Introduction to Linear Regression

Session 4

presented by:
Imen Hammami

OxPal Research Fellowship Series 2021

UNIVERSITY OF OXFORD

February, 2020

What will this session tell you?

- To understand correlation.
- To be able to fit and interpret the coefficients of a simple linear regression model.
- To be able to interpret ANOVA tables and use them to compare group means.
- To be able to check the assumptions of a linear regression model.

Your new friend

Jane Superbrain 2.0

- She steals the brains of top statisticians.
- She appears in red boxes to tell you really important things.

Content

Layout

1. Correlation
   - What is correlation?
   - Confounding variables
   - Pearson's correlation coefficient
2. Simple linear regression
   - The regression line
   - One continuous independent variable
   - One dichotomous independent variable
3. Comparing more than two means
   - The linear regression model
   - One-Way Analysis of Variance
   - ANOVA table and F-test
4. Regression diagnostics
   - Properties of the data
   - Properties of the residuals

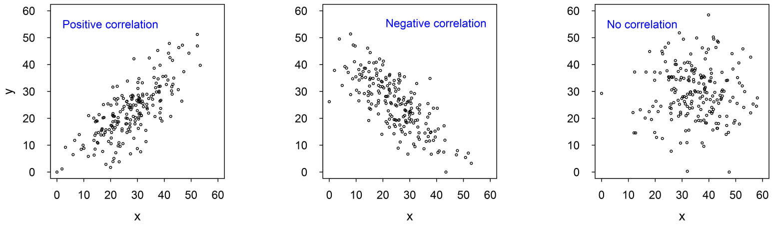
What is correlation?

Correlation

- A general term used to describe how to variables vary together.
- Imprecise, used loosely to describe a general relationship.

Types of correlation

- *Positive correlation:* an increase in one variable is accompanied by an increase in another variable.
- *Negative correlation:* an increase in one variable is accompanied by a decrease in another variable.
- *No correlation:* there is no relationship between the two variables.



Spurious correlation

Figure 1 shows the association between ice cream sales and swimming pool deaths.

Example based on simulated data. For more examples, http://www.tylervigen.com/spurious-correlations

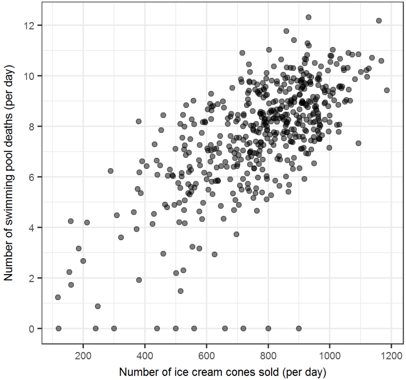


Figure 1: Daily swimming pool deaths versus number of ice cream cones sold.

Spurious correlation (Cont.)

What else could be causing this apparent relationship 🤔 ?

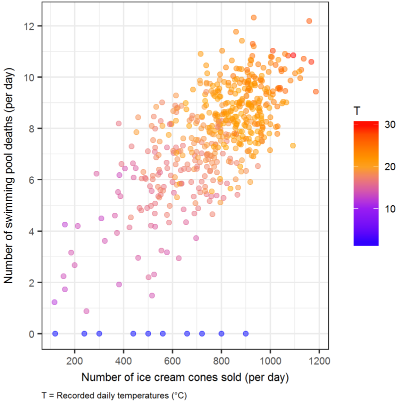


Figure 1 (Cont.): Daily swimming pool deaths versus number of ice cream cones sold.

Correlation does not imply causation!

After taking into account the recorded daily temperatures, the spurious relationship is eliminated, as we would intuitively expect.

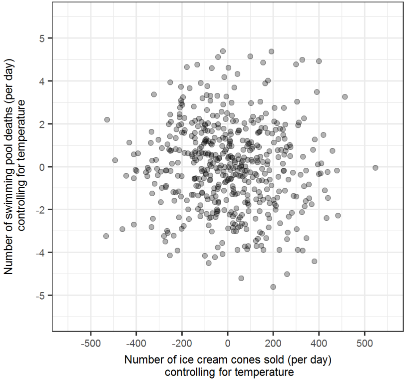


Figure 2: Daily swimming pool deaths versus number of ice cream cones sold (controlling for temperature).

Important points

Jane Superbrain 2.0

- What a correlation does not tell you is why two variables tend to vary together.
- A correlation might be coincidental, or it might be a result of both patterns being caused by a third factor (a confounder).

Pearson's correlation coefficient

Definition

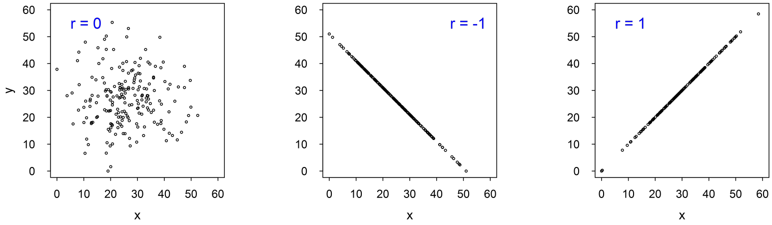
- The degree of association between two continuous variables is measured by a correlation coefficient, denoted by r.
- r is a measure of the strength of a linear association on a scale that varies from - 1 to $+1$.

Assumption

The association between the two continuous variables is linear (i.e. one variable increases or decreases a fixed amount for a unit increase or decrease in the other).

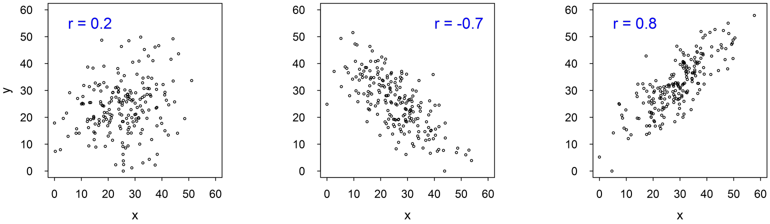
What does this correlation coefficient tell you?

- If $r = 0$, there is no linear relationship between X and Y.
- If $r = -1$,there is a perfect *negative* linear relationship between X and Y.
- If $r = 1$, there is a perfect *positive* linear relationship between X and Y.



What does this correlation coefficient tell you?

- The closer r is to 0, the weaker the linear relationship.
- The closer r is to -1, the stronger the negative linear relationship.
- the closer r is to 1, the stronger the positive linear relationship.

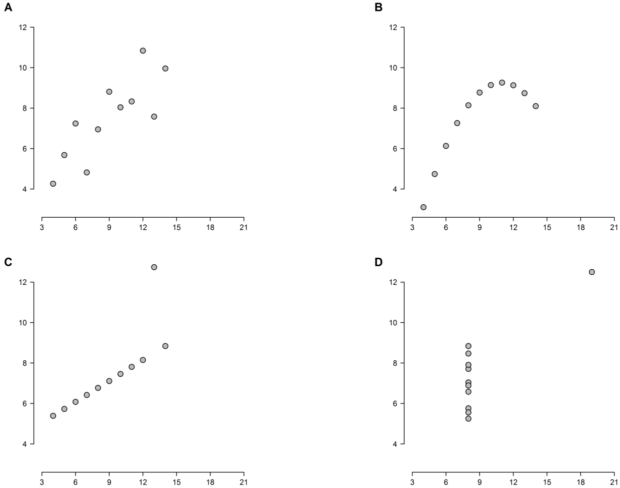


R code & output

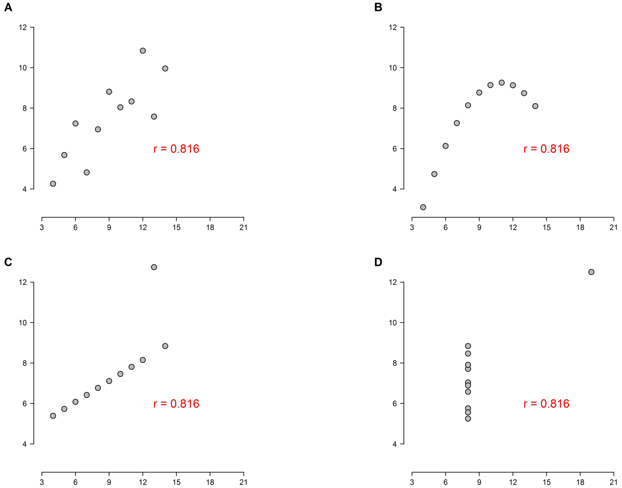
```
#Open data
white.data <- read.csv(file="<mypath>\\Whitehall_like-
    Baseline.csv", header=TRUE, na.strings = ".")
corrtable<-cor(white.data[, 11:17], use="complete.obs")
round(corrtable,2)
```

```
       HDLC  LDLC  APOB APOA1  CHOL   CRP  VitD
HDLC   1.00 -0.07 -0.43  0.82  0.11 -0.09  0.09
LDLC  -0.07  1.00  0.70  0.05  0.89 -0.06  0.05
APOB  -0.43  0.70  1.00 -0.11  0.75 -0.04  0.01
APOA1  0.82  0.05 -0.11  1.00  0.30 -0.10  0.09
CHOL   0.11  0.89  0.75  0.30  1.00 -0.09  0.07
CRP   -0.09 -0.06 -0.04 -0.10 -0.09  1.00 -0.06
VitD   0.09  0.05  0.01  0.09  0.07 -0.06  1.00
```

Don't be fooled! [1]



Don't be fooled! [1]



Advantages and disadvantages

Jane Superbrain 2.0

- ☺ Pearson's correlation coefficient is a useful summary statistic.
- ☺ It can provide good insight for further investigation.
- ☹ It can only be used when the relationship between two variables is linear.
- ☹ It is very sensitive to clustering and outliers.

Layout

Regression terminology

In linear regression, we use is a straight line to model the relationship between two variables X and Y.

Y is called the '*dependent variable*' or the '*response variable*', which is the measurement of interest that we want to estimate/predict.

X is called the '*independent variable*' or the '*explanatory variable*', which is the variable that we believe can be used to explain some of the variation in the response variable.

Regression terminology (Cont.)

We can describe the regression line knowing only the slope and the intercept.

'*The intercept*' β_0 is where the line cuts the y-axis. This is the expected value of Y when X equals 0.

'*The slope*' β_1 is the expected change in Y for a one unit increase in X.

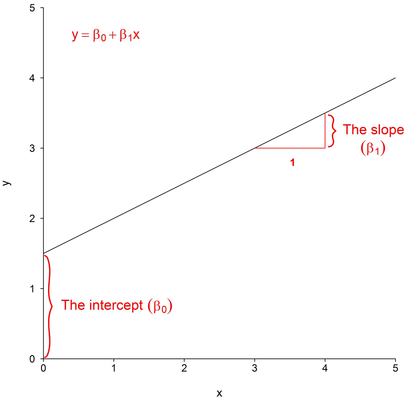


Figure 3: Graph of a straight line.

Regression terminology (Cont.)

Jane Superbrain 2.0

In simple linear regression, the equation of the regression line is given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$\hat{y}_i$ is the estimated mean value of Y for a given value of X.

β_0 and β_1 are the population parameters.
Estimates of these parameters are denoted by putting a “hat” over the Greek corresponding letter.

Example

Inheritance of height (Pearson and Lee, 1903) [2]

- Data were collected on the height of 1375 mothers in the United Kingdom under the age of 65 and one of their adult daughters over the age of 18.
- The objective was to examine the relationship between the heights of the mothers and the heights of their daughters.

The scatterplot

How would you describe the relationship between the heights of the mothers and the heights of their daughters 🤔 ?

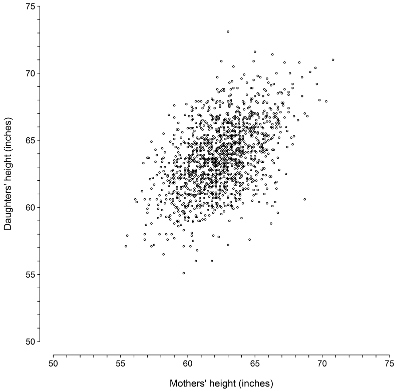


Figure 4: Daughter's height versus mother's height.

The regression line

If we were to assume a linear relationship between mothers' height and daughters' height, we can draw the regression line that describes that relationship.

We can use the regression line to estimate the average height of daughters with mothers of a given height.

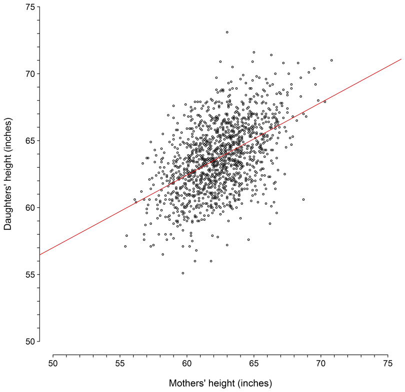


Figure 4 (Cont.): Daughter's height versus mother's height.

What does this line tell you?

The line goes through the point of averages (62.5; 63.8).

Mothers of average height tend to have daughters of average height.

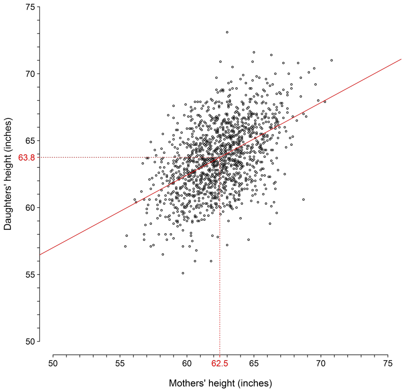


Figure 4 (Cont.): Daughter's height versus mother's height.

What does this line tell you?

The average height of daughters whose mothers are 58 inches tall is 61.3 inches.

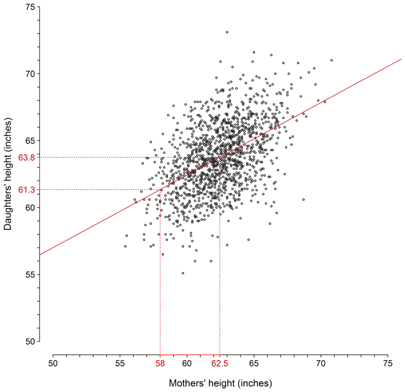


Figure 4 (Cont.): Daughter's height versus mother's height.

The residuals

The *observed* value of the height of particular daughters with mothers of a given height will typically not equal the *estimated* value (indicated by the regression line) for that given height.

The vertical distances between the observed values and the estimated values are known as *residuals*, denoted e.

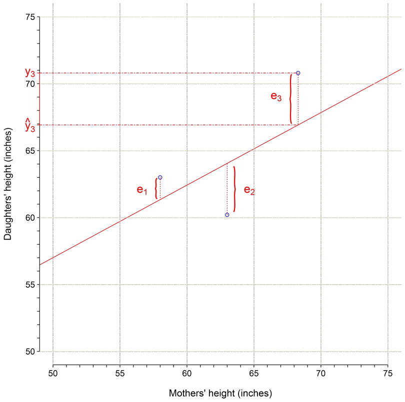


Figure 5: The regression line for Pearson's data.

The residuals (Cont.)

Data points fall both above and below the line, yielding both positive and negative differences.

The *regression line* is the line that results in the least amount of (squared) difference between the observed data points and the line.

If we sum positive and negative differences, they tend to cancel each other out, so we square them before adding them up. This method is known as 'Ordinary Least Squares' (OLS) regression.

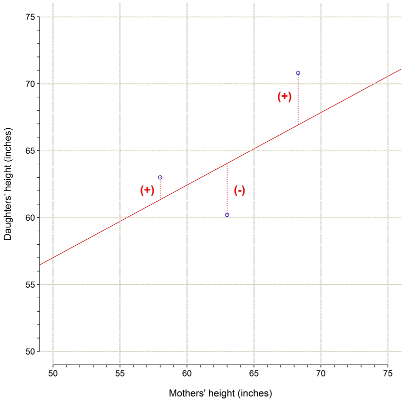


Figure 5 (Cont.): The regression line for Pearson's data.

Let's fit a linear regression model

Working example 1

1. Examine the relationship between systolic blood pressure (SBP) and age in the urban China workers dataset.
2. Using R, fit a linear regression model with SBP as the dependent variable and age as the independent variable.
3. Write down the final model and interpret its coefficients.

Examine the scatterplot

How would you describe the relationship between SBP and age?

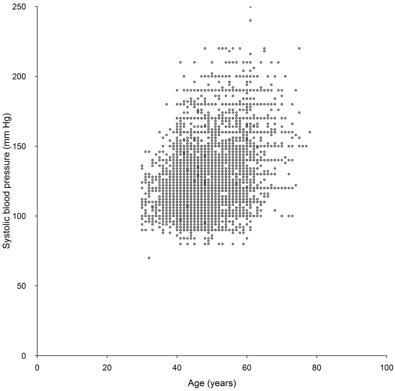


Figure 6: SBP versus age in the urban China workers data.

R code

R function

lm(*depvar*∼*indepvars*)

```
#Open data
shanghai <- read.csv(file="shanghai_psc.csv", header=TRUE, na.strings = ".")
m0 <- lm(sbp ~ ages, data=shanghai)
summary(m0)
```

R output

```
Call:
lm(formula = sbp ~ ages, data = shanghai)

Residuals:
    Min      1Q  Median      3Q     Max
-55.836 -14.318  -2.895  10.917 112.070
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.05956    1.50563   49.19   <2e-16 ***
ages         1.04706    0.03075   34.06   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 20.78 on 9015 degrees of freedom
Multiple R-squared:  0.114,    Adjusted R-squared:  0.1139
F-statistic:  1160 on 1 and 9015 DF,  p-value: < 2.2e-16
```

Interpretation

```
Call:
lm(formula = sbp ~ ages, data = shanghai)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.05956    1.50563   49.19   <2e-16 ***
ages         1.04706    0.03075   34.06   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The $\hat{\beta}$s are estimates of the population parameters so they have standard errors (*se*). In the R output, the *se* is denoted Std. Error.

The null hypothesis for the *t*-test states that the β is equal to zero, and the alternative hypothesis states that β is not equal to zero.

$$t\text{-}value = \frac{Coef.}{Std.Err.}$$

Interpretation

```
Call:
lm(formula = sbp ~ ages, data = shanghai)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.05956    1.50563   49.19   <2e-16 ***
ages         1.04706    0.03075   34.06   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Equation of the regression line

$$\hat{y}_i = \hat{\beta}_0 \quad + \hat{\beta}_1 \quad x_i$$

$$s\hat{b}p = 74.06 + 1.05 \; ages$$

Interpretation

Equation of the regression line

$$\hat{sbp} = 74.06 + 1.05\ ages$$

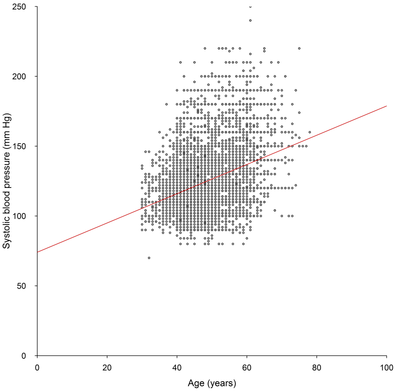


Figure 6 (Cont.): SBP versus age in the urban China workers data.

Interpretation

Equation of the regression line

$$\hat{sbp} = 74.06 + 1.05 \; ages$$

- $\hat{\beta}_0 = 74.06$ is the mean SBP when age equals zero.

- $\hat{\beta}_1 = 1.05$ represents the amount of change in SBP relative to a one unit change in age.
- If we compare two participants whose ages differ by 1 year, we would expect their SBP to differ by approximately 1.05 mm Hg (with the person with the higher age having the higher SBP as the slope is positive).

Let's fit a linear regression model

Working example 2

1. Examine the relationship between systolic blood pressure (SBP) and sex in the urban China workers data.
2. Using R, fit a linear regression model with SBP as the dependent variable and sex as the independent variable.
3. Write down the final model and interpret its coefficients.

Examine the scatterplot

How would you describe the relationship between SBP and sex?

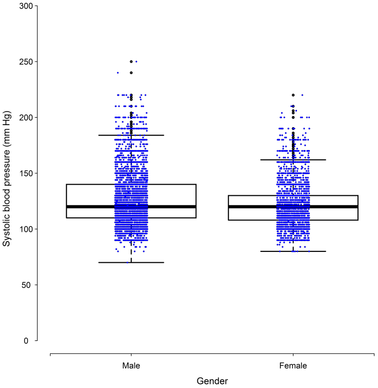


Figure 7: SBP versus sex in the urban China workers data.

Factor variables

Indicator (or dummy) variables are binary variables i.e. variables that take only two values.

The value 1 indicates the presence of some characteristic or attribute. The value 0 indicates the absence of that same characteristic or attribute.

The dichotomous variable sex which is coded as sex $= 0$ for males and sex $= 1$ for females could be defined using two dummy variables `sex.f0` $= 1$ if sex $= 0$, 0 otherwise; and `sex.f1` $= 1$ if sex $= 1$, 0 otherwise.

Factor variables (Cont.)

In R, you can use the `factor()` function to specify indicators for each level (category) of the categorical variable e.g. `factor(sex)`.

The level indicator variables are ‘virtual’ -not created in your dataset, saving lots of space.

R code

R function

`lm(`*depvar*~*indepvars*`)`

```
#Open data
shanghai <- read.csv(file="shanghai_psc.csv", header=TRUE, na.
    strings = ".")
shanghai$sex.f=factor(shanghai$sex)
m1 <- lm(sbp ~ sex.f, data=shanghai)
summary(m1)
```

R output

```
Call:
lm(formula = sbp ~ sex.f, data = shanghai)

Residuals:
    Min      1Q  Median      3Q     Max
-56.026 -16.026  -6.026  11.974 123.974

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 126.0261     0.2790 451.744  < 2e-16 ***
sex.f1       -3.9789     0.5006  -7.948 2.13e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 22 on 9015 degrees of freedom
Multiple R-squared:  0.006958,  Adjusted R-squared:  0.006848
F-statistic: 63.17 on 1 and 9015 DF,  p-value: 2.128e-15
```

R output (Cont.)

```
Call:
lm(formula = sbp ~ sex.f, data = shanghai)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 126.0261     0.2790 451.744  < 2e-16 ***
sex.f1       -3.9789     0.5006  -7.948 2.13e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

In this example, the '*sex*' variable is coded as $male = 0$ and $female = 1$. In the R output, `sex.f1` indicates the '*female*' category.

Equation of the regression line

$$\hat{sbp} = 126.03 - 3.98 \; \texttt{sex.f1}$$

$$\hat{sbp} = 126.03 - 3.98 \; female$$

Interpretation

Equation of the regression line

$$\hat{sbp} = 126.03 - 3.98 \ female$$

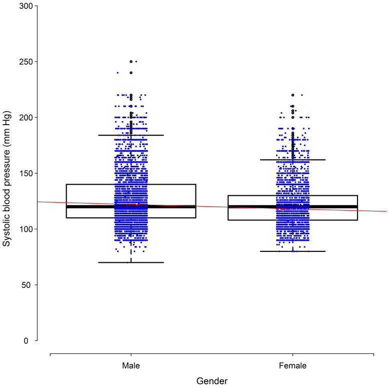


Figure 7 (Cont.): SBP versus sex in the urban China workers data.

Interpretation

Equation of the regression line

$$\hat{sbp} = 126.03 - 3.98\ female$$

$\hat{\beta}_0 = 126.03$ is the mean SBP in males ($\hat{sbp}_{male} = 126.03$ mm Hg).

```
aggregate(sbp~sex, mean, data=shanghai)
```

```
  sex      sbp
1   0 126.0261
2   1 122.0471
```

Interpretation

Equation of the regression line

$$\hat{sbp} = 126.03 - 3.98 \; female$$

The mean SBP in females is

$$\hat{sbp}_{female} = 126.03 - 3.98 \times 1 = 122.05 \text{ mm Hg}$$

```
aggregate(sbp~sex, mean, data=shanghai)
```

```
  sex      sbp
1   0 126.0261
2   1 122.0471
```

Interpretation

Equation of the regression line

$$\hat{sbp} = 126.03 - 3.98\ female$$

The regression coefficient ($\hat{\beta}_1 = -3.979$) associated with *female* represents the expected difference in mean SBP levels for '*female*' as compared to the reference category '*male*'.

$$\hat{sbp}_{female} - \hat{sbp}_{male} = 122.04 - 126.02 = -3.98 \text{ mm Hg}$$

Interpretation

The mean value of SBP in females is significantly different from the mean value of SBP in males ($\hat{\beta}_1 = -3.98$, $p < 0.001$).

```
Call:
lm(formula = sbp ~ sex.f, data = shanghai)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 126.0261     0.2790 451.744  < 2e-16 ***
sex.f1       -3.9789     0.5006  -7.948 2.13e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Interpretation

Performing a simple linear regression with one dichotomous independent variable is equivalent to performing a two-sample *t*-test.

```
t.test(sbp ~ sex, data = shanghai, var.equal=TRUE)
```

```
        Two Sample t-test

data:  sbp by sex
t = 7.9478, df = 9015, p-value = 2.128e-15
alternative hypothesis: true difference in means is not equal to
     0
95 percent confidence interval:
 2.997559 4.960271
sample estimates:
mean in group 0 mean in group 1
       126.0261        122.0471
```

Layout

Let's fit a linear regression model

Working example

1. Examine the relationship between systolic blood pressure (SBP) and body mass index (BMI) categories in the urban China workers data.

   BMI groups are < 18.5; $18.5 < 25$; $25 < 30$; and $\geq 30\ kg/m^2$.
2. Using R, fit a linear regression model with SBP as the dependent variable and BMI groups as the independent variable.
3. Interpret the coefficients of your model.

Examine the scatterplot

Are there any differences in the means of SBP across BMI categories?

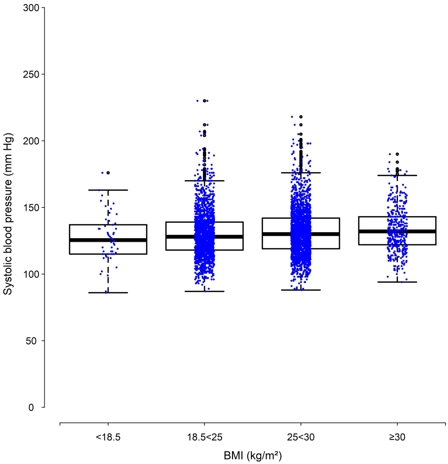


Figure 7: SBP versus BMI groups in the Whitehall data.

R code

```
Whitehall <- read.csv(file="<mypath>\\Whitehall.csv", header=
    TRUE, na.strings = ".", sep=",")
Whitehall$bmigp.f <- factor(Whitehall$bmigp)
Whitehall$bmigp.f <- relevel(Whitehall$bmigp.f, ref=2)
m0 <- lm(sbp ~ bmigp.f, data=Whitehall)
summary(m0)
```

R output

```
Call:
lm(formula = sbp ~ bmigp.f, data = Whitehall)

Residuals:
     Min       1Q   Median       3Q      Max
-43.366  -12.239   -1.570    9.634  100.430

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 129.5695     0.4134 313.389  < 2e-16 ***
bmigp.f1     -2.9095     2.5088  -1.160 0.246221
bmigp.f3      1.7966     0.5638   3.187 0.001449 **
bmigp.f4      3.6698     0.9926   3.697 0.000221 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 17.5 on 4297 degrees of freedom
Multiple R-squared:  0.00488,   Adjusted R-squared:  0.004185
F-statistic: 7.024 on 3 and 4297 DF,  p-value: 0.0001041
```

Interpretation

```
Call:
lm(formula = sbp ~ bmigp.f, data = Whitehall)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   129.5695     0.4134 313.389  < 2e-16 ***
bmigp.f1       -2.9095     2.5088  -1.160 0.246221
bmigp.f3        1.7966     0.5638   3.187 0.001449 **
bmigp.f4        3.6698     0.9926   3.697 0.000221 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Intercept $= 129.57$ mm Hg is the mean value of SBP in participants with BMI $\geq 18.5 < 25\ kg/m^2$.

Interpretation (Cont.)

```
aggregate(sbp~bmigp, mean, data=Whitehall)
```

```
   bmigp      sbp
1   <18.5 126.6600
2 18.5<25 129.5695
3   25<30 131.3661
4     30+ 133.2394
```

Interpretation (Cont.)

```
Call:
lm(formula = sbp ~ bmigp.f, data = Whitehall)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 129.5695     0.4134 313.389  < 2e-16 ***
bmigp.f1     -2.9095     2.5088  -1.160 0.246221
bmigp.f3      1.7966     0.5638   3.187 0.001449 **
bmigp.f4      3.6698     0.9926   3.697 0.000221 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The regression coefficient for a given BMI category represents the estimated difference in mean SBP levels for that category as compared to the reference group.

Interpretation (Cont.)

```
Call:
lm(formula = sbp ~ bmigp.f, data = Whitehall)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 129.5695     0.4134 313.389  < 2e-16 ***
bmigp.f1     -2.9095     2.5088  -1.160 0.246221
bmigp.f3      1.7966     0.5638   3.187 0.001449 **
bmigp.f4      3.6698     0.9926   3.697 0.000221 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

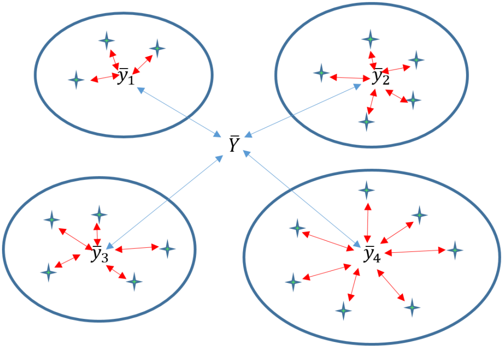
For example, the estimated difference in mean SBP between participants in the top BMI category ($\geq 30\ kg/m^2$) and participant in the reference BMI category ($\geq 18.5 < 25\ kg/m^2$) is 3.67 mm Hg.

R output

```
Call:
lm(formula = sbp ~ bmigp.f, data = Whitehall)

Residuals:
    Min      1Q  Median      3Q     Max
-43.366 -12.239  -1.570   9.634 100.430

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 129.5695     0.4134 313.389  < 2e-16 ***
bmigp.f1     -2.9095     2.5088  -1.160 0.246221
bmigp.f3      1.7966     0.5638   3.187 0.001449 **
bmigp.f4      3.6698     0.9926   3.697 0.000221 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 17.5 on 4297 degrees of freedom
Multiple R-squared:  0.00488,   Adjusted R-squared:  0.004185
F-statistic: 7.024 on 3 and 4297 DF,  p-value: 0.0001041
```

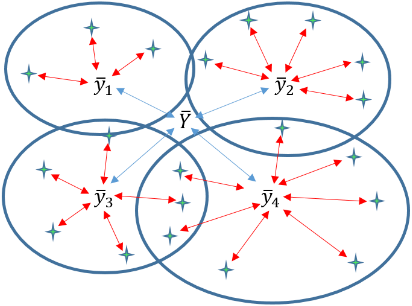
Motivating example

Examine the differences in mean SBP ($\hat{y}$) across the four BMI groups.



Motivating example (Cont.)

Examine the differences in mean SBP ($\hat{y}$) across the four BMI groups.



ANOVA

What is ANOVA?

In its simplest form, the ANalysis Of VAriance (ANOVA) provides a statistical test of whether or not the means of several groups are equal, and therefore generalises the t-test to more than two groups.

How it works?

ANOVA compares the variation between groups to the variation within groups. If the variation between groups is greater than the variation within groups, then there is evidence that the means are not equal across groups.

Assumptions

- The dependent variable is normally distributed in each of the groups.
- The variances across the groups are equal.

The ANOVA table

Source	Sum of Squares (SS)	Degrees of of Freedom	Mean Square (MS)	F-value
Model/Group	Between-group variation (MS_{group})	$k-1$	$MS_{group} = \frac{SS_{group}}{k-1}$	$\frac{MS_{group}}{MS_E}$
Residuals	Within-group variation (MS_E)	$n-k$	$MS_E = \frac{SS_E}{n-k}$	
Total	Overall variation	$n-1$		

Table 1: Summary ANOVA

k is the number of groups and n is the number of observations.

The Sum of Squares (SS) is the sum of the squared differences (a measure of variation).
The Mean Sum of Squares (MS) is a measure of variation per degree of freedom (MS=SS/df).

The F-test

Null hypothesis: All the k population group means are equal.
Alternative hypothesis: At least one of the k population means differs from all of the other.

- If the variances are similar, the F-value will be approximately 1.
- Large F-values are evidence of differences in means across groups.
- The F-distribution with $(k-1,\ n-k)$ df is used to get a P value (R will do it for you).
- When $k=2$, the F-test is equivalent to performing a t-test.

R code & output

```
Whitehall <- read.csv(file="<mypath>\\Whitehall.csv", header=
    TRUE, na.strings = ".", sep=",")
Whitehall$bmigp.f <- factor(Whitehall$bmigp)
m1 <- aov(sbp ~ bmigp.f, data=Whitehall)
summary(m1)
```

```
                Df  Sum Sq Mean Sq F value   Pr(>F)
bmigp.f          3    6451  2150.4   7.024 0.000104 ***
Residuals     4297 1315528   306.2
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The large F-value ($F(3, 4297) = 7.02$) means that the between-group variation (the model variance) exceeds the within-group variation (the residual variance) by a substantial amount.

We can conclude that not all the group means are equal ($p = 0.0001$).

Important points

Jane Superbrain 2.0

- The F-test associated with the ANOVA tables tests whether the means of all groups are equal.
- Just because the F-test indicates that there is a difference somewhere does not mean that all pairwise comparisons are significant.
- The F-test does not tell you about the differences between specific pairs of means.
- To determine which means are significantly different, you must compare all pairs -but be careful of increasing Type I error (use *Bonferroni* correction).

Layout

Context

We use the sample data to estimate the value of the parameters in the population.

We calculate an estimate of how well it represents the population such as a standard error or confidence interval.

We also test hypotheses about these parameters by computing test statistics and P values.

Sources of bias

Jane Superbrain 2.0

- Things that bias the parameter estimates.
- Things that bias standard errors and confidence intervals.
- Things that bias test statistics and P values.

(A short list of) Regression diagnostics

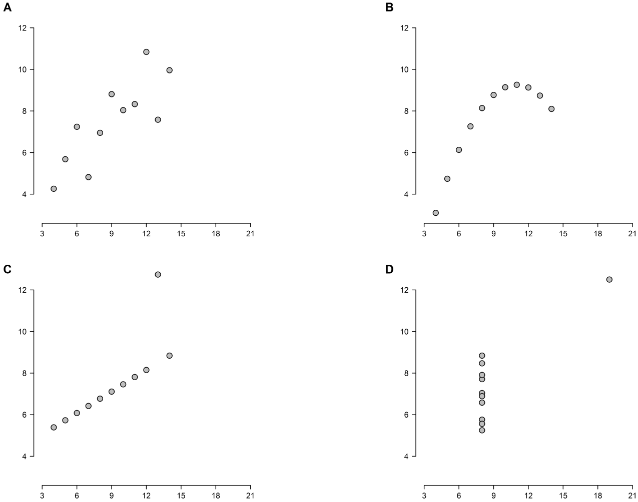
Spotting unusual and influential data ✓

Checking linearity ✓

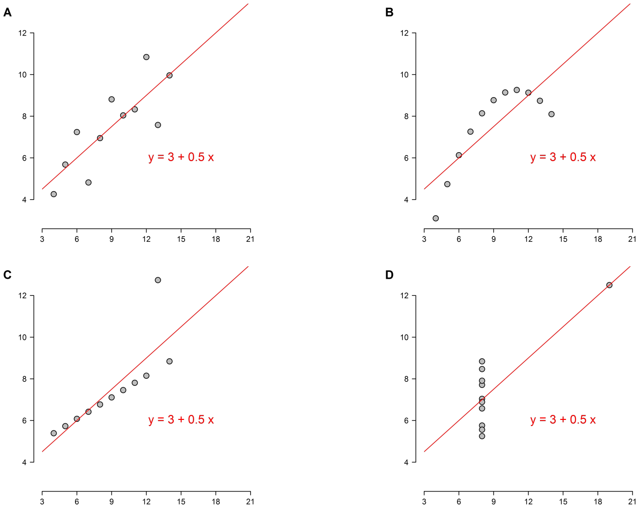
Checking normality of residuals ✓

Checking homoscedasticity ✓

Data visualisation matters [1]



Data visualisation matters (Cont.) [1]



Residual diagnostics

Residuals could show how poorly a model represents data.

They could reveal unexplained patterns in the data by the fitted model.

Using this information, you can check if model assumptions are met.

OLS Assumptions

Assumptions

The residuals are independent (uncorrelated); normally distributed and have constant variance (homoscedasticity).

Useful R functions

- `resid()` to extract the residuals from the fitted model.
- `fitted()` to extract fitted values ($\hat{y}_i$) from the fitted model.

Checking for normality

How?

You can use a histogram of the residuals.

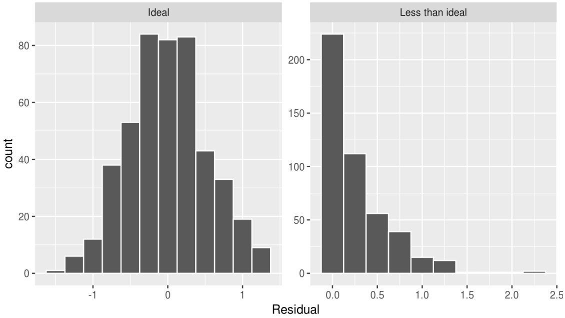


Figure 8: Histogram of the residuals.

Checking for normality (Cont.)

It is often hard to tell if a distribution is normal from just a histogram. Use $Q-Q$ plots!

A $Q-Q$ plot of the residuals displays the residuals versus their expected values when the distribution is normal.

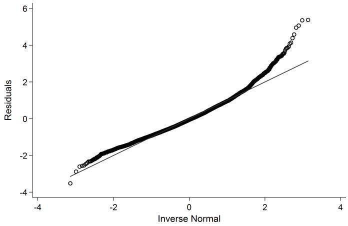


Figure 9: $Q-Q$ plot of the residuals.

Residuals versus fitted values

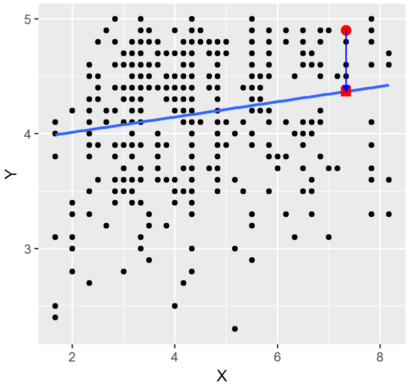


Figure 10: Observed values versus exposure.

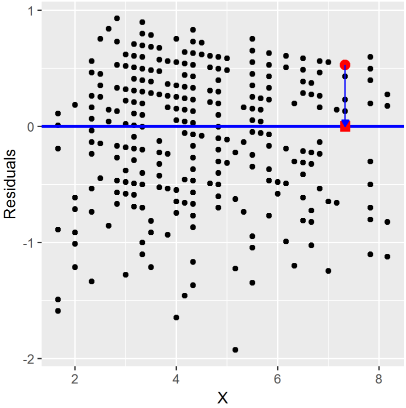


Figure 11: Residuals versus exposure.

Residuals versus fitted values (Cont.)

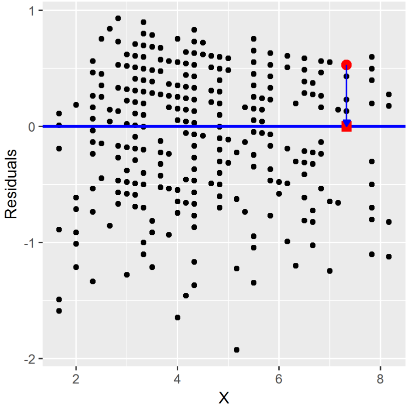


Figure 11 (Cont.): Residuals versus exposure.

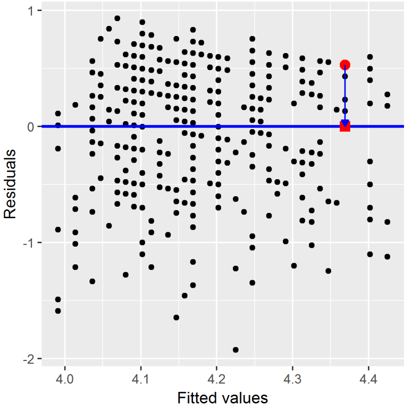


Figure 12: Residuals versus fitted values.

Checking for equal variance

We can check that the residuals do not vary systematically with the fitted values by inspecting the plot of the residuals against the fitted values.

We are looking for any evidence that residuals vary in a clear pattern.

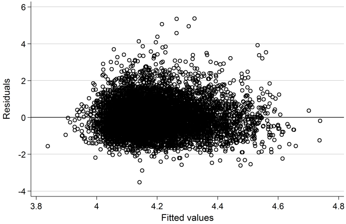
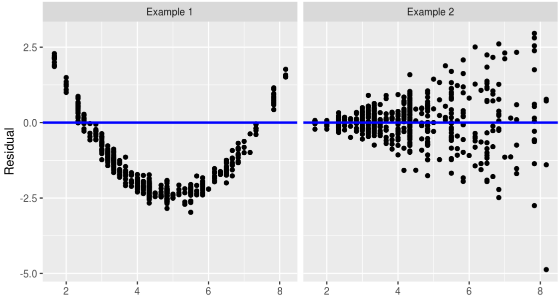


Figure 13: A graph of the residuals versus the fitted values.

Examples of violation of the OLS assumptions

Curvature in the pattern of the residuals in Example 1 suggests a violation of the linearity assumption.

The increasing variation in the residuals in Example 2 suggest a violation of the homoscedasticity assumption.



Important points

Jane Superbrain 2.0

- In a well-fitted model, there should be no pattern to the residuals plotted against the fitted values.
- Any pattern whatsoever indicates a violation of the OLS assumptions.

What to do if assumptions are violated?

- Checking for mistakes in your data.
- Assessing the impact of influential observations on the results.
- Using transformations.
- Using more advanced methods.

What did this session tell you?

- To understand correlation.
- To be able to fit and interpret the coefficients of a simple linear regression model.
- To be able to check the assumptions of a linear regression model.
- To be able to interpret ANOVA tables and use them to compare group means.

[1] F. J. Anscombe, "Graphs in statistical analysis," *The American Statistician*, vol. 27, no. 1, pp. 17–21, 1973.

[2] K. Pearson and A. Lee, "On the laws of inheritance in man: I. inheritance of physical characters," *Biometrika*, vol. 2, no. 4, pp. 357–462, 1903.