Question 1 : Linear Regression

We will fit linear regression models to the data in file `regression_part1.csv`.

(a) Describe the main properties of the data, focusing on the size, data ranges, and data types.

Your Answer Here
size: (50, 2)
data ranges: revision_time: [2.723, 48.011] exam_score: [14.731, 94.945]
data types: revision_time: float64 exam_score: float64

Question 1 continued on next page. . .

(b) Fit a linear model to the data so that we can predict `exam_score` from `revision_time`. Report the estimated model parameters $\mathbf{w}$. Describe what the parameters represent for this 1D data. For this part, you should use the sklearn implementation of Linear Regression.

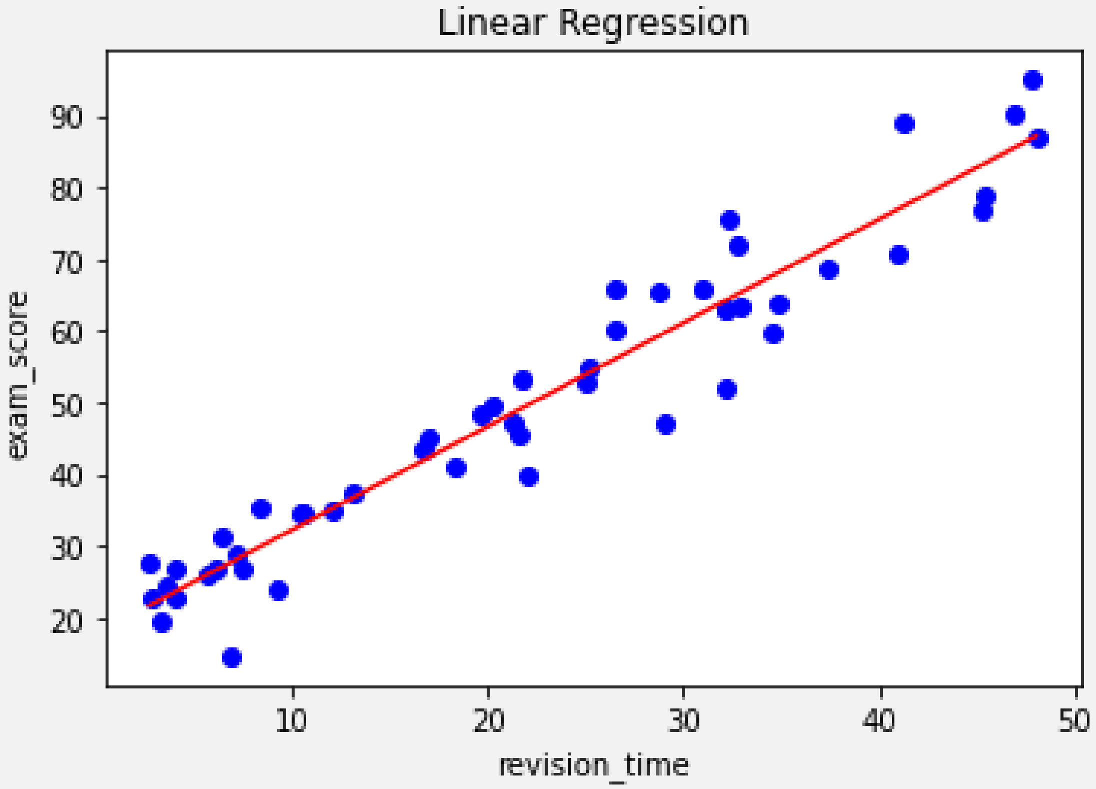
Hint: By default in sklearn `fit_intercept = True`*. Instead, set* `fit_intercept = False` *and pre-pend* 1 *to each value of* x_i *yourself to create* $\boldsymbol{\phi}(x_i) = [1, x_i]$.

Your Answer Here

$$\mathbf{w} = \begin{bmatrix} 17.8997 \\ 1.4411 \end{bmatrix}$$

(c) Display the fitted linear model and the input data on the same plot.

Your Answer Here
The fitted linear model below.



(d) Instead of using sklearn, implement the closed-form solution for fitting a linear regression model yourself using numpy array operations. Report your code in the answer box. It should only take a few lines (i.e. <5).
Hint: Only report the relevant lines for estimating **w** *e.g. we do not need to see the data loading code. You can write the code in the answer box directly or paste in an image of it.*

Your Answer Here

```
import numpy as np
poly = np.polyfit(x_true, y_true,deg=1)
y_pred = np.polyval(poly, x_true)
```

Question 1 continued on next page. . .

(e) Mean Squared Error (MSE) is a common metric used for evaluating the performance of regression models. Write out the expression for MSE and list one of its limitations. *Hint: For notation, you can use y for the ground truth quantity and $\hat{y}$ (`$\hat{y}$` in latex) in place of the model prediction.*

Your Answer Here

$MSE = \sum_{i=1}^{n} \frac{1}{n} w_i (y_i - \hat{y}_i)^2, w_i > 0$

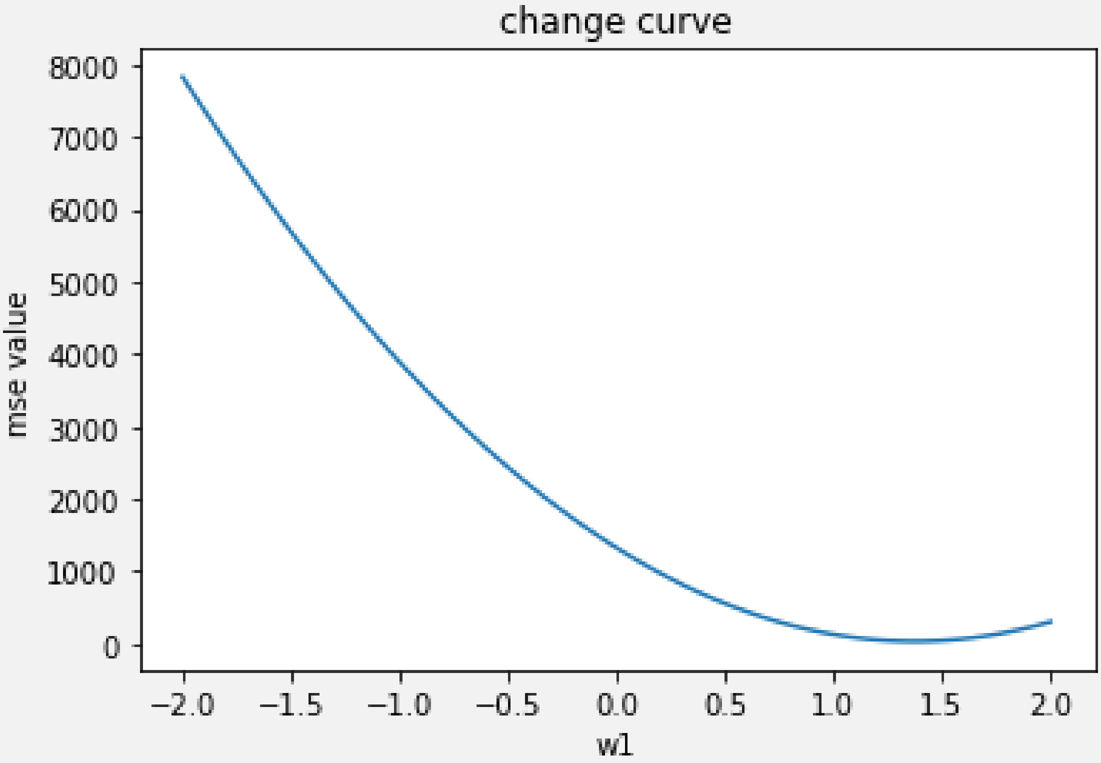
(f) Our next step will be to evaluate the performance of the fitted models using Mean Squared Error (MSE). Report the MSE of the data in `regression_part1.csv` for your prediction of `exam_score`. You should report the MSE for the linear model fitted using sklearn and the model resulting from your closed-form solution. Comment on any differences in their performance.

Your Answer Here
MSE for the linear model: 30.985
MSE for the model resulting from your closed-form solution: 30.985

Question 1 continued on next page. . .

(g) Assume that the optimal value of w_0 is 20, it is not but let's assume so for now. Create a plot where you vary w_1 from -2 to $+2$ on the horizontal axis, and report the Mean Squared Error on the vertical axis for each setting of $\mathbf{w} = [w_0, w_1]$ across the dataset. Describe the resulting plot. Where is its minimum? Is this value to be expected?
Hint: You can try 100 values of w_1 i.e. `w1 = np.linspace(-2,2, 100)`.

Your Answer Here
The MSE curve below.



When $w_1 = 1.354$, the MSE value reaches its minimum.
The minimum MSE value is 32.481.