In [2]:
```
import pandas as pd
from scipy import stats
import statsmodels.api as sm
import matplotlib.pyplot as plt
import numpy as np

```

The college placement office is developing a model to relate grade point average (GPA) to starting salary for liberal arts majors. Twenty recent graduates have been randomly selected and their GPAs and starting salaries are provided in HW10.xslx.

a. Fit a simple linear regression model to the data. What is the estimate for $\beta 0$ and $\beta 1$? Give the fitted regression model.

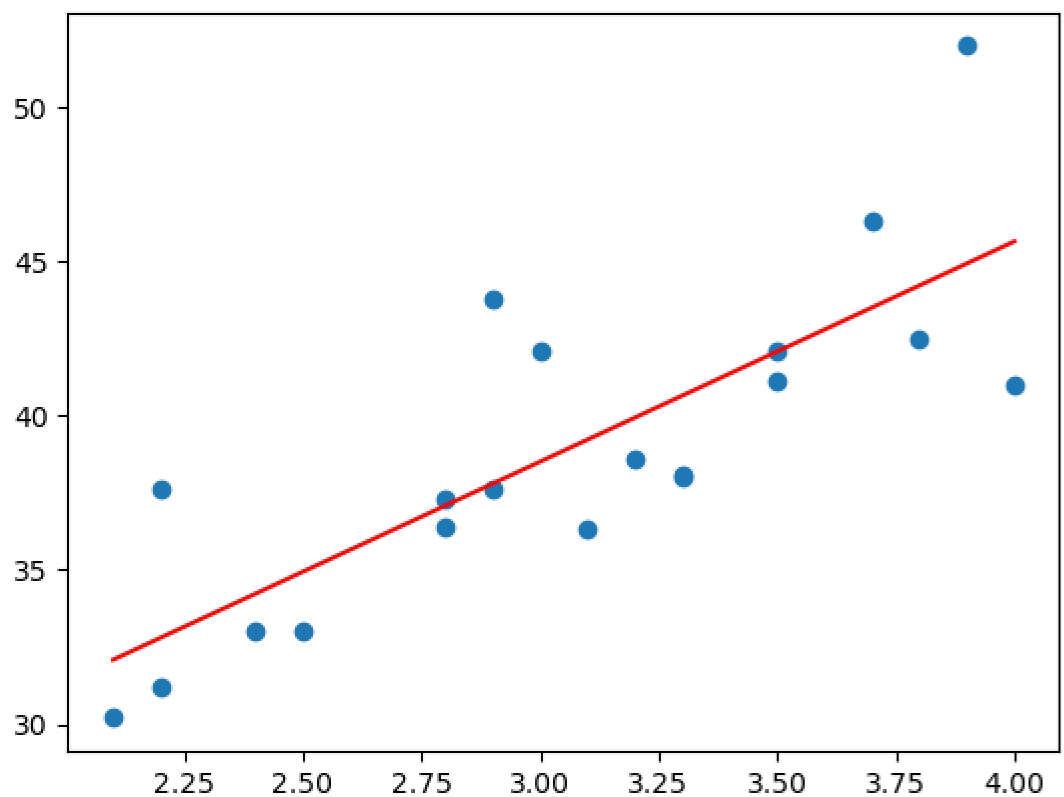
In [7]:

```
data = pd.read_excel("HW #11.xlsx",sheet_name='Problem 1')
data = data.sort_values(by='GPA')
x = np.array(data['GPA'])
y = np.array(data['Predicted Starting Salary (K)'])

slope, intercept, r_value, p_value, std_err = stats.linregress(x,y)

x_hat = np.linspace(x[0],x[-1])
y_hat = [slope*x_i+intercept for x_i in x_hat]
y_pred = [slope*x_i+intercept for x_i in x]
residuals = y - y_pred


plt.scatter(x,y)
plt.plot(x_hat,y_hat,'r')
plt.show()
```



$$\beta_1 = slope = 7.1439$$
$$\beta_0 = intercept = 17.0853$$
$$model:$$
$$\hat{y}_i = \beta_1 * x_i + \beta_0 = 7.1439 * x_i + 17.0853$$

---

b. Test the hypothesis $H0$:$\beta1$ = 0. Include the alternative hypothesis, test statistic, critical value, p-value, decision, and conclusion (with context to the problem).

In [8]:
```
stats.t.sf((slope/std_err),18)
```

Out[8]: 1.4296883024464914e-05

$$H_1 : \beta_1 \neq 0$$
$$Test\ Statistic = t_0 = \frac{\beta_1}{se_{(b1)}} = \frac{7.1439}{1.2868} = 5.55$$
$$Critical\ Value = t_{\frac{\alpha}{2},n-2} = 2.1$$
$$5.55 > 2.1$$
$$P = 2 * 1.42 * .00001 = 0.0000284$$

Since P value is very small we can reject the H0 and we can conclude that GPA of students has a role on their starting salary

---

c. What is $R2$? Interpret this value.

In [9]:
```
r_value
```

Out[9]: 0.7945440488521425

$$R^2 = 0.6313$$
$$R = \sqrt{R^2} = .7945$$

The regression model is explaining the 79.45% of the variability in the salary

---

d. Find the estimate for mean starting salary when GPA is 3.6.

In [10]:

```
7.1439 * 3.6 + 17.0853
```

Out[10]:

```
42.803340000000006
```

Since 3.6 is in the range of the GPA provided in the data we can have:

$$y = \beta_1 * x_i + \beta_0 = 7.1439 * x_i + 17.0853$$
$$y = 7.1439 * 3.6 + 17.0853 = 42.8033$$

---

e. What is $\hat{\sigma}2$?

In [11]:

```
sse = np.sum([i**2 for i in residuals])
print(sse)
print(len(y))
print(sse/(len(y)-2))
```

```
191.04472735782826
20
10.613595964323792
```

$$\sigma^2 = \frac{SS_e}{DF} = \frac{191.044}{20-2} = 10.613$$

Q2

---

2. A hospital administrator wishes to study the relation between patient satisfaction and patient's age, severity of illness, and anxiety level. The data is in HW11.xslx

a. Fit a linear regression model with the three predictor variables.

In [3]:

```
data = pd.read_excel("HW #11.xlsx",sheet_name='Problem 2')
data = data.sort_values(by='Satisfaction')
y = data['Satisfaction']
X = data[['Age', 'Severity of Illness', 'Anxiety Level']]

# Adding a constant to the model (for the intercept)
X = sm.add_constant(X)

# Fitting the linear regression model
model = sm.OLS(y, X).fit()

# Getting the summary of the model
model_summary = model.summary()
model_summary

```

Out[3]:

OLS Regression Results

Dep. Variable:	Satisfaction	**R-squared:**	0.682
Model:	OLS	**Adj. R-squared:**	0.659
Method:	Least Squares	**F-statistic:**	30.05
Date:	Mon, 04 Dec 2023	**Prob (F-statistic):**	1.54e-10
Time:	12:15:53	**Log-Likelihood:**	-169.36
No. Observations:	46	**AIC:**	346.7
Df Residuals:	42	**BIC:**	354.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	158.4913	18.126	8.744	0.000	121.912	195.071
Age	-1.1416	0.215	-5.315	0.000	-1.575	-0.708
Severity of Illness	-0.4420	0.492	-0.898	0.374	-1.435	0.551
Anxiety Level	-13.4702	7.100	-1.897	0.065	-27.798	0.858

Omnibus:	5.219	**Durbin-Watson:**	1.214
Prob(Omnibus):	0.074	**Jarque-Bera (JB):**	2.074
Skew:	-0.098	**Prob(JB):**	0.354
Kurtosis:	1.978	**Cond. No.**	782.

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

$$\hat{y}_i = 13.47 * x_3 + -.442 * x_2 + -1.1416 * x_1 + 157.4913$$

---

b. Using t tests, what can be implied about $\beta 1$, $\beta 2$, and $\beta 3$ (just conclusion required)?

Age:

the p value for age is 0 which is less than .05, that indicates that we can reject null hypothesis that there is no relation between age and saticfication. in other words it indicates a significant relationship between patient age and satisfaction.

Severity of Illness:

Here, the p-value is greater than 0.05, indicating that we fail to reject the null hypothesis. This implies that the coefficient for Severity of Illness is not significantly different from zero

Anxiety Level:

The p-value is slightly above the 0.05 threshold, which typically suggests that the coefficient is not significantly different from zero at the 5% significance level

c. Interpret the estimated value of $\beta 1$, the regression coefficient for patient's age.

the coefficient for age is -1.1416, which means with one year age increase the satisfication level decrease 1.1416 level

---

d. Obtain a prediction for patient satisfaction with respect to a 35-year-old patient, who has a 45 severity of illness index and a 2.2 anxiety level index.

$$\hat{y}_i = 13.47 * x_3 + -.442 * x_2 + -1.1416 * x_1 + 158.4913$$
$$-13.47 * 2.2 + -.442 * 45 + -1.1416 * 35 + 158.49 = 69.01$$

---

e. Construct a normal probability plot of the residuals, what can you conclude?

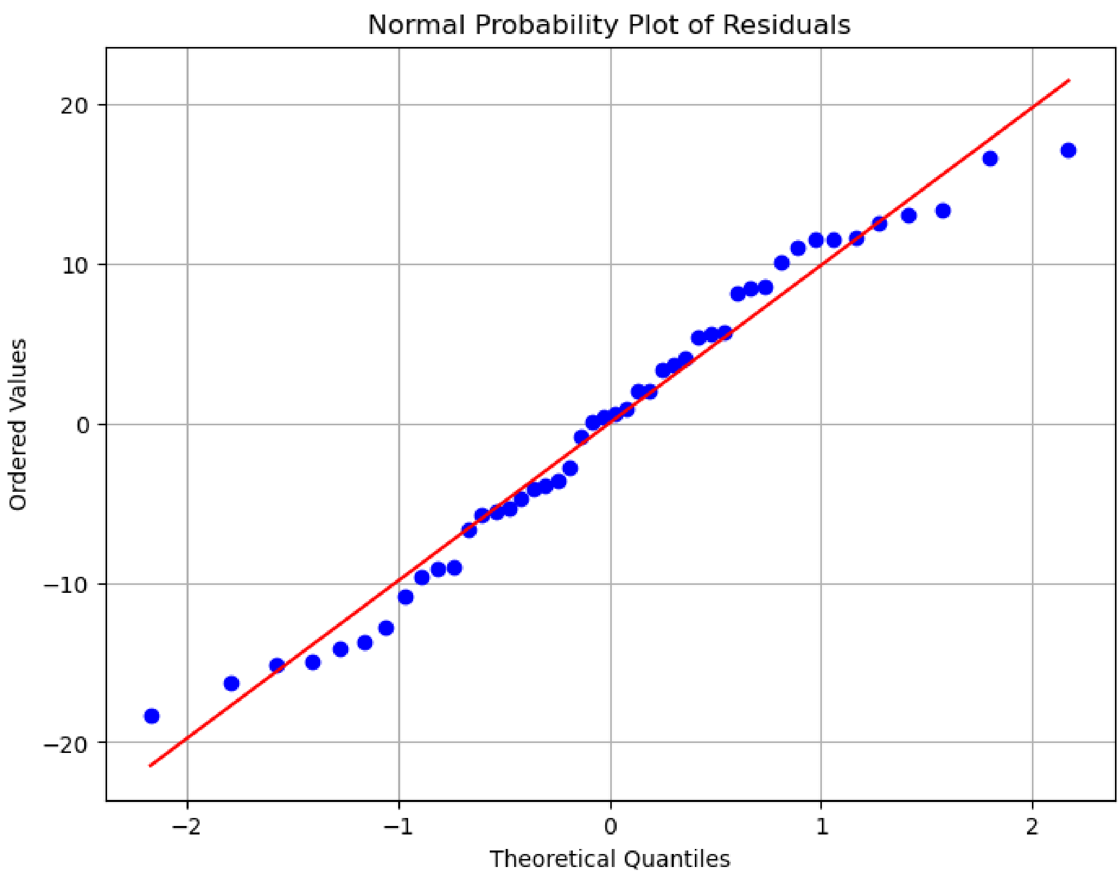
In [14]:

```
data['predicted_satisfaction'] = -13.47*data['Anxiety Level'] - 0.442
data['residuals'] = data['Satisfaction'] - data['predicted_satisfactio
```

In [15]:

```python
# Creating a normal probability plot
plt.figure(figsize=(8, 6))
stats.probplot(data.residuals, dist="norm", plot=plt)
plt.title("Normal Probability Plot of Residuals")
plt.xlabel("Theoretical Quantiles")
plt.ylabel("Ordered Values")
plt.grid(True)
plt.show()

```

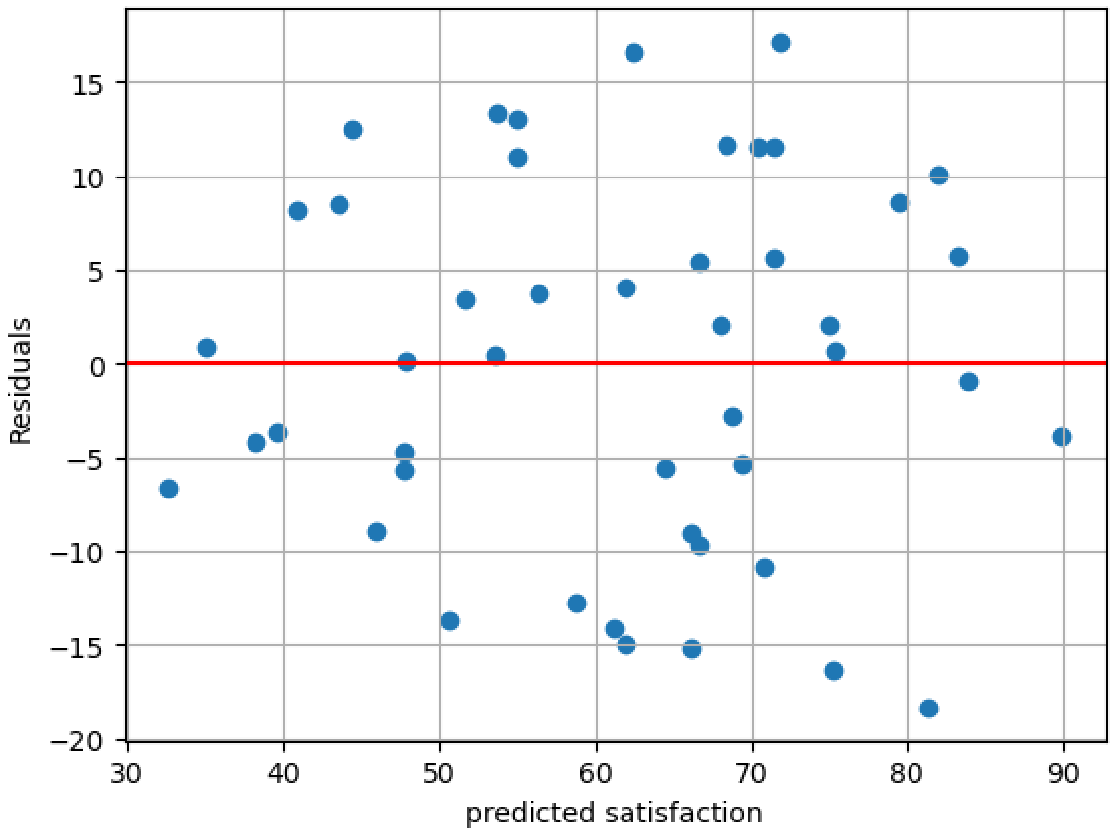


since the residuals almost perfectly align with the line we can understand that the residuals have a normal distribution

---

f. Construct a residual versus predicted plot, what can you conclude?

In [20]:
```
plt.scatter(data.predicted_satisfaction,data.residuals)
plt.ylabel("Residuals")
plt.xlabel("predicted satisfaction")
plt.axhline(y=0, color='r', linestyle='-')
plt.grid(True)
plt.show()
```



the residuals seems to be randomly distributed around the horizontal line so there is no distinct relation berweeen residuals and the predicted value

In [ ]:
```

```