TM-CM02 Biostatistics for Public Health
Lecture 3
Multiple linear regression

Kiesha Prem

Saw Swee Hock School of Public Health, National University of Singapore

Simple linear regression

Dependent variable

Independent variable

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Random error

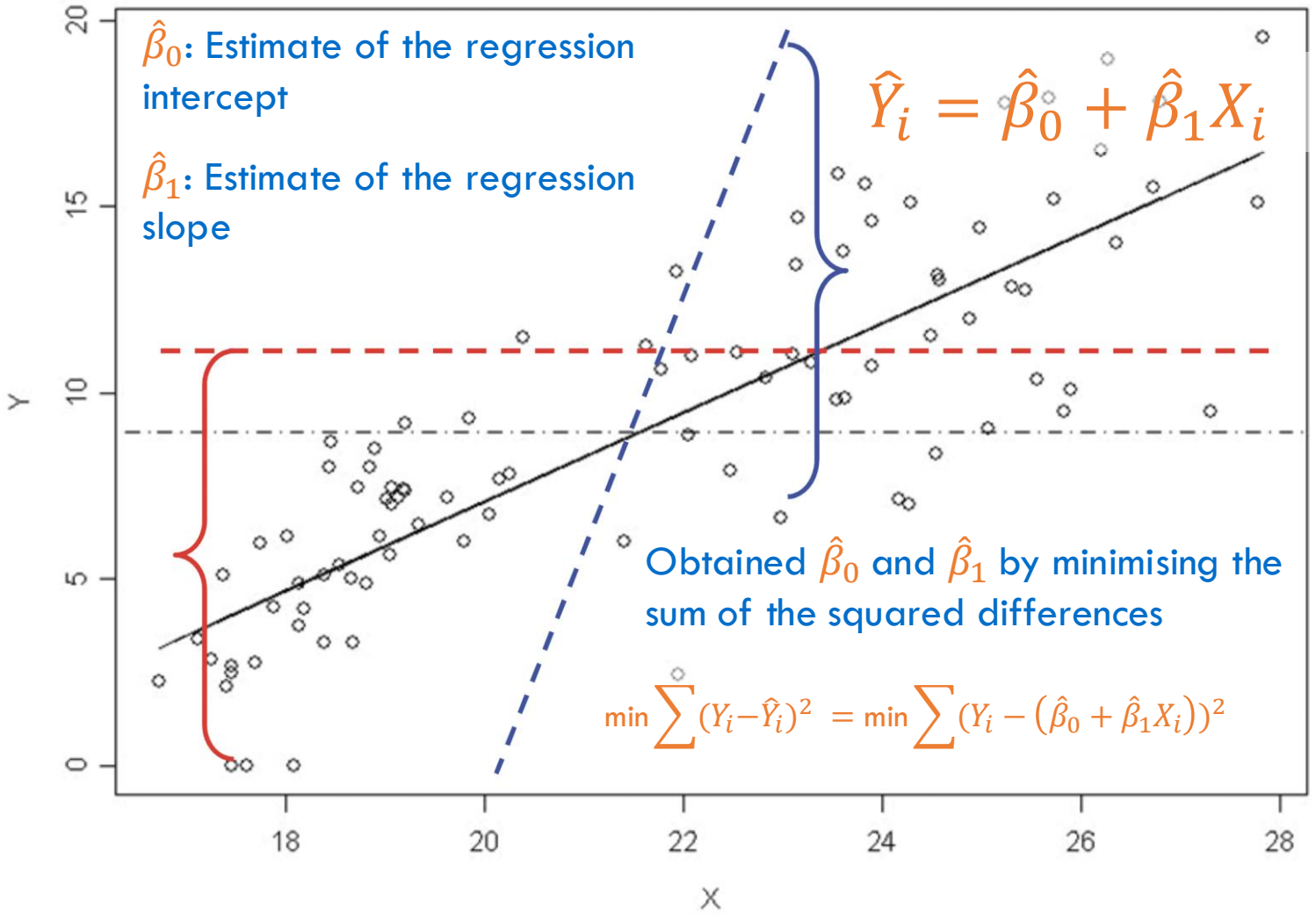
Intercept

Slope (regression) coefficient

Minimising error

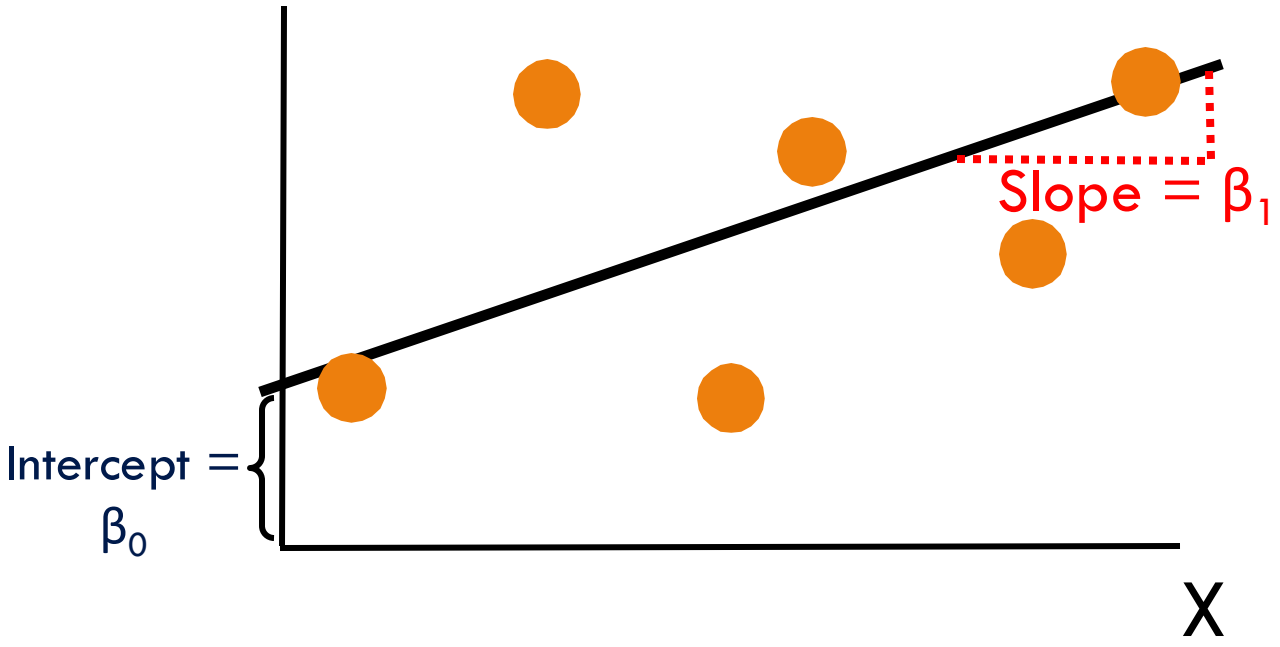
Least Squares Method

The equation provides an estimate of the population regression line.



Interpreting regression coefficients

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$



$\hat{\beta}_0$ least-square estimate of the regression intercept

- estimated mean value of Y when x = 0

$\hat{\beta}_1$ least-square estimate of the regression slope

- estimated change in y when x changes by one unit

Inference about the slope

NUS National University of Singapore | Saw Swee Hock School of Public Health

T-test

Null hypothesis

H_0: $\hat{\beta}_1 = 0$

no linear relationship

Alternative hypothesis

H_1: $\hat{\beta}_1 \neq 0$

linear relationship may exist

Test

$$t = \frac{(\hat{\beta}_1 - 0)}{SE(\hat{\beta}_1)}$$

with t distribution of d.f. = n - 2

Hypothesis tests

or confidence intervals

- To test the significance or "contribution" of an independent variable (*X*) to the dependent variable (*Y*)

- Is there a linear relationship between X and Y?

We are testing for zero slope:

If $\hat{\beta}_1 = 0$: then the X does not influence the value of Y

Determining the fit of the model

- We need to determine how well y is predicted by $\hat{y}(x_i) = \widehat{\beta_0} + \widehat{\beta_1} x_i$
- How much variability in y can be accounted for by the linear regression where x is the predictor?
- The R^2 indicates the proportion of the total variation explained by the model.
  - If x and y are independent (i.e., no relationship), then $R^2 = 0$.
  - If x and y are perfectly correlated (i.e., perfect relationship), then $R^2 = 1$.
- Hence, R^2 indicates how well the model is doing in explaining the response and ranges from 0 to 1.

```
> m_ldl_age = lm(chp$ldl ~ chp$age)
> summary(m_ldl_age)

Call:
lm(formula = chp$ldl ~ chp$age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.13569 -0.10771 -0.08370 -0.05917  0.34480

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -1.5666710  0.0167344  -93.62 <0.0000000000000002 ***
chp$age      0.0714981  0.0002448  292.05 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1714 on 9998 degrees of freedom
Multiple R-squared:  0.8951,    Adjusted R-squared:  0.8951
F-statistic: 8.529e+04 on 1 and 9998 DF,  p-value: < 0.00000000000000022
```

$$R^2 = 1 - \frac{SSE}{SST}$$

R-Squared & F-test

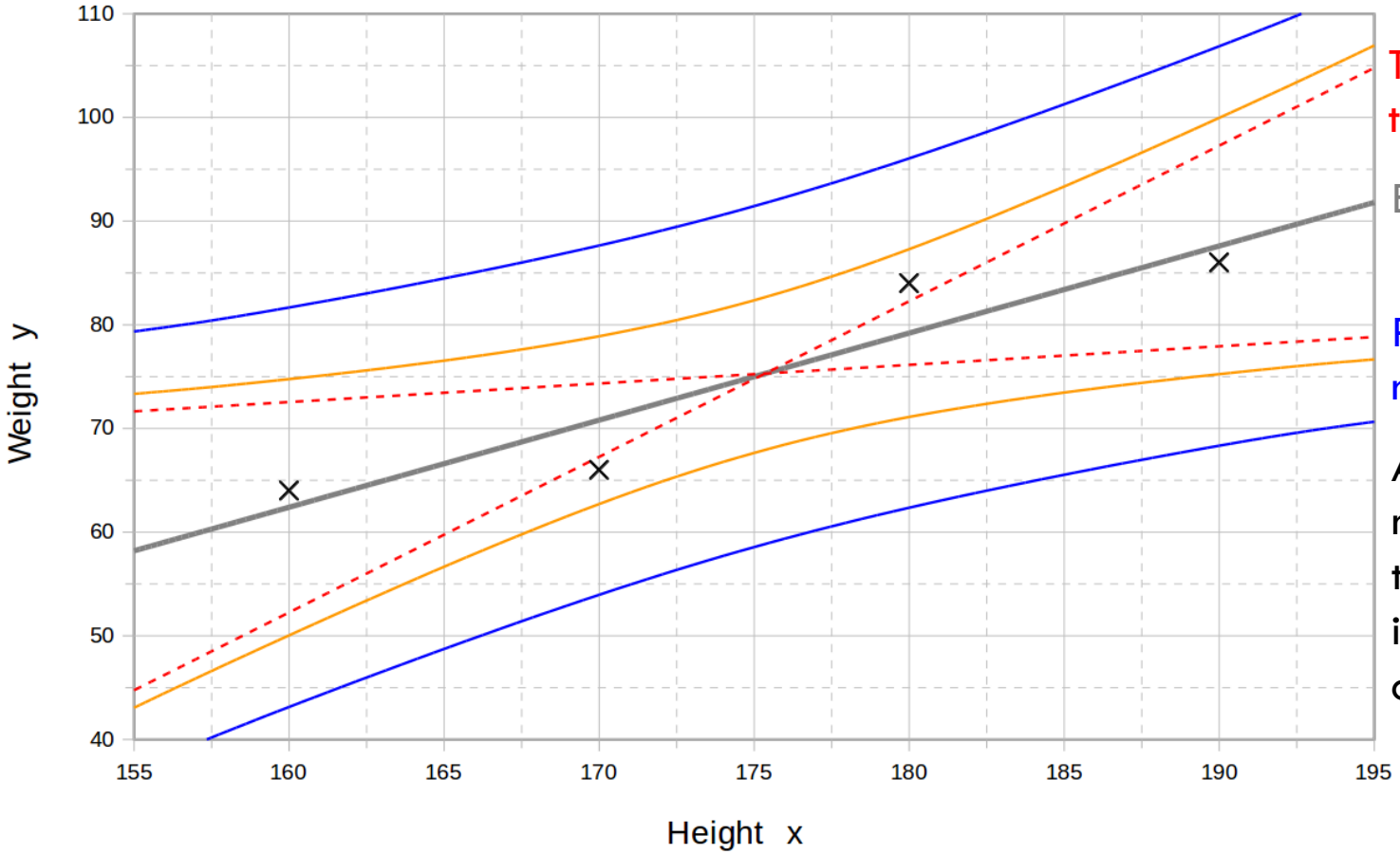
Determining the fit of the model

- We need to determine how well y is predicted by $\hat{y}(x_i) = \widehat{\beta_0} + \widehat{\beta_1} x_i$
- How much variability in y can be accounted for by the linear regression where x is the predictor?
- The R^2 indicates the proportion of the total variation explained by the model.
  - If x and y are independent (i.e., no relationship), then $R^2 = 0$.
  - If x and y are perfectly correlated (i.e., perfect relationship), then $R^2 = 1$.
- Hence, R^2 indicates how well the model is doing in explaining the response and ranges from 0 to 1.

```
> m_ldl_age = lm(chp$ldl ~ chp$age)
> summary(m_ldl_age)

Call:
lm(formula = chp$ldl ~ chp$age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.13569 -0.10771 -0.08370 -0.05917  0.34480

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -1.5666710  0.0167344  -93.62 <0.0000000000000002 ***
chp$age      0.0714981  0.0002448  292.05 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1714 on 9998 degrees of freedom
Multiple R-squared:  0.8951,    Adjusted R-squared:  0.8951
F-statistic: 8.529e+04 on 1 and 9998 DF,  p-value: < 0.00000000000000022
```

Age explains 89.5% of the variation in LDL cholesterol. From the F-statistic, the p-value is <0.001, indicating the regression model with age is significantly better than a model with intercept term only.

Confidence interval of mean outcome



The dotted lines represent the two extreme lines.

Best fit line

Prediction for a new measurement

A prediction interval that represents the uncertainty may accompany the point prediction. Such intervals tend to expand rapidly as the values of the independent variable(s) move outside the range covered by the observed data.

Confidence interval and prediction interval

The 95% CI of the slope

$$\widehat{\beta_1} \pm t_{n-2,0.025}\,\mathrm{SE}(\widehat{\beta_1})$$

The 95% CI of the intercept

$$\widehat{\beta_0} \pm t_{n-2,0.025}\,\mathrm{SE}(\widehat{\beta_0})$$

The 95% CI for the the predicted value at a specific x-value, x

$$\hat{y} \pm t_{n-2,0.025}\,\mathrm{SE}(\hat{y})$$

where

$$\mathrm{SE}(\hat{y}) = s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \text{ and } s = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\widehat{y_i})^2}{n-2}}$$

The 95% prediction interval for a predicted value $\hat{y}$ at a specific x-value, x

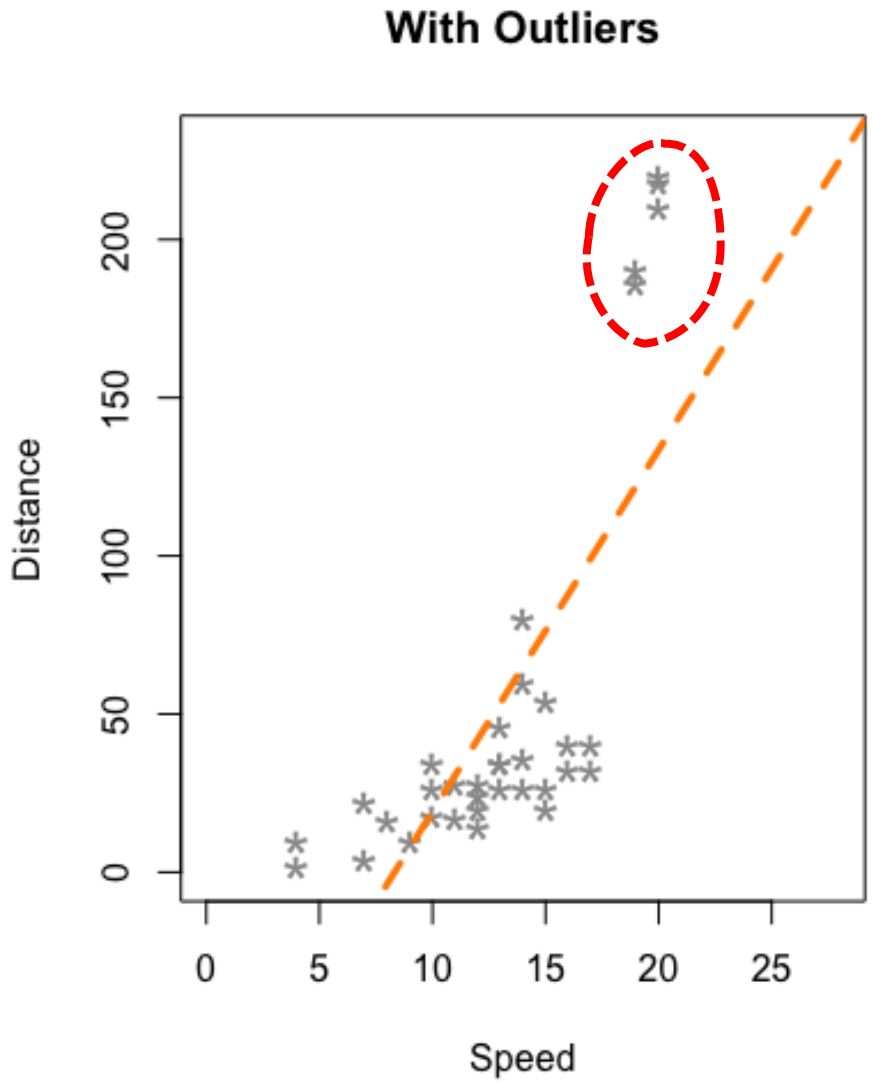
$$\hat{y} \pm t_{n-2,0.025}\,\mathrm{SE}_{pred}(\hat{y})$$

where

$$\mathrm{SE}_{pred}(\hat{y}) = s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

Reflecting both the uncertainty in the mean response and the natural variability of the data.

Problematic observations



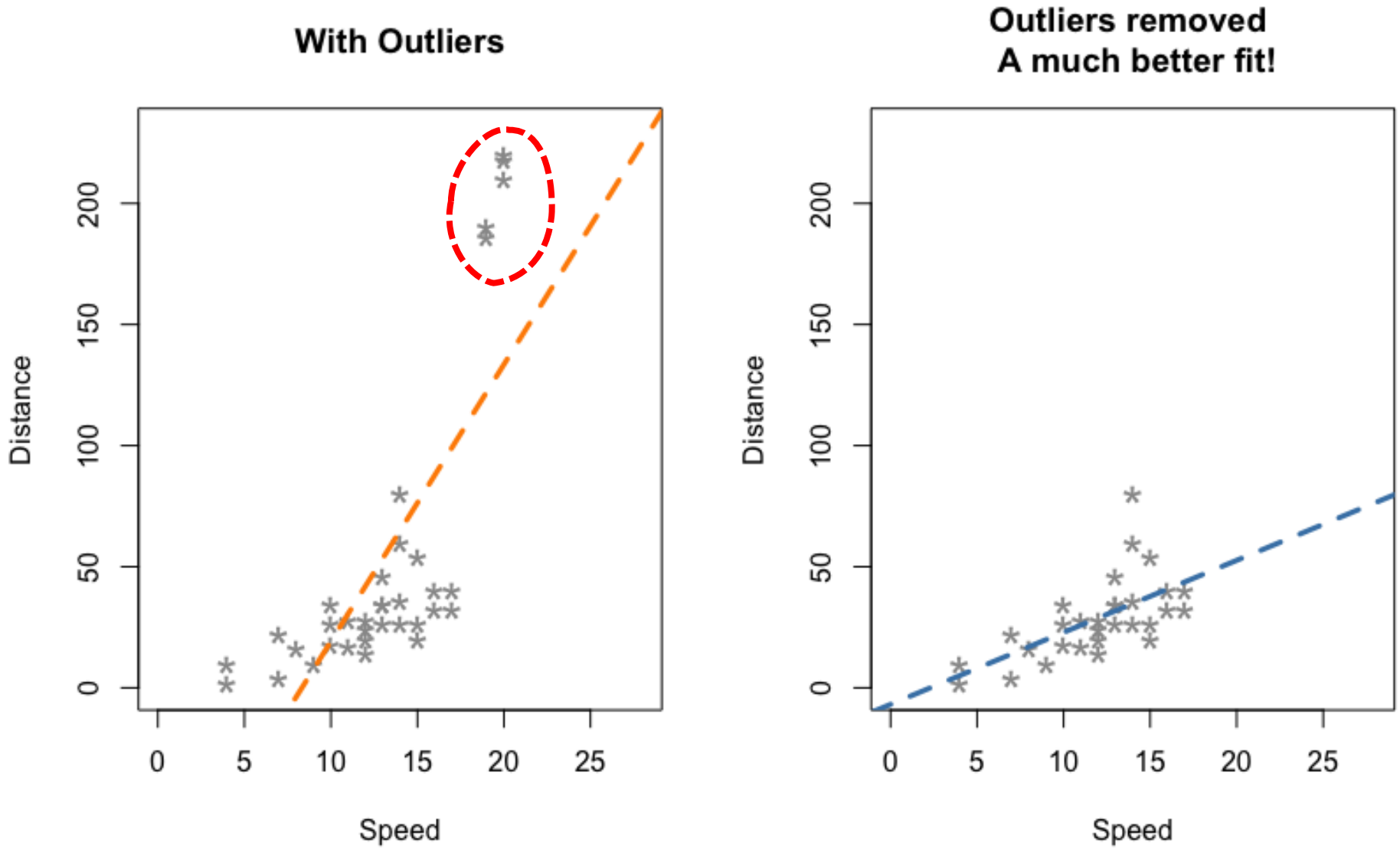
Ideally, each observation should have the same influence on the regression analysis.

If an observation has a significantly greater influence than the rest, it can potentially bias the results.

Some potential characteristics of an influential observation:

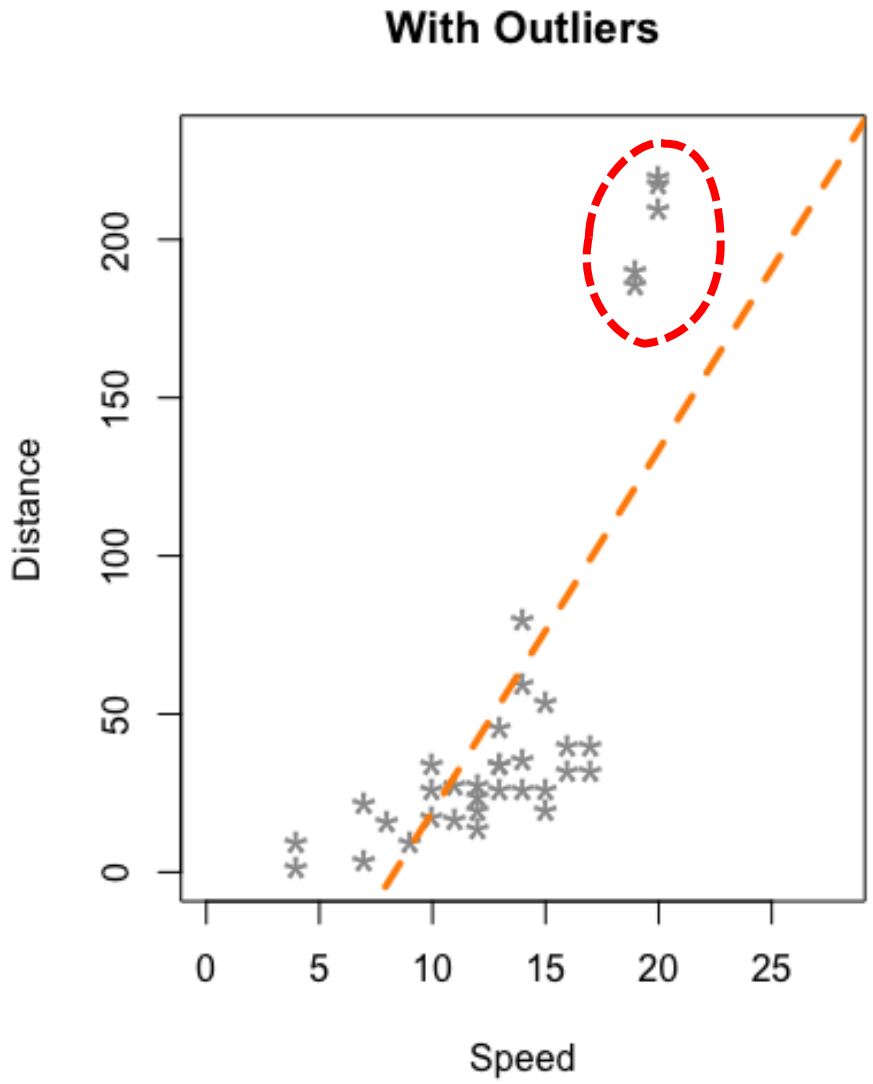
- it has a large absolute residual and
- it is far from mean x.

Problematic observations



The intercept and slope are different

Problematic observations

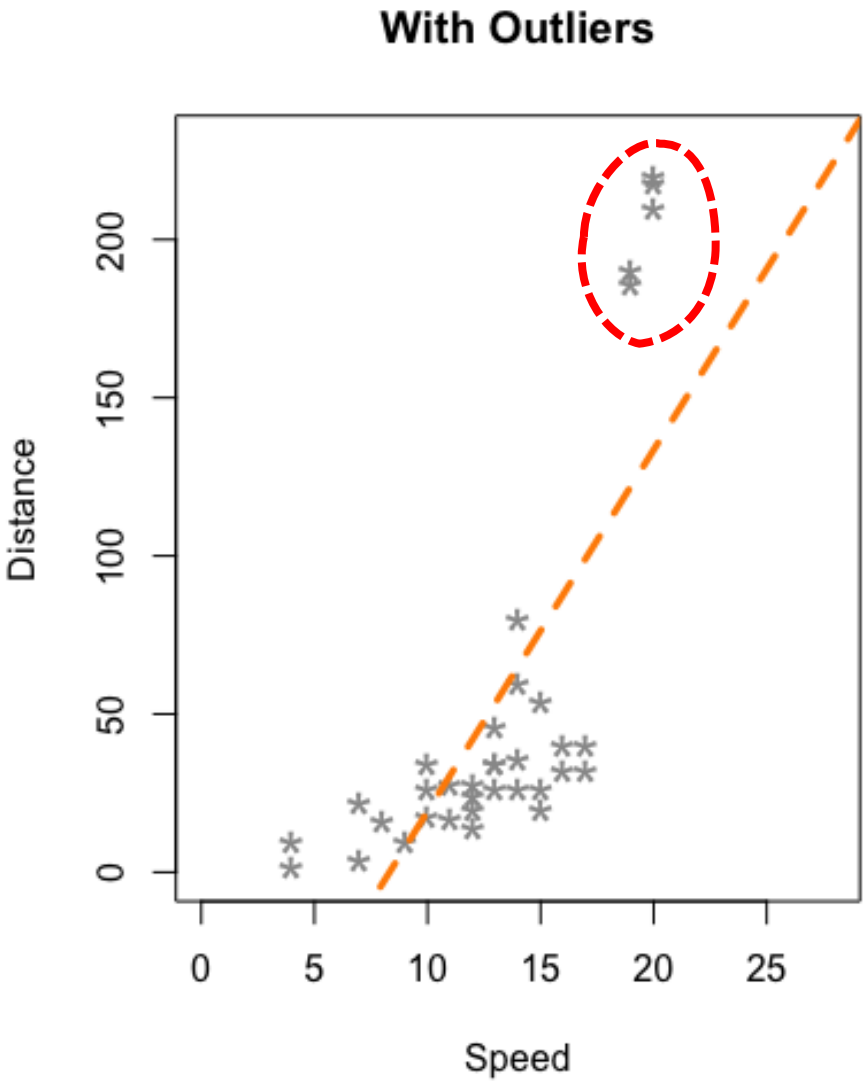


Leverage is used to quantity the potential of a predictor value influencing the results.
The leverage ranges from 1/n to 1.

The leverage is lowest when the observation is close to the mean of the predictor.

Observation with high leverage is influential if the associated residual is large.

Problematic observations

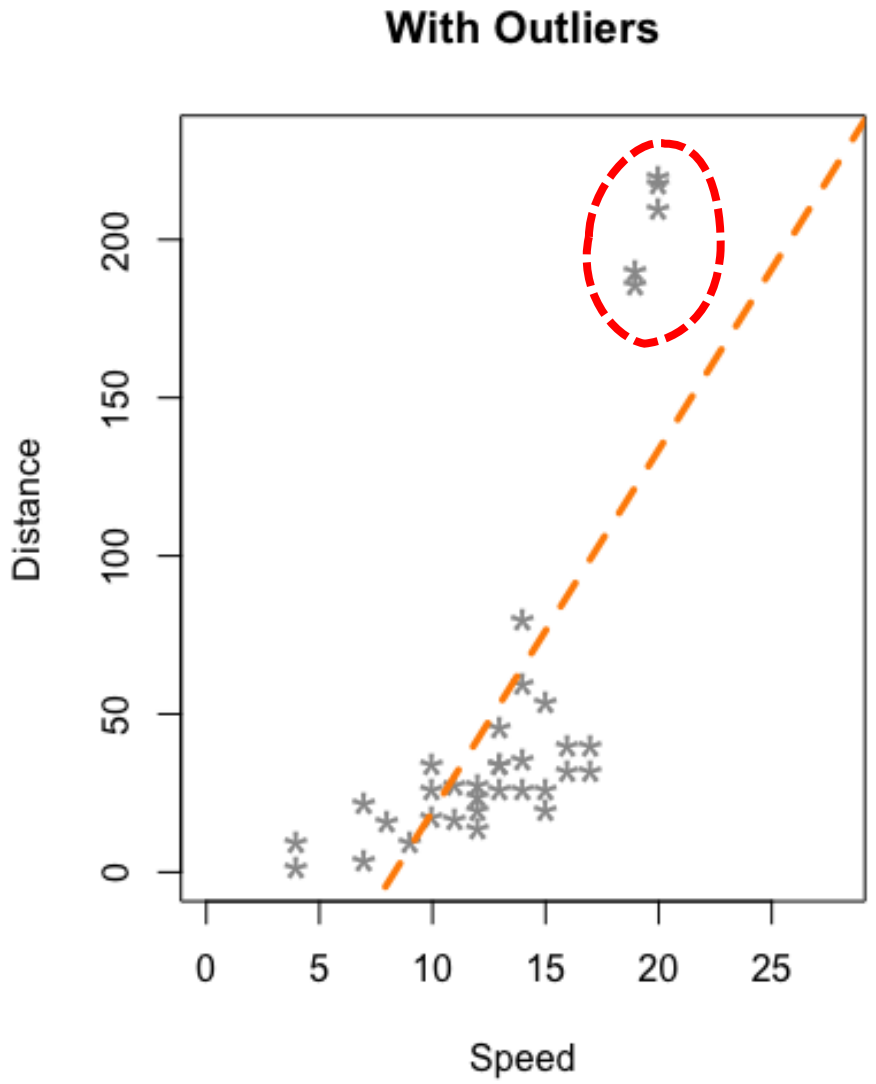


Outlier: An outlier is defined as an observation that has a large residual (i.e., absolute studentised or standardised residual exceeds 2 or 3), hence there is a large discrepancy between the observed outcome and predicted outcome from the linear regression model.

Leverage observations: A leverage observation is defined as an observation with a value of x far away from the mean of x (i.e., exceeding 2 or 3 times the average leverage).

Influential observations: An influential observation is defined as an observation that changes the estimates, and it usually has large residual and high leverage.

Influence statistics



Cook's Distance: A measure that combines the information of leverage and residual of the observation.

Measure	Value
Check the **Cook's distance**	> 4/n or >1
Check if **leverage** observation	> 2 × mean of leverage or > 3 × mean of leverage
|Studentised residual|	> 2 or >3

Simple linear regression

Dependent variable

Independent variable

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Random error

Intercept

Slope (regression) coefficient

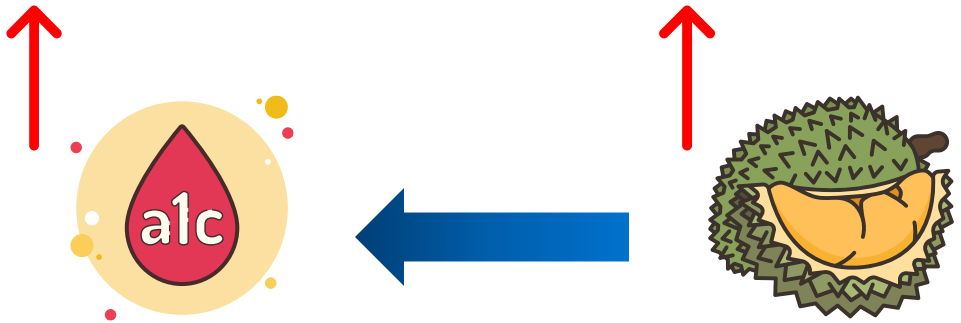
Simple linear regression

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Simple linear regression

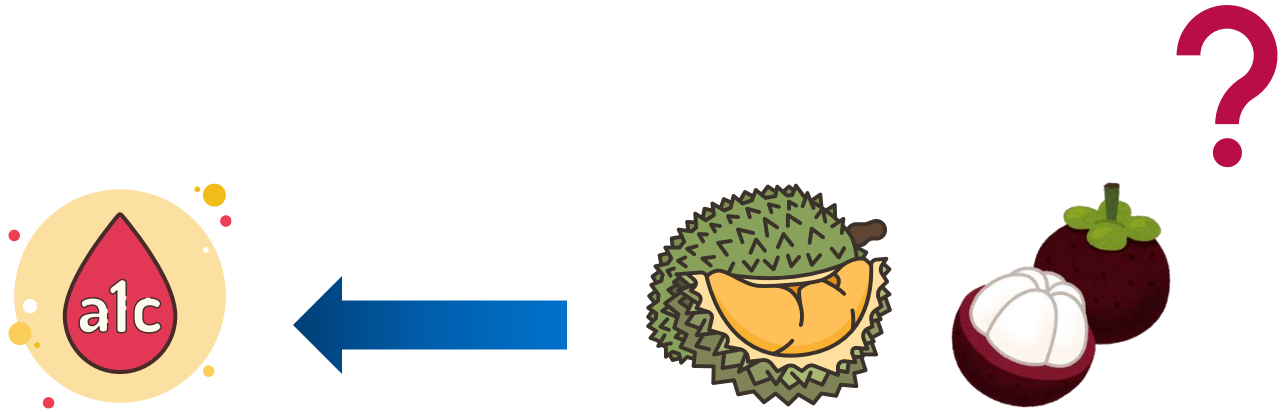
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Simple linear regression



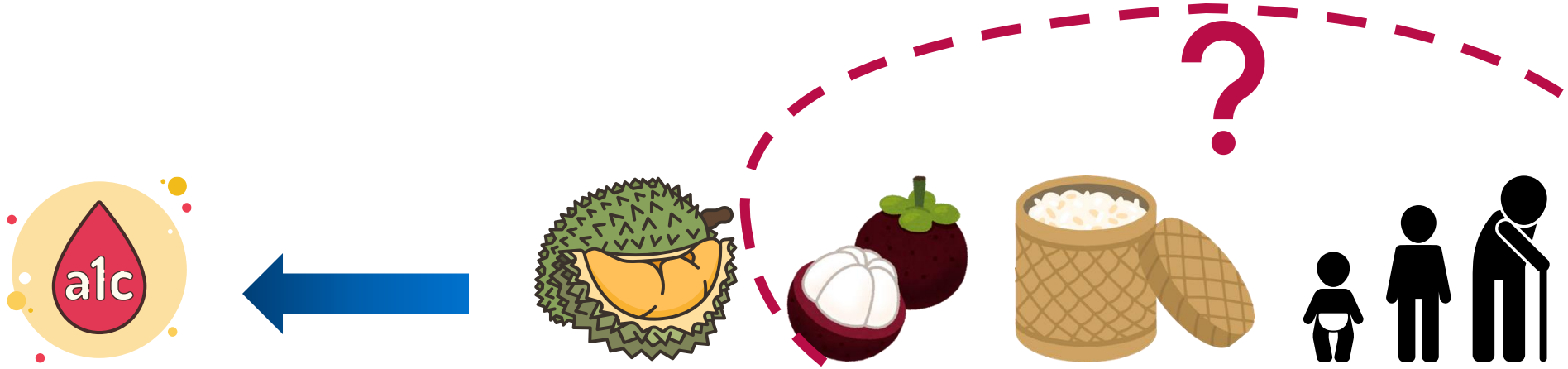
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Simple linear regression



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Simple linear regression



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Multiple linear regression

Dependent variable

Independent variables

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + \varepsilon_i$$

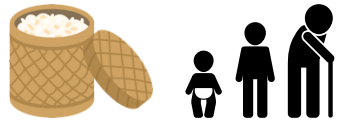
Random error

Multiple linear regression

Dependent variable

Independent variables

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + \varepsilon_i$$



Random error

Multiple linear regression

Dependent variable

Independent variables

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + \varepsilon_i$$

Intercept

Regression coefficients

Random error

Estimate the relationships between a dependent variable and independent variables

Multiple linear regression

Dependent variable

Independent variables

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + \varepsilon_i$$

Intercept

Regression coefficients

Random error

Estimate the relationships between a dependent variable and independent variables

Multiple linear regression

Multiple linear regression can improve our ability to predict an outcome when we have several predictors.

Controlling for confounding when investigating the relationship between the outcome Y_i and a predictor of interest $X_{i,1}$:

- A **confounding variable (or confounder)** is of little immediate interest but is correlated with $X_{i,1}$ and is independently related to the outcome.
- β_1 is the effect of $X_{i,1}$ on *yi* among subjects with the same values of $X_{i,2}$, $X_{i,3}$, …, $X_{i,p}$.
- The predictor of interest sometimes called the exposure.

How do you interpret the beta values?

Multiple linear regression

Multiple linear regression can improve our ability to predict an outcome when we have several predictors.

Controlling for confounding when investigating the relationship between the outcome Y_i and a predictor of interest $X_{i,1}$:

- A **confounding variable (or confounder)** is of little immediate interest but is correlated with $X_{i,1}$ and is independently related to the outcome.
- β_1 is the effect of $X_{i,1}$ on *yi* among subjects with the same values of $X_{i,2}$, $X_{i,3}$, …, $X_{i,p}$
- The predictor of interest sometimes called the exposure.

How do you interpret the beta values?

Linear regression

An example: factors associated with cardiovascular risk

Construct a multiple linear regression model


```
Call:
lm(formula = cvdData$ldl ~ cvdData$bmi + cvdData$age + factor(cvdData$race) +
    cvdData$gender)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1799 -0.5896 -0.0071  0.5171  2.0219

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.463024   0.514288   4.789 3.57e-06 ***
cvdData$bmi                  0.043112   0.017408   2.477  0.01422 *
cvdData$age                  0.013287   0.006313   2.105  0.03676 *
factor(cvdData$race)Chinese -0.100631   0.156395  -0.643  0.52078
factor(cvdData$race)Malays   0.188735   0.157285   1.200  0.23179
cvdData$genderFemale        -0.430744   0.128996  -3.339  0.00103 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8535 on 174 degrees of freedom
Multiple R-squared:  0.1172,    Adjusted R-squared:  0.09185
F-statistic: 4.621 on 5 and 174 DF,  p-value: 0.0005523
```

95% confidence intervals of the parameters


```
Call:
lm(formula = cvdData$ldl ~ cvdData$bmi + cvdData$age + factor(cvdData$race) +
    cvdData$gender)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1799 -0.5896 -0.0071  0.5171  2.0219

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   2.463024   0.514288   4.789 3.57e-06 ***
cvdData$bmi                   0.043112   0.017408   2.477  0.01422 *
cvdData$age                   0.013287   0.006313   2.105  0.03676 *
factor(cvdData$race)Chinese -0.100631   0.156395  -0.643  0.52078
factor(cvdData$race)Malays    0.188735   0.157285   1.200  0.23179
cvdData$genderFemale         -0.430744   0.128996  -3.339  0.00103 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8535 on 174 degrees of freedom
Multiple R-squared:  0.1172,    Adjusted R-squared:  0.09185
F-statistic: 4.621 on 5 and 174 DF,  p-value: 0.0005523
```

```
> confint(m1)
                                2.5 %      97.5 %
(Intercept)                1.4479788565  3.47806887
cvdData$bmi                0.0087547782  0.07746964
cvdData$age                0.0008266835  0.02574644
factor(cvdData$race)Chinese -0.4093052309  0.20804383
factor(cvdData$race)Malays  -0.1216975597  0.49916669
cvdData$genderFemale        -0.6853428156 -0.17614555
```

How do you interpret these results?

95% confidence intervals of the parameters

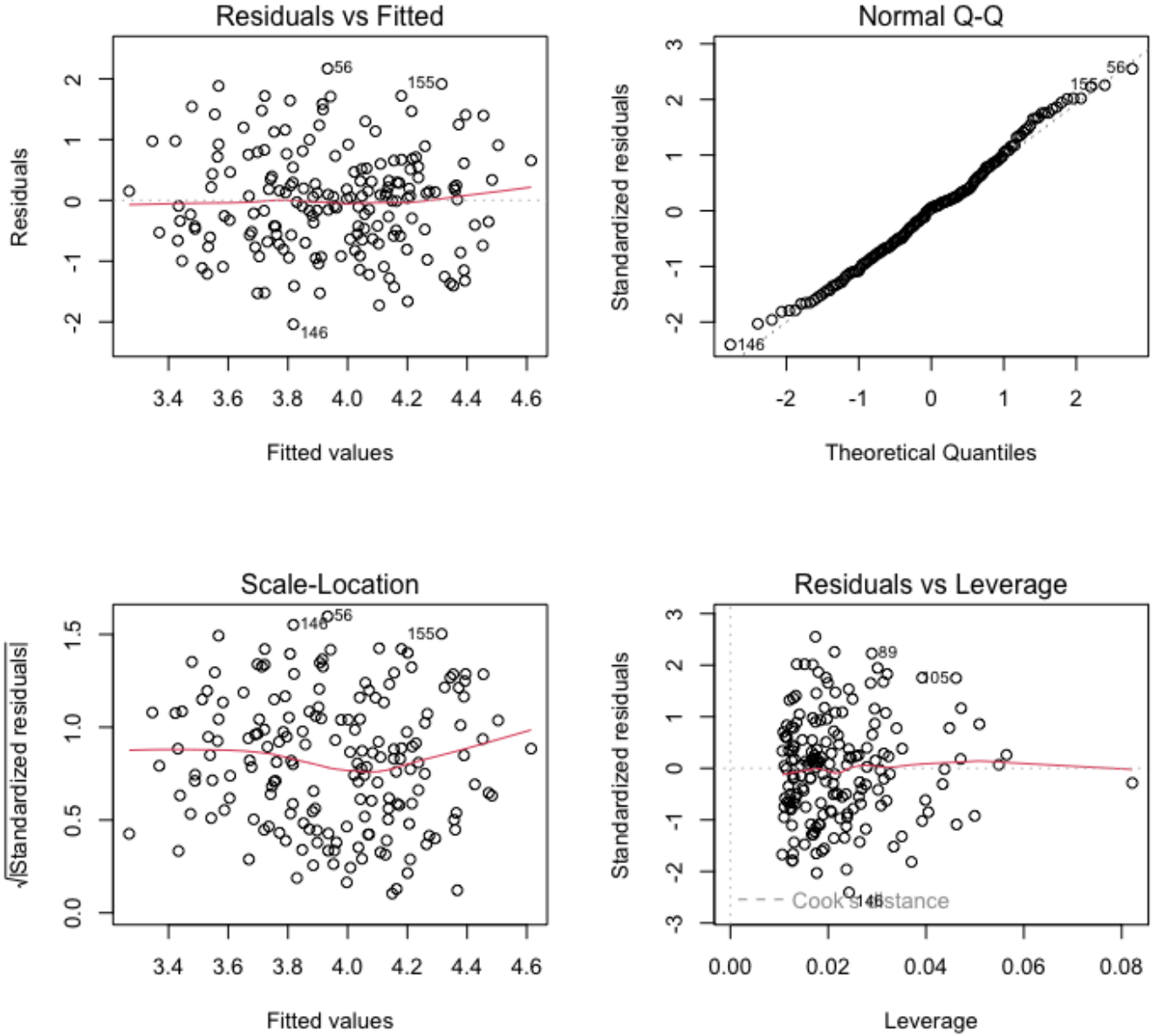
```
Call:
lm(formula = cvdData$ldl ~ cvdData$bmi + cvdData$age + factor(cvdData$race) +
    cvdData$gender)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1799 -0.5896 -0.0071  0.5171  2.0219

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   2.463024   0.514288   4.789 3.57e-06 ***
cvdData$bmi                   0.043112   0.017408   2.477  0.01422 *
cvdData$age                   0.013287   0.006313   2.105  0.03676 *
factor(cvdData$race)Chinese -0.100631   0.156395  -0.643  0.52078
factor(cvdData$race)Malays    0.188735   0.157285   1.200  0.23179
cvdData$genderFemale         -0.430744   0.128996  -3.339  0.00103 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8535 on 174 degrees of freedom
Multiple R-squared:  0.1172,    Adjusted R-squared:  0.09185
F-statistic: 4.621 on 5 and 174 DF,  p-value: 0.0005523
```

```
> confint(m1)
                                2.5 %      97.5 %
(Intercept)                1.4479788565  3.47806887
cvdData$bmi                0.0087547782  0.07746964
cvdData$age                0.0008266835  0.02574644
factor(cvdData$race)Chinese -0.4093052309  0.20804383
factor(cvdData$race)Malays  -0.1216975597  0.49916669
cvdData$genderFemale        -0.6853428156 -0.17614555
```

How do you interpret these results?

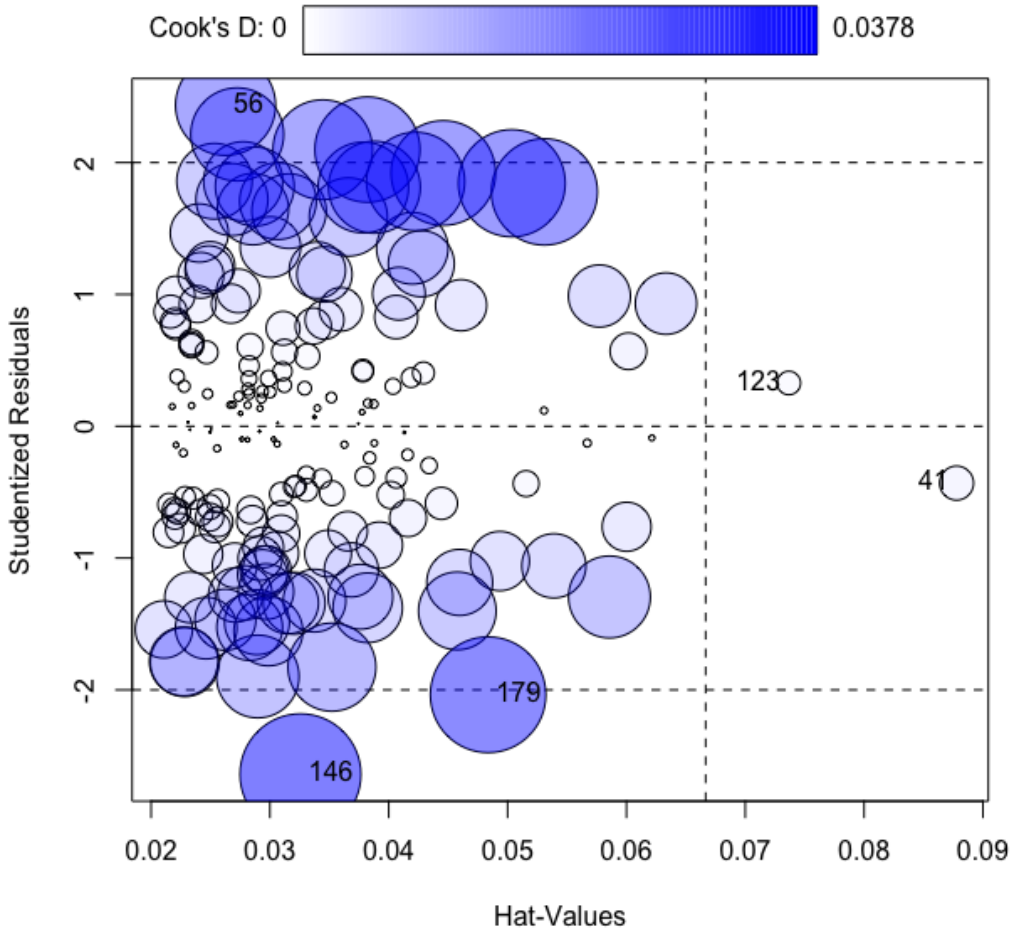
The effect of the age on LDL cholesterol is 0.0133 (95%CI: 0.0008–0.02575, p=0.0368) when the age increases by 1 year with an adjustment for BMI, race and sex.

Model diagnostics



Influential plot

Studentised residual vs leverage with Cook's distance




```
> influencePlot(m1, id=list(method="noteworthy",n=2))
       StudRes        Hat       CookD
41  -0.4313510 0.08778149 0.002998132
56   2.4345602 0.02626518 0.025912102
123  0.3278564 0.07366603 0.001432020
146 -2.6411234 0.03256909 0.037839694
179 -2.0369478 0.04835900 0.034516190
```

The observations listed are the top 2 largest absolute residuals (i.e., 56 and 146) or hat values (i.e., 41 and 123) or Cook's distances (i.e., 146 and 179). Let us remove these observations (i.e., 41st, 56th, 123th, 146th and 179th observations) and re-run the multiple linear regression analysis to compare the difference in the estimates.

Remove potentially influential observations

```
Call:
lm(formula = cvdData$ldl ~ cvdData$bmi + cvdData$age + factor(cvdData$race) +
    cvdData$gender)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1799 -0.5896 -0.0071  0.5171  2.0219

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   2.463024   0.514288   4.789 3.57e-06 ***
cvdData$bmi                   0.043112   0.017408   2.477  0.01422 *
cvdData$age                   0.013287   0.006313   2.105  0.03676 *
factor(cvdData$race)Chinese -0.100631   0.156395  -0.643  0.52078
factor(cvdData$race)Malays   0.188735   0.157285   1.200  0.23179
cvdData$genderFemale        -0.430744   0.128996  -3.339  0.00103 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8535 on 174 degrees of freedom
Multiple R-squared:  0.1172,    Adjusted R-squared:  0.09185
F-statistic: 4.621 on 5 and 174 DF,  p-value: 0.0005523
```

```
Call:
lm(formula = cvdDrop$ldl ~ cvdDrop$bmi + cvdDrop$age + cvdDrop$race +
    cvdDrop$gender)

Residuals:
     Min       1Q   Median       3Q      Max
-1.60240 -0.59470 -0.00231  0.49595  1.86078

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            2.473306   0.537547   4.601 8.22e-06 ***
cvdDrop$bmi            0.038703   0.017772   2.178  0.03081 *
cvdDrop$age            0.014817   0.006283   2.358  0.01950 *
cvdDrop$raceChinese   -0.094079   0.151985  -0.619  0.53675
cvdDrop$raceMalays     0.232770   0.154883   1.503  0.13474
cvdDrop$genderFemale -0.408124   0.125718  -3.246  0.00141 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8238 on 169 degrees of freedom
Multiple R-squared:  0.1256,    Adjusted R-squared:  0.09976
F-statistic: 4.856 on 5 and 169 DF,  p-value: 0.0003549
```

How do you interpret these results?

Remove potentially influential observations

```
Call:
lm(formula = cvdData$ldl ~ cvdData$bmi + cvdData$age + factor(cvdData$race) +
    cvdData$gender)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1799 -0.5896 -0.0071  0.5171  2.0219

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   2.463024   0.537547   4.582 8.22e-06 ***
cvdData$bmi                   0.043112   0.017408   2.477  0.01422 *
cvdData$age                   0.013287   0.005636   2.358  0.01950 *
factor(cvdData$race)Chinese  -0.100631   0.156395  -0.643  0.52078
factor(cvdData$race)Malays    0.188735   0.125718   1.501  0.13500
cvdData$genderFemale         -0.430744   0.128996  -3.339  0.00103 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8535 on 174 degrees of freedom
Multiple R-squared:  0.1172,    Adjusted R-squared:  0.09185
F-statistic: 4.621 on 5 and 174 DF,  p-value: 0.0005523
```

```
Call:
lm(formula = cvdDrop$ldl ~ cvdDrop$bmi + cvdDrop$age + cvdDrop$race +
    cvdDrop$gender)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1799 -0.5896 -0.0071  0.5171  2.0219

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   2.463024   0.537547   4.601 8.22e-06 ***
cvdDrop$bmi                   0.038703   0.017772   2.178  0.03081 *
cvdDrop$age                   0.013287   0.005636   2.358  0.01950 *
cvdDrop$raceChinese          -0.094079   0.151985  -0.619  0.53675
cvdDrop$raceMalays            0.188735   0.125718   1.503  0.13474
cvdDrop$genderFemale         -0.408124   0.125718  -3.246  0.00141 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8238 on 169 degrees of freedom
Multiple R-squared:  0.1256,    Adjusted R-squared:  0.09976
F-statistic: 4.856 on 5 and 169 DF,  p-value: 0.0003549
```

The adjusted R2 are comparable in the two models (i.e., Full: 9.2% vs Reduced: 10%). The significant findings correspond to the same variables (i.e., BMI, age and gender) with the same direction of the association and similar estimated values. Similarly, the non-significant findings correspond to the same variables (i.e., Chinese and Malays) with the same direction of association with estimated values from reduced data.

How do you interpret these results?

Multiple linear regression

Dependent variable

Independent variables

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + \varepsilon_i$$

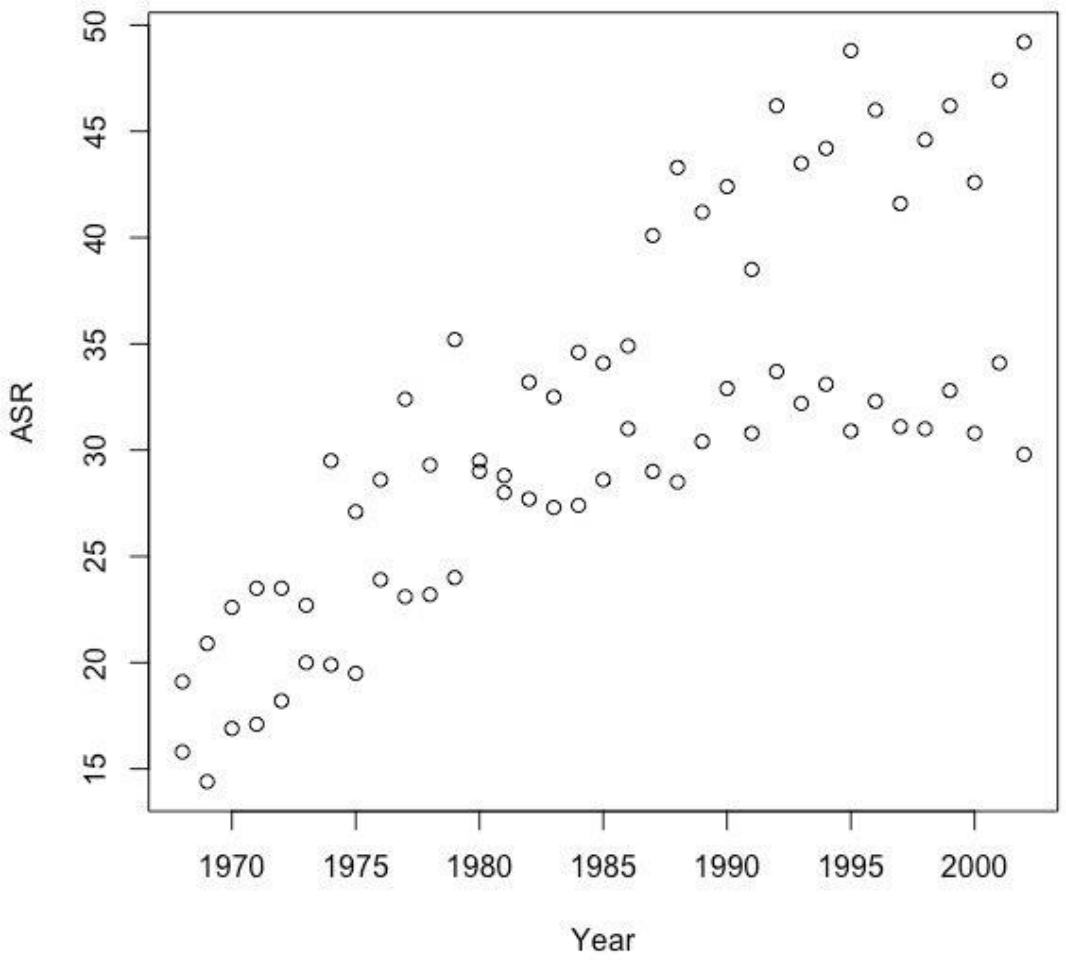
Intercept

Regression coefficients

Random error

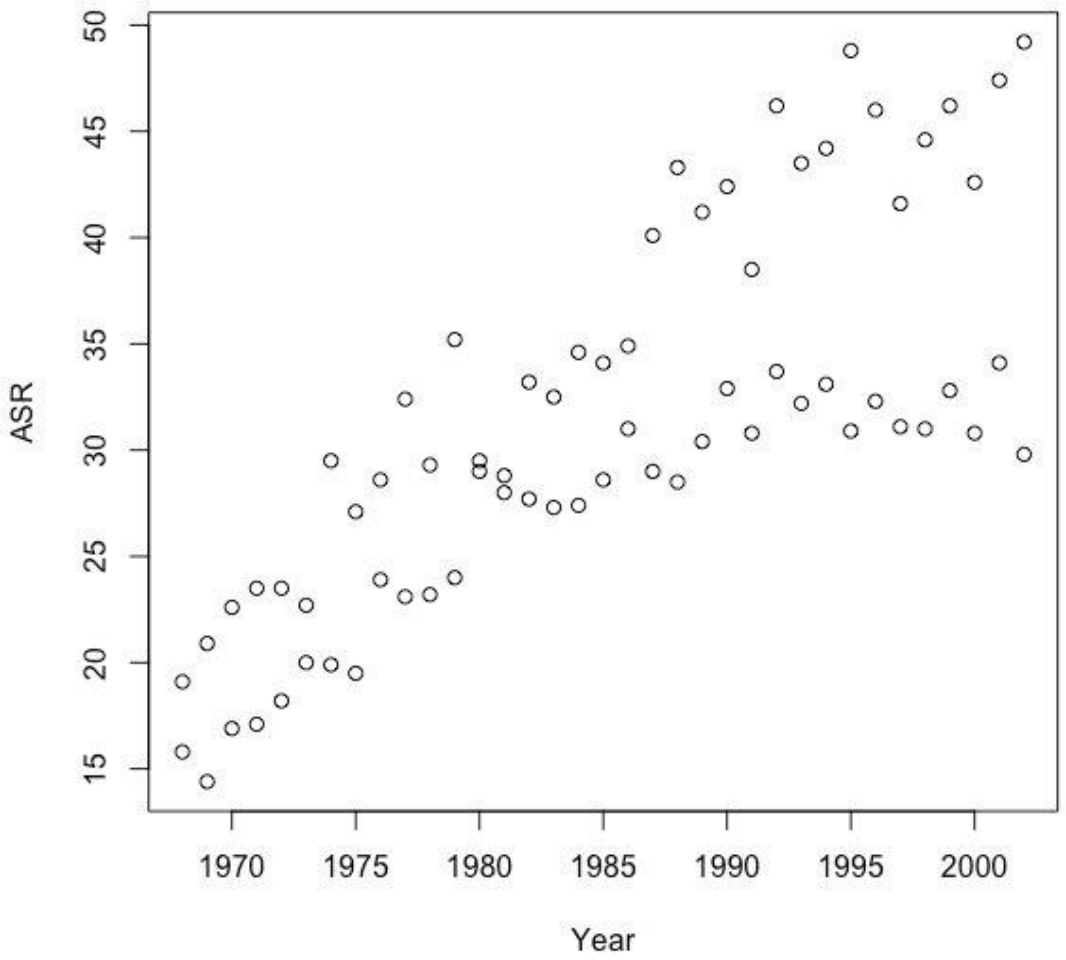
Estimate the relationships between a dependent variable and independent variables

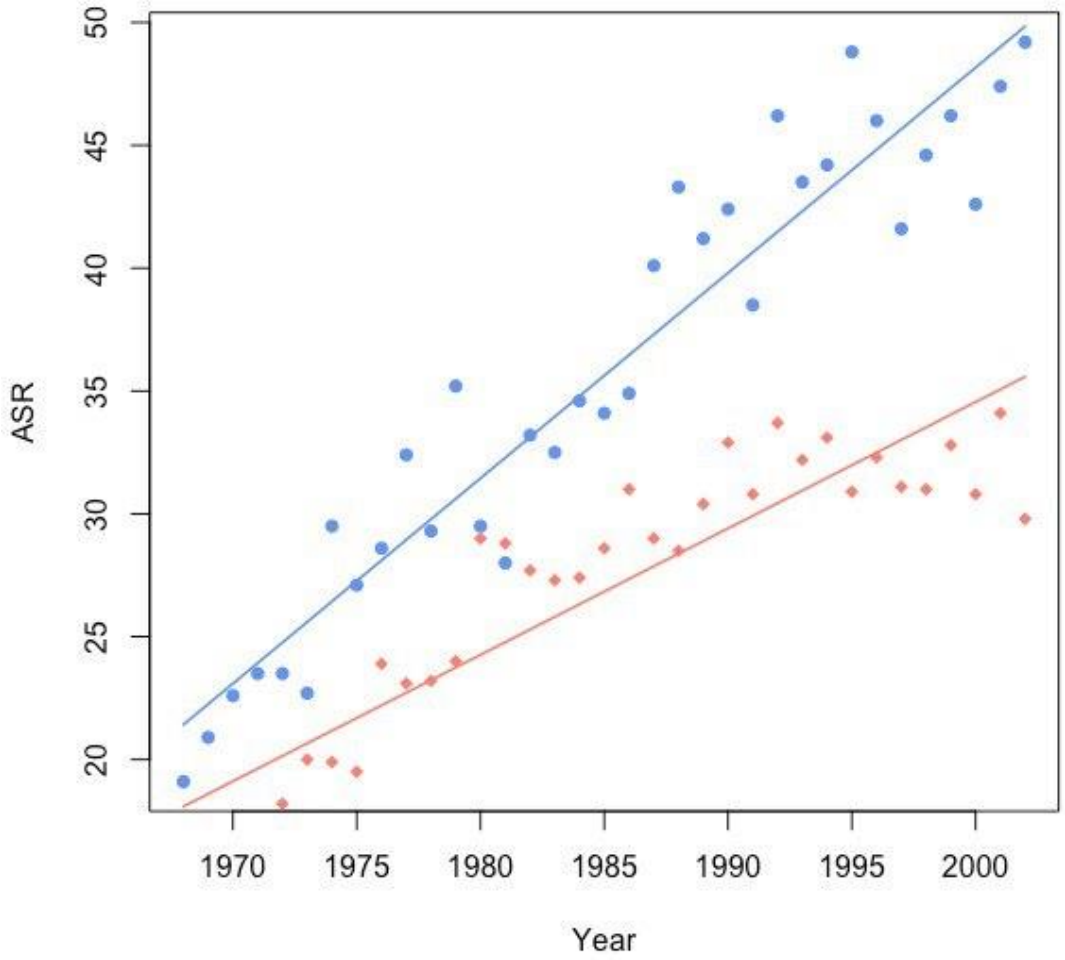
Colorectal cancer incidence in Singapore



Colorectal cancer incidence in Singapore

The linear temporal trend of rates from 1968 to 2002 is different between Singapore Chinese **males** and **females**.



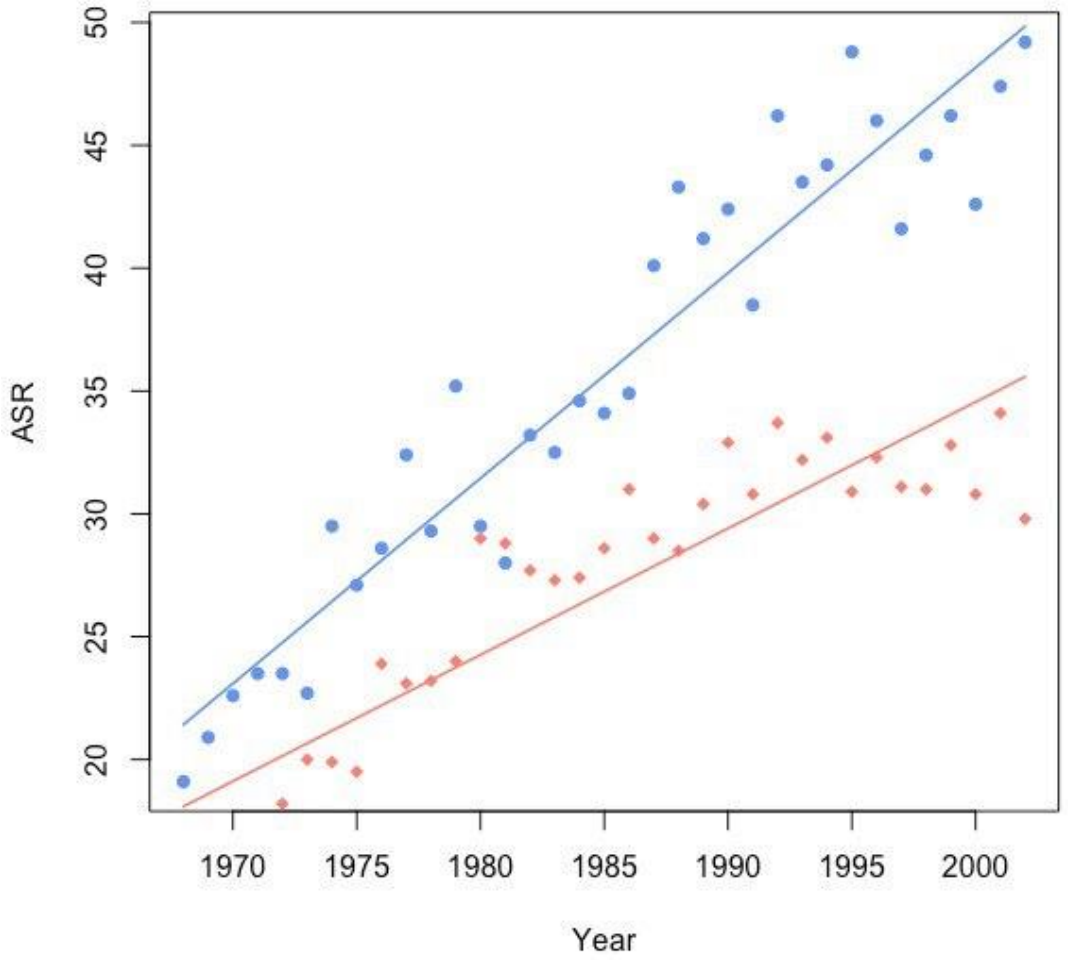


Modelling interaction

- In most scientific questions, for example, is smoking prevalence associated with cancer incidence rate?
- The answer is simple: **Yes**
- However, some answers to a scientific question can be complex: **It depends.**

Interaction (or effect modification) occurs when the magnitude of the effect of the exposure variable on an outcome (i.e., the association) differs depending on the level of a third variable (i.e. effect modifier).

The linear temporal trend of rates from 1968 to 2002 is different between Singapore Chinese **males** and **females**.



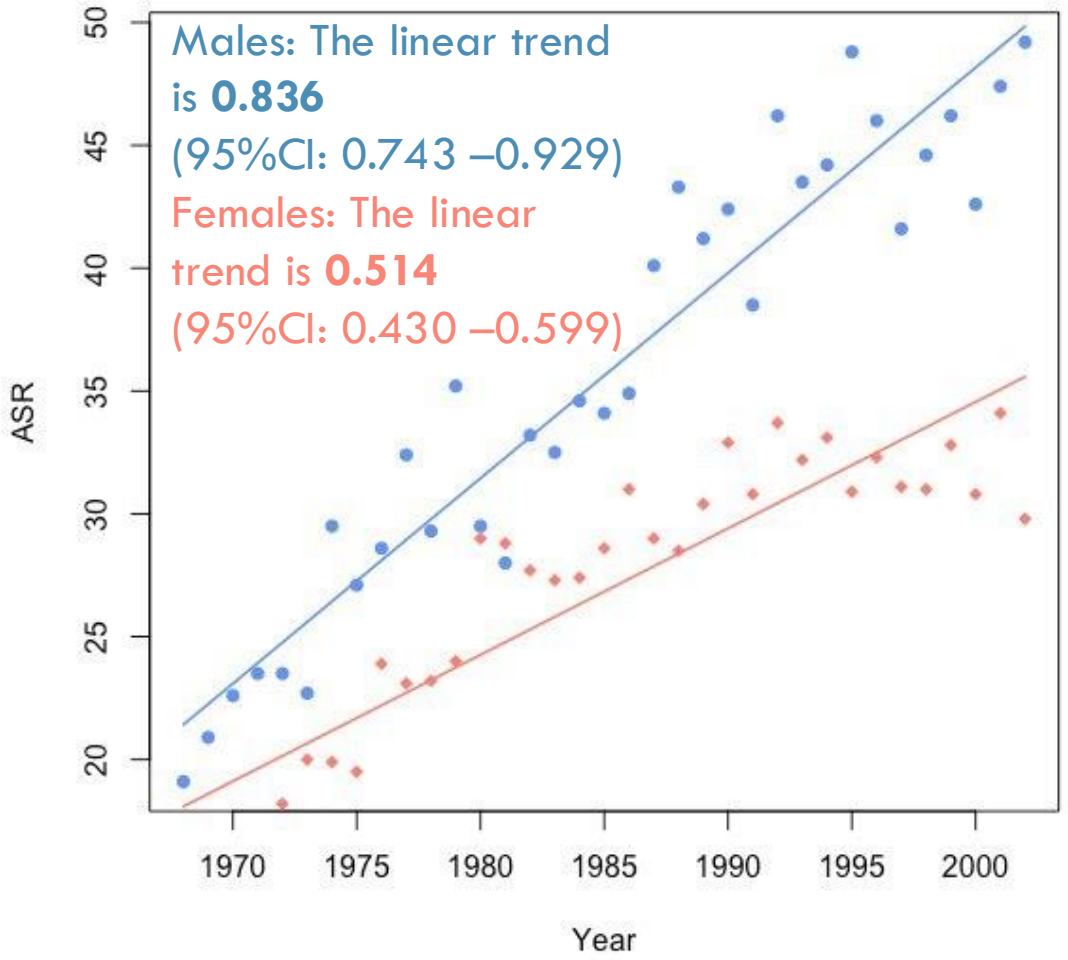
Modelling interaction: fit separate models?

NUS National University of Singapore | Saw Swee Hock School of Public Health

```
> summary(m_male)

Call:
lm(formula = both$ASR[both$Female %in% 0] ~ both$Year[both$Female %in%
    0])

Residuals:
    Min      1Q  Median      3Q     Max
-5.5643 -1.5821 -0.4786  1.7071  5.1643

Coefficients:
                              Estimate  Std. Error t value           Pr(>|t|)
(Intercept)                 -1623.26429    90.53351  -17.93 <0.0000000000000002 ***
both$Year[both$Female %in% 0]     0.83571     0.04561   18.32 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.725 on 33 degrees of freedom
Multiple R-squared:  0.9105,    Adjusted R-squared:  0.9078
F-statistic: 335.8 on 1 and 33 DF,  p-value: < 0.0000000000000022

> summary(m_female)

Call:
lm(formula = both$ASR[both$Female %in% 1] ~ both$Year[both$Female %in%
    1])

Residuals:
    Min      1Q  Median      3Q     Max
-5.7805 -1.9270  0.1223  1.5716  4.7381

Coefficients:
                              Estimate Std. Error t value           Pr(>|t|)
(Intercept)                 -994.4121    82.7739  -12.01 0.0000000000001355 ***
both$Year[both$Female %in% 1]    0.5145     0.0417   12.34 0.0000000000000657 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.492 on 33 degrees of freedom
Multiple R-squared:  0.8218,    Adjusted R-squared:  0.8164
F-statistic: 152.2 on 1 and 33 DF,  p-value: 0.0000000000006573
```

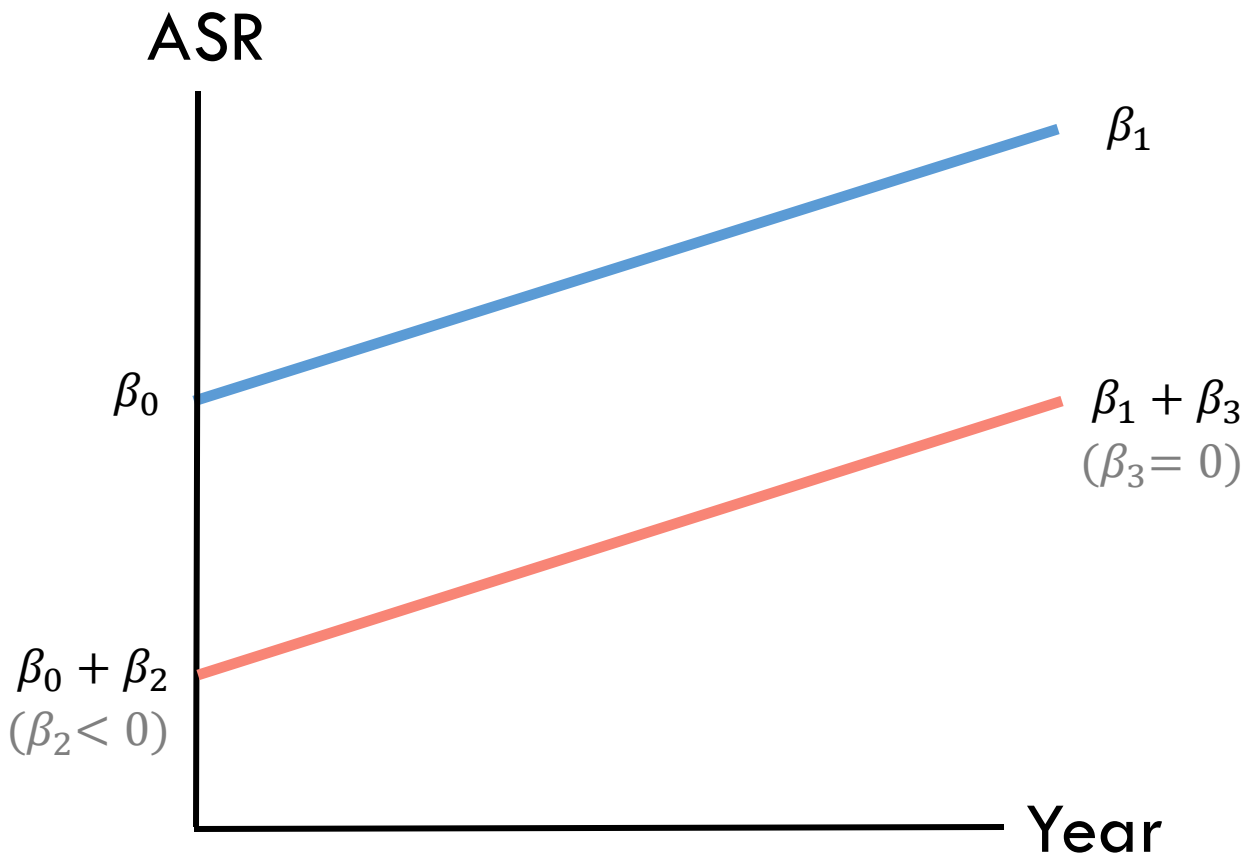
Stratified analysis

The linear temporal trend of rates from 1968 to 2002 is different between Singapore Chinese **males** and **females**.



Interpretation of model with interaction

$$\text{ASR}_i = \beta_0 + \beta_1 \text{Year}_i + \beta_2 \text{Female}_i + \beta_3 (\text{Year}_i \times \text{Female}_i) + \varepsilon_i$$



T-test

Null hypothesis

$H_0: \beta_3 = 0$

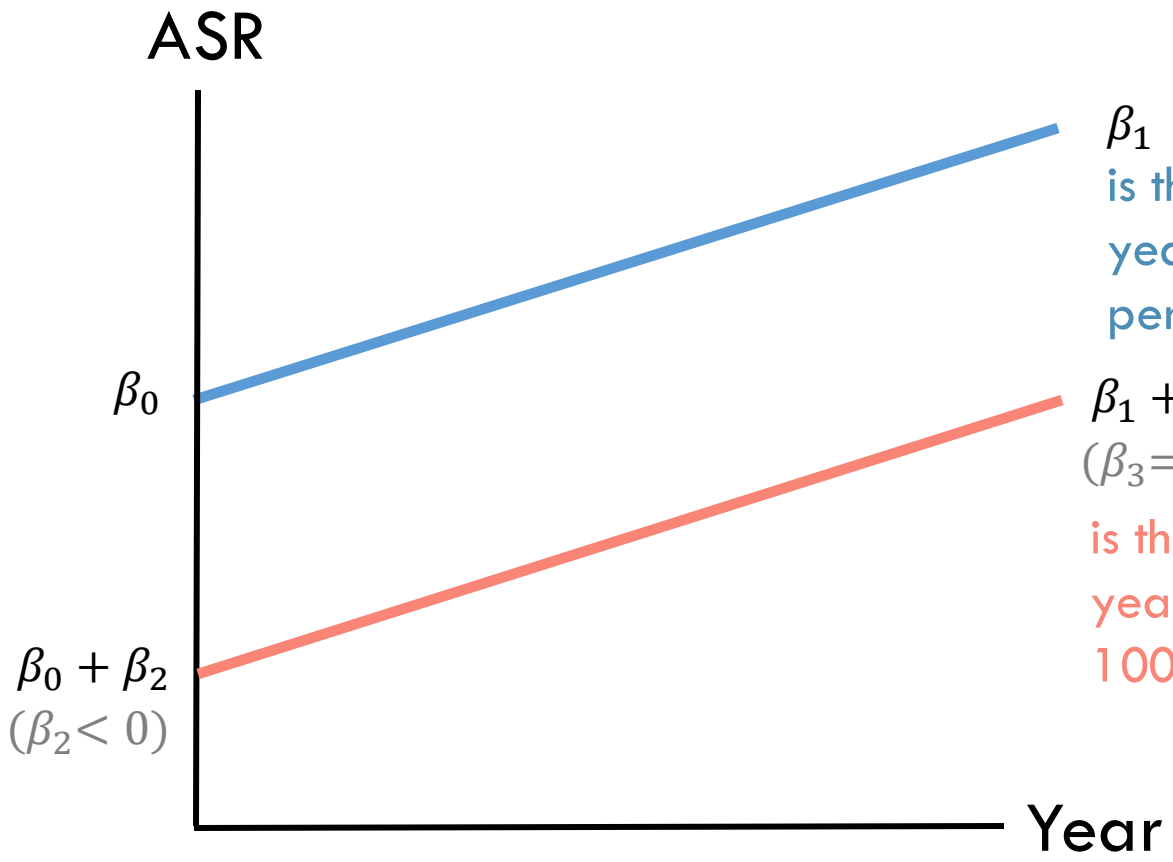
Alternative hypothesis

$H_1: \beta_3 \neq 0$

Interpretation of model with interaction

$$\text{ASR}_i = \beta_0 + \beta_1 \text{Year}_i + \beta_2 \text{Female}_i + \beta_3(\text{Year}_i \times \text{Female}_i) + \varepsilon_i$$

$$\text{ASR}_i = (\beta_0 + \beta_2 \text{Female}_i) + (\beta_1 + \beta_3 \text{Female}_i)\text{Year}_i + \varepsilon_i$$



β_1
is the linear trend for males, i.e., as Year_i changes by 1 year, ASR_i is expected to change by β_1 per 100,000 person-years

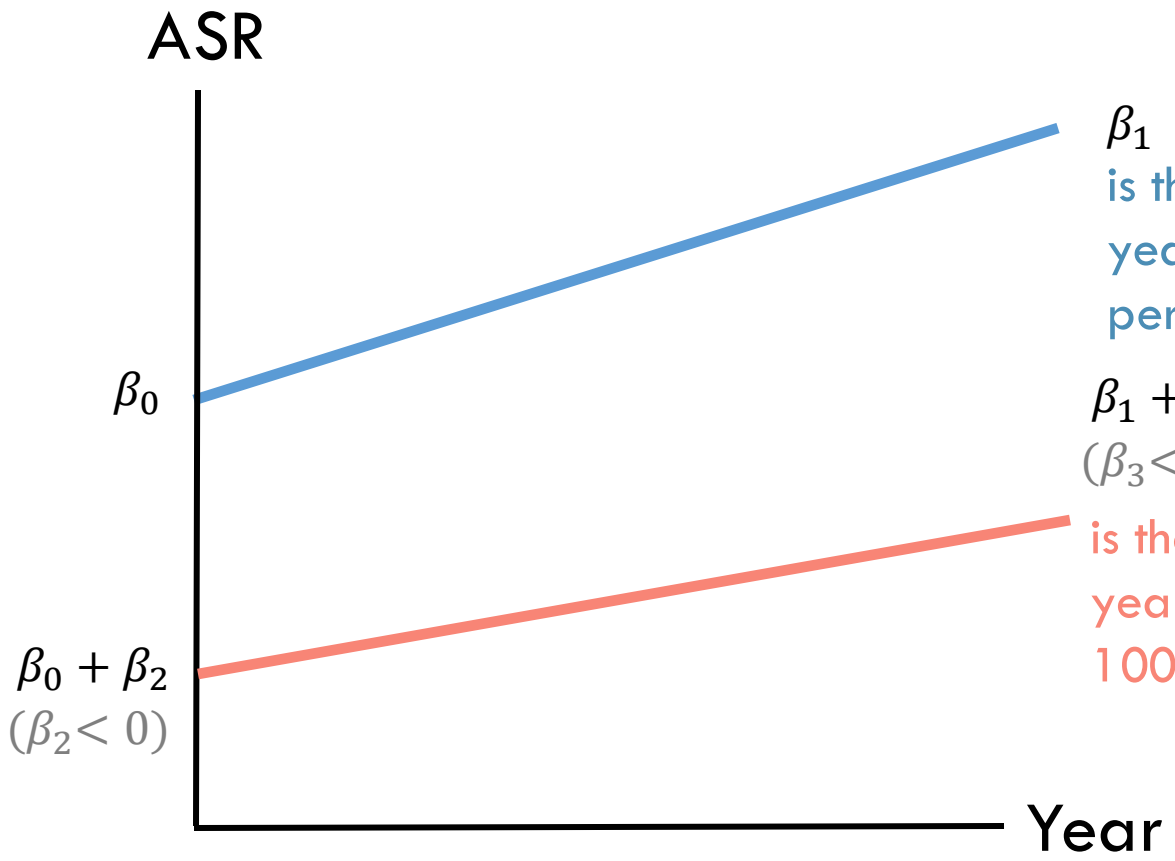
$\beta_1 + \beta_3$
$(\beta_3 = 0)$
is the linear trend for males, i.e., as Year_i changes by 1 year, ASR_i is expected to change by $(\beta_1 + \beta_3)$ per 100,000 person-years

If $\beta_3 = 0$ the linear trend between males and females is the same, i.e., parallel lines.

Interpretation of model with interaction

$$\text{ASR}_i = \beta_0 + \beta_1 \text{Year}_i + \beta_2 \text{Female}_i + \beta_3(\text{Year}_i \times \text{Female}_i) + \varepsilon_i$$

$$\text{ASR}_i = (\beta_0 + \beta_2 \text{Female}_i) + (\beta_1 + \beta_3 \text{Female}_i)\text{Year}_i + \varepsilon_i$$



β_1
is the linear trend for males, i.e., as Year_i changes by 1 year, ASR_i is expected to change by β_1 per 100,000 person-years

$\beta_1 + \beta_3$
$(\beta_3 < 0)$
is the linear trend for males, i.e., as Year_i changes by 1 year, ASR_i is expected to change by $(\beta_1 + \beta_3)$ per 100,000 person-years

If $\beta_3 \neq 0$ the linear trend between males and females is the different, i.e., non-parallel lines.

Modelling interaction: Intercept

```
Call:
lm(formula = both$ASR ~ both$Year.centered * both$Female.factor)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7805 -1.8196 -0.1932  1.6035  5.1643

Coefficients:
                                            Estimate Std. Error t value             Pr(>|t|)
(Intercept)                                  35.6286     0.4413  80.731 < 0.0000000000000002 ***
both$Year.centered                            0.8357     0.0437  19.125 < 0.0000000000000002 ***
both$Female.factorFemale                     -8.7943     0.6241 -14.091 < 0.0000000000000002 ***
both$Year.centered:both$Female.factorFemale  -0.3212     0.0618  -5.198          0.00000212 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.611 on 66 degrees of freedom
Multiple R-squared:  0.9142,    Adjusted R-squared:  0.9103
F-statistic: 234.3 on 3 and 66 DF,  p-value: < 0.00000000000000022
```

β_0

We centre the mean year to improve the interpretation of the intercept.

Recall, the linear model for **males**:

$$\text{ASR}_i = \beta_0 + \beta_1 \textbf{Year}_i + \varepsilon_i$$

With mean centred year:

$$\text{ASR}_i = \beta_0 + \beta_1 (\textbf{Year}_i - 1985) + \varepsilon_i$$

The intercept β_0 is the mean ASR among males at 1985, which is 35.6 per 100,000 person-years.

Modelling interaction: Interaction

```
Call:
lm(formula = both$ASR ~ both$Year.centered * both$Female.factor)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7805 -1.8196 -0.1932  1.6035  5.1643

Coefficients:
                                          Estimate Std. Error t value             Pr(>|t|)
(Intercept)                               35.6286     0.4413  80.731 < 0.0000000000000002 ***
both$Year.centered                         0.8357     0.0437  19.125 < 0.0000000000000002 ***
both$Female.factorFemale                  -8.7943     0.6241 -14.091 < 0.0000000000000002 ***
both$Year.centered:both$Female.factorFemale -0.3212   0.0618  -5.198             0.00000212 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.611 on 66 degrees of freedom
Multiple R-squared:  0.9142,    Adjusted R-squared:  0.9103
F-statistic: 234.3 on 3 and 66 DF,  p-value: < 0.00000000000000022
```

β_1

β_3

The interaction factor of two predictors, $\mathbf{Year}_i$ and $\mathbf{Female}_i$, is a multiplication of the two predictors.

Linear trend for **males**: β_1

$$\widehat{\beta_1} = 0.836$$

Linear trend for **females**: $\beta_1 + \beta_3$

$$\widehat{\beta_1} + \widehat{\beta_3} = 0.836 + (-0.321) = 0.515$$

The linear trend between males and females is **significantly different** (p-value<0.001).

Modelling interaction: How do we decide?

NUS National University of Singapore | Saw Swee Hock School of Public Health

1. How do we decide whether to **include an interaction in the model**?
   - ❑ Prior belief about the relationship between two predictors and outcome.
   - ❑ Exploratory data analysis (e.g., plots and assessing the significance of the interaction term).
2. So, how **many predictors** should we include in an interaction?
   - ❑ Seldom consider the interaction between 3 or more predictors (i.e., multiplying three or predictors together) because the explanation becomes difficult and can potentially be meaningless.
   - ❑ Here we will only consider interactions consisting of two predictors (i.e., two-way interactions)

ANOVA (Type I)

Type I compares the variables sequentially in the model.

Let us consider the model (m1) where the variables are specified as:

Year.centered + Female.factor + Year.centered:Female.factor

Note that mean centred year (Year.centered) was specified first, followed by female (Female.factor), and finally the interaction between the first two variables (Year.centered:Female.factor).

```
> m1 = lm(both$ASR~both$Year.centered + both$Female.factor + both$Year.centered:both$Female.factor)
> summary(m1)

Call:
lm(formula = both$ASR ~ both$Year.centered + both$Female.factor +
    both$Year.centered:both$Female.factor)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7805 -1.8196 -0.1932  1.6035  5.1643

Coefficients:
                                          Estimate Std. Error t value            Pr(>|t|)
(Intercept)                                35.6286     0.4413  80.731 < 0.0000000000000002 ***
both$Year.centered                          0.8357     0.0437  19.125 < 0.0000000000000002 ***
both$Female.factorFemale                   -8.7943     0.6241 -14.091 < 0.0000000000000002 ***
both$Year.centered:both$Female.factorFemale -0.3212     0.0618  -5.198          0.00000212 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.611 on 66 degrees of freedom
Multiple R-squared:  0.9142,    Adjusted R-squared:  0.9103
F-statistic: 234.3 on 3 and 66 DF,  p-value: < 0.00000000000000022

> anova(m1)
Analysis of Variance Table

Response: both$ASR
                                   Df Sum Sq Mean Sq F value               Pr(>F)
both$Year.centered                  1 3254.1  3254.1 477.367 < 0.00000000000000022 ***
both$Female.factor                  1 1353.4  1353.4 198.545 < 0.00000000000000022 ***
both$Year.centered:both$Female.factor  1  184.2   184.2  27.021          0.00000212 ***
Residuals                          66  449.9     6.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA (Type I)

Type I would provide the variation explained by:

1. Mean centred year with the model having mean centred year only (giving p-value for slope of mean centred year)
2. Female with the model having mean centred year and female only(giving p-value for slope of value for slope of female after adjusting for mean centred year)
3. Interaction with the model having mean centred year, female, and interaction (giving p-value for slope of interaction after adjusting for mean centred year and female)

```
> m1 = lm(both$ASR~both$Year.centered + both$Female.factor + both$Year.centered:both$Female.factor)
> summary(m1)

Call:
lm(formula = both$ASR ~ both$Year.centered + both$Female.factor +
    both$Year.centered:both$Female.factor)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7805 -1.8196 -0.1932  1.6035  5.1643

Coefficients:
                                          Estimate Std. Error t value            Pr(>|t|)
(Intercept)                                35.6286     0.4413  80.731 < 0.0000000000000002 ***
both$Year.centered                          0.8357     0.0437  19.125 < 0.0000000000000002 ***
both$Female.factorFemale                   -8.7943     0.6241 -14.091 < 0.0000000000000002 ***
both$Year.centered:both$Female.factorFemale -0.3212     0.0618  -5.198          0.00000212 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.611 on 66 degrees of freedom
Multiple R-squared:  0.9142,    Adjusted R-squared:  0.9103
F-statistic: 234.3 on 3 and 66 DF,  p-value: < 0.00000000000000022

> anova(m1)
Analysis of Variance Table

Response: both$ASR
                                   Df Sum Sq Mean Sq F value               Pr(>F)
both$Year.centered                  1 3254.1  3254.1 477.367 < 0.00000000000000022 ***
both$Female.factor                  1 1353.4  1353.4 198.545 < 0.00000000000000022 ***
both$Year.centered:both$Female.factor  1  184.2   184.2  27.021          0.00000212 ***
Residuals                          66  449.9     6.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA (Type I)

Now let us consider the model (m2) where the variables are specified as:

Year.centered:Female.factor + Female.factor + Year.centered

Note that interaction was specified first, followed by female, and finally mean centred year.

```
> m2 = lm(both$ASR~both$Year.centered.Female + both$Female.factor + both$Year.centered)
> summary(m2)

Call:
lm(formula = both$ASR ~ both$Year.centered.Female + both$Female.factor +
    both$Year.centered)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7805 -1.8196 -0.1932  1.6035  5.1643

Coefficients:
                           Estimate Std. Error t value            Pr(>|t|)
(Intercept)                 35.6286     0.4413  80.731 < 0.0000000000000002 ***
both$Year.centered.Female   -0.3212     0.0618  -5.198             0.00000212 ***
both$Female.factorFemale    -8.7943     0.6241 -14.091 < 0.0000000000000002 ***
both$Year.centered           0.8357     0.0437  19.125 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.611 on 66 degrees of freedom
Multiple R-squared:  0.9142,    Adjusted R-squared:  0.9103
F-statistic: 234.3 on 3 and 66 DF,  p-value: < 0.00000000000000022

> anova(m2)
Analysis of Variance Table

Response: both$ASR
                           Df  Sum Sq Mean Sq F value                Pr(>F)
both$Year.centered.Female   1  944.95  944.95  138.62 < 0.00000000000000022 ***
both$Female.factor          1 1353.44 1353.44  198.55 < 0.00000000000000022 ***
both$Year.centered          1 2493.35 2493.35  365.77 < 0.00000000000000022 ***
Residuals                  66  449.91    6.82
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA (Type I)

Type I would provide the variation explained by:

1. Interaction with the model having the interaction only (giving p-value for slope of interaction)
2. Female with the model having interaction and female only(giving p-value for slope of value for slope of female after adjusting for interaction)
3. Mean centred year with the model model having interaction, female and mean centred year (giving p-value for slope of mean centred year after adjusting for interaction and female)

Order matters!

```
> m2 = lm(both$ASR~both$Year.centered.Female + both$Female.factor + both$Year.centered)
> summary(m2)

Call:
lm(formula = both$ASR ~ both$Year.centered.Female + both$Female.factor +
    both$Year.centered)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7805 -1.8196 -0.1932  1.6035  5.1643

Coefficients:
                          Estimate Std. Error t value            Pr(>|t|)
(Intercept)                35.6286     0.4413  80.731 < 0.0000000000000002 ***
both$Year.centered.Female  -0.3212     0.0618  -5.198             0.00000212 ***
both$Female.factorFemale   -8.7943     0.6241 -14.091 < 0.0000000000000002 ***
both$Year.centered          0.8357     0.0437  19.125 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.611 on 66 degrees of freedom
Multiple R-squared:  0.9142,    Adjusted R-squared:  0.9103
F-statistic: 234.3 on 3 and 66 DF,  p-value: < 0.00000000000000022

> anova(m2)
Analysis of Variance Table

Response: both$ASR
                          Df  Sum Sq Mean Sq F value                Pr(>F)
both$Year.centered.Female  1  944.95  944.95  138.62 < 0.00000000000000022 ***
both$Female.factor         1 1353.44 1353.44  198.55 < 0.00000000000000022 ***
both$Year.centered         1 2493.35 2493.35  365.77 < 0.00000000000000022 ***
Residuals                 66  449.91    6.82
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA (Type I)

```
> m1 = lm(both$ASR~both$Year.centered + both$Female.factor + both$Year.centered:both$Female.factor)
> summary(m1)

Call:
lm(formula = both$ASR ~ both$Year.centered + both$Female.factor +
    both$Year.centered:both$Female.factor)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7805 -1.8196 -0.1932  1.6035  5.1643

Coefficients:
                                          Estimate Std. Error t value            Pr(>|t|)
(Intercept)                                35.6286     0.4413  80.731 < 0.0000000000000002 ***
both$Year.centered                          0.8357     0.0437  19.125 < 0.0000000000000002 ***
both$Female.factorFemale                   -8.7943     0.6241 -14.091 < 0.0000000000000002 ***
both$Year.centered:both$Female.factorFemale -0.3212     0.0618  -5.198          0.00000212 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.611 on 66 degrees of freedom
Multiple R-squared:  0.9142,    Adjusted R-squared:  0.9103
F-statistic: 234.3 on 3 and 66 DF,  p-value: < 0.00000000000000022
```

```
> m2 = lm(both$ASR~both$Year.centered.Female + both$Female.factor + both$Year.centered)
> summary(m2)

Call:
lm(formula = both$ASR ~ both$Year.centered.Female + both$Female.factor +
    both$Year.centered)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7805 -1.8196 -0.1932  1.6035  5.1643

Coefficients:
                          Estimate Std. Error t value            Pr(>|t|)
(Intercept)                35.6286     0.4413  80.731 < 0.0000000000000002 ***
both$Year.centered.Female  -0.3212     0.0618  -5.198          0.00000212 ***
both$Female.factorFemale   -8.7943     0.6241 -14.091 < 0.0000000000000002 ***
both$Year.centered          0.8357     0.0437  19.125 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.611 on 66 degrees of freedom
Multiple R-squared:  0.9142,    Adjusted R-squared:  0.9103
F-statistic: 234.3 on 3 and 66 DF,  p-value: < 0.00000000000000022
```

The two models gave the same estimates for the predictors and the same R^2.

ANOVA (Type I)

```
> anova(m1)
Analysis of Variance Table

Response: both$ASR
                                 Df Sum Sq Mean Sq F value              Pr(>F)
both$Year.centered                1 3254.1  3254.1 477.367 < 0.0000000000000022 ***
both$Female.factor                1 1353.4  1353.4 198.545 < 0.0000000000000022 ***
both$Year.centered:both$Female.factor  1  184.2   184.2  27.021          0.00000212 ***
Residuals                        66  449.9     6.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> anova(m2)
Analysis of Variance Table

Response: both$ASR
                          Df  Sum Sq Mean Sq F value              Pr(>F)
both$Year.centered.Female  1  944.95  944.95  138.62 < 0.0000000000000022 ***
both$Female.factor         1 1353.44 1353.44  198.55 < 0.0000000000000022 ***
both$Year.centered         1 2493.35 2493.35  365.77 < 0.0000000000000022 ***
Residuals                 66  449.91    6.82
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The Type I of m1 and m2 gave different sum of squares (i.e., Sum Sq) and p-values (esp. interaction).

Order is important for Type I

Collinearity

Predictors are naturally correlated.

Collinearity is a problem when adding correlated predictors into the model, resulting in a very large increase in the standard errors of some estimates.

The collinearity problem **entangles the effects of predictors and complicates the interpretation**.

Collinearity

Collinearity occurs when two or more predictors have a linear relationship (or are highly correlated).

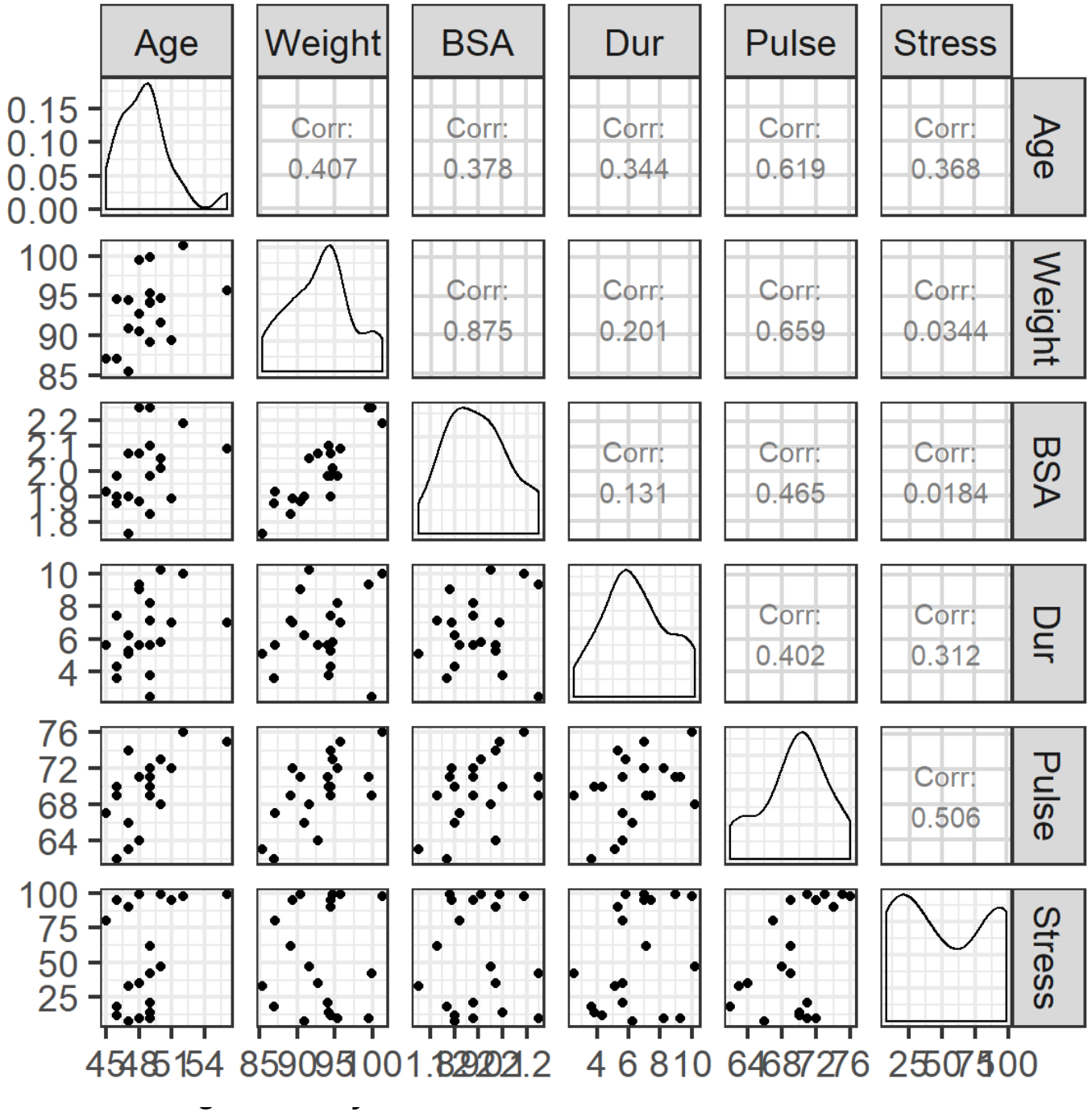
Among current smokers, these three variables:
(i) age at survey,
(ii) age started smoking, and
(iii) duration of smoking (in years), are related in a linear expression:

Age at survey = Age started smoking + Smoking duration

Estimates for these predictors are not uniquely estimated because of the exact linear expression (i.e., perfect collinearity).

Software may drop one of the predictors when there is perfect collinearity (e.g., the correlation between predictors is 1).

Pairwise scatterplot



Heat map of correlation coefficients

Variance inflation factor (VIF)

Quantifies how much the variance of the estimate (or standard error) is **inflated** because of the **collinearities among the predictors** (when compared with no collinearity)

a) **VIF of 1** for the j^{th} predictor means no correlation between this predictor and the remaining predictors in the model (i.e., no inflation in variance).
b) **VIF > 4** warrants further investigation.
c) **VIF > 10** is a sign of serious collinearity requiring correction.

Big VIF is bad → means inflated (high) uncertainty in estimates

- ❑ Remove problematic (highly correlated) variables
- ❑ Use linearly combine the predictor variables

Variance inflation factor (VIF)

VIF is not applicable to categorical variables with three or more groups.

A generalised variance inflation factor (GVIF) was proposed with degrees of freedom (df) of a categorical variable corresponding to:

$$df = \text{number of groups} - 1$$

The **GVIF$^{1/(2 \times df)}$** is reported, and it reduces to the square root of VIF when the variable is continuous or binary, suggesting:

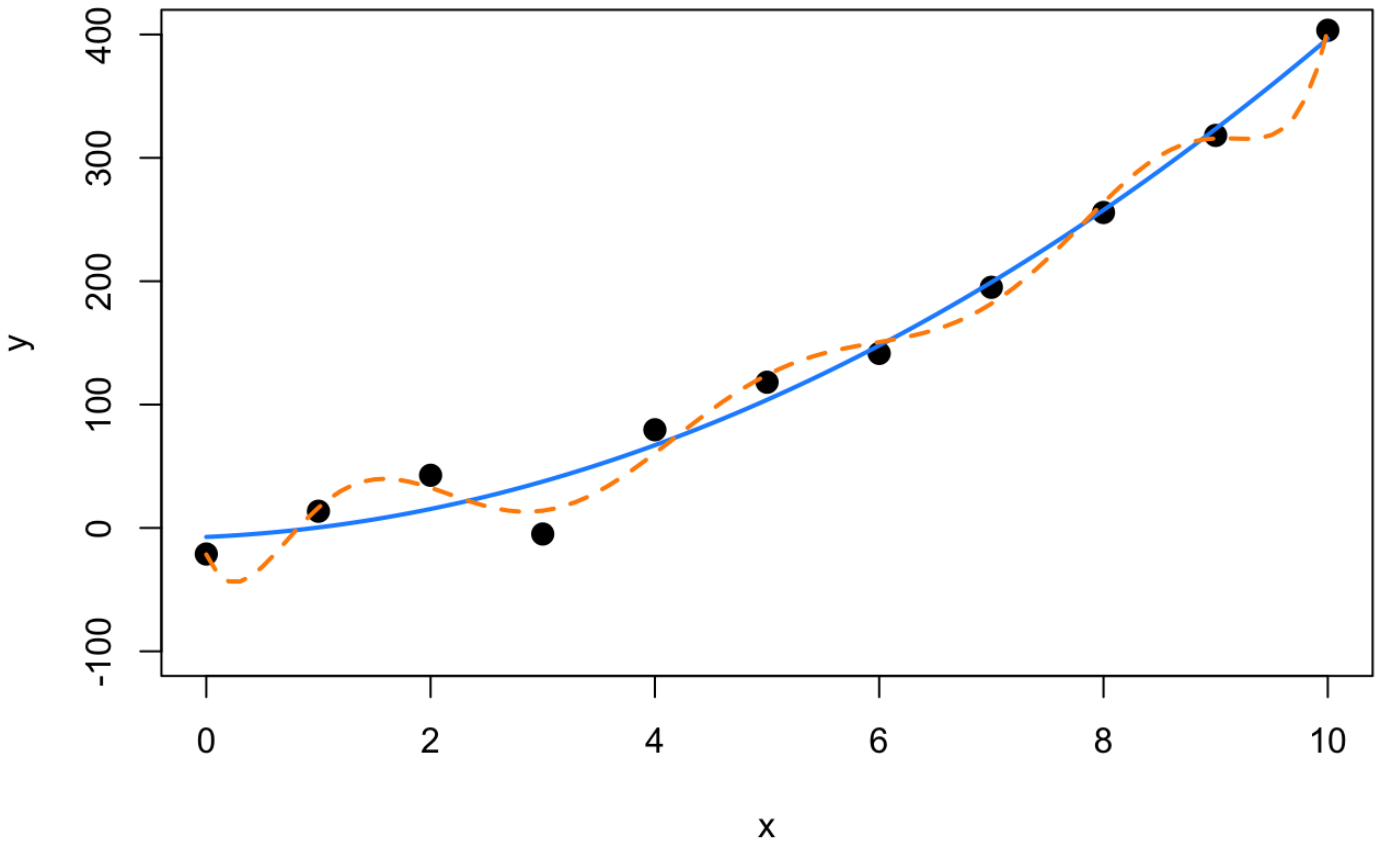
- a GVIF$^{1/(2 \times df)} > 2 = \text{sqrt}(4)$ warrant further investigation.
- a GVIF$^{1/(2 \times df)} > 3.2 \approx \text{sqrt}(10)$ is a sign of serious collinearity requiring correction.

"Choose well. Your choice is brief, and yet endless."
— **Johann Wolfgang von Goethe**

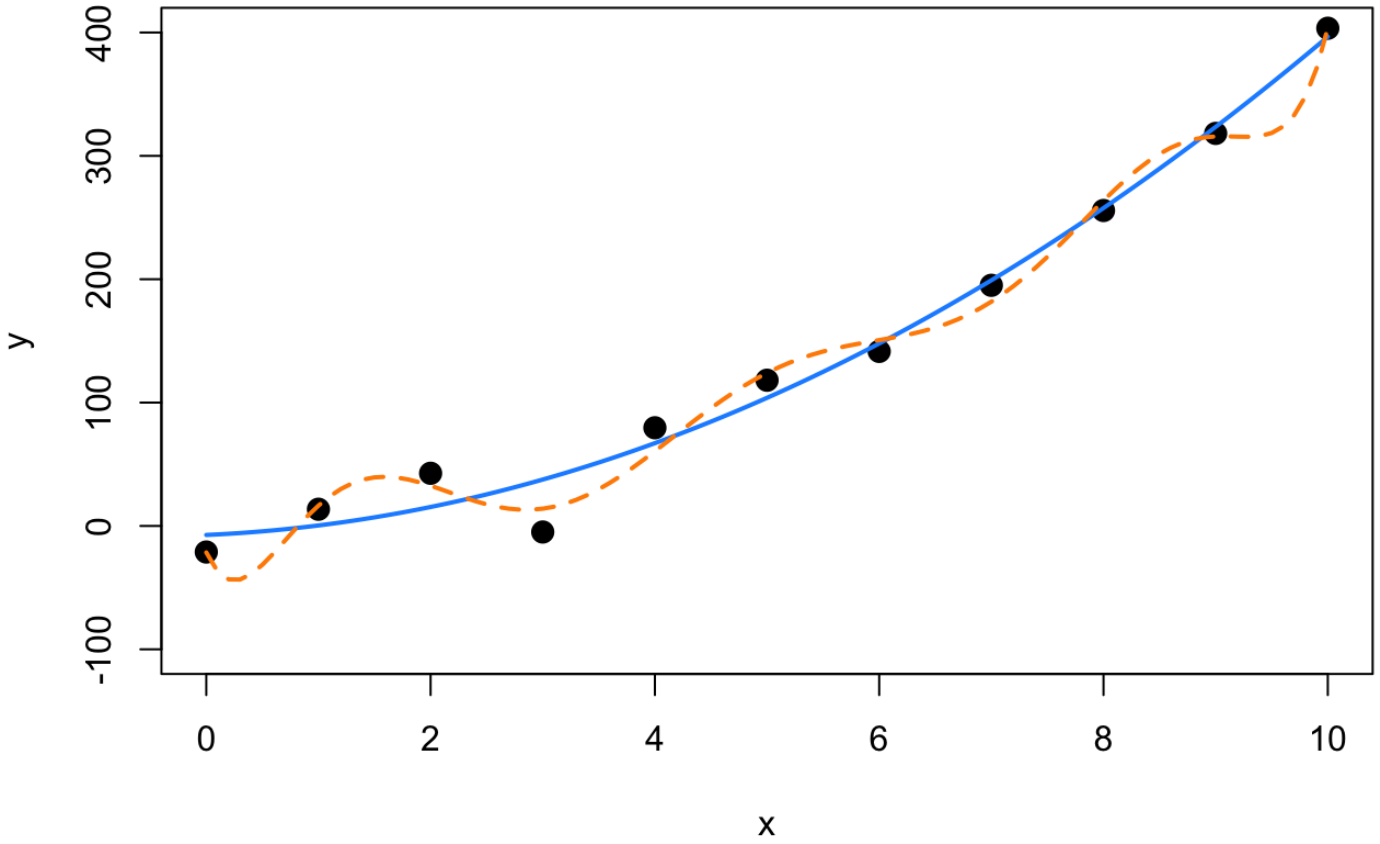
Variable and model selection

In linear regression

Which is better? orange line vs blue line



Which is better? orange line vs blue line



Are you **overfitting** your model?

Why do we need to **choose well**?

Trade-offs between **goodness-of-fit** and **model complexity**

- Use variable selection procedures to find a good model from a set of possible models (e.g., no collinearity issues)

Understand the two uses of models:

1. **explanation**
2. **prediction**

How can we **choose well**?

1. Stepwise (+ backward and forward) variable selection

   Easy to implement in R

2. Variable selection criterion:

- compare the models AIC, C_p and BIC
- Smaller AIC and BIC indicate better model fit

Multiple linear regression

Dependent variable

Independent variables

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + \varepsilon_i$$

Random error

Linear regression

An example: factors associated with cardiovascular risk

Task 0: Construct a multiple linear regression model

```
> # Task 0: Construct a multiple linear regression model
> mod = lm(tcData$chol ~ tcData$age+ tcData$bmi.1+ tcData$bmi.2+ tcData$gender+ tcData$smoker)
> # ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++ #
> # Task 1: Interpret the model output
> summary(mod)

Call:
lm(formula = tcData$chol ~ tcData$age + tcData$bmi.1 + tcData$bmi.2 +
    tcData$gender + tcData$smoker)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5015 -0.6799  0.0074  0.6644  3.5606

Coefficients:
                           Estimate Std. Error t value            Pr(>|t|)
(Intercept)                1.878677   0.137851  13.628 <0.0000000000000002 ***
tcData$age                 0.042216   0.001262  33.444 <0.0000000000000002 ***
tcData$bmi.1               0.010177   0.040052   0.254               0.7994
tcData$bmi.2               0.052434   0.040058   1.309               0.1906
tcData$gendermale          0.008556   0.028249   0.303               0.7620
tcData$smokerex smoker    -0.010881   0.035801  -0.304               0.7612
tcData$smokernever smoker -0.081607   0.032791  -2.489               0.0129 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.02 on 5206 degrees of freedom
Multiple R-squared:  0.1935,    Adjusted R-squared:  0.1926
F-statistic: 208.2 on 6 and 5206 DF,  p-value: < 0.00000000000000022
```

Fit a multiple linear regression model with chol as the outcome and the following variables as the predictors: age, bmi.1, bmi.2, gender, and smoker.

Task 1: Interpret the model output

```
> # Task 0: Construct a multiple linear regression model
> mod = lm(tcData$chol ~ tcData$age+ tcData$bmi.1+ tcData$bmi.2+ tcData$gender+ tcData$smoker)
> # +++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++ #
> # Task 1: Interpret the model output
> summary(mod)

Call:
lm(formula = tcData$chol ~ tcData$age + tcData$bmi.1 + tcData$bmi.2 +
    tcData$gender + tcData$smoker)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5015 -0.6799  0.0074  0.6644  3.5606

Coefficients:
                           Estimate Std. Error t value            Pr(>|t|)
(Intercept)                1.878677   0.137851  13.628 <0.0000000000000002 ***
tcData$age                 0.042216   0.001262  33.444 <0.0000000000000002 ***
tcData$bmi.1               0.010177   0.040052   0.254               0.7994
tcData$bmi.2               0.052434   0.040058   1.309               0.1906
tcData$gendermale          0.008556   0.028249   0.303               0.7620
tcData$smokerex smoker    -0.010881   0.035801  -0.304               0.7612
tcData$smokernever smoker -0.081607   0.032791  -2.489               0.0129 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.02 on 5206 degrees of freedom
Multiple R-squared:  0.1935,    Adjusted R-squared:  0.1926
F-statistic: 208.2 on 6 and 5206 DF,  p-value: < 0.00000000000000022
```

Based on the model you have fitted in Task 1, report and interpret the effect of body mass index on total cholesterol.

Task 1: Interpret the model output

```
> # Task 0: Construct a multiple linear regression model
> mod = lm(tcData$chol ~ tcData$age+ tcData$bmi.1+ tcData$bmi.2+ tcData$gender+ tcData$smoker)
> # ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++ #
> # Task 1: Interpret the model output
> summary(mod)

Call:
lm(formula = tcData$chol ~ tcData$age + tcData$bmi.1 + tcData$bmi.2 +
    tcData$gender + tcData$smoker)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5015 -0.6799  0.0074  0.6644  3.5606

Coefficients:
                           Estimate Std. Error t value            Pr(>|t|)
(Intercept)                1.878677   0.137851  13.628 <0.0000000000000002 ***
tcData$age                 0.042216   0.001262  33.444 <0.0000000000000002 ***
tcData$bmi.1               0.010177   0.040052   0.254              0.7994
tcData$bmi.2               0.052434   0.040058   1.309              0.1906
tcData$gendermale          0.008556   0.028249   0.303              0.7620
tcData$smokerex smoker    -0.010881   0.035801  -0.304              0.7612
tcData$smokernever smoker -0.081607   0.032791  -2.489              0.0129 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.02 on 5206 degrees of freedom
Multiple R-squared:  0.1935,    Adjusted R-squared:  0.1926
F-statistic: 208.2 on 6 and 5206 DF,  p-value: < 0.00000000000000022
```

A one kg/m2 increase in BMI would increase the mean total cholesterol by 0.010 mmol/l and 0.052 mmol/l for the first and second measurement of BMI, respectively, after adjusting for age, gender and smoking status. Hence, both BMI measurements have a positive association with total cholesterol.

However, their p-values are greater than 0.05. As such, the effect of BMI measured before and after physical examination are non-significant, suggesting BMI and total cholesterol are not associated after adjusting for age, gender and smoking status.

Task 2: Collinearity

Generate the variance inflation factors for the model built. Report the variable(s) with a collinearity problem and the corresponding variance inflation factor(s).

Justify your answer.

Task 2: Collinearity

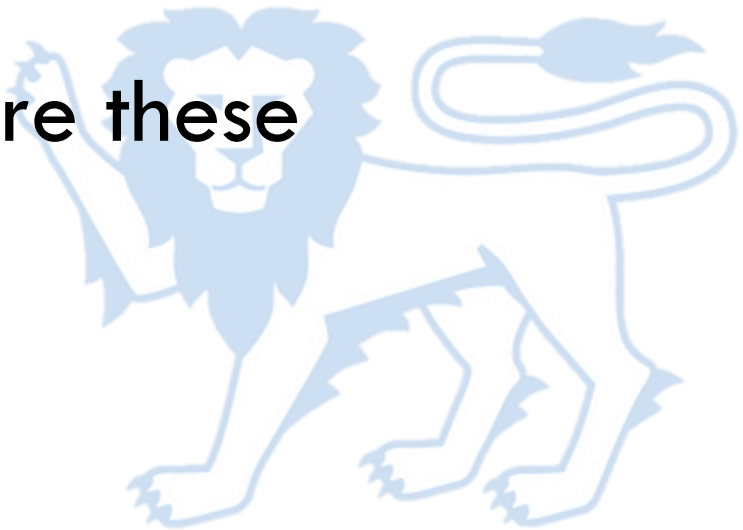
```
> vif(mod)
                  GVIF Df GVIF^(1/(2*Df))
tcData$age    1.000554  1        1.000277
tcData$bmi.1 52.108078  1        7.218593
tcData$bmi.2 52.107294  1        7.218538
tcData$gender 1.000404  1        1.000202
tcData$smoker 1.000724  2        1.000181
> round(cov2cor(vcov(mod)),3)
                       (Intercept) tcData$age tcData$bmi.1 tcData$bmi.2 tcData$gendermale tcData$smokerex smoker tcData$smokernever smoker
(Intercept)                  1.000     -0.377       -0.065       -0.061           -0.094                -0.098                    -0.108
tcData$age                  -0.377      1.000        0.012       -0.011            0.006                 0.017                     0.000
tcData$bmi.1                -0.065      0.012        1.000       -0.990           -0.010                 0.007                    -0.003
tcData$bmi.2                -0.061     -0.011       -0.990        1.000            0.008                -0.008                     0.003
tcData$gendermale           -0.094      0.006       -0.010        0.008            1.000                -0.007                    -0.012
tcData$smokerex smoker      -0.098      0.017        0.007       -0.008           -0.007                 1.000                     0.420
tcData$smokernever smoker   -0.108      0.000       -0.003        0.003           -0.012                 0.420                     1.000
```

Among the predictors, bmi.1 and bmi.2 may have a collinearity problem because their VIFs are 51.108 and 52.107, respectively, and larger than 10.

Task 3: Collinearity

Fit two multiple linear regression models similar to Task 0 where chol is the outcome and the following variables are still predictors: age, gender and smoker, BUT bmi.1 and bmi.2 are included as predictors **in two separate models.**

Based on the models you have fitted, compare these findings with those in Task 1.



Task 3: Collinearity

```
> mod1 = lm(tcData$chol ~ tcData$age+ tcData$bmi.1+ tcData$gender+ tcData$smoker)
> summary(mod1)

Call:
lm(formula = tcData$chol ~ tcData$age + tcData$bmi.1 + tcData$gender +
    tcData$smoker)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5129 -0.6752  0.0081  0.6625  3.5448

Coefficients:
                           Estimate Std. Error t value            Pr(>|t|)
(Intercept)                1.889766   0.137600  13.734 <0.0000000000000002 ***
tcData$age                 0.042235   0.001262  33.458 <0.0000000000000002 ***
tcData$bmi.1               0.062099   0.005550  11.190 <0.0000000000000002 ***
tcData$gendermale          0.008254   0.028250   0.292              0.7702
tcData$smokerex smoker    -0.010491   0.035802  -0.293              0.7695
tcData$smokernever smoker -0.081756   0.032793  -2.493              0.0127 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.02 on 5207 degrees of freedom
Multiple R-squared:  0.1932,    Adjusted R-squared:  0.1925
F-statistic: 249.4 on 5 and 5207 DF,  p-value: < 0.00000000000000022
```

BMI and total cholesterol has a positive association as the estimated effects are greater than 0 after adjusting for age, gender and smoking status.

Task 3: Collinearity

```
> mod2 = lm(tcData$chol ~ tcData$age+ tcData$bmi.2+ tcData$gender+ tcData$smoker)
> summary(mod2)

Call:
lm(formula = tcData$chol ~ tcData$age + tcData$bmi.2 + tcData$gender +
    tcData$smoker)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5010 -0.6792  0.0075  0.6644  3.5636

Coefficients:
                           Estimate Std. Error t value            Pr(>|t|)
(Intercept)                1.880941   0.137551  13.675 <0.0000000000000002 ***
tcData$age                 0.042212   0.001262  33.446 <0.0000000000000002 ***
tcData$bmi.2               0.062515   0.005549  11.265 <0.0000000000000002 ***
tcData$gendermale          0.008626   0.028245   0.305              0.7601
tcData$smokerex smoker    -0.010946   0.035797  -0.306              0.7598
tcData$smokernever smoker -0.081579   0.032787  -2.488              0.0129 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.019 on 5207 degrees of freedom
Multiple R-squared:  0.1935,    Adjusted R-squared:  0.1927
F-statistic: 249.9 on 5 and 5207 DF,  p-value: < 0.00000000000000022
```

BMI and total cholesterol has a positive association as the estimated effects are greater than 0 after adjusting for age, gender and smoking status.

0.062 (=0.010+0.052) is close to the adjusted effect of BMI :

- bmi.1 (i.e., 0.062 in mod1)
- bmi.2 (i.e., 0.063 in mod2).

Task 3: Collinearity

```
> mod2 = lm(tcData$chol ~ tcData$age+ tcData$bmi.2+ tcData$gender+ tcData$smoker)
> summary(mod2)

Call:
lm(formula = tcData$chol ~ tcData$age + tcData$bmi.2 + tcData$gender +
    tcData$smoker)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5010 -0.6792  0.0075  0.6644  3.5636

Coefficients:
                          Estimate Std. Error t value            Pr(>|t|)
(Intercept)               1.880941   0.137551  13.675 <0.0000000000000002 ***
tcData$age                0.042212   0.001262  33.446 <0.0000000000000002 ***
tcData$bmi.2              0.062515   0.005549  11.265 <0.0000000000000002 ***
tcData$gendermale         0.008626   0.028245   0.305              0.7601
tcData$smokerex smoker   -0.010946   0.035797  -0.306              0.7598
tcData$smokernever smoker -0.081579   0.032787  -2.488              0.0129 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.019 on 5207 degrees of freedom
Multiple R-squared:  0.1935,    Adjusted R-squared:  0.1927
F-statistic: 249.9 on 5 and 5207 DF,  p-value: < 0.00000000000000022
```

The SE of the adjusted effects of BMI in Task 1 is much larger than Task 3

Task 1:
0.04005 for bmi.1
0.04006 and bmi.2
vs
Task 3: 0.00555 for both bmi.1 and bmi.2.

Task 4: Collinearity

Generate the variance inflation factors for the model built in Task 3. Report the variable(s) with a collinearity problem and the corresponding variance inflation factor(s).

Justify your answer

Task 4: Collinearity

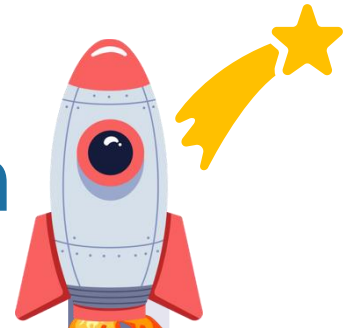
```
> vif(mod1)
                GVIF Df GVIF^(1/(2*Df))
tcData$age    1.000423  1        1.000211
tcData$bmi.1  1.000250  1        1.000125
tcData$gender 1.000338  1        1.000169
tcData$smoker 1.000595  2        1.000149
> vif(mod1)
                GVIF Df GVIF^(1/(2*Df))
tcData$age    1.000423  1        1.000211
tcData$bmi.1  1.000250  1        1.000125
tcData$gender 1.000338  1        1.000169
tcData$smoker 1.000595  2        1.000149
```

In both models, there are no variables with a collinearity problem as their VIF is close to 1 and less than 4 (or 10).

TM-CM02

Biostatistics for Public Health

- quizzes
- 2 assignments



4

Logistic regression

Data prep, descriptive & inferential statistics, logistics regression and others if we have time

3

Linear regression

Data prep, descriptive & inferential statistics, linear regression

2

Hypothesis testing

Inferential statistics

1

Categorical data analysis

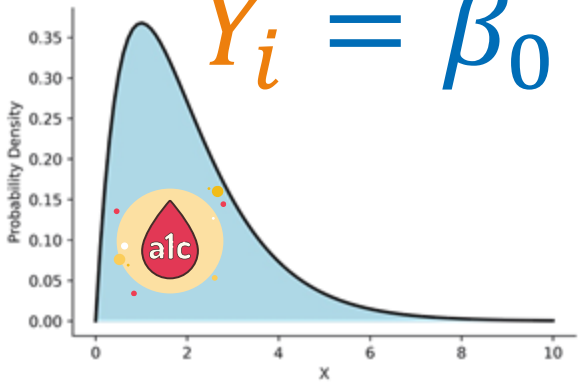
Data prep, descriptive & inferential statistics

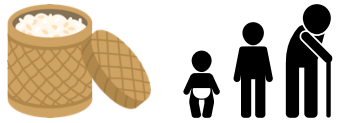
Multiple linear regression

Dependent variable

Independent variables

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + \varepsilon_i$$





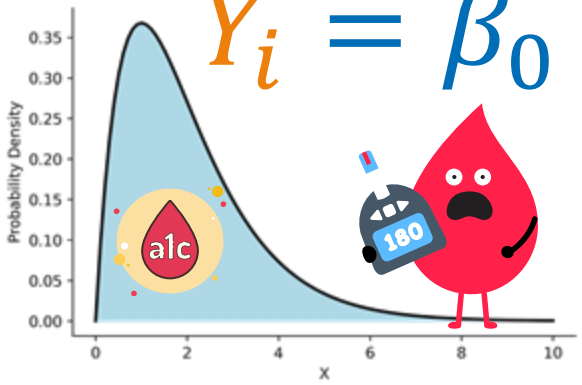
Random error

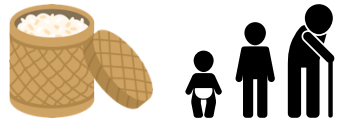
Logistic regression

Dependent variable

Independent variables

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + \varepsilon_i$$





Random error

Dependent variable

0: no T2DM

1: T2DM

What type of data? Which model?

Multiple linear regression is suitable for a numerical continuous outcome variable.

- **Model:** Multiple linear regression
- **What is modelled:** The mean value of the outcome

Example: To describe the linear relationship between the response variable(weight) and explanatory variables (age and sex).

- The model can be used to determine whether variations in weight could be explained by age and/or sex:

$$\text{Mean weight} = \beta_0 + \beta_1(\text{age}) + \beta_2(\text{sex})$$

In more general terms, a linear model is written:

$$\text{Mean}\,(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- where β_0 is the intercept and β_1 and β_2 are the slopes for variables X_1 and X_2 respectively

What type of data? Which model?

Linear regression

Only appropriate for continuous response variables, which have a Normal distribution at each level of the predictor variable but it is not appropriate for modelling the risk or prevalence of disease, measures which are more common in epidemiology.

We use a different type of statistical model when fitting for risk or prevalence studies. In this model, the log odds of disease are used as the measure of disease outcome.

Binary outcome

Model: Logistic regression

What is modelled: The log of the odds of the outcome

But why do we use log odds?

Thank you