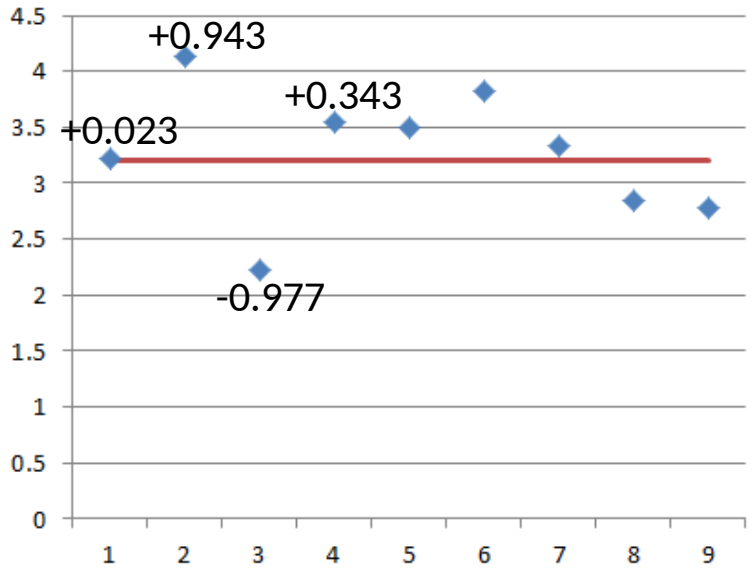


… (first 9 of 30 baby-mum pairs)

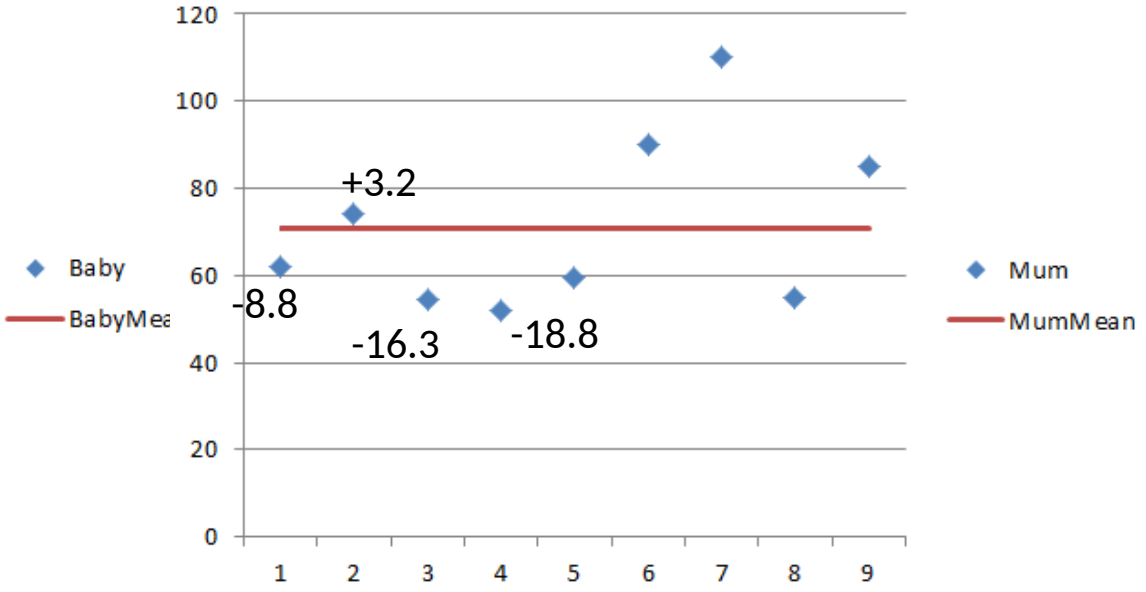
$$cov(x,y)=\frac{1}{29}\left[(+23)\cdot(-8800)+(+943)\cdot(+3200)+(-977)\cdot(-16300)+(+342)\cdot(-18800)+\cdots\right]\approx 6518344$$

Covariance (depends on scale!!!)

$$cov(x,y)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$



Weight[kg]



… (first 9 of 30 baby-mum pairs)

$$cov(x,y)=\frac{1}{29}\left[(+0.023)\cdot(-8.8)+(+0.943)\cdot(+3.2)+(-0.977)\cdot(-16.3)+(+0.342)\cdot(-18.8)+\cdots\right]\approx 6.518344$$

-> Covariance depends on scale!
so we use correlation

Correlation

$$cov(x,y)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

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

For example, weight in **g**:

$$cov(x,y)\approx 6518344$$

$$r=\frac{6518344}{529\cdot 19599}\approx 0.6286$$

For example, weight in **kg**:

$$cov(x,y)\approx 6.518344$$

$$r=\frac{6.518344}{0.529\cdot 19.599}\approx 0.6286$$

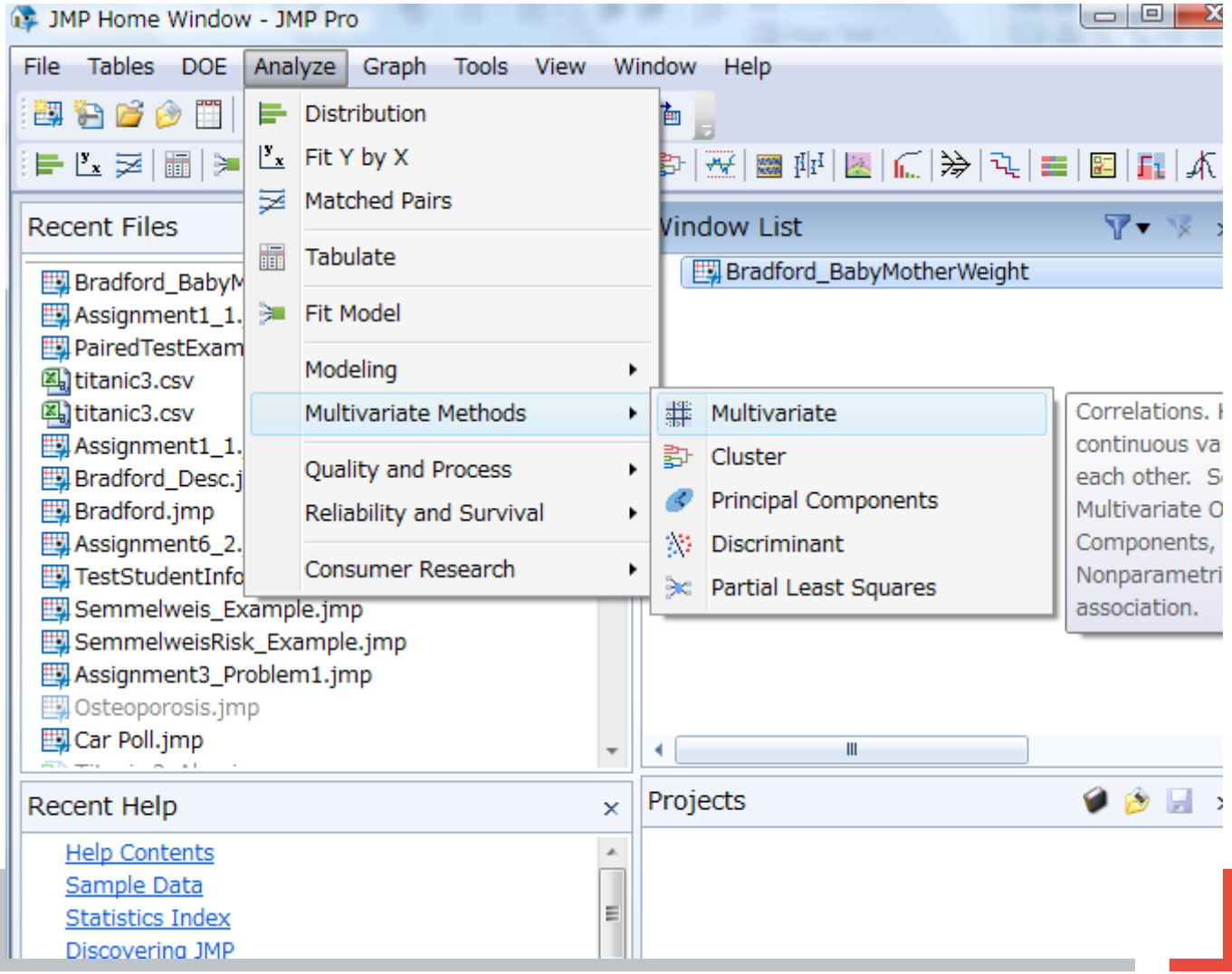
Correlation

$$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

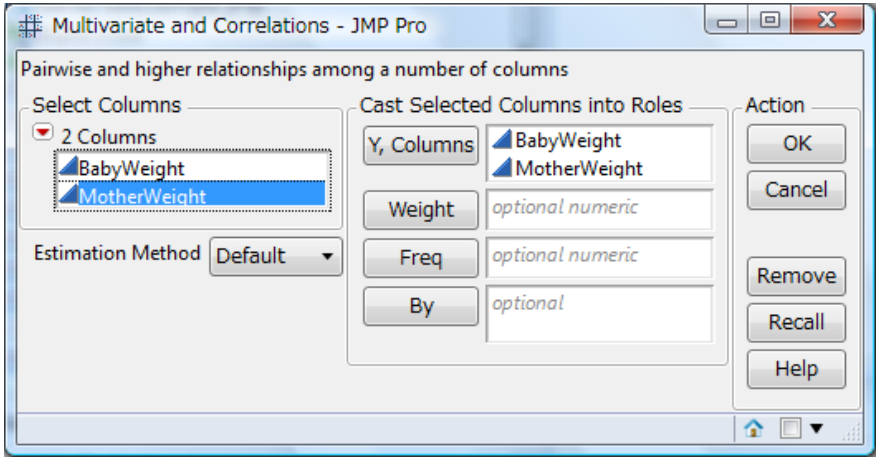
$$r = \frac{cov(x,y)}{s_x s_y} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

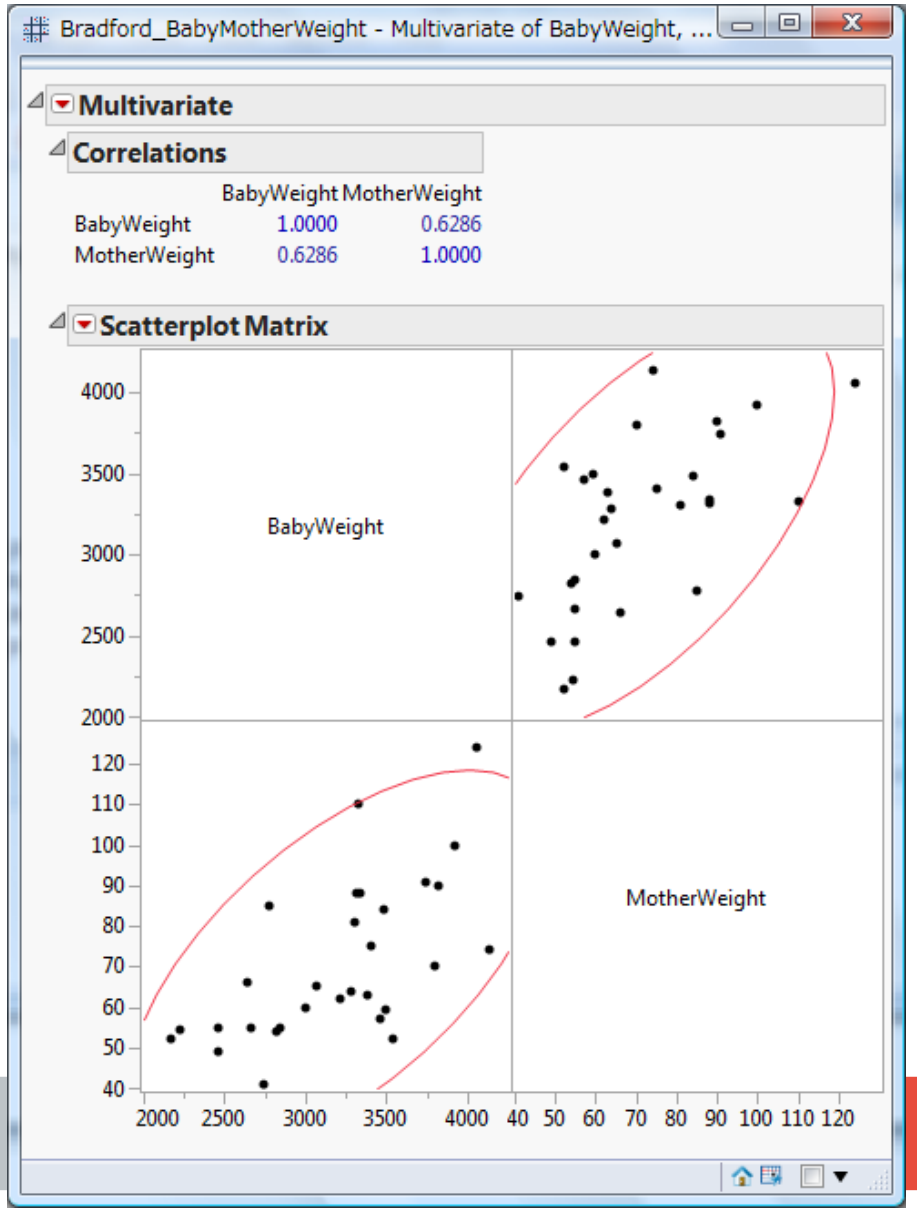
Correlation - JMP

In JMP: Analyze->Multivariate Methods->Multivariate



Correlation - JMP





Significance Test for Correlations

Assuming x and y come from bivariate normal distributions:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

t follows a t-distribution with n-2 df.

we can use this to test the null hypothesis:

H_0: null hypothesis: $\rho = 0$

H_a: alternative hypothesis: $\rho \neq 0$

(ρ is the 'true' correlation)

Significance Test for Correlations

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

t follows a t-distribution with n-2 df.

Example:

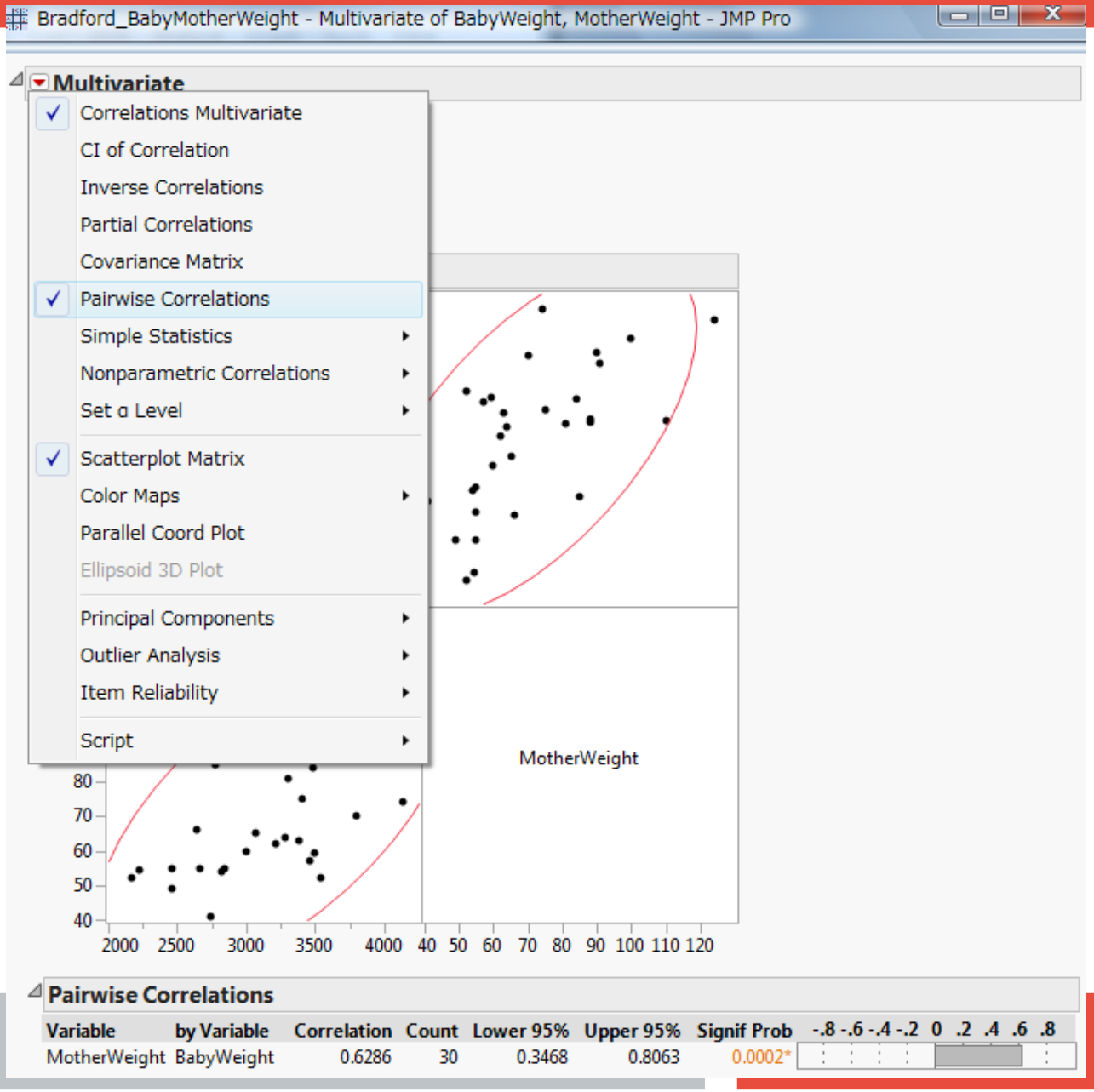
$t_{crit2}[28; 97.5\%] = +2.048$

$$t = \frac{0.6286\sqrt{28}}{\sqrt{1-0.6286^2}} = 4.2769 > t_{crit\,2}$$

→ Correlation statistically significant

Correlations (JMP)

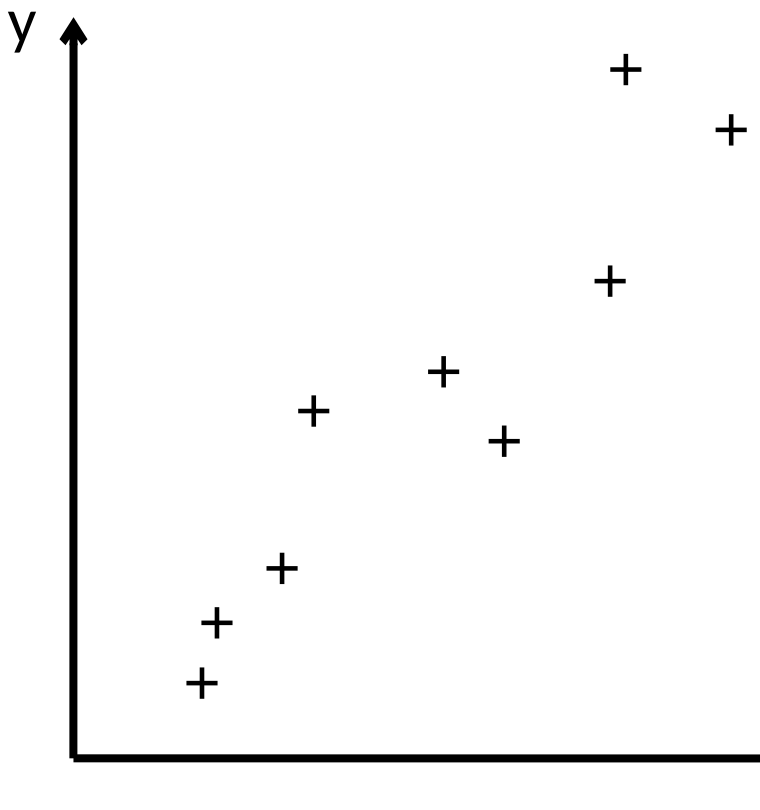
In JMP: Analyze-> Multivariate Methods-> Multivariate (like before)



Shows confidence interval and p-value.

Example of Correlations

Positive correlation
r > 0



e.g.

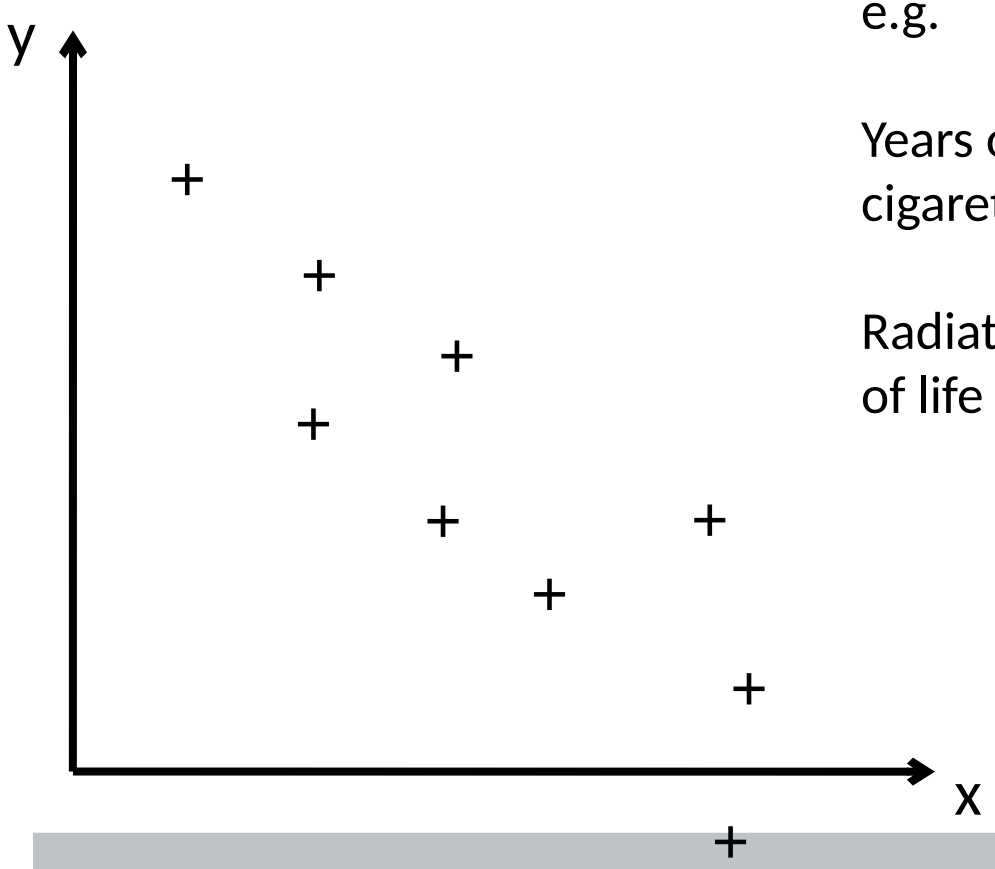
Amount studied ~ Exam score

Income parents ~ Income children

BMI ~ Blood pressure

Example of Correlations

Negative correlation
$r < 0$



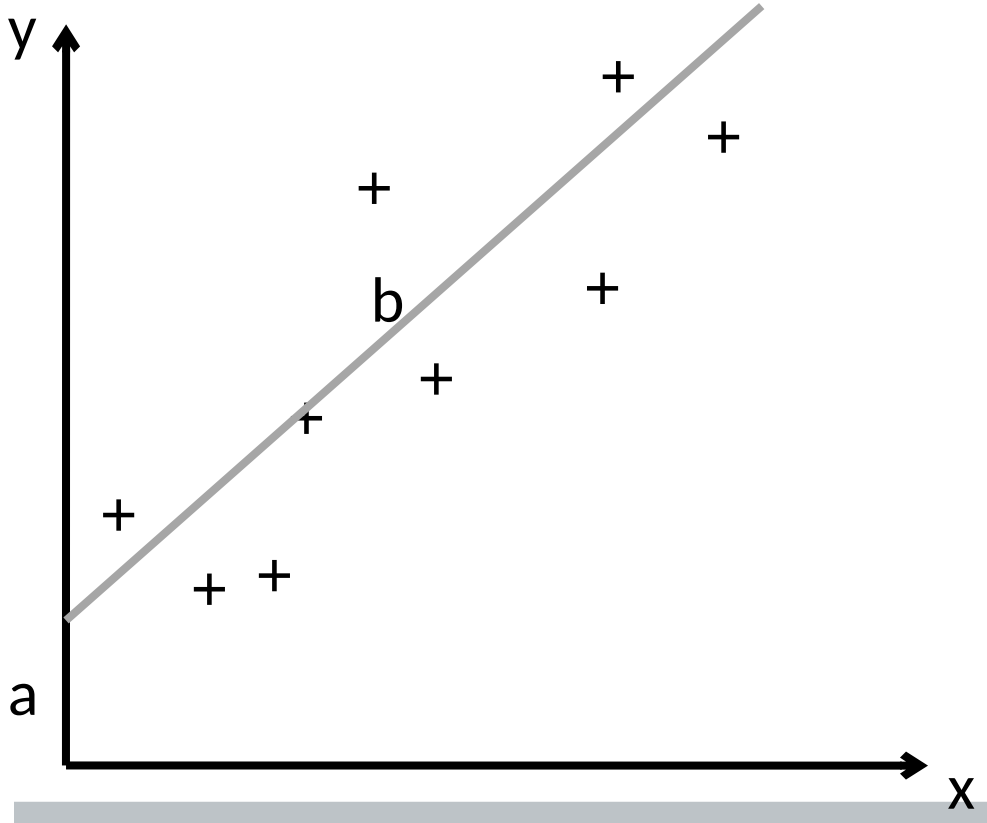
e.g.

Years of education ~ Numbers of cigarettes/day

Radiation dose ~ Remaining length of life

Linear Regression

Can we build a mathematical model for the relationship between x and y?



Linear model:

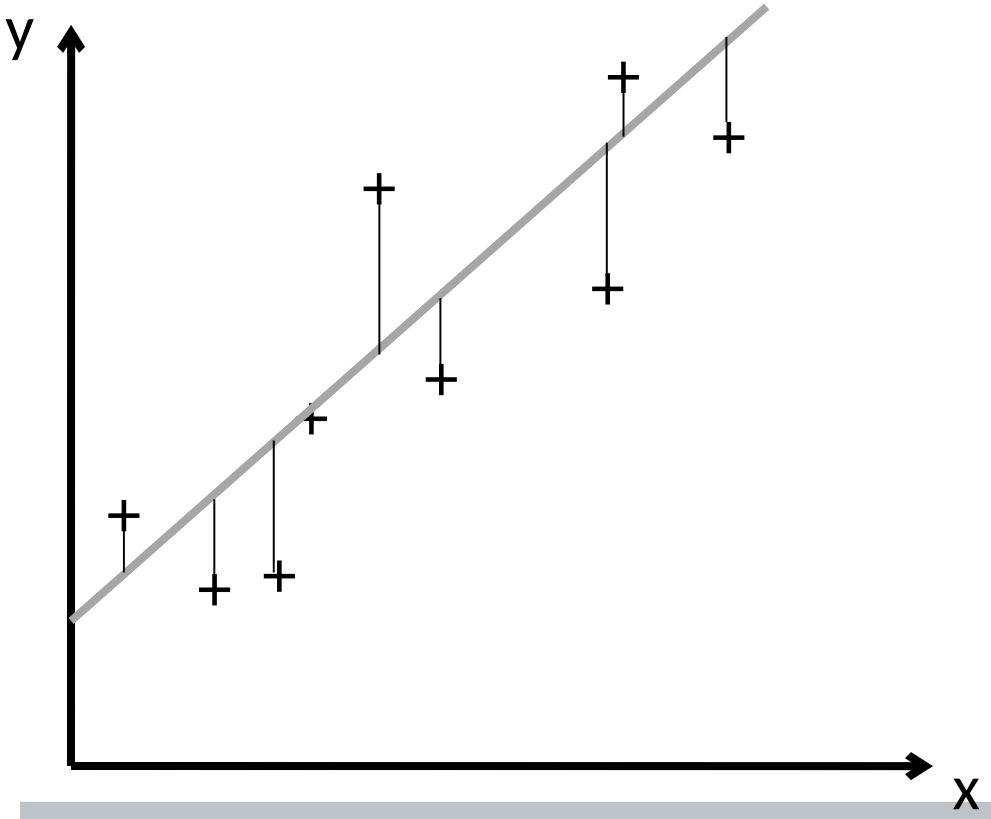
$y = a + bx$

a: intercept
b: slope

Linear Regression

How can we get the parameters a and b?

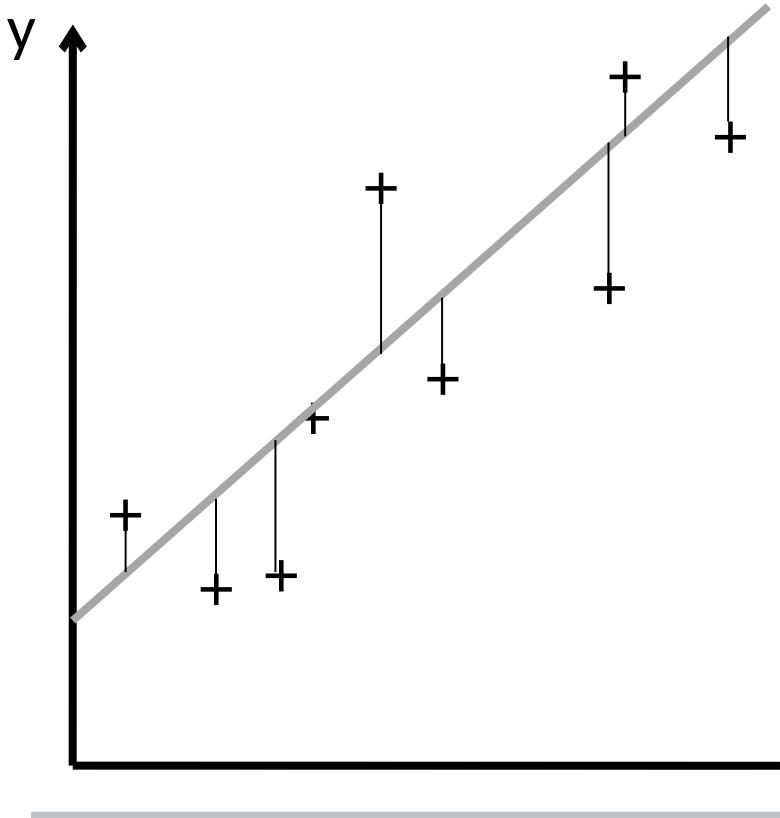
Method of least squares-> Minimizing the sum of squared deviations from the line.



Minimize:

$$s^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Linear Regression



We find:

$$a = \overline{y} - b\,\overline{x}$$

$$b = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$b = \frac{cov(x, y)}{s_x^2}$$

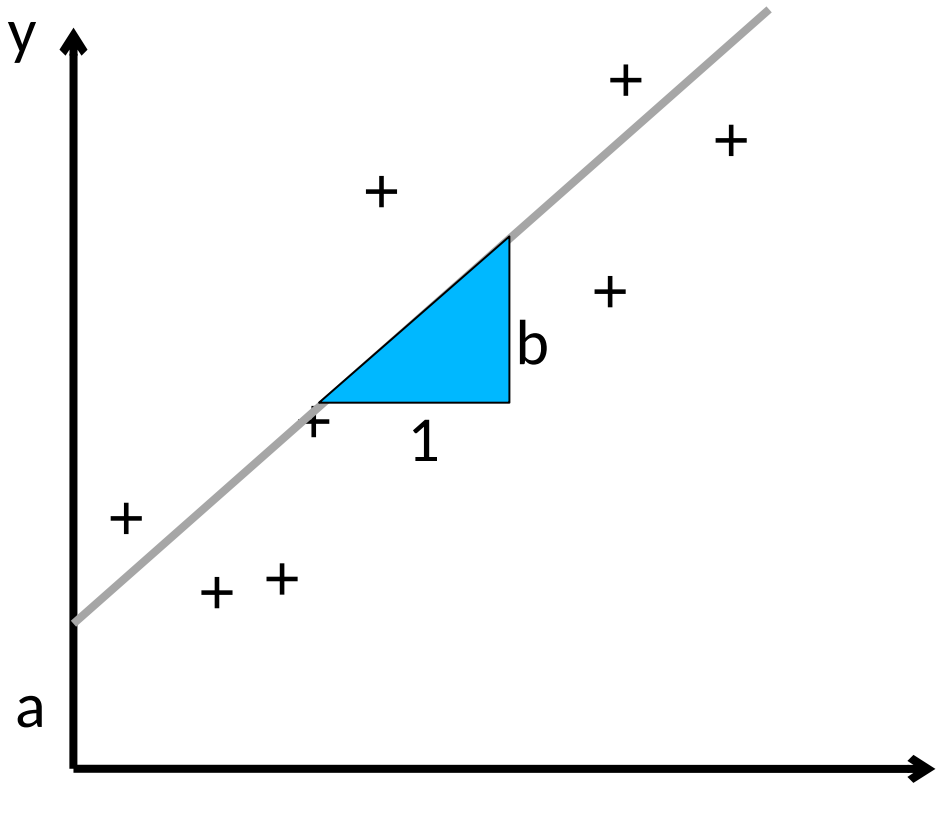
We also find:

$$b = r\,\frac{s_y}{s_x}$$

because:

$$r = \frac{cov(x, y)}{s_x\, s_y}$$

Linear Regression



$y_i = a + bx_i + e_i$

a: intercept b: slope

or:

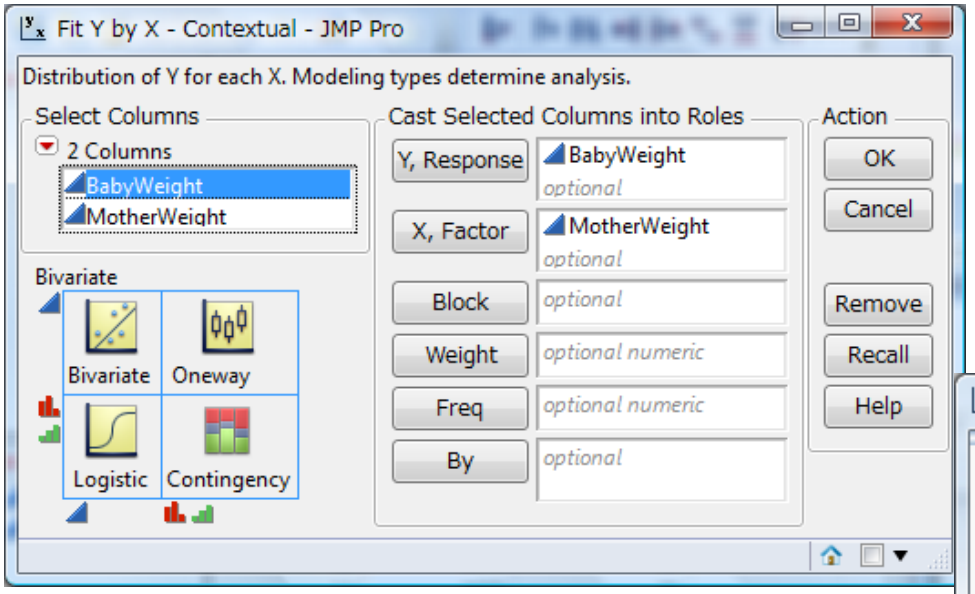
$y_i = b_0 + b_1 x_i + e_i$

b_0: intercept b_1: slope

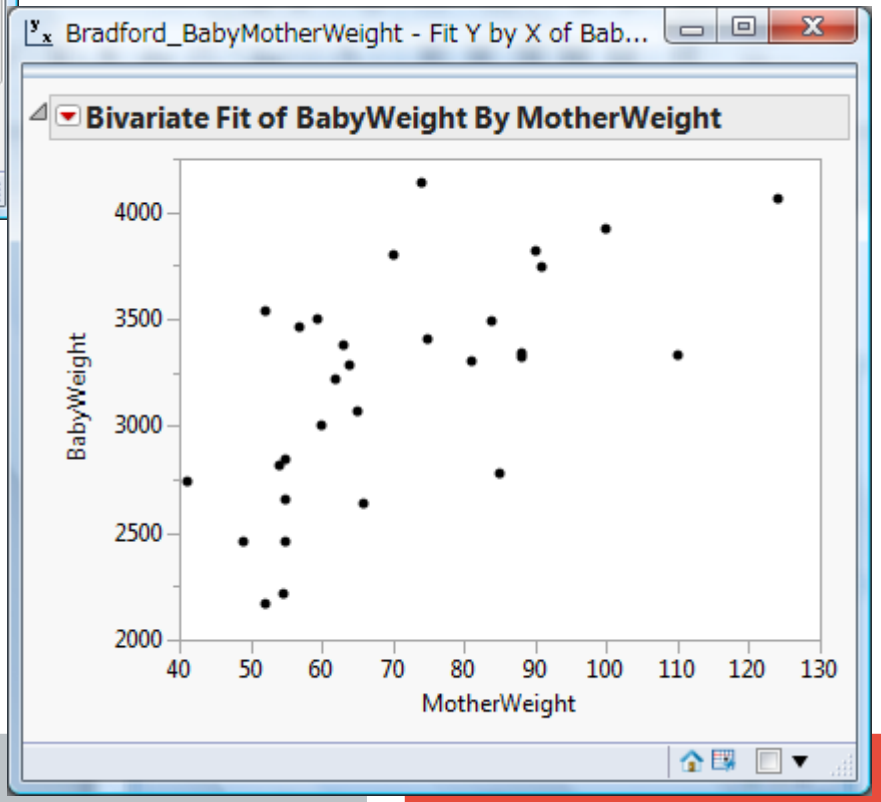
e_i: residual term

(not explained by model)

Linear Regression - JMP



In JMP: Analyze-> Fit Y by x



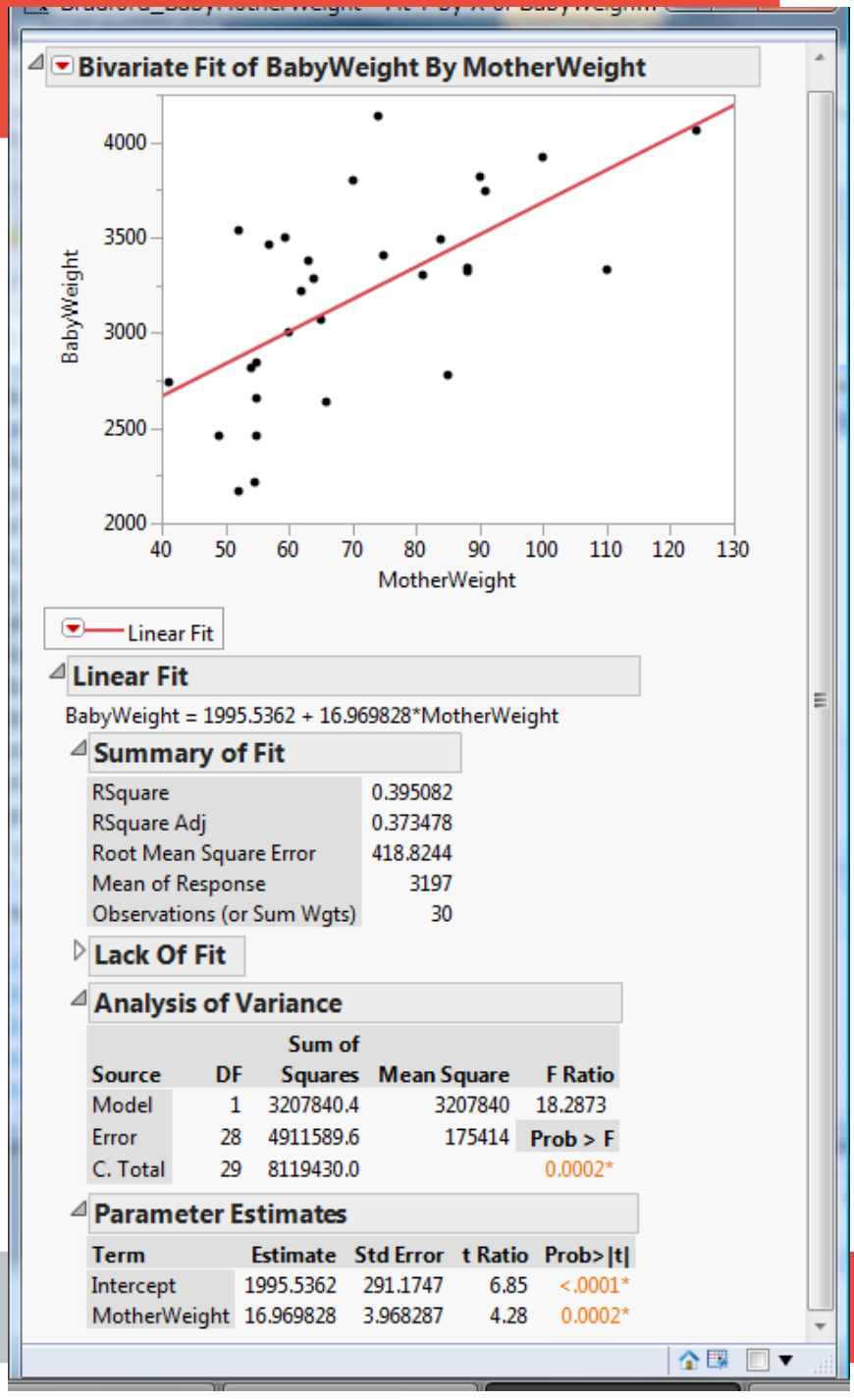
Linear Regression - JMP

Red Triangle -> Fit Line

Equation:

$y = 1996+16.97x$

Significance tests for estimated parameters



Linear Regression - JMP

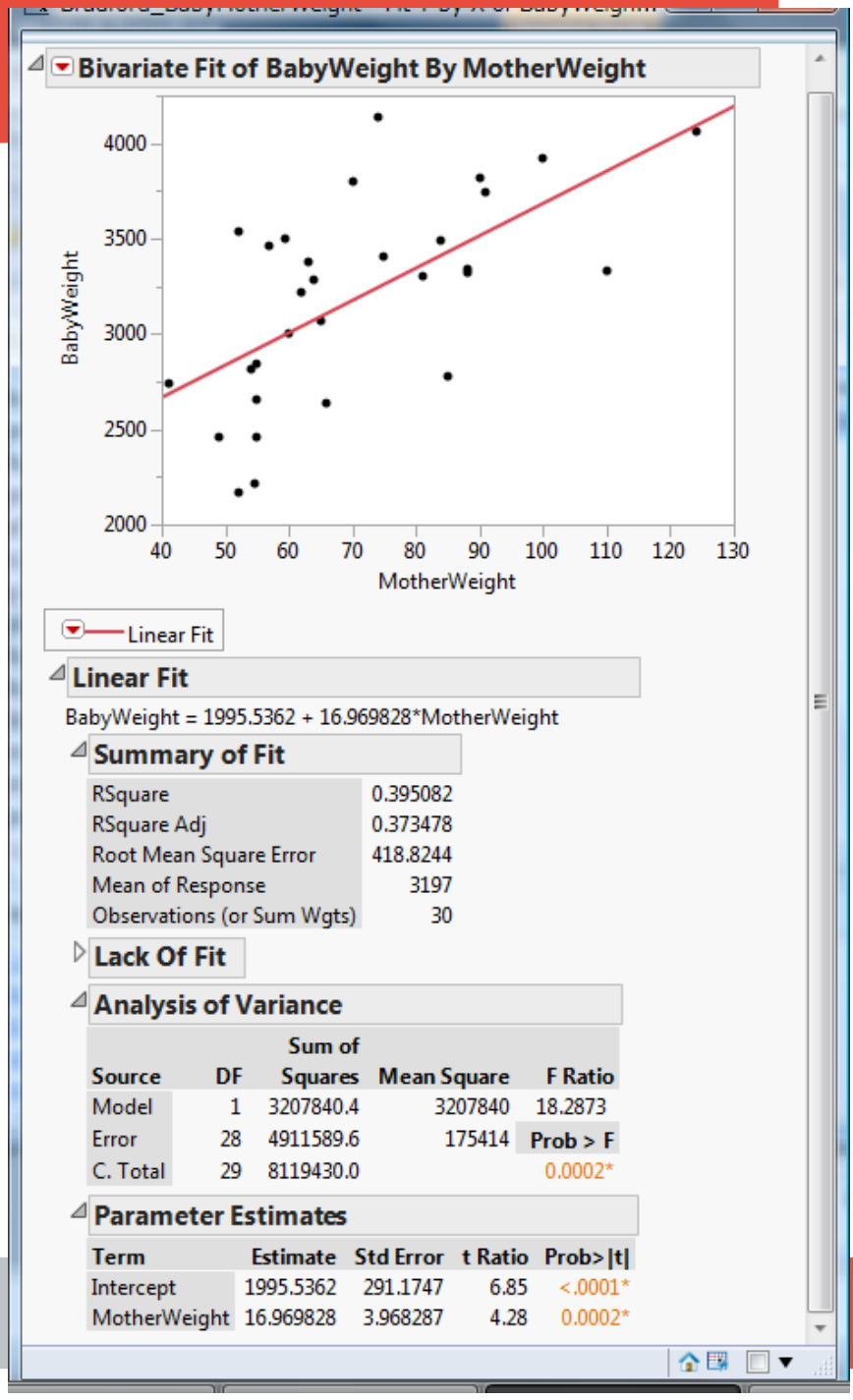
Testing the estimated parameters against the null hypothesis $b_i = 0$

b_0: intercept

b_1: slope

$$t = \frac{b_i}{SE_{b_i}}$$

is t-distributed with n-2 df.



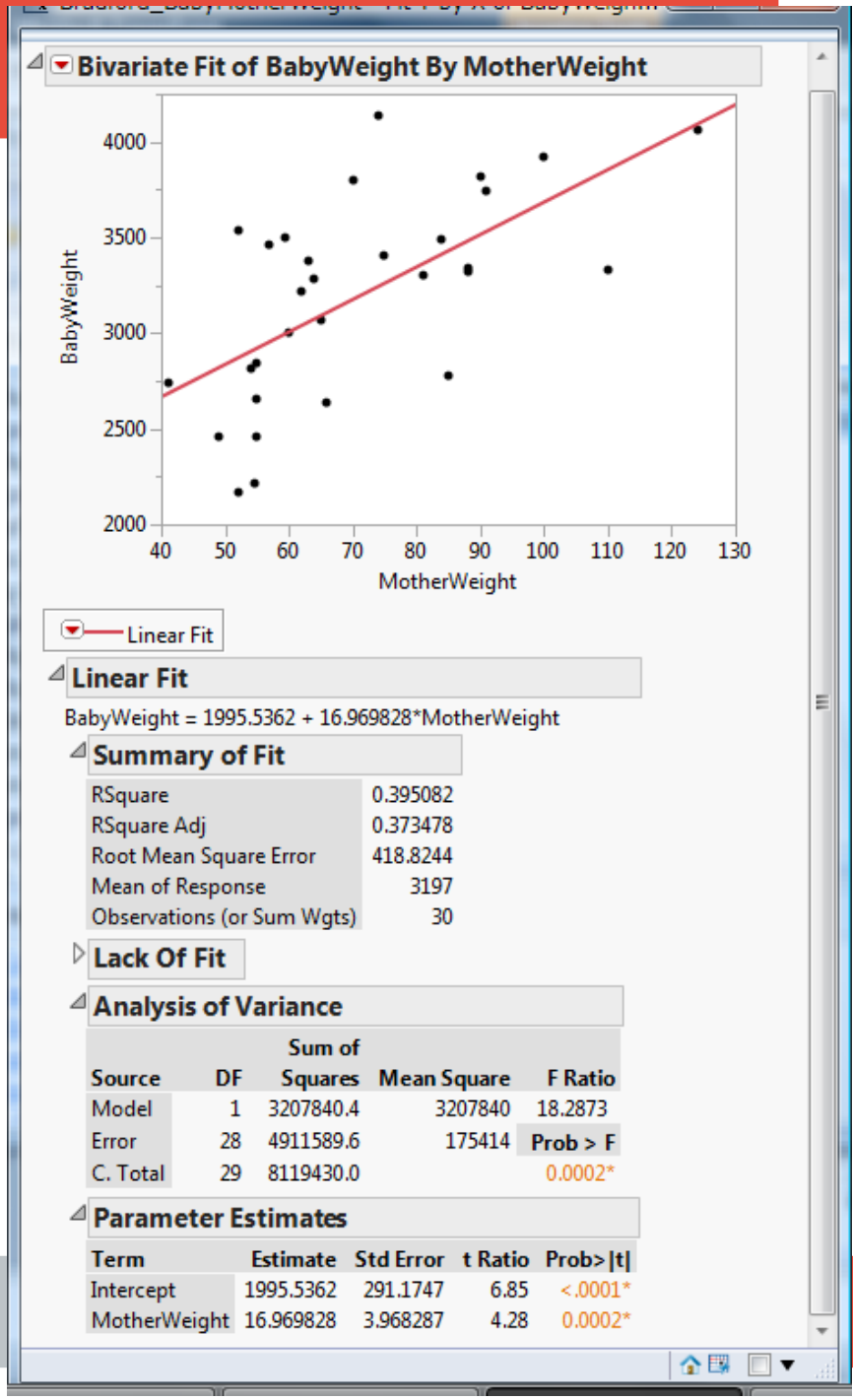
Linear Regression - JMP

RSquare (variance explained by model relative to total variance):

$$R^2 = \frac{SS_M}{SS_T} =$$

$$\frac{\sum_{i=1}^{n} (a + bx_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = r^2$$

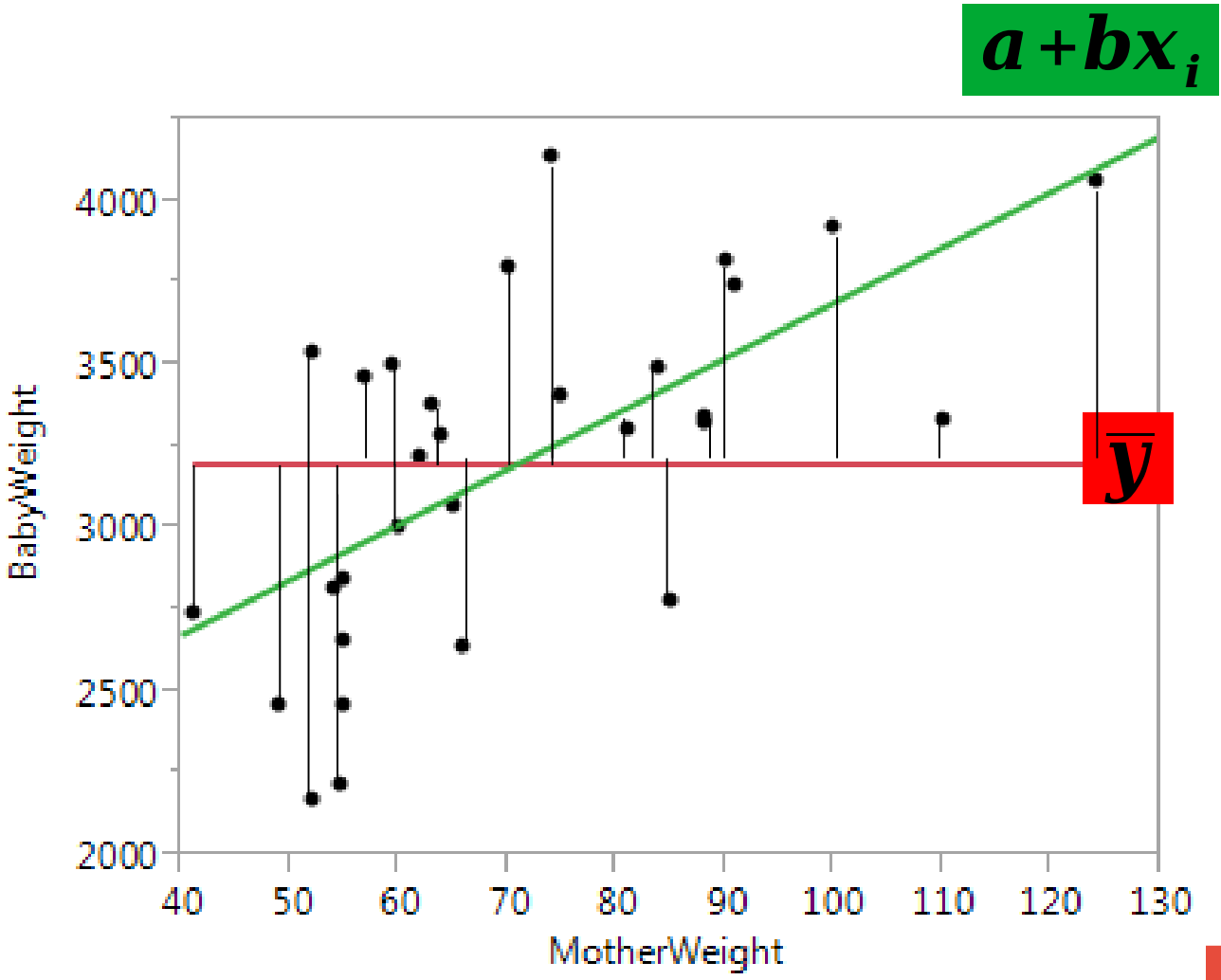
r as obtained from correlation



Testing the model – Linear Regression

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

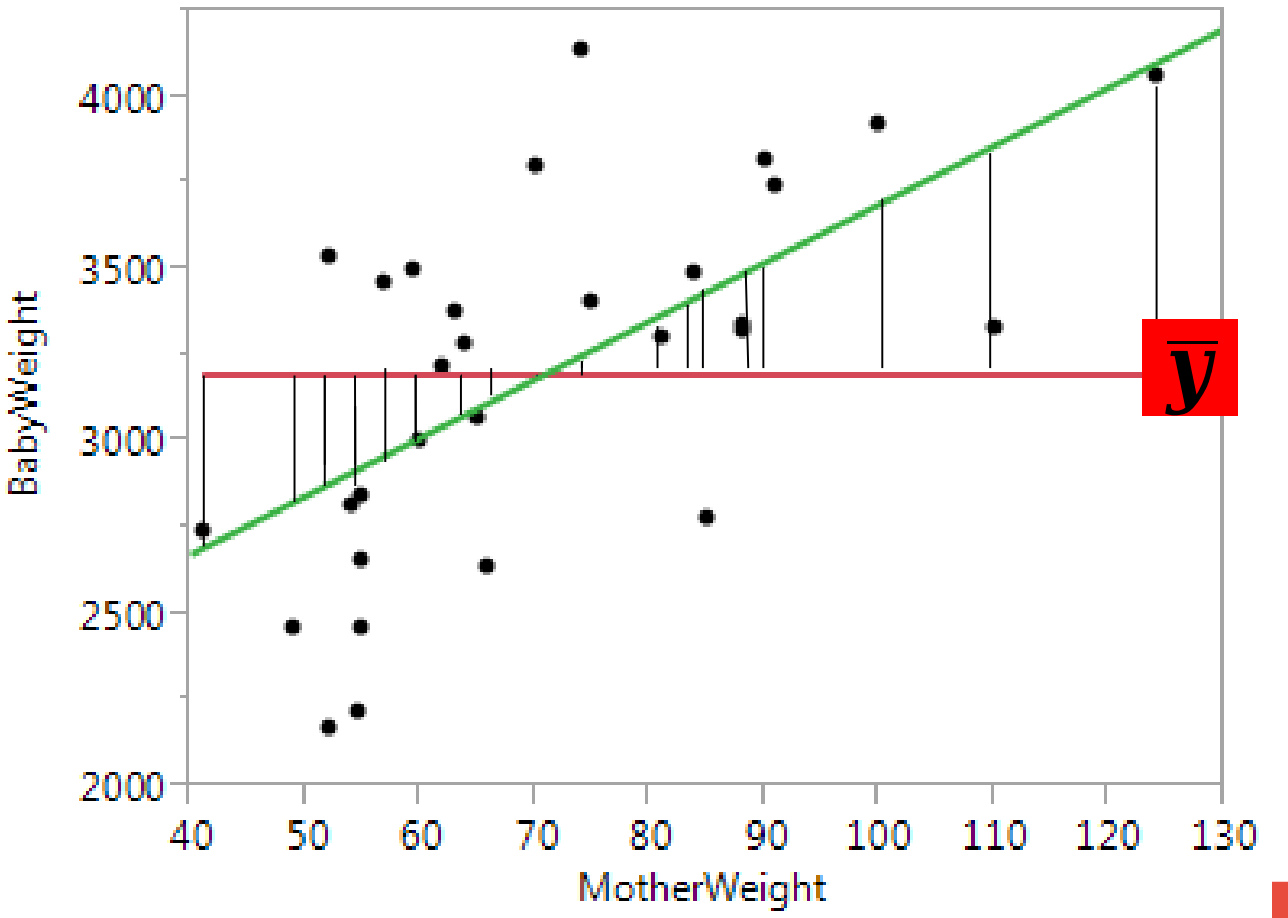
Total sum of squares



Testing the model – Linear Regression

$$SS_M = \sum_{i=1}^{n} (a + bx_i - \bar{y})^2$$

Model sum of squares



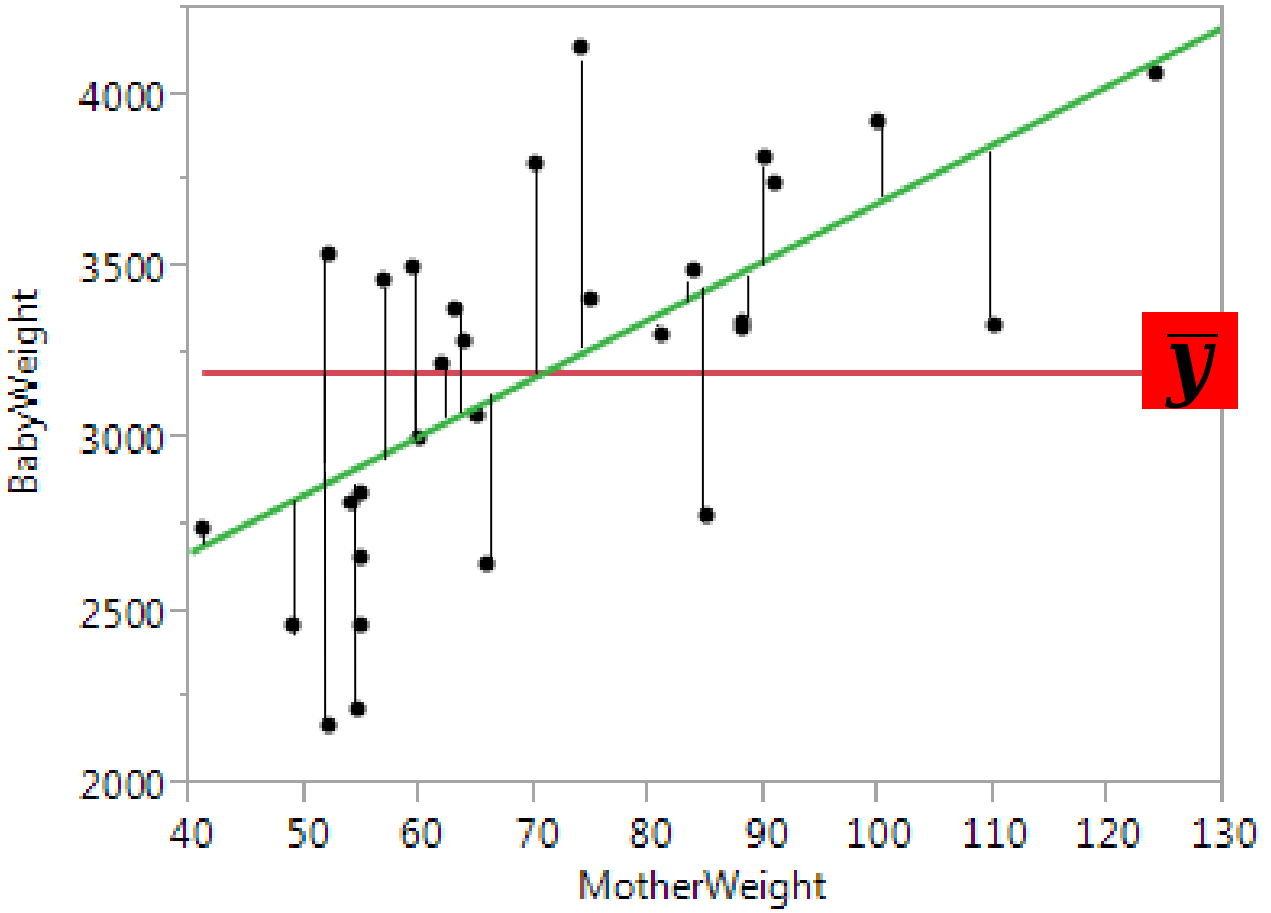
Testing the model – Linear Regression

$$SS_T = SS_M + SS_R$$

$$SS_R = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$a + bx_i$

Residual sum of squares



F-test

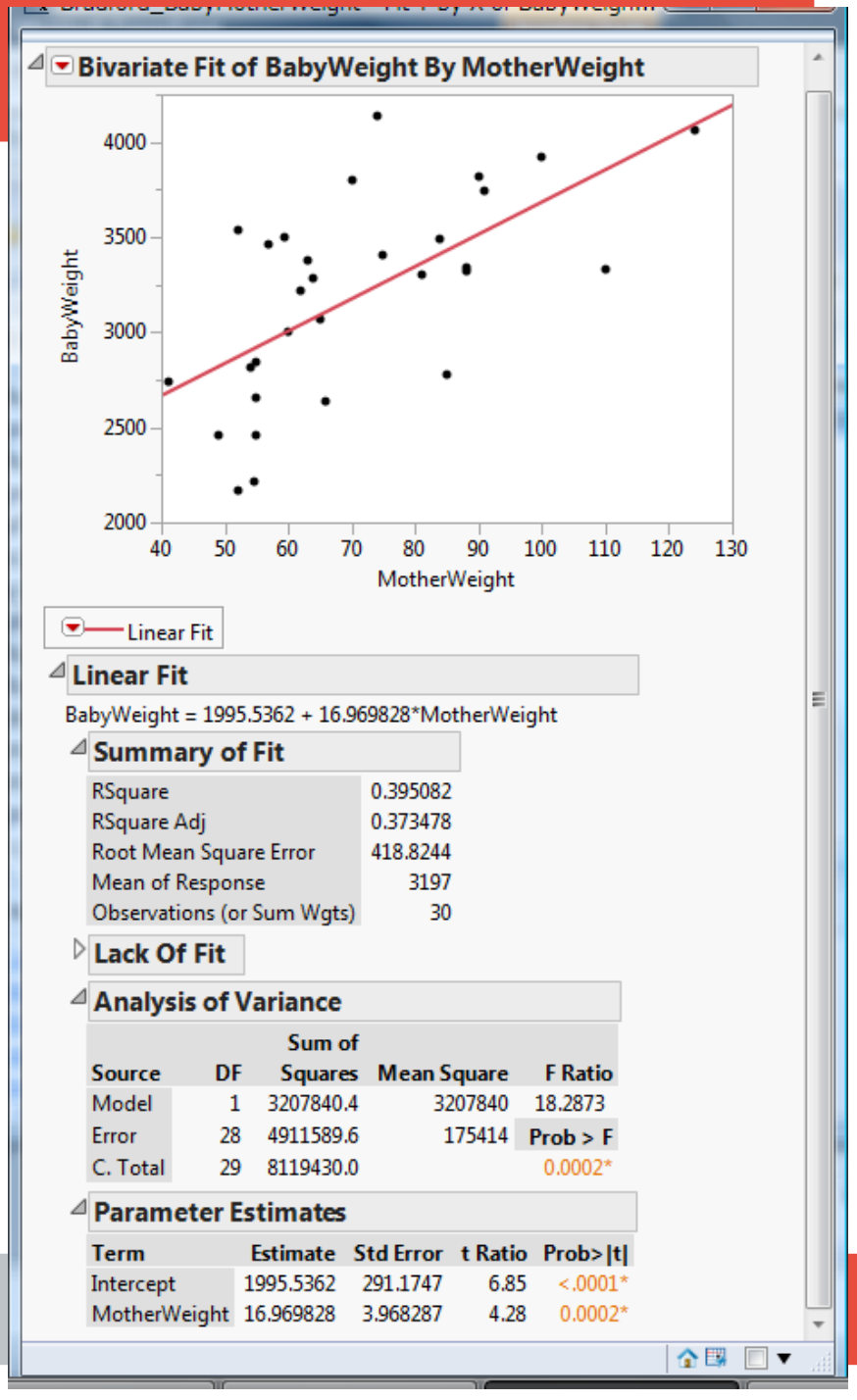
Significance test for the whole model:

$$F = \frac{MS_M}{MS_R} = \frac{SS_M / df_M}{SS_R / df_R}$$

$$\frac{3207840.1/1}{4911589.6/28} = 18.2873$$

df_M: *model degrees of freedom = number of x-variables = 1*

df_R: *error degrees of freedom = number of observations – number of estimated parameters (a,b) = 30-2*

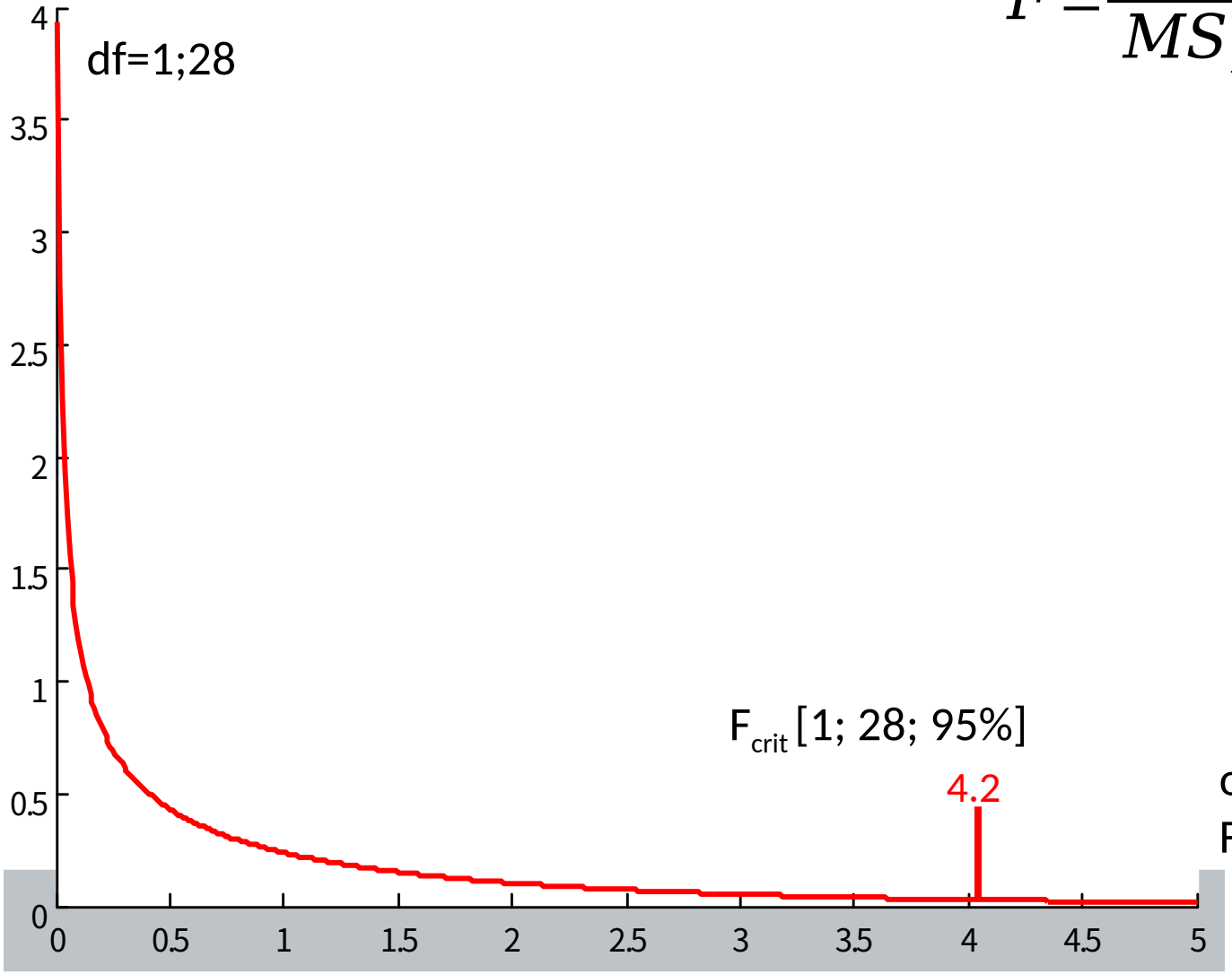


F-Test

Under the Null hypothesis (that b=0):

$P(F < F_{crit}) = \alpha = 0.05$

$$F = \frac{MS_M}{MS_R} = \frac{SS_M/df_M}{SS_R/df_R}$$

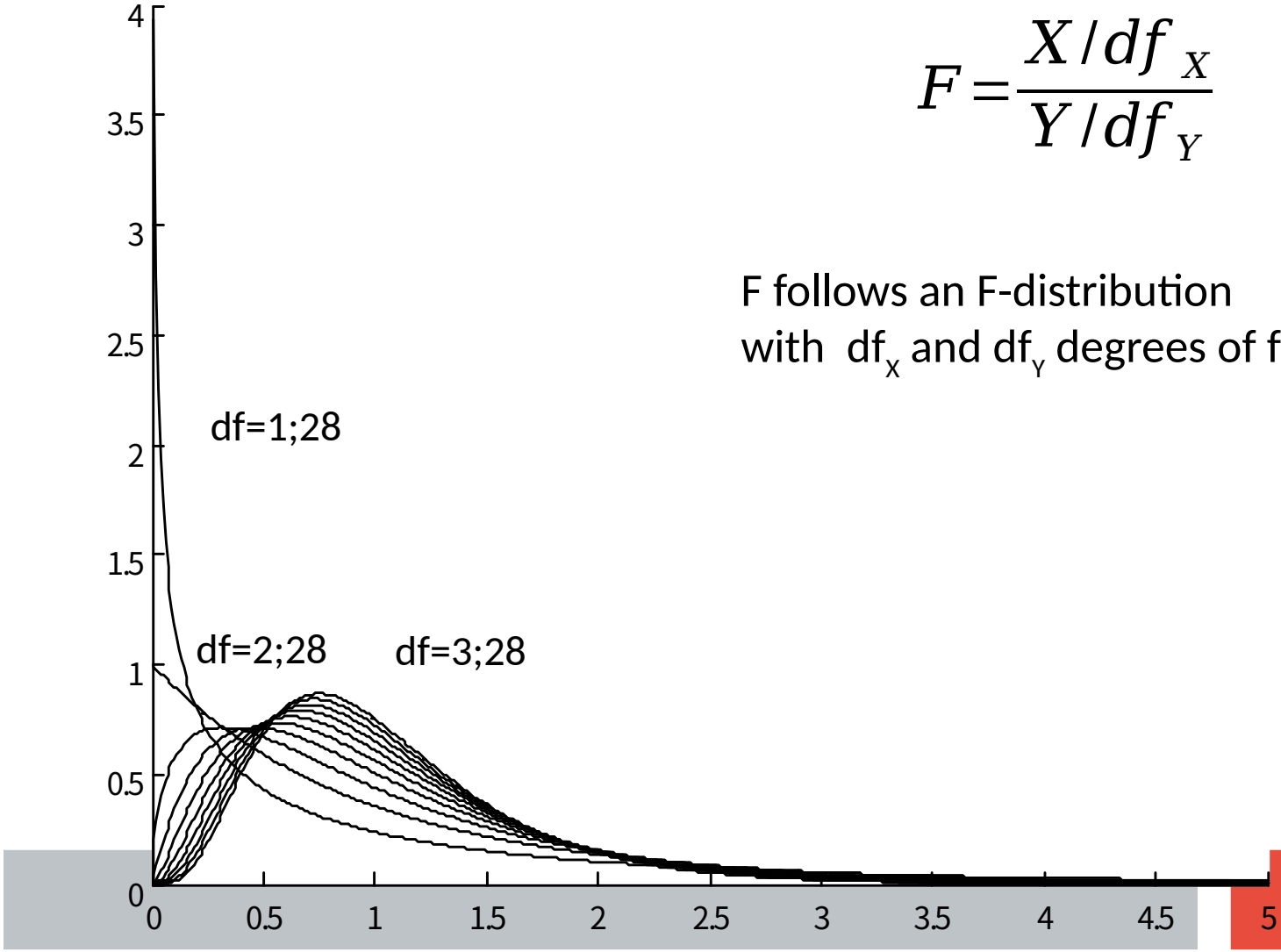


F-distribution

If X and Y are independent, χ^2-distributed random variables with df_X and df_Y degrees of freedom:

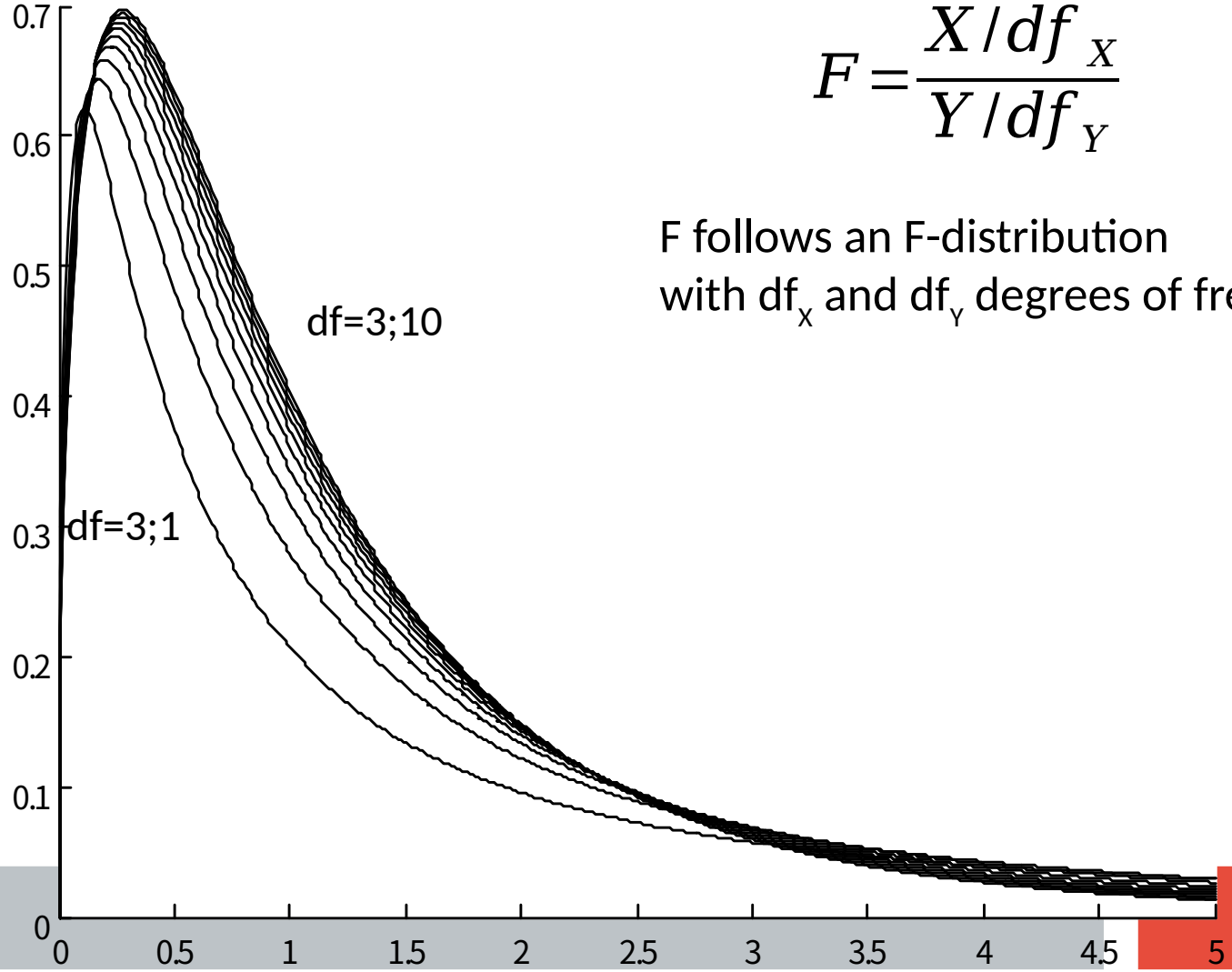
$$F = \frac{X/df_X}{Y/df_Y}$$

F follows an F-distribution with df_X and df_Y degrees of freedom.



F-distribution

If X and Y are independent, χ^2-distributed random variables with df_X and df_Y degrees of freedom:

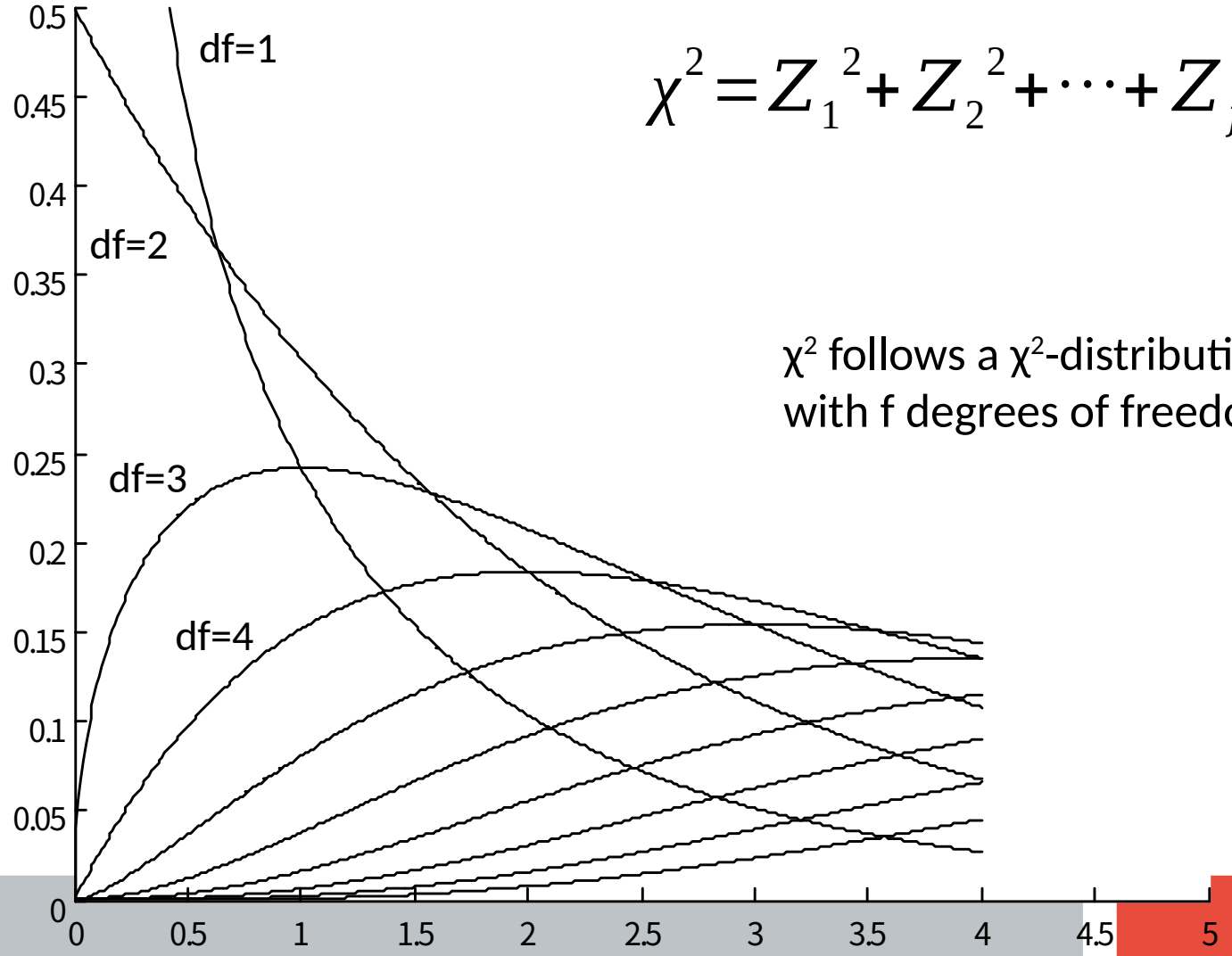


$$F = \frac{X/df_X}{Y/df_Y}$$

F follows an F-distribution with df_X and df_Y degrees of freedom.

χ^2-distribution

If Z_1, $Z_{2,\ldots}$ Z_f are independent standard normal random variables:



$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_f^2$$

χ^2 follows a χ^2-distribution with f degrees of freedom.

Summary about χ^2 , F, and t-distribution

If Z_1, $Z_{2,...}$ Z_f are independent standard normal random variables:

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_f^2$$

χ^2 follows a χ^2-**distribution** with f degrees of freedom.

If X and Y are independent, χ^2-distributed random variables with and degrees of freedom:

$$F = \frac{X / df_X}{Y / df_Y}$$

F follows an **F-distribution** with df_X and df_Y degrees of freedom.

If A and B are independent, A is a standard normal random variable, B is a χ^2-distributed random variable with f degrees of freedom:

$$t = \frac{A}{\sqrt{B / f}}$$

t follows at **t-distribution** with f degrees of freedom.

Linear Regression

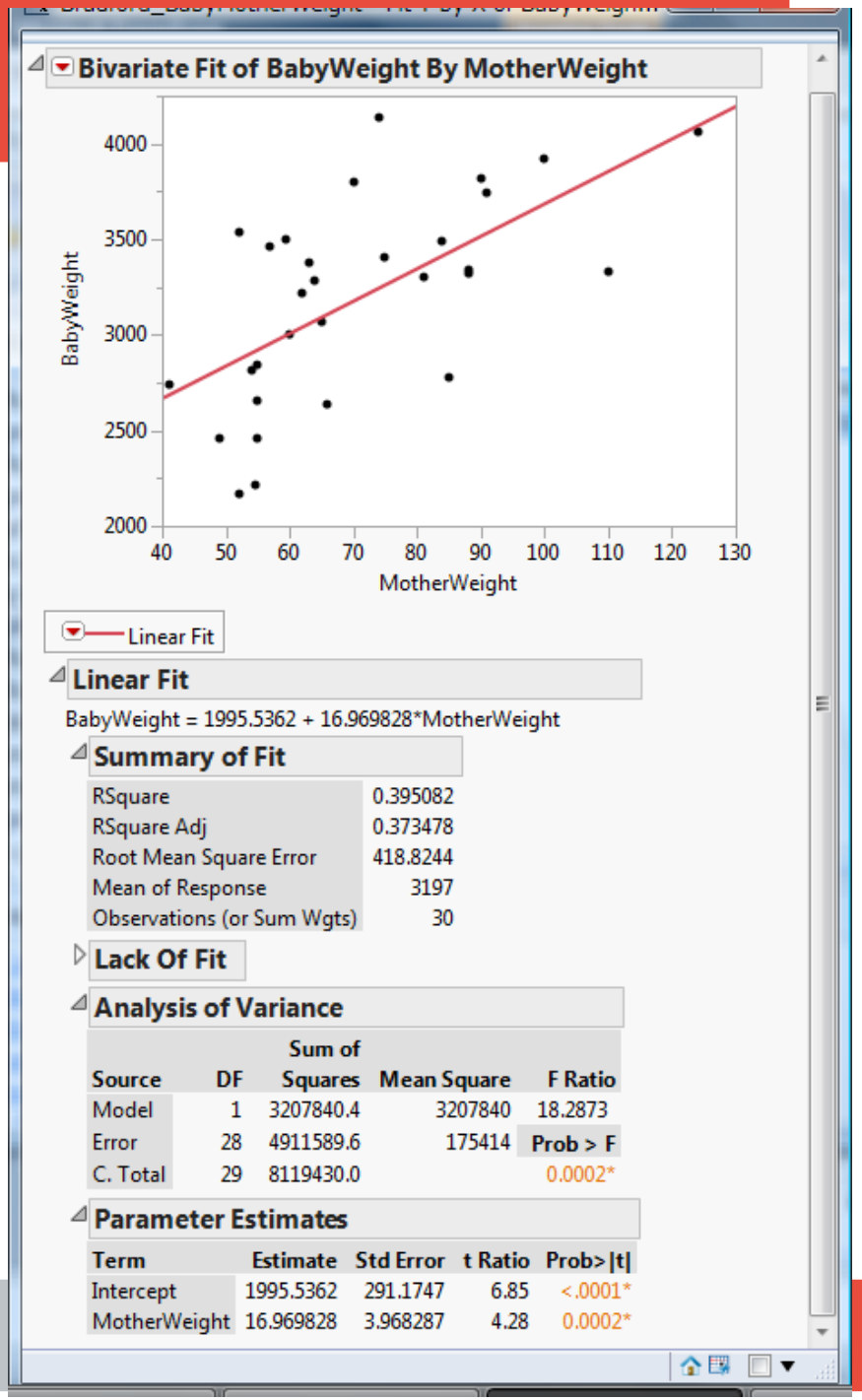
RSquare Adj:
like Rsquare, but penalizes increasing the number of variables.

$$R^2(adjusted) = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

n: number of observations
k: number of variables (not including the intercept)

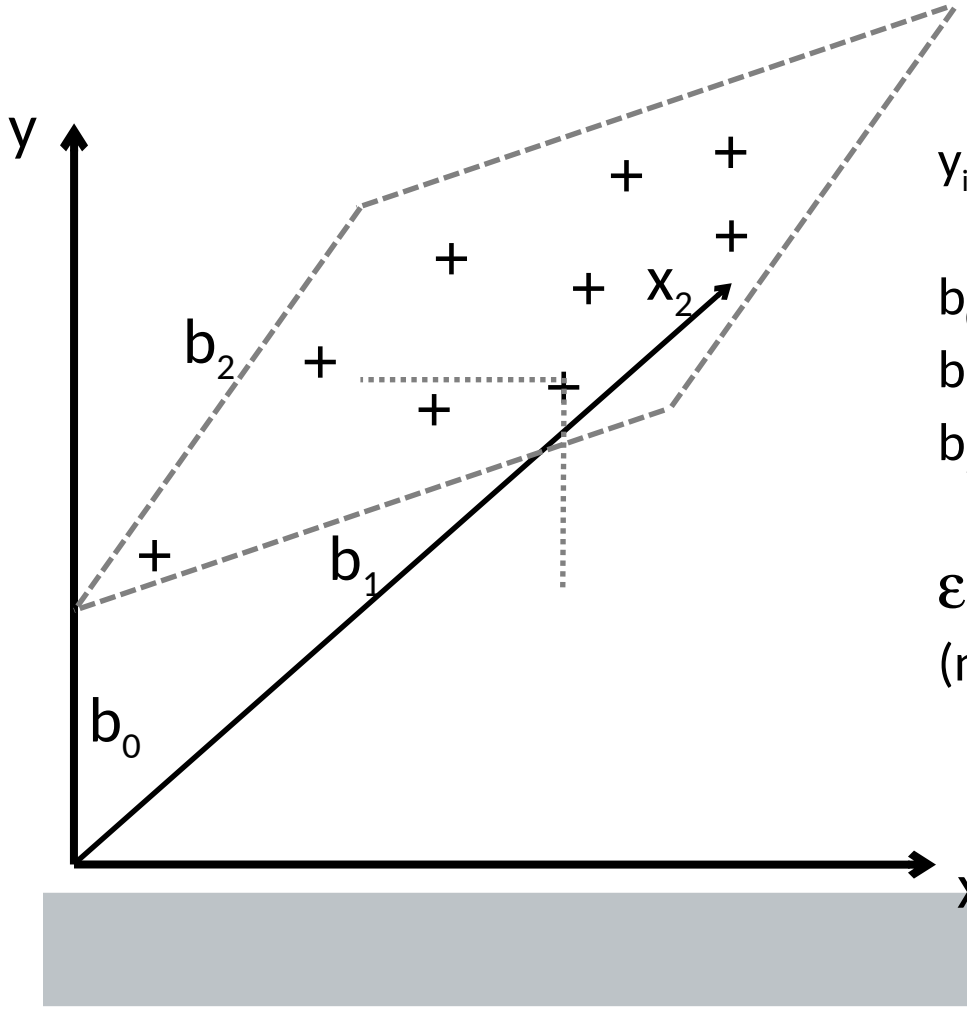
here:

$$R^2(adjusted) = \frac{(1-0.395082)(30-1)}{30-1-1}$$



Multiple Linear Regression

We can use regression to test the influence of more than one variable on an outcome -> multiple regression.



$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + \varepsilon_i$

b_0: intercept

b_1: slope

b_2 : slope

ε_i: residual term

(not explained by model)

Multiple Linear Regression

Example:

Lelong et al., Am J Hypertension, 2015

This study tested the Influence of various lifestyle variables on blood pressure. In particular, sodium (salt) has been previously reported to influence blood pressure and thus increase the risk of cardiovascular disease.
8670 French participants

Multiple Linear Regression – Example

Table 2. Age-adjusted associations between systolic blood pressure and lifestyle recommendations in women and men

Lifestyle recommendation parameter	Systolic blood pressure	
	β (standard error)	*P* value
Women		
Salt (g/day)	0.0001 (0.0001)	0.08
Potassium intake (mg/day)	−0.001 (0.0002)	0.01
Sodium-to-potassium ratio	2.3 (0.61)	0.0001
Alcohol intake (g/day)	0.058 (0.016)	0.0004
BMI (kg/m^2)	0.88 (0.04)	<0.0001
Fruits and vegetables intake (g/day)	−0.004 (0.0008)	<0.0001
Physical activity		0.17
Medium[a]	−0.59 (0.47)	0.21
High[a]	0.10 (0.50)	0.84
Sedentary (min/day)	−0.001 (0.001)	0.53
Men		
Salt (g/day)	0.0003 (0.0001)	0.01
Potassium intake (mg/day)	0.0002 (0.0003)	0.64
Sodium-to-potassium ratio	2.1 (1.03)	0.05
Alcohol intake (g/day)	0.051 (0.018)	0.006
BMI (kg/m^2)	1 (0.90)	<0.0001
Fruits and vegetables intake (g/day)	−0.0030 (0.0012)	0.01
Physical activity		0.31
Medium[a]	−1.19 (0.93)	0.20
High[a]	−0.28 (0.91)	0.75
Sedentary (min/day)	−0.002 (0.002)	0.18

Abbreviation: BMI, body mass index.
[a]Reference: physical activity = low.

Lelong, Am J Hypertension, 2014

Multiple Linear Regression – Example

Table 4. Multivariate association between lifestyle factors and systolic blood pressure in women and men

	Adjusted R^2 (%)	*P* value	
Model	**19.7**	**<0.0001**	
Parameters	**Partial r^2 (%)**	***β* (standard error)**	***P* value**
Women			
Age	8.6	0.33 (0.01)	<0.0001
BMI	6.8	0.87 (0.04)	<0.0001
Alcohol intake	0.2	0.05 (0.02)	0.002
Physical activity	0.3		0.008
Physical activity medium vs. low		0.40 (0.45)	0.38
Physical activity high vs. low		1.37 (0.49)	0.01
Salt intake	–	–0.00001 (0.0001)	0.86
Fruits and vegetables intake (g/day)	0.3	–0.003 (0.001)	<0.0001
Education level	0.3		<0.0001
Secondary vs. primary		–2.12 (1.09)	0.05
University vs. primary		–3.73 (1.08)	0.001
Men			
Age	4.1	0.22 (0.02)	<0.0001
BMI	5.1	0.97 (0.09)	<0.0001
Alcohol intake	–	0.02 (0.02)	0.32
Physical activity	–		0.50
Physical activity medium vs. low		–0.02 (0.91)	0.98
Physical activity high vs. low		0.71 (0.89)	0.42
Salt intake	–	0. 0001 (0.0001)	0.41
Fruits and vegetables intake (g/day)	–	–0.001 (0.001)	0.44
Education level	0.3		0.01
Secondary vs. primary		4.71 (1.74)	0.01
University vs. primary		3.34 (1.72)	0.05

Abbreviaton: BMI, body mass index.

Lelong, Am J Hypertension, 2014

Multiple Linear Regression – Example

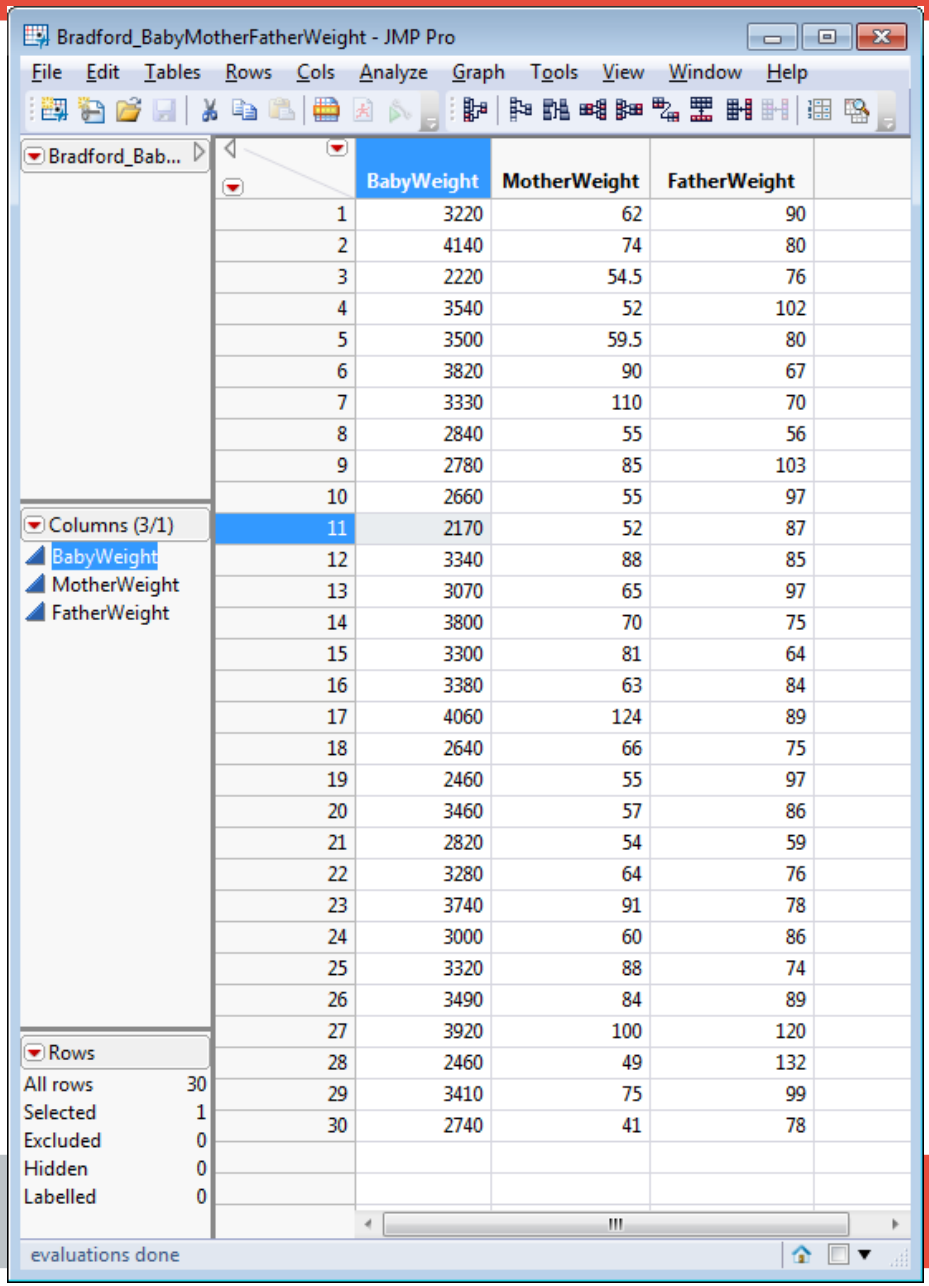
The main result of our study was that, aside from age, BMI had the strongest association with BP level; an SBP increase of almost 20 mm Hg across the BMI categories was shown. Furthermore, after multiple adjustments, BMI persisted as the main contributory modifiable factor in the SBP multivariate model.

Salt. Despite the widespread knowledge that salt consumption is a major nutritional factor associated with BP level,[21,22] we found that age-adjusted SBP was not significantly associated with salt intake in women and not associated with either sex when adjusted for all parameters. Nevertheless, hypertensive participants consumed salt at a significantly higher rate than did those who were nonhypertensive. Several hypotheses can explain this finding. First, in our study, people who were treated with antihypertensive drugs and who could have a greater salt consumption were excluded. This may have led to minimization of the salt consumption range and, thus, the expected association. Second, dietary questionnaires covered only sodium intake from foods; total consumption was estimated by adding a constant for cooking meals and eating, which could have reduced the interindividual variation and, hence, made the potential association more difficult to identify.

Lelong et al., Am J Hypertension, 2015

Multiple Linear Regression – JMP

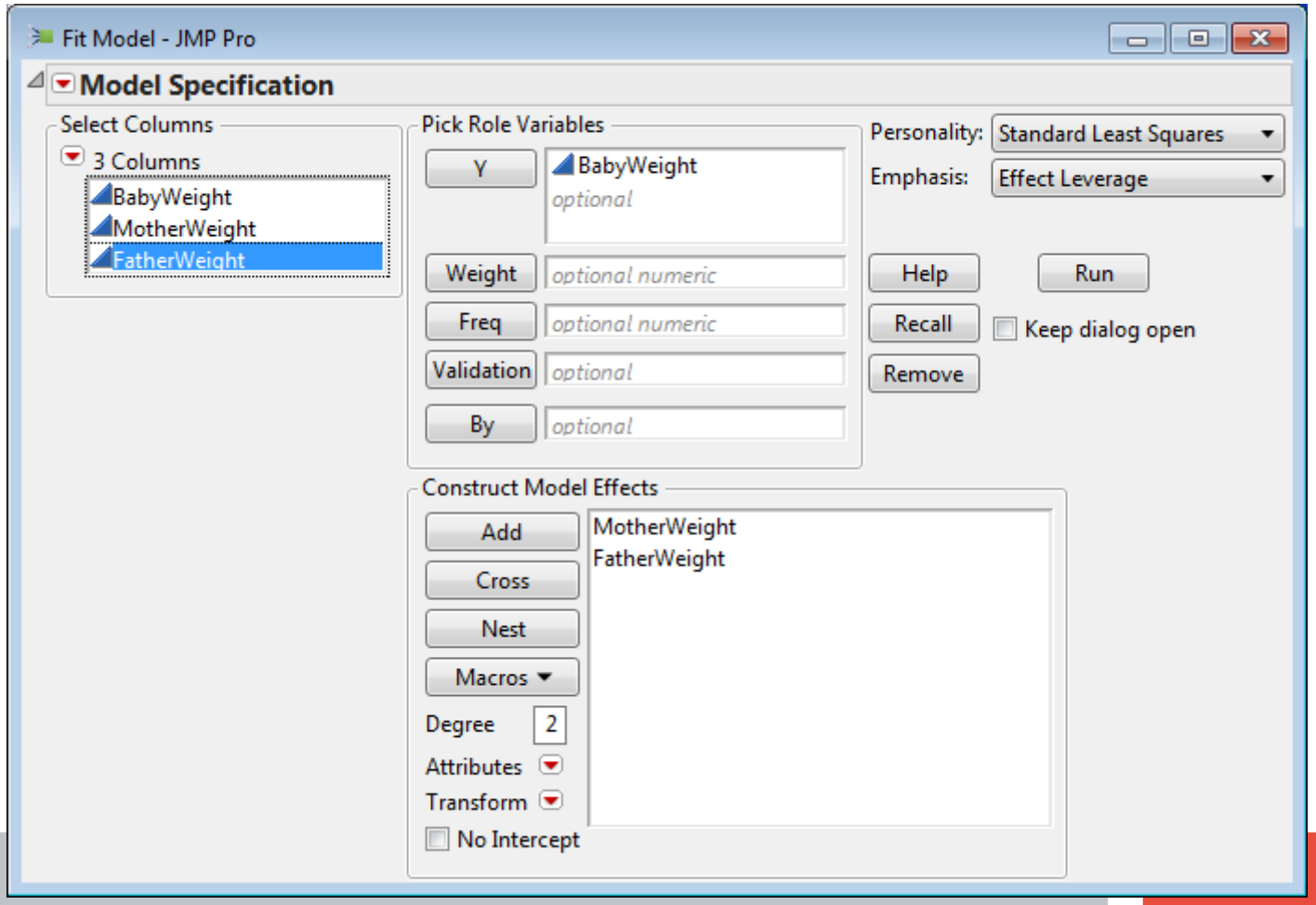
In JMP: Analyze-> Fit Model



	BabyWeight	MotherWeight	FatherWeight
1	3220	62	90
2	4140	74	80
3	2220	54.5	76
4	3540	52	102
5	3500	59.5	80
6	3820	90	67
7	3330	110	70
8	2840	55	56
9	2780	85	103
10	2660	55	97
11	2170	52	87
12	3340	88	85
13	3070	65	97
14	3800	70	75
15	3300	81	64
16	3380	63	84
17	4060	124	89
18	2640	66	75
19	2460	55	97
20	3460	57	86
21	2820	54	59
22	3280	64	76
23	3740	91	78
24	3000	60	86
25	3320	88	74
26	3490	84	89
27	3920	100	120
28	2460	49	132
29	3410	75	99
30	2740	41	78

Multiple Linear Regression – JMP

Y: variable to be explained by the multiple variables x_1, x_2, …, x_i in "Construct Model Effects" -> Run



Multiple Linear Regression – JMP

RSquare

Omnibus test for the model (null hypothesis: all slope parameters $b_1 = b_2 = 0$)

Tests for the single parameters

Bradford_BabyMotherFatherWeight - Fi...

Response BabyWeight

Summary of Fit

RSquare	0.39733
RSquare Adj	0.352688
Root Mean Square Error	425.7168
Mean of Response	3197
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	3226091.5	1613046	8.9003
Error	27	4893338.5	181235	Prob > F
C. Total	29	8119430.0		0.0011*

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	26	4873338.5	187436	9.3718
Pure Error	1	20000.0	20000	Prob > F
Total Error	27	4893338.5		0.2535
				Max RSq
				0.9975

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2128.5508	513.1164	4.15	0.0003*
MotherWeight	16.920625	4.036569	4.19	0.0003*
FatherWeight	-1.523297	4.800228	-0.32	0.7534

Effect Tests

Effect Details

Summary

- The correlation coefficient r quantifies a linear relationship between two metric variables in a range of -1 to 1 (0: no linear relationship).
- Statistical hypotheses about r can be tested using the t-distribution.
- In linear regression a relationship between two metric variables is modeled with:

$$y_i = b_0 + b_1 x_{1i} + \ldots. b_k x_{ki} + \varepsilon_i$$

- Statistical hypotheses about individual parameters b_j can be tested using the t-distribution, the whole model can be tested using the F-distribution.