Agenda for Week 7, Hour 2

Multiple regression: co-linearity, perturbations, correlation matrix

Consider the dataset 'angina.csv',

A dataset of made-up basic medical data from 32 adult white males.

It includes body mass index (BMI and QUET), age (AGE, AGESQ), systolic blood pressure (SBP) , and smoking status (SMK).

It’s taken from Page 77 of the textbook, and we’ll be using it several times in the next few weeks.

We're interested in modeling blood pressure as a function of other factors, perhaps to find the risk factors. We'll start with body-mass index and age. A healthy SBP is considered 120 or less.

```
mod1 = lm(SBP ~ BMI + AGE, data=angina)
summary(mod1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  55.3554    12.5466   4.412 0.000129 ***
BMI           1.3659     0.7639   1.788 0.084237 .
AGE           1.0461     0.3881   2.696 0.011571 *
```

The regression equation is

SBP = 55.35 + 1.36*BMI + 1.05*AGE + error ,

which means...

Someone with ***0 BMI and 0 AGE*** has blood pressure of 55.35

For each 1 BMI increase, we except blood pressure to increase by 1.36. (incomplete)

For each 1 year of AGE increase, we except blood pressure to increase by 1.05. (incomplete)

“Someone with 0 BMI and 0 AGE has blood pressure of 55.35”

Does the interpretation of the intercept make real-world sense? Not even close.

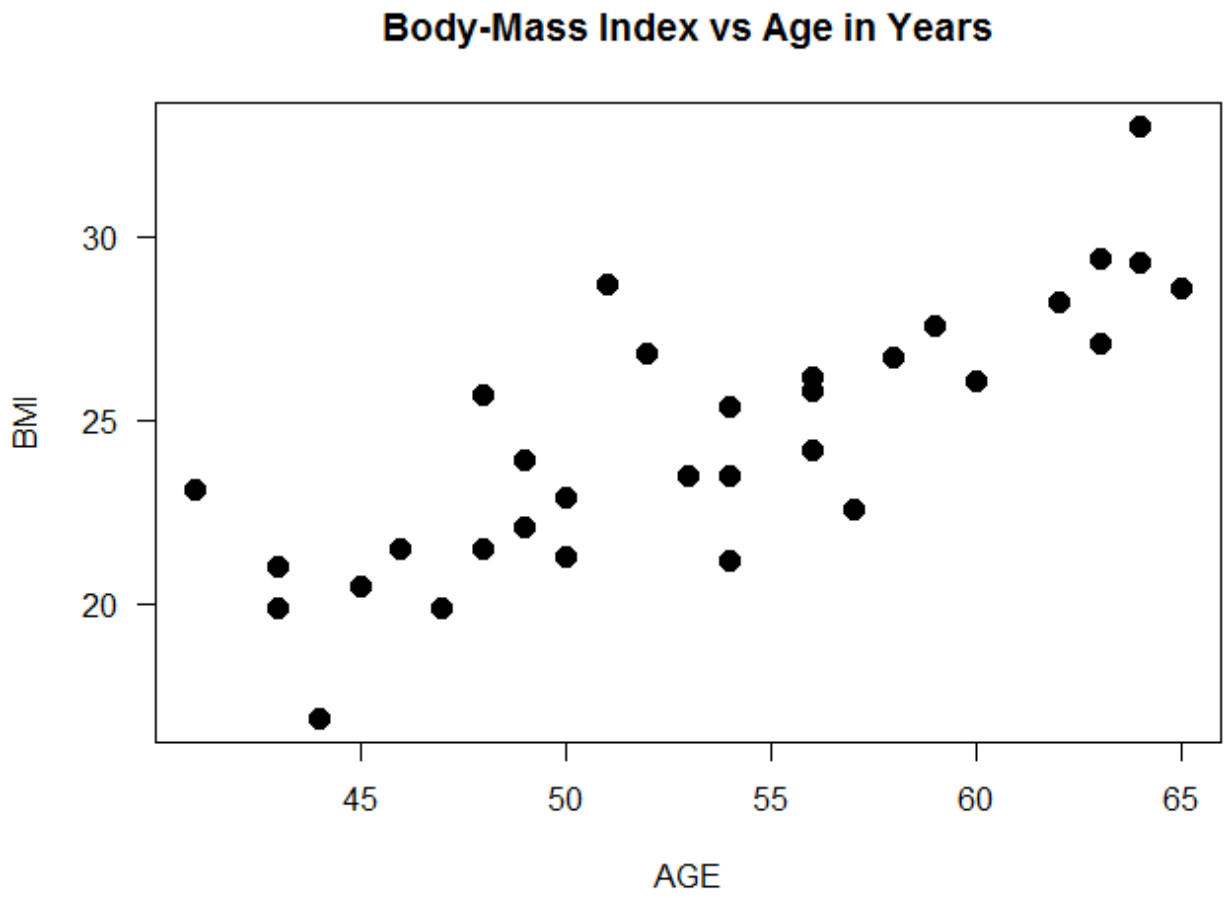
For someone to have 0 BMI, they would have to have literally no mass.

Also, the data for this model is based on adults, so someone with age 0 would be an extreme outlier and not belong in the data.

So no, the intercept doesn't make real-world sense, but that doesn't matter because 0 BMI, and 0 age is an extrapolation beyond the data.

Like in simple regression, the interpretation always makes mathematical sense, but not always real-world sense.

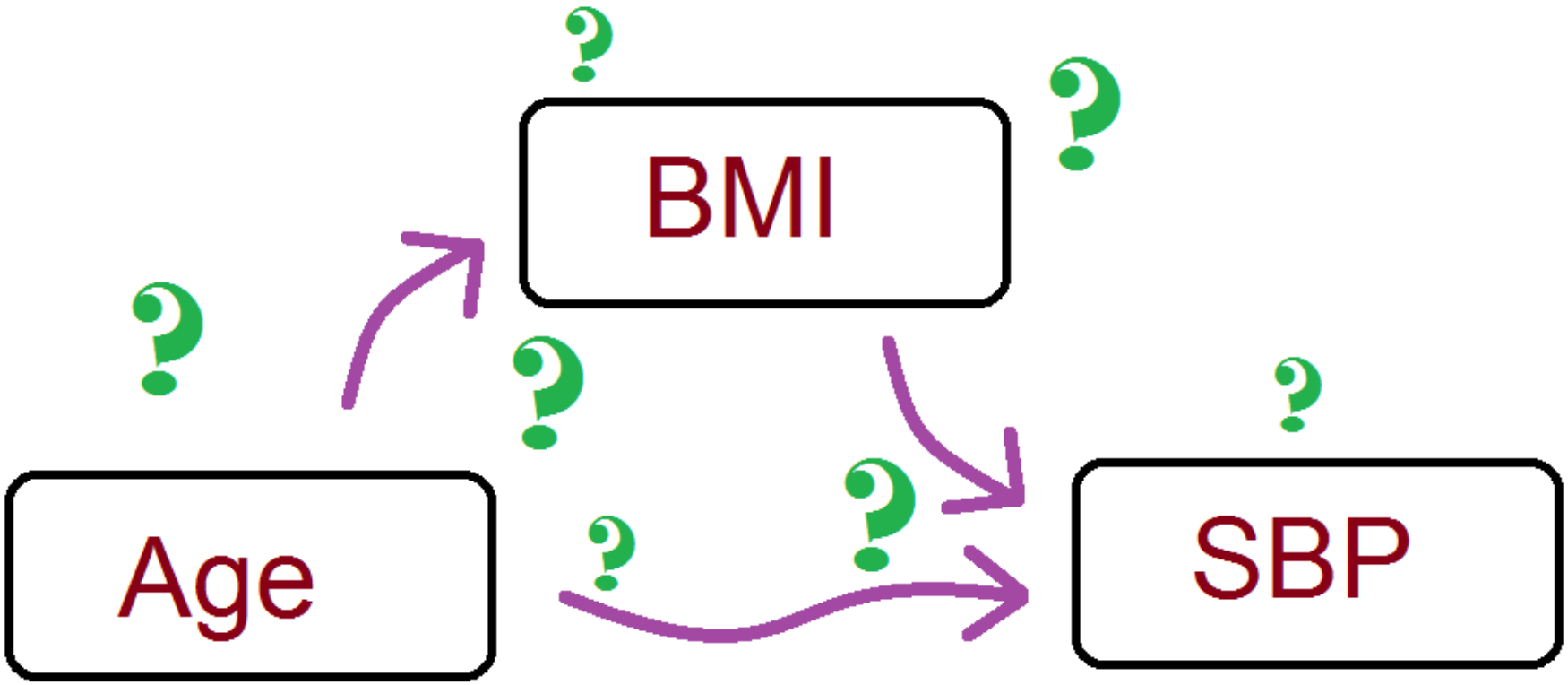
The slope coefficients for BMI and AGE seem straightforward, but there's one confounding issue,



BMI increases with AGE. In other words, men get fatter as they get older.

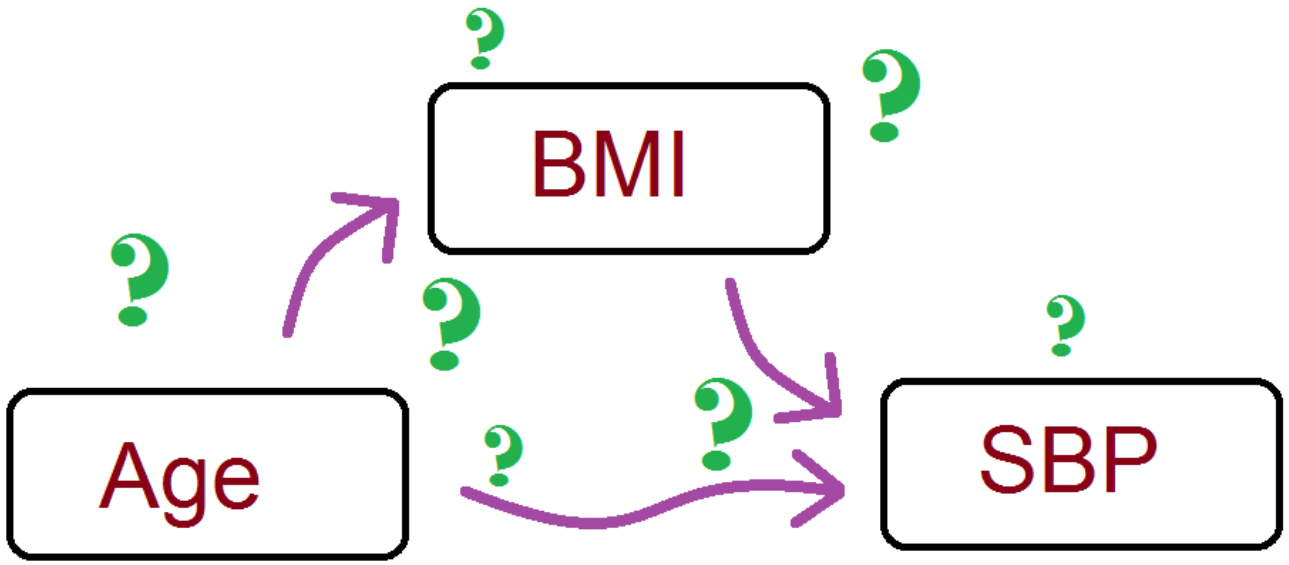
So when AGE increases by 1, blood pressure isn't the only variable that changes. BMI changes as well, and BMI has its own effect on SBP.

Is β_{AGE} , the 1.05 increase in SBP per year of AGE coming directly from age, or through BMI?



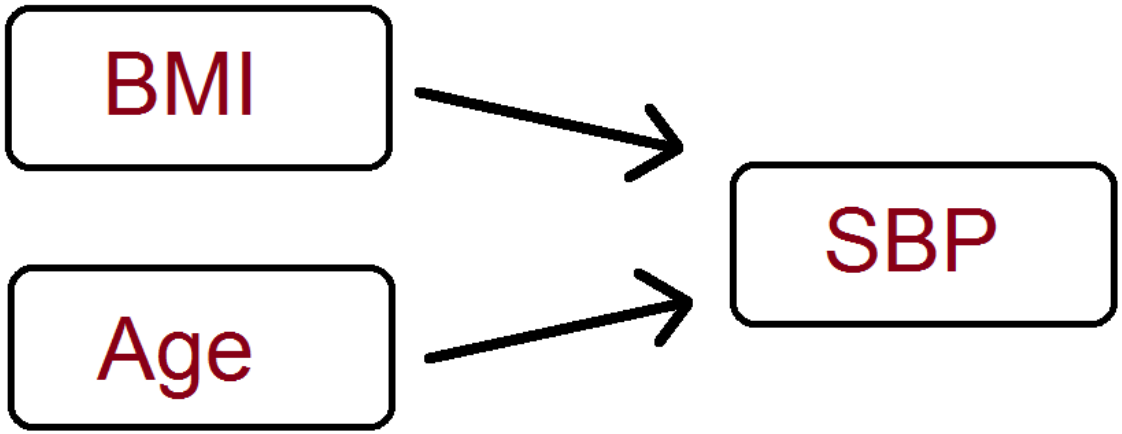
If both AGE and BMI increase by 1, would we expect blood pressure to increase by 1.05 + 1.36 ?

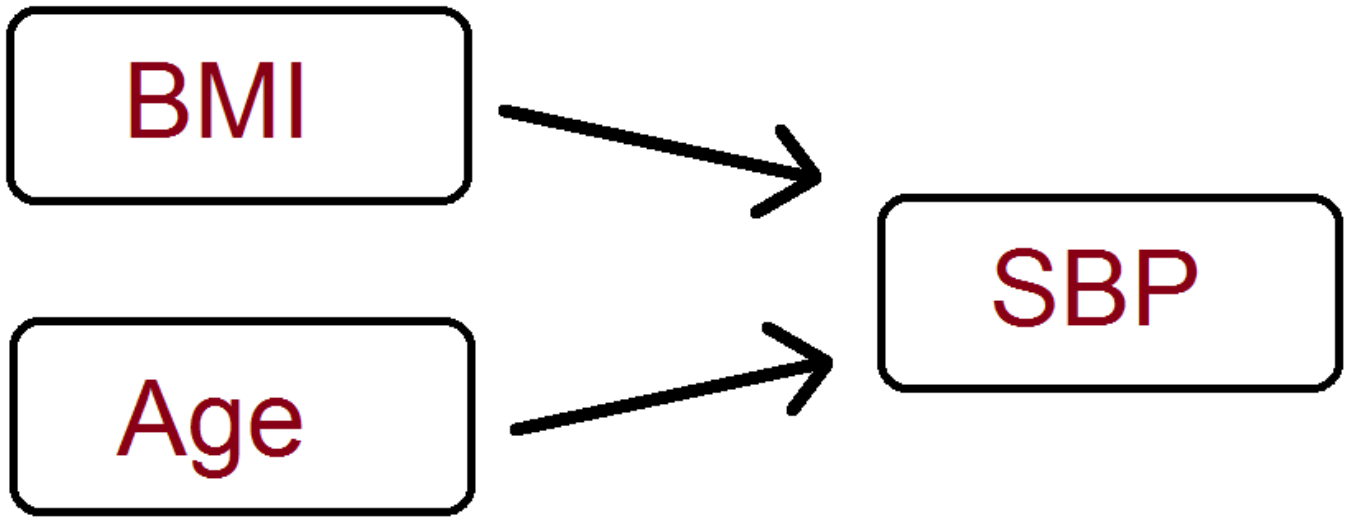
Would some of that increase in SBP already be 'covered' by the intervening effect from BMI?



This is why the previous interpretation is incomplete – we need to specify what's going with the other variables in the model when describing a slope coefficient.

Thankfully, the slope coefficients computed in a multiple regression describe the effect of each variable as a completely distinct phenomenon.





So a better interpretation of the slopes would be...

- For each 1 BMI increase, ***holding age constant***, we except blood pressure to increase by 1.36.

- For each 1 year of AGE increase, ***holding BMI constant***, we except blood pressure to increase by 1.05.

So even though there is it complicated relationship between Age, BMI, and SBP, the coefficients from a multiple regression describes these variables in terms of their separate effects on SBP.

A much more compact way to describe a slope coefficient, especially if there are many variables involved, would be like this:

- For each 1 BMI increase, ***all else being equal***, we except blood pressure to increase by 1.36.

Or...

- For each 1 BMI increase, ***controlling for other model variables***, we except blood pressure to increase by 1.36.

To answer a previous question:

Effects are ***additive***,

meaning that a change in both AGE and BMI together has the same effect as increasing each variable alone and adding the effects.

So increasing age by one year and BMI by one unit would indeed increase our expected blood pressure by 1.05 + 1.36.

Looking at the regression equal again, this should now be intuitive

$$SBP = 55.35 + 1.36*BMI + 1.05*AGE + error$$

Each effect is right there: Separate but additive.

But everything has a cost...

A multiple regression gives us estimates of the individual effects, but sometimes those effects are themselves correlated.

These effects are said to be ***co-linear***.

Consider the our angina model, SBP ~ AGE + BMI

The correlation matrix is show below

	BMI	AGE	SBP
BMI	1.000	0.805	0.742
AGE	0.805	1.000	0.775
SBP	0.742	0.775	1.000

	BMI	AGE	SBP
BMI	1.000	0.805	0.742
AGE	0.805	1.000	0.775
SBP	0.742	0.775	1.000

BMI and AGE have a strong correlation (r = 0.805), so when making a model that uses both of them, we have co-linearity.

Slope coefficients with co-linearity are sensitive to small changes in the data.

In our example, the regression equation

*SBP = 55.35 + 1.36*BMI + 1.05*AGE + error*

fits this data best, but perhaps another equation, like

SBP = ***40.86*** *+* ***2.14*****BMI +* ***0.99*****AGE + error*

would fit the data nearly as well.

Why does this happen? Because a multiple regression is built on how well it fits the data, but two very different looking models can fit the data in roughly the same way.

If we add more observations, or even do something as minor as round off another digit, the best fitting model could look a lot different.

In other terms, coefficients with co-linearity are ***sensitive to perturbations*** .

Let's compare the fitted values for the first five patients between the two models

Model A is the real model,

SBP = 55.35 + 1.36*BMI + 1.05*AGE + e

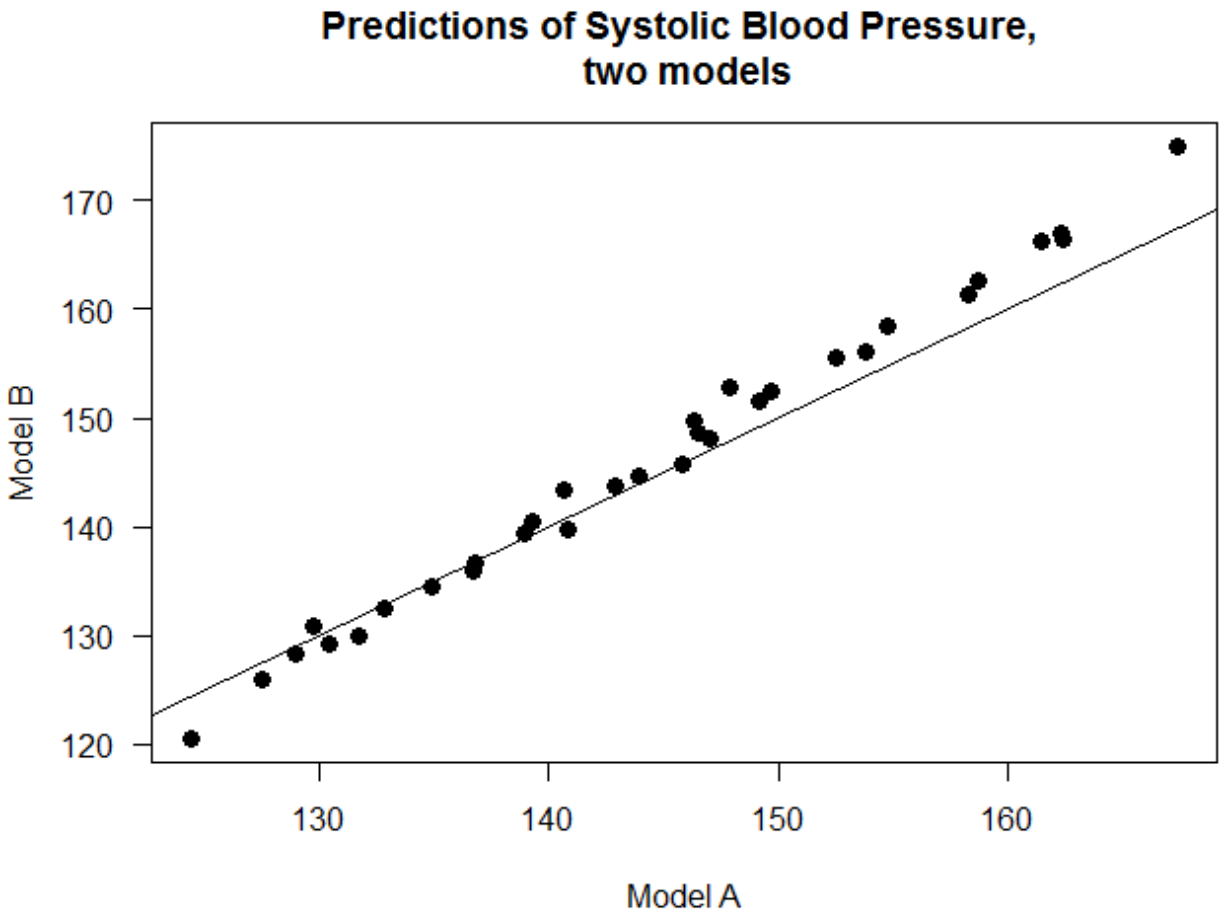
Model B is the 'perturbed' model,

SBP = **40.86** + **2.14***BMI + **0.99***AGE + e

Patient	BMI	Age	SBP, Model A	SBP, Model B
1	20.5	45	130.4	129.3
2	23.1	41	129.8	130.9
3	22.1	49	136.8	136.7
4	26.8	52	146.4	149.8
5	21.2	54	140.8	139.7

Here are all 32 fitted values in both models.

Points on the line are the same for both models.



Model A: SBP = 55.35 + 1.36*BMI + 1.05*AGE + e

Model B : SBP = **40.86** + **2.14***BMI + **0.99***AGE + e

Do these models 'look' the same? Not at all. One slope and the intercept are much different, the results are almost identical.

So co-linearity doesn't affect how well a model fits data, but it does make it hard to determine HOW the model is actually fitting that data.

Take-home message:

- Co-linearity is what happens when two explanatory variables are correlated with each other as well as the response variable.

- Co-linearity makes parameter estimates unstable and unreliable.

- Co-linearity does not affect estimates of predicted values.

One final note: Notice that the standard error for BMI gets worse when we include AGE.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  70.4612    12.3423   5.709 3.15e-06 ***
BMI           3.0229     0.4987   6.061 1.17e-06 ***

Multiple R-squared:  0.5505,
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  55.3554    12.5466   4.412 0.000129 ***
BMI           1.3659     0.7639   1.788 0.084237 .
AGE           1.0461     0.3881   2.696 0.011571 *

Multiple R-squared:  0.6406
```

But sometimes it all just works...

This sensitivity comes from the correlation between explanatory variables. It does NOT happen for every multiple regression model.

Consider the NHL teams data one more time.

Here is the correlation matrix for Wins, Goals For and Against.

	Wins	GFor	GAgainst
Wins	1.000	0.659	-0.650
GFor	0.659	1.000	-0.033
GAgainst	-0.650	-0.033	1.000

	Wins	GFor	GAgainst
Wins	1.000	0.659	-0.650
GFor	0.659	1.000	-0.033
GAgainst	-0.650	-0.033	1.000

Goals for and goals are almost uncorrelated.

In hockey terms, that means a team's offensive strength tells you nothing about their defensive strength.

In regression terms, this means there is no co-linearity between “Goals for” and “Goals against”.

Using the real model A, and a 'perturbed' model B, we get the same pattern of similar predictions.

Team	Goals For	Goals Against	Wins, Model A	Wins, Model B
Vancouver	249	198	49.7	50.1
NY Rangers	226	187	47.4	48.3
St Louis	21	165	48.2	49.6
Pittsburgh	282	221	51.7	51.3
Nashville	237	210	45.6	46.1

... but by applying the same kind of perturbation* as the angina data, only the intercept parameter changes much.

Real model A

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.30181    8.42725   5.376 1.11e-05 ***
GF           0.17601    0.02216   7.943 1.54e-08 ***
GA          -0.19835    0.03205  -6.190 1.29e-06 ***
```

Perturbed model B

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.51384    7.30960   4.859 4.45e-05 ***
GF           0.21878    0.02326   9.406 5.19e-10 ***
GA          -0.19371    0.01836 -10.550 4.45e-11 ***
```

*A 'resampling', but that's beyond this course.

Later we will talk in greater detail about the VIF, or ***variance inflation factor***.

The VIF is a measure of much the uncertainty of the parameter values is inflated by include a given co-linear variable.

Each explanatory variable has its own VIF.

Explanatory variables that have no co-linearity at all have a VIF of 1. Higher is worse.

Both “Goals for” and “Goals against” have a VIF of about 1.001

Both “BMI” and “AGE” have a VIF of about 2.78

Next time Polynomial Fits!

Code for making the first correlation matrix:

```
x = cor(cbind(angina$BMI, angina$AGE, angina$SBP))
colnames(x) = c("BMI","AGE","SBP")
rownames(x) = c("BMI","AGE","SBP")
round(x,3)
```