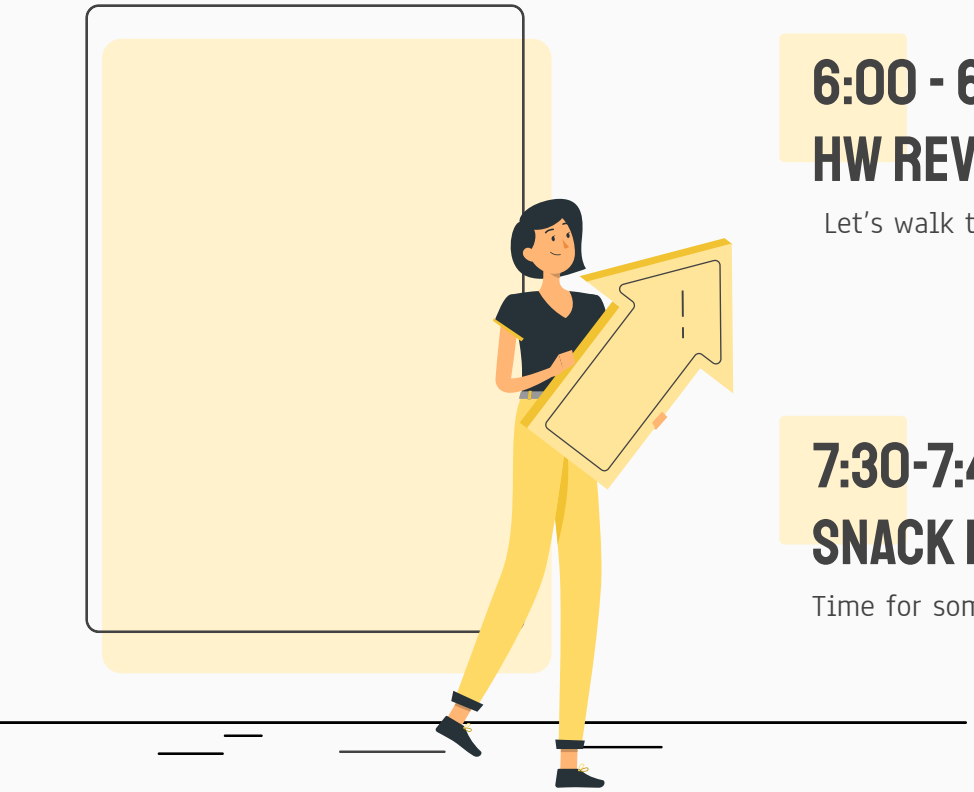
APPLIED DATA SCIENCE II

Week 3: EXTENDING LINEAR MODELS!

Kyle Scot Shank
WI-22



6:00 - 6:30
HW REVIEW

Let's walk through it!

6:30-7:30
TOPIC OVERVIEW

This week, we're talking about extending Linear models to be a little fancier!

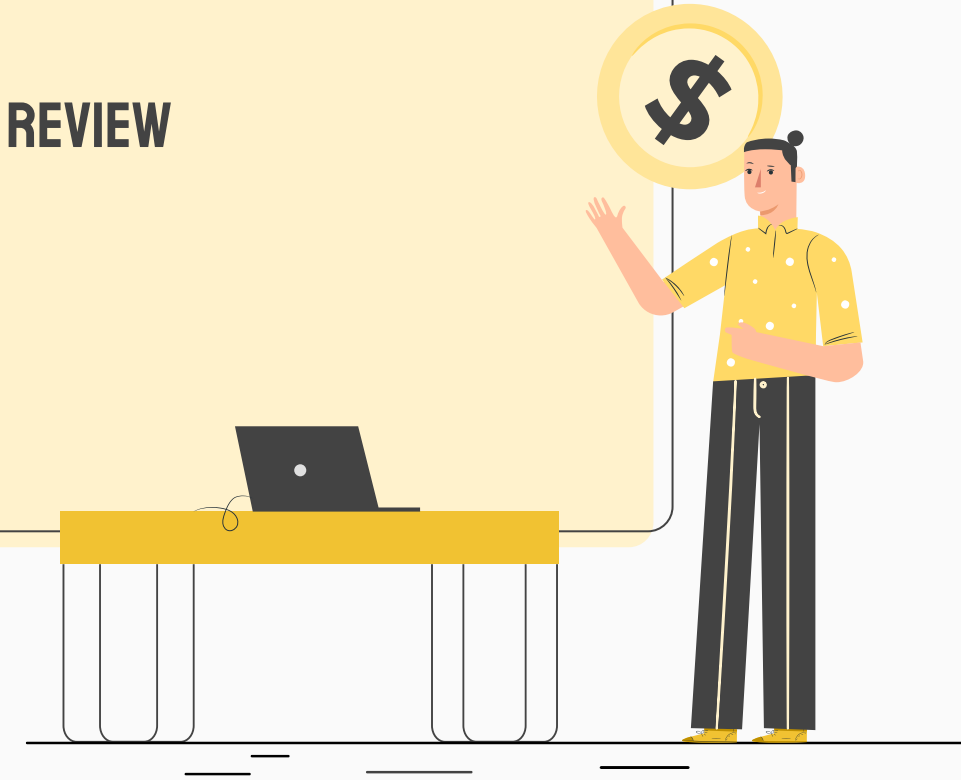
7:30-7:45
SNACK BREAK!

Time for some munchies

7:45 - 9:00
HANDS-ON CODE LAB

Work through stuff together

HW REVIEW



TOPIC OVERVIEW

EXTENDING LINEAR MODELS!

LET'S EXTEND LINEAR MODELS!

EXTENDING LINEAR REGRESSION

We remember that linear models take the following form:

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p + \varepsilon,$$

Despite its simplicity, the linear model has distinct advantages in terms of its interpretability and often shows good predictive performance.

Hence we discuss in this lecture some ways in which the simple linear model can be improved, by replacing ordinary least squares fitting with some alternative fitting procedures.

EXTENDING LINEAR REGRESSION

Why consider alternatives to ordinary least squares?

Prediction Accuracy: *especially when p > n, to control the variance.*

Model Interpretability*: By removing irrelevant features – that is, by setting the corresponding coefficient estimates to zero – we can obtain a model that is more easily interpreted. We will present some approaches for automatically performing feature selection.*

EXTENDING LINEAR REGRESSION

Let's talk about the three most common methods for feature selection!

Subset Selection: *We identify a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables*

Shrinkage: *We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as regularization) has the effect of reducing variance and can also perform variable selection.*

Dimension Reduction: *We project the p predictors into a M-dimensional subspace, where M < p. This is achieved by computing M different linear combinations, or projections, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares. (don't let this sound too scary!)*

BEST SUBSETS REGRESSION PROCESS!

This process (an algorithm!) follows some fairly simple rules:

1. Let M_0 denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For k = 1, 2, . . . p:

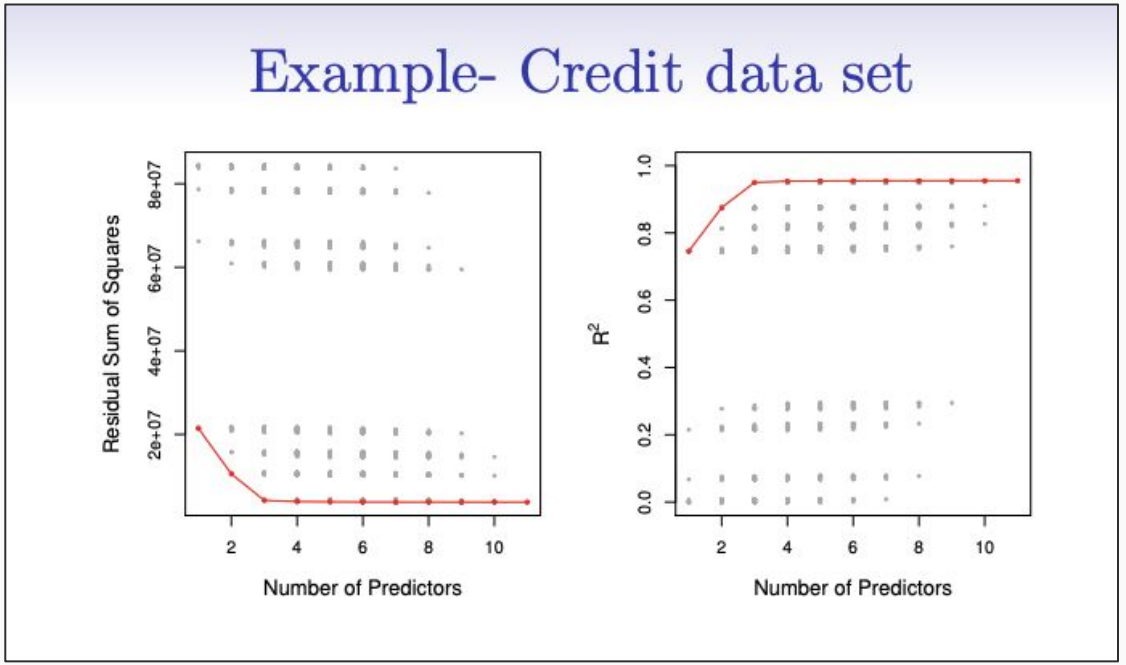
*(a) Fit all $\binom{p}{k}$ models that contain exactly **k** predictors.*

(b) Pick the best among these models, and call it M_k. Here best is defined as having the smallest RSS, or equivalently largest R2 .

3. Select a single best model from among $M_0, \ldots, M_p$ using cross-validated prediction error (we'll talk about more this more in two weeks!), AIC, BIC, or adjusted R2

BEST SUBSETS REGRESSION PROCESS!

For each possible model containing a subset of the ten predictors in the Credit data set, the RSS and R2 are displayed. The red frontier tracks the best model for a given number of predictors, according to RSS and R2 . Though the data set contains only ten predictors, the x-axis ranges from 1 to 11, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables



BEST SUBSETS REGRESSION PROCESS!

- *For computational reasons, best subset selection cannot be applied with very large p. **Why not**?*

- *Best subset selection may also suffer from statistical problems when p is large: larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.*

- *Thus an enormous search space can lead to overfitting and high variance of the coefficient estimates.*

- *For both of these reasons, stepwise methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection.*

STEPWISE SELECTION

- ***Forward stepwise selection*** *begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.*
- *In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model.*

- ***Backward stepwise selection*** *backward stepwise selection provides an efficient alternative to best subset selection.*
- *However, unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.*

STEPWISE SELECTION

Neither are guaranteed to find the globally "best" model for a subset of variables. Can you explain why?

SELECTING YOUR “BEST”

The model containing all of the predictors will always have the smallest RSS and the largest R2 , since these quantities are related to the training error.
We wish to choose a model with low ***test*** *error, not a model with low* ***training*** *error. Recall from the readings (esp. APM) that training error is usually a poor estimate of test error.*

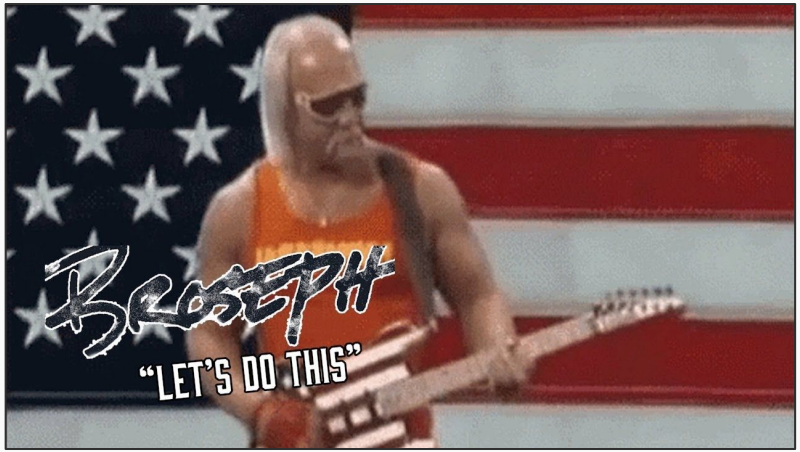
Therefore, RSS and R2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.

SELECTING YOUR "BEST"

We've got two options:

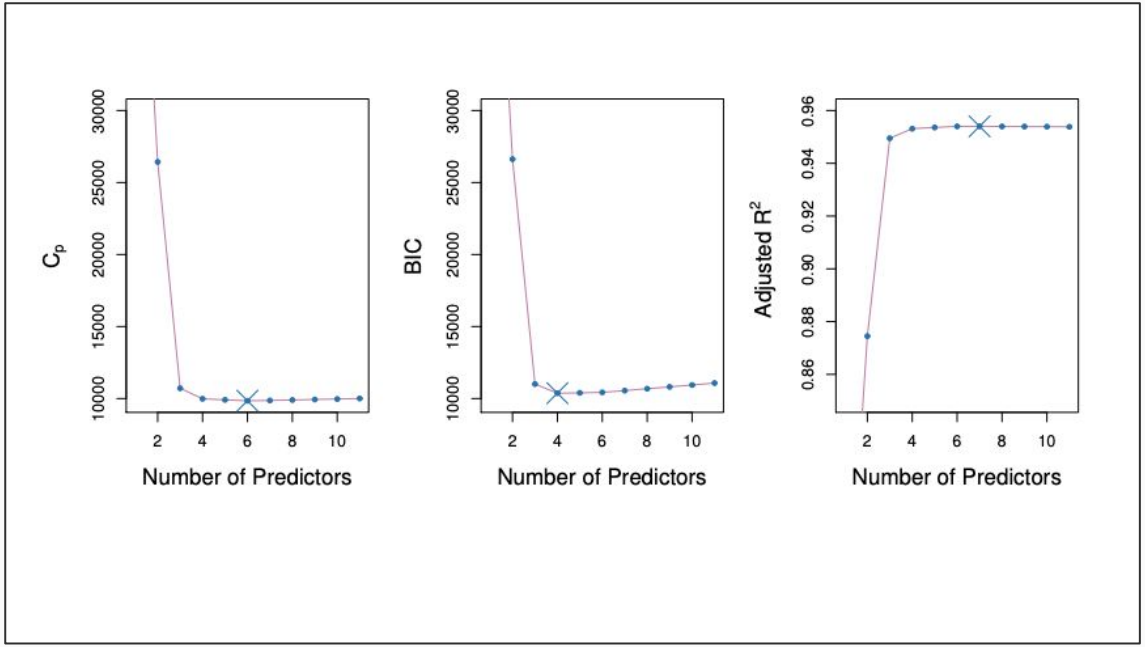
- *We can **indirectly** estimate test error by making an adjustment to the training error to account for the bias due to overfitting.*

- *We can **directly** estimate the test error, using either a validation set approach or a cross-validation approach (which we'll be discussing in week five!)*



SELECTING YOUR "BEST"

- *We'll first focus on indirect techniques. These techniques adjust the* ***training*** *error for the model size, and can be used to select among a set of models with different numbers of variables.*
- *The next figure displays C_p, BIC, and adjusted R^2 for the best model of each size p*



SELECTING YOUR "BEST"...METRICS!

Mallow's C_p:

$C_p = (1/n) * (RSS + 2d\sigma^2)$

where d is the total # of parameters used and σ^2 is an estimate of the variance of the error associated with each response measurement.

AIC:

$AIC = -2 \log L + 2 * d$

where d is the total # of parameters used and L is the maximized value of the likelihood function for the estimated model.

BIC:

$BIC = (1/n) (RSS + \log(n)d\sigma^2)$

Notice that this is very similar to C_p, but there is a log(n) term added!

*For all of these metrics, a **smaller** value indicates an expected lower test error.*

SELECTING YOUR "BEST"...METRICS!

Adjusted R^2

Adjusted R^2 *= 1 - (RSS/(n - d - 1)) / (TSS/(n - 1))*

where d is the total # of parameters used.

For this metric, unlike the previous 3 - a ***larger*** *value indicates an expected lower test error. Good thing to remember: a* ***standard R^2*** *will always go up if you include an additional variable, whereas the* ***adjusted R^2*** *may go up or down.*

SHRINKAGE METHODS

Now let's talk about shrinkage methods!

SHRINKAGE METHODS

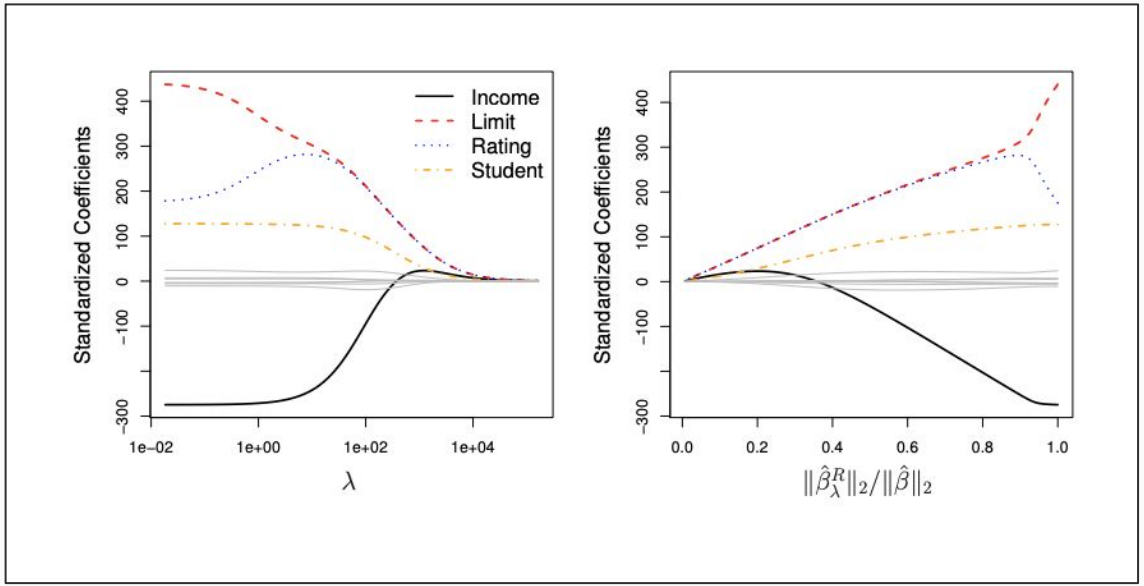
We're going to focus on two main shrinkage methods: ***ridge regression*** *and the* ***lasso***

- *The subset selection methods use least squares to fit a linear model that contains a subset of the predictors.*
- *As an alternative, we can fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.*
- *It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance.*

RIDGE REGRESSION

As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small.

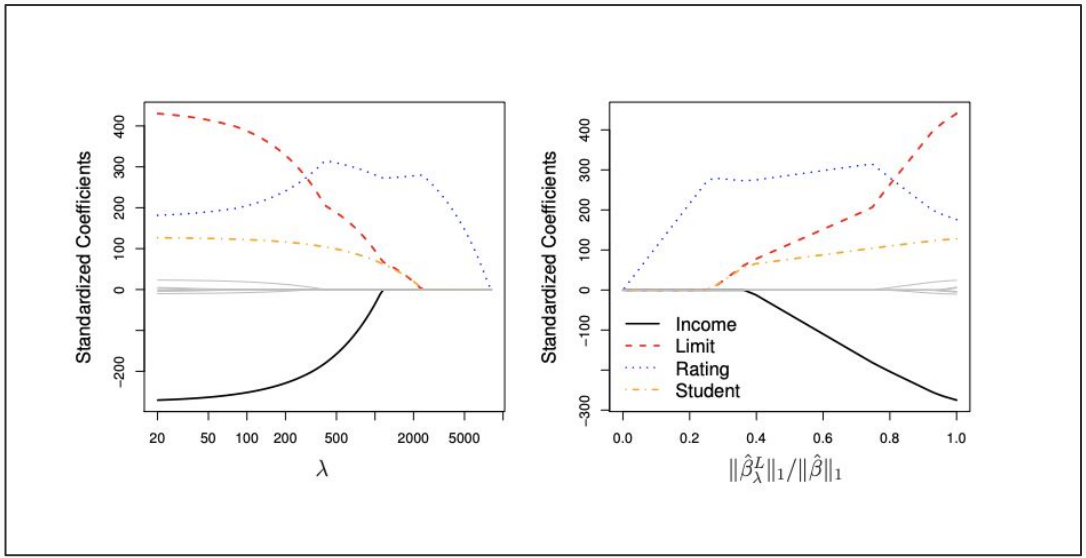
However, unlike standard least squares, ridge regression ads a the second term λ *to each estimated coefficient. This is called a shrinkage penalty. It is small when* $\beta_1, \ldots, \beta_p$ *are close to zero, and so it has the effect of shrinking the estimates of* β_j *towards zero.*



LASSO REGRESSION

Ridge regression does have one obvious disadvantage: unlike subset selection, ridge regression will include all p predictors in the final model and just shrink their coefficients to be really, really small.

The lasso model actually shrinks a coefficient all the way to zero, effectively allowing it to become a form of subset selection!



DIMENSION REDUCTION

The methods that we have discussed so far in this chapter have involved fitting linear regression models, via least squares or a shrunken approach, using the original predictors, $X_1, X_2, \ldots, X_p$.

We now explore a class of approaches that transform the predictors and then fit a least squares model using the transformed variables. We will refer to these techniques as ***dimension reduction methods.***

PRINCIPAL COMPONENTS REGRESSION (PCR)

*Here we apply **principal components analysis (PCA)** (discussed in Chapter 10 of the text if you're interested) to define the linear combinations of the predictors, for use in our regression.*

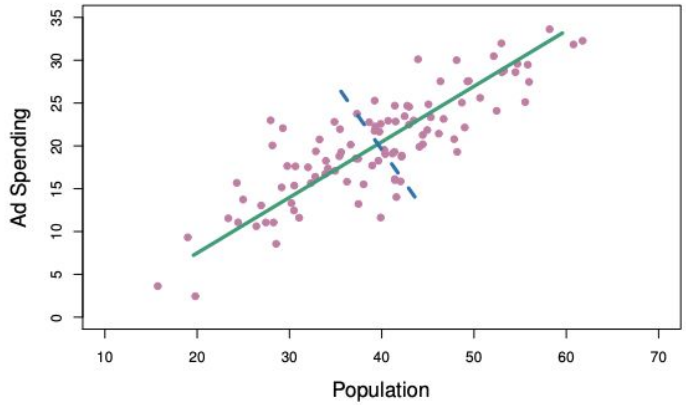
The first principal component is that (normalized) linear combination of the variables with the largest variance.

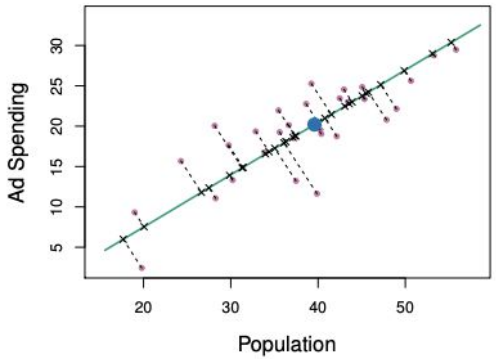
The second principal component has largest variance, subject to being uncorrelated with the first.

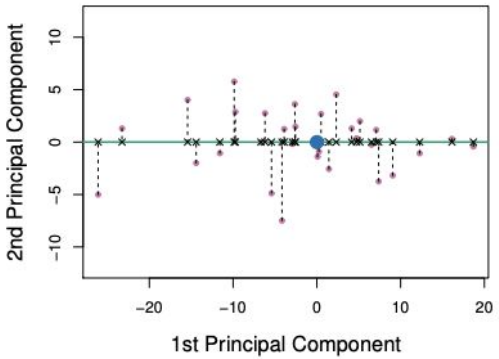
And so on.

Hence with many correlated original variables, we replace them with a small set of principal components that capture their joint variation.

PRINCIPAL COMPONENTS REGRESSION (PCR)







PCR is incredibly useful when dealing with highly correlated predictors, but suffers from a potentially serious drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

PARTIAL LEAST SQUARES REGRESSION (PLS)

Like PCR, PLS is a dimension reduction method, which first identifies a new set of features $Z_1, \ldots, Z_M$ that are linear combinations of the original features, and then fits a linear model via OLS using these M new features.

But unlike PCR, PLS identifies these new features in a supervised way – that is, it makes use of the response Y in order to identify new features that not only approximate the old features well, but also that are related to the response.

Roughly speaking, the PLS approach attempts to find directions that help explain ***both*** *the response and the predictors.*

RECAP + SUMMARY

Extending linear models:

- *There are times when you'll want to stretch beyond an ordinary least squares paradigm and into something more flexible to allow you to better model your data.*
- *Variable selection processes can be a great place to start, but aren't necessarily going to give you a global optimum (or might take a really, really long time to run if you've got a ton of data and variables).*
- *Shrinkage methods are* incredibly *popular (specifically lasso) because of how they introduce the idea of sparsity - which we'll talk about more later in the course with deep learning.*
- *Dimension reduction is a great way forward if you've got really, really correlated variables.*

SNACK BREAK!

COME BACK IN 15!

CODE LAB!

OPEN UP RSTUDIO