NHH

TRIAL EXAMINATION
TECH3

Spring, 2025

Date: April 08. 2025

Time: 09:00-13:00

Number of hours: 4

An invigilator can contact course responsible by phone: +47 xxx xx xxx

SUPPORT MATERIALS PERMITTED DURING THE EXAMINATION:

Calculator Yes ☑ No ☐

Dictionary: one bilingual dictionary permitted.

List materials permitted or write: No other support materials permitted.

Number of pages, including front page: 8

Problem 1:

Below is a list of statements. Give a **short** explanation why each are wrong.

a) A probability only has to be positive.

b) The Central Limit Theorem states that if n becomes large, any variable X_n will be asymptotically normally distributed.

c) If X and Y are correlated, X causes Y.

d) An estimator is consistent, if its variance approaches zero when the sample size becomes large.

e) When doing hypothesis testing, the significance level should always be 5%.

f) The code below is a Monte Carlo simulation.

```
from sklearn.model_selection import KFold
from statsmodels.formula.api import ols
import numpy as np
import pandas as pd

# Initiate 10-folds function:
kf = KFold(n_splits=10, shuffle=True, random_state=42)

# Initialize the prediction column
height['Predicted'] = 0.0

for train_index, test_index in kf.split(height):
    train_data = height.iloc[train_index]
    test_data = height.iloc[test_index]
    # Model and fit:
    model = ols(formula='height ~ age + gender', data=train_data)
    trainfit = model.fit()
    # Predict on test_data and assign predictions to correct locations
    height.iloc[test_index, height.columns.get_loc('Predicted')] = (
    trainfit.predict(test_data))

# Compute RMSE
rmse = np.sqrt(np.mean((height['height'] - height['Predicted']) ** 2))
print("RMSE:", rmse)
```

Problem 2:

Roulette is a game of chance where a roulette wheel has 37 numbered pockets where a ball can land. 18 of these pockets are red, 18 are black, and 1 is green (at least in Europe).

If, for example, you bet 1€, on red, and red occurs, then you win your bet back plus 1€, (i.e., a gain of 1€). But if black or green occurs, you lose the bet (a gain of -1€).

a) You bet 1€, on red. Answer the following:

   i) What is the expected gain?

   ii) What is the variance of this gain?

   iii) What is the probability of a positive gain?

b) You continue this strategy for $n = 50$ times. Use the Central Limit Theorem to answer the questions in part (a) for the average gain.

c) How would you solve (a) using simulation? Explain how such a simulation would be set up. You are not expected to write Python code, but describe the procedure you would use

Problem 3:

A premium light bulb producer claims their light bulbs lasts for more than 20,000 hours. We perform a hypothesis test for the claim for the premium light bulb using a one-sample t-test with the following test statistic:

$$T_1 = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}\text{-distributed.}$$

a) Set up H_0 and H_A.

b) Discuss how the test statistic "behaves" under the two hypothesis?

Having tested 45 light bulbs, you find that $\bar{x}_1 = 21\,253$ and $s_1 = 2200$.

c) Calculate the test statistic and find the p-value.

Hint: One of these printouts from Python may be useful.

```
from scipy import stats
print(stats.t.cdf(t1, df = 45))
```

```
## 0.9997971947217582
```

```
print(stats.t.cdf(t1, df = 44))
```

```
## 0.9997928824923898
```

```
print(stats.norm.cdf(t1))
```

```
## 0.9999334435960688
```

The premium light bulb producers claim that their product last 5000 hours longer than the economy one. You have therefore also tested 45 economy class light bulbs and find $\bar{x}_2 = 15\,990$ and $s_2 = 1500$.

d) Set up H_0 and H_A for testing if the difference between the two light bulbs is really as large as they claim.

The test statistic for this setup,

$$T_2 = \frac{\bar{X}_1 - \bar{X}_2 - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

is t-distributed with degrees of freedom (df):

$$\text{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_1^2/n_2)^2}{n_2 - 1}}$$

e) Discuss how the test statistic will behave under the various hypothesis.

f) The degrees of freedom is 77.6. Find the value of the test statistic. What is the p-value here?

Hint:

```
from scipy import stats
print(stats.t.cdf(t2, df = 77.6))
```

```
## 0.7452188872062666
```

g) The premium light bulb costs 50% more than economy one. Is the difference economically significant and which light bulb would you buy?

Problem 4:

A wind turbine placed on Fedje, an island off the Norwegian west coast, produces electricity. The company that owns the turbine, *Spinning Profits*, wants to make a prediction model for how much electricity they can expect to generate the following days based on weather forecasts from Yr. They read in a physics book that the power available in wind, P_{wind}, is given by:

$$P_{\text{wind}} = \frac{1}{2}\rho A v^3,$$

where

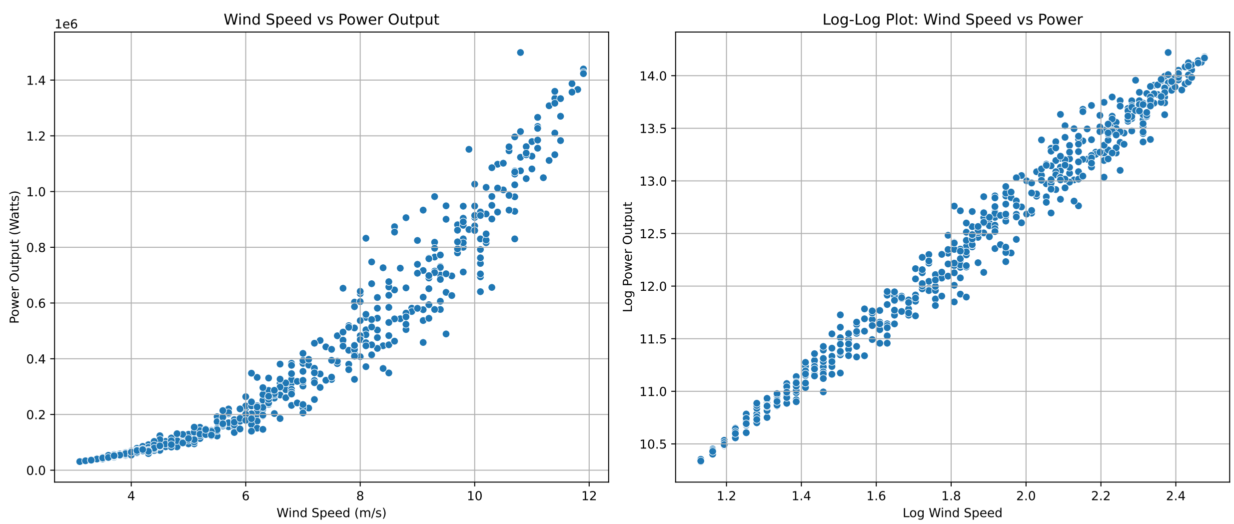
- P_{wind} is the theoretical power in watts
- v is the wind speed in m/s
- ρ is the air density (kg/m^3)
- A is the swept area of turbine blades in m^2.

The turbine has a radius of $r = 40m$. The parameters used are $\rho = 1.225\,\text{kg}/m^3$ and $A = \pi r^2 = \pi * (40m)^2 = 5027$. No wind turbine is able to produce this theoretical power. It should be considered a maximum.

a) Find an expression for $\log P_{\text{wind}}$. What kind of relationship is there between logarithmic wind power and logarithmic wind speed?

The company collects hourly data from their turbine for the last 30 days. The dataset contains two variables, $\mathtt{v}$ = wind speed (m/s) and P = power produced by the wind turbine (in Watt-hours). If there is too little wind (below 3 m/s) or too much wind (above 25 m/s), no energy is produced. At these high winds, the risk of damaging the turbine is too high. Between 12 m/s and 25 m/s, the turbine produces at it maximum, which is 15 megawatts (10^6 watts).

We start focusing on the transition between zero power and maximum production corresponding to wind speeds between 3 m/s and 12 m/s. In this range of wind speeds, we expect that the power produced will follow the theoretical relation between wind speed and power. In the figure below, we have plotted both the relationship between wind speed and power output, and the same quantities on the log-scale to the right.



In **Appendix A**, you find a regression summary from Python.

b) Set up the equation for the model that has been fitted here.

c) How much of the variance is explained by the model?

d) Give an interpretation of the estimated coefficients. Is the coefficient associated with wind speed consistent with the formula for the theoretical relationship?

In **Appendix B**, you find diagnostic plots for the assumptions of the models.

e) What are the assumptions of a linear regression model and are they fulfilled here?

The company is pleased with having this model for the relationship between wind speed and power output, and want to use it for prediction, given a wind forecast from Yr.no. Consider the two scenarios: i) Yr.no is forecasting 8 m/s and ii) Yr.no is forecasting 17 m/s.

f) Make point forecasts using the estimated model. What is the predicted power production in these scenarios according to the model? Comment on the two scenarios.

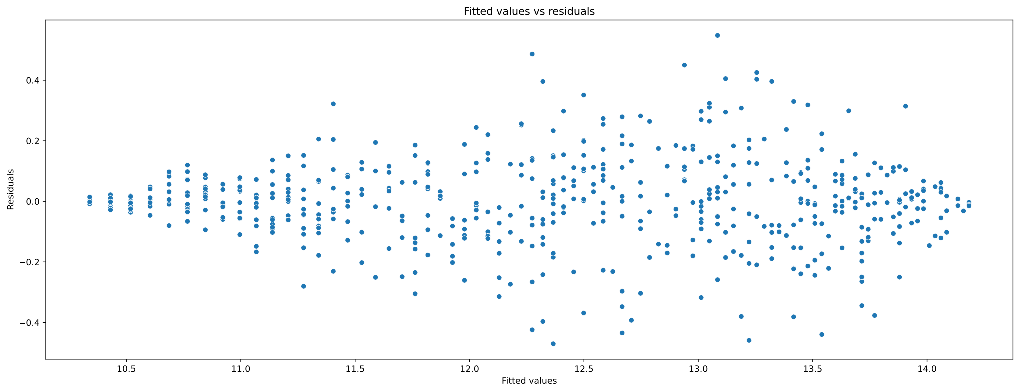
g) In **Appendix C** you will find a figure showing the fitted regression line with 95% Prediction interval and 95% Confidence interval. The latter is so narrow, it is almost impossible to distinguish from the regression line. Why is the prediction interval wider?

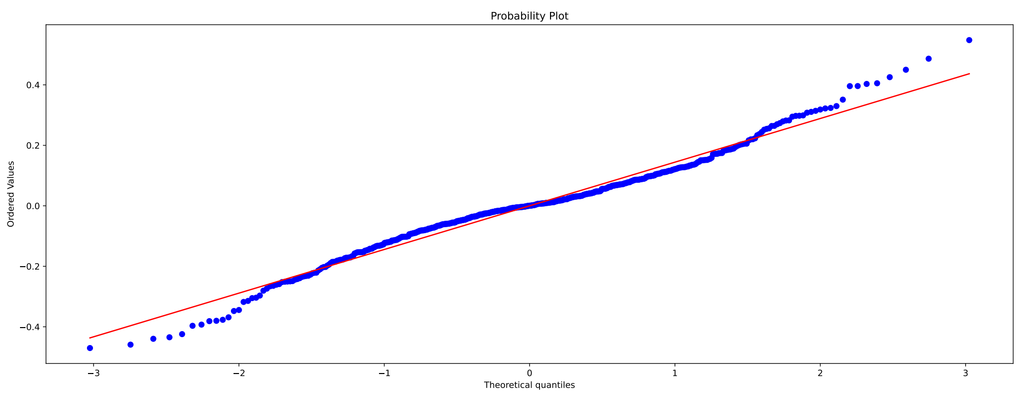
Appendix A:

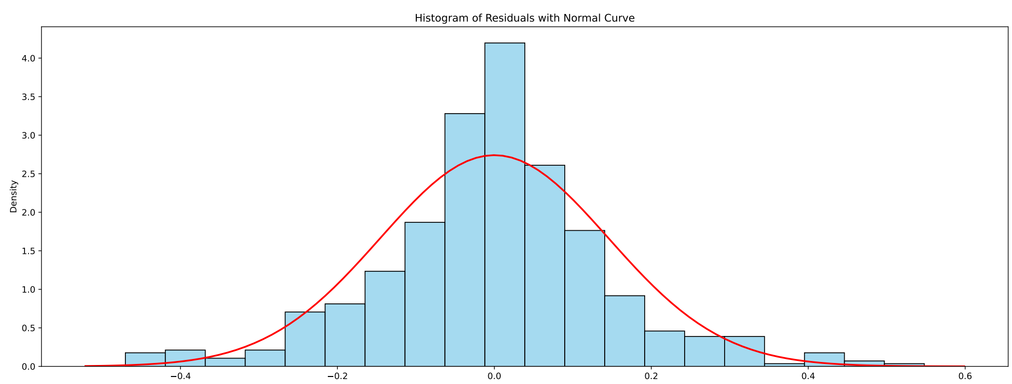
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  log_P   R-squared:                       0.983
Model:                            OLS   Adj. R-squared:                  0.983
Method:                 Least Squares   F-statistic:                 3.124e+04
Date:                Tue, 08 Apr 2025   Prob (F-statistic):               0.00
Time:                        12:57:24   Log-Likelihood:                 283.58
No. Observations:                 557   AIC:                            -563.2
Df Residuals:                     555   BIC:                            -554.5
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.1069      0.030    233.716      0.000       7.047       7.167
log_v          2.8574      0.016    176.745      0.000       2.826       2.889
==============================================================================
Omnibus:                       18.382   Durbin-Watson:                   2.040
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               42.164
Skew:                           0.008   Prob(JB):                     6.99e-10
Kurtosis:                       4.348   Cond. No.                         11.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
```

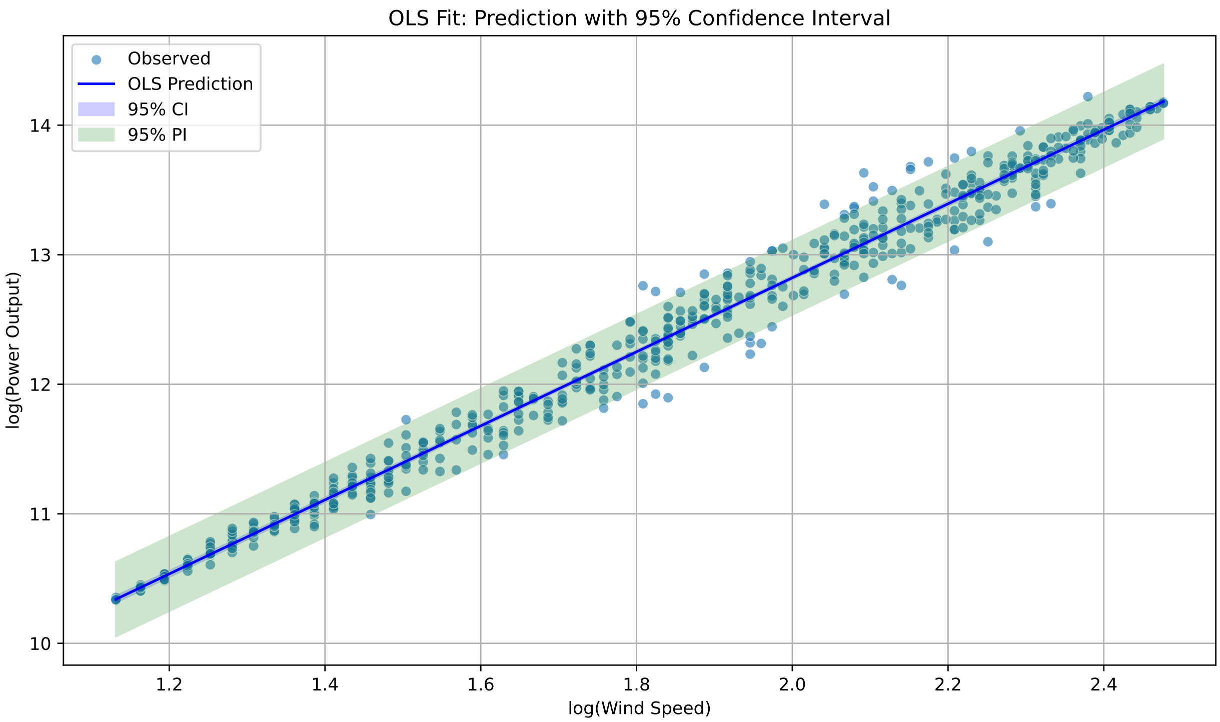
Appendix B:


Fitted values vs residuals
Residuals
Fitted values
0.4
0.2
0.0
−0.2
−0.4
10.5
11.0
11.5
12.0
12.5
13.0
13.5
14.0


Probability Plot
Ordered Values
Theoretical quantiles
0.4
0.2
0.0
−0.2
−0.4
−3
−2
−1
0
1
2
3


Histogram of Residuals with Normal Curve
Density
Residuals
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.4
−0.2
0.0
0.2
0.4
0.6

Appendix C:



OLS Fit: Prediction with 95% Confidence Interval
Observed
OLS Prediction
95% CI
95% PI
log(Power Output)
log(Wind Speed)
10
11
12
13
14
1.2
1.4
1.6
1.8
2.0
2.2
2.4