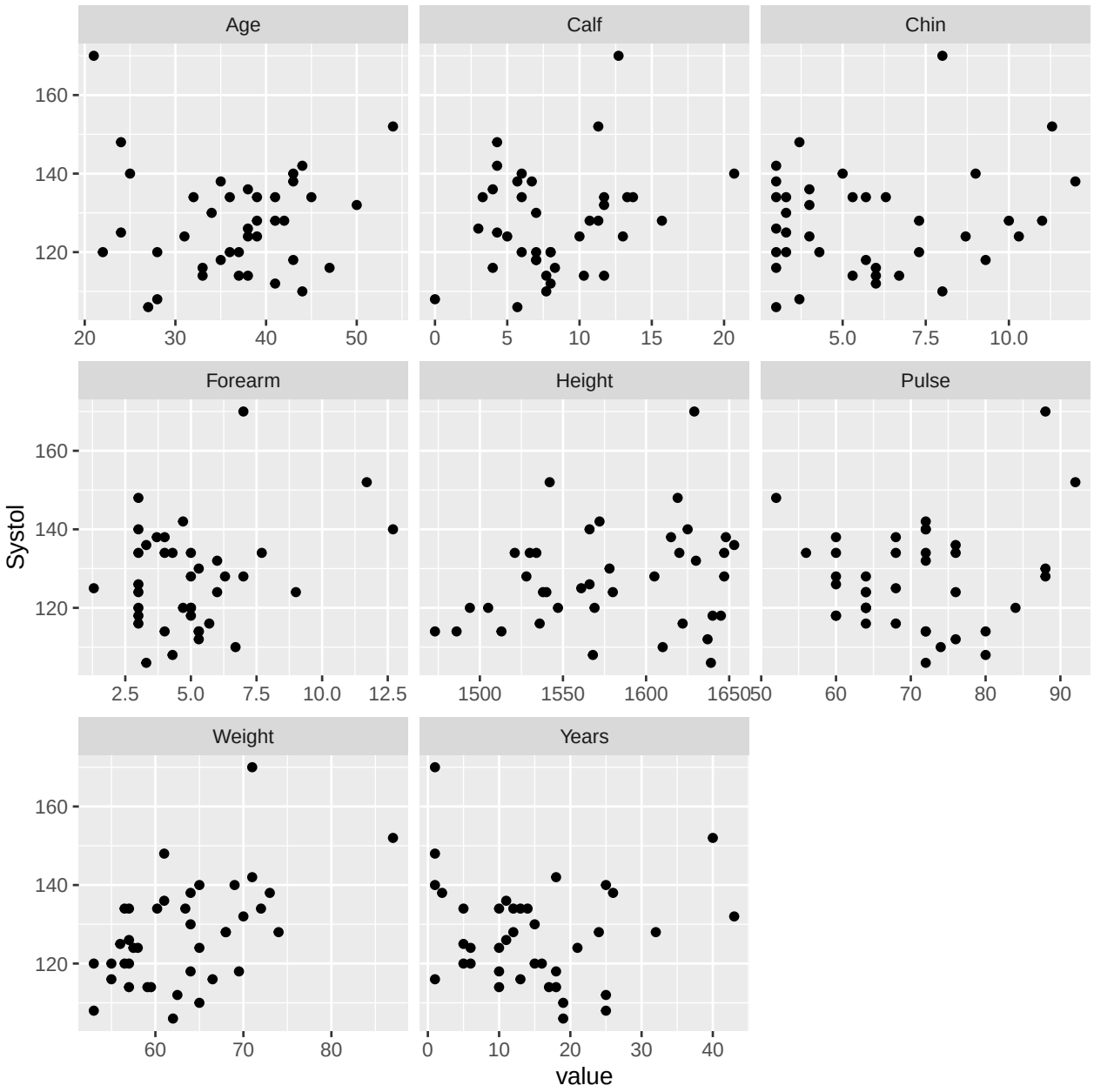
Assignment 5: Under (blood) pressure

Alyssa Garcia

2022-07-10

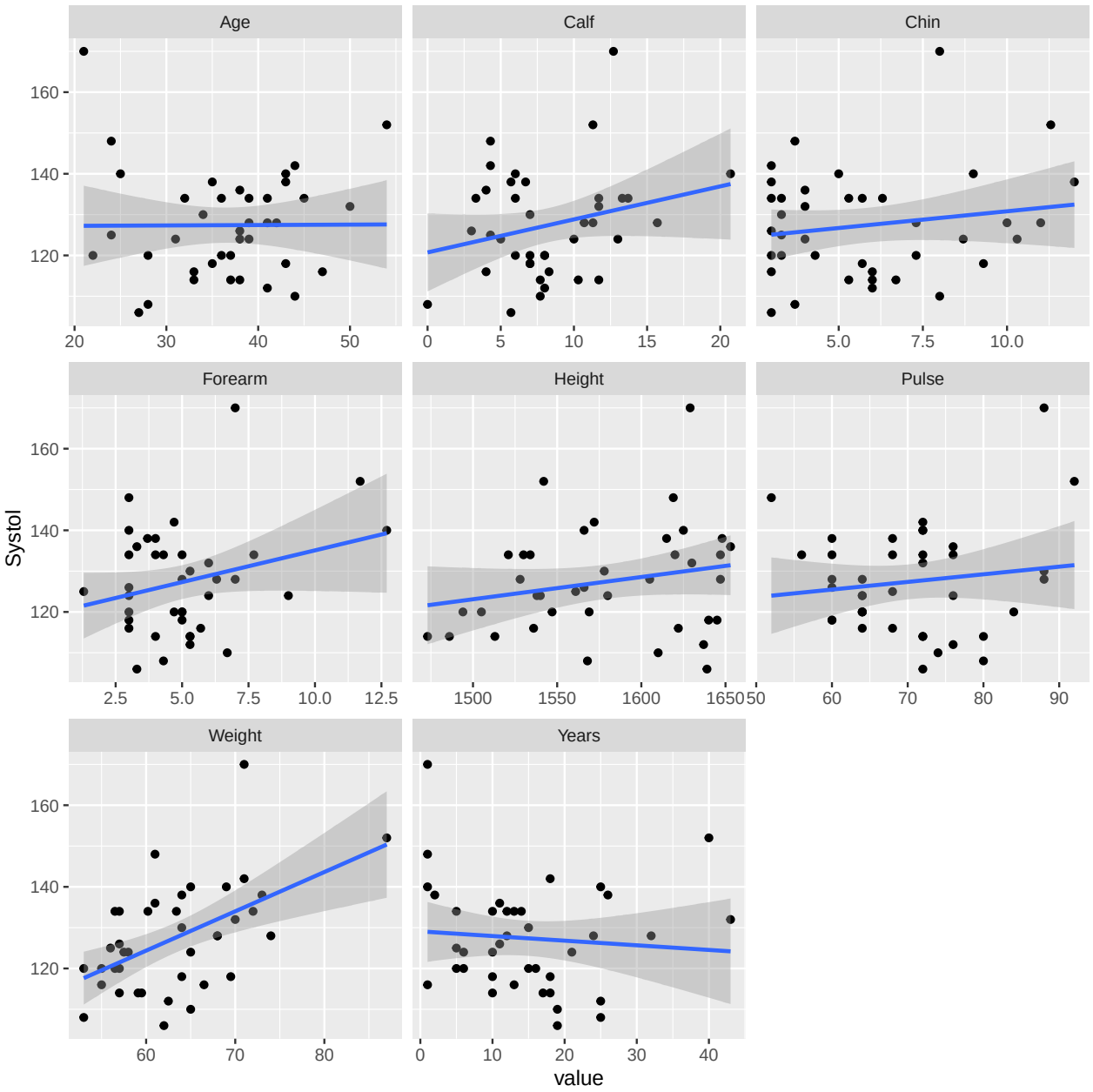
Exercise 1

```
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")
```




```
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")
```

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



Exercise 2

i. The correlation between the variables "Years" and "Systol" seems to have a slightly negative correlation.

ii. The variable "Weight" shows a moderate to strong positive correlation to "Systol".

Exercise 3

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

```
  )
```

Exercise 4

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

Exercise 5

```
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

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

r.squared
0.0762564

Exercise 6

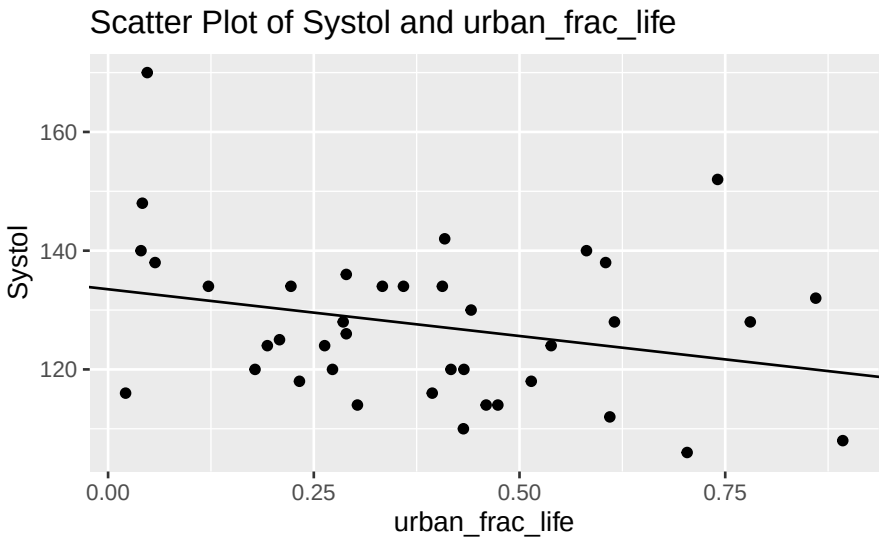
```
systol_urban_frac_df <- blood_pressure_updated %>%
  add_predictions(systol_urban_frac_model) %>%
  add_residuals(systol_urban_frac_model)
```

i. The name of the column that holds the response (y) values predicted by the model is "pred".

ii. The name of the column that holds the residuals for each observation is "resid".

Exercise 7

```
ggplot(systol_urban_frac_df) +
  geom_point(mapping = aes(x = urban_frac_life, y = Systol)) +
  geom_abline(slope = -15.75182, intercept = 133.5) +
  labs(
    title = "Scatter Plot of Systol and urban_frac_life",
    x = "urban_frac_life",
    y = "Systol"
  )
```



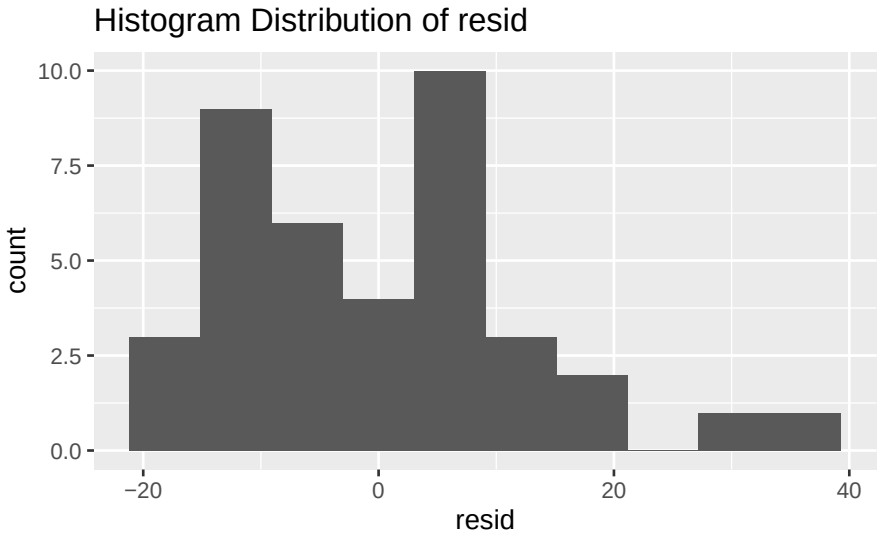
The model meets the first condition, linearity, as the points fall more or less along the line. There is a negative correlation between the two variables.

Exercise 8

The variability, in Exercise 8, does not look reasonably constant all the way along the line. There are mostly large residuals, while most of the residuals are relatively far. This means that the graph does not satisfy the linear model's assumption of constant variability of residuals.

Exercise 9

```
systol_urban_frac_df %>%
  ggplot() +
  geom_histogram(
    mapping = aes(x = resid),
    bins = 10
) +
  labs(
    title = "Histogram Distribution of resid",
    x = "resid"
  )
```



i. The shape of this distribution is right-skewed and bimodal, with a visible outlier.

ii. This histogram suggests that the nearly normal residuals condition is not met, because it lacks the "bell-shaped" curve, meaning the data is not symmetric. This histogram also displays a bimodal distribution, which does not meet the "singular peak" requirement for a roughly normal distribution.

Exercise 10

```
systol_weight_model <- lm(Systol ~ Weight, data = blood_pressure_updated)
```

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

r.squared
0.0762564

```
 systol_weight_model %>%
  glance() %>%
  select(r.squared)
```

r.squared
0.2718207

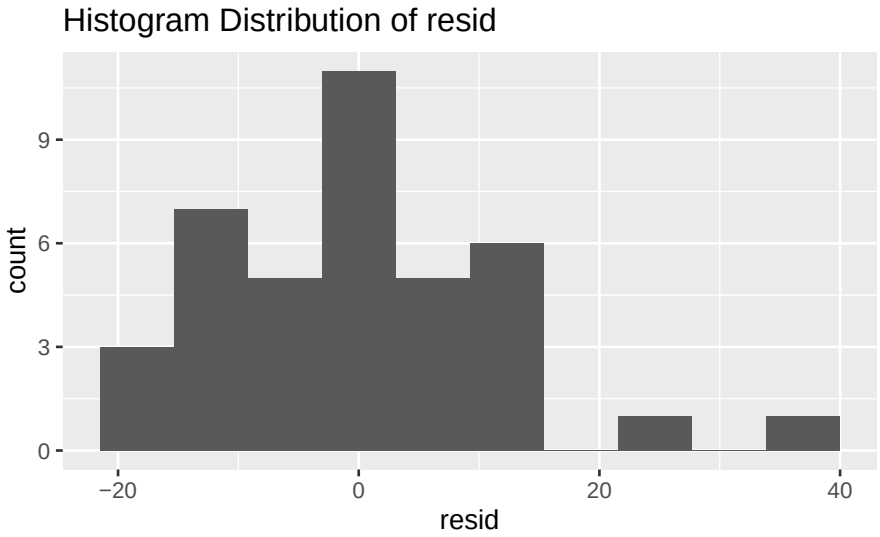
Yes, systol_weight_model seems to predict Systol better than urban_frac_life, because it is closer to 1.

Exercise 11

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

```
ggplot(systol_weight_model) +
  geom_point(mapping = aes(x = Weight, y = Systol)) +
  geom_abline(slope = 0.9628622 , intercept = 66.59687) +
  labs(
    title = "Scatter Plot of Weight and Systol",
    x = "Weight",
    y = "Systol"
  )
```


```
systol_weight_df %>%
  ggplot() +
  geom_histogram(
    mapping = aes(x = resid),
    bins = 10
) +
  labs(
    title = "Histogram Distribution of resid",
    x = "resid"
  )
```



We can conclude that this model is more reliable since more conditions are met, than the first model. This model meets the first condition, linearity, as the points fall along the line. The constant variation condition is met, as the variation in points look relatively constant. There are a few large residuals, but most are small. The nearly normal residuals condition is not met, as the histogram is right-skewed, which does not satisfy the condition for the "bell-shaped" curve.

Exercise 12

The second model explains the data better since two of the three conditions were met and R^2 was closer to 1 than the first model. We can see that the second model has a positive linearity with a constant variation, which are not conditions met with the first model. However, in both models we can see that nearly normal residuals condition is not met, because both modals are right-skewed, meaning it does not have a Normal distribution.