Econometrics: Multiple Regression and Applications

ECON4004: LAB 2

Duong Trinh

University of Glasgow

February 14, 2024

Intro

- Duong Trinh
  - PhD Student in Economics (Bayesian Microeconometrics)
  - Email: Duong.Trinh@glasgow.ac.uk
- ECON4004-LB01
  - Wednesday 10am -12 pm
  - 5 sessions (7-Feb, 14-Feb, 21-Feb, 28-Feb, 6-March)
  - ST ANDREWS:357
- ECON4004-LB02
  - Wednesday 12-2 pm
  - 5 sessions (7-Feb, 14-Feb, 21-Feb, 28-Feb, 6-March)
  - ST ANDREWS:357

Record Attendance

Plan for LAB 2

- Exercise 2: based on Stock and Watson, E13.1
- Exercise 1: based on Stock and Watson, E8.1

- We will focus on *incorporating interactions between two independent variables into a regression model*.

Exercise 2: based on Stock and Watson, E13.1

Picture the Scenario

- **Objective:** Examine Labor Market Discrimination: Are Emily and Greg More Employable Than Lakisha and Jamal?
- **Dataset:** `Names.dta`
  - Experimental data from research on US labor market.
- **Key variables:**
  - `call_back`: callback rate, measured by fraction of resumes that generate a phone call from prospective employer.
  - `black`: $= 1$ for "African American–sounding name" resumes, $= 0$ for "white-sounding name" resumes.
  - `female`: $= 1$ for women, $= 0$ for men.
  - `high`: $= 1$ for high-quality resumes, $= 0$ for low-quality resumes.

Picture the Scenario

Randomized Controlled Experiment (Bertrand and Mullainathan, 2004)

- A prospective employer receives two resumes: a resume from a white job applicant and a similar resume from an African American applicant. Is the employer more likely to call back the white applicant to arrange an interview?
- Because race is not typically included on a resume, they differentiated resumes on the basis of “white-sounding names” such as Emily Walsh or Gregory Baker) and “African American–sounding names” (such as Lakisha Washington or Jamal Jones).
- A large collection of fictitious resumes was created, and the presupposed “race” (based on the “sound” of the name) was randomly assigned to each resume.
- These resumes were sent to prospective employers to see which resumes generated a phone call (a callback) from the prospective employer.

Questions

Random assignment & Average effect

(»review)

(a) What was the callback rate for whites? For African Americans? Construct a 95% confidence interval for difference in callback rates. Is the difference statistically significant? Is it large in a real-world sense?

(d) Authors of study claim that race was assigned randomly to the resumes. Is there any evidence of nonrandom assignment?

Questions

Heterogeneous effects across subgroups

(b) Is the African American/white callback rate differential different for men than for women?

(c) What is the difference in callback rates for high-quality versus low quality resumes?
What is the high-quality/low-quality difference for white applicants? For African American applicants?
Is there a significant difference in this high-quality/low-quality difference for whites versus African Americans?

(a) Callback rates for Whites versus African Americans?

Approach 1: Linear regression model

◇ Model specification

$$call\_back_i = \beta_0 + \beta_1 \cdot black_i + u_i$$

$$\begin{aligned} E[call\_back \mid black = 1] &= \beta_0 + \beta_1 && \rightarrow \text{for African Americans} \\ E[call\_back \mid black = 0] &= \beta_0 && \rightarrow \text{for White} \\ \Delta &= \beta_1 && \rightarrow \text{the Difference} \end{aligned}$$

(a) Callback rates for Whites versus African Americans?

- OLS estimation results (»stata)

$$\underset{(se)}{\widehat{call\_back}} = \underset{(0.006)}{0.097} - \underset{(0.008)}{0.032} \cdot black$$

- *On average*, the call-back rate for whites is $\hat{\beta}_0 = 0.097$ and the call-back for blacks is $\hat{\beta}_0 + \hat{\beta}_1 = 0.097 - 0.032 = 0.065$. This implies that 9.7% of resumes with white-sounding names generated a call back, while only 6.5% of resumes with black-sounding names generated a call back.

- The difference is $\hat{\beta}_1 = -0.032$, which is statistically significant at the 1% significance level ($t - statistic = -4.11, p - value = 0.00 < 0.01$).

(a) Callback rates for Whites versus African Americans?

Approach 2: Two-sample t test

- Purpose: Test if two population means are equal (reference).
- The data may either be *paired* or *unpaired*. The variances of the two samples may be assumed to be *equal* or *unequal*.

- STATA note:

```
* ttest yvar, by(groupvar)
// Test if mean(yvar) equal between 2 groups defined by groupvar
```

```
* ttest yvar, by(groupvar) unequal
// Test if mean(yvar) equal between 2 groups defined by groupvar
// add option 'unequal' to assume unequal variances
```

(a) Callback rates for Whites versus African Americans?

Approach 2: Two-sample t test

```
. ttest call_back, by(black) unequal

Two-sample t test with unequal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
0	2,435	.0965092	.0059853	.295349	.0847724	.1082461
1	2,435	.0644764	.0049781	.2456501	.0547145	.0742382
Combined	4,870	.0804928	.0038988	.2720826	.0728493	.0881363
diff		.0320329	.007785		.0167707	.047295

```
    diff = mean(0) - mean(1)                                      t =   4.1147
H0: diff = 0                     Satterthwaite's degrees of freedom =   4711.6

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

⇒ Same conclusion as the first approach.

(d) Is there any evidence of nonrandom assignment of race to resumes?

- Idea: Is there statistically significant difference in other characteristics for two groups - black and white sounding names? (»review)

- Use any approach in (a), calculate estimated means of other characteristics for these two groups and test whether the difference is statistically significant.

- There are only two significant differences in the mean values: the call-back rate (the outcome variable of interest) and computer skills (for which black-named resumes had a slightly higher fraction that white-named resumes).

$\Rightarrow$ There is no evidence of non-random assignment.

(d) Is there any evidence of nonrandom assignment of race to resumes?

	Black-Sounding Names			White-Sounding Names			Black-White Difference		
Variable	**n**	$\bar{X}$	**se**$(\bar{X})$	**n**	$\bar{X}$	**se**$(\bar{X})$	$\bar{X}_b - \bar{X}_w$	**se**$(\bar{X}_b - \bar{X}_w)$	***t*-stat**
ofjobs	2435	3.658	1.219	2435	3.664	1.219	–0.006	0.035	–0.18
yearsexp	2435	7.830	5.011	2435	7.856	5.079	–0.027	0.145	–0.18
honors	2435	0.051	0.221	2435	0.054	0.226	–0.003	0.006	–0.45
volunteer	2435	0.414	0.493	2435	0.409	0.492	0.006	0.014	0.41
military	2435	0.102	0.303	2435	0.092	0.290	0.009	0.008	1.11
empholes	2435	0.446	0.497	2435	0.450	0.498	–0.004	0.014	–0.29
workinschool	2435	0.561	0.496	2435	0.558	0.497	0.003	0.014	0.20
email	2435	0.480	0.500	2435	0.479	0.500	0.001	0.014	0.06
computerskills	2435	0.832	0.374	2435	0.809	0.393	0.024	0.011	2.17
specialskills	2435	0.327	0.469	2435	0.330	0.470	–0.003	0.013	–0.21

...

(c1) Callback rates for high-quality versus low-quality resumes?

- OLS estimation results (»stata)

$$\underset{(se)}{\widehat{call\_back}} = \underset{(0.005)}{0.073} + \underset{(0.008)}{0.014} \cdot high$$

- On average, the call-back rate for low-quality resumes is 0.073 and for high-quality resumes is $0.073 + 0.014 = 0.087$.
- The difference is 0.014, which is not significant at the 5% level, but is at the 10% level ($t - statistic = 1.80$)

(c2) A significant difference in the high-quality/low-quality difference for whites versus African Americans?

Model specification

$$call\_back_i = \beta_0 + \beta_1 \cdot black_i + \beta_2 \cdot high_i + \beta_3 \cdot black_i \times high_i + u_i$$

$$E[call\_back \mid high = 1, black = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3 \rightarrow \text{for high-quality blacks}$$

$$E[call\_back \mid high = 0, black = 1] = \beta_0 + \beta_1 \rightarrow \text{for low-quality blacks}$$

$$\Delta^{HvL}_{black=1} = \beta_2 + \beta_3 \rightarrow \text{the h/l difference in black group}$$

$$E[call\_back \mid high = 1, black = 0] = \beta_0 + \beta_2 \rightarrow \text{for high-quality whites}$$

$$E[call\_back \mid high = 0, black = 0] = \beta_0 \rightarrow \text{for low-quality whites}$$

$$\Delta^{HvL}_{black=0} = \beta_2 \rightarrow \text{the h/l difference in white group}$$

$$\Delta^{HvL}_{black=1} - \Delta^{HvL}_{black=0} = \beta_3$$

(c2) A significant difference in the high-quality/low-quality difference for whites versus African Americans?

- OLS estimation results (»stata)

$$\underset{(se)}{\widehat{call\_back}} = \underset{(0.008)}{0.084} - \underset{(0.011)}{0.023} \cdot black + \underset{(0.012)}{0.023} \cdot high - \underset{(0.016)}{0.018} \cdot black \times high$$

- On average, the high-quality/low-quality difference for whites is $\hat{\beta}_2 = 0.023$ and for blacks is $\hat{\beta}_2 + \hat{\beta}_3 = 0.023 - 0.018 = 0.005$.

- The black-white difference is $\hat{\beta}_3 = -0.018$, which is not statistically significant at the 5% level ($t - statistic = -1.14$).

Table of Results

	Dependent Variable = call_back		
Regressor	**(a)**	**(c1)**	**(c2)**
black	-0.032 (0.008)		-0.023 (0.011)
high		0.014 (0.008)	0.023 (0.012)
black $\times$ *high*			-0.018 (0.016)
Intercept	0.097 (0.006)	0.073 (0.005)	0.084 (0.008)

Notes: Standard errors shown in parentheses.

Exercise 1: based on Stock and Watson, E8.1

Picture the Scenario

- **Objective:** Investigate the effect of *lead water pipes* on *infant mortality* (*with a focus on interaction effects*).
- **Dataset:** `lead_mortality.dta`
  - Data for 172 U.S. cities in 1900.
- **Key variables:**
  - `Lead`: type of water pipes (lead or nonlead).
  - `Inf`: the average infant mortality rate.
  - `pH`: water acidity.
  - several demographic variables.

Questions

(a) Compute the average infant mortality rate (`Inf`) for cities with lead pipes and for cities with nonlead pipes.
Is there a statistically significant difference in the averages?

Questions

(b) The amount of lead leached from lead pipes depends on the chemistry of the water running through the pipes. The more acidic the water is (i.e. the lower its pH), the more lead is leached. Run a regression of `Inf` on `Lead`, `pH`, and the interaction term `Lead` $\times$ `pH`.

1. Explain what coefficients measure.
2. Plot the estimated regression function relating `Inf` to `pH` for $Lead = 0$ and for $Lead = 1$.
3. Does `Lead` have a statistically significant effect on `Inf`? Explain.
4. Does the effect of `Lead` on `Inf` depend on `pH`? Is this dependence statistically significant?
5. What is the median value of `pH` in the sample? At this `pH` level, what is the estimated effect of `Lead` on `Inf`? What is the standard deviation of `pH`?
   Suppose the pH level is one standard deviation lower than the median level of pH in the sample: What is the estimated effect of Lead on infant mortality?
   What if pH is one standard deviation higher than the median value?

(a) Compute the average `Inf` for cities with lead pipes and for cities with nonlead pipes. Is there a statistically significant difference in the averages?

Two-sample t test

```
* ttest yvar, by(groupvar) unequal
// Test if mean(yvar) equal between 2 groups defined by groupvar
// add option 'unequal' to assume unequal variances
```

(a) Compute the average `Inf` for cities with lead pipes and for cities with nonlead pipes. Is there a statistically significant difference in the averages?

```
. ttest infrate, by(lead) unequal

Two-sample t test with unequal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       0 |      55    .3811679    .0199238    .1477588     .341223    .4211127
       1 |     117    .4032576    .0141529    .1530873    .3752259    .4312892
---------+--------------------------------------------------------------------
Combined |     172     .396194    .0115384    .1513249    .3734179    .4189701
---------+--------------------------------------------------------------------
    diff |           -.0220897     .024439               -.0705255    .0263461
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -0.9039
H0: diff = 0                     Satterthwaite's degrees of freedom =  109.292

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1840         Pr(|T| > |t|) = 0.3681          Pr(T > t) = 0.8160
```

(a) Compute the average `Inf` for cities with lead pipes and for cities with nonlead pipes. Is there a statistically significant difference in the averages?

	n	$\bar{Y}$	SE($\bar{Y}$)
Lead	117	0.403	0.014
No Lead	55	0.381	0.020
Difference		*0.022*	*0.024*

- ◇ The difference in the sample means is 0.022 with a standard error of 0.024.
- ◇ The estimate implies that cities with lead pipes have a larger infant mortality rate (by 0.02 deaths per 100 people in the population), but the standard error is large (0.024) and the difference is not statistically significant ($t = 0.022/0.024 \approx 0.9$).

(b) Regression of `Inf` on `Lead`, `pH`, and the interaction term `Lead` $\times$ `pH`

Model specification

$$Inf_i = \beta_0 + \beta_1 \cdot Lead_i + \beta_2 \cdot pH_i + \beta_3 \cdot Lead_i \times pH_i + u_i$$

$$(1) \quad E[Inf \mid Lead, pH] = \beta_0 + (\beta_1 \cdot Lead + \beta_2 \cdot pH + \beta_3 \cdot pH \times Lead)$$

$$(2) \quad E[Inf \mid Lead, pH] = (\beta_0 + \beta_2 \cdot pH) + (\beta_1 + \beta_3 \cdot pH) \times Lead$$

$$(3) \quad E[Inf \mid Lead, pH] = (\beta_0 + \beta_1 \cdot Lead) + (\beta_2 + \beta_3 \cdot Lead) \times pH$$

(b1) Understand what coefficients measure.

From (1): $E[Inf \mid Lead, pH] = \beta_0 + (\beta_1 \cdot Lead + \beta_2 \cdot pH + \beta_3 \cdot pH \times Lead)$

$$E[Inf \mid Lead = 0, pH = 0] = \beta_0$$

$\Rightarrow$ The intercept β_0 shows the level of `Inf` when $Lead = 0$ and $pH = 0$. It dictates the level of the regression line.

(b1) Understand what coefficients measure.

From (2): $E[Inf \mid Lead, pH] = (\beta_0 + \beta_2 \cdot pH) + (\beta_1 + \beta_3 \cdot pH) \times Lead$

$$
\begin{aligned}
E[Inf \mid Lead = 1, pH] &= (\beta_0 + \beta_2 \cdot pH) + (\beta_1 + \beta_3 \cdot pH) \\
E[Inf \mid Lead = 0, pH] &= (\beta_0 + \beta_2 \cdot pH) \\
\rightarrow \Delta_{pH \text{ fixed}}^{Lead-NoLead} &= \qquad\qquad\qquad\qquad (\beta_1 + \beta_3 \cdot pH)
\end{aligned}
$$

$\Rightarrow \beta_1$ and β_3 measure the effect of `Lead` on `Inf`. Comparing 2 cities, one with lead pipes ($Lead = 1$) and one without lead pipes ($Lead = 0$), but the same of `pH`, the difference in infant mortality rate on average is $\beta_1 + \beta_3 \cdot pH$.

(b1) Understand what coefficients measure.

From (3): $E[Inf \mid Lead, pH] = (\beta_0 + \beta_1 \cdot Lead) + (\beta_2 + \beta_3 \cdot Lead) \times pH$

$$
\begin{aligned}
E[Inf \mid pH = c+1, Lead] &= (\beta_0 + \beta_1 \cdot Lead) + (\beta_2 + \beta_3 \cdot Lead) \times (c+1) \\
E[Inf \mid pH = c, Lead] &= (\beta_0 + \beta_1 \cdot Lead) + (\beta_2 + \beta_3 \cdot Lead) \times c \\
\to \Delta^{\text{increase pH by 1}}_{Lead \text{ fixed}} &= \qquad\qquad\qquad\qquad (\beta_2 + \beta_3 \cdot Lead)
\end{aligned}
$$

$\Rightarrow$ β_2 and β_3 measure the effect of `pH` on `Inf`. Comparing 2 cities, with 1 unit differential in `pH`, but the same of `Lead`, the difference in infant mortality rate on average is $\beta_2 + \beta_3 \cdot Lead$.

(b1) Run the regression and Interpret.

OLS estimation results (»stata)

$$\underset{(se)}{\widehat{Inf}} = \underset{(0.150)}{0.919} + \underset{(0.208)}{0.462} \cdot Lead - \underset{(0.021)}{0.075} \cdot pH - \underset{(0.028)}{0.057} \cdot Lead \times pH$$

(b1) Run the regression and Interpret.

OLS estimation results (»stata)

$$\underset{(se)}{\widehat{Inf}} = \underset{(0.150)}{0.919} + \underset{(0.208)}{0.462} \cdot Lead - \underset{(0.021)}{0.075} \cdot pH - \underset{(0.028)}{0.057} \cdot Lead \times pH$$

- $\hat{\beta}_0 = 0.919$ shows the level of `Inf` when $Lead = 0$ and $pH = 0$. It dictates the level of the regression line.

(b1) Run the regression and Interpret.

OLS estimation results (»stata)

$$\underset{(se)}{\widehat{Inf}} = \underset{(0.150)}{0.919} + \underset{(0.208)}{0.462} \cdot Lead - \underset{(0.021)}{0.075} \cdot pH - \underset{(0.028)}{0.057} \cdot Lead \times pH$$

- $\hat{\beta}_0 = 0.919$ shows the level of `Inf` when $Lead = 0$ and $pH = 0$. It dictates the level of the regression line.

- Comparing 2 cities, one with lead pipes $Lead = 0$ and one without lead pipes $Lead = 0$, but the same of pH, the difference in predicted infant mortality rate is

$$(2') \quad \widehat{Inf}(Lead = 1, pH) - \widehat{Inf}(Lead = 0, pH) = 0.462 - 0.057 \cdot pH$$

(b1) Run the regression and Interpret.

OLS estimation results (»stata)

$$\underset{(se)}{\widehat{Inf}} = \underset{(0.150)}{0.919} + \underset{(0.208)}{0.462} \cdot Lead - \underset{(0.021)}{0.075} \cdot pH - \underset{(0.028)}{0.057} \cdot Lead \times pH$$

- ◇ $\hat{\beta}_0 = 0.919$ shows the level of `Inf` when $Lead = 0$ and $pH = 0$. It dictates the level of the regression line.

- ◇ Comparing 2 cities, one with lead pipes $Lead = 0$ and one without lead pipes $Lead = 0$, but the same of pH, the difference in predicted infant mortality rate is

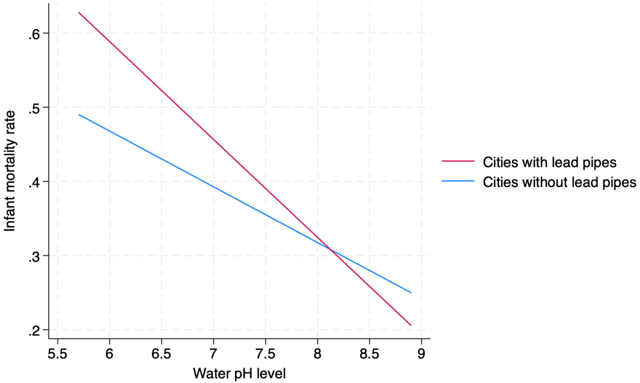
$$(2') \quad \widehat{Inf}(Lead = 1, pH) - \widehat{Inf}(Lead = 0, pH) = 0.462 - 0.057 \cdot pH$$

- ◇ Comparing 2 cities, one with $pH = 6$ and one with with $pH = 6$, but the same of `Lead`, the difference in predicted infant mortality rate is

$$(3') \quad \widehat{Inf}(pH = 6, Lead) - \widehat{Inf}(pH = 5, Lead) = -0.075 - 0.057 \cdot Lead$$

$\Rightarrow$ so the difference is -0.075 for cities without lead pipes and -0.132 for cities with lead pipes.

(b2) Plot the estimated regression function relating `Inf` to `pH` for *Lead* = 0 and for *Lead* = 1.



⇒ The infant mortality rate is higher for cities with lead pipes, but the difference declines as the pH level increases. (»stata)

(b2) The difference in infant mortality rates between cities with lead pipes and cites without lead pipes

- At the 10^{th} percentile of pH (6.4) is

$$\widehat{Inf}(Lead = 1, pH = 6.4) - \widehat{Inf}(Lead = 0, pH = 6.4) = 0.462 - 0.057 \times 6.4 \approx 0.097$$

- At the 50^{th} percentile of pH (7.5) is

$$\widehat{Inf}(Lead = 1, pH = 7.5) - \widehat{Inf}(Lead = 0, pH = 7.5) = 0.462 - 0.057 \times 7.5 \approx 0.035$$

- At the 90^{th} percentile of pH (8.2) is

$$\widehat{Inf}(Lead = 1, pH = 8.2) - \widehat{Inf}(Lead = 0, pH = 8.2) = 0.462 - 0.057 \times 8.2 \approx 0.005$$

- *Note: Refer to equation (2') in (b1).*

(b3) Does `Lead` have a statistically significant effect on `Inf`? Explain.

- Null Hypothesis: $H_0 : \beta_1 = \beta_3 = 0$ (»stata)
- The F-statistic for the coefficient on `Lead` and the interaction term is $F = 3.94$, which has a p-value of 0.02, so the coefficients are jointly statistically significantly different from zero at the 5% but not the 1% significance level.

(b4) Does the effect of `Lead` on `Inf` depend on `pH`? Is this dependence statistically significant?

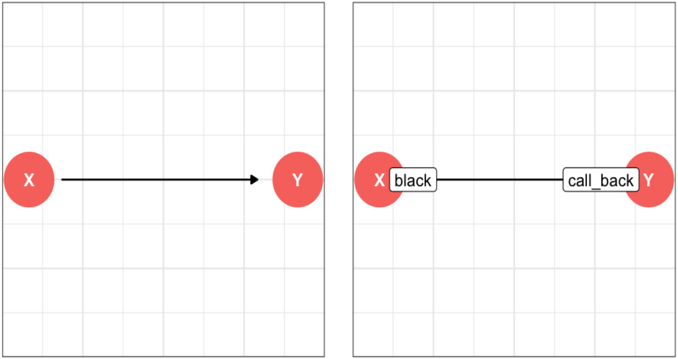
- Null Hypothesis: $H_0 : \beta_3 = 0$ (»stata)

- The interaction term has a t statistic of $t = -2.02$, so the coefficient is significant at the 5% but not the 1% significance level.

BRIEF REVIEW

Causal Graph (I)

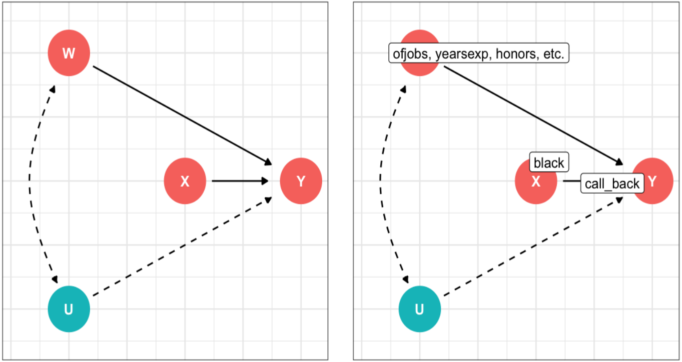
Randomized Experiment



(»back)

Causal Graph (II)

Randomized Experiment with additional characteristics



(»back)

STATA CODES & RESULTS

Exercise 2(a)

```
. regress call_back black, vce(robust)

Linear regression                               Number of obs     =      4,870
                                                F(1, 4868)        =      16.93
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0035
                                                Root MSE          =     .27164
```

call_back	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
black	-.0320329	.007785	-4.11	0.000	-.0472949	-.0167708
_cons	.0965092	.0059853	16.12	0.000	.0847753	.1082431

Exercise 2(a)

```
* ttest yvar, by(groupvar) unequal
// Test if mean(yvar) equal between 2 groups defined by groupvar
// add option 'unequal' to assume unequal variances
```

```
. ttest call_back, by(black) unequal

Two-sample t test with unequal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
0	2,435	.0965092	.0059853	.295349	.0847724	.1082461
1	2,435	.0644764	.0049781	.2456501	.0547145	.0742382
Combined	4,870	.0804928	.0038988	.2720826	.0728493	.0881363
diff		.0320329	.007785		.0167707	.047295

```
    diff = mean(0) - mean(1)                                      t =   4.1147
H0: diff = 0                     Satterthwaite's degrees of freedom =   4711.6

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Exercise 2(c1)

```
. *Regression using only the high quality resume as a regressor
. regress call_back high, vce(robust)

Linear regression                               Number of obs     =      4,870
                                                F(1, 4868)        =       3.25
                                                Prob > F          =     0.0713
                                                R-squared         =     0.0007
                                                Root MSE          =     .27202

------------------------------------------------------------------------------
             |               Robust
   call_back | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        high |   .0140574   .0077932     1.80   0.071    -.0012207    .0293356
       _cons |   .0734323   .0052991    13.86   0.000     .0630436     .083821
------------------------------------------------------------------------------
```

Exercise 2(c2)

```
. *Generating an interaction term for having a high quality resume and being black
. gen h_b = high*black

. *Regression including the interaction term
. regress call_back black high h_b, vce(robust)

Linear regression                               Number of obs     =      4,870
                                                F(3, 4866)        =       6.61
                                                Prob > F          =     0.0002
                                                R-squared         =     0.0044
                                                Root MSE          =     .27157
```

call_back	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
black	-.0231023	.0105901	-2.18	0.029	-.0438636	-.002341
high	.0229478	.0119584	1.92	0.055	-.000496	.0463917
h_b	-.0177808	.0155605	-1.14	0.253	-.0482864	.0127248
_cons	.0849835	.0080133	10.61	0.000	.0692739	.1006931

Exercise 1(a)

```
. ttest infrate, by(lead) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       0 |      55    .3811679    .0199238    .1477588     .341223    .4211127
       1 |     117    .4032576    .0141529    .1530873    .3752259    .4312892
---------+--------------------------------------------------------------------
Combined |     172     .396194    .0115384    .1513249    .3734179    .4189701
---------+--------------------------------------------------------------------
    diff |           -.0220897     .024439               -.0705255    .0263461
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -0.9039
H0: diff = 0                             Satterthwaite's degrees of freedom =  109.292

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1840         Pr(|T| > |t|) = 0.3681          Pr(T > t) = 0.8160
```

Exercise 1(b1)

```
. *Generating interaction of lead exposure with ph
. gen lead_ph = lead*ph

. *Regression of infant mortality on lead exposure, ph, and their interaction
. regress infrate lead ph lead_ph, vce(robust)

Linear regression                               Number of obs     =        172
                                                F(3, 168)         =      20.97
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2719
                                                Root MSE          =     .13027
```

infrate	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
lead	.4617985	.2076136	2.22	0.027	.0519309	.8716661
ph	-.0751792	.0209532	-3.59	0.000	-.1165447	-.0338136
lead_ph	-.0568622	.0280837	-2.02	0.044	-.1123047	-.0014197
_cons	.9189038	.1504941	6.11	0.000	.6218005	1.216007

Exercise 1(b2)

```
. quiet regress infrate lead ph lead_ph, vce(robust)

. predict inf_hat
(option xb assumed; fitted values)

. separate inf_hat, by(lead)

Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------------------------
inf_hat0        float   %9.0g                 inf_hat, lead == 0
inf_hat1        float   %9.0g                 inf_hat, lead == 1

. line inf_hat0 inf_hat1 ph, sort ytitle("Infant mortality rate")    ///
>                                                       xtitle("Water pH level")           ///
>                                 legend(col(1) order(2 "Cities with lead pipes" 1 "Cities without lead pipes")) ///
>                                                       xscale(range(5.5 9)) xlabel(5.5(0.5)9)
```

Exercise 1(b3)

```
. *Tesing the joint-significance of the coefficients of lead exposure and its interaction with ph
. test lead lead_ph

 ( 1)  lead = 0
 ( 2)  lead_ph = 0

       F(  2,   168) =    3.94
            Prob > F =    0.0214
```