Lecture 13: Model Selection

So far:

- Simple Linear Regression
- Multiple linear regression

Today:

- Collinearity
- Residual Plots and Log Plots
- Over-fitting Problem and Lasso Regression

Collinearity

- Collinearity: if prediction variables are **highly correlated** (either positively or negatively)
  - For collinear predictors, hard to separate and interpret individual effects

- **Example**: Estimating the price of a house (in Lecture 11-12)

We want to be able to predict the selling price of houses using values we can observe when we talk to the seller

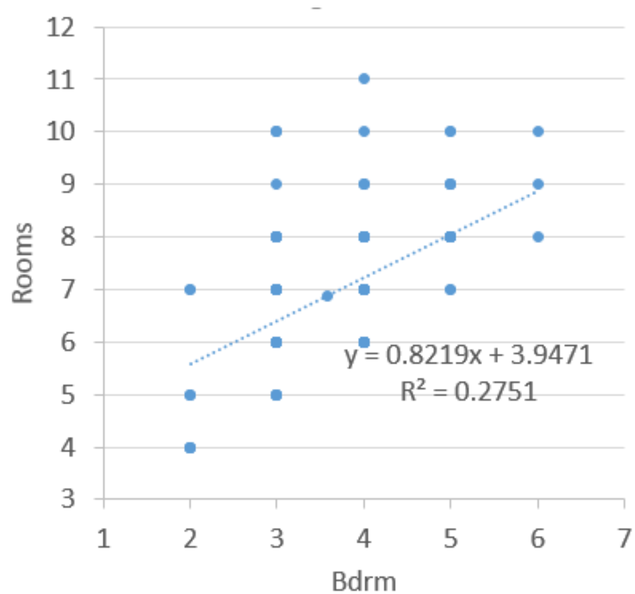
$$\text{price} = \beta_0 + \beta_{\text{ls}}(\text{lot size}) + \beta_{\text{bedr}}(\#\ \text{bedr}) + \cdots$$

$$\cdots + \beta_{\text{grg}}(\#\ \text{garage}) + \beta_{\text{location}} \quad + \quad \textit{error}$$

Collinearity

- The correlation matrix

	price	lotsz	bdrm	bath	rooms		
price	1						
lotsz	–.059	1					
bdrm	0.209	0.150	1				
bath	0.540	0.109	0.275	1			
rooms	0.481	0.171	0.525	0.466	1		
garg	0.233	0.127	0.048	0.260	0.236		
age	–.252	–.185	–.142	–.161	–.152	–.052	1



Regression Output with All Variables

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.841
R Square	0.708
Adj R Square	0.697
Standard Error	20.591
Observations	228

ANOVA

	df	*SS*	*MS*	*F*	*Signif F*
Regression	8	224753	28094	66.264	0.000
Residual	219	92850	424		
Total	227	317602			

	Coef	*Std Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	78.749	10.527	7.480	0.000	58.001	99.497
lot size	0.680	0.371	1.833	0.068	-0.051	1.411
bedrooms	-3.691	2.226	-1.658	0.099	-8.078	0.696
baths	19.045	2.804	6.791	0.000	13.518	24.572
rooms	8.492	1.493	5.689	0.000	5.550	11.433
age	-0.351	0.120	-2.920	0.004	-0.588	-0.114
garages	3.938	2.338	1.684	0.094	-0.670	8.547
e meadow	57.135	3.976	14.371	0.000	49.299	64.970
lvttwn	24.473	3.891	6.289	0.000	16.804	32.142

Regression Output without variable: rooms

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.815134013
R Square	0.66444346
Adj R Squa	0.653766661
Standard E	22.00966662
Observation	228

ANOVA

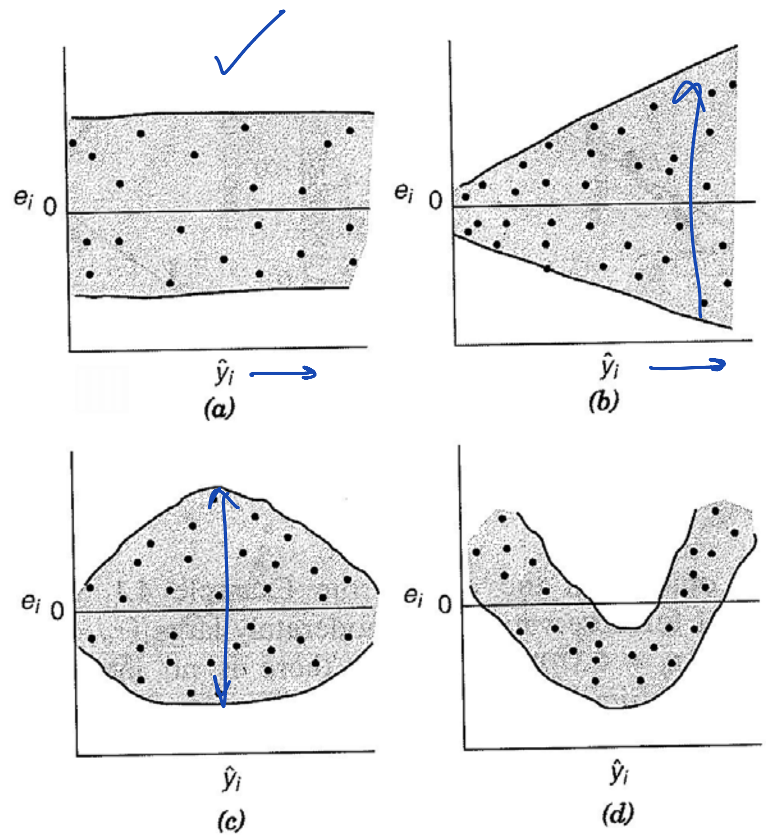
	df	SS	MS	F	Signif F
Regression	7	211029	30147	62.232	0.000
Residual	220	106574	484		
Total	227	317602			

	Coef	Std Err	t Stat	P-value	Lower 95%	Upper 95%
Intercept	105.384	10.079	10.456	0.000	85.521	125.248
lot size	0.851	0.395	2.154	0.032	0.072	1.630
bedrooms	1.985	2.127	0.933	0.352	-2.207	6.177
baths	24.227	2.835	8.545	0.000	18.640	29.815
age	-0.365	0.128	-2.842	0.005	-0.618	-0.112
garages	5.899	2.472	2.386	0.018	1.027	10.771
e meadow	58.824	4.238	13.881	0.000	50.472	67.176
lvttwn	24.657	4.159	5.928	0.000	16.460	32.854

Homoscedasticity?

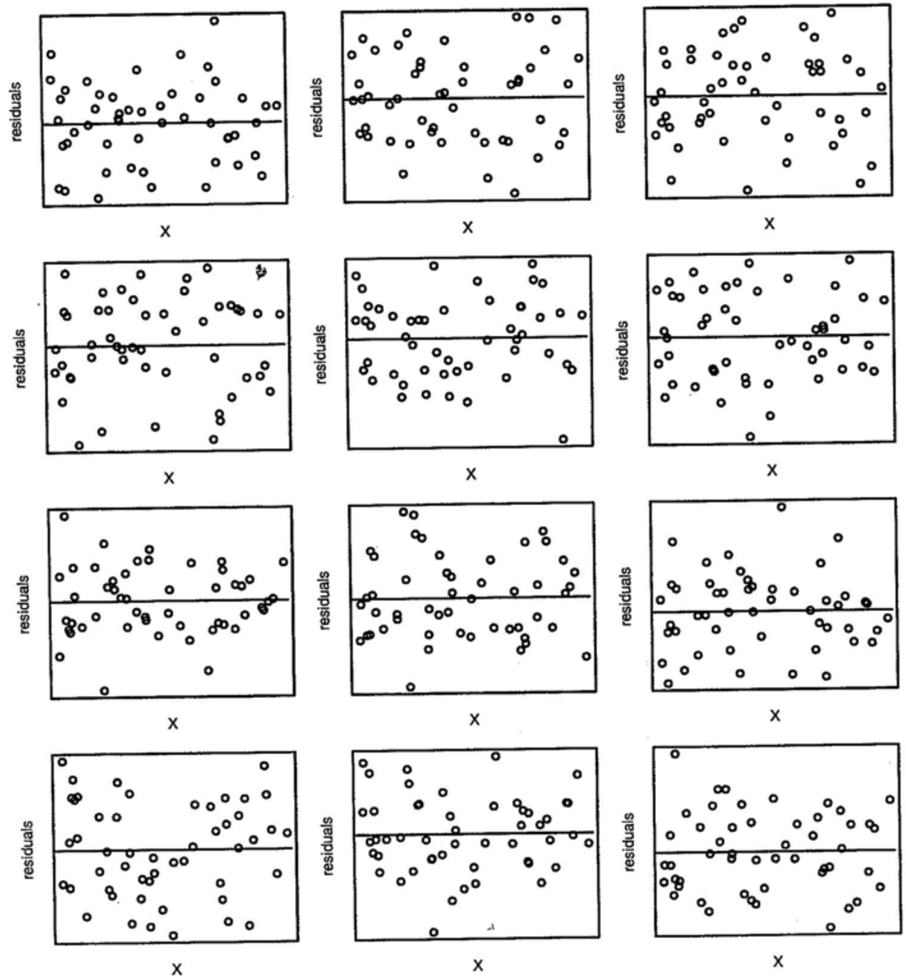
For each value in the X axis, dots in the residual plots should be "random" and have same spread above and below zero

- Reveal how well the linear equation explains the data
- (a) indicates homoscedasticity
- (b), (c) & (d) do not because one can see a pattern

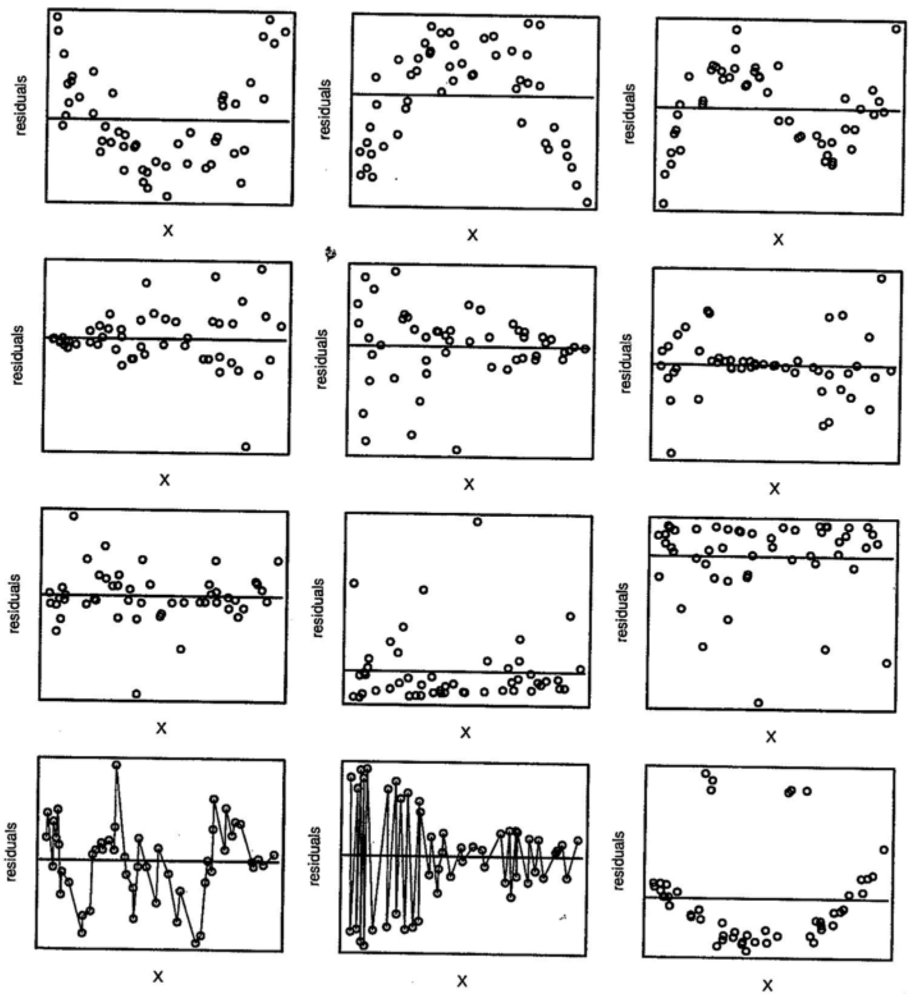


$e_i = \hat{y}_i - y_i$

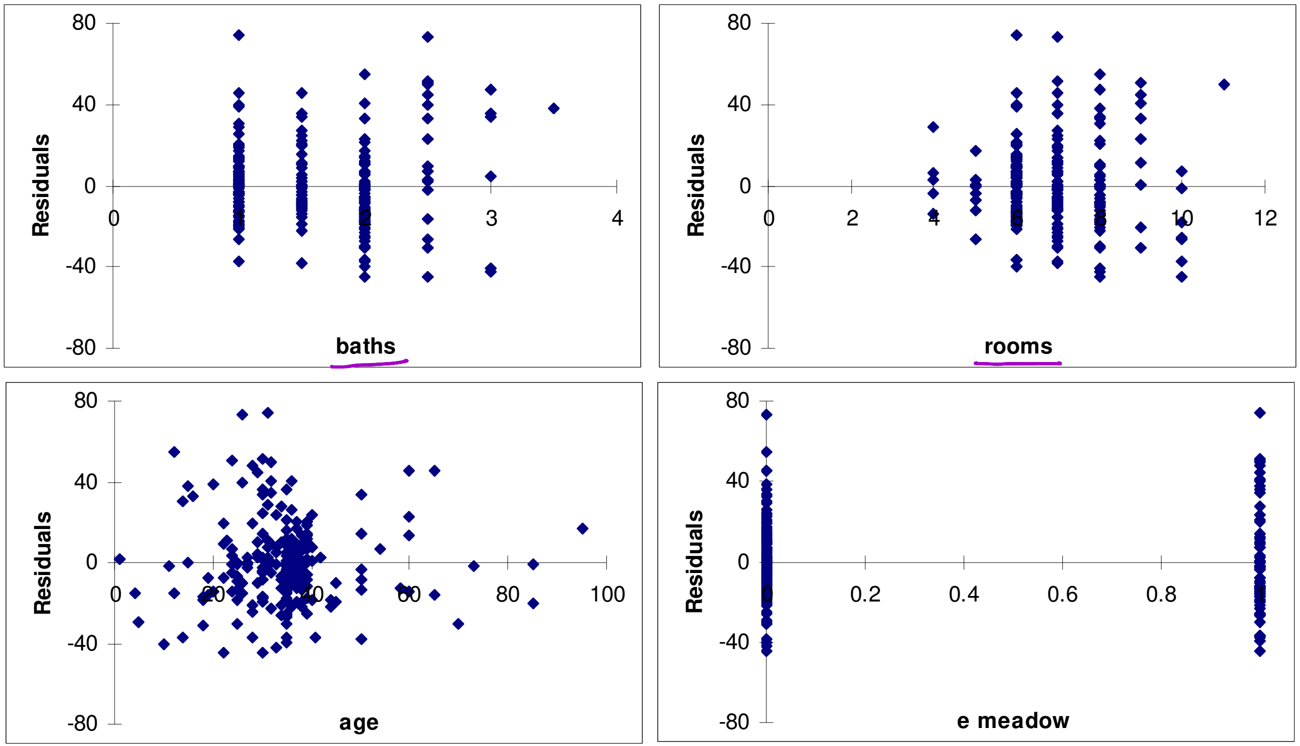
Residual Plots Seem OK



Residual Plots Do Not Seem OK

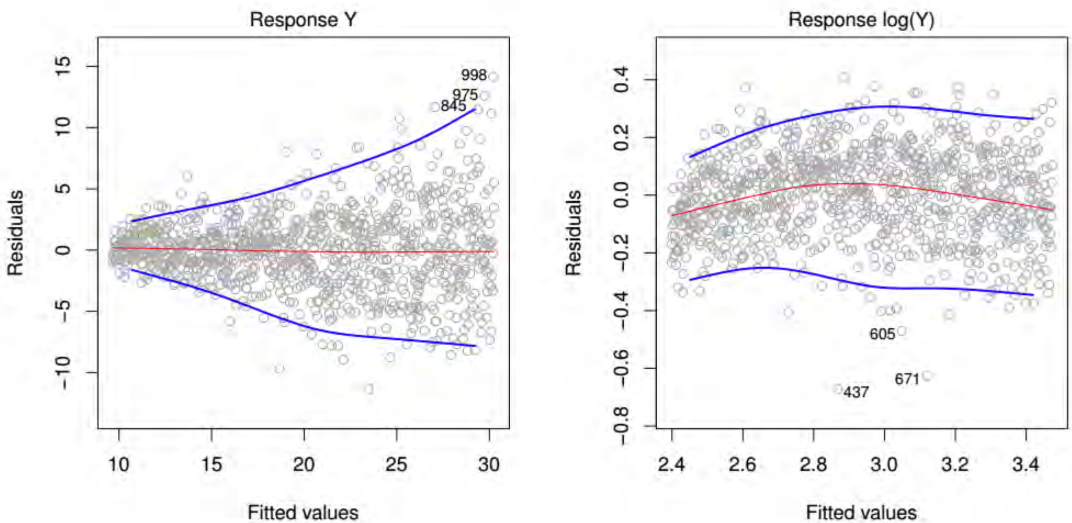


Residual Plots for Houses



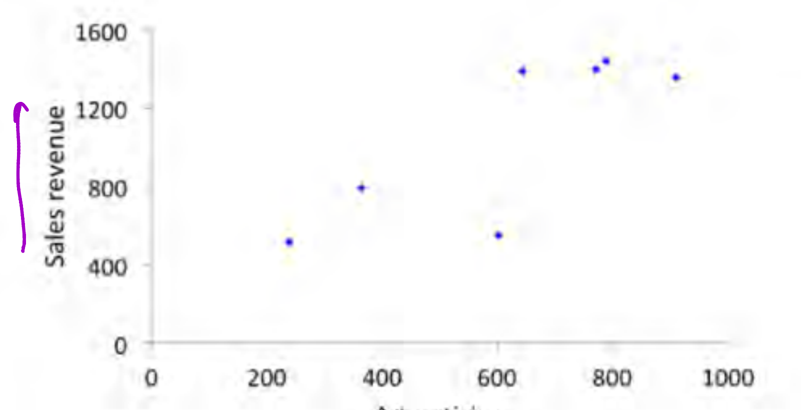
Non-constant variance in the residuals

- Also known as heteroscedasticity
- Consider a transformation of the dependent variable
  - For example replace Y with $log(Y)$, Y^2, or $\sqrt{Y}$.



Overfitting: Too Many Predictive Variables

- Example:



$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots \beta_6 x^6$

- Two possible models:

$y = 1.4234x + 182.78$

$R^2 = 65\%$

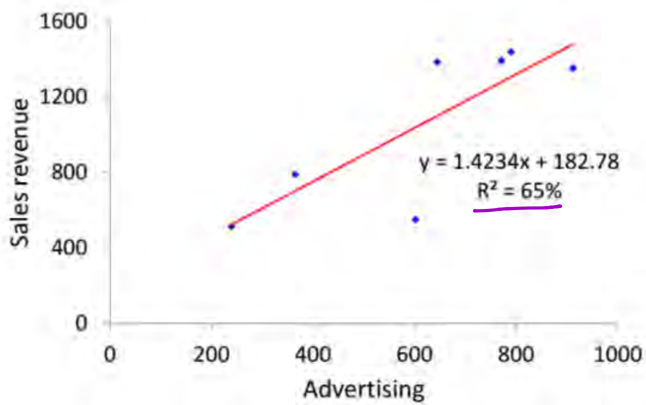
$y = -2.6 \times 10^{-11}x^6 + 9.7 \times 10^{-8}x^5 - 1.4 \times 10^{-4}x^4 + 1.1 \times 10^{-1}x^3 - 4.6 \times 10^{1}x^2 + 9.6 \times 10^{3}x - 7.9 \times 10^{5}$

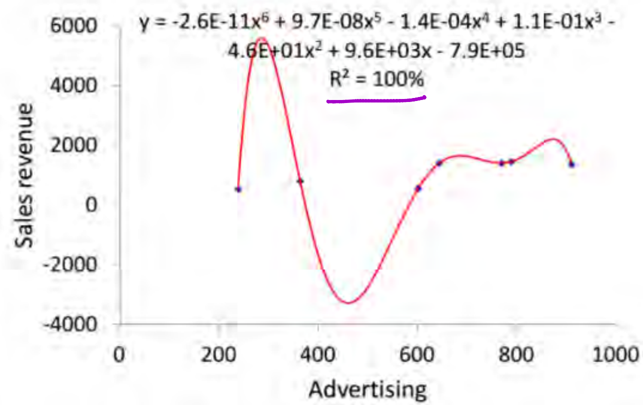
$R^2 = 100\%$

Which model do you select?

Overfitting: Too Many Predictive Variables

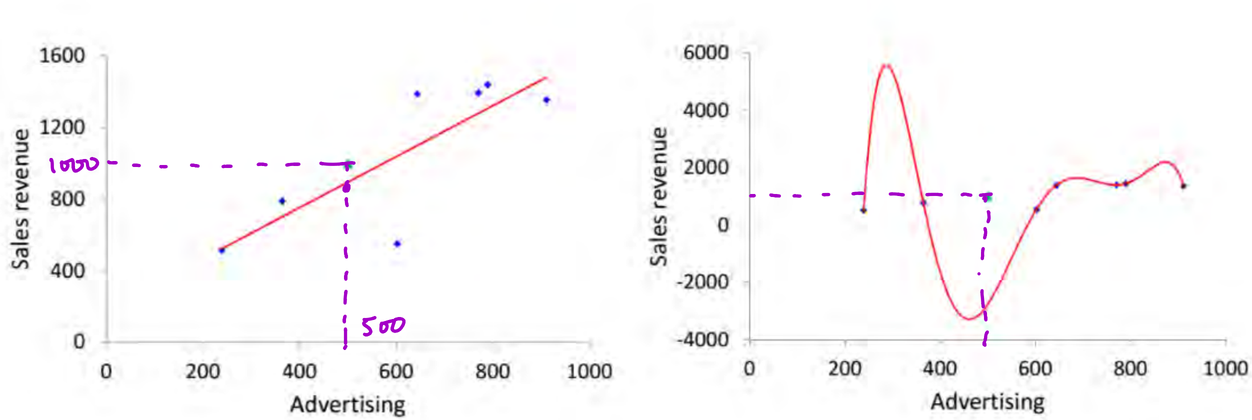
- Adding predictive variables gives a better and better fit to the data.





Overfitting: Too Many Predictive Variables

- Too many predictive variables may not do a good job of predicting out of sample

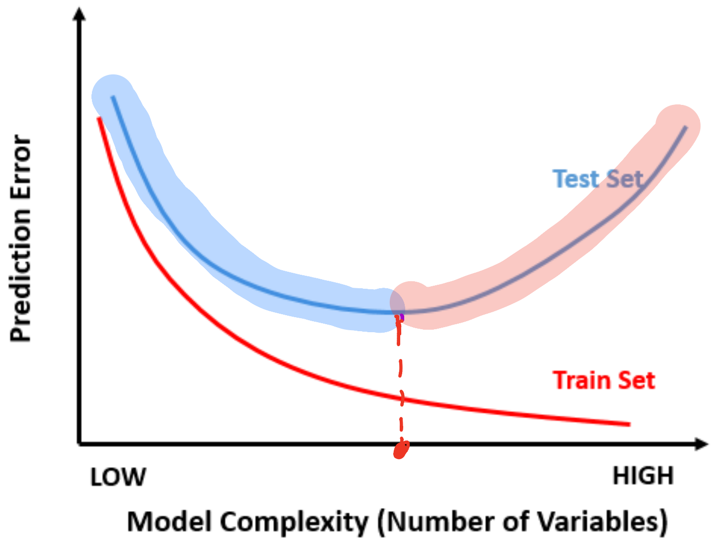


Overfitting: Too Many Predictive Variables

- Testing prediction error in training data will lead to overly optimistic performance assessments

Regularization

$(x_i, y_i) \quad i = 1 \ldots n$



$\|\beta\|_0$

$$\min_{\beta} \left[\underbrace{\sum_{i=1}^{n} (y_i - \beta^T x_i)^2} + \underbrace{\lambda \,(\#\,\text{non-zero coefficients in } \beta)} \right]$$

Lasso : $$\min_{\beta} \left(\sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \lambda \sum_{j=1}^{m} |\beta_j| \right)$$

$$\|\beta\|_1 = \sum_{i=1}^{m} |\beta_i|$$

Lasso Regression

- **Lasso Regression** is often formulated by dualizing the constraint $||\beta||_1 \leq t$, resulting in the problem

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta' x_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

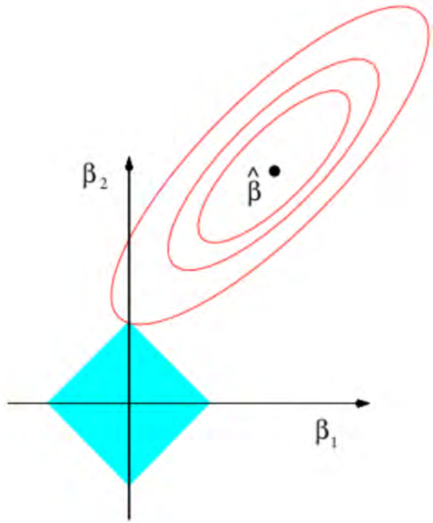
where the Lagrangian $\lambda \geq 0$ can be viewed as a tuning parameter, controlling the strength of the penalty.

- If a group of predictors re highly correlated among themselves, LASSO trends to pick only one of them and shrink other to zero.

- The red term is a **shrinkage penalty**.
  - If $\lambda = 0$ the penalty term has no effect, i.e., it produces the least squares estimates.
  - As λ increases, the flexibility decreases, i.e., variance decreases, but bias increases.
  - If $\lambda = \infty$, $\beta = 0$. Equivalent to the NULL model.

Lasso Regression

- Let $\hat{\beta}$ be the standard least squares estimate and $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j|$ be its L1 norm..
  - Value $t \geq t_0$ do **NOT** affect the least squares minimization.
  - $t < t_0$ leads to a **shrinkage** of the least squares solution.
  - Some coefficients will be 0 exactly, leading to variable selection and a simplification of the model.
  - If $t = 0$, all estimated coefficients are **shrunk to 0**



- **Bias and variance of the lasso**:
  Generally speaking,
  - The bias increases as λ (amount of shrinkage) increases
  - The variance decreases as λ (amount of shrinkage) increases

Lasso Regression for Housing Example