SDA - Assignment 7

Leon Lušić (2670440) - no group

Exercise 7.1

(a) The scatter plots of the photo counts versus the observer counts are presented below:

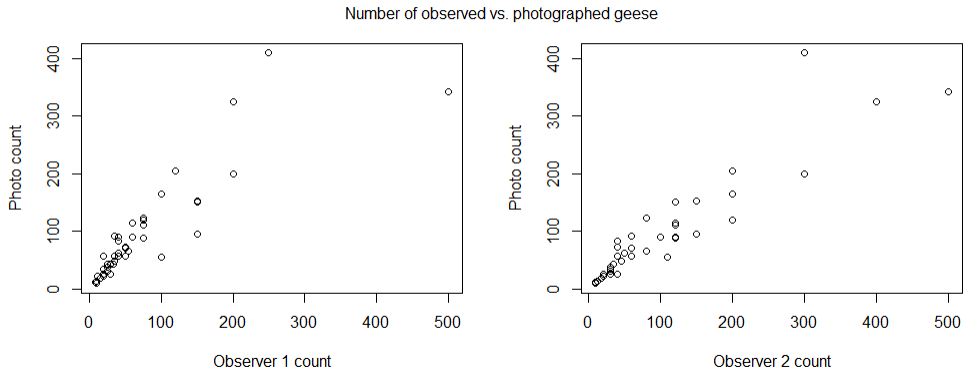

Figure 1: Observed vs. photo counts

The data appears to follow a linear pattern. Although the variability in the data appears to increase as either of the variables increases, the overall trend still seems to be linear.

(b) Looking at the p-values corresponding to the F-statistic of each of the two simple linear models, we find the following values:

Exp. variable	observer1	observer2
p-value	$1.573 \cdot 10^{-14}$	$2.2 \cdot 10^{-16}$
R^2	0.7503	0.8547

Table 1: Table of p-values and R^2 for the two simple linear models

As both of the p-values are significantly below 0.05, we confidently reject the null hypothesis $H_0 : \beta_1 = 0$ and accept the alternative $H_1 : \beta_1 \neq 0$. We also included the values of the determination coefficient (R^2) for the purposes of further analysis.

(c) Plotting the residuals of both models against the photo counts, we find the following distributions:

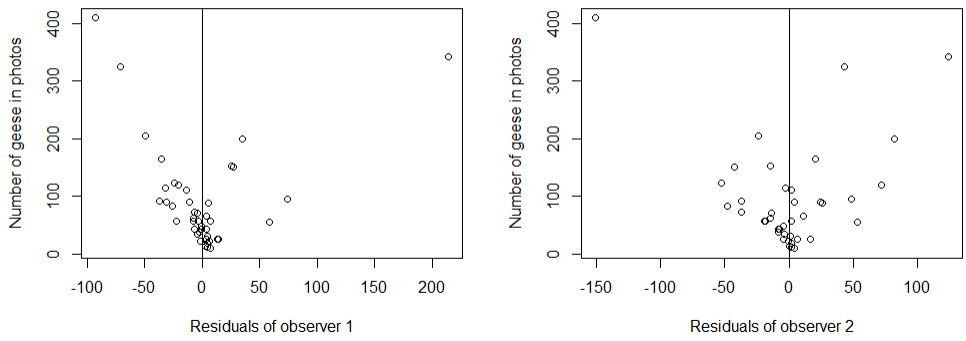

Figure 2: Residuals vs. photo counts

The vertical lines in the plots correspond to the line $x = 0$. As can be seen, the residuals are heavily grouped around the zero, with decreasing concentration the further we move from zero. Also, the residuals appear to be at least somewhat symmetrically distributed around $x = 0$. Based on these plots, the normality assumption on the residuals appears to be justified.

(d) In order to visually check the normality assumption, we plot two QQ-plots against the normal distribution:

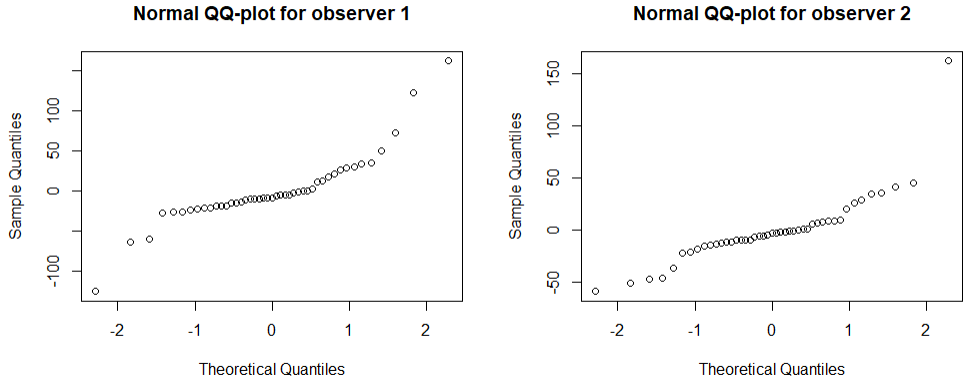

Figure 3: QQ-plots of the residuals against the normal distribution

As can be seen in the plots, neither show a strong linear patter, hence highly raising suspicion around the normality assumption. In order to further check this, we perform the Kolmogorov-Smirnov test on the residuals. The resultant p-values for the both datasets are $2.2 \cdot 10^{-16}$ (for both tests). As this is basically as significant as it gets, we confidently reject the null hypothesis and conclude that the residuals are not normally distributed.

(e) Now, we will repeat the steps **(a)** through **(d)**, but on the data obtained by a log transformation of all individual data points. First, we will look at the plots of the log of observed counts versus the log of the photo counts. The resultant plots are as follows:

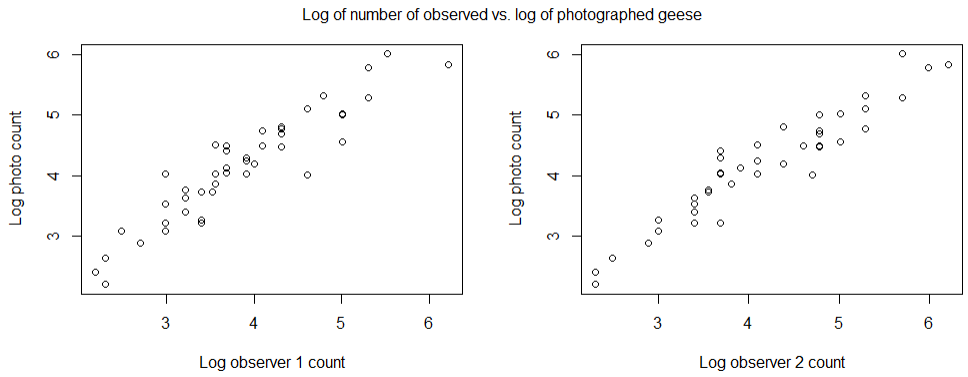

Figure 4: Log observed vs. log photo counts

As can be immediately seen, the data appears to be somewhat 'cleaner' and more regular; there is no apparent increase in variance as either of the variables increases.

Looking at the p-values corresponding to the F-statistic of each of the new log-log linear models (linear models between `log(photo)` and `log(observer1/2)`, we find the following values:

Exp. variable	log(observer1)	log(observer2)
p-value	$2.2 \cdot 10^{-16}$	$2.2 \cdot 10^{-16}$
R^2	0.8658	0.9051

Table 2: Table of p-values and R^2 for the two log-log linear models

Not only did the p-values decrease (in the first case, at least), but also the R^2 values increased. As the p-values are still non-significant, we reject the null hypothesis $H_0 : \beta_1' = 0$ in both cases and conclude that $\beta_1' \neq 0$ (as we chose $\vec{\beta}$ for the model in parts **(a)-(d)**, we choose $\vec{\beta}'$ for the log-log model).

Lastly, let us look at the plots of new residuals against the log of the photo counts, as well as the QQ-plot of the residuals against the normal distribution:

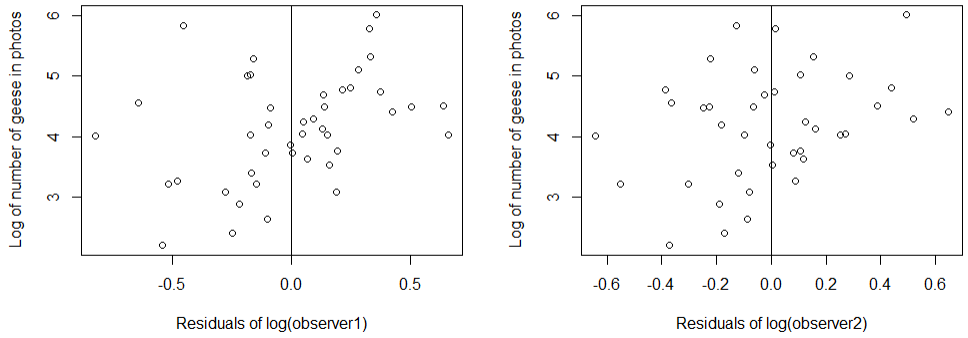

Figure 5: Residuals vs. log of photo counts

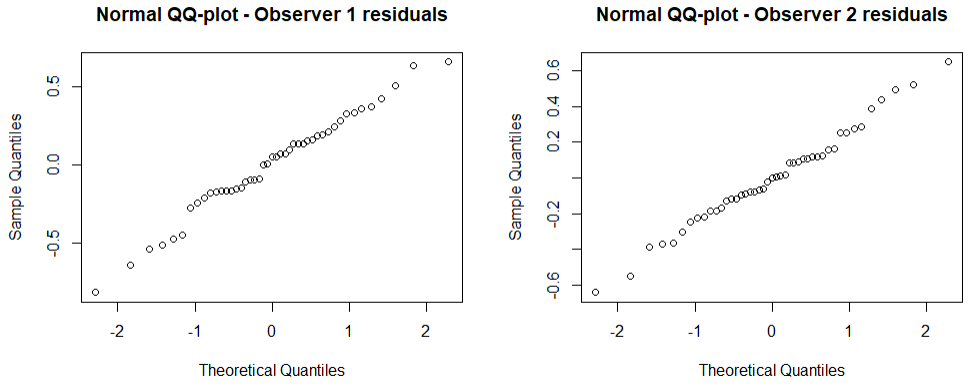

Figure 6: QQ-plot against the normal distribution

As can be seen from the plots of residuals, they still appear to be centered around $x = 0$, being less concentrated farther away, and more concentrated around the center. Furthermore, looking at the QQ-plots, we see that they appear to form a very straight line against the normal distribution quantiles, which justifies the normality assumption.

(f) Looking at the first two models, we see that the data appears to follow a straight line nearly perfectly for small counts, but increasingly strays away from it as we increase the counts. The QQ-plots show that the normality assumption might not be as valid since the tails appear to be too heavy for a normal distribution. On the other hand, based on the plots of the log-transformed data, all data points appear to follow a straight line very accurately. The variance in the data appears to be somewhat constant as the counts increase, while the errors seem to be normally distributed. Furthermore, the R^2 is greater for the model based on the log-transformed data. Based on these arguments, the second pair of models appears to be more trustworthy and accurate.

(g) While there is obviously a fairly strong linear relationship between the observer counts and the photo counts of the geese, the trustworthiness of it tends to diminish as the number of counts increase. While still following the same line, it appears that the photo counts tend to deviate increasingly more as the observer counts increase. However, although the guesses become worse as the number of geese increases, they do not do so in a random manner. If we look at the logarithm of the estimated guesses, as well as the logarithm of the photo counts, these tend to follow another very accurate linear model. Furthermore, in this model, the data truly appear to be normally distributed, with overall approximately constant variance based, regardless of the (logs of) observer and photo counts.

Exercise 7.2

(a) In order to get a feeling about any potential linear dependencies, we look at the scatter plots of all independent variables and the response variable against each other.

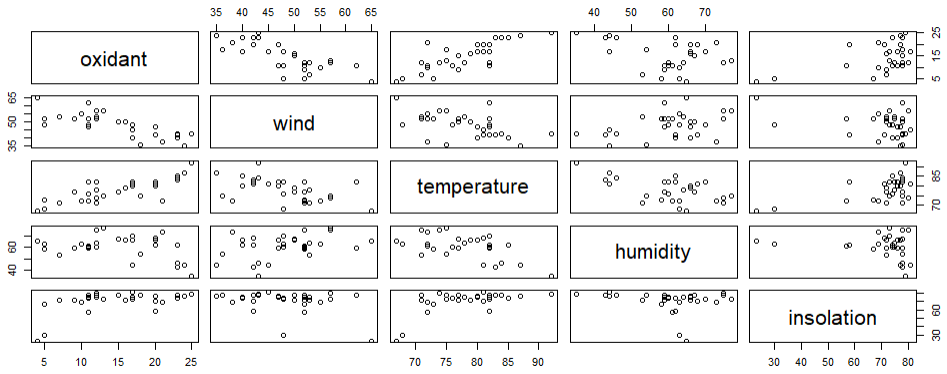

Figure 7: Plots of independent variables and response variable against each other

Although it might appear a bit unclear at first, we may notice that there appears to be negative linear correlation between `oxidant` (the response variable) and `wind`, as well as a positive linear correlation between `oxidant` and `temperature`. A linear relation to other variables appears absent, although we still need to test this.

(b) Looking at the determination coefficient R^2 for simple linear regression models for each of the candidate explanatory variables, we find the following values:

Exp. variable	Wind	Temperature	Humidity	Insolation
R^2	0.5863	0.576	0.124	0.2552

Table 3: Table of R^2 for simple lin. reg. models

The largest determination coefficient corresponds to `Wind` variable. The p-value corresponding to the t-test is smaller than 0.05, so we include `Wind` in the model. Next up, we find the following values of R^2 for the models with other variables along the `Wind` variable:

Exp. variables	Wind + Temperature	Wind + Humidity	Wind + Insolation
R^2	0.7773	0.5913	0.6613

Table 4: Table of R^2 for the first step-up lin. reg. models

The new largest R^2 corresponds to the model with `Temperature` variable added. The p-value corresponding to the t-test is smaller than 0.05, so we include `Temperature` in the model. Next up, we find the following values of R^2 for the models with other variables along the `Wind` and `Temperature` variables:

Exp. variables	Wind + Temperature + Humidity	Wind + Temperature + Insolation
R^2	0.7964	0.7816

Table 5: Table of R^2 for the second step-up lin. reg. models

The new largest R^2 corresponds to the model with `Humidity` variable added. The p-value corresponding to the t-test (0.131) is larger than 0.05, so we do not include `Humidity` in the model. Therefore, our final model uses only `Wind` and `Temperature` as explanatory variables.

(c) Performing the *overall* analysis of the full model, we find the p-value corresponding to the F statistic to be $2.279 \cdot 10^{-8}$. Therefore, we safely reject the null hypothesis and conclude that at least one variable should be included in the model.

(d) Looking at the `summary` for the full model, we find the following p-values for the t-test for each of the explanatory variables, as well as the intercept:

Exp. variable	(Intercept)	Wind	Temperature	Humidity	Insolation
p-value	0.26219	$2.85 \cdot 10^{-5}$	0.00041	0.16743	0.65728

Table 6: Table of p-values (for the t-test) for the full model

As can be seen, the largest p-value corresponds to the `Insolation` variable. Hence, we remove it from the model and repeat the process. The p-values for the model with all the variables except for `Insolation` are:

Exp. variable	(Intercept)	Wind	Temperature	Humidity
p-value	0.215	$1.78 \cdot 10^{-5}$	$2.47 \cdot 10^{-5}$	0.131

Table 7: Table of p-values for the first step-down model

In this case, the p-value corresponding to the intercept is the largest. However, it is usually preferable to include the intercept as it might have some physical significance. In particular, even if we choose to exclude it, it is guaranteed to appear if the temperature measurements were to switch to Celsius or some other unit. Hence, since the p-value

corresponding to `Humidity` is the second largest (while also being significant), we exclude it. Therefore, the p-values for the model with all the variables except for `Humidity` are:

Exp. variable	(Intercept)	Wind	Temperature
p-value	0.644	$3.58 \cdot 10^{-5}$	$5.05 \cdot 10^{-5}$

Table 8: Table of p-values for the second step-down model

Although the only significant p-value here is the one corresponding to `(Intercept)`, we do not remove it for the same reasons as explained in the previous step. Therefore, our final model turns out to be exactly the same as obtained using the step-up approach.

(e) As there is no difference in the two obtained models, we simply choose the model which includes `Wind` and `Temperature` (the one given by the two approaches).

(f)

(g) In order to check whether observation 4 is an outlier, we need to perform a t-test in the mean shift outlier model. We can do this by first augmenting the selected model by adding an explanatory variable $\vec{u}$, where $u_4 = 1$ and $u_i = 0$ for $i \neq 4$. Therefore, our model becomes

$$Y = \mathbf{X}\beta + u\delta + e,$$

where δ is the shifted mean. The null hypothesis for the following test is $\delta = 0$, whereas the alternate hypothesis is $\delta > 0$ (since visually it appears that observation 4 lies above the approximate best fit line).

Performing the t-test we arrive at the following p-values:

Exp. variable	Wind	Temperature	u
p-value	$1.09 \cdot 10^{-8}$	$1.81 \cdot 10^{-13}$	0.128

Table 9: Table of p-values for the mean shift outlier model

As can be seen in the table, the p-values corresponding to the added u variable is not significant. Therefore, we have no reason to reject the null hypothesis and cannot conclude that observation 4 is an outlier.

(h) In order to identify potential leverage points, we compute the diagonal entries of the *hat matrix*. We usually admit points X_i as leverage points if the diagonal entry h_{ii} is greater than $2(p+1)/n$, where p is the number of explanatory variables in the model and n is the number of data points. In this case, we look for values $h_{ii} > 0.2$.

Computing the diagonal entries, we find the following points with $h_{ii} > 0.2$:

Data point	4	8	23	25	30
$h_{i,i}$	0.236	0.277	0.242	0.212	0.242

Table 10: Table of diagonal entries of the hat matrix with values > 0.2

Hence, observations 4, 8, 23, 25, and 30 are leverage points.

In order to find potential influence points, we compute the Cook's distance for the dataset:

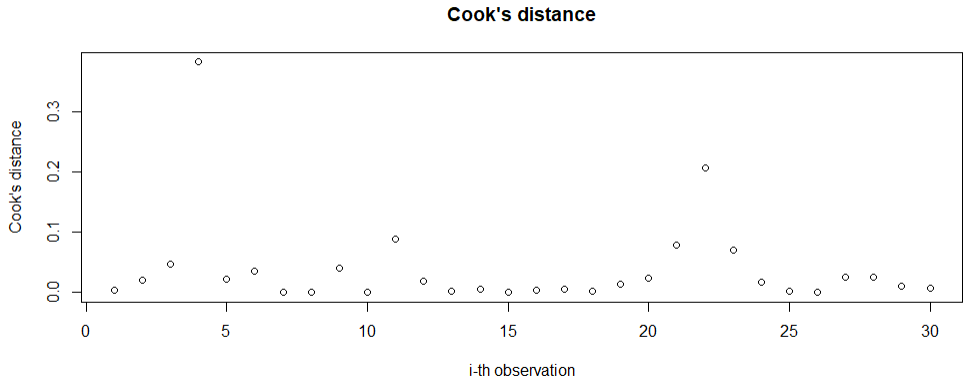

Figure 8: Cook's distance for the dataset

While observations 4 and 22 obviously stand out, neither of them is even close to 1 (being at around 0.35 and 0.2, respectively), which is a safe approximate threshold for an observation to be considered an influence point. Therefore, none of the observations can be considered influence points.

Finally, let us investigate collinearity among the explanatory variables. First, we look at the pairwise correlation:

	Wind	Temperature	Humidity	Insolation
Wind	1.00	-0.50	0.37	-0.32
Temperature	-0.50	1.00	-0.54	0.57
Humidity	0.37	-0.54	1.00	-0.18
Insolation	-0.32	0.57	-0.18	1.00

Table 11: Table of pairwise correlations

As can be seen, there appears to be some correlation between `Wind` and `Temperature`, `Temperature` and `Humidity`, as well as `Temperature` and `Insolation`. None of these values are above 0.6 and hence are not as indicative of the real potential correlation, although there is definitely some collinearity, which is even more noticeable if we look at the plots in **(a)**. However, the correlation between `Temperature` and `Insolation` appears to be somewhat misleading as the data for `Insolation` appears to have significant leverage points.

(i) In order to investigate the normality assumption, we look at the plot of residuals against the response variable, as well as the QQ-plot against the normal distribution:

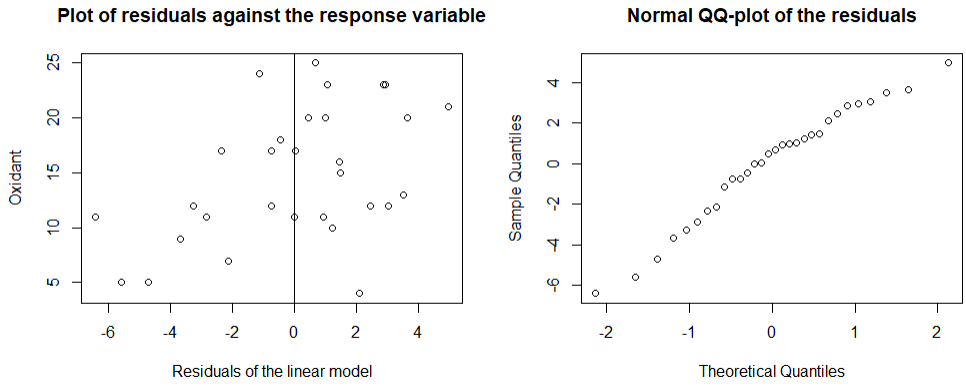

Figure 9: Plots of residuals

As can be seen in the plots, the residuals indeed appear to justify the normality assumption. In the first plot, they seem to be centered around $x = 0$, decreasing in concentration the farther away we go from the origin, while in the QQ-plot they appear to form a reasonably straight line.

(j) Based on all these arguments, it appears that our final model is appropriate for the data. The normality assumption has been verified, it has been checked that there are no influence points, some points have been identified as leverage points, but the required values of the diagonal entries do not appear too big, and we have not identified any particularly problematic problems with collinearity. The final model is given by the formula:

$$Y = \beta_0 + \beta 1 W + \beta_2 T,$$

Where Y is the response variable (`Oxidant`), W is `Wind`, and T is `Temperature`. The estimated parameters are: $\beta_0 = -5.20$, $\beta_1 = -0.43$, and $\beta_0 = 0.52$. The estimated standard error is 2.95 and the final R^2 is computed to be 0.7773.

Exercise 7.3

In order to perform a full regression analysis, we will first look at some plots of the data. Since we will use `expend` as the response variable and `bad`, `crime`, `lawyers`, `employ`, and `pop` as independent variables, we will first plot all of those against `expend`.

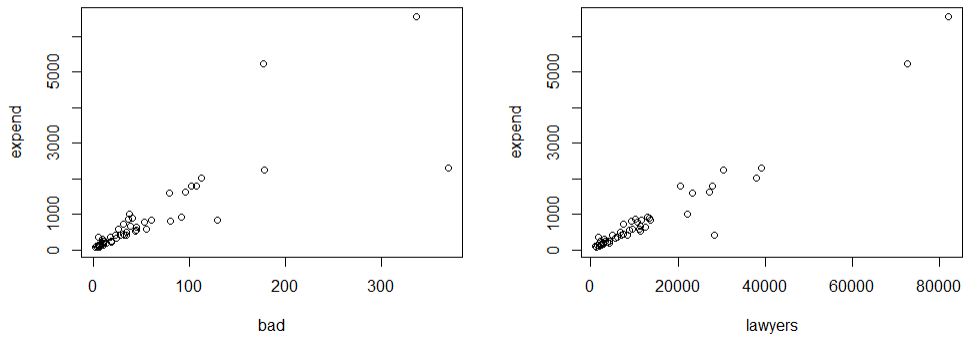

Figure 10: `bad` vs. `expend` and `lawyers` vs. `expend`

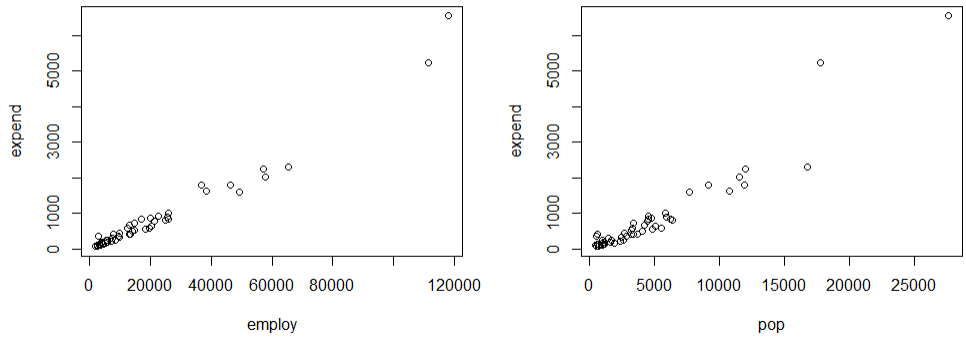

Figure 11: `employ` vs. `expend` and `pop` vs. `expend`

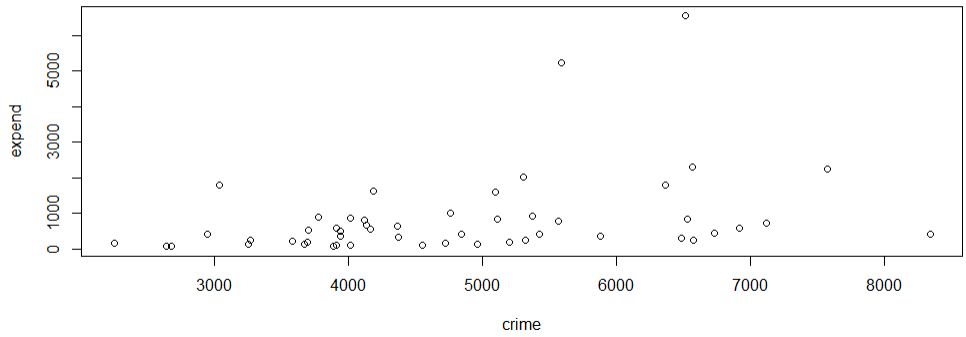

Figure 12: `crime` vs. `expend`

As can be seen, the state expenditure appears to be very linearly dependent on the number of people under judicial supervision (`bad`), number of lawyers in the state (`lawyers`), number of persons employed by and performing services for a government (`employ`), and state population in 1000 (`pop`), but not on crime rate per 100000 (`crime`). Next up, we will plot these 5 variables against each other (using the `pairs` function) in order to look for any collinearity.

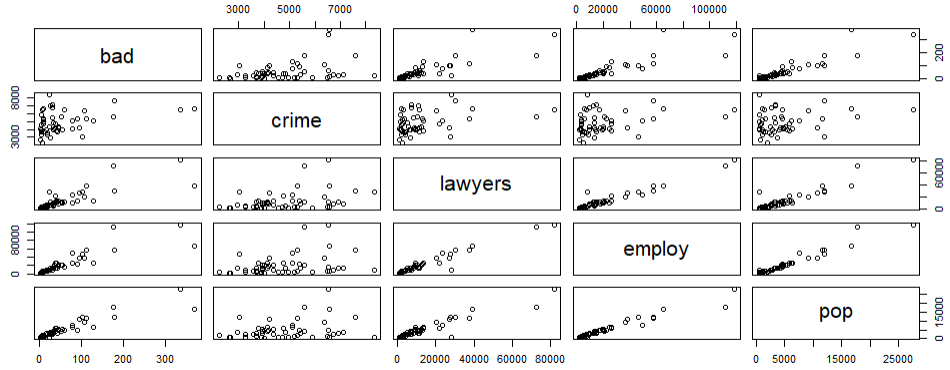

Figure 13: Plots of independent variables against each other

As can be seen, all variables appear to be collinear with each other, aside from the `crime` variable. This should not be particularly surprising, based on the plots before.

Now, we in order to find an appropriate model, we first try the step-up method. Hence, the determination coefficients (R^2) for each single linear model are:

Exp. variable	Bad	Crime	Lawyers	Employ	Pop
R^2	0.6964	0.1119	0.9373	0.9540	0.9073

Table 12: Table of R^2 for simple lin. reg. models

The largest determination coefficient corresponds to `employ` variable. The p-value corresponding to the t-test is smaller than 0.05, so we include `employ` in the model. Next up, we find the following values of R^2 for the models with other variables along the `employ` variable:

Exp. variable	Bad	Crime	Lawyers	Pop
R^2	0.9551	0.9551	0.9632	0.9543

Table 13: Table of R^2 for the first step-up lin. reg. models

The new largest R^2 corresponds to the model with `lawyers` variable added. The p-value corresponding to the

t-test is smaller than 0.05, so we include `lawyers` in the model. Next up, we find the following values of R^2 for the models with other variables along the `employ` and `lawyers` variables:

Exp. variable	Bad	Crime	Pop
R^2	0.9639	0.9632	0.9637

Table 14: Table of R^2 for the second step-up lin. reg. models

The new largest R^2 corresponds to the model with `bad` variable added. The p-value corresponding to the t-test (0.345) is larger than 0.05, so we do not include `bad` in the model. Therefore, our final step-up model uses only `employ` and `lawyers` as explanatory variables.

It turns out that if we repeat the same step-down process as in **Exercise 7.2**, we arrive at the same model. For the sake of brevity, the exact steps are omitted, but they follow the same algorithm.

Looking at the leverage points, we find three points that stand out. In particular, in this case we consider leverage points those observations for which the diagonal entries of the hat matrix exceed $2 * (2 + 1)/51 \approx 0.118$. Therefore, the leverage points are observation 5 ($h_{5,5} = 0.384$), 8 ($h_{8,8} = 0.643$), and 35 ($h_{35,35} = 0.292$).

Looking at the potential influence points, we may conclude that the influence points are those for which the Cook's distance exceeds 1. In particular, those points are again observations 5 ($d_{Cook} = 5.47$) and 8 ($d_{Cook} = 6.38$). Therefore, observations 5 and 8 are both leverage *and* influence points.

Finally, looking at the pairwise correlations between the variables, we find the following variables:

	Bad	Crime	Lawyers	Employ	Pop
Bad	1.00	0.37	0.83	0.87	0.92
Crime	0.37	1.00	0.38	0.31	0.28
Lawyers	0.83	0.38	1.00	0.97	0.93
Employ	0.87	0.31	0.97	1.00	0.97
Pop	0.92	0.28	0.93	0.97	1.00

Table 15: Table of pairwise correlations

As can be seen from the table, there is quite a few very correlated variables. Namely, `Pop` is extremely highly correlated to all variables except for `Crime`, `Employ` is in addition highly correlated with `Bad` and `Lawyers`, while `Lawyers` is also highly correlated with `Bad`. If we also take a look at the variance inflation factors and condition indices, we get even more insight into the data:

Exp. variable	Bad	Crime	Lawyers	Employ	Pop
VIF	8.36	1.49	16.97	33.59	32.94

Table 16: Table of VIF's

Exp. variable	Bad	Crime	Lawyers	Employ	Pop
Condition index	10.66	10.81	31.74	1345.31	171782.48

Table 17: Table of condition indices

As can be seen, the `Employ` and `Pop` variables appear to exhibit high multi-collinearity, which should not be exactly surprising based on the plots at the beginning of the exercise. Also, as can be expected from the plots, `Crime` appears to show very little collinearity with the rest of the variables.

Appendix

```
### Exercise 7.1 ###

source("functions_Ch8.txt")

geese = read.csv("geese.txt", sep="\t")
attach(geese)

# a)

par(mfrow=c(1,2))

plot(observer1,photo,
     xlab='Observer 1 count',
     ylab='Photo count')
plot(observer2,photo,
     xlab='Observer 2 count',
     ylab='Photo count')
mtext('Number of observed vs. photographed geese',
      side=3, line=-3, outer=TRUE)

# b)

observer1lm = lm(photo~observer1)
observer2lm = lm(photo~observer2)

summary(observer1lm)
summary(observer2lm)

# c)

par(mfrow=c(1,2))

plot(observer1lm$residuals, photo,
     xlab='Residuals of observer 1',
     ylab='Number of geese in photos')
abline(v=0)
plot(observer2lm$residuals, photo,
     xlab='Residuals of observer 2',
     ylab='Number of geese in photos')
abline(v=0)

# d)

par(mfrow=c(1,2))

qqnorm(observer1lm$residuals, main='Normal QQ-plot for observer 1')
qqnorm(observer2lm$residuals, main='Normal QQ-plot for observer 2')

ks.test(lm.norm.test(observer1,photo), pnorm)
ks.test(lm.norm.test(observer2,photo), pnorm)

# e)

logphoto = log(photo)
logobs1 = log(observer1)
```

```
logobs2 = log(observer2)

par(mfrow=c(1,2))

plot(logobs1,logphoto,
     xlab='Log observer 1 count',
     ylab='Log photo count')
plot(logobs2,logphoto,
     xlab='Log observer 2 count',
     ylab='Log photo count')
mtext('Log of number of observed vs. log of photographed geese',
      side=3, line=-3, outer=TRUE)

loglm1 = lm(logphoto~logobs1)
loglm2 = lm(logphoto~logobs2)

summary(loglm1)
summary(loglm2)

par(mfrow=c(1,2))

plot(loglm1$residuals, logphoto,
     xlab='Residuals of log(observer1)',
     ylab='Log of number of geese in photos')
abline(v=0)
plot(loglm2$residuals, logphoto,
     xlab='Residuals of log(observer2)',
     ylab='Log of number of geese in photos')
abline(v=0)

qqnorm(loglm1$residuals, main='Normal QQ-plot - Observer 1 residuals')
qqnorm(loglm2$residuals, main='Normal QQ-plot - Observer 2 residuals')

# f)

# g)

### Exercise 7.2 ###

pollution = read.csv("airpollution.txt", sep=" ")
attach(pollution)
alpha = 0.05

# a)

pairs(cbind(oxidant, wind, temperature, humidity, insolation))

# b)

windlm = lm(oxidant~wind)
temperaturelm = lm(oxidant~temperature)
humiditylm = lm(oxidant~humidity)
insolationlm = lm(oxidant~insolation)

summary(windlm)
summary(temperaturelm)
summary(humiditylm)
```

```
summary(insolationlm)

summary(lm(oxidant~wind+temperature))
summary(lm(oxidant~wind+humidity))
summary(lm(oxidant~wind+insolation))

summary(lm(oxidant~wind+temperature+humidity))
summary(lm(oxidant~wind+temperature+insolation))

oxi_stepup_lm = lm(oxidant~wind+temperature)

summary(oxi_stepup_lm)

# c)

lmfull = lm(oxidant~wind+temperature+humidity+insolation)

summary(lmfull)

# d)

summary(lm(oxidant~wind+temperature+humidity))
summary(lm(oxidant~wind+temperature))

oxi_stepdown_lm = lm(oxidant~wind+temperature)

summary(oxi_stepdown_lm)

# e)

plot(wind, oxidant)
plot(temperature, oxidant)

oxi_lm = lm(oxidant~wind+temperature)

# f)

plot(oxidant)
plot(oxidant-oxi_stepdown_lm$residuals, oxidant)
plot(oxi_stepup_lm$fitted.values)

plot(temperature, oxidant)

plot(lm(oxidant~day+wind+humidity+insolation)$residuals,
     lm(temperature~day+wind+humidity+insolation)$residuals,
     main='added varplot for Temperature',
     xlab='RXKXK',
     ylab='RYXK')

plot(lm(oxidant~temperature+wind+humidity+insolation)$residuals,
     lm(day~temperature+wind+humidity+insolation)$residuals,
     main='added varplot for Day',
     xlab='RXKXK',
     ylab='RYXK')

plot(lm(oxidant~day+temperature+humidity+insolation)$residuals,
     lm(wind~day+temperature+humidity+insolation)$residuals,
```

```
     main='added varplot for Wind',
     xlab='RXKXK',
     ylab='RYXK')

plot(lm(oxidant~day+wind+temperature+insolation)$residuals,
     lm(humidity~day+wind+temperature+insolation)$residuals,
     main='added varplot for Humidity',
     xlab='RXKXK',
     ylab='RYXK')

plot(lm(oxidant~day+wind+humidity+temperature)$residuals,
     lm(insolation~day+wind+humidity+temperature)$residuals,
     main='added varplot for Insolation',
     xlab='RXKXK',
     ylab='RYXK')

# g)

plot(oxidant, lm(oxidant~wind+temperature)$residuals)

u = c(rep(0,3),1,rep(0,length(oxidant)-4))
msolm = lm(oxidant~wind+temperature+u-1)
summary(msolm)

# h)

# leverage points
p = 2
n = 30

2*(p+1)/n

hii = round(hatvalues(oxi_lm), 3)

hii[hii>= 0.2]

# influence points
round(cooks.distance(oxi_lm), 2)

plot(cooks.distance(oxi_lm),
     main='Cook\'s distance',
     xlab='i-th observation',
     ylab='Cook\'s distance')

# collinearity
round(cor(pollution[,2:5]), 2)
round(varianceinflation(pollution[,2:5]), 2)
round(conditionindices(pollution[,2:5]), 2)
round(vardecomposition(pollution[,2:5]), 3)

# i)

par(mfrow=c(1,2))

plot(oxi_lm$residuals, oxidant,
     main='Plot of residuals against the response variable',
     xlab='Residuals of the linear model',
```

```
     ylab='Oxidant')
abline(v=0)

qqnorm(oxi_lm$residuals,
       main='Normal QQ-plot of the residuals')

# j)

summary(oxi_lm)
round(oxi_lm$coefficients, 2)

### Exercise 7.3 ###

expenses = read.csv("expensescrime.txt", sep=" ")
attach(expenses)

# plots against 'expend'
par(mfrow=c(1,2))

plot(bad, expend)
plot(lawyers, expend)
plot(employ, expend)
plot(pop, expend)

par(mfrow=c(1,1))

plot(crime, expend)

# looking for collinearity
pairs(cbind(bad, crime, lawyers, employ, pop))

# step-up method
summary(lm(expend~bad))
summary(lm(expend~crime))
summary(lm(expend~lawyers))
summary(lm(expend~employ))  # highest R^2
summary(lm(expend~pop))

summary(lm(expend~employ+bad))
summary(lm(expend~employ+crime))
summary(lm(expend~employ+lawyers))  # highest R^2
summary(lm(expend~employ+pop))

summary(lm(expend~employ+lawyers+bad))  # highest R^2, 'bad' not significant
summary(lm(expend~employ+lawyers+crime))
summary(lm(expend~employ+lawyers+pop))

expend_stepup_lm = lm(expend~employ+lawyers)

# step-down method

summary(lm(expend~bad+crime+lawyers+employ+pop))

# p-val highest for 'crime'

summary(lm(expend~bad+lawyers+employ+pop))
```

```
# p-val highest for 'pop'

summary(lm(expend~bad+lawyers+employ))

# p-val highest for 'bad'

summary(lm(expend~lawyers+employ))

# p-val highest for '(Intercept)', but significant

expend_stepdown_lm = lm(expend~lawyers+employ) # same as step-up

expend_lm = lm(expend~employ+lawyers)

logexpend_lm = lm(log(expend)~log(employ)+log(lawyers))

# residuals versus 'expend'

plot(expend_lm$residuals, expend,
     xlab='Residuals of the linear model',
     ylab='Expend')
abline(v=0)

qqnorm(expend_lm$residuals)

# leverage points
p = 2
n = 51

2*(p+1)/n

hii = round(hatvalues(expend_lm), 3)

hii[hii>= 2*(p+1)/n]

# influence points
expend_cook = round(cooks.distance(expend_lm), 2)

expend_cook[expend_cook >= 1]

plot(cooks.distance(oxi_lm),
     main='Cook\'s distance',
     xlab='i-th observation',
     ylab='Cook\'s distance')

# collinearity
round(cor(expenses[,3:7]), 2)
round(varianceinflation(expenses[,3:7]), 2)
round(conditionindices(expenses[,3:7]), 2)
round(vardecomposition(expenses[,3:7]), 3)
```