This problem set was based on the preliminary analysis in the paper


# Setup

```r
install.packages("np")
require(np)
data(oecdpanel)
```

1. **Answer:**

```r
# Fit a linear model of growth on initgdp
lm.1 = lm(growth ~ initgdp, data = oecdpanel)
> signif(lm.1$coefficients,3)

     (Intercept)       initgdp
  -0.02130000  0.00533333

The coefficient of `initgdp` in the linear model is about $5 \times 10^{-3}$, suggesting that higher growth rates are associated with higher initial levels of GDP, contrary to the claim we want to test.

2. **Answer:** See Figure 1

The black line suggests that choosing a bandwidth near 0.3 minimizes the MSE for predictions on new observations. The green line shows the MSE for the model fit after using the whole data as the training data. The important thing to notice from the green line is that the model fit always improves as the bandwidth decreases because the test set = the training set = the whole data ↔ no cross-validation. For finding the right bandwidth to optimize the model for prediction on new observations, it is important that the training set is separate from the test set.
Figure 1: The black solid line is the MSE for predictions on new observations in 5-fold cv. The green, dashed line is in-sample MSE when all of the data is used (so the training set = testing set = the whole data).
# Code for problem 2: start by running the supplied 5-fold cv code  
# Now modify the inner loop of supplied code for the "whole data" case (non-CV)  
whole.MSE = rep(NA, length(bandwidths))  
for (bw in bandwidths) {  
  # Fit to the training set = whole data  
  current.npr = npreg(growth ~ initgdp, data=oecdpanel,bws=bw)  
  # Predict on the test set = whole data, so predictions are just fitted values  
  predictions = fitted(current.npr)  
  # What’s the mean-squared error?  
  whole.MSE[which(bandwidths == bw)] = mean((oecdpanel$growth - predictions)^2)  
  # Note the alternative to the paste() trick for accessing bandwidths  
  # Also: one part of the objects returned by npreg is an MSE attribute, so  
  # we could have done  
  # whole.MSE[which(bandwidths == bw)] = current.npr$MSE  
  # but the more explicit approach will also work if we decide to start using  
  # a different smoothing package  
}  

# display the results (bandwidths.cv.mses and whole.MSE)  
pdf("graphics/p2.pdf")  
plot(bandwidths, bandwidths.cv.mses, type = "l", lwd = 2,  
ylim = c(min(c(bandwidths.cv.mses, whole.MSE)),  
max(c(bandwidths.cv.mses, whole.MSE))),  
main = "MSE vs. bandwidth  
\n5-fold CV (black, solid) and in-sample (green, dashed)",  
ylab = "MSE", xlab = "Kernel bandwidth")  
lines(bandwidths, whole.MSE, lwd = 2, col = "green", lty=2)  
dev.off()  

3. Answer: See Figure 2

The kernel regression curve does not give a strong indication of whether poorer countries grow faster than richer ones. Through most of the mid-section, rising initial wealth seems to correlate with increasing growth rates. However, the left tail of the curve suggests that the poorest countries grow just a bit faster than the “almost poorest” countries, and an analogous thing appears to happen in the right tail.

4. Answer:

# Fit a linear model of growth on initgdp, popgro, and inv  
lm.4 = lm(growth ~ initgdp+popgro+inv, data = oecdpanel)  
> signif(lm.4$coefficients,3)  
(Intercept)  initgdp  popgro  inv  
 0.08640  -0.00465  -0.00545   0.02290  

The new regression coefficients (above) show that including popgro and inv as predictors makes a big difference to how growth depends on initgdp.
Figure 2: The red line is the line for the linear model. The green points are the kernel regression fitted values with bandwidth = 0.3.
In particular, the sign of the coefficient for \texttt{initgdp} is opposite what we obtained in problem 1. This result suggests that, “holding constant \texttt{popgro} and \texttt{inv}”, high growth rates are associated with low initial levels of GDP, consistent with the claim we want to test.

5. **Answer:** Run the provided code:

```r
oecd.npr <- npreg(growth ~ initgdp + popgro + inv, data=oecdpanel, tol=0.1, ftol=0.1)
summary(oecd.npr)
```

#output contains this:

<table>
<thead>
<tr>
<th></th>
<th>initgdp</th>
<th>popgro</th>
<th>inv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth(s):</td>
<td>0.211</td>
<td>0.0591</td>
<td>0.117</td>
</tr>
</tbody>
</table>

The selected bandwidth for \texttt{initgdp} is quite a bit smaller than 0.3.

6. **Answer:**

```r
# Find and report the medians:
med.popgro = median(oecdpanel$popgro)
med.inv = median(oecdpanel$inv)
> signif(med.popgro,3)
[1] -2.64
> signif(med.inv,3)
[1] -1.78
```

There are several ways to get the predictions. One is to make a new data frame which has the values we want. Here we re-use the real \texttt{initgdp} values (we could also use an equally-spaced grid), but sort them in increasing order, and fix \texttt{popgro} and \texttt{inv} to their medians.

```r
# Make a new data set \texttt{med.oecdpanel} that contains growth, initgdp, popgro, # and inv, but replace popgro and inv and with their respective medians
# Use the actual values of initgdp, but sort them for easier plotting later
initgdp.sort = sort(oecdpanel$initgdp)
med.oecdpanel = data.frame(initgdp=initgdp.sort, popgro=med.popgro, inv=med.inv)
```

```r
# Find the predicted growth rates under the model from problem 4
mod4.med.preds = predict(lm.4, newdata = med.oecdpanel)
# Find the predicted growth rates under the model from problem 5
mod5.med.preds = predict(oecd.npr, newdata = med.oecdpanel)
```

```r
# Plot growth v.s. initgdp for \texttt{med.data} and add the fitted values for both models:
pdf("graphics/p6.pdf")
# Make scatterplot of initgdp versus growth:
plot(oecdpanel$initgdp, oecdpanel$growth, lwd = 2, cex = 0.3,
xlab = "initgdp", ylab = "growth",
```
The linear model (red) suggests that high growth rates follow from low initial levels of GDP. The kernel regression (grey) is more nuanced. Most of the kernel regression curve is downward-sloping, but the subsections near the 30th and 90th percentiles of `initgdp` are steeply upward-sloping. This suggests that either (1) other variables exist which, if included in the model, would help clarify the dependence of `growth` upon `initgdp`, or (2) the dependence of `growth` on `initgdp` is best understood to be non-monotonic, i.e., changes in `initgdp` correspond to (or even cause?) both positive and negative changes in `growth` depending on the starting point.

7. Answer:

```r
nfolds = 5
case.folds = rep(1:nfolds,length.out=nrow(oecdpanel))
case.folds = sample(case.folds)
linear.fold.mses = vector(length=nfolds)
```
kernel.fold.mses = vector(length=nfolds)
for (fold in 1:nfolds) {
  train = oecdpanel[case.folds!=fold,]
  test = oecdpanel[case.folds==fold,]
  # Fit the models
  linear.model = lm(growth ~ initgdp+popgro+inv, data=train)
  kernel.model = npreg(growth ~ initgdp + popgro + inv,
                        data=train, tol=0.1, ftol=0.1)
  # Predict on the test set
  linear.predictions = predict(linear.model, newdata=test)
  kernel.predictions = predict(kernel.model, newdata=test)
  linear.fold.mses[fold] = mean((test$growth - linear.predictions)^2)
  kernel.fold.mses[fold] = mean((test$growth - kernel.predictions)^2)
}

cv.linear.mod = mean(linear.fold.mses)
cv.kernel.mod = mean(kernel.fold.mses)

Based on the cross-validation scores (0.000782 for the linear model and 0.000754 for the kernel regression), one would prefer the kernel regression model.

8. **Answer:** The initial analysis, in problems 1 to 3, undermined the idea of catching up. In the subsequent problems, we began controlling for other variables, and found that the idea of catching up might be correct. Both analyses are legitimate, if interpreted carefully, but the thing that makes the second analysis interesting is that it is more suggestive of causality. If we can establish that having a lower initial GDP in some sense causes higher growth rates, then we might have arrived at a conclusion with a lot of policy relevance. It would suggest, for instance, that we should expect a country’s growth rate to decline as it got richer, and not attribute that to problems with the economy. It would also suggest that growth should be extra high after recessions and wars, which lower GDP, but that those rates couldn’t be sustained (and blowing up part of the country to raise the growth rate would be perverse).

The kernel regression model complicates matters by allowing us to model growth as a possibly non-linear function of initgdp. Hence, the question “Is the catching up theory” correct does not necessarily have a yes/no answer; it appears that the theory may hold true for specific ranges of initgdp but not everywhere.

It is important to keep in mind that while the idea of catching up may be legitimate, it seems unlikely to account for much of the variability in growth rates, as you might guess from the last plot.