Lecture 4

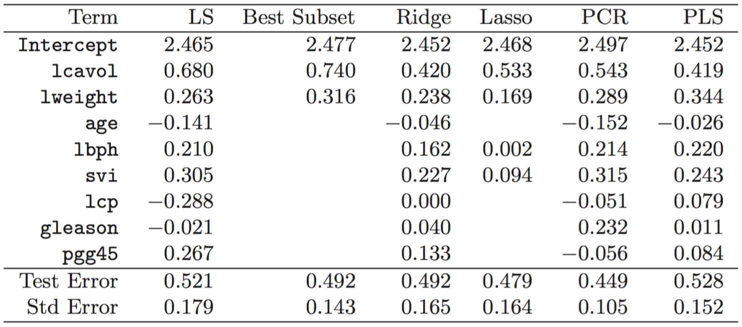
IDS575: Statistical Models and Methods
Theja Tulabandhula

Notes derived from the book titled "Elements of Statistical Learning [2nd edition] (Sections 3.4 and 4.1-4.4)

We have briefly discussed bias-variance trade-off in Lecture 3. We will again revisit it in Lecture 5 (Model Assessment and Selection).

1 Applications

1.1 Prostate Cancer Prediction



Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	−0.141		−0.046		−0.152	−0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	−0.288		0.000		−0.051	0.079
gleason	−0.021		0.040		0.232	0.011
pgg45	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

Figure 1: Prostate cancer dataset: Coefficients estimated using extensions of linear regression. The last two lines are post cross-validation and report numbers using the 30 observation test data.

Figure 1 shows the performance of certain biased methods (we saw subset selection last time) compared to vanilla linear regression.

1.2 Salary Calculator by Stackoverflow

Here is an example of regression task used to create an information tool[1] (Announced on Sept 19th 2017).

The folks at Stackoverflow, an online question answering website that also provides a job board, released a salary calculator. Its purpose is to help job seekers as well as employers find typical salaries based on input variables such as experience level, location, technologies used and educational qualification.

So they collected data using a survey. And the survey seeks the information related to the input variables defined above. As an example of some preliminary analysis, a plot of salary versus years of experience is shown in Figure 2. The authors mention that developers in San Francisco and Seattle have different salaries compared to the rest of the US. In the survey, they did not ask for the city. To overcome this, they used the IP addresses to geolocate respondents.

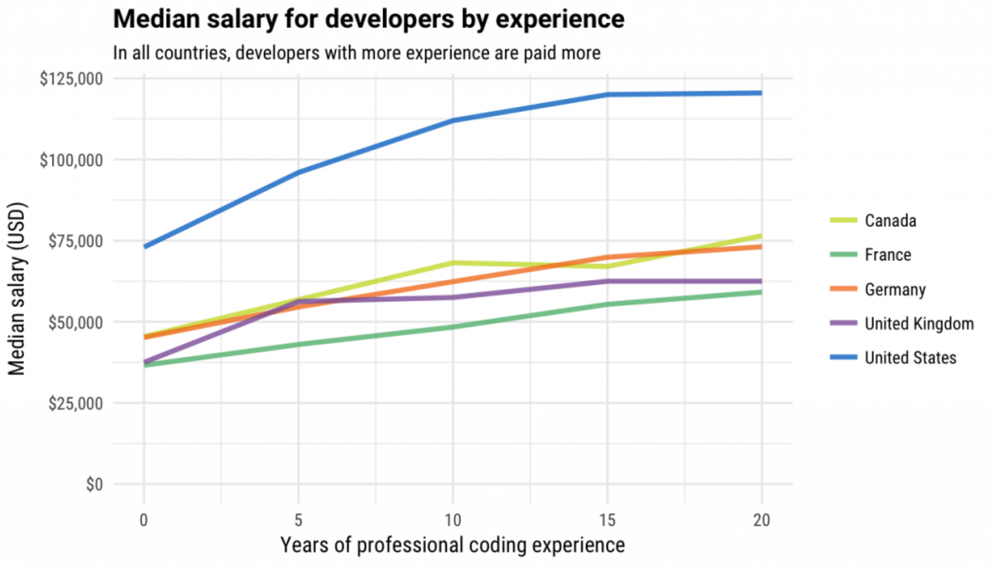


Figure 2: Exploratory plot from the salary survey (Image source).

Here is another exploratory plot (Figure 3) showing the impact of type of work (technologies used) on salaries.

The authors chose to create salary predictions only when there were enough observations for a given location (city or country). And the model they used is the familiar *linear regression*[2]. Specifically, they modeled the salaries in log scale, because they found the salary distribution was log-normal, 'with a long tail of very high salaries'.

Here is a plot (Figure 4) of the residuals, giving a hint to whether linear modeling is a good enough model choice here. The residuals seem mostly flat, except for the US where they are low around the ends. This is an indicator that a linear model is not sufficient here.

[1] See link for more information.

[2] They may have done an ad-hoc subset selection, which is not discussed.

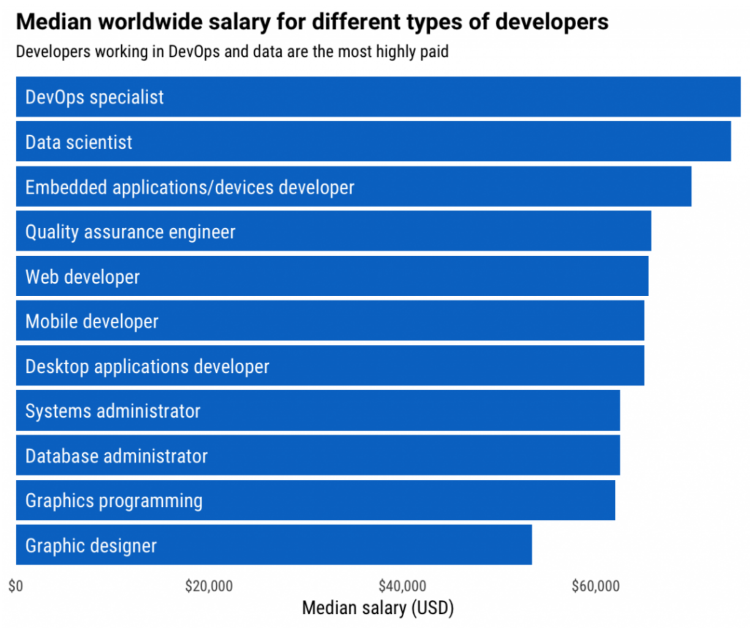


Figure 3: Another exploratory plot from the salary survey (Image source).

Here is a link to fun discussion page, by people who may use such a predictive tool!

2 Biasing Linear Regression via Ridge Regression and LASSO

Other than k-nearest neighbors, we have been focusing on linear methods for regression so far. Today, we will see linear methods for classification. This does not mean that linear methods are the only ones out there. One can use many other non-linear methods for both regression and classification. One of the reasons we have been focusing on linear methods is because they are intuitive and simpler to understand.

Subset selection is great, but is a discrete process: you either retain a coordinate of the input variable or discard it. Intuitively, variance can be further reduced if we don't completely discard these coordinates. One such attempt is via ridge regression and LASSO.

2.1 Ridge Regression

The idea of ridge regression is pretty straightforward. We want to shrink regression coefficients ($\widehat{\beta}_j$s) by imposing a penalty. What do we mean by this? Remember the RSS objective? We will just add another term to that:

$$RSS(\lambda, \beta) = (\mathbf{y} - \mathbf{X}\,\beta)^T(\mathbf{y} - \mathbf{X}\,\beta) + \lambda\beta^T\beta.$$

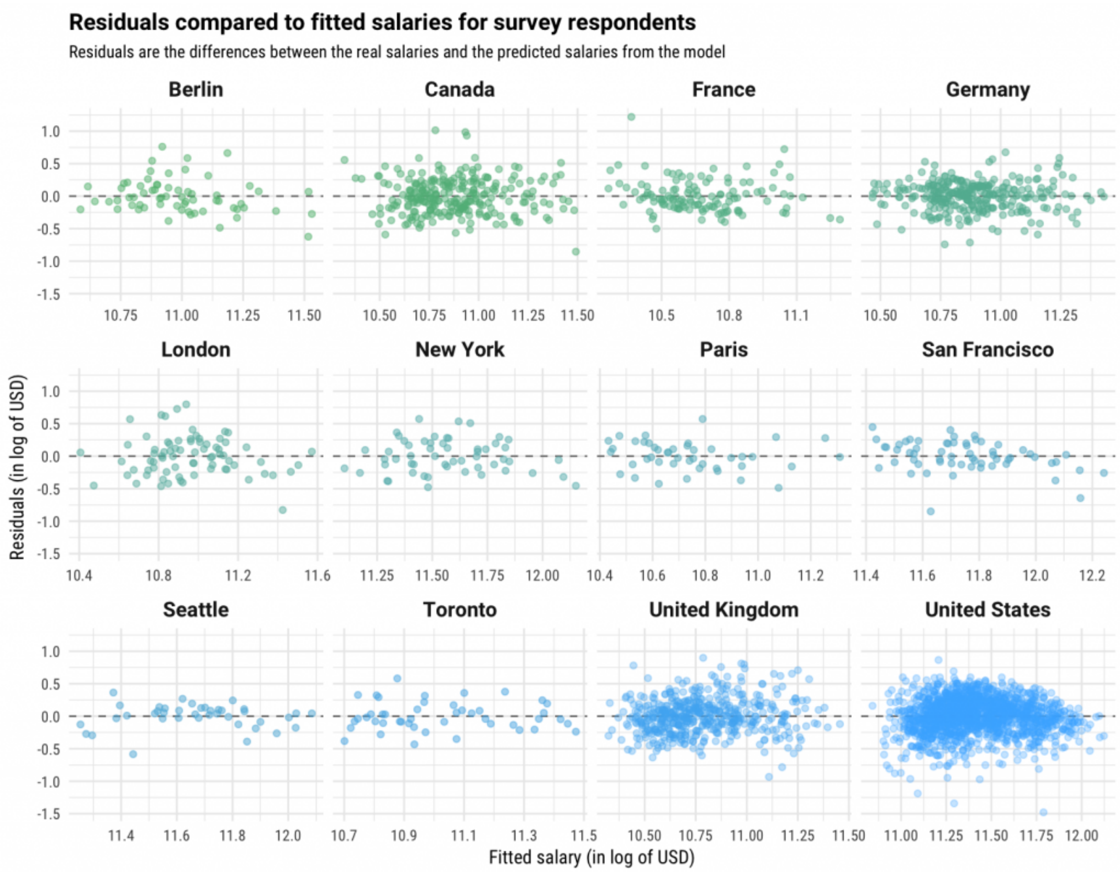


Figure 4: Plot of residuals with linear regression: salary dataset. (Image source).

What does λ do? It is a non-negative parameter. If it is large, it will make the coefficients shrink towards 0.

It can be shown that for a given λ, there exists a t such that the solution to the following problem is the same as that of $\text{RSS}(\lambda, \beta)$:

$$(\mathbf{y} - \mathbf{X}\,\beta)^T(\mathbf{y} - \mathbf{X}\,\beta)$$
$$\text{subject to } \beta^T\beta \leq t.$$

Why is this useful? It is useful because we can think of t as controlling the size of the coefficients directly. Recall that if we do this, then the chances of high positive coefficients on some variables and high negative coefficients on related variables is potentially mitigated.

Note 1. There is no point penalizing the intercept (corresponding to the coefficient of the last column of $\mathbf{X}$, which is all ones). If we *center* our data (i.e., subtract the feature means from training), we can compute the intercept separately. So, in our exposition here, lets assume data is centered and there is no intercept.

(Not so) surprisingly, the $\widehat{\beta}$ for ridge regression looks like the solution for linear regression:

$$\widehat{\beta} = (\mathbf{X}^T\,\mathbf{X} + \lambda I)^{-1}\,\mathbf{X}^T\,\mathbf{y}\,.$$

Figure 5 shows the ridge coefficients as a function of (another function [3] $df()$ of) λ. λ is like the k in subset selection, and will need to be determined using cross-validation.

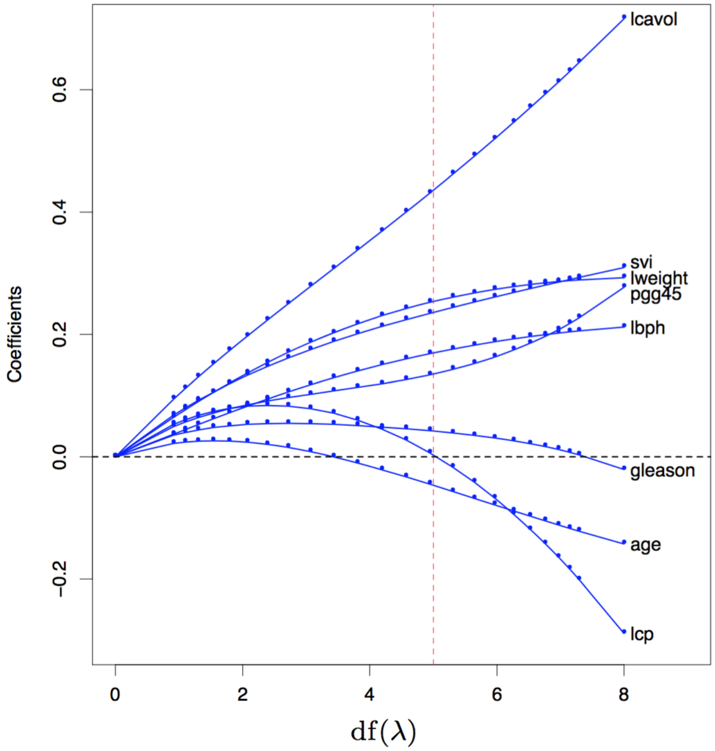


Figure 5: Prostate cancer dataset: Ridge regression coefficients as a function of (some function $df()$ of) λ.

Note 2. Just like linear regression is related to Maximum Likelihood Estimation, ridge regression is related to Maximum A posteriori Estimation (MAP). The latter estimation problem (MAP) is like MLE but with notions of priors and posteriors.

Note 3. Quick aside about the terms *prior* and *posterior*: If you remember Bayes rule, which is essentially about conditional distributions, it looks like this: $Pr(\theta|Z) = \frac{1}{Pr(Z)} Pr(Z|\theta) Pr(\theta)$. The second term in the numerator is called prior and the term on the left hand side is called

[3] This function is inversely related to λ.

posterior. This is because, we know that θ is distributed according to $Pr(\theta)$ before we observe Z (hence is called prior), and is distributed according to $Pr(\theta|Z)$ after observing Z (hence is called posterior). The term $Pr(Z|\theta)$ is called the likelihood!

Example 1. Say $Y \sim N(X^T\beta, \sigma^2)$ and $\beta \sim N(0, \tau^2 I)$. Then, assuming known τ^2, σ^2, the log posterior of β is proportional to the RSS(λ, β) above where $\lambda = \sigma^2/\tau^2$ (we will not show this here).

2.1.1 Interpretation using Singular Value Decomposition

Let the singular value decomposition (SVD) of matrix $\mathbf{X}$ be $\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^T$, where $\mathbf{D}$ is a diagonal ($d_1 \geq d_2 \geq ...$), $\mathbf{U}$ and $\mathbf{V}$ are orthogonal. Let us not worry about how this is computed. The thing to notice here is that:

- The columns of $\mathbf{U}$ span the column space of $\mathbf{X}$.
- The columns of $\mathbf{V}$ span the row space of $\mathbf{X}$.

Okay, assuming the decomposition, the training predictions by the ridge solution looks like:

$$\begin{aligned}\mathbf{X}\,\widehat{\beta} &= \mathbf{U}\,\mathbf{D}(\mathbf{D}^2 + \lambda I)^{-1}\,\mathbf{D}\,\mathbf{U}^T\,\mathbf{y} \\ &= \sum_{j=1}^{p} \mathbf{u}_j \, \frac{d_j^2}{d_j^2 + \lambda}\, \mathbf{u}_j^T\,\mathbf{y},\end{aligned}$$

where $\mathbf{u}_j$ is the j^{th} column of $\mathbf{U}$.

Note 4. The term $\frac{d_j^2}{d_j^2+\lambda} \leq 1$.

Like linear regression, ridge regression computes the coordinates of $\mathbf{y}$ with respect to the orthonormal basis $\mathbf{U}$. It then shrinks these coordinates by factors $\frac{d_j^2}{d_j^2+\lambda}$. There is a greater amount of shrinkage to coordinates of basis vectors with smaller d_j^2.

What does a smaller d_j^2 mean? Well, for that, lets understand the (unnormalized) sample covariance matrix $\mathbf{X}^T\,\mathbf{X} = \mathbf{V}\,\mathbf{D}^2\,\mathbf{V}^T$. The vectors $\mathbf{v}_j$s are the *principal component directions* of $\mathbf{X}$.

The first principal component direction $\mathbf{v}_1$ has the property that $\mathbf{z}_1 = \mathbf{X}\,\mathbf{v}_1$ has the largest sample variance among all linear combinations of columns of $\mathbf{X}$. This variance is equal to $\frac{d_1^2}{N}$.

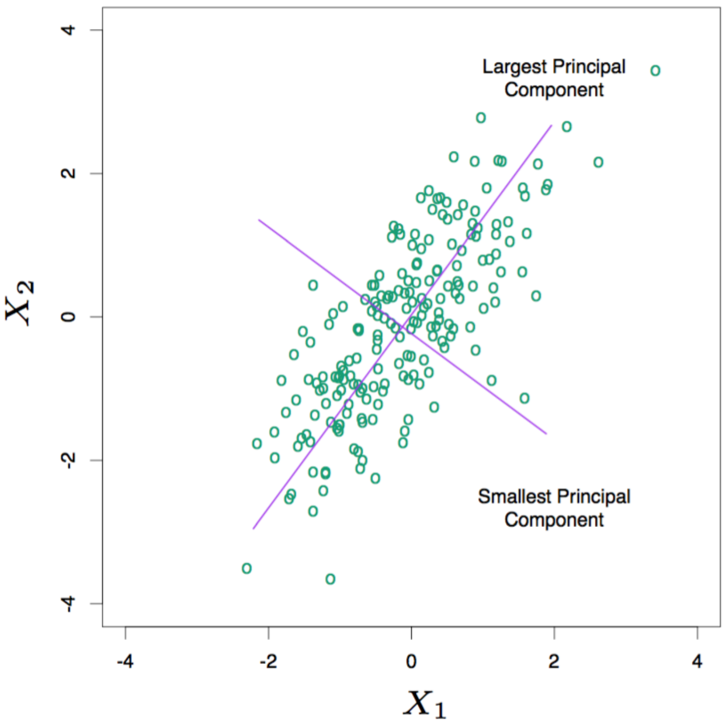


Figure 6: Principal components: illustration.

Note 5. $\mathbf{z}_1 = \mathbf{X}\,\mathbf{v}_1 = d_1\,\mathbf{u}_1$ is called the first principal component of $\mathbf{X}$. Subsequent principal components $\mathbf{z}_j$ have variance $\frac{d_j^2}{N}$, and are orthogonal to others.

Example 2. Figure 6 shows the principal components of a simulated 2-dimensional data.

Small singular values d_j correspond to directions in the column space of $\mathbf{X}$ that have small variance, and ridge regression shrinks these directions the most.

Note 6. When there are large number of input variables, another way to approach the dimensionality issue is to produce a small number of linear combinations $Z_m, m = 1, ..., M$ of the original inputs X_j and use these for regression. When Z_m are the principal components, the approach is called *principal component regression.* While ridge regression shrinks the coefficients of the principal components, principal component regression discards the $p - M$ smallest eigenvalue components.

2.2 LASSO

LASSO is short for *Least Absolute Shrinkage and Selection Operator.* Thats a mouthful!

It is very slightly different from ridge regression:

$$RSS(\lambda, \beta) = (\mathbf{y} - \mathbf{X}\,\beta)^T(\mathbf{y} - \mathbf{X}\,\beta) + \lambda \sum_{j=1}^{p} |\beta_j|.$$

The second term is called an ℓ_1-*penalty*[4].

Note 7. There is no closed form expression for the LASSO estimate $\widehat{\beta}$ in general.

Example 3. Figure 7 shows how $\widehat{\beta}_j$s vary for LASSO as (a function of) λ is varied.

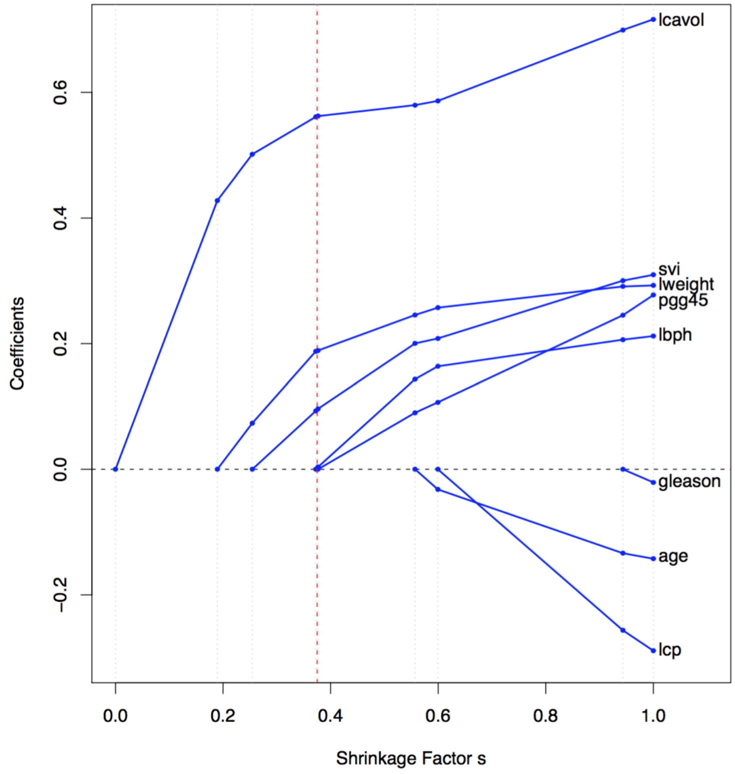


Figure 7: LASSO Coefficients as λ is varied (shrinkage factor above is inversely related to λ).

Lets contrast LASSO with ridge regression. See Figure 8 for an example 2-dimensional input setting. The residual sum of squares (without the penalty term) is depicted as elliptical contours (these are the level sets, where points on the curve lead to the same value). The LASSO level set looks like a diamond. Both methods find the first point where the elliptical contours hit the penalty regions. Since the diamond has corners, if the ellipse hits the penalty region at the corner, some $\widehat{\beta}_j$s will be zero!

Note 8. We have ignored how to compute LASSO model. The search for the best model is not difficult because the problem is *convex*. We may revisit this aspect later.

[4] Sometimes also denoted as L_1-penalty.

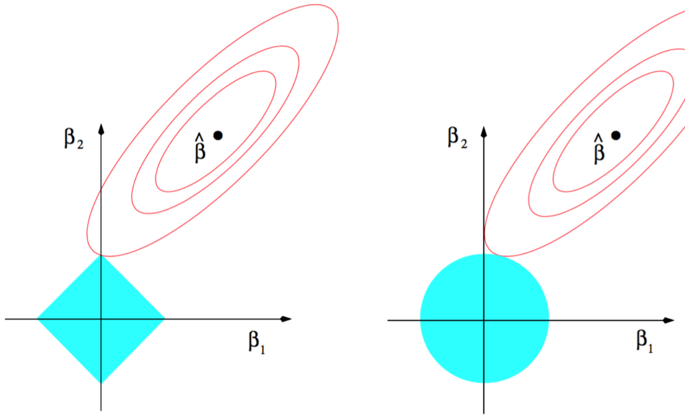


Figure 8: LASSO versus ridge regression: contour plots of the residual sum of squares term and the penalty terms.

3 Summary

We learned the following things:

- Ridge regression and reasoned what the penalty does from the SVD point of view.
- LASSO.

In the next lecture, we will discuss the nuances of classification. If time permits, we will also look at model assessment and selection: (a) the bias-variance decomposition, (b) Bayesian Information Criterion (BIC), (c) cross-validation, and (d) bootstrap methods.

A Sample Exam Questions

1. How does cross-validation improve over just breaking training data further into a validation set and using the remaining for actual training?

2. Why is validation necessary? Why can't we just use test set?

3. What is the motivation between ridge regression and LASSO? How do they relate to subset selection?