STATISTICAL MODELLING: Theory and practice

Project 1: Wind power data

GOALS: ASSIGNMENT 1

- **Descriptive statistics**
- **Compute simple models**

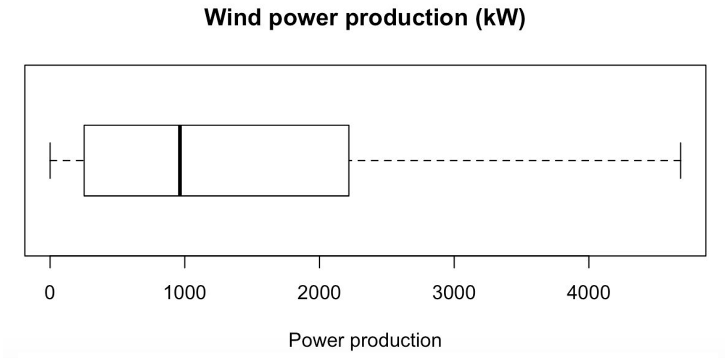
1. **Fit** different **probability density models** to wind power, wind speed and wind direction data.
2. Select **better model** for each variable.
3. **Report parameters** including assessment of their **uncertainty**.

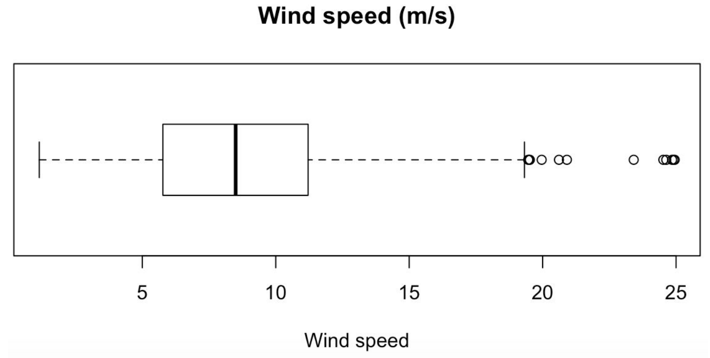
Descriptive statistics

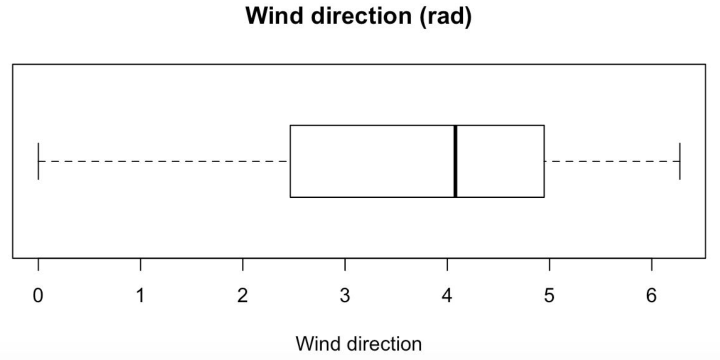
	Power production
Min	0.123
1st Qu	254.158
Median	964.123
Mean	1381.196
3rd Qu	2196.579
Max	4681.062

	Wind speed
Min	1.139
1st Qu	5.779
Median	8.498
Mean	9.112
3rd Qu	11.202
Max	24.950

	Wind Direction
Min	0.000095
1st Qu	2.474999
Median	4.079297
Mean	3.602390
3rd Qu	4.945443
Max	6.274642

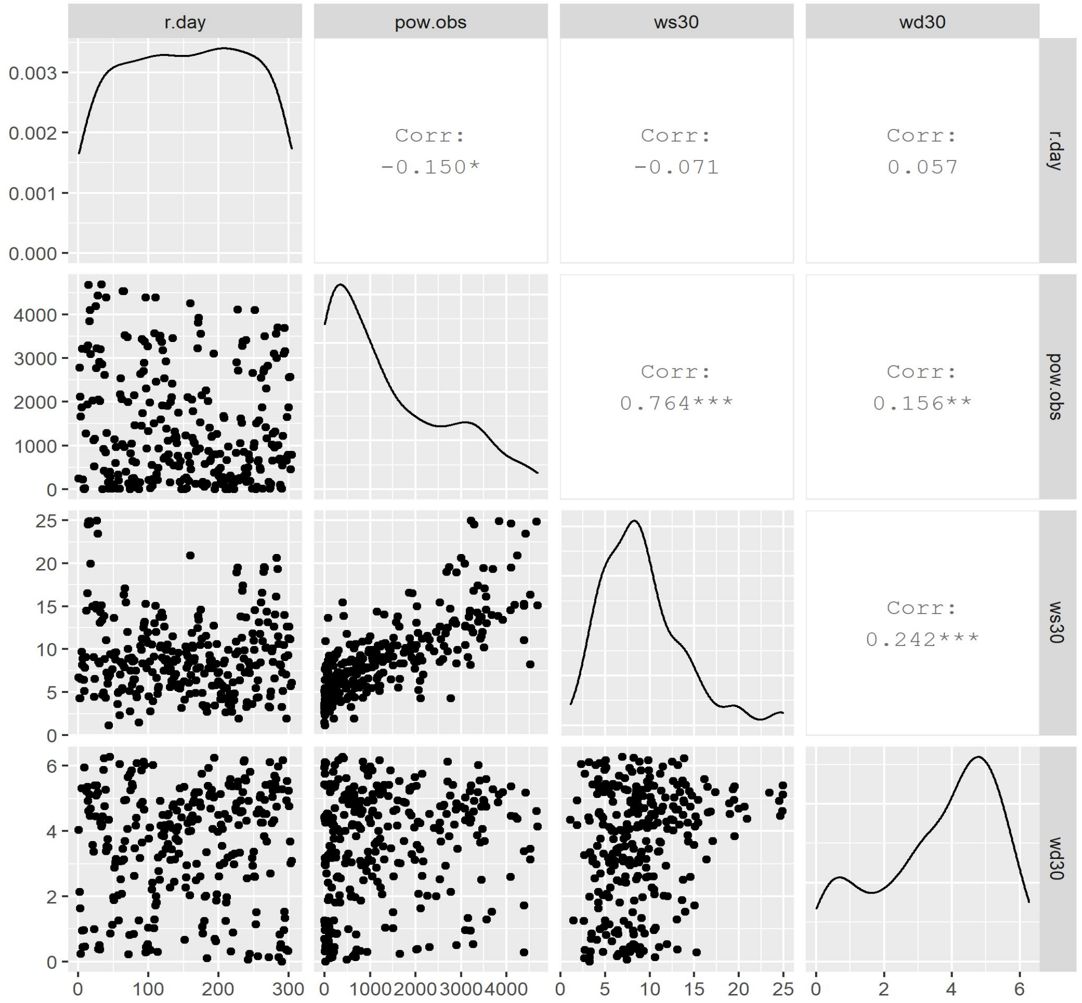




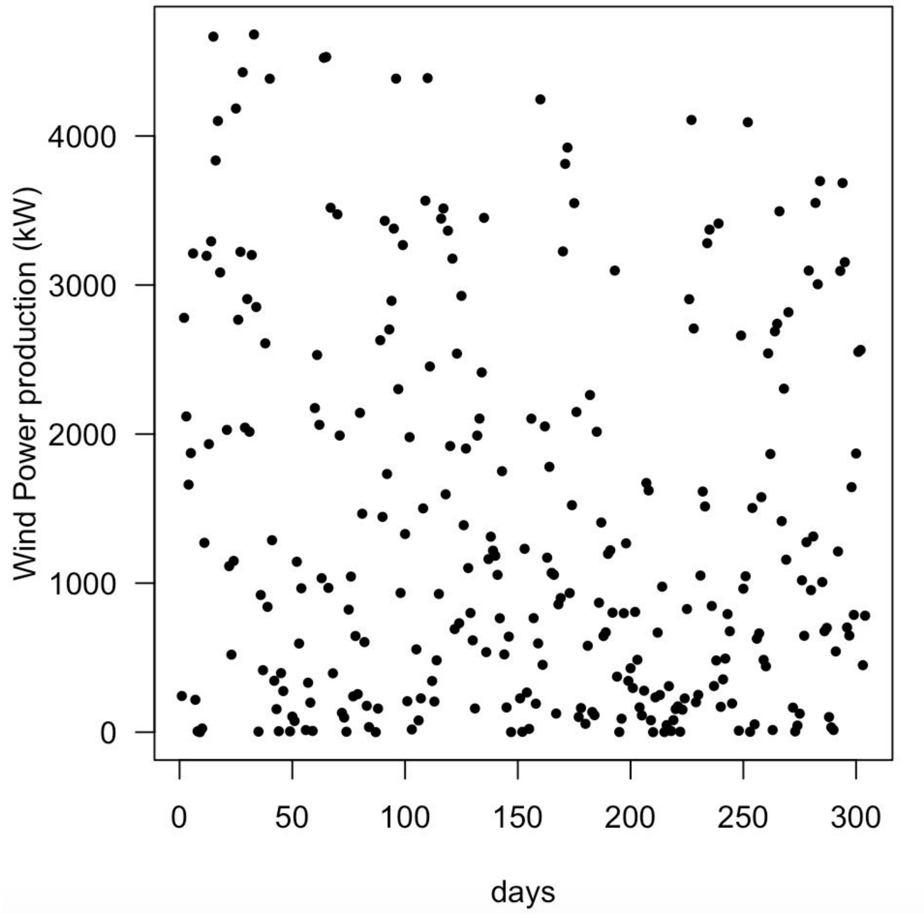


Descriptive statistics

Data correlation and variable frequencies



Wind power production over time



Normalization and data distribution

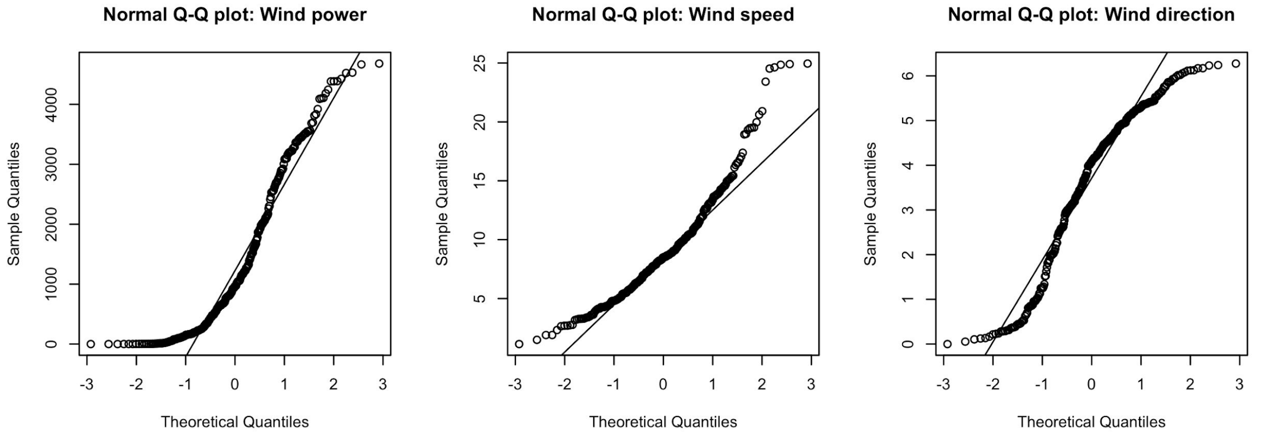
Normalization of power production:

$$power^* = \frac{power - min(power)}{max(power) - min(power)}$$

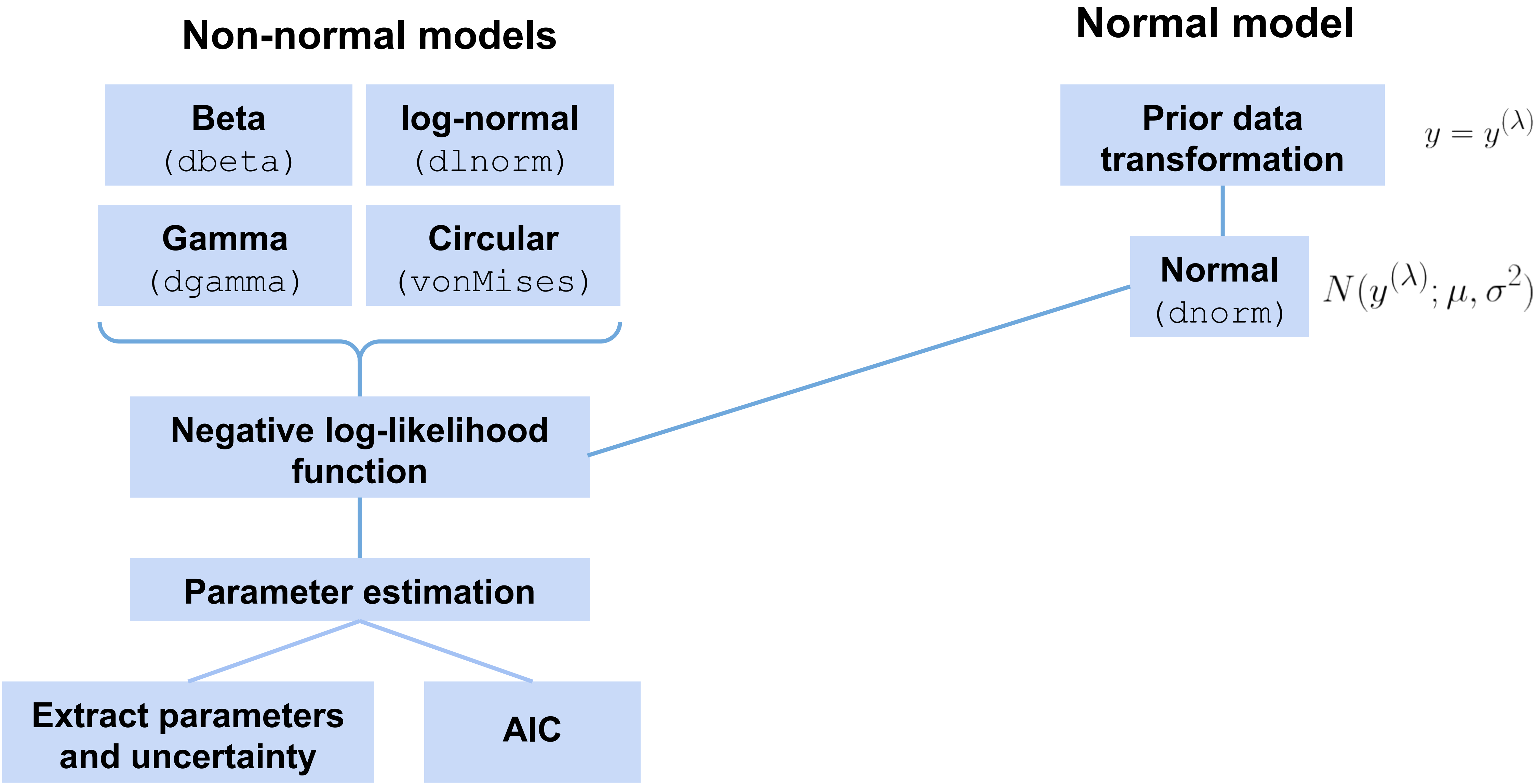
The normalization is based on the installed capacity, which maximum is 5000 kW.

Values that from 0 to 1 for power production.

Checking for normality of the data:



Modelling strategy



Data transformations

Box-Cox Transformation

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

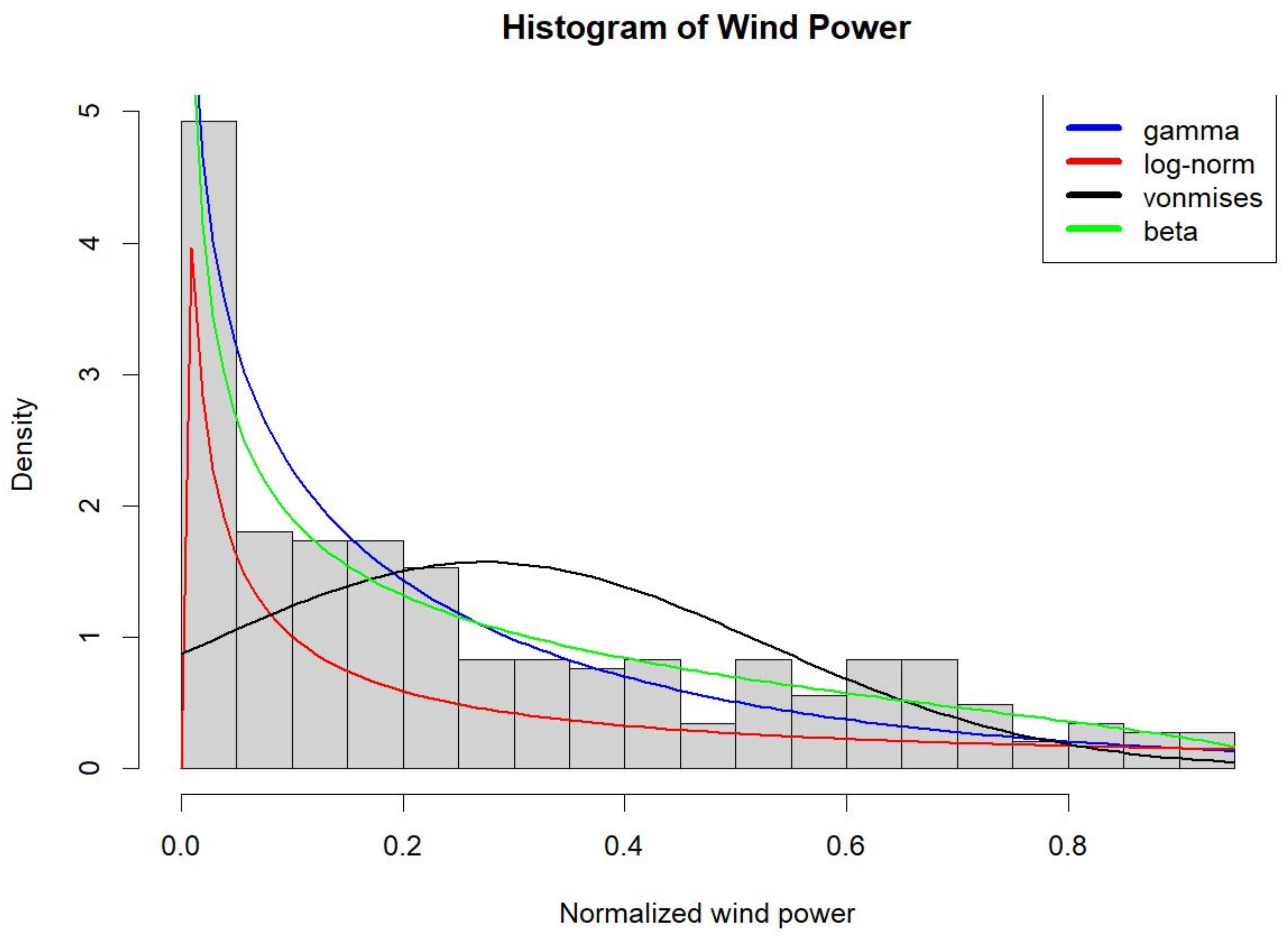
Transformation 1

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

Transformation 2

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

POWER PRODUCTION:
Non-normal models



Model	Parameter 1	CI	Parameter 2	CI
Gamma	0.6926	[0.6246 , 0,7606]	2.5073	[2.1594 , 2.8553]
Beta	0.5571	[0.4952 , 0.6189]	1.4918	[1.2863 , 1.6973]
VonMises	0.2738	[0.2443 , 0.3034]	15.7607	[13.2307 , 18.2907]
Log-normal	0.00	[-0.3335 , 0.3335]	2.8880	[2.6522 , 3.1239]

Model	AIC
Gamma	-190.7635
Beta	-239.3236
VonMises	36.7573
Log-normal	187.8728

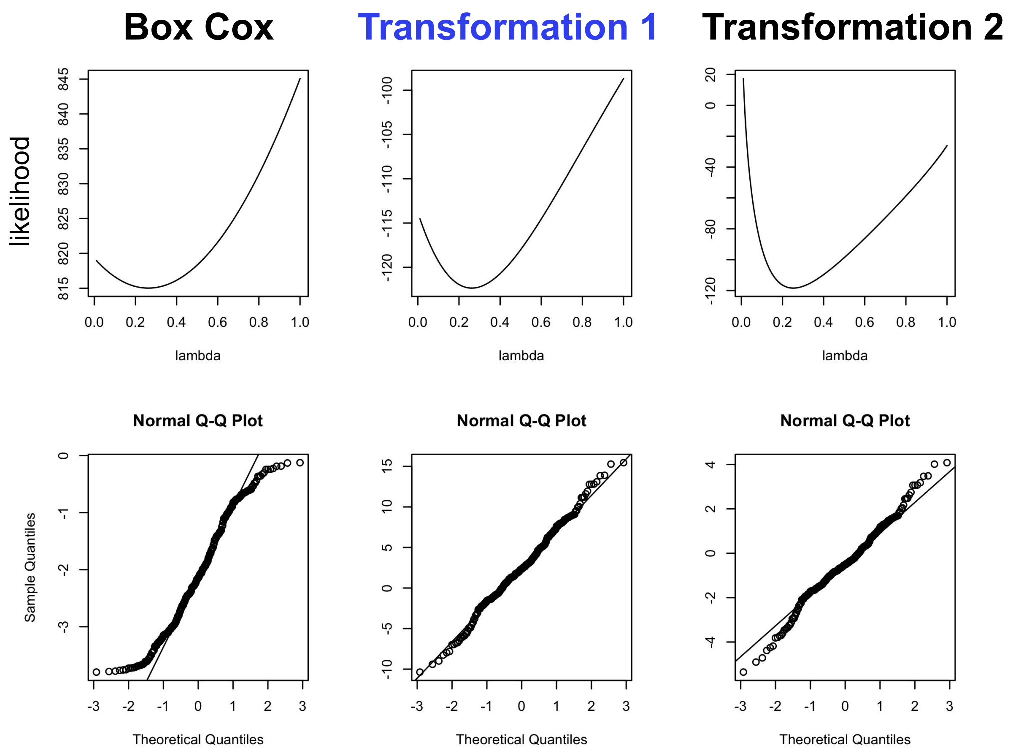
$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

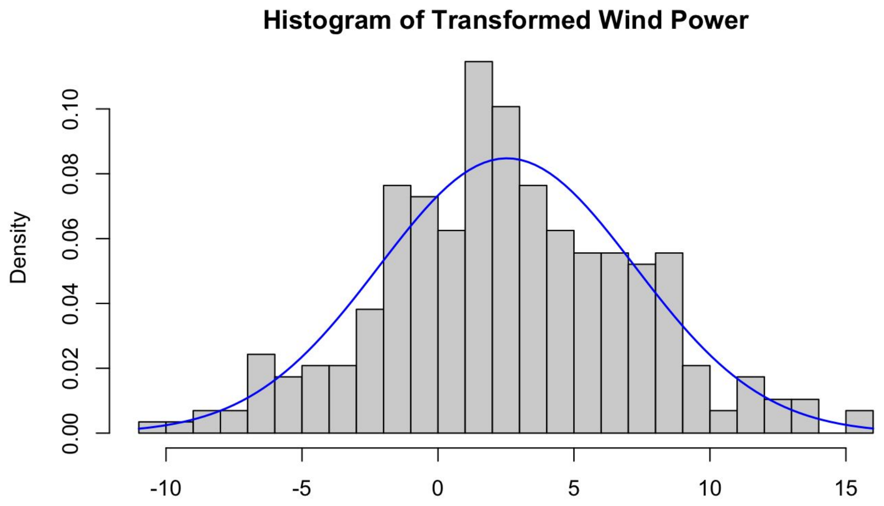
POWER PRODUCTION:
Non-normal models

Transformation	λ
Boxcox	0.3467
1	0.2620
2	0.2523

Profile likelihood :

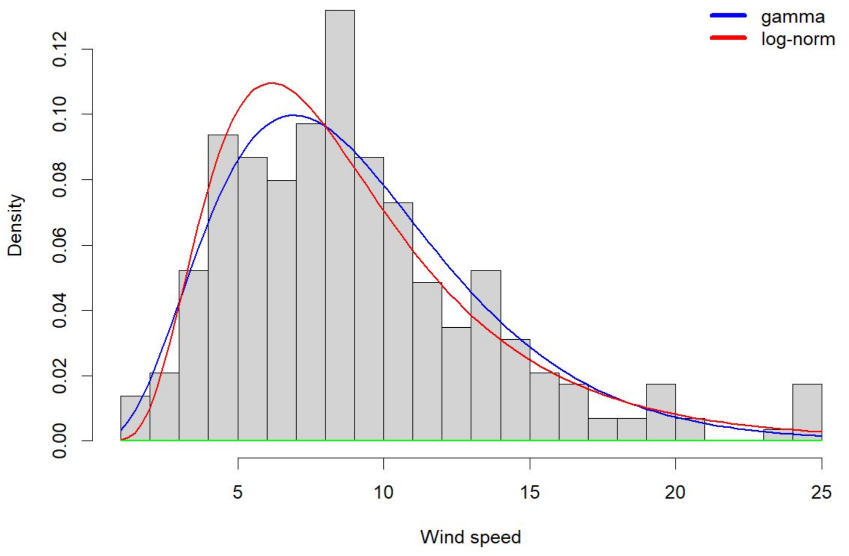
$$\log L(\lambda, \mu, \sigma^2) = -\frac{1}{2}\log\sigma^2 - \frac{(y_\lambda - \mu)^2}{2\sigma^2} + (\lambda - 1)\log y$$





AIC = 8.94

WIND SPEED:
Non-normal models

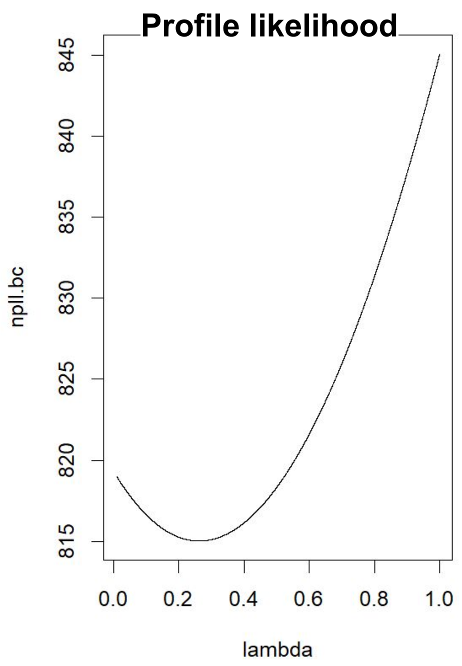


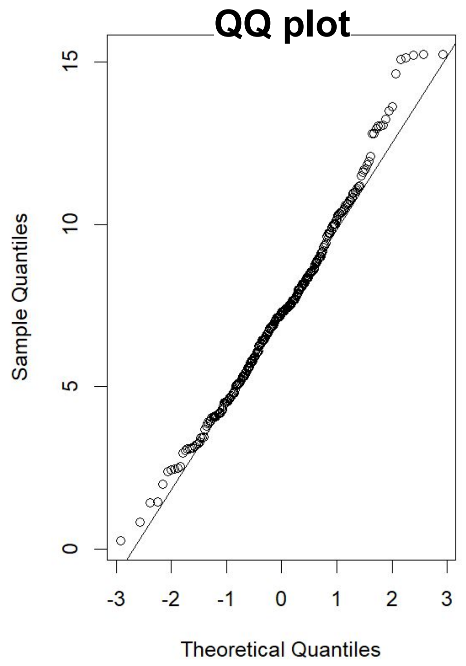
Log norm model	
μ	2.0844793
CI μ	[2.024706, 2.144253]
σ	0.5175574
CI σ	[0.4752910, 0.5598238]

AIC = 1642

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

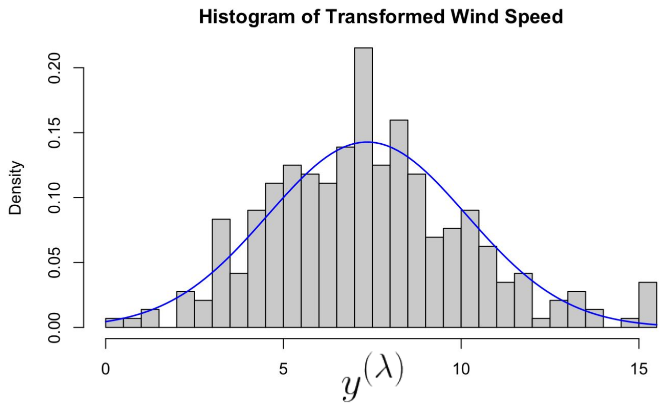
Box-Cox Transformation





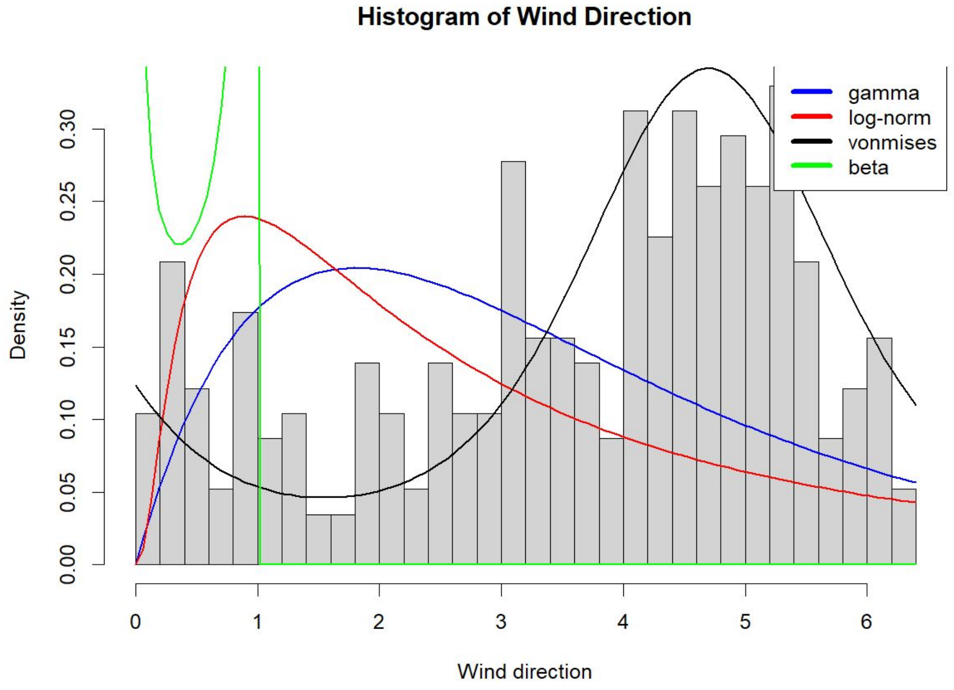
Lambda estimation

$\lambda = 0.2621765$



AIC = 7.89

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

GOALS: ASSIGNMENT 2

Formulate model for predicting wind power

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

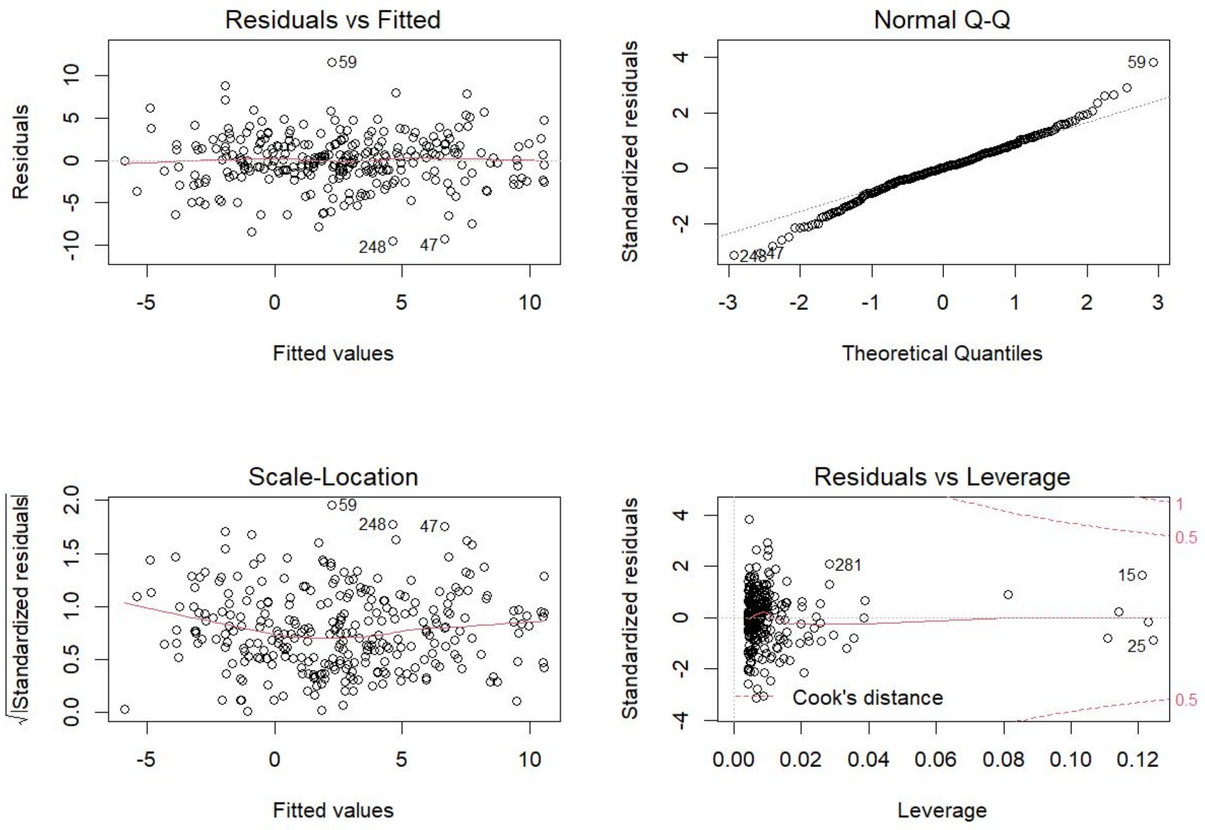
Formulate a normal model for wind power

Normal model:

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

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
β_0	-7.454787	0.735443	-10.136	< 2e-16	***
β_1	1.410711	0.138145	10.212	< 2e-16	***
β_2	-0.027604	0.005685	-4.855	1.98e-06	***



Formulate a model for wind power

Normal model: $\lambda = 0.2620668$

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

Series of expansions

(2) $\hat{y}^{(\lambda)} = \beta_0 + \beta_1 ws + \beta_2 ws^2 + \beta_3 ws^3$

(3) $\hat{y}^{(\lambda)} = \beta_0 + \beta_1 ws + \beta_2 ws^2 + \beta_3 ws^3 + \beta_4 ws^4$

Model	AIC
1	1461.802
2	1463.800
3	1465.755

Analysis of Variance Table

	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

Best model parameters (model (1))

Parameter	Estimate	CI
β_0	-7.4547	[-8.9023 , -6.0023]
β_1	1.4110	[1.1387, 1.6826]
β_2	-0.0276	[-0.0387 , - 0.0164]

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
```

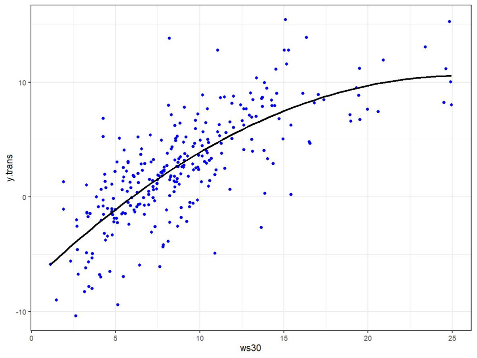
AIC = 1456.084

Formulate a model for wind power

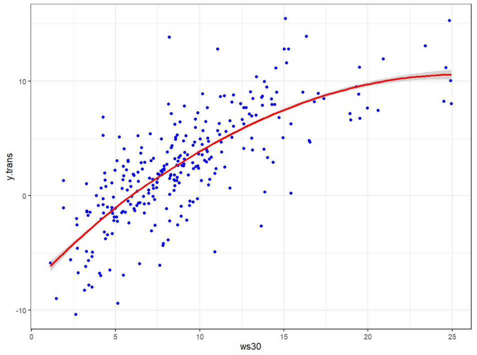
Normal model: Parameter uncertainty of the two best normal models

	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]

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



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

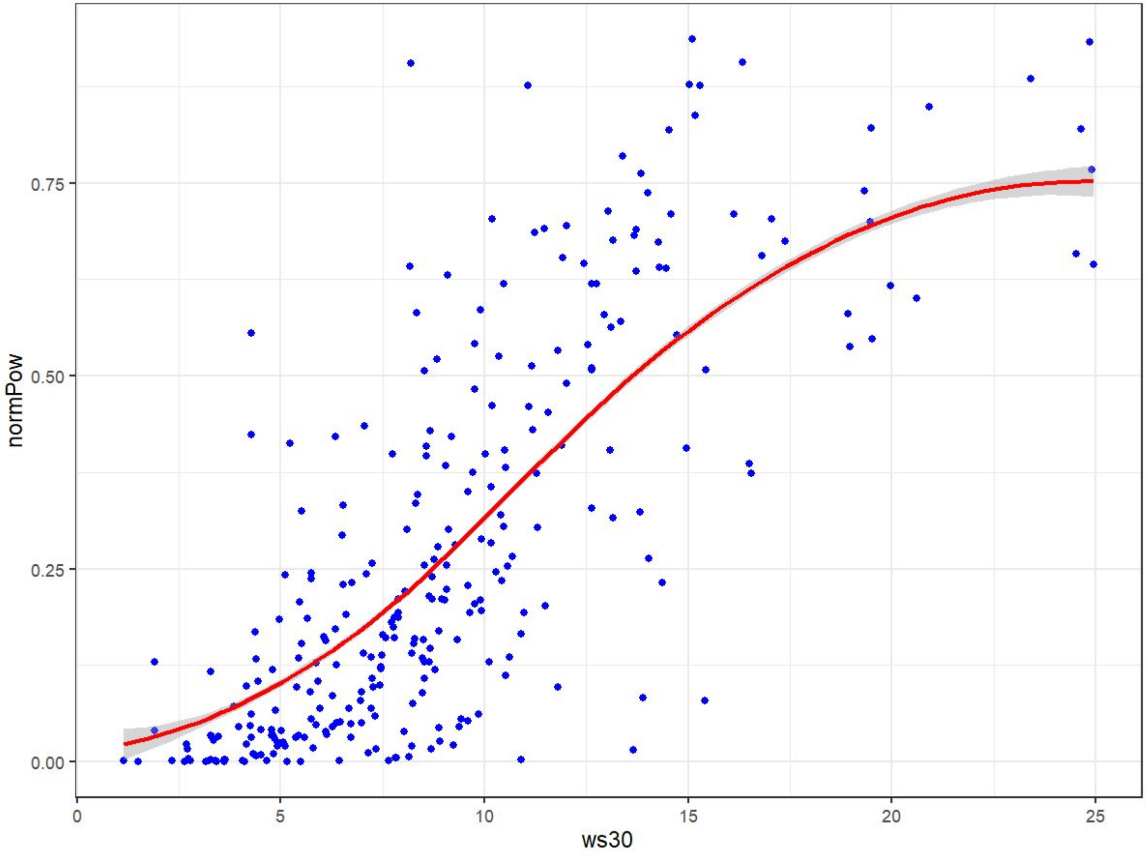


Formulate a model for wind power

Non-normal model: Beta regression

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

	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

GOALS: ASSIGNMENT 3

Analysis of autocorrelation AR(1)

- **Extract the residuals** from the previous model
- **Fit parameters** in the model $[e_i , e_{i+1}]$ T ~ N(0,Σ) and report:
    - Parameter **estimates and Wald intervals**
    - **Contour plot** of the likelihood
    - **Likelihood ratio test and wald test** (H_0: p = 0)

- Compare the **I(σ^2 ,ρ) calculated by numerical** methods with the algebraic form
- Estimate the **parameters of the AR(1)**
- Discuss the effect of **short and long term on AR(1)**

Wind power as a time series

Are our observations iid ?

Extract the residuals

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

Extract residuals

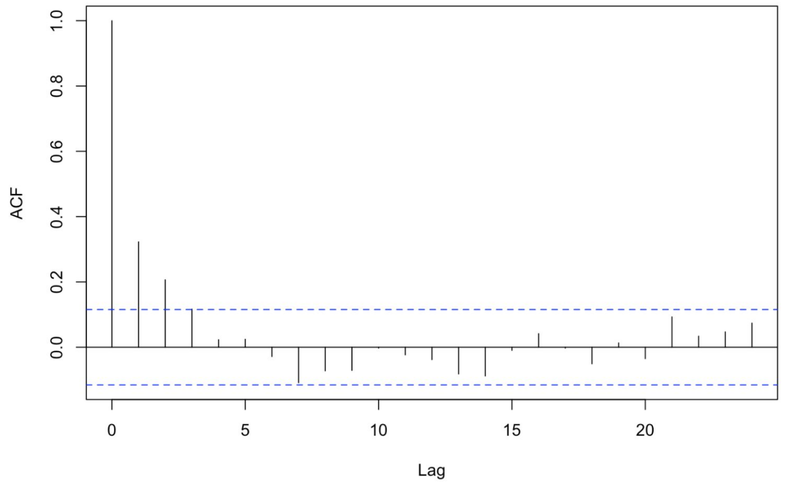
$$\varepsilon = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 \\ . & . \\ \varepsilon_{n-1} & \varepsilon_n \end{bmatrix}$$

Fit to the model

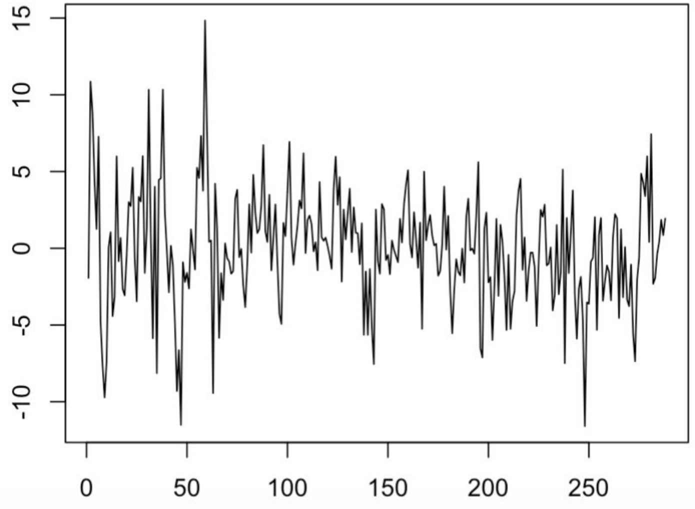
$$\left[\varepsilon_i,\ \varepsilon_{i+1}\right]^T \sim N(0,\ \Sigma)\ ; \quad \Sigma = \sigma^2 \begin{bmatrix} 1 & p \\ p & 1 \end{bmatrix}$$

***p* is the covariance between $[e_i; e_{i+1}]$**

Autocorrelation function plot



Residuals



Wind power as a time series

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

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

Extract residuals

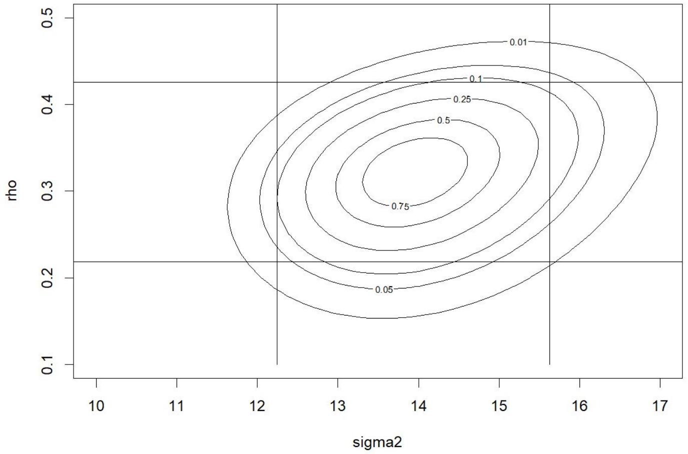
$$\varepsilon = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 \\ . & . \\ \varepsilon_{n-1} & \varepsilon_n \end{bmatrix}$$

Fit to the model

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

AR model estimates

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

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

- Wald test

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

Reject null hypothesis

There is a correlation between $[e_i\,,\, e_{i+1}]$

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

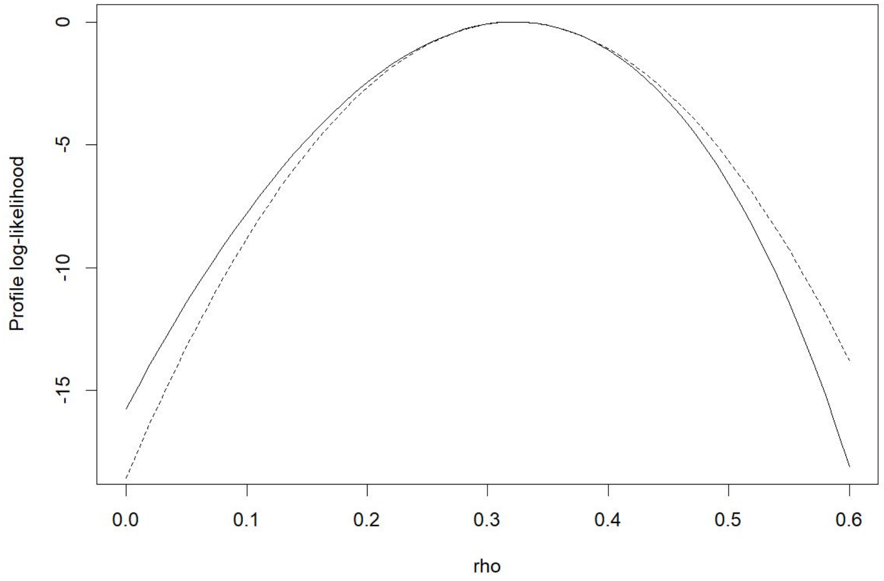
$$I(\sigma^2, \rho) = \begin{pmatrix} \frac{n}{\sigma^4} & -\frac{n\rho}{\sigma^2(1-\rho^2)} \\ -\frac{n\rho}{\sigma^2(1-\rho^2)} & \frac{n(1+\rho^2)}{(1-\rho^2)^2} \end{pmatrix} \rightarrow \begin{bmatrix} 1.477781 & -7.405631 \\ -7.405631 & 394.481927 \end{bmatrix}$$

Numerical method

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

Profile likelihood and quadratic approximation

$$l(\theta) - l(\hat{\theta}) \approx -\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2$$



Wind power as a time series

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

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

Combined model: $y^{\lambda} = \beta_0 + \beta_1 ws + \beta_2 ws2 + \varepsilon_{AR}$

ARIMA(1,0,0) = first-order autoregressive model

```
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

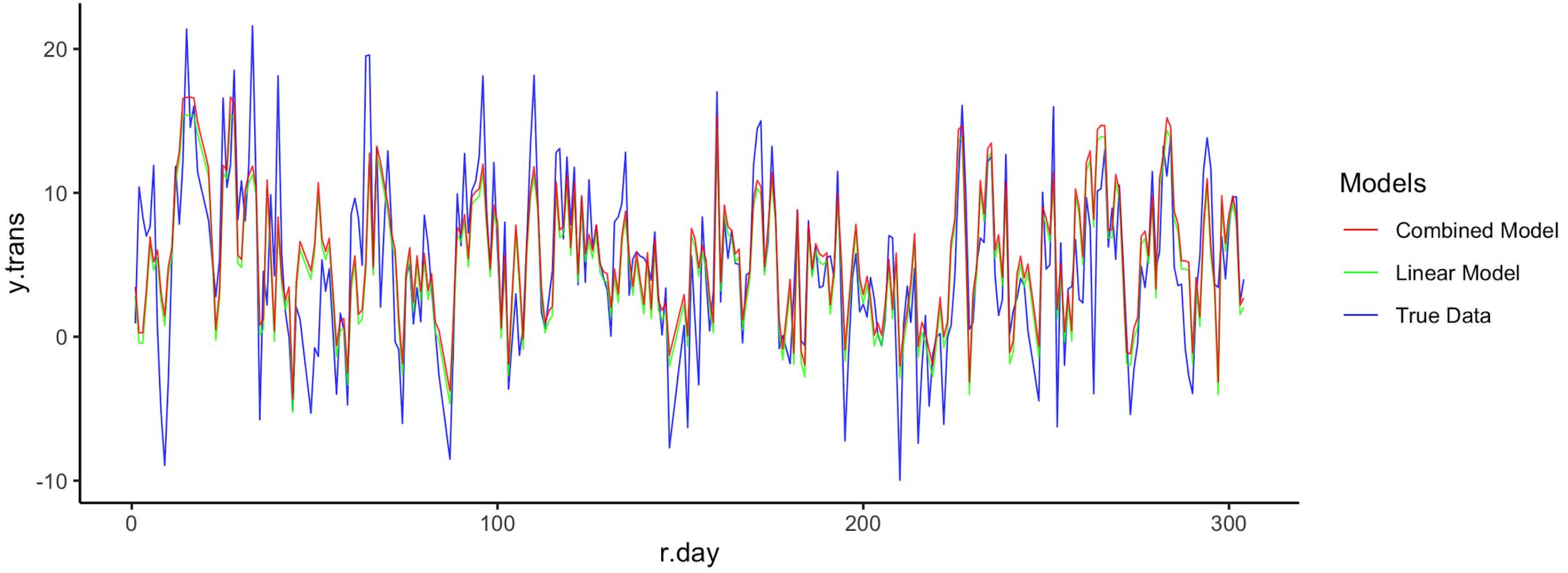
Better fit than linear model!

Wind power as a time series

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

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

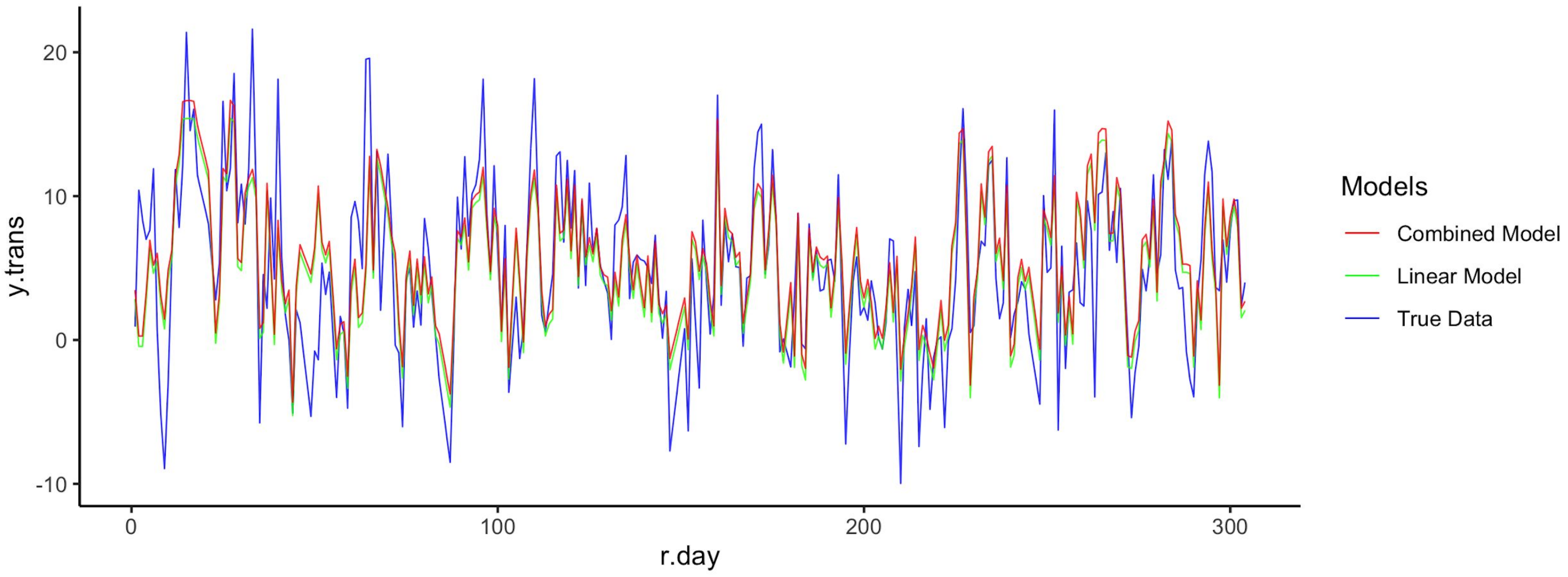
Combined model: $y^{\lambda} = \beta_0 + \beta_1 ws + \beta_2 ws2 + \varepsilon_{AR}$



Wind power as a time series

Mean Absolute error	Linear model	Combined model
Long term	2.7969	2.8448
Short term (3 days)	10.8596	9.8045

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