Linear models

Mark Andrews

April 5, 2017

Introduction

In linear models, we model the expected value of an outcome variable as a linear function of one or more predictor variables. Even when we know that this is not a great modelling assumption, linear models can still be very informative, especially for exploratory work. In any case, it is hard to progress to more complex and realistic models without first understanding linear models.

For the following, we will use some data from the `psych` package. So, first load that, and a few other goodies:

```
library(MASS)
library(car)
library(psych)
library(ggplot2)
library(pander) # for making nice tables

data(sat.act)
Df <- within(sat.act,
             {gender <- factor(gender)}
)
```

We'll start by predicting ACT (a standardized academic test) scores on the basis of education level (measured on a five point scale):

```
M <- lm(ACT ~ education, data=Df)
pander(summary(M))
```

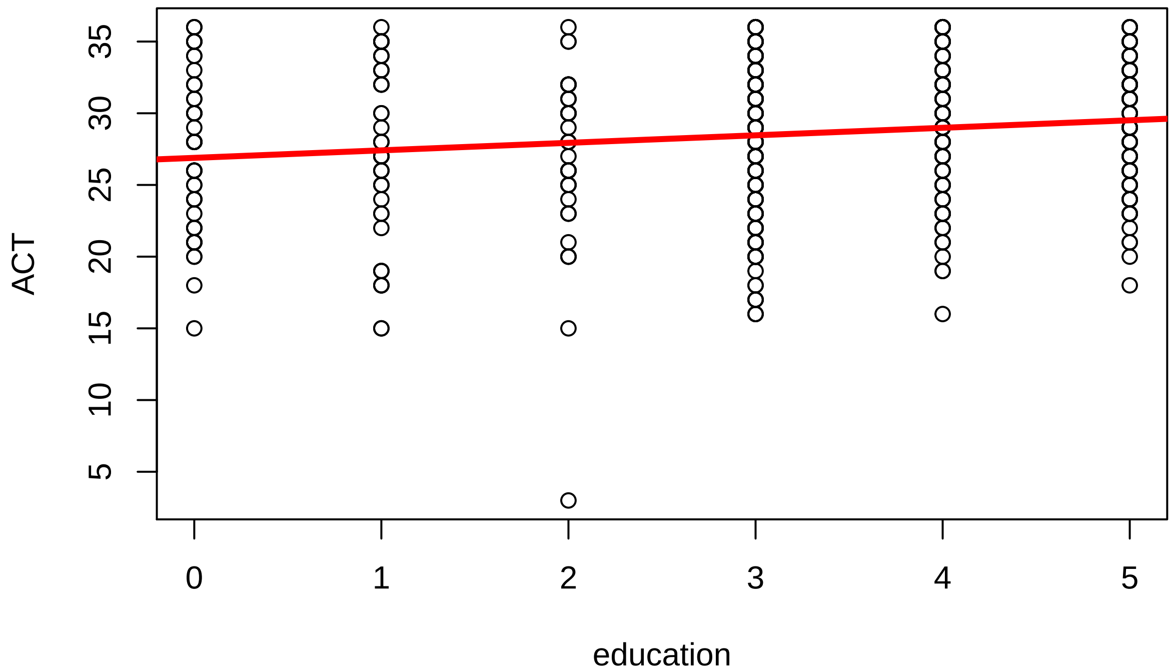
	Estimate	Std. Error	t value	Pr(>|t|)
education	0.524	0.1265	4.14	3.89e-05
(Intercept)	26.89	0.4391	61.23	6.733e-283

Table 2: Fitting linear model: ACT ~ education

Observations	Residual Std. Error	R^2	Adjusted R^2
700	4.769	0.02397	0.02257

We can visualize this as follows:

```
plot(ACT ~ education, data=Df)
abline(M, lwd=3, col='red')
```



Confidence intervals

We can get confidence intervals as follows:

```
confint(M)
```

```
##                  2.5 %     97.5 %
## (Intercept) 26.0270213 27.7513496
## education    0.2755026  0.7724163
```

Predictions

On the basis of our fitted model M, we can make predictions about possible values of the predictor variable.

```
hypothetical.data <- data.frame(education = c(1, 2, 5, 10, 15))
predict(M, newdata=hypothetical.data)
```

```
##        1        2        3        4        5
## 27.41314 27.93710 29.50898 32.12878 34.74858
```

Multiple linear regression

We can add as many predictor variables as we like:

```
M <- lm(ACT ~ education + age + gender, data=Df)
pander(summary(M))
```

	Estimate	Std. Error	t value	Pr(>|t|)
education	0.4789	0.1523	3.143	0.00174
age	0.01623	0.02278	0.7123	0.4765
gender2	-0.4861	0.3798	-1.28	0.2011
(Intercept)	26.93	0.5919	45.5	1.025e-210

Table 4: Fitting linear model: ACT ~ education + age + gender

Observations	Residual Std. Error	R^2	Adjusted R^2
700	4.768	0.0272	0.02301

Collinearity

We'll evaluate multicollinerity using Variance Inflation Factor (VIF):

```
vif(M)
```

```
## education       age    gender
##  1.450002  1.439585  1.014574
```

General linear models

We can use predictors that categorical as well as continuous in our model. Here, we investigate how the post treatment weight of a patient differs from their pre treatment weight, for three different types of therapy (control, CBT, family therapy):

```
M <- lm(Postwt ~ Prewt + Treat, data=anorexia)
pander(summary(M))
```

	Estimate	Std. Error	t value	Pr(>|t|)
Prewt	0.4345	0.1612	2.695	0.00885
TreatCont	-4.097	1.893	-2.164	0.034
TreatFT	4.563	2.133	2.139	0.03604
(Intercept)	49.77	13.39	3.717	0.0004101

Table 6: Fitting linear model: Postwt ~ Prewt + Treat

Observations	Residual Std. Error	R^2	Adjusted R^2
72	6.978	0.2777	0.2458

Model evaluation

We can compare any two linear models using the generic `anova` function. Here, we'll use this to test the significance of an interaction:

```
load('../data/beautyeval.Rda')

# Model 1: evaluation score as a function of beauty and sex
M.additive <- lm(eval ~ beauty + sex, data=beautydata)

# Model 2: evaluation score as a function of beauty and sex and their interaction
M.interaction <- lm(eval ~ beauty * sex, data=beautydata)

# Is there an interaction?
pander(anova(M.additive, M.interaction), missing='')
```

Table 7: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
460	132.8				
459	131.9	1	0.8912	3.101	0.07891