Ch. 11 - Simple Linear Regression

Objective: to describe a linear relationship between two quantitative variables using a model. The model fits a straight line to the data and can be used to make predictions on the **response variable** (y) given the **explanatory variable** (x).
e.g. Is there a relationship between age of car (explanatory variable, x) and value (response variable, y)? Hours of sleep (x) and score on a test (y)?

1 Scatterplots (ch. 2)

We can graphically examine the relationship between two quantitative variables using a **scatterplot**.

Let's examine blood fat data for 25 individuals aged between 20 to 60 years shown in the table below (D.G. Kleinbaum and L.L. Kupper, *Applied Regression Analysis and Other Multivariable Methods*). Blood fat can be used as a measurement for assessing cardiovascular health and high levels of blood fat in the body can lead to heart issues. Simple blood tests can be used to determine the blood fat level in a patient.

Subject	Weight	Age	Fat
1	84	46	354
2	73	20	190
3	65	52	405
4	70	30	263
5	76	57	451
6	69	25	302
7	63	28	288
8	72	36	385
9	79	57	402
10	75	44	365
11	27	24	209
12	89	31	290
13	65	52	346
14	57	23	254
15	59	60	395
16	69	48	434
17	60	34	220
18	79	51	374
19	75	50	308
20	82	34	220
21	59	46	311
22	67	23	181
23	85	37	274
24	55	40	303
25	63	30	244

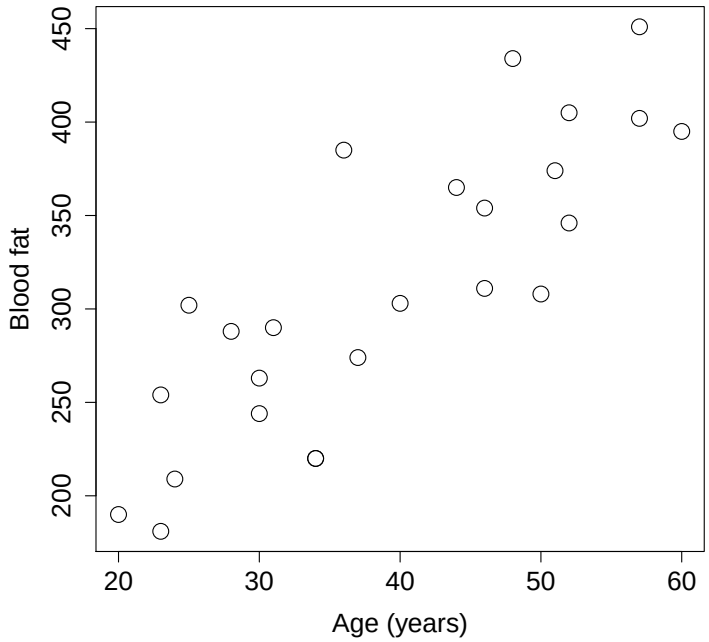


Figure 1: A scatterplot of blood fat vs. age

When we look at scatterplots we will look for:

1. Direction (positive, negative, none)

   (a) If large values of x are associated with large values of y, or as x increases the corresponding value of y tends to increase, then there is a **positive relationship** between x and y.
   e.g. the association between hours slept and score on a test likely has a weak, positive association. Generally, a well rested person is expected to score higher on a test. The relationship is weak, since there are other variables involved. Maybe a person got less sleep because they were up studying.

   (b) If small values of x are associated with large values of y, or as x increases the corresponding value of y tends to decrease, then there is a **negative relationship** between x and y.
   e.g. age of car and value may have a negative linear relationship. As cars get older, its value decreases.

2. Form (linear, curved, no clear form)

3. Scatter (strong relationship, weak or no relationship)

4. Any outliers?

2 Covariance and Correlation Coefficient (ch. 2)

We can quantify the degree of linear association between pairs of variables using the **covariance** and the **correlation coefficient** (r).

The covariance is a measure of how much two random variables change together. The sample covariance is

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

- If x and y are positively associated then when one of them is above (below) its mean, the other will also tend to be above (below) its mean. (i.e. variables tend to show similar behaviour)
  $\implies$ $Cov(x, y)$ will be large and positive

- If x and y are negatively associated then when one of them is above its mean, the other will tend to be below its mean. (i.e. variables tend to show opposite behaviour)
  $\implies$ $Cov(x, y)$ will be large and negative

- If the variables are not positively nor negatively associated, the quantity $(x_i - \bar{x})(y_i - \bar{y})$ will be positive and negative with approximately the same frequency.
  $\implies$ $Cov(x, y)$ will be small

However, $Cov(x, y)$ is dependent on the units of x and y so we standardize it so we get a measure that puts a number on the strength of the relationship but does not depend on the units we use (e.g. if x is height measured in feet and y is weight measured in pounds, the $Cov(x, y)$ would change if we converted feet into centimetres). The **correlation coefficient** gives us a numerical measurement of the strength of the linear relationship between two quantitative variables.

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

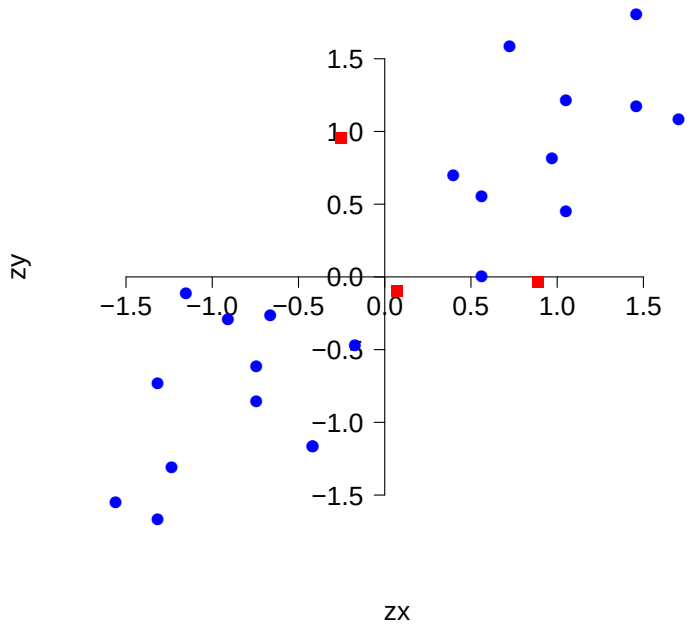


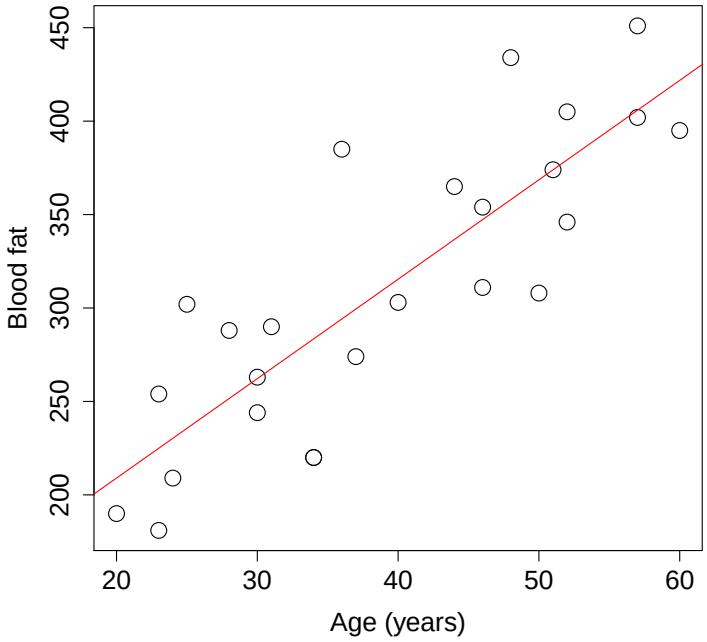
Figure 2: Scatterplot of the standardized blood fat and age. Both variables are standardized and the coordinates of a point are written as (z_x, z_y). Some points (blue circles) strengthen the impression of a positive association between height and weight. Other points (red squares) tend to weaken the positive association.

Properties of Correlation Coefficient

- $-1 \leq r \leq 1$ and has no units
- the sign of r tells us the direction of the relationship
- degree of positive correlation increases: r becomes closer to 1
  degree of negative correlation increases: r becomes closer to -1
- r close to 0 implies very weak or no linear relationship between the two variables (but this does not imply the two variables are not related in another way, e.g. non-linear relationship)

3 Linear Regression

The **regression line** is a line that best describes the relationship between x and y. Linear regression consists of finding the best-fitting straight line through the points. For our data, we could fit a regression line through the points as such:



The line that best describes the relationship between x and y is:

Regression line (y on x)
$\hat{y}$ = intercept + (slope) $\times$ (x)
where $\hat{y}$ is the predicted value

- The slope ("rise over run") tells us how much a change in y to expect for a unit increase in x. If the slope is positive (negative), y increases (decreases) with x.
- The intercept tells us the y value when x takes a value of 0. It is the point that crosses the y-axis.
- The line passes through the mean-mean point $(\bar{x}, \bar{y})$
- The value of $\hat{y}$ does not necessarily coincide with the value of y for any given x. (The line does not necessarily pass through all the points)

Residuals

The residual (e) is the difference between the observed y and the predicted value $\hat{y}$,

$$e = y - \hat{y}$$

- The linear model is obtained by minimizing the sum of the squared residuals (vertical distances from the observed points to the line). Thus we refer to the linear model as the **least squares regression line**.
- Aim: minimize $\sum_{i}^{n} e_i^2$ (see pg. 171 for proof)

The least squares estimates of the regression line is

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

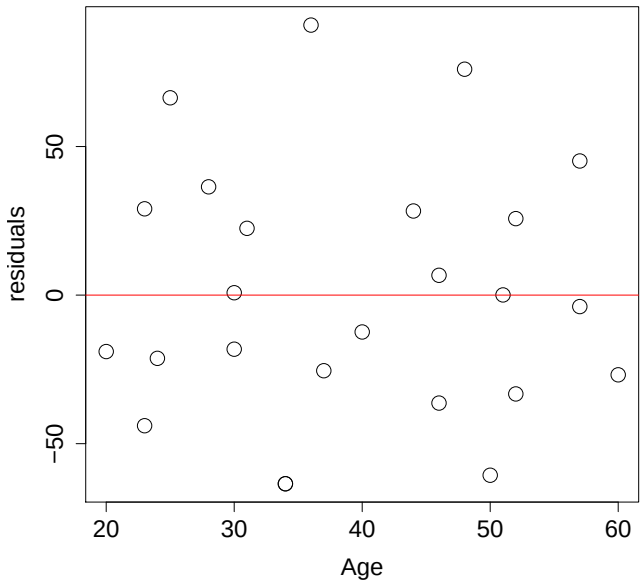
where $b_1 = \dfrac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} = \dfrac{r s_y}{s_x}$ where s_x and s_y are the standard deviations of x and y respectively and r is correlation

$$b_0 = \frac{\sum y_i - b_1 \sum x_i}{n} = \bar{y} - b_1 \bar{x}$$

Using our blood fat example, can you work out b_0 and b_1 yourself?
Ans: $b_0 = 102.6,\ b_1 = 5.32$

Example 1. Predict the blood fat level for an 26 year old individual.

Subject	Age x	Bloodfat y	Predicted $\hat{y} = 102.6 + 5.32x$	Residuals $e = y - \hat{y}$
1	46	354	347.33	6.67
2	20	190	208.99	-18.99
3	52	405	379.25	25.75
4	30	263	262.20	0.80
5	57	451	405.85	45.15
6	25	302	235.59	66.41
7	28	288	251.55	36.45
8	36	385	294.12	90.88
9	57	402	405.85	-3.85
10	44	365	336.68	28.32
11	24	209	230.27	-21.27
12	31	290	267.52	22.48
13	52	346	379.25	-33.25
14	23	254	224.95	29.05
15	60	395	421.82	-26.82
16	48	434	357.97	76.03
17	34	220	283.48	-63.48
18	51	374	373.93	0.07
19	50	308	368.61	-60.61
20	34	220	283.48	-63.48
21	46	311	347.33	-36.33
22	23	181	224.95	-43.95
23	37	274	299.44	-25.44
24	40	303	315.40	-12.40
25	30	244	262.20	-18.20



4 Inference for Regression

Now we want to know what can the regression model tell us beyond the 25 patients in this study? To do this we need to draw inferences. Suppose we draw a random sample and fit a linear model. We want to know whether the coefficients in our model are just about the data we have or whether they tell us something more fundamental? This is where hypothesis testing comes in.

If we draw several samples from a population and fit a regression line, each sample and least squares regression (and thus intercept and slope) will be different even if its drawn from the same population.

We imagine the underlying true linear relationship:

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

where ε represents model errors

We use Greek letters to denote the **coefficients** (intercept and slope) since these are parameters. If we had all the values in the population, we could find the slope and intercept of this idealized regression line by using least squares. The line can't actually match all the values in the population. Some ys lie above or below the line so for each data point, the model makes an error. The errors are random and can be positive, negative or zero.

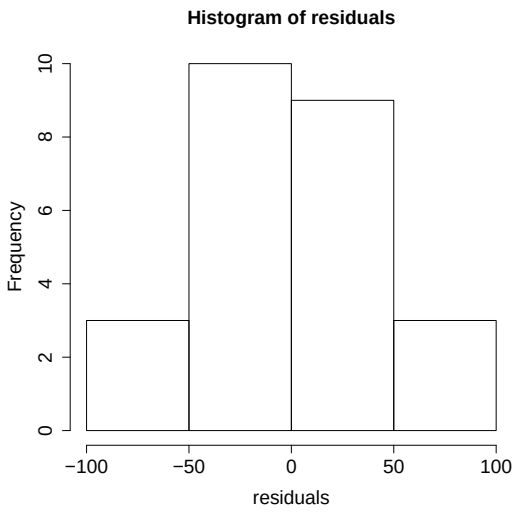
We estimate the true linear relationship with

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

This linear regression line is constructed from data. The slope of this regression model is a statistic and has a sampling distribution, which we can model. We have two models: the linear model, which describes the relationship between blood fat and age as well as a model for the sampling distribution. Whenever we use models we need to check assumptions and conditions.

Assumptions & Conditions

1. Linearity Assumption
    - If the true relationship is not linear and we use a straight line to fit the data our entire analysis is useless
    - Check this assumption with a scatterplot
2. Independence Assumption
    - The errors in the true regression model (ε's) must be independent
    - Check residual plot for patterns, trends or clumping, any of these would suggest a failure of independence.
3. Equal Variance Assumption (homoscedasticity)
    - the variability of y should be the same for all values of x
    - Check the spread around the line of your scatterplot is nearly constant (check for fan shapes or tendencies of the variation to grow or shrink in one part of the scatterplot). Or look at a residuals vs. predicted values ($\hat{y}$) plot and look for patterns, which would violate this assumption.
4. Normal Population Assumption
    - Errors around the idealized regression line at each value of x follow a Normal model.
    - Histogram of the residuals should be nearly normal



We can fit the regression model as long as the linearity assumption is satisfied. We need the other 3 conditions to use the sampling distribution model for inference.

Hypothesis Testing and Confidence Intervals Involving β_1

$$\frac{b_1 - \beta_1}{SE(b_1)} \sim t_{n-2}$$

$$SE(b_1) = \frac{s_e}{s_x\sqrt{n-1}}$$

$$s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$$

A $(1-\alpha) \times 100$ confidence interval for β_1 is:

$b_1 \pm t^*_{n-2} \times SE(b_1)$
where the critical value comes from the
t-distribution with $n-2$ degrees of freedom

In simple linear regression, we often wish to test:

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

(although $H_A : \beta_1 > 0$ or $H_A : \beta_1 < 0$ are also possibilities. In some instances, we test $H_0 : \beta_1 = a$ where a is a number).

A slope of 0 implies the mean value of y for any value of x is the same in which case it

means that x is not useful in predicting y.

What about the intercept?
The intercept usually isn't interesting so most hypothesis tests and confidence intervals for regression are about the slope.

$$\frac{b_0 - \beta_0}{SE(b_0)} \sim t_{n-2}$$

$$\text{where } SE(b_0) = s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x - \bar{x})^2}}$$

Here is some output from R for our blood fat data. You should be able to read output from R such as what is shown below.

```
> fit1 <- lm(Fat ~ Age, data = newDat)
> summary(fit1)

Call:
lm(formula = Fat ~ Age, data = newDat)

Residuals:
    Min      1Q  Median      3Q     Max
-63.478 -26.816  -3.854  28.315  90.881

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.5751    29.6376   3.461  0.00212 **
Age           5.3207     0.7243   7.346 1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.46 on 23 degrees of freedom
Multiple R-squared:  0.7012,  Adjusted R-squared:  0.6882
F-statistic: 53.96 on 1 and 23 DF,  p-value: 1.794e-07
```

Example 2. Consider the following data obtained in a simple linear regression study

x	3.27	1.26	4.55	0.86	4.07	4.79	3.25
y	16.67	19.93	14.65	17.48	18.18	13.58	15.70

(a) Find the estimated regression line

(b) Predict y when $x = 3$

(c) Conduct the hypothesis test, $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ with $\alpha = 0.01$. Is there any evidence that x is useful in predicting y?

(d) Redo part c using $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 < 0$ with $\alpha = 0.05$.

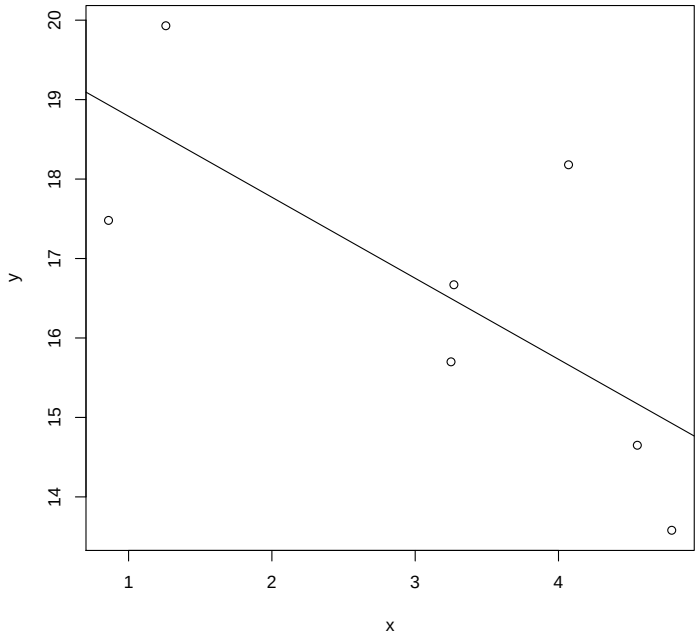
(e) Redo part c using $H_0 : \beta_1 = -0.5$ vs. $H_A : \beta_1 \neq -0.5$ with $\alpha = 0.05$.

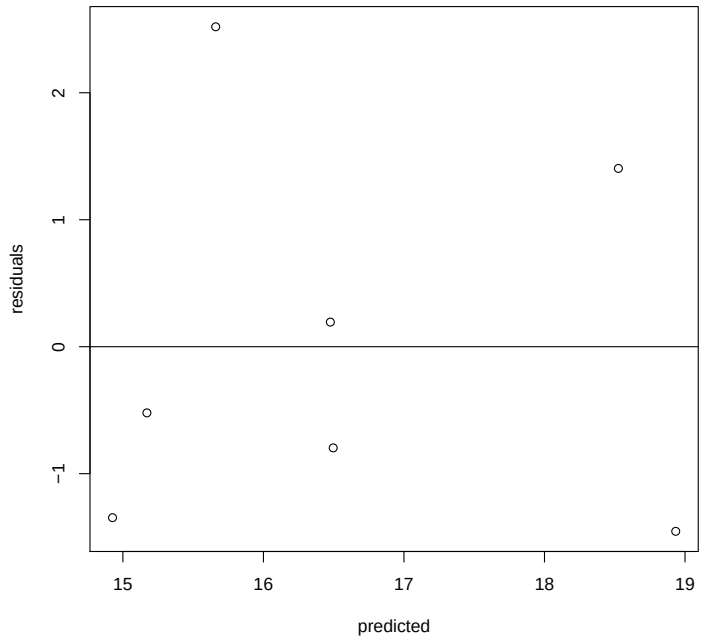
(f) Find a 95% confidence interval for β_1

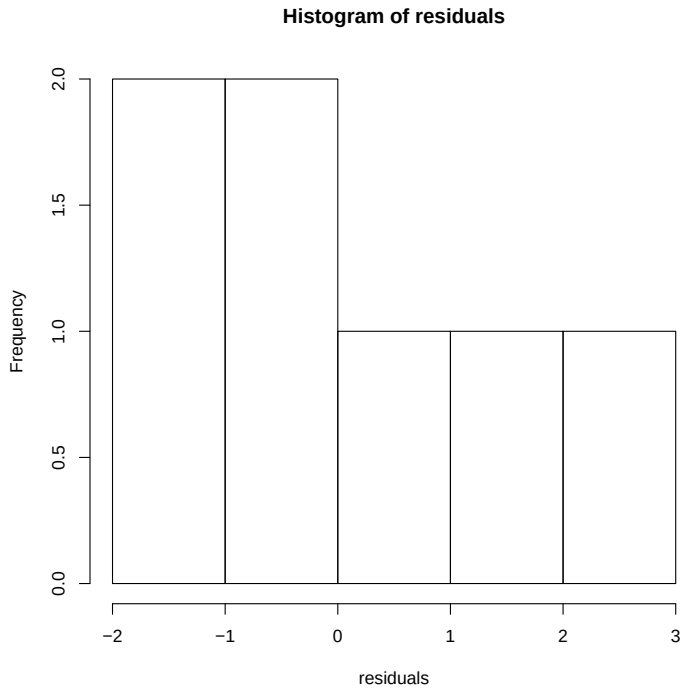
Solution: Instead of doing it by hand, let's use R to help get all the parameter estimates and standard error calculations.

```
> x <- c(3.27, 1.26, 4.55, 0.86, 4.07, 4.79, 3.25)
> y <- c(16.67, 19.93, 14.65, 17.48, 18.18, 13.58, 15.70)
> plot(x,y)
> fit <- lm (y~x)
> abline(fit)
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
      1       2       3       4       5       6       7
 0.1938  1.4041 -0.5209 -1.4538  2.5196 -1.3462 -0.7966

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.8108     1.4841  13.349 4.22e-05 ***
x            -1.0197     0.4289  -2.377   0.0634 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.624 on 5 degrees of freedom
Multiple R-squared:  0.5306,  Adjusted R-squared:  0.4367
F-statistic: 5.652 on 1 and 5 DF,  p-value: 0.06337
```

```
> plot(predict(fit), resid(fit), xlab = "predicted", ylab = "residuals")
> abline(h=0)
> hist(resid(fit), xlab = "residuals", main = "Histogram of residuals")
```




residuals
2
1
0
−1
15
16
17
18
19
predicted


Histogram of residuals
Frequency
2.0
1.5
1.0
0.5
0.0
−2
−1
0
1
2
3
residuals