PS3_Code

November 23, 2024

1 Pset 3

Muhammad Bashir

```
[60]: # Load Libraries
      import pandas as pd
      import numpy as np
      import matplotlib.pyplot as plt
      import statsmodels.api as sm
      import random
      # ignore warnings
      import warnings
      warnings.filterwarnings('ignore')
```

```
[61]: # set paths to working directory
      path = '/Users/muhammadbashir/GitHub/MuhammadCourses/Ec240a/Problem Sets'
      # load RPS_calorie_data.out data and read only columns YOtc and XOte.
      nlsy97ss = pd.read_csv('/Users/muhammadbashir/GitHub/MuhammadCourses/Ec240a/
        ↪Ec240a_Fall2023/Data/NLSY97/nlsy97ss.csv')
      nlsy97ss['LogEarn'] = np.log(nlsy97ss['avg_earn_2014_to_2018'])
```

```
[62]: def summary(data):
          # Create a table of summary statistics for avg_earn_2014_to_2018, LogEarn,␣
        ↪hgc_ever and asvab for this sub-sample.
          summary_stats = data[['avg_earn_2014_to_2018', 'LogEarn', 'hgc_ever',␣
        ↪'asvab']].describe()
          summary_stats = summary_stats.rename(columns={
              'avg_earn_2014_to_2018': 'Average Earnings (2014-2018)',
              'LogEarn': 'Log of Earnings',
              'hgc_ever': 'Highest Grade Completed',
              'asvab': 'ASVAB Score'
          })
          summary_stats.loc['count'] = summary_stats.loc['count'].astype('int64')
          summary_stats = summary_stats.rename(index={'count':'Number of␣
        ↪Observations','mean':'Mean','50%': 'Median', 'std': 'SD', 'min': 'Minimum',␣
        ↪'max': 'Maximum', '25%': 'Q1', '75%': 'Q3'})
          summary_stats.index.name = 'Statistic'
          summary_stats = summary_stats.round(2)
```

```
    print(summary_stats)
```

[63]:

```
# do least square fit of LogEarn on hgc_ever and a constant.
def LSQfit(data):
    X = data['hgc_ever']
    X = sm.add_constant(X)
    Y = data['LogEarn']
    model = sm.OLS(Y, X).fit(cov_type='HC3')
    return model
```

[64]:

```
# LS fit of LogEarn on hgc_ever and asvab, constant
def LSQfit2(data):
    """Least squares fit of LogEarn on hgc_ever and asvab"""
    X = data[['hgc_ever', 'asvab']]
    X = sm.add_constant(X)
    Y = data['LogEarn']
    model = sm.OLS(Y, X).fit(cov_type='HC3')
    return model
```

[65]:

```
# create variables for asvab-50, (asvab-50)*hgc_ever and then regression of␣
 ↪logearn on hgc_ever, asvab-50, (asvab-50)*hgc_ever and a constant
def LSQfit3(data):
    """Least squares fit of LogEarn on hgc_ever, asvab, and interaction term"""
    nlsy97ss1['asvab_50'] = nlsy97ss1['asvab'] - 50
    nlsy97ss1['asvab_50_hgc_ever'] = nlsy97ss1['asvab_50'] *␣
 ↪nlsy97ss1['hgc_ever']
    X = nlsy97ss1[['hgc_ever', 'asvab', 'asvab_50_hgc_ever']]
    X = sm.add_constant(X)
    Y = nlsy97ss1['LogEarn']
    model = sm.OLS(Y, X).fit(cov_type='HC3')
    return model
```

[66]:

```
# plot coeffcient estimates  0 +  0 (asvab − 50) against asvab
def predict_yhat(model,data):
    beta0 = model.params['hgc_ever']
    gamma0 = model.params['asvab_50_hgc_ever']
    data['asvab_50_hgc_ever_hat'] = beta0 + gamma0 * data['asvab_50']
    return data
```

[67]:

```
def bayesian_bootstrap(data, num_bootstraps):
    """
    Perform Bayesian bootstrap to estimate the distribution of OLS coefficients.

    Parameters:
    - Y: 1D array-like, dependent variable.
    - X: 2D array-like, independent variables (including a constant if needed).
    - num_bootstraps: int, number of bootstrap samples.
    Returns:
    - beta_hat: NumPy array of shape (num_bootstraps, number_of_parameters),
                containing bootstrap estimates of the coefficients.
    """
    beta0 = []
    gamma0 = []
    N = len(data)

    for i in range(num_bootstraps):
        # Draw weights from Gamma(1,1) and normalize to sum to 1
        W = np.random.gamma(1,1,N)
        W = np.array(W)/sum(W)
        # create variables for asvab-50, (asvab-50)*hgc_ever and then␣
↪regression of logearn on hgc_ever, asvab-50, (asvab-50)*hgc_ever and a␣
↪constant
        X = data[['hgc_ever', 'asvab', 'asvab_50_hgc_ever']]
        X = sm.add_constant(X)
        Y = data['LogEarn']
        model = sm.WLS(Y, X, weights=W).fit(cov_type='HC3')
        # Append the parameter estimates as a dictionary with variable names␣
↪beta0 = model.params['hgc_ever']
        beta0.append(model.params['hgc_ever'])
        gamma0.append(model.params['asvab_50_hgc_ever'])

    return beta0, gamma0
```

[68]:

```
# use each iteration of beta0 and gamma0 to predict the value of  0 +  0 (asvab␣
↪- 50) for each observation in the sample
def predict_LB_UB(data,beta0,gamma0,num_bootstraps):
    """ Predict the 95% confidence interval for  0 +  0 (asvab - 50)"""
    asvab_50_hgc_ever_hat = np.array([b0 + g0 * data['asvab_50'] for b0, g0 in␣
↪zip(beta0, gamma0)])
    upper_i = int(np.floor(num_bootstraps * .025))
    lower_i = int(np.floor(num_bootstraps * .975))
    lower_bound = []
    upper_bound = []
    for i in range(len(data['asvab_50'])):
        level_i_prediction = asvab_50_hgc_ever_hat[:, i]
        # sort the predictions
        level_i_prediction.sort()
        # get the 95% confidence interval
        lower_bound.append(level_i_prediction[lower_i])
        upper_bound.append(level_i_prediction[upper_i])

    data['asvab_LB'] = lower_bound
    data['asvab_UB'] = upper_bound
    return data
```

```
[69]: # plot the 95% confidence interval for  0 +  0 (asvab - 50) against asvab. Sort␣
       ↪the data by asvab before plotting.
      def plot_CI(data):
          data = data.sort_values(by='asvab')
          plt.scatter(data['asvab'], data['asvab_50_hgc_ever_hat'], color='blue',␣
       ↪s=10)
          plt.plot(data['asvab'], data['asvab_LB'], color='red')
          plt.plot(data['asvab'], data['asvab_UB'], color='red')
          plt.xlabel('ASVAB Score')
          plt.ylabel(' 0 +  0 (ASVAB - 50)')
          plt.title(' 0 +  0 (ASVAB - 50) against ASVAB')
          plt.legend(['Point Estimate from OLS', '95% Confidence Interval'])
          plt.show()
```

1.1 Using First subseting of data as in the question

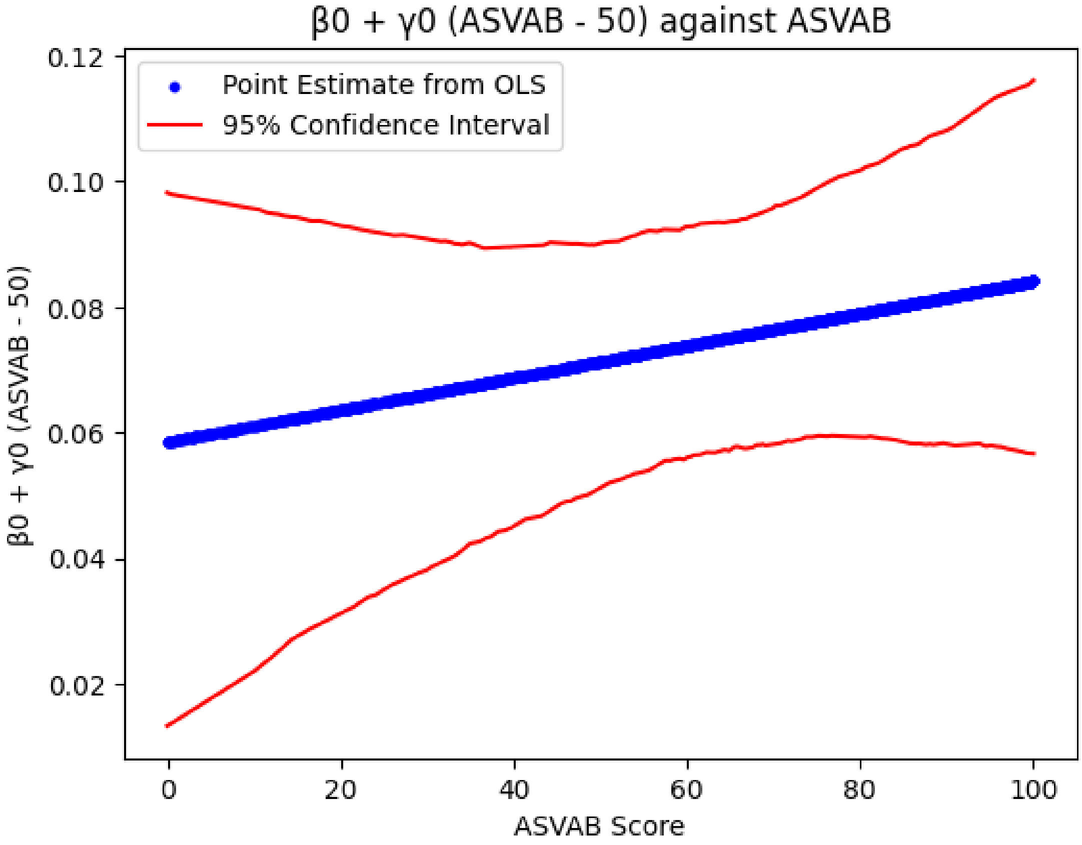
```
[70]: # subset to non-black,non-hispanic, non-female respondents with positive␣
       ↪earings in 2014-2018
      nlsy97ss1 = nlsy97ss[(nlsy97ss['black'] == 0) & (nlsy97ss['hispanic'] == 0) &␣
       ↪(nlsy97ss['female'] == 0) & (nlsy97ss['avg_earn_2014_to_2018'] > 0)]
      # summary statistics
      summary(nlsy97ss1)
      # LSQfit
      Lsq1 = LSQfit(nlsy97ss1)
      print("OLS Regression 1")
      print(Lsq1.summary())
      # LSQfit2
      lsq2 = LSQfit2(nlsy97ss1)
      print("OLS Regression 2")
      print(lsq2.summary())
      # LSQfit3
      lsq3 = LSQfit3(nlsy97ss1)
      print("OLS Regression 3")
      print(lsq3.summary())
      # plot coefficent estimates  0 +  0 (asvab - 50) against asvab
      nlsy97ss1 = predict_yhat(lsq3,nlsy97ss1)
      # Bayesian Bootstrap
      num_bootstraps=1000
      [beta0, gamma0]= bayesian_bootstrap(nlsy97ss1, num_bootstraps)
      nlsy97ss1 = predict_LB_UB(nlsy97ss1,beta0,gamma0,num_bootstraps)
      plot_CI(nlsy97ss1)
```

```
                        Average Earnings (2014-2018)  Log of Earnings  \
Statistic
Number of Observations                       1606.00          1606.00
```

```
Mean                                        75821.77            10.93
SD                                          59827.90             0.91
Minimum                                        58.45             4.07
Q1                                          38395.24            10.56
Median                                      61895.37            11.03
Q3                                          94180.04            11.45
Maximum                                    383978.89            12.86

                        Highest Grade Completed  ASVAB Score
Statistic
Number of Observations                  1606.00      1606.00
Mean                                      14.35        56.95
SD                                         3.01        28.40
Minimum                                    6.00         0.00
Q1                                        12.00        33.78
Median                                    14.00        59.53
Q3                                        16.00        82.03
Maximum                                   20.00       100.00
OLS Regression 1
                            OLS Regression Results
==============================================================================
Dep. Variable:                LogEarn   R-squared:                       0.103
Model:                            OLS   Adj. R-squared:                  0.102
Method:                 Least Squares   F-statistic:                     188.7
Date:                Sat, 23 Nov 2024   Prob (F-statistic):           1.14e-40
Time:                        15:46:53   Log-Likelihood:                -2048.6
No. Observations:                1606   AIC:                             4101.
Df Residuals:                    1604   BIC:                             4112.
Df Model:                           1
Covariance Type:                  HC3
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          9.5328      0.107     89.068      0.000       9.323       9.743
hgc_ever       0.0973      0.007     13.735      0.000       0.083       0.111
==============================================================================
Omnibus:                      778.410   Durbin-Watson:                   1.881
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7568.453
Skew:                          -2.035   Prob(JB):                         0.00
Kurtosis:                      12.825   Cond. No.                         71.7
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
OLS Regression 2
                            OLS Regression Results
==============================================================================
Dep. Variable:                LogEarn   R-squared:                       0.116
```

```
Model:                            OLS   Adj. R-squared:                  0.115
Method:                 Least Squares   F-statistic:                     107.1
Date:                Sat, 23 Nov 2024   Prob (F-statistic):           2.22e-44
Time:                        15:46:53   Log-Likelihood:                -2036.1
No. Observations:                1606   AIC:                             4078.
Df Residuals:                    1603   BIC:                             4094.
Df Model:                           2
Covariance Type:                  HC3
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          9.6195      0.108     89.463      0.000       9.409       9.830
hgc_ever       0.0732      0.009      7.973      0.000       0.055       0.091
asvab          0.0046      0.001      4.123      0.000       0.002       0.007
==============================================================================
Omnibus:                      780.475   Durbin-Watson:                   1.907
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7638.672
Skew:                          -2.040   Prob(JB):                         0.00
Kurtosis:                      12.875   Cond. No.                         322.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
OLS Regression 3
                            OLS Regression Results
==============================================================================
Dep. Variable:                LogEarn   R-squared:                       0.117
Model:                            OLS   Adj. R-squared:                  0.115
Method:                 Least Squares   F-statistic:                     75.74
Date:                Sat, 23 Nov 2024   Prob (F-statistic):           8.14e-46
Time:                        15:46:53   Log-Likelihood:                -2035.7
No. Observations:                1606   AIC:                             4079.
Df Residuals:                    1602   BIC:                             4101.
Df Model:                           3
Covariance Type:                  HC3
================================================================================
=====
                        coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
-----
const                 9.8137      0.275     35.694      0.000       9.275
10.353
hgc_ever              0.0713      0.010      7.202      0.000       0.052
0.091
asvab                 0.0010      0.004      0.222      0.824      -0.008
0.009
asvab_50_hgc_ever     0.0003      0.000      0.841      0.400      -0.000
```

```
0.001
==============================================================================
Omnibus:                      779.561   Durbin-Watson:                   1.908
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7612.557
Skew:                          -2.037   Prob(JB):                         0.00
Kurtosis:                      12.857   Cond. No.                     4.73e+03
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
[2] The condition number is large, 4.73e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```



2. Compute the least squares fit of LogEarn onto a constant and hgc_ever. Report the point estimate on the schooling variable as well as its heteroscedastic robust asymptotic standard error (you may use the StatsModels implementation of OLS to do this; later in the course we will construct our own program for these calculations).

The point estimate on schooling variable is 0.0973 and hetroskadestic robust standard error on this is 13.735. Note that schooling significantly predicts eanrings in this model and higher schooling leads to more earnings.

3. Compute the least squares fit of LogEarn on a constant, hgc_ever and asvab. Does the estimate coefficient on hgc_ever change?

Yes the coefficient changes as the value is now lower which means some of variations in earnings that was being captured buy education before is due to asvab.

4. Estimate the parameters of the following linear regression model by the method of least squares E∗[LogEarn| X] = 0 + 0hgc_ever + 0hgc_ever × (asvab − 50) + 0asvab where X = (hgc_ever, hgc_ever × (asvab − 50), asvab)'.

(a) Provide a semi-elasticity interpretation of 0.
(b) Provide a semi-elasticity interpretation of 0 + 0 (asvab − 50).

a. A one-year increase in schooling is associated with a 7.13% increase in earnings.
b. (0 + 0(asvab-50))*100 gives percentage change in earnings with one extra year of schooling for those with given level of asvab-50

5. Construct a plot with the OLS estimate of 0 + 0 (asvab − 50) on the y-axis and a grid of asvab values on the x-axis.

6. Using the Bayes' Bootstrap to approximate a posterior distribution for 0 + 0 (asvab − 50) at each value of asvab shown in your plot. Add (estimates of) the 0.025 and 0.975 quantiles, as well as the mean, of the posterior distribution of 0 + 0 (asvab − 50) to your plot.

The marginal impact increases as asvab score increases. But there is certain non-linearity into Bayesian confidence intervals. In general, for SEs in OLS, I had assume structure of error term to estimate error but in this case I did not need any specfication of error term to get CI. This is great about Bayesian. However, we had to assume gamma weights and I am not sure how sensitive results are to that.

1.2 Using 2nd subseting where instead of white males I look at white females

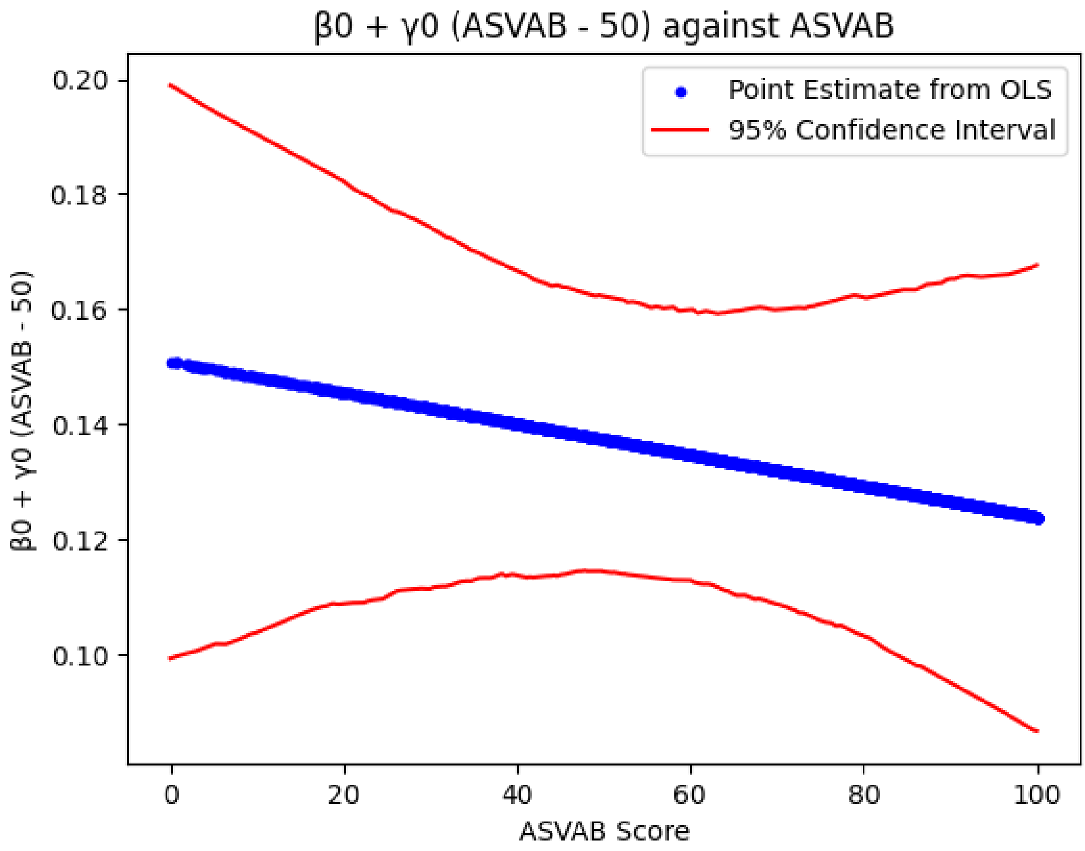
[73]:

```
# subset to non-black,non-hispanic, females respondents with positive earings␣
↪in 2014-2018
nlsy97ss1 = nlsy97ss[(nlsy97ss['black'] == 0) & (nlsy97ss['hispanic'] == 0) &␣
↪(nlsy97ss['female'] == 1) & (nlsy97ss['avg_earn_2014_to_2018'] > 0)]
# summary statistics
summary(nlsy97ss1)
# LSQfit
Lsq1 = LSQfit(nlsy97ss1)
print("OLS Regression 1")
print(Lsq1.summary())
# LSQfit2
lsq2 = LSQfit2(nlsy97ss1)
print("OLS Regression 2")
print(lsq2.summary())
# LSQfit3
lsq3 = LSQfit3(nlsy97ss1)
print("OLS Regression 3")
print(lsq3.summary())
# plot coefficent estimates  0 +  0 (asvab - 50) against asvab
```

```
nlsy97ss1 = predict_yhat(lsq3,nlsy97ss1)
# Bayesian Bootstrap
num_bootstraps=1000
[beta0, gamma0]= bayesian_bootstrap(nlsy97ss1, num_bootstraps)
predict_LB_UB(nlsy97ss1,beta0,gamma0,num_bootstraps)
plot_CI(nlsy97ss1)
```

```
                        Average Earnings (2014-2018)  Log of Earnings  \
Statistic
Number of Observations                       1449.00          1449.00
Mean                                        50839.69            10.38
SD                                          45172.37             1.18
Minimum                                        81.54             4.40
Q1                                          20749.86             9.94
Median                                      42422.72            10.66
Q3                                          66817.67            11.11
Maximum                                    336241.07            12.73

                        Highest Grade Completed  ASVAB Score
Statistic
Number of Observations                  1449.00      1449.00
Mean                                      15.18        59.36
SD                                         3.06        25.94
Minimum                                    0.00         0.00
Q1                                        13.00        39.78
Median                                    16.00        62.25
Q3                                        17.00        81.43
Maximum                                   20.00       100.00
OLS Regression 1
                            OLS Regression Results
==============================================================================
Dep. Variable:                LogEarn   R-squared:                       0.172
Model:                            OLS   Adj. R-squared:                  0.171
Method:                 Least Squares   F-statistic:                     204.2
Date:                Sat, 23 Nov 2024   Prob (F-statistic):           1.99e-43
Time:                        15:47:25   Log-Likelihood:                -2155.1
No. Observations:                1449   AIC:                             4314.
Df Residuals:                    1447   BIC:                             4325.
Df Model:                           1
Covariance Type:                  HC3
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          7.9640      0.180     44.266      0.000       7.611       8.317
hgc_ever       0.1594      0.011     14.289      0.000       0.138       0.181
==============================================================================
Omnibus:                      500.448   Durbin-Watson:                   1.977
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2007.828
```

```
Skew:                          -1.630   Prob(JB):                         0.00
Kurtosis:                       7.757   Cond. No.                         78.7
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
OLS Regression 2
                            OLS Regression Results
==============================================================================
Dep. Variable:                LogEarn   R-squared:                       0.185
Model:                            OLS   Adj. R-squared:                  0.184
Method:                 Least Squares   F-statistic:                     135.7
Date:                Sat, 23 Nov 2024   Prob (F-statistic):           9.41e-55
Time:                        15:47:25   Log-Likelihood:                -2143.0
No. Observations:                1449   AIC:                             4292.
Df Residuals:                    1446   BIC:                             4308.
Df Model:                           2
Covariance Type:                  HC3
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          7.9700      0.172     46.362      0.000       7.633       8.307
hgc_ever       0.1355      0.013     10.788      0.000       0.111       0.160
asvab          0.0060      0.001      4.480      0.000       0.003       0.009
==============================================================================
Omnibus:                      521.725   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2139.340
Skew:                          -1.699   Prob(JB):                         0.00
Kurtosis:                       7.887   Cond. No.                         337.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
OLS Regression 3
                            OLS Regression Results
==============================================================================
Dep. Variable:                LogEarn   R-squared:                       0.186
Model:                            OLS   Adj. R-squared:                  0.184
Method:                 Least Squares   F-statistic:                     91.41
Date:                Sat, 23 Nov 2024   Prob (F-statistic):           3.62e-54
Time:                        15:47:25   Log-Likelihood:                -2142.7
No. Observations:                1449   AIC:                             4293.
Df Residuals:                    1445   BIC:                             4315.
Df Model:                           3
Covariance Type:                  HC3
================================================================================
=====
                        coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
-----
const                  7.7549      0.379     20.438      0.000       7.011
8.499
hgc_ever               0.1374      0.012     11.048      0.000       0.113
0.162
asvab                  0.0100      0.007      1.463      0.143      -0.003
0.023
asvab_50_hgc_ever     -0.0003      0.000     -0.630      0.529      -0.001
0.001
==============================================================================
Omnibus:                      522.563   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2137.797
Skew:                          -1.703   Prob(JB):                         0.00
Kurtosis:                       7.879   Cond. No.                     4.81e+03
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
[2] The condition number is large, 4.81e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```


β0 + γ0 (ASVAB - 50) against ASVAB
Point Estimate from OLS
95% Confidence Interval
β0 + γ0 (ASVAB - 50)
ASVAB Score