MA 151-01: Multiple Linear Regression

7 December 2017

Session Materials: `http://bit.ly/ma151-mlr`

Learning Objectives

By the end of class today, you will be able to:

1. Predict an outcome using a given multiple linear regression.
2. Interpret the coefficients of a multiple linear regression as a data summary device.
3. Fit a multiple linear regression to data using Minitab and interpret Minitab's output.
4. Perform simple model selection using backward elimination.

1 Example 1: A Day in the Life of College Admissions

As a member of the college admissions staff at Super Omnia Gradibus University (SOG U), you have been put in charge of determining the students who will be offered admission to the university. SOG U's motto is 'Above all: grades,' and seeks to admit only those students who will achieve a certain threshold grade point average in their first semester of college. SOG U has access to the application material and first semester GPA for 4137 of its admitted freshman from the past four years[1].

The resident statistician, Dr. Regressus, has crunched the numbers and developed five linear regression models to predict the first semester GPA of a college freshman based on their overall SAT score, their class rank as a percentile, and the size of their graduating high school in hundreds of students. Overpaid and underworked, he provided the following table of regression coefficients summarizing each of the models:

[1] Actual data from *Introductory Econometrics: A Modern Approach* by Woolbridge (2012). Data file from `http://hedibert.org/wp-content/uploads/2014/02/gpa2-wooldridge.txt`.

Table 1: The regression coefficients for the five linear models developed by Dr. Regressus.

Model	b_0 (GPA points)	b_{SAT} $\left(\frac{\text{GPA points}}{\text{SAT point}}\right)$	$b_{\text{HS.PERC}}$ $\left(\frac{\text{GPA points}}{\text{percentile point}}\right)$	$b_{\text{HS.SIZE}}$ $\left(\frac{\text{GPA points}}{\text{100 students}}\right)$
Model 1	2.65	—	—	—
Model 2	0.66	0.0019	—	—
Model 3	1.28	—	0.017	—
Model 4	2.68	—	—	-0.011
Model 5	0.083	0.0015	0.014	-0.023

1.1 Simple Linear Regressions

The first four models provided by Dr. Regresssus are simple linear regressions of a student's first semester GPA on each of the predictors individually. In a rush, Dr. Regresssus forgot to provide any of the statistical details about standard errors and P-values, so all we know is that the simple linear models are

$$\textbf{Model 1: } \widehat{\text{GPA}} = \_\_\_\_\_ \tag{1}$$

$$\textbf{Model 2: } \widehat{\text{GPA}} = \_\_\_\_\_ + \_\_\_\_\_ \times \text{SAT} \tag{2}$$

$$\textbf{Model 3: } \widehat{\text{GPA}} = \_\_\_\_\_ + \_\_\_\_\_ \times \text{HS.PERC} \tag{3}$$

$$\textbf{Model 4: } \widehat{\text{GPA}} = \_\_\_\_\_ + \_\_\_\_\_ \times \text{HS.SIZE} \tag{4}$$

The first model is really a very special case when we assume that we know nothing about a student except that they are in the typical applicant pool for SOG U. In that case, our best guess at their first semester GPA is the average first semester GPA of a student in the applicant pool.

Notice that the intercepts (b_0) for each model are different. This makes sense if we recall that the intercept corresponds to our prediction of the first semester GPA when the predictor is 0: the expected GPA could be very different for someone from a very small high school (Model 4) compared to someone with a very low SAT score (Model 2). This observation should be generalized to all of the coefficients: they must always be interpreted in the context of the particular model under consideration.

1.2 Multiple Linear Regression

The first four models consider the predictors one-by-one or not at all. Depending on the choice of predictor we include, the predicted outcome can vary dramatically for a single student. Ideally, we would like to predict a student's GPA using all of the available predictors at once.

To do so, we can use Dr. Regressus's fifth model,

$$\textbf{Model 5: } \widehat{\text{GPA}} = \_\_\_\_\_ + \_\_\_\_\_ \times \text{SAT} + \_\_\_\_\_ \times \text{HS.PERC} + \_\_\_\_\_ \times \text{HS.SIZE} \tag{5}$$

If the multiple linear regression model is a good approximation to the data, it implies certain things about how a student's expected GPA varies with their SAT score, high school percentile, and high school size. In particular, it implies that the predicted score varies **independently** with respect to each predictor, and that a unit increase in each predictor will increase the predicted GPA by the amount given by the coefficients. For example, b_{SAT} has units of $\left[\frac{\text{GPA points}}{\text{SAT point}}\right]$, and thus for each additional SAT point, the model predicts an increase of b_{SAT} to the first semester GPA with HS.PERC and HS.SIZE fixed. In applications of multiple linear regression, this model assumption is often stated as 'varying predictor x_j with all else being equal,' and the multiple linear regression is thought of as 'controlling' for other predictors. In economics, the fancier *argumentum ad Latine* phrase *ceteris paribus* ("other things equal") is used.

Hands On 1

In the multiple linear regression model, the regression coefficient $b_{\text{SAT}} = 0.0015$ for SAT is positive, meaning that the model predicts an increase of 0.0015 to the first semester GPA for each additional SAT point. Does this mean that a student's parents should shell out their hard-earned cash to pay for SAT Prep in hopes of improving their child's first semester GPA? Why or why not?

Similarly, the regression coefficient $b_{\text{HS.SIZE}} = -0.023$, implying that the model predicts an increase of 0.023 to the first semester GPA for each 100 fewer students at the student's high school. Should the parents immediately put their child in home schooling to maximize their eventual first semester GPA? Why or why not?

1.3 Statistical Inferences from a Fitted Model

So far, we have only considered point predictions using Dr. Regressus's models. But as budding statisticians, we really want to know how much confidence we should place in the model. To answer that question, after a long lunch, Dr. Regressus shared the following summary table about Model 5:

Table 2: Table of inferential statistics for the the multiple linear regression provided by Dr. Regressus.

Term	Coef	SE Coef	T-Value	P-Value
Constant	0.08346	0.07035	1.186	0.236
SAT	0.001493	0.00006525	22.879	0.000
HS.PERC	0.01357	0.0005482	24.745	0.000
HS.SIZE	−0.02307	0.005027	−4.589	0.000

This table lists 4 inferential statistics:

1. **Coef** are precisely the coefficients originally provided by Dr. Regressus in Table 1, now with a few more decimal places.

2. **SE Coef** stands for the standard error in the estimate of the coefficient. A good rule of thumb, if we are willing to make a strong normality assumption, is that the interval

$$\text{Coef} \pm 2 \times \text{SE Coef} \tag{6}$$

will capture the true regression coefficient about 95% of the time, *i.e.* this is a rough 95% confidence interval[2]. For example, [0.00136250, 0.00162350] is a rough 95% confidence interval for the population coefficient on the SAT score.

3. **T-Value** is just $\frac{\text{Coef}}{\text{SE Coef}}$, which is used to compute the P-value for the coefficient.

4. Each **P-Value** is computed using the T-Value, and gives the probability of observing an absolute T-value that extreme or larger under a null model where all the other coefficients are estimated but that coefficient is fixed at 0. Thus, we should think of this as addressing the question, "Is the population coefficient for this predictor statistically significantly different from 0?" **Warning:** It is never correct to say, "The coefficient is statistically significant," since that leaves the underlying null model unspecified.

We see that all of the coefficients are statistically significantly different from 0 except for the intercept, which has a P-value of 0.236. This means that we would reject the null hypothesis

$$H_0 : \beta_0 = 0 \text{ in a model including GPA, HS.PERC, and HS.SIZE.} \tag{7}$$

Note that our rough 95% confidence interval for β_0 is [−0.05724, 0.22416]. In other words, we are not certain about its whereabouts, and therefore cannot reject the null hypothesis that it is different from 0.

There is also a P-value for the overall regression, which is computed under the null hypothesis that

$$H_0 : \beta_{\text{GPA}} = \beta_{\text{HS.PERC}} = \beta_{\text{HS.SIZE}} = 0. \tag{8}$$

[2]For a data set with N examples and k coefficients (not including the intercept), we should really use a t-based cutoff with $N - k - 1$ degrees of freedom. But for $N > 30$, the normal-based cutoffs are very nearly the same as the t-based cutoffs.

The overall P-value reported by Dr. Regressus is $< 2.2 \times 10^{-16}$, which is highly *statistically* significant. This P-value says that we would have to be crazy to assume that all of the regression coefficients are zero. But remember: statistical significance is not practical significance! Nothing we have seen so far has quantified how well any of Dr. Regressus's models perform at their actual job: prediction! We turn to that task next.

1.4 Model Summaries

After bugging Dr. Regressus one more time, he finally cops to how well his models perform at prediction with the following table:

Table 3: Table of Model Summary statistics for the multiple linear regression from Example 1.

Model	S	R-sq	R-sq(adj)
1	0.6586	0.0%	0.0%
2	0.6012	16.70%	16.68%
3	0.5952	18.36%	18.34%
4	0.6584	0.08156%	0.0574%
5	0.5602	27.71%	27.66%

If we denote the observed GPA of a student by y_i and the predicted GPA using a particular model by $\widehat{y}_i$, the **error** e_i for each prediction is the difference between the observed and predicted value,

$$e_i = y_i - \widehat{y}_i. \tag{9}$$

We have one of these error terms per each prediction. The components of Table 3 are computed using some combination of the observed outcomes, the predicted outcomes, and the error between the two:

1. **S** is the standard error of prediction,

$$S = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2} \tag{10}$$

$$\approx \sqrt{\mathsf{Var}(\{e_i\}}. \tag{11}$$

   The standard error of the prediction quantifies how far off, on average, the predicted value $\widehat{y}$ is from observed value y. A *smaller* value is better.

2. **R-squared** is a generalization of the coefficient of determination you are familiar with from simple linear regression,

$$R^2 = \frac{\mathsf{Var}(\{\widehat{y}_i\})}{\mathsf{Var}(\{y_i\})}. \tag{12}$$

In a perfect model, the predicted outcomes would perfectly match the observed outcomes, and thus the variances would be equal, giving $R^2 = 100\%$. At the opposite extreme, if the predicted outcomes do not vary at all with the predictors, then their sample variance is 0, giving $R^2 = 0\%$. Values between 0 and 1 indicate worse and better fits to the data[3].

3. **R-squared(adj)**, as the name suggests, is an adjustment of R^2 that reflects the fact that including more predictors will generically lead to a larger R^2, regardless of whether a predictor is truly associated with the outcome. The adjusted R^2 does account for this fact, and is thus more useful for comparing different multiple linear regression models on the same data.

We see that in this particular case, all three model summary statistics give the same ranking of the models, with the multiple linear regression (Model 5) performing best, and the simple linear regression of the graduating class size (Model 4) performing worst. Based on the standard error of prediction S, using Model 5 we expect our prediction to be within $\pm 2 \times 0.5602 \approx 1.1$ GPA points, while using Model 4 we only expect to be within $\pm 2 \times 0.6584 \approx 1.3$ GPA points. Neither model works particularly well *predictively*, given the standard breakdown of GPA into letter grades typically allots one to two letter grades per unit of GPA score.

1.5 Model Diagnostics

As with simple linear regression, to have any trust in the P-values or standard errors, we need for certain assumptions (linearity, normality, etc.) to hold about the data. You can (and should!) diagnose places where these assumptions fail to hold using the same diagnostic plots you learned about for simple linear regression.

2 The Multiple Linear Regression Model

The model from the previous section is one particular example of a **multiple linear regression**. Like a simple linear regression, we predict a **response / outcome** (y) as a linear function of k **predictors** $(x_1, x_2, \ldots, x_k)$ with the **intercept** b_0 and **coefficients** $(b_1, \ldots, b_k)$ using the model

$$\widehat{y} = b_0 + b_1 x_1 + \ldots + b_k x_k. \tag{13}$$

The mechanics for working with this model are precisely those learned in the previous hands on sections. For example, b_j corresponds to the change in $\widehat{y}$ for a unit change in x_j. The intercept and coefficients must either be assumed *a priori* or, more commonly, estimated from data via least squares. We turn to how to do this in Minitab next.

[3] If you ever encounter a multiple (or simple) linear regression in the wild, you will likely hear R^2 referred to as the proportion of the outcome variance 'explained' by the linear model. This is an unfortunate short hand for what R^2 actually measures, since it does not *really* have anything to do with an explanation as the term is commonly used.

3 Example 1 Revisited: Fitting a Multiple Linear Regression in Minitab

Dr. Regressus has graciously provided us with his data so we can take a look. The characteristics of the students are stored in the file `gpa2-wooldridge-subset.txt`. Use Minitab to open the data file `gpa2-wooldridge-subset.txt` in the data directory of the downloaded Github directory from `http://bit.ly/ma151-mlr` by selecting

> `File` → `Open...` from the menu bar.

from the menu bar.

3.1 Pairwise / Matrix Plot in Minitab

The first step in any data analysis is to look at the data in an appropriate representation for the eventual analysis. For a multiple linear regression, one appropriate representation is a matrix of plots that shows the pairwise relationship amongst each of the predictors and the outcome variable. To generate one such pairwise plot, use

> `Graph` → `Matrix Plot...` → `Matrix of plots` → `Simple`

Figure 1: The pairwise plots of the predictors (SAT, HS.PERC, and HS.SIZE) and outcome (GPA) for the 4137 students in the data set. Each point corresponds to a single student.

Hands On 2

Add your best guess at the lines of best fit to the bottom row of the matrix. How does your intuitive guess compare to the simple linear regressions from the first section? You can add these regressions to the plots by Right-clicking on the plot and selecting

```
Add → Regression Fit...
```

3.2 Estimated Coefficients and Inferential Statistics

Now we are ready to check Dr. Regressus's work, and get some additional details about the regression estimates. To fit the multiple linear regression in Minitab, select

`Stat` → `Regression` → `Regression` → `Fit Regression Model...`

from the menu bar. Choose GPA as the response, and SAT, HS.PERC, and HS.SIZE as the Continuous predictors. You should recognize the Model Summary, Coefficients, and Regression Equation printouts in the Session window. There is also an Analysis of Variance printout, which is largely redundant with the Model Summary printout. However, this printout does contain the overall statistical significance relative to the null model that all of the regression coefficients are 0, listed in the 'Regression' row.

3.3 Model Summary Statistics

The Model Summary statistics (standard error of the prediction, R^2, and adjusted R^2) are given in the Model Summary printout of the Session window. Minitab also provides a predictive R^2, which is an in-house special sauce model summary to address the overfitting issue of R^2. We will not address the predictive R^2 in this class.

4 Example 2: Banking on a Pay Gap

XYZ Bank[4] was brought to court over possible discrimination in hiring salaries based on gender. You have been called to the stand as a expert witness for the prosecution. The bank has provided you information about 116 of its employees, including:

- Starting salary (in dollars)
- Age (in months)
- Work experience (in months)
- Years of education
- Gender

These are stored in the Minitab project file `MR BANK WAGES.MPJ`.

[4] Actual data provided by Professor Bastian

Hands On 3

Load `MR BANK WAGES.MPJ` into Minitab. Create a matrix of plots like you did for the previous data set. Add simple linear regressions to the matrix of plots. Do the trends of how starting salary is associated with each of the employee characteristics agree with what you would expect? Why or why not?

Up until now, we have only considered numerical predictors in our multiple linear regression models. Gender, which for the sake of this exercise we will take as either Male or Female, does not naturally fall into a numerical scale. It is a **categorical** variable, since each instance comes from one of the two categories of Male or Female. In the data set, we have coded the Gender characteristic as 0 for Male and 1 for Female.

Hands On 4

Look back at the matrix of plots for this data set. Find the column corresponding to the Gender predictor. Based on the Gender predictor alone, does there appear to be a pay gap between Male and Female employees? Why might it be unwise to argue simply from this plot that this implies gender *discrimination* at XYZ Bank?

Minitab's regression model can naturally handle categorical predictors by including them in the `Categorical predictors` box. Let's fit a multiple linear regression model to the data

Hands On 5

What is the form of the multiple linear regression model estimated by Minitab? How does it handle the Gender predictor?

$$\widehat{\text{Salary}} =$$

The coefficient on Gender0 is most relevant to the question of pay gap between the male and female employees.

Hands On 6

How do you interpret the Gender0 coefficient? What component of the Minitab printout would argue for or against the presence of a pay gap between the employees? What conclusion do you draw from this statistic? What is the null hypothesis that is being tested?

5 Example 3: Applied Phrenology

A study from University of Texas Austin in 2002[5] sought to determine if an association existed between brain size and intelligence. To investigate this question, 38 undergraduate students were scored on the revised Wechsler Adult Intelligence Scale and their brain volume as assessed by Magnetic Resonance Imaging (MRI). Their height (recorded in inches) and and weight (recorded in pounds) were also measured[6].

[5] Data from Willerman, L., Schultz, R., Rutledge, J. N., and Bigler, E. D. (1991). *In vivo* brain size and intelligence. *Intelligence*, 15(2), 223-228. Data obtained from `https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/iqsize.txt`.

[6] There is a long history in statistics of collecting and analyzing body measurements, a field known as biometrics. One of the most prestigious journals in the field, *Biometrika*, was initially started to encourage the study of

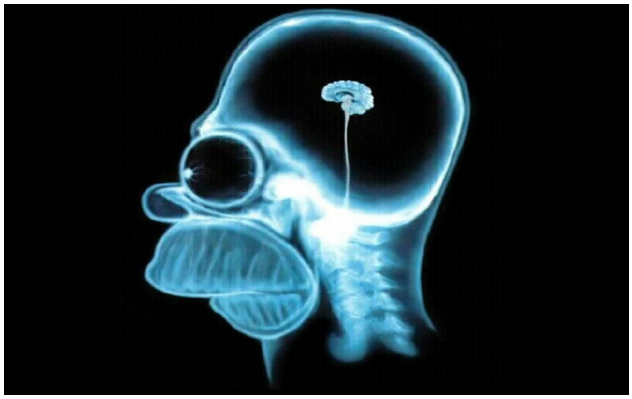

Figure 2: A potential outlier from the UT Austin study.

Hands On 7

Use the data in `iqsize.txt` to fit a multiple linear regression model to predict the IQ of a student using their brain volume, height, and weight. Record the estimated model in the table below.

Term	Coef	SE Coef	P-Value
Constant			
brain			
height			
weight			

What is the multiple linear regression model?

$$\widehat{\text{PIQ}} = \tag{14}$$

5.1 Model Selection and Backward Elimination

We saw in the previous section that the coefficient for weight in the multiple linear regression cannot be said to be statistically significantly different from 0. This *might* mean we can remove

biometrics.

it from our model[7]. Proceeding by sequentially removing the predictor with the largest P-value, greater than some threshold α, is called backward elimination, and is a particular form of model selection. Backward elimination proceeds in the following steps:

Backward Elimination

1. Fix a threshold α.
2. Fit the full multiple linear regression including all of the predictors.
3. While at least one of the P-values for a coefficient is greater than α:
   (a) Find the predictor with the largest P-value greater than α.
   (b) Refit the multiple linear regression, now excluding that predictor.

Hands On 8

Manually perform a backward elimination to select the predictors to include in the model. What predictors are included? How do S and the adjusted R^2 for the sub-model compare to the full model?

[7]Deciding to remove predictors based on their P-value is an old idea, and not necessarily a *good* one. Just because we cannot tell if a coefficient should be 0 using the data in hand does not mean that predictor is not relevant for prediction. We might just need better data!