```
## Warning: One bear with balalaika and vodka has joined your R session!
```

1. Рассмотрим следующую регрессионную модель зависимости логарифма заработной платы $ln(W)$ от уровня образования Edu, опыта работы Exp, Exp^2 и уровня образования родителей $Fedu$, $Medu$:

$$\widehat{ln(W)} = \hat{\beta}_1 + \hat{\beta}_2 Edu + \hat{\beta}_3 Exp + \hat{\beta}_4 Exp^2 + \hat{\beta}_5 Fedu + \hat{\beta}_6 Medu$$

Модель регрессии была отдельно оценена по выборкам из 35 мужчин и 23 женщин, и были получены остаточные суммы квадратов $RSS_1 = 34.4$ и $RSS_2 = 23.4$ соответственно. Остаточная сумма квадратов в регрессии, оценённой по объединённой выборке, равна 70.3. На уровне значимости 5% проверьте гипотезу об отсутствии дискриминации в оплате труда между мужчинами и женщинами

Решение

Упорядочим нашу выборку таким образом, чтобы наблюдения с номерами с 1 по 35 относились к мужчинам, а наблюдения с номерами с 36 по 58 относились к женщинам. Тогда уравнение

$$ln(W_i) = \beta_1 + \beta_2 Edu_i + \beta_3 Exp_i + \beta_4 Exp_i^2 + \beta_5 Fedu_i + \beta_6 Medu_i + \varepsilon_i, i = 1, ..., 35$$

соответствует регрессии, построенной для подвыборки из мужчин, а уравнение

$$ln(W_i) = \gamma_1 + \gamma_2 Edu_i + \gamma_3 Exp_i + \gamma_4 Exp_i^2 + \gamma_5 Fedu_i + \gamma_6 Medu_i + \varepsilon_i, i = 36, ..., 58$$

соответствует регрессии, построенной для подвыборки из женщин. Введем следующие переменные:

$$d_i = \begin{cases} 1, & \text{если } i\text{–ое наблюдение соответствует мужчине,} \\ 0, & \text{в противном случае;} \end{cases}$$

$$dum_i = \begin{cases} 1, & \text{если } i\text{–ое наблюдение соответствует женщине,} \\ 0, & \text{в противном случае.} \end{cases}$$

Рассмотрим следующее уравнение регрессии:

$$ln(W_i) = \beta_1 d_i + \gamma_1 dum_i + \beta_2 Edu_i d_i + \gamma_2 Edu_i dum_i + \beta_3 Exp_i d_i + \gamma_3 Exp_i dum_i + \beta_4 Exp_i^2 d_i +$$

$$+\gamma_4 Exp_i^2 dum_i + \beta_5 Fedu_i d_i + \gamma_5 Fedu_i dum_i + \beta_6 Medu_i d_i + \gamma_6 Medu_i dum_i + \varepsilon_i, i = 1, ..., 58$$

Гипотеза, которую требуется проверить в данной задаче, имеет вид

$$H_0 : \begin{cases} \beta_1 = \gamma_1, \\ \beta_2 = \gamma_2, \quad H_1 : |\beta_1 - \gamma_1| + |\beta_2 - \gamma_2| + \cdots + |\beta_6 - \gamma_6| > 0. \\ \dots \\ \beta_6 = \gamma_6 \end{cases}$$

Тогда регрессия

$$ln(W_i) = \beta_1 d_i + \gamma_1 dum_i + \beta_2 Edu_i d_i + \gamma_2 Edu_i dum_i + \beta_3 Exp_i d_i + \gamma_3 Exp_i dum_i + \beta_4 Exp_i^2 d_i +$$

$$+\gamma_4 Exp_i^2 dum_i + \beta_5 Fedu_i d_i + \gamma_5 Fedu_i dum_i + \beta_6 Medu_i d_i + \gamma_6 Medu_i dum_i + \varepsilon_i, i = 1, ..., 58$$

по отношению к основной гипотезе H_0 является регрессией без ограничений, а регрессия

$$ln(W_i) = \beta_1 + \beta_2 Edu_i + \beta_3 Exp_i + \beta_4 Exp_i^2 + \beta_5 Fedu_i + \beta_6 Medu_i + \varepsilon_i, i = 1, ..., 58$$

является регрессией с ограничениями.

Кроме того, для решения задачи должен быть известен следующий факт:

$RSS_{UR} = RSS_1 + RSS_2$, где RSS_{UR} — это сумма квадратов остатков в модели:

$$ln(W_i) = \beta_1 d_i + \gamma_1 dum_i + \beta_2 Edu_i d_i + \gamma_2 Edu_i dum_i + \beta_3 Exp_i d_i + \gamma_3 Exp_i dum_i + \beta_4 Exp_i^2 d_i +$$

$$+\gamma_4 Exp_i^2 dum_i + \beta_5 Fedu_i d_i + \gamma_5 Fedu_i dum_i + \beta_6 Medu_i d_i + \gamma_6 Medu_i dum_i + \varepsilon_i, i = 1, ..., 58$$

RSS_1 — это сумма квадратов остатков в модели:

$$ln(W_i) = \beta_1 + \beta_2 Edu_i + \beta_3 Exp_i + \beta_4 Exp_i^2 + \beta_5 Fedu_i + \beta_6 Medu_i + \varepsilon_i, i = 1, ..., 35$$

RSS_2 — это сумма квадратов остатков в модели:

$$ln(W_i) = \gamma_1 + \gamma_2 Edu_i + \gamma_3 Exp_i + \gamma_4 Exp_i^2 + \gamma_5 Fedu_i + \gamma_6 Medu_i + \varepsilon_i, i = 36, ..., 58$$

(a) Тестовая статистика:

$$T = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-m)},$$

где RSS_R – сумма квадратов остатков в модели с ограничениями;

RSS_{UR} – сумма квадратов остатков в модели без ограничений;

q – число линейно независимых уравнений в основной гипотезе H_0;

n – общее число наблюдений;

m – число коэффициентов в модели без ограничений

(b) Распределение тестовой статистики:

$$T \sim F(q, n-m)$$

(c) Наблюдаемое значение тестовой статистики:

$$T_{obs} = \frac{(70.3 - (34.4 + 23.4))/6}{(34.4 + 23.4)/(58 - 12)} = 1.66$$

(d) Область, в которой H_0 не отвергается:

$$[0; T_{cr}] = [0; 2.3]$$

(e) Статистический вывод:

Поскольку $T_{obs} \in [0; T_{cr}]$, то на основе имеющихся данных мы не можем отвергнуть гипотезу H_0 в пользу альтернативной H_1. Следовательно, имеющиеся данные не противоречат гипотезе об отсутствии дискриминации на рынке труда между мужчинами и женщинами

2. Рассмотрим следующую регрессионную модель зависимости логарифма заработной платы $ln(W)$ от уровня образования Edu, опыта работы Exp, Exp^2:

$$\widehat{ln(W)} = \hat{\beta}_1 + \hat{\beta}_2 Edu + \hat{\beta}_3 Exp + \hat{\beta}_4 Exp^2$$

Модель регрессии была отдельно оценена по выборкам из 20 мужчин и 20 женщин, и были получены остаточные суммы квадратов $RSS_1 = 49.4$ и $RSS_2 = 44.1$ соответственно. Остаточная сумма квадратов в регрессии, оценённой по объединённой выборке, равна 105.5. На уровне 5% проверьте гипотезу об отсутствии дискриминации в оплате труда между мужчинами и женщинами

3. Ниже приведены результаты оценивания спроса на молоко для модели $y_i = \beta_1 + \beta_2 I_i + \beta_3 P_i + \varepsilon_i$, где y_i – стоимость молока, купленного i–ой семьёй за последние 7 дней (в руб.), I_i – месячный доход i–ой семьи (в руб.), P_i – цена 1 литра молока (в руб.). Вычисления для общей выборки, состоящей из 2127 семей, дали $RSS = 8841601$. Для двух подвыборок, состоящих из 348 городских и 1779 сельских семей, соответствующие суммы квадратов остатков оказались следующими: $RSS_1 = 1720236$ и $RSS_2 = 7099423$. Можно ли считать зависимость спроса на молоко от его цены и дохода единой для городской и сельской местности? Ответ обоснуйте подходящим тестом

4. По 52 наблюдениям была оценена следующая зависимость цены квадратного метра квартиры $Price$ (в долларах) от площади кухни K (в квадратных метрах), времени в пути пешком до ближайшего метро M (в минутах), расстояния до центра города C (в км) и наличия рядом с домом лесопарковой зоны P (1 — есть, 0 — нет)

$$\underset{(s.e.)}{\widehat{Price}} = \underset{(3.73)}{16.12} + \underset{(0.14)}{1.7}\, K - \underset{(0.03)}{0.35} M - \underset{(0.12)}{0.46} C + \underset{(0.98)}{2.22} P$$

$$R^2 = 0.78, \sum_{i=1}^{52} (Price_i - \overline{Price})^2 = 278$$

Предположим, что все квартиры в выборке можно отнести к двум категориям: квартиры на севере города (28 наблюдений) и квартиры на юге города (24 наблюдения). Модель регрессии была оценена отдельно только по квартирам на севере и только по квартирам на юге. Ниже приведены результаты оценивания.

Для квартир на севере:

$$\underset{(s.e.)}{\widehat{Price}} = \underset{(3.3)}{14} + \underset{(0.23)}{1.6}\, K - \underset{(0.04)}{0.33} M - \underset{(0.22)}{0.4}\, C + \underset{(0.78)}{2.1}\, P, RSS = 21.8$$

Для квартир на юге:

$$\underset{(s.e.)}{\widehat{Price}} = \underset{(3.9)}{16.8} + \underset{(0.4)}{1.62} K - \underset{(0.12)}{0.29} M - \underset{(0.23)}{0.51} C + \underset{(1.28)}{1.98} P, RSS = 19.2$$

На уровне значимости 5% проверьте гипотезу о различии в ценообразовании квартир на севере и на юге

5. По 52 наблюдениям была оценена следующая зависимость цены квадратного метра квартиры $Price$ (в долларах) от площади кухни K (в квадратных метрах), времени в пути пешком до ближайшего метро M (в минутах), расстояния до центра города C (в км) и наличия рядом с домом лесопарковой зоны P (1 — есть, 0 — нет)

$$\underset{(s.e.)}{\widehat{Price}} = \underset{(3.73)}{16.12} + \underset{(0.14)}{1.7}\,K - \underset{(0.03)}{0.35}M - \underset{(0.12)}{0.46}C + \underset{(0.98)}{2.22}P$$

$$R^2 = 0.78, \sum_{i=1}^{52}(Price_i - \overline{Price})^2 = 278$$

Предположим, что все квартиры в выборке можно отнести к двум категориям: квартиры на севере города (28 наблюдений) и квартиры на юге города (24 наблюдения). Пусть S — это фиктивная переменная, равная 1 для домов в южной части города и 0 для домов в северной части города. Используя эту переменную, была оценена следующая регрессия:

$$\underset{(s.e.)}{\widehat{Price}} = \underset{(3.13)}{14.12} + \underset{(0.11)}{0.25}S + \underset{(0.13)}{1.65}K + \underset{(0.14)}{0.17}K\cdot S - \underset{(0.039)}{0.37}\,M + \underset{(0.0012)}{0.05}\,M\cdot S - \underset{(0.13)}{0.44}C - \underset{(0.18)}{0.06}C\cdot S + \underset{(0.88)}{2.27}P - \underset{(0.08)}{0.23}P\cdot S$$

$$R^2 = 0.85$$

На уровне значимости 5% проверьте гипотезу о различии в ценообразовании квартир на севере и на юге

6. На основе квартальных данных с 2003 по 2008 год было получено следующее уравнение регрессии, описывающее зависимость цены на товар P от нескольких факторов:

$$P = 3.5 + 0.4X + 1.1W, ESS = 70.4, RSS = 40.5$$

Когда в уравнение были добавлены фиктивные переменные, соответствующие первым трем кварталам года Q_1, Q_2, Q_3, оцениваемая модель приобрела вид:

$$P_t = \beta + \beta_X X_t + \beta_W W_t + \beta_{Q_{1t}} Q_{1t} + \beta_{Q_{2t}} Q_{2t} + \beta_{Q_{3t}} Q_{3t} + \varepsilon_t$$

При этом величина ESS выросла до 86.4. Сформулируйте и на уровне значимости 5% проверьте гипотезу о наличии сезонности

7. Рассмотрим следующую функцию спроса с сезонными переменными $SPRING$ (весна),

$SUMMER$ (лето), $FALL$ (осень):

$$\widehat{ln(Q)} = \hat{\beta}_1 + \hat{\beta}_2 \cdot ln(P) + \hat{\beta}_3 \cdot SPRING + \hat{\beta}_4 \cdot SUMMER + \hat{\beta}_5 \cdot FALL$$

$$R^2 = 0.37, n = 20$$

Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \beta_3 = \beta_5$. Дайте интерпретацию проверяемой гипотезе. Пусть для регрессии с ограничениями был вычислен коэффициент $R^2_R = 0.23$. На уровне значимости 5% проверьте нулевую гипотезу

8. Рассмотрим следующую функцию спроса с сезонными переменными $SPRING$ (весна), $SUMMER$ (лето), $FALL$ (осень):

$$\widehat{ln(Q)} = \hat{\beta}_1 + \hat{\beta}_2 \cdot ln(P) + \hat{\beta}_3 \cdot SPRING + \hat{\beta}_4 \cdot SUMMER + \hat{\beta}_5 \cdot FALL$$

$$R^2 = 0.24, n = 24$$

Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \begin{cases} \beta_3 = 0, \\ \beta_4 = \beta_5 \end{cases}$. Дайте интерпретацию проверяемой гипотезе. Пусть для регрессии с ограничениями был вычислен коэффициент $R^2_R = 0.13$. На уровне значимости 5% проверьте нулевую гипотезу

9. Исследователь собирается по выборке, содержащей данные за 2 года, построить модель линейной регрессии с константой и 3-мя объясняющими переменными. В модель предполагается ввести 3 фиктивные сезонные переменные $SPRING$ (весна), $SUMMER$ (лето) и $FALL$ (осень) на все коэффициенты регрессии. Однако в процессе оценивания статистический пакет вывел на экран компьютера следующее сообщение “insufficient number of observations”. Объясните, почему имеющегося числа наблюдений не хватило для оценивания параметров модели

10. По данным для 57 индивидов оценили зависимость длительности обучения индивида S от способностей индивида, описываемых обобщённой переменной IQ, и пола индивида, описываемого с помощью фиктивной переменной $MALE$ (равной 1 для мужчин и 0 для женщин), с помощью двух регрессий (в скобках под коэффициентами указаны оценки

стандартных отклонений):

$$\underset{(s.e.)}{\hat{S}} = \underset{(0.44)}{6.12} + \underset{(0.088)}{0.147} \cdot IQ, RSS = 2758.6$$

$$\underset{(s.e.)}{\hat{S}} = \underset{(0.73)}{6.12} + \underset{(0.014)}{0.147} \cdot IQ - \underset{(0.933)}{1.035} \cdot MALE + \underset{(0.018)}{0.0166} \cdot (MALE \cdot IQ), RSS = 2090.98$$

Зависит ли длительность обучения от пола индивида и почему?

11. По данным, содержащим 30 наблюдений, построена регрессия

$$\hat{y} = 1.3870 + 5.2587 \cdot x + 2.6259 \cdot d + 2.5955 \cdot x \cdot d,$$

где фиктивная переменная d определяется следующим образом:

$$d_i = \begin{cases} 1 & \text{при } i \in \{1, \ldots, 20\}, \\ 0 & \text{при } i \in \{21, \ldots, 30\}. \end{cases}$$

Найдите оценки коэффициентов в модели $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, построенной по первым 20-ти наблюдениям, т.е. при $i \in \{1, \ldots, 20\}$

12. Выборка содержит 30 наблюдений зависимой переменной y и независимой переменной x. Ниже приведены результаты оценивания уравнения регрессии $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ по первым 20-ти и последним 10-ти наблюдениям соответственно:

$$\hat{y} = 4.0039 + 2.6632 \cdot x$$

$$\hat{y} = 1.3780 + 5.2587 \cdot x$$

По имеющимся данным найдите оценки коэффициентов модели, рассчитанной по 30-ти наблюдениям $y_i = \beta_1 + \beta_2 x_i + \Delta\beta_1 \cdot d_i + \Delta\beta_2 \cdot x_i \cdot d_i + \varepsilon_i$, где фиктивная переменная d определяется следующим образом:

$$d_i = \begin{cases} 1 & \text{при } i \in \{1, \ldots, 20\}, \\ 0 & \text{при } i \in \{21, \ldots, 30\}. \end{cases}$$

13. Пусть регрессионная модель имеет вид $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \varepsilon_i, i = 1, \ldots, n$. Тестируемая гипотеза $H_0 : \beta_2 = \beta_3 = \beta_4$. Запишите, какой вид имеет модель «с ограничением» для тестирования указанной гипотезы

14. Пусть регрессионная модель имеет вид $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \varepsilon_i, i = 1, \ldots, n$. Тестируемая гипотеза $H_0 : \beta_3 = \beta_4 = 1$. Какая модель из приведённых ниже может выступать в качестве модели «с ограничением» для тестирования указанной гипотезы? Если ни одна из них, то запишите свою

    (a) $y_i - (x_{i2} + x_{i3}) = \beta_1 + \beta_2 x_{i1} + \varepsilon_i$

    (b) $y_i + (x_{i2} - x_{i3}) = \beta_1 + \beta_2 x_{i1} + \varepsilon_i$

    (c) $y_i + x_{i2} + x_{i3} = \beta_1 + \beta_2 x_{i1} + \varepsilon_i$

    (d) $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 + \beta_4 + \varepsilon_i$

15. Пусть регрессионная модель имеет вид $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \varepsilon_i, i = 1, \ldots, n$. Тестируемая гипотеза $H_0 : \begin{cases} \beta_2 + \beta_3 + \beta_4 = 1, \\ \beta_3 + \beta_4 = 0. \end{cases}$ Какая модель из приведённых ниже может выступать в качестве модели «с ограничением» для тестирования указанной гипотезы? Если ни одна из них, то запишите свою

    (a) $y_i - x_{i1} = \beta_1 + \beta_3(x_{i2} - x_{i3}) + \varepsilon_i$

    (b) $y_i - x_{i1} = \beta_1 + \beta_4(x_{i3} - x_{i2}) + \varepsilon_i$

    (c) $y_i + x_{i1} = \beta_1 + \beta_3(x_{i2} + x_{i3}) + \varepsilon_i$

    (d) $y_i + x_{i1} = \beta_1 + \beta_3(x_{i2} - x_{i3}) + \varepsilon_i$

16. Пусть регрессионная модель имеет вид $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \varepsilon_i, i = 1, \ldots, n$. Тестируемая гипотеза $H_0 : \begin{cases} \beta_2 - \beta_3 = 0, \\ \beta_3 + \beta_4 = 0. \end{cases}$ Какая модель из приведённых ниже может выступать в качестве модели «с ограничением» для тестирования указанной гипотезы? Если ни одна из них, то запишите свою

    (a) $y_i = \beta_1 + \beta_3(x_{i2} - x_{i1} - x_{i3}) + \varepsilon_i$

    (b) $y_i - x_{i1} = \beta_1 + \beta_4(x_{i3} - x_{i2}) + \varepsilon_i$

    (c) $y_i = \beta_1 + \beta_3(x_{i1} + x_{i2} + x_{i3}) + \varepsilon_i$

    (d) $y_i = \beta_1 + \beta_3(x_{i1} + x_{i2} - x_{i3}) + \varepsilon_i$

17. Известно, что P-значение для коэффициента регрессии равно 0.087, а уровень значимости 0.1. Является ли значимым данный коэффициент в регрессии?

18. Известно, что P-значение для коэффициента регрессии равно 0.078, а уровень значимости 0.05. Является ли значимым данный коэффициент в регрессии?

19. Известно, что P-значение для коэффициента регрессии равно 0.09. На каком уровне значимости данный коэффициент в регрессии будет признан значимым?

20. Ниже приведены результаты оценивания уравнения линейной регрессии зависимости количества смертей в автомобильных катастрофах от различных характеристик:

$$deaths_i = \beta_1 + \beta_2 drivers_i + \beta_3 popden_i + \beta_4 temp + \beta_5 fuel + \varepsilon_i$$

$$\widehat{deaths}_i = \underset{(222.8803)}{-27.1} + \underset{(0.3767)}{4.64} \cdot drivers_i - \underset{(0.0239)}{0.0228} \cdot popden_i + \underset{(4.6016)}{5.3} \cdot temp_i - \underset{(0.8679)}{0.663} \cdot fuel_i$$

	Estimate	St.Error	t value	P-value
Intercept	-27.10	222.88	-0.12	0.90
Drivers	4.64	0.38	12.30	0.00
Popden	-0.02	0.02	-0.95	0.35
Temp	5.30	4.60	1.15	0.26
Fuel	-0.66	0.87	-0.76	0.45

Перечислите, какие из переменных в регрессии являются значимыми и на каком уровне значимости