Lecture 20 - Model Selection

Sta 102

June 14, 2016

Colin Rundel & Mine Çetinkaya-Rundel

Inference for MLR

Modeling children's test scores

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are a subsample from the National Longitudinal Survey of Youth.

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
⋮					
5	115	yes	92.75	yes	27
6	98	no	107.90	no	18
⋮					
434	70	yes	91.25	yes	25

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* (2007)

Cambridge University Press.

Model output

```r
summary(lm(kid_score ~ mom_hs + mom_iq + mom_work + mom_age, data = cognitive))
```

```
##
## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq + mom_work + mom_age,
##     data = cognitive)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.134 -12.624   2.293  11.250  50.206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>♦t♦)
## (Intercept) 20.82261    9.18765   2.266   0.0239 *
## mom_hs       5.56118    2.31345   2.404   0.0166 *
## mom_iq       0.56208    0.06077   9.249   <2e-16 ***
## mom_work     0.13373    0.76763   0.174   0.8618
## mom_age      0.21986    0.33231   0.662   0.5086
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.17 on 429 degrees of freedom
## Multiple R-squared:  0.215,eeIAdjusted R-squared:  0.2077
## F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```

Inference for the model as a whole

Is the model as a whole significant?

Inference for the model as a whole

Is the model as a whole significant?

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$
$$H_A : \text{At least one of the } \beta_i \neq 0$$

Inference for the model as a whole

Is the model as a whole significant?

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$
$$H_A : \text{At least one of the } \beta_i \neq 0$$

```
F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```

Inference for the model as a whole

Is the model as a whole significant?

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$
$$H_A : \text{At least one of the } \beta_i \neq 0$$

```
F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```

Since p-value < 0.05, the model as a whole is significant.

Inference for the model as a whole

Is the model as a whole significant?

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$
$$H_A : \text{At least one of the } \beta_i \neq 0$$

```
F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```

Since p-value < 0.05, the model as a whole is significant.

- The F test yielding a significant result doesn't mean the model fits the data well, it just means at least one or more of the βs is non-zero. i.e. the combination of these variables overall yields a model that is better than the intercept only model.

ANOVA Table

```
anova(lm(kid_score~.,data=cognitive))
```

```
## Analysis of Variance Table
##
## Response: kid_score
##            Df Sum Sq Mean Sq F value    Pr(>F)
## mom_hs      1  10125 10125.0 30.6763 5.325e-08 ***
## mom_iq      1  28504 28504.1 86.3608 < 2.2e-16 ***
## mom_work    1     18    17.6  0.0533    0.8175
## mom_age     1    144   144.5  0.4377    0.5086
## Residuals 429 141595   330.1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA Table

```
anova(lm(kid_score~.,data=cognitive))
```

```
## Analysis of Variance Table
##
## Response: kid_score
##            Df Sum Sq Mean Sq F value    Pr(>F)
## mom_hs      1  10125 10125.0 30.6763 5.325e-08 ***
## mom_iq      1  28504 28504.1 86.3608 < 2.2e-16 ***
## mom_work    1     18    17.6  0.0533    0.8175
## mom_age     1    144   144.5  0.4377    0.5086
## Residuals 429 141595   330.1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$MS_{Reg} = (18 + 144 + 10125 + 28504)/4 = 9697.75$$

$$F_{Reg} = 9697.75/330.1 = 29.38$$

ANOVA Table

```
anova(lm(kid_score~.,data=cognitive))
```

```
## Analysis of Variance Table
##
## Response: kid_score
##            Df Sum Sq Mean Sq F value    Pr(>F)
## mom_hs      1  10125 10125.0 30.6763 5.325e-08 ***
## mom_iq      1  28504 28504.1 86.3608 < 2.2e-16 ***
## mom_work    1     18    17.6  0.0533    0.8175
## mom_age     1    144   144.5  0.4377    0.5086
## Residuals 429 141595   330.1
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$MS_{Reg} = (18 + 144 + 10125 + 28504)/4 = 9697.75$$
$$F_{Reg} = 9697.75/330.1 = 29.38$$

```
F-statistic: 29.38 on 4 and 429 DF,  p-value: < 2.2e-16
```

Inference for the slope(s)

Is whether or not a mother graduated from high school a significant predictor of kid's cognitive test score, given all other variables in the model?

Inference for the slope(s)

Is whether or not a mother graduated from high school a significant predictor of kid's cognitive test score, given all other variables in the model?

$H_0: \beta_1 = 0$, when all other variables are included in the model

$H_A: \beta_1 \neq 0$, when all other variables are included in the model

Inference for the slope(s)

Is whether or not a mother graduated from high school a significant predictor of kid's cognitive test score, given all other variables in the model?

$H_0: \ \beta_1 = 0$, when all other variables are included in the model

$H_A: \ \beta_1 \neq 0$, when all other variables are included in the model

```
            Estimate Std. Error t value Pr(>♦t♦)
(Intercept) 19.59241    9.21906   2.125   0.0341
mom_hsyes    5.09482    2.31450   2.201   0.0282
mom_iq       0.56147    0.06064   9.259   <2e-16
mom_workyes  2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101

Residual standard error: 18.14 on 429 degrees of freedom
```

Inference for the slope(s)

Is whether or not a mother graduated from high school a significant predictor of kid's cognitive test score, given all other variables in the model?

$H_0: \ \beta_1 = 0$, when all other variables are included in the model

$H_A: \ \beta_1 \neq 0$, when all other variables are included in the model

	Estimate	Std. Error	t value	Pr(>♦t♦)
(Intercept)	19.59241	9.21906	2.125	0.0341
mom_hsyes	5.09482	2.31450	2.201	0.0282
mom_iq	0.56147	0.06064	9.259	<2e-16
mom_workyes	2.53718	2.35067	1.079	0.2810
mom_age	0.21802	0.33074	0.659	0.5101

```
Residual standard error: 18.14 on 429 degrees of freedom
```

$$T = 2.201, \ df = n - k - 1 = 434 - 4 - 1 = 429, \ \text{p-value} = 0.0282$$

Inference for the slope(s)

Is whether or not a mother graduated from high school a significant predictor of kid's cognitive test score, given all other variables in the model?

$H_0: \; \beta_1 = 0$, when all other variables are included in the model

$H_A: \; \beta_1 \neq 0$, when all other variables are included in the model

```
            Estimate Std. Error t value Pr(>♦t♦)
(Intercept) 19.59241    9.21906   2.125   0.0341
mom_hsyes    5.09482    2.31450   2.201   0.0282
mom_iq       0.56147    0.06064   9.259   <2e-16
mom_workyes  2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101

Residual standard error: 18.14 on 429 degrees of freedom
```

$$T = 2.201, \; df = n - k - 1 = 434 - 4 - 1 = 429, \; \text{p-value} = 0.0282$$

Since p-value < 0.05, whether or not mom went to high school is a significant predictor of kid's test score, given all other variables in the model.

Interpreting the slope

What is the correct interpretation of the slope for `mom_work`?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

Interpreting the slope

What is the correct interpretation of the slope for `mom_work`?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.59	9.22	2.13	0.03
mom_hs:yes	5.09	2.31	2.20	0.03
mom_iq	0.56	0.06	9.26	0.00
mom_work:yes	2.54	2.35	1.08	0.28
mom_age	0.22	0.33	0.66	0.51

All else being equal, children whose mothers worked during the first three years of the child's life are estimated to score 2.54 points higher than those whose mothers did not work.

Inference Recap from SLR

Inference for the slope for a SLR model (only one explanatory variable):

Inference Recap from SLR

Inference for the slope for a SLR model (only one explanatory variable):

- Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$

Inference Recap from SLR

Inference for the slope for a SLR model (only one explanatory variable):

- Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$

- Confidence interval:

$$b_1 \pm t^{\star}_{df} \times SE_{b_1}$$

Inference Recap from SLR

Inference for the slope for a SLR model (only one explanatory variable):

- Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$

- Confidence interval:

$$b_1 \pm t^{\star}_{df} \times SE_{b_1}$$

For MLR we now have multiple tests (one for each b_i), the sampling distribution is Normal / T but with *df* and *SE* calculated slightly differently.

In multiple linear regression b_i has a normal sampling distribution with

$$E(b_i) = \beta_i$$

$$SE^2_{b_i} = Var(b_i) = \frac{1}{1 - R_i^2} \frac{\sigma_e^2}{\sum_j (x_{ij - \bar{x}_i})^2}$$

where R_i^2 is the Multiple R^2 of the regression of X_i on all other predictors.

In multiple linear regression b_i has a normal sampling distribution with

$$E(b_i) = \beta_i$$

$$SE^2_{b_i} = Var(b_i) = \frac{1}{1 - R_i^2} \frac{\sigma_e^2}{\sum_j (x_{ij - \bar{x}_i})^2}$$

where R_i^2 is the Multiple R^2 of the regression of X_i on all other predictors.

Since we plug in s_e for σ_e we end up with a t distribution which has $df = n - p - 1$ where p is the number of slope parameters.

$$\frac{b_i - \beta_i}{SE_{b_i}} \sim T_{df=n-p-1}$$

CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$b_{mw} \quad \pm \quad t^{\star} SE_{b_{mw}}$$

CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$
\begin{aligned}
b_{mw} &\pm t^\star SE_{b_{mw}} \\
df &= n - k - 1 = 434 - 4 - 1 = 429 \to 400
\end{aligned}
$$

CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$
\begin{aligned}
b_{mw} &\pm t^\star SE_{b_{mw}} \\
df &= n - k - 1 = 434 - 4 - 1 = 429 \rightarrow 400 \\
2.54 &\pm 1.97 \times 2.35
\end{aligned}
$$

CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$
\begin{aligned}
b_{mw} &\pm t^{\star} SE_{b_{mw}} \\
df &= n - k - 1 = 434 - 4 - 1 = 429 \rightarrow 400 \\
2.54 &\pm 1.97 \times 2.35 \\
2.54 &\pm 4.63
\end{aligned}
$$

CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$
\begin{aligned}
b_{mw} &\pm t^{\star} SE_{b_{mw}} \\
df &= n - k - 1 = 434 - 4 - 1 = 429 \rightarrow 400 \\
2.54 &\pm 1.97 \times 2.35 \\
2.54 &\pm 4.63 \\
(-2.0895 &, 7.1695)
\end{aligned}
$$

CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$
\begin{aligned}
b_{mw} &\pm t^{\star} SE_{b_{mw}} \\
df &= n - k - 1 = 434 - 4 - 1 = 429 \rightarrow 400 \\
2.54 &\pm 1.97 \times 2.35 \\
2.54 &\pm 4.63 \\
(-2.0895 &, 7.1695)
\end{aligned}
$$

Interpretation?

CI for the slope

Construct a 95% confidence interval for the slope of `mom_work`.

$$
\begin{aligned}
b_{mw} &\pm t^{\star} SE_{b_{mw}} \\
df &= n - k - 1 = 434 - 4 - 1 = 429 \rightarrow 400 \\
2.54 &\pm 1.97 \times 2.35 \\
2.54 &\pm 4.63 \\
(-2.0895 &, 7.1695)
\end{aligned}
$$

Interpretation?

We are 95% confident that, all else being equal, children whose mothers worked during the first three years of the child's life are estimated to score between -2.0895 and 7.1695 points higher than those whose mothers did not work.

Inference for the slope(s) (cont.)

Given all variables in the model, which variables are significant predictors of kid's cognitive test score?

```
            Estimate Std. Error t value Pr(>♦t♦)
(Intercept) 19.59241    9.21906   2.125   0.0341
mom_hsyes    5.09482    2.31450   2.201   0.0282
mom_iq       0.56147    0.06064   9.259   <2e-16
mom_workyes  2.53718    2.35067   1.079   0.2810
mom_age      0.21802    0.33074   0.659   0.5101
```

Inference for the slope(s) (cont.)

Given all variables in the model, which variables are significant predictors of kid's cognitive test score?

	Estimate	Std. Error	t value	Pr(>♦t♦)
(Intercept)	19.59241	9.21906	2.125	0.0341
mom_hsyes	5.09482	2.31450	2.201	0.0282
mom_iq	0.56147	0.06064	9.259	<2e-16
mom_workyes	2.53718	2.35067	1.079	0.2810
mom_age	0.21802	0.33074	0.659	0.5101

`mom_hs` *and* `mom_iq` *are significant*

`mom_work` *and* `mom_age` *are not.*

Model selection

Modeling kid's test scores (revisited)

Predicting cognitive test scores of three- and four-year-old children using characteristics of their mothers. Data are from a survey of adult American women and their children - a subsample from the National Longitudinal Survey of Youth.

	kid_score	mom_hs	mom_iq	mom_work	mom_age
1	65	yes	121.12	yes	27
⋮	⋮	⋮	⋮	⋮	⋮
5	115	yes	92.75	yes	27
6	98	no	107.90	no	18
⋮	⋮	⋮	⋮	⋮	⋮
434	70	yes	91.25	yes	25

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* (2007) Cambridge University Press.

Model output

```
cog_full = lm(kid_score ~ mom_hs + mom_iq + mom_work + m
              data = cognitive)
summary(cog_full)
```

```
##              Estimate Std. Error t value Pr(>♦t♦)
## (Intercept) 19.59241    9.21906   2.125   0.0341
## mom_hsyes    5.09482    2.31450   2.201   0.0282
## mom_iq       0.56147    0.06064   9.259   <2e-16
## mom_workyes  2.53718    2.35067   1.079   0.2810
## mom_age      0.21802    0.33074   0.659   0.5101
##
## Residual standard error: 18.14 on 429 degrees of free
## Multiple R-squared: 0.2171, Adjusted R-squared: 0.209
## F-statistic: 29.74 on 4 and 429 DF,  p-value: < 2.2e-
```

Backward-elimination

Adjusted R^2 approach:

- Start with the full model
- Drop one variable at a time and record R^2_{adj} of each smaller model
- Pick the model with the largest increase in R^2_{adj}
- Repeat until none of the reduced models yield an increase in R^2_{adj}

Backward-elimination

Adjusted R^2 approach:

- Start with the full model
- Drop one variable at a time and record R^2_{adj} of each smaller model
- Pick the model with the largest increase in R^2_{adj}
- Repeat until none of the reduced models yield an increase in R^2_{adj}

When removing a categorical variable all levels should be included or removed *at the same time*

Backward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	*0.2098*

Backward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	*0.2098*
Step 1	kid_score ~ mom_iq + mom_work + mom_age	0.2027

Backward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	*0.2098*
Step 1	kid_score ~ mom_iq + mom_work + mom_age	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	0.0541

Backward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	*0.2098*
Step 1	kid_score ~ mom_iq + mom_work + mom_age	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	0.2095

Backward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	kid_score ˜ mom_hs + mom_iq + mom_work + mom_age	*0.2098*
Step 1	kid_score ˜ mom_iq + mom_work + mom_age	0.2027
	kid_score ˜ mom_hs + mom_work + mom_age	0.0541
	kid_score ˜ mom_hs + mom_iq + mom_age	0.2095
	kid_score ˜ mom_hs + mom_iq + mom_work	*0.2109*

Backward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	*0.2098*
Step 1	kid_score ~ mom_iq + mom_work + mom_age	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	0.2095
	kid_score ~ mom_hs + mom_iq + mom_work	*0.2109*
Step 2	kid_score ~ mom_iq + mom_work	0.2024

Backward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	kid_score ˜ mom_hs + mom_iq + mom_work + mom_age	*0.2098*
Step 1	kid_score ˜ mom_iq + mom_work + mom_age	0.2027
	kid_score ˜ mom_hs + mom_work + mom_age	0.0541
	kid_score ˜ mom_hs + mom_iq + mom_age	0.2095
	kid_score ˜ mom_hs + mom_iq + mom_work	*0.2109*
Step 2	kid_score ˜ mom_iq + mom_work	0.2024
	kid_score ˜ mom_hs + mom_work	0.0546

Backward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	*0.2098*
Step 1	kid_score ~ mom_iq + mom_work + mom_age	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	0.2095
	kid_score ~ mom_hs + mom_iq + mom_work	*0.2109*
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_hs + mom_work	0.0546
	kid_score ~ mom_hs + mom_iq	*0.2105*

Backward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	*0.2098*
Step 1	kid_score ~ mom_iq + mom_work + mom_age	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	0.2095
	kid_score ~ mom_hs + mom_iq + mom_work	*0.2109*
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_hs + mom_work	0.0546
	kid_score ~ mom_hs + mom_iq	*0.2105*
Step 3*	kid_score ~ mom_hs	0.2024

Backward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Full	kid_score ~ mom_hs + mom_iq + mom_work + mom_age	*0.2098*
Step 1	kid_score ~ mom_iq + mom_work + mom_age	0.2027
	kid_score ~ mom_hs + mom_work + mom_age	0.0541
	kid_score ~ mom_hs + mom_iq + mom_age	0.2095
	kid_score ~ mom_hs + mom_iq + mom_work	*0.2109*
Step 2	kid_score ~ mom_iq + mom_work	0.2024
	kid_score ~ mom_hs + mom_work	0.0546
	kid_score ~ mom_hs + mom_iq	*0.2105*
Step 3*	kid_score ~ mom_hs	0.2024
	kid_score ~ mom_iq	0.0546

Forward-selection

Adjusted R^2 approach:

- Start with regression of response vs. each explanatory variable
- Pick the model with the highest R^2_{adj}
- Add the remaining variables one at a time to the existing model, and once again pick the model with the highest R^2_{adj}
- Repeat until the addition of any of the remaining variables does not result in a higher R^2_{adj}

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Step 1	kid_score ~ mom_hs	0.0539

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Step 1	kid_score ~ mom_hs	0.0539
	kid_score ~ mom_work	0.0097

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
Step 1	kid_score ˜ mom_hs	0.0539
	kid_score ˜ mom_work	0.0097
	kid_score ˜ mom_age	0.0062

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
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*

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
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

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
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

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
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

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
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*

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
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*

Forward-selection: R^2_{adj} approach

Step	Variables included	R^2_{adj}
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_iq + mom_hs + mom_age + mom_work	0.2098

Final model choice

```
cog_final = lm(kid_score ~ mom_hs + mom_iq, data = kid)
summary(cog_final)
```

```
## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq, data = kid)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>♦t♦)
## (Intercept) 25.73154    5.87521   4.380 1.49e-05 ***
## mom_hsyes    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
```