Lab 04A: Week 11

Contents

The **specific aims** of this lab are:

- perform simple and multiple regression
- identify which variables are significant and perform model selection
- check whether or not the assumptions for linear regression are met
- interpret the coefficients from a linear regression model
- use an estimated regression model to predict the outcomes for a new observation

The unit **learning outcomes** addressed are:

- LO1 Formulate domain/context specific questions and identify appropriate statistical analysis.
- LO3 Construct, interpret and compare numerical and graphical summaries of different data types including large and/or complex data sets.
- LO6 Formulate, evaluate and interpret appropriate linear models to describe the relationships between multiple factors.

1 Questions

1.1 Wind

The data in `pollut.txt` are `WS` (wind speeds), `Temp` (temperature), `H` (humidity), `In` (insolation) and `O` (ozone) for 30 days.

```
pollut = read_csv("https://raw.githubusercontent.com/DATA2002/data/master/pollut.txt")
glimpse(pollut)
```

```
Rows: 30
Columns: 5
$ WS   <dbl> 50, 47, 57, 38, 52, 57, 53, 62, 52, 42, 47, 40, 42, 40,…
$ Temp <dbl> 77, 80, 75, 72, 71, 74, 78, 82, 82, 82, 82, 80, 81, 85,…
$ H    <dbl> 67, 66, 77, 73, 75, 75, 64, 59, 60, 62, 59, 66, 68, 62,…
$ In   <dbl> 78, 77, 73, 69, 78, 80, 75, 78, 75, 58, 76, 76, 71, 74,…
$ O    <dbl> 15, 20, 13, 21, 12, 12, 12, 11, 12, 20, 11, 17, 20, 23,…
```

1. Generate a pairs plot of the data using `pairs()` or the `ggpairs()` function from the **GGally** package (Schloerke et al. 2021).
2. Perform a multiple regression of `ozone` on the other variables using `lm()`.
3. Does it look like any variables can be dropped from the model? If you were doing backwards selection using the `drop1()` function which would you drop first? Write down a the workflow for a formal hypothesis test to see if the coefficient for insolation is significantly different to zero.
4. Rather than dropping variables using their individual p-values, we can instead consider using an information criterion. Use the `step()` function to perform selection using the AIC starting from the full model.
5. Write down the fitted model for the model selected by the step-wise procedure.
6. Check the linear regression assumptions for the stepwise model.
7. What proportion of the variability of ozone is explained by the explanatory variables in the step-wise selected model?
8. Use the model to estimate the average `ozone` for days when `WS=40`, `Temp=80` and `H=50`. Is a confidence interval or a prediction interval most appropriate here? Write down the interval you think is most appropriate.

1.2 Diabetes

Efron et al. (2004) introduced the diabetes data set with 442 observations and 11 variables. It is often used as an exemplar data set to illustrate new model selection techniques. The following commands will help you get a feel for the data.

```
# install.packages('mplot')
data("diabetes", package = "mplot")
# help('diabetes', package='mplot')
```

```
glimpse(diabetes)  # glimpse the structure of the diabetes
pairs(diabetes)  # traditional pairs plot
GGally::ggpairs(diabetes)  # ggplotified pairs plot
boxplot(diabetes)  # always a good idea to check for gross outliers
boxplot(scale(diabetes))  # always a good idea to check for gross outliers
```

```
# OPTIONAL!!
# install.packages(c("pairsD3","heatmaply","skimr"))
pairsD3::shinypairs(diabetes) # interactive pairs plot of the data set
heatmaply::heatmaply(cor(diabetes))
skimr::skim(diabetes) # summary of the diabetes data
```

We can fit the null model (without any variables) and the full model as follows:

```
M0 = lm(y ~ 1, data = diabetes)  # Null model
M1 = lm(y ~ ., data = diabetes)  # Full model
```

We can compare the results side by side using the **stargazer** package (Hlavac 2018).

```
# stargazer::stargazer(M0, M1, type = 'latex', header = FALSE)
stargazer::stargazer(M0, M1, type = "html")
```

	Dependent variable:	
	y	
	(1)	(2)
age		-0.036
		(0.217)
sex		-22.860***
		(5.836)
bmi		5.603***
		(0.717)
map		1.117***
		(0.225)
tc		-1.090*
		(0.573)
ldl		0.746
		(0.531)
hdl		0.372
		(0.782)
tch		6.534
		(5.959)
ltg		68.483***
		(15.670)
glu		0.280

		(0.273)
Constant	152.133***	-334.567***
	(3.667)	(67.455)
Observations	442	442
R^2	0.000	0.518
Adjusted R^2	0.000	0.507
Residual Std. Error	77.093 (df = 441)	54.154 (df = 431)
F Statistic		46.272*** (df = 10; 431)
Note:		*p<0.1;* ***p<0.05;*** p<0.01

1. Try doing backward selection using AIC first.
2. Explore the forwards selection technique, which works very similarly to backwards selection, just set `direction = "forward"` in the `step()` function. When using `direction = "forward"` you need to specify a scope parameter: `scope = list(lower = M0, upper = M1)`.
3. Try using the `add1()` and `drop1()` functions. The general form is `add1(fitted.model, test = "F", scope = M1)` or `drop1(fitted.model, test = "F")`
4. What if you try backwards selection using an individual p-value approach, i.e. using `drop1()` from the full model.
5. Are you satisfied with the model you have arrived at? Check the assumptions.
6. Write down your final fitted model and interpret the estimated coefficients.

2 For practice after the computer lab

2.1 Predicting capital value

The data in `rentcap.txt` shows the capital value and annual rental value of 96 domestic properties in Auckland in 1991. The aim was to explore their relationship in the hope of being able to predict capital value from rental value.

```
rent = read_tsv("https://raw.githubusercontent.com/DATA2002/data/master/rentcap.txt")
glimpse(rent)
```

```
Rows: 96
Columns: 3
$ Capital <dbl> 61500, 67500, 75000, 75000, 76000, 77000, 80000, 810…
$ ...2    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
```

```
$ Rental  <dbl> 6656, 6864, 4992, 7280, 6656, 4576, 4992, 6656, 4784…
```

1. Fit a simple linear regression to the data to assess whether the rental value has an influence on the capital value.
2. Obtain a predicted capital value from the rental value of 7500 and the corresponding 90% prediction interval for your predicted capital value.
3. Use various visualisations to comment on whether the assumptions for the prediction interval are satisfied. If not, find an appropriate transformation and re-fit the linear regression.

References

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. "Least Angle Regression." *The Annals of Statistics* 32 (2): 407–51. https://doi.org/10.1214/009053604000000067.

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). https://CRAN.R-project.org/package=stargazer.

Schloerke, Barret, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. 2021. *GGally: Extension to 'Ggplot2'*. https://CRAN.R-project.org/package=GGally.