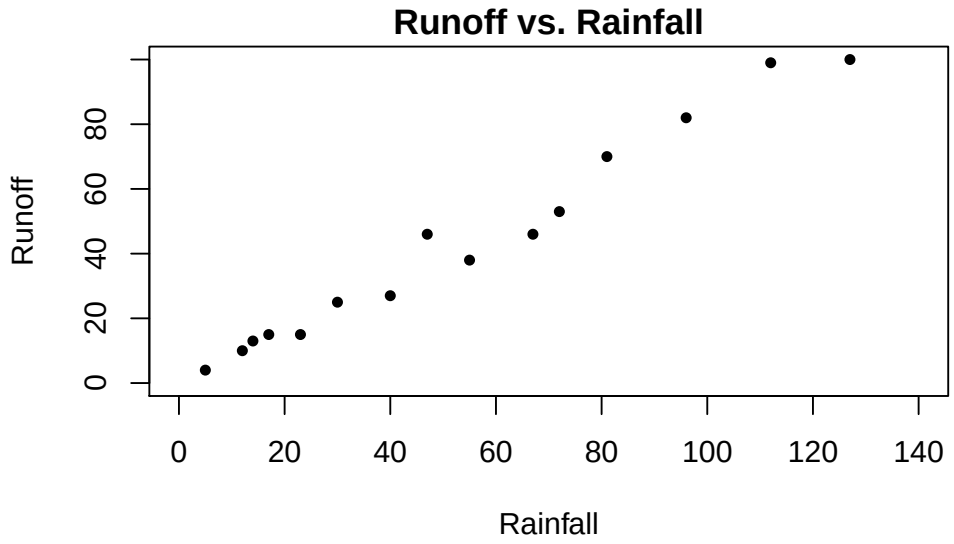
Solution to Series 3

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.

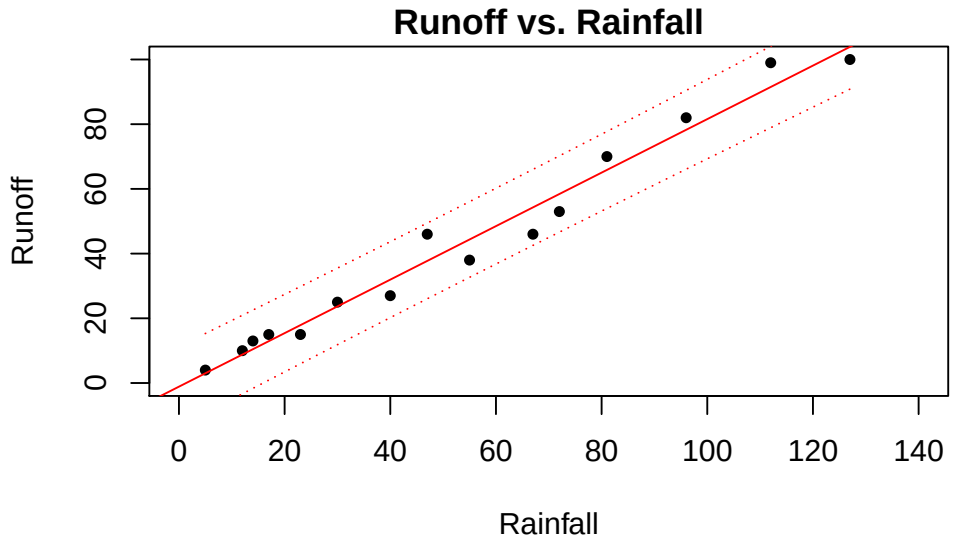
```
> rainfl <- c(5, 12, 14, 17, 23, 30, 40, 47, 55, 67, 72, 81, 96, 112, 127)
> runoff <- c(4, 10, 13, 15, 15, 25, 27, 46, 38, 46, 53, 70, 82,  99, 100)
> hrunof <- data.frame(rainfall=rainfl, runoff=runoff)
```



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=hrunof)
> 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.

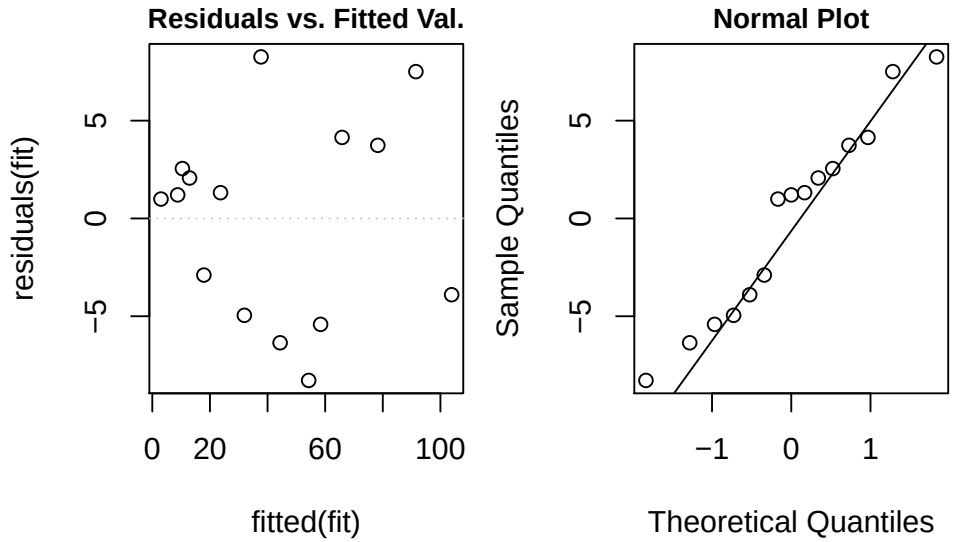


c) A R^2 of 0.98 is extremely high, i.e. that a huge part of the variation in the data can be attributed to the simple linear association between runoff and rainfall volume.

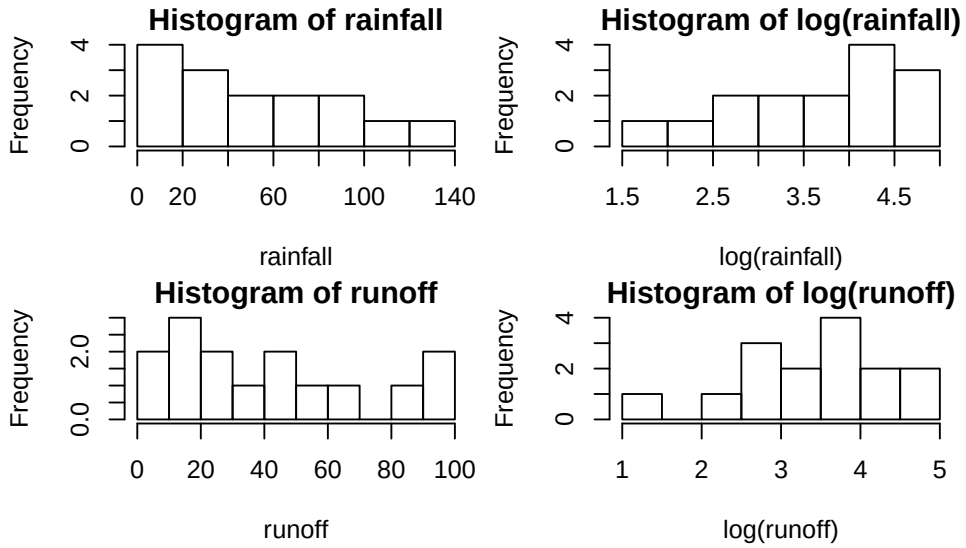
d) The null hypothesis $\beta_1 = 0$ is clearly rejected. However, the confidence interval for β_1 does not contain $\beta_1 = 1$, i.e. that a null hypothesis of $\beta_1 = 1$ would be rejected, too. Therefore, we conclude that no 1 : 1 relation between rainfall and runoff holds. We suspect that part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

```
> summary(fit)

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

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

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

e) From the Tukey-Anscombe plot (residuals vs. fitted values) we observe a non-constant variance of the residuals. With increasing runoff the residuals increase. This is a somewhat typical situation and physically plausible.

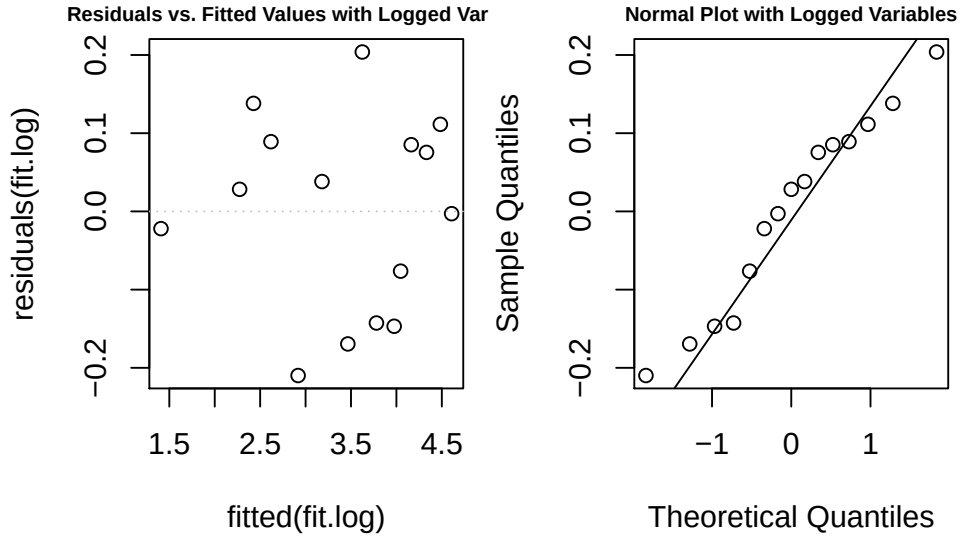


f) Although the histograms of the original data do not strongly point to a log-transformation, we try it and will see that it turns out to be useful.



Despite of the huge R^2 the model on the original scale does not fit very well. The constant variance assumption is violated. From the diagnostic plots we can see that the model on the transformed scale performs better.

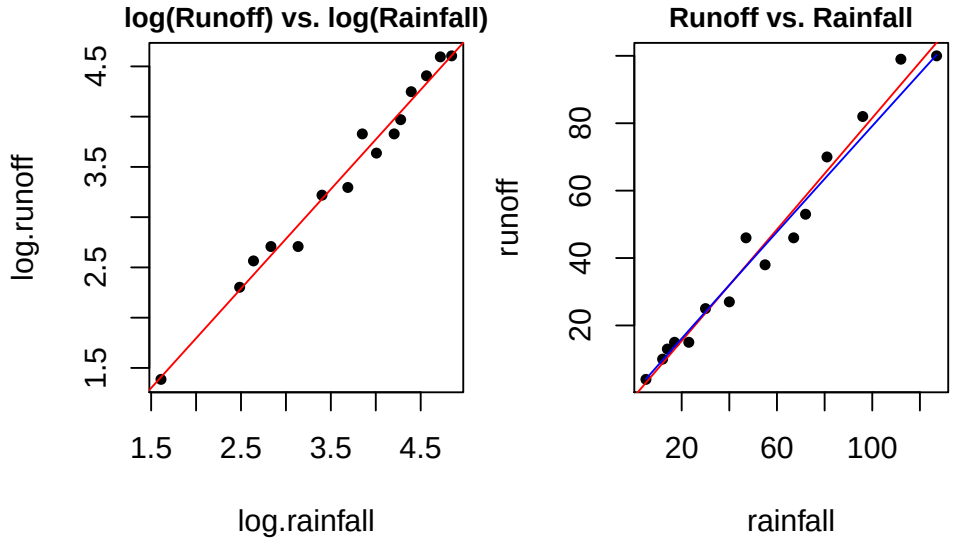
```
> hrunof <- within(hrunof, {
    log.runoff <- log(runoff)
    log.rainfall <- log(rainfall)
  })
> fit.log <- lm(log.runoff ~ log.rainfall, data=hrunof)
> par(mfrow = c(1,2))
> with(fit.log, {
    plot(fitted(fit.log), residuals(fit.log),
         main="Residuals vs. Fitted Values with Logged Variables",
         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.

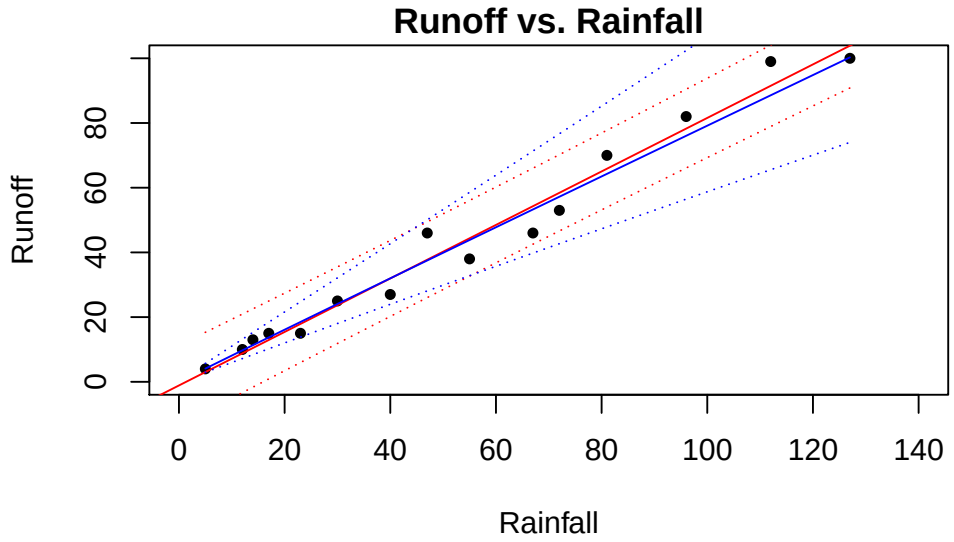
```
> par(mfrow=c(1,2))
> ## Scatterplot on the log scale
> with(hrunof, {
```

```
   plot(log.rainfall, log.runoff,
        main=c("log(Runoff) vs. log(Rainfall)"), cex.main=0.9,
        pch=20)
   abline(fit.log, col="red")
 })
> ## Scatterplot on original scale
> with(hrunof, {
   plot(rainfall, runoff, main = c("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 trsf model on the original scale
> with(hrunof, {
   plot(rainfall, runoff, pch=20,
        xlab="Rainfall", ylab="Runoff",
        xlim=c(0,140), ylim=c(0,100))
   title("Runoff vs. Rainfall")
   abline(fit, col="red")
   lines(rainfall, exp(predict(fit.log)), col="blue")

   interval <- predict(fit, interval="prediction")
   lines(rainfall, interval[,2], lty=3, col="red")
   lines(rainfall, interval[,3], lty=3, col="red")

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



2. **a)** Yes, the p-value (0.000475) is smaller than 5%.
Furthermore, we can from the estimation of the coefficient of C that for each additional cow the income increases by 20$.

```
farm <- read.table("http://stat.ethz.ch/Teaching/Datasets/farm.dat",header=TRUE)
attach(farm)
I <- Dollar
C <- cows
A <- acres

Call: lm(formula = I ~ C, 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 ***
C             20.111      4.725   4.256 0.000475 ***
---
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 degrees of freedom,
p-value: 0.0004751
```

b)

```
> xx <- data.frame(C=c(0,20,mean(C)))
> predc <- predict(mod1,xx,interval="confidence")
> predc
        fit      lwr       upr
1  694.0189 588.8902  799.1476
2 1096.2361 971.3953 1221.0768
3  872.0000 814.2627  929.7373

> predp <- predict(mod1,xx,interval="prediction")
> predp
        fit      lwr       upr
1  694.0189 415.2286  972.8092
```

```
2 1096.2361 809.4309 1383.0412
3  872.0000 607.4143 1136.5857
```

c)
```
> summary(mod1)
Call: lm(formula = I ~ A, 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 ***
A             0.0234     0.7066   0.033    0.974
---
Residual standard error: 174.1 on 18 degrees of freedom
Multiple R-Squared: 6.09e-005,  Adjusted R-squared: -0.05549
F-statistic: 0.001096 on 1 and 18 degrees
of freedom,   p-value: 0.974
> summary(mod2)
Call:
lm(formula = A ~ C, data = farm)

Residuals:
     Min       1Q   Median       3Q      Max
-76.1038 -31.0814  -0.7132  31.4186  89.7221

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  191.059     18.641   10.25  6.1e-09 ***
C             -5.826      1.760   -3.31   0.0039 **
---
Residual standard error: 45.78 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
> summary(mod3)
Call: lm(formula = I ~ A + C, 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 **
A             2.1384     0.3936   5.434 4.47e-05 ***
C            32.5690     3.7276   8.737 1.08e-07 ***
---
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 degrees of freedom, p-value: 5.165e-007
```

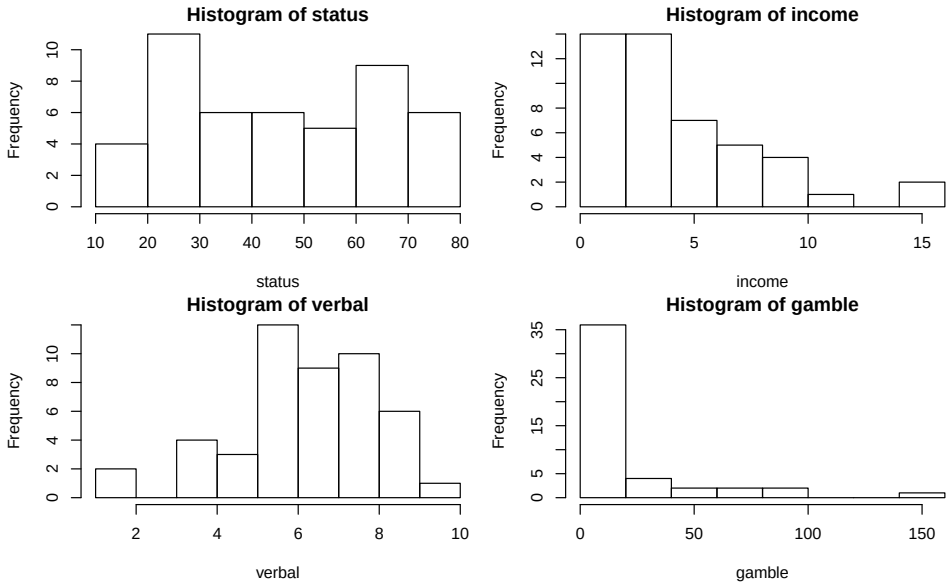
The income source *land* can only be identified if we control for the number of cows, i.e., comparing like with like.
In colloquial terms, the positive correlation of I and C and the negative correlation of C and A cancel each other out. Thus, the variable A is not considered significant in a univariate regression

of I onto A.
It holds that $I \approx 285.5 + 2.1A + 32.6C$, because $C \approx 17.9 - 0.065A$ it follows that $I \approx 285.5 + 2.1A + 32.6(17.9 - 0.065A) = 869.04 - 0.019A$. This means variable A is no longer significant.

3. **a)** First we read in the data and account for the categorial nature of `sex`.

```
> ## Load data
> file <- url("http://stat.ethz.ch/education/semesters/as2011/asr/teengamb.rda")
> load(file)
> ## Choose correct data type for sex
> teengamb <- within(teengamb, sex <- factor(sex, labels=c("male", "female")))
```

Then we draw histograms for the different variables.



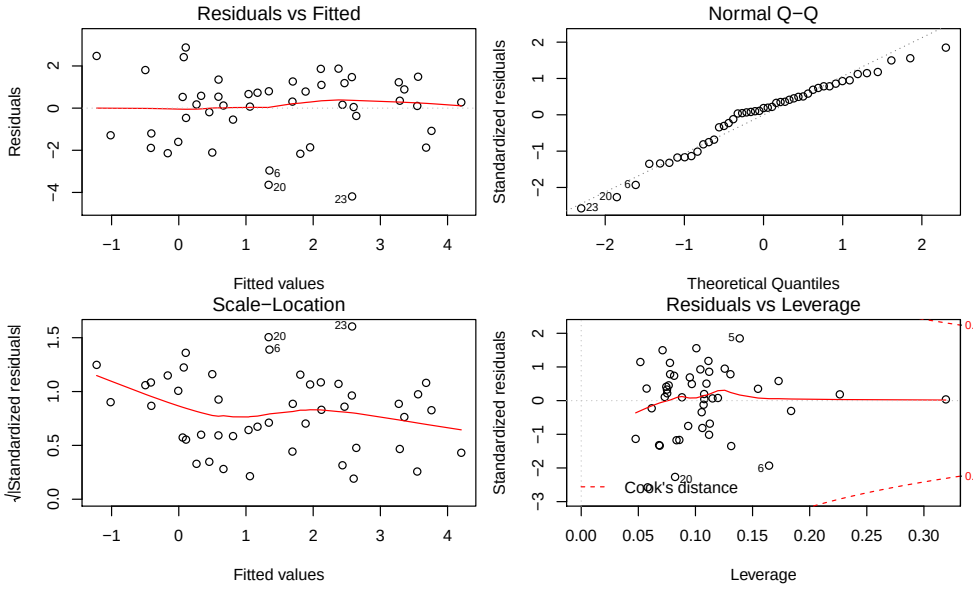
The histograms of `income` and `gamble` show skewed distributions. Therefore, we perform a log transformation. Due to the fact that 4 data points of `gamble` are zero, we need to add a constant (here: 0.1) prior to transformation.

```
> ## Transformations
> any(teengamb$income==0)   # log trsf directly possible
[1] FALSE
> any(teengamb$gamble==0)   # any zeros?
[1] TRUE
> sum(teengamb$gamble==0)   # how many zeros? before log, need to add 0.1
[1] 4
> teengamb <- within(teengamb, {
   log.income <- log(teengamb$income)
   log.gamble <- log(teengamb$gamble+0.1)
 })
```

b) After having transformed `gamble` and `income`, we fit a linear regression model to the data. Note, that we have also fitted the original model. For further inside, compare the results of the original model with the transformed one.

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

c) Only a small part of the total variation in the response can be explained by the predictors, i.e. that the R^2 is only 0.43.
However, if we compare the diagnostic plots of the transformed model with the original model (not shown), we see that the model assumptions are violated in case of the original data. Due to the log transformation we accounted for the non-constant variance and non-zero expectation of the residuals of the original model.



Furthermore, we see that for the original model negative gambling expenditures are predicted (not shown), which is an artifact. Contrarily, the transformed model predicts only positive expenses on the original scale.

d) 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.

```
> mx.ind <- which.max(resid(fit.trsf))
> teengamb[mx.ind,]
     sex status income verbal gamble log.gamble log.income
5 female     65      2      8   19.6   2.980619  0.6931472
> 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.gamble        log.income
 Min.   :-2.3026   Min.   :-0.5108
 1st Qu.: 0.1788   1st Qu.: 0.6931
 Median : 1.8083   Median : 1.1787
 Mean   : 1.4412   Mean   : 1.2747
 3rd Qu.: 2.9704   3rd Qu.: 1.8256
 Max.   : 5.0505   Max.   : 2.7081
```

e) In contrast to the median, the mean of the residuals is always zero. This is a consequence of the least squares method that minimizes the residual sum of squares.

f) In line with the previous task, the least squares method guarantees that the residuals are neither correlated with the predictors nor the fitted values.

```
> cor(resid(fit.trsf), fitted(fit.trsf))
> cor(resid(fit.trsf), teengamb$log.income)
> cor(resid(fit.trsf), teengamb$status)
```

g) 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.

h) 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)
```