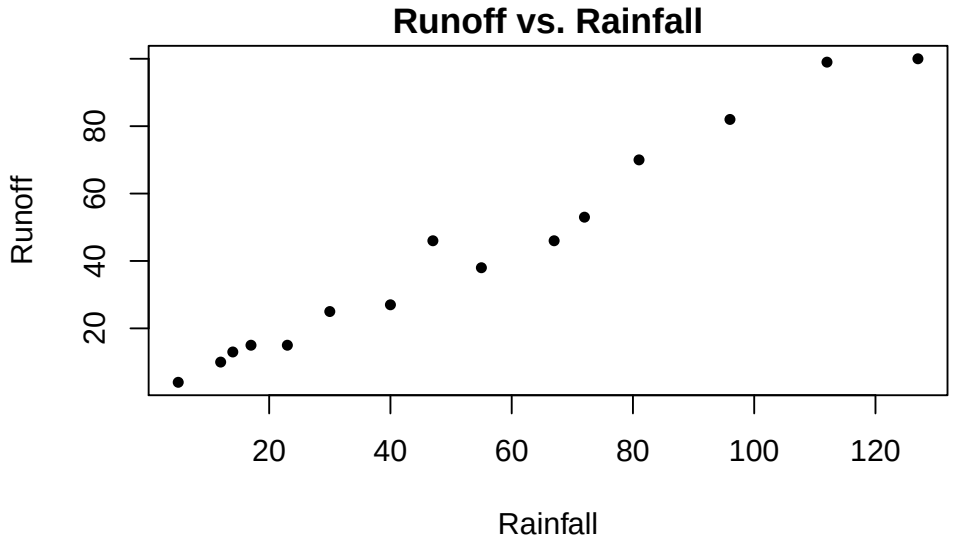
Solution to Series 3

1. **a)** First we type in the data. The scatterplot of `runoff` versus `rainfall` suggests that a linear relationship holds. Therefore, one would guess that the R^2 should be large, i.e. close to 1.

```
> rainfall <- c(5, 12, 14, 17, 23, 30, 40, 47, 55, 67, 72, 81, 96, 112, 127)
> runoff <- c(4, 10, 13, 15, 15, 25, 27, 46, 38, 46, 53, 70, 82,  99, 100)
> data <- data.frame(rainfall=rainfall, runoff=runoff)
> plot(data$runoff ~ data$rainfall, pch=20, xlab="Rainfall", ylab="Runoff",
      main="Runoff vs. Rainfall")
```



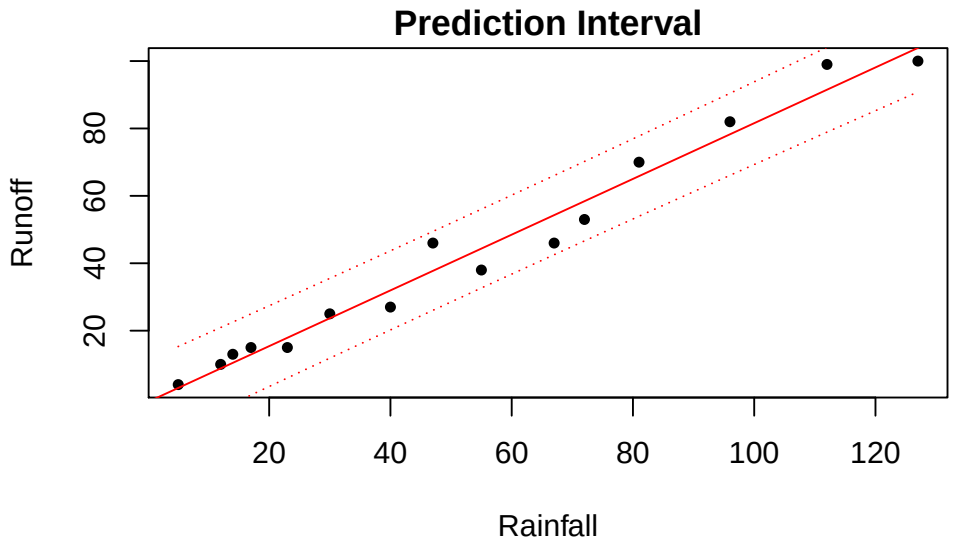
b) We fit a linear model with `runoff` as response and `rainfall` as predictor. We are then able to use this model for prediction.

```
> fit <- lm(runoff ~ rainfall, data=data)
> pred <- predict(fit, newdata=data.frame(rainfall=50), interval="prediction")
```

If the rainfall volume takes a value of 50 we find a runoff volume of 40.22 with a 95% prediction interval of [28.53,51.92].

We can also draw the regression line and the 95% prediction interval to the data.

```
> plot(data$runoff ~ data$rainfall, pch=20, xlab="Rainfall", ylab="Runoff",
      main="Prediction Interval")
> abline(fit, col="red")
> interval <- predict(fit, interval="prediction")
> lines(data$rainfall, interval[,2], lty=3, col="red")
> lines(data$rainfall, interval[,3], lty=3, col="red")
```

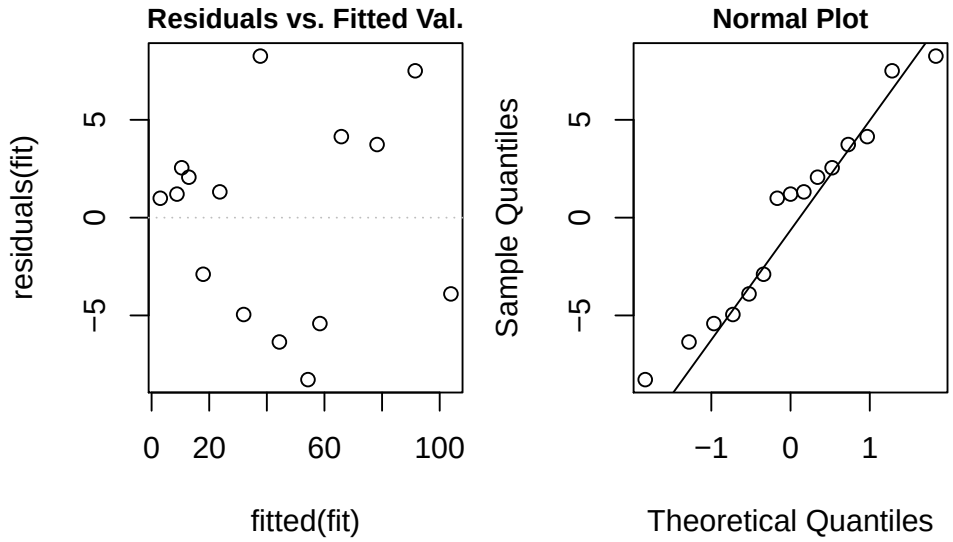


c) An R^2 of 0.98 is extremely high, i.e. a huge part of the variation in the data can be attributed to the linear association between runoff and rainfall volume.

d)
```
> summary(fit)

Call:
lm(formula = runoff ~ rainfall, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-8.279 -4.424  1.205  3.145  8.261

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.12830    2.36778  -0.477    0.642
rainfall     0.82697    0.03652  22.642  7.9e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.24 on 13 degrees of freedom
Multiple R-squared: 0.9753,        Adjusted R-squared: 0.9734
F-statistic: 512.7 on 1 and 13 DF,  p-value: 7.896e-12

> ## Confidence intervals for the coefficients
> confint(fit)
                 2.5 %    97.5 %
(Intercept) -6.2435879 3.9869783
rainfall     0.7480677 0.9058786
```

There is a significant linear association between runoff and rainfall volume, since the null hypothesis $\beta_1 = 0$ is clearly rejected. However, the confidence interval for β_1 does not contain $\beta_1 = 1$, i.e. a null hypothesis of $\beta_1 = 1$ would be rejected, too. Therefore, we conclude that no $1:1$ relation between rainfall and runoff holds. We suspect that part of the rain evaporates or trickles away.

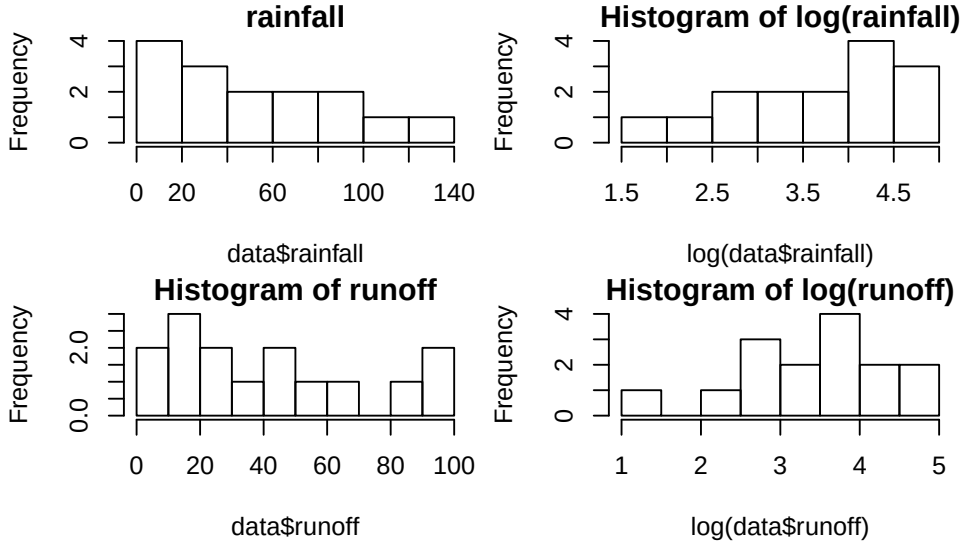
e)
```
> par(mfrow=c(1,2))
> plot(fitted(fit), residuals(fit), main="Residuals vs. Fitted Val.", cex.main=0.9)
> abline(h=0, col="grey", lty=3)
> qqnorm(residuals(fit), main="Normal Plot", cex.main=0.9)
> qqline(residuals(fit))
```



From the Tukey-Anscombe plot (residuals vs. fitted values) we observe a non-constant variance of the residuals. With increasing runoff the residuals increase.

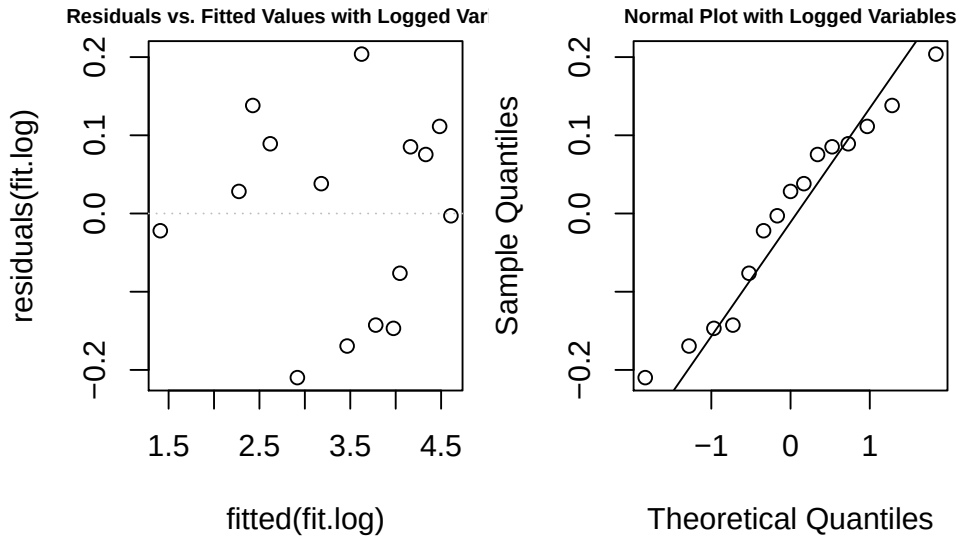
f) Although the histograms of the original data do not strongly point to a log-transformation, we try it and will see that it turns out to be useful.

```
> par(mfrow=c(2,2))
> hist(data$rainfall, 8, main="rainfall")
> hist(log(data$rainfall), 8, main="Histogram of log(rainfall)")
> hist(data$runoff, 8, main="Histogram of runoff")
> hist(log(data$runoff), 8, main="Histogram of log(runoff)")
```



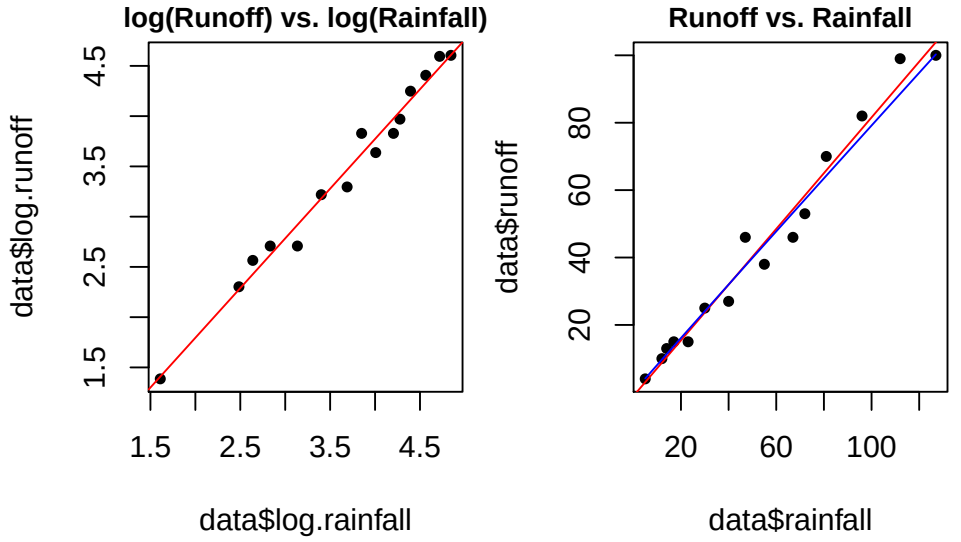
From the diagnostic plots we can see that the model on the transformed scale performs better, and the constant variance assumption seems more justified.

```
> data$log.runoff <- log(data$runoff)
> data$log.rainfall <- log(data$rainfall)
> fit.log <- lm(log.runoff ~ log.rainfall, data=data)
> par(mfrow = c(1,2))
> plot(fitted(fit.log), residuals(fit.log),
      main="Residuals vs. Fitted Values with Logged Variables",
      cex.main=0.7)
> abline(h=0, col="grey", lty=3)
> qqnorm(residuals(fit.log), main="Normal Plot with Logged Variables", cex.main=0.7)
> qqline(residuals(fit.log))
```



However, differences between the two models are small.

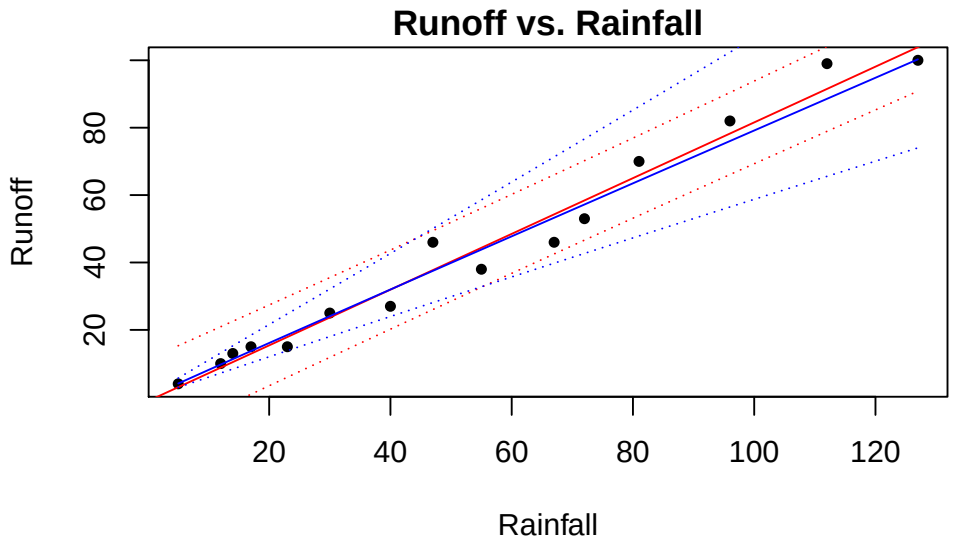
```
> par(mfrow=c(1,2))
> ## Scatterplot on the log scale
> plot(data$log.rainfall, data$log.runoff,
      main="log(Runoff) vs. log(Rainfall)", cex.main=0.9,
      pch=20)
> abline(fit.log, col="red")
> ## Scatterplot on original scale
> plot(data$rainfall, data$runoff, main = "Runoff vs. Rainfall", cex.main=0.9,
      pch=20)
> abline(fit, col="red")
> lines(rainfall, exp(predict(fit.log)), col="blue")
```



g) On the original scale the prediction interval of the log-transformed model is of the form of a trumpet (blue dot lines). This is more realistic, especially since fitted values and the prediction interval of the log-transformed model have positive values. Negative runoff values, as seen on the original scale, are impossible that is why the log-transformed model is superior to the original one. Although differences in the diagnostic plots seem small and problems appear to be more academic than fundamental, the log-transformed model resulting from a thorough statistical analysis pays off.

```
> ## Prediction intervals of the transformed model on the original scale
> plot(data$rainfall, data$runoff, pch=20, xlab="Rainfall", ylab="Runoff",
      main="Runoff vs. Rainfall")
> abline(fit, col="red")
> lines(data$rainfall, exp(predict(fit.log)), col="blue")
```

```
> interval <- predict(fit, interval="prediction")
> lines(data$rainfall, interval[,2], lty=3, col="red")
> lines(data$rainfall, interval[,3], lty=3, col="red")
> interval.log <- predict(fit.log, interval="prediction")
> lines(data$rainfall, exp(interval.log[,2]), lty=3, col="blue")
> lines(data$rainfall, exp(interval.log[,3]), lty=3, col="blue")
```



2. **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="confidence")
       fit      lwr      upr
1 694.0189 588.8902 799.1476
2 872.0000 814.2627 929.7373
```

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
```

It turns out that the covariates are collinear:

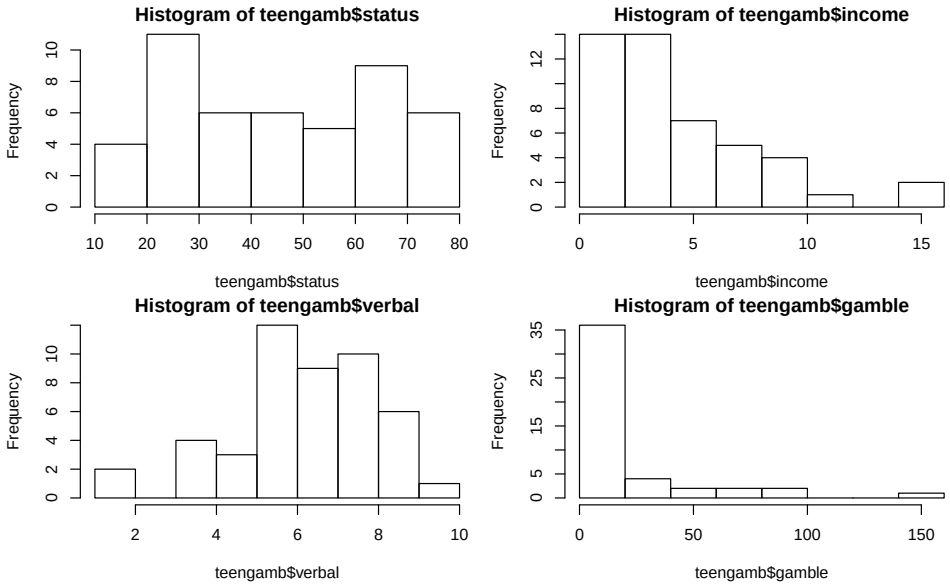
```
> fit3 <- lm(cows~acres, data=farm)
> summary(fit3)

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

Residuals:
    Min      1Q  Median      3Q     Max
-9.1163 -2.7169 -0.2916  4.1108  7.7800

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.90905    2.94280   6.086 9.46e-06 ***
acres       -0.06494    0.01962  -3.310   0.0039 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 4.834 on 18 degrees of freedom
Multiple R-squared: 0.3783,        Adjusted R-squared: 0.3438
F-statistic: 10.95 on 1 and 18 DF,  p-value: 0.003897
```

The income source *farm size* can only be identified 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.

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) Only a small part of the total variation in the response can be explained by the predictors, since R^2 is only 0.43.

e)
```
> mx.ind <- which.max(resid(fit.trsf))
> teengamb[mx.ind,]
     sex status income verbal gamble log.income log.gamble
5 female     65      2      8   19.6  0.6931472   2.980619
> summary(teengamb)
     sex         status          income           verbal          gamble
 male  :28   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0
 female:19   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1
             Median :43.00   Median : 3.250   Median : 7.00   Median :  6.0
             Mean   :45.23   Mean   : 4.642   Mean   : 6.66   Mean   : 19.3
             3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00   3rd Qu.: 19.4
             Max.   :75.00   Max.   :15.000   Max.   :10.00   Max.   :156.0
  log.income        log.gamble
 Min.   :-0.5108   Min.   :-2.3026
 1st Qu.: 0.6931   1st Qu.: 0.1788
 Median : 1.1787   Median : 1.8083
 Mean   : 1.2747   Mean   : 1.4412
 3rd Qu.: 1.8256   3rd Qu.: 2.9704
 Max.   : 2.7081   Max.   : 5.0505
```

The largest residual is associated with a female gambler that has a high socioeconomic status (based on the parents' occupation), good verbal communication skills, but low income and high gambling expenses compared to the average gambler.

f)
```
> median(resid(fit.trsf))
[1] 0.2745462
> mean(resid(fit.trsf))
[1] 1.708426e-17
```

In contrast to the median, the mean of the residuals is always zero. This is a consequence of the least squares method (the residuals are orthogonal to the columns in the design matrix, including (1,1,...,1)).

g)
```
> cor(resid(fit.trsf), fitted(fit.trsf))
[1] 2.434641e-16
```

```
> cor(resid(fit.trsf), teengamb$log.income)
[1] 8.067987e-17
```

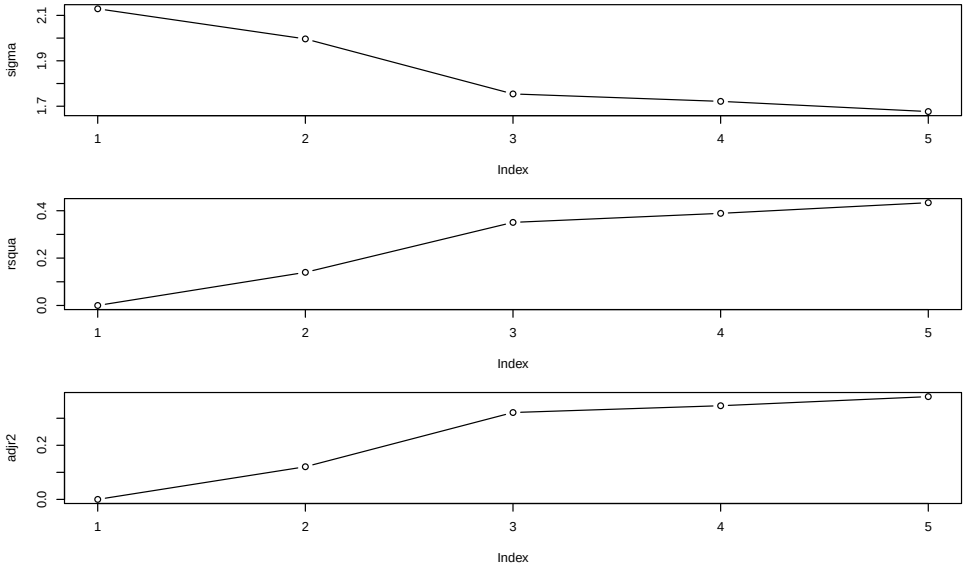
The correlations are practically zero. Again, this is a consequence of the least squares method.

h)
```
> 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.

i) 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)
```


sigma
2.1
1.9
1.7
1
2
3
4
5
Index
rsqua
0.4
0.2
0.0
1
2
3
4
5
Index
adjr2
0.2
0.0
1
2
3
4
5
Index