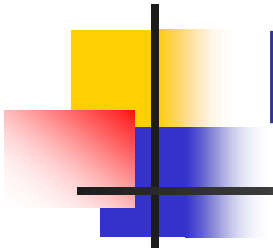
Session 4
Resampling Methods

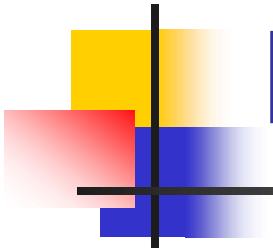
An Introduction to Machine Learning for the Behavioural and Social Sciences

Gyeongcheol Cho & Heungsun Hwang



Problem situations

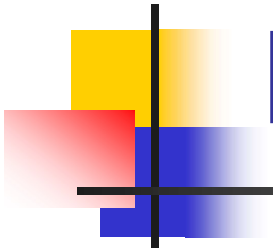
- A researcher trained a linear regression model on a sample. Now he would like to obtain the test error of the model but does not have a designated test sample.
- A researcher would like to apply KNN regression to her dataset and but is not sure which value she should choose for *K*.
- A researcher trained a linear regression model and would like to conduct hypothesis tests for individual coefficients. However, the residuals seem to be far from the normal distribution.



Resampling Methods

- Resampling methods involve repeatedly drawing samples from a training set and refitting a model of interest on each sample to obtain addition information about the fitted model.

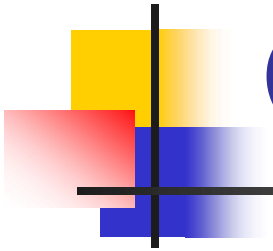
- The Information that can be obtained
  - Test error estimate
  - Standard error / confidence interval



Resampling Methods

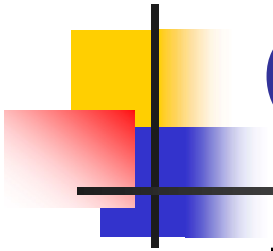
- Cross validation aims to estimate the test error associated with a given statistical method to evaluate its performance (model assessment), or to select the appropriate level of flexibility (model selection).

- The bootstrap is used in several contexts, most commonly to provide a measure of accuracy of a parameter estimate or of a statistical method.



Cross Validation

- Cross validation refers to a class of resampling methods that estimate the test error by holding out a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.



Cross Validation

- The training error is the average error that occurs when a trained model is used for prediction on the training sample.

- The test error is the average error that results from using a statistical learning method to predict the response on a new observation, i.e., a measurement that was not used in training the method.

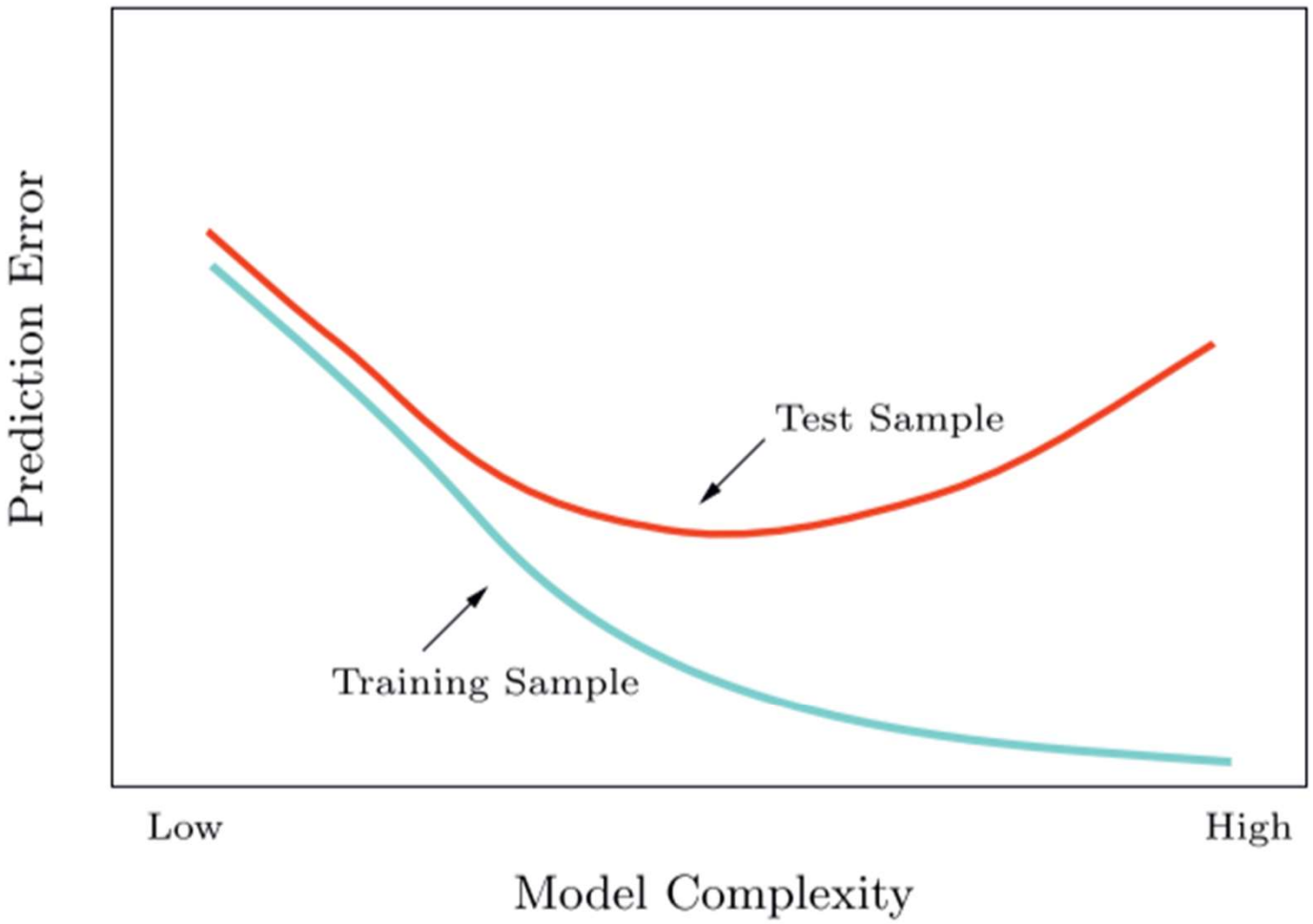
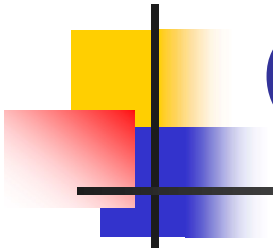


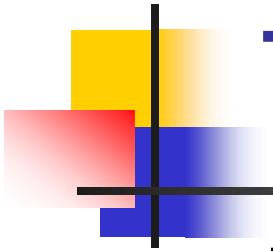
FIGURE 2.11. *Test and training error as a function of model complexity.*

Hastie, Tibshirani, & Friedman (2001, p. 38)



Cross Validation

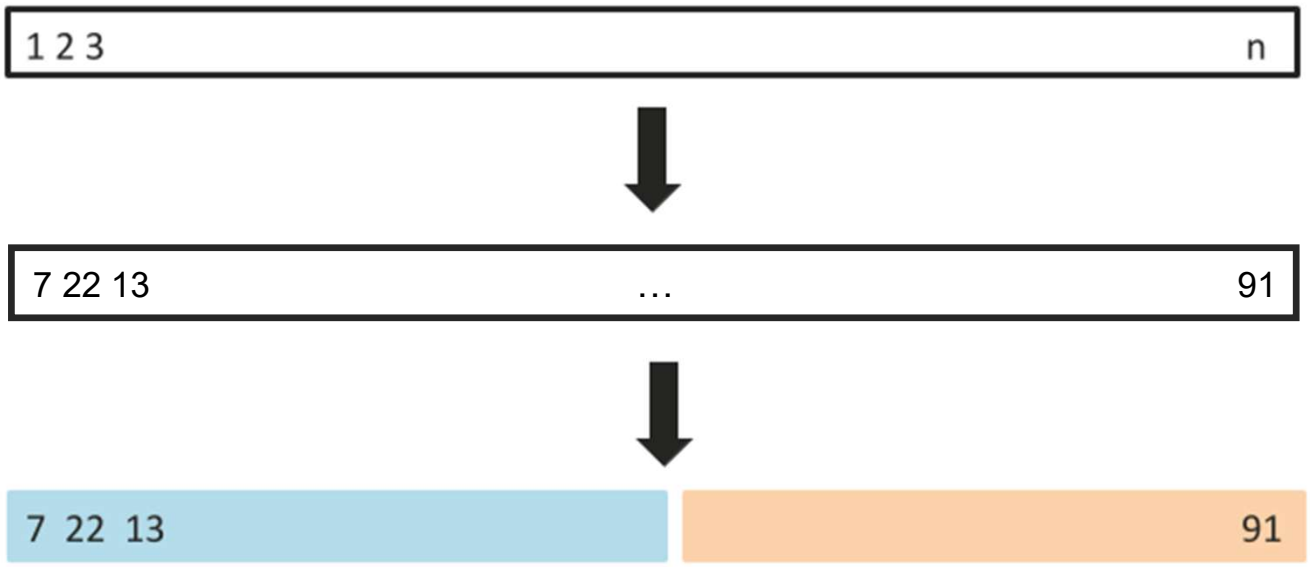
- Cross validation refers to a class of resampling methods that estimate the test error by holding out a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.
- Three types of cross validation
  - The validation set approach (holdout method)
  - Leave-one-out cross validation
  - K-fold cross validation



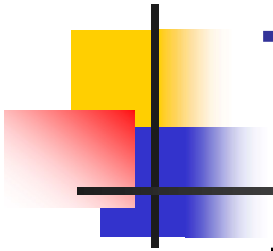
The Validation Set Approach

- This approach involves randomly dividing a set of observations into two subsets of comparable size, a training set and a validation (hold-out) set.

The Validation Set Approach



Revised from Figure 5.1 in James, Witten, Hastie, & Tibshirani, & Friedman (2013, p. 199)



The Validation Set Approach

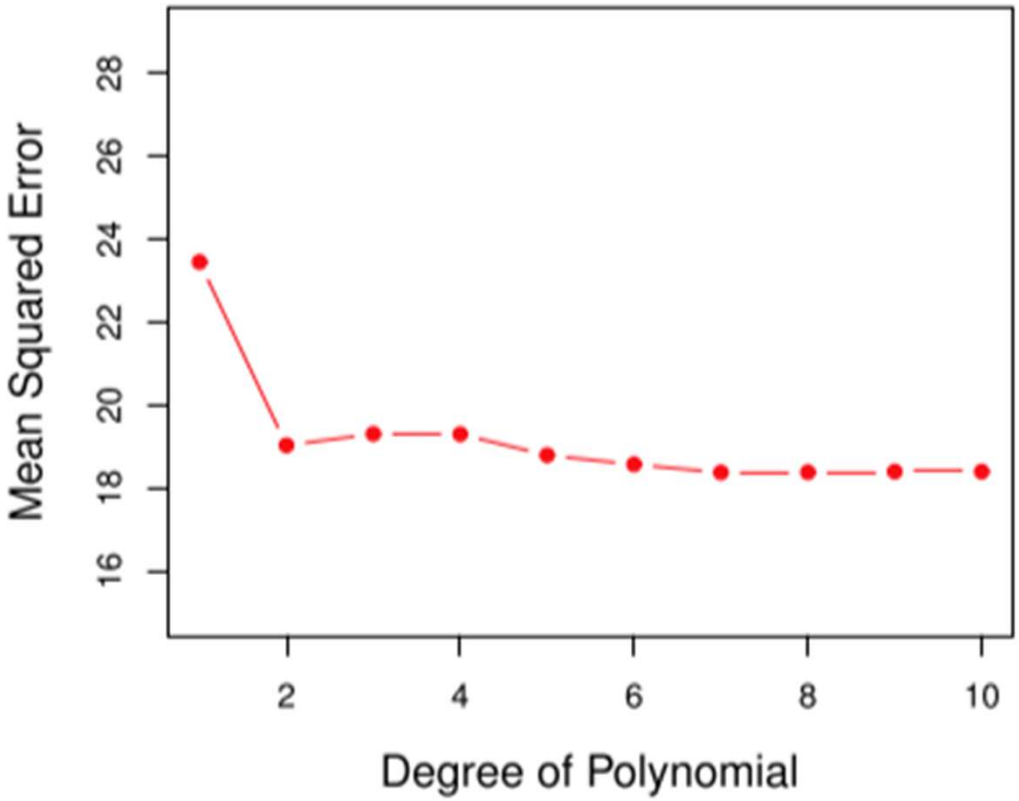
- This approach involves randomly dividing a set of observations into two subsets of comparable size, a training set and a validation (hold-out) set.

- The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set. The resulting validation set error rate (e.g., MSE in the case of quantitative responses) provides an estimate of the test error rate.

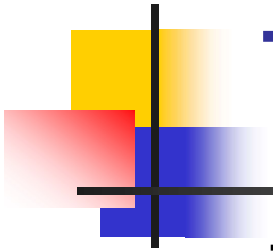
Example: The Validation Set Approach

- Auto MPG data set (auto_mpg.csv)
  - This datasets was initially reported in the 1983 American Statistical Association Exposition.
  - It includes eight attributes of cars that can be used to predict its mile per gallon (mpg)
    - 1. mpg: continuous
      2. cylinders: multi-valued discrete
      3. displacement: continuous
      4. horsepower: continuous
      5. weight: continuous
      6. acceleration: continuous
      7. model year: multi-valued discrete
      8. origin: multi-valued discrete
      9. car name: string (unique for each instance)
  - It has 392 observations
  - Original dataset can be downloaded from https://archive.ics.uci.edu/ml/datasets/auto+mpg

Example: The Validation Set Approach



James, Witten, Hastie, & Tibshirani, & Friedman (2013, p. 200)

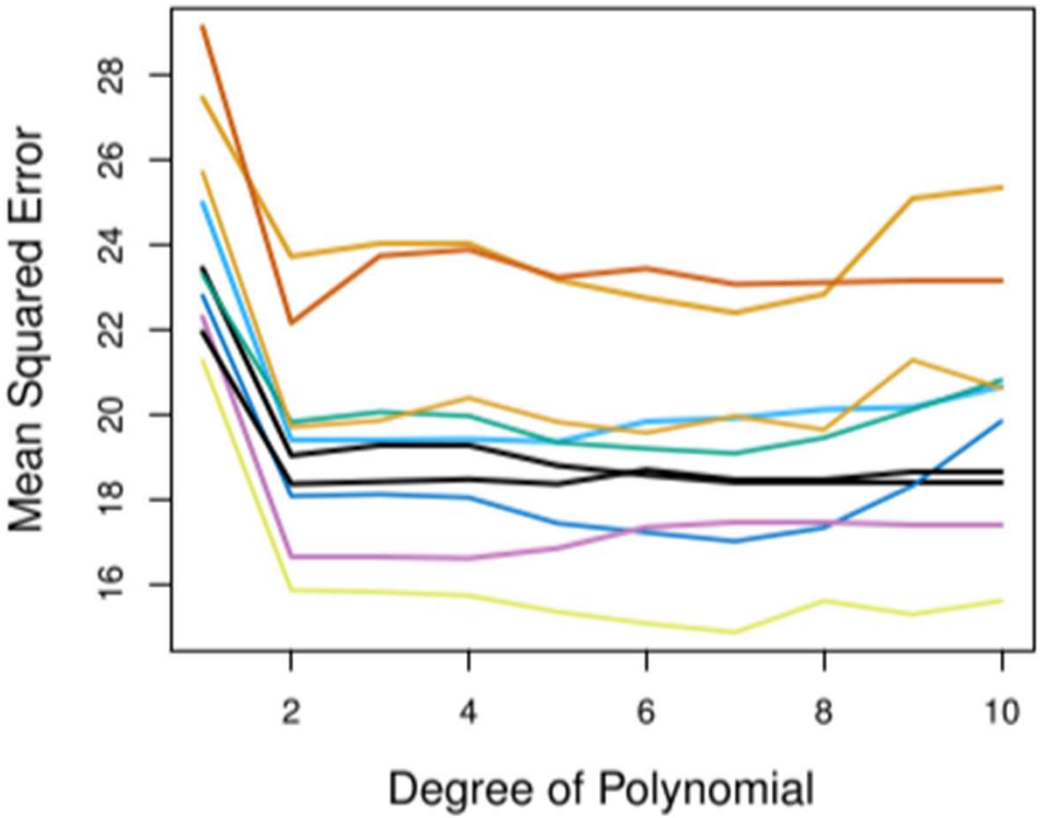


The Validation Set Approach

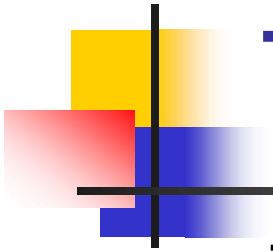
- The validation set approach is conceptually simple and is easy to implement.

- But it has two potential limitations:
  - The validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are in the validation set.

Example: The Validation Set Approach



James, Witten, Hastie, & Tibshirani, & Friedman (2013, p. 200)



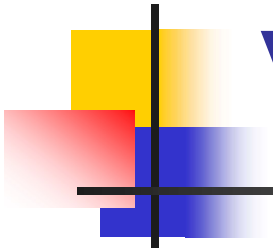
The Validation Set Approach

- The validation set approach is conceptually simple and is easy to implement.

- But it has two potential limitations:
  - The validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are in the validation set.

  - Only a subset of the observations (in the training set) are used to fit the model. Since statistical methods tend to perform worse when trained on fewer observations, this suggests that the validation set error rate may tend to be a biased estimate of the test error for the model fit on the entire data set.

Leave-One-Out Cross Validation



- LOOCV involves splitting the set of observations into two subsets. However, instead of creating two subsets of comparable size, a single observation is used for the validation set and the remaining observations make up the training set.
- The statistical learning method is fit on the n-1 training observations, and a prediction of a single response is made for the excluded observation. We can repeat the procedure n times, using each observation as the validation set.

Leave-One-Out Cross Validation

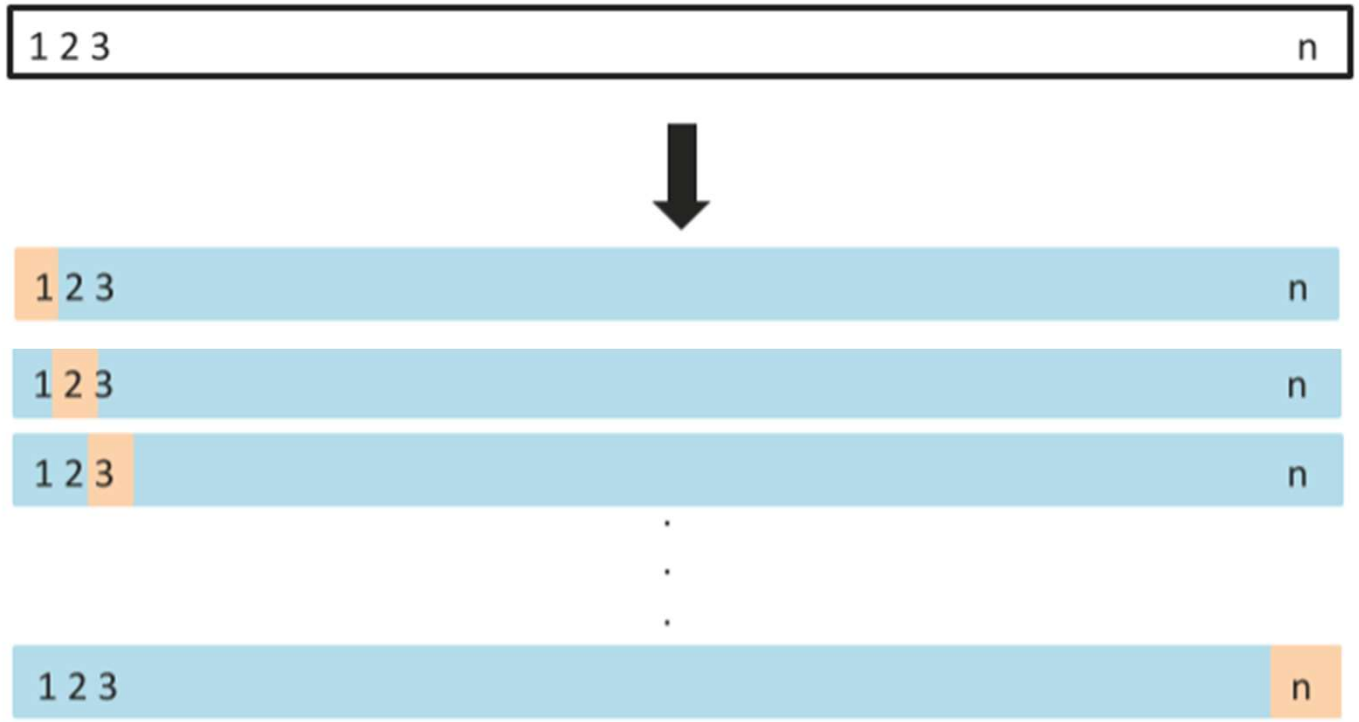
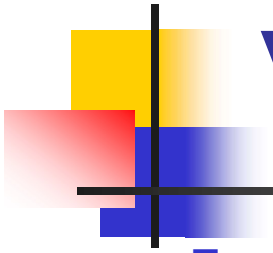
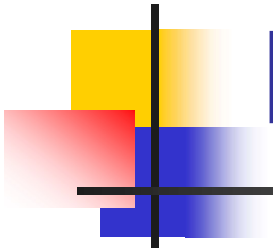


Figure 5.3 in James, Witten, Hastie, & Tibshirani, & Friedman (2013, p. 201)

Leave-One-Out Cross Validation



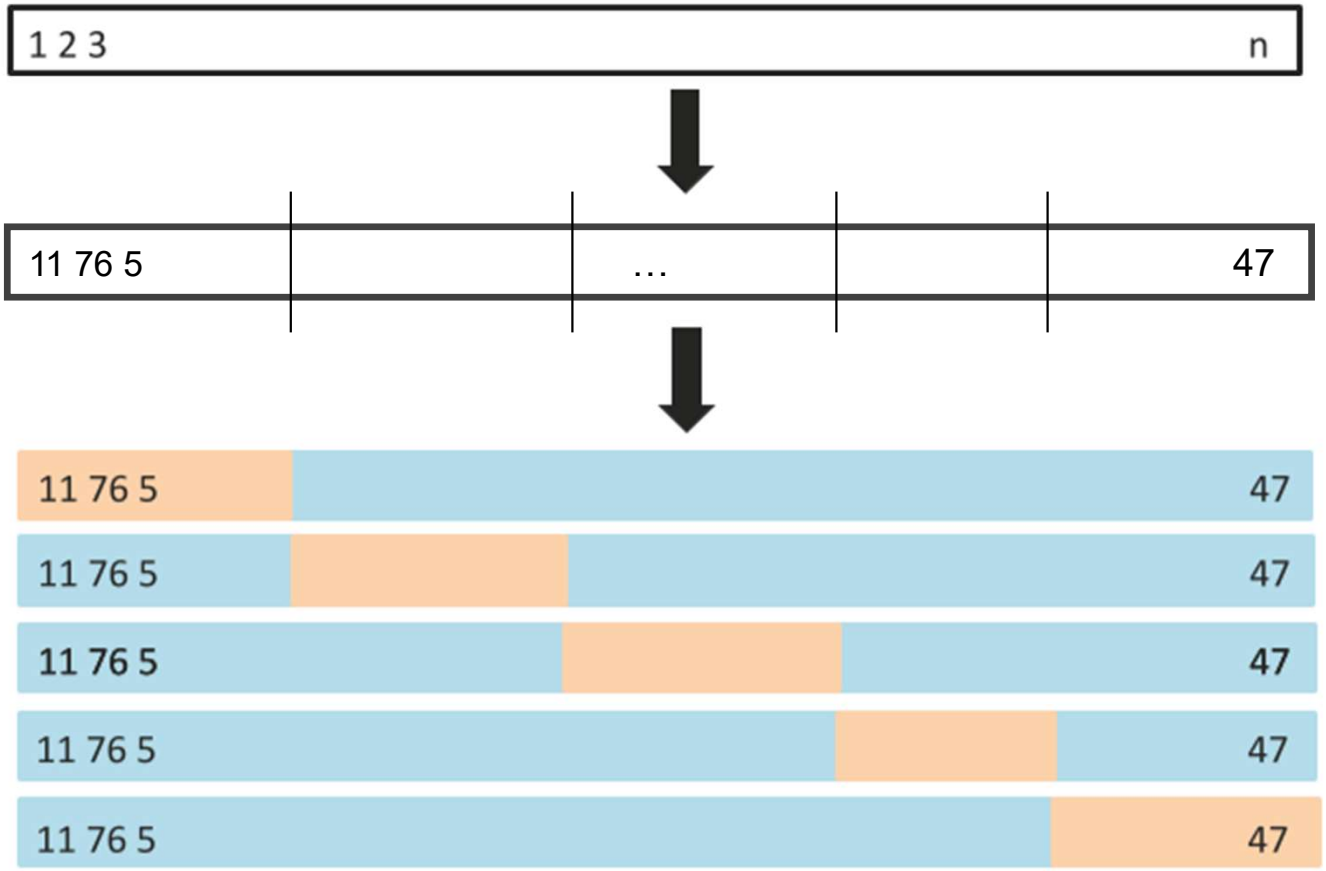
- LOOCV has two major advantages over the validation set approach.
  - It has far less bias because a statistical learning method is fit on training sets that contain n-1 observations, almost as many as are in the entire dataset.
  - Performing LOOCV multiple times will always yield the same results.
- But LOOCV has the potential to be expensive to implement because the model must be fit n times. This can be very time consuming if n is large, and if each individual model is slow to fit.
  - With least squares linear or polynomial regression, there exists an amazing shortcut makes the cost of LOOCV the same as that of a single model fit (see equation 5.2 in James et al., 2013, p. 202).



K-Fold Cross Validation

- An alternative to LOOCV is k-fold CV.
- This approach involves randomly dividing a set of observations into k groups (or folds) of approximately equal size. The first set is treated as a validation set and the method is fit on the remaining k-1 folds.
- The MSE is then computed on the observations in the validation set. The procedure is repeated k times; each time, the kth fold is treated as a validation set.

K-Fold Cross Validation



Revised from Figure 5.5 in James, Witten, Hastie, & Tibshirani, & Friedman (2013, p. 203)

K-Fold Cross Validation

- In practice, k = 5 or 10 is used.
  - when k = n, LOOCV = k-fold CV.
- K-fold CV is computationally more efficient than LOOCV especially if n is very large.
- Also, k-fold CV often gives a more accurate estimate of the test error rate than LOOCV.
  - The LOOCV estimate tends to be less biased yet be more highly variable than the k-fold CV estimate (James et al., 2013, pp. 183-184).

Example: LOOCV and 10 fold CV

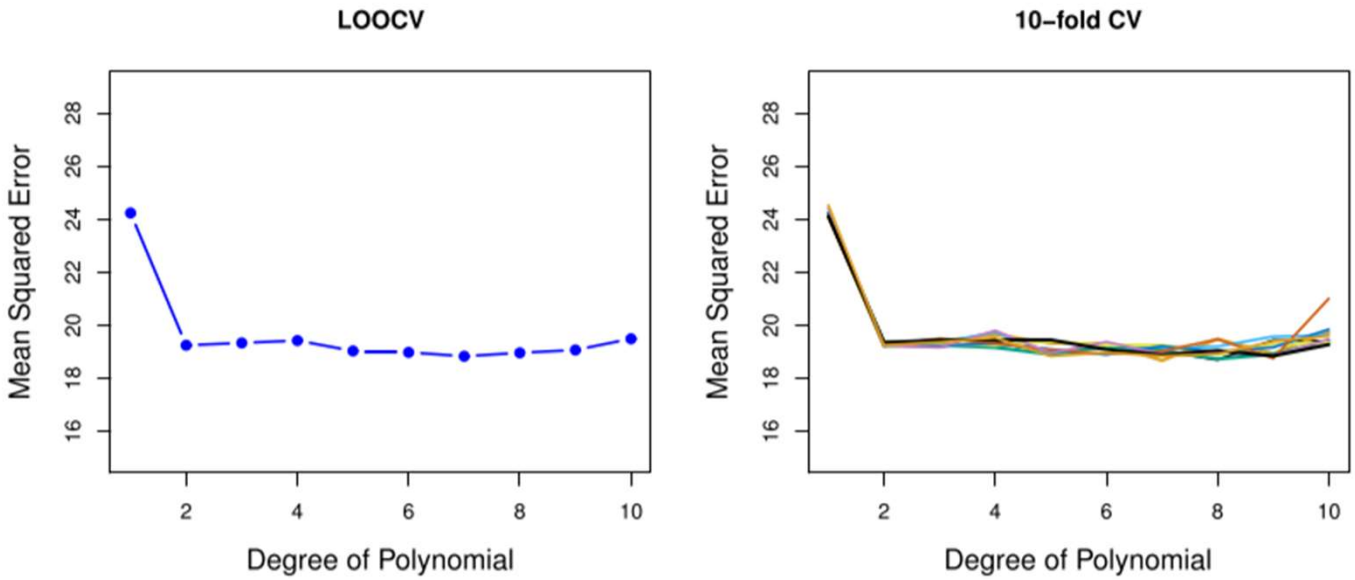


FIGURE 5.4. *Cross-validation was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. Left: *The LOOCV error curve.* Right: 10-*fold CV was run nine separate times, each with a different random split of the data into ten parts. The figure shows the nine slightly different CV error curves.*

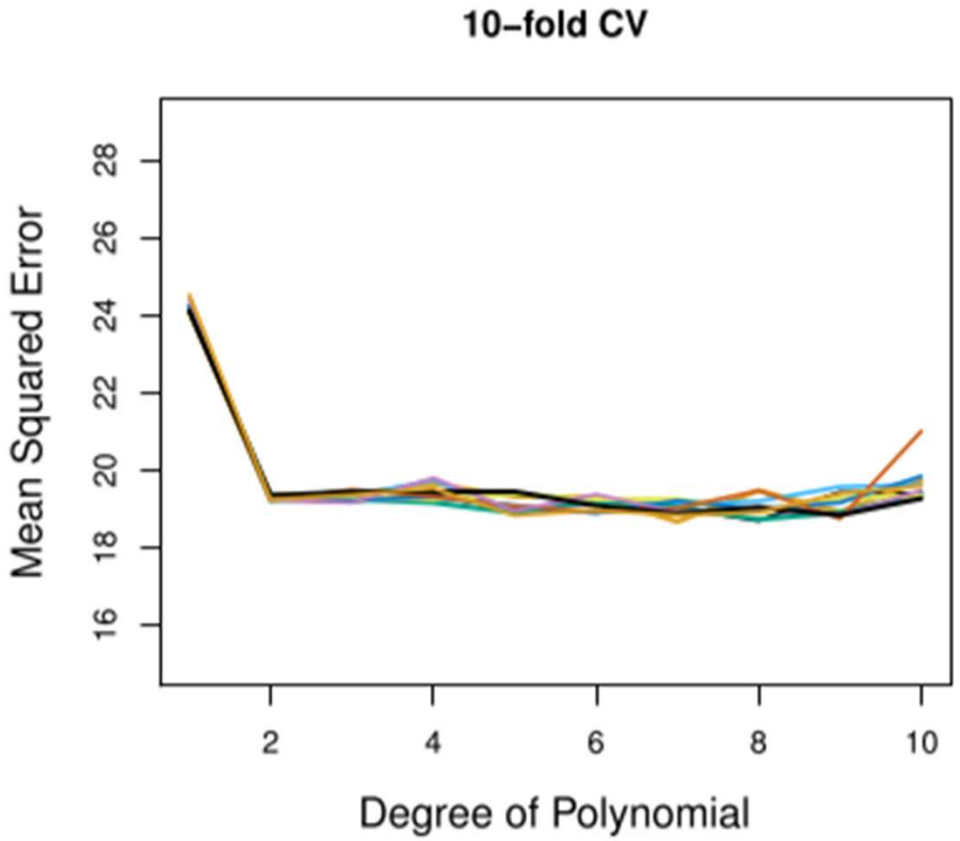
James, Witten, Hastie, & Tibshirani, & Friedman (2013, p. 202)

Problem situations

- A researcher trained a linear regression model on a sample. Now he would like to know the test error but do not have a designated test sample.
- A researcher would like to apply a KNN regression to her dataset and but is not sure which value she should choose for *K*.

Q. Is it acceptable to perform CV on the same dataset for determining the hyperparameter value of a model and evaluating the test error of the model?

Example: 10-fold Cross Validation



James, Witten, Hastie, & Tibshirani, & Friedman (2013, p. 202)

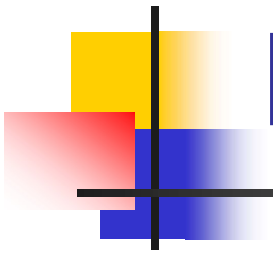
Q. Is it acceptable to perform CV on the same dataset for determining the hyperparameter value of a model and evaluating the test error of the model?

- Not acceptable
  - the test error may be underestimated because the validation samples has been already used for determining the hyperparameter value of the model.

- Solution?
  - To hold out some portion of a dataset in advance for testing the model and perform K-fold cross validation on the remaining portion of the dataset to determine its hyperparameter value.

  - Nested K-fold cross validation

    (e.g., https://weina.me/nested-cross-validation/).

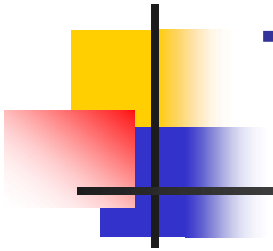


Problem situations

- A researcher trained a linear regression model on a sample. Now he would like to know the test error but do not have a designated test sample.

- A researcher would like to apply a KNN regression to her dataset and but is not sure which value she should choose for *K*.

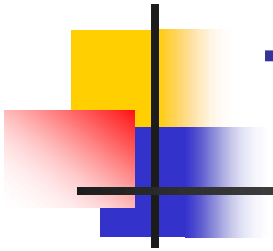
- A researcher trained a linear regression model and would like to estimate standard errors of regression coefficients. However, residuals seem to be far from the normal distribution.



The Bootstrap

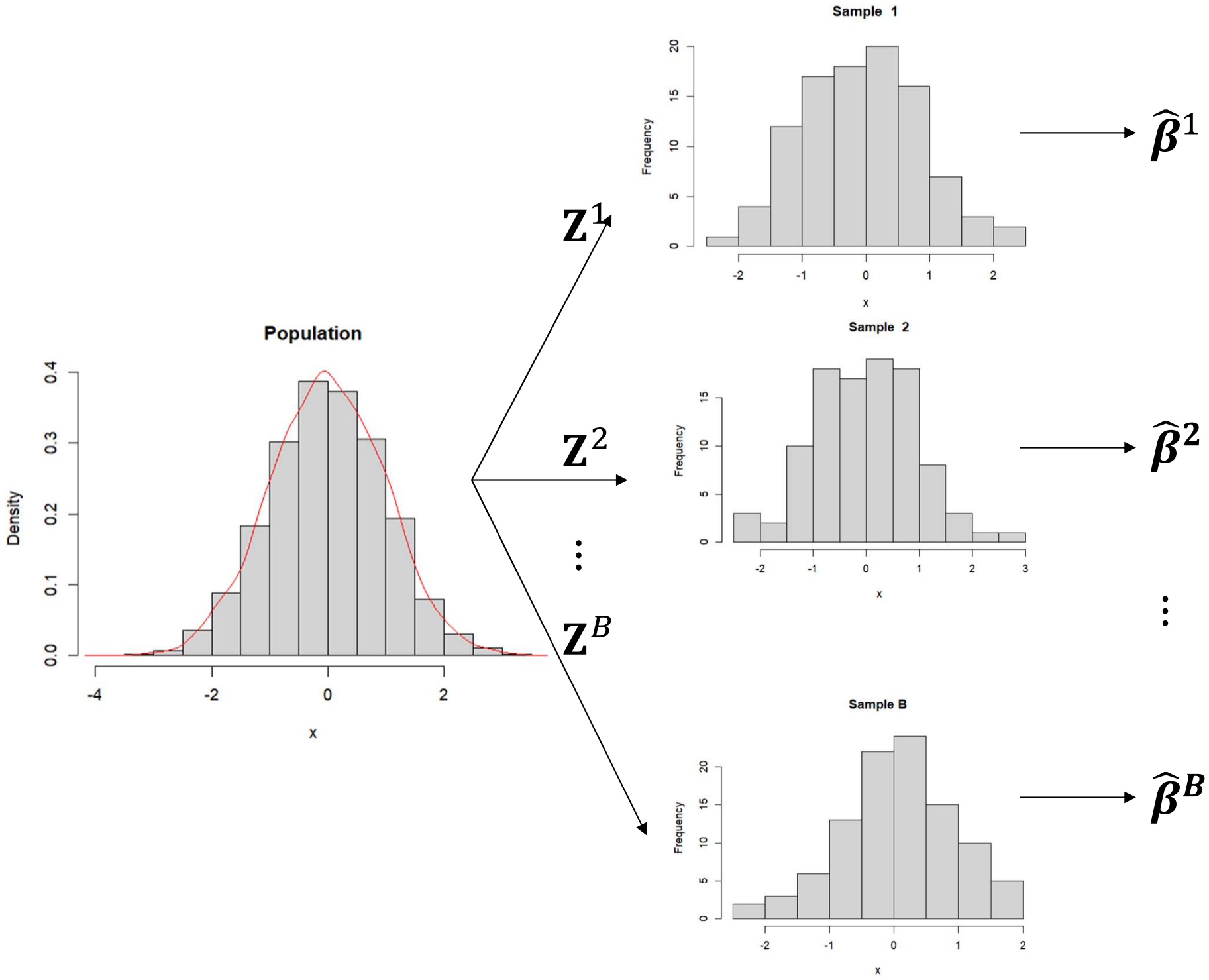
- The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given statistical method or estimator.

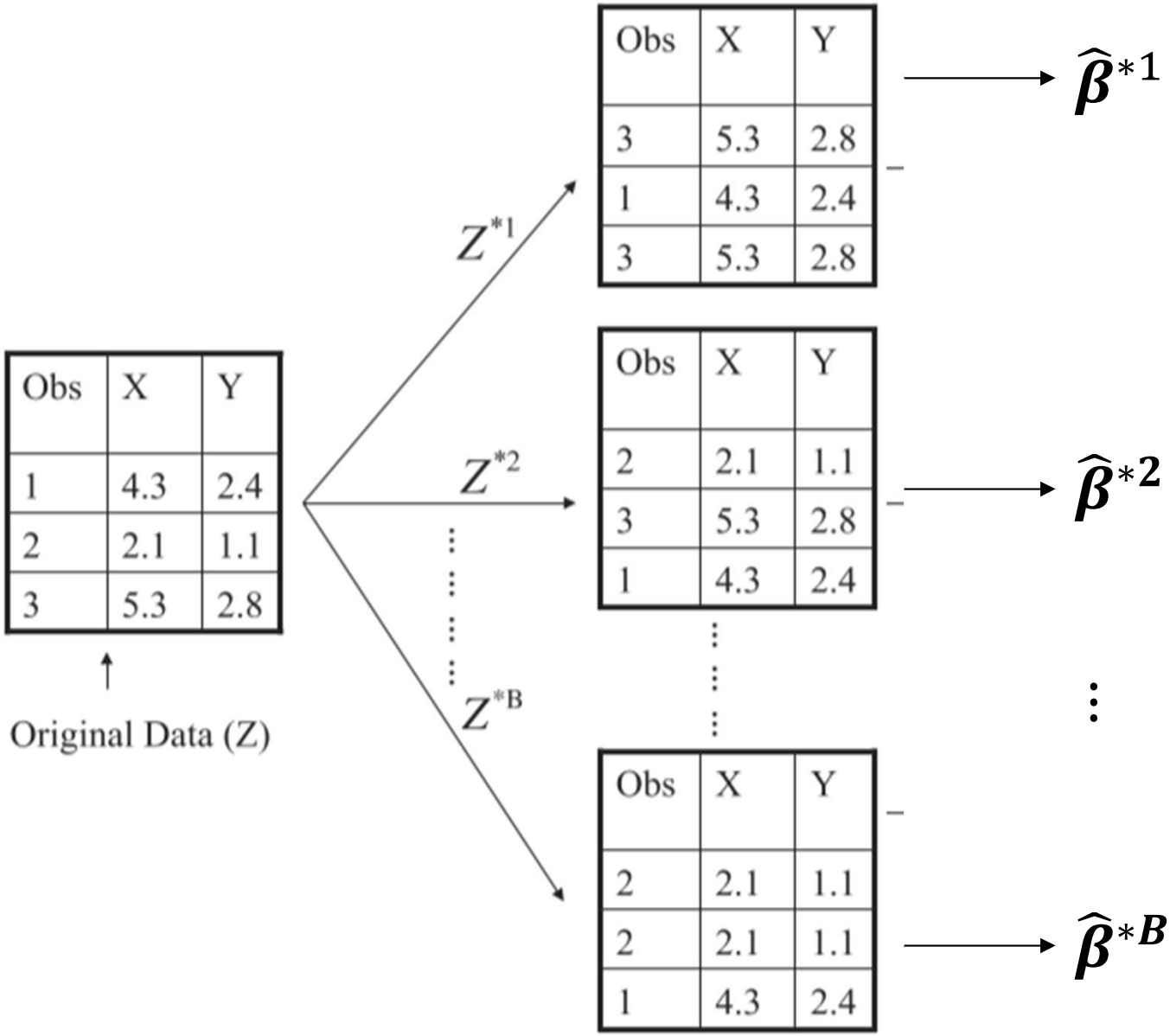
- For example, it can be used to estimate the standard errors of parameter estimates or their confidence intervals.



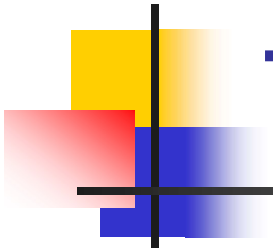
The Bootstrap

- The bootstrap allows us to use a computer to emulate the process of obtaining new sample sets, so that we can estimate the variability of a parameter estimate without generating additional samples.


Population
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
x
$\mathbf{Z}^1$
$\mathbf{Z}^2$
$\mathbf{Z}^B$
Sample 1
Frequency
0
5
10
15
20
-2
-1
0
1
2
x
$\widehat{\boldsymbol{\beta}}^1$
Sample 2
Frequency
0
5
10
15
-2
-1
0
1
2
3
x
$\widehat{\boldsymbol{\beta}}^2$
Sample B
Frequency
0
5
10
15
20
-2
-1
0
1
2
x
$\widehat{\boldsymbol{\beta}}^B$



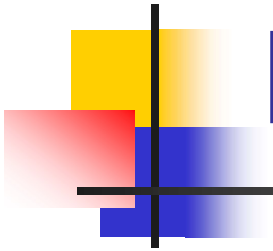
Revised from Figure 5.11 in James, Witten, Hastie, & Tibshirani, & Friedman (2013, p. 212)



The Bootstrap

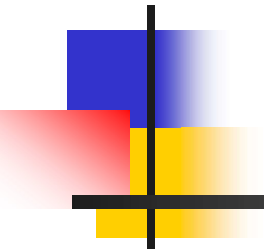
- The bootstrap allows us to use a computer to emulate the process of obtaining new sample sets, so that we can estimate the variability of a parameter estimate without generating additional samples.

- Rather than repeatedly obtaining independent data sets from the population, the bootstrap obtains distinct data sets by repeatedly sampling observations from the original data set "with replacement."



Reference

- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (2nd ed.). Springer.

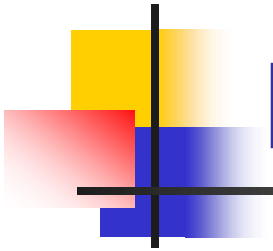


Lab: Resampling Methods

Session 5
Regularization Methods

An Introduction to Machine Learning for the Behavioural and Social Sciences

Heungsun Hwang & Gyeongcheol Cho



Linear Regression Model

- In the regression setting, the standard linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P + e$$

is commonly used to describe the relationship between a response Y and a set of predictors.

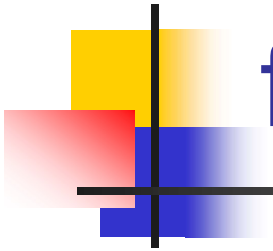
How to estimate coefficients?

- We typically use least squares to estimate the coefficients in the regression model.
  - This method chooses the values of the intercept and slopes that make the sum of the squared residuals as small as possible.
  - In other words, the coefficients are estimated to minimize

$$\text{SS(Residual)} = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

$$= \sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \cdots - \hat{\beta}_P X_P)^2$$

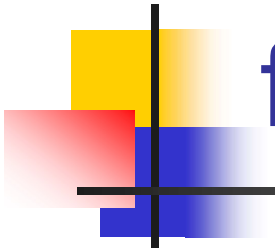
Alternative Fitting Procedures for Linear Models



- We will discuss some ways of improving linear models by replacing least squares fitting with alternative fitting procedures.

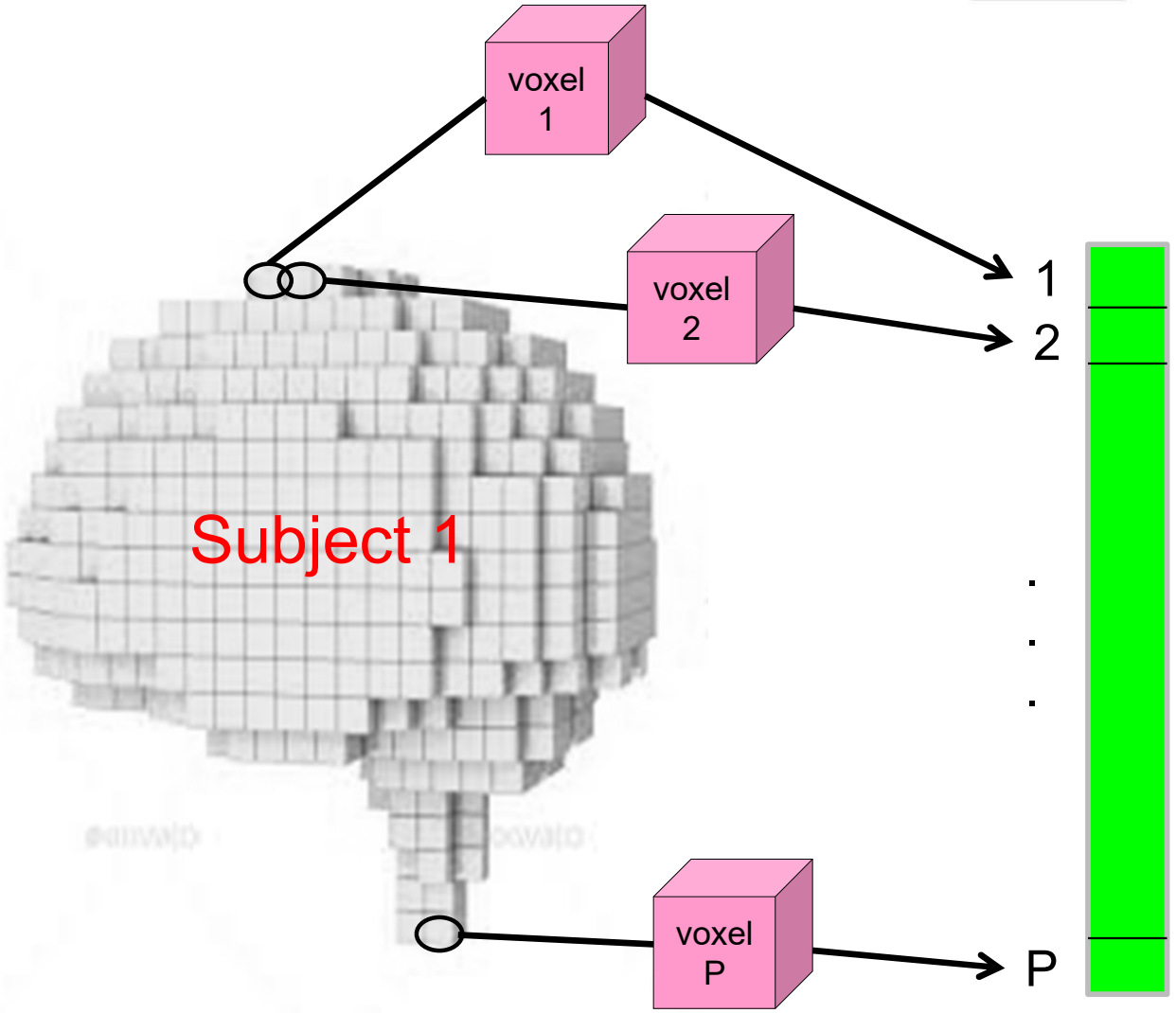
- Alternative fitting procedures may yield better prediction accuracy and/or model interpretability.

Alternative Fitting Procedures for Linear Models



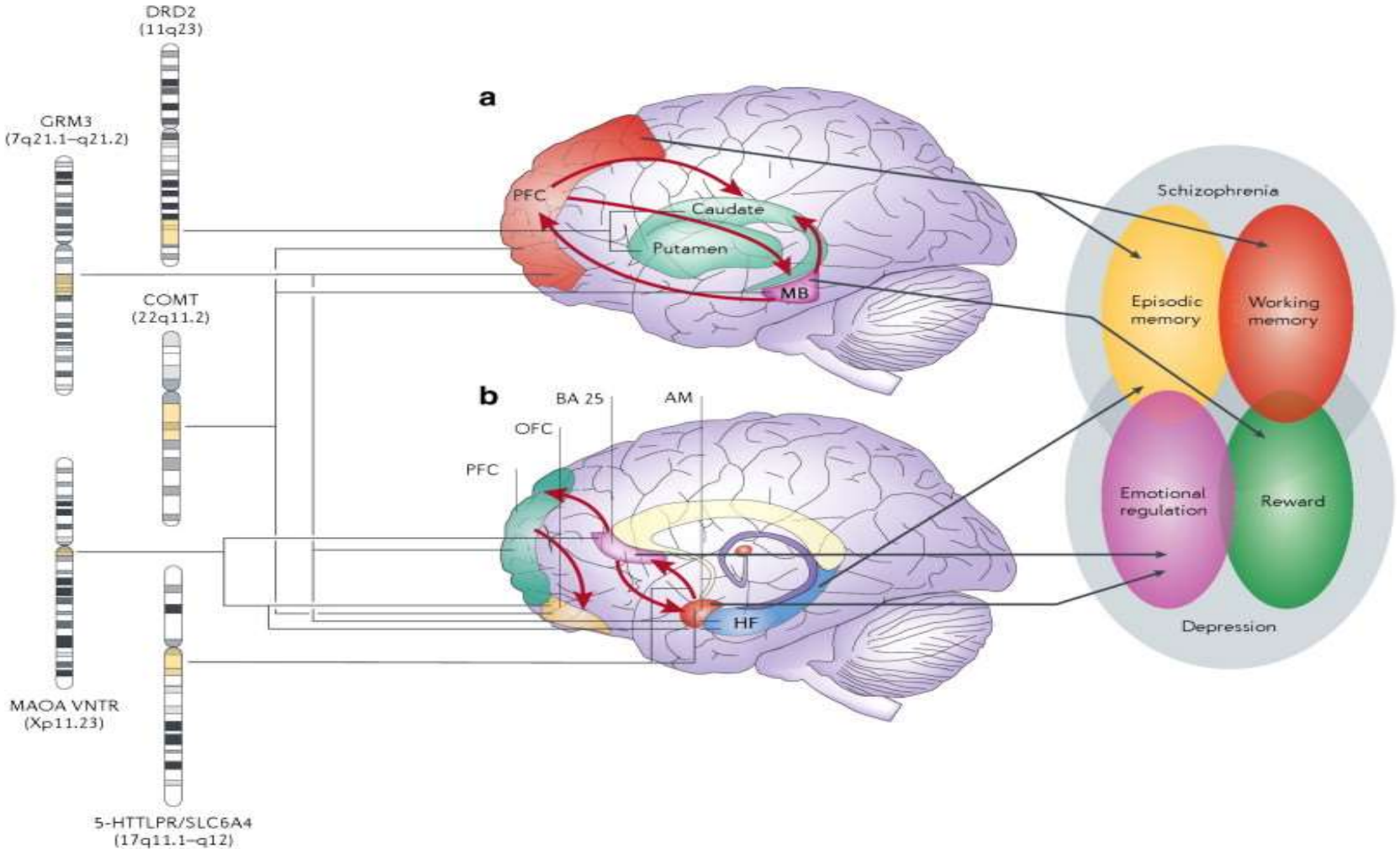
- Prediction accuracy: If sample size (N) is not much larger than the number of predictors (P), then there can be a lot of variability in the least squares fit, resulting in overfitting and consequently poor predictions on future observations. If P > N, there is no longer a unique least squares solution, so the method cannot be used at all.

High-Dimensional Data



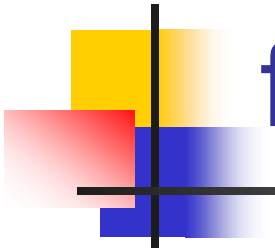
- e.g., P = 23,621 voxels and N = 4 subjects (Hwang et al., 2012)

High-Dimensional Data



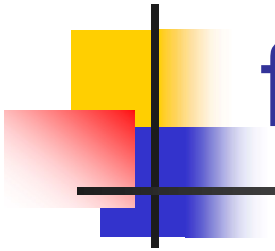
Tost et al. (2012, *Neuroimage*)

Alternative Fitting Procedures for Linear Models

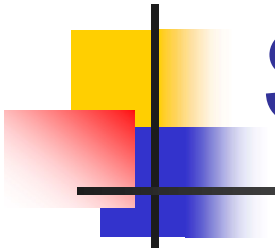


- Model interpretability: Including irrelevant variables leads to unnecessary complexity in the model. By removing these variables (i.e., by setting the corresponding coefficient estimates to zero), we can obtain a model that is more easily interpreted. Least squares is extremely unlikely to lead any coefficient estimates to be exactly zero. Excluding irrelevant variables is called variable selection or feature selection.

Alternative Fitting Procedures for Linear Models

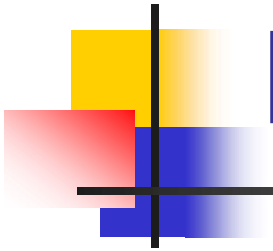


- We will discuss an alternative to using least squares.
  - Shrinkage (Regularization): This approach involves fitting a model involving all predictors. However, the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage has the effect of reducing variance. A type of shrinkage can also lead some coefficient estimates to be exactly zero (variable selection).



Shrinkage Methods

- We fit a model containing all P predictors using a technique that constraints or regularizes the coefficient estimates, or equivalently, that shrinks the estimates towards zero.
- Shrinking the coefficient estimates can significantly reduce their variance.
- Two best known shrinkage methods are ridge regression and the lasso.

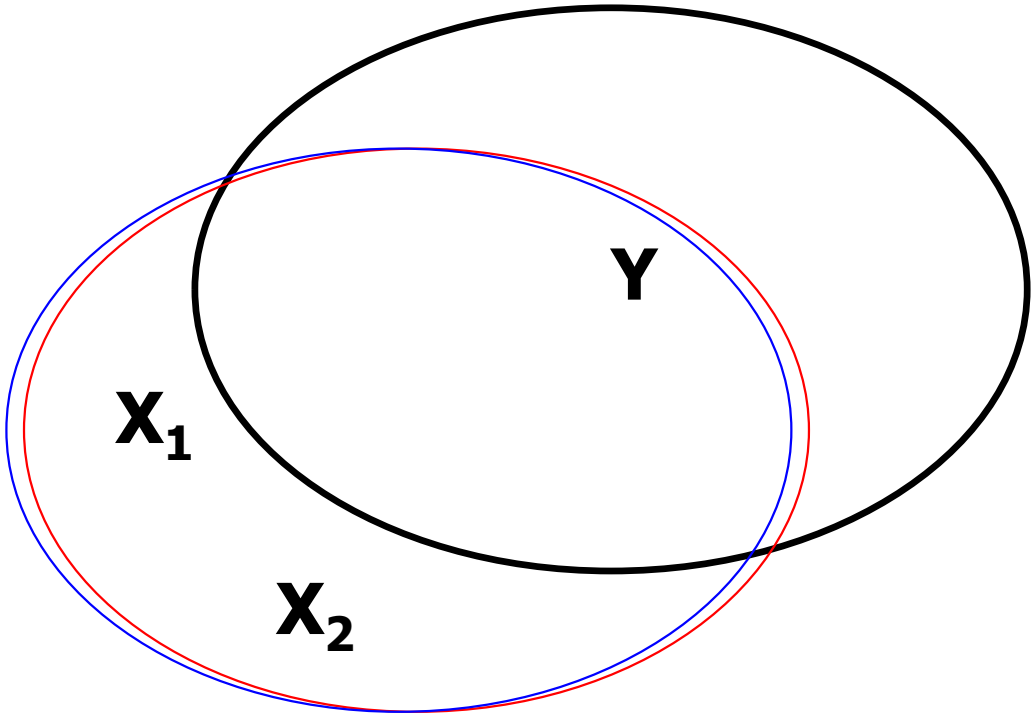


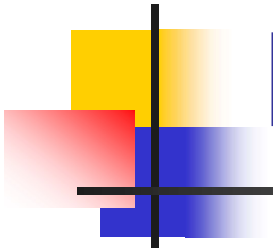
Ridge Regression

- Ridge regression was developed to address multicollinearity (Hoerl & Kennard, 1970).

- Multicollinearity, also termed collinearity or ill-conditioning, generally refers to a data problem that two or more predictors are highly correlated.

Multicollinearity





Multicollinearity

- Consequences:
  - Make it difficult to interpret the unique influence of a given predictor variable
  - Unstable regression coefficient estimates
    - Large variances of estimates (lower t ratios), leading them to be deviated farther from the true parameters on average
  - Unexpected signs of regression coefficient estimates
  - Computationally, the matrix inversion problem

Checking Multicollinearity

- Check by means of the correlation matrix of predictors
- Tolerance = $1 - R_p^2$
  - R^2 for the regression of one predictor on the other predictors, ignoring the DV. A high correlation among the predictors will lead the tolerance to approach zero.
  - Tolerance values < .1 indicate multicollinearity
- Variance Inflation Factor (VIF) = 1/Tolerance
  - VIF > 10 (Myers, 1990) suggests multicollinearity

Example: Multicollinearity

- Field's (2005) supermodel data (N = 231)
  - Salary: Salary per day on days when models were working
  - Age
  - Years: How many years they had worked as a model
  - Beauty: The attractiveness of each model as a percentage with 100% being perfectly attractive (rated by a panel of experts from modeling agencies)

Example: Multicollinearity

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	Collinearity Statistics	
							Tolerance	VIF
	(Intercept)	-60.890	16.497		-3.691	< .001		
	AGE	6.234	1.411	0.942	4.418	< .001	0.079	12.653
	YEARS	-5.561	2.122	-0.548	-2.621	0.009	0.082	12.157
	BEAUTY	-0.196	0.152	-0.083	-1.289	0.199	0.867	1.153

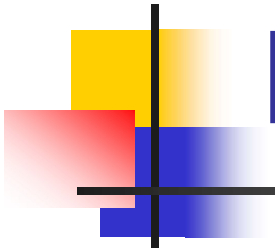
Ridge Regression

- Linear regression estimates the coefficients by minimizing the RSS or SS(Residual).

$$\mathrm{RSS} = \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{p=1}^{P}\beta_p x_{ip}\right)^2$$

- Ridge regression estimates the coefficients by minimizing a slightly different quantity.

$$\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{p=1}^{P}\beta_p x_{ip}\right)^2 + \lambda\sum_{p=1}^{P}\beta_p^2$$



Ridge Regression

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

- However, the second term, called a shrinkage (or ridge, L_2-norm, or quadratic) penalty, has the effect of shrinking their estimates towards zero.

Ridge Regression

- The tuning parameter λ controls for the relative impact of these two terms on the estimation of coefficients.
- When λ = 0, the penalty term has no effect, and ridge regression will produce the LS estimates.
- As λ → ∞, the term's impact grows, and the ridge regression estimates will approach zero.
- Ridge regression will produce a different set of coefficient estimates for each value of λ. Selecting a good value of λ is critical (cross validation).

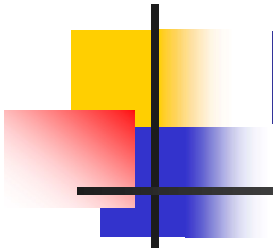
Why Does Ridge Regression Improve over Least Squares?

- To give further insight into how ridge regression works, we can consider an alternative way of obtaining ridge regression estimates. That is, the same problem can be solved by minimizing the RSS

$$\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{p=1}^{P}\beta_p x_{ip}\right)^2$$

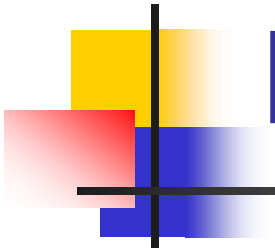
subject to $\sum_{p=1}^{P}\beta_p^2 \leq \tau$ (Hastie et al., 2001, p. 59).

Why Does Ridge Regression Improve over Least Squares?



- The size constraint ($\sum_{p=1}^{P} \beta_p^2 \leq \tau$) is imposed on the regression coefficients.

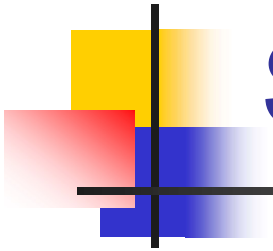
- When two predictor variables are highly correlated, a very large positive regression coefficient of one predictor variable can be offset by a very large negative regression coefficient of the other predictor variable. By imposing the size constraint, ridge regression keeps the magnitudes of the regression coefficients within a certain range.



Ridge Regression

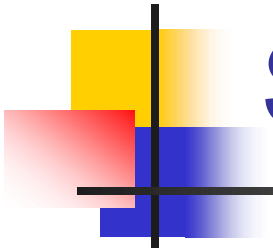
- Note that the shrinkage penalty is applied to $\beta_1, \ldots \beta_p$, but not the intercept β_0.
- We want to shrink the estimated association of each predictor with the response. However, we do not want to shrink the intercept, which is simply a measure of the mean value of the response when all predictors are zero.
- For any fixed value of λ, ridge regression is computationally as efficient as least squares.

Ridge Regression vs. Least Squares



- Ridge regression's advantages over least squares is rooted in the bias-variance trade-off.
- As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.
- In general, in situations when the relationship between the response and the predictors is close to linear, the least squares estimates will have low bias but may have high variance.

Ridge Regression vs. Least Squares



- Ridge regression works best when the least squares estimates have high variance, e.g., when the number of predictors (P) is almost as large as the number of observations (N) or when there is a high level of multicollinearity, the least squares estimates will be extremely variable.

- If P > N, the least squares estimates do not even have a unique solution.

Example: Ridge Regression

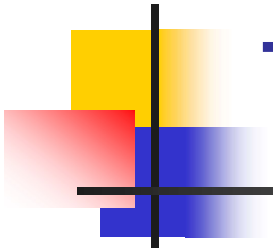
- The Major League Baseball data (Hitters.csv) that include records and salaries for baseball players in the 1986 and 1987 seasons
- DV:
  - Salary: 1987 annual salary on opening day in thousands of dollars
- Predictors:

- AtBat: # of times at bat in 1986
- Hits: # of hits in 1986
- HmRun: # of home runs in 1986
- Runs: # of runs in 1986
- RBI: # of runs batted in in 1986
- Walks: # of walks in 1986
- Years: # of years in the major leagues
- CAtBat: # of times at bat during his career
- CHits: # of hits during his career
- CHmRun: # of home runs during his career
- CRuns: # of runs during his career
- CRBI: # of runs batted in during his career
- CWalks: # of walks during his career
- League: player's league at the end of 1986 (A and N)
- Division: player's division at the end of 1986 (E and W)
- PutOuts: # of put outs in 1986
- Assists: # of assists in 1986
- Errors: # of errors in 1986
- NewLeague: player's league at the beginning of 1987 (A and N)

https://cran.r-project.org/web/packages/ISLR/ISLR.pdf

Example: Ridge Regression

	lam=10^9	lam=10^6	lam=10^3	lam=10^0
(Intercept)	5.501681e+02	5.479665e+02	87.08331332	211.29676862
AtBat	6.950346e-07	6.947008e-04	0.12316656	-2.52020209
Hits	2.464619e-06	2.462385e-03	0.62191426	9.08804182
HmRun	9.792868e-06	9.781230e-03	1.27359344	-3.55051191
Runs	3.944192e-06	3.937542e-03	0.79695583	-1.56243688
RBI	4.114842e-06	4.107052e-03	0.83256809	0.28953783
Walks	5.009139e-06	4.999945e-03	1.25990685	5.81827123
Years	1.969997e-05	1.968102e-02	2.46487778	-2.22774082
CAtBat	5.565417e-08	5.556640e-05	0.01031190	-0.15450071
CHits	2.080608e-07	2.076415e-04	0.04450730	0.05660042
CHmRun	1.617496e-06	1.613645e-03	0.39097718	0.96365607
CRuns	4.195439e-07	4.183310e-04	0.09052010	1.14332672
CRBI	4.272296e-07	4.258274e-04	0.10079606	0.86889954
CWalks	4.497039e-07	4.479279e-04	0.07525672	-0.72653672
League	-1.479685e-05	-1.463203e-02	8.89517753	26.05212188
Division	-1.039619e-04	-1.037738e-01	-51.15351556	-124.87961918
PutOuts	3.153370e-07	3.147069e-04	0.14430748	0.41476962
Assists	1.294450e-07	1.291365e-04	0.06009958	0.68090169
Errors	3.179735e-07	3.149243e-04	-0.35207093	-5.48217814
NewLeague	-2.225218e-07	-1.045981e-04	12.92124231	36.43239183



The Lasso (Lasso Regression)

- Ridge regression keeps all P predictors and produces their coefficient estimates.
- The ridge penalty will shrink all coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$).
- This may not be a problem for prediction accuracy. But, it can create a challenge in model interpretation when P is quite large.
- The lasso is an alternative to ridge regression that overcomes this disadvantage.

The Lasso

- The Lasso estimates the coefficients by minimizing the following quantity.

$$\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{p=1}^{P}\beta_p x_{ip}\right)^2 + \lambda\sum_{p=1}^{P}|\beta_p|$$

- As compared to ridge regression, the lasso replaces the ridge penalty with the lasso (or L_1-norm) penalty.

- This L1 penalty has the effect of forcing some of the coefficient estimates to be exactly zero when λ is sufficiently large.

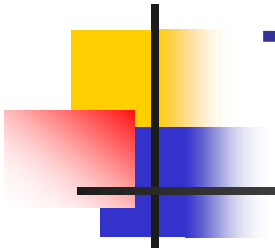
The Lasso

- The lasso problem is equivalent to minimizing the RSS

$$\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{p=1}^{P}\beta_p x_{ip}\right)^2$$

subject to $\sum_{p=1}^{P}|\beta_p| \leq \tau$.

- When we perform the lasso, we are trying to find the set of coefficient estimates that lead to the smallest RSS, subject to the size constraint.

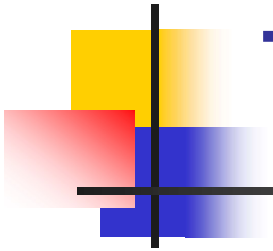


The Lasso

- When τ is very large, then this constraint is not difficult to satisfy, and the coefficient estimates can be large.

- In fact, if τ is large enough, then the lasso will simply yield the least squares solution.
  - If τ is greater than the sum of the least squares estimates in absolute value, the constraint has no effect.

- If τ is small, some of coefficient estimates are to be exact zero to satisfy the constraint.

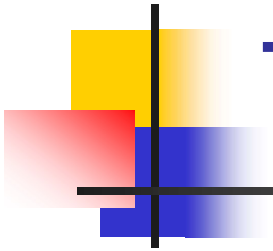


The Lasso

- Thus, the lasso performs variable selection. That is, it tends to produce simpler models than those produced by ridge regression, improving interpretability.
- We say that the lasso yields sparse models, i.e., models that involve only a subset of predictors.
- As in ridge regression, selecting a good value of λ is critical (cross validation).

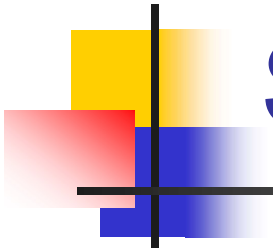
Example: The Lasso

	lam=10^1	lam=10^0	lam=10^-1	lam=10^-2
(Intercept)	74.4405180	205.1015733	208.86737289	209.14451375
AtBat	.	-2.4740268	-2.65866553	-2.66575403
Hits	2.0177886	8.8966278	9.77462213	9.79731222
HmRun	-0.5976880	-3.0099117	-2.44009624	-2.50739018
Runs	.	-1.0495697	-2.23773090	-2.25485461
RBI	.	.	.	0.02416709
Walks	1.9368036	5.5398404	6.06865524	6.08937697
Years	.	-1.7990503	.	0.03521957
CAtBat	.	-0.1365554	-0.18797197	-0.18810447
CHits	.	.	-0.01301876	-0.01446144
CHmRun	0.6330213	0.7567654	0.55116389	0.55651867
CRuns	.	1.0626726	1.44962706	1.45111478
CRBI	0.5258825	0.9470540	1.04269728	1.04505266
CWalks	.	-0.6837595	-0.78845582	-0.79163410
League	10.2640608	23.2187281	23.77164154	23.95545730
Division	-132.9306928	-125.0830367	-121.65380042	-121.82959313
PutOuts	0.3513209	0.4107125	0.41818134	0.41835661
Assists	0.1499942	0.6380844	0.70811357	0.71164465
Errors	-0.2243985	-4.7520735	-5.40594651	-5.46737391
NewLeague	24.9382535	37.0937356	40.58121438	40.63039496



The Lasso vs. Ridge Regression

- The lasso has an advantage over ridge regression, in that it can produce simpler and more interpretable models that involve only a subset of predictors.
- In terms of prediction accuracy, either of the two methods will not universally dominate the other. In general, the lasso is expected to perform better when a relatively small number of predictors have substantially large, whereas ridge regression is to perform better when all coefficients are of roughly equal size.
- CV can be used to compare the two methods on a particular data set.



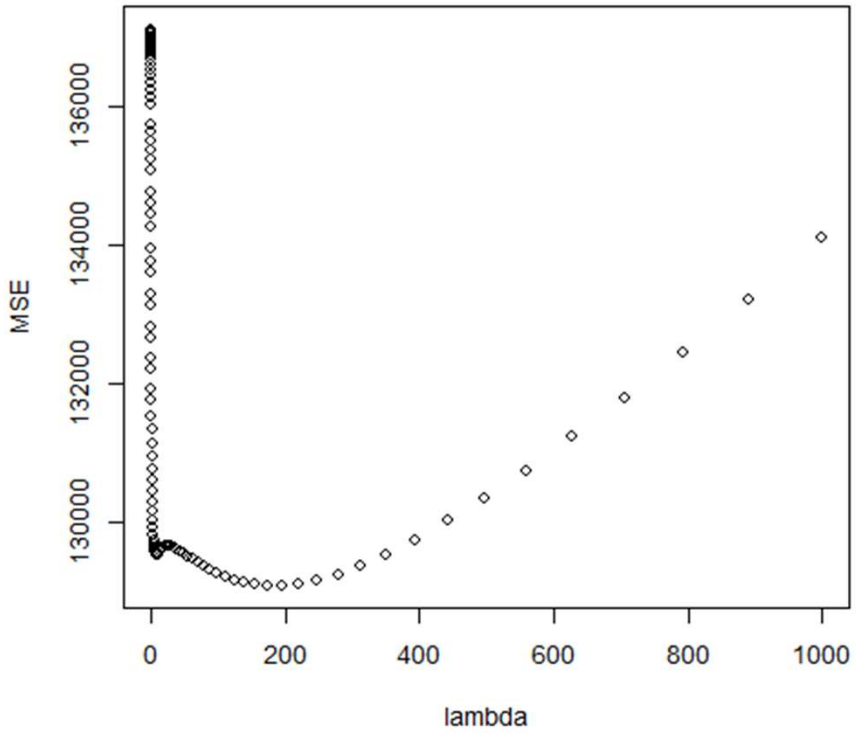
Selecting the Tuning Parameter

- Cross validation provides a simple way of selecting a value for λ in ridge regression and the lasso.
- We choose a grid of λ values and compute the cross-validation error for each value of λ. We then select the tuning parameter for which the cross-validation error is smallest.
- Finally, the model is re-fit using all the available observations and the selected value of the tuning parameter.

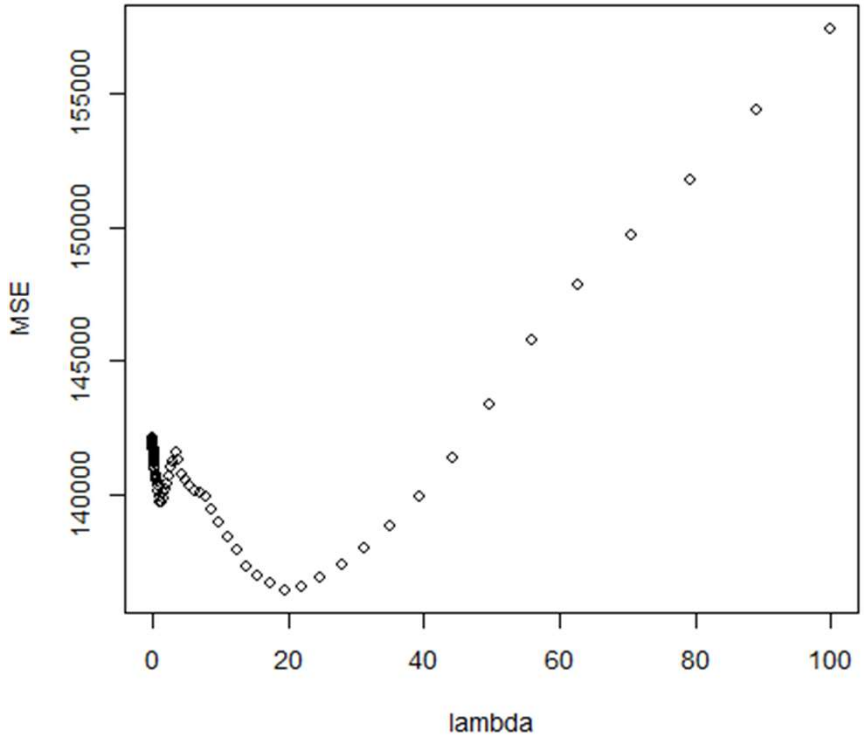
Example: Cross Validation for Ridge and Lasso

Hitters_training.csv

Ridge



Lasso



Example: Ridge and Lasso

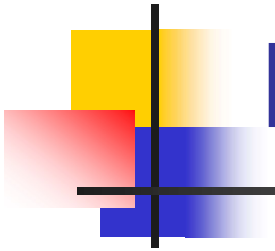
Hitters_training.csv

	Ridge	Lasso	No Reg
(Intercept)	5.445742e+01	114.90815640	152.20667028
AtBat	-1.036336e-02	.	-2.73455318
Hits	1.180396e+00	1.88302825	10.16845282
HmRun	-9.543557e-01	.	-2.04325704
Runs	8.032000e-01	.	-2.55611277
RBI	9.660890e-01	.	-0.14001318
Walks	1.932747e+00	1.92631459	6.23602955
Years	-1.997343e+00	.	0.01603933
CAtBat	7.559330e-03	.	-0.16397743
CHits	6.447880e-02	.	-0.19808368
CHmRun	7.511652e-01	0.32265343	0.20762834
CRuns	1.364149e-01	.	1.60133635
CRBI	1.846534e-01	0.57158603	1.20316567
CWalks	-3.657721e-03	.	-0.84351392
League	2.043242e+01	.	20.82926895
Division	-1.084553e+02	-114.19332716	-122.31543710
PutOuts	2.985322e-01	0.32270596	0.41945478
Assists	1.917919e-01	0.05638781	0.70481669
Errors	-2.046842e+00	.	-5.30400295
NewLeague	2.690478e+01	9.22924278	43.79582769

Example: Ridge and Lasso

Hitters_test.csv

MSE_Ridge:	94256.1
MSE_Lasso:	97640.2
MSE_No Reg:	116358.2



Lab

- Ridge Regression
- Lasso