In [ ]:

```
#50 years ocean fishing data
```

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import preprocessing,svm
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

In [11]:

```
df=pd.read_csv("bottle1.csv",encoding='ISO-8859-1')
df.head(2)
```

Out[11]:

| | Cst_Cnt | Btl_Cnt | Sta_ID | Depth_ID | Depthm | T_degC | Salnty | O2ml_L | STheta | O2Sat | . |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 054.0 056.0 | 19-4903CR-HY-060-0930-05400560-0000A-3 | 0 | 10.50 | 33.44 | NaN | 25.649 | NaN | . |
| 1 | 1 | 2 | 054.0 056.0 | 19-4903CR-HY-060-0930-05400560-0008A-3 | 8 | 10.46 | 33.44 | NaN | 25.656 | NaN | . |

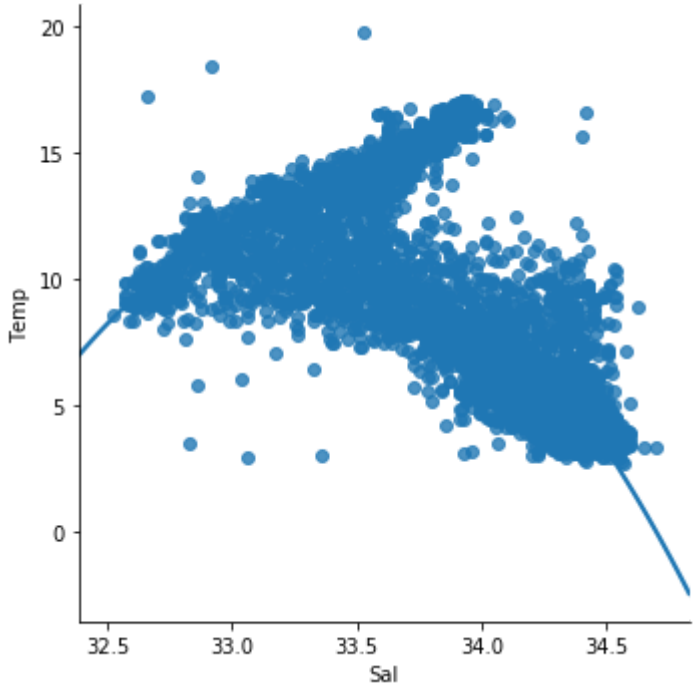
2 rows × 73 columns

In [19]:

```
df_binary = df[['Salnty', 'T_degC']]
df_binary
# Taking only the selected two attributes from the dataset
df_binary.columns=['Sal', 'Temp'] #renameing the column
df_binary.head() #taking only two columns
# Plotting the data scatter
sns.lmplot(x ="Sal", y ="Temp", data = df_binary, order = 2, ci = None)
#looks like negative correlation b/w temp and salinity
#Looks like dataset has to be cleaned for regression
```

Out[19]:

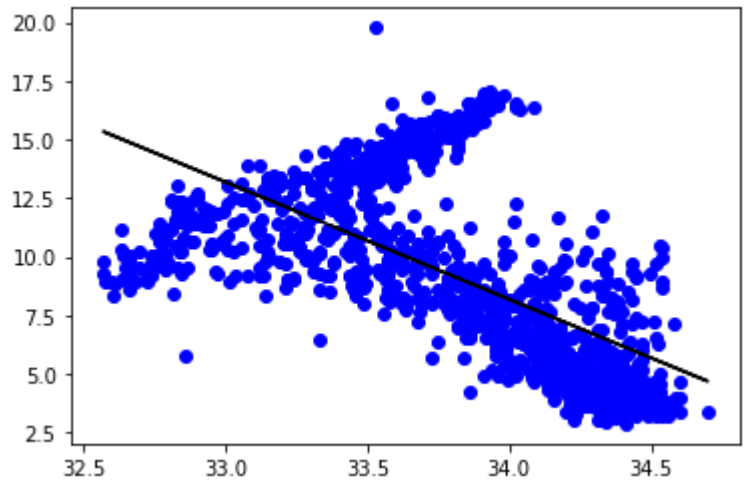
```
<seaborn.axisgrid.FacetGrid at 0x2a4291db808>
```



In [35]:

```
#Data cleaning
# Eliminating NaN or missing input numbers
df_binary.fillna(method ='ffill', inplace = True)  #forward fill valid value
#Training our model
X = np.array(df_binary['Sal']).reshape(-1, 1)
y = np.array(df_binary['Temp']).reshape(-1, 1)
df_binary.dropna(inplace = True)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25)
# Splitting the data into training and testing data
regr = LinearRegression()
regr.fit(X_train, y_train)
print(regr.score(X_test, y_test)) #0.40593972369001285
#Exploring our results
y_pred = regr.predict(X_test)
plt.scatter(X_test, y_test, color ='b')
plt.plot(X_test, y_pred, color ='k')

plt.show()
```

```
C:\Users\hp\Anaconda3\lib\site-packages\pandas\core\frame.py:4244: Setting
WithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  **kwargs
C:\Users\hp\Anaconda3\lib\site-packages\ipykernel_launcher.py:7: SettingWi
thCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  import sys

0.41216341574778725
```




In [ ]:

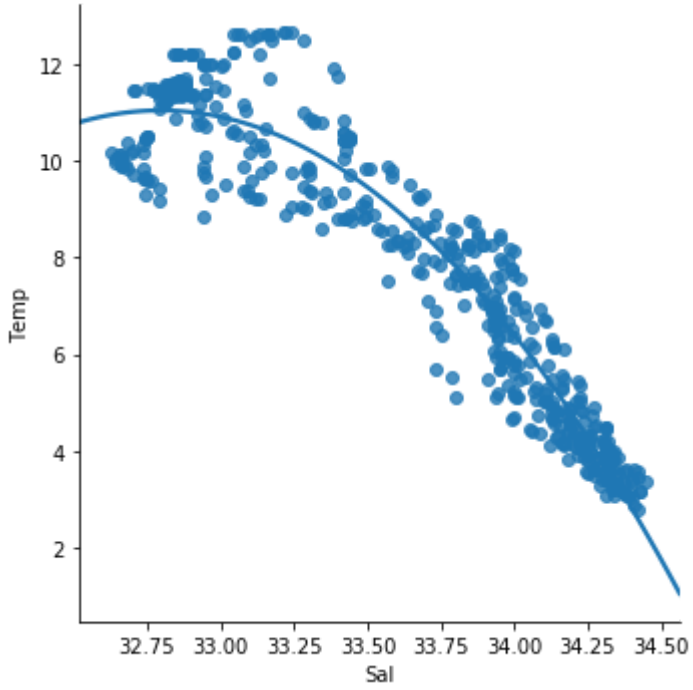
```
#Working with a smaller dataset
```

In [48]:

```
df_binary500=df_binary[:][:500]
# Selecting the 1st 500 rows of the data
sns.lmplot(x="Sal",y="Temp",data=df_binary500,order=2,ci=None)
```

Out[48]:

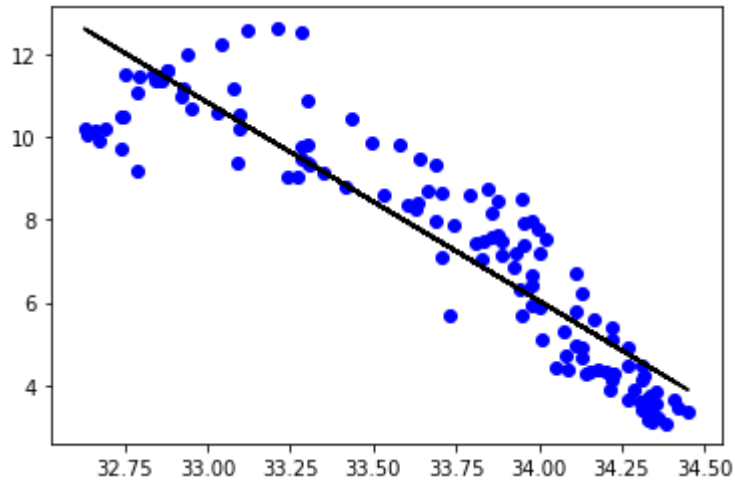
```
<seaborn.axisgrid.FacetGrid at 0x2a42aa51b88>
```



In [55]:

```
df_binary500.fillna(method='ffill',inplace=True)
X=np.array(df_binary500['Sal']).reshape(-1,1)
y=np.array(df_binary500['Temp']).reshape(-1,1)
df_binary500.dropna(inplace=True)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25)
regr = LinearRegression()
regr.fit(X_train, y_train)
print(regr.score(X_test, y_test)) #0.8772627816262462
#scatter
y_pred = regr.predict(X_test)
plt.scatter(X_test, y_test, color ='b')
plt.plot(X_test, y_pred, color ='k')

plt.show()
```

```
0.8303676579347041
```




In [1]:

```
pip install xelatex
```

```
Collecting xelatex
Note: you may need to restart the kernel to use updated packages.

  ERROR: Could not find a version that satisfies the requirement xelatex
(from versions: none)
ERROR: No matching distribution found for xelatex
```

In [ ]:

```

```