ENV 710

model selection

Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

roadmap

- questions?
- download from Week 9: `kidiq.csv` & `cong.csv`

where we are

multivariate linear models

model selection/reduction

interactions

explore your data ①

head off problems before they occur and know what to look out for as you proceed

- look for surprises (outliers, typos, etc.)
- look for gross deviations from assumptions (e.g., lack of normality, lack of linear relationship between DV and IV's, strong multicollinearity
- determine correct model based on type of DV (e.g., continuous or discrete variable)

build your model ②

based on hypotheses, study design, & DV's

- determine model structure
- should IV's be treated as continuous or nominal?
- do you need to standardize IV's?
- do you need to include interactions, exponentials or polynomials?
- do you need to include random effects?
- do you need to transform DV's?

refine your model ③

determine best model based on hypotheses

- do you need the most parsimonious model?
- remove non-significant terms?
- use F-test for comparing nested models
- use AIC for non-nested models

verify your model ④

evaluate model assumptions

- check residuals plots for normality, homoscedasticity and influential data points
- check for other assumptions, such as overdispersion (Poisson GLM)
- if assumptions aren't met, need to adapt by potentially transforming data, standardizing IV's or restructuring the model

interpret your model ⑤

make your conclusions

- is omnibus H0 statistically significant?
- does model explain much variance in DV?
- assess effect sizes (β's)
- graph results to demonstrate effects
- accept/reject hypotheses
- make predictions

model selection

model selection is a trade-off between

complexity and fit

number of explanatory variables in the model

how well the model fits the data

reflects conflicting interests…

- describe the data reasonably well
- build a model simple enough to be interpretable

statistical modeling basics

- null model: the mean is the only parameter → no explanatory power

$$Y = \beta_0 + \varepsilon$$

- saturated model: includes a parameter for every data point ($k = n$) → no explanatory power

- maximal model: contains all factors, interactions and covariates of interest

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

- minimum adequate model: only includes explanatory variables that improve the fit of the model to the data

prefer the most parsimonious model

compare models in 3 potential ways:

- partial F-test
- Akaike Information Criterion (AIC)
- adjusted R^2

1 – kid cognition

Do the characteristics of mothers, specifically their age, IQ, work, and high school affect their children's cognition scores?

load the data

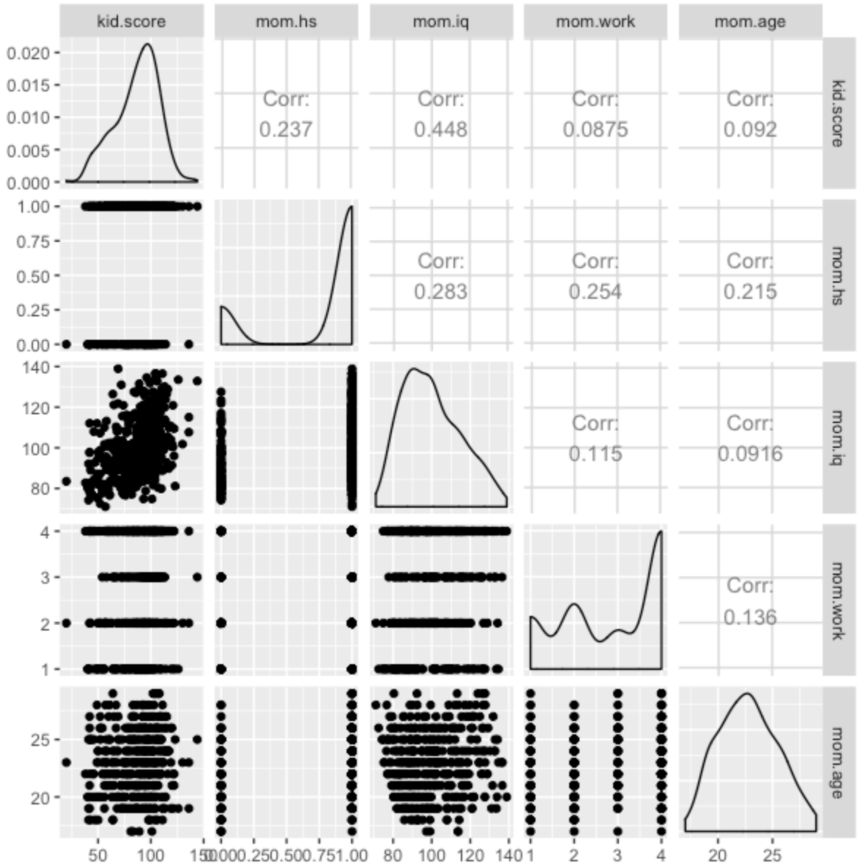
```
kdat <- read.csv("kidiq.csv", header = T)
 sapply(kdat, class)
```

1 – kid cognition

Do the characteristics of mothers, specifically their age, IQ, work, and high school affect their children's cognition scores?

look at the data

```
ggpairs(kdat)
```



1 – kid cognition

Do the characteristics of mothers, specifically their age, IQ, work, and high school affect their children's cognition scores?

run a model

```
lm1 <- lm(kid.score ~ factor(mom.hs) +
         mom.iq + mom.age +
          factor(mom.work))
```

1. what do the coefficients represent?
2. would you reduce this model?

```
Call:
lm(formula = kid.score ~ factor(mom.hs) + mom.iq + mom.age +
    factor(mom.work))

Residuals:
    Min      1Q  Median      3Q     Max
-54.414 -12.095   2.015  11.653  49.100

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       20.27273    9.39320   2.158   0.0315 *
factor(mom.hs)1    5.43466    2.32518   2.337   0.0199 *
mom.iq             0.55288    0.06138   9.008   <2e-16 ***
mom.age            0.21629    0.33351   0.649   0.5170
factor(mom.work)2  2.98266    2.81289   1.060   0.2896
factor(mom.work)3  5.48824    3.25239   1.687   0.0922 .
factor(mom.work)4  1.41929    2.51621   0.564   0.5730
---

Residual standard error: 18.14 on 427 degrees of freedom
Multiple R-squared:  0.2213, Adjusted R-squared:  0.2103
F-statistic: 20.22 on 6 and 427 DF,  p-value: < 2.2e-16
```

1 – kid cognition

run a new model

```
lm2 <- lm(kid.score ~ factor(mom.hs) +
        mom.iq + factor(mom.work))
```

```
lm(formula = kid.score ~ factor(mom.hs) + mom.iq + factor(mom.work))

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       24.89567    6.11293   4.073 5.54e-05 ***
factor(mom.hs)1    5.71119    2.28420   2.500   0.0128 *  
mom.iq             0.55348    0.06133   9.025  < 2e-16 ***
factor(mom.work)2  2.88285    2.80678   1.027   0.3050    
factor(mom.work)3  5.61569    3.24425   1.731   0.0842 .  
factor(mom.work)4  1.48960    2.51217   0.593   0.5535    
---
Residual standard error: 18.13 on 428 degrees of freedom
Multiple R-squared:  0.2205, Adjusted R-squared:  0.2114 
F-statistic: 24.22 on 5 and 428 DF,  p-value: < 2.2e-16
```

would you reduce this model?

I – kid cognition

run another new model

```
lm3 <- lm(kid.score ~ factor(mom.hs) +
       mom.iq)
```

```
Call:
lm(formula = kid.score ~ factor(mom.hs) + mom.iq)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     25.73154    5.87521   4.380 1.49e-05 ***
factor(mom.hs)1  5.95012    2.21181   2.690  0.00742 ** 
mom.iq           0.56391    0.06057   9.309  < 2e-16 ***
---

Residual standard error: 18.14 on 431 degrees of freedom
Multiple R-squared:  0.2141,	Adjusted R-squared:  0.2105 
F-statistic: 58.72 on 2 and 431 DF,  p-value: < 2.2e-16
```

would you reduce this model?

1 – kid cognition

compare models

```
AIC(lm1, lm2, lm3)
    df      AIC
lm1  8 3756.033
lm2  7 3754.460
lm3  4 3751.989
```

which model do you keep?

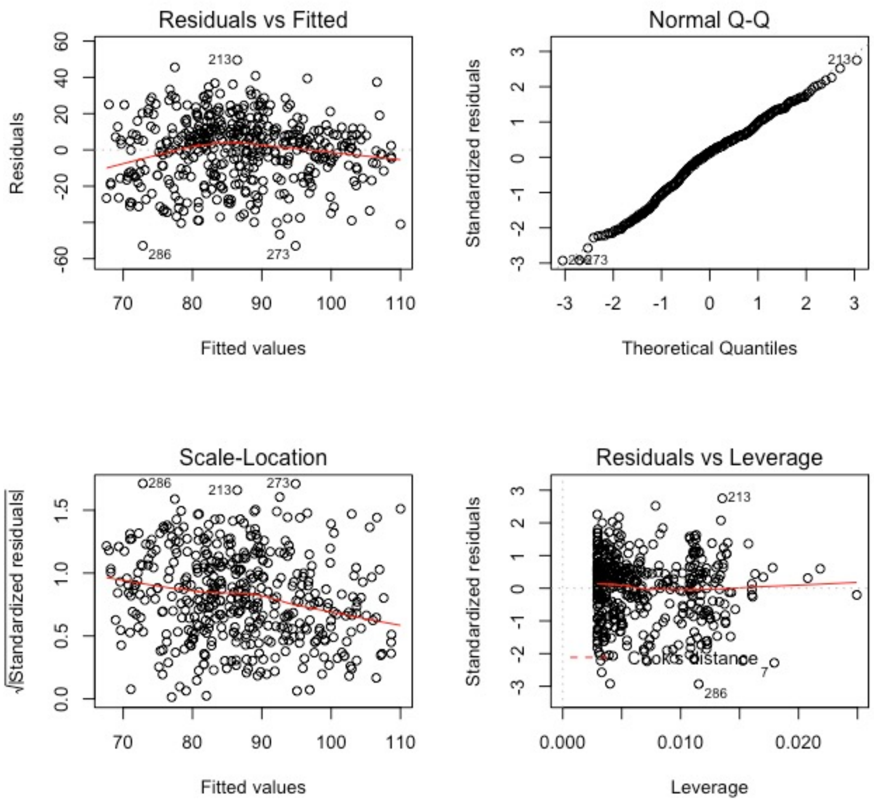
```
c(summary(lm1)$adj.r.squared, summary(lm2)$adj.r.squared, summary(lm3)$adj.r.squared)
[1] 0.2103350 0.2114041 0.2104999

anova(lm1, lm2, lm3)
Analysis of Variance Table

Model 1: kid.score ~ factor(mom.hs) + mom.iq + mom.age + factor(mom.work)
Model 2: kid.score ~ factor(mom.hs) + mom.iq + factor(mom.work)
Model 3: kid.score ~ factor(mom.hs) + mom.iq
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    427 140471
2    428 140609 -1   -138.36 0.4206 0.5170
3    431 141757 -3  -1147.93 1.1632 0.3235
```

1 – kid cognition

```
lm3 <- lm(kid.score ~ factor(mom.hs) + mom.iq)
```

```
Call:
lm(formula = kid.score ~ factor(mom.hs) + mom.iq)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     25.73154    5.87521   4.380 1.49e-05 ***
factor(mom.hs)1  5.95012    2.21181   2.690  0.00742 **
mom.iq           0.56391    0.06057   9.309  < 2e-16 ***
---
Residual standard error: 18.14 on 431 degrees of freedom
Multiple R-squared:  0.2141, Adjusted R-squared:  0.2105
F-statistic: 58.72 on 2 and 431 DF,  p-value: < 2.2e-16
```

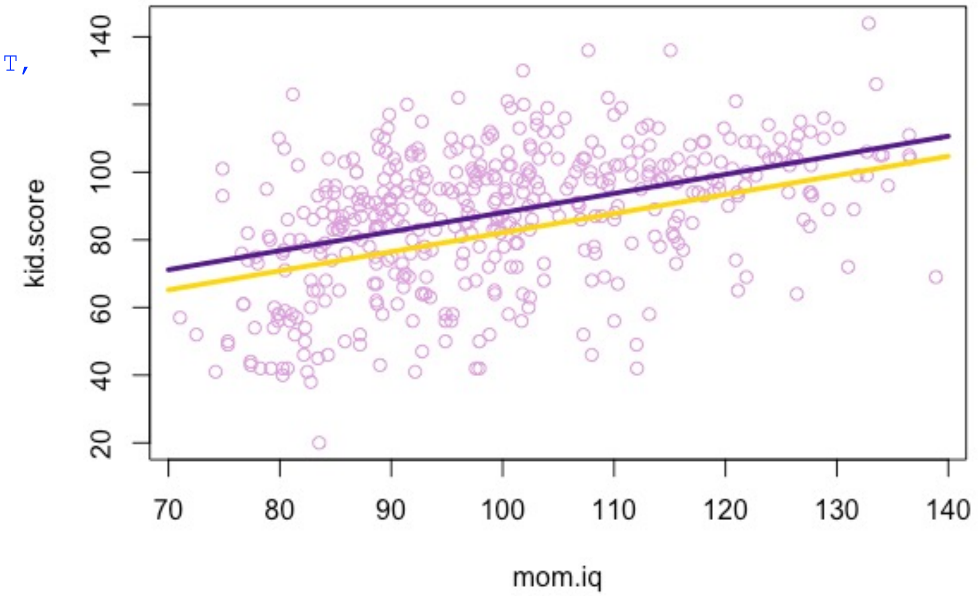
sketch the relationship between Mom IQ and kid score given the mom did or did not graduate high school



I – kid cognition

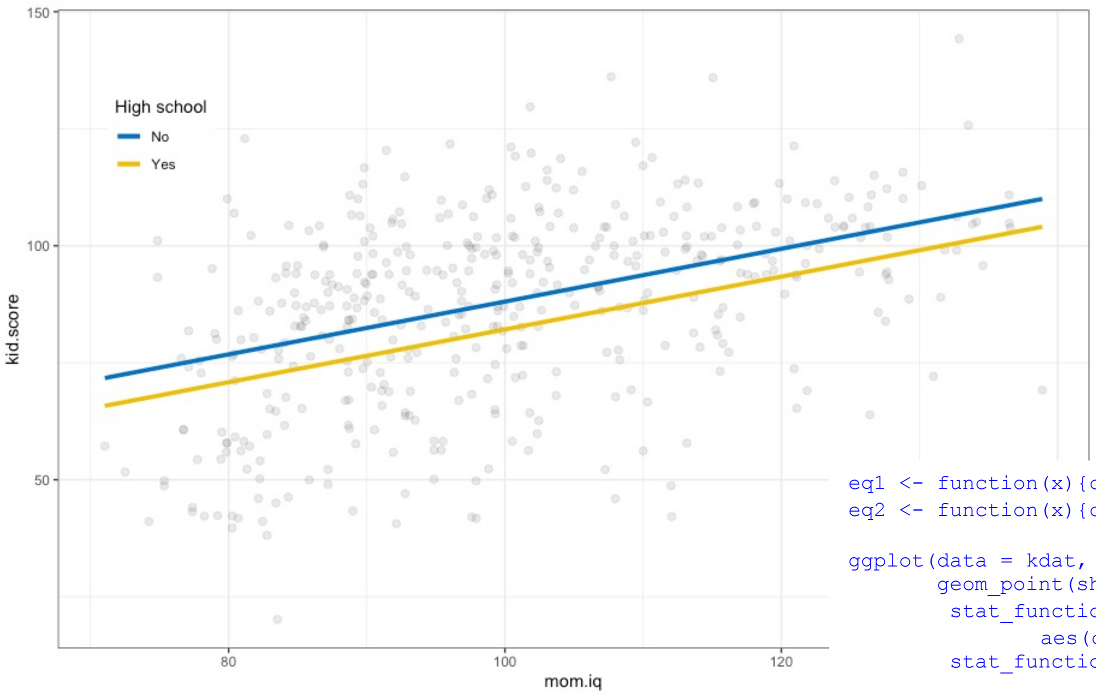
plotting a change in intercept

```
lm3 <- lm(kid.score ~ factor(mom.hs) + mom.iq)
 cf <- coef(lm3)

plot(x = mom.iq, y = kid.score, col = "plum")

 curve(cf[1] + cf[2] + cf[3]*x, 70, 140, add = T,
       col = "purple4", lwd = 3)

  curve(cf[1] + cf[2]*0 + cf[3]*x, 70, 140,
        add = T, col = "gold", lwd = 3)
```



I – kid cognition

plotting a change in intercept



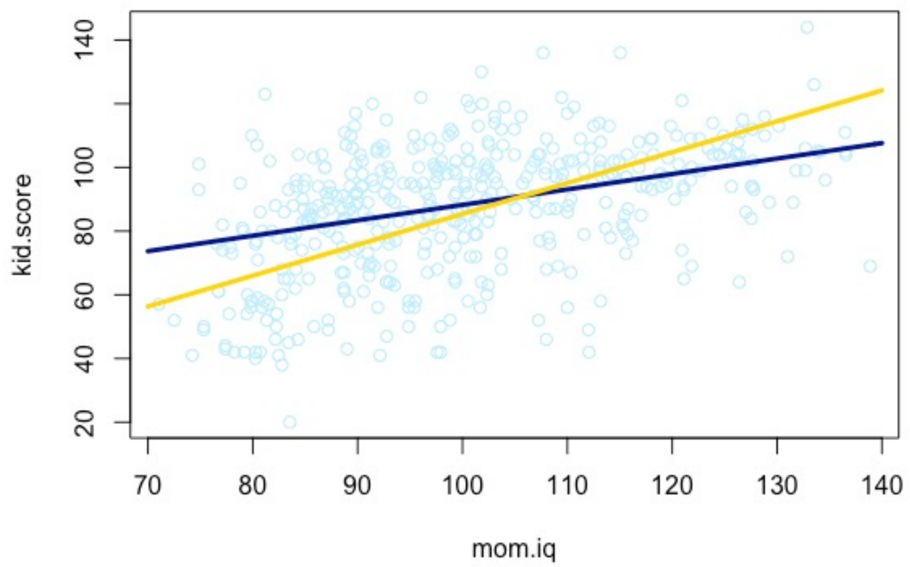

```
my_jco <- c("#0073C2FF", "#EFC000FF",
            "#868686FF", "#CD534CFF", "#7AA6DCFF",
            "#003C67FF", "#8F7700FF", "#3B3B3BFF",
            "#A73030FF", "#4A6990FF")
```

```
eq1 <- function(x){cf[1] + cf[2] + cf[3]*x}
eq2 <- function(x){cf[1] + cf[2]*0 + cf[3]*x}

ggplot(data = kdat, aes(x = mom.iq, y = kid.score)) +
       geom_point(shape = 19, size = 2, fill = my_jco[3], alpha = 0.1) +
        stat_function(fun = eq1, geom="line", linewidth = 1.3,
               aes(colour = "No")) +
        stat_function(fun = eq2, geom="line", linewidth = 1.3,
                          aes(colour = "Yes")) +
      theme_bw() + theme(legend.position = c(.1, .8)) +
      scale_colour_manual("High school", values = c(my_jco[1:4]))
```

1 – kid cognition

```
lm4 <- lm(kid.score ~ factor(mom.hs) * mom.iq)
```

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -11.4820    13.7580  -0.835 0.404422
factor(mom.hs)1          51.2682    15.3376   3.343 0.000902 ***
mom.iq                    0.9689     0.1483   6.531 1.84e-10 ***
factor(mom.hs)1:mom.iq   -0.4843     0.1622  -2.985 0.002994 **
```



teaser: plotting an interaction

```
cf <- coef(lm4)

plot(x = mom.iq, y = kid.score, col = "plum")

 curve(cf[1] + cf[2] + (cf[3] + cf[4])*x), 70, 140,
       add = T, col = "darkblue", lwd = 3)

   curve(cf[1] + cf[2]*0 + cf[3]*x, 70, 140,
       add = T, col = "gold", lwd = 3)
```

2 – model selection

What are the effects of average tree characteristics on plot-level aboveground biomass? Tree characteristics are basal area, number of trees, and number of recruits.

load the data

```
cong <- read.csv("cong.csv", header = T)
```

2 – model selection

Model the effects of basal area (BasalArea), number of trees (Trees), and number of recruits (Recruits) on tree biomass (AGB). Which model is best? How do you know?

load the data

```
cong <- read.csv("cong.csv", header = T)
```

run these models and select the best model

```
lm1 <- lm(AGB ~ BasalArea*Trees*Recruits, data = cong)
lm4 <- lm(AGB ~ BasalArea + Trees + Recruits + BasalArea:Trees, data = cong)
```

2 – model selection

Model the effects of basal area (BasalArea), number of trees (Trees), and number of recruits (Recruits) on tree biomass (AGB). Which model is best? How do you know?

load the data

```
cong <- read.csv("cong.csv", header = T)
```

reduce `lm1` to the minimum adequate model

```
lm1 <- lm(AGB ~ BasalArea*Trees*Recruits, data = cong)
```

2 – model selection

reduce `lm4` to the minimum adequate model

```
lm1 <- lm(AGB ~ BasalArea*Trees*Recruits, data = cong)
   lm2 <- update(lm1, .~.-BasalArea:Trees:Recruits)
    lm3 <- update(lm2, .~. -BasalArea:Recruits)
     lm4 <- update(lm3, .~. -Trees:Recruits)
      lm5 <- update(lm4, .~.-BasalArea:Trees)
       lm6 <- update(lm5, .~. -Recruits)
```

```
summary(lm6)
Call: lm(formula = AGB ~ BasalArea + Trees, data = cong)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.6499    45.7773  -0.473   0.6401
BasalArea    18.5675     1.3895  13.363 2.04e-13 ***
Trees        -0.3408     0.1263  -2.698   0.0119 *
---
Residual standard error: 36.15 on 27 degrees of freedom
Multiple R-squared:  0.8764,    Adjusted R-squared:  0.8672
F-statistic: 95.71 on 2 and 27 DF,  p-value: 5.535e-13
```

Questions?