Regression Diagnostics[1]

STA302 Fall 2020

[1] See last slide for copyright information.

Overview

$\widehat{\boldsymbol{\epsilon}}$ estimates $\boldsymbol{\epsilon}$.

- If $\widehat{\boldsymbol{\epsilon}}$ does not act like $\boldsymbol{\epsilon}$ should, investigate.
- Perhaps fix the model or the data.

$\widehat{\boldsymbol{\epsilon}}$ estimates $\boldsymbol{\epsilon}$?

$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$

- First of all, it's a little strange because $\boldsymbol{\epsilon}$ is random.
- But they are analogous.
  - $\widehat{\epsilon}_i$ are vertical distances of the y_i from the estimated regression plane.
  - ϵ_i are vertical distances of the y_i from the true regression plane.
- The vector of residuals is defined as

$$
\begin{aligned}
& & \widehat{\boldsymbol{\epsilon}} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\
& \Rightarrow & \mathbf{y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\epsilon}} \\
\text{Compare} & & \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}
\end{aligned}
$$

- Is it a good estimate?

Distribution: $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, $\widehat{\boldsymbol{\epsilon}} \sim N_n(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$

- Both are multivariate normal with expected value zero.
- $\widehat{\epsilon}_i$ do not have equal variance.
- $\widehat{\epsilon}_i$ are not independent.
- It's not as bad as it seems, because most of $\mathbf{H}$ goes to zero as $n \rightarrow \infty$.

Is $\widehat{\epsilon}_i$ close to ϵ_i? Look at $\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}$.

- $\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}$ is multivariate normal.
- $E(\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}) = \mathbf{0} - \mathbf{0} = \mathbf{0}$.

$$
\begin{aligned}
cov(\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}) &= cov(\mathbf{y} - \widehat{\mathbf{y}} - \boldsymbol{\epsilon}) \\
&= cov(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} - \widehat{\mathbf{y}} - \boldsymbol{\epsilon}) \\
&= cov(\mathbf{X}\boldsymbol{\beta} - \widehat{\mathbf{y}}) \\
&= cov(-\widehat{\mathbf{y}}) \\
&= cov(-\widehat{\mathbf{y}}, -\widehat{\mathbf{y}}) \\
&= cov(\widehat{\mathbf{y}}) \\
&= \sigma^2\mathbf{H}
\end{aligned}
$$

$\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{H})$

- Denoting $\mathbf{H}$ by $[h_{ij}]$, $Var(\widehat{\epsilon}_i - \epsilon_i) = \sigma^2 h_{ii}$.
- Diagonal elements h_{ii} of the hat matrix are sometimes called "hat values."
- Most of the hat values are small. Recall

$$\begin{aligned} tr(\mathbf{H}) &= tr\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) \\ &= tr\left(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) \\ &= tr\left(\mathbf{I}_{k+1}\right) \\ &= k+1 \end{aligned}$$

- So $\sum_{i=1}^n h_{ii} = k+1$ even as n increases.
- The average hat value goes to zero.
- For large samples, $Var(\widehat{\epsilon}_i - \epsilon_i) = \sigma^2 h_{ii}$ is very small most of the time, and $\widehat{\epsilon}_i$ is probably close to ϵ_i.

How about Independence?

- $cov(\widehat{\boldsymbol{\epsilon}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$, so the residuals are not independent.
- Let the $n \times 1$ vector $\mathbf{v}_j$ be all zeros except for a one in position j.
- Construct a *selection matrix*: the $n \times 2$ partitioned matrix $\mathbf{S} = (\mathbf{v}_i | \mathbf{v}_j)$.

$$\mathbf{S}'\mathbf{H}\mathbf{S} = \left(\begin{array}{cc} h_{ii} & h_{ij} \\ h_{ij} & h_{jj} \end{array} \right) = \mathbf{M}$$

- $\mathbf{M}$ is non-negative definite because $\mathbf{a}'\mathbf{M}\mathbf{a} = \mathbf{a}'\mathbf{S}'\mathbf{H}\mathbf{S}\mathbf{a} = (\mathbf{S}\mathbf{a})'\mathbf{H}(\mathbf{S}\mathbf{a}) = \mathbf{v}'\mathbf{H}\mathbf{v} \geq 0$.
- So the eigenvalues of $\mathbf{M}$ are ≥ 0.
- $\Longrightarrow |\mathbf{M}| = h_{ii}h_{jj} - h_{ij}^2 \geq 0$.
- $\Longrightarrow h_{ii}h_{jj} \geq h_{ij}^2$.
- $\Longrightarrow |h_{ij}| \leq \sqrt{h_{ii}h_{jj}}$
- And $h_{ij} \to 0$ if either $h_{ii} \to 0$ or $h_{jj} \to 0$.

Conclusion: For large samples,

- $\widehat{\epsilon}_i$ is a good approximation of ϵ_i, as long as h_{ii} is small.
- $\widehat{\epsilon}_i$ and $\widehat{\epsilon}_j$ are almost independent if either h_{ii} is small or h_{jj} is small (or both).
- In this case, the $\widehat{\epsilon}_i$ should behave very much like the ϵ_i if the model is correct.
- This is the basis of residual plots, where $\widehat{\epsilon}_i$ are treated as if they were ϵ_i.

Another good thing about small hat values

Theorem 5.1 on p. 106 of Sen and Srivastava's *Regression Analysis*

If $\lim_{n \to \infty} \max_i h_{ii} = 0$, then the distribution of $\widehat{\boldsymbol{\beta}}$ approaches a multivariate normal $N_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, even if the distribution of the ϵ_i is not normal.

In this case, tests and confidence intervals based on the normal distribution are roughly okay for large samples (details omitted).

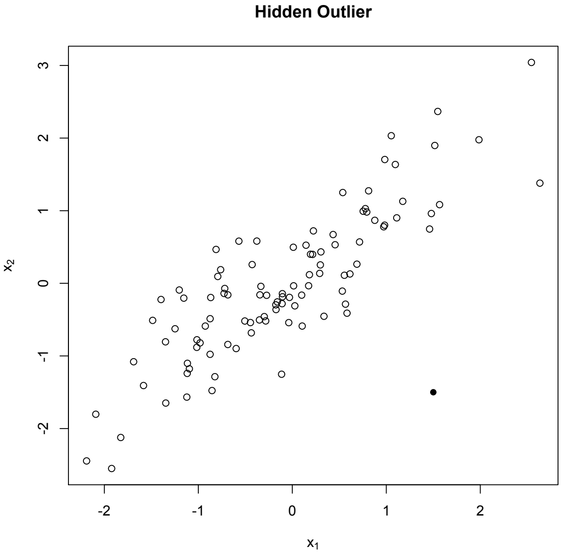
What is a "small" hat value?

- Because $\mathbf{H}$ is non-negative definite, $h_{ii} \geq 0$.
- Because $\mathbf{I} - \mathbf{H}$ is non-negative definite, $1 - h_{ii} \geq 0 \iff h_{ii} \leq 1$.
- So mathematically, $0 \leq h_{ii} \leq 1$.
- Rule of thumb: Worry about $h_{ii} > \frac{2(k+1)}{n}$ (Page 236).
- Another rule of thumb (for multivariate normality of $\widehat{\boldsymbol{\beta}}$) is worry about $h_{ii} > 0.2$.
- Or just look at a histogram of hat values (Page 236).

What causes large h_{ii} values?

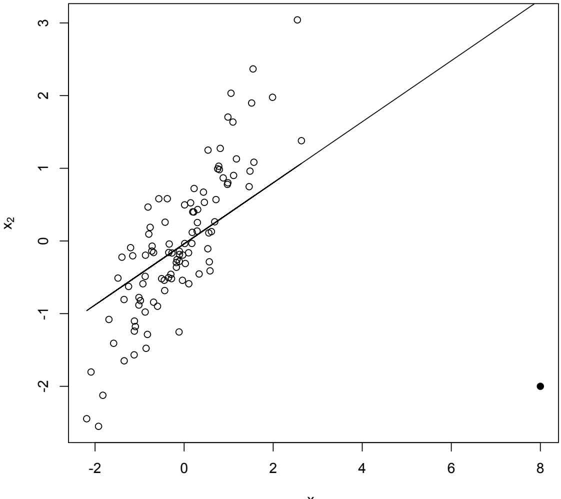
- They correspond to multivariate outliers in the x variables.
- The hat value h_{ii} is an increasing function of the distance from the vector $\mathbf{x}_i'$ and the vector of sample means $(1, \overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \ldots, \overline{\mathbf{x}}_k)'$.
- See Theorem 9.2 (iii) on p. 231.

Multivariate outliers can be hard to spot



Leverage

Hat values h_{ii} are sometimes called “leverage” values



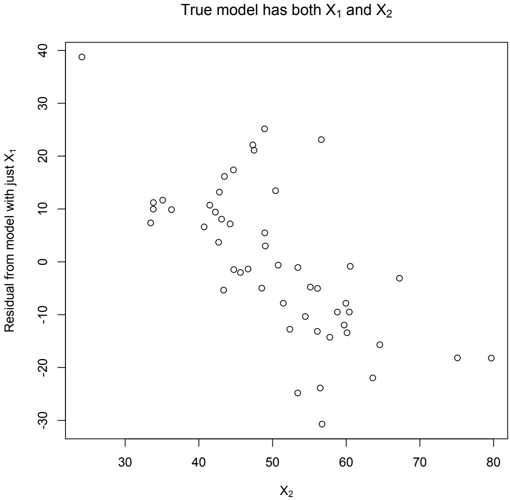
Easy Moral of the Story

- Start by checking for large hat values.
- Look for $h_{ii} > \frac{2(k+1)}{n}$ or $h_{ii} > 0.2$.
- Plots are useful – maybe just a histogram.
- If hat values are big, look at the x values.
- If the hat values are okay, start looking at residuals.

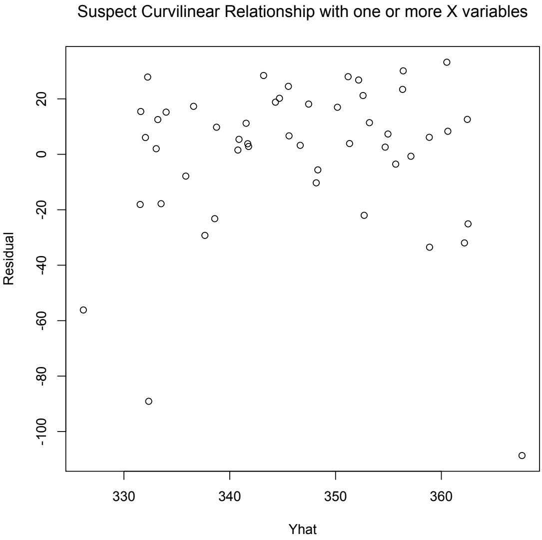
Plotting residuals can be helpful

- Against predicted y.
- Against explanatory variables not in the equation.
- Against explanatory variables in the equation.
- Against time.
- Look for serious departures from normality, outliers.

Plot Residuals Against Explanatory Variables Not in the Equation

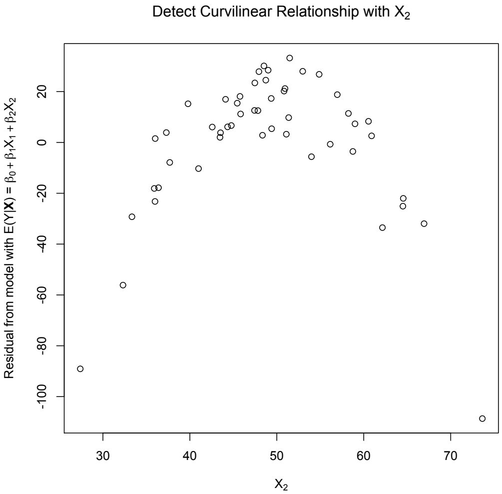


Plot Residuals Against $\widehat{y}$



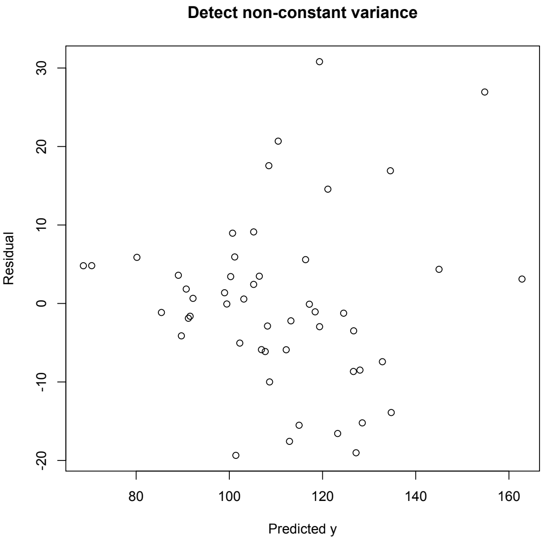
Plot Residuals Against Explanatory Variables in the Equation

Plot versus X_1 showed nothing



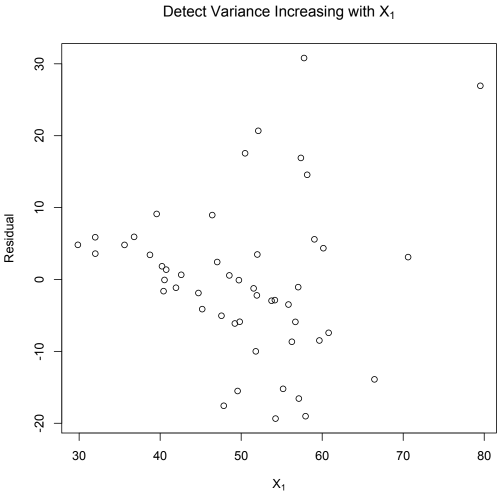
Plot Residuals Against Predicted y

Can show non-constant variance



Plot Residuals Against Explanatory Variables in the Equation

Can show non-constant variance



Faraway has some suggestions

Linear models with R

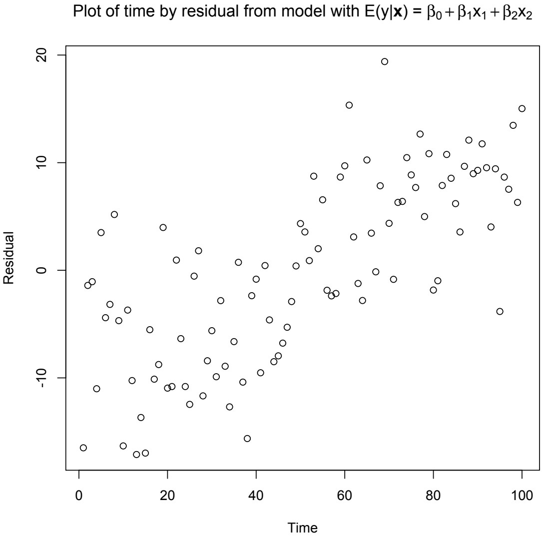
- Plot $\widehat{y}$ by $|\widehat{\epsilon}|$, making change in spread easier to see.
- Do a regression with $x = \widehat{y}$ and $y = |\widehat{\epsilon}|$, and look at the test of $H_0 : \beta_1 = 0$.
- If the model is correct, $\widehat{y}$ and $|\widehat{\epsilon}|$ should be independent.
- The distribution theory behind the test does not quite work, but it can give a rough indication to supplement your inspection of the plots.

Plot Residuals Against Time, if the data are time ordered

- You really need to watch out for time ordered data.
- Regression methods from this course may not be appropriate.
- The problem is that ϵ represents all other variables that are left out of the regression equation.
- Some of them could be time dependent.
- This would make the ϵ_i non-independent, possibly yielding misleading results.

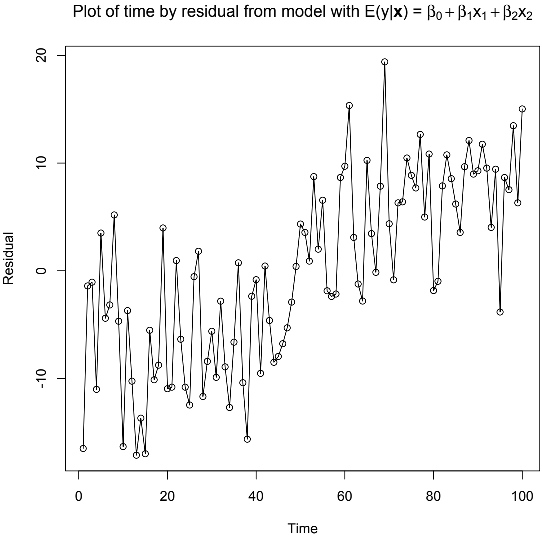
Plot Residuals Against Time

There should be no visible pattern

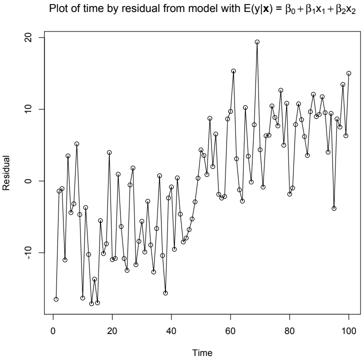


Plot Residuals Against Time

There should be no visible pattern



It's not always so easy



- Looks like an increasing trend. We will include time in the model.
- But it's not always so clear.
- A test would be nice.
- The key is that the higher $\widehat{\epsilon}_t$ is, the higher $\widehat{\epsilon}_{t+1}$ tends to be.
- This is typical of many time series structures, not just trends.

Lagged variables

Assume the observations (cases) are ordered in time

- The value of a variable lag one is the value of the variable one time period ago.
- Like yesterday's high temperature.
- The value of a variable lag six is the value of the variable six time periods ago.
- Regression with lagged x variables (and maybe un-lagged as well) is a natural thing to do.
- If a lagged x variable is related to y, it could be called a "leading indicator."
- Like a leading indicator of number of deaths from covid-19 could be number of covid-19 infections four weeks ago.

Lagged Residuals

Residual	Residual Lag One
$\widehat{\epsilon}_1$	NA
$\widehat{\epsilon}_2$	$\widehat{\epsilon}_1$
$\widehat{\epsilon}_3$	$\widehat{\epsilon}_2$
$\widehat{\epsilon}_4$	$\widehat{\epsilon}_3$
$\vdots$	$\vdots$
$\widehat{\epsilon}_{n-1}$	$\widehat{\epsilon}_{n-2}$
$\widehat{\epsilon}_n$	$\widehat{\epsilon}_{n-1}$

Compute the sample correlation.

Sample autocorrelation

- Correlation between a variable and its lag one is called an *autocorrelation.*
- Specifically, the first order autocorrelation.
- Correlation with lag 2 is the second order autocorrelation, etc.
- The sample autocorrelation is an estimate of the population autocorrelation.
- Of the ϵ_i values, not just the $\widehat{\epsilon}_i$.
- Can reveal lack of independence.
- The most common form is positive autocorrelation.
- The colder is was yesterday, the colder it will probably be today.

The Durbin-Watson Statistic

Assuming the data are in time order

$$d = \frac{\sum_{i=1}^{n}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=2}^{n}\hat{\epsilon}_i^2}$$

If successive $\widehat{\epsilon}_i$ are too close together, d will be small.

More about Durbin-Watson

$$d = \frac{\sum_{i=1}^{n}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=2}^{n}\hat{\epsilon}_i^2}$$

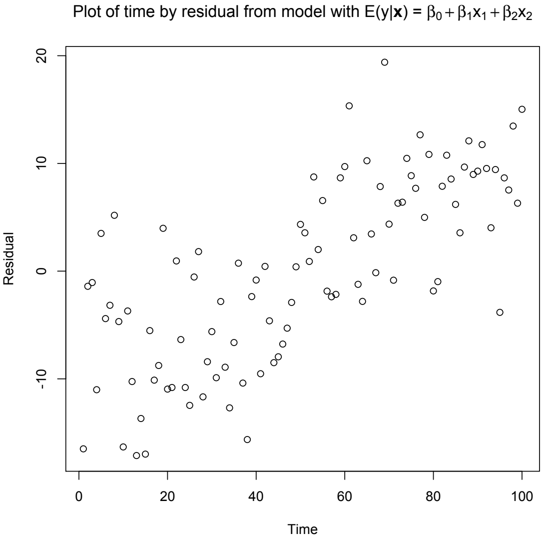
- $d \approx 2(1-\widehat{\rho})$, where $\widehat{\rho}$ is the sample autocorrelation of the residuals.
- $d = 2$ means zero autocorrelation.
- $0 \leq d \leq 4$.
- Small values of d mean positive autocorrelation.
- Rule of thumb is worry if $d < 1$.

Positive autocorrelation

When the ϵ_i values are positively autocorrelated,

- $\widehat{\boldsymbol{\beta}}$ is still unbiased and consistent.
- But MSE underestimate σ^2.
- Confidence intervals and prediction intervals are too narrow.
- Tests are too likely to reject a true null hypothesis.
- The Durbin-Watson test is really useful.

Back to the Example



Original Model

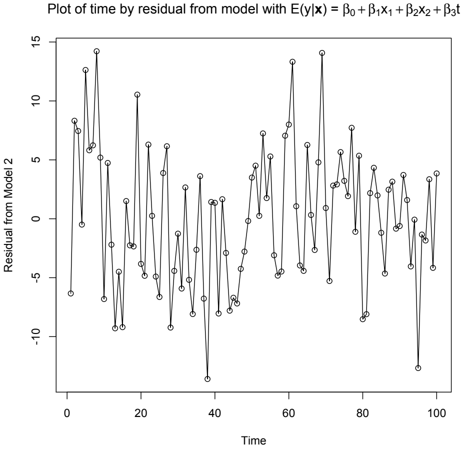
```
> tmod1 = lm(Y~X1+X2)
> # Only need to install package once
> # install.packages("lmtest", dependencies=TRUE)
> # Wow, a lot of stuff.
> library(lmtest)
> # help(dwtest)
> dwtest(tmod1)

Durbin-Watson test

data:  tmod1
DW = 0.74561, p-value = 1.106e-10
alternative hypothesis: true autocorrelation is greater than 0
```

Add Time to the Model

```
> tmod2 = lm(Y~X1+X2+Time);  Residual2 = residuals(tmod2)
> summary(tmod2)

Call:
lm(formula = Y ~ X1 + X2 + Time)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5978  -4.4346   0.0868   3.8548  14.2158

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.89219    3.74819   1.038    0.302
X1           1.04889    0.07201  14.567   <2e-16 ***
X2          -0.99279    0.06939 -14.308   <2e-16 ***
Time         0.21564    0.02065  10.443   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.916 on 96 degrees of freedom
Multiple R-squared:  0.7921,Adjusted R-squared:  0.7856
F-statistic: 121.9 on 3 and 96 DF,  p-value: < 2.2e-16
```

Plot Residuals Against Time

```
> plot(Time,Residual2,ylab = 'Residual from Model 2')
> lines(Time,Residual2)
> tstring = expression(paste('Plot of time by residual from model with
+ E(y|',bold(x),') = ', beta[0]+beta[1]*x[1]+beta[2]*x[2]+beta[3]*t ))
> title(tstring)
```



Durbin-Watson

```
> dwtest(tmod2)

Durbin-Watson test

data:  tmod2
DW = 1.5432, p-value = 0.007792
alternative hypothesis: true autocorrelation is greater than 0
```

Try a Cubic in Time

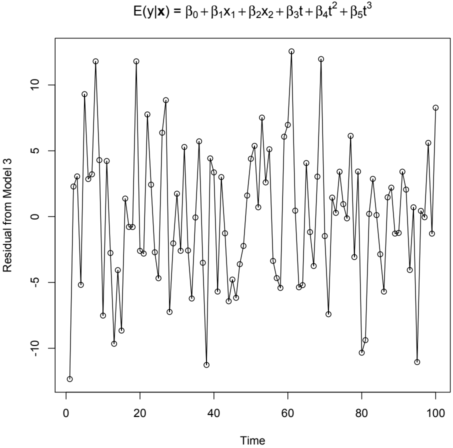
```
> # Okay, maybe a cubic
> Time2 = Time^2; Time3 = Time^3
> tmod3 = lm(Y~X1+X2+Time+Time2+Time3)
> Residual3 = residuals(tmod3)
> anova(tmod2,tmod3)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + Time
Model 2: Y ~ X1 + X2 + Time + Time2 + Time3
  Res.Df    RSS Df Sum of Sq     F   Pr(>F)
1     96 3359.9
2     94 2928.4  2    431.54 6.926 0.001563 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Summary

```
> summary(tmod3)

Call:
lm(formula = Y ~ X1 + X2 + Time + Time2 + Time3)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.3439  -3.6473   0.1622   3.4106  12.5547 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.183e+01  4.225e+00   2.801 0.006192 ** 
X1           1.081e+00  6.857e-02  15.770  < 2e-16 ***
X2          -1.040e+00  6.692e-02 -15.539  < 2e-16 ***
Time        -5.288e-01  2.018e-01  -2.621 0.010238 *  
Time2        1.702e-02  4.614e-03   3.688 0.000378 ***
Time3       -1.064e-04  2.993e-05  -3.556 0.000591 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.582 on 94 degrees of freedom
Multiple R-squared:  0.8188,Adjusted R-squared:  0.8091 
F-statistic: 84.94 on 5 and 94 DF,  p-value: < 2.2e-16
```

Plot Residuals

```
> plot(Time,Residual3,ylab = 'Residual from Model 3')
> lines(Time,Residual3)
> tstring = expression(paste('E(y|',bold(x),
+ ') = ', beta[0]+beta[1]*x[1]+beta[2]*x[2]+beta[3]*t+beta[4]*t^2+beta[5]*t^3))
> title(tstring)
```

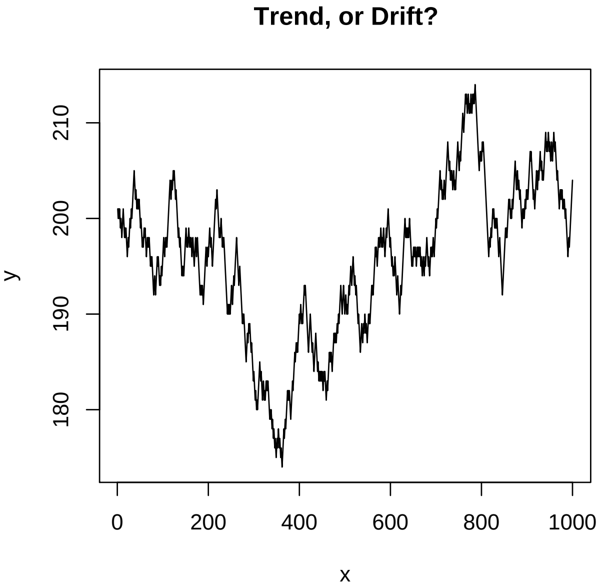


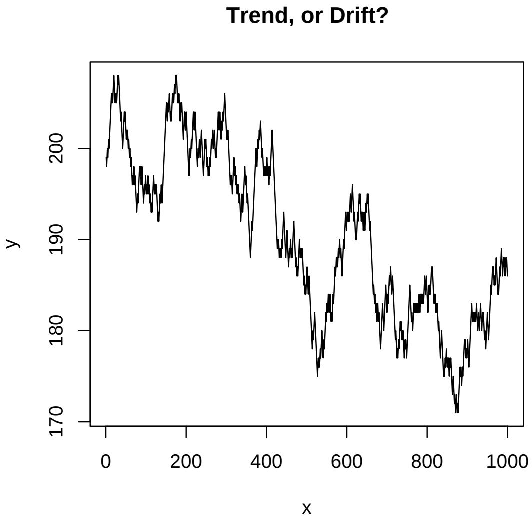
Durbin-Watson Test

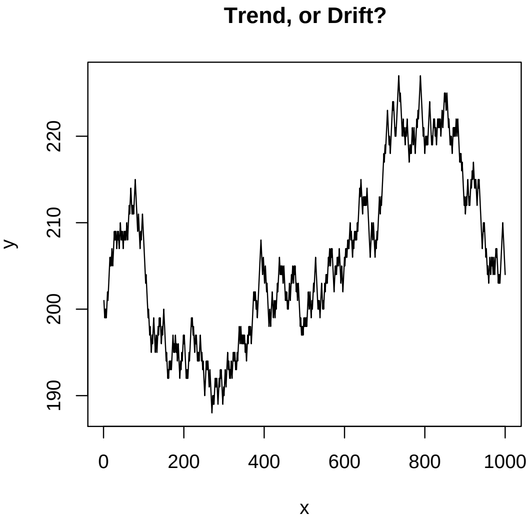
```
dwtest(tmod3)

Durbin-Watson test

data:  tmod3
DW = 1.7636, p-value = 0.06464
alternative hypothesis: true autocorrelation is greater than 0
```

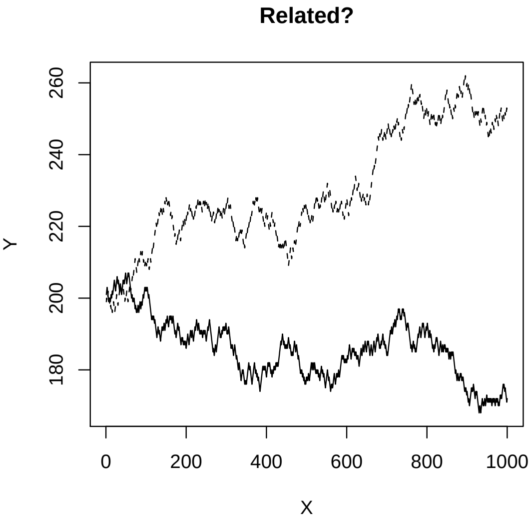
Watch Out

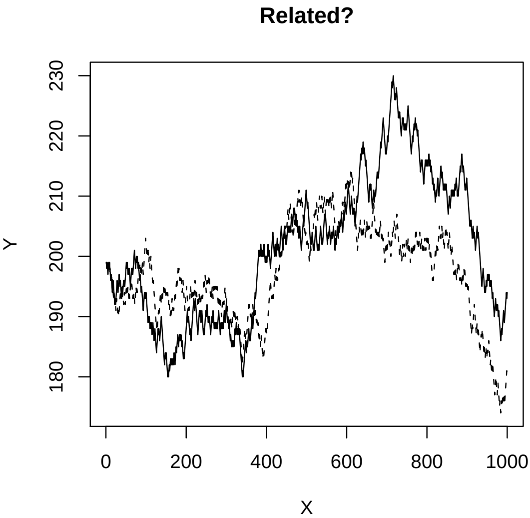
- Including time in the model is primitive, but effective if all you have is a trend.
- You really do have to be careful about using ordinary least squares regression on time series data.
- With positive autocorrelation, each observation tends to be close to the last one, and they can drift.


Trend, or Drift?
210
200
190
180
y
0
200
400
600
800
1000
x


Trend, or Drift?
y
170
180
190
200
x
0
200
400
600
800
1000


Trend, or Drift?
y
220
210
200
190
0
200
400
600
800
1000
x


Related?
Y
260
240
220
200
180
X
0
200
400
600
800
1000


Related?
Y
X
230
220
210
200
190
180
0
200
400
600
800
1000

Random walk

Sometimes called Drunkard's walk

- Take a step left or right at random.
- Steps could be of variable length.
- Location at time t depends on location at time $t-1$.

$$X_t = X_{t-1} + \epsilon_t$$

$\epsilon_1, \epsilon_2, \ldots$ all independent and identically distributed.

Correlations: 50 pairs of independent random walks, $n = 1000$ steps

Need around $|r| = 0.13$ for significance

-0.28175	-0.22242	-0.32170	-0.45053	0.07866	0.59167	-0.27414	-0.82570
-0.62175	0.43537	0.84147	0.04103	-0.17502	-0.89710	-0.19116	-0.53865
-0.50889	0.42855	-0.91074	0.90577	0.22818	0.84834	-0.52501	0.82583
-0.06838	-0.00234	0.16084	0.81393	-0.07063	-0.09908	-0.38405	-0.28510
0.24850	0.12445	0.33509	0.33586	0.41241	-0.33482	-0.32021	-0.73808
0.14045	-0.03618	-0.67757	0.81121	-0.39379	-0.58832	-0.26866	0.16687
0.38541	0.12433						

If you do ordinary regression on time series data

- Plot the residuals against time.
- Look at the Durbin-Watson test.
- Try to include relevant time-varying predictor variables.
- Learn about genuine time series methods (STA457).
- If you study time series, don't stick your nose up at univariate time series methods. Apply them to the residuals!

Outlier detection

- Big residuals may be outliers. What's "big?"
- Consider standardizing.
- But note that variances of $\widehat{\epsilon}_i$ are not all the same.
- Semi-Studentized: Estimate $Var(\widehat{\epsilon}_i)$ and divide by square root of that: $\frac{\widehat{\epsilon}_i}{\sqrt{MSE\,(1-h_{i,i})}}$
- In R, this is produced with `rstandard`.

Studentized deleted residuals

The idea

- An outlier will make MSE big.
- In that case, the standardized (Semi-Studentized) residual $\frac{\widehat{\epsilon}_i}{\sqrt{MSE\,(1-h_{i,i})}}$ will be too small – less noticeable.
- So calculate $\widehat{y}$ for each observation based on all the other observations, but not that one. Leave one out.
- Predict each observed y based on all the others, and assess error of prediction (divided by standard error).
- Big values suggest that the expected value of y_i is not what it should be.
- Maybe that observation is from a different domain – investigate.

Apply prediction interval technology

$$t = \frac{y_0 - \mathbf{x}_0' \widehat{\boldsymbol{\beta}}}{\sqrt{MSE(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}} \sim t(n-k-1)$$

- Note that y_i is now being called y_0.
- If the "prediction" is too far off there is trouble.
- Use t as a test statistic.
- Need to change the notation.

Studentized deleted residual

$$t_i = \frac{y_i - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}_{(i)}}{\sqrt{MSE_{(i)}(1 + \mathbf{x}_i'(\mathbf{X}_{(i)}' \mathbf{X}_{(i)})^{-1}\mathbf{x}_i)}} \sim t(n-k-2)$$

- In R, this is produced with `rstudent`.
- There is a more efficient formula.
- Use t_i as a test statistic of $H_0 : E(y_i) = \mathbf{x}_i'\boldsymbol{\beta}$.
- If H_0 is rejected, investigate.
- We are doing n tests.
- If all null hypotheses are true (no outliers), there is still a good chance of rejection at least one H_0.
- Type I errors are very time consuming and disturbing.
- How about a Bonferroni correction?

Bonferroni Correction for Multiple Tests

- Do the tests as usual, obtaining n test statistics.
- For each test, use the significance level α/n instead of α.
- Use the critical value $t_{\frac{\alpha}{2n}}(n-k-2)$.
- Even for large n it is not overly conservative.
- If you locate an outlier, **investigate**!

Normality

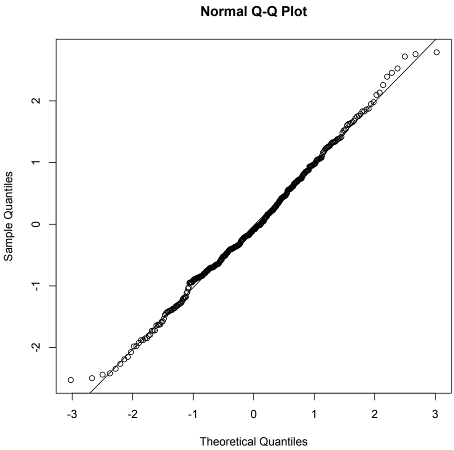
- Instead of checking the residuals for normality, I like to check the Studentized deleted residuals (`rstudent`).
- Their variances are all equal.
- And for a healthy sample size, t is almost z.
- Start with `hist()`.

QQ Plots

- Plot ordered values of a variable against the expected values of the order statistics under normality.
- If the distribution is normal, the plot should be approximately straight line.

qqnorm

```
> help(qqnorm)
> x1 = rnorm(400)
> qqnorm(x1); qqline(x1)
```

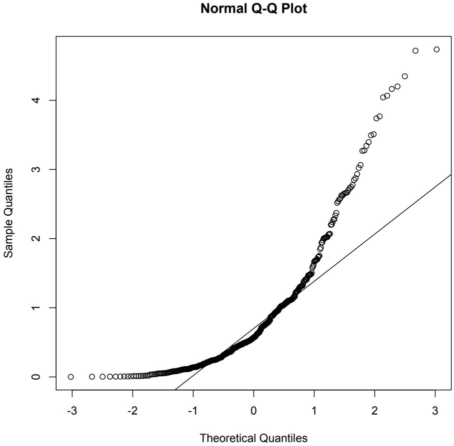


The Shapiro-Wilk Test for Normality

```
> help(shapiro.test)
> shapiro.test(x1)
Shapiro-Wilk normality test

data:  x1
W = 0.99574, p-value = 0.3527
```

Non-normal data

```
> x2 = rexp(400)
> qqnorm(x2); qqline(x2)
```



Test for Normality

x2 is exponential

```
> shapiro.test(x2)
Shapiro-Wilk normality test

data:  x2
W = 0.81528, p-value < 2.2e-16
```

Influential Observations

- Based on the idea of leverage, look for large hat values h_{ii}.
- If $h_{ii} > 0.2$ or $h_{ii} > \frac{2(k+1)}{n}$, investigate.
- Other methods are based on leave-one-out technology.

Leave One Out

- DFBETA $= \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\widehat{\epsilon}_i}{1-h_{ii}}$.
- DFBETAS: Use t_i instead of $\widehat{\epsilon}_i$.
- DFFIT $= \widehat{y}_i - \widehat{y}_{(i)} = \frac{h_{ii}\widehat{\epsilon}_i}{1-h_{ii}}$.
- DFFITS: Use t_i instead of $\widehat{\epsilon}_i$.
- Cook's distance: $D_i = \frac{\sum_{i=1}^n(\widehat{y}_i - \widehat{y}_{(i)})^2}{MSE(k+1)} = \left(\frac{1}{k+1}\right) t_i^2 \left(\frac{h_{ii}}{1-h_{ii}}\right)$.
- They say worry about $D_i > 1$.

- If any of these measures is a lot bigger than the others, investigate.

Copyright Information