STAT2402: Notes on Week 2 Computer Laboratory

In this lab we will first examine the data in the first week's lectures, specifically:

1. some exercises from the `R` lecture notes; and
2. fitting a linear regression model.

Getting Started:

Log in at one of the PCs and start up the software package `R`; either directly or via RStudio. If you have problems either logging in or starting `R` ask for help.

Recall: when typing in `R` commands you can use the arrow keys to speed things up. The 'up' arrow gives you the previous command that you typed. The usual prompt sign for `R` is `>`. If you get a `+` prompt sign instead, it means that `R` is awaiting the completion of the previous command that you typed in. This can happen because you have forgotten to close parentheses, for instance. Just type in the remainder of the command. **Note also that `R` is case sensitive.**

Use scripting an save your code.

Exercise 1: From `R` lecture notes An engineer is designing a battery for use in a device that will be subjected to some extreme variations in temperature. He has three possible choices for the plate material. For testing purposes he selects three temperatures. Four batteries are tested at each combination of plate material and temperature and the tests are run in random order. The battery life (hours) under each set of conditions is given in Table 1.

	Temperature (°C)					
Material	-10		20		55	
1	130	155	34	40	20	70
	74	180	80	75	82	58
2	150	188	136	122	25	70
	159	126	106	115	58	45
3	138	110	174	120	96	104
	168	160	150	139	82	60

Table 1: Life (in hours) data for the battery design example.

1. Examine the data and report your observations.

   Solution
   When examining two-dimensional tables, we look for patterns across the columns and down the rows. Here we see that across the table (as temperature increases) the lifetime decreases, and down the table (as material type changes) the lifetime increases. The one exception is that lifetimes for Material 2 at Temperature -10 is higher than that for Material 3. Overall, Material 3 seems to be the best performing at all temperatures.

2. Write the `R` code to enter this data into `R`.

   Solution

```
Material <- factor(rep(c(1, 2, 3), each = 12, length = 36))
Temperature <- factor(rep(c(-10, 20, 55), each = 4, length = 36))
Life <- c(130, 155, 74, 180, 34, 40, 80, 75, 20, 70, 82, 58, 150,
    188, 159, 126, 136, 122, 106, 115, 25, 70, 58, 45, 138, 110, 168,
    160, 174, 120, 150, 139, 96, 104, 82, 60)
Battery <- data.frame(Material, Temperature, Life)
```

Note that **both** Material and Temperature are factors (categorical).

3. Find the summary statistics for this data. First think about what sort of statistics you should be interested in.
   Solution
   We want the mean lifetime at each temperature for each material type.

```
with(Battery, tapply(Life, list(Material, Temperature), mean))

##      -10     20   55
## 1 134.75  57.25 57.5
## 2 155.75 119.75 49.5
## 3 144.00 145.75 85.5
```

   The table of means confirms our earlier observation. Material 3 is the best performing overall at every temperature.

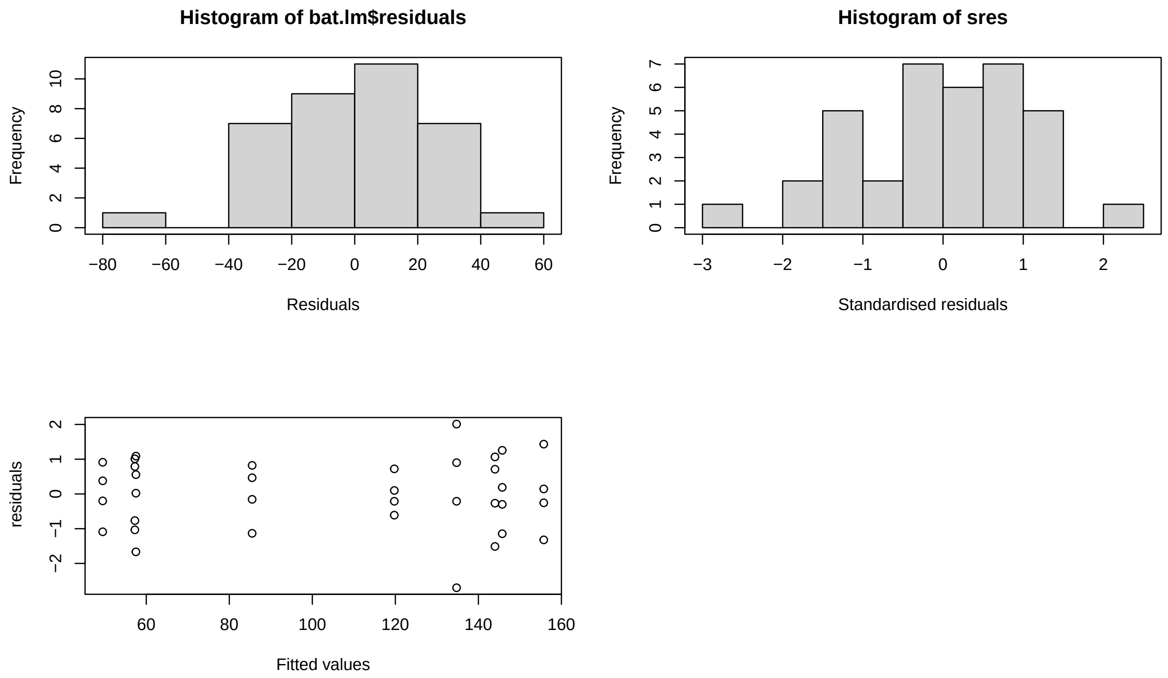
4. Now fit a linear model to the battery lifetimes. Investigate any interaction terms.
   Solution

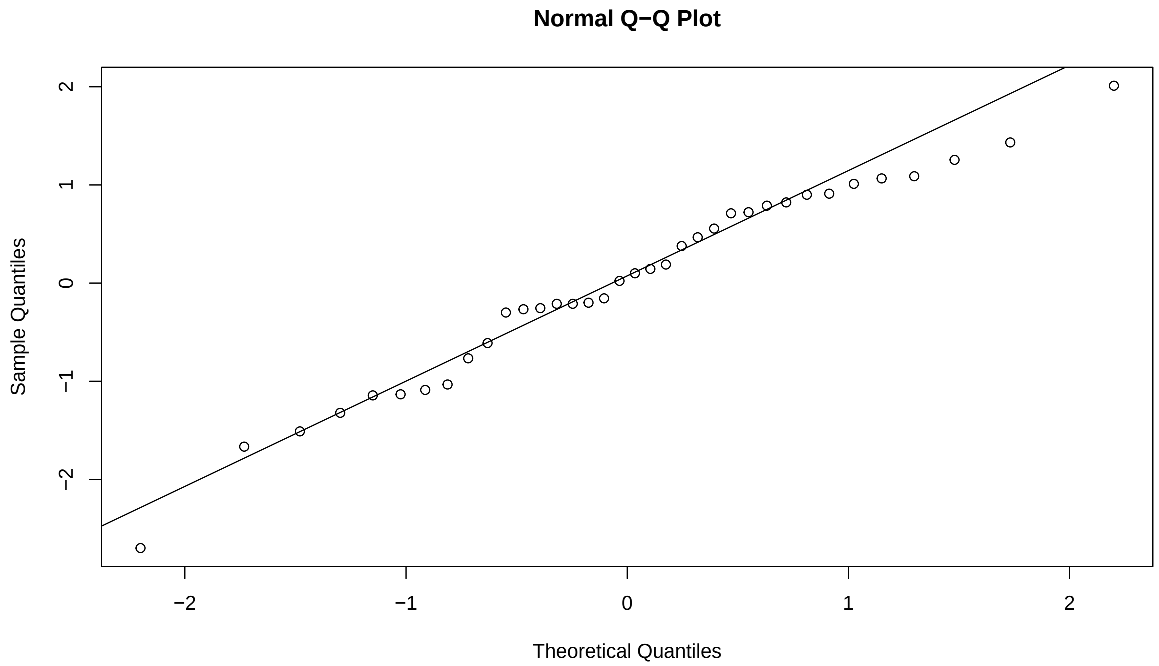
```
bat.lm <- lm(Life ~ Material * Temperature, data = Battery)
summary(bat.lm)

##
## Call:
## lm(formula = Life ~ Material * Temperature, data = Battery)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.750 -14.625   1.375  17.938  45.250
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                134.75      12.99  10.371 6.46e-11 ***
## Material2                   21.00      18.37   1.143 0.263107
## Material3                    9.25      18.37   0.503 0.618747
## Temperature20              -77.50      18.37  -4.218 0.000248 ***
## Temperature55              -77.25      18.37  -4.204 0.000257 ***
## Material2:Temperature20     41.50      25.98   1.597 0.121886
## Material3:Temperature20     79.25      25.98   3.050 0.005083 **
## Material2:Temperature55    -29.00      25.98  -1.116 0.274242
## Material3:Temperature55     18.75      25.98   0.722 0.476759
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.98 on 27 degrees of freedom
## Multiple R-squared:  0.7652,Adjusted R-squared:  0.6956
## F-statistic:    11 on 8 and 27 DF,  p-value: 9.426e-07
```

5. Select the best model based on your analysis.
   Solution
   The best model includes interaction between Temperature and Material.

6. Perform appropriate model diagnostics.
   Solution

```
oldpar <- par(mfrow = c(2, 2))
hist(bat.lm$residuals, xlab = "Residuals")
box()
sres <- stdres(bat.lm)
hist(sres, xlab = "Standardised residuals")
box()
plot(sres ~ bat.lm$fitted.values, xlab = "Fitted values", ylab = "Standardised
     residuals")
par(oldpar)
qqnorm(sres)
qqline(sres)
```

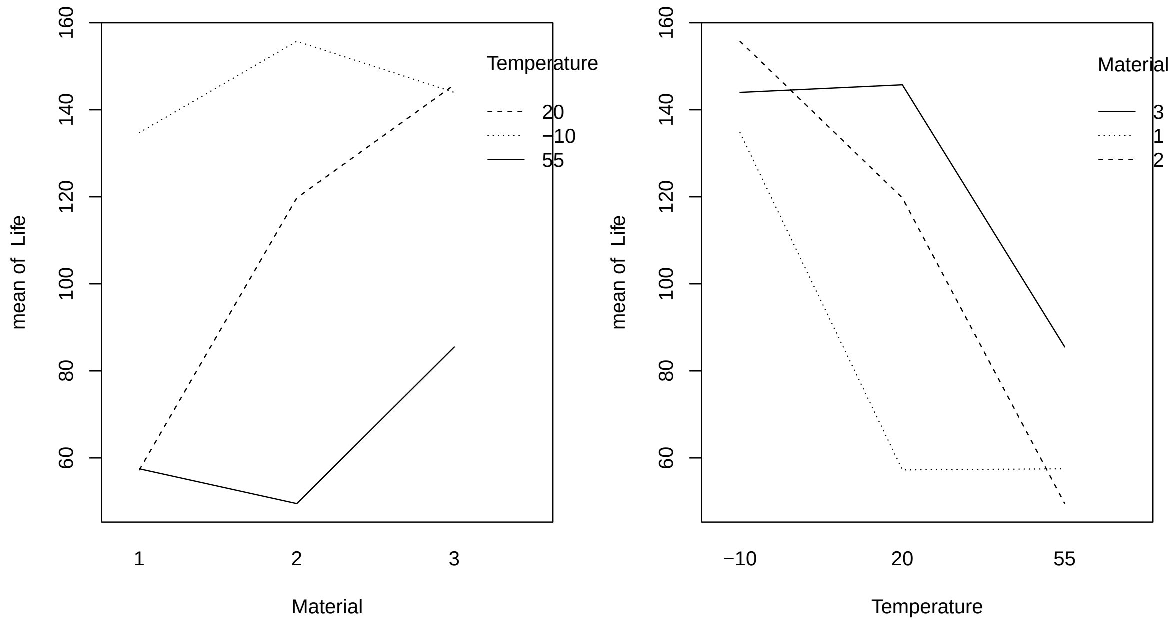




Based on the diagnostics, there is not evidence against the model assumptions (normality, homogeneous variance and linear model).

7. Produce an interaction plot for the mean battery lifetimes.
Solution

```
oldpar <- par(mfrow = c(1, 2))
with(Battery, interaction.plot(Material, Temperature, Life))
with(Battery, interaction.plot(Temperature, Material, Life))
par(oldpar)
```



8. Interpret your model.
Solution
Based on the linear model and interaction plot, the mean lifetime is higher for material 3 at a temperature of 20 degrees. At the other temperatures there is no significant difference, although the mean is again higher for material 3 at 55 degrees.

9. Which material would you recommend for the batteries? Justify your selection.
Solution
Based on the analyis we recommend Material 3 for the batteries.

Exercise 2: Data manipulation Consider the following grouped data on seatbelt use and the severity of injury in an accident.

	worn	not worn	unknown
fatal	35	6	15
severe	1142	48	328
minor	7969	76	764
unknown	11404	24	38570

1. Enter the data into `R` using the variables `Injury`, `SeatBelt` and `Frequency`.
Solution

```
Injury <- gl(n = 4, labels = c("fatal", "severe", "minor", "unknown"),
    k = 3, length = 12)
Seatbelt <- gl(n = 3, labels = c("worn", "not worn", "unknown"), k = 1,
    length = 12)
Frequency <- c(35, 6, 15, 1142, 48, 328, 7969, 76, 764, 11404, 24,
    38570)
Accident <- data.frame(Injury, Seatbelt, Frequency)
## Check data entry
xtabs(Frequency ~ Injury + Seatbelt)

##          Seatbelt
## Injury     worn not worn unknown
##   fatal      35        6      15
##   severe   1142       48     328
##   minor    7969       76     764
##   unknown 11404       24   38570
```

2. Now we want to create date that contains one record for each case. That is, we need to create 35 entries corresponding to a fatal injury where the seat belt was worn, 6 for when the seat belt was worn, and 15 for unknown. Similarly for the other levels of injuries. Write a short (2 lines!) of `R` code to achieve this, and test your code (for example, by producing a table from your new data).

   Solution
   One way to achieve this is the following.

```
ind <- rep(1:NROW(Accident), Accident$Frequency)
NewAccident <- data.frame(Accident$Injury[ind], Accident$Seatbelt[ind])
colnames(NewAccident) <- c("Injury", "Seatbelt")
```

   To see what this code does, let us look at a table for `ind`:

```
table(ind)

## ind
##     1     2     3     4     5     6     7     8     9    10
##    35     6    15  1142    48   328  7969    76   764 11404
##    11    12
##    24 38570

nrow(Accident)

## [1] 12

length(ind)

## [1] 60381

sum(Frequency)

## [1] 60381
```

   The variable `ind` has the length equal to the number of cases, and contains the row numbers for the dataframe `Accident`. The dataframe `NewAccident` is formed by taking the appropriate `Injury` type and `Seatbelt` status. We can check that this is the same dataset in a different format by looking at a table.

```
table(NewAccident$Injury, NewAccident$Seatbelt)

##
##             worn not worn unknown
##   fatal       35        6      15
```

```
##   severe   1142       48     328
##   minor    7969       76     764
##   unknown 11404       24   38570
```

Another more direct solution is given below.

```
NInjury <- rep(Accident$Injury, Accident$Frequency)
NSeatbelt <- rep(Accident$Seatbelt, Accident$Frequency)
NAccident <- data.frame(NInjury, NSeatbelt)
table(NAccident$NInjury, NAccident$NSeatbelt)

##
##           worn not worn unknown
##   fatal      35        6      15
##   severe   1142       48     328
##   minor    7969       76     764
##   unknown 11404       24   38570
```

Exercise 3: Fish data

1. The folder Data in the Computer Labs folder contains the data set `fish.txt`. Download the file and read the data into R. The variables are:

   - Code: fish species code
   - Weight: weight of the fish in grammes
   - Length1: length from the nose to the beginning of the tail (cm)
   - Length2: length from nose to notch of tail (cm)
   - Length3: length from nose to the end of the tail (cm)
   - Height: maximum height as a percentage of Length3
   - Width: maximum width as a percentage of Length3

   (a) Summarise the data and check for any data errors.
   Solution

```
fish <- read.table("../Data/Fish.txt", header = T)
summary(fish)

##       Code           Weight          Length1
##  Min.   :1.000   Min.   :   0.0   Min.   : 7.50
##  1st Qu.:2.250   1st Qu.: 120.0   1st Qu.:19.02
##  Median :5.000   Median : 272.5   Median :25.10
##  Mean   :4.519   Mean   : 398.7   Mean   :26.23
##  3rd Qu.:7.000   3rd Qu.: 650.0   3rd Qu.:32.70
##  Max.   :7.000   Max.   :1650.0   Max.   :59.00
##     Length2         Length3          Height          Width
##  Min.   : 8.40   Min.   : 8.80   Min.   :14.50   Min.   : 8.70
##  1st Qu.:21.00   1st Qu.:23.12   1st Qu.:24.23   1st Qu.:13.40
##  Median :27.15   Median :29.35   Median :27.00   Median :14.60
##  Mean   :28.39   Mean   :31.19   Mean   :28.26   Mean   :14.12
##  3rd Qu.:35.75   3rd Qu.:39.67   3rd Qu.:37.70   3rd Qu.:15.30
##  Max.   :63.40   Max.   :68.00   Max.   :44.50   Max.   :20.90
```

   Notice that the minimum weight is 0.

   (b) You will note a weight of 0. Determine which data record this corresponds to and omit it. Use the commands `which` and `fish1 <- fish[-x,]`, where `x` corresponds to the number of the record in error. Check that the record with the error has been removed.
   Solution

```r
x <- which(fish$Weight == 0)
fish1 <- fish[-x, ]
summary(fish1)
```

```
##       Code           Weight          Length1
##  Min.   :1.000   Min.   :   5.9   Min.   : 7.50
##  1st Qu.:2.000   1st Qu.: 120.0   1st Qu.:19.10
##  Median :5.000   Median : 273.0   Median :25.20
##  Mean   :4.529   Mean   : 401.2   Mean   :26.27
##  3rd Qu.:7.000   3rd Qu.: 650.0   3rd Qu.:32.70
##  Max.   :7.000   Max.   :1650.0   Max.   :59.00
##     Length2         Length3          Height          Width
##  Min.   : 8.40   Min.   : 8.80   Min.   :14.50   Min.   : 8.70
##  1st Qu.:21.00   1st Qu.:23.20   1st Qu.:24.20   1st Qu.:13.40
##  Median :27.30   Median :29.40   Median :26.90   Median :14.60
##  Mean   :28.44   Mean   :31.24   Mean   :28.26   Mean   :14.12
##  3rd Qu.:36.00   3rd Qu.:39.70   3rd Qu.:37.80   3rd Qu.:15.30
##  Max.   :63.40   Max.   :68.00   Max.   :44.50   Max.   :20.90
```

The record with zero weight has been removed.

(c) Note that `Code` for the species of fish. This is currently numerical and needs to be converted to a factor. Use the code `fish1$Code <- factor(fish1$Code)`. (Note that if `Code` is left as numerical, the model will estimate a single coefficient for it. The contribution of this variable will then be linear in this coefficient. So for example, the effect of a value 2 for `Code` is twice that for a value 1. This is not correct.)

Solution

```r
fish1$Code <- factor(fish1$Code)
```

(d) Fit a linear regression model with `Weight` as response against the other covariates.

Solution

```r
weight.lm <- lm(Weight ~ Code + Length1 + Length2 + Length3 + Height +
    Width, data = fish1)
summary(weight.lm)
```

```
##
## Call:
## lm(formula = Weight ~ Code + Length1 + Length2 + Length3 + Height +
##     Width, data = fish1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -185.90  -56.46  -14.48   36.35  411.95
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1139.213    210.083  -5.423  2.4e-07 ***
## Code2         100.918    100.913   1.000 0.318952
## Code3         118.322     98.156   1.205 0.229993
## Code4         130.239     70.235   1.854 0.065724 .
## Code5         515.362    144.989   3.554 0.000512 ***
## Code6        -121.299    153.787  -0.789 0.431549
## Code7         149.591    128.067   1.168 0.244696
## Length1       -64.577     35.801  -1.804 0.073338 .
## Length2        64.899     44.943   1.444 0.150889
## Length3        33.240     27.961   1.189 0.236473
## Height          5.089      5.816   0.875 0.383005
## Width           6.757      8.313   0.813 0.417655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 91.69 on 145 degrees of freedom
```

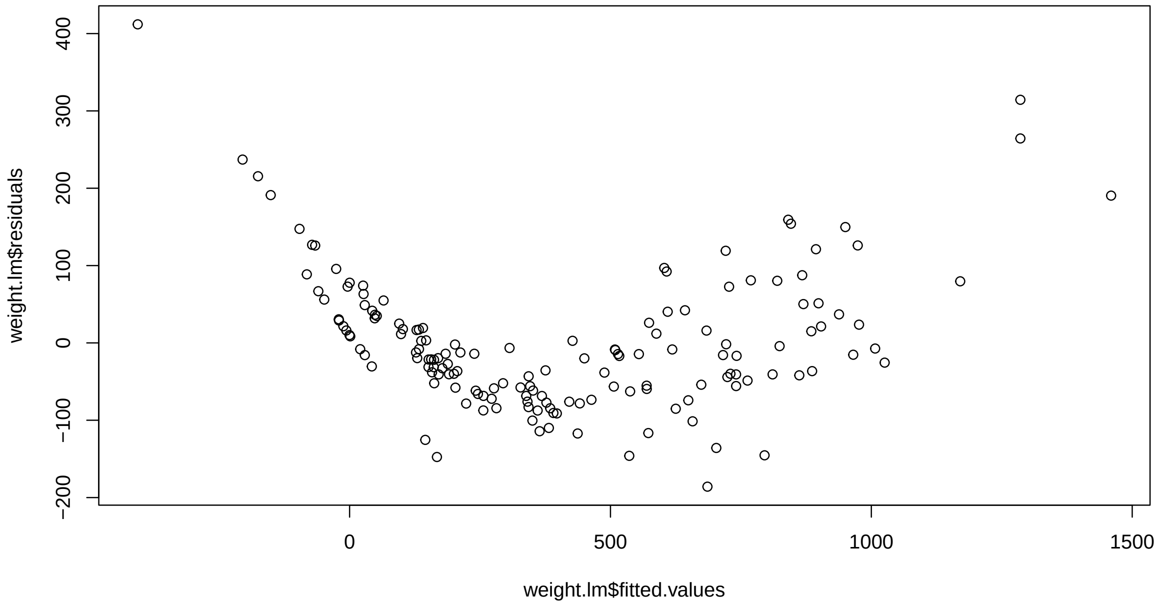
```
## Multiple R-squared:  0.9393,Adjusted R-squared:  0.9347
## F-statistic:   204 on 11 and 145 DF,  p-value: < 2.2e-16
```

(e) Investigate interaction terms in the model.
Solution

```
weight.lm <- lm(Weight ~ Code + Length1 + Length2 + Length3 + Height +
    Width, data = fish1)
summary(weight.lm)

##
## Call:
## lm(formula = Weight ~ Code + Length1 + Length2 + Length3 + Height +
##     Width, data = fish1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -185.90  -56.46  -14.48   36.35  411.95
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1139.213    210.083  -5.423  2.4e-07 ***
## Code2         100.918    100.913   1.000 0.318952
## Code3         118.322     98.156   1.205 0.229993
## Code4         130.239     70.235   1.854 0.065724 .
## Code5         515.362    144.989   3.554 0.000512 ***
## Code6        -121.299    153.787  -0.789 0.431549
## Code7         149.591    128.067   1.168 0.244696
## Length1       -64.577     35.801  -1.804 0.073338 .
## Length2        64.899     44.943   1.444 0.150889
## Length3        33.240     27.961   1.189 0.236473
## Height          5.089      5.816   0.875 0.383005
## Width           6.757      8.313   0.813 0.417655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 91.69 on 145 degrees of freedom
## Multiple R-squared:  0.9393,Adjusted R-squared:  0.9347
## F-statistic:   204 on 11 and 145 DF,  p-value: < 2.2e-16
```

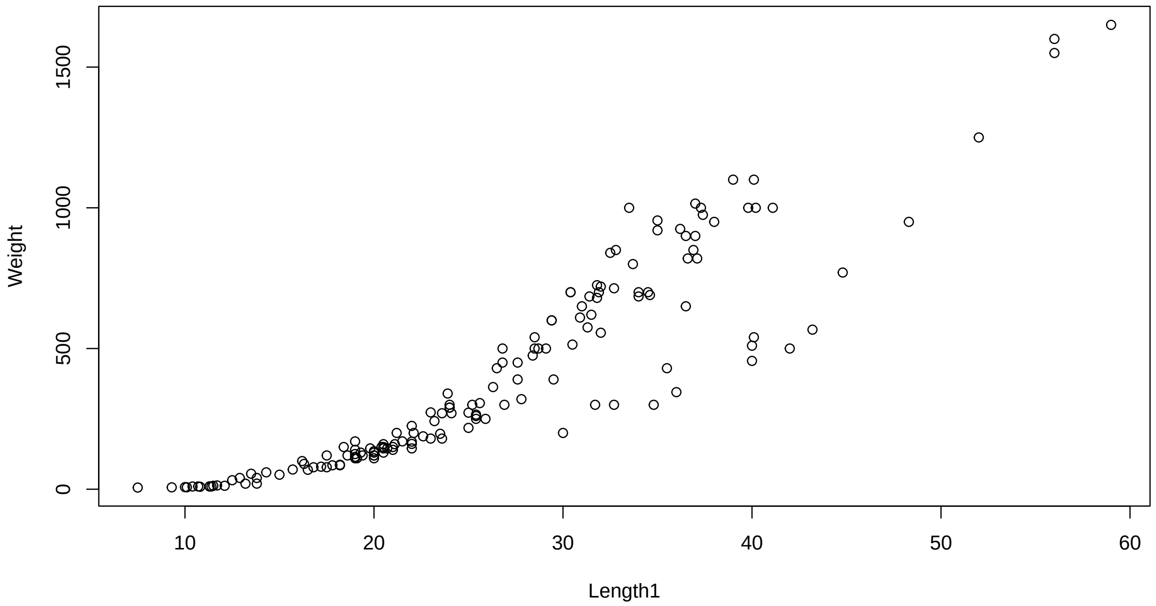
(f) Perform model diagnostics. In particular, examine the plot of residuals against fitted values for any patterns (indicating issues with a linear model fit) or change in spread (indicating a violation of homogeneous variance assumption).
Solution
You can see the variables stored under the linear model object by

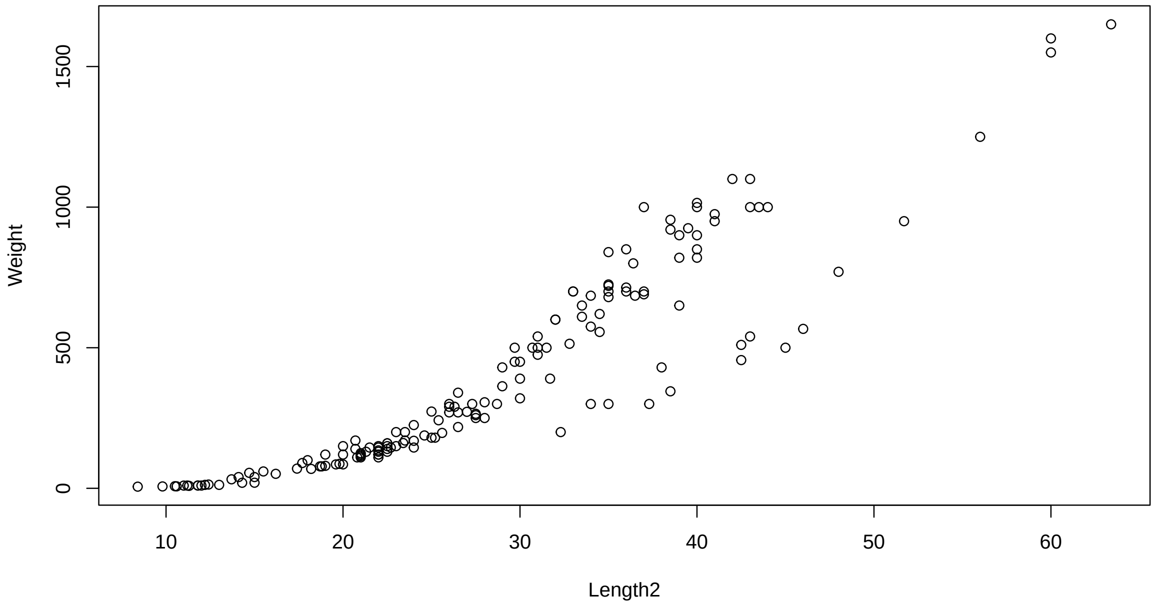
```
names(weight.lm)

##  [1] "coefficients"  "residuals"     "effects"
##  [4] "rank"          "fitted.values" "assign"
##  [7] "qr"            "df.residual"   "contrasts"
## [10] "xlevels"       "call"          "terms"
## [13] "model"

plot(weight.lm$residuals ~ weight.lm$fitted.values)
```



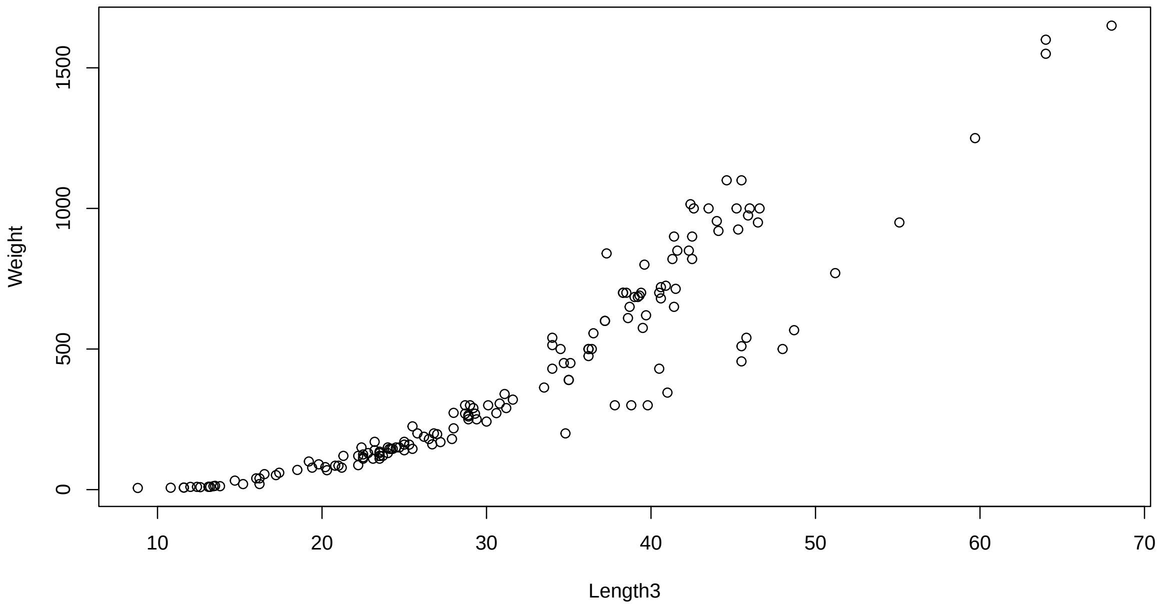
(g) By examining plots of the explanatory variables against the response variable, determine an appropriate transformation of data to improve the model for weight against the other morphological measurements.
Solution

```
with(fish1, plot(Weight ~ Length1))
with(fish1, plot(Weight ~ Length2))
with(fish1, plot(Weight ~ Length3))
with(fish1, plot(Weight ~ Height))
with(fish1, plot(Weight ~ Width))
```

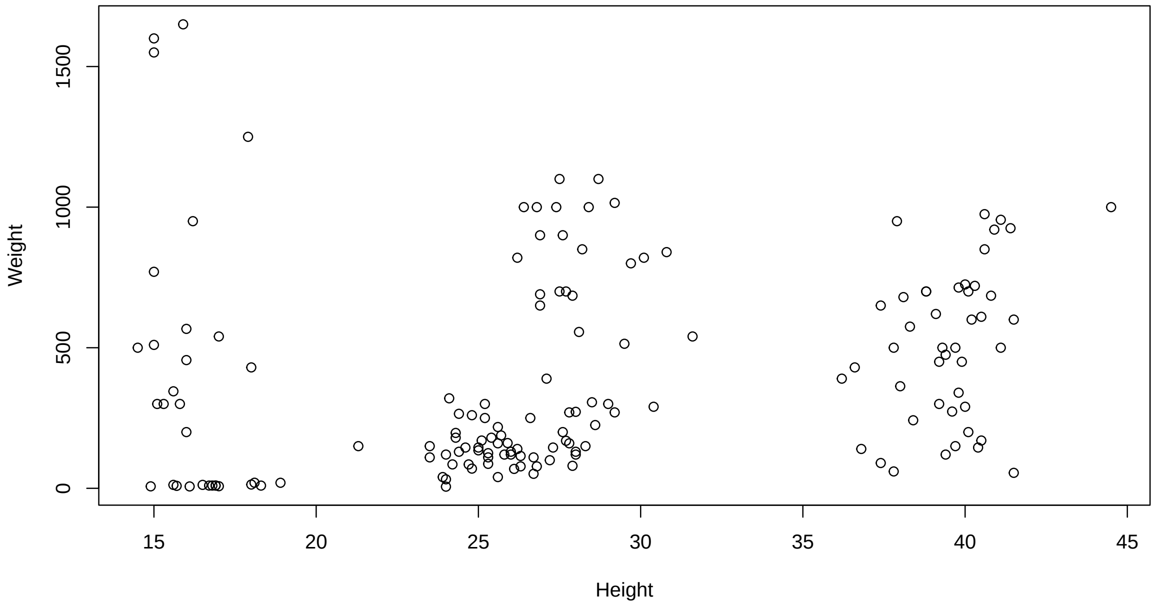


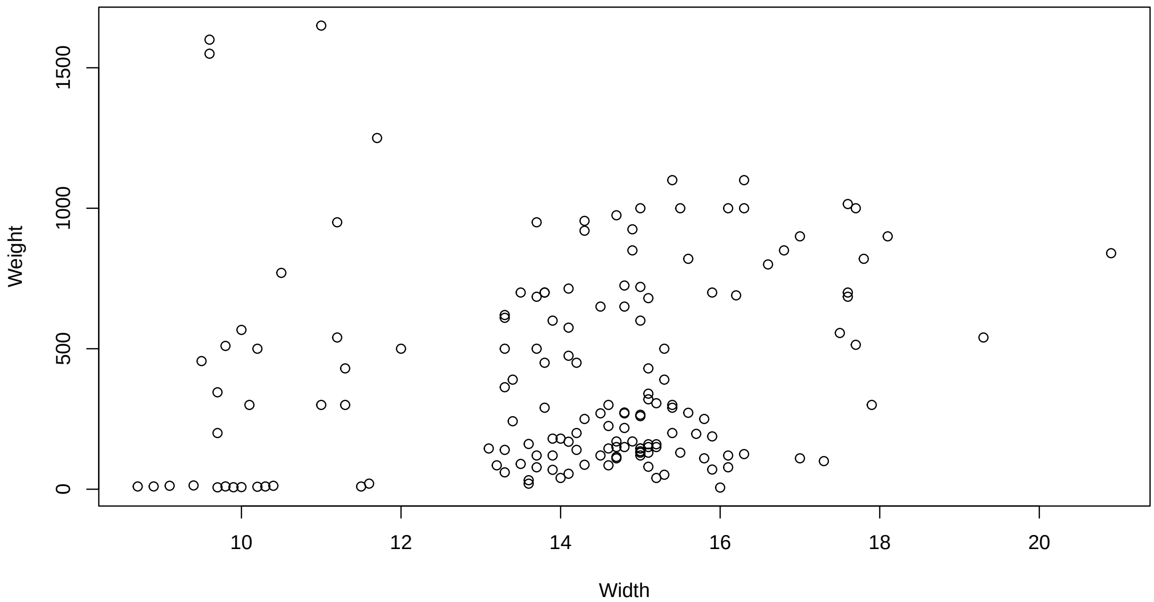


Weight
0
500
1000
1500
10
20
30
40
50
60
Length2



Weight
0
500
1000
1500
10
20
30
40
50
60
70
Length3



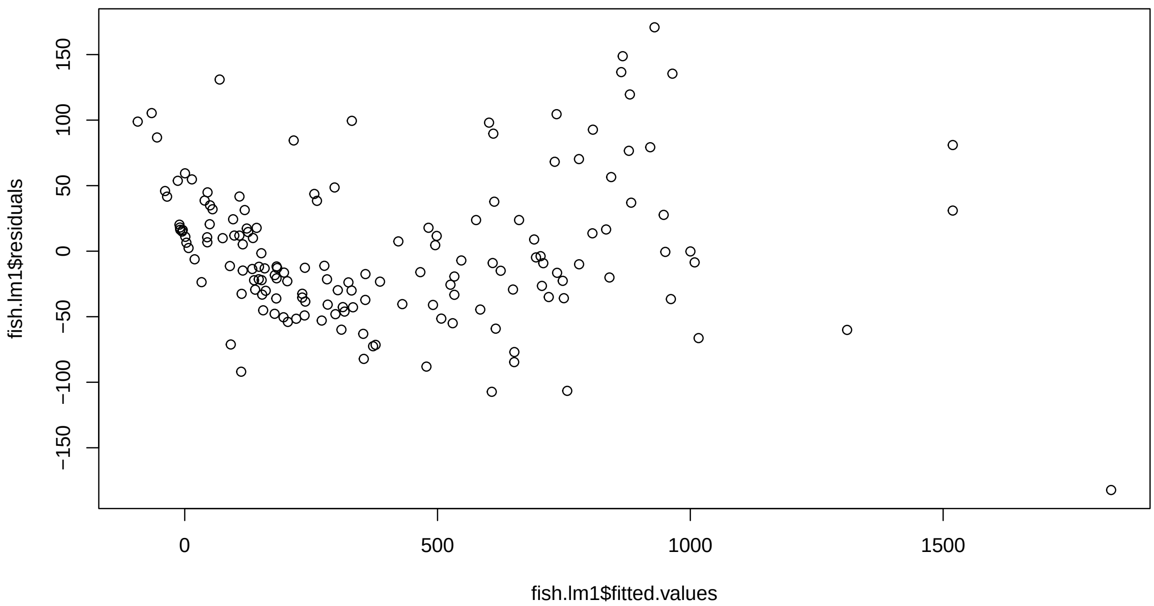


It seems that `Length1, Length2, Length3` have a quadratic or exponential relationship with `Weight`. So we first fit the qaudratic terms.

```
fish.lm1 <- lm(Weight ~ Code + I(Length1^2) + I(Length2^2) + I(Length3^2) +
    Width + Height, data = fish1)
summary(fish.lm1)

##
## Call:
## lm(formula = Weight ~ Code + I(Length1^2) + I(Length2^2) + I(Length3^2) +
##     Width + Height, data = fish1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -182.23  -34.99  -10.00   24.34  170.80
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -923.4640   125.0874  -7.383 1.12e-11 ***
## Code2         211.0505    59.0143   3.576 0.000474 ***
## Code3         176.3270    59.0938   2.984 0.003341 **
## Code4          71.1070    41.3197   1.721 0.087403 .
## Code5         475.4728    84.0179   5.659 7.87e-08 ***
## Code6          19.6182    90.0876   0.218 0.827916
## Code7         248.0920    71.5359   3.468 0.000690 ***
## I(Length1^2)    0.1695     0.3832   0.442 0.658923
## I(Length2^2)   -0.5031     0.4430  -1.136 0.257950
## I(Length3^2)    0.8222     0.2038   4.035 8.81e-05 ***
## Width          14.0911     5.0960   2.765 0.006430 **
## Height         13.3023     3.5586   3.738 0.000266 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.69 on 145 degrees of freedom
## Multiple R-squared:  0.9768,Adjusted R-squared:  0.975
## F-statistic:   555 on 11 and 145 DF,  p-value: < 2.2e-16
```

```
plot(fish.lm1$residuals ~ fish.lm1$fitted.values)
```

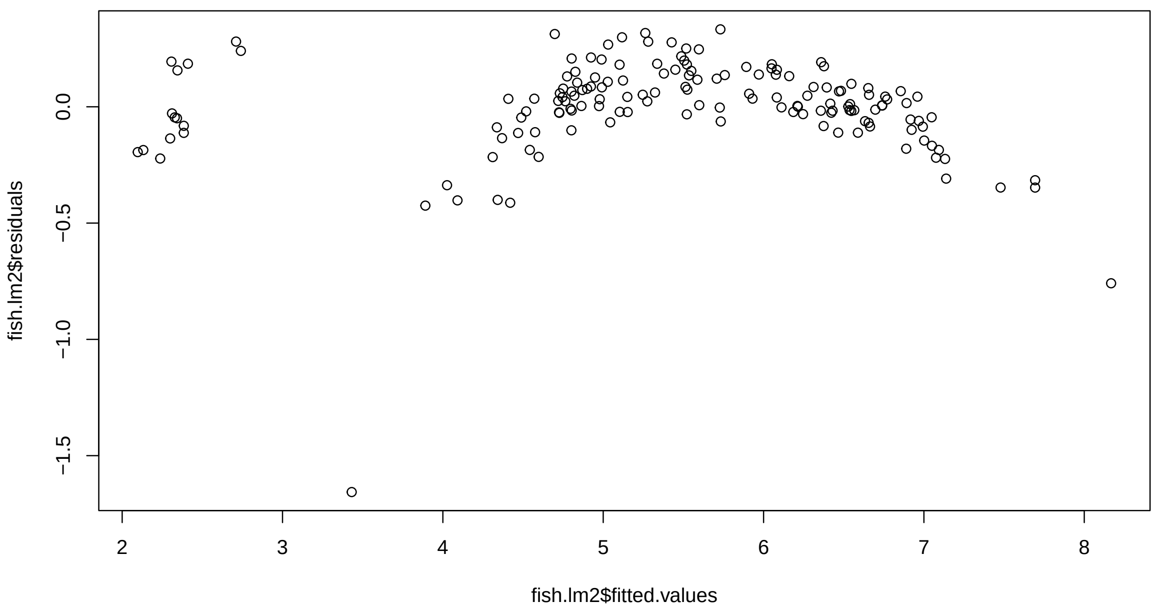


The residuals look better, but there is still a pattern, and the values of the residuals are quite large. We next fit an exponential model.

```
fish.lm2 <- lm(log(Weight) ~ Code + Length1 + Length2 + Length3 +
    Width + Height, data = fish1)
summary(fish.lm2)
```

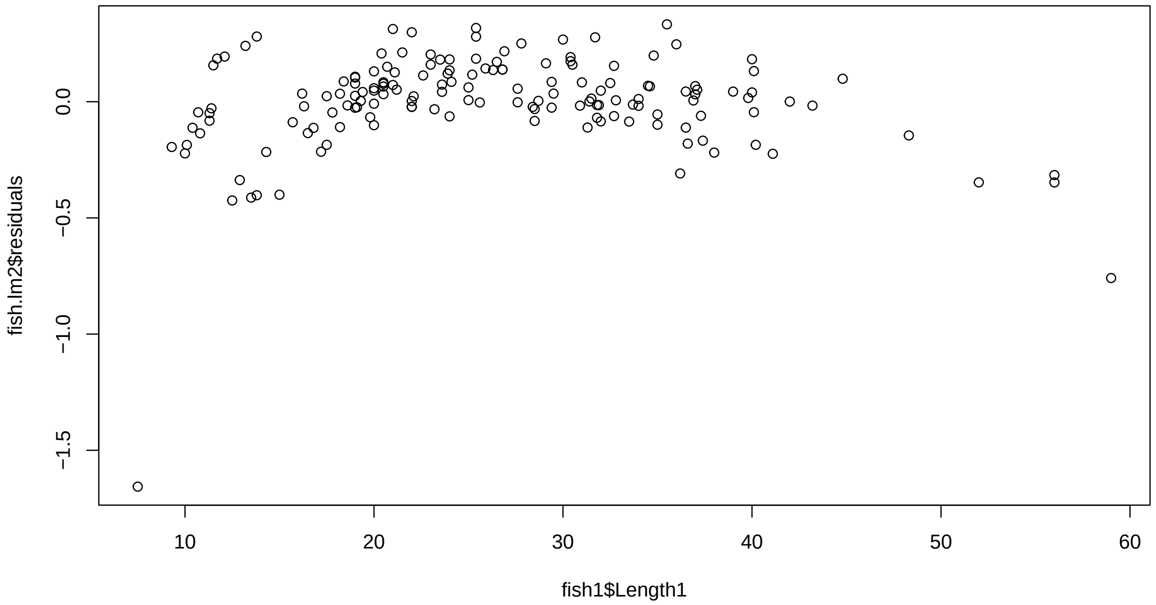
```
##
## Call:
## lm(formula = log(Weight) ~ Code + Length1 + Length2 + Length3 +
##     Width + Height, data = fish1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65654 -0.06310  0.02332  0.12087  0.33294
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

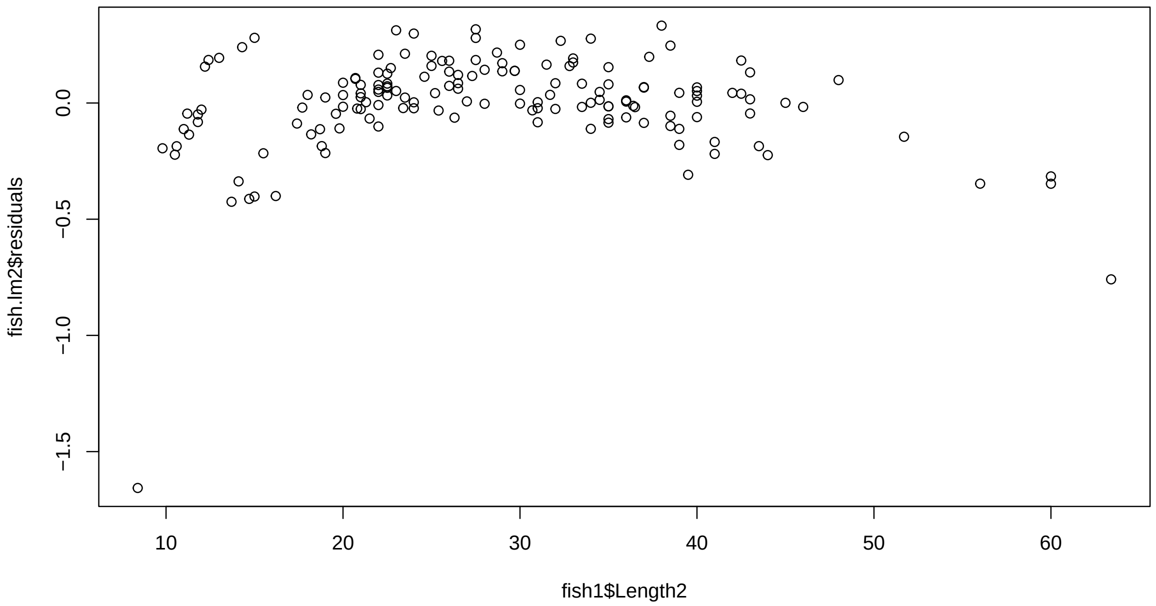
```
## (Intercept)  0.26271    0.50679   0.518   0.6050
## Code2        0.62686    0.24344   2.575   0.0110 *
## Code3        0.40891    0.23679   1.727   0.0863 .
## Code4        0.06833    0.16943   0.403   0.6873
## Code5       -0.40189    0.34976  -1.149   0.2524
## Code6        0.35798    0.37099   0.965   0.3362
## Code7        0.57882    0.30894   1.874   0.0630 .
## Length1      0.16580    0.08636   1.920   0.0568 .
## Length2     -0.22471    0.10842  -2.073   0.0400 *
## Length3      0.15843    0.06745   2.349   0.0202 *
## Width        0.04729    0.02005   2.358   0.0197 *
## Height       0.04513    0.01403   3.217   0.0016 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2212 on 145 degrees of freedom
## Multiple R-squared:  0.9743,Adjusted R-squared:  0.9723
## F-statistic: 499.5 on 11 and 145 DF,  p-value: < 2.2e-16

plot(fish.lm2$residuals ~ fish.lm2$fitted.values)
```

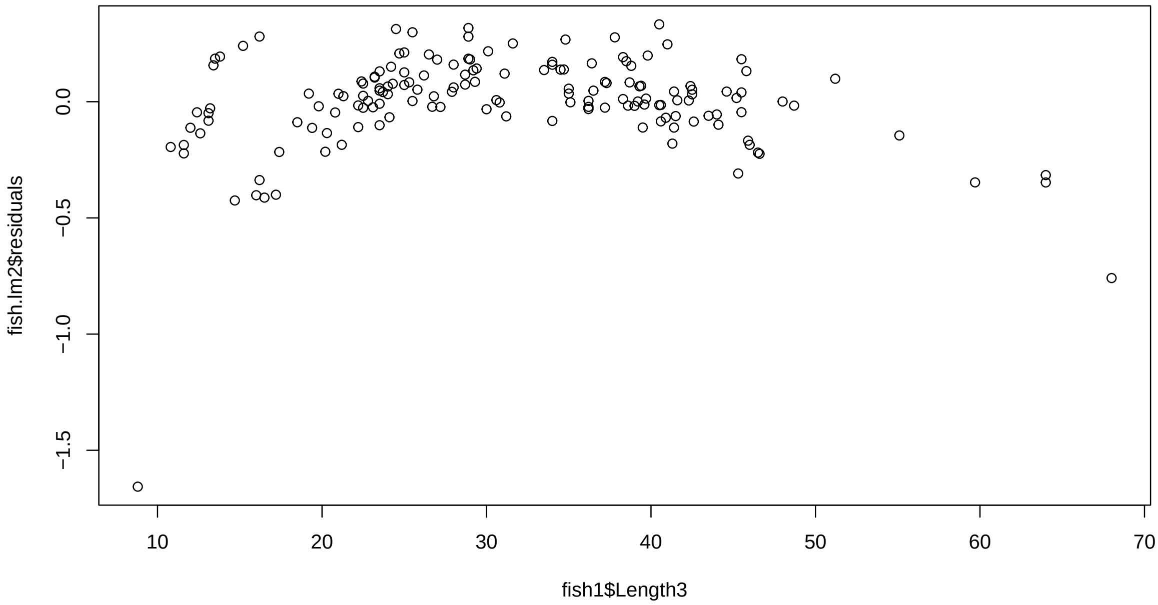


Residual plot is a lot better, and the residuals are smaller in value, but this may be the effect of taking the log of the data. The curvature in the residuals plot can be further investigated by plotting residuals against the covariates, to determine any further relationships.

```
plot(fish.lm2$residuals ~ fish1$Length1)
plot(fish.lm2$residuals ~ fish1$Length2)
plot(fish.lm2$residuals ~ fish1$Length3)
```


fish.lm2$residuals
fish1$Length1
0.0
−0.5
−1.0
−1.5
10
20
30
40
50
60


fish.lm2$residuals
fish1$Length2
0.0
−0.5
−1.0
−1.5
10
20
30
40
50
60



It appears that there is still a quadratic term in `Length1`, `Length2` and `Length3`. But this is not important, as the values of the residuals are quite small compared to the values of `Weight`.

(h) Fit your selected model.

Solution

Done as above.

(i) Reduce the model removing non-significant variables one by one, until a model with only significant terms is left. Use `update` command. For example, `fish.lm1 <- update(fish.lm .~.-Length1)`.

Solution

We first remove `Length1`.

```
fish.lm3 <- update(fish.lm2, . ~ . - Length1)
summary(fish.lm3)

##
## Call:
## lm(formula = log(Weight) ~ Code + Length2 + Length3 + Width +
##     Height, data = fish1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63333 -0.06780  0.02185  0.11417  0.35498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.23677    0.51125   0.463  0.64397
## Code2        0.63885    0.24558   2.601  0.01024 *
## Code3        0.47441    0.23646   2.006  0.04667 *
## Code4        0.14633    0.16599   0.882  0.37946
## Code5       -0.30731    0.34945  -0.879  0.38062
## Code6        0.37306    0.37430   0.997  0.32057
## Code7        0.61131    0.31130   1.964  0.05146 .
## Length2     -0.06782    0.07190  -0.943  0.34712
## Length3      0.15811    0.06807   2.323  0.02158 *
## Width        0.04447    0.02018   2.203  0.02914 *
## Height       0.04279    0.01410   3.034  0.00286 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2232 on 146 degrees of freedom
## Multiple R-squared:  0.9736,Adjusted R-squared:  0.9718
## F-statistic: 539.2 on 10 and 146 DF,  p-value: < 2.2e-16
```

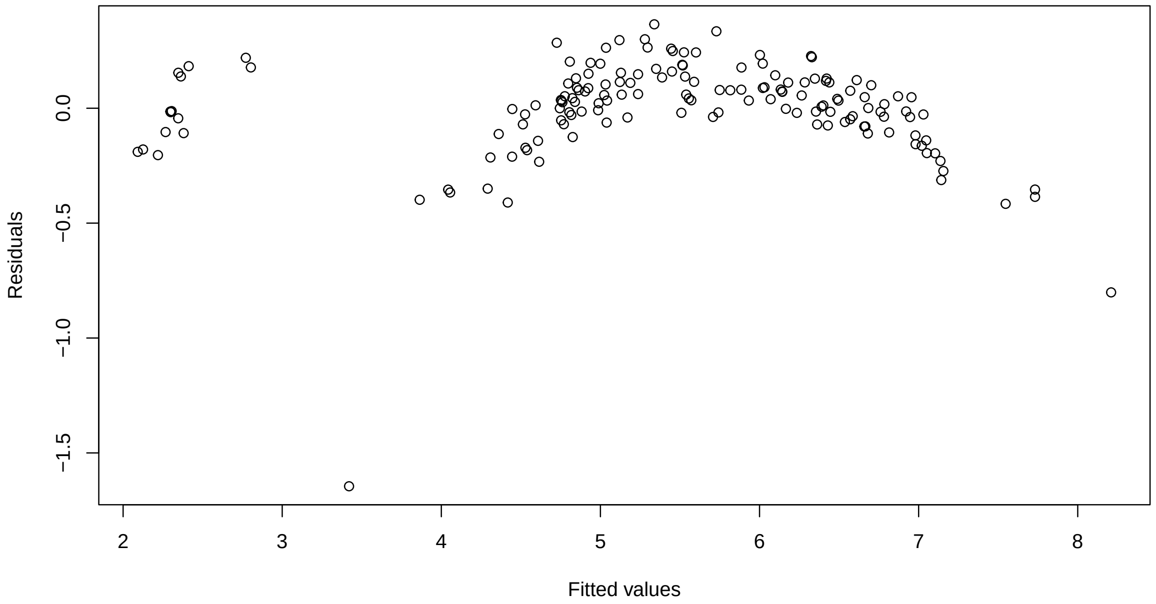
```
fish.lm4 <- update(fish.lm3, . ~ . - Length2)
summary(fish.lm4)
```

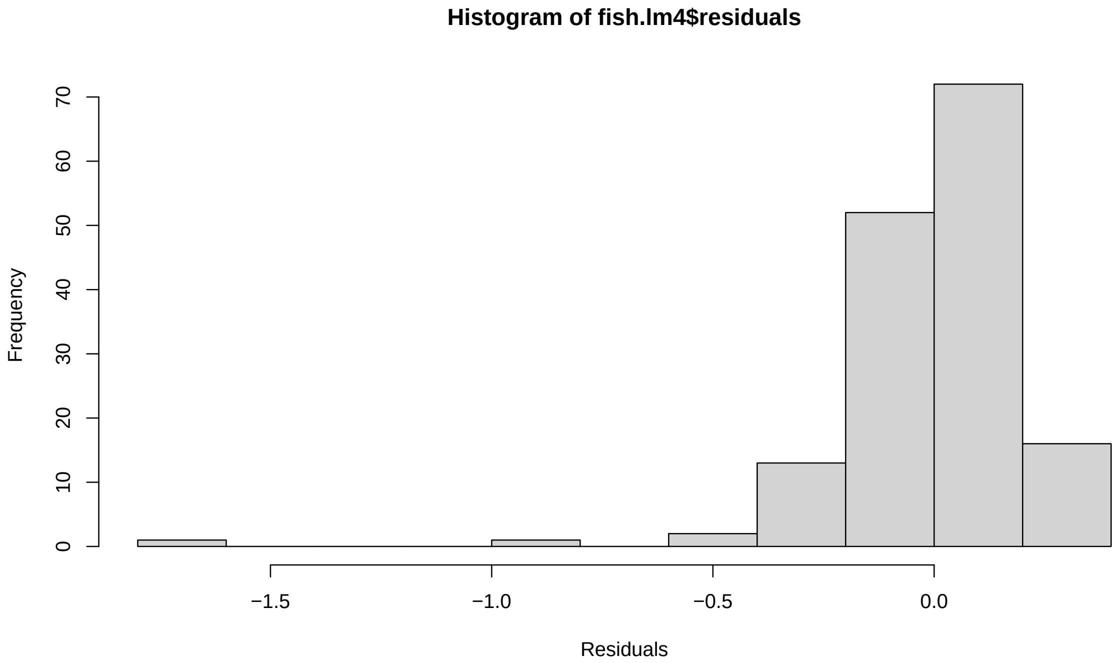
```
##
## Call:
## lm(formula = log(Weight) ~ Code + Length3 + Width + Height, data = fish1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.64509 -0.06965  0.03323  0.11493  0.36520
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.447596   0.459629   0.974  0.33175
## Code2        0.494618   0.192103   2.575  0.01102 *
## Code3        0.345587   0.192956   1.791  0.07535 .
## Code4        0.013673   0.088132   0.155  0.87692
## Code5       -0.498258   0.284729  -1.750  0.08222 .
## Code6        0.194399   0.322718   0.602  0.54785
## Code7        0.393712   0.208940   1.884  0.06149 .
## Length3      0.093954   0.002773  33.885  < 2e-16 ***
## Width        0.045940   0.020116   2.284  0.02382 *
## Height       0.042371   0.014092   3.007  0.00311 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2231 on 147 degrees of freedom
## Multiple R-squared:  0.9735,Adjusted R-squared:  0.9719
## F-statistic: 599.5 on 9 and 147 DF,  p-value: < 2.2e-16
```

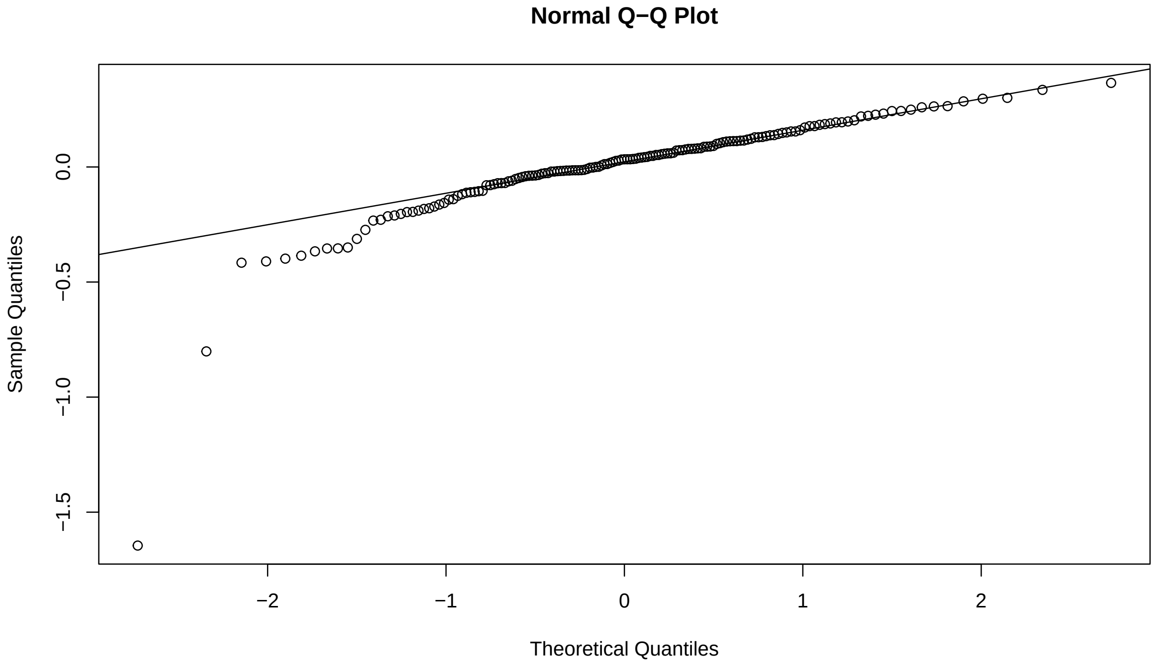
`Length3` is now significant. This is the final model.

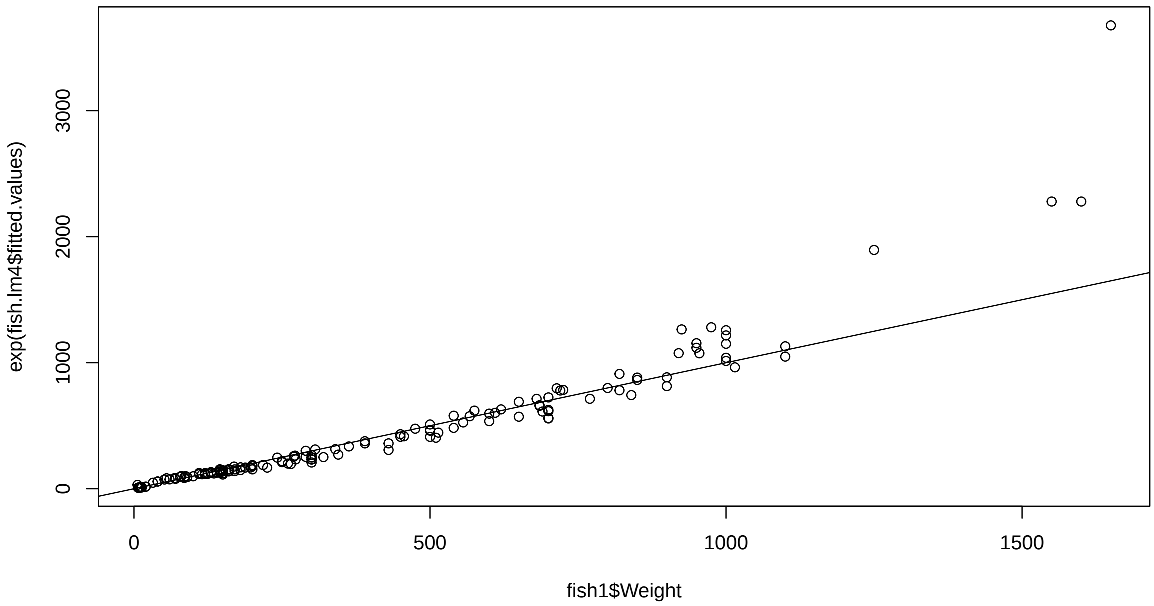
(j) Perform model diagnostics. For this, plot a histogram of the residuals and a scatter plot of the residuals against the fitted values. Comment on whether the model assumptions are satisfied.
Solution

```
plot(fish.lm4$residuals ~ fish.lm4$fitted.values, xlab = "Fitted values",
    ylab = "Residuals")
hist(fish.lm4$residuals, xlab = "Residuals")
qqnorm(fish.lm4$residuals)
qqline(fish.lm4$residuals)
plot(exp(fish.lm4$fitted.values) ~ fish1$Weight)
abline(0, 1)
```


Residuals
Fitted values
0.0
−0.5
−1.0
−1.5
2
3
4
5
6
7
8

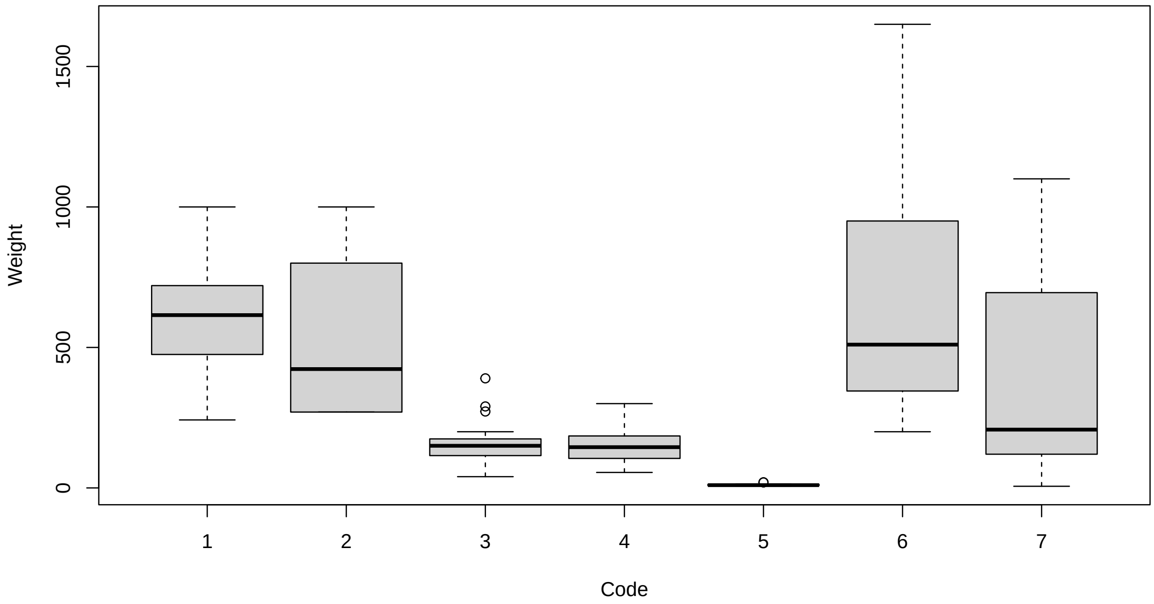

Histogram of fish.lm4$residuals
Frequency
Residuals
0
10
20
30
40
50
60
70
−1.5
−1.0
−0.5
0.0





The plot of residuals against fitted values shows a sligh curvature, but this is not important given the small values of residuals. The histogram of residuals shows a small outlier, but otherwise does not look too different from that expected for a normal distribution. The normal probability plot of residuals has some departure from a straight line at the lower end, indicating some right skewness, but this may be due to the outlier. The plot of fitted values against observed values shows a good fit for values of `Weight` uptil around 1,000.

```
fish1[which(fish1$Weight > 1100), ]

##     Code Weight Length1 Length2 Length3 Height Width
## 99     6   1250      52    56.0    59.7   17.9  11.7
## 100    6   1600      56    60.0    64.0   15.0   9.6
## 101    6   1550      56    60.0    64.0   15.0   9.6
## 102    6   1650      59    63.4    68.0   15.9  11.0

plot(fish1$Weight ~ fish1$Code, xlab = "Code", ylab = "Weight")
```



The fish with larger weights are all `Code` 6. A boxplot of `Weight` against `Code` shows outliers for `Code` 6.

Overall the model is satisfactory.

(k) Explore the data further and decide how the model can be improved.

Solution

We could fit a quadratic term for `Length1`. Also, the initial plot of `Weight` against `Height` indicated some three groups in `Height`, so we could categorise this variable. However, the overall fit of the final model is good, and these improvements may not not much difference.

(l) Report your findings on the dependence of the weight of the fish on the explanatory variables.

Solution

The fitted model equation is

$$\log \text{Weight} \ = 0.4476 + 0.4946\text{Code2} \ + 0.0940\text{Length3} \ + 0.0459\text{Width} \ + 0.0424\text{Height}$$

If we remove the log by taking exponential of both sides, we get

$$\begin{aligned}\text{Weight} \ &= \exp 0.4476 + 0.4946\text{Code2} \ + 0.0940\text{Length3} \ + 0.0424\text{Height} \\ &= \exp(0.4476) \times \exp(0.4946 \times \text{Code2} \ ) \times \exp(0.0940 \times \text{Length3} \ )\times \\ &= \quad \exp(0.0459 \times \text{Width} \ \times \exp(0.0424 \times \text{Height}).\end{aligned}$$

If all the other variable are held constant and the fish has Code $= 2$, then the weight of the fish, compared with other values of Code, is $\exp(0.4946) = 1.64$ times larger. Similarly, if all other variables are kept fixed and Length3 increases by 1 cm, then the Weight is $\exp(0.0940) = 1.099$ times larger. These effects are summarised below.

```
exp(fish.lm4$coefficients[c(2, 8:10)])

##     Code2  Length3    Width   Height
## 1.639872 1.098509 1.047012 1.043281
```

Fish species of `Code` 2 have higher mean weights by a factor of 1.6, compared with the other species. For every cm increase in `Length3`, the weight of the fish increases by a factor of 1.099. Similarly for every cm increase in `Width` the weight increases by a factor of 1.05, and for every cm increase in `Width` the weight increases by a factor of 1.04.

Exercise 4: Bank data The folder Data in the Computer Labs folder contains the data set `Bank.txt`. The female employees are suing the bank for gender discrimation in salary. Download the file and read the data into `R`. For each employee the Bank Data as the following variables.

- EducLev: education level, a categorical variable with categories 1 (finished high school), 2 (finished some tertiary education), 3 (obtained a bachelor's degree), 4 (took some postgraduate courses), 5 (obtained a postgraduate degree).
- Job Grade: a categorical variable indicating the current job level, the possible levels being 1 (lowest) to 6 (highest).
- YrHired: year employee was hired.
- YrBorn: year employee was born.
- Gender: a categorical variable with values "Female" and "Male".
- YrsPrior: number of years of work experience at another bank prior to working at First National.
- PCJob: a categorical yes/no variable indicating whether the employees current job is PC related.
- Salary: current salary in thousands of dollar.

1. Summarise the data and check for any data errors.
2. Fit a linear model to the Salary. Do not include any interactions.
3. Reduce the model to only significant terms.
4. Perform appropriate model diagnostics.
5. Is there a gender bias in salaries? Justify your decision.
6. Now include appropriate interaction terms. You may have to consider which interactions are meaningful.
7. Again reduce the model to only significant terms.
8. Perform model diagnostics.
9. Under this new model, is there gender bias in salaries? Justify your decision.
10. Produce a scatterplot of fitted salaries against observed salaries. The plotting character should be Sex (M or F), and the colour code should be by education level.
11. Comment your findings from the plot.
12. What form of discrimination can you detect from your analysis?
    Solution

```
## The following objects are masked from bank (pos = 3):
##
##     EducLev, Employee, Exp, Gender, JobGrade, PCJob, Salary, YrBorn,
##     YrHired, YrsPrior

## The following objects are masked from bank (pos = 4):
##
##     EducLev, Employee, Exp, Gender, JobGrade, PCJob, Salary, YrBorn,
##     YrHired, YrsPrior

##
## Call:
## lm(formula = Salary ~ YrsPrior + Exp + YrBorn + Gender + PCJob +
##     EducLev + JobGrade, data = bank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.117  -2.359  -0.397   1.778  23.958
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     -9.533e+02  1.888e+02  -5.049 1.02e-06 ***
## YrsPrior         1.677e-01  1.404e-01   1.194   0.2338
## Exp              5.156e-01  9.798e-02   5.262 3.77e-07 ***
## YrBorn           8.962e-03  5.770e-02   0.155   0.8767
## GenderMale       2.554e+00  1.012e+00   2.524   0.0124 *
## PCJobYes         4.923e+00  1.474e+00   3.340   0.0010 **
## EducLevTE       -4.856e-01  1.399e+00  -0.347   0.7289
## EducLevBach      5.279e-01  1.358e+00   0.389   0.6978
## EducLevPGrad     2.852e-01  2.405e+00   0.119   0.9057
## EducLevPGDegree  2.691e+00  1.621e+00   1.660   0.0985 .
## JobGrade2        1.564e+00  1.186e+00   1.319   0.1886
## JobGrade3        5.219e+00  1.262e+00   4.134 5.30e-05 ***
## JobGrade4        8.595e+00  1.496e+00   5.745 3.53e-08 ***
## JobGrade5        1.366e+01  1.874e+00   7.288 7.86e-12 ***
## JobGrade6        2.383e+01  2.800e+00   8.512 4.75e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.648 on 193 degrees of freedom
## Multiple R-squared:  0.7652,Adjusted R-squared:  0.7482
## F-statistic: 44.94 on 14 and 193 DF,  p-value: < 2.2e-16
##
## Call:
## lm(formula = Salary ~ YrsPrior + Exp + Gender + PCJob + EducLev +
##     JobGrade, data = bank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.106  -2.395  -0.390   1.726  23.977
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -938.01896  160.67099  -5.838 2.19e-08 ***
## YrsPrior           0.16225    0.13560   1.197 0.232959
## Exp                0.50781    0.08404   6.042 7.63e-09 ***
## GenderMale         2.56288    1.00798   2.543 0.011784 *
## PCJobYes           4.91462    1.46917   3.345 0.000987 ***
## EducLevTE         -0.47126    1.39211  -0.339 0.735335
## EducLevBach        0.58537    1.30287   0.449 0.653724
## EducLevPGrad       0.31070    2.39306   0.130 0.896832
## EducLevPGDegree    2.73904    1.58685   1.726 0.085925 .
## JobGrade2          1.56939    1.18237   1.327 0.185961
## JobGrade3          5.21285    1.25852   4.142 5.13e-05 ***
## JobGrade4          8.58911    1.49180   5.758 3.29e-08 ***
## JobGrade5         13.65249    1.86902   7.305 7.03e-12 ***
## JobGrade6         23.78011    2.77258   8.577 3.08e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.634 on 194 degrees of freedom
## Multiple R-squared:  0.7652,Adjusted R-squared:  0.7495
## F-statistic: 48.64 on 13 and 194 DF,  p-value: < 2.2e-16
## Analysis of Variance Table
##
## Model 1: Salary ~ YrsPrior + Exp + YrBorn + Gender + PCJob + EducLev +
##     JobGrade
## Model 2: Salary ~ YrsPrior + Exp + Gender + PCJob + EducLev + JobGrade
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    193 6156.9
## 2    194 6157.6 -1  -0.76965 0.0241 0.8767
##
## Call:
## lm(formula = Salary ~ Exp + Gender + PCJob + EducLev + JobGrade,
##     data = bank)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.036  -2.310  -0.358   1.763  23.898
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -919.74746  160.12057  -5.744 3.50e-08 ***
## Exp                0.49839    0.08377   5.950 1.22e-08 ***
## GenderMale         2.60166    1.00857   2.580 0.010629 *
## PCJobYes           5.22461    1.44774   3.609 0.000391 ***
## EducLevTE         -0.17163    1.37092  -0.125 0.900498
## EducLevBach        0.45461    1.29971   0.350 0.726885
## EducLevPGrad       0.04650    2.38549   0.019 0.984466
## EducLevPGDegree    2.48850    1.57472   1.580 0.115663
## JobGrade2          1.68894    1.17944   1.432 0.153751
## JobGrade3          5.46275    1.24244   4.397 1.80e-05 ***
## JobGrade4          8.78830    1.48412   5.922 1.42e-08 ***
## JobGrade5         14.03735    1.84317   7.616 1.10e-12 ***
## JobGrade6         23.90777    2.77359   8.620 2.29e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.64 on 195 degrees of freedom
## Multiple R-squared:  0.7635,Adjusted R-squared:  0.7489
## F-statistic: 52.46 on 12 and 195 DF,  p-value: < 2.2e-16
## Analysis of Variance Table
##
## Model 1: Salary ~ YrsPrior + Exp + Gender + PCJob + EducLev + JobGrade
## Model 2: Salary ~ Exp + Gender + PCJob + EducLev + JobGrade
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    194 6157.6
## 2    195 6203.1 -1    -45.44 1.4316  0.233
##
## Call:
## lm(formula = Salary ~ Exp + Gender + PCJob + JobGrade, data = bank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.948  -2.456  -0.448   2.209  23.940
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -811.81992  144.94668  -5.601 7.03e-08 ***
## Exp            0.44190    0.07592   5.821 2.32e-08 ***
## GenderMale     2.78568    0.99967   2.787 0.005842 **
## PCJobYes       5.37626    1.42269   3.779 0.000208 ***
## JobGrade2      2.08424    1.15309   1.808 0.072190 .
## JobGrade3      6.18730    1.13061   5.473 1.33e-07 ***
## JobGrade4     10.06050    1.33027   7.563 1.42e-12 ***
## JobGrade5     16.05248    1.50849  10.641  < 2e-16 ***
## JobGrade6     26.58457    2.34785  11.323  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.647 on 199 degrees of freedom
## Multiple R-squared:  0.758,Adjusted R-squared:  0.7483
## F-statistic: 77.93 on 8 and 199 DF,  p-value: < 2.2e-16
## Analysis of Variance Table
##
## Model 1: Salary ~ Exp + Gender + PCJob + EducLev + JobGrade
## Model 2: Salary ~ Exp + Gender + PCJob + JobGrade
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    195 6203.1
## 2    199 6345.8 -4   -142.78 1.1221 0.3473
```

The final model suggests a gender bias in salaries, after adjusting for the effects of the other variables. BUT before we concluded that gender bias exists, we need to perform model diagnostics.

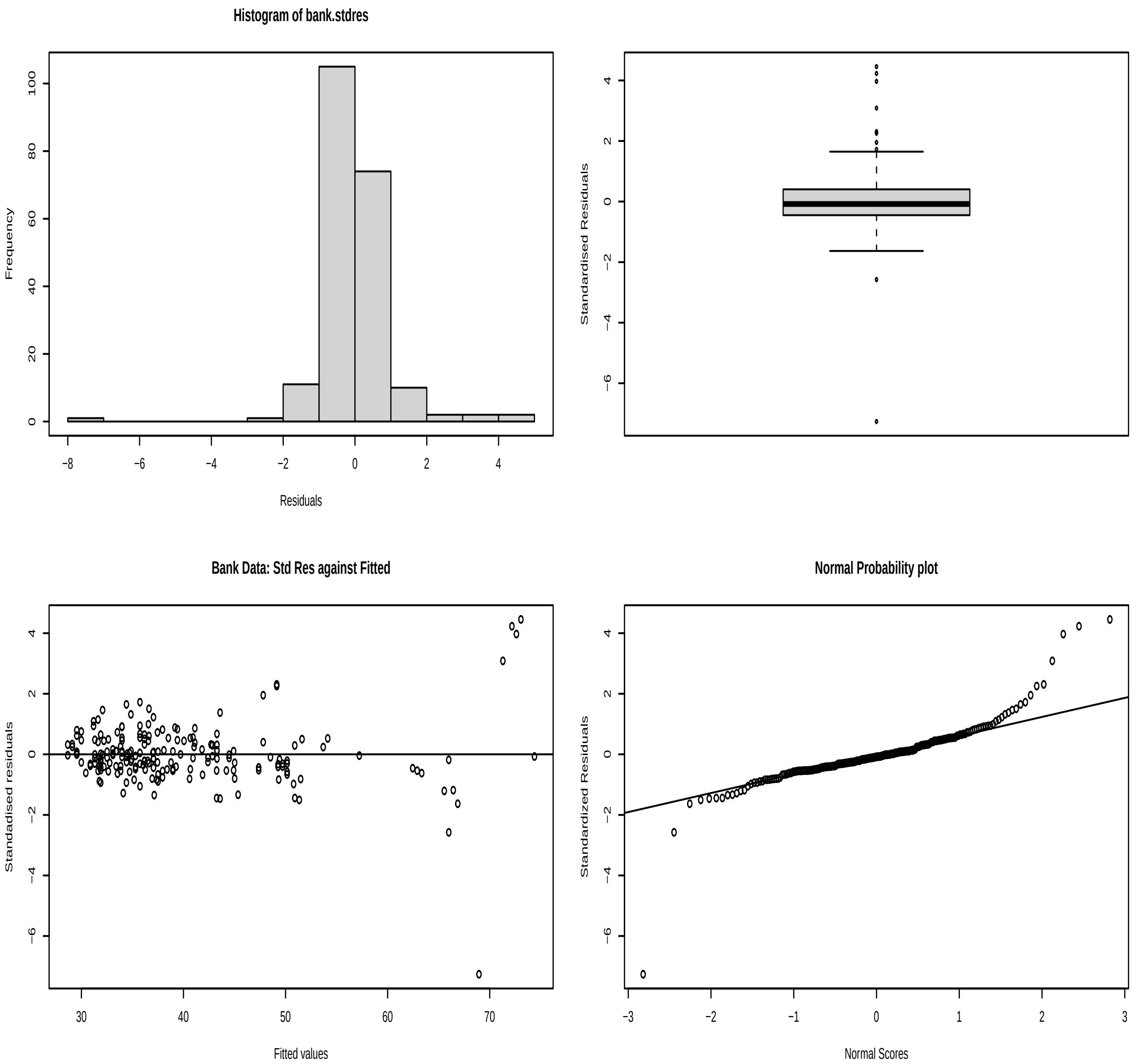


Figure 1: Diagnostic plots for Bank data linear model.

Main Observations: The central issue is the presence of outliers. The perceived issues with normality and constant variance are a result of these outliers. Once the outlier issue is fixed, these other issues should also be resolved, but of course we will need to re-investigate the model diagnostics.

What do we do with the outliers? First, we need to identify what they represent.

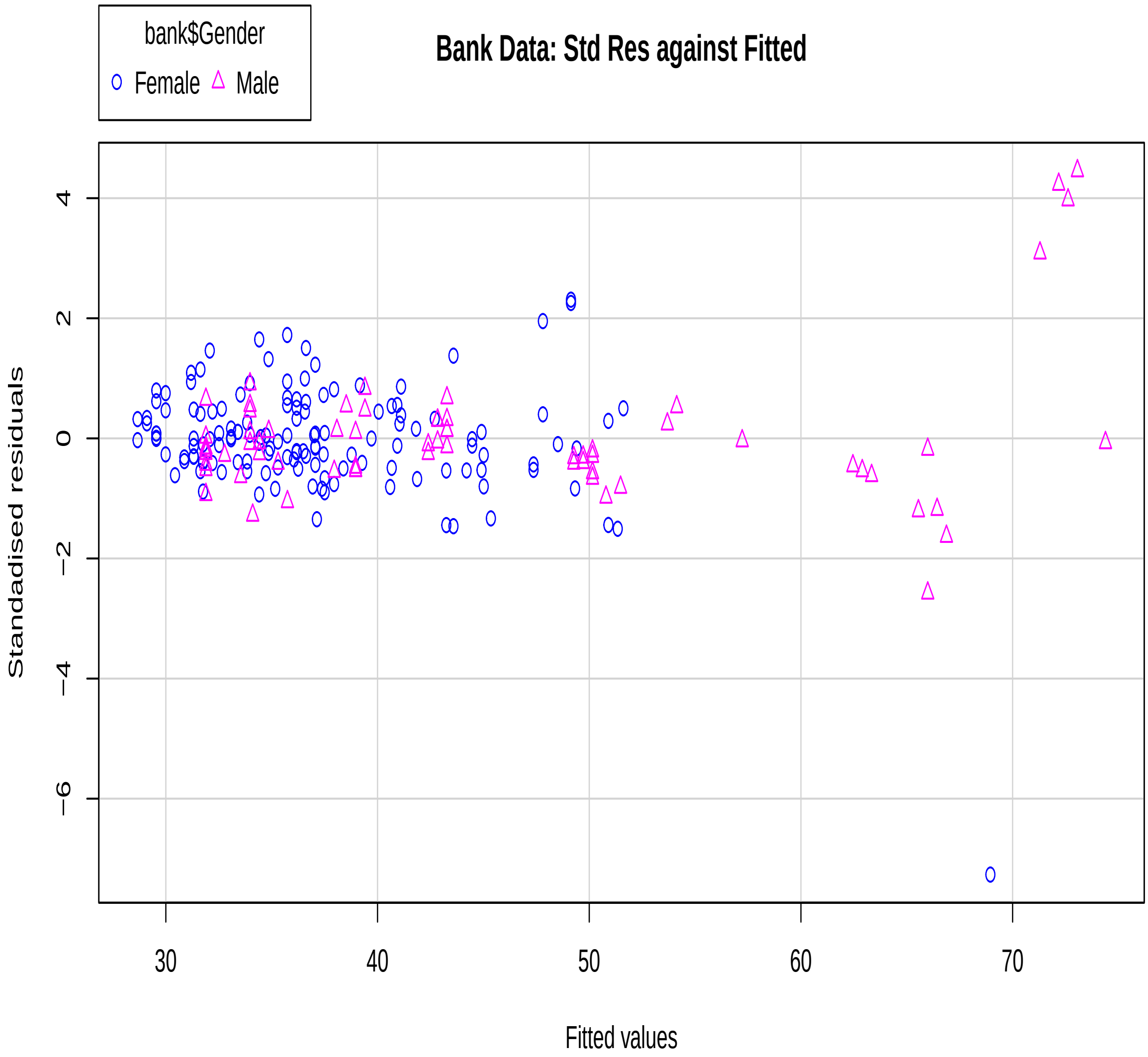


Figure 2: Scatter plot of residuals by Gender of employee.

The discovery is that the outliers are sex-related. In particular, the large residuals that represent large salaries correspond to males at Job Grade 6, while the small (negative) residual represents a small salary for a solitary female employee at Job Grade 6.

Question Why is this female employee being paid so little?

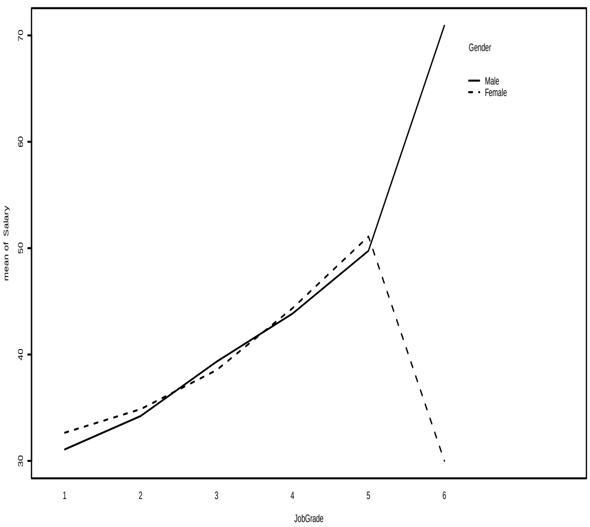
Answer Upon investigation it transpired that this female employee was close to retirement and so had chosen to work only half-time. Her salary is therefore half of the salary at this Job Grade. Once this is adjusted for, there is no difference in salary between the male and female employees at Job Grade 6.

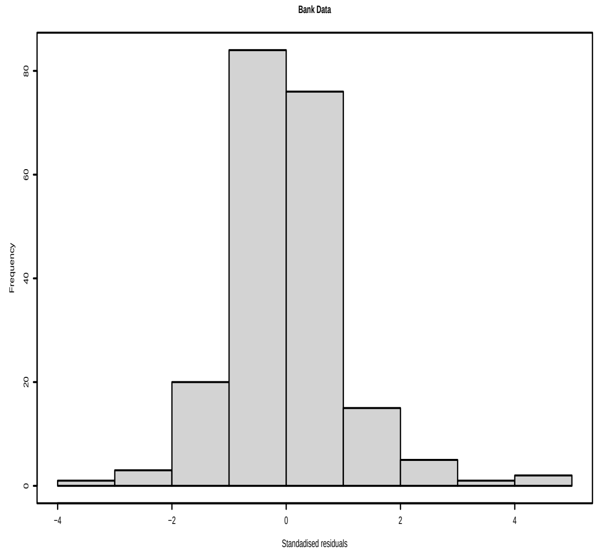
BUT, is there still a gender bias in salaries?

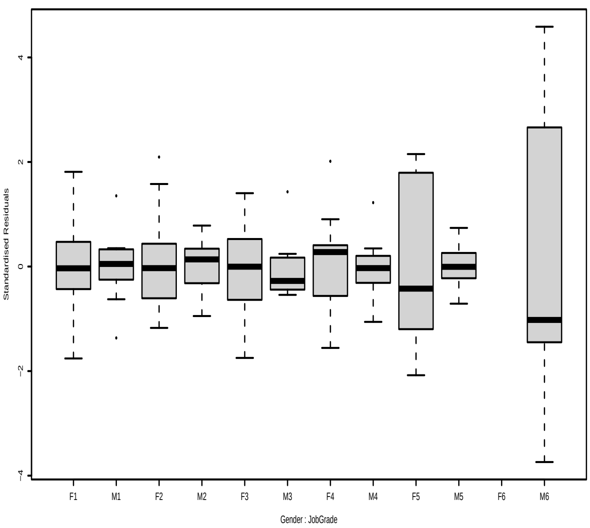
0.1 Interaction terms

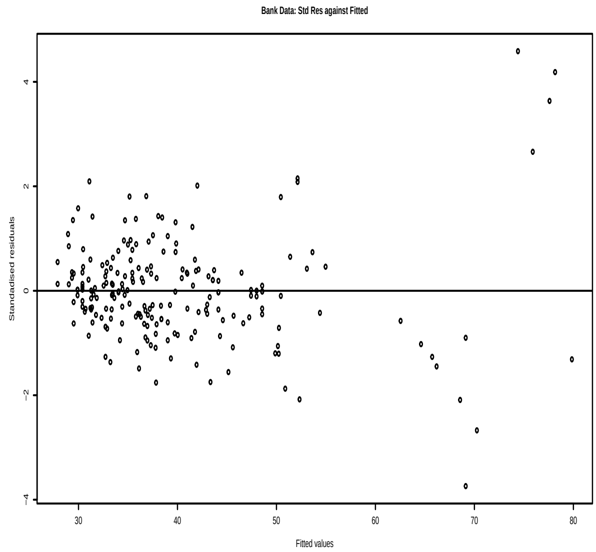
So far we have not investigated interactions. In particular, a data set that contains several categorical variables can expect to exhibit interactions.

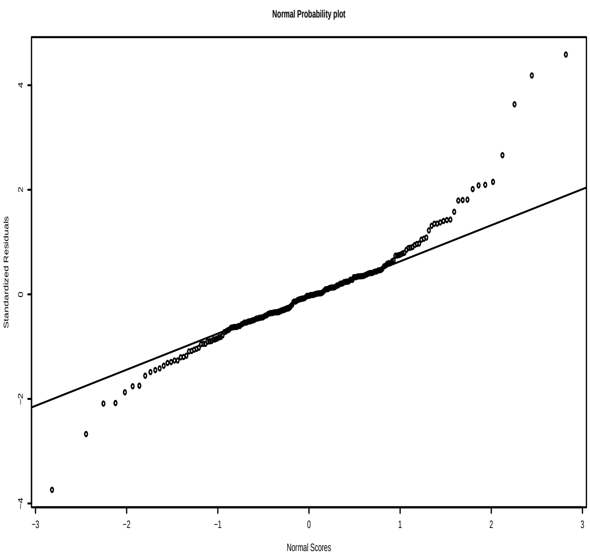
```
##
## Call:
## lm(formula = Salary ~ Exp + Gender + PCJob + EducLev + JobGrade +
##     Gender:JobGrade, data = bank)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1215  -2.3838  -0.1085   1.8712  20.6010
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.046e+03  1.371e+02  -7.632 1.09e-12 ***
## Exp                   5.646e-01  7.169e-02   7.875 2.55e-13 ***
## GenderMale            1.058e+00  1.597e+00   0.662 0.508549
## PCJobYes              5.109e+00  1.239e+00   4.125 5.54e-05 ***
## EducLevTE             7.768e-02  1.172e+00   0.066 0.947225
## EducLevBach           9.622e-01  1.123e+00   0.857 0.392622
## EducLevPGrad          8.206e-02  2.043e+00   0.040 0.967996
## EducLevPGDegree       3.589e+00  1.346e+00   2.665 0.008357 **
## JobGrade2             1.509e+00  1.155e+00   1.306 0.192963
## JobGrade3             4.794e+00  1.142e+00   4.197 4.14e-05 ***
## JobGrade4             7.904e+00  1.475e+00   5.357 2.43e-07 ***
## JobGrade5             1.509e+01  1.967e+00   7.669 8.76e-13 ***
## JobGrade6            -1.890e+01  5.349e+00  -3.534 0.000513 ***
## GenderMale:JobGrade2  9.046e-01  2.241e+00   0.404 0.686906
## GenderMale:JobGrade3  2.052e+00  2.530e+00   0.811 0.418304
## GenderMale:JobGrade4  1.513e+00  2.456e+00   0.616 0.538578
## GenderMale:JobGrade5 -2.378e+00  2.661e+00  -0.894 0.372678
## GenderMale:JobGrade6  4.540e+01  5.224e+00   8.691 1.66e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.779 on 190 degrees of freedom
## Multiple R-squared:  0.8345,Adjusted R-squared:  0.8197
## F-statistic: 56.36 on 17 and 190 DF,  p-value: < 2.2e-16
## Analysis of Variance Table
##
## Model 1: Salary ~ Exp + Gender + PCJob + EducLev + JobGrade
## Model 2: Salary ~ Exp + Gender + PCJob + EducLev + JobGrade + Gender:JobGrade
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    195 6203.1
## 2    190 4340.2  5    1862.8 16.309 2.238e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Gender
Male
Female
mean of Salary
30
40
50
60
70
1
2
3
4
5
6
JobGrade


Bank Data
Frequency
0
20
40
60
80
−4
−2
0
2
4
Standadised residuals


Standardised Residuals
−4
−2
0
2
4
F1
M1
F2
M2
F3
M3
F4
M4
F5
M5
F6
M6
Gender : JobGrade


Bank Data: Std Res against Fitted
Standadised residuals
−4
−2
0
2
4
30
40
50
60
70
80
Fitted values


Normal Probability plot
Standardized Residuals
−4
−2
0
2
4
−3
−2
−1
0
1
2
3
Normal Scores

What did we discover?

There is no gender bias in salary, **BUT** a gender bias in *promotions* does exist in the bank. The Bank was reprimanded for its promotion regime and instructed to address this issue.

Correct analysis of data reveals the truth!

Finishing Off:

When you've finished, close down R by typing `q()`. Choose 'Save' when prompted as to whether you want to retain your workspace. Remember to log off from your computer before leaving.