3AS/3AS4: Applied Statistics

Assignment 1

1. In the data set `igfdata.csv`, measurements on age, sex and insulin-like growth factor (`igf`) for a group of people are available. The data set can be downloaded from canvas. The original source is: *J. Clin. Endocrinol. Metab.* 78(3): 744–752, March 1994. Each row in the data set corresponds to one individual. You need to download the file in your computer in a suitable folder of your choice. Then start `RStudio` and set that folder containing the data as your working directory from the "Session" menu.Finally, import the data in `R` using the following:

```
igfdata = read.csv("igfdata.csv", header=T)
```

For all the following questions, include your R codes, plots, and outputs in the solution.

(a) Make a suitable plot for the distribution of `igf` and discuss your findings.

```
igfdata = read.csv("D:/BIRMINGHAM/RStudio/igfdata.csv", header=T)
library(ggplot2)
library(dplyr)
ggplot(igfdata, aes(x=age, y=igf, color=sex)) +
  geom_point()+ labs(x="Age, years", y="Insulin-like growth factor")
```

Plot 1. Scatterplot of igf against age and sex

It is better to look at the distribution of igf against age and sex, because age and sex are the only numerical variables in the data that has a value almost in each row of the data.

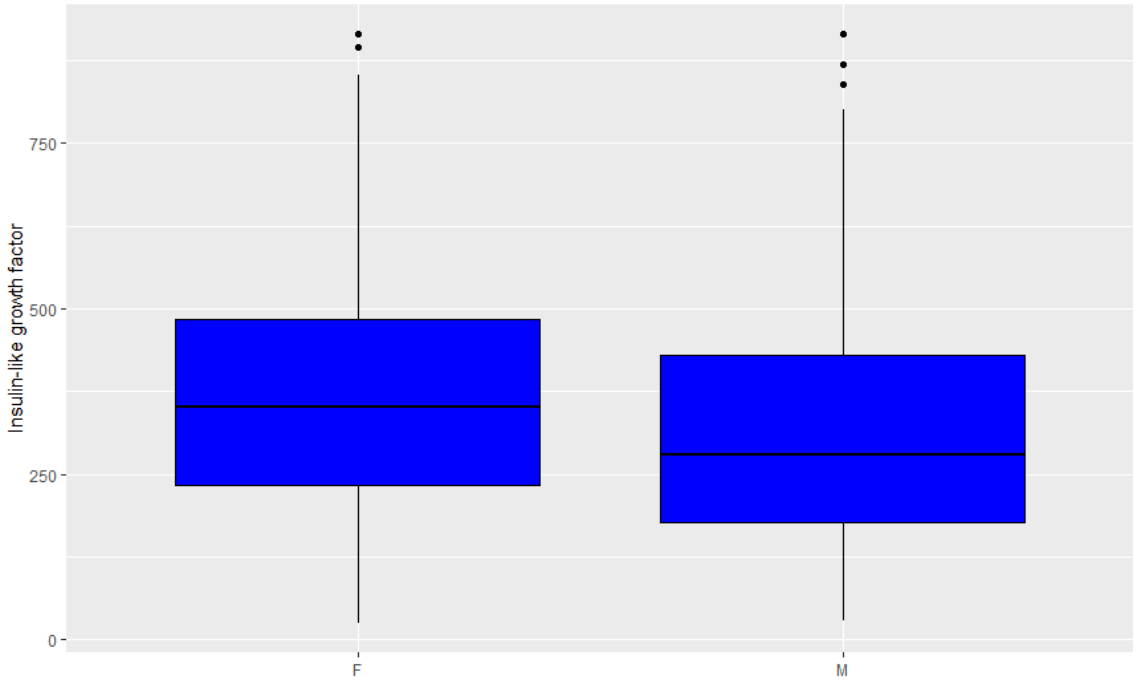
Scatterplot provides visual information about distribution of igf, showing its absolute values and demonstrating how its values change with years.

Plot 1 displays that the highest igf values have people who are between 10 and 20 years old. The lowest values of igf have people who were recently born or people who are around 80 years old.

Moreover, Plot 1 shows distribution of igf for different sex. Pink points represent females' value of igf, blue points – males' value of igf. Looking at the plot it is visible that the character of the igf distributions is similar for both sexes. Maximums achieves between 10 and 20 years and minimums are located at the ends of the age-scale.

(b) Compare the igf for males and females using boxplots. Discuss your findings.

```
ggplot(igfdata,aes(x=factor(sex), y=igf)) +geom_boxplot(colour="black", fill="blue")+
  labs(x=" ", y="Insulin-like growth factor")
```

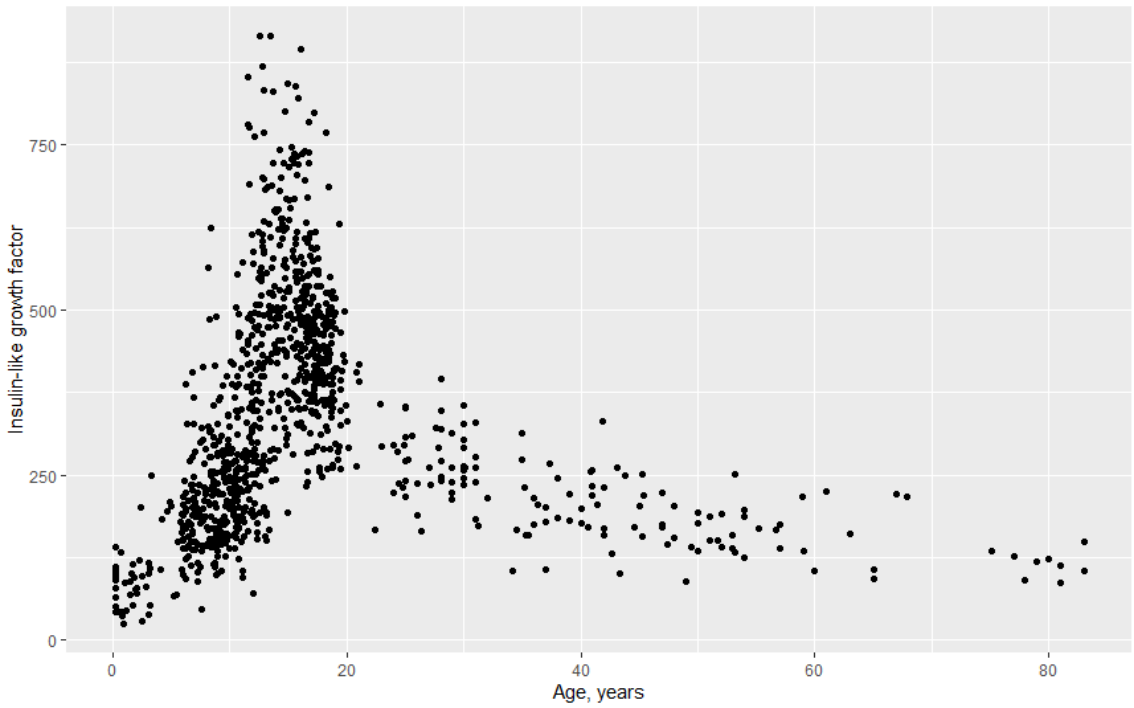


Plot 2. Boxplot of igf against gender

Plot 2 shows that females are more likely to have a greater igf value rather than males. So, females have higher hinge values. The difference between females' and males' hinge values could be estimated around 60. However, in general the distribution of igf for each categorical variable looks quit similarly.

(c) Make a scatterplot of igf against age. Comment on how igf changes with age.

```
ggplot(igfdata, aes(x=age, y=igf)) +
  geom_point()+ labs(x="Age, years", y="Insulin-like growth factor")
```



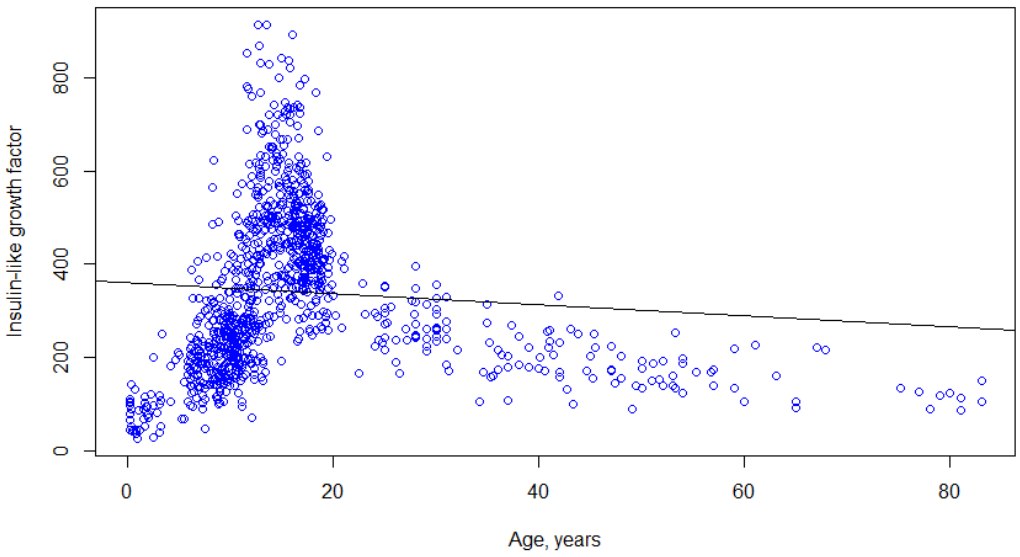
Plot 3. Scatterplot of igf against age

Plot 3 displays that the highest igf values have people who are between 10 and 20 years old. The lowest values of igf have people who were recently born or people who are around 80 years old.

(d) Fit a simple linear regression to predict igf using age. Is age a significant variable in this regression? Justify your answer.

```
linear_model1 = lm(igf~age, data = igfdata)
plot(igfdata$age,igfdata$igf, xlab="Age, years",
     ylab="Insulin-like growth factor", col="blue")
abline(linear_model1)
summary(linear_model1)
```

```
Call:
lm(formula = igf ~ age, data = igfdata)

Residuals:
    Min      1Q  Median      3Q     Max
-334.78 -134.38  -25.29  121.52  570.16

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 360.8602     9.0219  39.998  < 2e-16 ***
age          -1.1918     0.4408  -2.704  0.00697 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 170.3 on 1011 degrees of freedom
  (321 пропущенное наблюдение удалено)
Multiple R-squared:  0.00718,   Adjusted R-squared:  0.006198
F-statistic: 7.311 on 1 and 1011 DF,  p-value: 0.006967
```

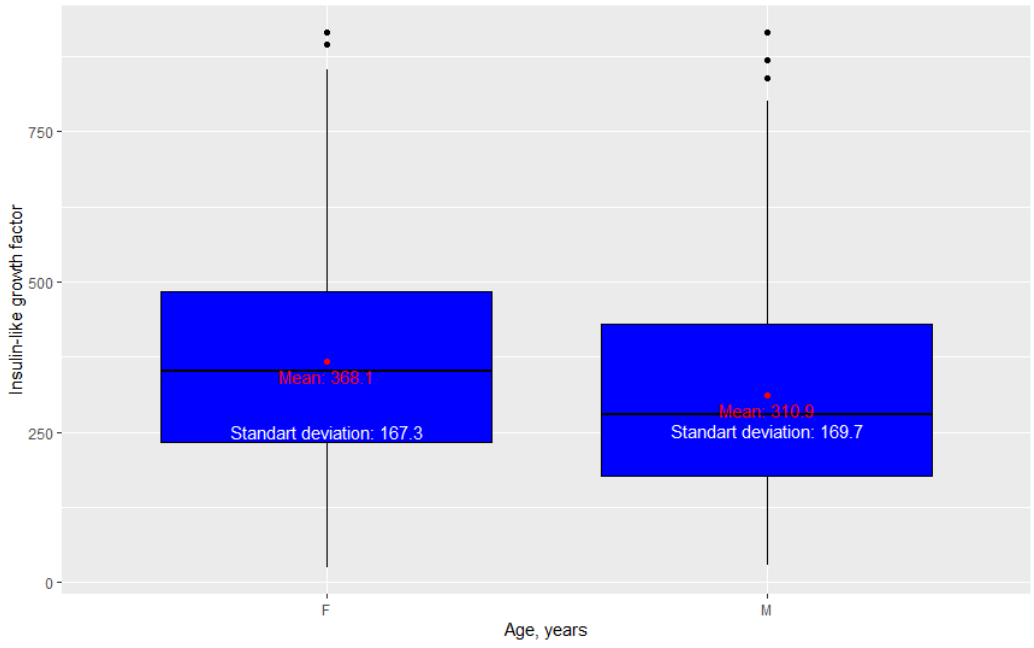


Plot 4. Simple linear regression to predict igf using age

P-value is equal to 0.006967 that is less than 0.05 (classical level of significance). That means that null hypothesis is false, so age is a significant value.

(c) Report the mean and standard deviation of `igf` for males and females separately.

```
ggplot(igfdata,aes(x=factor(sex), y=igf)) +geom_boxplot(colour="black", fill="blue")+
  labs(x=" ", y="Insulin-like growth factor") +
  stat_summary(fun = mean, geom = "point", col = "red") +
  stat_summary(fun = mean, geom = "text", col = "red",vjust = 1.5,
               aes(label = paste("Mean:", round(..y.., digits = 1))))+
  stat_summary(fun = sd, geom = "text", col = "white", vjust = -3,
              aes(label = paste("Standart deviation:",round(..y.., digits = 1))))
```



Plot 5. Boxplot of igf against gender. Included the mean and standard deviation

(f) Using linear regression or otherwise check if there is a significant difference in mean igf for males and females. Use level of significance $\alpha = 0.05$.

Assume that null hypothesis tells that there is not a significant difference in mean, so they are equal.

Test null hypothesis

```
m1 = 368.1
m2 = 310.9
sd1 = 167.3
sd2 = 169.7
num1 <- filter(igfdata, igfdata$sex == "M")
num1 = nrow(num1)
num2 <- filter(igfdata, igfdata$sex == "F")
num2 = nrow(num2)
se <- sqrt(sd1*sd1/num1+sd2*sd2/num2)
t <- (m1-m2)/se
pt(-t,min(num1,num2)-1)<0.05
pt(-t,min(num1,num2)-1)
```

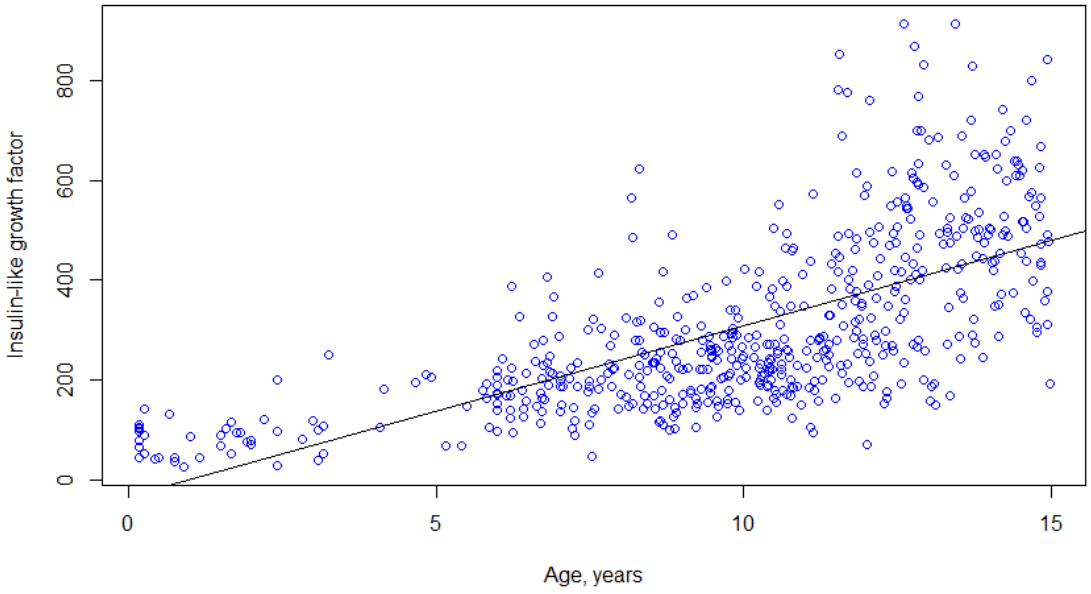
```
> m1 = 368.1
> m2 = 310.9
> sd1 = 167.3
> sd2 = 169.7
> num1 <- filter(igfdata, igfdata$sex == "M")
> num1 = nrow(num1)
> num2 <- filter(igfdata, igfdata$sex == "F")
> num2 = nrow(num2)
> se <- sqrt(sd1*sd1/num1+sd2*sd2/num2)
> t <- (m1-m2)/se
> pt(-t,min(num1,num2)-1)<0.05
[1] TRUE
> pt(-t,min(num1,num2)-1)
[1] 2.743712e-18
```

Result: P-value is less than 0.05, so null hypothesis is false. Thus, there is a significant difference in mean igf for males and females.

(g) Consider the subset of the data with age less than or equal to 15 years. For this subset of people, use linear regression with age and sex as predictors to predict igf. Comment on the significance of the variables age and sex.

```
igfdata <- igfdata %>%
  mutate(gender= if_else(sex == "M", 1, 0))
subset_age <- filter(igfdata, age <= 15)
head(subset_age)
linear_model2 = lm(igf~age+gender, data = subset_age)
plot(subset_age$age,subset_age$igf, xlab="Age, years", ylab="Insulin-like growth factor", col="blue")
abline(linear_model2)
summary(linear_model2)
```

```
Call:
lm(formula = igf ~ age + gender, data = subset_age)

Residuals:
    Min      1Q  Median      3Q     Max
-300.07  -88.95  -18.82   73.33  517.57

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.895     18.555  -0.102    0.919
age           33.064      1.612  20.507  < 2e-16 ***
gender       -45.048     10.974  -4.105 4.64e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 128.2 on 563 degrees of freedom
  (290 пропущенных наблюдений удалены)
Multiple R-squared:  0.4616,    Adjusted R-squared:  0.4597
F-statistic: 241.4 on 2 and 563 DF,  p-value: < 2.2e-16
```

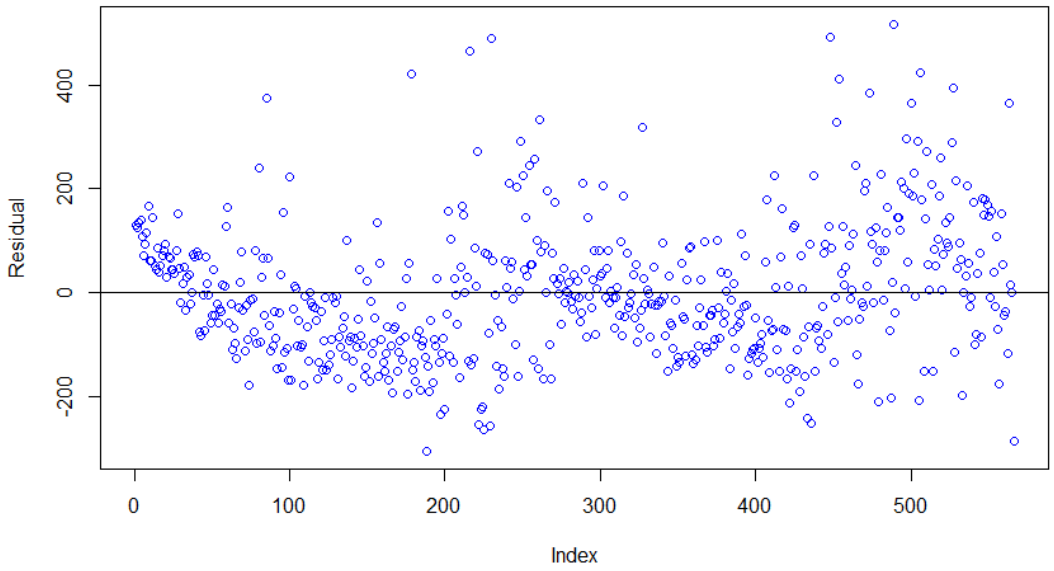


Plot 6. Linear regression with age and sex as predictors. Age<=15

P-value is less than 0.05, so null hypothesis is false. Thus, variables age and sex are significant. Moreover, according to R-squared the variables explain around 50% of original data.

(h) Use residual plot to check if the nonlinearity assumption is violated and if so, use an alternative model to fit the data.

```
res <- resid(linear_model2)
plot(res, xlab="Index", ylab="Residual", col="blue")
abline(0,0)
```

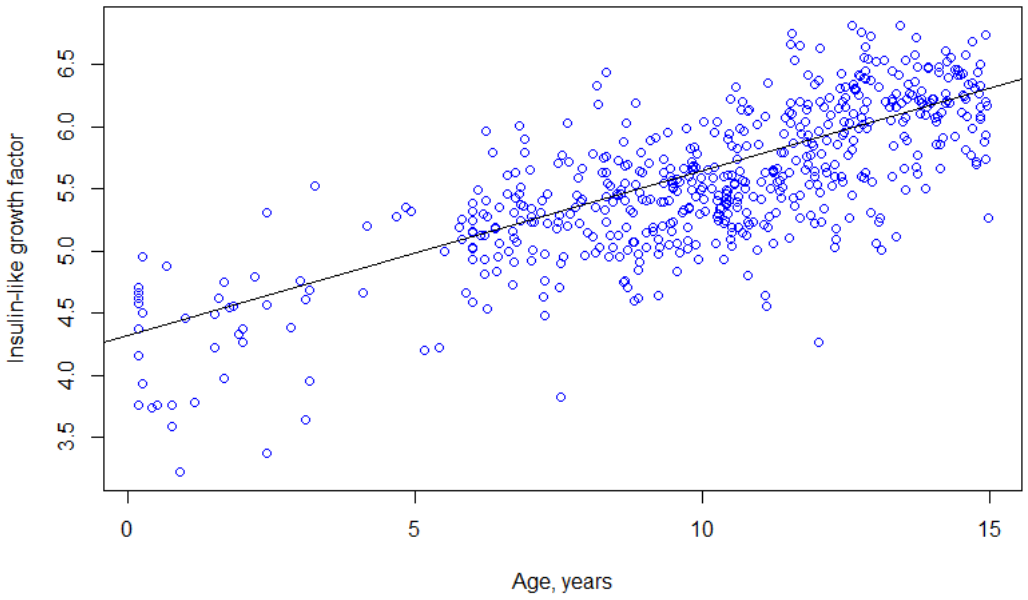


Plot 7. Residual plot. Age<=15

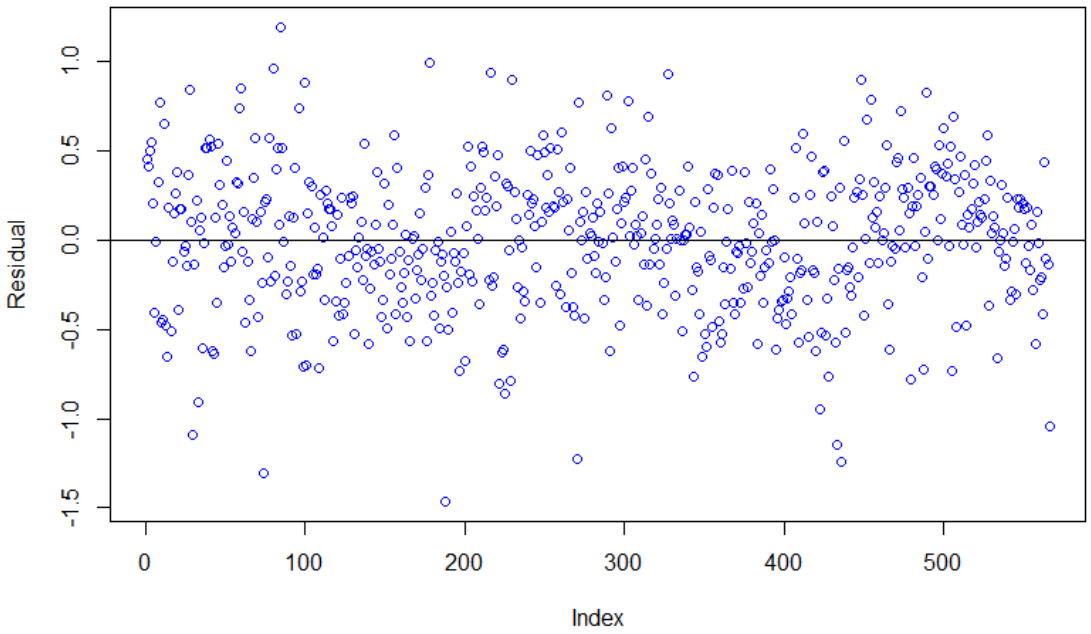
The points in a residual plot are not exactly randomly dispersed around the horizontal axis. Let us try to use a logarithmic model:

```
igfdata <- igfdata %>%
  mutate(gender= if_else(sex == "M", 1, 0))
subset_age <- filter(igfdata, age <= 15)
head(subset_age)
linear_model2 = lm(log(igf)~age+gender, data = subset_age)
plot(subset_age$age,log(subset_age$igf), xlab="Age, years", ylab="Insulin-like growth factor", col="blue")
abline(linear_model2)
summary(linear_model2)
```

```
Call:
lm(formula = log(igf) ~ age + gender, data = subset_age)

Residuals:
     Min       1Q   Median       3Q      Max
-1.46589 -0.23971  0.00906  0.25097  1.19674

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.323799   0.056236  76.886  < 2e-16 ***
age          0.132205   0.004887  27.053  < 2e-16 ***
gender      -0.183004   0.033262  -5.502 5.71e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3887 on 563 degrees of freedom
  (290 пропущенных наблюдений удалены)
Multiple R-squared:  0.5993,    Adjusted R-squared:  0.5979
F-statistic:   421 on 2 and 563 DF,  p-value: < 2.2e-16
```



Plot 8. Linear regression with age and sex as predictors for the logarithmic model. Age<=15



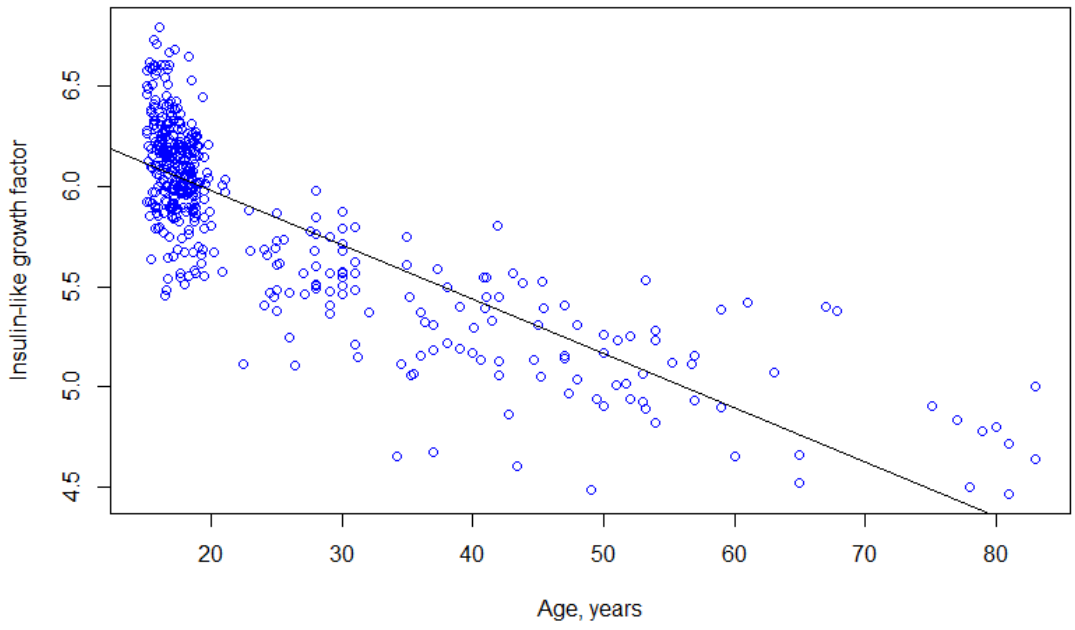
Plot 8. Residual plot for the logarithmic model. Age<=15

The points in a residual plot are randomly dispersed around the horizontal axis, so the logarithmic model fits better. And according to R-squared the variables explain around 60% of original data, that is better than before.

(i) Fit a similar model to predict `igf` for the subset of people with age greater than 15 years.

```
subset_age <- filter(igfdata, age >15)
linear_model3 = lm(log(igf)~age+gender, data = subset_age)
plot(subset_age$age,log(subset_age$igf), xlab="Age, years", ylab="Insulin-like growth factor", col="blue")
abline(linear_model3)
summary(linear_model3)
```

```
Call:
lm(formula = log(igf) ~ age + gender, data = subset_age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.94248 -0.17131  0.00849  0.16166  0.71829

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.5213903  0.0272218  239.56   <2e-16 ***
age         -0.0270689  0.0009056  -29.89   <2e-16 ***
gender       0.0109838  0.0255206    0.43    0.667
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2674 on 444 degrees of freedom
  (31 пропущенное наблюдение удалено)
Multiple R-squared:  0.6704,    Adjusted R-squared:  0.6689
F-statistic: 451.5 on 2 and 444 DF,  p-value: < 2.2e-16
```

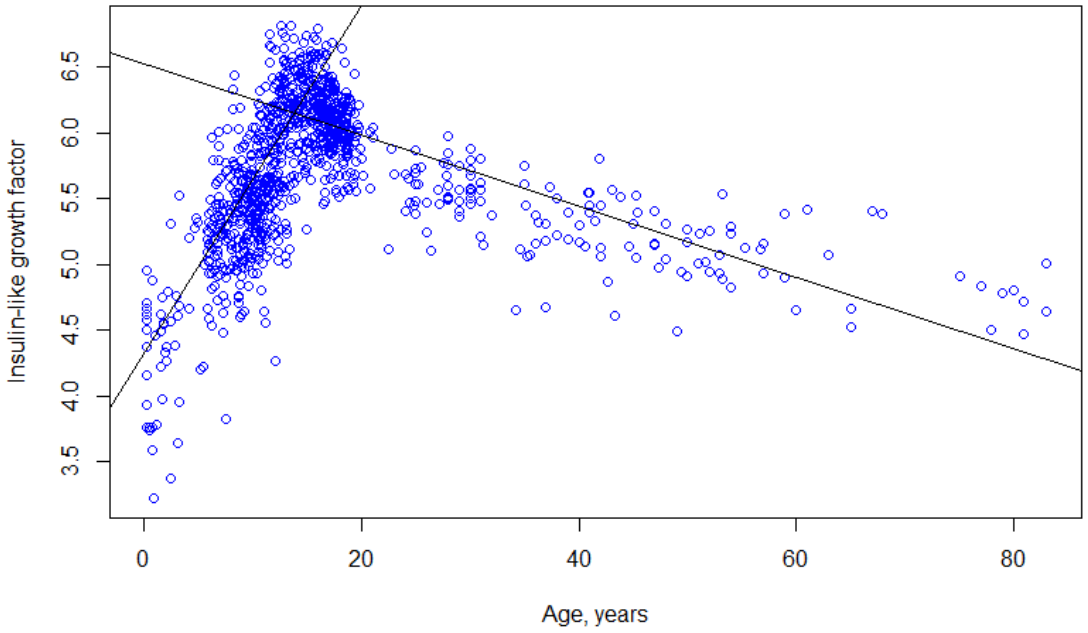


Plot 9. Linear regression with age and sex as predictors. Age>15

P-value is less than 0.05, so null hypothesis is false. Thus, variables age and sex are significant. Moreover, according to R-squared the variables explain around 70% of original data.

(j) Is there a significant difference in the models for people with age less than or equal to 15 years and for people with age greater than 15 years.

```
plot(igfdata$age, log(igfdata$igf), xlab="Age, years", ylab="Insulin-like growth factor", col="blue")
abline(linear_model2)
abline(linear_model3)
```



Plot 10. Combination of Linear regression models

Plot 10 shows that there is a significant difference in the models. One predicts that values of igf gets increase, another – those values of igf gets decrease.

2. $n = 100$ observations; a single predictor; quantitative response
fit a linear regression model to the data and a separate cubic regression

(a) Threre relationship between X and Y is linear, ie. $Y = \beta_0 + \beta_1 X + \varepsilon$

Consider the training residual sum of squares (RSS) for the linear regression and for the cubic regression.
Would we expect one to be lower than the other?

$$e_i = y_i - \hat{y}_i$$
$$RSS = \sum_{i=1}^{n} e_i^2$$
$$MSE = \frac{1}{n} \cdot RSS, \text{ where } \frac{1}{n} = \text{const} \quad (1)$$

We know, that with the increasing model flexibility MSE for training data decreases.
Due to (1), if MSE gets lower, RSS get lower either.
Hence, the cubic regression has lower RSS than the linear regression.

(b) For test data we already know, that the true relationship between X and Y is linear.
So, test RSS for linear regression is lower than for cubic regression. Because, in this case, the linear regression has low bias and lower variance than the cubic regression model have.

(c) The true relationship between X and Y is not linear.
Consider the training RSS for the linear regression and for cubic regression. Would we expect one to be lower than the other?
No matter what model we use, the training RSS gets lower with the increasing flexibility.
So the training RSS for the cubic regression is lower

(d) For test data:

There is not enough information to tell. Since for test data we consider two parameters: variance and bias.

If the flexibility of model increases, the variance of the model increases and its bias decreases.

In this case, we don't know which of the two parameters has a more significant inpact.

3. Consider a simple linear regression

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, \dots n$

$E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, $E(\varepsilon_i \varepsilon_j) = 0$, $i \neq j$

- $\hat{\beta}_0, \hat{\beta}_1$ - the least squares estimator of β_0 and β_1
- $\hat{y}_0$ is the predicted value of y for a new observation $x = x_0$.

(a) $$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i - \bar{y} \sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} =$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}, \text{ i.e. } \sum_{i=1}^n (x_i - \bar{x}) = 0$$

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 =$$

$$= \bar{y} + \hat{\beta}_1 (x_0 - \bar{x}) =$$

$$= \frac{1}{n} \sum_{i=1}^n y_i + \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} (x_0 - \bar{x}) =$$

$$= \sum_{i=1}^n \underbrace{\left(\frac{1}{n} + \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} (x_0 - \bar{x}) \right)}_{c_i} y_i = \sum_{i=1}^n c_i y_i$$

(b) $$Bias(\hat{y}_0) = E(\hat{y}_0) - y_0 = \sum_{i=1}^n c_i E(y_i) - y_0 =$$

$$= \sum_{i=1}^n c_i E(\beta_0 + \beta_1 x_i + \varepsilon_i) - y_0 =$$

$$= \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i - y_0$$

$$Var(\hat{y}_0) = Var\left(\sum_{i=1}^n c_i y_i \right) = \sum_{i=1}^n c_i \, var(y_i) =$$

$$= \sum_{i=1}^n c_i^2 \, var(\beta_0 + \beta_1 x_i + \varepsilon_i) = \sigma^2 \sum_{i=1}^n c_i^2$$

(c) The true model is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

$$y_0 = \beta_0 + \beta_1 x_0 + \beta_2 x_0^2 + \varepsilon$$

$$Bias(\hat{y}_0) = \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i - \beta_0 - \beta_1 x_0 - \beta_2 x_0^2 - \varepsilon$$

$$Var(\hat{y}_0) = \sigma^2 \sum_{i=1}^n c_i^2$$