Ordinary Least Squares Linear Regression Review

Lizzy Huang
Duke University

Week 4

Recall from Course 3 **Linear Regression and Modeling**, we discussed how to use R functions to obtain the point estimates of the coefficients in a simple/multiple linear regression minimizing the ordinary least squares. We call this type of linear regression model the Ordinary Least Squares (OLS) Linear Regression. Moreover, we also discussed hypothesis testing for coefficients using results from R, model selection using p-value and adjusted R^2.

However, in the Bayesian framework, we hope to obtain not only point estimates, but also the distributions of these regression coefficients, to account for uncertainty. Since using p-values has its own defect in hypothesis testing, we would like to use a more robust way to interpret the model, and to perform model selection. Before going into Week 4, let us go over some concepts that we have covered in Course 3. Then we will extend these ideas to the Bayesian methods.

Simple Linear Regression

We first start with the most simple case, the simple linear regression. Suppose we have a response variable Y, and we would like to explain this variable with only 1 explanatory variable X. Before seeing the data, these two variables are random variables. After seeing the data, $x_1, x_2, \cdots, x_n$, and $y_1, y_2, \cdots, y_n$ (or simply x and y), we woule like to see how a linear relationship could explain the relationship between these two. Therefore, we set up a simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \cdots, n.$$

Here, β_0 and β_1 are the unknown coefficients (or the parameters), ϵ_i is the unknown error. Remember, we can only perform estimates for these coefficients and errors, since we only have a sample of data. We will never be able to obtain the true values of these parameters for the entire population.

The reason that we call the model the OLS linear regression is because, the metric we use to judge whether the model is a good fit or not is the (ordinary) least squares or the sum of squares error (SSE)

$$\text{SSE} = \epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_n^2 = \sum_{i=1}^{n} \epsilon_i^2. \tag{1}$$

We obtain the point estimates of β_0 and β_1, denoted as $\hat{\beta}_0$ and $\hat{\beta}_1$, by minimizing this least squares (1). The shortcut formulas (and you may perform simple derivatives and look for the critical points of SSE) we provided are

$$\begin{aligned}\hat{\beta}_1 =& \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\ \hat{\beta}_0 =& \bar{y} - \hat{\beta}_1 \bar{x},\end{aligned}$$

where $\bar{x}$ and $\bar{y}$ are the sample mean of x and y.

Notations

We want to first talk about the notations we will be sticking with before going into more details.

Name	Notation
Explanatory variable and response variable	x, y
Observed data points	$x_1, x_2, \cdots, x_n,\ \ y_1, y_2, \cdots, y_n$
Coefficients (they are the parameters)	β_0, β_1
Point estimates of the coefficients (they are the statistics)	$\hat{\beta}_0$, $\hat{\beta}_1$
Predicted or fitted values from the model	$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \cdots n$
Errors	$\epsilon_1, \epsilon_2, \cdots, \epsilon_n$
Residuals from the model	$\hat{\epsilon}_i = y_i - \hat{y}_i$ (observed − predicted)

Moreover, we have other aggregation results and some shortcut equations known as the corrected sums of squares and crossproducts

$$\begin{aligned}
\text{sample mean of explanatory variable:} \quad & \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \\
\text{sample mean of response variable:} \quad & \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \\
\text{sum of squares in } x\text{:} \quad & \mathrm{S}_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
\text{sample variance of } x\text{:} \quad & \mathrm{s}_x^2 = \frac{1}{n-1}\mathrm{S}_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
\text{sum of squares in } y\text{:} \quad & \mathrm{S}_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
\text{sample variance of } y\text{:} \quad & \mathrm{s}_y^2 = \frac{1}{n-1}\mathrm{S}_{yy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\
\text{crossproduct between } X \text{ and } Y\text{:} \quad & \mathrm{S}_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})
\end{aligned}$$

With this notation, we can rewrite the shortcut formulas for the point estimate $\hat{\beta}_1$ to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{S}_{xy}}{\mathrm{S}_{xx}}.$$

We will use these notations in the upcoming Week 4 lectures.

Conditions

It is very important to keep in mind of the conditions or assumptions before we use any linear regression models

Linearity

Relationship between the explanatory variable and the response variable should be linear. We can check this using a scatterplot between x and y, or the *residual plot*.

Nearly Normal Residuals

Residuals should be nearly normally distributed, centered at 0. We can check this using a histogram (check out `hist` in R) or a *normal probability plot* (check out `qqnorm` in R) of the residuals.

Constant variability (homoscedasticity)

Variability of points around the ordinary least squares line should be roughly constant, which implies the variability of residuals around the 0 line should be roughly constant. We can check this using the *residual plot*.

The 2nd and the 3rd conditions imply that

$$\epsilon_i \sim \mathsf{N}(0, \sigma^2), \qquad i = 1, \cdots, n,$$

with the same variance σ^2.

Variability Partitioning

We can partition the total variability in y into the variability in y explained by x, and the variability in y not explained by x. Consider the following sums of squares:

$$\begin{aligned}
\text{total variability in } y: &\qquad \text{SST (total sum of squares)} = \mathrm{S}_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
\text{unexplained variability in } y: &\qquad \text{SSE (error sum of squares)} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{\epsilon}_i^2 \\
\text{explained variability in } y: &\qquad \text{model/regression sum of squares} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2
\end{aligned}$$

We have the relationship

$$\text{SST} = \mathrm{S}_{yy} = \text{SSE} + \text{SSR}.$$

Moreover, we can calculate the mean squared error:

$$\text{mean squared error} = \text{MSE} = \frac{1}{n-2}\text{SSE} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \hat{\sigma}^2.$$

Since we use the mean squared error (MSE) to estimate the error variance (which is σ^2), we also denote MSE to be $\hat{\sigma}^2$.

R^2 (R-Squared)

With the partitioning of the variability, we can come back to discuss R^2, a metric that we are very familiar with when conducting linear regression. R^2 is the ratio between the explained variability in y and the total variability in y. That is

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

R^2 is always a number between 0 and 1. Consider a very extreme case when we use a constant to fit the response y:

$$y_i = \beta_0 + \epsilon_i, \qquad i = 1, \cdots, n.$$

We can view this as a special case of the simple linear regression when $\hat{\beta}_1 = 0$, which leads to the point estimate of β_0 to be

$$\hat{\beta}_0 = \bar{y}.$$

That means, under this model, we are trying to use a horizontal line $y = \bar{y}$ to fit all the observed responses $y_1, y_2, \cdots, y_n$. So the predicted or fitted value in this case will be $\hat{y}_i = \bar{y},\ i = 1, \cdots, n$. Then the explained variability in y will be

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (\bar{y} - \bar{y})^2 = 0.$$

Therefore, in this extreme case, $R^2 = 0$. This means leaving only the constant in the model does not really explain the variability in y. **When adding more explanatory variables, R^2 increases.**

Correlation

The correlation of two variables describe the strenght of the **linear association** between two variables. The correlation between x and y, is calculated by

$$\text{corr}(x, y) = \frac{\text{S}_{xy}}{\text{S}_x \text{S}_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}.$$

Geometrically speaking, correlation calculates the cosine of the "angle" between the two variables x and y under some nice metric, if we view $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$ as vectors. When there is only 1 explanatory variable, that is, **only under the simple linear regression**, $R^2 = (\text{corr}(x, y))^2$.

We need to be cautious when using correlation:

- When the correlation between two variables is small, it does not imply there is no association between the two variables, because correlation only measure the linear association.
- When the two variables are independent, their correlation must be 0. The converse is not true, that is, if the correlation between two variables is 0, it does not imply the two variables are independent.

Inference on Regression Slope β_1

We would like to ask, is the explanatory variable x a significant predictor of the response variable y? This is equivalent to ask, is the parameter β_1 not equal to 0?

We can set up a hypothesis test for the regression slope β_1:

$$H_0 : \beta_1 = 0, \qquad H_A : \beta_1 \neq 0.$$

We then calculate the t-score (t-statistics)

$$t = \frac{\text{point estimate - null value}}{\text{standard error}} = \frac{\hat{\beta}_1 - 0}{\text{se}_{\hat{\beta}_1}},$$

and find the p-value in the t-table with degress of freedom $n - 2$.

This can be done through R, when we run the `lm` function. We just want to provide the formula for the standard error of $\hat{\beta}_1$, which we will use in the upcoming video.

$$\text{se}_{\hat{\beta}_1} = \sqrt{\frac{\text{MSE}}{\text{S}_{xx}}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Example 1

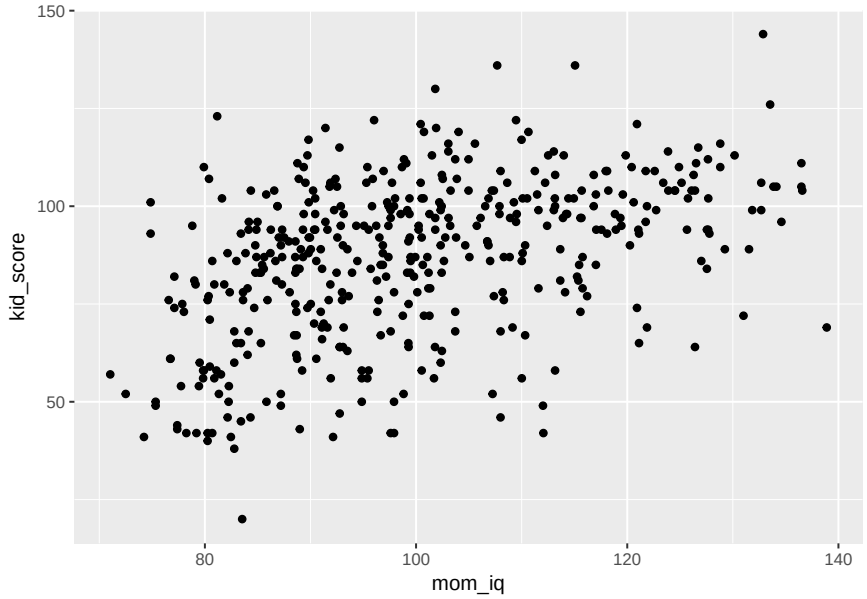
We demonstrate a simple OLS linear regression using R. The data set is the `cognitive` data set we used in Course 3 Week 3 Video **Inference for MLR** and we will be using in this upcoming week's lectures. We first read in the data set

```
cognitive = read.csv("http://bit.ly/dasi_cognitive")
colnames(cognitive)
```

```
## [1] "kid_score" "mom_hs"    "mom_iq"    "mom_work"  "mom_age"
```

There are 1 response variable `kid_score`, and 4 explanatory variables: `mom_hs` (mother's high school status), `mom_iq` (mother's IQ score), `mom_work` (whether the mother works in the first 3 years of the kid's life), and `mom_age` (mother's age).

We first predict `kid_score` using `mom_iq`. The following scatterplot shows that the two variables roughly have a positive linear relationship.



The two variables have a correlation of about 0.448.

We then run the simple linear regression using the `lm` function in R.

```
cognitive.slr = lm(kid_score ~ mom_iq, data = cognitive)
summary(cognitive.slr)
```

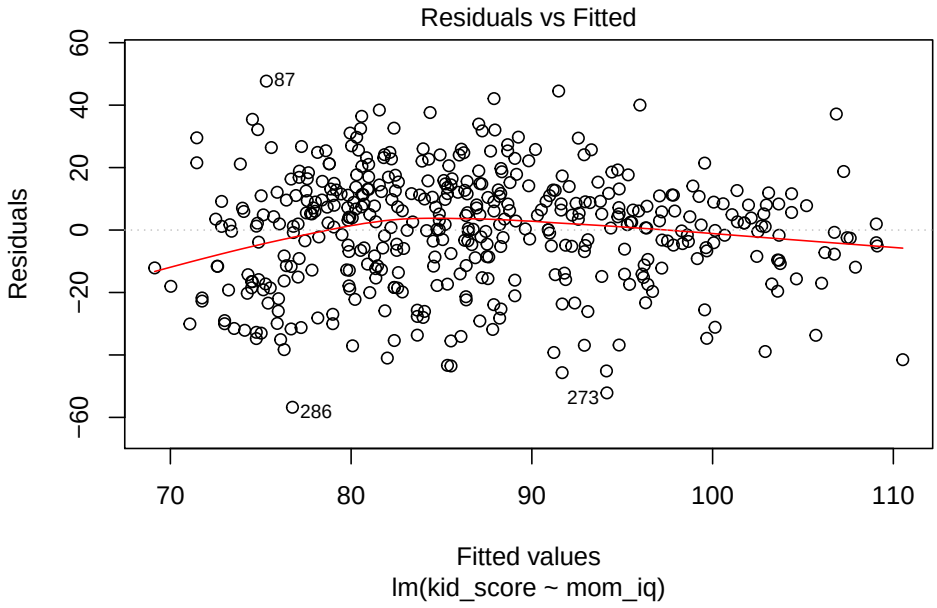
```
##
## Call:
## lm(formula = kid_score ~ mom_iq, data = cognitive)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -56.753 -12.074   2.217  11.710  47.691
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.79978    5.91741    4.36 1.63e-05 ***
## mom_iq       0.60997    0.05852   10.42  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.27 on 432 degrees of freedom
## Multiple R-squared:  0.201,  Adjusted R-squared:  0.1991
## F-statistic: 108.6 on 1 and 432 DF,  p-value: < 2.2e-16
```

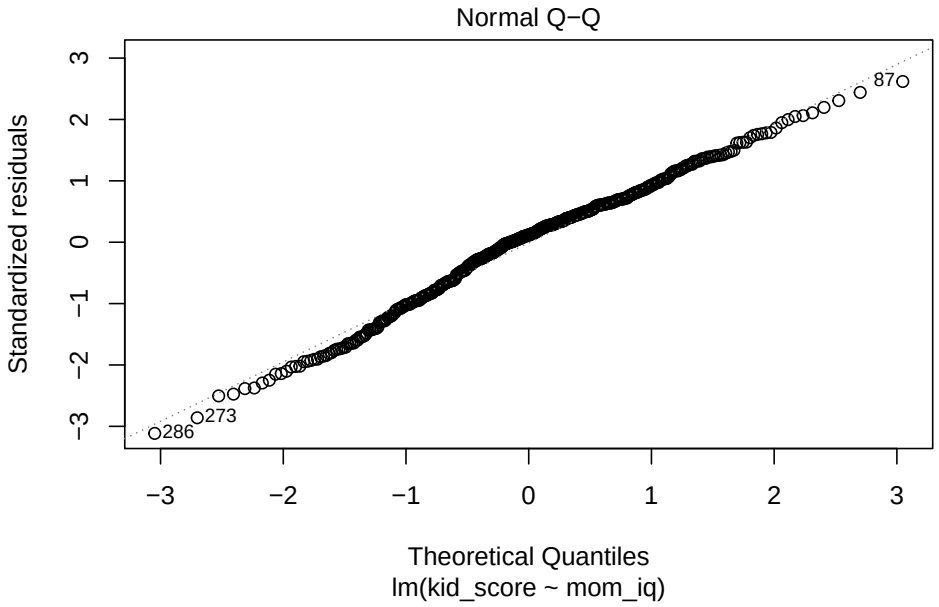
The summary of the results shows that the R^2 is about 0.201, and the t-score for the significance of the variable `mom_iq` is about 10.423, which leads to a very small p-value.

The *residual plot* and *normal probability plot* of the residuals can be obtained by

```
# residual plot
plot(cognitive.slr, which = 1)
```




```
# normal probability plot of residuals
plot(cognitive.slr, which = 2)
```



The residuals plot shows that the residuals are centered at 0 and have a rough constant variance. The normal probability plot shows that the residuals are nearly normal, with some outliers identified on both plots. We will improve the regression model by adding more explanatory variables.

Multiple Linear Regression

We can add more explanatory into the linear regression, which becomes a multiple linear regression.

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_p x_{p,i} + \epsilon_i, \qquad i = 1, 2, \ldots, n.$$

Here we have p predictor variables $x_1, x_2, \ldots, x_p$, with $p+1$ coefficients $\beta_0, \beta_1, \ldots, \beta_p$. We sometimes may denote all the coefficients using a vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$.[1]

Collinearity and Parsimony

When we are adding more and more explanatory variables, we would like to ask whether these additional explanatory variables really give us new information in the response variable, i.e., whether the additional explanatory variables explain more variability in y necessarily.

We call the two predictors are collinear if they are correlated with each other. When we add more explanatory variables, we expect these variables are independent variables, which means, they should not be collinear. Inclusion of collinear predictors (which also called multicollinearity) complicates the model so the estimates coming out of the model may not be reliable.

We want to avoid adding predictors that are associate with each other because the addition of such variable often will not bring any new information about the response, and may even dilute the original information that we have (the estimates of the coefficients may be biased). We prefer the simplest best model, that is, the parsimonious model. This model has the highest predictive power, with the lowest number of predictors.

[1] The superscript T means we will transpose the vector so that it becomes a $(p+1) \times 1$column vector.

Model Selection

To perform model selection, meaning, to select the explanatory variables so that the model will become parsimonious, we can perform **backword elimination** or **forward selection**.

Adjusted R^2

Since R^2 will keep increasing when the number of explanatory variables increases, we are not going to use R^2 to be the measure of parsimony. Instead, we introduced the adjusted R^2

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-p-1}\frac{\text{SSE}}{SST},$$

where p is the number of predictors (excluding the constant term). We select the model with the highest adjusted R^2.

Since $n-p-1 \leq n-1$, $\dfrac{n-1}{n-p-1} \geq 1$. Therefore, R^2_{adj} is always smaller than R^2, unless $p = 0$.

The factor $\dfrac{n-1}{n-p-1}$ serves as a penalty, that if we are adding too many predictors, leading too large p, the adjusted R^2 will drop instead.

AIC (new)

Another criterion we may use is the Akaike Information Criterion (AIC). It is defined to be

$$\text{AIC} = n\ln(1-R^2) + 2(p+1) + \text{some constant},$$

where p is the number of predictors. Here "some constant" only depends on the number of observations, so we usually ignore it and define

$$\text{AIC} = n\ln(1-R^2) + 2(p+1).$$

We select model with the least AIC. When adding more explanatory variables, we increase R^2 and decease $\ln(1-R^2)$. However, we may also increase $p+1$. Therefore, the second term in the AIC definition serves as the penalty term.

We also introduced using p-value as one of the selection criteria. In Bayesian statistics, we will focus on using BIC (Bayesian Information Criterion) and Bayes factors for model selection.

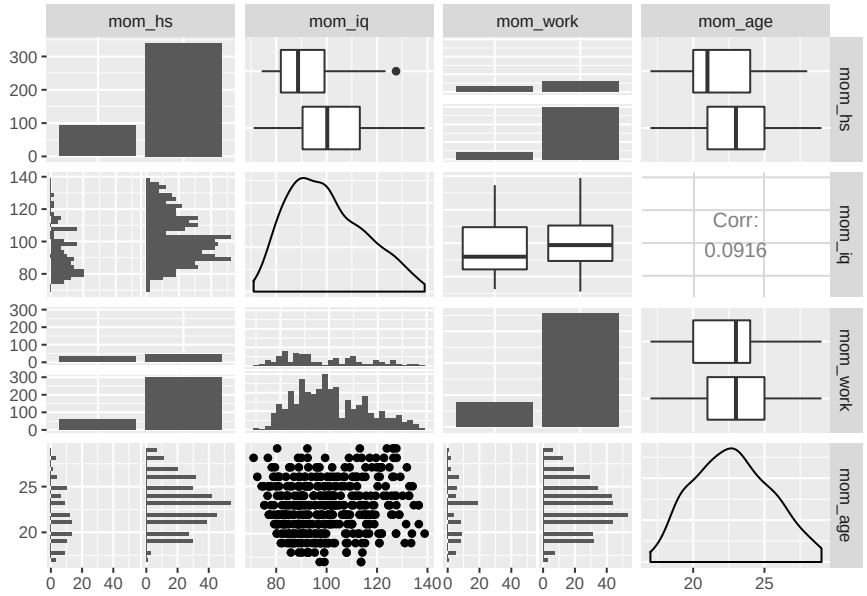
Example 2

The backward elimination and forward selection using adjusted R^2 and p-value were demonstrated in Course 3 Week 3 Video **Model Selection**. So we are not going to repeat these methods. Here, we would like to show how to perform model selection using AIC. R has a very simple function `stepAIC` in the `MASS` package to finish this process.

We still use the kid's `cognitive` score data set as our example. We performed a simple linear regression using just `mom_iq` variable. The summary shows that the R^2 is about 0.201 and adjusted R^2 is about 0.199. From the residual plot, we see that the residuals are not really randomly scattered around the 0 line, so there may be some improvement in the model.

We will use a backward elimination using AIC. Before that, we first examine the collinearity between each variable.

```
library(GGally)
ggpairs(cognitive[2:5]) # only the explanatory variables
```



The ggpairs function generate correlation if the two variables are both numeric, as well as other summary statistics plots if one of them is categorical.

It seems there is not any patterns between any of the two variables. Then we start with the full model and call stepAIC to select variables.

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

# AIC
library(MASS)
stepAIC(cognitive.mlr)
```

```
## Start:  AIC=2520.71
## kid_score ~ mom_hs + mom_iq + mom_work + mom_age
##
##            Df Sum of Sq    RSS    AIC
## - mom_age   1     143.0 141365 2519.2
## - mom_work  1     383.5 141605 2519.9
## <none>                  141222 2520.7
## - mom_hs    1    1595.1 142817 2523.6
## - mom_iq    1   28219.9 169441 2597.8
##
## Step:  AIC=2519.15
## kid_score ~ mom_hs + mom_iq + mom_work
##
##            Df Sum of Sq    RSS    AIC
## - mom_work  1     392.5 141757 2518.3
## <none>                  141365 2519.2
## - mom_hs    1    1845.7 143210 2522.8
## - mom_iq    1   28381.9 169747 2596.6
##
## Step:  AIC=2518.35
```

```
## kid_score ~ mom_hs + mom_iq
##
##          Df Sum of Sq    RSS    AIC
## <none>                141757 2518.3
## - mom_hs  1    2380.2 144137 2523.6
## - mom_iq  1   28504.1 170261 2595.9

##
## Call:
## lm(formula = kid_score ~ mom_hs + mom_iq, data = cognitive)
##
## Coefficients:
## (Intercept)    mom_hsyes       mom_iq
##     25.7315       5.9501       0.5639
```

From the stepwise selection, we see that AIC first eliminates `mom_age`, which will result a smaller AIC (from the full model 2520.71 to 2519.15). Then AIC eliminates `mom_work`, which lower the AIC to 2518.35. Finally, `stepAIC` provides the final model to be

$$\text{kid\_score} \sim \text{mom\_hs} + \text{mom\_iq}.$$

AIC selects a different model than the one selected by adjusted R^2 and p-value in the **Model Selection** video. It is very common that we have different models under different selection criterion. To make the final decision, we may utilize decision-making theory by calculating the associated loss functions, or follow expert opinion.