STAT 213
Model Selection

Colin Reimer Dawson

Oberlin College

March 26, 2018

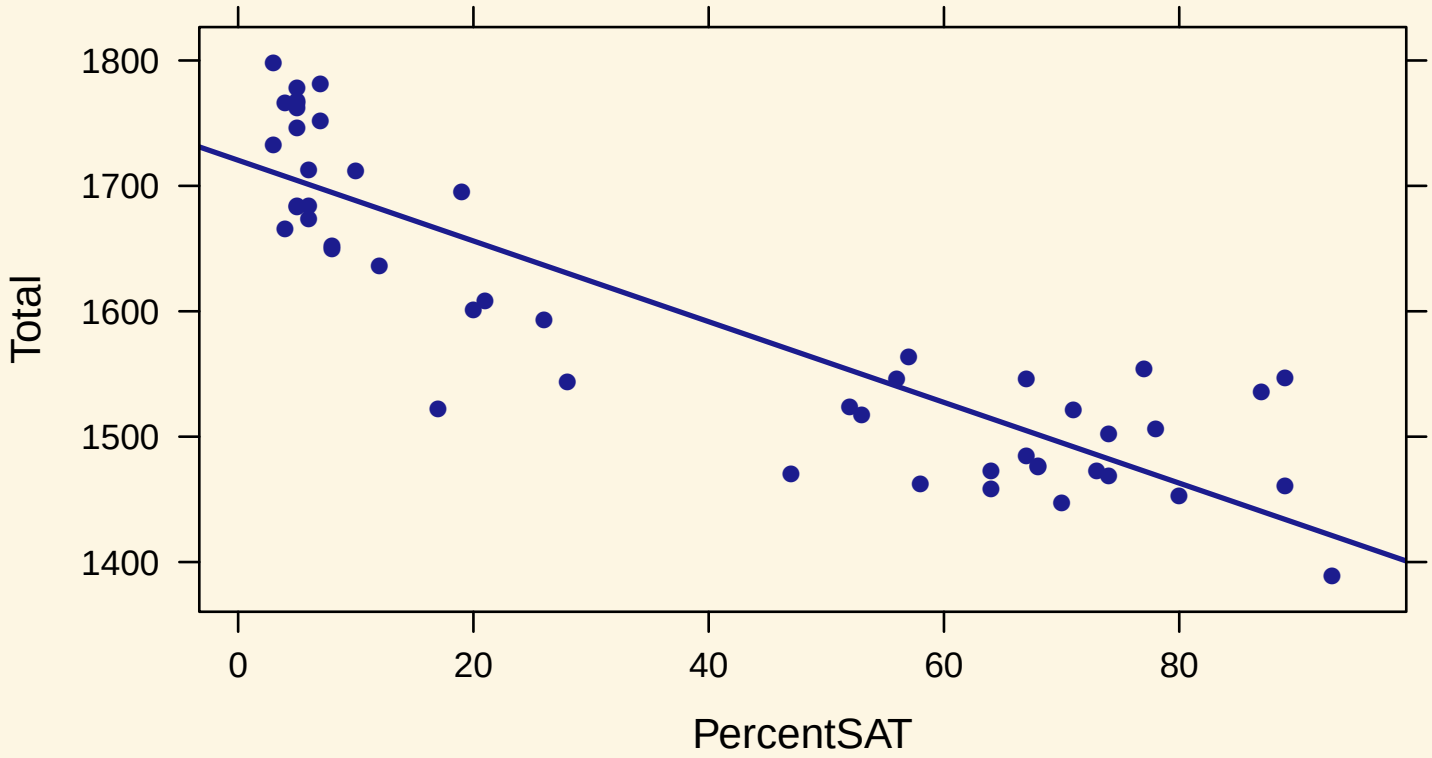
Outline

Outline

SATs and Teacher Salary

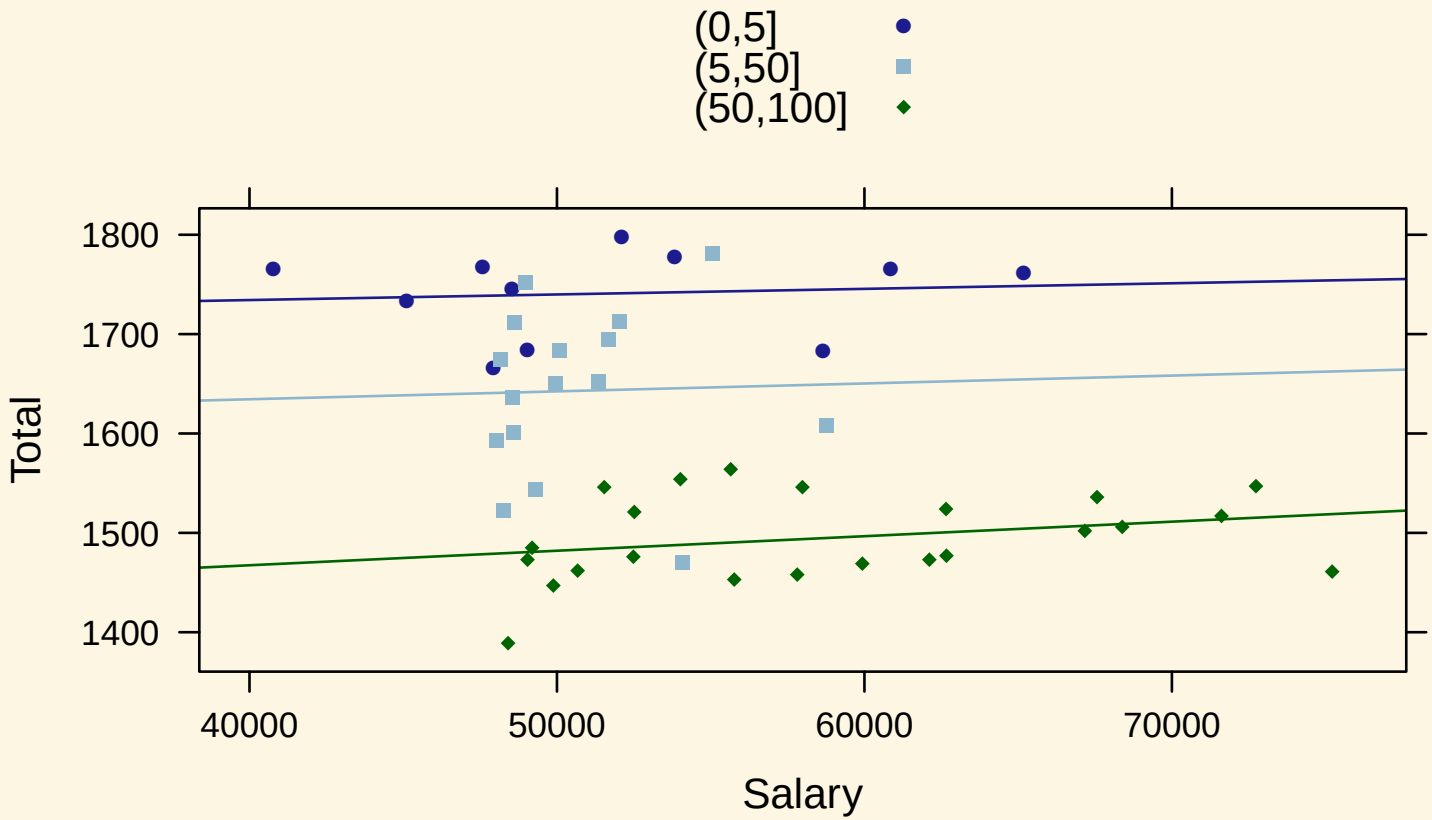
Overall, states that pay teachers more have lower mean SAT scores.

SATs and Teacher Salary



Overall, states with higher participation rates have lower mean SAT scores.

Controlling for Participation



Controlling for participation, the more teachers are paid, the higher the mean score.

Controlling for Participation

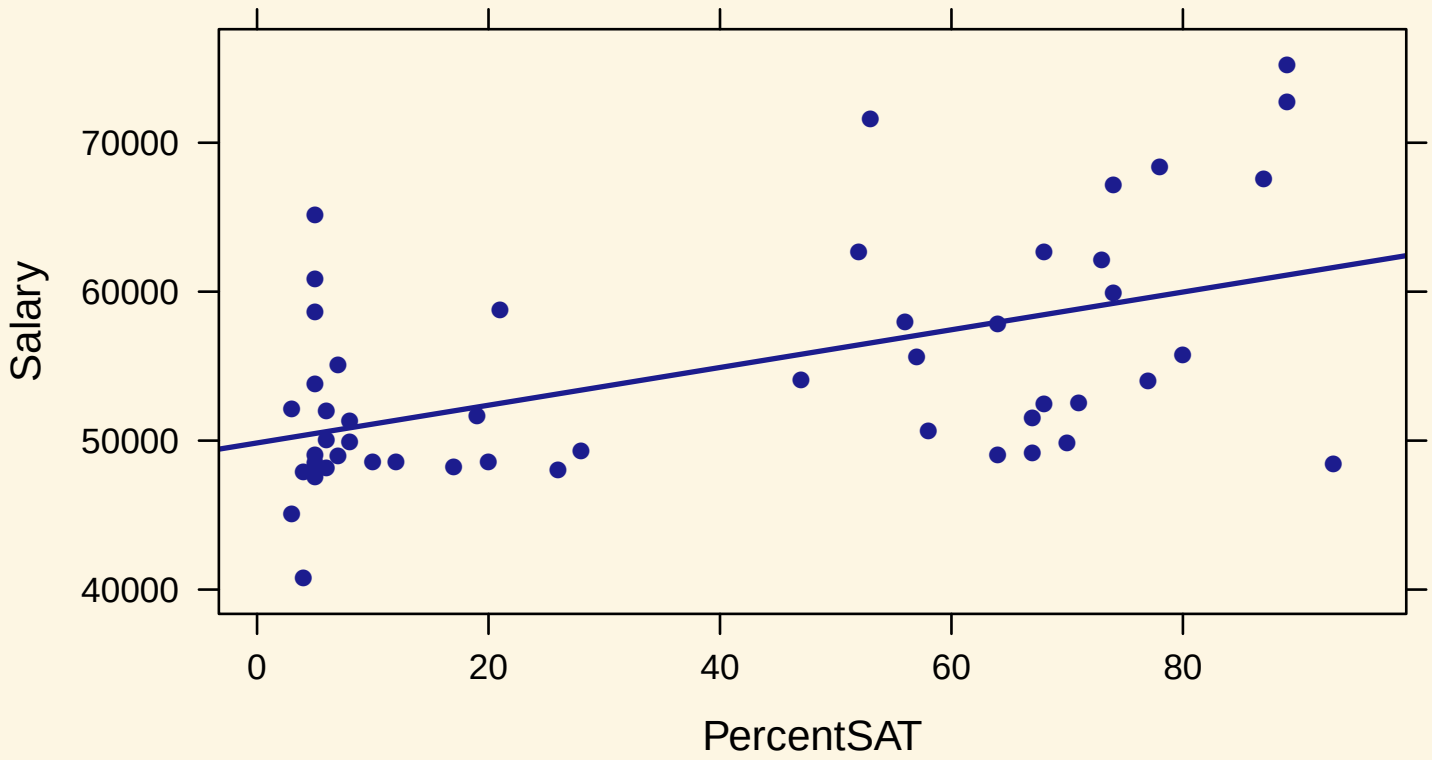
```
m.salary.percent <- lm(Total ~ Salary + PercentSAT, data = SATdata)
m.salary.percent %>% summary() %>% coefficients() %>% round(3)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1589.007     58.471  27.176    0.000
Salary         0.003      0.001   2.295    0.026
PercentSAT    -3.553      0.278 -12.756    0.000
```

In the full model the `Salary` coefficient has the opposite sign from the simple model! This is an instance of **Simpson's Paradox**

What's Going On?

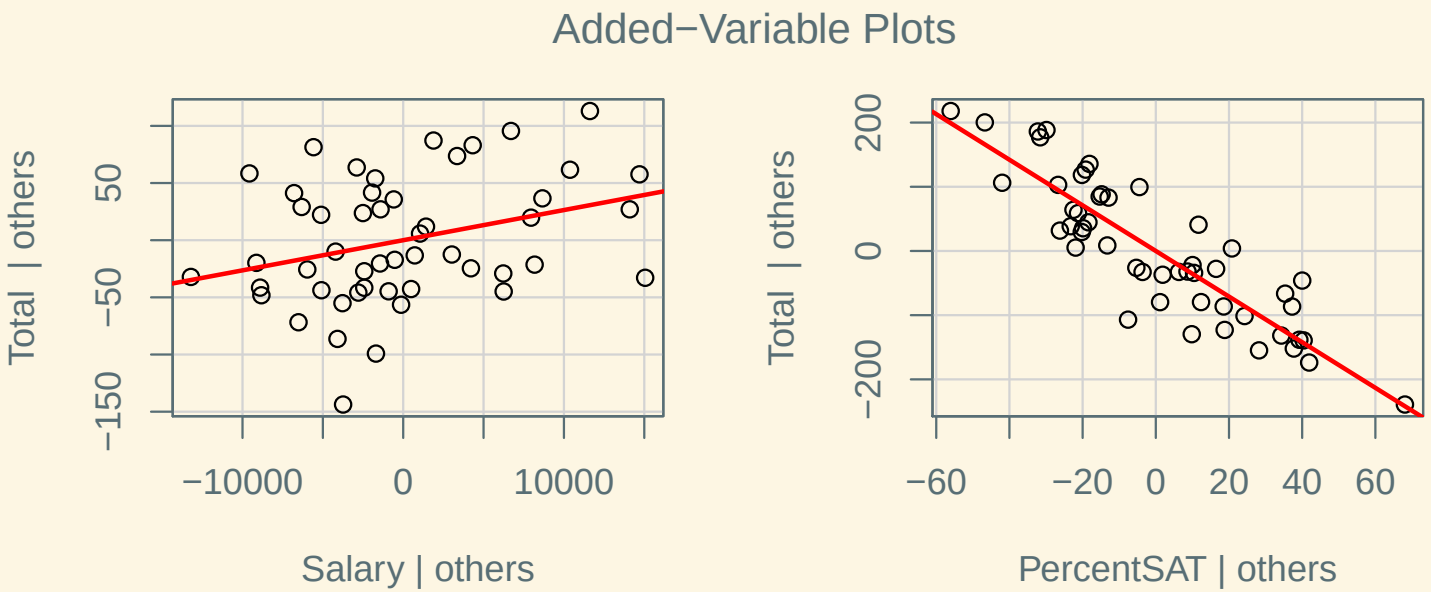
```
xyplot(Salary ~ PercentSAT, data = SATdata, type = c("p", "r"))
```



PercentSAT was a **confounding variable**

Outline

Visualizing "Added Value"



To see the unique contribution of a predictor, X_k, we can plot two sets of residuals against each other

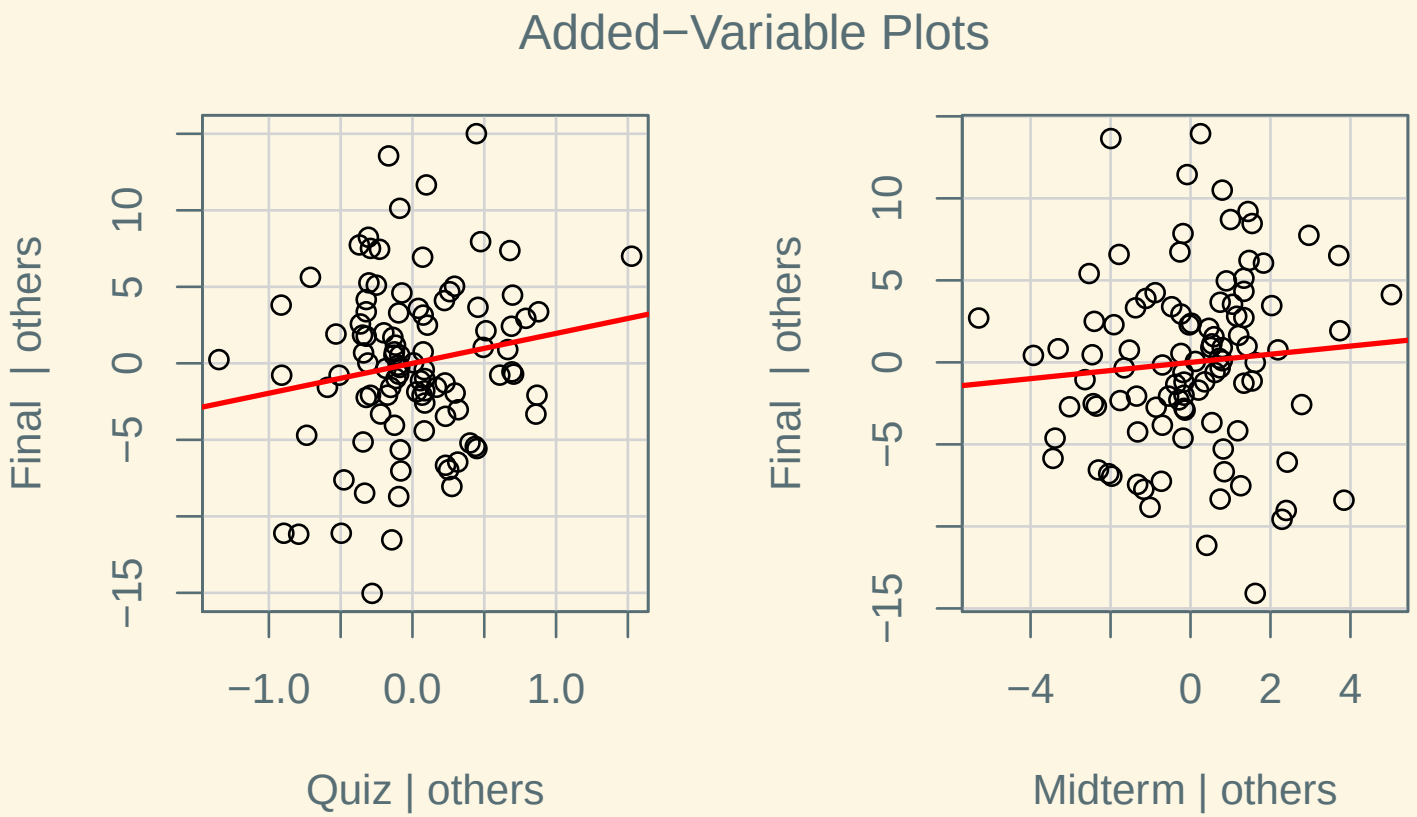
1. Residuals from model predicting Y from other predictors
2. Residuals from model predicting X_k from other predictors

Added Variable Plots

- Also called "partial regression plots"
- Idea: Capture the *nonredundant* information provided by X_k
- How much information does X_k provide about Y that we couldn't have already predicted?

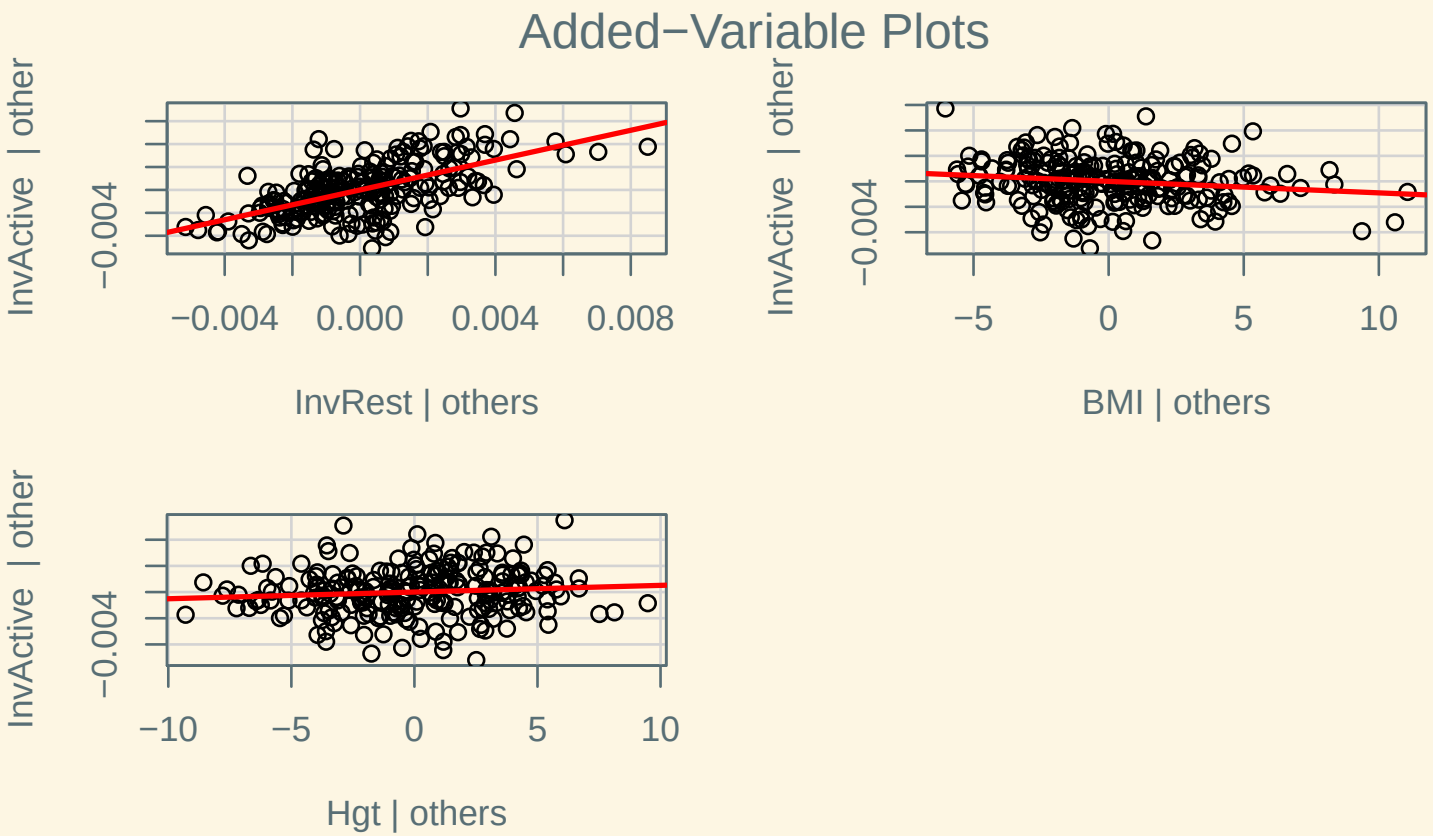
Example: Quiz and Exam Scores

```
m.both <- lm(Final ~ Quiz + Midterm, data = Scores)
avPlots(m.both)
```



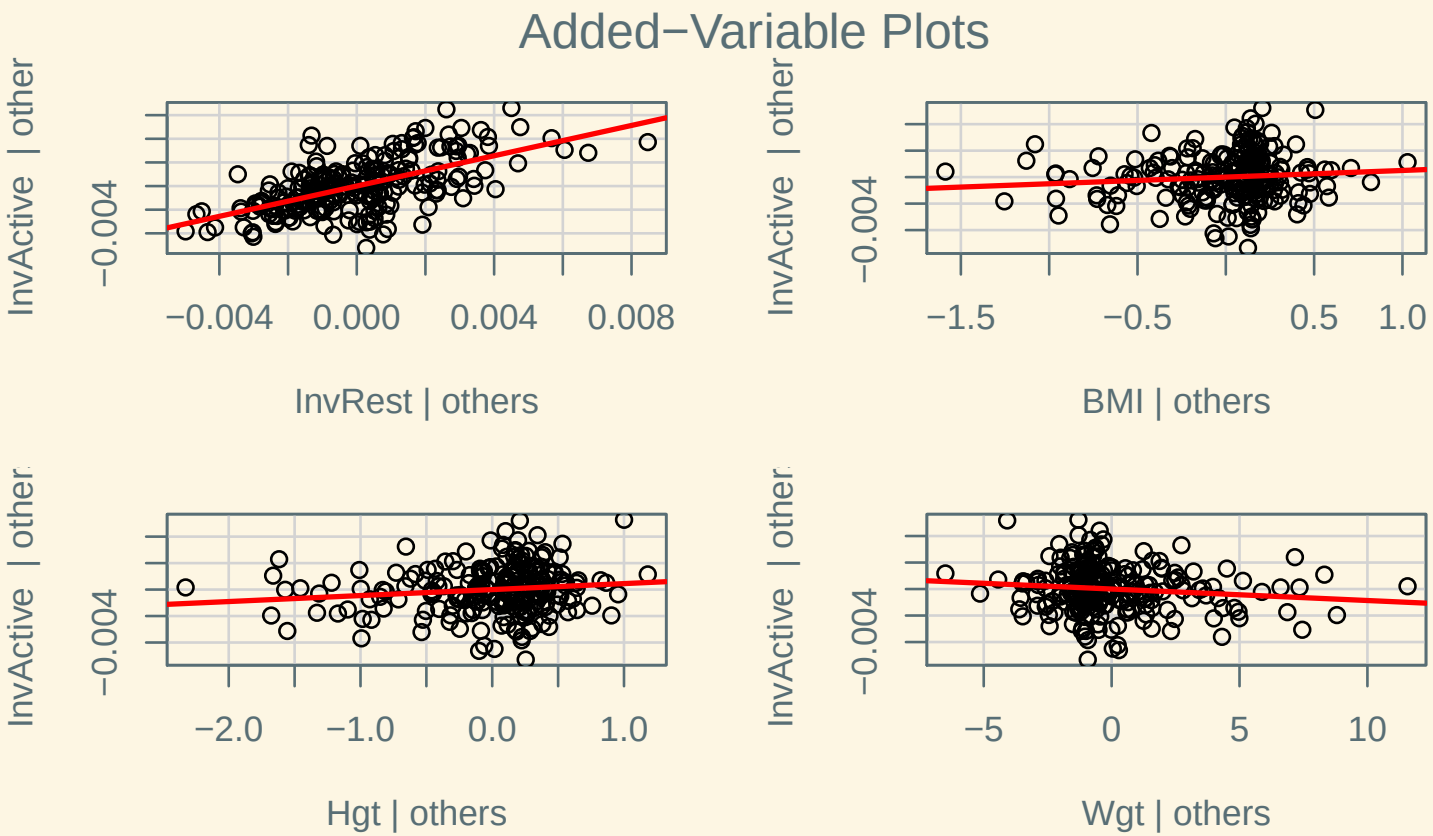
Example: Pulse Rates

```
m.pulse <- lm(InvActive ~ InvRest + BMI + Hgt, data = Pulse2)
avPlots(m.pulse)
```



Example: Pulse Rates (With Multicollinearity)

```
m.pulse2 <- lm(InvActive ~ InvRest + BMI + Hgt + Wgt, data = Pulse2)
avPlots(m.pulse2)
```



Note the differences in x axis range!

Outline

So many models...

- How to decide among all these models?
  1. Understand the subject area! Build sensible models.
  2. Nested F-tests to make *targeted* comparisons of interest
  3. Model quality measures

ASSESS: Coefficient of Determination

Can we use R^2 to identify the best model?

As before, $R^2 = \frac{SS_{Model}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$

What Happens if We Add Useless Predictors?

Simulation Handout

What Makes a "Good" Model?

Fit

High R^2

Small SSE (equivalent)

Validity

Satisfies stated conditions

Strong evidence for predictors

Simple (Parsimonious)

Generalizes to new data

Why Does Parsimony Matter?

Don't we just care about good predictions?

Not exclusively...

- We also use models to *understand* the world (harder with more complexity)

And even so...

- We really care about making predictions for data we *haven't seen yet*.

Balancing Fit and Parsimony

- R^2 can only go up as we add a predictor, because at worst, we can choose $\beta_{K+1} = 0$ and get the same SSE. Usually we can pick coefficients to do somewhat better.
- Would like to "penalize" unnecessary predictors.

Adjusted R^2 (Three equivalent formulas)

$$R^2_{adj} = 1 - \frac{SS_{Error}/(N-K-1)}{SS_{Total}/(N-1)}$$

$$= 1 - \frac{\hat{\sigma}^2_\varepsilon}{s^2_Y}$$

$$1 - R^2_{adj} = \left(1 - R^2\right) \frac{df_{Total}}{df_{Error}}$$

Criteria to "score" models

1. high R^2/low SSE/low $\hat{\sigma}_\varepsilon^2$: always favors more complex models
2. Adj. R^2: balances fit and complexity
3. Mallow's C_p / Akaike Information Criterion (AIC): estimates mean squared prediction error based on $\hat{\sigma}_\varepsilon^2$ from a "full" model
4. Out-of-sample predictive accuracy

Mallow's C_p / AIC

Two measures that reduce to the same thing in the case of MLR with independent, equal variance, Normal residuals. For a "reduced" model with $p_{reduced}$ total parameters (including the intercept) which is nested in a "full" model with p_{full} parameters, both fit using n observations:

$$C_p = \frac{SSE_{\text{reduced}}}{MSE_{\text{full}}} + 2p_{\text{reduced}} - n \quad (1)$$

$$= p_{\text{reduced}} + \frac{SS_{nested}}{MSE_{full}} \quad (2)$$

where SS_{nested} is the (positive) difference between the sums of squares of the two models (as for a nested F test)

Should we prefer larger or smaller values?