hw6_project

February 19, 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, 19 Feb 2026   Prob (F-statistic):               0.00
Time:                        12:06:46   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 will use the final Lantz model as from HW4. The standard errors of the model coefficients as automatically provided by the linear model fit under the standard assumptions are:

[5]:
```
model.bse
```

[5]:
```
const                1362.751128
age                    59.824155
bmi                    34.266049
children              105.883122
sex_male              244.365939
smoker_yes            439.949051
region_northwest      349.274576
region_southeast      351.635152
region_southwest      350.528454
age_squared             0.746299
bmi_obese             422.840162
obese_smoker          604.656665
dtype: float64
```

2 2

Residual bootstrap assumes that the model is correctly specified, errors are i.i.d and homoskedastic, and the X is fixed. However, from our previous analysis of multiple different models, we know that the insurance dataset almost certainly violates homoskedasticity. This is why our model has nonlinear terms and interaction effects. Thus, we prefer to use case bootstrap, as this doesn't make any homoskedasticity assumptions, does not require fixed regressors, and is robust to model misspecification. Given that this is observational insurance data, treating X as random is more realistic, which is also why we prefer to resample entire cases.

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

B = 1000
n = len(insurance)

boot_coefs = []

np.random.seed(0)

for _ in range(B):
    sample_idx = np.random.choice(n, n, replace=True)
    X_boot = X_full.iloc[sample_idx]
    y_boot = y.iloc[sample_idx]

    boot_model = sm.OLS(y_boot, X_boot).fit()
    boot_coefs.append(boot_model.params.values)
```

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

comparison = pd.DataFrame({
    "Classical SE": model.bse,
    "Bootstrap SE": boot_ses
})

comparison
```

[ ]:

	Classical SE	Bootstrap SE
const	1362.751128	1422.671331
age	59.824155	63.580815
bmi	34.266049	34.889666
children	105.883122	109.062810
sex_male	244.365939	244.995048
smoker_yes	439.949051	365.790444
region_northwest	349.274576	358.299651
region_southeast	351.635152	373.412189
region_southwest	350.528454	365.346416
age_squared	0.746299	0.782413
bmi_obese	422.840162	438.391203
obese_smoker	604.656665	551.425311

We see that the bootstrap estimates are quite close to the expected coefficients. Of course, if we increased the number of draws, the bootstrap coefficients would get closer and close to those of the linear model fit. We can see some notable differences in `smoker_yes` and `obese_smoker`, which reinforces the fact that classical standard error may rely on homoskedasticity assumptions that aren't perfectly satisfied. This again shows why we should resample on cases and not residuals.

3 3

[9]:
```
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)
selected = X.columns[lasso_cv.coef_ != 0]
```

```
X_reduced = sm.add_constant(X[selected])
reduced_model = sm.OLS(y, X_reduced).fit()

B = 1000
n = len(y)

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.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.OLS(y_boot, Xb_reduced).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)
```

```
                            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 lasso cv dropped the linear age predictor, which was expected as it was not statistically significant in the previous model. We see tht the SEs are larger than the reduced model SEs. This is because variable selection was redone in each bootstrap sample, accounting for selection variability. Again, we see large differences in the SEs for `smoker_yes`, `bmi_obese`, and `obese_smoker`, as expected. This confirms that the interaction effects are important and the errors are not homoskedastic. Overall we are coming to the expected conclusions by Lantz that smokers and obese individuals should disproportionately experience increased medical charges.

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.

1: LLM was not used in this problem.
2: LLM was not used in this problem.
3: LLM was consulted for advice on variable selection process.