Statistical Modeling

Fundamental Techniques in Data Science

Utrecht University

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Statistical Modeling

Different Flavors of Statistical Modeling

- Data Modeling
- Algorithmic Modeling

Statistical Reasoning

Statistics and data science are used to answer questions about hypothetical populations.

- Do men have higher job satisfaction than women?
- Can I predict your voting behavior?
- Can I detect groups of people who share similar attitudes towards climate change?

To answer these questions, we need to use *statistical reasoning*.

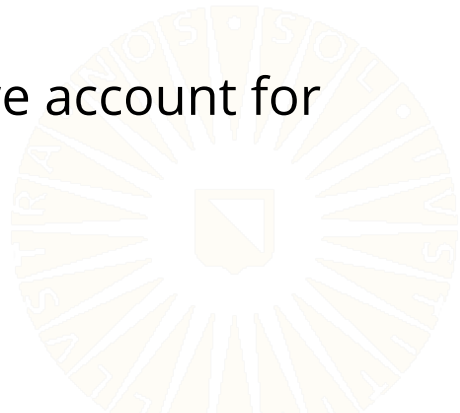
- The foundation of all good statistical analyses is a deliberate, careful, and thorough consideration of uncertainty.

Statistical Reasoning

If I measure a mean satisfaction rating for men of 5.6 and a mean satisfaction rating for women of 5.1, does that imply higher job satisfaction for men?

- Maybe…
- If the satisfaction ratings are highly variable, with respect to the size of the mean difference, we may not care much about the observed mean difference.
- The *observed* mean difference may not represent a *true* mean difference in the population.

The purpose of statistics is to systematize the way that we account for uncertainty when making data-based decisions.

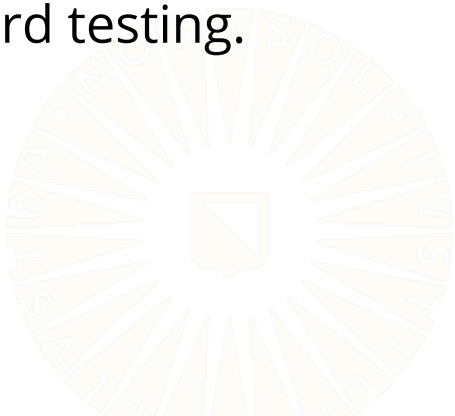


Statistical Modeling

To implement this "statistical reasoning," we could use two different approaches: *statistical testing* or *statistical modeling*.

- In experimental contexts, real-world "messiness" is controlled through random assignment, and statistical testing is a sufficient method of knowledge generation.
- Apart from A/B testing, data scientists rarely have the luxury of being able to conduct experiments.
- Data scientists work with messy observational data and often don't have questions that lend themselves to straight-forward 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.

Different Flavors of Statistical Modeling

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

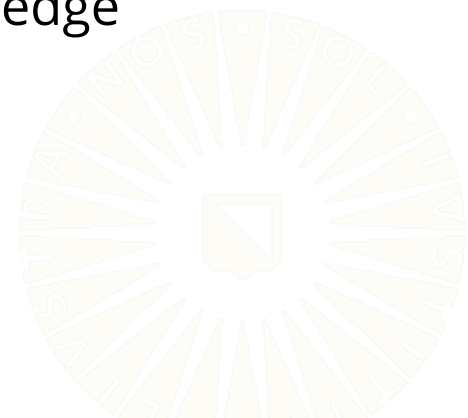
Breiman (2001) defines two cultures of statistical modeling:

- 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:

```
library(dplyr)
library(rockchalk)

## Load the data:
diabetes <- readRDS("../data/diabetes.rds")
diabetes <- rename(diabetes, sex = sexF)

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

Data Modeling Example

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

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -174.7986    27.0004  -6.474 2.58e-10
bmi            7.2106     0.8922   8.082 6.34e-15
sexmale      -90.1718    35.1134  -2.568   0.0106
age            0.1691     0.2322   0.728   0.4670
bp             1.4032     0.2385   5.884 7.97e-09
bmi:sexmale    3.0257     1.3090   2.311   0.0213

Residual standard error: 59.68 on 436 degrees of freedom
Multiple R-squared:  0.4075,Adjusted R-squared:  0.4007
F-statistic: 59.98 on 5 and 436 DF,  p-value: < 2.2e-16
```

Data Modeling Example

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

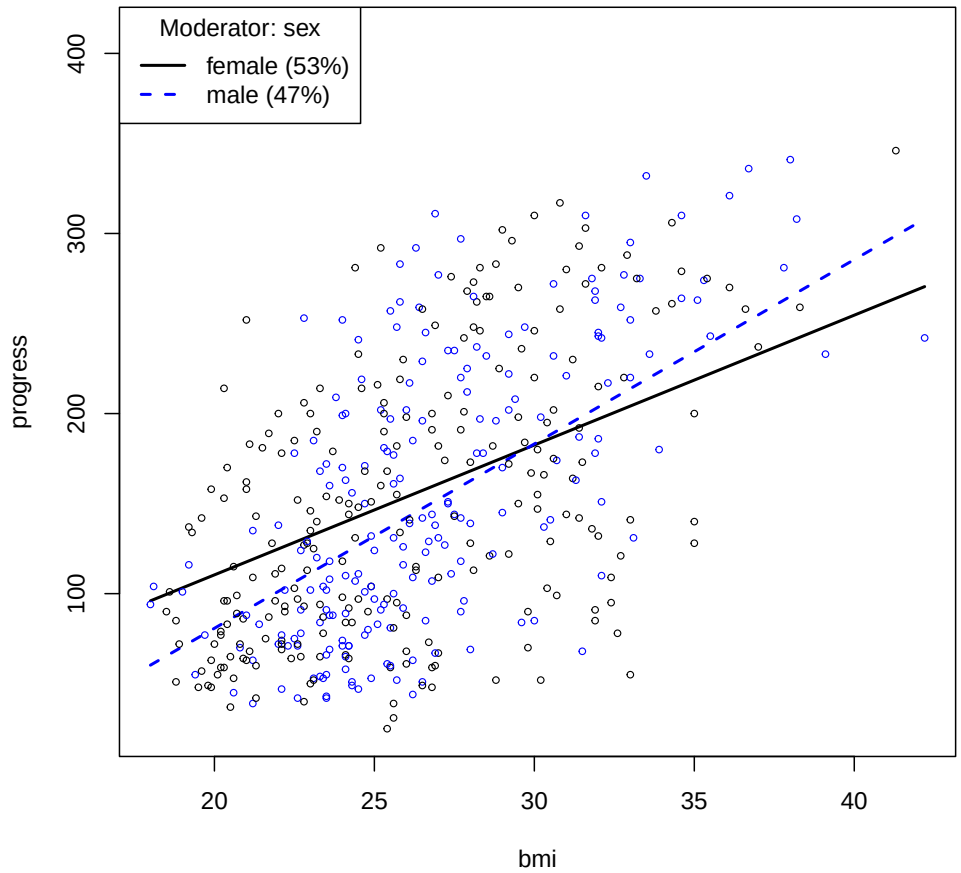
```
psOut <- plotSlopes(fit, plotx = "bmi", modx = "sex")
tsOut <- testSlopes(psOut)
```

```
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)

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

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

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

Algorithmic Modeling Example

```
sum$outmat

           age sexN bmi bp  tc  ldl hdl tch ltg glu sexmale
1   ( 1 )  " " " "  "*" " " " " " " " " " " " " " " " " " "
2   ( 1 )  " " " "  "*" " " " " " " " " " " " " "*" " " " "
3   ( 1 )  " " " "  "*" "*" " " " " " " " " " " "*" " " " "
4   ( 1 )  " " " "  "*" "*" "*" " " " " " " " " "*" " " " "
5   ( 1 )  " " " "  "*" "*" " " " " " " "*" " " "*" " " "*"
6   ( 1 )  " " " "  "*" "*" "*" "*" " " " " " " "*" " " "*"
7   ( 1 )  " " "*"  "*" "*" "*" "*" " " " " "*" "*" " " " "
8   ( 1 )  " " " "  "*" "*" "*" "*" " " " " "*" "*" "*" "*"
9   ( 1 )  " " " "  "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
10  ( 1 ) "*" "*"  "*" "*" "*" "*" "*" "*" "*" "*" " " " "
```

Algorithmic Modeling Example

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

[1] "bmi" "bp"  "hdl" "ltg" "sex"

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

[1] "bmi" "bp"  "tc"  "ldl" "tch" "ltg" "glu" "sex"

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

[1] "bmi" "bp"  "tc"  "ldl" "ltg" "sex"
```

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(diabetes) - 1)
for(rp in 1 : 100) {
    ## Resample the data:
    tmp <- diabetes[sample(1 : nrow(diabetes), nrow(diabetes), 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])
}
```

Algorithmic Modeling Example

```
colMeans(bic)

    age    sexN     bmi      bp      tc     ldl     hdl
   0.01    0.40    1.00    1.00    0.59    0.32    0.40
    tch     ltg     glu sexmale
   0.27    1.00    0.06    0.51

colMeans(r2)

    age    sexN     bmi      bp      tc     ldl     hdl
   0.28    0.54    1.00    1.00    0.89    0.67    0.36
    tch     ltg     glu sexmale
   0.61    1.00    0.50    0.46

colMeans(cp)

    age    sexN     bmi      bp      tc     ldl     hdl
   0.16    0.44    1.00    1.00    0.86    0.50    0.26
    tch     ltg     glu sexmale
   0.50    1.00    0.35    0.54
```

Algorithmic Modeling Example

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

    age    sexN     bmi      bp      tc     ldl     hdl
  0.150   0.460   1.000   1.000   0.780   0.497   0.340
    tch     ltg     glu sexmale
  0.460   1.000   0.303   0.503

preds <- xNames[votes > 0.5]; preds

[1] "bmi" "bp"  "tc"  "ltg" "sex"

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

Algorithmic Modeling Example

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

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -335.11146   25.68289 -13.048  < 2e-16
bmi            6.47376    0.68565   9.442  < 2e-16
bp             1.05016    0.21789   4.820 1.99e-06
tc            -0.29836    0.08833  -3.378 0.000796
ltg           60.36010    6.49158   9.298  < 2e-16
sexmale      -14.14306    5.40833  -2.615 0.009231

Residual standard error: 54.83 on 436 degrees of freedom
Multiple R-squared:  0.4999,Adjusted R-squared:  0.4941
F-statistic: 87.15 on 5 and 436 DF,  p-value: < 2.2e-16
```

References

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