STATISTICAL MODELLING: Theory and practice

Project 1: Wind power data

GOALS: ASSIGNMENT 1

Descriptive statistics

Simple models

1. Fit different probability density models to wind power, wind speed and wind direction data.
2. Conclude on the most appropriate model for each variable.
3. Report parameters including assessment of their uncertainty.

Descriptive statistics

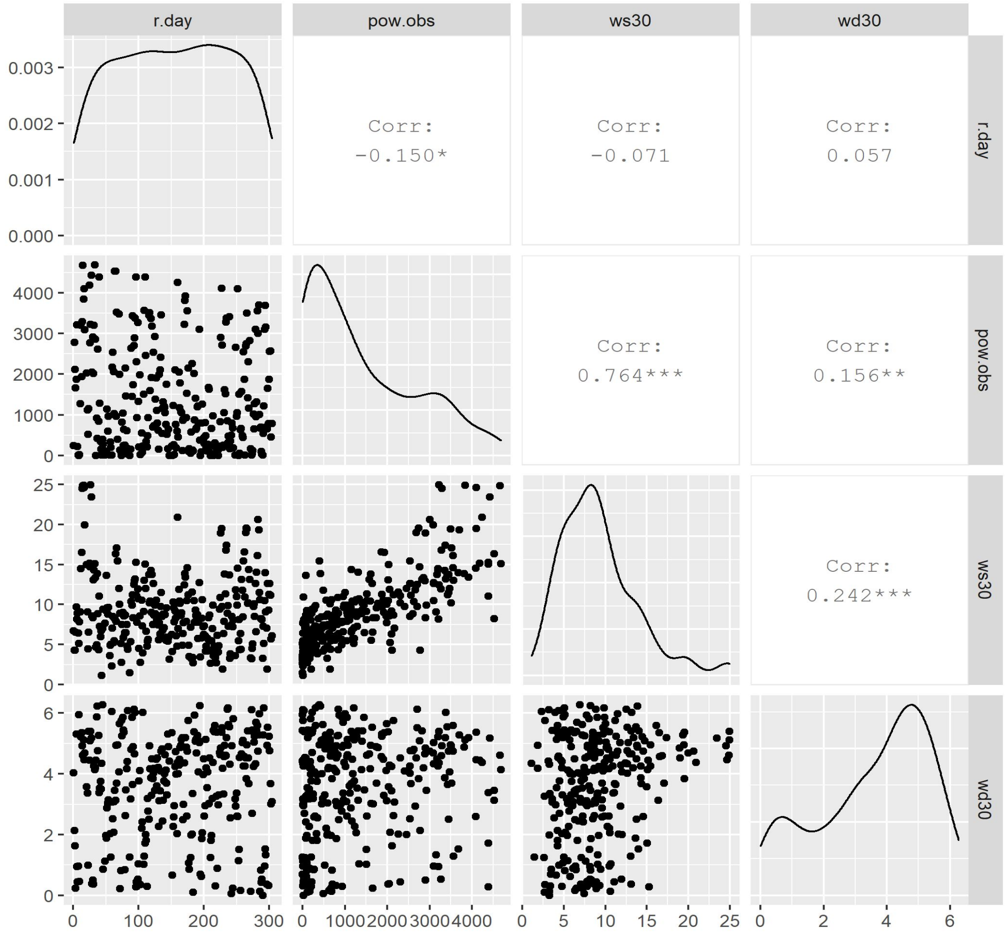
r.day	month	day	pow.obs	ws30	wd30
1	1	1	243.0277778	6.723611	4.0343405
2	1	2	2780.0136986	4.272603	2.1365208
3	1	3	2118.6164384	4.272603	1.6240318
4	1	4	1660.8767123	6.541096	0.2269022
5	1	5	1872.7945205	9.713699	5.3161852
6	1	6	3212.2602740	8.161644	0.9522963

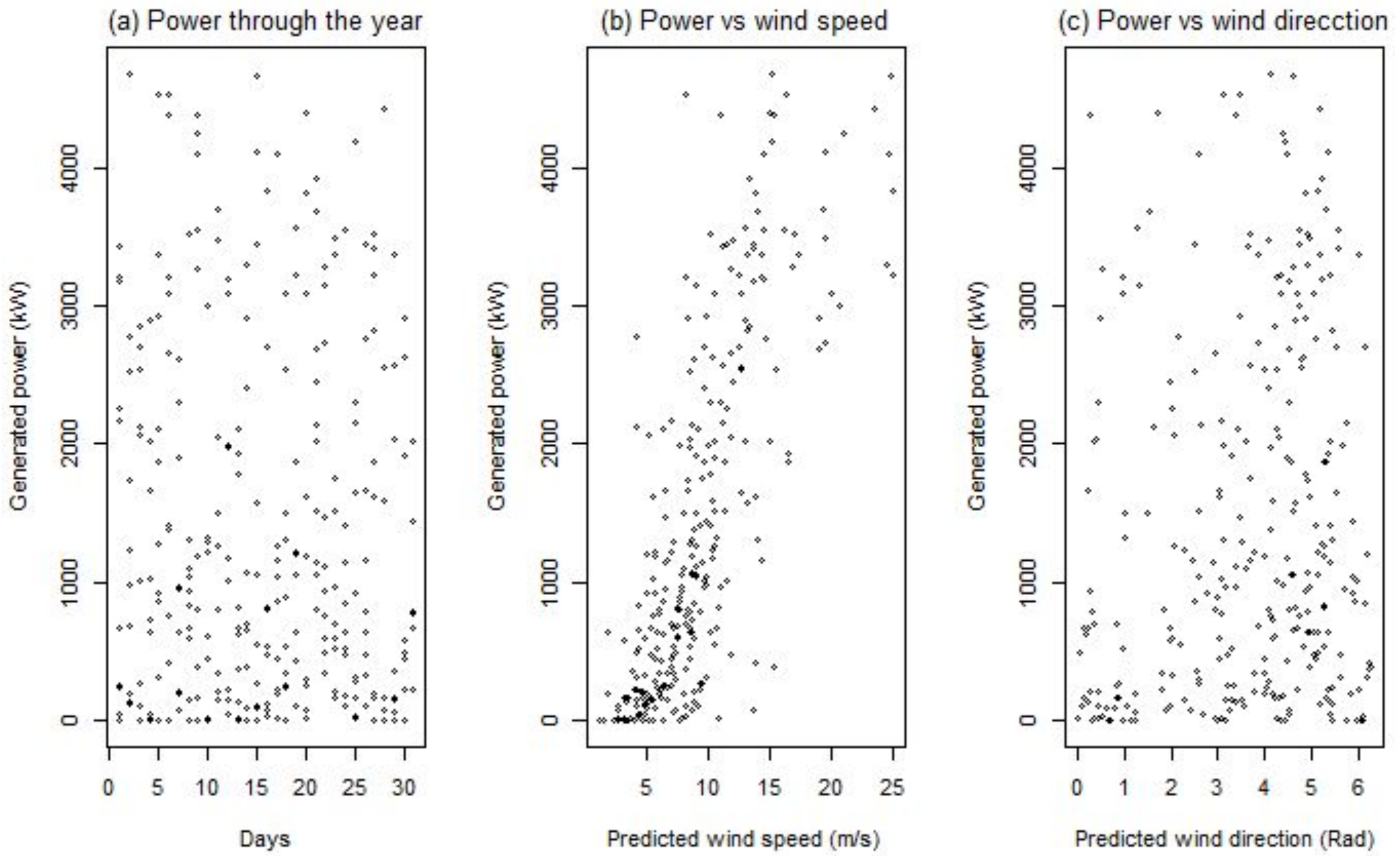
288 x 6

Variable	Meaning	Unit
`r.day:`	Days since 1/1 2003	days
`month:`	Month in year	
`day:`	Day in month	
`pow.obs:`	Average daily wind power production	kW
`ws30:`	Predicted wind speed 30 meters above ground level	m/s
`wd30:`	Predicted wind direction (0 north, $\pi/2$ east) 30 meters above ground level	rad

```
pow.obs             ws30             wd30
Min.   :   0.123   Min.   : 1.139   Min.   :0.000095
1st Qu.: 254.158   1st Qu.: 5.779   1st Qu.:2.474999
Median : 964.123   Median : 8.498   Median :4.079297
Mean   :1381.196   Mean   : 9.112   Mean   :3.602390
3rd Qu.:2196.579   3rd Qu.:11.202   3rd Qu.:4.945443
Max.   :4681.062   Max.   :24.950   Max.   :6.274642
```

Descriptive statistics





Simple models

Normalization of power production:

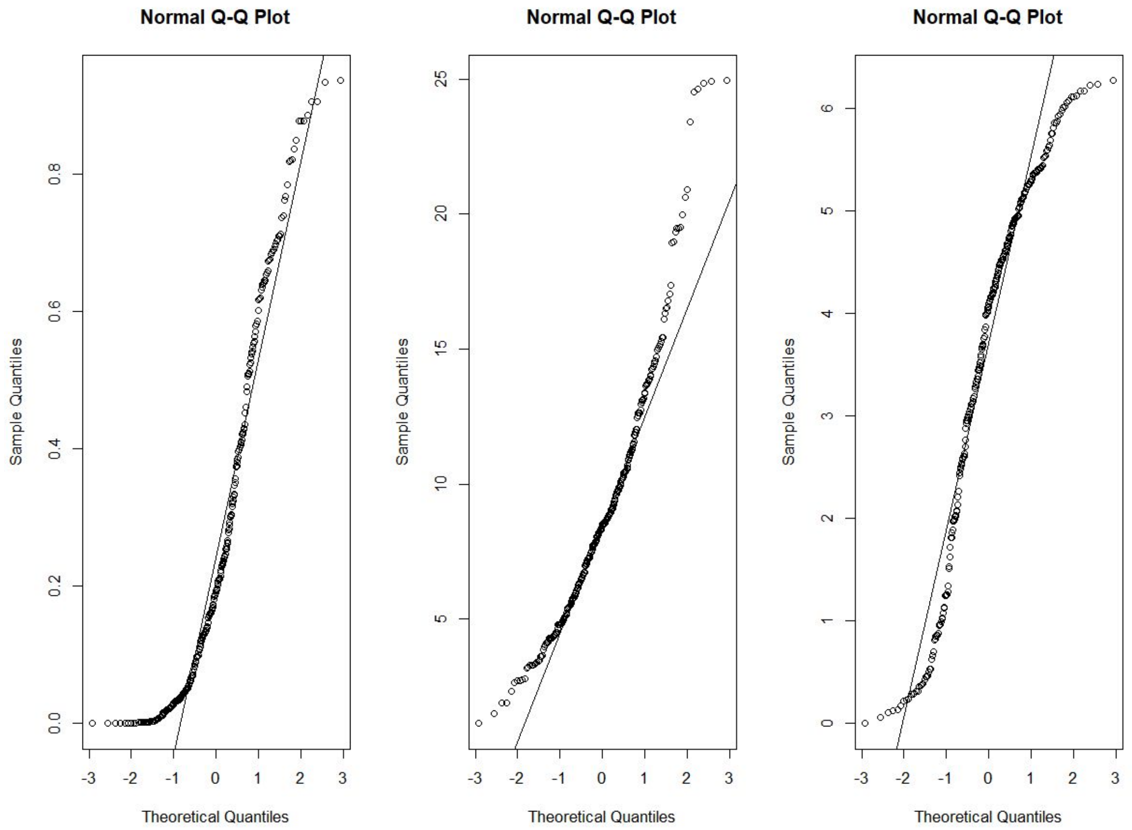
$$normpow = \frac{pow.obs - min(pow.obs)}{max(pow.obs) - min(pow.obs)}$$

The normalization is based on the installed capacity, which maximum is 5000 kW.

Values that from 0 to 1 for power production.

Checking for normality of the data:

Power production | Wind speed | Wind direction



Simple models

Model	Parameter 1	CI Parameter 1	Parameter 2	CI Parameter 2
Gamma	0.6926402	[0.6246109, 0.7606695]	2.5073938	[2.159441, 2.855347]
Beta	0.5571045	[0.4952471, 0.6189619]	1.4918277	[1.286340, 1.697315]
VonMises	0.2738893	[0.2443170, 0.3034616]	15.7607520	[13.23072, 18.29079]
Log-normal	0.0000000	[-0.3335482, 0.3335482]	2.888063	[2.652209, 3.123918]

Power production: Non-normal models

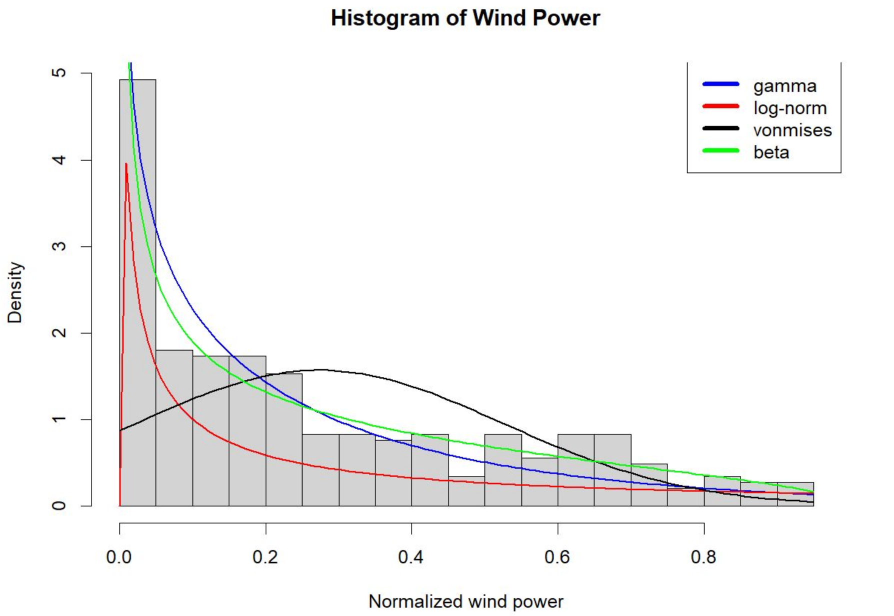
Gamma (dgamma) | Beta (dbeta) | Circular (vonMises) | log-normal (dlnorm)

→ Negative log-likelihood function

→ Optimization (nlminb)

→ Extract parameters and uncertainty | AIC

Model	Parameters	AIC
Gamma	[0.6926402, 2.5073938]	-190.7635
Beta	[0.5571045, 1.4918277]	-239.3236
VonMises	[0.2738893, 15.7607520]	36.75738
Log-normal	[0.000000, 2.888063]	187.8728



Simple models

Power production: Normal models

Box-Cox Transformation

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda}{}_{i} - 1}{\lambda} & \lambda \neq 0 \\ \log(y_i) & \lambda = 0 \end{cases}$$

$$\lambda = 0.3467256$$

Transformation 1

$$y^{(\lambda)} = \frac{1}{\lambda} \log\left(\frac{y^{\lambda}}{1 - y^{\lambda}}\right); \quad \lambda > 0$$

$$\lambda = 0.2620668$$

Transformation 2

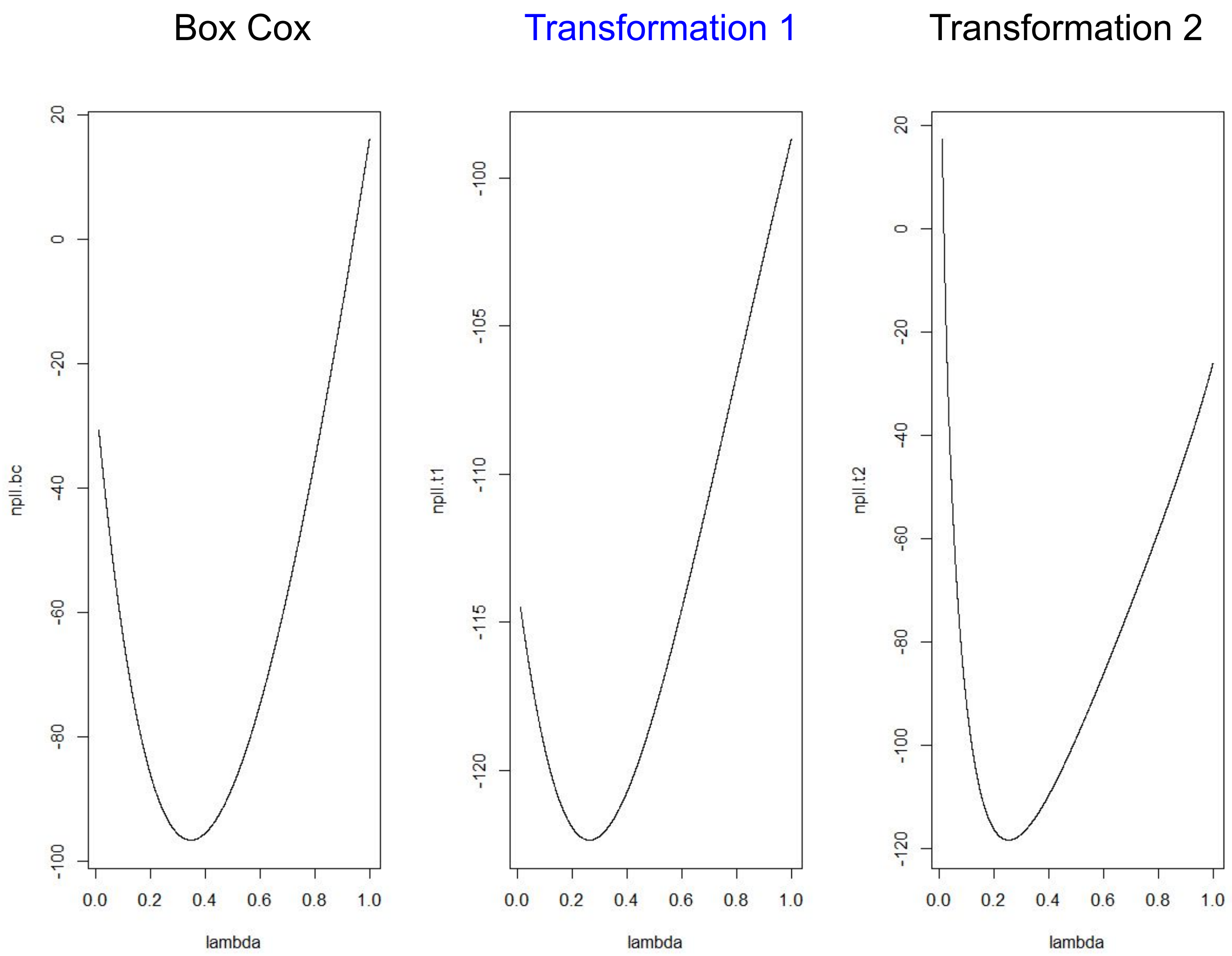
$$y^{(\lambda)} = 2\log\left(\frac{y^{\lambda}}{(1 - y)^{1-\lambda}}\right); \quad \lambda \in (0, 1)$$

$$\lambda = 0.2523665$$

Simple models

Power production: Normal models

Transformation	λ	min(npll)
Box Cox	0.3467256	-96.61813
1	0.2620668	-122.3459
2	0.2523665	-118.4156

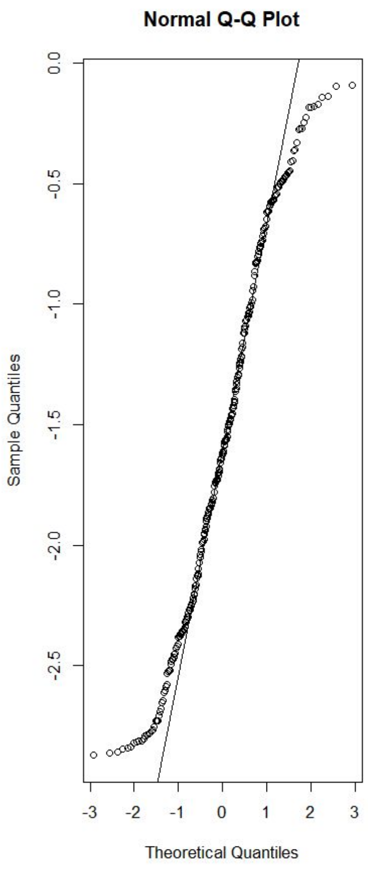


Simple models

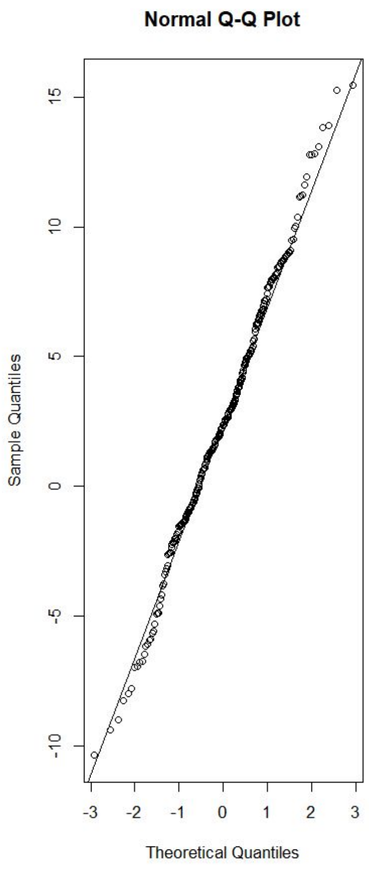
Power production: Normal models

Transformation 1 is more suitable for the variable *Power production.*

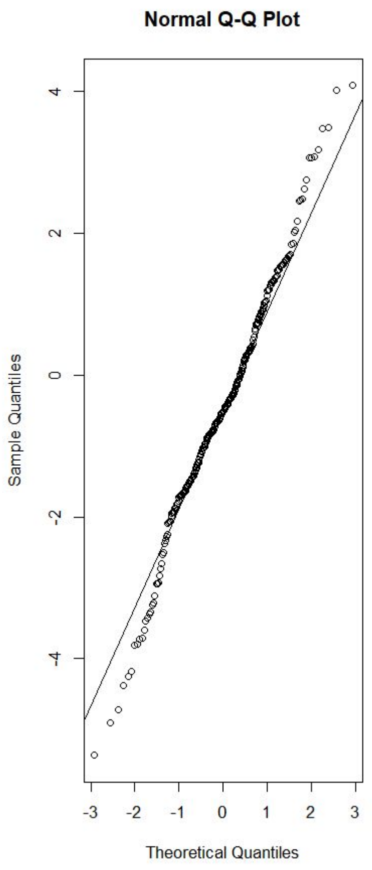
Box Cox



Transformation 1



Transformation 2



Simple models

Power production:

Transformation 1 to
Power production

NON- NORMAL DISTRIBUTION

FIT NON-NORMAL MODELS

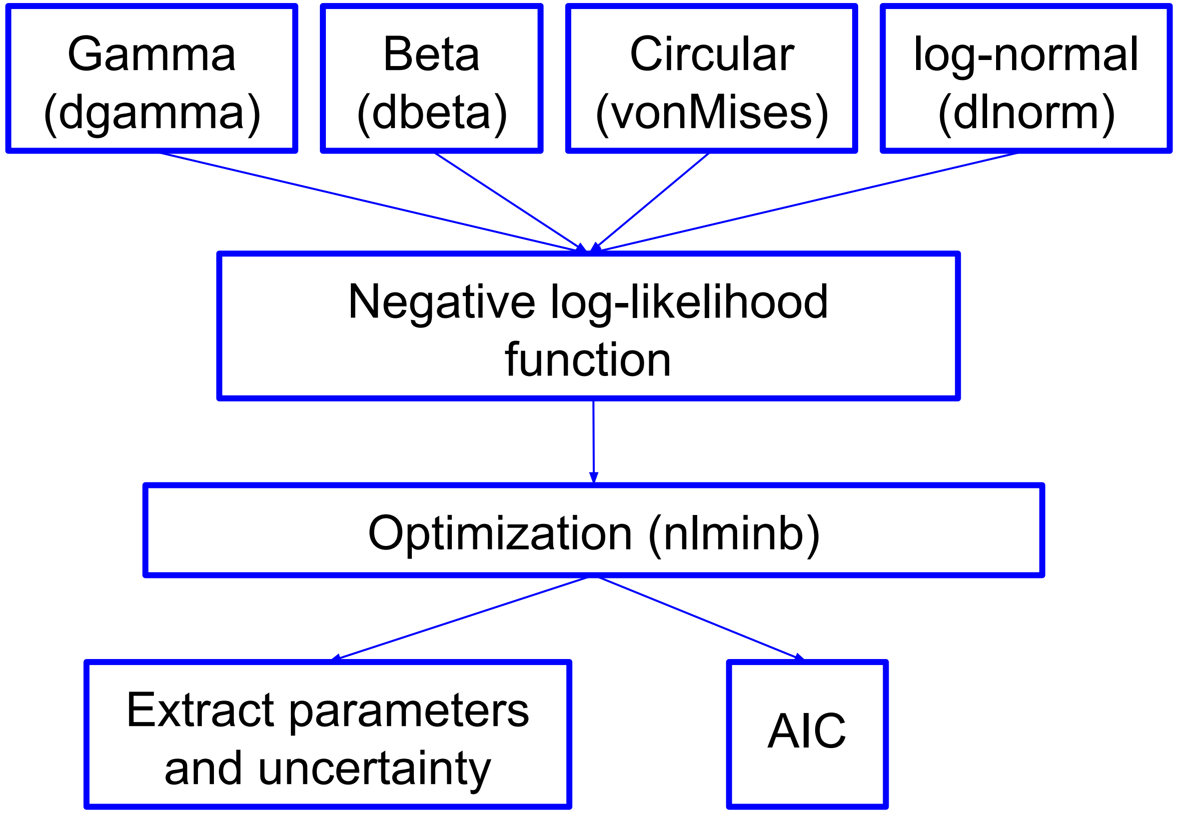
NORMAL DISTRIBUTION

NORMAL MODEL

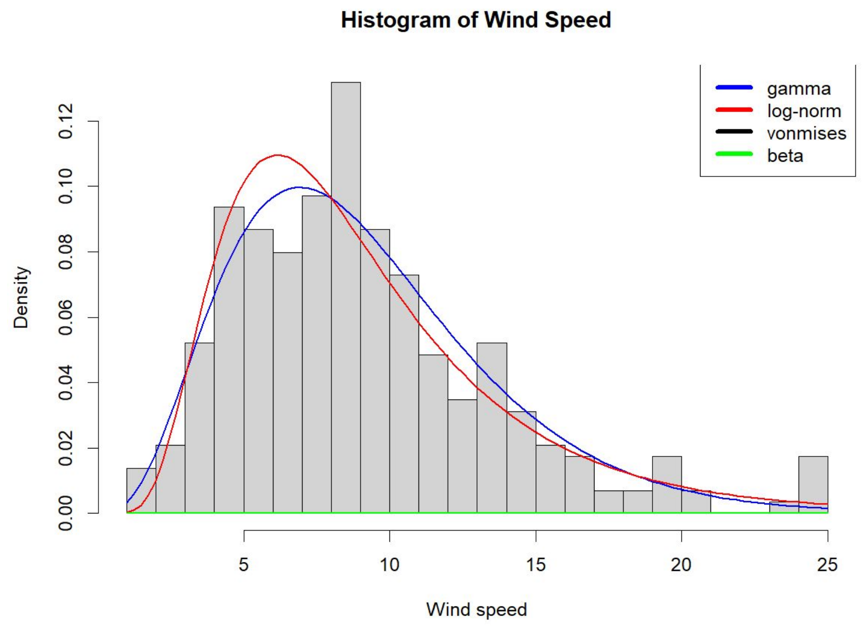
BETA REGRESSION fits better to the variable *Power production.*

Simple models

Wind speed: non-normal models



Log norm model	
μ	2.0844793
CI μ	[2.024706, 2.144253]
σ	0.5175574
CI σ	[0.4752910, 0.5598238]



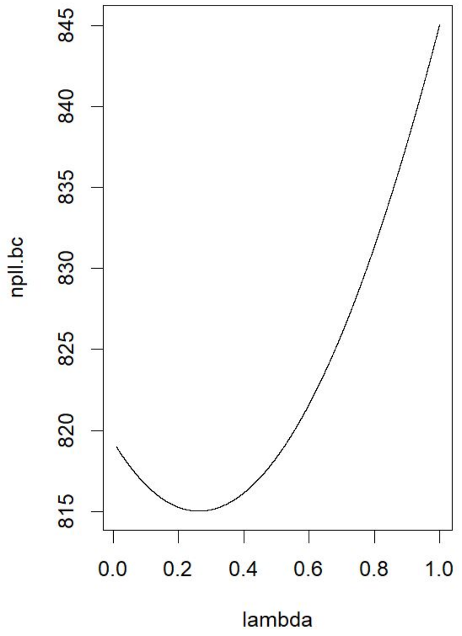
Lognorm is the most appropriate model for **wind speed**

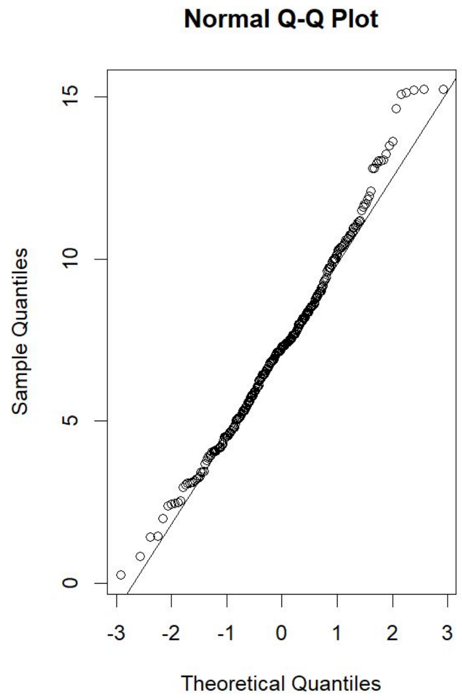
Simple models

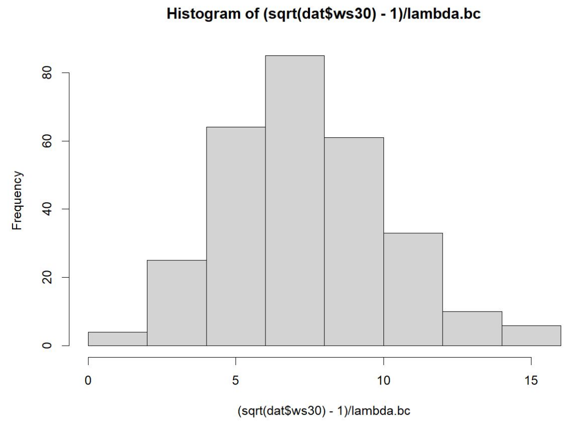
Wind speed: Normal models

Box-Cox Transformation

$$y^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \log(y_i) & \lambda = 0 \end{cases} \quad \longrightarrow \quad \lambda = 0.2621765$$

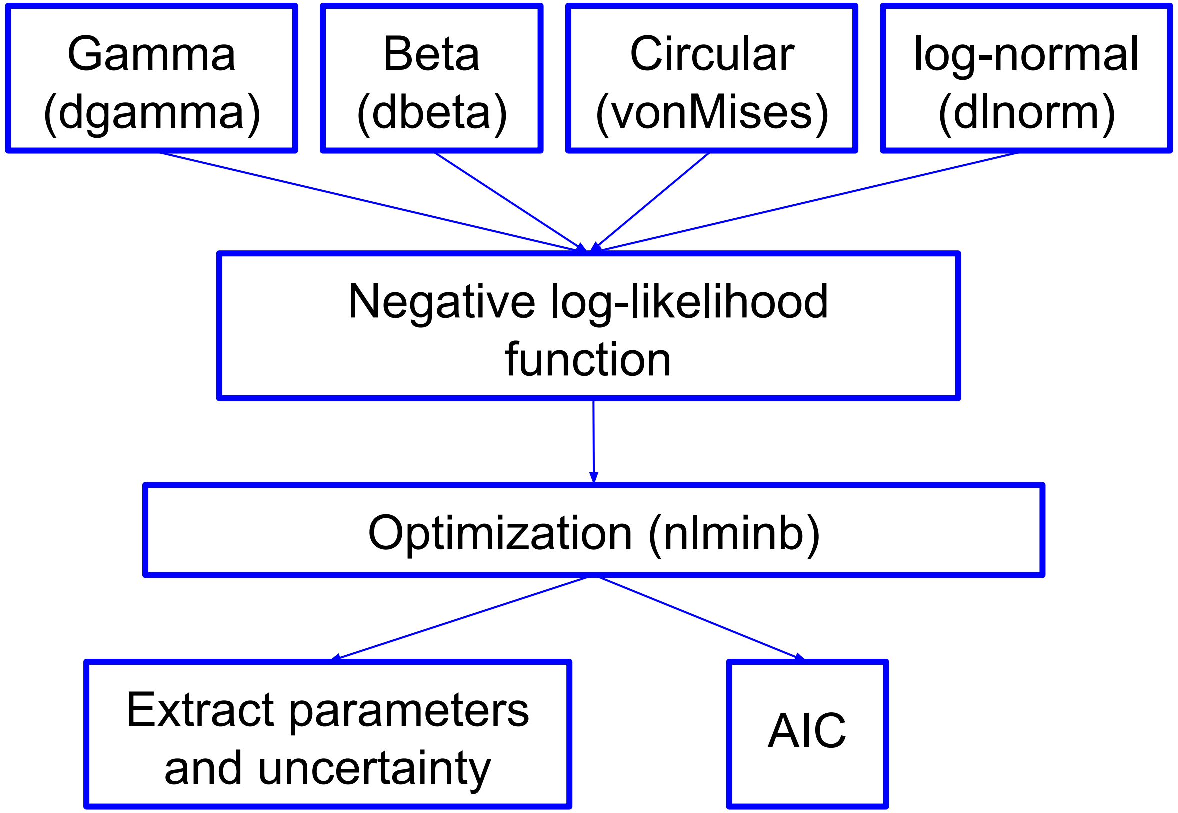




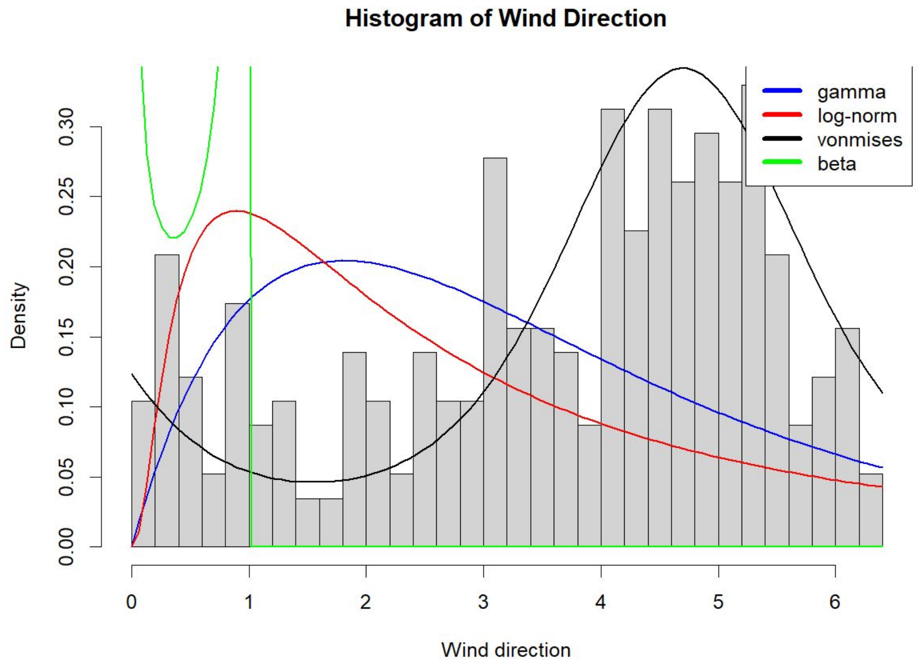


Simple models

Wind direction: non-normal models



Von Mises model	
μ	4.696011
CI μ	[4.474360, 4.917661]
κ	1.000000
CI κ	[0.8059838, 1.1940162]
AIC	1042.121



Von Mises is the most appropriate model for **wind direction**.

Simple models

Wind speed: non-normal models

"'A new circular probability distribution which is based on **GvM (generalization of the von Mises)** is proposed. The new distribution is used to construct a joint probability distribution **which is applied to fit joint distribution of linear and circular variables such as wind speed and wind direction.** The results of several numerical experiments show that compared with the existing distribution models, the new circular distribution and the new constructed joint distribution in the paper can provide higher degree of the fit for the wind data under study'"

(Qin et al., 2010)

www.ccsenet.org/jmr Journal of Mathematics Research Vol. 2, No. 3; August 2010

A New Circular Distribution and Its Application to Wind Data

Xu Qin (Corresponding author)
Faculty of Science and State Key Laboratory for Manufacturing Systems Engineering
Xi'an Jiaotong University, 28 Xian Ning Street, Xi'an 710049, China
Tel: 86-29-8267-3324 E-mail: jiayouxuxu@gmail.com

Jiangshe Zhang
Faculty of Science and State Key Laboratory for Manufacturing Systems Engineering
Xi'an Jiaotong University, 28 Xian Ning Street, Xi'an 710049, China
Tel: 86-29-8266-5961 E-mail: jszhang@mail.xjtu.edu.cn

Xiaodong Yan
Key Laboratory of Regional Climate-Environment Research for Temperature East Asia
Institute of Atmospheric Physics, Chinese Academy of Science
Qi Jia Huo Zi, Beijing 100029, China
Tel: 86-10-8299-5275 E-mail: yxd@tea.ac.cn

GOALS: ASSIGNMENT 2

Formulate model for predicting wind power

1. Consider non-normal models.
2. Present parameters of the final model and their uncertainty.
3. Interpretation of parameters and series expansions.
4. Graphical representation of predictions.

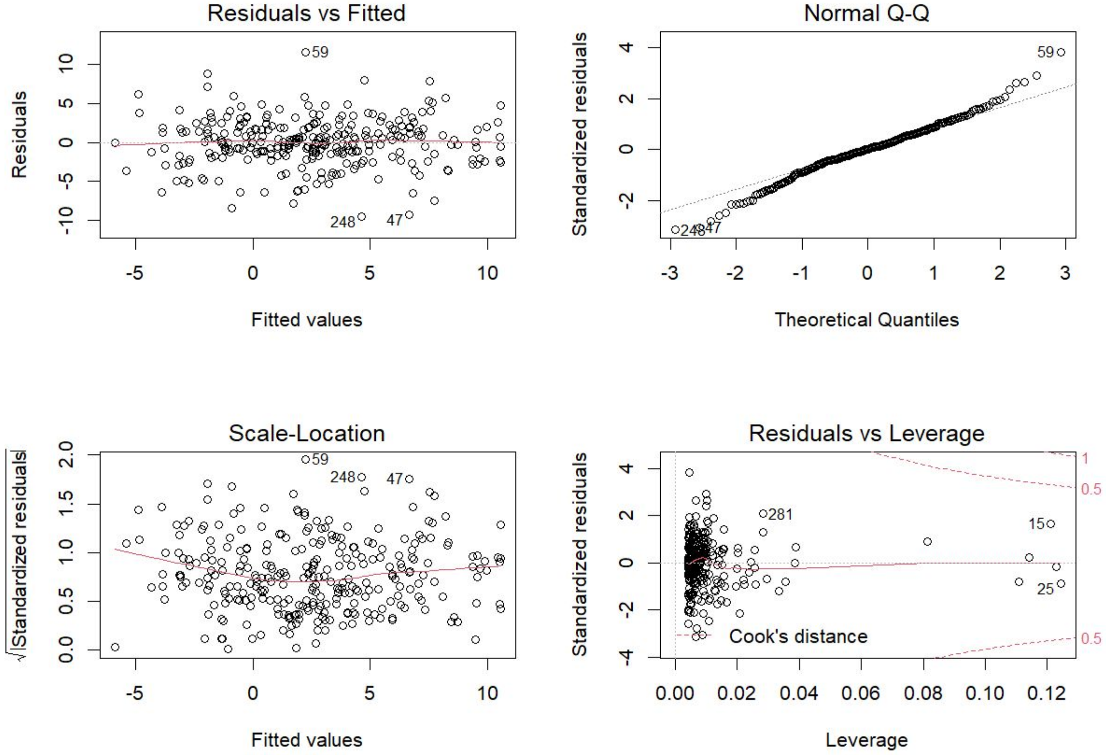
Formulate a model for wind power

Normal model: $$\hat{y}^{(\lambda)} = \beta_0 + \beta_1 ws + \beta_2 ws^2; \quad \lambda = 0.2620668$$

```
lm(formula = y.trans ~ dat$ws30 + I(dat$ws30^2), data = dat,
   family = "Gaussian")
Residuals:
   Min      1Q  Median      3Q     Max
-9.5458 -1.4894  0.0397  1.7834 11.5948

Coefficients:
                 Estimate   Std. Error   t value Pr(>|t|)
(Intercept)     -7.454787   0.735443   -10.136  < 2e-16 ***
dat$ws30         1.410711   0.138145    10.212  < 2e-16 ***
I(dat$ws30^2)   -0.027604   0.005685    -4.855 1.98e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.035 on 285 degrees of freedom
Multiple R-squared:  0.587,    Adjusted R-squared:  0.5841
F-statistic: 202.6 on 2 and 285 DF,  p-value: < 2.2e-16
```



Formulate a model for wind power

Normal model: $\hat{y}^{(\lambda)} = \beta_0 + \beta_1 ws + \beta_2 ws^2; \quad \lambda = 0.2620668$

Series of expansions

$$\left\{ \begin{array}{l} \hat{y}^{(\lambda)} = \beta_0 + \beta_1 ws + \beta_2 ws^2 + \beta_3 ws^3 \\ \hat{y}^{(\lambda)} = \beta_0 + \beta_1 ws + \beta_2 ws^2 + \beta_3 ws^3 + \beta_4 ws^4 \end{array} \right.$$

	df	AIC
Model 1	4	1461.802
Model 2	5	1463.800
Model 3	6	1465.755

Analysis of Variance Table

```
Model 1: y.trans ~ dat$ws30 + I(dat$ws30^2)
Model 2: y.trans ~ dat$ws30 + I(dat$ws30^2) + I(dat$ws30^3)
Model 3: y.trans ~ dat$ws30 + I(dat$ws30^2) + I(dat$ws30^3) + I(dat$ws30^4)
  Res.Df    RSS   Df Sum of Sq Pr(>Chi)
1    285  2625.4
2    284  2625.4  1   0.01904   0.9639
3    283  2625.0  1   0.41022   0.8334
```

Model 1	Estimate	CI
β_0	-7.45478717	[-8.90237687, -6.00719746]
β_1	1.41071121	[1.13879808, 1.68262433]
β_2	-0.02760426	[-0.03879467, -0.01641385]

Formulate a model for wind power

Normal model:

Including variable Wind direction

Significative p-value when chisq test

$$\hat{y}^{(\lambda)} = \beta_0 + \beta_1 ws + \beta_2 ws^2 + \beta_3 wd$$

```
Call:
lm(formula = y.trans ~ dat$ws30 + I(dat$ws30^2) + dat$wd30,
data = dat,
   family = "Gaussian")

Residuals:
    Min      1Q  Median      3Q     Max
-9.5542 -1.4635  0.0329  1.7768 11.5957

Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    -7.501578   0.796976  -9.413  < 2e-16 ***
dat$ws30        1.409471   0.138616  10.168  < 2e-16 ***
I(dat$ws30^2) -0.027616   0.005696  -4.849  2.05e-06 ***
dat$wd30        0.016457   0.106935   0.154   0.878
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.04 on 284 degrees of freedom
Multiple R-squared:  0.5871,  Adjusted R-squared:  0.5827
F-statistic: 134.6 on 3 and 284 DF,  p-value: < 2.2e-16
```

AIC = 1463.778

$$\hat{y}^{(\lambda)} = \beta_0 + \beta_1 ws + \beta_2 ws^2 + \beta_3 wd + \beta_4 wd^2$$

```
Call:
lm(formula = y.trans ~ dat$ws30 + I(dat$ws30^2) + dat$wd30 +
   I(dat$wd30^2), data = dat, family = "Gaussian")

Residuals:
    Min      1Q  Median      3Q     Max
-9.8995 -1.6671  0.0294  1.7429 10.9922

Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    -8.648971   0.867302  -9.972  < 2e-16 ***
dat$ws30        1.382843   0.136811  10.108  < 2e-16 ***
I(dat$ws30^2) -0.026749   0.005617  -4.762  3.07e-06 ***
dat$wd30        1.250221   0.410146   3.048   0.00252 **
I(dat$wd30^2) -0.197036   0.063304  -3.113   0.00204 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.995 on 283 degrees of freedom
Multiple R-squared:  0.6007,  Adjusted R-squared:  0.5951
F-statistic: 106.5 on 4 and 283 DF,  p-value: < 2.2e-16
```

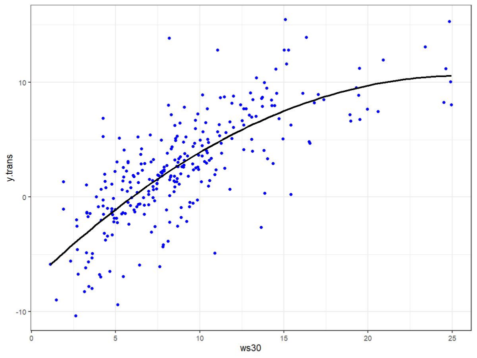
AIC = 1456.084

Formulate a model for wind power

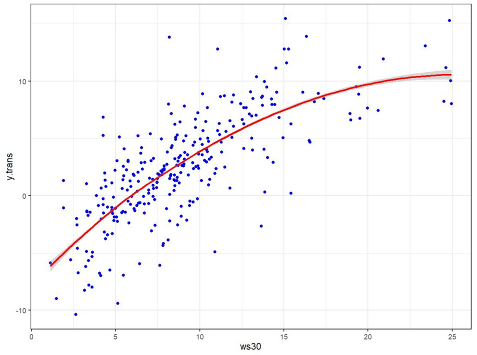
Normal model: Parameter uncertainty of the two best normal models

	Model 1 (without wd)	Model 2 (with wd)
β0	-7. 45478717	-8.6489713
CI β0	[-8.90237687, -6.00719746]	[-10.35615292, -6.94178972]
β1	1.41071121	1.3828433
CI β1	[1.13879808, 1.68262433]	[1.11354616, 1.65214036]
β2	-0.02760426	-0.0267491
CI β2	[-0.03879467, -0.01641385]	[-0.03780599, -0.01569222]
β3	-	1.2502209
CI β3	-	[0.44289656, 2.05754528]
B4	-	-0.1970357
CI β4	-	[-0.32164310, -0.07242825]

$$\hat{y}^{(\lambda)} = \beta_0 + \beta_1 ws + \beta_2 ws^2; \quad \lambda = 0.2620668$$



$$\hat{y}^{(\lambda)} = \beta_0 + \beta_1 ws + \beta_2 ws^2 + \beta_3 wd + \beta_4 wd^2$$



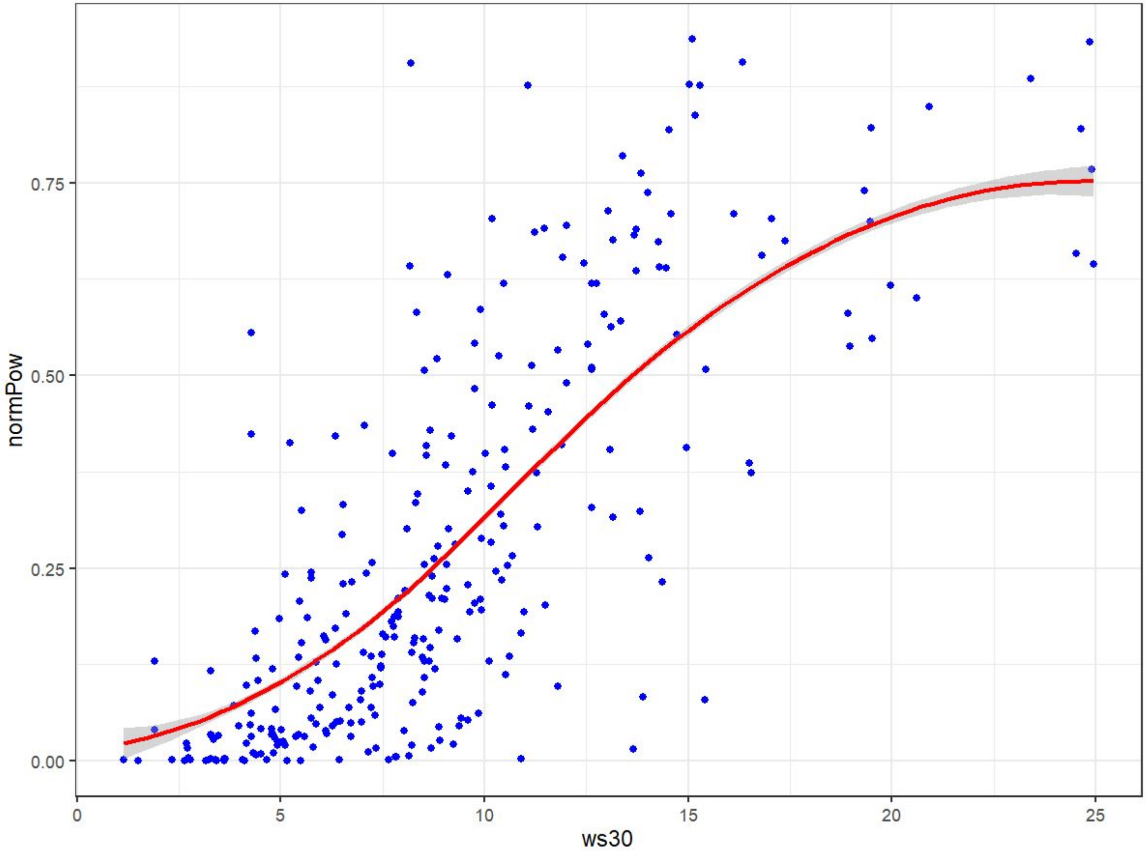
Formulate a model for wind power

Non-normal model: Beta regression

	Estimate value	CI
β0	-4.418622074	[-4.98896968, -3.84827447]
β1	0.402275796	[0.31733284, 0.48721875]
β2	-0.007969534	[-0.01135872, -0.00458035]
β3	0.349155870	[0.10717155, 0.59114020]
B4	-0.052658876	[-0.08983082, -0.01548693]
Φ	5.298284	[4.42934017, 6.16722725]

Best model! AIC = -484.3728

$$y = \beta_0 + \beta_1 ws + \beta_2 ws^2 + \beta_3 wd + \beta_4 wd^2$$



GOALS: ASSIGNMENT 3

Analysis of autocorrelation AR(1)

Wind power as a time series

AR(1) model: $\varepsilon_i = \phi\varepsilon_{i-1} + u_i\,;\quad u_i \sim N\left(0,\ \sigma_u^{\,2}\right)$

$$Y^{(0.2)} = \beta_0 + \beta_1 ws + \beta_2 ws^2 + \varepsilon \quad ; \varepsilon \sim N\left(0,\ \sigma^2\right)$$

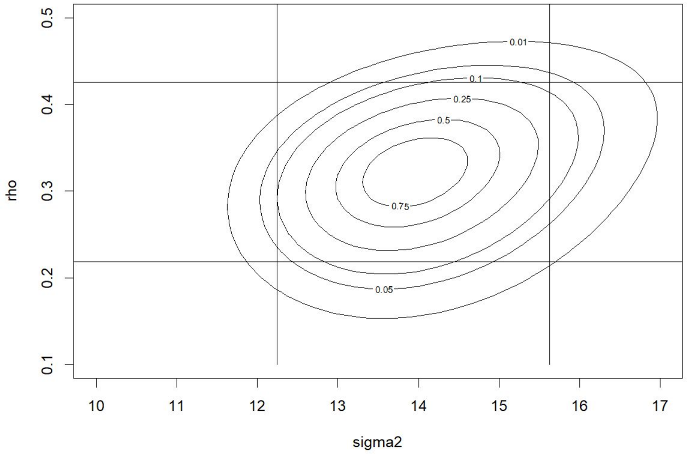
Extract residuals

$$\varepsilon = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 \\ . & . \\ \varepsilon_{n-1} & \varepsilon_n \end{bmatrix}$$

Fit to the model

$$\left[\varepsilon_i,\ \varepsilon_{i+1}\right]^T \sim N(0,\ \Sigma)\,; \quad \Sigma = \sigma^2 \begin{bmatrix} 1 & p \\ p & 1 \end{bmatrix}$$

Parameter	Estimate value	CI
σ^2	13.93593	[12.24160 , 15.63025]
p	0.3222538	[0.2185932, 0.4259144]



Hypothesis testing: $H_0 : \rho = 0$

- Wilk's likelihood ratio statistic

$W = 2\log\left(\frac{L(\rho_0)}{L(\hat{\rho})}\right) \rightarrow X^2 \rightarrow$ p-value =1.468168e-07

- Wald test

$z = \frac{\hat{\rho} - \rho_0}{se(\hat{\rho})} \rightarrow$ N(0,1) $\rightarrow$ p-value = 9.658089e-09

Reject null hypothesis

Wind power as a time series

$$[\varepsilon_i, \varepsilon_{i+1}]^T \sim N(0, \Sigma); \quad \Sigma = \sigma^2 \begin{bmatrix} 1 & p \\ p & 1 \end{bmatrix}$$

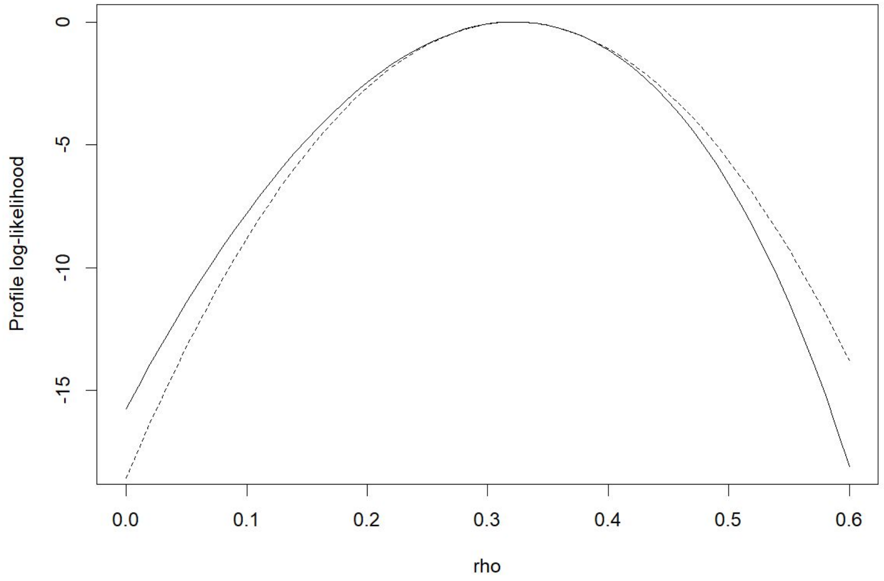
Analytical method

$$I(\sigma^2, \rho) = \begin{pmatrix} \frac{n}{\sigma^4} & -\frac{n\rho}{\sigma^2(1-\rho^2)} \\ -\frac{n\rho}{\sigma^2(1-\rho^2)} & \frac{n(1+\rho^2)}{(1-\rho^2)^2} \end{pmatrix} \rightarrow \begin{bmatrix} 1.477781 & -7.405631 \\ -7.405631 & 394.481927 \end{bmatrix}$$

Numerical method

Hessian $\rightarrow \begin{bmatrix} 1.477192 & -7.407708 \\ -7.407708 & 394.641620 \end{bmatrix}$

Profile likelihood and quadratic approximation



Wind power as a time series

AR(1) model: $\varepsilon_i = \phi \varepsilon_{i-1} + u_i\,;\;\; u_i \sim N\left(0,\, \sigma_u^{\,2}\right)$

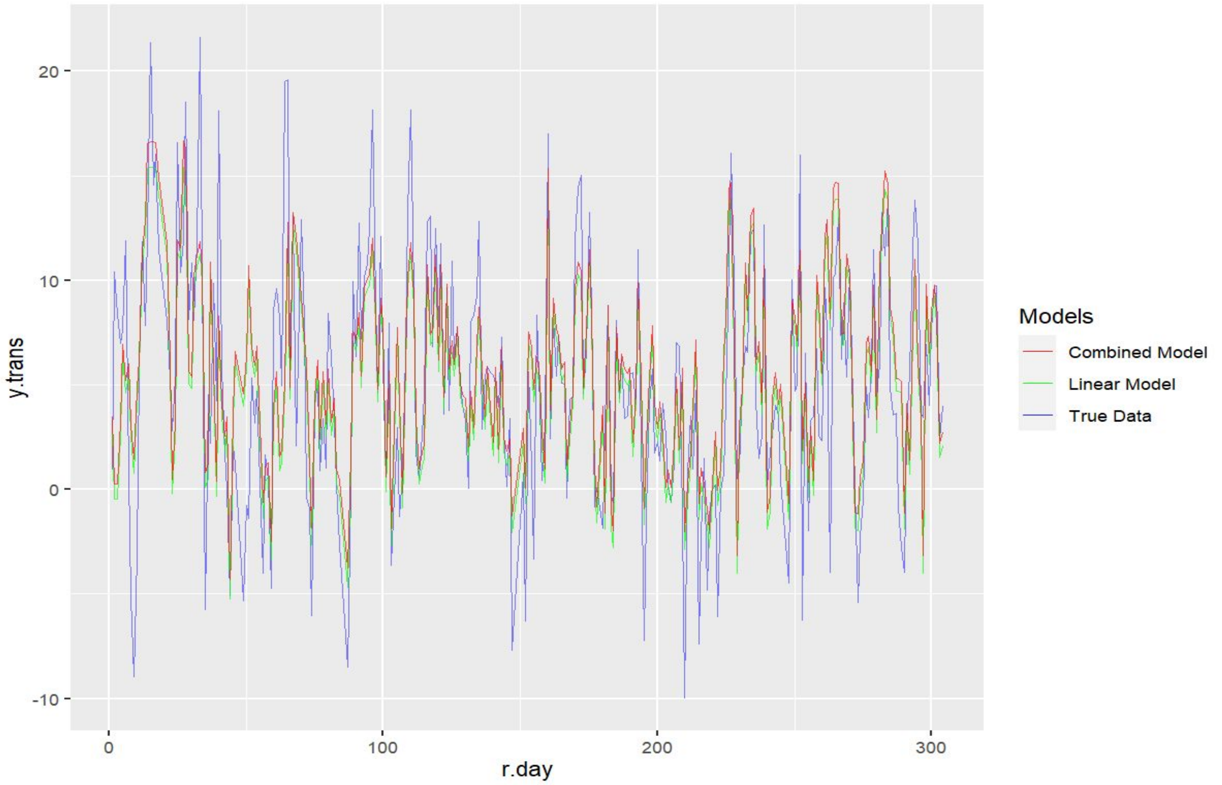
```
Call:
arima(x = y.trans, order = c(1, 0, 0), xreg = xreg)

Coefficients:
         ar1     intercept    ws30     ws30sq
      0.3252    -6.7491     1.6170   -0.0282
s.e.  0.0559     0.9351     0.1731    0.0074

sigma^2 estimated as 12.44:  log likelihood = -771.69,  aic = 1553.39
                2.5 %          97.5 %
ar1          0.21558497     0.4347553
Intercept   -8.58184971    -4.9163993
ws30         1.27766793     1.9563241
ws30sq      -0.04262742    -0.0138018
```

AIC linear model: 1583.305
AIC combined model: 1553.390

Better fit than linear model!

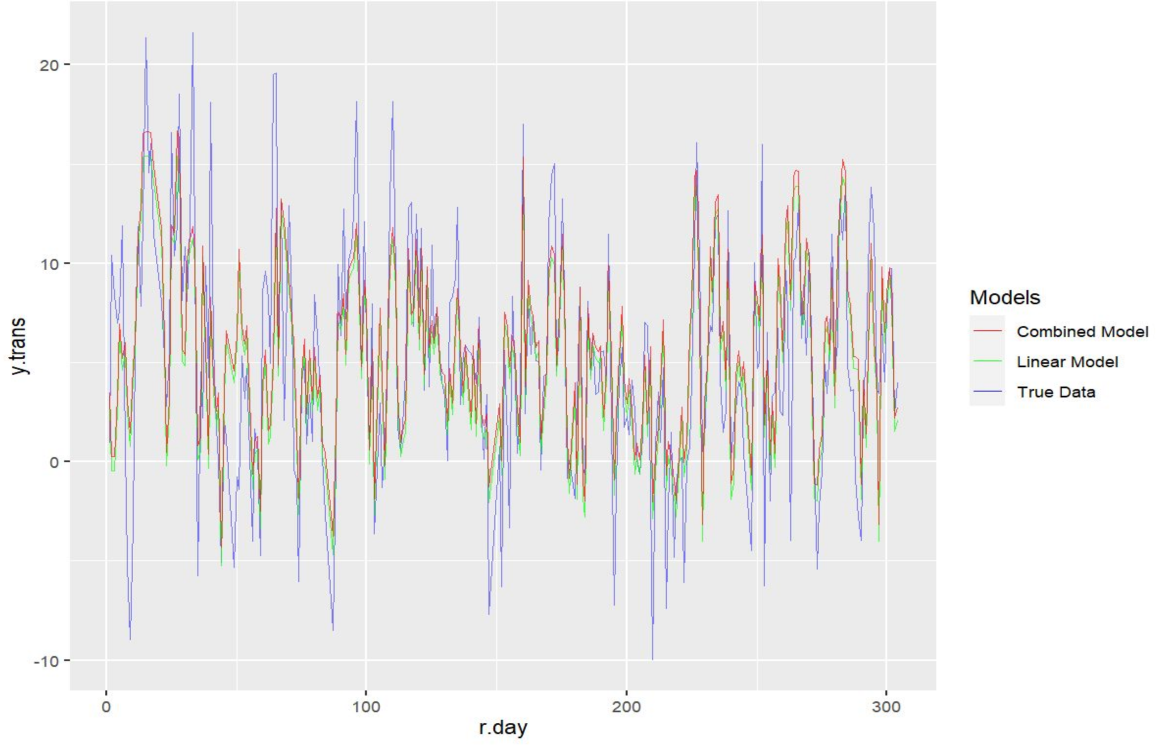


Wind power as a time series

AR(1) model: $\varepsilon_i = \phi\varepsilon_{i-1} + u_i\,;\;\; u_i \sim N\left(0,\, \sigma_u{}^2\right)$

MAE (Mean Absolute Error)	Linear model	Combined model
Long term	2.796916	2.848373
Short term (last 3 days)	10.85968	9.804569

AR(1) model more suitable for short term and Linear model more suitable for long term.



References

Pawitan Y. In All Likelihood: Statistical Modelling and Inference Using Likelihood. OUP Oxford; 2001. (Oxford science publications)

Code for the project can be found at [Statistical Modelling]

DTU