Lab 10 - Linear Regression 2

Nick Sumpter (Edited by Eddie-Williams Owiredu & Guy Twa)

2023-09-20

- Loading Packages and Data
- Assumptions of Simple Linear Regression
    - Independence
    - Linearity
    - Homoscedasticity of Model Residuals
    - Normality of the Model Residuals
- Assumptions of a Multiple Linear Regression
    - Multicollinearity
- Independent Practice

Loading Packages and Data

For this lab, we will need the `tidyverse`, `pastecs`, and `car` packages. We will also be using the same Framingham Heart Study dataset from last lab (`framingham.RData`).

```
library(pastecs)
library(car)
library(tidyverse)

theme_set(theme_bw())

setwd("/Users/eddie-williamsowiredu/Desktop/grd770_23/Lab10")

load("framingham.RData")

source("functions.R")
```

Assumptions of Simple Linear Regression

Let's review our simple linear regression model of systolic blood pressure (sysBP) as a function of age alone. We will recalculate this using `lm`.

```
sysBPvsAge <- lm(formula = sysBP ~ age, data = fhs)

summary(sysBPvsAge)
```

```
##
## Call:
## lm(formula = sysBP ~ age, data = fhs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.293  -9.533   1.273   5.580  57.010
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  89.8743    12.9313   6.950 8.77e-09 ***
## age           0.8853     0.2705   3.272  0.00198 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.57 on 48 degrees of freedom
## Multiple R-squared:  0.1824, Adjusted R-squared:  0.1653
## F-statistic: 10.71 on 1 and 48 DF,  p-value: 0.001981
```

We know that we have a significant main effect of age on systolic blood pressure. But does our model pass the assumptions of simple linear regression?

There are 4 assumptions that apply to simple linear regression:

- Independence
- Linearity
- Homoscedasticity of the model residuals
- Normality of the model residuals

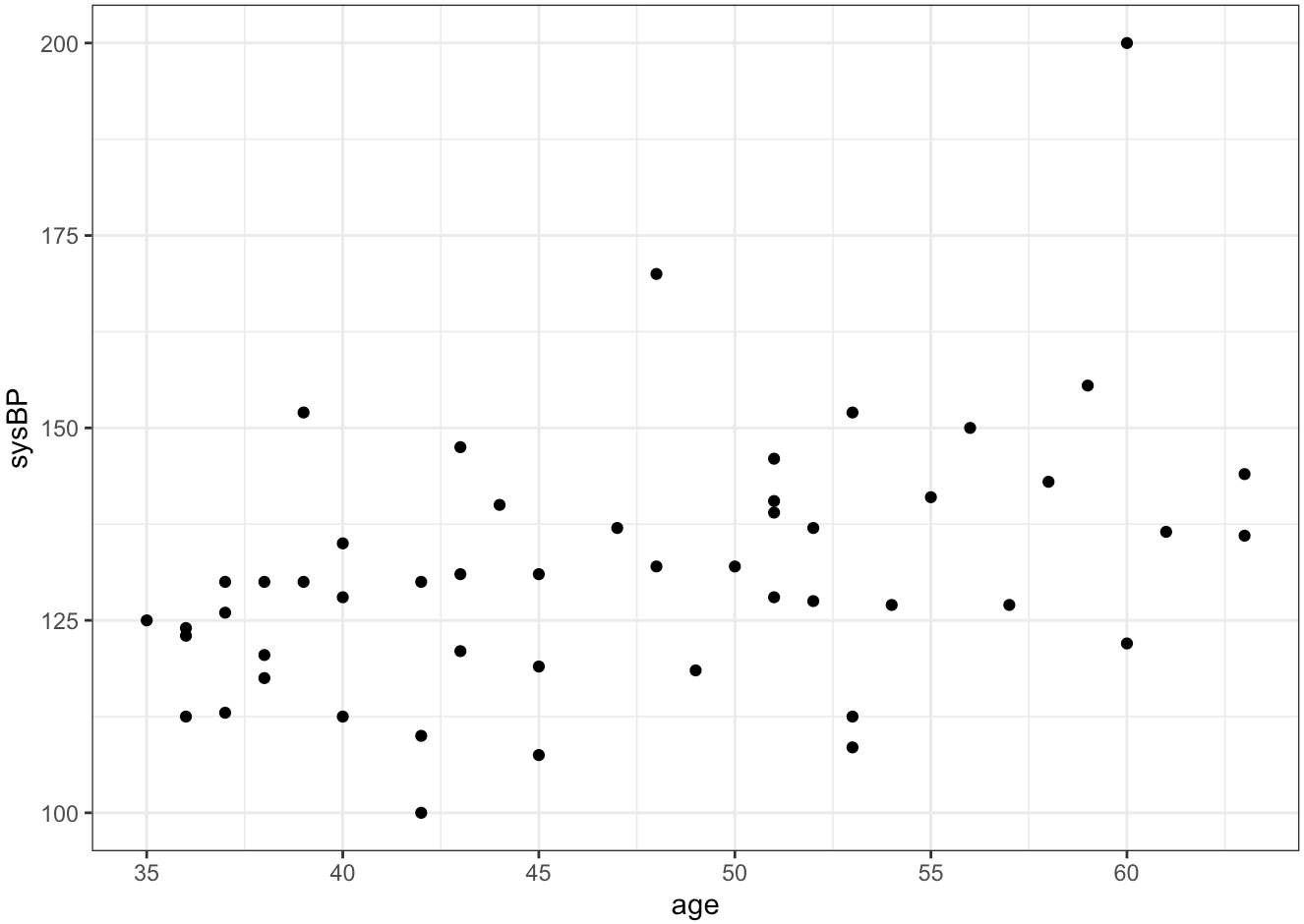
Independence

There are no statistical tests for the independence assumption. However, from the data and the way it is collected, you can determine whether each sample is independent from the other samples in the dataset. You can say something along the lines of: "The independence assumption is met because the data for each sample was collected independently from the other samples". In our case, we know that all 50 rows represent 50 different individuals and we just have to assume that these 50 individuals are independent of one another (i.e. ideally not related).

Linearity

As for correlation analyses, we assess linearity using a scatterplot.

```
ggplot(data = fhs, mapping = aes(x = age, y = sysBP)) +
  geom_point()
```



You can see the general increasing trend between age and sysBP for this model, and it seems to be linear (i.e. there is no clear curve in the relationship). However, the data still appears relatively scattered, so use this test with caution.

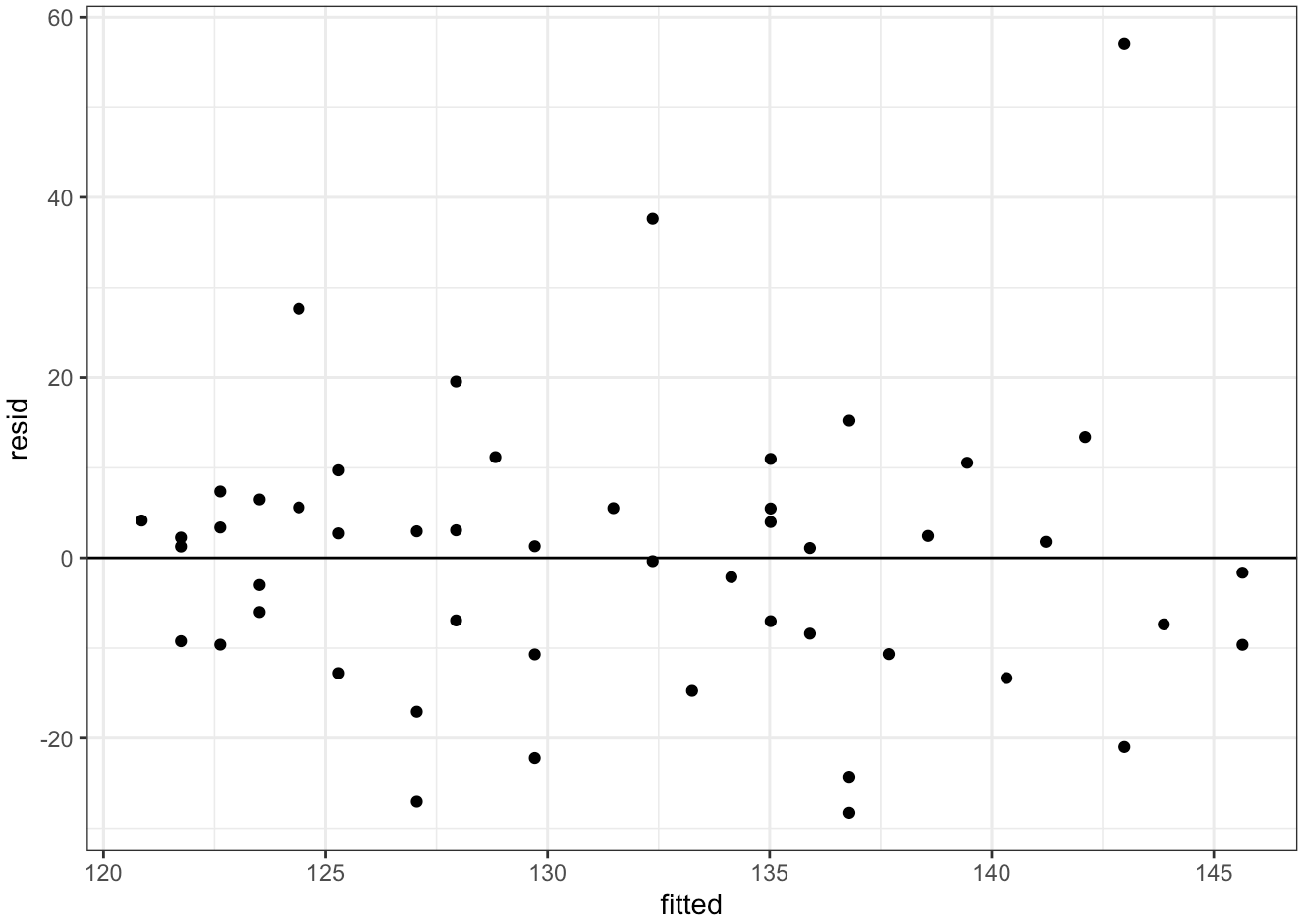
Homoscedasticity of Model Residuals

In previous labs, we used the `leveneTest` to test the homoscedasticity (homogeneity of variance) assumption. We used this test to determine whether the variance of a single variable between two groups was similar. For linear regression, we will need to test homoscedasticity of the model residuals. Residuals represent the point-by-point error between the true value of the outcome variable and what the linear model predicts based on the value of the predictor variable. The residuals are automatically calculated by the `lm` function, and we can add them as a new variable in the fhs dataset using the `residuals` function. We'll also add the predicted (fitted) values using `predict`.

```
fhs2 <- fhs %>%
  mutate(fitted = predict(sysBPvsAge),
         resid = residuals(sysBPvsAge))
```

To plot our residuals, we will be using the `fitted` and the `resid` variables. Let's make a scatterplot using `geom_point` with `fitted` on the x axis and `resid` on the y axis. Also, we can add a horizontal line at 0 using `geom_hline`.

```
ggplot(data = fhs2, aes(x = fitted, y = resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



In order for this data to meet the homoscedasticity assumption, the following 3 criteria must be met:

1. Points must be randomly distributed above and below the 0 line
2. There should be a similar residual variance at all fitted values
3. There should be no obvious pattern in the residuals

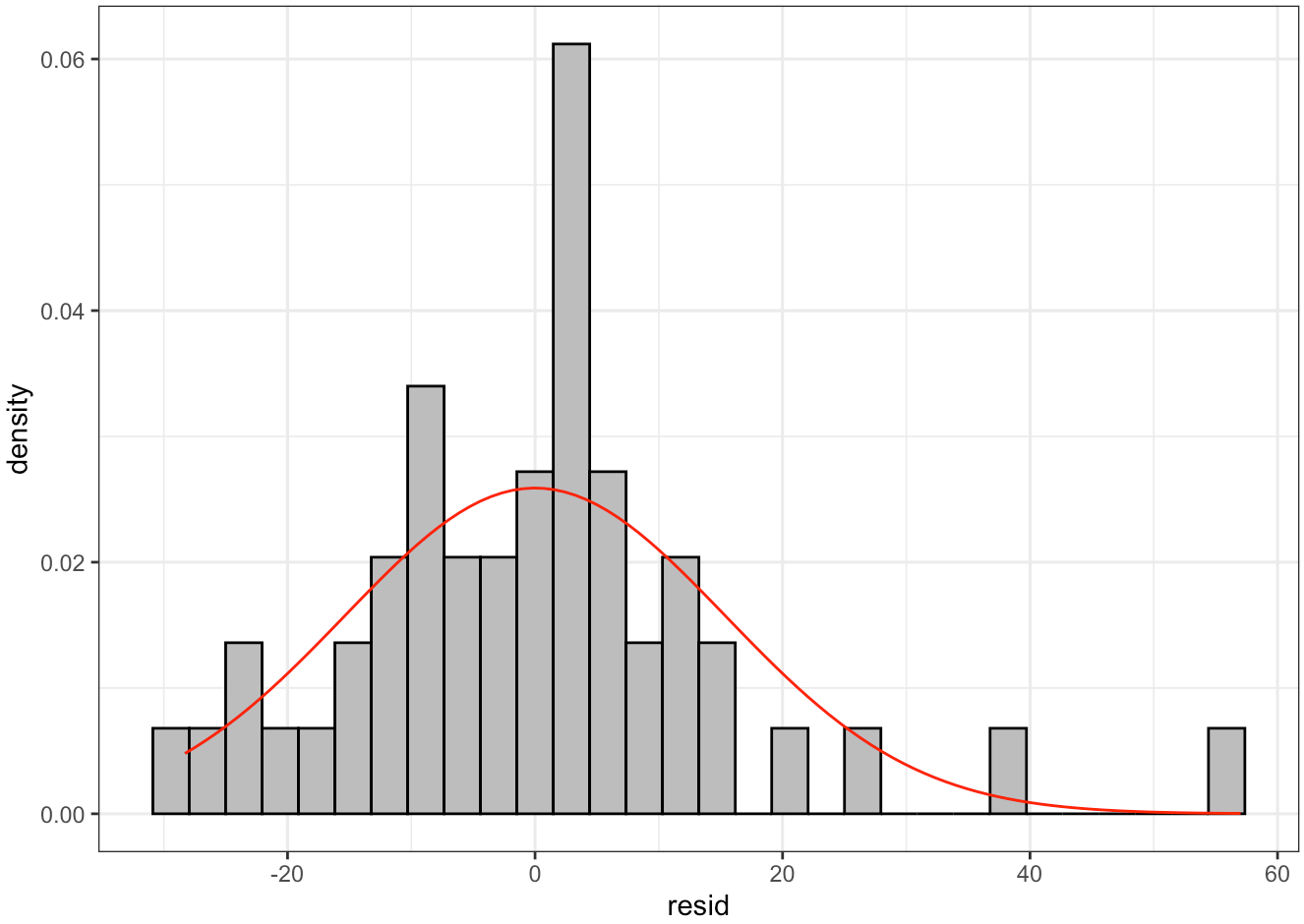
Does our data meet the homoscedasticity assumption?

Yes, our data meets all 3 criteria. All of the residuals are nicely distributed, there is no noticeable change in the variance at a given location, and there is no discernible pattern.

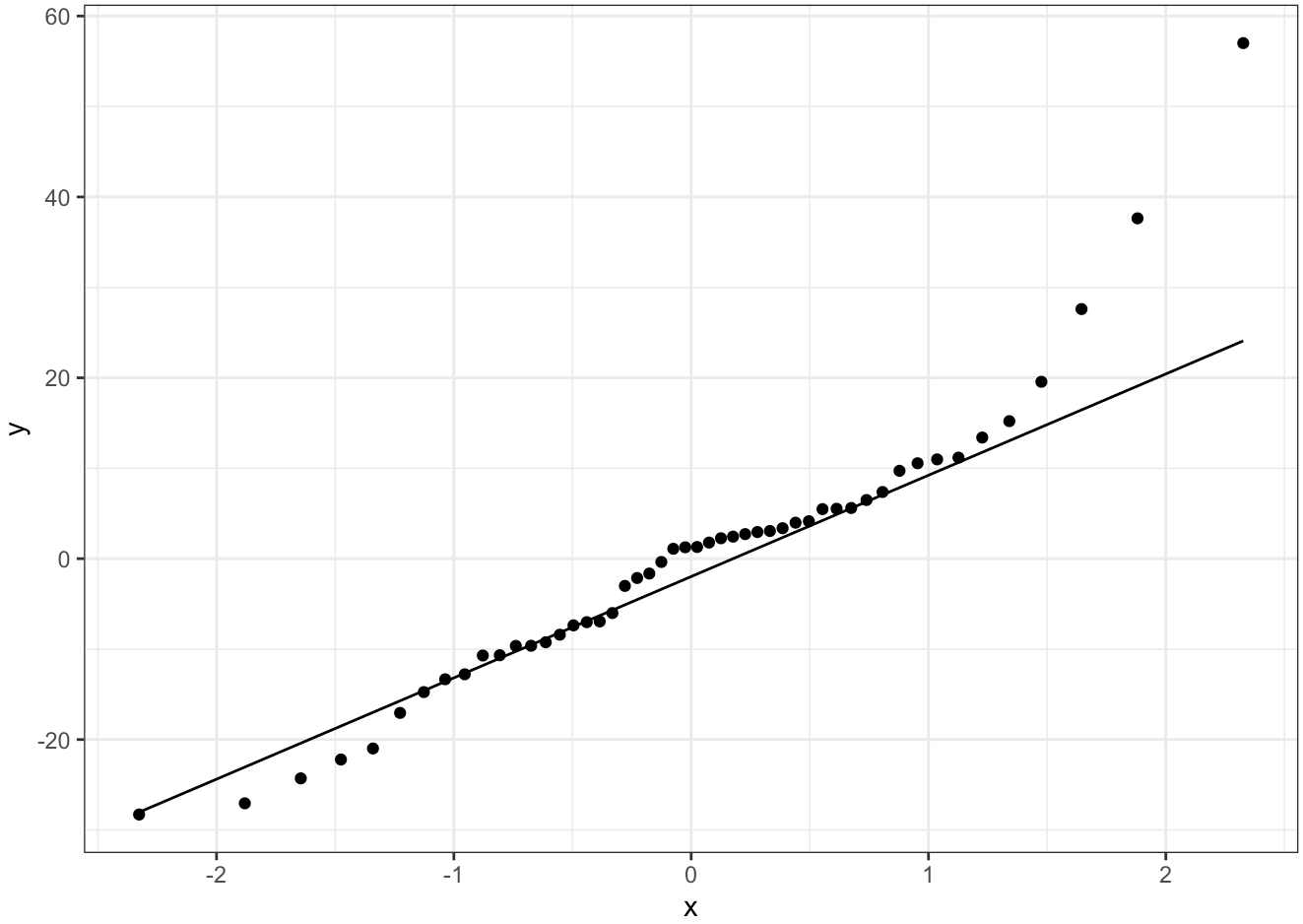
Normality of the Model Residuals

We will now assess normality of the resid variable in the same way we have done previously: histograms, Q-Q plots, boxplots, and statistical tests.

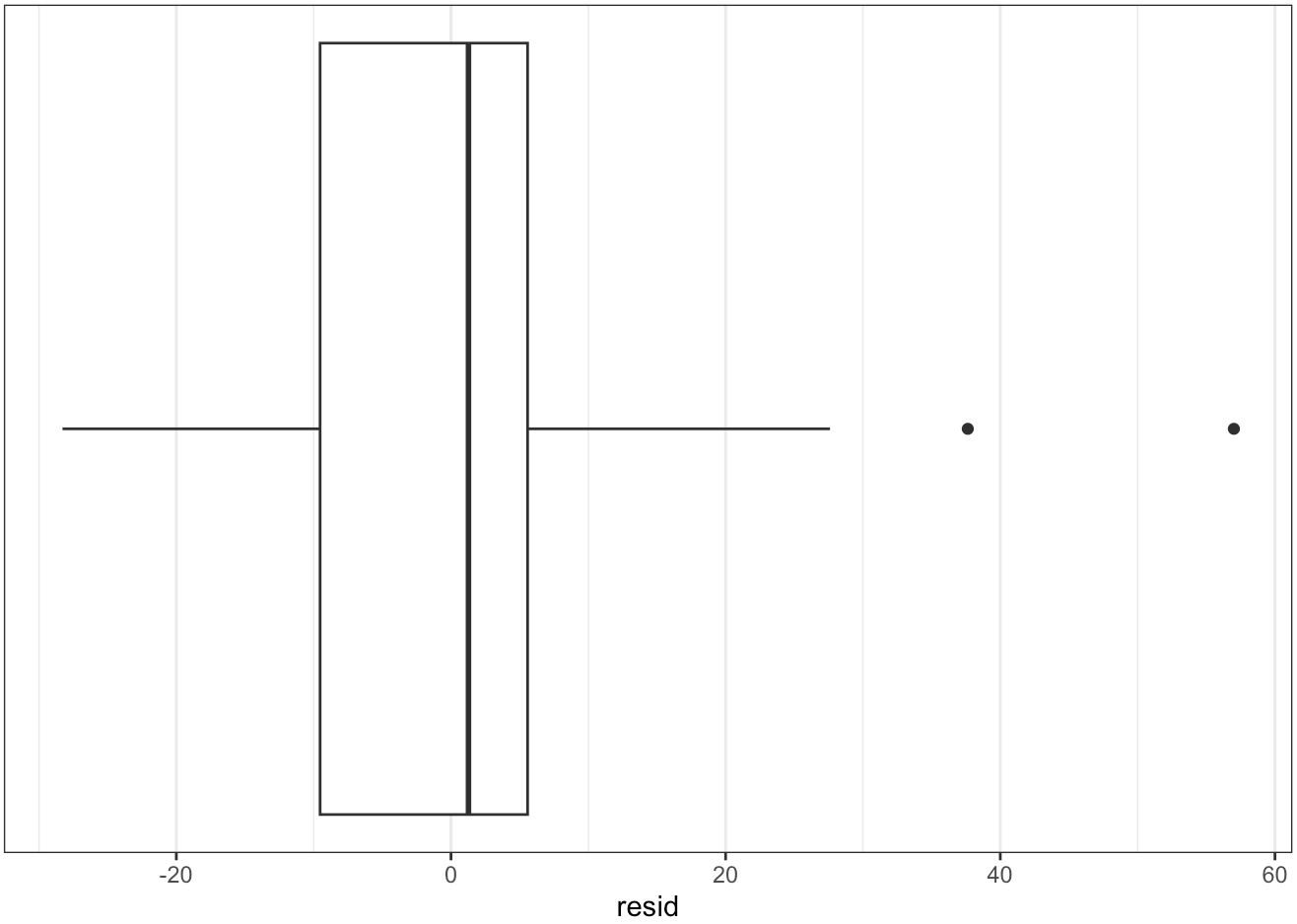
```
ggplot(data = fhs2, mapping = aes(x = resid)) +
  geom_histogram(mapping = aes(y = ..density..), bins = 30, fill = 'gray', color = 'blac
k') +
  stat_function(fun = dnorm,
                args = list(mean = mean(fhs2$resid), sd = sd(fhs2$resid)),
                color = 'red')
```




```
ggplot(data = fhs2, mapping = aes(sample = resid)) +
  geom_qq() +
  geom_qq_line()
```




```
ggplot(data = fhs2, mapping = aes(x = resid)) +
  geom_boxplot() +
  theme(axis.ticks.y = element_blank(),
        axis.text.y = element_blank(),
        panel.grid.major.y = element_blank(),
        panel.grid.minor.y = element_blank())
```



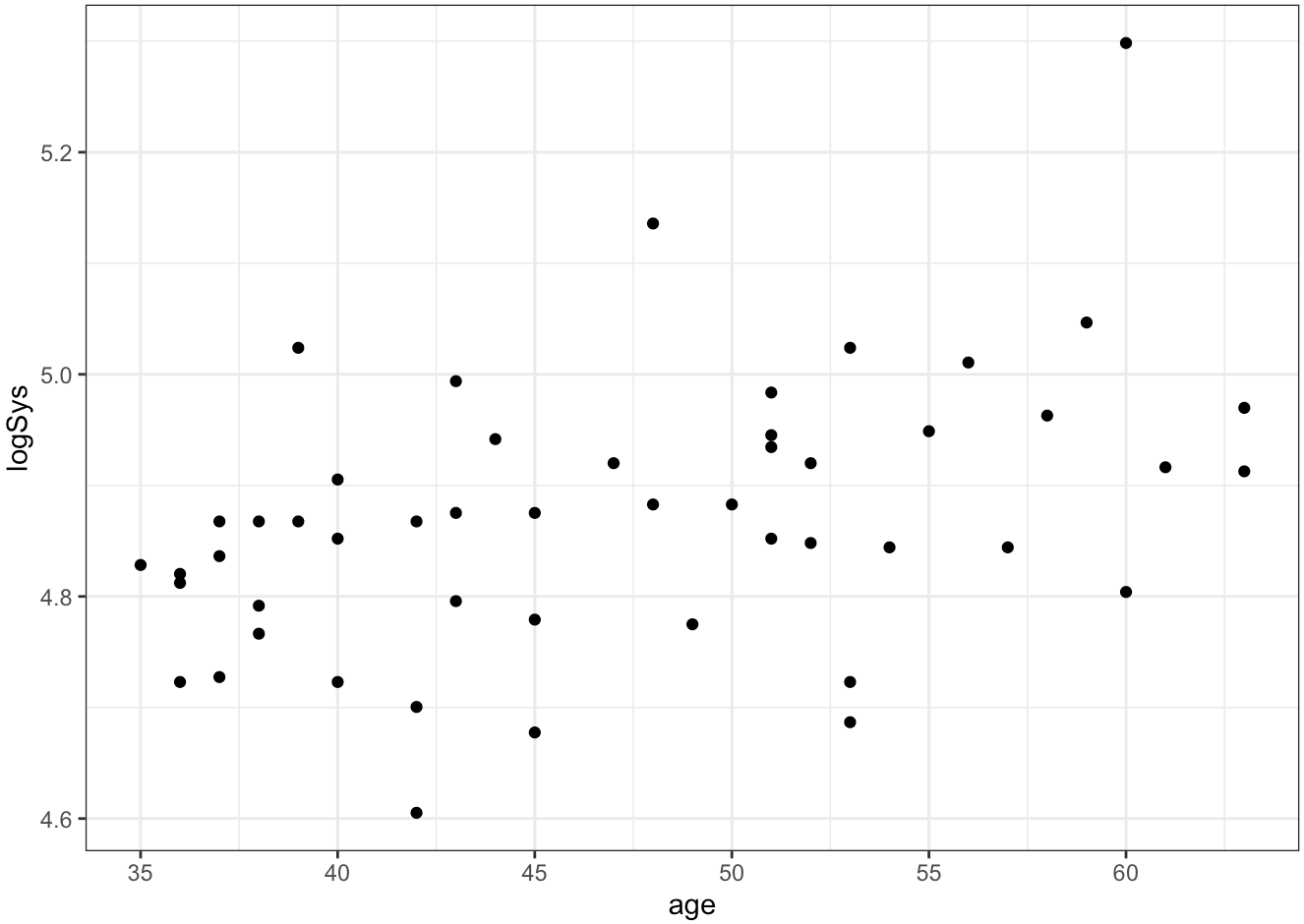

```
stat.desc.clean(dataset = fhs2, variable = resid)
```

```
## # A tibble: 1 × 6
##   skewness skew.2SE kurtosis kurt.2SE normtest.W normtest.p
##      <dbl>    <dbl>    <dbl>    <dbl>      <dbl>      <dbl>
## 1    0.999     1.48     2.56     1.93      0.926    0.00408
```

Based on these results, there actually does appear to be some deviations from normality for the residuals of the plot. However, they do seem to be mostly driven by a couple of outliers, and so we could possibly try removing those outliers and re-testing the model. Another option would be to transform our outcome and/or the predictor variable(s) and re-run the model. Finally, a robust form of linear regression could be used instead, but this won't be covered in this course.

However, let's try to log transform the outcome variable to see if we get normaly distributed residuals.

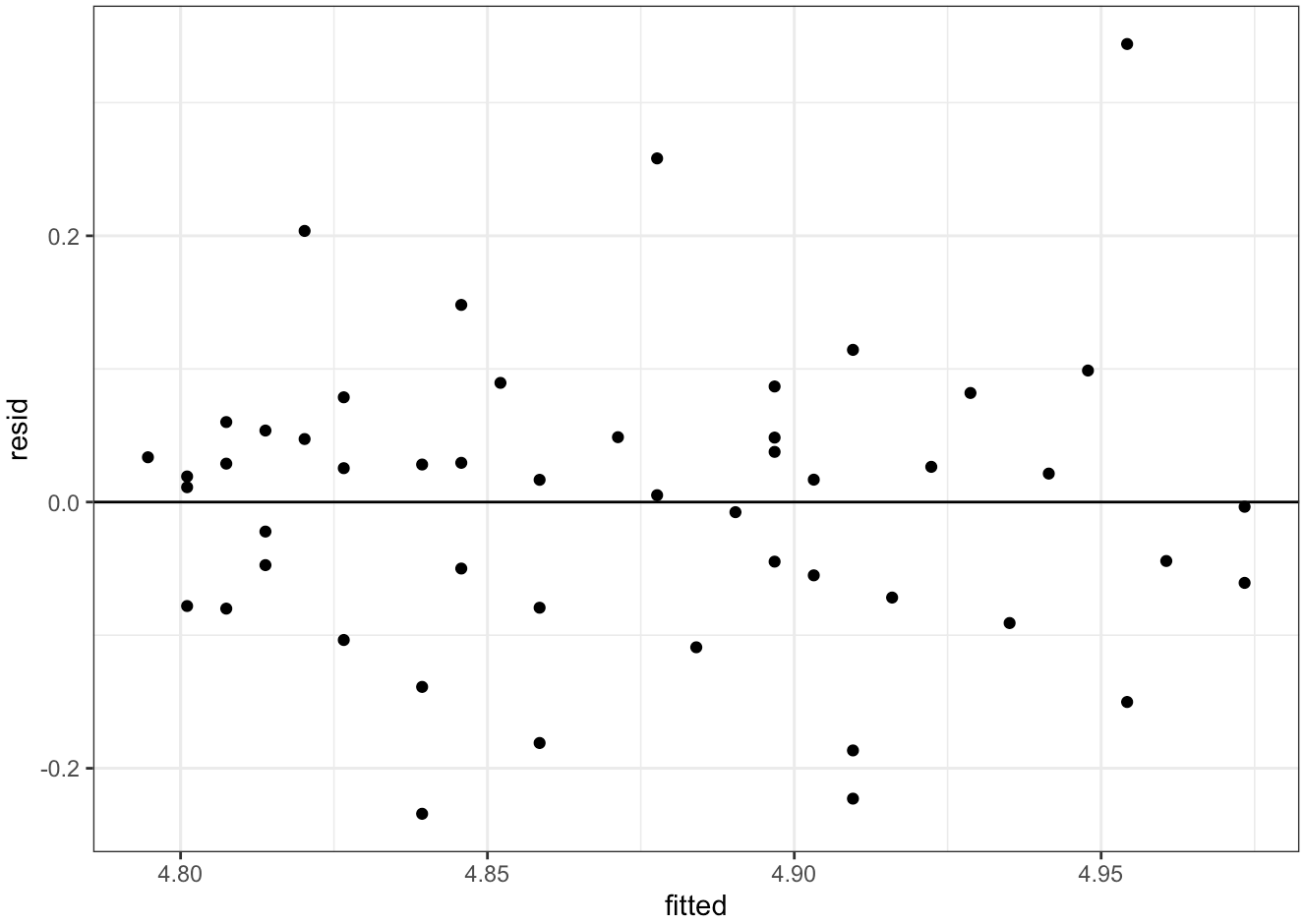
```
#Log-transformation of response/outcome variable?
fhs <- fhs %>% mutate(logSys = log(sysBP))

# Linearity
ggplot(data = fhs, mapping = aes(x = age, y = logSys)) +
  geom_point()
```



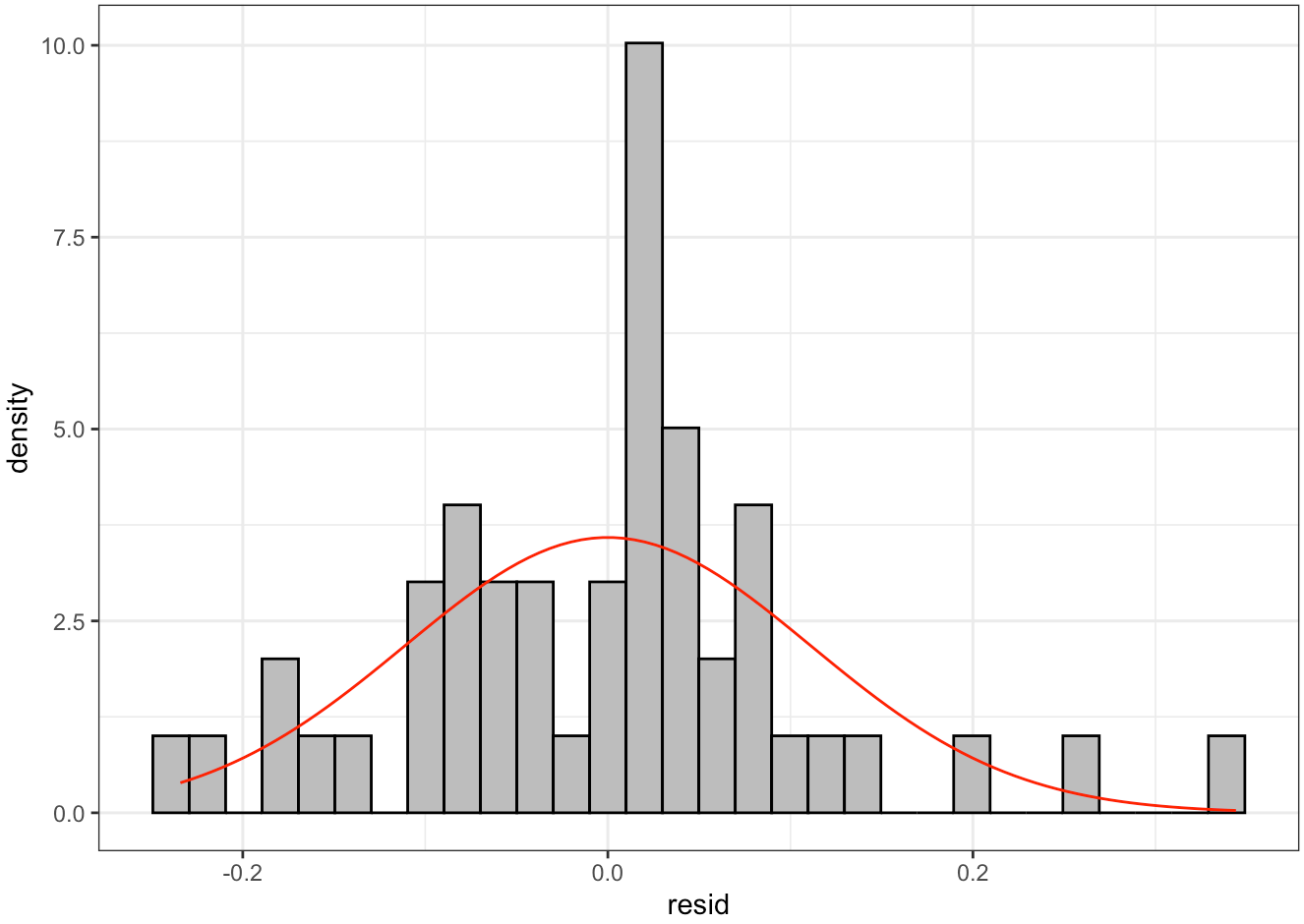

```
# Homoscedasticity of the model residuals
# Need to run the model first then extract the model residuals
sysBPvsAge <- lm(formula = logSys ~ age, data = fhs)
summary(sysBPvsAge)
```

```
##
## Call:
## lm(formula = logSys ~ age, data = fhs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23419 -0.06900  0.01675  0.04864  0.34408
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.571308   0.093333  48.978   <2e-16 ***
## age         0.006382   0.001953   3.268    0.002 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1124 on 48 degrees of freedom
## Multiple R-squared:  0.182,  Adjusted R-squared:  0.165
## F-statistic: 10.68 on 1 and 48 DF,  p-value: 0.002003
```

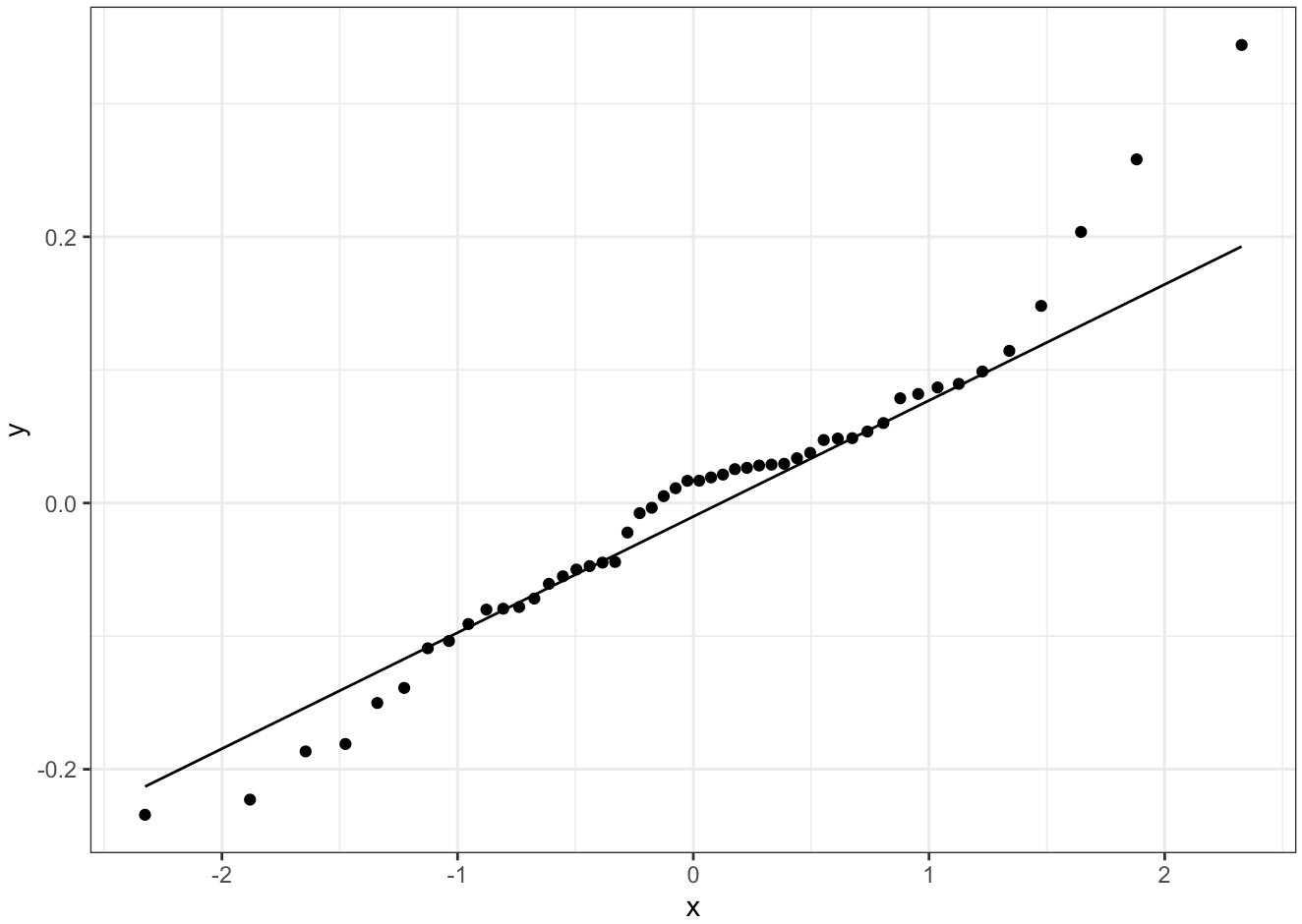
```
fhs2 <- fhs %>%
  mutate(fitted = predict(sysBPvsAge),
         resid = residuals(sysBPvsAge))

ggplot(data = fhs2, aes(x = fitted, y = resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



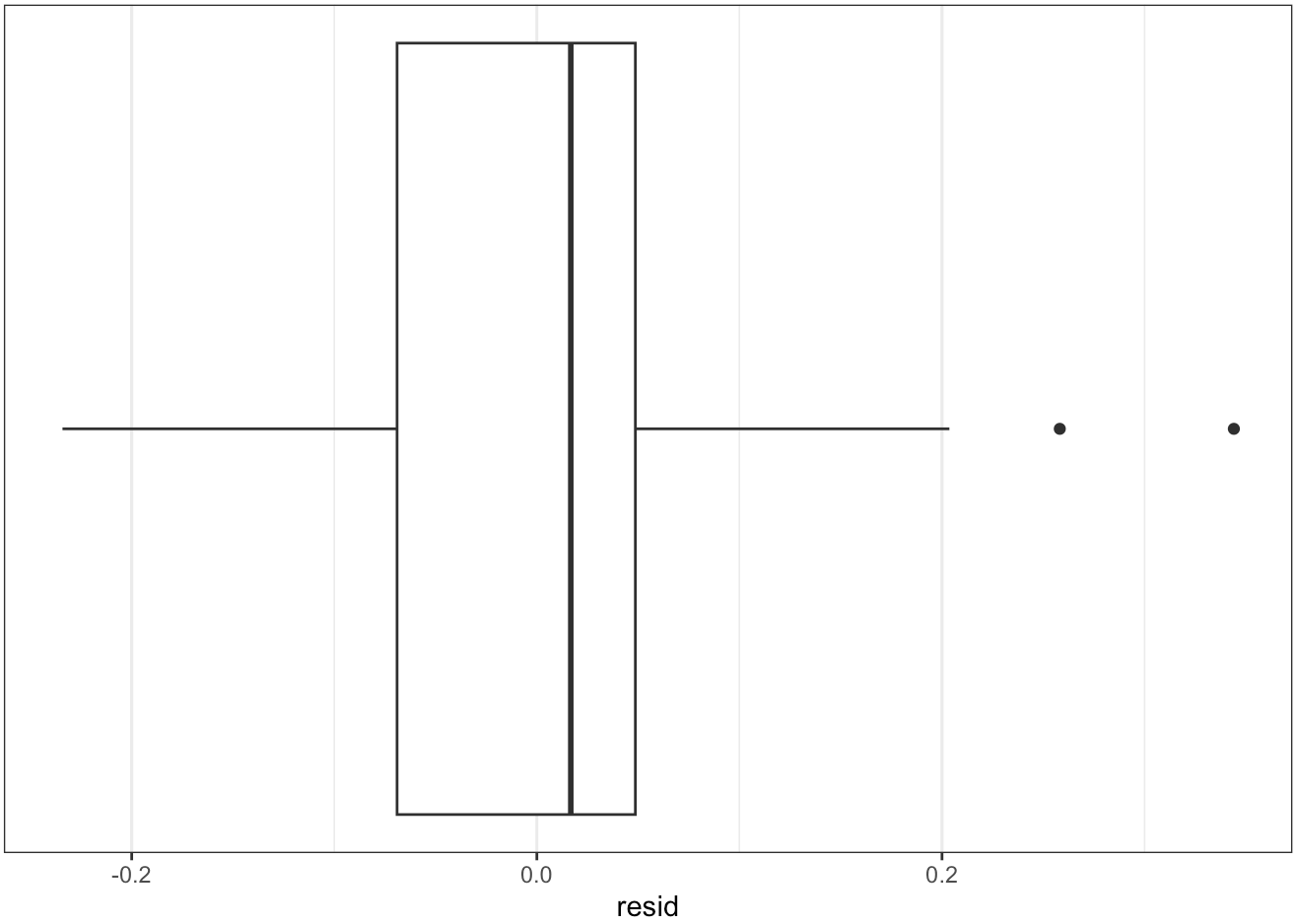

```
# Normality of the model residuals
# Histogram
ggplot(data = fhs2, mapping = aes(x = resid)) +
  geom_histogram(mapping = aes(y = ..density..), bins = 30, fill = 'gray', color = 'blac
k') +
  stat_function(fun = dnorm,
                args = list(mean = mean(fhs2$resid), sd = sd(fhs2$resid)),
                color = 'red')
```




```
#Q-Q plot
ggplot(data = fhs2, mapping = aes(sample = resid)) +
  geom_qq() +
  geom_qq_line()
```




```
ggplot(data = fhs2, mapping = aes(x = resid)) +
  geom_boxplot() +
  theme(axis.ticks.y = element_blank(),
        axis.text.y = element_blank(),
        panel.grid.major.y = element_blank(),
        panel.grid.minor.y = element_blank())
```




```
stat.desc.clean(dataset = fhs2, variable = resid)
```

```
## # A tibble: 1 × 6
##   skewness skew.2SE kurtosis kurt.2SE normtest.W normtest.p
##      <dbl>    <dbl>    <dbl>    <dbl>      <dbl>      <dbl>
## 1    0.396    0.588    0.967    0.730      0.963      0.123
```

Log transforming the outcome variable satisfies all assumptions of linear regression including normality of the model residuals.

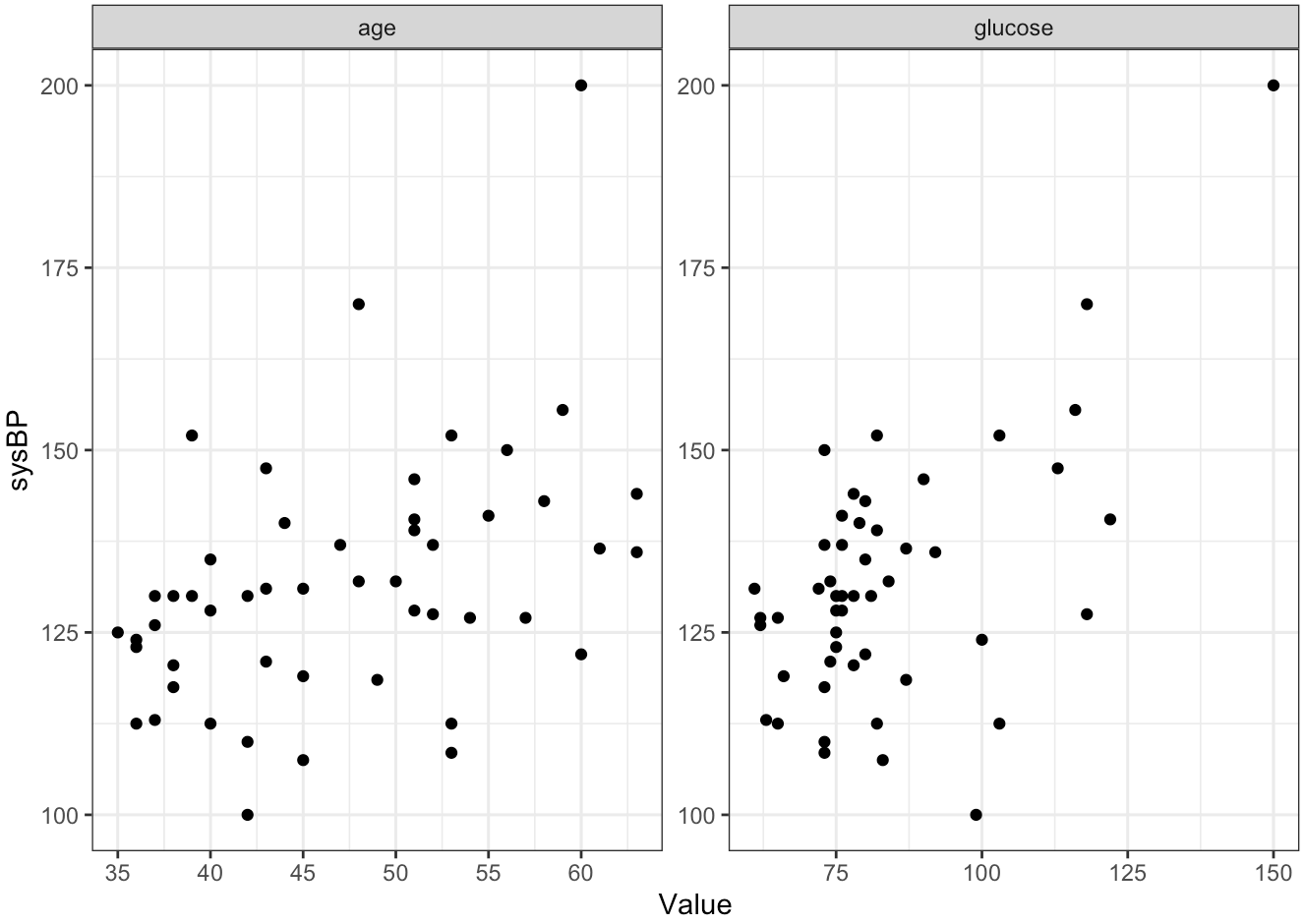
Assumptions of a Multiple Linear Regression

These are the same as for a simple linear regression, with the addition of a new assumption: no multicollinearity. Let's very quickly run the same model from last lab and test the assumptions as above.

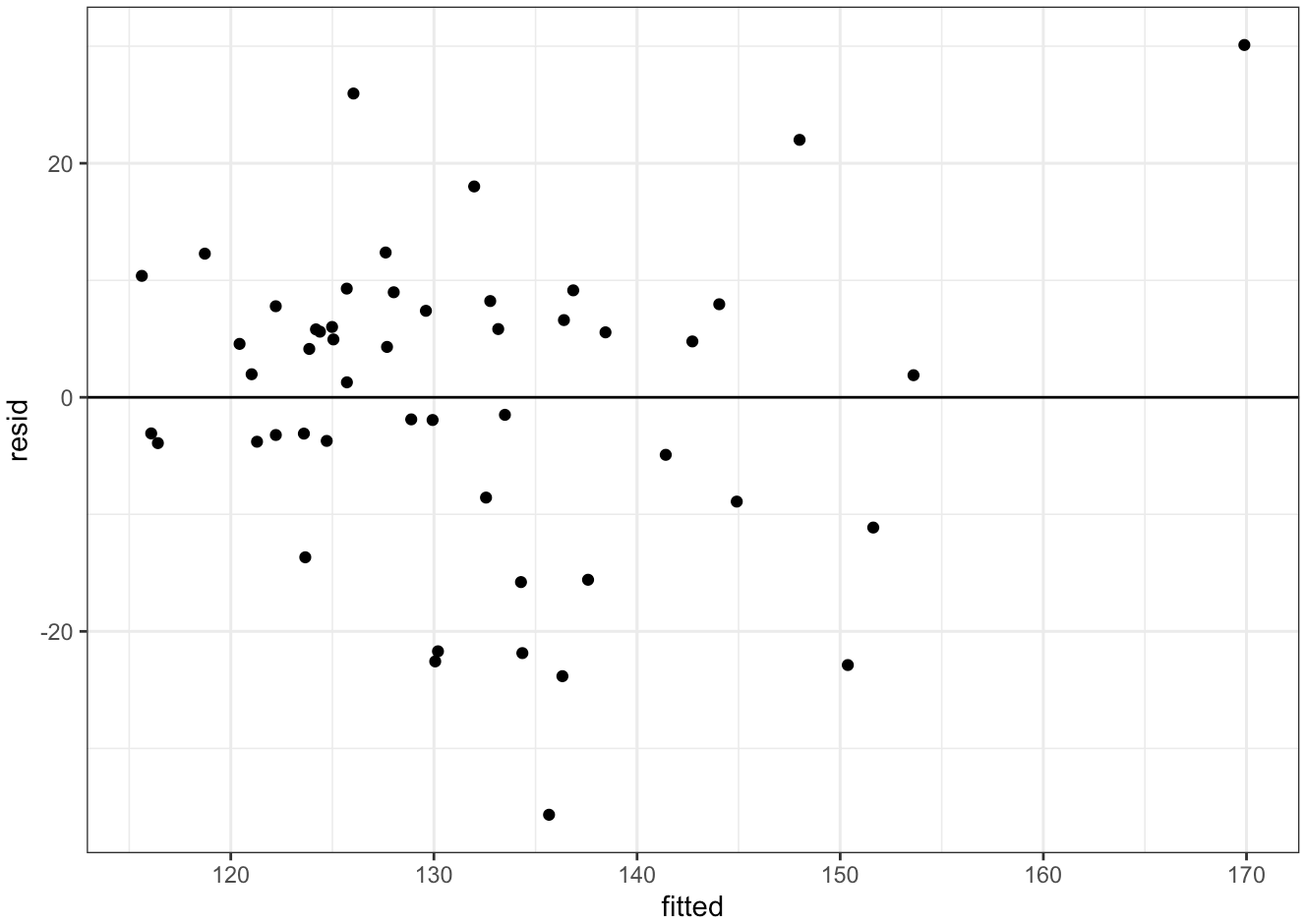
```
# Running our model
sysBPvsAgeandGluc <- lm(formula = sysBP ~ age + glucose, data = fhs)

summary(sysBPvsAgeandGluc)
```

```
##
## Call:
## lm(formula = sysBP ~ age + glucose, data = fhs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.667  -4.662   3.054   7.687  30.106
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.0347    12.6703   5.133 5.36e-06 ***
## age           0.5939     0.2440   2.433 0.018815 *
## glucose       0.4615     0.1110   4.158 0.000135 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.45 on 47 degrees of freedom
## Multiple R-squared:  0.4023, Adjusted R-squared:  0.3768
## F-statistic: 15.81 on 2 and 47 DF,  p-value: 5.598e-06
```

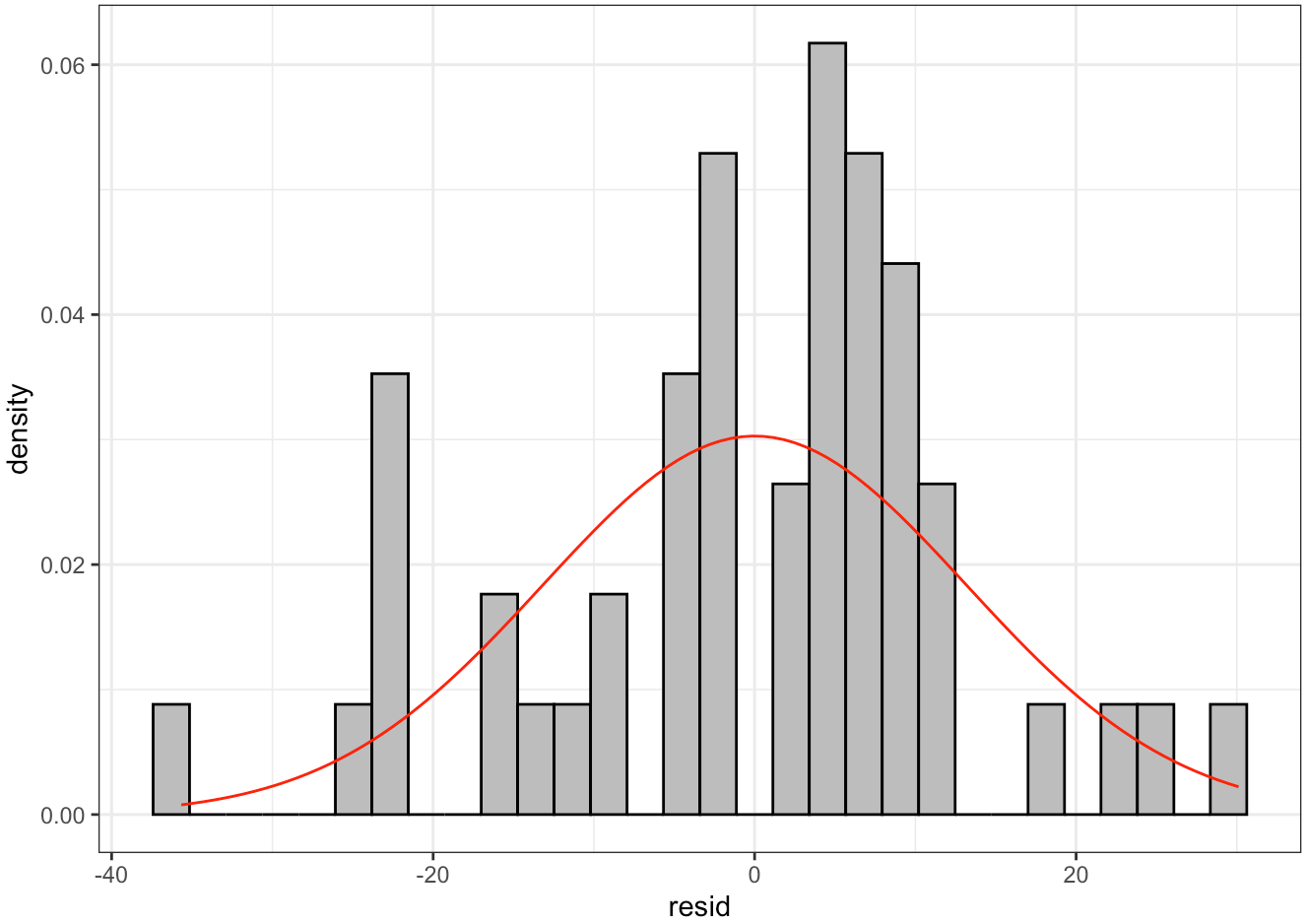
```
# We assume independence is met based on the model design

# Linearity
fhs3 <- fhs %>%
  select(-gender) %>%
  pivot_longer(cols = c(age, glucose),
               names_to = "Metric",
               values_to = "Value")

ggplot(data = fhs3, mapping = aes(x = Value, y = sysBP)) +
  geom_point() +
  facet_wrap(~ Metric, scales = "free")
```



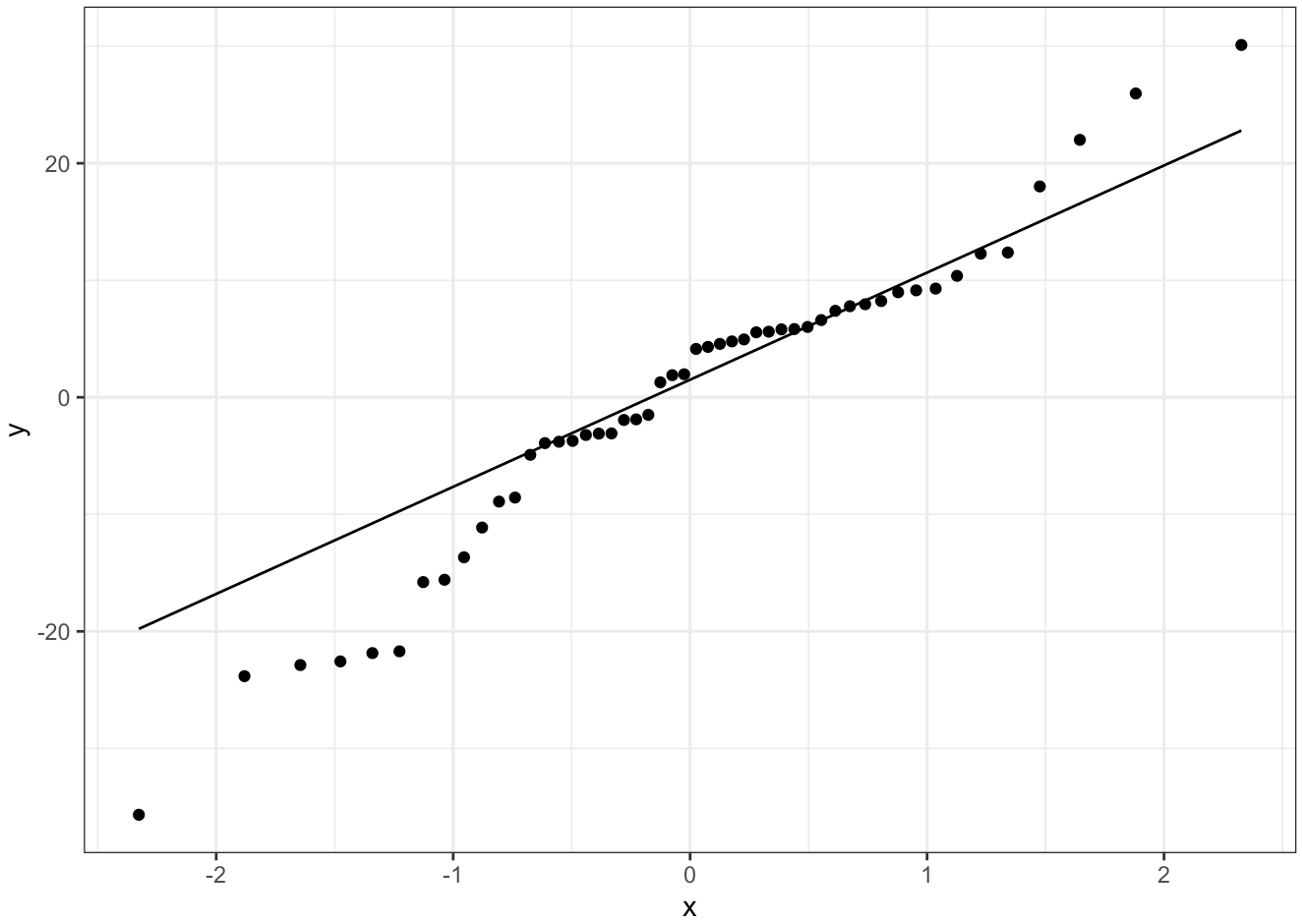

```
# Linearity seems to be met for both variables vs sysBP, even though it is harder to see
for glucose vs sysBP

fhs4 <- fhs %>%
  mutate(fitted = predict(sysBPvsAgeandGluc),
         resid = residuals(sysBPvsAgeandGluc))

# Homoscedasticity of residuals
ggplot(data = fhs4, aes(x = fitted, y = resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



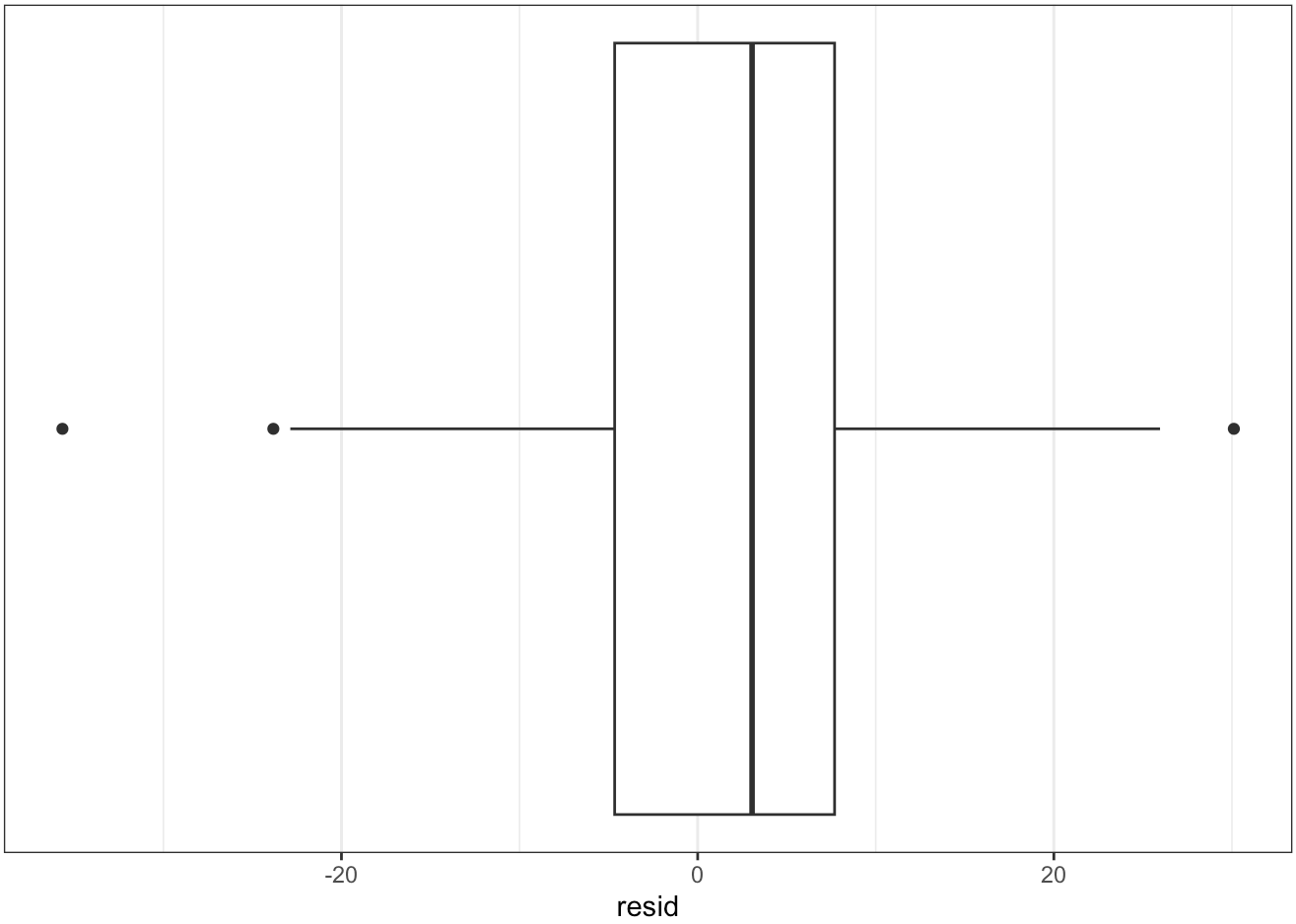

```
# This assumption seems to be mostly met, but there seems to be an outlier

# Normality of residuals
ggplot(data = fhs4, mapping = aes(x = resid)) +
  geom_histogram(mapping = aes(y = ..density..), bins = 30, fill = 'gray', color = 'blac
k') +
  stat_function(fun = dnorm,
                args = list(mean = mean(fhs4$resid), sd = sd(fhs4$resid)),
                color = 'red')
```




```
ggplot(data = fhs4, mapping = aes(sample = resid)) +
  geom_qq() +
  geom_qq_line()
```




```
ggplot(data = fhs4, mapping = aes(x = resid)) +
  geom_boxplot() +
  theme(axis.ticks.y = element_blank(),
        axis.text.y = element_blank(),
        panel.grid.major.y = element_blank(),
        panel.grid.minor.y = element_blank())
```




```
stat.desc.clean(dataset = fhs4, variable = resid)
```

```
## # A tibble: 1 × 6
##   skewness skew.2SE kurtosis kurt.2SE normtest.W normtest.p
##      <dbl>    <dbl>    <dbl>    <dbl>      <dbl>      <dbl>
## 1   -0.387   -0.576    0.206    0.156      0.960     0.0932
```

```
# This assumption seems to be met
```

Multicollinearity

Multicollinearity occurs when predictor variables in a **multiple** linear regression are highly correlated. This causes problems with both parameter estimation and prediction, so we need to make sure our predictors are fairly unique. We will test multicollinearity by calculating the variance inflation factor using the `vif` function from the `car` package.

```
vif(sysBPvsAgeandGluc)
```

```
##      age  glucose
## 1.089871 1.089871
```

Variables with VIF values above 10 should be removed from the model. We can see that neither of our predictors approach that so multicollinearity is not an issue in our multiple linear regression model. Thus, the multicollinearity assumption is met.

Independent Practice

Run the same model from last week to assess the effect of both glucose and age on diastolic blood pressure, then test your assumptions as above. Follow the complete process below:

1. Get to know your data: describe your variables and their distribution
2. Run the multiple linear regression test
3. Assumptions
    a. Determine whether assumptions are met
    b. If assumptions are not met, describe what you will do to account for this. If possible, modify your model to meet the assumptions
4. Report your findings as you would describe them in the results section