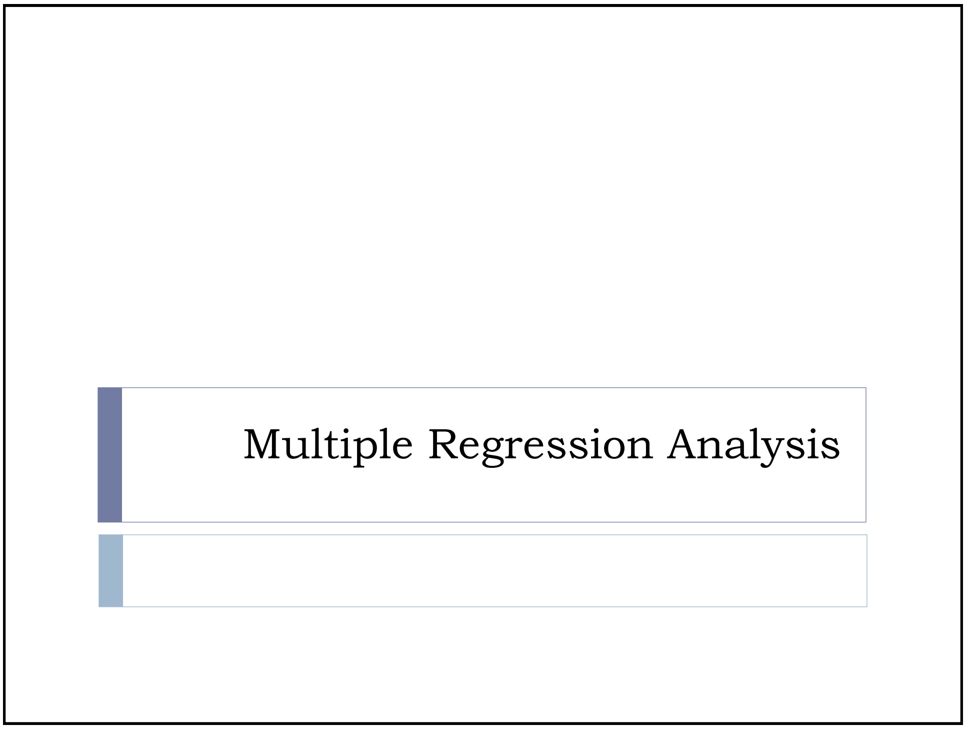
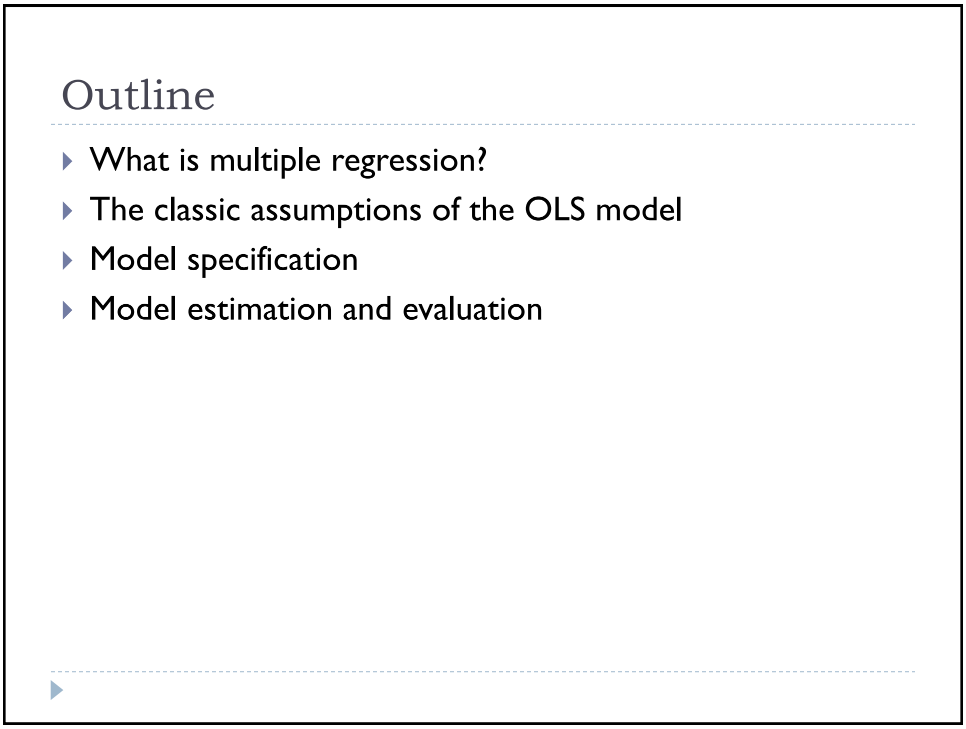


Multiple Regression Analysis



Outline

- What is multiple regression?
- The classic assumptions of the OLS model
- Model specification
- Model estimation and evaluation

What is multiple regression?

- Where as simple linear regression has 2 variable (1 dependent, 1 independent):

$$\hat{y} = a + bx$$

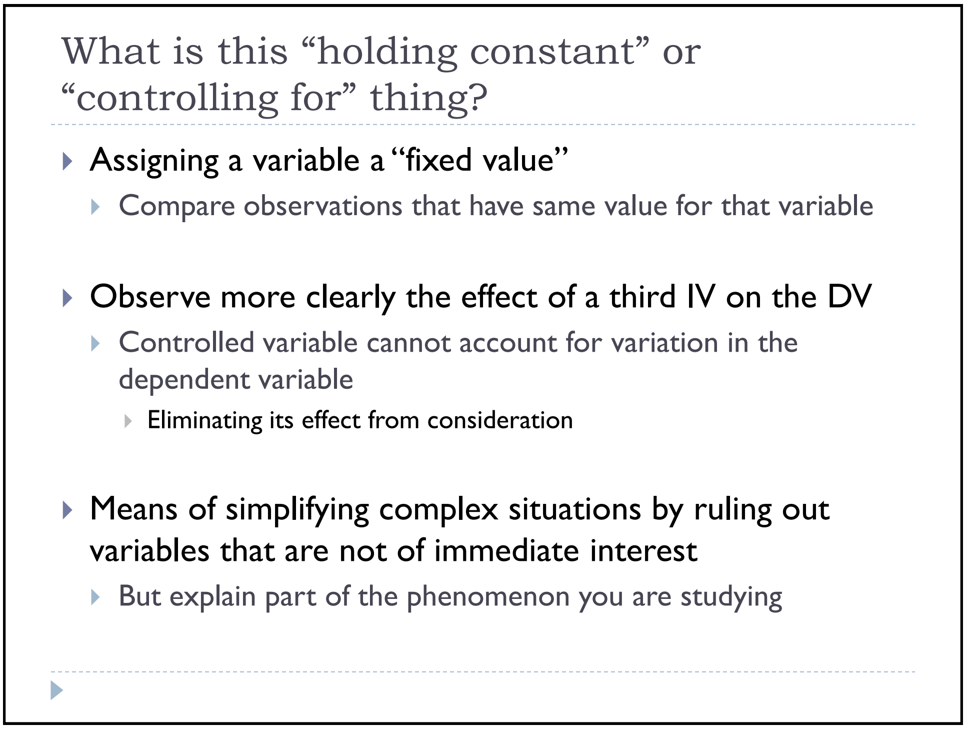
- Multiple linear regression has >2 variables (1 dependent, many independent):

$$\hat{y} = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

- The problems and solutions are the same as bivariate regression, except there are more parameters to estimate
  - Estimate a slope (*b*) for each variable

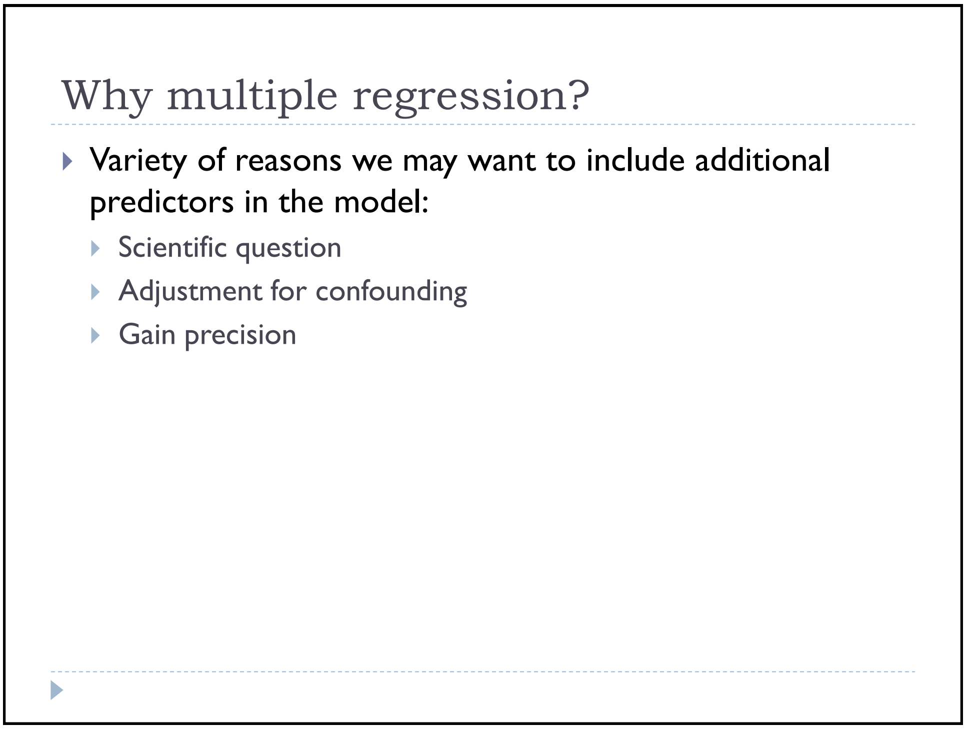
Parameters

- Slope (β_n)
  - Estimated change in y (DV) for a 1 unit increase in x_n **holding all other variables constant**
- y-intercept (β_0)
  - Average value of *y* when x_n = 0



What is this "holding constant" or "controlling for" thing?

- Assigning a variable a "fixed value"
  - Compare observations that have same value for that variable
- Observe more clearly the effect of a third IV on the DV
  - Controlled variable cannot account for variation in the dependent variable
    - Eliminating its effect from consideration
- Means of simplifying complex situations by ruling out variables that are not of immediate interest
  - But explain part of the phenomenon you are studying



Why multiple regression?

- Variety of reasons we may want to include additional predictors in the model:
  - Scientific question
  - Adjustment for confounding
  - Gain precision

Study question

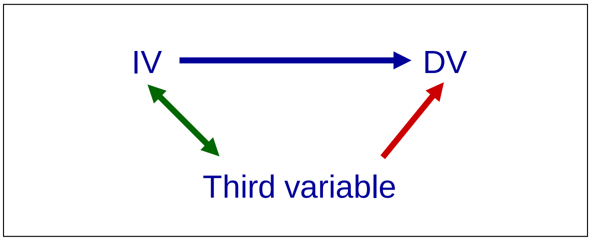
- May dictate inclusion of particular predictors
  - Predictors of interest
    - The scientific factor under investigation can/should/must be modeled by multiple predictors (e.g., dummy variables, etc.)
  - Confounders
    - The scientific question can't be answered without adjusting for known (or suspected) confounders
  - Effect modifiers
    - The scientific question may relate to detection of effect modification

Confounding

- Sometimes the study question of greatest interest is confounded by associations in the data
- A third factor which is related to both IV and DV, and which accounts for some or all of the observed relationship between the two
  - *"In general, confounding exists if meaningfully different interpretations of the relationship of interest result when an extraneous variable is ignored or included in the analysis"*
- Causes both Type I and Type II error

Confounding

- To be a confounding factor, two conditions must be met:



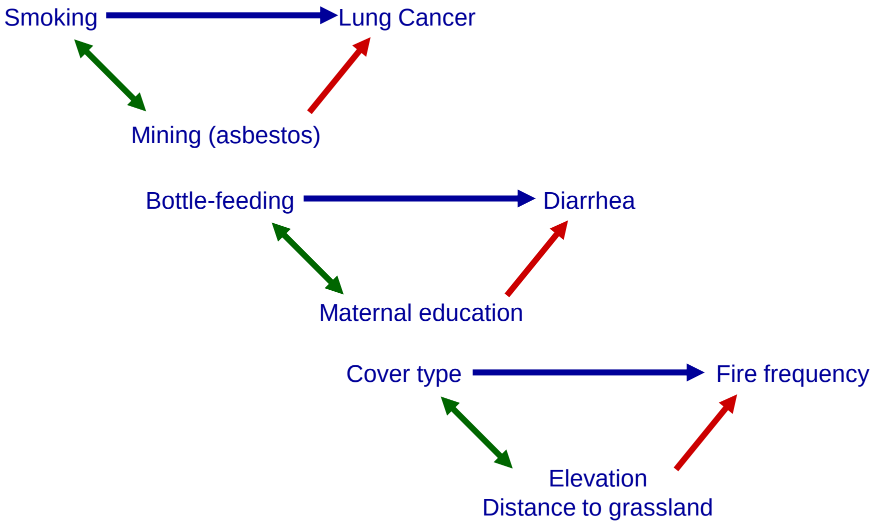
IV

DV

Third variable

1. Be associated with IV
   - without being the consequence of IV
2. Be associated with DV
   - independently of IV (not an intermediary)

Confounding examples



Smoking

Lung Cancer

Mining (asbestos)

Bottle-feeding

Diarrhea

Maternal education

Cover type

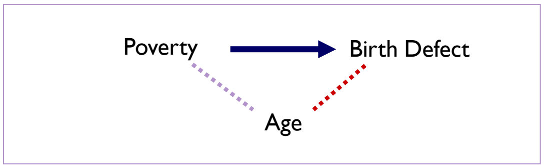
Fire frequency

Elevation
Distance to grassland

Effect modification

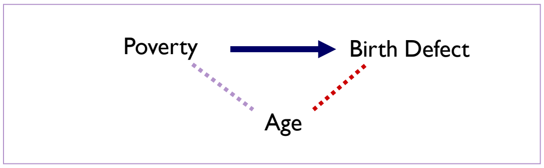
- Sometimes a study question if affected by (changes) due to effect modifiers
- The strength of the association varies over different categories of a third variable
    - The third variable is changing the effect of the IV
- There is no adjustment for effect modification
    - Must use interaction terms OR stratified analysis

Confounding or effect modification



- Can age be responsible for the poverty association with birth defects?
    - Is it correlated with poverty?
    - Is it correlated with the birth defect independently of poverty?
    - Is it associated with the birth defect even if poverty is low?
    - Is age distribution uneven in comparison groups?

Confounding or effect modification

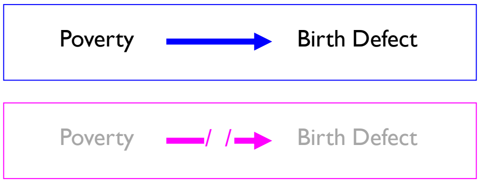


Poverty → Birth Defect

Age

Women in poverty have a 50% increase in risk of a birth defect

Does poverty association differ in strength according to sex?



Mother <20 years

Poverty → Birth Defect

Women in poverty have a 80% increase in risk of a birth defect

Mother ≥20 years

Poverty —/ /→ Birth Defect

Women in poverty have a 10% decrease in risk of a birth defect

Confounding vs. effect modification

- Effect modifier
  - Belongs to nature (something inherently different between strata)
  - Coefficients different (different effects) for different strata
  - Useful
  - Increases knowledge of causal mechanism
- Confounding factor
  - Belongs to study (poor data collection/sampling)
  - Coefficients different before and after confounder introduced
  - Distortion of effect
  - Creates confusion in data

Precision

- Adjusting for an additional IVs (covariates, confounders, etc) changes the standard error of the slope estimate
  - Standard error is decreased by having smaller within group variance
  - Standard error is increased by having correlations between the predictor of interest and other covariates in the model
  - Changes significance level of coefficients

A few general comments

- Can be difficult to choose the "best" model, since many reasonable candidates may exist
- More difficult to visualize the fitted model
- More difficult to interpret the fitted model

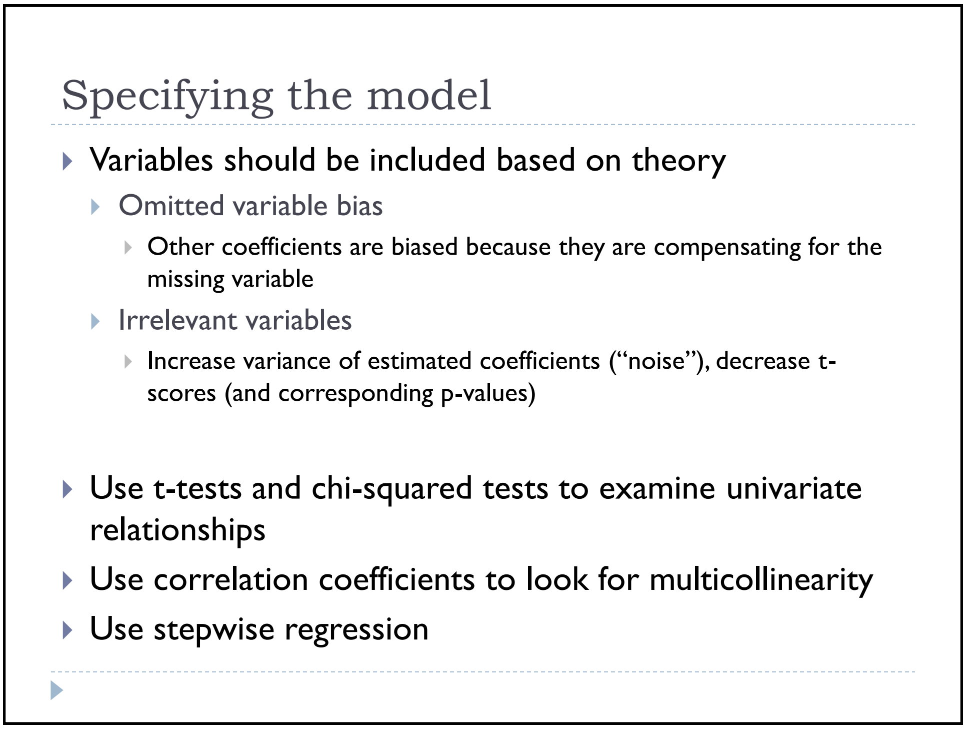
Assumptions of multiple regression

1. The relationship between y (DV) and x (IV) is (approximately) linear
2. None of the independent variables are highly correlated
   - Multicollinearity
3. The errors (residuals) are normally distributed and have a 0 population mean
4. The residuals do not vary with x
   - Constant variance, no heteroskedacticity
5. The residuals are uncorrelated with each other
   - Serial correlation, as with time series

Steps in applied regression analysis

1. Review the literature and develop a theoretical model
2. Specify the model
   - The independent variables and how they should be measured
   - The mathematical form of the variables
     - mistake=specification error
3. Hypothesize the expected signs of the coefficients
4. Estimate the regression and evaluate the results

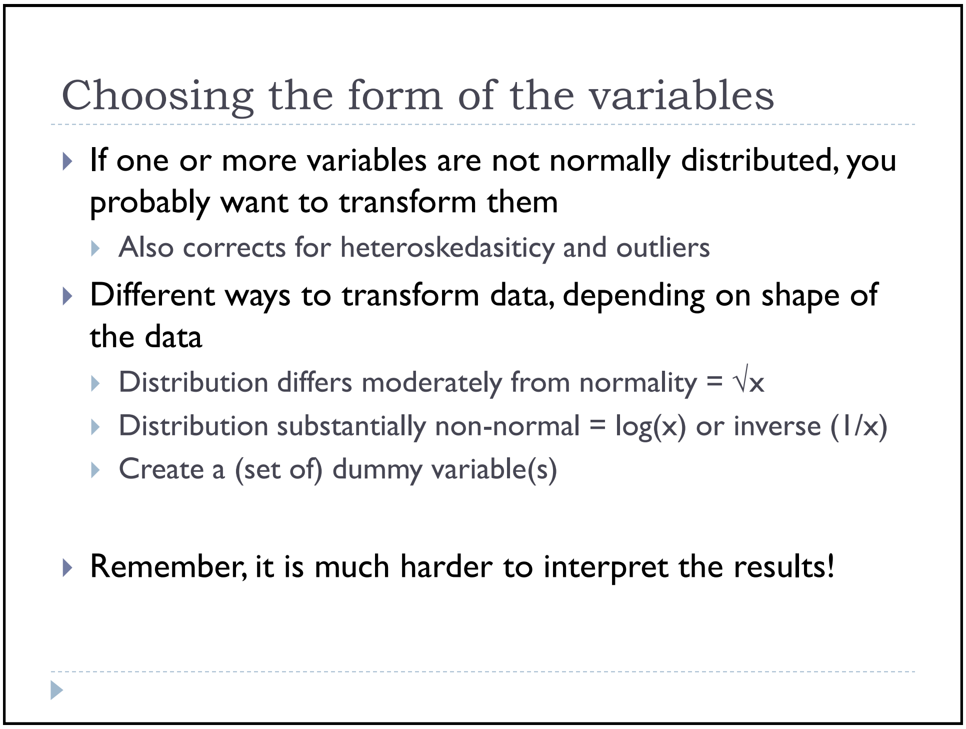
- Note: we are learning Ordinary Least Square (OLS) regression
  - Goal of minimizing the summed squared residuals



Specifying the model

- Variables should be included based on theory
  - Omitted variable bias
    - Other coefficients are biased because they are compensating for the missing variable
  - Irrelevant variables
    - Increase variance of estimated coefficients (“noise”), decrease t-scores (and corresponding p-values)

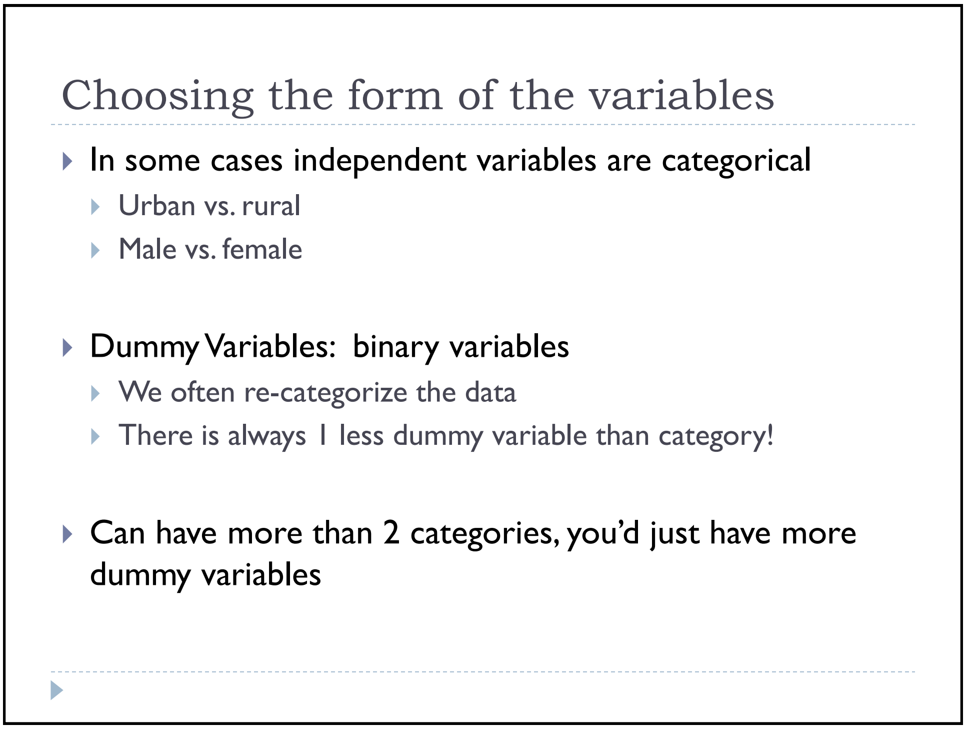
- Use t-tests and chi-squared tests to examine univariate relationships
- Use correlation coefficients to look for multicollinearity
- Use stepwise regression



Choosing the form of the variables

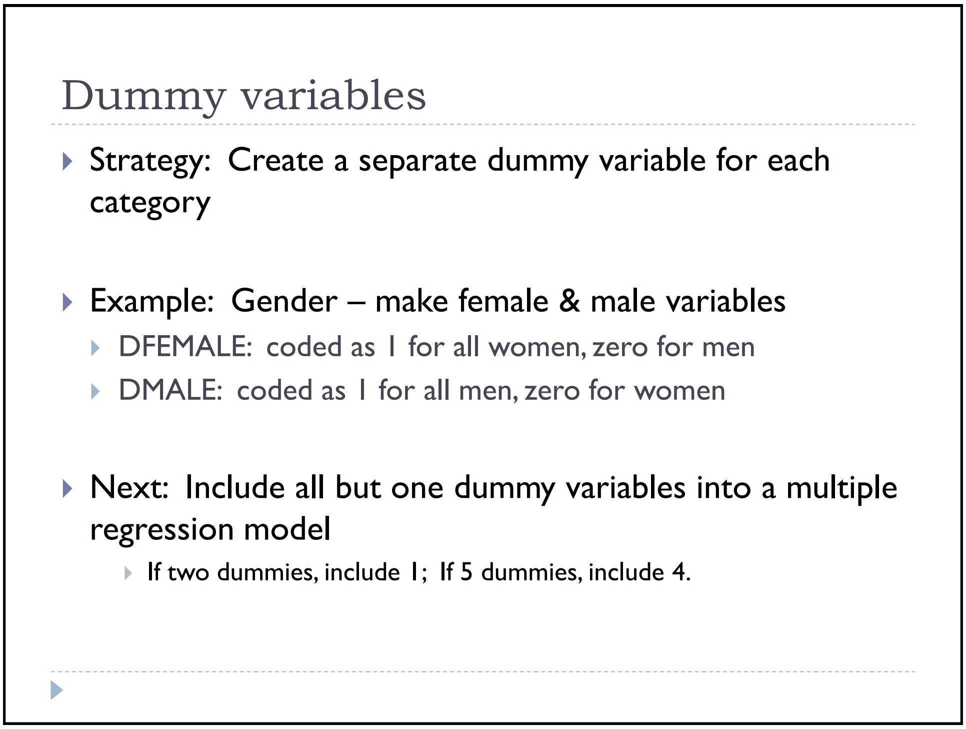
- If one or more variables are not normally distributed, you probably want to transform them
  - Also corrects for heteroskedasiticy and outliers
- Different ways to transform data, depending on shape of the data
  - Distribution differs moderately from normality = √x
  - Distribution substantially non-normal = log(x) or inverse (1/x)
  - Create a (set of) dummy variable(s)

- Remember, it is much harder to interpret the results!



Choosing the form of the variables

- In some cases independent variables are categorical
  - Urban vs. rural
  - Male vs. female
- Dummy Variables: binary variables
  - We often re-categorize the data
  - There is always 1 less dummy variable than category!
- Can have more than 2 categories, you'd just have more dummy variables



Dummy variables

- Strategy: Create a separate dummy variable for each category
- Example: Gender – make female & male variables
  - DFEMALE: coded as 1 for all women, zero for men
  - DMALE: coded as 1 for all men, zero for women
- Next: Include all but one dummy variables into a multiple regression model
  - If two dummies, include 1; If 5 dummies, include 4.

Dummy variables

- Question: Why can't you include DFEMALE and DMALE in the same regression model?
- Answer: They are perfectly correlated (negatively): r = -1
  - Result: Regression model "blows up"
- For any set of nominal categories, a full set of dummies contains redundant information
  - DMALE and DFEMALE contain same information
  - Dropping one removes redundant information

Interpreting dummy variables

- Consider the following regression equation:

$$Y_i = a + b_1 INCOME_i + b_2 DFEMALE_i + e_i$$

- Question: What if the case is a male?
- Answer: DFEMALE is 0, so the entire term becomes zero
  - Result: Males are modeled using the familiar regression model: a + b_1X + e

Interpreting dummy variables

- Consider the following regression equation:

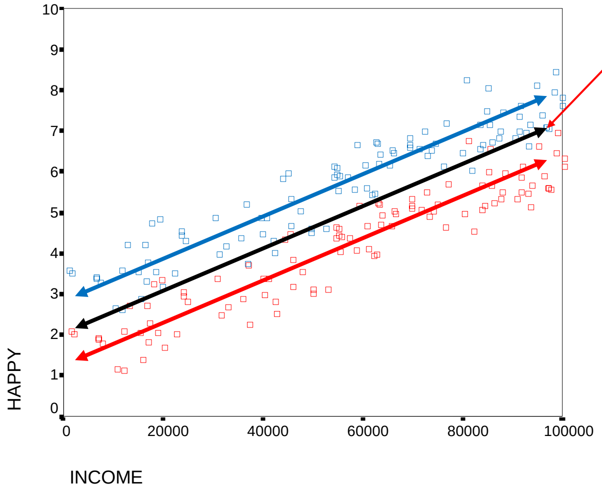
$$Y_i = a + b_1 INCOME_i + b_2 DFEMALE_i + e_i$$

- Question: What if the case is a female?

- Answer: DFEMALE is 1, so b_2(1) stays in the equation (and is added to the constant)
  - Result: Females are modeled using a different regression line: $(a+b_2) + b_1X + e$
  - Thus, the coefficient of b_2 reflects difference in the constant for women

Interpreting dummy variables

- Women = blue, Men = red



Overall slope for all data points

Line for men, women have same slope…but one is high other is lower

The constant differs!

If women=1, men=0:

Dummy coefficient (b) reflects increase for women (relative to men)

Interaction terms

- Sometimes we want to know if the relationship between the DV and one of our IVs differs depending on the value of another variable
  - Example: the relationship between income (IV) and happiness (DV)…does it differ by sex
    - Perhaps men are more materialistic - an extra dollar increases their happiness a lot
    - If women are less materialistic, each dollar has a smaller effect on income (compared to men)
  - The slope of a variable coefficient (for income) differs across groups

Interaction terms

- Interaction effects: Differences in the relationship (slope) between two variables for each category of a third variable

- Option #1: Analyze each group separately
- Option #2: Multiply the two variables of interest: (DFEMALE, INCOME) to create a new variable
  - Called: DFEMALE*INCOME
  - Add that variable to the multiple regression model
  - This is called an interaction term

Interaction terms

- Consider the following regression equation:

$$y_i = a + b_1 INCOME_i + b_2 DFEM * INC_i + e_i$$

- Question: What if the case is male?

- Answer: DFEMALE is 0, so b_2 (DFEM*INC) drops out of the equation
  - Result: Males are modeled using the ordinary regression equation: $a + b_1X + e$

Interaction terms

- Consider the following regression equation:

$$y_i = a + b_1 INCOME_i + b_2 DFEM * INC_i + e_i$$

- Question: What if the case is female?

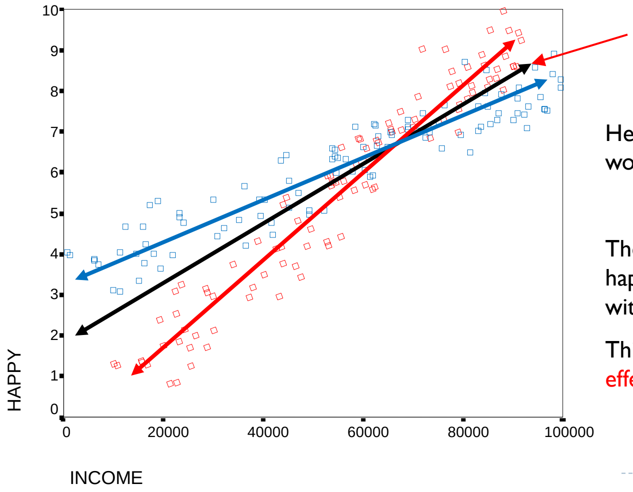
- Answer: DFEMALE is 1, so b_2(DFEM*INC) becomes b_2*INCOME, which is added to b_1
  - Result: Females are modeled using a different regression line: $a + (b_1+b_2)\ X + e$
  - Thus, the coefficient of b_2 reflects difference in the slope of INCOME for women

Interpreting interaction terms

- A positive *b* for DFEMALE*INCOME indicates the slope for income is higher for women vs. men
  - A negative effect indicates the slope is lower
  - Size of coefficient indicates actual difference in slope
- Example: DFEMALE*INCOME. Observed *b*'s:
  - INCOME: *b* = .5
  - DFEMALE * INCOME: *b* = -.2
- Interpretation: Slope is .5 for men, .3 for women.

Interpreting interaction terms

- Women = blue, Men = red

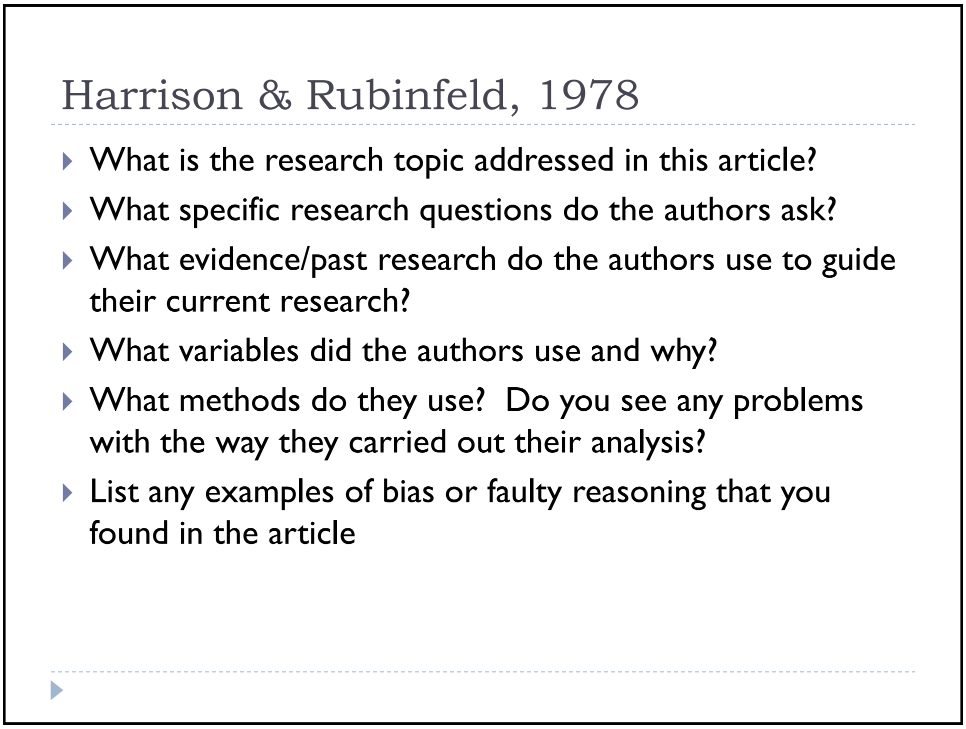


Overall slope for all data points

Here, the slope for men and women differs

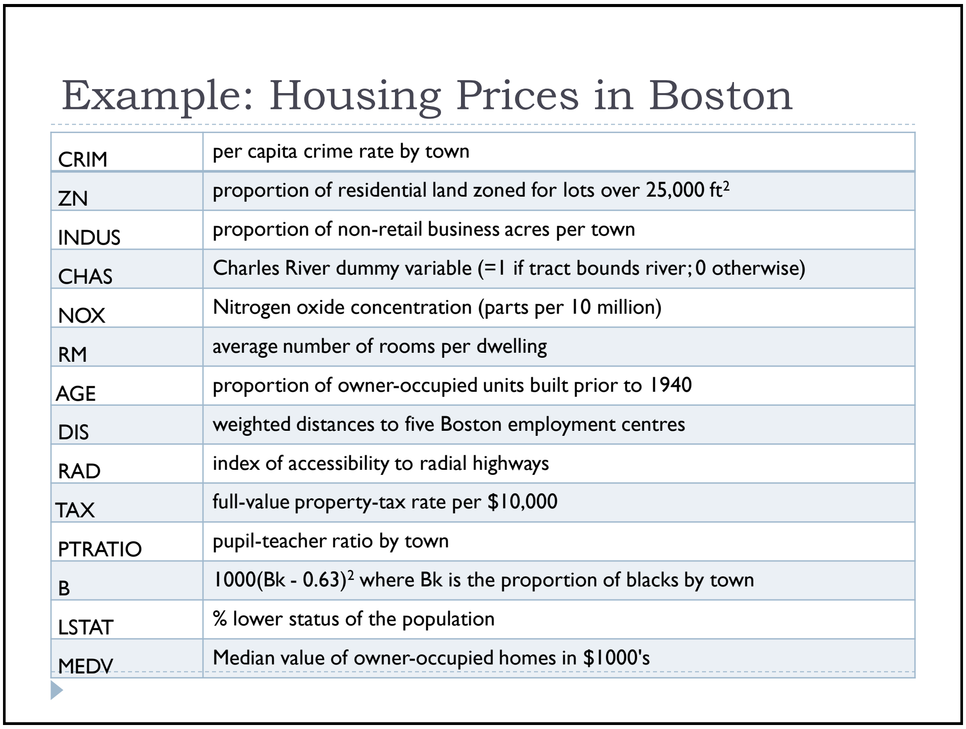
The effect of income on happiness (X_1 on Y) varies with gender (X_2)

This is called an "interaction effect"



Harrison & Rubinfeld, 1978

- What is the research topic addressed in this article?
- What specific research questions do the authors ask?
- What evidence/past research do the authors use to guide their current research?
- What variables did the authors use and why?
- What methods do they use? Do you see any problems with the way they carried out their analysis?
- List any examples of bias or faulty reasoning that you found in the article



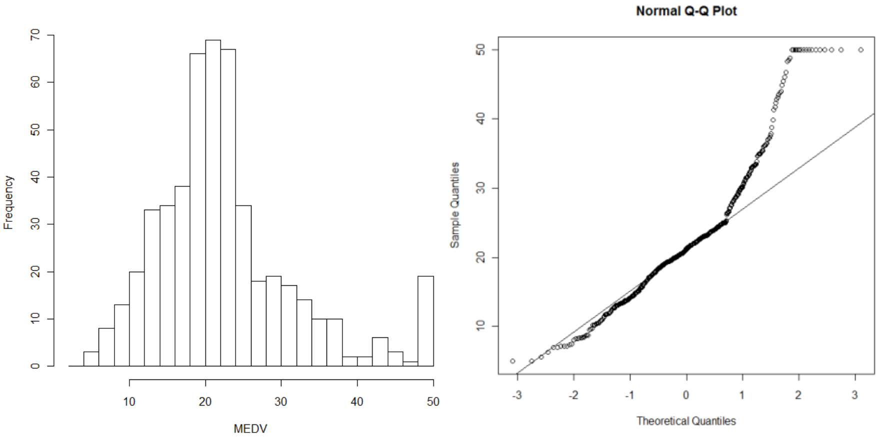
Example: Housing Prices in Boston

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 ft^2
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (=1 if tract bounds river; 0 otherwise)
NOX	Nitrogen oxide concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

Read data in, look at distribution of DV

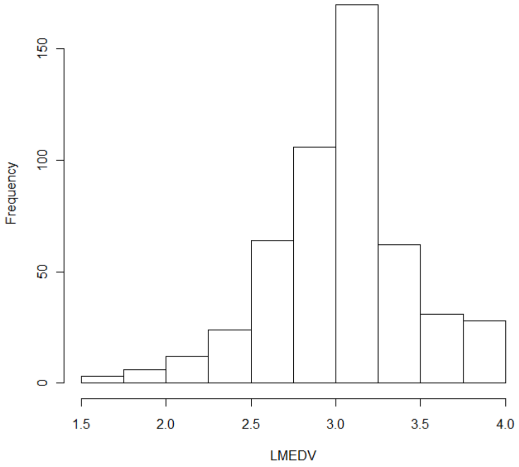
```
> boston<-read.csv("C:/Users/Elisabeth Root/Desktop/Quant/
  R/boston.csv",header=T)
> names(boston)
 [1] "OBS."    "TOWN"    "TOWN."   "TRACT"   "LON"     "LAT"     "MEDV"
 [8] "CMEDV"   "CRIM"    "ZN"      "INDUS"   "CHAS"    "NOX"     "RM"
[15] "AGE"     "DIS"     "RAD"     "TAX"     "PTRATIO" "B"       "LSTAT‟

> hist(MEDV)
> qqnorm(MEDV)
> qqline(MEDV)
```

Histogram and QQPlot



Histogram transformed

```
> boston$LMEDV<-log(boston$MEDV)
> hist(LMEDV)
```



Now what?

- Specify the equation
  - Examine simple correlation matrix
    - `cor(boston[,c(9:22)])`
  - Which variables would you include? Why?
  - What signs (+/-) do you expect the coefficients to have? Why?
  - Do we need to transform any of the other variables?

- Get in groups and specify an equation…

```
> bost<-lm(LMEDV ~ RM + LSTAT + CRIM + ZN + CHAS + DIS)
> summary(bost)
```

Regression Diagnostics

- Departures from the underlying assumptions cannot be detected using any of the summary statistics we've examined so far such as the t or F statistics or R^2
  - Linearity
  - Constant variance
  - Normality
  - Independence/correlation
- The diagnostic methods we'll be exploring are based primarily on the residuals
  - Recall, the residual is defined as:

$$e_i = (y_i - \hat{y}_i)$$

How to use residuals for diagnostics

- Residual analysis is usually done graphically using:
  - Quantile plots: to assess normality
  - Histograms and boxplots
  - Scatterplots: to assess model assumptions, such as constant variance and linearity, and to identify potential outliers
  - Cook's D: to check for influential observations

Checking the normality of the error terms

- To check if the population mean of residuals=0

```
> mean(bost$residuals)
[1] 3.915555e-18
```

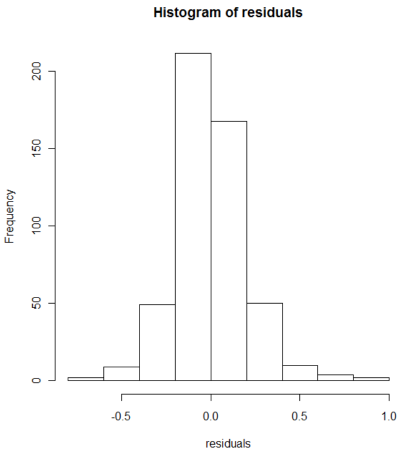
- histogram of residuals

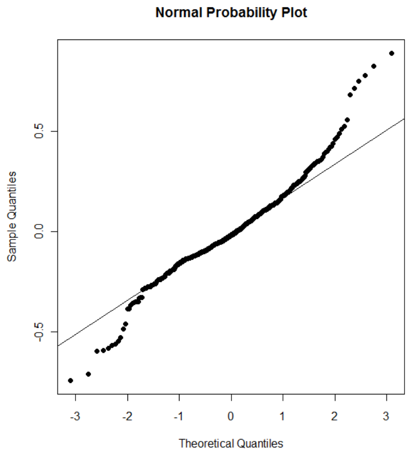
```
> hist(bost$residuals, xlab="residuals", main="Histogram
  of residuals")
```

- normal probability plot, or QQ-plot

```
> qqnorm(bost$residuals, main="Normal Probability Plot",
  pch=19)
> qqline(bost$residuals)
```

Result





Checking: linear relationship, error has a constant variance, error terms are not independent

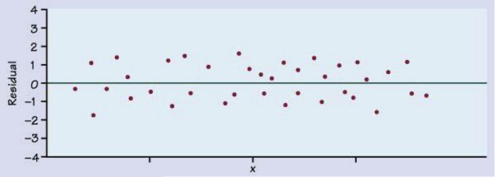
- plot residuals against each predictor (x=LSTAT)

```
> plot(boston$LSTAT, bost$residuals, main="Residuals
  vs. Predictor", xlab="% in Lower Status",
  ylab="Residuals", pch=19)
> abline(h=0)
```

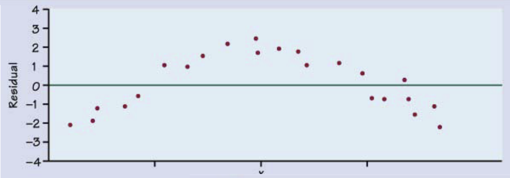
- plot residuals against fitted values (Y-hat)

```
> plot(bost$fitted.values, bost$residuals, main="Residuals
  vs. Fitted", xlab="Fitted values", ylab="Residuals",
  pch=19)
> abline(h=0)
```

Patterns in the residuals

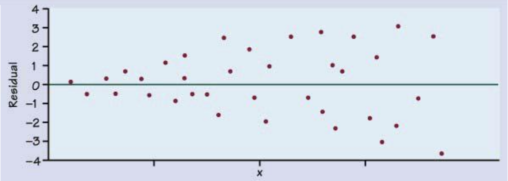


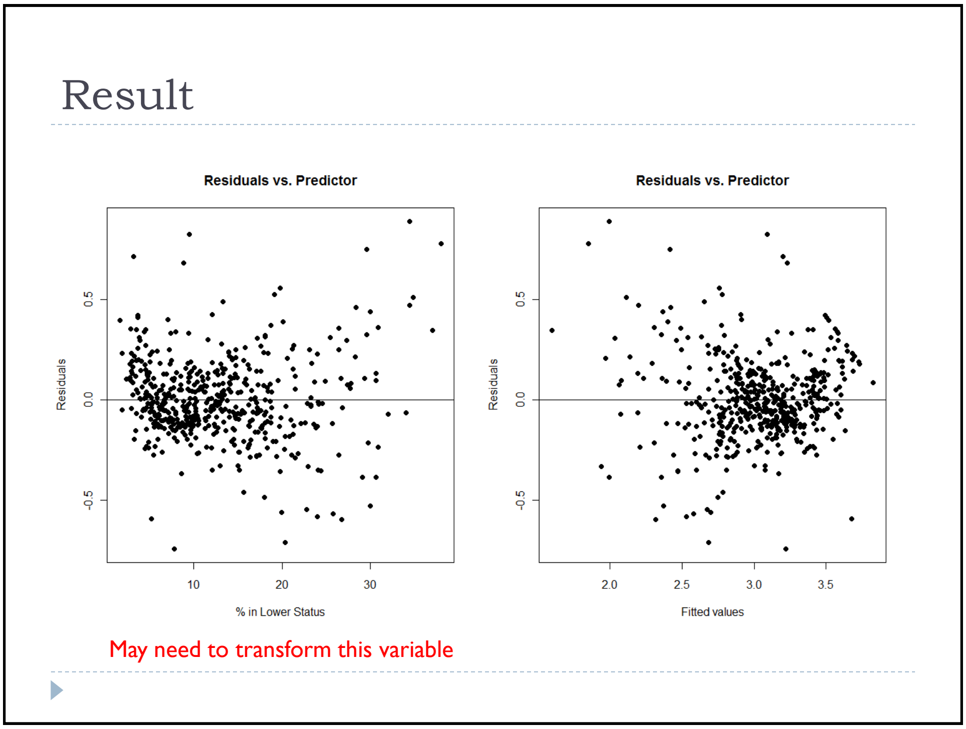
Totally random



Nonlinear

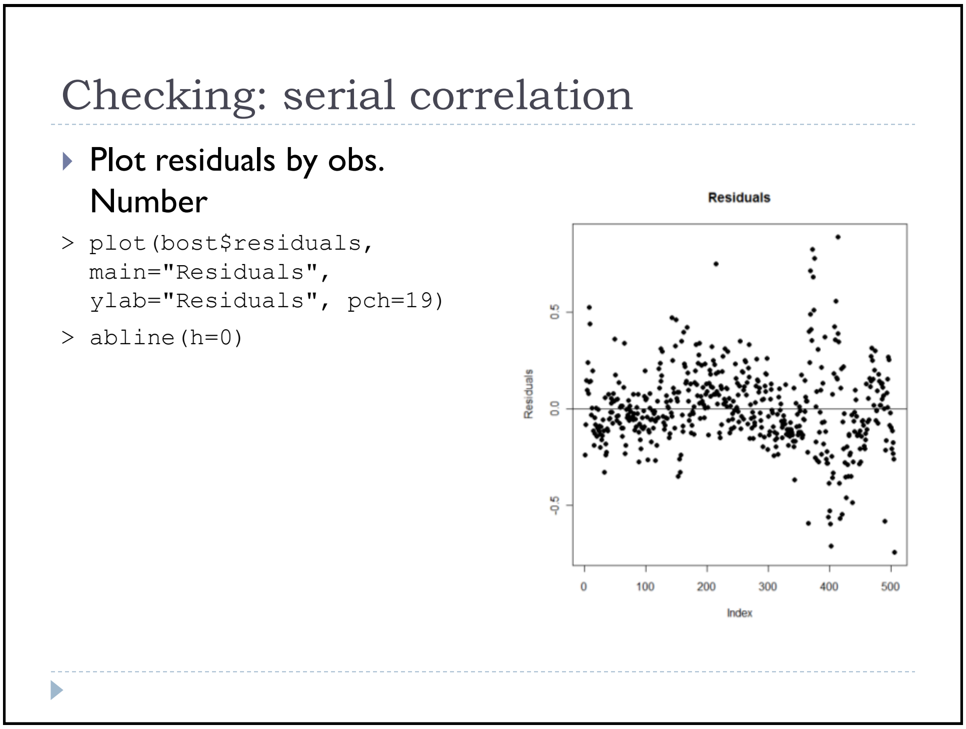
Heteroskedastic
(increasing variance with value)





Result

May need to transform this variable



Checking: serial correlation

- Plot residuals by obs. Number

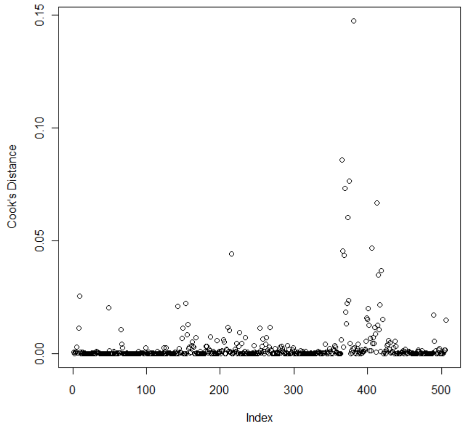
```
> plot(bost$residuals,
  main="Residuals",
  ylab="Residuals", pch=19)
> abline(h=0)
```

Checking: influential observations

- Cook's D measures the influence of the *i*th observation on all n fitted values
- The magnitude of D_i is usually assessed as:
  - if the percentile value is less than 10 or 20 % than the *i*th observation has little apparent influence on the fitted values
  - if the percentile value is greater than 50%, we conclude that the *i*th observation has significant effect on the fitted values

Cook's D in R

```
> cd=cooks.distance(bost)
> plot(cd, ylab="Cook's
  Distance")
> abline(h=qf(c(.2,.5),6,
  499))
> ic=(1:506)[cd>qf(c(.2,.5),
  6,499)]
> text(ic,cd[ic],
  as.character(boston$OBS
  [ic]),adj=c(1,1))

Error in text.default(ic,
  cd[ic], as.character
  (boston$OBS[ic]), adj =
  c(1,  : zero length
  'labels'
```



“Fixing” heteroskedasticity

- Non-constant variance can often be remedied using appropriate transformations
  - Ideally, we would choose the transformation based on some prior scientific knowledge, but this might not always be available
- Commonly used variance stabilizing techniques:
  - natural log (log(y) or ln(y))
  - log base 10 (log10)
  - reciprocal or inverse (1/y)
  - square root ($\sqrt{y}$)

Model building

- There is no “best” regression model
  - What is “best”?
  - There are a number of ways we can choose the “best” – they will not all yield the same results.
  - What about the other potential problems with the model that might have been ignored while selecting the “best” model?

Techniques for model building

- All possible regressions
- Stepwise regression methods
  - Forward selection
  - Backward elimination
  - Stepwise regression
- How can we evaluate and compare different candidate models?
  - Adjusted R^2 (largest)
  - Residual mean square (smallest)
  - Change in AIC (largest reduction)

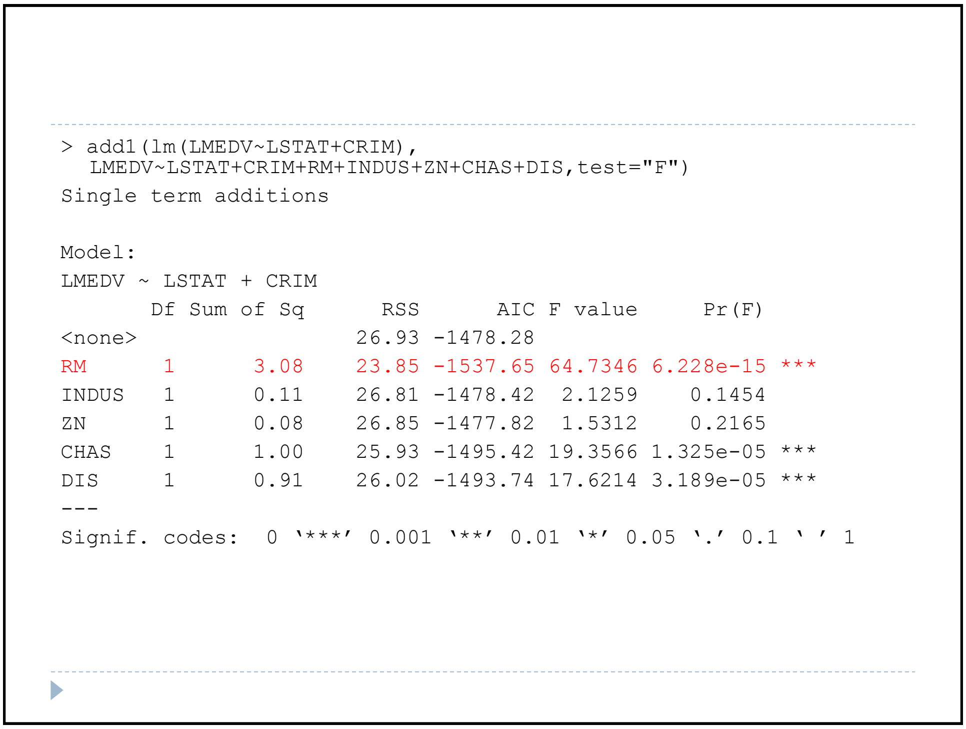
Forward selection

1. Begin with the assumption that there are no regressors in the model
2. Check models with all possible regressors added individually
3. Add the regressor that most changes your criterion in the correct direction…Go back to 2
4. If none of the regressors have a positive effect on your criterion, stop with the regressors you have.
   - This is your final model

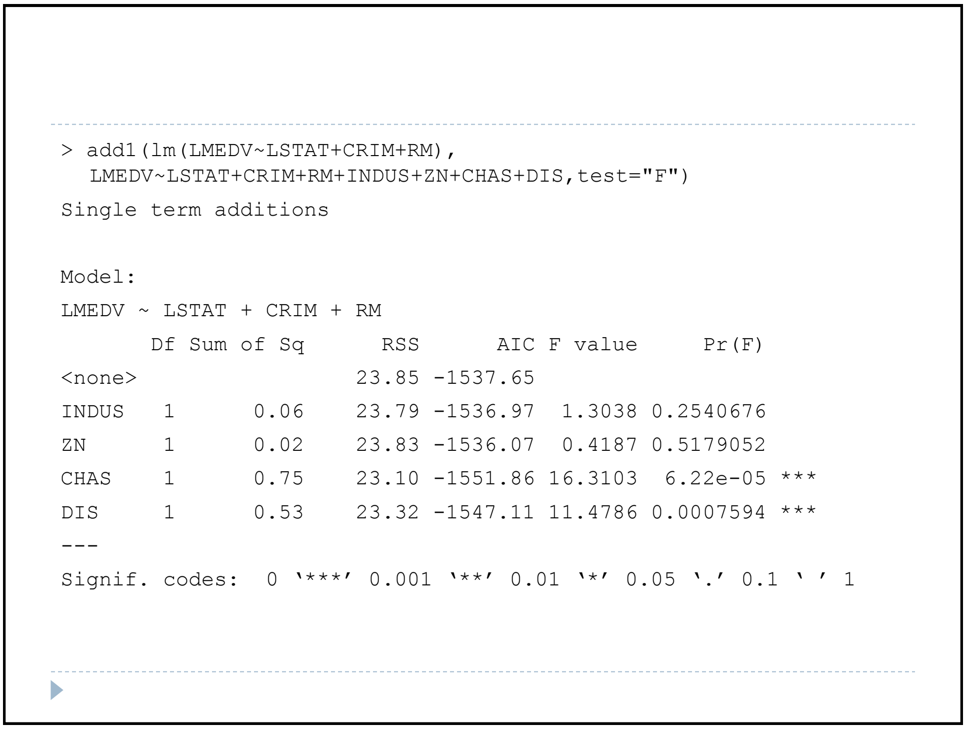
Example using R

```
> add1(lm(LMEDV~1),
  LMEDV~RM+LSTAT+CRIM+INDUS+ZN+CHAS+DIS,test="F")
Single term additions

Model:
LMEDV ~ 1
       Df Sum of Sq     RSS      AIC F value     Pr(F)
<none>                84.38  -904.37
RM      1     33.70   50.67 -1160.39 335.232 < 2.2e-16 ***
LSTAT   1     54.68   29.69 -1430.81 928.143 < 2.2e-16 ***
CRIM    1     23.52   60.86 -1067.70 194.765 < 2.2e-16 ***
INDUS   1     24.75   59.63 -1078.02 209.157 < 2.2e-16 ***
ZN      1     11.14   73.24  -974.01  76.658 < 2.2e-16 ***
CHAS    1      2.12   82.26  -915.23  12.973 0.0003473 ***
DIS     1      9.91   74.46  -965.62  67.104 2.136e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> add1(lm(LMEDV~LSTAT),
  LMEDV~LSTAT+RM+CRIM+INDUS+ZN+CHAS+DIS,test="F")
Single term additions

Model:
LMEDV ~ LSTAT
       Df Sum of Sq     RSS      AIC F value     Pr(F)
<none>                29.69 -1430.81
RM      1      2.57   27.12 -1474.69 47.7396 1.478e-11 ***
CRIM    1      2.77   26.93 -1478.28 51.6560 2.403e-12 ***
INDUS   1      0.41   29.29 -1435.83  7.0194  0.008317 **
ZN      1      0.10   29.60 -1430.47  1.6475  0.199889
CHAS    1      1.12   28.57 -1448.25 19.6991 1.115e-05 ***
DIS     1      0.37   29.33 -1435.13  6.3145  0.012287 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
> add1(lm(LMEDV~LSTAT+CRIM),
   LMEDV~LSTAT+CRIM+RM+INDUS+ZN+CHAS+DIS,test="F")
Single term additions

Model:
LMEDV ~ LSTAT + CRIM
       Df Sum of Sq     RSS     AIC F value     Pr(F)
<none>                26.93 -1478.28
RM      1      3.08   23.85 -1537.65 64.7346 6.228e-15 ***
INDUS   1      0.11   26.81 -1478.42  2.1259    0.1454
ZN      1      0.08   26.85 -1477.82  1.5312    0.2165
CHAS    1      1.00   25.93 -1495.42 19.3566 1.325e-05 ***
DIS     1      0.91   26.02 -1493.74 17.6214 3.189e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



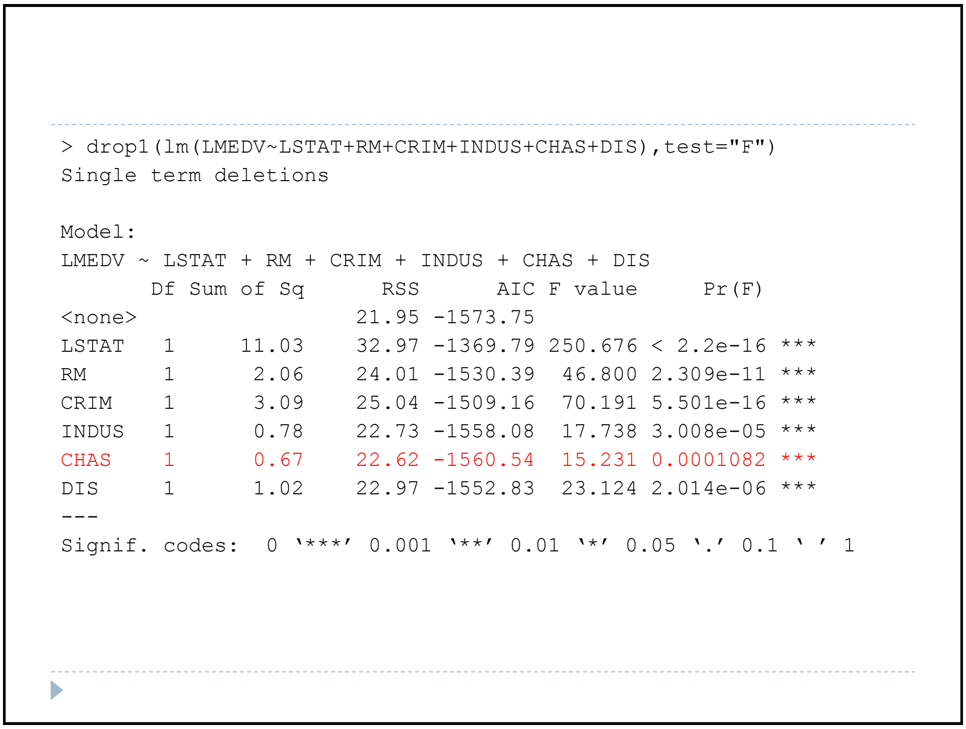
```
> add1(lm(LMEDV~LSTAT+CRIM+RM),
   LMEDV~LSTAT+CRIM+RM+INDUS+ZN+CHAS+DIS,test="F")
Single term additions

Model:
LMEDV ~ LSTAT + CRIM + RM
       Df Sum of Sq     RSS     AIC F value     Pr(F)
<none>                23.85 -1537.65
INDUS   1      0.06   23.79 -1536.97  1.3038 0.2540676
ZN      1      0.02   23.83 -1536.07  0.4187 0.5179052
CHAS    1      0.75   23.10 -1551.86 16.3103  6.22e-05 ***
DIS     1      0.53   23.32 -1547.11 11.4786 0.0007594 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Backward elimination

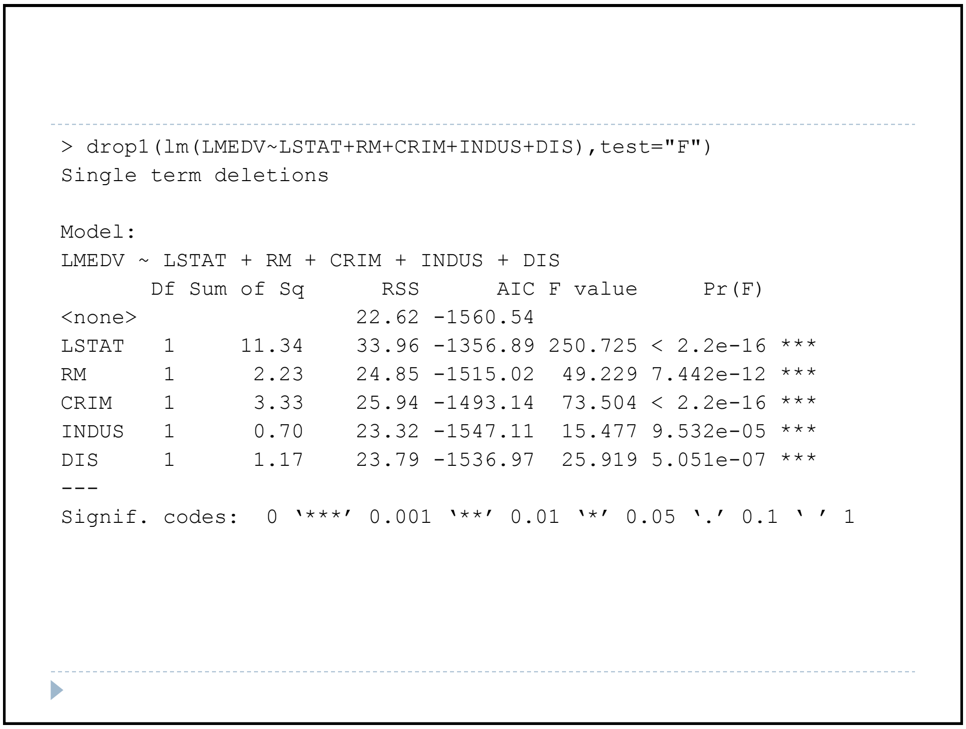
1. Start with all candidate regressors in the model
2. Drop the predictor that improves your selection criterion the least
3. Continue until there is no predictor that can be dropped and result in an improvement of your selection criterion, then all the remaining predictors define your final model

Example using R

```
> drop1(lm(LMEDV~LSTAT+RM+CRIM+INDUS+ZN+CHAS+DIS),test="F")
Single term deletions

Model:
LMEDV ~ LSTAT + RM + CRIM + INDUS + ZN + CHAS + DIS
       Df Sum of Sq     RSS      AIC F value     Pr(F)
<none>                21.49 -1582.44
LSTAT   1     11.03   32.52 -1374.88 255.509 < 2.2e-16 ***
RM      1      1.64   23.13 -1547.20  38.032 1.441e-09 ***
CRIM    1      3.34   24.83 -1511.41  77.315 < 2.2e-16 ***
INDUS   1      0.71   22.20 -1568.00  16.443 5.816e-05 ***
ZN      1      0.46   21.95 -1573.75  10.625 0.0011922 **
CHAS    1      0.65   22.14 -1569.33  15.092 0.0001163 ***
DIS     1      1.47   22.96 -1550.95  34.069 9.595e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
> drop1(lm(LMEDV~LSTAT+RM+CRIM+INDUS+CHAS+DIS),test="F")
Single term deletions

Model:
LMEDV ~ LSTAT + RM + CRIM + INDUS + CHAS + DIS
       Df Sum of Sq     RSS      AIC F value     Pr(F)
<none>                21.95 -1573.75
LSTAT   1     11.03   32.97 -1369.79 250.676 < 2.2e-16 ***
RM      1      2.06   24.01 -1530.39  46.800 2.309e-11 ***
CRIM    1      3.09   25.04 -1509.16  70.191 5.501e-16 ***
INDUS   1      0.78   22.73 -1558.08  17.738 3.008e-05 ***
CHAS    1      0.67   22.62 -1560.54  15.231 0.0001082 ***
DIS     1      1.02   22.97 -1552.83  23.124 2.014e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
> drop1(lm(LMEDV~LSTAT+RM+CRIM+INDUS+DIS),test="F")
Single term deletions

Model:
LMEDV ~ LSTAT + RM + CRIM + INDUS + DIS
       Df Sum of Sq     RSS      AIC F value     Pr(F)
<none>                22.62 -1560.54
LSTAT   1     11.34   33.96 -1356.89 250.725 < 2.2e-16 ***
RM      1      2.23   24.85 -1515.02  49.229 7.442e-12 ***
CRIM    1      3.33   25.94 -1493.14  73.504 < 2.2e-16 ***
INDUS   1      0.70   23.32 -1547.11  15.477 9.532e-05 ***
DIS     1      1.17   23.79 -1536.97  25.919 5.051e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

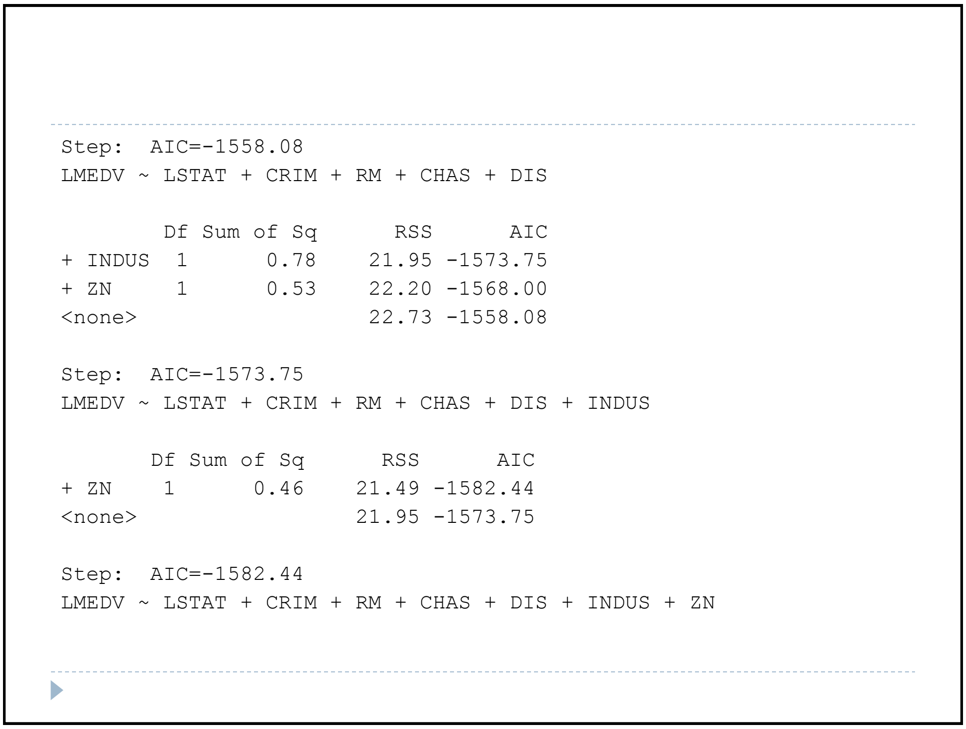
Stepwise Regression

- General stepwise regression techniques are usually a combination of backward elimination and forward selection, alternating between the two techniques at different steps
- Typically uses the AIC at each step to select the "next" variable to add

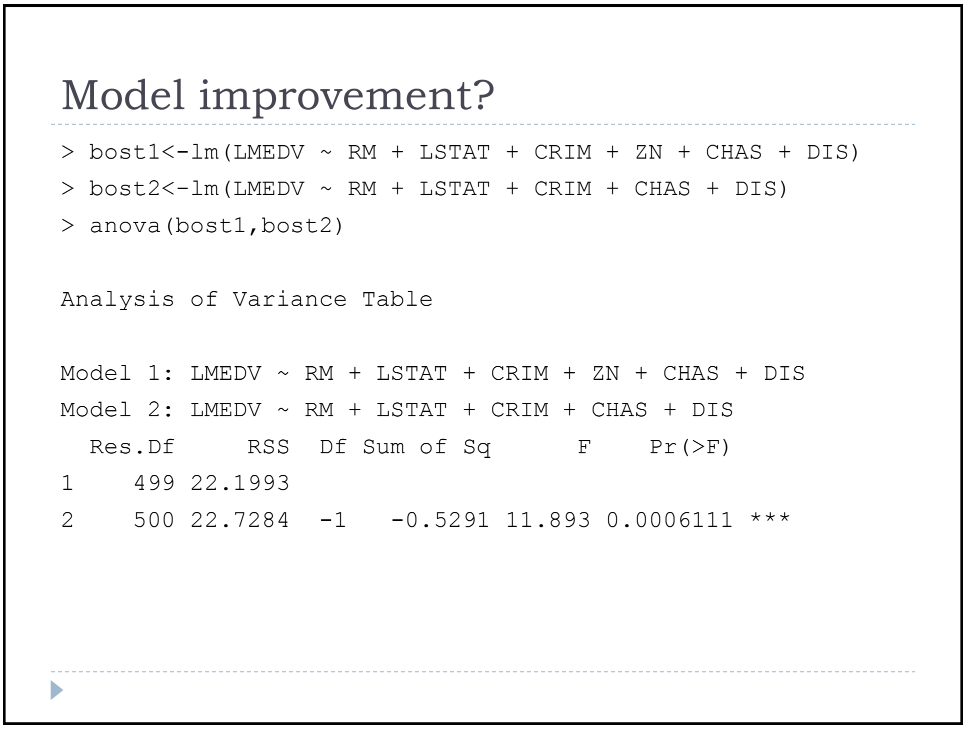
```
> step(lm(LMEDV~1), LMEDV~LSTAT+RM+CRIM+
  INDUS+ZN+CHAS+DIS,direction="forward")
Start:  AIC=-904.37
LMEDV ~ 1

        Df Sum of Sq     RSS      AIC
+ LSTAT  1     54.68   29.69 -1430.81
+ RM     1     33.70   50.67 -1160.39
+ INDUS  1     24.75   59.63 -1078.02
+ CRIM   1     23.52   60.86 -1067.70
+ ZN     1     11.14   73.24  -974.01
+ DIS    1      9.91   74.46  -965.62
+ CHAS   1      2.12   82.26  -915.23
<none>                 84.38  -904.37
```

```
Step:  AIC=-1430.81
LMEDV ~ LSTAT

        Df Sum of Sq     RSS      AIC
+ CRIM   1      2.77   26.93 -1478.28
+ RM     1      2.57   27.12 -1474.69
+ CHAS   1      1.12   28.57 -1448.25
+ INDUS  1      0.41   29.29 -1435.83
+ DIS    1      0.37   29.33 -1435.13
<none>                 29.69 -1430.81
+ ZN     1      0.10   29.60 -1430.47

Step:  AIC=-1478.28
LMEDV ~ LSTAT + CRIM

        Df Sum of Sq     RSS      AIC
+ RM     1      3.08   23.85 -1537.65
+ CHAS   1      1.00   25.93 -1495.42
+ DIS    1      0.91   26.02 -1493.74
+ INDUS  1      0.11   26.81 -1478.42
<none>                 26.93 -1478.28
+ ZN     1      0.08   26.85 -1477.82
```

```
Step:  AIC=-1537.65
LMEDV ~ LSTAT + CRIM + RM

        Df Sum of Sq     RSS      AIC
+ CHAS   1      0.75   23.10 -1551.86
+ DIS    1      0.53   23.32 -1547.11
<none>                 23.85 -1537.65
+ INDUS  1      0.06   23.79 -1536.97
+ ZN     1      0.02   23.83 -1536.07

Step:  AIC=-1551.86
LMEDV ~ LSTAT + CRIM + RM + CHAS

        Df Sum of Sq     RSS      AIC
+ DIS    1      0.37   22.73 -1558.08
+ INDUS  1      0.14   22.97 -1552.83
<none>                 23.10 -1551.86
+ ZN     1      0.04   23.06 -1550.83
```



```
Step:  AIC=-1558.08
LMEDV ~ LSTAT + CRIM + RM + CHAS + DIS

        Df Sum of Sq      RSS      AIC
+ INDUS  1      0.78    21.95 -1573.75
+ ZN     1      0.53    22.20 -1568.00
<none>                  22.73 -1558.08

Step:  AIC=-1573.75
LMEDV ~ LSTAT + CRIM + RM + CHAS + DIS + INDUS

       Df Sum of Sq      RSS      AIC
+ ZN    1      0.46    21.49 -1582.44
<none>                 21.95 -1573.75

Step:  AIC=-1582.44
LMEDV ~ LSTAT + CRIM + RM + CHAS + DIS + INDUS + ZN
```



Model improvement?

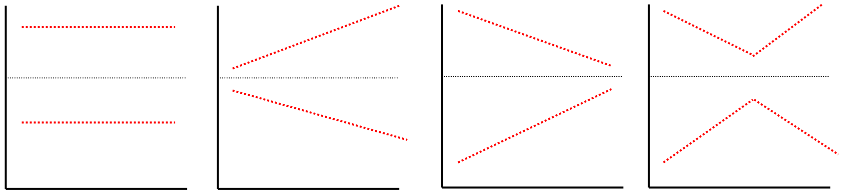
```
> bost1<-lm(LMEDV ~ RM + LSTAT + CRIM + ZN + CHAS + DIS)
> bost2<-lm(LMEDV ~ RM + LSTAT + CRIM + CHAS + DIS)
> anova(bost1,bost2)

Analysis of Variance Table

Model 1: LMEDV ~ RM + LSTAT + CRIM + ZN + CHAS + DIS
Model 2: LMEDV ~ RM + LSTAT + CRIM + CHAS + DIS
  Res.Df     RSS  Df Sum of Sq      F    Pr(>F)
1    499 22.1993
2    500 22.7284  -1   -0.5291 11.893 0.0006111 ***
```

Comment on stepwise regression techniques

- The techniques we've discussed in this lecture are all quantitative/computational and therefore have many scientific drawbacks
- They should NEVER replace careful scientific thought and consideration in model building
- It yields R^2 values that are biased high

A note on heteroskedastic error terms

- Harrison, et al. noticed heteroskedastic error terms
  - Used Park-Glejser test
- If the errors are heteroskedastically distributed
  - The SE may inefficient, i.e. either too small or large
- BEST thing to do is figure out which variables are causing heteroskedasticity and remove or transform them



Homoskedastic

Heteroskedastic (+)

Heteroskedastic (-)

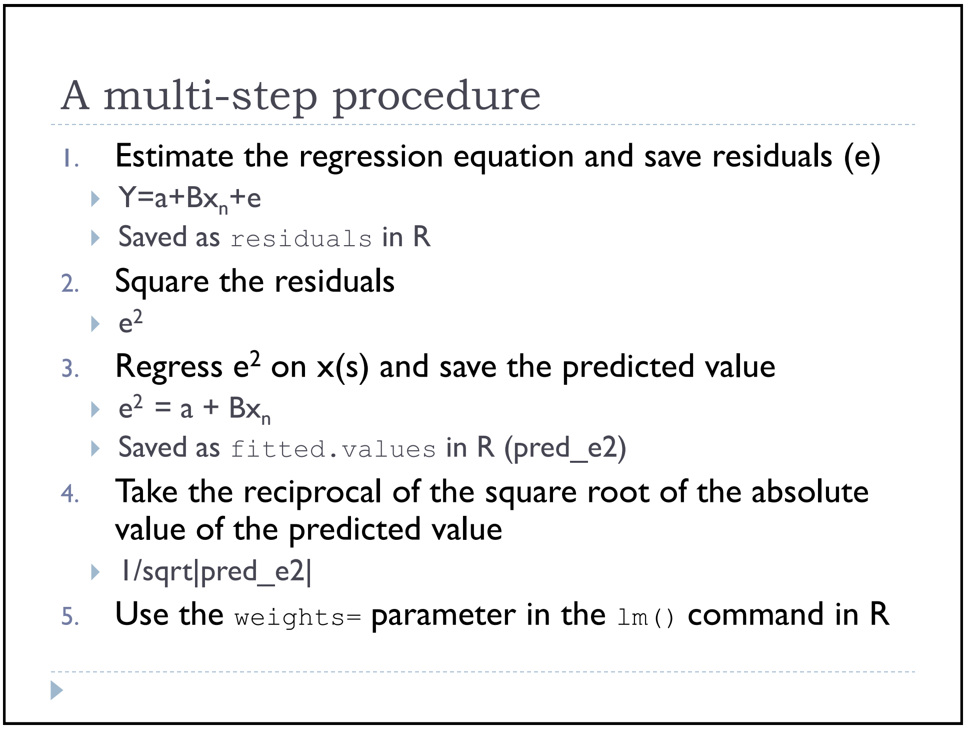
Heteroskedastic
(Hour-Glass Distribution)

How do we correct for heteroskedasticity?

- But we can also use Weighted Least-Squares (WLS) Regression
- The logic of WLS Regression:
  - Find a weight (w_i) that can be used to modify the influence of large errors on the estimation of the 'best' fit values of:
    - The regression constant (a)
    - The regression coefficients (b_n)
- OLS is designed to minimize $\sum(y-\hat{y})^2$
- In WLS, values of a and b_n are estimated to minimize $\sum w_i(y-\hat{y})^2$
  - Minimizes the influence of an observation with a large error on the estimation of a and b_n
  - Maximizes the influence of an observation with a small error on the estimation of a and b_n

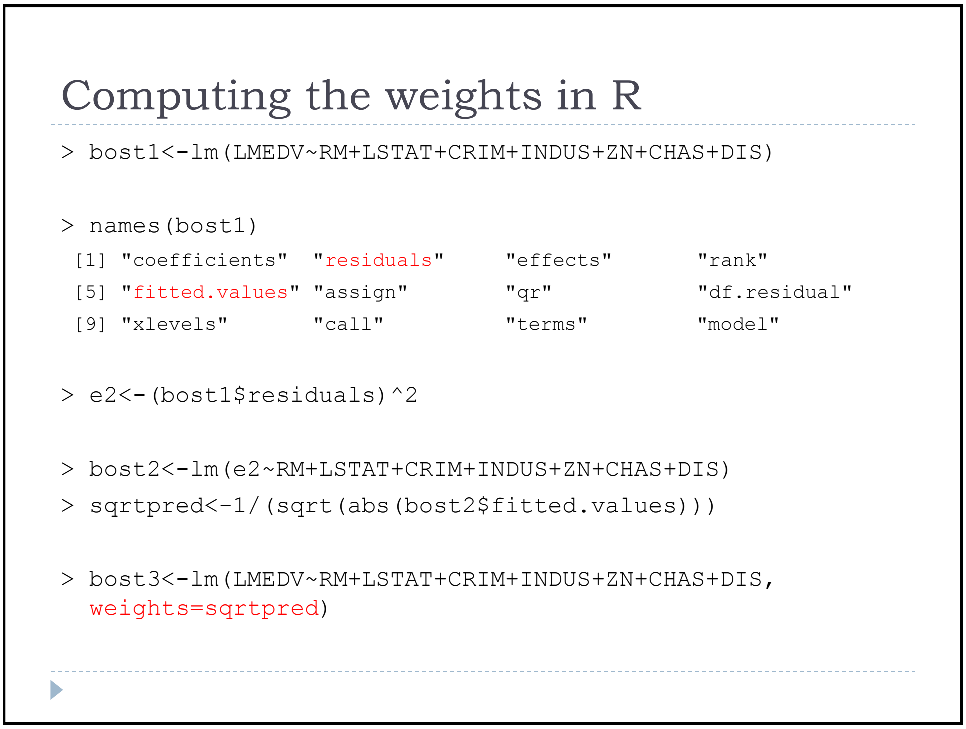
How do we determine the weight?

- From theory, the literature, or experience gained in prior research
  - Rarely will this approach prove successful, except by trial and error
- Estimate w_i by regressing e^2 on the independent variables x and weighting the values of x and y
  - This is called residualizing the variables x $w_i = \frac{1}{\hat{e}_i^2}$
- Use log-likelihood estimation to determine a suitable value of w_i



A multi-step procedure

1. Estimate the regression equation and save residuals (e)
   - $Y=a+Bx_n+e$
   - Saved as `residuals` in R
2. Square the residuals
   - e^2
3. Regress e^2 on x(s) and save the predicted value
   - $e^2 = a + Bx_n$
   - Saved as `fitted.values` in R (pred_e2)
4. Take the reciprocal of the square root of the absolute value of the predicted value
   - 1/sqrt|pred_e2|
5. Use the `weights=` parameter in the `lm()` command in R



Computing the weights in R

```
> bost1<-lm(LMEDV~RM+LSTAT+CRIM+INDUS+ZN+CHAS+DIS)

> names(bost1)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

> e2<-(bost1$residuals)^2

> bost2<-lm(e2~RM+LSTAT+CRIM+INDUS+ZN+CHAS+DIS)
> sqrtpred<-1/(sqrt(abs(bost2$fitted.values)))

> bost3<-lm(LMEDV~RM+LSTAT+CRIM+INDUS+ZN+CHAS+DIS,
  weights=sqrtpred)
```

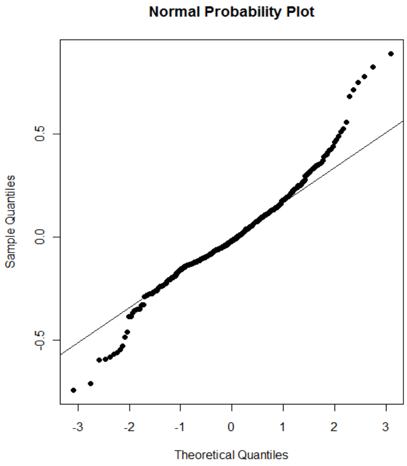
Results (compared to OLS)

```
Call:
lm(formula = LMEDV ~ RM + LSTAT + CRIM + INDUS + ZN +
    CHAS + DIS, weights = sqrtpred)


Residuals:
     Min       1Q   Median       3Q      Max
-1.68471 -0.27229 -0.04659  0.24892  1.79510


Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.6074931  0.1298848  20.075  < 2e-16 ***
RM           0.1677726  0.0159926  10.491  < 2e-16 ***
LSTAT       -0.0305643  0.0020397 -14.985  < 2e-16 ***
CRIM        -0.0112150  0.0015522  -7.225 1.89e-12 ***
INDUS       -0.0070068  0.0020115  -3.483 0.000539 ***
ZN           0.0011824  0.0003987   2.966 0.003162 **
CHAS         0.1262277  0.0376466   3.353 0.000860 ***
DIS         -0.0378324  0.0057583  -6.570 1.27e-10 ***
---
Residual standard error: 0.4502
Multiple R-squared:  0.76, Adjusted R-squared: 0.7566
F-statistic: 225.3 on 7 and 498 DF,  p-value: < 2.2e-16
```

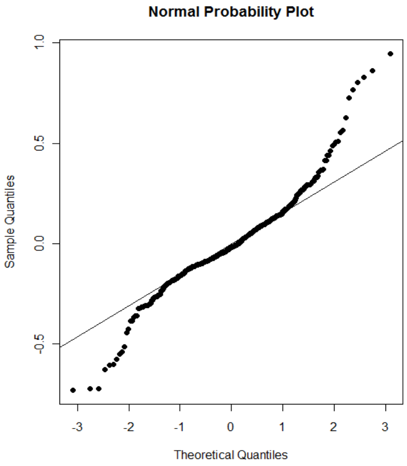
```
Call:
lm(formula = LMEDV ~ RM + LSTAT + CRIM + INDUS + ZN +
    CHAS + DIS)


Residuals:
     Min       1Q   Median       3Q      Max
-0.74280 -0.11824 -0.01867  0.10990  0.88834


Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.0381747  0.1394897  21.781  < 2e-16 ***
RM           0.1075084  0.0174328   6.167 1.44e-09 ***
LSTAT       -0.0322258  0.0020160 -15.985  < 2e-16 ***
CRIM        -0.0110130  0.0012525  -8.793  < 2e-16 ***
INDUS       -0.0086536  0.0021340  -4.055 5.82e-05 ***
ZN           0.0017965  0.0005511   3.260 0.001192 **
CHAS         0.1439320  0.0370502   3.885 0.000116 ***
DIS         -0.0436777  0.0074830  -5.837 9.59e-09 ***
---
Residual standard error: 0.2077
Multiple R-squared: 0.7453, Adjusted R-squared: 0.7417
F-statistic: 208.2 on 7 and 498 DF,  p-value: < 2.2e-16
```

Results

OLS Regression



WLS Regression

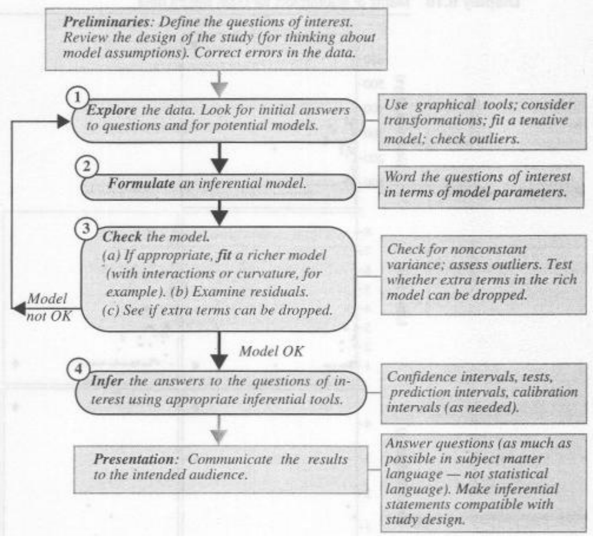


A comment on regression vs. ANOVA

- We use analysis of variance to help us compare the means for several groups (low, middle, high income)
  - Determine if the group variable explains a sufficient amount of the variation in the data

- We use regression to help us determine if there is some kind of linear relationship between an explanatory and response variable
  - Determine if the explanatory variable explains a sufficient amount of the variation in the data

A strategy for data analysis using statistical models



***Preliminaries**: Define the questions of interest. Review the design of the study (for thinking about model assumptions). Correct errors in the data.*

1 ***Explore** the data. Look for initial answers to questions and for potential models.*

Use graphical tools; consider transformations; fit a tenative model; check outliers.

2 ***Formulate** an inferential model.*

Word the questions of interest in terms of model parameters.

3 ***Check** the model.*
*(a) If appropriate, **fit** a richer model (with interactions or curvature, for example). (b) Examine residuals.*
(c) See if extra terms can be dropped.

Check for nonconstant variance; assess outliers. Test whether extra terms in the rich model can be dropped.

Model not OK

Model OK

4 ***Infer** the answers to the questions of interest using appropriate inferential tools.*

Confidence intervals, tests, prediction intervals, calibration intervals (as needed).

***Presentation**: Communicate the results to the intended audience.*

Answer questions (as much as possible in subject matter language — not statistical language). Make inferential statements compatible with study design.

Source: Ramsey and Shafer, pp. 251