Diagnostics in SLR

"Things to check behind those significant T-test and F test"

Diagnostics – Methods to check whether our model is reasonable for our data and representative of the system that we are studying.

Some diagnostics check the *assumptions* of our model. Other diagnostics check the *influence* of different data points.

Remedies – Analytic strategies used to fix problems identified by the diagnostics.

Why do we need to check the model?

> All models are wrong. Some models are useful.
>
> — George Box

The goal of building a model is to:

- *learn something* about the real world
- *predict outcomes* in the real world

To use a model successfully, we need to know its limitations:

- Does it adequately describe the functional relationship of interest?
- Is there reason to worry that inferences about the parameters might be flawed?
- Is the error distribution appropriate?

What do we need to check?

- Is the *functional form* of the model appropriate?
- Do any of the data points have a *disproportionate influence* on the parameter estimates?
- Are there *outliers?*
- Are the residuals in the data consistent with our model for random error:
  - Errors are *independent*,
  - Errors all have the *same variance*, σ^2
  - Errors are *normally distributed*.

How do we check?

- Diagnostic plots and tests on **residuals (major content in this topic)**
  - Plot of residual vs predictor variable.
  - Plot of residual against fitted value
  - Plot of residual against time or other sequence
  - Plot of residual against omitted predictor variable.
  - Normal probability plot of residual
  - Brown-Forsythe test for non-constant variance (**heteroscedasticity)**
  - **Shapiro test for non-Normality**

- Diagnostic plots and tests **on dependent and independent variables**

Diagnostics based on residuals

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

The random errors ε_i are *independent* , *normal* , and should have *constant variance* .

$$\varepsilon \sim Normal(0, \sigma)$$

The residuals are computed from sample to simulate ε_i,

$$e_i = Y_i - \widehat{Y_i}$$

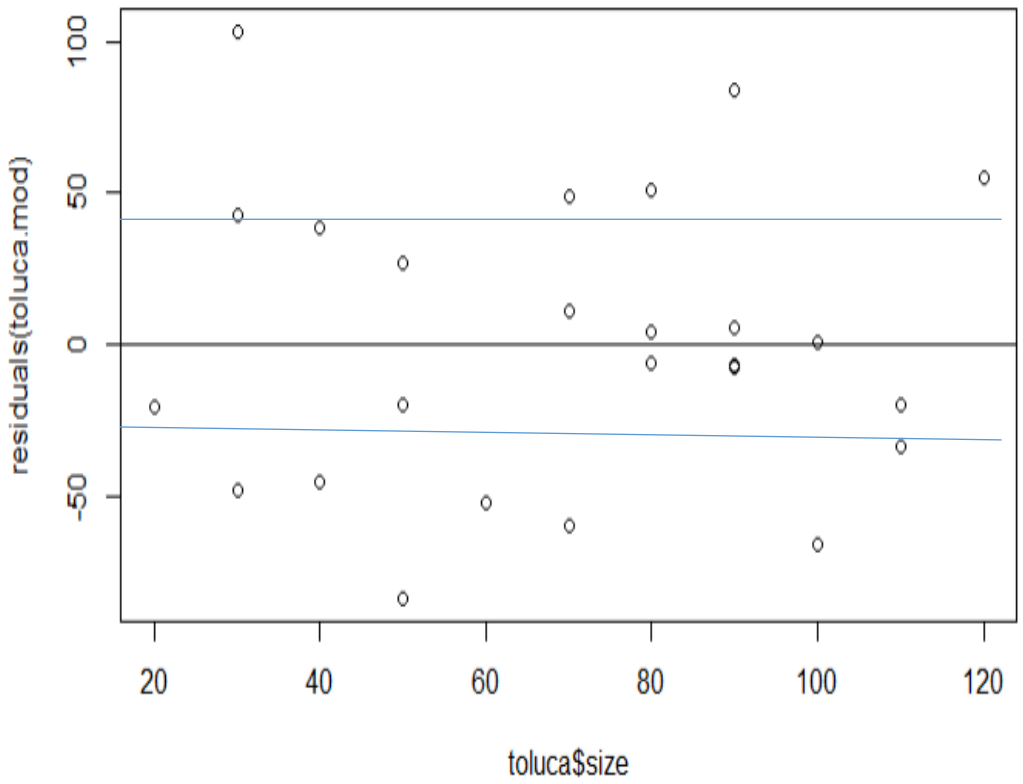
If the model fits, e_i should reflects the properties assumed for ε_i.

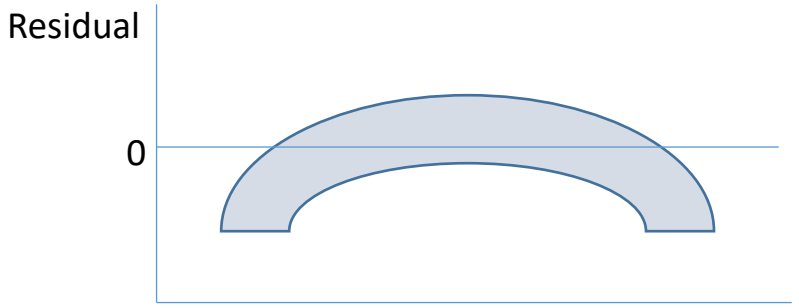
Questions Addressed by Diagnostics from Residuals plot

- Is the functional form appropriate (in SLR, that means linear)? →is there no curvature pattern?
- Does the variance depend on X ? → is there increase or decrease in average magnitude with the fitted values?
- Are the errors normal? → The normal plot of residual is straight?
- Are there outliers? → is there relatively large residual?
- Are the errors appearing independent? → Any patterns in the residual plot?

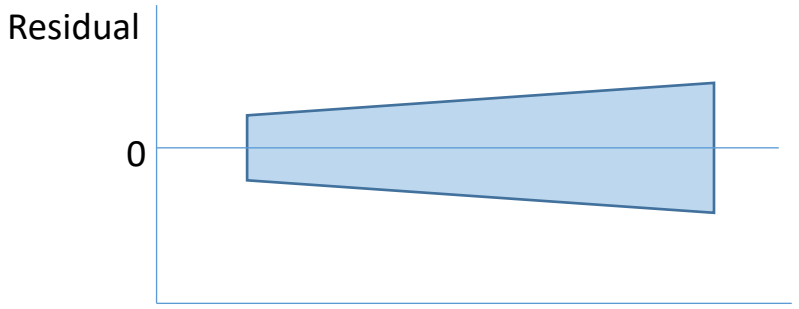
Note that $e_i = Y_i - \hat{Y}_i$ are not independent since each $\hat{Y}_i$ is computed with the same b_0 and b_1.
However, when the sample size (n) is much larger than the number of parameters (p),
We can ignore the minor dependence.

Prototype of good or problematic residual plot

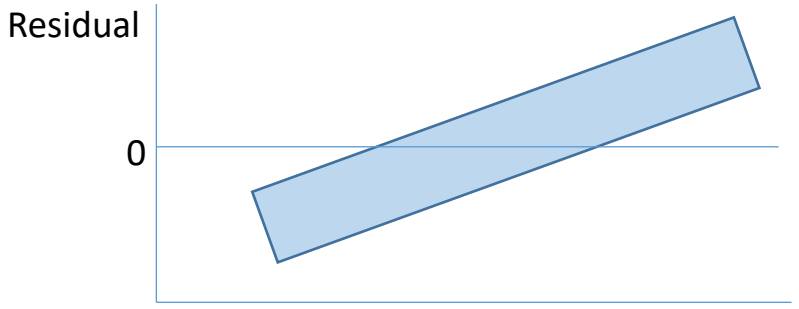




Curvilinear



Non-constant Variance



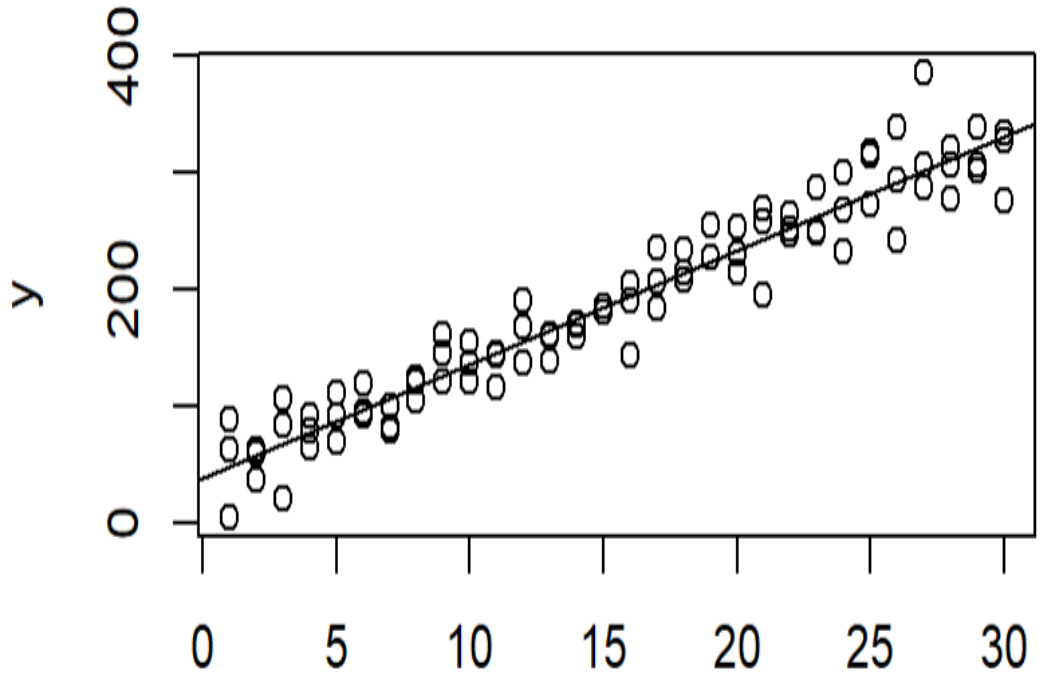
Dependent (omitting of Important predictor)

Some example of violations and diagnostics from residual plot

- linear relationship
- constant variance
- normal errors
- Independence
- outliers

A case with perfect linear relationship

$$Y = 10X + 30 + N(0,25)$$

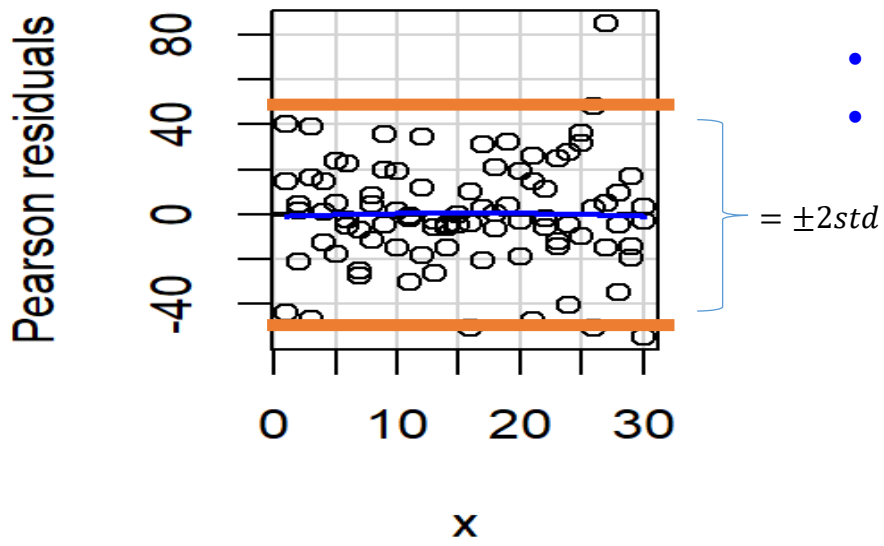



```r
x<-rep(seq(1:30),3)
y<-10*x+30+rnorm(90, 0,25)
SLRdata<-data.frame(x,y)
SLRdataRM<-lm(y~x, SLRdata)
plot(x,y)
abline(SLRdataRM)
summary(SLRdataRM)
anova(SLRdataRM)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.0032     5.1771   7.341 1.01e-10 ***
x             9.7570     0.2916  33.458  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 23.95 on 88 degrees of freedom
Multiple R-squared:  0.9271,    Adjusted R-squared:  0.9263
F-statistic:  1119 on 1 and 88 DF,  p-value: < 2.2e-16
```

A case with perfect linear relationship

$$Y = 10X + 30 + N(0,25)$$



$= \pm 2std$

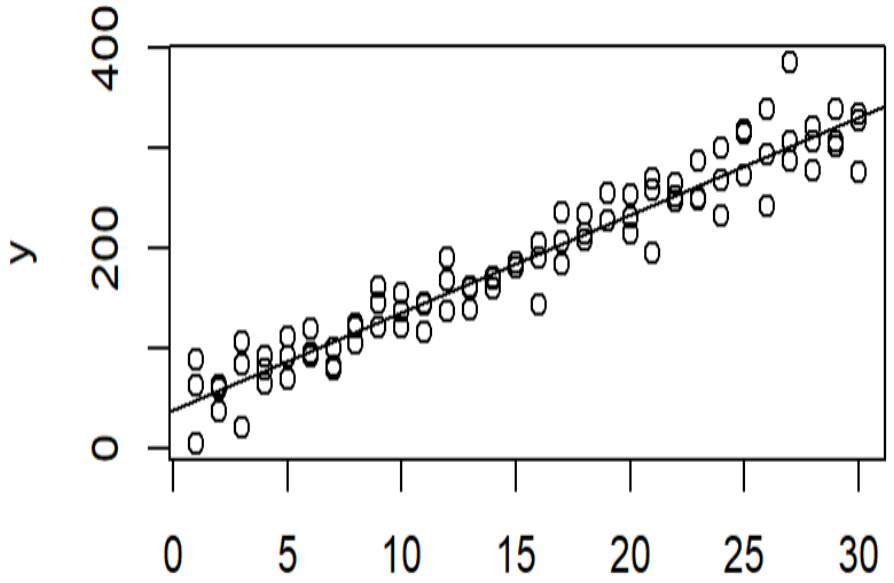
- The residual plot shows the residuals scatter evenly around a flat line of 0,
- The variance is constant across X, no obvious pattern.

```
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x           1 641876  641876  1119.4 < 2.2e-16 ***
Residuals  88  50458     573
```

- $\sqrt{MSE} = s\{residual\} = \sqrt{573} = 24$ which is a good estimate of the actual s{random error}=25. The model is **efficient**.

A case with perfect linear relationship

$$Y = 10X + 30 + N(0,25)$$



- From the scatter plot and the residual plot, we can observe that the mean response prediction and the actual mean For each X level are close. Hence no lack-of-fit is expected.

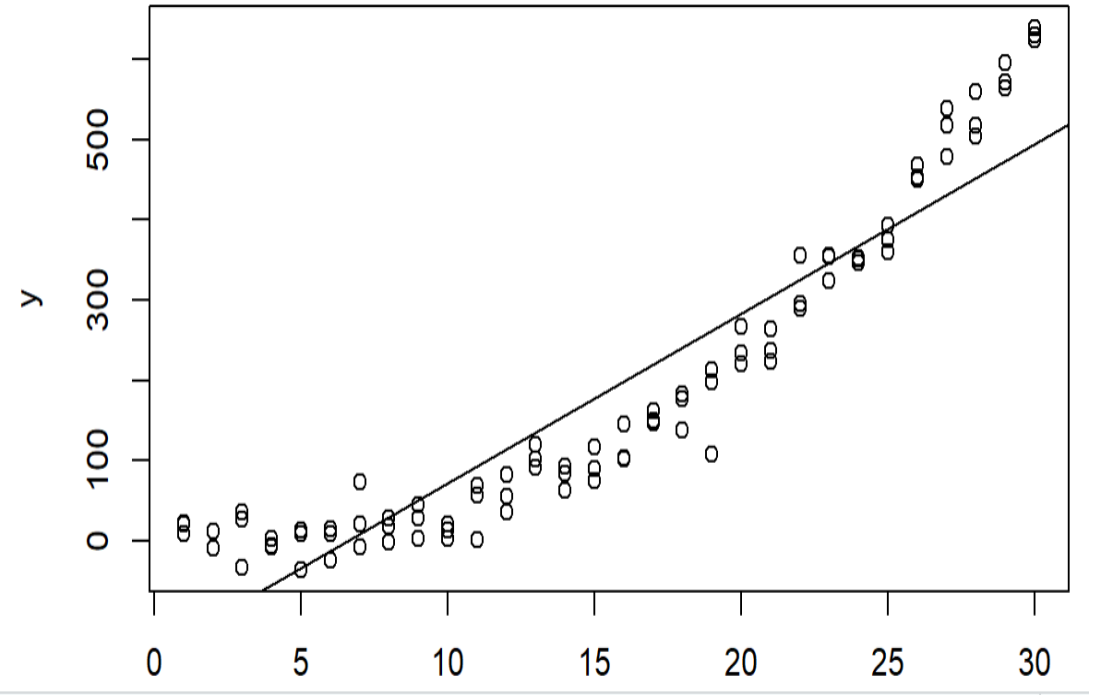
```
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ as.factor(x)
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     88 50458
2     60 41055 28    9403.5 0.4908 0.9793
```

- $Fs = \frac{MSLF}{MSPE} = \frac{9403/28}{41055/60} = \frac{336}{684} = 0.491$
- $\sqrt{MSPE} = s\{pure\ error\} = \sqrt{684} = 26.15$ which is also a good estimate of the actual s{random error}=25

Make a dataset that we know is quadratic, not linear.

$Y = 30 - 10X + X^2 + N(0, 25)$

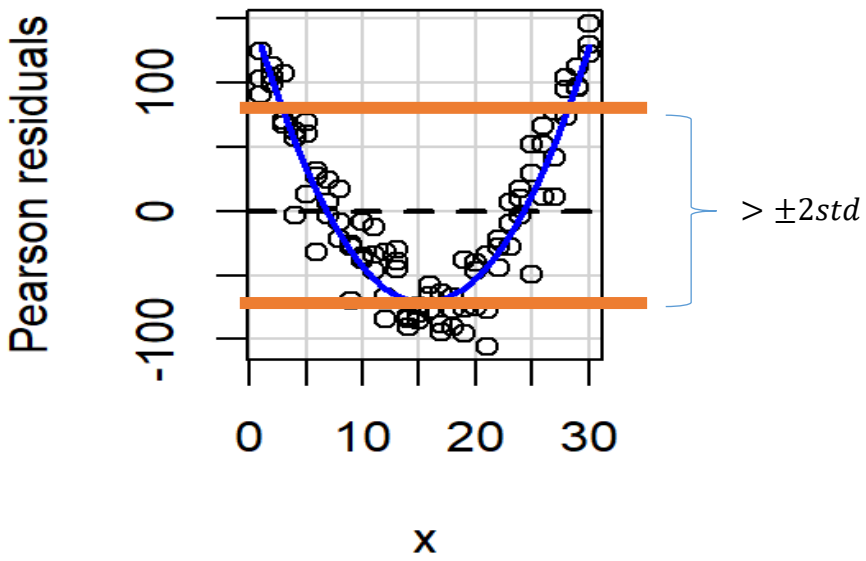
```
x<-rep(seq(1:30),3)
y<-x^2-10*x+30+rnorm(90, 0,25)
nl<-data.frame(x,y)
nonlinearRM<-lm(y~x, nl)
plot(x,y)
abline(nonlinearRM)
summary(nonlinearRM)
anova(nonlinearRM)
```




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -138.6100    15.4200  -8.989 4.32e-14 ***
x             21.0999     0.8686  24.292  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 71.32 on 88 degrees of freedom
Multiple R-squared:  0.8702,    Adjusted R-squared:  0.8688
F-statistic: 590.1 on 1 and 88 DF,  p-value: < 2.2e-16
```

- The summary shows that the linear impact is significant, the R-square is 88.26%, and the model is significant.
- Is the model good?

$$Y = 30 - 10X + X^2 + N(0, 25)$$

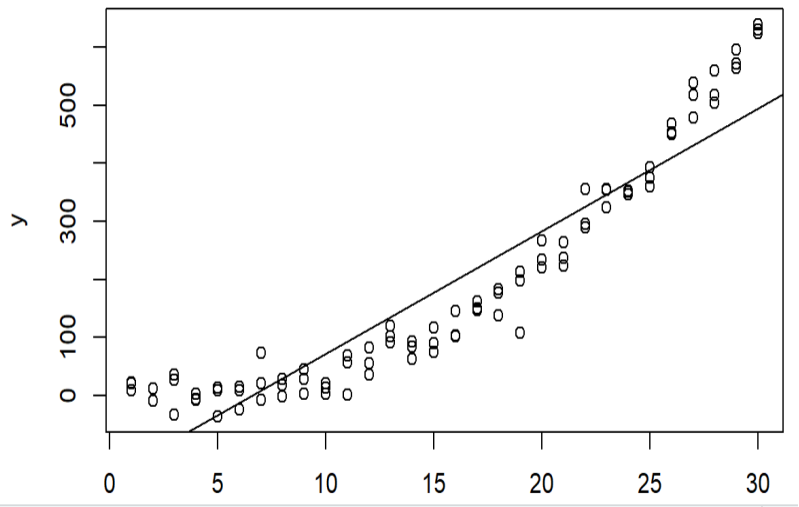


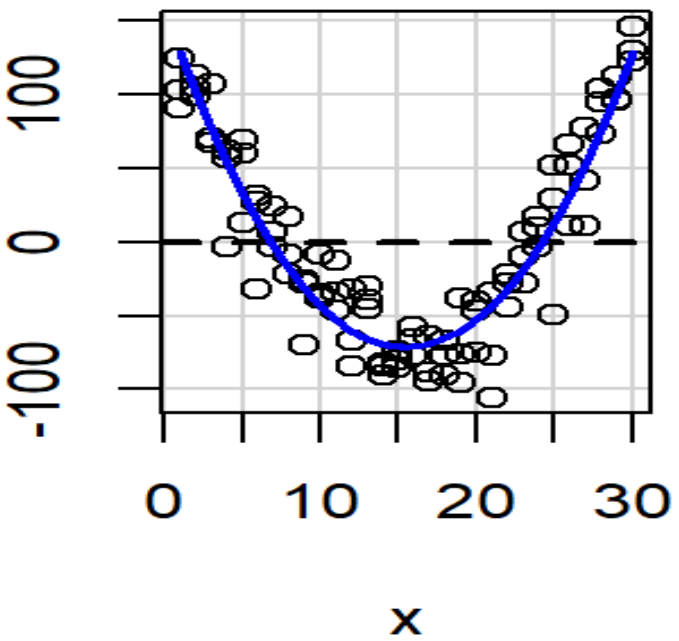
- The residual plot shows that the residuals do not scatter evenly around the flat line of 0.
- The variances are not constant across X.

```
Response: y
           Df  Sum Sq Mean Sq F value    Pr(>F)
x           1 3001798 3001798  590.11 < 2.2e-16 ***
Residuals  88  447641    5087
```

- $\sqrt{MSE} = s\{residual\} = \sqrt{5087} = 71$, while the actual S{random error}=25.
- The actual deviation is overestimated. The model is **inefficient**.

$$Y = 30 - 10X + X^2 + N(0, 25)$$





- From the scatter plot and the residual plot, we can observe that the mean response prediction and the actual mean For each X level are always off. Hence lack-of-fit is expected.

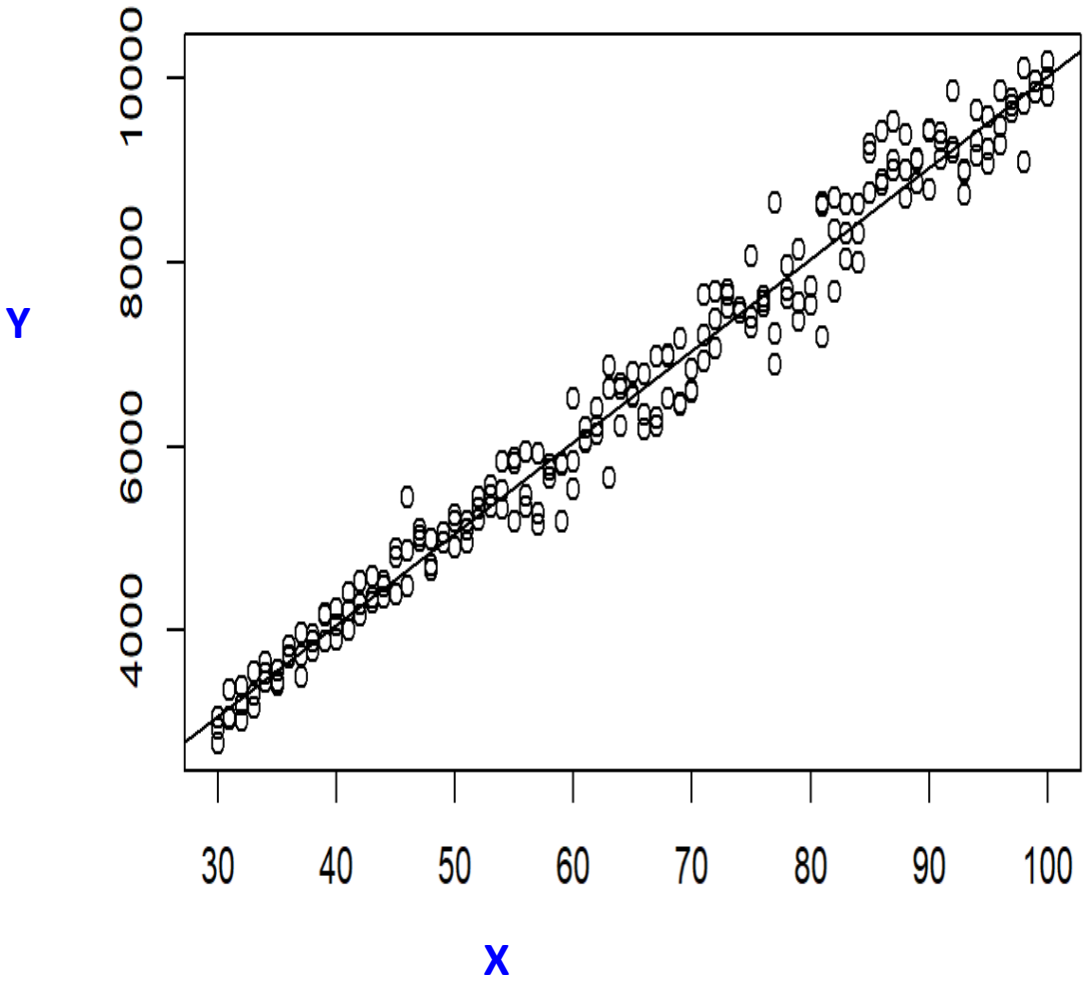
The lack of fit test

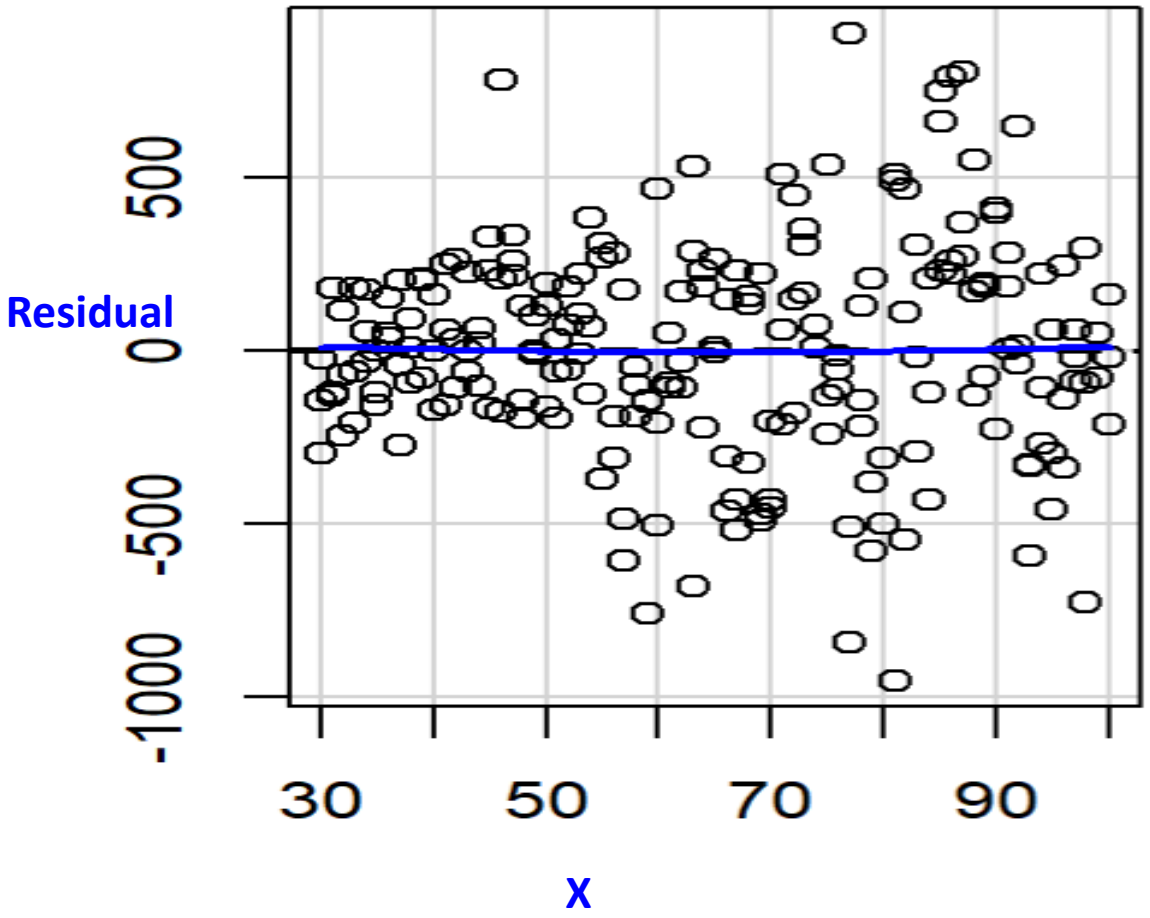
```
Model 1: y ~ x
Model 2: y ~ as.factor(x)
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     88 447641
2     60  35301 28    412340 25.03 < 2.2e-16 ***
```

- $Fs = \frac{MSLF}{MSPE} = \frac{412340/28}{35301/60} = \frac{14726}{588} = 25.03$
- $\sqrt{MSPE} = s\{pure\ error\} = \sqrt{588} = 24.25$, close to 25.
- This is because the Pure error is the variation among Y in each X value, $\sigma\{\mathrm{Y}\} = \sigma\{30 - 10X + X^2 + \varepsilon\} = \sigma\{\varepsilon\} = 25$

A case with non-constant variance (heteroscedasticity)

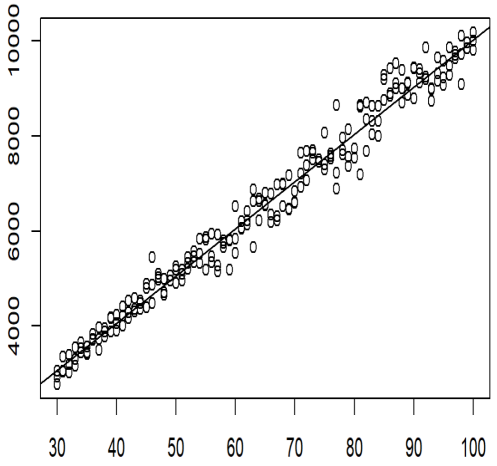
$Y = 30 + 100X + N(0, 5x)$, X ranges from 30 to 100

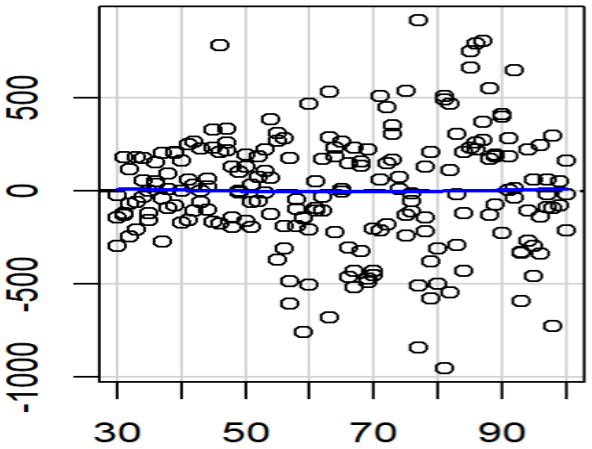




Fit a SLR on the heteroscedasticity data

$Y = 30 + 100X + N(0, 5x)$, X ranges from 30 to 100





- The actual random error deviates from 150 to 500, or 650/2=325 on average.

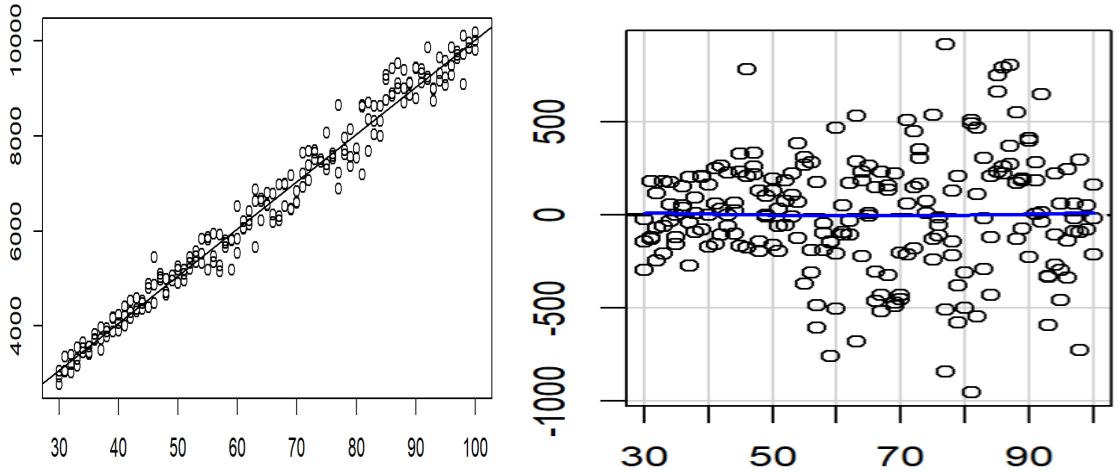
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    77.614     82.189   0.944    0.346
x              98.787      1.206  81.918   <2e-16 ***

Residual standard error: 360.7 on 211 degrees of freedom
Multiple R-squared:  0.9695,    Adjusted R-squared:  0.9694
F-statistic:  6711 on 1 and 211 DF,  p-value: < 2.2e-16
```

```
Response: y
            Df    Sum Sq   Mean Sq F value    Pr(>F)
x            1 873033514 873033514  6710.5 < 2.2e-16 ***
Residuals  211  27450926    130099
```

- The linear impact (slope) is estimated well: 98.8 ± 1.2
- The intercept is not estimated well: 77.6 ± 82.2, with a P-value of 0.346.
- The residual standard error is 360.7, this is a good estimate for the **average** random error standard deviation. But not a good estimate for the actual random error standard deviation which has a changing value.

The lack of fit test on the heteroscedasticity data

$Y = 30 + 100X + N(0, 5x)$, X ranges from 30 to 100



- The actual random error deviates from 150 to 500, or 650/2=325 on average.

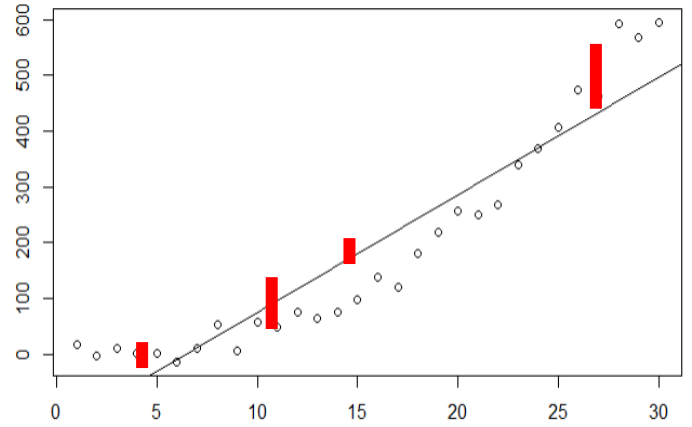
```
Model 1: y ~ x
Model 2: y ~ as.factor(x)
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1    211 27450926
2    142 19030337 69   8420590 0.9106 0.6641
```

- $\sqrt{MSPE} = s\{pure\ error\} = \sqrt{19030337/142} = 366.08$
- Similar as the SLR model, this is a good estimate for the **average** random error standard deviation.

Compare the consequence with the non-linear and the non-constant variance

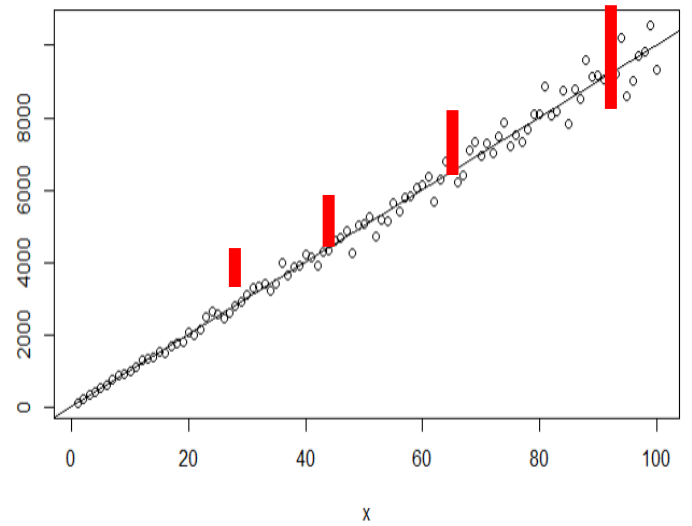
Systematic deviations from the functional form of the model and non-constant variance are both examples of *model misspecification* , and are both magnified in residual plots, they both cause prediction problems. Specifically,

- Systematic deviation from the linear form
  1. Different systematic biases at different values of *X* (i.e., $b_0 + b_1X$ is no good).
  2. A higher overall estimate of error variance (the model is inefficient).



- Non-constant variance
  1. Does **not** cause bias in the point estimates. (i.e., $b_0 + b_1X$ could still be good)
  2. **But it does** invalidate estimates for the standard errors of the parameters.

  For example, $s\{b_0\}$ is large.



Statistical test for heteroscedasticity

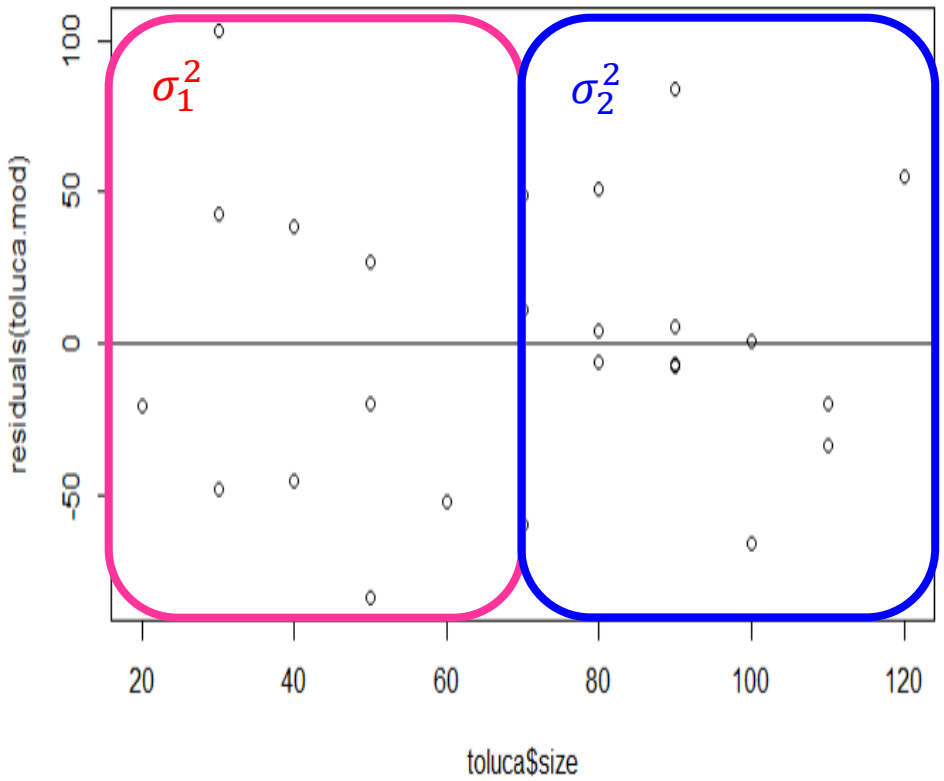
$Ho: residuals\ have\ constant\ variance \quad Ha: rediuals\ have\ non-constant\ variances$

1. Brown Forsythe (BF) test
   - Does not depend on normality of the error terms.
   - BF test is usually used for case with categorical predictors (X).
   - We need to adjust the continuous prediction variable X in the LR **into two or n groups** or categories.
   - Convenient for SLR, but not for MLR.
2. Breusch-Pagan (BP) test
   - Assumes that the error terms are independent and normal and that the variance of the error terms is related to X.
   - No need to group X so more convenient for MLR.

Brown-Forsythe test (2 groups)

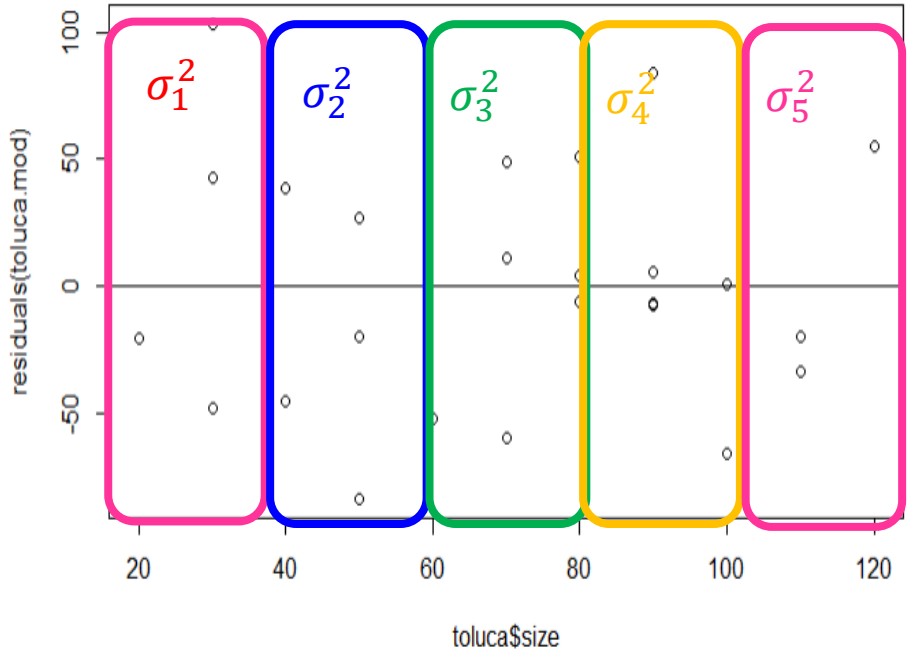
```
library(ALSM)
g<-rep(1,25)
g[toluca$size<=70]=0      #form two groups
bftest(lm(hour~size, toluca),g)
```

```
      t.value   P.value alpha df
[1,] 1.316482 0.2009812  0.05 23
```



Brown-Forsythe test (n groups)

```
library(onewaytests)
toluca$group<-cut(toluca$size, 5)  #form five groups
toluca$residual<-toluca.mod$residuals
bf.test(residual~group, toluca )
```




```
  Brown-Forsythe Test
-----------------------------------------------------
  data : residual and group

  statistic   : 0.5567856
  num df      : 4
  denom df    : 16.50945
  p.value     : 0.6970538

  Result      : Difference is not statistically significant.
-----------------------------------------------------
```

Breusch-Pagan Test (BP test)

```
library(lmtest)
bptest(lm(hour~size, toluca))
```

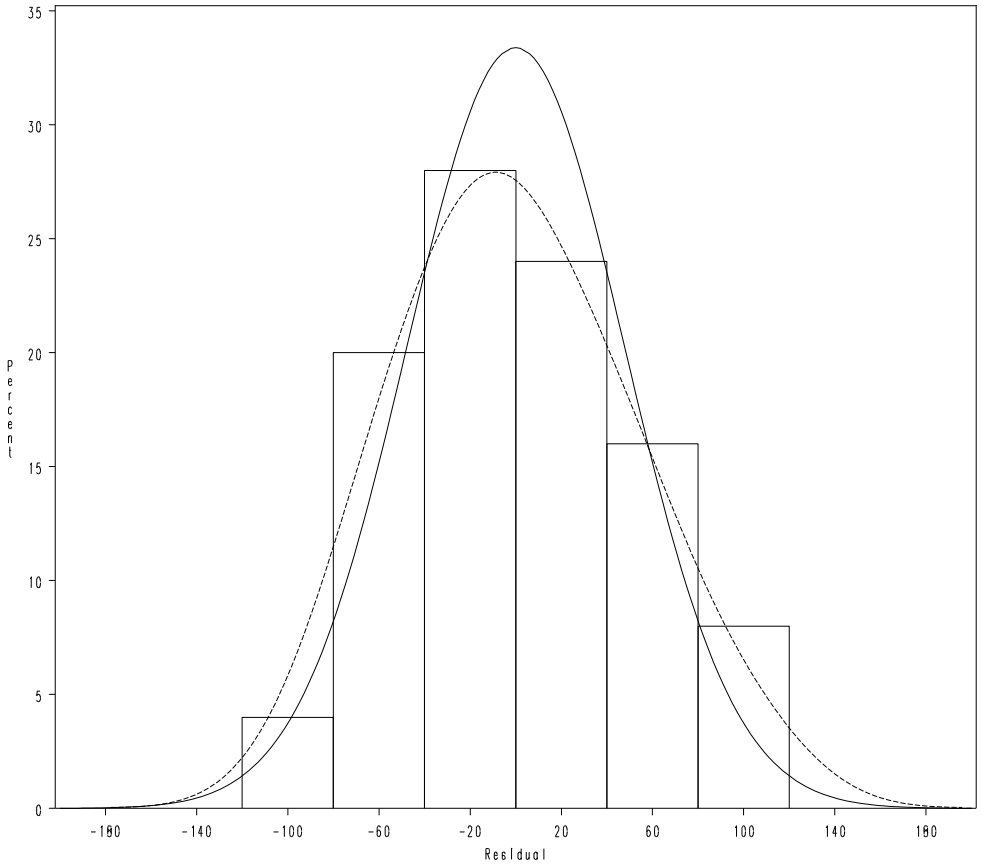
```
        studentized Breusch-Pagan test

data:  lm(hour ~ size, toluca)
BP = 1.1326, df = 1, p-value = 0.2872
```

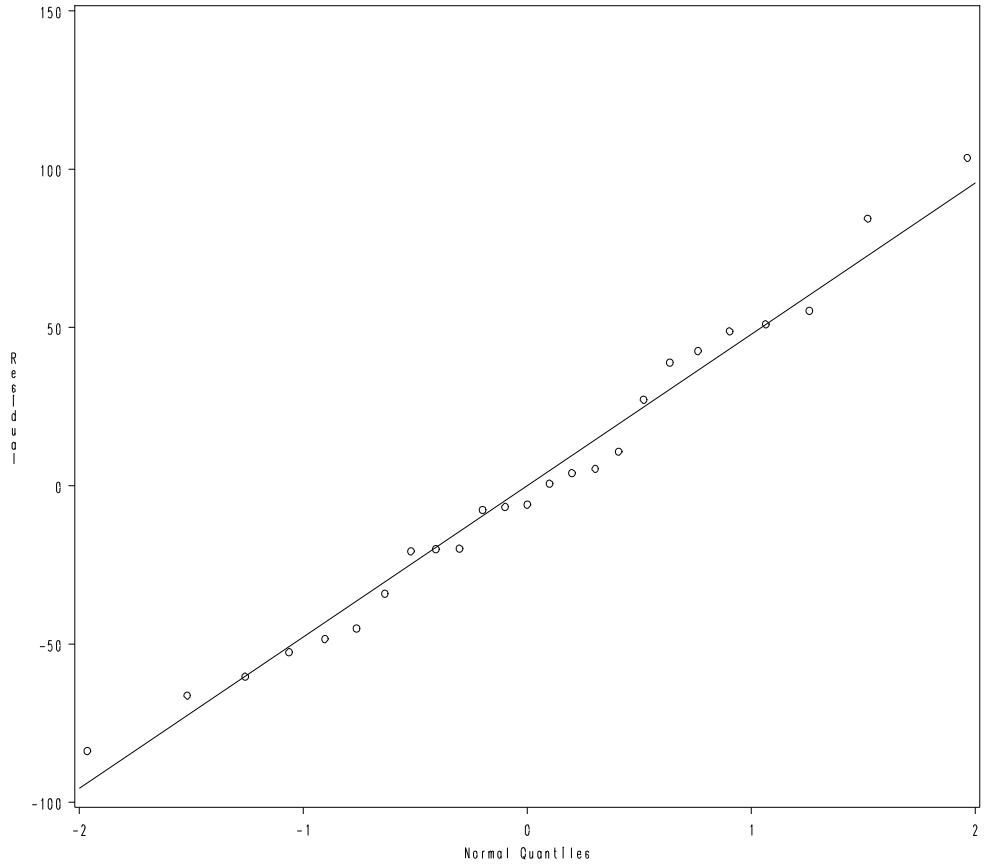
Normal assumption and diagnostics (check residuals)

“Are the errors normal?" is not actually a very helpful question. More useful questions are:

- How far is the distribution of the errors from normal?
- In what way is it non-normal? Is it heavy-tailed? light-tailed? skewed? discrete?
- How will the non-normality affect our inferences? Is it bad enough to invalidate our confidence intervals and hypothesis tests?

Normal quantile plot appears to be a straight line



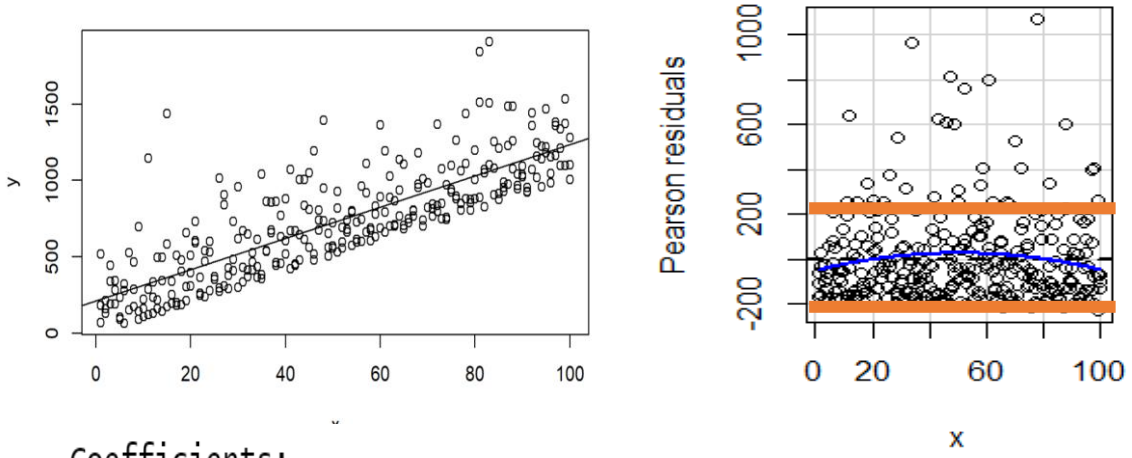


The Toluca example looks pretty good.

A case with non-normal errors

$$Y = \beta_0 + \beta_1 X + \varepsilon, \varepsilon \sim \exp\left(\frac{1}{200}\right), \mu\{\varepsilon\} = 200, \sigma\{\varepsilon\} = 200$$

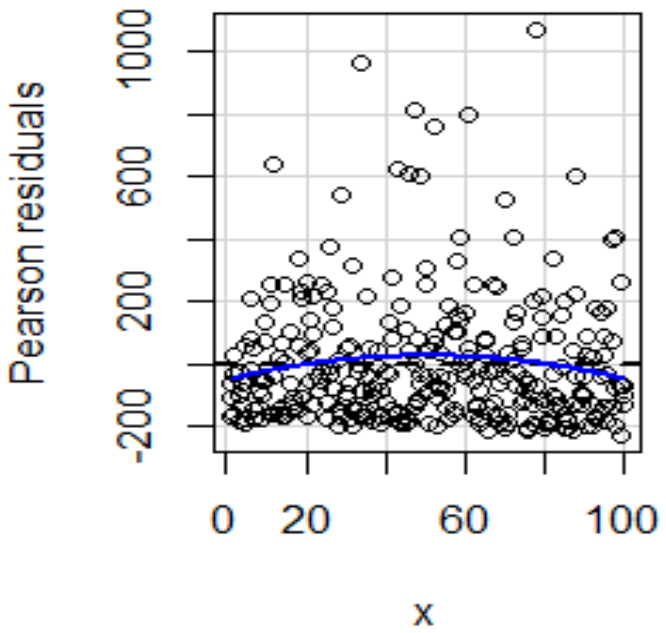
X ranges from 1 to 100, replicate=3




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 212.4916    24.5219   8.665 2.92e-16 ***
x            10.2293     0.4216  24.265  < 2e-16 ***

Residual standard error: 221.8 on 298 degrees of freedom
Multiple R-squared:  0.644,     Adjusted R-squared:  0.6428
F-statistic: 539.2 on 1 and 298 DF,  p-value: < 2.2e-16

Model 1: y ~ x
Model 2: y ~ as.factor(x)
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1    298 14659698
2    200  9815558 98   4844140 1.0072 0.4757
```

- The problem with this case is the outliers.

Fit a SLR to a non-normal data

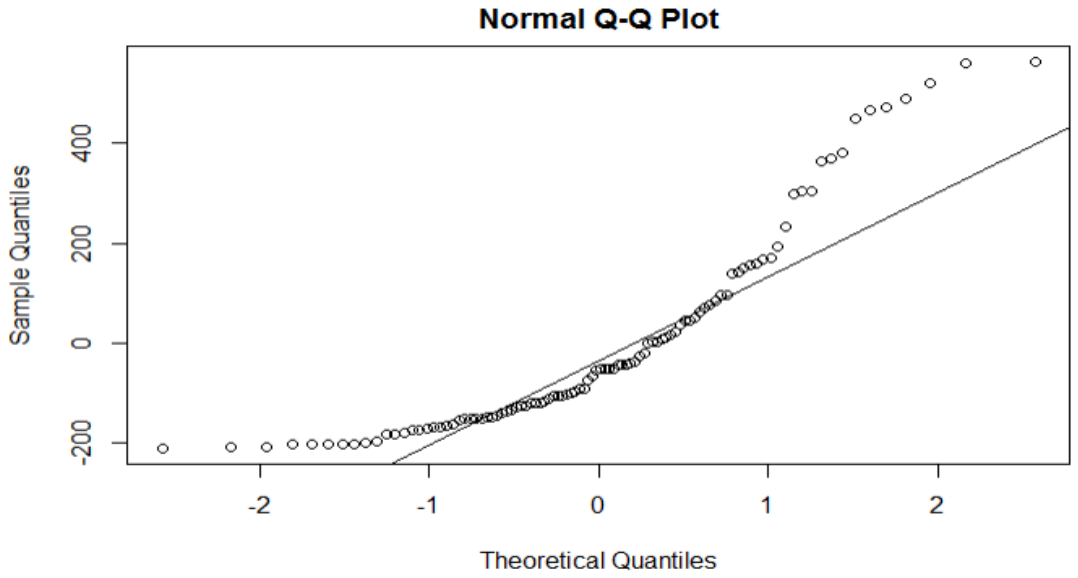



```
Model 1: y ~ x
Model 2: y ~ as.factor(x)
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1    298 14659698
2    200  9815558 98   4844140 1.0072 0.4757
```

Shapiro test for normality

Ho: Data follows normal distribution.
Ha: Data violates from normal distribution.

```
exporesid<-residuals(lm(y~x, data))
shapiro.test(exporesid)
qqnorm(exporesid)
qqline(exporesid)
```




```
        Shapiro-Wilk normality test

data:  exporesid
W = 0.79732, p-value < 2.2e-16
```

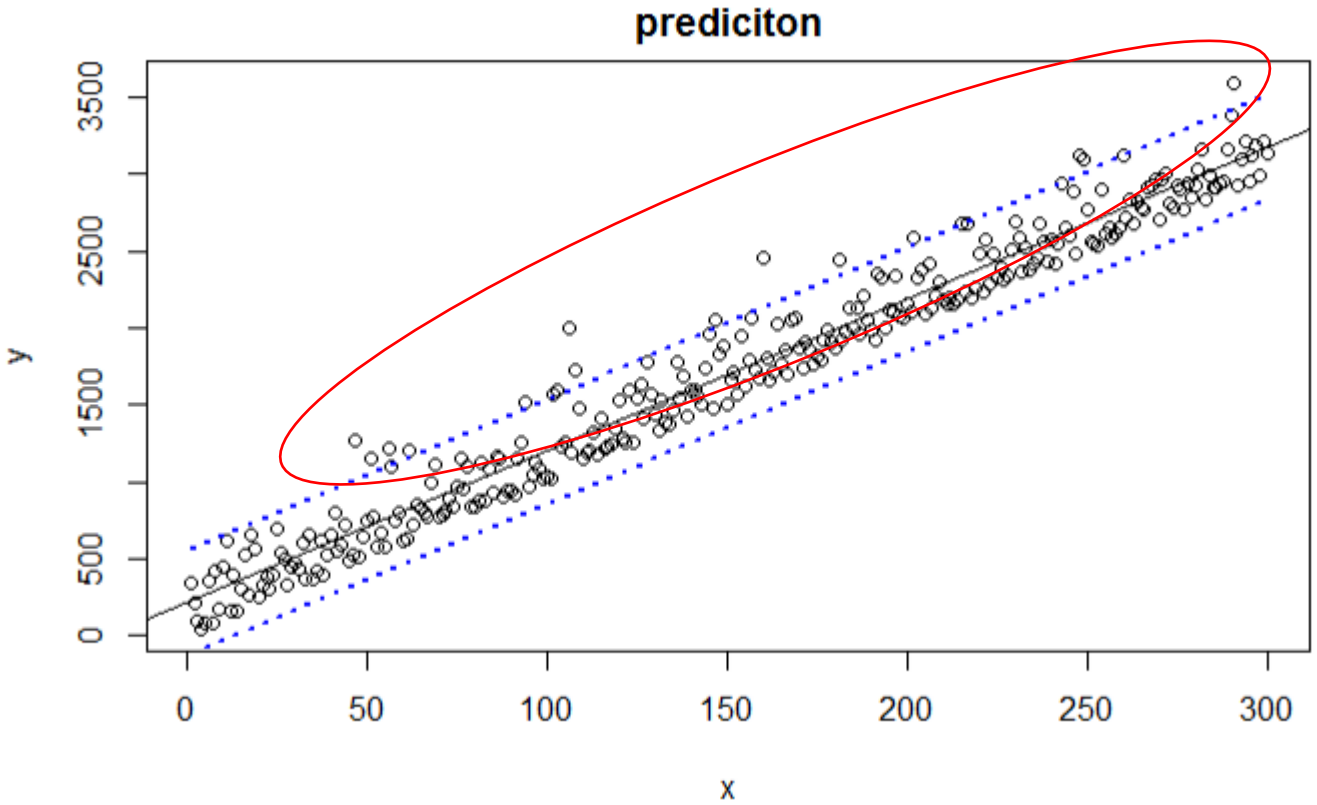
Simulating consequence with non-normal violation (another dataset)

```
library(ALSM)
x<-seq(1:300)
y<-10*x+rexp(300, rate=1/200)
expo<-data.frame(x,y)
expo.mod<-lm(y~x, expo)
cin<-ci.reg(expo.mod, expo$x, type='n',alpha=0.05)
plot(y~x, expo, main="prediciton")
abline(expo.mod)
lines(expo$x, cin$Lower.Band,col="blue", lwd=2, lty=3)
lines(expo$x, cin$Upper.Band, col="blue", lwd=2, lty=3)
```

We create confidence band for predicting a single response variable, $\hat{Y}_h\{new\}$

$$\hat{Y}_h\{new) \pm ts^2\{pred\}$$

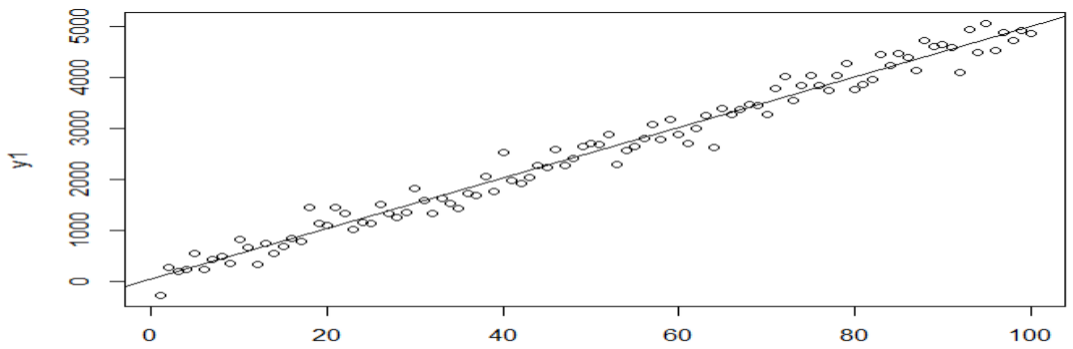
Where $s^2_{\{pred\}} = s^2_{\{\hat{Y}_h\}} + s^2 = s^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2} + 1\right]$



Diagnostic procedure on outliers (the informal procedure)

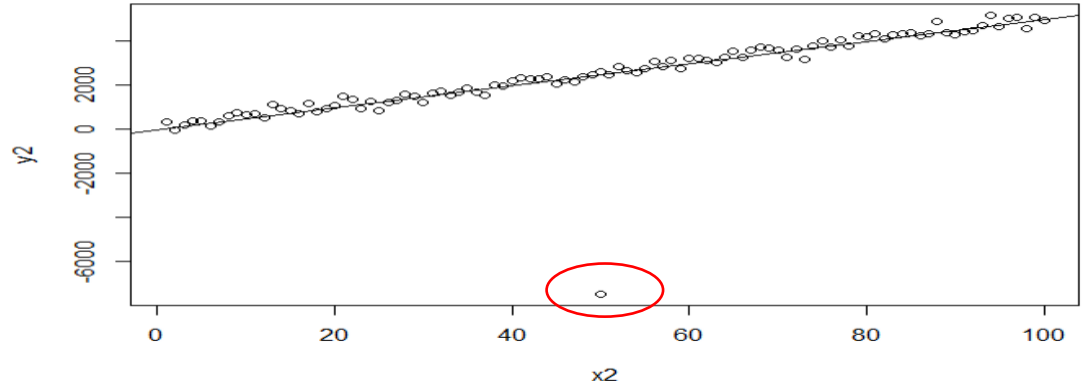
- Outliers are extreme observations.
- Residual outliers can be identified from ***residual plot***, ***boxplot, stem-and-leaf plot*** etc.
- Under the least square (LS) method, a fitted line may be pulled disproportionately toward an outlier.
- Outlier may convey significant information because of an interaction with another predictor variable.
- Outlier that stand *near and far to $\bar{X}$ has different impact*

Impact of outliers: *true model* $Y = 30 + 50X + N(0, 200)$



Without outlier

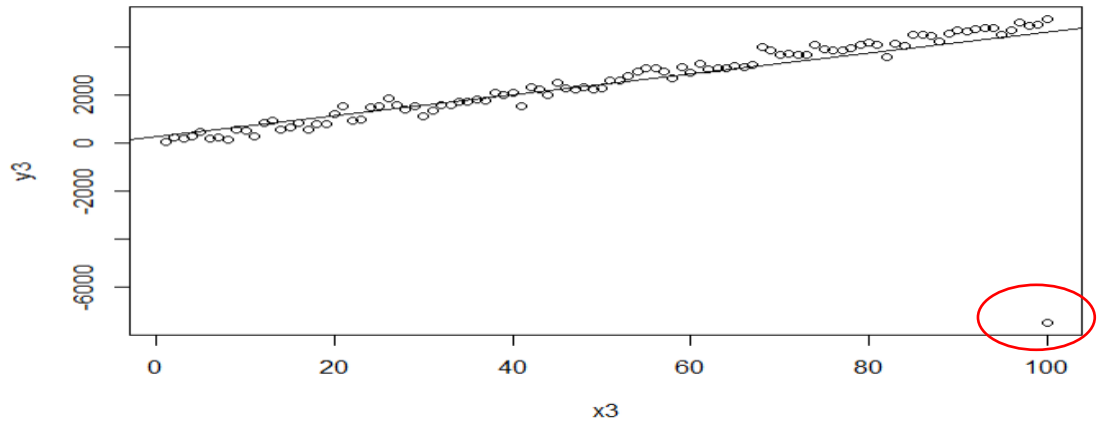
$$\hat{Y} = 42.03 + 49.8X$$
$$R^2 = 0.99, s = 191.8 \ (\sigma = 200)$$



With outlier near $\bar{X}$

$$\hat{Y} = -59.98 + 49.8X$$
$$R^2 = 0.67 \ , s = 1018$$

With outlier near $X\ max$



$$\hat{Y} = 235.1 + 43.5X$$
$$R^2 = 0.51 \ , s = 1251$$

Outlier nears the edge has more impact.

Different kinds of outliers and impact

Outliers near the mean of $\bar{X}$ can *influence* the intercept but lack the *leverage* to strongly affect the slope. However, they still inflate the standard errors for both parameters.

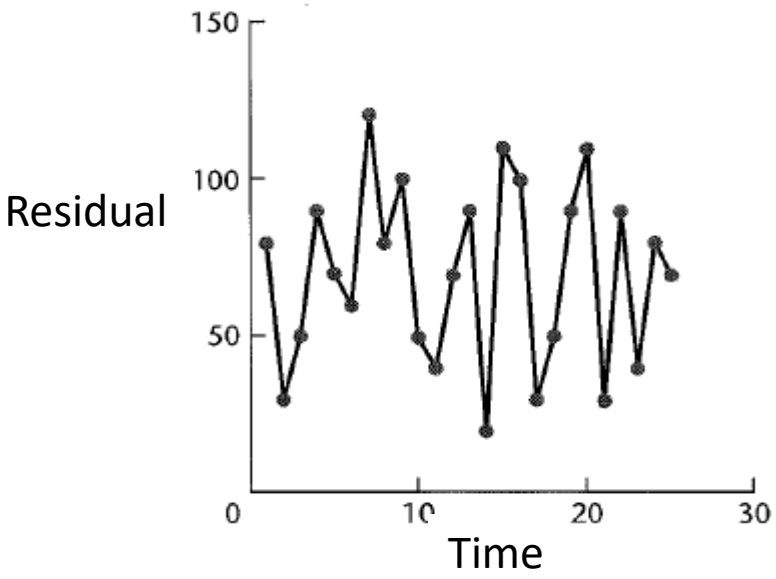
Outliers that are located further from $\bar{X}$ has *greater leverage*, and thus a greater effect on the estimated slope for the same "extremeness". They increase MSE and reduce the precision of estimates.

Diagnostic procedure on dependent Errors

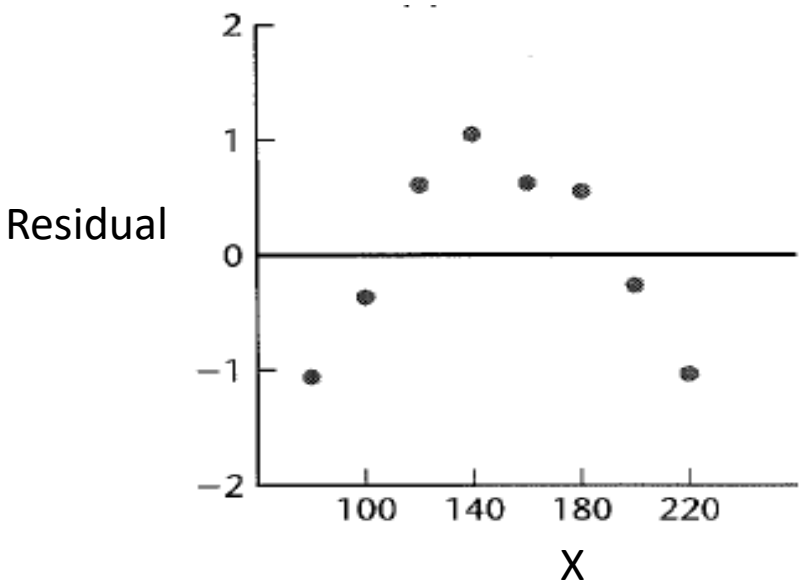
- Ideally, any potential source of dependence is handled at the experimental design stage, so that it is either eliminated by randomization or explicitly included in the data.
- Always watch out for trends or cyclical patterns in the residuals, and plot the residuals **against time, collection order, spatial coordinates, etc.** if you think these might affect the data.
- More subtle dependences can be difficult to detect, especially if the information needed to detect them has not been included with the dataset.

Example

Residual plot A



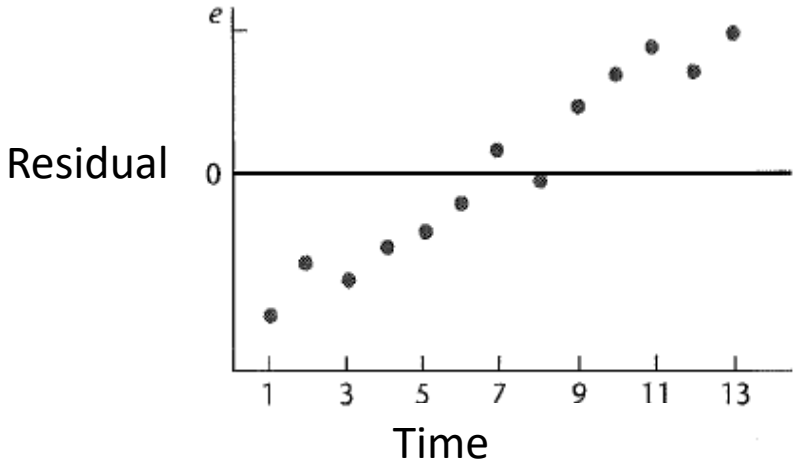
Residual plot B



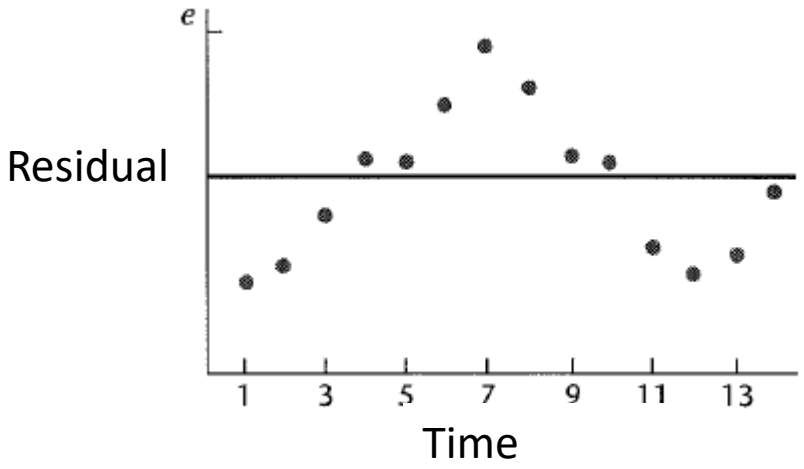
Which of the residual plots provide(s) strong evidence of dependent errors?

--C and D

Residual plot C



Residual plot D



Diagnostic procedure on distribution of predictors

Linear models do not make any assumptions about X ,

but the distribution of X in the data can affect

- the *scope* of the model
- the *accuracy* of inferences for the values of parameters
- the *efficiency* (and accuracy) of inferences for $\hat{Y}_h$ and $Y_{h(new)}$

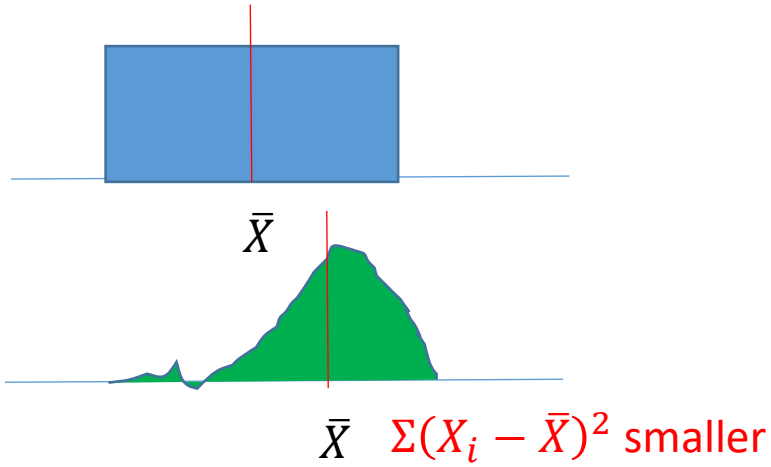
Why does the distribution of X matter?

Even though we make no assumptions about the distribution of X, our estimators depend on its sample mean and variance (technically, on $SS_X = \Sigma(X_i - \bar{X})^2$.

If the range of X in the data is held constant, then relative to a dataset where X is uniformly distributed, a more skewed distribution (or outliers) will:

- pull $\bar{X}$ toward the *body* of the distribution
- cause SS_X to be smaller

As a result . . .



Datasets in which X is skewed will generally yield *less precise estimates for the slope* parameter(s) compared with datasets in which X is more uniformly distributed:

$$s^2\{b_1\} = \frac{MSE}{\Sigma(X_i - \bar{X})^2}$$

If the estimate for β_1 is less precise, then estimates for $\hat{Y}_h$ and $Y_{h(new)}$ will also suffer.

$$s^2_{\{\hat{Y}_h\}} = s^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right] \text{ and } s^2_{\{pred\}} = s^2_{\{\hat{Y}_h\}} + s^2 = s^2$$

In addition, outliers or skew increase the risk that a small number of highly influential data points can dominate the fit and degrade accuracy.

Summary

- We discuss model departures one at a time, although in actuality, several types of departures may occur together. For instance, a linear regression function may be a poor fit and the variance of the error terms may not be constant.

- Graphic analysis of residual analysis is one informal method of analysis, but in many cases, it suffices for examining the aptness of a model.

- The basic approach to residual analysis explained here applies not only simple linear regression but also to more complex regression and other types of statistical models.

Summary

- Several types of departures from the simple linear regression model have been identified by diagnostic tests of the residuals. Model misspecification due to either nonlinearity or the omission of important predictor variables tends to be serious, leading to biased estimates of the regression parameters and error variances.

- Non-constancy of errors variance tends to be less serious, leading to less efficient estimates and invalid error variance estimates.

- The presence of outliers can be serious for smaller data sets when their influence is large.

- Finally, the dependence of error terms results in estimators that are unbiased but whose variances are seriously biased. These problems will be discussed later in depth.