Multiple Testing & Resampling

December 29, 2024

Table of Contents

The Challenge of Multiple Testing

The Challenge of Multiple Testing

- Given one Hypothesis, we know that α (the significance level) controls the probability for a Type I error.
- Assume we want to test m null hypothesis at $\alpha = 0.01$
- If we reject all null hypotheses for which the p-value falls below 0.01, then how many Type I errors should we expect to make?

The Challenge of Multiple Testing

- Assume we flip 1024 fair coins ten times each and define H_0 : The coin is fair
- Then one possible outcome is that one coin comes up all tails (Probability: $\frac{1}{2^{10}} = \frac{1}{1024}$ that a single coin will come up all tails.)
- If one coin comes up all tails, then we might suspect that this coin is *not* fair (p-Value for such a HT: 0.002, Thus we reject H_0).
- **BUT**: It would be incorrect to conclude that the coin is *not* fair, as we just have gotten ten tails in a row by chance.
- The challenge of Multiple Testing:

  The more you search, the more you find, purely by chance!

- Note: Type I errors are considered more harmful than Type II errors, as they are connected with discoveries and subsequently with more/further time and effort wrt. research.

The Challenge of Multiple Testing

- In more general terms...
  - When testing a huge number of null hypotheses, we are bound to get some very small p-values by chance! (e.g. 10 tails in a row)
  - Thus we might end up rejecting a large number of Hypotheses
  - Assume $m = 10,000$ hypotheses, when $\alpha = 0.01$, then we expect to falsely reject 100 null hypothesis. That are a lot of Type I errors.

The Challenge of Multiple Testing

	H_0 is True	H_0 is False	Total
Reject H_0	V	S	R
Do not reject H_0	U	W	$m - R$
Total	m_0	$m - m_0$	m

Assume m null hypotheses:

- A given null hypothesis is either true or false, and a test of that null hypothesis can either reject or fail to reject it. Then let:
  - V: Number of Type I Errors
  - W: Number of Type II Errors
  - U and S: Number of correct decisions.
  - Note: In practice, the individual values of V, S, U, and W are unknown.
- **BUT**, we have access to $R = V + S$ and $m - R = U + W$, which are the numbers of null hypotheses rejected and not rejected, respectively.

Questions

1. Explain the meaning of a Type I error.
2. Explain the meaning of a Type II error.
3. What happens to the Type I error and the Type II error, respectively, when you decrease your significance level α?

The Family-Wise Error Rate

The Family-Wise Error Rate

- The family-wise error rate (FWER) is the probability of making at least one Type I error: FWER $= Pr(V \geq 1)$
- Controlling each null hypothesis at level α, leads to:

$$\begin{aligned}\text{FWER} &= 1 - Pr(V = 0) = 1 - Pr(\text{do not falsely reject any null hypothesis}) \\ &= 1 - Pr\left(\bigcap_{j=1}^{m} (\text{do not falsely reject } H_{0j})\right)\end{aligned}$$

- Assuming independence among the hypothesis tests:

$$\text{FWER}(\alpha) = 1 - \prod_{j=1}^{m}(1-\alpha) = 1 - (1-\alpha)^m$$

- Note:
  - For one HT: $\text{FWER}(\alpha) = 1 - (1-\alpha)^1 = \alpha$
  - For $m = 100$ and $\alpha = 0.05$: $\text{FWER}(\alpha) = 1 - (1-0.05)^{100} = 0.994$.
  - We are virtually guaranteed to make at least one Type I error.

The Family-Wise Error Rate

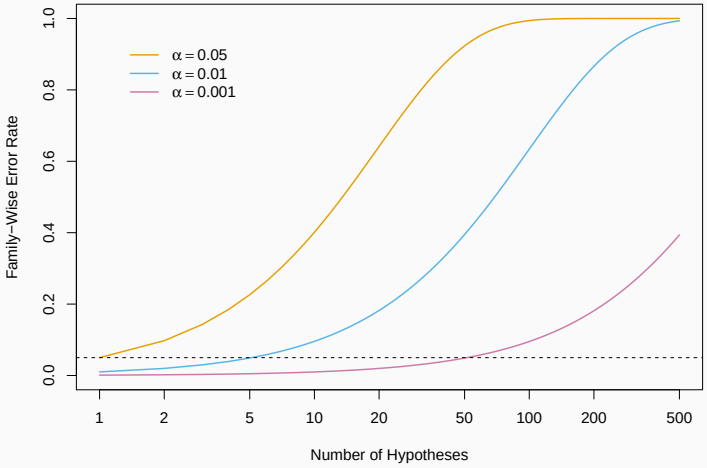

Figure 1: The family-wise error rate, as a function of the number of hypotheses tested (displayed on the log scale), for three values of α : $\alpha = 0.05$ (orange), $\alpha = 0.01$ (blue), and $\alpha = 0.001$ (purple). The dashed line indicates 0.05. For example, in order to control the FWER at 0.05 when testing $m = 50$ null hypotheses, we must control the Type I error for each null hypothesis at level $= 0.001$.

The Family-Wise Error Rate

- How to control the FWER?
  1. The Bonferroni method
  2. Holm's Step-Down Procedure
- Both can be applied whenever m p-values have been computed, independent of the form of hypotheses, the choice of test statistics or the independence of the p-values.

The Family-Wise Error Rate – The Bonferroni method

- Let A_j be the event that we make a Type I error for the jth null hypothesis:

$$\begin{aligned}\text{FWER}(\alpha) &= Pr(\text{falsely reject at least one null hypothesis}) \\ &= Pr(\bigcup_{j=1}^{m} A_j) \\ &\leq \sum_{j=1}^{m} Pr(A_j)\end{aligned}$$

- Note: $Pr(A \cup B) \leq Pr(A) + Pr(B)$ regardless of whether A and B are independent.
- The *Bonferroni* method then sets the threshold for rejecting each hypothesis test to $\frac{\alpha}{m}$, such that: $Pr(A_j) \leq \frac{\alpha}{m}$:

$$FWER(\alpha) \leq m \times \frac{\alpha}{m} = \alpha$$

- For example: To control the FWER at level 0.1, while testing $m = 100$ null hypotheses, Bonferroni requires us to reject all null hypotheses for which the p-value is below 0.001.

Manager	Mean, $\bar{x}$	Standard Deviations	t-statistic	p-value
One	3.0	7.4	2.86	0.006
Two	-0.1	6.9	-0.10	0.918
Three	2.8	7.5	2.62	0.012
Four	0.5	6.7	0.53	0.601
Five	0.3	6.8	0.31	0.756

Table 1: Summary statistics for the first five managers in the Fund dataset. The last two columns provide the t-statistic and associated p-value for testing $H_{0j} : \mu_j = 0$, the null hypothesis that the (population) mean return for the *jth* hedge fund manager equals zero.

Questions:

a) Define the FWER.

b) Which manager fail to show statistically significant returns at $\alpha = 5\%$, when you test each manager independently?

c) Use the Bonferroni method at $\alpha = 5\%$ to test, whether any of the managers show statistically significant returns.

Holm's Step-Down Procedure

- Bonferroni can be quite conservative (true FWER might be lower than target FWER).
- Holm is less restrictive in that it rejects more null hypotheses, resulting in fewer Type II errors.

Algorithm 13.1: *Holm's Step-Down Procedure to Control the FWER*

1. Specify α at the level at which to control the FWER
2. Compute p-values $p_1, ..., p_m$ for the m null hypotheses $H_{01}, ..., H_{0m}$
3. Order the m p-values so that $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$
4. Define: $L = \min\{j : p_{(j)} > \frac{\alpha}{m+1-j}\}$
5. Reject all null hypotheses H_{0j} for which $p_j < p_{(L)}$

Holm's Step Down Procedure – Manager Example

Manager	Mean, $\bar{x}$	Standard deviation	t-statistic	p-value
One	3.0	7.4	2.86	0.006
Two	−0.1	6.9	−0.10	0.918
Three	2.8	7.5	2.62	0.012
Four	0.5	6.7	0.53	0.601
Five	0.3	6.8	0.31	0.756

Table 2: Summary statistics for the first five managers in the Fund dataset. The last two columns provide the t-statistic and associated p-value for testing $H_{0j} : \mu_j = 0$, the null hypothesis that the (population) mean return for the *jth* hedge fund manager equals zero.

a) Use Holm's Step Down procedure at $\alpha = 5\%$ to test whether any of the managers show statistically significant returns.

b) Compare and contrast Holm's Step Down procedure with the Bonferroni method.

The Family-Wise Error Rate

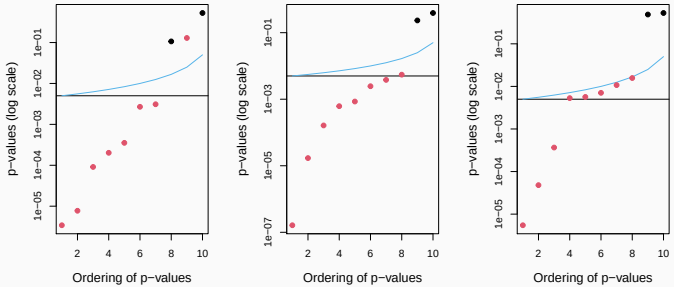

Figure 2: Each panel displays, for a separate simulation, the sorted p-values for tests of $m = 10$ null hypotheses. The p-values corresponding to the $m_0 = 2$ true null hypotheses are displayed in black, and the rest are in red. When controlling the FWER at level 0.05, the Bonferroni procedure rejects all null hypotheses that fall below the black line, and the Holm procedure rejects all null hypotheses that fall below the blue line. The region between the blue and black lines indicates null hypotheses that are rejected using the Holm procedure but not using the Bonferroni procedure. In the center panel, the Holm procedure rejects one more null hypothesis than the Bonferroni procedure. In the right-hand panel, it rejects five more null hypotheses.

The Family-Wise Error Rate

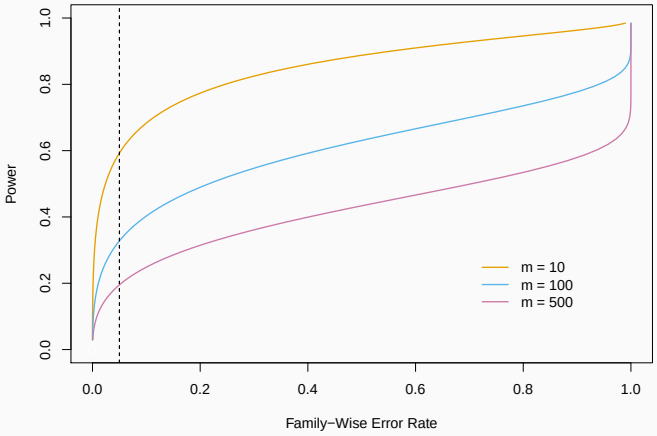

Figure 3: In a simulation setting in which 90% of the m null hypotheses are true, we display the power (the fraction of false null hypotheses that we successfully reject) as a function of the family-wise error rate. The curves correspond to $m = 10$ (orange), $m = 100$ (blue), and $m = 500$ (purple). As the value of m increases, the power decreases. The vertical dashed line indicates a FWER of 0.05.

The False Discovery Rate

The False Discovery Rate

- Controlling the FWER at level α helps ensure that we do not reject any true null hypotheses. This might be too stringent.
- We might want to tolerate a few false positives in the interest of making more discoveries.
- We might try to make sure that the ratio of False Positives (V) to Total Positives ($R = V + S$) is sufficiently low, so that most of the rejected null hypotheses are not false positives.

The False Discovery Rate

- To understand the *false discovery rate*, let us first look at the *false discovery proportion* (FDP):

$$\text{FDP} = \frac{\text{False positives}}{\text{Total positives}} = \frac{V}{R} = \frac{V}{V+S}$$

- Example: FDP $= 20\% \Rightarrow$ No more than 20% of the rejected null hypotheses are false positives.
- **BUT**: The data analyst might guarantee that $Pr(V \geq 1) \leq \alpha$ for any pre-specified α, but she cannot guarantee $V = 0$ on any dataset.
- Thus, we control the False Discovery Rate (FDR) instead:

$$\text{FDR} = \mathbb{E}(\text{FDP}) = \mathbb{E}(\frac{V}{R})$$

- We say, we control the FDR at the level q, whereas $q \in [0, 1]$.
- Note: The expected value is taken over the population.

The False Discovery Rate

- For instance, suppose we control the FDR for m null hypothesis at $q = 0.2$.
- If we repeat the experiment a huge number of times and each time control the FDR at $q = 0.2$, then we should expect that, on average 20% of the rejected null hypotheses will be false positives.
- How to control the FDR?
  - Benjamini Hochberg (BH) procedure (around since mid-1990s)

The False Discovery Rate

Algorithm 13.2: *BH Procedure to control the FDR*

1. Specify q, the level at which to control the FDR.
2. Compute p-vales $p_1, ..., p_m$ for the m null hypotheses $H_{01}, ..., H_{0m}$
3. Order the m p-values so that $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$
4. Define: $L = \min\{j : p_{(j)} < \frac{q \times j}{m}\}$
5. Reject all null hypotheses H_{0j} for which $p_j \leq p_{(L)}$

- As long as the m p-values are independent or only mildly dependent, BH guarantees that $FDR \leq q$.
- Holds independent of how many null hypotheses are true and regardless of the distribution of the p-values.
- Note:
  - Bonferroni is independent of the data
  - BH is data dependent (i.e. we reject the p-value that is less than or equal to the Lth smallest p-value.)

BH Procedure – Manager Example

Manager	Mean, $\bar{x}$	Standard deviation	t-statistic	p-value
One	3.0	7.4	2.86	0.006
Two	−0.1	6.9	−0.10	0.918
Three	2.8	7.5	2.62	0.012
Four	0.5	6.7	0.53	0.601
Five	0.3	6.8	0.31	0.756

Table 3: Summary statistics for the first five managers in the Fund dataset.

a) Use the Benjamini Hochberg (BH) Procedure to control the FDR at $q = 5\%$ and for testing the null hypothesis, whether any of the managers show statistically significant returns.

b) Compare and contrast the Bonferroni method with the Benjamini Hochberg method for dealing with the problem of multiple testing.

The False Discovery Rate

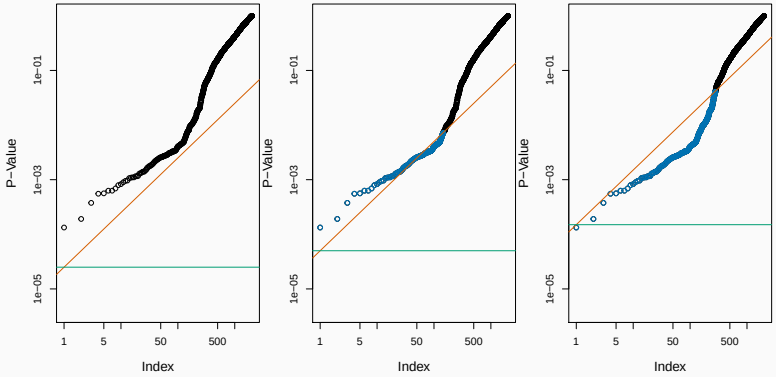

Figure 4: Each panel displays the same set of $m = 2,000$ ordered p-values for the Fund data. The green lines indicate the p-value thresholds corresponding to FWER control, via the Bonferroni procedure, at levels $\alpha = 0.05$ (left), $\alpha = 0.1$ (center), and $\alpha = 0.3$ (right). The orange lines indicate the p-value thresholds corresponding to FDR control, via Benjamini-Hochberg, at levels $q = 0.05$ (left), $q = 0.1$ (center), and $q = 0.3$ (right). When the FDR is controlled at level $q = 0.1$, 146 null hypotheses are rejected (center); the corresponding p-values are shown in blue. When the FDR is controlled at level $q = 0.3$, 279 null hypotheses are rejected (right); the corresponding p-values are shown in blue.

Resampling

Resampling

- In the section we discuss two resampling methods:
  1. Cross-Validation
  2. The Bootstrap
- These methods refit a model of interest to samples formed from the training set, in order to obtain additional information about the fitted model.
- For example, they provide estimates of test-set prediction error, and the standard deviation and bias of our parameter estimates

Resampling

- Recall the distinction between the test error and the training error:
  - *The test error* is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method.
  - In contrast, the *training error* can be easily calculated by applying the statistical learning method to the observations used in it's training.
  - But the training error rate is often quite different from the test error rate, and in particular the former can dramatically underestimate the latter.

Recall the analysis from Lecture 1...

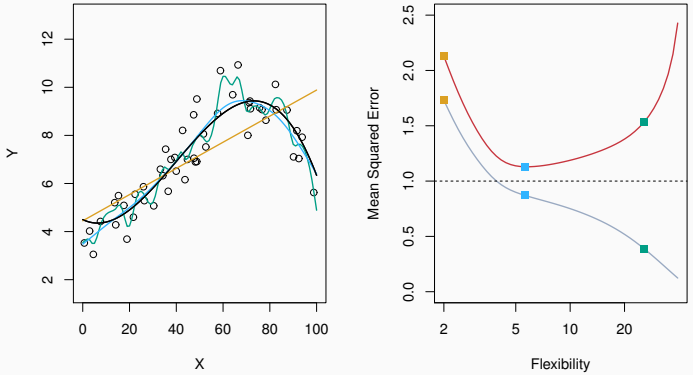

Figure 5: Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

Questions

1. Explain the U-Shape of the red test MSE curve.
2. What is the meaning of the horizontal grey dotted line in the figure on the right.
3. Do you always want to choose the method that provides you with the lowest training MSE? Why? Why not?

Cross Validation

Cross Validation

- Best solution: a large designated test set. However: Often simply not available.
- Some methods make a mathematical adjustment to the training error rate in order to estimate the test error rate. These include the Mallow's C_p statistic, the *Akaike Information Criteria* (AIC) and the *Bayesian Information Criteria* (BIC). They are discussed elsewhere in the course.
- Here we instead consider a class of 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

Validation-Set Approach

Validation-Set Approach

- Randomly divide the available set of samples into two parts: a training set and a *validation* or *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 provides an estimate of the test error. This is typically assessed using MSE in the case of a quantitative response and misclassification rate in the case of a qualitative (discrete) response.

Validation-Set Approach

Figure 6: A random splitting into two halves: left part is training set, right part is validation set

Validation-Set Approach

- Example: Automobile Data
  - We want to compare linear vs. higher-order polynomial terms in a linear regression
  - Randomly split the 392 observations into two sets, a training set containing 196 of the data points, and a validation set containing the remaining 196 observations.

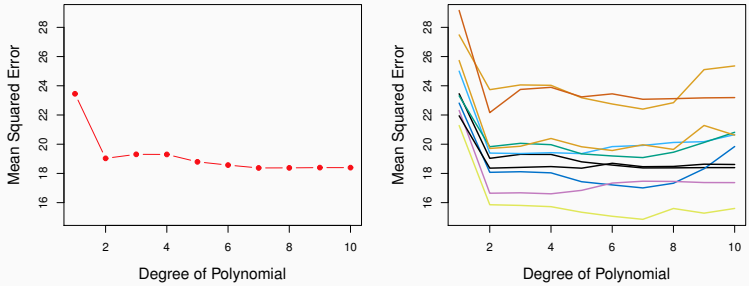

Figure 7: Left panel shows single split; right panel shows multiple splits.

Validation-Set Approach

- Drawbacks of validation set approach:
  - 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 included in the validation set.
  - In the validation approach, only a subset of the observations — those that are included in the training set rather than in the validation set — are used to fit the model.
  - This suggests that the validation set error may tend to overestimate the test error for the model fit on the entire data set. Why?

Cross Validation

K-Fold Cross-Validation

K-Fold Cross-Validation

- The idea is to randomly divide the data into K equal-sized parts. First, leave out part k and fit the model to the other $K-1$ parts (combined). For this fit, obtain the predictions and test MSE for the left-out kth part.
- This is done in turn for each part $k = 1, 2, ...K$, and then the average MSE across all K folds is computed. results are combined.
- Let the K parts be $C_1, C_2, ...C_K$, where C_k denotes the indices of the observations in part k. There are n_k observations in part k:

$$CV_{(K)} = \frac{1}{K} \sum_{k=1}^{K} MSE_k$$

  where $MSE_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$ and $\hat{y}_i$ is the fit for observation i, obtained from the data with part k removed.
- Setting $K = n$ yields n-fold or `leave-one out cross-validation` (LOOCV)

K-Fold Cross-Validation

- Divide data into K roughly equal-sized parts (here $K = n$)

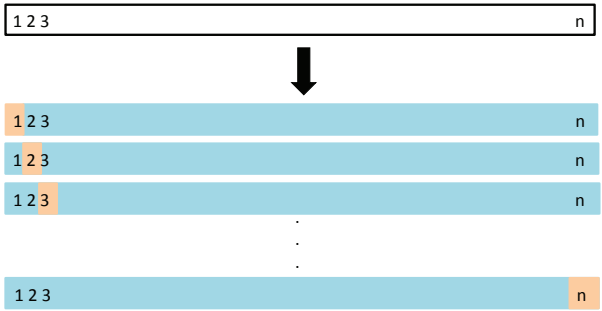

Figure 8: A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.

K-Fold Cross-Validation

- An example with $K = 5$:

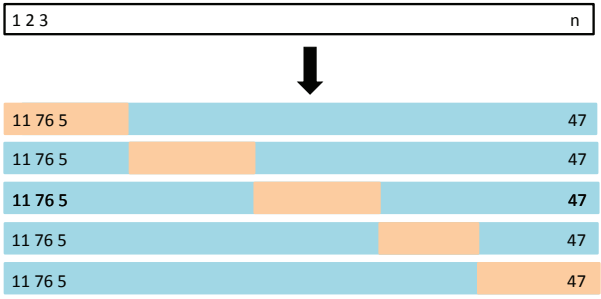

Figure 9: A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.

K-Fold Cross-Validation

Some notes about LOOCV and choosing a higher K-fold CV:

- Validation Set Approach (VSA): (a) Training set is typically around half the size of the original data set and (b) VSA will yield different results when applied repeatedly due to randomness in the training/validation set splits
- In LOOCV, we repeatedly fit the statistical learning method using training sets that contain $n - 1$ observations, almost as many as are in the entire data set.
- Consequence: (a) LOOCV tends not to overestimate the test error rate as much as the validation set approach does and (b) no randomness in the training/validation split.
- But: LOOCV has the potential to be expensive to implement, since the model has to be fit n times. This can be very time consuming if n is large, and if each individual model is slow to fit.
- Solution: Choosing $K = 5$ or $K = 10$ might serve as a better idea.

K-Fold Cross-Validation

- Auto data revisited:

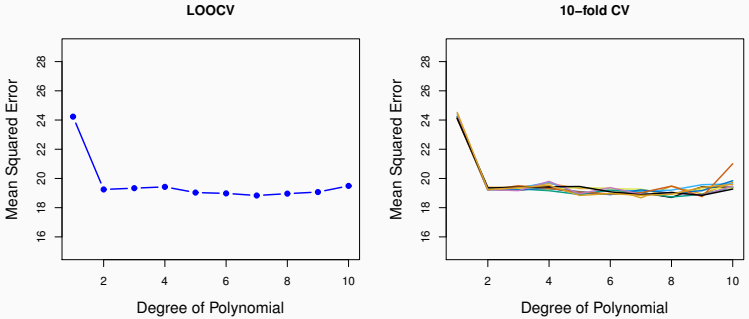

Figure 10: Cross-validation was used on the Auto data set in order to estimate the test error that results from predicting mpg using polynomial functions of horsepower. Left: The LOOCV error curve. Right: 10-fold CV was run nine separate times, each with a different random split of the data into ten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

K-Fold Cross-Validation

- When considering real-data we do not know the true test MSE, which makes it difficult to determine the accuracy of the cross-validation approach.
- Reconsider the simulated data set from Lecture 1 (provided again in the next three slides), in which we fitted different smoothing splines to the data.
- Let's apply cross validation to this simulated data to find the optimal amount of flexibility that the smoothing spline should have.
- Note that in such situations, we are not necessarily interested in the magnitude of the test MSE, but rather want to find the amount of flexibility that leads to the lowest MSE.

K-Fold Cross-Validation

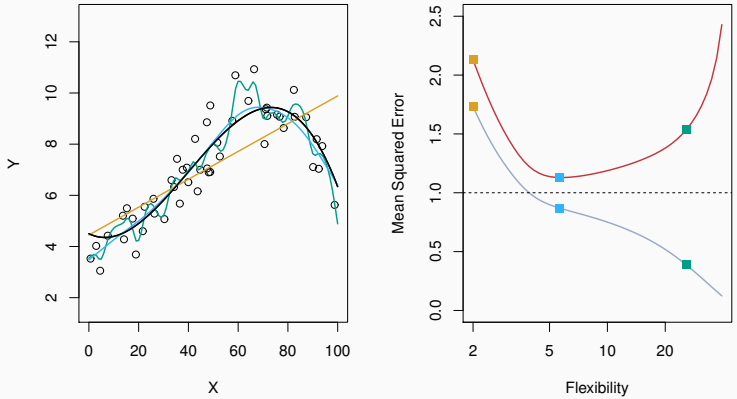

Figure 11: Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

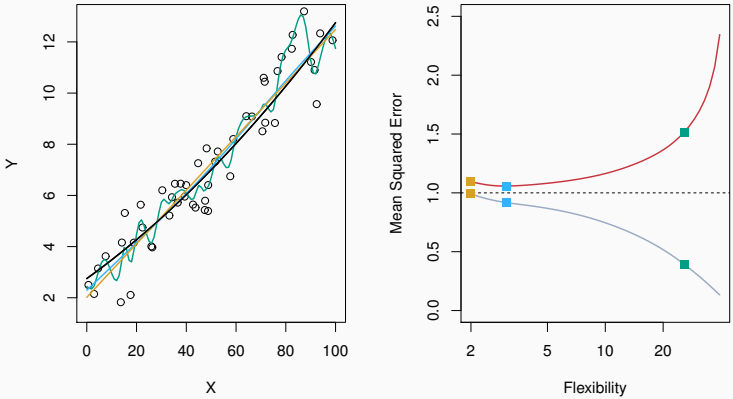

Figure 12: Details are as before, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.

K-Fold Cross-Validation

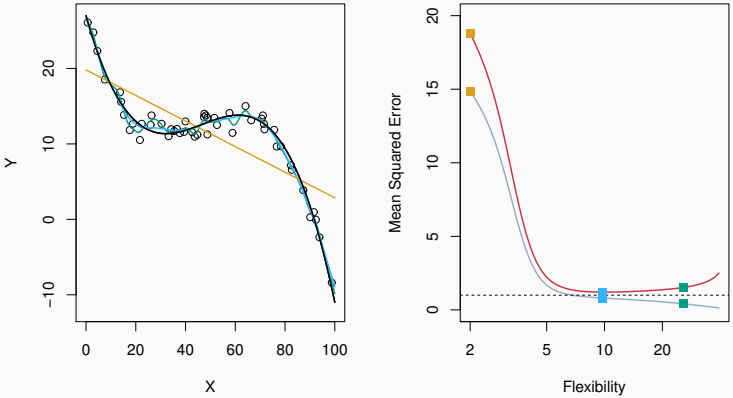

Figure 13: Details are as in Figure 2.9, using a different f that is far from linear. In this setting, linear regression provides a very poor fit to the data.

K-Fold Cross-Validation

- True and estimated test MSE for the simulated data:

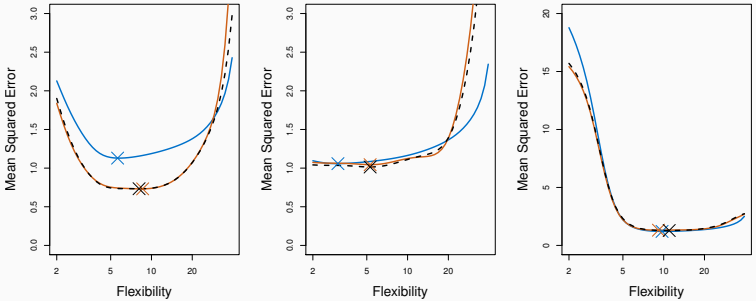

Figure 14: True and estimated test MSE for the simulated data sets. The true test MSE is shown in blue, the LOOCV estimate is shown as a black dashed line, and the 10-fold CV estimate is shown in orange. The crosses indicate the minimum of each of the MSE curves.

Cross Validation

Bias-Variance Trade-Off

K-Fold Cross-Validation - Bias-Variance Trade-Off

- Using k-fold CV with $k < n$ has a computational advantage to LOOCV and often also gives more accurate estimates of the test error rate than LOOCV. This has to do with a bias-variance trade-off.
- Test Error Bias: LOOCV $<$ k-fold CV
- Test Error Variance: k-fold CV $<$ LOOCV
- LOOCV: train n different models on nearly identical observations $\Rightarrow$ High Correlation among the outputs
- k-fold CV: average the outputs of k fitted models that are less correlated with each other (smaller overlap between the training sets).
- Recall: the mean of many highly correlated quantities has higher variance than does the mean of many quantities that are not as highly correlated.
- Thus $K = 5$ or $K = 10$ provides a good compromise for this bias-variance trade-off.

Cross Validation

Cross Validation in the Classification Setting

K-Fold Cross-Validation - Classification

- So far: Cross Validation for Regression. Now let's look at classification.
- Here we use the number of missclassified observations.
- We divide the data into K roughly equal-sized parts $C_1, C_2, ...C_K$. C_k denotes the indices of the observations in part k. There are n_k observations in part k, then:

$$CV_K = \frac{1}{K} \sum_{k=1}^{K} ERR_k$$

where $ERR_K = \sum_{i \in C_k} I(y_i \neq \hat{y}_i)/n_k$.

K-Fold Cross-Validation - Classification

- Example: Let's reconsider the data from Lecture 2.

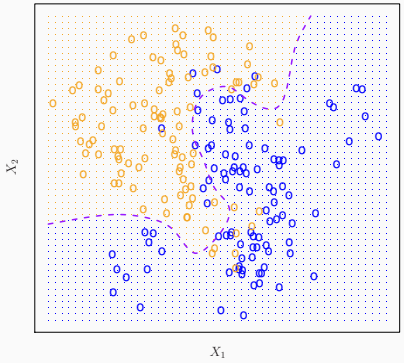

Figure 15: A simulated data set consisting of 100 observations in each of two groups, indicated in blue and in orange. The purple dashed line represents the Bayes decision boundary. The orange background grid indicates the region in which a test observation will be assigned to the orange class, and the blue background grid indicates the region in which a test observation will be assigned to the blue class.

K-Fold Cross-Validation - Classification

- Let's fit a logistic regressions with different degrees of polynomials, i.e. degrees 1-4.
- For example, we can fit a quadratic logistic regression model of the form:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 \tag{1}$$

K-Fold Cross-Validation - Classification

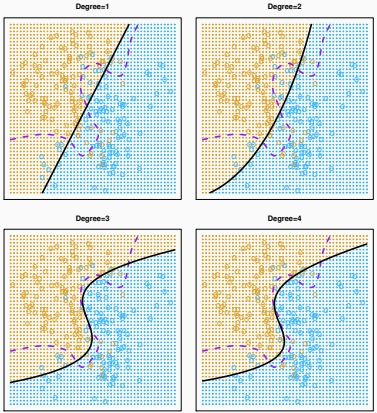

Figure 16: Logistic regression fits on the two-dimensional classification data. The Bayes decision boundary is represented using a purple dashed line. Estimated decision boundaries from linear, quadratic, cubic and quadratic (degrees 1–4) logistic regressions are displayed in black. The test error rates for the four logistic regression fits are respectively 0.201, 0.197, 0.160, and 0.162, while the Bayes error rate is 0.133.

K-Fold Cross-Validation - Classification

- How to decide among the four logistic regressions?
- We can use Cross Validation:

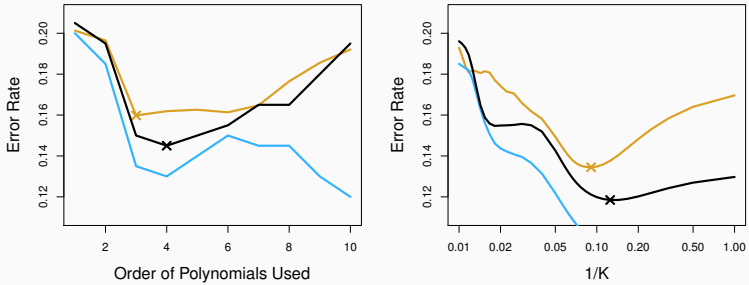

Figure 17: Test error (brown), training error (blue), and 10-fold CV error (black) on the two-dimensional classification data. Left: Logistic regression using polynomial functions of the predictors. The order of the polynomials used is displayed on the x-axis. Right: The KNN classifier with different values of K, the number of neighbors used in the KNN classifier. *NOTE: We have not discussed the KNN in this course and it will thus also not be part of the exam.*

The Bootstrap

Bootstrapping

- The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- For example, it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.
- Loosely speaking: We want to repeatedly create smaller new samples from our original large sample, given that we do not know anything about the distribution of the population.

Bootstrapping

- The use of the term bootstrap derives from the phrase to pull oneself up by one's bootstraps, widely thought to be based on one of the eighteenth century "The Surprising Adventures of Baron Munchausen" by Rudolph Erich Raspe:

  The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.

Bootstrapping

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y, respectively, where X and Y are random quantities.
- We will invest a fraction α of our money in X, and will invest the remaining $(1 - \alpha)$ in Y.
- We wish to choose α to minimize the total risk, or variance, of our investment. In other words, we want to minimize $Var(\alpha X + (1 - \alpha) Y)$.
- One can show that the value that minimizes the risk is given by:

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}, \tag{2}$$

where $\sigma_X^2 = Var(X)$, $\sigma_Y^2 = Var(Y)$, $\sigma_{XY} = Cov(X, Y)$.

Bootstrapping

- But the values of σ_X^2 ,σ_X^2 and σ_{XY} are unknown.
- We can compute estimates for these quantities, $\hat{\sigma}_X^2$, $\hat{\sigma}_X^2$, and $\hat{\sigma}_X^2$, using a data set that contains measurements for X and Y.
- We can then estimate the value of α that minimizes the variance of our investment using:

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}} \tag{3}$$

- Note: This is a simulation study - so we assume that we know the true population distribution of X and Y. For the following simulations the parameters were set to $\sigma_X^2 = 1, \sigma_Y^2 = 1.25, \sigma_{XY} = 0.5$.

Bootstrapping

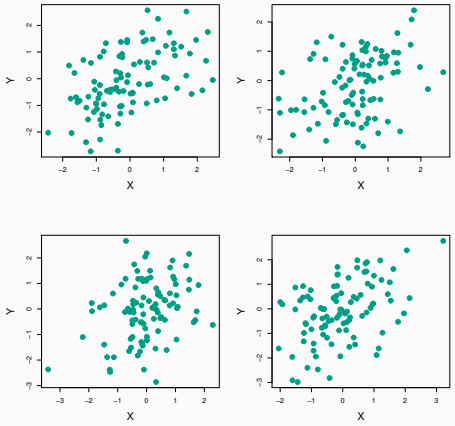

Figure 18: Each panel displays 100 simulated returns for investments X and Y. From left to right and top to bottom, the resulting estimates for α are $0.576, 0.532, 0.657$ and 0.651.

Bootstrapping

- To estimate the *standard deviation* of $\hat{\alpha}$ we repeated the process of simulating 100 paired observations of X and Y, and estimate α 1000 times.
- We thereby obtained 1000 estimates for α, which we can call $\hat{\alpha}_1, \hat{\alpha}_2, ...\hat{\alpha}_{1000}$.
- Remember we know the simulation parameters, which are $\sigma_X^2 = 1, \sigma_Y^2 = 1.25, \sigma_{XY} = 0.5$, and so we know that the true value of α is 0.6 (indicated by the red line in the figure below).

Bootstrapping

- The mean over all 1000 estimates for α is:

$$\overline{\alpha} = \frac{1}{1000}\sum_{r=1}^{1000}\hat{\alpha}_r = 0.5996,$$

  very close to $\alpha = 0.6$, and the standard deviation of the estimates is:

$$SE(\hat{\alpha}) = \sqrt{\frac{1}{1000}\sum_{r=1}^{1000}(\hat{\alpha}_r - \overline{\alpha})^2} = 0.083,$$

- This gives us a very good idea of the accuracy of $\hat{\alpha}$: $SE(\hat{\alpha}) \sim 0.083$
- So roughly speaking, for a random sample from the population, we would expect $\hat{\alpha}$ to differ from α by approximately 0.08, on average.

Bootstrapping

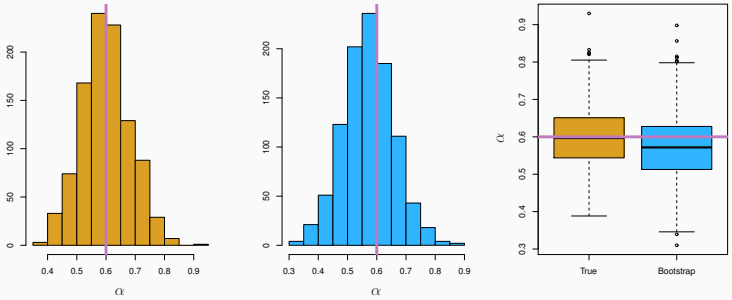

Figure 19: Left: A histogram of the estimates of α obtained by generating 1000 simulated data sets from the true population. Center: A histogram of the estimates of α obtained from 1000 bootstrap samples from a single data set. Right: The estimates of α displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of α.

Bootstrapping

- Now back to the real world
  - The procedure outlined above cannot be applied, because for real data we cannot generate new samples from the original population.
  - However, the bootstrap approach allows us to use a computer to mimic the process of obtaining new data sets, so that we can estimate the variability of our estimate without generating additional samples.
  - Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set with replacement.
  - Each of these "bootstrap data sets" is created by sampling with replacement, and is the same size as our original dataset. As a result some observations may appear more than once in a given bootstrap data set and some not at all.

Bootstrapping

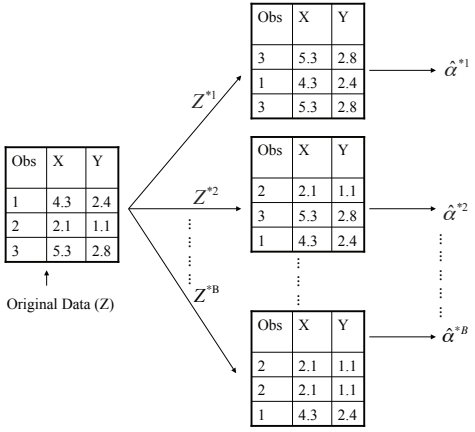

Figure 20: A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations. Each bootstrap data set contains n observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of α

Bootstrapping

- Denoting the first bootstrap data set by Z^{*1}, we use Z^{*1} to produce a new bootstrap estimate for α, which we call $\hat{\alpha}^{*1}$.
- This procedure is repeated B times for some large value of B (say 100 or 1000), in order to produce B different bootstrap data sets, Z^{*1}, Z^{*1}, ..., Z^{*B} and B corresponding α estimates, $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \hat{\alpha}^{*3}, ..., \hat{\alpha}^{*B}$.
- We estimate the standard error of these bootstrap estimates using the formula:

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}(\hat{\alpha}^{*r} - \overline{\hat{\alpha}^{*}})}$$

- This serves as an estimate of the standard error of $\hat{\alpha}$ estimated from the original data set. For this example $SE_B(\hat{\alpha}) = 0.087$.

Bootstrapping

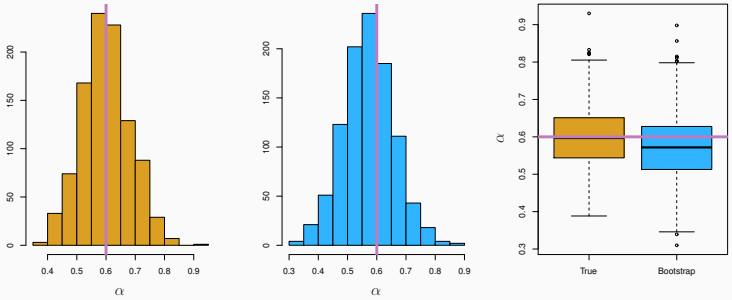

Figure 21: Left: A histogram of the estimates of α obtained by generating 1000 simulated data sets from the true population. Center: A histogram of the estimates of α obtained from 1000 bootstrap samples from a single data set. Right: The estimates of α displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of α.

True or False?

a) __ LOOCV is always the preferred method, when aiming to reduce the test MSE.

b) __ The validation set approach has a lower bias than the LOOCV approach.

c) __ The correlation among training datasets is lower for k-fold CV, than for LOOCV.

d) __ Bootstrapping refers to the situation, in which we create new samples by drawing observations from the original sample, without replacement.

e) __ Bootstrapping can only be applied, if we know the distribution of the population.

True or False?

a) F LOOCV is always the preferred method, when aiming to reduce the test MSE.

b) F The validation set approach has a lower bias than the LOOCV approach.

c) T The correlation among training datasets is lower for k-fold CV, than for LOOCV.

d) F Bootstrapping refers to the situation, in which we create new samples by drawing observations from the original sample, without replacement.

e) F Bootstrapping can only be applied, if we know the distribution of the population.

Disclaimer: This material has been prepared by Philipp Kremer and Constantin Lisson in 2021 and draws very extensively on James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). *An introduction to statistical learning* and the corresponding lecture slides available from these authors.

Solutions to review questions

Solutions to review questions

Review questions from Slide 6

a) The Type I Error represents the Probability of rejecting H_0, if in fact H_0 is true.

b) The Type II Error measure the Probability of failing to reject H_0, if in fact H_0 is false.

c) When we decrease the significance level α, i.e. the probabilits of a Type I error, then the probability of a Type II error will increase.

Solutions to review questions

Review questions from Slide 11

a) The Family Wise Error Rate is the probability of making at least one Type I error, that is rejecting H_0, although H_0 is true.

b) Testing each manager individually, we can reject Manager Two, Four and Five, as all of the p-Values are greater than $\alpha = 0.05\%$.

c) Using Bonferroni: $\alpha/m = 0.05/5 = 0.01 \Rightarrow$ Reject the null hypothesis only for the first manager, since all other p-values exceed 0.01.

Solutions to review questions

Review questions from Slide 13

a) Here $m = 5$. Associated ordered p-values:

1. $p_{(1)} = 0.006 < 0.05/(5+1-1) = 0.01$
2. $p_{(2)} = 0.012 < 0.05/(5+1-2) = 0.0125$
3. $p_{(3)} = 0.601 > 0.05/(5+1-3) = 0.0167$
4. $p_{(4)} = 0.756 > 0.05/(5+1-4) = 0.025$
5. $p_{(5)} = 0.918 > 0.05/(5+1-5) = 0.05$

- Holm rejects the first two null hypotheses, which in turn implies $L = 3$.

b) The Bonferroni Method is stricter in the sense that we only find one statistically significant manager, as compared to Holm's Set Down procedure, in which we find two statistically significant manager.

Solutions to review questions

Review questions from Slide 20

a) Here $m = 5$. Associated ordered p-values:

1. $p_{(1)} = 0.006 < 0.05 \times \frac{1}{5}$
2. $p_{(2)} = 0.012 < 0.05 \times \frac{2}{5}$
3. $p_{(3)} = 0.601 > 0.05 \times \frac{3}{5}$
4. $p_{(4)} = 0.756 > 0.05 \times \frac{4}{5}$
5. $p_{(5)} = 0.918 > 0.05 \times \frac{5}{5}$

- To control the FDR at 5%, we reject the null hypothesis that the first and the third fund manager perform no better than chance.

b) The Bonferroni Method is stricter in the sense that we only find one statistically significant manager, as compared to Benjamini Hochberg approach, in which we find two statistically significant manager.

Solutions to review questions

Review questions from Slide 25:

a) As we choose a more flexible method, the bias associated with each observed data point is reduced and the red line starts to decrease. As we try to model more information from the training data, we continue to reduce the bias, but increase the variance of the method. The overfit to the training data leads to a higher variance and an increasing red test MSE curve, as the more complex procedure performs inferior on previously unseen test observations.

b) The horizontal grey dotted line represents the irreducible error. It is the error that can not be eliminated, even when we know the true underlying functional form of $f(X)$ (For the distinction between reducible and irreducible error, recall Lecture 1).

c) No, choosing the method with the lowest *training* MSE, will lead to choosing the method that has potentially overfit the data. Applying such method to previously unobserved data will lead to a high test MSE.

This material draws extensively on James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). *An introduction to statistical learning* and the lecture slides available from these authors.