Tools from Modern Statistics: Review and Roadmap

Data Science for Marketing Decision Making
Günter J. Hitsch
Chicago Booth

Winter 2017

Overview

1. Goals of statistical inference and learning
2. What you already need to know about linear regression
3. Consistency, bias, and the variance-bias trade-off
4. Modern statistical tools and machine learning: Big data and flexibility
5. Causality

Terminology

Data-driven marketing applications all have this structure:

- We observe variables $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, called *covariates*, *predictors*, *independent variables* (more traditional statistics language), *inputs*, or *features* (machine learning language)
- We observe y_i, called the *dependent variable* or *output*
- y_i can be quantitative (market share) or qualitative ("customer has a Facebook account")

Our sample (data set) is $((y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n))$

Example: Marketing mix modeling

- Advertising/media, price, promotions, distribution
- Sales units

Goals of statistical inference and learning

The data are drawn from a distribution F,

$$(Y, X) \sim F$$

F captures everything about the relationship between X and Y that we would like to know

- Expected spending and variance of spending given demographics
- Likelihood of default given credit history

Goal: Learn F or some aspects of F from the data

In marketing applications:

- Rarely: Learn probability density $f(y, \boldsymbol{x})$
- Frequently: Learn some aspect(s) of the distribution F, such as the conditional expectation $\mathbb{E}(Y|X)$

Regression analysis

Model the relationship between X and Y using the regression function, $r(\boldsymbol{x}) = \mathbb{E}(Y|X = \boldsymbol{x})$

Using the regression function we can express the relationship between the inputs and the output:

$$Y = \mathbb{E}(Y|X) + \epsilon, \qquad \mathbb{E}(\epsilon|X) = 0$$

- ϵ (the error term) captures all factors that affect the output and that we do not measure

Applications:

- Prediction of Y given the inputs/dependent variables X
- Study the relationship between X and Y. E.g., does competitor advertising increase or decrease sales of our pharmaceutical drug?
- Hypothesis testing. For example, which of the inputs have a non-zero relationship with the output?

Special case: The linear regression model

How I (and probably you, too) learned statistics:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- Assumption: $\mathbb{E}(\epsilon|X_1, \ldots, X_p) = 0$
- The error term (= unmeasured factors related to Y) is *mean-independent* of the measured factors X_k, and consequently also uncorrelated with each input

Checklist:

- Interpretation of regression coefficients
- Standard errors, t-statistics, p-values, confidence intervals
- R^2 and F-statistic (typically useless in marketing applications)
- Dummy variables to code categorical inputs (gender, ZIP code, past default status, ...)

Example: Income and soda consumption

```
fit = lm(volume_capita ~ income, data = purchases_hh[category=="CSD"])
summary(fit)

Call:
lm(formula = volume_capita ~ income, data = purchases_hh[category ==
    "CSD"])

Residuals:
   Min     1Q Median     3Q    Max
 -8.16  -4.60  -2.75   1.23 574.41

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.285494   0.085384   97.04   <2e-16 ***
income      -0.042948   0.001271  -33.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.184 on 51043 degrees of freedom
Multiple R-squared:  0.02188,Adjusted R-squared:  0.02186
F-statistic:  1142 on 1 and 51043 DF,  p-value: < 2.2e-16
```


Desirable properties of estimators

$\hat{\theta}_n$ is an **estimator** for some unknown quantity of interest, θ

- $\hat{\theta}_n$ depends on the data
- Quantity of interest could be a model parameter (e.g. a regression coefficient β_k) or multiple parameters, or the prediction of some future outcome Y

Bias of $\hat{\theta}_n$:

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$$

- Average, systematic difference between the estimate and the true quantity of interest

An estimator is **unbiased** if

$$\mathbb{E}(\hat{\theta}_n) = \theta$$

- No systematic difference between estimate and the truth

Desirable properties of estimators

$\hat{\theta}_n$ is **consistent** for θ if

$$\hat{\theta}_n \xrightarrow{p} \theta$$

- This expression means that the estimate $\hat{\theta}_n$ converges in probability to the true value θ when the sample size, n, grows large
- To be precise: For any $\delta > 0$, $\Pr\{|\hat{\theta}_n - \theta| < \delta\} \to 1$ as $n \to \infty$
- More intuitively, for a large sample size n the probability that the estimate $\hat{\theta}_n$ is arbitrarily close to to the true quantity θ is arbitrarily close to 1

For a consistent estimator, the bias goes to zero when the sample size grows large

- Consistent estimators are **asymptotically unbiased**

Consistency and bias

Consistency is a key requirement for a good estimator

- It would be highly worrisome if even for an arbitrarily large data set the estimator would never be close to the quantity θ with high probability

On the other hand, how much should we care about bias?

Old-school view: Elimination of bias is of central importance. 25 years ago we only estimated linear models, and if the true relationship between X and Y is described by a linear regression model then the standard estimation method, OLS (method of ordinary least squares), yields an unbiased estimator for the true regression coefficients, even with a small sample size n

The *modern view*: The price to be paid for an unbiased estimator can be very high. We can often predict much better if we allow for some small amount of bias in exchange for reduced variance.

The bias-variance trade-off

A standard way to evaluate goodness-of-fit of an estimator is the **mean squared error** (MSE),

$$\text{MSE} = \mathbb{E}\left[(\hat{\theta}_n - \theta)^2\right]$$

- Average squared difference between estimator and true quantity

The MSE can be expressed as follows:

$$\text{MSE} = \Big(\text{bias}(\hat{\theta}_n)\Big)^2 + \text{var}(\hat{\theta}_n)$$

- MSE = sum of squared bias and variance of estimator

Some key tools developed in the modern statistics/machine learning literature allow for some bias in exchange for lower variance, and thus yield better predictive power

Modern statistics and big data

"Old school," 25 years ago

- Linear model, small number of inputs
- Functional form assumed to be true (and even if not there was no way around a restrictive model structure)

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Statistics in the era of big data

- Big data, large p (number of inputs) relative to n (number of observations) — maybe even $p > n$
- The form of the regression function, $\mathbb{E}(Y|X = \boldsymbol{x})$, need not be linear
  - If regression function is mis-specified we get biased estimates and poor predictions
- Hence, the goal is to allow for flexible functional forms

Modern tools to estimate flexible relationships using big data are often called **machine learning methods**

Flexible functional forms in the linear regression model

The linear regression model is linear in the parameters, not in the inputs

- You can transform the inputs (and the output) in any way you want

Example: Polynomials

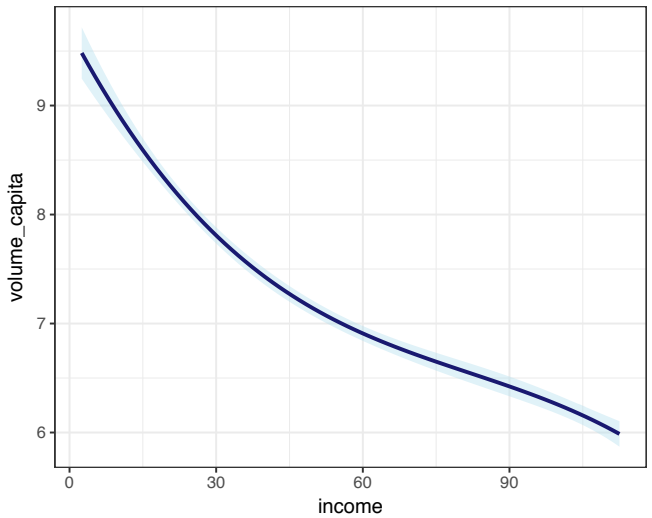
$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \dots$$

```
fit_polynomial = lm(volume_capita ~ poly(income, 3),
                    data = purchases_hh[category=="CSD"])

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.74847    0.04062 141.520  < 2e-16 ***
poly(income, 3)1 -310.31217    9.17720 -33.813  < 2e-16 ***
poly(income, 3)2   76.53826    9.17720   8.340  < 2e-16 ***
poly(income, 3)3  -28.16112    9.17720  -3.069  0.00215 **
```


Example: Polynomial fit

```
ggplot(purchases_hh, aes(x = income, y = volume_capita)) +
   geom_smooth(method = "lm", formula = y ~ poly(x, 3),
               color = "midnightblue", fill = "lightblue2", alpha = 0.4) +
   theme_bw()
```



Use any non-linear transformation, such as the logarithm:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.967804   0.035435   55.53   <2e-16 ***
log(income) -0.289032   0.008982  -32.18   <2e-16 ***
```

Limitations

- With many variables and polynomials of higher order and interactions the number of inputs explodes:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \beta_6 X_1^2 X_2 + \beta_7 X_1 X_2^2$$

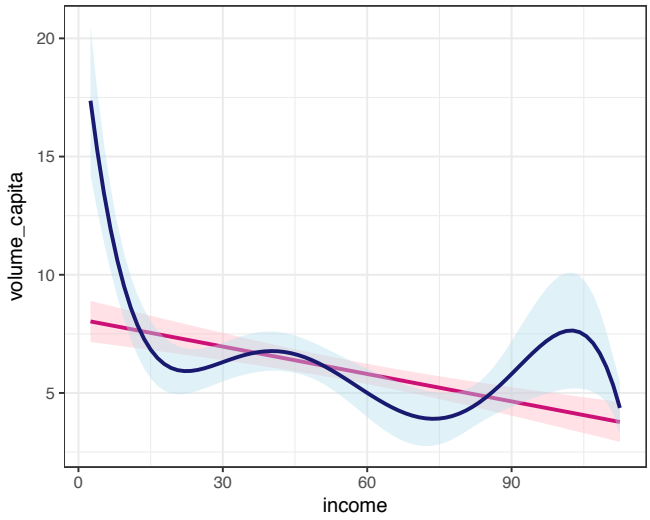
- Interactions are terms such as $\beta_k X_1 X_2$ — for example, to allow for income effects to be moderated by age or occupation

With a large number of terms we run into two key problems:

- **Feasibility**: Number of inputs larger than number of observations
- **Overfitting**: Flexibility of model reduces bias, but increases variance and yields poor out-of-sample fit

Illustration of overfitting

Estimation based on small subsample of the beverage consumption data: Polynomial (blue) overfits the data



Modern solutions

The goals:

1. Flexibly estimate relationships such as the regression function, $\mathbb{E}(Y|X = \boldsymbol{x})$
2. Incorporate "big data" — a large number of inputs with p large relative to n
3. Make accurate out-of-sample predictions and avoid overfitting

Goals 1. and 2. typically conflict with goal 3. — we buy flexibility and reduce bias at the cost of a high variance (bias-variance trade-off)

Modern statistics (machine learning) accomplishes all three goals using regularization and variable selection methods

We will encounter two important tools:

1. The LASSO
2. Random forests

LASSO and random forests

The LASSO is based on the linear regression model and incorporates regularization and variable selection methods (i.e. set some coefficients $\beta_k = 0$). The results are easy to interpret (like linear regression).

A random forest is an ensemble method based on trees, and automatically detects nonlinear relationships in the data. It is a feasible nonparametric method and achieves good predictive fit. Unlike the LASSO, random forests are harder to interpret and more of a "black box."

In this class:

- First part focuses on marketing mix modeling—use of linear regression with an emphasis on achieving consistent estimates
- Second part focuses on customer relationship management—ideal application for modern machine learning tools

Causality

Causality is a key concept in applications of statistics to business (and the social sciences in general), but interestingly only a niche subject among statisticians

We'll have a "deeper" discussion of causality in a later lecture.

For now, the basic intuition:

$$Y = \mathbb{E}(Y|X = \boldsymbol{x}) + \epsilon$$

Suppose we can **manipulate** one (or multiple) X_k's, i.e. "move" X_k while ϵ is fixed. Then the corresponding change in Y is the causal effect of the change in X_k.

- If the regression model is linear, then β_k is the causal effect of increasing X_k by 1

The important point is manipulation!

Example: Advertising and sales

Use data from *An Introduction to Statistical Learning*, http://www-bcf.usc.edu/~gareth/ISL/

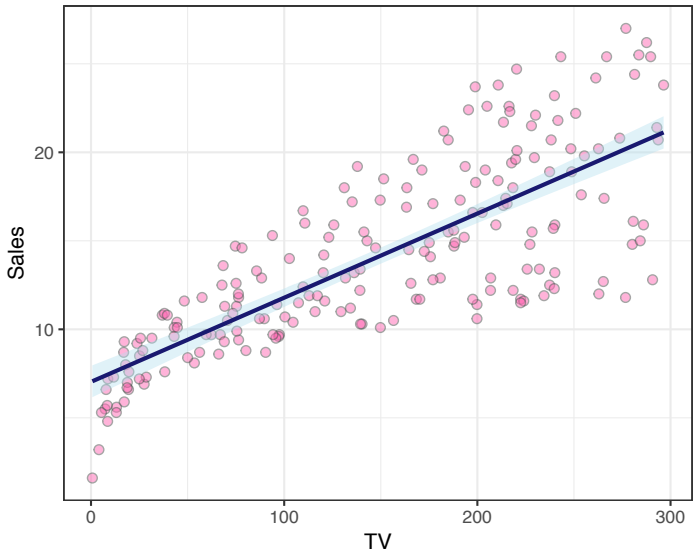
Regression of sales on three advertising variables using cross-sectional data from 200 markets:

```
advertising = fread("http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv")

fit = lm(Sales ~ TV + Radio + Newspaper, data = advertising)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
```

Regression fit:



Is the estimated relationship causal?

The question is, what is the source of the variation in the advertising levels across markets. One possible source: In some markets sales are larger than in other markets, hence the firm spends more on advertising, for example if the advertising budget is a fixed percentage of sales.

If true, the estimated relationship is not causal—we spend more on advertising if sales are typically large.

What went wrong?

$$Y_i = \beta_0 + \beta_1 \text{TV}_i + \cdots + \epsilon_i$$

(i is the index for a market)

In our example, whenever TV_i is large, ϵ_i is also large. This data-variation **does not correspond to manipulation**!
Mean-independence, $\mathbb{E}(\epsilon|X) = 0$, is violated. Thus, we cannot estimate the true, *structural* regression coefficients from the data. We still get some estimates, but these estimates only describe the statistical relationship between the variables in the data, not a causal effect.

Correlation does not imply causation

The title is almost a cliché these days—but the advertising example illustrates its meaning.

The example is one of **omitted variables** are **unmeasured confounders**:

- In some markets demand is large, hence sales are large and advertising is large

Can we fix this problem? — Yes, if we **experimentally manipulate** advertising.

As good as random variation

But even without experimental variation we can infer causal effects, if we rely on variation in the data that is **as good as random**.

Example: If the organic differences in demand in a market are constant over time, we can proxy them using a dummy coefficient δ_i:

$$Y_i = \beta_0 + \beta_1 \mathsf{TV}_i + \cdots + \delta_i + \tilde{\epsilon}_i$$

If advertising changes over time *within* a market because of changes in advertising costs, then these changes are unrelated to organic changes in demand, and hence as good as random. Then we are able to infer the causal effect of advertising from the data.

Summary

1. Goal of statistical inference and learning is to estimate relationships between the inputs and outputs in the data, e.g. a regression function
2. Need to know all about linear regression analysis
3. Understand properties of estimators:
   - Consistency and bias
   - Variance-bias trade-off
4. Modern machine learning tools achieve model flexibility, the incorporation of big data, and good out-of-sample model predictions
   - Requires regularization variable selection
5. Causal vs. statistical relationships
   - Causality requires manipulation
   - Based on actual experiments, or variation that is as good as random