hw4_project

February 6, 2026

[1]:
```python
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import statsmodels.api as sm
from scipy.stats import norm
```

[2]:
```python
insurance = pd.read_csv("../datasets/insurance.csv")
insurance.head()
```

[2]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

1 1

[3]:
```python
# continuous predictor variables only
X = insurance[['age', 'bmi', 'children']]
y = insurance['charges']

X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.120
Model:                            OLS   Adj. R-squared:                  0.118
Method:                 Least Squares   F-statistic:                     60.69
Date:                Fri, 06 Feb 2026   Prob (F-statistic):           8.80e-37
Time:                        10:18:53   Log-Likelihood:                -14392.
No. Observations:                1338   AIC:                         2.879e+04
Df Residuals:                    1334   BIC:                         2.881e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
```

```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -6916.2433   1757.480     -3.935      0.000   -1.04e+04   -3468.518
age          239.9945     22.289     10.767      0.000     196.269     283.720
bmi          332.0834     51.310      6.472      0.000     231.425     432.741
children     542.8647    258.241      2.102      0.036      36.261    1049.468
==============================================================================
Omnibus:                      325.395   Durbin-Watson:                   2.012
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              603.372
Skew:                           1.520   Prob(JB):                    9.54e-132
Kurtosis:                       4.255   Cond. No.                         290.
==============================================================================

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

We see that the DW score is 2, which doesn't indicate any signs of autocorrelation or nonlinearity. The omnibus, skew, and kurtosis scores are quite high however, indicating non-Gaussian residuals. Overall, however, as noted in the previous project as well, this doesn't indicate an ill-fitted *non-linear* relationship, but rather just an omission of more important predictor variables such as the categorical ones that Lantz analyzes is more important towards the model.

2 2

[6]:
```
X = insurance.drop(columns=['charges'])
X = pd.get_dummies(X, drop_first=True).astype(float)
y = insurance['charges']

X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.749
Method:                 Least Squares   F-statistic:                     500.8
Date:                Fri, 06 Feb 2026   Prob (F-statistic):               0.00
Time:                        10:24:23   Log-Likelihood:                -13548.
No. Observations:                1338   AIC:                         2.711e+04
Df Residuals:                    1329   BIC:                         2.716e+04
Df Model:                           8
Covariance Type:            nonrobust
================================================================================
====
                       coef    std err          t      P>|t|      [0.025
0.975]
```

```
--------------------------------------------------------------------------------
----
const              -1.194e+04    987.819    -12.086      0.000    -1.39e+04
-1e+04
age                 256.8564     11.899     21.587      0.000     233.514
280.199
bmi                 339.1935     28.599     11.860      0.000     283.088
395.298
children            475.5005    137.804      3.451      0.001     205.163
745.838
sex_male           -131.3144    332.945     -0.394      0.693    -784.470
521.842
smoker_yes         2.385e+04    413.153     57.723      0.000      2.3e+04
2.47e+04
region_northwest   -352.9639    476.276     -0.741      0.459   -1287.298
581.370
region_southeast -1035.0220    478.692     -2.162      0.031   -1974.097
-95.947
region_southwest   -960.0510    477.933     -2.009      0.045   -1897.636
-22.466
==============================================================================
Omnibus:                      300.366   Durbin-Watson:                   2.088
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              718.887
Skew:                           1.211   Prob(JB):                   7.86e-157
Kurtosis:                       5.651   Cond. No.                         311.
==============================================================================

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

This is a better model including all variables. We can see that only `age`, `bmi`, `chidlren`, and `smoker_yes` are significant predictors. Likewise to Lantz, we can convert `bmi` to a categorical binary predictor instead, indicating `1` for individuals that are obese (bmi > 30), and `0` otherwise.

[8]:
```python
# model with only significant predictors
X = insurance.drop(columns=['charges'])
X = pd.get_dummies(X, drop_first=True).astype(float)
X = X[['age', 'bmi', 'children', 'smoker_yes']]

# convert bmi to categorical variable (obese or not)
X['bmi_obese'] = (X['bmi'] >= 30).astype(float)
X = X.drop(columns=['bmi'])

X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
print(model.summary())
```

```
                            OLS Regression Results
```

```
==============================================================================
Dep. Variable:                charges   R-squared:                       0.753
Model:                            OLS   Adj. R-squared:                  0.753
Method:                 Least Squares   F-statistic:                     1018.
Date:                Fri, 06 Feb 2026   Prob (F-statistic):               0.00
Time:                        10:28:54   Log-Likelihood:                -13541.
No. Observations:                1338   AIC:                         2.709e+04
Df Residuals:                    1333   BIC:                         2.712e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -4549.7577    531.337     -8.563      0.000   -5592.105   -3507.410
age          260.3762     11.783     22.097      0.000     237.261     283.492
children     475.9851    136.795      3.480      0.001     207.628     744.342
smoker_yes  2.383e+04    408.245     58.366      0.000     2.3e+04    2.46e+04
bmi_obese   4184.2284    331.129     12.636      0.000    3534.637    4833.820
==============================================================================
Omnibus:                      327.599   Durbin-Watson:                   2.092
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              831.658
Skew:                           1.293   Prob(JB):                    2.56e-181
Kurtosis:                       5.869   Cond. No.                         140.
==============================================================================

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

This model doesn't really improve anything. We can try again by adding an `age^2` term and interaction between obesity and smoking like Lantz does.

[13]:
```python
X = insurance.drop(columns=['charges'])
X = pd.get_dummies(X, drop_first=True).astype(float)
X = X[['age', 'bmi', 'children', 'smoker_yes']]
X['bmi_obese'] = (X['bmi'] >= 30).astype(float)
X = X.drop(columns=['bmi'])
X['age_squared'] = X['age'] ** 2
X['obese_smoker'] = X['bmi_obese'] * X['smoker_yes']
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.864
Model:                            OLS   Adj. R-squared:                  0.863
Method:                 Least Squares   F-statistic:                     1404.
Date:                Fri, 06 Feb 2026   Prob (F-statistic):               0.00
```

```
Time:                        10:33:28   Log-Likelihood:                -13145.
No. Observations:                1338   AIC:                         2.630e+04
Df Residuals:                    1331   BIC:                         2.634e+04
Df Model:                           6
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const         2237.3818   1083.910      2.064      0.039     111.023    4363.741
age            -24.5114     60.287     -0.407      0.684    -142.779      93.757
children       669.3870    106.688      6.274      0.000     460.092     878.682
smoker_yes    1.339e+04    442.587     30.253      0.000    1.25e+04    1.43e+04
bmi_obese       42.3927    276.907      0.153      0.878    -500.828     585.614
age_squared      3.6643      0.752      4.870      0.000       2.188       5.140
obese_smoker  1.976e+04    608.631     32.465      0.000    1.86e+04     2.1e+04
==============================================================================
Omnibus:                      880.074   Durbin-Watson:                   2.062
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7491.608
Skew:                           3.125   Prob(JB):                         0.00
Kurtosis:                      12.763   Cond. No.                     1.84e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.84e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

3 3

This is a much better model than our one from project 3. The R^2 is significantly higher.

Comparing the coefficients:
Intercept: 2237.3818, SE 1083.910, p 0.039
Still significant, albeit with a completely different value. No big change.

Age: 024.5114, SE 60.287, p 0.684
Age has now become a statistically insignificant predictor. This is expected as we now added in an age^2 term.

We now have additional coefficients such as `smoker_yes`, `bmi_obese`, and `obese_smoker`. These coefficients are all relatively well estimated and statistically significant with the exception of `bmi_obese`.

[ ]:
```
X = insurance.drop(columns=['charges'])
y = insurance['charges']
X = pd.get_dummies(X, drop_first=True).astype(float)
X['age_squared'] = X['age'] ** 2
X['bmi_obese'] = (X['bmi'] >= 30).astype(float)
```

```python
X['obese_smoker'] = X['bmi_obese'] * X['smoker_yes']
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.866
Model:                            OLS   Adj. R-squared:                  0.865
Method:                 Least Squares   F-statistic:                     781.7
Date:                Fri, 06 Feb 2026   Prob (F-statistic):               0.00
Time:                        12:25:25   Log-Likelihood:                -13131.
No. Observations:                1338   AIC:                         2.629e+04
Df Residuals:                    1326   BIC:                         2.635e+04
Df Model:                          11
Covariance Type:            nonrobust
================================================================================
====
                       coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
----
const              134.2509   1362.751      0.099      0.922   -2539.132
2807.634
age                -32.6851     59.824     -0.546      0.585    -150.045
84.675
bmi                120.0196     34.266      3.503      0.000      52.798
187.241
children           678.5612    105.883      6.409      0.000     470.844
886.278
sex_male          -496.8245    244.366     -2.033      0.042    -976.210
-17.438
smoker_yes          1.34e+04    439.949     30.469      0.000    1.25e+04
1.43e+04
region_northwest  -279.2038    349.275     -0.799      0.424    -964.395
405.987
region_southeast  -828.5467    351.635     -2.356      0.019   -1518.369
-138.725
region_southwest -1222.6437    350.528     -3.488      0.001   -1910.294
-534.993
age_squared          3.7316      0.746      5.000      0.000       2.268
5.196
bmi_obese        -1000.1403    422.840     -2.365      0.018   -1829.649
-170.632
obese_smoker      1.981e+04    604.657     32.764      0.000    1.86e+04
2.1e+04
==============================================================================
Omnibus:                      890.976   Durbin-Watson:                   2.064
```

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7722.759
Skew:                           3.170   Prob(JB):                         0.00
Kurtosis:                      12.916   Cond. No.                     2.34e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 2.34e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

We can also generate a model similar to Lantz's final model and find similar results. There are varying degrees of statistical significance between the added coefficients, however overall model performance is the same.

Adding in the `age^2` term and interaction effects between obesity indication and smoking have significantly improved our model compared to just using the continuous predictor variables.

LLM Usage: All work was done by myself in VSCode with GitHub Copilot integration. The integration "provides code suggestions, explanations, and automated implementations based on natural language prompts and existing code context," and also offers autonomous coding and an in-IDE chat interface that is able to interact with the current codebase. Only the Copilot provided automatic inline suggestions for both LaTex and Python in `.tex` and `.ipynb` Jupyter notebook files respectively were taken into account / used.