Solution to Series 4

1. a)
```
> farm <- read.table("http://stat.ethz.ch/Teaching/Datasets/farm.dat",header=TRUE)
> fit <- lm(Dollar~cows, data=farm)
> summary(fit)
```

```
Call:
lm(formula = Dollar ~ cows, data = farm)

Residuals:
    Min      1Q  Median      3Q     Max
-204.68  -80.02   15.48   54.57  284.43

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  694.019     50.039  13.869 4.75e-11 ***
cows          20.111      4.725   4.256 0.000475 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 122.9 on 18 degrees of freedom
Multiple R-squared:  0.5016,        Adjusted R-squared:  0.4739
F-statistic: 18.11 on 1 and 18 DF,  p-value: 0.0004751
```

There is a significant dependence (e.g. on the 5% level) between income and number of cows, since the p-value of the regression coefficient is very small (0.000475).

b)
```
> predict(fit, newdata=data.frame(cows=c(0,20,8.85)), interval="confidence")
```

```
        fit      lwr       upr
1  694.0189 588.8902  799.1476
2 1096.2361 971.3953 1221.0768
3  872.0000 814.2627  929.7373
```

```
> predict(fit, newdata=data.frame(cows=c(0,8.85)), interval="prediction")
```

```
       fit      lwr       upr
1 694.0189 415.2286  972.8092
2 872.0000 607.4143 1136.5857
```

c) We first try to explain I with A:

```
> fit1 <- lm(Dollar~acres, data=farm)
> summary(fit1)
```

```
Call:
lm(formula = Dollar ~ acres, data = farm)

Residuals:
    Min      1Q  Median      3Q     Max
-281.54 -113.94  -28.18   94.28  387.05

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 868.7363   105.9796   8.197 1.73e-07 ***
acres         0.0234     0.7066   0.033    0.974
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

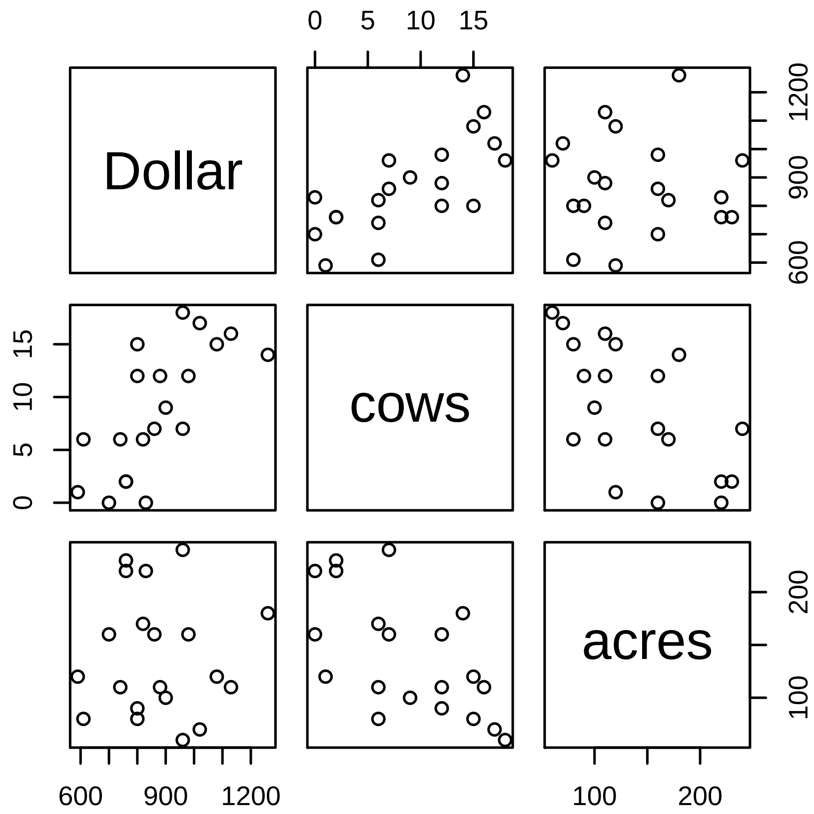
```
Residual standard error: 174.1 on 18 degrees of freedom
Multiple R-squared:  6.09e-05,        Adjusted R-squared:  -0.05549
F-statistic: 0.001096 on 1 and 18 DF,  p-value: 0.974
```

There seems to be no significant dependence. However, if we add C as a covariate, both variables are significant!

```
> fit2 <- lm(Dollar~acres+cows, data=farm)
> summary(fit2)
Call:
lm(formula = Dollar ~ acres + cows, data = farm)

Residuals:
     Min       1Q   Median       3Q      Max
-145.064  -46.719   -9.992   55.149  133.664

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 285.4572    81.3793   3.508   0.0027 **
acres         2.1384     0.3936   5.434 4.47e-05 ***
cows         32.5690     3.7276   8.737 1.08e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 76.45 on 17 degrees of freedom
Multiple R-squared:  0.8179,        Adjusted R-squared:  0.7965
F-statistic: 38.17 on 2 and 17 DF,  p-value: 5.165e-07
```

d) Using `pairs(farm)` we can observe the pairwise relationships between variables.

```
> pairs(farm)
```



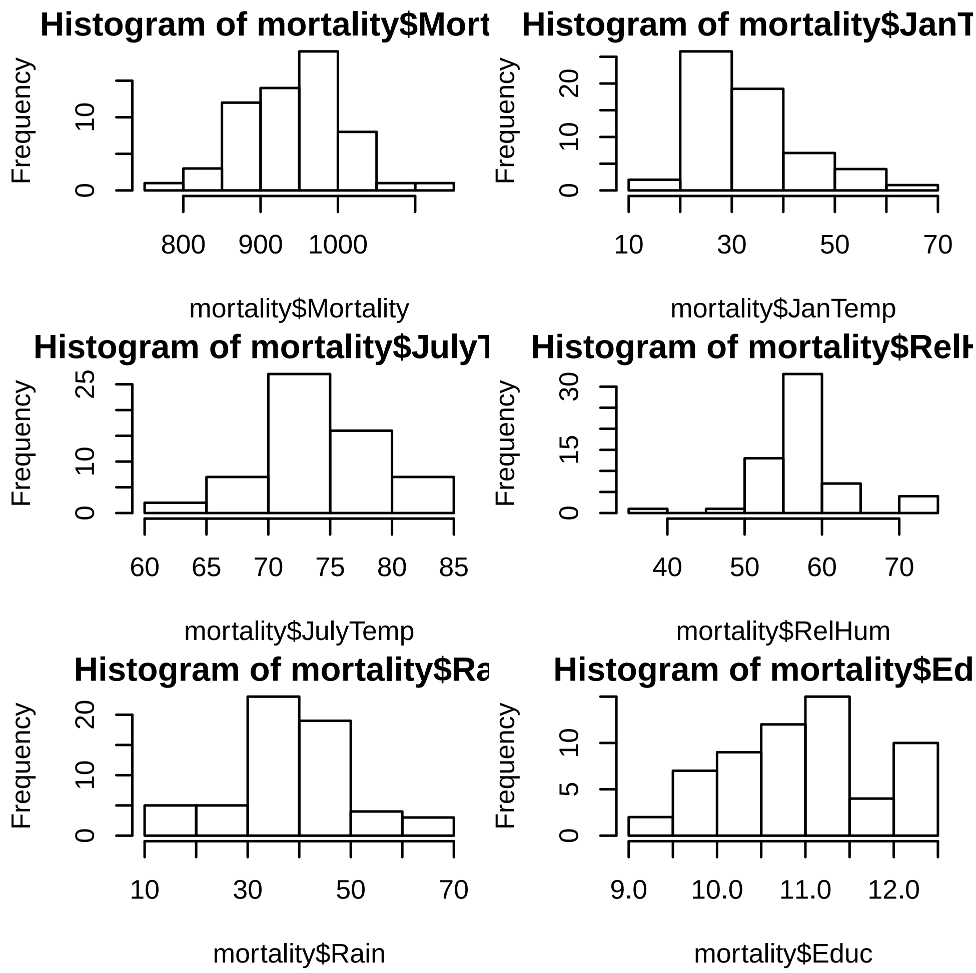
From the pairs() plot we can observe a positive correlation between I and C and a negative correlation between C and A. However, there seems to be no correlation between I and A. We can only identify the influence of farm size on income if we control for the number of cows, i.e. comparing like with like. In colloquial terms, the positive correlation of I and C and the negative correlation of C and A cancel each other out. Thus the variable A is not considered significant in a univariate regression of I and A.

2. a)

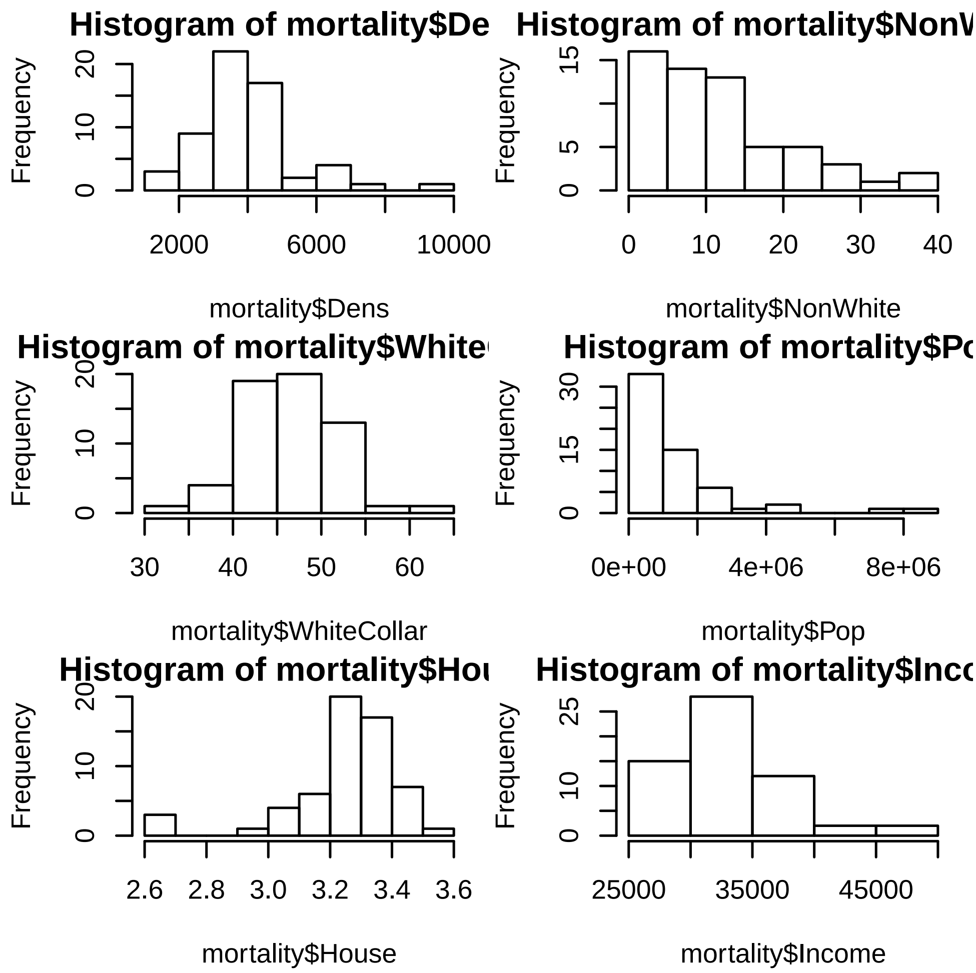
```
> mortality <- read.csv("http://stat.ethz.ch/Teaching/Datasets/mortality.csv",
                        header=TRUE)
> str(mortality)
'data.frame':        59 obs. of  16 variables:
 $ City       : Factor w/ 59 levels "Akron, OH","Albany-Schenectady-Troy, NY",..: 1 2 3 4 5 6
 $ Mortality  : num  922 998 962 982 1071 ...
 $ JanTemp    : int  27 23 29 45 35 45 30 30 24 27 ...
 $ JulyTemp   : int  71 72 74 79 77 80 74 73 70 72 ...
 $ RelHum     : int  59 57 54 56 55 54 56 56 61 59 ...
 $ Rain       : int  36 35 44 47 43 53 43 45 36 36 ...
 $ Educ       : num  11.4 11 9.8 11.1 9.6 10.2 12.1 10.6 10.5 10.7 ...
 $ Dens       : int  3243 4281 4260 3125 6441 3325 4679 2140 6582 4213 ...
 $ NonWhite   : num  8.8 3.5 0.8 27.1 24.4 38.5 3.5 5.3 8.1 6.7 ...
 $ WhiteCollar: num  42.6 50.7 39.4 50.2 43.7 43.1 49.2 40.4 42.5 41 ...
 $ Pop        : int  660328 835880 635481 2138231 2199531 883946 2805911 438557 1015472 404421
 $ House      : num  3.34 3.14 3.21 3.41 3.44 3.45 3.23 3.29 3.31 3.36 ...
 $ Income     : int  29560 31458 31856 32452 32368 27835 36644 47258 31248 29089 ...
 $ HC         : int  21 8 6 18 43 30 21 6 18 12 ...
 $ NOx        : int  15 10 6 8 38 32 32 4 12 7 ...
 $ SO2        : int  59 39 33 24 206 72 62 4 37 20 ...
> rownames(mortality) <- mortality$City
> mortality <- mortality[,-1]
```

We set the city as row names and look at the histograms of the other variables to determine whether they require transformations:

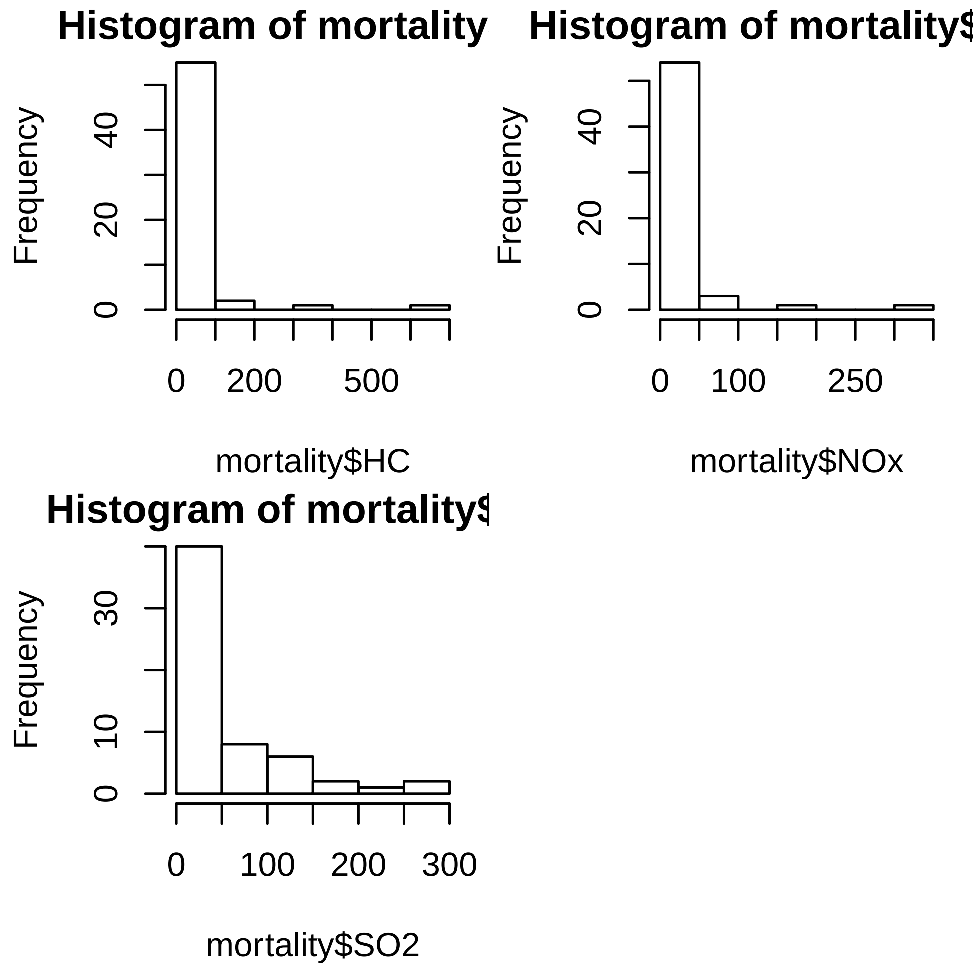
```
> par(mfrow=c(3,2))
> hist(mortality$Mortality)   ## ok, no transformation
> hist(mortality$JanTemp)     ## right-skewed, log transformation recommendable
> hist(mortality$JulyTemp)    ## ok, no transformation
> hist(mortality$RelHum)      ## ok, no transformation
> hist(mortality$Rain)        ## ok, no transformation
> hist(mortality$Educ)        ## ok, no transformation
```




```
> par(mfrow=c(3,2))
> hist(mortality$Dens)        ## right skewed, log-tranformation recommendable
> hist(mortality$NonWhite)    ## percentage, arcsin-transformation recommendable
> hist(mortality$WhiteCollar) ## percentage, arcsin-transformation recommendable
> hist(mortality$Pop)         ## right skewed, log-tranformation recommendable
> hist(mortality$House)       ## ok, no transformation
> hist(mortality$Income)      ## right skewed, log-tranformation recommendable
```




```
> par(mfrow=c(2,2))
> hist(mortality$HC)          ## strongly right skewed, log-tranformation mandatory
> hist(mortality$NOx)         ## strongly right skewed, log-tranformation mandatory
> hist(mortality$SO2)         ## strongly right skewed, log-tranformation mandatory
```



We transform the following variables:

```
> mortality$JanTemp     <- log(mortality$JanTemp)
> mortality$Dens        <- log(mortality$Dens)
> mortality$NonWhite    <- asin(sqrt(mortality$NonWhite/100))
> mortality$WhiteCollar <- asin(sqrt(mortality$WhiteCollar/100))
> mortality$Pop         <- log(mortality$Pop)
> mortality$Income      <- log(mortality$Income)
> mortality$HC          <- log(mortality$HC)
> mortality$NOx         <- log(mortality$NOx)
> mortality$SO2         <- log(mortality$SO2)
```

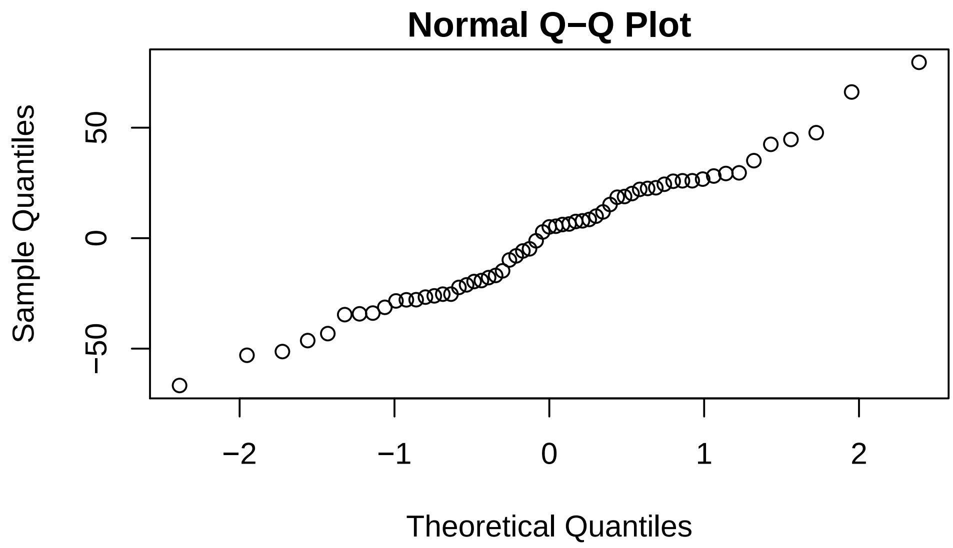
b) Full model:

```
> fit <- lm(Mortality ~ ., data=mortality)
> summary(fit)
```

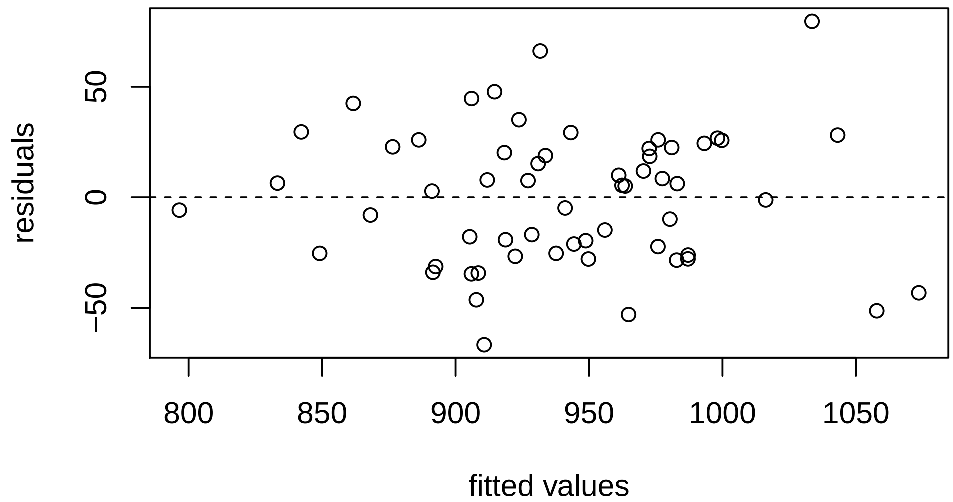
```
Call:
lm(formula = Mortality ~ ., data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-66.668 -25.338   5.108  22.670  79.594

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1514.05643  592.42867   2.556  0.01413 *
JanTemp      -65.90878   27.23547  -2.420  0.01972 *
JulyTemp      -2.18908    2.06935  -1.058  0.29589
RelHum         0.04771    1.08381   0.044  0.96509
Rain           1.70646    0.58318   2.926  0.00541 **
Educ         -12.26491    8.87953  -1.381  0.17417
```

```
Dens           16.05653   16.29979   0.985  0.32997
NonWhite      321.61186   64.66123   4.974 1.05e-05 ***
WhiteCollar -154.16478  114.47231  -1.347  0.18496
Pop             2.34899    7.79886   0.301  0.76468
House         -28.18972   37.85883  -0.745  0.46047
Income        -17.90976   48.47305  -0.369  0.71354
HC            -23.84947   15.27338  -1.562  0.12557
NOx            34.00128   14.51624   2.342  0.02375 *
SO2            -1.35604    6.90926  -0.196  0.84531
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.86 on 44 degrees of freedom
Multiple R-squared:  0.7634,        Adjusted R-squared:  0.6881
F-statistic: 10.14 on 14 and 44 DF,  p-value: 1.373e-09
```

```
> qqnorm(fit$resid)
```



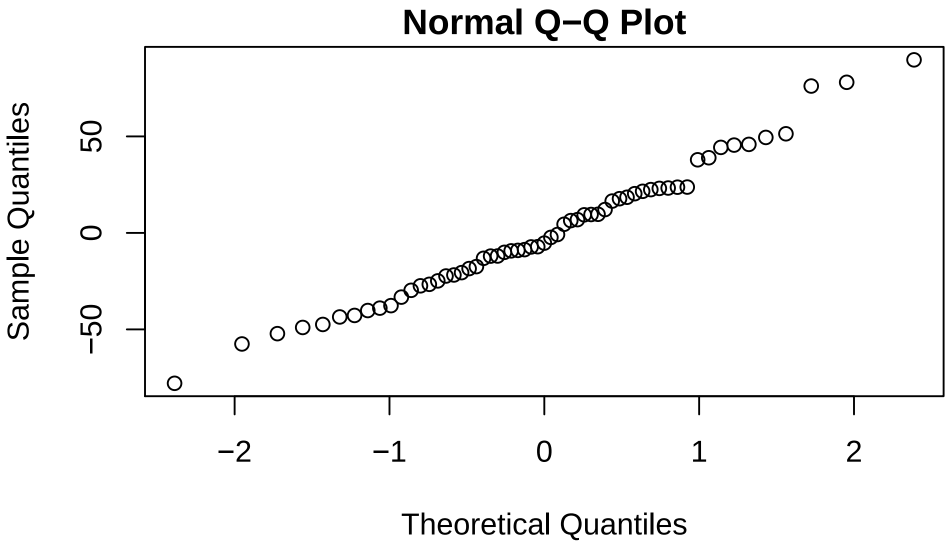

```
> plot(fit$fitted,fit$resid,xlab="fitted values",ylab="residuals")
> abline(h=0,lty=2)
```



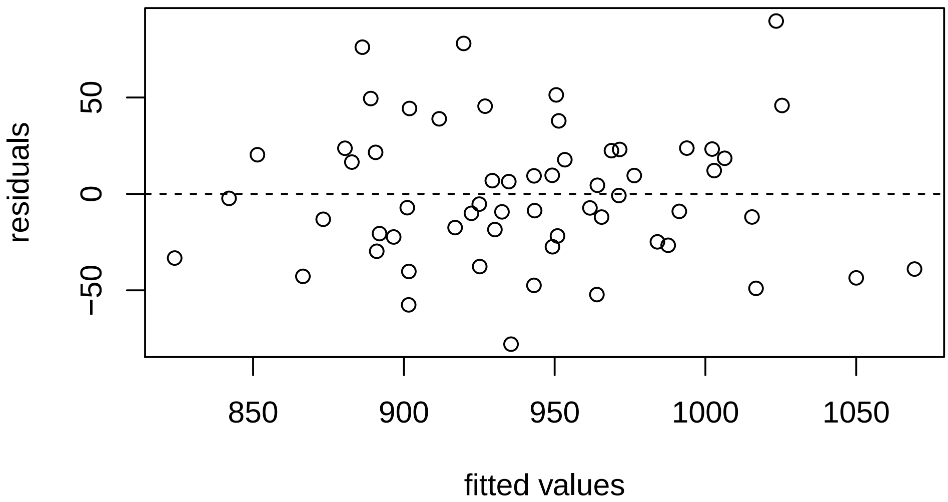
Even though most of the predictors seem to have no significant effect on the response, the model fits quite well. We do not see any violation of the model assumptions.

c) Now we just use the significant variables:

```
> fit2 <- lm(Mortality ~ JanTemp + Rain + NonWhite + NOx, data=mortality)
> summary(fit2)
Call:
lm(formula = Mortality ~ JanTemp + Rain + NonWhite + NOx, data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-77.919 -23.592  -5.281  22.011  89.691

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 980.8357    62.7178  15.639  < 2e-16 ***
JanTemp     -79.8471    18.8162  -4.244 8.70e-05 ***
Rain          2.5434     0.4822   5.275 2.40e-06 ***
NonWhite    276.2770    42.5363   6.495 2.72e-08 ***
NOx          20.9886     4.6856   4.479 3.92e-05 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.32 on 54 degrees of freedom
Multiple R-squared:  0.6847,       Adjusted R-squared:  0.6614
F-statistic: 29.32 on 4 and 54 DF,  p-value: 5.674e-13
> qqnorm(fit2$resid)
```




```
> plot(fit2$fitted,fit2$resid,xlab="fitted values",ylab="residuals")
> abline(h=0,lty=2)
```



Now all the variables are highly significant. As expected with fewer variables, the residuals are a little bigger now and R^2 decreased slightly. However, the difference in adjusted R^2 is very small, indicating that we have not lost much explanatory power.

Even though leaving out all of the non-significant variable at once worked quite well here, this is not a good strategy in general. If the predictors are not mutually independent, leaving out one can have a huge effect on the significance of the others. A better way of pruning the model thus is to leave out predictors step by step, one at a time.

d)

```
> fit.reduc <- fit
> fit.reduc <- update(fit.reduc, ~.-RelHum)     ; summary(fit.reduc)
Call:
lm(formula = Mortality ~ JanTemp + JulyTemp + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
    SO2, data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-66.738 -25.325   5.229  22.785  79.521

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1522.5940   553.5340   2.751  0.00854 **
JanTemp      -66.0256    26.8036  -2.463  0.01766 *
JulyTemp      -2.2342     1.7771  -1.257  0.21516
Rain           1.7110     0.5678   3.014  0.00423 **
Educ         -12.2876     8.7657  -1.402  0.16784
Dens          16.0014    16.0704   0.996  0.32472
NonWhite     322.3336    61.8501   5.212 4.53e-06 ***
WhiteCollar -154.1022   113.1870  -1.361  0.18014
Pop            2.3599     7.7080   0.306  0.76089
House        -28.3888    37.1684  -0.764  0.44898
Income       -18.0148    47.8743  -0.376  0.70847
HC           -23.8440    15.1026  -1.579  0.12138
NOx           34.0558    14.3021   2.381  0.02155 *
SO2           -1.4567     6.4474  -0.226  0.82228
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.47 on 45 degrees of freedom
Multiple R-squared:  0.7634,        Adjusted R-squared:  0.695
F-statistic: 11.17 on 13 and 45 DF,  p-value: 3.976e-10
```

```
> fit.reduc <- update(fit.reduc, ~.-SO2)        ; summary(fit.reduc)
Call:
lm(formula = Mortality ~ JanTemp + JulyTemp + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx,
    data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-67.414 -24.501   3.764  22.349  84.136

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1476.3654   508.9942   2.901  0.00570 **
JanTemp      -62.6563    22.0407  -2.843  0.00665 **
JulyTemp      -2.1685     1.7349  -1.250  0.21766
Rain           1.6932     0.5565   3.043  0.00387 **
Educ         -11.7713     8.3749  -1.406  0.16658
Dens          15.3827    15.6712   0.982  0.33143
NonWhite     319.5287    59.9631   5.329 2.89e-06 ***
WhiteCollar -155.2406   111.9024  -1.387  0.17204
Pop            2.1424     7.5683   0.283  0.77839
House        -26.6033    35.9420  -0.740  0.46296
Income       -15.4399    46.0158  -0.336  0.73875
HC           -23.8494    14.9459  -1.596  0.11740
NOx           32.8564    13.1427   2.500  0.01605 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.12 on 46 degrees of freedom
Multiple R-squared:  0.7631,        Adjusted R-squared:  0.7013
F-statistic: 12.35 on 12 and 46 DF,  p-value: 1.119e-10
> fit.reduc <- update(fit.reduc, ~.-Pop)        ; summary(fit.reduc)
Call:
lm(formula = Mortality ~ JanTemp + JulyTemp + Rain + Educ + Dens +
    NonWhite + WhiteCollar + House + Income + HC + NOx, data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-68.002 -25.180   3.806  23.184  84.056

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1464.677    502.328   2.916  0.00542 **
JanTemp      -63.036     21.784  -2.894  0.00575 **
JulyTemp      -2.074      1.686  -1.230  0.22471
Rain           1.677      0.548   3.060  0.00365 **
Educ         -11.567      8.262  -1.400  0.16806
Dens          15.518     15.510   1.000  0.32219
NonWhite     321.751     58.862   5.466 1.71e-06 ***
WhiteCollar -154.170    110.739  -1.392  0.17042
House        -28.564     34.922  -0.818  0.41752
Income       -11.935     43.883  -0.272  0.78683
HC           -24.039     14.784  -1.626  0.11063
NOx           33.618     12.738   2.639  0.01124 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 33.78 on 47 degrees of freedom
Multiple R-squared:  0.7627,	Adjusted R-squared:  0.7071
F-statistic: 13.73 on 11 and 47 DF,  p-value: 3.024e-11
```

> fit.reduc <- update(fit.reduc, ~.-Income)   ; summary(fit.reduc)

```
Call:
lm(formula = Mortality ~ JanTemp + JulyTemp + Rain + Educ + Dens +
    NonWhite + WhiteCollar + House + HC + NOx, data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-68.184 -25.120   4.127  22.528  83.274

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1351.8460   280.5051   4.819 1.49e-05 ***
JanTemp      -63.7347    21.4218  -2.975  0.00457 **
JulyTemp      -2.0778     1.6695  -1.245  0.21934
Rain           1.6935     0.5392   3.141  0.00288 **
Educ         -12.2927     7.7434  -1.588  0.11896
Dens          15.5653    15.3586   1.013  0.31592
NonWhite     322.5924    58.2112   5.542 1.25e-06 ***
WhiteCollar -157.8965   108.8227  -1.451  0.15330
House        -28.2564    34.5651  -0.817  0.41769
HC           -23.6377    14.5676  -1.623  0.11122
NOx           33.0513    12.4445   2.656  0.01070 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.45 on 48 degrees of freedom
Multiple R-squared:  0.7623,	Adjusted R-squared:  0.7128
F-statistic: 15.39 on 10 and 48 DF,  p-value: 7.686e-12
```

> fit.reduc <- update(fit.reduc, ~.-House)   ; summary(fit.reduc)

```
Call:
lm(formula = Mortality ~ JanTemp + JulyTemp + Rain + Educ + Dens +
    NonWhite + WhiteCollar + HC + NOx, data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-72.137 -25.144   4.209  24.152  83.480

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1176.7896   180.5674   6.517 3.71e-08 ***
JanTemp      -55.2844    18.6991  -2.957  0.00477 **
JulyTemp      -1.9777     1.6593  -1.192  0.23906
Rain           1.7423     0.5341   3.262  0.00202 **
Educ         -10.4655     7.3886  -1.416  0.16298
Dens          18.9748    14.7313   1.288  0.20378
NonWhite     299.6942    50.8559   5.893 3.42e-07 ***
WhiteCollar -156.1713   108.4334  -1.440  0.15616
HC           -21.5406    14.2914  -1.507  0.13817
NOx           31.7474    12.3000   2.581  0.01289 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 33.34 on 49 degrees of freedom
Multiple R-squared:  0.759,	Adjusted R-squared:  0.7147
F-statistic: 17.15 on 9 and 49 DF,  p-value: 2.444e-12
```

> fit.reduc <- update(fit.reduc, ~.-JulyTemp)   ; summary(fit.reduc)

```
Call:
lm(formula = Mortality ~ JanTemp + Rain + Educ + Dens + NonWhite +
    WhiteCollar + HC + NOx, data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-74.697 -26.160   0.063  20.863  83.863

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1056.2316   150.2029   7.032 5.35e-09 ***
JanTemp      -60.2590    18.3038  -3.292  0.00183 **
Rain           1.7576     0.5361   3.278  0.00190 **
Educ          -9.3189     7.3565  -1.267  0.21111
Dens          18.3262    14.7830   1.240  0.22088
NonWhite     261.7294    39.8105   6.574 2.78e-08 ***
WhiteCollar -180.9759   106.8639  -1.694  0.09658 .
HC           -14.3194    12.9978  -1.102  0.27588
NOx           29.0735    12.1444   2.394  0.02046 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.48 on 50 degrees of freedom
Multiple R-squared:  0.752,	Adjusted R-squared:  0.7123
F-statistic: 18.95 on 8 and 50 DF,  p-value: 1.05e-12
```

> fit.reduc <- update(fit.reduc, ~.-HC)   ; summary(fit.reduc)

```
Call:
lm(formula = Mortality ~ JanTemp + Rain + Educ + Dens + NonWhite +
    WhiteCollar + NOx, data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-76.495 -25.543   4.253  19.846  84.672

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1067.5033   150.1677   7.109 3.66e-09 ***
JanTemp      -64.0371    18.0173  -3.554 0.000828 ***
Rain           1.8825     0.5251   3.585 0.000754 ***
Educ         -11.1702     7.1770  -1.556 0.125799
Dens          18.7825    14.8081   1.268 0.210418
NonWhite     264.7197    39.8010   6.651 1.94e-08 ***
WhiteCollar -179.4981   107.0791  -1.676 0.099797 .
NOx           16.8616     4.9716   3.392 0.001350 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.55 on 51 degrees of freedom
Multiple R-squared:  0.746,	Adjusted R-squared:  0.7111
F-statistic:  21.4 on 7 and 51 DF,  p-value: 3.851e-13
```

> fit.reduc <- update(fit.reduc, ~.-Dens)   ; summary(fit.reduc)

```
Call:
lm(formula = Mortality ~ JanTemp + Rain + Educ + NonWhite + WhiteCollar +
    NOx, data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-80.854 -26.449   3.159  18.654  84.961

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1217.1646    93.4291  13.028  < 2e-16 ***
JanTemp      -66.8959    17.9801  -3.721 0.000489 ***
Rain           1.9731     0.5233   3.771 0.000418 ***
Educ         -13.1443     7.0471  -1.865 0.067797 .
NonWhite     261.3019    39.9414   6.542 2.66e-08 ***
WhiteCollar -142.8799   103.7157  -1.378 0.174224
NOx           19.5735     4.5146   4.336 6.69e-05 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.74 on 52 degrees of freedom
Multiple R-squared:  0.738,        Adjusted R-squared:  0.7078
F-statistic: 24.41 on 6 and 52 DF,  p-value: 1.59e-13
```

```
> fit.reduc <- update(fit.reduc, ~.-WhiteCollar); summary(fit.reduc)
```

```
Call:
lm(formula = Mortality ~ JanTemp + Rain + Educ + NonWhite + NOx,
    data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-82.794 -25.435   6.366  20.410  77.977

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1183.4856    90.9344  13.015  < 2e-16 ***
JanTemp      -70.9168    17.8912  -3.964 0.000222 ***
Rain           1.8185     0.5154   3.528 0.000874 ***
Educ         -17.9858     6.1597  -2.920 0.005131 **
NonWhite     268.4084    39.9410   6.720 1.27e-08 ***
NOx           18.4360     4.4759   4.119 0.000134 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.03 on 53 degrees of freedom
Multiple R-squared:  0.7284,        Adjusted R-squared:  0.7028
F-statistic: 28.43 on 5 and 53 DF,  p-value: 6.945e-14
```

Now we stop because all of the remaining variables are significant. We now see that in part c) we missed out one significant variable (Educ).

e) Fitting the model without the meteo-variables:

```
> fit.without.meteo <- lm(Mortality ~ .-JanTemp-JulyTemp-RelHum-Rain, data=mortality)
> anova(fit, fit.without.meteo)
```

```
Analysis of Variance Table

Model 1: Mortality ~ JanTemp + JulyTemp + RelHum + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
    SO2
```

```
Model 2: Mortality ~ (JanTemp + JulyTemp + RelHum + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
    SO2) - JanTemp - JulyTemp - RelHum - Rain
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     44 53474
2     48 71705 -4    -18230 3.7501 0.01038 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With the function anova() one carries out an F-test in order to compare two models. In this case, the null-hypothesis gets rejected on the 5% level. That is, the bigger model (the one with the meteo-variables) is significantly better.

Fitting the model without the air pollution-variables:

```
> fit.without.air <- lm(Mortality ~ .-HC-NOx-SO2, data=mortality)
> anova(fit, fit.without.air)
Analysis of Variance Table

Model 1: Mortality ~ JanTemp + JulyTemp + RelHum + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
    SO2
Model 2: Mortality ~ (JanTemp + JulyTemp + RelHum + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
    SO2) - HC - NOx - SO2
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     44 53474
2     47 62715 -3   -9240.3 2.5344 0.06905 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, the partial F-test is not significant on the 5% level, however, only slightly so. This seems to contradict the fact that NOx is a significant predictor, as seen from our analysis in part d). The thing to note is that the F-test only compares two models, i.e. in this case the full model and the full model minus *all* pollution variables. In this context, we do not seem to lose much by throwing away those variables, *if we keep all the others in the model* (possibly because there is another variable correlated with NOx).

Fitting the model without the demographic-variables:

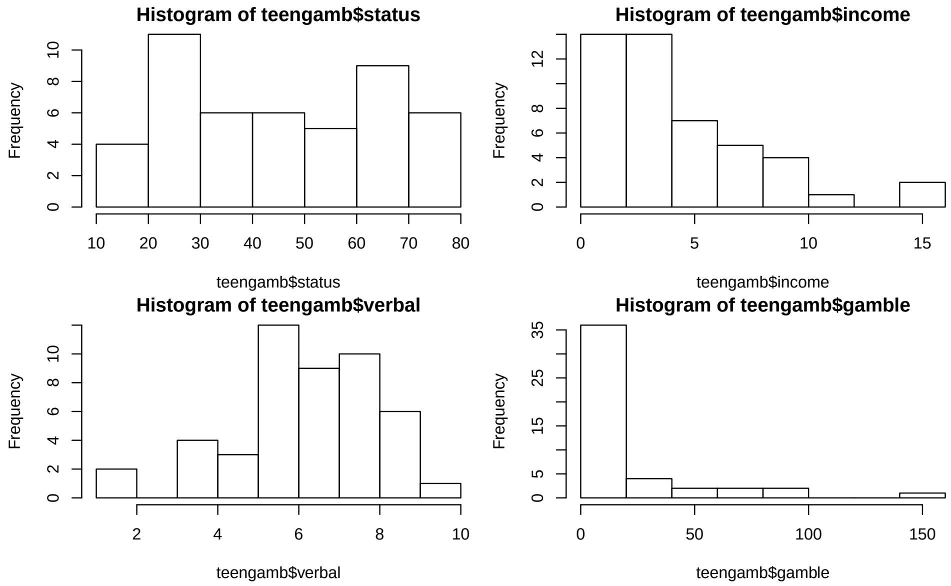
```
> fit.without.demographic <- lm(Mortality ~ .-Educ-Dens-NonWhite-WhiteCollar-Pop-House
                           -Income, data=mortality)
> anova(fit, fit.without.demographic)
Analysis of Variance Table

Model 1: Mortality ~ JanTemp + JulyTemp + RelHum + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
    SO2
Model 2: Mortality ~ (JanTemp + JulyTemp + RelHum + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
    SO2) - Educ - Dens - NonWhite - WhiteCollar - Pop - House -
    Income
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     44  53474
2     51 103411 -7    -49936 5.8698 7.524e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, the null hypothesis gets rejected, that is we cannot leave out the demographic-variables.

3. **a)**

```
> ## Load data
> file <- url("http://stat.ethz.ch/education/semesters/as2011/asr/teengamb.rda")
```

```
> load(file)
> ## Histograms
> par(mfrow=c(2,2))
> hist(teengamb$status)
> hist(teengamb$income)
> hist(teengamb$verbal)
> hist(teengamb$gamble)
```



The histograms of `income` and `gamble` show skewed distributions. Therefore, we perform a log transformation. Due to the fact that 4 data points of `gamble` are zero, we need to add a constant (here: 0.1) prior to transformation.

```
> ## Transformations
> any(teengamb$income==0)   # log trsf directly possible
[1] FALSE
> any(teengamb$gamble==0)   # any zeros?
[1] TRUE
> teengamb$log.income <- log(teengamb$income)
> teengamb$log.gamble <- log(teengamb$gamble+0.1)
```

b)

```
> ## Choose correct data type for sex
> teengamb$sex <- factor(teengamb$sex, labels=c("male", "female"))
```

c) After having transformed `gamble` and `income`, we fit a linear regression model to the data.

```
> fit.trsf <- lm(log.gamble ~ sex + status + log.income + verbal, data=teengamb)
> summary(fit.trsf)
Call:
lm(formula = log.gamble ~ sex + status + log.income + verbal,
    data = teengamb)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1889 -1.1400  0.2745  1.1436  2.8771

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.49053    1.27810   1.166  0.25011
sexfemale   -1.50261    0.58908  -2.551  0.01448 *
status       0.03705    0.02030   1.825  0.07510 .
log.income   1.13326    0.35438   3.198  0.00263 **
verbal      -0.38478    0.16046  -2.398  0.02101 *
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.677 on 42 degrees of freedom
Multiple R-squared:  0.4338,	Adjusted R-squared:  0.3799
F-statistic: 8.046 on 4 and 42 DF,  p-value: 6.554e-05
```

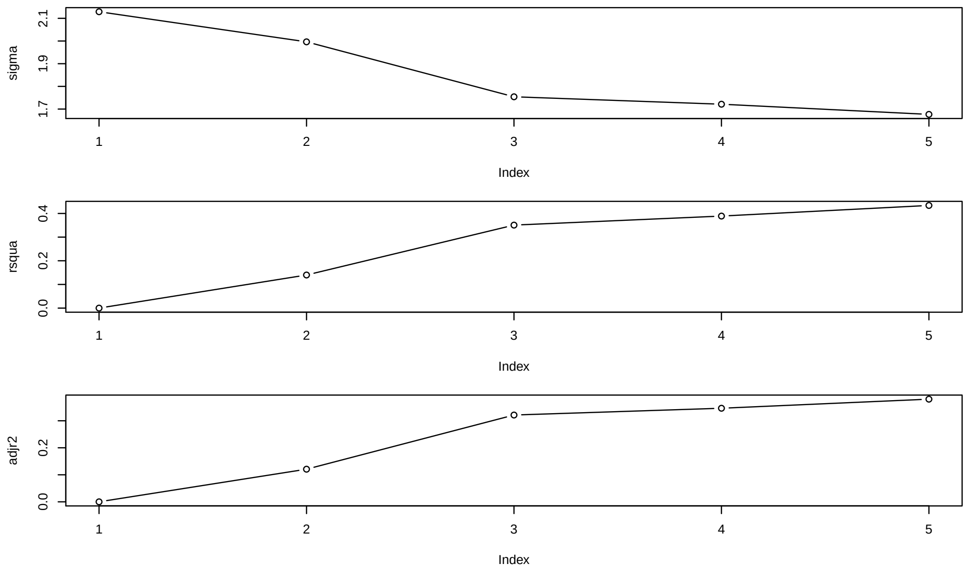
d)

```
> coeftr <- coef(fit.trsf)
> coeftr["sexfemale"]
sexfemale
-1.502611
> conf <- confint(fit.trsf)
> conf
                   2.5 %      97.5 %
(Intercept) -1.088767239  4.06983303
sexfemale   -2.691424093 -0.31379839
status      -0.003917605  0.07801884
log.income   0.418090989  1.84842855
verbal      -0.708604704 -0.06095770
```

The predicted (log) gambling expenses decrease by -1.5 when looking at female gamblers instead of males. The 95% confidence interval [-2.69,-0.31] suggests that this decrease is significant.

e) The more predictors we add the lower the standard deviation of the residuals but the higher the R^2 and adjusted R^2. This means that we can explain more and more variance in the response by adding these predictors.

```
> fit   <- lm(log.gamble ~ 1, data=teengamb)
> sigma <- summary(fit)$sigma
> rsqua <- summary(fit)$r.squared
> adjr2 <- summary(fit)$adj.r.squared
> fit   <- lm(log.gamble ~ log.income, data=teengamb)
> sigma <- c(sigma, summary(fit)$sigma)
> rsqua <- c(rsqua, summary(fit)$r.squared)
> adjr2 <- c(adjr2, summary(fit)$adj.r.squared)
> fit   <- lm(log.gamble ~ log.income + sex, data=teengamb)
> sigma <- c(sigma, summary(fit)$sigma)
> rsqua <- c(rsqua, summary(fit)$r.squared)
> adjr2 <- c(adjr2, summary(fit)$adj.r.squared)
> fit   <- lm(log.gamble ~ log.income + sex + verbal, data=teengamb)
> sigma <- c(sigma, summary(fit)$sigma)
> rsqua <- c(rsqua, summary(fit)$r.squared)
> adjr2 <- c(adjr2, summary(fit)$adj.r.squared)
> fit   <- lm(log.gamble ~ log.income + sex + verbal + status, data=teengamb)
> sigma <- c(sigma, summary(fit)$sigma)
> rsqua <- c(rsqua, summary(fit)$r.squared)
> adjr2 <- c(adjr2, summary(fit)$adj.r.squared)
```



4. **a)** True.

b) False. The log transform is recommended when a predictor has a right skewed distribution. A way of dealing with a left skewed predictor variable x might be to take $x' = (-1)x$ and then try to apply the log transformation to x'. However, this will not always be possible, but true left-skeness occurs very rarely in practice.

c) True.

d) True.

e) False. Multiple R^2 will get larger whenever you add predictors to a model, so you will always end up choosing the biggest model. This can lead to the inclusion of non significant predictors in the model and over-fitting.