Lecture 19

Strategies for Fitting Models

Last time we talked about model diagnostics. This time we show some examples, and also talk about strategies for selecting models, which means choosing which variables belong and which do not.

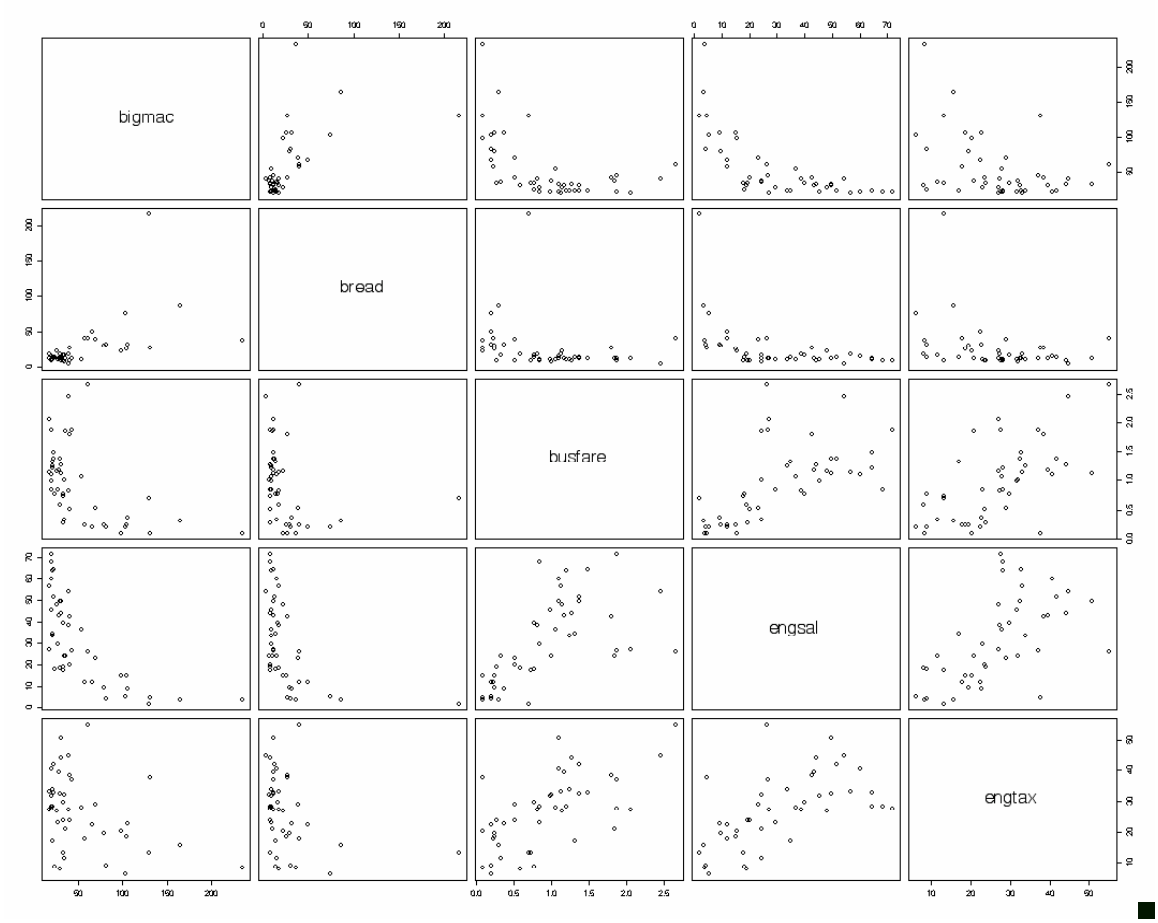
The dataset we use is the "big mac" data, which was collected by The Economist magazine to compare costs of the Big Mac around the world. Why this product? Because it's pretty much the same all over the world, and so differences in price should reflect the inefficiencies of currency exchange. This data shows how the price depends on

- bread -- minutes of labor required to buy one kilogram of bread
- busfare the lowest cost of a 10 K bus, tram o rsubway ticket, in US dollars
- EngSal The average annual salary of an electrical engineer in 1000 US dollars
- Eng TAx The average tax rate paid by engineers
- Service annual cost of 19 services, primarily relevant to Euopre and N. America
- TeachSal The average annual salary of a primary school teacher, in US dollars
- TeachTAx The average tax rate paid by primary teachers
- VacDays Average days of vacation per year
- WorkHrs Average hours worked per year
- BigMac minutes of labor required by an average worker to buy a BigMac and fries.

```
names(bigmac.table)
 [1] "bigmac"   "bread"    "busfare"  "engsal"   "engtax"
"service"
```

Lets focus on a subset, for the sake of discussion.

```
> smallbigmac <- data.frame(bigmac.table[,-
c(6,7,8,9,10,11)])
> detach(bigmac.table)
> attach(smallbigmac)
> names(smallbigmac)
[1] "bigmac"  "bread"   "busfare" "engsal"  "engtax"
> pairs(smallbigmac)
```



The top row tells us the relation between Bigmac and the predictors. What do you see?

Let's fit the basic model so we can see whether or diagnostics tell us what we thought they would.

```
> naive <- lm(bigmac~., data=smallbigmac)
> summary(naive)

Call:
lm(formula = bigmac ~ ., data = smallbigmac)

Residuals:
    Min      1Q  Median      3Q     Max
-36.148 -20.286  -5.303  11.277 143.346

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

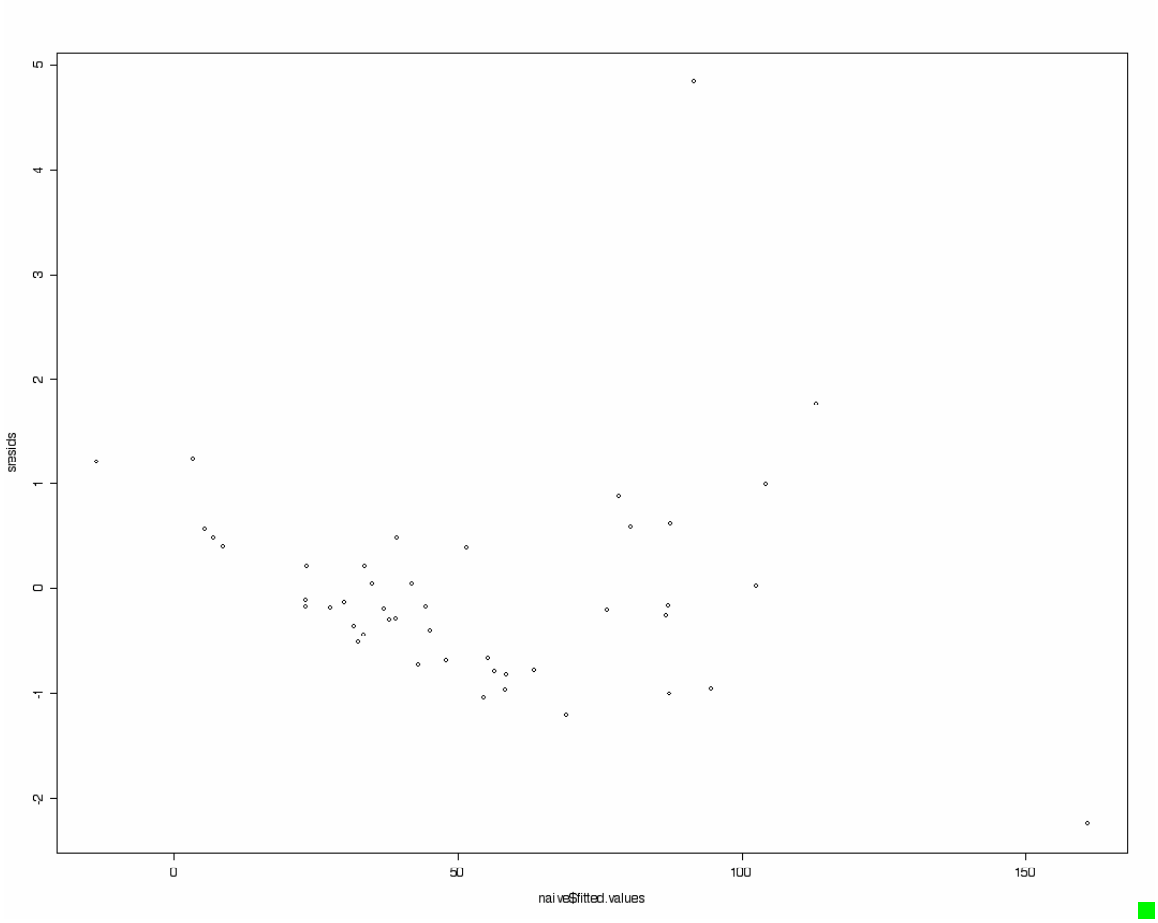
```
(Intercept)  76.7458    14.3856   5.335 4.05e-06 ***
bread         0.4189     0.1587   2.640  0.01177 *
busfare     -17.6372    10.0915  -1.748  0.08819 .
engsal       -1.0849     0.3410  -3.181  0.00283 **
engtax        0.6094     0.5521   1.104  0.27633
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 31.04 on 40 degrees of freedom
Multiple R-Squared: 0.569,    Adjusted R-squared: 0.526
F-statistic:  13.2 on 4 and 40 DF,  p-value: 6.044e-07
```

So how did we do?

We said earlier that it's better to look at the standardized residuals rather than the residuals:

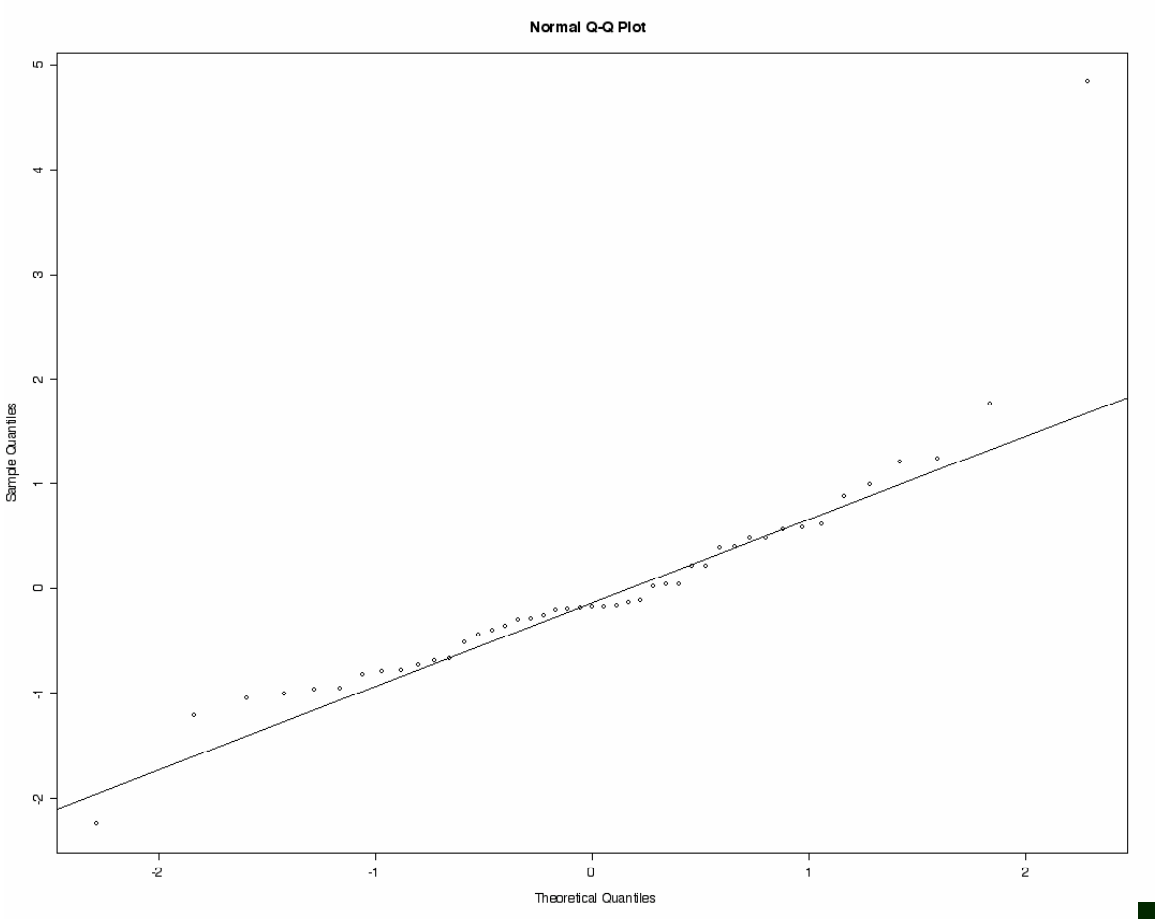
```
> sresids <- rstandard(naive)
> plot(naive$fitted.values, sresids)
```



Pretty bad.

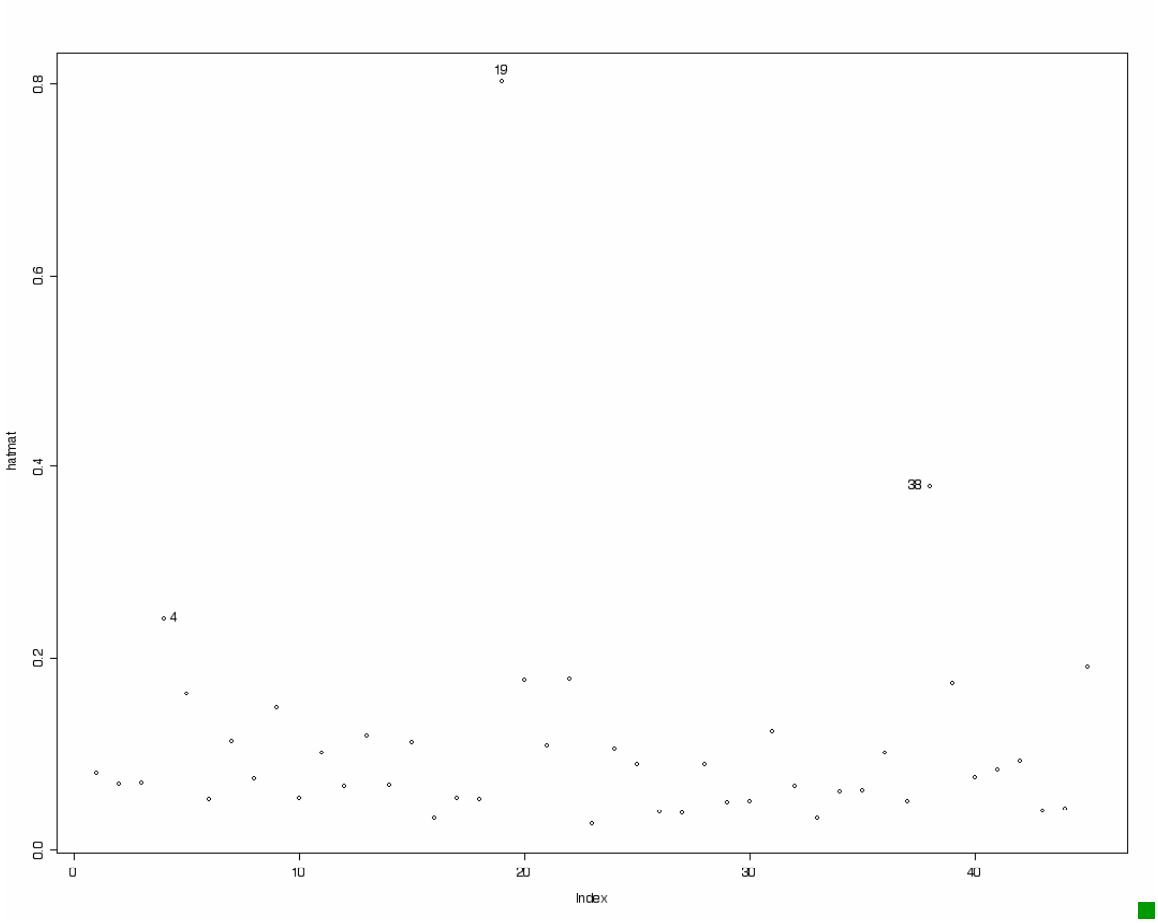
Oddly, normality isn't much of a problem.

```
> qqnorm(sresids)
> qqline(sresids)
```



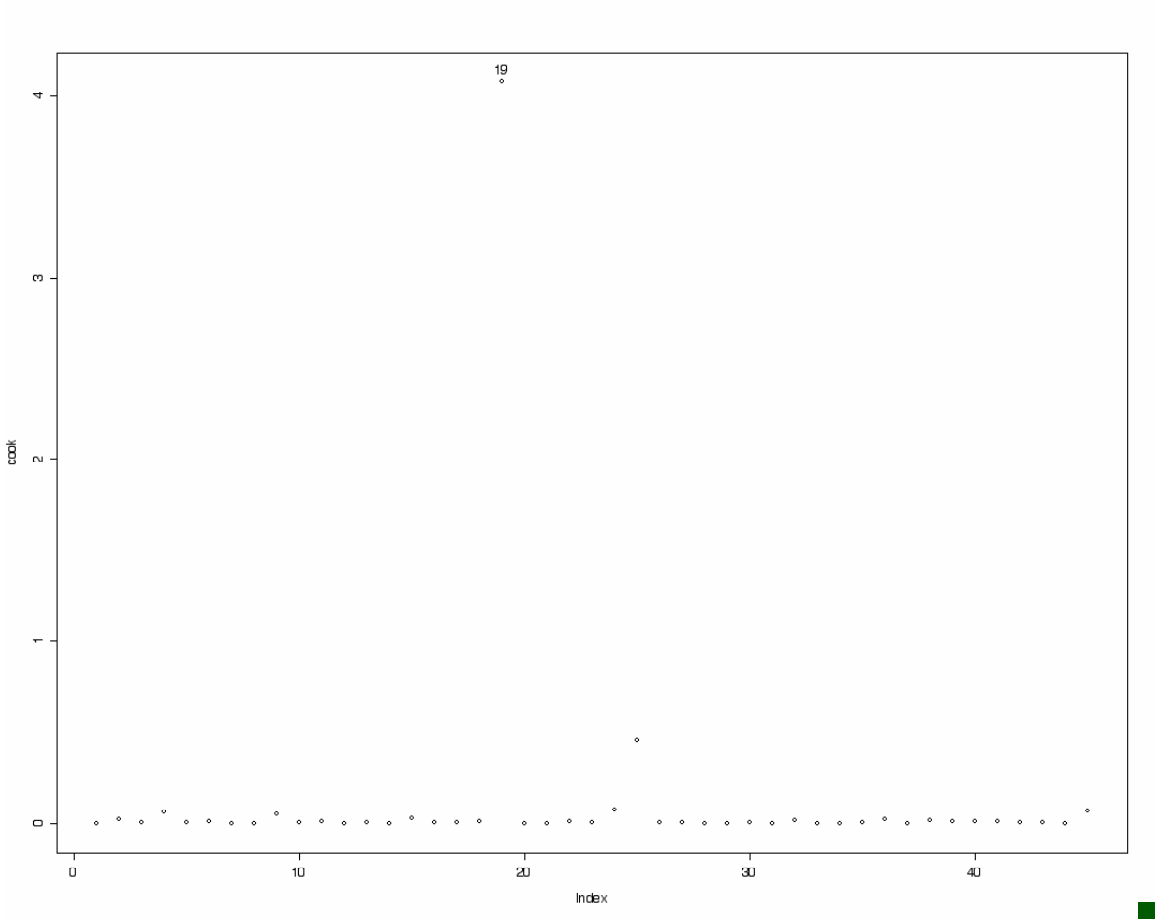
We can see the leverages by looking at the diagonals of the hat matrix:

```
> hatmat <- hatvalues(naive)
> plot(hatmat)
> identify(hatmat)
[1]  4 19 38
> bigmac$city[c(4,19,38)]
[1] Bombay    Lagos     Stockholm
```



And we can look at the Cook's distances:

```
> cook <- cooks.distance(naive)
> plot(cook)
```



Lagos is the only influential point.

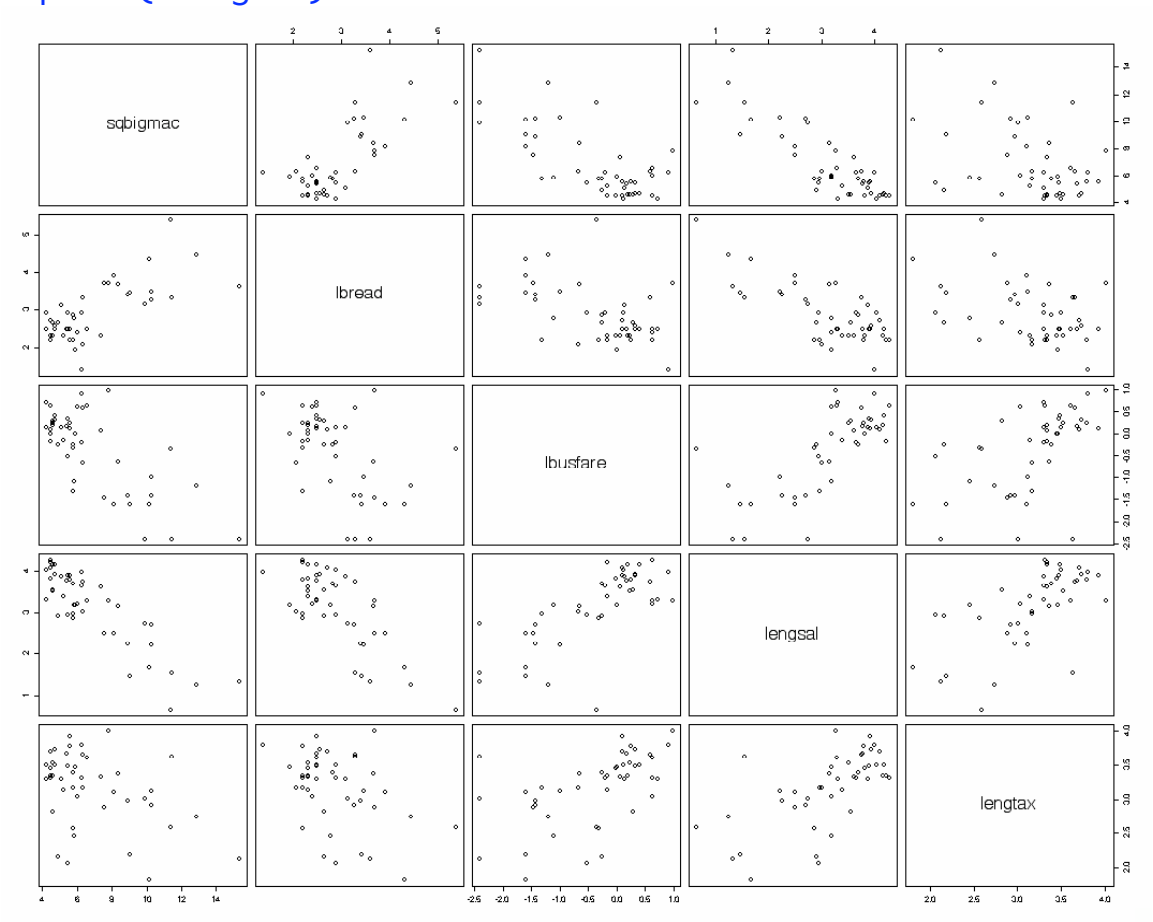
We know from our initial scatterplots and our diagnostics that something is wrong.Our first decision is what to do with Lagos. Lagos is so different from the others, we're going to try to fit a model without it.

```
> smallbigmac <- smallbigmac[-19,]
> city <- bigmac$city[-19]
```

At this point, we need to re-plot the data and try a transformation so that the assumption of linearity is more or less true.
I don't want to go through this hear, so here are the transformations I'm going to use:

```
> lbread <- log(bread)
> lbusfare <- log(busfare)
> lengsal <- log(engsal)
> lengtax <- log(engtax)
> rm(bigmac)
sqbigmac <- sqrt(bigmac)
```

```
> t2bigmac <-
data.frame(sqbigmac,lbread,lbusfare,lengsal,lengtax)
> pairs(t2bigmac)
```



A fit of the model at this point shows pretty good diagnostics. (Try it).

Let's turn our attention now to another phenomenon.

Suppose we did not use the log of engineers salary in our model. We would get a summary that looks like this:

```
> summary(lm(sqbigmac~lbread+busfare+lengtax))

Call:
lm(formula = sqbigmac ~ lbread + busfare + lengtax)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4333 -1.0977 -0.4554  1.0399  6.0273
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.95910    2.43865   0.803   0.4264
lbread       2.05780    0.39738   5.178  6.3e-06 ***
busfare     -1.10735    0.51629  -2.145   0.0379 *
lengtax      0.00582    0.63122   0.009   0.9927
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 1.733 on 41 degrees of freedom
Multiple R-Squared: 0.5832,   Adjusted R-squared: 0.5527
F-statistic: 19.13 on 3 and 41 DF,  p-value: 6.516e-08
```

From this we learn that the tax rate for engineers is not a predictor of the cost of bigmacs, but the cost of bread is. But suppose we now add the log of engineers salary:

```
lm(formula = sqbigmac ~ ., data = t2bigmac)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7939 -0.6597 -0.1644  0.5501  4.1589

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.2109     2.2259   3.240  0.00241 **
lbread        0.5675     0.3595   1.579  0.12232
lbusfare     -0.7915     0.2882  -2.747  0.00898 **
lengsal      -2.0039     0.4060  -4.936 1.45e-05 ***
lengtax       1.2594     0.4462   2.822  0.00739 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 1.171 on 40 degrees of freedom
Multiple R-Squared: 0.8142,   Adjusted R-squared: 0.7956
F-statistic: 43.83 on 4 and 40 DF,  p-value: 4.148e-14
```

Now the taxrate is important, but the cost of bread is not. What's going on here? What can we believe?

The problem is collinearity: engineers salary and tax rates are correlated. Not terribly strongly (about .6 for the log transforms), but enough to cause some problems.

You can think about this in terms of the interpretation. This full model says that among cities in which engineers pay the same tax, an increase of salary corresopnds to a decrease in the cost of a BigMac. But are there any such cities? The correlation means that cities in which engineers make more also charge more taxes, and so this is somewhat of a "false" interpretation. This is why it is very difficult to get causal conclusions out of these observational studies -- we simply don't observe enough cities that charge the same tax rates for engineers who make different salaries.

The added variable plot can be useful. First, fit the response without lengtax and save the residuals. Then, fit lengsal on lengtax and save residuals. Now plot the residuals of the first (everything in bigmac prices without the affect of engineers' tax) against the residuals of the second (everything with engineer's salary that is not influenced by engineers' tax.) If this is a vertical plot, then the two variables are a linear combination of each other. (Because essentially it means that engineer's tax removes all of the variability from engineer's salary.)

This all speaks to the need for some sort of concerted strategy for comparing models and deciding which we use and which we don't, and deciding which variables to use.