Lecture 4.2 - Model Building

Professor MacDonald

2025-04-29

Table of contents

Practicing model building - Average test score in schools

Provided is a dataset to continue building your skills in model building. The dataset includes many variables related to the quality of schools in California. Please use `testscr` as your response variable.

The definition of all of the variables can be found here.

Learning objectives

- Continue building skills in interpreting variables
- Practice model building, this time by adding p values to the process
- Interpret regression results, including p values

Hypothesis development

First, we want to examine the basic relationships in our data.

- Make a list of hypothesized relationships to `testscr`. For key variables, list what you expect its relationship to `testscr`.
- Select an alpha value for your hypotheses and provide a justification.
- Also note, based on the summary information of the key predictor variables, generally speaking, how large of an impact do you expect the predictor variable to have.

Variable Exploration

For this part of the activity, you should explore the distribution of all of the relevant variables via histograms.

To combine multiple plots into one image, you can use the `grid.arrange` function as shown in the sample code below:

```
library(gridExtra)

p1 <- ggplot(schools, aes(x=testscr))+geom_histogram()+ggtitle("Test scores")+ylab("Average t
p2 <- ggplot(schools, aes(x=mealpct))+geom_histogram()+ggtitle("Percent of students given fre
p3 <- ggplot(schools, aes(x=enrltot))+geom_histogram()+ggtitle("Enrollment total")+ylab("Coun

grid.arrange(p1, p2, p3)
```

- Make note of any outliers or non-normal distributions that may cause problems for your later statistical test.
- Also consider if any variables need to be recoded or transformed. For variables that are normally distributed, you don't need to make a lot of remarks – you only need to note the unusual cases or the ones in which a transformation is necessary.

Two-Way Data Relationships

Based on your hypothesized relationship between each of the predictor variables and `testscr`, check the two-way relationships to see your expectations are met or not. You may also want to check the correlations between all variables. As a reminder, one way to create a correlation matrix plot using the `ggcorrplot` library is:

```
schools.subset <- schools %>%
  select(c(testscr, mealpct, enrltot))

testcors <- cor(schools.subset, use="complete.obs")
ggcorrplot(testcors)
```

More details can be found here

- Develop some plots and/or tables to summarize your two-way relationships – again, you only need to remark on the unusual cases or where the relationship looks strong and you think it warrants further investigation.
- Did any of the two-way relationships surprise you? Which ones and why?

Model Building

Now, armed with your hypotheses, enter in the variables you wish to test and any other variables you think relevant and work with your partner to create the best regression model possible.

- First, check if the regression requirements (linear relationship, independence, does the plot thicken?, nearly normal residuals)
- If they violate the conditions, make sure to adjust them
- Are any of the variables collinear? How do you think that will change the relationship between the predictor and response variables? Draw a causal diagram of how you think the causation will work in your model.

Next, check to make sure the model appears to be a reasonable model overall by viewing the residual plots and paying attention to the model diagnostics.

- Create and interpret residual plots (see code below)
- Carefully interpret your coefficients
- Examine your R^2 and S_e
- View and interpret the partial residual plots
- If you notice any problems, correct the variable choice or variable expression

As a reminder, to generate residuals, you can do so with the `augment()` function from the `broom` library:

```
fit <- lm(data=ca.school, testscr ~ mealpct)
aug.fit <- augment(fit)
```

You can then graph them with the following commands:

```
ggplot(lmodel.data, aes(x=.fitted, y=.resid)) +
  geom_point() +
  geom_hline(yintercept=0)
```

As for partial residual plots, remember the way to plot partial residuals is:

```
library(visreg)

eff <- visreg(lmodel, "mealpct", gg=TRUE)
eff+ggtitle("Partial Residual Plot")
```

Statistical significance of the predictors

As discussed in lecture, the null hypothesis for a regression is that $\beta_1 = 0$; that is, the slope of the predictor variable is equal to zero. The alternative hypothesis is that $\beta_1 \neq 0$. If the slope of the regression coefficient is 0, that means that there is no relationship between the predictor variable and the response variable.

- Using this knowledge, interpret the statistical significance of your predictor variables
- Next, interpret each predictor variables' practical significance. You can do this with the Q1/Q3 method discussed in previous lectures
- Finally, evaluate your hypotheses with respect to the outcome of your model. Were you surprised by any of the results? Why or why not?