Chapter 9: Multivariable Methods

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Associations

We often are interested in the association between variables.

We often say **correlation**, with little thought to an actual definition.

We often say trend or **linear** relationship without defining how determine this relationship.

We define y to be the response or **dependent** (on x) **variable** and x to be the explanatory or **independent variable**. i.e. y depends on x (or several x's).

9.3 Introduction to Correlation and Regression Analysis

Formally, correlation is a measure of “linear” association between two continuous variables x and y. Sample value r and population value ρ.

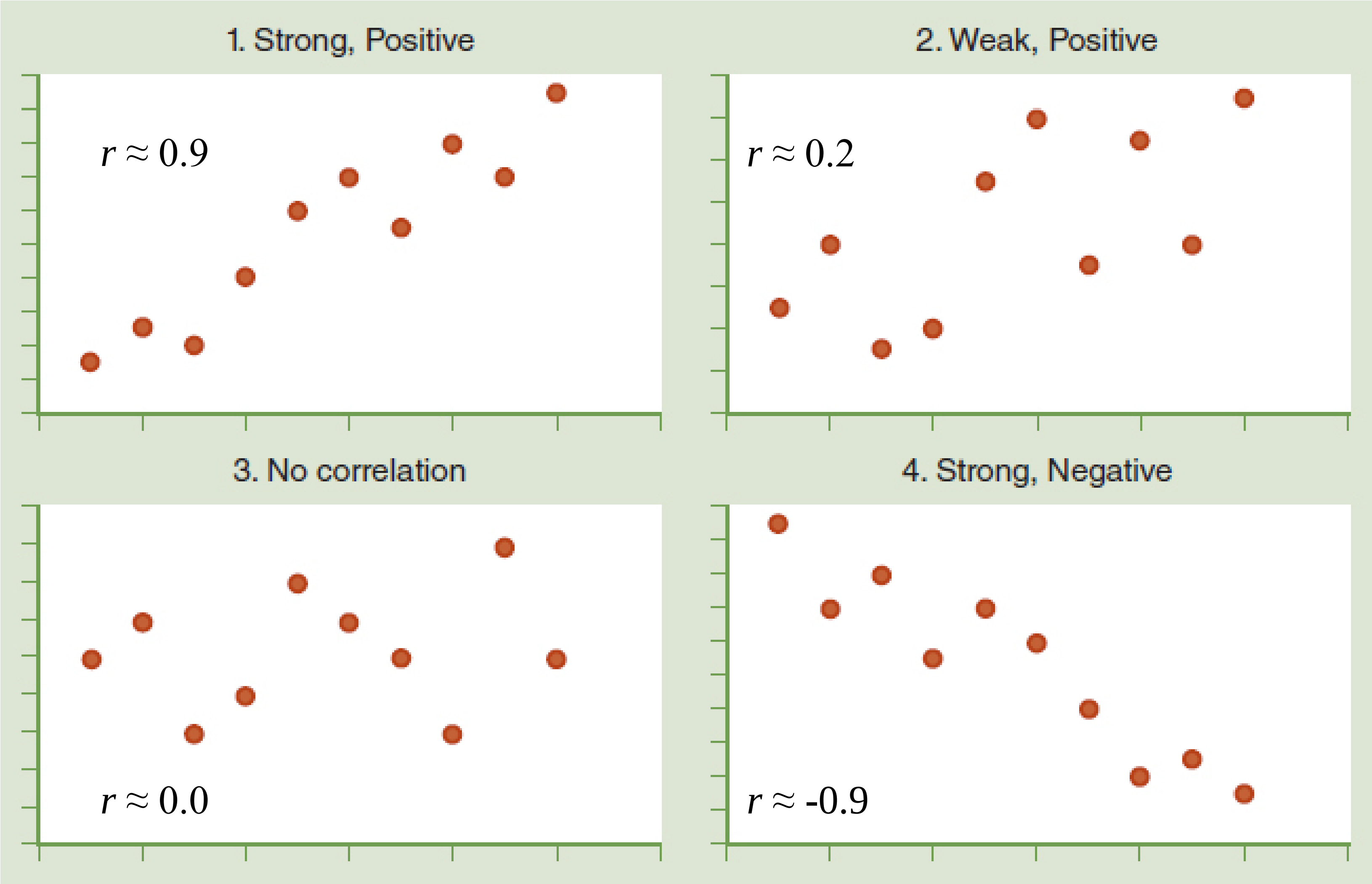
Correlations are between -1 and 1. $-1 \le r \le 1$ (will calculate soon)

Values close to $+1$ or -1 mean a stronger association.
Values close to 0 mean a weak association.

Positive values mean a positive relationship, as x increases so does y.
Negative values mean a negative relationship, x increases y decreases.

9.3 Introduction to Correlation and Regression Analysis

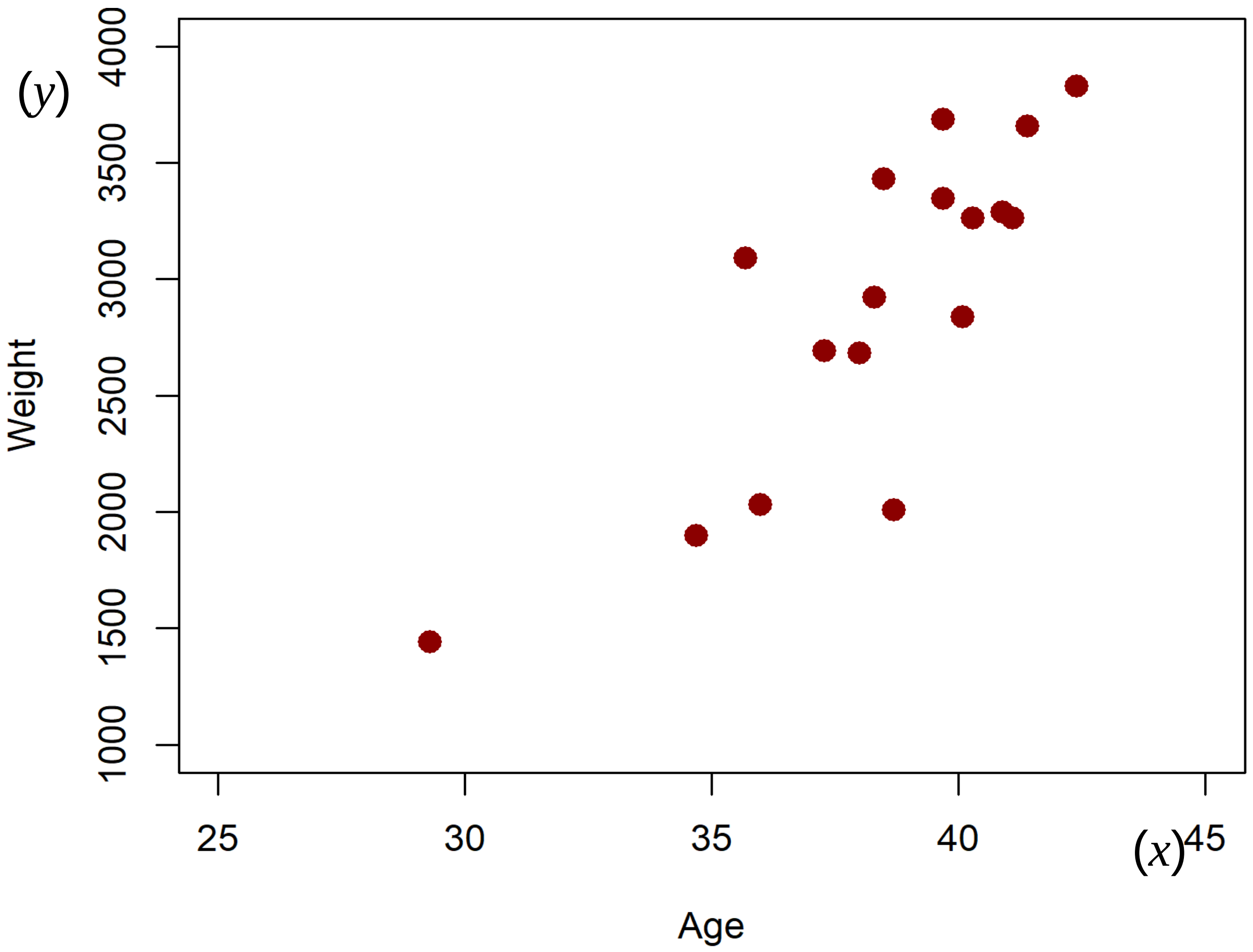
A scatter diagram is a plot of the independent variable (x) and the dependent variable (y). From it we can glean if there is an association.



9.3 Introduction to Correlation and Regression Analysis

Example: A small study … to investigate the association between gestational age and birth weight. A scatter diagram is constructed.

Infant Identification Number	(X) Gestational Age (weeks)	(Y) Birth Weight (grams)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005



Is there a relationship between age and weight?

In general as age increases does weight increase?

9.3 Introduction to Correlation and Regression Analysis-Correlation

Correlations r are between -1 and 1, $-1 \le r \le 1$.

$$r = \frac{\text{cov}(x,y)}{\sqrt{s_x^2 s_y^2}}$$

Variance of X

$$s_x^2 = \frac{1}{n-1}\sum (X-\bar{X})^2 = \frac{1}{n-1}\left[\sum X^2 - \frac{1}{n}\left(\sum X\right)^2\right]$$

Variance of Y

$$s_y^2 = \frac{1}{n-1}\sum (Y-\bar{Y})^2 = \frac{1}{n-1}\left[\sum Y^2 - \frac{1}{n}\left(\sum Y\right)^2\right]$$

CoVariance of X&Y

$$\text{cov}(x,y) = \frac{1}{n-1}\sum (Y-\bar{Y})(X-\bar{X}) = \frac{1}{n-1}\left[\sum XY - \frac{1}{n}\left(\sum Y\right)\left(\sum X\right)\right]$$

Not in book

9.3 Introduction to Correlation and Regression Analysis-Correlation

We are going to calculate the correlation in column format with sums.

n	X	X^2	Y	Y^2	XY
1	34.7		1895		
2	36.0		2030		
3	29.3		1440		
4	40.1		2835		
5	35.7		3090		
6	42.4		3827		
7	40.3		3260		
8	37.3		2690		
9	40.9		3285		
10	38.3		2920		
11	38.5		3430		
12	41.4		3657		
13	39.7		3685		
14	39.7		3345		
15	41.1		3260		
16	38.0		2680		
17	38.7		2005		

5 Sums

$\sum X =$ $\qquad$ $\sum X^2 =$

$\sum Y =$ $\qquad$ $\sum Y^2 =$

$\sum XY =$

9.3 Introduction to Correlation and Regression Analysis-Correlation

We are going to calculate the correlation in column format with sums.

n	X	X^2	Y	Y^2	XY
1	34.7	1204.1	1895	3591025.0	65756.5
2	36.0	1296.0	2030	4120900.0	73080.0
3	29.3	858.5	1440	2073600.0	42192.0
4	40.1	1608.0	2835	8037225.0	113683.5
5	35.7	1274.5	3090	9548100.0	110313.0
6	42.4	1797.8	3827	14645929.0	162264.8
7	40.3	1624.1	3260	10627600.0	131378.0
8	37.3	1391.3	2690	7236100.0	100337.0
9	40.9	1672.8	3285	10791225.0	134356.5
10	38.3	1466.9	2920	8526400.0	111836.0
11	38.5	1482.3	3430	11764900.0	132055.0
12	41.4	1714.0	3657	13373649.0	151399.8
13	39.7	1576.1	3685	13579225.0	146294.5
14	39.7	1576.1	3345	11189025.0	132796.5
15	41.1	1689.2	3260	10627600.0	133986.0
16	38.0	1444.0	2680	7182400.0	101840.0
17	38.7	1497.7	2005	4020025.0	77593.5

5 Sums

$$\sum X = \qquad \sum X^2 =$$

$$\sum Y = \qquad \sum Y^2 =$$

$$\sum XY =$$

9.3 Introduction to Correlation and Regression Analysis-Correlation

We are going to calculate the correlation in column format with sums.

n	X	X^2	Y	Y^2	XY
1	34.7	1204.1	1895	3591025.0	65756.5
2	36.0	1296.0	2030	4120900.0	73080.0
3	29.3	858.5	1440	2073600.0	42192.0
4	40.1	1608.0	2835	8037225.0	113683.5
5	35.7	1274.5	3090	9548100.0	110313.0
6	42.4	1797.8	3827	14645929.0	162264.8
7	40.3	1624.1	3260	10627600.0	131378.0
8	37.3	1391.3	2690	7236100.0	100337.0
9	40.9	1672.8	3285	10791225.0	134356.5
10	38.3	1466.9	2920	8526400.0	111836.0
11	38.5	1482.3	3430	11764900.0	132055.0
12	41.4	1714.0	3657	13373649.0	151399.8
13	39.7	1576.1	3685	13579225.0	146294.5
14	39.7	1576.1	3345	11189025.0	132796.5
15	41.1	1689.2	3260	10627600.0	133986.0
16	38.0	1444.0	2680	7182400.0	101840.0
17	38.7	1497.7	2005	4020025.0	77593.5
	652.1	25173.2	49334.0	150934928.0	1921162.6

5 Sums

$$\sum X = 652.1 \qquad \sum X^2 = 25173.2$$

$$\sum Y = 49334.0 \qquad \sum Y^2 = 150934928.0$$

$$\sum XY = 1921162.6$$

9.3 Introduction to Correlation and Regression Analysis-Correlation

We are going to calculate the correlation in column format with sums.

n	X	X^2	Y	Y^2	XY
1	34.7	1204.1	1895	3591025.0	65756.5
2	36.0	1296.0	2030	4120900.0	73080.0
3	29.3	858.5	1440	2073600.0	42192.0
4	40.1	1608.0	2835	8037225.0	113683.5
5	35.7	1274.5	3090	9548100.0	110313.0
6	42.4	1797.8	3827	14645929.0	162264.8
7	40.3	1624.1	3260	10627600.0	131378.0
8	37.3	1391.3	2690	7236100.0	100337.0
9	40.9	1672.8	3285	10791225.0	134356.5
10	38.3	1466.9	2920	8526400.0	111836.0
11	38.5	1482.3	3430	11764900.0	132055.0
12	41.4	1714.0	3657	13373649.0	151399.8
13	39.7	1576.1	3685	13579225.0	146294.5
14	39.7	1576.1	3345	11189025.0	132796.5
15	41.1	1689.2	3260	10627600.0	133986.0
16	38.0	1444.0	2680	7182400.0	101840.0
17	38.7	1497.7	2005	4020025.0	77593.5
	652.1	25173.2	49334.0	150934928.0	1921162.6

5 Sums

$$\sum X = 652.1 \qquad \sum X^2 = 25173.2$$

$$\sum Y = 49334.0 \qquad \sum Y^2 = 150934928.0$$

$$\sum XY = 1921162.6$$

$$\operatorname{cov}(x,y) = \frac{1}{n-1}\left[\sum XY - \frac{1}{n}\left(\sum Y\right)\left(\sum X\right)\right]$$

$$s_x^2 = \frac{1}{n-1}\left[\sum X^2 - \frac{1}{n}\left(\sum X\right)^2\right]$$

$$r = \frac{\operatorname{cov}(x,y)}{\sqrt{s_x^2 s_y^2}}$$

$$s_y^2 = \frac{1}{n-1}\left[\sum Y^2 - \frac{1}{n}\left(\sum Y\right)^2\right]$$

9.3 Introduction to Correlation and Regression Analysis-Correlation

We are going to calculate the correlation in column format with sums.

n	X	X^2	Y	Y^2	XY
1	34.7	1204.1	1895	3591025.0	65756.5
2	36.0	1296.0	2030	4120900.0	73080.0
3	29.3	858.5	1440	2073600.0	42192.0
4	40.1	1608.0	2835	8037225.0	113683.5
5	35.7	1274.5	3090	9548100.0	110313.0
6	42.4	1797.8	3827	14645929.0	162264.8
7	40.3	1624.1	3260	10627600.0	131378.0
8	37.3	1391.3	2690	7236100.0	100337.0
9	40.9	1672.8	3285	10791225.0	134356.5
10	38.3	1466.9	2920	8526400.0	111836.0
11	38.5	1482.3	3430	11764900.0	132055.0
12	41.4	1714.0	3657	13373649.0	151399.8
13	39.7	1576.1	3685	13579225.0	146294.5
14	39.7	1576.1	3345	11189025.0	132796.5
15	41.1	1689.2	3260	10627600.0	133986.0
16	38.0	1444.0	2680	7182400.0	101840.0
17	38.7	1497.7	2005	4020025.0	77593.5
	652.1	25173.2	49334.0	150934928.0	1921162.6

5 Sums

$$\sum X = 652.1 \qquad \sum X^2 = 25173.2$$

$$\sum Y = 49334.0 \qquad \sum Y^2 = 150934928.0$$

$$\sum XY = 1921162.6$$

$$\text{cov}(x,y) = \frac{1}{17-1}\left[1921162.6 - \frac{1}{17}(49334.0)(652.1)\right]$$

$$s_x^2 = \frac{1}{17-1}\left[25173.2 - \frac{1}{17}(652.1)^2\right] \qquad r = \frac{\text{cov}(x,y)}{\sqrt{s_x^2 s_y^2}}$$

$$s_y^2 = \frac{1}{17-1}\left[150934928.0 - \frac{1}{17}(49334.0)^2\right]$$

9.3 Introduction to Correlation and Regression Analysis-Correlation

We are going to calculate the correlation in column format with sums.

n	X	X^2	Y	Y^2	XY
1	34.7	1204.1	1895	3591025.0	65756.5
2	36.0	1296.0	2030	4120900.0	73080.0
3	29.3	858.5	1440	2073600.0	42192.0
4	40.1	1608.0	2835	8037225.0	113683.5
5	35.7	1274.5	3090	9548100.0	110313.0
6	42.4	1797.8	3827	14645929.0	162264.8
7	40.3	1624.1	3260	10627600.0	131378.0
8	37.3	1391.3	2690	7236100.0	100337.0
9	40.9	1672.8	3285	10791225.0	134356.5
10	38.3	1466.9	2920	8526400.0	111836.0
11	38.5	1482.3	3430	11764900.0	132055.0
12	41.4	1714.0	3657	13373649.0	151399.8
13	39.7	1576.1	3685	13579225.0	146294.5
14	39.7	1576.1	3345	11189025.0	132796.5
15	41.1	1689.2	3260	10627600.0	133986.0
16	38.0	1444.0	2680	7182400.0	101840.0
17	38.7	1497.7	2005	4020025.0	77593.5
	652.1	25173.2	49334.0	150934928.0	1921162.6

5 Sums

$$\sum X = 652.1 \qquad \sum X^2 = 25173.2$$

$$\sum Y = 49334.0 \qquad \sum Y^2 = 150934928.0$$

$$\sum XY = 1921162.6$$

$$\text{cov}(x, y) = 1798.0$$

$$s_x^2 = 9.9638$$

$$s_y^2 = 485478.8$$

$$r = \frac{1798.0}{\sqrt{(10.0)(485478.8)}}$$

$$r = 0.82$$

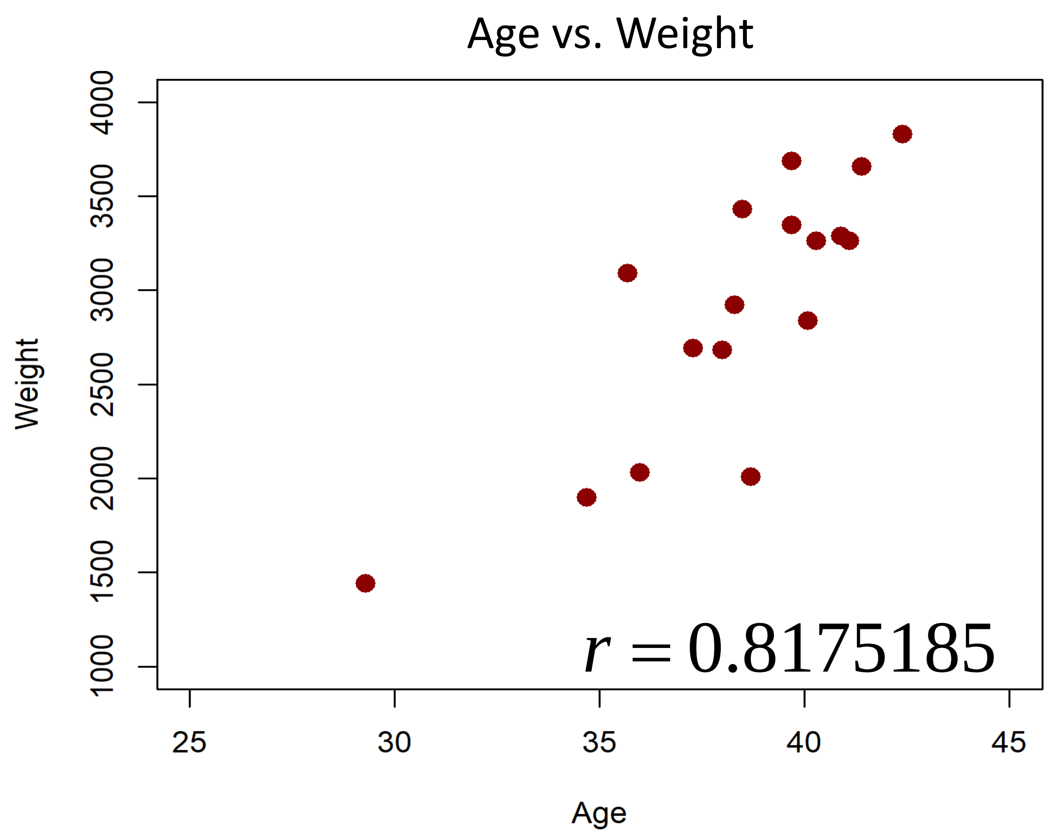
9.3 Introduction to Correlation and Regression Analysis-Correlation

It is important to know how the correlation is calculated, critical thinker. However, in practice we use a software package such as ***R***.

```
# Gestational Data
xx <- c(34.7,36.0,29.3,40.1,35.7,42.4,40.3,37.3,40.9,38.3,38.5,41.4,39.7,39.7,41.1,38.0,38.7)
yy <- c(1895,2030,1440,2835,3090,3827,3260,2690,3285,2920,3430,3657,3685,3345,3260,2680,2005)

plot(x = xx,y = yy,xlab = "Age",ylab = "Weight", , main = "Age vs. Weight",
xlim = c(25,45),ylim = c(1000,4000), col = "darkred", cex = 1.5pch = 16)

sx2 <- var(xx)
sy2 <- var(yy)
sxy <- cov(xx,yy)
r    <- cor(xx,yy)
```



Can also form test statistic $z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ to test H_0: $\rho=0$ vs. H_1: $\rho\neq 0$.

9.3 Introduction to Correlation and Regression Analysis-Regression

We often want to estimate the linear association between the independent variable (x) and the dependent variable (y).
From it we can more quantitively describe an x - y association

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

From data, we are going to estimate β_0 by b_0 and β_1 by b_1 and denote the estimated relationship by

$\hat{y} = b_0 + b_1 x$. This is simple linear regression.

Of note x does not have to be continuous in regression but y does.

9.3 Introduction to Correlation and Regression Analysis-Regression

$s_x^2 = 9.9638$

$s_y^2 = 485478.8$

$r = 0.82$

We can estimate the y-intercept and slope from what we have already computed for the correlation.

The slope is estimated as $b_1 = r\dfrac{s_y}{s_x}$ and $b_0 = \bar{Y} - b_1\bar{X}$.

Point slope formula

Line goes through $(\bar{X},\bar{Y})$. Note b_1 has same sign as r.

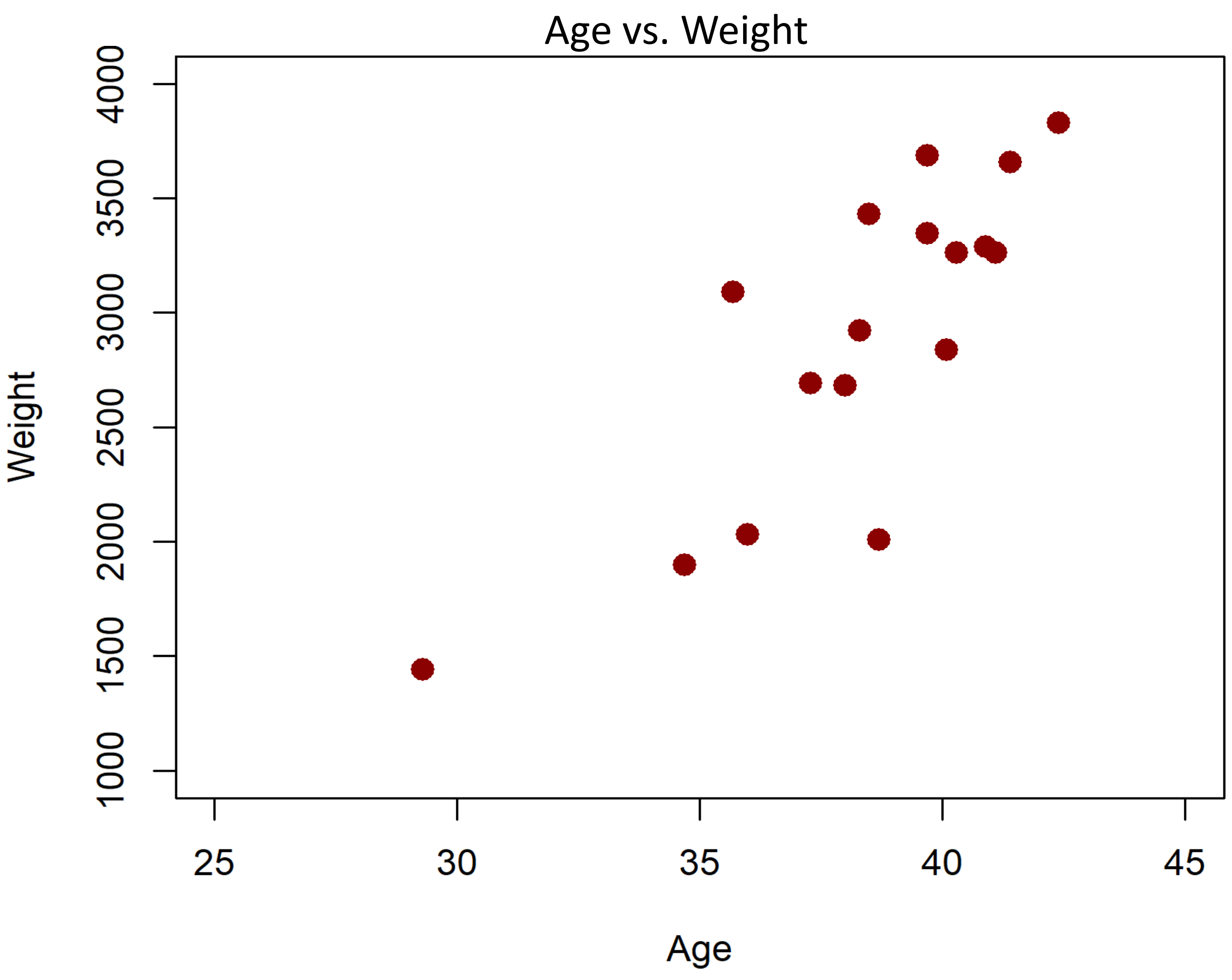
And hence we have determined our regression line.

$\hat{y} = b_0 + b_1 x$

9.3 Introduction to Correlation and Regression Analysis-Regression

Example: Continuing the small study … to investigate the association between gestational age and birth weight.

Infant Identification Number	(X) Gestational Age (weeks)	(Y) Birth Weight (grams)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005



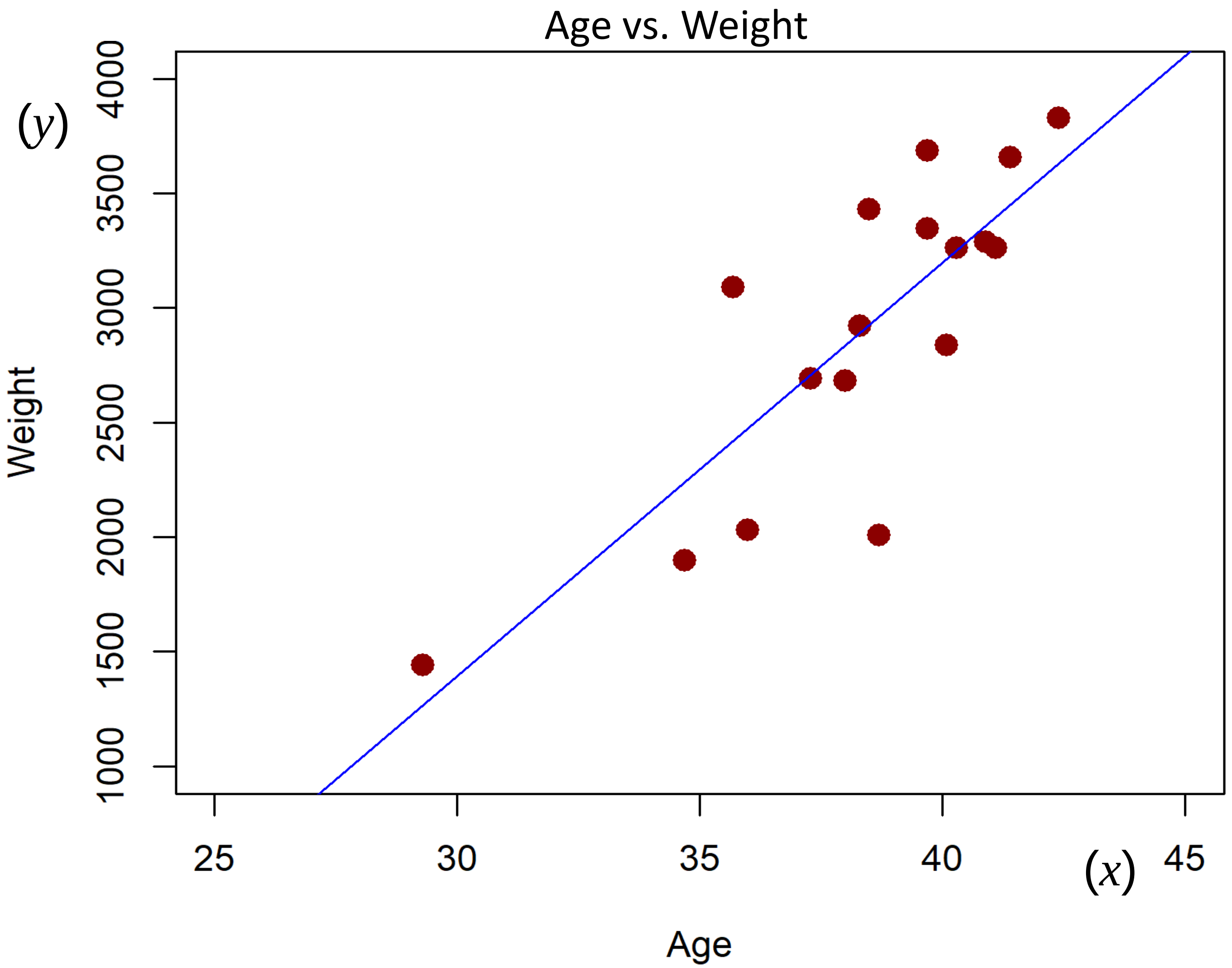
9.3 Introduction to Correlation and Regression Analysis-Regression

Example: Continuing the small study … to investigate the association between gestational age and birth weight.

$$s_x^2 = 9.9638$$

$$s_y^2 = 485478.8$$

$$r = 0.82$$



$$b_1 = r\frac{s_y}{s_x}$$

$$b_1 = 0.82\frac{696.8}{3.2}$$

$$b_1 = 180.5$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_0 = 2902 - (180.5)(38.4)$$

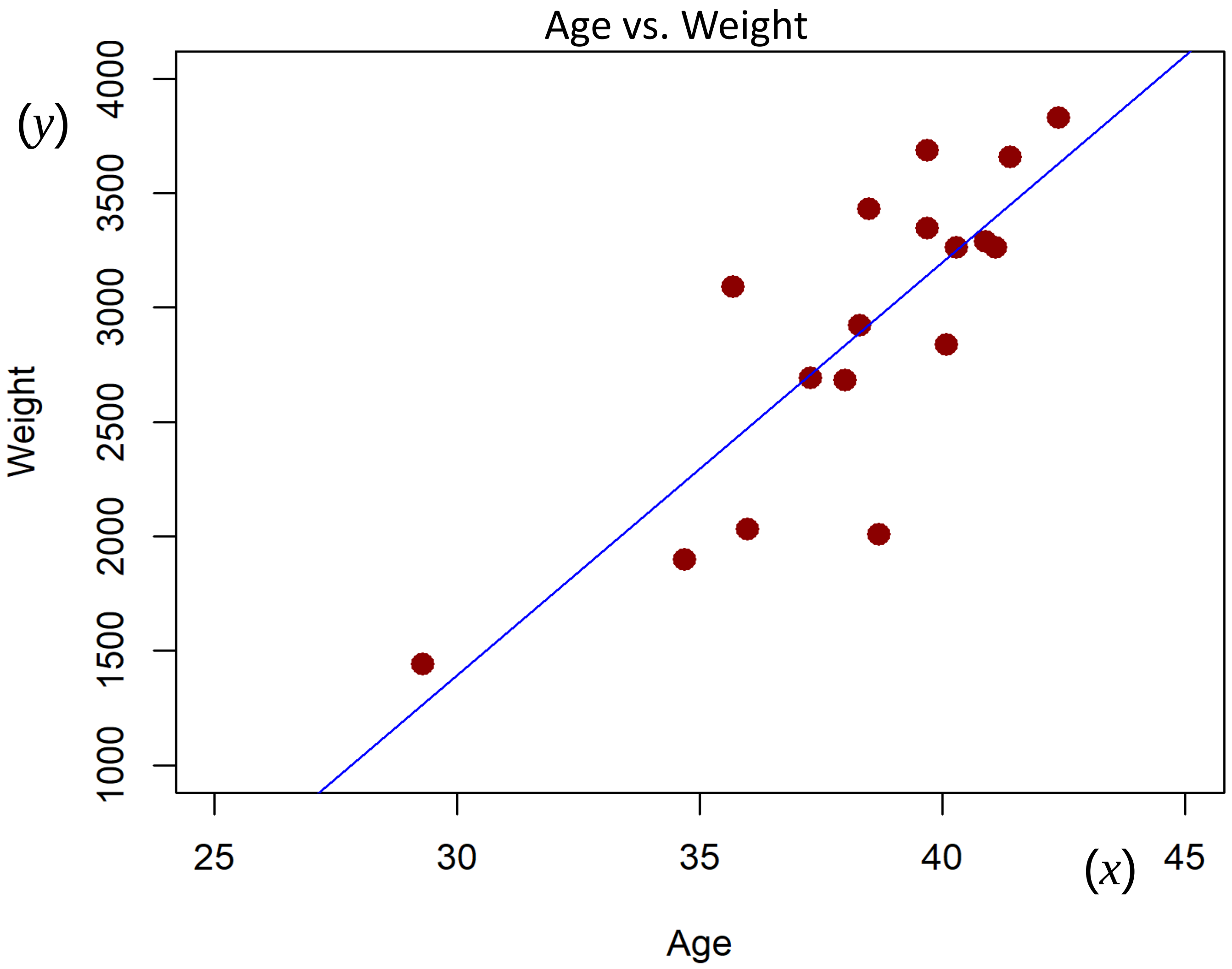
$$b_0 = -4029.2$$

$$\hat{Y} = -4029.2 + 180.5x$$

9.3 Introduction to Correlation and Regression Analysis-Regression

Example: Continuing the small study … to investigate the association between gestational age and birth weight.

A one-week increase in gestational age on average results in a 180.5 gram increase in weight.



$$b_1 = r\frac{s_y}{s_x}$$

$$b_1 = 180.5$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_0 = -4029.2$$

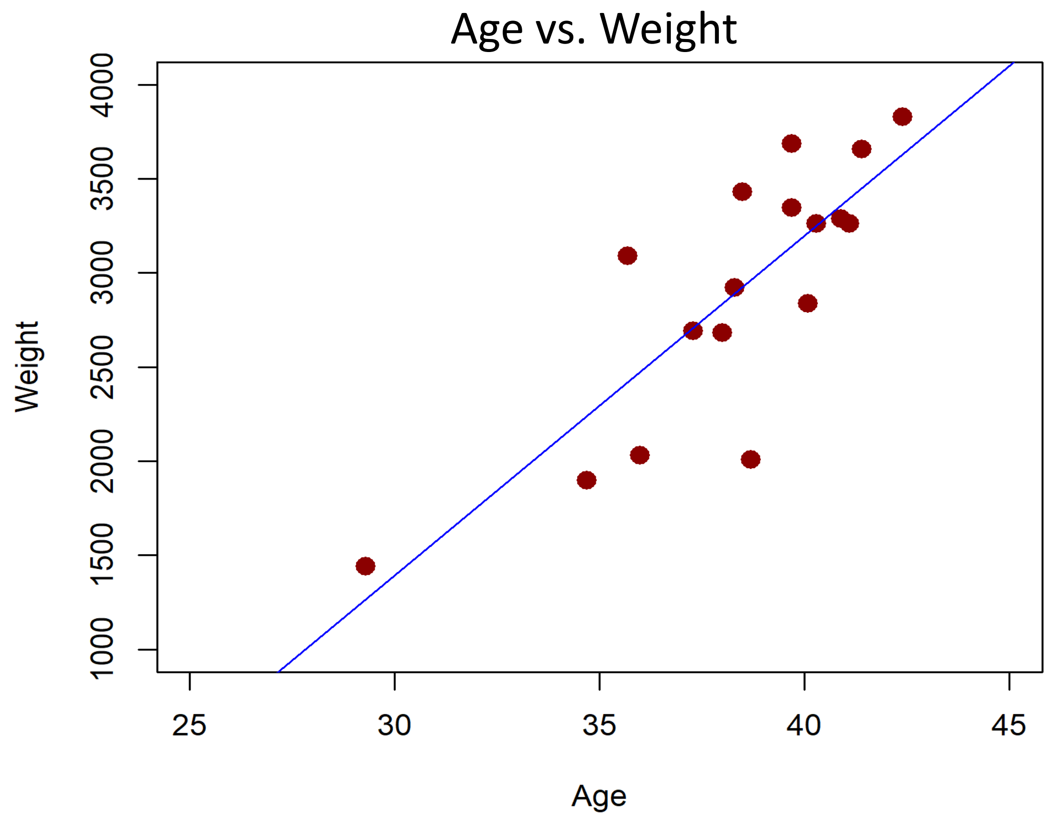
$$\hat{Y} = -4029.2 + 180.5x$$

9.3 Introduction to Correlation and Regression Analysis-Correlation

It is important to know how the regression line is calculated, critical thinker. However, in practice we use a software package such as ***R***.

```
# Gestational Data
xx <- c(34.7,36.0,29.3,40.1,35.7,42.4,40.3,37.3,40.9,38.3,38.5,41.4,39.7,39.7,41.1,38.0,38.7)
yy <- c(1895,2030,1440,2835,3090,3827,3260,2690,3285,2920,3430,3657,3685,3345,3260,2680,2005)

#scatter plot with line
plot(x = xx,y = yy,xlab ="Age",ylab = "Weight", xlim = c(25,45),ylim = c(1000,4000),
col = "darkred", cex = 1.5, main = "Age vs. Weight", pch = 16)
reg<-lm(yy ~ xx)
abline(reg, col = "blue")
coeff = coefficients(reg)
```



$$\hat{Y} = -4029.2 + 180.5x$$

9.4 Multiple Linear Regression Analysis

Our variable y might depend on more than one x, $x_1, x_2, \ldots, x_p$.

So we want to be able to estimate a relationship such as

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

However, the estimation of the regression coefficients, the b's is now much more complicated.

We would need to use a software package such as ***R***.

9.4 Multiple Linear Regression Analysis

Example: Suppose we want to know the relationship between systolic blood pressure (SBP) and the variables BMI, AGE, Male Sex (MLS), and Treatment for Hypertension (TFH).

We have measured SBP, BMI, AGE, MLS, and TFH on n=3959 study participants. Male Sex and Treatment are 0/1 variables.

A multiple regression analysis is run and coefficients estimated.

9.4 Multiple Linear Regression Analysis

Example: SBP and BMI, Age, Male Sex, and TFH.

A multiple regression analysis is run and coefficients estimated.

$$SBP = 68.15 + 0.58BMI + 0.65AGE + 0.94MLS + 6.44TFH$$

Independent Variable	Regression Coefficient	t	p-value
Intercept	b_0=68.15	t_0=26.33	0.0001=p_0
BMI	b_1= 0.58	t_1=10.30	0.0001=p_1
Age	b_2= 0.65	t_2=20.22	0.0001=p_2
Male sex	b_3= 0.94	t_3= 1.58	0.1133=p_3
Treatment for hypertension	b_4= 6.44	t_4= 9.74	0.0001=p_4

You will often see this type of output.

The t statistic is for H_0: $\beta_j=0$, H_1: $\beta_j \neq 0$.

The p-*value* is the probability of getting this coefficient estimate or larger in abs if it were truly 0.

$$t_j = \frac{b_j - 0}{\sqrt{\text{var}(b_j)}}$$

$df=n\text{-}p\text{-}1$

9.5 Multiple Logistic Regression Analysis

The probability p of an event E can depend on an independent variable x, such as the probability p of getting an A on the final depends on the number of hours that you study x.

If you study x=10 hours then your probability $p(x)$ of getting an A might be $p(10)=0.25$,
but if you study x=30 hours then your probability $p(x)$ of getting an A might be $p(30)=0.75$.

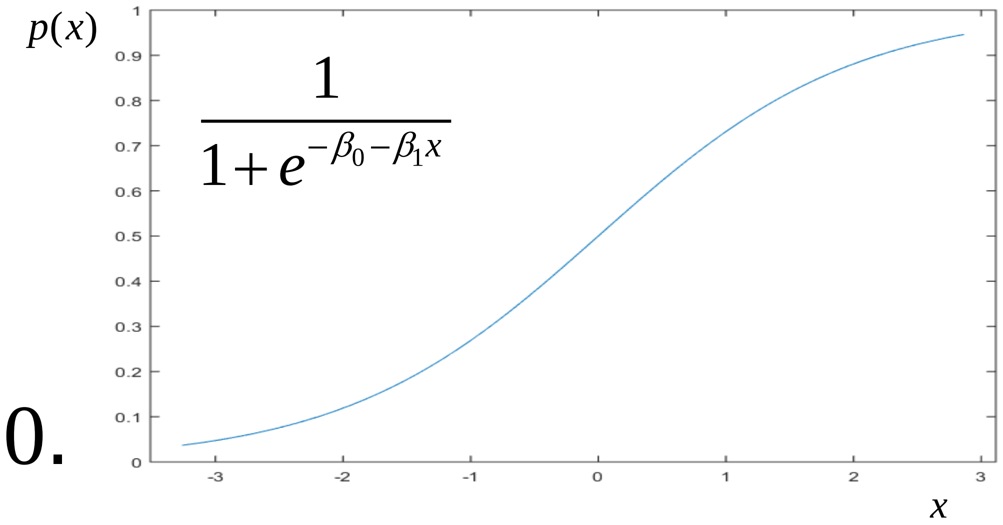
i.e. as x increases so does p..

Hours (x)	**A (y)**
6	0
8	0
10	0
12	0
14	0
16	1
18	0
20	0
22	0
24	0
26	1
28	0
30	0
32	1
34	1
36	1
38	1
40	1

9.5 Multiple Logistic Regression Analysis

This dependency of a probability $p(x)$, $0 \leq p(x) \leq 1$, on an independent variable x, $-\infty < x < \infty$, is generally described through the logistic mapping function

$$p = p(x) = \frac{1}{1+e^{-\beta_0-\beta_1 x}} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} .$$



If the event E occurs, then we say y=1 and if not y=0. $P(y$=1)=p and $P(y$=0)=1-p. …

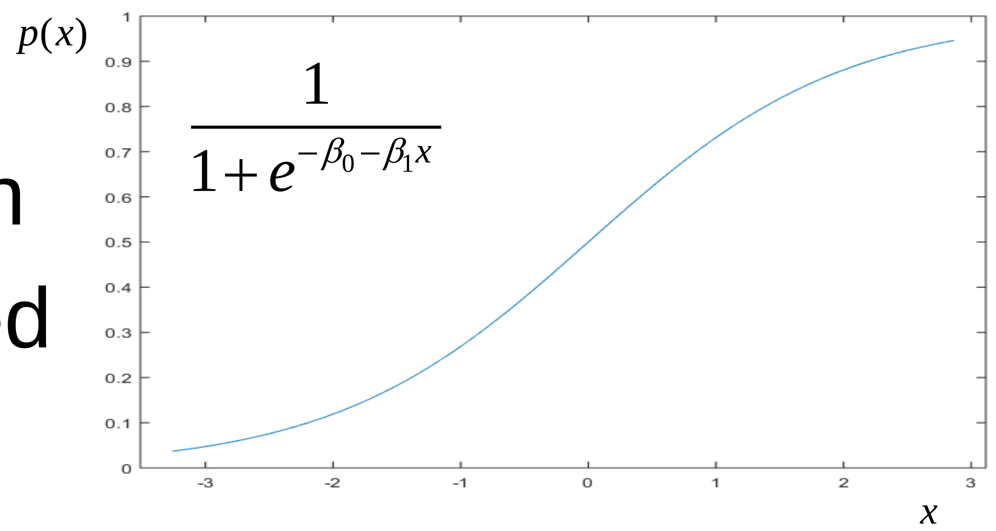
This is a Binomial trial with n=1 and whose probability of success depends on x.

Verhulst, 1838; Ostwald, 1883; .., Fisher,1935,….

9.5 Multiple Logistic Regression Analysis

Sometimes the logistic regression is written as log odds

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = b_0 + b_1 x_1 + \ldots + b_p x_p$$



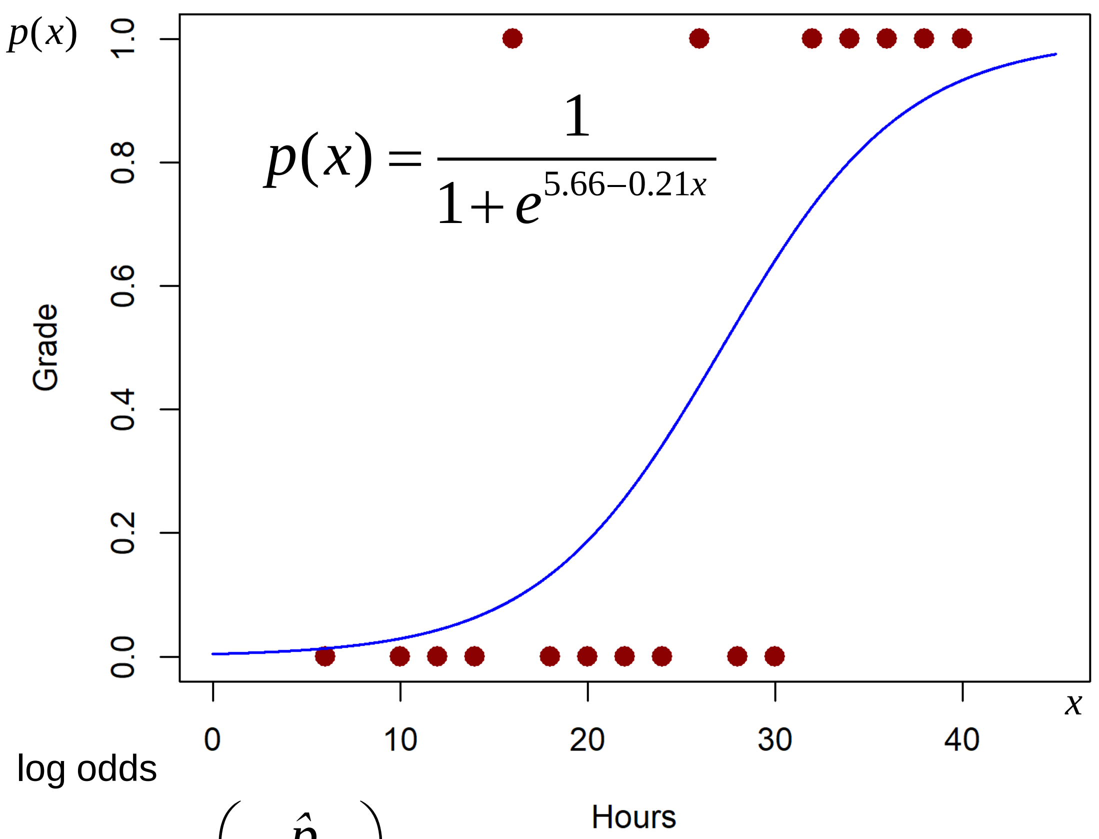
and it looks like we can then use Linear Regression to estimate the coefficients. It turns out that we need to find the coefficient values that maximize

$$LL = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} ln(1+e^{\beta_0+\beta_1 x_i}) \quad .$$

We need to use a software package such as ***R***.

9.5 Multiple Logistic Regression Analysis

Using *R*,



log odds

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -5.66 + 0.21x$$

output

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.65549    2.54651  -2.221   0.0264
xx           0.20778    0.09302   2.234   0.0255
```

Hours (x)	A (y)
6	0
8	0
10	0
12	0
14	0
16	1
18	0
20	0
22	0
24	0
26	1
28	0
30	0
32	1
34	1
36	1
38	1
40	1

```
# grade data
xx <- c(6, 8,10,12,14,16,18,20,22,24,26,28,30,32,34,36,38,40)
yy <- c(0, 0, 0, 0, 0, 1, 0, 0, 0, 0 ,1, 0, 0, 1, 1, 1, 1, 1)

#scatter plot plot(x = xx,y = yy,xlab = “Hours",ylab = “Grade",
xlim = c(0,45),ylim = c(0,1),col = "darkred",
  cex = 1.5, main = "Hours vs. Grade", pch = 16)

logistic_model <- glm(yy~xx, family=binomial(link="logit"))
summary(logistic_model)
b0    <- logistic_model$coefficients[1]
b1    <- logistic_model$coefficients[2]
phat <- round(1/(1+exp(-b0-b1*xx)), digits = 4)
O     <- round(phat/(1-phat)          , digits = 4)
df    <- data.frame(xx,yy,phat,O)
df

#scatter plot with curve
xhat <- (1:4500)/100
yhat <- 1/(1+exp(-b0-b1*xhat))
plot(x = xx,y = yy,xlab = “Hours",ylab = “Grade",
  xlim = c(0,45),ylim = c(0,1),col = "darkred",
  cex = 1.5, main = "Hours vs. Grade", pch = 16)
points(xhat,yhat,cex = .1,col = "blue")
```

9.5 Multiple Logistic Regression Analysis

Once we have $\hat{\beta}_0$ and $\hat{\beta}_1$, insert them back into

$$\hat{p}_i = \frac{1}{1+e^{-\hat{\beta}_0-\hat{\beta}_1 x_i}}$$ for estimated probabilities

and also for odds $\hat{o}_i = \dfrac{\hat{p}_i}{1-\hat{p}_i} = e^{\hat{\beta}_0+\hat{\beta}_1 x_i}$

OR for a difference in x

and for odds ratio $\hat{OR} = e^{\hat{\beta}_0+\hat{\beta}_1 x_b} / e^{\hat{\beta}_0+\hat{\beta}_1 x_a} = e^{\hat{\beta}_1 \Delta}$, $\Delta = x_b - x_a$.

$$\hat{\beta}_0 = -5.66$$
$$\hat{\beta}_1 = 0.21$$

$$\hat{OR} = e^{(0.21)(2)} = 1.5220$$

OR for a difference of x=2

Hours (x)	A (y)	$\hat{p}$	$\hat{o}$
6	0	0.0120	0.0122
8	0	0.0181	0.0184
10	0	0.0272	0.0279
12	0	0.0406	0.0423
14	0	0.0603	0.0641
16	1	0.0886	0.0972
18	0	0.1284	0.1473
20	0	0.1824	0.2232
22	0	0.2527	0.3381
24	0	0.3388	0.5124
26	1	0.4371	0.7764
28	0	0.5405	1.1764
30	0	0.6406	1.7824
32	1	0.7298	2.7008
34	1	0.8036	4.0923
36	1	0.8611	6.2008
38	1	0.9038	9.3957
40	1	0.9344	14.2365

Study 2 more hours and *OR* increases by 1.5.

9.6 Summary

Correlation

$$\text{cov}(x,y)=\frac{1}{n-1}\left[\sum XY-\frac{1}{n}\left(\sum Y\right)\left(\sum X\right)\right]$$

$$s_x^2=\frac{1}{n-1}\left[\sum X^2-\frac{1}{n}\left(\sum X\right)^2\right]$$

$$s_y^2=\frac{1}{n-1}\left[\sum Y^2-\frac{1}{n}\left(\sum Y\right)^2\right]$$

$$r=\frac{\text{cov}(x,y)}{\sqrt{s_x^2 s_y^2}}$$

Linear Regression

$$b_1=r\frac{s_y}{s_x}$$

$$b_0=\bar{Y}-b_1\bar{X}$$

$$\hat{y}=b_0+b_1x$$

Logistic Regression

$$\hat{p}=\frac{1}{1+e^{-b_0-b_1x_1-\ldots-b_px_p}}$$ logistic probability

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right)=b_0+b_1x_1+\ldots+b_px_p$$ log odds

$$\hat{O}R=e^{\hat{\beta}_1\Delta_1+\ldots+\hat{\beta}_p\Delta_p}$$ odds ratio for difference Δ_j in x_j

Questions?

Homework 9

Read Chapter 9.

Problems *, 6, 13

*Given (x,y) points $(1,1),(3,2),(2,3),(4,4)$,
a) Plot the points
b) Find r, b_0 and b_1 by hand with sums.
c) Draw the fitted regression line on the same graph as points.
d) What do b_0 and b_1 mean?

$$\hat{y} = b_0 + b_1 x$$

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = b_0 + b_1 x_1 + \ldots + b_p x_p$$