# Ограничения исследования

* Неполнота данных или недоступность некоторых важных переменных.
* Возможные проблемы с мультиколлинеарностью, автокорреляцией или гетероскедастичностью.
* Ограниченность выборки по времени или географии.
* Возможные ошибки измерения переменных.

# Проблема пропущенной переменной в эконометрике

Проблема пропущенной переменной возникает, когда в регрессионной модели отсутствует важная объясняющая переменная, которая имеет влияние на зависимую переменную.

Это может привести к смещению оценок коэффициентов, так как пропущенная переменная может быть коррелирована с включенными переменными, что нарушает одно из ключевых предположений МНК — отсутствие корреляции между ошибками и объясняющими переменными.

Если пропущенная переменная (Z) влияет на (Y) и коррелирована с (X), то математическое ожидание (Y) при фиксированном (X) будет искажено

Контрольные переменные включаются в модель для учета влияния других факторов, которые могут искажать оценку интересующего нас эффекта, а не для замены отсутствующих переменных. Если важная переменная отсутствует, это может вызвать проблему пропущенной переменной, и контрольные переменные могут помочь уменьшить смещение оценок, но они не заменяют полностью пропущенные переменные.

Пропущенная переменная считается проблемой, если она одновременно коррелирует с зависимой переменной и с одним или несколькими регрессорами. Если пропущенная переменная влияет на зарплату, но не коррелирует с полом, это не приведет к смещению оценки эффекта пола на зарплату. Однако, если пропущенная переменная коррелирует с полом, это может привести к смещению оценок.

1. **Корреляция с зависимой переменной и регрессорами**:
   * Если пропущенная переменная коррелирует как с зависимой переменной, так и с регрессорами, это приводит к смещению оценок коэффициентов регрессоров. Это происходит потому, что эффект пропущенной переменной "переносится" на регрессоры, искажая их истинные эффекты.
2. **Корреляция только с зависимой переменной**:
   * Если пропущенная переменная коррелирует только с зависимой переменной, но не коррелирует с регрессорами, это не приводит к смещению оценок коэффициентов регрессоров. В этом случае пропущенная переменная будет учтена в случайной ошибке ((\epsilon)), но оценки регрессоров останутся несмещенными. Может увеличивать дисперсию случайно ошибки.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

# Контрольная переменная

Контрольная переменная — это переменная, которая включается в регрессионную модель для учета влияния других факторов, которые могут искажать оценку интересующего нас эффекта. Например, если мы исследуем влияние образования на заработную плату, контрольными переменными могут быть возраст, опыт работы и регион проживания.

Контрольная переменная включается в модель не для замены отсутствующих переменных, а для учета других факторов, которые могут влиять на зависимую переменную и тем самым искажать оценку интересующего нас эффекта. Например, если мы исследуем влияние образования на заработную плату, контрольными переменными могут быть возраст, опыт работы и регион проживания, чтобы учесть их влияние и получить более точную оценку эффекта образования.

Контрольные переменные включаются в модель для учета влияния других факторов, которые могут искажать оценку интересующего нас эффекта. Они помогают уменьшить смещение оценок и повысить точность модели. Однако они не всегда должны быть интерпретированы, их основная цель — улучшить модель.

# Теорема Гаусса-Маркова

Теорема Гаусса-Маркова утверждает, что при выполнении определенных предпосылок оценки коэффициентов, полученные методом наименьших квадратов (МНК), являются лучшими линейными несмещенными оценками (BLUE).

#### **Предпосылки теоремы Гаусса-Маркова:**

1. **Линейность модели**: Модель должна быть линейной по параметрам.
2. **Случайный отбор**: Данные должны быть случайной выборкой из генеральной совокупности.
3. **Независимость наблюдений**: Наблюдения должны быть независимыми друг от друга.
4. **Гомоскедастичность**: Дисперсия ошибок должна быть постоянной для всех наблюдений.
5. **Отсутствие мультиколлинеарности**: Независимые переменные не должны быть линейно зависимы между собой.

Если есть мультиколлинеарность, теорема Гаусса-Маркова не выполняется, и оценки могут быть неэффективными.

# Проверка предпосылок теоремы Гаусса-Маркова

Для проверки предпосылок теоремы Гаусса-Маркова необходимо:

1. **Линейность модели**: Убедиться, что модель линейна по параметрам.
2. **Случайный отбор**: Проверить, что данные являются случайной выборкой.
3. **Независимость наблюдений**: Использовать тесты на автокорреляцию, такие как тест Дарбина-Уотсона.
4. **Гомоскедастичность**: Применить тесты на гетероскедастичность, например тест Бройша-Пагана или тест Уайта.
5. **Отсутствие мультиколлинеарности**: Проверить мультиколлинеарность с помощью VIF (Variance Inflation Factor).

# Тест Хьюбера-Уайта

**Тест Хьюбера-Уайта** (или просто **робастные стандартные ошибки Хьюбера-Уайта**) — это метод, который корректирует стандартные ошибки коэффициентов регрессии для учета гетероскедастичности. Это не тест на гетероскедастичность, а метод оценки, который делает стандартные ошибки более надежными в присутствии гетероскедастичности.

# ВМНК и МНК

**Метод наименьших квадратов (МНК)** — это стандартный метод для оценки параметров линейной регрессии, который минимизирует сумму квадратов разностей между наблюдаемыми значениями зависимой переменной и предсказанными значениями.

**Взвешенный метод наименьших квадратов (ВМНК)** — это обобщение МНК, который применяется, когда дисперсии ошибок не являются постоянными (гетероскедастичность). ВМНК присваивает каждому наблюдению вес, обратный его дисперсии, что позволяет минимизировать влияние наблюдений с высокой дисперсией ошибок.

# Предпосылки МНК

1. **Линейность модели**.
2. **Случайный отбор**.
3. **Независимость наблюдений**.
4. **Гомоскедастичность**.
5. **Нормальность ошибок**.
6. **Отсутствие мультиколлинеарности**.

Некоторые предпосылки МНК можно проверить до построения модели, а некоторые — только после.

# Предпосылки, которые можно проверить до построения модели:

1. **Линейность модели**: Проверить, что предполагаемая модель линейна по параметрам.
2. **Случайный отбор**: Убедиться, что данные являются случайной выборкой.
3. **Отсутствие мультиколлинеарности**: Начальная проверка с использованием корреляционной матрицы.

# Предпосылки, которые можно проверить после построения модели:

1. **Независимость наблюдений**: Проверка на автокорреляцию (тест Дарбина-Уотсона).
2. **Гомоскедастичность**: Проверка на гетероскедастичность (тест Бройша-Пагана).
3. **Нормальность ошибок**: Проверка нормальности остатков (тест Шапиро-Уилка, Q-Q plot).

Корреляционная матрица может дать начальное представление о мультиколлинеарности, показывая корреляции между независимыми переменными. Однако VIF (Variance Inflation Factor) предоставляет более точную оценку мультиколлинеарности, так как он учитывает влияние всех других переменных в модели. VIF выше 10 указывает на серьезные проблемы мультиколлинеарности.

# Последствия невыполнения предпосылок

1. **Линейность модели**: Если модель нелинейна по параметрам, оценки могут быть несостоятельными и смещенными.
2. **Случайный отбор**: Нарушение этого предположения может привести к нерепрезентативности выборки и смещению оценок.
3. **Независимость наблюдений**: Нарушение независимости (автокорреляция) приводит к неэффективности оценок, хотя они могут оставаться несмещенными.
4. **Гомоскедастичность**: Нарушение гомоскедастичности (гетероскедастичность) приводит к неэффективности оценок и недостоверности стандартных ошибок и тестов значимости.
5. **Отсутствие мультиколлинеарности**: Высокая мультиколлинеарность приводит к неэффективности оценок и увеличению стандартных ошибок, что затрудняет интерпретацию результатов.

# Проверка мультиколлинеарности, автокорреляции и гетероскедастичности

1. **Мультиколлинеарность**:
   * Использование VIF (Variance Inflation Factor). Значения VIF выше 10 указывают на серьезные проблемы мультиколлинеарности.
   * Анализ корреляционной матрицы независимых переменных.
2. **Автокорреляция**:
   * Тест Дарбина-Уотсона.
   * Тест Бреуша-Годфри.
3. **Гетероскедастичность**:
   * Тест Бройша-Пагана.
   * Тест Уайта.

# Регрессия с одним регрессором

Вы можете построить регрессию с одним регрессором (пол человека) и зависимой переменной (доход человека за последние 30 дней). Это даст вам оценку среднего различия в доходах между мужчинами и женщинами. Однако такая модель может быть недостаточной по нескольким причинам:

1. **Пропущенные переменные**: Другие факторы (образование, опыт работы, отрасль и т.д.) также влияют на доход. Если они не включены в модель, это может привести к смещению оценок.
2. **Контроль за другими факторами**: Включение других регрессоров позволяет учесть влияние этих факторов и получить более точную оценку эффекта пола на доход.
3. **Улучшение модели**: Включение дополнительных регрессоров может повысить объяснительную силу модели (увеличение ( R^2 )).

### 0. Титульный слайд

### 1. Введение

* Цели исследования
* Описание проблемы гендерного разрыва в заработных платах

### 2. Данные

* Источник данных
* Описание переменных (зависимая, независимые, контрольные)
* Статистический обзор данных (средние значения, стандартные отклонения и т.д.)

### 3. Ограничения исследования

* Возможные ограничения данных (неполнота, недоступность важных переменных)
* Возможные ошибки измерения переменных

### 4. Проблема пропущенных переменных

* Объяснение проблемы пропущенных переменных
* Влияние пропущенных переменных на результаты модели
* Методы учета пропущенных переменных (включение контрольных переменных)

### 5. Предпосылки МНК и теоремы

* Линейность модели
* Случайный отбор
* Независимость наблюдений
* Гомоскедастичность
* Нормальность ошибок (для проведения тестов)
* Отсутствие мультиколлинеарности
* Проверка выполнения предпосылок (тесты и результаты)

### 6. Выбор регрессоров и построение модели

* Обоснование выбора регрессоров (основные и контрольные переменные)
* Описание построенной модели (формула регрессии)
* Проверка на мультиколлинеарность (VIF), автокорреляцию (тест Дарбина-Уотсона) и гетероскедастичность (тест Бройша-Пагана)

### 7. Результаты

* Представление результатов моделей для каждого года и региона
* Интерпретация ключевых коэффициентов
* Влияние контрольных переменных

На подумать (множество возможных альтернативных моделей, включающих дополнительные переменные)