Семинар 8.
Фиктивные переменные. Тест Чоу.

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

$$\widehat{\ln(W_i)} = \hat{\beta}_1 + \hat{\beta}_2 Edu_i + \hat{\beta}_3 Exp_i + \hat{\beta}_4 Exp_i^2 + \hat{\beta}_5 Fedu_i + \hat{\beta}_6 Medu_i.$$

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, оцененной по объединенной выборке, равна 70.3. Протестируйте на 5% уровне значимости гипотезу об отсутствии дискриминации в оплате труда между мужчинами и женщинами.

Решение:

Запишем гипотезу H_0:

$$\beta_1^M = \beta_1^F,$$

$$\dots$$

$$\beta_6^M = \beta_6^F.$$

Число параметров $k = 6$, число ограничений $q = k = 6$.

Проверим гипотезу с помощью теста Чоу:

$$F = \frac{[RSS_{pooled} - (RSS_M + RSS_F)]/(k)}{(RSS_M + RSS_F)/(n_M + n_F - 2k)}$$

$$F = \frac{(70.3 - 57.8)/6}{57.8/46} = \frac{12.5/6}{57.8/46} = \frac{2.08333}{1.25652} \approx 1.658$$

Критическое значение $F_{0.05}(6, 46) \approx 2.30$.

Так как $1.658 < 2.30$, гипотеза об отсутствии структурного сдвига (одинаковости уравнений для мужчин и женщин) не отвергается на уровне значимости 5%.

2. (Универсиада по эконометрике, МГУ, 2018 год, отборочный этап). В некоторой отрасли заработная плата работника следующим образом зависит от его опыта работы и пола:

$$\ln W_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i D_i + \varepsilon_i,$$

где W_i — заработная плата i-го работника в рублях в месяц, X_i — стаж i-го работника в годах, D_i — бинарная переменная, равная единице для женщин и нулю для мужчин, ε_i — случайные ошибки. Предполагается, что все предпосылки классической линейной модели множественной регрессии выполнены. Оценка параметров модели при помощи МНК на основе данных о десяти тысячах наблюдений позволила получить следующие результаты:

$$\widehat{\ln W_i} = 10 + 60X_i - 3X_i^2 - 12X_iD_i.$$

Оценка ковариационной матрицы вектора оценок коэффициентов имеет вид:

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \begin{pmatrix} 51 & -20 & 2 & 0 \\ -20 & 9 & -1 & 0 \\ 2 & -1 & 0.1 & 0 \\ 0 & 0 & 0 & 0.4 \end{pmatrix}.$$

(a) Опираясь на полученные оценки параметров, изобразите на одном рисунке графики логарифмов заработной платы типичного работника и типичной работницы в зависимости от их стажа. Интерпретируйте полученный результат.

(b) Аналитик Афанасий предполагает, что женщины в данной отрасли достигают максимума своей производительности при стаже, равном 10 годам. Соответственно и их заработная плата максимальна именно в этот момент. Сформулируйте (в терминах коэффициентов модели) гипотезу, которая соответствует предположению Афанасия, и проверьте её при уровне значимости 5%.

(c) Аналитик Евгения утверждает, следующее: «Если обозначить X_M^* — стаж работы, при котором зарплата мужчины максимальна, а X_W^* — стаж, при котором зарплата женщины максимальна, то окажется, что $X_M^* = X_W^* + 1$. То есть женщины достигают пика своей зарплаты на год раньше мужчин». Сформулируйте (в терминах коэффициентов модели) гипотезу, которая соответствует утверждению Евгении, и проверьте её на уровне значимости 5%.

Решение:

(a) Модель:

$$\ln W_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i D_i + \varepsilon_i$$

Оценённое уравнение:

$$\widehat{\ln W_i} = 10 + 60X_i - 3X_i^2 - 12X_iD_i$$

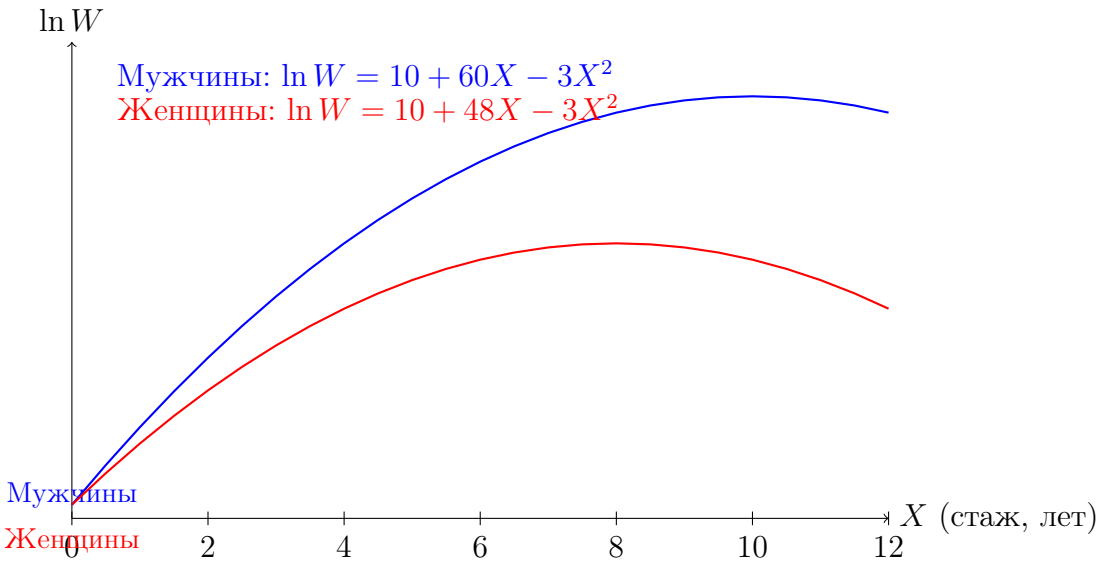
– Для мужчин ($D_i = 0$):

$$\widehat{\ln W_M}(X) = 10 + 60X - 3X^2$$

– Для женщин ($D_i = 1$):

$$\widehat{\ln W_W}(X) = 10 + 60X - 3X^2 - 12X = 10 + 48X - 3X^2$$

Оба графика – параболы с ветвями вниз ($\beta_3 = -3 < 0$). У мужчин коэффициент при X больше (60 против 48), значит, начальный рост зарплаты у мужчин круче. Из-за $-12XD_i$ зарплата женщин ниже при любом стаже $X > 0$, разрыв растёт с ростом X до некоторого момента.



(b) Для женщин ($D = 1$):

$$\ln W = 10 + 48X - 3X^2$$

Производная по X:

$$\frac{\partial \ln W}{\partial X} = 48 - 6X$$

Приравниваем к нулю: $48 - 6X = 0 \Rightarrow X_W^* = 8$.

Афанасий предполагает $X_W^* = 10$.

В общем виде:

$$\ln W = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 XD$$

Для женщин: $\beta_2 + \beta_4$ при X, β_3 при X^2.

Производная: $(\beta_2 + \beta_4) + 2\beta_3 X = 0 \Rightarrow X^*_W = -\frac{\beta_2+\beta_4}{2\beta_3}$.

Гипотеза Афанасия:

$$H_0 : -\frac{\beta_2 + \beta_4}{2\beta_3} = 10$$

Или:

$$H_0 : \beta_2 + \beta_4 + 20\beta_3 = 0$$

Подставляем оценки:

$$\hat{\beta}_2 = 60, \quad \hat{\beta}_3 = -3, \quad \hat{\beta}_4 = -12$$

$$\hat{\beta}_2 + \hat{\beta}_4 + 20\hat{\beta}_3 = 60 - 12 + 20 \cdot (-3) = 48 - 60 = -12$$

Найдём оценку дисперсии оценки для $\beta_2 + \beta_4 + 20\beta_3$:

$$\widehat{Var}(\hat{\beta}_2+\hat{\beta}_4+20\hat{\beta}_3) = \widehat{Var}(\hat{\beta}_2)+\widehat{Var}(\hat{\beta}_4)+400\,\widehat{Var}(\hat{\beta}_3)+2\widehat{Cov}(\hat{\beta}_2,\hat{\beta}_4)+40\widehat{Cov}(\hat{\beta}_2,\hat{\beta}_3)+40\widehat{C}$$

Из оценки ковариационной матрицы:

$$\widehat{Var}(\hat{\beta}_2) = 9, \quad \widehat{Var}(\hat{\beta}_3) = 0.1, \quad \widehat{Var}(\hat{\beta}_4) = 0.4$$

$$\widehat{Cov}(\hat{\beta}_2, \hat{\beta}_4) = 0, \quad \widehat{Cov}(\hat{\beta}_2, \hat{\beta}_3) = -1, \quad \widehat{Cov}(\hat{\beta}_4, \hat{\beta}_3) = 0$$

Подставляем:

$$\widehat{Var}(\hat{h}) = 9 + 0.4 + 400 \cdot 0.1 + 2 \cdot 0 + 40 \cdot (-1) + 40 \cdot 0$$

$$= 9 + 0.4 + 40 - 40 = 9.4$$

Наблюдаемое значение статистики:

$$t = \frac{-12}{3.065} \approx -3.915$$

Критическое значение $t_{0.025}(\infty) \approx 1.96$.

$|t| > 1.96 \Rightarrow$ отвергаем H_0 на уровне 5%.

Вывод: предположение Афанасия не согласуется с данными.

(c) Для мужчин ($D = 0$):

$$\frac{\partial \ln W}{\partial X} = \beta_2 + 2\beta_3 X = 0 \Rightarrow X^*_M = -\frac{\beta_2}{2\beta_3}$$

Для женщин:

$$X_W^* = -\frac{\beta_2 + \beta_4}{2\beta_3}$$

Евгения утверждает: $X_M^* = X_W^* + 1$.

Подставляем:

$$-\frac{\beta_2}{2\beta_3} = -\frac{\beta_2 + \beta_4}{2\beta_3} + 1$$

Умножаем на $2\beta_3$:

$$-\beta_2 = -(\beta_2 + \beta_4) + 2\beta_3$$

$$-\beta_2 = -\beta_2 - \beta_4 + 2\beta_3$$

$$0 = -\beta_4 + 2\beta_3$$

$$H_0 : 2\beta_3 - \beta_4 = 0$$

Оценка:

$$2\hat{\beta}_3 - \hat{\beta}_4 = 2 \cdot (-3) - (-12) = -6 + 12 = 6$$

Дисперсия:

$$\widehat{Var}(2\hat{\beta}_3 - \hat{\beta}_4) = 4\widehat{Var}(\hat{\beta}_3) + \widehat{Var}(\hat{\beta}_4) - 4\widehat{Cov}(\hat{\beta}_3, \hat{\beta}_4)$$

$$= 4 \cdot 0.1 + 0.4 - 4 \cdot 0 = 0.4 + 0.4 = 0.8$$

Наблюдаемое значение статистики:

$$t = \frac{6}{0.894} \approx 6.708$$

$|t| > 1.96 \Rightarrow$ отвергаем H_0 на 5% уровне значимости.

Вывод: утверждение Евгении не согласуется с данными.

3. (компьютерный практикум). В файле $Chow\_2.xls$ содержатся данные об экономике Баккардии в период с 1 квартала 2015 года по 4 квартал 2022 года. Показатели выражены в миллиардах баккардийских крон 2015 года.

   a) Оцените следующую модель регрессии:

   $$C_t = \beta_1 + \beta_2 Y_t + \beta_3 D2_t + \beta_4 D3_t + \beta_5 D4_t + \varepsilon_t,$$

   где C_t — конечное потребление в момент времени t, Y_t — доход в момент времени t, Dj_t — дамми переменная на квартал ($j = 2, 3, 4$).
   Проинтерпретируйте полученные результаты.

   b) На уровне значимости 5% проверьте гипотезу о наличии сезонности. Сформулируйте нулевую и альтернативную гипотезы.

c) Оцените модель в следующем виде:

$$C_t = \beta_1 Y_t + \beta_2 D1_t + \beta_3 D2_t + \beta_4 D3_t + \beta_5 D4_t + \varepsilon_t.$$

Сравните полученные результаты с предыдущим пунктом.

d) Попробуйте улучшить модель, включив в нее переменные взаимодействия:

$$C_t = \beta_1 + \beta_2 Y_t + \beta_3 D2_t + \beta_4 D3_t + \beta_5 D4_t + \beta_6 (Y * D2) + \beta_7 (Y * D3) + \beta_8 (Y * D4) + \varepsilon_t.$$

Проинтерпретируйте полученные результаты.