10.1 Multiple Regression

Section 2.6 showed us how to fit a line to describe a response variable using a single explanatory variable.

In Section 9.1 we learned how to test to see if the slope is significantly different from 0 – that is, the variable helps explain the response.

What if more than one variable is helpful in explaining the response? We use multiple regression.

Simple linear regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Multiple linear regression with k variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

For the simple model our estimate is

$$\hat{y} = b_0 + b_1 X$$

For the multivariable model we get

$$\hat{y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$$

Computers are happy to fit either model.

Ex: (St. Lawrence GPA)

Spose we wish to model student GPAs as reported in a student survey.

Among the available variables are:

GPA = college GPA

Height = height in inches

TV = hrs of tv/wk

Piercings = number of body piercings

MathSAT = Math SAT score

VerbalSAT = Verbal SAT score

We first plot the data to look for patterns
– if it's not linear don't use a line

Pairs plots

xyplots

bwplot

Let R fit the model.

$$\widehat{GPA} = 2.438 - .010\,\text{Height} - .005\,\text{TV} + .006\,\text{Piercing} + .001\,\text{MSAT} + .001\,\text{VSAT}$$

A little professional advice... start by checking that the signs make sense.

↑HT ↓GPA ?

Males are taller and have lower GPAs. Confounding.

↑TV ↓GPA

makes sense

↑Piercings ↑GPA

Women get more piercings and have higher GPAs. Confounding.

↑MSAT ↑GPA

We hope

↑VSAT ↑GPA

We hope

Prediction and residual

Suppose we have a woman with a 3.13 GPA who has 0 piercings, is 71 in tall, watches 1 hr TV/wk, MSAT = 600 and VSAT = 610. What is $\widehat{GPA}$?

$$\widehat{GPA} = 2.438 - .010(71) - .005(1) + .006(0) + .001(600) + .001(610)$$

$$= 2.933$$

$$e = obs - pred$$

$$= 3.13 - 2.933$$

$$= .1970$$

Note that $R^2 = .1606$ means that the model explains only 16.06% of the variability in GPA.

The F-stat is another measures of the quality of the model. Large values of F are good.

$F_{5,332} = 12.7 \Rightarrow$ p-value $< .0001$

so we rej $H_0: \beta_0 = \beta_1 = \ldots = \beta_k = 0$

and conclude at least one $\beta \neq 0$.

Which variables are really helping?

	t_{n-k-1}	p-val
HT	-1.792	.0740
TV	-1.259	.2091
Piercings	0.568	.5705
MSAT	2.801	.0054
VSAT	4.609	$< .001$

Piercings not helpful

TV not helpful

HT getting there — but may be gender

MSAT seems to be working

VSAT " " " "

Get rid of "bad" variables

we prefer parsimonious models

$\widehat{GPA} = 2.692 - 0.013\,HT + 0.001\,MSAT + 0.002\,VSAT$

For our 71" tall 600, 610 student we get

$$\widehat{GPA} = 2.692 - 0.013(71) + 0.001(600) + 0.002(610)$$
$$= 3.589$$

$R^2 = .1554$ $\quad F_{3,334} = 20.49$ $\quad$ p-value $< .0001$

$.1554 < .1606$ but not by much

	t_{n-k-1}	p-val
HT	-2.760	0.0061
MSAT	2.646	0.0085
VSAT	5.076	$< .0001$

To test we need normality and independence.

Are the residuals okay?

Residual plots look okay

Normal prob plot okay

Leverage indicates a possible influential point

3D plane for fun

Extra example

Golden State Warriors point scoring.

See MultipleRegression.RMD