Multiple Linear Regression

STA 198: Introduction to Health Data Science

Yue Jiang

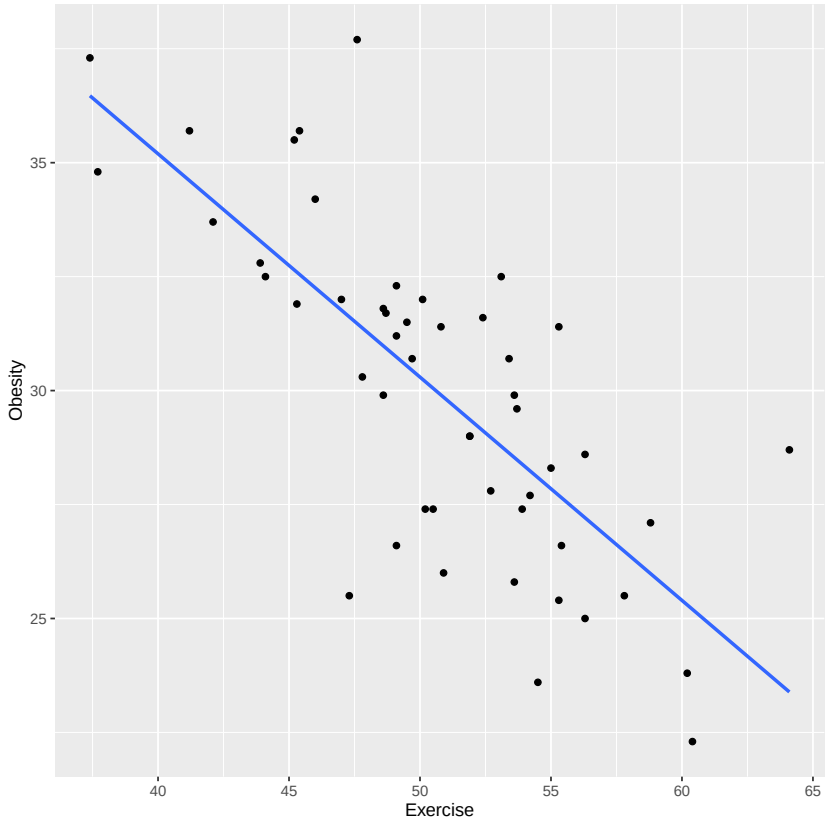
July 13, 2020

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

Back to Lab 01...

Last time we examined the regression of obesity percentage on adequate exercise percentage.



Do we think this is the only factor that can help us predict obesity percentage?

Multiple regression

The multiple regression model extends the simple linear regression model by incorporating more than one explanatory variable. The assumptions are similar to those of the simple linear regression model. This type of model is often called a multivariable (**not multivariate**) model.

A multiple regression model is often used to control for confounders or predictors that explain important variability in the response. Example:

- Knowing that a state has above average adequate exercise percentage might tell you something about the obesity percentage
- However, if you also knew the state's smoking rate, you might be able to predict obesity percentage more accurately
- If you also know that state's HDI category, you might be able to do better still!

Multiple regression

The model is given by $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i$, where

- p is the total number of predictor or explanatory variables
- y_i is the outcome (dependent variable) of interest
- β_0 is the intercept parameter
- $\beta_1, \beta_2, \cdots, \beta_p$ are the slope parameters
- $x_{1i}, x_{2i}, \cdots, x_{pi}$ are predictor variables
- ϵ_i is the error (like the βs, it is not observed)
- Assumptions are essentially the same as in simple linear regression
- **Interpretations are conditional on other covariates in model** (more next)

Multiple regression

Consider the model $Obesity_i = \beta_0 + \beta_1 Exercise_i + \beta_2 Smoking_i + \epsilon_i$. The parameter interpretations are below.

- β_0 represents the expected obesity percentage for a state with a value of 0 for all other predictors (i.e., adequate exercise and smoking % of 0)
- β_1 represents the expected increase in obesity percentage for a 1 percentage point increase in adequate exercise, *holding all other variables constant*
- β_2 represents the expected increase in obesity percentage for a 1 percentage point increase in smoking, *holding all other variables constant*

Multiple regression

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.10015    3.86311   9.345 2.72e-12 ***
Exercise    -0.30899    0.05579  -5.538 1.34e-06 ***
Smoking      0.54249    0.08716   6.224 1.23e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.837 on 47 degrees of freedom
Multiple R-squared:  0.7546,    Adjusted R-squared:  0.7442
F-statistic: 72.28 on 2 and 47 DF,  p-value: 4.572e-15
```

Multiple regression

$$Obesity_i = \beta_0 + \beta_1 Exercise_i + \beta_2 Smoking_i + \epsilon_i.$$

- $\hat{\beta}_0 = 36.1$
- $\hat{\beta}_1 = -0.31$
- $\hat{\beta}_2 = 0.54$

How might we interpret these estimates?

Hypotheses of interest

Hypotheses of interest may include the following:

- $H_0 : \beta_1 = \beta_2 = 0$ vs. H_1 : at least one of β_1 or β_2 is not 0. This tests whether any of the predictors have any association together with the outcome and is called an overall or group test.
- $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. This tests whether exercise % is associated with obesity %, controlling for smoking %.
- $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$. This tests whether smoking % is associated with obesity %, controlling for exercise %.

Overall F test

The *overall F test* tests the hypothesis
$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (all non-intercept parameters are 0).
This is a test of whether any of our predictors are related to the response. This is an F test like those we used in ANOVA and has numerator df equal to the number of *predictors* being tested (p), and denominator df equal to the total sample size minus the number of mean parameters in the model ($n - p - 1$ because the intercept is also a parameter).

In the CDC data, our p-value for the overall F test is < 0.0001 and we conclude that at least one predictor is statistically significantly related to obesity percentage.

Particular predictors

We can also test parameters individually, as we did in simple linear regression. For instance, the p-value corresponding to $\hat{\beta}_1$ was significant at the $\alpha = 0.05$ level in this dataset. Thus, we reject H_0 and conclude that exercise % is related to obesity % after controlling for smoking %.

We can also use the standard error to construct confidence intervals. For instance, $SE_1 = 0.06$. Thus, **at the same smoking %**, a state with one percentage point exercise % higher than another would be expected to have an obesity % that is -0.31 (-0.37, -0.25) greater (i.e., 0.31 percentage points less) than another.

R^2

In our model, $R^2 = 0.7546$, suggesting that about 75% of the variability in obesity percentage can be explained by our model.

However, unadjusted R^2 can never decrease when variables are added to a model, even if they are useless.

Thus, we can use *adjusted* $R^2 \leq R^2$, where the adjustment is made to account for the number of predictors.

The adjusted R^2 incorporates a penalty for each additional variable in a model, so that the adjusted R^2 will go down if a new variable does not improve prediction much, and it will go up if the new variable does improve prediction, conditional on the other variables already in the model.

Interaction effects

Sometimes, the effect of one variable depends on the value of another. For example, the effect of exercise % on obesity may be different for smokers vs. non-smokers. To model such a relationship (often called effect modification because one variable modifies the effect of another), we create an interaction term.

This is created simply by multiplying two predictors x_1 and x_2 to create a new predictor, $x_1 x_2$. When interaction terms are in a model, interpretations can become tricky.

Interaction

We can see whether the effect of exercise on obesity is modified by smoking as follows. First we create an interaction term by multiplying the predictors. The model is then

$$Obesity_i = \beta_0 + \beta_1 Exercise_i + \beta_2 Smoking_i + \beta_3 Exercise_i Smoking_i + \epsilon_i$$

Interaction

- $\hat{\beta}_0 = 38.6$
- $\hat{\beta}_1 = -0.36$
- $\hat{\beta}_2 = 0.41$
- $\hat{\beta}_3 = 0.03$

How do we interpret these estimates?

Interaction

How do we interpret the estimates? For a given exercise (ex) and smoking (sm) percentage, our predicted obesity (ob) percentage is

$$\widehat{ob}_i = \hat{\beta}_0 + \hat{\beta}_1 ex_i + \hat{\beta}_2 sm_i + \hat{\beta}_3 ex_i sm_i$$

and for a state at the same smoking % but 1 percentage point higher in exercise %, the predicted obesity percentage is

$$\begin{aligned}\widehat{ob}_{i'} &= \hat{\beta}_0 + \hat{\beta}_1(ex_i + 1) + \hat{\beta}_2(sm_i) + \hat{\beta}_3(ex_i + 1)sm_i \\ &= \hat{\beta}_0 + \hat{\beta}_i ex_i + \hat{\beta}_1 + \hat{\beta}_2 sm_i + \hat{\beta}_3 ex_i sm_i + \hat{\beta}_3 sm_i.\end{aligned}$$

Subtracting, we have $\widehat{ob}_{i'} - \widehat{ob}_i = \hat{\beta}_1 + \hat{\beta}_3 sm_i$, which is the expected change in obesity % for a 1% change in exercise %.

The effect of exercise % depends on the level of smoking % in that state.

Collinearity

One common problem in multiple regression is *collinearity*, which occurs when multiple highly correlated variables are used as predictors. In this case, the model can become unstable (often seen as standard errors that get huge and lead to huge confidence interval estimates), and it can be hard to assess the impact of the predictors.

Diagnosing collinearity

If nothing is predictive, we have some clues:

1. Individual predictors are significant in simple linear regression models,
2. but standard errors and interval estimates are huge,
3. and the overall F test is significant

A significant overall F test with no significant individual variable test is a typical sign of collinearity. We can check out the correlations among the three predictors.

Collinearity

There is no fixed criterion for correlation to exclude a variable for collinearity. It is possible to construct examples where the correlation is very high, but collinearity is not a problem because the information about the outcome in the two variables is different.

Dummy variables

In regression settings, we can account for categorical variables by creating dummy variables, which are indicator variables for certain conditions happening. For instance, there are three categories of HDI in the dataset: bottom ten, middle, and top ten.

When considering categorical variables, one variable is taken to be the baseline or reference value. All other categories will be compared to it.

Dummy variables

Suppose the "bottom ten" category is taken to be the referent value. Then we can create two dummy variables:

- HDI == Middle: 1 if this condition is true; 0 otherwise
- HDI == Top Ten: 1 if this condition is true; 0 otherwise

Dummy variable interpretation

Consider the model
$Obesity_i = \beta_0 + \beta_1(HDI == Middle)_i + \beta_2(HDI == TopTen)_i + \epsilon_i$.
The parameter interpretations are below.

- β_0 represents the expected obesity percentage for a state with 0 for the two dummy variables. That is, in the bottom ten HDI
- β_1 represents the expected difference in obesity percentage for a state in the middle HDI category, compared to the bottom ten
- β_2 represents the expected difference in obesity percentage for a state in the top ten HDI category, compared to the bottom ten

Dummy variable interpretation

Let's consider a different referent category for the following model:

$$Obesity_i = \beta_0 + \beta_1(HDI == Middle)_i + \beta_2(HDI == BottomTen)_i + \epsilon_i$$

How might you interpret the following estimates?

- $\hat{\beta}_0 = 34.4$
- $\hat{\beta}_1 = -4.8$
- $\hat{\beta}_2 = -8.1$

Interactions with categorical predictors

Luckily, interpretation of interaction terms with categorical predictors is easier than with continuous predictors. Since categorical predictors are based on dummy variables, they can only take on the values of 0 or 1 in the model. Again, an interaction effect implies that the regression coefficient for an explanatory variable would change depending on the value of another predictor.