



# *Chapter 4*

## *Multiple Linear Regression*

# STAT210/410 Study Plan

Topic	Weeks covered	Readings	Assessment
<b>Topic 1: Simple Linear regression (SLR)</b>	Wk 1	Chapter 3	Online Quiz due 9 <sup>th</sup> March
<b>Topic 2: Multiple Linear Regression (MLR)</b>	Wk2 & 3	Chapter 4	Written Assessment A2 due 23 <sup>rd</sup> March
<b>Topic 3: Model building</b>	Wk 4	Chapter 5	
<b>Topic 4: Variable Screening and regression pitfalls</b>	Wk 5	Chapters 6, 7	
<b>Topic 5: Residual Analysis</b>	Wk 6	Chapter 8	Written Assessment A3 due 13 <sup>th</sup> April
<b>Topic 6 Generalised Linear Models (GLMs)</b>	Wk 9 & 10	Chapter 9	
<b>Topic 7: Principles of Experimental Design</b>	Wk 11	Chapter 11	Written Assessment A4 due 11 <sup>th</sup> May
<b>Topic 8: ANOVA, contrasts</b>	Wk 12 & 13	Chapter 12	
<b>STAT410 ONLY</b>			
<b>ART: Nonparametric Regression</b>		Section 9.9	Written Assessment ART due 18 <sup>th</sup> May



# Chapter 4 Outline

## Lecture 1

- ❖ Intro to MLR
- ❖ Fitting the model, testing the overall utility of a model
- ❖ Interpreting regression coefficients

## Lecture 2

- ❖ Inferences about the individual  $\beta_i$
- ❖ Multiple Coefficients of determination,  $R^2$  and  $R^2_{\text{adj}}$
- ❖ Using the model for estimation and prediction

## Lecture 3

- ❖ An interaction model with quantitative predictors

## Lecture 4

- ❖ Models with qualitative predictors

NB: Sections 4.11, 4.13 and 4.14 of the text will **not** be covered



# Lecture 1

Multiple Linear Regression

# Chapter 4 Outline

## Lecture 1

- ❖ Intro to MLR
- ❖ Fitting the model, testing the overall utility of a model
- ❖ Interpreting regression coefficients

## Lecture 2

- ❖ Inferences about the individual  $\beta_i$
- ❖ Multiple Coefficients of determination,  $R^2$  and  $R^2_{\text{adj}}$
- ❖ Using the model for estimation and prediction

## Lecture 3

- ❖ An interaction model with quantitative predictors

## Lecture 4

- ❖ Models with qualitative predictors

NB: Sections 4.11, 4.13 and 4.14 of the text will **not** be covered

# Intro to MLR



## General Form of the Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where  $y$  is the dependent variable

$x_1, x_2, \dots, x_k$  are the independent variables

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$  is the deterministic portion of the model

$\beta_i$  determines the contribution of the independent variable  $x_i$

*Note:* The symbols  $x_1, x_2, \dots, x_k$  may represent higher-order terms for quantitative predictors (e.g.,  $x_2 = x_1^2$ ) or terms for qualitative predictors.

# MLR model assumptions

Multiple regression model:

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}_{\text{Deterministic portion}} + \underbrace{\epsilon}_{\text{random error}}$$

## Assumptions

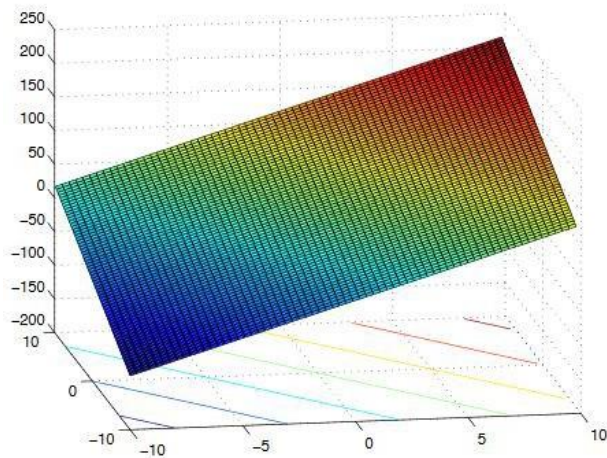
- ❖ residuals  $\epsilon$  are independent
- ❖ residuals  $\epsilon \sim N(0, \sigma^2)$ 
  - Normally distributed
  - mean 0
  - Variance  $\sigma^2$  are constant w.r.t.  $x$

# MLR



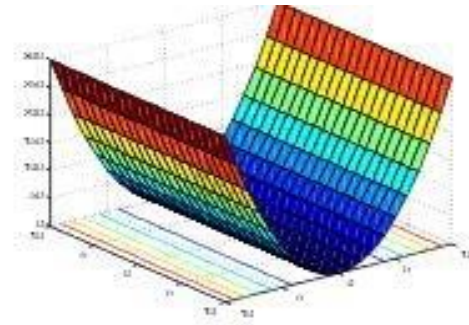
*First-order model (main effects model)*

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \epsilon$$

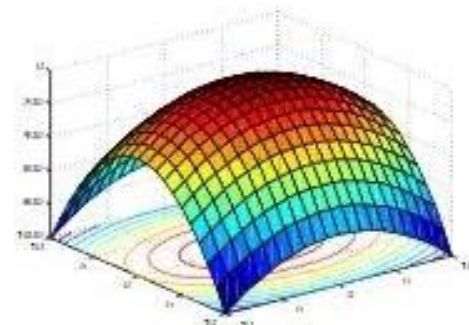


*Second-order models*

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2$$



$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{12}x_1x_2 + \epsilon$$



# Fitting the model

The method of fitting MLR is identical to that of the straight-line model  
i.e. the method of least squares

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

To estimate  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$  we minimize the  $SSE = \sum (y_i - \hat{y}_i)^2$

i.e. minimize  $SSE = \sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k))^2$

The sample estimate  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  are the solutions to the set of simultaneous equations

$$\frac{\partial SSE}{\partial \beta_0} = 0; \quad \frac{\partial SSE}{\partial \beta_1} = 0 \dots; \quad \frac{\partial SSE}{\partial \beta_k} = 0 \quad (\text{proof omitted, see Appendix B})$$

# Example

The dataset *SampCountries.txt* contains information about 49 countries. There are 5 variables:

- LifeExp = life expectancy
- Population = population (in millions)
- Health = the percentage of government expenditure on health care
- Internet = percentage of people having internet
- BirthRate = birth rate (births per 1000)

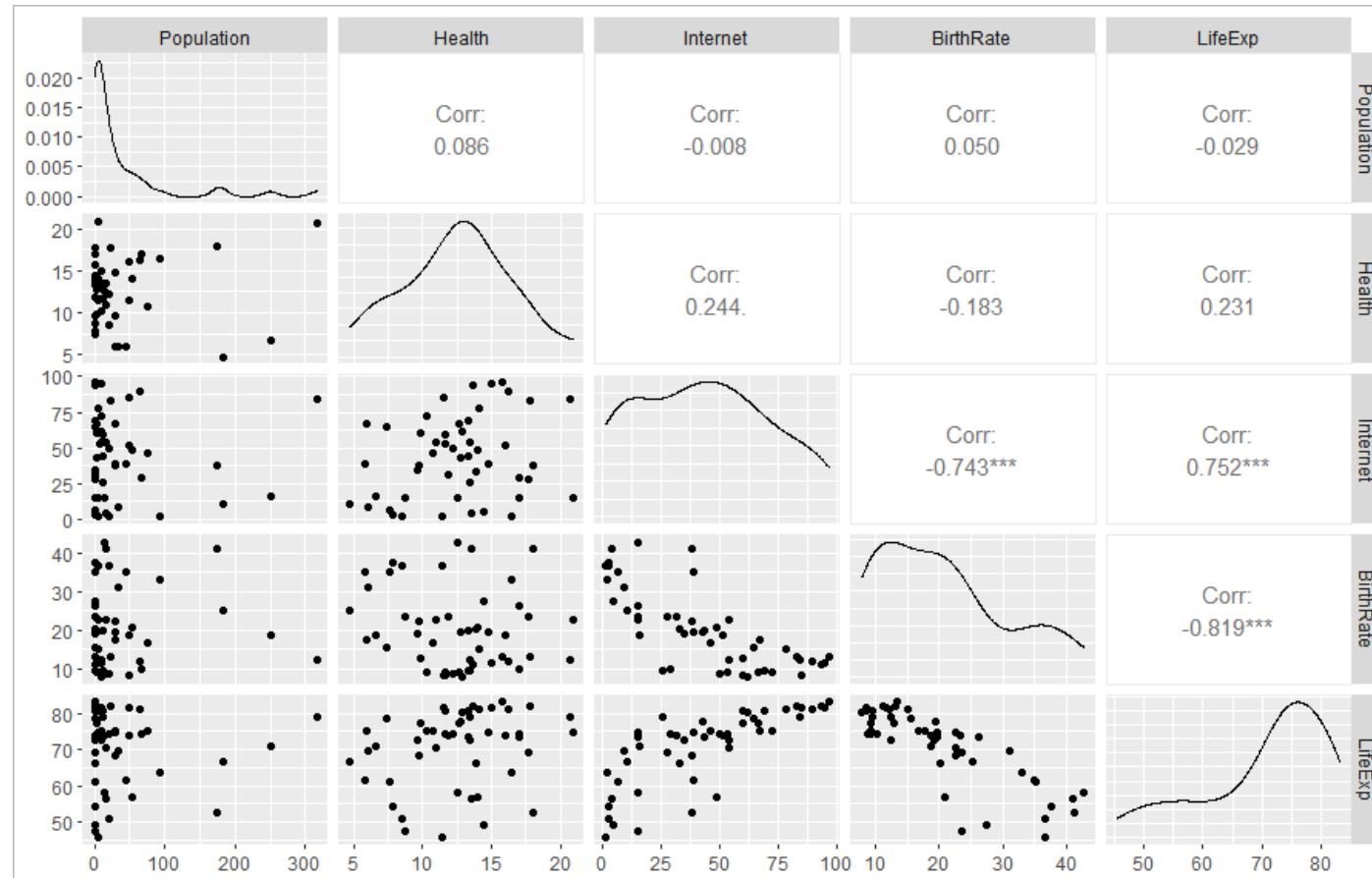
Fit a model of the form

$$\text{LifeExp} = \beta_0 + \beta_1 \text{Population} + \beta_2 \text{Health} + \beta_3 \text{Internet} + \beta_4 \text{BirthRate} + \epsilon$$

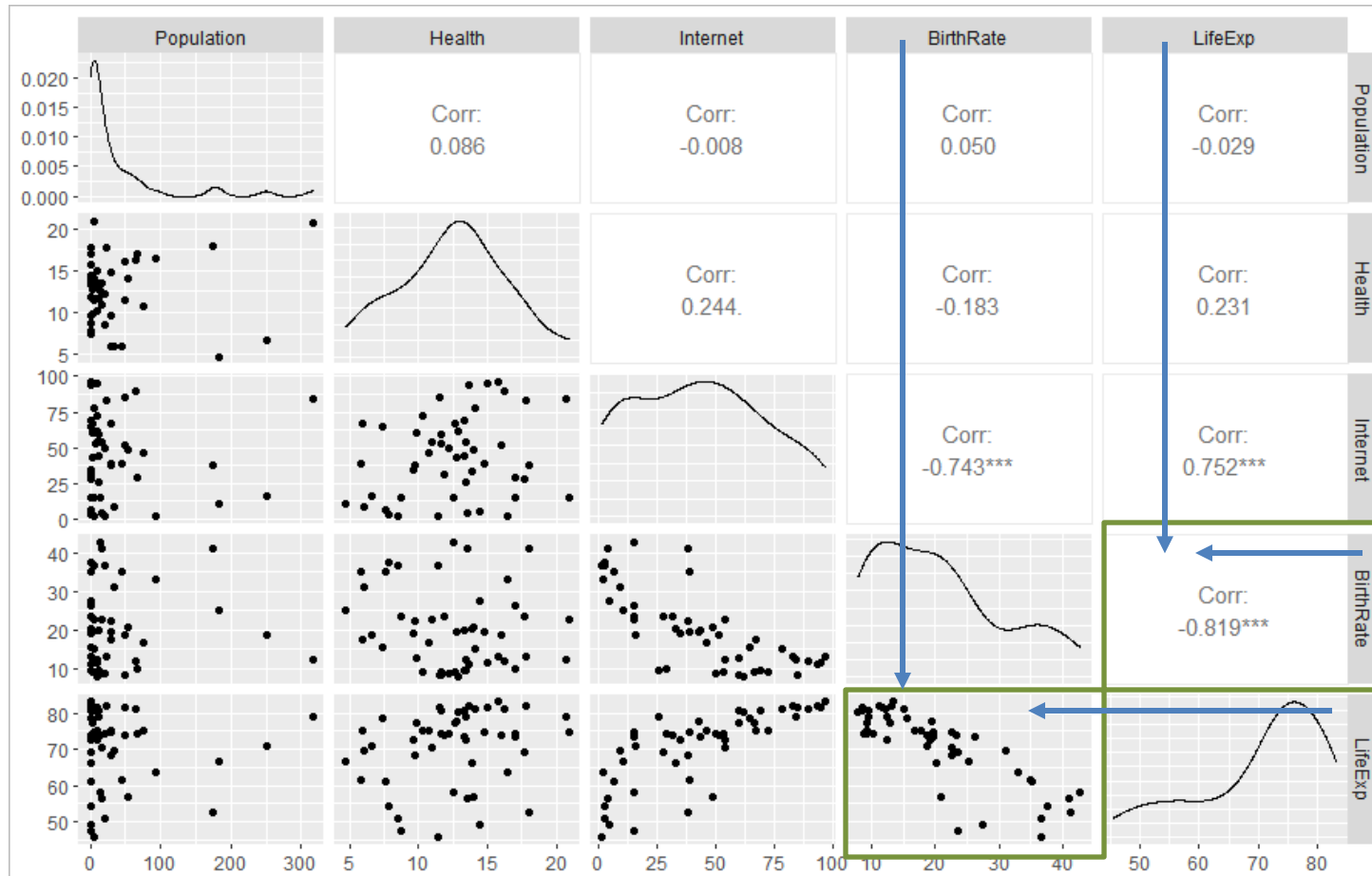
# Pairs Plot

Don't forget to install/load the GGally package before producing the plot

```
data <- read.table("SampCountries.txt", header=T)
library(GGally)
ggpairs(data)
```



# Exploratory Plot



**Q:** Interpret this pairs plot.

# Exploratory Plot

From the pairs plot, we see that

- The response variable *LifeExp* has a strong positive correlation with Internet ( $r = 0.75$ ) and a strong negative correlation with BirthRate ( $r = -0.82$ ).
- There appears a weak correlation between LifeExp with both Population and Health
- Between the 4 predictors, it's noticed that Internet is strongly correlated with BirthRate ( $r = -0.74$ ).

# LifeExp summary table

```
mod1<-lm(LifeExp ~Population+Health + Internet + BirthRate, data =countries.df)
summary(mod1)
```

```
#####
```

Estimated of  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	76.239603	4.834725	15.77	< 2e-16
Population	-0.000333	0.012823	-0.03	0.979
Health	0.131375	0.222374	0.59	0.558
Internet	0.111761	0.043780	2.55	0.014
BirthRate	-0.594465	0.122328	-4.86	1.5e-05

**Residual standard error: 5.72** on 44 degrees of freedom

Multiple R-squared: 0.72, Adjusted R-squared: 0.694

F-statistic: 28.2 on 4 and 44 DF, p-value: 1.19e-11

$$\widehat{\text{Life expectancy}} = 76.24 - 0.0003\text{Population} + 0.13\text{Health} + 0.11\text{Internet} - 0.59\text{BirthRate}$$

## Estimator of $\sigma^2$ for Multiple Regression Model with $k$ Independent Variables

$$\hat{\sigma}^2 = s^2 = MSE = \frac{SSE}{n - (k + 1)}$$

$k$  = Number of  $\beta$  parameters fitted to the model. EG number of  $\hat{\beta}_i$

# LifeExp: ANOVA table

```
anova(mod1)
```

```
#####
```

```
Analysis of Variance
```

```
Response: LifeExp
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Population	1	4.44	4.44	0.1357	0.714323
Health	1	280.95	280.95	8.5966	0.005327 **
Internet	1	2634.46	2634.46	80.6088	1.664e-11 ***
BirthRate	1	771.81	771.81	23.6159	1.531e-05 ***
Residuals	44	1438.01	32.68		

$$\text{MSE} = 32.68 = \hat{\sigma}^2$$

$$\text{residual std error} = s = 5.72$$

$$s^2 = 5.72^2 = 32.71$$

Estimate of  $\sigma^2$

# Model's utility

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

## Testing **Global Usefulness** of the Model: The Analysis of Variance *F*-Test

$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$  (All model terms are unimportant for predicting  $y$ )

$H_a$ : At least one  $\beta_i \neq 0$  (At least one model term is useful for predicting  $y$ )

$$\begin{aligned} \text{Test statistic: } F &= \frac{(\text{SS}_{yy} - \text{SSE})/k}{\text{SSE}/[n - (k + 1)]} = \frac{R^2/k}{(1 - R^2)/[n - (k + 1)]} \\ &= \frac{\text{Mean square (Model)}}{\text{Mean square (Error)}} \end{aligned}$$

where  $n$  is the sample size and  $k$  is the number of terms in the model.

*Rejection region:*  $F > F_\alpha$ , with  $k$  numerator degrees of freedom and  $[n - (k + 1)]$  denominator degrees of freedom.

or

$\alpha > p\text{-value}$ , where  $p\text{-value} = P(F > F_c)$ ,  $F_c$  is the computed value of the test statistic.

*Assumptions:* The standard regression assumptions about the random error component (Section 4.2).

# LifeExp summary table

```
mod1<-lm(LifeExp ~Population+Health + Internet + BirthRate, data =countries.df)
summary(mod1)
```

```
#####
```

Estimated of  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	76.239603	4.834725	15.77	< 2e-16
Population	-0.000333	0.012823	-0.03	0.979
Health	0.131375	0.222374	0.59	0.558
Internet	0.111761	0.043780	2.55	0.014
BirthRate	-0.594465	0.122328	-4.86	1.5e-05

**Residual standard error: 5.72** on 44 degrees of freedom

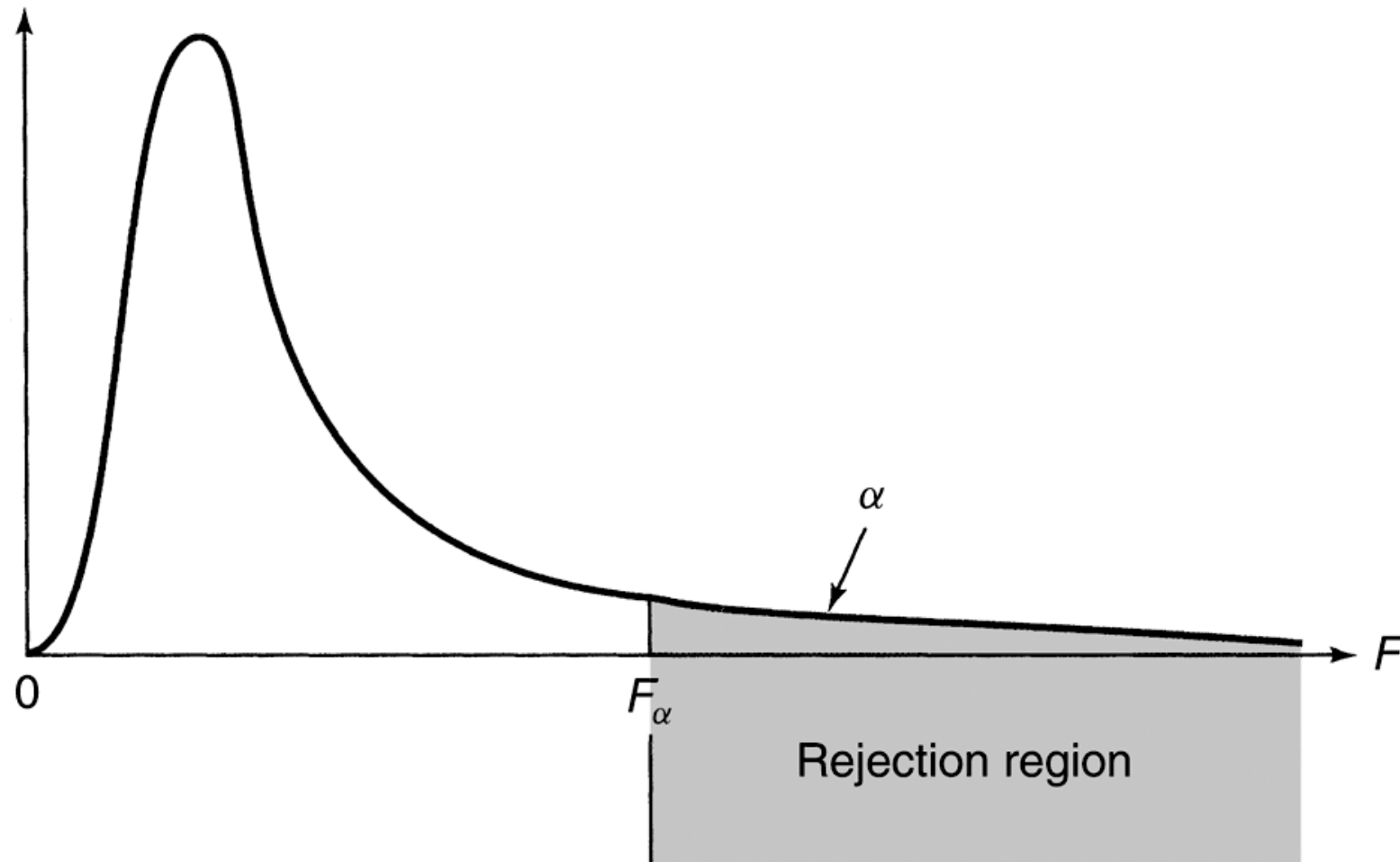
Multiple R-squared: 0.72, Adjusted R-squared: 0.694

F-statistic: 28.2 on 4 and 44 DF, p-value: 1.19e-11

$$\widehat{\text{Life expectancy}} = 76.24 - 0.0003\text{Population} + 0.13\text{Health} + 0.11\text{Internet} - 0.59\text{BirthRate}$$

## Figure 4.4

### Rejection region for the global $F$ -test



# LifeExp model's utility

$\text{LifeExp} = 76.24 - 0.0003 * \text{Population} + 0.13 * \text{Health} + 0.11 * \text{Internet} - 0.59 * \text{BirthRate}$

```
mod1<-lm(LifeExp ~Population+Health + Internet + BirthRate, data =countries.df)
summary(mod1)
```

#####

Estimated of  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	76.239603	4.834725	15.77	< 2e-16
Population	-0.000333	0.012823	-0.03	0.979
Health	0.131375	0.222374	0.59	0.558
Internet	0.111761	0.043780	2.55	0.014
BirthRate	-0.594465	0.122328	-4.86	1.5e-05

**Residual standard error: 5.72** on 44 degrees of freedom

Multiple R-squared: 0.72, Adjusted R-squared: 0.694

**F-statistic: 28.2 on 4 and 44 DF, p-value: 1.19e-11**

Conclusion about the model's utility?

# LifeExp model's utility



**Test the overall utility of the model using the global F-test at  $\alpha = 0.05$ .**

The null hypothesis for the global test is

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$  or all of the predictors are not useful.

$H_a$  : At least 1 of the coefficients is non-zero (or at least 1 of the predictors is useful)

From the summary table, the test statistic is  $F = 28.2$  on 4 and 44 df,  $p\text{-value} = 1.19 \times 10^{-11} \approx 0$ ).

*Since the  $p$ -value is much smaller than the threshold  $\alpha = 0.05$ , the data provides strong evidence against  $H_0$ . We reject  $H_0$  and conclude that at least 1 of the model coefficients is non-zero or at least 1 of the predictors is useful to predict LifeExp. The overall model appears to be statistically useful for predicting life expectancy.*

# Caution When interpreting global F value



- ❖ Global usefulness does not equal best model
- ❖ Better ways to choose “best” model (Week 4)
- ❖ Models must pass this test first before checking individual  $\beta_i$
- ❖ Tells you ONLY that at least one parameter in the model is useful, not which ones

# Interpretation of regression coefficients

- ❖ In SLR,  $\beta_1$  can be interpreted as the expected increase/decrease in the response variable,  $y$ , when the predictor,  $x$ , is increased by *one unit*.
- ❖ In *MLR* (i.e., more than one quantitative predictor),  $\beta_i$  is the expected change in the response when the value of  $x_i$  is increased by one unit ***provided the other predictors remain unchanged***.

Q: Interpret regression coefficient for Internet?

# Q: Interpret regression coefficient for Internet?

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	76.239603	4.834725	15.77	< 2e-16
Population	-0.000333	0.012823	-0.03	0.979
Health	0.131375	0.222374	0.59	0.558
Internet	0.111761	0.043780	2.55	0.014
BirthRate	-0.594465	0.122328	-4.86	1.5e-05

$$\text{LifeExp} = 76.24 - 0.0003 * \text{Population} + 0.13 * \text{Health} + 0.11 * \text{Internet} - 0.59 * \text{BirthRate}$$

# Interpretation of regression coefficients

For every 1% increase in internet usage in the population, life expectancy increases by 0.11 years **provided that the health, birth rate and population remain unchanged.**

Life expectancy is expected to increase by 0.11 years for every 1% increase in internet use in the population **provided that the other variables in the model remain unchanged.**

$$\text{LifeExp} = 76.24 - 0.0003 * \text{Population} + 0.13 * \text{Health} + 0.11 * \text{Internet} - 0.59 * \text{BirthRate}$$

Q: Interpret regression coefficient for BirthRate?

# Q: Interpret regression coefficient for BirthRate?

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	76.239603	4.834725	15.77	< 2e-16
Population	-0.000333	0.012823	-0.03	0.979
Health	0.131375	0.222374	0.59	0.558
Internet	0.111761	0.043780	2.55	0.014
BirthRate	-0.594465	0.122328	-4.86	1.5e-05

$$\text{LifeExp} = 76.24 - 0.0003 * \text{Population} + 0.13 * \text{Health} + 0.11 * \text{Internet} - 0.59 * \text{BirthRate}$$

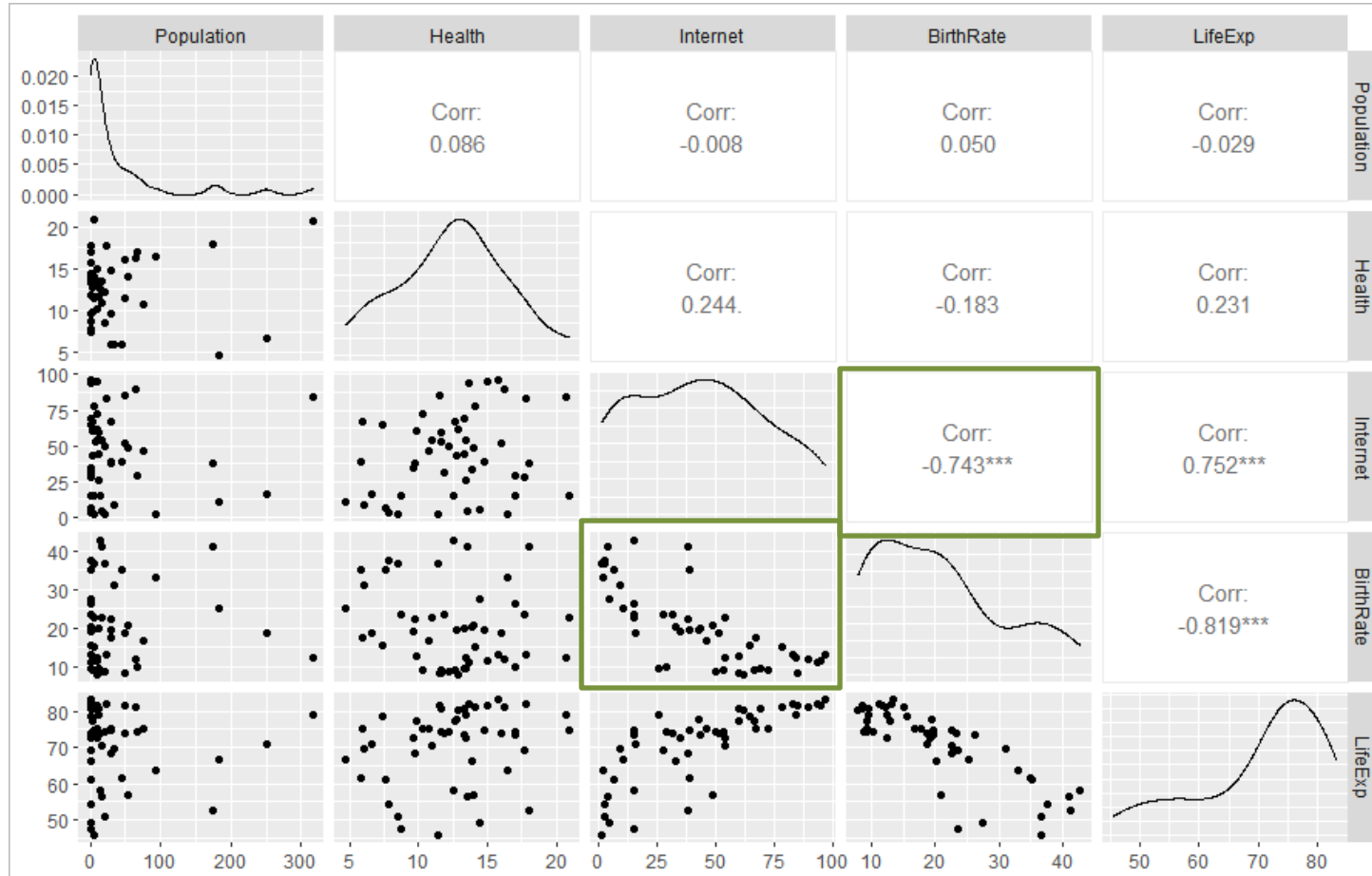
# ***Partial** regression coefficients*

Hence the parameters,  $\beta_1, \dots, \beta_k$ ,  
are called ***partial** regression coefficients*.

**Caution:** trying to interpret partial regression parameters by holding all other predictors constant is very dangerous when the predictor variables are *correlated*.

A change in one predictor variable will result in changes to some (or all) of the other predictors.

# Exploratory Plot



# Next lecture

## Lecture 1

- ❖ Intro to MLR
- ❖ Fitting the model, testing the overall utility of a model
- ❖ Interpreting regression coefficients

## Lecture 2

- ❖ Inferences about the individual  $\beta_i$
- ❖ Multiple Coefficients of determination,  $R^2$  and  $R^2_{\text{adj}}$
- ❖ Using the model for estimation and prediction

## Lecture 3

- ❖ An interaction model with quantitative predictors

## Lecture 4

- ❖ Models with qualitative predictors

NB: Sections 4.11, 4.13 and 4.14 of the text will **not** be covered



# Lecture 2

Multiple Linear Regression

# Inference about an individual parameter $\beta_i$



Multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

1.  $H_0: \beta_i=0$ , *given that the other predictors have already been fitted in the model*
2. Test statistic,  $t = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$
3.  $df = n-(k+1)$
4. p-value
5. Conclusion (reject or fail to reject  $H_0$ )

# LifeExp

$$LifeExp = \beta_0 + \beta_1 Population + \beta_2 Health + \beta_3 Internet + \beta_4 BirthRate + \varepsilon$$

```
mod1<-lm(LifeExp ~Population+Health + Internet + BirthRate, data =countries.df)
```

```
summary(mod1)
```

```
#####
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	76.239603	4.834725	15.77	< 2e-16
Population	-0.000333	0.012823	-0.03	0.979
Health	0.131375	0.222374	0.59	0.558
Internet	0.111761	0.043780	2.55	0.014
BirthRate	-0.594465	0.122328	-4.86	1.5e-05

**Residual standard error: 5.72** on 44 degrees of freedom

Multiple R-squared: 0.72, Adjusted R-squared: 0.694

F-statistic: 28.2 on 4 and 44 DF, p-value: 1.19e-11

# LifeExp



```
mod1<-lm(LifeExp ~Population+Health + Internet + BirthRate, data =countries.df)
```

```
summary(mod1)
```

```
#####
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	76.239603	4.834725	15.77	< 2e-16
Population	-0.000333	0.012823	-0.03	0.979
Health	0.131375	0.222374	0.59	0.558
Internet	0.111761	0.043780	2.55	0.014
BirthRate	-0.594465	0.122328	-4.86	1.5e-05

Tests  $H_0: \beta_i=0$ ,  
given that the  
other predictors  
have been fitted

**Residual standard error: 5.72** on 44 degrees of freedom

Multiple R-squared: 0.72, Adjusted R-squared: 0.694

F-statistic: 28.2 on 4 and 44 DF, p-value: 1.19e-11

# LifeExp

```
mod1<-lm(LifeExp ~Population+Health + Internet + BirthRate, data =countries.df)
```

```
summary(mod1)
```

```
#####
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	76.239603	4.834725	15.77	< 2e-16
Population	-0.000333	0.012823	-0.03	0.979
Health	0.131375	0.222374	0.59	0.558
Internet	0.111761	0.043780	2.55	0.014
BirthRate	-0.594465	0.122328	-4.86	1.5e-05

Tests  $H_0: \beta_i=0$ ,  
given that the  
other predictors  
have been fitted

**Residual standard error: 5.72** on 44 degrees of freedom

Multiple R-squared: 0.72, Adjusted R-squared: 0.694

F-statistic: 28.2 on 4 and 44 DF, p-value: 1.19e-11

$$t = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)} = \frac{-0.594465}{0.122328} = -4.86$$

# Inference about an individual parameter $\beta_i$



Which variables are significant predictors of life expectancy in this model, at a 5% level? Write down and test appropriate hypotheses.

## Population:

$H_0 : \beta_{\text{population}} = 0$  or *Population* is not a significant predictor, given Health, Internet and BirthRate are already in the model

$t = -0.03$ ,  $p = 0.979$ . The p-value is more than the threshold of 0.05, so it is not significant, fail to reject the null hypothesis.

...

## Health:

$H_0 : \beta_{\text{Health}} = 0$  or *Health* is not a significant predictor, given Birthrate, Population and Internet are already in the model.

$t = 0.59$ ,  $p = 0.558$ . The p-value is more than the threshold 0.05, thus it's not significant, and so we fail to reject the null hypothesis.

Conclusion: *Population* and *Health* are not useful/significant predictors for life expectancy.

# Inference about an individual parameter $\beta_i$



Which variables are significant predictors of life expectancy in this model, at a 5% level? Write down and test appropriate hypotheses.

## **Internet:**

$H_0 : \beta_{\text{Internet}} = 0$  or *Internet* is not a significant predictor, given Health, Population and BirthRate are already in the model  
 $t = 2.55$ ,  $p = 0.014$ . The p-value is less than the threshold 0.05, thus it's significant, and so we reject the null hypothesis.

...

## **Birth rate:**

$H_0 : \beta_{\text{BirthRate}} = 0$  or *BirthRate* is not a significant predictor, given Health, Population and Internet are already in the model.  
 $t = -4.86$ ,  $p = 1.5 \times 10^{-5}$ . The p-value is less than the threshold 0.05, thus it's significant, and so we reject the null hypothesis.

Conclusion: *Internet and BirthRate* are useful/significant predictors for life expectancy.

# Inference about an individual parameter $\beta_i$



Which variables are significant predictors of life expectancy in this model, at a 5% level? Write down and test appropriate hypotheses.

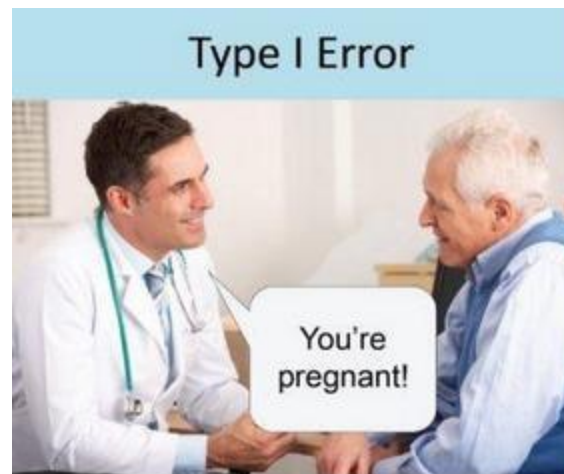
	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	76.239603	4.834725	15.77	< 2e-16	Not a good linear predictor
Population	-0.000333	0.012823	-0.03	0.979	
Health	0.131375	0.222374	0.59	0.558	
Internet	0.111761	0.043780	2.55	0.014	A good linear predictor
BirthRate	-0.594465	0.122328	-4.86	1.5e-05	

Conclusion: Of the predictor variables fitted, only *Internet* and *BirthRate* are useful/significant predictors for life expectancy.

# Inference about an individual parameter $\beta_i$

- ❖ Extreme care should be taken when conducting t-test on individual  $\beta_i$
- ❖ If you fail to reject  $H_0: \beta_i = 0$ , it could be:
  - No relationship between  $y$  and  $x_i$
  - A linear relationship between  $y$  and  $x_i$  exists (holding other variables constant) but a type II of error occurred
  - A relationship between  $y$  and  $x_i$  exists **but it is not linear**.
- ❖ So what we should say if we do not reject  $H_0$ ?  
→ There is insufficient evidence of a linear relationship between  $y$  and  $x_i$

Image source:  
[unbiasedresearch.blogspot.com](http://unbiasedresearch.blogspot.com)



# Refit the model

```
mod2 <- lm(LifeExp ~ Internet + BirthRate, data = countries.df)
```

```
summary(mod2)
```

```
#####
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	77.6921	4.0881	19.00	< 2e-16
Internet	0.1160	0.0424	2.74	0.0088
BirthRate	-0.5949	0.1198	-4.96	9.9e-06

```
Residual standard error: 5.61 on 46 degrees of freedom
```

```
Multiple R-squared: 0.717, Adjusted R-squared: 0.705
```

```
F-statistic: 58.4 on 2 and 46 DF, p-value: 2.37e-13
```

$$\widehat{Life\ Expectancy} = 77.69 + 0.12Internet - 0.59BirthRate$$



**A 100 (1 –  $\alpha$ )% Confidence Interval for a  $\beta$  Parameter**

$$\hat{\beta}_i \pm (t_{\alpha/2})s_{\hat{\beta}_i}$$

where  $t_{\alpha/2}$  is based on  $n - (k + 1)$  degrees of freedom and

$n$  = Number of observations

$k + 1$  = Number of  $\beta$  parameters in the model

```
> qt(0.025, df=46)
[1] -2.012896
```

### A 100 (1 - $\alpha$ )% Confidence Interval for a $\beta$ Parameter

$$\hat{\beta}_i \pm (t_{\alpha/2})s_{\hat{\beta}_i}$$

where  $t_{\alpha/2}$  is based on  $n - (k + 1)$  degrees of freedom and

$n$  = Number of observations

$k + 1$  = Number of  $\beta$  parameters in the model

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	77.6921	4.0881	19.00	< 2e-16
Internet	0.1160	0.0424	2.74	0.0088
BirthRate	-0.5949	0.1198	-4.96	9.9e-06

$$CI_{BirthRate} = -0.59 \pm -2.01 * 0.12$$

# 95% CIs for the regression coefficients (parameters)



```
mod2 <- lm(LifeExp ~ Internet + BirthRate, data = countries.df)
confint(mod2)
```

	2.5 %	97.5 %
(Intercept)	69.46	85.92
<b>Internet</b>	<b>0.03</b>	<b>0.20</b>
BirthRate	-0.84	-0.35

The 95% CI for  $\beta_{\text{internet}}$  is (0.03, 0.2) suggests that if birth rate is held constant then for each percentage **increase** in having internet access, the average life expectancy will increase between 0.03 and 0.2 years

# 95% CIs for the regression coefficients (parameters)

	2.5 %	97.5 %
(Intercept)	69.46	85.92
Internet	0.03	0.20
<b>BirthRate</b>	<b>-0.84</b>	<b>-0.35</b>

The 95% CI for  $\beta_{\text{Birthrate}}$  is (-0.84, -0.35) suggests that if percentage of having internet access is held constant, then for each unit increase in birth rate, the average life expectancy will be **decreased** between 0.35 and 0.84 years

# $R^2$ and $R^2_{\text{adj}}$

❖ In SLR, the coefficient of determination,  $R^2$ , measures how well the straight- line fit to the data. Smaller  $R^2$  means a poor fitting model.

In MLR, we have  $R^2$  and  $R^2_{\text{adj}}$

**Definition 1** The **multiple coefficient of determination**,  $R^2$ , is defined as

$$R^2 = 1 - \frac{\text{SSE}}{\text{SS}_{yy}} \quad 0 \leq R^2 \leq 1$$

where  $\text{SSE} = \sum (y_i - \hat{y}_i)^2$ ,  $\text{SS}_{yy} = \sum (y_i - \bar{y})^2$ , and  $\hat{y}_i$  is the predicted value of  $y_i$  for the multiple regression model.

❖  $R^2$  represents the fraction of the sample variation (in the response variable) that is explained by the model.

❖  $R^2$  increases when we increase the number of parameters  $\rightarrow$  this leads to overfitting

e.g.  $R^2 = 1$  when we fit number of parameters  $k = \text{sample size } n$

# $R^2$ and $R^2_{\text{adj}}$

**Definition 2** The **adjusted multiple coefficient of determination** is given by

$$\begin{aligned} R_a^2 &= 1 - \left[ \frac{(n-1)}{n-(k+1)} \right] \left( \frac{\text{SSE}}{\text{SS}_{yy}} \right) \\ &= 1 - \left[ \frac{(n-1)}{n-(k+1)} \right] (1 - R^2) \end{aligned}$$

*Note:*  $R_a^2 \leq R^2$  and, for poor-fitting models  $R_a^2$  may be negative.

- ❖  $R^2_{\text{adj}}$  takes into account both sample size  $n$  and number of parameters  $k$ .
- ❖  $R^2_{\text{adj}}$  is always smaller than  $R^2$
- ❖ In MLR,  $R^2_{\text{adj}}$  should be used to measure model's adequacy (after testing the utility of the model using the F-test).

# $R^2$ and $R^2_{\text{adj}}$



$$\widehat{\text{Life expectancy}} = 76.24 - 0.0003\text{Population} + 0.13\text{Health} + 0.11\text{Internet} - 0.59\text{BirthRate}$$

Multiple R-squared: 0.720, Adjusted R-squared: 0.694

$$\widehat{\text{Life Expectancy}} = 77.69 + 0.12\text{Internet} - 0.59\text{BirthRate}$$

Multiple R-squared: 0.717, Adjusted R-squared: 0.705

$$R^2_a = 1 - \frac{(1 - R^2)(n - 1)}{n - (k + 1)} = 1 - \frac{(1 - 0.717)(49 - 1)}{49 - (2 + 1)} = 1 - 0.295 = 0.705$$

# LifeExp

## Exercise: Summarise these results

```
mod2 <- lm(LifeExp ~ Internet + BirthRate, data =countries.df)
summary(mod2)
#####
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	77.6921	4.0881	19.00	< 2e-16
Internet	0.1160	0.0424	2.74	0.0088
BirthRate	-0.5949	0.1198	-4.96	9.9e-06

Residual standard error: 5.61 on 46 degrees of freedom

Multiple R-squared: 0.717, Adjusted R-squared: 0.705

F-statistic: 58.4 on 2 and 46 DF, p-value: 2.37e-13

# LifeExp

## Exercise: Summarise these results

```
mod2 <- lm(LifeExp ~ Internet + BirthRate, data =countries.df)
```

```
summary(mod2) Life Expectancy = 77.69 + 0.12Internet - 0.59BirthRate
```

```
#####
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	77.6921	4.0881	19.00	< 2e-16
Internet	0.1160	0.0424	2.74	0.0088
BirthRate	-0.5949	0.1198	-4.96	9.9e-06

Both internet and Birthrate are useful predictors of life expectancy when the other is included in the model.

Adjusted R<sup>2</sup>: approx. 70.5% of variability explained by the model.

5.61 on 46 degrees of freedom

17, Adjusted R-squared: **0.705**

F-statistic: 58.4 on 2 and 46 DF, p-value: 2.37e-13

Global F-test: at least one-predictor is useful

# LifeExp

## Exercise: Summarise these results

It was found that the percentage of people having internet and the birth rate (births per 1000) are useful in predicting a country's life expectancy, while the percentage of government expenditure on health care and the country's population are not having significant effects on the life expectancy.

The final model is:

$$E(\text{LifeExp}) = 77.69 + 0.12 * \text{Internet} - 0.59 * \text{BirthRate}$$

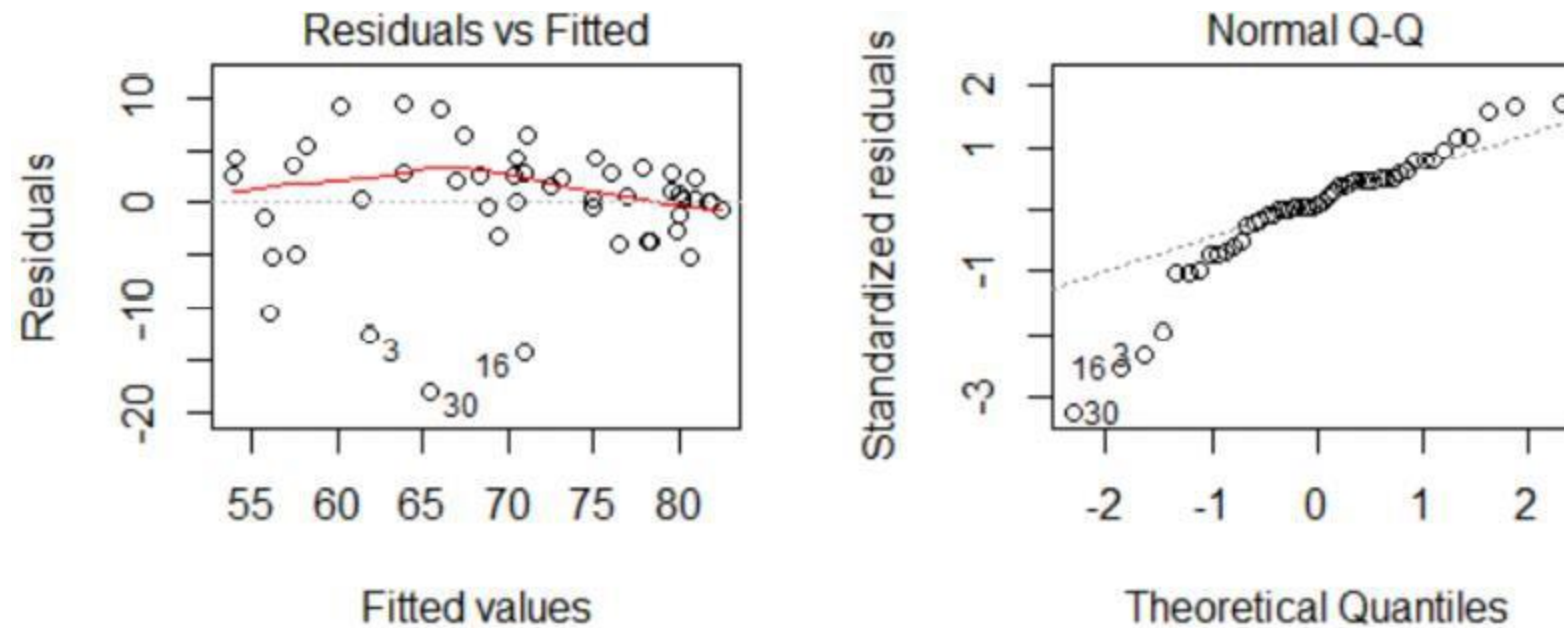
The average life expectancy will increase when more people having access to the internet, given that birth rate remains the same. However, for the same percentage of internet access, the average life expectancy decreases for higher birth rate.

The model using Internet and BirthRate explains about 70.5% of variability in life expectancy.

*Note: we still need to check model assumptions by assessing the residuals plots.*

# Warning

Still need to check model assumptions. This is covered in detail more detail in Topic 5 (Chapter 8).



# Prediction

IF the model assumptions were valid ...

Find a *confidence interval* and a *prediction interval* for life expectancy when

- BirthRate = 20.5, Internet = 39.2
- BirthRate = 45.6, Internet = 39.2



# predict () in R

```
#### add new data frame
Internet <- c(39.2, 39.2)
BirthRate <- c(20.5, 45.6)
new.df <- data.frame(Internet,BirthRate)
```

```
# 95 % CI for mean life expectancy
> predict(mod2,newdata=new.df, interval="confidence")
```

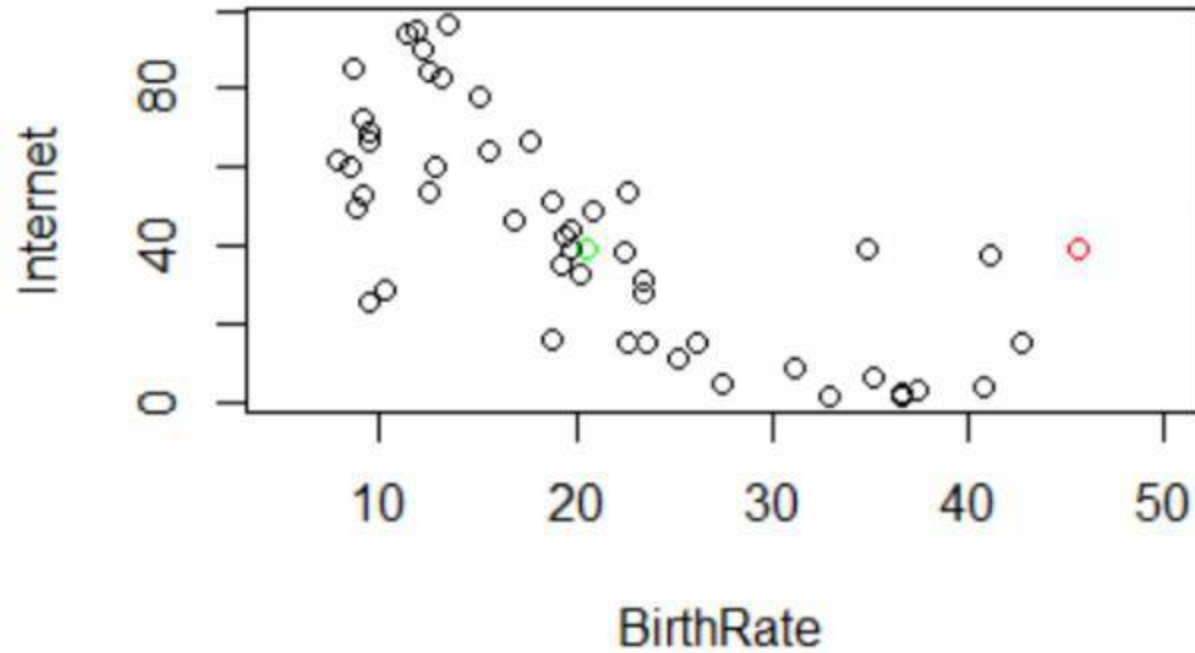
	fit	lwr	upr
1	70.0	68.4	71.7
2	55.1	49.1	61.1

```
# 95% prediction intervals for life expectancy
> predict(mod2, newdata = new.df, interval="prediction")
```

	fit	lwr	upr
1	70.0	58.6	81.5
2	55.1	42.3	67.9

- ✓ **Prediction intervals always wider/ more variability compared to the confidence intervals of the mean.**

# Exploratory Plot



Predictions for

- BirthRate = 20.5, Internet = 39.2
- BirthRate = 45.6, Internet = 39.2

# Next lecture

## Lecture 1

- ❖ Intro to MLR
- ❖ Fitting the model, testing the overall utility of a model
- ❖ Interpreting regression coefficients

## Lecture 2

- ❖ Inferences about the individual  $\beta_i$
- ❖ Multiple Coefficients of determination,  $R^2$  and  $R^2_{\text{adj}}$
- ❖ Using the model for estimation and prediction

## Lecture 3

- ❖ An interaction model with quantitative predictors

## Lecture 4

- ❖ Models with qualitative predictors

NB: Sections 4.11, 4.13 and 4.14 of the text will **not** be covered