Homework 7 Solution

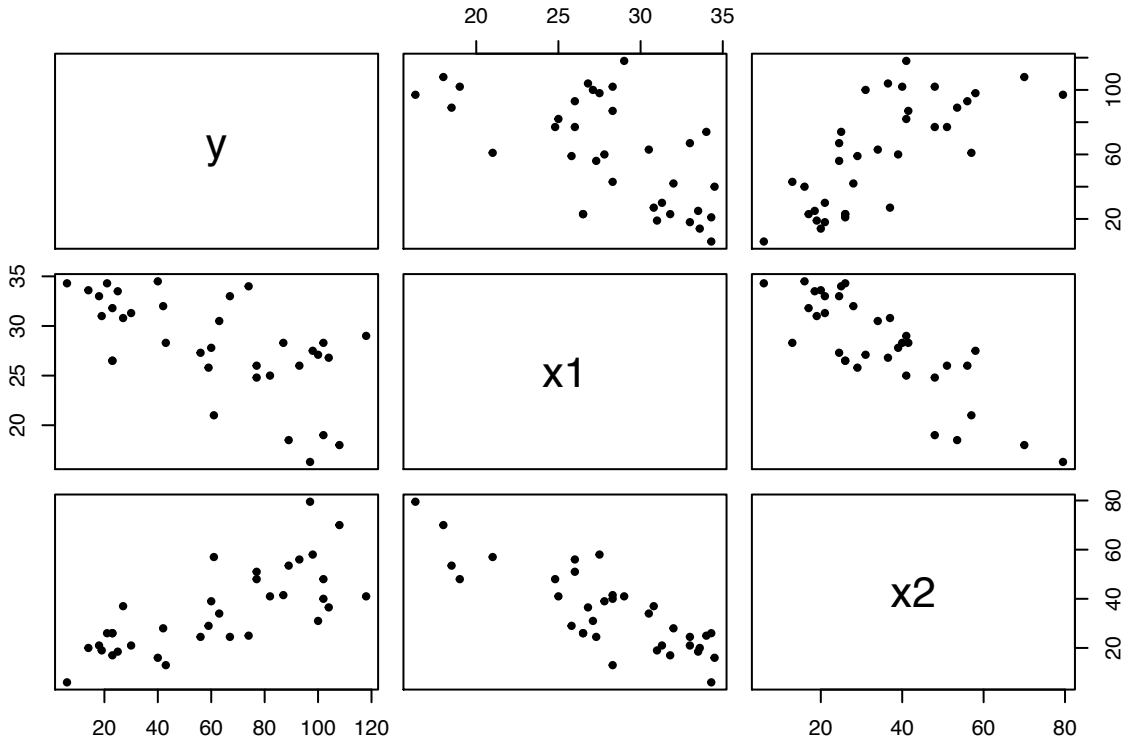
Christian Schmid

October 23, 2019

```
# read in data
aphid <- read.csv("aphideData.csv")
colnames(aphid)<- c("y", "x1", "x2")
```

1 (6 points, 3 for plot, 3 for comment)

```
plot(aphid, pch=20)
```



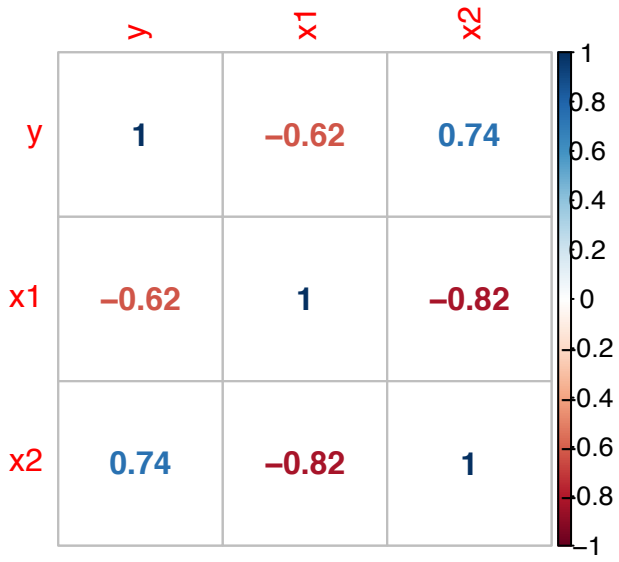
The response and x1 appear to be negatively correlated and the response and x2 appear to be positively correlated. The two variables x1 and x2 appear to strongly positively correlated, which would violate the model assumptions.

2 (6 points, 3 for plot, 3 for comment)

```
# obtain correlation matrix
aphid.cor.mat <- cor(aphid)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot(aphid.cor.mat, method="number")
```

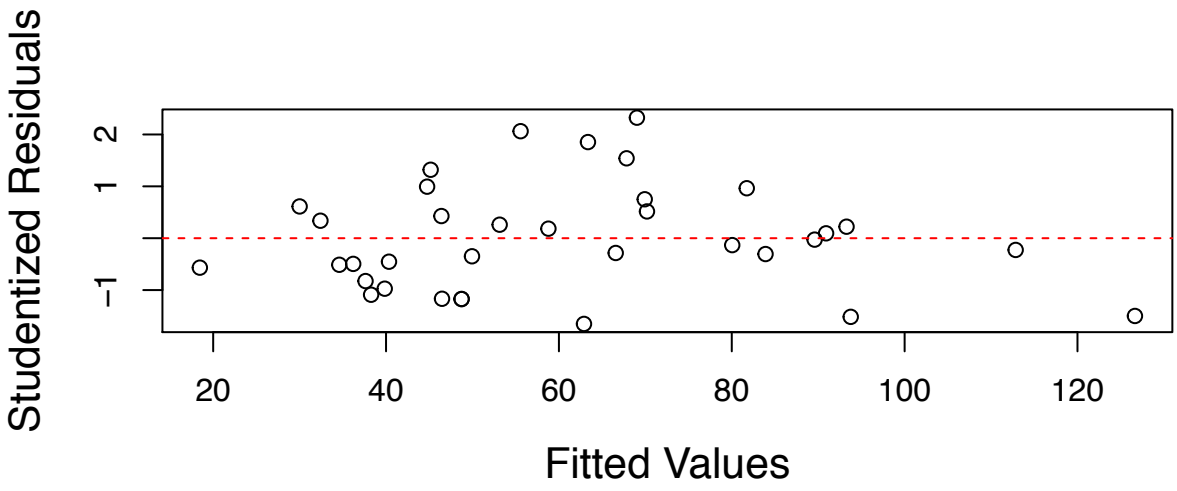


As observed in 1), there is a negative correlation between y and x1 and a positive correlation between y and x2. In addition, the two predictor variables x1 and x2 are highly correlated.

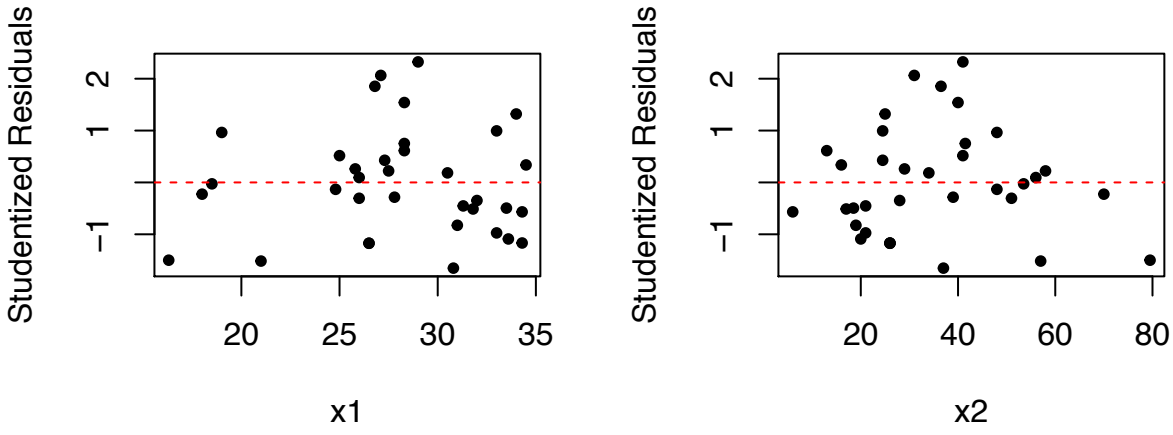
3 (7 points, 3 for model fit, 2 for linearity, constant variance and normality, 2 for multicollinearity)

```
model1 <- lm(y~ x1+x2, data=aphid)

#obtaining studentized residuals
#or standardized residuals: std.residuals = rstandard(fullModel)
library(MASS)
stud.residuals = studres(model1)
fitted.values = fitted.values(model1)
#Plot of residuals Vs. fitted values
plot(x = fitted.values, y = stud.residuals, xlab = "Fitted Values",
     ylab = "Studentized Residuals", cex.lab=1.3)
abline(h=0, col="red",lty=2)
abline(h=-3,col="blue",lty=3)
abline(h=3,col="blue",lty=3)
```




```
#Plots of residuals vs. each predictor
par(mfrow=c(1,2)) #This line split the plot window in to 1 by 2 small windows
plot(x = aphid$x1, y = stud.residuals, xlab = "x1", ylab = "Studentized Residuals", pch=20)
abline(h=0, col="red",lty=2)
abline(h=-3,col="blue",lty=3)
abline(h=3,col="blue",lty=3)
plot(x = aphid$x2, y = stud.residuals, xlab = "x2", ylab = "Studentized Residuals", pch=20)
abline(h=0, col="red",lty=2)
abline(h=-3,col="blue",lty=3)
abline(h=3,col="blue",lty=3)
```




```
par(mfrow=c(1,1)) #I want to go back to the single plot window setting
qqnorm(stud.residuals)
abline(a=0, b=1, col="red")
library(nortest)
ad.test(stud.residuals)
```

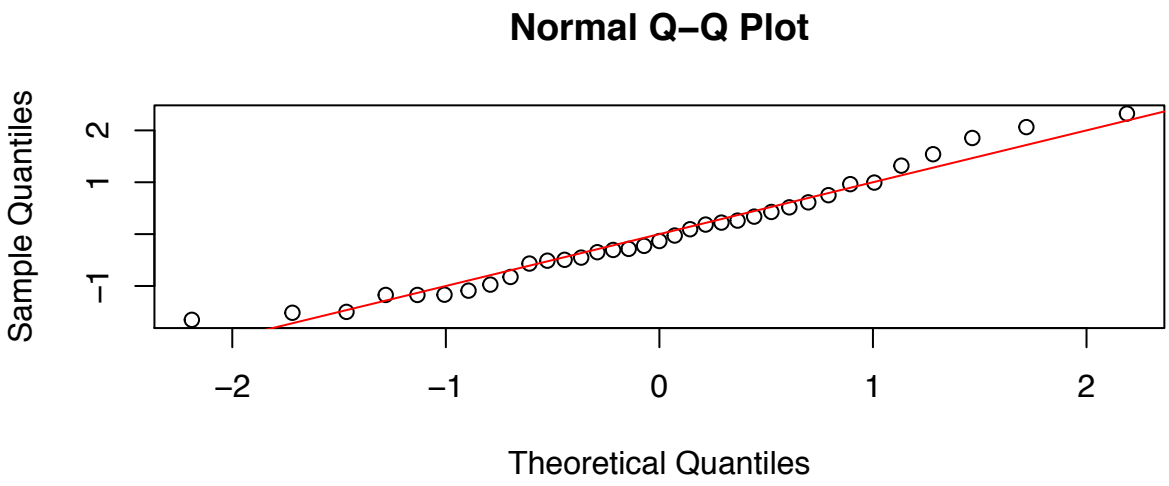
```
##
##  Anderson-Darling normality test
##
## data:  stud.residuals
## A = 0.29295, p-value = 0.5832
```

```
#Breush-Pagan test
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```




```
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 0.37287, df = 2, p-value = 0.8299
```

```
#vif values
library(car)
```

```
## Loading required package: carData
```

```
vif(model1)
```

```
##       x1       x2
## 3.076691 3.076691
```

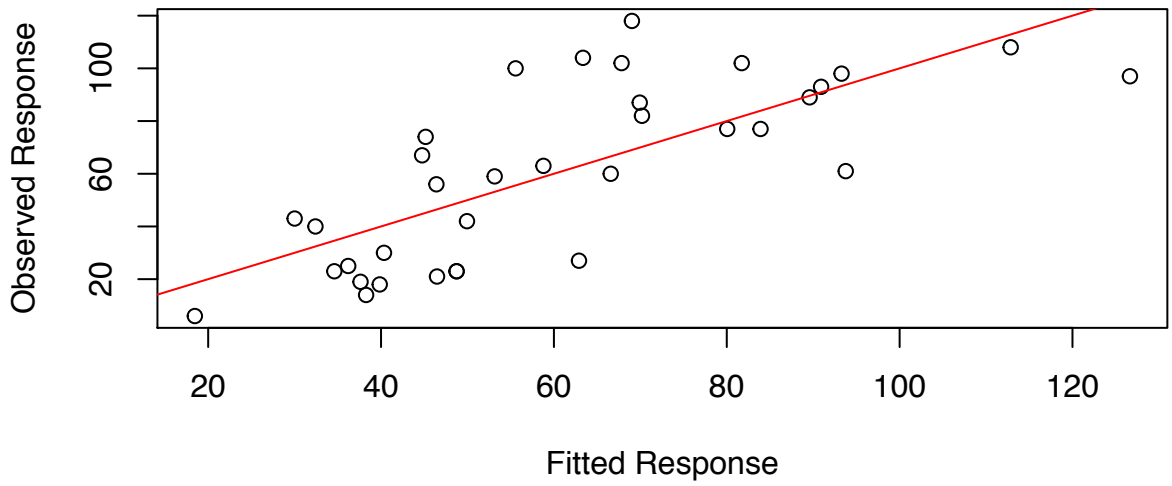
Linearity and Constant Variance: Since all the points in the residual plots: Studentized residual Vs. Fitted and Studentized residual Vs. each predictor variable are within (-3, 3) limits randomly scattered around zero horizontal line without any particular pattern, we don't have any concerns of possible violations of

linearity. They do not suggest any departure from homogeneity either. Breush-Pagan test also support the homogeneity of the errors (p-value=0.8299>0.05).

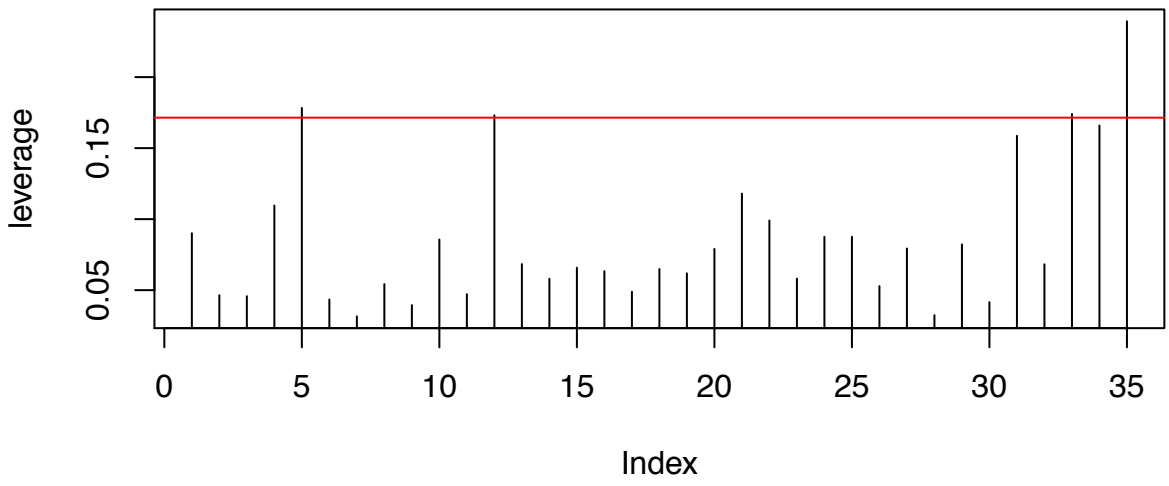
Normality: Points in the normal QQ plot stay close to the reference line supporting the normality of the errors. This can be further confirmed by looking at the p-value for the Anderson-Darling test (p-value 0.5832>0.05)

Multicolinearity could be an issue since the VIF is close to 3. However, it still would not be an serious issue since it is less than 10.

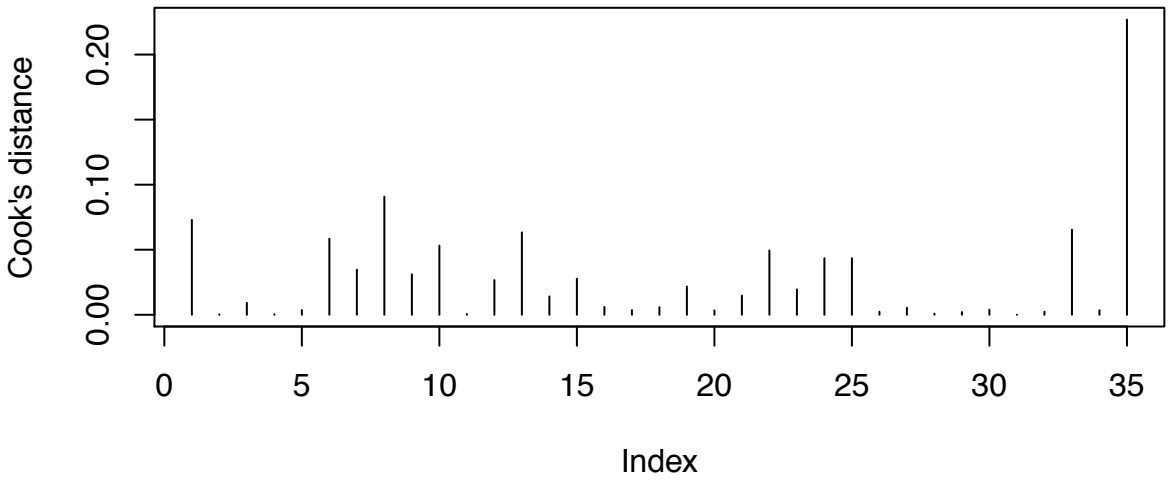
```
#Plot of fitted and observed values
plot(fitted.values,aphid$y, xlab="Fitted Response", ylab="Observed Response")
abline(a=0,b=1,col=2)
```




```
#plot of leverage values
plot(hatvalues(model1), type = "h", ylab="leverage")
n = nrow(aphid)
p = length(coefficients(model1))
cutLev = 2*p/n
abline(h=cutLev, col="red")
```




```
#cook's distance
plot(cooks.distance(model1), ylab="Cook's distance", type="h")
abline(h=1, col="red")
```



Outliers: From the plot of fitted and observed values, we can see they agree with each other most the time.

High-leverage: There is one observation that seems to have relatively large leverage value.

Although, there seems to be at least one observation with high-leverage value, none of the observations seems to be influential.

4 (9 points, 3 for model, 2 each for interpreation of β_0, β_1 and β_2)

```
summary(model1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = aphid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.901 -15.541  -3.053  12.404  48.973
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.0026    52.1131   0.384  0.70364
## x1           -0.2902     1.4058  -0.206  0.83779
## x2            1.4010     0.4143   3.382  0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.11 on 32 degrees of freedom
## Multiple R-squared:  0.5483, Adjusted R-squared:  0.5201
## F-statistic: 19.42 on 2 and 32 DF,  p-value: 3.001e-06
```

The estimated model is

$$\hat{y} = 20.00 - 0.29 \cdot x1 + 1.40 \cdot x2$$

Interpretation of intercept (β_0): If the mean temperature (x1) is 0 and the mean relative humidity is 0, we expect the infestation rate to be 20 aphids/100 leaves.

Interpretation of β_1: The expected change in the infestation rate for a one unit increase in x1, when the x2-value is being held constant, is -0.29.

Interpretation of β_2: The expected change in the infestation rate for a one unit increase in x2, when the x1-value is being held constant, is 1.40.

5 (10 points, 3 hypothesis, 3 ANOVA table, 2 p-value, 2 conclusion)

```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x1         1 14643.0 14643.0  27.410 1.002e-05 ***
## x2         1  6109.7  6109.7  11.437  0.001914 **
## Residuals 32 17094.9   534.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Hypotheses are

$$H_0 : \beta_1 = \beta_2 = 0 \quad , \quad H_1 : \text{At least one } \beta_j \neq 0 \quad j = 1, 2$$

ANOVA table:

Under the null hypothesis, this follows a F-distribution with 2 and 32 degrees of freedom. Thus, the p-value is

Source	df	SS	MS	F
Regression	2	20752.7	10376.35	19.42352
Error	32	17094.9	534.22	
Total	34	37847.6		

```
1-pf(19.42352, df1=2, df2=32)
```

```
## [1] 3.000905e-06
```

The p-value is less than $\alpha = 0.05$. Thus, we reject H_0 and conclude that this model fits the data better than the model with no predictor variables, meaning that there is a significant linear relationship between the response and the predictor variables.

6 (3 points)

$R^2 = \frac{SSR}{SSTO} = \frac{20752.7}{37847.6} = 0.5483$

7 (8 points, 2 each for hypothesis, test statistic, p-value, and conclusion)

Hypothesis:

$$H_0 : \beta_2 = 0 \quad , \quad H_1 : \beta_2 \neq 0$$

Test statistic:

$$t = \frac{\hat{\beta}_2 - \beta_2}{\hat{SE}(\hat{\beta}_2)} = \frac{1.40 - 0}{0.41} = 3.38$$

Under the null this follows a t-distribution with $n - k - 1 = 35 - 2 - 1 = 32$ degrees of freedom. Thus, the p-value is

```
2*(1-pt(3.38, df=32))
```

```
## [1] 0.001923039
```

The p-value is smaller than our significance level $\alpha = 0.05$. Therefore, we reject the null and conclude that x2 is a significant predictor.

8 (8 points, 2 each for hypothesis, test statistic, p-value, and conclusion)

Hypothesis:

$$H_0 : \beta_1 = 0 \quad , \quad H_1 : \beta_1 \neq 0$$

Test statistic:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\hat{SE}(\hat{\beta}_1)} = \frac{-0.29 - 0}{1.41} = -0.206$$

Under the null this follows a t-distribution with $n - k - 1 = 35 - 2 - 1 = 32$ degrees of freedom. Thus, the p-value is

```
2*(pt(-0.206, df=32))
```

```
## [1] 0.8380959
```

The p-value is larger than our significance level $\alpha = 0.05$. Therefore, we fail to reject the null and conclude that x1 is not a significant predictor.

9 (3 points)

Since x1 turned out to be insignificant, the next step would be to remove this variable from our model.

10 (6 points, 3 for model fit, 3 for written model)

```
model2 <- lm(y~x2, data=aphid)
summary(model2)
```

```
##
## Call:
## lm(formula = y ~ x2, data = aphid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.847 -15.303  -3.031  13.364  48.268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.4132     8.9942   1.047    0.303
## x2            1.4712     0.2327   6.322 3.75e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.78 on 33 degrees of freedom
## Multiple R-squared:  0.5477, Adjusted R-squared:  0.534
## F-statistic: 39.96 on 1 and 33 DF,  p-value: 3.755e-07
```

The SLR model is

$$y = 9.41 + 1.47 * x2$$

11 (9 points, 2 hypothesis, 3 ANOVA, 2 p-value, 2 conclusion)

Hypotheses:

$$H_0 : \beta_1 = 0 \quad , \quad H_1 : \beta_1 \neq 0$$

```
anova(model2, model1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x2
## Model 2: y ~ x1 + x2
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     33 17118
## 2     32 17095  1    22.756 0.0426 0.8378
```

The test statistic is $F = 0.0426$ and the corresponding p-value is 0.8378. The p-value is larger than our significance level $\alpha = 0.05$, which means we fail to reject the null hypothesis.

12 (4 points)

In 11) we failed to reject $H_0 : \beta_1 = 0$. For this reason, we would choose model 2, since adding x1 does not significantly improve our model and our objective is to find a good and parsimonious model.

13 (10 points, 3 for model fit, 3 for written model, 2 for right choice, 2 for reasoning)

```
model3<- lm( y~ x1, data=aphid)
summary(model3)
```

```
##
## Call:
## lm(formula = y ~ x1, data = aphid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.262 -19.554  -2.183  18.081  61.228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 178.4601    26.1704   6.819 8.85e-08 ***
## x1           -4.1962     0.9195  -4.563 6.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.52 on 33 degrees of freedom
## Multiple R-squared:  0.3869, Adjusted R-squared:  0.3683
## F-statistic: 20.82 on 1 and 33 DF,  p-value: 6.638e-05
```

Model 2 has a R^2-value of 0.55, while model 3 has a R^2-value of 0.39. Based on these two R^2-values I would choose model 2 over model 3, since variable x2 explains more of the variability in y than x1.

14 (6 points, 3 for right choice, 3 for reasoning)

```
max(aphid$x2); min(aphid$x2)
```

```
## [1] 79.5
```

```
## [1] 6
```

A mean relative humidity of x2 = 115 lies outside the range of observed values for x2 and would result in model extrapolation. For this reason, I would highly recommend not using model 2 to predict the average infestation rate at x2 = 115.

15 (5 points)

We are trying to predict the infestation for a single cotton plan for a given average value of relative humidity. Thus, we need to calculate a confidence interval.

```
# create data frame with new x2 value
new.aphid <- data.frame(x2=65)
# calculate confidence interval
predict(model2, new.aphid, interval="confidence", level=0.95)
```

```
##       fit      lwr      upr
## 1 105.041 88.79078 121.2912
```

The confidence interval is (88.79, 121.29).