homework-04

October 19, 2023

1 Homework 4

1.1 References

- Lectures 13-16 (inclusive).

1.2 Instructions

- Type your name and email in the "Student details" section below.
- Develop the code and generate the figures you need to solve the problems using this notebook.
- For the answers that require a mathematical proof or derivation you should type them using latex. If you have never written latex before and you find it exceedingly difficult, we will likely accept handwritten solutions.
- The total homework points are 100. Please note that the problems are not weighed equally.

```python
import numpy as np
np.set_printoptions(precision=3)
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set(rc={"figure.dpi":100, "savefig.dpi":300})
sns.set_context("notebook")
sns.set_style("ticks")

import scipy
import scipy.stats as st
import urllib.request
import os

def download(
    url : str,
    local_filename : str = None
):
    """Download a file from a url.

    Arguments
    url            -- The url we want to download.
    local_filename -- The filemame to write on. If not
                      specified
```

```
    """
    if local_filename is None:
        local_filename = os.path.basename(url)
    urllib.request.urlretrieve(url, local_filename)
```

1.3 Student details

- **First Name:** Stav
- **Last Name:** Zeliger
- **Email:** szeliger@purdue.edu

2 Problem 1 - Estimating the mechanical properties of a plastic material from molecular dynamics simulations

First, make sure that this dataset is visible from this Jupyter notebook. You may achieve this by either:

- Downloading the data file, and then mannually upload it on Google Colab. The easiest way is to click on the folder icon on the left of the browser window and click on the upload button (or just drag and drop the file). Some other options are here.
- Downloading the file to the working directory of this notebook with this code:

```
[ ]: url = "https://github.com/PredictiveScienceLab/data-analytics-se/raw/master/
     ↪lecturebook/data/stress_strain.txt"
     download(url)
```

It's up to you what you choose to do. If the file is in the right place, the following code should work:

```
[ ]: data =  np.loadtxt('stress_strain.txt')
```

The dataset was generated using a molecular dynamics simulation of a plastic material (thanks to Professor Alejandro Strachan for sharing the data!). Specifically, Strachan's group did the following: - They took a rectangular chunk of the material and marked the position of each one of its atoms; - They started applying a tensile force along one dimension. The atoms are coupled together through electromagnetic forces and they must all satisfy Newton's law of motion. - For each value of the applied tensile force they marked the stress (force be unit area) in the middle of the materail and the corresponding strain of the material (percent enlogation in the pulling direction). - Eventually the material entered the plastic regime and then it broke. Here is a visualization of the data:

```
[ ]: # Strain
     x = data[:, 0]
     # Stress in MPa
     y = data[:, 1]

     plt.figure()
     plt.plot(
         x,
```
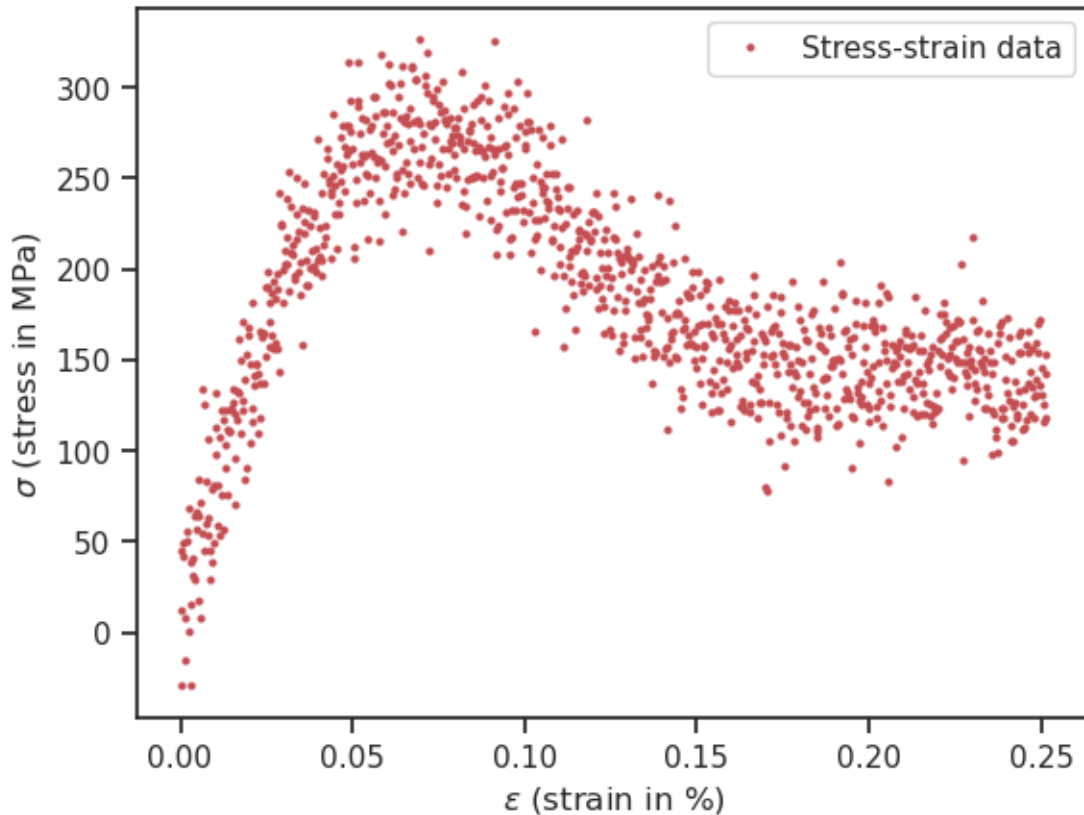
```
    y,
    'ro',
    markersize=2,
    label='Stress-strain data'
)
plt.xlabel('$\epsilon$ (strain in %)')
plt.ylabel('$\sigma$ (stress in MPa)')
plt.legend(loc='best');
```


Note that for each particular value of the strain, you don't necessarily get a unique stress. This is because in molecular dynamics the atoms are jiggling around due to thermal effects. So there is always this "jiggling" noise when you are trying to measure the stress and the strain. We would like to process this noise in order to extract what is known as the stress-strain curve of the material. The stress-strain curve is a macroscopic property of the the material which is affeted by the fine structure, e.g., the chemical bonds, the crystaline structure, any defects, etc. It is a required input to mechanics of materials.

2.1 Part A - Fitting the stress-strain curve in the elastic regime

The very first part of the stress-strain curve should be linear. It is called the *elastic regime*. In that region, say $\epsilon < \epsilon_l = 0.04$, the relationship between stress and strain is:

$$\sigma(\epsilon) = E\epsilon.$$

The constant E is known as the *Young modulus* of the material. Assume that you measure ϵ without any noise, but your measured σ is noisy.

2.1.1 Subpart A.I

First, extract the relevant data for this problem, split it into training and validation datasets, and visualize the training and validation datasets using different colors.

```python
# The point at which the stress-strain curve stops being linear
epsilon_l = 0.04
# Relevant data (this is nice way to get the linear part of the stresses and␣
 ↪straints)
x_rel = x[x < 0.04]
y_rel = y[x < 0.04]

# Visualize to make sure you have the right data
plt.figure()
plt.plot(
    x_rel,
    y_rel,
    'ro',
    markersize=2,
    label='Stress-strain data'
)
plt.xlabel('$\epsilon$ (strain in %)')
plt.ylabel('$\sigma$ (stress in MPa)')
plt.legend(loc='best');
```
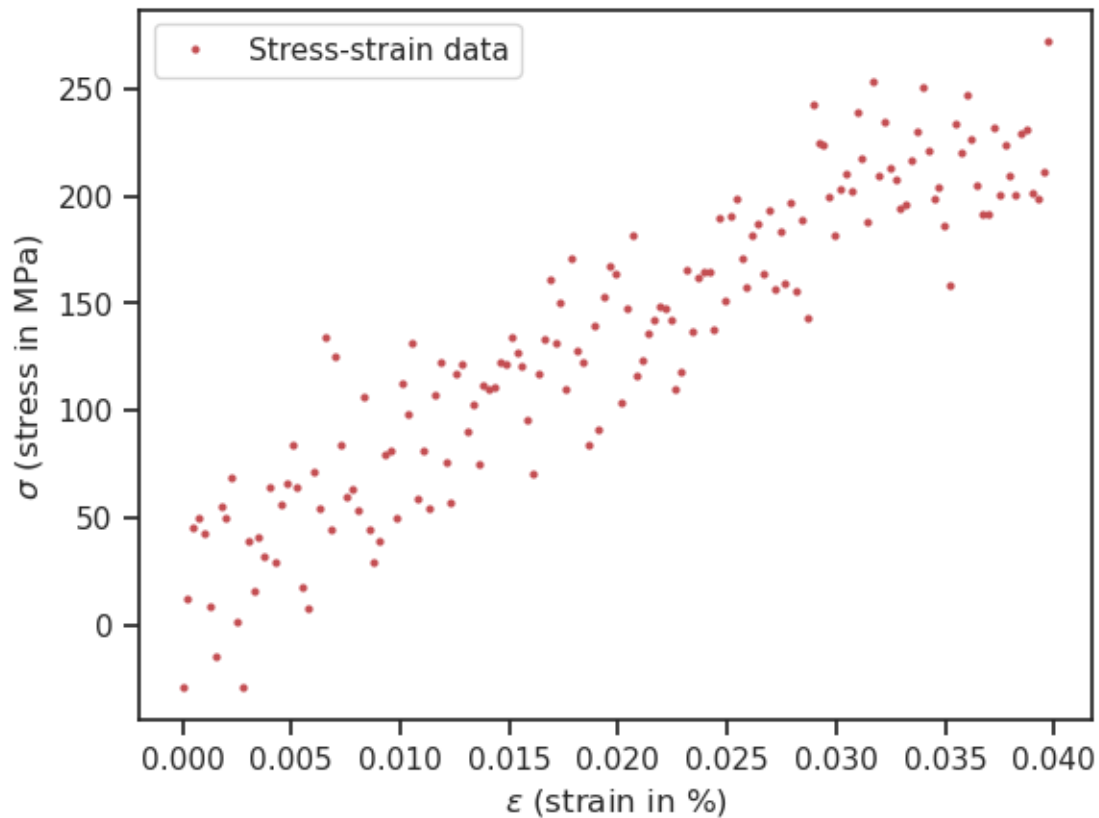
Split your data into training and validation.

Hint: You may use sklearn.model_selection.train_test_split if you wish.

```python
# Split the data into training and validation datasets
# Hint: Consult the lecture notes

import sklearn as skl
x_train, x_valid, y_train, y_valid = skl.model_selection.
 ↪train_test_split(x_rel,y_rel,test_size=0.33)
```
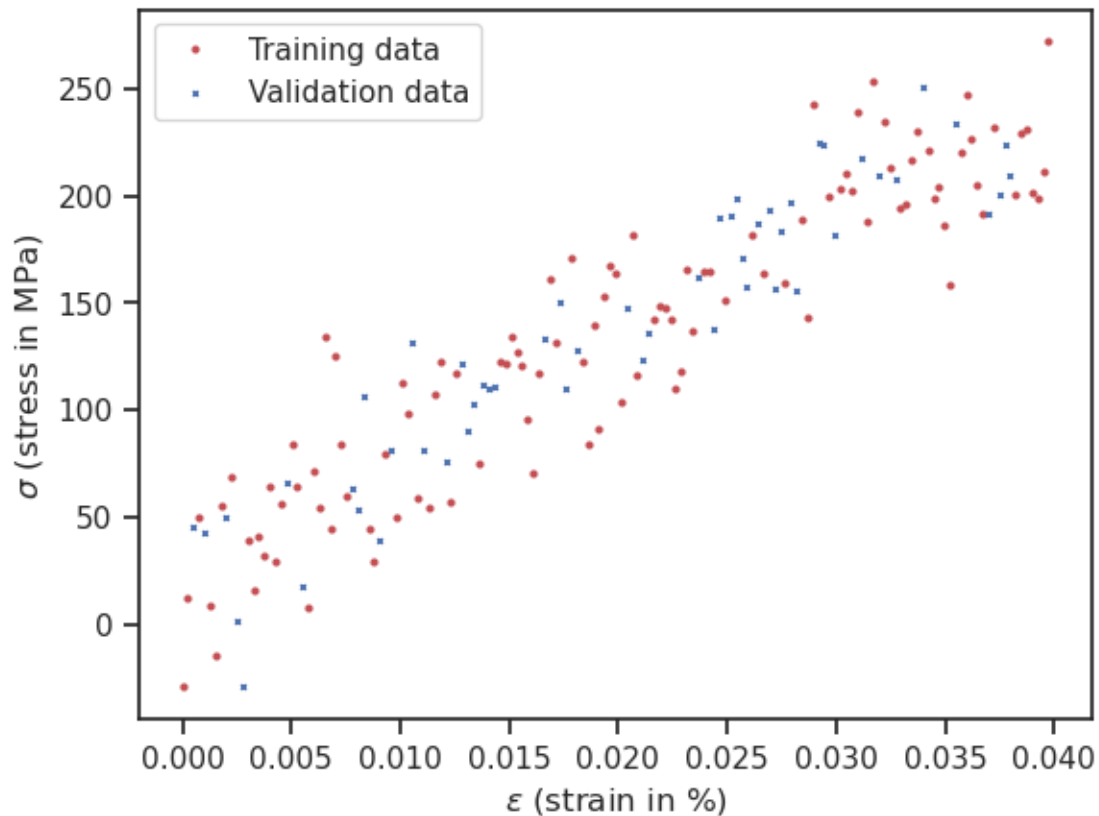
Use the following to visualize your split:

```python
plt.figure()
plt.plot(
    x_train,
    y_train,
    'ro',
    markersize=2,
    label='Training data'
)
plt.plot(
```

```
    x_valid,
    y_valid,
    'bx',
    markersize=2,
    label='Validation data'
)
plt.xlabel('$\epsilon$ (strain in %)')
plt.ylabel('$\sigma$ (stress in MPa)')
plt.legend(loc='best');
```


2.1.2 Subpart A.II

Perform Bayesian linear regression with the evidence approximation to estimate the noise variance
and the hyperparameters of the prior.

```
[ ]: from sklearn.linear_model import ARDRegression

def get_polynomial_design_matrix(x, degree):
    """Return the polynomial design matrix of ``degree`` evaluated at ``x``.

    Arguments:
```

```
    x       -- A 2D array with only one column.
    degree -- An integer greater than zero.
    """
    assert isinstance(x, np.ndarray), 'x is not a numpy array.'
    assert x.ndim == 2, 'You must make x a 2D array.'
    assert x.shape[1] == 1, 'x must be a column.'
    cols = []
    for i in range(degree+1):
        cols.append(x ** i)

    return np.hstack(cols)

# Get Phi
Phi = get_polynomial_design_matrix(x_train[:, None], 1)

# Fit
regressionModel = ARDRegression(fit_intercept = False).fit(Phi,y_train)

# Get Variance
sigma = np.sqrt(1.0 / regressionModel.alpha_)
lamb = regressionModel.lambda_
print(f'Sigma = {sigma}')
print(f'Hyperparameters = {lamb}')
```

```
Sigma = 27.8869185172524
Hyperparameters = [1.469e-03 3.408e-08]
```

2.1.3 Subpart A.III

Calculate the mean square error of the validation data.

```
[ ]: # your code here
     Phi_valid = get_polynomial_design_matrix(x_valid[:, None], 1)

     m = regressionModel.coef_
     S = regressionModel.sigma_
     w_post = st.multivariate_normal(
         mean=m,
         cov=S + np.eye(S.shape[0])
     )
     w = w_post.rvs()


     y_prediction = Phi_valid @ w
     error= np.mean((y_prediction - y_valid)**2)
     print(f'Mean Squared Error: {error}')
```

```
Mean Squared Error: 585.0597030865363
```

2.1.4 Subpart A.IV

Make the observations vs predictions plot for the validation data.

```
[ ]: # your code here

plt.figure()
plt.plot(y_prediction,y_valid,'bo',)
yys = np.linspace(
    y_valid.min(),
    y_valid.max(),
    100)
plt.plot(yys, yys, 'r-')

plt.xlabel('Predictions')
plt.ylabel('Observations')
plt.title('Observations vs predictions for order 1 polynomial')
sns.despine(trim=True)
```


Observations vs predictions for order 1 polynomial
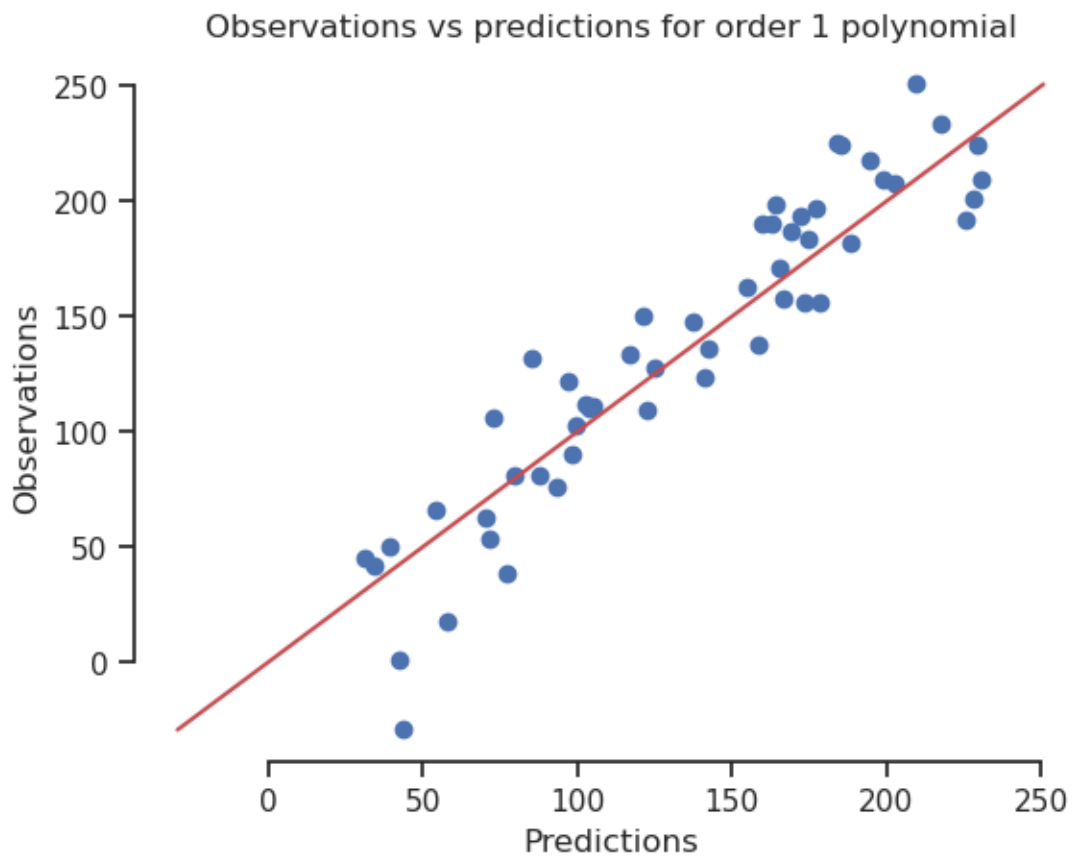
2.1.5 Subpart A.V

Compute and plot the standarized errors for the validation data.

```
[ ]: # your code here
     eps = (y_valid - y_prediction) / sigma

     idx = np.arange(1, eps.shape[0] + 1)

     fig, ax = plt.subplots()
     ax.plot(idx, eps, 'o', label='Standarized errors')
     ax.plot(idx, 1.96 * np.ones(eps.shape[0]), 'r--')
     ax.plot(idx, -1.96 * np.ones(eps.shape[0]), 'r--')
     ax.set_xlabel('$i$')
     ax.set_ylabel('$\epsilon_i$')
     sns.despine(trim=True);
```
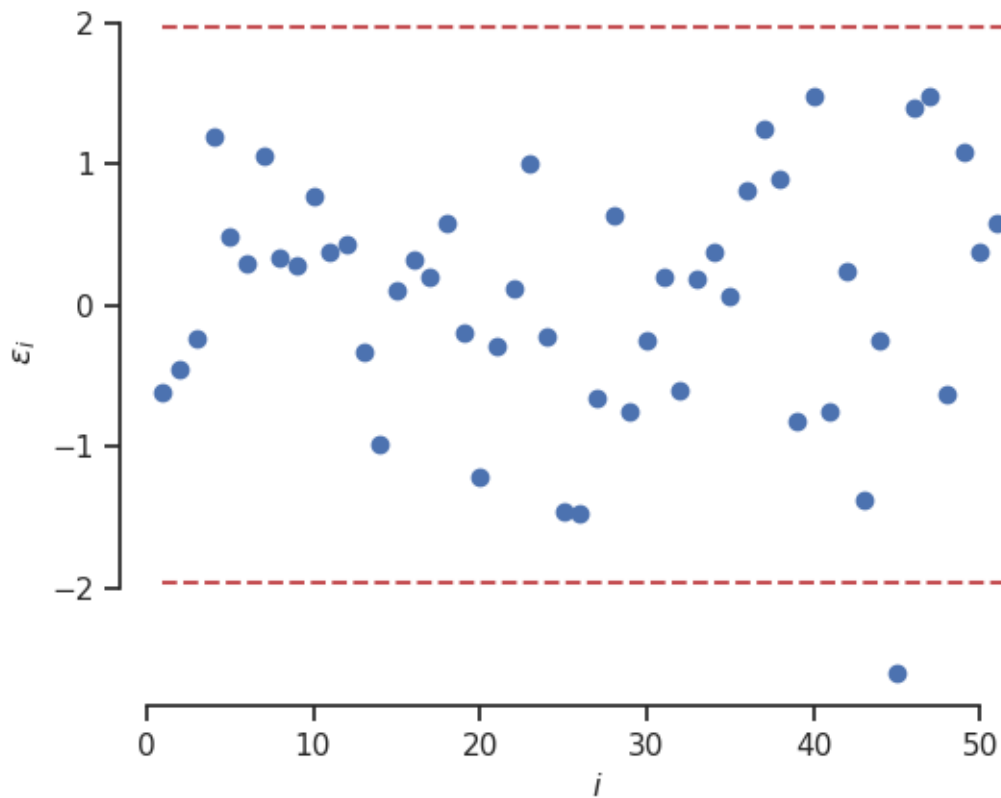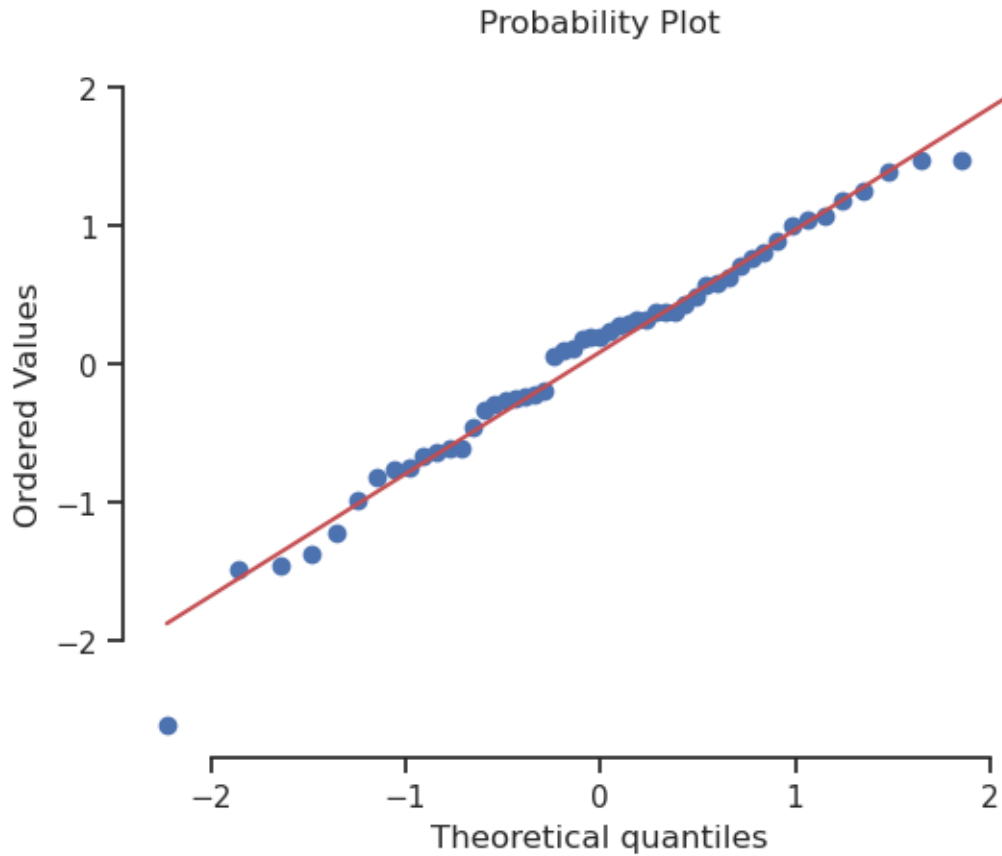

2.1.6 Subpart A.VI

Make the quantile-quantile plot of the standarized errors.

```
[ ]:  # your code here
      fig, ax = plt.subplots()
      st.probplot(eps, dist=st.norm, plot=ax)
      sns.despine(trim=True);
```


Probability Plot

2.1.7 Subpart A.VII

Visualize your epistemic and the aleatory uncertainty about the stress-strain curve in the elastic regime.

```
[ ]:  # your code here
      def plot_posterior_predictive(
          model,
          xx,
          phi_func,
          phi_func_args=(),
          y_true=None
      ):
          """Plot the posterior predictive separating
          aleatory and espitemic uncertainty.
```

```python
    Arguments:
    model    -- A trained model.
    xx       -- The points on which to evaluate
                the posterior predictive.
    phi_func -- The function to use to compute
                the design matrix.

    Keyword Arguments:
    phi_func_args -- Any arguments passed to the
                     function that calculates the
                     design matrix.
    y_true        -- The true response for plotting.
    """
    Phi_xx = phi_func(
        xx[:, None],
        *phi_func_args
    )
    yy_mean, yy_measured_std = model.predict(
        Phi_xx,
        return_std=True
    )
    sigma = np.sqrt(1.0 / model.alpha_)
    yy_std = np.sqrt(yy_measured_std ** 2 - sigma**2)
    yy_le = yy_mean - 2.0 * yy_std
    yy_ue = yy_mean + 2.0 * yy_std
    yy_lae = yy_mean - 2.0 * yy_measured_std
    yy_uae = yy_mean + 2.0 * yy_measured_std

    fig, ax = plt.subplots()
    ax.plot(xx, yy_mean, 'r', label="Posterior mean")
    ax.fill_between(
        xx,
        yy_le,
        yy_ue,
        color='red',
        alpha=0.25,
        label="95% epistemic credible interval"
    )
    ax.fill_between(
        xx,
        yy_lae,
        yy_le,
        color='green',
        alpha=0.25
    )
    ax.fill_between(
```

```python
        xx,
        yy_ue,
        yy_uae,
        color='green',
        alpha=0.25,
        label="95% epistemic + aleatory credible interval"
    )
    ax.plot(x_rel, y_rel, 'kx', label='Observed data')
    if y_true is not None:
        ax.plot(xx, y_true, "--", label="True response")
    ax.set_xlabel('$x$')
    ax.set_ylabel('$y$')
    plt.legend(loc="upper left", frameon=False)
    sns.despine(trim=True)




xx = np.linspace(0, 0.04, 100)

plot_posterior_predictive(
    regressionModel,
    xx,
    get_polynomial_design_matrix,
    phi_func_args=(1,)
)
```
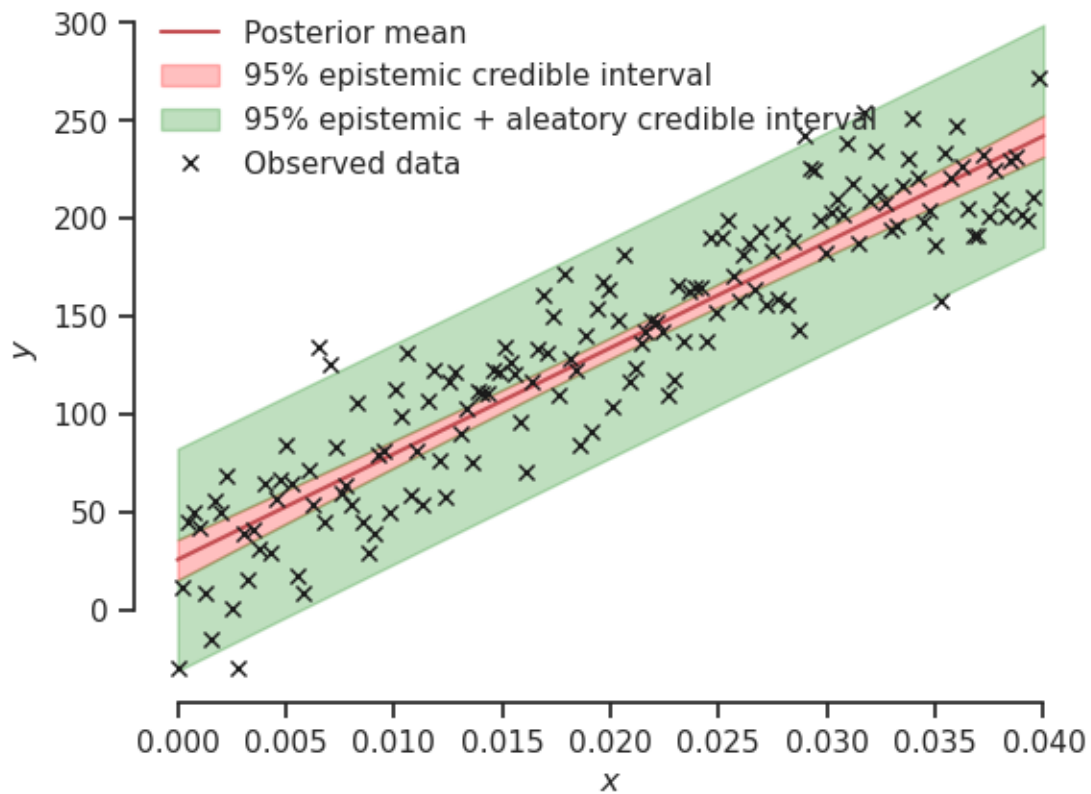
2.1.8 Subpart A. VIII

Visualize the posterior of the Young modulus E conditioned on the data.

```python
# your code here
import scipy.stats as st

sigma = regressionModel.sigma_
print(sigma)
m = regressionModel.coef_
ww = np.linspace(4500, 6500, 100)


fig, ax = plt.subplots()
wj_post = st.norm(
    loc=m[1],
    scale=np.sqrt(sigma[1,1])
)
ax.plot(
    ww,
    wj_post.pdf(ww),
```
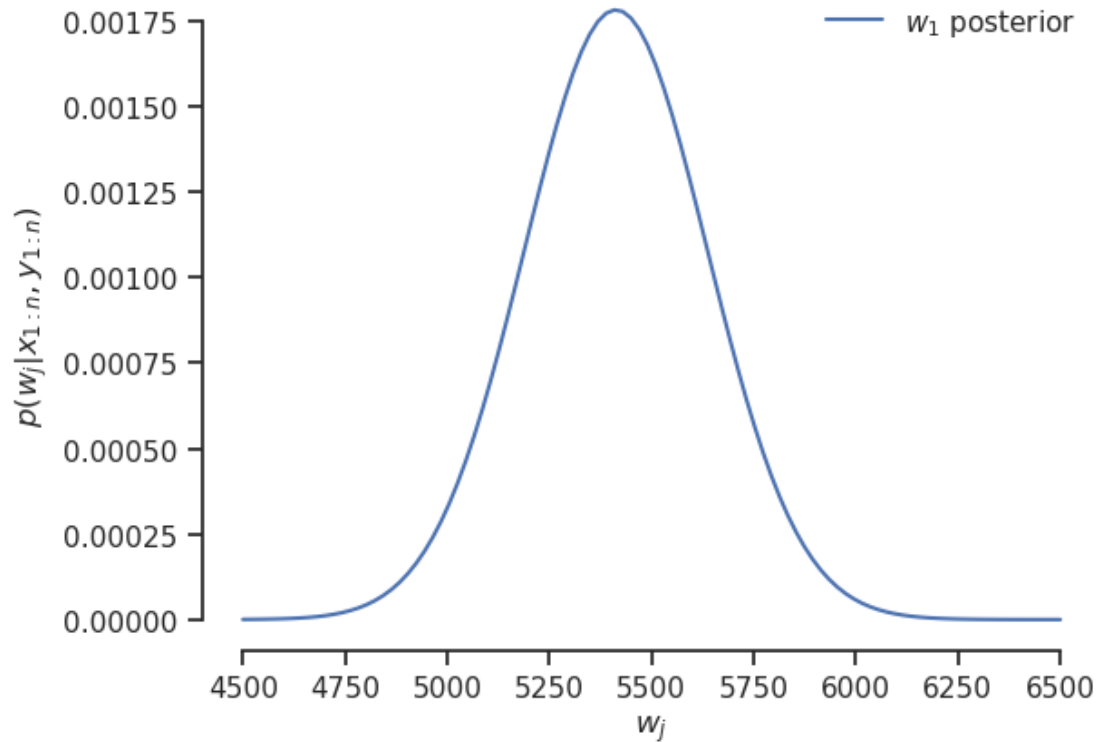
```
     label=f'$w_{{{1}}}$ posterior')
ax.set_xlabel("$w_j$")
ax.set_ylabel("$p(w_j|x_{{1:n}}, y_{{1:n}})$")
plt.legend(loc='best', frameon=False)
sns.despine(trim=True);
```

```
[[ 2.664e+01 -9.851e+02]
 [-9.851e+02  5.008e+04]]
```


2.1.9 Subpart A.IX

Take five samples of stress-strain curve in the elastic regime and visualize them.

```
[ ]: # your code here

def plot_posterior_samples(
    model,
    xx,
    phi_func,
    phi_func_args=(),
    num_samples=10,
    y_true=None,
```

```python
    nugget=1e-6
):
    """Plot posterior samples from the model.

    Arguments:
    model     -- A trained model.
    xx        -- The points on which to evaluate
                 the posterior predictive.
    phi_func -- The function to use to compute
                 the design matrix.

    Keyword Arguments:
    phi_func_args -- Any arguments passed to the
                     function that calculates the
                     design matrix.
    num_samples   -- The number of samples to take.
    y_true        -- The true response for plotting.
    nugget        -- A small number to add the covariance
                     if it is not positive definite
                     (numerically).
    """
    Phi_xx = phi_func(
        xx[:, None],
        *phi_func_args
    )
    m = model.coef_
    S = model.sigma_
    w_post = st.multivariate_normal(
        mean=m,
        cov=S + nugget * np.eye(S.shape[0])
    )
    fig, ax = plt.subplots()
    for _ in range(num_samples):
        w_sample = w_post.rvs()
        yy_sample = Phi_xx @ w_sample
        ax.plot(xx, yy_sample, 'r', lw=0.5)
    ax.plot([], [], "r", lw=0.5, label="Posterior samples")
    ax.plot(x_rel, y_rel, 'kx', label='Observed data')
    ax.set_xlabel('$x$')
    ax.set_ylabel('$y$')
    plt.legend(loc="best", frameon=False)
    sns.despine(trim=True)

plot_posterior_samples(
    regressionModel,
    xx,
    get_polynomial_design_matrix,
```
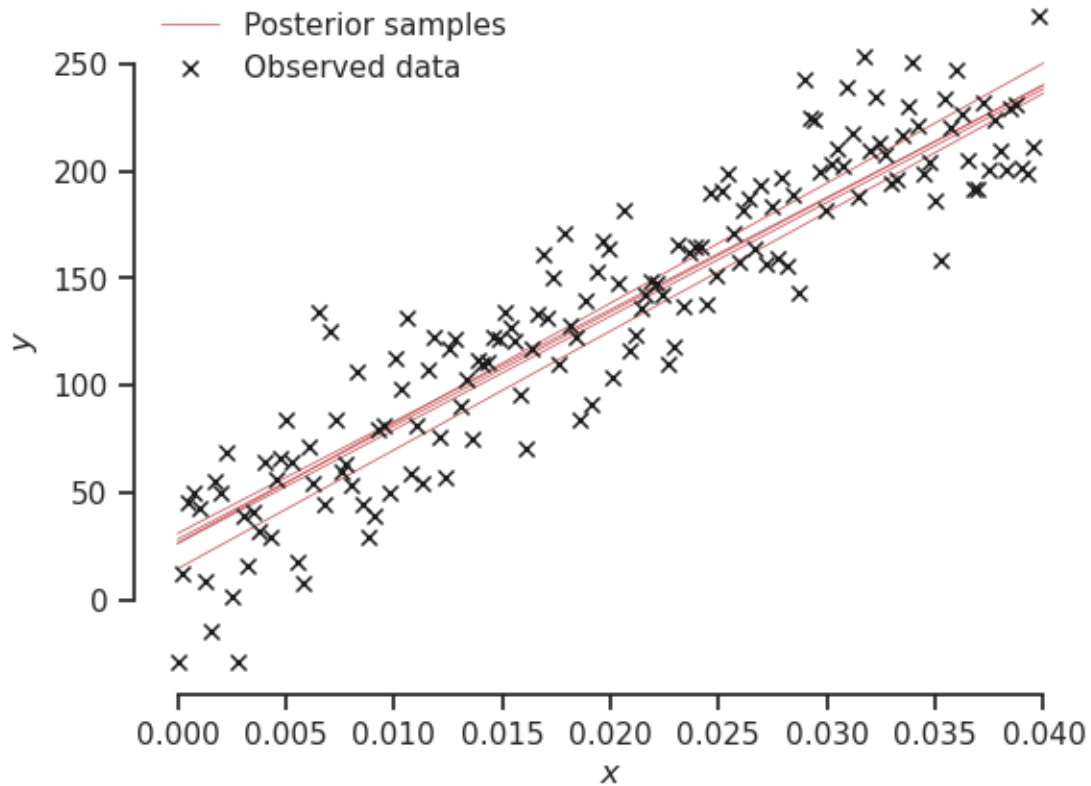
```
        phi_func_args=(1,),
        y_true=None,
        num_samples=5
)
```


2.1.10 Subpart A.X

Find the 95% centered credible interval for the Young modulus E.

```
[ ]: # your code here
     lb = wj_post.ppf(0.05)
     ub = wj_post.ppf(0.95)

     print(f'lower bound: {lb}')
     print(f'upper bound: {ub}')
```

```
lower bound: 5044.555882992406
upper bound: 5780.7513388955695
```

2.1.11 Subpart A.XI

If you had to pick a single value for the Young modulus E, what would it be and why?

```
[ ]: YoungModulus = np.mean([ub,lb])
     print(f'Mean of upper and lower bound: {YoungModulus}')
```

Mean of upper and lower bound: 5412.653610943988

I would choose this value for E, since it is the mean of the upper and lower bound of the 95% credible interval for it.

Your answer here

2.2 Part B - Estimate the ultimate strength

The pick of the stress-strain curve is known as the ultimate strength. We will like to estimate it.

2.2.1 Subpart B.I - Extract training and validation data

Extract training and validation data from the entire dataset.

```
[ ]: # your code here - Repeat as many text and code blocks as you like
     x_train, x_valid, y_train, y_valid = skl.model_selection.
      ↪train_test_split(x,y,test_size=0.33)
```

Use the following to visualize your split:

```
[ ]: plt.figure()
     plt.plot(
         x_train,
         y_train,
         'ro',
         markersize=2,
         label='Training data'
     )
     plt.plot(
         x_valid,
         y_valid,
         'bx',
         markersize=2,
         label='Validation data'
     )
     plt.xlabel('$\epsilon$ (strain in %)')
     plt.ylabel('$\sigma$ (stress in MPa)')
     plt.legend(loc='best');
```
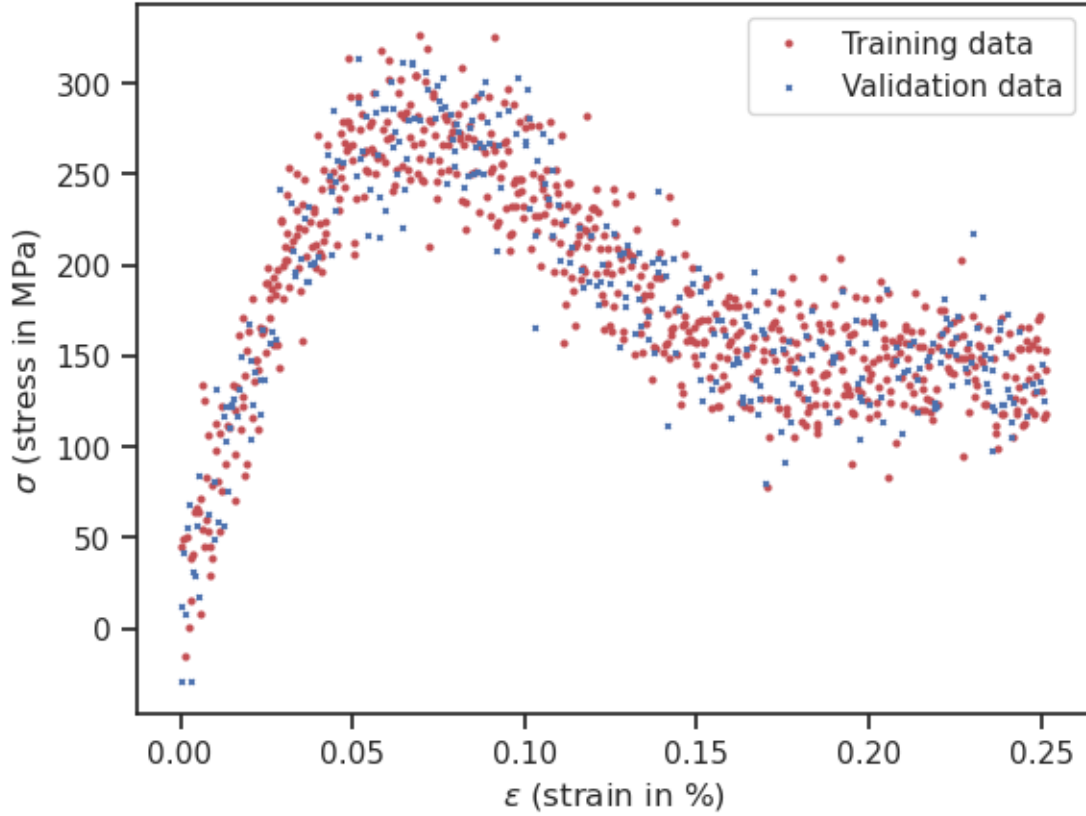
2.2.2 Subpart B.II - Model the entire stress-strain relationship.

To do this, we will set up a generalized linear model that can capture the entire stress-strain relationship. Remember, you can use any model you want as soon as: + it is linear in the parameters to be estimated, + it clearly has a well-defined elastic regime (see Part A).

I am going to help you set up the right model. We are goint to use the Heavide step function to turn on or off models for various ranges of ϵ. The idea is quite simple: We will use a linear model for the elastic regime and we are going to turn to a non-linear model for the non-linear regime. Here is a model that has the right form in the elastic regime and an arbitrary form in the non-linear regime:

$$f(\epsilon; E, \mathbf{w}_g) = E\epsilon \left[(1 - H(\epsilon - \epsilon_l)] + g(\epsilon; \mathbf{w}_g)H(\epsilon - \epsilon_l),$$

where

$$H(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{otherwise}, \end{cases}$$

and g is any function linear in the parameters $\mathbf{w}_g$.

You can use any model you like for the non-linear regime, but let's use a polynomial of degree d:

$$g(\epsilon) = \sum_{i=0}^{d} w_i \epsilon^i.$$

The full model can be expressed as:

$$f(\epsilon) = \begin{cases} h(\epsilon) = E\epsilon, \ \epsilon < \epsilon_l, \\ g(\epsilon) = \sum_{i=0}^{d} w_i \epsilon^i, \epsilon \geq \epsilon_l \end{cases}$$
$$= E\epsilon \left(1 - H(\epsilon - \epsilon_l)\right) + \sum_{i=0}^{d} w_i \epsilon^i H(\epsilon - \epsilon_l).$$

We could proceed with this model, but there is a small problem: It is discontinuous at $\epsilon = \epsilon_l$. This is unphysical. We can do better than that!

To make the model nice, we force the h and g to match up to the first derivative, i.e., we demand that:

$$h(\epsilon_l) = g(\epsilon_l)$$
$$h'(\epsilon_l) = g'(\epsilon_l).$$

The reason we include the first derivative is so that we don't have a kink in the stress-strain. That would also be unphysical. The two equations above become:

$$E\epsilon_l = \sum_{i=0}^{d} w_i \epsilon_l^i$$
$$E = \sum_{i=1}^{d} i w_i \epsilon_l^{i-1}.$$

We can use these two equations to eliminate two weights. Let's eliminate w_0 and w_1. All you have to do is express them in terms of E and $w_2, \dots, w_d$. So, there remain d parameters to estimate. Let's get back to the stress-strain model.

Our stress-strain model was:

$$f(\epsilon) = E\epsilon \left(1 - H(\epsilon - \epsilon_l)\right) + \sum_{i=0}^{d} w_i \epsilon^i H(\epsilon - \epsilon_l).$$

We can now use the expressions for w_0 and w_1 to rewrite this using only all the other parameters. I am going to spare you the details... The end result is:

$$f(\epsilon) = E\epsilon + \sum_{i=2}^{d} w_i \left[(i-1)\epsilon_l^i - i\epsilon\epsilon_l^{i-1} + \epsilon^i\right] H(\epsilon - \epsilon_l).$$

Okay. This is still a generalized linear model. This is nice. Write code for the design matrix:

```python
# Complete this code to make your model:
def compute_design_matrix(Epsilon, epsilon_l, d):
    """Compute the design matrix for the stress-strain curve problem.

    Arguments:
        Epsilon     -    A 1D array of dimension N.
        epsilon_l   -    The strain signifying the end of the elastic regime.
        d           -    The polynomial degree.

    Returns:
        A design matrix N x d
    """
    # Sanity check
    assert isinstance(Epsilon, np.ndarray)
    assert Epsilon.ndim == 1, 'Pass the array as epsilon.flatten(), if it is⏎
↪two dimensional'
    n = Epsilon.shape[0]
    # The design matrix:
    Phi = np.ndarray((n, d))
    # The step function evaluated at all the elements of Epsilon.
    # You can use it if you want.
    Step = np.ones(n)
    Step[Epsilon < epsilon_l] = 0
    # Build the design matrix
    Phi[:, 0] = epsilon_l*Epsilon
    for i in range(2, d+1):
        Phi[:, i-1] = ((i-1)*epsilon_l**i -⏎
↪i*Epsilon*epsilon_l**(i-1)+Epsilon**i) * Step
    return Phi
```

Visualize the basis functions here:

```python
d = 4
eps = np.linspace(0, x.max(), 100)
Phis = compute_design_matrix(eps, epsilon_l, d)
fig, ax = plt.subplots(dpi=100)
ax.plot(eps, Phis)
ax.set_xlabel('$\epsilon$ (strain in %))')
ax.set_ylabel('$\phi_i(\epsilon)$');
```
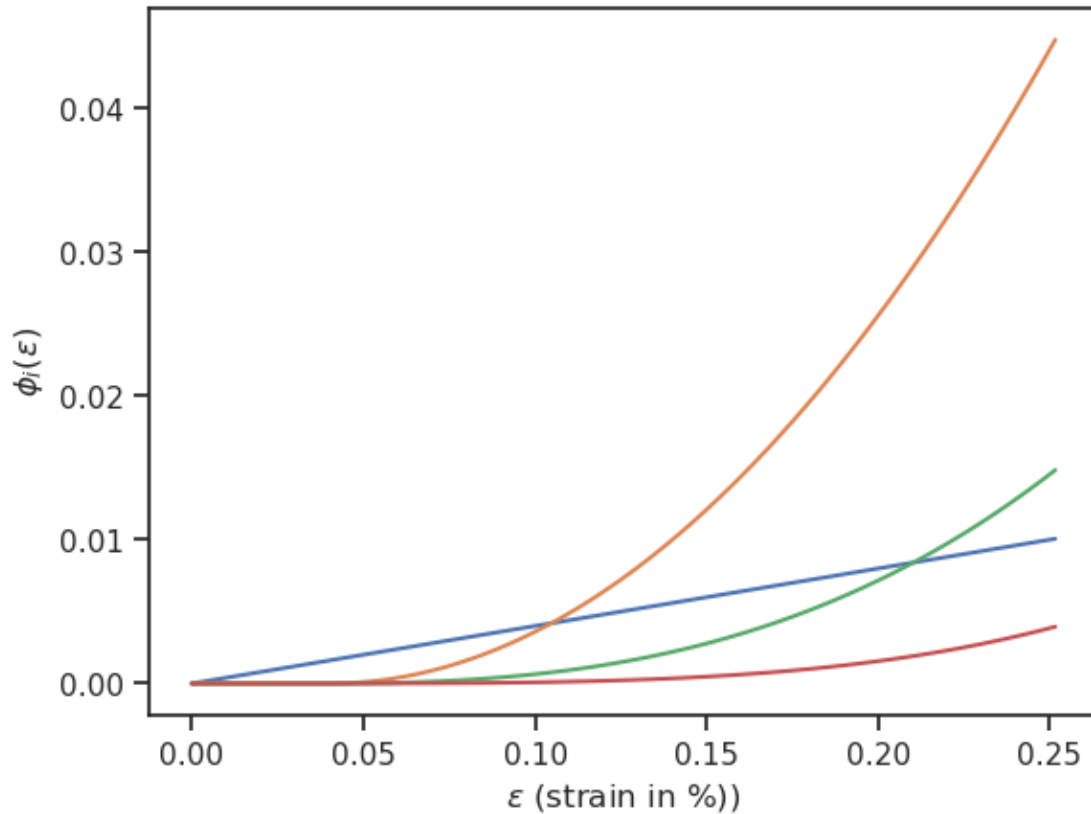
2.2.3 Subpart B.III

Fit the model using automatic relevance determination and demonstrate that it works well by doing all the things we did above (MSE, observations vs predictions plot, standarized errors, etc.).

```python
#### Sigma and Hyperparams #####
# Get Phis
Phis = compute_design_matrix(x_train, epsilon_l, d)

# Fit
model = ARDRegression(fit_intercept = False).fit(Phis,y_train)

# Get Variance
sigma = np.sqrt(1.0 / model.alpha_)
lamb = model.lambda_
print(f'Sigma = {sigma}')
print(f'Hyperparameters = {lamb}')


#### Mean Squared Error ####
```

```python
Phis_validation = compute_design_matrix(x_valid, epsilon_l, d)

m = model.coef_
S = model.sigma_
w_post = st.multivariate_normal(
    mean=m,
    cov=S + np.eye(S.shape[0])
)
w = w_post.rvs()



y_prediction = Phis_validation @ w
error= np.mean((y_prediction - y_valid)**2)
print(f'Mean Squared Error: {error}')



##### Observations vs Predictions ####
plt.figure()
plt.plot(y_prediction,y_valid,'bo',)
yys = np.linspace(
    y_valid.min(),
    y_valid.max(),
    100)
plt.plot(yys, yys, 'r-')

plt.xlabel('Predictions')
plt.ylabel('Observations')
plt.title(f'Observations vs predictions for order {d} polynomial')
sns.despine(trim=True)



#### Standardized Errors ####
eps = (y_valid - y_prediction) / sigma

idx = np.arange(1, eps.shape[0] + 1)

fig, ax = plt.subplots()
ax.plot(idx, eps, 'o', label='Standarized errors')
ax.plot(idx, 1.96 * np.ones(eps.shape[0]), 'r--')
ax.plot(idx, -1.96 * np.ones(eps.shape[0]), 'r--')
ax.set_xlabel('$i$')
ax.set_ylabel('$\epsilon_i$')
sns.despine(trim=True);



fig, ax = plt.subplots()
st.probplot(eps, dist=st.norm, plot=ax)
```
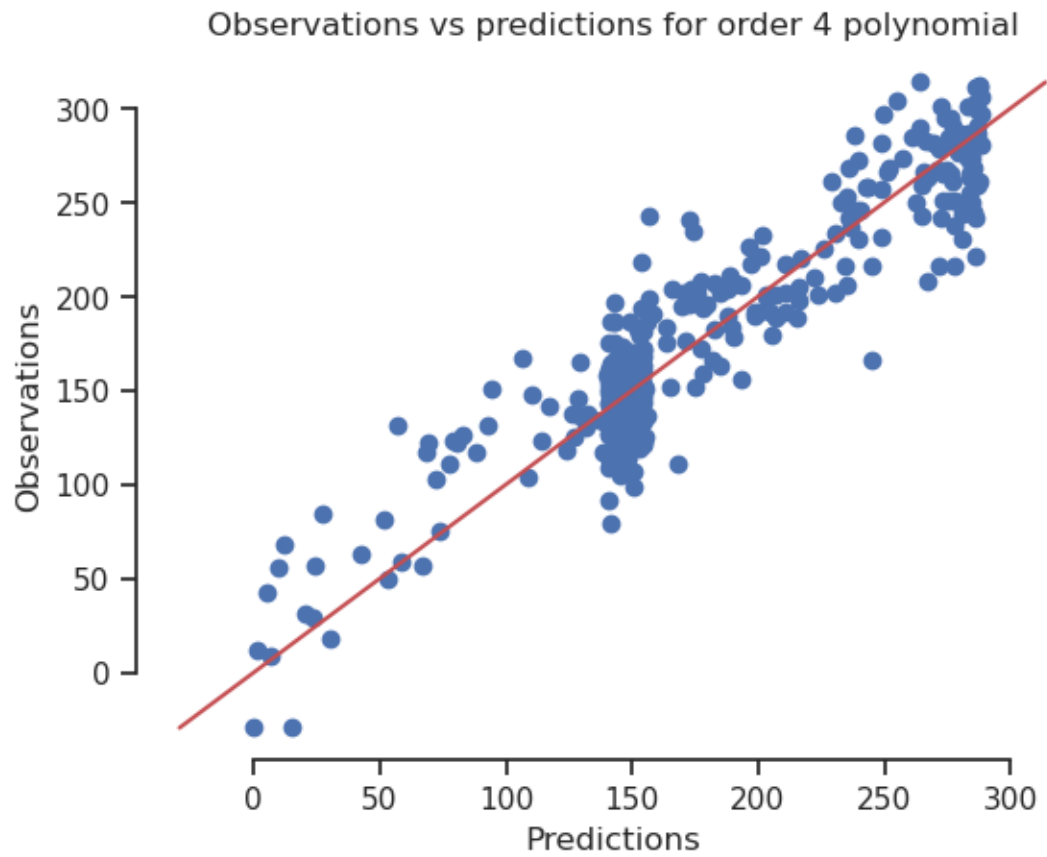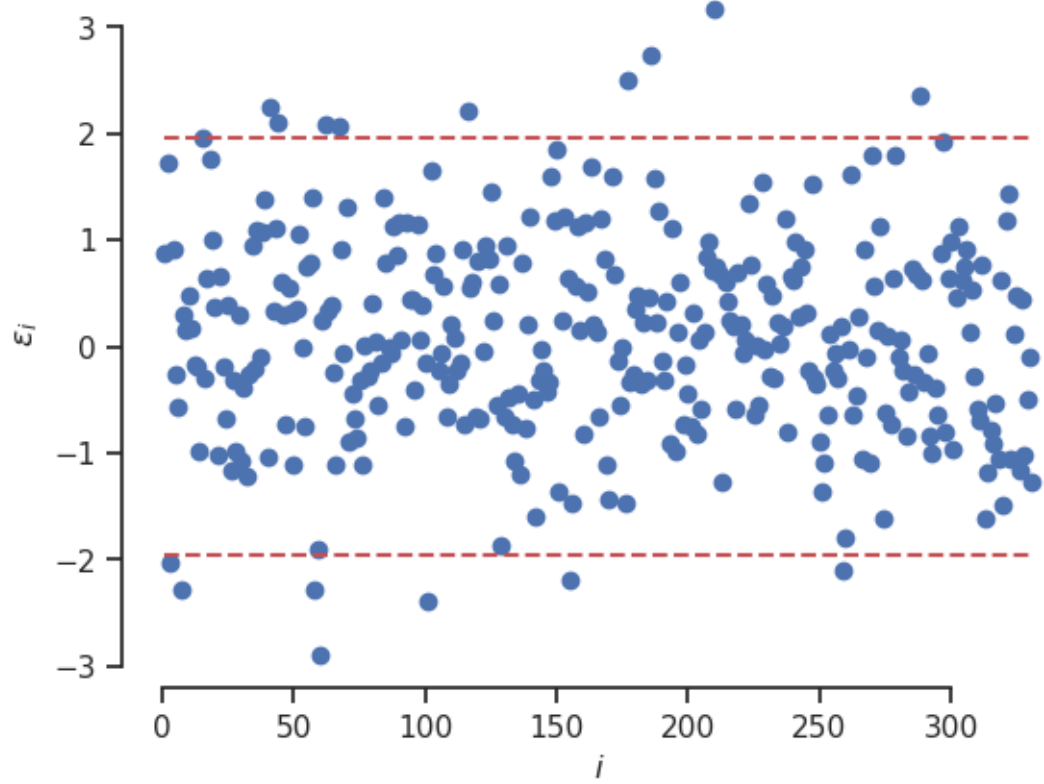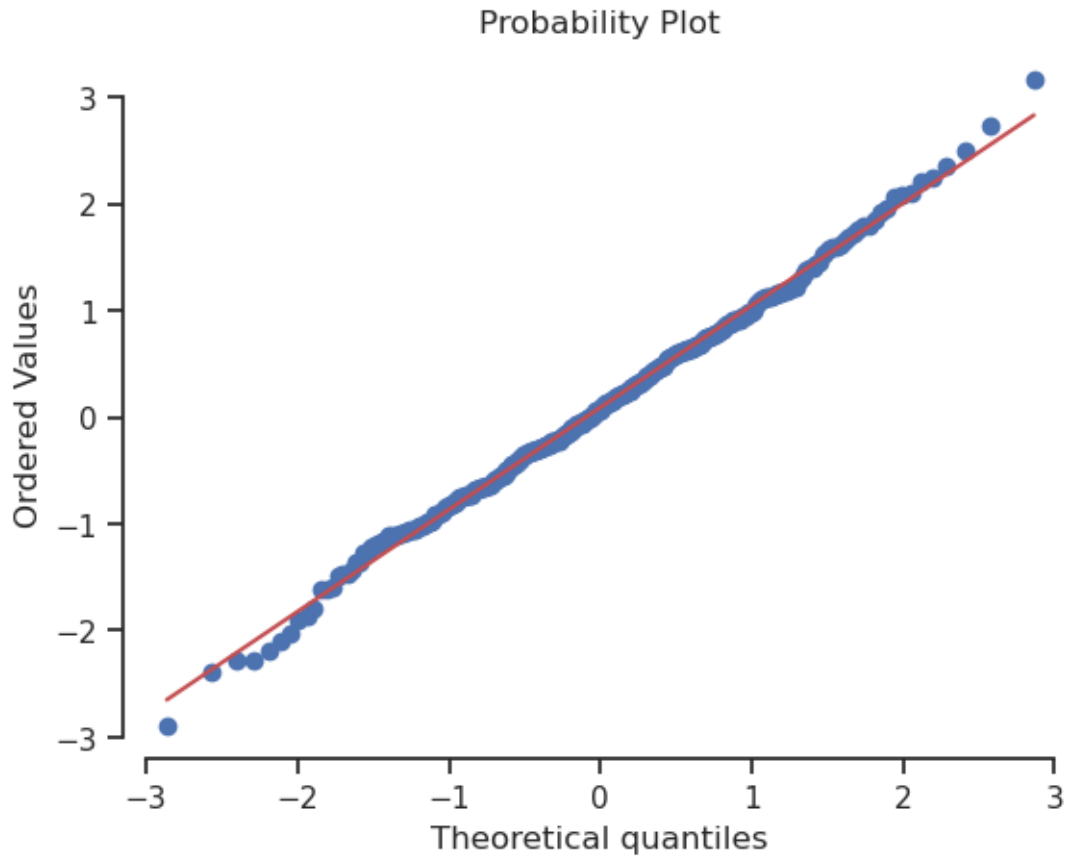
```
sns.despine(trim=True)
```

Sigma = 27.245588788453265
Hyperparameters = [5.416e-11 1.691e-11 7.844e-13 3.076e-13]
Mean Squared Error: 678.3994288343642

Observations vs predictions for order 4 polynomial

Probability Plot

2.2.4 Subpart B.IV

Visualize epistemic and aleatory uncertainty in the stess-strain relation.

```python
# Your code here
def plot_posterior_predictive(
    model,
    xx,
    phi_func,
    phi_func_args=(),
    y_true=None
):
    """Plot the posterior predictive separating
    aleatory and espitemic uncertainty.

    Arguments:
    model    -- A trained model.
    xx       -- The points on which to evaluate
                the posterior predictive.
    phi_func -- The function to use to compute
```

```python
            the design matrix.

    Keyword Arguments:
    phi_func_args -- Any arguments passed to the
                     function that calculates the
                     design matrix.
    y_true        -- The true response for plotting.
    """
    Phi_xx = phi_func(
        xx,
        *phi_func_args
    )
    yy_mean, yy_measured_std = model.predict(
        Phi_xx,
        return_std=True
    )
    sigma = np.sqrt(1.0 / model.alpha_)
    yy_std = np.sqrt(yy_measured_std ** 2 - sigma**2)
    yy_le = yy_mean - 2.0 * yy_std
    yy_ue = yy_mean + 2.0 * yy_std
    yy_lae = yy_mean - 2.0 * yy_measured_std
    yy_uae = yy_mean + 2.0 * yy_measured_std

    fig, ax = plt.subplots()
    ax.plot(xx, yy_mean, 'r', label="Posterior mean")
    ax.fill_between(
        xx,
        yy_le,
        yy_ue,
        color='red',
        alpha=0.25,
        label="95% epistemic credible interval"
    )
    ax.fill_between(
        xx,
        yy_lae,
        yy_le,
        color='green',
        alpha=0.25
    )
    ax.fill_between(
        xx,
        yy_ue,
        yy_uae,
        color='green',
        alpha=0.25,
        label="95% epistemic + aleatory credible interval"
```

```
    )
    ax.plot(x, y, 'kx', label='Observed data')
    if y_true is not None:
        ax.plot(xx, y_true, "--", label="True response")
    ax.set_xlabel('$x$')
    ax.set_ylabel('$y$')
    plt.legend(loc="upper left", frameon=False)
    sns.despine(trim=True)
xx = np.linspace(0, 0.25, y.size)
print(epsilon_l)
plot_posterior_predictive(
    model,
    xx,
    compute_design_matrix,
    phi_func_args=(epsilon_l,d),
    y_true=None
)
```

0.04
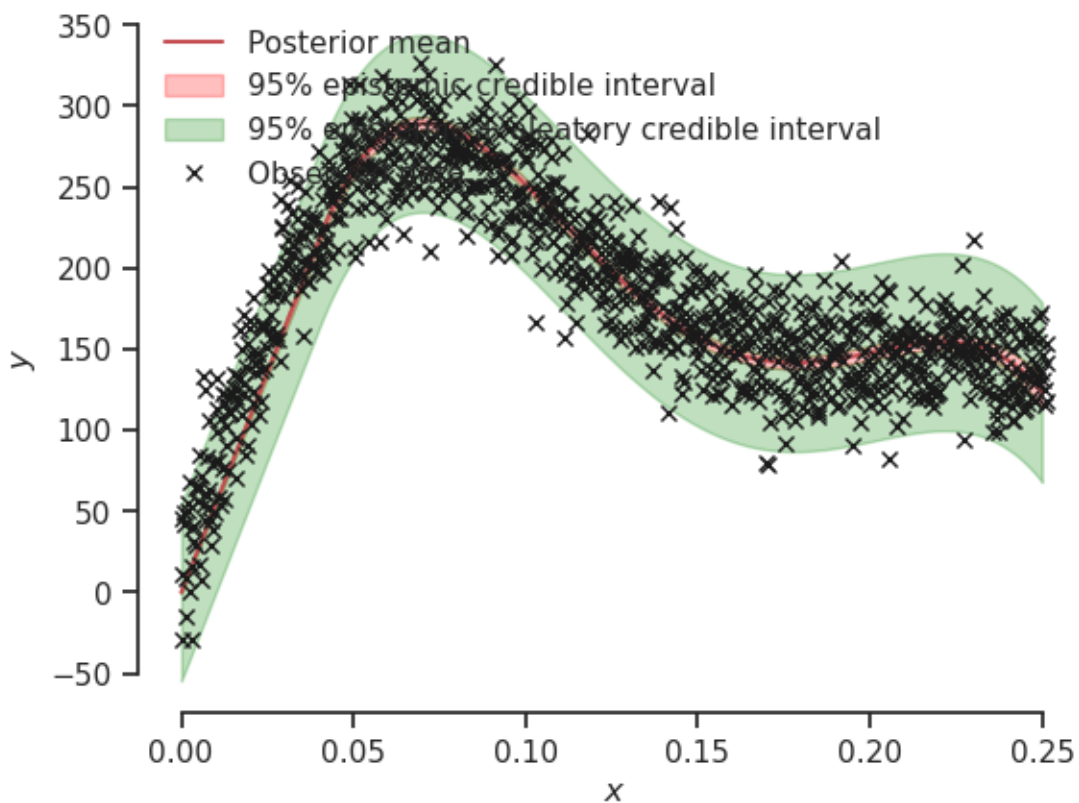
2.2.5 Subpart B.V - Extract the ultimate strength

Now, you are going to quantify your epistemic uncertainty about the ultimate strength. The ultimate strength is the maximum of the stress-strain relationship. Since you have epistemic uncertainty about the stress-strain relationship, you also have epistemic uncertainty about the ultimate strength.

Do the following: - Visualize posterior of the ultimate strength. - Find a 95% credible interval for the ultimate strength. - Pick a value for the ultimate strength.

Hint: To characterize your epistemic uncertainty about the ultimate strength, you would have to do the following: - Define a dense set of strain points between 0 and 0.25. - Repeatedly: + sample from the posterior of the weights of your model + for each sample evaluate the stresses at the dense set of strain points defined earlier + for each sampled stress vector, find the maximum. This is a sample of the ultimate strength.

```python
[ ]: xx = np.linspace(0,0.25,10000)
num_samples = 10000
Phi_xx = compute_design_matrix(xx,epsilon_l,d)
samples = np.zeros(num_samples)
m = model.coef_
S = model.sigma_
w_post = st.multivariate_normal(
    mean=m,
    cov=S + np.eye(S.shape[0])
)
for i in range(num_samples):
    w_sample = w_post.rvs()
    yy_sample = Phi_xx @ w_sample
    samples[i] = np.max(yy_sample)

strengthMean = np.mean(samples)
strengthSigma = np.var(samples)

strength_rv = st.norm(
    loc=strengthMean,
    scale=np.sqrt(strengthSigma)
)

ww = np.linspace(strengthMean - 3*np.sqrt(strengthSigma),strengthMean + 3*np.
  ↪sqrt(strengthSigma),1000)

fig, ax = plt.subplots()
ax.plot(
    ww,
    strength_rv.pdf(ww),
    label=f'$w_{{{1}}}$ posterior')
lb = wj_post.ppf(0.05)
ub = wj_post.ppf(0.95)
```
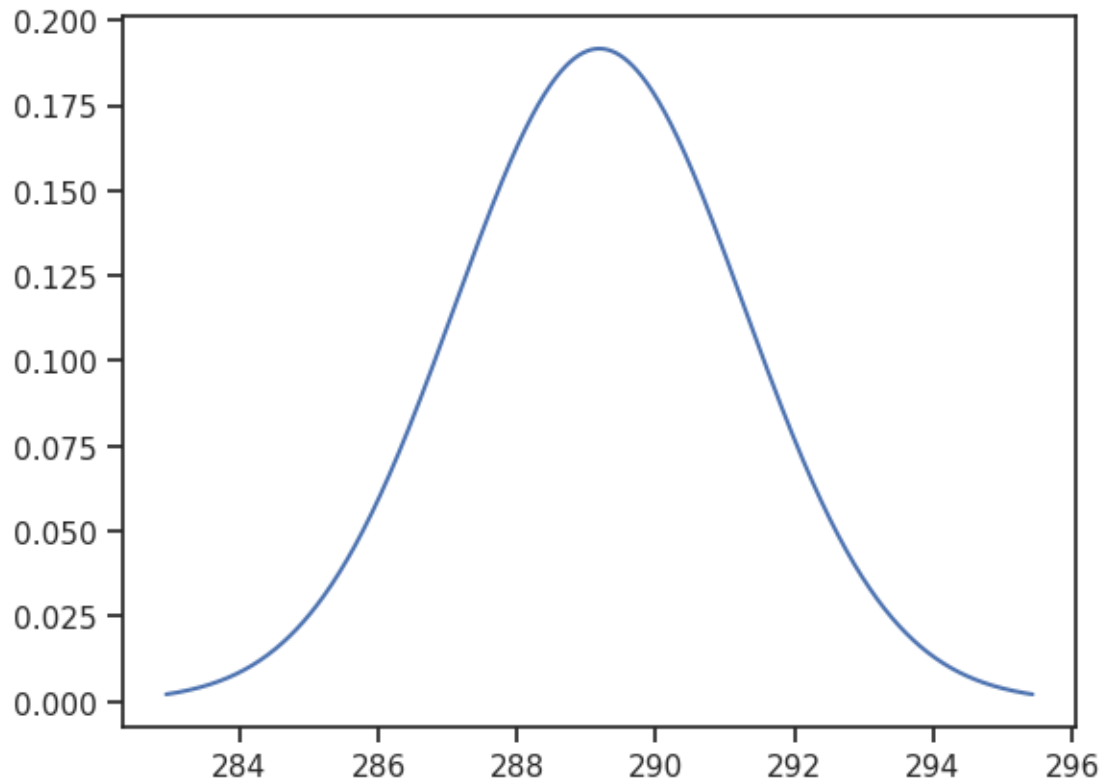
```
print(f'95% credibile interval [{lb}, {ub}], found using {num_samples} samples')
```

95% credibile interval [285.78374458582954, 292.58126472987334], found using
10000 samples

3   Problem 2 - Optimizing the performance of a compressor

In this problem we are going to need this dataset. The dataset was kindly provided to us by
Professor Davide Ziviani. As before, you can either put it on your Google drive or just download
it with the code segment below:

```
[ ]: url = "https://github.com/PredictiveScienceLab/data-analytics-se/raw/master/
     ↪lecturebook/data/compressor_data.xlsx"
     download(url)
```

Note that this is an Excell file, so we are going to need pandas to read it. Here is how:

```
[ ]: import pandas as pd
     data = pd.read_excel('compressor_data.xlsx')
     data
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
File ~/.local/lib/python3.10/site-packages/pandas/compat/_optional.py:132, in
 ↪import_optional_dependency(name, extra, errors, min_version)
    131 try:
--> 132     module = importlib.import_module(name)
    133 except ImportError:

File /usr/lib/python3.10/importlib/__init__.py:126, in import_module(name,
 ↪package)
    125         level += 1
--> 126 return _bootstrap._gcd_import(name[level:], package, level)

File <frozen importlib._bootstrap>:1050, in _gcd_import(name, package, level)

File <frozen importlib._bootstrap>:1027, in _find_and_load(name, import_)

File <frozen importlib._bootstrap>:1004, in _find_and_load_unlocked(name,
 ↪import_)

ModuleNotFoundError: No module named 'openpyxl'

During handling of the above exception, another exception occurred:

ImportError                               Traceback (most recent call last)
/home/stav/Documents/data-analytics-se/lecturebook/homework/homework-04.ipynb
 ↪Cell 60 line 2
       <a href='vscode-notebook-cell:/home/stav/Documents/data-analytics-se/
 ↪lecturebook/homework/homework-04.ipynb#Y112sZmlsZQ%3D%3D?line=0'>1</a> import
 ↪pandas as pd
----> <a href='vscode-notebook-cell:/home/stav/Documents/data-analytics-se/
 ↪lecturebook/homework/homework-04.ipynb#Y112sZmlsZQ%3D%3D?line=1'>2</a> data =
 ↪pd.read_excel('compressor_data.xlsx')
       <a href='vscode-notebook-cell:/home/stav/Documents/data-analytics-se/
 ↪lecturebook/homework/homework-04.ipynb#Y112sZmlsZQ%3D%3D?line=2'>3</a> data

File ~/.local/lib/python3.10/site-packages/pandas/io/excel/_base.py:504, in
 ↪read_excel(io, sheet_name, header, names, index_col, usecols, dtype, engine,
 ↪converters, true_values, false_values, skiprows, nrows, na_values,
 ↪keep_default_na, na_filter, verbose, parse_dates, date_parser, date_format,
 ↪thousands, decimal, comment, skipfooter, storage_options, dtype_backend,
 ↪engine_kwargs)
    502 if not isinstance(io, ExcelFile):
    503     should_close = True
--> 504     io = ExcelFile(
    505         io,
    506         storage_options=storage_options,
    507         engine=engine,
    508         engine_kwargs=engine_kwargs,
```

```
    509        )
    510 elif engine and engine != io.engine:
    511     raise ValueError(
    512         "Engine should not be specified when passing "
    513         "an ExcelFile - ExcelFile already has the engine set"
    514     )

File ~/.local/lib/python3.10/site-packages/pandas/io/excel/_base.py:1580, in␣
 ↪ExcelFile.__init__(self, path_or_buffer, engine, storage_options,␣
 ↪engine_kwargs)
   1577 self.engine = engine
   1578 self.storage_options = storage_options
-> 1580 self._reader = self._engines[engine](
   1581     self._io,
   1582     storage_options=storage_options,
   1583     engine_kwargs=engine_kwargs,
   1584 )

File ~/.local/lib/python3.10/site-packages/pandas/io/excel/_openpyxl.py:552, in␣
 ↪OpenpyxlReader.__init__(self, filepath_or_buffer, storage_options,␣
 ↪engine_kwargs)
    534 @doc(storage_options=_shared_docs["storage_options"])
    535 def __init__(
    536     self,
    (…)
    539     engine_kwargs: dict | None = None,
    540 ) -> None:
    541     """
    542     Reader using openpyxl engine.
    543
    (…)
    550         Arbitrary keyword arguments passed to excel engine.
    551     """
--> 552     import_optional_dependency("openpyxl")
    553     super().__init__(
    554         filepath_or_buffer,
    555         storage_options=storage_options,
    556         engine_kwargs=engine_kwargs,
    557     )

File ~/.local/lib/python3.10/site-packages/pandas/compat/_optional.py:135, in␣
 ↪import_optional_dependency(name, extra, errors, min_version)
    133 except ImportError:
    134     if errors == "raise":
--> 135         raise ImportError(msg)
    136     return None
    138 # Handle submodules: if we have submodule, grab parent module from sys.
 ↪modules
```

```
ImportError: Missing optional dependency 'openpyxl'.  Use pip or conda to␣
 ↪install openpyxl.
```

The data are part of a an experimental study of a variable speed reciprocating compressor. The experimentalists varied two temperatures T_e and T_c (both in C) and they measured various other quantities. Our goal is to learn the map between T_e and T_c and measured Capacity and Power (both in W). First, let's see how you can extract only the relevant data.

```
[ ]: # Here is how to extract the T_e and T_c columns and put them in a single numpy␣
      ↪array
     x = data[['T_e','T_c']].values
     x
```

```
[ ]: # Here is how to extract the Capacity
     y = data['Capacity'].values
     y
```

Fit the following multivariate polynomial model to **both the Capacity and the Power**:

$$y = w_1 + w_2 T_e + w_3 T_c + w_4 T_e T_c + w_5 T_e^2 + w_6 T_c^2 + w_7 T_e^2 T_c + w_8 T_e T_c^2 + w_9 T_e^3 + w_{10} T_c^3 + \epsilon,$$

where ϵ is a Gaussian noise term with unknown variance.

Hints: + You may use sklearn.preprocessing.PolynomialFeatures to construct the design matrix of your polynomial features. Do not program the design matrix by hand. + You should split your data into training and validation and use various validation metrics to make sure that your models make sense. + Use ARD Regression to fit any hyperparameters and the noise.

3.0.1 Subpart A.I - Fit the capacity

Please don't just fit blindly. Split in training and test and use all the usual diagnostics.

```
[ ]: # your code here - Repeat as many text and code blocks as you like
```

3.0.2 Subpart A.II

What is the noise variance you estimated for the Capacity?

```
[ ]: # your code here
```

3.0.3 Subpart A.III

Which features of the temperatures (basis functions of your model) are the most important for predicting the Capacity?

```
[ ]: # your code here
```

3.0.4 Subpart B.I - Fit the Power

Please don't just fit blindly. Split in training and test and use all the usual diagnostics.

```
[ ]: # your code here - Repeat as many text and code blocks as you like
```

3.0.5 Subpart B.II

What is the noise variance you estimated for the Power?

```
[ ]: # your code here
```

3.0.6 Subpart B.III

Which features of the temperatures (basis functions of your model) are the most important for predicting the Power?

```
[ ]: # your code here
```

4 Problem 3 - Explaining the challenger disaster

On January 28, 1986, the Space Shuttle Challenger disintegrated after 73 seconds from launch. The failure can be traced on the rubber O-rings which were used to seal the joints of the solid rocket boosters (required to force the hot, high-pressure gases generated by the burning solid propellant through the nozzles thus producing thrust).

It turns out that the performance of the O-ring material was particularly sensitive on the external temperature during launch. This dataset contains records of different experiments with O-rings recorded at various times between 1981 and 1986. Download the data the usual way (either put them on Google drive or run the code cell below).

```
[ ]: url = "https://github.com/PredictiveScienceLab/data-analytics-se/raw/master/
     ↪lecturebook/data/challenger_data.csv"
     download(url)
```

Even though this is a csv file, you should load it with pandas because it contains some special characters.

```
[ ]: raw_data = pd.read_csv('challenger_data.csv')
     raw_data
```

The first column is the date of the record. The second column is the external temperature of that day in degrees F. The third column labeled `Damage Incident` is has a binary coding (0=no damage, 1=damage). The very last row is the day of the Challenger accident.

We are going to use the first 23 rows to solve a binary classification problem that will give us the probability of an accident conditioned on the observed external temperature in degrees F. Before we proceed to the analysis of the data, let's clean the data up.

First, we drop all the bad records:

```
[ ]: clean_data_0 = raw_data.dropna()
     clean_data_0
```

We also don't need the last record. Just remember that the temperature the day of the Challenger accident was 31 degrees F.

```
[ ]: clean_data = clean_data_0[:-1]
     clean_data
```

Let's extract the features and the labels:

```
[ ]: x = clean_data['Temperature'].values
     x
```

```
[ ]: y = clean_data['Damage Incident'].values.astype(np.float)
     y
```

4.1 Part A - Perform logistic regression

Perform logistic regression between the temperature (x) and the damage label (y). Do not bother doing a validation because there are not a lot of data. Just use a very simple model so that you don't overfit.

```
[ ]: # your code here - Repeat as many text and code blocks as you like
```

4.2 Part B - Plot the probability of damage as a function of temperature

Plot the probability of damage as a function of temperature.

```
[ ]: # your code here
```

4.3 Part C - Decide whether or not to launch

The temperature the day of the Challenger accident was 31 degrees F. Start by calculating the probability of damage at 31 degrees F. Then, use formal decision-making (i.e., define a cost matrix and make decisions by minimizing the expected loss) to decide whether or not to launch on that day. Also, plot your optimal decision as a function of the external temperature.

```
[ ]: # your code here - Repeat as many text and code blocks as you like
```