Introduction to Regression in R

1. Start RStudio, set your working directory and load the data set carain.csv from Module 5.

```
> carain <- read.csv("carain.csv", header=TRUE, sep=",")
```

2. Inspect the variables in your data frame object

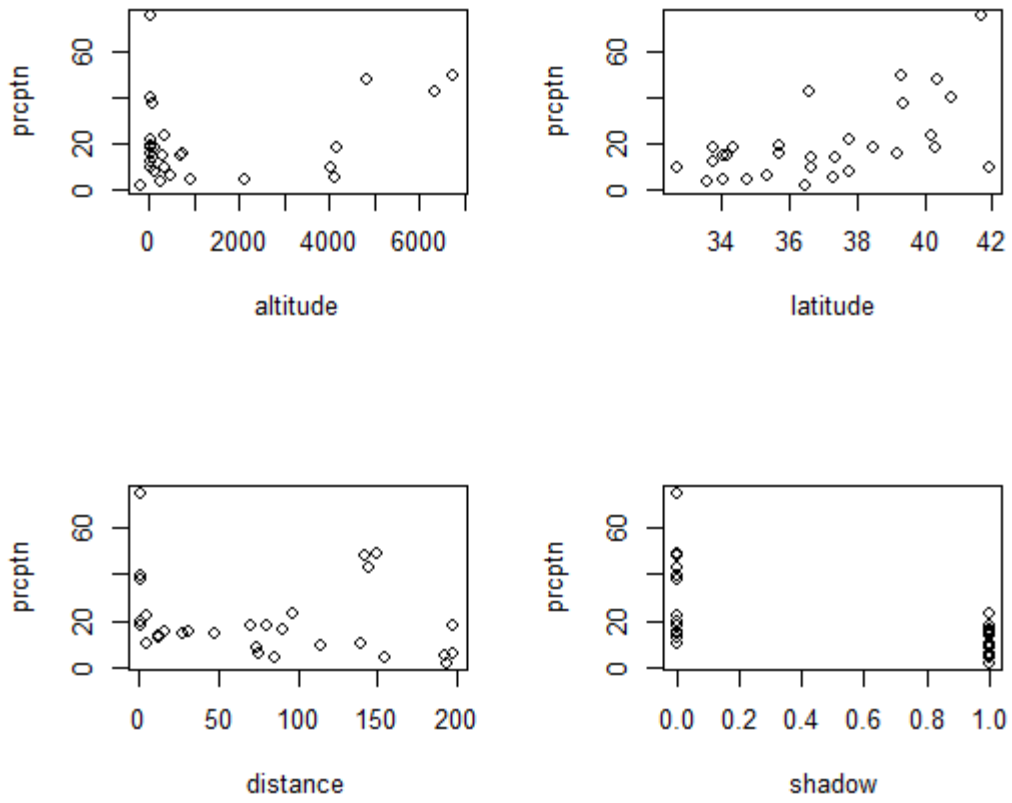
```
> str(carain)
'data.frame':    30 obs. of  6 variables:
 $ station : int  1 2 3 4 5 6 7 8 9 10 ...
 $ prcptn  : num  39.6 23.3 18.2 37.5 49.3 ...
 $ altitude: int  43 341 4152 74 6752 52 25 95 6360 74 ...
 $ latitude: num  40.8 40.2 33.8 39.4 39.3 ...
 $ distance: int  1 97 70 1 150 5 80 28 145 12 ...
 $ shadow  : int  0 1 1 0 0 0 1 1 0 1 ...
```

3. So you have 30 climate stations at which precipitation and other climate data were recorded across the state of California in a specific year. Precipitation (prcptn) is our dependent variable. We seek to understand the variance in prcptn values across the state. Regression is the statistical tool that we can use to help in this task. Regression uses the variance in one or more independent variables to try and "explain", in a statistical sense, the variance in a dependent variable. In the carain dataset, we have 4 independent variables – altitude (height above sea-level where climate observations were measured), latitude (distance from the equator of the climate stations), distance (distance from the coast of the climate stations), shadow (a rain shadow effect). You likely know that all else equal in the state, it rains more at higher elevations, it rains more as you move north, it rains less as you move away from the coast, and locations to the east of the Sierras receive less rainfall than those west of the mountains.

4. Attach the carain object and then generate some quick plots of the independent variables against the dependent variable. It's always good to plot your data and have a look at things!

```
> attach(carain)
> par(mfrow=c(2,2))
> plot(altitude, prcptn)
> plot (latitude, prcptn)
> plot (distance, prcptn)
> plot (shadow, prcptn)
```

The plots below hint at the right sort of relationships, but more investigation is required. Note that the dependent variable is always plotted on the Y-axis.



5. Let's have a look at the correlation between altitude and precptn

```
> cor(altitude, prcptn)
[1] 0.3020067
```

So a moderate positive relationship. Of course it does not matter which way around you order the variables in correlation

```
> cor(prcptn, altitude)
[1] 0.3020067
```

6. So now let's think about regression. Here the order of variables is critical. The dependent variable is always the first variable listed in the regression function

```
> regression1 <- lm(prcptn ~ altitude, data=carain)
```

So we create an object (regression1) to store the results of the regression output. lm stands for linear model and in parentheses we show the function we want to examine, prcptn as a function (~) of altitude). To check the output

```
> summary(regression1)

Call:
lm(formula = prcptn ~ altitude, data = carain)

Residuals:
    Min      1Q  Median      3Q     Max
-20.620  -8.479  -2.729   4.555  58.271
```

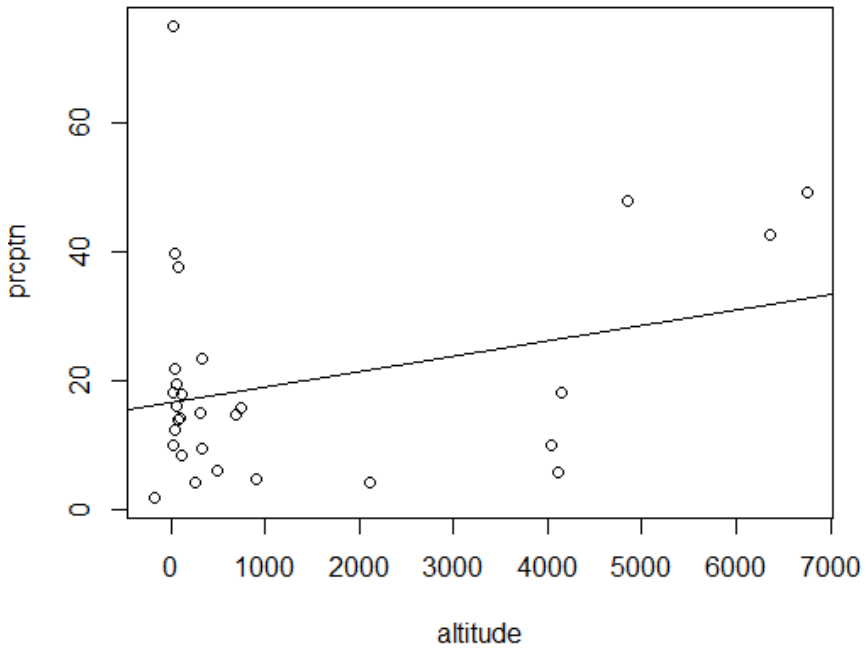
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.514799   3.539141   4.666  6.9e-05 ***
altitude     0.002394   0.001428   1.676    0.105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.13 on 28 degrees of freedom
Multiple R-squared:  0.09121, Adjusted R-squared:  0.05875
F-statistic:  2.81 on 1 and 28 DF,  p-value: 0.1048
```

7. What do we look at in terms of the output? Well the overall goodness of fit of the model (the share of the variance in prcptn explained by the variance in altitude is the first R^2 (or r^2) term = 0.0912. This makes sense as the correlation coefficient (r) between altitude and prcptn was 0.302 and 0.302*0.302 = 0.0912. So altitude seems to explain about 9% of the variance in prcptn in the data.

8. We also want to check our regression equation: Y = 16.515 + 0.002394 altitude
This states that if altitude = 0, then prcptn = 16.515 inches (but beware we cannot reliably interpret the intercept). Our real focus is on the slope of the regression function (0.002394). This variable means that for every 1 unit increase in altitude (every additional 1 foot above sea-level) prcptn on average will rise by 0.002394 inches. You could check this function in another plot

```
> plot(altitude, prcptn)
> abline(regression1)
```



9. So 0.002394 is our sample slope coefficient. Is this value significantly different from zero? To check this, we look at the p-value for the slope coefficient. In this case the p-value = 0.105. So we would not reject the null hypothesis that this sample slope coefficient could come from a population where the slope parameter between altitude and precptn was equal to zero. In other words, we say that altitude has no significant impact on prcptn.

10. Let's jump right in to running another regression, this time with two independent variables, altitude and latitude.

```
> regression2 <- lm(prcptn ~ altitude + latitude, data=carain)
> summary(regression2)

Call:
lm(formula = prcptn ~ altitude + latitude, data = carain)

Residuals:
    Min      1Q  Median      3Q     Max
-29.807  -7.117  -0.783   6.490  41.356

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.057e+02  3.616e+01  -2.924  0.00692 **
altitude     1.414e-03  1.252e-03   1.129  0.26866
latitude     3.338e+00  9.841e-01   3.392  0.00215 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.75 on 27 degrees of freedom
Multiple R-squared:  0.3628,  Adjusted R-squared:  0.3156
F-statistic: 7.686 on 2 and 27 DF,  p-value: 0.00228
```

11. OK, with this regression we are now explaining 36.28% of the variance in the dependent variable. The variable altitude is still not significant with a p-value = 0.26866, but latitude is statistically significant with a p-value = 0.00215. So we would reject the null hypothesis that the sample slope coefficient on latitude could come from a population with a slope parameter on latitude equal to zero. We also get an F-statistic that has a p-value of 0.00228, meaning that the overall regression is significant. No one really looks at the F-statistic, though. We are really concerned with the R^2 and with the significance and sign of the independent variables. As well as being significant, we do want the slope coefficients to have the right sign.

12. Let's add distance from the coast to our model.

```
> regression3 <- lm(prcptn ~ altitude + latitude + distance, data=carain)
```

Note that the order of the independent variables does not matter – check this!

```
> summary(regression3)
```

```
Call:
lm(formula = prcptn ~ altitude + latitude + distance, data = carain)

Residuals:
    Min      1Q  Median      3Q     Max
-28.723  -5.604  -0.532   3.509  33.317

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.024e+02  2.920e+01  -3.505 0.001673 **
altitude     4.091e-03  1.218e-03   3.358 0.002428 **
latitude     3.451e+00  7.947e-01   4.343 0.000190 ***
distance    -1.429e-01  3.634e-02  -3.932 0.000559 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.1 on 26 degrees of freedom
Multiple R-squared:  0.6004,  Adjusted R-squared:  0.5542
F-statistic: 13.02 on 3 and 26 DF,  p-value: 2.201e-05
```

13. The R^2 is now up to 60%, so our extended model is explaining more of the variance in prcptn. Look at our 3 slope coefficients, they are now all statistically significant with the expected sign (negative on distance from the coast and positive for altitude and latitude). How do we interpret the slope coefficient when there is more than one independent variable? Well the slope coefficients are now called partial regression coefficients and each of them shows the influence of a single variable on prcptn with the influence of other independent variables held constant. So the coefficient for distance means that every mile you move away from the coast, on average precipitation declines by 0.1429 inches with the influence of altitude and latitude held constant.

14. The slope coefficient for altitude was previously insignificant. If an independent variable appears in a model that is correlated with other independent variables that are not included in the model, a portion of the variance of the omitted variables (say distance) has an influence on the slope coefficient of the variable that is included (altitude). This is why it is so important to make sure that you have included all the independent variables that theory states are important in the regression model.

15. Let's think about this a bit more, because this will really help you understand the meaning of a partial regression coefficient. We can also get predicted values of Y from the regression and a set of residuals which represent the difference between the actual values of Y and the predicted values. To get the predicted values and the residuals for the regression3 model

```
> predict3 <- predict(regression3)
> residual3 <-resid(regression3)
```

Now back to thinking about partial regression coefficients. I am going to regress prcptn on latitude and distance and capture the residuals. Those residuals represent the portion of the variance in precipitation that is unrelated to latitude and distance (because their influence has been captured in the regression).

```
> regression11 <- lm(prcptn ~ latitude + distance, data=carain)
```

```
> resid11 <- resid(regression11)
```

I am now going to regress altitude against latitude and distance and capture the residuals. I know you are protesting – how can altitude be influenced by latitude and distance from the coast? Altitude does not make sense as a dependent variable here, but R doesn't know that. When we run this regression we are simply removing the influence of latitude and distance from altitude.

```
> regression12 <- lm(altitude ~ latitude + distance, data=carain)
> resid12 <- resid(regression12)
```

Now finally, regress the residuals from regression 11 against the residuals from regression 12 (so here I am regressing prcptn with the influence of latitude and distance removed on the variable altitude with the effects of latitude and distance removed)

```
> regression13 <- lm(resid11 ~ resid12, data=carain)
> summary(regression13)

Call:
lm(formula = resid11 ~ resid12, data = carain)

Residuals:
    Min      1Q  Median      3Q     Max
-28.723  -5.604  -0.532   3.509  33.317

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.054e-16  1.952e+00   0.000  1.00000
resid12     4.091e-03  1.174e-03   3.485  0.00164 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.69 on 28 degrees of freedom
Multiple R-squared:  0.3025,  Adjusted R-squared:  0.2776
F-statistic: 12.14 on 1 and 28 DF,  p-value: 0.00164
```

I don't care about most of the output here, just look at the slope coefficient on resid12. Now compare this to the slope coefficient on altitude for regression 3 above. Yes, they are identical! You might have to think about what we've done here a little.

16. Finally, back to our regression3 model again. Now I am going to add the rain shadow variable.

```
> regression4 <- lm(prcptn ~ altitude + latitude + distance + shadow, data=carain)
> summary(regression4)

Call:
lm(formula = prcptn ~ altitude + latitude + distance + shadow,
    data = carain)

Residuals:
     Min       1Q   Median       3Q      Max
-22.3166  -4.8036   0.1003   3.1063  28.7263
```

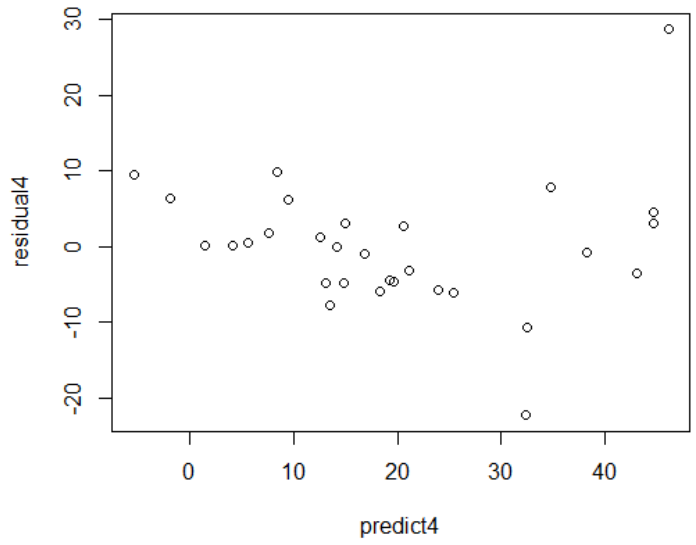
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -97.878166  24.134158  -4.056 0.000429 ***
altitude      0.002206   0.001132   1.949 0.062615 .
latitude      3.453198   0.655987   5.264 1.88e-05 ***
distance     -0.053670   0.038787  -1.384 0.178681
shadow      -15.853895   4.371013  -3.627 0.001282 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.16 on 25 degrees of freedom
Multiple R-squared:  0.7382,  Adjusted R-squared:  0.6963
F-statistic: 17.62 on 4 and 25 DF,  p-value: 5.437e-07
```

See how the partial regression coefficients have shifted again. This is simply the result of some correlation between the rain shadow effect and the existing variables in the model (altitude, latitude and distance). Now note that the rain shadow effect is significant and with the correct sign, latitude is also significant, but distance is not significant and altitude is close. Also note how the overall variance explained by the model has improved. Regression can be tricky!

17. Finally, a very useful plot in regression analysis is to examine the relationship between the residuals from the regression and the predicted values. We plot the predicted values on the X-axis

```
> predict4 <- predict(regression4)
> residual4 <- resid(regression4)
> plot(predict4, residual4)
```



Overall, we seem to be doing a reasonable job in explaining the variance in precipitation. There are two outliers, is that a data error or something else? Note that you **do not** want to see an obvious trend in this plot, for that would suggest something important is missing from the model.