Lab 11 - Multiple Linear Regression

ENVX1002 Handbook

Semester 1, 2025

Learning outcomes

- Learn to perform MLR and interpret the results using R;
- Undertake hypothesis testing to determine if model is significant
- Undertake hypothesis testing to determine if the true partial regression slope $\neq 0$
- Check assumptions are filled prior to assessing model output
- Assess model summary in terms of fit and P-values
- Consider more parsimonious models

Before you begin

Create your Quarto document and save it as `Lab-11.Rmd` or similar. The following data files are required:

1) ENVX1002_wk11_practical_data_Regression.xlsx

Last week you explored simple linear regression and assessed the output of your models.

This week we will build upon this and venture into multiple linear regression.

Before you begin, ensure you have a project set up in your desired folder. Then open up a fresh R markdown and save the file within this folder.

Don't forget to save as you go!

Exercise 1: Corn yields

Data: *Corn* spreadsheet

In this data

- y = P content of corn
- x_1 = inorganic P content of soil
- x_2 = organic P content of soil
- n = 17 sites (The original data had 18 sites, one is removed here.)

Aim of our investigation: Understand the relationship between Organic and Inorganic phosphorous contents in the soil, and the phosphorous content of corn. This will allow us to see which type of phosphorous is being taken up by the corn.

```
library(readxl)
Corn <- read_xlsx("data/ENVX1002_practical_wk11_data_Regression.xlsx", "Corn")
head(Corn)
```

```
# A tibble: 6 × 3
  CornP InorgP  OrgP
  <dbl>  <dbl> <dbl>
1    64    0.4    53
2    60    0.4    23
3    71    3.1    19
4    61    0.6    34
5    54    4.7    24
6    77    1.7    65
```

1.1 Examine the correlations

Some people find it difficult to visually interpret graphical summaries of data in more than 2 dimensions; however, 3-dimensional surface plots are reasonably common in statistics although not usually in descriptive statistics.

Initially we will examine the pairwise correlations to "get a feel" for the data. We will then make a 3-dimensional surface plot using the `lattice` package .

Using R, we can calculate the correlation matrix quite easily.

Note the use of `round()` to limit the number of significant digits.

```
round(cor(Corn),3)
```

```
       CornP InorgP  OrgP
CornP  1.000  0.720 0.212
InorgP 0.720  1.000 0.399
OrgP   0.212  0.399 1.000
```

a) What do the results of the correlation matrix tell you?

Solution

---

a) These results tell you that the correlation between CornP and InorgP the highest (r = 0.720) The correlation between CornP and OrgP (r = 0.212) or between OrgP and InorgP (r = 0.399) is much weaker.

---

b) Based on the correlation matrix, if we were to fit a single predictor model involving EITHER InorgP OR OrgP, then which model would be more successful?

(Hint, the r^2 is exactly that for a single predictor regression, the square of the correlation, r).

Solution

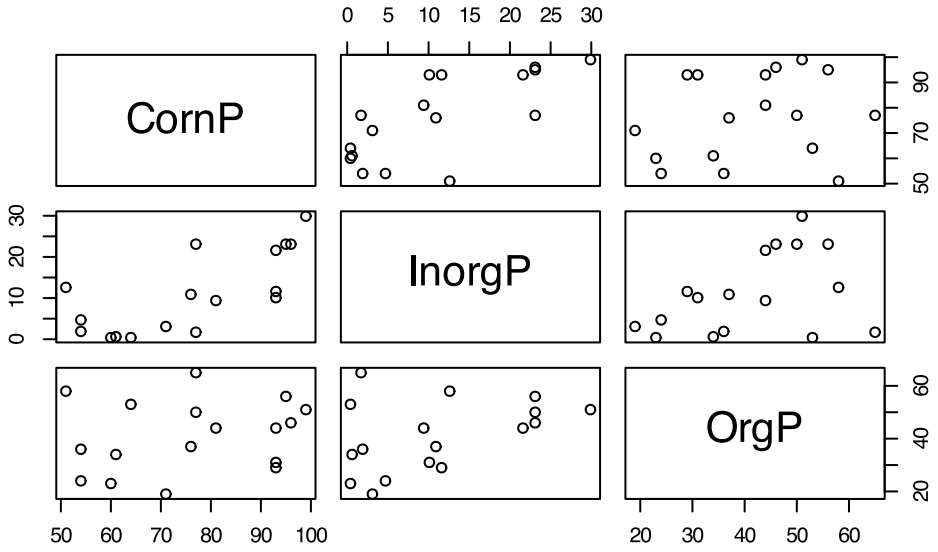
---

b) The more successful model would be $CornP = \beta_0 + \beta_1 InorgP$ as the correlation between CornP and InorgP is much higher than the relationship than CornP and OrgP.

The pairs plot creates scatterplots between each possible pair of variables. Like a single scatterplot the pairs plot allows us to visually observe any trends.

c) Observing the pairs plot below, do you see any strong trends? How well does this link to your correlation matrix?

```
pairs(Corn)
```



Solution

---

c) We read the pairs plot like a correlation matrix. We can observe a slight trend between CornP and InorgP. The plots of CornP and OrgP, and that of OrgP and InorgP show much weaker trends and a more even scatter of points.

Simple 3-D plot

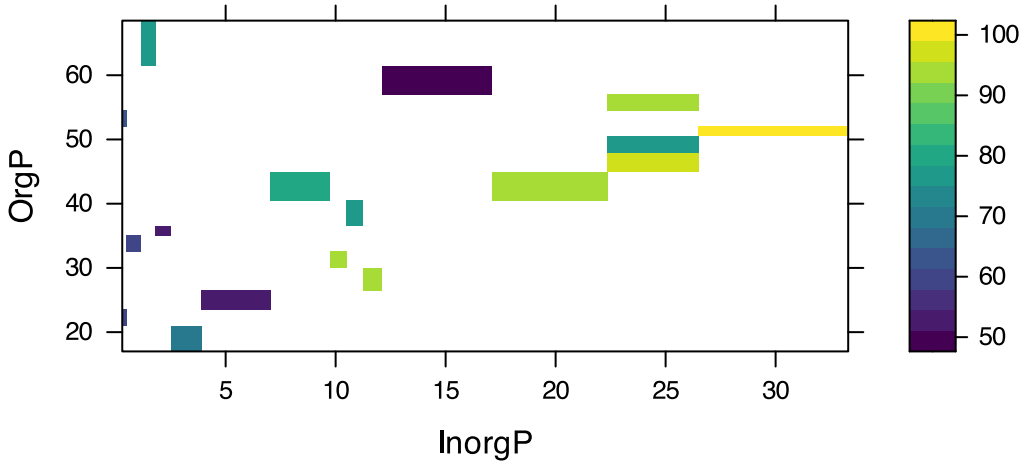
Unlike a simple plot we can creat for a simple linear regression, it is a bit more complex to visualise a model with more predictors. One way we can visualise the relationship is with a 3-D plot, which can be made using the function `levelplot()` in lattice.

Here we plot the OrgP and InorgP in the axes and the levels in the plot are CornP.

Note the package Viridis has been called, this is through personal choice.

The Viridis package has a range of assembled colour ramps which are easier for the reader to differentiate the colours, especially when printed in grayscale, or if the reader is colourblind.

```
library(lattice, quiet = T)
library(viridis, quiet = T) # will need to install.packages("viridis")

levelplot(CornP ~ InorgP + OrgP, data = Corn
          , col.regions = viridis(100))
```



The level plot shows us the x (InorgP) and y variable (OrgP), with the colour scale representing the z variable, which in this case is Phosphorous being taken up by the corn (cornP). From the plot we can see that with higher levels of OrgP and InorgP in the soil, the Phosphorous content in the corn is generally higher.

It is clear that the 3-D surface plot does not have colours everywhere, but this relates of course to the underlying data. In this case we don't have continuous data in both directions, so the response (the colour) is only plotted where we have input variables.

If we did have continuous data in both directions the plot would look more like a heatmap, here are some examples.

1.2 Fit the model

We will now use regression to estimate the joint effects of both inorganic phosphorus and organic phosphorus on the phosphorus content of corn.

$CornP = \beta_0 + \beta_1 InorgP + \beta_2 OrgP + error$

This is fairly simple and follows the same structure as simple linear regression and uses `lm()`.

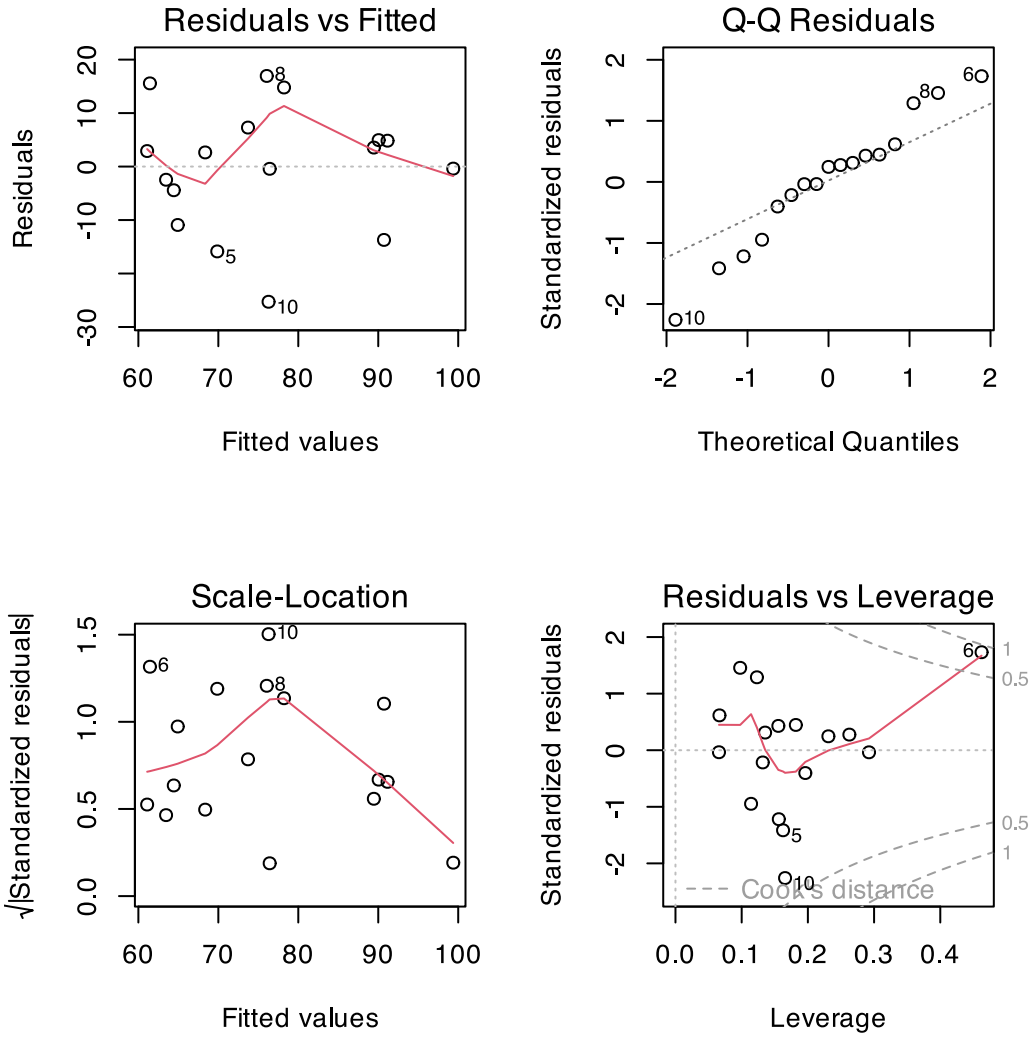
```
MLR.Corn <- lm(CornP ~ InorgP + OrgP, data=Corn)
```

1.3 Check assumptions

Let's check the assumptions of regression are met via residual diagnostics.

a) Are there any apparent problems with normality of CornP residuals or equality of variance for this small data set?

```
par(mfrow=c(2,2))
plot(MLR.Corn)
```



Solution

---

a) No there is not as the distribution is approximately normal and the variance is constant. In the residuals vs Leverage plot there is a point sitting close to the top right corner, but this is most likely because our dataset is small.

Interpretation of the assumptions is quite subjective and often comes down to time and experience. If, for example in your Project 3 it is not obvious whether the assumptions have been met, make your decision and provide justification so that the viewer is able to come to the same conclusion as you. e.g. if you have a smaller dataset as you see in this exercise, it may be harder to tell whether assumptions have been met.

1.4 Model output

After checking our assumptions and we are happy with them, we can interpret our model output.

a) Incorporating the partial regression coefficient estimates, what is the model equation?

```
summary(MLR.Corn)
```

```
Call:
lm(formula = CornP ~ InorgP + OrgP, data = Corn)

Residuals:
    Min      1Q  Median      3Q     Max
-25.282  -4.428   2.645   4.949  16.946

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  66.4654     9.8496   6.748 9.35e-06 ***
InorgP        1.2902     0.3428   3.764  0.00209 **
OrgP         -0.1110     0.2486  -0.447  0.66195
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.25 on 14 degrees of freedom
Multiple R-squared:  0.5253,    Adjusted R-squared:  0.4575
F-statistic: 7.746 on 2 and 14 DF,  p-value: 0.005433
```

Solution

---

a) CornP = 66.4654 + 1.2902 * InorgP - 0.1110 * OrgP

Simple linear regression allowed us to describe the relationship incorporating our regression coefficient estimate. We would interpret it as follows:

"As x increases by 1, y decreases by b_1 units" (depending on the direction of the relationship).

This week it is a bit different because we are dealing with **partial** regression coefficients instead.

Instead, we would say:

"as x_1 increases by 1, y decreases by b_1 units given all other partial regression coefficients are held constant".

Applied to our model, if we wanted to describe the relationship between InorgP and CornP, we would say:

"As InorgP increases by 1, CornP increases by 1.2902, given OrgP is held constant."

b) Given the above, how would you interpret the relationship between OrgP and CornP?

Solution

---

b) As OrgP increases by 1, CornP decreases by 0.1110, given InorgP is held constant.

1.5 Is the model useful?

Remember now that the F-test and T-test are testing slightly different things.

F-test

H_0 : all $\beta_k = 0$, i.e. $\beta_1 = \beta_2 = 0$

H_1 : at least 1 $\beta_k \neq 0$, i.e. our model is significant

We will find the P-value for this test at the end of the summary output.

t-test

$H_0 : \beta_k = 0$

$H_1 : \beta_k \neq 0$

Where β_k refers to one of the model partial regression coefficients. In the case of our model we only have 2: b_1 (InorgP) and b_2 (OrgP).

Now it is your turn:

a) Looking at the `summary()` output, is our overall model significant?

Solution

---

a) Our model is significant as the P-value derived from the F-test statistic is less than 0.05 (P = 0.005). As P < 0.05, we reject our null hypothesis and can say at least one of our predictors are significant.

---

b) Which independent variable is a significant predictor of corn yield?

Solution

---

b) InorgP (P = 0.002)

1.6 How good is the model?

a) How much of the variation in CornP content is explained by the two independent variables?

Solution

---

a) This question refers to the r-squared adjusted. As the adjusted r-squared = 0.4575, we can say that the model (containing our two independent variables) explains 45.8% of variation in CornP.

---

b) Run the model again but this time with *only* the significant independent variable. How do the model performance criteria (r^2-adj, P-values, Residual Standard Error) change?

Solution

---

b) By removing OrgP (deemed non-significant as P = 0.662) and only retaining InorgP (P = 0.002), we can see that the r-squared adjusted has increased to 0.4864, indicating an improvement in model fit.

1.7 Our conclusions

Writing up conclusions for multiple linear regression are similar to simple linear regression, just with a couple of extra P-values to state.

We would first mention that our overall model is significant as we rejected the null hypothesis (P = 0.05). We could then describe the hypothesis test results for our predictor variables. Finally, we would describe the model fit and our adjusted-r^2.

Remember the scientific conclusion then relates our findings back to the context, answering aims.

a) What would our statistical conclusion be?

Solution

---

a) Statistical conclusion:

The F-test indicated our model is significant (P < 0.05), allowing us to reject the null hypothesis and conclude at least one of our partial regression coefficients has a slope not equal to 0.

Our t-test upon each of our partial regression coefficients supports this, as the P-value for InorgP was 0.002, much smaller than 0.05 and we can therefore reject the null hypothesis that this coefficient is equal to 0. In contrast, the P-value for OrgP was greater than 0.05 (P = 0.662), and so we fail to reject the null hypothesis, concluding OrgP is not a significant predictor in this model.

The fit of this model is moderate, with a residual standard error of 12.25 and an r^2-adj = 0.4575.

Furthermore this fit could be improved by the removal of the OrgP predictor within the model (r^2-adj = 0.4864).

---

b) What would our Scientific conclusion be?

Solution

---

b) Scientific conclusion:

Inorganic phosphorous is a significant predictor of phosphorous content in corn ($P < 0.05$), whereas organic phosphorous is not ($P > 0.05$). The model accounts for 45.8% of variation in the phosphorous content of corn.

Exercise 2: Water quality

Data: *Microbes* spreadsheet

This exercise will use data from the NOAA Fisheries data portal. The dataset contains the results of microbial study of Pudget Sound, an estuary in Seattle, U.S.A.

The dataset contains the following variables:

- total_bacteria = Total bacteria (cells per ml) –> this will be our response variable
- water_temp = Water temperature (℃)
- carbon_L_h = microbial production (µg carbon per L per hour)
- NO3 = Nitrate (µm)

First thing to do, is read in the data. This time we will be using the *Microbes* spreadsheet:

```
mic      <-      read_xlsx("data/ENVX1002_practical_wk11_data_Regression.xlsx",
"Microbes")

# Check the structure
str(mic)
```

```
tibble [55 × 4] (S3: tbl_df/tbl/data.frame)
 $ total_bacteria: num [1:55] 498353 529135 529911 603798 632847 ...
 $ water_temp    : num [1:55] 11.2 8.5 8.6 11.2 8.4 8.9 11.3 8.5 9 8.3 ...
 $ carbon_L_h    : num [1:55] 0.286 0.37 0.324 0.287 0.637 ...
 $ NO3           : num [1:55] 21.1 23.5 23.5 21.8 23 ...
```

2.1 Examine the correlations

For this dataset we may expect to see some correlations;

- Warmer water temperature we would expect to see a higher amount of bacterial growth

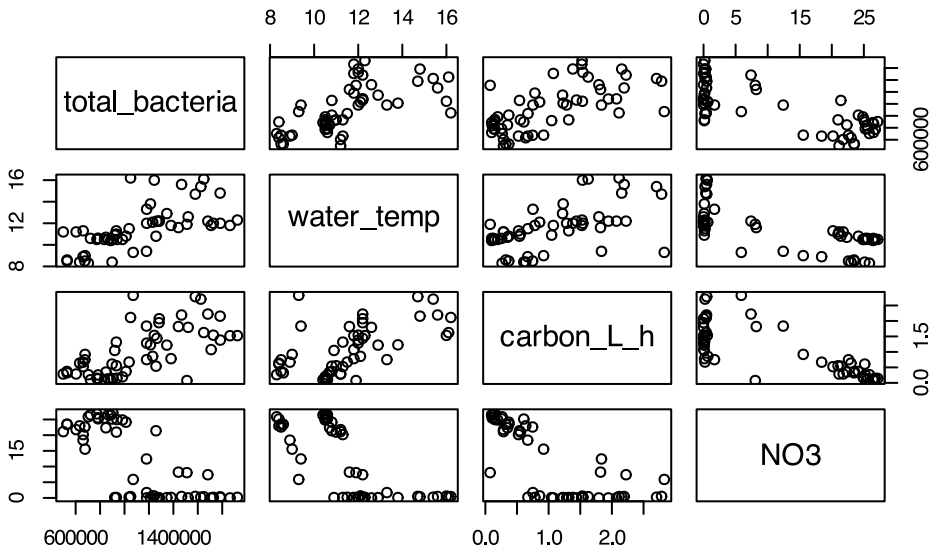
- Carbon is a proxy for microbial production, so if we see a higher rate of carbon production, we would expect to see higher levels of bacteria
- NO3 (Nitrate) is an essential nutrient for plants and some bacteria species metabolise this

a) Let's test this. Observe the correlation matrix and pairs plots. Do you notice any strong correlations?

```
cor(mic)
```

```
               total_bacteria water_temp carbon_L_h        NO3
total_bacteria      1.0000000  0.6445878  0.6629503 -0.7587849
water_temp          0.6445878  1.0000000  0.5883947 -0.6958635
carbon_L_h          0.6629503  0.5883947  1.0000000 -0.7653497
NO3                -0.7587849 -0.6958635 -0.7653497  1.0000000
```

```
pairs(mic)
```



Solution

---

a) The strongest correlation with total_bacteria is with NO3 (r = −0.759), followed by Carbon (r = 0.663) and water_temp (r = 0.645).

Observing the pairs plot, there is something weird happening with NO3, so there may be another kind of relationship occurring here.

There are also some correlations between independent variables, but they are not strong enough to exclude from the model.

2.2 Fit the model

We can now fit the model to see how much these predictors account for the variation in total bacteria.

$$totalbacteria = \beta_0 + \beta_1 watertemp + \beta_2 carbon + \beta_3 NO3 + error$$

There are two forms the lm code can take; you can either specify which variables you want to include by naming each one, or if only your desired variables are within your dataset, you can use the ~. to specify all columns.

```
names(mic) # tells us column names within the dataset
```

```
[1] "total_bacteria" "water_temp"     "carbon_L_h"     "NO3"
```

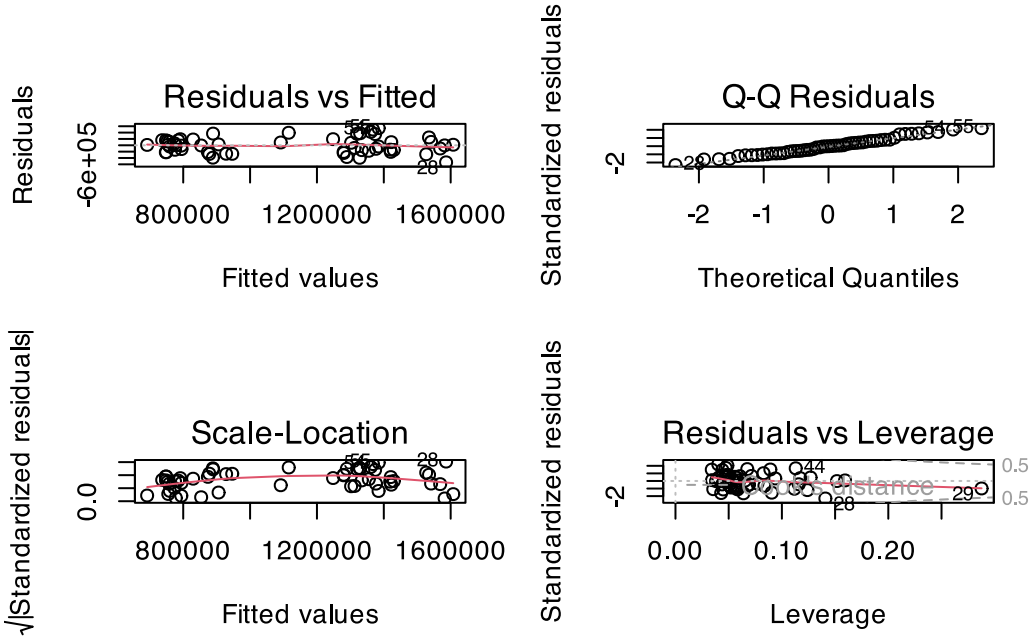
```
# Form 1:
mic.lm <- lm(total_bacteria ~ water_temp + carbon_L_h + NO3, data = mic)

# Form 2
mic.lm <- lm(total_bacteria ~ ., data = mic)
```

2.3 Check assumptions

Let's check the assumptions of regression are met via residual diagnostics.

a) Are there any apparent problems with normality of total_bacteria residuals or equality of variance for this data set?

```
par(mfrow=c(2,2))
plot(mic.lm)
```



Solution

---

a) Residuals vs fitted looks ok, as does normality.

There seems to be some slight clustering but most likely due to the data being obtained from two sampling periods (both within the same season), or potentially from different points. Water quality and microbial activity is highly variable, so this is not entirely unexpected. Nevertheless we can move on with this in mind.

Note this question did not mention the Residuals vs Leverage plot. There is one value in the top right corner outside the first threshold. This indicates there may be a potential outlier within the predictors influencing the regression model, or maybe the current model is not the best for this data.

You are welcome to investigate this further, but for the sake of our example, we will move onto interpretation of the output.

2.4 Model output

After investigating the assumptions, they seem to be ok, so we can move onto the model summary.

```
summary(mic.lm)
```

```
Call:
```

```
lm(formula = total_bacteria ~ ., data = mic)

Residuals:
    Min      1Q  Median      3Q     Max
-536948 -191145   -2584  154144  539394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   766294     314957   2.433  0.01852 *
water_temp     40051      23547   1.701  0.09505 .
carbon_L_h     84425      67427   1.252  0.21624
NO3           -16586       5256  -3.156  0.00268 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 250500 on 51 degrees of freedom
Multiple R-squared:  0.614, Adjusted R-squared:  0.5913
F-statistic: 27.04 on 3 and 51 DF,  p-value: 1.318e-10
```

a) Incorporating the partial regression coefficient estimates, what is the equation for this model?

Solution

---

a) total_bacteria = 766294 + 40051 * water_temp + 84425 * carbon_L_h - 16586 * NO3

---

b) Like you did in Exercise 1.4, how would you interpret the relationship between total_bacteria and water_temp?

Solution

---

b) As water_temp increases by 1, total_bacteria increases by 40051, given all other partial regression coefficients remain constant.

2.5 Is the model useful?

a) Observing the P-value of the F-statistic in the summary, can we say our model is significant?

Solution

---

a) The P-value for the F-statistic is 1.318e-10, which is smaller than 0.05. We therefore reject the null hypothesis and can say our model is indeed significant as at least one of our predictors are significant.

---

b) Are any predictors significant?

Solution

---

b) Yes; as the P-value for N03 is 0.00268 (P < 0.05), we can reject the null hypothesis that the true partial regression slope for this variable is not zero.

Water_temp and carbon_L_h are not significant predictors (P > 0.05), i.e. as we cannot reject the null hypothesis that the true partial regression slope of each is non-zero.

2.6 How good is the model?

a) How much of the variation in total bacteria is explained by the three independent variables?

Solution

---

a) The model is not the best; the Residual Standard error is huge, even in terms of the response variable, and the r^2-adjusted is only 0.5913.

The model accounts for 59.1% of variation in total bacteria counts.

---

b) Run the model again but this time excluding the variable with the largest P-value. How do the model performance criteria (r^2-adj, P-values, Residual Standard Error) change?

Solution

---

b) The least significant predictor is carbon_L_h (P = 0.21624).

After removing carbon_L_h, the r^2-adj has decreased (r^2-adj = 0.5868) and Residual Standard Error has increased (251900). In terms of p-values, that of water_temp has decreased, but is still non-significant. The P-value for NO3 has decreased substantially.

From this, we can see the model has not been improved by eliminating the predictor.

```r
summary(lm(total_bacteria ~ water_temp + NO3, data = mic))
```

```
Call:
lm(formula = total_bacteria ~ water_temp + NO3, data = mic)

Residuals:
    Min      1Q  Median      3Q     Max
-506706 -203093  -11950  157707  533145

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   858466     307901   2.788  0.00739 **
water_temp     43611      23502   1.856  0.06917 .
NO3           -20620       4175  -4.938 8.54e-06 ***
```

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

Residual standard error: 251900 on 52 degrees of freedom
Multiple R-squared:  0.6021,    Adjusted R-squared:  0.5868
F-statistic: 39.34 on 2 and 52 DF,  p-value: 3.927e-11
```

2.7 Conclusions

a) What would the statistical conclusion be for this model?

Solution

---

a) Statistical conclusion:

The F-test indicated our model is significant ($P < 0.05$), allowing us to reject the null hypothesis and conclude at least one of our partial regression coefficients has a slope not equal to 0.

Our t-test upon each of our partial regression coefficients supports this, as the P-value for NO3 was 0.003, smaller than 0.05 and we can therefore reject the null hypothesis that this coefficient is equal to 0. In contrast, the P-value for water_temp and carbon_L_h were greater than 0.05 ($P = 0.095$ and 0.216, respectively), and so we fail to reject the null hypothesis, concluding water_temp and carbon_L_h were not significant predictors within this model.

The fit of this model is moderate, with a residual standard error of 250500 and an $r^2 = 0.5913$.

Note: remember residual standard error is on the same scale as the response variable, so although it is a huge number, it is not as large when we think of the scale.

Furthermore this fit was not substantially improved by the removal of the water_temp predictor within the model (r^2-adj = 0.587).

---

b) What would our scientific conclusion be?

Solution

---

b) Scientific conclusion:

Nitrate is a significant predictor of total bacteria count in ($P < 0.05$), whereas water temperature and Carbon production rate were not is not in this case ($P > 0.05$). The model accounts for 59.1% of variation in the bacteria counts measured in water samples collected from Puget Sound.

Exercise 3: Dippers

Data: *Dippers* spreadsheet

We will revisit the Dippers dataset from last week, but now incorporating other factors which may be influencing the distribution.

The file, Breeding density of dippers, gives data from a biological survey which examined the nature of the variables thought to influence the breeding of British dippers.

Dippers are thrush-sized birds living mainly in the upper reaches of rivers, which feed on benthic invertebrates by probing the river beds with their beaks.

Twenty-two sites were included in the survey. Variables are as follows

- water hardness
- river-bed slope
- the numbers of caddis fly larvae
- the numbers of stonefly larvae
- the number of breeding pairs of dippers per 10 km of river

In the analyses, the four invertebrate variables were transformed using a Log(Number+1) transformation.

Now it is your turn to work through the steps as above. What other factors are influencing the number of breeding pairs of Dippers?

a) Read in the data from today's Excel sheet, the corresponding sheet name is "Dippers"

Solution

---

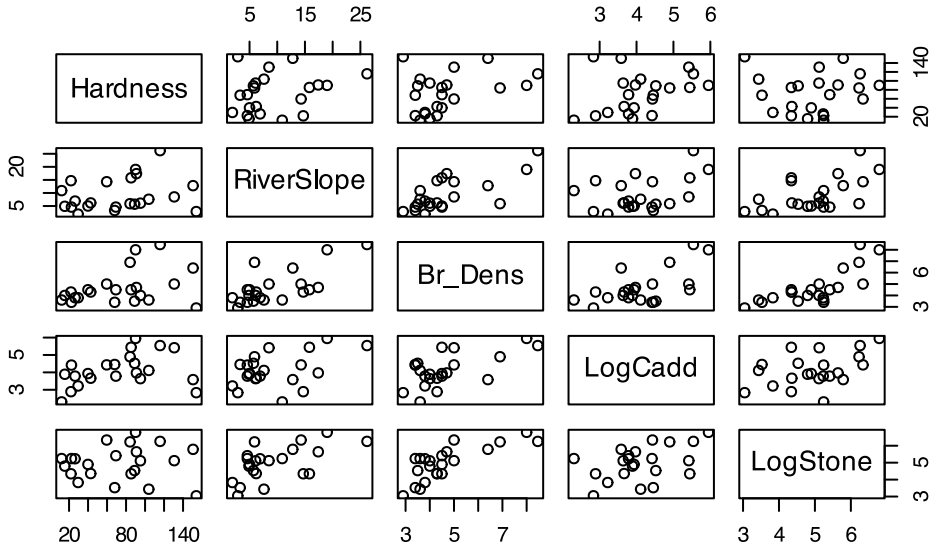
a) Use the following code:

```
library(readxl, quietly = TRUE)
Dippers <- read_xlsx("data/ENVX1002_practical_wk11_data_Regression.xlsx"
                     ,"Dippers")
```

---

b) Investigate a correlation matrix and pairs plot of the dataset, are there signs of a relationship between breeding pair density and other independent variables?

```
pairs(Dippers)
```




```
cor(Dippers)
```

```
             Hardness RiverSlope   Br_Dens   LogCadd   LogStone
Hardness   1.00000000  0.2420874 0.3503482 0.3337586 0.02951218
RiverSlope 0.24208735  1.0000000 0.7102216 0.4313081 0.57479242
Br_Dens    0.35034817  0.7102216 1.0000000 0.6126891 0.76294367
LogCadd    0.33375857  0.4313081 0.6126891 1.0000000 0.44347440
LogStone   0.02951218  0.5747924 0.7629437 0.4434744 1.00000000
```

Solution

---

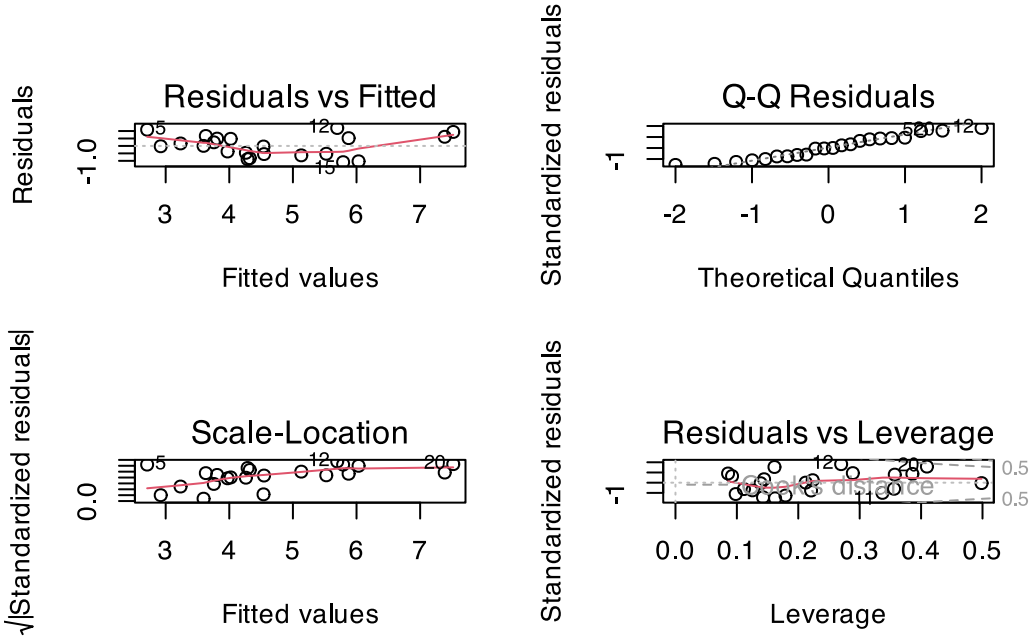
b) Looking at the pairs scatterplots there seems to be a positive relationship between Br_Dens and RiverSlope, LogCadd, and with LogStone. We can follow this up using correlation, which indicates there is a moderate positive relationship of 0.710 between Br_Dens and RiverSlope, 0.763 Br_Dens and LogStone, and 0.613 between Br_Dens and LogCadd.

---

c) Let's investigate further. Run the model incorporating all of our predictors, but before looking at our model output, are the assumptions ok?

```
# Run model
dipper.lm <- lm(Br_Dens ~ ., data=Dippers)

# Check assumptions
```

```
par(mfrow = c(2,2))
plot(dipper.lm)
```



Solution

---

c) Assumptions look ok:

- Residuals vs fitted: Points evenly scattered around mean.
- Normal Q-Q: Points follow line.
- Scale-Location: Points evenly scattered, no fanning.
- Residuals vs Leverage: No points occurring at top right or bottom right corners outside the dotted red lines. A couple of points are close, but not outside, so all is good.

We can continue to interpreting the output!

Once you are happy assumptions are good, you can interpret the model output using `summary()`.

Solution

```
summary(dipper.lm)
```

```
Call:
lm(formula = Br_Dens ~ ., data = Dippers)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08910 -0.54070 -0.00874  0.51623  1.20326

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.530702   1.057244  -1.448  0.16586
Hardness     0.006916   0.004315   1.603  0.12740
RiverSlope   0.067924   0.034769   1.954  0.06741 .
LogCadd      0.317611   0.219858   1.445  0.16674
LogStone     0.752987   0.220851   3.409  0.00334 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7753 on 17 degrees of freedom
Multiple R-squared:  0.7799,    Adjusted R-squared:  0.7281
F-statistic: 15.06 on 4 and 17 DF,  p-value: 1.971e-05
```

---

d) What is the equation for our model, incorporating our partial regression coefficients?

Solution

---

d) Br_Dens = −1.530702 + 0.006916 * Hardness + 0.067924 * RiverSlope + 0.317611 * LogCadd + 0.752987 * LogStone

---

e) Based on the F-statistic output, is the model significant? How can we tell? Is it different to the significance of LogCadd this time?

Solution

---

e) The F-test statistic tells us whether our overall model is significant, which in this case it is (P = 1.971e-05). This time, unlike when we only used LogCadd as the predictor, our P-value from the F-Statistic is now different to the P-value of LogCadd (P = 0.16674).

As our P - value for the model is < 0.05, we can reject the null hypothesis and conclude our model is significant and at least one of our partial regression slopes are non-zero.

---

f) Is LogCadd still a significant predictor of Dipper breeding pair density?

Solution

f) Unlike last week, the P-value for LogCadd is now greater than 0.05 (P = 0.16674), and so we fail to reject the null hypothesis that the partial regression slope is equal to zero. We can therefore say LogCadd not a significant predictor of Dipper breeding pair density.

---

g) What are the significant predictors of this model?

Solution

---

g) The only significant predictor for this model is LogStone (P < 0.05).

---

h) How good is the fit of our model?

Solution

---

h) Our model fit is much better compared to the simple linear model from last week.

Our residual standard is (Residual SE = 0.7753) and the r^2-adjusted is closer to 1 than 0 (r^2-adj = 0.7281).

Our residual SE is a low value, and relative to the range of Br_Dens (min = 2.9, max = 8.4), which is pretty good and also much better than last week.

---

i) What might we do to improve the model fit?

Solution

---

i) Consider elimination of a non-significant predictor variable.

---

j) What statistical and scientific conclusions can we make from this model output?

Solution

---

j) Statistical:

As the F-test statistic is less than 0.05, we can reject the null hypothesis that all partial regression coefficients are equal to zero. Furthermore, LogStone was the only significant predictor (P < 0.05) and therefore we can only reject the null for this predictor.

The model fit is good, with an r^2-adjusted value of 0.7281 and Residual standard error of 0.7753.

Scientific:

The number of Stonefly Larvae is a significant predictor of Dipper breeding pair density (P < 0.05), whereas other predictors used, such as water hardness, river slope and the number of Caddis Fly

larvae are not significant. The overall model accounts for 72.8% of variation in Dipper breeding pair density.

The significance of Stonefly larvae over Caddis fly larvae, and the positive relationship in the partial regression coefficient, suggests that Dippers may have a preference for Stonefly Larvae over Caddis Fly larvae.

Another week of linear models done! Great work fitting Multiple Linear Regression! Next week we break away and explore non-linear functions.

Bonus Take Home Exercises

Use the template below to test the hypotheses for each exercise.

1. Scatterplot and correlation
2. Fit the model
3. Check assumptions
4. P-value and model fit
   a) Is the model significant?
   b) Are the predictors significant?
   c) How good is the model fit?
5. Conclusions

Exercise 1: House Prices

Data: - Housing

This exercise will use a dataset from kaggleto explore the variables affecting house prices.

Solution

---

1. State Aim: > To
2. Scatterplot and correlation

```
#Read in data
housing<- read_csv("data/Housing.csv")
```

```
Rows: 5000 Columns: 7
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (1): Address
dbl (6): area_income, area_house_age, room_no, bedroom_no, area_pop, Price
```

```
ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
str(housing) #There is a character variable that will mess with the analysis
```

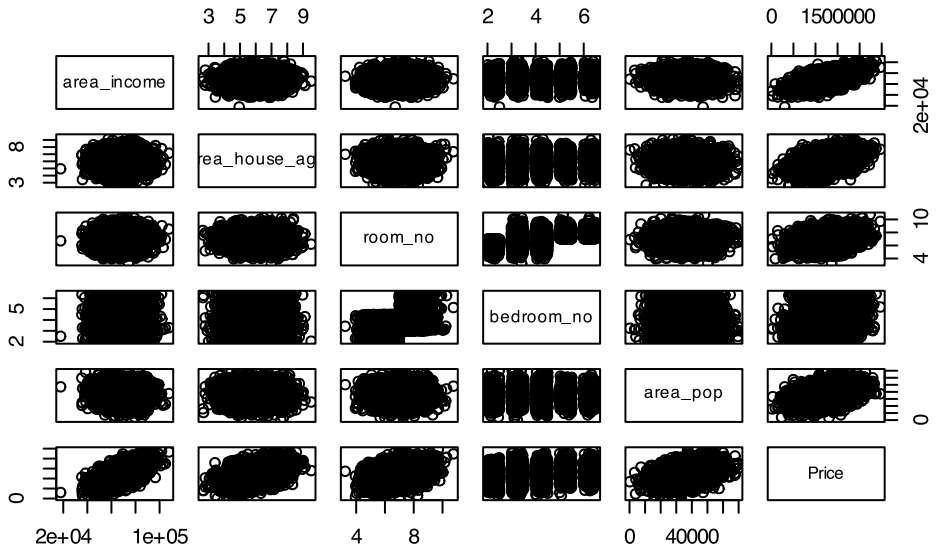
```
spc_tbl_ [5,000 × 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ area_income   : num [1:5000] 79545 79249 61287 63345 59982 ...
 $ area_house_age: num [1:5000] 5.68 6 5.87 7.19 5.04 ...
 $ room_no       : num [1:5000] 7.01 6.73 8.51 5.59 7.84 ...
 $ bedroom_no    : num [1:5000] 4.09 3.09 5.13 3.26 4.23 4.04 3.41 2.42 2.3
6.1 ...
 $ area_pop      : num [1:5000] 23087 40173 36882 34310 26354 ...
 $ Price         : num [1:5000] 1059034 1505891 1058988 1260617 630943 ...
 $ Address          : chr [1:5000] "208 Michael Ferry Apt. 674\nLaurabury,
NE 37010-5101" "188 Johnson Views Suite 079\nLake Kathleen, CA 48958" "9127
Elizabeth Stravenue\nDanieltown, WI 06482-3489" "USS Barnett\nFPO AP 44820" ...
 - attr(*, "spec")=
  .. cols(
  ..   area_income = col_double(),
  ..   area_house_age = col_double(),
  ..   room_no = col_double(),
  ..   bedroom_no = col_double(),
  ..   area_pop = col_double(),
  ..   Price = col_double(),
  ..   Address = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
```

```
#Clean data
housing<- housing%>%
  select(-Address)

str(housing)
```

```
tibble [5,000 × 6] (S3: tbl_df/tbl/data.frame)
 $ area_income   : num [1:5000] 79545 79249 61287 63345 59982 ...
 $ area_house_age: num [1:5000] 5.68 6 5.87 7.19 5.04 ...
 $ room_no       : num [1:5000] 7.01 6.73 8.51 5.59 7.84 ...
 $ bedroom_no    : num [1:5000] 4.09 3.09 5.13 3.26 4.23 4.04 3.41 2.42 2.3
6.1 ...
 $ area_pop      : num [1:5000] 23087 40173 36882 34310 26354 ...
 $ Price         : num [1:5000] 1059034 1505891 1058988 1260617 630943 ...
```

Looking at the correlation matrix and the plots, it looks like the average number of bedrooms for a house in the area (`bedroom_no`) is fairly correlated to the average number of rooms for a house in the area (`room_no`). Additionally, `bedroom_no` has a very weak relationship to Price, so I will leave it out of the model to try to keep all the variables independent from each other. Everything else will go in the model.

```
pairs(housing)
```




```
round(cor(housing), 3)
```

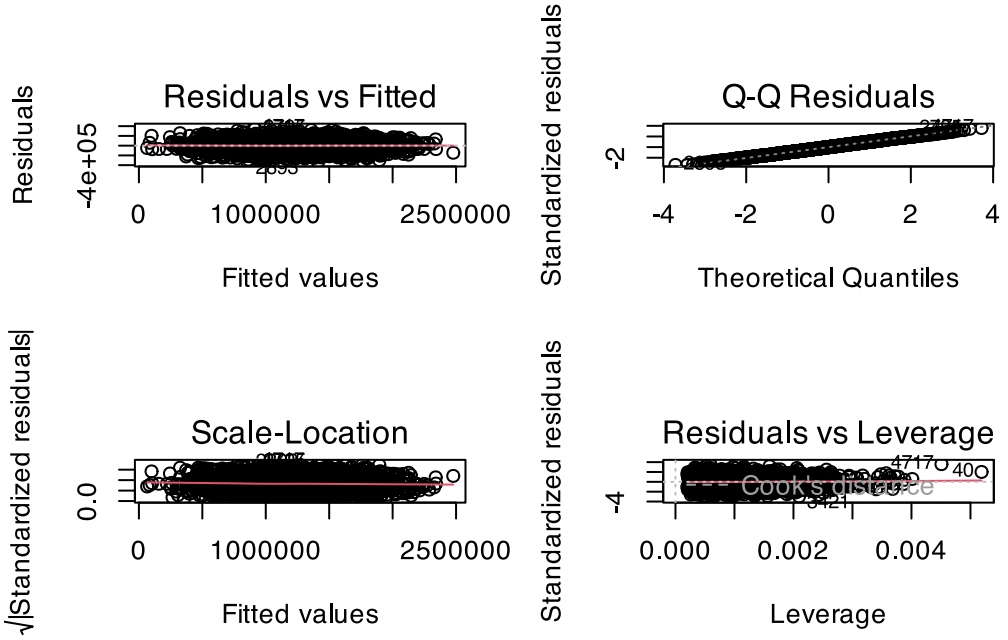
```
               area_income area_house_age room_no bedroom_no area_pop Price
area_income          1.000         -0.002  -0.011      0.020   -0.016 0.640
area_house_age      -0.002          1.000  -0.009      0.006   -0.019 0.453
room_no             -0.011         -0.009   1.000      0.463    0.002 0.336
bedroom_no           0.020          0.006   0.463      1.000   -0.022 0.171
area_pop            -0.016         -0.019   0.002     -0.022    1.000 0.409
Price                0.640          0.453   0.336      0.171    0.409 1.000
```

2. Fit the model

```
housing_mlr <- lm(Price~ area_income+ area_house_age + room_no + area_pop, data
= housing)
```

3. Check assumptions

```
par(mfrow = c(2,2))
plot(housing_mlr)
```




```
par(mfrow = c(1,1))
```

Assumptions are beautiful and prefect. Relationship is linear, data is normally distributed, there is no fanning in the scale-location plot, and the residuals vs leverage doesn't even show the Cook's distance line.

4. P-value and model fit

```
summary(housing_mlr)
```

```

Call:
lm(formula = Price ~ area_income + area_house_age + room_no +
    area_pop, data = housing)

Residuals:
    Min      1Q  Median      3Q     Max
-338419  -70058     132   69074  362025

Coefficients:
```

```
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.638e+06  1.716e+04 -153.73   <2e-16 ***
area_income     2.158e+01  1.343e-01  160.74   <2e-16 ***
area_house_age  1.657e+05  1.443e+03  114.77   <2e-16 ***
room_no         1.216e+05  1.423e+03   85.48   <2e-16 ***
area_pop        1.520e+01  1.442e-01  105.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 101200 on 4995 degrees of freedom
Multiple R-squared:  0.918, Adjusted R-squared:  0.9179
F-statistic: 1.398e+04 on 4 and 4995 DF,  p-value: < 2.2e-16
```

```
a)   Is the model significant? Yes, the f tests shows p < 0.05

b)   Are the predictors significant? All the predictors are significant, with
p-values < 0.05

c)  How good is the model fit? The model has an excellent fit. The R^2^ = 0.9179,
which indicates that the model explains ~92% of the variation in the data.
```

5. Conclusions

The F-test indicated our model is significant (P < 0.05), allowing us to reject the null hypothesis and conclude at least one of our partial regression coefficients has a slope not equal to 0.

Our t-test upon each of our partial regression coefficients supports this, as area income, area house age, room number and area population all have p-values below 0.05.

The fit of the model is excellent, with ~92% of the variation explained by our model.

Exercise 2: Energy use

Data:

- energy

Use the dataset from kaggle to explore which variables affect energy consumption.

Solution

```
#Read data
energy<- read_csv("data/energy_data.csv")
```

```
Rows: 100 Columns: 7
── Column specification ──────────────────────────────────────────────────────
Delimiter: ","
chr (2): building_type, day_of_week
dbl (5): square_footage, occupant_number, appliances_used, avg_temp, energy_...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

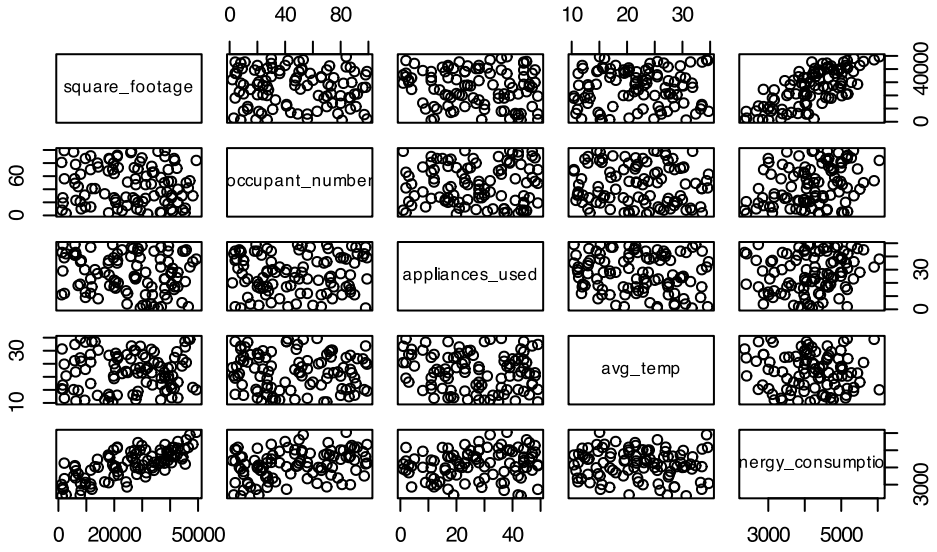
```
str(energy)
```

```
spc_tbl_ [100 × 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ building_type       : chr [1:100] "Residential" "Commercial" "Commercial"
"Residential" ...
 $ square_footage    : num [1:100] 24563 27583 45313 41625 36720 ...
 $ occupant_number   : num [1:100] 15 56 4 84 58 47 18 21 24 90 ...
 $ appliances_used   : num [1:100] 4 23 44 17 47 28 44 19 16 35 ...
 $ avg_temp          : num [1:100] 28.5 23.1 33.6 27.4 17.1 ...
 $ day_of_week       : chr [1:100] "Weekday" "Weekend" "Weekday" "Weekend" ...
 $ energy_consumption: num [1:100] 2866 4284 5068 4624 4821 ...
 - attr(*, "spec")=
  .. cols(
  ..   building_type = col_character(),
  ..   square_footage = col_double(),
  ..   occupant_number = col_double(),
  ..   appliances_used = col_double(),
  ..   avg_temp = col_double(),
  ..   day_of_week = col_character(),
  ..   energy_consumption = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

```
#Keep only the numeric variables
energy_numeric <- energy%>%
  select(-building_type, -day_of_week)
```

1. Scatterplot and correlation

```
pairs(energy_numeric)
```




```
round(cor(energy_numeric),3)
```

```
                  square_footage occupant_number appliances_used avg_temp
square_footage             1.000          -0.080          -0.146    0.031
occupant_number           -0.080           1.000           0.034   -0.064
appliances_used           -0.146           0.034           1.000   -0.136
avg_temp                   0.031          -0.064          -0.136    1.000
energy_consumption         0.724           0.309           0.176   -0.080
                  energy_consumption
square_footage                 0.724
occupant_number                0.309
appliances_used                0.176
avg_temp                      -0.080
energy_consumption             1.000
```

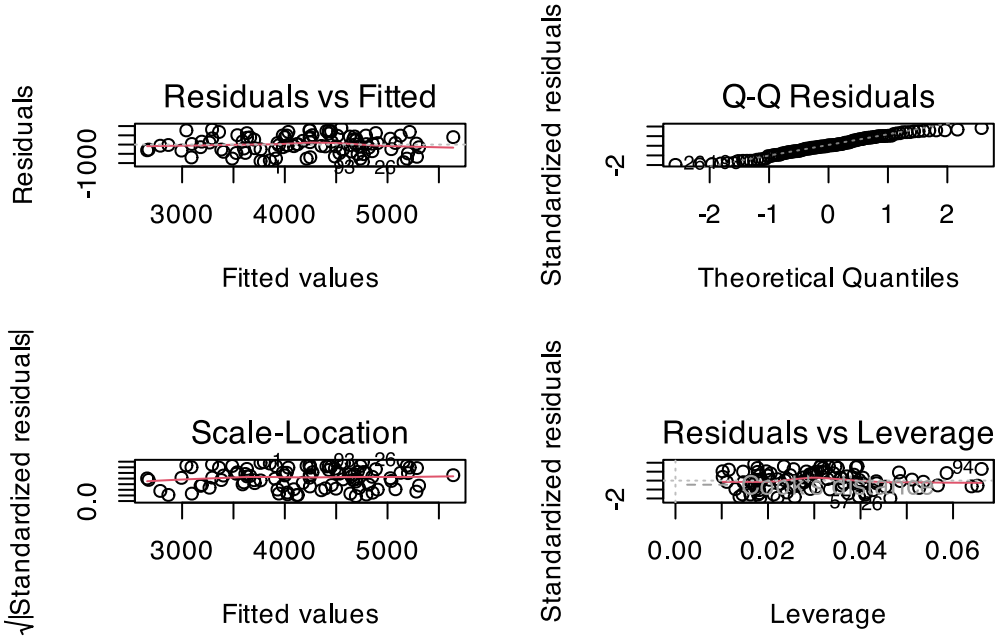
Not too much correlation between independent variables, but `appliances_used` has a very weak relationship with `energy_consumption`. I'm going to leave it out of my model.

2. Fit the model

```
energy_lm <-lm(energy_consumption~ square_footage + occupant_number, data =
energy_numeric )
```

3. Check assumptions

```
par(mfrow = c(2,2))
plot(energy_lm)
```




```
par(mfrow = c(1,1))
```

All the assumptions look fine. The points are starting to veer off the line at the edges of the QQ plot, but I'm not worried about it.

4. P-value and model fit

```
summary(energy_lm)
```

```
Call:
lm(formula = energy_consumption ~ square_footage + occupant_number,
    data = energy_numeric)

Residuals:
    Min      1Q  Median      3Q     Max
-951.19 -312.04   18.46  400.32  895.85

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)     2.517e+03  1.357e+02  18.555  < 2e-16 ***
square_footage  4.577e-02  3.608e-03  12.684  < 2e-16 ***
occupant_number 1.028e+01  1.654e+00   6.214  1.3e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 490.7 on 97 degrees of freedom
Multiple R-squared:  0.6597,    Adjusted R-squared:  0.6527
F-statistic: 94.02 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
a)   Is the model significant? yes, the f-test shows p< 0.05

b)  Are the predictors significant? yes, the t tests for theindividual predictors
all have p < 0.05

c)   How good is the model fit? The model has a decent fit, with an adjusted
R^2^ of 0.65
```

5. Conclusions > The F-test indicated our model is significant (P < 0.05), allowing us to reject the null hypothesis and conclude at least one of our partial regression coefficients has a slope not equal to 0.

   Our t-test upon each of our partial regression coefficients supports this, as area income, area house age, room number and area population all have p-values below 0.05.

   The fit of the model is good, with 65% of the variation explained by our model.

Exercise 3: Sales vs advertising budget

Data:

- advertising budget

Use the dataset from kaggle to explore how different advertising budgets affect sales.

Solution

```
sales<- read_csv("data/Advertising_Budget_and_Sales.csv")
```

```
New names:
Rows: 200 Columns: 5
── Column specification
──────────────────────────────────────────────────────── Delimiter: "," dbl
```

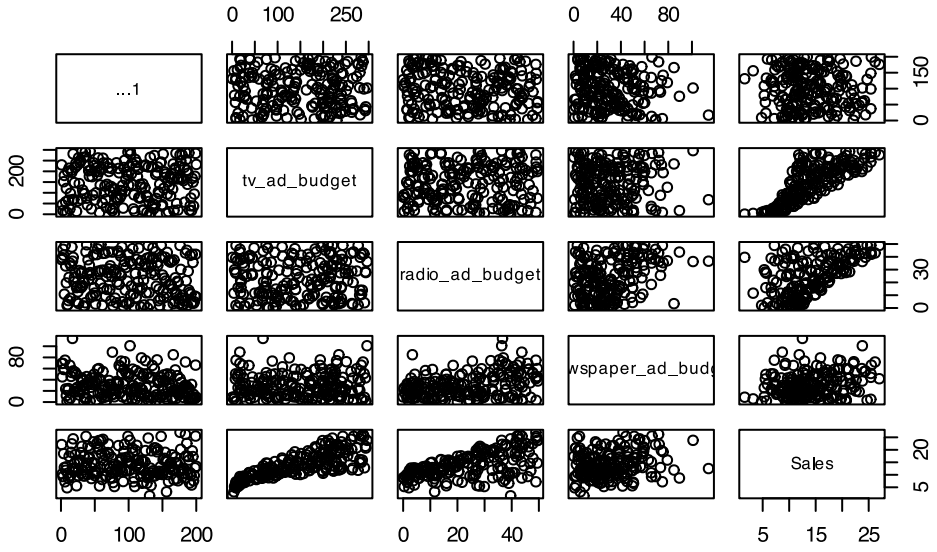
```
(5): ...1, tv_ad_budget, radio_ad_budget, newspaper_ad_budget, Sales
ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
Specify the column types or set `show_col_types = FALSE` to quiet this message.
• `` -> `...1`
```

```
str(sales)
```

```
spc_tbl_ [200 × 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ ...1               : num [1:200] 1 2 3 4 5 6 7 8 9 10 ...
 $ tv_ad_budget       : num [1:200] 230.1 44.5 17.2 151.5 180.8 ...
 $ radio_ad_budget    : num [1:200] 37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1
2.6 ...
 $ newspaper_ad_budget: num [1:200] 69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1
21.2 ...
 $ Sales              : num [1:200] 22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8
10.6 ...
 - attr(*, "spec")=
  .. cols(
  ..   ...1 = col_double(),
  ..   tv_ad_budget = col_double(),
  ..   radio_ad_budget = col_double(),
  ..   newspaper_ad_budget = col_double(),
  ..   Sales = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

1. Scatterplot and correlation

```
pairs(sales)
```




```
round(cor(sales),3)
```

```
                    ...1 tv_ad_budget radio_ad_budget newspaper_ad_budget
...1                1.000        0.018          -0.111               -0.155
tv_ad_budget        0.018        1.000           0.055                0.057
radio_ad_budget    -0.111        0.055           1.000                0.354
newspaper_ad_budget -0.155       0.057           0.354                1.000
Sales               -0.052       0.782           0.576                0.228
                    Sales
...1               -0.052
tv_ad_budget        0.782
radio_ad_budget     0.576
newspaper_ad_budget 0.228
Sales               1.000
```

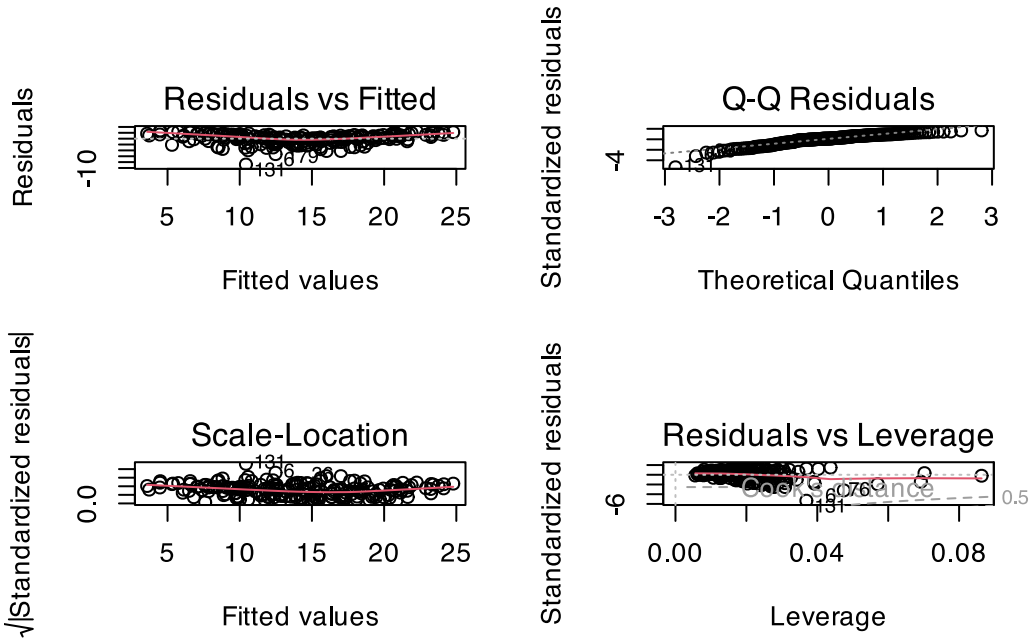
The first variable is just the row ID, so I'm going to ignore it. All the independent variables look like they have a relationship with the target variable, though the relationship with `newspaper_ad_budget` is rather weak. There aren't any particularly strong correlations between the budgets, so I will include all of them in my model.

2. Fit the model

```
sales_mlr<- lm(Sales~tv_ad_budget+radio_ad_budget+ newspaper_ad_budget, data =
sales)
```

3. Check assumptions

```
par(mfrow = c(2,2))
plot(sales_mlr)
```




```
par(mfrow = c(1,1))
```

All the assumptions look fine. The residuals vs fitted plot is a bit squished, but the line is relatively straight, so I'm not worried about it.

4. P-value and model fit

```
summary(sales_mlr)
```

```
Call:
lm(formula = Sales ~ tv_ad_budget + radio_ad_budget + newspaper_ad_budget,
    data = sales)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.938889   0.311908   9.422   <2e-16 ***
tv_ad_budget        0.045765   0.001395  32.809   <2e-16 ***
radio_ad_budget     0.188530   0.008611  21.893   <2e-16 ***
newspaper_ad_budget -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
a)   Is the model significant? Yes, the F-test shows p< 0.05.

b)    Are the predictors significant? The tv and radio ad budgets are both
significant, but the newspaper ad budget was not significant (p = 0.86)

c)   How good is the model fit? The model has an excellent fit with an adjusted
R^2^ of 0.8956
```

5. Conclusions > The F-test indicated our model is significant (P < 0.05), allowing us to reject the null hypothesis and conclude at least one of our partial regression coefficients has a slope not equal to 0.

   Our t-test upon each of our partial regression coefficients supports this, as `tv_ad_budget` and `radio_ad_budget` both have p-values below 0.05. However, `newspaper_ad_budget` was not a significant predictor, with p = 0.86.

   The fit of the model is excellent, with ~90% of the variation explained by our model. Since the `newspaper_ad_budget` was not significant, we could try running the model again without this variable, to see if it imporves the fit.
