Stat 306:
Finding Relationships in Data.

Lecture 17

Sections 4.4 and 4.5

BONUS FOR LEAVE-ONE-OUT Cross Validation

The cross-validated root mean square (prediction) error is:

$$CVRMSE_{\text{leaveoneout}}(x_1, \ldots, x_p) = \sqrt{n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_{i|-i})^2}. \tag{4.26}$$

$CVRMSE_{\text{leaveoneout}}(\mathbf{x}_J)$ can be also defined for the subset of the explanatory variables indexed by $J \subset \{1, \ldots, p\}$.

It turns out there is a simple formula for cross-validation leave-one-out residuals (without having to compute n different regressions). An identity is

$$y_i - \hat{y}_{i|-i} = (y_i - \hat{y}_i)/(1 - P_{ii}), \quad P_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i, \tag{4.27}$$

where $\mathbf{x}_i^T = (1, x_{i1}, \ldots, x_{ip})$. Alternatively, P_{ii} is the ith diagonal element of the projection matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. See Section 4.3 for why this quantity is called the projection matrix.

Chapter 4 – Variable selection and additional diagnostics

4.1 Variable Selection algorithms

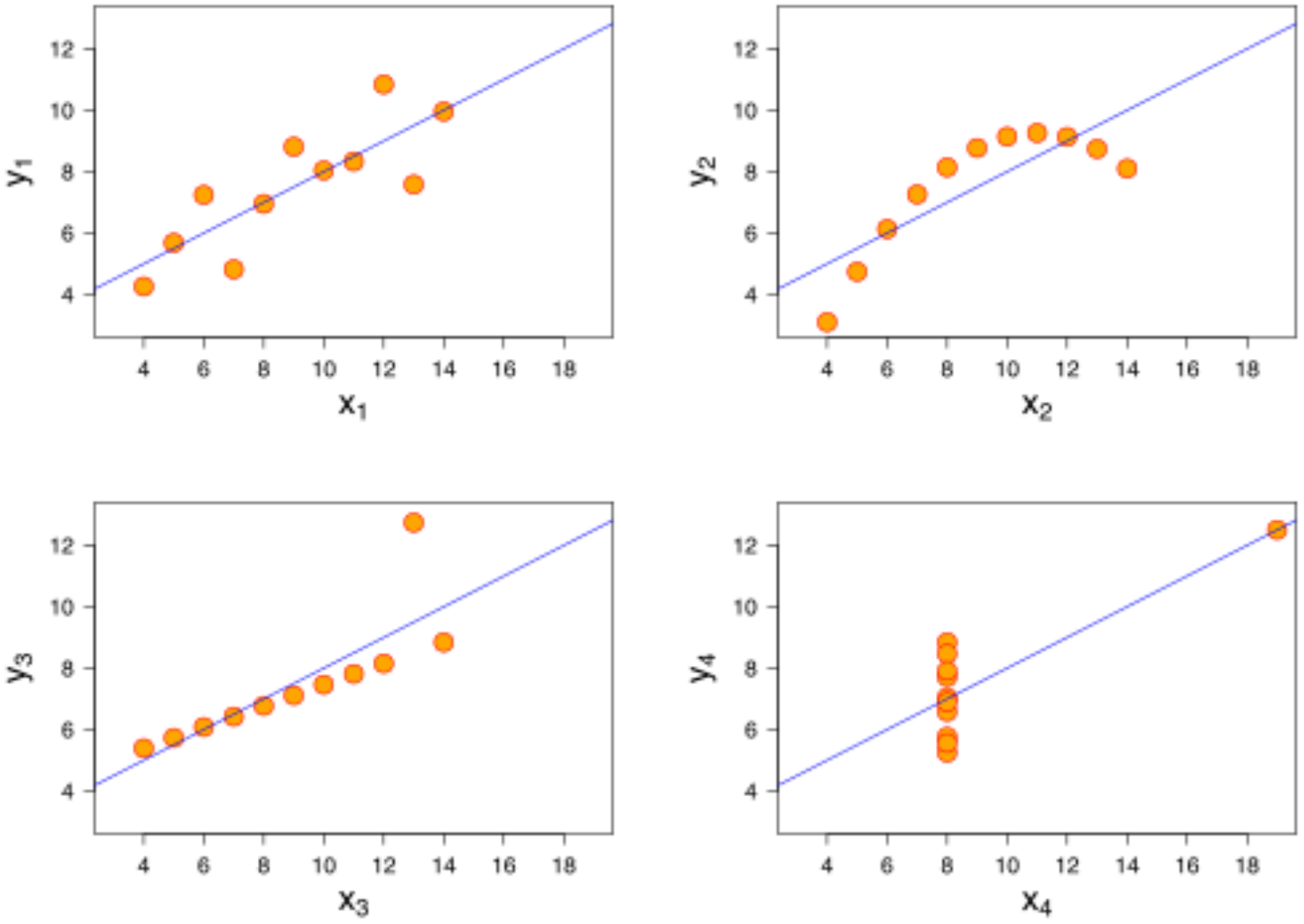
4.2 Cross-validation and out-of sample assessment

4.3 Additional diagnostics

4.4 Transforms and nonlinearity

4.5 Diagnostics for data collected sequentially in time

Classic example: Anscombe's quartet



- https://www.refsmmat.com/regression/regression.html

Chapter 4 – Variable selection and additional diagnostics

4.1 Variable Selection algorithms

4.2 Cross-validation and out-of sample assessment

4.3 Additional diagnostics

4.4 Transforms and nonlinearity

4.5 Diagnostics for data collected sequentially in time

4.4 Transforms and nonlinearity

In this section, we discuss when transforms of variables might be needed and how to handle diagnostics that suggest that g in (3.33) is nonlinear in the explanatory variables. The general prediction model is:

$$Y_i = g(x_{i1}, \ldots, x_{ip}) + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \sim N(0, \sigma^2(\mathbf{x}_i)) \text{ independently}, \tag{4.32}$$

where now we indicate heteroscedasticity with $\sigma^2(\mathbf{x})$ being a function of the explanatory variables.

Sometimes after a transform of y or one or more of the explanatory variables might lead to the homoscedastic model with g linear being a better approximation to the data.

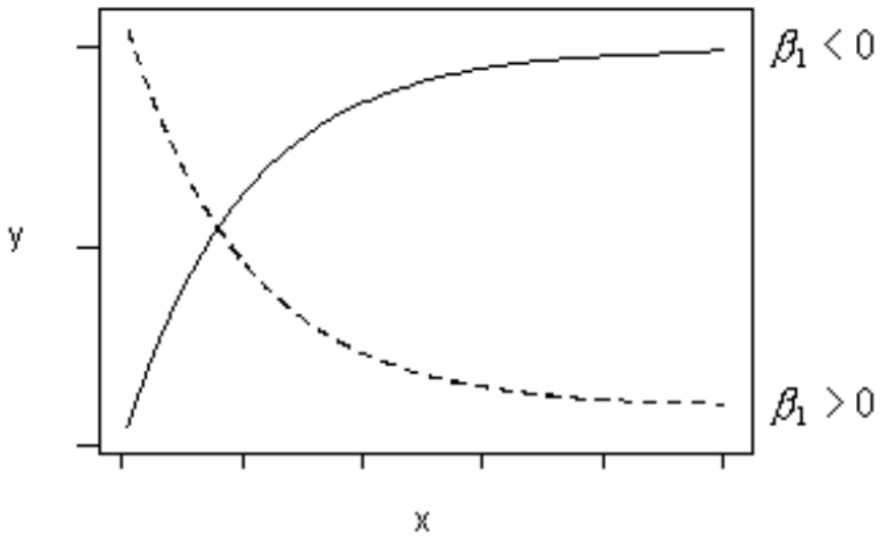
In particular, if diagnostics suggest heteroscedasticity with more variability and y increases, then a log or square root or power transform of y can be useful (assuming y is positive-valued). Examples of this are shown in Exercise 2.7 and in the case study in Chapter 5.

4.4 Transforms and nonlinearity

If the trend in your data follows either of these patterns, you could try fitting this regression function:

$$\mu_Y = \beta_0 + \beta_1 e^{-x}$$

to your data.



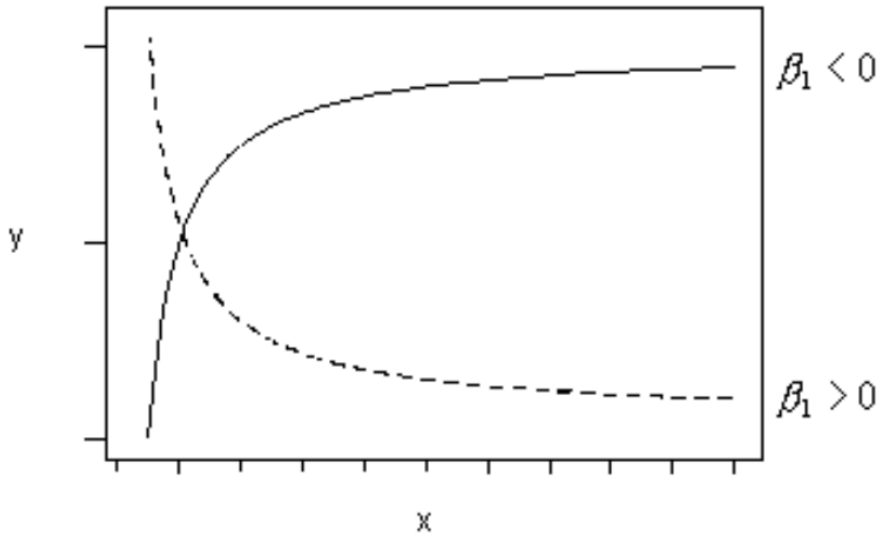
source: https://onlinecourses.science.psu.edu/stat501/node/322

4.4 Transforms and nonlinearity

Or, if the trend in your data follows either of these patterns, you could try fitting this regression function:

$$\mu_Y = \beta_0 + \beta_1 \left(\frac{1}{x}\right)$$

to your data. (This is sometimes called a "reciprocal" transformation.)



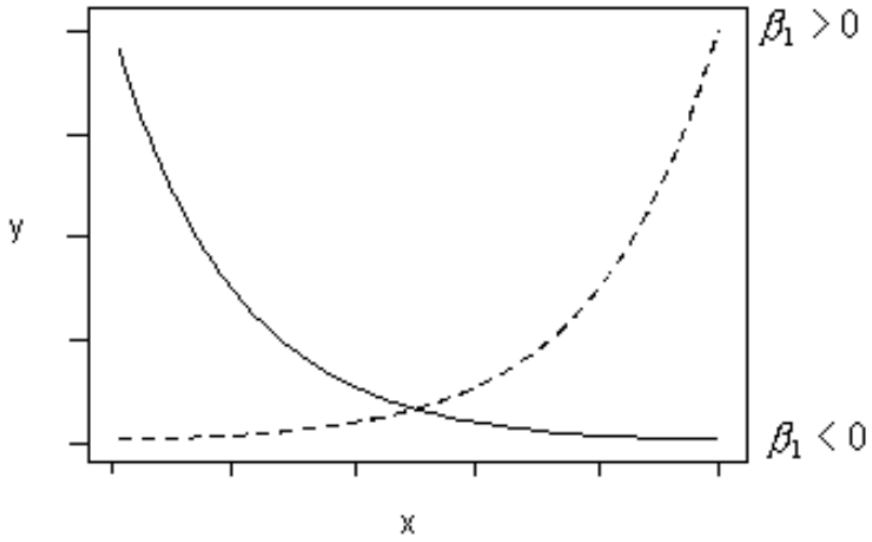
source: https://onlinecourses.science.psu.edu/stat501/node/322

4.4 Transforms and nonlinearity

Or, if the trend in your data follows either of these patterns, try fitting this regression function:

$$\mu_{lnY} = \beta_0 + \beta_1 x$$

to your data. That is, fit the model with $ln(y)$ as the response and x as the predictor.



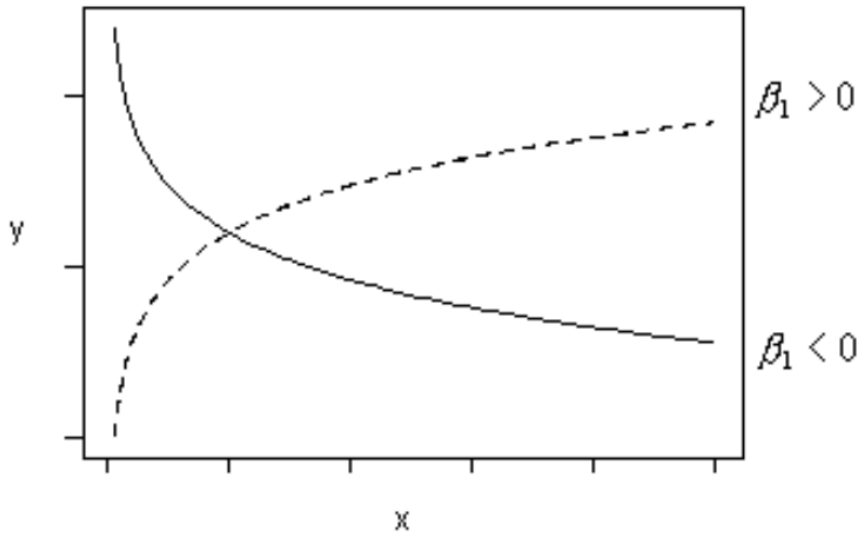
source: https://onlinecourses.science.psu.edu/stat501/node/322

4.4 Transforms and nonlinearity

Or, try fitting this regression function:

$\mu_Y = \beta_0 + \beta_1 ln(x)$

if the trend in your data follows either of these patterns. That is, fit the model with y as the response and $ln(x)$ as the predictor.



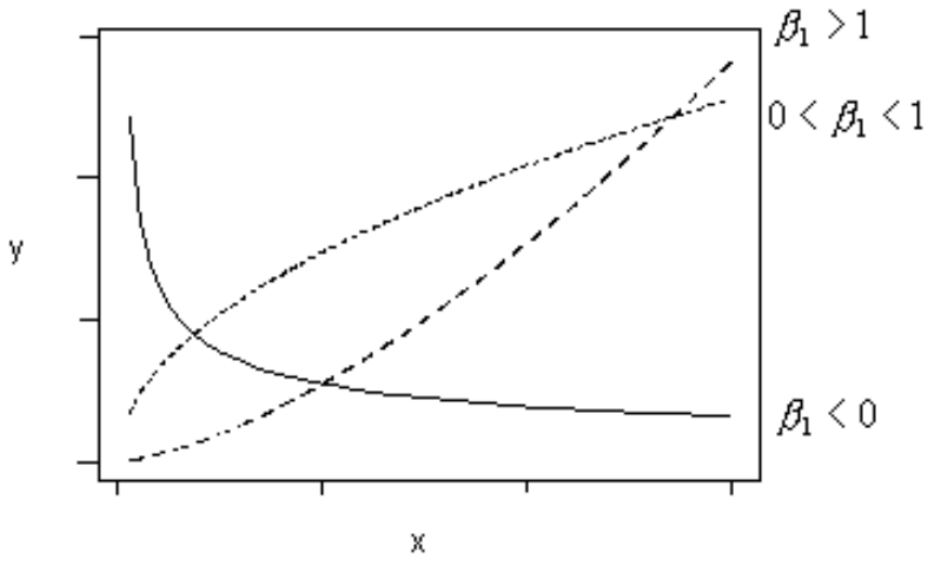
source: https://onlinecourses.science.psu.edu/stat501/node/322

4.4 Transforms and nonlinearity

And, finally, try fitting this regression function:

$$\mu_{lnY} = \beta_0 + \beta_1 ln(x)$$

if the trend in your data follows any of these patterns. That is, fit the model with $ln(y)$ as the response and $ln(x)$ as the predictor.



source: https://onlinecourses.science.psu.edu/stat501/node/322

4.4 Transforms and nonlinearity

It's easy to make transformations.

If you see that the variance of the residuals is not constant with the mean of the fitted values… you just try a transform. For example: $y \rightarrow \log(y)$
Do the residuals now look any better?

Interpretation is where things get tough.

Good explanations at:

https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-variables-are-log-transformed/

4.4 Transforms and nonlinearity

$$\text{The } \textit{linear} \text{ model: } y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \quad i = 1, ..., n$$

Interpretation:

For every one unit increase in x, the mean of y increases by β_1 units.

4.4 Transforms and nonlinearity

The *linear* model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \quad i = 1, ..., n$

Interpretation:

For every one unit increase in x, the mean of y increases by β_1 units.

```
> lm(y~x1+x2)

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
     14.030        0.539        0.124
```

Conclusion: For every additional year in age, an individual's amount of money increases on average by an estimated of $0.54 (95% C.I. = [$0.01, $1.06]), when adjusted for income (x2).

4.4 Transforms and nonlinearity

The *linear-log* model: $y_i = \beta_0 + \beta_1 log(x_i) + \epsilon_i \quad , \quad i = 1, ..., n$

Interpretation:

For every one unit increase in log(x), the mean of y increases by β_1 units.

or:

A *p*% increase in X is associated with an increase in the mean of y of $\beta_1$log([100+p]/100) units.

4.4 Transforms and nonlinearity

The *linear-log* model: $y_i = \beta_0 + \beta_1 log(x_i) + \epsilon_i \quad , \quad i = 1, ..., n$

Interpretation:

A *p*% increase in X is associated with an increase in the mean of y of $\beta_1 \log([100+p]/100)$ units.

```
> lm(y~I(log(x1))+x2)

Call:
lm(formula = y ~ I(log(x1)) + x2)

Coefficients:
(Intercept)   I(log(x1))          x2
   -24.4235      17.8226      0.0725
```

Consider a 22 year old:

$\exp(y_{22}) = \exp(\beta_0) + \exp(\beta_1 \log(22))$

and a 20 year old:

$\exp(y_{20}) = \exp(\beta_0) + \exp(\beta_1 \log(20))$

We have:

$\exp(y_{22})/\exp(y_{20}) = (22/20)^{\beta 1}$

or equivalently:

$\exp(y_{22}-y_{20}) = (22/20)^{\beta 1}$

or equivalently:

$y_{22}-y_{20} = \log[(22/20)^{\beta 1}] =$ **1.7**

Conclusion: For any 10% increase in age, an individual's amount of money increases on average by an estimated of $1.7 (95% C.I. = [$-2, $37]), when adjusted for income (x2).

4.4 Transforms and nonlinearity

The *log-linear* model: $log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \quad i = 1, ..., n$

Interpretation:

For every one unit increase in x, the mean of log(y) increases by β_1 units.

or:

The effect of a p-unit increase in X is to multiply the mean of y by $\exp(p\beta_1)$.

4.4 Transforms and nonlinearity

The *log-linear* model: $log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \quad i = 1, ..., n$

Interpretation:
The effect of a p-unit increase in X is to multiply the mean of y by exp(pβ_1).

```
> lm(log(y)~x1+x2)

Call:
lm(formula = log(y) ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
    2.71759      0.01580      0.00484
```

Consider a 21 year old:
$y_{22} = \exp(\beta_0 + \beta_1 21)$

and a 20 year old:
$y_{20} = \exp(\beta_0 + \beta_1 20)$

We have:
$y_{21}/y_{20} = \exp(\beta_1 (21\text{-}20))$

or equivalently:
$y_{21} = y_{20} (\exp(\beta_1 (1)))$

or equivalently:
$y_{21} = y_{20} (1.02)$

Conclusion: For every additional year in age, an individual's amount of money increases on average by a factor of 1.02 (95% C.I. = [1.00, 1.04]), when adjusted for income (x2).

4.4 Transforms and nonlinearity

The *log-linear* model: $log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \quad i = 1, ..., n$

Interpretation:
The effect of a p-unit increase in X is to multiply the mean of y by exp(pβ_1).

```
> lm(log(y)~x1+x2)

Call:
lm(formula = log(y) ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
    2.71759      0.01580      0.00484
```

Consider a 21 year old:
$y_{22} = \exp(\beta_0 + \beta_1 21)$

and a 20 year old:
$y_{20} = \exp(\beta_0 + \beta_1 20)$

We have:
$y_{21}/y_{20} = \exp(\beta_1 (21\text{-}20))$

or equivalently:
$y_{21} = y_{20} (\exp(\beta_1 (1)))$

or equivalently:
$y_{21} = y_{20} (1.02)$

Conclusion: For every additional year in age, an individual's amount of money increases on average by a factor of 1.02 (95% C.I. = [1.00, 1.04]), when adjusted for income (x2).

4.4 Transforms and nonlinearity

The *log-log* model: $log(y_i) = \beta_0 + \beta_1 log(x_i) + \epsilon_i \quad , \quad i = 1, ..., n$

Interpretation:

For every one unit increase in log(x), the mean of log(y) increases by β_1 units.

or:

For every p-unit increase in X, the mean of y is multiplied by $\exp(a\beta_1)$, where a=log([100+p]/100) .

4.4 Transforms and nonlinearity

The *log-log* model: $log(y_i) = \beta_0 + \beta_1 log(x_i) + \epsilon_i \quad , \quad i = 1, ..., n$

Interpretation:
For every p-unit increase in X, the mean of y is multiplied by $\exp(a\beta_1)$, where $a=\log([100+p]/100)$.

```
> lm(log(y)~I(log(x1))+x2)

Call:
lm(formula = log(y) ~ I(log(x1)) + x2)

Coefficients:
(Intercept)   I(log(x1))           x2
    1.54652      0.53611     0.00328
```

Exercise:
How would you interpret this?
How would you obtain a the proper confidence interval?

4.4 Transforms and nonlinearity

Good explanations and examples from economics:

http://kenbenoit.net/assets/courses/ME104/logmodels2.pdf

4.5 Diagnostics for data collected sequentially in time

4.5 Diagnostics for data collected sequentially in time

If data are collected sequentially in time, the assumption of independent ϵ_i in model (3.36) might not be appropriate. In this situation, sometimes ϵ_i are serially correlated. If the ϵ_i are positive serially correlated, then SEs of $\hat{\beta}$'s derived under the assumption of independent ϵ_i are generally too small.

4.5 Diagnostics for data collected sequentially in time

4.5 Diagnostics for data collected sequentially in time

If data are collected sequentially in time, the assumption of independent ϵ_i in model (3.36) might not be appropriate. In this situation, sometimes ϵ_i are serially correlated. If the ϵ_i are positive serially correlated, then SEs of $\hat{\beta}$'s derived under the assumption of independent ϵ_i are generally too small.

4.5 Diagnostics for data collected sequentially in time

As before, suppose least square has been applied to get $\hat{\boldsymbol{\beta}}$ and residuals $e_1, \ldots, e_n$. One additional residual plot is e_t versus t for $t = 1, \ldots, n$. If there are trends or patterns in this plot, then the assumption of independent ϵ_t is violated. There can be several patterns, and two of them are as follows; see also Figure 4.5.

(a) A positive e_t is often followed by a negative residual and a negative e_t is often followed by a positive residual. This is negative serial correlation, and the plot shows with many crossings of the horizontal line if consecutive residuals are joined by line segments.

(b) A positive e_t is often followed by a positive residual and a negative e_t is often followed by a negative residual. This is positive serial correlation, and the plot shows alternating runs of several positive residuals and runs of several negative residuals, with few crossings of the horizontal line if consecutive residuals are joined by line segments.

4.5 Diagnostics for data collected sequentially in time

As before, suppose least square has been applied to get $\hat{\boldsymbol{\beta}}$ and residuals $e_1, \ldots, e_n$. One additional residual plot is e_t versus t for $t = 1, \ldots, n$. If there are trends or patterns in this plot, then the assumption of independent ϵ_t is violated.

Important Question: Is there any "serial correlation" ?

To find out, first plot the *residuals vs. time*.

4.5 Diagnostics for data collected sequentially in time

Below is the table of the lagged residual e_{t-1} versus e_t

e_1	e_2
e_2	e_3
$\vdots$	$\vdots$
e_{n-1}	e_n

The lag 1 serial correlation (with $\bar{e} = 0$) is

$$\hat{\rho}_1 = \frac{\sum_{t=2}^{n} e_{t-1} e_t}{\sqrt{\sum_{t=2}^{n} e_{t-1}^2 \sum_{t=2}^{n} e_t^2}} \in (-1, 1).$$

One could also look at scatterplot of e_{t-1} versus e_t to check if there appears to be significant serial correlation

4.5 Diagnostics for data collected sequentially in time

The Durbin-Watson statistic is

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}. \tag{4.33}$$

If n is large, then

$$DW = \frac{\sum_{t=2}^{n} e_t^2 + \sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_{t-1}e_t}{\sum_{t=1}^{n} e_t^2} \approx 2 - 2\hat{\rho}_1.$$

Hence $DW \approx 2$ for serially uncorrelated residuals, $DW \approx 0$ for strong positive serial correlation ($\hat{\rho}_1$ near 1) and $DW \approx 4$ for strong negative serial correlation ($\hat{\rho}_1$ near -1), since

$$-1 \leq \hat{\rho}_1 \leq 1 \Rightarrow 0 \leq 2 - 2\hat{\rho}_1 \leq 4. \tag{4.34}$$

https://www.youtube.com/watch?v=QGN073tpDXE&ab_channel=BenLambert

4.7 Summary of variable selection

Variable selection methods are algorithms for finding a good subset of explanatory variables. Cross-validation, either leave-one-out or with training/holdout subsets, provide an out-of-sample comparison of different prediction equations; the assessment of predictability is applied to data that are not used in the estimation of regression coefficients. Cross-validated and out-of-sample root mean squared errors combined with residual plots and other diagnostics could provide information that transforms of variables are needed or that the response variables in nonlinear in the explanatory variables.