Introduction to Statistical Modeling

Theory Construction and Statistical Modeling

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Motivating Example

Imagine you are working for an F1 team. You're job is to use data from past seasons to optimize the baseline setup of your team's car.

- Suppose you have two candidate setups that you want to compare.
- For each setup, you have 100 past lap times.
- How do you distill those 200 lap times into a succinct decision between the two setups?

Motivating Example

Suppose I tell you that the mean lap time for Setup A is 118 seconds and the mean lap time for Setup B is 110 seconds.

- Can you confidently recommend Setup B?
- What caveats might you consider?

Motivating Example

Suppose I tell you that the standard deviation for the times under Setup A is 7 seconds and the standard deviation for the times under Setup B is 5 seconds.

- How would you incorporate this new information into your decision?

Motivating Example

Suppose I tell you that the standard deviation for the times under Setup A is 7 seconds and the standard deviation for the times under Setup B is 5 seconds.

- How would you incorporate this new information into your decision?

Suppose, instead, that the standard deviation of times under Setup A is 35 seconds and the standard deviation under setup B is 25 seconds.

- How should you adjust your appraisal of the setups’ relative benefits?

Statistical Reasoning

The preceding example calls for *statistical reasoning*.

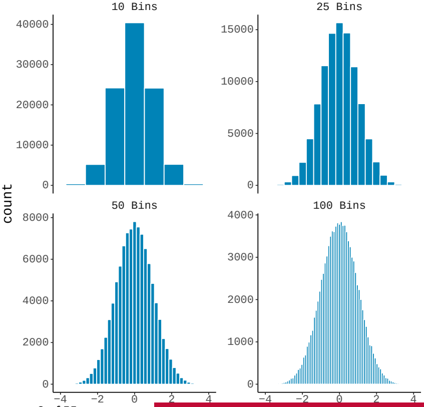
- The foundation of all good statistical analyses is a deliberate, careful, and thorough consideration of uncertainty.
- In the previous example, the mean lap time for Setup A is clearly longer than the mean lap time for Setup B.
- If the times are highly variable, with respect to the size of the mean difference, we may not care much about the mean difference.
- The purpose of statistics is to systematize the way that we account for uncertainty when making data-based decisions.

Probability Distributions

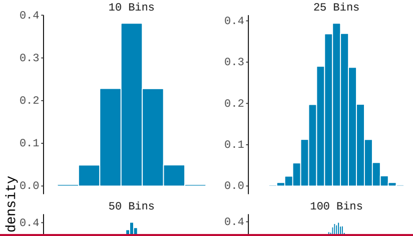
Statisticians (and anyone who uses statistics) quantify uncertainty using probability distributions.

- Probability distributions quantify how likely it is to observe each possible value of some probabilistic entity.
- Probability distributions are re-scaled frequency distributions.
- We can build up the intuition of a probability density by beginning with a histogram.

Probability Distributions



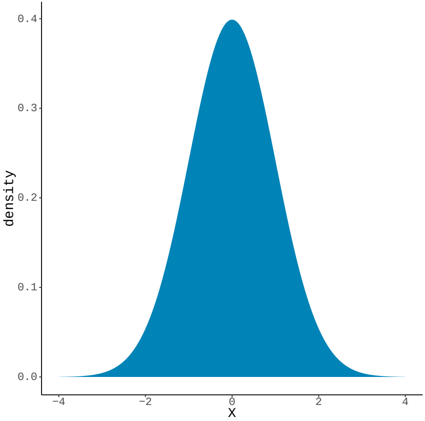
```
## Warning:  The dot-dot notation
(‘..density..‘) was deprecated in
ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘
instead.
## This warning is displayed once
every 8 hours.
## Call
‘lifecycle::last_lifecycle_warnings()‘
to see where this warning was
generated.
```



Probability Distributions

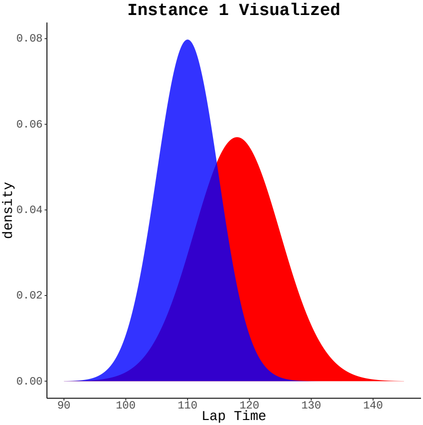
With an infinite number of bins, a histogram smooths into a continuous curve.

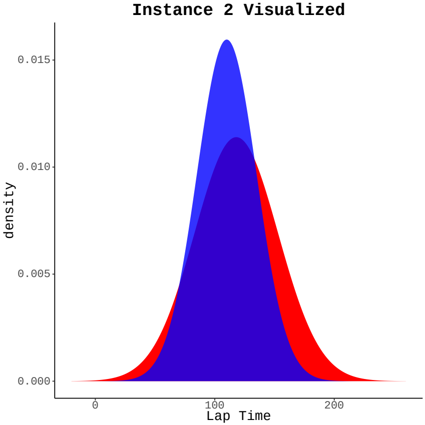
- In a loose sense, each point on the curve gives the probability of observing the corresponding X value in any given sample.
- The area under the curve must integrate to 1.0.



Reasoning with Distributions

We will gain insight by conceptualizing our example problem in terms of the underlying distributions of lap times.


Instance 1 Visualized


Instance 2 Visualized

Statistical Testing

In practice, we may want to distill the information in the preceding plots into a simple statistic so we can make a judgment.

- One way to distill this information and control for uncertainty when generating knowledge is through statistical testing.
    - When we conduct statistical tests, we weight the estimated effect by the precision of the estimate.
- A common type of statistical test, the *Wald Test*, follows this pattern:

$$T = \frac{\textit{Estimate} - \textit{Null-Hypothesized Value}}{\textit{Variability}}$$

Statistical Testing

If we want to test the null hypothesis of a zero mean difference, applying Wald test logic to control for the uncertainty in our estimate results in the familiar *t-test*:

$$t = \frac{\left(\bar{X}_A - \bar{X}_B\right) - 0}{\sqrt{S^2_{A-B}\left(n_A^{-1} + n_B^{-1}\right)}}$$

where

$$Estimate = \bar{X}_A - \bar{X}_B$$

and

$$\begin{aligned} Variability &= \sqrt{S^2_{A-B}\left(n_A^{-1} + n_B^{-1}\right)} \\ &= \sqrt{\frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} \end{aligned}$$

Statistical Testing

Applying the preceding formula to the first instantiation of our example problem produces:

$$t = \frac{118 - 110 - 0}{\sqrt{\frac{(100-1)7^2+(100-1)5^2}{100+100-2}\left(\frac{1}{100}+\frac{1}{100}\right)}}$$

$$\approx \frac{8}{0.86}$$

$$\approx 9.30$$

Statistical Testing

If we consider the second instantiation of our example problem, the effect does not change, but our measure of variability does:

$$V = \sqrt{\frac{(100-1)35^2 + (100-1)25^2}{100+100-2}\left(\frac{1}{100}+\frac{1}{100}\right)}$$
$$\approx 4.30$$

As a results, our test statistic changes to reflect our decreased certainty:

$$t \approx \frac{8}{4.30} \approx 1.86$$

Statistical Testing

Of course, we can do the same analysis in R:

```r
xA <- scale(rnorm(100)) * 7 + 118
xB <- scale(rnorm(100)) * 5 + 110

mean(xA); sd(xA)

## [1] 118
## [1] 7

mean(xB); sd(xB)

## [1] 110
## [1] 5
```

Statistical Testing

```r
out <- t.test(x = xA, y = xB, var.equal = TRUE)
wrap(out)

##
##  Two Sample t-test
##
## data:  xA and xB
## t = 9.2998, df = 198, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not
## equal to 0
## 95 percent confidence interval:
##  6.303606 9.696394
## sample estimates:
## mean of x mean of y
##       118       110
```

Statistical Testing

We can also consider the second version of our problem:

```
xA2 <- scale(rnorm(100)) * 35 + 118
xB2 <- scale(rnorm(100)) * 25 + 110

mean(xA2); sd(xA2)

## [1] 118
## [1] 35

mean(xB2); sd(xB2)

## [1] 110
## [1] 25
```

Statistical Testing

```
out <- t.test(x = xA2, y = xB2, var.equal = TRUE)
wrap(out)

##
##  Two Sample t-test
##
## data:  xA2 and xB2
## t = 1.86, df = 198, p-value = 0.06437
## alternative hypothesis: true difference in means is not
## equal to 0
## 95 percent confidence interval:
##  -0.4819679 16.4819679
## sample estimates:
## mean of x mean of y
##       118       110
```

Statistical Testing

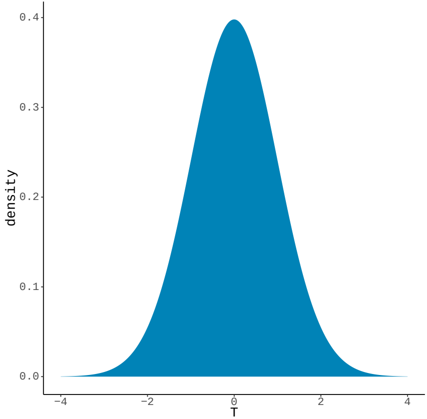
We’ve computed a test statistic, but how do we use it to compare lap times under Setups A and B?

- A test statistic, by itself, is just an arbitrary number.
- To conduct the test, we need to compare the test statistic to some objective reference.
- This objective reference needs to tell us something about how exceptional our test statistic is.
- The specific reference we will be employing is known as a *sampling distribution* of the test statistic.

Sampling Distribution

A sampling distribution is simply the probability distribution of a parameter.

- The *population* is defined by an infinite sequence of repeated tests.
    - The sampling distribution quantifies the possible values of the test statistic over infinite repeated sampling.
- The area of a region under the curve represents the probability of observing a *test statistic* within the corresponding interval.



Sampling Distributions

Note that a sampling distribution is a slightly different concept than the distribution of a random variable.

- The sampling distribution quantifies the possible values of a statistic (e.g., mean, t-statistic, correlation coefficient, etc.).
- The distribution of a random variable quantifies the possible values of a variable (e.g., age, gender, income, movie preferences, etc.).

Sampling Distributions

Note that a sampling distribution is a slightly different concept than the distribution of a random variable.

- The sampling distribution quantifies the possible values of a statistic (e.g., mean, t-statistic, correlation coefficient, etc.).
- The distribution of a random variable quantifies the possible values of a variable (e.g., age, gender, income, movie preferences, etc.).

The t-test we've been considering is a way to summarize the comparison of two variables' distributions.

- The t-statistic also has a sampling distribution that quantifies the possible t-values we could get if we repeatedly drew samples from the variables' distributions and re-computed a t-statistic each time.
- `http://onlinestatbook.com/stat_sim/sampling_dist/`

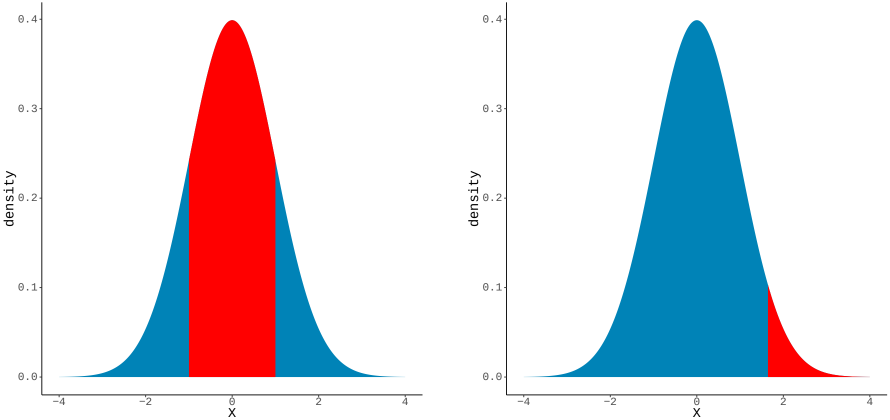
Statistical Testing

To quantify how exceptional our estimated t-statistic is, we compare the estimated value to a sampling distribution of t-statistics *assuming no effect*.

- This distribution quantifies the *null hypothesis*.
  - The special case of a null hypothesis of no effect is called the *nil-null*.
- If our estimated statistic would be very unusual in a population where the null hypothesis is true, we reject the null and claim a “statistically significant” effect.

Computing the Probability of Events

We can find the probability associated with a range of values (i.e., a range of possible events, variable values, or statistics) by computing the area of the corresponding slice from the distribution.



P-Values

By calculating the area in the null distribution that exceeds our estimated test statistic, we can compute the probability of observing the given test statistic, or one more extreme, if the null hypothesis were true.

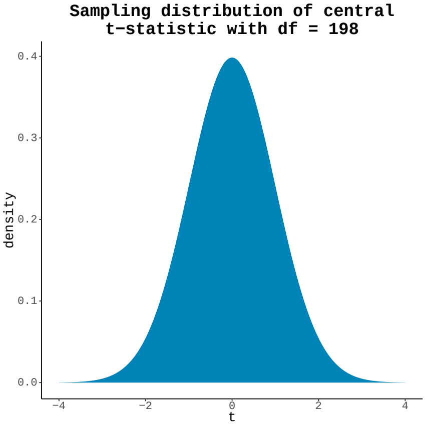
- In other words, we can compute the probability of having sampled the data we observed, or more unusual data, from a population wherein there is no true mean difference in lap times.

This value is the infamous *p-value*.

P-Values

```
tOut <-
    t.test(x         = xA2,
           y         = xB2,
           var.equal = TRUE)
tHat <- tOut$statistic
tHat

##        t
## 1.859962
```



Sampling distribution of central t-statistic with df = 198

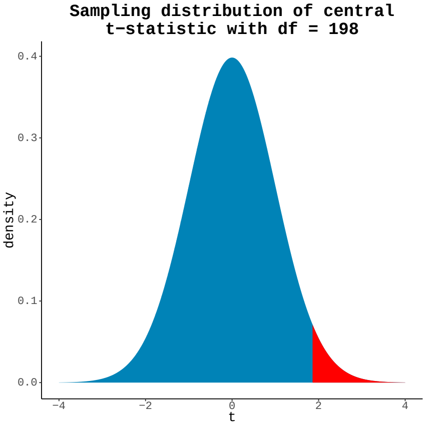
P-Values

Find the area higher than $\hat{t}$:

```
pt(q          = tHat,
   df         = 198,
   lower.tail = FALSE)

##          t
## 0.03218702
```

Hmm...this value looks too small. Why?



Sampling distribution of central t-statistic with df = 198

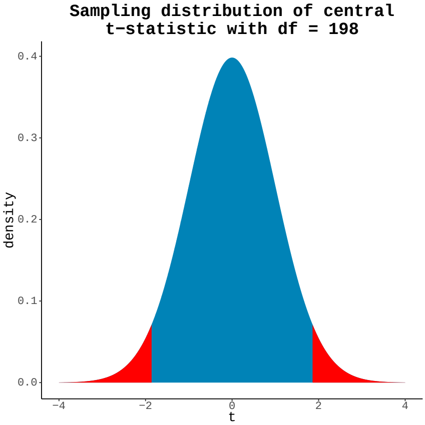
P-Values

The preceding test is *one-tailed*.

- We use a one-tailed test when we have directional hypotheses.
- Since we didn’t expect Setup B to out-perform Setup A, we need to use a two-tailed test.

```
2 * pt(q          = tHat,
       df         = 198,
       lower.tail = FALSE)

##          t
## 0.06437404
```


Sampling distribution of central t−statistic with df = 198

Interpreting P-Values

Consider the one-tailed test for our estimated test-statistic of $\hat{t} = 1.86$ that produces a p-value of $p = 0.032$.

- We *cannot* say that there is a 0.032 probability that the true mean difference is greater than zero.
- We *cannot* say that there is a 0.032 probability that the alternative hypothesis is true.
- We *cannot* say that there is a 0.032 probability that the null hypothesis is false.
- We *cannot* say that there is a 0.032 probability that the observed result is due to chance alone.
- We *cannot* say that there is a 0.032 probability of replicating the observed effect in future studies.

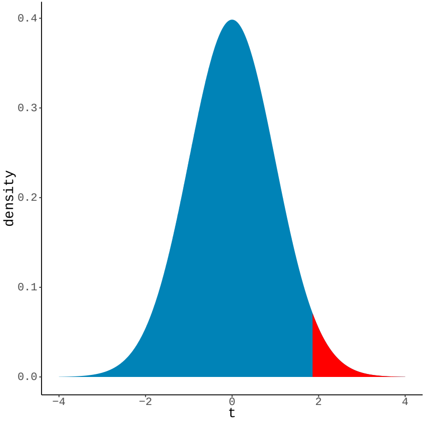
Interpreting P-Values

The p-value tells us $P(t \geq \hat{t} | H_0)$

- What we really want to know is $P(H_0 | t \geq \hat{t})$.

All that we can say is that there is a 0.032 probability of observing a test statistic at least as large as $\hat{t}$, if the null hypothesis is true.

- Our test uses the same logic as *proof by contradiction*.



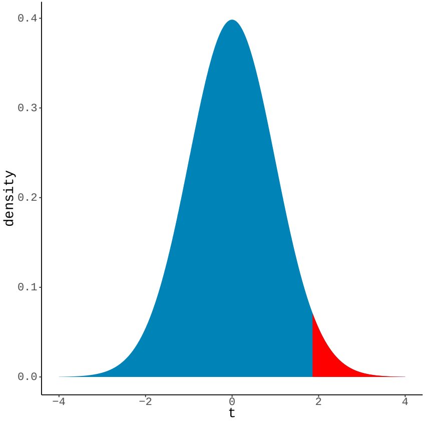
Interpreting P-Values

Note that $P(t \geq \hat{t}|H_0) \neq P(t = \hat{t}|H_0)$

- We *cannot* say that there is a 0.032 probability of observing $\hat{t}$, if the null hypothesis is true.

The probability of observing any individual point on a continuous distribution is exactly zero.

- $P(t = \hat{t}|H_0) = 0$



Statistical Modeling

Statistical testing is a very useful tool, but it quickly reaches a limit.

- In experimental contexts, real-world “messiness” is controlled through random assignment, and statistical testing is a sufficient method of knowledge generation.
- Data scientists rarely have the luxury of being able to conduct experiments.
- Data scientists work with messy observational data and usually don’t have questions that lend themselves to rigorous testing.

Data scientists need *statistical modeling*.

Statistical Modeling

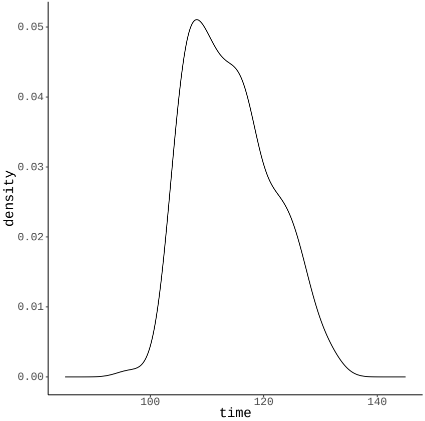
- Modelers attempt to build a mathematical representation of the (interesting aspects) of a data distribution.
- The model succinctly describes whatever system is being analyzed.
- Beginning with a model ensures that we are learning the important features of a distribution.
- The modeling approach is especially important in messy data science applications where clear a priori hypotheses are rare.

Statistical Modeling

To apply a modeling approach to our example problem we consider the combined distribution of lap times.

- The model we construct will explain variation in lap times based on interesting features.

- In this simple case, the only feature we consider is the type of setup.



Modeling our Example

Let's say we're willing to assume that the (conditional) distribution of lap times is normal.

$$Y_{time} \sim \mathrm{N}\left(\mu, \sigma^2\right)$$

To get the same answer as our statistical test, we model the mean of the distribution of lap times, μ, using a single grouping factor.

$$\mu = \beta_0 + \beta_1 X_{setup}$$

$$Y_{time} \sim \mathrm{N}\left(\beta_0 + \beta_1 X_{setup}, \sigma^2\right)$$

Modeling our Example

Since we're mostly interested in describing the mean lap time, we can express the above differently:

$$Y_{time} = \beta_0 + \beta_1 X_{setup} + \varepsilon$$

$$\varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$$

After we fit this model to a sample, the parameters β_0 and β_1 are replaced by estimated statistics.

$$\begin{aligned}\hat{Y}_{time} &= \hat{\beta}_0 + \hat{\beta}_1 X_{setup} \\ &= 110 + 8X_{setup}\end{aligned}$$

Modeling our Example

We can easily fit this model in R:

```r
lmOut <- lm(time ~ setup, data = exData)

partSummary(lmOut, -c(1, 2))

## Error in summary(x, ...)  %>% quiet():  could not find function "%>%"
```

Two Modeling Traditions

Breiman (2001) defines two cultures of statistical modeling:

- Data models
- Algorithmic models

Two Modeling Traditions

Breiman (2001) defines two cultures of statistical modeling:

- Data models
- Algorithmic models

Data scientists use both types of models.

- Both types of model have strengths and weaknesses.
    - Data models tend to support a priori hypothesis testing more easily.
    - Data models also tend to provide more interpretable results.
    - Algorithmic models can’t be beat for pure power.

Two Modeling Traditions

Breiman (2001) defines two cultures of statistical modeling:

- Data models
- Algorithmic models

Data scientists use both types of models.

- Both types of model have strengths and weaknesses.
    - Data models tend to support a priori hypothesis testing more easily.
    - Data models also tend to provide more interpretable results.
    - Algorithmic models can’t be beat for pure power.
- Algorithmic models are currently preferred in cutting edge prediction/classification applications.

Two Modeling Traditions

Breiman (2001) defines two cultures of statistical modeling:

- Data models
- Algorithmic models

Data scientists use both types of models.

- Both types of model have strengths and weaknesses.
  - Data models tend to support a priori hypothesis testing more easily.
  - Data models also tend to provide more interpretable results.
  - Algorithmic models can’t be beat for pure power.
- Algorithmic models are currently preferred in cutting edge prediction/classification applications.
- Many models can be viewed as data models or algorithmic models, depending on how they’re used.

Characteristics of Models

Data models share several core features:

- Data models are built from probability distributions.
    - Data models are modular.
- Data models encode our hypothesized understanding of the system we're exploring.
    - Data models are constructed in a "top-down", theory-driven way.

Characteristics of Models

Data models share several core features:

- Data models are built from probability distributions.
  - Data models are modular.
- Data models encode our hypothesized understanding of the system we're exploring.
  - Data models are constructed in a "top-down", theory-driven way.

Algorithmic models are distinct from data models in several ways:

- Algorithmic models do not have to be built from probability distributions.
  - Often, they are based on a set of decision rules (i.e., an algorithm).
- Algorithmic models begin with an objective (i.e., a problem to solve) and seek the optimal solution, given the data.
  - They are built in a "bottom-up", data-driven way.

Data Modeling Example

Suppose we believe the following:

1. BMI is positively associated with disease progression in diabetic patients after controlling for age and average blood pressure.
2. After controlling for age and average blood pressure, the effect of BMI on disease progression is different for men and women.

We can represent these beliefs with a moderated regression model:

$$Y_{prog} = \beta_0 + \beta_1 X_{BMI} + \beta_2 X_{sex} + \beta_3 X_{age} + \beta_4 X_{BP} + \beta_5 X_{BMI} X_{sex} + \varepsilon$$

Data Modeling Example

We can use R to fit our model to some patient data:

```r
library(rockchalk)

##
## Attaching package:  ’rockchalk’
## The following object is masked from ’package:plyr’:
##
##     summarize

## Load the data:
dataDir <- "../../data/"
dDat    <- readRDS(paste0(dataDir, "diabetes.rds"))

## Fit the regression model:
fit <- lm(progress ~ bmi * sex + age + bp, data = dDat)
```

Data Modeling Example

```r
partSummary(fit, -c(1, 2))

## Error in summary(x, ...)  %>% quiet():  could not find function "%>%"
```

Data Modeling Example

We can do a simple slopes analysis to test the group-specific effects of BMI on disease progression:

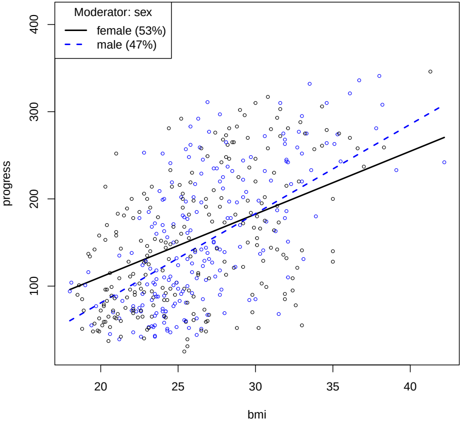
```
psOut <- plotSlopes(fit, plotx = "bmi", modx = "sex")
tsOut <- testSlopes(psOut)
```

```
tsOut$hypotests[ , -1]

##              slope Std. Error  t value     Pr(>|t|)
## female  7.210575  0.8921929 8.081856 6.335264e-15
## male   10.236323  1.0328739 9.910525 5.137409e-21
```

Data Modeling Example

We can also visualize the simple slopes:



Algorithmic Modeling Example

Suppose we want to find the best predictors of disease progression among the variables contained in our dataset:

- Age
- BMI
- Blood Pressure
- Blood Glucose
- Sex
- Total Cholesterol
- LDL Cholesterol
- HDL Cholesterol
- Triglycerides
- Lamorigine

We could try *best-subset selection*.

- Fit a series of regression models wherein disease progression is predicted by all possible subsets of X variables.
- Choose the set of X variables that minimizes the prediction error.

Algorithmic Modeling Example

```
library(leaps)

## Error in library(leaps):  there is no package called ’leaps’

## Save the predictor variables' names:
xNames <- grep(pattern = "progress",
               x       = colnames(dDat),
               invert  = TRUE,
               value   = TRUE)

## Train the models:
fit <- regsubsets(x     = progress ~ .,
                  data  = dDat,
                  nvmax = ncol(dDat) - 1)

## Error in regsubsets(x = progress ~ ., data = dDat, nvmax = ncol(dDat) -
:  could not find function "regsubsets"

## Summarize the results:
sum <- summary(fit)
```

Algorithmic Modeling Example

```
sum$outmat

## NULL
```

Algorithmic Modeling Example

```
## Variables selected by BIC:
xNames[with(sum, which[which.min(bic), -1])]

## Error in eval(substitute(expr), data, enclos = parent.frame()):  object
’bic’ not found

## Variables selected by Adjusted R^2:
xNames[with(sum, which[which.max(adjr2), -1])]

## Error in eval(substitute(expr), data, enclos = parent.frame()):  object
’adjr2’ not found

## Variables selected by Mallow's Cp:
xNames[with(sum, which[which.min(cp), -1])]

## Error in eval(substitute(expr), data, enclos = parent.frame()):  object
’cp’ not found
```

Algorithmic Modeling Example

The results seem to be highly sensitive to the error measure. What should we do?

Algorithmic Modeling Example

The results seem to be highly sensitive to the error measure. What should we do?

- We could pick our favorite error measure and use its results.
- We could throw our hands up in defeat and quit.
- We could look at the results and pick the answer we like best.
  - The previous two suggestions are sub-optimal, but this one is actually unethical. Don't do this!

Algorithmic Modeling Example

The results seem to be highly sensitive to the error measure. What should we do?

- We could pick our favorite error measure and use its results.
- We could throw our hands up in defeat and quit.
- We could look at the results and pick the answer we like best.
  - The previous two suggestions are sub-optimal, but this one is actually unethical. Don't do this!

If we think like a data scientist and get creative, we don't need to settle for these ambiguous results.

- We could implement a more robust method of calculating prediction error like *K-fold cross validation*.
- We can use resampling methods to quantify uncertainty in the variable selection process.

Algorithmic Modeling Example

```
bic <- r2 <- cp <- matrix(NA, 100, ncol(dDat) - 1)
for(rp in 1 : 100) {
    ## Resample the data:
    tmp <- dDat[sample(1 : nrow(dDat), nrow(dDat), TRUE), ]

    ## Train the models:
    fit <- regsubsets(x     = progress ~ .,
                      data  = tmp,
                      nvmax = ncol(tmp) - 1)
    sum <- summary(fit)

    ## Save the optimal selections:
    bic[rp, ] <- with(sum, which[which.min(bic), -1])
    r2[rp, ]  <- with(sum, which[which.max(adjr2), -1])
    cp[rp, ]  <- with(sum, which[which.min(cp), -1])
}

## Error in regsubsets(x = progress ~ ., data = tmp, nvmax = ncol(tmp) - :
could not find function "regsubsets"
```

Algorithmic Modeling Example

```
colMeans(bic)

##  [1] NA NA NA NA NA NA NA NA NA NA

colMeans(r2)

##  [1] NA NA NA NA NA NA NA NA NA NA

colMeans(cp)

##  [1] NA NA NA NA NA NA NA NA NA NA
```

Algorithmic Modeling Example

```
## Find the best subset via majority vote:
votes <- colMeans(rbind(bic, r2, cp)); round(votes, 3)

##  [1] NA NA NA NA NA NA NA NA NA NA

preds <- xNames[votes > 0.5]; preds

##  [1] NA NA NA NA NA NA NA NA NA NA

## Fit the winning model to the original data:
form <- paste0("progress ~ ",
               paste(preds, collapse = " + ")
               )
fit  <- lm(form, data = dDat)

## Error in terms.formula(formula, data = data):  invalid model formula in
ExtractVars
```

Algorithmic Modeling Example

```
partSummary(fit, -c(1, 2))

## Error in summary(x, ...)  %>% quiet():  could not find function "%>%"
```

Model-Based Prediction

So far, our discussion has centered on inference about estimated model parameters.

- The mean difference between lap times under Setups A and B.
- We modeled the system and scrutinized $\hat{\beta}_1$ to make inferences about the mean difference in lap times.

Model-Based Prediction

So far, our discussion has centered on inference about estimated model parameters.

- The mean difference between lap times under Setups A and B.
- We modeled the system and scrutinized $\hat{\beta}_1$ to make inferences about the mean difference in lap times.

In data science applications, we're often more interested in predicting the outcome for new observations.

- After we estimate $\hat{\beta}_0$ and $\hat{\beta}_1$, we can plug in new predictor data and get a predicted outcome value for any new case.
- In our example, these predictions represent the projected lap times under the different setups.

Inference vs. Prediction

When doing statistical inference, we focus on how certain variables relate to the outcome.

- Do men have higher job-satisfaction than women?
- Does increased spending on advertising correlate with more sales?
- Is there a relationship between the number of liquor stores in a neighborhood and the amount of crime?

Inference vs. Prediction

When doing statistical inference, we focus on how certain variables relate to the outcome.

- Do men have higher job-satisfaction than women?
- Does increased spending on advertising correlate with more sales?
- Is there a relationship between the number of liquor stores in a neighborhood and the amount of crime?

When doing prediction, we want to build a tool that can accurately guess future values.

- Will it rain tomorrow?
- Will this investment turn a profit within one year?
- Will increasing the number of contact hours improve grades?

References

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.