

Simple Linear Regression

Collin Nolte

April 13, 2021

Last week we covered the case of simple linear regression

- X and Y are continuous variables
- Assume a linear relationship between them
- $Y = \beta_0 + \beta_1 X + \epsilon$
- $\hat{\beta} \sim N(\beta, \text{var}(\beta))$
- $\hat{\beta}/\text{sd}(\hat{\beta}) \sim t_{n-1}$
- Model assumptions, checked with residual plots

Multiple Regression

Previously, we had only considered the relationship between two variables, which resulted in our fit being a line,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

As we add more explanatory variables (X), the dimension of our fit increases. For example, with two explanatory variables, instead of a line, we will have a square

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

Despite the "squareness" of this new model, we still consider it a linear function (and consequently, we are still doing linear regression)

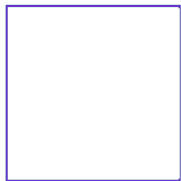
Dimensions



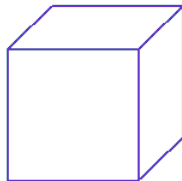
0



1



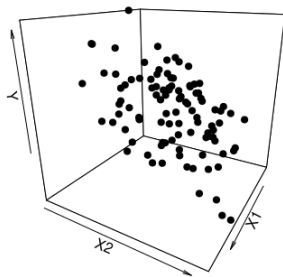
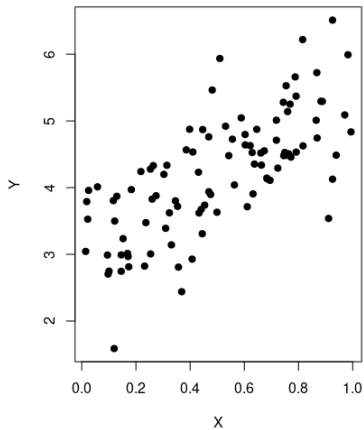
2



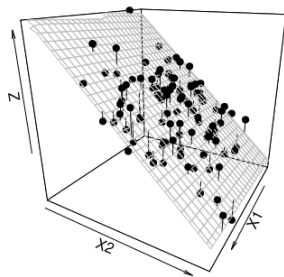
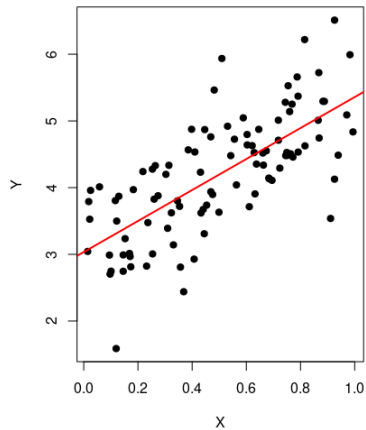
3

Dimensions

Observed Data



Fit



Interpretation

Although we have increased our dimensions, everything in multiple regression is analagous to what was done in simple regression, including interpretation and model assumptions.

For example, given

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2,$$

we would interpret $\hat{\beta}_2$ as a unit change in X_2 results in a $\hat{\beta}_2$ change in \hat{y} , *with the value of X_1 being fixed*

Iris dataset

- Collected 50 flowers from each of three species of iris flowers
- Measurements taken on the length and width of the petals and sepals, taken in centimeters
- Species include *Iris setosa*, *versicolor*, and *virginica*
- Ignoring species for now, we will try to fit a model for predicting sepal length, given sepal width and petal dimensions



Iris dataset

The fitted model can be written as

$$\hat{Y} = 1.85 + 0.65X_1 + 0.71X_2 - 0.56X_3$$

where Y = sepal length, and X_1 , X_2 , and X_3 are sepal width, petal length, and petal width, respectively

We could interpret as follows: *with other X values being fixed*, a centimeter change in sepal width leads to a 0.65 centimeter increase in sepal length. Similarly, a centimeter change in petal width corresponds to a -0.56 centimeter change in sepal length

Types of covariates

Until this point, we have only considered covariates that are continuous in nature, such as petal length, or muscle mass

Often, however, we might wish to include a *categorical* variable in our regression model, for example, sex or treatment type. In doing so, we are able to consider regression values in different groups

We will consider an example in which odontoblasts (cells responsible for tooth growth) were measured in 60 guinea pigs, each receiving one of three doses of vitamin C (0.5, 1, 2 mg/day) by one of two methods of delivery (orange juice or ascorbic acid)

Categorical variables are often coded with *indicators*, with a value of 1 for one group and a value of 0 for others

Guinea pigs

Here, we have the fitted model for the guinea pig data

$$\hat{y} = 9.27 + 9.76 \times \text{Dose} - 3.7 \times \text{AbsorbicAcid}$$

We see from this that, within each group, a milligram increase in the dose of vitamin C resulted in a 9.76 micron increase in the length of odontoblasts

The categorical variable in this case has the value 1 for guinea pigs receiving absorbic acid, indicating that, as a whole, this group had odontoblasts that were 3.7 microns shorter than the orange juice group.

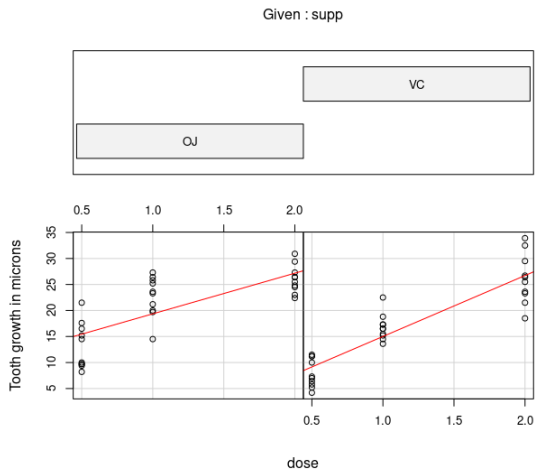
For a vitamin C dose at 1mg/day, we would then predict

$$\hat{y} = 9.27 + 9.76 = 19.03 \text{ microns}$$

$$\hat{y} = 9.27 + 9.76 - 3.7 = 15.33 \text{ microns}$$

for guinea pigs with orange juice and absorbic acid, respectively

Guinea pigs



Control Variables

We will also often be interested in including *control variables* in our model, which may not be variables of interest, but seek to control confounding in our model

Put another way, our outcome variable has some total amount of variance (SS_{Total}), and we include covariates in order to "account" for this variance. The more variance explained by a covariate, the more likely it is to have a relationship with the outcome. Including control variables is a productive way to mop up this excess variance or, more critically, control for confounding

For this example, we will consider data extracted from the 1974 Motor Trend magazine, measuring fuel consumption along with 10 additional aspect of vehicle design for 32 cars. We are interested in investigating the relationship between mpg and vehicle weight

mpg vs weight

Call:

```
lm(formula = mpg ~ wt, data = mtcars)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.543	-2.365	-0.125	1.410	6.873

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	37.285	1.878	19.86	< 0.0000000000000002 ***
wt	-5.344	0.559	-9.56	0.00000000013 ***

Residual standard error: 3.05 on 30 degrees of freedom

Multiple R-squared: 0.753, Adjusted R-squared: 0.745

F-statistic: 91.4 on 1 and 30 DF, p-value: 0.000000000129

mpg vs weight + controls

Call:

```
lm(formula = mpg ~ wt + disp + carb, data = mtcars)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.074	-1.839	-0.352	1.310	5.684

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	35.49063	2.01903	17.58	<0.00000000000000002	***
wt	-2.87249	1.09765	-2.62	0.014	*
disp	-0.01697	0.00853	-1.99	0.056	.
carb	-0.79718	0.33286	-2.39	0.024	*

Residual standard error: 2.7 on 28 degrees of freedom

Multiple R-squared: 0.818, Adjusted R-squared: 0.799

F-statistic: 42 on 3 and 28 DF, p-value: 0.00000000017

Controlling variables

As we saw, including displacement and the number of carburetors decreased the effect that weight had on vehicle mileage, while each of these in turn had effects in the same direction (that is, an increase in either resulted in negative impact on mpg)

This makes some sense: we might imagine that larger vehicles (which weigh more) would also have larger engine displacement and more carburetors

By accounting for these in our model, we are able to get a more accurate idea of what the true impact of weight might be on mileage

Inference

For each of the models just considered, there were $n = 32$ total observations. When our data is written as a matrix, this indicates that we have 32 total rows

The number of covariates in our model, then, makes up the number of columns, designated p . In the first model, with only weight, we had $p = 1$. After adding displacement and the number of carburetors, we had $p = 3$.

The relationship of n to p is of critical importance: the larger n is relative to p , the better a fit (and the smaller the variance) we will have in our model. For typical regression, we will always require that $n > p$, though there are special methods for handling the $p > n$ case, which is common when performing regression on genetic arrays

Inference

The most immediate consequence of the relation of n to p comes in the t -statistic generated by the parameter estimates. In the simple regression case, we indicated that

$$\frac{\hat{\beta}}{sd(\hat{\beta})} \sim t_{n-1}$$

However, in the case of multiple regression, it follows that

$$\frac{\hat{\beta}}{sd(\hat{\beta})} \sim t_{n-p}$$

Recall that as $n - p$ gets larger, the variance of this distribution gets smaller. As n is usually fixed, we are limited by the number of covariates we can include. It's worth asking, then, if the addition of an extra covariate is worth reducing the value of $n - p$

Last week, we introduced the concept of R^2 , giving information on how much variance is captured in the model

$$R^2 = 1 - \frac{SS_{Residual}}{SS_{Total}}$$

$$SS_{Total} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad SS_{Residual} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

As we add more and more variables, \hat{y}_i will never get further away from y_i . It can either make our estimate much better, or more or less the same, but never worse.

Only considering R^2 , it will always appear that adding more variables is better

Adjusted R^2

We might then consider a value known as *adjusted* R^2 , or R_{adj}^2 , given

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

The algebra doesn't work out nicely in comparing to the original R^2 , but we can illustrate with an example: suppose $n = 10$, and $p = \{1, 2, 3\}$

$$\frac{10 - 1}{10 - 1 - 1} = 1.125, \quad \frac{10 - 1}{10 - 2 - 1} = 1.285, \quad \frac{10 - 1}{10 - 3 - 1} = 1.5$$

From $p = 1$ to $p = 2$, this inflation factor increases by 0.16. From $p = 2$ to $p = 3$, by 0.21. Each additional covariate increases the inflation by a marginally greater amount. In other words, the more covariates we already have, the greater the justification we need to add another

Adding covariates - Example

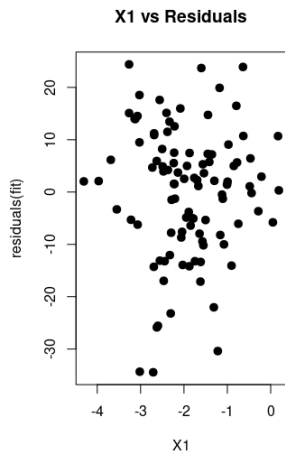
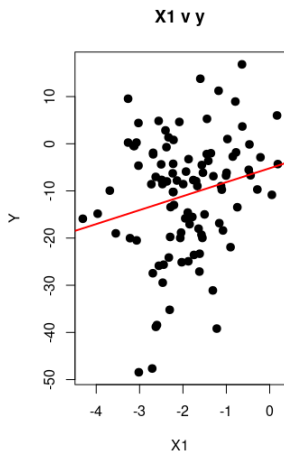
An interesting way to consider adding variables is to consider plots of residuals. For this contrived example with random data, suppose we have two covariates X_1 and X_2 , and a true model that looks like

$$y = 1 + 3X_1 - 4X_2$$

We might start with a model that only includes X_1 , that is, our first model is

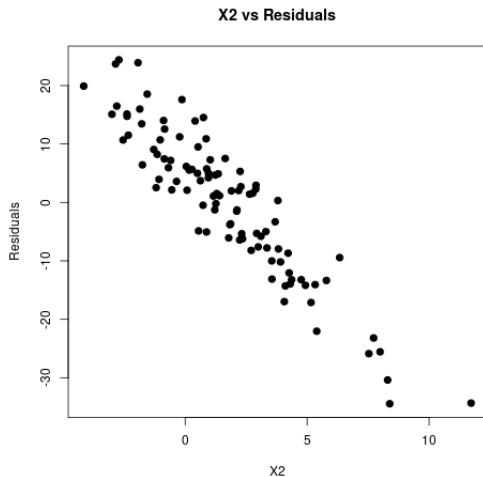
$$y = \beta_0 + \beta_1 X_1$$

Adding covariates - Example



Adding covariates - Example

$$y = 1 + 3X_1 - 4X_2$$



Model Selection

Consider what is being shown in this second plot. The residuals represent a measure of the outcome that was not captured by X_1 alone

In particular, we saw that there was a strong linear relationship between the residuals and the covariate that was not in the model, X_2 .

In general, then, a way to consider the process of adding covariates is to ask yourself, “How much of the model residual is accounted for by this missing covariate?”

Although this is only a slight simplification, the process of *model selection* is itself an entire sub-discipline of statistics

Multicollinearity

As we increase the number of covariates in our model, there are a number of potential pitfalls to be on the lookout for, the most significant of which is the issue of *multicollinearity*

In the simplest case, we say that two covariates X_1 and X_2 are (perfectly) collinear if there is an exact linear relationship between them, i.e., if

$$X_1 = a + bX_2$$

There are a number of ways to interpret how this can cause issues, and we will consider a few in detail. Although the interpretations will be slightly different, the underlying phenomenon is the same in each case

Women height

We have previously considered a dataset in which we recorded the average height in inches and weight in pounds of american women aged 30-39. The range of heights is 58-72 inches

$$\hat{y} = -87.52 + 3.45X_1$$

Now suppose that we included a variable X_2 , which measured women's height in feet as well (with 1ft = 12in), and consider the model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

As mentioned previously, we interpret the value of $\hat{\beta}_2$ to be “for every 1ft change in height, weight changes by $\hat{\beta}_2$, *everything else being fixed.*”

Of course, in this situation, it would be *impossible* for X_2 to change without X_1 changing, as $X_1 = 12 \cdot X_2$

Some linear algebra

Behind the scenes, these problems are solved with linear algebra. Suppose that we have two covariates, where $X_2 = 2X_1$, and we wish to estimate β_1 and β_2

$$\begin{array}{r} 4 = 2\beta_1 + 4\beta_2 \\ - 3 = \beta_1 + 2\beta_2 \\ \hline 1 = \beta_1 + 2\beta_2 \end{array}$$

Here, there are an infinite number of solutions: $\beta_1 = 0$ and $\beta_2 = 1/2$ would be one, and $\beta_1 = 1$ and $\beta_2 = 0$ would be another; and while all would be able to estimate Y the same, we have no idea which of these is “correct”. This is problematic when we are specifically interested in knowing the true value of β

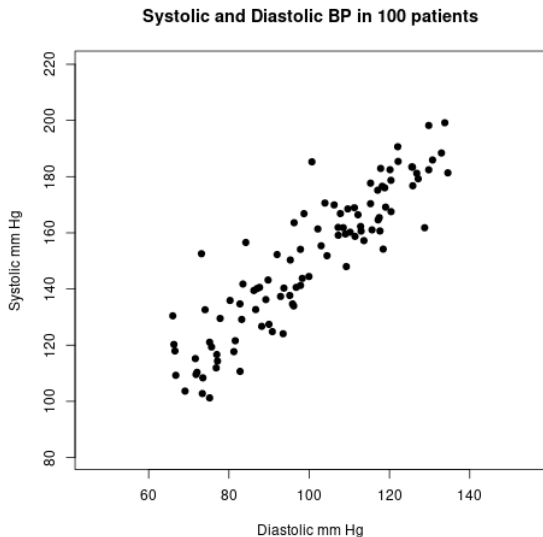
Hidden extrapolation

Perhaps a nicer way to see this is to consider a statement we made last lecture: it's important to not attempt to make predictions outside of the range of X . When X was a line, this was simple; we only had to consider the range of X .

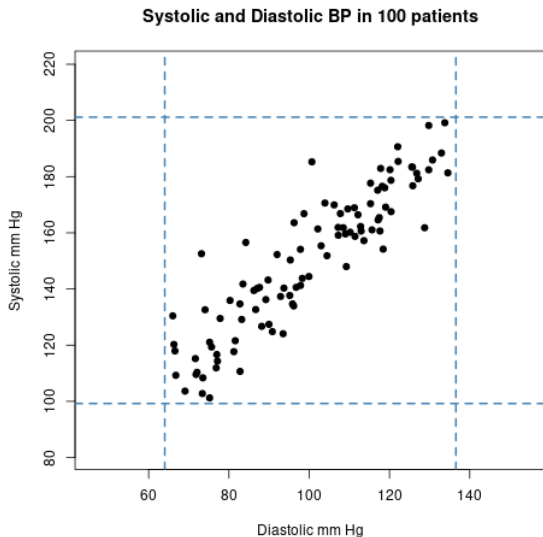
In the case of multiple variables, the issue is a bit trickier. Now let's consider a more realistic case in which X_1 and X_2 are no longer multiples of each other, but rather, are highly correlated.

For example, suppose a study collected both systolic and diastolic blood pressure on patients. Even though these are not directly related, they should be highly correlated

Hidden Extrapolation

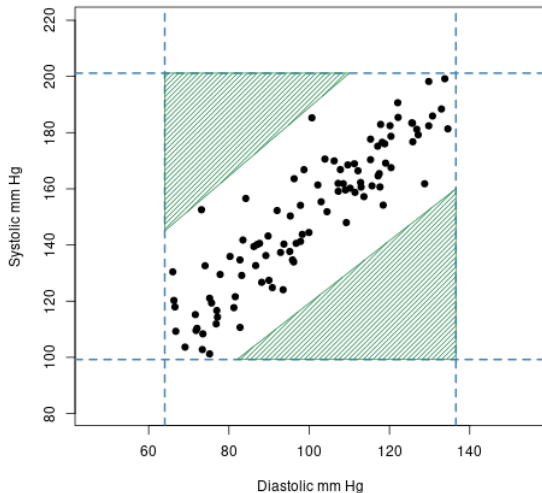


Hidden Extrapolation

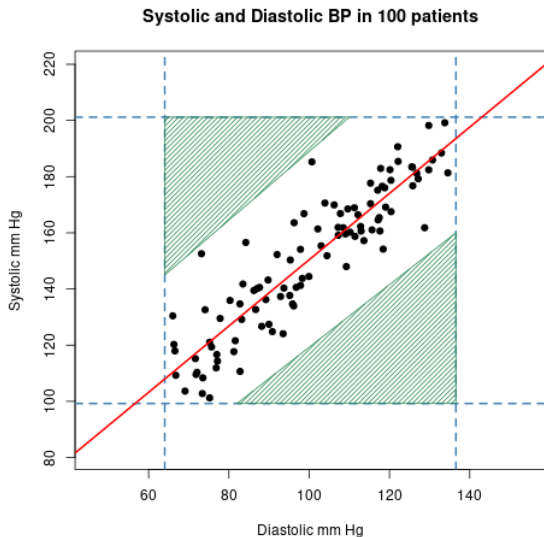


Hidden Extrapolation

Systolic and Diastolic BP in 100 patients



Hidden Extrapolation



Hidden Extrapolation

From this, there are two primary things to keep in mind:

1. Only considering the ranges of the X values respectively, we might consider it safe to predict, say, the outcome for an individual with diastolic BP of 80 and Systolic of 200 – however, we don't actually have any observations that fall in this range
2. More broadly, we see that what is ostensibly a box is also like a line. In other words, “on paper” we have increase our dimension from one to two, but in reality, it's more akin to something like one and a half. This idea will be *especially* relevant next week

A few notes

Nearly all of what we discussed last week in terms of residuals and model assumptions is true in the multivariate case

As one might imagine, even the cases discussed here have generalizations. For example, *logistic regression* is a case in which the outcome Y is binary (such as disease status), and the regression coefficients tell us about the change in *odds* given changes in the covariates

One may even change the assumption on error terms, and assume different underlying distributions. This falls under the category of *generalized linear models*

Lastly, there are cases addressing high-dimensional situations, where the number of covariates exceeds the number of outcomes. This is known as *penalized regression*

- *Applied Linear Statistical Models, 5th Edition*, Kutner, Nachtsheim, Neter, Li (2005)
- Crampton, E. W. (1947). The growth of the odontoblast of the incisor teeth as a criterion of vitamin C intake of the guinea pig. *The Journal of Nutrition*, 33(5), 491–504. doi: 10.1093/jn/33.5.491.