Hat Notation: Estimates Versus Known Values

The “hat” notation is used to differentiate between estimates and

known values. So the symbol *b* (“b-hat”) is an estimate of the

unknown parameter *b*. Why do statisticians differentiate between

the estimate and the true value? The estimate has uncertainty,

whereas the true value is fixed.2 **G**

Regression Terminology

When analysts and researchers use the term *regression* by itself,

they are typically referring to linear regression; the focus is usually

on developing a linear model to explain the relationship between

predictor variables and a numeric outcome variable. In its formal

statistical sense, regression also includes nonlinear models that

yield a functional relationship between predictors and outcome

variables. In the machine learning community, the term is also

occasionally used loosely to refer to the use of any predictive model

that produces a predicted numeric outcome (as opposed to classification

methods that predict a binary or categorical outcome). **B**

In addition to the t-statistic, *R* and other packages will often report

a *p-value* (Pr(>|t|) in the *R* output) and *F-statistic*. Data scientists

do not generally get too involved with the interpretation of these

statistics, nor with the issue of statistical significance. Data scientists

primarily focus on the t-statistic as a useful guide for whether

to include a predictor in a model or not. High t-statistics (which go

with p-values near 0) indicate a predictor should be retained in a

model, while very low t-statistics indicate a predictor could be

dropped. See “p-Value” on page 106 for more discussion.  **O**

AIC, BIC, and Mallows Cp

The formula for AIC may seem a bit mysterious, but in fact it is

based on asymptotic results in information theory. There are several

variants to AIC:

*AICc*

A version of AIC corrected for small sample sizes.

*BIC or Bayesian information criteria*

Similar to AIC, with a stronger penalty for including additional

variables to the model.

*Mallows Cp*

A variant of AIC developed by Colin Mallows.

These are typically reported as in-sample metrics (i.e., on the training

data), and data scientists using holdout data for model assessment

do not need to worry about the differences among them or

the underlying theory behind them. **O**

Prediction Interval or Confidence Interval?

A prediction interval pertains to uncertainty around a single value,

while a confidence interval pertains to a mean or other statistic calculated

from multiple values. Thus, a prediction interval will typically

be much wider than a confidence interval for the same value.

We model this individual value error in the bootstrap model by

selecting an individual residual to tack on to the predicted value.

Which should you use? That depends on the context and the purpose

of the analysis, but, in general, data scientists are interested in

specific individual predictions, so a prediction interval would be

more appropriate. Using a confidence interval when you should be

using a prediction interval will greatly underestimate the uncertainty

in a given predicted value. **O**

Different Factor Codings

There are several different ways to encode factor variables, known

as *contrast coding* systems. For example, *deviation coding*, also

known as *sum contrasts*, compares each level against the overall

mean. Another contrast is *polynomial coding*, which is appropriate

for ordered factors; see the section “Ordered Factor Variables” on

page 169. With the exception of ordered factors, data scientists will

generally not encounter any type of coding besides reference coding

or one hot encoder. **O**

Multicollinearity is not such a problem for nonlinear regression

methods like trees, clustering, and nearest-neighbors, and in such

methods it may be advisable to retain *P* dummies (instead of *P* – 1).

That said, even in those methods, nonredundancy in predictor

variables is still a virtue. **B**

Model Selection with Interaction Terms

In problems involving many variables, it can be challenging to

decide which interaction terms should be included in the model.

Several different approaches are commonly taken:

• In some problems, prior knowledge and intuition can guide

the choice of which interaction terms to include in the model.

• Stepwise selection (see “Model Selection and Stepwise Regression”

on page 156) can be used to sift through the various

models.

• Penalized regression can automatically fit to a large set of possible

interaction terms.

• Perhaps the most common approach is to use *tree models*, as

well as their descendants, *random forest* and *gradient boosted*

*trees*. This class of models automatically searches for optimal

interaction terms; see “Tree Models” on page 249. **G**

Why Would a Data Scientist Care About Heteroskedasticity?

Heteroskedasticity indicates that prediction errors differ for different

ranges of the predicted value, and may suggest an incomplete

model. For example, the heteroskedasticity in lm\_98105 may indicate

that the regression has left something unaccounted for in highand

low-range homes. **G**

Scatterplot Smoothers

Regression is about modeling the relationship between the

response and predictor variables. In evaluating a regression model,

it is useful to use a *scatterplot smoother* to visually highlight relationships

between two variables.

For example, in Figure 4-7, a smooth of the relationship between

the absolute residuals and the predicted value shows that the variance

of the residuals depends on the value of the residual. In this

case, the loess function was used; loess works by repeatedly fitting

a series of local regressions to contiguous subsets to come up

with a smooth. While loess is probably the most commonly used

smoother, other scatterplot smoothers are available in *R*, such as

super smooth (supsmu) and kernel smoothing (ksmooth). In

*Python*, we can find additional smoothers in scipy (wiener or sav)

and statsmodels (kernel\_regression). For the purposes of evaluating

a regression model, there is typically no need to worry about

the details of these scatterplot smooths. **B**

Nonlinear Regression

When statisticians talk about *nonlinear regression*, they are referring

to models that can’t be fit using least squares. What kind of

models are nonlinear? Essentially all models where the response

cannot be expressed as a linear combination of the predictors or

some transform of the predictors. Nonlinear regression models are

harder and computationally more intensive to fit, since they

require numerical optimization. For this reason, it is generally preferred

to use a linear model if possible.  **G**