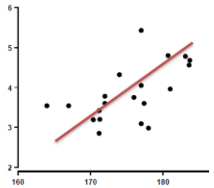
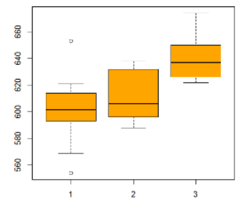
Review - Simple Linear Regression

A <u>continuous</u> predictor (x): $\hat{y} = b_0 + b_1x$



The **slope (b$_1$)** indicates the association between X and Y, and it is the change in y for every unit increase in x

A single <u>categorical</u> predictor with 3 levels, we need to set a reference level (group 1) and create two dummy variables:



$$\hat{y} = b_0 + b_1z_i + b_2z_2$$

b_1 is difference between group 2 and the reference group 1

b_2 is difference between group 3 and the reference group 1

Review- Regression Coefficients for a Categorical Variable

- A categorical variable with *l* levels needs *l*-1 dummy variables to represent it

If *l*=3, we need to create two dummy variables: z_1 and z_2

$$\hat{y} = b_0 + b_1 z_i + b_2 z_2$$

z1	z2
0	0
0	0
0	0
0	0
1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1

b_0 *control (reference group)*

$b_0 + b_1$ *low jump grp*

$b_0 + b_2$ *High jump grp*

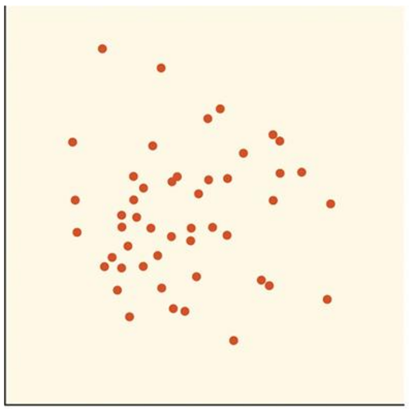
b_1 is difference between low jump group and control

b_2 is difference between high jump group and control

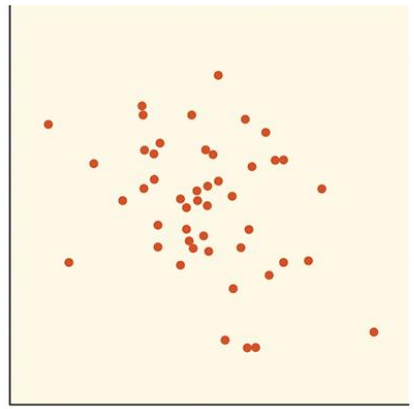
Pearson and Spearman correlation

Correlation Coefficient

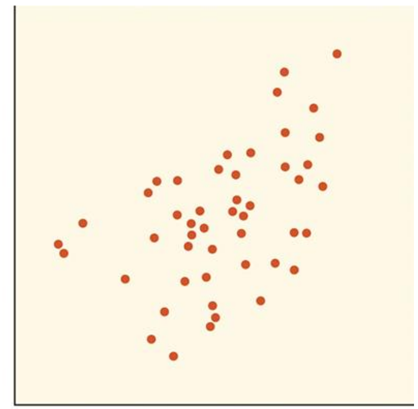
We say a linear relationship is strong if the points lie close to a straight line and weak if they are widely scattered about a line.



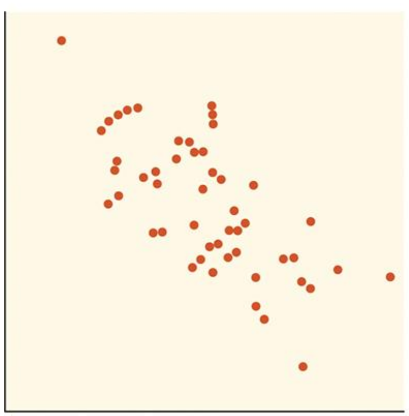
Correlation $r = 0$



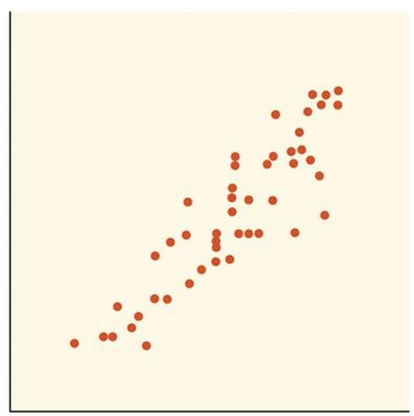
Correlation $r = -0.3$



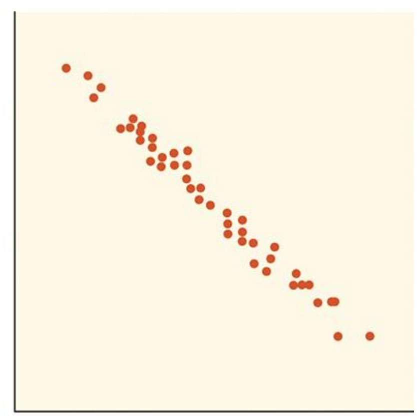
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$

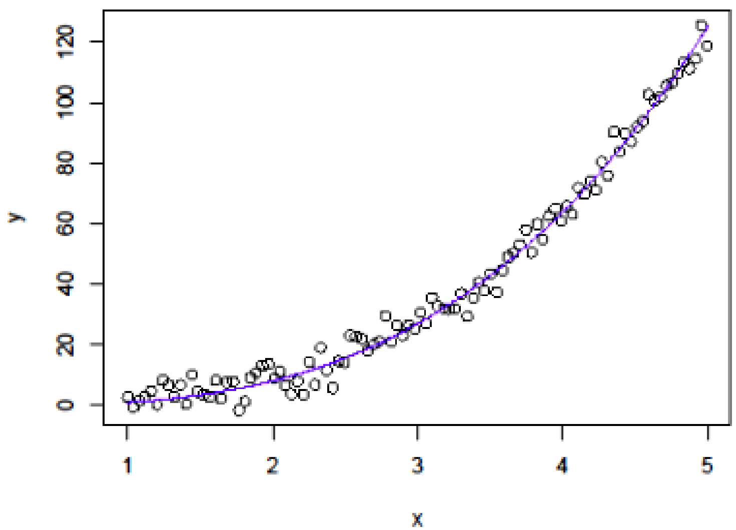


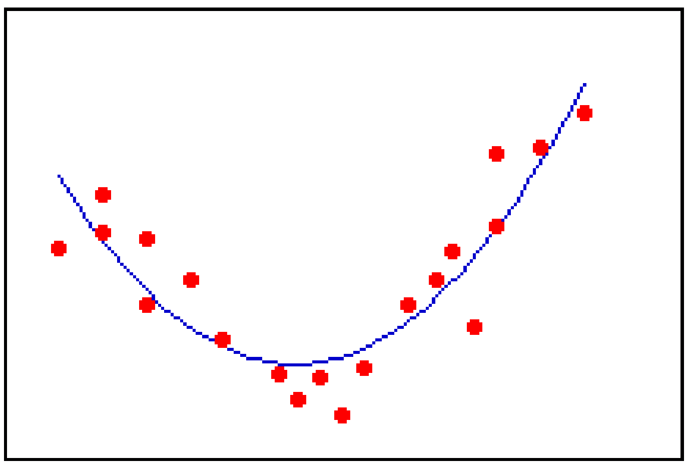
Correlation $r = -0.99$

Cautions

- Pearson Correlation requires that both variables be quantitative.
- Pearson Correlation does not describe *curved relationships* between variables
- The Pearson correlation *r* is not robust to outliers.

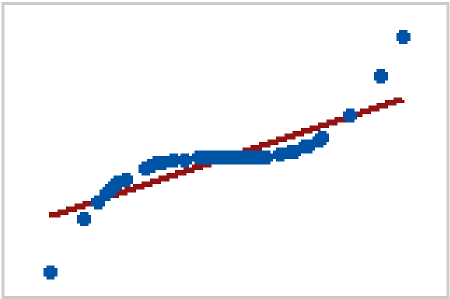
Examples of curved relationships



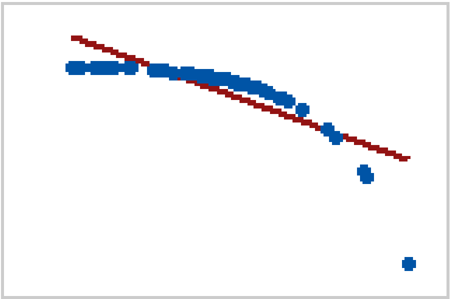


Spearman Correlation

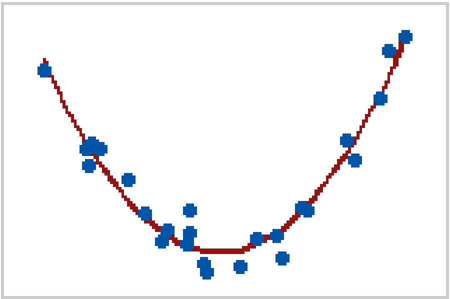
- The **Spearman correlation** (rank-based) evaluates the **monotonic** relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate.



Pearson = +0.85
Spearman = +1



Pearson = -0.80
Spearman = -1



Pearson = 0
Spearman = 0

Multiple Linear Regression

Multiple Regression

- Why use multiple predictors?
  - There are many problems in which a knowledge of more than one explanatory variable is necessary in order to obtain a better understanding and better prediction of a particular response (reduce unexplainable variability)
  - Gives more realistic interpretations by allowing us to adjust for other explanatory factors while isolating the effect of one variable
    - For example, whether a biomarker is an independent predictor of survival for patients with ovarian cancer (after removing the effect of age, tumor grade and stage)?

Data for Multiple Regression

The data for a simple linear regression problem consist of *n* observations (x_i, y_i) of two variables.

Data for multiple linear regression consist of the value of a response variable *y* and *p* explanatory variables $(x_1, x_2, \ldots, x_p)$ on each of *n* cases.

We write the data and enter them into software in the form:

	Variables				
Case	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	…	$\boldsymbol{x_p}$	$\boldsymbol{y}$
1	x_{11}	x_{12}	…	x_{1p}	y_1
2	x_{21}	x_{22}	…	x_{2p}	y_2
…	…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{np}	y_n

Motivating Example

outcome predictors

Pt	BP	Age	Weight	BSA	Dur	Pulse	Stress
1	105	47	85.4	1.75	5.1	63	33
2	115	49	94.2	2.10	3.8	70	14
3	116	49	95.3	1.98	8.2	72	10
4	117	50	94.7	2.01	5.8	73	99
5	112	51	89.4	1.89	7.0	72	95
6	121	48	99.5	2.25	9.3	71	10
7	121	49	99.8	2.25	2.5	69	42
8	110	47	90.9	1.90	6.2	66	8
9	110	49	89.2	1.83	7.1	69	62
10	114	48	92.7	2.07	5.6	64	35
11	114	47	94.4	2.07	5.3	74	90

Multiple Linear Regression

- Up to this point, we have considered the linear regression model in which the response, *y*, is related to **one predictor variable** *x*:

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$

- In multiple regression, the response, *y*, depends on ***p*** **predictor variables**:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$$

The **error** ε_i are independent and Normally distributed $N(0, s)$

Estimation of Parameters

The least-squares regression method chooses $b_0, b_1,\ldots,b_p$ to minimize the sum of squared deviations $(y_i - \hat{y}_i)^2$, where

$$\hat{y}_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip}$$

There is only one yhat no matter how many predictors there are

As with simple linear regression, the constant b_0 is the **intercept**.

- The regression **coefficients ($b_1,\ldots, b_p$)** describe the effect of each predictor on the *y* variable while **removing** the effects of other predictors (i.e. holding other predictors constant)

Interpret Parameters Using Examples

- $\hat{y} = b_0 + b_1 x_1$ **(one predictor: age)**

First subject is 32 and second subject is 33 years old.

$$\left.\begin{aligned} \hat{y}_1 &= b_0 + b_1 \times 32 \\ \hat{y}_2 &= b_0 + b_1 \times 33 \end{aligned}\right\} \hat{y}_2 - \hat{y}_1 = b_1$$

✓ b_1 represents the increase in y for every one year increase in age

- $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ **(two predictors: age and weight)**

Given the two subjects have the same weight x_2 (e.g, $x_2 = 140$)

$$\left.\begin{aligned} \hat{y}_1 &= b_0 + b_1 \times 32 + b_2 x_2 \\ \hat{y}_2 &= b_0 + b_1 \times 33 + b_2 x_2 \end{aligned}\right\} \hat{y}_2 - \hat{y}_1 = b_1$$

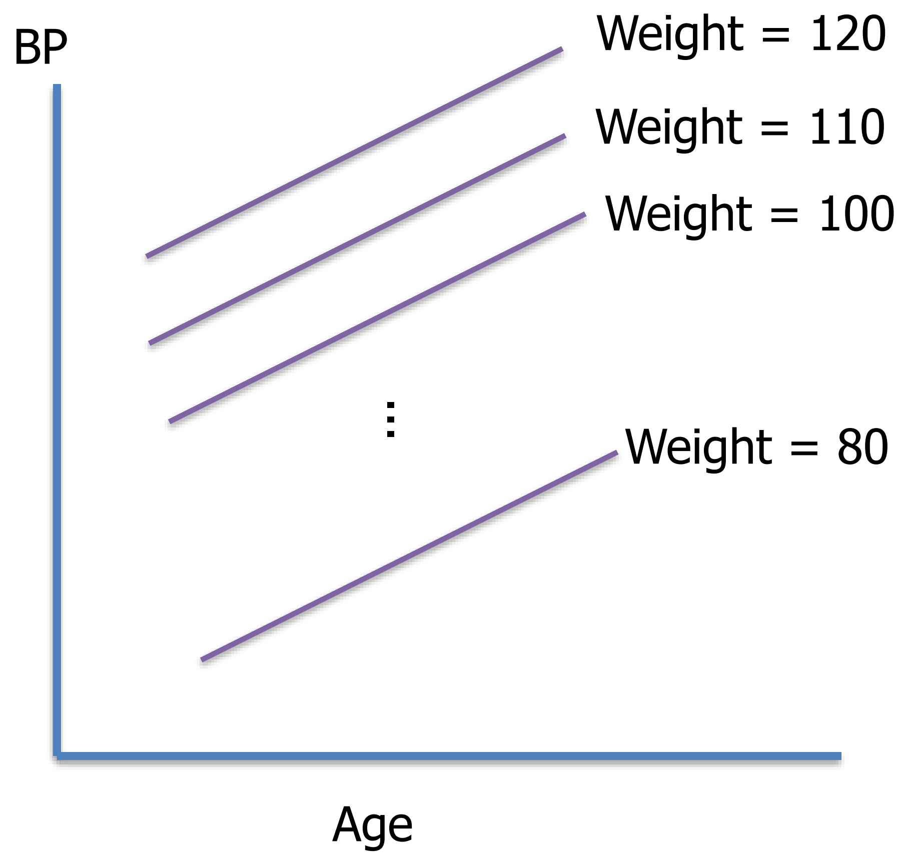
✓ b_1 represents the increase in y for every one year increase in age **given the same value of x_2**

✓ We also call b_1 the effect of age after adjusting for (or removing the effect of) weight

Interpret Parameters Using Graph

$\hat{y}_1 = b_0 + b_1 \times Age + b_2 Weight$

Interpret Age adjusting for Weight

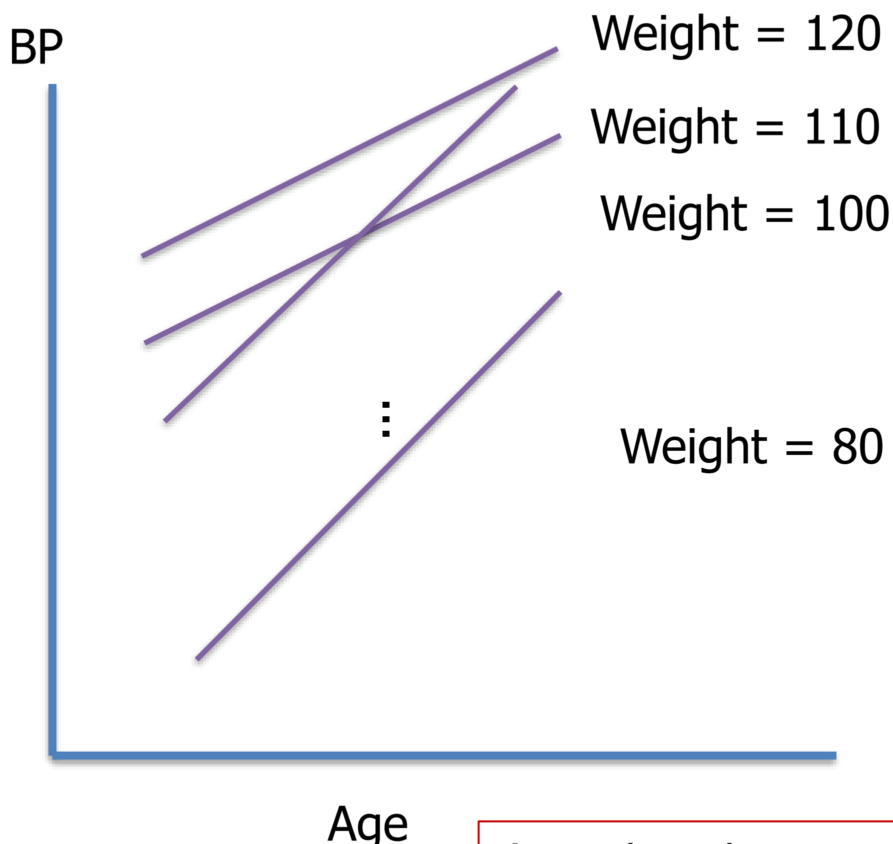


b_1 shows the relationship between age and BP while holding Weight constant

Similarly, b_2 shows the relationship between Weight and BP while holding Age constant

Multiple Regression with an Interaction

Two continuous covariates



$$\hat{y}_1 = b_0 + b_1 \times Age + b_2 Weight + \boldsymbol{b_3\, Age \times Weight}$$

Analysis Results with 2 Predictors

Parameter Estimates				
Variable	Parameter Estimate	Standard Error	t Value	p value
Intercept	-16.57937	3.00746	-5.51	<.0001
Age	0.70825	0.05351	13.23	<.0001
Weight	1.03296	0.03116	33.15	<.0001

Both age and weight are significant in the multiple regression

$$\hat{y} = -16.58 + 0.71x_{age} + 1.03\, x_{weight}$$

- 0.71 is the effect of age after controlling for weight
- 1.03 is the effect of weight after controlling for age
- *-16.58 is the value when both age and weight are 0*

Analysis Results with 6 Predictors

Parameter Estimates						
Variable	Parameter Estimate	Standard Error	t Value	P value	95% Confidence Limits	
Intercept	-12.87048	2.55665	-5.03	0.0002	-18.39378	-7.34717
Age	0.70326	0.04961	14.18	<.0001	0.59609	0.81043
Weight	0.96992	0.06311	15.37	<.0001	0.83358	1.10626
BSA	3.77649	1.58015	2.39	0.0327	0.36278	7.19020
Dur	0.06838	0.04844	1.41	0.1815	-0.03627	0.17303
Pulse	-0.08448	0.05161	-1.64	0.1256	-0.19598	0.02701
Stress	0.00557	0.00341	1.63	0.1265	-0.00180	0.01294

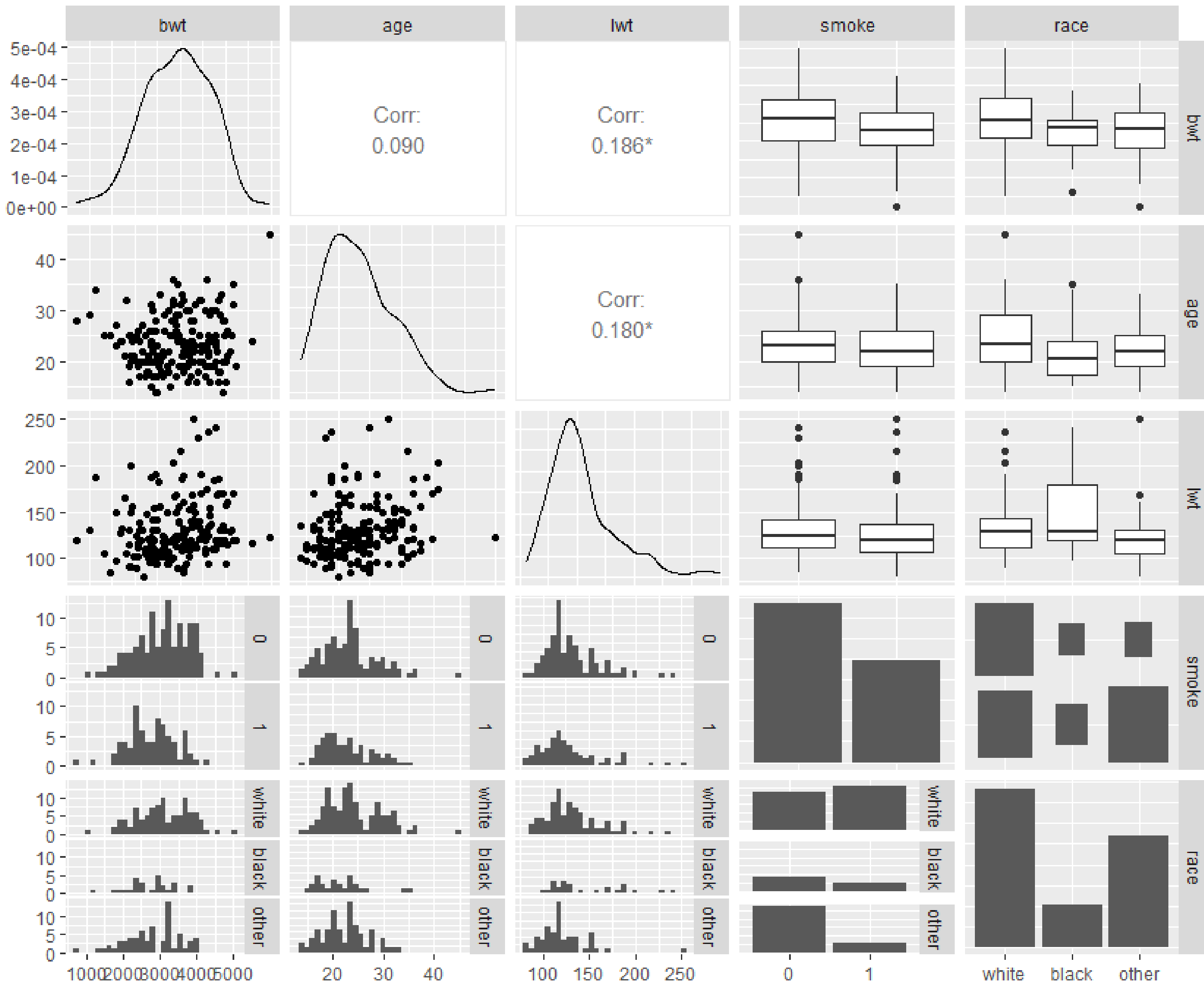
Conclusion: Age, Weight and BSA are significant in the multiple regression. For Age: Age has a significant effect on BP after controlling for Weight, BSA, Dur, Pulse and Stress, and BP is increased by 0.7 (95%CI: [0.6,0.8], p <0.0001) for every one year's increase in age when all other covariates are fixed.

Real Data Example- birthweight

The Excel spreadsheet 'birthweight.xls' contains data on 189 births collected at Baystate Medical Center, Springfield, Mass. during 1986. The goal of the study was to assess the relationship between infant birthweight (bwt) and mother's age, weight (lwt), smoking status during pregnancy (smoke), and race. The dataset consists of the following 5 variables:

- b*wt:* birth weight (in grams) (**The dependent variable**)
- *age:* mother's age in years
- *lwt:* mother's weight in pounds at last menstrual period
- *smoke:* smoking status during pregnancy
- *race:* mother's race ("white", "black", "other")

Exploratory Data Analysis



Analysis Results

```
> ## multiple regression model
> m_lm_multiple = lm(bwt ~ age+ lwt+smoke+race, data = df_birthweight)
>
> # put results into a table
> tbl_summary =
+   cbind(summary(m_lm_multiple)$coefficients[,c(1,4)], confint(m_lm_multiple))
> tbl_summary
               Estimate     Pr(>|t|)        2.5 %       97.5 %
(Intercept) 2839.433435 8.196552e-16 2205.2392620 3473.627608
age           -1.947841 8.429898e-01  -21.3230509   17.427369
lwt            3.999938 2.249357e-02    0.5708088    7.429068
smoke1      -401.720488 3.095917e-04 -617.2537916 -186.187185
raceblack   -510.501493 1.373456e-03 -820.4159434 -200.587043
raceother   -398.643859 1.037171e-03 -634.5750995 -162.712619
```

<u>Conclusion</u>: in the multiple regression, lwt, race and smoke are significantly associated with birthweight. For example, Black race has lower weight baby than White race (difference is 511, 95% CI; 201, 820, p=0.0014)

R square and Partial R square

Partitioning Total Variability

- It can be shown that

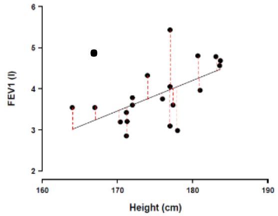
total variation | Variation explained by Model | Variation explained by Model

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

$$SS_{Total} = SS_{Model} + SS_{Error}$$

$$SSTotal = SSM + SSE$$

$\sum_{i=1}^{n} {e_i}^2$



$\hat{y} = b_0 + b_1x$

R^2 is the proportion of the variance of Y that is explained by the model. $0 \le R^2 \le 1$

R^2 is the square of the Pearson correlation coefficient.

R^2 and Partial R^2

Intercept-only model in simple linear regression

$$R^2 = \frac{SSTotal - SSE(full)}{SSTotal}$$

$$R^2_{partial} = \frac{SSE(reduced) - SSE(full)}{SSE(reduced)}$$

In multiple regression with multiple predictors (x1, x2, x3), a reduced model has a subset of the predictors, for example, a reduced model has x1 alone, (x1, x2), or (x2, x3).