Bayesian Inference and Decision Theory

Unit 10: Hypothesis Tests, Bayes Factors, and Bayesian Model Averaging
Learning Objectives for Unit 10

• Describe the decision theoretic approach to hypothesis testing and compare with the frequentist approach
• Define model uncertainty and parameter uncertainty
• Use Bayesian model averaging to account for model uncertainty in parameter estimates and predictions
• Wrapup: Bayesian inference and decision theory
Hypothesis Testing

- Typical hypothesis testing problem: use data to evaluate support for two hypotheses
  - “Null” hypothesis $H_0 : \theta \in \Omega_0$ (given benefit of doubt)
  - “Alternative” hypothesis $H_A : \theta \notin \Omega_0$
  - Goal is to determine whether to reject null hypothesis in favor of alternative
  - Must balance two types of error:
    - False alarm or Type I error: incorrectly reject null hypothesis
    - Miss or Type II error: incorrectly fail to reject null hypothesis

- Bayesian decision theorists are less concerned with whether a hypothesis is “true” than with what action should be taken
  - Specify loss function for actions under different hypotheses for the parameter
  - Find optimal action

Null hypotheses are usually known in advance to be false, and the point of significance tests is usually to find out whether they are nevertheless approximately true. - I. J. Good
Decision Theoretic View of Hypothesis Testing

- Testing a hypothesis is a decision problem
  - Reject hypothesis means act “as if” it is false
  - Do not reject hypothesis means act “as if” it is true
    - Note: frequentists do not accept hypotheses!

- Beware the “false dichotomy”
  - Sometimes we can defer action until more information is available
  - We can use Bayesian decision theory to model information decisions about whether and how to collect information

- Decision-theoretic model of hypothesis requires a loss function
  - Cost of failing to reject a false null hypothesis (Type I error)
  - Cost of rejecting a true null hypothesis (Type II error)
  - Cost of indecision if indecision is allowable

- Optimal decision minimizes expected loss
A Simple Model for Hypothesis Testing

• Loss function depends only on whether you declare the correct hypothesis (not on how far off you are):
  - \( m = \) loss for “miss” (saying \( H_0 \) if \( H_A \) is true - Type II error)
  - \( f = \) loss for “false alarm” (saying \( H_A \) if \( H_0 \) is true - Type I error)
  - \( 0 = \) loss for correct decision
  - This loss function is rarely a "correct" representation of the losses in a real decision problem but is often a reasonable simplification
• Compare expected losses given current evidence \( X = \chi \):
  - Option R: Reject null hypothesis:
    - Loss is 0 if \( H_A \) is true; \( f \) if \( H_0 \) is true
    - Expected loss is: \( f \cdot P(H_0 | \chi) \)
  - Option N: Do not reject null hypothesis
    - Loss is \( m \) if \( H_A \) is true; 0 if \( H_0 \) is true
    - Expected loss is: \( m \cdot P(H_A | \chi) \)
• It is optimal to reject null hypothesis if:
  - \( f \cdot P(H_0 | \chi) < m \cdot P(H_A | \chi) \) or \( \frac{P(H_A | \chi)}{P(H_0 | \chi)} > \frac{f}{m} \) or \( P(H_A | \chi) > \frac{f}{f+m} \)
• As the penalty for false alarm (\( f \)) increases or the penalty for misses (\( m \)) decreases, we become more conservative, requiring a higher probability of \( H_A \) in order to reject \( H_0 \).
Quality Control Example: The Problem

- Due to a malfunction in one of three box packing lines, one of every three boxes of a consumer product is missing a component required for assembly
- It is not known which of the boxes waiting to be shipped are missing the component
- It is expensive to open and inspect every box
- Solution: weigh boxes and open underweight ones
- What should be our weight cutoff for opening boxes?
Quality Control Example: Optimal Decision

- Statistical model for box weights:
  - Weights for correctly packed boxes are normally distributed with mean 2.3kg and standard deviation 0.1kg
  - Weights for incorrectly packed boxes are normally distributed with mean 1.9 and standard deviation .1kg
  - Prior probability of incorrectly packed box is 1/3

- Loss function:
  - Shipping a box missing the component is 10 times loss for inspecting a correctly packed box

- Optimal policy:
  - Let $X$ be the measured weight of the box
  - Calculate normal density values:
    - $f_A(X) = f(X \mid 1.9, 0.1)$ normal density with mean 1.9, st dev 0.1
    - $f_0(X) = f(X \mid 2.3, 0.1)$ normal density with mean 2.3, st dev 0.1
  - Posterior probability that box is incorrectly packed is:
    - $P(H_A \mid X) = 0.33 \times N(x \mid 1.9,0.1)/(0.33 \times N(x \mid 1.9,0.1) + 0.67 \times N(x \mid 2.3,0.1))$
  - Inspect box if $P(H_A \mid X) > 1/11 = 9.1\% \text{ (or } X < 2.14 \text{ kg})$
Quality Control Example: Decision Threshold

- Inspect box if:
  - Posterior probability that part is missing exceeds 1/11
  - Weight is less than 2.14 kg
  - The threshold is the value at which the two choices have equal losses
How Good is a Test?

- Examples:
  - How effective is a test for diagnosing a disease?
  - How well does our fingerprint recognition system work in matching fingerprints to suspects?
  - How accurate is our sonar classification system?
  - How well does the polygraph examination process perform at detecting deception?

- “Percent correct” is often used as a measure of accuracy of a test
  - The same test can have different values for percent correct depending on:
    - Base rate of condition in the population
    - Threshold for declaring a positive result

- Is there an intrinsic measure of how well a test diagnoses a condition?

- Moving threshold to left decreases false negatives but increases false positives
- Moving threshold to right decreases false positives but increases false negatives
Receiver Operating Characteristic (ROC)

- ROC shows tradeoff between false positives and false negatives as threshold varies
  - Horizontal axis is false positive rate (1 - specificity)
  - Vertical axis is true positive rate (sensitivity)
  - Different thresholds yield different points on curve
- Area under curve is intrinsic measure of accuracy
  - A perfect test has area 1.0
  - A random test has area 0.5
Area under curve is 0.975

Sensitivity = Specificity = 0.977
(0.023, 0.977)
(0.058, 0.992)

Optimal Threshold
Sensitivity = 0.992
Specificity = 0.942

ROC for Box Weight
ROC Examples

Area 0.999

Area 0.71

Sensitivity & Specificity = 0.99

Sensitivity & Specificity = 0.67

©Kathryn Blackmond Laskey  Spring 2022  Unit 10v3 - 12 -
Rare Conditions

• When target condition is very rare, even a highly accurate test yields many false positives for each true positive

• Hypothetical example from The Polygraph and Lie Detection, National Academy Press, 2002:
  • Screen a population of mostly truthful employees to identify possible spies
  • Assume polygraph has area under ROC of 0.9 and base rate of approximately 1 in 1,000
  • Test can be expected to falsely accuse approximately 200 innocent people for every spy detected

<table>
<thead>
<tr>
<th>Examinee’s true condition</th>
<th>Spy</th>
<th>Nonspy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Fail” test</td>
<td>8</td>
<td>1,598</td>
<td>1,606</td>
</tr>
<tr>
<td>“Pass” test</td>
<td>2</td>
<td>8,392</td>
<td>8,394</td>
</tr>
<tr>
<td>TOTAL</td>
<td>10</td>
<td>9,990</td>
<td>10,000</td>
</tr>
</tbody>
</table>

• We can lower the threshold to reduce false positives but then we fail to catch most of the spies
How to Improve a Test

• For any test, sensitivity and specificity must be traded off against each other

• A more informative test can increase both sensitivity and specificity
  • More informative test has greater area under ROC
  • We can increase area under ROC by increasing separation of density functions $f_0(x)$ and $f_A(x)$ of data under null and alternative hypotheses

• How to make a test more informative
  • If test is based on sampling, we can increase the number of observations
  • If test is based on evidential reasoning from a set of features we can add additional features
    • Multi-source fusion is in principle more accurate than single-source discrimination
    • Achieving the potential depends on
      • Identifying diagnostic features with independent sources of error
      • Accurately modeling the interaction among features
      • Achieving tractable computational performance
Value of Information and Sample Size

- If a test is based on averaging many observations, then we can improve accuracy by taking more samples.
- Often we can model cost of gathering information as sum of fixed cost plus per-sample cost.
- Taking more samples reduces error due to random noise but cannot reduce “nonsampling error.”
  - There may be systematic bias in measuring instrument.
  - Degree of bias may depend on unmeasurable environmental factors.
  - What we are measuring may be only indirectly related to hypothesis we are testing.
  - We may have a poor model of the phenomenon.

EXAMPLE: OBLIQUELY VIEWED OBJECT

- Objective: measure length of object.
- Object looks shorter when viewed obliquely.
- Our sensor measures apparent length of object.
- No matter how accurately we measure apparent length, we cannot measure length accurately unless we know the view angle.
Samples from an object are collected to analyze whether it contains a given compound:
- If compound is present, observations are iid normal with mean 1 and standard deviation 5.
- If compound is absent, observations are iid normal with mean -1 and standard deviation 5.
- If we take enough observations, we can diagnose with near certainty.

Prior probability is 10% that the compound is present.

Loss function:
- \( L_M = 100 \) Miss cost (report absent when compound is present)
- \( L_{FA} = 30 \) False alarm cost (report present when compound is absent)
- \( L_T = 5 \) Fixed cost of taking measurements
- \( L_S = 0.01 \) Per-observation cost of measurement

Questions:
- Should we take any measurements?
- If we take measurements, how many observations should we collect?

Optimal policy minimizes expected loss.
Example: Expected Loss with No Observations

- Expected loss for reporting compound is present
  - Compound present: Probability 0.1, Loss 0
  - Compound absent: Probability 0.9, Loss 30 (false positive)
  - Expected loss: 0.9 x 30 + 0.1 x 0 = 27

- Expected loss for reporting compound is absent
  - Compound present: Probability 0.1, Loss 100 (false negative)
  - Compound absent: Probability 0.9, Loss 0
  - Expected loss: 0.9 x 0 + 0.1 x 100 = 10

- Optimal action is not to reject $H_0$ (report compound is absent)
- Expected loss is 10
Example: Expected Loss with 25 Observations

- Observation has normal distribution:
  - Compound present: mean 1, standard deviation $5/25^{1/2} = 1$
  - Compound absent: mean -1, standard deviation $5/25^{1/2} = 1$

- Posterior probability that compound is present is
  \[ p = \frac{0.1 \times f_N(x | \mu=1, \sigma=1)}{0.1 \times f_N(x | \mu=1, \sigma=1) + 0.9 \times f_N(x | \mu=-1, \sigma=1)} \]

- Optimal action is to report present if $30 \times (1-p) < 100 \ p \ \Rightarrow \ p > 0.231 \ \Rightarrow \ x > 0.497$

<table>
<thead>
<tr>
<th>Condition</th>
<th>Observation</th>
<th>Action</th>
<th>Probability</th>
<th>Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>$X &gt; 0.497$</td>
<td>Reject $H_0$</td>
<td>$0.1 \times 0.693$</td>
<td>0</td>
</tr>
<tr>
<td>Present</td>
<td>$X \leq 0.497$</td>
<td>Do not reject $H_0$</td>
<td>$0.1 \times 0.307$</td>
<td>100</td>
</tr>
<tr>
<td>Absent</td>
<td>$X &gt; 0.497$</td>
<td>Reject $H_0$</td>
<td>$0.9 \times 0.067$</td>
<td>30</td>
</tr>
<tr>
<td>Absent</td>
<td>$X \leq 0.497$</td>
<td>Do not reject $H_0$</td>
<td>$0.9 \times 0.933$</td>
<td>0</td>
</tr>
</tbody>
</table>

- Total Expected Loss is sum of expected loss from misses, expected loss from false alarms and cost of observations:
  - $0.1 \times 0.307 \times 100 + 0.9 \times 0.067 \times 30 + 5 + 25 \times 0.01 = 10.13$

- This is greater than the expected loss of no observations
Example: Expected Loss with 100 Observations

- Observation has normal distribution:
  - Compound present: mean 1, standard deviation 5/\sqrt{100} = 1/2
  - Compound absent: mean -1, standard deviation 5/\sqrt{100} = 1/2

- Posterior probability that compound is present is
  \[ p = \frac{0.1 \times f_N(x|\mu=1, \sigma=1/2)}{0.1 \times f_N(x|\mu=1, \sigma=1/2) + 0.9 \times f_N(x|\mu=-1, \sigma=1/2)} \]

- Optimal action is to report present if
  \[ 30 \times (1-p) < 100 p \Rightarrow p > 0.231 \Rightarrow x > 0.124 \]

<table>
<thead>
<tr>
<th>Condition</th>
<th>Observation</th>
<th>Action</th>
<th>Probability</th>
<th>Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>X &gt; 0.124</td>
<td>Reject $H_0$</td>
<td>0.1 x 0.960</td>
<td>0</td>
</tr>
<tr>
<td>Present</td>
<td>X ≤ 0.124</td>
<td>Do not reject $H_0$</td>
<td>0.1 x 0.040</td>
<td>100</td>
</tr>
<tr>
<td>Absent</td>
<td>X &gt; 0.124</td>
<td>Reject $H_0$</td>
<td>0.9 x 0.012</td>
<td>30</td>
</tr>
<tr>
<td>Absent</td>
<td>X ≤ 0.124</td>
<td>Do not reject $H_0$</td>
<td>0.9 x 0.988</td>
<td>0</td>
</tr>
</tbody>
</table>

- Total Expected Loss is sum of expected loss from misses, expected loss from false alarms and cost of observations:
  \[ 0.1 \times 0.040 \times 100 + 0.9 \times 0.012 \times 30 + 5 + 100 \times 0.01 = 6.73 \]

- This is less than the expected loss of no observations
Losses as Function of Sample Size

Optimal sample size balances error reduction against observation cost.
Summary: Optimal Policy

- Optimal policy is to collect a sample
- Optimal decision threshold: declare condition is present if posterior probability is greater than or equal to 0.23
- Optimal sample size is 122
- Expected loss for optimal sample size is 6.658
Parameter and Model Uncertainty

- **Parameter uncertainty** refers to uncertainty about the unknown parameters of a model with known structure
  - Example: What are the values of regression coefficients in a linear regression?
  - Example: What is the mean of a normally distributed random variable?

- **Model uncertainty** (also called **structure uncertainty**) refers to uncertainty about the structure of a model
  - Example: is the relationship of a dependent variable on an independent variable linear or quadratic?
  - Example: does a dependent variable depend on a particular independent variable (equivalently, is the regression coefficient equal to zero)?
  - Do two normally distributed data sets have the same mean?
Example: Improvised Explosive Devices

- Rechtia is conducting military operations in Impestuostria to support international peacekeeping efforts
- Zappist insurgent groups plant improvised explosive devices (IEDs) to disrupt peacekeeping activities
- Rechtia’s Chief Statistician, Dr. Chi Square, has analyzed IED incidents over the past year and found that:
  - A Poisson distribution is a good approximation to counts of aggregate IED incidents occurring on major supply routes
  - The rate of occurrence is approximately 5.2 incidents per 20km segment per month
- The young analyst Anthony Prentice has been asked to investigate IED incidents on a 20km segment of Hardrod Highway
  - Nationwide weekly average is $5.2 \times 12 / 52 = 1.2$ incidents per week
  - If this segment followed the national pattern we would expect 7.2 incidents in 6 weeks
  - Data for the past 6 weeks: 3, 2, 1, 2, 4, 3 (15 incidents)
  - Is Poisson distribution with $\lambda = 1.2$ incidents per week an acceptable model for incident counts on this segment of highway?
Hypothesis Test for IED Incidence Rate

• A. Prentice wishes to evaluate the following hypotheses:
  • $H_0$: The distribution of IED incidents on this segment of highway follows a Poisson distribution with rate 7.2 incidents per 6-week period (null hypothesis)
  • $H_A$: The distribution of IED incidents on this segment of highway follows a Poisson distribution with unknown rate (alternate hypothesis)

• A. Prentice wishes to find the probabilities $P(H_0 \mid \text{data})$ and $P(H_A \mid \text{data})$

• We need the following ingredients:
  • Prior probability of $H_0$ (needs to be specified)
  • Likelihood of data under $H_0$ (this is known: Poisson with parameter 7.2)
  • Likelihood of data under $H_A$ (likelihood given $\Lambda$ is known; prior distribution for $\Lambda$ needs to be specified)

• Note that $H_0$ and $H_A$ have different dimensions

• This is a case of model (structural) uncertainty
  • $H_0$ is a zero-dimensional model ($\Lambda$ is known)
  • $H_A$ is one-dimensional model ($\Lambda$ takes on values on the positive real line)
Graphical Model of Hypothesis Testing Problem

- To specify the model fully, A. Prentice makes the following judgments:
  - There is a 75% chance this segment is like national aggregate: $P(H_0) = 0.75$
  - Prior distribution for $\Lambda$
    - shape $\alpha = 2$ and scale $\beta = 0.6$
    - Expected value is same as national aggregate
    - Not very informative prior: “like” starting with uniform prior ($\alpha=1, \beta=\infty$) and seeing 1 incident in 1.67 weeks

Hypothesis Problem

$H_0$ or $H_A$

Given $H_0$: $\Lambda = 1.2 / \text{week}$

Given $H_A$: $\Lambda \sim \text{Gamma}(\alpha, \beta)$

$x_i \sim_{iid} \text{Poisson}(\Lambda)$

$\sum x_i \sim \text{Poisson}(K), K = 6\Lambda$
• Find marginal likelihood for sufficient statistic $Y = \sum_i X_i$
  • Under $H_0$: Poisson distribution with rate $6 \times 1.2 = 7.2$ incidents per 6 weeks
  • Under $H_A$: Poisson-Gamma (negative binomial) marginal likelihood with size $\alpha = 2$ and probability $1/(1+n\beta) = 1/(1+6\times0.6) = 0.2174$
• Insert observed value $Y = 15$ into marginal likelihood
• Use Bayes rule to find posterior probability of each hypothesis

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Marginal Likelihood</th>
<th>Marginal Likelihood Value</th>
<th>Posterior Probability when $P(H_0)=0.75$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$: $\Lambda=1.2$, $K=7.2$</td>
<td>Poisson(7.2)</td>
<td>0.00414</td>
<td>39.3%</td>
</tr>
<tr>
<td>$H_A$: $\Lambda\sim\text{Gamma}(2,0.6)$, $K\sim\text{Gamma}(2,7.2)$</td>
<td>NBinom(2,0.2174)</td>
<td>0.0191</td>
<td>60.7%</td>
</tr>
</tbody>
</table>

$$0.393 = \frac{0.75 \times 4.14 \times 10^{-3}}{0.75 \times 4.14 \times 10^{-3} + 0.25 \times 1.91 \times 10^{-2}}$$
$$0.607 = \frac{0.025 \times 1.91 \times 10^{-2}}{0.75 \times 4.14 \times 10^{-3} + 0.25 \times 1.91 \times 10^{-2}}$$
Bayes Factor

• The posterior odds ratio of $H_A$ to $H_0$ is equal to the prior odds ratio times the likelihood ratio
  • For the IED problem the prior odds ratio is $0.25/0.75$ and the likelihood ratio is $1.91 \times 10^{-2} / 4.14 \times 10^{-3}$
  • Posterior odds ratio $P(H_A|Y)/P(H_0|Y) = \frac{1.91 \times 10^{-2}}{4.14 \times 10^{-3}} \frac{0.25}{0.75} = 60.7/39.3$

• When computing the likelihood of $H_A$ and/or $H_0$ requires integrating over an unknown parameter, the likelihood ratio is called the Bayes factor

Bayes factor for $H_A$ vs $H_0$ is 
$1.92 \times 10^{-2} / 4.14 \times 10^{-3} = 4.63$

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Marginal Likelihood</th>
<th>Marginal Likelihood Value</th>
<th>Posterior Probability when $P(H_0)=0.75$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$: $\Lambda=1.2$, $K=7.2$</td>
<td>Poisson(7.2)</td>
<td>0.00414</td>
<td>39.3%</td>
</tr>
<tr>
<td>$H_A$: $\Lambda$~Gamma(2,0.6), $K$~Gamma(2,7.2)</td>
<td>NBinom(6,1(1+7.2))</td>
<td>0.0191</td>
<td>60.7%</td>
</tr>
</tbody>
</table>
Sufficient Statistic or Full Data Set?

• We computed the posterior probability using the ratio $f_A(Y)/f_0(Y)$ of likelihoods of the sufficient statistic $Y = \sum_i X_i$
  • $Y$ has a negative binomial distribution with size $\alpha = 2$ and probability 0.2174
• We could have used the ratio $f_A(X)/f_0(X)$ of joint likelihoods of the observations
  • $f_0(X)$ is the product of six Poisson(1.2) likelihoods
  • $f_A(X)$ is the product of six negative correlated negative binomial likelihoods (under $H_A$ the observations depend on each other through their shared dependence on the unknown Poisson mean $\Lambda$)
• These ratios have the same value (that’s what it means for a sufficient statistic to be sufficient)
Varying the Virtual Sample Size

- A. Prentice thinks the national aggregate distribution seems to be a good choice for the expected value of $\Lambda$

- A. Prentice was unsure about how to choose the virtual sample size for the prior distribution of $\Lambda$

- To investigate sensitivity to choice of $\alpha$, we fix the expected value $\alpha\beta = 1.2$ and vary $\alpha$
  - As $\alpha$ tends to zero, the pmf at $Y = 0$ tends to 1, and the Bayes factor $f_A(15)/f_0(15)$ tends to zero
  - As $\alpha$ tends to infinity the pmf tends to the Poisson distribution and the Bayes factor $f_A(15)/f_0(15)$ tends to 1

- In the IED problem with $\alpha\beta = 1.2$ and $Y = 15$
  - The maximum Bayes factor is 4.627 at $\alpha = 2.07$
  - The Bayes factor for the value $\alpha = 2$ chosen by A. Prentice is 4.626
  - A. Prentice has given $H_A$ nearly the maximum opportunity to overcome $H_0$
Poisson-Gamma Marginal Likelihood with $\alpha \beta = 1.2$ and Different Values of $\alpha$

**Posterior probability of $H_0$ increases if $Y$ is in range expected under $H_0$.**

**Observed count: 15 in 6 weeks**

**Posterior probability of $H_0$ decreases if $Y$ is outside range expected under $H_0$.**
The Loss Function and the Decision Problem

- Miss penalty: cost of acting as if the local rate is the same as the national rate when it is not
- False alarm penalty: cost of acting as if the local rate is different from the national rate if it is the same
- It seems reasonable to suppose that misses cost more than false alarms (we would like to know about local variation)
  - Reject if \( P(H_A|Y) / P(H_0|Y) > f / m \)
  - If misses cost more than false alarms then \( f / m < 1 \)
  - \( P(H_A|Y) / P(H_0|Y) = 0.61 / 0.39 > 1 > f / m \)
  - If misses cost more than false alarms, it is optimal to reject \( H_0 \) in favor of \( H_A \)
Bayesian Model Averaging for IED Prediction

- We may be more interested in predicting IED incidents than in testing the hypothesis that the rate is different from the national aggregate rate
  - Acting “as if” $H_0$ is correct would mean using the national aggregate rate
  - Acting “as if” $H_A$ is correct would mean estimating rate using posterior distribution given $H_A$
  - It may be desirable not to commit to either hypothesis

- Bayesian model averaging keeps both hypotheses active
  - Predict future incidents as weighted average of posterior predictive distributions under the two hypotheses
    $$f_{BMA}(y) = 0.383f_{Pois}(y, 7.2) + 0.607f_{NegBinom}(y, 6, 0.561)$$
In regression problems, we often want to know whether there is a linear relationship between dependent variable $Y$ and independent variable $X$.

For a simple linear regression:
- Hypothesis $H_0$ (no relationship) $\beta = 0$
- Hypothesis $H_A$ (linear relationship) $\beta \neq 0$

$H_0$ and $H_A$ have different dimensions.

The question of which regression coefficients are zero / non-zero is a question of model uncertainty.
Selecting Explanatory Variables

• Often there are very many potentially relevant variables
• Too many explanatory variables leads to collinearity and non-identifiability
• It is common in applied work to do exploratory analysis to select the “most important” variables and then report results of regression using these variables
  • Example: stepwise regression
  • Standard hypothesis testing theory is not valid when this kind of exploratory analysis was performed to select variables to include in regression
  • Actual significance levels are much lower than nominal significance levels
• Section 9.3 of Hoff discusses the pitfalls of some common model selection procedures
Bayesian Model Averaging for Regression Predictor Selection

- We can compare the posterior probability of hypotheses in which we set some coefficients $\beta_i = 0$ and other coefficients $\beta_j \neq 0$
  - $H_{0i}$: no dependence on $X_i (\beta_i = 0)$
  - $H_{Ai}$: dependence on $X_i (\beta_i \neq 0)$
- Rewrite the regression equation
  - $E[y_i] = z_1 x_1 \beta_1 + \cdots + z_p x_p \beta_p$
  - $z_i = 0$ or $1$ indicates whether predictor $i$ affects dependent variable
- Use Bayes rule to compare the relative posterior probabilities of the hypotheses $z_i = 0$ or $1$
- When there are more than a few predictors this is typically done via Monte Carlo
- Bayesian model averaging is less prone to overfitting than many commonly applied methods based on significance tests, especially when there are many potential predictors
Bayesian Model Selection: Example

- Diabetes example (Section 9.3 of Hoff text)
  - Baseline data collected for 10 predictor variables $x_1, \ldots, x_{10}$ and disease progression measure $y$ for 442 diabetes patients
    - All variables centered and scaled to have mean 0 and variance 1
  - 64 possible predictor variables:
    - 10 main effects + 9 quadratic terms + $10 \times 9 / 2 = 45$ interaction terms
- Divide data into 342 training cases and 100 test cases
- Mean squared prediction error (MSPE) on test sample:
  - “No relationship” (predict 0 for all cases): $\text{MSPE}_0 = 0.97$
  - Ordinary least squares regression with all 64 variables: $\text{MSPE}_{\text{ols}} = 0.67$
  - Backward elimination (start with all predictors and remove one by one until t-statistic is above cutoff): $\text{MSPE}_{\text{be}} = 0.53$
  - Bayesian model averaging: $\text{MSPE}_{\text{bma}} = 0.45$
- Backward elimination typically finds some variables with a positive relationship even when there is no relationship!
Bayesians formulate hypothesis testing as a decision problem
- Decision: Act as if $H_0$ is true or act as if $H_A$ is true
- Utilities: Loss for miss and loss for false alarm
- Probabilities:
  - Specify prior distribution for each hypothesis and likelihood for observations given each hypothesis
  - Use Bayes rule to obtain posterior probability for each hypothesis
- Reject $H_0$ if $m / f > P(H_0|X) / P(H_A|X)$
- The “aggressiveness” of the hypothesis test depends on the ratio of $m$ to $f$

We considered using value of information in deciding whether to collect data and how much to collect.

When we integrate over parameter spaces of different dimensions the likelihood ratio is called the Bayes factor
  - The Bayes factor is sensitive to the prior distribution
  - Results can be misleading if the prior is chosen badly
  - We investigated sensitivity of the Bayes factor to the virtual sample size

Bayesian model averaging keeps all hypotheses under consideration and weights them by their posterior probabilities
Recap: The Bayesian Approach

• The Bayesian approach is:
  • A way of thinking about problems of inference and decision-making under uncertainty
  • A set of tools for applying this way of thinking to practical problems

• A Bayesian can answer the questions a decision-maker cares about:
  • What is the probability this event will happen?
  • What is the probability this parameter falls in a certain range?
  • What is my best decision in these circumstances?

• When we have a reasonable amount of data and weak prior information, we can give many standard statistical tools a Bayesian interpretation (at least approximately)

• Recent advances in computation have made Bayesian methods practical for many complex real-world problems

\[ g(\theta|x) = \frac{f(x|\theta)g(\theta)}{f(x)} \]

Thomas Bayes (1702-1762)
Bayesian and Frequentist Statistics

- Most statistics courses are taught from the frequentist perspective
- Frequentists
  - View probability as objective property of random processes
  - Assign probabilities to collectives but *not* individual events
  - Condition on parameters, treat data as probabilistic
- Subjectivists
  - View probability as rational degrees of belief about uncertain phenomena
  - Assign probability to any unknown, including individual events
  - Condition on knowns, treat unknowns as probabilistic
- Frequentist analyses can often be given a Bayesian interpretation
  - Often good approximation if large sample and weak prior information
Fundamental Ideas of the Bayesian Approach
[1 of 3]

- Probability expresses rational degrees of belief about uncertain phenomena
- Rational decision makers update beliefs by Bayes Rule and make choices according to maximum expected utility
- Inference is belief dynamics
  - Prior beliefs are updated with evidence
  - Posterior beliefs at one stage become prior beliefs for next stage
  - Predict next stage using all knowledge up to present stage
- “Probability is common sense reduced to calculation” - Laplace
- Whether to gather information can be treated as a decision problem
Fundamental Ideas of the Bayesian Approach
[2 of 3]

• Conjugate prior / likelihood pairs simplify Bayesian inference
  • Posterior distribution and predictive distribution can be found exactly
  • Convenient and useful when a good model for the data

• Approximation methods are important when exact results are unavailable
  • There is a rapidly growing literature in approximation methods for Bayesian inference
  • Markov Chain Monte Carlo (MCMC) is a general-purpose class of approximation methods that helped spark the Bayesian revolution
  • MCMC diagnostics can help to assess whether a MCMC sampler is behaving well
  • There are many special purpose exact and Monte Carlo approximation methods
Fundamental Ideas of the Bayesian Approach [3 of 3]

- Hierarchical models use structural assumptions to achieve better statistical power without sacrificing realism
  - Information sharing among related parameters
- “All models are wrong but some models are useful” – Box
  - Always check model assumptions against data
- Posterior predictive model evaluation can help assess whether model is adequate for the intended purpose