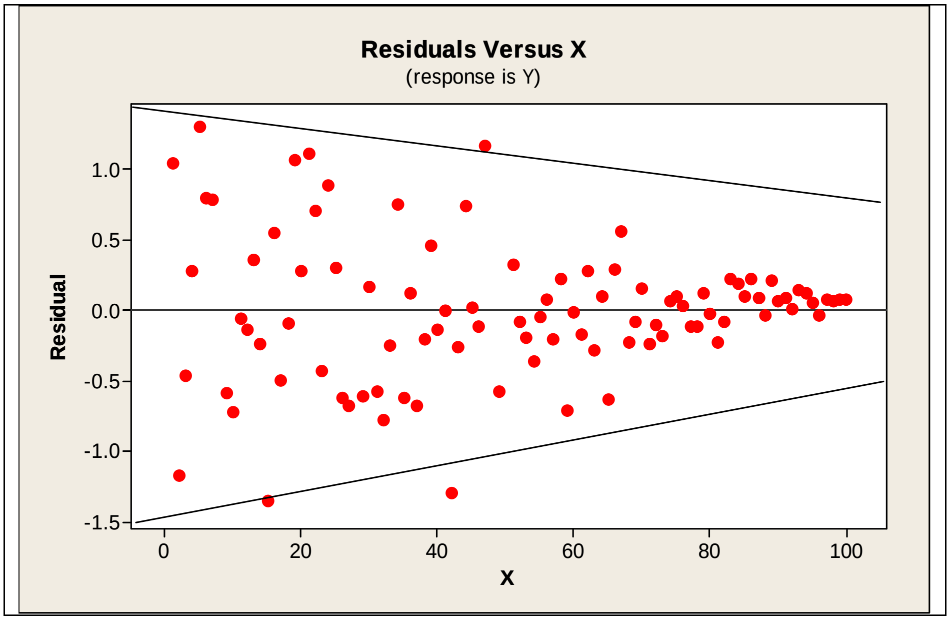
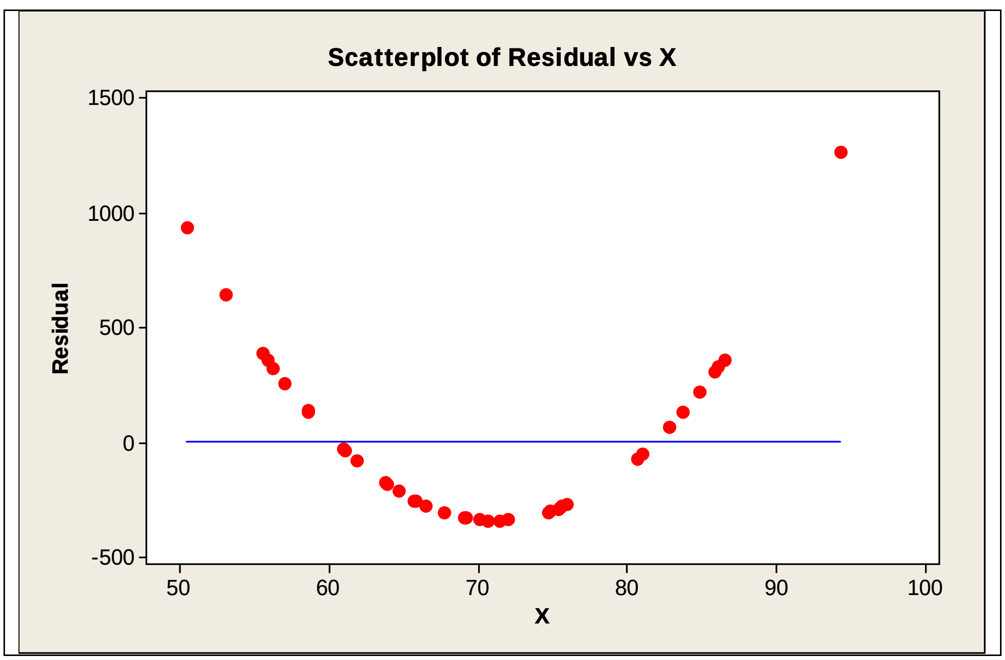
Homework 3 Solutions:
Mon, Oct 12: Chapter 3: #2, 9, 18, 20
Wed, Oct 14: Chapter 2: #57 and Chapter 3: #15, 16a

Assigned Mon, Oct 12:

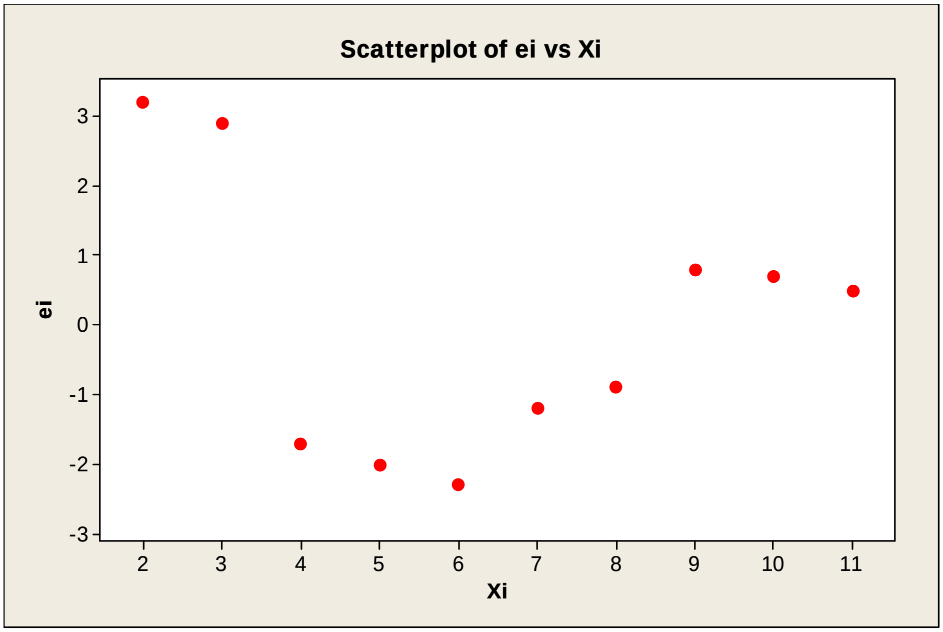
3.2 **(1)** Prototype of plot in which error variance decreases with X would be like the following picture:



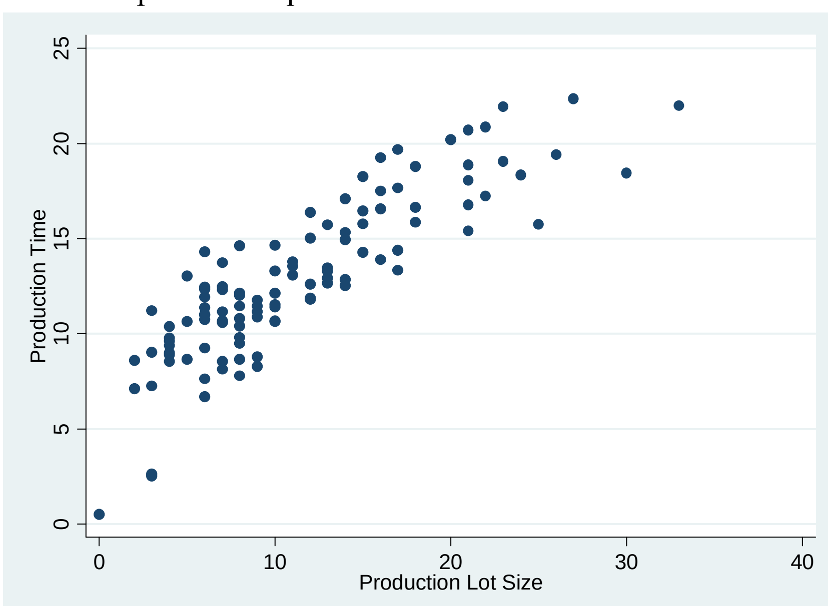
(2) For the second situation, the residual plot (of residuals versus either X or the fitted values) for the situation in which the true regression is U shaped but linear regression is fit would be U-shaped as well. Here is an extreme example:



3.9 The plot of residuals (e_i) versus X_i is below. The problem appears to be two large outliers at the X values of 2 and 3. A transformation would not help. Those houses have large positive residuals, which means that the actual electricity consumption was higher than we would expect based on the small number of rooms. Perhaps houses with only a total of 2 or 3 rooms fit a different model, in which case it could be argued that they should be removed from the dataset.



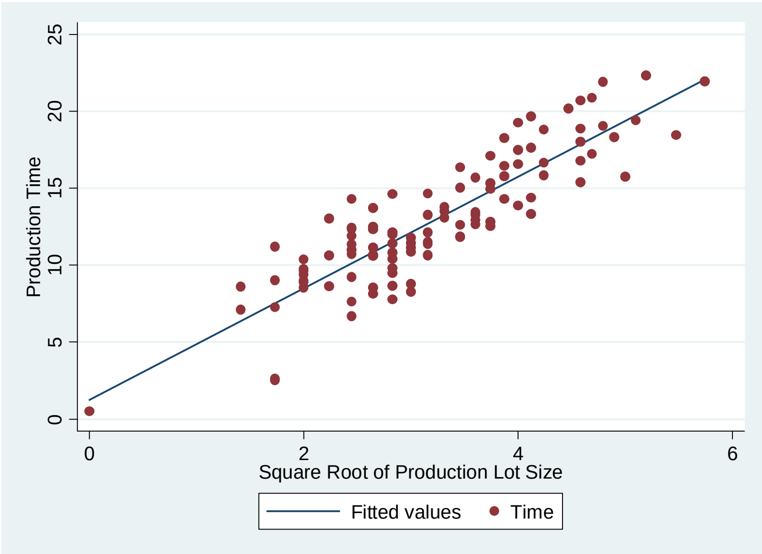
3.18 **a.** Here is a scatter plot of Y = production time in hours versus X = lot size:



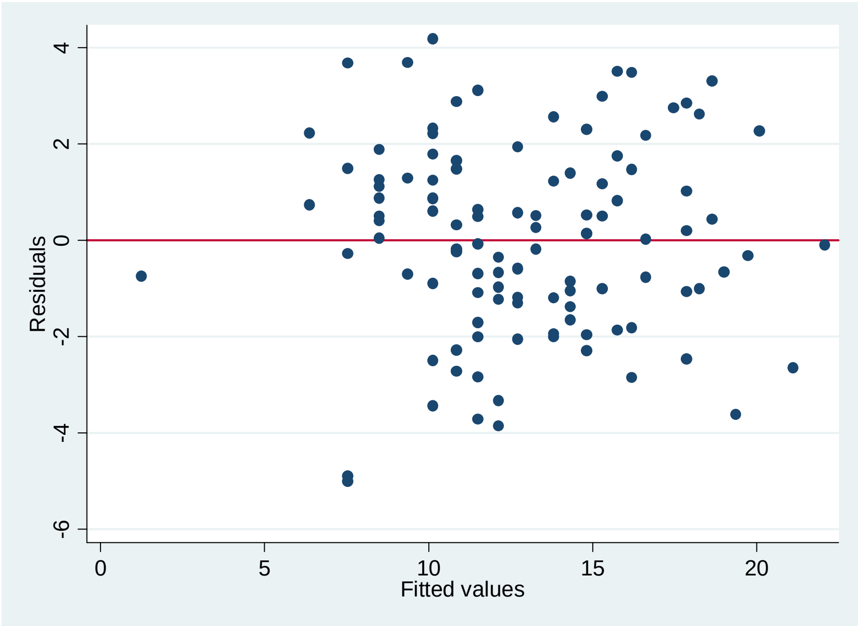
A linear relation does not appear to be adequate here. The regression relation in the scatter-plot appears to be curvilinear. (This would be even more obvious with a residual plot.) The variability across different X levels appears to be fairly constant, thus a transformation to X is more suitable.

b. $\hat{Y} = 1.2547 + 3.6235\,X'$, where $X' = \sqrt{X}$.

c. Yes, the scatter-plot (below) shows a reasonably linear relationship, the estimated regression line appears to be a good fit to the transformed data.

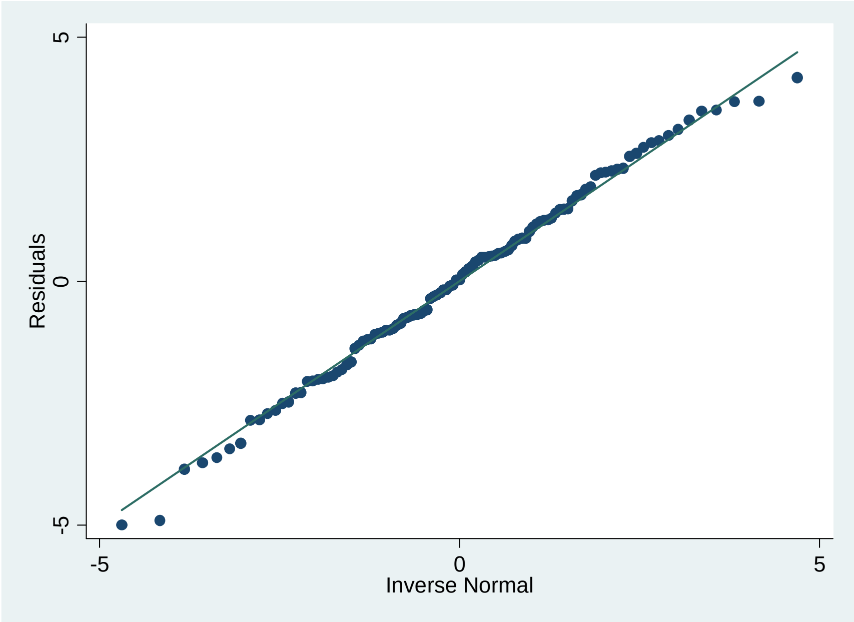


d. To plot the residuals versus fitted values in Stata, use rvfplot, yline(0):



The residuals vs. fitted values plot shows that points are spread out without a systematic pattern, implying that the model fit is reasonable; points fall within a horizontal band centered around 0, so the error variances appear to be stable. Two possible outliers are detected on the bottom left side of the plot, between fitted values of 5 and 10. The other apparent outlier near fitted value of 1 isn't a problem because it's residual is small.

The normal probability plot (below) shows that points fall reasonably close to a straight line, with very small tails deviating from the line, suggesting that the distribution of the error terms is approximately normal.



e. $\hat{Y} = 1.2547 + 3.6235\sqrt{X}$ where X is the lot size and Y is time in hours.

3.20 The error terms after the transformation $X' = 1/X$ will still be normally distributed because changing the X-axis doesn't change the vertical spread. The error terms after the transformation $Y' = 1/Y$ will not be normally distributed.

Assigned Wed, Oct 14:

2.57 **a.** The reduced model is $Y_i = \beta_0 + 5X_i + \varepsilon_i$ or equivalently (and necessary if you wanted to run this in a software package) $Y_i - 5X_i = = \beta_0 + \varepsilon_i$. The df = n – 1 because only one parameter is being estimated.
b. This seems like a silly model to test, but it would be $Y_i = 2 + 5X_i + \varepsilon_i$ or equivalently (and necessary if you wanted to run this in a software package) $Y_i - 2 - 5X_i = \varepsilon_i$. The df = n because no parameters are being estimated.

3.15 **a.** The regression equation, from the Stata output below, is $\hat{Y} = 2.575 - .324X$.

```
------------------------------------------------------------------------------
    Solution |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Time |      -.324   .0432987    -7.48   0.000    -.4175412   -.2304588
       _cons |   2.575333   .2487321    10.35   0.000      2.03798    3.112686
------------------------------------------------------------------------------
```

b. The results below were obtained using the Stata maxr2 command. Clearly we reject the null hypothesis (linear regression model) because the p-value is 0.0000. We conclude that *there is* lack of fit for the linear regression model, and that the individual means are a much better fit.

```
SSLF (df) = 2.767253 (3) MSLF = .92241767
SSPE (df) = .15739998 (10) MSPE = .01574

F (dfn, dfd) for lack-of-fit test (MSLF/MSPE) =   58.6034 (3,10)
                                     prob > F =    0.0000
```

c. The test does not indicate what regression function is appropriate. It simply tells us that the linear regression model is not a good fit. To find out what model to try next, we need to look at a scatter plot (which we now do for Problem 3.16a.)

3.16 **a.** Based on the scatter plot below and the pictures shown in Figure 3.15, it appears that a log transformation may be appropriate. It looks somewhat like the middle picture in Figure 3.15.