ST430: Introduction to Regression Analysis, Chapter 9, Sections 1-4

Luo Xiao

November 16, 2015
Special Topics
Some complex model-building problems can be handled using the linear regression approach covered up to this point.

- For example, piecewise regression, including piecewise linear regression and spline regression.

Some require more general nonlinear approaches.

- For example, logistic and probit regression for binary responses.
Sometimes, theory may suggest that a relationship is a straight line, but with different slopes in different parts of the data.

For example,

\[
E(Y) = \begin{cases} 
\beta_0 + \beta_1 x & x \leq \xi, \\
\beta_0^* + \beta_1^* x & x > \xi.
\end{cases}
\]

Usually we want the model to be continuous at \( x = \xi \):

\[
\beta_0 + \beta_1 \xi = \beta_0^* + \beta_1^* \xi
\]

Instead of imposing this as a constraint, the model is usually reparameterized as

\[
E(Y) = \beta_0 + \beta_1 x + \beta_2 \max (x - \xi, 0).
\]
Piecewise regression

Example 9.1

- $Y$: compressive strength of concrete;
- $X$: ratio of water to cement (by mass);
- Note: the compressive strength decreases as the ratio increases.

R code (output in “output1.txt“):

```r
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("CEMENT.RData")
with(CEMENT, plot(STRENGTH~RATIO, pch=20))

lq = lm(STRENGTH ~ RATIO + I(RATIO^2), CEMENT)
summary(lq)
curve(predict(lq, data.frame(RATIO = x)), add = TRUE)
```
R graph

![Graph showing the relationship between STRENGTH and RATIO with a curved line and scattered points.](image)
In the example, suppose the theory suggests that $\xi = 70$ is the critical ratio:

R code (output in “output2.txt“):

```r
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("CEMENT.RData")
with(CEMENT, plot(STRENGTH~RATIO, pch=20))

xi = 70
lp70 = lm(STRENGTH ~ RATIO + pmax(RATIO - xi, 0), CEMENT)
summary(lp70)
curve(predict(lp70, data.frame(RATIO = x)), add = TRUE, col = "red")
```

Note the slightly higher $R^2$ and $R^2_a$. 
Special Topics
Sometimes the lines are *not* constrained to be continuous.

Example 9.2:

- $Y$: reading test score;
- $X$: age of children/young adults.

R code:

```r
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("READSCORES.RData")

with(READSCORES, plot(SCORE~AGE, pch=20))
```
Scatter plot
Models

We may fit a quadratic regression.

To fit separate lines for $AGE \leq 14$ and $AGE > 14$, include the interaction of $AGE$ and the indicator variable for $AGE > 14$.

The piecewise regression model has higher $R^2_a$ than the quadratic model.
R code:

```r
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("READSCORES.RData")
with(READSCORES, plot(SCORE~AGE,pch=20))

l14 = lm(SCORE ~ AGE * (AGE > 14), READSCORES)
lq = lm(SCORE~AGE + I(AGE^2), READSCORES)

curve(predict(l14, data.frame(AGE = x)), to = 14,
     add = TRUE, col = "blue")
curve(predict(l14, data.frame(AGE = x)), from = 15,
     add = TRUE, col = "blue")
curve(predict(lq, data.frame(AGE = x)), add=TRUE, col="red")
```
R fits

- piecewise linear
- quadratic
Inverse prediction

One use of a regression model

\[ E(Y) = \beta_0 + \beta_1 X \]

is to predict \( Y \) for a new \( X, x_0 \).

Sometimes, instead, we observe a new \( y_0 \), and want to make an inference about the new \( x_0 \).

Often \( X \) is expensive to measure, but \( Y \) is cheap; the relationship is determined from a \textit{calibration} dataset.
Because

\[ y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0, \]

we can solve for \( x_0 \):

\[ x_0 = \frac{y_0 - \beta_0 - \epsilon_0}{\beta_1}. \]

We do not observe \( \epsilon_0 \), but we know that \( E(\epsilon_0) = 0 \).

Similarly, we do not know \( \beta_0 \) and \( \beta_1 \), but we have estimates \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \).
This suggests the estimate

\[ \hat{x}_0 = \frac{y_0 - \hat{\beta}_0}{\hat{\beta}_1}. \]

This is known as inverse prediction.

An approximate $100(1 - \alpha)\%$ prediction interval for $x_0$ is:

\[ \hat{x}_0 \pm t_{\alpha/2} \times \frac{s}{\hat{\beta}_1} \times \sqrt{1 + \frac{1}{n} + \frac{(\hat{x}_0 - \bar{x})^2}{SS_{xx}}} . \]
An alternative approach is to fit the *inverse regression*:

\[
X = \gamma_0 + \gamma_1 Y + \epsilon.
\]

Then use the standard prediction interval

\[
\hat{x}_0 \pm t_{\alpha/2} \times s_{x|y} \times \sqrt{1 + \frac{1}{n} + \frac{(y_0 - \bar{y})^2}{SS_{yy}}}
\]

where

\[
\hat{x}_0 = \hat{\gamma}_0 + \hat{\gamma}_1 y_0.
\]

This is *not* supported by the standard theory, because, in the calibration data, \(X\) is fixed and \(Y\) is random.

But it has been shown to work well in practice.
Weighted least squares

Recall the linear regression equation

\[ E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \]

We have estimated the parameters \( \beta_0, \beta_1, \beta_2, \ldots, \beta_k \) by minimizing the sum of squared residuals

\[
SSE = \sum_{i=1}^{n} \left( Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} \left[ Y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \hat{\beta}_2 X_{i,2} + \cdots + \hat{\beta}_k X_{i,k} \right) \right]^2.
\]
Sometimes we want to give some observations more weight than others.

We achieve this by minimizing a weighted sum of squares:

$$WSSE = \sum_{i=1}^{n} w_i (Y_i - \hat{Y}_i)^2$$

$$= \sum_{i=1}^{n} w_i \left[ Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \hat{\beta}_2 X_{i,2} + \cdots + \hat{\beta}_k X_{i,k}) \right]^2$$

The resulting $\hat{\beta}$s are called weighted least squares (WLS) estimates, and the WLS residuals are

$$\sqrt{w_i}(Y_i - \hat{Y}_i).$$
Why use weights?

Suppose that the variance is not constant:

$$\text{var}(Y_i) = \sigma_i^2.$$ 

If we use weights

$$w_i \propto \frac{1}{\sigma_i^2},$$

the WLS estimates have smaller standard errors than the ordinary least squares (OLS) estimates.

That is, the OLS estimates are *inefficient*, relative to the WLS estimates.
In fact, using weights proportional to $1/\sigma_i^2$ is optimal: no other weights give smaller standard errors.

When you specify weights, regression software calculates standard errors on the assumption that they are proportional to $1/\sigma_i^2$. 
How to choose the weights

- If you have many replicates for each unique combination of $\mathbf{X}$s, use $s_i^2$ to estimate $\text{var}(Y|\mathbf{X}_i)$.
- Often you will not have enough replicates to give good variance estimates.

The text suggests grouping observations that are “nearest neighbors”.

Alternatively you can use the regression diagnostic plots.
Example 9.4: Florida road contracts

- \( Y \): winning (lowest) bid price
- \( X \): length of new road

R code (run the code!):

```r
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("DOT11.RData")
fit = lm(BIDPRICE~LENGTH,DOT11)
plot(fit)
```
The first plot uses \textit{unweighted} residuals $y_i - \hat{y}_i$, but the others use weighted residuals.

Also recall that they are “Standardized residuals”

$$z_i^* = \frac{y_i - \hat{y}_i}{s \sqrt{1 - h_i}}.$$ 

which are called Studentized residuals in the text.

With weights, the standardized residuals are

$$z_i^* = \sqrt{w_i} \left( \frac{y_i - \hat{y}_i}{s \sqrt{1 - h_i}} \right).$$
Note that the “Scale-Location” plot shows an increasing trend.

Try weights that are proportional to powers of $x = \text{LENGTH}$:

```r
fit1 = lm(BIDPRICE~LENGTH, DOT11, weights = 1/LENGTH)
plot(fit1)

fit2 = lm(BIDPRICE~LENGTH, DOT11, weights = 1/LENGTH^2)
plot(fit2)
```

`summary()` shows that the fitted equations are all very similar, and `weights = 1/LENGTH` gives the smallest standard errors.
Note

Standard errors are computed as if the weights are known constants.

But for the “nearest neighbors" method in textbook, weights are based on a preliminary OLS fit.

Theory shows that *in large samples* the standard errors are also valid with estimated weights.
Note

When you specify weights $w_i$, \texttt{lm()} fits the model

$$\sigma_i^2 = \frac{\sigma^2}{w_i}$$

and the “Residual standard error” $s$ is an estimate of $\sigma$:

$$s^2 = \frac{\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2}{n - k - 1}$$

If you change the weights, the meaning of $\sigma$ (and $s$) changes.

You 	extit{cannot} compare the residual standard errors for different weighting schemes (c.f. page 488, foot).