Some Notes on Regression Model Building

"Essentially, all models are wrong, but some are useful."

--- Box, George E. P.; Norman R. Draper (1987). Empirical Model-Building and Response Surfaces, p. 424, Wiley. ISBN 0471810339.

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." John W. Tukey 1962 The future of data analysis. Annals of Mathematical Statistics 33: 1-67 (see pp.13-14)

Modeling is difficult, and there often is more than one answer to a set of data. In this set of notes, we try to give a basic framework for how to attack a regression modeling problem. Some items are left out since we wont cover them until class next Monday.

Step 1: Look at univariate plots of the Y and all the X variables. This means a dotplot, histogram or boxplot of every variable you have. You're looking for data entry errors, and obvious weird points (outliers). If a point seems extreme, consider removing it.

Step 1b: We haven't covered this topic yet, but if the Y variable is positively skewed, as income and real estate prices often are, we want to log the Y variable to make it more symmetric. Normally we don't like to modify the Y variable but there are exceptions to every rule.

Step 2: We haven't covered this yet also, but we don't want our X variables highly correlated with each other (we want them highly correlated with Y of course). Compute the correlation matrix of all the X variables and look for correlations above 0.8 in absolute value. If say variables X2 and X4 have a correlation of 0.9, you don't need both in the model and can just use one (probably the one with a higher correlation with Y).

Step 3: Examine scatter plots of Y versus each X. By its ubiquitous name, linear regression requires the relationship between Y and each X variable to be linear! If the plot of Y versus X does not look linear, you want to see if you can find a transformation to X to make it a linear relationship (common ones are log(x), sqrt(x), 1/x). It might also appear that there is explosive growth in the relationship between Y and X and you will need higher powers of X (such as X^2).

Step 4: Fit a big model with your current crop of X variables (whatever came out of Step 2). Calculate the residuals from this model and look at a histogram of the residuals. The histogram should look normal or at least symmetric. If it doesn't and you haven't already, log the Y variable and try this step again. Look also for outlier values of Y.

Step 5: Run backward stepwise regression (automatically in Stata using the sw regression command).

Step 6: From the final model, do the following:

1) Run the rvfplot command (residuals versus fitted) to look at the over all residual diagnostic plot. Does it look random? Is there a pattern? If there is a pattern you will need to plot residuals versus each X to find the problem.

2) Are any standardized residuals larger than 2 in absolute value? You should expect up to 5% of the residuals to be flagged as outliers, but examine the larger ones to see if those observations are truly outliers and can possibly be eliminated.
3) Examine the residuals to make sure they are symmetric. You can run sktest on the residuals [null hypothesis is normality]
4) You can see if you need any transformations by running the ovtest [null hypothesis is no transformations of the X variables are needed]
5) If you delete observations or transform variables run Step 6 again.

Some Definitions and Further Notes

Residuals: $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_{1i} + \cdots + b_k X_{ki})$

Plots (see prototype plots in class):

- Plot of e_i vs $\hat{Y}_i$ Can be used to check for linear relation, constant variance
  - If relation is nonlinear, U-shaped pattern appears
  - If error variance is non constant, funnel shaped pattern appears
  - If assumptions are met, random cloud of points appears

- Plot of e_i vs X_{ji} for each j Can be used to check for linear relation with respect to X_j
  - If relation is nonlinear, U-shaped pattern appears
  - If assumptions are met, random cloud of points appears

- Histogram of e_i
  - If distribution is normal, histogram of residuals will be mound-shaped, around 0

Useful Stata Commands:

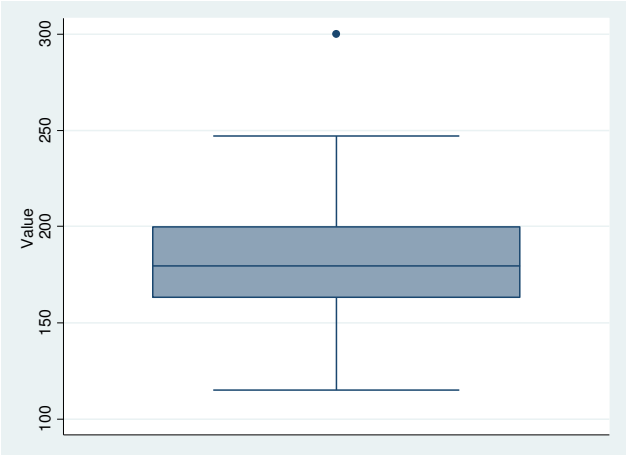
Stata Command	What it does
`scatter y x`	**Plot of y versus x**
`corr x1 x2 x3 x4`	**Generate a correlation matrix of the x variables**
`regress y x1 x2 x3 x4`	**Runs a multiple regression**
`sw regress y x2 x3 x3 x4,pr(.05)`	**Backward stepwise**
`predict yhat`	**Computes the fitted values**
`predict res,r`	**Computes the residuals**
`generate sres=res/[put s_e value here]`	**Computes the standardized residuals**
`rvfplot`	**Produces the residuals versus fitted plot**
`scatter sres x1`	**Residual plot of residuals versus x**
`sktest sres`	**Normality test for the residuals**
`ovtest`	**Test if any transformations needed**
`hettest`	**Tests for heteroskedasticity**

Example

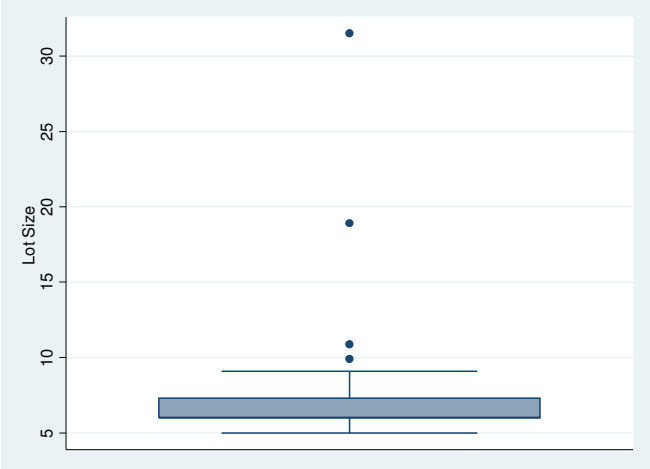
[Note-this is a quick runthough of a data set-more like stream of consciousness. It is provided to give you a sense of hwo to do this in Stata and what to look for.]

Using dataset houseexample.xls we will model the Value of a house as a function of the variables: Lot Size, Beds, Baths, Rooms, Age, and, Taxes.

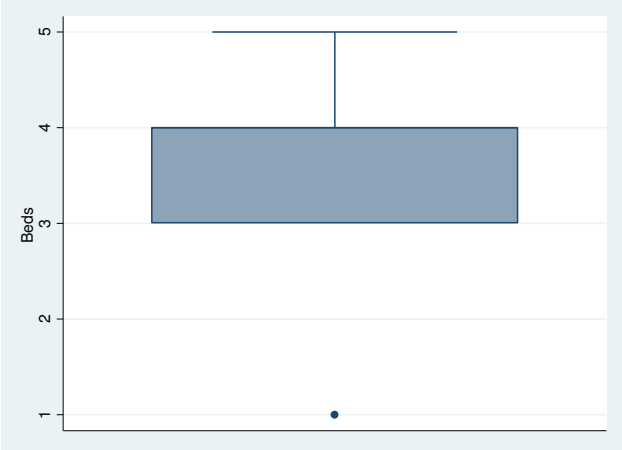
Boxplot of Value: Looks symmetric though there is an outlier-may want to investigate that point.



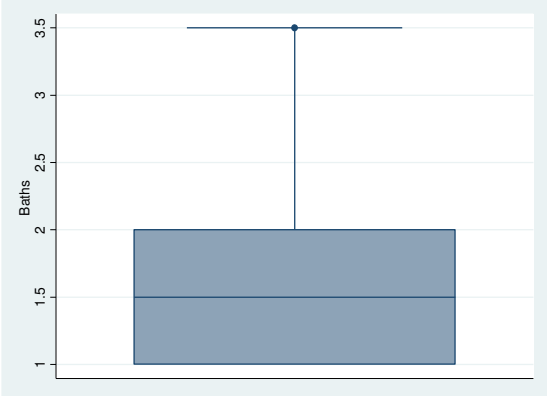
Boxplot of lotsize: A few outliers that bear looking into.



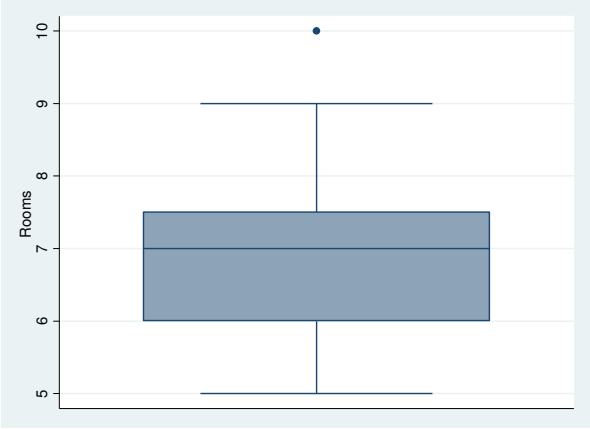
Boxplot of beds: That one bedroom house sure looks unusual:



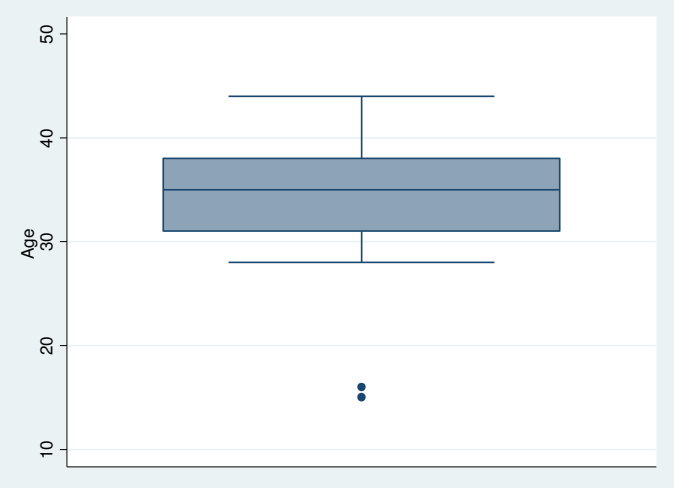
Boxplot of baths:



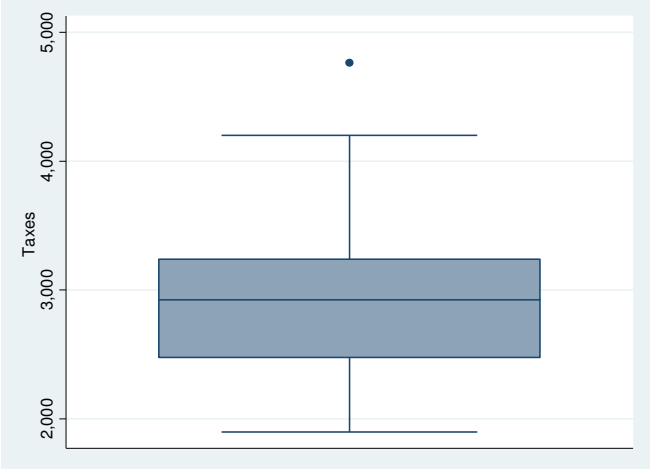
Boxplot of rooms: The ten room house is an outlier



Boxplot of age. Curious-two homes are much younger than the rest of the data set. Might have to see what their corresponding prices are.



Boxplot of taxes: One outlier. Taxes are related to size of the house and lot, so also need to check the correlation of the X variables.



We are not going to drop any observations at this time, but if for example, you wanted to drop the observation with the extreme tax value, you would do the following command in Stata,

```
. drop if taxes> 4500

(1 observation deleted)
```

We will now examine the correlation matrix of all the X variables to see if any of them are highly related:

```
. corr lotsize beds baths rooms age taxes
(obs=40)

             |  lotsize     beds    baths    rooms      age    taxes
-------------+------------------------------------------------------
     lotsize |   1.0000
        beds |   0.1902   1.0000
       baths |   0.0656   0.3101   1.0000
       rooms |   0.3247   0.5081   0.5254   1.0000
         age |  -0.3652   0.1421  -0.3924  -0.2173   1.0000
       taxes |   0.2569   0.1699   0.6955   0.5610  -0.4046   1.0000
```

One rule of thumb is if an absolute value of a correlation is above 0.80, we don't need both variables in the model. Not surprisingly, taxes and baths are highly correlated, but not high enough to warrant eliminating one right now, so we will keep them all in.

We will now do backward stepwise regression:

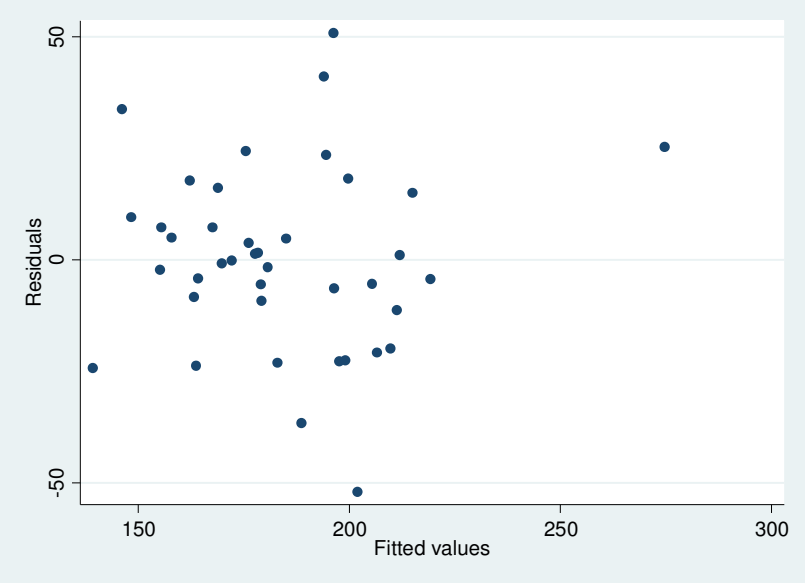
```
. sw regress value lotsize beds baths rooms age taxes,pr(.05)
                    begin with full model
p = 0.3002 >= 0.0500  removing age
p = 0.2011 >= 0.0500  removing rooms
p = 0.1139 >= 0.0500  removing beds

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  3,    36) =   18.04
       Model |   24926.446     3  8308.81535           Prob > F      =  0.0000
    Residual |  16582.2105    36  460.616959           R-squared     =  0.6005
-------------+------------------------------           Adj R-squared =  0.5672
       Total |  41508.6566    39  1064.32453           Root MSE      =  21.462

------------------------------------------------------------------------------
       value |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lotsize |  -1.782171   .7992875    -2.23   0.032    -3.403201   -.1611406
       taxes |   .0279309   .0081516     3.43   0.002     .0113987    .0444632
       baths |   17.97271   8.270458     2.17   0.036     1.199446    34.74598
       _cons |   88.03484   17.12445     5.14   0.000     53.30485    122.7648
------------------------------------------------------------------------------
```

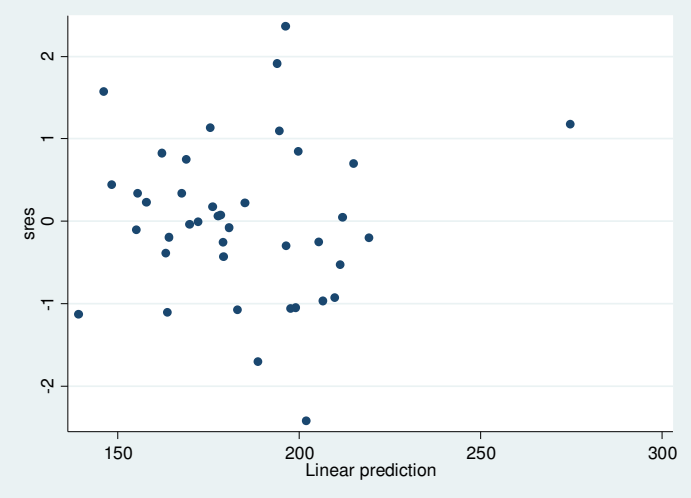
Three variables are removed and three remain. We will now look at some diagnostics.

The quickest diagnostic is to simply do the Stata command, `rvfplot.` There is one extreme fitted value, but otherwise the plot doesn't seem to have any of the unusual shapes we know signal issues.



We should create the standardized residuals and look for outliers.

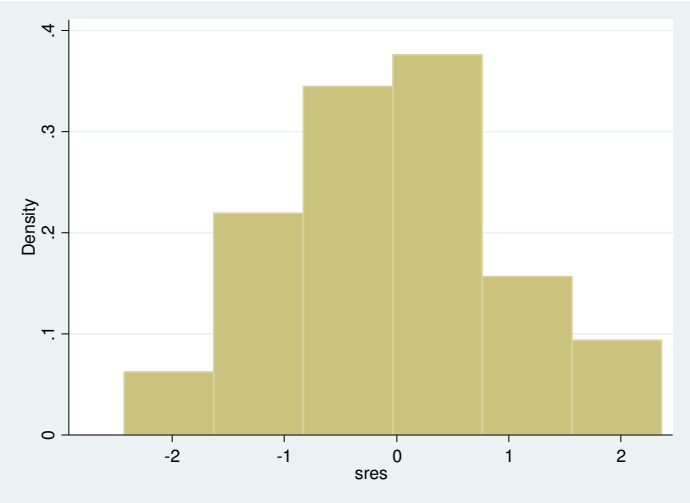
```
. predict yhat,xb
. predict res,r
. generate sres=res/21.462
. scatter sres yhat
```



Only two outliers (standardized residuals larger than 2 in absolute value). For now we will keep them but you could drop them and see if you get a better fit.

We also need to look at a histogram of the residuals to see if they look normal. If not we would log the Y variable and start all over with stepwise. It looks fine here.

```
. hist sres
```



We can formally test if the residuals are normally distributed using the Stata command `sktest`:

```
. sktest sres

                    Skewness/Kurtosis tests for Normality
                                                              ------- joint ------
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------------
        sres |     40      0.8117         0.3431         1.00          0.6059
```

The null hypothesis is normality, and we fail to reject the null due to the large pvalue (0.6059). If normality is rejected, we would log the Y variable and begin again with a new stepwise regression.

So far everything appears ok in this model. We can run the Stata command `ovtest`, which tests the null hypothesis that transformations of the X variable are required.

```
. ovtest

Ramsey RESET test using powers of the fitted values of value
       Ho:  model has no omitted variables
                  F(3, 33) =      0.88
                  Prob > F =      0.4615
```

The large pvalue here (.4615) means we don't reject the null, and no transformations of the X variables are indicated. If we obtained a low pvalue here (below 0.05), we would have to investigate the individual residual diagnostic plots to see which variable requires a transformation.

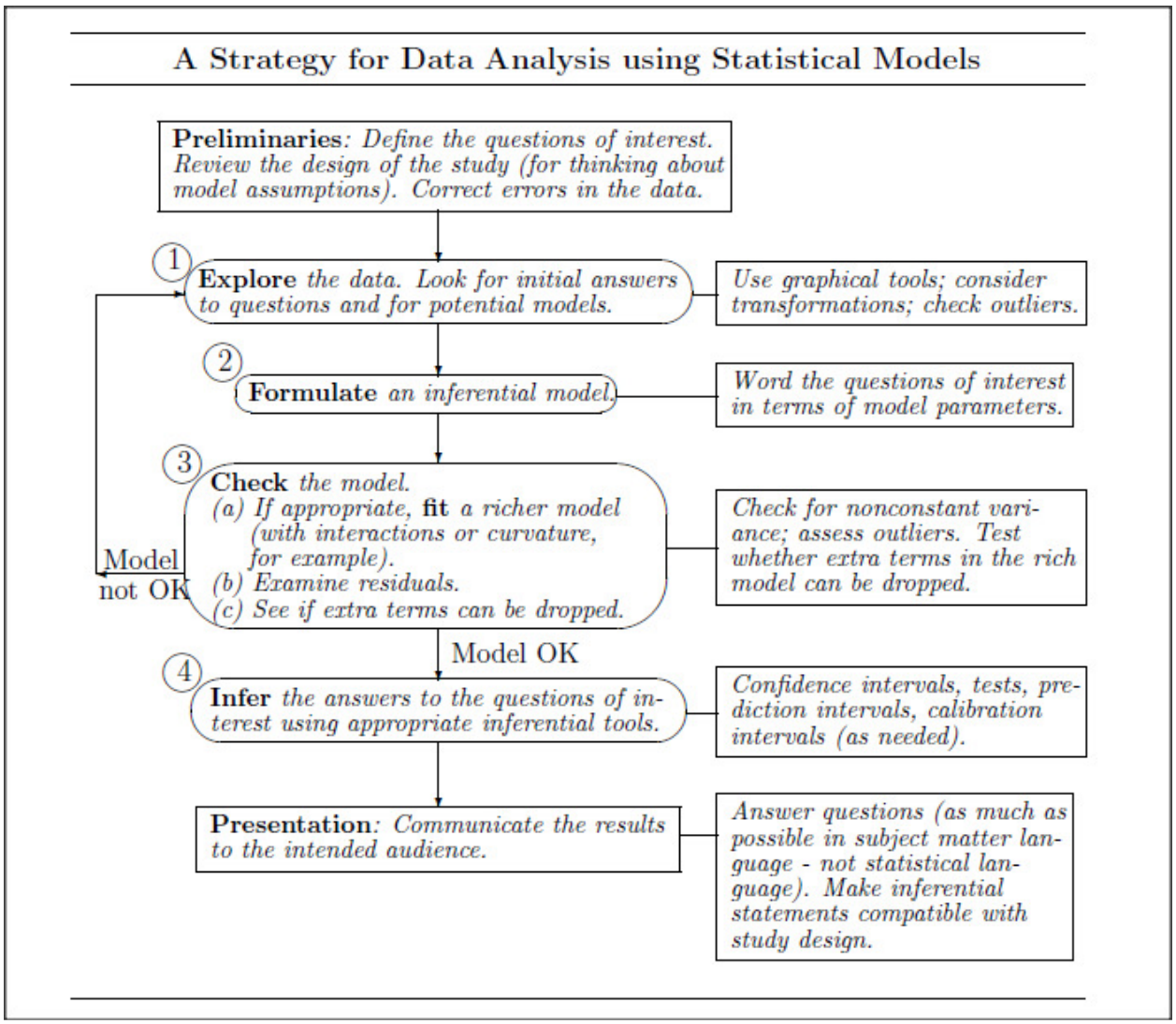
At this point the model looks fine and we could interpret the coefficients and make predictions.

Some additional reading and observations.

The book chapter "Model Building with Multiple Regression" (available on course website under the project tab) is from an introductory book by Alan Agresti, and a useful guide for those students wanting a little more guidance. It goes into a little more detail with some diagnostics than we have time for.

The paper "On Strategies in Regression Analysis" (available on course website under the project tab) is a philosophical treatise on modeling, written by a long time statistician with much experience. The paper is too deep for our purposes, but the general guidelines given on page 7-9 of the paper are useful. I recommend a brief read of the strategy though parts of it may not make complete sense.

The following diagram is from the book "The Statistical Sleuth" which is used in Stat 139. As you can see, regression modeling is an iterative process. Step 3 in the flowchart is what we tried to describe in the previous pages.



ditional reading and observations.