Отчёт выполнил Царюк Артём 21МО

Блоки II-IV

Конфирматорный факторный анализ

# Задача 1.

**Анализ близнецовых данных**

#### Задание 1.1.

**Провести конфирматорный факторный анализ близнецовых данных, используя электронную таблицу *Excel.***

***Метод анализа и модель***

В качестве наблюдаемых переменных ***Pi*** использовать результаты физиологических исследований(*вариант 1*) или результаты тестирования темперамента (*вариант 2*), а в качестве латентных переменных ***Ai***, ***Di***, ***Ci*** и ***Ei*** – величины, обусловленные аддитивными генетическими эффектами (различия между гомозиготами), доминированием (гетерозиготные эффекты), а также влиянием общей и индивидуальной сред. Связь этих переменных для каждой пары близнецов описать одномерной генетической моделью:

***A1***

***D1***

***C1***

***E1***

***A2***

***D2***

***C2***

***E2***

***P1***

***P2***

***a d c e 0 0 0 0***

***0 0 0 0 a d c e***

=

Физиологические характеристикии характеристики темперамента рассматриваются как полигенные признаки, для которых выполняются принципы менделевского наследования.

Полагая ***Ai***, ***Di***, ***Ci*** и ***Ei*** (***i=1,2***) независимыми случайными величинами с единичной дисперсией, фенотипические дисперсии ***Pi*** можно представить в виде суммы следующих компонентов:

***Var(Pi)*= *a2 + d2 + c2 + e2 .***

Величина ***a2***определяет вклад аддитивных генетических эффектов, ***d2*** – эффектов доминирования, ***c2***и***e2***– влияние соответственно общей и индивидуальной сред. Аддитивные генетические корреляции для моно- и дизиготных близнецов равны соответственно 1 и 0,5; доминантные – 1 и 0,25; корреляции факторов общей среды равны 1 для обоих типов близнецов (т. е. предполагается, что они воспитывались в одной и той же семье); факторы индивидуальной среды для разных близнецов не коррелируют по определению.

Данную модель удобно представить графически в виде следующей путевой диаграммы:

## 1

*1 (МЗ) или 0,5 (ДЗ)*

*1 (МЗ) или 0,25 (ДЗ)*

*1 (МЗ и ДЗ)*

***a***

***a***

***d***

***d***

***c***

***c***

***e***

***e***

## Близнец 1

## Близнец 2

## 1

## 1

## 1

## 1

## 1

## 1

## 1

***P1***

***P2***

Коэффициенты ***a***, ***d***, ***c*** и ***e*** рассматриваются как свободные параметры модели. Найти их оценки методом максимального правдоподобия, использовав в качестве минимизируемого критерия функцию

***F = [ln |Σ| – ln |S| + tr(SΣ-1) – p] (N-1) ,***

где ***S*** – выборочная ковариационная матрица наблюдаемых переменных, ***Σ*** –прогнозируемая ковариационная матрица наблюдаемых переменных, ***|Σ|*** и ***|S|*** - определители матриц ***Σ*** и ***S***, ***tr(SΣ-1)*** – след матрицы ***(SΣ-1)***, ***N*** – объем выборки, использованной для вычисления матрицы ***S***, ***p*** – число наблюдаемых переменных.

В случае многомерного нормального распределения наблюдаемых переменных значения критерия ***F*** описываются распределением *хи-квадрат* (χ2). Для проверки того, насколько хорошо модель согласуется с результатами измерений, использовать критерии χ2 и *AIC* (*Akaike’s Information Criterion*). Второй критерий вычисляется как

***χ2 – 2df***

и отдает предпочтение тем моделям, которые согласуются с результатами измерений при меньшем количестве параметров (модель с наименьшим значением этого критерия рассматривается как наилучшая.) Разумеется, то, что считается наилучшим в смысле критерия χ2,не обязательно является наилучшим с точки зрения критерия *AIC*.

**Реализовать процедуру нахождения указанных оценок в виде макроса для электронной таблицы Excel.**

***Рекомендации***

Матрицу ***Σ*** выразить через свободные параметрыследующим образом:

,

***a2 + d2 + c2 + e2 xa2 + yd2 + c2***

***xa2 + yd2 + c2 a2 + d2 + c2 + e2***

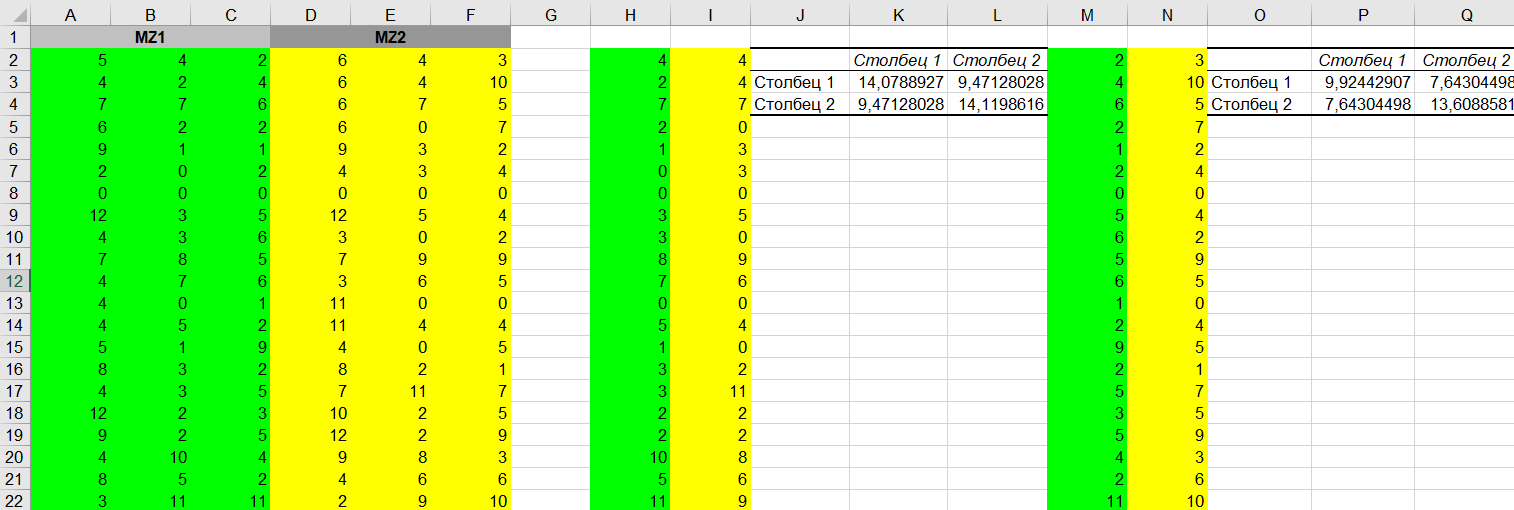
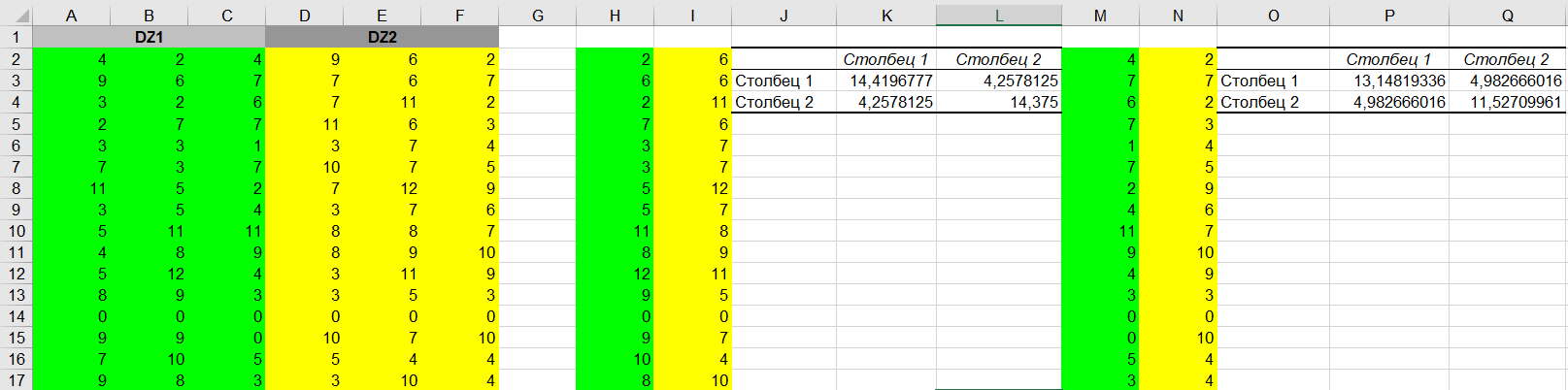
где для монозиготных близнецов ***x=y=1***, а для дизиготных - ***x=0,5***, ***y=0,25.***

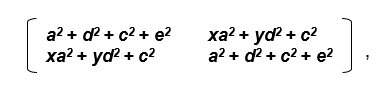
Для определения оценок свободных параметров необходимо, чтобы их число не превышало числа независимых статистик, определенных по результатам наблюдений. В нашем случае число этих статистик равно ***p(p+1)/2*** – столько различных между собой элементов имеется в симметричной матрице ***S***. Разность между числом независимых статистик, определенных по результатам наблюдений, и числом свободных параметров есть число степеней свободы задачи(***df***). Если ***df = 0***, то найти решение можно, но нельзя судить о степени согласованности модели с результатами наблюдений.

Чтобы получить достаточное количество статистик, искомые характеристики оценить одновременно для монозиготных и дизиготных пар близнецов (каждому типу близнецов соответствует своя матрица ***Σ***). Минимизируемая функция при этом составляется из соответствующих функций ***FMZ***  и ***FDZ*** для моно- и дизиготных пар:

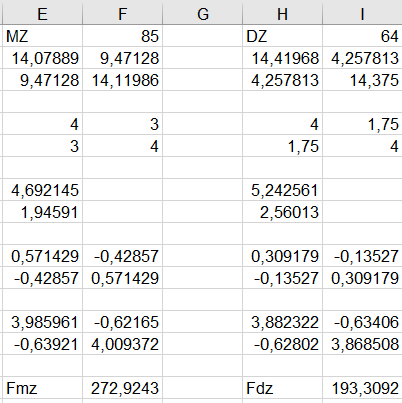
F = FMZ + FDZ .

#### Решение 1.1

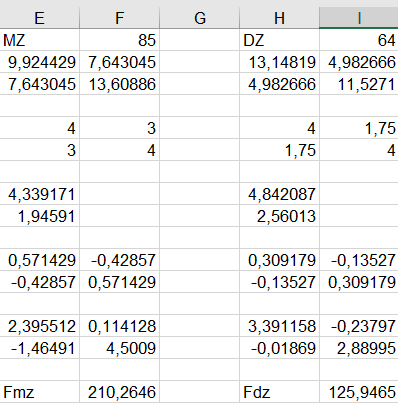
Изначально имеем два листа с параметрами для MZ и DZ. Из них вынес пары 2 и 3 характеристик. После вызвал инструмент ковариация и в качестве областей выбирал пары. Провел это по два раза для MZ и DZ(т.к. вынесено две характеристики на каждом листе). Прикрепляю результаты для MZ и DZ соответственно:  

Далее я скопировал матрицы ковариаций и перенес их на отдельные листы для 2 параметра и 3 параметра. Объем выборок равен 85 и 64 соответственно. Т.к. дана формула для прогнозируемой матрицы ковариаций я построил ее. Для этого задал значения a, d, c, e = 1. Формула: 

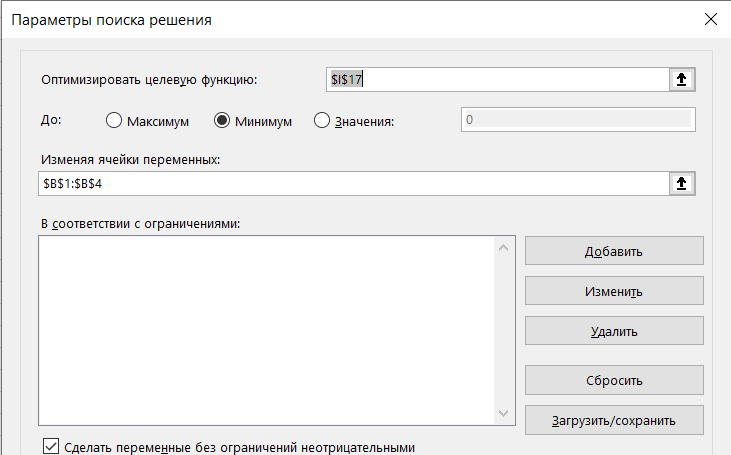
Далее я посчитал определитель изначальной матрицы ковариаций, определитель прогнозируемой матрицы ковариаций + ее обратную матрицу. После этого изначальную матрицу ковариаций умножил на обратную прогнозируемой матрицы ковариаций. Теперь я воспользовался данной выше формулой для нахождения F: 

Результат последних действий по порядку: 

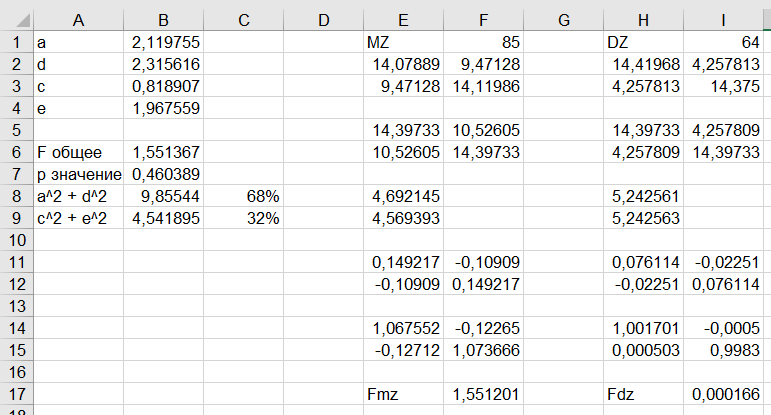
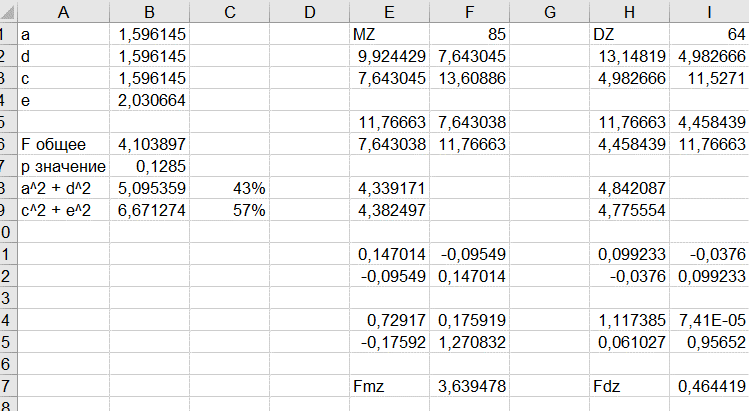
Аналогично для 3 параметра



После этого вызвал Поиск решения, где задал следующее:



Поиск решения нашел значения a,d,c,e. После этого я нашел F общее, p значение, а также оценил степень влияния наследственности и среды на изменчивость. Так как p значение больше 0,05, то модель адекватна. Результат для 2 и 3 параметра:

#### Задание 1.2.

**Проверить нулевую гипотезу.**

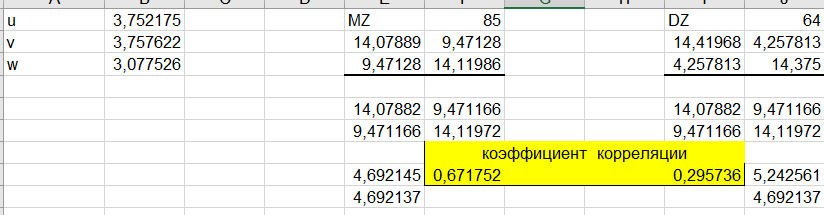
Используя метод максимального правдоподобия, проверить нулевую гипотезу о равенстве корреляций для ***MZ-*** и ***DZ-*** близнецов, а именно: проверить гипотезу о том, что ковариационные матрицы для ***MZ-*** и ***DZ-*** близнецов являются выборочными оценками одной и той же ковариационной матрицы вида:

***u2 w***

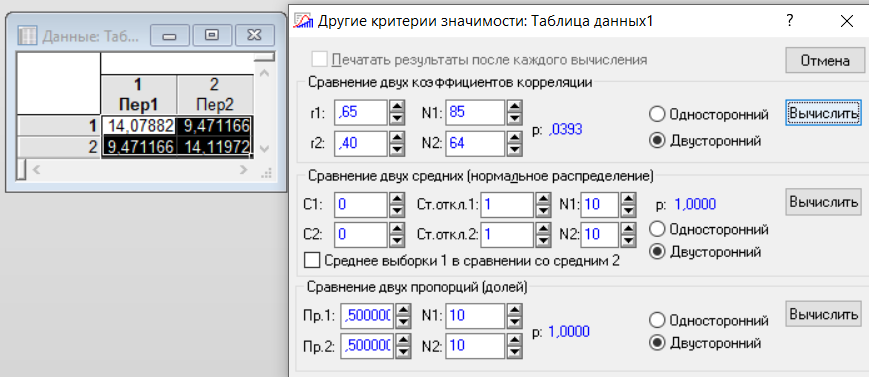
***wv2***

Коэффициент корреляции при этом фактически полагается свободным параметром, причем его значение подбирается таким образом, чтобы обеспечить наилучшее согласование приведенной выше прогнозируемой матрицы с выборочными ковариационными матрицами для ***MZ-*** и ***DZ-*** близнецов одновременно.

#### Решение 1.2

Для проверки выбран 2 параметр. Нашел коэффициент корреляции и провел такие же действия, как в задаче 1.1. С помощью поиска решения нашел u,v,w. Результат: 

После этого я перешел в статистику, создал таблицу 2 на 2 и перенес в нее прогнозируемую матрицу. С помощью пакета Другие критерии значимости я нашел p значение, которое меньше 0,05, а значит могу сказать, что гипотеза неверна. Результат:



#### Задание 1.3.

**Провести анализ упрощенных моделей и выбрать модель, наилучшим образом согласующуюся с наблюдениями.**

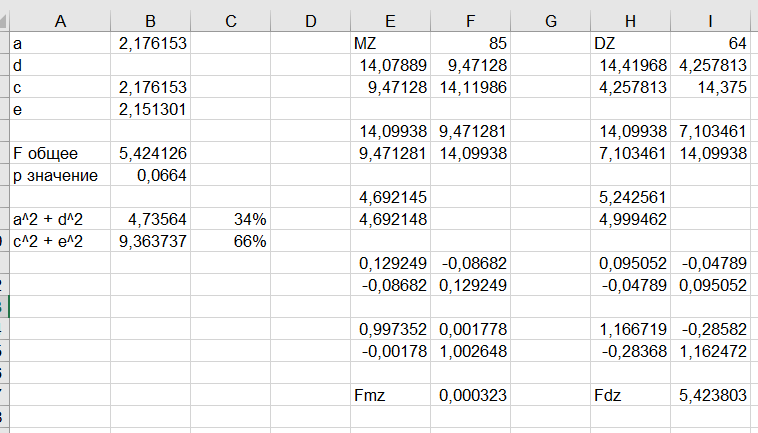
Провести конфирматорный факторный анализ близнецовых данных с помощью упрощенных моделей (***ace***, ***ade***, ***dce, ae, ce***). Используя критерии *хи-квадрат* и *AIC*, выбрать модель, наилучшим образом согласующуюся с наблюдениями. При выборе оптимальных моделей использовать то, что критерий *хи-квадрат* позволяет сравнивать между собой альтернативные модели, а именно: оценку согласованности для полной модели можно сравнивать с той же характеристикой для упрощенных моделей, в которых некоторые из искомых параметров полагаются равными нулю. Поскольку разность в значениях критерия χ2 для полной и упрощенных моделей сама асимптотически описывается распределением χ2, эта величина может быть использована для исследования статистической значимости удаленных из модели параметров.

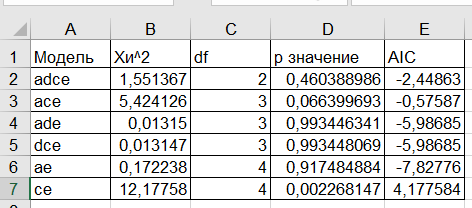
Для каждого исследуемого параметра найти (в %), какие части фенотипической дисперсии обусловлены аддитивными генетическими эффектами, какие - эффектами доминирования, а какие - влиянием общей и индивидуальной сред.

**Результаты расчетов представить в виде следующих таблиц** (для каждого параметра - отдельно):



#### Решение 1.3

Аналогично предыдущей задаче взял 2 параметр. Для построения упрощенных моделей нужно следующее, на примере модели ace: не задавать значение напротив отсутствующей переменной и не изменять ее в поиске решения, а так действия как в задаче 1.1: 

Провел подобное для оставшихся моделей и построил итоговую таблицу: 

Наиболее предпочтительна модель ae, ведь в ней критерий AIC самый маленький.

# Задача 2.

**Симплекс-метод**

#### Задание 2.1.

**Построить ковариационные матрицы для анализа результатов лонгитюдного исследования *IQ* МЗ- и ДЗ-близнецов, используя симплекс-модель, представленную путевой диаграммой.**

## 1

*1 (МЗ) или 0,5 (ДЗ)*

*1 (МЗ и ДЗ)*

***g1***

***g1***

***c1***

***c1***

***e1***

***e1***

## 1-й близнец: 13 лет

## 2-й близнец: 13 лет

## 1

## 1

## 1

## 1

## 1

## 1

*1 (МЗ и ДЗ)*

***g2***

***g2***

***c2***

***c2***

***e2***

***e2***

## 1-й близнец: 16 лет

## 2-й близнец: 16 лет

## 1

## 1

## 1

## 1

## 1

*1 (МЗ) или 0,5 (ДЗ)*

***e12***

***e12***

***c12***

***c12***

***g12***

***g12***

***P1***

***P1***

***P2***

***P2***

Путевая диаграмма, представляющая симплекс-модель для лонгитюдного исследования ***IQ*** (измерения в 13 и 16 лет.)

Наблюдаемые переменные – показатели общего (вариант 1) и вербального (вариант 2) интеллекта. Они обозначаются, соответственно, как ***GIQ*** и ***VIQ***. Латентные переменные обусловлены генетическими эффектами, влиянием общей и индивидуальной сред. Показатели интеллекта рассматриваются как полигенные признаки, для которых выполняются принципы менделевского наследования.

На приведенной диаграмме:

* наблюдаемые переменные обозначаются прямоугольниками;
* латентные переменные обозначаются кругами;
* однонаправленные стрелки (или пути) используются для представления предполагаемых причинных связей между переменными;
* двунаправленные стрелки используются для представления ковариационных связей между переменными (в частности, они же используются и для представления дисперсий);
* ***G1*** и ***G2*** – генетические влияния на ***IQ*** в 13 и 16 лет;
* ***C1*** и ***C2*** – влияния общей среды в 13 и 16 лет;
* ***E1*** и ***E2*** – влияния индивидуальной среды в 13 и 16 лет;
* путевые коэффициенты ***g1***, ***c1***, ***e1*** и ***g2***, ***c2***, ***e2*** описывают влияния латентных переменных для указанных возрастов;
* путевые коэффициенты ***g12***, ***c12*** и ***e12*** описывают влияния переменных ***G1***, ***C1*** и ***E1*** на ***IQ*** в 16 лет (такое представление моделирует ковариацию между показателями ***IQ*** в 13 и 16 лет).

Генетические корреляции для моно- и дизиготных близнецов равны соответственно 1 и 0,5; корреляции общей среды равны 1 для обоих типов близнецов (т. е. предполагается, что они воспитывались в одной и той же семье); индивидуальная среда для разных близнецов не коррелирует по определению.

Результатом работы должны быть аналитические выражения для элементов двух прогнозируемых ковариационных матриц, соответствующих заданной симплекс-модели (отдельно для МЗ- и ДЗ-близнецов; размерность матриц - 4×4).

#### Решение 2.1

Построил две таблицы для MZ и DZ. Они содержат по 4 столбца и строки, т.к. матрица 4 на 4.

Таблица MZ:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 близнец P1 | 2 близнец P1 | 1 близнец P2 | 2 близнец P2 |
| 1 близнец P1 | g1^2+c1^2+e1^2 |  |  |  |
| 2 близнец P1 | 1\*g1^2+1\*c1^2 | g1^2+c1^2+e1^2 |  |  |
| 1 близнец P2 | g1\*g12+c1\*c12+e1\*e12 | g1\*1\*g12+c1\*1\*c12 | g2^2+c2^2+e2^2+g12^2+c12^2+e12^2 |  |
| 2 близнец P2 | g2^2\*g12\*1\*g1+g12\*1\*g1+c2^2\*c12\*1\*c1+c12\*1\*c1 | g1\*g12+c1\*c12+e1\*e12 | 1\*g2^2+1\*c2^2 | g2^2+c2^2+e2^2+g12^2+c12^2+e12^2 |

Таблица DZ:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 близнец P1 | 2 близнец P1 | 1 близнец P2 | 2 близнец P2 |
| 1 близнец P1 | g1^2+c1^2+e1^2 |  |  |  |
| 2 близнец P1 | 0,5\*g1^2+1\*c1^2 | g1^2+c1^2+e1^2 |  |  |
| 1 близнец P2 | g1\*g12+c1\*c12+e1\*e12 | g1\*0,5\*g12+c1\*1\*c12 | g2^2+c2^2+e2^2+g12^2+c12^2+e12^2 |  |
| 2 близнец P2 | g2^2\*g12\*0,5\*g1+g12\*0,5\*g1+c2^2\*c12\*1\*c1+c12\*1\*c1 | g1\*g12+c1\*c12+e1\*e12 | 0,5\*g2^2+1\*c2^2 | g2^2+c2^2+e2^2+g12^2+c12^2+e12^2 |

#### Задание 2.2.

**Провести анализ результатов расчетов, используя таблицы с оценками, полученными методом максимального правдоподобия.**

Приведенные далее оценки свободных параметров модели были вычислены с помощью метода максимального правдоподобия (см. критерий в задаче 1, прилагаемую книгу в формате *Excel* и таблицы 1 и 2). Для оценки адекватности моделей использовались критерии *χ2* и *AIC* (см. определения в задаче 1 и таблицы 1 и 2).

Чтобы получить приемлемое количество статистик, искомые характеристики оценивались одновременно для монозиготных и дизиготных пар близнецов (каждому типу близнецов соответствовали свои прогнозируемые и наблюдаемые матрицы с размерностью 4×4). В качестве минимизируемой функции использовалась сумма соответствующих функций ***FMZ***  и ***FDZ*** для моно- и дизиготных пар.

После анализа полной модели из нее последовательно исключались компоненты ***G*** и ***С***. Исключение ***G*** дает модель, в которой все сходство между близнецами можно отнести за счет общей среды. Исключение ***C*** предполагает, что все внутрипарное сходство обусловлено наследственностью.

На следующем этапе последовательно оценивалось, можно ли исключить из модели новые генетические и средовые влияния в 16 лет и можно ли приравнять нулю генетические и средовые ковариации (т. е. можно ли предполагать полную независимость генетических и средовых переменных в разном возрасте).

Критерий *χ2* позволяет сравнивать между собой альтернативные модели, а именно: оценка согласованности полной модели сравнивается с той же характеристикой для упрощенных моделей, в которых некоторые из искомых параметров полагаются равными нулю. Поскольку разность в значениях критерия χ2 для полной и упрощенных моделей сама описывается распределением χ2, она используется для исследования статистической значимости удаленных из модели компонентов.

Результаты расчетов частично приведены в таблицах 1 и 2.

#### Решение 2.2

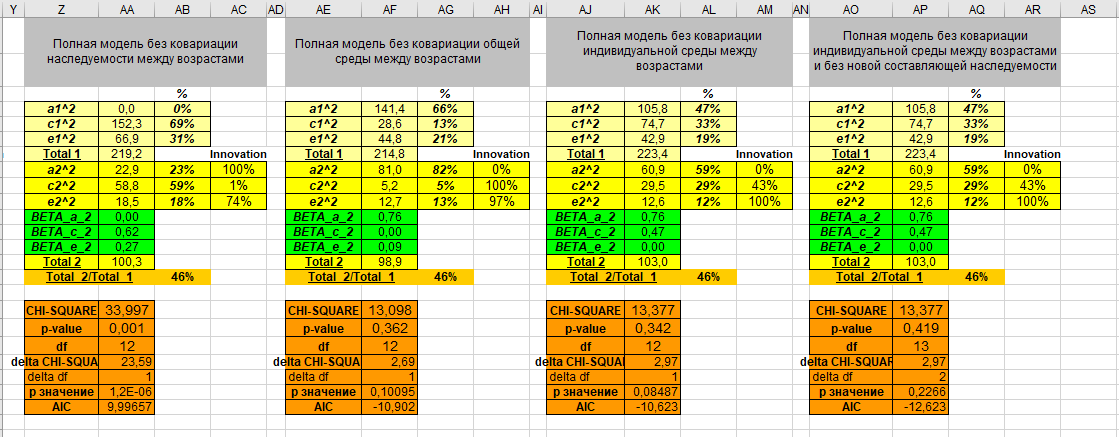
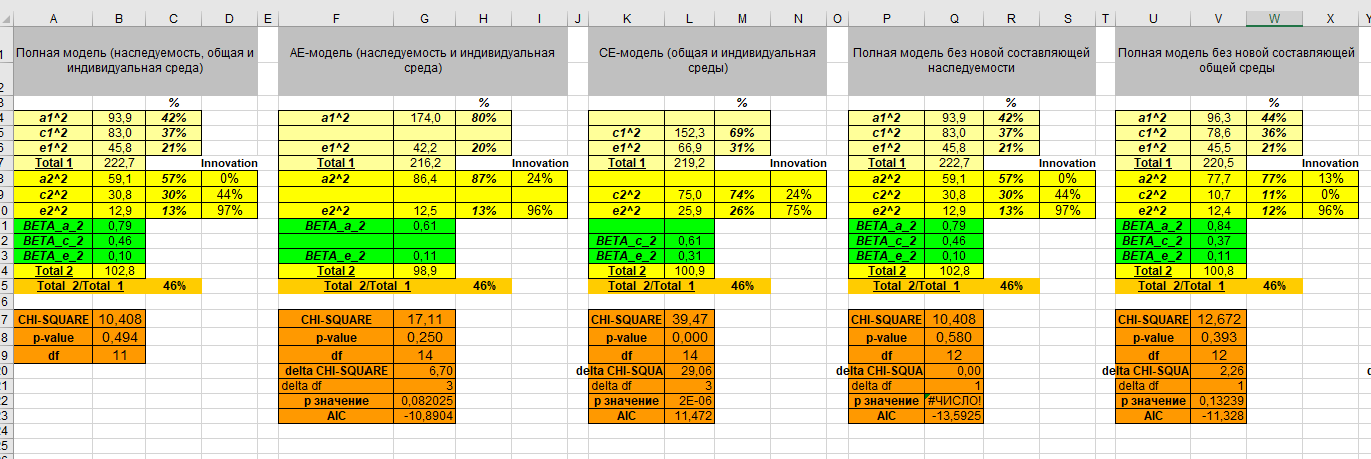
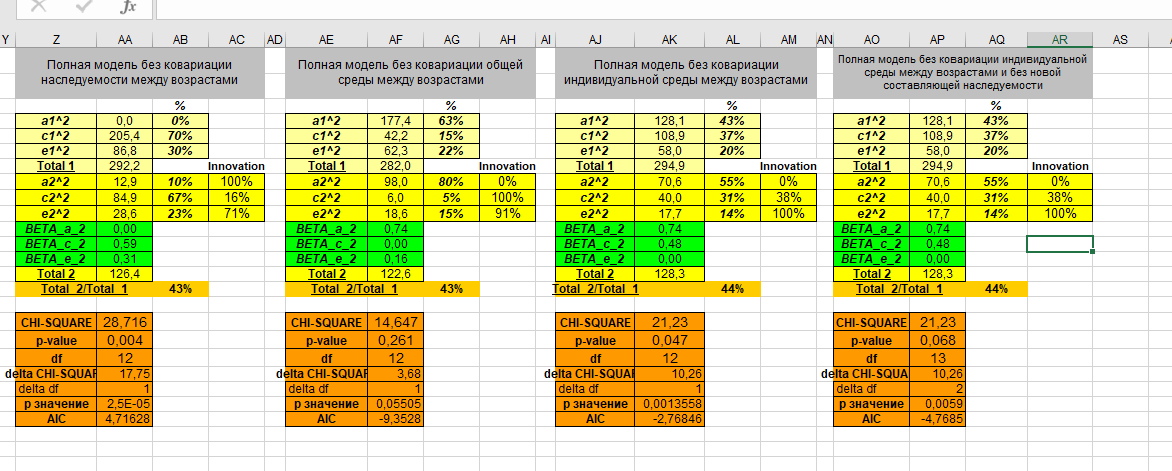
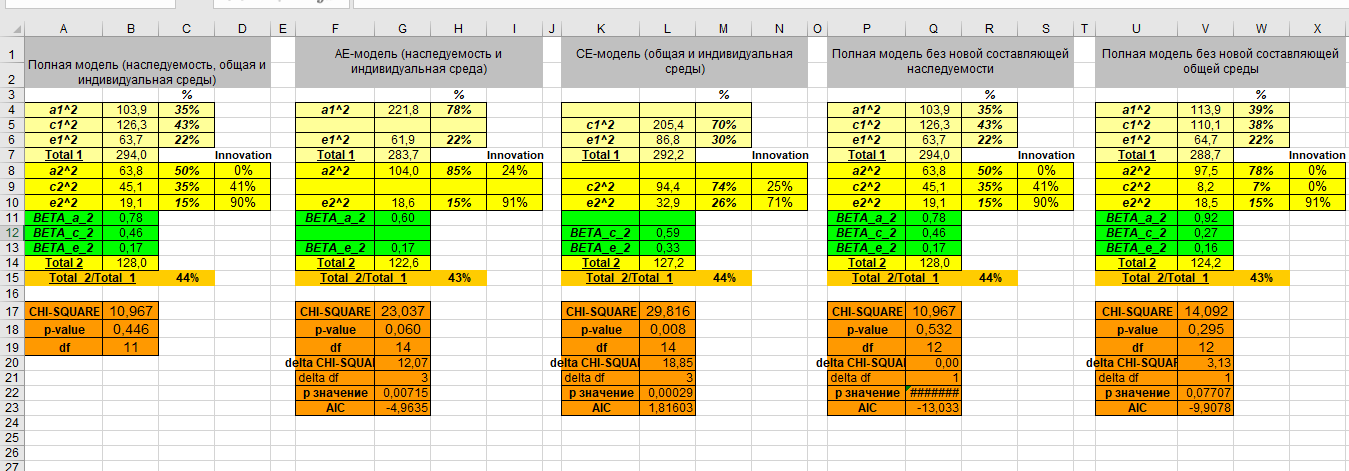
Таблица 1. Соответствие моделей близнецовым данным: ***GIQ***. 



Таблица 2. Соответствие моделей близнецовым данным: ***VIQ***. 



Вывод: предпочтение отдаю модели “Полная модель без новой составляющей наследуемости” в связи с тем, что критерий AIC самый маленький.

Таблица 1а. Оценки параметров для наилучшей модели: ***GIQ***.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Компоненты дисперсии и ковариации** | | | **Всего** |
| **Дисперсия и ковариация** | **Наследуемость** | **Общая**  **среда** | **Индивиду-альная среда** |
| Дисперсия - 13 лет | 93,9 | 83 | 45,8 | 222,7 |
| Дисперсия – 16 лет | 59,1 | 30,8 | 12,9 | 102,8 |
| Ковариация между 13 и 16 годами | 0,79 | 0,46 | 0,1 | 1,35 |
| Новые влияния (инновации) | 0 | 13,5 | 12,5 | 26 |
| **Нормализованные оценки** | **Наследуемость** | **Общая**  **среда** | **Индивиду-альная среда** | **Всего** |
| Дисперсия - 13 лет | 42% | 37% | 21% | 100% |
| Дисперсия – 16 лет | 57% | 30% | 13% | 100% |
| Ковариация между 13 и 16 годами | 59% | 34% | 7% | 100% |
| Новые влияния (инновации) в % к соответствующей составляющей дисперсии | 0% | 44% | 97% | - |

Таблица 2а. Оценки параметров для наилучшей модели: ***VIQ***.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Компоненты дисперсии и ковариации** | | | **Всего** |
| **Дисперсия и ковариация** | **Наследуемость** | **Общая**  **среда** | **Индивиду-альная среда** |
| Дисперсия - 13 лет | 103,9 | 126,3 | 63,7 | 294 |
| Дисперсия – 16 лет | 63,8 | 45,1 | 19,1 | 128 |
| Ковариация между 13 и 16 годами | 0,78 | 0,46 | 0,17 | 1,41 |
| Новые влияния (инновации) | 0 | 18,5 | 17,2 | 35,7 |
| **Нормализованные оценки** | **Наследуемость** | **Общая**  **среда** | **Индивиду-альная среда** | **Всего** |
| Дисперсия - 13 лет | 35% | 43% | 22% | 100% |
| Дисперсия – 16 лет | 50% | 35% | 15% | 100% |
| Ковариация между 13 и 16 годами | 55% | 33% | 12% | 100% |
| Новые влияния (инновации) в % к соответствующей составляющей дисперсии | 0% | 41% | 90% | - |