Start with review, some new definitions, and pictures on the white board.

Assumptions in the Normal Linear Regression Model

A1: There is a ***linear*** relationship between X and Y.

A2: The error terms (and thus the Y's at each X) have ***constant variance***.

A3: The error terms are ***independent***.

A4: The error terms (and thus the Y's at each X) are ***normally distributed*****.**
Note: In practice, we are looking for a fairly symmetric distribution with no major outliers.

Other things to check (**Q**uestions to ask):

Q5: Are there any major ***outliers*** in the data (X, or combination of (X,Y))?

Q6: Are there ***other possible predictors*** that should be included in the model?

Useful Plots for Checking Assumptions and Answering These Questions

Reminders:

Residual = $e_i = Y_i - \hat{Y}_i$ = observed Y_i – predicted Y_i

where predicted Y_i = $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, also called "fitted Y_i"

Definitions and formulas; we will use R to compute these:

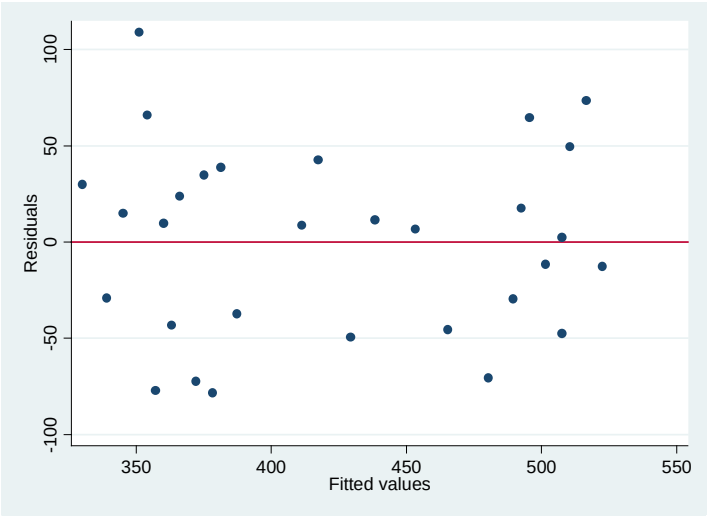
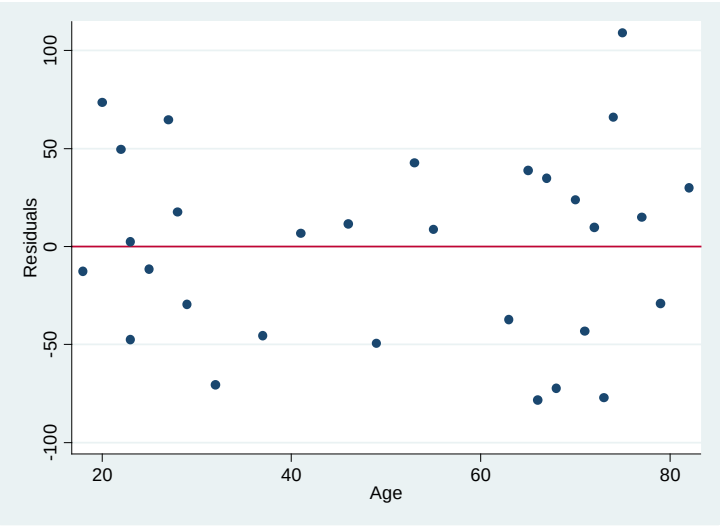
Standardized residual for unit *i* is $e_i^* = \frac{e_i}{standard\ error(e_i)} = \frac{e_i}{\sqrt{MSE}(1-h_i)}$

h_i defined later in course, but usually is close to 0, so denominator $\approx \sqrt{MSE}$

Externally studentized residual r_i for unit *i* is the same, except use MSE from the model fit *without* unit *i*.

Plot	Useful for
Dotplot, stemplot, histogram of X	**Q5** Outliers in X; range of X values
Residuals e_i versus X_i or predicted $\hat{Y}_i$	**A1** Linear, **A2** Constant var., **Q5** outliers
e_i* or r_i versus X_i or predicted $\hat{Y}_i$	As above, but a better check for outliers
Dotplot, stemplot, histogram of e_i	**A4** Normality assumption
Residuals e_i versus time (if measured)	**A3** Dependence across time
Residuals e_i versus other predictors	**Q6** Predictors missing from model
"Normal probability plot" of residuals	**A4** Normality assumption

Example: Highway sign data
Plot of:
Residuals versus predicted ("fitted") values
Residuals vs Age
NOTE: Plot of residuals versus predictor variable X should look the same except for the scale on the X axis, because fitted values are linear transform of X's. However, when the slope is negative, one will be a mirror image of the other.

Residuals vs fitted values	Residuals vs age
	

Comments: These are good "residual plots." Points look randomly scattered around 0. No evidence of *nonlinear* pattern or *unequal variances*.

Some other plots of the residuals, used to check the "normality" assumption

Stemplot of standardized residuals

To generate them in R, for the linear model called "HWModel":

```
>Highway$StResids <- rstandard(HWModel)
```

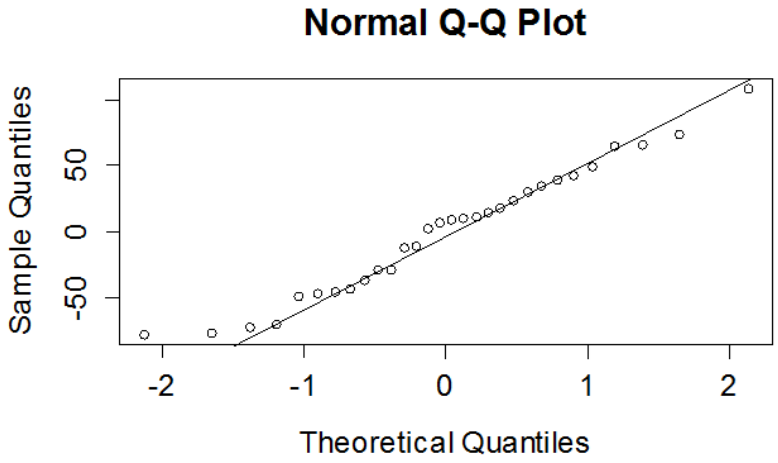
```
> stem(Highway$StResids)
The decimal point is at the |

  -1 | 6655
  -1 | 00
  -0 | 99866
  -0 | 32
   0 | 1122234
   0 | 56789
   1 | 034
   1 | 6
   2 | 3
```

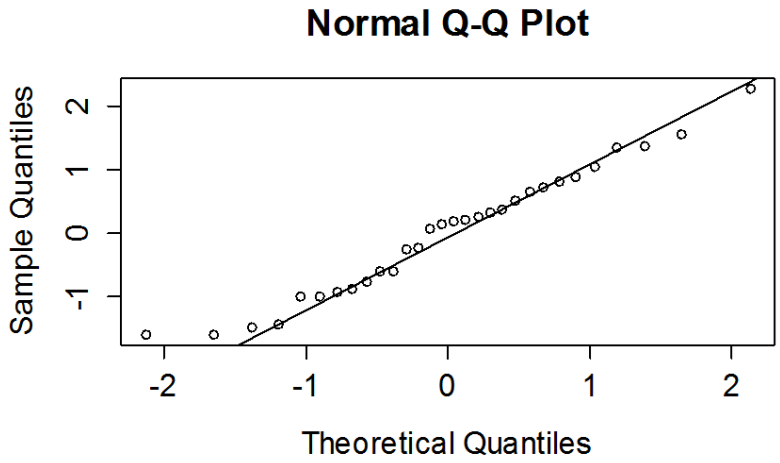
Further confirmation that the residuals are relatively symmetric with no major outliers. Residual of 2.3 is for a driver with x = 75 years, y = 460 feet, $\hat{y} = 577 - 3(75) = 352$, so residual = 460 – 352 = 108 feet.

Normal probability plot of residuals and standardized residuals for highway sign data, also to check normality assumption **A4** (see Figure 1.7 in textbook for examples with normal and quite non-normal residuals).

Using residuals

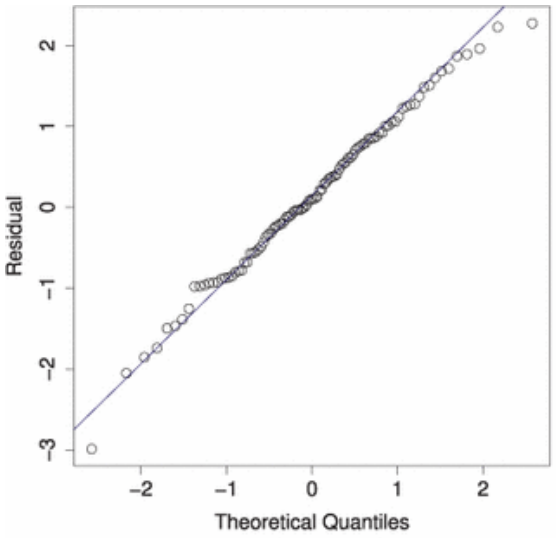


Using standardized residuals

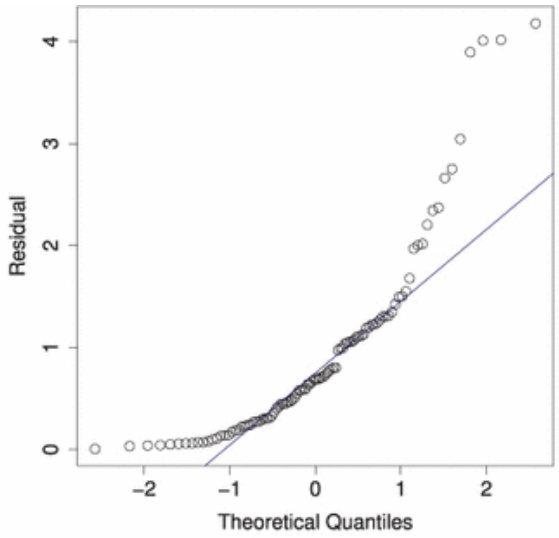


- Explanation of what these are will be given on the white board.
- These are pretty good plots. There is one point at the lower end that is slightly off, and might be investigated, but no major problems.
- Note that the plot looks the same using residuals or standardized residuals. It doesn't matter which ones you use.

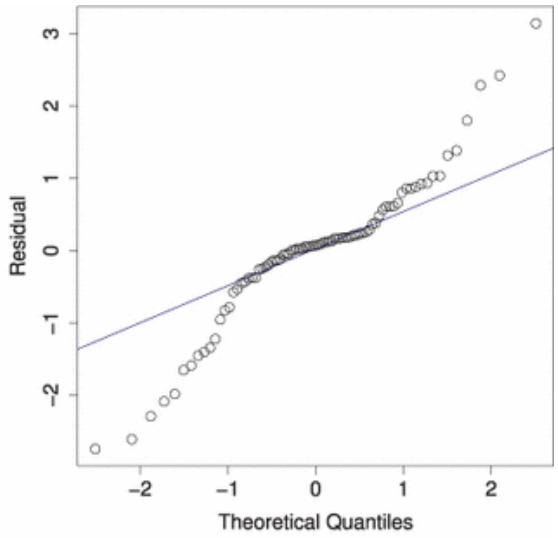
Example from textbook of good and bad normal probability plots:



(a) Normal residuals



(b) Skewed right residuals



(c) Long-tailed residuals

What to do when assumptions aren't met

Assumption 1:
Relationship is linear.

How to detect a problem:
Plot *y* versus *x* and also plot residuals versus fitted values or residuals versus *x*. If you see a pattern, there is a problem with the assumption.

What to do about the problem:
Transform the **X** values, $X' = f(X)$. (Read as "X-prime = a function of X.")
Then do the regression using X' instead of X:

$$Y = \beta_0 + \beta_1 X' + \varepsilon$$

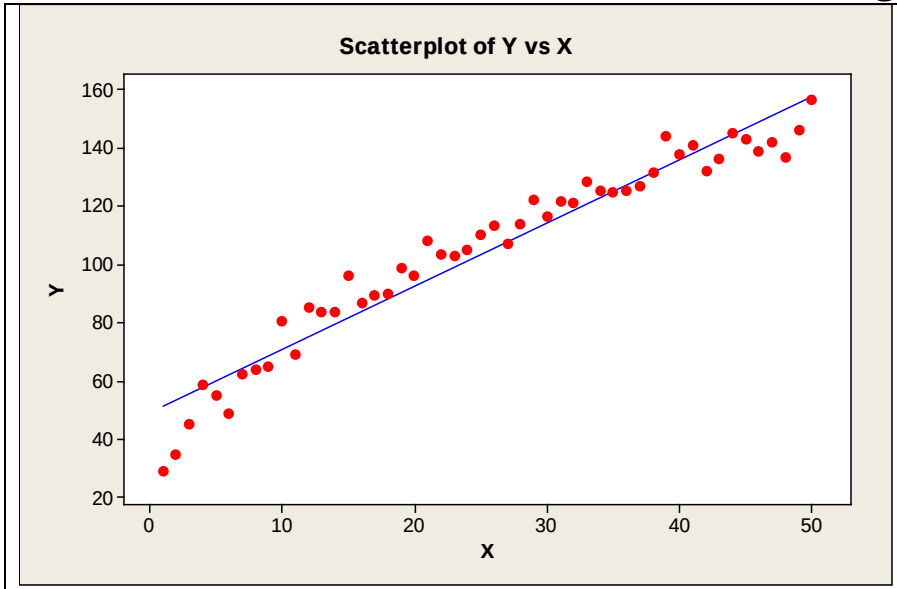
where we still assume the ε are $N(0, \sigma^2)$.

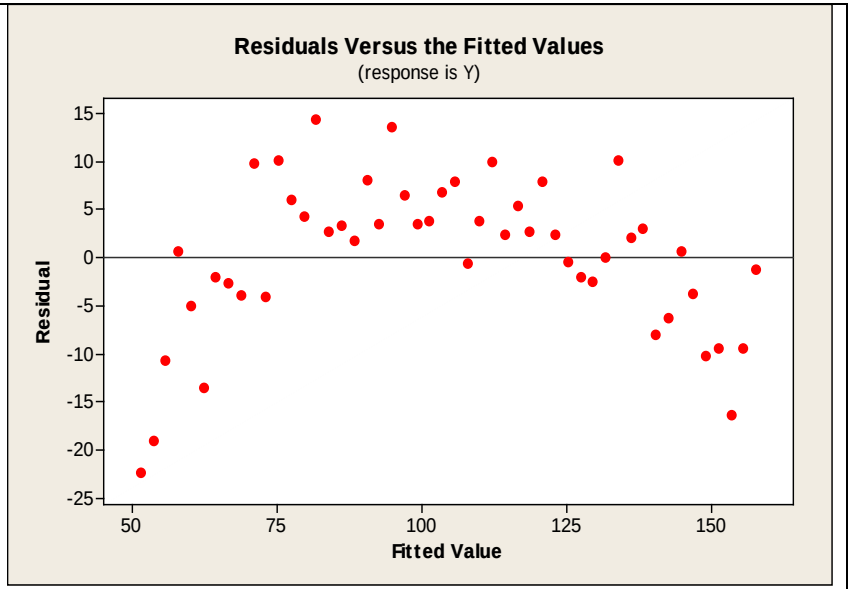
NOTE: Only use this "solution" if non-linearity is the *only* problem, not if it also looks like there is non-constant variance or non-normal errors. For those, we will transform Y.

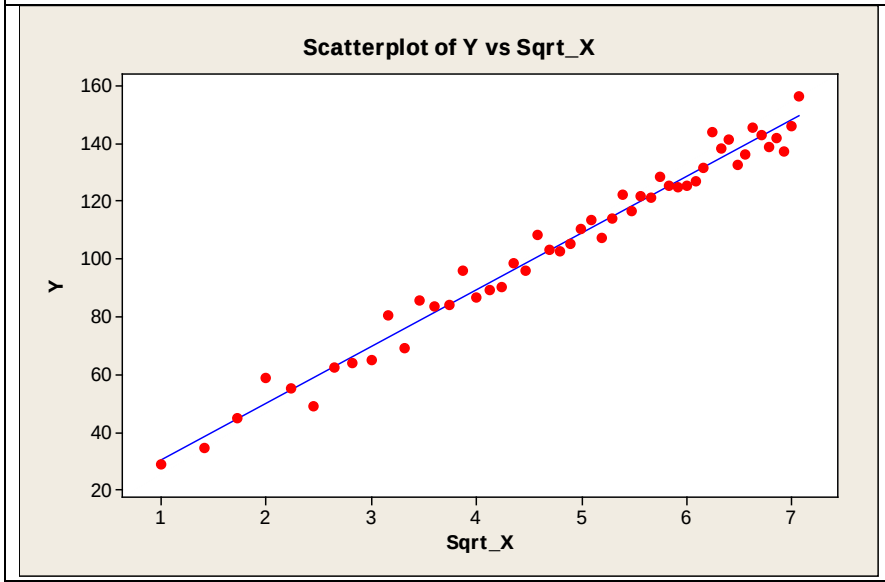
REASON: The errors are in the vertical direction. Stretching or shrinking the X-axis doesn't change those, so if they are normal with constant variance, they will stay that way.

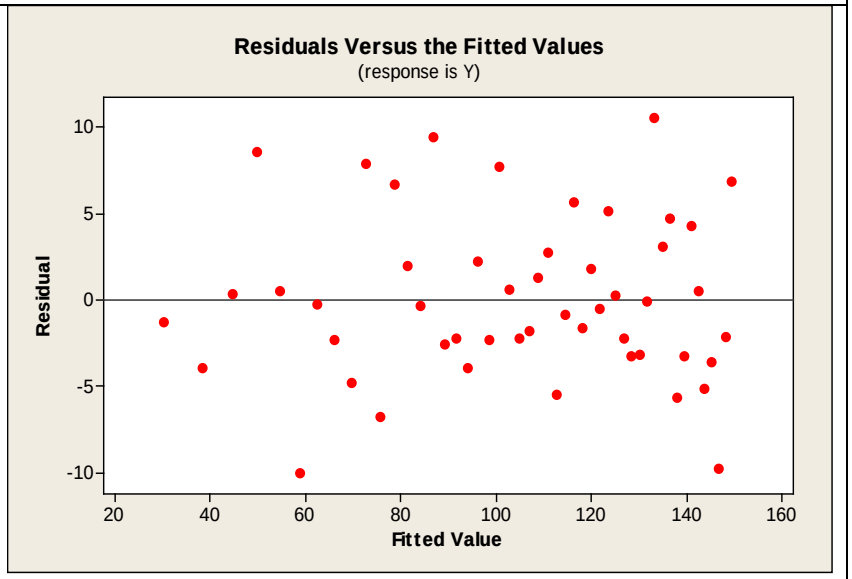
Let's look at what kinds of transformations to use.

Residuals are inverted U, use $X' = \sqrt{X}$ or $\log_{10} X$

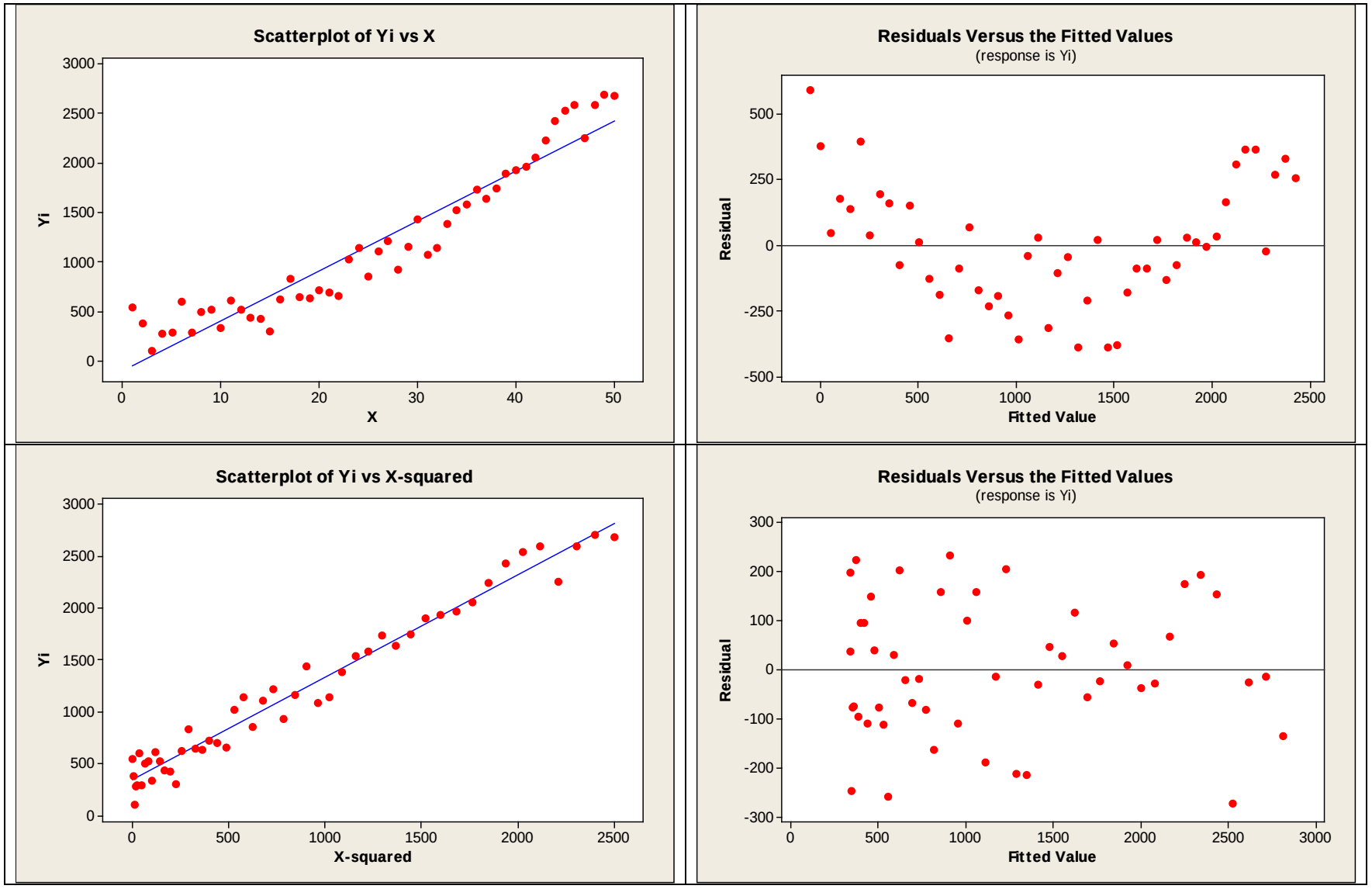




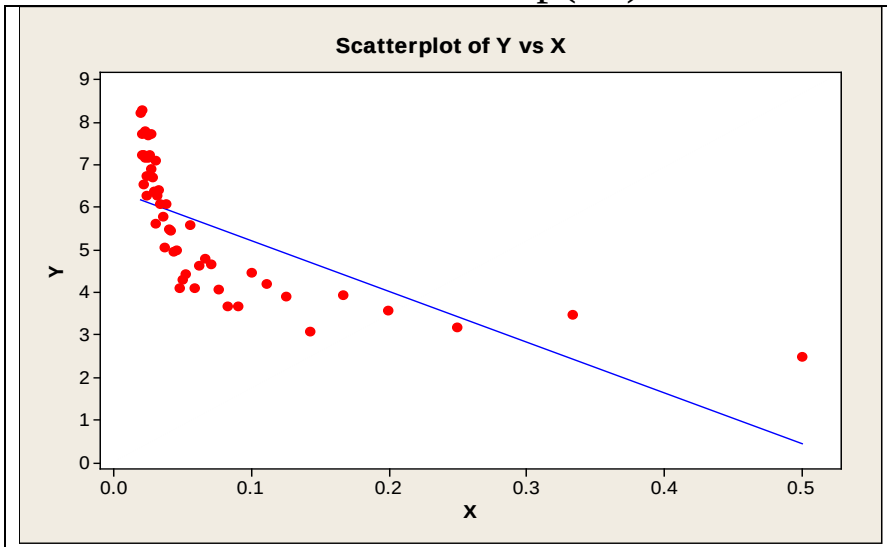


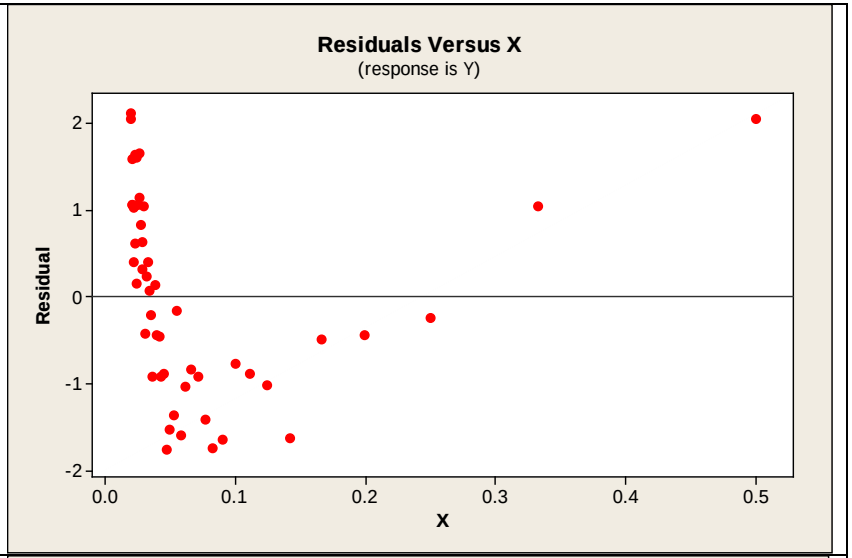


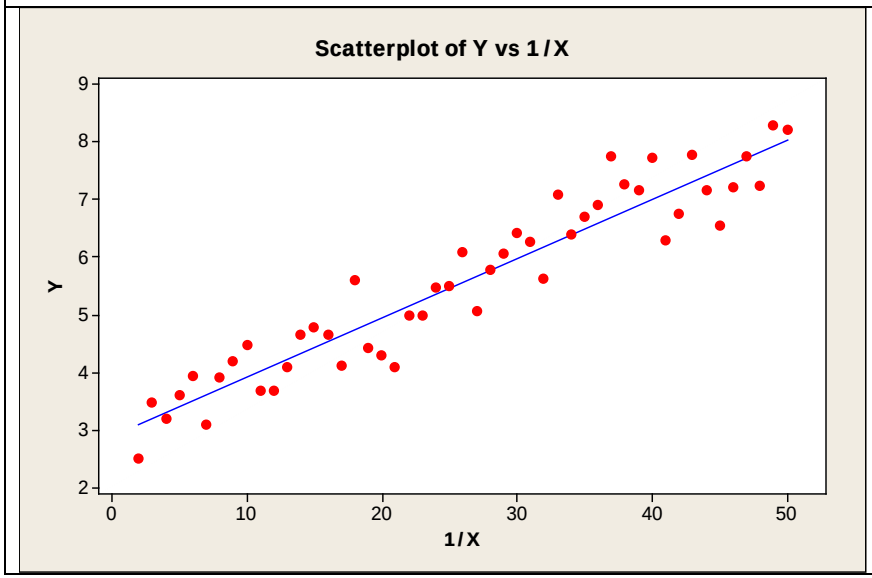
Residuals are U-shaped and association between X and Y is positive: Use $X' = X^2$

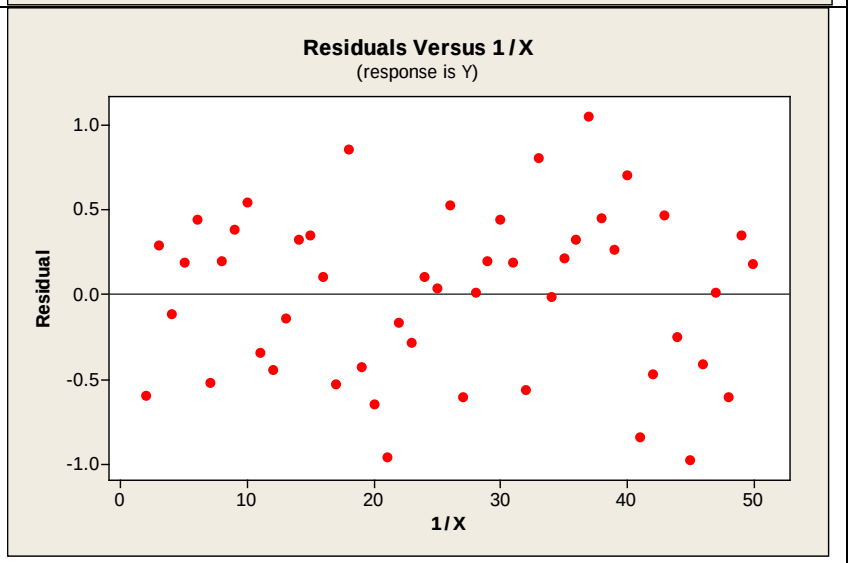


Residuals are U-shaped and association between X and Y is negative:
Use $X' = 1/X$ or $X' = \exp(-X)$





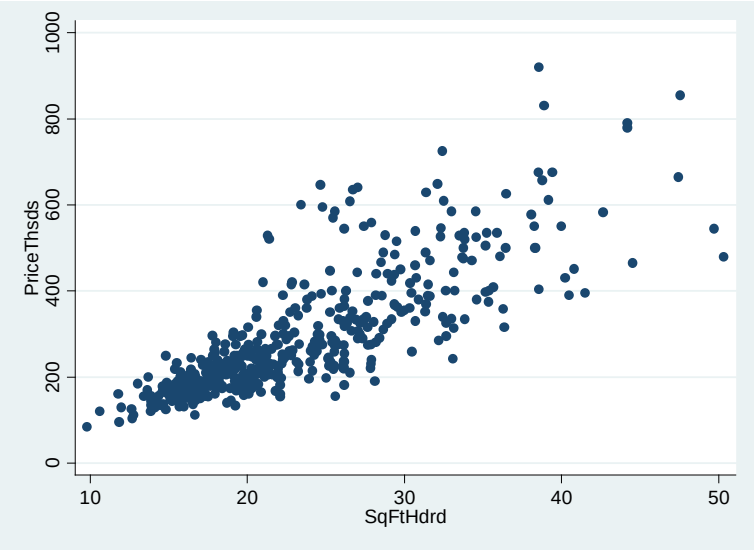
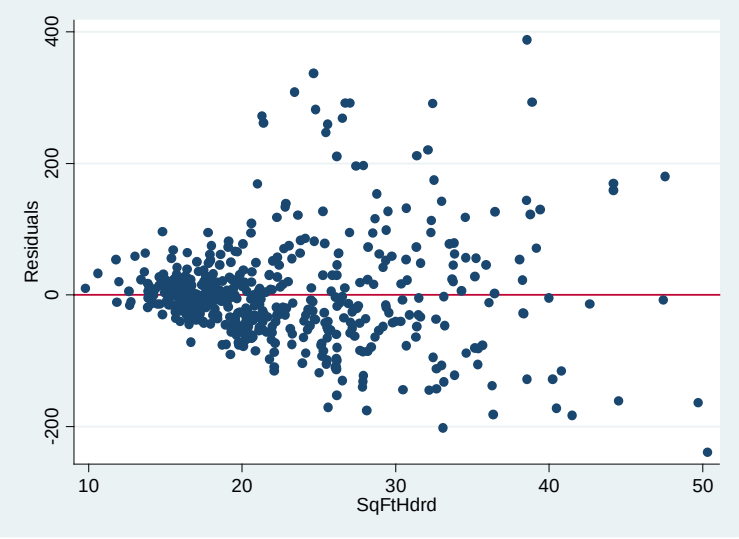




Assumption 2: Constant variance of the errors across X values.

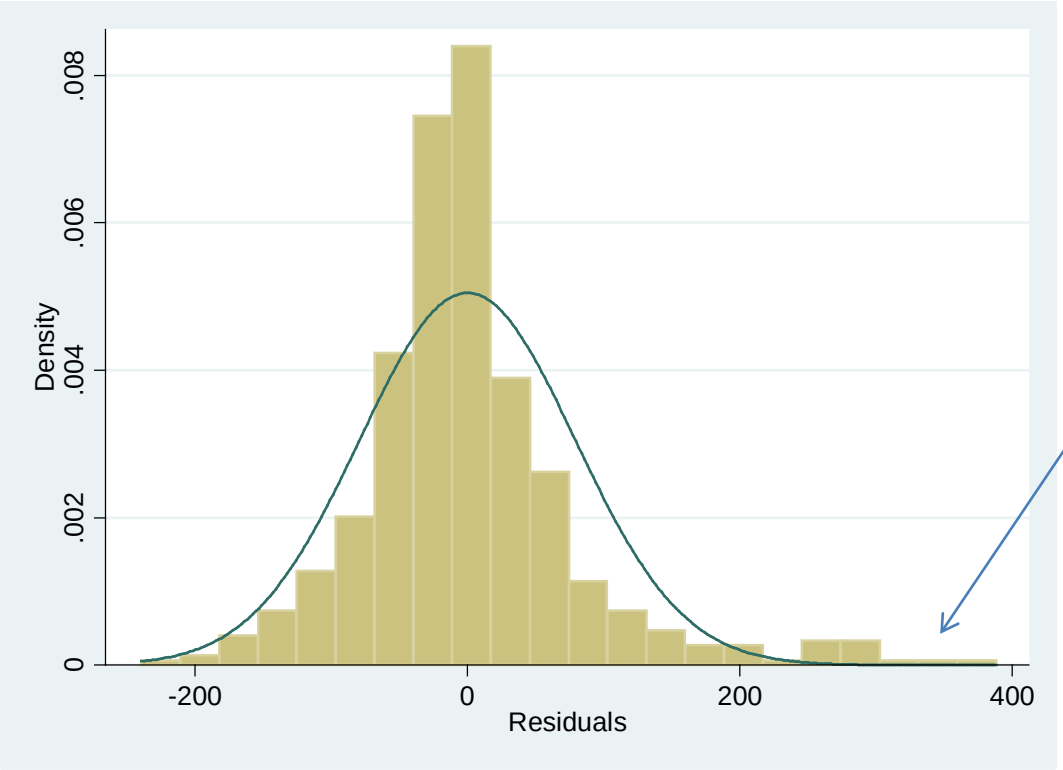
How to detect a problem:
Plot residuals versus fitted values. If you see increasing or decreasing spread, there is a problem with the assumption.

Example: Real estate data for n = 522 homes sold in a Midwestern city. Y = Sales price (thousands); X = Square feet (in hundreds). (*Source:* Applied Linear Regression Models)

Original data:	Residual plot:
	

Clearly, the variance is increasing as house size increases!

NOTE: Usually increasing variance and skewed distribution go together. Here is a histogram of the residuals, with a superimposed normal distribution. Notice the residuals extending to the right.

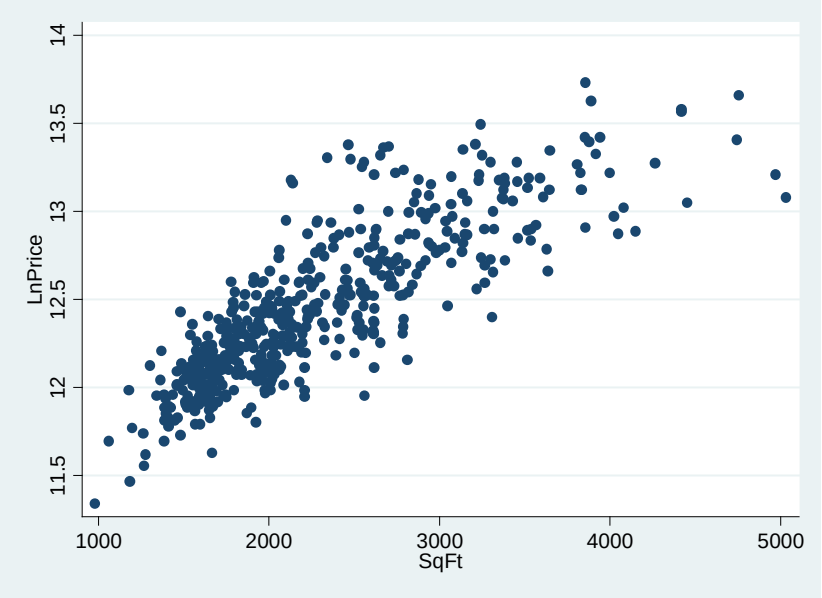


What to do about the problem:

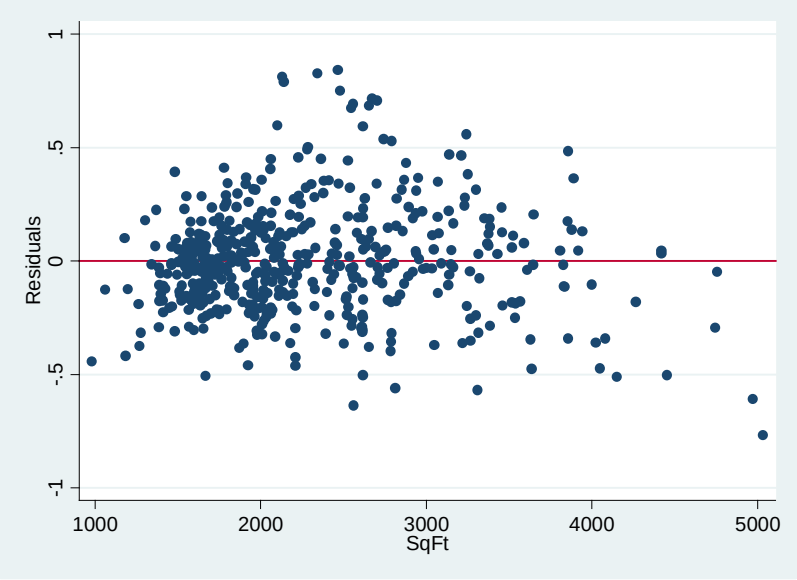
Transform the Y values, or both the X and Y values.

Example: Real estate sales, transform Y values to Y′ = ln (Y) = natural log of Y

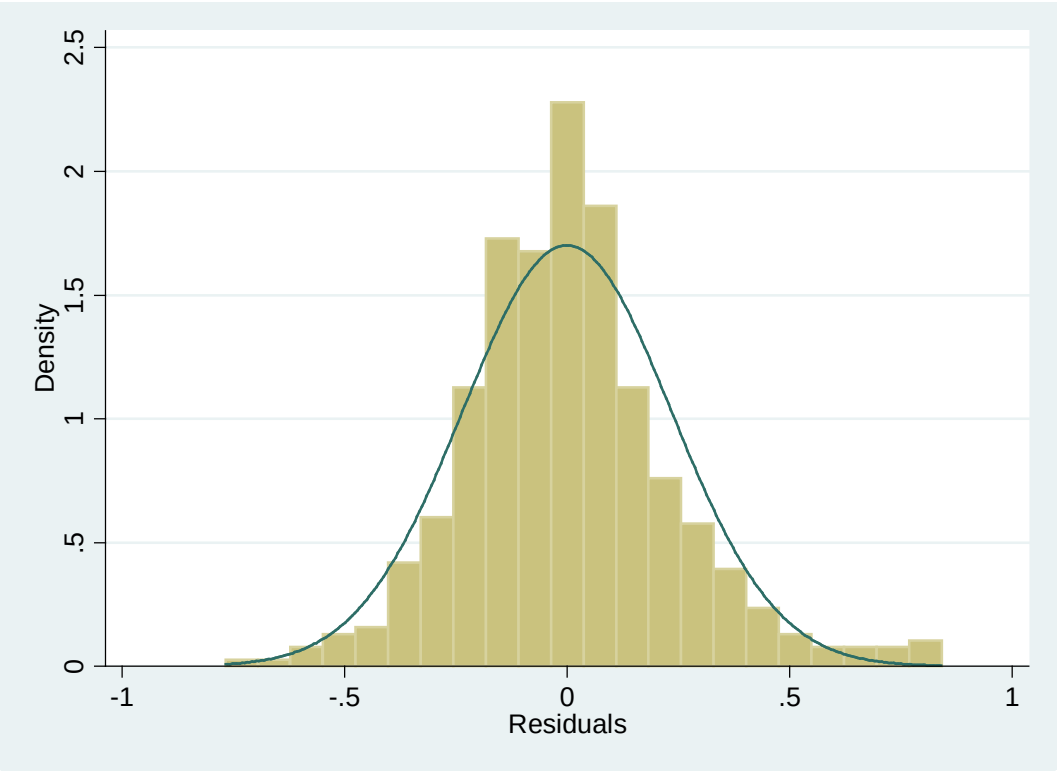
Scatter plot of ln(Price) vs Square feet



Residuals versus Square feet:



Looks like one more transformation might help – use square root of size. But we will leave it as this for now. See histogram of residual on next page.



This looks better – more symmetric and no outliers.

Using models after transformations

Transforming X only:

Use transformed X for future predictions: $X' = f(X)$.
Then do the regression using X' instead of X:

$$Y = \beta_0 + \beta_1 X' + \varepsilon$$

where we still assume the ε are $N(0, \sigma^2)$.

For example, if $X' = \sqrt{X}$ then the predicted values are:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \sqrt{X}$$

Transforming Y (and possibly X):

Everything must be done in transformed values. For confidence intervals and prediction intervals (to be covered next week), get the intervals *first* and then transform the endpoints back to original units.

Example: Predicting house sales price using square feet. Regression equation is:

Predicted Ln(Price) = 11.2824 + 0.051(Square feet in hundreds)
For a house with 2000 square feet = 20 hundred square feet:

$\hat{Y}' = 11.2824 + 0.051(20) = 12.3024$

So predicted price = exp(12.3024) = $220,224 (because taking exp reverses a natural log transformation).

Assumption 3: Independent errors

1. The main way to check this is to understand how the data were collected. For example, suppose we wanted to predict blood pressure from amount of fat consumed in the diet. If we were to sample entire families, and treat them as independent, that would be wrong. If one member of the family has high blood pressure, related members are likely to have it as well. Taking a random sample is one way to make sure the observations are independent.

2. If the values were collected over time (or space) it makes sense to plot the residuals versus order collected, and see if there is a trend or cycle.

OUTLIERS

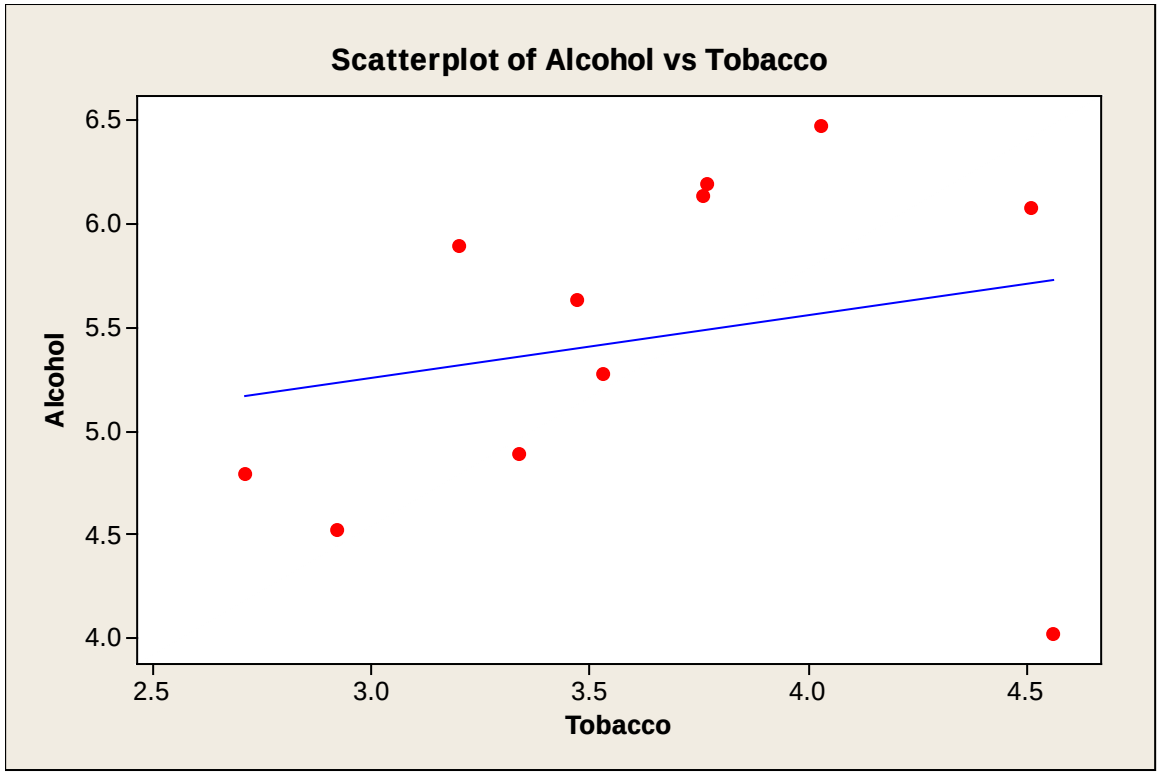
Some reasons for outliers:

1. A mistake was made. If it's obvious that a mistake was made in recording the data, or that the person obviously lied, etc., it's okay to throw out an outlier and do the analysis without it. For example, a height of 7 inches is an obvious mistake. If you can't go back and figure out what it should have been (70 inches? 72 inches? 67 inches?) you have no choice but to discard that case.

2. The person (or unit) belongs to a different population, and should not be part of the analysis, so it's okay to remove the point(s). An example is for predicting house prices, if a data set has a few mansions (5000+ square feet) but the other houses are all smaller (1000 to 2500 square feet, say), then it makes sense to predict sales prices for the smaller houses only. In the future when the equation is used, it should be used only for the range of data from which it was generated.

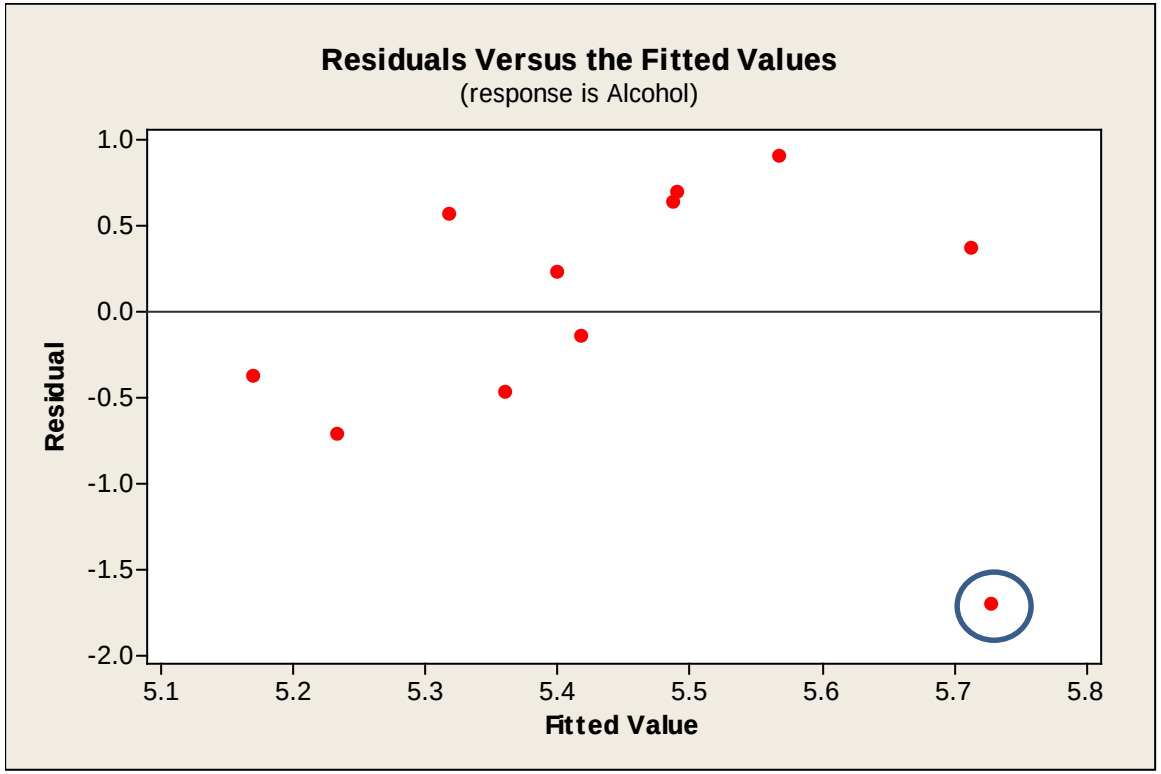
3. Sometimes outliers are simply the result of natural variability. In that case, it is NOT okay to discard them. If you do, you will underestimate the variance.

(Story of Alcohol and Tobacco from DASL)

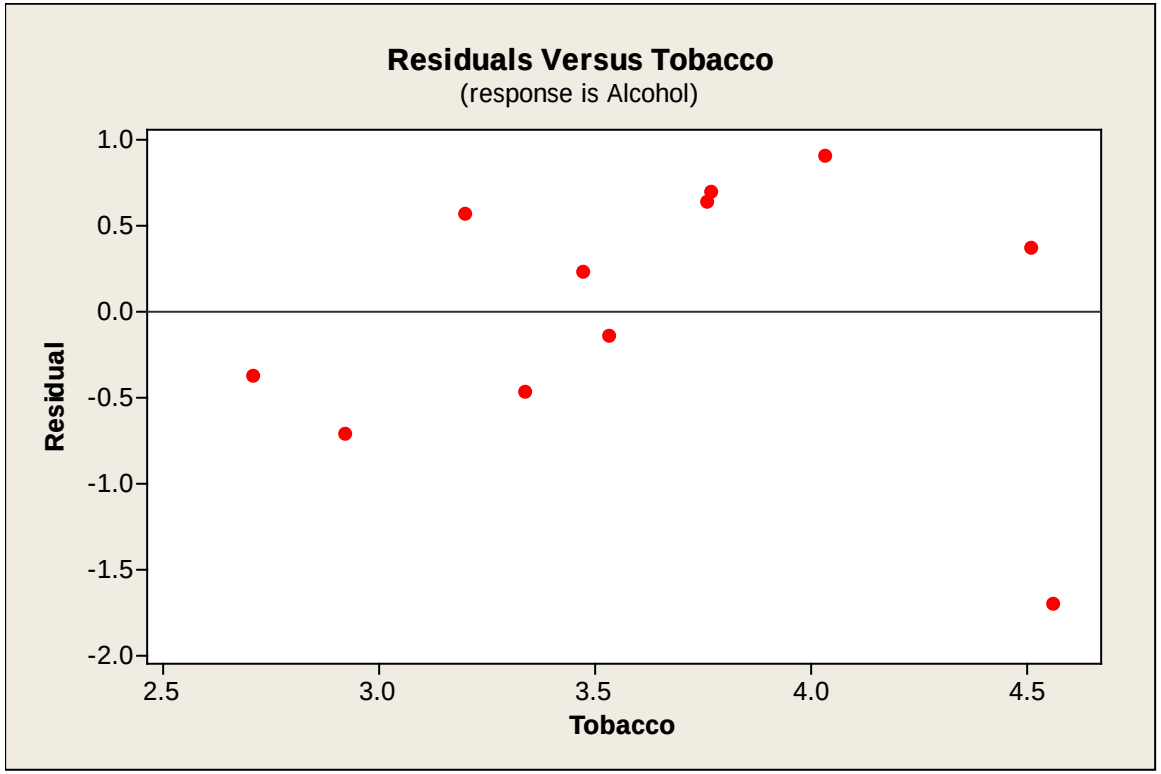


Notice Northern Ireland in lower right corner – a definite outlier, based on the combined (X,Y) values.

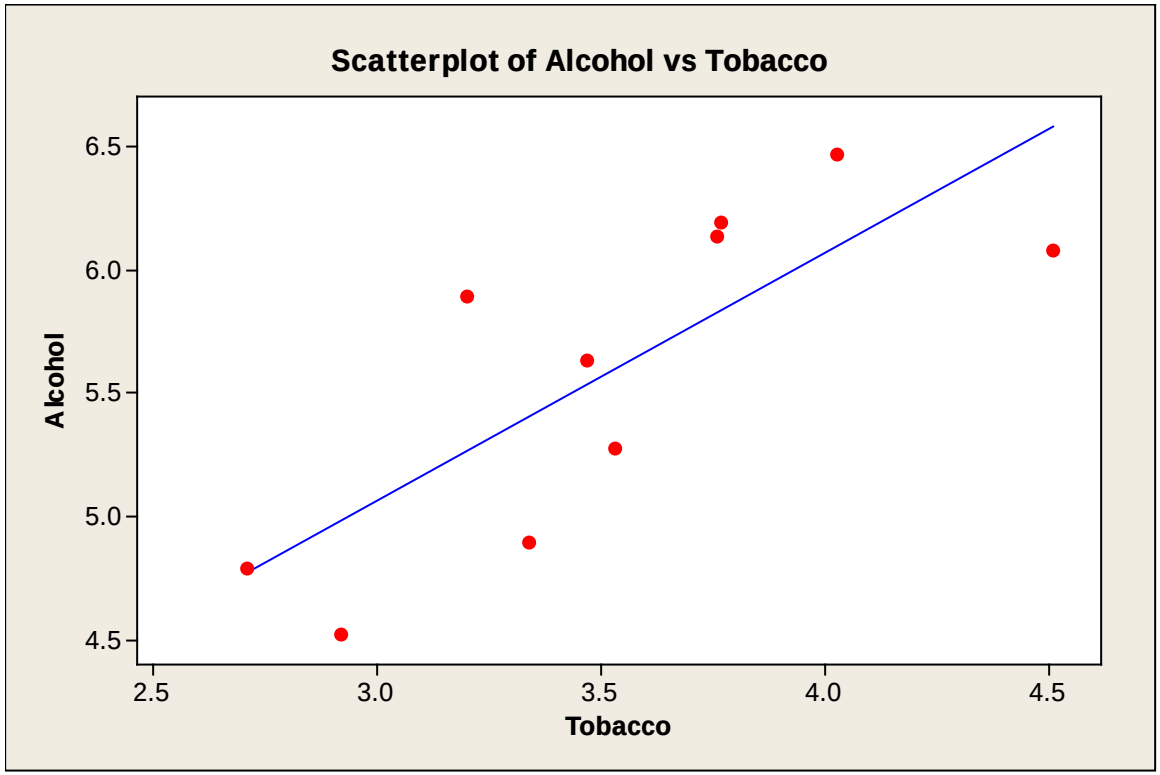
Why is it an outlier? It represents a different religion than other areas of Britain.



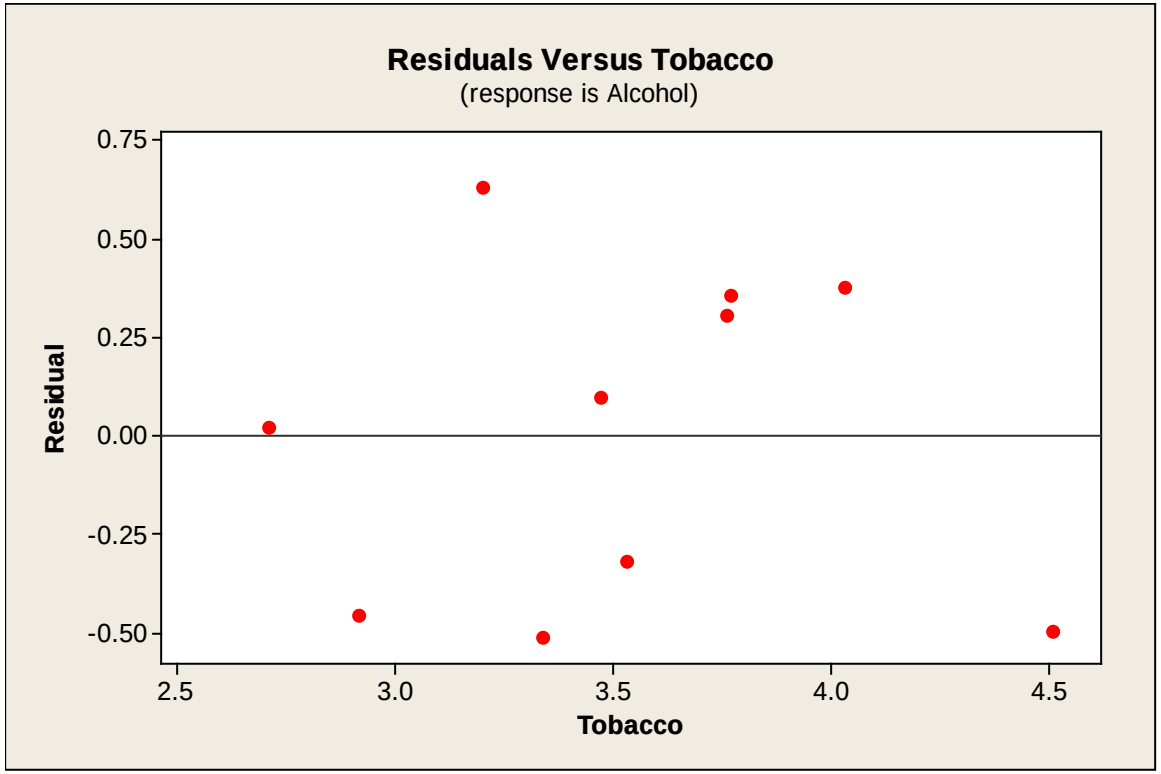
In the plot of residuals versus fitted values, it's even more obvious that the outlier is a problem.



The plot of residuals versus the X variable is very similar to residuals vs fitted values. Again the problem is obvious.



Here is a plot with Northern Ireland removed.



Here is a residual plot with Northern Ireland removed.

Notice how much the analysis changes when the outlier is removed. “R-sq” /100 is the square of the correlation between Alcohol and Tobacco, so the correlation goes from $\sqrt{.05} = .22$ to $\sqrt{.615} = .78$. Intercept and slope change too:

With Outlier (Northern Ireland)

```
The regression equation is
Alcohol = 4.35 + 0.302 Tobacco

Predictor    Coef  SE Coef     T      P
Constant    4.351    1.607  2.71  0.024
Tobacco    0.3019   0.4388  0.69  0.509

S = 0.819630   R-Sq = 5.0%   R-Sq(adj) = 0.0%
```

Without Outlier

```
The regression equation is
Alcohol = 2.04 + 1.01 Tobacco

Predictor    Coef  SE Coef     T      P
Constant    2.041    1.001  2.04  0.076
Tobacco    1.0059   0.2813  3.58  0.007

S = 0.446020   R-Sq = 61.5%   R-Sq(adj) = 56.7%
```

Transformations in R:

If you want to **transform** the response variable *Y* into some new variable *Y'*, you can add a new column to the data table consisting of the new variable.

For the data table named *MyData*, to square the response variable *GPA* and add it to the data table, type:

```
> MyData <- cbind(MyData, MyData$GPA^2)
```

To take its *square root*, type:

```
> MyData <- cbind(MyData, sqrt(MyData$GPA) )
```

To take its *natural logarithm*, type:

```
> MyData <- cbind(MyData, log(MyData$GPA) )
```

To take its *common logarithm (base 10)*, type:

```
> MyData <- cbind(MyData, log10(MyData$GPA) )
```

To take its *reciprocal*, type:

```
> MyData <- cbind(MyData, 1/MyData$GPA )
```

To take its *reciprocal square root*, type:

```
> MyData <- cbind(MyData, 1/sqrt(MyData$GPA) )
```

And so on. You will want to give the new column in the data table an appropriate name. You can then run a linear model using the transformed response variable and the original predictor. More in discussion on Friday!