model selection

- stepwise model selection
- p-value & adjusted R^2
- expert opinion

Duke UNIVERSITY

Dr. Mine Çetinkaya-Rundel
Duke University

stepwise model selection

- backwards elimination: start with a full model (containing all predictors), drop one predictor at a time until the parsimonious model is reached.
- forward selection: start with an empty model and add one predictor at a time until the parsimonious model is reached.
- criteria:
  - p-value, adjusted R^2
  - AIC, BIC, DIC, Bayes factor, Mallow's C_p (beyond the scope of this course)

backwards elimination - adjusted R^2

- Start with the full model
- Drop one variable at a time and record adjusted R^2 of each smaller model
- Pick the model with the highest increase in adjusted R^2
- Repeat until none of the models yield an increase in adjusted R^2

backwards elimination - adjusted R^2

step	variables included	removed	adjusted R
FULL	`kid_score ~ mom_hs + mom_iq + mom_work + mom_age`		**0.2098**
STEP 1	`kid_score ~ mom_iq + mom_work + mom_age`	`[-mom_hs]`	0.2027
	`kid_score ~ mom_hs + mom_work + mom_age`	`[-mom_iq]`	0.0541
	`kid_score ~ mom_hs + mom_iq + mom_age`	`[-mom_work]`	0.2095
	`kid_score ~ mom_hs + mom_iq + mom_work`	`[-mom_age]`	0.2109
STEP 2	`kid_score ~ mom_iq + mom_work`	`[-mom_hs]`	0.2024
	`kid_score ~ mom_hs + mom_work`	`[-mom_iq]`	0.0546
	`kid_score ~ mom_hs + mom_iq`	`[-mom_work]`	0.2105

backwards elimination - p-value

- Start with the full model
- Drop the variable with the highest p-value and refit a smaller model
- Repeat until all variables left in the model are significant

backwards elimination - p-value

FULL	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.5924	9.2191	2.13	0.0341
mom_hs:yes	5.0948	2.3145	2.20	0.0282
mom_iq	0.5615	0.0606	9.26	0.0000
mom_work:yes	2.5372	2.3507	1.08	0.2810
~~mom_age~~	~~0.2180~~	~~0.3307~~	~~0.66~~	~~0.5101~~

STEP 1	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.1794	6.0432	4.00	0.0001
mom_hs:yes	5.3823	2.2716	2.37	0.0183
mom_iq	0.5628	0.0606	9.29	0.0000
~~mom_work:yes~~	~~2.5664~~	~~2.3487~~	~~1.09~~	~~0.2751~~

STEP 2	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.7315	5.8752	4.38	0.0000
mom_hsyes	5.9501	2.2118	2.69	0.0074
mom_iq	0.5639	0.0606	9.31	0.0000

practice

The following model uses data from the American Community Survey to predict income from hours worked per week, race, and gender. Which variable (if any) should be dropped from the model first when doing backwards elimination using the p-value approach?

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2782.5726	6676.5534	0.42	0.6770	
hrs_work	1247.2128	146.2013	8.53	0.0000	✓
race:black	-9565.3090	6393.2168	-1.50	0.1350	
race:asian	35816.6156	8690.3484	4.12	0.0000	✓
race:other	-11112.8617	7213.3220	-1.54	0.1238	
gender:female	-16430.0916	3803.4700	-4.32	0.0000	✓

don't drop any variables

adjusted R^2 vs. p-value

- p-value: significant predictors
- adjusted R^2: more reliable predictions
- p-value method depends on the (somewhat arbitrary) 5% significance level cutoff
  - different significance level → different model
  - used commonly since it requires fitting fewer models (in the more commonly used backwards-selection approach)

forward selection - adjusted R^2

- Start with single predictor regressions of response vs. each explanatory variable
- Pick the model with the highest adjusted R^2
- Add the remaining variables one at a time to the existing model, and pick the model with the highest adjusted R^2
- Repeat until the addition of any of the remaining variables does not result in a higher adjusted R^2

forward selection - adjusted R^2

step	variables included	adjusted R
STEP 1	`kid_score ~ mom_hs`	0.0539
	`kid_score ~ mom_work`	0.0097
	`kid_score ~ mom_age`	0.0062
	`kid_score ~ mom_iq`	0.1991
STEP 2	`kid_score ~ mom_iq + mom_work`	0.2024
	`kid_score ~ mom_iq + mom_age`	0.1999
	`kid_score ~ mom_iq + mom_hs`	0.2105
STEP 3	`kid_score ~ mom_iq + mom_hs + mom_age`	0.2095
	`kid_score ~ mom_iq + mom_hs + mom_work`	0.2109
STEP 4	`kid_score ~ mom_hs + mom_iq + mom_work + mom_age`	0.2098

forward selection - p-value

- Start with single predictor regressions of response vs. each explanatory variable
- Pick the variable with the lowest significant p-value
- Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
- Repeat until any of the remaining variables do not have a significant p-value

expert opinion

- Variables can be included in (or eliminated from) the model based on expert opinion
- If you are studying a certain variable, you might choose to leave it in the model regardless of whether it's significant or yield a higher adjusted R^2

final model

R

```
> cog_final = lm(kid_score ~ mom_hs + mom_iq + mom_work, data = cognitive)
> summary(cog_final)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.17944    6.04319   4.001 7.42e-05 ***
mom_hsyes    5.38225    2.27156   2.369   0.0183 *
mom_iq       0.56278    0.06057   9.291  < 2e-16 ***
mom_workyes  2.56640    2.34871   1.093   0.2751

Residual standard error: 18.13 on 430 degrees of freedom
Multiple R-squared:  0.2163, Adjusted R-squared:  0.2109
F-statistic: 39.57 on 3 and 430 DF,  p-value: < 2.2e-16
```