Statistics 2

Multiple regression

Casper Albers & Jorge Tendeiro

Lecture 4, 2019 – 2020

university of groningen

Overview

Literature for this lecture

Read:

Agresti, Sections 11.1 – 11.3

Multiple regression

Simple Linear Regression ⟷ Multiple Regression

- Simple Linear Regression:
  Predict values of a y-variable through a linear relation with one x-variable.
- Multiple Linear Regression:
  Predict values of a y-variable through linear relations with multiple x-variables.

So, think of multiple linear regression as a natural extension of simple linear regression.

Multiple regression

Population regression equation:

$$\boxed{E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}$$

Or in terms of individual scores y:

$$\boxed{y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i}$$

- One dependent variable y.
- p independent variables $x_1, \ldots x_p$.
- Residual (error, deviation, noise, ...) ε_i:
  - $\mathcal{N}(0, \sigma)$, with σ constant.
  - Independent of all x-variables.
- Partial regression coefficients:
  - α: Intercept.
  - $\beta_1, \ldots, \beta_p$: Slopes.
  - σ: Residual SD.

Conceptual differences between simple and multiple regression:

- (There are none!)

Example – Overclaiming

Atir, Rosenzweig, and Dunning (2015) studied whether experts overrate the extent of their expertise[1].

- Dependent variable
  - y: OVCLAIM, overclaiming based on defining 15 terms (of which 3 do not exist).
- Independent variables ($p = 2$)
  - x_1: SPKNOW, based on a questionnaire assessing self-perceived knowledge.
  - x_2: ACCUR, accuracy operationalized as the ability to distinguish between the 12 real terms and the 3 fake terms.

Sample size = 202.

[1] Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science*, *26*, 1295-1303. doi: 10.1177/0956797615588195

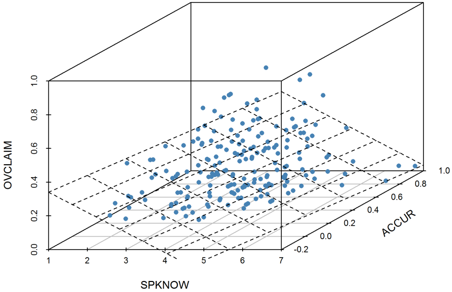
Example – Overclaiming (correlations)

		OVCLAIM	SPKNOW	ACCUR
OVCLAIM	Pearson's r	–		
	p-value	–		
SPKNOW	Pearson's r	0.481	–	
	p-value	$< .001$	–	
ACCUR	Pearson's r	-0.672	0.033	–
	p-value	$< .001$	0.645	–

We expect that:

- OVCLAIM is linearly related to either predictor.
- The predictors SPKNOW and ACCUR are not strongly linearly related.

Example – Overclaiming (3D plot)



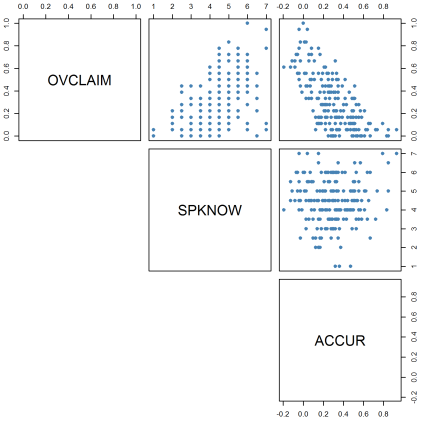
Interpretation of 3D plots is very difficult.

Observe that:

- In simple linear regression: Regression line plotted in 2D.
- In multiple linear regression: Regression (hyper) plane plotted in $(p+1)D$.

Example – Overclaiming (scatterplot matrix)

Better option: Scatterplot matrix.



Example – Overclaiming (partial plot)

One other plot alternative: Partial plots.

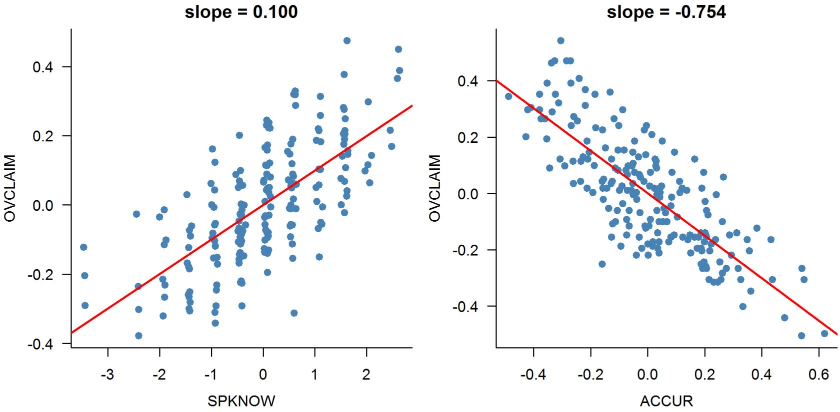
Idea:

Look at the relation between y and a predictor x_i, by removing (i.e., partialing out) the effects of all remaining predictors from both y and x_i.

For example, to look at the relation between y = OVCLAIM and x_1 = SPKNOW (and denote x_2 = ACCUR):

1. Partial x_2 from y:
   - $\widehat{y} = a + bx_2$
   - Save the residuals: $y^{(\text{res})} = y - \widehat{y}$.
2. Similarly, partial x_2 from x_1:
   - $\widehat{x_1} = a + bx_2$
   - Save the residuals: $x_1^{(\text{res})} = x_1 - \widehat{x_1}$.
3. Plot $y^{(\text{res})}$ on vertical axis, $x_1^{(\text{res})}$ on horizontal axis.

Example – Overclaiming (partial plot)



Multiple Regression Model

Population

Population regression equation: $E(y) = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p$

↓ link individual score and mean

Statistical model: $y_i = \alpha + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i$

$\Rightarrow$ population parameters: $\alpha, \beta_1, \ldots, \beta_p$ and $\mathcal{N}(0, \sigma)$

Sample

Estimation in sample (OLS): $a, b_1, \ldots, b_p$ and s

Estimated equation: $\widehat{y_i} = a + b_1 x_{1i} + \ldots + b_p x_{pi}$

Population

Population unknown

Probability statements about the unknown population

Tests and CI's $\longleftarrow$ Assumption: $\varepsilon \sim \mathcal{N}(0, \sigma)$

Sample

Estimation in sample (OLS) — $a, b_1, \ldots, b_p$ and s

Estimated equation — $\widehat{y_i} = a + b_1 x_{1i} + \ldots + b_p x_{pi}$

Multiple Regression Model

- Statistical model: $y_i = \alpha + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + e_i$.
  - α: Intercept; expected value for y if *all* x are 0.
  - β_j: Slope for x_j ($j = 1, \ldots, p$); change in y if x_j increases one unit *whilst other predictors stay the same*.
- Estimation using OLS: Minimize the sum of squared errors (SSE).

$$\min \text{SSE} = \min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 .$$

- Provides

$$\widehat{y}_i = a + b_1 x_{1i} + \ldots + b_p x_{pi}.$$

- Estimate σ from the residuals: $s = \sqrt{\frac{\text{SSE}}{n-p-1}}$.

  In JASP, s is also known as the root mean square error (RMSE).

Example – Overclaiming (parameter estimates)

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	0.089	0.037		2.420	0.016
	SPKNOW	0.100	0.008	0.504	13.072	$<$ .001
	ACCUR	−0.754	0.042	−0.688	−17.869	$<$ .001

$$\boxed{\widehat{\text{OVCLAIM}} = 0.089 + 0.100\,\text{SPKNOW} - 0.754\,\text{ACCUR}}$$

Interpret regression coefficients:

- $a = 0.089$: The expected OVCLAIM score is equal to 0.089 when both SPKNOW and ACCUR are equal to 0.
- $b_1 = 0.100$: OVCLAIM increases by 0.100 units when SPKNOW increases by 1 unit, controlling for ACCUR (i.e., keeping ACCUR fixed).
- $b_2 = -0.754$: OVCLAIM decreases by 0.754 units when ACCUR increases by 1 unit, controlling for SPKNOW (i.e., keeping SPKNOW fixed).

Example – Overclaiming (root mean square error)

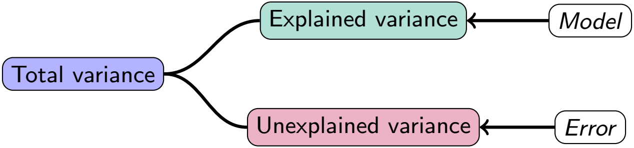
Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.840	0.705	0.702	0.127

Therefore $s = \sqrt{\frac{\text{SSE}}{n-p-1}} = 0.127$.

Sum-of-squares partitioning (ANOVA)

How much of the variance in y can be explained by the model?



$$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y}_i)^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

$$\text{TSS} = \text{RSS} + \text{SSE}$$

- TSS = Total sum of squares
- RSS = Regression sum of squares
- SSE = Sum of squares of residuals (errors)

Sum-of-squares partitioning (ANOVA)

Source	SS	df	MS	F	p
Regression	RSS$=\sum(\widehat{y_i}-\bar{y})^2$	p	MSR$=\frac{RSS}{p}$	$\frac{MSR}{MSE}$	(Table)
Residual	SSE$=\sum(y_i-\widehat{y})^2$	$n-p-1$	MSE$=\frac{SSE}{n-p-1}$		
Total	TSS$=\sum(y_i-\bar{y})^2$	$n-1$			

- df = degrees of freedom
- MSR = Mean squares of regression
- MSE = Mean squares of errors

We will learn more about sum-of-squares partitioning when ANOVA is introduced later in this course.

For the 'overclaiming' dataset, in JASP:

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	7.641	2	3.821	237.7	$<$ .001
	Residual	3.198	199	0.016		
	Total	10.840	201			

Sample multiple correlation:

$$\boxed{R = cor(y, \widehat{y})}$$

- Assess linear association between y and the set of predictors collectively (via $\widehat{y} = a + b_1 x_1 + \cdots + b_p x_p$).
- R is always between 0 (when all $b_j = 0$, $j = 1, \ldots, p$) and 1 (perfect linear relationship).

'Overclaiming' dataset:

OVCLAIM
0.4444
0.5556
0.1667
...
0.1667
0.2778

SPKNOW	ACCUR	$\widehat{\text{OVCLAIM}}$
5.5	0.2500	0.4491
4.5	0.1944	0.3912
3.5	0.3472	0.1763
...	...	...
2.5	0.1250	0.2441
4.0	0.4306	0.1633

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.840	0.705	0.702	0.127

(Recall: $\widehat{\text{OVCLAIM}} = 0.089 + 0.100\,\text{SPKNOW} - 0.754\,\text{ACCUR}$.)

Coefficient of multiple determination:

$$\boxed{R^2 = \frac{RSS}{TSS} = \frac{TSS - SSE}{TSS}}$$

- $R^2 =$ square of the multiple correlation R.
- Assess proportion of total variation in y that is explained by all preditors collectively (via $\widehat{y} = a + b_1 x_1 + \cdots + b_p x_p$).
- R^2 is always between 0 (when all $b_j = 0$, $j = 1, \ldots, p$) and 1 (perfect linear relationship).

'Overclaiming' dataset:

ANOVA

	SS	df	Mean Square	F	p
Regression	7.641	2	3.821	237.7	$< .001$
Residual	3.198	199	0.016		
Total	10.840	201			

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.840	0.705	0.702	0.127

$$R^2 = \frac{7.641}{10.840} = \frac{10.840 - 3.198}{10.840} = .705.$$

Adjusted R^2

R^2 has two drawbacks:

1. R^2 cannot decrease:
   If p increases then R^2 will also increase, *even if the new variables are unimportant*.
2. R^2 overestimates the population value:
   Because the computation of R^2 is optimal for the current sample (but not necessarily for other samples).

Therefore, adjusted R^2 is used in multiple regression.

Most commonly used: Wherry's R^2.

$$\boxed{R^2_{adj} = R^2 - \left(\frac{p}{n-p-1}\right)(1 - R^2)}$$

‘Overclaiming’ (adjusted R^2)

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.840	0.705	0.702	0.127

- $n = 202$
- $p = 2$

$$\begin{aligned} R^2_{adj} &= R^2 - \left(\frac{p}{n-p-1}\right)(1 - R^2) \\ &= .705 - \frac{2}{202-2-1}(1 - .705) \\ &= .702 \end{aligned}$$

Inference in multiple regression

There are two main types of inferential questions that can be asked in multiple regression:

1. Globally:

   Are the predictors, jointly, associated with the response variable?

Hypotheses of interest:

$$\boxed{\mathcal{H}_0 : R^2 = 0 \quad \text{versus} \quad \mathcal{H}_a : R^2 > 0}$$

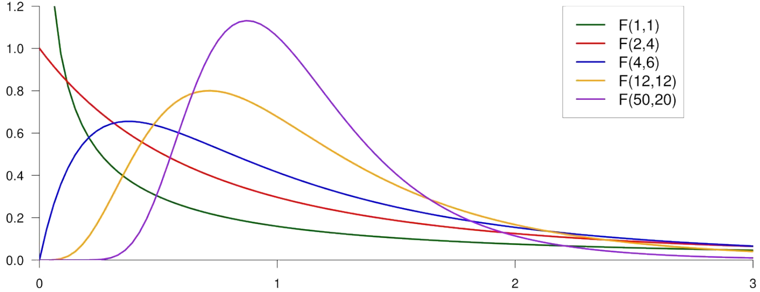
or, equivalently,

$$\boxed{\mathcal{H}_0 : \beta_1 = \cdots = \beta_p = 0 \quad \text{versus} \quad \mathcal{H}_a : \text{At least one } \beta_j \neq 0}$$

Test statistic:

$$\boxed{F = \frac{MSR}{MSE} = \frac{R^2/p}{(1-R^2)/(n-p-1)} \underset{\mathcal{H}_0}{\sim} F(p, n-p-1)}$$

The F distribution



'Overclaiming' (global test)

ANOVA

	SS	df	Mean Square	F	p
Regression	7.641	2	3.821	237.7	$< .001$
Residual	3.198	199	0.016		
Total	10.840	201			

Model Summary

Model	R	R²	Adjusted R²	RMSE
1	0.840	0.705	0.702	0.127

- $df_1 = p = 2$
- $df_2 = n - p - 1 = 202 - 2 - 1 = 199$
- $$F = \frac{R^2/p}{(1 - R^2)/(n - p - 1)} = \frac{.705/2}{(1 - .705)/199} = 237.7.$$
- $P(F \geq 237.7 | F \sim F_{2,199}) < .001$: Reject $\mathcal{H}_0$.

Inference in multiple regression

There are two main types of inferential questions that can be asked in multiple regression:

2. Locally:

   Which predictors are associated with the response variable?

Observe that rejecting the null hypothesis $\mathcal{H}_0 : R^2 = 0$ does not imply that all predictors have a partial effect on y.

Hypotheses of interest:

$$\boxed{\mathcal{H}_0 : \beta_j = 0 \quad \text{versus} \quad \mathcal{H}_a : \beta_j \neq 0,}$$

one test for each predictor j ($j = 1, \ldots, p$).

Test statistic

$$\boxed{t = \frac{b_j}{SE_j} \underset{\mathcal{H}_0}{\sim} t(n - p - 1)}$$

Confidence interval

$$\boxed{b_j \pm t^* SE_j}$$

t^* = critical value from $t(n - p - 1)$

'Overclaiming' (local tests)

Coefficients

						95% CI	
	Unstandardized	Standard Error	Standardized	t	p	Lower	Upper
(Intercept)	0.089	0.037		2.420	0.016	0.016	0.161
SPKNOW	0.100	0.008	0.504	13.072	$<$ .001	0.085	0.115
ACCUR	−0.754	0.042	−0.688	−17.869	$<$ .001	−0.837	−0.671

For example, test for partial effect ACCUR ($\alpha = .05$):

- $t = \frac{-0.754}{0.042} = -17.869$
- $df = n - p - 1 = 202 - 2 - 1 = 199$
- $t^* = t_{.975,199} = 1.972$
- $|t| > t^* \longrightarrow$ Reject $\mathcal{H}_0 : b_{\text{ACCUR}} = 0$
  p-value $= P(|t| \geq 17.869 | t \sim t(199)) < .001$

‘Overclaiming’ (local tests)

Coefficients

	Unstandardized	Standard Error	Standardized	t	p	95% CI Lower	95% CI Upper
(Intercept)	0.089	0.037		2.420	0.016	0.016	0.161
SPKNOW	0.100	0.008	0.504	13.072	$<$ .001	0.085	0.115
ACCUR	-0.754	0.042	-0.688	-17.869	$<$ .001	-0.837	-0.671

For example, 95% CI test for partial effect ACCUR:

- $df = n - p - 1 = 202 - 2 - 1 = 199$
- $t^* = t_{.975,199} = 1.972$
- $B_{\text{ACCUR}} \pm t^* \times SE_{\text{ACCUR}} = -0.754 \pm 1.972 \times 0.042 = [-0.837, -0.671]$

Next week: Multiple regression (interaction effects)

Agresti, Section 11.4 - 11.5