SC472 Econometrics Data analysis Using R **Date** – April 8, 2023

Assignment - 1

Name – Divya Kirtikumar Patel

Student ID - 202001420

3 The following table contains the *ACT* scores and the *GPA* (grade point average) for eight college students. Grade point average is based on a four-point scale and has been rounded to one digit after the decimal.

Student	*GPA*	*ACT*
1	2.8	21
2	3.4	24
3	3.0	26
4	3.5	27
5	3.6	29
6	3.0	25
7	2.7	25
8	3.7	30

(i) Estimate the relationship between *GPA* and *ACT* using OLS; that is, obtain the intercept and slope estimates in the equation

$$\widehat{GPA} = \hat{\beta}_0 + \hat{\beta}_1 ACT.$$

Comment on the direction of the relationship. Does the intercept have a useful interpretation here? Explain. How much higher is the *GPA* predicted to be if the *ACT* score is increased by five points?

3 Let y_i be GPA_i and $x_i = ACT_i$. Here $n = 8$

Here $\Sigma x_i = 21 + 24 + 26 + 27 + 25 + 29 + 25 + 30$

$$\bar{x} = \frac{\Sigma x_i}{n}$$

$$\bar{x} = \frac{207}{8}$$

$$\bar{x} = 25.875$$

Similarly $\bar{y} = \frac{27.5}{8} = 3.2125$

Now $\Sigma (x - x_i)(y - y_i) = \Sigma (x_i - 25.875)(y_i - 3.2125)$

$= 5.8125$

Also, $\Sigma (x_i - \bar{x})^2 = 56.875$

$$\hat{\beta}_1 = \frac{cov(x,y)}{var(x)} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{5.8125}{56.875} = 0.1022$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= (3.2125) - (0.1022)(25.875)$$

$$\hat{\beta}_0 = 0.5681$$

$$\widehat{GPA} = 0.5681 + 0.1022 ACT$$

The intercept does not have useful interpretation as ACT is not near 0 for population in question.

Increase in GPA if ACT is increased by 5:-

$$\Delta GPA = \beta_1 \Delta ACT$$

$$\Delta GPA = (0.1022)(5) = 0.511$$

(ii) Compute the fitted values and residuals for each observation, and verify that the residuals (approximately) sum to zero.

ii) Fitted Value:- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Residual $u_i = y_i - \hat{y}_i$

No.	GPA	$\widehat{GPA}$	$\hat{u}_i$
1	2.8	2.7143	0.0857
2	3.4	3.0209	0.3791
3	3.0	3.2253	-0.2253
4	3.5	3.3275	0.1725
5	3.6	3.5319	0.0681
6	3.0	3.1231	-0.1231
7	2.7	3.1231	-0.4321
8	3.7	3.6341	0.0659

$$\Sigma \hat{u}_i = -0.002 \approx 0$$

(iii) What is the predicted value of *GPA* when $ACT = 20$?

iii) $ACT = 20$

$$\widehat{GPA} = \hat{\beta}_0 + \hat{\beta}_1 ACT$$

$$\widehat{GPA} = 0.5681 + 0.1022(20)$$

$$\boxed{\widehat{GPA} = 2.61}$$

(iv) How much of the variation in *GPA* for these eight students is explained by *ACT*? Explain.

iv) Co-efficient of Determination $-R^2$:-

$$R^2 = 1 - \frac{SSE}{SST}$$

$$SSE = \sum \hat{u}_i^2 = 0.4347$$

$$SST = \sum (y_i - \bar{y})^2 = 1.0288$$

$$R^2 = 1 - \frac{0.4347}{1.0288} \approx 0.577$$

$\therefore$ 57.7% Variation in GPA is explained by ACT.

4 The data set BWGHT.RAW contains data on births to women in the United States. Two variables of interest are the dependent variable, infant birth weight in ounces (*bwght*), and an explanatory variable, average number of cigarettes the mother smoked per day during pregnancy (*cigs*). The following simple regression was estimated using data on $n = 1{,}388$ births:

$$\widehat{bwght} = 119.77 - 0.514\, cigs$$

(i) What is the predicted birth weight when $cigs = 0$? What about when $cigs = 20$ (one pack per day)? Comment on the difference.

(i) a/ bweight = 119.77 ounces [cigs = 0]

b/ When cigs = 20

bweight = 119.77 − 0.514 (20)

bweight = 109.49 ounces

Therefore when one pack of cigarette is smoked on average daily, there is difference of 8.6 % of baby's weight.

(ii) Does this simple regression necessarily capture a causal relationship between the child's birth weight and the mother's smoking habits? Explain.

ii) To some extent yes but there are other factors such as genes, quality of pre-natal care, etc..

(iii) To predict a birth weight of 125 ounces, what would *cigs* have to be? Comment.

iii) To predict for 125 ounces.

125 = 119.77 − 0.514 cigs

⇒ $\boxed{cigs \approx -10}$

↳ This makes no sense. This happened because we are modelling on single variable and childbirth is much more complex variable. Also we have small dataset as seen by equation that states maximum bweight can be 119.77 ounces

(iv) The proportion of women in the sample who do not smoke while pregnant is about .85. Does this help reconcile your finding from part (iii)?

iv) If we are provided data regarding smoker pregnant women, a better model can be prepared which will be more robust.

5 In the linear consumption function

$$\widehat{cons} = \hat{\beta}_0 + \hat{\beta}_1 inc,$$

the (estimated) *marginal propensity to consume* (MPC) out of income is simply the slope, $\hat{\beta}_1$, while the *average propensity to consume* (APC) is $\widehat{cons}/inc = \hat{\beta}_0/inc + \hat{\beta}_1$. Using observations for 100 families on annual income and consumption (both measured in dollars), the following equation is obtained:

$$\widehat{cons} = -124.84 + 0.853\ inc$$
$$n = 100,\ R^2 = 0.692.$$

(i) Interpret the intercept in this equation, and comment on its sign and magnitude.

(5)

(i) This consumption function might not be a very good predictor for lower income levels. If we say that income is \$0 then the consumption is \$ -124.48. where as consumption is never negative.

(ii) What is the predicted consumption when family income is \$30,000?

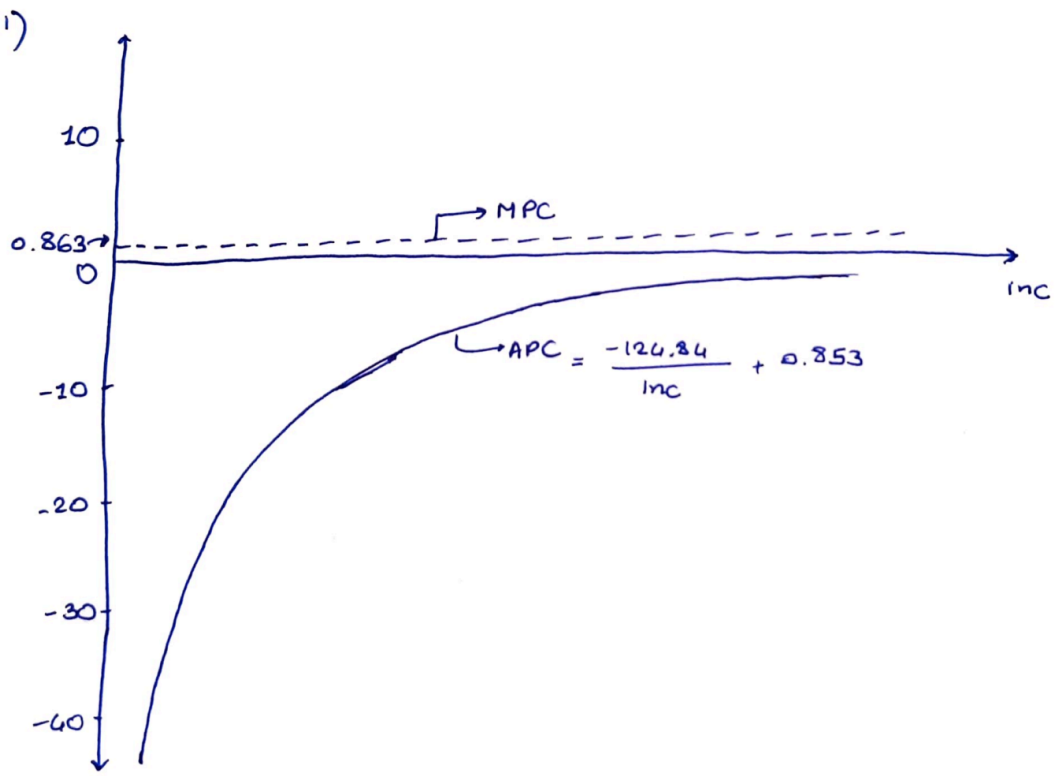
(ii) inc = \$30,000

$\widehat{cons} = -124.84 + 0.853(30,000) = 25465.1$

$\widehat{cons} = \$25465.16$

(iii) With *inc* on the *x*-axis, draw a graph of the estimated MPC and APC.

(iii)



8 Consider the standard simple regression model $y = \beta_0 + \beta_1 x + u$ under the Gauss-Markov Assumptions SLR.1 through SLR.5. The usual OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for their respective population parameters. Let $\tilde{\beta}_1$ be the estimator of β_1 obtained by assuming the intercept is zero (see Section 2.6).

(i) Find $E(\tilde{\beta}_1)$ in terms of the x_i, β_0, and β_1. Verify that $\tilde{\beta}_1$ is unbiased for β_1 when the population intercept (β_0) is zero. Are there other cases where $\tilde{\beta}_1$ is unbiased?

(i)

$$\tilde{y} = \tilde{\beta}_1 x$$

To obtain best slope :- we minimize sum of Residuals (squared sum)

$$\min\left(\Sigma (y_i - \tilde{\beta}_1 x_i)^2\right)$$

$$W = \Sigma (y_i - \tilde{\beta}_1 x_i)^2$$

$$\frac{dw}{d\beta_1} = (-2) \Sigma x_i (y_i - \tilde{\beta}_1 x_i) = 0$$

or $$\Sigma x_i (y_i - \tilde{\beta}_1 x_i) = 0$$

$$\tilde{\beta}_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

i) $y_i = \beta_0 + \beta_1 x_i + u_i$

$$\tilde{\beta}_1 = \frac{\Sigma x_i (\beta_0 + \beta_1 x_i + u_i)}{\Sigma x_i^2}$$

$$\tilde{\beta}_1 = \frac{\beta_0 \Sigma x_i}{\Sigma x_i^2} + \beta_1 + \frac{\Sigma u_i x_i}{\Sigma x_i^2}$$

$$E(\tilde{\beta}_1) = \frac{\beta_0 \Sigma x_i}{\Sigma x_i^2} + \beta_1$$

Thus $\tilde{\beta}_1$ is unbiased when $\beta_0 = 0$. It is also unbiased when $\Sigma x_i = 0$

(ii) Find the variance of $\tilde{\beta}_1$. (*Hint:* The variance does not depend on β_0.)

ii) $\text{Var}(\tilde{\beta}_1)$

$$\text{Var}(\tilde{\beta}_1) = \text{Var}\left(\frac{\Sigma u_i x_i}{\Sigma x_i^2}\right)$$

$$= \text{Var}\left((\Sigma x_i^2)^{-1} \cdot (\Sigma u_i x_i)\right)$$

$$= [\Sigma x_i^2]^{-2} [\Sigma x_i^2 \, \text{Var}(u_i)]$$

Let $\text{Var}(u_i) = \sigma^2$

$$\text{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{\Sigma x_i^2}$$

(iii) Show that $\text{Var}(\tilde{\beta}_1) \leq \text{Var}(\hat{\beta}_1)$. [*Hint:* For any sample of data, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$, with strict inequality unless $\bar{x} = 0$.]

iii) Prove :- $\text{Var}(\tilde{\beta}_1) \leq \text{Var}(\hat{\beta}_1)$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma (x_i - \bar{x})^2}$$

$$\text{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{\Sigma x_i^2}$$

As $\Sigma x_i^2 \geq \Sigma (x_i - \bar{x})^2$

$$\frac{1}{\Sigma x_i^2} \leq \frac{1}{\Sigma (x_i - \bar{x})^2}$$

$$\therefore \frac{\sigma^2}{\Sigma x_i^2} \leq \frac{\sigma^2}{\Sigma (x_i - \bar{x})^2}$$

$$\text{Var}(\tilde{\beta}_1) \leq \text{Var}(\hat{\beta}_1)$$

(iv) Comment on the tradeoff between bias and variance when choosing between $\hat{\beta}_1$ and $\tilde{\beta}_1$.

iv) For the given sample size bias in $\tilde{\beta}_1$ increases as the mean increases. But the variance of $\hat{\beta}_1$ also increases. When β_0 is small, bias in $\tilde{\beta}_1$ is also small.

10 Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the OLS intercept and slope estimators, respectively, and let $\bar{u}$ be the sample average of the errors (not the residuals!).

(i) Show that $\hat{\beta}_1$ can be written as $\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i$ where $w_i = d_i/\text{SST}_x$ and $d_i = x_i - \bar{x}$.

(i) Show $\hat{\beta}_1 = \beta_1 + \Sigma w_i u_i$ where $w_i = \frac{d_i}{SST_x}$

$d_i = x_i - \bar{x}$

$$\hat{\beta}_1 = \frac{\text{cov}(x_i, y_i)}{\text{var}(x_i)}$$

$$= \frac{\Sigma (x_i - \bar{x})\, y_i}{\Sigma (x_i - \bar{x})^2}$$

$$= \frac{\Sigma (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\Sigma (x_i - \bar{x})^2}$$

Now $\Sigma (x_i - \bar{x})^2 = SST_x$

$$\hat{\beta}_1 = \frac{\Sigma (x_i - \bar{x})\beta_0 + \Sigma (x_i - \bar{x})\beta_1 x_i + \Sigma (x_i - \bar{x}) u_i}{SST_x}$$

We have $\Sigma (x_i - \bar{x}) = 0$

$$\hat{\beta}_1 = \frac{\beta_1 \Sigma (x_i - \bar{x}) x_i + \Sigma (x_i - \bar{x}) u_i}{SST_x}$$

Also $\Sigma (x_i - \bar{x})(x_i) = \Sigma (x_i - \bar{x})^2 = SST$

$\hat{\beta}_1 =$

$$\hat{\beta}_1 = \frac{\beta_1 \cdot SST_x + \Sigma d_i u_i}{SST_x}$$

$$\boxed{\hat{\beta}_1 = \beta_1 + \Sigma w_i u_i}$$

(ii) Use part (i), along with $\sum_{i=1}^{n} w_i = 0$, to show that $\hat{\beta}_1$ and $\bar{u}$ are uncorrelated. [*Hint*: You are being asked to show that $\mathrm{E}[(\hat{\beta}_1 - \beta_1) \cdot \bar{u}] = 0$.]

(ii) $\Sigma w_i = 0$

Prove $\hat{\beta}$ and $\bar{u}$ are uncorrelated

$$\text{Corr}(\hat{\beta}, \bar{u}) = \frac{E[(\hat{\beta}_1 - \beta_1)(\bar{u} - \bar{\bar{u}})]}{\sigma_{\hat{\beta}_1} \cdot \sigma_{\bar{u}}}$$

$\because \quad \bar{u} = \bar{\bar{u}}$

$$E[(\hat{\beta}_1 - \beta_1)(\bar{u} - \bar{\bar{u}})] = 0$$

$$\text{Corr}(\hat{\beta}, \bar{u}) = 0$$

(iii) Show that $\hat{\beta}_0$ can be written as $\hat{\beta}_0 = \beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1)\bar{x}$.

iii) $\hat{\beta}_0 = \beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1)\bar{x}$

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u} \quad \text{— ①}$$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \quad \text{— ②}$$

From ① & ②

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \beta_0 + \beta_1 \bar{x} + \bar{u}$$

$$\hat{\beta}_0 = \beta_0 + \bar{u} + (\beta_1 - \hat{\beta}_1)\bar{x}$$

(iv) Use parts (ii) and (iii) to show that $\text{Var}(\hat{\beta}_0) = \sigma^2/n + \sigma^2(\bar{x})^2/\text{SST}_x$.

iv) Prove :- $\text{var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2(\bar{x})^2}{SST_x}$

$\hat{\beta}_0 = \beta_0 + \bar{u} - \bar{x}(\hat{\beta}_1 - \beta_1)$

$\text{var}(\hat{\beta}_0) = \text{var}(\beta_0) + \text{var}(\bar{u}) + \text{var}(-\bar{x}(\hat{\beta}_1 - \beta_1))$

$= 0 + \text{var}(\bar{u}) + (-\bar{x})^2 \text{var}(\hat{\beta}_1 - \beta_1)$

$= \frac{\sigma^2}{n} + (\bar{x})^2 \frac{\sigma^2}{SST_x}$

$\therefore \text{var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + (\bar{x})^2 \frac{\sigma^2}{SST_x}$

(v) Do the algebra to simplify the expression in part (iv) to equation (2.58).
[*Hint*: $\text{SST}_x/n = n^{-1}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$.]

v) $SST_x = \Sigma(x_i - \bar{x})^2$

$SST_x = \Sigma(x_i)^2 - 2\Sigma(\bar{x}) \cdot x_i + \Sigma(\bar{x})^2$

$= \Sigma(x_i)^2 - 2\bar{x}\Sigma x_i + n(\bar{x})^2$

$= \Sigma(x_i)^2 - n(\bar{x})^2$

$\text{var}(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{(\bar{x})^2}{SST_x}\right]$

$\text{var}(\hat{\beta}_0) = \sigma^2\left[\frac{SST_x + n(\bar{x})^2}{n SST_x}\right]$

using value of SST_x from (1),

$\text{var}(\hat{\beta}_0) = \frac{\sigma^2 n^{-1}\Sigma x_i^2 - n\bar{x}^2 + n\bar{x}^2}{SST_x} = \frac{\sigma^2 \Sigma x_i^2}{n \cdot SST_x}$

$\text{var}(\hat{\beta}_0) = \frac{\sigma^2 n^{-1}\Sigma x_i^2}{\Sigma(x_i - \bar{x})^2}$

C1. The data in 401K.RAW are a subset of data analyzed by Papke (1995) to study the relationship between participation in a 401(k) pension plan and the generosity of the plan. The variable prate is the percentage of eligible workers with an active account; this is
the variable we would like to explain. The measure of generosity is the plan match rate, mrate. This variable gives the average amount the firm contributes to each worker's plan for each $1 contribution by the worker. For example, if mrate 5 0.50, then a $1
contribution by the worker is matched by a 50¢ contribution by the firm.

(i) Find the average participation rate and the average match rate in the sample of plans.

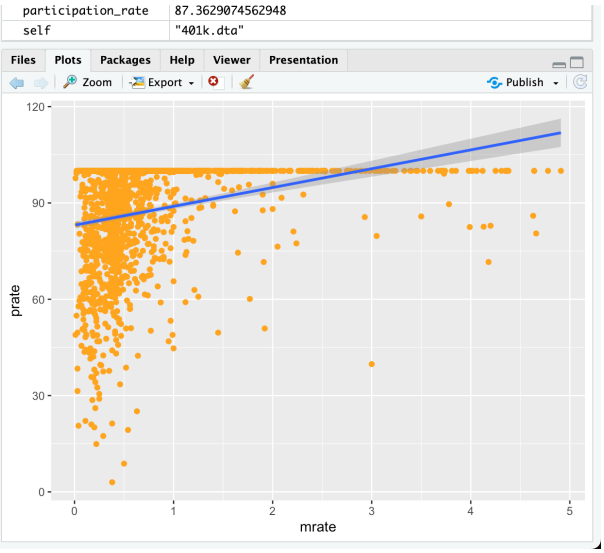
```
#C1
load("/Users/divya/Documents/Semester-6/Econometric Data Analysis with R/R data sets for 5e/401k.RData")
k401k_data <- k401k # Define the data set

# (i) Find the average participation rate and the average match rate in the sample of plans.
participation_rate <- mean(k401k_data$prate) # The average of participation rate
participation_rate
match_rate <- mean(k401k_data$mrate) # The average of match rate
match_rate
```

```
> # (i) Find the average participation rate and the average match rate in the sample of plans.
> participation_rate <- mean(k401k_data$prate) # The average of participation rate
> participation_rate
[1] 87.36291
> match_rate <- mean(k401k_data$mrate) # The average of match rate
> match_rate
[1] 0.7315124
>
```

(ii) Now, estimate the simple regression equation
$\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1 mrate.$
and report the results along with the sample size and R-squared.

```
# (ii) Now, estimate the simple regression equation
regression1 <- lm(prate ~ mrate, data = k401k_data) # Regression prate on mrate
ggplot(k401k_data, aes(x = mrate, y = prate)) +
  geom_point(color = 'orange') +
  geom_smooth(method = 'lm') +
  labs(x = 'mrate', y = 'prate') # A scatterplot and SRF for mrate & prate
summary(regression1) # The report for results (sample size and R-squared)
```

```
> summary(regression1) # The report for results (sample size and R-squared)

Call:
lm(formula = prate ~ mrate, data = k401k_data)

Residuals:
    Min      1Q  Median      3Q     Max
-82.303  -8.184   5.178  12.712  16.807

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.0755     0.5633  147.48   <2e-16 ***
mrate         5.8611     0.5270   11.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.09 on 1532 degrees of freedom
Multiple R-squared:  0.0747,    Adjusted R-squared:  0.0741
F-statistic: 123.7 on 1 and 1532 DF,  p-value: < 2.2e-16
```



(iii) Interpret the intercept in your equation. Interpret the coefficient on mrate.

The intercept implies that, even if mrate = 0, the predicted participation rate is 83.05 percent. The coefficient on mrate implies that a one-dollar increase in the match rate – a fairly large increase – is estimated to increase prate by 5.86 percentage points. This assumes, of course, that this change prate is possible (if, say, prate is already at 98, this interpretation makes no sense).

(iv) Find the predicted prate when mrate 5 3.5. Is this a reasonable prediction? Explain what is happening here.

If we plug mrate = 3.5 into the equation we get prate= 83.05 + 5.86(3.5) = 103.59. This is impossible, as we can have at most a 100 percent participation rate. This illustrates that, especially when dependent variables are bounded, a simple regression model can give strange predictions for extreme values of the independent variable. (In the sample of 1,534 firms, only 34 have mrate ≥ _3.5.)

```
> 83.0755 + 5.8611 * (3.5) # It isn't reasonable because of the characteristics of data 'prate'.
[1] 103.5894
>
```

(v) How much of the variation in prate is explained by mrate? Is this a lot in your opinion?

mrate explains about 7.47% of the variation in prate. This is not much, and suggests that many other factors influence 401(k) plan participation rates. This is a relatively low value and indicates that there is not a strong relationship between the two variables in the sample. In this case, it is important to consider other factors that may affect participation rates, such as age, income, education, job tenure, etc.

C2. The data set in CEOSAL2.RAW contains information on chief executive officers for U.S. corporations. The variable salary is annual compensation, in thousands of dollars, and ceoten is prior number of years as company CEO.

(i) Find the average salary and the average tenure in the sample.

```
50  # C2
51  load("/Users/divya/Documents/Semester-6/Econometric Data Analysis with R/R data sets for 5e/ceosal2.Rdata")
52  ceosal2_data <- ceosal2 # Define the data set
53
54  # (i)
55  mean(ceosal2_data$salary) # The average of salary
56  mean(ceosal2_data$ceoten) # The average of tenure
57
53:1  (Top Level)                                                    R Scr
```

```
Console  Terminal  Background Jobs
R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/
> ceosal2_data <- ceosal2 # Define the data set
> # (i)
> mean(ceosal2_data$salary) # The average of salary
[1] 865.8644
> mean(ceosal2_data$ceoten) # The average of tenure
[1] 7.954802
>
```

(ii) How many CEOs are in their first year as CEO (that is, ceoten 5 0)? What is the longest tenure as a CEO?

```
58  # (ii)
59  first_tenure <- subset(ceosal2_data, ceoten == '0')
60  length(first_tenure) # Number of first conten
61  max(ceosal2_data$ceoten) # Max of conten
62
60:45  (Top Level)                                                    R
```

```
Console  Terminal  Background Jobs
R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/
> # (ii)
> first_tenure <- subset(ceosal2_data, ceoten == '0')
> length(first_tenure) # Number of first conten
[1] 15
> max(ceosal2_data$ceoten) # Max of conten
[1] 37
>
```

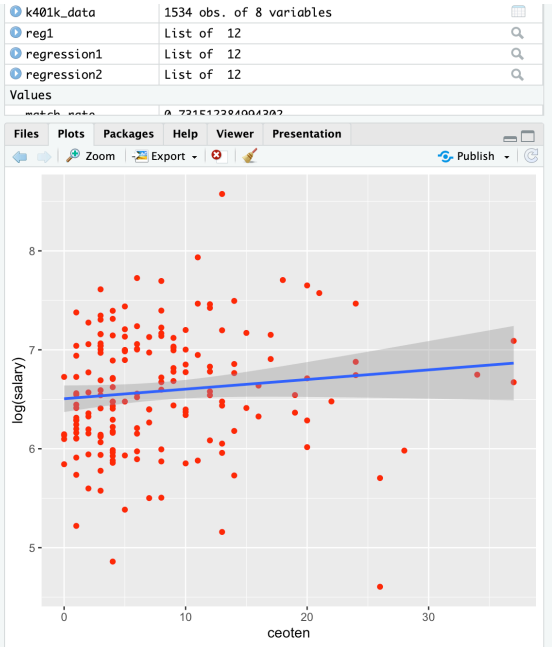
(iii) Estimate the simple regression model

$$\log(salary) = \beta_0 + \beta_1 ceoten + u,$$

and report your results in the usual form. What is the (approximate) predicted percentage increase in salary given one more year as a CEO?

```
# (iii)
regression2 <- lm(log(salary) ~ ceoten, data = ceosal2_data)
ggplot(ceosal2_data, aes(x = ceoten, y = log(salary))) +
  geom_point(color = 'red') +
  geom_smooth(method = 'lm') +
  labs(x = 'ceoten', y = 'log(salary)') # A scatterplot and SRF for ceoten & log(salary)
summary(regression2) # The report for results(sample size and R-squared)

## The (approximate) predicted percentage increase in salary given one more year as a CEO
regression2$coefficients[2] * 100
```

```
Call:
lm(formula = log(salary) ~ ceoten, data = ceosal2_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15314 -0.38319 -0.02251  0.44439  1.94337

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.505498   0.067991  95.682   <2e-16 ***
ceoten      0.009724   0.006364   1.528    0.128
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6038 on 175 degrees of freedom
Multiple R-squared:  0.01316,   Adjusted R-squared:  0.007523
F-statistic: 2.334 on 1 and 175 DF,  p-value: 0.1284

>
> ## The (approximate) predicted percentage increase in salary given one more year as a CEO
> regression2$coefficients[2] * 100
   ceoten
0.9723632
>
```



Predicted percentage increase in salary given one more year as a CEO: **0.97236 %**

C3. Use the data in SLEEP75.RAW from Biddle and Hamermesh (1990) to study whether there is a tradeoff between the time spent sleeping per week and the time spent in paid work. We could use either variable as the dependent variable. For concreteness, estimate the model

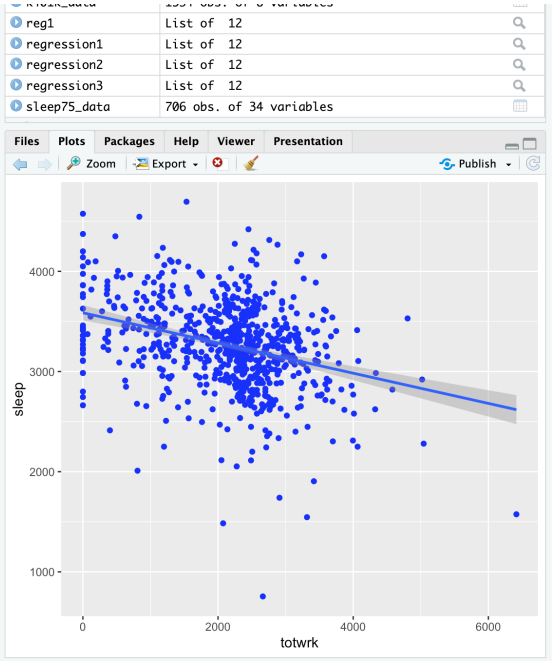
$$sleep = \beta_0 + \beta_1 totwrk + u,$$

where sleep is minutes spent sleeping at night per week and totwrk is total minutes worked during the week.

(i) Report your results in equation form along with the number of observations and R2. What does the intercept in this equation mean?

```
# C3
load("/Users/divya/Documents/Semester-6/Econometric Data Analysis with R/R data sets for 5e/sleep75.Rdata")
sleep75_data <- sleep75
# (i)
regression3 <- lm(sleep ~ totwrk, data = sleep75_data)
ggplot(sleep75_data, aes(x = totwrk, y = sleep)) +
  geom_point(color = 'blue') +
  geom_smooth(method = 'lm') +
  labs(x = 'totwrk', y = 'sleep') # A scatterplot and SRF for totwrk & sleep
summary(regression3) # The report for results(sample size and R-squared)
```

```
> ggplot(sleep75_data, aes(x = totwrk, y = sleep)) +
+   geom_point(color = 'blue') +
+   geom_smooth(method = 'lm') +
+   labs(x = 'totwrk', y = 'sleep') # A scatterplot and SRF for totwrk & sleep
`geom_smooth()` using formula = 'y ~ x'
> summary(regression3) # The report for results(sample size and R-squared)

Call:
lm(formula = sleep ~ totwrk, data = sleep75_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2429.94  -240.25     4.91   250.53  1339.72

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3586.37695   38.91243  92.165   <2e-16 ***
totwrk        -0.15075    0.01674  -9.005   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 421.1 on 704 degrees of freedom
Multiple R-squared:  0.1033,    Adjusted R-squared:  0.102
F-statistic: 81.09 on 1 and 704 DF,  p-value: < 2.2e-16
```



The intercept implies that the estimated amount of sleep per week for someone who does not work is 3,586.4 minutes, or about 59.77 hours. This comes to about 8.5 hours per night.

(ii) If totwrk increases by 2 hours, by how much is sleep estimated to fall? Do you find this to be a large effect?

```
> # (ii)
> # Two more hour
> regression3$coefficients[2] * 2*60
  totwrk
-18.0895
>
> # One more hour 5 days
> regression3$coefficients[2] * 1*60*5
   totwrk
-45.22375
```

If someone works two more hours per week then Δtotwrk = 120 (because totwrk is measured in minutes), and so Δ sleep = –0.151(120) = –18.12 minutes.

If someone were to work one more hour on each of five working days, Δsleep = –45.22 minutes.

C4 Use the data in WAGE2.RAW to estimate a simple regression explaining monthly salary (wage) in terms of IQ score (IQ).
(i) Find the average salary and average IQ in the sample. What is the sample standard deviation of IQ? (IQ scores are standardized so that the average in the population is 100 with a standard deviation equal to 15.)

```
 98  #C4
 99  load("/Users/divya/Documents/Semester-6/Econometric Data Analysis with R/R data sets for 5e/wage2.Rdata")
100  w<-wage2 #regressand=wage, regressor=IQ
101
102  #(i)
103  mean(w$wage)
104  mean(w$IQ)
105  sd(w$IQ)
```

103:13 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/

```
> w<-wage2 #regressand=wage, regressor=IQ
>
> #(i)
> mean(w$wage)
[1] 957.9455
> mean(w$IQ)
[1] 101.2824
> sd(w$IQ)
[1] 15.05264
>
```

(ii) Estimate a simple regression model where a one-point increase in IQ changes wage by a constant dollar amount. Use this model to find the predicted increase in wage for an increase in IQ of 15 points. Does IQ explain most of the variation in wage?

```
#(ii)
reg4<-lm(wage~IQ, data=w)
ggplot(w, aes(x=IQ, y=wage))+geom_point(color='red')+geom_smooth(method = 'lm')+labs(x='IQ',y='wage') #A scatte
summary(reg4) #The report for results(sample size and R-squared)

##if IQ increases about 15, then wage changes....
reg4$coefficients[2]*15

## the variation in wage explained by IQ is...

SSE_w<- sum(predict(reg4)-mean(w$wage)) # SSE for wage(You can also check the R-squared)
SSE_w
```

117:54 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/

```
Call:
lm(formula = wage ~ IQ, data = w)

Residuals:
   Min     1Q Median     3Q    Max
-898.7 -256.5  -47.3  201.1 2072.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.9916    85.6415   1.366    0.172
IQ            8.3031     0.8364   9.927   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 384.8 on 933 degrees of freedom
Multiple R-squared:  0.09554,   Adjusted R-squared:  0.09457
F-statistic: 98.55 on 1 and 933 DF,  p-value: < 2.2e-16

>
> ##if IQ increases about 15, then wage changes....
> reg4$coefficients[2]*15
     IQ
124.546
>
> ## the variation in wage explained by IQ is...
>
> SSE_w<- sum(predict(reg4)-mean(w$wage)) # SSE for wage(You can also check the R-squared)
> SSE_w
[1] 2.117645e-09
>
```

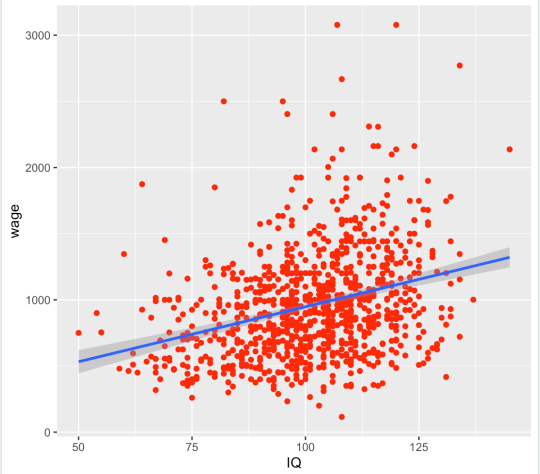
R | Global Environment

first_tenure	5 obs. of 15 variables
k	1534 obs. of 8 variables
k401k_data	1534 obs. of 8 variables
reg1	List of 12
reg4	List of 12
regression1	List of 12
regression2	List of 12
regression3	List of 12
sleep75_data	706 obs. of 34 variables
w	935 obs. of 17 variables

Values

Files Plots Packages Help Viewer Presentation

Zoom Export Publish



Increase in wage = 124.546

The R-squared value of the regression is 0.09554, which means that IQ explains only **9.5%** of the variation in wage. This suggests that there are other factors beyond IQ that affect salary.

(iii) Now, estimate a model where each one-point increase in IQ has the same percentage effect on wage. If IQ increases by 15 points, what is the approximate percentage increase in predicted wage?

```
#(iii)
reg4_1<-lm(lwage~IQ, data=w)
ggplot(w, aes(x=IQ, y=lwage))+geom_point(color='brown')+geom_smooth(method = 'lm')+labs(x='IQ',y='log(wage)') #
summary(reg4_1) #The report for results(sample size and R-squared)

##the approximate percentage increase in predicted wage if IQ increases by 15 points
reg4_1$coefficients[2]*100*15
```

122:43 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/

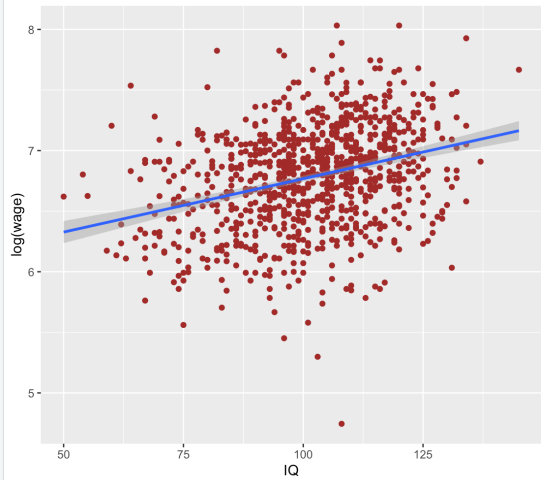
```
> #(iii)
> reg4_1<-lm(lwage~IQ, data=w)
> ggplot(w, aes(x=IQ, y=lwage))+geom_point(color='brown')+geom_smooth(method = 'lm')+labs(x='IQ',y='log(wage)') #A sca
tterplot and SRF for IQ & log(wage)
`geom_smooth()` using formula = 'y ~ x'
> summary(reg4_1) #The report for results(sample size and R-squared)

Call:
lm(formula = lwage ~ IQ, data = w)

Residuals:
     Min       1Q   Median       3Q      Max
-2.09324 -0.25547  0.02261  0.27544  1.21487

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.8869944  0.0890206   66.13   <2e-16 ***
IQ          0.0088072  0.0008694   10.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3999 on 933 degrees of freedom
Multiple R-squared:  0.09909,   Adjusted R-squared:  0.09813
F-statistic: 102.6 on 1 and 933 DF,  p-value: < 2.2e-16

>
> ##the approximate percentage increase in predicted wage if IQ increases by 15 points
> reg4_1$coefficients[2]*100*15
      IQ
13.21073
>
```

reg4	List of 12
reg4_1	List of 12
regression1	List of 12
regression2	List of 12
regression3	List of 12
sleep75_data	706 obs. of 34 variables
w	935 obs. of 17 variables

Values

Files Plots Packages Help Viewer Presentation

Zoom Export Publish



Percentage Increase in Wage = 13.21

C5. For the population of firms in the chemical industry, let rd denote annual expenditures on research and development, and let sales denote annual sales (both are in millions of dollars).
(i) Write down a model (not an estimated equation) that implies a constant elasticity between rd and sales. Which parameter is the elasticity?

The constant elasticity model is a log-log model:

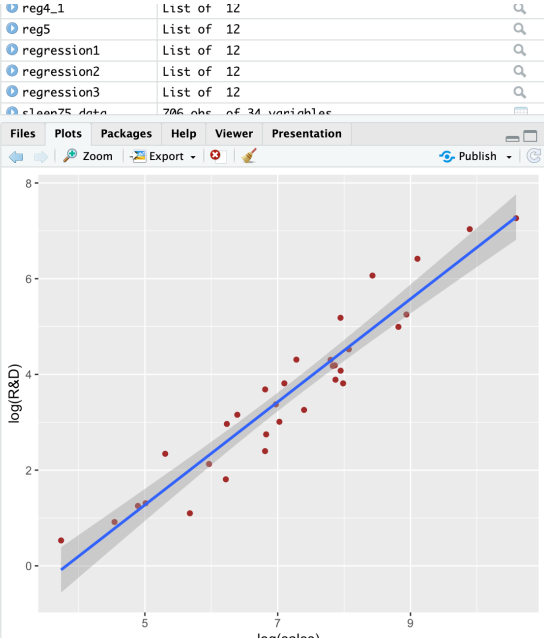
$$\log(rd) = \beta_0 + \beta_1 \log(sales) + u,$$

where β_1 is the elasticity of rd with respect to sales.

(ii) Now, estimate the model using the data in RDCHEM.RAW. Write out the estimated equation in the usual form. What is the estimated elasticity of rd with respect to sales? Explain in words what this elasticity means.

```
#C5
load("/Users/divya/Documents/Semester-6/Econometric Data Analysis with R/R data sets for 5e/rdchem.Rdata")
r<-rdchem
#(ii)
reg5<-lm(lrd~lsales, data=r)
ggplot(r, aes(x=lsales, y=lrd))+geom_point(color='brown')+geom_smooth(method = 'lm')+labs(x='log(sales)',y='log
summary(reg5) #The report for results(sample size and R-squared)
```

```
> #(ii)
> reg5<-lm(lrd~lsales, data=r)
> ggplot(r, aes(x=lsales, y=lrd))+geom_point(color='brown')+geom_smooth(method = 'lm')+labs(x='log(sales)',y='log(R&
D)') #A scatterplot and SRF for log(sales) & log(rd)
`geom_smooth()` using formula = 'y ~ x'
> summary(reg5) #The report for results(sample size and R-squared)

Call:
lm(formula = lrd ~ lsales, data = r)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90406 -0.40086 -0.02178  0.40562  1.10439

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.10472    0.45277  -9.066 4.27e-10 ***
lsales       1.07573    0.06183  17.399  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5294 on 30 degrees of freedom
Multiple R-squared:  0.9098,    Adjusted R-squared:  0.9068
F-statistic: 302.7 on 1 and 30 DF,  p-value: < 2.2e-16
```



$$\widehat{\log(rd)} = -4.105 + 1.076 \log(sales)$$

$R^2 = 0.9098$
The estimated elasticity of rd with respect to sales is 1.07573, which is just above one. A one increase in sales is estimated to increase rd by about 1.07573%.

C8. To complete this exercise you need a software package that allows you to generate data from the uniform and normal distributions.
(i) Start by generating 500 observations xi – the explanatory variable – from the uniform distribution with range [0,10]. (Most statistical packages have a command for the Uniform[0,1] distribution; just multiply those observations by 10.) What are the sample mean and sample standard deviation of the xi?

```
177 #(i)
178 x<-runif(500, min = 0, max = 10) #Uniform distribution(0,10), n=500
179 head(x)
180 mean(x)
181 sd(x)
182
```

182:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/

```
> #(i)
> x<-runif(500, min = 0, max = 10) #Uniform distribution(0,10), n=500
> head(x)
[1] 4.340518 6.210572 6.311716 0.866398 1.562267 4.375621
> mean(x)
[1] 4.893856
> sd(x)
[1] 2.867174
>
```

(ii) Randomly generate 500 errors, ui, from the Normal[0,36] distribution. (If you generate a Normal[0,1], as is commonly available, simply multiply the outcomes by six.) Is the sample average of the ui exactly zero? Why or why not? What is the sample standard deviation of the ui?

```
184 #(ii)
185 u<-rnorm(500,0,36) #Normal distribution(0,36), n=500
186 mean(u) #Why isn't it zero??
187 sd(u)
```

186:29 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/

```
> #(ii)
> u<-rnorm(500,0,36) #Normal distribution(0,36), n=500
> mean(u) #Why isn't it zero??
[1] 3.796335
> sd(u)
[1] 34.75609
>
```

(iii) Now generate the yi as

$$y_i = 1 + 2x_i + u_i \equiv \beta_0 + \beta_1 x_i + u_i;$$

that is, the population intercept is one and the population slope is two. Use the data to run the regression of yi on xi. What are your estimates of the intercept and slope? Are they equal to the population values in the above equation? Explain

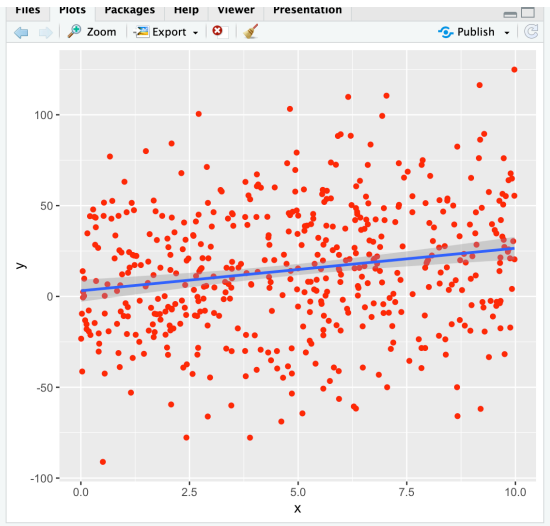
```
189 #(iii)
190 y=1+2*x+u
191 reg8<-lm(y~x)
192 dat<-data.frame(x,y)
193 head(dat)
194
195 ggplot(dat, aes(x=x, y=y))+geom_point(color='red')+geom_smooth(method = 'lm')+labs(x='x',y='y') #A scatterplot
196 summary(reg8) #The report for results(sample size and R-squared)
```

183:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/

```
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-95.383 -22.729  -0.998  24.346  98.385

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.119      3.079   1.013    0.312
x              2.343      0.543   4.315 1.93e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.78 on 498 degrees of freedom
Multiple R-squared:  0.03603,   Adjusted R-squared:  0.0341
F-statistic: 18.61 on 1 and 498 DF,  p-value: 1.93e-05

>
```



(iv) Obtain the OLS residuals, u^i, and verify that equation (2.60) hold (subject to rounding error).

```
198 #(iv)
199 sum(reg8$residuals)
200 sum(reg8$residuals*x)
201
```

199:20 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/

```
> #(iv)
> sum(reg8$residuals)
[1] -7.549517e-14
> sum(reg8$residuals*x)
[1] -9.841017e-13
>
```

(v) Compute the same quantities in equation (2.60) but use the errors ui in place of the residuals. Now what do you conclude?

```
202 #(v)
203 sum(u)
204 sum(u*x)
205
```

203:7 (Top Level) R Scr

Console Terminal Background Jobs

R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/

```
> #(v)
> sum(u)
[1] 1898.168
> sum(u*x)
[1] 10695.23
>
```

(vi) Repeat parts (i), (ii), and (iii) with a new sample of data, starting with generating the xi. Now what do you obtain for $\hat{b}_0$ and $\hat{b}_1$? Why are these different from what you obtained in part (iii)?

```
177 #(i)
178 x<-runif(500, min = 0, max = 10) #Uniform distribution(0,10), n=500
179 head(x)
180 mean(x)
181 sd(x)
```

169:39 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/

```
> head(x)
[1] 0.3511547 1.9960870 7.3540410 2.0132090 0.7831408 2.4621845
> mean(x)
[1] 5.07236
> sd(x)
[1] 2.923667
>
```

```
184 #(ii)
185 u<-rnorm(500,0,36) #Normal distribution(0,36), n=500
186 mean(u) #Why isn't it zero??
187 sd(u)
188
```

185:53 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/

```
> #(ii)
> u<-rnorm(500,0,36) #Normal distribution(0,36), n=500
> mean(u) #Why isn't it zero??
[1] 0.5854473
> sd(u)
[1] 35.82763
>
```

```r
#(iii)
y=1+2*x+u
reg8<-lm(y~x)
dat<-data.frame(x,y)
head(dat)

ggplot(dat, aes(x=x, y=y))+geom_point(color='red')+geom_smooth(method = 'lm')+labs(x='x',y='y') #A scatterplot
summary(reg8) #The report for results(sample size and R-squared)
```

195:64 (Top Level) R Script

Console Terminal Background Jobs

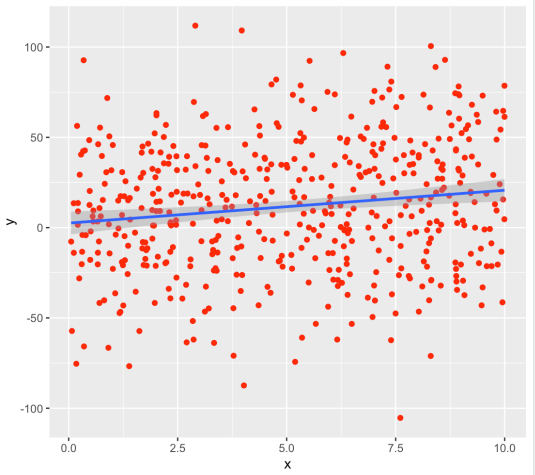
R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-121.697  -23.866    0.523   25.153  104.043

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5020     3.2138   0.779 0.436634
x             1.8193     0.5491   3.313 0.000989 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.86 on 498 degrees of freedom
Multiple R-squared:  0.02157,   Adjusted R-squared:  0.01961
F-statistic: 10.98 on 1 and 498 DF,  p-value: 0.0009887
```

y num [1:500] -65.8 -21.2 76.5 63.4 15.3 ...

Files Plots Packages Help Viewer Presentation

Zoom Export Publish




```r
#(iv)
sum(reg8$residuals)
sum(reg8$residuals*x)

#(v)
sum(u)
sum(u*x)
```

202:5 (Top Level) R Script

Console Terminal Background Jobs

R 4.2.2 · ~/Documents/Semester-6/Econometric Data Analysis with R/

```
Residual standard error: 35.86 on 498 degrees of freedom
Multiple R-squared:  0.02157,   Adjusted R-squared:  0.01961
F-statistic: 10.98 on 1 and 498 DF,  p-value: 0.0009887

> #(iv)
> sum(reg8$residuals)
[1] -1.33582e-12
> sum(reg8$residuals*x)
[1] -2.341949e-11
>
> #(v)
> sum(u)
[1] 292.7237
> sum(u*x)
[1] 714.0666
>
```