Шишлевская Валерия Васильевна
БГУ222(в ведомости)

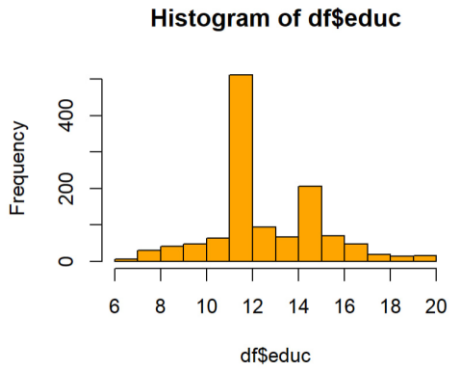
Задача: рассмотреть зависимость величины почасовой оплаты в городских районах от образования и опыта работы на примере данных о работающих мужчинах США в 1991 году. И ответить на вопрос: *что вносит больший вклад в зарплату, опыт работы или число лет обучения?*

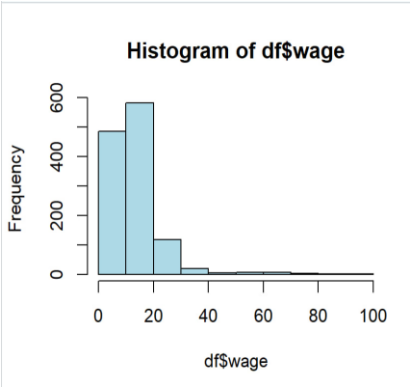
Обзор переменных:

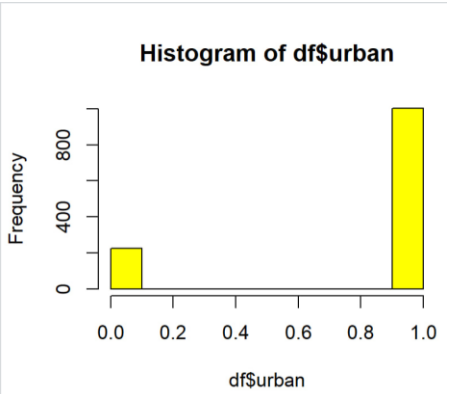
Зависимая переменная:	wage - почасовая оплата в $
Независимые переменные:	educ - число лет обучения к 1991 году
	exper - опыт работы к 1991 году
Дамми-переменная:	1, городская зона
(независимая)	0, не городская зона

Описательная статистика:

wage	educ	exper	urban
Min. : 1.024	Min. : 6.00	Min. : 1.00	Min. :0.0000
1st Qu.: 7.942	1st Qu.:12.00	1st Qu.: 9.00	1st Qu.:1.0000
Median :11.543	Median :12.00	Median :11.00	Median :1.0000
Mean :13.288	Mean :13.04	Mean :10.74	Mean :0.8171
3rd Qu.:16.026	3rd Qu.:15.00	3rd Qu.:13.00	3rd Qu.:1.0000
Max. :91.309	Max. :20.00	Max. :21.00	Max. :1.0000







Посмотрим на корреляцию, используя метод Пирсона

Между переменными wage и educ, p-value =2.2e-16 < 0.05, что свидетельствует о статистической зависимости, а корреляция 0.36 показывает умеренную положительную связь и следовательно, при увеличении числа лет обучения, заработная плата увеличивалась.

Между переменными wage и exper, p-value =1.948e-16 < 0.05, что свидетельствует о статистической зависимости, а корреляция -0.18 показывает слабую отрицательную связь и следовательно, что при увеличении опыта работы, заработная плата незначительно уменьшалась.

Спецификация модели ***множественной регресси***и следующая:
$$wage = a_0 + a_1 * educ + a_2 * exper + a_3 * urban + \varepsilon$$

```
Call:
lm(formula = wage ~ educ + exper + urban, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-18.551  -4.481  -1.052   2.606  69.130

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.3304     2.8305  -8.242 4.30e-16 ***
educ          1.9814     0.1376  14.398  < 2e-16 ***
exper         0.7470     0.1069   6.986 4.64e-12 ***
urban         3.3806     0.6350   5.324 1.21e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.27 on 1226 degrees of freedom
Multiple R-squared:  0.1728,    Adjusted R-squared:  0.1708
F-statistic:  85.4 on 3 and 1226 DF,  p-value: < 2.2e-16
```

T
Ответ на вопрос: Рассмотрим доверительные интервалы для модели множественной регрессии:

```
> confint(model1, level=0.95)
                 2.5 %      97.5 %
(Intercept) -28.8836397 -17.7771943
educ          1.7114142   2.2514021
exper         0.5372113   0.9568045
urban         2.1348345   4.6263372
```

Нижняя граница доверительного интервала для образования(1.71) больше, чем верхняя граница для стажа(0.95) и следовательно, образование вносит больший вклад в заработную плату.

Мне первоначально показалось странным, что при увеличении опыта работы, зарплата уменьшается, но скорее всего это говорит о том, что при увеличении опыта, растет возраст сотрудника, и в какой-то момент зависимость перестает быть линейной, и становится квадратичной (при построении регрессии необходимо будет это учесть)

С помощью ***полиномиальной регрессии*** учтем возраст:

Проверим стандартные нулевые гипотезы для коэффициентов регрессии:

Спецификация модели следующая:
$$wage = a_0 + a_1 * educ + a_2 * exper + a_3 * exper^2 + a_4 * urban + \varepsilon$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.71972    3.56323  -7.779 1.54e-14 ***
educ          2.01583    0.13849  14.555  < 2e-16 ***
exper         1.55594    0.41377   3.760 0.000178 ***
I(exper^2)   -0.03747    0.01852  -2.024 0.043229 *
urban         3.30419    0.63529   5.201 2.32e-07 ***
```

Заметим, что все коэффициенты статистически значимые, то есть нулевые гипотезы отвергаем
Интерпретация:
1) при увеличении числа лет обучения на 1 год, зарплата в среднем увеличивается на 2.015$ в час выше.
2-3) для интерпретации используем предельный эффект. Предельный эффект равен $a_2 + 2*a_3*exper$. То есть изменение зависит от того, насколько опытен человек. Оценим знак
$1.55 - 0.03*2*exper > 0$
$exper < 1.55/0.06 \sim 26$
То есть, если опыт работы больше 26 лет, то каждый новый год влияет негативно, а если меньше, то позитивно. Скорее всего, это связано, с тем, что люди начинают выходить на пенсию, а также с увеличивающимися рисками лишиться работы, как следствие, начинают соглашаться на меньшую з/п
4) В среднем городские жители получают на 3,3 доллара в час выше сельских

Хочу рассмотреть зарплату, как дамми переменную в которой индикатором служит следующее:
wage_above = 1, если зарплата выше 13 долларов в час
wage_abobe = 0, если зарплата ниже 13 долларов в час

wage_above - зависимая переменная и дамми, следовательно, построим ***логит-модель***

Линейный индекс задаём так
wage_above = a_0 + a_1*educ + a_2*exper + a_3*exper^2 + a_4*urban + ε

Доверительные интервалы:

	2.5 %	97.5 %
(Intercept)	0.0000149544	0.0008091848
exper	1.1778814300	1.8618469776
I(exper^2)	0.9776180687	0.9984185149
educ	1.3798927709	1.6156605702
urban	1.5992279783	3.2575446816

Каждая переменная статистически значима, поскольку доверительные интервалы не пересекают 1

(Intercept)	exper	I(exper^2)	educ	urban
0.0001134128	1.4752133889	0.9881696228	1.4911104862	2.2678864928

1) На каждый дополнительный год получения образования, шансы принадлежать к группе, зарабатывающей больше 12 долларов в час, увеличивается на 1.49
2) urban = 2.26. У человека, живущего в городе, шансы зарабатывать больше 12 долларов в час, увеличиваются на 2.26 по сравнению с человеком, живущем не в городе
3) Для интерпретации образования надо посмотреть на средний предельный эффект

Расчет предельных эффектов:

factor	AME	SE	z	p	lower	upper
educ	0.0835	0.0071	11.7738	0.0000	0.0696	0.0974
exper	0.0287	0.0056	5.1260	0.0000	0.0177	0.0397
urban	0.1712	0.0368	4.6530	0.0000	0.0991	0.2433

Как видим средний предельный эффект по стажу(exper) равен 0,02, что является положительным числом, то есть с увеличением стажа вероятность иметь зарплату выше средней повышается

```
Call:
glm(formula = wage_above ~ exper + I(exper^2) + educ + urban,
    family = binomial, data = dt)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.084476   1.017219  -8.931  < 2e-16 ***
exper        0.388803   0.116564   3.336 0.000851 ***
I(exper^2)  -0.011901   0.005363  -2.219 0.026471 *
educ         0.399521   0.040210   9.936  < 2e-16 ***
urban        0.818848   0.181212   4.519 6.22e-06 ***
```

*Рассмотрим модель **усеченной регрессии:***

Усечённая регрессия - модель регрессии в условиях, когда выборка осуществляется только из тех наблюдений, которые удовлетворяют ограничениям. В нашем случае, я поставила ограничение снизу равным 0, т.к зарплата выражена положительным числом.

```
Coefficients (location model):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -59.63255    6.67886  -8.929  < 2e-16 ***
exper         2.77351    0.72420   3.830 0.000128 ***
educ          3.30088    0.25833  12.778  < 2e-16 ***
urban         5.94877    1.20599   4.933 8.11e-07 ***
I(exper^2)   -0.07245    0.03336  -2.172 0.029886 *
```

С помощью модели усеченной регрессии найдем матожидание заработной платы для произвольного индивида:

$E(wage) = a_0 + a_1*educ + a_2*exper + a_3*exper^2 + a_4*urban + \sigma*\lambda$,

$\lambda = f(lin/\sigma) / F(lin/\sigma)$

$lin = a_0 + a_1*educ + a_2*exper + a_3*exper^2 + a_4*urban$
f - плотность стандартного нормального распределения
F - функция распределения стандартного нормального распределения

```
Coefficients (scale model with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.37433    0.03613   65.72   <2e-16 ***
```

sigma - это экспонента от этого коэффициента

Сравним все 3 модели по информационному ***критерию Акаике***

```
linear    8693.648
poly      8691.543
truncated 8376.543
```

Видим, что усеченная регрессия лучше работает, поэтому для предсказаний заработной платы, лучше использовать формулу для матожидания, данную ранее

В работе я рассмотрела данные с помощью линейной регрессии, полиномиальной регрессии, логистической регрессии, а также усеченной регрессии и расписала интерпретацию для каждого случая. Также благодаря критерию Акаике было установлено, что наиболее точные результаты дает усеченная регрессия. Отвечая на поставленный вопрос, число лет обучения вносит больший вклад в заработную плату, нежели опыт работы. А у стажа есть максимальная точка. Это 26 лет, после достижения данного опыта работы, прирост зарплаты носит отрицательный характер.