

## 10. Alternative case influence statistics

- a. Alternative to  $D_i$ : *dffits<sub>i</sub>* (and others)
- b. Alternative to  $\text{studres}_i$ : *externally-studentized residual*
- c. Suggestion: use whatever is convenient with the statistical computer package you're using

11. Note:  $D_i$  only detects influence of single-cases; influential *pairs* may go undetected

## D. Partial Residual Plots

1. **A problem:** a scatterplot of  $y$  vs  $x_2$  gives information regarding  $m(y/x_2)$  about (a) whether  $x_2$  is a useful predictor of  $y$ , (b) non-linearity in  $x_2$  and (c) outliers and influential observations.

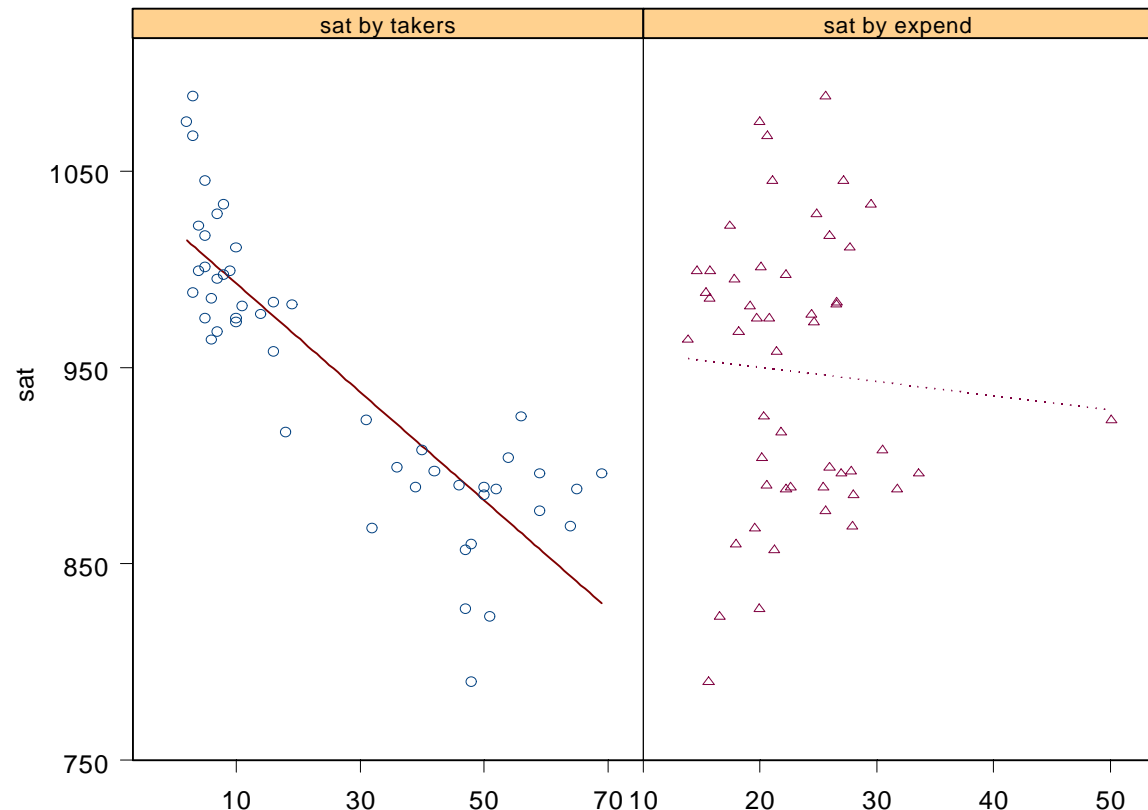
We would like a plot revealing (a), (b), and (c) for  $m(y/x_1, x_2, x_3)$  (e.g. what is the effect of  $x_2$ , *after accounting for*  $x_1$  and  $x_3$ ?)

2. **Example (Case 1201)** Is the distribution of state average SAT scores associated with state expenditure on public education, after accounting for percentage of high school students who take the SAT test?
3. We would like to visually explore the function  $f(\text{expend})$  in  $m(\text{SAT}|\text{takers}, \text{expend}) = \beta_0 + \beta_1 \text{takers} + f(\text{expend})$

“Marginal” plots  
 $y$  vs.  $x_1$   
and  $y$  vs.  $x_2$

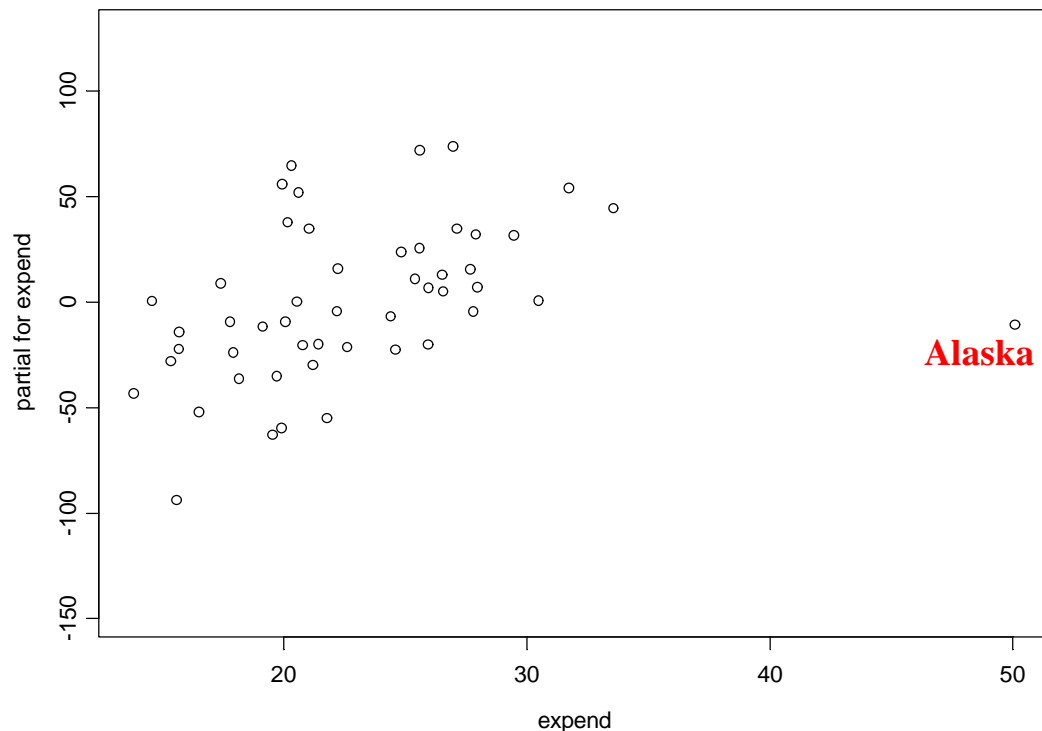
Is the state with  
largest expenditure  
influential?

Is there an assoc-  
iation of expend and  
SAT, after accounting  
for takers?



4. We'd like to plot  $y$  versus  $x_2$  but with the effect of  $x_1$  subtracted out; i.e. plot  $y - b_0 - b_1x_1$  versus  $x_2$
5. To approximate this, get the *partial residual for  $x_2$* :

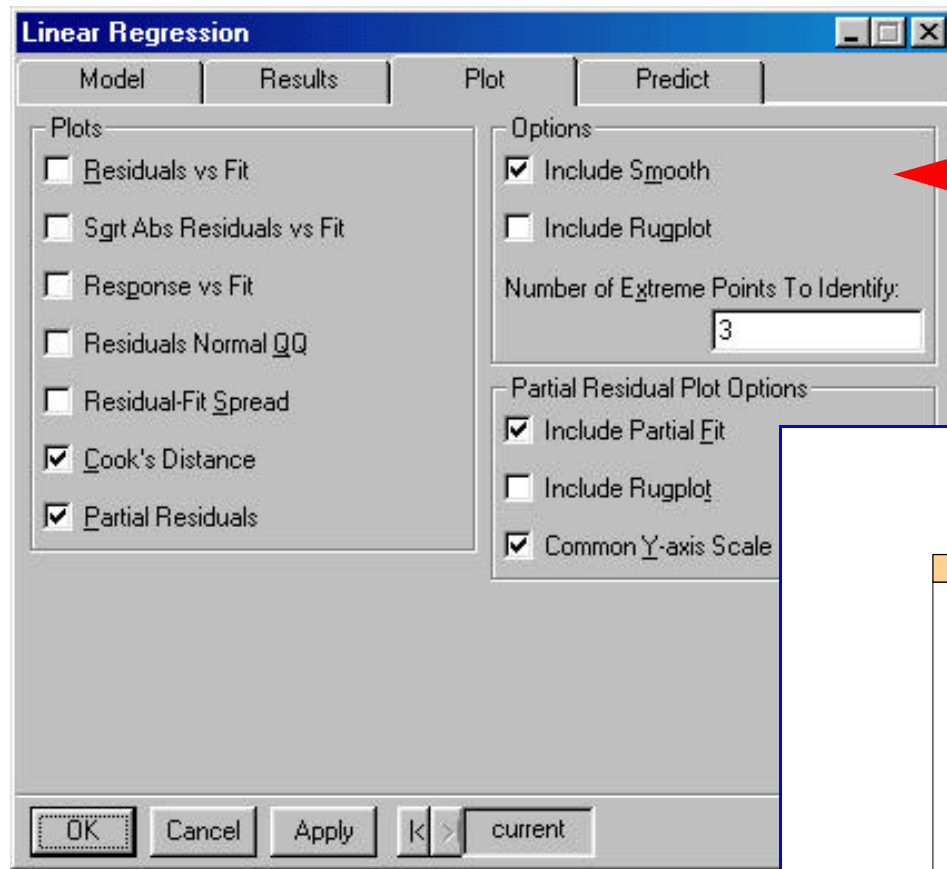
- a. Get  $\hat{b}_0, \hat{b}_1, \hat{b}_2$  in  $m(y | x_1, x_2) = b_0 + b_1x_1 + b_2x_2$
- b. Compute the partial residual as  $pres = y - \hat{b}_0 - \hat{b}_1x_1$
6. This is also called a *component plus residual*; if  $res$  is the residual from 3a,  $pres = res + \hat{b}_2x_2$  (easier computation)



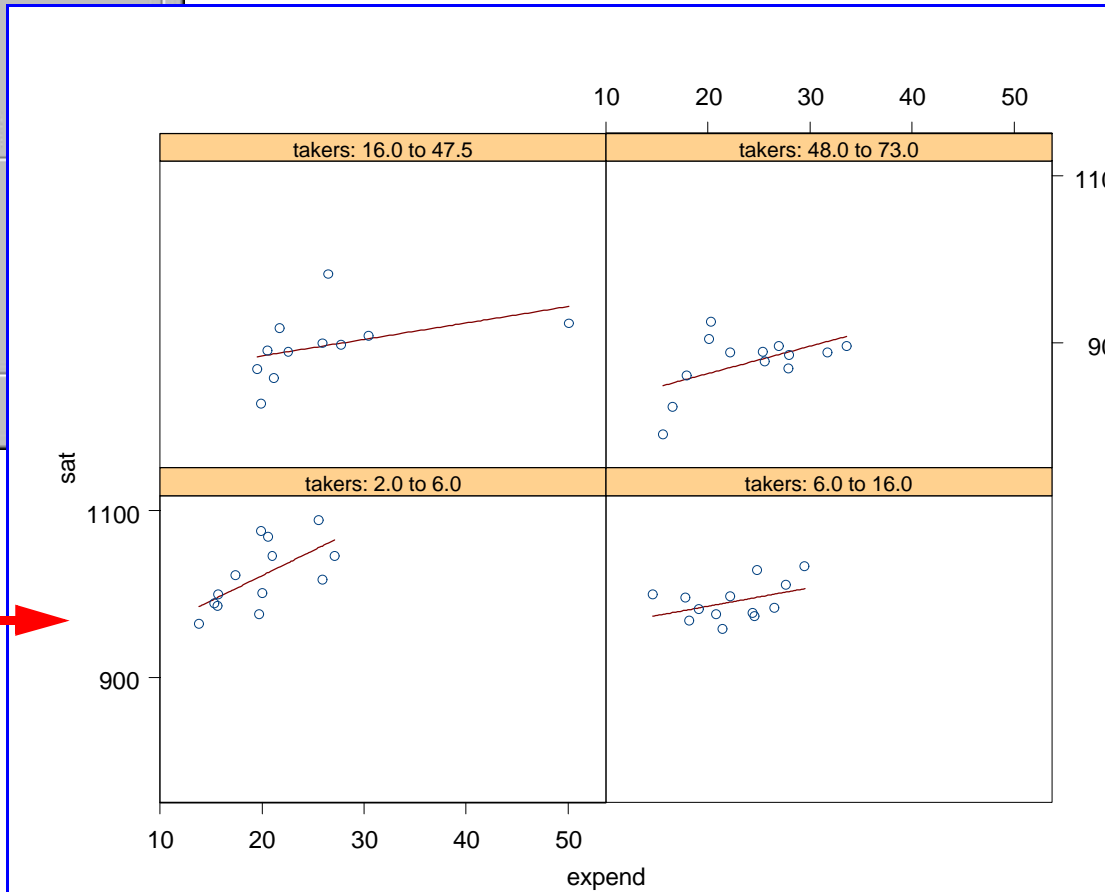
**Notice:**

**1. Alaska is unusual in its expenditure, and is apparently quite influential**

**2. After accounting for % of students who take SAT, there is a positive association between expenditure and mean SAT score**



**Getting partial residual plots in S-PLUS**



**A Trellis Graph is another way to look at  $y$  vs  $x_2$  for roughly fixed values of  $x_1$**

## E. Weighted regression for certain types of non-constant variance

1. Suppose  $m(y | x_1, x_2) = b_0 + b_1x_1 + b_2x_2$ ,  $\text{var}(y | x_1, x_2) = \mathcal{S}^2 / w_i$



and the  $w_i$ 's are known

2. *Weighted least squares* is the appropriate tool for this model; it minimizes the weighted sum of squared residuals

$$\sum_{i=1}^n w_i (y_i - \hat{b}_1 x_{1i} - \hat{b}_2 x_{2i})^2$$

3. In statistical computer programs: use linear regression in the usual way, specify the column  $w$  as a *weight*, read the output in the usual way

#### **4. Important special cases where this is useful**

**a.  $y_i$  is an average based on a sample of size  $m_i$**

**In this case, the weights are  $w_i = 1/m_i$**

**b. the variance is proportional to  $x$ ;**

**so  $w_i = 1/x_i$**

## F. Measurement errors in $x$ 's

- 1. Fact: least squares estimates are biased and inferences about  $m(y | x_1, x_2) = b_0 + b_1x_1 + b_2x_2$  can be misleading if the available data for estimating the regression are observations  $y, x_1, x_2^*$ , where  $x_2^*$  is an imprecise measurement of  $x_2$  (even though it may be an unbiased measurement)**
- 2. This is an important problem to be aware of; general purpose solutions do not exist in standard statistical programs**
- 3. Exception: if the purpose of the regression is to predict future  $y$ 's from future values of  $x_1$  and  $x_2^*$  then there is no need to worry about  $x_2^*$  being a measurement of  $x_2$**
- 4. The available remedies are beyond the level of this course**

# V. Variable Selection

## A. Introduction

1. **The problem: we may be faced with a fairly large number of potential explanatory variables (5, 10, 20, 50)**
2. **Two good reasons for seeking a subset:**
  - a. **General principle: smaller is better (Occam's razor)**
  - b. **Unnecessary terms add imprecision to inferences**
3. **Computer assisted tools**
  - a. **Judge fits of all possible models (include or excluding each  $X$ ); compare with these statistics:  $C_p$ , AIC, or BIC**
  - b. **Stepwise regression (search along favorable directions)**
4. **But don't expect a BEST or a TRUE model or a law of nature**

## **B. Objectives when there are many $X$ 's (12.2.1)**

- 1. Assessment of one  $X$ , after accounting for many others**
  - a. Ex: Do males receive higher salaries than females, after accounting for legitimate determinants of salary?**
  - b. Strategy: first find a good set of  $X$ 's to explain salary; then see if the sex indicator is significant when added in**
- 2. Fishing for association; i.e. what are the important  $X$ 's?**
  - a. The trouble with this: we can find several subsets of  $X$ 's that explain  $Y$ ; but that doesn't imply importance or causation**
  - b. Best attitude: use this for hypothesis generation**
- 3. Prediction (this is a straightforward objective)**

**Find a useful set of  $X$ 's; no interpretation required**

## C. Loss of precision due to multicollinearity

1. Review: variance of L.S. estimator of slope in simple reg. =

$$\frac{s^2}{(n-1)s_x^2}$$

2. Fact: variance of L.S. estimator of coef. of  $X_j$  in mult. reg. =

$$\frac{s^2}{(n-1)s_j^2(1-R_j^2)}$$

3. So variance of an estimated coef. will tend to be larger if there are other  $X$ 's in the model that can predict  $X_j$

4. **The S.E. of prediction will also tend to be larger if there are unnecessary or redundant  $x$ 's in the model**
5. ***Multicollinearity*: the situation in which  $s_j^2(1 - R_j^2)$  is small for one or more  $j$ 's (usually characterized by highly correlated  $X$ 's)**
6. **Strategy: there isn't a real need to decide whether multicollinearity is or isn't present, as long as one tries to find a subset of  $X$ 's that adequately explains  $\mu(Y)$ , without redundancies**
7. **“Good” subsets of  $x$ 's: (a) lead to a small  $\hat{S}^2$  (b) with as few  $x$ 's as possible (Criteria  $C_p$ , AIC, and BIC formalize this)**

## **D. Strategy for dealing with many $X$ 's**

1. **Identify objectives; identify relevant set of  $X$ 's**

- 2. Exploration: matrix of scatterplots; correlation matrix; residual plots after fitting tentative models**
- 3. Resolve transformation and influence before variable selection**
- 4. Computer-assisted variable selection**
  - a. Best: Compare all possible subset models using either  $C_p$ , AIC, or BIC; find some model with a fairly small value**
  - b. Next best: Use sequential variable selection, like stepwise regression (this doesn't look at all possible subset models, but may be more convenient with some statistical programs)**

## **E. Sequential variable selection**

### **1. Forward selection**

- a. Start with no  $X$ 's "in" the model**

- b. Find the “most significant” additional  $X$  (with an F-test)**
- c. If its p-value is less than some cutoff (like .05) add it to the model (and re-fit the model with the new set of  $X$ 's)**
- d. Repeat (b) and (c) until no further  $X$ 's can be added**

## **2. Backward elimination**

- a. Start with all  $X$ 's “in” the model**
- b. Find the “least significant” of the  $X$ 's currently in the model**
- c. If it's p-value is greater than some cutoff (like .05) drop it from the model (and re-fit with the remaining  $x$ 's)**
- d. Repeat until no further  $X$ 's can be dropped**

## **3. Stepwise regression**

- a. Start with no  $X$ 's “in”**

- b. Do one step of forward selection**
- c. Do one step of backward elimination**
- d. Repeat (b) and (c) until no further  $X$ 's can be added or dropped**

#### **4. Notes**

- a. Add and drop factor indicator variables *as a group***
- b. Don't take p-values and CI's for selected variables seriously—because of serious data snooping (not a problem for objectives 1 and 3)**
- c. A drawback: the product is a single model. This is deceptive. Think not: “here is the best model.” Think instead: “here is one, possibly useful model.”**

## F. Criteria for comparing models

| <b>Criterion<br/>to minimize</b> | <div style="border: 1px solid blue; border-radius: 15px; padding: 5px; display: inline-block; margin-bottom: 5px;"> <i>Some function<br/>of the mean square<br/>of residuals</i> </div><br>$=$ | $f(\hat{S}^2)$   | $+$ | <div style="border: 1px solid blue; border-radius: 15px; padding: 5px; display: inline-block; margin-bottom: 5px;"> <i>Some function<br/>of the number of<br/>b's in the model</i> </div><br>$g(p)$ |
|----------------------------------|--|--|-----|---|
| <b>Cp</b>                        | $=$  | $\frac{(n-p)(\hat{S}^2 - \hat{S}_{full}^2)}{\hat{S}_{full}^2}$ | $+$ | $p$   |
| <b>BIC</b>                       | $=$  | $n \log(\hat{S}^2)$  | $+$ | $p \log(n)$   |
| <b>AIC</b>                       | $=$  | $n \log(\hat{S}^2)$  | $+$ | $2p$  |

Idea: favor models  
with small mean  
square of residuals

but penalize for  
too many x's

- 1. The proposed criteria: Mallows's Cp Statistic, Schwarz's Bayesian Information Criterion (BIC or SBC), and Akaike's Information Criterion (AIC)**
- 2. The idea behind these is the same, but the theory for arriving at the trade-off between small  $\hat{S}^2$  and small  $p$  differs**
- 3. My opinion: there's no way to truly say that one of these criteria is better than the others**
- 4. Computer programs: Fit all possible models; report the best 10 or so according to the selected criteria**
- 5. Note: one other criteria:  $\hat{S}^2 + 0$  (sometimes used; but isn't as good). An equivalent criterion is  $-R_{adjusted}^2$  (see Sect. 10.4.1)**

## **G. Cross Validation (12.6.4)**

- 1. If tests, CIs, or prediction intervals are needed after variable selection and if  $n$  is large, maybe try:**
- 2. Cross validation**
  - a. Randomly divide the data into 75% for model construction and 25% for inference**
  - b. Perform variable selection with the 75%**
  - c. Refit the same model (don't drop or add anything) on the remaining 25% and proceed with inference using that fit**

## H. Review

1. **“Regression”, “regression model”, “linear regression model”, “regression analysis”**
2. **Fitted values, residuals, least squares method of estimation**
3. **Properties of least squares; tests and confidence intervals for individual coefficients; prediction intervals; extra SS F-tests (full and reduced models)**
4. **Model building and refinement: transformation, indicator variables,  $x^2$ , interaction, variable selection**
5. **Influence and case-influence statistics**
6. **Variable selection**

**7. A note on the difference between “confounding variable” and “interaction”**

- a. Is there an association between gestation and mean brain weight after accounting for body weight?**

$$\mu\{\text{brain}\} = \beta_0 + \beta_1\text{body} + \beta_2\text{gest}$$

**( $\beta_2$  represents the association of gestation with mean brain weight after accounting for body weight.)**

- b. Is the association between gestation and brain weight different for animals of different body sizes?**

$$\mu\{\text{brain}\} = \beta_0 + \beta_1\text{body} + \beta_2\text{gest} + \beta_3\text{body}\times\text{gest}$$

**(There is an interactive effect of body and gest on brain)**

## 8. What about all those F-tests?

- a. **All** F-tests we've considered are special cases of the extra sum of squares F-test (Sect. 10.3)
- b. *F-test for overall significance of regression*  
Full: a model of interest  
Reduced: model with  $\beta_0$  only
- c. *F-test for lack-of fit*  
Full: one-way anova (separate means for each distinct combination of x's)  
Reduced: a model of interest
- d. *Partial F-test* is an F-test for a single  $\beta$

e. ***One-way ANOVA F-test***

**Full:** model with a separate mean for each group  
i.e.  $\beta_0$  and  $k-1$  indicators to distinguish  $k$  groups

**Reduced:**  $\beta_0$  only (single mean model)

f. ***“Type III” F-tests*** (a computer package term)

**Full:** model that has been specified

**Reduced:** model without a particular term

g. ***“Sequential” F-tests*** (depends on order that  $x$ 's are listed)  
(S-PLUS gives these in ANOVA output)

i. **Full:** intercept and  $x_1$

**Reduced:** intercept

ii. **Full:** intercept,  $x_1$ , and  $x_2$

**Reduced:** intercept and  $x_1$

iii. **Full:** intercept,  $x_1, x_2$ , and  $x_3$

**Reduced:** intercept,  $x_1$ , and  $x_2$

**9. In “linear regression,” what does “linear in  $\beta$ 's” mean?**

**a.  $\beta_0 \times \text{something} + \beta_1 \times \text{something} + \beta_2 \times \text{something} + \dots$**

**b. Ex. of nonlinear regression:  $\mu(y|x) = \beta_0 x^{\beta_1}$**

**10. A note about “mean response.” It is useful to explicitly write  $\mu(y|x_1, x_2, x_3)$  to talk about the mean of  $y$  as a function of  $x_1, x_2,$  and  $x_3$ . Sometimes we abbreviate this to “the mean of the response” if it’s clear what  $x$ 's we’re talking about.**

**11. Partial residuals**

**a. You *may* find a plot of partial residuals vs.  $x_1$  to be useful when it is desired to study the relationship between  $y$  and  $x_1$ , after getting the effects of  $x_2, x_3,$  etc. out of the way, especially**

**if the effect of  $x_1$  is relatively small (in which case the plot of  $y$  versus  $x_1$  does not reveal much).**

- b. For example: How is mammal brain weight related to litter size, after accounting for body weight?**
- c. Suppose  $\mu(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ . A plot of  $y$  versus  $x_1$  won't show a linear relationship whose slope is  $\beta_1$  if  $x_1$  and  $x_2$  are correlated. However, a plot of  $y - (\beta_0 + \beta_2 x_2)$  versus  $x_1$  will show a pattern whose slope is  $\beta_1$ .**
- d. So, the partial residuals are  $y_i - (\hat{b}_0 + \hat{b}_2 x_{2i})$ , where the  $\beta$ 's are the estimates from the regression of  $y$  on  $x_1$  and  $x_2$ .**