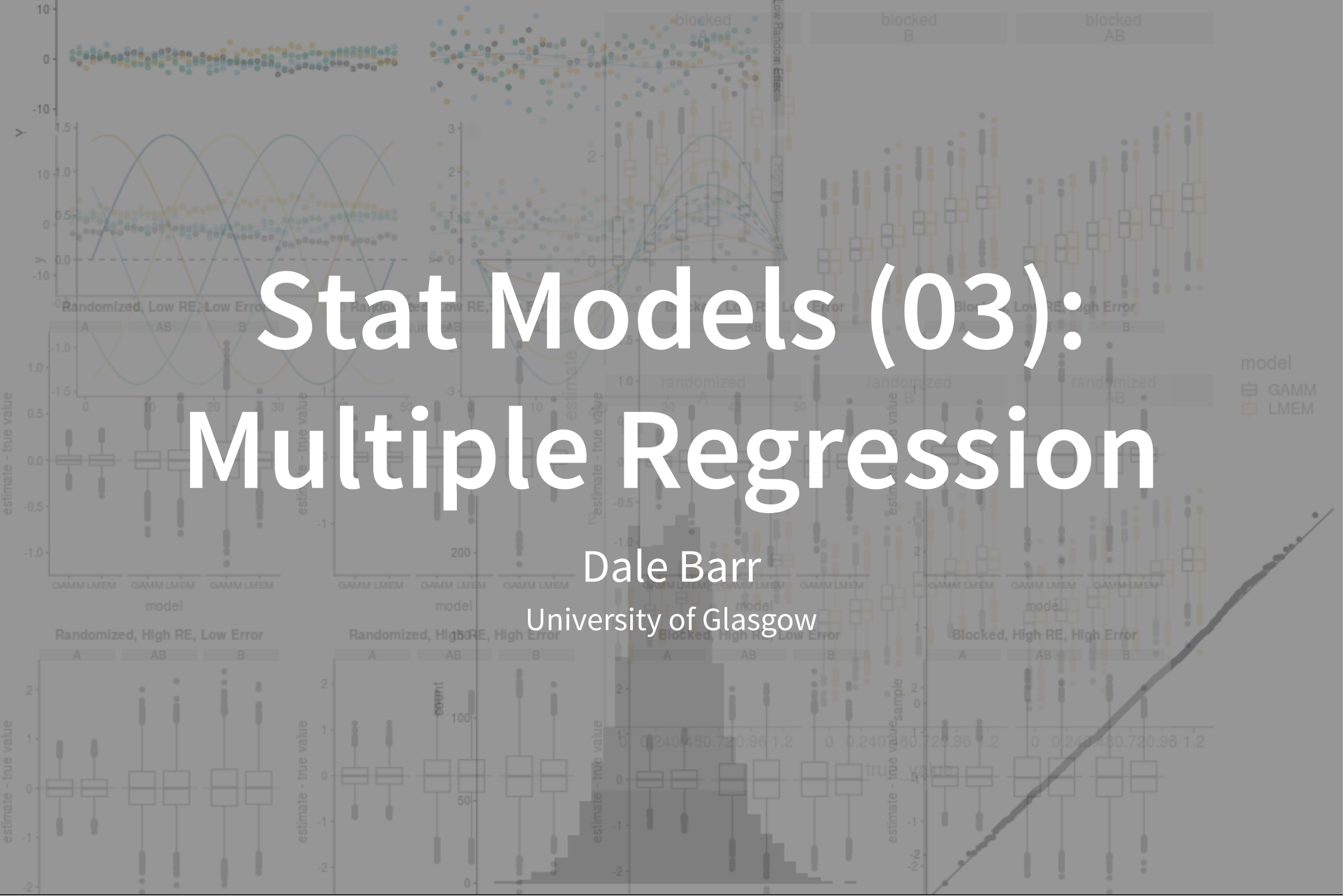


Stat Models (03): Multiple Regression

Dale Barr

University of Glasgow

moving beyond simple regression

- correlation matrices
- multiple regression
- comparing models
- coding categorical predictors

correlation matrices

how many correlations for n variables?

Note that $\rho_{xy} = \rho_{yx}$.

For any n measures, you can calculate $\frac{n(n-1)}{2}$ unique pairwise correlations between measures. So, if you have six measurements, you have

$$\frac{6(6-1)}{2} = \frac{30}{2} = 15$$

unique correlations.

grades

`grades.csv`

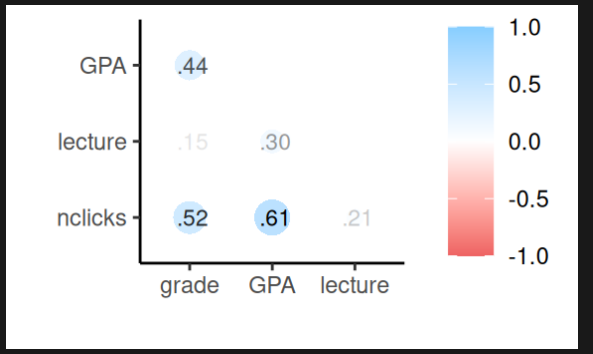
grade	GPA	lecture	nclicks
4.00	2.52	10	108
3.02	2.73	7	93
1.47	1.55	4	71
1.21	2.55	10	101
1.90	2.46	9	84
3.38	2.25	6	93
4.00	3.45	8	135
3.47	2.96	6	126
2.59	3.22	7	109
1.87	2.64	7	74

- 100 rows (students)
- `grade`: grade at end of semester
- `lecture`: number of lectures attended (out of 10)
- `nclicks`: engagement with online materials

How well does engagement (measured by lecture attendance / clicks on materials) predict grade?

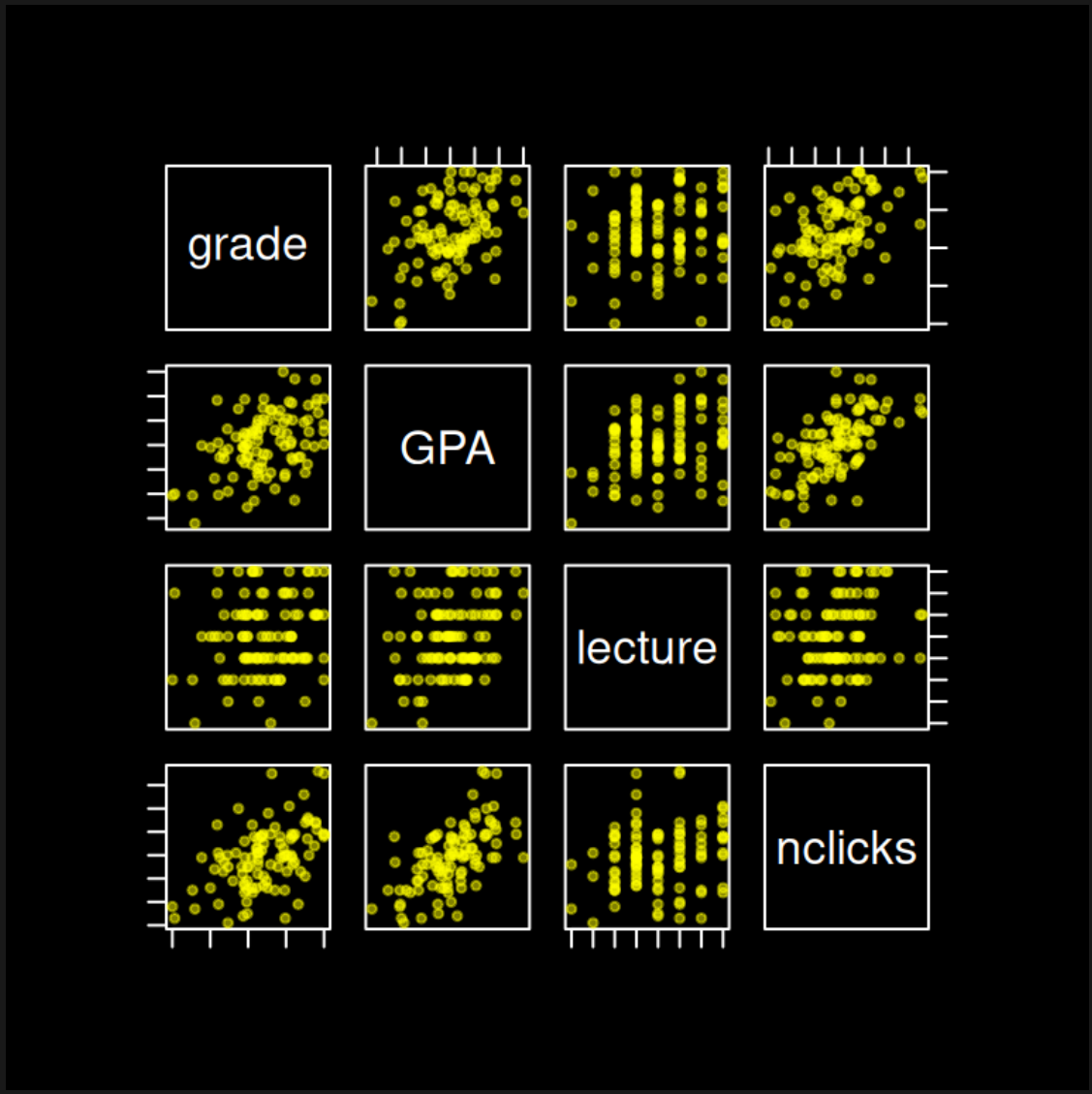
correlation matrix

term	grade	GPA	lecture	nclicks
grade	1.00	.44	.15	.52
GPA	.44	1.00	.30	.61
lecture	.15	.30	1.00	.21
nclicks	.52	.61	.21	1.00



- Each row x col entry corresponds to the **bivariate correlation** between those variables
- upper & lower triangle; diagonal
- Symmetric

scatterplots



multiple regression

General model for data with m predictors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_m X_{mi} + e_i$$

individual Xs can be any combination of continuous and categorical predictors (and their interactions)

Each β_j is the *partial effect of X_j on Y_i holding all other Xs constant*

NB: These models are only valid for data with *one* observation on the DV (response variable) per participant. In experimental psychology, this is very rare. Also it assumes there are no interactions in the model.

example questions

- Are lecture attendance and engagement with online materials associated with higher course grades?
- Does this relationship hold after controlling for overall GPA?

estimation

$$grade_i = \beta_0 + \beta_1 lectures_i + \beta_2 nclicks_i + e_i$$

```
mod <- lm(grade ~ lecture + nclicks, data = grades)
summary(mod)
```

```
Call:
lm(formula = grade ~ lecture + nclicks, data = grades)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77648 -0.45633  0.04778  0.49755  1.54089

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.100047   0.609603  -1.805   0.0743 .
lecture      0.024888   0.045623   0.546   0.5867
nclicks      0.033941   0.005903   5.750 1.04e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7729 on 97 degrees of freedom
Multiple R-squared:  0.272, Adjusted R-squared:  0.257
F-statistic: 18.12 on 2 and 97 DF,  p-value: 2.061e-07
```

mean centering predictors

- makes y-intercept more interpretable
- new pred = old pred - mean(old pred)
  - `lecture_c = lecture - mean(lecture)`
  - `nclicks_c = nclicks - mean(nclicks)`

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.433696   0.077288  31.489  < 2e-16 ***
lecture_c   0.024888   0.045623   0.546    0.587
nclicks_c   0.033941   0.005903   5.750 1.04e-07 ***
```

which predictor is more important?

To compare β weights, you need to standardize predictors

- `lecture_z = (lecture - mean(lecture)) / sd(lecture)`
- `nclicks_z = (nclicks - mean(nclicks)) / sd(nclicks)`

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.43370    0.07729  31.489  < 2e-16 ***
lecture_z    0.04332    0.07942   0.546    0.587
nclicks_z    0.45665    0.07942   5.750 1.04e-07 ***
```

- standardized

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.43370    0.07729  31.489  < 2e-16 ***
lecture_z    0.04332    0.07942   0.546    0.587
nclicks_z    0.45665    0.07942   5.750 1.04e-07 ***
```

- mean-centered

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.433696   0.077288  31.489  < 2e-16 ***
lecture_c   0.024888   0.045623   0.546    0.587
nclicks_c   0.033941   0.005903   5.750 1.04e-07 ***
```

- original (raw)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.100047   0.609603  -1.805   0.0743 .
lecture      0.024888   0.045623   0.546   0.5867
nclicks      0.033941   0.005903   5.750 1.04e-07 ***
```

relation between multiple regression and correlation

Each β_j corresponds to the partial correlation between X_j and Y controlling for other predictors.

model comparison

Is engagement (as measured by lecture attendance and downloads) positively associated with final course grade *above and beyond* student ability (as measured by GPA)?

model comparison strategy

Compare “base” to “augmented” model with focal predictors

- Base

  $grade_i = \beta_0 + \beta_1 GPA_i + e_i$
- Augmented

  $grade_i = \beta_0 + \beta_1 GPA_i + \beta_2 lecture_i + \beta_3 nclicks_i + e_i$

$$H_0 : \beta_2 = \beta_3 = 0$$

F-test on residual sum of squares (RSS). If $p < \alpha$, reject H_0.

- Base

  $grade_i = \beta_0 + \beta_1 GPA_i + e_i$
- Augmented

  $grade_i = \beta_0 + \beta_1 GPA_i + \beta_2 lecture_i + \beta_3 nclicks_i + e_i$

model	RSS	df
base	64.0	98
augmented	56.1	96

$F(2, 96) = 6.772$, $p = 0.002$

categorical (nominal) predictors

dummy coding two-level nominal variables

Arbitrarily assign one level to 0; assign the other to 1.

R will do this automatically if a predictor is of type “character” or “factor” instead of numeric (choosing as baseline the level that would be alphabetized before all other levels)

NB: sign of the variable depends on the choice of baseline!

factors with $k > 2$

Arbitrarily choose one level as "baseline" level. Need $k - 1$ predictors, each contrasting a target level with baseline.

$k = 3$

	A2v1	A3v1
A_1	0	0
A_2	1	0
A_3	0	1

$k = 4$

	A2v1	A3v1	A4v1
A_1	0	0	0
A_2	1	0	0
A_3	0	1	0
A_4	0	0	1

Bodyweight over the seasons

season	bodyweight_kg
winter	96.8707
winter	102.0794
winter	101.3670
winter	106.5152
winter	106.0500
spring	108.9893

- four levels: winter, spring, summer, fall

Coding the predictor

season	spring_v_winter	summer_v_winter	fall_v_winter
winter	0	0	0
spring	1	0	0
summer	0	1	0
fall	0	0	1

Fitting the model

$$BW_i = \beta_0 + \beta_1 SPvW + \beta_2 SUvW + \beta_3 FvW + e_i$$

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     102.57645    1.71619  59.770   <2e-16 ***
spring_v_winter  -0.03665    2.42705  -0.015    0.988
summer_v_winter   1.02200    2.42705   0.421    0.679
fall_v_winter    -0.98818    2.42705  -0.407    0.689
```

Main effect of season?

Use model comparison.

- Base:

  $BW_i = \beta_0 + e_i$
- Augmented:

  $BW_i = \beta_0 + \beta_1 SPvW + \beta_2 SUvW + \beta_3 FvW + e_i$

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

model	RSS	df
base	245.7	19
augmented	235.6	16

$F(3, 16) = 0.229$, $p = 0.875$

This is identical to a one-way ANOVA!

Watch out for nominal variables disguised as numbers!

Imagine you got a dataset with `season` coded as a single variable where: 1=winter, 2=spring, 3=summer, 4=fall. R will treat this as a single numeric predictor and the output will be nonsense.