In [2]:
```
import pandas as pd
import numpy as np

raw = pd.read_csv("boston.txt", sep=r"\s+", skiprows=22, header=None)
X = np.hstack([raw.values[::2, :], raw.values[1::2, :2]])
y = raw.values[1::2, 2]
columns = [ "crim", "zn", "indus", "chas", "nox", "rm", "age",
"dis", "rad", "tax", "ptratio", "b", "lstat"]
df = pd.DataFrame(X, columns=columns)
df["medv"] = y
```

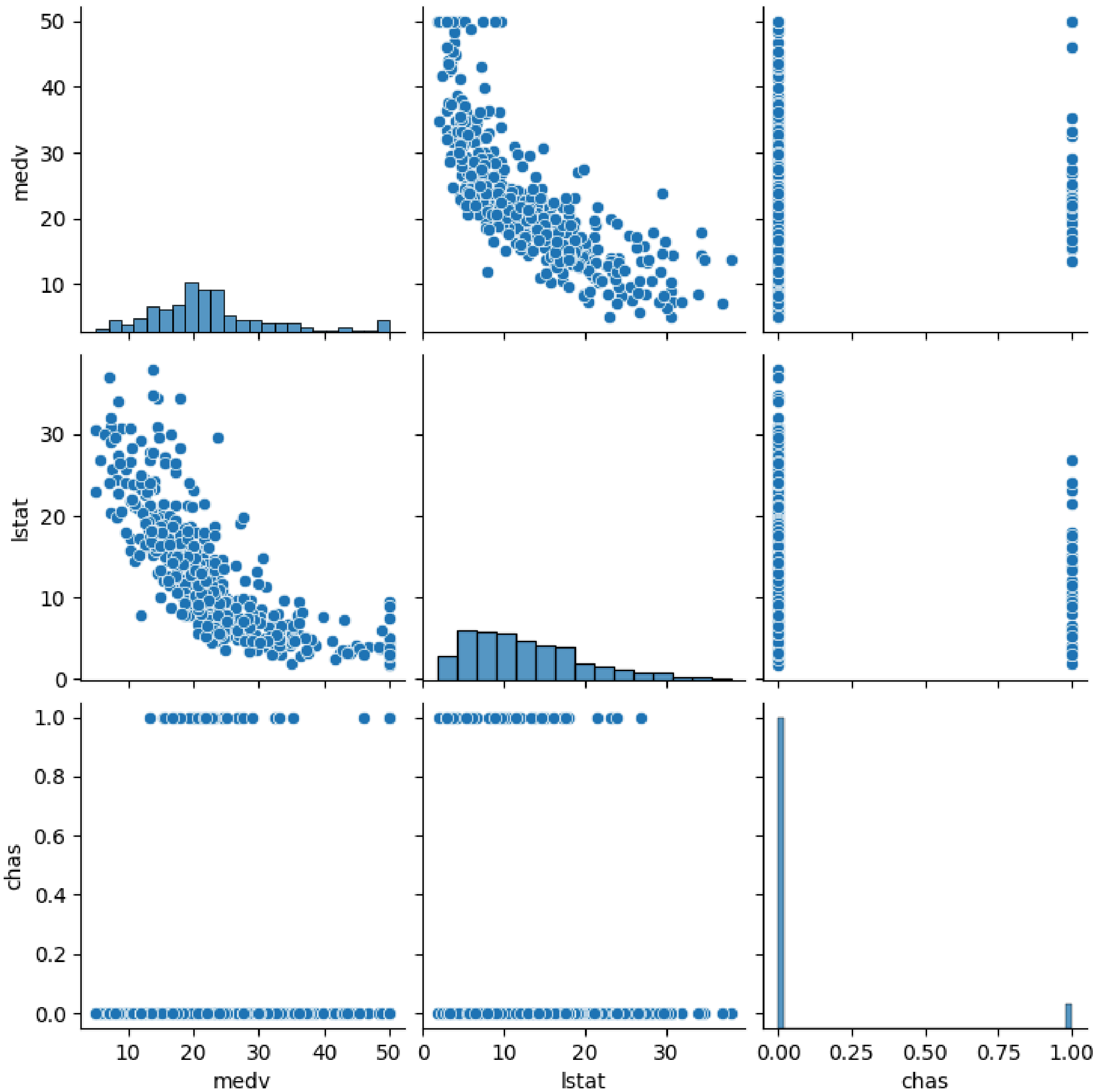
a

In [5]:
```
import matplotlib.pyplot as plt
from seaborn import pairplot
import statsmodels.api as sm

pairplot(df[["medv", "lstat", "chas"]])
X = sm.add_constant(df[["lstat", "chas"]])
model = sm.OLS(df["medv"], X).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   medv   R-squared:                       0.563
Model:                            OLS   Adj. R-squared:                  0.561
Method:                 Least Squares   F-statistic:                     323.4
Date:                Tue, 10 Feb 2026   Prob (F-statistic):           4.93e-91
Time:                        17:06:14   Log-Likelihood:                -1631.1
No. Observations:                 506   AIC:                             3268.
Df Residuals:                     503   BIC:                             3281.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         34.0941      0.561     60.809      0.000      32.993      35.196
lstat         -0.9406      0.038    -24.729      0.000      -1.015      -0.866
chas           4.9200      1.069      4.601      0.000       2.819       7.021
==============================================================================
Omnibus:                      131.896   Durbin-Watson:                   0.975
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              275.510
Skew:                           1.406   Prob(JB):                     1.49e-60
Kurtosis:                       5.272   Cond. No.                         57.8
==============================================================================

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



The pair plots indicate a non linear relationship between `lstat` and `medv`, `medv` drops quickly for low-mid `lstat`, and then flattens out. Usually taking the log, square root, or inverse of `lstat` can help fix this. For this problem, we will use a log transformation.

Intercept: 34.0941, SE 0.561, $p < 0.0001$
This is the predicted `medv` for an `lstat` and `chas` value of 0. This isn't meaningful realistically as `lstat` 0 isn't possible, but this anchors the regression line. The value is statistically significant and well estimated from the low SE and p value.

`lstat` **coefficient**: -0.9406, SE 0.038, $p < 0.0001$
Holding `chas` constant, a 1 percentage point increase in `lstat` is attributed to about a $940.6 decrease in `medv`. The value is statistically significant and well estimated from the low SE and p value.

`chas` **coefficient**: 4.92, SE 1.069, $p < 0.0001$ Holding `lstat` constant, properties that border the Charles River are attributed to a $4,920 increase in `medv`. The value is statistically significant and well estimated from the low SE and p value.

b

In [6]:
```
# add interaction term to model

df["lstat_chas"] = df["lstat"] * df["chas"]
X = sm.add_constant(df[["lstat", "chas", "lstat_chas"]])
model_interaction = sm.OLS(df["medv"], X).fit()
print(model_interaction.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                   medv   R-squared:                       0.569
Model:                            OLS   Adj. R-squared:                  0.566
Method:                 Least Squares   F-statistic:                     220.8
Date:                Tue, 10 Feb 2026   Prob (F-statistic):           2.68e-91
Time:                        17:16:58   Log-Likelihood:                -1627.4
No. Observations:                 506   AIC:                             3263.
Df Residuals:                     502   BIC:                             3280.
Df Model:                           3                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         33.7672      0.570     59.222      0.000      32.647      34.887
lstat         -0.9150      0.039    -23.478      0.000      -0.992      -0.838
chas           9.8251      2.103      4.672      0.000       5.693      13.957
lstat_chas    -0.4329      0.160     -2.703      0.007      -0.748      -0.118
==============================================================================
Omnibus:                      130.570   Durbin-Watson:                   1.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              275.834
Skew:                           1.383   Prob(JB):                     1.27e-60
Kurtosis:                       5.330   Cond. No.                         114.
==============================================================================

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

Intercept: 33.7672, SE 0.570, p < 0.0001
This is the predicted `medv` for an `lstat` and `chas` value of 0. This isn't meaningful realistically as `lstat` 0 isn't possible, but this anchors the regression line. The value is statistically significant and well estimated from the low SE and p value.

`lstat` coefficient: -0.9150, SE 0.039, p < 0.0001
Holding `chas` constant, a 1 percentage point increase in `lstat` is attributed to about a $915 decrease in `medv`. The value is statistically significant and well estimated from the low SE and p value. The coefficient is slightly smaller than the model without the interaction term.

`chas` **coefficient**: 9.8251, SE 2.103, p < 0.0001 Holding `lstat` constant, properties that border the Charles River are attributed to a $9,825.10 increase in `medv`. The value is statistically significant and well estimated from the low SE and p value. The coefficient is significantly larger than the model without the interaction term.

`lstat:chas` **coefficient**: -0.4329, SE 0.160, p 0.007
The negative effect of `lstat` is stronger for properties bordering the Charles River. The value is statistically significant and well estimated from the low SE and p value. Specifically, for properties that border the Charles River, a 1 percentage point increase in `lstat` decreased `medv` by $1,347.90 compared to the$915 elsewhere.

Given that the new R squared value of this model (0.569, 0.566) is barely an improvement from the previous model (0.563, 0.561), interactions do not improve the model, but the term is real and statistically meaningful.

C

In [7]:
```
# fit two separate models of medv as a function of lstat
# one model for chas = 0 and one model for chas = 1
# compare to joint interaction model to determine which is better

df0 = df[df["chas"] == 0]
df1 = df[df["chas"] == 1]

X0 = sm.add_constant(df0["lstat"])
model0 = sm.OLS(df0["medv"], X0).fit()
print(model0.summary())

X1 = sm.add_constant(df1["lstat"])
model1 = sm.OLS(df1["medv"], X1).fit()
print(model1.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                   medv   R-squared:                       0.552
Model:                            OLS   Adj. R-squared:                  0.551
Method:                 Least Squares   F-statistic:                     577.2
Date:                Tue, 10 Feb 2026   Prob (F-statistic):           9.59e-84
Time:                        17:25:11   Log-Likelihood:                -1504.9
No. Observations:                 471   AIC:                             3014.
Df Residuals:                     469   BIC:                             3022.
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         33.7672      0.557     60.603      0.000      32.672      34.862
lstat         -0.9150      0.038    -24.026      0.000      -0.990      -0.840
==============================================================================
Omnibus:                      139.989   Durbin-Watson:                   1.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              330.529
Skew:                           1.523   Prob(JB):                     1.68e-72
Kurtosis:                       5.751   Cond. No.                         30.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
cified.
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                   medv   R-squared:                       0.582
Model:                            OLS   Adj. R-squared:                  0.569
Method:                 Least Squares   F-statistic:                     45.90
Date:                Tue, 10 Feb 2026   Prob (F-statistic):           1.01e-07
Time:                        17:25:11   Log-Likelihood:                -120.33
No. Observations:                  35   AIC:                             244.7
Df Residuals:                      33   BIC:                             247.8
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         43.5923      2.593     16.814      0.000      38.318      48.867
lstat         -1.3479      0.199     -6.775      0.000      -1.753      -0.943
==============================================================================
Omnibus:                        1.398   Durbin-Watson:                   1.269
Prob(Omnibus):                  0.497   Jarque-Bera (JB):                1.349
Skew:                           0.395   Prob(JB):                        0.509
Kurtosis:                       2.453   Cond. No.                         25.9
==============================================================================

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

Model	n	R^2	Adj R^2	AIC	BIC
Interaction	506	0.569	0.566	3263	3281

Model	n	R^2	Adj R^2	AIC	BIC
`chas` = 0	471	0.552	0.551	3014	3022
`chas` = 1	35	0.582	0.569	244.7	247.8

Note that the slopes for each `chas` specific model are the same as the indicated slopes in the interaction model when accounting for the interaction term. Additionally, the `chas` = 1 model only has 35 samples. Thus, we prefer the interaction model due to it using all 506 observations, and provides a direct statistical test for slope differences in the `chas` specific models with better estimates. The comparison given by the model is straighforward and captures the majority of the information given by both groups.

d

In [ ]:
```
df_restrict = df[df["rad"] <= 8]

X_num = sm.add_constant(df_restrict[["lstat", "rad"]].astype(float))
model_num = sm.OLS(df_restrict["medv"].astype(float), X_num).fit()
print(model_num.summary())

rad_dummies = pd.get_dummies(df_restrict["rad"], prefix="rad")
X_cat = sm.add_constant(pd.concat([df_restrict["lstat"].astype(float), rad_dummies.
model_cat = sm.OLS(df_restrict["medv"].astype(float), X_cat).fit()
print(model_cat.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   medv   R-squared:                       0.458
Model:                            OLS   Adj. R-squared:                  0.455
Method:                 Least Squares   F-statistic:                     156.7
Date:                Tue, 10 Feb 2026   Prob (F-statistic):           4.77e-50
Time:                        17:37:08   Log-Likelihood:                -1212.6
No. Observations:                 374   AIC:                             2431.
Df Residuals:                     371   BIC:                             2443.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         33.0491      1.096     30.164      0.000      30.895      35.204
lstat         -0.9524      0.054    -17.592      0.000      -1.059      -0.846
rad            0.3819      0.197      1.937      0.054      -0.006       0.770
==============================================================================
Omnibus:                       85.140   Durbin-Watson:                   0.926
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              147.053
Skew:                           1.313   Prob(JB):                     1.17e-32
Kurtosis:                       4.593   Cond. No.                         44.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
cified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                   medv   R-squared:                       0.500
Model:                            OLS   Adj. R-squared:                  0.489
Method:                 Least Squares   F-statistic:                     45.67
Date:                Tue, 10 Feb 2026   Prob (F-statistic):           1.40e-50
Time:                        17:37:08   Log-Likelihood:                -1197.4
No. Observations:                 374   AIC:                             2413.
Df Residuals:                     365   BIC:                             2448.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         30.5985      0.585     52.271      0.000      29.447      31.750
lstat         -0.9116      0.054    -16.761      0.000      -1.019      -0.805
rad_1.0        0.4850      1.239      0.391      0.696      -1.952       2.922
rad_2.0        5.3733      1.141      4.708      0.000       3.129       7.618
rad_3.0        5.6042      0.930      6.026      0.000       3.775       7.433
rad_4.0        1.9095      0.646      2.958      0.003       0.640       3.179
rad_5.0        4.8236      0.617      7.824      0.000       3.611       6.036
rad_6.0        1.5964      1.116      1.430      0.154      -0.599       3.792
rad_7.0        3.7895      1.335      2.839      0.005       1.165       6.414
rad_8.0        7.0170      1.139      6.159      0.000       4.777       9.257
==============================================================================
Omnibus:                       93.551   Durbin-Watson:                   1.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              181.753
Skew:                           1.350   Prob(JB):                     3.41e-40
Kurtosis:                       5.093   Cond. No.                     1.10e+17
```

```
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spe
cified.
[2] The smallest eigenvalue is 4.55e-30. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Model	n	R^2	Adj R^2	AIC	BIC
Numerical `rad`	374	0.458	0.455	2431	2443
Categorical `rad`	374	0.500	0.489	2413	2448

Using `rad` as categorical significantly improves R^2 by 0.4. The AIC strongly favors the categorical specification, as the delta of 18 shows that allowing non-linear, level-specific effects of rad improves the fit. BIC slightly favors numerical specification however, as rad is heavily penalized with extra parameters as a categorical variable. Looking at the coefficients, the numerical rad value is barely statistically significant with a p-value of 0.054. However for the categorical model, different rads have varying statistical significance levels. This matches how rad is constructed realistically as it is an index, not a continuous economic quantity. Thus overall, we prefer the categorical specification due to higher adjusted R^2 and better representation of the predictor variables.

e

In [13]:

```python
b3 = model_cat.params["rad_3.0"]
b4 = model_cat.params["rad_4.0"]
diff = b3 - b4
print(f"Estimated difference in expected value of medv between houses w/ rad = 3 an

cov = model_cat.cov_params()
var_diff = (
    cov.loc["rad_3.0", "rad_3.0"]
    + cov.loc["rad_4.0", "rad_4.0"]
    - 2 * cov.loc["rad_3.0", "rad_4.0"]
)
se_diff = np.sqrt(var_diff)
print(f"Estimated standard error of the difference: {se_diff}")
```

```
Estimated difference in expected value of medv between houses w/ rad = 3 and rad =
4: 3.6947186903988825
Estimated standard error of the difference: 1.1451240401723881
```