R Notebook

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

knitr::opts\_chunk$set(echo = TRUE)

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

### Dependency

##library(tidyverse)  
library(readxl)

#load necessary libraries  
library(ggplot2)  
library(ggpubr)

# Problem 2.152

## A) Solution:

### Read xls file

file\_name = "ex02-152meis.xls"  
data = read\_xls(file\_name)

### Preview tibble

data

## # A tibble: 21 x 4  
## Country Sales Production DwellPermit  
## <chr> <dbl> <dbl> <dbl>  
## 1 Australia 137 109 116  
## 2 Belgium 105 112 125  
## 3 Canada 122 101 224  
## 4 Czech Republic 134 162 178  
## 5 Denmark 126 109 121  
## 6 Finland 136 125 105  
## 7 France 121 104 145  
## 8 Germany 100 119 54  
## 9 Greece 136 102 117  
## 10 Hungary 140 155 109  
## # ... with 11 more rows

summary(data)

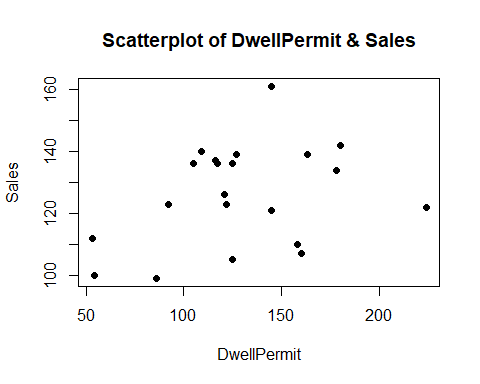
## Country Sales Production DwellPermit   
## Length:21 Min. : 99.0 Min. : 94.0 Min. : 53.0   
## Class :character 1st Qu.:112.0 1st Qu.:108.0 1st Qu.:109.0   
## Mode :character Median :126.0 Median :112.0 Median :125.0   
## Mean :126.1 Mean :120.4 Mean :128.8   
## 3rd Qu.:137.0 3rd Qu.:125.0 3rd Qu.:158.0   
## Max. :161.0 Max. :162.0 Max. :224.0

fit<- lm(data$Sales~ data$DwellPermit)  
summary(fit)

##   
## Call:  
## lm(formula = data$Sales ~ data$DwellPermit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.036 -16.122 1.556 11.397 32.859   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 109.8204 11.5582 9.501 1.19e-08 \*\*\*  
## data$DwellPermit 0.1263 0.0857 1.474 0.157   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.7 on 19 degrees of freedom  
## Multiple R-squared: 0.1027, Adjusted R-squared: 0.05543   
## F-statistic: 2.174 on 1 and 19 DF, p-value: 0.1568

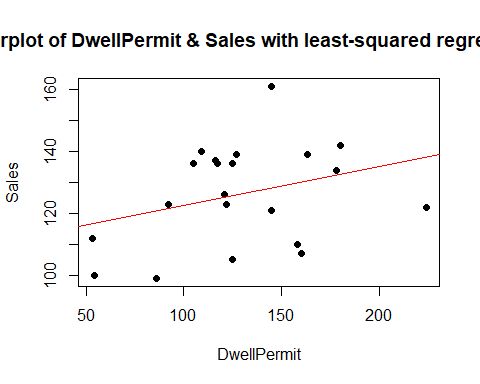
b = unname(fit$coefficients[1])  
a = unname(fit$coefficients[2])

# Scatterplot of DwellPermit & Sales  
plot(data$DwellPermit, data$Sales, main="Scatterplot of DwellPermit & Sales",  
 xlab="DwellPermit", ylab="Sales", pch=19)

 *There seems be a week positive association and one outlier near x=145 (Luxembourg)*

## B) Solution:

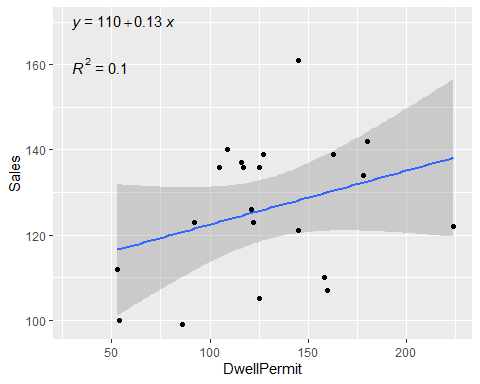
# Scatterplot of DwellPermit & Sales  
  
plot(data$DwellPermit, data$Sales, main="Scatterplot of DwellPermit & Sales with least-squared regression line",  
 xlab="DwellPermit", ylab="Sales", pch=19)  
# Add fit lines  
abline(lm(Sales~DwellPermit, data = data), col="red") # regression line (y~x)



### Using ggplot2

#create plot with regression line, regression equation, and R-squared  
ggplot(data=data, aes(x=DwellPermit, y=Sales)) +  
 geom\_smooth(method="lm") +  
 geom\_point() +  
 stat\_regline\_equation(label.x=30, label.y=170) +  
 stat\_cor(aes(label=..rr.label..), label.x=30, label.y=160)

## `geom\_smooth()` using formula 'y ~ x'



## C) Solution:

x = 160  
sales = a\*x + b  
print(paste("Predicted value of sales for a country that has an index of 160 for dwelling permits is", sales))

## [1] "Predicted value of sales for a country that has an index of 160 for dwelling permits is 130.036089058185"

## D) Solution:

id = which(data$Country == "Netherlands" & data$DwellPermit==160)

#List of residual for Netherlands (15th on the list)  
print(paste("Residual for Netherlands is",resid(fit)[id]))

## [1] "Residual for Netherlands is -23.0360890581847"

## E) Solution:

data$xbar\_diff = data$DwellPermit- mean(data$DwellPermit)  
data$ybar\_diff = data$Sales- mean(data$Sales)  
data$xbar\_diffs = (data$DwellPermit- mean(data$DwellPermit))^2  
data$ybar\_diffs = (data$Sales- mean(data$Sales))^2  
data$xy = data$xbar\_diff \* data$ybar\_diff  
data

## # A tibble: 21 x 9  
## Country Sales Production DwellPermit xbar\_diff ybar\_diff xbar\_diffs  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Australia 137 109 116 -12.8 10.9 164.   
## 2 Belgium 105 112 125 -3.81 -21.1 14.5  
## 3 Canada 122 101 224 95.2 -4.10 9061.   
## 4 Czech Republic 134 162 178 49.2 7.90 2420.   
## 5 Denmark 126 109 121 -7.81 -0.0952 61.0  
## 6 Finland 136 125 105 -23.8 9.90 567.   
## 7 France 121 104 145 16.2 -5.10 262.   
## 8 Germany 100 119 54 -74.8 -26.1 5596.   
## 9 Greece 136 102 117 -11.8 9.90 139.   
## 10 Hungary 140 155 109 -19.8 13.9 392.   
## # ... with 11 more rows, and 2 more variables: ybar\_diffs <dbl>, xy <dbl>

r = (sum(data$xy))/sqrt(sum(data$ybar\_diffs) \* sum(data$xbar\_diffs))  
r^2

## [1] 0.1026555

cor(data$DwellPermit, data$Sales)

## [1] 0.320399

### Alternate solution:

rsq <- cor(data$DwellPermit, data$Sales) ^ 2  
rsq

## [1] 0.1026555

*So 10.27% of the variance of sales can be explained by dwelling permits*

# Problem 2.153

## A) Solution:

### Read xls file

file\_name = "ex02-152meis.xls"  
data = read\_xls(file\_name)

### Preview tibble

data

## # A tibble: 21 x 4  
## Country Sales Production DwellPermit  
## <chr> <dbl> <dbl> <dbl>  
## 1 Australia 137 109 116  
## 2 Belgium 105 112 125  
## 3 Canada 122 101 224  
## 4 Czech Republic 134 162 178  
## 5 Denmark 126 109 121  
## 6 Finland 136 125 105  
## 7 France 121 104 145  
## 8 Germany 100 119 54  
## 9 Greece 136 102 117  
## 10 Hungary 140 155 109  
## # ... with 11 more rows

summary(data)

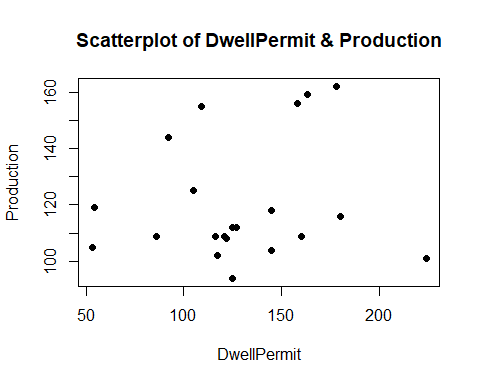
## Country Sales Production DwellPermit   
## Length:21 Min. : 99.0 Min. : 94.0 Min. : 53.0   
## Class :character 1st Qu.:112.0 1st Qu.:108.0 1st Qu.:109.0   
## Mode :character Median :126.0 Median :112.0 Median :125.0   
## Mean :126.1 Mean :120.4 Mean :128.8   
## 3rd Qu.:137.0 3rd Qu.:125.0 3rd Qu.:158.0   
## Max. :161.0 Max. :162.0 Max. :224.0

fit<- lm(data$Production~ data$DwellPermit)  
summary(fit)

##   
## Call:  
## lm(formula = data$Production ~ data$DwellPermit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.344 -11.883 -8.249 6.361 38.021   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 110.95813 15.88466 6.985 1.18e-06 \*\*\*  
## data$DwellPermit 0.07315 0.11778 0.621 0.542   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.57 on 19 degrees of freedom  
## Multiple R-squared: 0.0199, Adjusted R-squared: -0.03168   
## F-statistic: 0.3858 on 1 and 19 DF, p-value: 0.5419

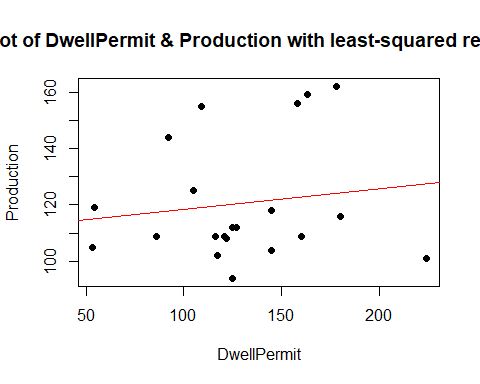
b = unname(fit$coefficients[1])  
a = unname(fit$coefficients[2])

# Scatterplot of DwellPermit & Production  
plot(data$DwellPermit, data$Production, main="Scatterplot of DwellPermit & Production",  
 xlab="DwellPermit", ylab="Production", pch=19)

 **There seems be a very week positive association and one outlier near x=224 (Canada)**

## B) Solution:

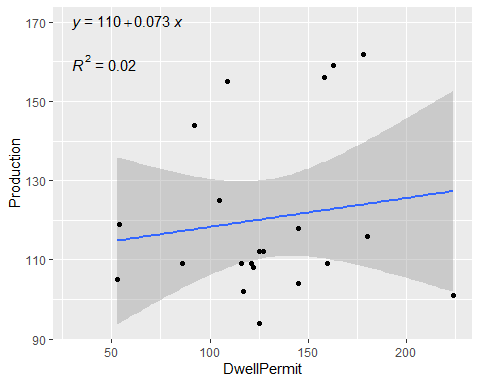
# Scatterplot of DwellPermit & Production  
  
plot(data$DwellPermit, data$Production, main="Scatterplot of DwellPermit & Production with least-squared regression line",  
 xlab="DwellPermit", ylab="Production", pch=19)  
# Add fit lines  
abline(lm(Production~DwellPermit, data = data), col="red") # regression line (y~x)



### Using ggplot2

#create plot with regression line, regression equation, and R-squared  
ggplot(data=data, aes(x=DwellPermit, y=Production)) +  
 geom\_smooth(method="lm") +  
 geom\_point() +  
 stat\_regline\_equation(label.x=30, label.y=170) +  
 stat\_cor(aes(label=..rr.label..), label.x=30, label.y=160)

## `geom\_smooth()` using formula 'y ~ x'



## C) Solution:

x = 160  
sales = a\*x + b  
print(paste("Predicted value of Production for a country that has an index of 160 for dwelling permits is", sales))

## [1] "Predicted value of Production for a country that has an index of 160 for dwelling permits is 122.662632802455"

## D) Solution:

id = which(data$Country == "Netherlands" & data$DwellPermit==160)

#List of residual for Netherlands (15th on the list)  
print(paste("Residual for Netherlands is",resid(fit)[id]))

## [1] "Residual for Netherlands is -13.6626328024547"

## E) Solution:

cor(data$DwellPermit, data$Production)

## [1] 0.1410671

rsq <- cor(data$DwellPermit, data$Production) ^ 2  
rsq

## [1] 0.01989992

*So 1.99% of the variance of Production can be explained by dwelling permits. In earlier exercise we observed a higher variance explanations between sales and dwelling permit which was due to higher association among the variables.*

# Problem 2.154

## A) Solution:

### Read xls file

file\_name = "ex02-152meis.xls"  
data = read\_xls(file\_name)

### Preview tibble

data

## # A tibble: 21 x 4  
## Country Sales Production DwellPermit  
## <chr> <dbl> <dbl> <dbl>  
## 1 Australia 137 109 116  
## 2 Belgium 105 112 125  
## 3 Canada 122 101 224  
## 4 Czech Republic 134 162 178  
## 5 Denmark 126 109 121  
## 6 Finland 136 125 105  
## 7 France 121 104 145  
## 8 Germany 100 119 54  
## 9 Greece 136 102 117  
## 10 Hungary 140 155 109  
## # ... with 11 more rows

summary(data)

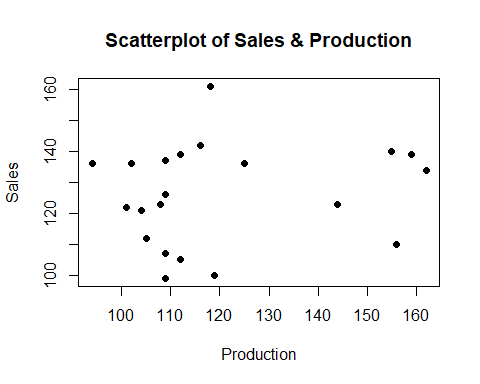
## Country Sales Production DwellPermit   
## Length:21 Min. : 99.0 Min. : 94.0 Min. : 53.0   
## Class :character 1st Qu.:112.0 1st Qu.:108.0 1st Qu.:109.0   
## Mode :character Median :126.0 Median :112.0 Median :125.0   
## Mean :126.1 Mean :120.4 Mean :128.8   
## 3rd Qu.:137.0 3rd Qu.:125.0 3rd Qu.:158.0   
## Max. :161.0 Max. :162.0 Max. :224.0

fit<- lm(data$Sales~ data$Production)  
summary(fit)

##   
## Call:  
## lm(formula = data$Sales ~ data$Production)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -25.936 -12.327 1.213 12.018 35.178   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 112.2568 21.0657 5.329 3.83e-05 \*\*\*  
## data$Production 0.1150 0.1725 0.667 0.513   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 16.38 on 19 degrees of freedom  
## Multiple R-squared: 0.02285, Adjusted R-squared: -0.02858   
## F-statistic: 0.4443 on 1 and 19 DF, p-value: 0.5131

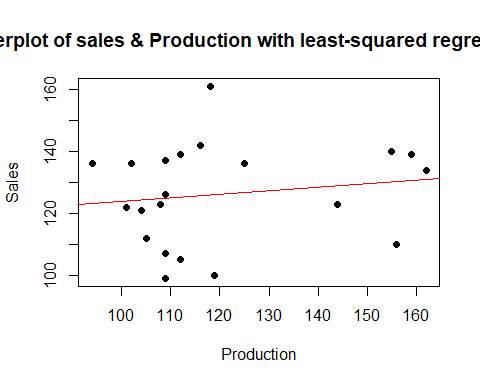
b = unname(fit$coefficients[1])  
a = unname(fit$coefficients[2])

# Scatterplot of Sales & Production  
plot(data$Production, data$Sales, main="Scatterplot of Sales & Production",  
 xlab="Production", ylab="Sales", pch=19)

 **From the plot we get no clear relationship between the two variables.There is an outlier in the data and that is related to y=161 and x=118 (Luxembourg)**

## B) Solution:

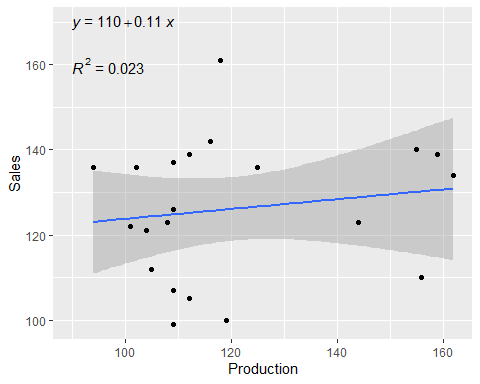
# Scatterplot of DwellPermit & Production  
  
plot(data$Production, data$Sales, main="Scatterplot of sales & Production with least-squared regression line",  
 xlab="Production", ylab="Sales", pch=19)  
# Add fit lines  
abline(lm(Sales~Production, data = data), col="red") # regression line (y~x)



### Using ggplot2

#create plot with regression line, regression equation, and R-squared  
ggplot(data=data, aes(x=Production, y=Sales)) +  
 geom\_smooth(method="lm") +  
 geom\_point() +  
 stat\_regline\_equation(label.x=90, label.y=170) +  
 stat\_cor(aes(label=..rr.label..), label.x=90, label.y=160)

## `geom\_smooth()` using formula 'y ~ x'



## C) Solution:

x = 125  
sales = a\*x + b  
print(paste("Predicted value of Production for a country that has an index of 125 for Production is", sales))

## [1] "Predicted value of Production for a country that has an index of 125 for Production is 126.626223071321"

## D) Solution:

id = which(data$Country == "Finland" & data$Production ==125)

#List of residual for Netherlands (15th on the list)  
print(paste("Residual for Finland is",resid(fit)[id]))

## [1] "Residual for Finland is 9.37377692867924"

## E) Solution:

cor(data$Production, data$Sales)

## [1] 0.1511681

rsq <- cor(data$Production, data$Sales) ^ 2  
rsq

## [1] 0.02285178

*So 2.29% of the variance of sales can be explained by productions. So Sales has more association with dwelling permits than productions.*

# Problem 2.157

## A) Solution:

### Read xls file

file\_name = "ex02-157fos.xls"  
data = read\_xls(file\_name)

### Preview tibble

data

## # A tibble: 5 x 8  
## Field Canada France Germany Italy Japan UK US  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 SsBL 64 153 66 125 250 152 878  
## 2 SME 35 111 66 80 136 128 355  
## 3 AH 27 74 33 42 123 105 397  
## 4 Ed 20 45 18 16 39 14 167  
## 5 Other 30 289 35 58 97 76 272

data <- rbind(data, c("Total", colSums(data[,2:8])))  
data

## # A tibble: 6 x 8  
## Field Canada France Germany Italy Japan UK US   
## <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>  
## 1 SsBL 64 153 66 125 250 152 878   
## 2 SME 35 111 66 80 136 128 355   
## 3 AH 27 74 33 42 123 105 397   
## 4 Ed 20 45 18 16 39 14 167   
## 5 Other 30 289 35 58 97 76 272   
## 6 Total 176 672 218 321 645 475 2069

data = transform(data, Field = as.character(Field), Canada = as.numeric(Canada),France = as.numeric(France),Germany = as.numeric(Germany),Italy = as.numeric(Italy),  
 Japan = as.numeric(Japan),UK = as.numeric(UK),US = as.numeric(US))

# Marginal totals  
data$total = rowSums(data[2:8])  
data

## Field Canada France Germany Italy Japan UK US total  
## 1 SsBL 64 153 66 125 250 152 878 1688  
## 2 SME 35 111 66 80 136 128 355 911  
## 3 AH 27 74 33 42 123 105 397 801  
## 4 Ed 20 45 18 16 39 14 167 319  
## 5 Other 30 289 35 58 97 76 272 857  
## 6 Total 176 672 218 321 645 475 2069 4576

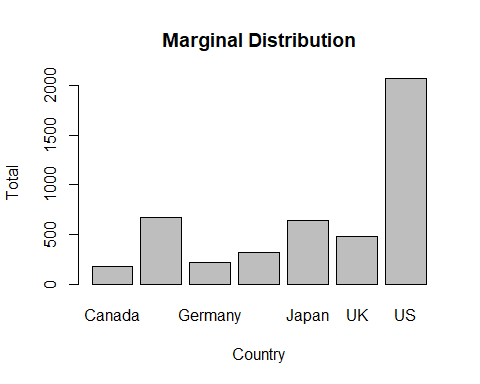
## B) Solution:

library(data.table)

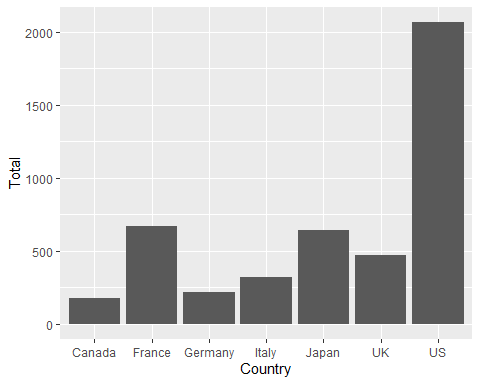
# Marginal distribution of country  
df = as.data.frame(transpose(tail(data[,2:8], n=1)))  
colnames(df) <- c("Total")  
df$country <- colnames(data[,2:8])  
df

## Total country  
## 1 176 Canada  
## 2 672 France  
## 3 218 Germany  
## 4 321 Italy  
## 5 645 Japan  
## 6 475 UK  
## 7 2069 US

barplot(df$Total,names.arg = df$country, main="Marginal Distribution", xlab="Country", ylab = "Total")



ggplot(df, aes(x=country, y=Total)) + geom\_bar(stat="identity") +   
 labs(x="Country", y="Total")

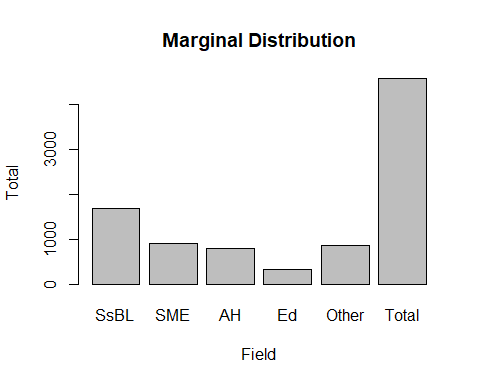


## C) Solution:

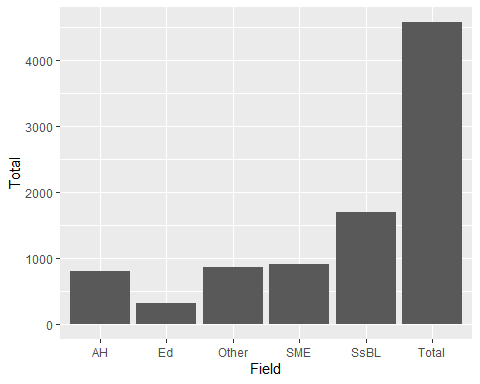
df2 = data[ , ncol(data), drop = FALSE]  
df2$field = data$Field  
df2

## total field  
## 1 1688 SsBL  
## 2 911 SME  
## 3 801 AH  
## 4 319 Ed  
## 5 857 Other  
## 6 4576 Total

barplot(df2$total,names.arg = df2$field, main="Marginal Distribution", xlab="Field", ylab = "Total")



ggplot(df2, aes(x=field, y=total)) + geom\_bar(stat="identity") +   
 labs(x="Field", y="Total")



# Problem 2.162

### Read xls file

file\_name = "ex02-162raises.xls"  
data = read\_xls(file\_name)

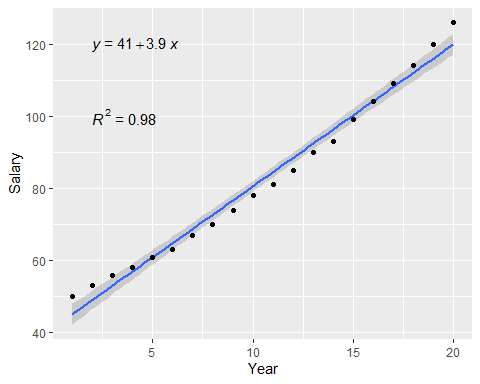
### Preview tibble

data

## # A tibble: 20 x 2  
## Year Salary  
## <dbl> <dbl>  
## 1 1 50  
## 2 2 53  
## 3 3 56  
## 4 4 58  
## 5 5 61  
## 6 6 63  
## 7 7 67  
## 8 8 70  
## 9 9 74  
## 10 10 78  
## 11 11 81  
## 12 12 85  
## 13 13 90  
## 14 14 93  
## 15 15 99  
## 16 16 104  
## 17 17 109  
## 18 18 114  
## 19 19 120  
## 20 20 126

ggplot(data=data, aes(x=Year, y=Salary)) +  
 geom\_smooth(method="lm") +  
 geom\_point() +  
 stat\_regline\_equation(label.x=2, label.y=120) +  
 stat\_cor(aes(label=..rr.label..), label.x=2, label.y=100)

## `geom\_smooth()` using formula 'y ~ x'



## A) Solution:

*From the graph we can see that there is a strong positive association between salary and year.*

## B) Solution:

rsq <- round(cor(data$Year, data$Salary) ^ 2,4)  
rsq

## [1] 0.9832

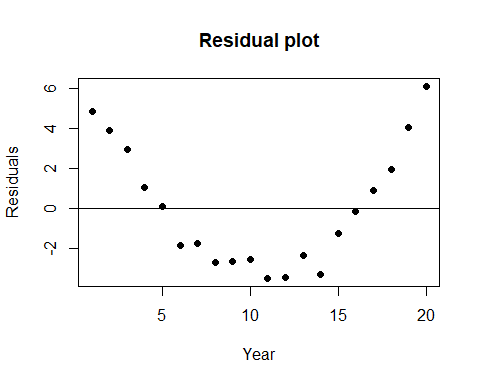
*As the R2 value is 0.9832, 98.32% of variation in salary is explained by the year*

# Problem 2.162

## A) Solution:

data.lm = lm(Salary ~ Year, data=data)   
data.res = resid(data.lm)

plot(data$Year, data.res, ylab="Residuals", xlab="Year", main="Residual plot", pch=19)   
abline(0, 0)



mean(data.res)

## [1] 1.000502e-16

*In the graph above, we can see that the residual values are not randomly distributed around the horizon (y=0)(which should have been the case had the data been perfectly linear) but follow a pattern. This happened due to inconsistent percentage of raise (<=5%). But the mean of deviations is pretty close to zero.*

## B) Solution:

*This residual plot indicates that the independent variable (year) does not capture the entire deterministic component. Unfortunately, some of the explanatory data has leaked over to the supposedly random error. The model obviously has room for improvement.*

# Problem 2.166

### Read xls file

file\_name = "ex02-166faculty.xls"  
data = read\_xls(file\_name)

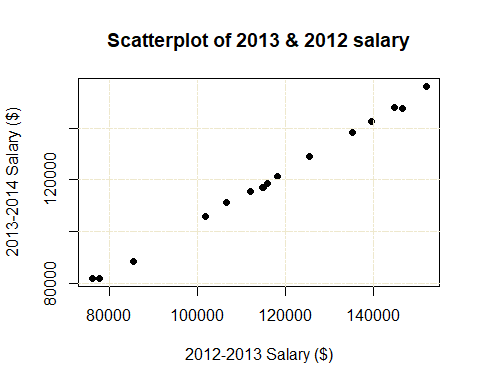
### Preview tibble

data

## # A tibble: 16 x 2  
## Salary2012 Salary2013  
## <dbl> <dbl>  
## 1 146600 147700  
## 2 115800 118600  
## 3 112000 115500  
## 4 101700 105800  
## 5 115000 117180  
## 6 114790 117240  
## 7 106500 111100  
## 8 152000 156080  
## 9 139650 142350  
## 10 135160 138485  
## 11 77792 82072  
## 12 76000 82000  
## 13 85500 88700  
## 14 144850 147830  
## 15 125506 128906  
## 16 118100 121200

## A) Solution:

# Scatterplot of 2013 & 2021 salary  
plot(data$Salary2012, data$Salary2013, main="Scatterplot of 2013 & 2012 salary",  
 xlab="2012-2013 Salary ($)", ylab="2013-2014 Salary ($)", pch=19)  
grid (NULL,NULL, lty = 6, col = "cornsilk2")



## B) Solution:

*The relationship is linear in nature. It is showing an increasing trend. As the 2012-2013 Salary increases, so does the 2013-2014 Salary. There seems to be a very strong positive association between the two variables.*

## B) Solution:

rsq <- cor(data$Salary2012, data$Salary2013) ^ 2  
rsq

## [1] 0.9984137

*So 99.84% of the variance of 2013-2014 Salary can be explained by 2012-2013 Salary.*

# Problem 2.177

### Read xls file

file\_name = "ex02-177csize.xls"  
data = read\_xls(file\_name)

### Preview tibble

data

## # A tibble: 16 x 4  
## Department Year Size Count  
## <chr> <chr> <chr> <dbl>  
## 1 A First Large 2  
## 2 A First Small 0  
## 3 A Second Large 9  
## 4 A Second Small 1  
## 5 A Third Large 5  
## 6 A Third Small 15  
## 7 A Fourth Large 4  
## 8 A Fourth Small 16  
## 9 B First Large 18  
## 10 B First Small 2  
## 11 B Second Large 40  
## 12 B Second Small 10  
## 13 B Third Large 4  
## 14 B Third Small 16  
## 15 B Fourth Large 2  
## 16 B Fourth Small 14

colnames(data)

## [1] "Department" "Year" "Size" "Count"

# Total number of classes in department A  
a\_tot = sum(data[data$Department=="A",]["Count"])  
print(paste("Total number of classes in department A =",a\_tot))

## [1] "Total number of classes in department A = 52"

# Total number of small classes in department A  
a\_s\_tot = sum(data[data$Size=='Small' & data$Department=="A",]["Count"])  
print(paste("Total number of small classes in department A =",a\_s\_tot))

## [1] "Total number of small classes in department A = 32"

#The percentage of small classes in Department A  
print(paste("The percentage of small classes in department A is", round(a\_s\_tot/a\_tot\*100,2), "%"))

## [1] "The percentage of small classes in department A is 61.54 %"

# Total number of classes in department B  
b\_tot = sum(data[data$Department=="B",]["Count"])  
print(paste("Total number of classes in department B =",b\_tot))

## [1] "Total number of classes in department B = 106"

# Total number of small classes in department B  
b\_s\_tot = sum(data[data$Size=='Small' & data$Department=="B",]["Count"])  
print(paste("Total number of small classes in department B =",b\_s\_tot))

## [1] "Total number of small classes in department B = 42"

#The percentage of small classes in Department B  
print(paste("The percentage of small classes in department B is", round(b\_s\_tot/b\_tot\*100,2), "%"))

## [1] "The percentage of small classes in department B is 39.62 %"

*It seems that department A has a higher % of small classes.*

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':  
##   
## between, first, last

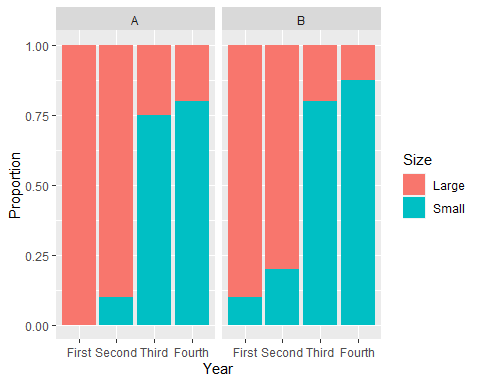
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data = data %>% mutate(row = row\_number())  
data

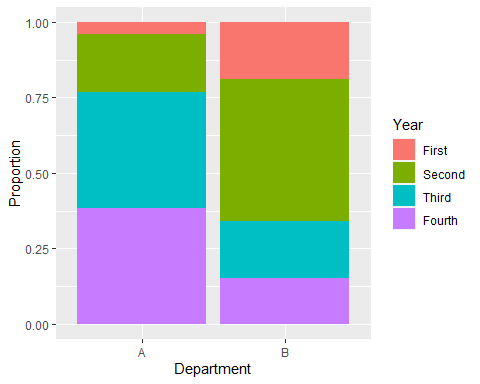
## # A tibble: 16 x 5  
## Department Year Size Count row  
## <chr> <chr> <chr> <dbl> <int>  
## 1 A First Large 2 1  
## 2 A First Small 0 2  
## 3 A Second Large 9 3  
## 4 A Second Small 1 4  
## 5 A Third Large 5 5  
## 6 A Third Small 15 6  
## 7 A Fourth Large 4 7  
## 8 A Fourth Small 16 8  
## 9 B First Large 18 9  
## 10 B First Small 2 10  
## 11 B Second Large 40 11  
## 12 B Second Small 10 12  
## 13 B Third Large 4 13  
## 14 B Third Small 16 14  
## 15 B Fourth Large 2 15  
## 16 B Fourth Small 14 16

ggplot(data=data, aes(x=reorder(Year,row), y=Count, fill=Size)) +  
geom\_bar(stat="identity", position = position\_fill()) + xlab("Year") +  
 ylab("Proportion") + facet\_wrap(~Department)



*In the y-axis we have the years, in x-axis are the proportions, and bars are stacked with the percentage of large and small classes in each year. Plot on the left is for Department A, and plot on the right is for Department B (see letters above each plot). In fact, the plot shows what Department A pointed out, Third and Fourth Years have a large proportion of small classes. If Department A has a higher proportion of smaller classes, we may think that it has a higher proportion of third and Fourth year classes. We can change the plot a bit to check for this assumption. Plotting the percentages of classes per year in each department.*

ggplot(data=data, aes(x=Department, y=Count, fill=reorder(Year,row))) +  
geom\_bar(stat="identity", position = position\_fill()) +  
ylab("Proportion") + scale\_fill\_discrete(name = "Year")



*So, Department A has large proportion of third and fourth years classes.*

*Finally, we can test if there is an statistical association between years and size of the class, controlling by department. In other words, if the counts of large and small classes are statistically different between years, inside each department. This is done by analyzing this data as a contingency table. The test that evaluates association between two factors (Size and Year in this case), controlling for a third factor (department in this case) is the Mantel-Haenzel test. As in a chi-squared statistics, this evaluates deviations between observed and expected frequencies.*

*The hypothesis tested are:*

*H0 (null): There is not an association between year and size*

*H1 (alternative): There is an association between year and size*

*As in classical hypothesis testing, a p-value for the value of the statistic obtained in computed. The p-value is the probability of finding a value of the statistic equal or higher than the one obtained. The decision about which hypothesis the data supports is made based on a cutoff (alpha, level of significance, usually = 0.05). If p-value is smaller than alpha we reject the null hypothesis, if p-value is higher than alpha we do not reject the null hypothesis.*

CT\_data <- xtabs(data$Count~data$Size+data$Year+data$Department)  
  
mantelhaen.test(CT\_data)

##   
## Cochran-Mantel-Haenszel test  
##   
## data: CT\_data  
## Cochran-Mantel-Haenszel M^2 = 62.344, df = 3, p-value = 1.855e-13

*So, the p-value is below 0.05. We can reject the null hypothesis. So there are statistical differences in the proportions of small and large classes between years within each department. As we saw in the first figure, larger counts are associated with years third and fourth, and these differences are significant, and the differences of small classes proportion between two departments is due to the fact that department A has higher proportions of third and fourth classes (as shown in second figure)*