• We’ve been looking at individual outliers and the influence they can have on the fitted model.

• To greatly impact the fitted regression coefficients, a point must be an outlier with respect to the independent variables (i.e. have high leverage), and its Y-value must be an outlier for the given independent variables (i.e. have a large studentized residual.)

• In the Duncan data set, the data point ‘minister’ had the most influence.

  This data point had a large positive studentized residual... the prestige for the ‘minister’ was higher than expected for the given income and education (seems reasonable).

  The ‘minister’ also had an odd combination of income/education with a low income for their given amount of education (also seems reasonable).
• The fitted regression coefficients with the ‘minister’ data point:

> summary(lm.out)

Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| -6.06466 | 4.27194    | -1.420  | 0.163    |
| education  | 0.54583  | 0.09825    | 5.555   | 1.73e-06 *** |
| income     | 0.59873  | 0.11967    | 5.003   | 1.05e-05 *** |

• The fitted regression coefficients without the ‘minister’ data point:

> summary(lm(prestige[-6]~education[-6]+income[-6]))

Coefficients:

|                  | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------|----------|------------|---------|----------|
| (Intercept)      | -6.62751 | 3.88753    | -1.705  | 0.0958   . |
| education[-6]    | 0.43303  | 0.09629    | 4.497   | 5.56e-05 *** |
| income[-6]       | 0.73155  | 0.11674    | 6.266   | 1.81e-07 *** |

The income coefficient is higher without ‘minister’.
The education coefficient is lower without ‘minister’.
(these relationships are seen in the added variable plots.)
Jointly influential data points

- The previous section considered the impact that individual data points can have on the fitted regression coefficients.

- Their impact can be seen with leave-one-out deletion diagnostics. See...

  > influence.measures(lm.out)

- But data points can have a joint influence on the fitted model.

- Graphical methods can help identify such subsets.

- For this purpose, we will use the Partial Regression Plot that we saw earlier (in the multiple regression introduction). Also known as Added Variable Plot.
Partial Regression Plot for $X_i$

For $X_i$,

1. get the residuals from the regression of $Y$ on all predictors except $X_i$, and plot on the Y axis (what’s left for $X_i$ to explain).
2. get the residuals from the regression of $X_i$ on all other predictors, and plot on the X axis (the non-redundant part of $X_i$ in the model).

The regression of $Y$ on $X$ from the plot above gives the multiple regression coefficient for $X_i$.

After we fit the regression to this plot, the residuals are the same as the multiple regression residuals.

As this plot is used to fit the multiple regression coefficient, the relationship should be linear.
Example: Returning to Duncan model:

> av.plots(lm.out, education, labels=row.names(Duncan))
> av.plots(lm.out, income, labels=row.names(Duncan))

Here we see how the multiple regression coefficient for education is higher when ‘minister’ is included (influential point).

The ‘conductor’ data point influences it in the same direction (removal of both at once may show a large change in the regression coefficient).
Here we see that ‘minister’ and ‘conductor’ work together to *flatten* the income multiple regression coefficient. Removal of both at once may show a large change in the income regression coefficient.
Removing outliers (careful)

- Removal of an outlier should only be done after careful investigation, and perhaps discussion.

- It shouldn’t be removed if it’s not an error, or can’t be shown that the data point belong to a different ‘population’.

- Duncan Data set: When reporting the analysis, one could report both analyses (with and without ‘minister’). Otherwise, if all involved agree that ‘minister’ is not representative of the population at hand, it could be removed.