Class 25: Multivariable Regression (Text: Sections 11.1)

Multivariable Regression

To predict lung cancer deaths, we had one explanatory variable—the number of cigarettes smoked per capita—and we found $y = 0.23x + 65.7$.

What if we thought that more than one variable affected y, as is usually the case? Then instead of using

$$y = \beta_0 + \beta_1 x,$$

we could use a model with two explanatory variables,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

or with three or more variables:

$$\underbrace{y}_{\text{Response variable}} = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n}_{\text{Explanatory variables}} + \varepsilon$$

The ***population parameters*** for this model are $\beta_0, \beta_1, \beta_2$ .... etc. and σ.

Then values of the xs are used to predict the values of y. The notation for the ***sample regression line*** is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

The "hat" on the $\hat{y}$ means it was estimated from the data. The coefficients in the sample regression equation are estimated from the data using Excel

- b_0 is the value of y when all the explanatory variables are zero.
- b_1, b_2, b_3,…. are the slopes, or rates of change, of the response variable when each explanatory variable changes *with the other variables held constant*.
- R^2 tell us how close the data is to a line and how reliable the predictions are

(As before, Greek letters are population parameters; ordinary letters are the sample statistics.)

Example: Women's Heptathlon

The women's heptathlon consists of seven track and field events, the 200-m and 800-m runs, 100-m high hurdles, shot put, javelin, high jump, and long jump. Carolina Klüft of Sweden won the high jump and the long jump; Austra Skujyté of Lithuania won the shot put and the 800-meter race. Which of these contributes to who wins? We will compare the contributions of the high jump, the long jump, and the 800-meter run. The dependent variable is the Total number of Heptathlon points. The winning score was 6952.

Women's Heptathlon: First we look at the High Jump alone.

(a) Use the table below to write the regression equation.

Total points	*Coefficients*	*Standard Error*	*95% lower*	*95% upper*	*t Stat*	*p-level*
Intercept	2,767.6	992.1	720.0	4,815.3	2.8	0.010
High Jump (m)	1,929.5	561.9	769.9	3,089.1	3.4	0.002

(b) Interpret the coefficient of High Jump.

(c) Does the High Jump make a significant contribution to the scores? Give three reasons.

(d) What does it mean to "keep the other variables constant"? How does this affect the conclusion?

Now we look at the Long Jump alone.

(a) Use the table below to write the regression equation.

Total points	*Coefficients*	*Standard Error*	*95% lower*	*95% upper*	*t Stat*	*p-level*
Intercept	1,788.9	1,021.6	-319.5	3,897.4	1.8	0.0927
Long Jump (m)	712.6	166.0	370.0	1,055.2	4.3	0.0003

(b) Interpret the coefficient of Long Jump.

(c) Does the Long Jump make a significant contribution to the scores? Give three reasons.

Now we look at the 800-meter alone.
(a) Use the table below to write the regression equation.

Total points	*Coefficients*	*Standard Error*	*95% lower*	*95% upper*	*t Stat*	*p-level*
Intercept	9,557.0	1,390.9	6,686.4	12,427.6	6.9	0.000
800-m (sec)	−24.8	10.2	−45.8	−3.8	−2.4	0.023

(b) Interpret the coefficient 800-meters.

(c) Does the 800-meter run make a significant contribution to the scores? Give three reasons.

So the three variables each make a significant contribution to the score.

- What does it tell us that the coefficient of High Jump is greater than the coefficient of the other two?
- What happens if all three variables are used together?
- Must they be significant? Must each one make a significant contribution?

Size of Coefficients: What affects them?
The size of the coefficient can be changed by changing the units of measurement, so is not relevant. For example, measuring the high jump in centimeters instead of meters would make the coefficient a hundred times smaller, 19.295.

Does changing units affect the p-value? Why?

Does the size of a coefficient affect whether or not it is significant?

Women's Heptathlon: Three variables together:
We could show the regression with pairs of variables, but will jump to three variables right away:

Total points	*Coefficients*	*Standard Error*	*95% lower*	*95% upper*	*t Stat*	*p-level*
Intercept	2,244.3	1,932.2	−1,762.8	6,251.5	1.2	0.258
High Jump (m)	1,106.3	521.7	24.4	2,188.1	2.1	0.045
800-m (sec)	−9.6	8.4	−27.0	7.9	−1.1	0.269
Long Jump (m)	533.4	160.0	201.6	865.2	3.3	0.003

(a) Use the table above to write the regression equation.

(b) Does each variable make a significant contribution to the scores? Give three reasons.

(c) Does the largest coefficient in the three separate regressions correspond to the smallest p-value when the regression is done with all three together?

(d) Interpret the coefficient High Jump.

(e) Interpret the coefficient Long Jump.

(f) Interpret the coefficient 800-meters.

(g) What might you worry about in the interpretation of the coefficient of 800-meters?

The whole table for three Total Points, High Jump, Long Jump, 800-meter run.

Regression Statistics	
R	0.77
R Square	0.59
Adjusted R Square	0.53
S	151.60
Total number of observations	26

ANOVA

	d.f.	*SS*	*MS*	*F*	*p-level*
Regression	3	728,795.9	242,932.0	10.6	0.0002
Residual	22	505,598.4	22,981.7		
Total	25	1,234,394.3			

	Coefficients	*Standard Error*	*95% lower*	*95% upper*	*t Stat*	*p-level*
Intercept	2,244.3	1,932.2	−1,762.8	6,251.5	1.2	0.258
High Jump (m)	1,106.3	521.7	24.4	2,188.1	2.1	0.045
800-m (sec)	−9.6	8.4	−27.0	7.9	−1.1	0.269
Long Jump (m)	533.4	160.0	201.6	865.2	3.3	0.003

(a) What does the p-level in ANOVA tell you?

(b) How is the p-level in ANOVA different from the individual p-values?

(c) What does the R^2 tell us?

What do you expect for the pairwise regressions?

High Jump and 800-meter:

Total points	*Coefficients*	*Standard Error*	*95% lower*	*95% upper*	*t Stat*	*p-level*
Intercept	5,312.1	2,038.5	1,095.1	9,529.1	2.6	0.016
High Jump (m)	1,583.4	601.9	338.3	2,828.5	2.6	0.015
800-m (sec)	-14.2	10.0	-34.8	6.5	-1.4	0.169

Long Jump and 800-meter:

Total points	*Coefficients*	*Standard Error*	*95% lower*	*95% upper*	*t Stat*	*p-level*
Intercept	4,451.5	1,747.4	836.8	8,066.2	2.5	0.0180
800-m (sec)	-15.6	8.5	-33.2	2.0	-1.8	0.0793
Long Jump (m)	626.5	165.2	284.8	968.1	3.8	0.0009

High Jump and Long Jump

Total points	*Coefficients*	*Standard Error*	*95% lower*	*95% upper*	*t Stat*	*p-level*
Intercept	402.1	1,053.2	-1,776.6	2,580.8	0.4	0.706
High Jump (m)	1,307.0	493.8	285.5	2,328.4	2.6	0.014
Long Jump (m)	563.1	158.8	234.6	891.7	3.5	0.002