Dem 11.36

Muhammad Reza Fahlevi

Problems

The dataset consists of variable relating to blood pressure of 15 Peruvians $(n = 15)$ who have moved from rural, high-altitude areas to urban, lower altitude areas. The variables in this data sets are: Systolic blood pressure (Y), weight (X_1), height (X_2), and pulse.

Weight_kg <dbl>	Heigh_mm <dbl>	Pulse_per_minute <dbl>	Systolic_pressure_mmHg <dbl>
71.0	1629	88	170
56.5	1569	64	120
56.0	1561	68	125
61.0	1619	52	148
65.0	1566	72	140
62.0	1639	72	106
53.0	1494	64	120
53.0	1568	80	108
65.0	1540	76	124
57.0	1530	60	134

1-10 of 15 rows Previous **1** 2 Next

i. Determine if weight and systolic blood pressure are in a linear relationship, that is, test whether $H_0 : \beta_{1.0} = 0$, where β_1 is the slope of the regressor variable.
ii. Perform a lack-of-fit test to determine if linear relationship between weight and systolic blood pressure is adequate. Draw conclusions.
iii. Determine if pulse rate influences systolic blood pressure in a linear relationship. Which regressor variable is the better predictor of the systolic blood pressure?

Demonstrandum

Let

$$x_1 \stackrel{def}{=} \text{weight}$$
$$x_2 \stackrel{def}{=} \text{height}$$
$$p \stackrel{def}{=} \text{pulse}$$

The simple linear regression for given data $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$ is defined as

$$Y = \beta_0 + \beta_1 x$$

Y is estimated by

$$\hat{y} = b_0 + b_1 x$$

where b_0 and b_1 are regression's coefficient estimator for β_0 and β_1, respectively. These estimator is computed as follows,

$$

$$\begin{aligned} b_0 &= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \qquad \text{and} \\ b_1 &= \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} \end{aligned}$$

$$

Take a Glimpse to the Data

The following output is the summarize of the given data

```
##    Weight_kg        Heigh_mm    Pulse_per_minute Systolic_pressure_mmHg
##  Min.   :53.00   Min.   :1486   Min.   :52.0     Min.   :106.0         
##  1st Qu.:56.75   1st Qu.:1550   1st Qu.:62.0     1st Qu.:117.0         
##  Median :62.00   Median :1569   Median :68.0     Median :124.0         
##  Mean   :61.51   Mean   :1580   Mean   :69.6     Mean   :127.4         
##  3rd Qu.:65.00   3rd Qu.:1626   3rd Qu.:74.0     3rd Qu.:136.0         
##  Max.   :71.00   Max.   :1648   Max.   :88.0     Max.   :170.0         
```

Hypothesis Testing on the Slope for Regressor Variable x_1

The hypothesis that's being tested is the slope of the regression line $\hat{y} = \beta_0 + \beta_1 x$

$$\begin{cases} H_0 : \beta_{1.0} = 0 \\ H_1 : \beta_{1.1} \neq 0 \end{cases}$$

In order to make decision with regards to the hypothesis, the analysis of variances is performed.

Step 1. Construct the linear model for X_1,

```
## 
## Call:
## lm(formula = "Systolic_pressure_mmHg~Weight_kg", data = dem_11_36)
## 
## Coefficients:
## (Intercept)    Weight_kg  
##      44.398        1.349  
```

then for $\hat{y} = b_0 + b_1 x_1$,

$$\hat{y} = 44.398 + 1.349x_1$$

Step 2. Compute the *one-way ANOVA*

```
anova(lmodels_X1)
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
Weight_kg	1	805.8589	805.8589	3.364495	0.08959975
Residuals	13	3113.7411	239.5185	NA	NA

2 rows

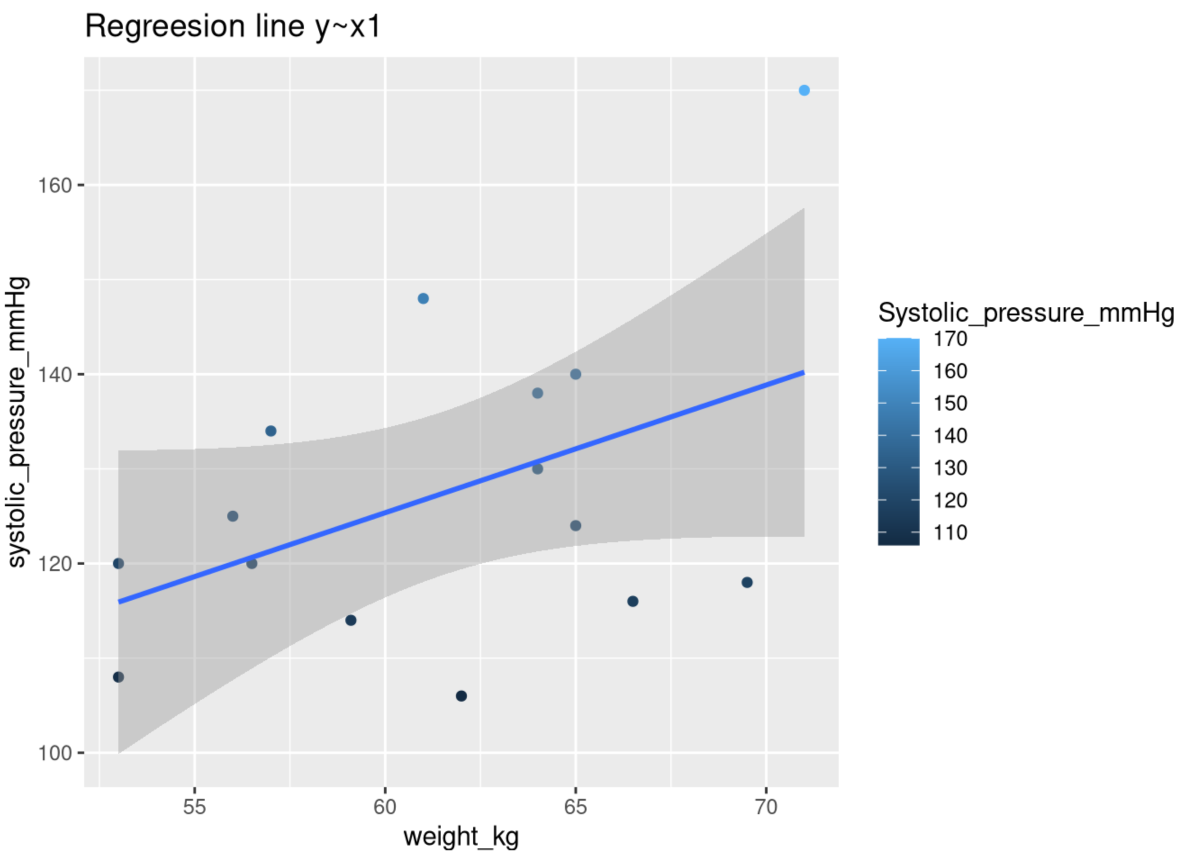
Step 3. Conclusion. Let $\alpha = 0.05$, the critical value for $f_\alpha(1, n-2)$

```
## [1] 4.667193
```

According to *one-way ANOVA* tables, the computed f-values $f_{\text{reg}} = 3.3645$. Hence, $f_{\text{reg}} < f_{0.05}(1, 13)$. Therefore, the hypothesis testing lead to *do not reject $H_0 : \beta_{1.0} = 0$ at $\alpha = 0.05$ level of significance*.

ANOVA for Testing Linearity of Regression $Y \sim x_1$

In order to determine the linear relationship between weight (x_1) and the systolic blood pressure (Y) is adequate or not, the *ANOVA for testing linearity of regression* is performed.



Step 1. Compute the sum of square SSR, SSE, SSE(pure) and lack-of-fit. The SSR is computed as follows,

$$\begin{aligned}\text{SSR} &= \sum_{i=1}^{n} (\hat{y_i} - \bar{y})^2 \\ &= 805.86\end{aligned}$$

on 1 degrees of freedom. The SSE is computed as follows,

$$
\begin{aligned}
\text{SSE} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= 3113.74
\end{aligned}
$$

on $n-2=13$ degrees of freedom. For data which contain k-groups, the $\text{SSE}(\text{pure})$ is computed as follows,

$$
\text{pure error } := \sum_{i=1}^{k}\sum_{j=1}^{n}(y_j^{(i)} - \bar{y}^{(i)})^2
$$

where

$$
\bar{y}^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_j^{(i)}
$$

for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$. In order to compute the pure error, the data must be group by the regressor variable (x_1).

Weight_kg <dbl>	Systolic_pressure_mmHg <dbl>
53.0	120
53.0	108
56.0	125
56.5	120
57.0	134
59.1	114
61.0	148
62.0	106
64.0	130
64.0	138

1-10 of 15 rows Previous **1** 2 Next

then

Weight_kg <dbl>	mean_y_ith <dbl>	sum_sqr_y_ith <dbl>
53.0	114	72
56.0	125	0
56.5	120	0
57.0	134	0
59.1	114	0
61.0	148	0
62.0	106	0
64.0	134	32
65.0	132	128
66.5	116	0

from the tables, $k = 12$ groups, the second column is equals to $\bar{y}^{(i)}$, the third column is the value of

$$\sum_{j=1}^{n_i}(y_j^{(i)} - \bar{y}^{(i)})$$

and the sum of third column is equals to the so called pure error, thus,

$$\text{pure error} := 232.00$$

on $n - k = 3$ degrees of freedom.

The differences between SSE and SSE(pure) is equals to the so called *lack-of-fit*, hence,

$$\begin{aligned}
\text{lack-of-fit} := \text{SSE} - \text{SSE(pure)} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 - \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_j^{(i)} - \bar{y}^{(i)})^2 \\
&= 3113.74 - 232.00 \\
&= 2881.74
\end{aligned}$$

on $k - 2 = 10$ degrees of freedom.

Step 2. Compute the mean square for $\text{SSE}, \text{SSE(pure)}$ and lack-of-fit. The mean square for SSE,

$$\begin{aligned}
s^2 &= \frac{\text{SSE}}{n-2} = \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2} \\
&= \frac{3113.74}{15-2} = \frac{3113.74}{13} \\
s^2 &= 239.52
\end{aligned}$$

The mean square for *pure-error*,

$$\begin{aligned}
s^2_{\text{pure}} &= \frac{\text{pure error}}{n-k} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{(y_j^{(i)} - \bar{y}^{(i)})^2}{n-k} \\
&= \frac{232.00}{15-12} \\
&= \frac{232.00}{3} \\
s^2_{\text{pure}} &= 77.33
\end{aligned}$$

The mean square for *lack-of-fit*,

$$\begin{aligned}
s^2_{\text{lack-of-fit}} &= \frac{\text{lack-of-fit}}{k-2} = \frac{\text{SSE} - \text{SSE(pure)}}{k-2} \\
&= \frac{1}{k-2}\left\{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 - \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_j^{(i)} - \bar{y}^{(i)})^2\right\} \\
&= \frac{3113.74 - 232.00}{12-2} \\
&= \frac{2881.74}{10} \\
s^2_{\text{lack-of-fit}} &= 288.17
\end{aligned}$$

Step 3. Compute the f-values. For f_{reg}

$$
\begin{aligned}
f_{\text{reg}} &= \frac{\text{SSE}}{s^2_{\text{pure}}} \\
&= \frac{3113.74}{77.33} \\
f_{\text{reg}} &= 10.4206
\end{aligned}
$$

on 1 and $n-2=13$ degrees of freedom, and for $f_{\text{lack-of-fit}}$,

$$
\begin{aligned}
f_{\text{lack-of-fit}} &= \frac{\text{lack-of-fit}}{s^2_{\text{pure}}(k-2)} \\
&= \frac{\text{SSE} - \text{SSE(pure)}}{s^2_{\text{pure}}(k-2)} \\
&= \frac{2881.74}{77.33 \times (12-2)} \\
&= \frac{2881.74}{77.33 \times 10} \\
&= \frac{2881.74}{773.3} \\
f_{\text{lack-of-fit}} &= 3.7264
\end{aligned}
$$

on $k-2=10$ and $n-k=3$ degrees of freedom.

Step 4. Compute the P-values

```
## [1] "P-values for f_regression := 0.00660075471987764"
```

```
## [1] "P-values for f_lack-of-fit :=  0.153221390704174"
```

Step 5. Summarize altogether computation as table of *ANOVA for testing for linearity of regression.*

```
EnvStats::anovaPE(lmodels_X1)
```

	Df <dbl>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
Weight_kg	1	805.8589	805.85886	10.420589	0.0482858
Lack of Fit	10	2881.7411	288.17411	3.726389	0.1532219
Pure Error	3	232.0000	77.33333	NA	NA

3 rows

Step 6 (Conclusion). Let $\alpha = 0.05$, then the critical value for

$$f_\alpha(1, n-2) = f_\alpha(1, 13) = 4.667193$$

, and

$$f_\alpha(k-2, n-k) = f_\alpha(10, 3) = 8.785525$$

From the table of *ANOVA for testing the linearity of regression*, $f_{\text{reg}} > f_\alpha(1,13)$ is **true**, and $f_{\text{lack-of-fit}} > f_\alpha(10,3)$ is **false**. Therefore, there are significant amount of variation accounted for by linear model (*reject* $H_0 : \beta_{1.0} = 0$) and insignificant amount due to lack of fit. Thus, the experimental data do not seem to suggest the need to consider terms higher than first order in the model, and the null hypothesis is not rejected.

ANOVA for Testing Linearity of Regression $Y \sim p$

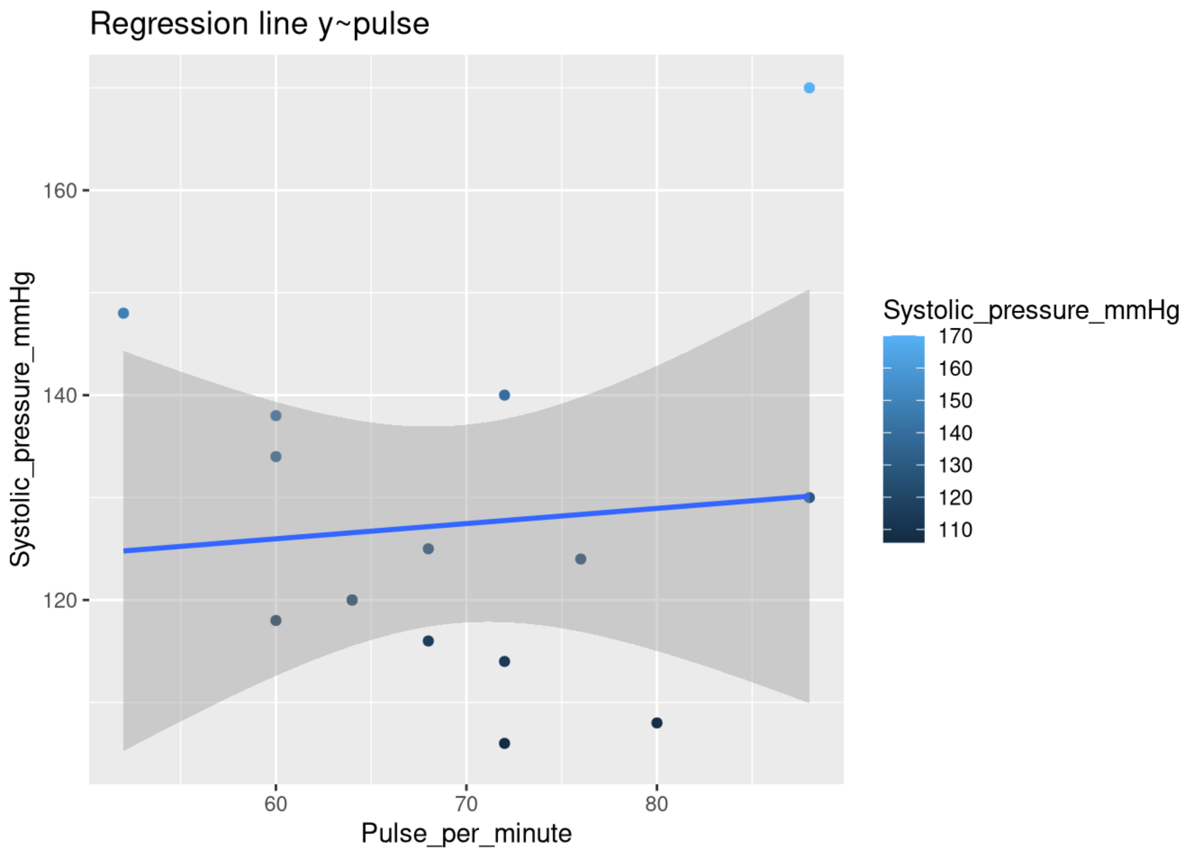
As usual, in order to determine the linear relationship between pulse rate (p) and the systolic blood pressure (Y) is adequate or not, the *Analysis of Variances for testing linearity of regression* is performed.

Step 1. Construct the linear model $\hat{y} = b_0 + b_1 p$

```
##
## Call:
## lm(formula = "Systolic_pressure_mmHg~Pulse_per_minute", data = dem_11_36)
##
## Coefficients:
##      (Intercept)  Pulse_per_minute
##         117.0641            0.1485
```

therefore,

$$\hat{y} = 117.0641 + 0.1485p$$



Step 2. Compute the SSR, SSE, pure-error, and lack-of-fit.

Step 3. Compute the mean square error for SSE, pure-error, and lack-of-fit.

Step 4. Compute the f-value for regression and lack-of-fit.

Step 5. Compute the $P\text{-values}$ for f_{reg} and $f_{\text{lack-of-fit}}$

Step 6. Summarize altogether results into table of *ANOVA for testing linearity of regression*.

```
EnvStats::anovaPE(lmodels_pulse)
```

	Df <dbl>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
Pulse_per_minute	1	33.02735	33.02735	0.1362755	0.7229254
Lack of Fit	6	2190.07265	365.01211	1.5060918	0.3007204
Pure Error	7	1696.50000	242.35714	NA	NA

3 rows

Step 7 (Conclusion). Let $\alpha = 0.05$, recall the critical value for $f_\alpha(1, 13)$ and $f_\alpha(10, 3)$. From the table of ANOVA for testing the linearity of regression, $f_{\text{reg}} > f_\alpha(1, 13)$ is **false**, and $f_{\text{lack-of-fit}} > f_\alpha(10, 3)$ is **false**. Therefore, there are insignificant amount of variation accounted for by linear model (*do not reject* $H_0 : \beta_{1.0} = 0$) and insignificant amount due to lack of fit. Thus, the experimental data do not seem to suggest the need to consider terms higher than the first order.

Hypothesis Testing on the Slope for Regressor Variable x_2

The hypothesis that's being tested is the slope of the regression line $\hat{y} = \beta_0 + \beta_1 x_2$, such that

$$\begin{cases} H_0 : \beta_{1.0} = 0 \\ H_1 : \beta_{1.1} \neq 0 \end{cases}$$

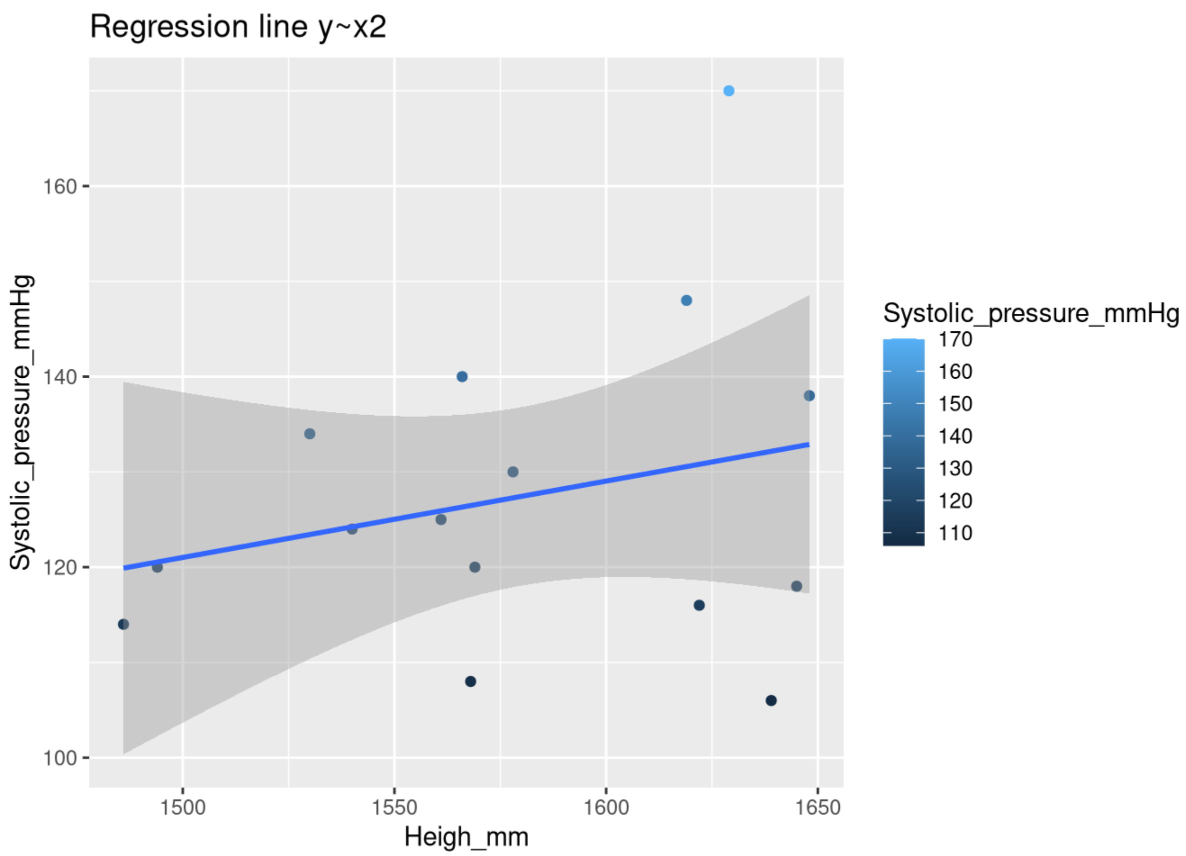
in order decide which hypothesis should be chosen, the *one-way analysis of variances* is used.

Step 1. Construct the linear model $\hat{y} = b_0 + b_1 x_2$.

```
## 
## Call:
## lm(formula = "Systolic_pressure_mmHg~Heigh_mm", data = dem_11_36)
## 
## Coefficients:
## (Intercept)     Heigh_mm  
##     0.80328      0.08014
```

therefore,

$$\hat{y} = 0.80328 + 0.08014 x_2$$



Step 2. Compute the SSR and SSE.

Step 3. Compute the f-values.

Step 4. Compute the P-values.

Step 5. Summarize altogether computation as table of *One-way ANOVA*.

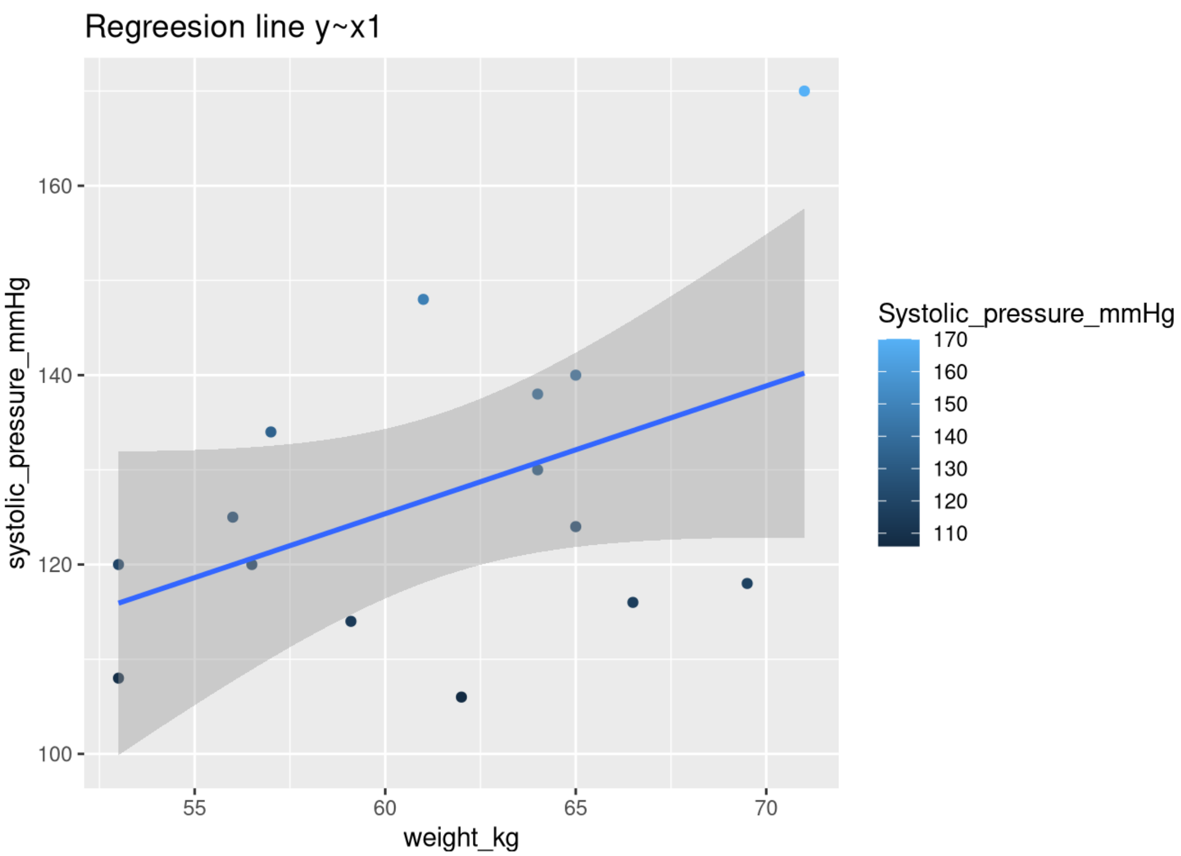
	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
Heigh_mm	1	251.6066	251.6066	0.891737	0.3622279
Residuals	13	3667.9934	282.1533	*NA*	*NA*

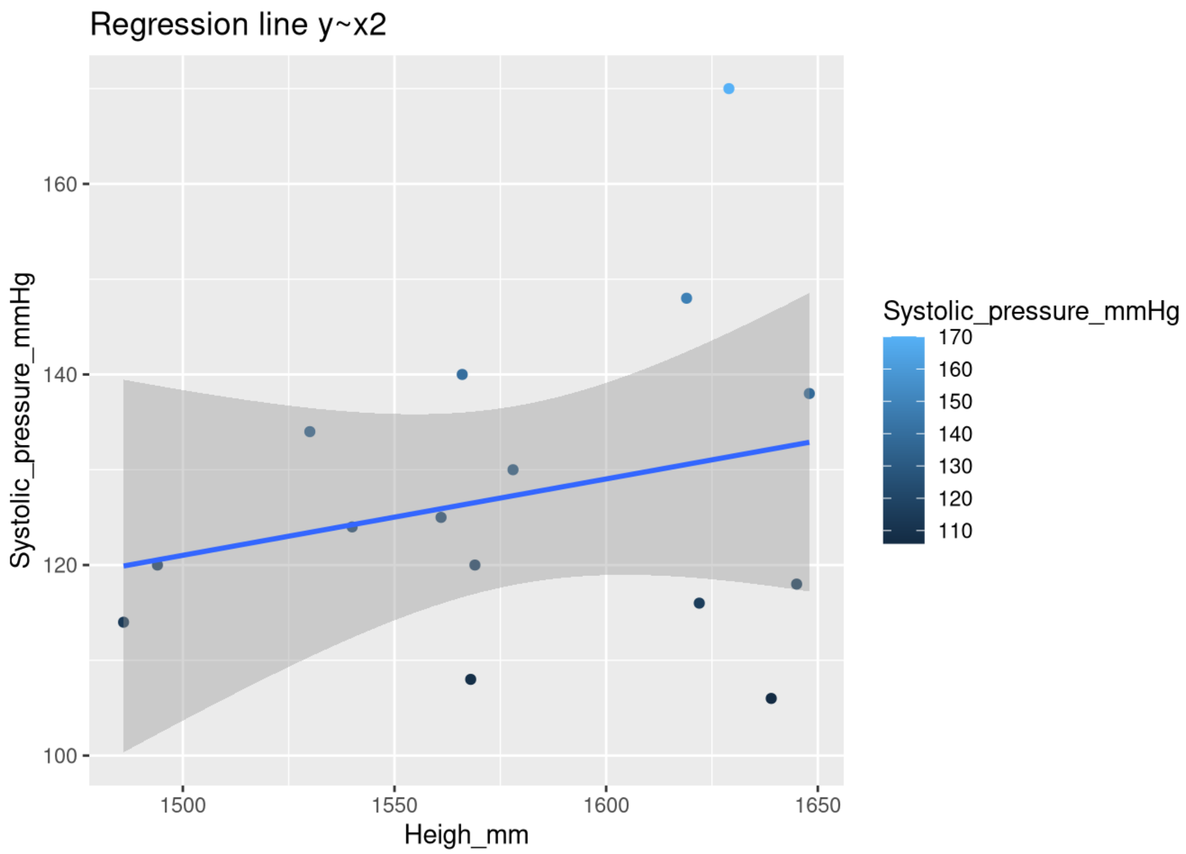
2 rows

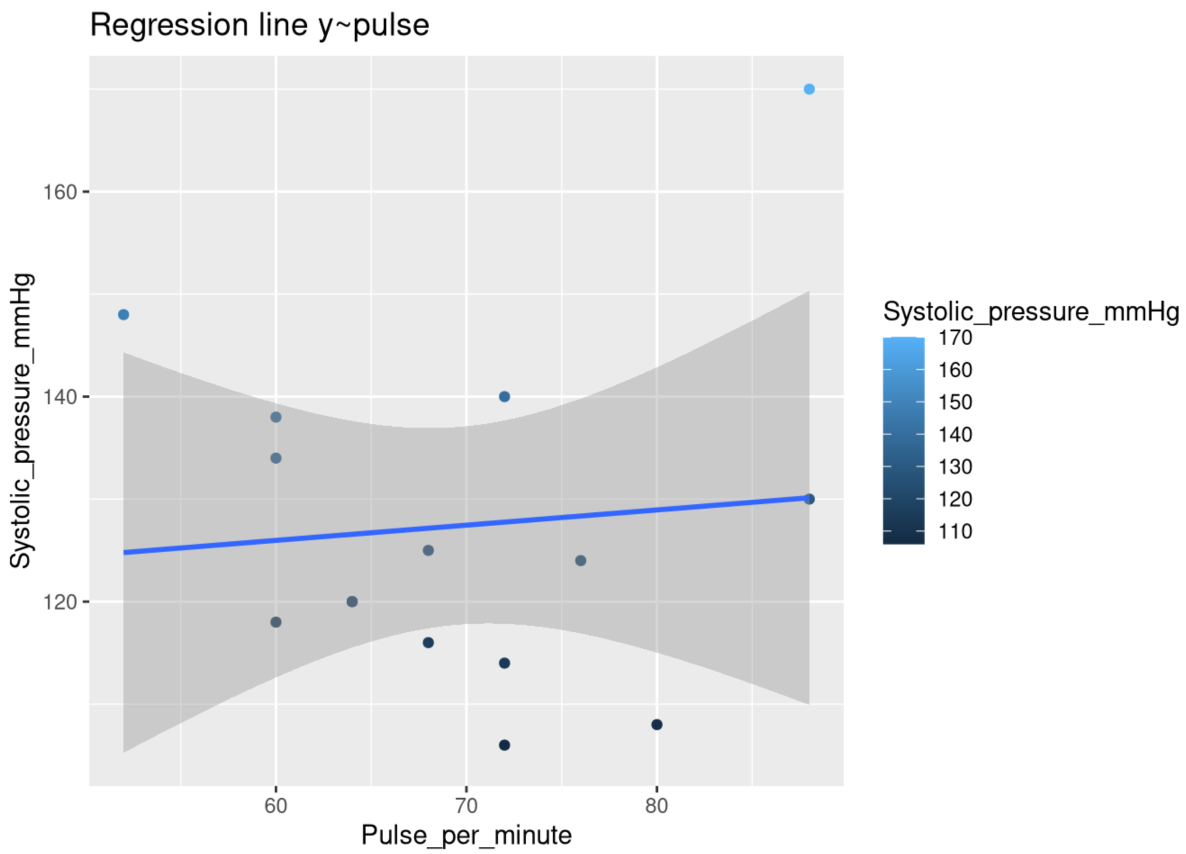
Step 6 (Conclusion). Let $\alpha = 0.05$, recal the critical value for $f_\alpha(1, n-2)$ or $f_\alpha(1, 13)$. From the table of one-way ANOVA, $f\text{-values} > f_\alpha(1, 13)$ is **false**. Therefore, there are insignificant amount of variation accounted for by linear model (*do not reject* $H_0 : \beta_{1.0} = 0$).

Determine Which Regressor Variable is the Better Predictor of Y

Visualize altogether linear models







With regards to which regressor variable is better, from the computed $P\text{-values}$ after performed ANOVA for variable x_1, x_2 and p, for regressor variable x, $P\text{-values} \approx 0$, then x is a better regressor variable for Y. Therefore, variable weight (x_1) is better regressor variable for predictor of the systolic blood pressure (Y), with $P\text{-values} = 0.04829$.