CHECKING ASSUMPTIONS/CONDITIONS AND FIXING PROBLEMS

Recall, linear model is:

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

where we assume:

$E\{\varepsilon\} = 0$, Variance of the "errors" = $\sigma^2\{\varepsilon\} = \sigma^2$, same value at any X

Sometimes assume errors are normally distributed ("normal errors" model).

After we fit the model using data, the *residuals* $e_i = Y_i - \hat{Y}_i$ will be useful for checking the assumptions and conditions.

But what is a "large" residual? If discussing GPA, 0.5 could be large but if discussion SAT scores (possible scores of 200 to 800), 0.5 would be tiny.

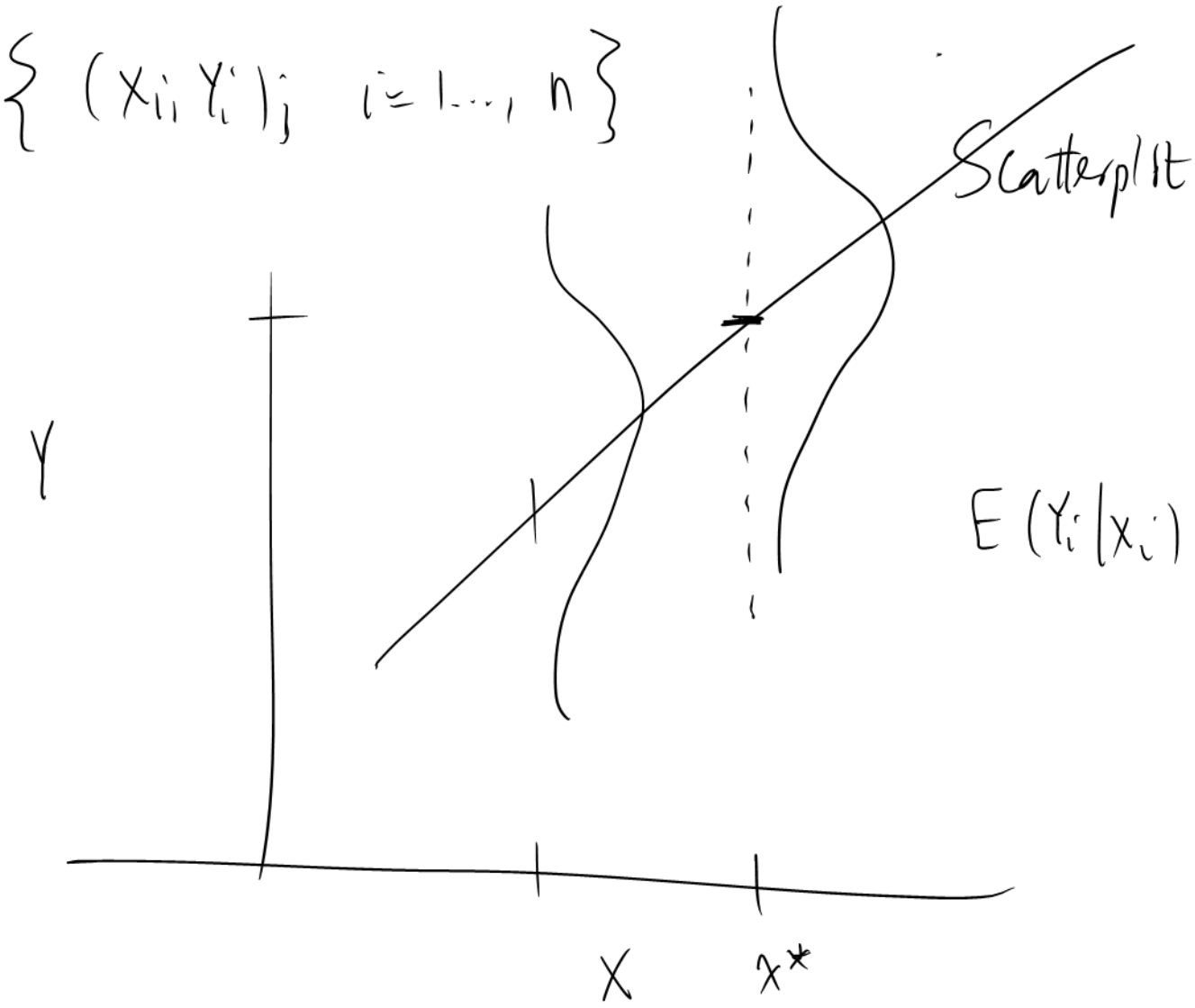
Need to create equivalent of a *z*-score. Recall estimate of σ is $s = \sqrt{MSE}$

"Semi-studentized" residuals (p. 103) $e_i^* = \dfrac{e_i}{\sqrt{MSE}}$

Note these are called *standardized residuals* in R.

Later, will learn another version, which R calls *studentized residuals*.

DATA : $\{(X_i, Y_i)_i \quad i = 1, \ldots, n\}$



$$E(Y_i | X_i) = \beta_0 + \beta_1 X_i$$

Assumptions in the Normal Linear Regression Model

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

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?

Applet for illustrating the effect of outliers on the regression line and correlation: http://illuminations.nctm.org/LessonDetail.aspx?ID=L455

Useful Plots for Checking Assumptions and Answering These Questions

Reminders:

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

Predicted $Y_i = \hat{Y}_i = b_0 + b_1 X_i$, also called "fitted Y_i"

Recall the *semi-studentized residual* for unit *i* is $e_i^* = \frac{e_i}{\sqrt{MSE}}$

Plot	Useful for
Dotplot, stemplot, histogram of X's	**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* 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

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Dist

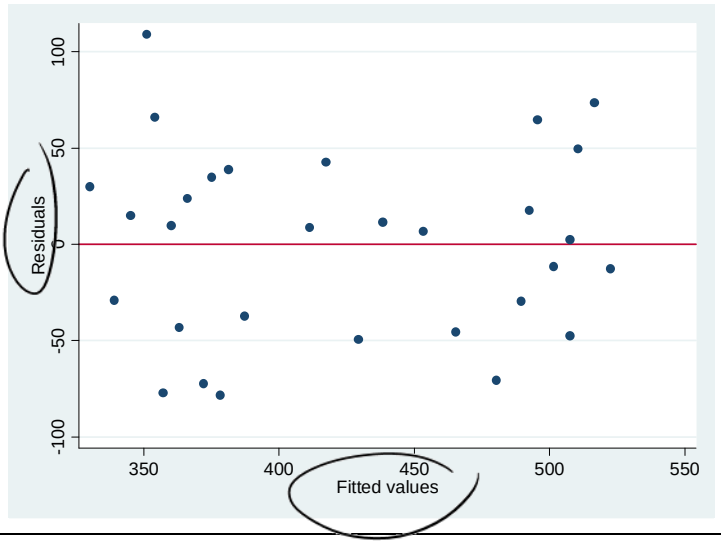
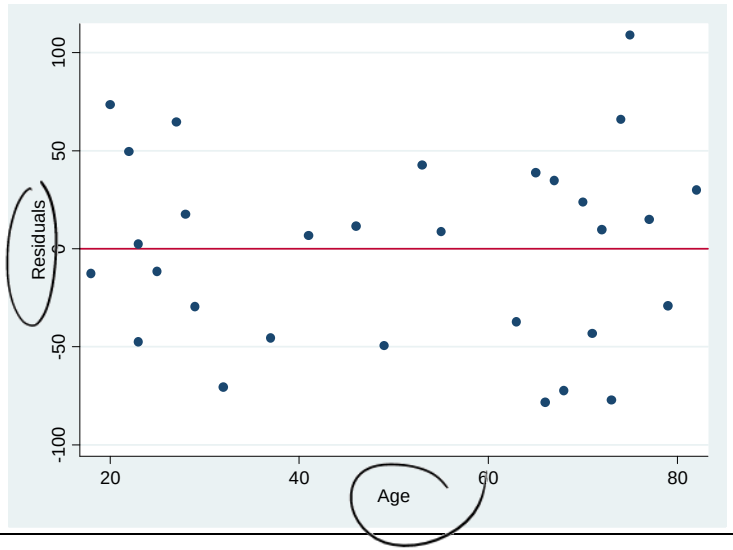
Age

Plot of:

residuals versus predicted ("fitted") values

residuals vs Age

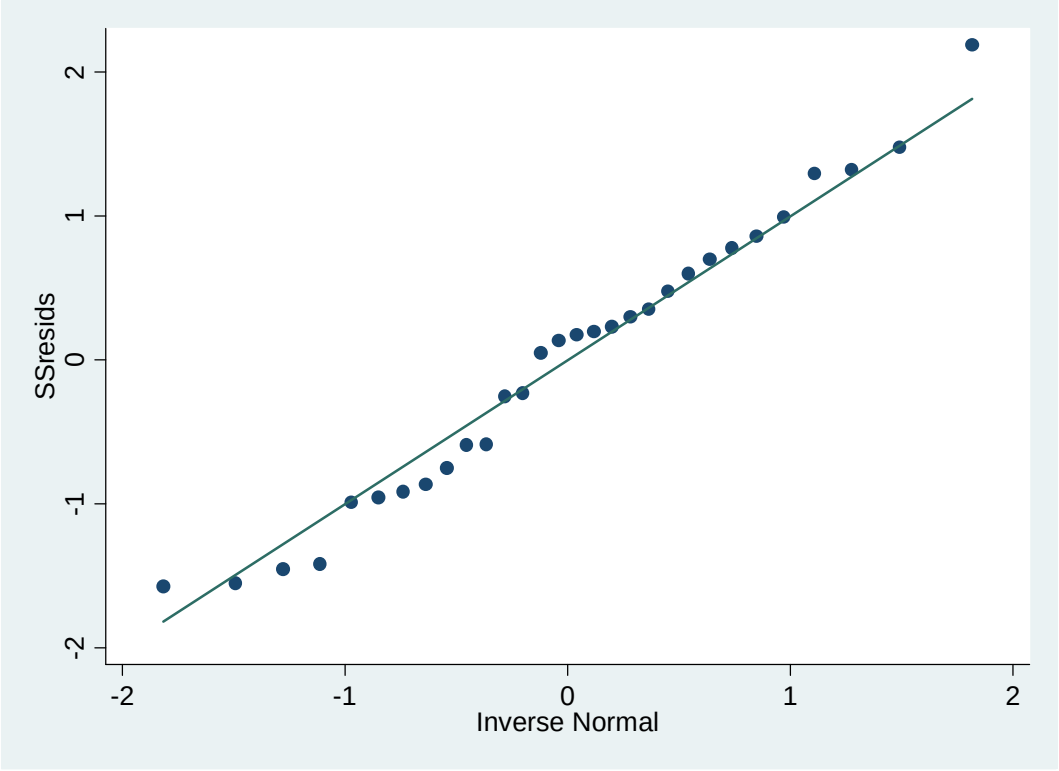
NOTE: Plot of residuals versus predictor variable X should look the same except for the scale on the X axis, because fitted values are linear transform of X's. However, when the slope is negative, one will be a mirror image of the other.

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:

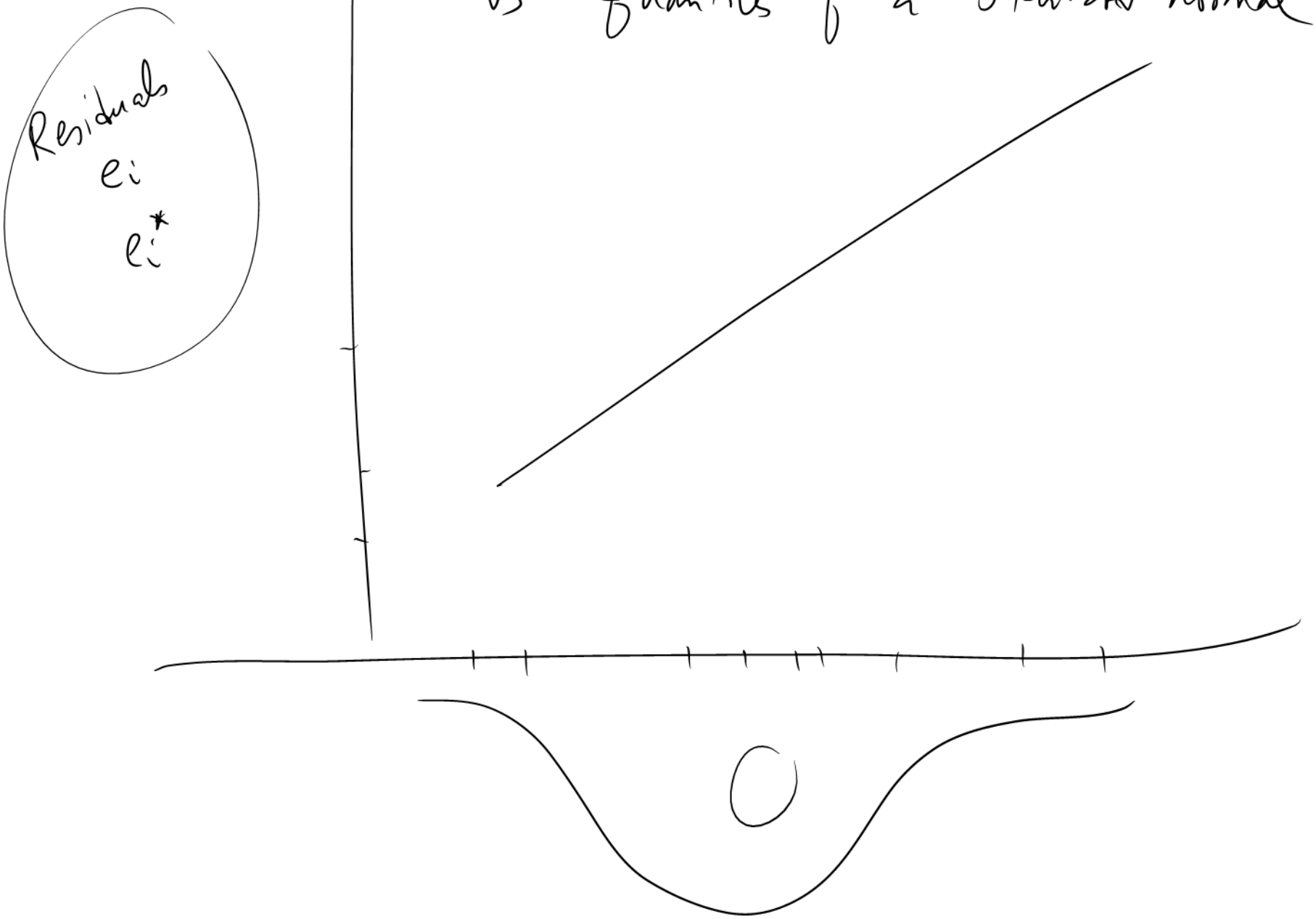
- **Normal probability plot** of standardized residuals, to check normality, **Assumption A4** (see Figure 3.2d, p. 104 in textbook); explanation on white board.



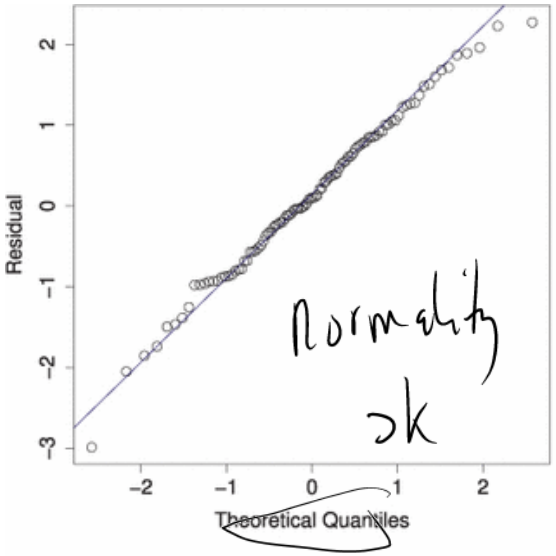
This is a pretty good plot. There is one point at each end that is slightly off, that might be investigated, but no major problems.

QQ plot

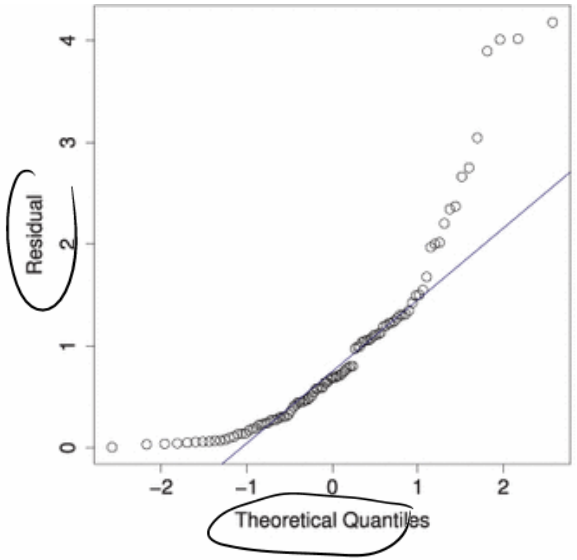
Compare quantiles of the residuals (e_i, e_i^*) vs quantiles of a standard normal



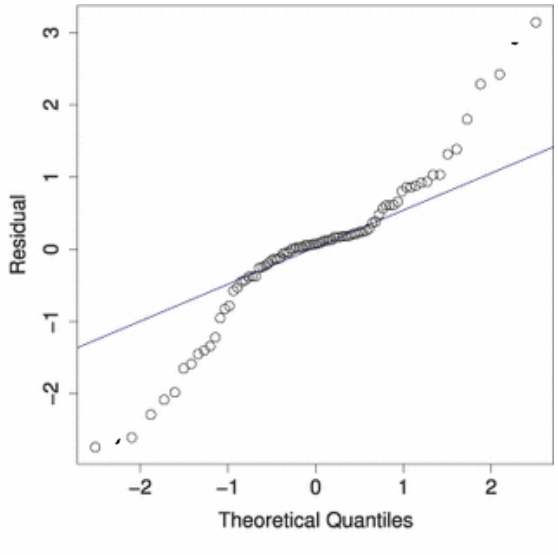
Examples of good and bad normal probability plots (also see Figure 3.9, p. 112):



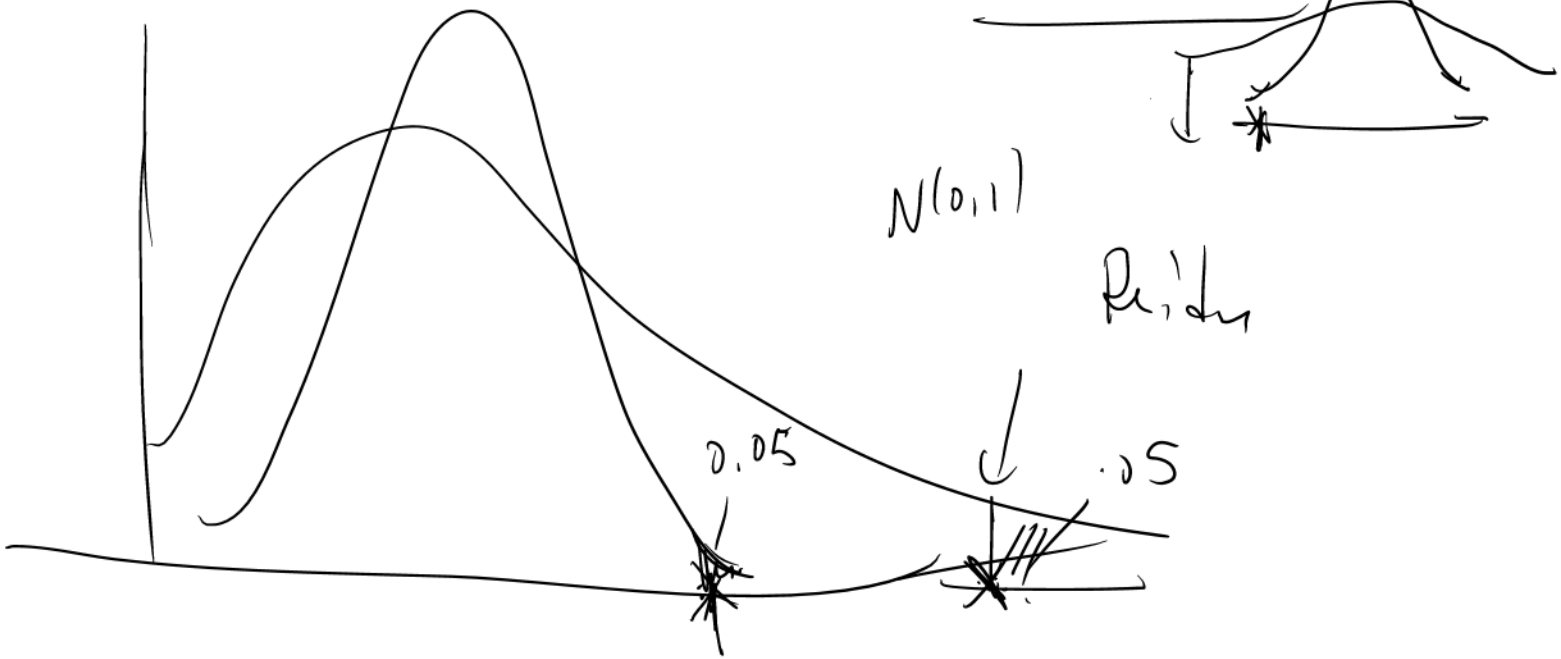
(a) Normal residuals



(b) Skewed right residuals



(c) Long-tailed residuals



Stemplot of standardized residuals (to check normality assumption):
To generate them in R, for the linear model called "HWModel":
> Highway$StResids <- rstandard(HWModel)

```
> stem.leaf(Highway$StResids)
1 | 2: represents 1.2
 leaf unit: 0.1
            n: 30
   2     -1. | 66
   6     -1* | 0044
  11     -0. | 66789
  13     -0* | 22
  (8)     0* | 01122334
   9      0. | 6778
   5      1* | 033
   2      1. | 5
   1      2* | 2
```

This is further confirmation that the residuals are relatively symmetric with no major outliers. The 2.2 is for a driver with X = 75 years, Y = 460 feet.

What to do when assumptions aren't met

Assumption 1:
Relationship is linear.

How to detect a problem:
Plot residuals versus fitted values. 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). 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. (Also see Figure 3.13, p. 130.)

Truth $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$

Fit $Y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i$

$\hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_i$

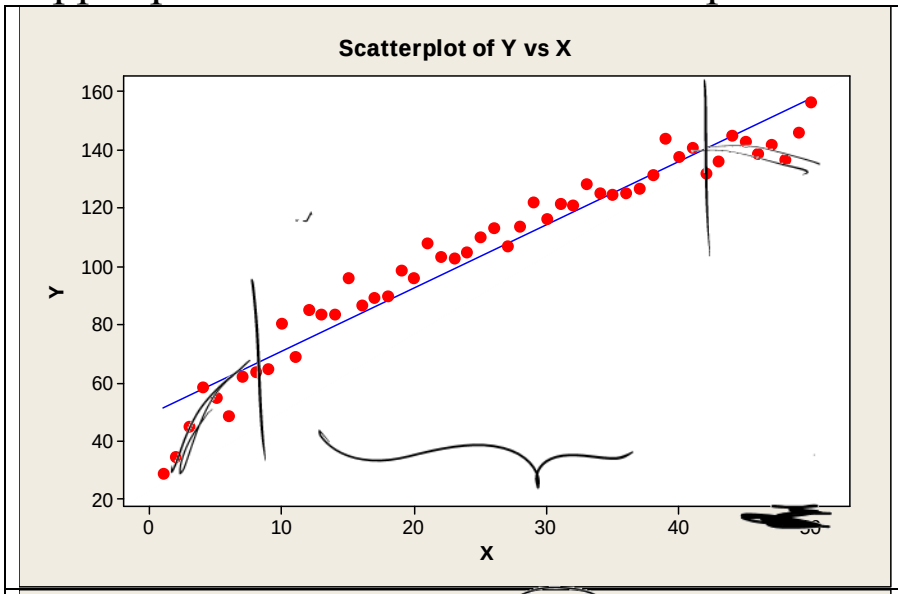
$$e_i = Y_i - \hat{Y}_i = \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i\right) - \left(\hat{\alpha}_0 + \hat{\alpha}_1 x_i\right)$$

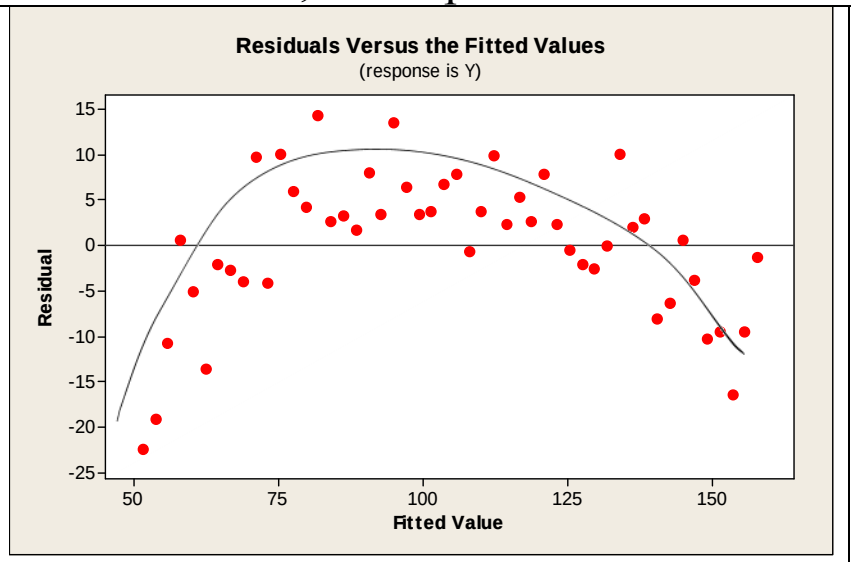
$$= (\beta_0 - \hat{\alpha}_0) + (\beta_1 - \hat{\alpha}_1) x_i + \beta_2 x_i^2 + \varepsilon_i$$

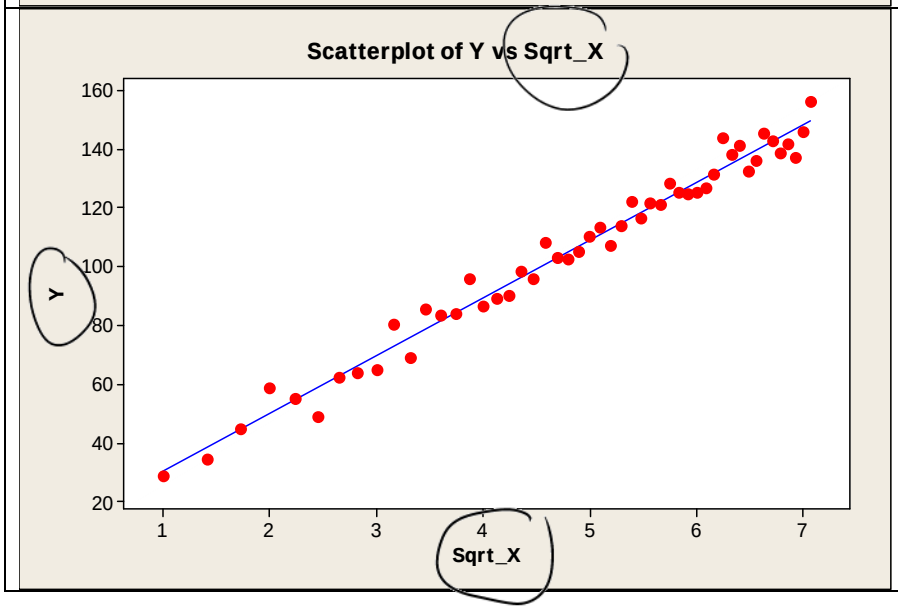
(where $\beta_0 - \hat{\alpha}_0 \approx 0$, $\beta_1 - \hat{\alpha}_1 \approx 0$, and β_2 remains)

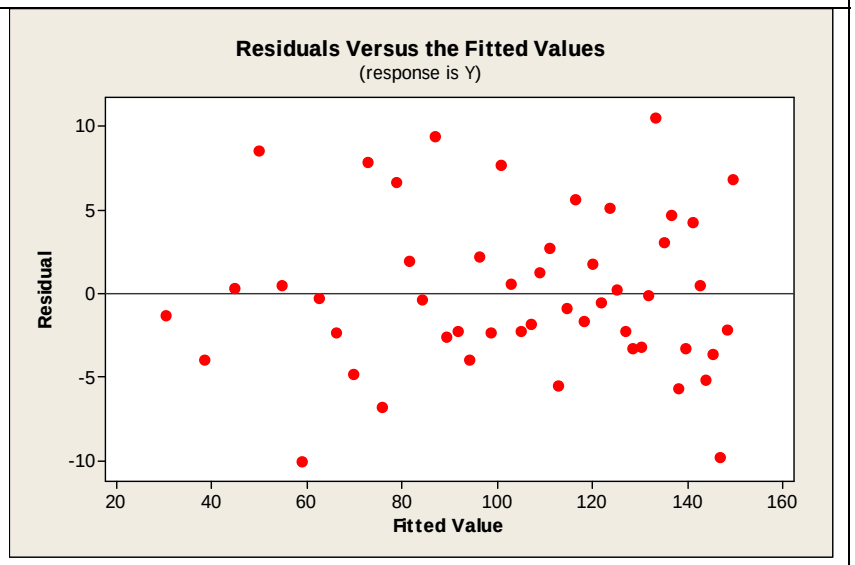
Residuals are inverted U, use X' = $\sqrt{X}$ or $\log_{10}$ X

Upper plots show data and residual plot before transformation; lower plots show after.

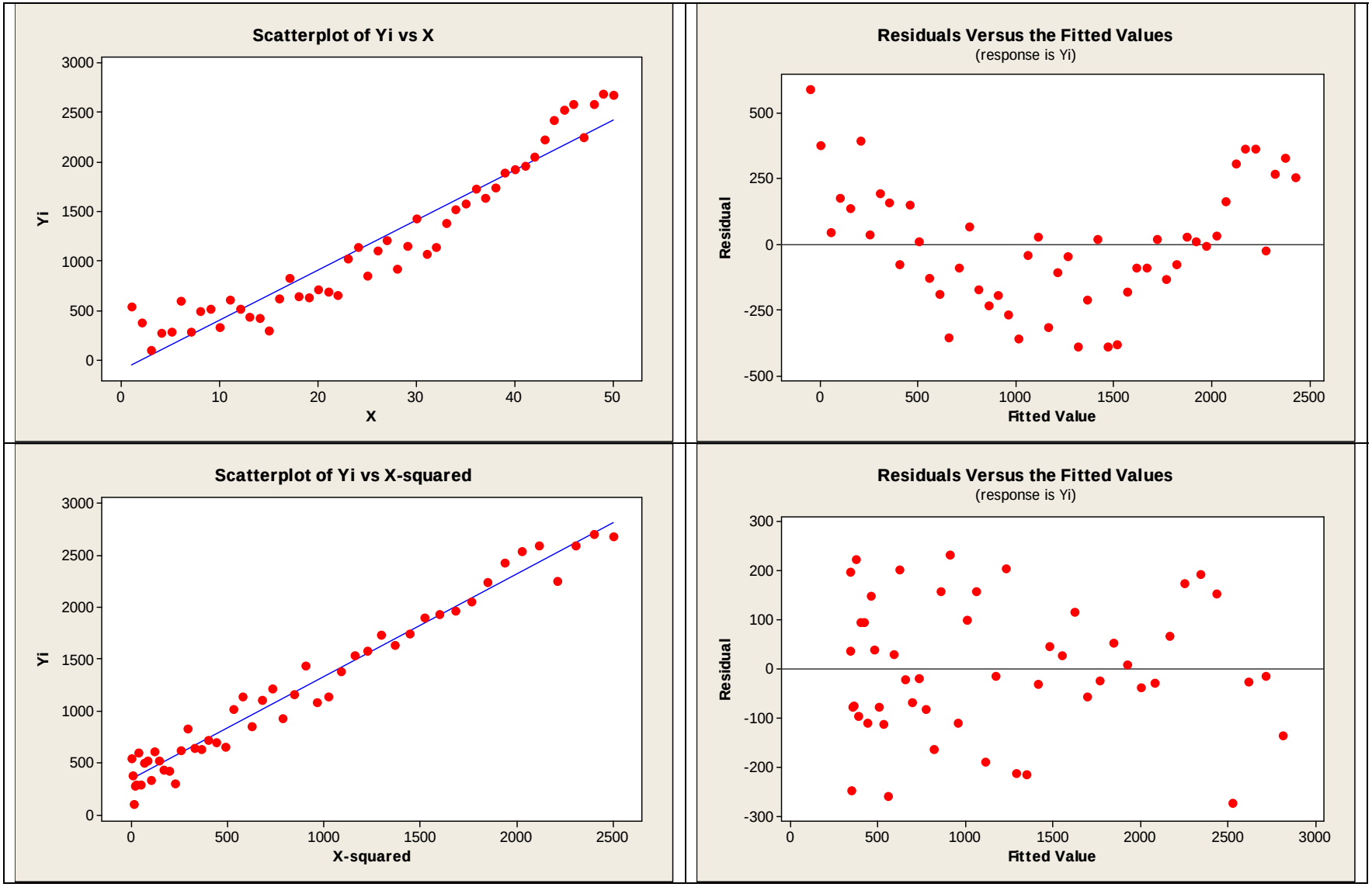




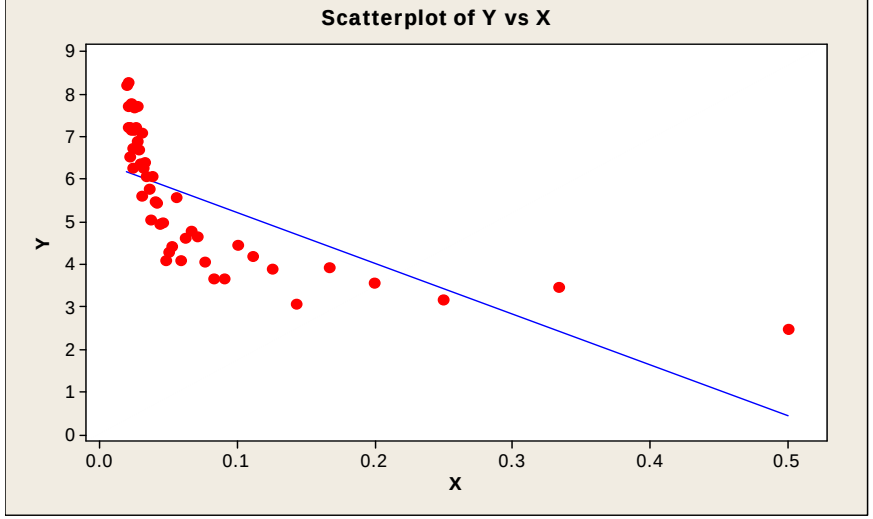


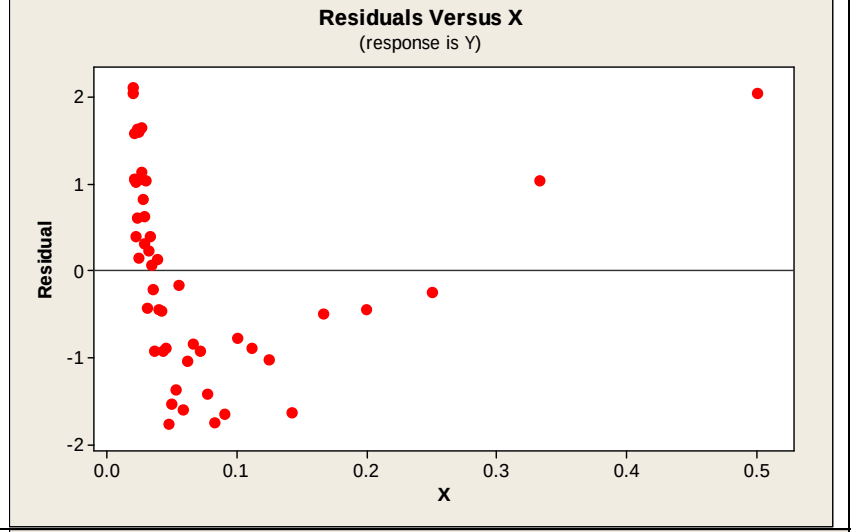


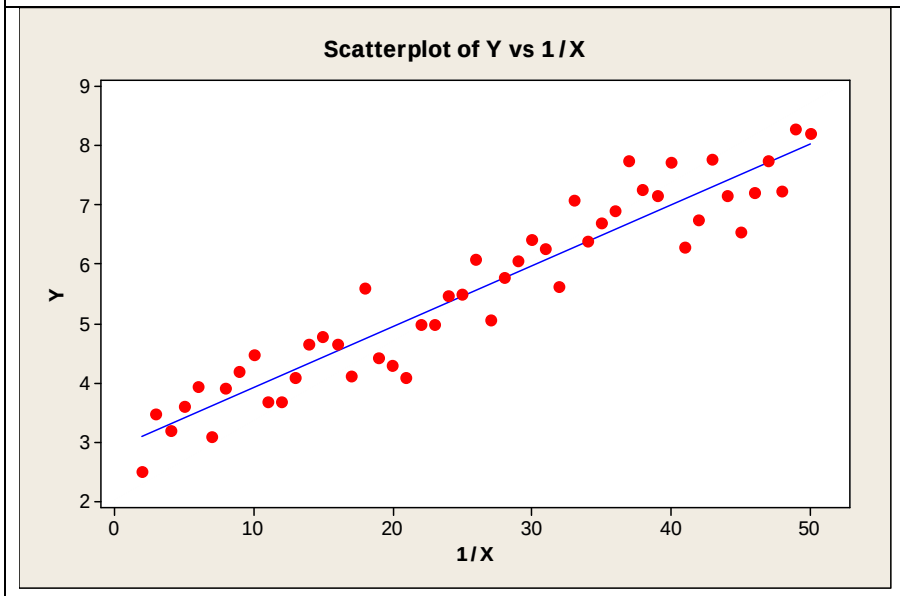
Residuals are U-shaped and association between X and Y is positive: Use $X' = X^2$

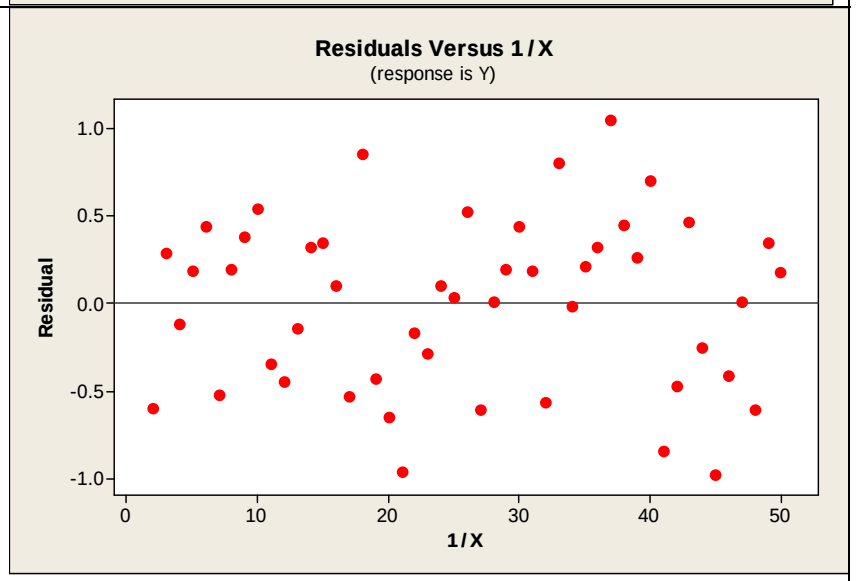


Residuals are U-shaped and association between X and Y is negative:
Use X' = 1/X or X' = exp(-X)





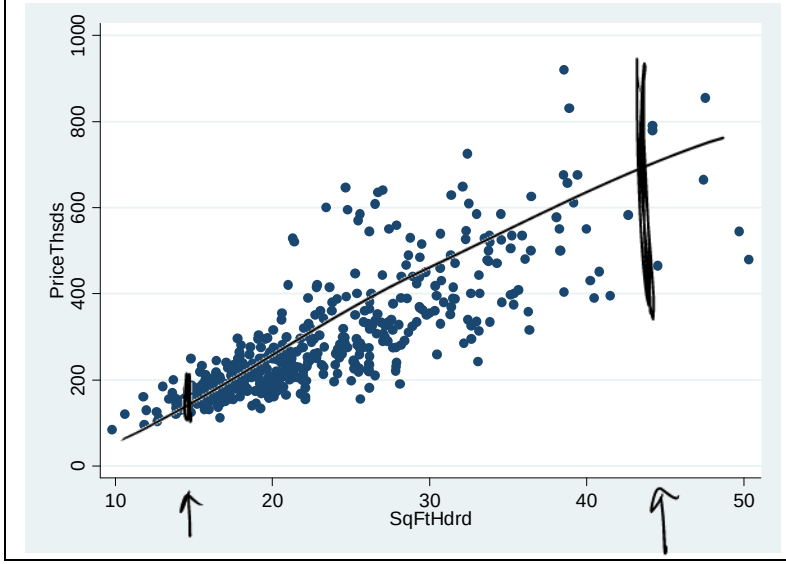




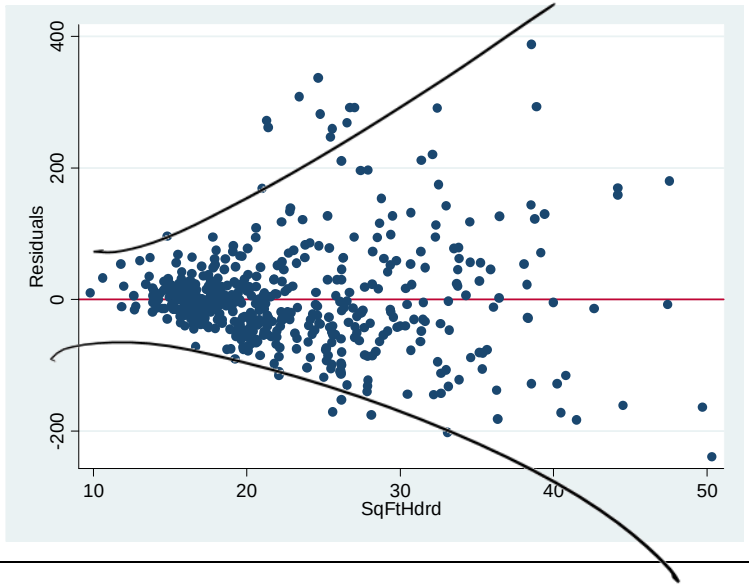
Assumption 2: Constant variance of the errors across X values.
Often correct **Assumption 3** (normality) using the same methods.
How to detect a problem:
Plot residuals versus fitted values. If you see increasing or decreasing spread, there is a problem with the assumption.

Example: Real estate data for $n = 522$ homes sold in a Midwestern city. Y = Sales price (thousands); X = Square feet (in hundreds). (*Data described in Appendix C.7*)

Original data:

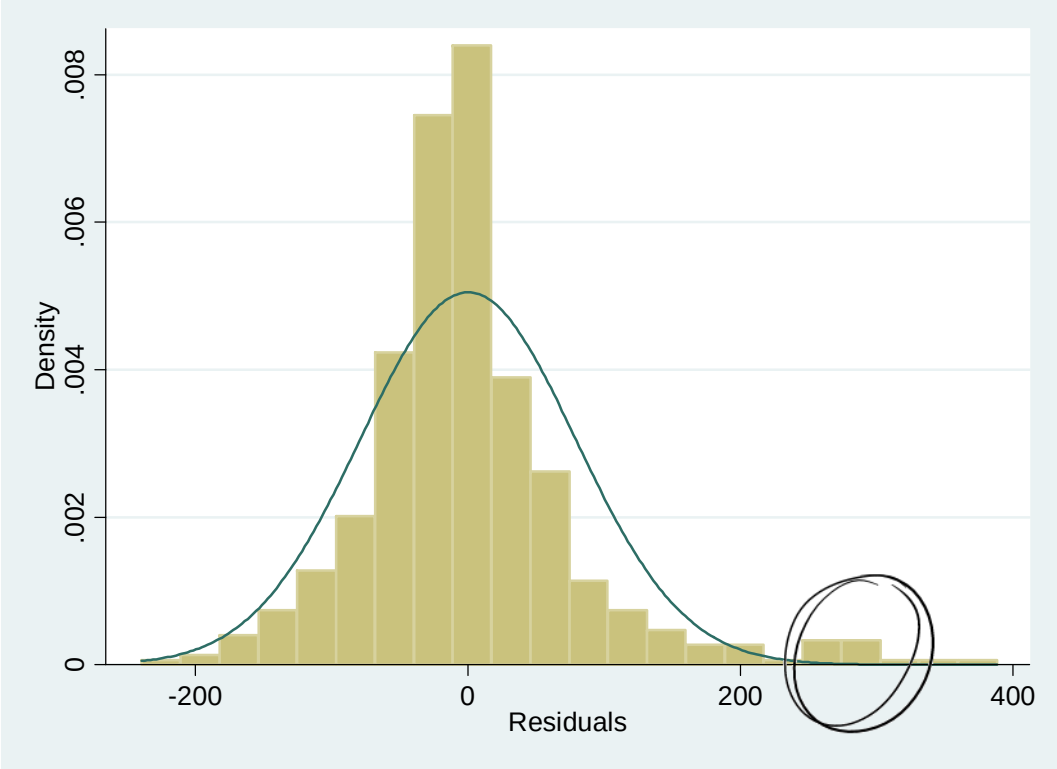


Residual plot:



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.

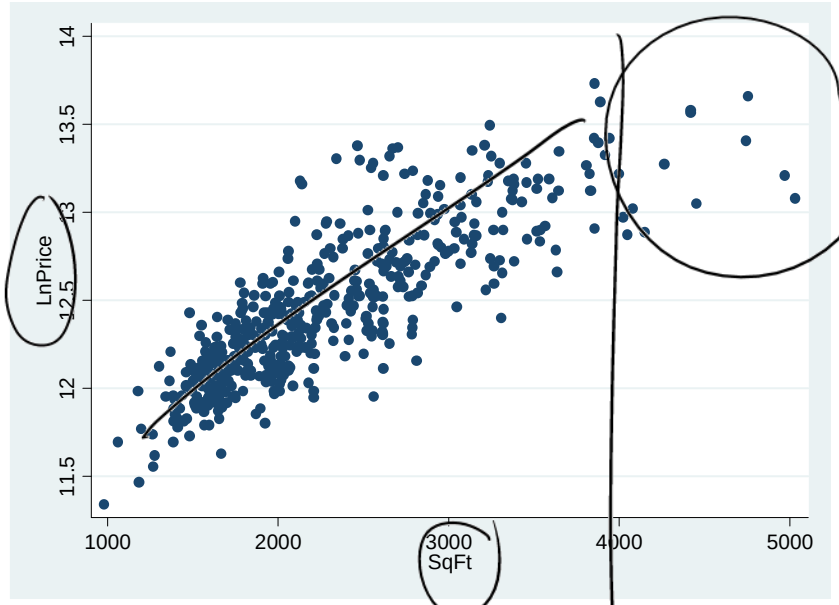


What to do about the problem:

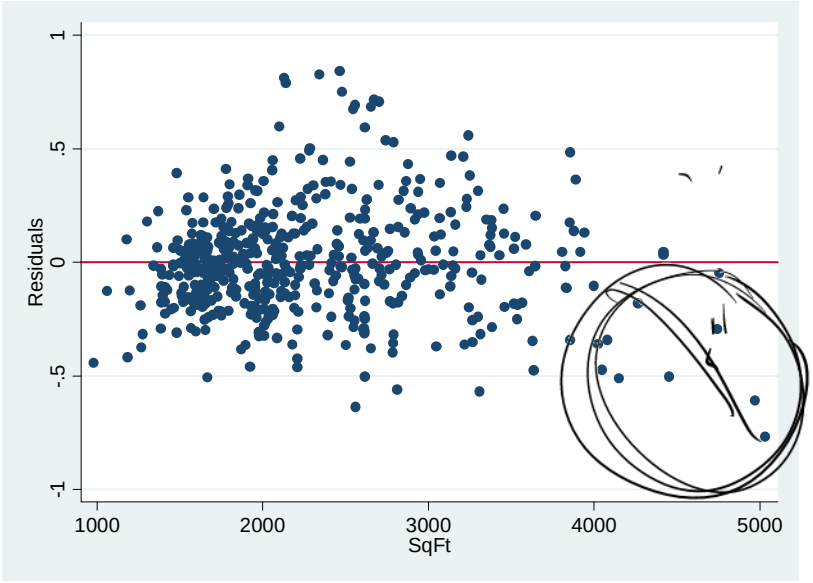
Transform the Y values, or both the X and Y values.

Example: Real estate sales, transform Y values to Y' = ln (Y)

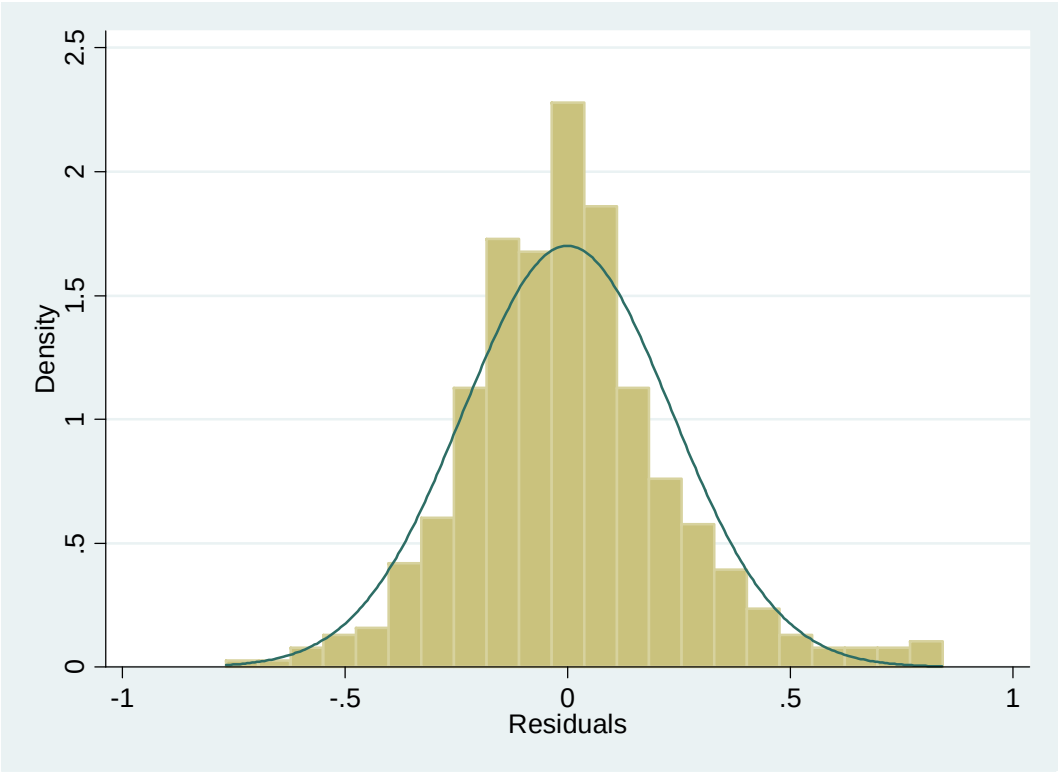
Scatter plot of ln(Price) vs Square feet



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.

See Figure 3.15, p. 132 for prototypes of regression patterns requiring various transformations of the Y-variable.

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, σ^2).

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

$$\hat{Y} = b_0 + b_1\sqrt{X}$$

Transforming Y (and possibly X):

Everything must be done in transformed values. For confidence intervals and prediction intervals, 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.

95% prediction interval for Ln(Price) is 11.8402, 12.7634. Transform back to dollars:

Exp(11.8402) = $138,718
Exp(12.7634) = $349,200

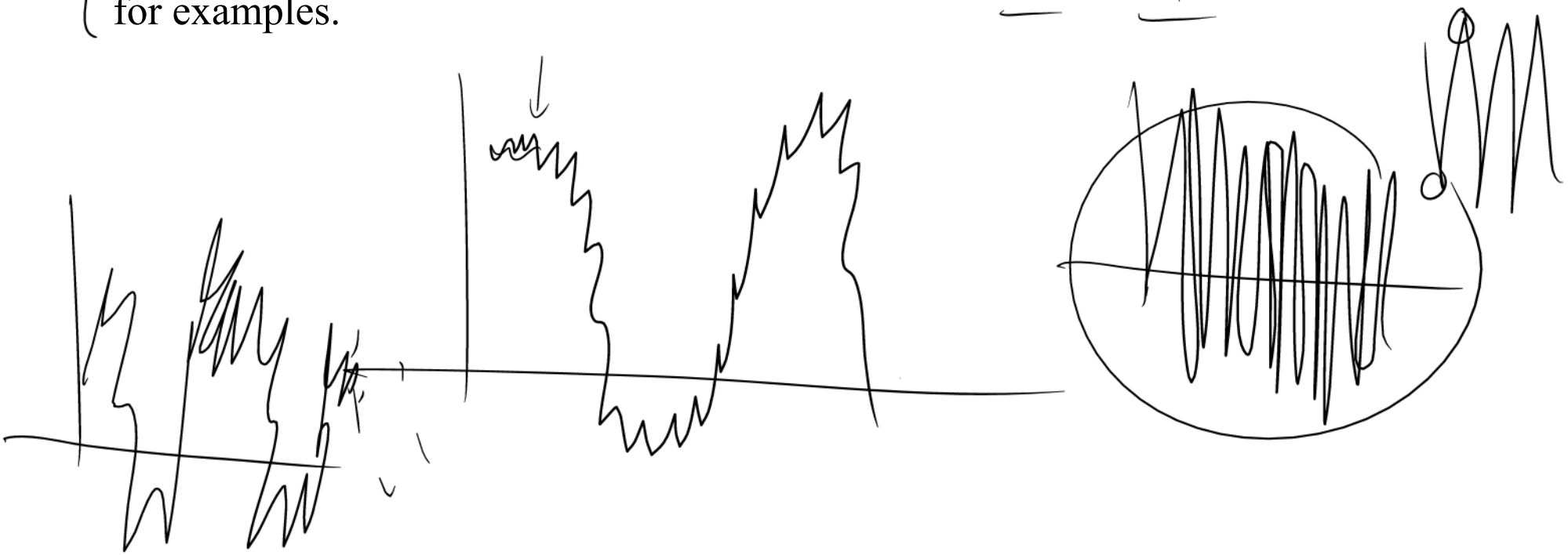
95% confidence interval for the *mean* Ln(Price) is 12.2803, 12.3233

Exp(12.2803) = $215,410
Exp(12.3233) = $224,875

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. See page 109 for examples.



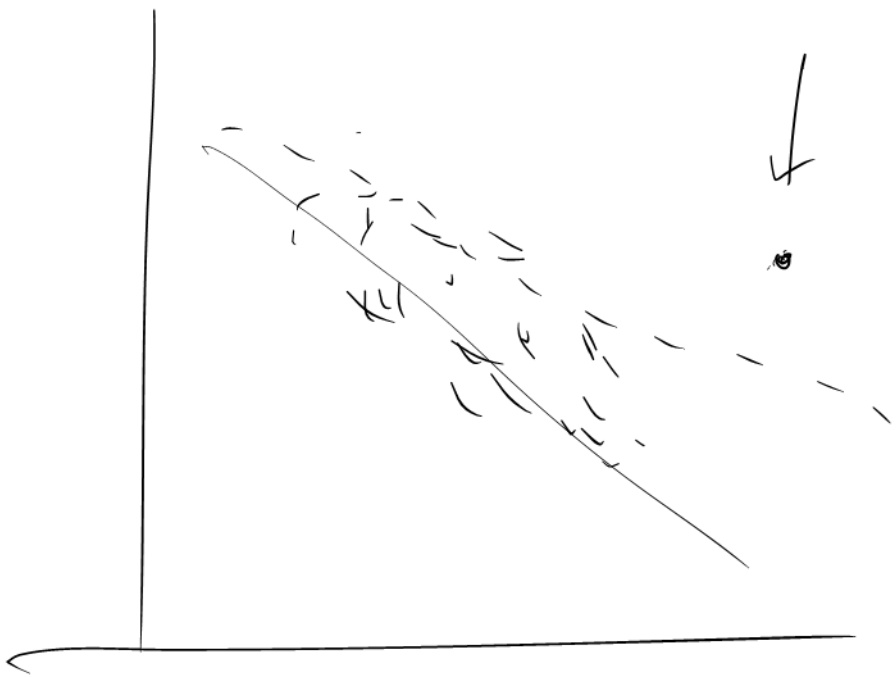
OUTLIERS

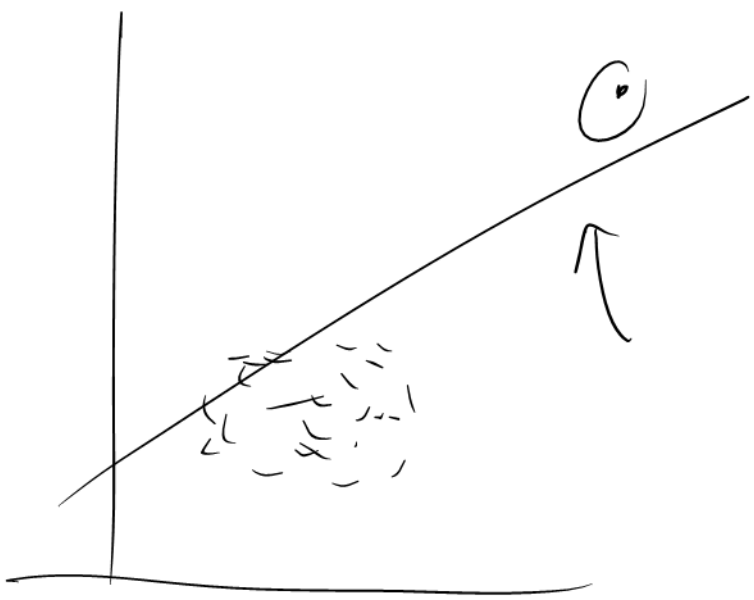
Some reasons for outliers:

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.

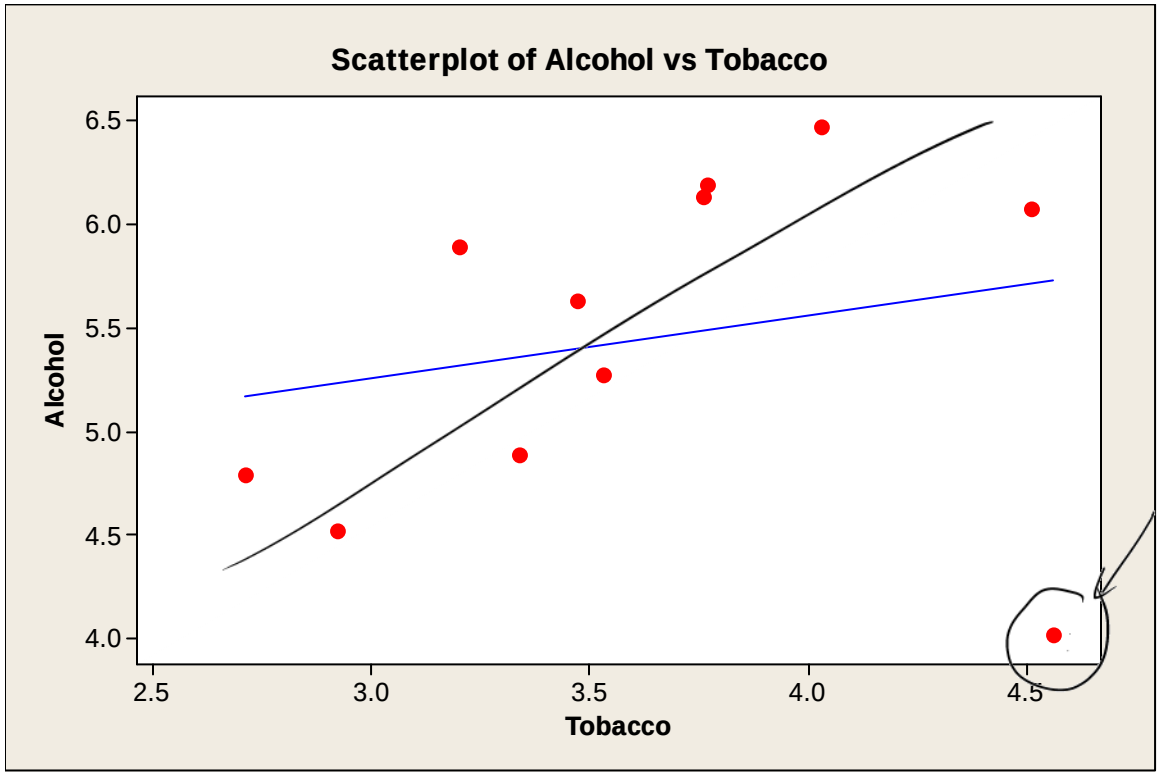
2. The person (or unit) belongs to a different population, and should not be part of the analysis, so it's okay to remove the point(s). An example is for predicting house prices, if a data set has a few mansions (5000+ square feet) but the other houses are all smaller (1000 to 2500 square feet, say), 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.



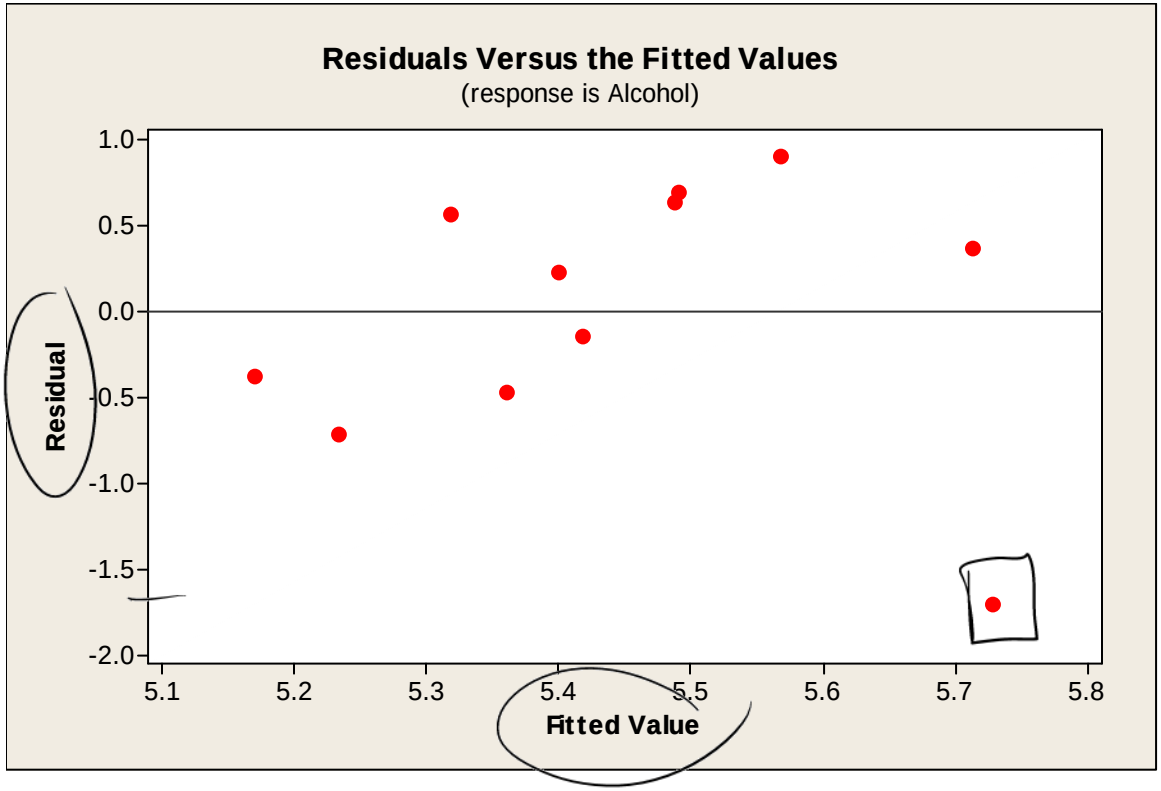


Story of Alcohol and Tobacco from DASL: Household spending on Tobacco and Alcohol for 11 regions in Britain, including Northern Ireland.
http://lib.stat.cmu.edu/DASL/Stories/AlcoholandTobacco.html

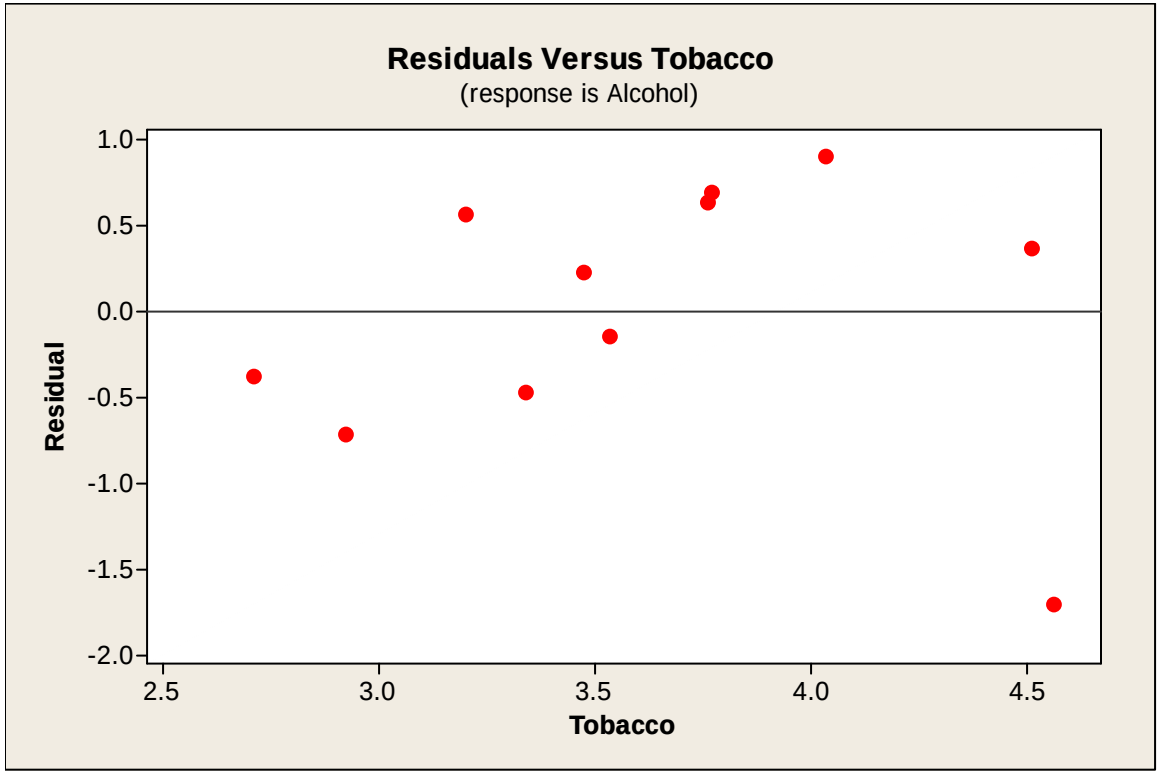


Notice Northern Ireland in lower right corner – a definite outlier, based on the combined (X,Y) values.

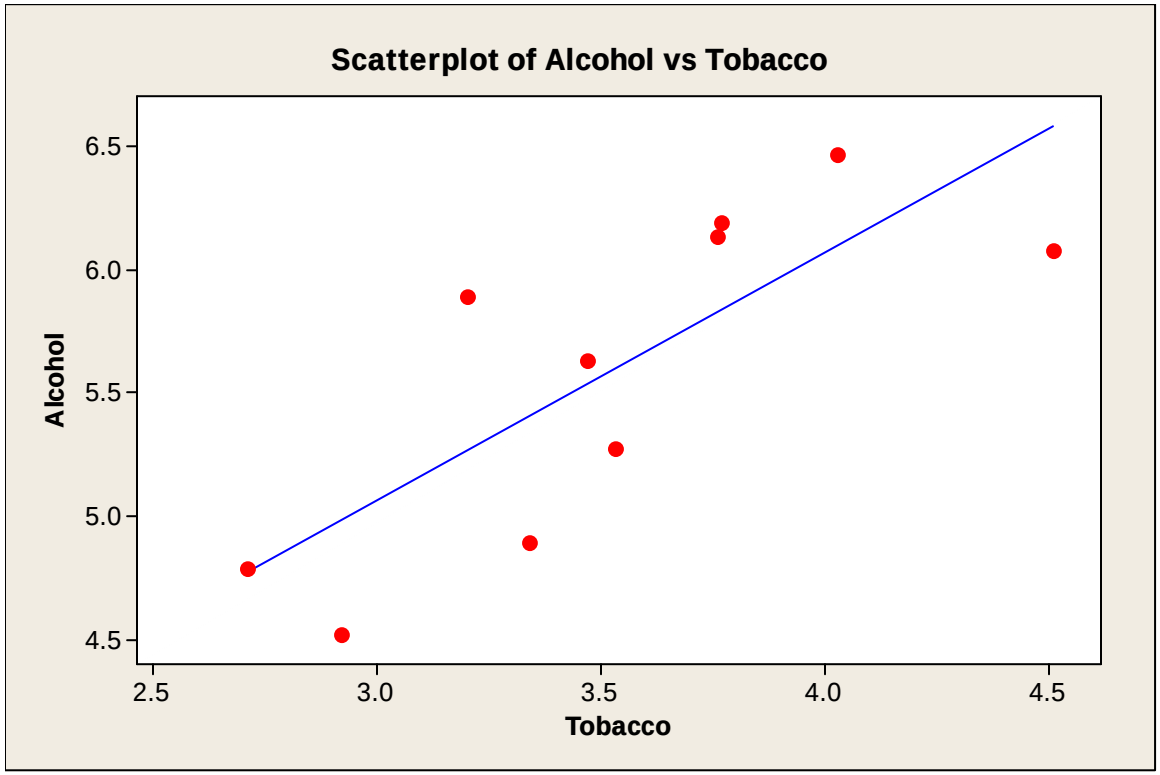
Why is it an outlier? It represents a different religion than other areas of Britain.



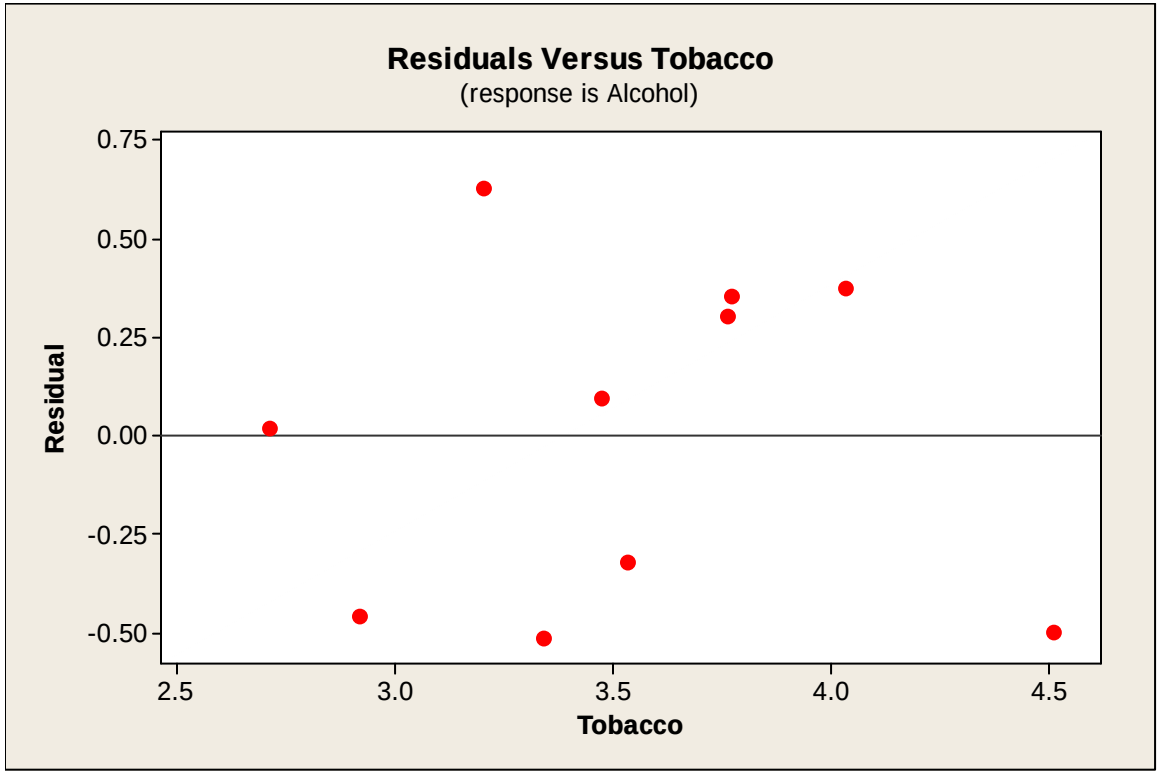
In the plot of residuals versus fitted values, it's even more obvious that the outlier is a problem.



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:

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

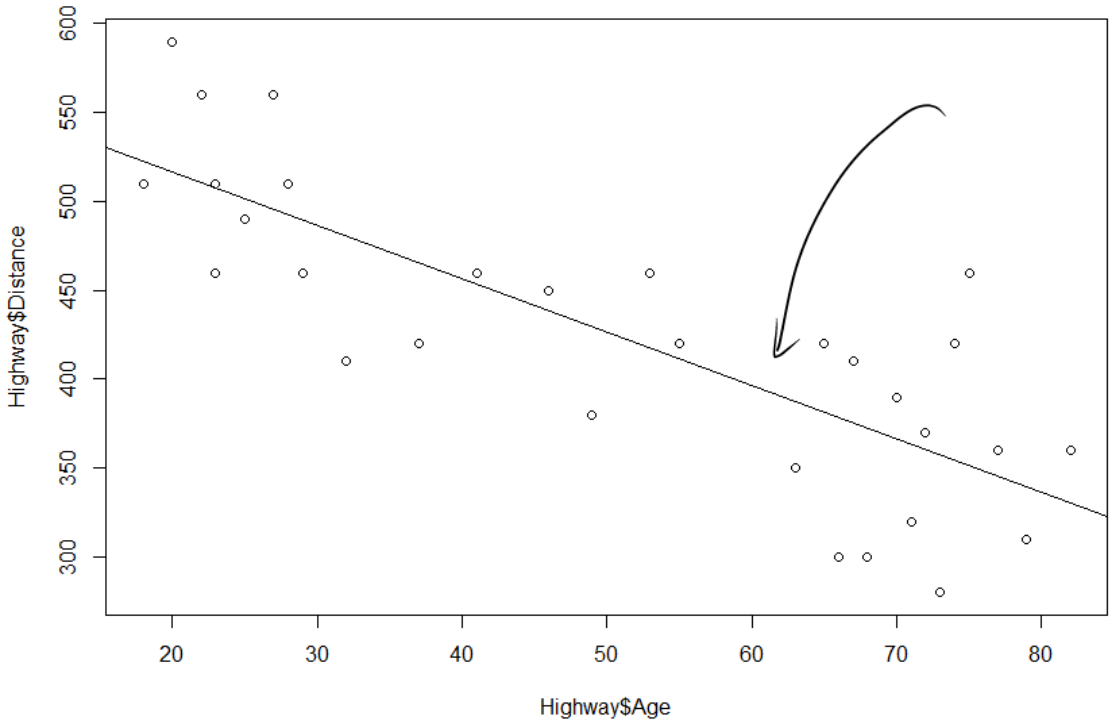
```
The regression equation is
Alcohol = 2.04 + 1.01 Tobacco

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%
```

R Commands for Highway Sign Data (as an example; posted on website too)
NOTE: This is the R session used to find the regression equation, and some plots for the Highway sign data. Download the file from the webpage.

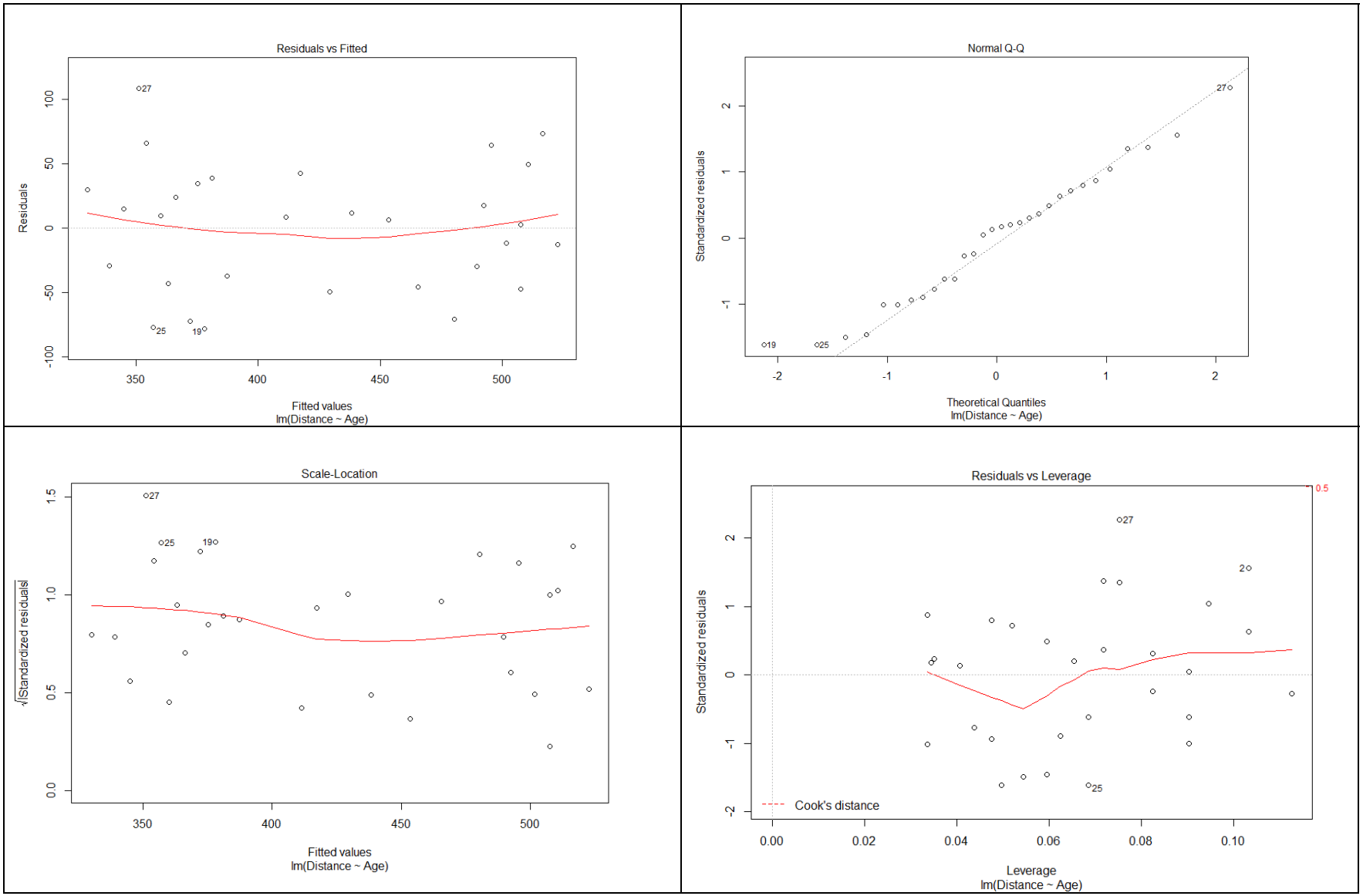
```
> #Read in the data.
> #sep="\t" shows that the columns are separated with
a tab.
> #header=F says there is no beginning line with
variable names.
> #col.names provides names for the two columns.
> Highway<-read.table("HighwaySign.txt", header=F,
sep="\t", col.names=c("Age","Distance"))

> #Make sure it worked by printing out first 6 lines:
> head(Highway)
  Age Distance
1  18      510
2  20      590
3  22      560
4  23      510
5  23      460
6  25      490
```

```
> #Create the regression model. Call it "HWModel"
> HWModel<-lm(Distance~Age,data=Highway)
> #See a summary of the model, including
coefficients, etc.
> summary(HWModel)
Call:
lm(formula = Distance ~ Age, data = Highway)

Residuals:
    Min      1Q  Median      3Q     Max
-78.231 -41.710   7.646  33.552 108.831

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 576.6819    23.4709  24.570  < 2e-16 ***
Age          -3.0068     0.4243  -7.086 1.04e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 49.76 on 28 degrees of freedom
Multiple R-squared: 0.642,      Adjusted R-squared: 0.6292
F-statistic: 50.21 on 1 and 28 DF,  p-value: 1.041e-07
```

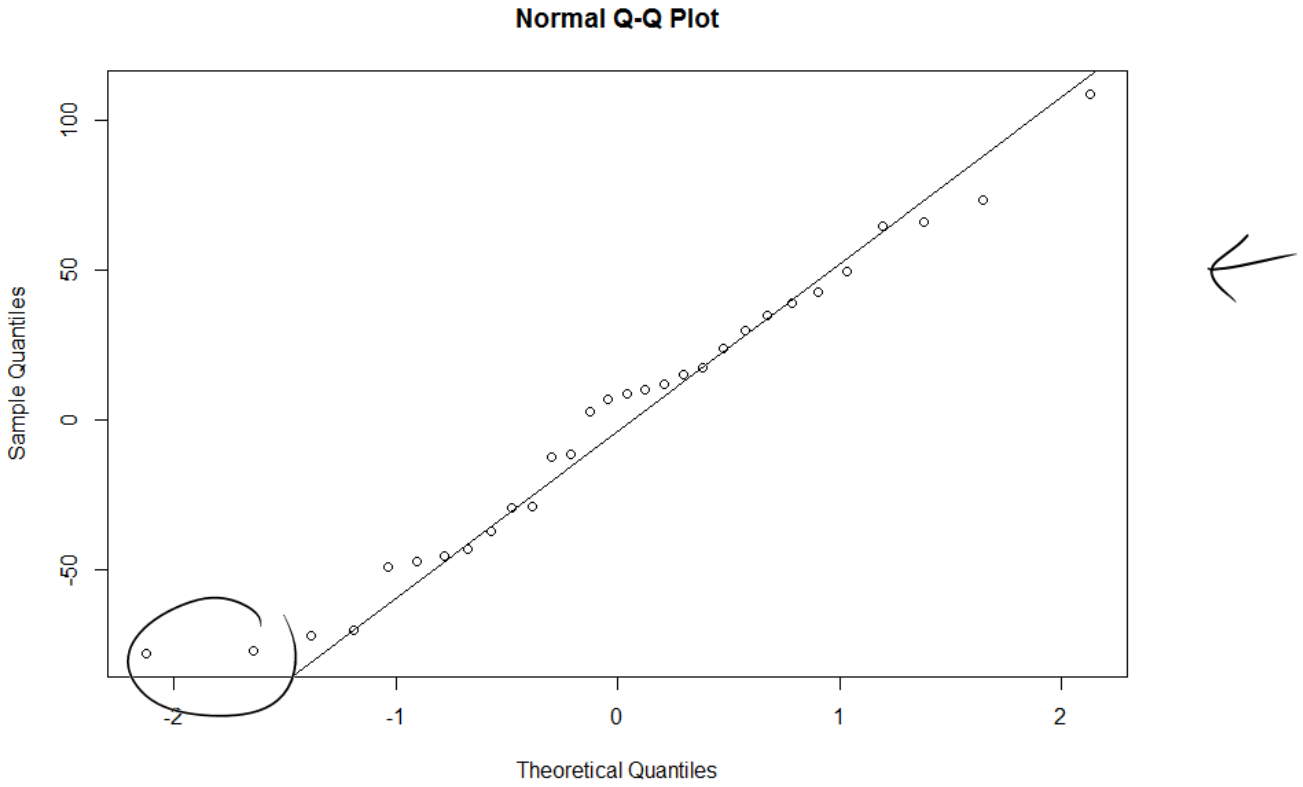
```
> #Create semi-studentized residuals
> Highway$StandardResids <- rstandard(HWModel)

> #Plot Age versus distance; add least squares line
> plot(Highway$Age,Highway$Distance)
> abline(HWModel)
```



```
> #Get four residual plots all together (“leverage” explained later in course)
> plot(HWModel)
```



```
> #Get a normal probability plot of the residuals and
add a line
> qqnorm(HWModel$resid)
> qqline(HWModel$resid)
```



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 *Data*, to square the response variable *GPA* and add it to the data table, type:

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

To take its *square root*, type:

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

To take its *natural logarithm*, type:

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

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

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

To take its *reciprocal*, type:

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

To take its *reciprocal square root*, type:

```
> Data <- cbind(Data, 1/sqrt(Data$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.

OLD $\quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$$\begin{cases} \mu(x_i) = E(Y_i|x_i) = \beta_0 + \beta_1 x_i \\ \mu((x_i+1)) \Leftarrow \beta_0 + \beta_1 (x_i+1) \\ \mu(x_i) = \beta_0 + \beta_1 (x_i) \end{cases}$$

$$\Rightarrow \mu(x_{i+1}) - \mu(x_i) = \beta_1 \;//$$

$$x_i^* = \sqrt{x_i}$$

$$Y_i = \beta_0 + \beta_1 x_i^* + \varepsilon_i$$

$$\beta_1 = \mu(\sqrt{x_i}+1) - \mu(\sqrt{x_i})$$