TEXAS McCombs

MASTER OF SCIENCE IN BUSINESS ANALYTICS

Dummy variables and model selection

Outline

Dummy variables

Model selection and regularization

Example: Detecting Sex Discrimination

Imagine you are a trial lawyer and you want to file a suit against a company for salary discrimination... you gather the following data...

	Gender	Salary
1	Male	32.0
2	Female	39.1
3	Female	33.2
4	Female	30.6
5	Male	29.0
...	...	...
208	Female	30.0

Detecting sex discrimination

You want to relate salary (Y) to gender (X)... how can we do that?

Gender is an example of a **categorical variable**. The gender variable separates our data into 2 groups or categories.

The question we want to answer is: *"how is your salary related to which group you belong to...?"*

Could we think about additional examples of categories potentially associated with salary?

- UT education vs. not
- foreign or domestic born citizen
- quarterback vs. wide receiver

Detecting sex discrimination

We can use regression to answer these questions, but first we need to recode the categorical variable into a **dummy variable**:

```
 Gender    Salary  Sex
1      Male  32.00   1
2    Female  39.10   0
3    Female  33.20   0
4    Female  30.60   0
5      Male  29.00   1
... ... ...
208  Female  30.00   0
```

Note:

This can be done implicitly in `R` by, `Sex = factor(Gender)`. This tells `R` that Sex is a variable separated into its unique levels, in this case just 2!

Detecting sex discrimination

Now you can present the following model in court:

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \epsilon_i$$

How do you interpret β_1? What is your predicted salary for males and females?

$$\begin{aligned} E[\text{Salary}|\text{Sex} = 0] &= \beta_0 \\ E[\text{Salary}|\text{Sex} = 1] &= \beta_0 + \beta_1 \end{aligned}$$

β_1 is the male-female difference!

Detecting sex discrimination

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \epsilon_i$$

```
data = read.table("SalaryData.txt",header=T)
Sex = (data$Gender=="Male")
data$Sex = Sex
fit = lm(Salary~Sex,data)
summary(fit)
```

```
##
## Call:
## lm(formula = Salary ~ Sex, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.805  -6.434  -1.860   4.115  51.495
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2099     0.8945  41.597  < 2e-16 ***
## SexTRUE       8.2955     1.5645   5.302 2.94e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\hat{\beta}_1 = b_1 = 8.29...$ on average, a male makes approximately \$8,300 more than a female in this firm.

We've seen this before!

What is the below code demonstrating?

```
data = read.table("SalaryData.txt",header=T)
Sex = (data$Gender=="Male")
data$Sex = Sex
fit = lm(Salary~Sex,data)
coef(fit)[2] # regression coefficient
```

```
##  SexTRUE
## 8.295513
```

```
DiM = mean(data$Salary[Sex==1]) - mean(data$Salary[Sex==0])
DiM # what is this?
```

```
## [1] 8.295513
```

Detecting sex discrimination

How can the defense attorney try to counteract the plaintiff's argument?

Perhaps, the observed difference in salaries is related to other variables in the background and NOT to gender discrimination...

Obviously, there are many other factors which we can legitimately use in determining salaries:

- education
- job productivity
- experience

How can we use regression to incorporate additional information?

Detecting sex discrimination

Let's add a measure of experience...

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \beta_2 \text{Exp}_i + \epsilon_i$$

What does that mean? Write out the model for each gender separately:

$$\begin{aligned} E[\text{Salary}|\text{Sex} = 0, \text{Exp}] &= \beta_0 + \beta_2 \text{Exp} \\ E[\text{Salary}|\text{Sex} = 1, \text{Exp}] &= (\beta_0 + \beta_1) + \beta_2 \text{Exp} \end{aligned}$$

Detecting sex discrimination

Here is our data with this additional variable:

	Exp	Gender	Salary	Sex
1	3	Male	32.00	1
2	14	Female	39.10	0
3	12	Female	33.20	0
4	8	Female	30.60	0
5	3	Male	29.00	1
...	...	...		
208	33	Female	30.00	0

Detecting sex discrimination

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Sex}_i + \beta_2 \text{Exp} + \epsilon_i$$

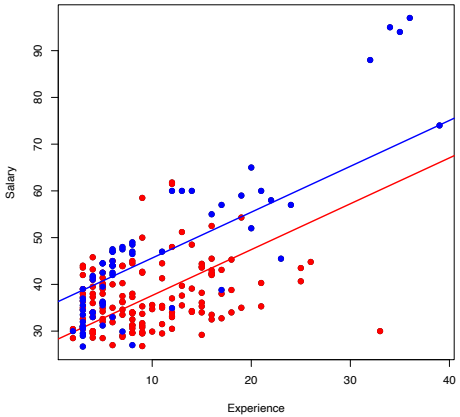
```
fit = lm(Salary~Sex+Exp,data)
summary(fit)
```

```
##
## Call:
## lm(formula = Salary ~ Sex + Exp, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.1899  -5.7484  -0.6046   4.8129  25.8548
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.81190    1.02789  27.057  < 2e-16 ***
## SexTRUE      8.01189    1.19309   6.715 1.81e-10 ***
## Exp          0.98115    0.08028  12.221  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\rightarrow$ $\text{Salary}_i = 27 + 8\text{Sex}_i + 0.98\text{Exp}_i + \epsilon_i$, ... **Is this good or bad news for the defense?**

Detecting sex discrimination

$$\text{Salary}_i = \begin{cases} 27 + 0.98\text{Exp}_i + \epsilon_i & \text{females} \\ 35 + 0.98\text{Exp}_i + \epsilon_i & \text{males} \end{cases}$$



More than two categories

We can use dummy variables in situations in which there are more than two categories. Dummy variables are needed for each category except one, designated as the "base" category.

Why? **Remember that the numerical value of each category has no quantitative meaning!**

House prices revisted

We want to evaluate the difference in house prices in a couple of different neighborhoods.

	Nbhd	SqFt	Price
1	2	1.79	114.3
2	2	2.03	114.2
3	2	1.74	114.8
4	2	1.98	94.7
5	2	2.13	119.8
6	1	1.78	114.6
7	3	1.83	151.6
8	3	2.16	150.7
...	...	...	...

House prices revisted

Let's create the dummy variables dn1, dn2 and dn3...

	Nbhd	SqFt	Price	dn1	dn2	dn3
1	2	1.79	114.3	0	1	0
2	2	2.03	114.2	0	1	0
3	2	1.74	114.8	0	1	0
4	2	1.98	94.7	0	1	0
5	2	2.13	119.8	0	1	0
6	1	1.78	114.6	1	0	0
7	3	1.83	151.6	0	0	1
8	3	2.16	150.7	0	0	1
...	...	...				

House prices revisted

$$\text{Price}_i = \beta_0 + \beta_1 \text{dn1}_i + \beta_2 \text{dn2}_i + \beta_3 \text{SqFt}_i + \epsilon_i$$

$$\begin{aligned}
E[\text{Price}|\text{dn1} = 1, \text{SqFt}] &= \beta_0 + \beta_1 + \beta_3 \text{SqFt} \quad \text{(Nbhd 1)} \\
E[\text{Price}|\text{dn2} = 1, \text{SqFt}] &= \beta_0 + \beta_2 + \beta_3 \text{SqFt} \quad \text{(Nbhd 2)} \\
E[\text{Price}|\text{dn1} = 0, \text{dn2} = 0, \text{SqFt}] &= \beta_0 + \beta_3 \text{SqFt} \quad \text{(Nbhd 3)}
\end{aligned}$$

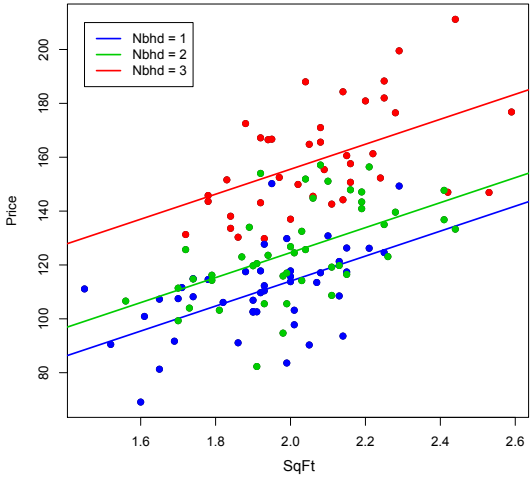
Model output

```
fit = lm(Price~dn1+dn2+SqFt,data)
summary(fit)
```

```
##
## Call:
## lm(formula = Price ~ dn1 + dn2 + SqFt, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.107 -10.924  -0.305   9.643  38.506
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.776     14.248   4.406 2.25e-05 ***
## dn1          -41.535      3.534 -11.754  < 2e-16 ***
## dn2          -30.967      3.369  -9.192 1.13e-15 ***
## SqFt          46.386      6.746   6.876 2.67e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.26 on 124 degrees of freedom
## Multiple R-squared:  0.6851, Adjusted R-squared:  0.6774
## F-statistic: 89.91 on 3 and 124 DF,  p-value: < 2.2e-16
```

$$\text{Price} = 62.78 - 41.54 * \text{dn1} - 30.97 * \text{dn2} + 46.39 * \text{SqFt} + \epsilon$$

What do these models look like?



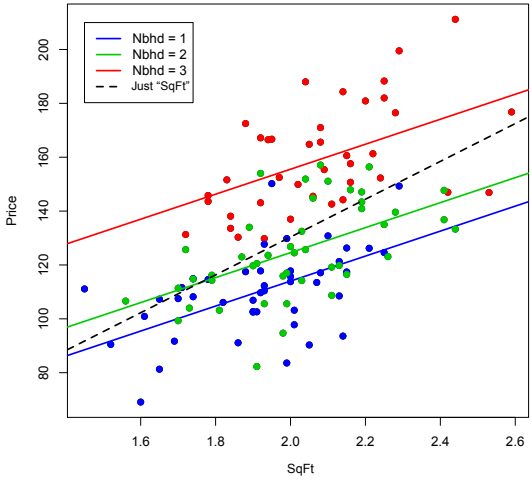
Model output only with “SqFt”

```
fit = lm(Price~SqFt,data)
summary(fit)
```

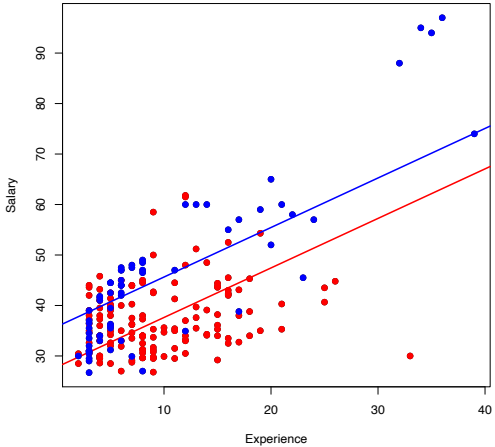
```
##
## Call:
## lm(formula = Price ~ SqFt, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46.59 -16.64  -1.61  15.12  54.83
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -10.091     18.966  -0.532    0.596
## SqFt          70.226      9.426   7.450  1.3e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.48 on 126 degrees of freedom
## Multiple R-squared:  0.3058, Adjusted R-squared:  0.3003
## F-statistic:  55.5 on 1 and 126 DF,  p-value: 1.302e-11
```

$$\text{Price} = -10.09 + 70.23 * \text{SqFt} + \epsilon$$

What do these models look like?



Back to the sex discrimination case



Does it look like the effect of experience on salary is the same for males and females?

Back to the sex discrimination case

Could we try to expand our analysis by allowing a different slope for each group?

Yes... Consider the following model:

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Exp}_i + \beta_2 \text{Sex}_i + \beta_3 \text{Exp}_i \times \text{Sex}_i + \epsilon_i$$

For Females:

$$\text{Salary}_i = \beta_0 + \beta_1 \text{Exp}_i + \epsilon_i$$

For Males:

$$\text{Salary}_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\text{Exp}_i + \epsilon_i$$

How are these models different from each other?

Sex discrimination case

We are just creating a new variable!

	Exp	Gender	Salary	Sex	Exp*Sex
1	3	Male	32.00	1	3
2	14	Female	39.10	0	0
3	12	Female	33.20	0	0
4	8	Female	30.60	0	0
5	3	Male	29.00	1	3
...	...	...			
208	33	Female	30.00	0	0

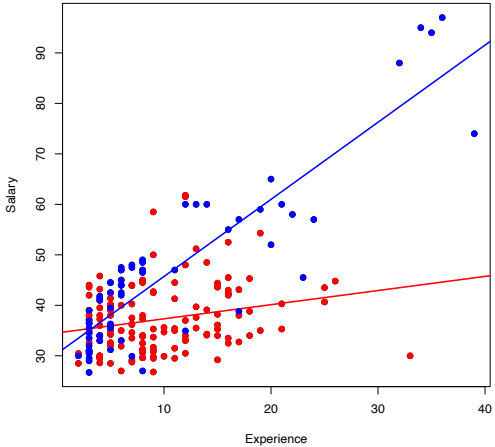
Sex discrimination case

```
fit = lm(Salary~Sex+Exp+ExpSex,data)
summary(fit)
```

```
##
## Call:
## lm(formula = Salary ~ Sex + Exp + ExpSex, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.0685  -4.6506  -0.7679   4.4034  23.9122
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.5283     1.1380  30.342  < 2e-16 ***
## SexTRUE      -4.0983     1.6658  -2.460  0.01472 *
## Exp           0.2800     0.1025   2.733  0.00684 **
## ExpSex        1.2478     0.1367   9.130  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.816 on 204 degrees of freedom
## Multiple R-squared:  0.6386, Adjusted R-squared:  0.6333
## F-statistic: 120.2 on 3 and 204 DF,  p-value: < 2.2e-16
```

$$\text{Salary} = 34 - 4 * \text{Sex} + 0.28 * \text{Exp} + 1.24 * \text{Exp}*\text{Sex} + \epsilon$$

Sex discrimination case



Is this good or bad news for the plaintiff?

Variable interaction

The effect of experience on salary is different for males and females... in general, when the effect of the variable X_1 onto Y depends on another variable X_2 we say that X_1 and X_2 **interact** with each other.

We can extend this notion by the inclusion of multiplicative effects through interaction terms.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) + \epsilon$$

$$\frac{\partial E[Y|X_1, X_2]}{\partial X_1} = \beta_1 + \beta_3 X_2$$

Variable selection and regularization

When working with linear regression models where the number of X variables is large, we need to think about strategies to select what variables to use...

We will focus on 2 ideas:

- Subset Selection
- Shrinkage

Subset selection

The idea here is very simple: fit as many models as you can and compare their performance based on some criteria!

Issues:

- How many possible models? Total number of models = 2^p
  Is this large?
- What criteria to use?
  Just as before, if prediction is what we have in mind, out-of-sample predictive ability should be the criteria

Information criteria

Another way to evaluate a model is to use Information Criteria metrics which attempt to quantify how well our model would have predicted the data (regardless of what you've estimated for the β_j's).

A good alternative is the BIC: Bayes Information Criterion, which is based on a "Bayesian" philosophy of statistics.

$$BIC = n\log(s^2) + p\log(n)$$

You want to choose the model that leads to minimum BIC.

Information criteria

One nice thing about the BIC is that you can interpret it in terms of model probabilities.

Given a list of possible models $\{M_1, M_2, \ldots, M_R\}$, the probability that model i is correct is

$$P(M_i) \approx \frac{e^{-\frac{1}{2}BIC(M_i)}}{\sum_{r=1}^{R} e^{-\frac{1}{2}BIC(M_r)}} = \frac{e^{-\frac{1}{2}[BIC(M_i)-BIC_{min}]}}{\sum_{r=1}^{R} e^{-\frac{1}{2}[BIC(M_r)-BIC_{min}]}}$$

(Subtract $BIC_{min} = \min\{BIC(M_1) \ldots BIC(M_R)\}$ for numerical stability.)

Similar, alternative criteria include AIC, C_p, adjusted R^2...

Bottom line: these are only useful if we lack the ability to compare models based on their out-of-sample predictive ability!

Search strategies: Stepwise regression

One computational approach to build a regression model step-by-step is “stepwise regression” There are 3 options:

- Forward: adds one variable at the time until no remaining variable makes a significant contribution (or meet a certain criteria... could be out of sample prediction)
- Backwards: starts will all possible variables and removes one at the time until further deletions would do more harm them good
- Stepwise: just like the forward procedure but allows for deletions at each step

Shrinkage methods

An alternative way to deal with selection is to **work with all p predictors at once** while placing a constraint on the size of the estimated coefficients

This idea is a regularization technique that reduces the variability of the estimates and tend to lead to better predictions.

The hope is that by having the constraint in place, the estimation procedure will be able to focus on "the important β's"

Ridge regression

Ridge regression is a modification of the least squares criteria that minimizes (as a function of β's)

$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

for some value of $\lambda > 0$

- The "blue" part of the equation is the traditional objective function of LS
- The "red" part is the shrinkage penalty, ie, something that makes costly to have big values for β

Ridge regression

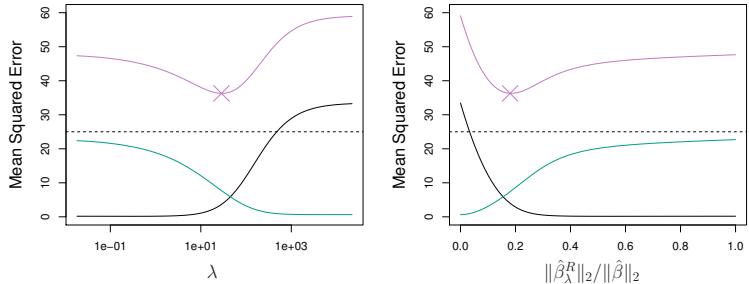
$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

- if $\lambda = 0$ we are back to least squares
- when $\lambda \to \infty$, it is "too expensive" to allow for any β to be different than 0...
- So, for different values of λ we get a different solution to the problem

Ridge regression

- What ridge regression is doing is exploring the bias-variance trade-off! The larger the λ the more bias (towards zero) is being introduced in the solution, ie, the less flexible the model becomes... at the same time, the solution has less variance
- As always, the trick to find the "right" value of λ that makes the model not too simple but not too complex!
- Whenever possible, we will choose λ by comparing the out-of-sample performance (usually via cross-validation)

Ridge regression



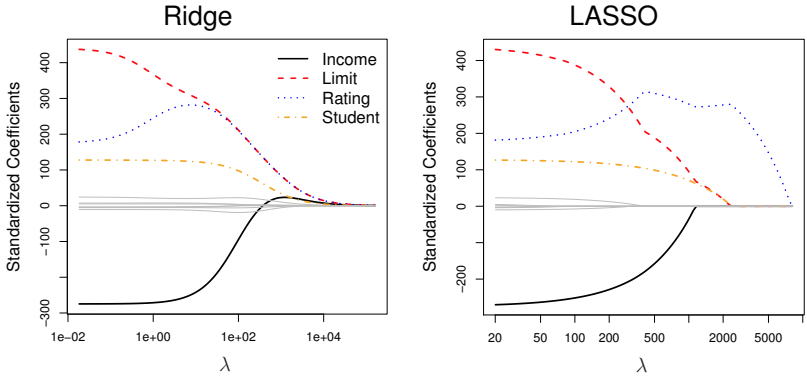
bias^2 (black), *variance* (green), test MSE (purple)

$\rightarrow$ Ridge is computationally very attractive as the "computing cost" is almost the same of least squares (contrast that with subset selection!)

$\rightarrow$ It's a good practice to always center and scale the X's

LASSO

The LASSO is a shrinkage method that performs automatic selection. It is similar to ridge but it will provide solutions that are sparse, ie, some β's exactly equal to 0! This facilitates interpretation of the results...



LASSO

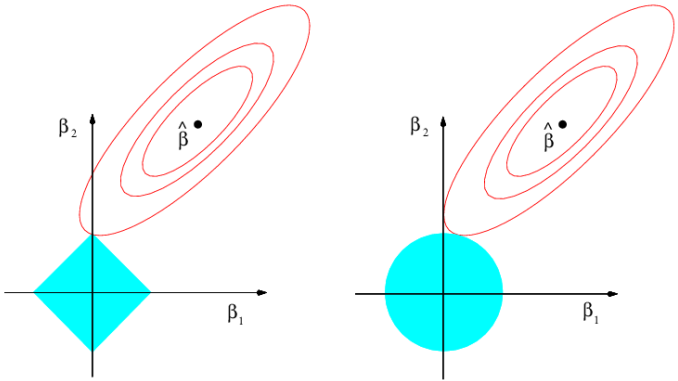
The LASSO solves the following problem:

$$\arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Once again, λ controls how flexible the model gets to be
- Still a very efficient computational strategy
- Whenever possible, we will choose λ by comparing the out-of-sample performance (usually via cross-validation)

Ridge vs. LASSO

Why does the LASSO outputs zeros?



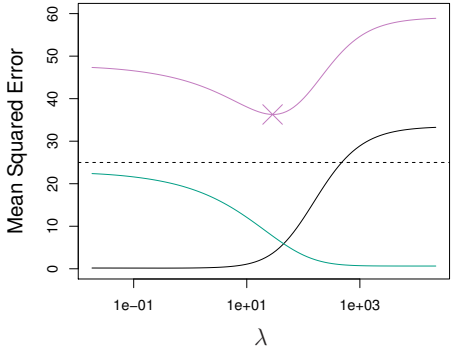
Ridge vs. LASSO

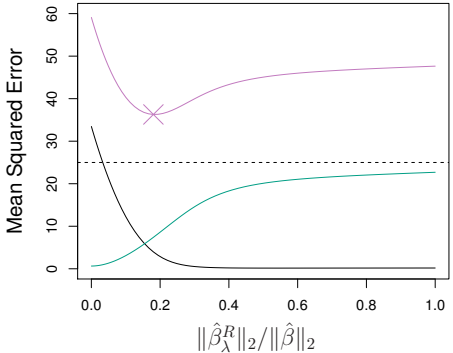
Which one is better?

- **It depends**...
- In general LASSO will perform better than Ridge when a relative small number of predictors have a strong effect in Y while Ridge will do better when Y is a function of many of the *X*'s and the coefficients are of moderate size
- LASSO can be easier to interpret (the zeros help!)
- But, if prediction is what we care about the only way to decide which method is better is comparing their out-of-sample performance

Choosing λ

The idea is to solve the ridge or LASSO objective function over a grid of possible values for λ...





How to do this in R

Check out the package `glmnet`. You can easily do Ridge, LASSO, and much more!