STA302 week 11

Mark Ebden 2018. Chapter 6, November 22–29

With grateful acknowledgment to Alison Gibbs

Overview

Chapter 6:

- **MLR Diagnostics**: the Seven C's
- A deeper look at Added Variable Plots
- Introduction to a dataset to cut your teeth on: house prices

Recap of SLR diagnostics: The Seven C's (7 Checks)

1. Plot *standardized residuals* to help determine whether the proposed regression model is a valid model
2. Identify any *leverage points*
3. Identify any *outliers*
4. Identify any *influential points*
5. Assess the assumption of *error homoscedasticity*
6. For time series: examine whether the data are *correlated over time*
7. Assess the assumption of *normal errors*

Are the Seven Seas Uniquely Defined?

Ya'qubi was a historian living in ninth-century Asia and Africa. He wrote, "Whoever wants to go to China must cross seven seas, each one with its own colour, wind, fish and breeze." He mentioned:

Persian Gulf	Strait of Malacca	Gulf of Thailand
Arabian Sea	Singapore Strait	South China Sea
Bay of Bengal		

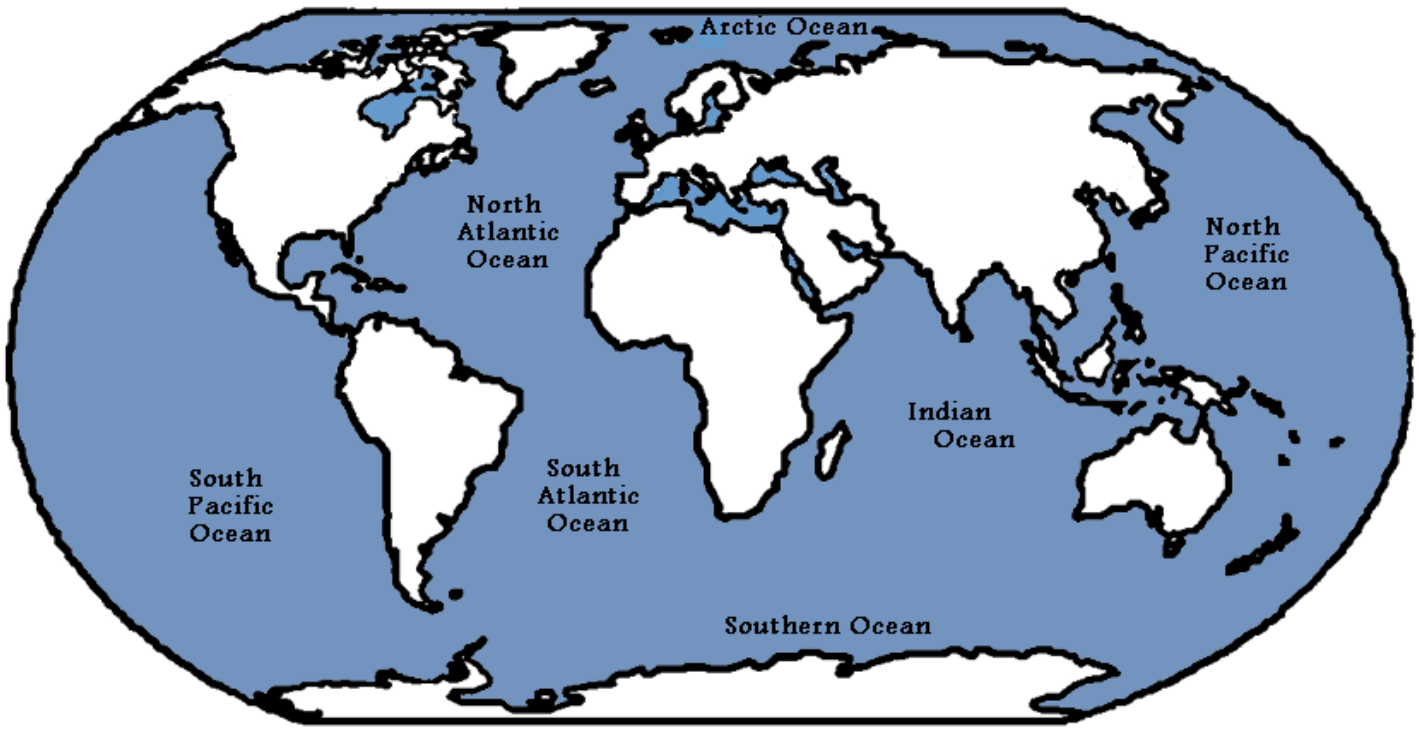
Optional material

Alternative Seven Seas

Other lists of Seven Seas date from 2300 B.C. to the present, e.g.:

- Pacific Ocean
- Atlantic Ocean
- Indian Ocean
- Arctic Ocean
- Mediterranean Sea
- Caribbean Sea
- Gulf of Mexico

or:



Optional material

MLR diagnostics: The Seven C's (7 Checks) revisited

Chapter 6 of our textbook lists seven C's for MLR. Five of them are similar to those for **SLR**:

1. Plot *standardized residuals* to help determine whether the proposed regression model is a valid model. Do further analysis, not covered here.

2. Identify any *leverage points*

3. Identify any *outliers*

6. Assess the assumption of *error homoscedasticity*

7. For time series: examine whether the data are *correlated over time*

But, two checks don't apply to SLR and are **new**:

4. Assess the effect of each X on Y

5. Assess *multicollinearity*

Five traditional C's, now applied to MLR

1. The ith standardized residual is, as before,

$$r_i = \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}}$$

However, $s = \sqrt{\text{RSS}/(n-p-1)}$ is the MLR estimate of σ; more important, there are superior techniques available, beyond the scope of this course.

2. To find leverage, we compute **H** as per earlier lectures. The threshold is:

$$h_{ii} > \frac{2(p+1)}{n}$$

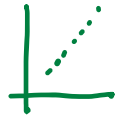
e.g. $\frac{2(1+1)}{n} = \frac{4}{n}$ for SLR

3. As before, outliers may have $|r_i| > 2$ etc, depending on the size of the dataset.

6. The constancy of variance is checked in a new way for MLR (see Check 1).

7. Correlations over time can be assessed as they were for SLR.

What about the two missing C's?



In MLR, we should still assess the assumption of **normal errors**: for this course, we can use the normal quantile plot of residuals as with SLR.

We can also still identify any **influential points**. The influence statistics are the same as those for simple linear regression:

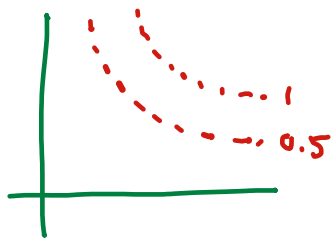
still a scalar in MLR

$$\text{DFBETAS}_{ik} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\text{s.e. of } \hat{\beta}_{k(i)}} \qquad \text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\text{s.e. of } \hat{y}_{i(i)}}$$

$$D_i = \frac{\sum_{j=1}(\hat{y}_{j(i)} - \hat{y}_j)^2}{2s^2} = \frac{r_i^2 \, h_{ii}}{2(1 - h_{ii})}$$

However, the threshold formulae are generalized for MLR:

- DFBETAS $> 2/\sqrt{n}$
- DFFITS $> 2\sqrt{\frac{p+1}{n}}$
- Cook's distance $D > \frac{4}{n-(p+1)}$

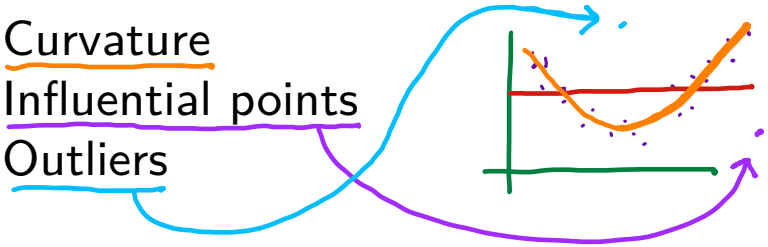


...and as usual, a large gap between the point's metric and the others is also indicative, and some authors set a threshold of 1, etc. As with SLR, there isn't universal agreement.

Focus on MLR Check 1: Residual plots

Plotting (standardized) residuals versus x_j for $j \in \{1, \ldots p\}$ helps us to look for:

- Curvature
- Influential points
- Outliers



Plotting (standardized) residuals versus $\hat{y}$ helps us to look for:

- Nonconstant variance
- Outliers

Plotting residuals versus other potential predictors can help expand our model as appropriate, as mentioned in our SLR work. L eg Z, walkability

And as with SLR, residual plots are not the only way to assess model assumptions. For example, plots of y vs x_j help us answer:

- Is the linear model appropriate?
- Are there unusual points? (e.g. potential outliers or influential points)

We can also look at the added variable plots (Check 4, to come).

MLR Check 5: Multicollinearity

eg Light intensity: $X_1, X_2, \ldots X_6$

Multicollinearity occurs when there is heavy correlation among the X's. We use the term interchangeably with "ill-conditioning".

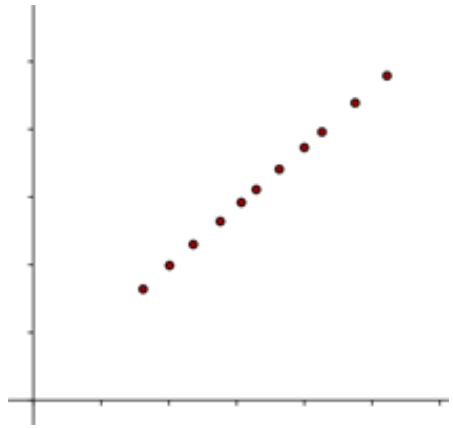
When explanatory variables are highly correlated, it's difficult or impossible to measure the individual variable's influence on the response.

The fitted equation is unstable:

- The estimated regression coefficients vary widely from data set to data set (even if the data sets are very similar), and depending on which other predictor variables are in the model
- An estimated coefficient may have opposite sign to what you'd expect
- A coefficient might not be statistically significantly different from zero even though there is a strong relationship between the X and Y when only considering X and Y

Multicollinearity

Recall: $\hat{\beta} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. If some the X's are perfectly correlated, $\mathbf{X}'\mathbf{X}$ is singular and we can't calculate $\mathbf{b}$.



Put another way, in terms of the Week 8 SLR material: if $\mathbf{X}$ contains linearly dependent columns, $\mathbf{X}$ has a rank below $p+1$. Therefore $(\mathbf{X}'\mathbf{X})^{-1}$ has a rank below $p+1$. A matrix must have full rank to be invertible.

In the case of an $\mathbf{X}'\mathbf{X}$ which is *close* to singular, the determinant of $\mathbf{X}'\mathbf{X}$ will be *near* 0. Therefore, $\text{var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ will be large. This means that the standard errors of the estimated coefficients will be large, so we'll have "inefficient" estimates. (We can't make precise statements about their values.)

Quantifying Multicollinearity

eg $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$

If $j=1$: $x_1 | \ldots | x_2$

Let R_j^2 represent the coefficient of multiple determination obtained when the jth predictor variable is regressed against the other predictor variables.
The **variance inflation factor** is

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

Because $\text{var}(\hat{\beta}_j) \propto \frac{s^2}{1 - R_j^2}$

A large VIF_j is a sign of multicollinearity. Rules of thumb:

- If $5 \lesssim \text{VIF}_j \lesssim 10$, the effects of multicollinearity might be seen (this is a warning)
- If $\text{VIF}_j \gtrsim 10$, there is a serious problem

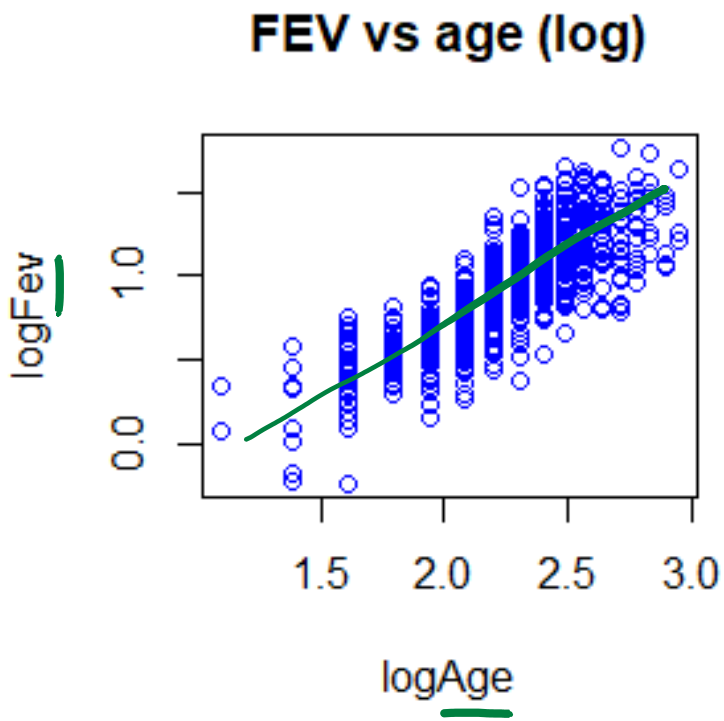
However, if the jth variable's coefficient is statistically significantly different from zero, this can be an indication not to worry as much about a high VIF_j.

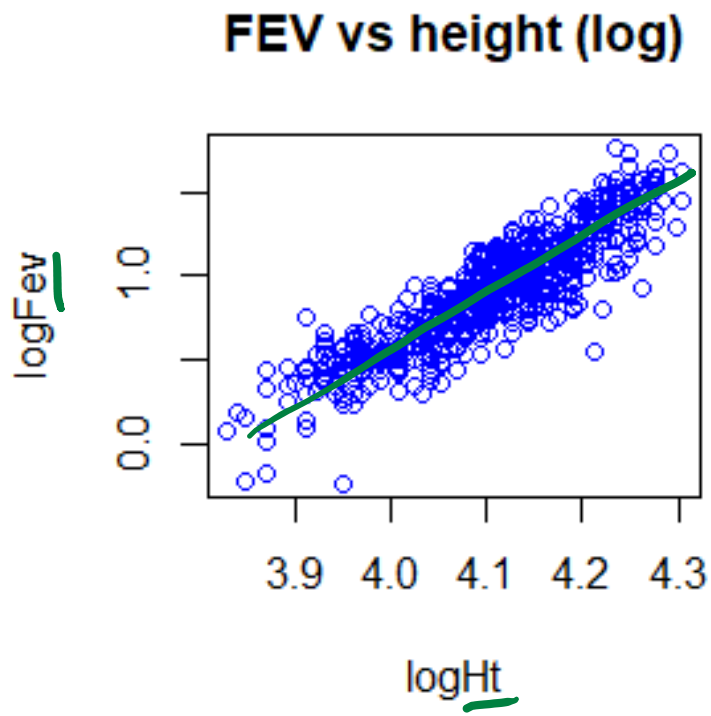
Optional material: **Tolerance** is defined as $1/\text{VIF}_j$

Multicollinearity in the FEV dataset

We'd calculated last week that $\beta_1 \approx 0.18$ for the relationship between `logFEV` and `logAge` in a particular MLR context.

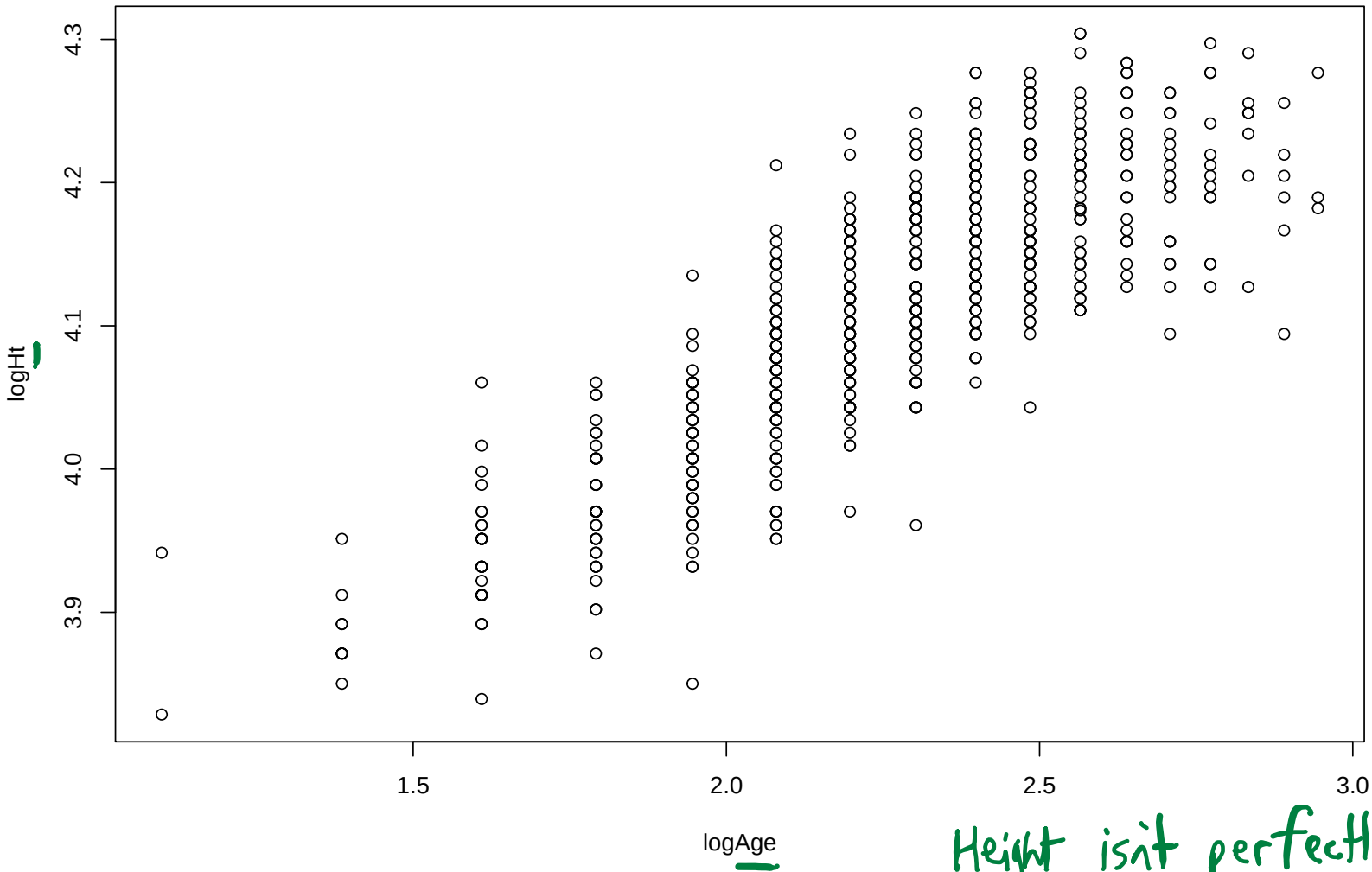
Here are the graphs again:





For this dataset, $\text{VIF}_1 = \text{VIF}_2 \approx 3.3$.

Scatterplot of the predictor variables



Height isn't perfectly predictable from age!
VIF is below the threshold

You aren't responsible for knowing the maths — just that this method exists!

Proposed solution 1 to multicollinearity: Ridge regression

When fitting regression models with serious multicollinearity, you may try ridge regression: $\hat{\beta} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$

$\lambda I =$

$$\begin{bmatrix} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \end{bmatrix}$$

- "Beefing up" the diagonal of the matrix being inverted makes the problem better-conditioned
- There exist principled ways to choose λ — See Kevin Murphy's 'Machine Learning' Chapter 7 — Regularization

Again you won't be tested on the maths. Just know this remedial measure exists.

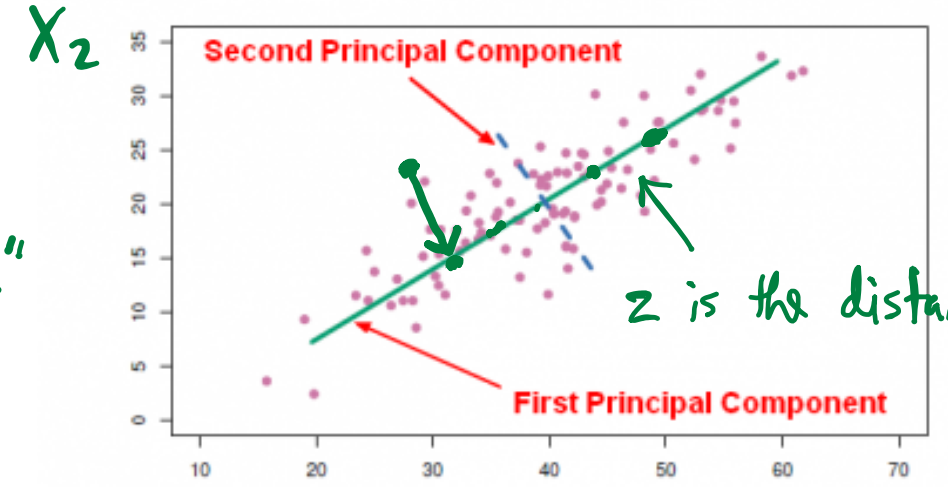
Proposed solution 2: PCR

In **principal component regression**, $\hat{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$

- $\mathbf{Z}$ is a lower-dimensional version of $\mathbf{X}$, e.g. $n \times 2$ instead of $n \times 3$
- Suppose $\mathbf{X}$ consists of two predictor variables that are noticeably correlated, and you run PCA (principal component analysis) to discover the direction of maximum variation (along the first principal component, a 2×1 vector):

$\underset{\sim}{X} = \begin{bmatrix} 1 & X_1 & X_2 \\ 1 & \vdots & \vdots \\ 1 & & \end{bmatrix}$

Example of feature "extraction"



orthogonal regression (week 2, slide 15)

z is the distance along the teal line

- Projecting the corresponding two n-columns of $\mathbf{X}$ onto the principal component results in a single n-column (distances along the teal line)
- This n-column goes into forming the matrix $\mathbf{Z}$ (dimension $n \times 2$ here)
- You then run SLR on $\mathbf{Z}$ which has just one hybridized predictor variable

The MLR approach in full

1. Plot *standardized residuals* to help determine whether the proposed regression model is a valid model

2. Identify any *leverage points*

3. Identify any *outliers*

4. Assess the effect of each X on Y

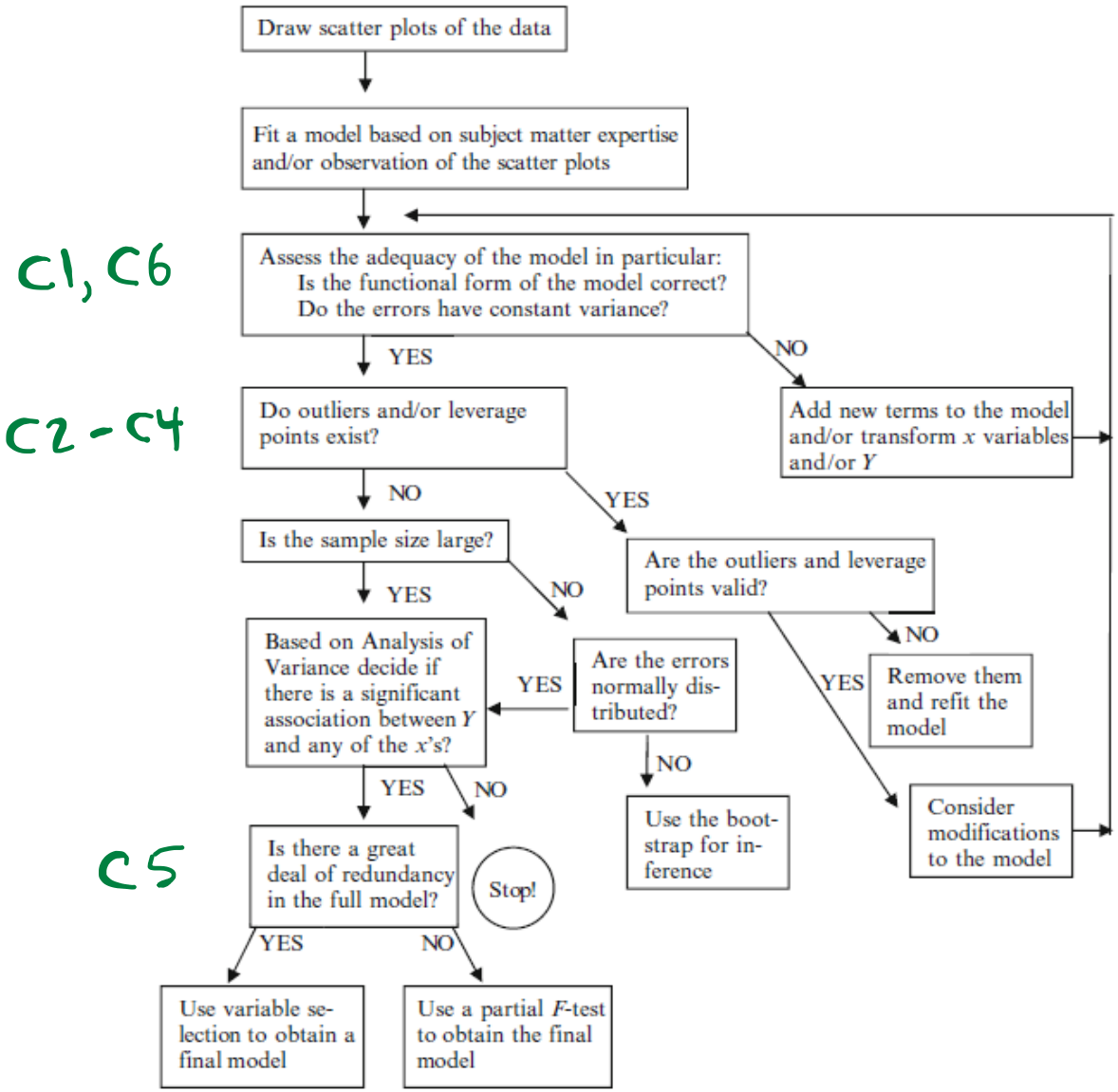
5. Assess *multicollinearity*

6. Assess the assumption of *error homoscedasticity*

7. For time series: examine whether the data are *correlated over time*

The Seven C's can also be viewed as fitting into a broader **flowchart approach to MLR** provided on the next slide.

The MLR flowchart, from p 252

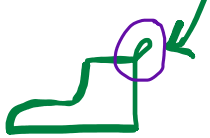


Variable Selection? Bootstrap?

Variable Selection is a topic for Chapter 7 next week.

As with SLR, the MLR flowchart mentions the "bootstrap", which you're not responsible for.

The bootstrap uses data resampling to numerically approximate the sampling distribution of the test statistic under H_0 (rather than using theoretical results, which we did assuming normally distributed errors).



eg $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$ may not be normally distributed

optional material →

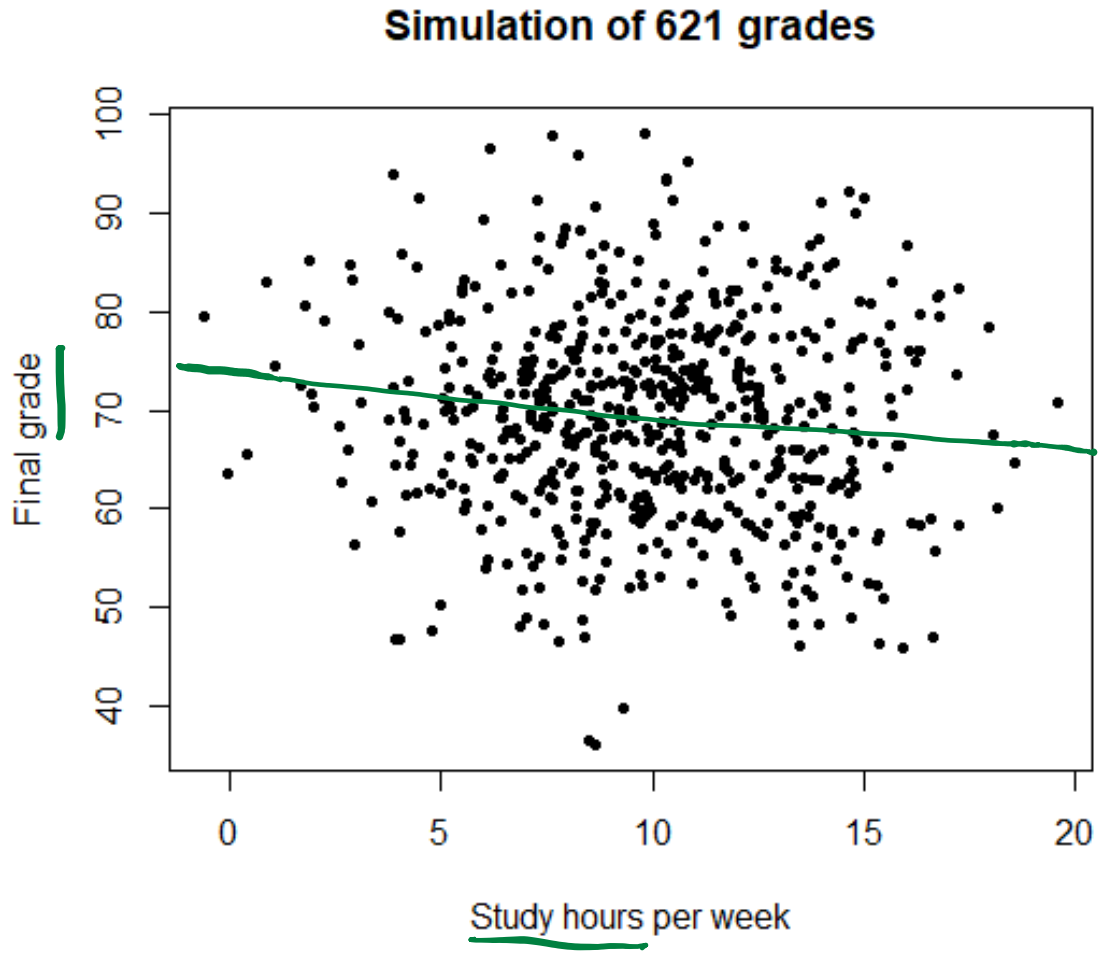
Example bootstrap: Calculate $\hat{\beta}_1$ 1000 times based on different samplings of the available data. The 0.025^{th} & 0.975^{th} quantiles of the 1000 $\hat{\beta}_1$'s can function as 95% CI.

Overview

Chapter 6:

- MLR Diagnostics: the Seven C's
- **A deeper look at Added Variable Plots**
- Introduction to a dataset to cut your teeth on: house prices

Detailed example of partial correlation



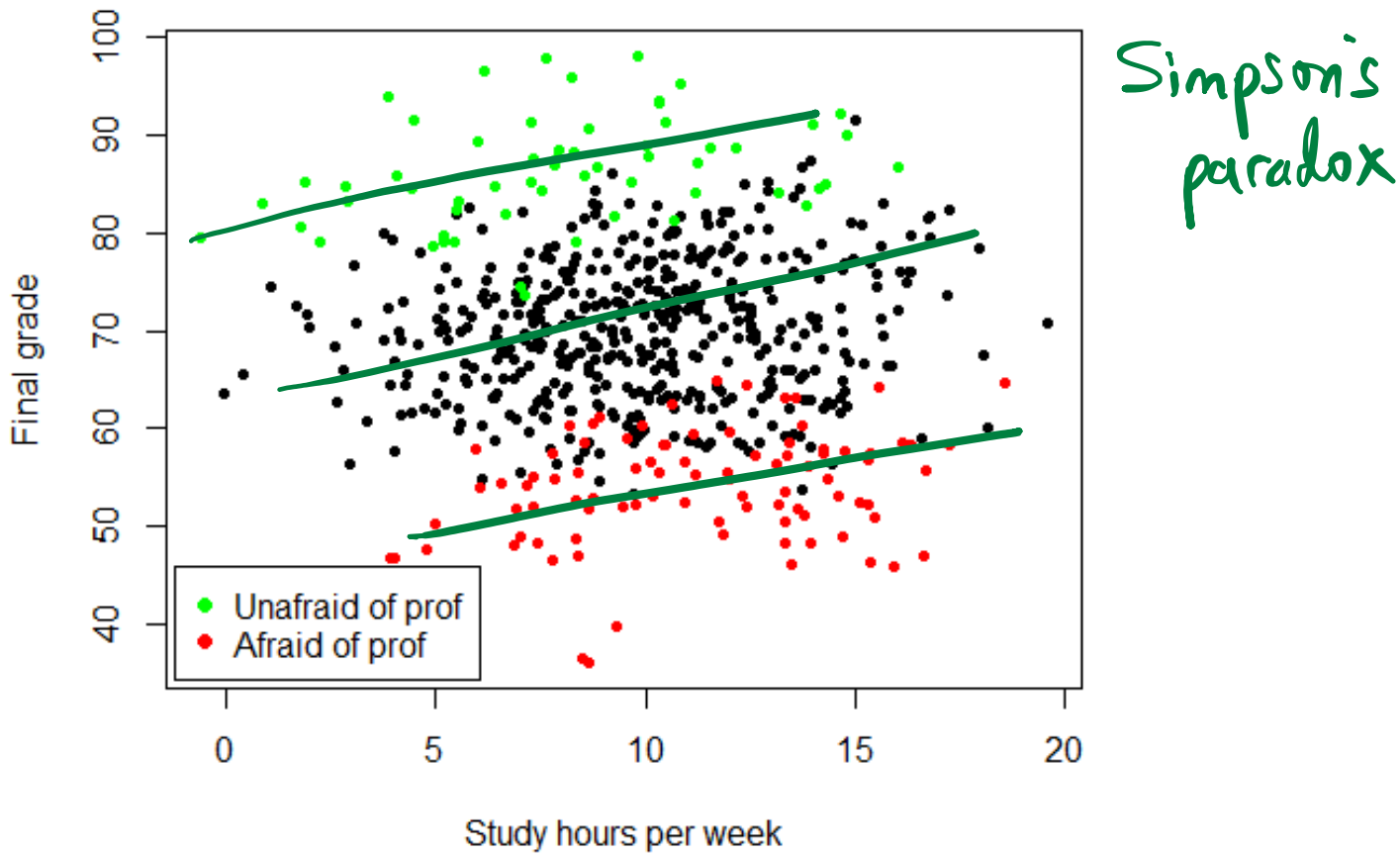
It may appear that studying more doesn't improve your final grade:

- No pattern is distinguishable
- $r \approx -0.06$, with a p-value of ~ 0.14

Detailed example of partial correlation

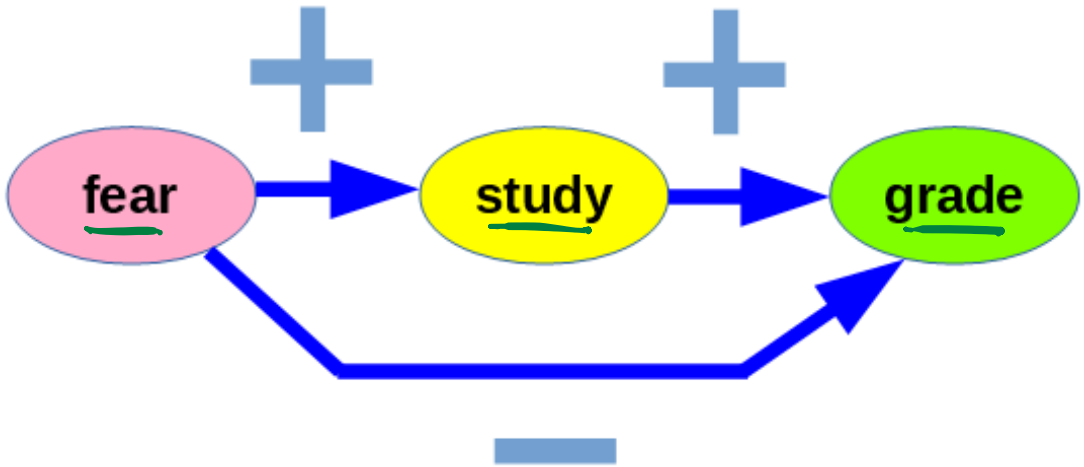
Students answer a question from 0 to 10: "Were you afraid of the prof?"

- Green points: Score of 2.5 or lower
- Red points: Score of 6 or higher



What happened?

As a graphical model (which you aren't responsible for in this course):



Partial correlations:

- Between Fear and Study: Significantly positive
- Between Fear and Grade: Significantly negative
- Between Study and Grade: Significantly positive

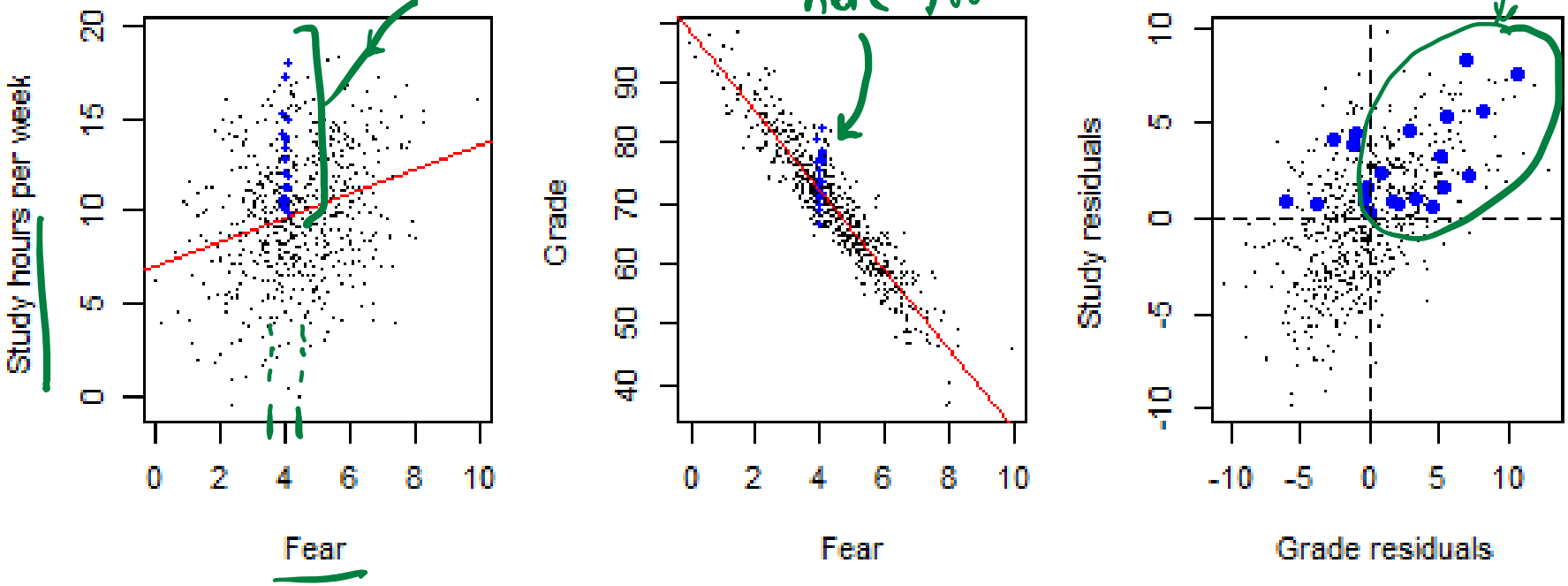
Regressing to produce the two sets of residuals

The partial correlation between `Study` and `Grade` given `Fear` is the r between:

- The residuals $\hat{e}_{\text{Study}}$ resulting from the regression of `Study` versus `Fear`
- The residuals $\hat{e}_{\text{Grade}}$ resulting from the regression of `Grade` versus `Fear`

busy quadrant

these points tend to be above the red line here too



MLR Check 4: Added variable plots

An **added variable plot** allows you to *visualize* the relationship between a response variable and an explanatory variable over and above the other explanatory variables.

Such plots are also known as *partial regression plots*, *adjusted variable plots*, or *partial residual plots*.

The technique is based on the concept of **partial correlation** from a few slides earlier. The partial correlation between X_1 and X_2 given a set of n other variables $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_n\}$ is the correlation between:

- The residuals $\hat{e}_{X_1}$ resulting from the linear regression of X_1 versus $\mathbf{Z}$
- The residuals $\hat{e}_{X_2}$ resulting from the linear regression of X_2 versus $\mathbf{Z}$

Added variable plots

For the jth added variable plot, we first divide X into X_j and the other X's (excluding X_j). Then we plot:

- x-axis: Residuals from the regression of X_j versus the other X's
- y-axis: Residuals from the regression of Y versus the other X's

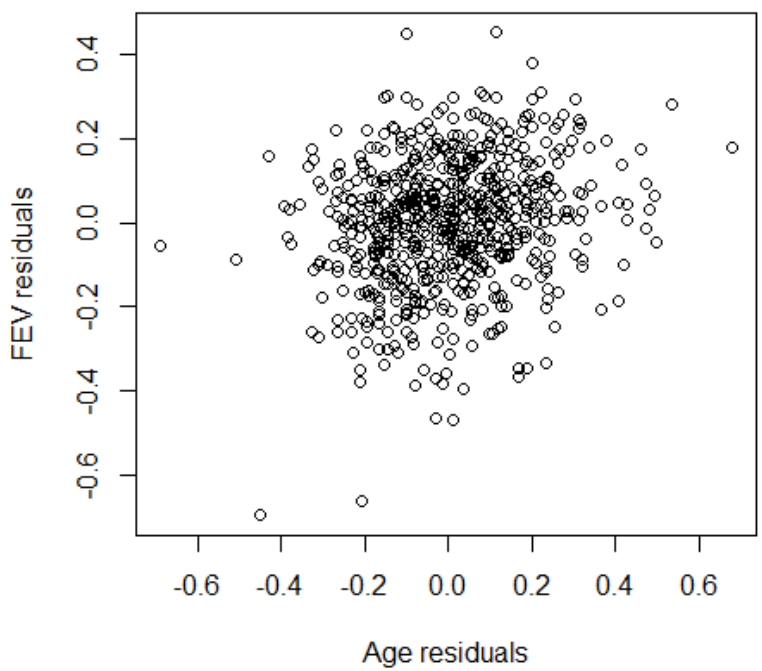
The correlation between the x- and y-axes here is a famous example of a partial correlation.

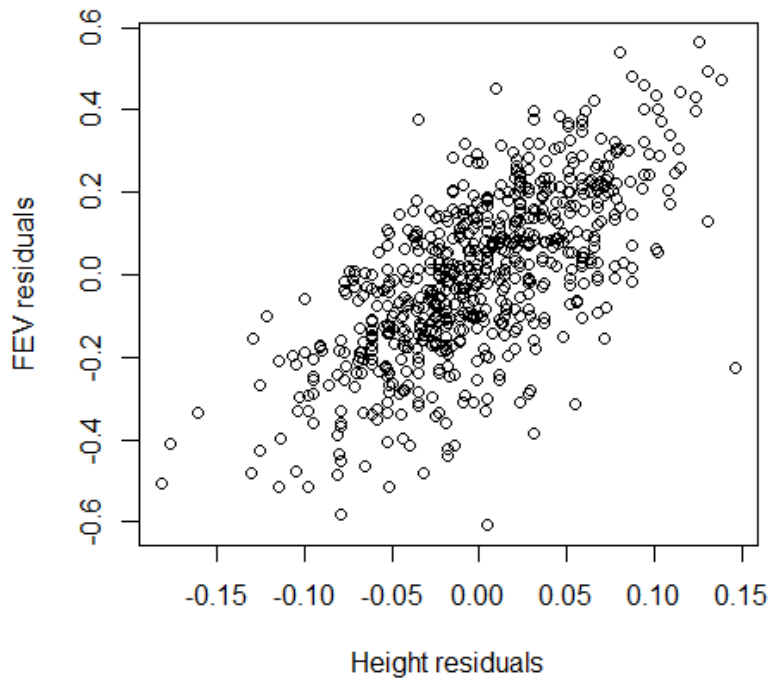
In an added variable plot, a linear pattern indicates that X_j is useful in the model over and above the other explanatory variables. In other words, the plot shows the strength of linear relationship Y and X_j over and above other variables.

The plot is also useful for detecting nonlinear relationships (polynomial in X_j for example), outliers, nonconstant variance, and influential points.

Example: FEV data

```
par(mfrow=c(1,2))
xAxis <- lm(logAge ~ logHt); yAxis <- lm(logFev ~ logHt)
plot(xAxis$residuals,yAxis$residuals,xlab="Age residuals",
     ylab="FEV residuals")
xAxis <- lm(logHt ~ logAge); yAxis <- lm(logFev ~ logAge)
plot(xAxis$residuals,yAxis$residuals,xlab="Height residuals",
     ylab="FEV residuals")
```





Overview

Chapter 6:

- MLR Diagnostics: the Seven C's
- A deeper look at Added Variable Plots
- **Introduction to a dataset to cut your teeth on**: house prices

Your mission: Explore a house-price dataset

For 26 houses sold in Chicago, a long time ago, we know the selling price Y as well as eight characteristics of each house: square footage, parking information, etc.

Reference: Ashish Sen and Muni Srivastava, *Regression Analysis: Theory, Methods and Applications*, 2013.

A peek at the house-price data

```
Q <- read.csv("houses.txt",sep=""); print(Q)
```

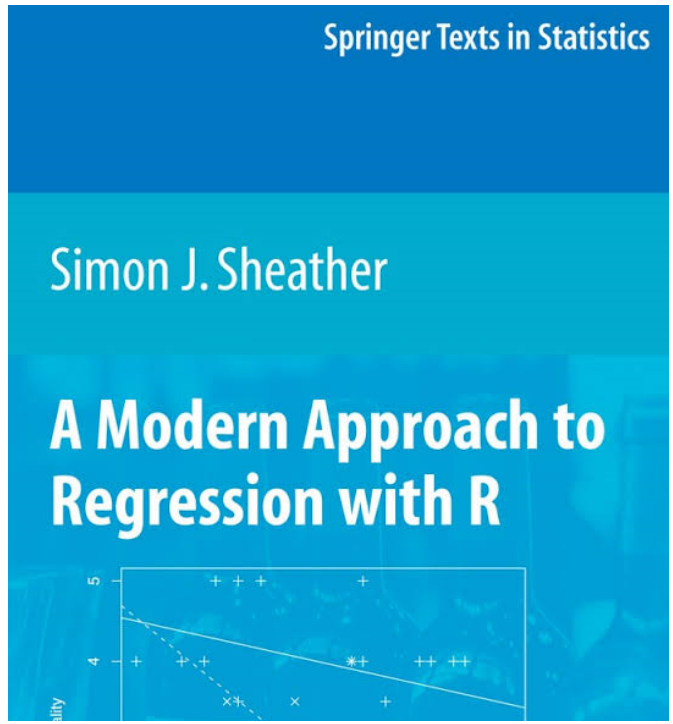
```
##      Y bdr  flr fp rms st lot bth gar
## 1   53   2  967  0   5  0  39 1.5 0.0
## 2   55   2  815  1   5  0  33 1.0 2.0
## 3   56   3  900  0   5  1  35 1.5 1.0
## 4   58   3 1007  0   6  1  24 1.5 2.0
## 5   64   3 1100  1   7  0  50 1.5 1.5
## 6   44   4  897  0   7  0  25 2.0 1.0
## 7   49   5 1400  0   8  0  30 1.0 1.0
## 8   70   3 2261  0   6  0  29 1.0 2.0
## 9   72   4 1290  0   8  1  33 1.5 1.5
## 10 82   4 2104  0   9  0  40 2.5 1.0
## 11 85   8 2240  1  12  1  50 3.0 2.0
## 12 45   2  641  0   5  0  25 1.0 0.0
## 13 47   3  862  0   6  0  25 1.0 0.0
## 14 49   4 1043  0   7  0  30 1.5 0.0
## 15 56   4 1325  0   8  0  50 1.5 0.0
## 16 60   2  782  0   5  1  25 1.0 0.0
## 17 62   3 1126  0   7  1  30 2.0 0.0
## 18 64   4 1226  0   8  0  37 2.0 2.0
## 19 66   2  929  1   5  0  30 1.0 1.0
## 20 35   4 1137  0   7  0  25 1.5 0.0
```

The response variable and eight predictors

Variable	Meaning
`Y`	Selling price in thousands of dollars
`bdr`	Number of bedrooms
`flr`	Floor space in square feet
`fp`	Number of fireplaces
`rms`	Number of rooms
`st`	Storm windows present (indicator variable)
`lot`	Frontage in feet
`bth`	Number of bathrooms
`gar`	Number of garage parking spaces

Next week, we'll use the techniques on these slides to analyse the dataset. For those wanting a head start, the data are available on Quercus (click).

Chapter 6 coverage



- We *don't* cover Marginal Model Plots, Inverse Response Plots, or Box-Cox transformations
- Use the lecture slides as a guide to Chapter 6 coverage. This means we cover more or less pp 151–4, 162–6, and 195–225, but omitting any content in §6.5 associated with the material we'd skipped earlier

Next steps

- Try all questions (1 to 5) in Chapter 6, using the techniques from our lecture slides. Solutions will be posted
- Continuing in week 12 with new slides, we'll cover some of Chapter 7, and (time permitting) Chapter 4 which is quite short
- The exam on 14 December will contain more than 20% pre-midterm content: questions which you could answer without having attended lectures after mid-October or having read beyond Chapter 3