

# Linear models

Mark Andrews

July 5, 2018

## 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(dplyr)
library(ggplot2)
library(pander) # for making nice tables

Df <- mutate(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))
```

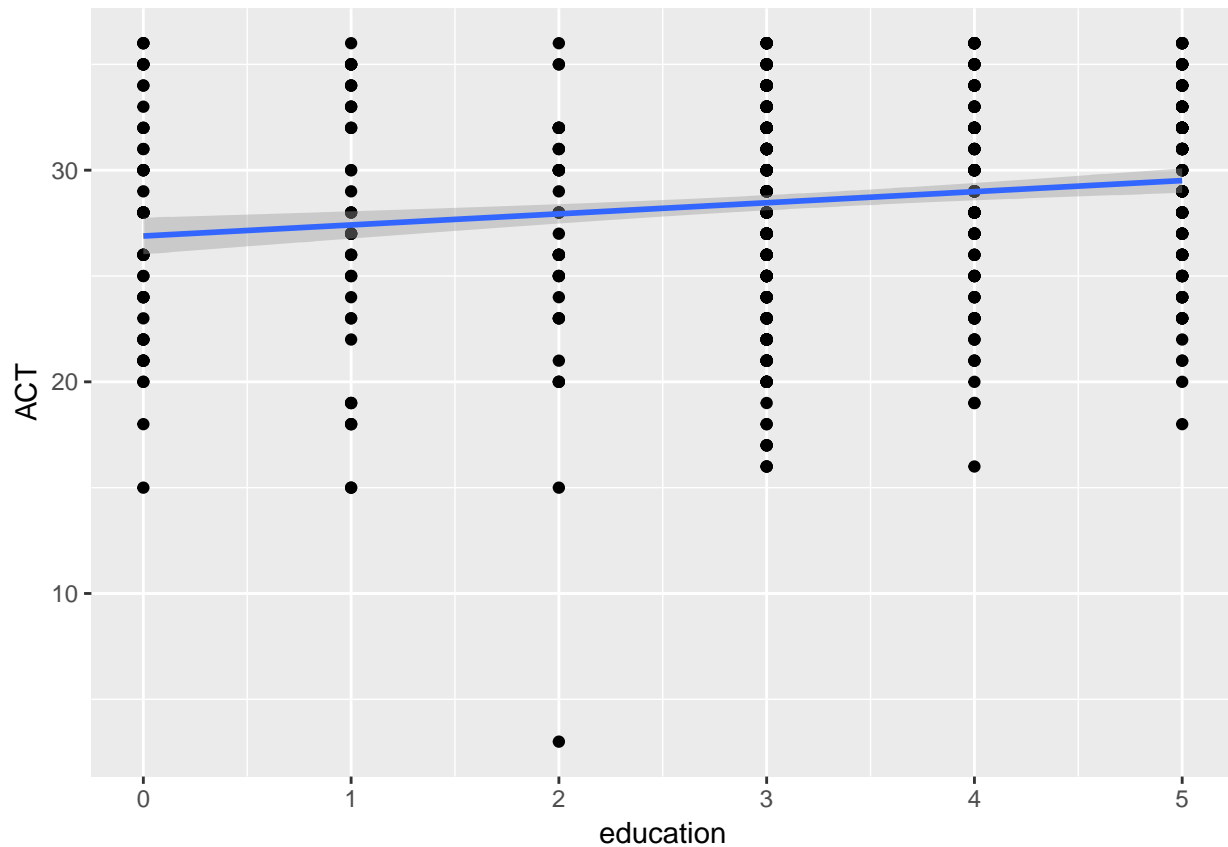
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	26.89	0.4391	61.23	6.733e-283
education	0.524	0.1265	4.14	3.89e-05

Table 2: Fitting linear model:  $ACT \sim education$

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

We can visualize this as follows:

```
ggplot(Df,
       aes(x= education, y=ACT)) +
  geom_point() +
  stat_smooth(method='lm')
```



## 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 )
<b>(Intercept)</b>	26.93	0.5919	45.5	1.025e-210
<b>education</b>	0.4789	0.1523	3.143	0.00174
<b>age</b>	0.01623	0.02278	0.7123	0.4765
<b>gender2</b>	-0.4861	0.3798	-1.28	0.2011

Table 4: Fitting linear model:  $ACT \sim education + age + gender$

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

## Collinearity

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

```
vif(M)
```

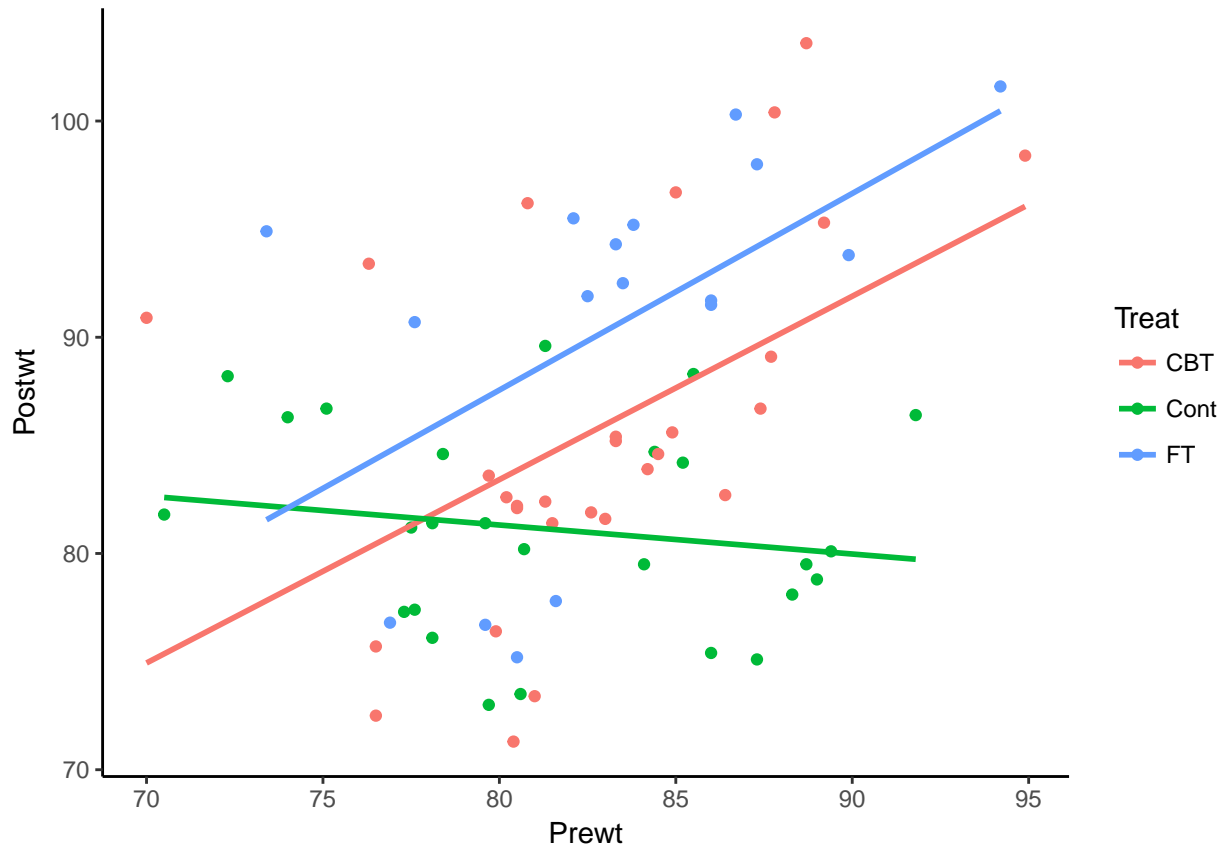
```
## education      age      gender
##  1.450002  1.439585  1.014574
```

## 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).

First, we'll visualize the data (we'll turn off standard error shading to allow the lines to be seen more easily):

```
ggplot(anorexia,
       aes(x = Prewt, y = Postwt, col=Treat)) +
  geom_point() +
  stat_smooth(method='lm', se=F) +
  theme_classic()
```



Now, we'll do a *varying intercept*, which is also known as an *ANCOVA*:

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

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

Table 6: Fitting linear model:  $\text{Postwt} \sim \text{Prewt} + \text{Treat}$

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

We also do a *varying slopes and varying intercepts* model. This is a type of interaction model:

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

	Estimate	Std. Error	t value	Pr(> t )
<b>(Intercept)</b>	15.58	21.21	0.7345	0.4653
<b>Prewt</b>	0.848	0.2561	3.312	0.001507

	Estimate	Std. Error	t value	Pr(> t )
<b>TreatCont</b>	76.47	28.35	2.698	0.008852
<b>TreatFT</b>	-0.7575	34.55	-0.02192	0.9826
<b>Prewt:TreatCont</b>	-0.9822	0.3442	-2.853	0.005776
<b>Prewt:TreatFT</b>	0.06124	0.4155	0.1474	0.8833

Table 8: Fitting linear model: Postwt ~ Prewt \* Treat

Observations	Residual Std. Error	$R^2$	Adjusted $R^2$
72	6.565	0.3794	0.3324

## Model evaluation

We can compare any two linear models using the generic `anova` function. Here, we'll use this to test whether the varying slopes and intercepts model is a better fit to the data than the just varying intercepts model:

```
model_comparison <- anova(M, M_interaction)
pander(model_comparison, missing='')
```

Table 9: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
68	3311				
66	2845	2	466.5	5.411	0.006666