Домашнее задание №1

Агалян Роберт БПТ214

Сначала загрузим библиотеку **ggplot2** для визуализации, откроем необходимый датафрейм и загрузим его в переменную **insure**. Построим модель линейной регрессии и сохраним её в переменную **fit1**

```
library(ggplot2)
insure <- read.csv(file.choose())
fit1 <- lm(charges ~ age, data=insure)
```

Задание 1

Пункт 1

Зависимая переменная - *charges*
Независимая переменная - *age*

```
fit1 <- lm(charges ~ age, data=insure)
summary(fit1)
```

```
##
## Call:
## lm(formula = charges ~ age, data = insure)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8627  -7155  -6258   5350  47213
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2268.26    1262.90   1.796   0.0729 .
## age           293.23      30.48   9.621   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12150 on 798 degrees of freedom
## Multiple R-squared:  0.1039, Adjusted R-squared:  0.1028
## F-statistic: 92.56 on 1 and 798 DF,  p-value: < 2.2e-16
```

Пункт 2

intercept Если возраст получателя медицинской страховки равен 0, то, согласно нашей модели, средний показатель расходов на медицинские услуги равен $2268.26
age Если мы будем сравнивать получателей медицинской страховки, которые отличаются лишь по возрасту, там, где возраст будет больше на единицу, ожидаемый расход на медицинские услуги, в среднем, выше на $293.23
На 5%-м уровне значимости при p-value = <2e-16 (крайне близко к нулю, приблизительно 0.0000000000000002) связь между переменными статистически значима.

Также мы можем обратить внимание на “код значимости” справа от p-value коэффициента *age*. В данном случае это три звёздочки, что свидетельствует о том, что коэффициент значим на уровне значимости 0.001 (0.1%), следовательно он значим и на 5% уровне значимости.

Пункт 3

$cha\hat{r}ges = 2268.26 + 293.23 * age$

Пункт 4

Найдём среднее значение возраста:

```
summary(insure$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   18.00   26.00   39.00   38.96   50.00   64.00
```

$cha\hat{r}ges = 2268.26 + 293.23 * 38.96 = 13692.5$

Задание 2

Для начала добавим в датафрейм столбец с остатками модели:

```
insure$residuals <- fit1$residuals
```

Выведем описательные статистики по остаткам модели:

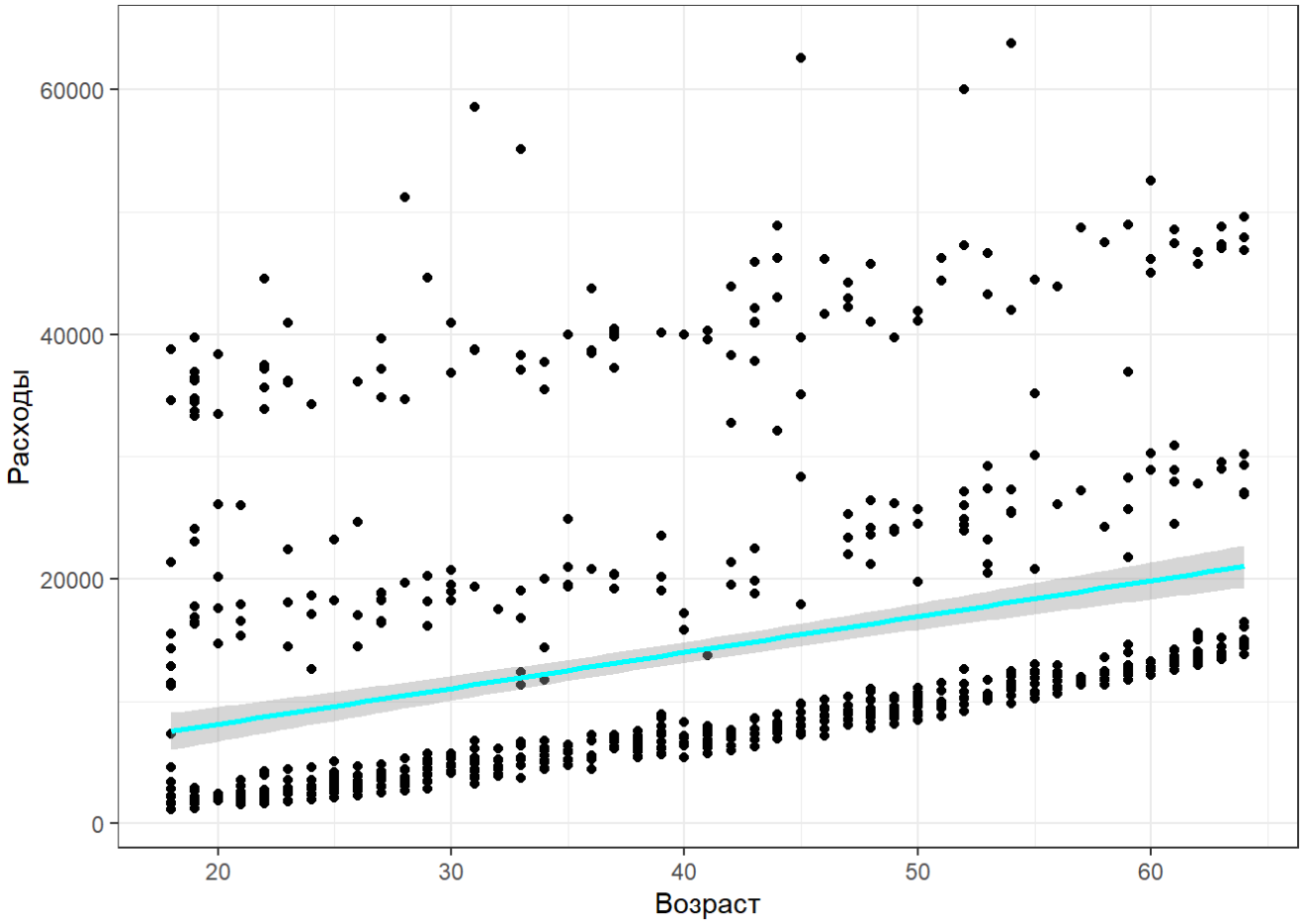
```
summary(insure$residuals)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   -8627   -7155   -6258       0    5350   47213
```

Среднее равно нулю, однако лучше уточним эти данные исходя из графиков.

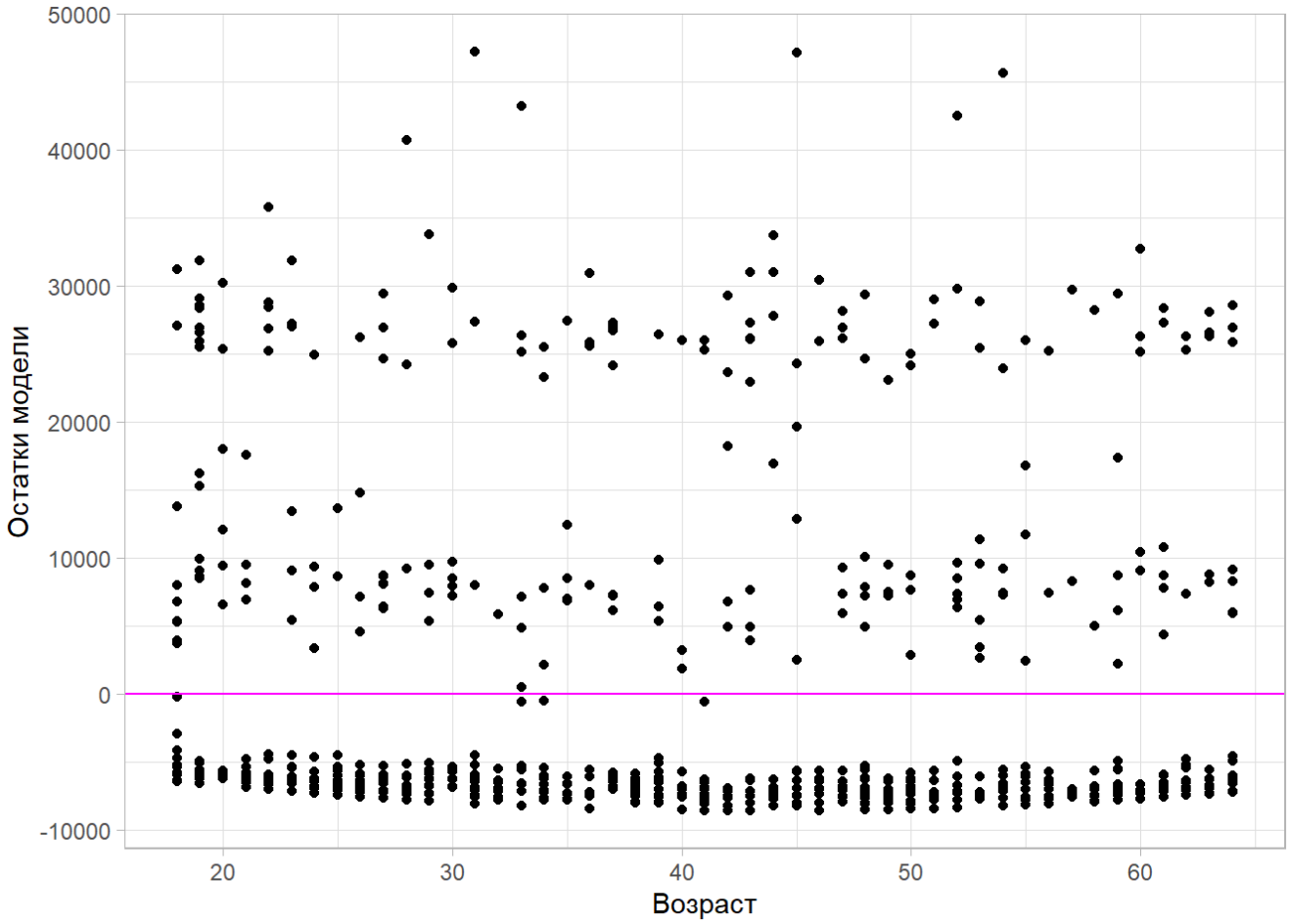
```
ggplot(data = insure, aes(x = age, y = charges)) +
  geom_point() +
  labs(x = "Возраст", y = "Расходы") +
  geom_smooth(method = "lm", color = "cyan") +
  theme_bw()
```

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



Итак, нельзя определённо сказать, что количество точек под и над регрессионной прямой одинаково. Скорее нет, чем да. Условие не выполняется.

```
ggplot(data = insure, aes(x = age, y = residuals)) +
  geom_point() +
  labs(x = "Возраст", y = "Остатки модели") +
  geom_hline(yintercept = 0, color = "magenta") +
  theme_light()
```

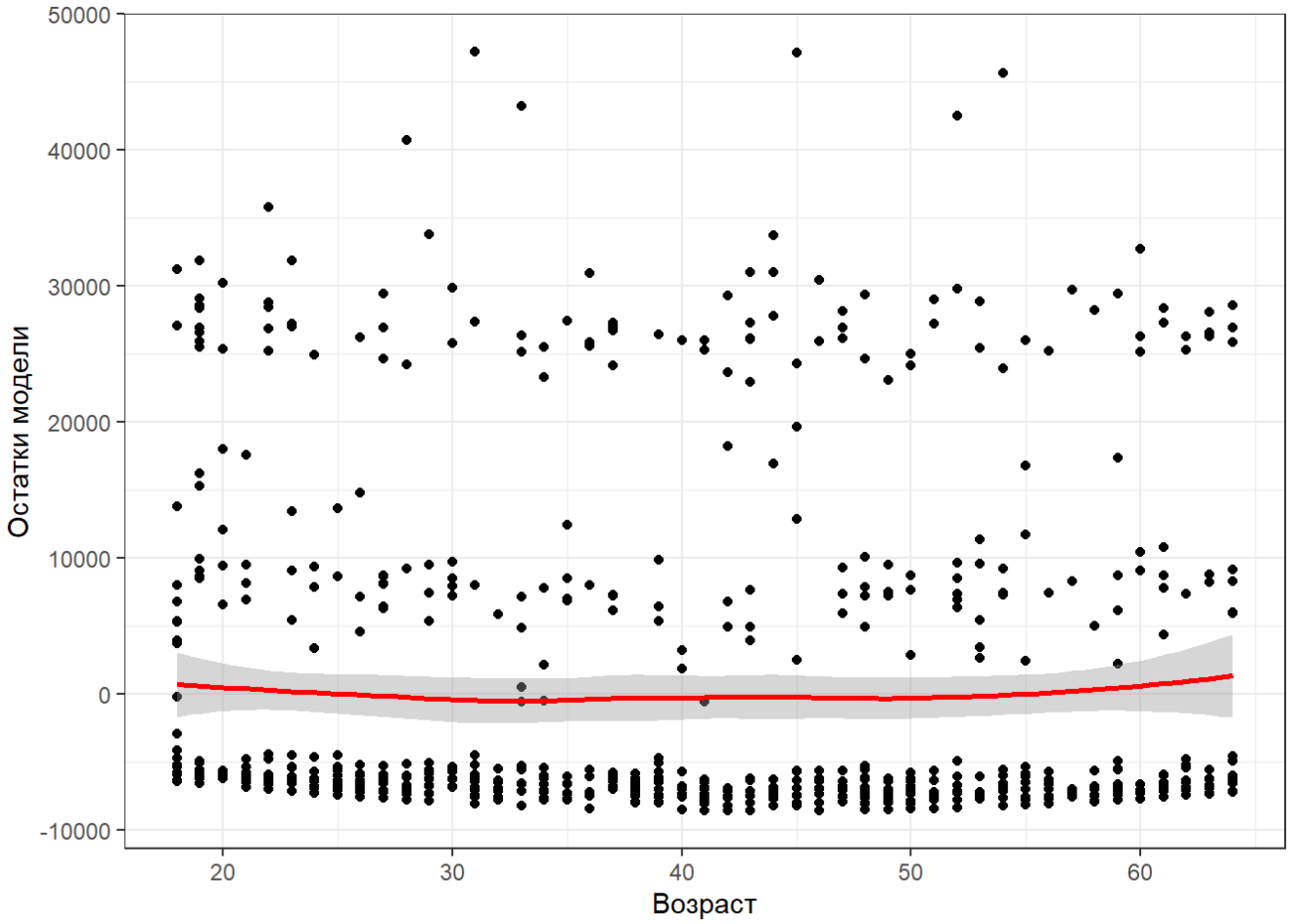


Исходя из этого графика можем сделать вывод об практически полной гомоскедастичности. Изменчивость остатков практически постоянно. Условие выполняется.

Также можем сказать, что остатки независимы от переменной *age*, однако для уточнения этого построим следующий график:

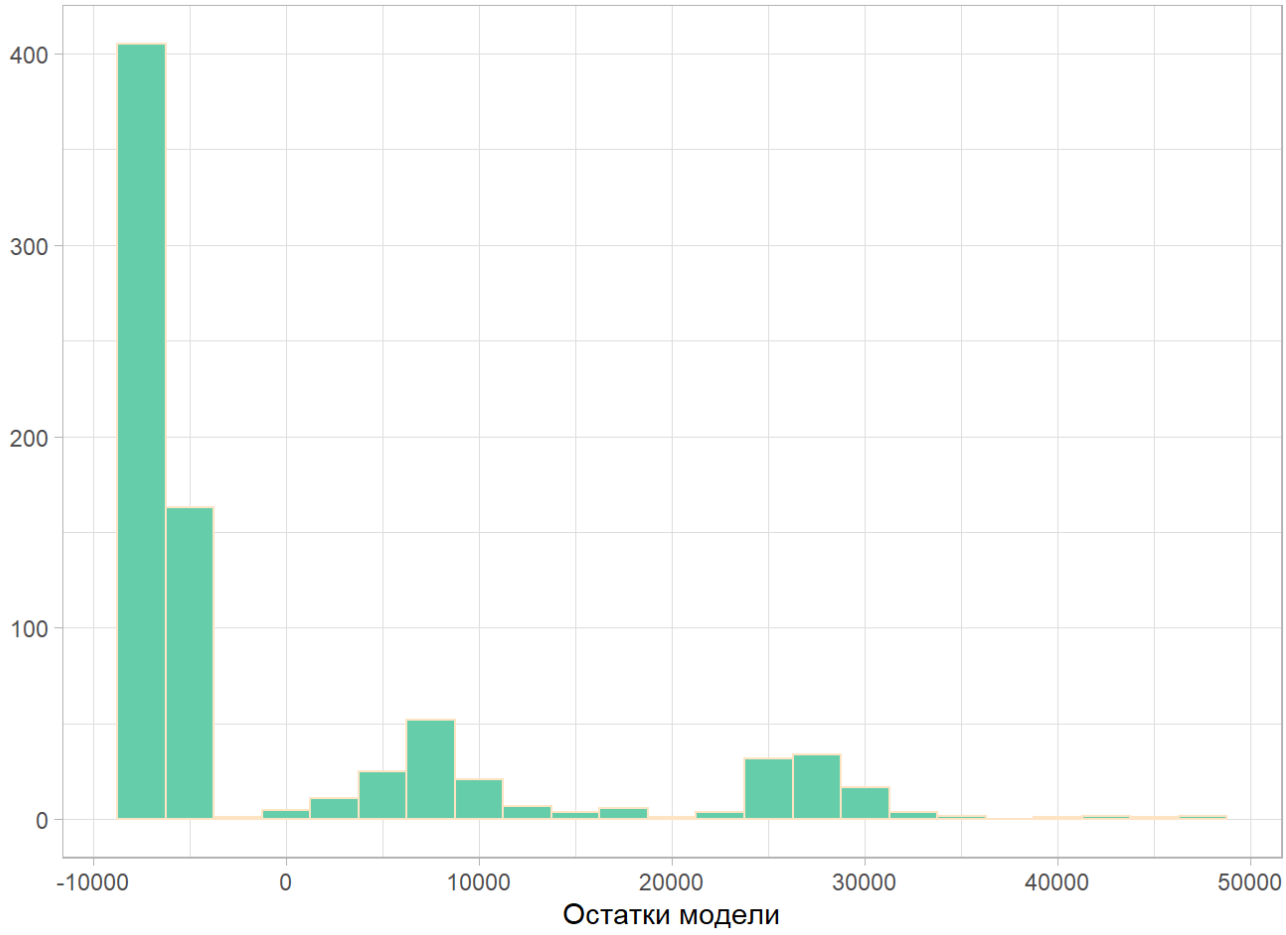
```
ggplot(data = insure, aes(x = age, y = residuals)) +
  labs(x = "Возраст", y = "Остатки модели") +
  geom_point() +
  geom_smooth(color = "red") +
  theme_bw()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Построив линию взвешенной регрессии, можем увидеть, что зависимость между переменной *age* и остатками практически отсутствует, кроме позиций, где возраст превышает 60. Однако в целом условие экзогенности соблюдается.

```
ggplot(data = insure, aes(x = residuals)) +
  geom_histogram(fill = "aquamarine3", color = "bisque1", binwidth = 2500) +
  labs(x = "Остатки модели", y = "") +
  theme_light()
```



Условия о нормальном распределении остатков не соблюдается, распределение не соответствует нормальному.

Задание 3

В выдаче по модели мы уже получали коэффициента детерминации полученной модели. Вывеедем его ещё раз:

```
summary(fit1)$r.squared
```

```
## [1] 0.1039363
```

Исходя из выдачи получается, что модель объясняет ~10.4% изменчивости зависимой переменной. Модель объясняет малую долу изменчивости.