ISOM 2600 Business Analytics

TOPIC 6: MULTIPLE LINEAR REGRESSION (PART 2)

XUHU WAN

HKUST

JANUARY 15, 2022

Case study: Retail profits Continued

Step 3: Evaluate the initial model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Profit   R-squared:                       0.756
Model:                            OLS   Adj. R-squared:                  0.742
Method:                 Least Squares   F-statistic:                     53.12
Date:                Tue, 28 Jan 2020   Prob (F-statistic):           2.41e-29
Time:                        10:33:06   Log-Likelihood:                -1199.0
No. Observations:                 110   AIC:                             2412.
Df Residuals:                     103   BIC:                             2431.
Df Model:                           6
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                     1.316e+04   1.91e+04      0.689      0.493   -2.47e+04    5.11e+04
Income                       0.5986      0.589      1.017      0.312      -0.569       1.766
Disposable Income            2.5350      0.732      3.464      0.001       1.084       3.986
Birth Rate (per 1,000)    1703.8657    563.673      3.023      0.003     585.954    2821.777
Soc Security (per 1,000)   -47.5162    110.213     -0.431      0.667    -266.097     171.065
CV Death (per 100,000)     -22.6821     31.464     -0.721      0.473     -85.084      39.720
% 65 or Older             7713.8505   1316.210      5.861      0.000    5103.458    1.03e+04
==============================================================================
Omnibus:                        0.673   Durbin-Watson:                   1.546
Prob(Omnibus):                  0.714   Jarque-Bera (JB):                0.789
Skew:                          -0.097   Prob(JB):                        0.674
Kurtosis:                       2.634   Cond. No.                     4.85e+05
==============================================================================
```

F-test and ANOVA table

Q: Is the regression model useful?

The p-value on ANOVA table is used for testing

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_1 : \text{at least one of } \beta_1, \beta_2, \ldots, \beta_k \neq 0$$

What does this null hypothesis imply about the relationship between $\mathbf{y}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_k$?

This hypothesis is tested with p-value:

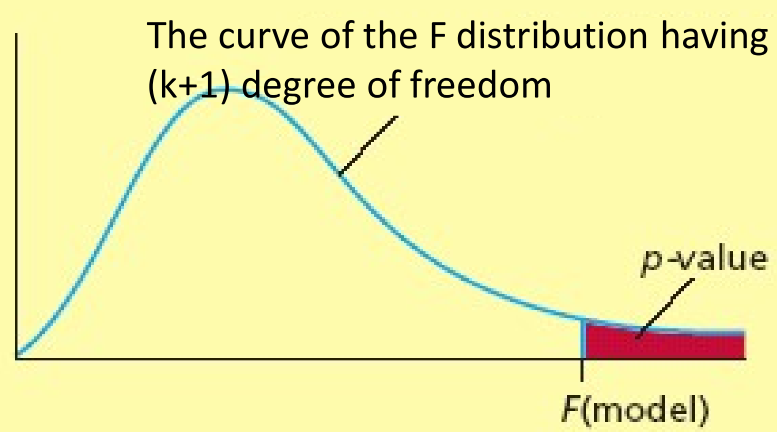
- e.g. If the p-value < 0.05, then $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ can be rejected at 5% significance.

The test is based on an F-ratio. The formula is

$$F(model) = \frac{SSR/k}{SSE/[n-(k+1)]} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

Reject H_0 if p-value < α

- p-value is based on F distribution with k numerator and n-(k+1) denominator degrees of freedom



- (Optional) F distribution with

degree of freedom d_1 and d_2 is determined by the following pdf.

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} {d_2}^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}}{x\, B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} = \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2} x\right)^{\frac{d_1+d_2}{2}}, \qquad x \in [0, +\infty)$$

where the Beta function $B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt$

Case study: Retail profits Continued

Step 3: Evaluate the initial model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Profit   R-squared:                       0.756
Model:                            OLS   Adj. R-squared:                  0.742
Method:                 Least Squares   F-statistic:                     53.12
Date:                Tue, 28 Jan 2020   Prob (F-statistic):           2.41e-29
Time:                        10:33:06   Log-Likelihood:                -1199.0
No. Observations:                 110   AIC:                             2412.
Df Residuals:                     103   BIC:                             2431.
Df Model:                           6
Covariance Type:            nonrobust
============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                   1.316e+04   1.91e+04      0.689      0.493   -2.47e+04    5.11e+04
Income                     0.5986      0.589      1.017      0.312      -0.569       1.766
Disposable Income          2.5350      0.732      3.464      0.001       1.084       3.986
Birth Rate (per 1,000)  1703.8657    563.673      3.023      0.003     585.954    2821.777
Soc Security (per 1,000) -47.5162    110.213     -0.431      0.667    -266.097     171.065
CV Death (per 100,000)   -22.6821     31.464     -0.721      0.473     -85.084      39.720
% 65 or Older           7713.8505   1316.210      5.861      0.000    5103.458    1.03e+04
==============================================================================
Omnibus:                        0.673   Durbin-Watson:                   1.546
Prob(Omnibus):                  0.714   Jarque-Bera (JB):                0.789
Skew:                          -0.097   Prob(JB):                        0.674
Kurtosis:                       2.634   Cond. No.                     4.85e+05
==============================================================================
```

R-square

Q: How useful is the regression model?

As in simple linear regression, R^2 in multiple regression is "the proportion of variation in y explained by the regression"

The formula is

$$R^2 = \frac{Explained\ variation}{Total\ variation} = \frac{\mathrm{SS}R}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

or

$$(1 - R^2)s_y^2 \approx RMSE^2$$

R^2 cannot decrease when another independent variable x is added to the regression.

Case study: Retail profits Continued

Step 3: Evaluate the initial model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Profit   R-squared:                       0.756
Model:                            OLS   Adj. R-squared:                  0.742
Method:                 Least Squares   F-statistic:                     53.12
Date:                Tue, 28 Jan 2020   Prob (F-statistic):           2.41e-29
Time:                        10:33:06   Log-Likelihood:                -1199.0
No. Observations:                 110   AIC:                             2412.
Df Residuals:                     103   BIC:                             2431.
Df Model:                           6
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                     1.316e+04   1.91e+04      0.689      0.493   -2.47e+04    5.11e+04
Income                       0.5986      0.589      1.017      0.312      -0.569       1.766
Disposable Income            2.5350      0.732      3.464      0.001       1.084       3.986
Birth Rate (per 1,000)    1703.8657    563.673      3.023      0.003     585.954    2821.777
Soc Security (per 1,000)   -47.5162    110.213     -0.431      0.667    -266.097     171.065
CV Death (per 100,000)     -22.6821     31.464     -0.721      0.473     -85.084      39.720
% 65 or Older             7713.8505   1316.210      5.861      0.000    5103.458    1.03e+04
==============================================================================
Omnibus:                        0.673   Durbin-Watson:                   1.546
Prob(Omnibus):                  0.714   Jarque-Bera (JB):                0.789
Skew:                          -0.097   Prob(JB):                        0.674
Kurtosis:                       2.634   Cond. No.                     4.85e+05
==============================================================================
```

Inference about $\beta_1, \beta_2, \ldots, \beta_k$

Q: Which predictors are useful?

Once we have checked the assumptions of the MRM, we can proceed to inference. Tests and confidence intervals used in simple regression generalize naturally to multiple regression.

Fact:

Under MRM, the sampling distributions of b_0, b_1, …, b_k are normal with means β_0, β_1 ….. β_k

Hypothesis testing for β_j

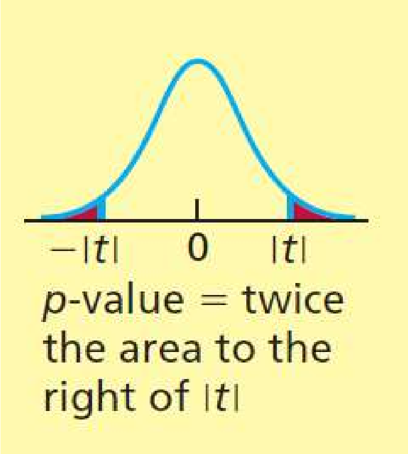
For testing

$$H_0 : \beta_j = c \text{ vs. } H_1: \beta_j \neq c$$

If

- p-value < 0.05 or
- 95% CI for β_j does not contain c,

then reject H_0 at 0.05 level of significance.



$$t = \frac{b_j - c}{s_{b_j}}$$

Null hypothesis of the form $H_0 : \beta_j = 0$ are usually of most interest. Why?

Confidence interval

Confidence Intervals for β_j

Approximately 95% CI's for β_0, β_1 ..... β_k are given by the same procedure as in simple regression, namely

$$b_j \pm t_{\alpha/2,n-k-1}\, s_{b_j},$$

- Note: for hand calculation of 95% CI, you can use 2 to approximate $t_{\alpha/2,n-k-1}$

Case study: Retail profits Continued

Step 3: Evaluate the initial model

```
                         OLS Regression Results
==============================================================================
Dep. Variable:                 Profit   R-squared:                       0.756
Model:                            OLS   Adj. R-squared:                  0.742
Method:                 Least Squares   F-statistic:                     53.12
Date:                Tue, 28 Jan 2020   Prob (F-statistic):           2.41e-29
Time:                        10:33:06   Log-Likelihood:                -1199.0
No. Observations:                 110   AIC:                             2412.
Df Residuals:                     103   BIC:                             2431.
Df Model:                           6
Covariance Type:            nonrobust
============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                    1.316e+04   1.91e+04      0.689      0.493   -2.47e+04    5.11e+04
Income                      0.5986      0.589      1.017      0.312      -0.569       1.766
Disposable Income           2.5350      0.732      3.464      0.001       1.084       3.986
Birth Rate (per 1,000)   1703.8657    563.673      3.023      0.003     585.954    2821.777
Soc Security (per 1,000)  -47.5162    110.213     -0.431      0.667    -266.097     171.065
CV Death (per 100,000)    -22.6821     31.464     -0.721      0.473     -85.084      39.720
% 65 or Older            7713.8505   1316.210      5.861      0.000    5103.458    1.03e+04
==============================================================================
Omnibus:                        0.673   Durbin-Watson:                   1.546
Prob(Omnibus):                  0.714   Jarque-Bera (JB):                0.789
Skew:                          -0.097   Prob(JB):                        0.674
Kurtosis:                       2.634   Cond. No.                     4.85e+05
==============================================================================
```

Interpreter the Estimated Coefficients

Slope (b_i)

- Estimated $\mathbf{y}$ changes by b_i unit for each 1 unit increase in $\mathbf{x}_i$, ***holding the other predictors constant /when the other predictors do not change / when compare to Y with the same*** $\mathbf{x}_j$ ***, j ≠i./ after allowing for the linear effects of the other predictors / after accounting for other predictors / For given values of other predictors.***

$\mathbf{y}$ -Intercept (b_0)

- Average value of $\mathbf{y}$ when all $\mathbf{x}_j$ = 0

Case study: Retail profits Continued

Step 3: Evaluate the initial model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Profit   R-squared:                       0.756
Model:                            OLS   Adj. R-squared:                  0.742
Method:                 Least Squares   F-statistic:                     53.12
Date:                Tue, 28 Jan 2020   Prob (F-statistic):           2.41e-29
Time:                        10:33:06   Log-Likelihood:                -1199.0
No. Observations:                 110   AIC:                             2412.
Df Residuals:                     103   BIC:                             2431.
Df Model:                           6
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                     1.316e+04   1.91e+04      0.689      0.493   -2.47e+04    5.11e+04
Income                       0.5986      0.589      1.017      0.312      -0.569       1.766
Disposable Income            2.5350      0.732      3.464      0.001       1.084       3.986
Birth Rate (per 1,000)    1703.8657    563.673      3.023      0.003     585.954    2821.777
Soc Security (per 1,000)   -47.5162    110.213     -0.431      0.667    -266.097     171.065
CV Death (per 100,000)     -22.6821     31.464     -0.721      0.473     -85.084      39.720
% 65 or Older             7713.8505   1316.210      5.861      0.000    5103.458    1.03e+04
==============================================================================
Omnibus:                        0.673   Durbin-Watson:                   1.546
Prob(Omnibus):                  0.714   Jarque-Bera (JB):                0.789
Skew:                          -0.097   Prob(JB):                        0.674
Kurtosis:                       2.634   Cond. No.                     4.85e+05
==============================================================================
```

Multicollinearity/Collinearity

When there is multicollinearity in the sample there are strong linear relationships between $\mathbf{x}_1$,..., $\mathbf{x}_k$ in the sample

Effects of collinearity:

- Coefficient standard errors increase
- Fewer statistically significant slopes (t-ratios decrease and p-values increase)
- Difficulty interpreting coefficients
- Coefficients change as others come and go.

These effects become more evident as collinearity grows stronger.

Variance inflation factor (VIF)

$$\mathrm{VIF} = 1/(1 - R_j^2),$$

where R_j^2 is the R-square of a multiple regression without involving the original response variable y :

using $\mathbf{x}_j$ as the dependent variable and all other $\mathbf{x}$-variables $\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_k$ as independent variables.

Thus R_j^2 measures to what extent x_j depends linearly on other $\mathbf{x}$-variables on a scale from 0 to 1. Since VIF is an increasing function of R_j^2, it measures the degrees of collinearity in the data on a scale from 1 to ∞.

How to deal with multicollinearity?

If the VIF's of a $\mathbf{x}$-variable of the multiple regression is large (>10), indicating serious multicollinearity in the data, we can

- Drop the $\mathbf{x}$-variable from the regression
- Combine it with other $\mathbf{x}$-variable(s), or
- Keep it in the regression (even though the coefficient of the $\mathbf{x}$-variable may not be reliable, the prediction of the regression still is), if your goal is prediction.
- Other regression method (e.g. variable selection)

Case study: Retail profits

Continued

Calculation of VIF

```python
def getvif(X):
    X = sm.add_constant(X)
    vif = pd.DataFrame()
    vif["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
    vif["Predictors"] = X.columns
    return(vif.drop(index = 0).round(2))
```

```python
# For each X, calculate VIF and save in dataframe
from statsmodels.stats.outliers_influence import variance_inflation_factor
getvif(X)
```

	VIF	Predictors
1	2.95	Income
2	3.30	Disposable Income
3	1.70	Birth Rate (per 1,000)
4	10.04	Soc Security (per 1,000)
5	4.71	CV Death (per 100,000)
6	11.39	% 65 or Older

	VIF Factor	Predictors
0	132.6	Income
1	122.4	Disposable Income
2	16.4	Birth Rate (per 1,000)
3	155.6	Soc Security (per 1,000)
4	72.5	CV Death (per 100,000)
5	156.5	% 65 or Older

Adjusted R^2

Q: How do I know which model have better prediction performance?

Adjusted R^2

$$R^2_{adj} = \left[R^2 - \frac{k}{n-1}\right]\left[\frac{n-1}{n-k-1}\right]$$

- Each additional variable reduces adjusted R^2, unless SSE goes down enough to compensate

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} \leq 1 - \frac{SSE}{SST} = R^2$$

- Better estimate of the importance of the independent variables

Case study: Retail profits Continued

Step 4: the Final Model

```python
# drop insignificant x-variables
X_new = df_p.drop(columns=['Profit','Income','Soc Security (per 1,000)', 'CV Death (per 100,000)'])
```

```python
# Refit multiple regression model and show summary of fit
model_fit = sm.OLS(Y,sm.add_constant(X_new)).fit()
print(model_fit.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Profit   R-squared:                       0.752
Model:                            OLS   Adj. R-squared:                  0.745
Method:                 Least Squares   F-statistic:                     107.2
Date:                Tue, 28 Jan 2020   Prob (F-statistic):           5.73e-32
Time:                        10:58:46   Log-Likelihood:                -1199.8
No. Observations:                 110   AIC:                             2408.
Df Residuals:                     106   BIC:                             2418.
Df Model:                           3
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                 1.004e+04   1.53e+04      0.655      0.514   -2.04e+04    4.05e+04
Disposable Income        3.2386      0.414      7.828      0.000       2.418       4.059
Birth Rate (per 1,000) 1874.0454   526.501      3.559      0.001     830.206    2917.885
% 65 or Older         6619.2075    465.534     14.219      0.000    5696.241    7542.174
==============================================================================
```

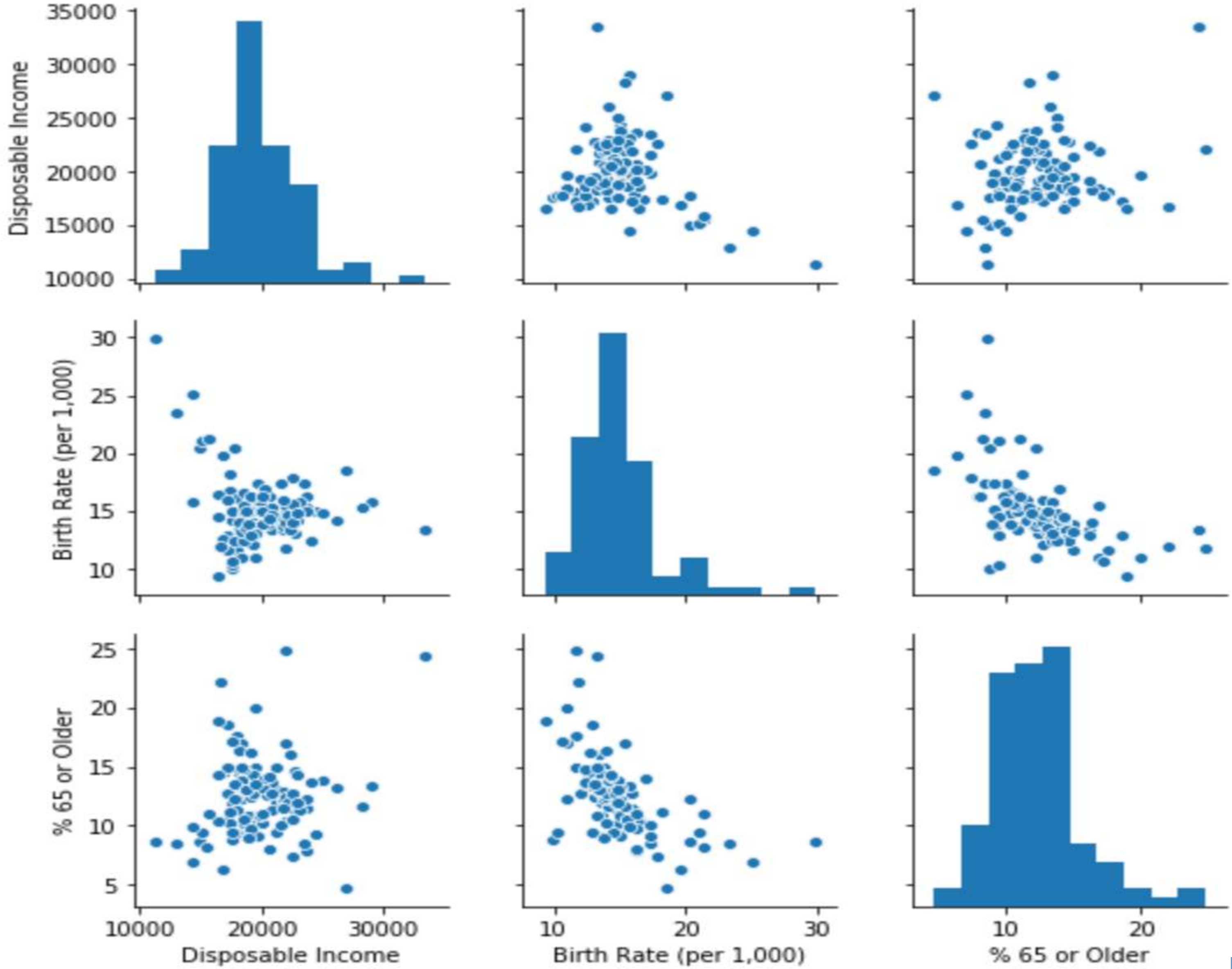
Note: please refer to Appendix 1 for the other Statistics listed in the Python output.

Impact of the selected variables

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.004e+04	1.53e+04	0.655	0.514	-2.04e+04	4.05e+04
Disposable Income	3.2386	0.414	7.828	0.000	2.418	4.059
Birth Rate (per 1,000)	1874.0454	526.501	3.559	0.001	830.206	2917.885
% 65 or Older	6619.2075	465.534	14.219	0.000	5696.241	7542.174

- For locations with the same proportion of residents 65 and older and comparable birth rate, pharmacy profits increase on average $2,400 to $4,100 per $1,000 increase in the median local disposable income.
- Comparing sites with the same disposable income and proportion of residents 65 and older , we expect $830 to $2,900 more in profits on average for each 1% increase in the birth rate by 1 per 1,000.
- Comparing sites with the same disposable income and birth rate, profits increase on average from $5,700 to $7,500 for each 1% increase in the percentage of the local population above 65.

Choosing new communities for expansion



(c) Examine sales at current locations to identify underperforming sites

Observation #	Standardized residual	Location
31	-2.29	Denver-Boulder, CO
90	-2.21	Rockford, IL

Step 5: Prediction

What is the estimated profit for a new store in Kansas City, MO? (median disposable income $22,642, 14.4 births per 1,000, and 11.4% 65 or more years old)

$$\text{Predicted Profit} = 10045 + 3.24\,(22{,}642) + 1874\,(14.4) + 6619\,(11.4) = \$185{,}819$$

```
Xnew = np.column_stack((1,22642,14.4,11.4))
Ynew = model_fit.predict(Xnew)

model_fit_get_prediction = model_fit.get_prediction(Xnew)
model_fit_get_prediction.summary_frame()
```

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	185818.710508	1809.324546	182231.548617	189405.872398	158901.859001	212735.562014

Business Implication

Three characteristics of the local community affect estimated profits: disposable income, age and birth rates. Increases in each of these lead to higher profits.

The data show that the pharmacy chain will have to trade off these characteristics in selecting a site for expansion.

Two current locations are underperforming sites: Denver-Boulder-Greeley, CO. and Rockford, IL.

Split the data in to training and testing set

How to pick models that can predict better?

Make comparisons through **training set** and **testing set**:

- 1. Divide the whole data set into training set and testing set
- 2. Use the **training set** to fit the models using different methods
- 3. Apply the fitted models to the **testing set**, and compare the prediction performance

Compare prediction performance using Adjusted R^2 or testing set RMSE

Adjusted R^2

$$R^2_{adjusted} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

where $R^2 = 1 - \frac{SSE}{SST}$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$$

RMSE

$$RMSE = \sqrt{SSE/(n-k-1)}$$

where $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Case study: Retail profits Continued

```python
# Split the data: training set 80%, test set 20%
training_data = df_p.sample(frac=0.8, random_state=25)
testing_data = df_p.drop(training_data.index)

print(f"No. of training examples: {training_data.shape[0]}")
print(f"No. of testing examples: {testing_data.shape[0]}")
```

```
No. of training examples: 88
No. of testing examples: 22
```

```python
## Full model
# Fit the full model in training set
model_fit_full = sm.OLS(Y_train, sm.add_constant(X_train)).fit()

# Predict Y on test set using the full model build in training set
Y_pred_full = model_fit_full.predict(sm.add_constant(X_test))
```

```python
## Final model
# drop insignificant x-variables
X_train_new = X_train.drop(columns=['Income','Soc Security (per 1,000)', 'CV Death (per 100,000)'])
X_test_new = X_test.drop(columns=['Income','Soc Security (per 1,000)', 'CV Death (per 100,000)'])

# build the final model on the training set
model_fit_final = sm.OLS(Y_train, sm.add_constant(X_train_new)).fit()

# Predict Y on test set using the final model build in training set
Y_pred_final = model_fit_final.predict(sm.add_constant(X_test_new))
```

```python
# Calculate the RMSE
from sklearn.metrics import mean_squared_error
mse_full = np.sqrt(mean_squared_error(Y_test, Y_pred_full))
mse_final = np.sqrt(mean_squared_error(Y_test, Y_pred_final))
print(f"Test set RMSE (Full model) : {mse_full}")
print(f"Test set RMSE (Final model): {mse_final}")
```

```
Test set RMSE (Full model) : 14007.541348265508
Test set RMSE (Final model): 13989.19276250275
```

Take away from Topic 6

Analysis using multiple linear regression

Scatter plot matrix

- Model assumption
- The over all model is useful?
  - Overall F-test,
  - R^2
- Check for multicollinearity – VIF
- Predictor useful?
  - Individual t-test,
  - Coefficient estimate, CI for coefficient

Choose final model

Compare prediction performance using adjusted R^2

Use model for prediction or explanation

Appendix 1

Statistical Methods and Concepts Reported in Python Computer Output

Omnibus test is a statistical test of residual normality, the smaller the test statistic value the better, or equivalently, the larger P(Omnibus) the better.

Skewness, the standardized 3rd central moment, is a measure of symmetry of residuals – the closer to zero the better.

Kurtosis, the standardized 4th central moment, is a measure of heaviness of residual tails – the closer to 3 the better.

Jarque-Bera test is another test of residual normality based on skewness and kurtosis. The smaller the JB test statistic value the better, or equivalently, the larger P(JB) the better.

Durbin-Watson test is a statistical test of residual independence based on lag one autocorrelation. The closer DW statistic to 2 the better.

Condition number is the ratio of maximum to minimum eigenvalue of Gramian matrix X^TX, where X has n rows (correspond to n observation) and p columns (corresponds to p x-variables). A large conditional number indicates collinearity among independent variables.