Coefficient Plot (Optional)

(Jing Li, Miami University)

1. This note introduces coefficient plot that visualizes point estimate and its uncertainity measured by standard error. Data is cps92_12.xlsx [1] and the first six observations are

```
> setwd("/Users/lij14/Dropbox/teaching/311r")
> library("readxl")
> data = read_excel("cps92_12.xlsx")
> head(data)
   year   ahe bachelor female   age
  <dbl> <dbl>    <dbl>  <dbl> <dbl>
1  1992 11.2          1      0    29
2  1992 10            1      0    33
3  1992  5.77         0      0    30
4  1992  1.56         0      0    32
5  1992 15.0          1      0    31
6  1992  8.66         1      1    26
> data$rahe = data$ahe
> data$rahe[data$year==1992] = data$ahe[data$year==1992]/140.4*229.6
```

   where year is either 1992 or 2012; ahe is average hourly earning (nominal wage); bachelor is a dummy variable equal to one for a person with bachelor diploma; female is a dummy variable equal to one for female worker. We generate the real wage (rahe) in 2012 dollars by adjusting the nominal wage with CPI in 1992 and 2012.

2. For teaching purpose, I create two dummy variables called irrelevantA and irrelevantB that are independent of real wage. Note that the latter has less variation than the former.

```
> set.seed(12345)
> data$irrelevantA = rbinom(nrow(data), size = 1, prob = 0.5)
> data$irrelevantB = rbinom(nrow(data), size = 1, prob = 0.9)
> var(data$irrelevantA)
[1] 0.2500159
```

[1] the data is downloaded from the webpage for Student Resources for Stock and Watson's Introduction to Econometrics, 3rd Updated Edition.

```
> var(data$irrelevantB)
[1] 0.08988902
> attach(data)
```

3. Then I run six bivariate regressions

$$rahe = \beta_0 + \beta_1 female + u \tag{1}$$

$$rahe = \beta_0 + \beta_1 bachelor + u \tag{2}$$

$$rahe = \beta_0 + \beta_1 D_{2012} + u \tag{3}$$

$$rahe = \beta_0 + \beta_1 irrelevantA + u \tag{4}$$

$$rahe = \beta_0 + \beta_1 irrelevantA + u \qquad (i = 1, 2, ...50) \tag{5}$$

$$rahe = \beta_0 + \beta_1 irrelevantB + u \tag{6}$$

where D_{2012} in regression (3) is a dummy variable equal to one for 2012, and regression (5) only uses 50 observations.

4. Because the regressor in each regression is dummy variable, the slope coefficient β_1 in each regression measures a wage gap between two groups determined by the dummy variable. For instance, β_1 in model (1) measures gender wage gap (female minus male), and in model (2) it measures wage gap between people with and without bachelor degree. The results of first three regressions are

```
> summary(lm(rahe~female))$coef
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 20.57177  0.1061298 193.83598 0.000000e+00
female      -2.76905  0.1627461 -17.01455 2.556913e-64
> summary(lm(rahe~bachelor))$coef
              Estimate Std. Error   t value Pr(>|t|)
(Intercept) 16.039501  0.1028647 155.92814        0
bachelor     7.300134  0.1517416  48.10899        0
> summary(lm(rahe~as.factor(year)))$coef
                       Estimate Std. Error    t value     Pr(>|t|)
(Intercept)          18.9973298  0.1141315 166.451200 0.000000e+00
as.factor(year)2012   0.8029283  0.1623365   4.946073 7.654268e-07
```

We see the slope coefficients (wage gaps) are all statistically significant (t value = -17.01455, 48.10899, 4.946073), and the slope coefficients indicate that on average female earns 2.76905 less than male, people with bachelor degree earn 7.300134 more than without the degree, and real wage in 2012 is more than 1992 by 0.8029283.

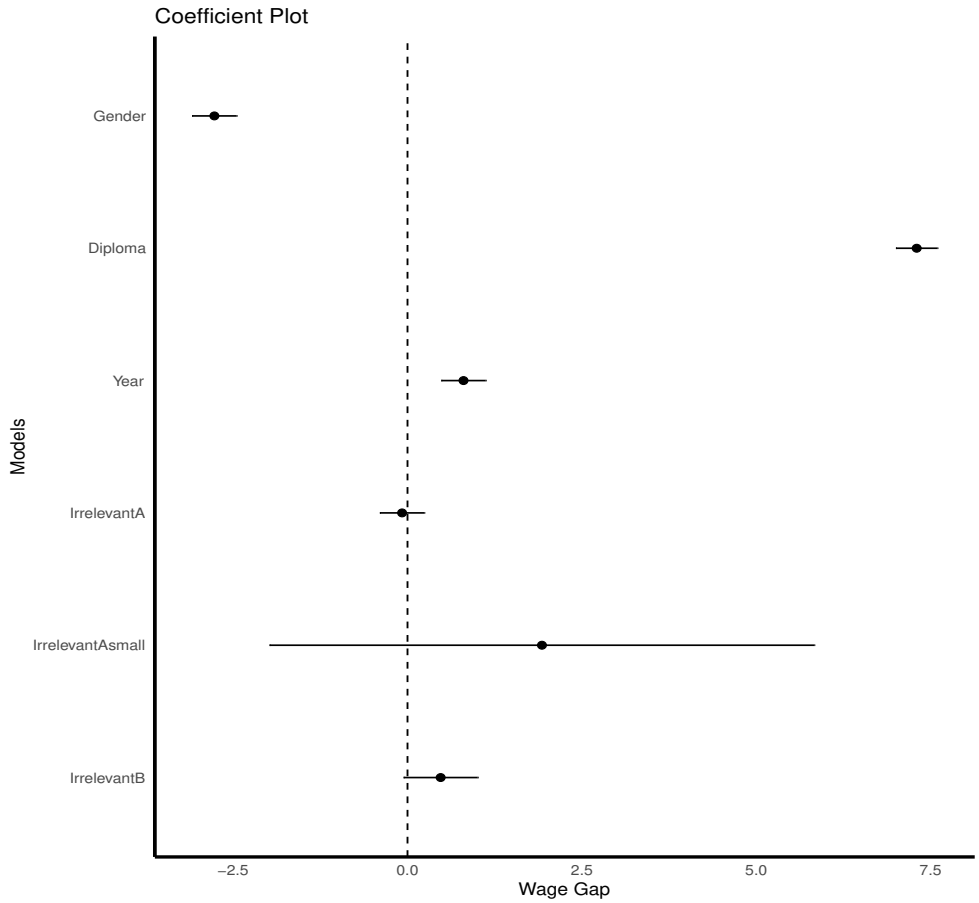
5. By contrast, the slope coefficients are insignificant in regression (4), (5), (6) shown below, which agrees with the fact that they use a regressor that has no bearing on wage

```
> summary(lm(rahe~irrelevantA))$coef
               Estimate Std. Error     t value  Pr(>|t|)
(Intercept) 19.43303942  0.1149735 169.0219105 0.0000000
irrelevantA -0.07753204  0.1624568  -0.4772471 0.6331931
> summary(lm(rahe[1:50]~irrelevantA[1:50]))$coef
                   Estimate Std. Error    t value     Pr(>|t|)
(Intercept)       15.628475   1.351220 11.5661991 1.751576e-15
irrelevantA[1:50]  1.927509   1.992265  0.9674963 3.381458e-01
> summary(lm(rahe~irrelevantB))$coef
              Estimate Std. Error   t value   Pr(>|t|)
(Intercept) 18.9662037  0.2570300 73.789851 0.00000000
irrelevantB  0.4754814  0.2709114  1.755118 0.07925949
```

Let us focus on the standard error of β_1. We find se in regression (5) is greater than (4) ($1.99226 > 0.1624568$) because it uses a small sample of only 50 observations. Also se in regression (6) is greater than (4) ($0.2709114 > 0.1624568$) because irrelevantB has smaller variance than irrelevantA. In fact those findings are implied by the formula

$$se(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)} = \sqrt{\frac{\sigma^2}{(n-1)S_{x1}^2(1-R_{1ststep}^2)}} \tag{7}$$

6. In the past researchers used to summarize the regression results in a table in which each column represents a regression. Recently it becomes increasingly popular to use coefficient plot to visualize and compare regression output. For instance, the coefficient plot below compares the wage gap (slope coefficient) in the six regressions



More specifically, the coefficient plot displays the 95 percent confidence interval (CI) of the slope coefficient—the solid dot in the middle is the point estimate $\hat{\beta}_1$, and the width of the line indicates the range of confidence interval. A vertical line that passes 0 is drawn for easy comparison.

7. Mathematically, the 95 percent CI is computed as

$$95\%\ confidence\ interval = (\hat{\beta}_1 - 1.96 * se,\ \ \hat{\beta}_1 + 1.96 * se) \tag{8}$$

where the standard error (se) is computed with formula (7). Recall two important facts about CI

   (a) A point estimate is statistically significant (rejecting zero value) if 0 is outside CI

   (b) A point estimate is precise (with small se) if CI is narrow

In light of those two facts, the coefficient plot above tells us that

(a) Wage gap associated with gender, bachelor diploma, and year are statistically significant. All are precisely estimated thanks to their narrow CI.

(b) Wage gap in regression (4), (5), (6) are all insignificant (they should be!) because their CI contains value 0. Regression (4) is more precise than (5) and (6) by having narrower CI.

Below are the codes to draw the coefficient plot

```
model1 = lm(rahe~female)
model2 = lm(rahe~bachelor)
model3 = lm(rahe~as.factor(year))
model4 = lm(rahe~irrelevantA)
model5 = lm(rahe[1:50]~irrelevantA[1:50])
model6 = lm(rahe~irrelevantB)

output <- data.frame(
    model = c("Gender", "Diploma", "Year", "IrrelevantA","IrrelevantAsmall","Irrele
    beta1 = c(coef(model1)[2], coef(model2)[2], coef(model3)[2], coef(model4)[2], c
    se = c(summary(model1)$coefficients[2, "Std. Error"],
            summary(model2)$coefficients[2, "Std. Error"],
            summary(model3)$coefficients[2, "Std. Error"],
            summary(model4)$coefficients[2, "Std. Error"],
            summary(model5)$coefficients[2, "Std. Error"],
            summary(model6)$coefficients[2, "Std. Error"])
)

output$min = output$beta1 - 1.96*output$se
output$max = output$beta1 + 1.96*output$se
output$model = factor(output$model,levels=c("IrrelevantB", "IrrelevantAsmall", "Irr

library(ggplot2)
ggplot(output, aes(x = beta1, y = model)) +
    geom_point(size = 2) +
    geom_errorbarh(aes(xmin = min, xmax = max), height = 0) +
    geom_vline(xintercept = 0,linetype = "dashed") +
```

```
    labs(title = "Coefficient Plot",
         x = "Wage Gap",
         y = "Models") +
    theme_minimal() +
    theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.line = element_line(size = 1, color = "black")
  )
```

You may start learning ggplot with this webpage

https://cran.r-project.org/web/packages/ggplot2/vignettes/ggplot2.html

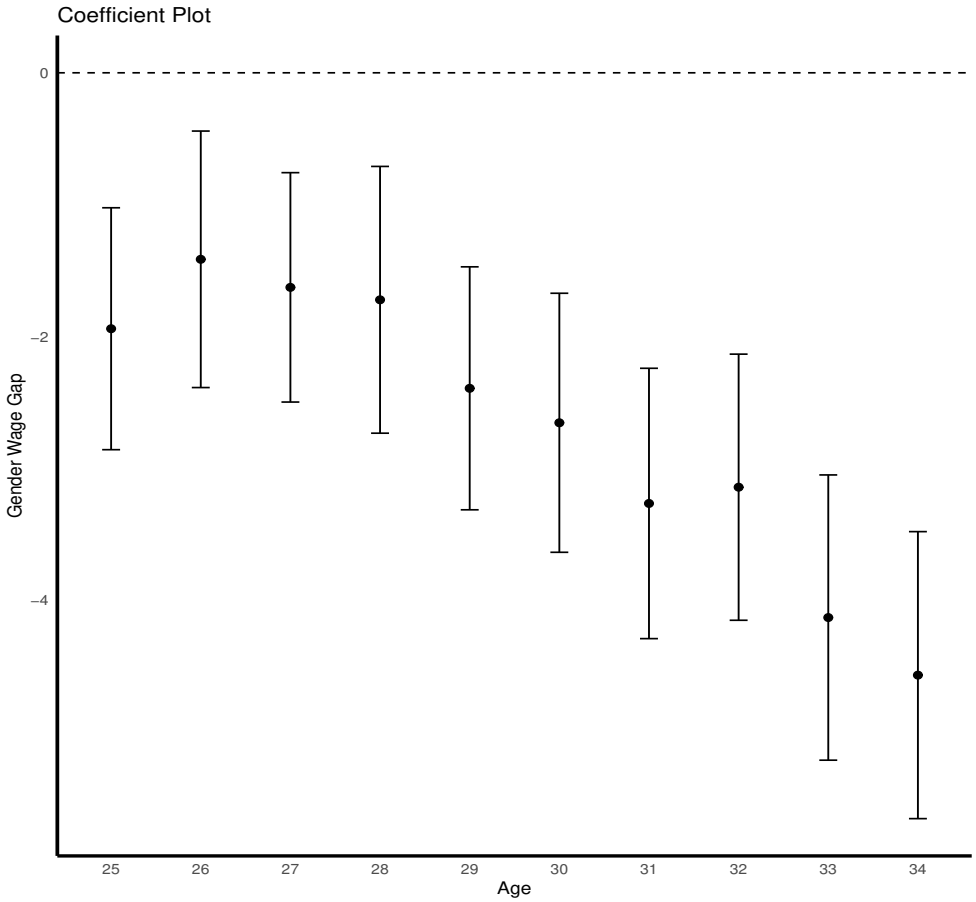
In my computer, I sometime need to run the ggplot block twice to generate the graph.

8. Other than a "portrait" coefficient plot, we can rotate and create a "landscape" plot. For instance, codes below compares the gender wage gap at each age

```
sort(unique(age))
 [1] 25 26 27 28 29 30 31 32 33 34

i = 1; beta1 = NULL; se = NULL
while (i<=length(unique(age))) {
        m = lm(rahe~female,data=data[age==sort(unique(age))[i],])
        beta1[i] = coef(m)[2]
        se[i] = summary(m)$coefficients[2, "Std. Error"]
        i = i + 1
}

output = data.frame(model = sort(unique(age)),beta1,se)
output$min = output$beta1 - 1.96*output$se
output$max = output$beta1 + 1.96*output$se
output$model = factor(output$model,levels=c("25","26","27","28","29","30","31","32"

ggplot(output, aes(x = model, y = beta1)) +
```

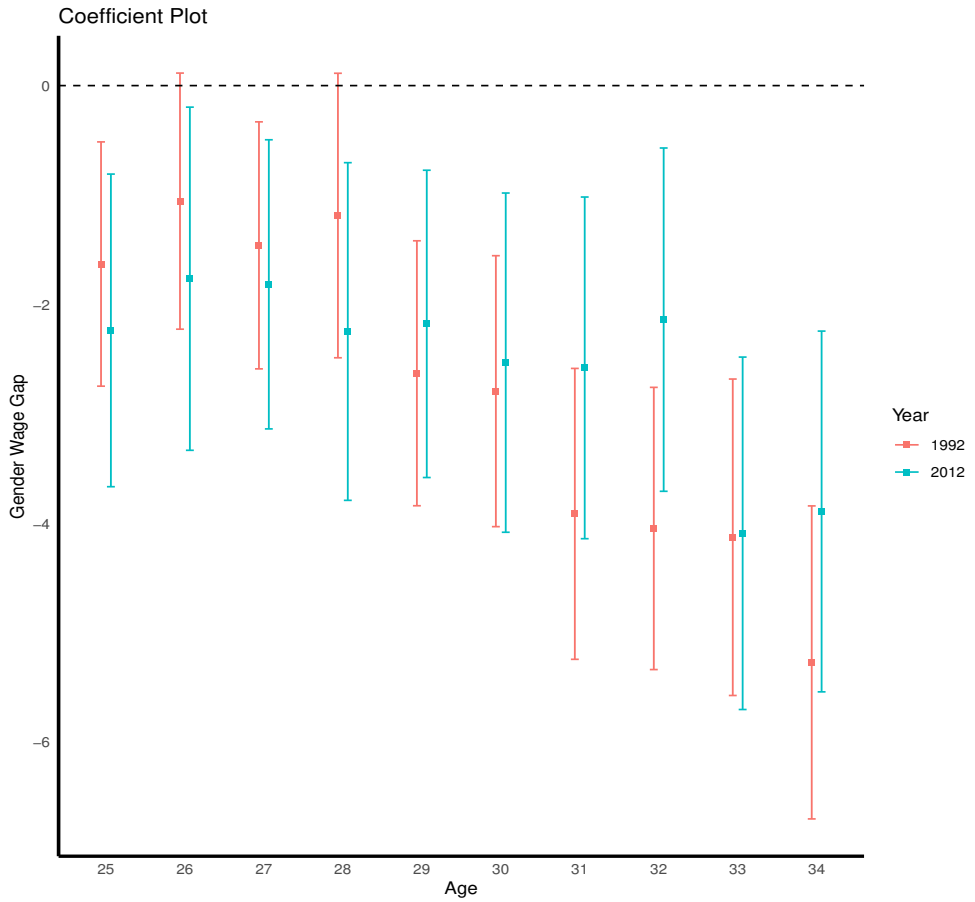
```
    geom_point(size = 2) +
    geom_errorbar(aes(ymin = min, ymax = max), width = 0.2) +
    geom_hline(yintercept = 0,linetype = "dashed") +
    labs(title = "Coefficient Plot",
         x = "Age",
         y = "Gender Wage Gap") +
    theme_minimal() +
    theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.line = element_line(size = 1, color = "black")
  )
```

The coefficient plot looks like



We see all the wage gaps are significant (excluding 0). It is clear that the wage gap gets bigger as age rises (there is a downward trend connecting those solid dots).

9. The graph above does not distinguish 1992 and 2012. Alternatively, we can redraw the coefficient plot to compare gender wage gap at each age across two years. The new plot is



The codes are below

```
# Two plots together
i = 1; beta1 = NULL; se = NULL
while (i<=length(unique(age))){
        m= lm(rahe~female,data=data[age==sort(unique(age))[i]&year==1992,])
        beta1[i] = coef(m)[2]
        se[i] = summary(m)$coefficients[2, "Std. Error"]
        i = i + 1
}
output = data.frame(model = sort(unique(age)),beta1,se)
output$min = output$beta1 - 1.96*output$se
output$max = output$beta1 + 1.96*output$se
```

```
output$model = factor(output$model,levels=c("25","26","27","28","29","30","31","32"
output$Year = "1992"

i = 1; beta1 = NULL; se = NULL
while (i<=length(unique(age))) {
        m= lm(rahe~female,data=data[age==sort(unique(age))[i]&year==2012,])
        beta1[i] = coef(m)[2]
        se[i] = summary(m)$coefficients[2, "Std. Error"]
        i = i + 1
}
output2 = data.frame(model = sort(unique(age)),beta1,se)
output2$min = output2$beta1 - 1.96*output2$se
output2$max = output2$beta1 + 1.96*output2$se
output2$model = factor(output2$model,levels=c("25","26","27","28","29","30","31","3
output2$Year = "2012"

comb = rbind(output,output2)
ggplot(comb, aes(x = model, y = beta1)) +
    geom_point(aes(x=model, y=beta1, colour = Year),
              shape=15, position = position_dodge(width = 0.25)) +
    geom_errorbar(aes(ymin = min, ymax = max, colour = Year), width = 0.2, position
    geom_hline(yintercept = 0,linetype = "dashed") +
    labs(title = "Coefficient Plot",
         x = "Age",
         y = "Gender Wage Gap") +
    theme_minimal() +
    theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.line = element_line(size = 1, color = "black")
  )
```

Can you guess why the 1992 gender wage gaps at ages 26 and 28 are insignificant?
Hint: see formula (7)

10. (Exercise) please use R to draw this coefficient plot