1.1 **Описательная статистика. Корреляционный анализ**

Цель работы: изучение корреляционного анализа и явлений ложной корреляции.

Ход работы:

В качестве набора данных была выбрана статистика об автомобилях. Переменные: длина, размер двигателя, количество лошадиных сил, потребление топлива в городе.

Результат дескриптивного анализа представлен на рисунке 1.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
length	1	205	174.05	12.34	173.2	173.79	10.23	141.1	208.1	67	0.15	-0.14	0.86
engine_size	2	205	126.91	41.64	120.0	120.58	34.10	61.0	326.0	265	1.92	5.07	2.91
horsepower	3	203	104.26	39.71	95.0	99.22	37.06	48.0	288.0	240	1.37	2.48	2.79
city_mpg	4	205	25.22	6.54	24.0	24.76	7.41	13.0	49.0	36	0.65	0.50	0.46

Рисунок 1 – Результат дескриптивного анализа

```
$length

        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.98209, p-value = 0.01036


$engine_size

        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.82944, p-value = 3.057e-14


$horsepower

        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.88464, p-value = 2.324e-11


$city_mpg

        Shapiro-Wilk normality test

data:  newX[, i]
W = 0.95722, p-value = 7.824e-06
```

Рисунок 2 - Результат теста Шапиро-Уилка

Поскольку значения p-value переменных меньше чем уровень значимости a = 0.05, то нулевая гипотеза отвергается.

На рисунках 3-6 представлены гистограммы атрибутов. Для построения гистограмм была использована формула Стерджесса:

$$k = 1 + 3.322 * log_{10}(n)$$

где k - количество интервалов, n - объём выборки.

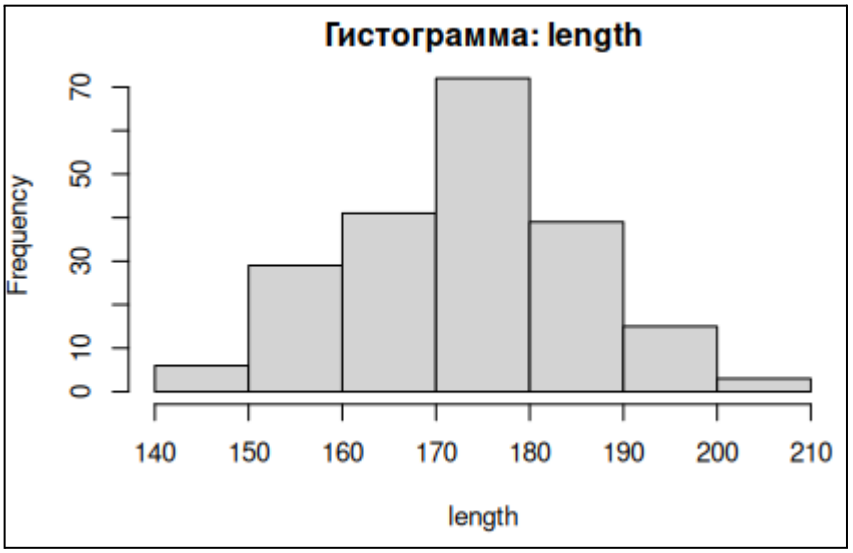

Рисунок 3 – Гистограмма атрибута «length»

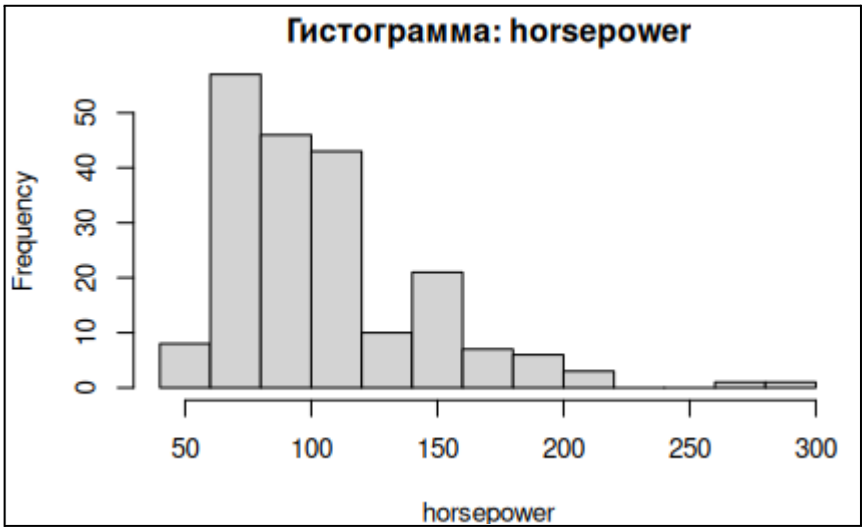

Рисунок 4 – Гистограмма атрибута «horsepower»

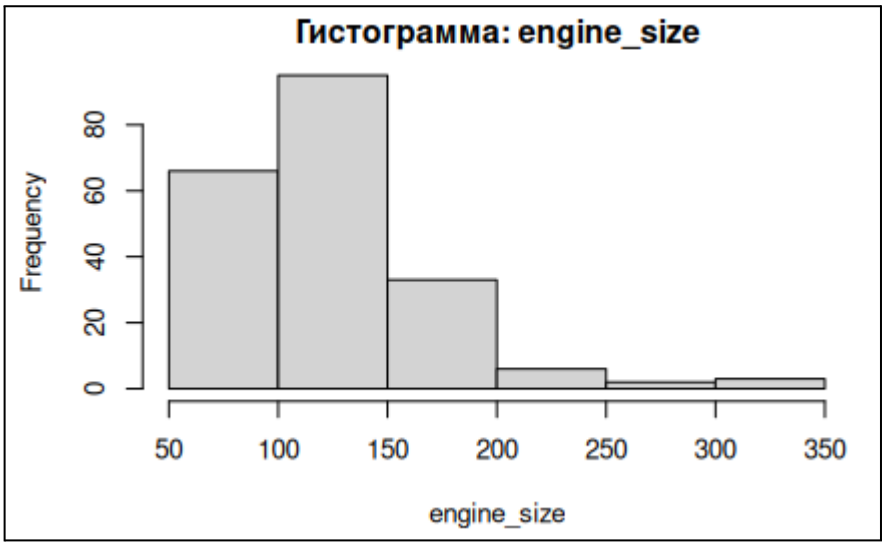

Рисунок 5 – Гистограмма атрибута «engine.size»

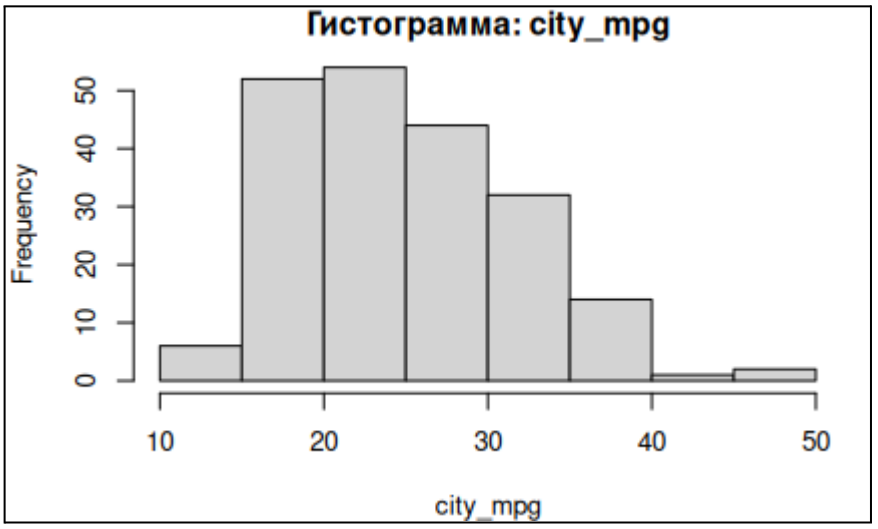

Рисунок 6 – Гистограмма атрибута «city.mpg»

Для построения таблицы сопряженности, переведем переменные «размер двигателя» и «лошадиные силы» в категориальные. Полученная таблица сопряженности представлена на рисунке 7.

```
          HPless110 HPmore110
ESless125       120         10
ESmore125        11         62
>
```

Рисунок 7 – Таблица сопряженности

Для проверки независимости «размер двигателя» от «лошадиные силы» проведем тест хи-квадрат, результат которого можно увидеть на рисунке 8.

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_nomiss
X-squared = 118.5, df = 1, p-value < 2.2e-16
```

Рисунок 8 – Результат теста хи-квадрат

Судя по результату теста хи-квадрат, а именно по значению p-value, можно сказать, что переменные «размер двигателя» и «лошадиные силы» зависят друг от друга.

Проведем попарный анализ и получим диаграммы рассеивания, которые можно увидеть на рисунке 9.

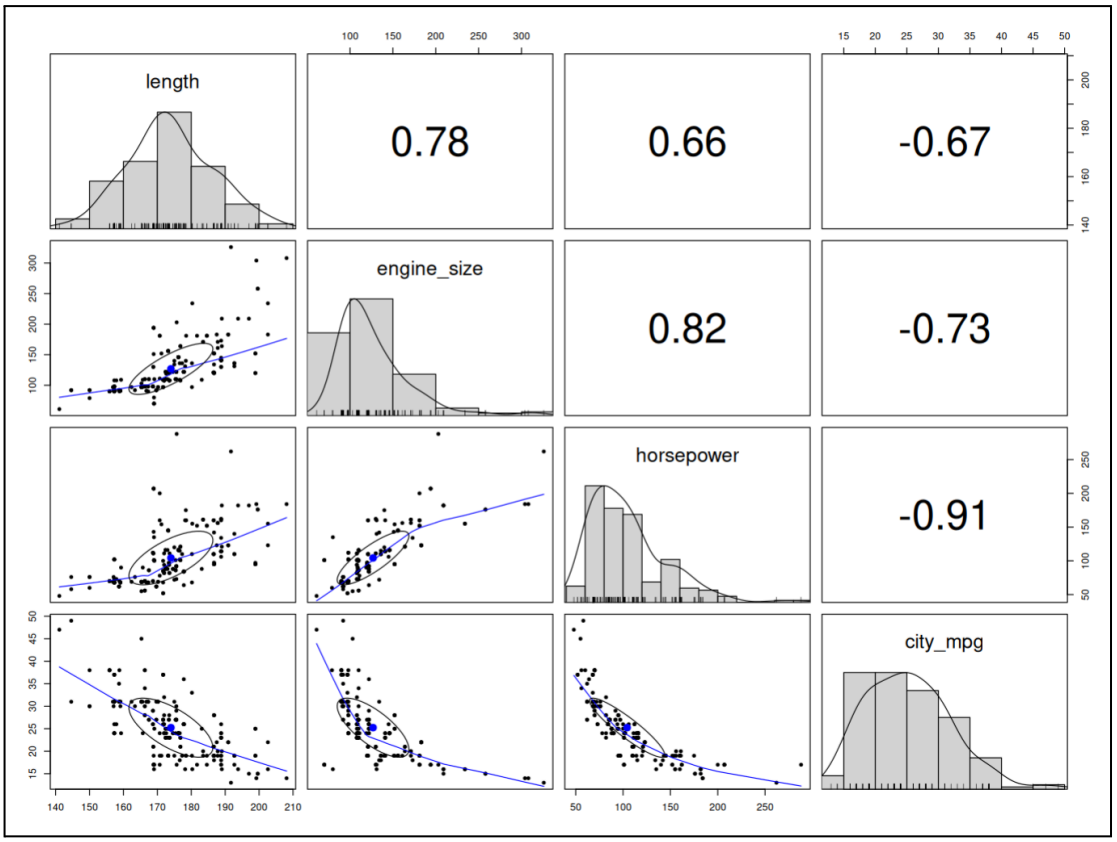

Рисунок 9 – Диаграммы рассеивания

Судя по диаграммам рассеивания, можно предположить наличие корреляции между «длина» и «размер двигателя», «размер двигателя» и «лошадиные силы», «потребление топлива в городе» и «лошадиные силы».

Сформулируем гипотезу о равенстве коэффициента корреляции нулю у выбранных пар переменных. Чтобы подтвердить ее или опровергнуть проведем корреляционный анализ.

```
        Spearman's rank correlation rho

data:  df_eh$engine_size and df_eh$horsepower
S = 251625, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8195209
```

Рисунок 10 – Корреляционный анализ переменных «размер двигателя» и «лошадиные силы»

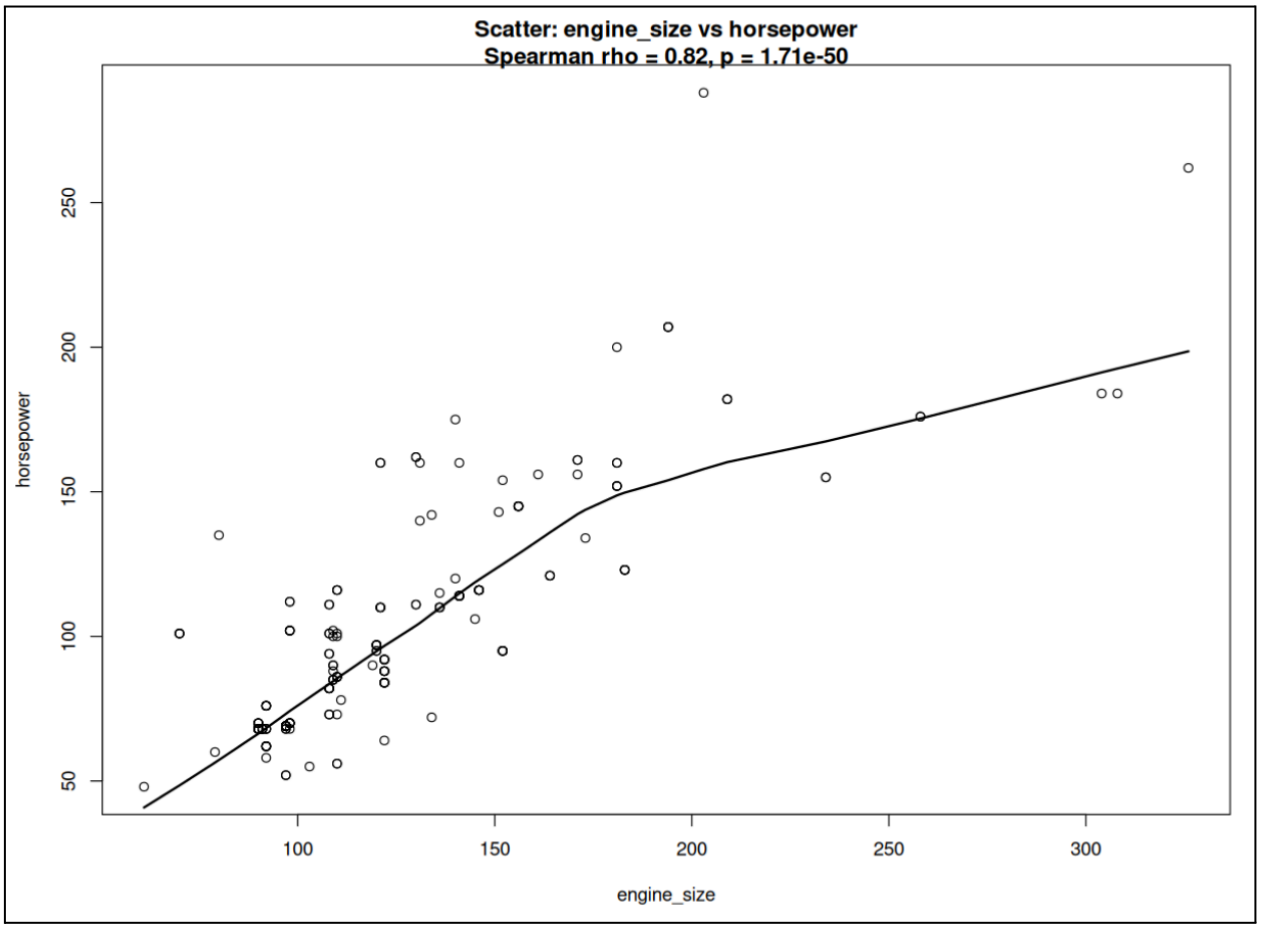

Рисунок 11 – Диаграмма рассеивания переменных «размер двигателя» и «лошадиные силы»

Т.к. p-value < 0.05, то гипотеза о равенстве коэффициента корреляции нулю у выбранных пар переменных опровергается.

```
        Spearman's rank correlation rho

data:  df_le$length and df_le$engine_size
S = 312124, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7826164
```

Рисунок 12 – Корреляционный анализ переменных «длина» и «размер двигателя»

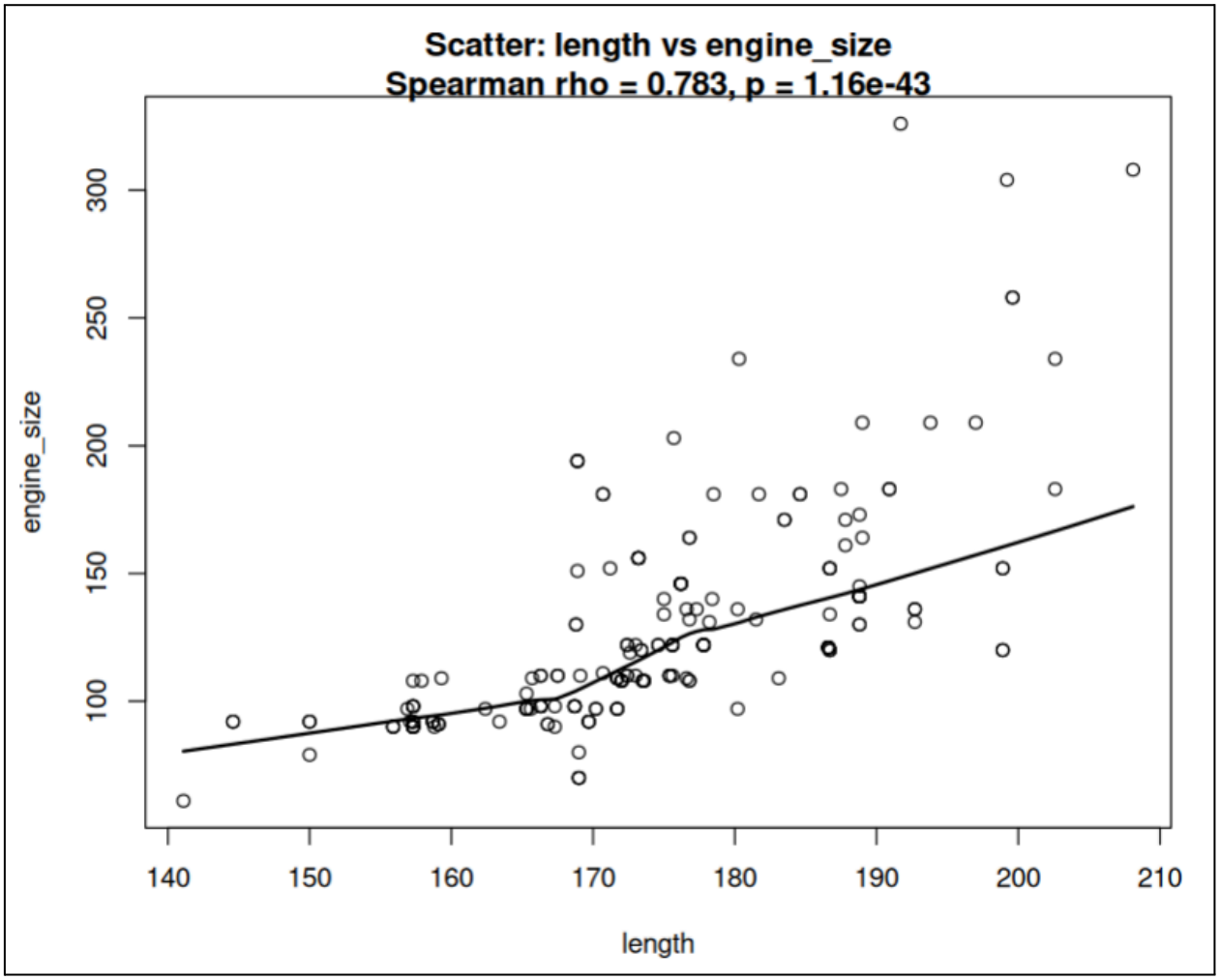

Рисунок 13 – Диаграмма рассеивания переменных «длина» и «размер двигателя»

Т.к p-value<0.05, то гипотеза о равенстве коэффициента корреляции нулю у выбранных пар переменных опровергается.

```
        Spearman's rank correlation rho

data:  df_hc$horsepower and df_hc$city_mpg
S = 2666422, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.9125048
```

Рисунок 14 – Корреляционный анализ переменных «лошадиные силы» и «потребление топлива в городе»

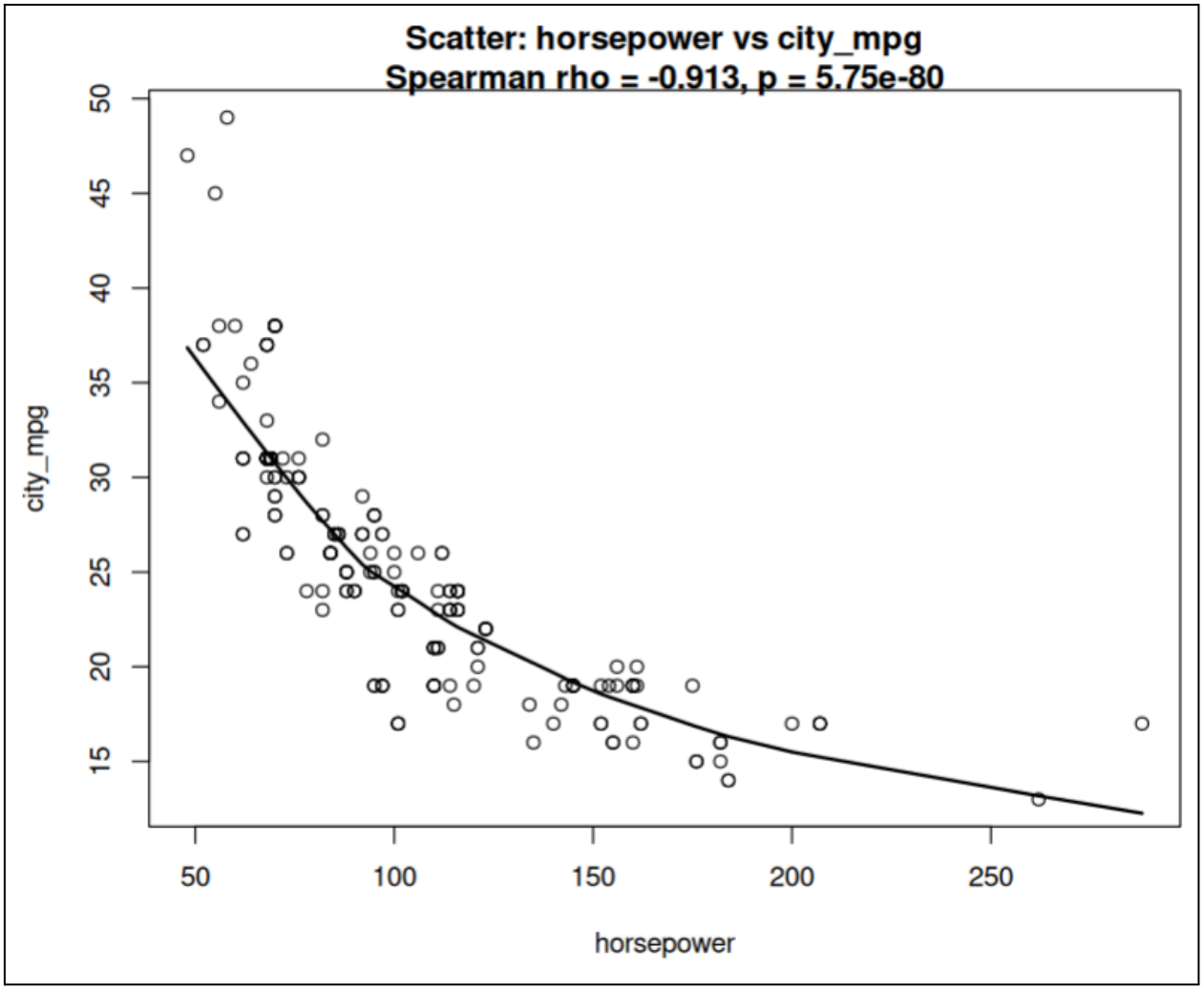

Рисунок 15 – Диаграмма рассеивания переменных «лошадиные силы» и «потребление топлива в городе»

Т.к. p-value<0.05, то гипотеза о равенстве коэффициента корреляции нулю у выбранных пар переменных опровергается.

Исходя из результатов анализа, можно сказать, что пара переменных «лошадиные силы» и «потребление топлива в городе» наиболее

коррелированные, а пара переменных «длина» и «размер двигателя» наименее.

1.2 Регрессионный анализ

Цель работы: изучение регрессионного анализа.

Ход работы:

Изменять набор данных не будем. Построим линейную регрессионную модель для переменных «размер двигателя» и «лошадиные силы». В качестве независимой переменной будет «размер двигателя», а зависимой «лошадиные силы».

```
Residuals:
    Min      1Q  Median      3Q     Max
-80.825  -8.274   0.156   8.437 113.020

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.7938     4.8519   7.789 3.52e-13 ***
horsepower    0.8543     0.0435  19.637  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.56 on 201 degrees of freedom
Multiple R-squared:  0.6574,    Adjusted R-squared:  0.6556
F-statistic: 385.6 on 1 and 201 DF,  p-value: < 2.2e-16
```

Рисунок 16 – линейная регрессионная модель

По полученным p-value < 0.05 можно судить о значимости свободного члена и horsepower. А также по полученному p-value можно сказать о значимости модели. На рисунках 17 и 18 можно увидеть график модели и график остатков.

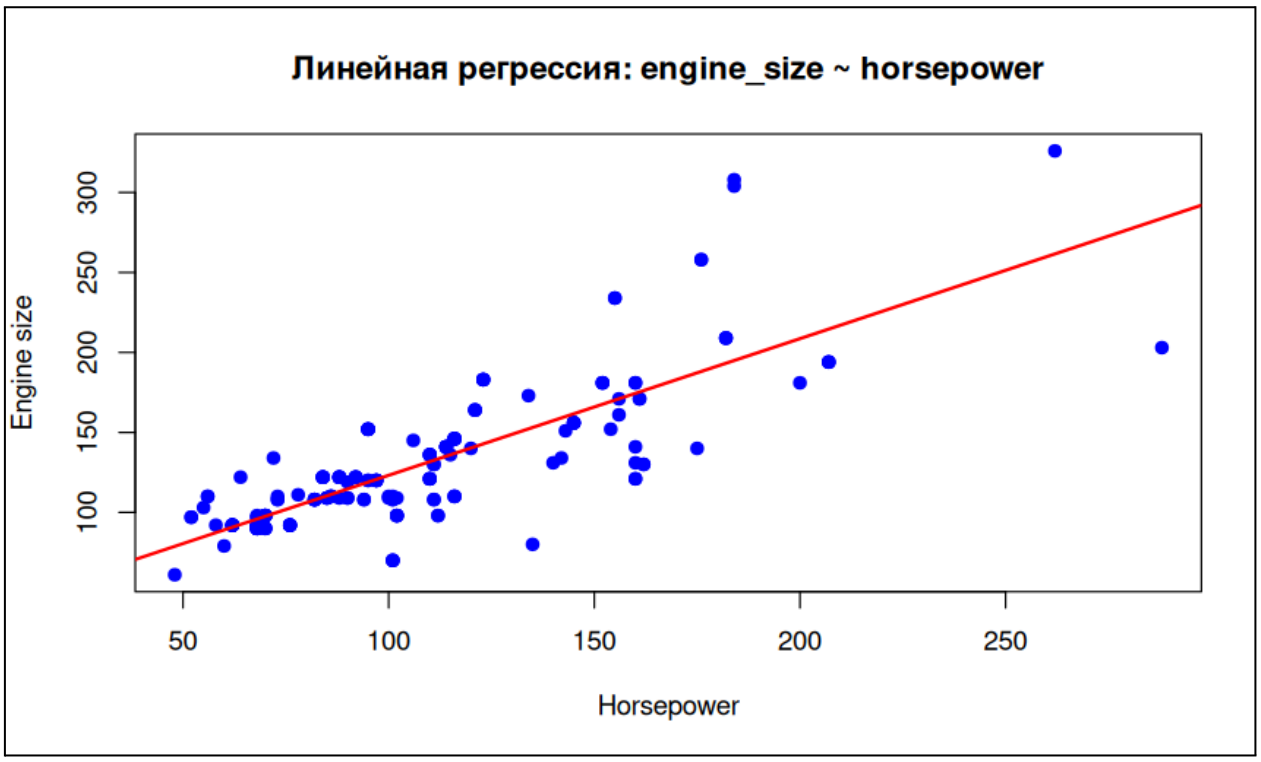

Рисунок 17 – График модели

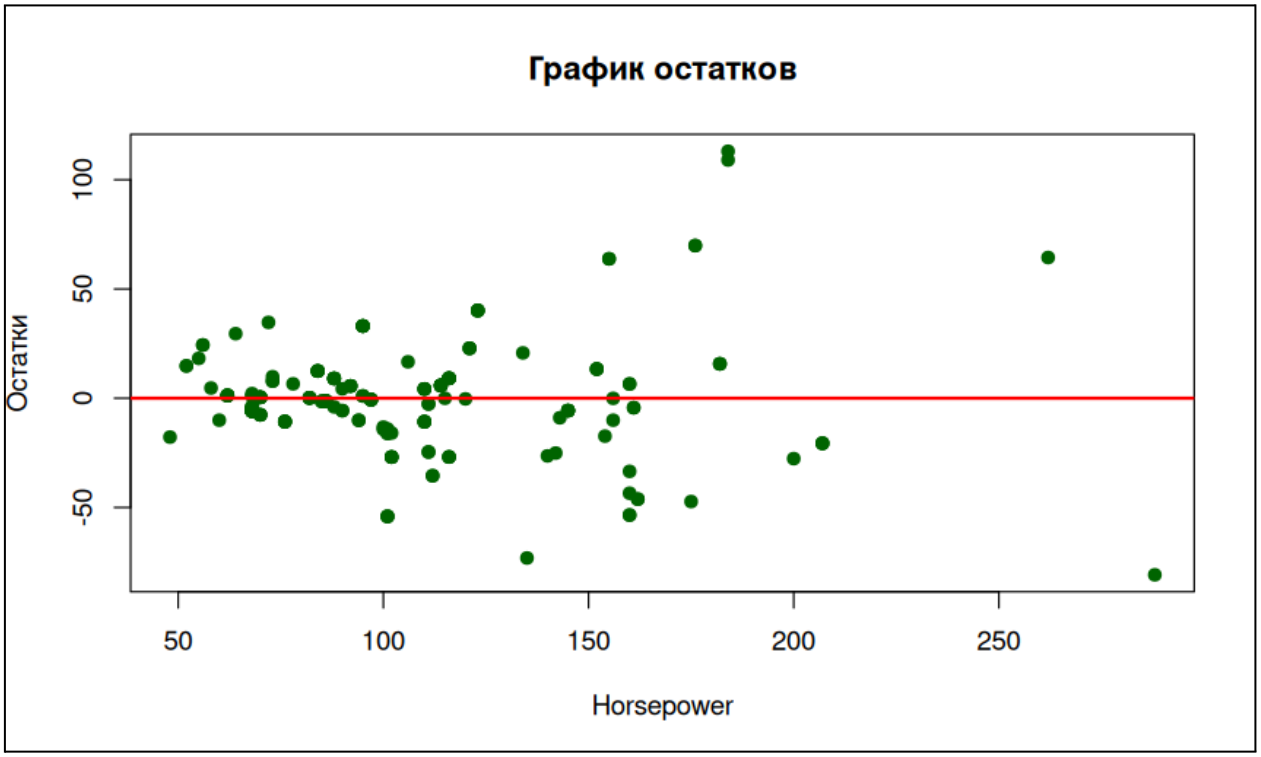

Рисунок 18 – График остатков

По графику остатков можно сказать, что дисперсия не постоянна: при малых значениях horsepower разброс относительно небольшой, а при

больших значениях — заметно возрастает. Среднее можно считать постоянным.

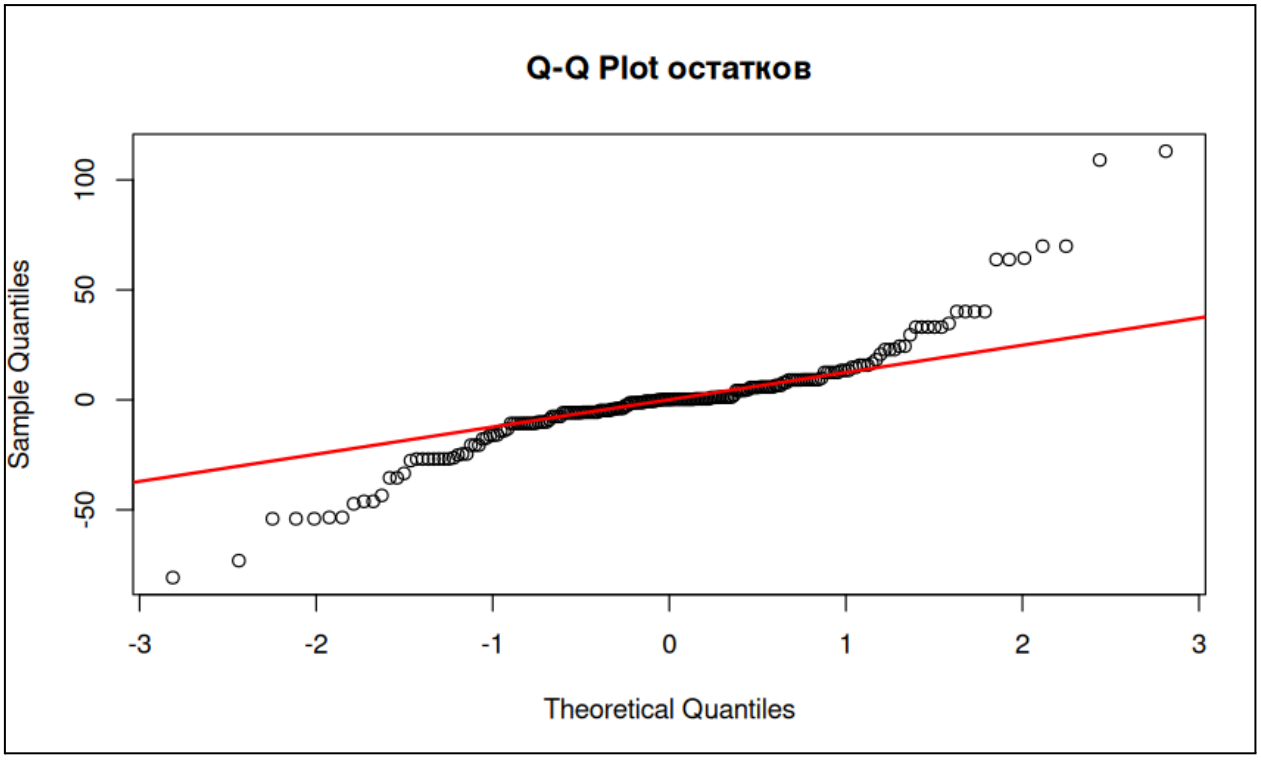

Рисунок 19 – График Квантиль-Квантиль остатков

По этому графику можно предположить, что остатки приблизительно нормальны в центральной части распределения. Но есть отклонения от нормальности в хвостах, что говорит о наличии выбросов или ненормальных крайних значений.

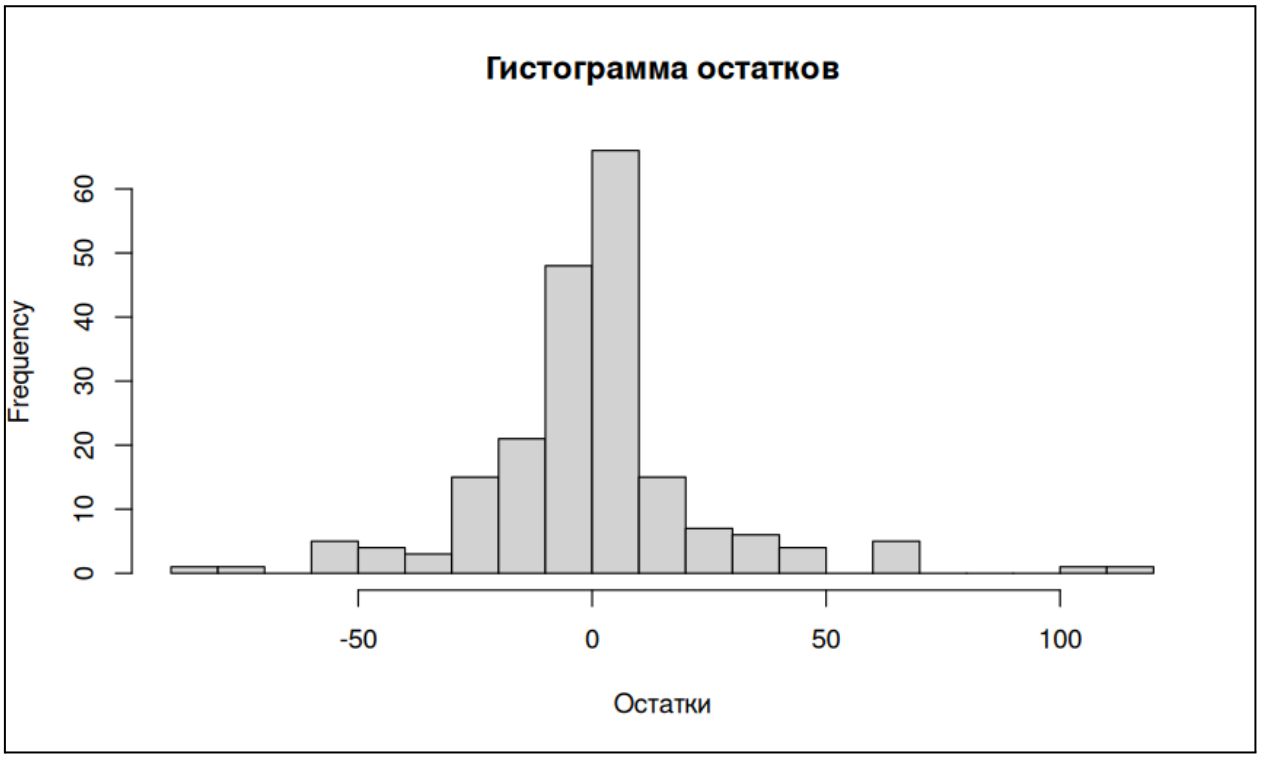

Рисунок 20 – Гистограмма остатков

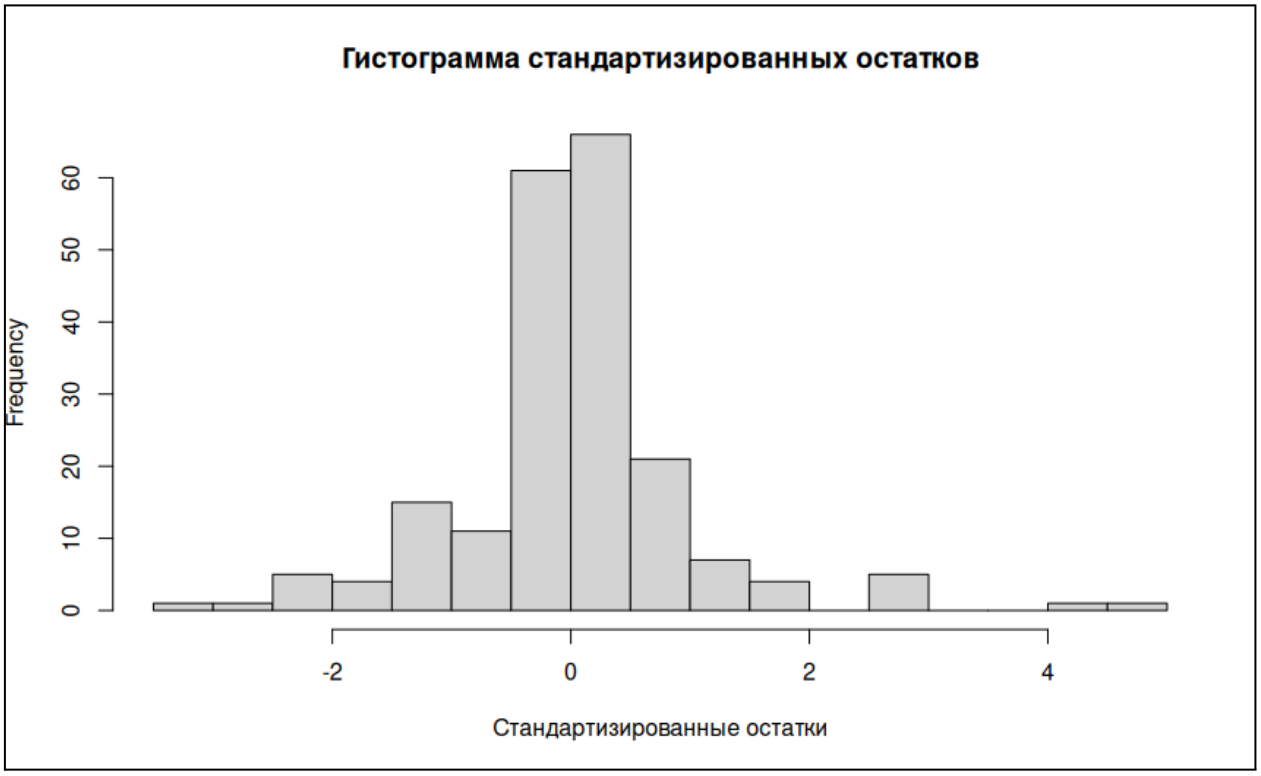

Рисунок 21 – Гистограмма стандартизированных остатков

Остатки приблизительно соответствуют нормальному распределению, особенно в центральной части. Однако есть отклонения в хвостах

(наблюдаются выбросы). Об этом же говорит график Квантиль-Квантиль остатков.

Получим следующее уравнение регрессии:

```
Engine_size = 37.79 + 0.85 * Horsepower
```

Коэффициент 0.85 показывает, что при увеличении л.с. на 1 единицу размер двигателя в среднем возрастает на 0.85 куб. дюймов. Константа 37.79 отражает теоретическое значение размера двигателя при нулевом числе лошадиных сил (в данном случае – лишь формальный параметр).

На рисунке 16 видно, что значение детерминации R^2 равно $R^2 = 65.6\%$. Оно говорит о среднем качестве модели.

1.3 Дисперсионный анализ

Цель работы: изучение дисперсионного анализа.

Ход работы:

Изменять набор данных не будем. Гипотеза H0 – **количество лошадиных сил не зависит от размера двигателя**. Поделим «размер двигателя» на 3 равных группы по квантилям (0; 0.33; 0.66; 1):

```
Small Medium  Large
   80     56     67
```

Проведем однофакторный дисперсионный анализ, разделив «лошадиные силы» по «размеру двигателя». Результат анализа видно на рисунке 22.

```
              Df Sum Sq Mean Sq F value Pr(>F)
engine_group   2 164571   82286   106.8 <2e-16 ***
Residuals    200 154029     770
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Рисунок 22 – Однофакторный дисперсионный анализ

По результатам однофакторного дисперсионного анализа (ANOVA) получено значение F = 106.8 при p < 0.001. Это означает, что нулевая гипотеза отвергается, и число лошадиных сил статистически значимо зависит от объема двигателя.

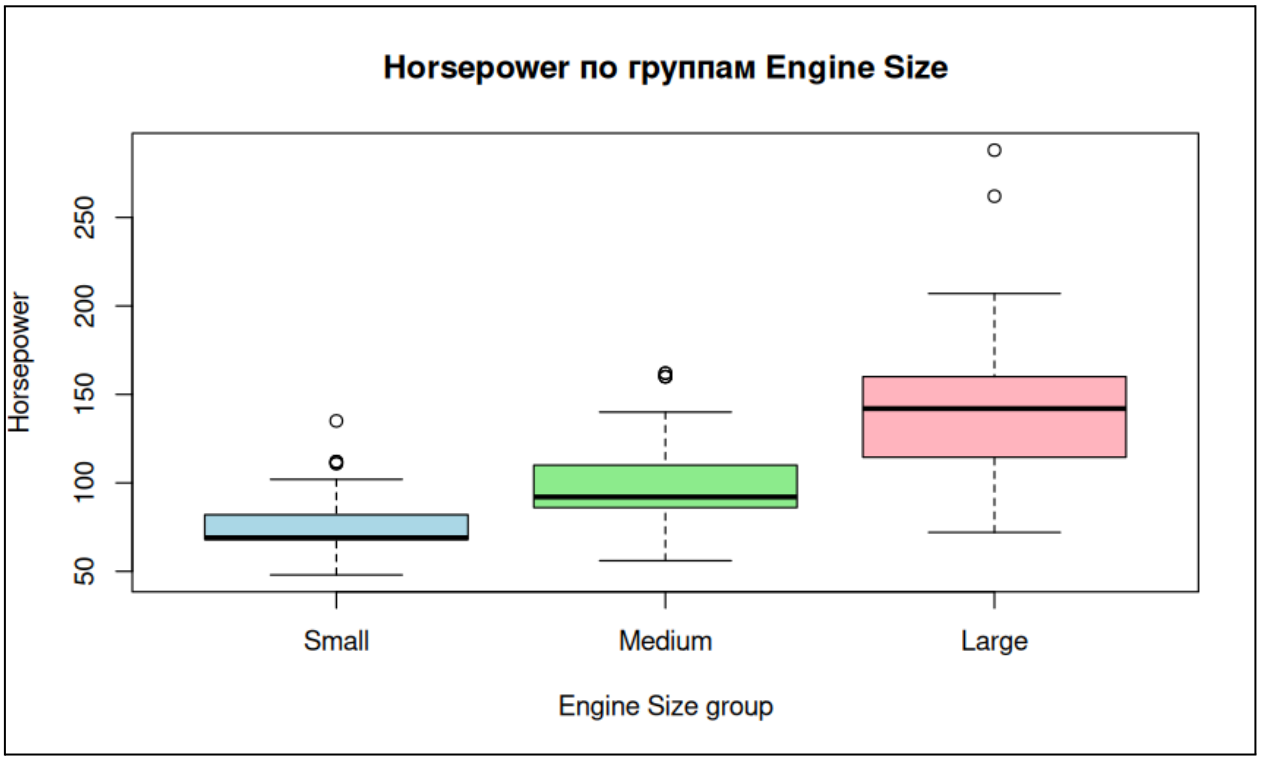

Рисунок 23 – Распределение «лошадиных сил» по «размеру двигателя»

Исходя из результатов анализа, гипотезу можно отвергнуть, т.к p-value<0.05, т.е. число лошадиных сил явно зависит от размера двигателя.

Проведем двухфакторный дисперсионный анализ с engine_group и fuel_type:

```
                     Df Sum Sq Mean Sq F value   Pr(>F)
engine_group          2 164571   82286 127.439  < 2e-16 ***
fuel_type             1  24352   24352  37.714 4.44e-09 ***
engine_group:fuel_type   2   2477    1238   1.918     0.15
Residuals           197 127201     646
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Рисунок 24 – Двухфакторный дисперсионный анализ

Двухфакторный дисперсионный анализ показал, что оба фактора - engine_size и fuel_type - оказывают статистически значимое влияние на число лошадиных сил ($p < 0.001$). При этом взаимодействие факторов (engine_size:fuel_type) оказалось статистически незначимым ($p = 0.15$).

Таким образом, каждый из факторов по отдельности влияет на число лошадиных сил, но их совместный эффект не имеет дополнительного вклада.

По результатам однофакторного анализа нулевая гипотеза была отвергнута, по результатам двухфакторного анализа гипотеза о независимости числа лошадиных сил от типа топлива и от размера двигателя была отвергнута.

1.4 Логарифмический регрессионный анализ

В качестве независимой переменной будет «размер двигателя», а зависимой «лошадиные силы». Проведем логарифмический регрессионный анализ, результат которого видно на рисунке 25.

```
Call:
lm(formula = engine_size ~ log(horsepower), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-77.417  -8.217  -1.162   6.627 121.117

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -309.350     24.113  -12.83   <2e-16 ***
log(horsepower)   95.156      5.245   18.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.83 on 201 degrees of freedom
Multiple R-squared:  0.6208,    Adjusted R-squared:  0.619
F-statistic: 329.1 on 1 and 201 DF,  p-value: < 2.2e-16
```

Рисунок 25 – Логарифмический регрессионный анализ

По полученным p-value < 0.05 можно судить о значимости свободного члена и horsepower. А также, по полученному p-value можно сказать о

значимости модели. На рисунках 26 и 27 можно увидеть график модели и график остатков.

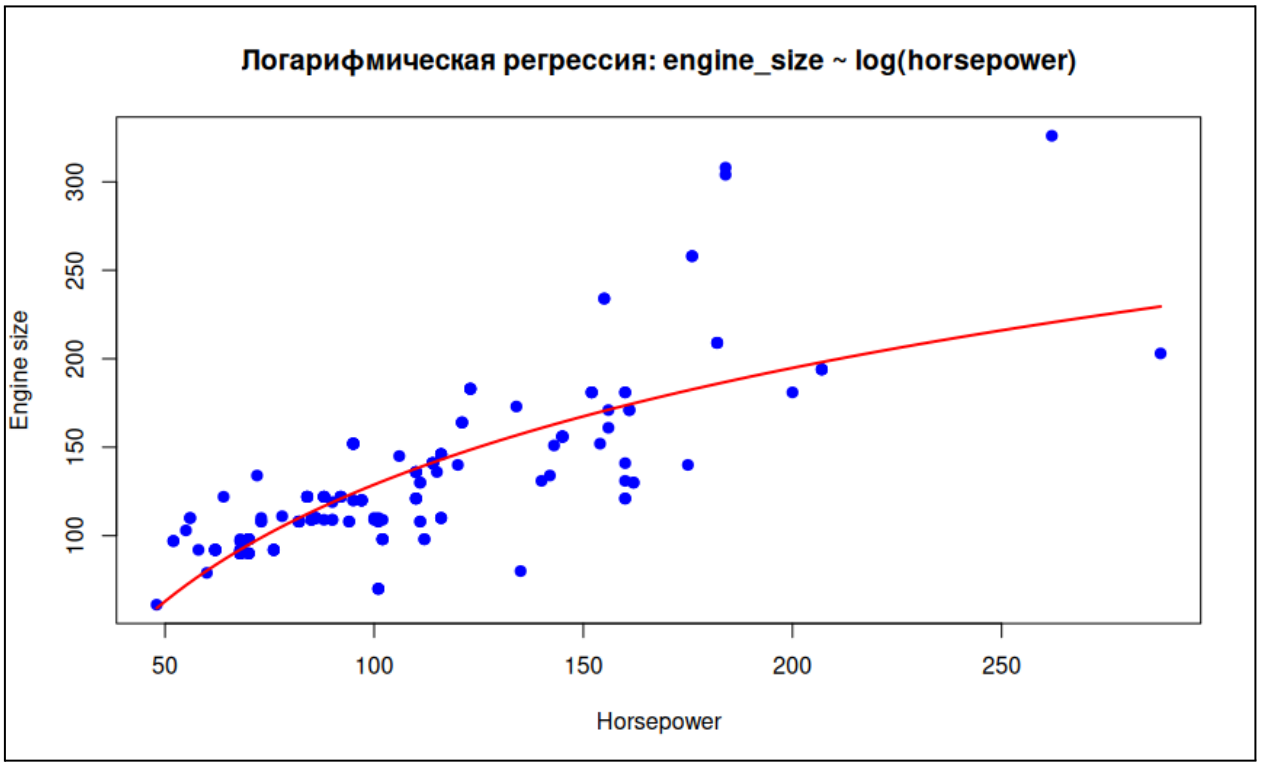

Рисунок 26 – График модели

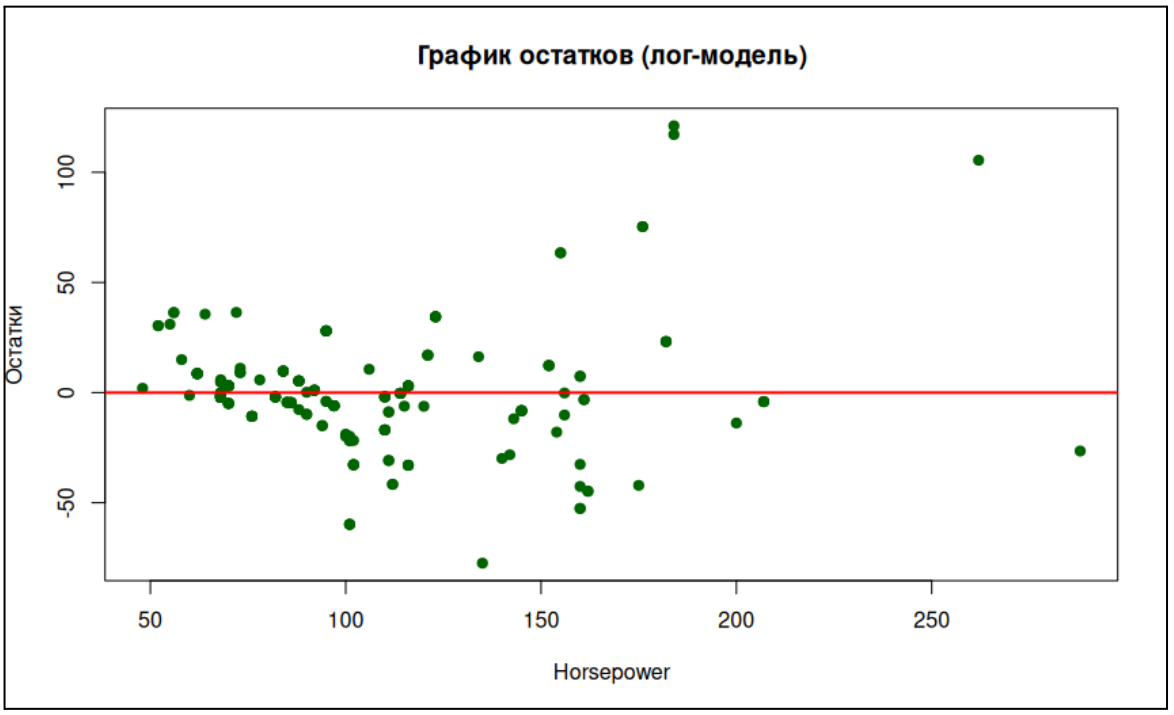

Рисунок 27 – График остатков

По графику остатков можно сказать, что дисперсия не постоянна: при малых значениях horsepower разброс относительно небольшой, а при больших значениях — заметно возрастает. Среднее можно считать постоянным.

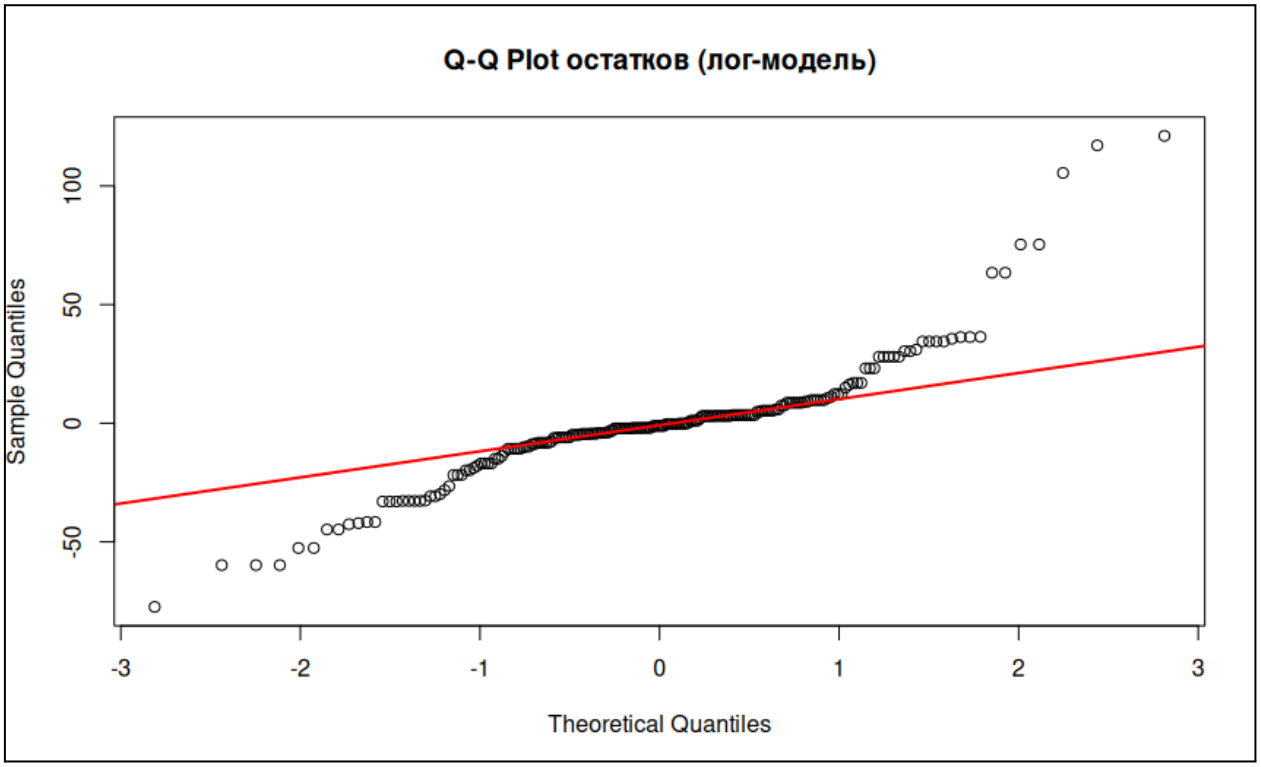

Рисунок 28 – График Квантиль-Квантиль остатков

По этому графику можно предположить, что остатки приблизительно нормальны в центральной части распределения. Но есть отклонения от нормальности в хвостах, что говорит о наличии выбросов или ненормальных крайних значений.

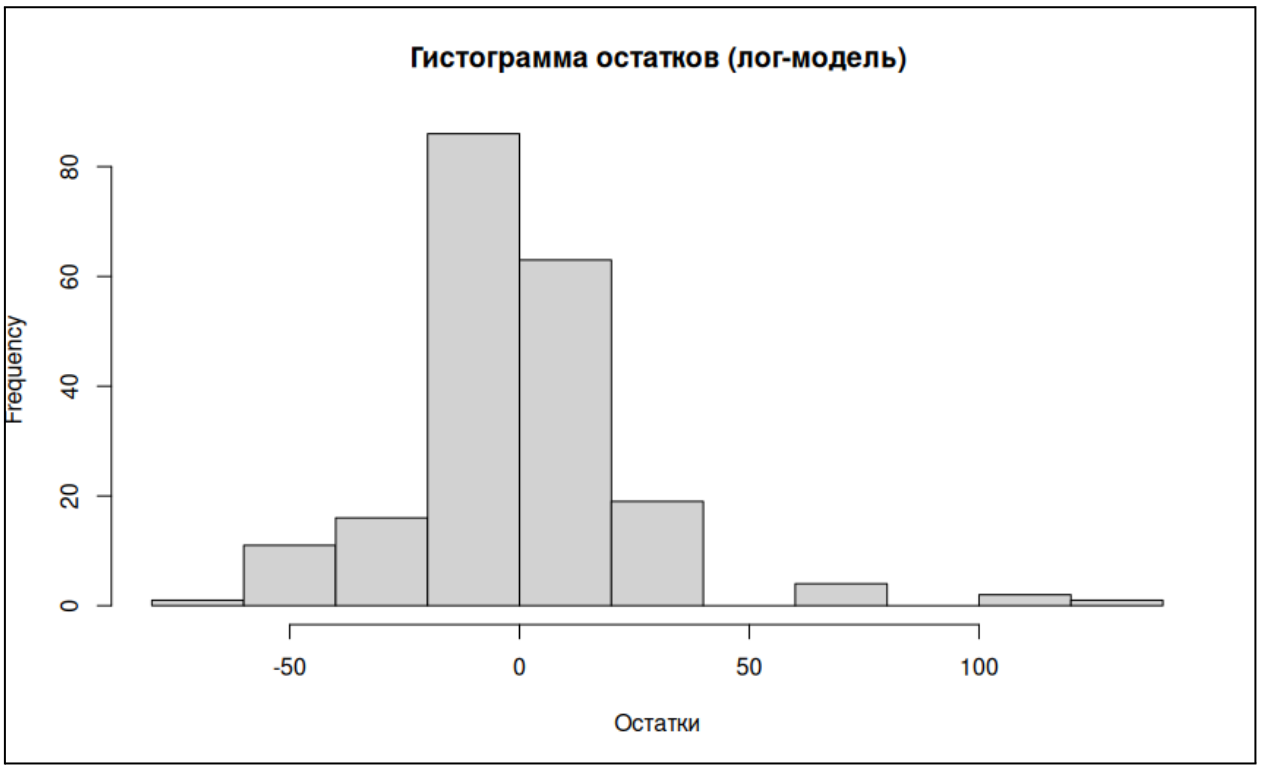

Рисунок 29 – Гистограмма остатков

Исходя из полученной гистограммы, можно сказать, что остатки не подчиняются нормальному распределению.

Получили следующее уравнение логарифмической регрессии:

```
Engine_size = -309.35 + 95.16 * log(Horsepower)
```

На рисунке 25 видно, что значение детерминации равно 61.9%. Оно говорит о среднем качестве модели. Данное значение меньше рассчитанного значения линейной модели (65.6%). Из этого можно сделать вывод, что модель, основанная на линейной регрессии является более точной.