Contents

[simple and multiple regression coefficients can be quite different 1](#_Toc521583289)

[Omitted Variable Bias 2](#_Toc521583290)

[F Test for over all significance j 3](#_Toc521583291)

[Deciding on Important Variable – T Test 4](#_Toc521583292)

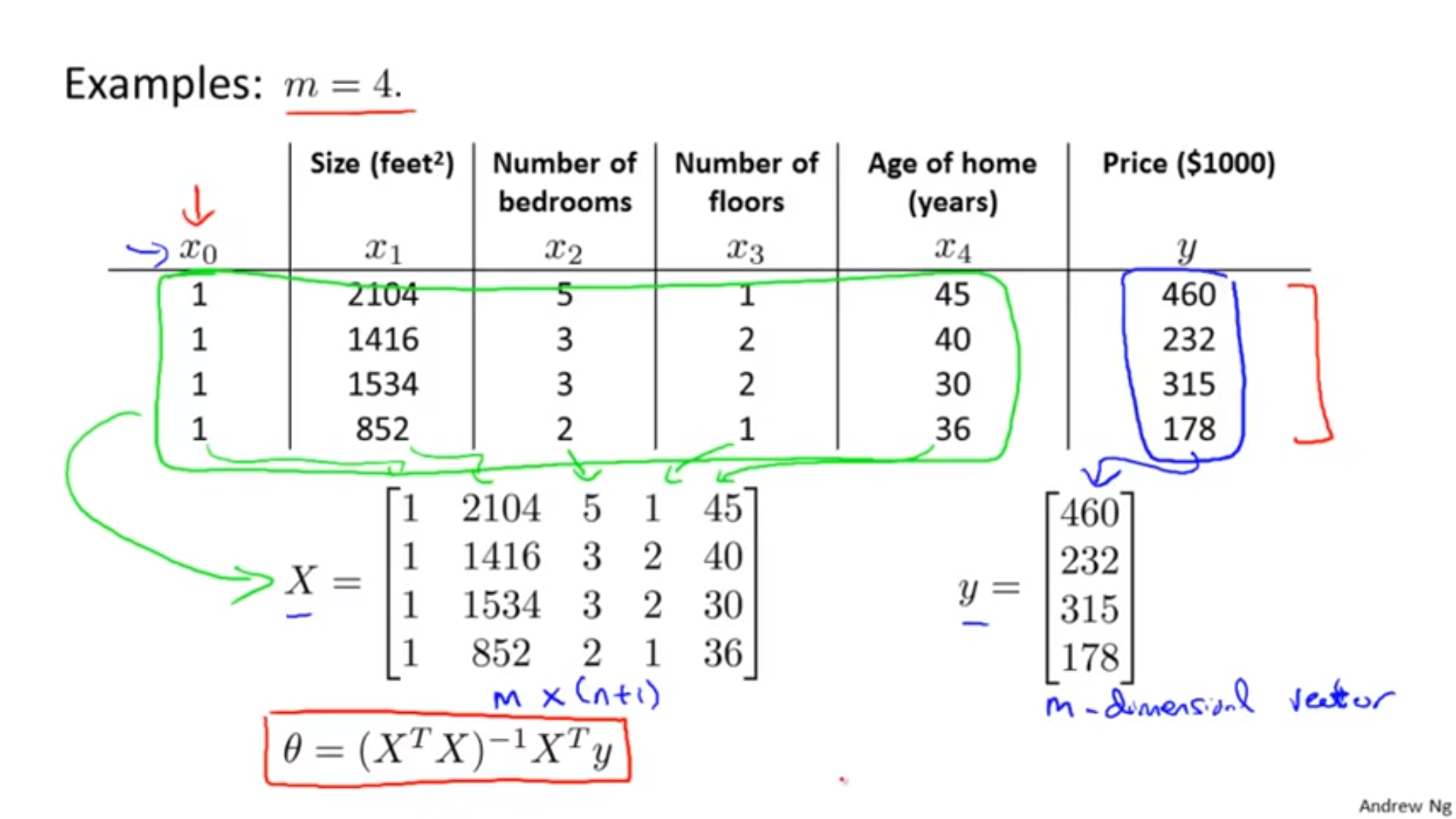
[Interview Questions 5](#_Toc521583293)

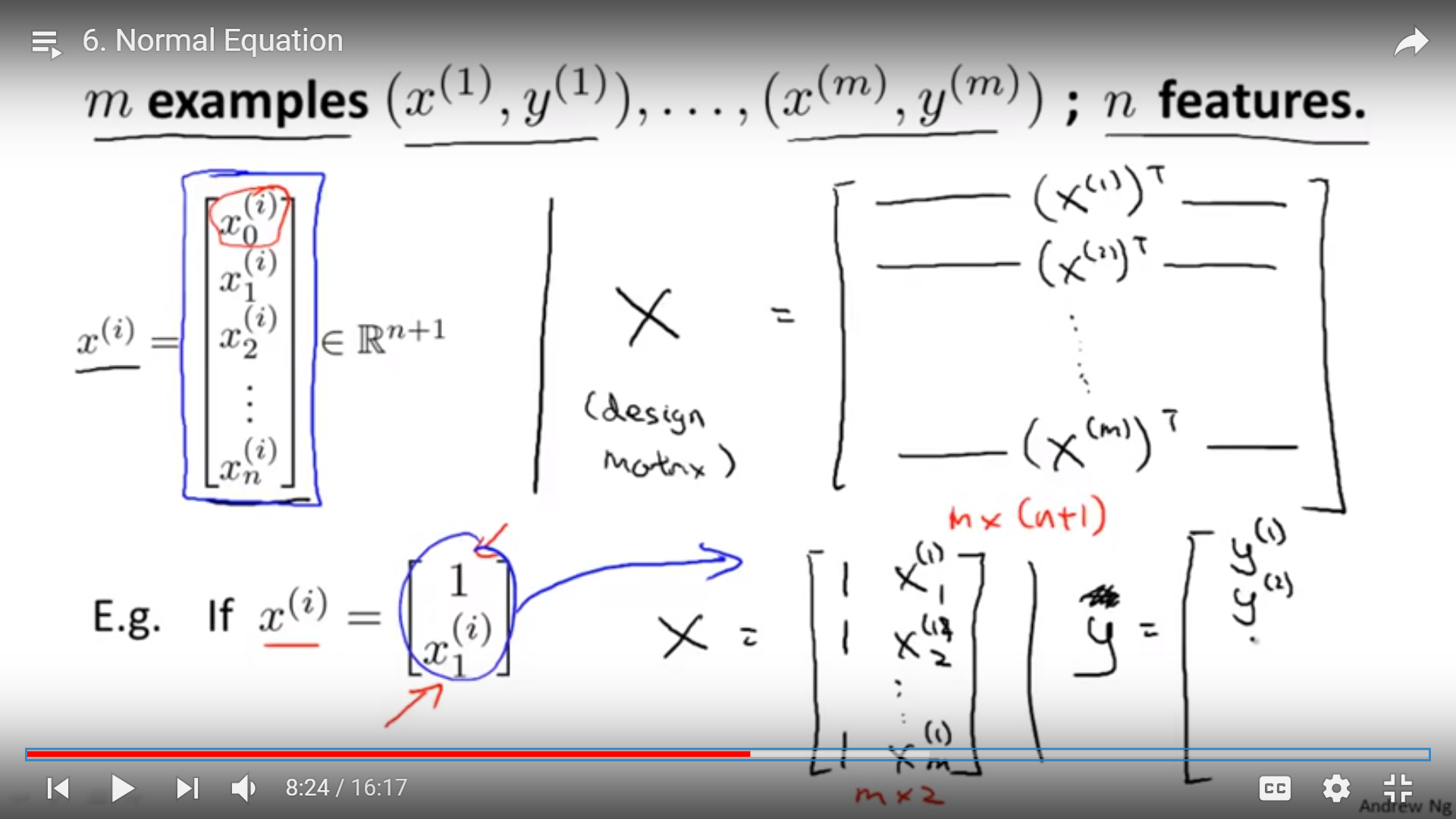
[What is the difference between a multiple linear regression and a multivariate regression? 5](#_Toc521583294)

[You have several variables that are positively correlated with your response, and you think combining all of the variables could give you a good prediction of your response. However, you see that in the multiple linear regression, one of the weights on the predictors is negative. What could be the issue? 5](#_Toc521583295)

[Regression coefficients that flip sign after including other predictors 6](#_Toc521583296)

# Multiple linear regression matrix form





Here in this example x(i)0 represent - 0 row ith columnx(column )(row)

# simple and multiple regression coefficients can be quite different

However, while the newspaper regression coefficient estimate in

Table 3.3 was significantly non-zero, the coefficient estimate for newspaper

in the multiple regression model is close to zero, and the corresponding

p-value is no longer significant, with a value around 0*.*86. This illustrates

that the simple and multiple regression coefficients can be quite different.

This difference stems from the fact that in the simple regression case, the

slope term represents the average effect of a $1*,*000 increase in newspaper

advertising, **ignoring other predictors such as TV and radio**. In contrast, in

the multiple regression setting, the coefficient for newspaper represents the

average effect of increasing newspaper spending by $1*,*000 while holding TV

and radio fixed.

**Does it make sense for the multiple regression to suggest no relationship between sales and newspaper while the simple linear regression implies the opposite?**

**In fact it does**.

Consider the correlation matrix for the three

predictor variables and response variable, displayed in Table 3.5. Notice

that the correlation between radio and newspaper is 0*.*35. This reveals a

tendency to spend more on newspaper advertising in markets where more

is spent on radio advertising.

Now suppose that the multiple regression is

correct and newspaper advertising has no direct impact on sales, but radio

advertising does increase sales. Then in markets where we spend more

on radio our sales will tend to be higher, and as our correlation matrix

shows, we also tend to spend more on newspaper advertising in those same

markets. Hence, in a simple linear regression which only examines sales

versus newspaper, we will observe that higher values of newspaper tend to be

associated with higher values of sales, even though newspaper advertising

does not actually affect sales. So newspaper sales are a surrogate for radio

advertising; newspaper gets “credit” for the effect of radio on sales.

# Omitted Variable Bias

In [statistics](https://en.wikipedia.org/wiki/Statistics), **omitted-variable bias** (**OVB**) occurs when a model created incorrectly leaves out one or more important factors. The "bias" is created when the model compensates for the missing factor by over- or underestimating the effect of one of the other factors.

Suppose the true cause-and-effect relationship is given by

{\displaystyle y=a+bx+cz+u}with parameters *a, b, c*, dependent variable *y*, independent variables *x* and *z*, and error term *u*. We wish to know the effect of *x* itself upon *y* (that is, we wish to obtain an estimate of *b*).

Two conditions must hold true for omitted-variable bias to exist in [linear regression](https://en.wikipedia.org/wiki/Linear_regression):

* the omitted variable must be a [determinant](https://en.wikipedia.org/wiki/Determinant) of the dependent variable (i.e., its true regression coefficient is not zero); and
* the omitted variable must be correlated with an independent variable specified in the regression (i.e., cov(*z*,*x*), is not equal to zero).

But suppose that we omit *z* from the regression, and suppose the relation between *x* and *z* is given by

{\displaystyle z=d+fx+e}

with parameters *d*, *f* and error term *e*. Substituting the second equation into the first gives

{\displaystyle y=(a+cd)+(b+cf)x+(u+ce).}

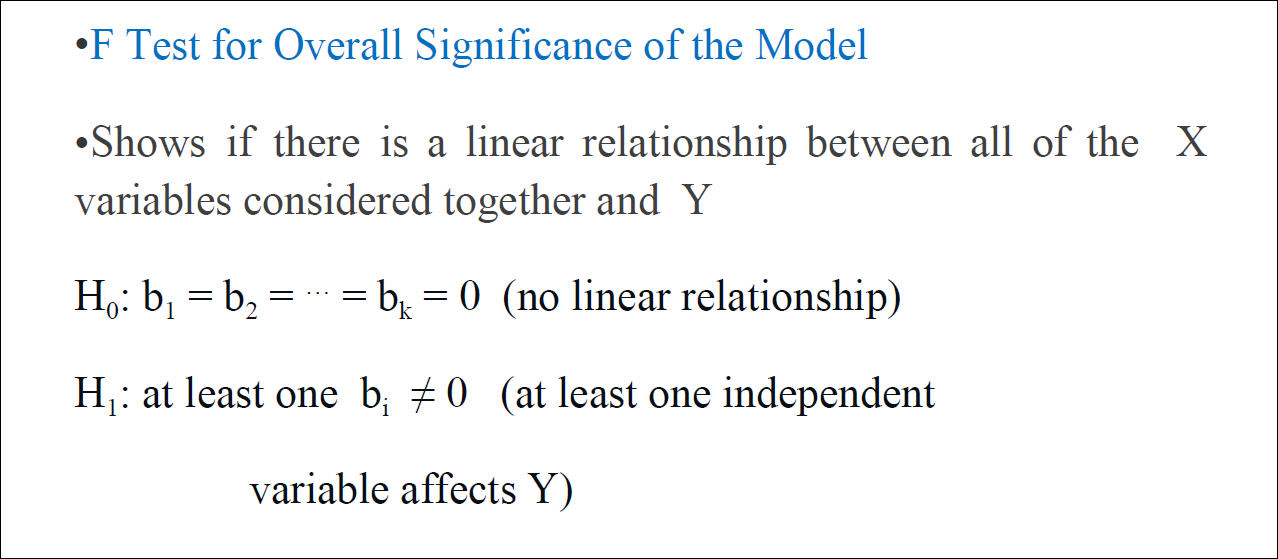
If a regression of *y* is conducted upon *x* only, this last equation is what is estimated, and the regression coefficient on *x* is actually an estimate of (*b* + *cf* ), giving not simply an estimate of the desired direct effect of *x* upon *y* (which is *b*), but rather of its sum with the indirect effect (the effect *f* of *x* on *z* times the effect *c* of *z* on *y*).

**Example**

Income is directly propotional to height

But in reality it is propotional to gender

# F Test for over all significance j



When *n* is large, an F-statistic that is just a little larger than 1 might still provide evidence against *H*0. In contrast,

a larger F-statistic is needed to reject *H*0 if *n* is small. When *H*0 is true and the errors *\_i* have a normal distribution, the F-statistic follows an F-distribution.

For any given value of *n* and *p*, any statistical software

package can be used to compute the p-value associated with the F-statistic

using this distribution. Based on this p-value, we can determine whether

or not to reject

Given these individual p-values for each variable, why do we need to look

at the overall F-statistic? After all, it seems likely that if any one of the

p-values for the individual variables is very small, then *at least one of the*

*predictors is related to the response*. However, this logic is flawed, especially

when the number of predictors *p* is large.

For instance, consider an example in which *p* = 100 and *H*0 : *β*1 = *β*2 =

*. . .* = *βp* = 0 is true, so no variable is truly associated with the response. In

this situation, about 5% of the p-values associated with each variable (of

the type shown in Table 3.4) will be below 0*.*05 by chance. In other words,

we expect to see approximately five *small* p-values even in the absence of

any true association between the predictors and the response. In fact, we

are almost guaranteed that we will observe at least one p-value below 0*.*05

by chance! Hence, if we use the individual t-statistics and associated pvalues

in order to decide whether or not there is any association between

the variables and the response, there is a very high chance that we will

incorrectly conclude that there is a relationship. However, the F-statistic

does not suffer from this problem because it adjusts for the number of

predictors. Hence, if *H*0 is true, there is only a 5% chance that the Fstatistic

will result in a p-value below 0*.*05, regardless of the number of

predictors or the number of observations.

The approach of using an F-statistic to test for any association between

the predictors and the response works when *p* is relatively small, and certainly

small compared to *n*. However, sometimes we have a very large number

of variables. If *p > n* then there are more coefficients *βj* to estimate

than observations from which to estimate them. In this case we cannot

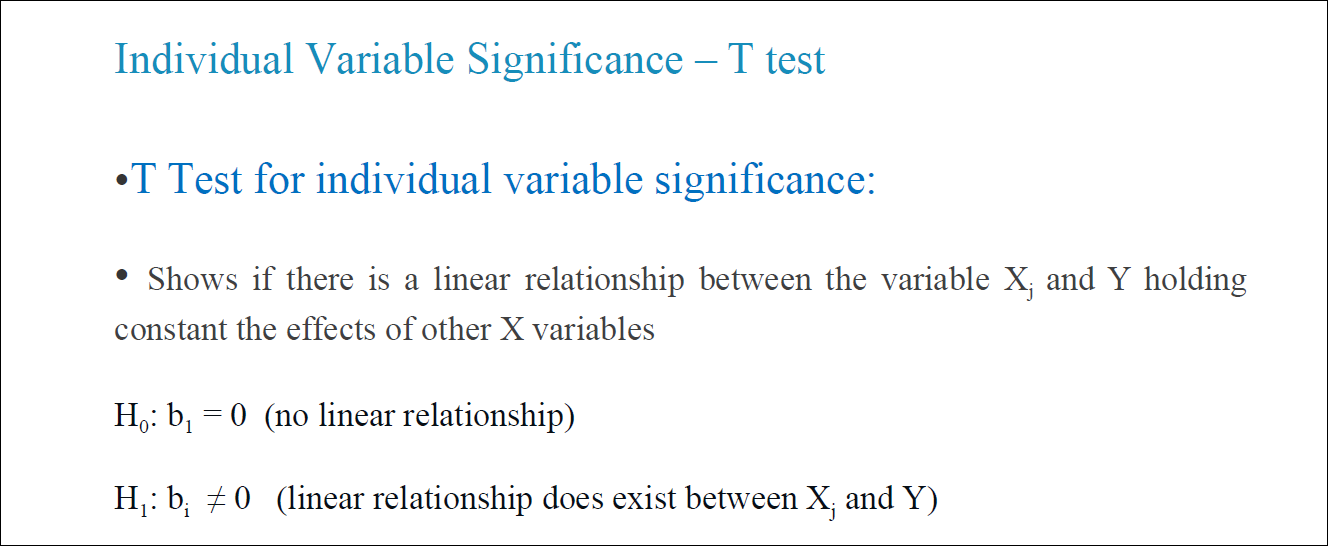
even fit the multiple linear regression model using least squares, so the F-statistic cannot be used, and neither can most of the other concepts that

we have seen so far in this chapter. When p is large, some of the approaches

discussed in the next section, such as forward selection, can be used. This

high-dimensional setting is discussed in greater detail in Chapter 6.

# Deciding on Important Variable – T Test



As discussed in the previous section, the first step in a multiple regression

analysis is to compute the F-statistic and to examine the associated pvalue.

If we conclude on the basis of that p-value that at least one of the

predictors is related to the response, then it is natural to wonder *which* are

the guilty ones! We could look at the individual p-values as in Table 3.4,

but as discussed, if *p* is large we are likely to make some false discoveries.

It is possible that all of the predictors are associated with the response,

but it is more often the case that the response is only related to a subset of

the predictors. The task of determining which predictors are associated with

the response, in order to fit a single model involving only those predictors,

is referred to as *variable selection*. The variable selection problem is studied

variable extensively in Chapter 6, and so here we will provide only a brief outline selection

of some classical approaches.

# Interview Questions

## What is the difference between a multiple linear regression and a multivariate regression?

**multiple**regression has more than one X in *one formula.*

A **multivariate**regression has more than one Y, but in *different formulae*. And a **multivariate multiple**regression has *multiple X’s to predict multiple Y’s* with each Y in a different formula, usually based on the same data.

You fit a multiple regression to examine the effect of a particular variable a worker in another department is interested in. The variable comes back insignificant, but your co-worker says that this is impossible as it is known to have an effect. What would you do?

This is a *multiple* regression. Presumably there are other variables. It may be the case that the specific variable in question has an effect on the response, but in the presence of the other variables its effect is not significant. I'd use partial least squares pairing the variable in question with the other variables to determine which other variables are "explaining away" the effect of the target variable and report my findings back to the contentious party, in addition to teaching them a little bit about multiple regression.

## You have several variables that are positively correlated with your response, and you think combining all of the variables could give you a good prediction of your response. However, you see that in the multiple linear regression, one of the weights on the predictors is negative. What could be the issue?

* Multicollinearity refers to a situation in which two or more explanatory variables in a [multiple regression](https://en.wikipedia.org/wiki/Multiple_regression) model are highly linearly related.
* Leave the model as is, despite multicollinearity. The presence of multicollinearity doesn't affect the efficiency of extrapolating the fitted model to new data provided that the predictor variables follow the same pattern of multicollinearity in the new data as in the data on which the regression model is based.
* principal component regression

# [Regression coefficients that flip sign after including other predictors](https://stats.stackexchange.com/questions/1580/regression-coefficients-that-flip-sign-after-including-other-predictors)

* You run a linear regression with four numeric predictors (IV1, ..., IV4)
* When only IV1 is included as a predictor the standardised beta is +.20
* When you also include IV2 to IV4 the sign of the standardised regression coefficient of IV1 flips to -.25 (i.e., it's become negative).

This gives rise to a few questions:

* With regards to terminology, do you call this a "suppressor effect"?
* What strategies would you use to explain and understand this effect?
* Do you have any examples of such effects in practice and how did you explain and understand these effects?

Multicollinearity is the usual suspect as JoFrhwld mentioned. Basically, if your variables are positively correlated, then the coefficients will be negatively correlated, which can lead to a wrong sign on one of the coefficients.

One check would be to perform a principal components regression or ridge regression. This reduces the dimensionality of the regression space, handling the multicollinearity. You end up with biased estimates but a possibly lower MSE and corrected signs. Whether you go with those particular results or not, it's a good diagnostic check. If you still get sign changes, it may be theoretically interesting.

I believe effects like these are frequently caused by collinearity (see [this question](https://stats.stackexchange.com/questions/1149/is-there-an-intuitive-explanation-why-multicollinearity-is-a-problem-in-linear-re)). I think the book on multilevel modeling by Gelman and Hill talks about it. The problem is that IV1 is correlated with one or more of the other predictors, and when they are all included in the model, their estimation becomes erratic.

If the coefficient flipping is due to collinearity, then it's not really interesting to report, because it's not due to the relationship between your predictors to the outcome, but really due to the relationship between predictors.

What I've seen suggested to resolve this problem is residualization. First, you fit a model for IV2 ~ IV1, then take the residuals of that model as rIV2. If all of your variables are correlated, you should really residualize all of them. You may choose do to so like this

**You have built a multiple regression model. Your model R² isn’t as good as you wanted. For improvement, your remove the intercept term, your model R² becomes 0.8 from 0.3. Is it possible? How?**

**Answer:** Yes, it is possible. We need to understand the significance of intercept term in a regression model. The intercept term shows model prediction without any independent variable i.e. mean prediction. The formula of R² = 1 – ∑(y – y´)²/∑(y – ymean)² where y´ is predicted value.

When intercept term is present, R² value evaluates your model wrt. to the mean model. In absence of intercept term (ymean), the model can make no such evaluation, with large denominator, ∑(y - y´)²/∑(y)² equation’s value becomes smaller than actual, resulting in higher R².