Econometrics - 1

Data Assignment

Name: Varun Khurana

Roll No.: 2019124

Question-1

Let us fetch the data set…

```
df = read.csv("C:\\Users\\varun\\Desktop\\eco_assignment\\eco_dataset.csv")
head(df, 5)
```

```
                         STATE    YEAR     BT    EL    CA     GT     DW      GER
1 Andaman & Nicobar Islands 2013-14  94.52 88.86 53.06  93.44  98.69 95.68000
2 Andaman & Nicobar Islands 2014-15 100.00 88.89 57.25 100.00  99.52 82.56333
3 Andaman & Nicobar Islands 2015-16 100.00 90.10 57.00 100.00 100.00 91.39667
4             Andhra Pradesh 2013-14  56.88 90.34 29.57  81.31  90.35 80.20333
5             Andhra Pradesh 2014-15  65.34 92.76 28.06  98.07  93.74 75.32333
```

(a) Average GER of India from 2013-14 to 2015-16 = 87.02994

```
a = mean(x = df$GER)
print(a)
```

```
[1] 87.02994
```

(b) 70.04178% of schools on an average are electrified.

```r
b = mean(x = df$EL)
print(b)
```

```
[1] 70.04178
```

(c) 94.89187% of schools on an average have drinking water facility.

```r
c = mean(x = df$DW)
print(c)
```

```
[1] 94.89187
```

(d) 91.92449% of schools on an average have boys toilets.

```r
d = mean(x = df$BT)
print(d)
```

```
[1] 91.92449
```

(e) 94.96692% of schools on an average have girls toilets.

```r
e = mean(x = df$GT)
print(e)
```

```
[1] 94.96692
```

(f) 40.40252% of schools on an average have computer labs.

```r
f = mean(x = df$CA)
print(f)
```

```
[1] 40.40252
```

(g) Variance for above variables:

(1) GER

```r
a = var(df$GER)
print(a)
```

```
[1] 127.6561
```

(2) Percentage of schools electrified

```r
a = var(df$EL)
print(a)
```

```
[1] 949.3671
```

(3) Percentage of schools having drinking water supply

```r
a = var(df$DW)
print(a)
```

```
[1] 63.40203
```

(4) Percentage of schools having boys toilets

```r
a = var(df$BT)
print(a)
```

```
[1] 140.7956
```

(5) Percentage of schools having girls toilets

```r
a = var(df$GT)
print(a)
```

```
[1] 72.51242
```

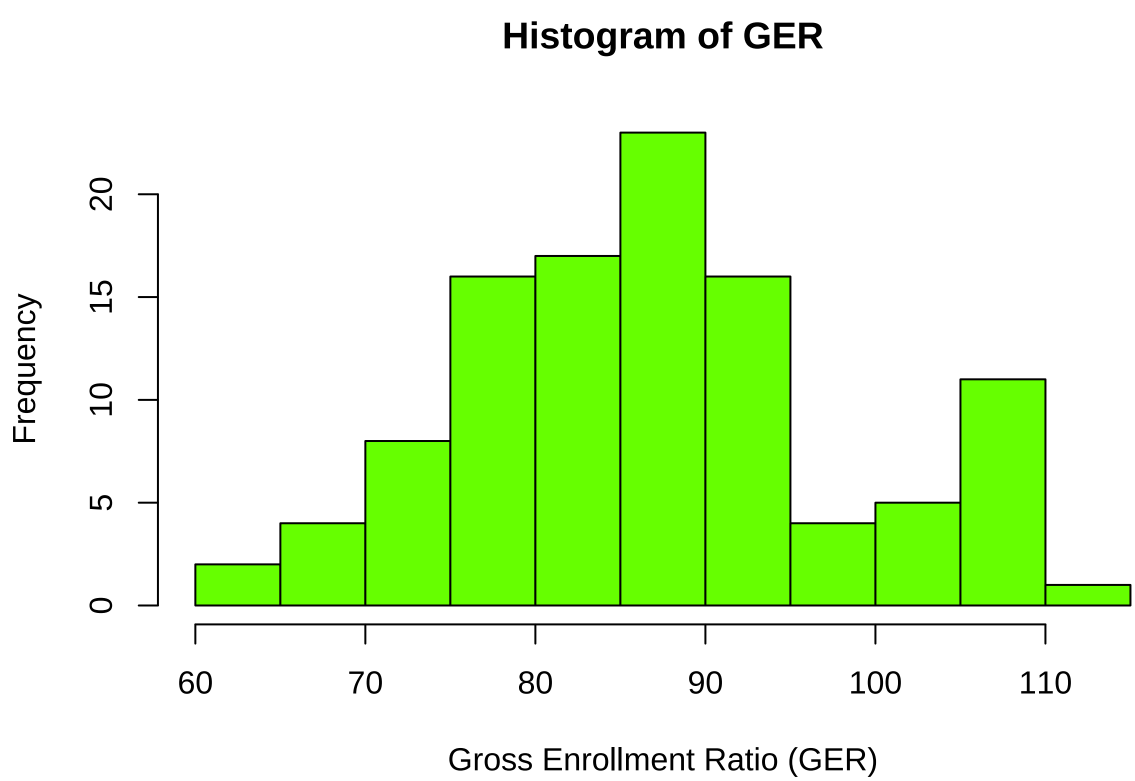
(6) Percentage of schools having computer facilities

```r
a = var(df$CA)
print(a)
```

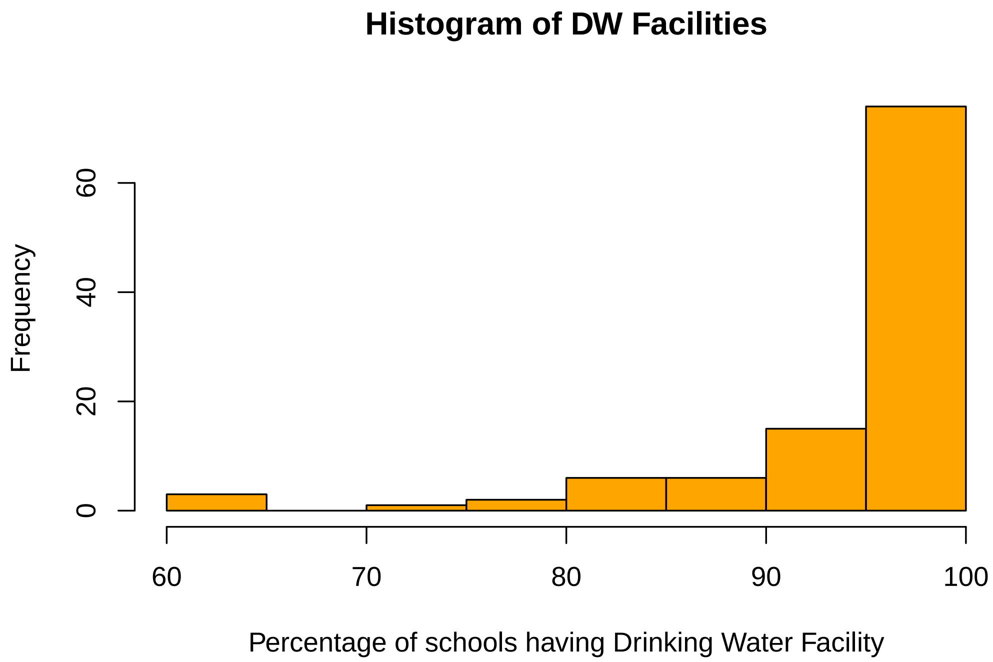
```
[1] 780.8061
```

(h) HISTOGRAMS:

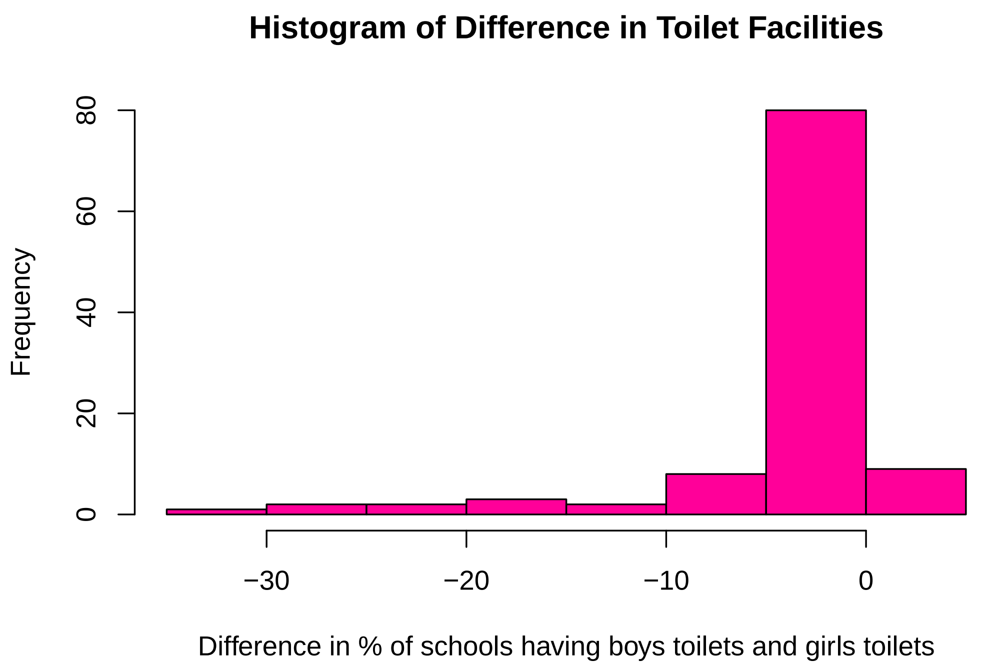
(1) Gross Enrollment Ratio (GER)



(2) Drinking Water Facilities



(3) Difference of boys and girls toilet availability



(i) Inference:

- The Gross Enrollment Ratio (GER): It is fairly high. The average GER is 87.02%.
- A high percentage of schools in India are equipped with Drinking Water Facilities.
- The difference between availability of boys toilets and girls toilets in schools is very narrow.

Question-2

Let us fetch the Literacy dataset and merge it with the current dataset.

Note that the Literacy data is for the year 2011. However, Telangana was created as a new state in 2014. So it's literacy data is not available in the dataset.

```
df2 = df
df3 = read.csv("C:\\Users\\varun\\Desktop\\eco_assignment\\literacy_rates_dataset.csv")
df2 = merge(df2, df3, by.x = "STATE", by.y = "State")
```

Let us drop the redundant columns in the merged dataset and rename the Literacy rate column as "LIT".

```
df2 = subset(df2, select = -c(Male, Female, S.No., X..Change))
library(dplyr)
df2 = rename(df2, LIT = Literacy)
head(df2, 5)
```

```
                      STATE    YEAR     BT    EL    CA     GT     DW      GER
1 Andaman & Nicobar Islands 2013-14  94.52 88.86 53.06  93.44  98.69 95.68000
2 Andaman & Nicobar Islands 2014-15 100.00 88.89 57.25 100.00  99.52 82.56333
3 Andaman & Nicobar Islands 2015-16 100.00 90.10 57.00 100.00 100.00 91.39667
4            Andhra Pradesh 2014-15  65.34 92.76 28.06  98.07  93.74 75.32333
5            Andhra Pradesh 2015-16  99.69 93.50 30.59  99.72  95.37 73.10333
```

```
    LIT
1 86.63
2 86.63
3 86.63
4 67.02
5 67.02
```

We will now split the States into 3 categories – HIGH, MEDIUM and LOW, based on their literacy rates.

LOW: 0-33 percentile of literacy rate

MEDIUM: 34-67 percentile of literacy rate

HIGH: 68-100 percentile of literacy rate

```r
a0 <- quantile(df2$LIT, 0.00)    # 0 %ile (min literacy rate)
a <- quantile(df2$LIT, 0.33)     # 33 %ile
b <- quantile(df2$LIT, 0.67)     # 67 %ile
c <- quantile(df2$LIT, 1.00)     # 100 %ile (max literacy rate)
```

```
Percentiles:

  0%
61.8

  33%
74.43

  67%
82.34

100%
  94
```

(a) Finding the mean of various parameters for these Literacy categories:

- **Mean Gross Enrollment Ratio (GER)**

LIT_CATEGORY	GER
HIGH	91.57172
MEDIUM	87.61093
LOW	82.43046

- **Mean Boys Toilet Availability**

LIT_CATEGORY	BT
HIGH	97.35818
MEDIUM	96.40778
LOW	82.75083

- **Mean Girls Toilet Availability**

LIT_CATEGORY	GT
HIGH	99.02697
MEDIUM	97.97167
LOW	88.18028

- **Mean Drinking Water Facilities**

LIT_CATEGORY	DW
HIGH	98.01333
MEDIUM	96.69139

LIT_CATEGORY	DW
LOW	90.29167

- **Mean Electricity Availability**

LIT_CATEGORY	EL
HIGH	89.36424
MEDIUM	81.11639
LOW	40.07306

- **Mean Computer Facilities**

LIT_CATEGORY	CA
HIGH	63.69515
MEDIUM	44.02750
LOW	15.63139

(b) Pattern in the three Literacy Categories in terms of school enrollment and school infrastructure:

- We can clearly observe that the High literacy states perform best in all metrics i.e. Gross Enrollment Ratio, Boys Toilet Availability, Girls Toilet Availability, Electricity Availability, Drinking Water Facilities and Computer Facilities, followed by Medium literacy states which perform better than the Low literacy states.
- There is an evident correspondence between the literacy rates and the school enrollment, infrastructure, facilities.
- Thus, we conclude that in order to improve the literacy rate in the country, we require

greater investment on school infrastructure and better facilities for students so that they yield better learning outcomes and result in higher school enrollment.

(c) Now let us categorise the states according to the geographical and administrative criteria into the following groups:

- **NEHS**: North-east and Hilly States
- **UTC**: Union Territories and City states
- **SS**: Southern States
- **OTH**: Other Major states

We will add a column "STATE_CATEGORY" to indicate these categories for each state.

```
df5 = df
df5$STATE_CATEGORY = c("")
```

```
head(df5, 5)
```

```
                     STATE    YEAR     BT    EL    CA     GT     DW      GER
1 Andaman & Nicobar Islands 2013-14  94.52 88.86 53.06  93.44  98.69 95.68000
2 Andaman & Nicobar Islands 2014-15 100.00 88.89 57.25 100.00  99.52 82.56333
3 Andaman & Nicobar Islands 2015-16 100.00 90.10 57.00 100.00 100.00 91.39667
4            Andhra Pradesh 2013-14  56.88 90.34 29.57  81.31  90.35 80.20333
5            Andhra Pradesh 2014-15  65.34 92.76 28.06  98.07  93.74 75.32333
  STATE_CATEGORY
1            UTC
2            UTC
3            UTC
4             SS
5             SS
```

Now, let us compare the mean and variance of various parameters for each of these State groups.

- Gross Enrollment Ratio **(GER)**

STATE_CATEGORY	Mean_GER	Variance_GER
NEHS	96.25456	111.90833
UTC	85.08136	165.18985
OTH	82.90656	25.20695
SS	82.00867	83.71654

- Drinking Water Facility **(DW)**

STATE_CATEGORY	Mean_DW	Variance_DW
UTC	98.84815	8.173816
SS	97.88400	8.261057
OTH	97.19067	7.151738
NEHS	87.03767	121.823494

- Electricity Facility **(EL)**

STATE_CATEGORY	Mean_EL	Variance_EL
UTC	90.39926	535.1793
SS	85.44150	574.1554
OTH	60.10100	949.0531
NEHS	51.39433	653.7177

- Boys Toilet Availability **(BT)**

STATE_CATEGORY	Mean_BT	Variance_BT
UTC	95.78741	70.43388
OTH	92.64533	64.72936
SS	91.25800	170.77966
NEHS	88.17133	245.55624

- Girls Toilet Availability **(GT)**

STATE_CATEGORY	Mean_GT	Variance_GT
UTC	97.56593	37.47178
SS	96.48400	28.13214
OTH	94.51867	46.95898
NEHS	92.06467	149.26198

- Computer Facilities **(CA)**

STATE_CATEGORY	Mean_CA	Variance_CA
UTC	66.20889	803.8429
SS	46.68450	629.4039
OTH	26.70633	493.4141
NEHS	26.68500	191.5431

In general, we observe that the UTC group has relatively higher values of mean for most parameters. This suggests that Union Territories and City States fare the best in terms of school infrastructure followed by the Southern States. They also have high school enrollment ratio standing second after the North Eastern and Hilly States.

Question-3

(a) Regression Model for Gross Enrollment Ratio (GER)

$$GER_i = \beta_0 + \beta_1 DW_i + \beta_2 BT_i + \beta_3 GT_i + \beta_4 EL_i + \beta_5 CA_i + u_i$$

```
Call:
lm(formula = GER ~ DW + BT + GT + EL + CA, data = df8)


Residuals:
     Min       1Q   Median       3Q      Max
-27.3641  -6.6369  -0.0921   5.2118  22.4992


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 112.61642   16.81466   6.698 1.22e-09 ***
DW           -0.67651    0.19105  -3.541 0.000605 ***
BT            0.09368    0.17564   0.533 0.594945
GT            0.26601    0.27269   0.975 0.331649
EL            0.02865    0.06239   0.459 0.647085
CA            0.06753    0.05901   1.144 0.255138
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1


Residual standard error: 10.74 on 101 degrees of freedom
Multiple R-squared:  0.1394,    Adjusted R-squared:  0.09675
F-statistic: 3.271 on 5 and 101 DF,  p-value: 0.00886
```

(i) Regression coefficients:

$\beta_0 = 112.6164$

$\beta_1 = -0.67651$

$\beta_2 = 0.09368$

$\beta_3 = 0.26601$

$\beta_4 = 0.02865$

$\beta_5 = 0.06753$

(ii) Variance of regression coefficients:

$$V[\hat{\beta}_{OLS}] = \hat{\sigma}^2 (X'X)^{-1}$$

This gives the variance-covariance matrix of $\hat{\beta}_{OLS}$.

In this matrix each entry $a_{i,j}$:

$$a_{i,j} = \begin{cases} cov(\beta_i, \beta_j) & ; i \neq j \\ var(\beta i) & ; i = j \end{cases} \tag{1}$$

	(Intercept)	DW	BT	GT	EL	CA
(Intercept)	282.7328	-1.8333	0.8036	-2.2484	0.4784	-0.0383
DW	-1.8333	0.0365	-0.0024	-0.0122	-0.0042	0.0010
BT	0.8036	-0.0024	0.0308	-0.0367	0.0026	-0.0026
GT	-2.2484	-0.0122	-0.0367	0.0744	-0.0052	0.0022
EL	0.4784	-0.0042	0.0026	-0.0052	0.0039	-0.0024
CA	-0.0383	0.0010	-0.0026	0.0022	-0.0024	0.0035

The diagonal entries of this matrix will give the variance of the regression coefficients.

	Variance
(Intercept)	282.7328
DW	0.0365
BT	0.0308
GT	0.0744
EL	0.0039
CA	0.0035

(iii) Estimate of $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = 115.3048$$

(b) Regression model for GER by incorporating dummy variables for Literacy rate groups. In this case, let H_i, M_i, L_i be dummy variables such that:

$$\begin{cases} H_i = 1, M_i = 0, L_i = 0 & \text{for High literacy State} \\ H_i = 0, M_i = 1, L_i = 0 & \text{for Medium literacy State} \\ H_i = 0, M_i = 0, L_i = 1 & \text{for Low literacy State} \end{cases} \tag{2}$$

If we include all 3 dummy variables, H_i, M_i, L_i along with intercept in the regression model, then it will lead to **dummy variable trap**. Consequently, we will not be able to estimate the regression.

This is because, for any i^{th} observation, $H_i + M_i + L_i = 1$ always. If X denotes the matrix of explanatory variables, then the columns of X will not be linearly independent, i.e. X will not have full rank.

So, let us include only 2 dummy variables, H_i, M_i in the regression model specified as follows:

$$GER_i = \beta_0 + \beta_1 DW_i + \beta_2 BT_i + \beta_3 GT_i + \beta_4 EL_i + \beta_5 CA_i + \beta_6 H_i + \beta_7 M_i + u_i$$

```
Call:
lm(formula = GER ~ DW + BT + GT + EL + CA + H + M, data = df9)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-29.4356  -6.1092  -0.0241   5.6684  21.0364

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.552942  16.304069   7.149 1.65e-10 ***
DW           -0.591844   0.189135  -3.129  0.00232 **
BT            0.002202   0.179188   0.012  0.99022
GT            0.229264   0.266100   0.862  0.39105
EL           -0.025867   0.066432  -0.389  0.69785
CA           -0.002953   0.062353  -0.047  0.96232
H            12.609311   3.825660   3.296  0.00137 **
M             7.838742   3.356706   2.335  0.02159 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.37 on 97 degrees of freedom
Multiple R-squared:  0.2277,    Adjusted R-squared:  0.172
F-statistic: 4.086 on 7 and 97 DF,  p-value: 0.000572
```

(i) Regression coefficients:

$\beta_0 = 116.552942$

$\beta_1 = -0.591844$

$\beta_2 = 0.002202$

$\beta_3 = 0.229264$

$\beta_4 = -0.025867$

$\beta_5 = -0.002953$

$\beta_6 = 12.609311$

$\beta_7 = 7.838742$

(ii) Variance of regression coefficients:

$$V[\hat{\beta}_{OLS}] = \hat{\sigma}^2(X'X)^{-1}$$

This gives the variance-covariance matrix of $\hat{\beta}_{OLS}$.

In this matrix each entry $a_{i,j}$:

$$a_{i,j} = \begin{cases} cov(\beta_i, \beta_j) & ; i \neq j \\ var(\beta i) & ; i = j \end{cases} \tag{3}$$

	(Intercept)	DW	BT	GT	EL	CA	H	M
(Intercept)	265.8227	-1.6975	0.7104	-2.0963	0.4362	-0.0627	3.9905	2.5742
DW	-1.6975	0.0358	-0.0037	-0.0118	-0.0050	0.0008	0.1135	0.1197
BT	0.7104	-0.0037	0.0321	-0.0355	0.0034	-0.0023	-0.1111	-0.1402
GT	-2.0963	-0.0118	-0.0355	0.0708	-0.0046	0.0023	-0.0453	-0.0117
EL	0.4362	-0.0050	0.0034	-0.0046	0.0044	-0.0022	-0.0714	-0.0870
CA	-0.0627	0.0008	-0.0023	0.0023	-0.0022	0.0039	-0.0755	-0.0160
H	3.9905	0.1135	-0.1111	-0.0453	-0.0714	-0.0755	14.6357	9.2974
M	2.5742	0.1197	-0.1402	-0.0117	-0.0870	-0.0160	9.2974	11.2675

The diagonal entries of this matrix will give the variance of the regression coefficients.

	Variance
(Intercept)	265.8227
DW	0.0358
BT	0.0321
GT	0.0708

	Variance
EL	0.0044
CA	0.0039
H	14.6357
M	11.2675

(iii) Estimate of $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = 107.6256$$

(c) Regression model for GER by incorporating dummy variables for geographical and administrative groups of States. In this case, let N_i, U_i, S_i, O_i be dummy variables such that:

$$\begin{cases} N_i = 1, U_i = 0, S_i = 0, O_i = 0 & \text{for North-Eastern \& Hilly States} \\ N_i = 0, U_i = 1, S_i = 0, O_i = 0 & \text{for Union Territories \& City States} \\ N_i = 0, U_i = 0, S_i = 1, O_i = 0 & \text{for Southern States} \\ N_i = 0, U_i = 0, S_i = 0, O_i = 1 & \text{for Other States} \end{cases} \tag{4}$$

If we include all 4 dummy variables, N_i, U_i, S_i, O_i along with intercept in the regression model, then it will lead to **dummy variable trap**. Consequently, we will not be able to estimate the regression.

This is because, for any i^{th} observation, $N_i + U_i + S_i + O_i = 1$ always. If X denotes the matrix of explanatory variables, then the columns of X will not be linearly independent, i.e. X will not have full rank.

So, let us include only 3 dummy variables, N_i, U_i, S_i in the regression model specified as follows:

$$GER_i = \beta_0 + \beta_1 DW_i + \beta_2 BT_i + \beta_3 GT_i + \beta_4 EL_i + \beta_5 CA_i + \beta_6 N_i + \beta_7 U_i + \beta_8 S_i + u_i$$

```
Call:
lm(formula = GER ~ DW + BT + GT + EL + CA + N + U + S, data = df10)

Residuals:
    Min      1Q  Median      3Q     Max
-24.756  -5.995   0.797   5.573  23.655

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.28151   16.29031   4.376 3.03e-05 ***
DW           0.07606    0.20497   0.371   0.7114
BT           0.05626    0.15197   0.370   0.7121
GT          -0.07686    0.24018  -0.320   0.7497
EL           0.04622    0.05377   0.860   0.3921
CA           0.13133    0.05570   2.358   0.0204 *
N           14.58855    2.96472   4.921 3.48e-06 ***
U           -4.48196    2.94191  -1.523   0.1309
S           -4.51642    2.88603  -1.565   0.1208
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.093 on 98 degrees of freedom
Multiple R-squared:  0.4012,    Adjusted R-squared:  0.3523
F-statistic: 8.208 on 8 and 98 DF,  p-value: 1.906e-08
```

(i) Regression coefficients:

$\beta_0 = 71.28151$

$\beta_1 = 0.07606$

$\beta_2 = 0.05626$

$\beta_3 = -0.07686$

$\beta_4 = 0.04622$

$\beta_5 = 0.13133$

$\beta_6 = 14.58855$

$\beta_7 = -4.48196$

$\beta_8 = -4.51642$

(ii) Variance of regression coefficients:

$$V[\hat{\beta}_{OLS}] = \hat{\sigma}^2(X'X)^{-1}$$

This gives the variance-covariance matrix of $\hat{\beta}_{OLS}$.

In this matrix each entry $a_{i,j}$:

$$a_{i,j} = \begin{cases} cov(\beta_i, \beta_j) & ; i \neq j \\ var(\beta i) & ; i = j \end{cases} \tag{5}$$

	(Intercept)	DW	BT	GT	EL	CA	N	U	S
(Intercept)	265.374	-2.279	0.484	-1.134	0.359	-0.010	-23.393	-7.518	-7.168
DW	-2.279	0.042	-0.001	-0.017	-0.003	0.001	0.355	0.030	0.057
BT	0.484	-0.001	0.023	-0.027	0.002	-0.002	0.041	0.050	0.090
GT	-1.134	-0.017	-0.027	0.058	-0.003	0.001	-0.179	-0.004	-0.075
EL	0.359	-0.003	0.002	-0.003	0.003	-0.002	-0.008	-0.009	-0.028
CA	-0.010	0.001	-0.002	0.001	-0.002	0.003	-0.009	-0.070	-0.025
N	-23.393	0.355	0.041	-0.179	-0.008	-0.009	8.790	3.192	3.307
U	-7.518	0.030	0.050	-0.004	-0.009	-0.070	3.192	8.655	4.437
S	-7.168	0.057	0.090	-0.075	-0.028	-0.025	3.307	4.437	8.329

The diagonal entries of this matrix will give the variance of the regression coefficients.

	Variance
(Intercept)	265.3741
DW	0.0420
BT	0.0231
GT	0.0577
EL	0.0029
CA	0.0031
N	8.7895
U	8.6549
S	8.3292

(iii) Estimate of $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = 82.67839$$