Two Geometric Views of Least Squares Fits

Scatterplot view: We can plot the data points, (x_i, y_i), in $p+1$ dimensional space. The least squares fit defines a hyperplane that minimizes the sum of squared distances to points in the y direction (the squared residuals).

See Figure 3.1 in the text.

Projection view: Instead of plotting data points, we can plot the variables (inputs and response) as points in N dimensional space. Each variable is represented by a vector from the origin to the point whose coordinates are its values in the N training cases.

The least squares fit produces a vector of fitted values, $\hat{\mathbf{y}}$, that is the projection of $\mathbf{y}$ onto the space spanned by $x_1, \ldots, x_p$.

(Actually, if we have an intercept, β_0, we need a "variable" for it too, which is represented by the vector from the origin to $(1, 1, \ldots, 1, 1)$.)

See Figure 3.2 in the text.

Why Least Squares?

Using the least squares estimate for β has two justifications:

> It is the *maximum likelihood estimate* if we assume that the "noise" in the relationship of y to x is normal (Gaussian), independently for each x_i, with the same variance.
>
> It is the *unbiased linear estimate with smallest variance* if the noise is independent — for any distribution of noise with finite variance.

However, the normal assumption is often dubious, and there's no particular reason why we should want an estimate that's linear or unbiased.

There's also a pragmatic justification — the least squares estimate is easy to compute.

Statistical Inference for LS Linear Regression

If the noise is normal, or N is big enough for the Central Limit Theorem to take effect, we can find *standard errors* for estimates, and do tests of "significance".

Prostate example: Training set of size 67, predicting `lpsa` from 8 inputs.

```
> summary(lm(lpsa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,data=pt))

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.429170   1.553588   0.276  0.78334
lcavol       0.576543   0.107438   5.366 1.47e-06 ***
lweight      0.614020   0.223216   2.751  0.00792 **
age         -0.019001   0.013612  -1.396  0.16806
lbph         0.144848   0.070457   2.056  0.04431 *
svi          0.737209   0.298555   2.469  0.01651 *
lcp         -0.206324   0.110516  -1.867  0.06697 .
gleason     -0.029503   0.201136  -0.147  0.88389
pgg45        0.009465   0.005447   1.738  0.08755 .

Residual standard error: 0.7123 on 58 degrees of freedom
```

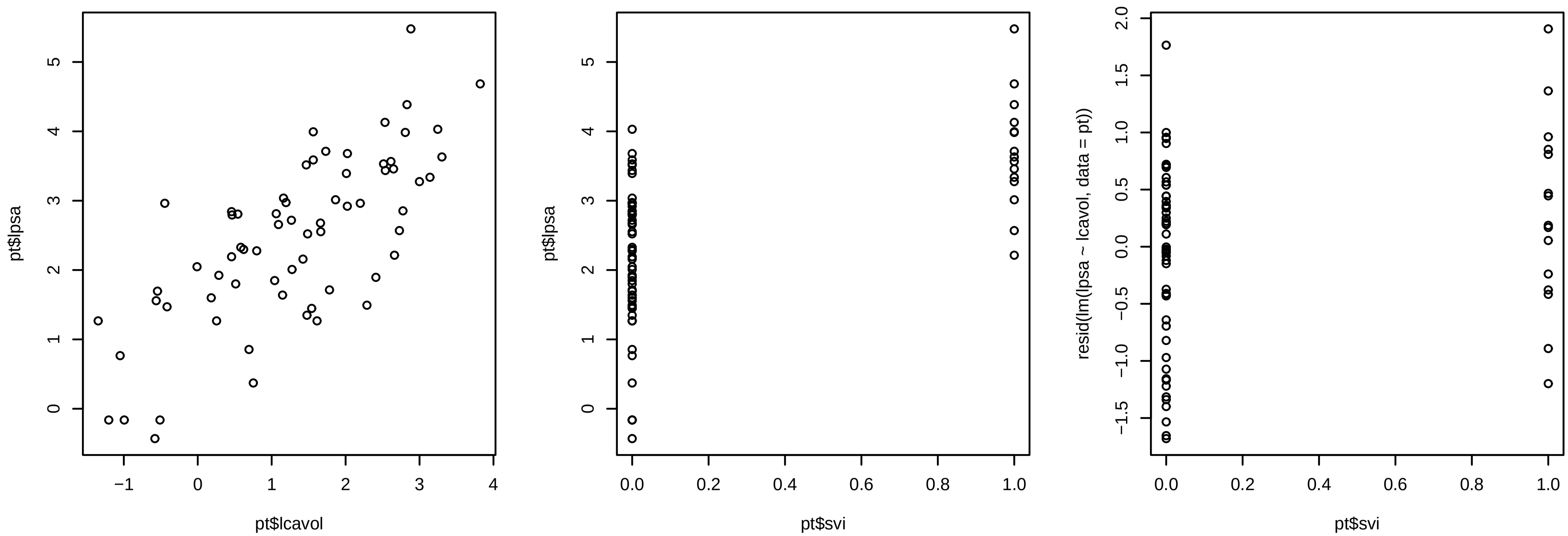
Significance of Inputs in Different Models

Whether an input has a significant effect varies with the other inputs used. What matters is whether the input contains *additional* information about the response.

```
> summary(lm(lpsa~svi,data=pt))

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0938     0.1402  14.936  < 2e-16 ***
svi           1.6015     0.2963   5.406 9.88e-07 ***

Residual standard error: 1.011 on 65 degrees of freedom

> summary(lm(lpsa~lcavol+svi,data=pt))

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5376     0.1456  10.561 1.17e-15 ***
lcavol        0.6040     0.1000   6.039 8.70e-08 ***
svi           0.5418     0.2959   1.831   0.0718 .

Residual standard error: 0.8131 on 64 degrees of freedom
```

Relating an Input to Residuals Using Other Inputs

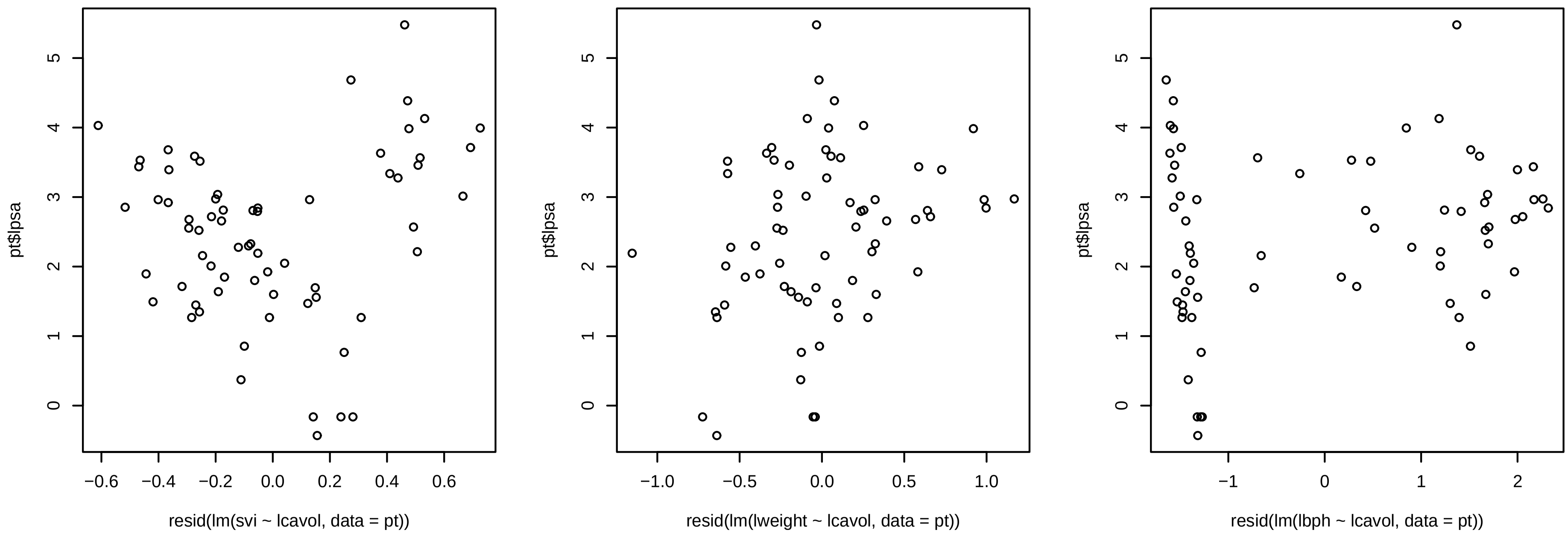
One way of visualizing whether an input gives additional information about the response is to plot it against the *residuals* after regressing on other inputs:



From the left and centre plots, it's clear that `lpsa` is related to both `lcavol` and `svi`. But the right plot shows that the residuals from regressing `lpsa` on `lcavol` are at most weakly related to `svi`.

Looking at the Residuals of an Input Given Other Inputs

Another way of visualizing whether an input gives additional information is to plot the response against the residuals from regressing this input on *other inputs*:



We see on the left that there's at most a weak linear relationship of `lpsa` to the residual of `svi` on `lcavol`. (But maybe there's a non-linear relationship?)

There do seem to be linear relationships of `lpsa` to the residuals of `lweight` and `lbph` on `lcavol`. (But is something non-linear going on with `lbph` too?)

Subset Selection Strategies

Because of these sort of effects, it is difficult to tell what subset of inputs should be included in a linear model.

Three strategies:

> Look at **all subsets**.
>
> Use a **forward selection** procedure: Start with nothing, and add inputs one at a time.
>
> Use a **backward selection** procedure: Start with all inputs, eliminate them one at a time.

When looking at all subsets, we need a criterion for deciding which is best. We can't use RSS, since it's always smaller with more inputs. We might use the unbiased estimate of residual standard deviation; many other possibilities.

With forward and backward selection, we need to know when to stop. We could use the same criterion as above, or we might use significance tests.