Module 2
Lesson 3 - Multiple Regression

WESLEYAN
UNIVERSITY

```python
# center quantitative IVs for regression analysis
sub1['numbercigsmoked_c'] = (sub1['numbercigsmoked'] - sub1['numbercigsmoked'].mean())

print (sub1['numbercigsmoked_c'].mean())
```

IPython console

Console 2/A

```
In [24]: print (sub1['numbercigsmoked_c'].mean())
    ...:
-1.8371020199707542e-14

In [25]:
```

0.00000000000001837

```
# bivariate bar graph
seaborn.factorplot(x="MAJORDEPLIFE", y="NDSymptoms", data=sub2, kind="bar", ci=None)
plt.xlabel('Major Life Depression')
plt.ylabel('Mean Number Nicotine Dependence Symptoms')


# center quantitative IVs for regression analysis
sub1['numbercigsmoked_c'] = (sub1['numbercigsmoked'] - sub1['numbercigsmoked'].mean())

print (sub1['numbercigsmoked_c'].mean())


reg2 = smf.ols('NDSymptoms ~ MAJORDEPLIFE + numbercigsmoked_c', data=sub1).fit()
print (reg2.summary())
```

IPython console

Console 2/A

```
In [25]: reg2 = smf.ols('NDSymptoms ~ MAJORDEPLIFE + numbercigsmoked_c', data=sub1).fit()
    ...: print (reg2.summary())
    ...:
                            OLS Regression Results
==============================================================================
Dep. Variable:             NDSymptoms   R-squared:                       0.132
Model:                            OLS   Adj. R-squared:                  0.131
Method:                 Least Squares   F-statistic:                     99.87
Date:                Fri, 23 Oct 2015   Prob (F-statistic):           4.28e-41
Time:                        13:47:53   Log-Likelihood:                -2593.9
No. Observations:                1313   AIC:                             5194.
Df Residuals:                    1310   BIC:                             5209.
Df Model:                           2
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------------
Intercept             2.1946      0.056     38.940      0.000         2.084     2.305
MAJORDEPLIFE          1.3424      0.109     12.327      0.000         1.129     1.556
numbercigsmoked_c     0.0358      0.006      6.432      0.000         0.025     0.047
==============================================================================
Omnibus:                       70.355   Durbin-Watson:                   2.066
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               50.154
Skew:                           0.372   Prob(JB):                     1.29e-11
Kurtosis:                       2.398   Cond. No.                         20.4
==============================================================================

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


Disthymia
Number Nicotine Dependence symptoms

IPython console

reg3 = smf.ols('NDSymptoms ~ DYSLIFE', data=sub1).fit()
print (reg3.summary())

```
    ...:
                            OLS Regression Results
==============================================================================
Dep. Variable:             NDSymptoms   R-squared:                       0.023
Model:                            OLS   Adj. R-squared:                  0.022
Method:                 Least Squares   F-statistic:                     30.35
Date:                Fri, 23 Oct 2015   Prob (F-statistic):           4.34e-08
Time:                        13:49:37   Log-Likelihood:                -2679.2
No. Observations:                1317   AIC:                             5362.
Df Residuals:                    1315   BIC:                             5373.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      2.4785      0.053     46.958      0.000         2.375     2.582
DYSLIFE        1.1378      0.207      5.509      0.000         0.733     1.543
==============================================================================
Omnibus:                      123.634   Durbin-Watson:                   2.079
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               63.384
Skew:                           0.374   Prob(JB):                     1.72e-14
Kurtosis:                       2.228   Cond. No.                         4.07
==============================================================================

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

IPython console

reg4 = smf.ols('NDSymptoms ~ DYSLIFE + MAJORDEPLIFE', data=sub1).fit()
print (reg4.summary())

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             NDSymptoms   R-squared:                       0.106
Model:                            OLS   Adj. R-squared:                  0.105
Method:                 Least Squares   F-statistic:                     78.23
Date:                Fri, 23 Oct 2015   Prob (F-statistic):           7.98e-33
Time:                        13:53:26   Log-Likelihood:                -2620.2
No. Observations:                1317   AIC:                             5246.
Df Residuals:                    1314   BIC:                             5262.
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
Intercept        2.1808      0.057     38.152      0.000         2.069     2.293
DYSLIFE          0.3477      0.210      1.656      0.098        -0.064     0.760
MAJORDEPLIFE     1.2993      0.117     11.103      0.000         1.070     1.529
==============================================================================
Omnibus:                       70.772   Durbin-Watson:                   2.061
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               51.246
Skew:                           0.380   Prob(JB):                     7.45e-12
Kurtosis:                       2.402   Cond. No.                         4.61
==============================================================================

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

In [28]:
```

Object inspector | Variable explorer | File explorer

reg5 = smf.ols('NDSymptoms ~ DYSLIFE + MAJORDEPLIFE + numcigsmoked_c + age_c + SEX", data=sub1).fit()
print (reg4.summary())

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             NDSymptoms   R-squared:                       0.136
Model:                            OLS   Adj. R-squared:                  0.133
Method:                 Least Squares   F-statistic:                     41.08
Date:                Fri, 23 Oct 2015   Prob (F-statistic):           2.39e-39
Time:                        13:56:07   Log-Likelihood:                -2591.2
No. Observations:                1313   AIC:                             5194.
Df Residuals:                    1307   BIC:                             5226.
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------------
Intercept             2.2550      0.156     14.480      0.000         1.949     2.561
DYSLIFE               0.2746      0.209      1.316      0.188        -0.135     0.684
MAJORDEPLIFE          1.2975      0.116     11.161      0.000         1.069     1.526
numbercigsmoked_c     0.0353      0.006      6.257      0.000         0.024     0.046
age_c                -0.0400      0.022     -1.806      0.071        -0.083     0.003
SEX                  -0.0439      0.099     -0.442      0.658        -0.238     0.151
==============================================================================
Omnibus:                       69.558   Durbin-Watson:                   2.075
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               48.596
Skew:                           0.361   Prob(JB):                     2.80e-11
Kurtosis:                       2.394   Cond. No.                         38.5
==============================================================================

Warnings:
```

```
reg5 = smf.ols('NDSymptoms ~ DYSLIFE + MAJORDEPLIFE + numcigsmoked_c + age_c + SEX', data=sub1).fit()
print (reg4.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             NDSymptoms   R-squared:                       0.138
Model:                            OLS   Adj. R-squared:                  0.135
Method:                 Least Squares   F-statistic:                     41.88
Date:                Fri, 23 Oct 2015   Prob (F-statistic):           2.38e-39
Time:                        13:56:07   Log-Likelihood:                -2591.2
No. Observations:                1313   AIC:                             5194.
Df Residuals:                    1307   BIC:                             5226.
Df Model:                           5
Covariance Type:            nonrobust
==================================================================================
                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept           2.2550      0.156     14.480      0.000         1.949     2.561
DYSLIFE             0.2746      0.209      1.316      0.188        -0.135     0.684
MAJORDEPLIFE        1.2975      0.116     11.161      0.000         1.069     1.526
numbercigsmoked_c   0.0353      0.006      6.257      0.000         0.024     0.046
age_c              -0.0400      0.022     -1.806      0.071        -0.083     0.003
SEX                -0.0439      0.099     -0.442      0.658        -0.238     0.151
==============================================================================
Omnibus:                       69.558   Durbin-Watson:                   2.075
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               48.596
Skew:                           0.361   Prob(JB):                     2.80e-11
Kurtosis:                       2.394   Cond. No.                         38.5
==============================================================================

Warnings:
```

Module 2

Lesson 4 - Confidence Intervals

WESLEYAN UNIVERSITY

```
    ...]
                            OLS Regression Results
==============================================================================
Dep. Variable:             NDSymptoms   R-squared:                       0.136
Model:                            OLS   Adj. R-squared:                  0.133
Method:                 Least Squares   F-statistic:                     41.08
Date:                Fri, 23 Oct 2015   Prob (F-statistic):           2.39e-39
Time:                        13:56:07   Log-Likelihood:                -2591.2
No. Observations:                1313   AIC:                             5194.
Df Residuals:                    1307   BIC:                             5226.
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                         coef    std err          t      P>|t|   [95.0% Conf. Int.]
-------------------------------------------------------------------------------------
Intercept              2.2550      0.156     14.400      0.000       1.949     2.561
DYSLIFE                0.2746      0.209      1.316      0.188      -0.135     0.684
MAJORDEPLIFE           1.2975      0.116     11.161      0.000       1.069     1.526
numbercigsmoked_c      0.0353      0.006      6.257      0.000       0.024     0.046
age_c                 -0.0400      0.022     -1.806      0.071      -0.083     0.003
SEX                   -0.0439      0.099     -0.442      0.658      -0.238     0.151
==============================================================================
Omnibus:                       69.558   Durbin-Watson:             [illegible]
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               48.596
Skew:                           0.361   Prob(JB):                     2.80e-11
Kurtosis:                       2.394   Cond. No.                         38.5
==============================================================================

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

Permissions: RW    End-of-lines: CRLF    Encoding: UTF-8    Line: 81    Column: 1    Memory: 28 %

```
   ...:
                            OLS Regression Results
==============================================================================
Dep. Variable:             NDSymptoms   R-squared:                       0.136
Model:                            OLS   Adj. R-squared:                  0.133
Method:                 Least Squares   F-statistic:                     41.08
Date:                Fri, 23 Oct 2015   Prob (F-statistic):           2.39e-39
Time:                        13:56:07   Log-Likelihood:                -2591.2
No. Observations:                1313   AIC:                             5194.
Df Residuals:                    1307   BIC:                             5226.
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                         coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------------
Intercept              2.2550      0.156     14.480      0.000         1.949     2.561
DYSLIFE                0.2746      0.209      1.316      0.188        -0.135     0.684
MAJORDEPLIFE           1.2975      0.116     11.161  [illegible]       1.069     1.526
numbercigsmoked_c      0.0353      0.006      6.257      0.000         0.024     0.046
age_c                 -0.0400      0.022     -1.806      0.071        -0.083     0.003
SEX                   -0.0439      0.099     -0.442      0.658        -0.238     0.151
==============================================================================
Omnibus:                       69.558   Durbin-Watson:                   2.075
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               48.596
Skew:                           0.361   Prob(JB):                     2.80e-11
Kurtosis:                       2.394   Cond. No.                         38.5
==============================================================================

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

Permissions: RW   End-of-lines: CRLF   Encoding: UTF-8   Line: 81   Column: 1   Memory: 28 %

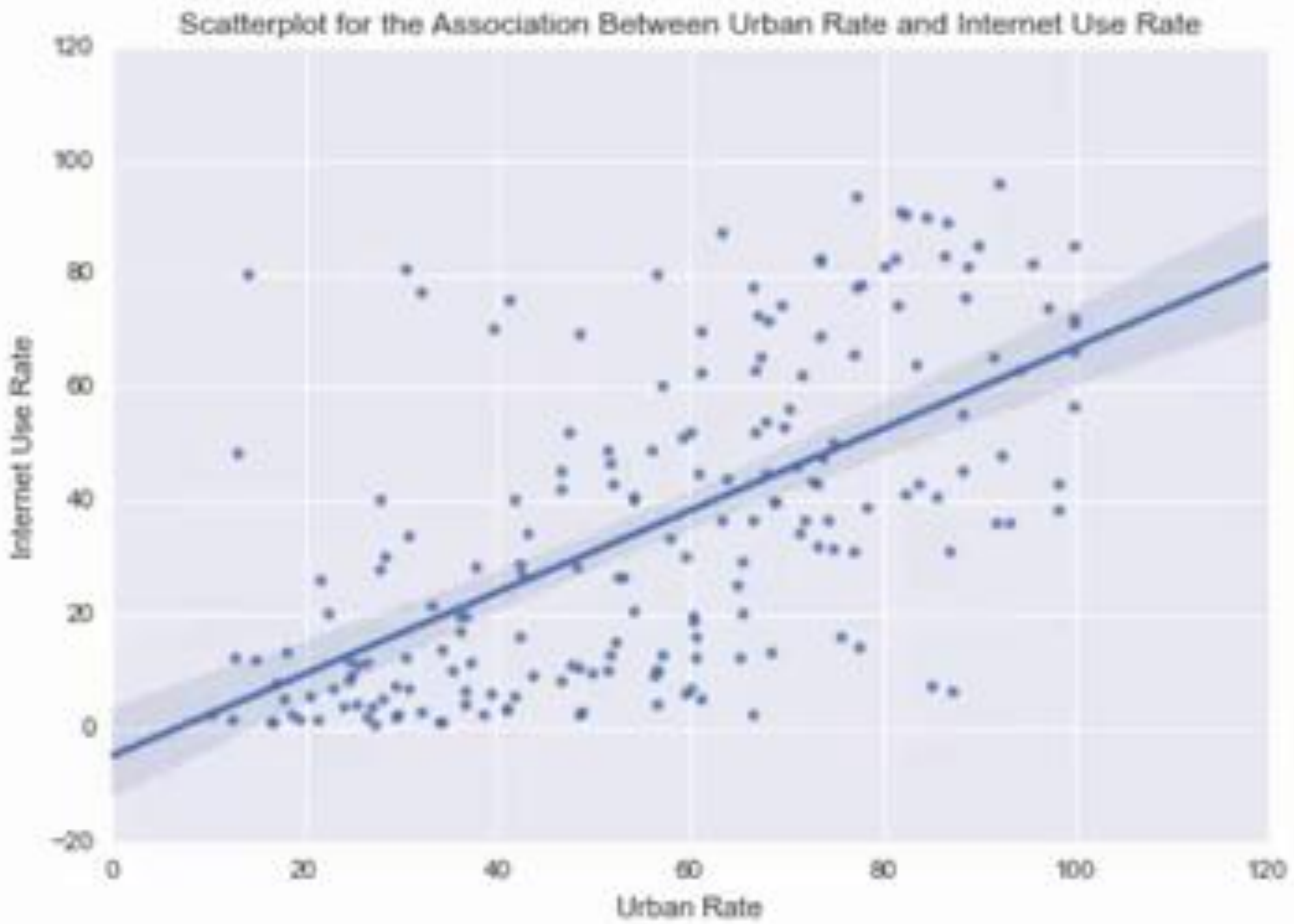
Module 3

Lesson 3 - Polynomial Regression

WESLEYAN UNIVERSITY

IPython console

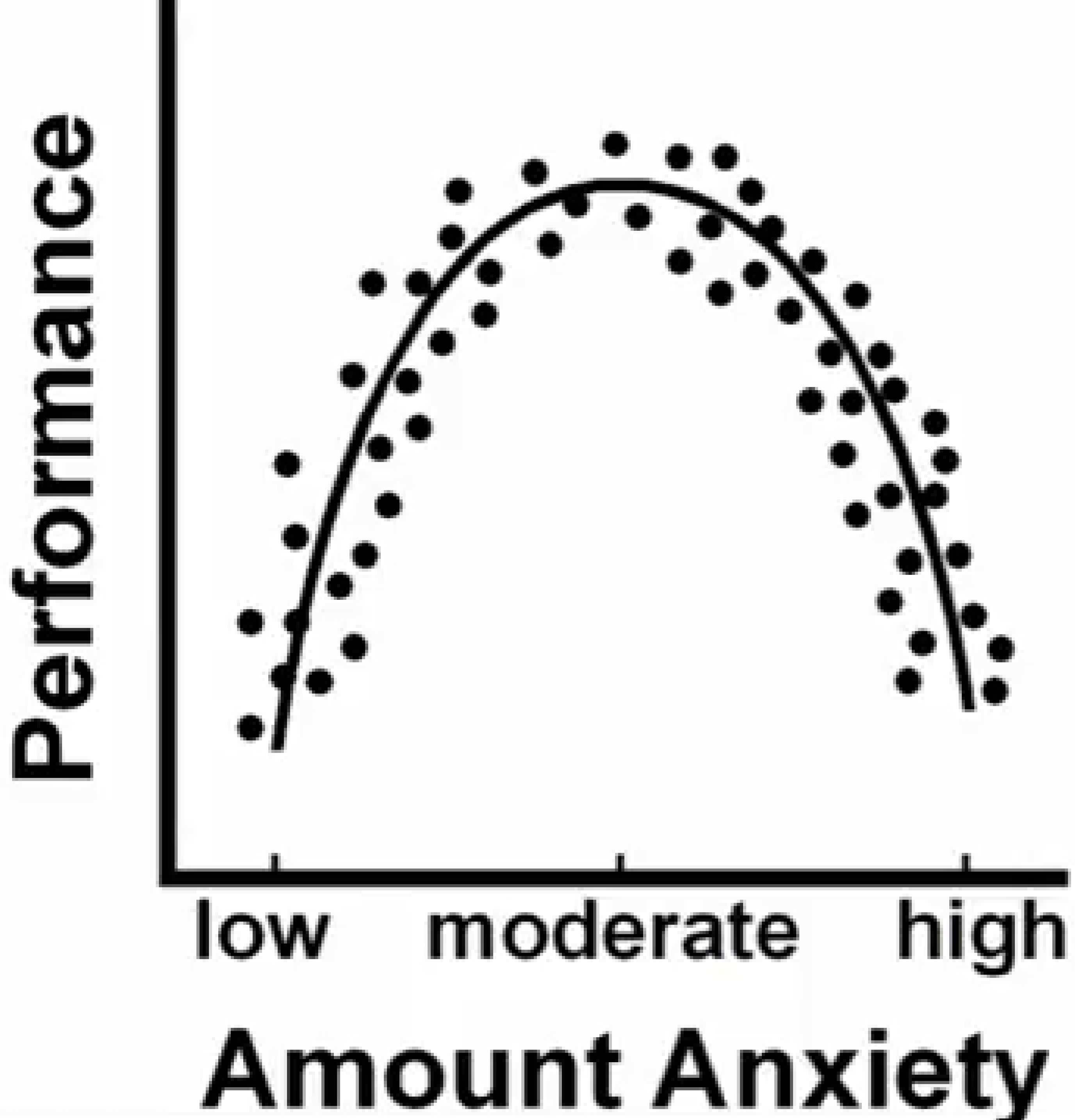
Console 1/A




```
In [3]:
```

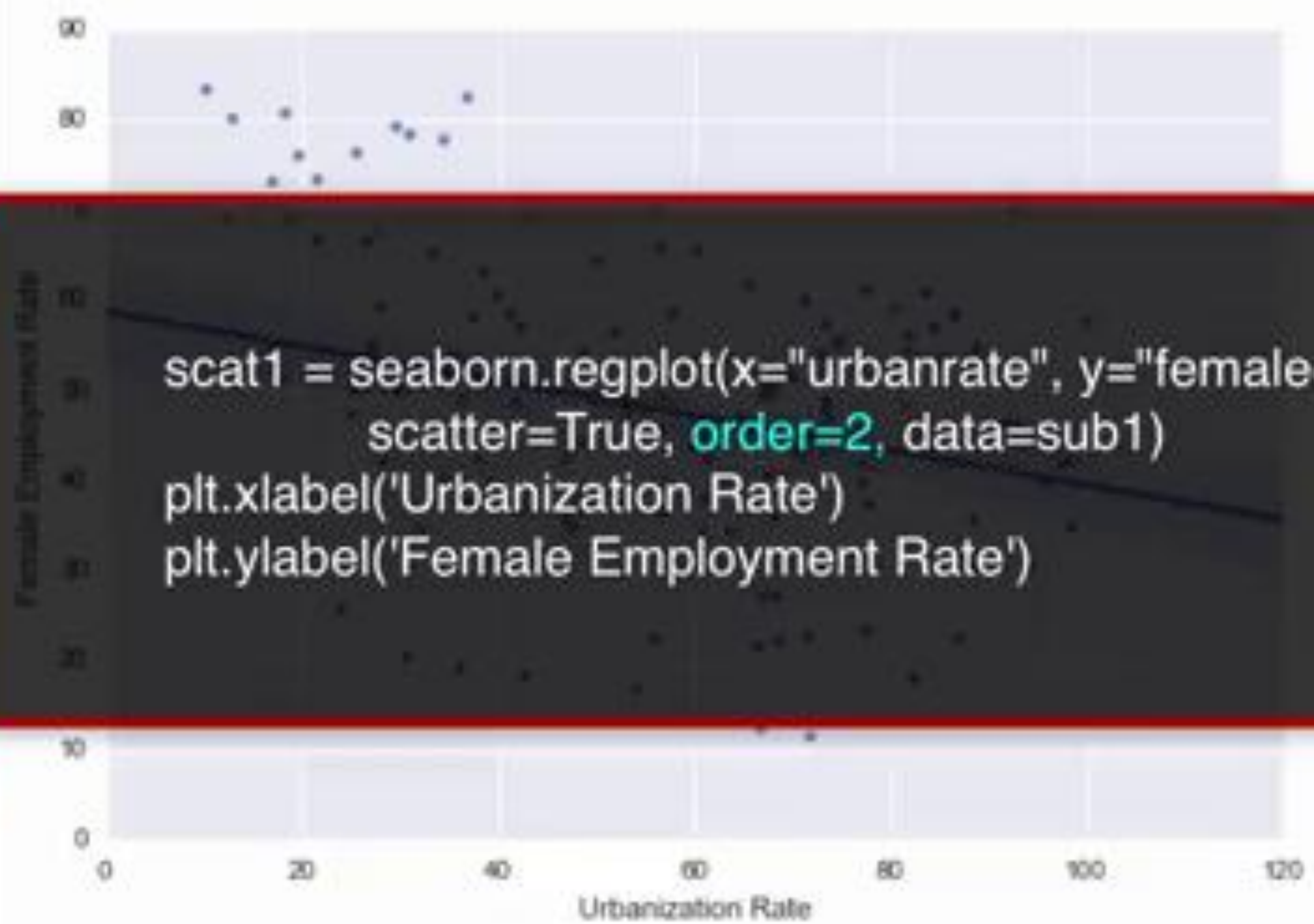
Console | IPython console | History log

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 27 Column: 1 Memory: 83 %


Performance
low
moderate
high
Amount Anxiety

IPython console

Console 1/A    Console 2/A

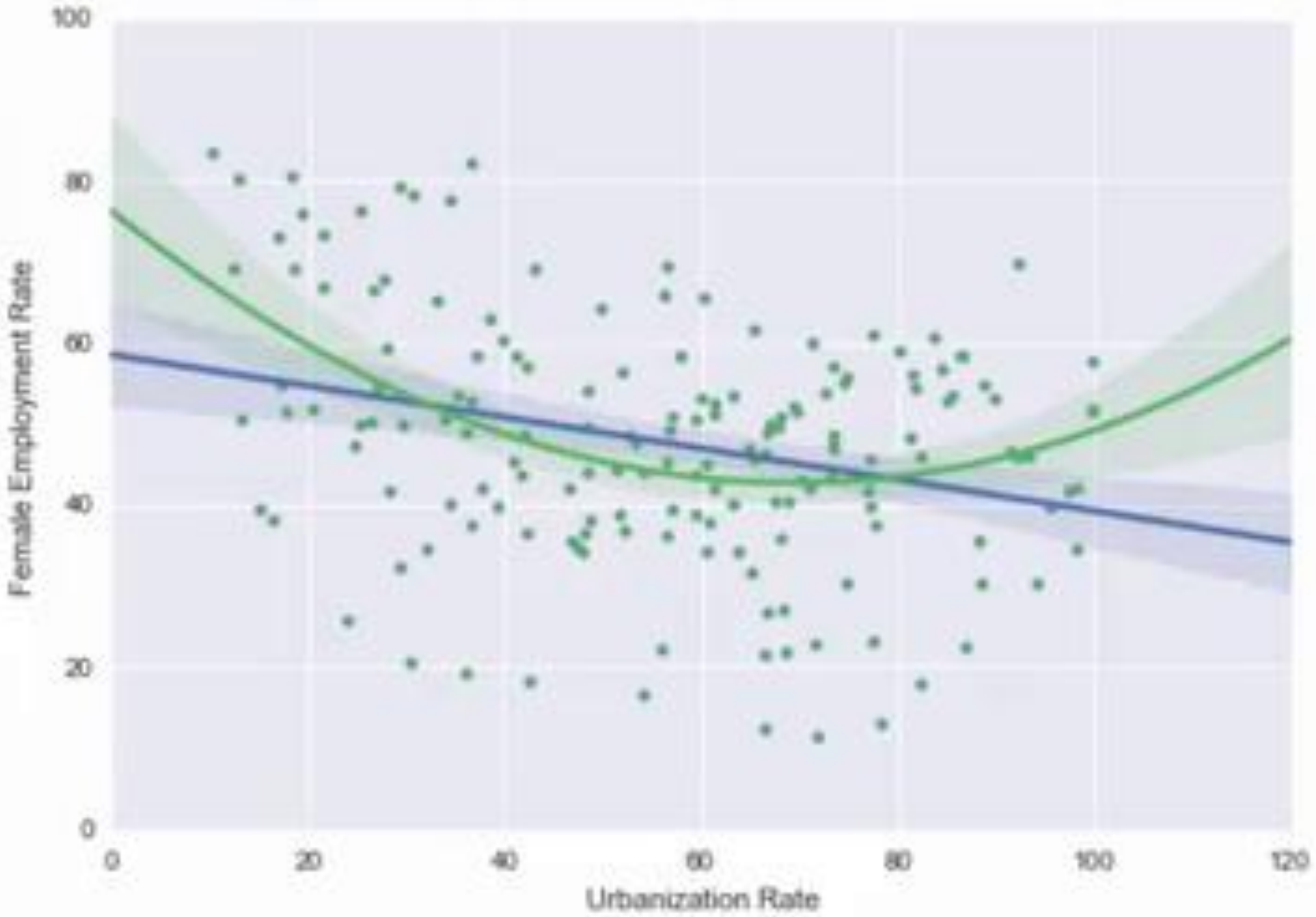



```
scat1 = seaborn.regplot(x="urbanrate", y="femaleemployrate",
              scatter=True, order=2, data=sub1)
plt.xlabel('Urbanization Rate')
plt.ylabel('Female Employment Rate')
```

In [14]:

IPython console

Console 1/A Console 2/A




```
In [16]:
```

```
data['femaleemployrate'] = pandas.to_numeric(data['femaleemployrate'], errors='coerce')
del I

sub1 = data[['urbanrate', 'femaleemployrate', 'internetuserate']].dropna()


scat1 = seaborn.regplot(x="urbanrate", y="femaleemployrate", scatter=True, data=sub1)
plt.xlabel('Urbanization Rate')
plt.ylabel('Female Employment Rate')

scat1 = seaborn.regplot(x="urbanrate", y="femaleemployrate", scatter=True, order=2, data=sub1)
plt.xlabel('Urbanization Rate')
plt.ylabel('Female Employment Rate')

# center quantitative IVs for regression analysis
sub1['urbanrate_c'] = (sub1['urbanrate'] - sub1['urbanrate'].mean())
sub1['internetuserate_c'] = (sub1['internetuserate'] - sub1['internetuserate'].mean())

reg1 = smf.ols('femaleemployrate ~ urbanrate_c', data=sub1).fit()
print (reg1.summary())
```

Console 1/A Console 2/A

```
In [21]: reg1 = smf.ols('femaleemployrate ~ urbanrate_c', data=sub1).fit()
    ...: print (reg1.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:       femaleemployrate   R-squared:                       0.092
Model:                            OLS   Adj. R-squared:                  0.086
Method:                 Least Squares   F-statistic:                     16.69
Date:                Fri, 23 Oct 2015   Prob (F-statistic):           6.84e-05
Time:                        14:41:44   Log-Likelihood:                -678.68
No. Observations:                 167   AIC:                             1361.
Df Residuals:                     165   BIC:                             1368.
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------
Intercept      47.6024      1.096     43.416      0.000        45.438    49.767
urbanrate_c    -0.1927      0.047     -4.086      0.000        -0.286    -0.100
==============================================================================
Omnibus:                        2.347   Durbin-Watson:                   1.868
Prob(Omnibus):                  0.309   Jarque-Bera (JB):                2.409
Skew:                          -0.269   Prob(JB):                        0.300
Kurtosis:                       2.763   Cond. No.                         23.2
==============================================================================

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

In [22]:
```

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 58 Column: 1 Memory: 22 %

Console 1/A Console 2/A

```
In [21]: reg1 = smf.ols('femaleemployrate ~ urbanrate_c', data=sub1).fit()
    ...: print (reg1.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:       femaleemployrate   R-squared:                       0.092
Model:                            OLS   Adj. R-squared:                  0.086
Method:                 Least Squares   F-statistic:                     16.69
Date:                Fri, 23 Oct 2015   Prob (F-statistic):           6.84e-05
Time:                        14:41:44   Log-Likelihood:                -678.68
No. Observations:                 167   AIC:                             1361.
Df Residuals:                     165   BIC:                             1368.
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------
Intercept      47.6024      1.096     43.416      0.000        45.438    49.767
urbanrate_c    -0.1927      0.047     -4.086      0.000        -0.286    -0.100
==============================================================================
Omnibus:                        2.347   Durbin-Watson:                   1.868
Prob(Omnibus):                  0.309   Jarque-Bera (JB):                2.409
Skew:                          -0.269   Prob(JB):                        0.300
Kurtosis:                       2.763   Cond. No.                         23.2
==============================================================================

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

In [22]:
```

Permissions: RW   End-of-lines: CRLF   Encoding: UTF-8   Line: 58   Column: 1   Memory: 22 %

```
scat1 = seaborn.regplot(x="urbanrate", y="femaleemployrate", scatter=True, order=2, data=sub1)
plt.xlabel('Urbanization Rate')
plt.ylabel('Female Employment Rate')

# center quantitative IVs for regression analysis
sub1['urbanrate_c'] = (sub1['urbanrate'] - sub1['urbanrate'].mean())
sub1['internetuserate_c'] = (sub1['internetuserate'] - sub1['internetuserate'].mean())

reg1 = smf.ols('femaleemployrate ~ urbanrate_c', data=sub1).fit()
print (reg1.summary())


# regression model with second order polynomial (quadratic term)
reg2 = smf.ols('femaleemployrate ~ urbanrate_c + I(urbanrate_c**2)', data=sub1).fit()
print (reg2.summary())
```

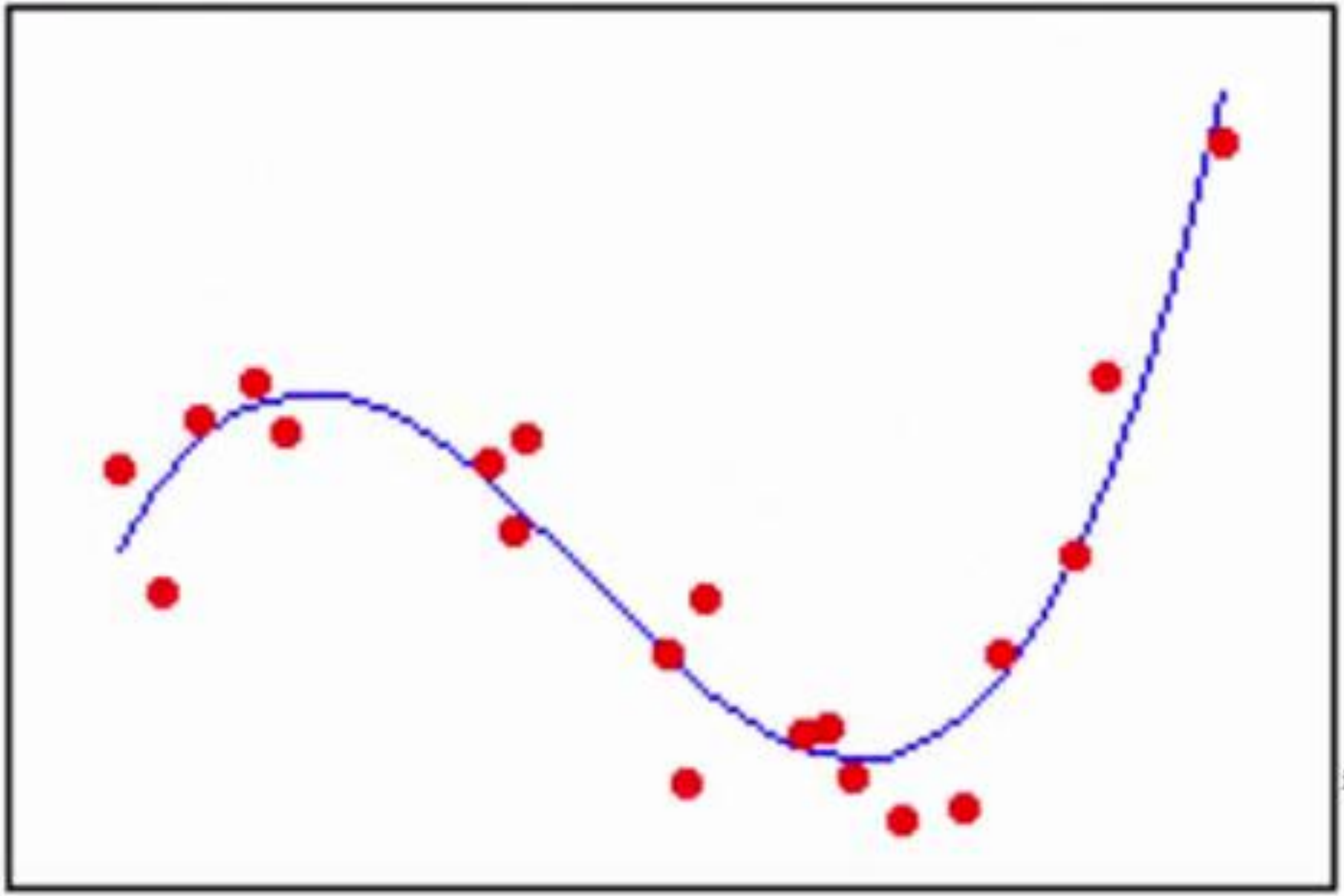
```
                            OLS Regression Results
==============================================================================
Dep. Variable:      femaleemployrate   R-squared:                       0.160
Model:                           OLS   Adj. R-squared:                  0.150
Method:                Least Squares   F-statistic:                     15.60
Date:               Fri, 23 Oct 2015   Prob (F-statistic):           6.30e-07
Time:                       14:45:30   Log-Likelihood:                -672.19
No. Observations:                167   AIC:                             1350.
Df Residuals:                    164   BIC:                             1360.
Df Model:                          2
Covariance Type:           nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------------
Intercept              43.8428      1.478     29.659      0.000        40.924    46.762
urbanrate_c            -0.1751      0.046     -3.827      0.000        -0.266    -0.085
I(urbanrate_c ** 2)     0.0070      0.002      3.641      0.000         0.003     0.011
==============================================================================
Omnibus:                       3.627   Durbin-Watson:                   1.898
Prob(Omnibus):                 0.163   Jarque-Bera (JB):                3.677
Skew:                         -0.351   Prob(JB):                        0.159
Kurtosis:                      2.811   Cond. No.                     1.08e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.08e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

In [24]:
```

Permissions: RW   End-of-lines: CRLF   Encoding: UTF-8   Line: 58   Column: 1   Memory: 22 %

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      femaleemployrate   R-squared:                       0.160
Model:                           OLS   Adj. R-squared:                  0.150
Method:                Least Squares   F-statistic:                     15.60
Date:               Fri, 23 Oct 2015   Prob (F-statistic):           6.30e-07
Time:                       14:45:30   Log-Likelihood:                -672.19
No. Observations:                167   AIC:                             1350.
Df Residuals:                    164   BIC:                             1360.
Df Model:                          2
Covariance Type:           nonrobust
========================================================================================
                          coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------------
Intercept              43.8428      1.478     29.659      0.000        40.924    46.762
urbanrate_c            -0.1751      0.046     -3.827      0.000        -0.266    -0.085
I(urbanrate_c ** 2)     0.0070      0.002      3.641      0.000         0.003     0.011
==============================================================================
Omnibus:                       3.627   Durbin-Watson:                   1.898
Prob(Omnibus):                 0.163   Jarque-Bera (JB):                3.677
Skew:                         -0.351   Prob(JB):                        0.159
Kurtosis:                      2.811   Cond. No.                     1.08e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.08e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

In [24]:
```

Permissions: RW    End-of-lines: CRLF    Encoding: UTF-8    Line: 58    Column: 1    Memory: 22 %

OLS Regression Results

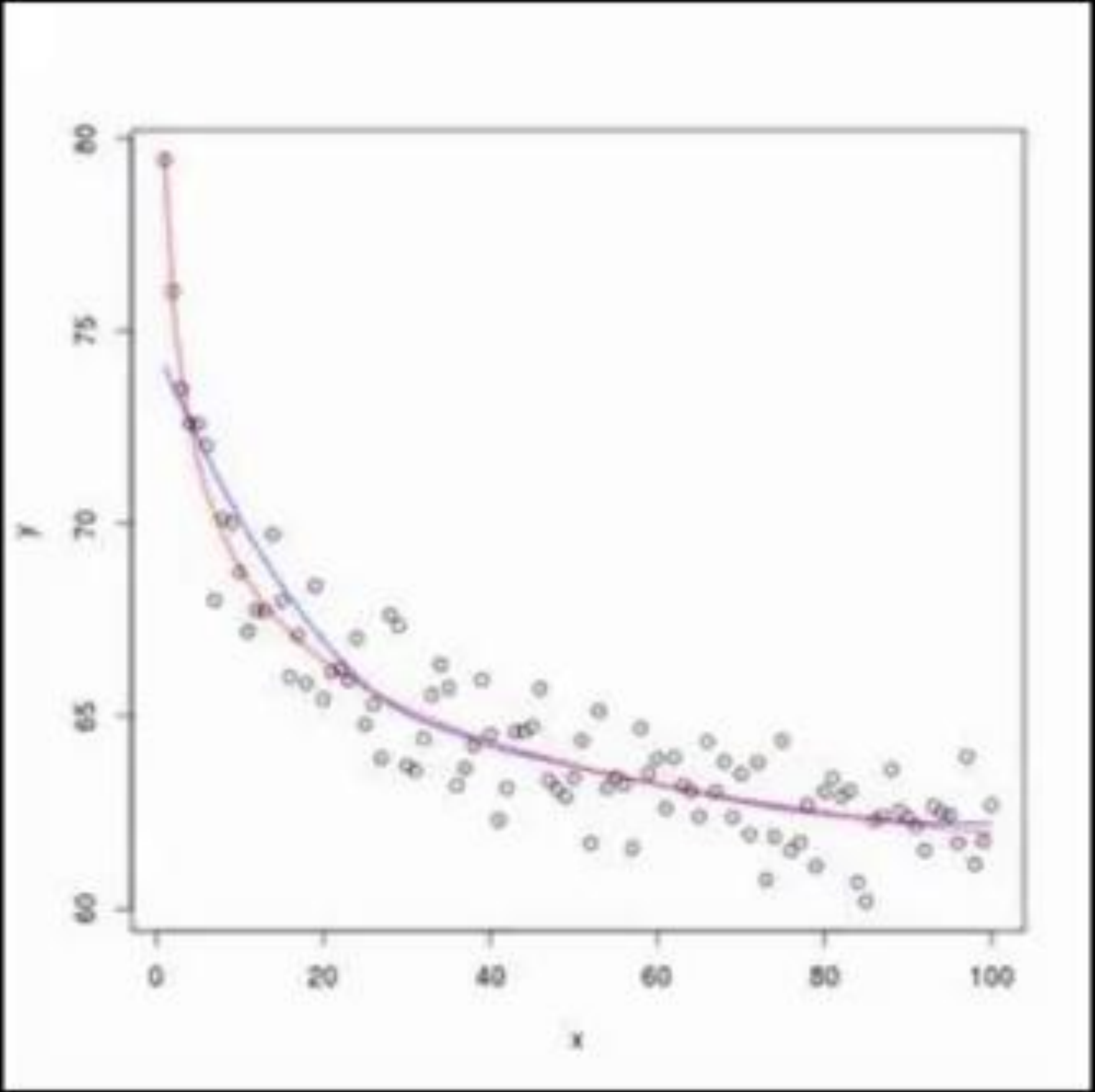


ied.

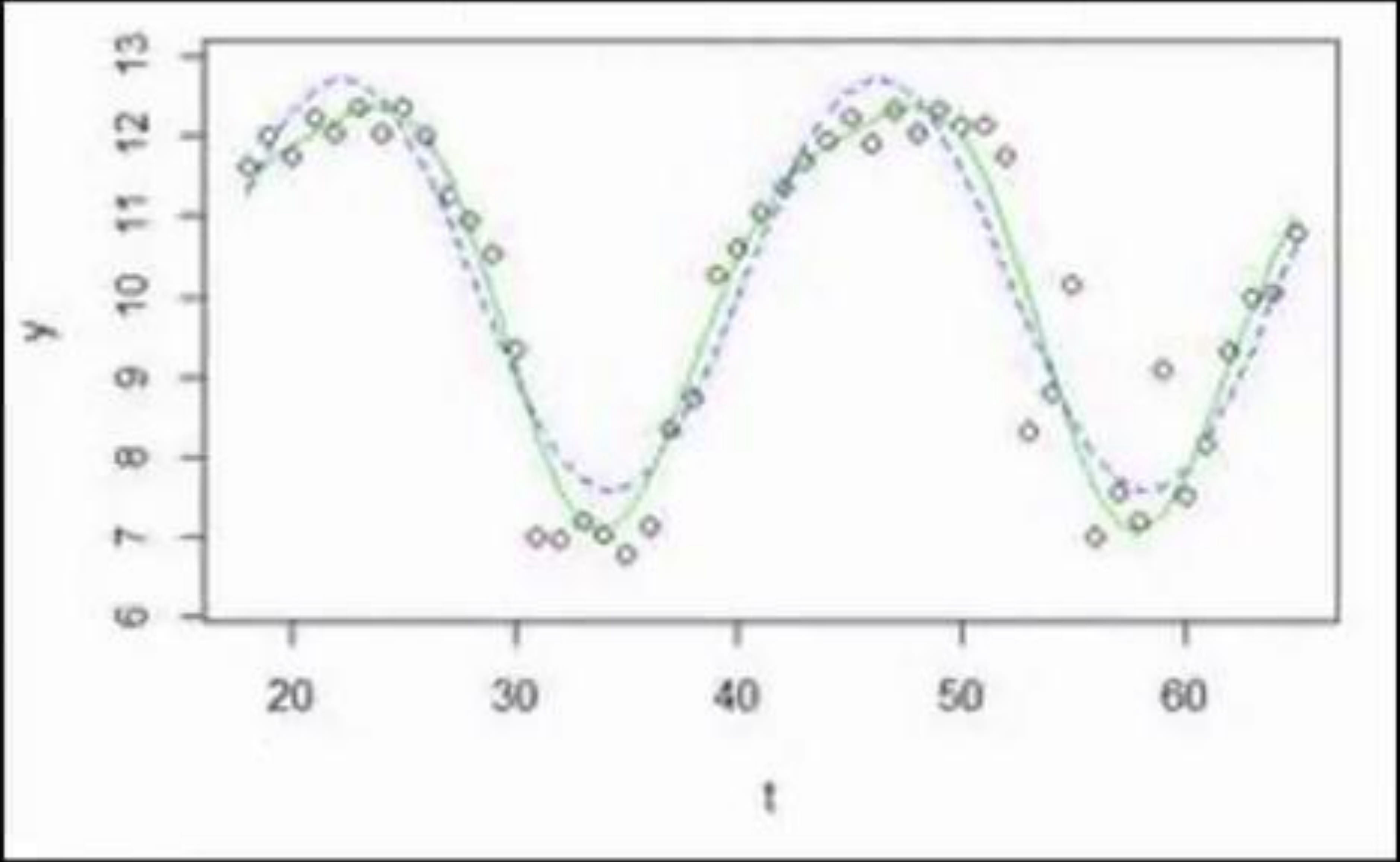
In [24]: |

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 58 Column: 1 Memory: 22 %

# Overfitting



# Overfitting

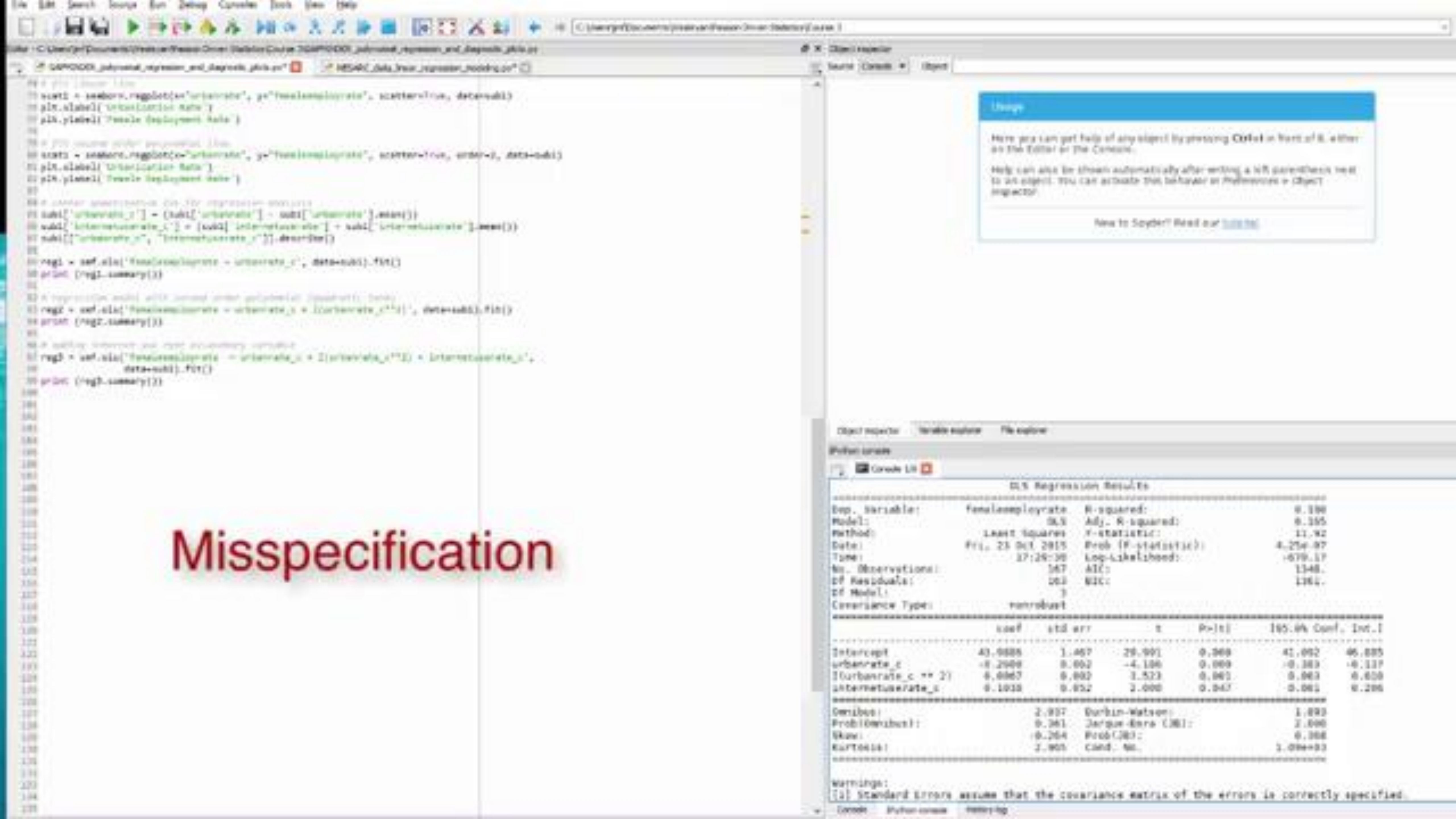


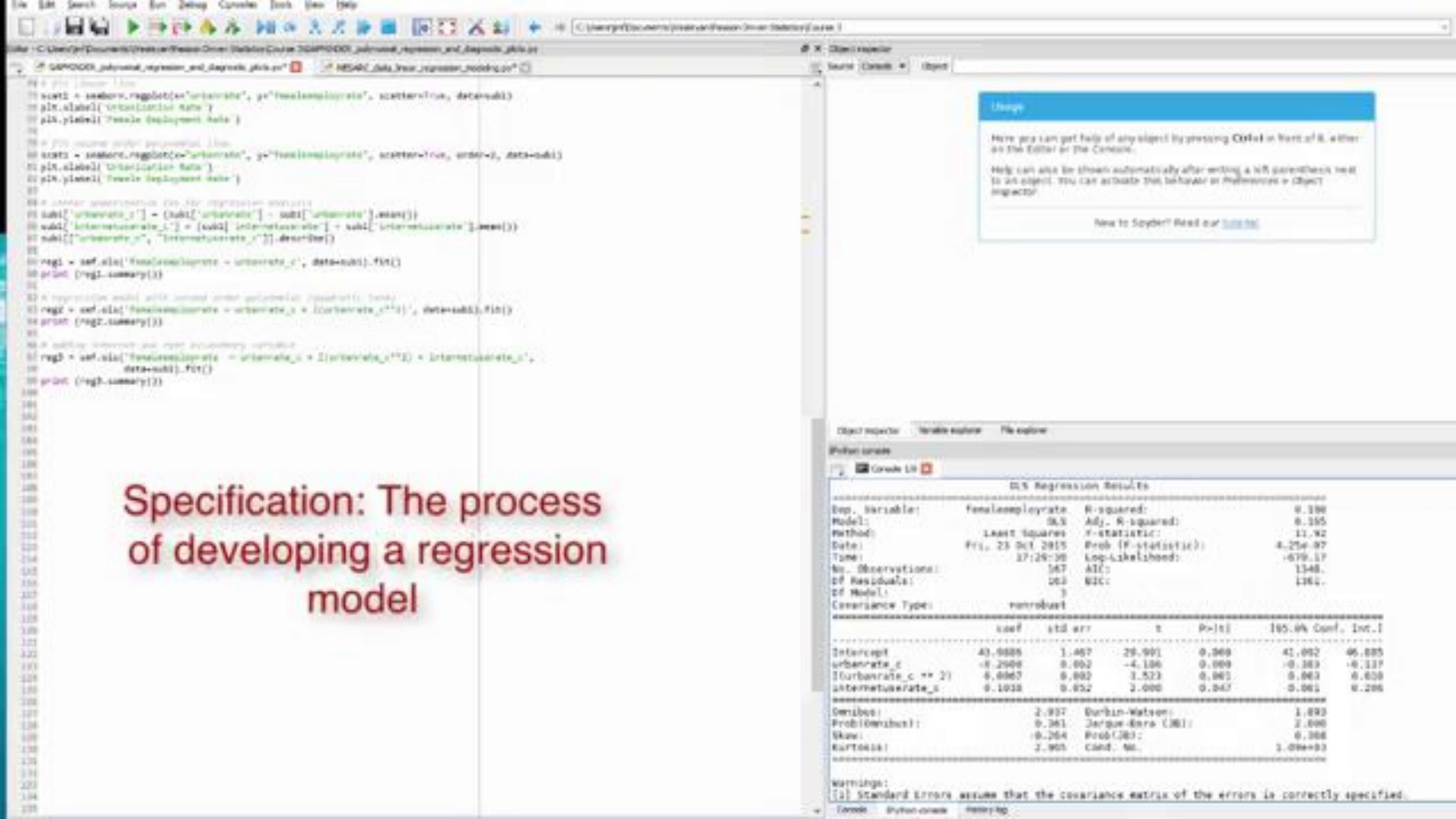
# Bias-Variance Tradeoff

Module 3

Lesson 4 - Evaluating Model Fit, part 1

WESLEYAN UNIVERSITY


Misspecification
OLS Regression Results


Specification: The process of developing a regression model
OLS Regression Results

```python
# center quantitative IVs for regression analysis
sub1['urbanrate_c'] = (sub1['urbanrate'] - sub1['urbanrate'].mean())
sub1['internetuserate_c'] = (sub1['internetuserate'] - sub1['internetuserate'].mean())
sub1[["urbanrate_c", "internetuserate_c"]].describe()

reg1 = smf.ols('femaleemployrate ~ urbanrate_c', data=sub1).fit()
print (reg1.summary())

# regression model with second order polynomial (quadratic term)
reg2 = smf.ols('femaleemployrate ~ urbanrate_c + I(urbanrate_c**2)', data=sub1).fit()
print (reg2.summary())

# adding internet use rate exlanatory variable
reg3 = smf.ols('femaleemployrate  ~ urbanrate_c + I(urbanrate_c**2) + internetuserate_c',
               data=sub1).fit()
print (reg3.summary())
```

DFFITS Statistic

Bartlett's Test

Kolmogorov-Smirnoff Test

Variance Inflation Factor

Durbin-Watson Test

```python
# center quantitative IVs for regression analysis
sub1['urbanrate_c'] = (sub1['urbanrate'] - sub1['urbanrate'].mean())
sub1['internetuserate_c'] = (sub1['internetuserate'] - sub1['internetuserate'].mean())
sub1[["urbanrate_c", "internetuserate_c"]].describe()

reg1 = smf.ols('femaleemployrate ~ urbanrate_c', data=sub1).fit()
print (reg1.summary())

# regression model with second order polynomial (quadratic term)
reg2 = smf.ols('femaleemployrate ~ urbanrate_c + I(urbanrate_c**2)', data=sub1).fit()
print (reg2.summary())

# adding internet use rate exlanatory variable
reg3 = smf.ols('femaleemployrate  ~ urbanrate_c + I(urbanrate_c**2) + internetuserate_c',
                data=sub1).fit()
print (reg3.summary())
```

$$\text{Femaleemployrate} = \beta_0 + \beta_1(\text{urbanrate}) + \beta_2(\text{urbanrate})^2 + \beta_3(\text{internetuserate}) + \varepsilon$$

Console 1/A

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      femaleemployrate   R-squared:                       0.180
Model:                           OLS   Adj. R-squared:                  0.165
Method:                Least Squares   F-statistic:                     11.92
Date:               Fri, 23 Oct 2015   Prob (F-statistic):           4.25e-07
Time:                       17:29:30   Log-Likelihood:                -670.17
No. Observations:                167   AIC:                             1348.
Df Residuals:                    163   BIC:                             1361.
Df Model:                          3
Covariance Type:           nonrobust
=========================================================================================
                          coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------------
Intercept              43.9886      1.467     29.991      0.000        41.092    46.885
urbanrate_c            -0.2600      0.062     -4.186      0.000        -0.383    -0.137
I(urbanrate_c ** 2)     0.0067      0.002      3.523      0.001         0.003     0.010
internetuserate_c       0.1038      0.052      2.000      0.047         0.001     0.206
==============================================================================
Omnibus:                       2.037   Durbin-Watson:                   1.893
Prob(Omnibus):                 0.361   Jarque-Bera (JB):                2.000
Skew:                         -0.264   Prob(JB):                        0.368
Kurtosis:                      2.905   Cond. No.                     1.09e+03
==============================================================================

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

Console IPython console History log

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 136 Column: 1 Memor

```
reg3 = smf.ols('femaleemployrate  ~ urbanrate_c + I(urbanrate_c**2) + internetuserate_c',
                 data=sub1).fit()
print (reg3.summary())



##########################################################################################
# Regression diagnostic plots
##########################################################################################

#Q-Q plot for normality
fig1=sm.qqplot(reg3.resid, line='r')
```

What we're looking for is to see if the points follow a straight line.

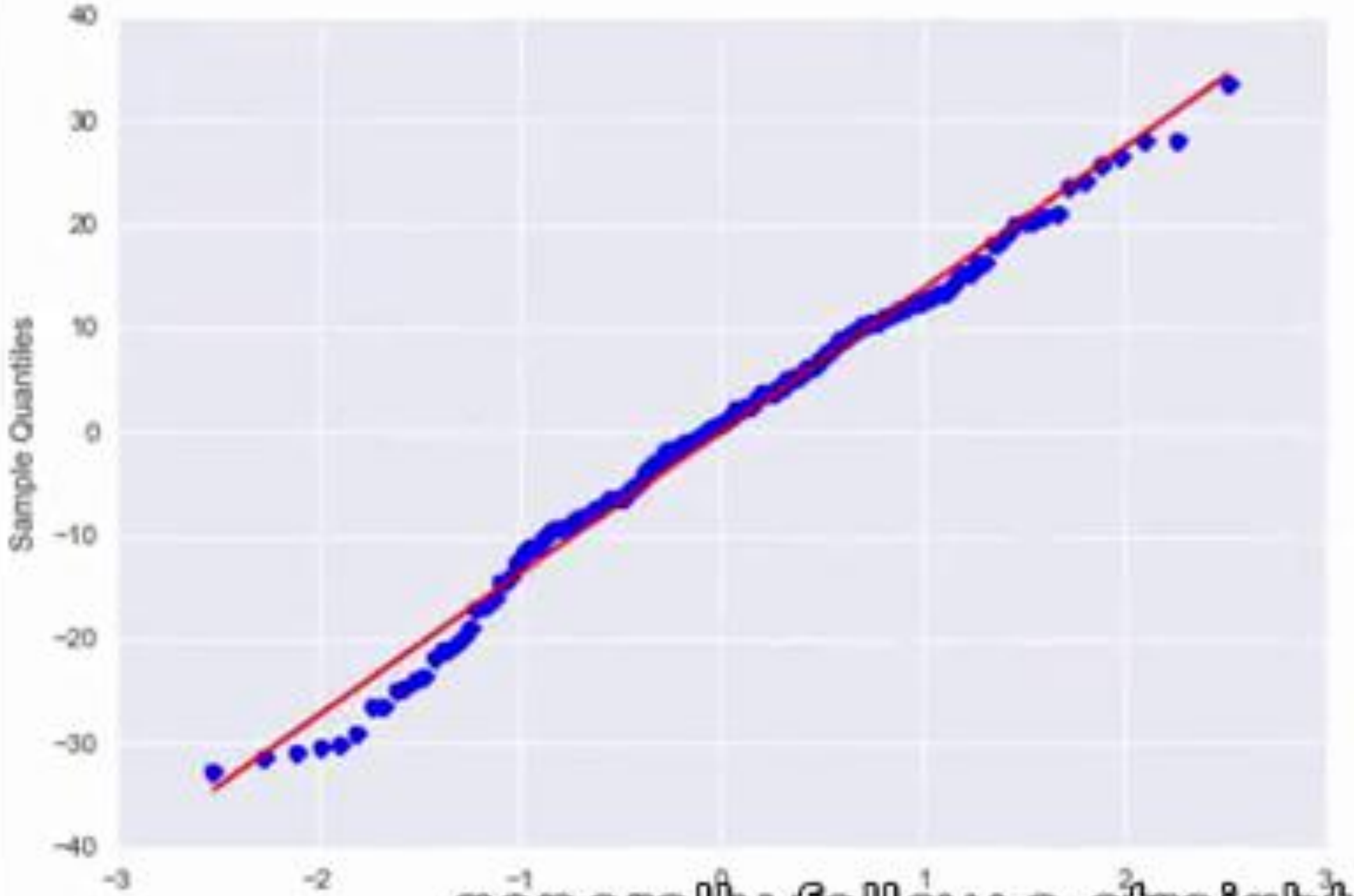
```
reg3 = smf.ols('femaleemployrate  ~ urbanrate_c + I(urbanrate_c**2) + internetuserate_c',
               data=sub1).fit()
print (reg3.summary())



##########################################################################################
# Regression diagnostic plots
##########################################################################################

#Q-Q plot for normality
fig1=sm.qqplot(reg3.resid, line='r')
```

Meaning that the model estimated residuals are what we would expect

```
reg3 = smf.ols('femaleemployrate  ~ urbanrate_c + I(urbanrate_c**2) + internetuserate_c',
                data=sub1).fit()
print (reg3.summary())



#############################################################################
# Regression diagnostic plots
#############################################################################

#Q-Q plot for normality
fig1=sm.qqplot(reg3.resid, line='r')
```

if the residuals were
normally distributed.

Console 1/A

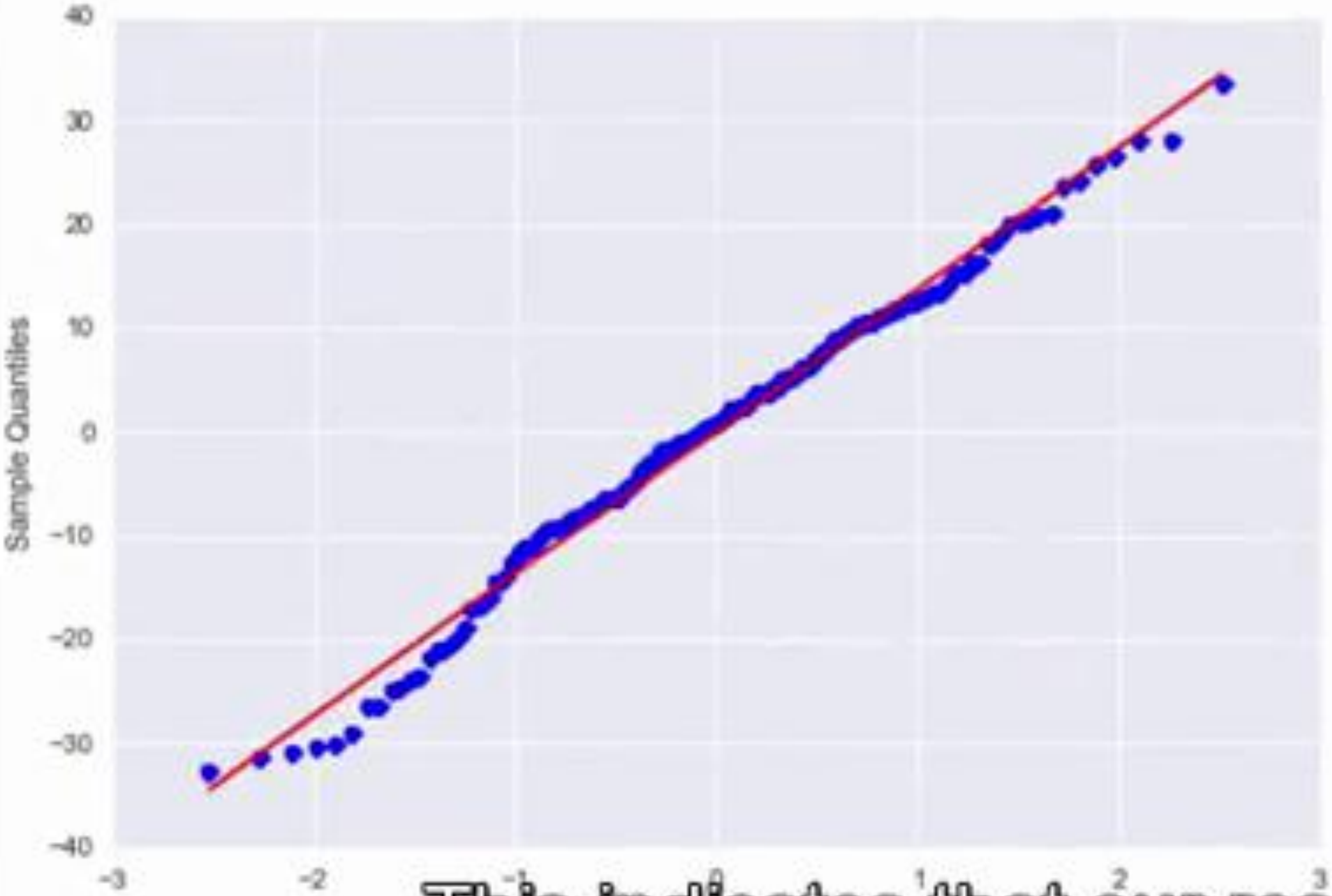



generally follow a straight line, but
deviate at the lower and higher quantiles.

In [95]:

Console IPython console History log

Console 1/A




This indicates that our residuals did not follow perfect normal distribution.

In [95]:

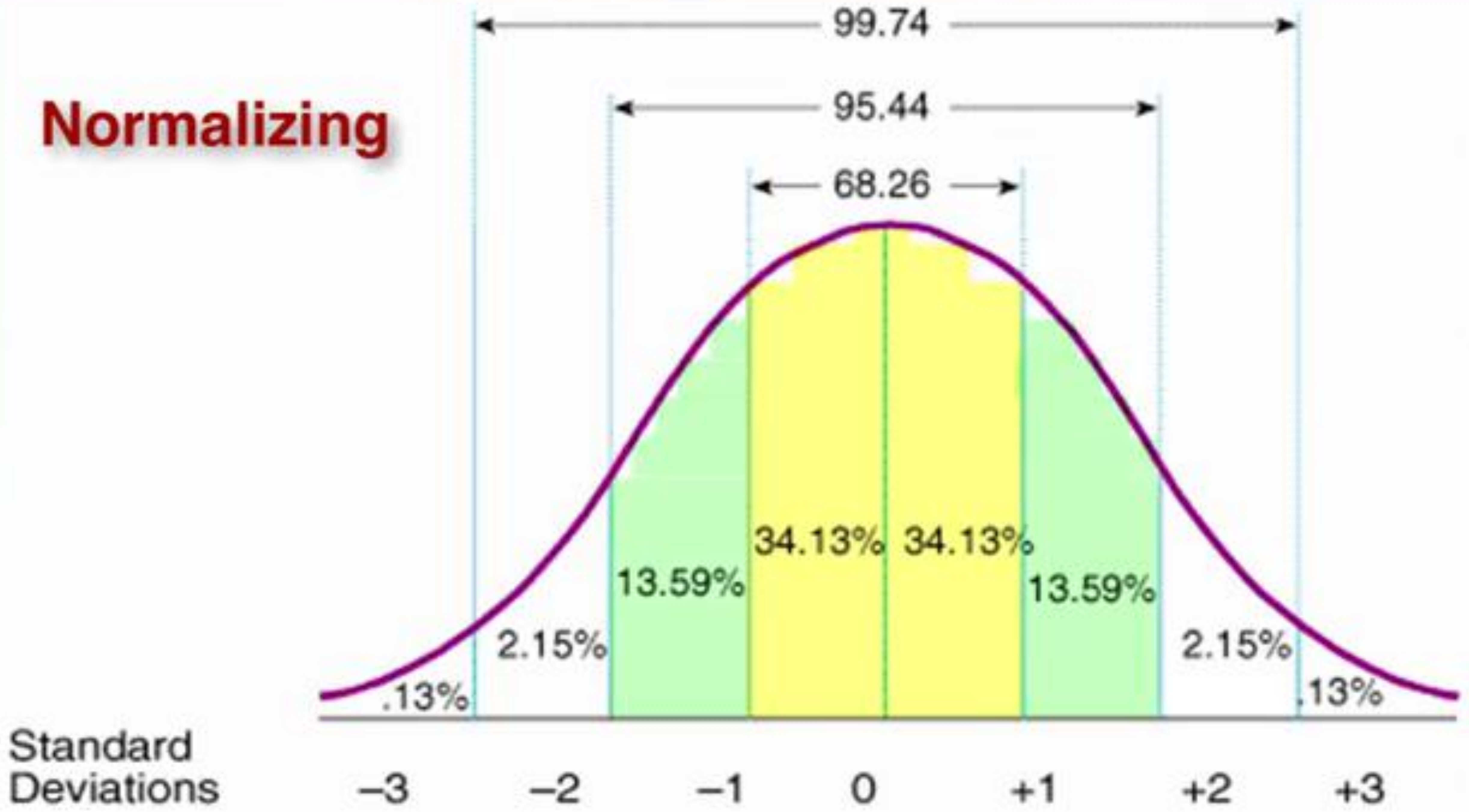
Console IPython console History log

Module 3

Lesson 5 - Evaluating Model Fit, part 2

WESLEYAN UNIVERSITY

# Normalizing



```
# Regression diagnostic plots

fig1=sm.qqplot(reg3.resid, line='r')

stdres=pandas.DataFrame(reg3.resid_pearson)
fig2=plt.plot(stdres, 'o', ls='None')
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')
print(fig2)
```

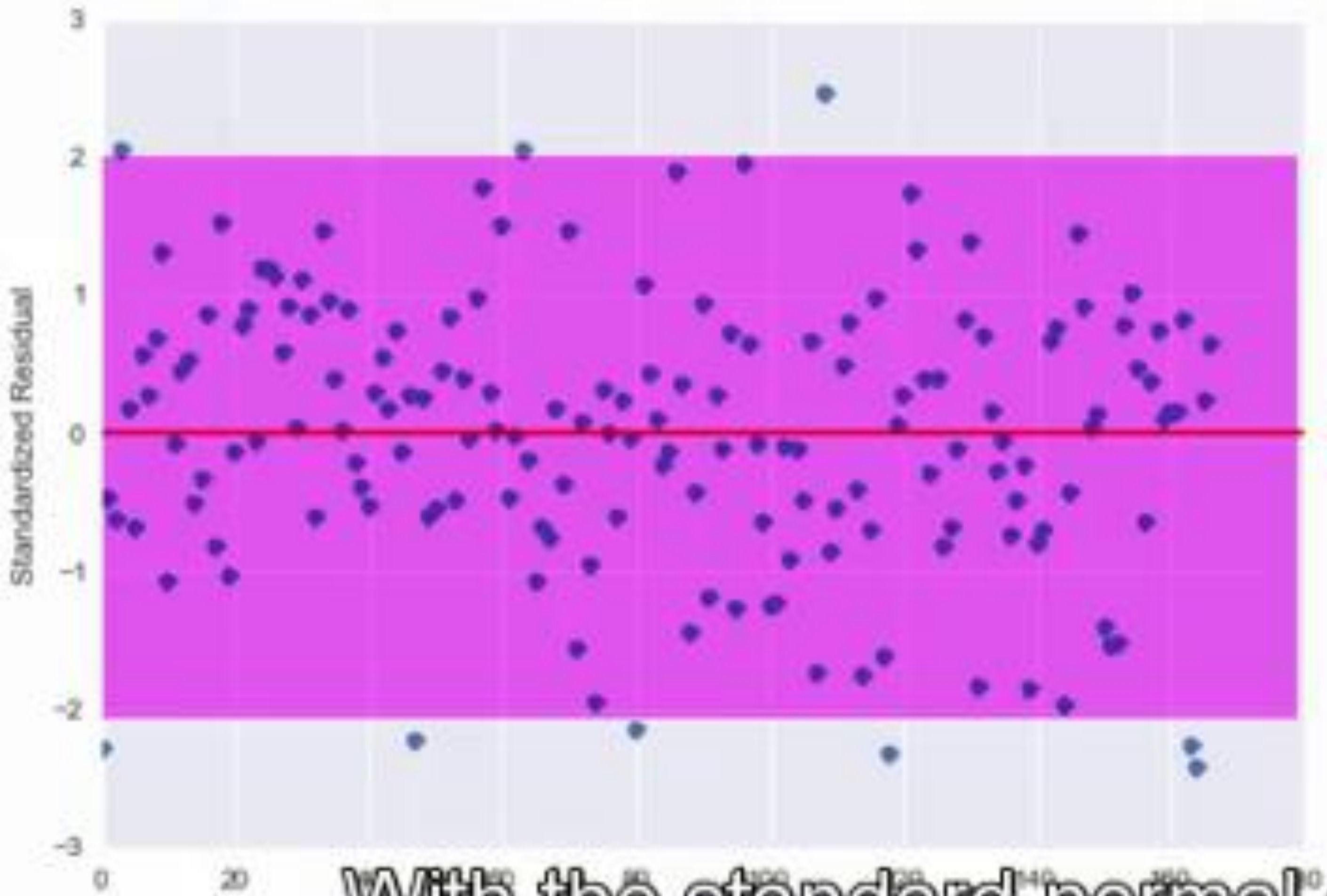
```
print(fig1)


# simple plot of residuals
stdres=pandas.DataFrame(reg3.resid_pearson)
fig2=plt.plot(stdres, 'o', ls='None')
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')
print(fig2)
```

```
# additional regression diagnostic plots
fig3 = plt.figure(figsize(12,8))
fig3 = sm.graphics.plot_regress_exog(reg3,  "internetuserate_c", fig=fig3)
```

IPython console

Console 1/A

```
   ....:
[<matplotlib.lines.Line2D object at 0x000000000BD56C18>]
```





With the standard normal distribution,
we would expect 95% of the values of

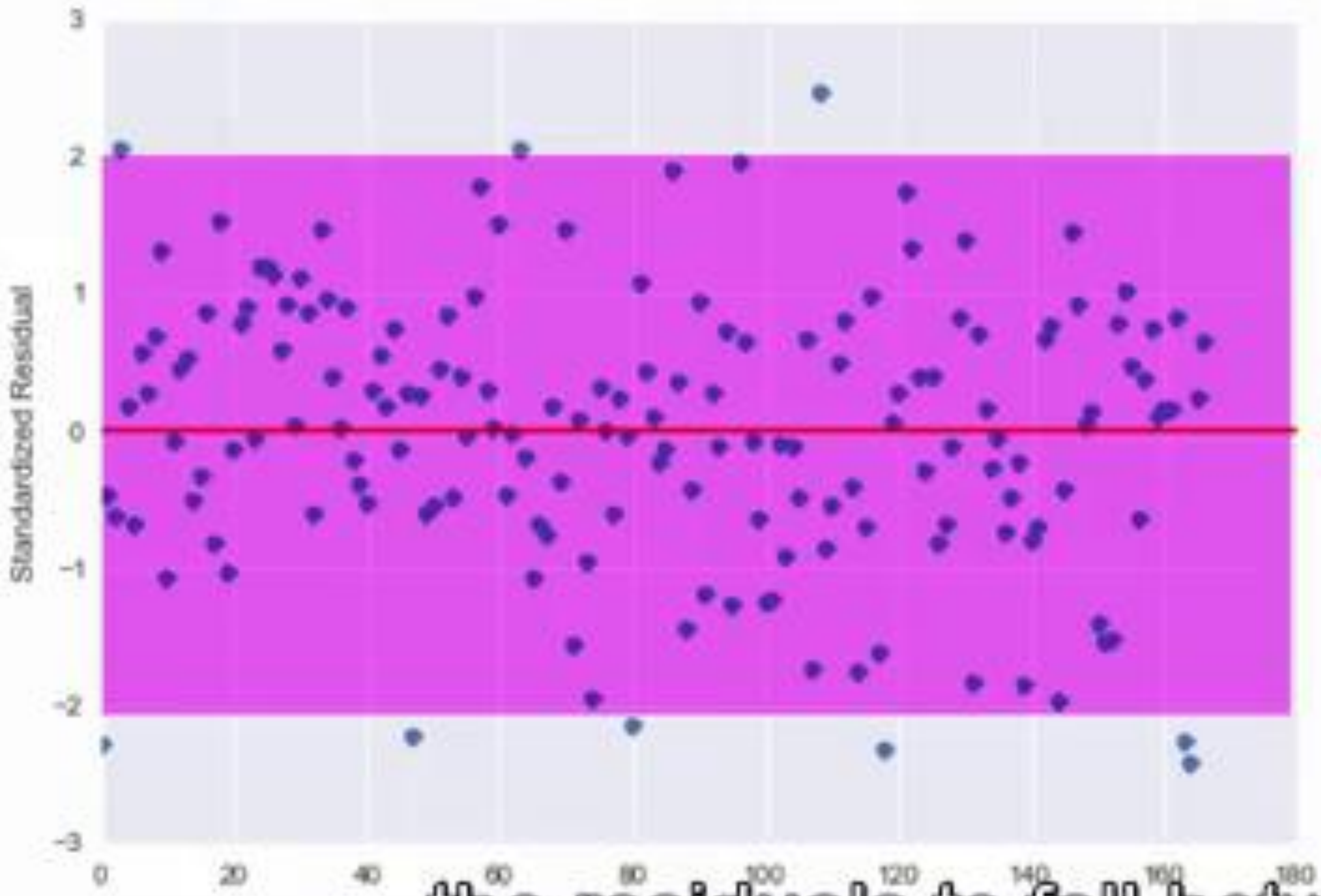
Console | IPython console

Permissions: RW | End-of-lines: CRLF | Encoding: UTF-8 | Line: 148 | Column: 1 | Memory: 71

IPython console

Console 1/A

```
...:
[<matplotlib.lines.Line2D object at 0x000000000BD56C18>]
```





the residuals to fall between two standard deviations of the mean.

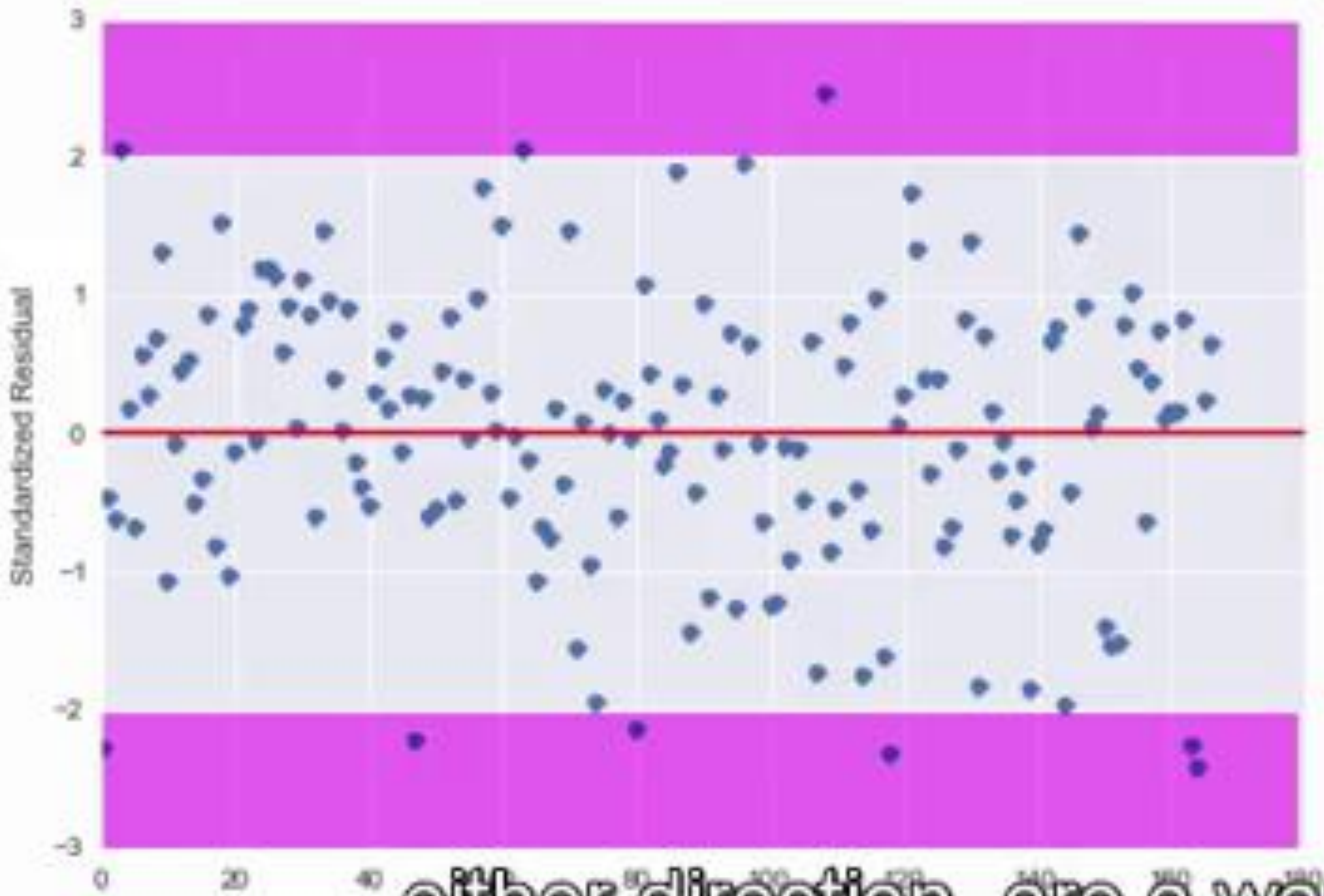
Console IPython console

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 148 Column: 1 Memory: 71

IPython console

Console 1/A

```
....
[<matplotlib.lines.Line2D object at 0x000000000BD56C18>]
```

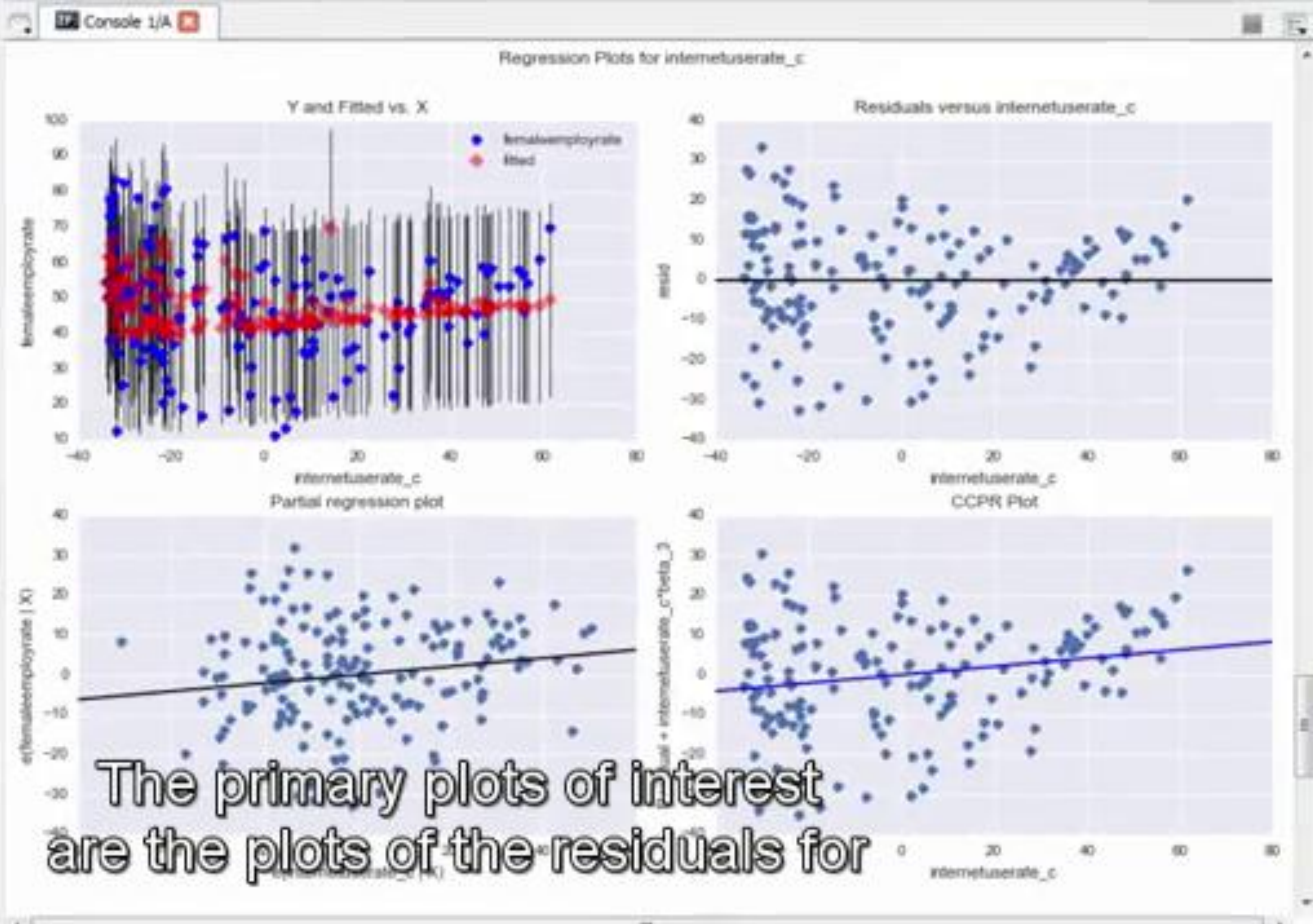



either direction, are a warning sign
that we may have some outliers.

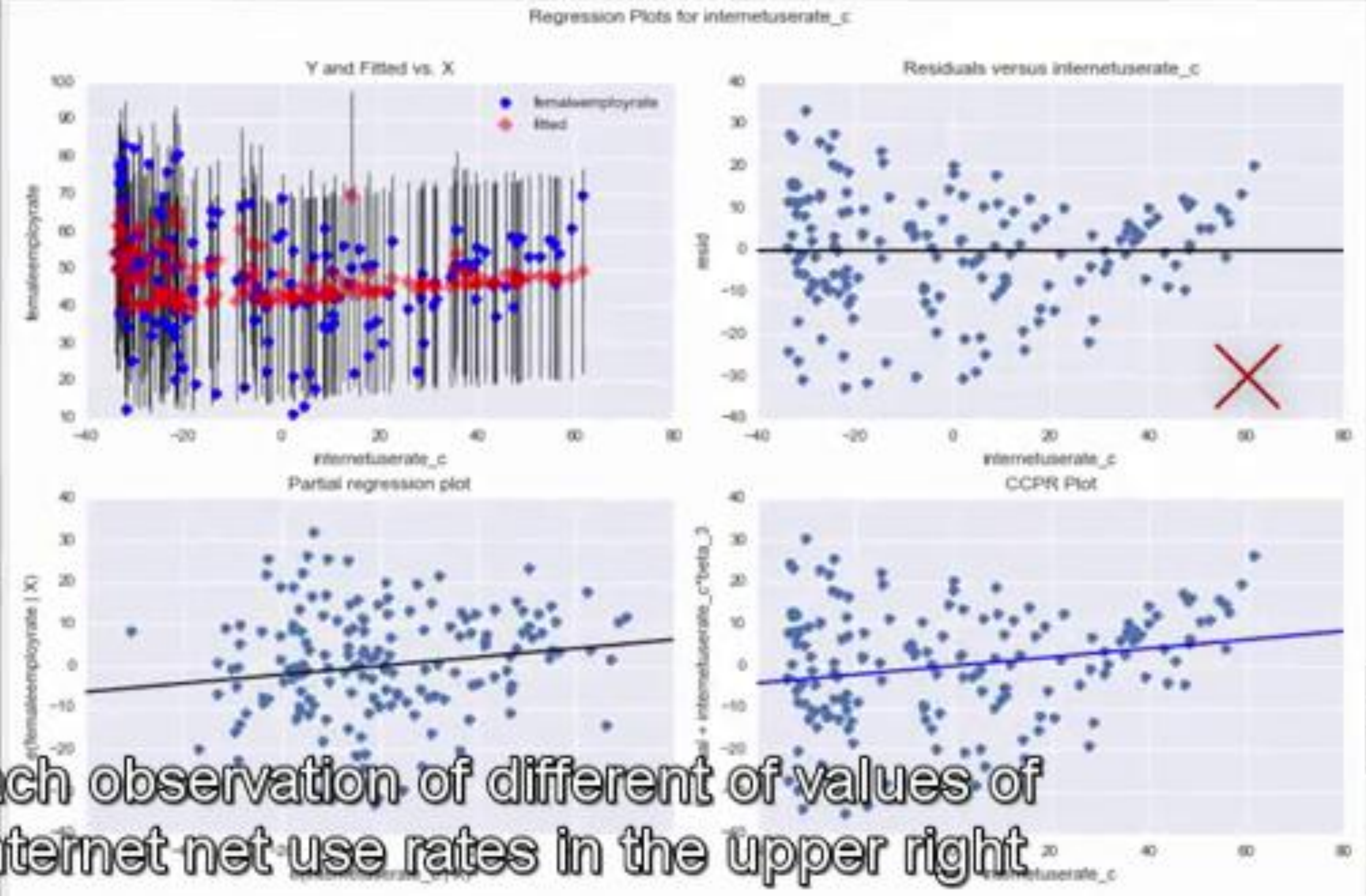
Console IPython console

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 148 Column: 1 Memory: 71

```python
print(fig1)


# simple plot of residuals
stdres=pandas.DataFrame(reg3.resid_pearson)
fig2=plt.plot(stdres, 'o', ls='None')
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')
print(fig2)


# additional regression diagnostic plots
fig3 = plt.figure(figsize(12,8))
fig3 = sm.graphics.plot_regress_exog(reg3,  "internetuserate_c", fig=fig3)
```

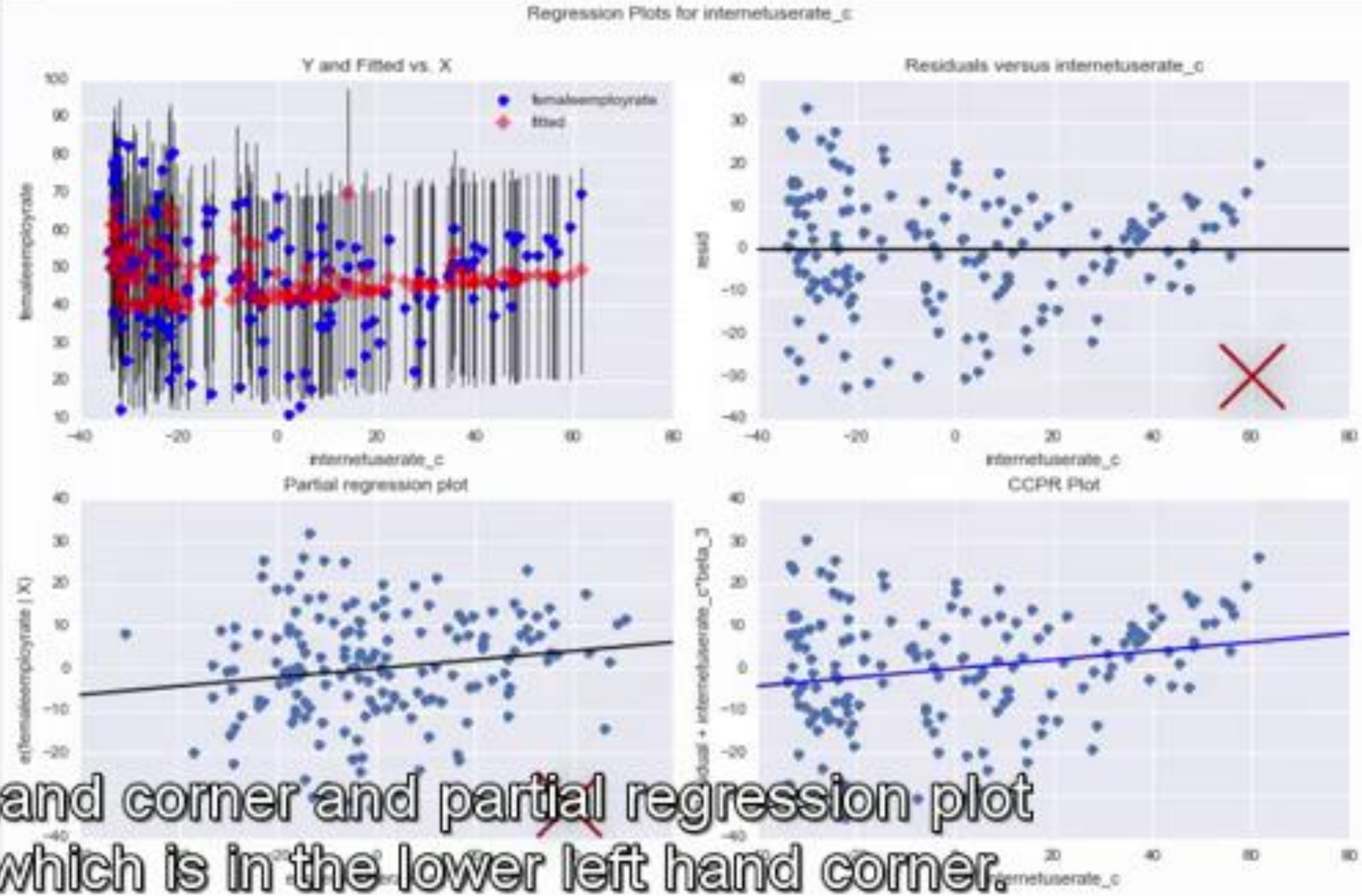



The primary plots of interest
are the plots of the residuals for

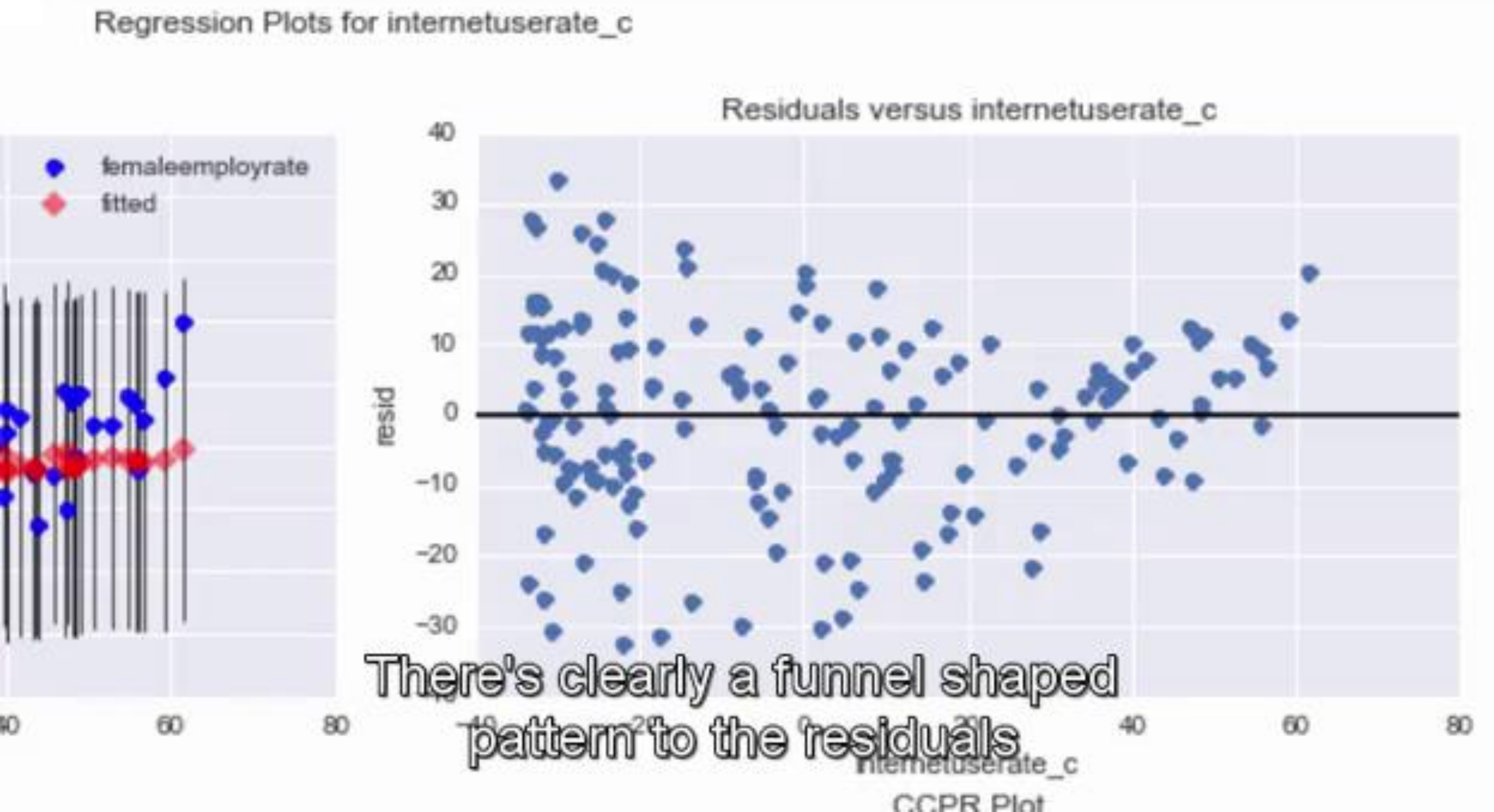



each observation of different of values of
Internet net use rates in the upper right

Console 1/A

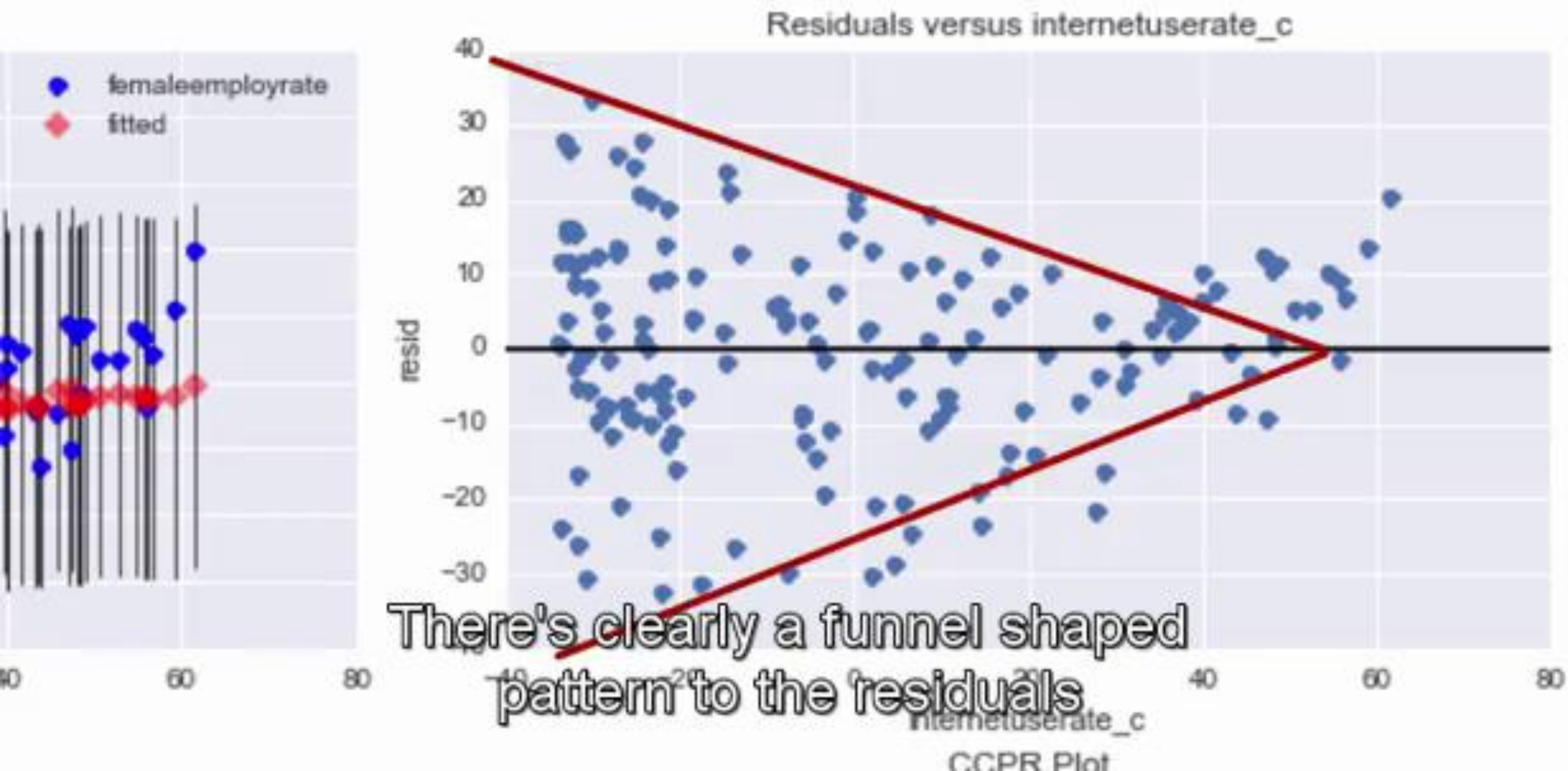



hand corner and partial regression plot
which is in the lower left hand corner.



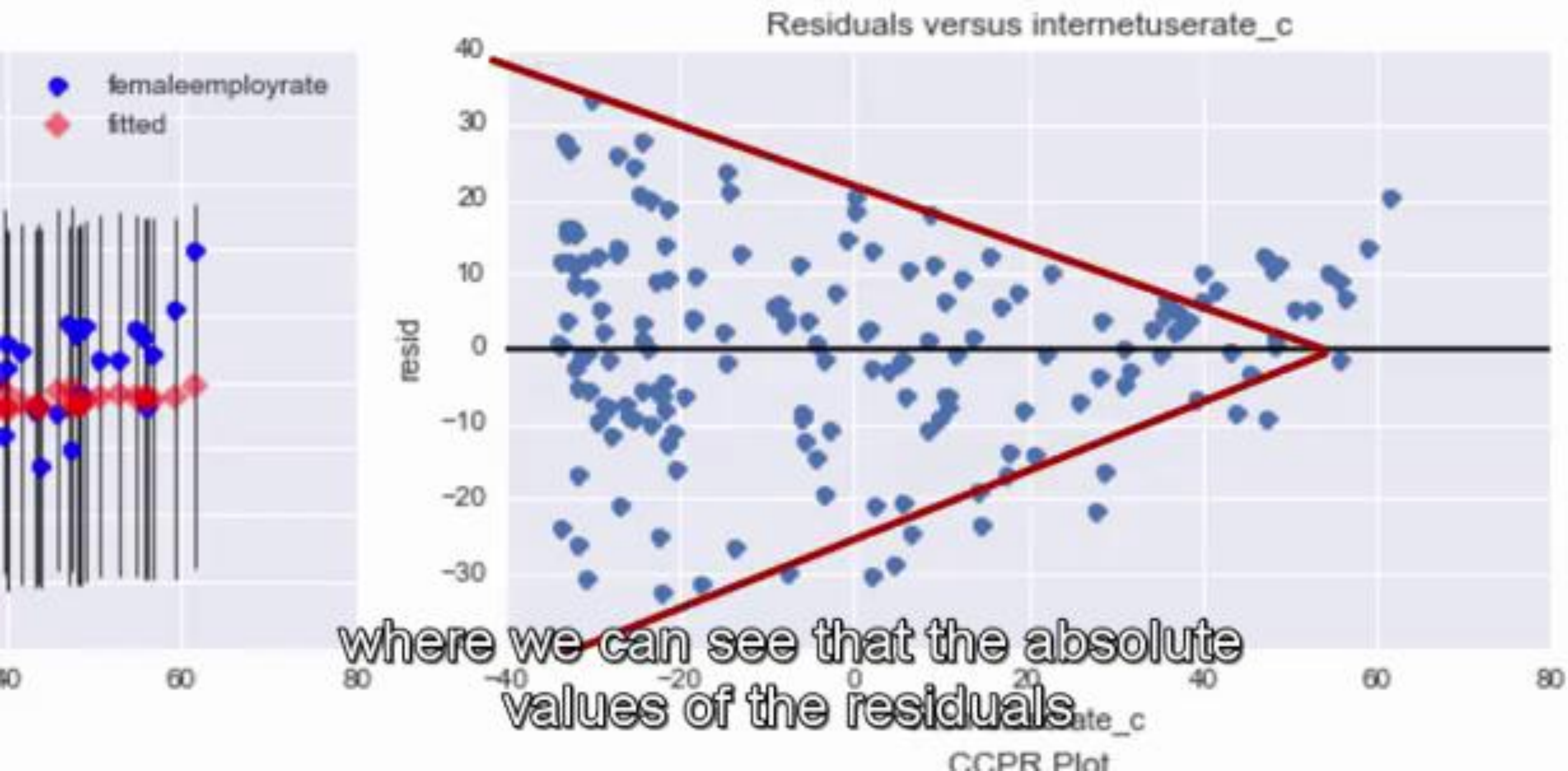

There's clearly a funnel shaped pattern to the residuals

Regression Plots for internetuserate_c

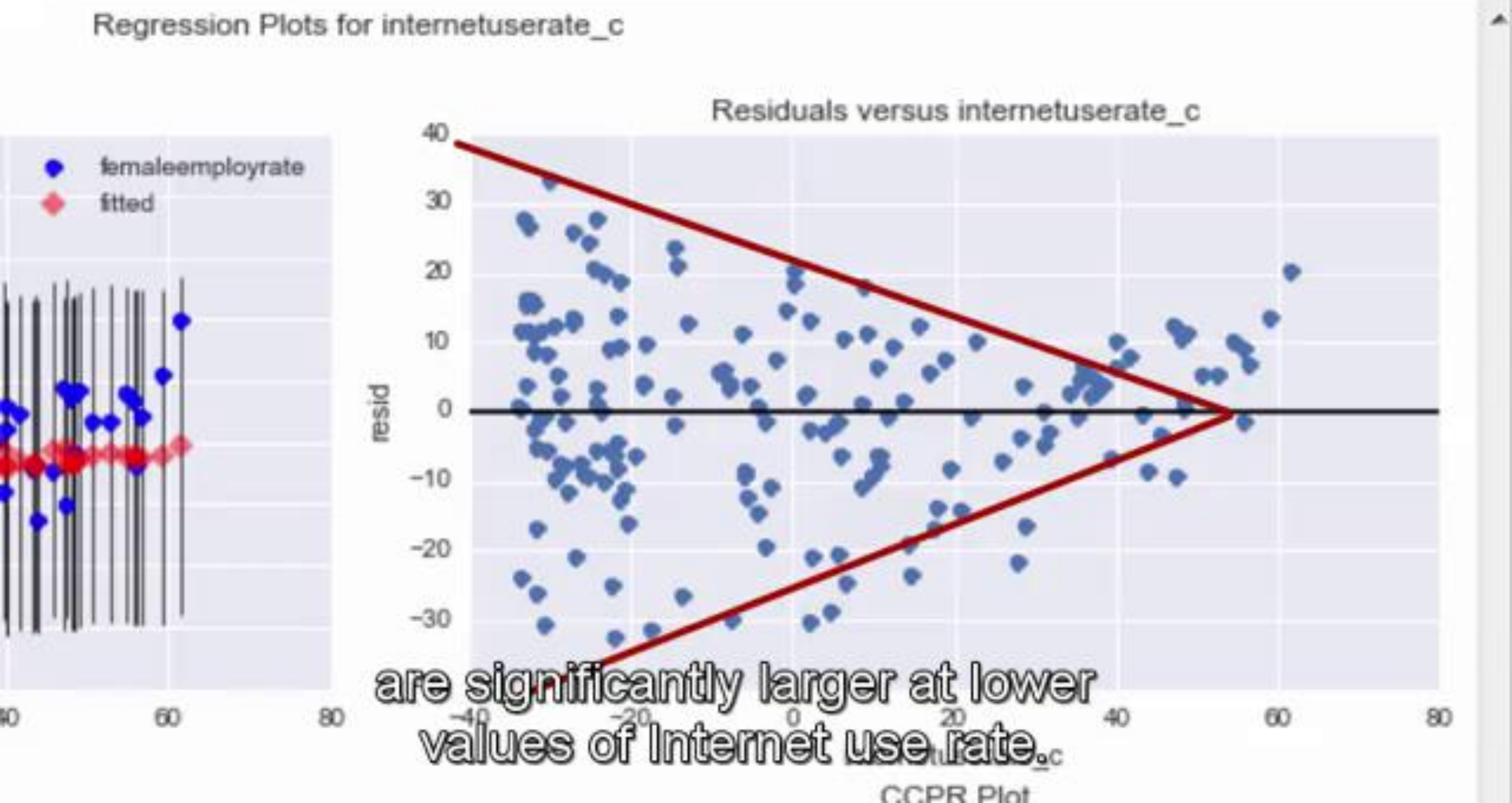



There's clearly a funnel shaped pattern to the residuals

Regression Plots for internetuserate_c

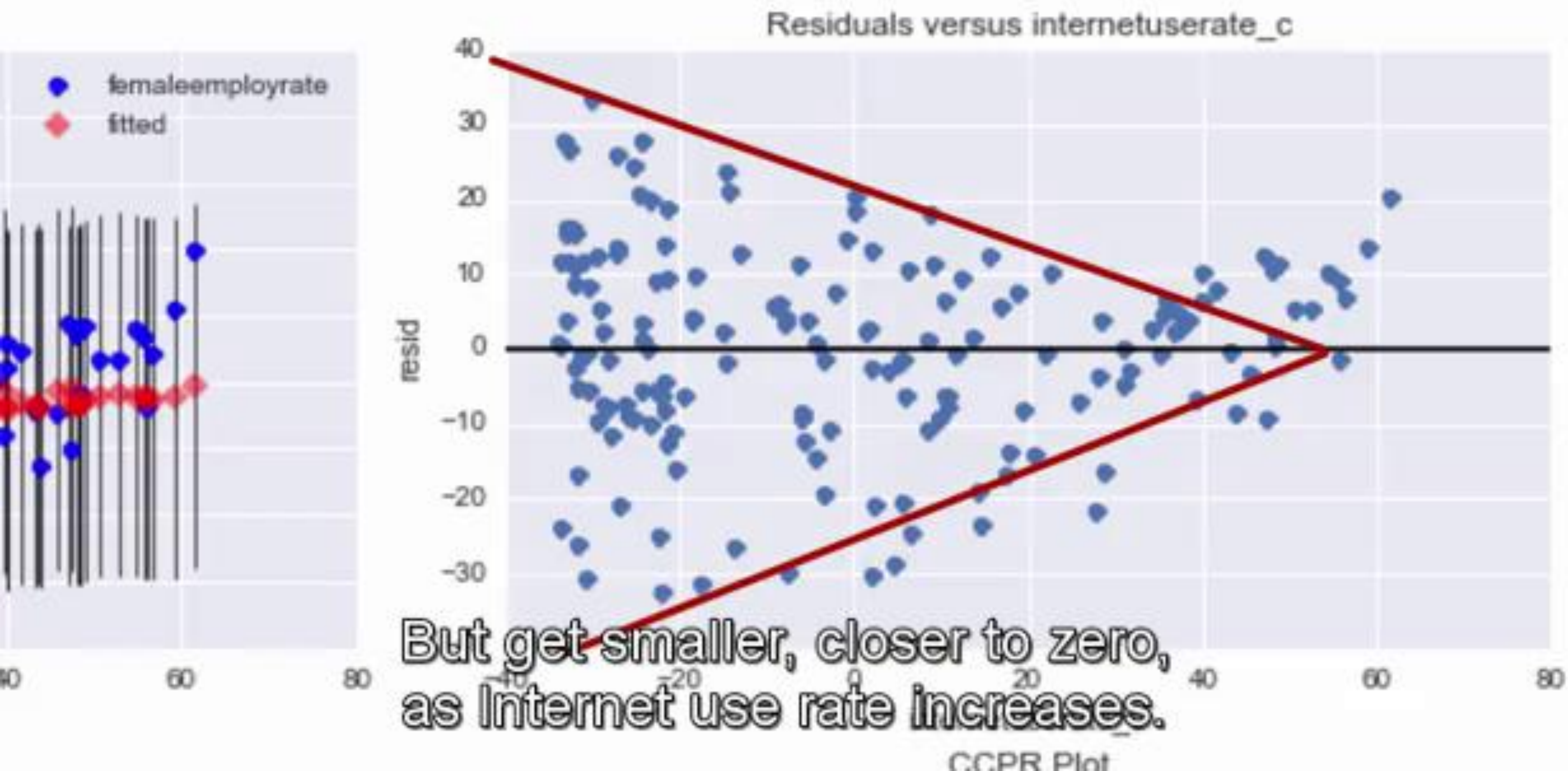



where we can see that the absolute values of the residuals



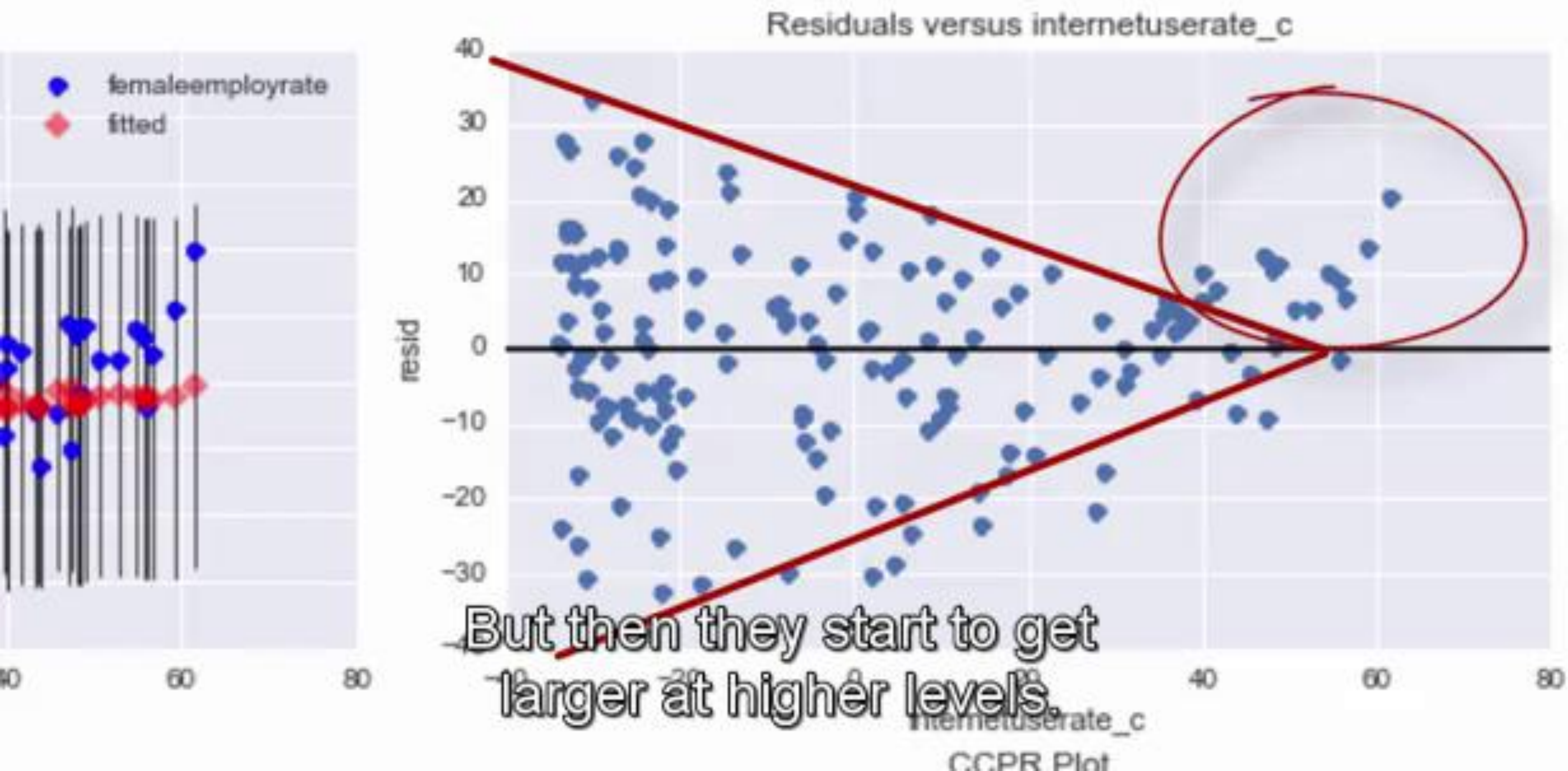

are significantly larger at lower values of Internet use rate.

Regression Plots for internetuserate_c

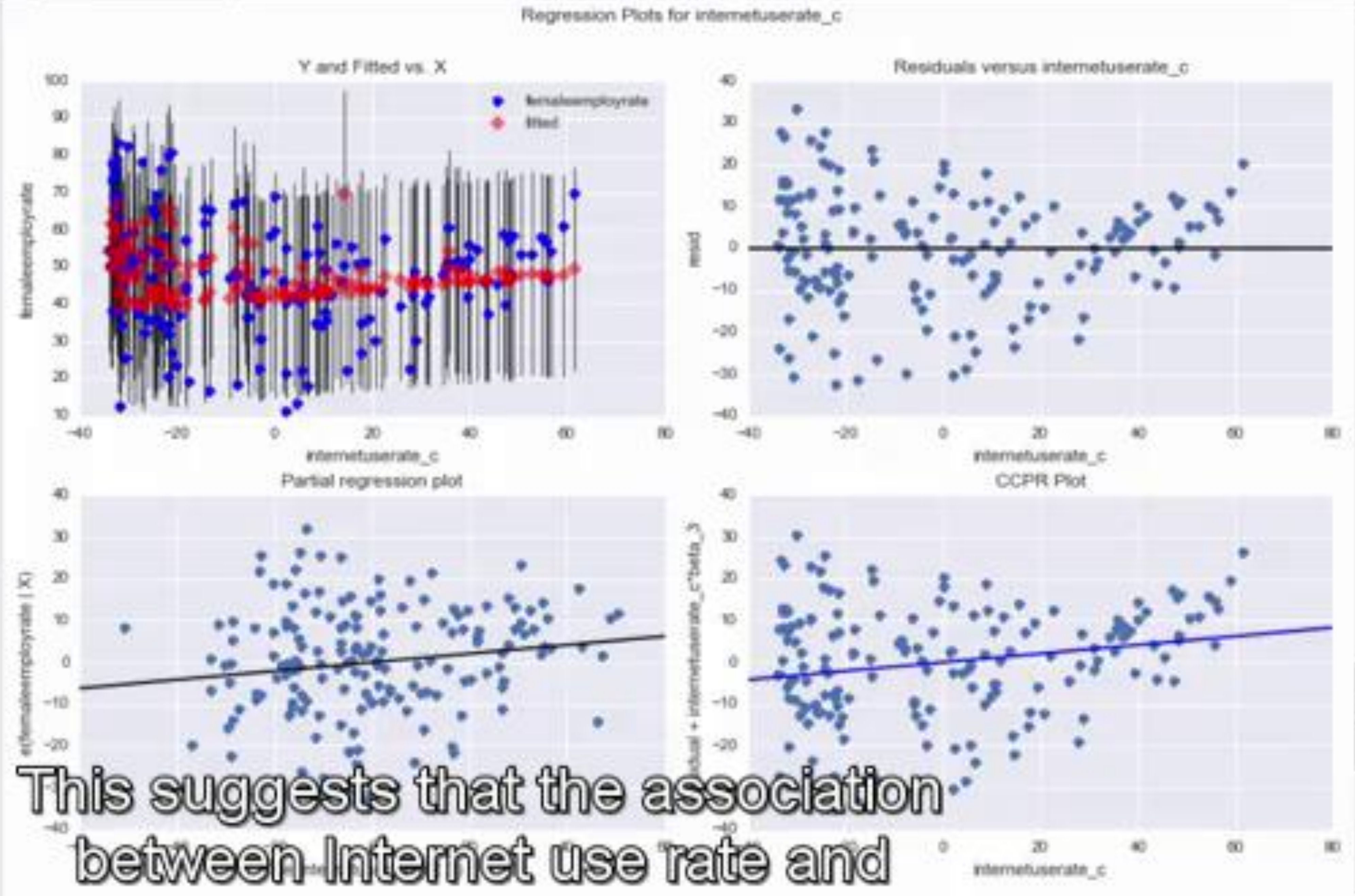



But get smaller, closer to zero,
as Internet use rate increases.

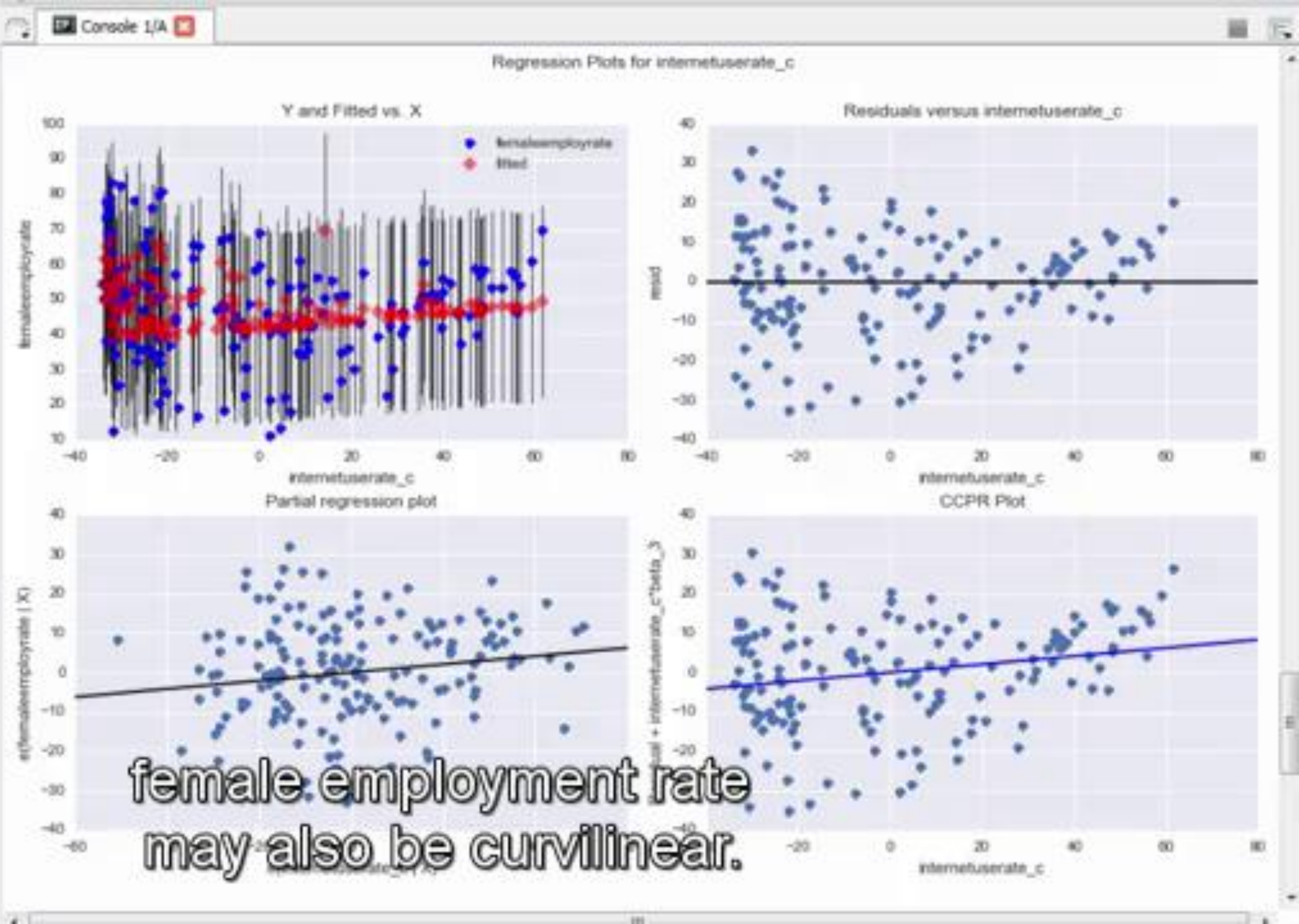
Regression Plots for internetuserate_c

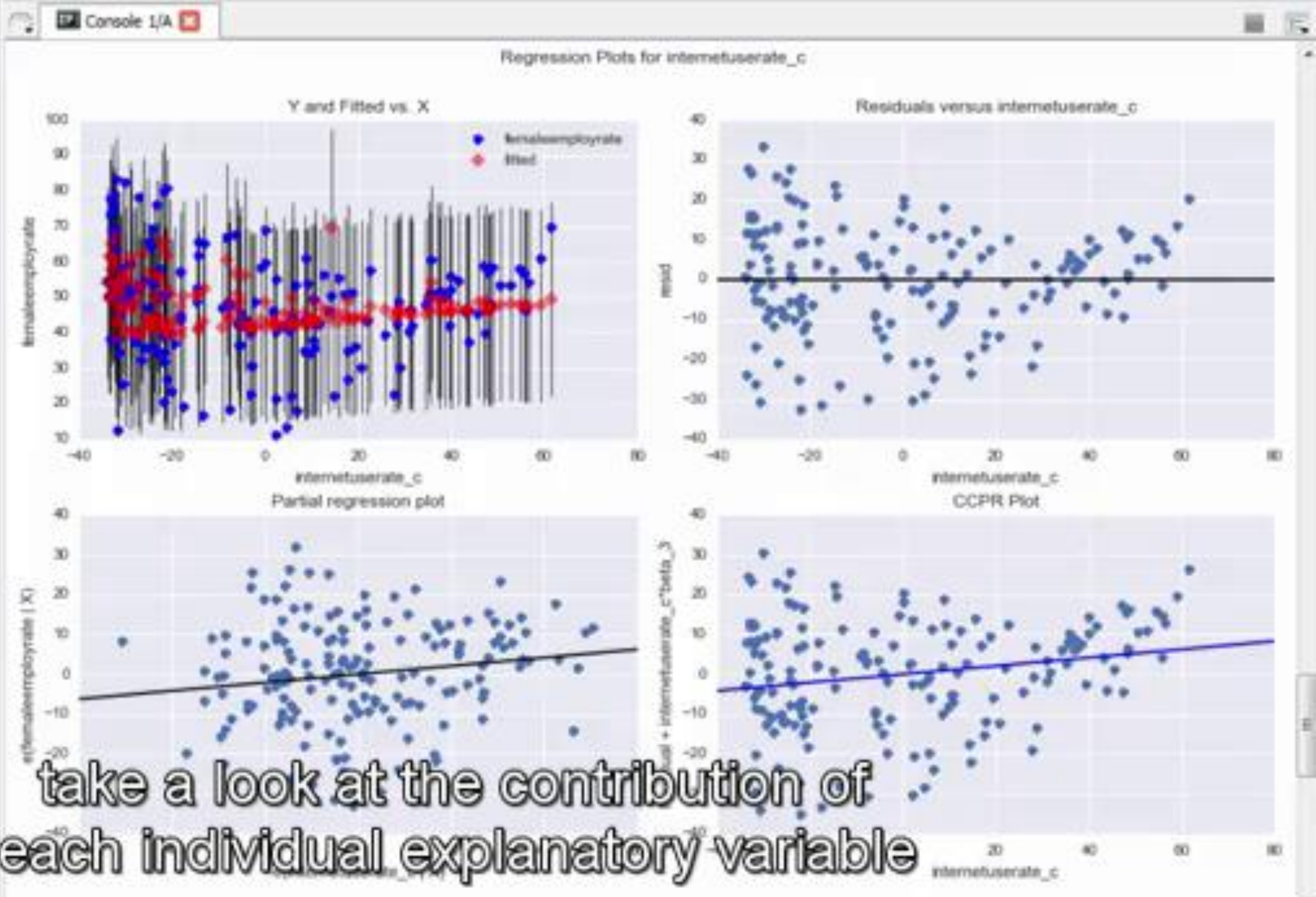



But then they start to get larger at higher levels.

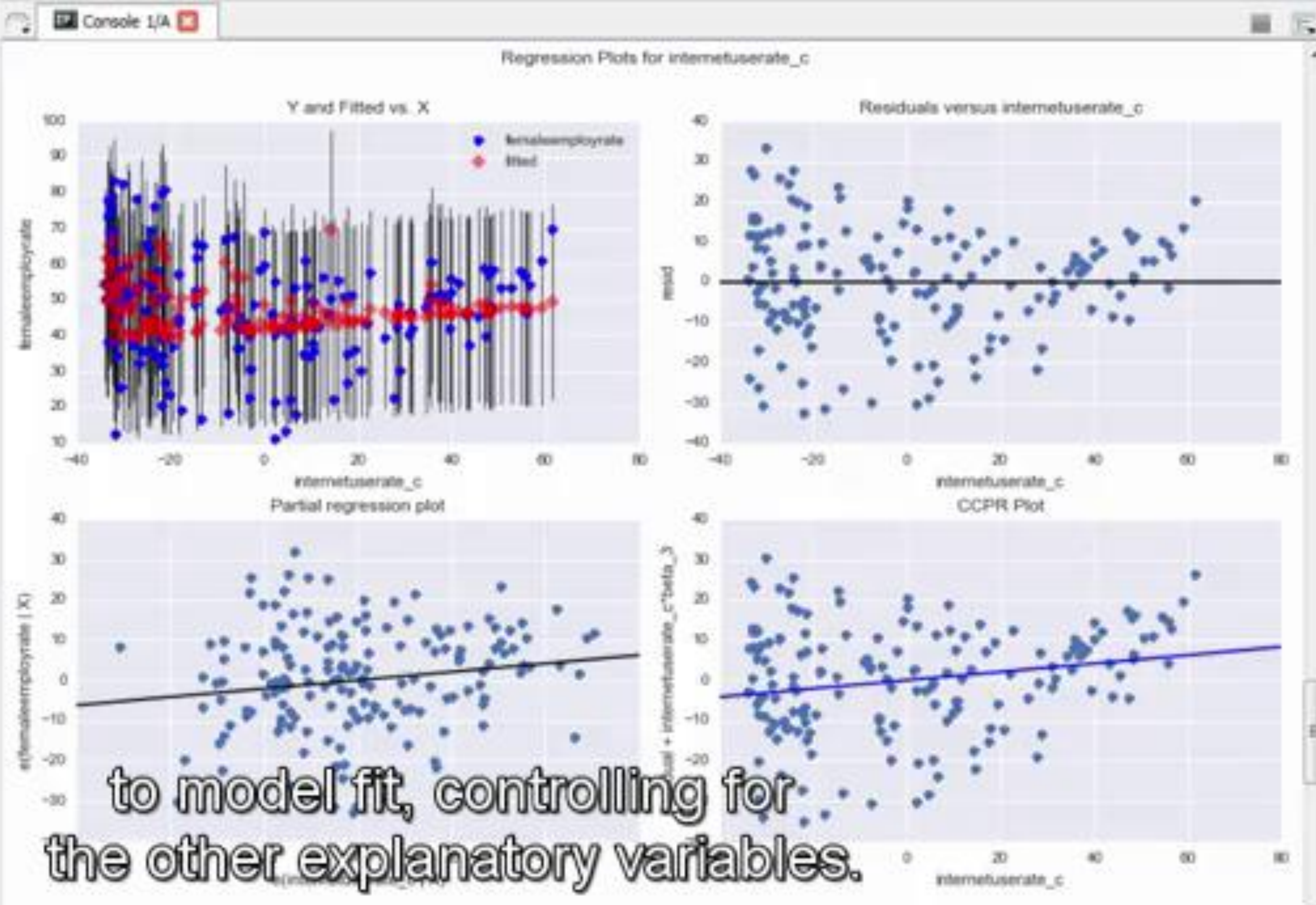



This suggests that the association
between Internet use rate and

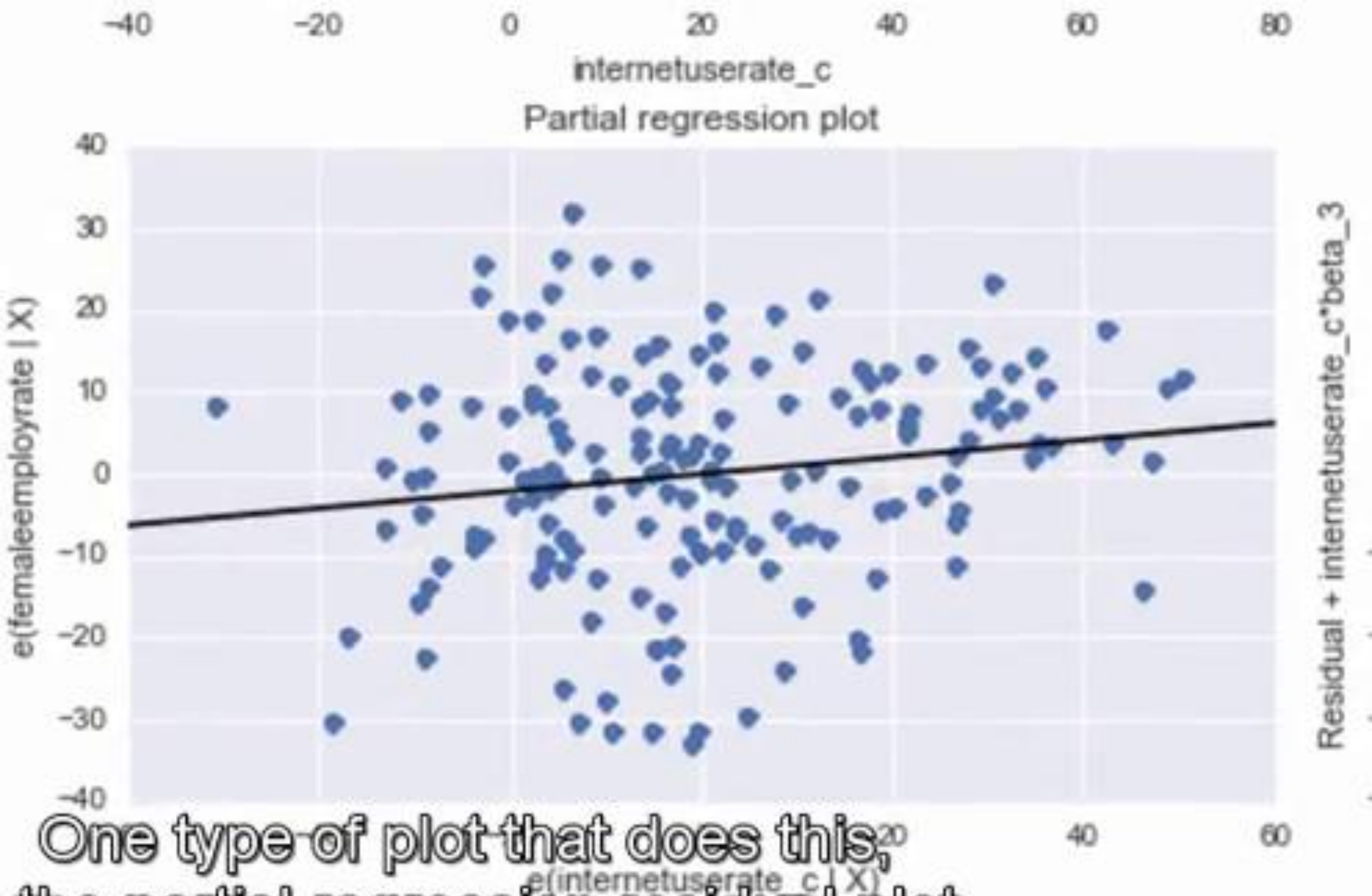

Console 1/A
Regression Plots for internetuserate_c
Y and Fitted vs. X
femaleemployrate
fitted
internetuserate_c
Residuals versus internetuserate_c
resid
internetuserate_c
Partial regression plot
CCPR Plot
internetuserate_c
female employment rate
may also be curvilinear.



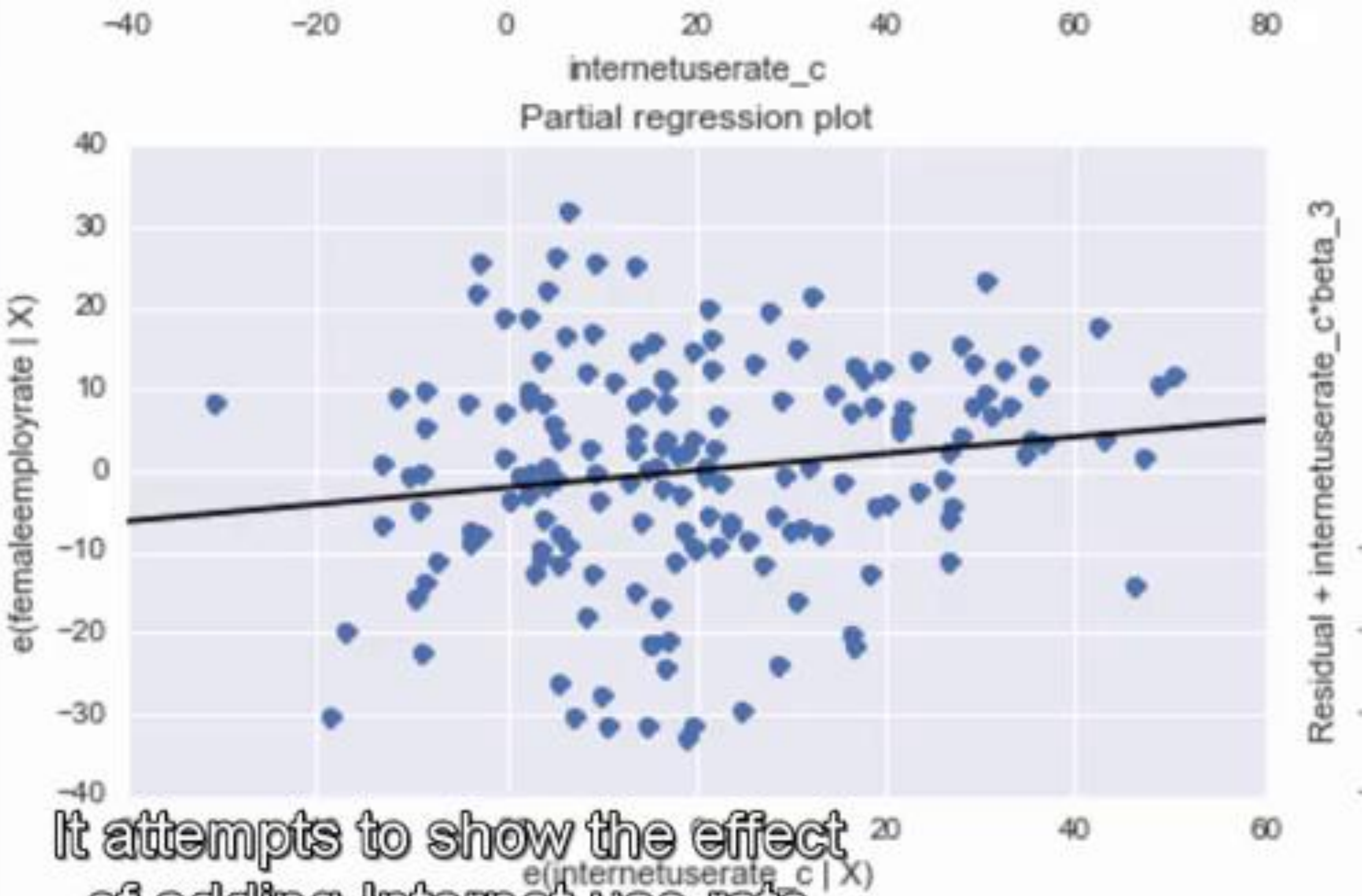

take a look at the contribution of
each individual explanatory variable



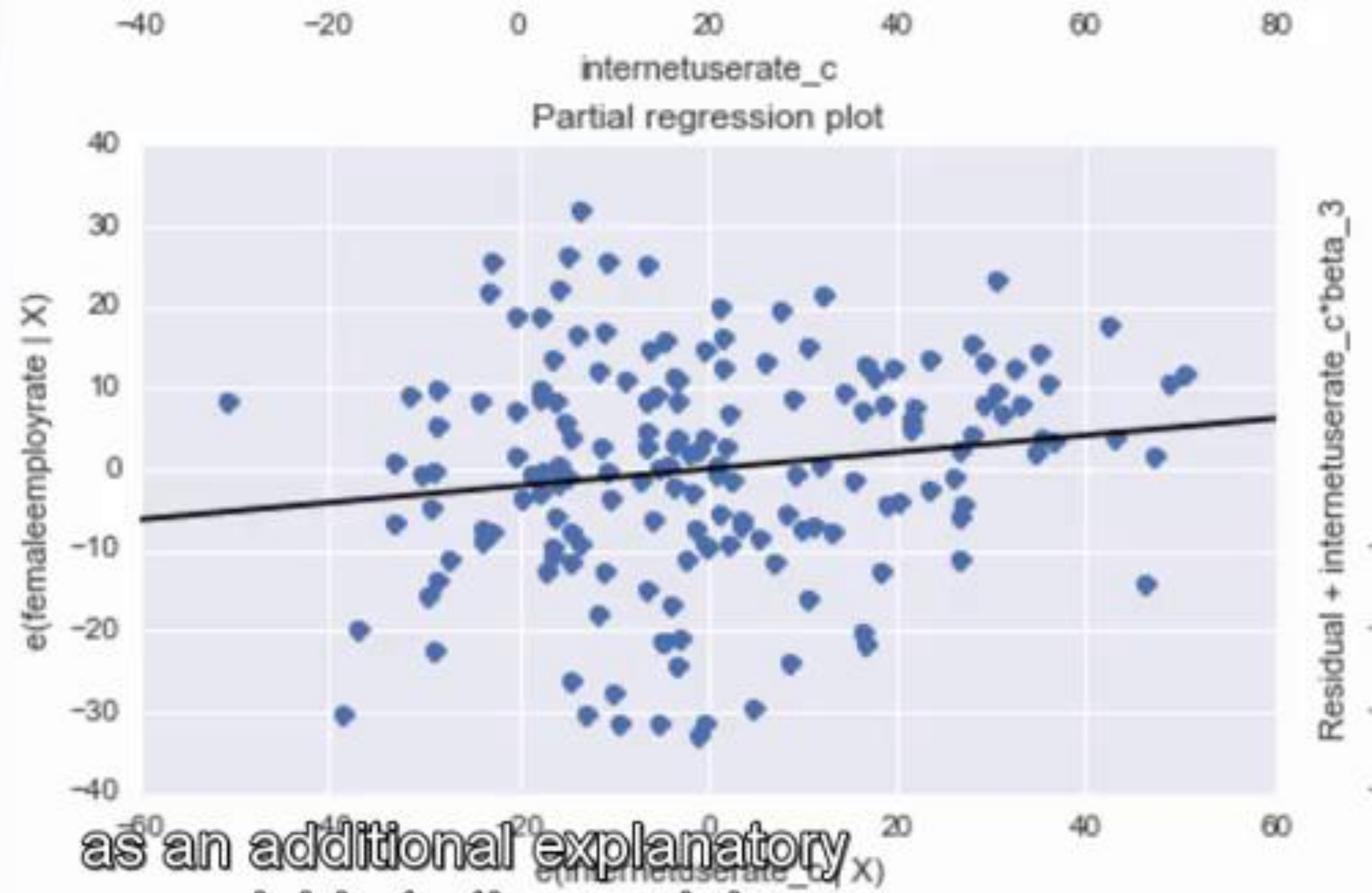

to model fit, controlling for
the other explanatory variables.



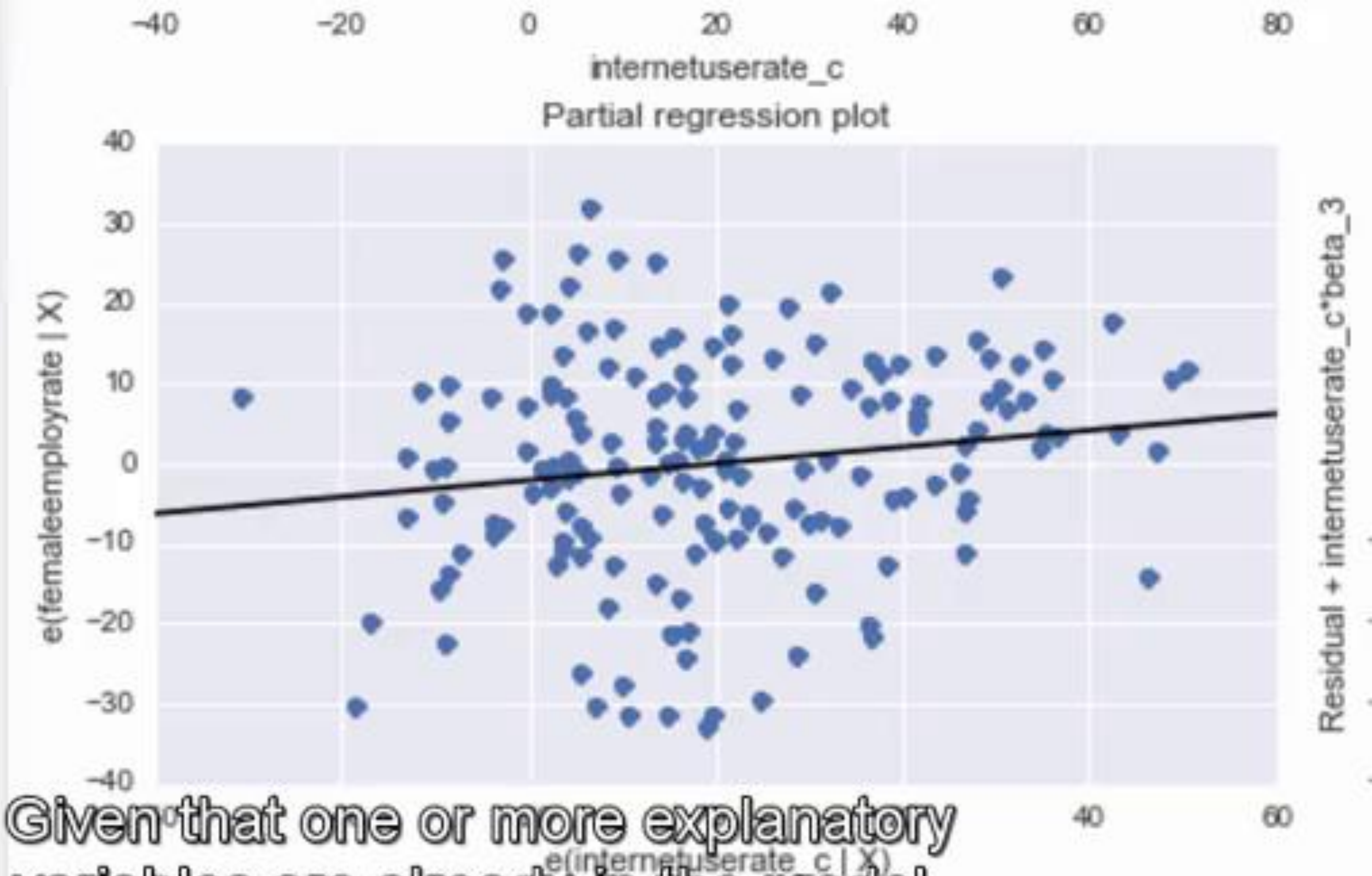

One type of plot that does this,
is the partial regression residual plot.



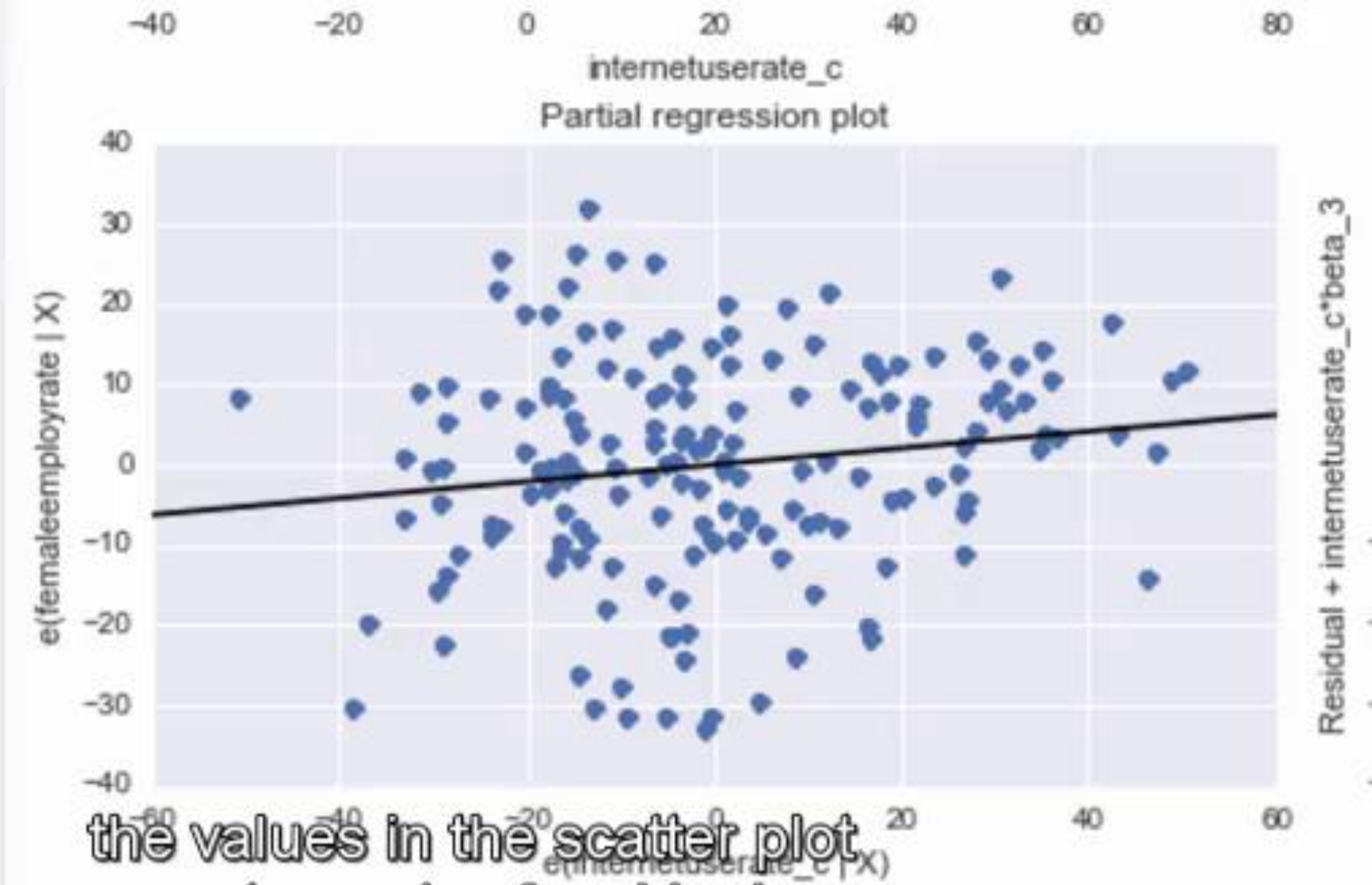

It attempts to show the effect of adding Internet use rate



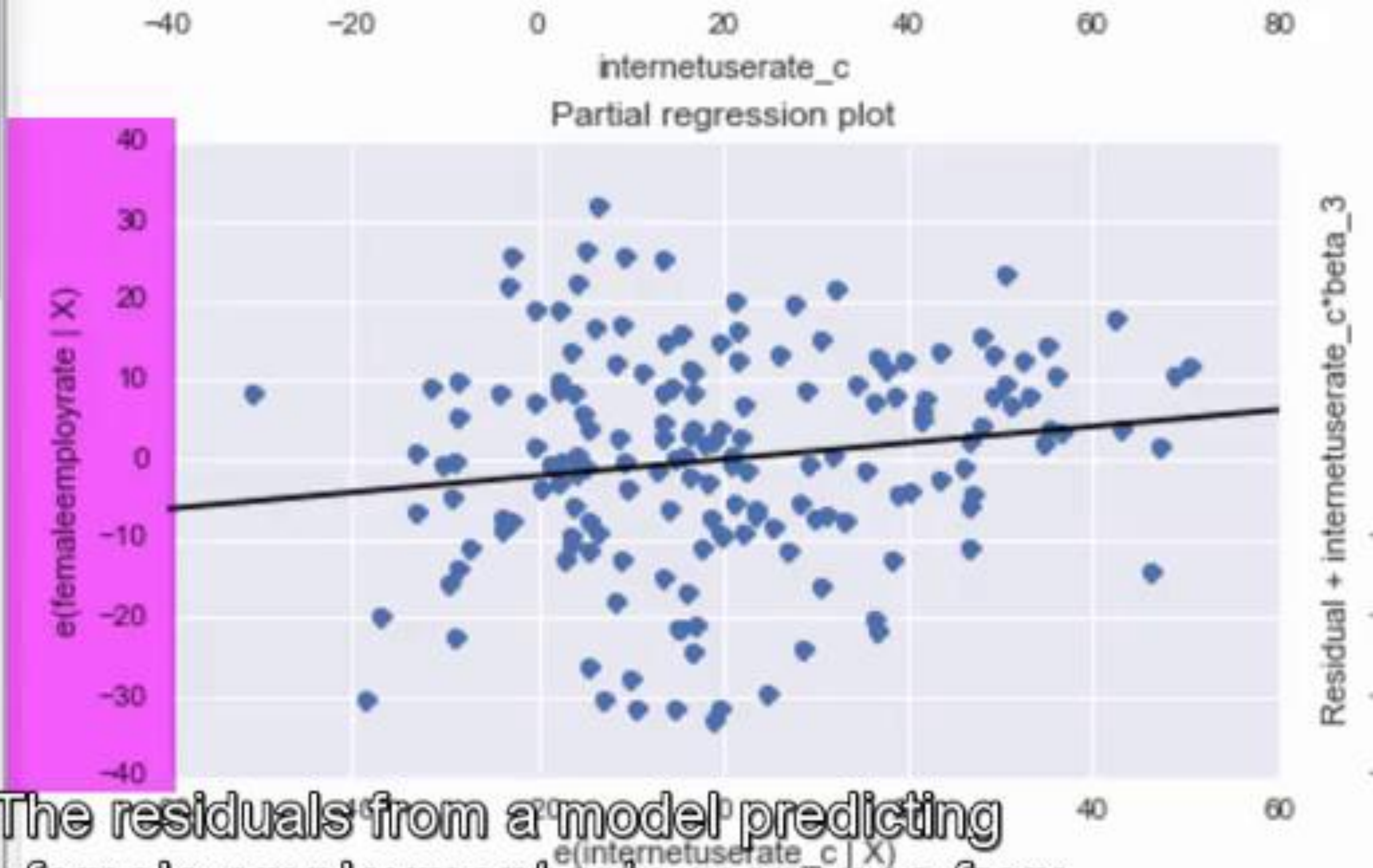

as an additional explanatory variable to the model.



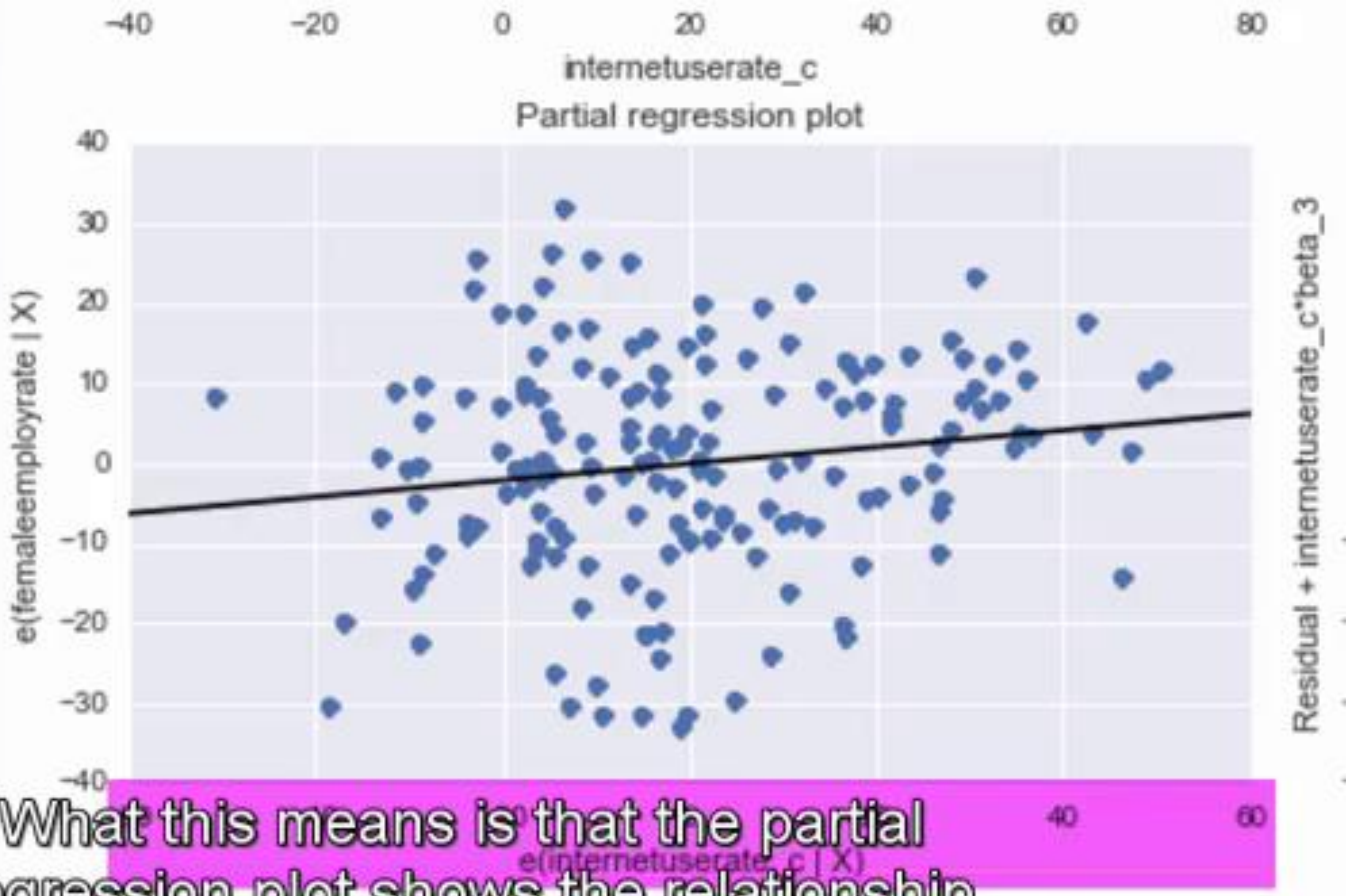

Given that one or more explanatory variables are already in the model.



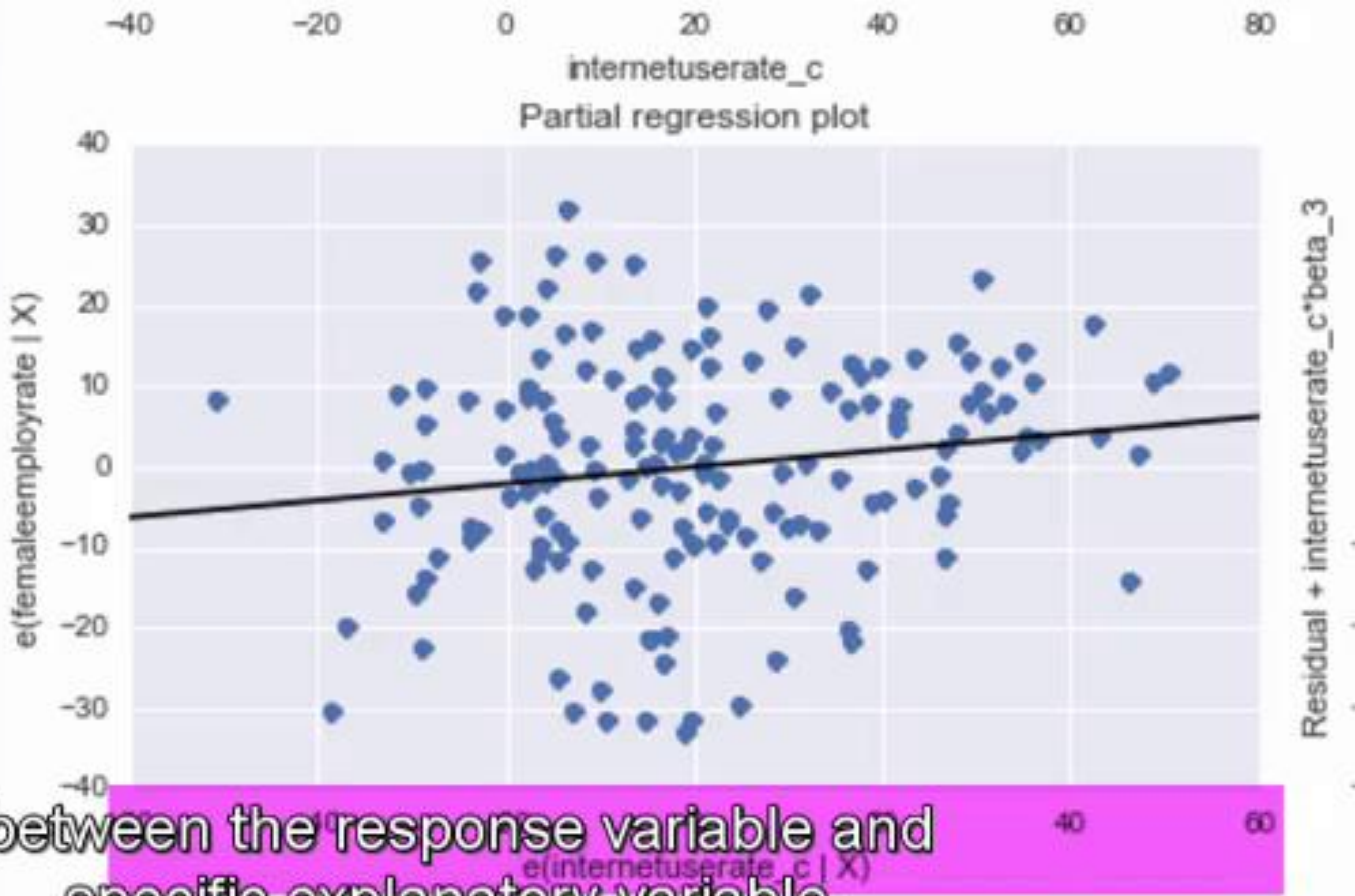

the values in the scatter plot
are two sets of residuals.



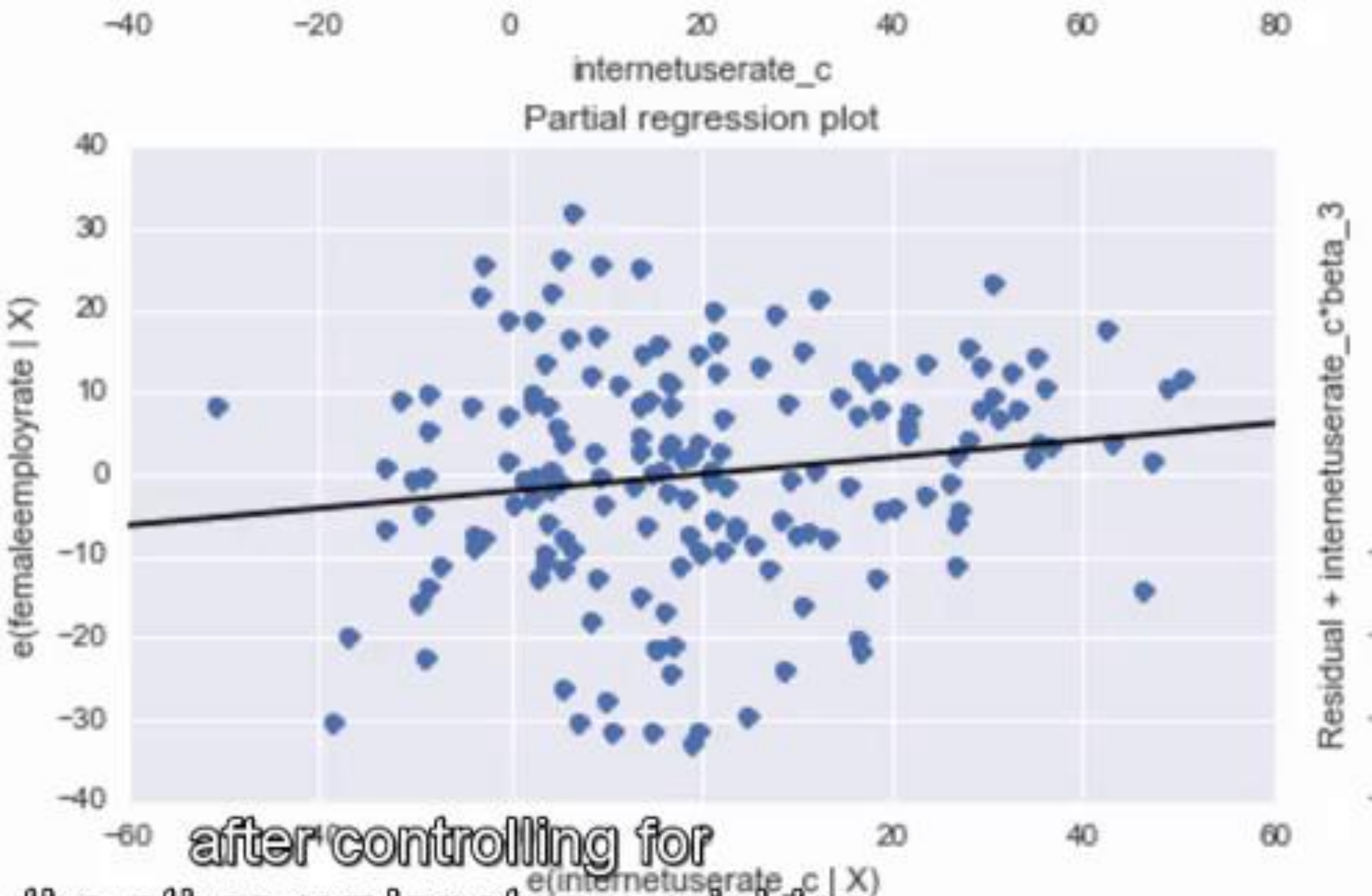

The residuals from a model predicting
the female employment rate response from



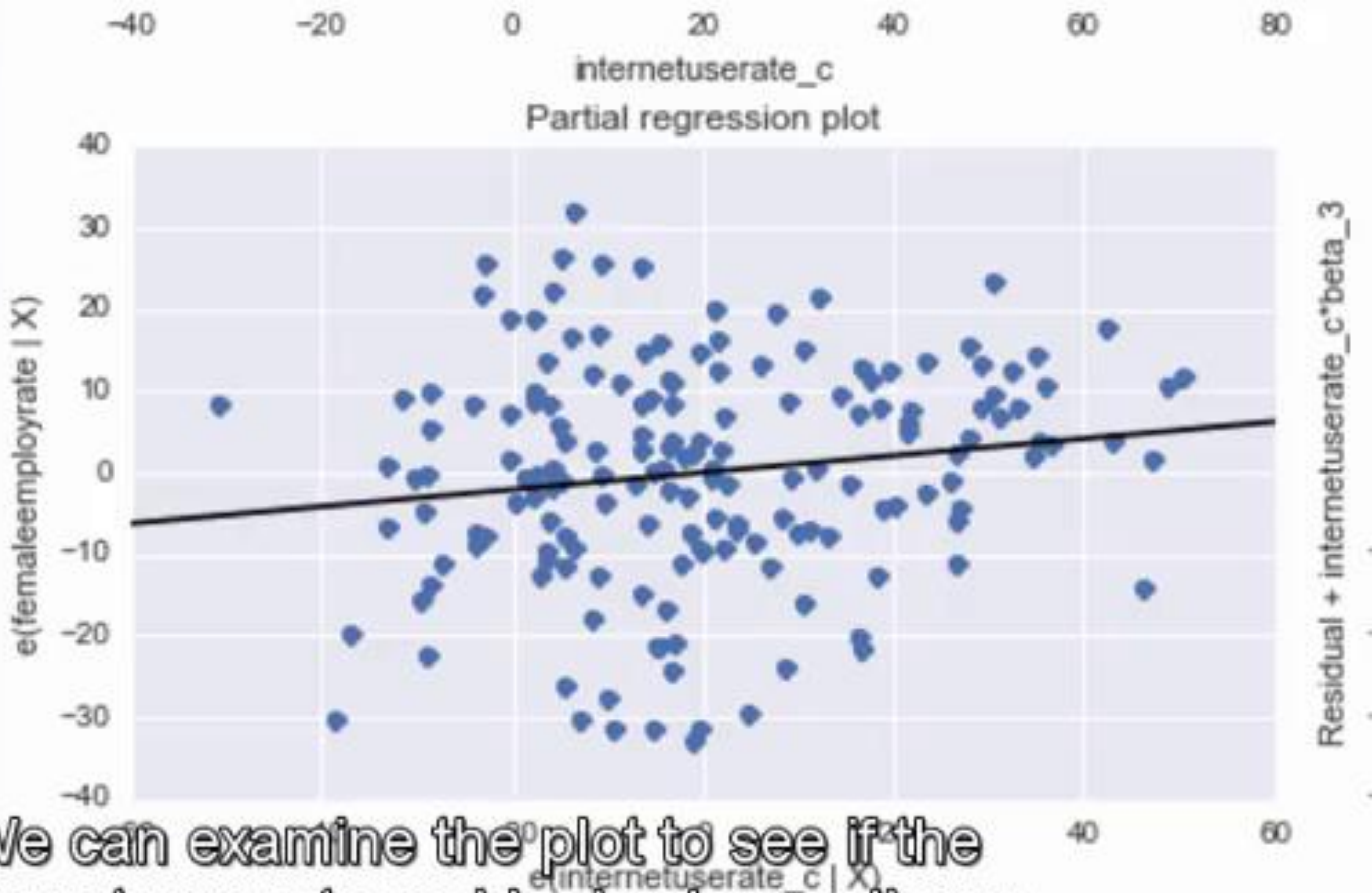

What this means is that the partial regression plot shows the relationship



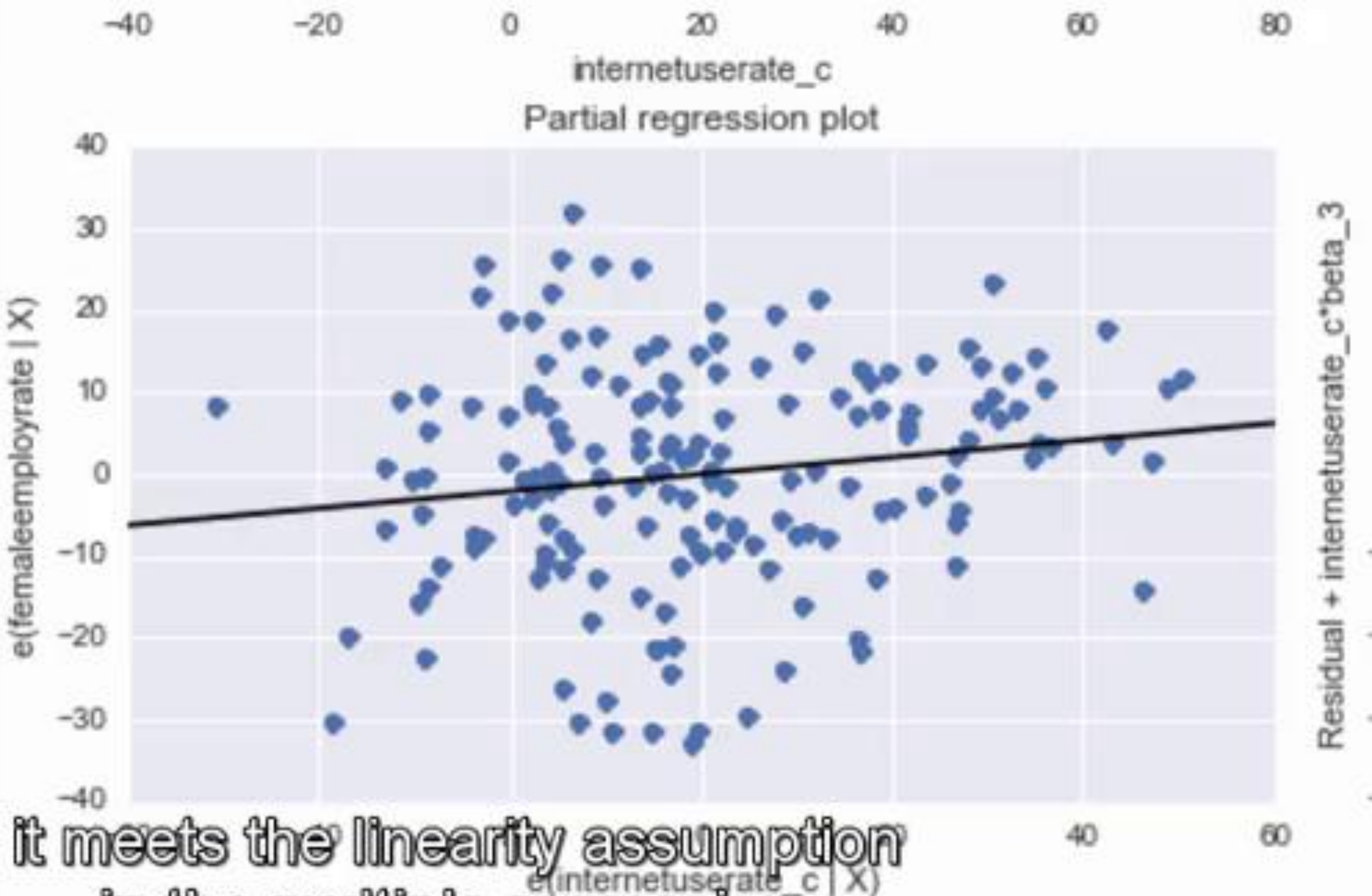

between the response variable and
specific explanatory variable,



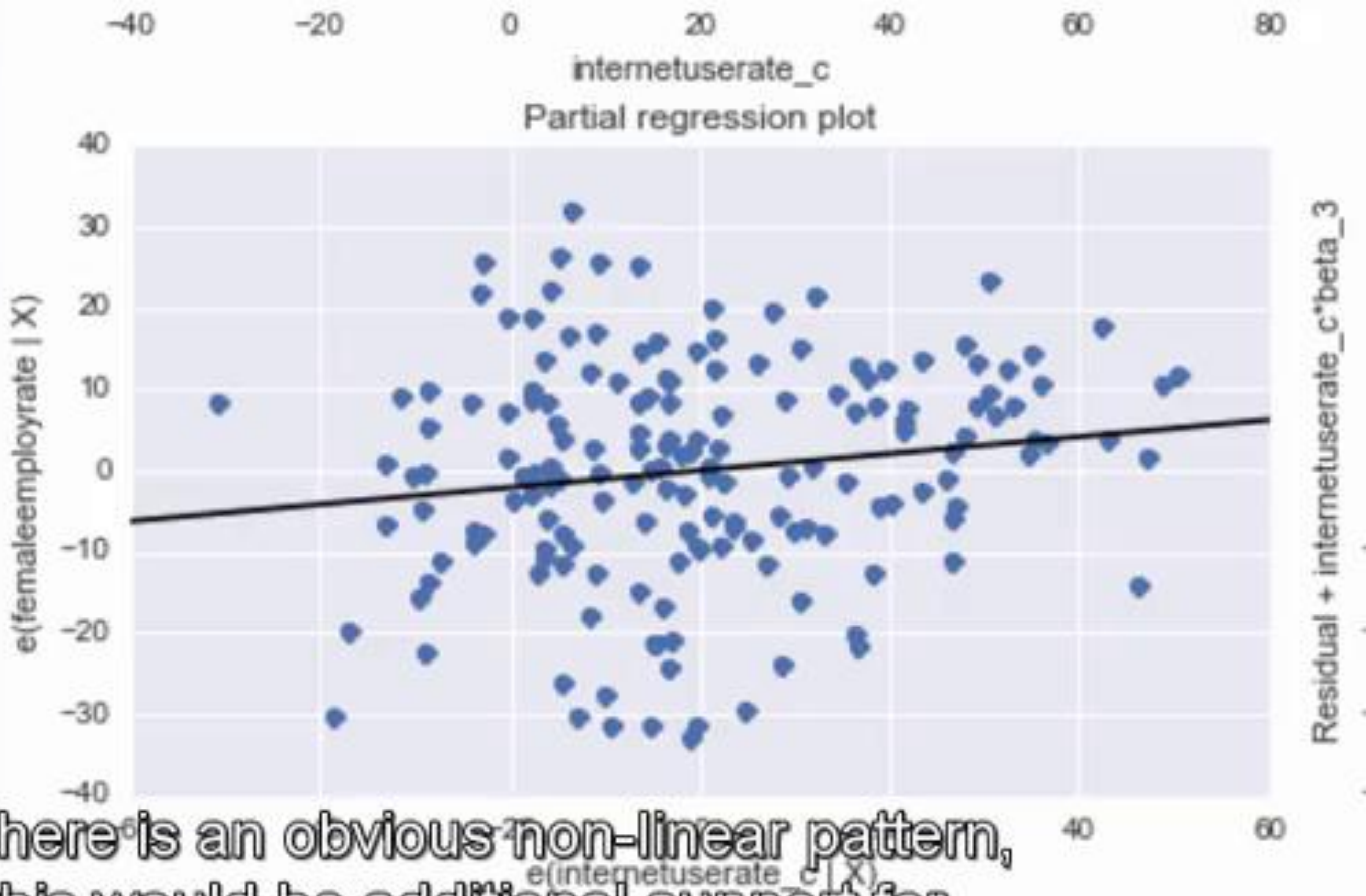

after controlling for
the other explanatory variables.



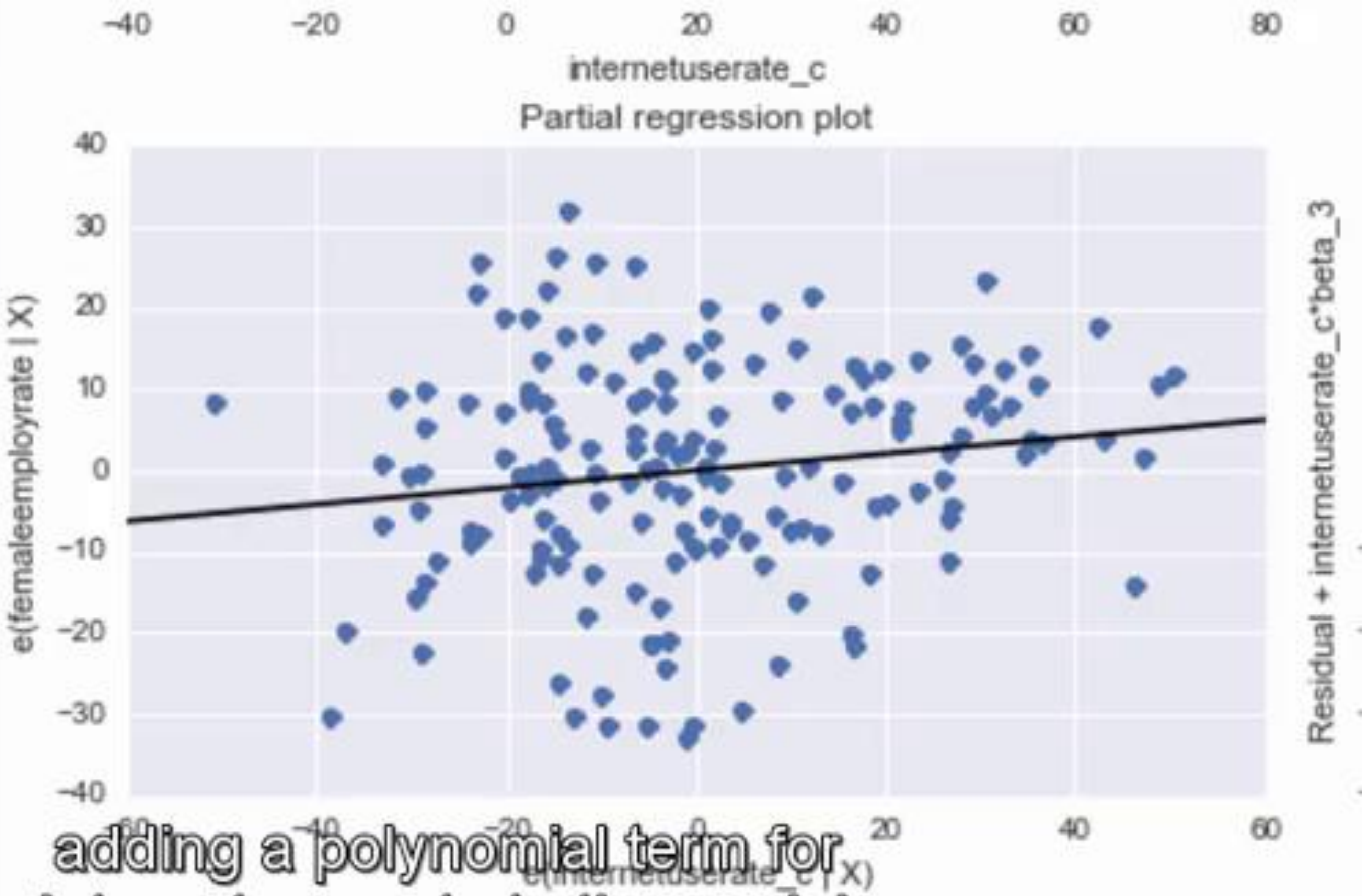

We can examine the plot to see if the
Internet use rate residuals show a linear,



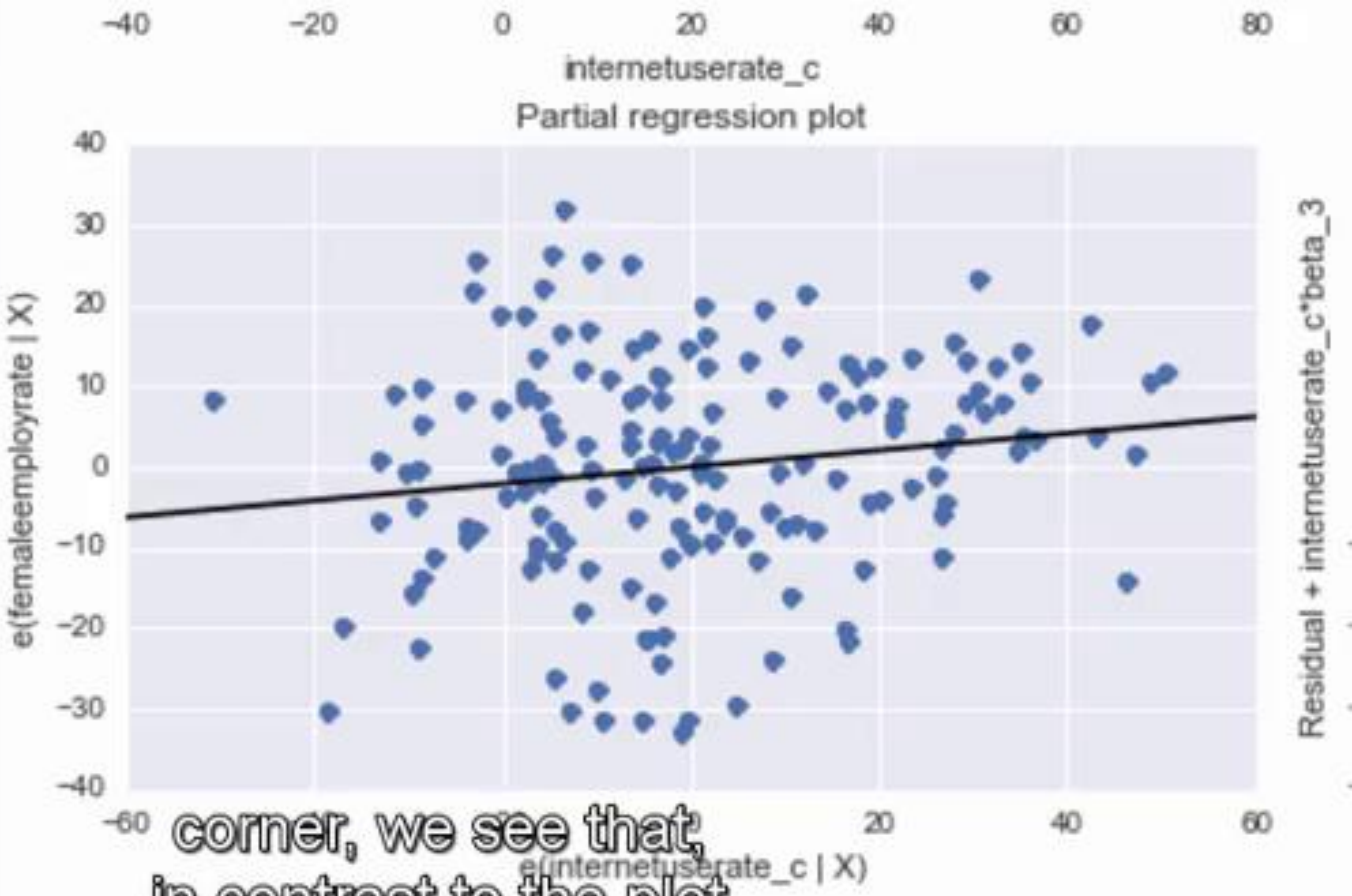

it meets the linearity assumption
in the multiple regression.



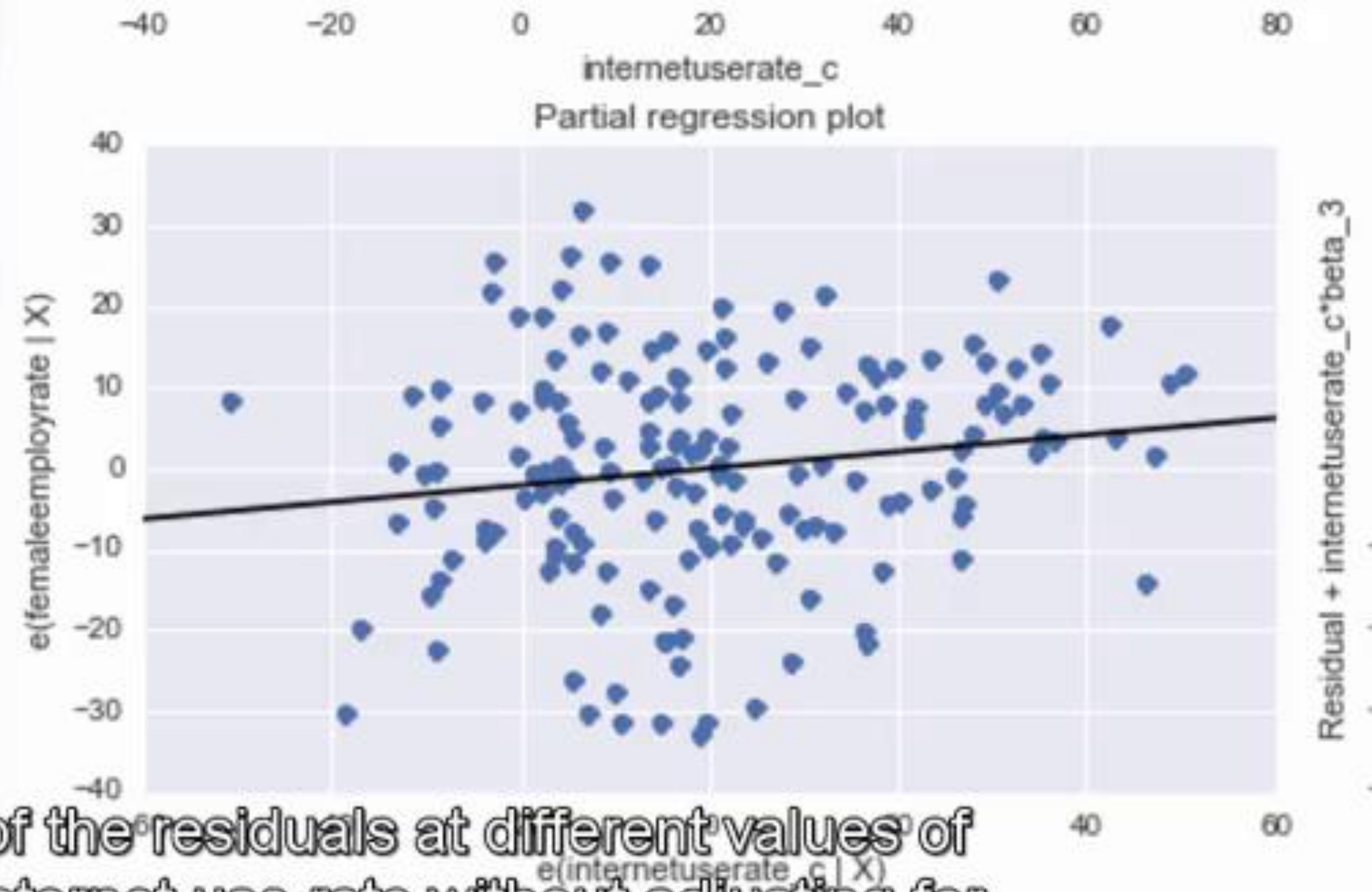

If there is an obvious non-linear pattern,
this would be additional support for



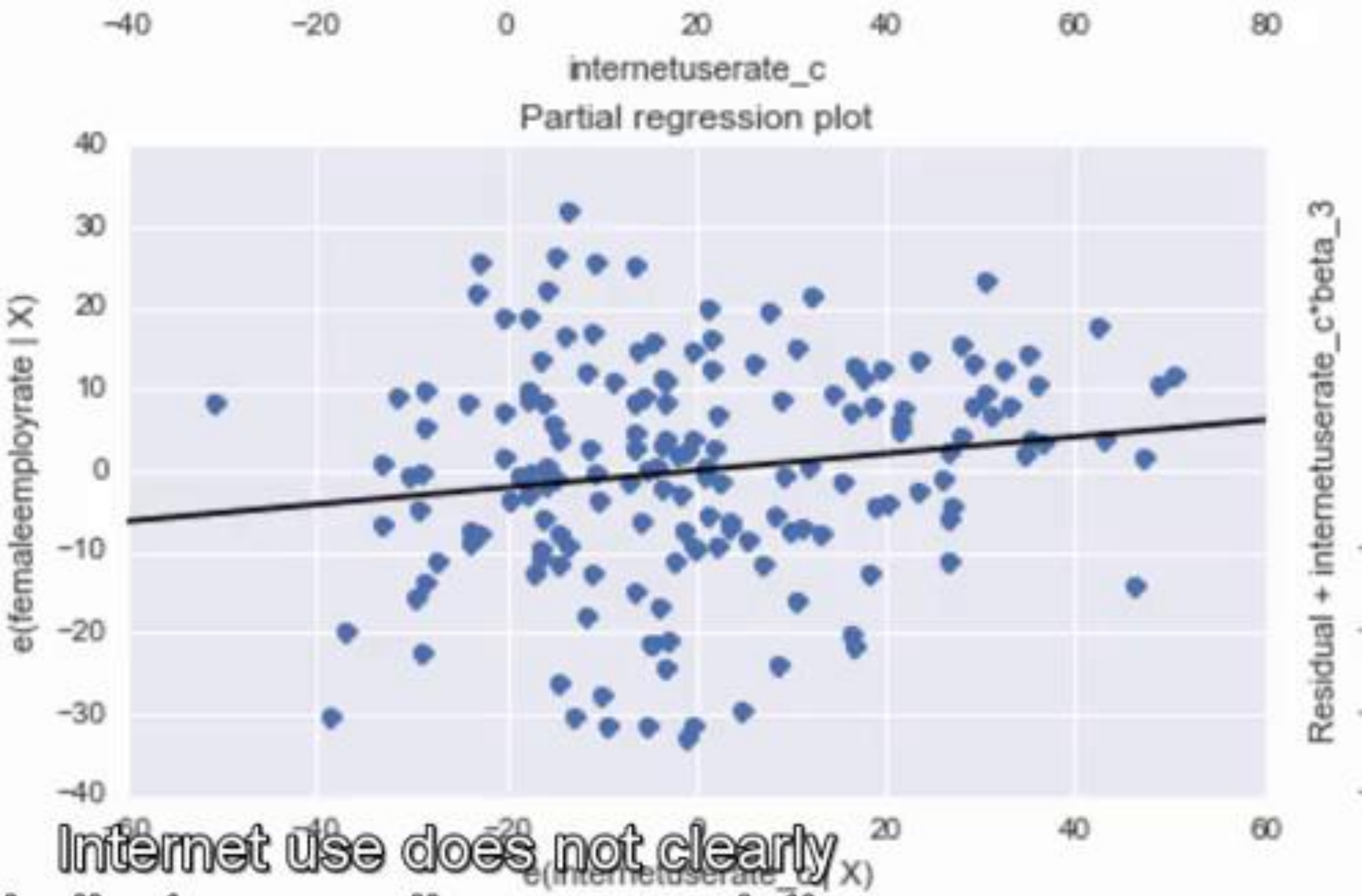

adding a polynomial term for
Internet use rate to the model.



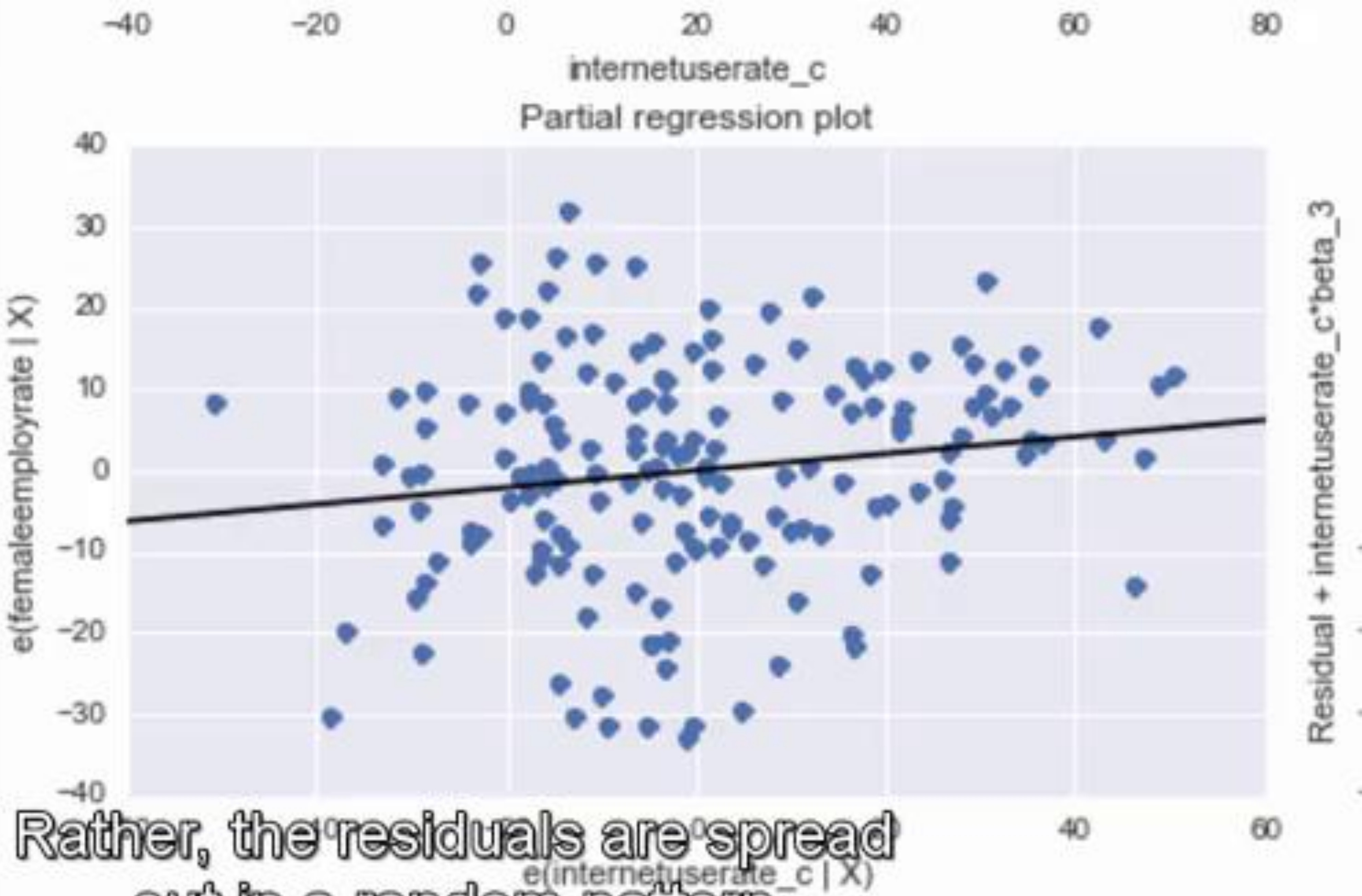

corner, we see that,
in contrast to the plot



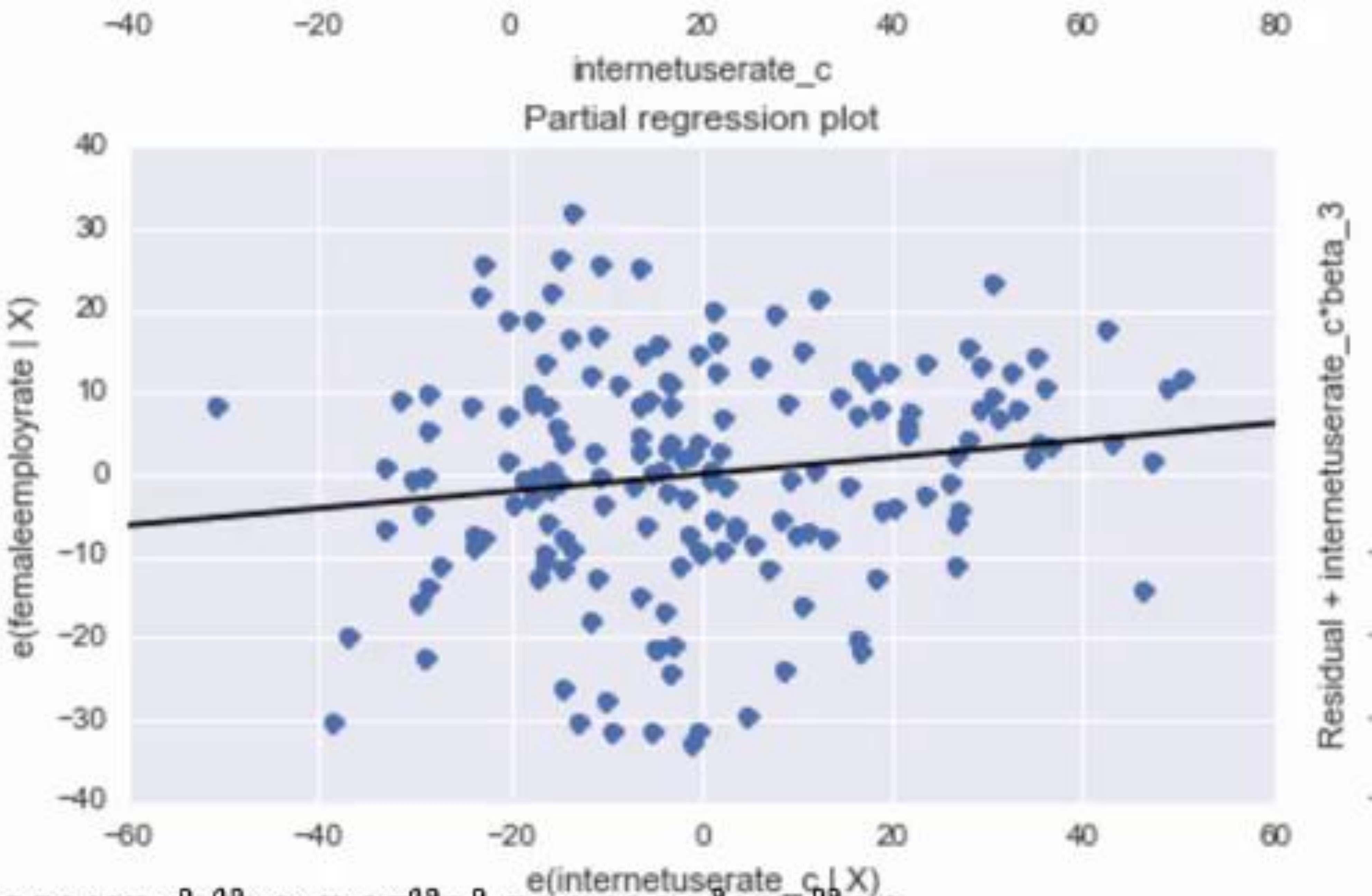

of the residuals at different values of
Internet use rate without adjusting for



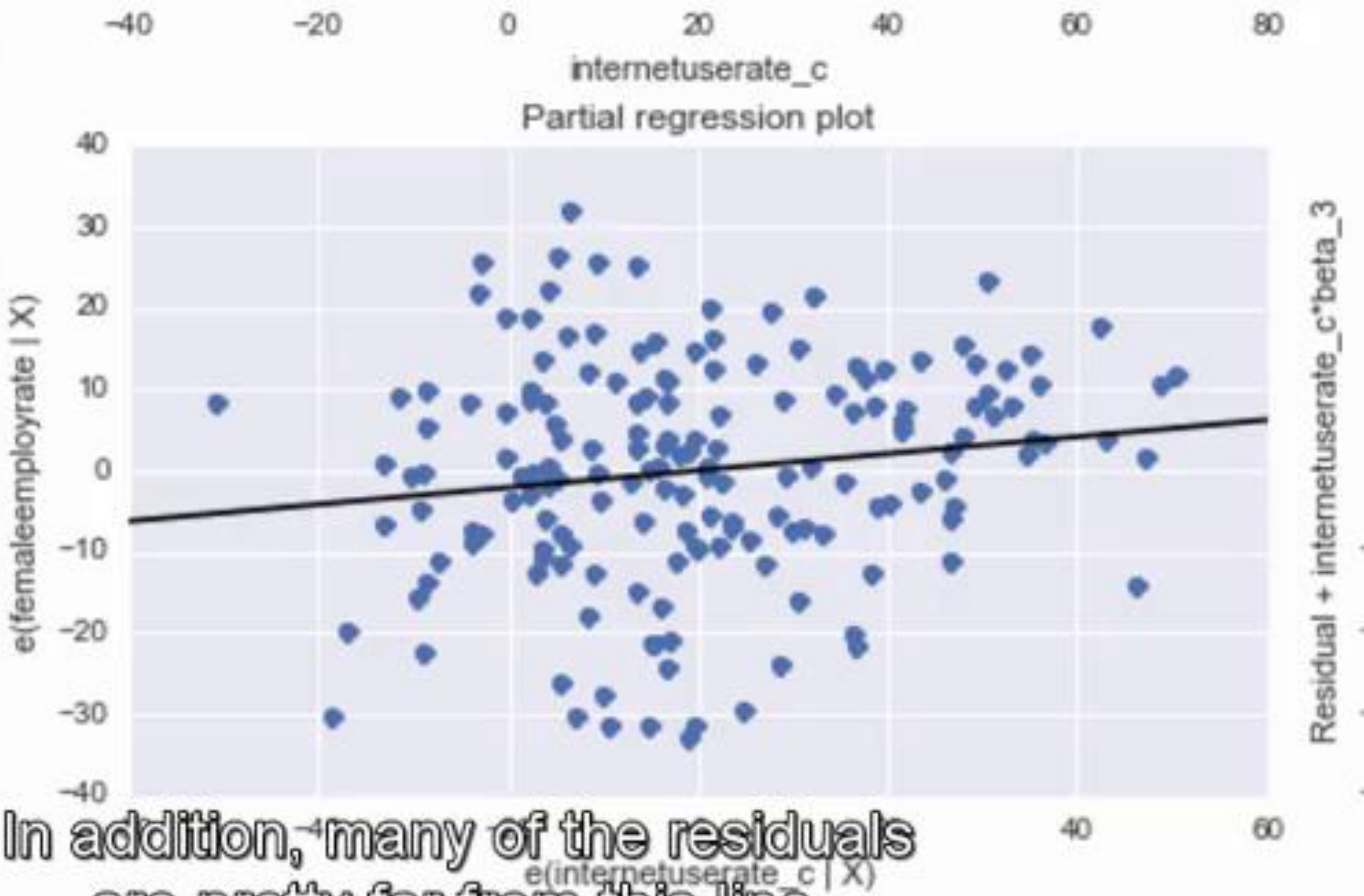

Internet use does not clearly indicate a non-linear association.



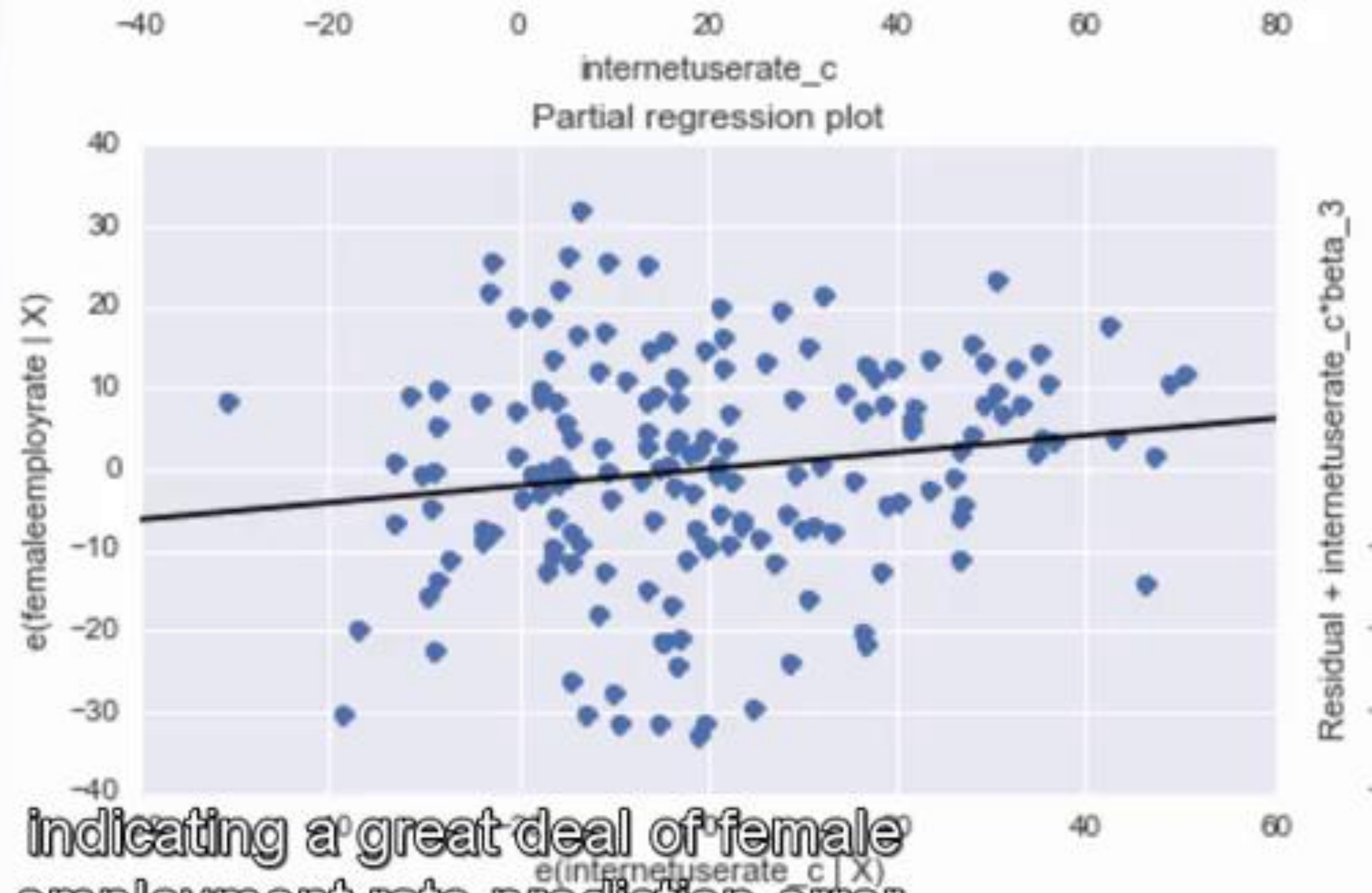

Rather, the residuals are spread out in a random pattern



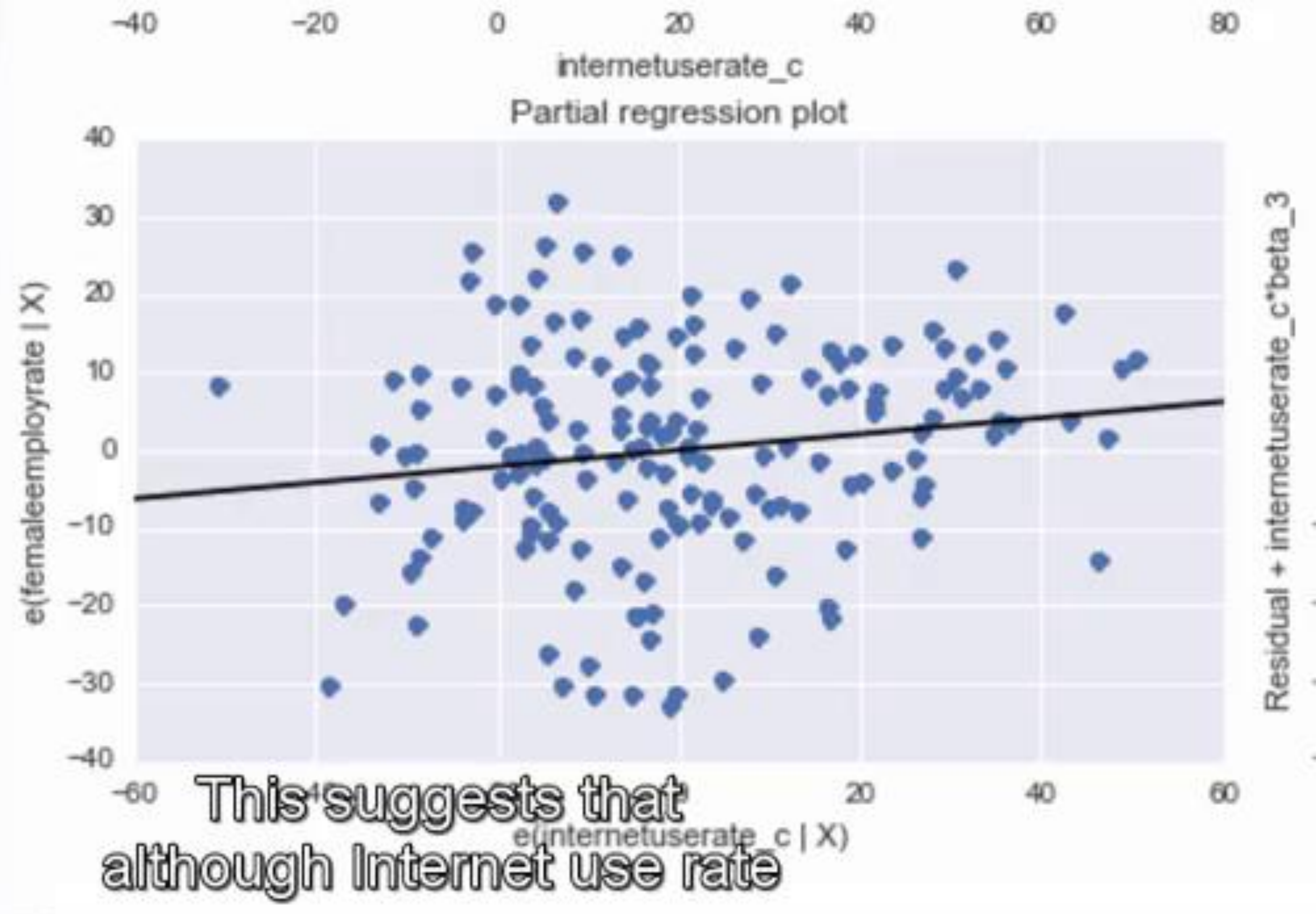

around the partial regression line.



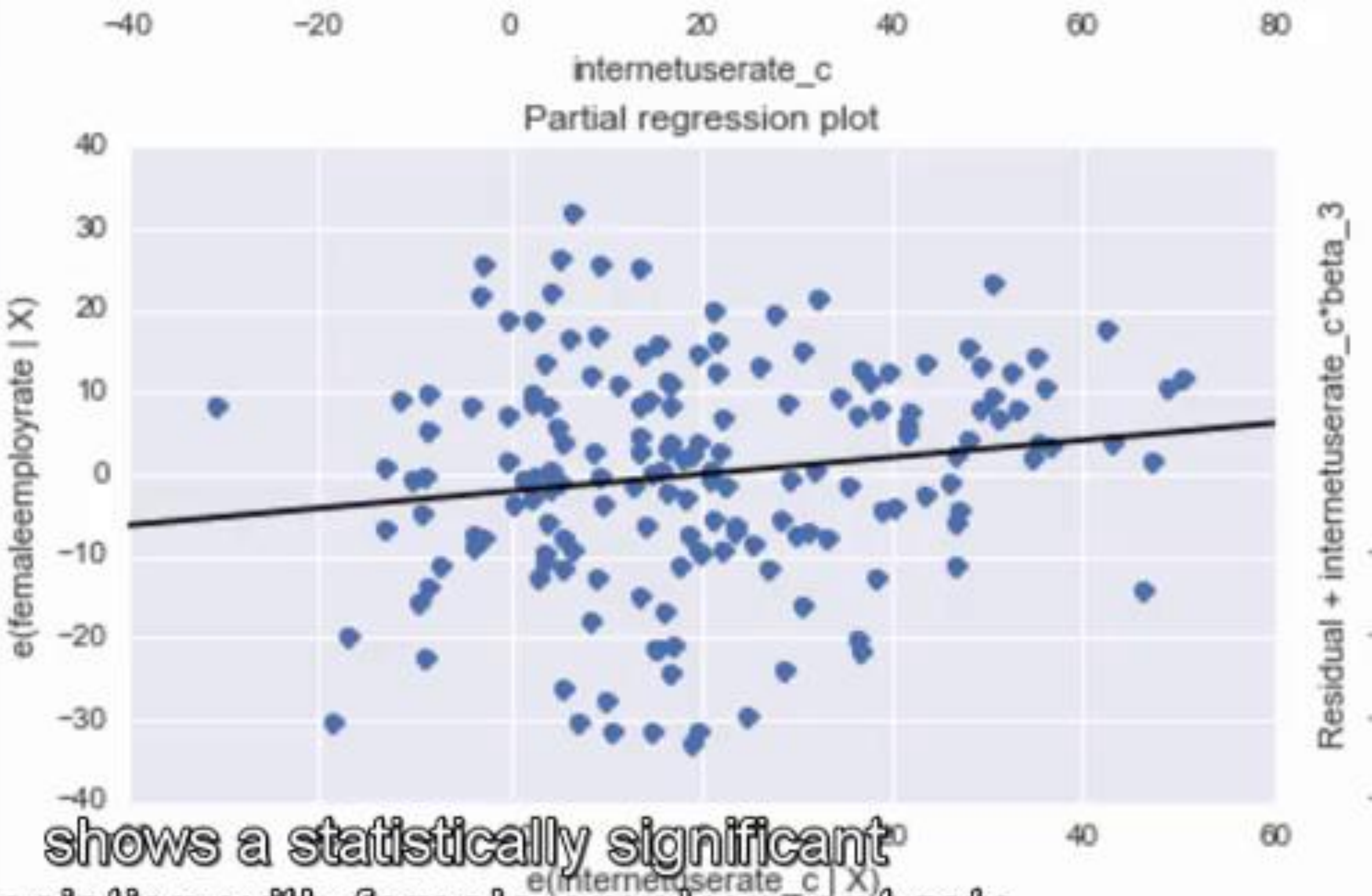

In addition, many of the residuals
are pretty far from this line,



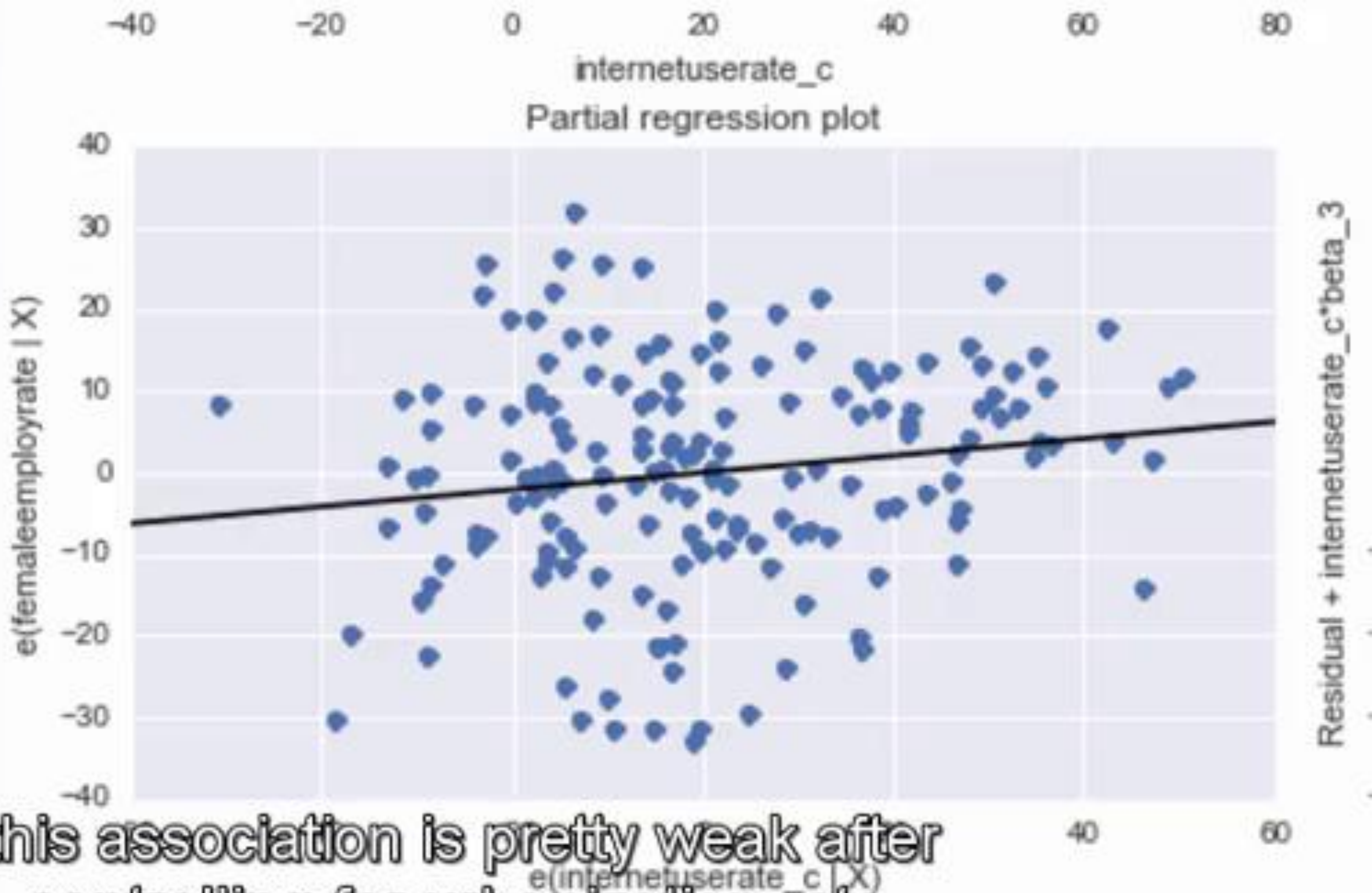

indicating a great deal of female employment rate prediction error.



-40
-20
0
20
40
60
80
internetuserate_c
Partial regression plot
e(femaleemployrate | X)
Residual + internetuserate_c*beta_3
e(internetuserate_c | X)
This suggests that
although Internet use rate




shows a statistically significant
association with female employment rate,




this association is pretty weak after controlling for urbanization rate.

```python
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')
print(fig2)


# additional regression diagnostic plots
fig3 = plt.figure(figsize(12,8))
fig3 = sm.graphics.plot_regress_exog(reg3,  "internetuserate_c", fig=fig3)


# leverage plot
fig4=sm.graphics.influence_plot(reg3, size=8)
print(fig4)
```

Finally, we can examine a leverage plot
to identify observations that have

```
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')
print(fig2)


# additional regression diagnostic plots
fig3 = plt.figure(figsize(12,8))
fig3 = sm.graphics.plot_regress_exog(reg3,  "internetuserate_c", fig=fig3)


# leverage plot
fig4=sm.graphics.influence_plot(reg3, size=8)
print(fig4)
```

an unusually large influence on
the estimation of the predicted value of

```
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')
print(fig2)


# additional regression diagnostic plots
fig3 = plt.figure(figsize(12,8))
fig3 = sm.graphics.plot_regress_exog(reg3,  "internetuserate_c", fig=fig3)


# leverage plot
fig4=sm.graphics.influence_plot(reg3, size=8)
print(fig4)
```

of how much the predicted scores for
the other observations would differ

```
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')
print(fig2)


# additional regression diagnostic plots
fig3 = plt.figure(figsize(12,8))
fig3 = sm.graphics.plot_regress_exog(reg3,  "internetuserate_c", fig=fig3)


# leverage plot
fig4=sm.graphics.influence_plot(reg3, size=8)
print(fig4)
```

if the observations in question
were not included in the analysis.

```python
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')
print(fig2)


# additional regression diagnostic plots
fig3 = plt.figure(figsize(12,8))
fig3 = sm.graphics.plot_regress_exog(reg3,  "internetuserate_c", fig=fig3)


# leverage plot
fig4=sm.graphics.influence_plot(reg3, size=8)
print(fig4)
```

The leverage always takes on values between zero and one.

```python
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')
print(fig2)


# additional regression diagnostic plots
fig3 = plt.figure(figsize(12,8))
fig3 = sm.graphics.plot_regress_exog(reg3,  "internetuserate_c", fig=fig3)


# leverage plot
fig4=sm.graphics.influence_plot(reg3, size=8)
print(fig4)
```

A point with zero leverage has no effect on the regression model.

```
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')
print(fig2)


# additional regression diagnostic plots
fig3 = plt.figure(figsize(12,8))
fig3 = sm.graphics.plot_regress_exog(reg3,  "internetuserate_c", fig=fig3)


# leverage plot
fig4=sm.graphics.influence_plot(reg3, size=8)
print(fig4)
```

And outliers are observations with residuals greater than 2 or less than -2.

```
fig2 ...
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')
print(fig2)


# additional regression diagnostic plots
fig3 = plt.figure(figsize(12,8))
fig3 = sm.graphics.plot_regress_exog(reg3,  "internetuserate_c", fig=fig3)


# leverage plot
fig4=sm.graphics.influence_plot(reg3, size=8)
print(fig4)
```

but this time we use
the code influence_plot.

```
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')
print(fig2)


# additional regression diagnostic plots
fig3 = plt.figure(figsize(12,8))
fig3 = sm.graphics.plot_regress_exog(reg3,  "internetuserate_c", fig=fig3)


# leverage plot
fig4=sm.graphics.influence_plot(reg3, size=8)
print(fig4)
```

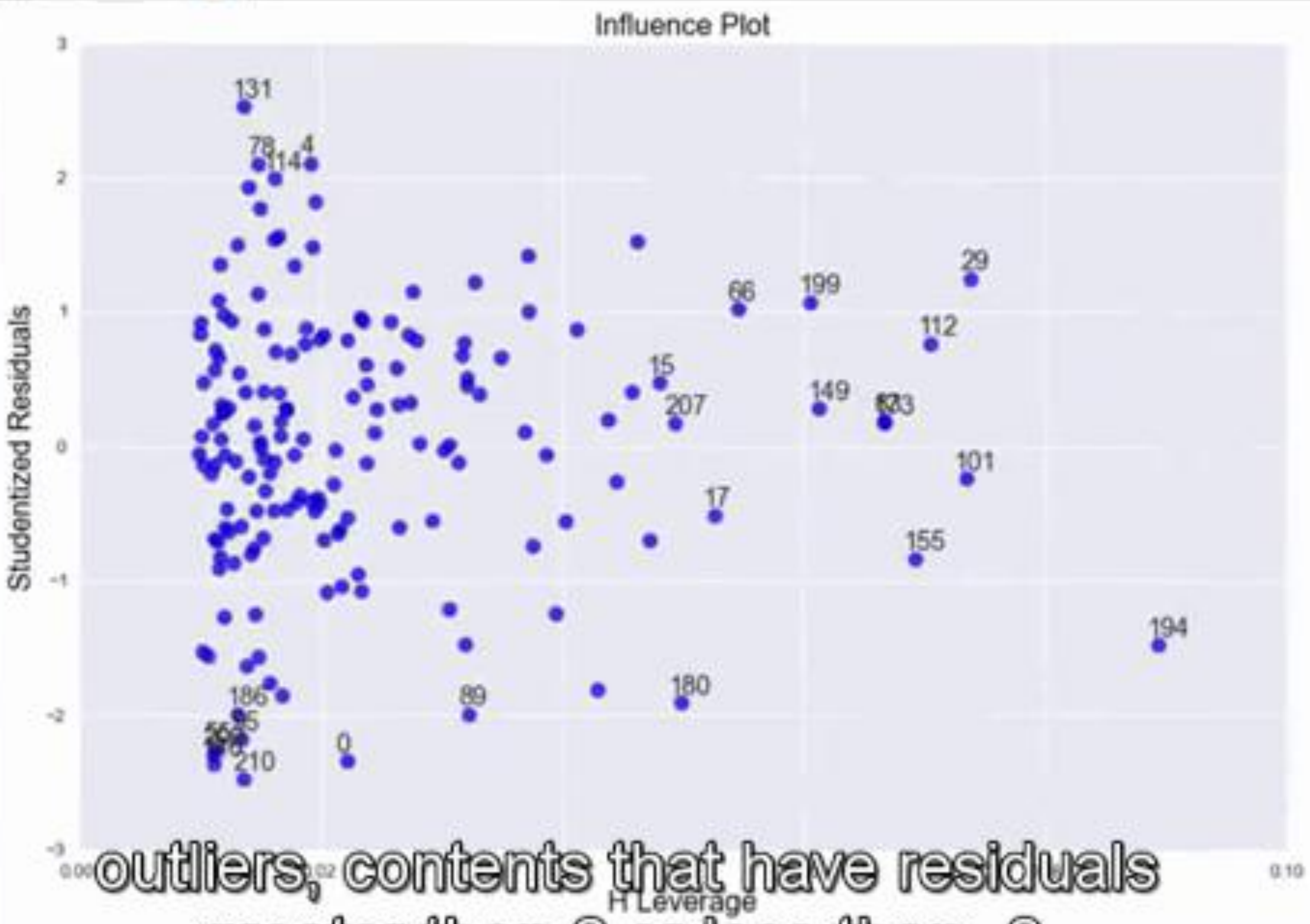
Size=8 is an option to make the points on the plot smaller than the default size so

```
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')
print(fig2)


# additional regression diagnostic plots
fig3 = plt.figure(figsize(12,8))
fig3 = sm.graphics.plot_regress_exog(reg3,  "internetuserate_c", fig=fig3)


# leverage plot
fig4=sm.graphics.influence_plot(reg3, size=8)
print(fig4)
```

that they're easier to distinguish.

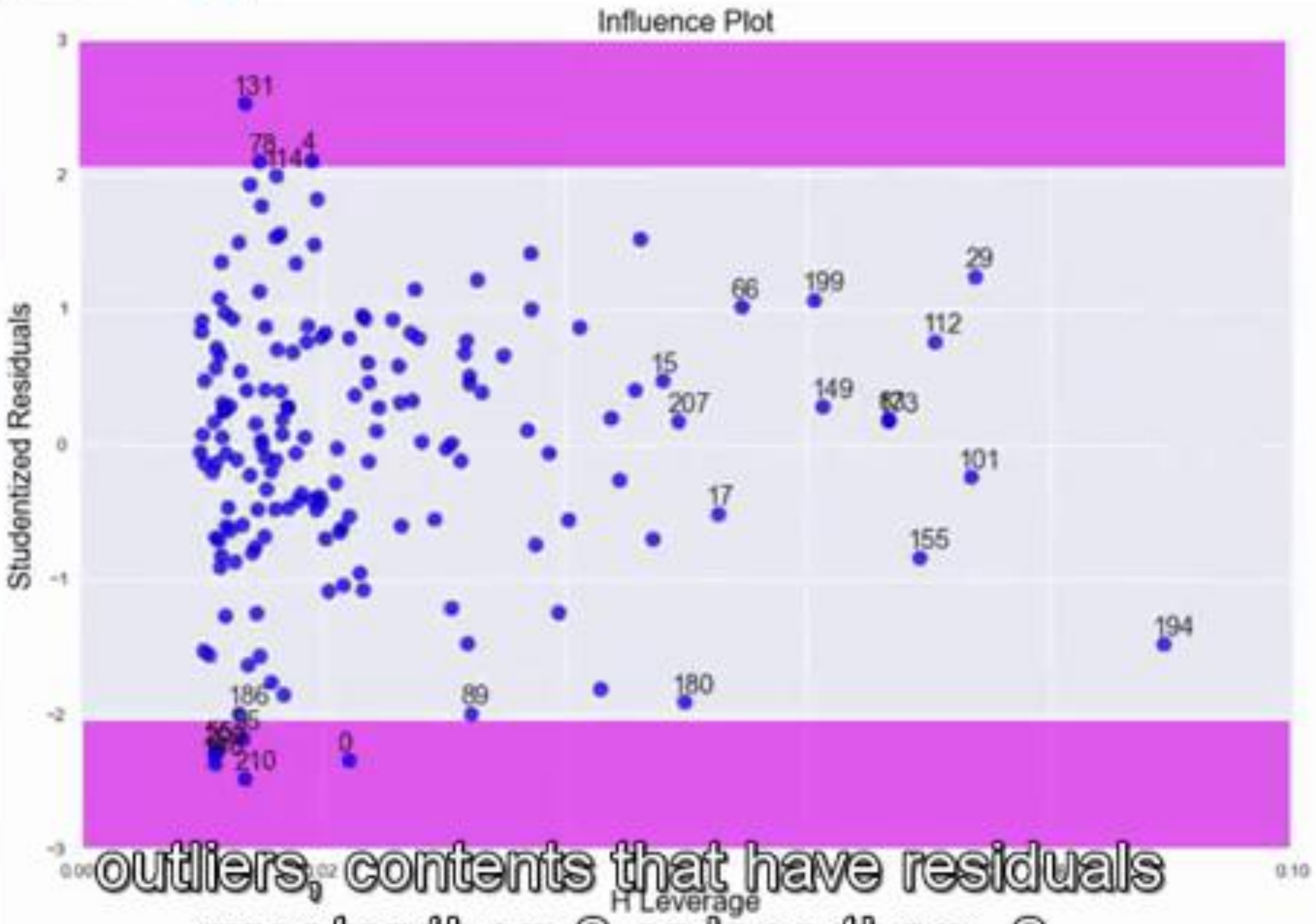

One of the first things we see in the leverage plot is that we have a few

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 78 Column: 1 Memory: 29 %



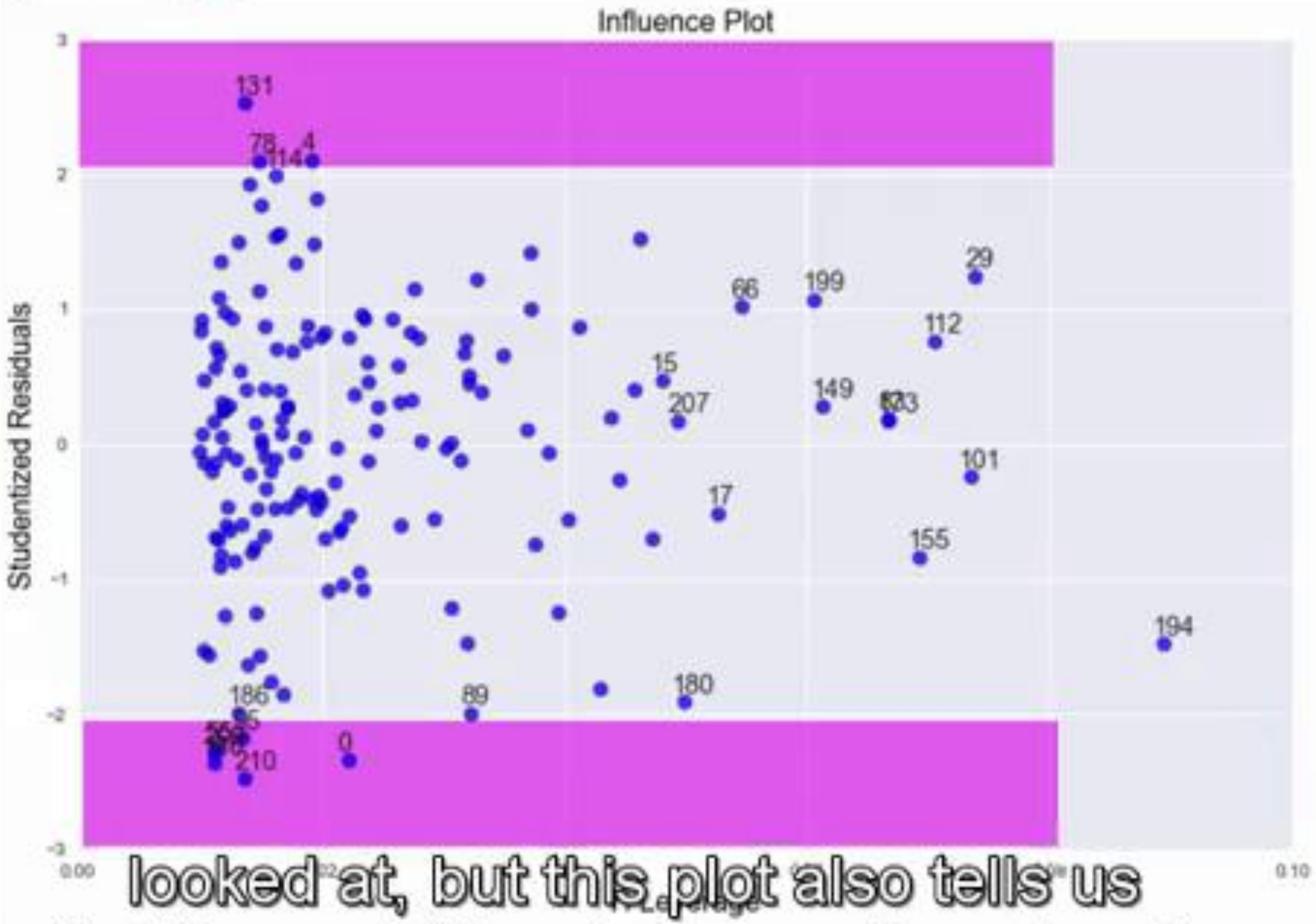

outliers, contents that have residuals greater than 2 or less than -2.

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 78 Column: 1 Memory: 29 %



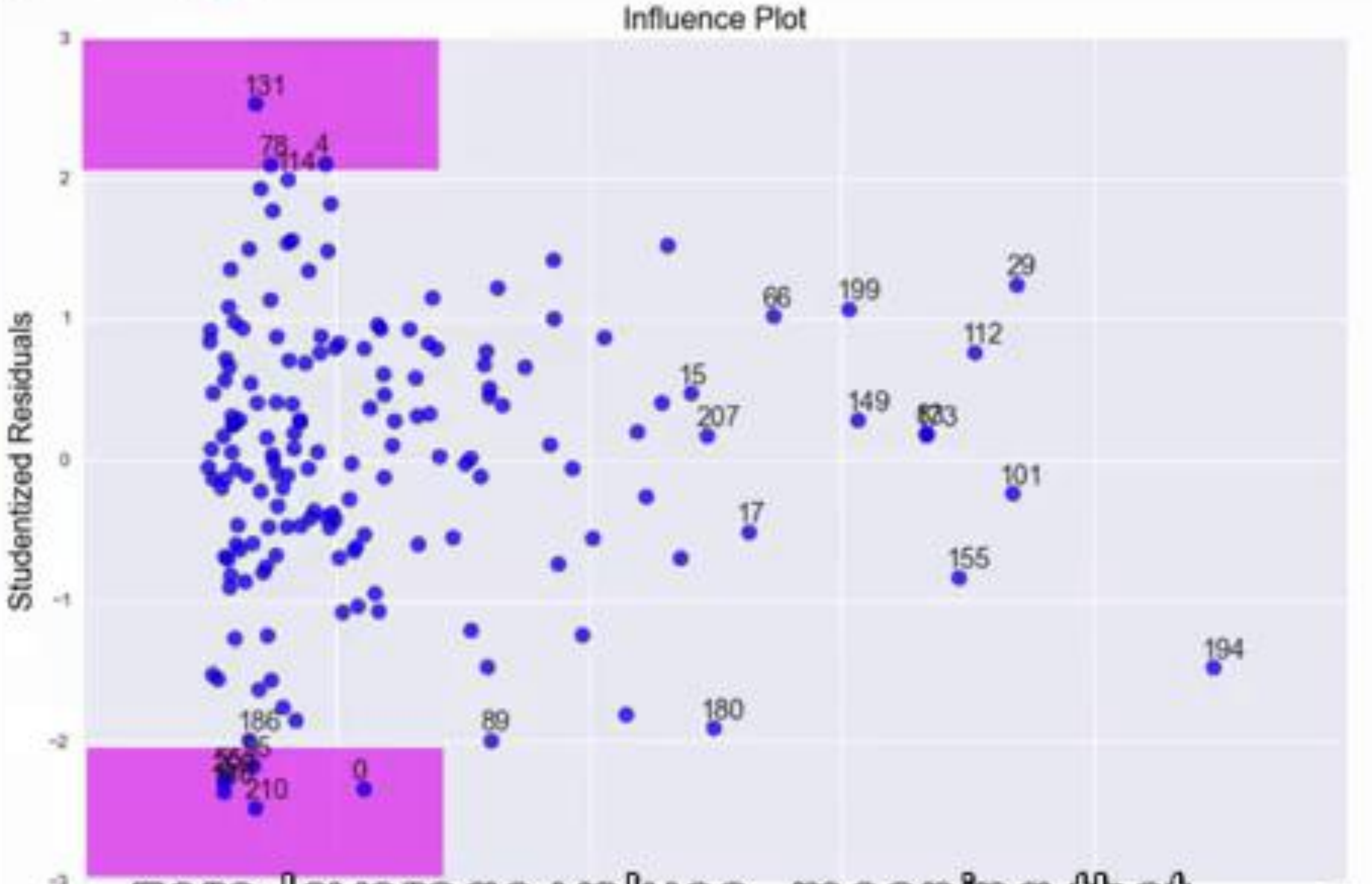

outliers, contents that have residuals greater than 2 or less than -2.

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 78 Column: 1 Memory: 29 %



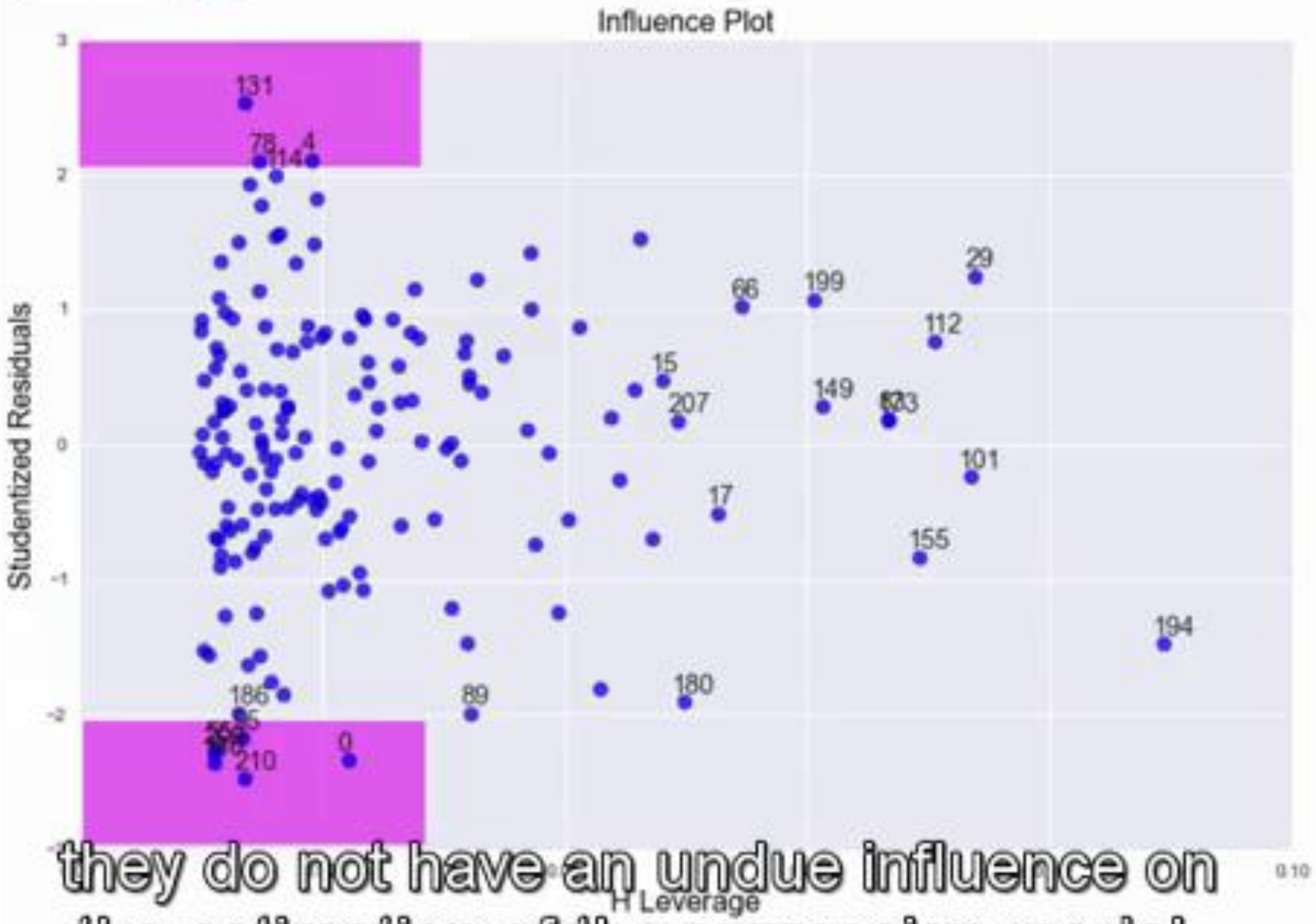

looked at, but this plot also tells us
that these outliers have small or close to

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 78 Column: 1 Memory: 29 %



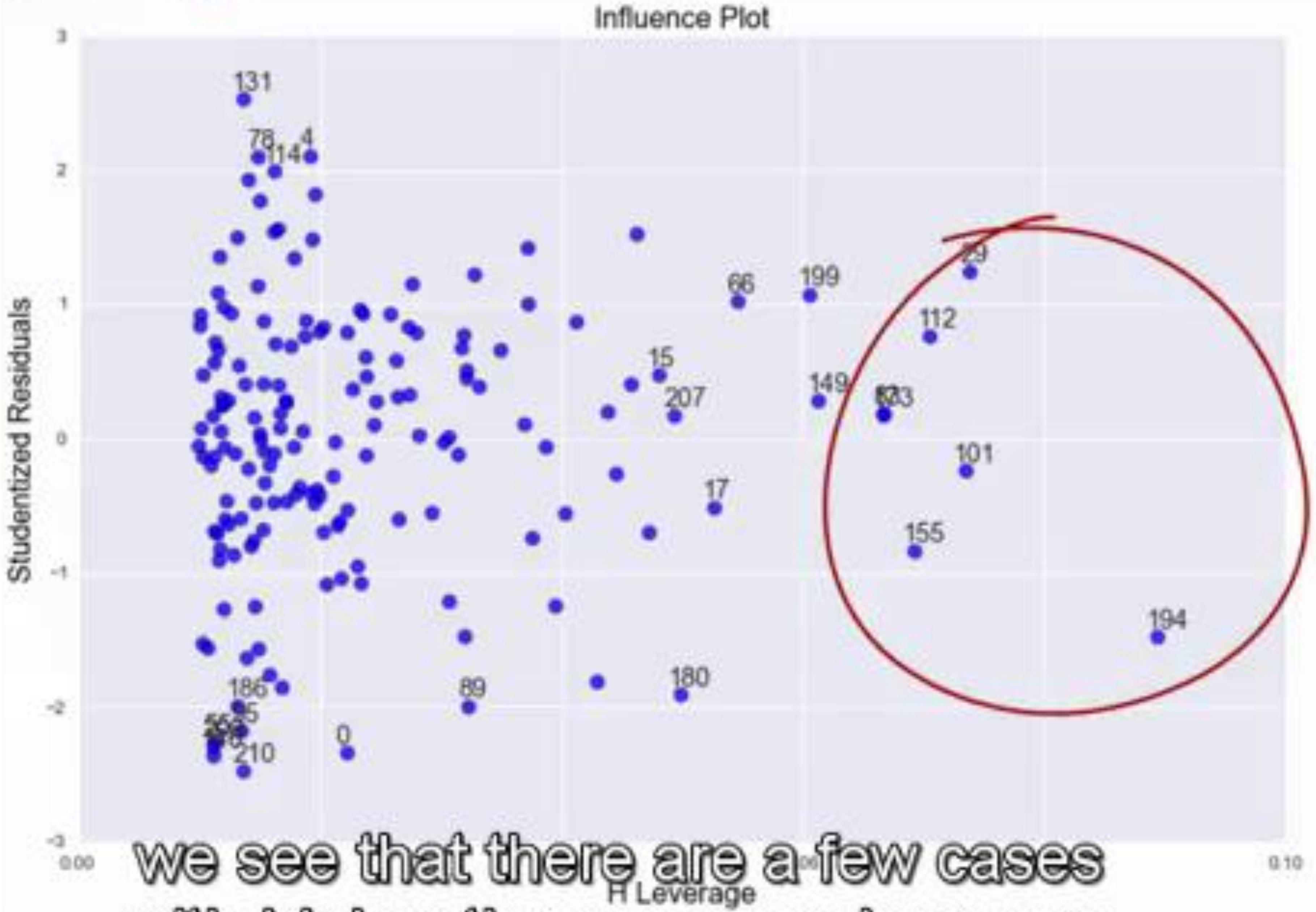

zero leverage values, meaning that
although they are outlying observations,

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 78 Column: 1 Memory: 29 %



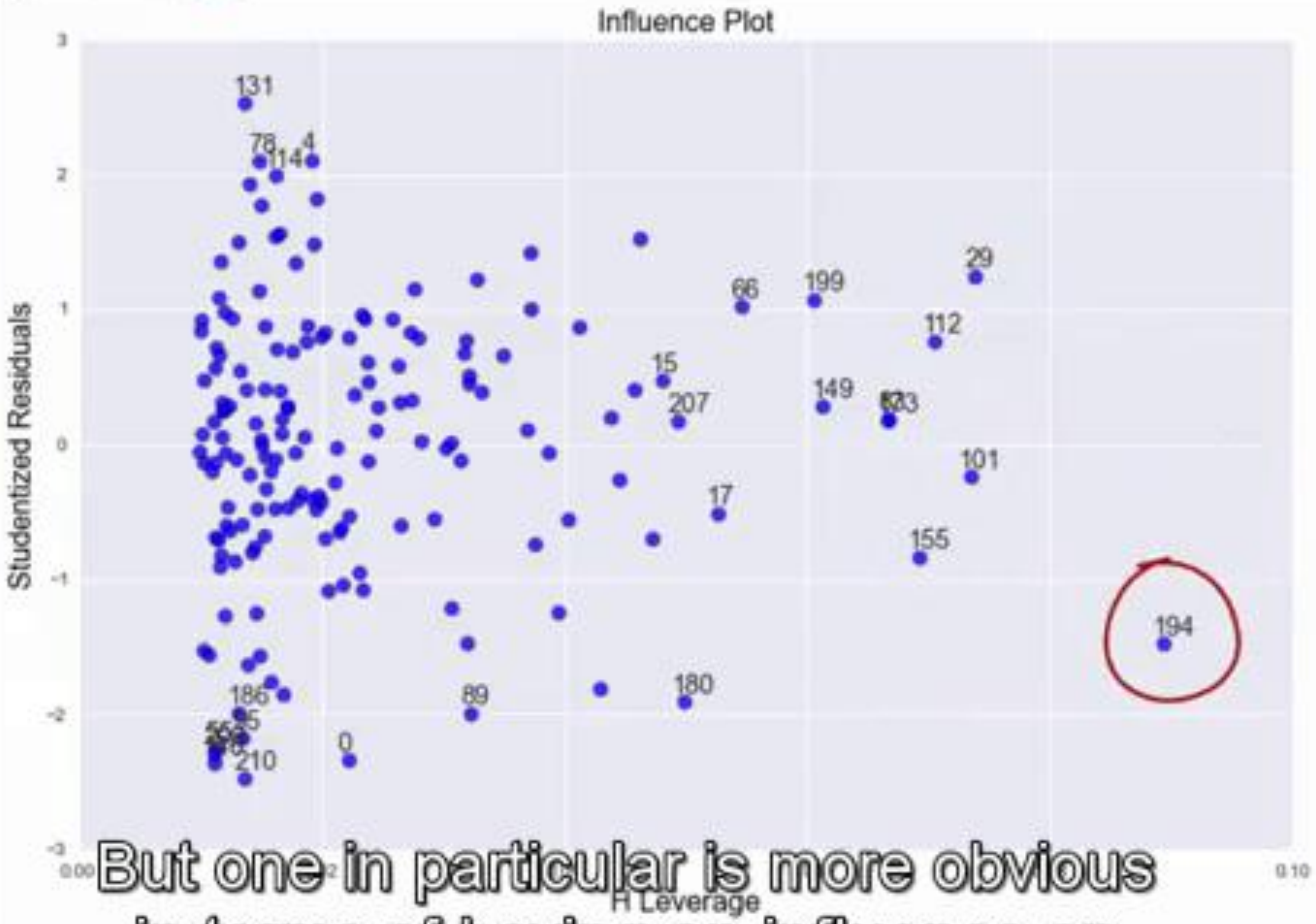

they do not have an undue influence on the estimation of the regression model.

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 78 Column: 1 Memory: 29 %




we see that there are a few cases
with higher than average leverage.

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 78 Column: 1 Memory: 29 %

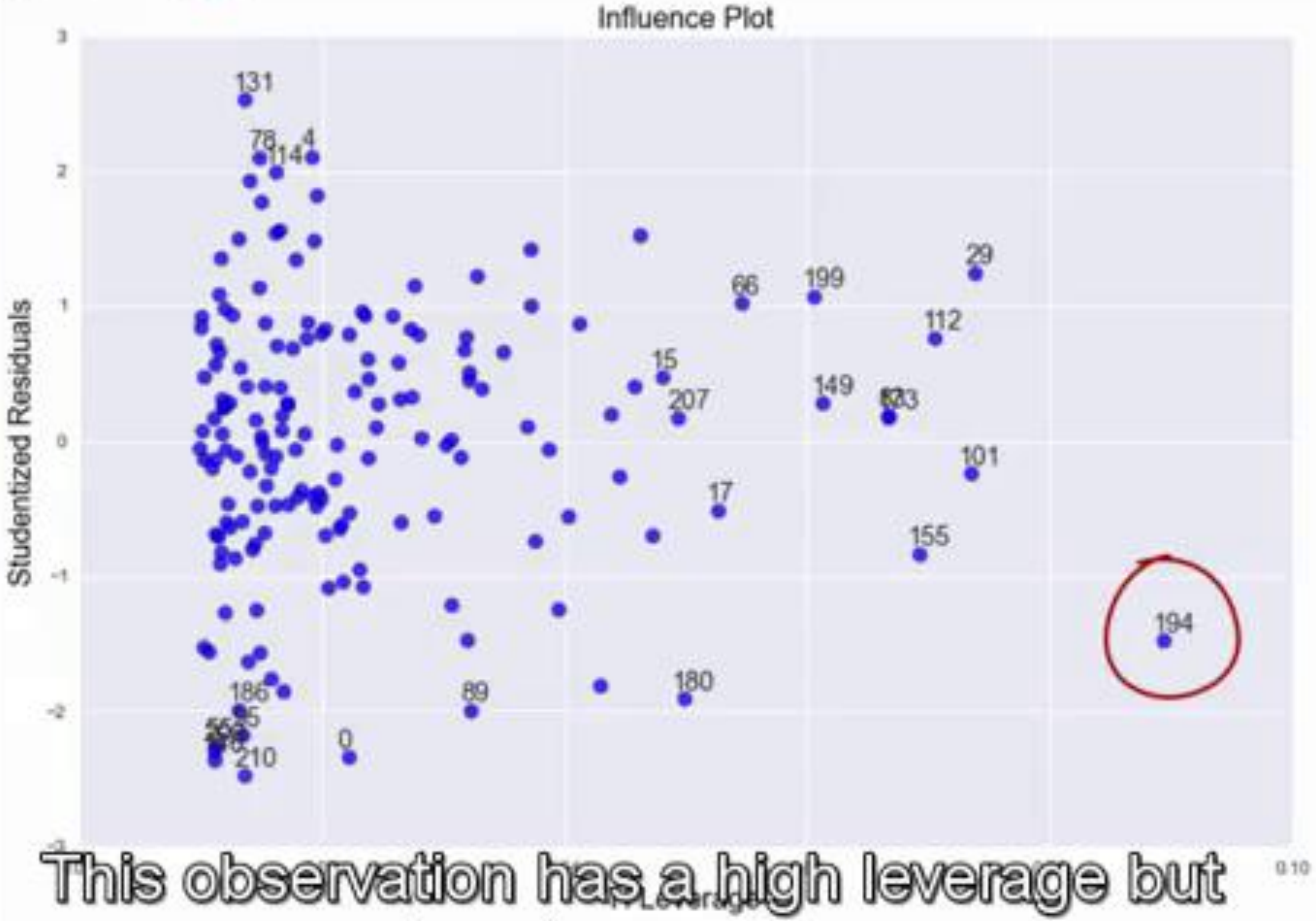



But one in particular is more obvious
in terms of having an influence on

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 78 Column: 1 Memory: 29 %


the estimation of the predicted
value of female employment rate.

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 78 Column: 1 Memory: 29 %




This observation has a high leverage but is not an outlier.

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 78 Column: 1 Memory: 29 %


We don't have any observations that
are both high leverage and outliers.

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 78 Column: 1 Memory: 29 %