STA302/STA1001, Weeks 11-12

Mark Ebden, 28–30 November

With grateful acknowledgment to Alison Gibbs

Overview

- Practical example: Analysing house-price data
- Chapter 7: Variable selection, and going with the flow
- Key MLR aspects

Exploring 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.

```
##     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
## 21 38   3  743  0   6  0  25 1.0 0.0
## 22 43   3  596  0   5  0  50 1.0 0.0
## 23 46   2  803  0   5  0  27 1.0 0.0
```

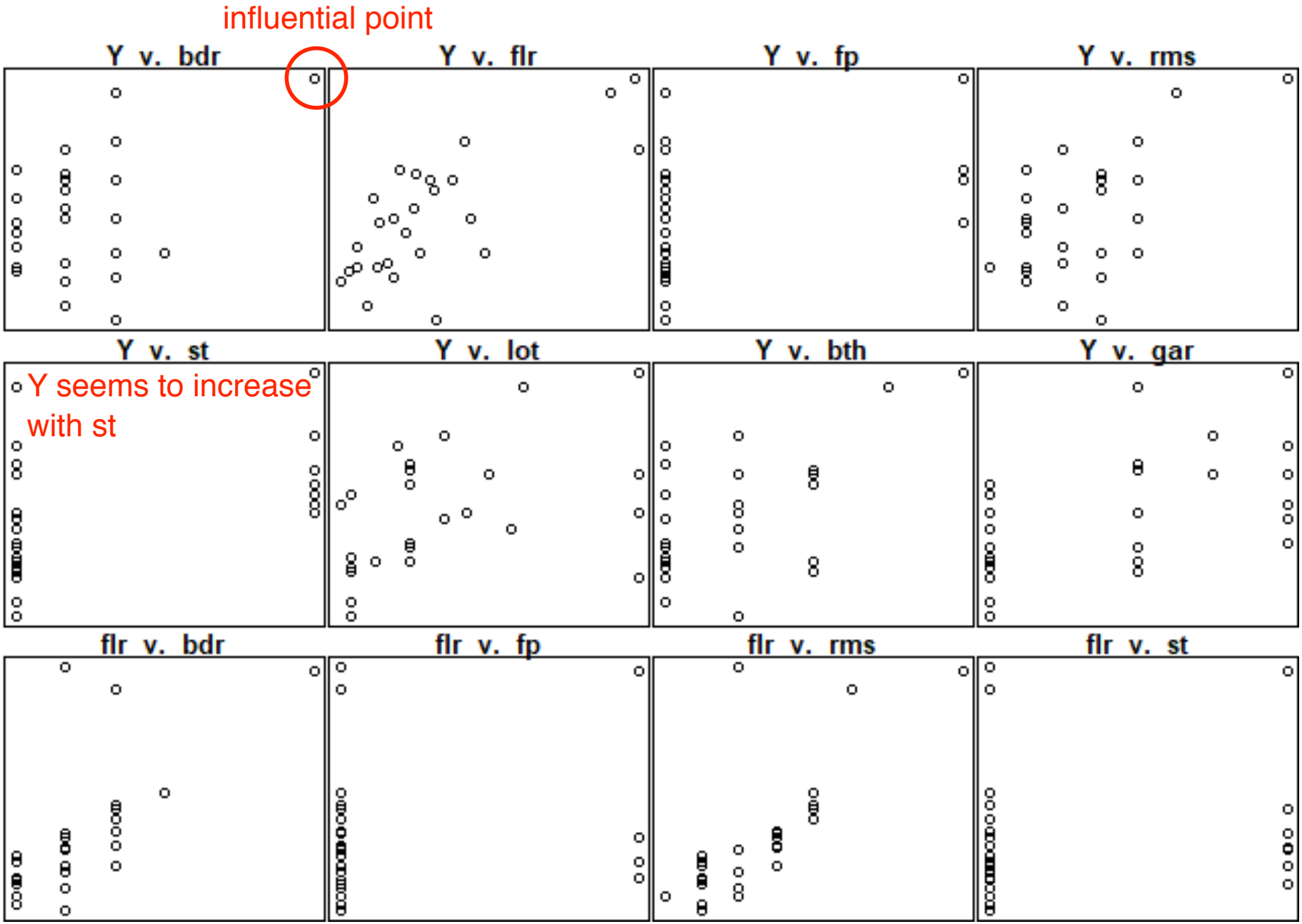
The response variable and eight predictors

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

We'll use the techniques on last week's slides to analyse the dataset. The data are available on Portal.

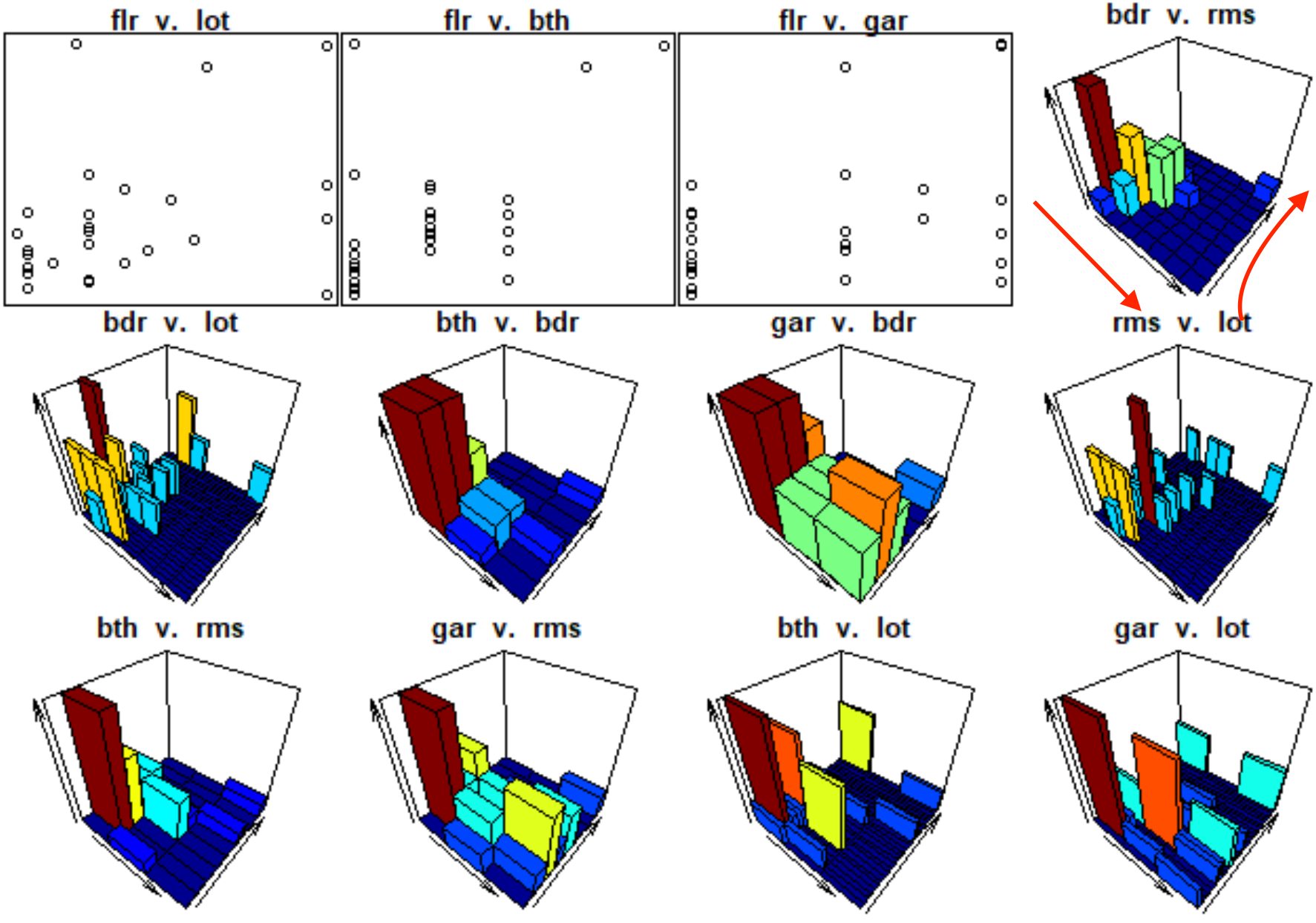
Pairwise scatter plots

9C2 possible pairwise combination

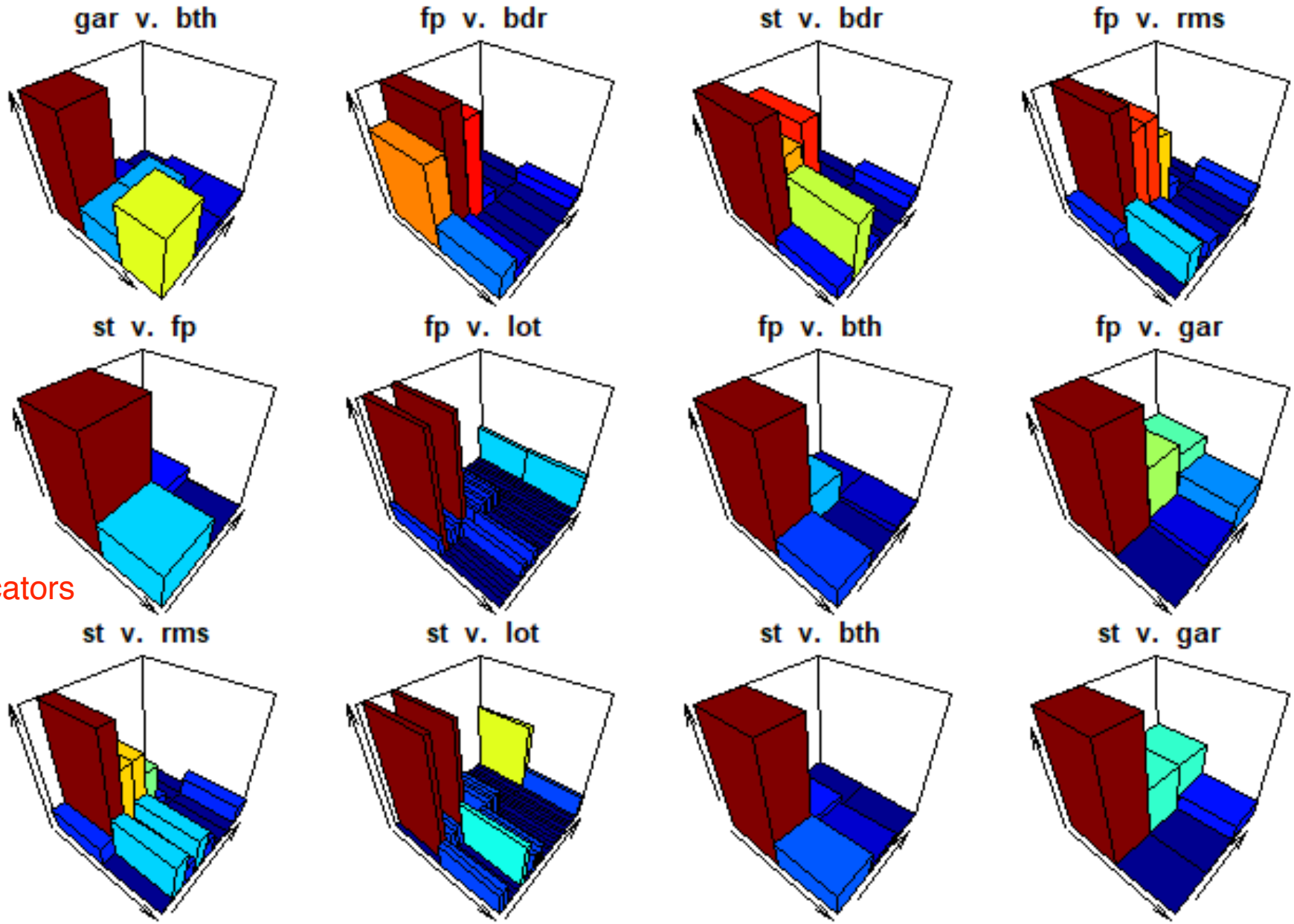


Pairwise 3D histograms

seems to have linear relation, through the diagonals



Pairwise 3D histograms



2 indicators

Initial findings from the scatter plots

- All X's seem to have some relationship with Y
- No obvious nonlinear relationship between Y and any of the predictors; no obvious need for transformations
- For later when we do the 7 C's: perhaps there exists an influential point; namely, the very first plot shows a house with eight bedrooms
- There are also relationships among some of the X's: for example, between `bdr` and `rms` (middle slide of the three: top right)
- But first, let's continue to explore the dataset systematically

Examining the correlations r and their p-values

mostly positive correlation

pairwise correlation

	Y	bdr	flr	fp	rms	st	lot	bth	gar
Y		0.41	0.74	0.39	0.58	0.46	0.42	0.55	0.54
bdr	1.41		0.68	0.17	0.92	0.23	0.40	0.63	0.24
flr	4.74	3.81		0.16	0.74	0.13	0.34	0.53	0.40
fp	1.32				0.19	-0.02	0.39	0.12	0.41
rms	2.73	10.37	4.81			0.23	0.43	0.69	0.30
st	1.72						-0.04	0.35	0.17
lot	1.50	1.39		1.34	1.55			0.37	0.14
bth	2.47	3.28	2.23		3.98				0.26
gar	2.39		1.40	1.44					

larger by common sense

blank = nonsignificant

lower p-value, larger -log(p-value)

Upper right: Pearson's r. **Lower left:** $-\log_{10}(p\text{-value})$, e.g. 2 means $p = 0.01$

?

if there is correlation or not

- Each correlation p-value is equal to an SLR β_1 t-test's
- We mainly consider the p-values and signs here, and needn't pay as much attention to the magnitude of each r
- All X's have a statistically significant r (or R^2) with Y
- Question: Have we run an appropriate test for st?

Performing MLR

The next slide describes the MLR. (What can we conclude from the F-test?)

The fitted equation is:

$$\hat{Y} = 18.6 - 7.70\ \text{bdr} + 0.0176\ \text{flr} + 6.91\ \text{fp} + 3.90\ \text{rms} + 10.8187\ \text{st} + 0.2635\ \text{lot} + 2.37\ \text{bth} + 1.77\ \text{gar}$$

Interpreting the coefficients:

- Recall that b_j represents the estimated change in mean selling price due to a unit change in X_j with all other variables held constant
- e.g. A 100-square-foot increase corresponds to a \$1760 increase in mean price, with everything else held constant
- A 1-bedroom increase corresponds to a \$7700 decrease in mean price (despite the fact that the correlation between price and `bdr` is positive)

doesnt make sense ... Y~bdr has significant (p-value=1.41) positive relation (0.4)

Despite the significant correlation between `Y` and `bth`, the t-test on the next slide gives a p-value of $0.366 > 0.05$. Why? (Hint: slide 26 from last week)

Y~bth seems to have positive linear relationship (p-value=2.47)
but bth correlated with some other predictors, and MLR favors more valuable predictors and leave less important predictor out

```
## Call:
## lm(formula = Q)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3058  -2.8417  -0.1511   3.2882   7.9518
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.637664   5.240957   3.556 0.002429 **
## bdr         -7.697444   1.829426  -4.208 0.000592 ***
## flr          0.017570   0.003235   5.431 4.49e-05 ***
## fp           6.909765   3.083583   2.241 0.038680 *
## rms          3.904374   1.615617   2.417 0.027194 *
## st          10.818663   2.300203   4.703 0.000205 ***
## lot          0.263522   0.135109   1.950 0.067808 .
## bth          2.374591   2.557865   0.928 0.366221
## gar          1.770861   1.404310   1.261 0.224334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.717 on 17 degrees of freedom
## Multiple R-squared:  0.9044, Adjusted R-squared:  0.8595
## F-statistic: 20.11 on 8 and 17 DF,  p-value: 3.147e-07
```

response variable is first column, predictor as other cols

The Seven C's (7 Checks) for STA302 MLR diagnostics

In their proper order this time:

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*

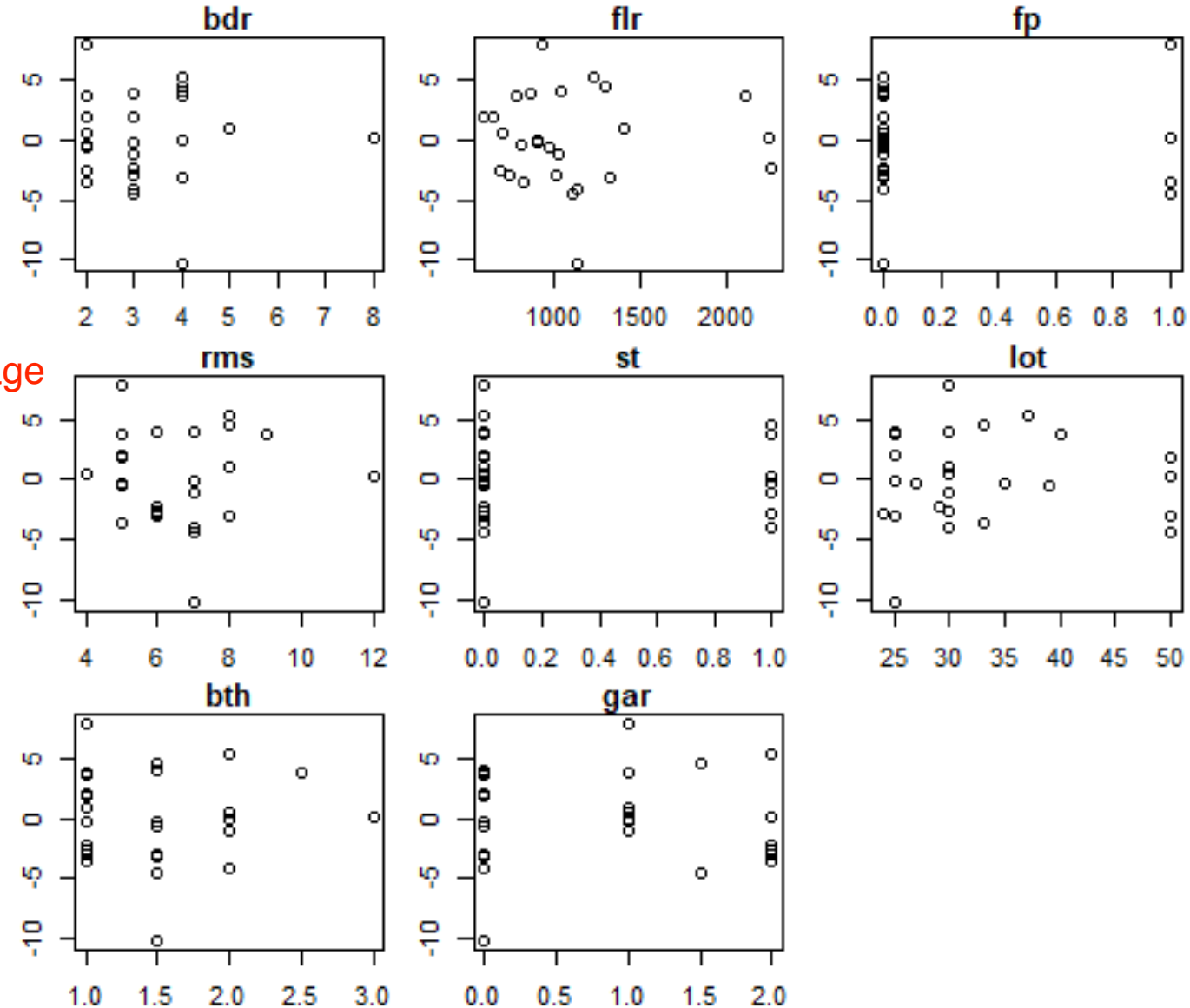
7. For time series: examine whether the data are *correlated over time*

Residual plots in the house-price example

residual vs predictor plots

Recall that plotting residuals versus X_j for $j \in \{1, \ldots p\}$ helps us to look for curvature, influential points, and outliers.

no obvious pattern here, there is some outliers/leverage pt though

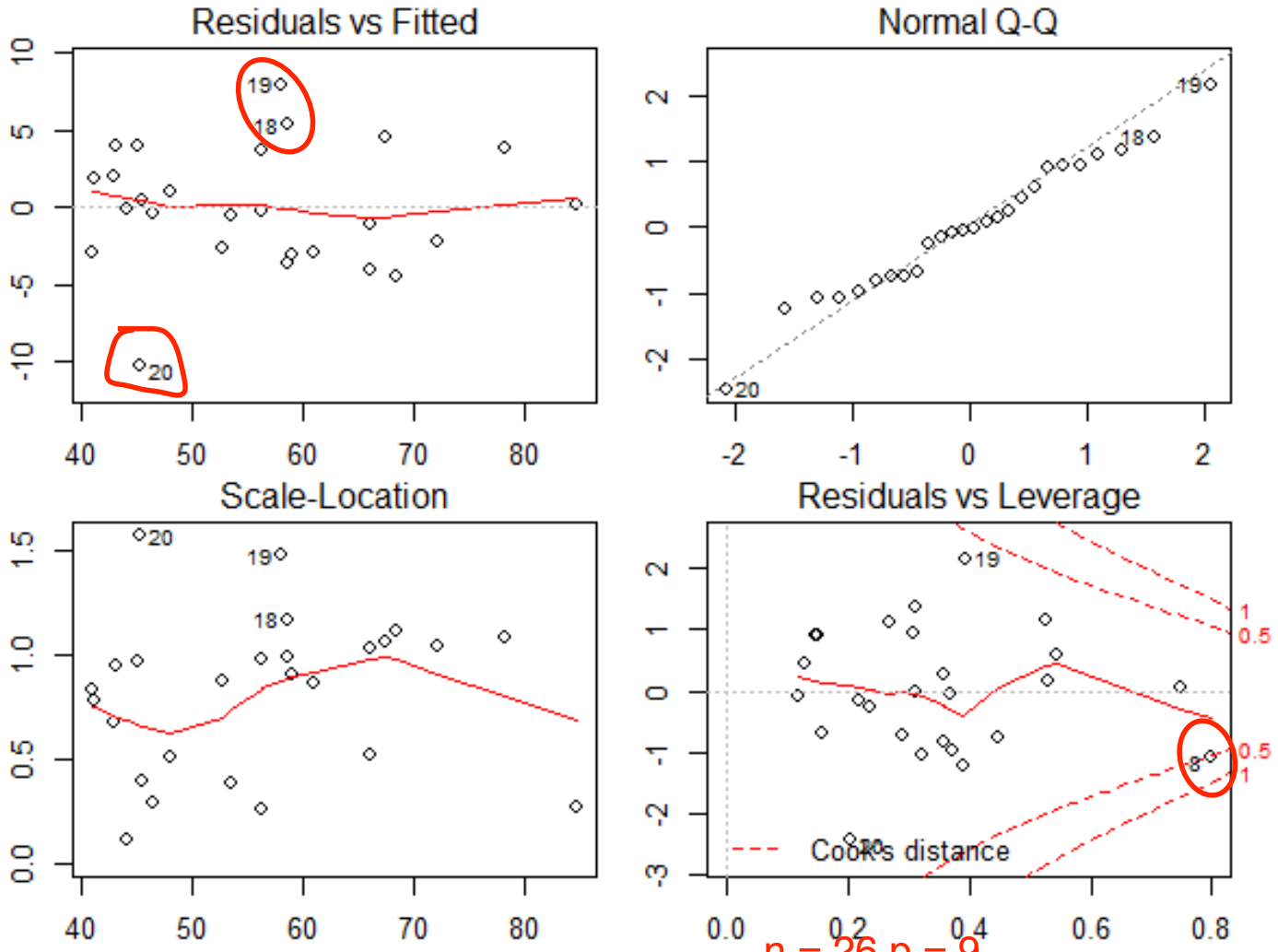


Output of `plot(lm(..))`

- One house with large negative residual (perhaps one large positive residual)
- At least one influential point according to the $4/(n-p-1)$ criterion: large Cook's distance in the bottom-right of the diagnostics panel

standardized residue here

outliers



influential

n = 26 p = 9

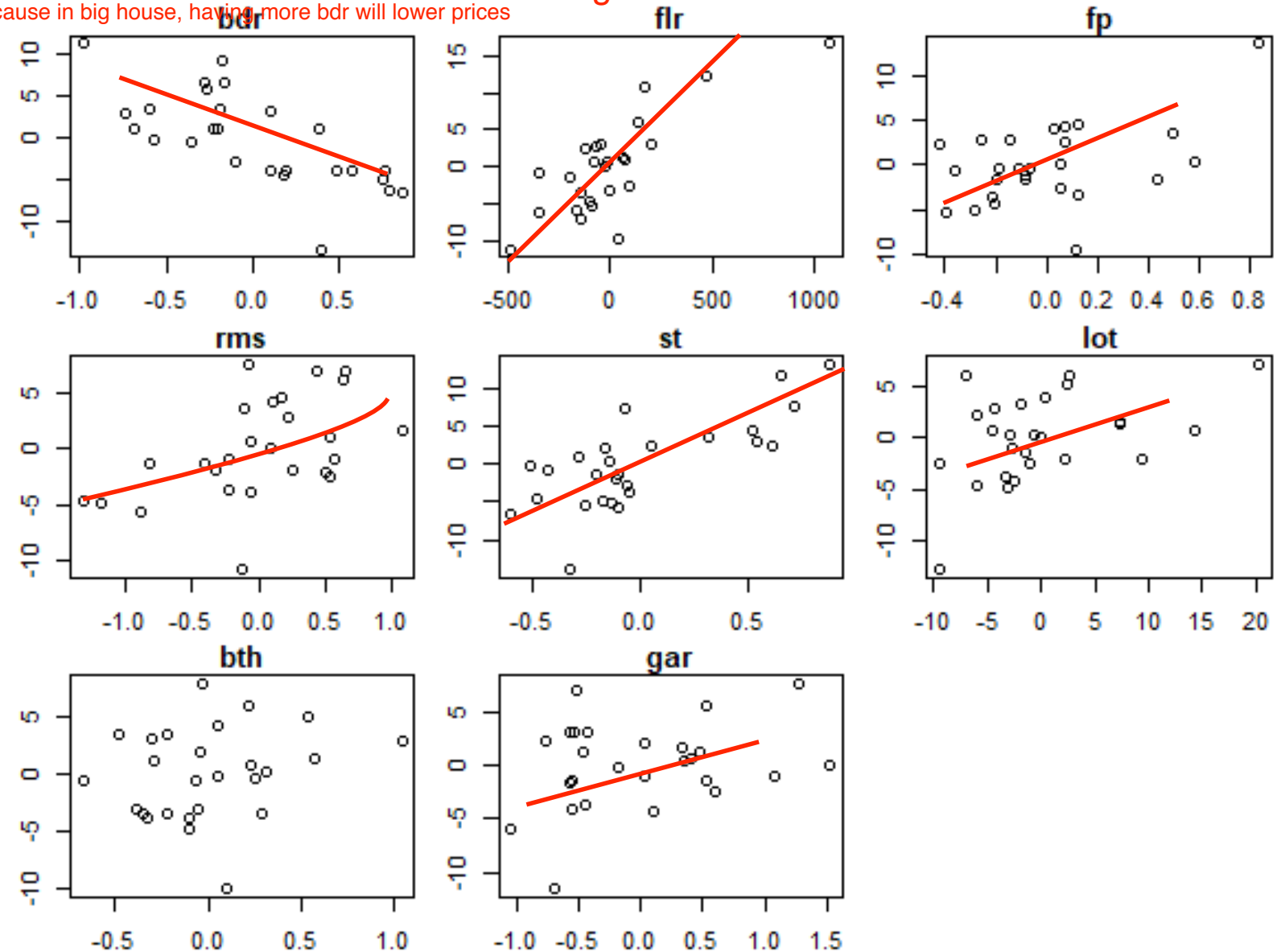
4 / (26–9-1) ~ 1/4

check for effect of each X on Y visually

Check 4 in the house-price example: Added variable plots

negative partial correlation
Y go up as bdr go up,
but thats because in big house, having more bdr will lower prices

flr adding value to the model...



probably not adding value to model

Assessing assumptions in the House-price example

We view the added variable plots in context. Consider the plot of Y versus `bdr`:

- Not a strong linear relationship; otherwise, no concerns with non-linearity
- Potential influential point: one house has 8 bedrooms seems ok
- SLR may not be a good idea for the price - bedroom relationship but may be ok in multiple regression probably not a good idea for SLR

Consider the added variable plot:

- The linear relationship is stronger (although it's now negative)
- We want to check if the influential is here. However, the large negative residual is not the house with 8 bedrooms

Summary of concerns:

- Large negative residual in the added variable plots
- Potential influential point in the added variable plot for `flr`

Summary of unusual observations

House 8:

- Influential
- Largest Cook's distance
- Very large floor space, everything else including price (Y) was of moderate space

House 19:

- Large positive residual
- Small house but middle-ish price; nice house?

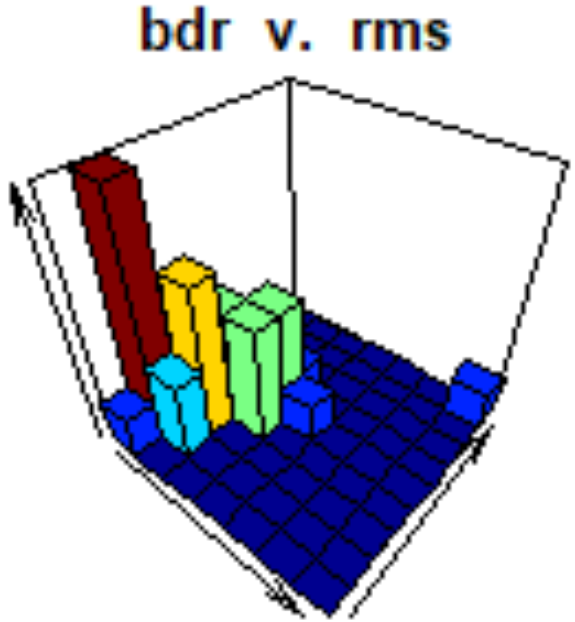
House 20:

- Large negative residual
- Smallest price, but all features middle-ish

House 11 *isn't* a concern:

- It has eight bedrooms but is neither an outlier nor influential
- It's expensive, but everything is big

Check 5 in the house-price example: Multicollinearity variance inflation factor



- $\text{VIF}_{\text{rms}} = 8.5$, $\text{VIF}_{\text{bdr}} = 6.5$, $\text{VIF}_{\text{flr}} = 2.5$, $\text{VIF}_{\text{bth}} = 2.1$
- Indeed we saw what might have been the effects of multicollinearity: the coefficient for `bdr` wasn't of the expected sign
- Details of executing the two remedial measures introduced last week (ridge regression, principal component regression) are beyond STA302's scope
- For now, awareness of possible multicollinearity informs our choice of how to **simplify the model**

Model simplification

Assumption is Y can be modeled with 8 predictors,

- If you remove variables during MLR, the β_j values of the remaining predictors become biased, and the p-values from t- and F- statistics generally shrink below their true values
- Therefore, when reducing the dimensionality of your MLR problem, take into account your reduction when reporting CIs and other results

keeping more predictor tend to keep p-value high,

Model simplification

method 1 conservative, more predictor leads to problem of overfitting.

Method 1: If there is multicollinearity in your data set, perform **variable selection**. e.g. choose the subset of predictors that maximizes R^2_{adj}.

Method 2: If there isn't multicollinearity, you can conduct a **partial F-test** on predictors you suspect of being unhelpful. Remove them if the result is non-significant.

method 2 more aggressive

Applying the two methods to the house-price data:

1. The subset turns out to be everything but `bth`. Therefore, `bth` would removed by Method 1.
   - Among the choices of removing `gar` and/or `bth`, the effect on R^2_{adj} is small in all cases (well under 1% change). Method 1 naively maximizes R^2_{adj} without regard to this; the method is known for its timidity in removing predictors.

2. Removing `bth` and `gar` leads to a non-significant partial F-test ($p > 0.05$). Therefore, `bth` and `gar` would be removed by Method 2. Removing any other predictor(s) instead of or in addition to these leads to a significant p-value for the partial F-test.

The MLR approach

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*

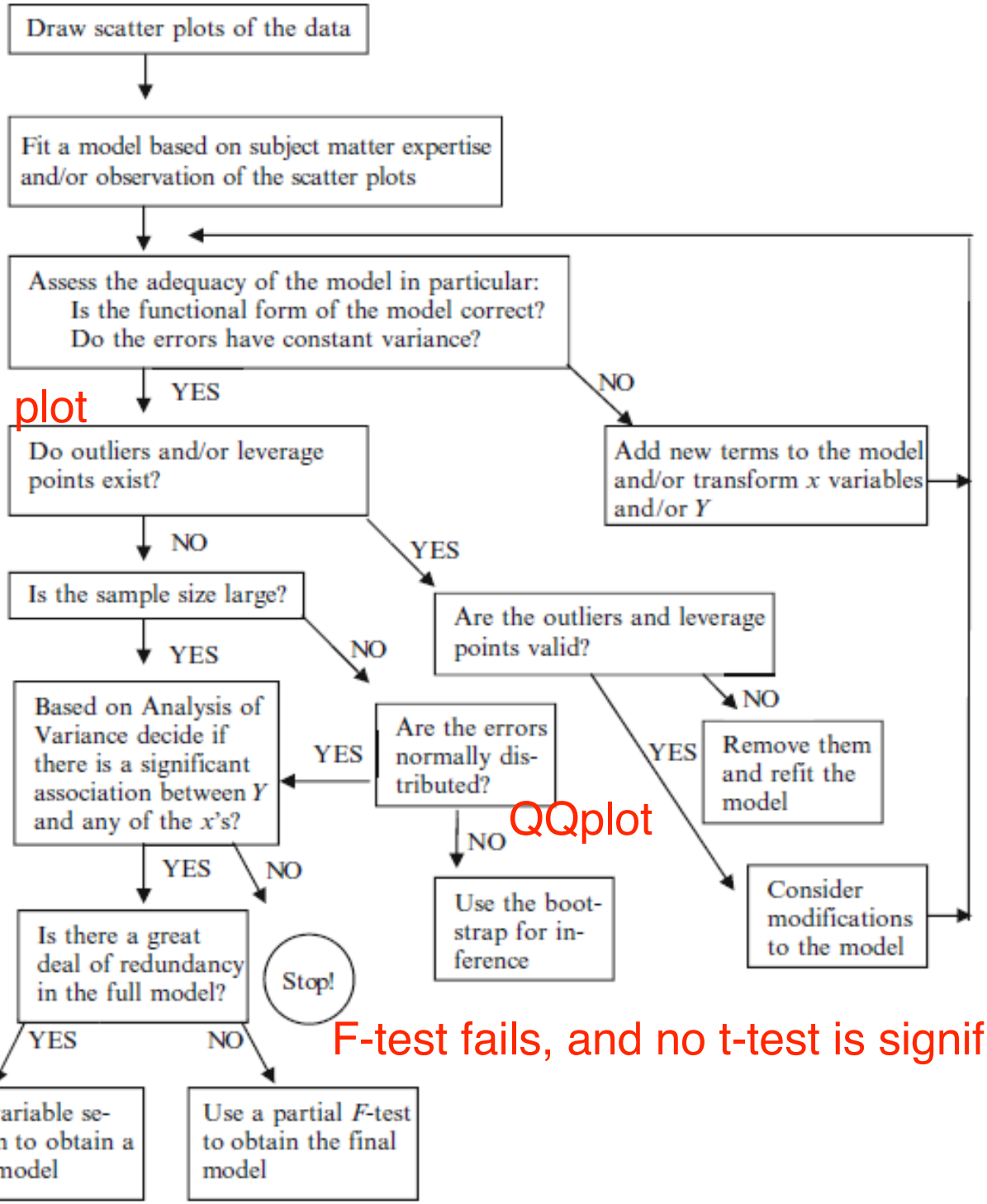
6. Assess the assumption of *error homoscedasticity*

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

cant really do check 7, since data is a slice in time

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

- The 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)


Draw scatter plots of the data
Fit a model based on subject matter expertise and/or observation of the scatter plots
Assess the adequacy of the model in particular:
Is the functional form of the model correct?
Do the errors have constant variance?
YES
NO
Add new terms to the model and/or transform x variables and/or Y
Do outliers and/or leverage points exist?
NO
YES
Is the sample size large?
YES
NO
Are the outliers and leverage points valid?
NO
YES
Remove them and refit the model
Based on Analysis of Variance decide if there is a significant association between Y and any of the x's?
YES
Are the errors normally distributed?
NO
Use the bootstrap for inference
Consider modifications to the model
YES
NO
Is there a great deal of redundancy in the full model?
Stop!
YES
NO
Use variable selection to obtain a final model
Use a partial F-test to obtain the final model
residual plot
F-test
t-test
QQplot
multicolinearity
minimize r^2
F-test fails, and no t-test is significant

Overview

- Practical example: Analysing house-price data
- Chapter 7: Variable selection, and going with the flow
- **Key MLR aspects**

Why do regression? Two main reasons

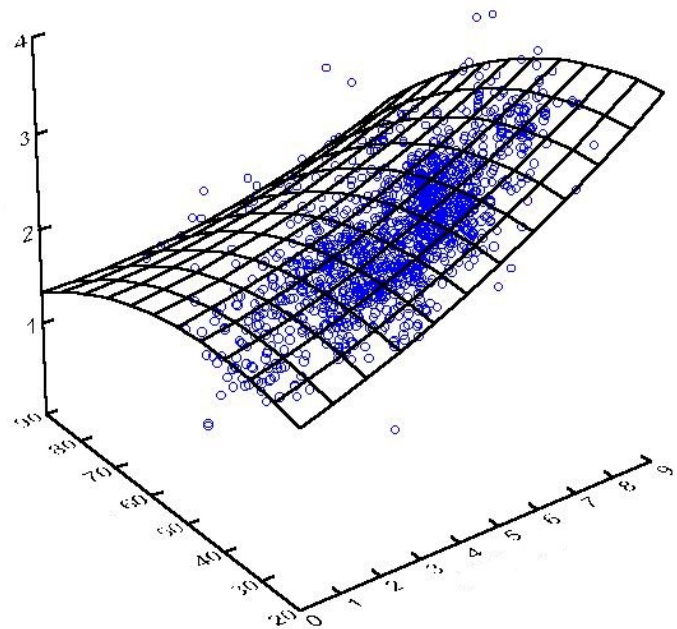
1. Predict the value of a variable given the value(s) of 1+ others
- For interpolation or extrapolation, for example in house prices
- It doesn't matter which variables are in the model or how they were chosen (model building/variable selection OK)
- However, we must consider the level of smoothing that's appropriate. If we overfit, predictions may not work well on new data

2. Understanding an underlying mechanism
- Describe the relationships among variables
- What are the effects of predictor(s) on the response?
- Describe differences in relationships between different sets of data
- For example, data sets for the Montreal Protocol (CFC data), Meadowfoam, etc
- This is *inferential modelling*. The form of the model is assumed to be known but there is uncertainty about the value(s) of the coefficients

It can be risky doing inferential modelling on the *same* data that you used to *build* a model (to describe how Y might originate in general).

Why do *multiple* regression?

- Multiple X's give a better prediction of Y, e.g. house prices
- MLR allows polynomial fits
- MLR allows comparison of regression line for two groups, ancova
  e.g. meadowfoam data
- You can control for some X's, e.g. a baseline value in a study, such as mammals and brain size

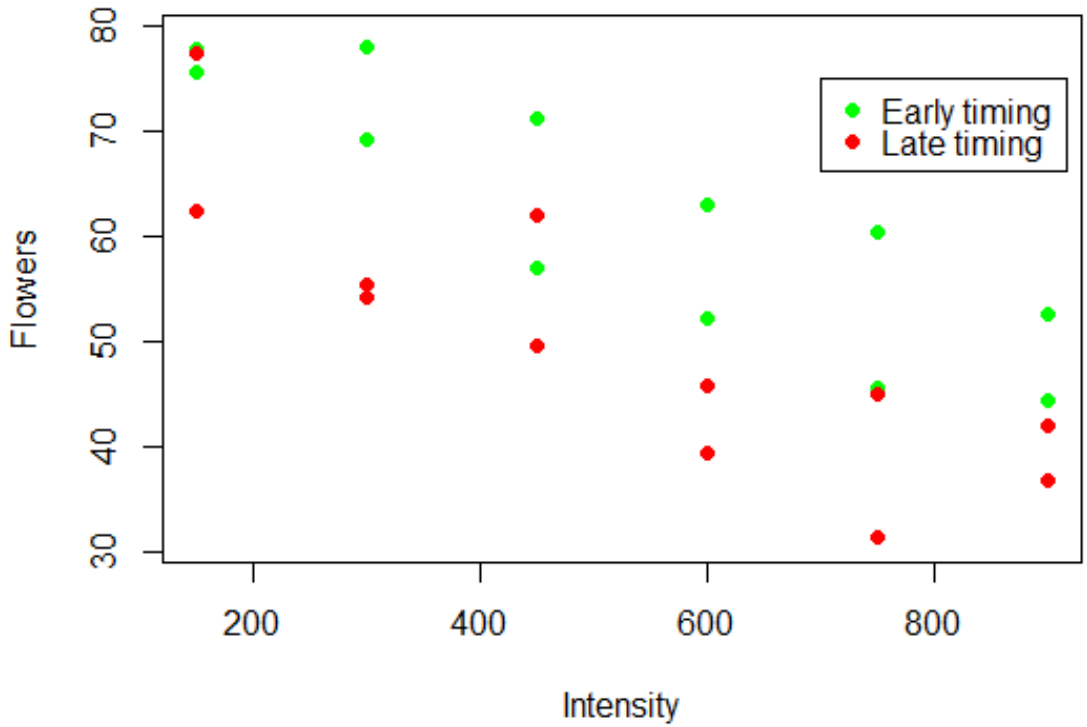


whats the regression equation like

Experiments vs Observational Studies

Experiment (e.g. for the meadowfoam dataset):
A treatment is imposed on experimental units. Advantages:

- Strength of conclusion
- If properly randomized, you can make cause-and-effect conclusions
- e.g. you can say that increasing light intensity causes a decrease in the number of flowers per plant on average

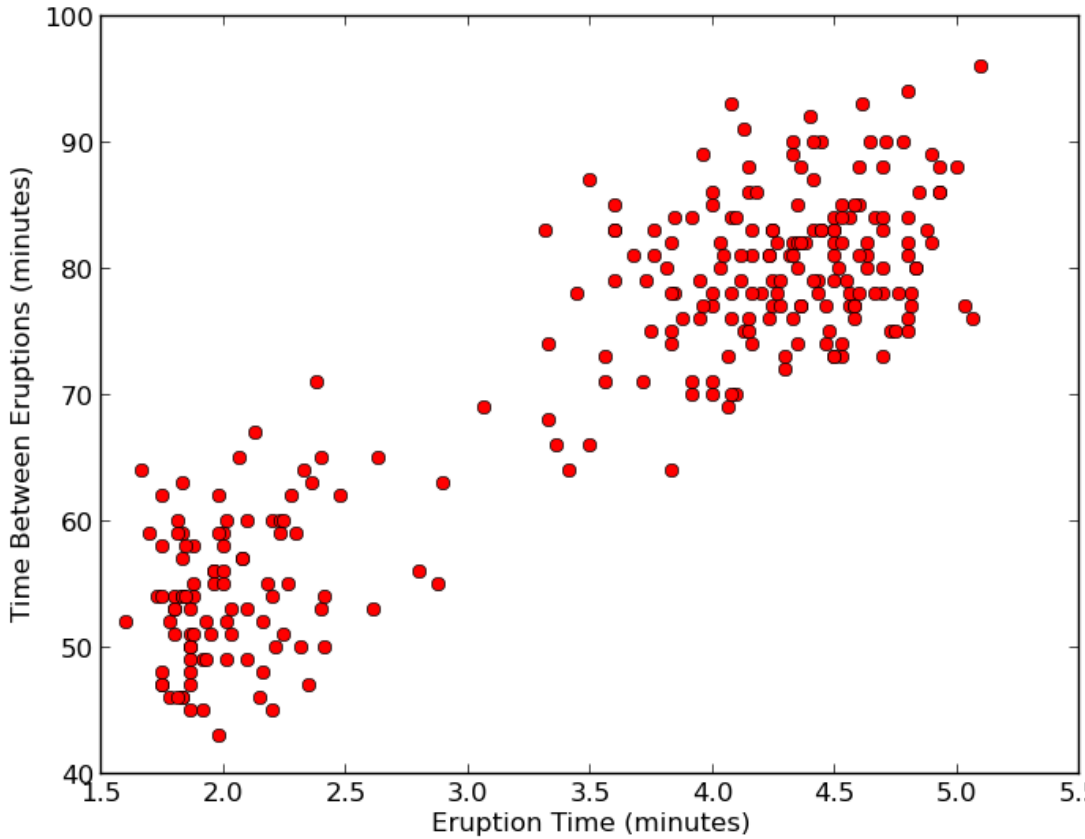


Experiments vs Observational Studies

Observational study (e.g. for the CFC data or house-price data):

Data are measured without an intervention. associations only

- You can say that there's an association, but you can't say that X causes Y
- Perhaps they are both related to a third variable
- e.g. Old Faithful data: you can't say that a longer duration causes a large interval: perhaps both are related to some other geological phenomenon



Multiple Regression

Comparison of:

- t-tests ($H_0 : \beta_j = 0$, all other X's in the model)
- The ANOVA F-test ($H_0 : \beta_1 = \cdots = \beta_p = 0$, a test for the overall model)

It's possible to have:

F-test significant ($p < 0.05$), and all or some t-tests significant:

- *The model has useful variables for explaining Y*

F-test non-significant, and all t-tests non-significant:

- *There are no explanatory variables useful for explaining Y*

strength in numbers…

F-test significant, and all t-tests non-significant:

- *Multicollinearity; individual X's don't predict Y over and above the others*

F-test non-significant, and some t-tests significant:

- *There are two possible ways this can happen (next slide)*

Significant t-test results accompanying a non-significant F-test result

t-test falsely significant due to type 1 error

1. The model has no predictive value, but in your t-test analysis there were some Type I errors (falsely rejecting H_0). This is always a risk with multiple tests

too many predictors diluted effect of 1 useful predictor

2. The predictors were chosen poorly. e.g. if there was one useful predictor among many not useful predictors, the contribution from the useful one may not be enough for the F-test to be significant

Problems with Regression Models and their Consequences

You might have the wrong model for several reasons:

1. **The true relationship is nonlinear:**

- Biased $\hat{\beta}$'s
- Meaningless t- and F-tests / CIs

2. **You omitted an important predictor variable:**

- Biased $\hat{\beta}$'s
- Meaningless t- and F-tests / CIs

$$\text{Actual model: } Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

$$\text{Fitted model: } \hat{Y} = b_0 + b_1 x_1 \quad \text{and thus } \mathrm{E}(b_1) \neq \beta_1$$

3. **You included an unimportant predictor variable:**

- $\hat{\beta}$'s aren't biased
- The variance is unnecessarily large (inefficient)

add more noise

Problems with Regression Models and their Consequences

The Gauss-Markov conditions might not be met:

1. **Non-constant variance:**

- $\hat{\beta}$'s aren't biased
- S^2 is a biased estimate
- Invalid t- and F-tests / CIs

2. **Correlated errors:**

- $\hat{\beta}$'s aren't biased
- S^2 isn't estimating the right thing
- Invalid t- and F-tests / CIs

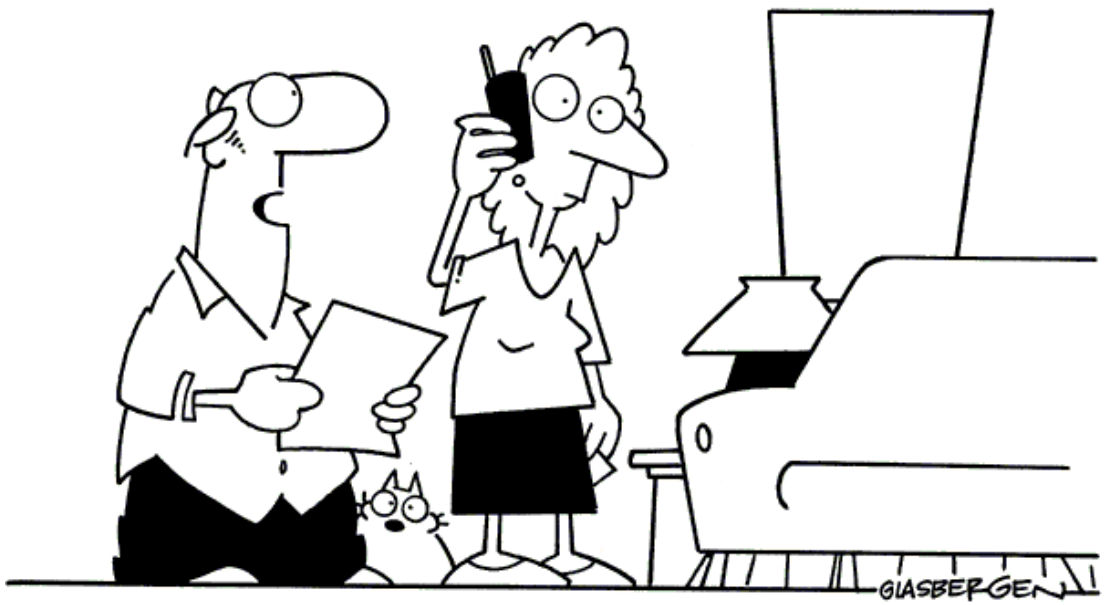
3. **Non-normal errors:**

- $\hat{\beta}$'s aren't biased
- S^2 is an unbiased estimate for σ^2
- Invalid t- and F-tests / CIs, except for large sample sizes
- Invalid PIs

Other problems with Regression Models and their Consequences

1. **Multicollinearity:**
   - var($\hat{\beta}$) will be high — inefficient

2. **The predictor variables were measured with error:**
   - Biased $\hat{\beta}$ and S^2 Sxx inflated
   - Invalid t- and F-tests / CIs
   - Generally ok provided the variance in X's is small
   - To fix: take multiple measurements, or use Structural Equation Modelling



"Ask the realtor if we can list the litter box as a third bathroom."

Exam results by Section

After the exams are marked, I'll check for any significant differences between the grades of Sections 1 and 2. If one exists then the marks may be tweaked to eliminate this.

If you're in Section 1 but have been attending Section 2's lectures and wish to be viewed within Section 2 for the purposes of exam marks adjustment (or vice versa), please let me know directly by Friday 8 December and give your reasons.

The expectation is that Sections 1 and 2 will perform similarly on the shared exam, having been taught the same. Therefore, hopefully this slide is pointless!

Course Evaluations

FAS Fall 2017 Undergrad for Meth Data Analysis STA302H1-F-LEC0101

Medium	Online
Timing	Scheduled

- Start Date 2017-11-23 00:00
- End Date 2017-12-08 23:59

Response Rate

	Responded	Invited	% Rate
Students	39	341	11.44%

Please visit https://courseevaluations.utoronto.ca

Next steps

- Solutions to Chapter 6 have been posted on Portal under 'Homework'
- Chapter 7: You may wish to attempt the R^2_{adj} portion of 1(a), 2(a), or 3(a), but not all three
- No more homework!