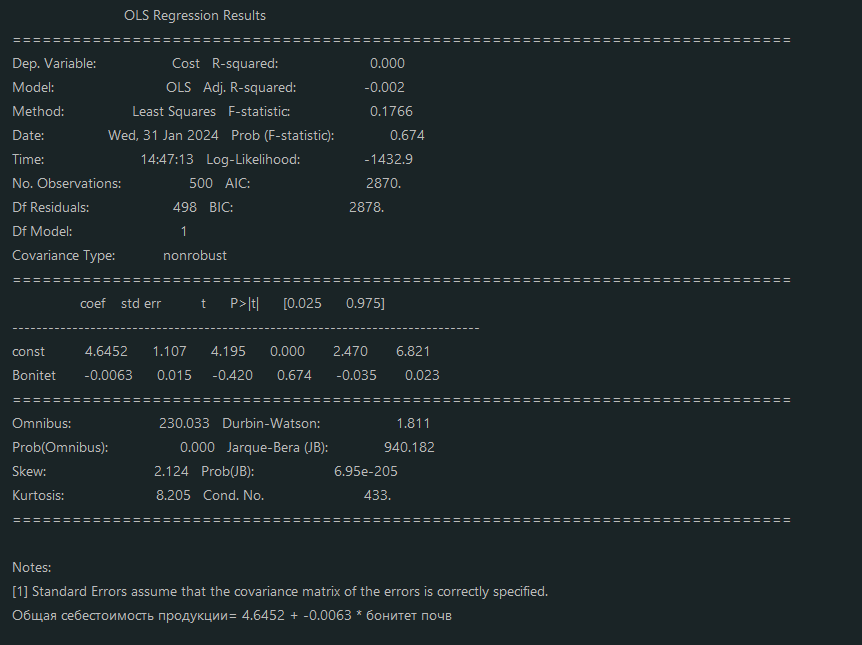
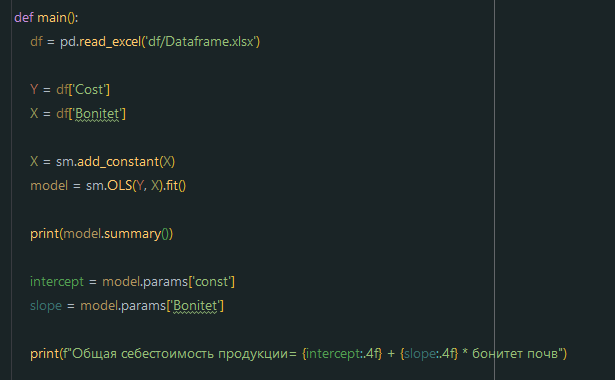
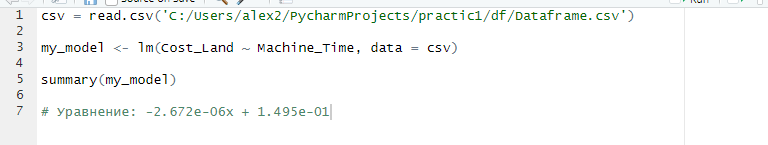
Вариант 3

1. Постройте в Python уравнение линейной регрессии по представленному датафрейму, если в качестве зависимой переменной выступает общая себестоимость продукции(Cost), независимой переменной - бонитет почв(Bonitet)



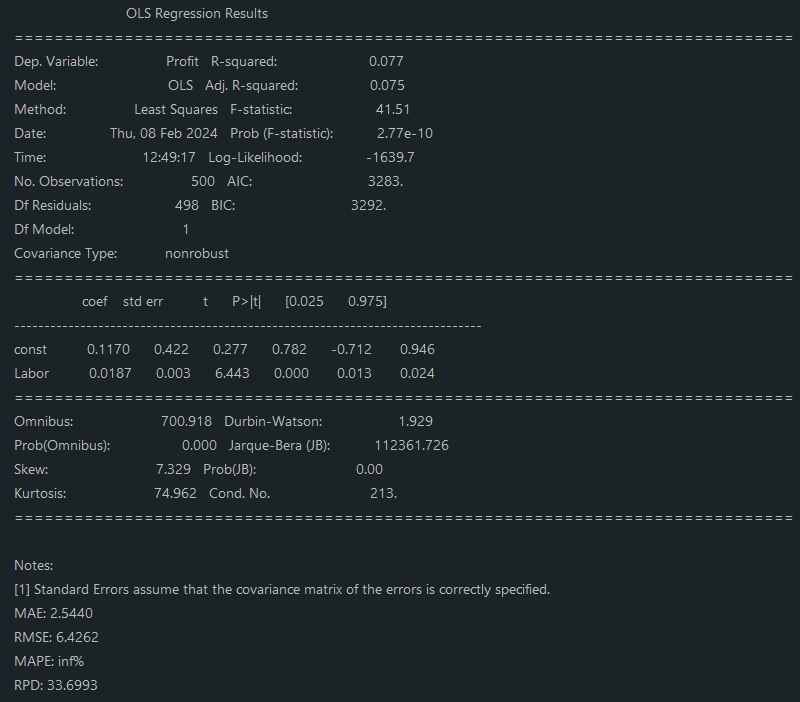


2. Постройте в R уравнение линейной регрессии по представленному датафрейму, а также рассчитайте в обоих программах все возможные классические метрики ошибок; в качестве зависимой переменной выступает величина затрат на сельскохозяйственные угодия(Cost\_Land), независимой переменной - общее время работы машин на сельскохозяйственных работах(Machine\_Time)



3. Постройте в R и Python два уравнения линейной регрессии по представленному датафрейму, в качестве зависимой переменной примите общая величина прибыли(Profit). Независимой переменной в первой модели будет выступать средняя температура почвы в вегетативный период(Temp), во второй модели - количество работников(Labor). Рассчитайте для обоих моделей все возможные классические метрики ошибок и сделайте вывод о том, какая модель лучшая.

import pandas as pd  
import statsmodels.api as sm  
from sklearn.metrics import mean\_absolute\_error, mean\_squared\_error  
import numpy as np  
  
  
def main():  
 df = pd.read\_excel('df/Dataframe.xlsx')  
 Y = df['Profit']  
 X = df['Labor']  
  
 X = sm.add\_constant(X)  
 model = sm.OLS(Y, X).fit()  
  
 print(model.summary())  
  
 Y\_pred = model.predict(X)  
 residuals = Y - Y\_pred  
 mae = mean\_absolute\_error(Y, Y\_pred)  
 rmse = np.sqrt(mean\_squared\_error(Y, Y\_pred))  
 mape = np.mean(np.abs((Y - Y\_pred) / Y)) \* 100  
 rpd = np.max(Y) / np.mean(np.abs(residuals))  
  
 print(f"MAE: {mae:.4f}")  
 print(f"RMSE: {rmse:.4f}")  
 print(f"MAPE: {mape:.4f}%")  
 print(f"RPD: {rpd:.4f}")  
  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 main()



====================================================================================

library(yardstick)

csv = read.csv("C:/Users/alex2/PycharmProjects/practic1/df/Dataframe.csv")

model\_first <- lm(Profit ~ Temp, data = csv)

summary(model\_first)

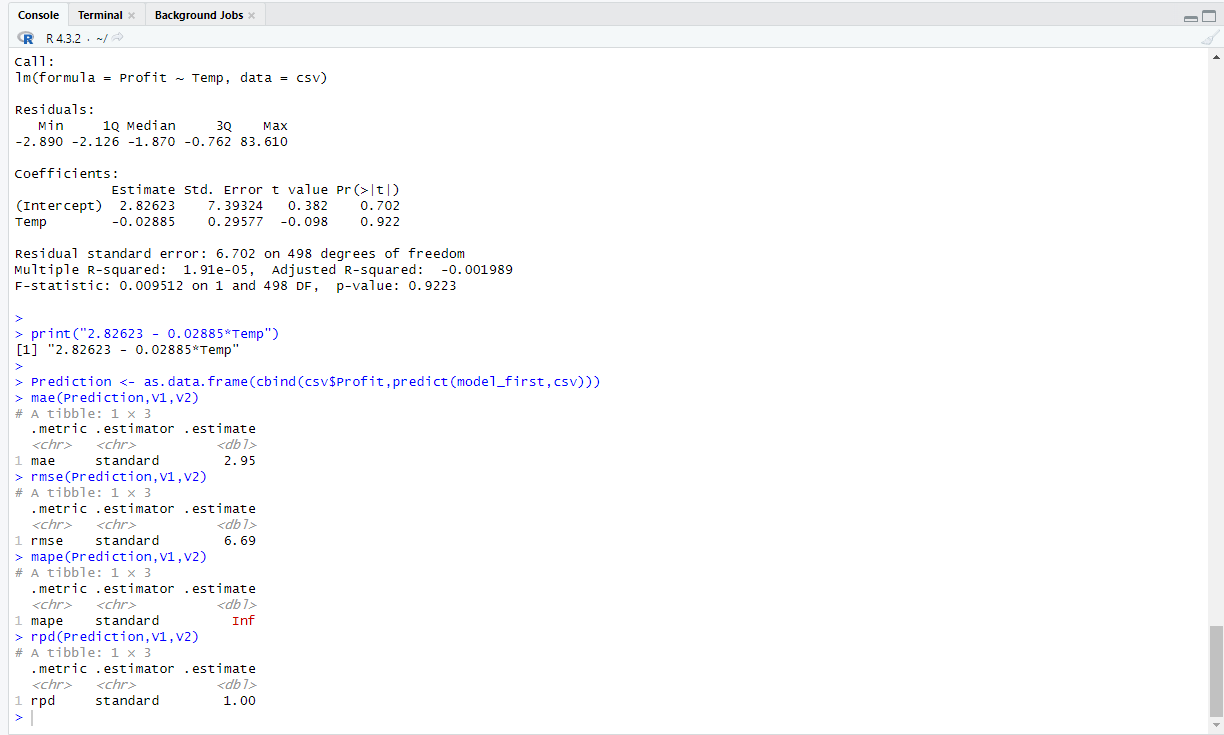
print("2.82623 - 0.02885\*Temp")

Prediction <- as.data.frame(cbind(csv$Profit,predict(model\_first,csv)))

mae(Prediction,V1,V2)

rmse(Prediction,V1,V2)

mape(Prediction,V1,V2)

rpd(Prediction,V1,V2)  
  


4. Постройте в R и Python уравнение регрессии как в задании 3, только объедините обе независимые переменные в третьей модели. Рассчитайте для нее все возможные классические метрики ошибок и сделайте вывод о том, какая модель лучшая, учитывая результаты задания 3.  
====================================================================================

library(yardstick)

csv = read.csv("C:/Users/alex2/PycharmProjects/practic1/df/Dataframe.csv")

model\_first <- lm(Profit ~ Temp + Labor, data = csv)

summary(model\_first)

print("0.468813 - 0.014083\*Temp + 0.018718\*Labor")

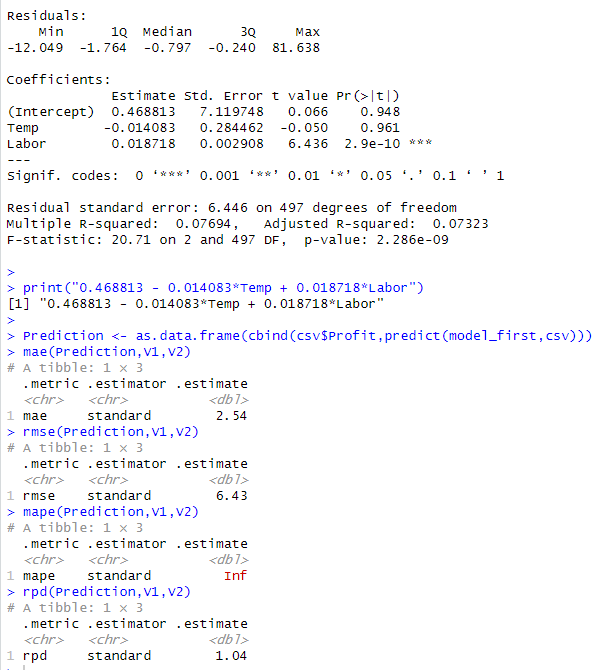
Prediction <- as.data.frame(cbind(csv$Profit,predict(model\_first,csv)))

mae(Prediction,V1,V2)

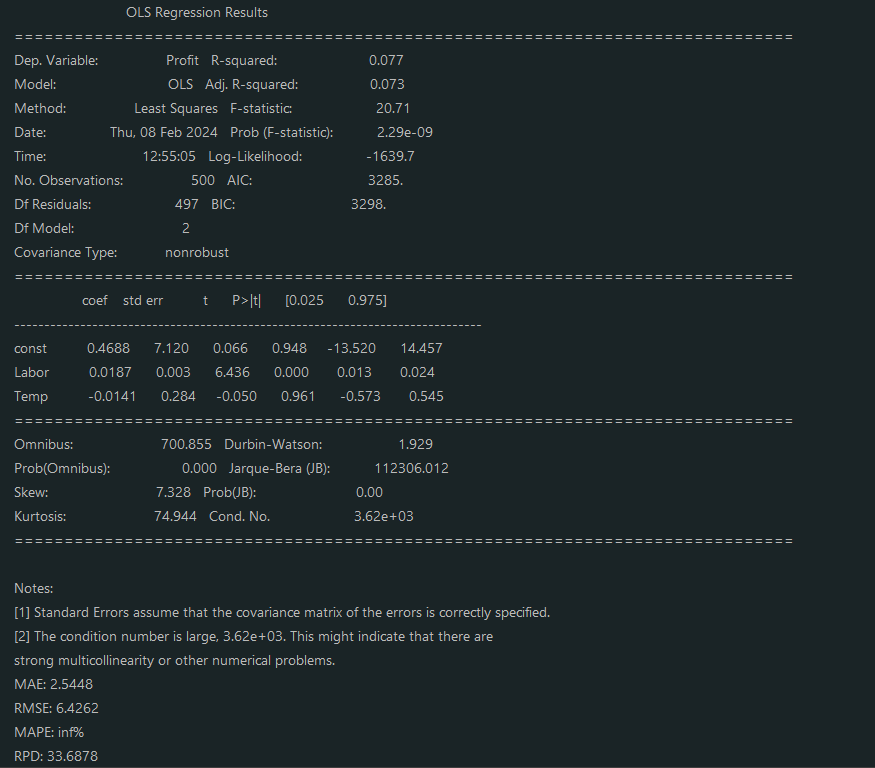
rmse(Prediction,V1,V2)

mape(Prediction,V1,V2)

rpd(Prediction,V1,V2)

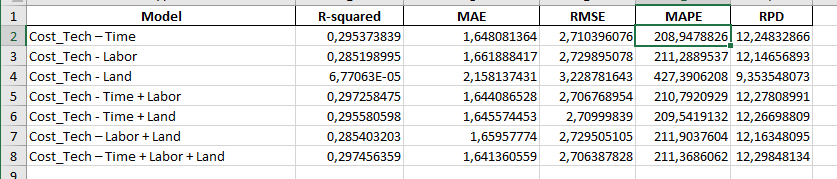


import pandas as pd  
import statsmodels.api as sm  
from sklearn.metrics import mean\_absolute\_error, mean\_squared\_error  
import numpy as np  
  
  
def main():  
 df = pd.read\_excel('df/Dataframe.xlsx')  
 Y = df['Profit']  
 X = df[['Labor', 'Temp']]  
  
 X = sm.add\_constant(X)  
 model = sm.OLS(Y, X).fit()  
  
 print(model.summary())  
  
 Y\_pred = model.predict(X)  
 residuals = Y - Y\_pred  
 mae = mean\_absolute\_error(Y, Y\_pred)  
 rmse = np.sqrt(mean\_squared\_error(Y, Y\_pred))  
 mape = np.mean(np.abs((Y - Y\_pred) / Y)) \* 100  
 rpd = np.max(Y) / np.mean(np.abs(residuals))  
  
 print(f"MAE: {mae:.4f}")  
 print(f"RMSE: {rmse:.4f}")  
 print(f"MAPE: {mape:.4f}%")  
 print(f"RPD: {rpd:.4f}")  
  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 main()



5. Ваш начальник считает, что на переменную величина затрат на технику(Cost\_Tech) линейно могут влиять три фактора: время работы сельскохозяйственного персонала(Time), количество работников(Labor), общая площадь посевов(Land). Найдите лучшую модель из возможных, включающую данные переменные в качестве независимых. Язык - на ваш выбор.

import pandas as pd  
import statsmodels.api as sm  
from sklearn.metrics import mean\_absolute\_error, mean\_squared\_error, r2\_score  
import numpy as np  
  
  
def calculate\_metrics(X, Y):  
 X = sm.add\_constant(X)  
 model = sm.OLS(Y, X).fit()  
 Y\_pred = model.predict(X)  
 residuals = Y - Y\_pred  
 r2 = r2\_score(Y, Y\_pred)  
 mae = mean\_absolute\_error(Y, Y\_pred)  
 rmse = np.sqrt(mean\_squared\_error(Y, Y\_pred))  
 mape = np.mean(np.abs((Y - Y\_pred) / Y)) \* 100  
 rpd = np.max(Y) / np.mean(np.abs(residuals))  
 return r2, mae, rmse, mape, rpd  
  
  
def main():  
 df = pd.read\_excel('df/Dataframe.xlsx')  
 results = []  
 feature\_combinations = {  
 'Cost\_Tech – Time': ['Time'],  
 'Cost\_Tech - Labor': ['Labor'],  
 'Cost\_Tech - Land': ['Land'],  
 'Cost\_Tech - Time + Labor': ['Time', 'Labor'],  
 'Cost\_Tech - Time + Land': ['Time', 'Land'],  
 'Cost\_Tech – Labor + Land': ['Labor', 'Land'],  
 'Cost\_Tech – Time + Labor + Land': ['Time', 'Labor', 'Land']  
 }  
  
 for title, features in feature\_combinations.items():  
 X = df[features]  
 Y = df['Cost\_Tech']  
 r2, mae, rmse, mape, rpd = calculate\_metrics(X, Y)  
 results.append([title, r2, mae, rmse, mape, rpd])  
 df\_results = pd.DataFrame(results, columns=['Model', 'R-squared', 'MAE', 'RMSE', 'MAPE', 'RPD'])  
 print(df\_results)  
 df\_results.to\_csv('metrics.csv', index=False)  
  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 main()



6. Посмотрите ноутбук с решением задачи линейной регрессии по ссылке https://www.kaggle.com/code/mohamedelnahry/salary-prediction-lr. Укажите, какая модель была получена, и какие показатели качества по ней были рассчитаны. Оцените качество ноутбука.

DataFrame – Salary Dataset.  
P.S… С

7. У вас есть база данных https://www.kaggle.com/datasets/mayurdalvi/simple-linear-regression-placement-data. Оцените, какая переменная должна считаться зависимой, какая - независимой, постройте в R линейное уравнение регрессии, выведите все классические показатели качества. Разместите модель на kaggle.

8. У вас есть база данных https://www.kaggle.com/datasets/shubham47/students-score-dataset-linear-regression. Оцените, какая переменная должна считаться зависимой, какая - независимой, постройте в Python линейное уравнение регрессии, выведите все классические показатели качества. Разместите модель на kaggle.