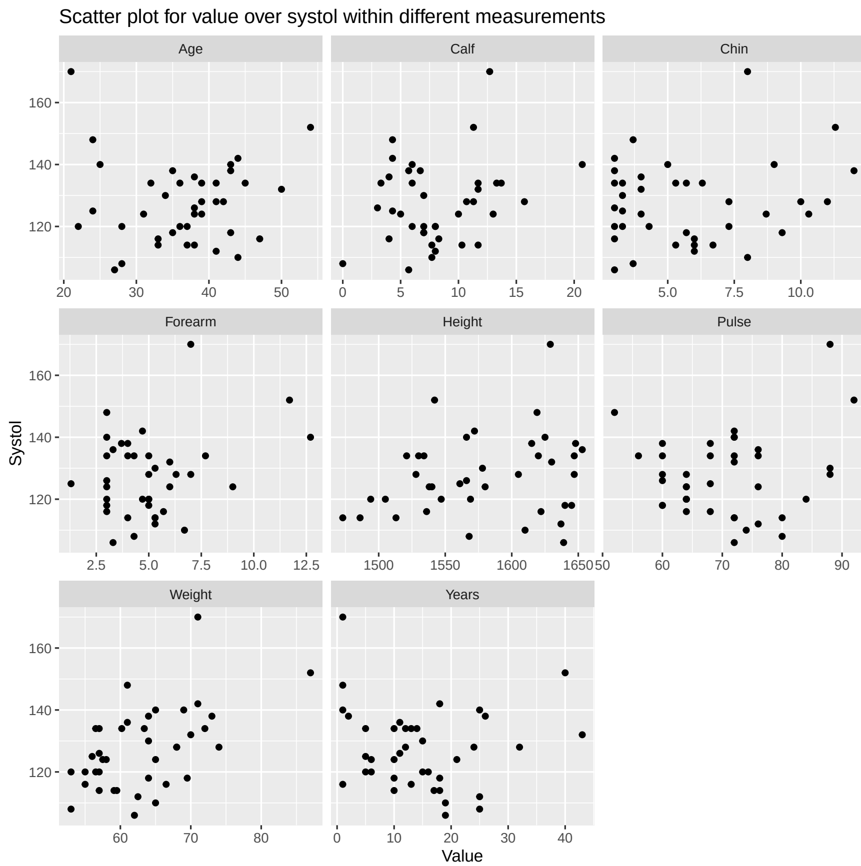
Assignment 5: Under (blood) pressure

JamesJeong_MinsuKang

2022-07-06

Exercise 1

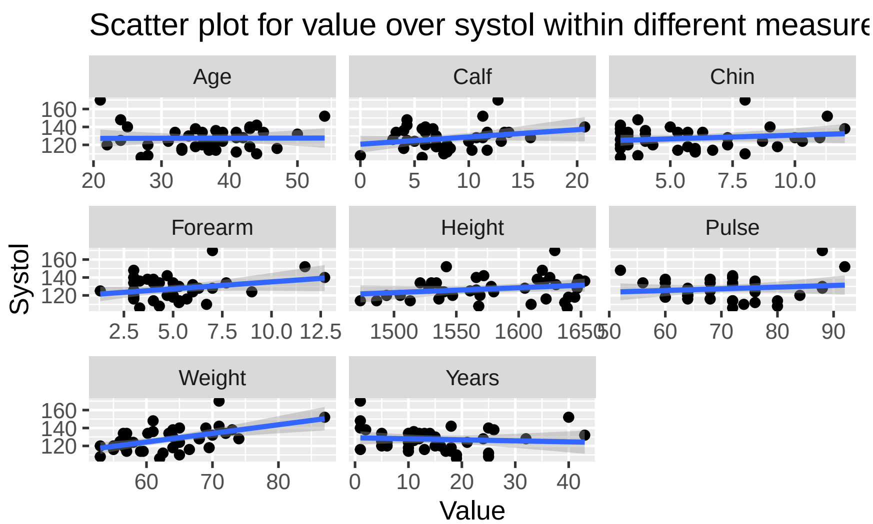
```
blood_pressure %>%
  pivot_longer(cols = Age:Pulse, names_to = "measurement", values_to = "value") %>%
  ggplot() +
    geom_point(mapping = aes(x = value, y = Systol)) +
    facet_wrap(~ measurement, scales = "free_x") +
  labs(title = "Scatter plot for value over systol within different measurements",
       x = "Value",
       y = "Systol"
       )
```



Exercise 2

```
blood_pressure %>%
  pivot_longer(cols = Age:Pulse, names_to = "measurement", values_to = "value") %>%
  ggplot() +
    geom_point(mapping = aes(x = value, y = Systol)) +
    facet_wrap(~ measurement, scales = "free_x") +
    geom_smooth(mapping = aes(x = value, y = Systol), method = "lm") +
   labs(title = "Scatter plot for value over systol within different measurements",
       x = "Value",
       y = "Systol"
       )
```

```
## `geom_smooth()` using formula 'y ~ x'
```



i. Years variable have negative correlation with Systol since it is a negative slope.
ii. Weight.

Exercise 3

```r
blood_pressure_updated <- blood_pressure %>%
  mutate(urban_frac_life = Years / Age)
```

Exercise 4

```r
systol_urban_frac_model <- lm(Systol ~ urban_frac_life, data = blood_pressure_updated)
```

Exercise 5

```r
systol_urban_frac_model %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	133.49572	4.038011	33.059770	0.0000000
urban_frac_life	-15.75182	9.012962	-1.747686	0.0888139

```r
systol_urban_frac_model %>%
  glance() %>%
  select(r.squared:sigma)
```

r.squared	adj.r.squared	sigma
0.0762564	0.0512904	12.76966

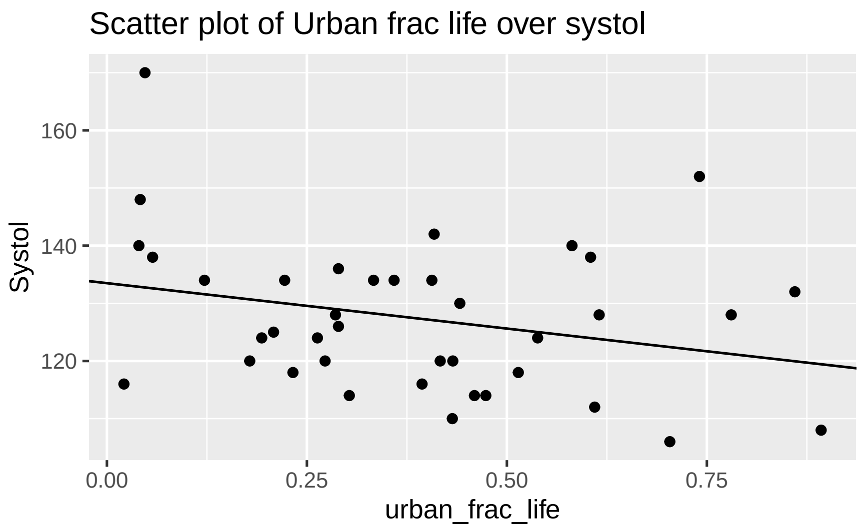
Exercise 6

```
systol_urban_frac_df <- blood_pressure_updated %>%
  add_predictions(systol_urban_frac_model) %>%
  add_residuals(systol_urban_frac_model)
```

i. "pred" is the name of the column that holds response (y) values predicted by the model.
ii. "resid" is the name of the column that holds the residuals for each observation.

Exercise 7

```
ggplot(systol_urban_frac_df) +
  geom_point(mapping = aes(x = urban_frac_life, y = Systol)) +
  geom_abline(slope = systol_urban_frac_model$coefficients[2], intercept = systol_urban_frac_mo
   labs(title = "Scatter plot of Urban frac life over systol",
       x = "urban_frac_life",
       y = "Systol"
       )
```



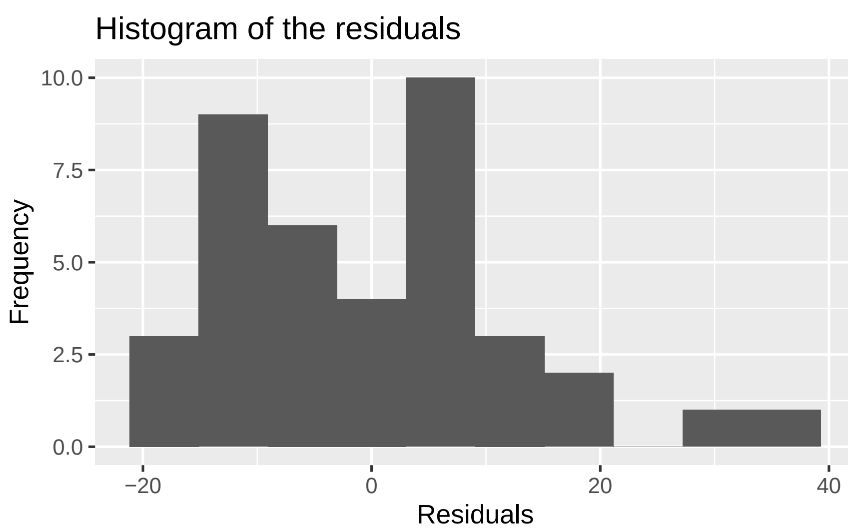
- The model is linear as it is a straight line.

Exercise 8

- Condition of constant variability is that the arrangement of the residuals should all be roughly be similarly distributed above and below the line. However, the graph does not seem to be so; therefore, it violates the third condition resulting to be unreliable.

Exercise 9

```
systol_urban_frac_df %>%
  ggplot() +
  geom_histogram(mapping = aes(x=resid), bins=10) +
   labs(title = "Histogram of the residuals",
       x = "Residuals",
       y = "Frequency"
       )
```



i. The distribution is right-skewed based on the bin width of 10 which makes it smoother to see whether the shape of the distribution is skedwed or not.
ii. As the distribution shows skewness rather than being normal, it violates the second condition.

Exercise 10

```
systol_weight_model <- lm( Systol ~ Weight, data=blood_pressure_updated)
weight_r <- systol_weight_model %>% glance() %>% select(r.squared)
urban_r <- systol_urban_frac_model %>% glance() %>% select(r.squared)
 weight_r > urban_r
```

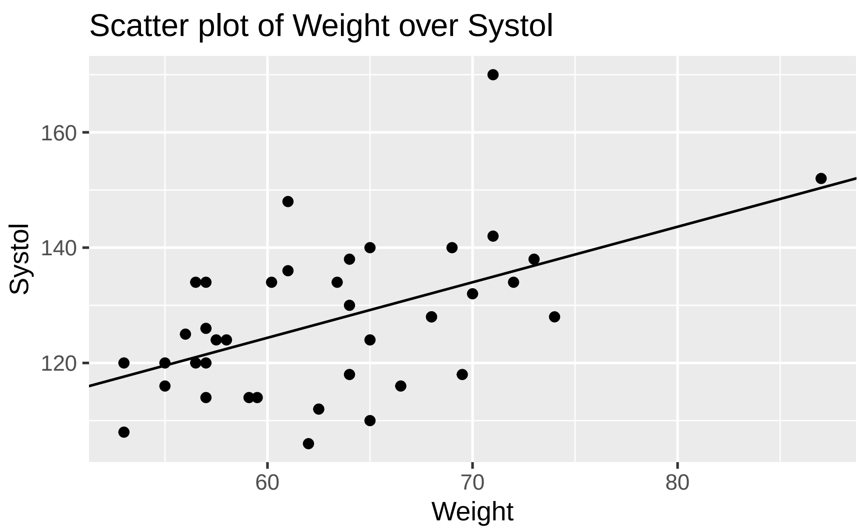
```
##      r.squared
## [1,]      TRUE
```

- The new model perform better than the previous model as it has higher r.squared value.

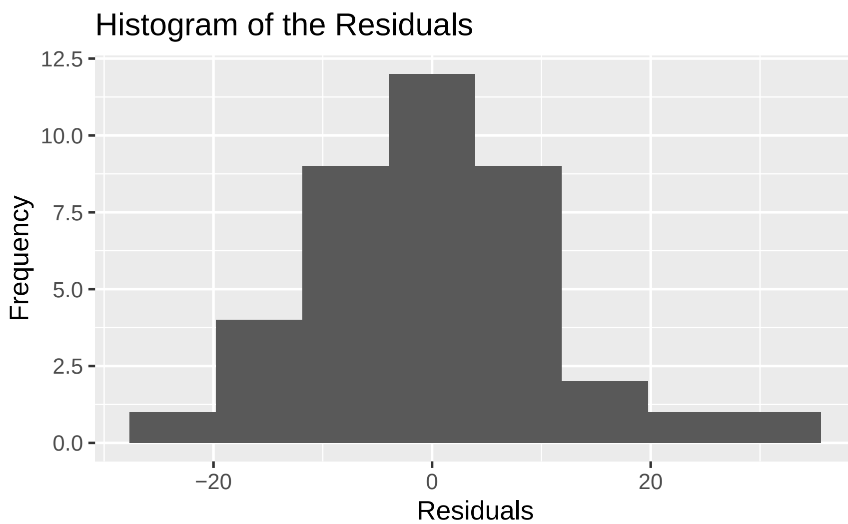
Exercise 11

```
systol_weight_df <- blood_pressure_updated %>%
  add_predictions(systol_weight_model) %>%
  add_residuals(systol_weight_model)
```

```
ggplot(systol_weight_df) +
  geom_point(mapping = aes(x = Weight, y = Systol)) +
  geom_abline(slope = systol_weight_model$coefficients[2], intercept = systol_weight_model$coe
   labs(title = "Scatter plot of Weight over Systol",
       x = "Weight",
       y = "Systol"
       )
```




```
systol_weight_df %>%
  ggplot() +
  geom_histogram(mapping = aes(x=resid), bins=8) +
  labs(title = "Histogram of the Residuals",
       x = "Residuals",
       y = "Frequency")
```



- Although the scatter plot meets the first condition by being linear, it violates the third condition as the residuals are quite fluctuating over the lines. Furthermore the histogram shows that the graph is nearly normal which means that the new model can somewhat be reliable.

Exercise 12

```
systol_weight_model %>%
  glance()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.2718207	0.2521402	11.33764	13.81166	0.0006654	1	-149.009	304.0181	309.0088	4756.056	37	39

```
systol_urban_frac_model %>%
  glance()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.0762564	0.0512904	12.76966	3.054406	0.0888139	1	-153.6478	313.2957	318.2864	6033.372	37	39

- Considering the fact that the second model has 0.27 of r.squared value and the first one has 0.076, it could simply be compared that the second model is better in terms of how well they explain the data as it is within the level of 'okay' for the r.squared values while the first model does not even reach to the 'weak' and stays where it cannot even describe the correlation. Furthermore, the second model has passed the second condition while the first models fails to pass any. Therefore, it could be said that the second model is more reliable compared to the first model.