Data Analysis
Week 6 Task Solutions

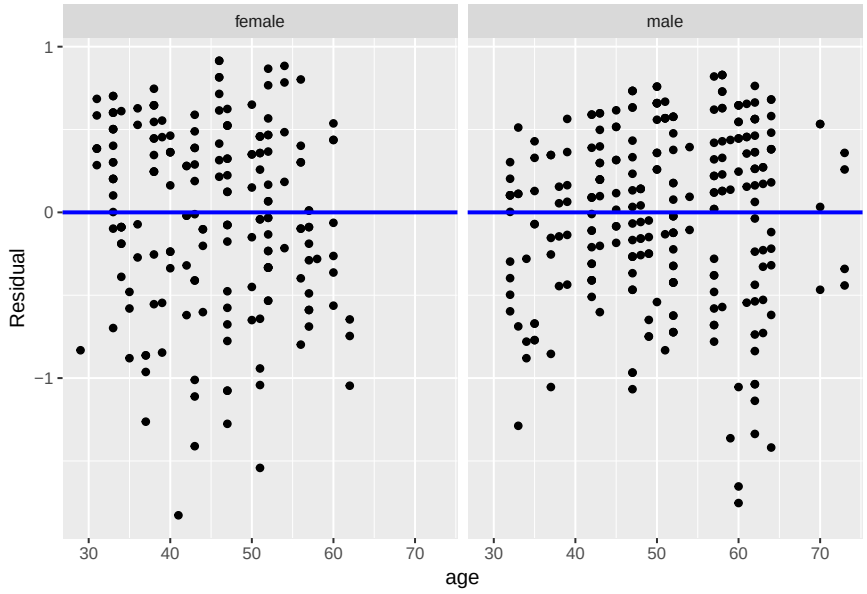
Tasks

1. Assess the model assumptions for the parallel regression lines model. Do they appear valid?

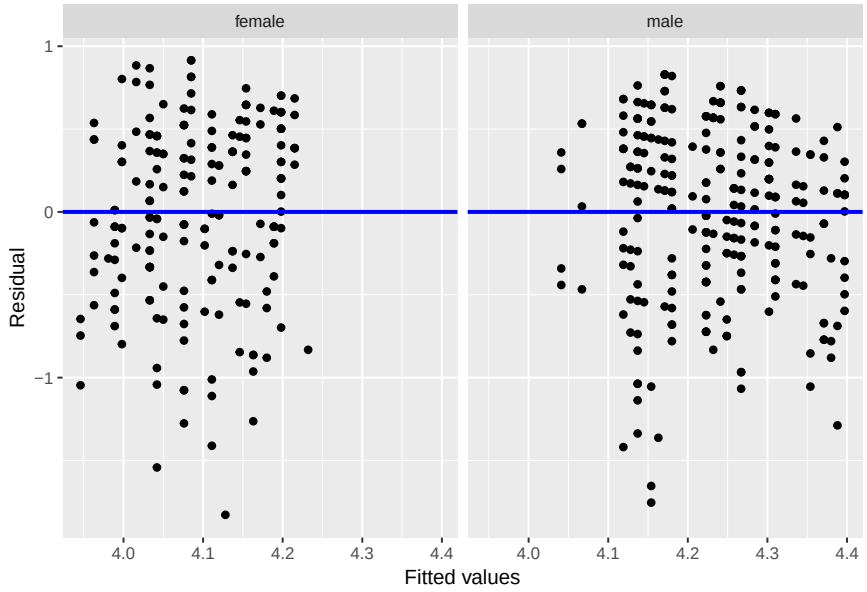
```
par.model <- lm(score ~ age + gender, data = eval.score)
regression.points <- get_regression_points(par.model)
```

```
Warning: package 'bindrcpp' was built under R version 3.4.4
```

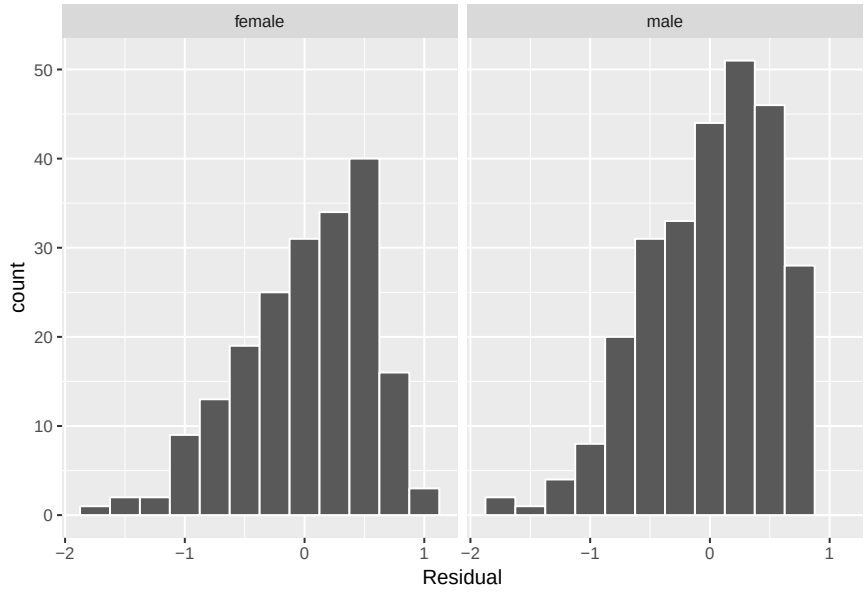
```
ggplot(regression.points, aes(x = age, y = residual)) +
  geom_point() +
  labs(x = "age", y = "Residual") +
  geom_hline(yintercept = 0, col = "blue", size = 1) +
  facet_wrap(~ gender)
```




```
ggplot(regression.points, aes(x = score_hat, y = residual)) +
geom_point() +
  labs(x = "Fitted values", y = "Residual") +
  geom_hline(yintercept = 0, col = "blue", size = 1) +
  facet_wrap(~ gender)
```




```
ggplot(regression.points, aes(x = residual)) +
  geom_histogram(binwidth = 0.25, color = "white") +
  labs(x = "Residual") +
  facet_wrap(~gender)
```



2. Return to the `Credit` data set and fit a multiple regression model with `Balance` as the outcome variable, and `Income` and `Age` as the explanatory variables, respectively. Assess the assumptions of the multiple regression model.

```
Cred <- Credit %>%
  select(Balance, Income, Age)

# skim_with(integer = list(hist = NULL)) # This supresses the histograms
# Cred %>%
#   skim()

Cred$Balance <- as.numeric(Cred$Balance)
Cred$Age <- as.numeric(Cred$Age)

skim_with(numeric = list(hist = NULL, missing = NULL, complete = NULL))
Cred %>%
  skim_to_list() %>%
  .$numeric %>%
  kable(col.names = c("Variable", "n", "Mean", "SD", "Minimum", "1st quartile", "Median",
                      "3rd quartile","Maximum"), caption =
          '\\label{tab:summary} Summary statistics on Credit Card Balance, Income and Age.',
        booktabs = TRUE, format = "latex") %>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 1: Summary statistics on Credit Card Balance, Income and Age.

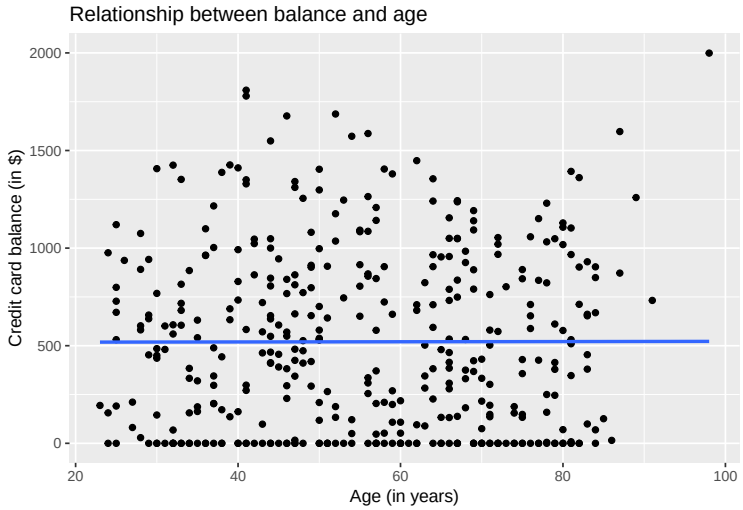
Variable	n	Mean	SD	Minimum	1st quartile	Median	3rd quartile	Maximum
Age	400	55.67	17.25	23	41.75	56	70	98
Balance	400	520.01	459.76	0	68.75	459.5	863	1999
Income	400	45.22	35.24	10.35	21.01	33.12	57.47	186.63

```
Cred %>%
  cor() %>%
  kable(caption =
          '\\label{tab:cor} Correlation Coefficients between Credit Card Balance,
        Income and Age.', booktabs = TRUE, format = "latex") %>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 2: Correlation Coefficients between Credit Card Balance, Income and Age.

	Balance	Income	Age
Balance	1.0000000	0.4636565	0.0018351
Income	0.4636565	1.0000000	0.1753384
Age	0.0018351	0.1753384	1.0000000

```
ggplot(Cred, aes(x = Age, y = Balance)) +
  geom_point() +
  labs(x = "Age (in years)", y = "Credit card balance (in $)",
       title = "Relationship between balance and age") +
  geom_smooth(method = "lm", se = FALSE)
```



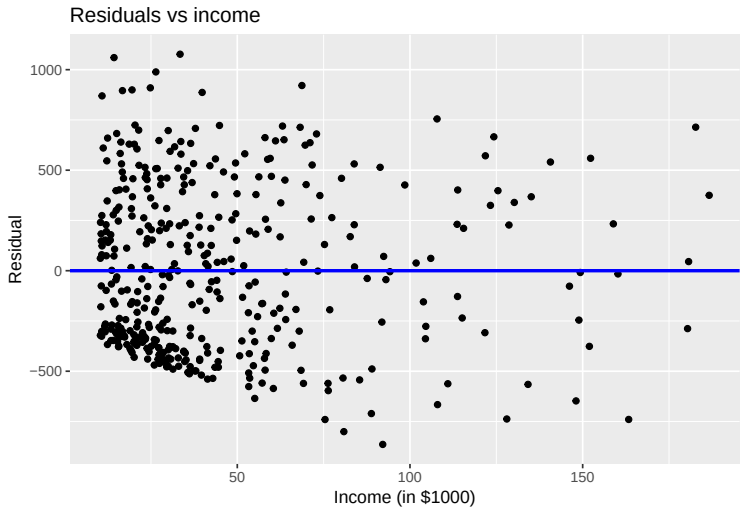

```
Balance.model <- lm(Balance ~ Age + Income, data = Cred)
get_regression_table(Balance.model) %>%
  kable(caption =
          '\\label{tab:reg} Estimated Coefficients from the fitted model
        Balance = Age + Income ', booktabs = TRUE, format = "latex") %>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 3: Estimated Coefficients from the fitted model Balance = Age + Income

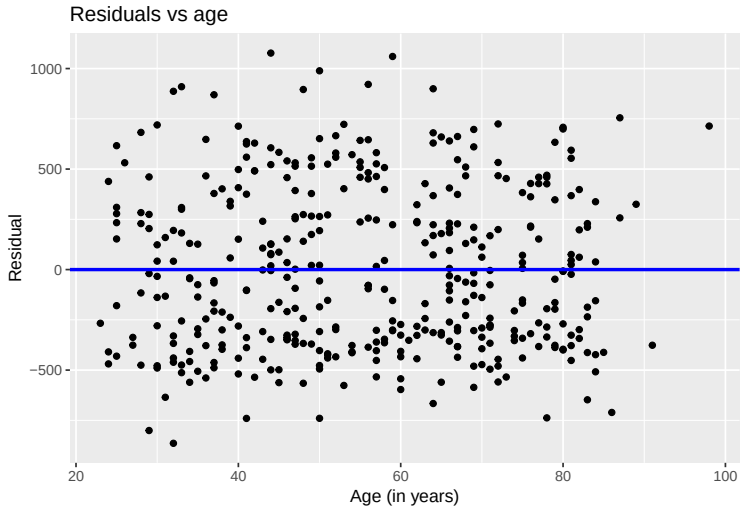
term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	359.673	70.358	5.112	0.000	221.351	497.994
Age	-2.185	1.199	-1.823	0.069	-4.542	0.172
Income	6.236	0.587	10.628	0.000	5.082	7.389

```
regression.points <- get_regression_points(Balance.model)

ggplot(regression.points, aes(x = Income, y = residual)) +
  geom_point() +
  labs(x = "Income (in $1000)", y = "Residual", title = "Residuals vs income")  +
  geom_hline(yintercept = 0, col = "blue", size = 1)
```

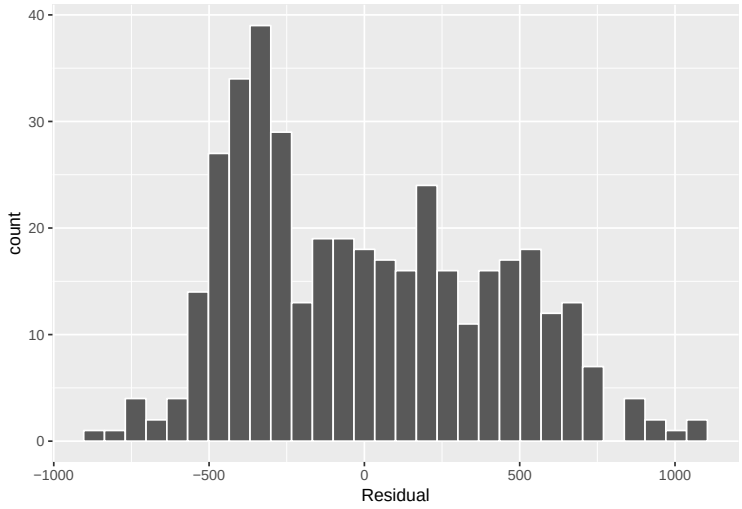



```
ggplot(regression.points, aes(x = Age, y = residual)) +
  geom_point() +
  labs(x = "Age (in years)", y = "Residual", title = "Residuals vs age")  +
  geom_hline(yintercept = 0, col = "blue", size = 1)
```




```
ggplot(regression.points, aes(x = residual)) +
  geom_histogram(color = "white") +
  labs(x = "Residual")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


40
30
20
10
0
count
−1000
−500
0
500
1000
Residual

3. Return to the `Credit` data set and fit a parallel regression lines model with `Balance` as the outcome variable, and `Income` and `Student` as the explanatory variables, respectively. Assess the assumptions of the fitted model.

```
Cred <- Credit %>%
  select(Balance, Income, Student)

# Cred %>%
#   skim()

Cred %>%
  group_by(Student) %>%
  summarise(n()) %>%
  kable(col.names = c("Student", "n"),caption =
          '\\label{tab:T3Student} Numbers of students and non-students',
        booktabs = TRUE, format = "latex") %>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 4: Numbers of students and non-students

Student	n
No	360
Yes	40

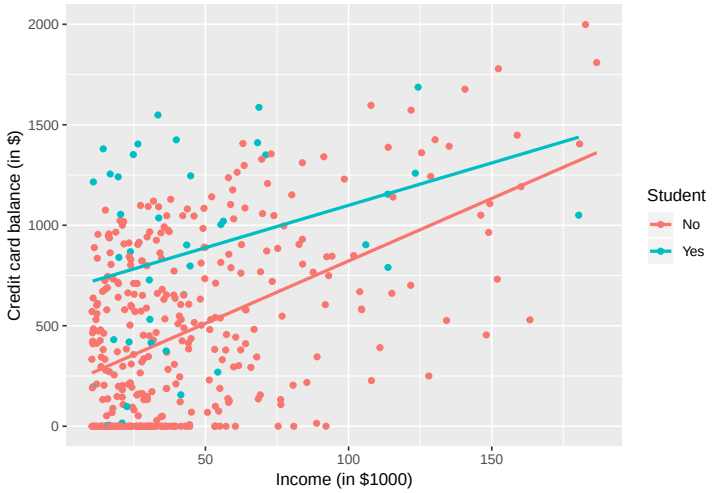
```
Cred$Balance <- as.numeric(Cred$Balance)

skim_with(numeric = list(hist = NULL, missing = NULL, complete = NULL))
Cred %>%
  skim_to_list() %>%
  .$numeric %>%
  kable(col.names = c("Variable", "n", "Mean", "SD", "Minimum", "1st quartile", "Median",
                      "3rd quartile","Maximum"), caption =
          '\\label{tab:T3summary} Summary statistics on Credit Card Balance and Income.',
        booktabs = TRUE, format = "latex") %>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 5: Summary statistics on Credit Card Balance and Income.

Variable	n	Mean	SD	Minimum	1st quartile	Median	3rd quartile	Maximum
Balance	400	520.01	459.76	0	68.75	459.5	863	1999
Income	400	45.22	35.24	10.35	21.01	33.12	57.47	186.63

```
ggplot(Cred, aes(x = Income, y = Balance, color = Student)) +
  geom_jitter() +
  labs(x = "Income (in $1000)", y = "Credit card balance (in $)", color = "Student") +
  geom_smooth(method = "lm", se = FALSE)
```



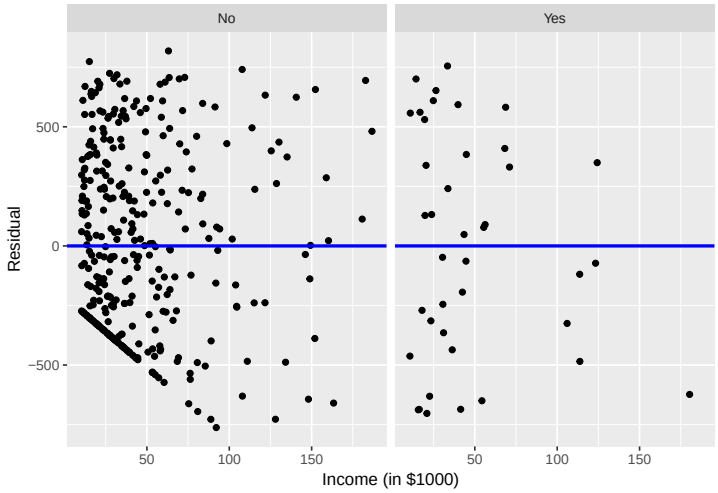

```
par.model <- lm(Balance ~ Income + Student, data = Cred)
get_regression_table(par.model) %>%
  kable(caption =
          '\\label{tab:T3reg} Estimated Coefficients from the fitted model
        Balance = Income + Student', booktabs = TRUE, format = "latex") %>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 6: Estimated Coefficients from the fitted model Balance = Income + Student

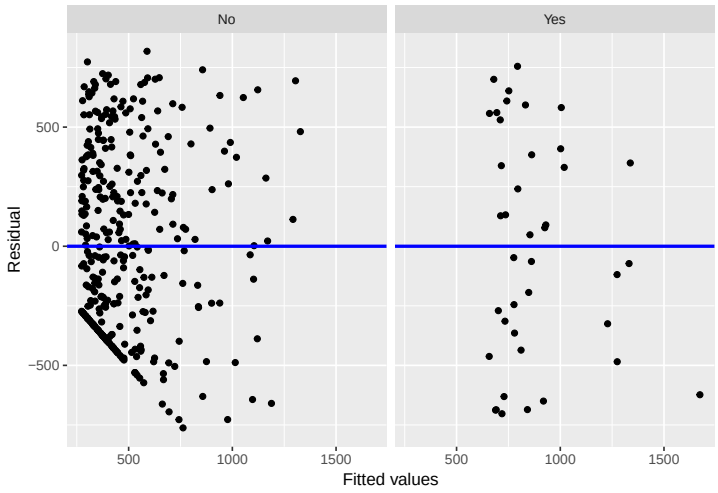
term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	211.143	32.457	6.505	0	147.333	274.952
Income	5.984	0.557	10.751	0	4.890	7.079
StudentYes	382.671	65.311	5.859	0	254.272	511.069

```
regression.points <- get_regression_points(par.model)

ggplot(regression.points, aes(x = Income, y = residual)) +
  geom_point() +
  labs(x = "Income (in $1000)", y = "Residual") +
  geom_hline(yintercept = 0, col = "blue", size = 1) +
  facet_wrap(~ Student)
```

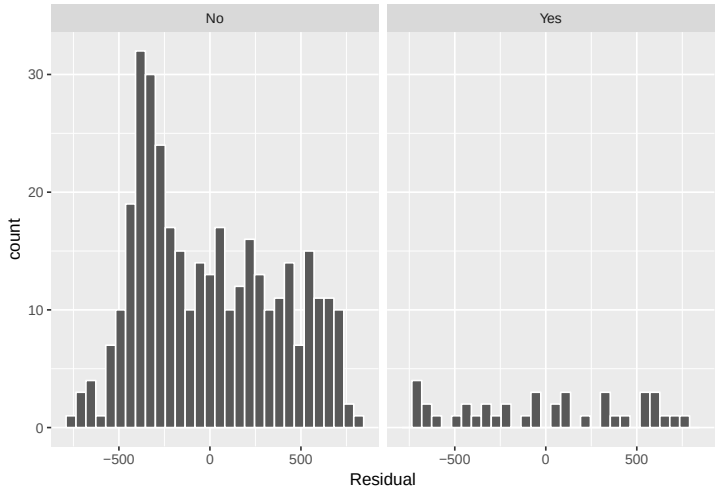



```
ggplot(regression.points, aes(x = Balance_hat, y = residual)) +
  geom_point() +
  labs(x = "Fitted values", y = "Residual") +
  geom_hline(yintercept = 0, col = "blue", size = 1) +
  facet_wrap(~ Student)
```




```
ggplot(regression.points, aes(x = residual)) +
  geom_histogram(color = "white") +
  labs(x = "Residual") +
  facet_wrap(~Student)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


No
Yes
30
20
10
0
count
−500
0
500
Residual

Trickier

4. Load the library `datasets` and look at the `iris` data set of Edgar Anderson containing measurements (in centimetres) on 150 different flowers across three different species of iris. Fit an interaction model with `Sepal.Width` as the outcome variable, and `Sepal.Length` and `Species` as the explanatory variables. Assess the assumptions of the fitted model.

```
library(datasets)

Irs <- iris %>%
  select(Sepal.Width, Sepal.Length, Species)

# Irs %>%
#   skim()

Irs %>%
  group_by(Species) %>%
  summarise(n()) %>%
  kable(col.names = c("Species", "n"),caption =
          '\\label{tab:T4Species} Numbers of different species',
        booktabs = TRUE, format = "latex") %>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 7: Numbers of different species

Species	n
setosa	50
versicolor	50
virginica	50

```
skim_with(numeric = list(hist = NULL, missing = NULL, complete = NULL))
Irs %>%
  skim_to_list() %>%
  .$numeric %>%
  kable(col.names = c("Variable", "n", "Mean", "SD", "Minimum", "1st quartile", "Median",
                      "3rd quartile","Maximum"), caption =
          '\\label{tab:T4summary} Summary statistics on Iris variables.',
        booktabs = TRUE, format = "latex") %>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 8: Summary statistics on Iris variables.

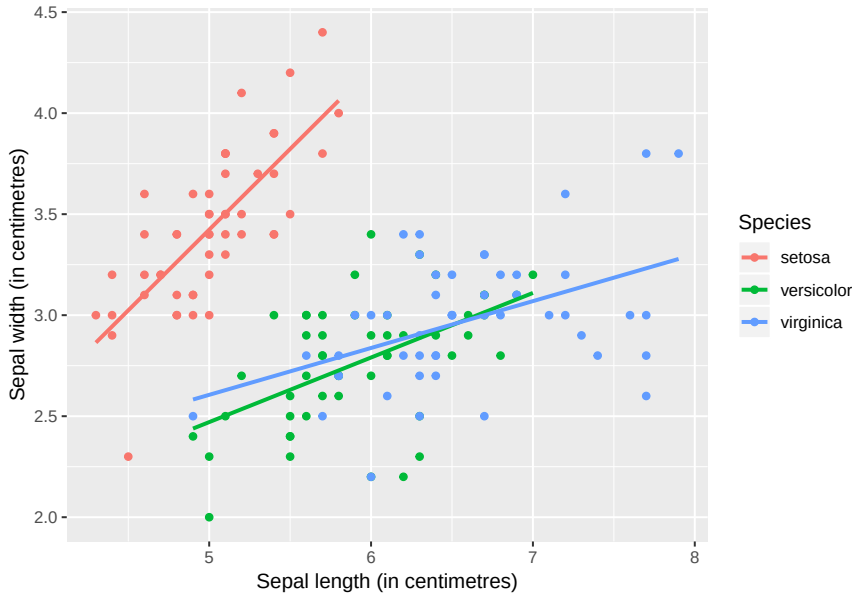
Variable	n	Mean	SD	Minimum	1st quartile	Median	3rd quartile	Maximum
Sepal.Length	150	5.84	0.83	4.3	5.1	5.8	6.4	7.9
Sepal.Width	150	3.06	0.44	2	2.8	3	3.3	4.4

```
Irs %>%
  get_correlation(formula = Sepal.Width ~ Sepal.Length) %>%
  kable(caption =
          '\\label{tab:T4cor} Correlation Coefficient bewteen Sepal.Width and Sepal.Length',
        booktabs = TRUE, format = "latex") %>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 9: Correlation Coefficient bewteen Sepal.Width and Sepal.Length

correlation
-0.1175698

```
ggplot(Irs, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +
  geom_point() +
  labs(x = "Sepal length (in centimetres)", y = "Sepal width (in centimetres)",
       color = "Species") +
  geom_smooth(method = "lm", se = FALSE)
```




```
int.model <- lm(Sepal.Width ~ Sepal.Length * Species, data = Irs)
get_regression_table(int.model)  %>%
  kable(caption =
          '\\label{tab:T4reg} Estimated Coefficients from the fitted model
        Sepal.Width = Sepal.Length . Species', booktabs = TRUE, format = "latex") %>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

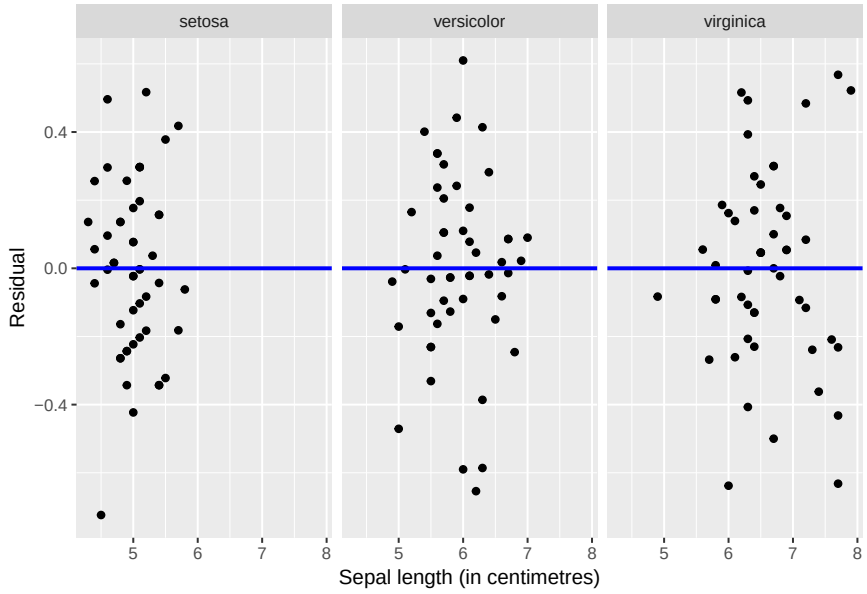
Table 10: Estimated Coefficients from the fitted model Sepal.Width = Sepal.Length . Species

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-0.569	0.554	-1.028	0.306	-1.664	0.525
Sepal.Length	0.799	0.110	7.235	0.000	0.580	1.017
Speciesversicolor	1.442	0.713	2.022	0.045	0.032	2.851
Speciesvirginica	2.016	0.686	2.938	0.004	0.660	3.372
Sepal.Length:Speciesversicolor	-0.479	0.134	-3.582	0.000	-0.743	-0.215
Sepal.Length:Speciesvirginica	-0.567	0.126	-4.490	0.000	-0.816	-0.317

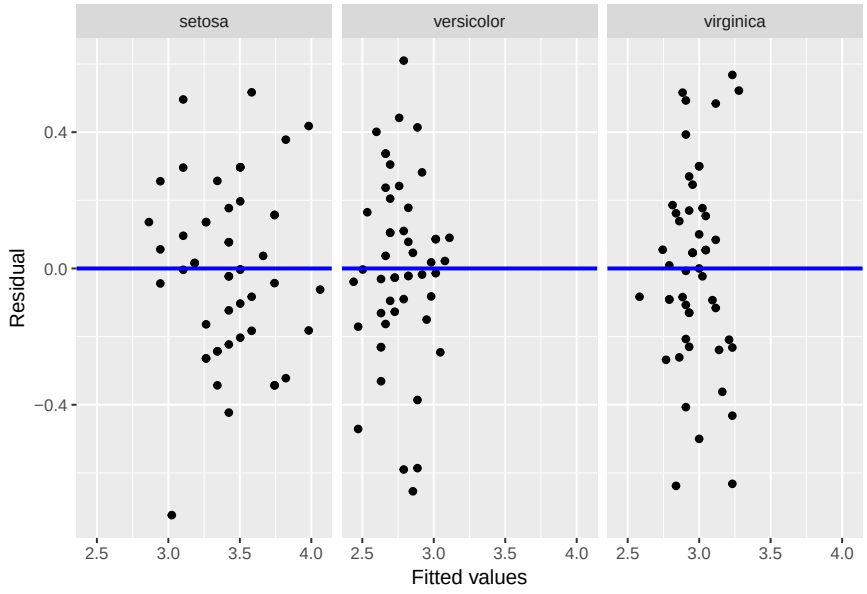
```
regression.points <- get_regression_points(int.model)

ggplot(regression.points, aes(x = Sepal.Length, y = residual)) +
```

```
  geom_point() +
  labs(x = "Sepal length (in centimetres)", y = "Residual") +
  geom_hline(yintercept = 0, col = "blue", size = 1) +
  facet_wrap(~ Species)
```




```
ggplot(regression.points, aes(x = Sepal.Width_hat, y = residual)) +
  geom_point() +
  labs(x = "Fitted values", y = "Residual") +
  geom_hline(yintercept = 0, col = "blue", size = 1) +
  facet_wrap(~ Species)
```




```
ggplot(regression.points, aes(x = residual)) +
  geom_histogram(color = "white") +
  labs(x = "Residual") +
  facet_wrap(~ Species)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```