SF2930 - Regression analysis

KTH Royal Institute of Technology, Stockholm

Lecture 5 – Prediction accuracy and model assessment (MPV 4)

January 28, 2022

Todays lecture

- Prediction and hidden extrapolation
- Rescaled residuals
- Testing the normality assumptions
- Different useful plots of residuals

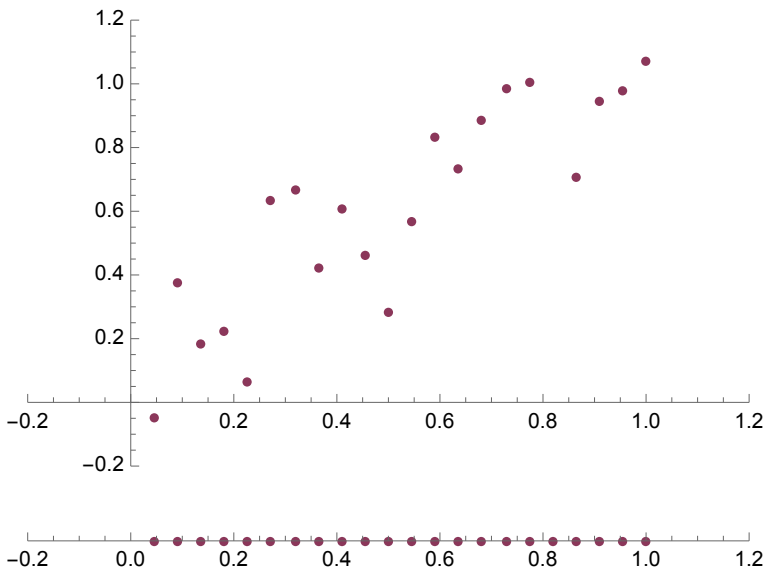
Prediction of new observations

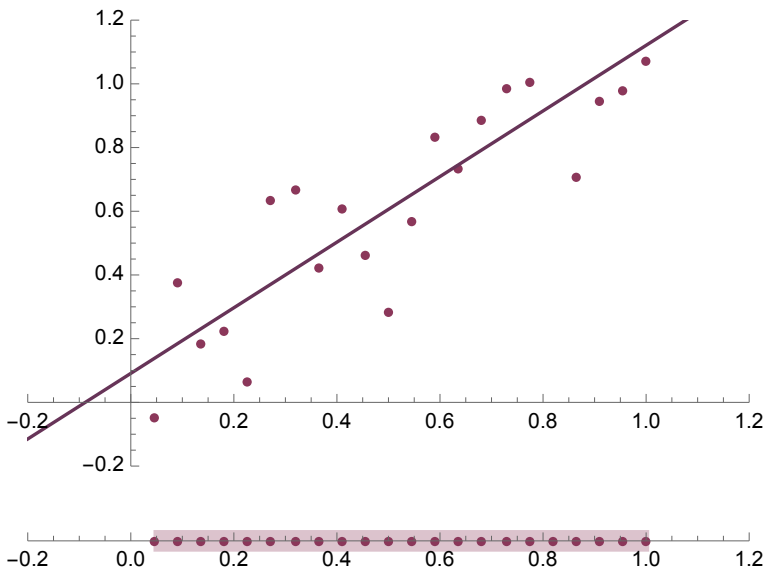
Assume that we want to predict y_0 at a point $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{0n})$. Then a point estimate is given by $\hat{y}_0 = \mathbf{x}_0^T \hat{\boldsymbol{\beta}}$, and a confidence interval is given by

$$y_0 = \hat{y}_0 \pm t_{\alpha/2, n-k-1} \sqrt{\hat{\sigma}^2 (1 - \mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0)}$$

Extrapolation

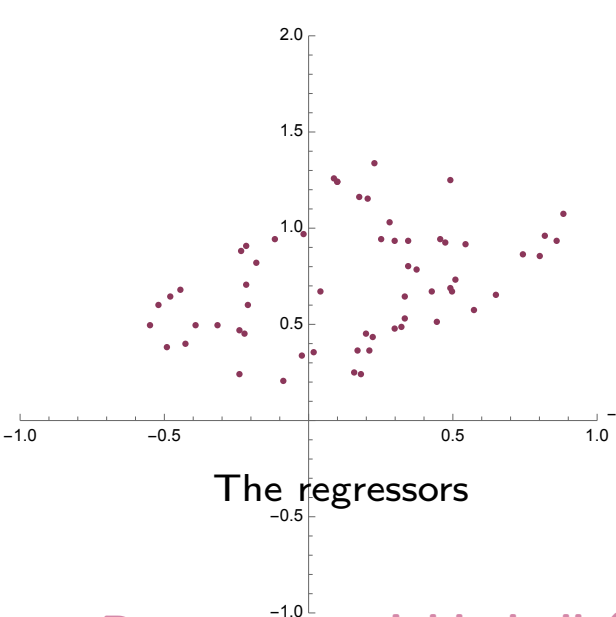
"We extrapolate when we predict a response y_0 at a point x outside of the set containing the regressors."



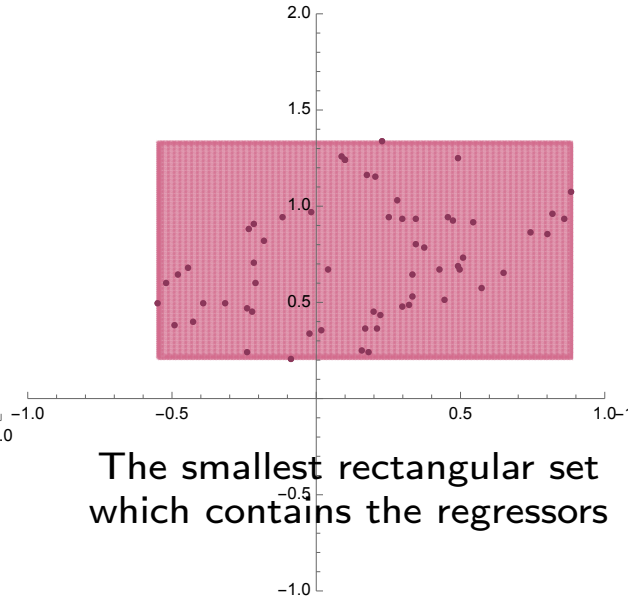


Extrapolation

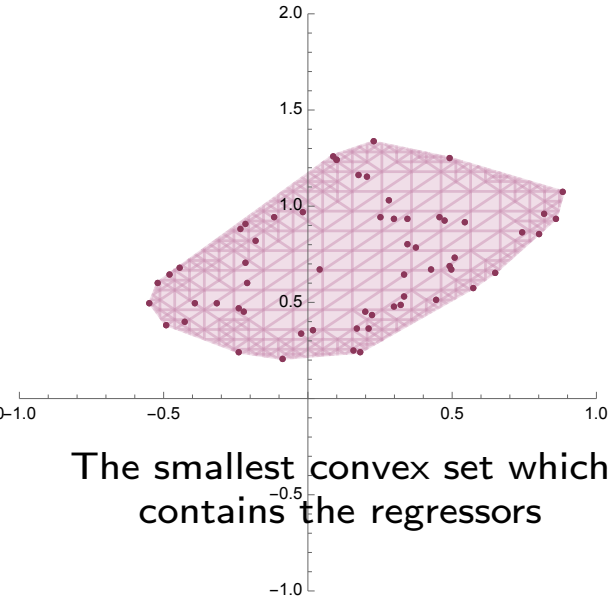
"We extrapolate when we predict a response y_0 at a point $\mathbf{x}_0$ outside of the set containing the regressors."



The regressors



The smallest rectangular set which contains the regressors



The smallest convex set which contains the regressors

Regressor variable hull (RVH)

The smallest convex set containing all the original data points.

Intrapolation vs extrapolation

Prediction at a data point inside the RVH is called *interpolation*, and prediction outside the RVH is called *extrapolation*.

Hidden extrapolation

Extrapolation at points which are in the rectangular hull, but not in the RVH.

Extrapolation

If we have many regressors, then it is hard to visualize the RVH. How can we avoid hidden extrapolation?

Ellipsical approximations of the RVH

Note that

- $\mathbf{x}_j^T (X^T X)^{-1} \mathbf{x}_j = X^T (X^T X)^{-1} X(j,j) = H(j,j)$
- $\{\mathbf{x} \colon \mathbf{x}^T (X^T X)^{-1} \mathbf{x} \leq t\}$ is always an ellipse.

Hence the set

$$\text{ERVH} = \{\mathbf{x} \colon \mathbf{x}^T (X^T X)^{-1} \mathbf{x} \leq \max \operatorname{diag} H\}$$

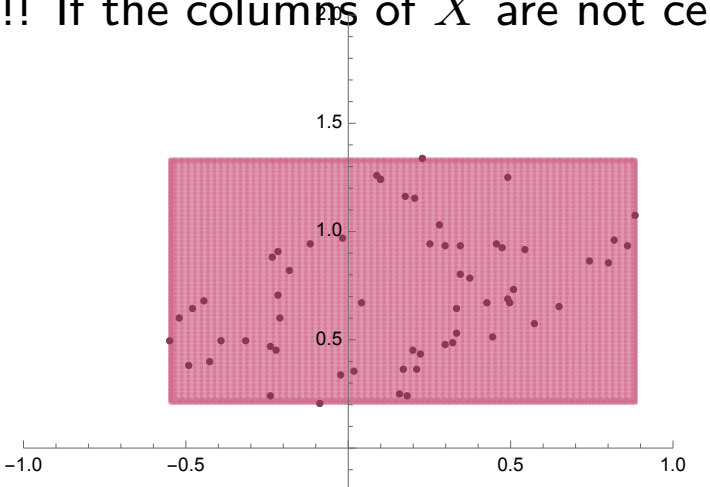
contains the RVH. Equivalently, ERVH is the smallest ellipse, centered at the origin, which contains all the points in the RVH.

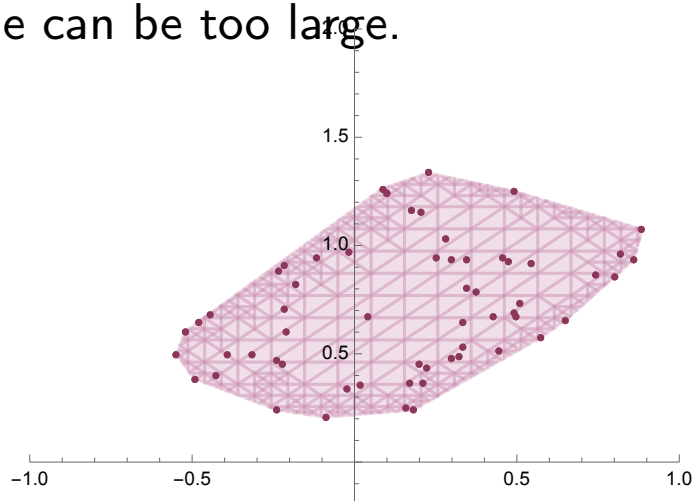
$\rightarrow$ It is (much!) easier to check if a point belongs to ERVH than if it belongs to RVH.

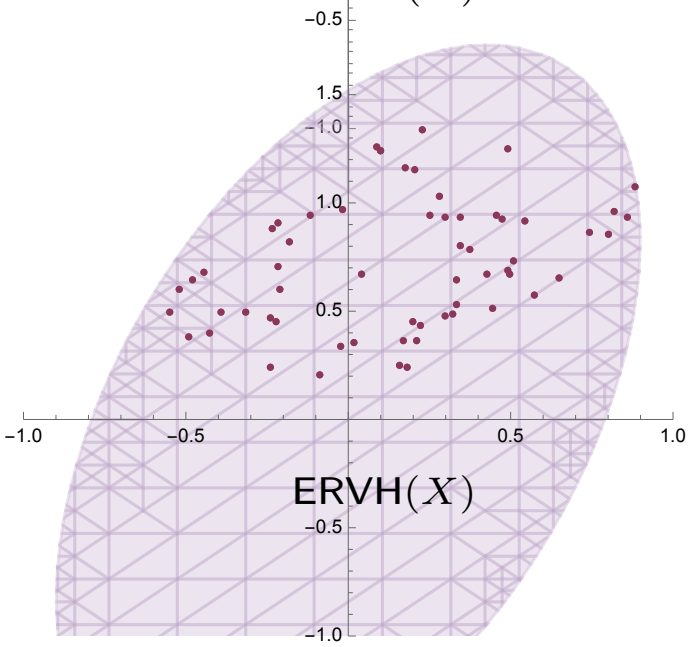
Extrapolation

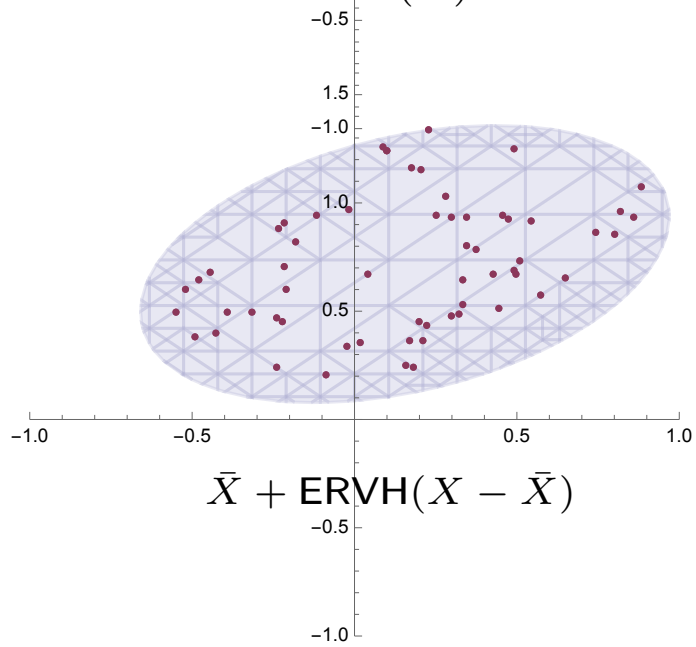
Comparison of approximations of the RVH

!! If the columns of X are not centered, the can be too large.









How to check that the model assumptions hold?

Which assumptions have we made?

- There is a linear relationship between the regressors and the response
- The errors ε_i have mean 0, variance σ^2, and are uncorrelated
- The error $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$.

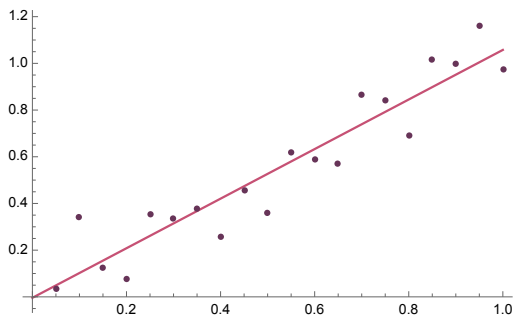
Even though the fit of the model, using the statistics from previous lectures, are good, this will be meaningless if the model assumptions are violated.

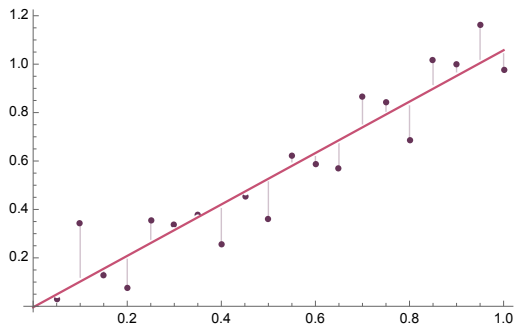
Reminder on residuals

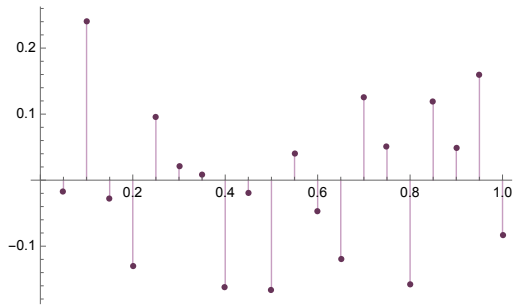
Residuals

Measures deviation between data and fit.

$$e_i := y_i - \hat{y}_i$$







Properties of the residuals

- $\mathbb{E}[e_i] = 0$
- $\mathrm{Var}(e_i) \approx \frac{\sum (e_i - \bar{e})^2}{n-k-1} = \frac{SS_{Res}}{n-k-1}$
- *Not* independent (this can be seen from the degree of freedom, which should be n if they were independent).

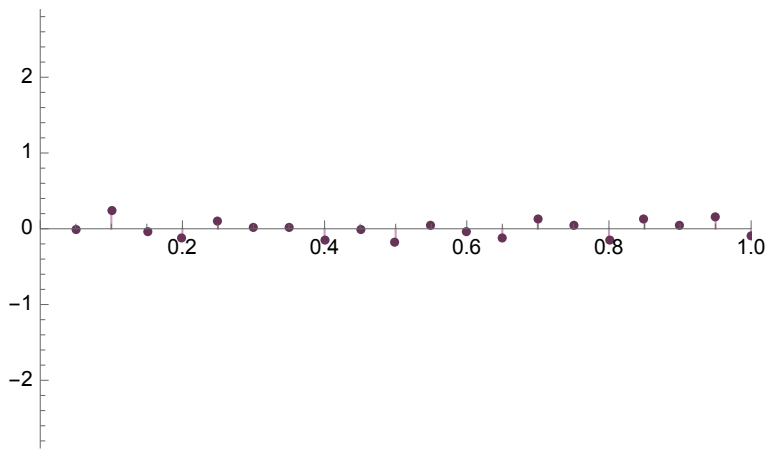
Rescaled residuals

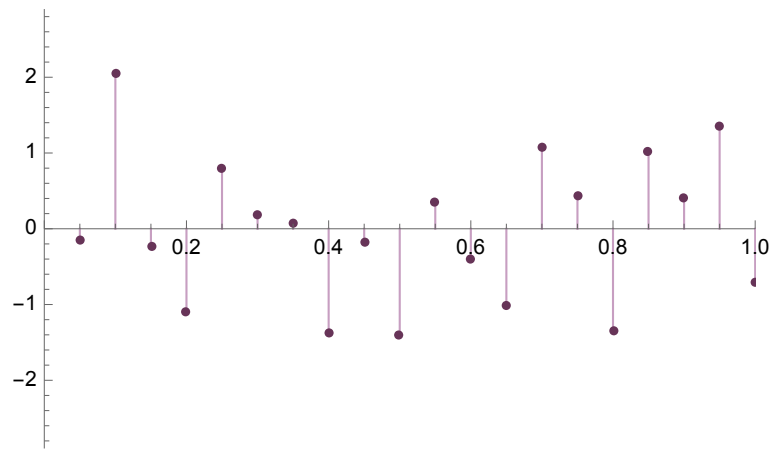
Standardized residuals

Rescaled to have approximate variance 1. Makes sense to rescale to be able to interpret residual size.

$$d_i := \frac{e_i}{\sqrt{MS_{Res}}} = \frac{e_i}{\sqrt{SS_{Res}/(n-k-1)}}.$$

standardized residuals





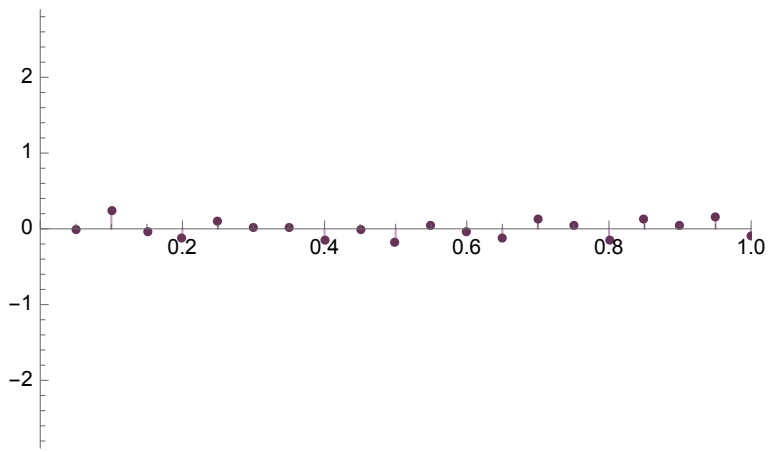
Rescaled residuals

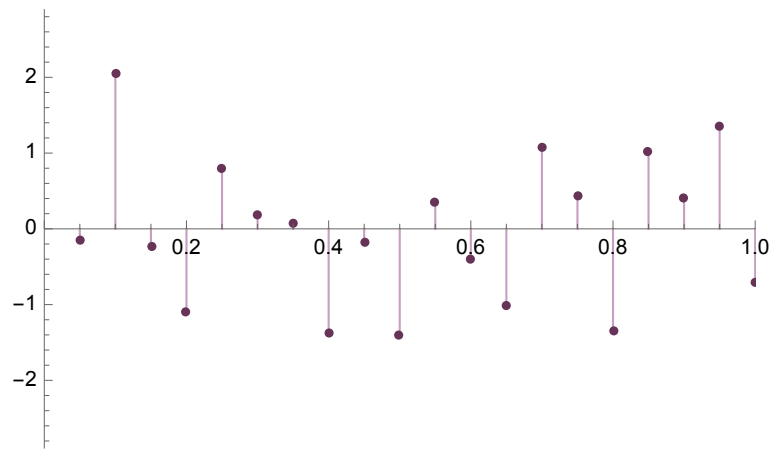
Studentized residuals

We previously showed that $\operatorname{Cov} \mathbf{e} = \sigma^2(I - H)$ (not diagonal, since correlated!) and thus $\operatorname{Var} e_i = \sigma^2(1 - h_{ii})$. In particular, the variance is not the same for all points! Approximating σ with MS_{Res}, we define

$$r_i := \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}} = \frac{e_i}{\sqrt{MS_{Res}(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}})}}.$$

studentized residuals





Rescaled residuals

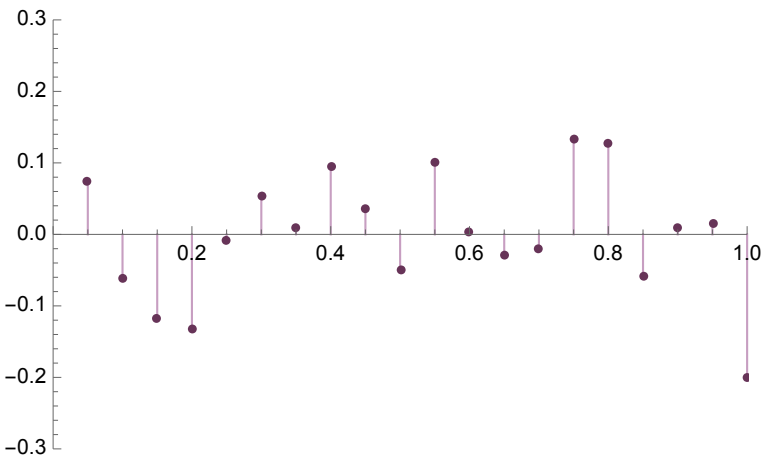
PRESS residuals

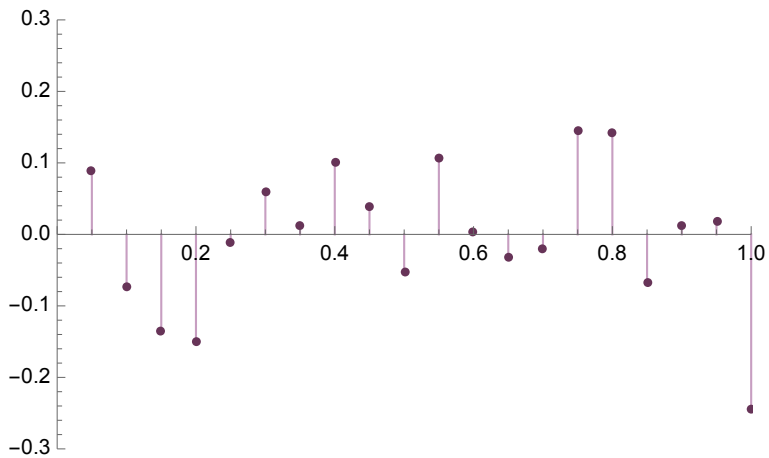
If observation i is an outlier (which thus violates the model assumptions), then it is likely that removing the outlier will affect the fitted model. In other words, if we let $\hat{y}_{(i)}$ be the predicted value at $\mathbf{x}_i$ if we remove $(\mathbf{x}_i, y_i)$ from our data set, then

$$e_{(i)} = y_i - \hat{y}_{(i)} = \frac{e_i}{1 - h_{ii}}$$

PRESS residuals

should be large.





Rescaled residuals

Externally studentized residuals (R-student)

Recall that the studentized residuals were given by

$$r_i := e_i / \sqrt{MS_{Res}(1 - h_{ii})}.$$

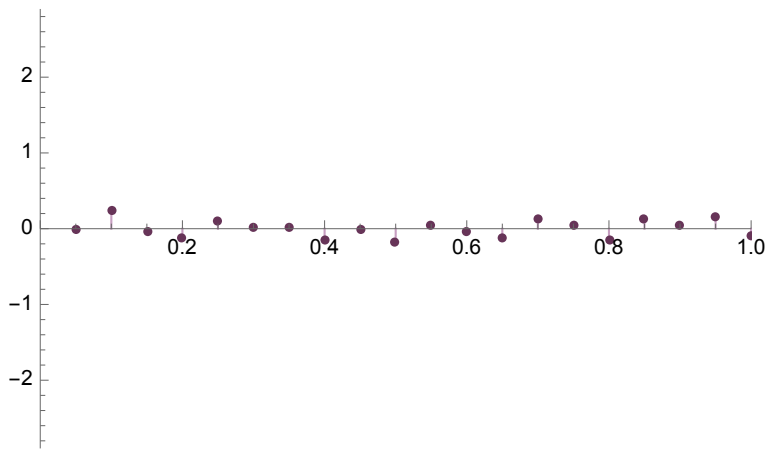
internally studentized residuals

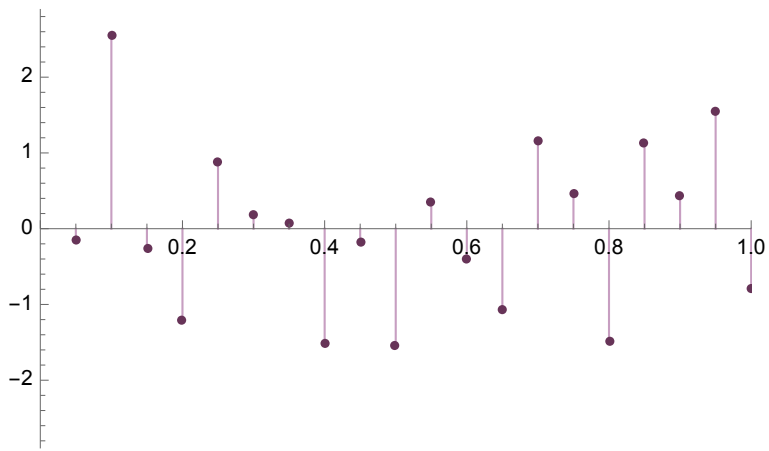
Since the ith oberservation was used to estimate σ^2, we say that these are *internally studentized residuals.*

If $(\mathbf{x}_i, y_i)$ is an outlier, then it might effect the estimate $\hat{\sigma}^2$ much. This motivates removing this point when estimating σ^2 before normalizing e_i;

$$t_i := \frac{e_i}{\sqrt{S^2_{(i)}(1 - h_{ii})}} \sim t_{n-k-1-1}.$$

externally studentized residuals





How can we check the normality assumption?

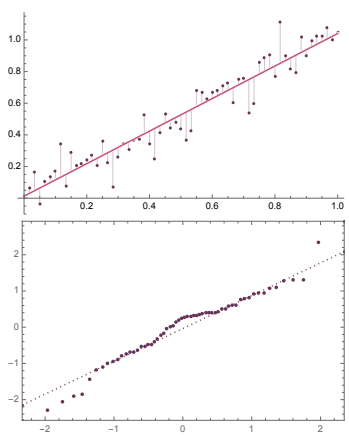
If $\varepsilon_i \sim N(0, \sigma^2)$, then we have

$$t_i := \frac{e_i}{\sqrt{S^2_{(i)}(1 - h_{ii})}} \sim t_{n-k-1-1} \overset{n-k-2=\infty}{\sim} N(0, 1)$$

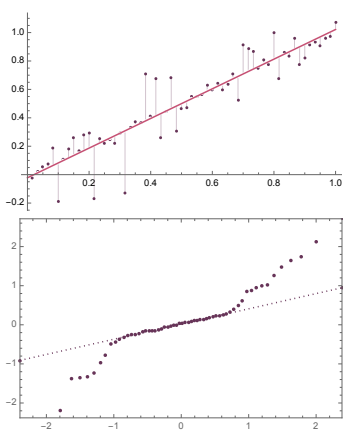
externally studentized residuals

1. Order the externally studentized residuals $t^{(1)} \leq t^{(2)} \leq \ldots t^{(n)}$.
2. Sample n points from an independent normal distribution, and order them $z^{(1)} \leq z^{(2)} \leq \ldots \leq z^{(n)}$.
3. Plot the points $(z^{(1)}, t^{(1)})$, $(z^{(2)}, t^{(2)})$, $\ldots$, $(z^{(n)}, t^{(n)})$.

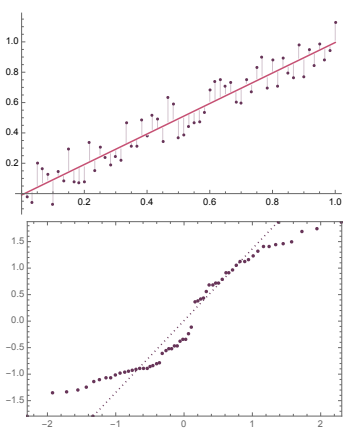
If the normality assumption is correct, then these should lie on a straight line.



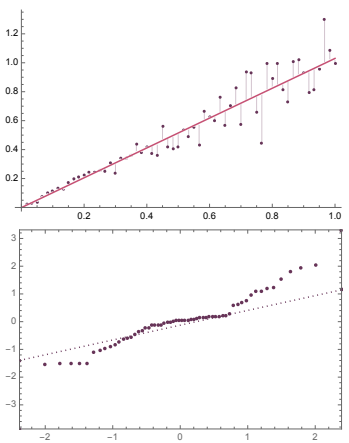
$y = x + \varepsilon$



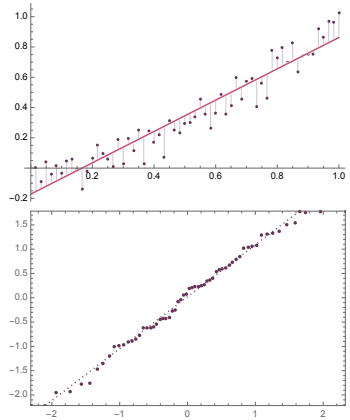
$y = x + |\varepsilon|^2 \operatorname{sgn} \varepsilon$



$y = x + \sqrt{|\varepsilon|} \operatorname{sgn} \varepsilon$



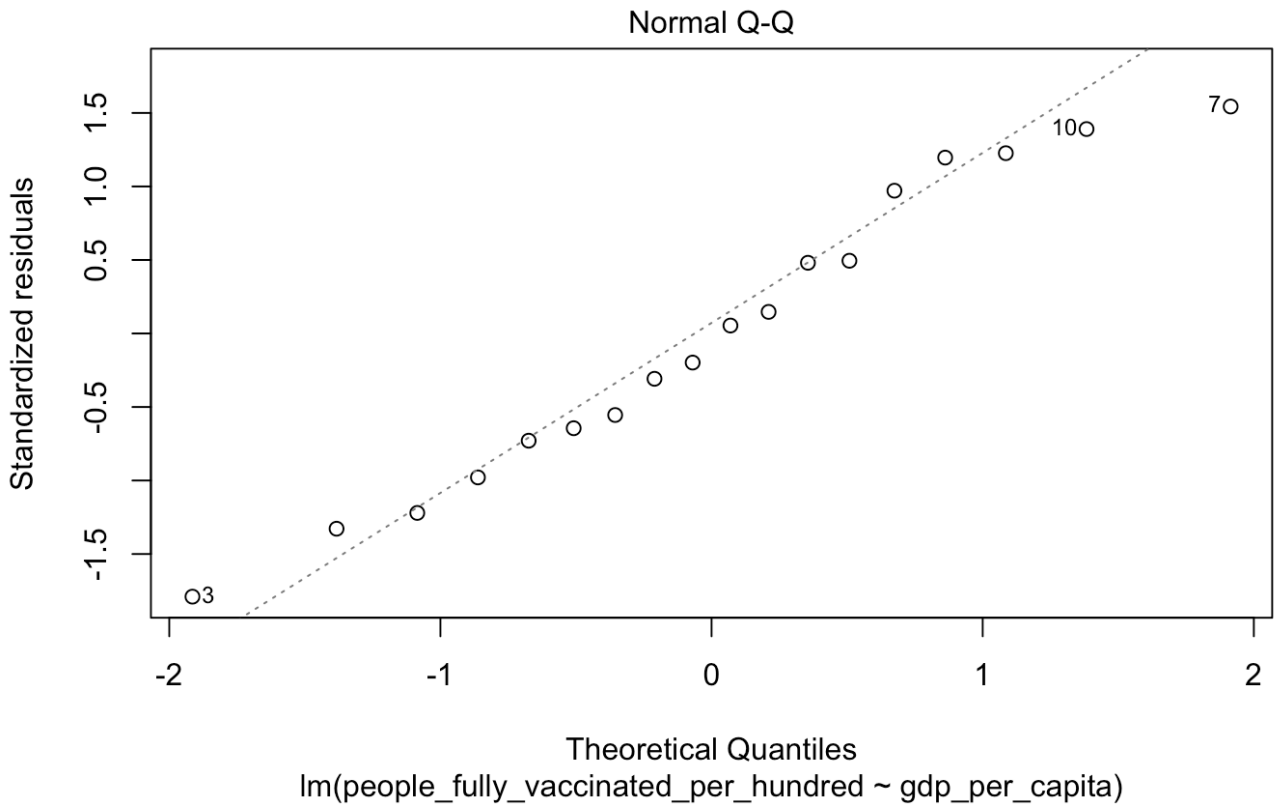
$y = x + x\varepsilon$



$y = x^2 + \varepsilon$

Example: Normality plot

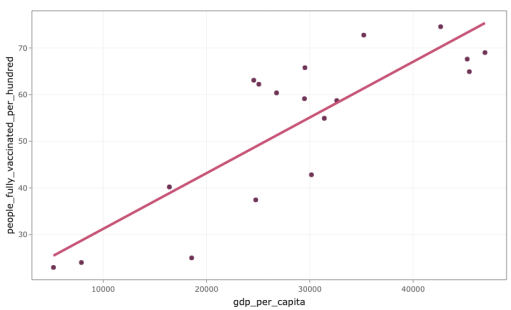
```
1 plot(df00.model,2)
```



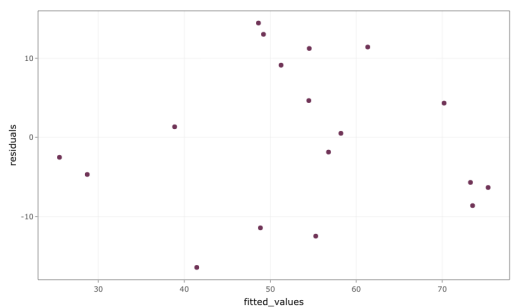
Example: Plot of residuals vs fitted values

If the model is correct, residuals should be evenly spread out in a horisontal band.

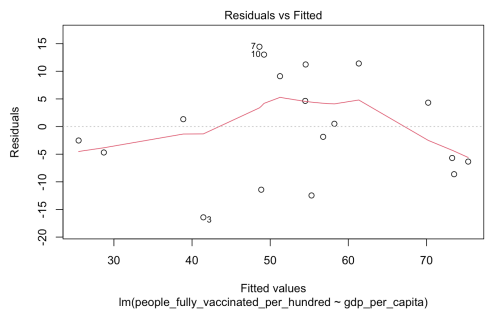
```
pp <- ggplot(model00, aes(x=fitted_values, y=residuals)) +
    geom_point(aes(text=location),color="#703457") + theme_
    bw()
ggplotly(pp, tooltip="text")

plot(df00.model,1)
```



Data and fitted line.



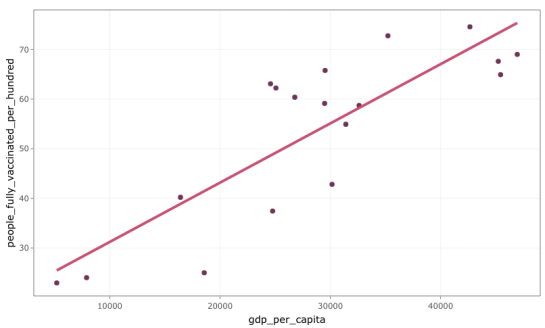
Residuals vs. fitted values.



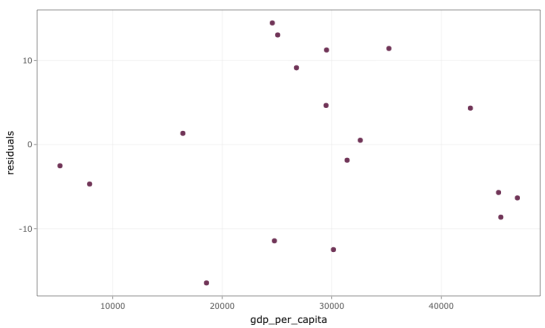
Residuals vs. fitted values.

Example: Plots of residuals vs regressors

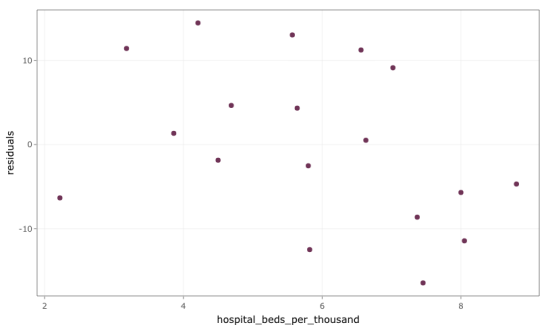
```
pp <- ggplot(model00, aes(x=gdp_per_capita, y=residuals)) +
    geom_point(aes(text=location),color="#703457") + theme_
    bw()
ggplotly(pp, tooltip="text")

pp <- ggplot(model00, aes(x=hospital_beds_per_thousand, y=
    residuals)) + geom_point(aes(text=location),color="
    #703457") + theme_bw()
ggplotly(pp, tooltip="text")
```



Data and fitted line.



Residuals vs. one of the regressors included in the model.



Residuals vs. one of the regressors not included in the model.

The PRESS

Recall that the press residuals were computed as $e_{(i)} = y_i - \hat{y}_{(i)}$
PRESS statistic:

$$PRESS = \sum (y_i - \hat{y}_{(i)})^2 = \sum \left(\frac{e_j}{1 - h_{ii}} \right)^2$$

A measure of how well the model can predict new data. In general, we want PRESS to be small.

$$R^2_{prediction} := 1 - \frac{PRESS}{SS_T}$$

"the model explains about $100R^2\%$ of the variability in predicting new observations

Example

```r
# the press residuals
pr <- resid(df00.model)/(1 - lm.influence(df00.model)$hat)
PRESS <- sum(pr^2)
PRESS

# the total sum of squares
SSt <- sum((df00$people_fully_vaccinated_per_hundred - mean(
    df00$people_fully_vaccinated_per_hundred))^2)

# The R2_prediction statistic
R2 <- 1 - PRESS/SSt
R2
```

```
[1] 1792.621
[1] 0.6393454
```