Moving beyond linearity in predictors

UBCO MDS — DATA 570

Introducing Categorical Predictors

- So far we've assumed that our predictors are continuous valued when we've fit a regression.

- But there is no real problem if instead we have categorical values

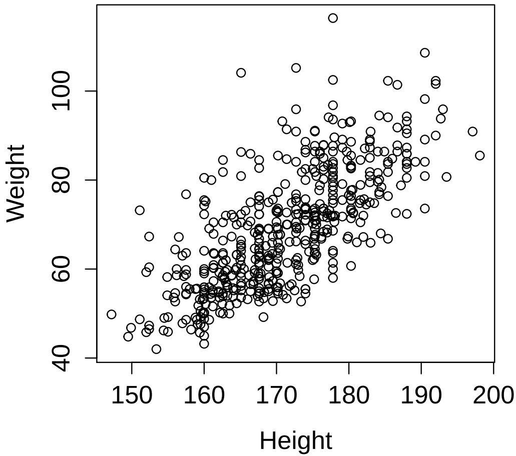
- Let's motivate this through an example...

Example

- Data was collected on 507 adult participants with respect to height (in cm) and weight (in kg). You can find this in the `gclus` library as `body`.

	Weight	Height
1	65.60	174.00
2	71.80	175.30
3	80.70	193.50
4	72.60	186.50
⋮	⋮	⋮

- Scatterplot Weight vs height

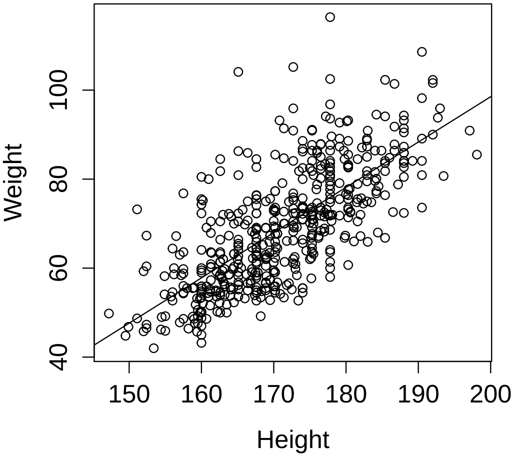


- Weight $= b_o + b_1(\text{Height})$

	Coefficient	t value	Sig
(Intercept)	-105.0113	-13.93	0.0000
Height	1.0176	23.13	0.0000

- $r^2 = 0.5145$

- Weight vs Height with Regression Line

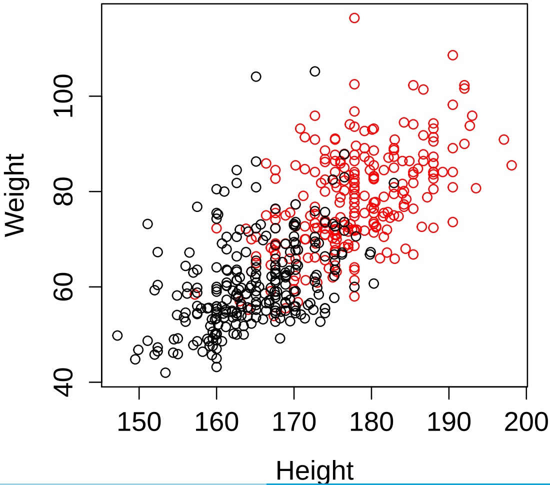


We Know More Though!

- Also, their gender was recorded.

	Weight	Height	Gender
1	65.60	174.00	Male
2	71.80	175.30	Male
3	80.70	193.50	Male
4	72.60	186.50	Female
5	78.80	187.20	Male
⋮	⋮	⋮	⋮

- Weight vs Height coloured by Gender



Moving Forward

- The question now becomes: how do we incorporate categorical variables into this model?
- We need to create 'dummy' variables.
- A dummy variable is a variable created to assign numerical value to levels of categorical variables. Each dummy variable represents one category of the predictor variable.

Moving Forward

- For a binary (two-option) variable like Gender, it's easy.
- Gender: 1 - male, 0 - female
- Weight $= b_0 + b_1$(Height)$+ b_2$(Gender)$+$Error
- b_0 is the average weight among females when Heights take on the value 0,

  $b_0 + b_1$ is the average weight among males...

  and b_1 is the average difference in weight between males and females...

Example

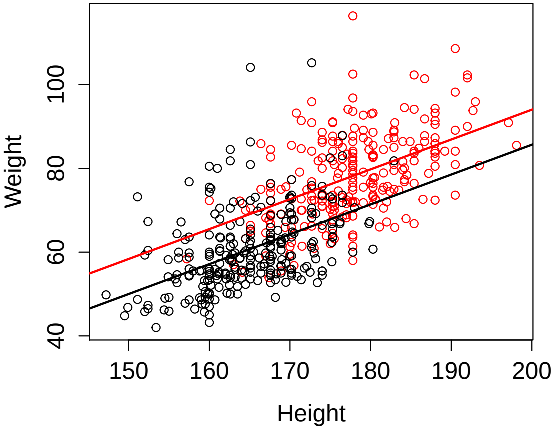
- Weight $= b_o + b_1(\text{Height}) + b_2(\text{Gender})$

	Coefficient	t value	Sig
(Intercept)	-56.9495	-6.04	0.0000
Height	0.7130	12.49	0.0000
Gender	8.3660	7.80	0.0000

- Give the estimated equation of a line for females:
  $\hat{y} = -56.9495 + 0.7130(Height)$

- Give the estimated equation of a line for males:
  $\hat{y} = -56.9495 + 8.3660 + 0.7130(Height)$

Weight vs Height coloured by Gender with separate lines plotted (red=male, black=female)



Linear Regression with Interactions

- With a continuous response Y, and (at least) two predictors X_1, X_2 we assume that we can write the following model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- Y is the response variable at i
- X_j is the j^{th} predictor variable
- β_0 is the true intercept (unknown)
- β_j are the true slopes (unknown)
- ϵ is the true error, assumed $\epsilon_i \sim N(0, \sigma^2)$.

Assumptions and Interpretation

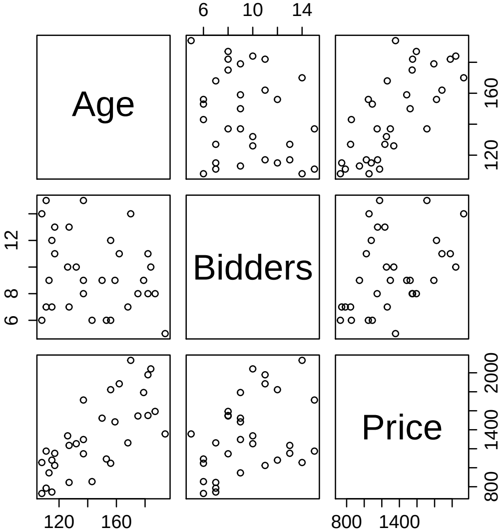
- The assumptions of multiple linear regression with interactions are not different from simple linear regression or multiple linear regression.
- Interpretation of β_j becomes trickier than in 'regular' multiple regression.
- Why?

- The data give the selling price at auction of 32 antique grandfather clocks. Also recorded is the age of the clock and the number of people who made a bid.

- It is available in `https://drive.google.com/open?id=1yvDDNzmpJrDfbLI744fNSgrlXwvY-bLm`

	Age	Bidders	Price
1	127	13	1235
2	115	12	1080
⋮	⋮	⋮	⋮

Scatterplot Matrix — shows all pairwise scatterplots.



Example

Model	Coefficients	t	Sig.
(Intercept)	322.7544	1.10	0.2806
Age	0.8733	0.43	0.6688
Bidders	-93.4099	-3.14	0.0039
Age:Bidders	1.2979	6.15	0.0000

- $r_a^2 = 0.9495$
- Age is not considered significant (p-value = 0.6688) but the interaction effect is...can we remove Age as a predictor?

Example

- Price = Age ($r_a^2 = 0.5177$)

Model	Coefficients	t	Sig.
(Intercept)	-191.6576	-0.73	0.4733
Age	10.4791	5.85	0.0000

Example

- Price = Bidders ($r_a^2 = 0.1276$)

Model	Coefficients	t	Sig.
(Intercept)	806.4049	3.50	0.0015
Bidders	54.6362	2.35	0.0254

Example

- Price = Age + Bidders ($r_a^2 = 0.8853$)

Model	Coefficients	t	Sig.
(Intercept)	-1336.7221	-7.71	0.0000
Age	12.7362	14.11	0.0000
Bidders	85.8151	9.86	0.0000

Example

- Price = Bidders +Age*Bidders ($r_a^2 = 0.9509$)

Model	Coefficients	t	Sig.
(Intercept)	447.0965	7.841	0.0000
Bidders	-105.6312	-11.713	0.0000
Bidders:Age	1.3850	22.449	0.0000

Non-linearity

- Suppose we have a case where the response has a non-linear relationship with the predictor(s).
- We will treat these cases in more detail eventually...
- But for a first go-around, what if there's a quadratic relationship?

Non-linearity

- Easy. Fit a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

- Basically, if we square (or otherwise transform) the original predictor, we can still fit a 'linear' model for the response.

- Though it's important to keep in mind the change in interpretation for β_1 and β_2, for example.

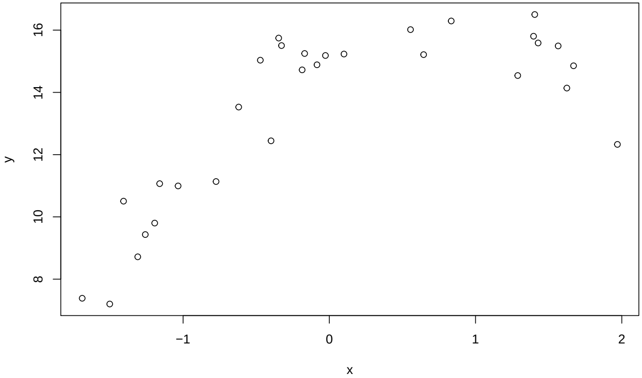
Example: Quadratic Simulation

- We simulate 30 values from the following model

$$Y = 15 + 2.3x - 1.5x^2 + \epsilon$$

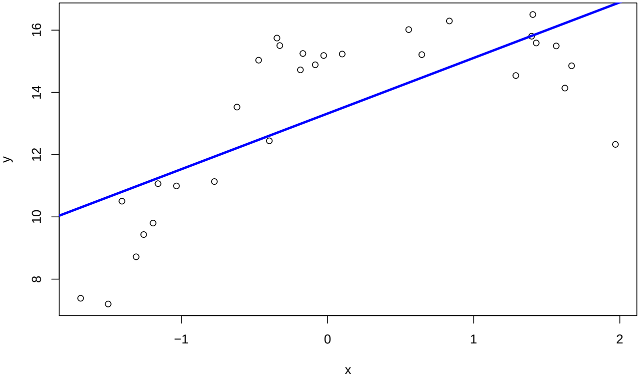
- Where x and ϵ are standard normally distributed.

Example: Quadratic Simulation



```
Call: lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5110 -1.3096 -0.1721  1.7138  3.0435

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.3206     0.3592  37.081  < 2e-16 ***
x             1.7866     0.3248   5.501 7.06e-06 ***
---
Signif. codes:  0‘***’,  0.001 ‘**’, 0.01 ‘*’ 0.05 ‘.’,  0.1 ‘’,  1

Residual standard error: 1.967 on 28 degrees of freedom
Multiple R-squared:  0.5194,Adjusted R-squared:  0.5023
F-statistic: 30.26 on 1 and 28 DF,  p-value: 7.059e-06
```

Example: Quadratic Simulation

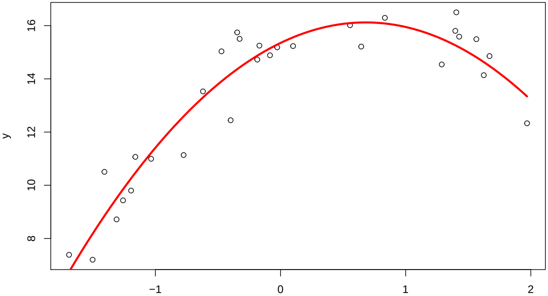


Note: $x2 = x^2$

```
Call: lm(formula = y ~ x + x2) or lm(formula = y ~ x + I(x^2))

Residuals:
     Min       1Q   Median       3Q      Max
-1.73833 -0.49510 -0.09868  0.58328  1.66181

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.3556     0.2509   61.20  < 2e-16 ***
x             2.2773     0.1529   14.89 1.54e-14 ***
x2           -1.6702     0.1578  -10.59 4.14e-11 ***
---
Signif. codes:  0‘***’,  0.001 ‘**’, 0.01 ‘*’ 0.05 ‘.’,  0.1 ‘’,  1

Residual standard error: 0.8829 on 27 degrees of freedom
Multiple R-squared:  0.9067,Adjusted R-squared:  0.8998
F-statistic: 131.2 on 2 and 27 DF,  p-value: 1.244e-14
```

Example: Quadratic Simulation

```
xx <- seq(-2,2, length.out=250)
lines(xx, predict(lm(y~x+x+I(x^2)),
data.frame(x=xx)), col='red')
```



Piecewise Polynomials

- Instead of a single polynomial in X over its whole domain, we can rather use different polynomials in regions defined by knots.
- $y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \varepsilon_i & \text{if } x_i < c \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \varepsilon_i & \text{if } x_i \geq c \end{cases}$
- c is the knot or breakpoint/changepoint.

Piecewise Polynomials

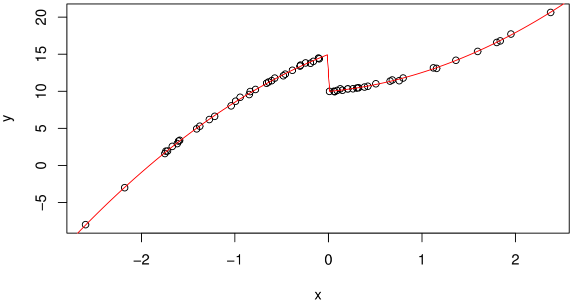
```
x=rnorm(60)
e=rnorm(60,0,0.1)
y=(15+5*x-1.5*x^2)*I(x<=0)
+(10+1*x+1.5*x^2)*I(x>0)+e
fit=lm(y~ I(x>0)+I(x*(x<=0))+I(x*(x>0))
+I(x^2*(x<=0))+I(x^2*(x>0)))
```

Piecewise Polynomials

```
summary(fit)
Coefficients:
                       Estimate      t value  Pr(>|t|)
(Intercept)            14.97680       316.56   < 2e-16
I(x > 0)TRUE           -4.99958       -81.08   < 2e-16
I(x * (x <= 0))         4.93112        51.95   < 2e-16
I(x * (x > 0))          1.14078        10.56  9.68e-15
I(x^2 * (x <= 0))      -1.51413       -38.04   < 2e-16
I(x^2 * (x > 0))        1.41463        28.45   < 2e-16
```

Piecewise Polynomials

```
plot(x,y)
xx <- seq(-3,3, length.out=250)
lines(xx, predict(fit, data.frame(x=xx)), col='re
```



Fitting splines

```
library(npreg);ordering <- order(x);
xs=x[ordering]; ys=y[ordering]
mod.ss <- ss(xs, ys, nknots =6)
mod.smsp <- smooth.spline(x, y, nknots = 6)
plot(xs, ys,xlab="x",ylab="y")
lines(xx,predict(fit,data.frame(x=xx)),col=1)
lines(xs,mod.ss$y, lty = 2, col = 2, lwd = 2)
lines(xs, mod.smsp$y, lty = 3, col = 3, lwd = 2)
legend("bottomright",legend = c("Piecewise",
"ss", "smooth.spline"),lty = 1:3, col = 1:3,
lwd = 2, bty = "n")
```

Fitting splines