hw7_project

February 26, 2026

[1]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import statsmodels.api as sm
```

[2]:
```
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]:
```
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)
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:                Thu, 26 Feb 2026   Prob (F-statistic):               0.00
Time:                        15:55:26   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 again start with the Lantz model from HW4.

[6]:

```
threshold = y.median()
y_binary = (y > threshold).astype(int)
```

```python
logit_model = sm.GLM(y_binary, X, family=sm.families.Binomial()).fit()
print(logit_model.summary())

print(np.exp(logit_model.params))
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                charges   No. Observations:                 1338
Model:                            GLM   Df Residuals:                     1326
Model Family:                Binomial   Df Model:                           11
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -287.73
Date:                Thu, 26 Feb 2026   Deviance:                       575.46
Time:                        16:03:48   Pearson chi2:                 1.46e+03
No. Iterations:                    25   Pseudo R-squ. (CS):             0.6157
Covariance Type:            nonrobust
==============================================================================
====
                       coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
----
const                7.8018      1.395      5.592      0.000       5.067
10.536
age                 -0.7961      0.078    -10.157      0.000      -0.950
-0.642
bmi                  0.0484      0.030      1.612      0.107      -0.010
0.107
children             0.5113      0.090      5.706      0.000       0.336
0.687
sex_male            -0.5040      0.219     -2.297      0.022      -0.934
-0.074
smoker_yes          29.0072   2.49e+04      0.001      0.999   -4.89e+04
4.89e+04
region_northwest    -0.4304      0.296     -1.456      0.145      -1.010
0.149
region_southeast    -1.0420      0.312     -3.345      0.001      -1.652
-0.431
region_southwest    -0.9675      0.307     -3.154      0.002      -1.569
-0.366
age_squared          0.0126      0.001     11.424      0.000       0.010
0.015
bmi_obese           -0.2167      0.357     -0.608      0.543      -0.915
0.482
obese_smoker        -0.1575   3.36e+04  -4.69e-06      1.000   -6.58e+04
6.58e+04
==============================================================================
====
const                2.445119e+03
age                  4.511004e-01
bmi                  1.049562e+00
children             1.667511e+00
sex_male             6.041287e-01
smoker_yes           3.959766e+12
region_northwest     6.502211e-01
region_southeast     3.527627e-01
region_southwest     3.800250e-01
age_squared          1.012727e+00
bmi_obese            8.051868e-01
obese_smoker         8.542758e-01
dtype: float64
```

All coefficients are interpreted in the following manner: log-odds, odds ratio, standard error

Intercept: 7.8018, 2,445, 1.395
The intercept of the model, the baseline when all predictors are zero which isn't realistic or meaningful in this model. Anchors the regression line.

Age: -0.7961, 4.511, 0.078
Age^2: 0.0126, 1.103, 0.001
The negative linear term and positive quadratic term imply a U-shaped relationship in log-odds. This means at younger ages, the probability of high cost decreases quickly, but as age increases, the quadratic term dominates, increasing the probability of a higher cost. Both terms are significant and precise (low SE), thus age is an important non linear predictor.

BMI: 0.0484, 1.05, 0.03
A one unit increase in BMI while holding all other predictors constant increases the odds of high cost medical charge by around 5%. The coefficient is well estimated due to low SE however is not statistically significant with a p-value of 0.107.

Children: 0.5113, 1.668, 0.09
Each additional child increases the odds of being high cost by around 67%. This is a strong, precise, and statistically significant effect.

Sex (male): -0.504, 0.6, 0.219
Males have around a 40% lower odds of being high cost compared to females while holding all other variables constant. This is a strong, well estimated, and statistically significant effect.

Region:
NW: -0.4304, 0.65, 0.296
SE: -1.0420, 0.35, 0.312
SW: -0.9675, 0.38, 0.307
People outside the northeast are generally less likely to be in the high cost group. These are reasonably precise estimates, however northwest seems to be not statistically significant due to the high p value of 0.145.

Smoker (yes): 29.0072, 3.959766e+12, 2.49e+04
The enormous standard errors which tells us that the model is experiencing quasi-complete separation. This means that almost every smoker is above the median cahrge, thus smoking nearly perfectly predicts high cost. In a logistic regression, this causes the coefficient to spike, as well as

its stanard error, causing the model to become numerically unstable. We can also see this issue with people who are obese and smokers. Logically this also makes sense as having both of these issues generally puts you toward higher medical insurance cost charges.

BMI_obese: -0.2167, 0.8, 0.357
An indicator that people who are obese are more likely to experience higher insurane cost charges. The value is well estimated but not statistically significant due to its p-value of 0.543.

2 2

[7]:

```python
from sklearn.linear_model import LassoCV, Lasso
from sklearn.preprocessing import StandardScaler

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)
X['obese_smoker'] = X['bmi_obese'] * X['smoker_yes']

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

lasso_cv = LassoCV(cv=5, random_state=0).fit(X_scaled, y_binary)
selected = X.columns[lasso_cv.coef_ != 0]

X_reduced = sm.add_constant(X[selected])
reduced_model = sm.GLM(y_binary, X_reduced, family=sm.families.Binomial()).fit()

B = 1000
n = len(y_binary)

boot_coefs = []

np.random.seed(0)
for _ in range(B):
    sample_idx = np.random.choice(n, n, replace=True)
    X_boot = X.iloc[sample_idx]
    y_boot = y_binary.iloc[sample_idx]

    scaler_b = StandardScaler()
    Xb_scaled = scaler_b.fit_transform(X_boot)

    lasso_b = Lasso(alpha=lasso_cv.alpha_)
    lasso_b.fit(Xb_scaled, y_boot)
    selected_b = X.columns[lasso_b.coef_ != 0]
    if len(selected_b) == 0:
        continue

    Xb_reduced = sm.add_constant(X_boot[selected_b])
    model_b = sm.GLM(y_boot, Xb_reduced, family=sm.families.Binomial()).fit()

    coef_series = pd.Series(0.0, index=X_reduced.columns)
    for name in model_b.params.index:
        if name in coef_series.index:
            coef_series[name] = model_b.params[name]

    boot_coefs.append(coef_series.values)

boot_coefs = np.array(boot_coefs)
boot_ses = pd.Series(boot_coefs.std(axis=0), index=X_reduced.columns)

print(reduced_model.summary())
print(boot_ses)
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                charges   No. Observations:                 1338
Model:                            GLM   Df Residuals:                     1326
Model Family:                Binomial   Df Model:                           11
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -287.73
Date:                Thu, 26 Feb 2026   Deviance:                       575.46
Time:                        16:55:41   Pearson chi2:                 1.46e+03
No. Iterations:                    25   Pseudo R-squ. (CS):             0.6157
Covariance Type:            nonrobust
================================================================================
====
                       coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
----
const                7.8018      1.395      5.592      0.000       5.067
10.536
age                 -0.7961      0.078    -10.157      0.000      -0.950
-0.642
bmi                  0.0484      0.030      1.612      0.107      -0.010
0.107
children             0.5113      0.090      5.706      0.000       0.336
0.687
sex_male            -0.5040      0.219     -2.297      0.022      -0.934
-0.074
smoker_yes          29.0072   2.49e+04      0.001      0.999   -4.89e+04
4.89e+04
```

```
region_northwest    -0.4304      0.296     -1.456      0.145      -1.010
0.149
region_southeast    -1.0420      0.312     -3.345      0.001      -1.652
-0.431
region_southwest    -0.9675      0.307     -3.154      0.002      -1.569
-0.366
age_squared          0.0126      0.001     11.424      0.000       0.010
0.015
bmi_obese           -0.2167      0.357     -0.608      0.543      -0.915
0.482
obese_smoker        -0.1575   3.36e+04  -4.69e-06      1.000   -6.58e+04
6.58e+04
================================================================================
====
const               1.432633
age                 0.082923
bmi                 0.029217
children            0.094576
sex_male            0.225012
smoker_yes          0.324555
region_northwest    0.285357
region_southeast    0.322889
region_southwest    0.309513
age_squared         0.001190
bmi_obese           0.350160
obese_smoker        0.292672
dtype: float64
```

Here we see that the Lasso CV kept all the predictors. This tells us that there is no strong sparsity structure in this problem. However, what is important is the difference in the bootstrap SE's versus the reduced model coefficients. We see that for the majority of the coefficients the SE's are roughly the same. This is what we want for simple predictors. However as we previously identified, smoking yes and obese smoker have drastically different SE's in the bootstrap versus the reduced model. The classical SE is huge, while the bootstrap SE is tiny. This is because in bootstrap samples, smoking is consistently selected, leading to a large and positive coefficient, however there is low variability across samples. This once again tells us that smoking is an extremely dominant predictor. The effect of obese and smoking is a little less. In general, our bootstrap SE's are more reliable and overall tell us that strong predictors are primarily smoking, followed by age (non linear), number of children, region, and sex. BMI and obesity factors are relatively weak.

3 3

Our model from project 6.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.866
Model:                            OLS   Adj. R-squared:                  0.865
```

```
Method:                 Least Squares   F-statistic:                     860.3
Date:                Thu, 19 Feb 2026   Prob (F-statistic):               0.00
Time:                        12:26:52   Log-Likelihood:                -13131.
No. Observations:                1338   AIC:                         2.628e+04
Df Residuals:                    1327   BIC:                         2.634e+04
Df Model:                          10
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const             -414.4423    920.902     -0.450      0.653   -2221.024    1392.140
bmi                119.2718     34.230      3.484      0.001      52.122     186.422
children           661.1365    100.939      6.550      0.000     463.119     859.154
sex_male          -495.7472    244.293     -2.029      0.043    -974.991     -16.504
smoker_yes         1.34e+04    439.832     30.476      0.000    1.25e+04    1.43e+04
region_northwest  -277.7562    349.172     -0.795      0.426    -962.746     407.233
region_southeast  -827.1352    351.533     -2.353      0.019   -1516.756    -137.515
region_southwest -1222.6434    350.436     -3.489      0.001   -1910.112    -535.175
age_squared          3.3282      0.109     30.581      0.000       3.115       3.542
bmi_obese         -985.5475    421.884     -2.336      0.020   -1813.180    -157.915
obese_smoker      1.981e+04    604.495     32.771      0.000    1.86e+04     2.1e+04
==============================================================================
Omnibus:                      890.985   Durbin-Watson:                   2.064
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7730.895
Skew:                           3.170   Prob(JB):                         0.00
Kurtosis:                      12.924   Cond. No.                     1.65e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.65e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
const               1371.678675
bmi                   35.145301
children             102.991410
sex_male             246.225509
smoker_yes           368.468643
region_northwest     372.975856
region_southeast     380.098283
region_southwest     351.463406
age_squared            0.711913
bmi_obese            438.910109
obese_smoker         527.960025
dtype: float64
```

The key differenecs between the two models lies in what they are predicting. The continuous model predicts the given charge for a patient with a set of certain characteristics, while the binary model predicts the log-odds of being in a high cost group (experience high medical charges). A coefficient in the continuous model represents the predicted change in insurance charge, while a coefficient in the binary model represents a change in the log-odds of being high cost. We also see that the continuous model actually dropped the linear age predictor.

We can see for smokers, the OLS model tells us that they are predicted to pay $13,400 more, while an obese smoker pays $19,810 more. Likewise, the binary model tells us that smokers / obese smokers are significantly more likely to pay more. Smoking deterministically pushes someone into high cost. We can see similar effects for age, number of children, and region, but the two models tell similar stories.

The OLS model R^2 is 0.866, which is pretty high. This means that the OLS model is able to explain 86.6% of variance in charges, however multicollinearity is present. Even then, it is handled well and OLS can estimate the magnitude of impacts for each coefficient pretty well. The logistic model on the other hand, has a pseudo R^2 of 0.61. This means that it can't capture the magnitude of variation inside each class well, which makes sense as it only models the probability of crossing the median. It boils down to the fact that the OLS continuous model answers what drives the actual dollar amount of insurance costs, while the binary model answers what predicts being in the expensive half of patients. Binarizing a continuous outcome arguably throws away useful information used for regression, reduces statistical efficiency, and can inflate instatiblity as we see with the high SE's. We overall prefer to use the OLS continuous model to predict insurance pricing, but we can also use the logistic sometimes if we care about risk stratification and identifying potential high-cost patients.

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.

Problem 1: LLM was used to swap model from Logit to GLM binomial as we encountered singular matrix issues. This ended up to be the case as smoking nearly perfectly predicted the high cost class.
Problem 2: LLM was not used in this problem (code taken from project 6).
Problem 3: LLM was used in this problem to explain the key differences in efficiency / statistical analysis when it comes down to what each model is doing / predicting.