To Do

Read Sections 6.1 – 6.3.

Assignment 4 is due Friday November 25.

Last Class

(1) Confidence Interval for Slope β and Testing H_0: $\beta = \beta_0$

(2) Test of No Relationship between Response and Explanatory Variates

(3) Confidence interval for the mean response $\mu(x) = \alpha + \beta x$

(4) Prediction Interval for an Individual Response Y

Simple Linear Regression Model

For data (x_i, y_i), $i = 1,2,\ldots,n$
we assume the model

$Y_i \backsim G(\alpha + \beta x_i, \sigma)$ for $i = 1,2,\ldots,n$

independently and where the
x_i's, $i = 1,2,\ldots,n$
are assumed to be known constants.

Today's Class

(1) General Form of a Gaussian Response Model

(2) Linear Regression Models

(3) Checking the Assumptions of the Simple Linear Regression Model

Simple Linear Regression Model

Simple linear regression model

$Y_i \backsim G(\alpha + \beta x_i, \sigma)$ for $i = 1,2,\ldots,n$

independently and where the x_i's, $i = 1,2,\ldots,n$ are assumed to be known constants.

This model is a member of a larger family of models called Gaussian response models.

Gaussian Response Models

The general form of a Gaussian response model is

$$Y_i \backsim G(\mu(x_i), \sigma) \quad \text{for} \quad i = 1,2,\ldots,n$$

independently and where the x_i's, $i = 1,2,\ldots,n$ are assumed to be known constants (possibly vectors).

In this model

$$E(Y_i) = \mu(x_i)$$

depends on the explanatory variate x_i, but

$$sd(Y_i) = \sigma$$

does not.

Gaussian Response Model

The Gaussian Response Model

$Y_i \backsim G(\mu(x_i), \sigma)$ for $i = 1,2,\ldots,n$ independently can also be written in the form

$Y_i = \mu(x_i) + R_i$ where $R_i \sim G(0,\sigma)$, $i = 1,2,\ldots,n$ independently.

Y_i is a sum of two components.

The first component, $\mu(x_i)$, is a deterministic component (not a random variable) and the second component, R_i , is a random component or random variable.

Linear Regression Models and STAT 331/STAT 371/STAT 373

In many examples the deterministic component takes the form

$$E(Y_i) = \mu(\boldsymbol{x}_{\mathbf{i}}) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

so E($Y_{\mathbf{i}}$) is a linear function of
$\boldsymbol{x_i}$ = (x_{i1},x_{i2},…,x_{ik}), a vector of explanatory variates for unit *i*, and the unknown parameters $\boldsymbol{\beta_0}$,$\boldsymbol{\beta_1}$,…, $\boldsymbol{\beta_k}$.

These models are called linear regression models.
The $\boldsymbol{\beta_j}$ '*s* are called the regression coefficients.
The $\boldsymbol{x_i}$'s are called covariates.

Model Checking

There are two main assumptions for Gaussian linear response models:

(1) Y_i (given covariates x_i) has a Gaussian distribution with standard deviation σ which does not depend on the covariates.

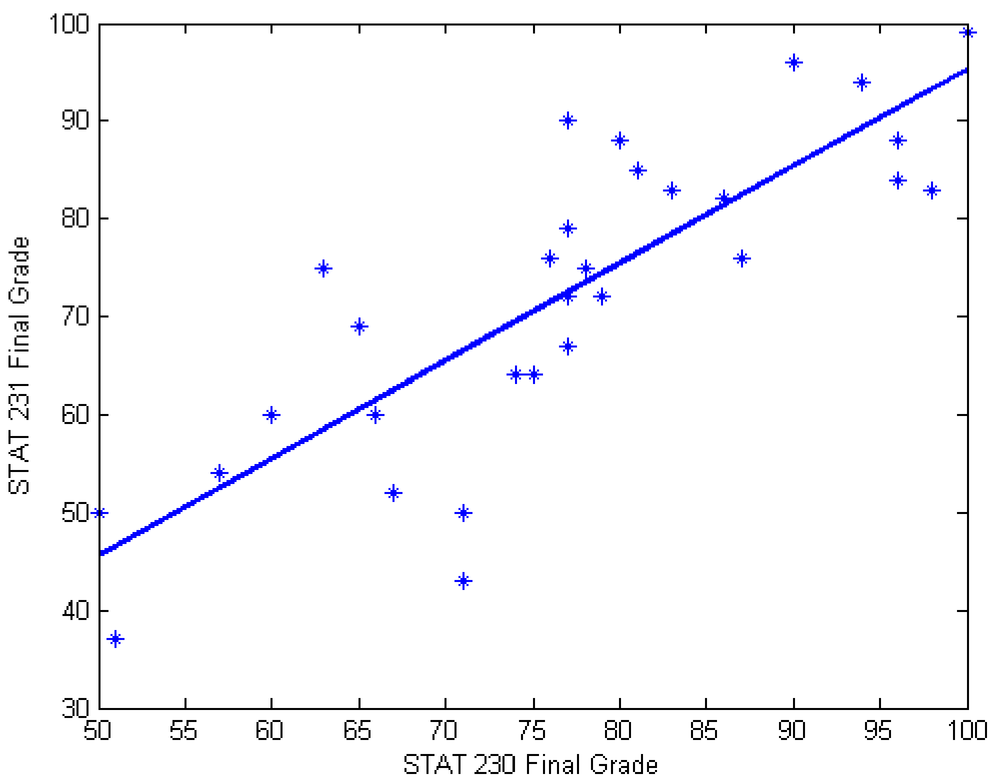
(2) $E(Y_i) = \mu(x_i)$ is a linear combination of known covariates $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$, and unknown regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$.

MODEL ASSUMPTIONS SHOULD ALWAYS BE CHECKED!!!

We use graphical methods to do this.

Model Checking Method 1

In simple linear regression, a scatterplot of the data with the fitted line superimposed indicates how well the model fits the data.



Model Checking Method 2 - Residual Plots

Residual plots are very useful for model checking when there are 2 or more covariates. For the simple linear regression model let

$$\hat{\mu}_i = \hat{\alpha} + \hat{\beta}_i x_i$$

(often called the "fitted" response)

and

$$\hat{r}_i = y_i - \hat{\mu}_i$$

The $\hat{r}_i's$ **are called residuals since** $\hat{r}_i$ **represents what is "left" after the model has been "fitted" to the data.**

Residual Plots

The idea behind the $\hat{r}_i$'s is that they can be thought of as "observed" R_i's in the model

$Y_i = \mu_i + R_i$ where $R_i \sim G(0,\sigma)$, $i = 1,2,\ldots,n$

independently.

This isn't exactly correct since we are using $\hat{\mu}_i$ instead of μ_i.

However if the model is correct, then the $\hat{r}_i$'s should behave roughly like a random sample from the $G(0,\sigma)$ distribution.

Residual Plots

Recall

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \quad \text{or} \quad \overline{y} - \hat{\alpha} - \hat{\beta}\overline{x} = 0$$

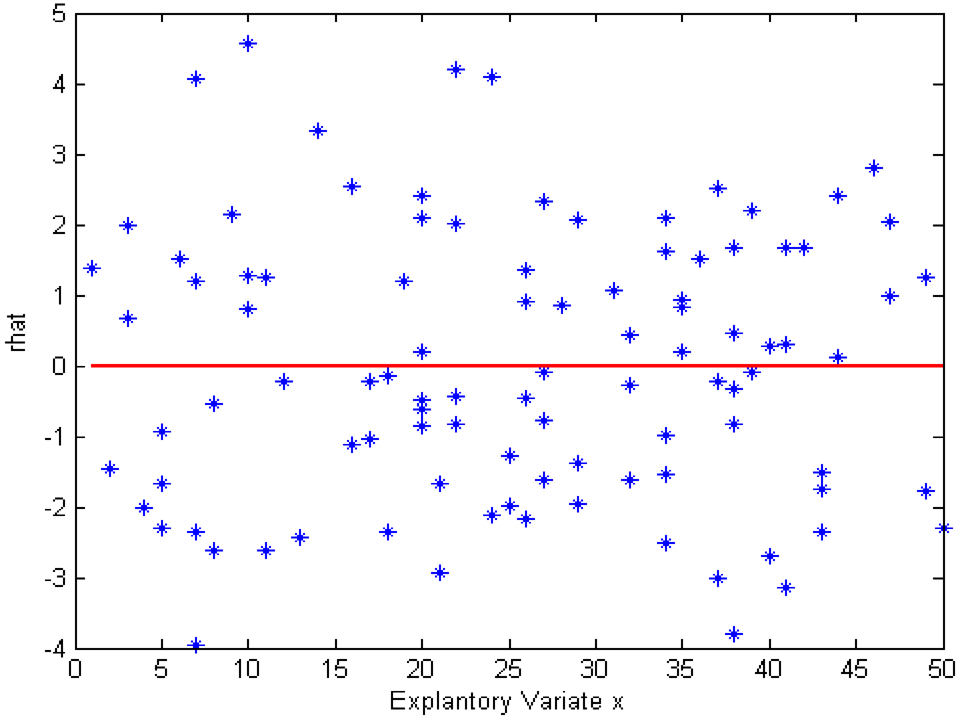
which implies

$$0 = \overline{y} - \hat{\alpha} - \hat{\beta}\overline{x} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = \frac{1}{n}\sum_{i=1}^{n}\hat{r}_i$$

so that the average of the residuals is always zero.

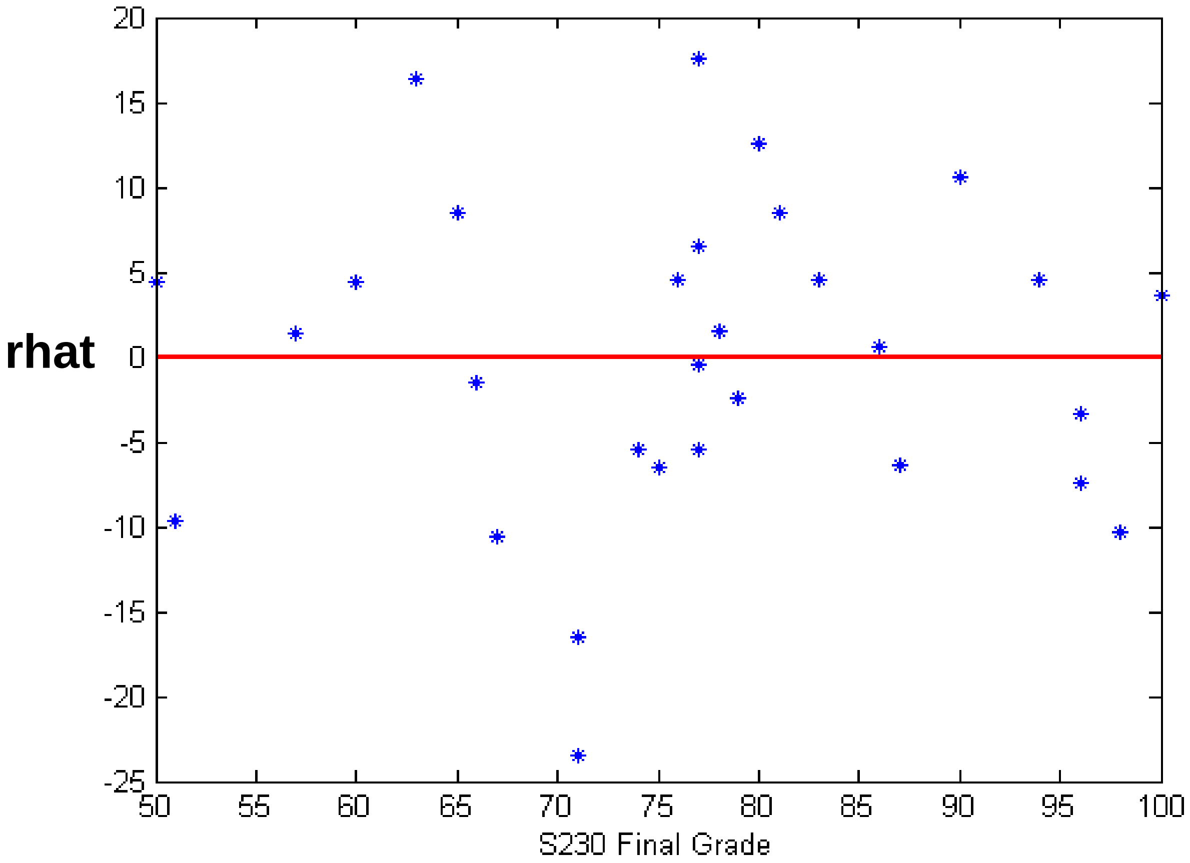
Residual Plots

If the model assumptions hold then a plot of the points $(x_i, \hat{r}_i),\ i = 1,2,\ldots,n$ should lie more or less within a horizontal band or belt around the line $\hat{r}_i = 0$ showing no obvious pattern.



STAT 231/230 Residual Plot

What would you conclude?



Standardized Residual Plots

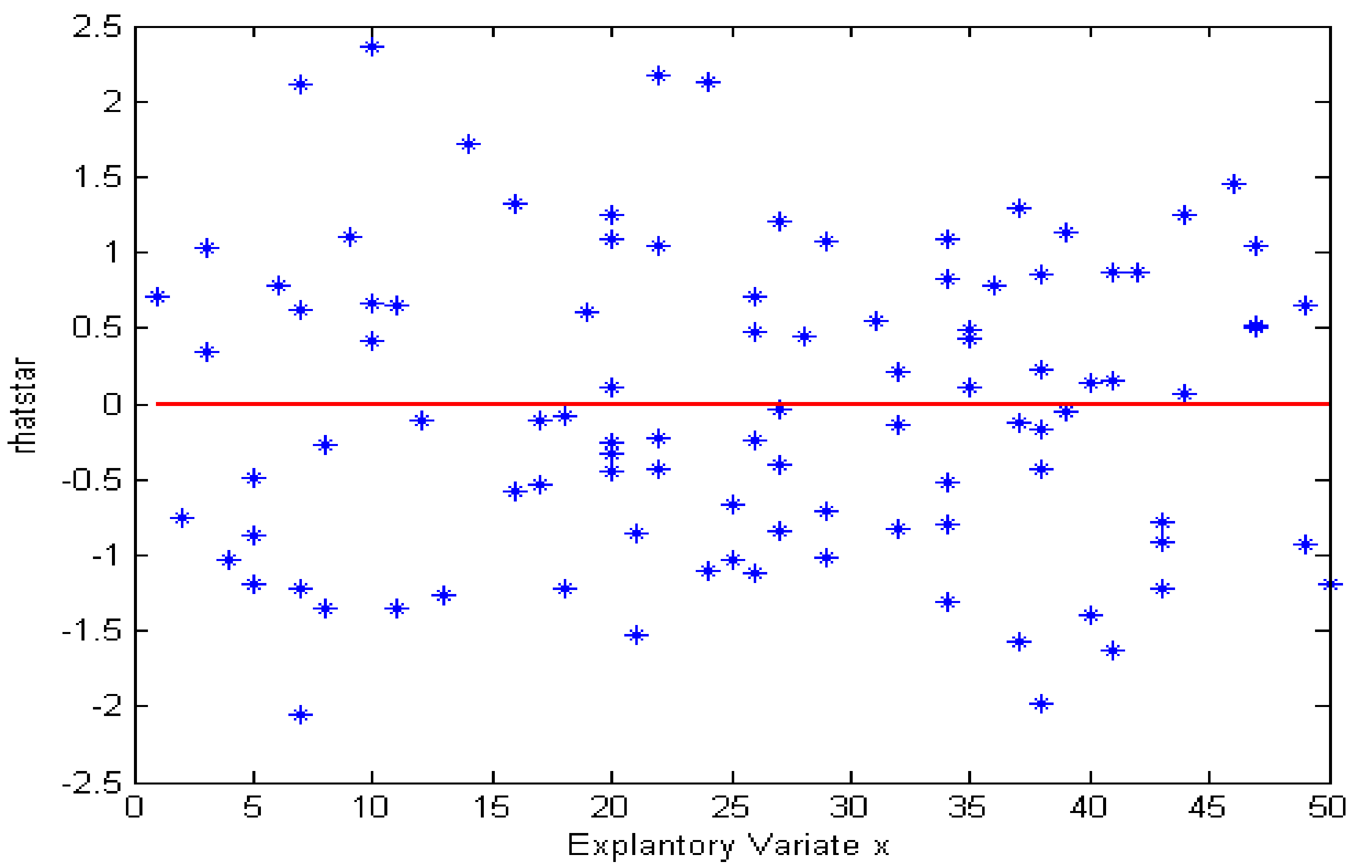
Define the standardized residuals

$$\hat{r}_i^* = \frac{\hat{r}_i}{s_e} = \frac{y_i - \hat{\mu}_i}{s_e} \qquad i = 1,2,\ldots,n$$

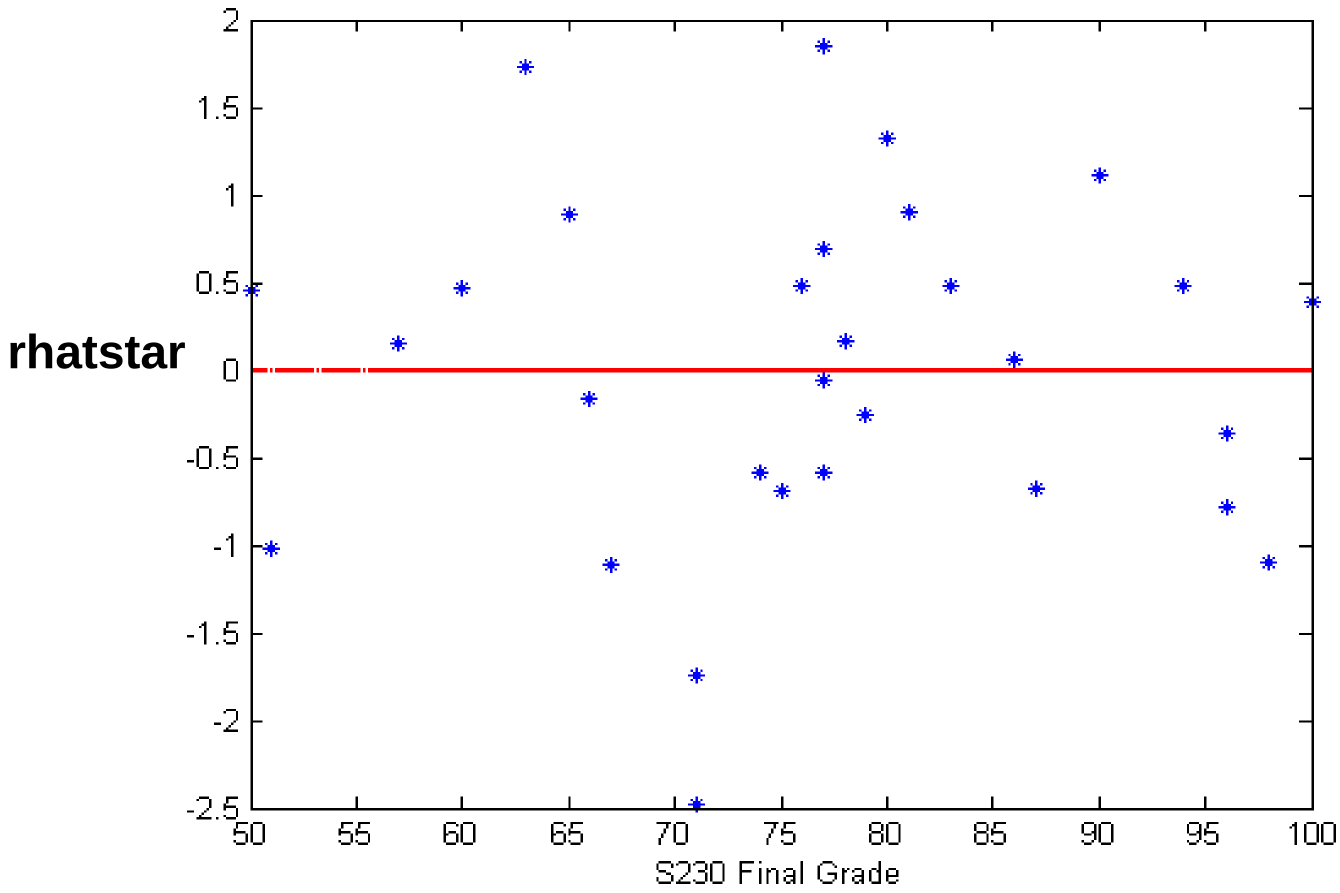
What is the only difference between a plot of the points $(x_i, \hat{r}_i),\ i = 1,2,\ldots,n$ **and a plot of the points** $(x_i, \hat{r}_i^*),\ i = 1,2,\ldots,n$ **?**

If the model is correct then the $\hat{r}_i^*$ **values will lie in the range (-3,3). Why is this?**

Example – Standardized Residual Plot



STAT 231/230 Standardized Residual Plot



Residual Plot Type 2

Another type of residual plot consists of plotting the points

$$\left(\hat{\mu}_i, \hat{r}_i^*\right), \;\; i = 1, 2, \ldots, n$$

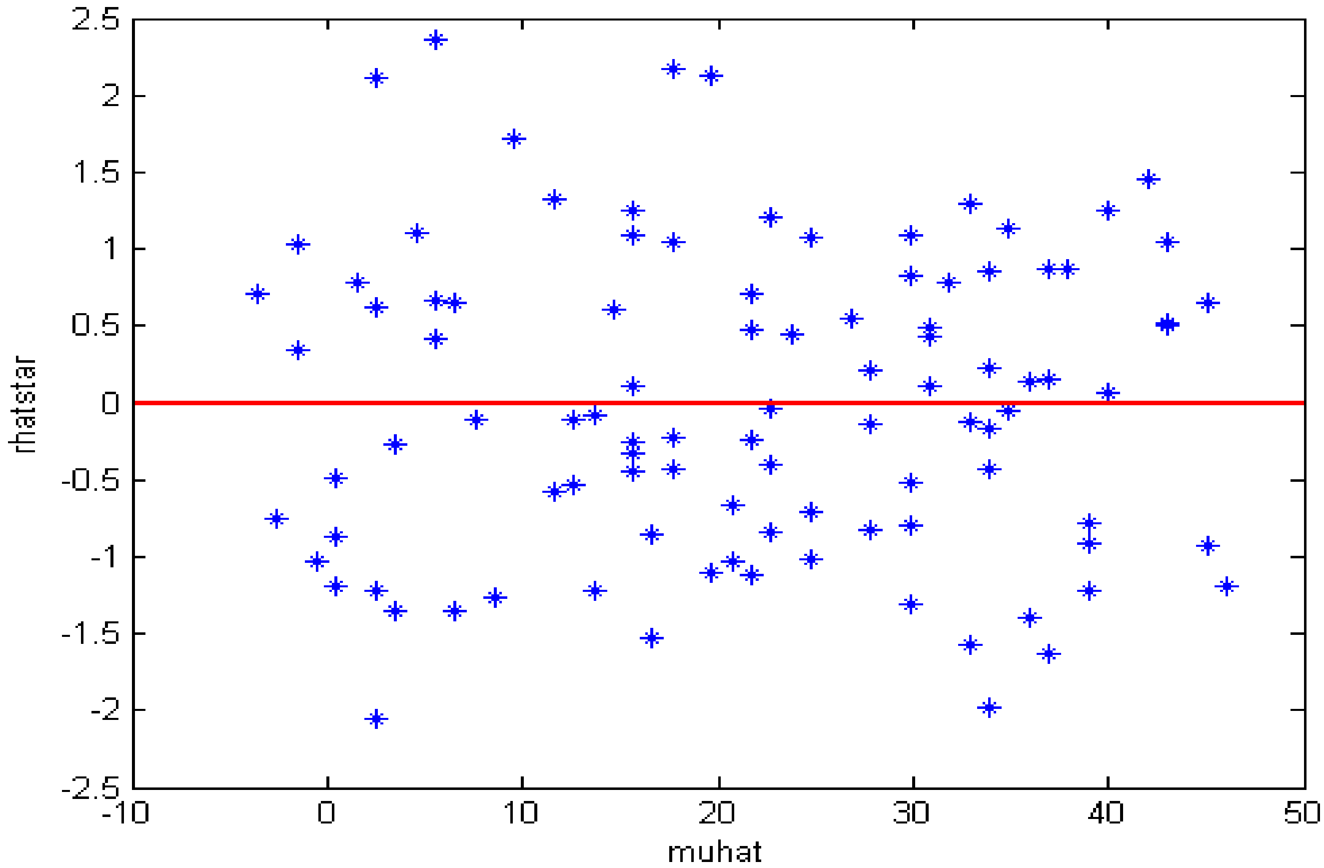
Such a plot can be used to check the assumption about the form of the mean E(Y_i) = $\mu(x_i)$.

For the simple linear regression model we are checking whether the assumed mean E(Y_i) = $\mu(x_i)$ = $\alpha + \beta x_i$ is reasonable.

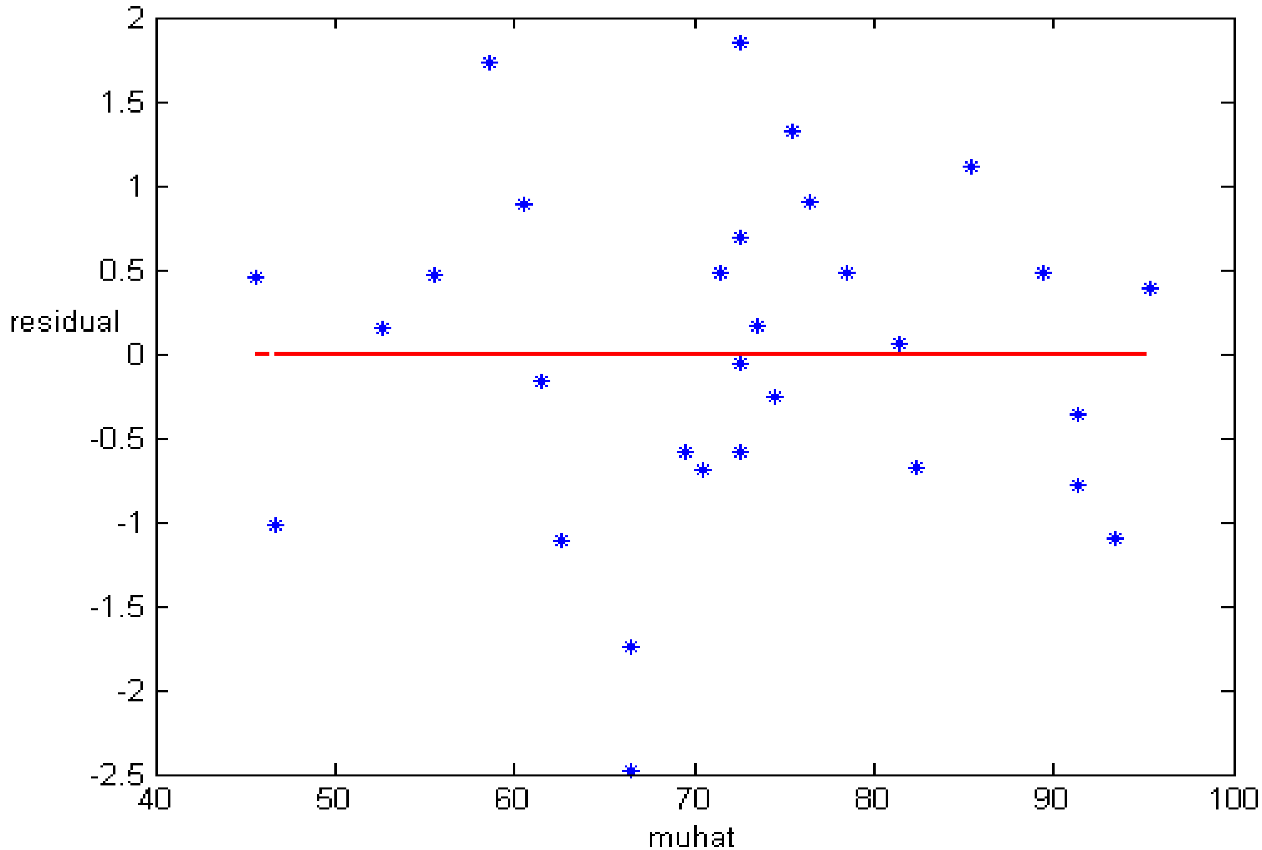
If the assumed mean is reasonable we should see approximately a horizontal band around the line

$\hat{r}_i^* = 0$.

Example – Standardized Residual Plot Using Muhat



STAT 231/230 Standardized Residual Plot with Muhat



Qqplot of Residuals

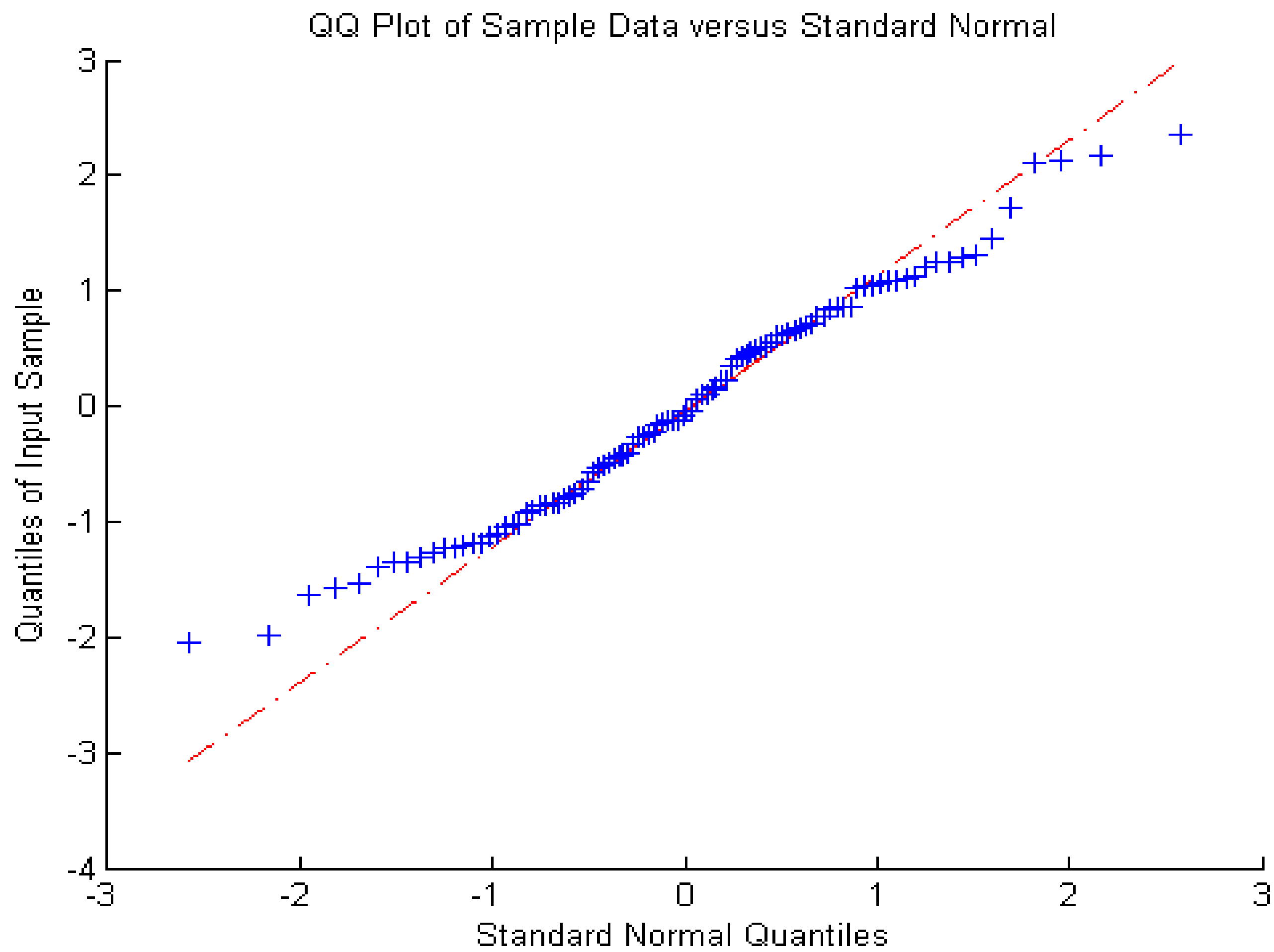
To check the Gaussian assumption we use a qqplot of the standardized residuals.
Since our assumed model is

$$R_i / \sigma = (Y_i - \mu_i)/\sigma \sim G(0,1)$$

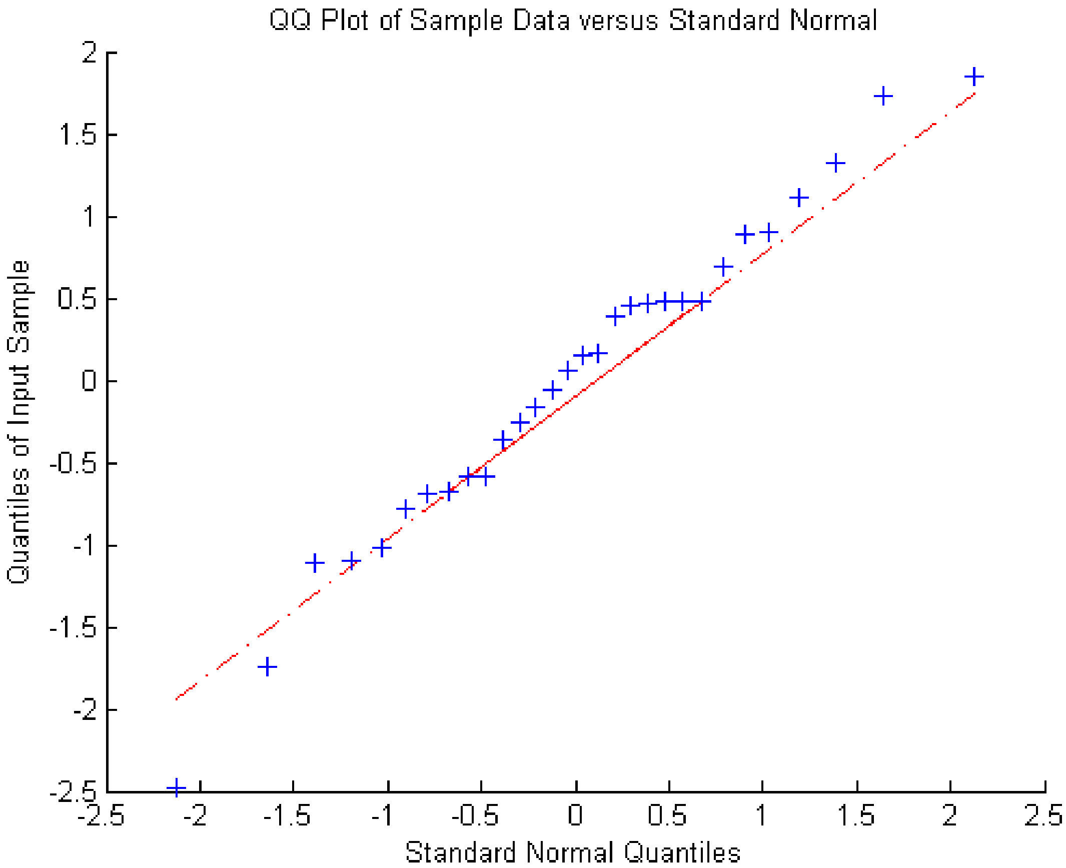
the $\hat{r}_i^{*}$**'s should roughly represent a sample from the G(0,1) distribution.**

Therefore a qqplot of the $\hat{r}_i^{*}$**'s should give approximately a straight line if the model assumptions hold.**

Example - Qqplot



STAT 231/230 Qqplot



Interpreting Residual Plots

If a plot of the points

$$\left(\hat{\mu}_i, \hat{r}_i^*\right),\ i = 1,2,\ldots,n$$

or

$$\left(\hat{\mu}_i, \hat{r}_i\right),\ i = 1,2,\ldots,n$$

shows a distinctive pattern then this suggests the assumed form for $E(Y_i) = \mu(x_i)$ may be inappropriate.

Interpreting Residual Plots

If a plot of the points

$$\left(\hat{\mu}_i, \hat{r}_i^*\right),\ i = 1,2,\ldots,n$$

indicates that the variability in the $\hat{r}_i^*$**'**s

is bigger for large values of $\hat{\mu}_i$

than for small values of $\hat{\mu}_i$

(or vice versa) then there is evidence to suggest that the assumption of constant variance, Var(Y_i**) = Var(**R_i**) =** σ^2**,** i**=1,2,…,n does not hold.**

Interpreting Residual Plots

If the points in the qqplot do not lie roughly in a straight line then this suggests the Gaussian assumption may not hold.

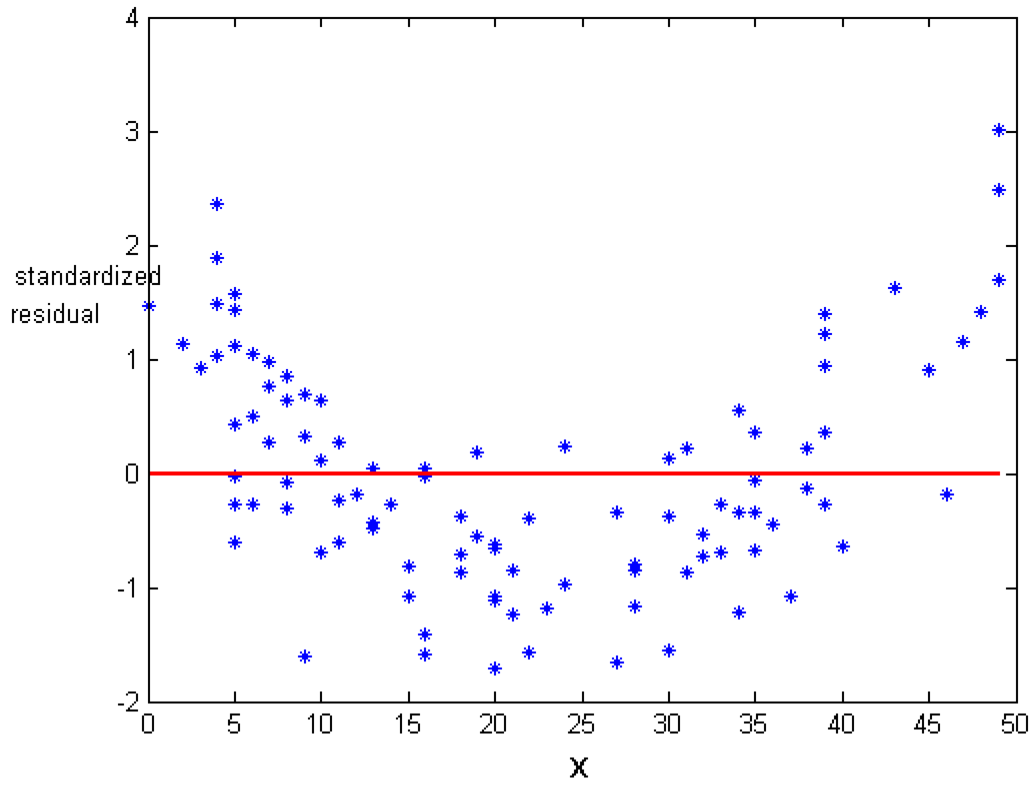
Interpreting Residual Plots - Warning

Reading these plots takes practice and you should try not to read too much into plots especially if the plots are based on a small number of points.

The plots on the next slides exhibit patterns.

Examples of Residual Plots with Patterns

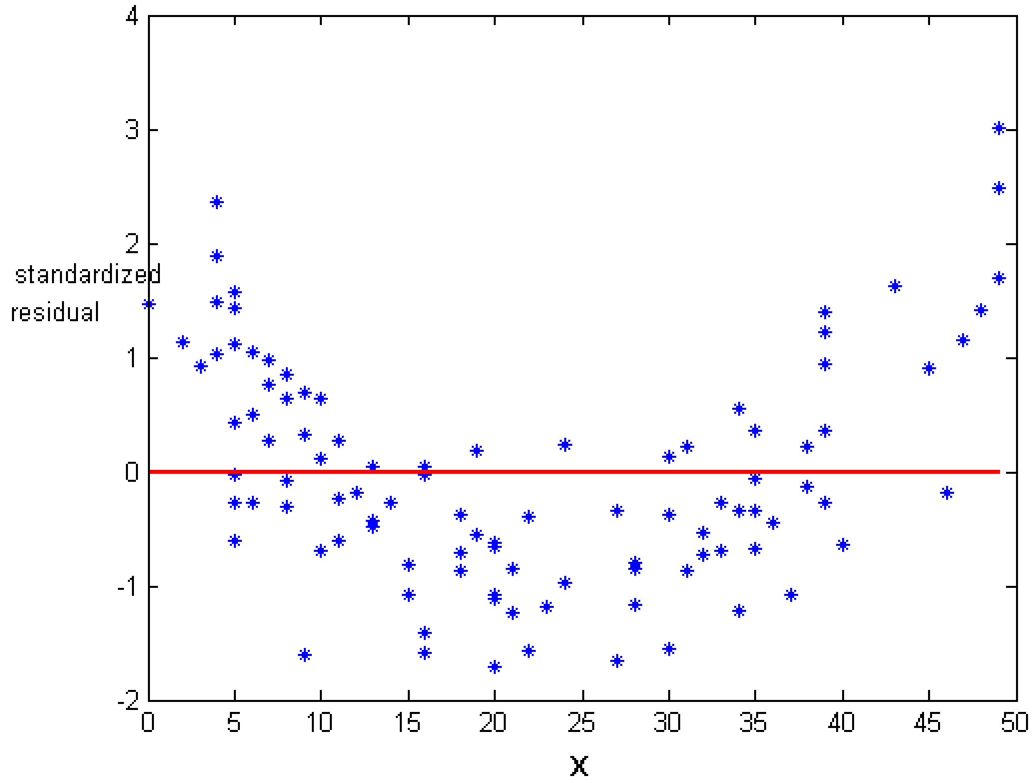
This plot suggests that the function $\mu(x_i)$ is not correctly specified. Can you suggest a better model?



Examples of Residual Plots with Patterns

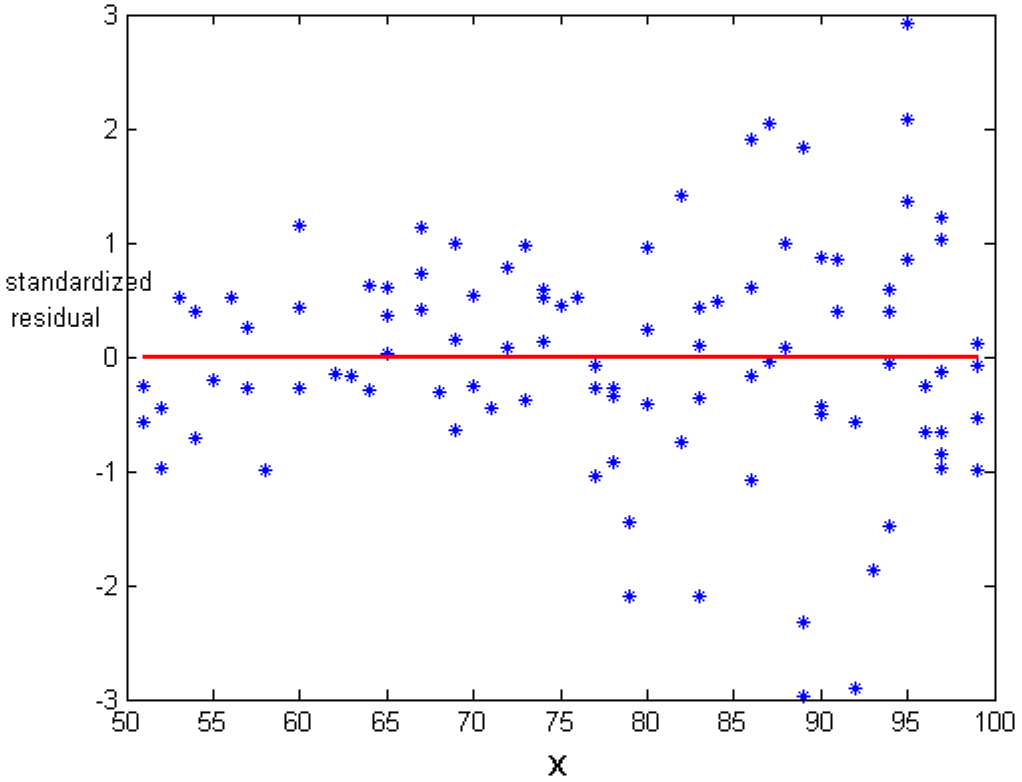
Assume a quadratic model for the mean:

$\mu(x_i) = \alpha + \beta x_i + \gamma x_i^2$ **rather than** $\mu(x_i) = \alpha + \beta x_i$



Examples of Residual Plots with Patterns

What do you notice?



Scatterplot for same data