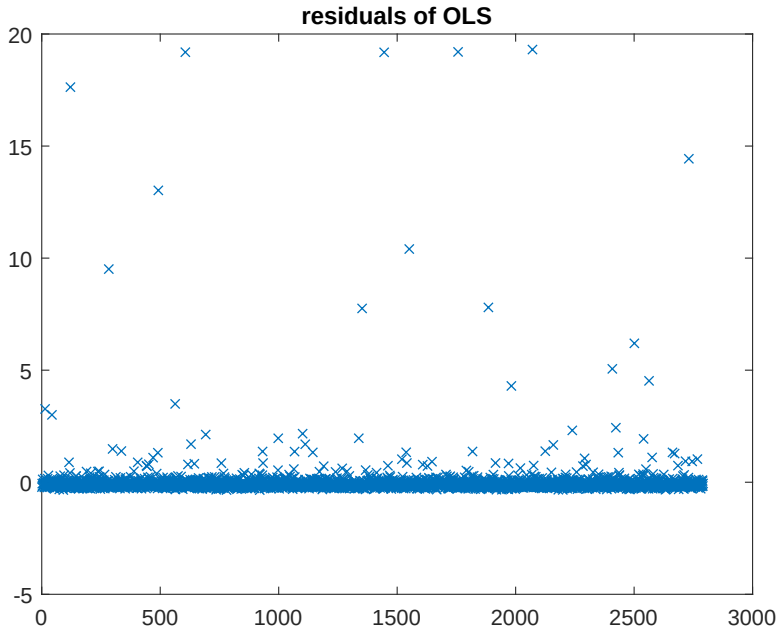
First, we run regression.

```
% Title:  Matlab script for group project, STATS 506
% Author: Daxuan Deng
% Funtion: explore the relationship between health condition and drinking
%          alcohol, using data NHANES 2005-2006 data.
% Date: 11/30/2019

% load data
alq = xptread('ALQ_D.XPT');
demo = xptread('DEMO_D.XPT');
hsq = xptread('HSQ_D.XPT');

% select variables
alq = alq(:, {'SEQN', 'ALQ120Q'});
demo = demo(:, {'SEQN', 'RIAGENDR', 'RIDAGEYR', 'DMDEDUC2', 'INDFMPIR'});
hsq = hsq(:, {'SEQN', 'HSD010'});

% merge data
dt = join(alq, demo, 'Keys', 'SEQN');
dt = join(dt, hsq, 'Keys', 'SEQN');

% rename columns
dt.Properties.VariableNames = ...
["id", "alcohol", "sex", "yr", "edu", "pir", "health"];

% drop invalid values
dt = rmmissing(dt);
dt(dt.alcohol > 365, :) = [];
dt(dt.yr < 21, :) = [];
dt(dt.edu > 5, :) = [];
dt(dt.health > 3, :) = [];

% centralize and factorize
dt.alcohol = (dt.alcohol - mean(dt.alcohol)) ./ std(dt.alcohol);
dt.sex = categorical(dt.sex);
dt.yr = (dt.yr - mean(dt.yr)) ./ std(dt.yr);
dt.edu = categorical(dt.edu);
dt.pir = (dt.pir - mean(dt.pir)) ./ std(dt.pir);
dt.health = categorical(dt.health);
% set 'Good' as base level
dt.health = reordercats(dt.health, {'3','1','2'});

% run OLS
md = fitlm(dt, 'alcohol ~ sex + yr + edu + pir + health');
```

```
% extract fitted values and residuals
fit = predict(md, dt(:, 3:7));
res = md.Residuals.Raw;
coef = md.Coefficients(:,1);

% plot residuals
plot(1:height(dt), res, 'x'), title('residuals of OLS')
```



The plot reveals that the data is skewed. So residual bootstrap will be an appropriate way to estimate standard error.

```
% bootstrap
rng(506);
nboot = 1000;

% resample residuals
func = @(x)x;
res_boot = bootstrp(nboot, func, res);

dt_boot = dt(:, 3:7);
beta_boot = zeros(nboot, 10);

for i=1:nboot
    % generate new response
    dt_boot.alcohol = fit + res_boot(i,:)';

    % fit new model
```

```
    md_boot = fitlm(dt_boot, 'alcohol ~ sex + yr + edu + pir + health');

    % extract new estimate
    beta_boot(i,:) = table2array(md_boot.Coefficients(:,1))';
end

% calculate std err
se = std(beta_boot);

% summary
result = coef;
result.se = se';
result.t = result.Estimate ./ result.se;
result.pvalue = 1-tcdf(abs(result.t),1);

result

result =

  10x4 table
```

	Estimate	se	t	pvalue
(Intercept)	-0.097662	0.074424	-1.3122	0.20728
sex_2	-0.066717	0.037603	-1.7742	0.16337
yr	-0.0055428	0.020298	-0.27307	0.41515
edu_2	0.10222	0.090457	1.13	0.23059
edu_3	0.10132	0.078252	1.2948	0.20934
edu_4	0.14169	0.080976	1.7498	0.16527
edu_5	0.087332	0.085008	1.0273	0.24571
pir	0.016709	0.022218	0.75205	0.29475
health_1	0.059674	0.062881	0.94899	0.25833
health_2	0.049529	0.041953	1.1806	0.2237

We could learn from the table that health does not have strong relationship with drinking alcohol.