Chapter 4

DJM

14 February 2017

Workflow for doing statistics

1. Choose a family of models.
2. For each model:
    1. Calculate CV to get estimates of the risk.
    2. Choose the tuning parameter that gets the lowest estimate of the risk.
3. Choose a model by picking the **model** with the lowest estimate of the risk.
4. Evaluate and describe your model. Make plots, interpret coefficients, make predictions, etc.
5. If you see things if 4 you don't like, propose a new model(s) to handle these issues and return to step 2.

Linear smoothers

- Recall S431:

  The "Hat Matrix" puts the hat on Y: $\hat{Y} = HY$.

- If I want to get fitted values from the linear model

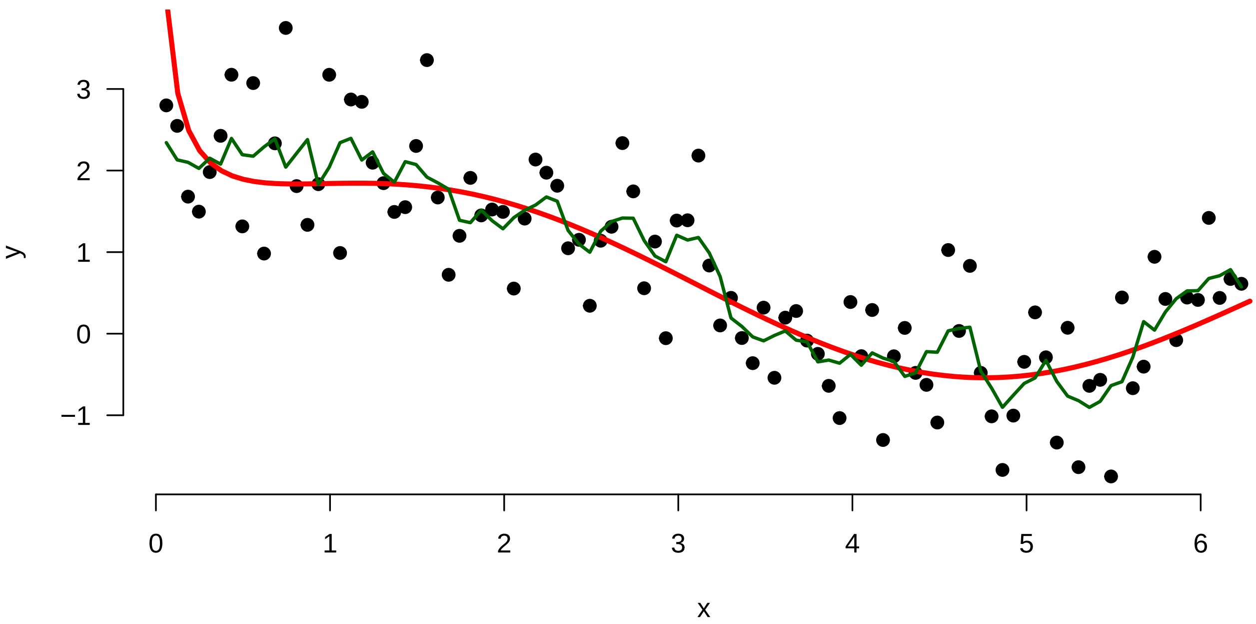
$$\hat{Y} = X\hat{\beta} = X\left[(X^\top X)^{-1}X^\top Y\right] = HY$$

- We generalize this to arbitrary matrices:

  A linear smoother is any predictor f that gives fitted values via $f(X) = WY$.

- Today, we will other ways of predicting Y from X.
- If I can get the fitted values at my original datapoints X by multiplying Y by a matrix, then that is a linear smoother.

Example



At each x, find 2 points on the left, and 2 on the right. Average their y values with that of your current point.

```
W = toeplitz(c(rep(1,3),rep(0,97)))
W = sweep(W, 1, rowSums(W), '/')
Yhat = W %*% y
lines(x, Yhat, col='darkgreen', lwd=2)
```

This is a linear smoother. What is W?

What is W?

- I actually built this one directly into the code.
- An example with a 10 x 10 matrix:

```
W = toeplitz(c(rep(1,3),rep(0,7)))
round(sweep(W, 1, rowSums(W), '/'), 2)
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,] 0.33 0.33 0.33 0.00  0.0  0.0 0.00 0.00 0.00  0.00
##  [2,] 0.25 0.25 0.25 0.25  0.0  0.0 0.00 0.00 0.00  0.00
##  [3,] 0.20 0.20 0.20 0.20  0.2  0.0 0.00 0.00 0.00  0.00
##  [4,] 0.00 0.20 0.20 0.20  0.2  0.2 0.00 0.00 0.00  0.00
##  [5,] 0.00 0.00 0.20 0.20  0.2  0.2 0.20 0.00 0.00  0.00
##  [6,] 0.00 0.00 0.00 0.20  0.2  0.2 0.20 0.20 0.00  0.00
##  [7,] 0.00 0.00 0.00 0.00  0.2  0.2 0.20 0.20 0.20  0.00
##  [8,] 0.00 0.00 0.00 0.00  0.0  0.2 0.20 0.20 0.20  0.20
##  [9,] 0.00 0.00 0.00 0.00  0.0  0.0 0.25 0.25 0.25  0.25
## [10,] 0.00 0.00 0.00 0.00  0.0  0.0 0.00 0.33 0.33  0.33
```

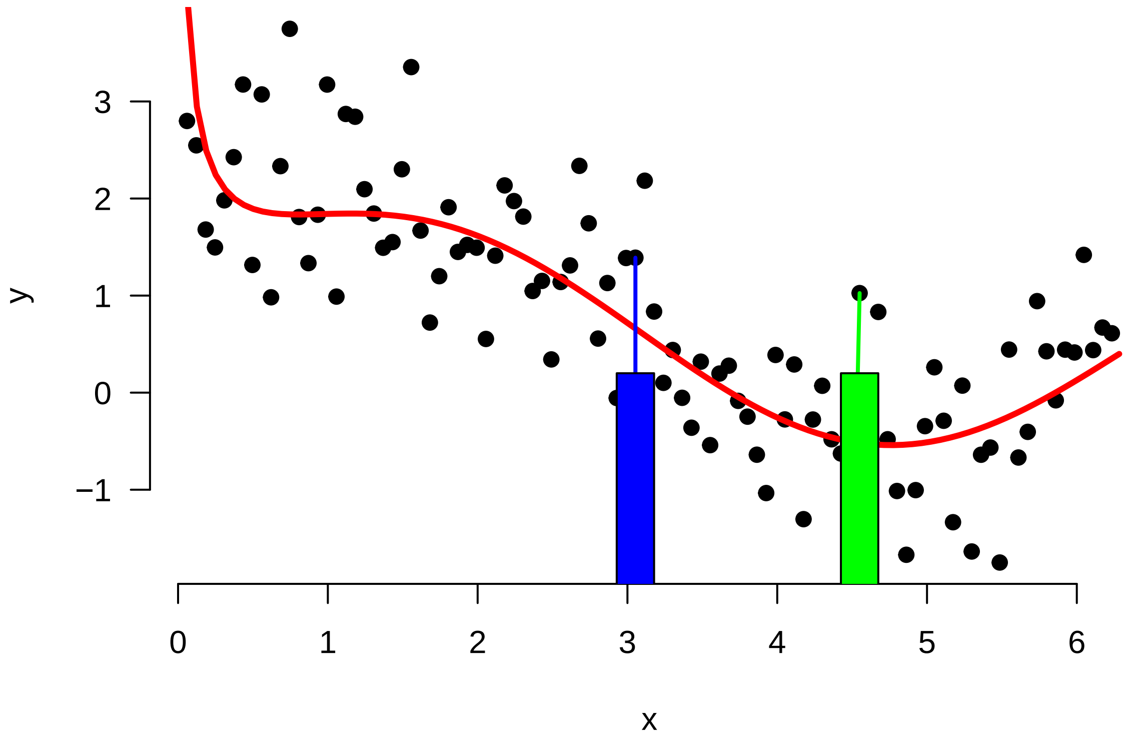
- This is a "kernel" smoother.

What is a "kernel" smoother?

- The mathematics:

A kernel is any function K such that for any u, $K(u) \geq 0$, $\int duK(u) = 1$ and $\int uK(u)du = 0$.

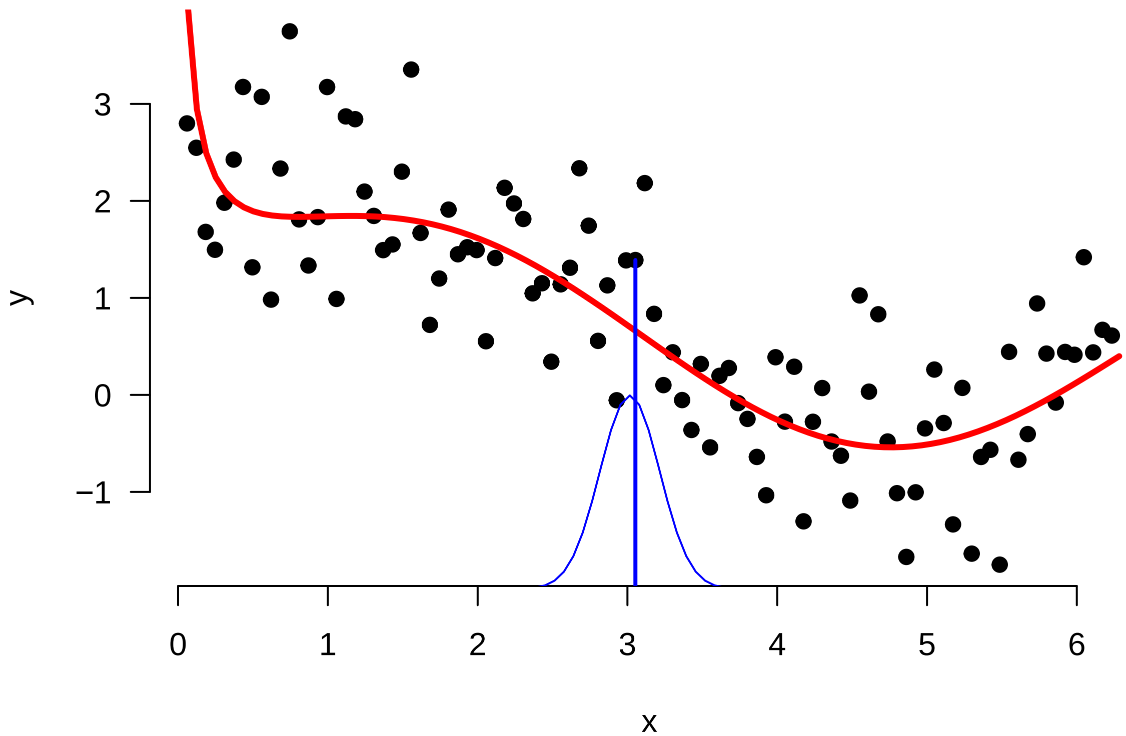
- The idea: a kernel is a nice way to take weighted averages. The kernel function gives the weights.
- The previous example is called the **boxcar** kernel. It looks like this:



- Notice that the kernel gets centered at each x. The weights of the average are determined by the shape of the kernel.
- For the boxcar, all the points inside the box get the same weight, all the rest get 0.

Other kernels

- Most of the time, we don't use the boxcar because the weights are weird.
- A more common one is the Gaussian kernel:



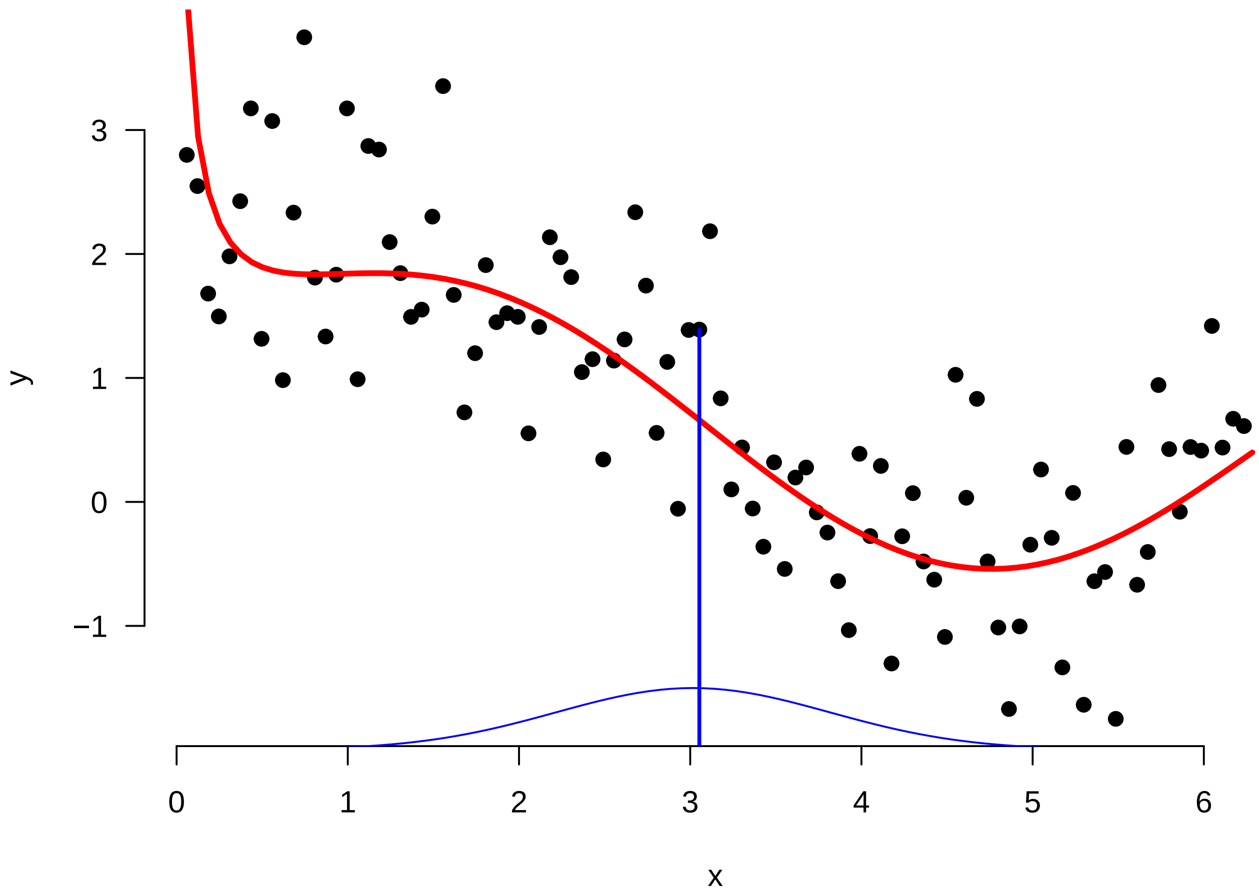
- Let's look at row 49 of the W matrix here:

$$W_{49,j} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_j - x_{49})\right)$$

- For the plot, I made $\sigma = .2$.

Other kernels

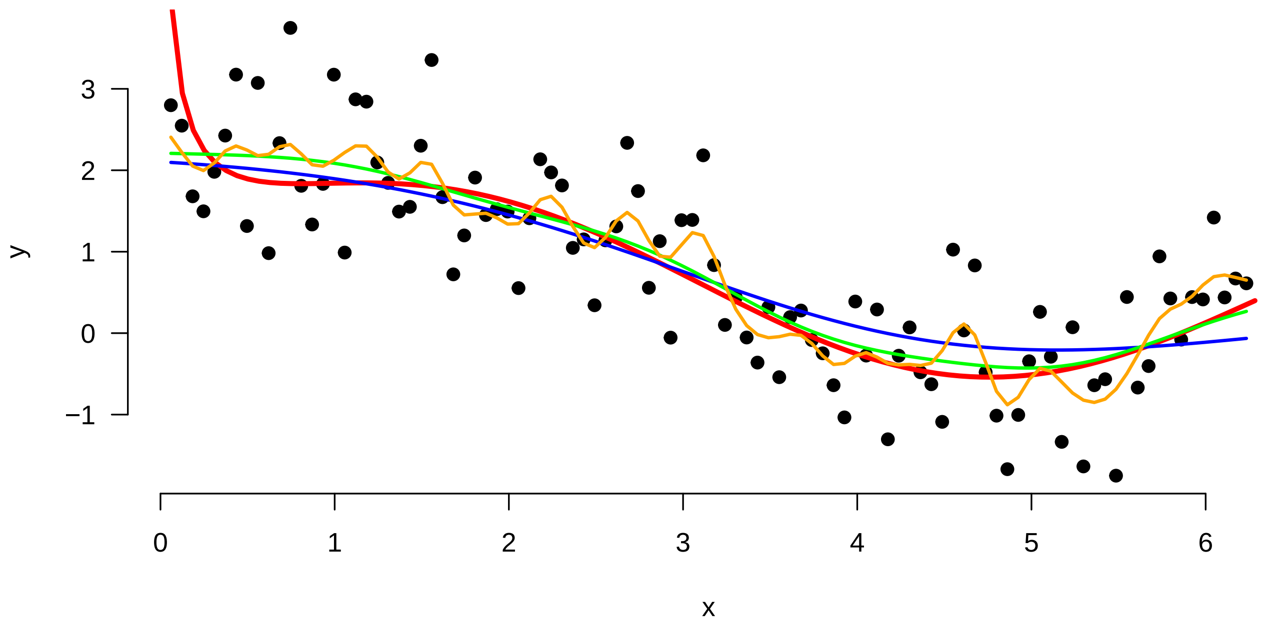
- What if I made $\sigma = 0.8$?



- Before, points far from x_{49} got very small weights for predicting at x_{49}, now they have more influence.
- For the Gaussian kernel, σ determines something like the "range" of the smoother.

Many Gaussians

- Using my formula for W, I can calculate different linear smoothers with different σ

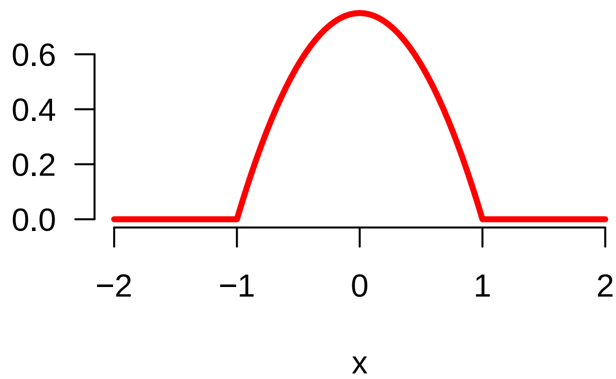



```
dmat = as.matrix(dist(x))
Wgauss <- function(sig){
  gg = exp(-dmat^2/(2*sig^2)) / (sig * sqrt(2*pi))
  sweep(gg, 1, rowSums(gg),'/')
}
W1 = Wgauss(1)
W.5 = Wgauss(.5)
W.1 = Wgauss(.1)
lines(x, W1%*%y, col='blue',lwd=3, lty=1)
lines(x, W.5%*%y, col='green',lwd=3, lty=2)
lines(x, W.1%*%y, col='orange',lwd=3, lty=3)
```

The bandwidth

- Choosing σ is **very** important.
- This "range" parameter is called the **bandwidth**.
- Most practitioners will tell you that it is way more important than which kernel you use.
- The default kernel is something called 'Epanechnikov':

```
epan <- function(x) 3/4*(1-x^2)*(abs(x)<1)
curve(epan(x),-2,2,col=2, lwd=3, mar=c(3,2,0,0), bty='n', las=1, cex.lab=1, cex.axis=1, ylab='')
```



How do you choose the bandwidth?

- Cross validation of course!
- Now the trick:

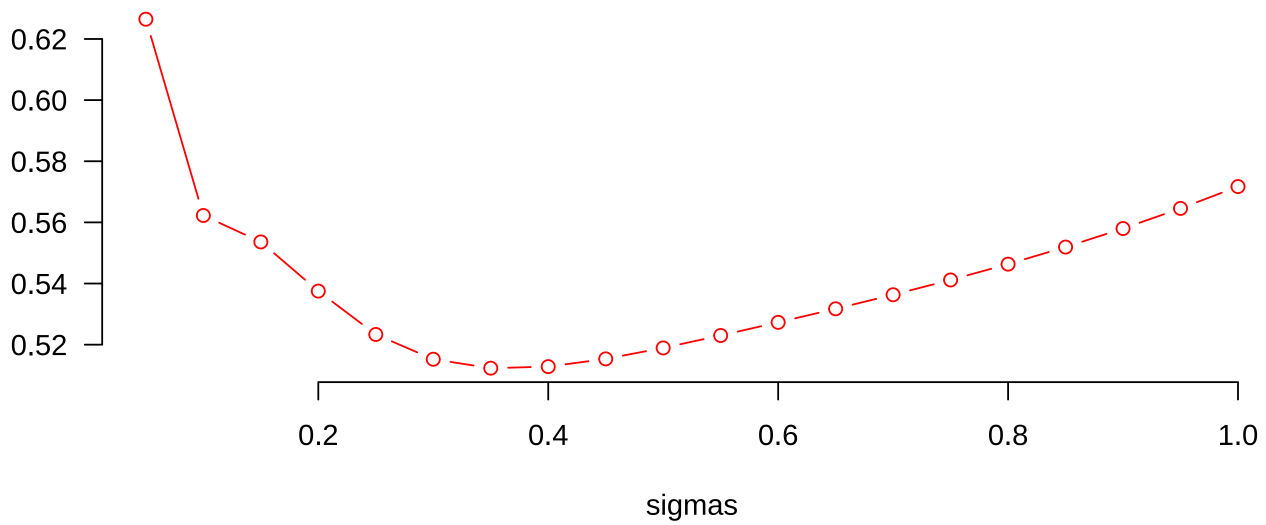
  For linear smoothers, one can show (after pages of tedious algebra which I wouldn't wish on my worst enemy, but might, in a fit of rage assign to a belligerent graduate student) that for $\hat{Y} = WY$,

$$\text{LOO-CV} = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{(1-w_{ii})^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{e}_i^2}{(1-w_{ii})^2}.$$

- This trick means that you only have to fit the model once rather than n times!
- You still have to calculate this for each model!

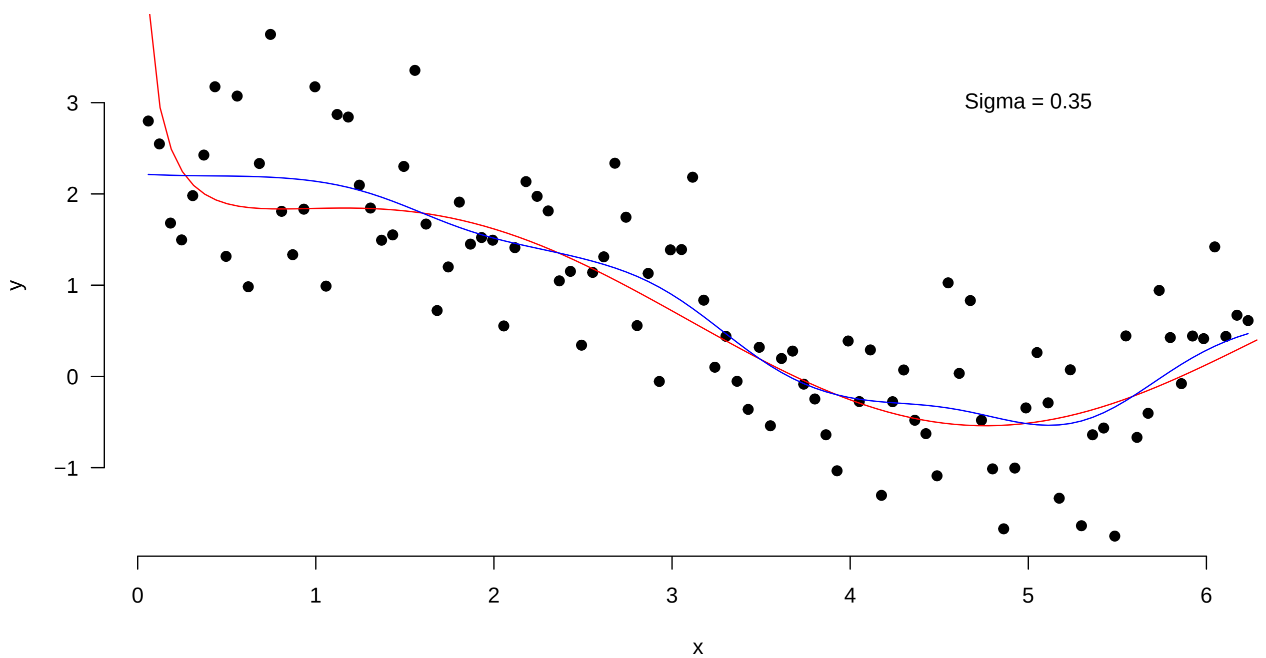
Back to my Gaussian example

```
looCV <- function(y, W){
  n = length(y)
  resids2 = ((diag(n)-W) %*% y)^2
  denom = (1-diag(W))^2
  return(mean(resids2/denom))
}
looCVs = double(20)
sigmas = seq(.05, 1, length.out=length(looCVs))
for(i in 1:length(looCVs)){
  W = Wgauss(sigmas[i])
  looCVs[i] = looCV(y, W)
}
plot(sigmas, looCVs, type='b', col=2, las=1, mar=c(4,5,0,0), bty='n', ylab='', cex.axis=1, cex.lab=1)
```



Back to my Gaussian example

```
par(mar=c(4,4,0,0))
plot(x, y, pch=19, bty='n', las=1, cex.lab=1, cex.axis=1)
curve(trueFunction(x), 0, 2*pi, col=2, lwd=1, add=TRUE)
Wstar = Wgauss(sigmas[which.min(looCVs)])
lines(x, Wstar %*% y, col='blue', lwd=1, lty=1)
text(5, 3,paste('Sigma =', round(sigmas[which.min(looCVs)],2)), cex=1)
```



Some of those ugly formulas

- These are things like (4.10)-(4.12) and (4.14)

- The purpose of these formulas is to illustrate **VERY GENERALLY** how to trade bias and variance with Kernel smoothers.
- The highest level overview is equation (4.16):

$$MSE - \sigma^2(x) = O(h^4) + O(1/nh).$$

- The first term on the left is the **squared bias** while the second term on the right is the **variance**.
- Note: we have moved **irreducible noise** to the left of =.
- The "big-Oh" notation means we have removed a bunch of constants that don't depend on n or h.

[They DO depend on the properties of the Kernel, and the distribution which generated the data.]

- The **Optimal Bandwidth** minimizes the MSE:

$$\begin{aligned} h_{opt} &= \arg\min_h C_1 h^4 + \frac{C_2}{nh} \\ \Rightarrow 0 &\overset{set}{=} 4C_1 h^3 - \frac{C_2}{nh^2} \\ \Rightarrow h^5 &= O\left(\frac{1}{n}\right) \\ \Rightarrow h_{opt} &= O\left(\frac{1}{n^5}\right). \end{aligned}$$

Ok, you asked for the algebra.

- You don't want the algebra.
- Like the formula for LOO-CV, if I were a horrible, soul destroying person, I would wade through it for the next two hours (to get (4.10)).
- Believe me, I've done it. Not fun. The hand wavy, "big-Oh" stuff is what you should keep in mind.
- If you really want it, I will write up a document with all the work.

Kernels and interactions

- In multivariate kernel regressions, you estimate a **surface** over the input variables.
- This is trying essentially to find $\hat{f}(x_1, \ldots, x_p)$.
- Therefore, this function **by construction** includes interactions, handles categorical data, etc. etc.
- This is contrast with **linear models** which need you to specify these things.
- This extra complexity (automatically including interactions, as well as other things) comes with tradeoffs.

Issue 1

- More complicated functions (smooth Kernel regressions vs. linear models) tend to have **lower bias** but **higher variance**.
- For $p = 1$, equations (4.19) and (4.20) show this:
- **Bias**
    1. The bias of using a linear model when it is wrong is a number $b(x, \theta_0)$ which doesn't depend on n.

2. The bias of using kernel regression is $O(1/n^{4/5})$. This goes to 0 as $n \to \infty$.

- **Variance**
  1. The variance of using a linear model is $O(1/n)$
  2. The variance of using kernel regression is $O(1/n^{4/5})$.
- To conclude: bias of kernels goes to zero (not for lines) but variance of lines goes to zero faster than for kernels.
- If the linear model is right, you win. But if it's wrong, you (eventually) lose.
- How do you know if you have enough data? Do model selection (CV to choose models).
- Compare of the kernel version with CV-selected tuning parameter (the CV estimate of the risk), with the CV estimate of the risk for the linear model.

Issue 2

- For $p > 1$, there is more trouble.
- First, lets look again at
$$MSE(h) - \sigma^2(x) = O(1/n^{4/5}).$$
  That is for $p = 1$. It's not **that much** slower than $O(1/n)$, the variance for linear models.
- If $p > 1$ similar calculations show,
$$MSE(h) - \sigma^2(x) = O(1/n^{4/(4+p)}) \qquad MSE(\theta_0) - \sigma^2(x) = b(x, \theta_0) + O(p/n).$$
- What if p is big?
  1. Then $O(1/n^{4/(4+p)})$ is still big.
  2. But $O(p/n)$ is small.
  3. So unless $b(x, \theta_0)$ is big, we should use the linear model.
- How do you tell? Use CV to decide.

Issue 3

- When p is big, `npreg` is slow.
- Not much to do about that.
- Chapter 8 has some compromises that people use.
- A **very, very** questionable rule of thumb: if $p > \log(n)$, this may not work.

Some `npreg` discussion

- `npreg` is using CV and optimization to try to choose the bandwidth for you.
- The `tol` and `ftol` arguments control how close the solution needs to be to an optimum.
- Very basic minimization (called Gradient descent):
  - Suppose I want to minimize $f(x) = (x - 6)^2$ numerically.
  - If I start at a point (say $x_1 = 23$), vaguely, I want to "go" in the negative direction of the gradient.
  - The gradient (at $x_1 = 23$) is $f'(23) = 2(23 - 6) = 34$.
  - Gradient descent says, ok go that way by some small amount: $x_2 = x_1 - \gamma 34$, for γ. small.

  - In general, $x_{n+1} = x_n - \gamma f'(x_n)$.

```
niter = 10
gam = 0.1
x = double(niter)
x[1] = 23
grad <- function(x) 2*(x-6)
for(i in 2:niter) x[i] = x[i-1] - gam*grad(x[i-1])
x
```

```
##  [1] 23.000000 19.600000 16.880000 14.704000 12.963200 11.570560 10.456448
##  [8]  9.565158  8.852127  8.281701
```

- How do I decide if I'm done? The easiest way is to check how much I'm moving.

Fixing my gradient descent code

```
maxiter = 1000
conv = FALSE
gam = 0.1
x = 23
tol = 1e-3
grad <- function(x) 2*(x-6)
for(iter in  1:maxiter){
  x.new = x - gam * grad(x)
  conv = (x - x.new < tol)
  x = x.new
  if(conv) break
}
x
```

```
## [1] 6.003531
```

```
iter
```

```
## [1] 38
```

- What happens if I change `tol` to `1e-7`?