

Лабораторная работа 1. Первичный (разведочный) анализ данных

Лабораторная работа выполняется с помощью Python (библиотеки Scipy + NumPy, Pandas) или пакета R. Теоретический материал ниже приводится для R.

Оценка выборочных характеристик описательной статистики

Благодаря наличию набора специально созданных для этого функций, расчет параметров описательной статистики в R не составляет труда. Продемонстрируем их использование на примере данных по характеристикам 32 моделей автомобилей (таблица mtcars, входящая в стандартный набор данных R):

```
data(mtcars)
head(mtcars)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1

Каждая модель автомобиля описана по 11 признакам, два из которых (vs и am) являются номинальными переменными (факторами) с уровнями, закодированными в виде 0 и 1.

Для расчета среднего арифметического, медианы, дисперсии, стандартного отклонения, а также минимального и максимального значений в R служат функции mean(), median(), var(), sd(), min() и max() соответственно. Используем эти функции в отношении, например, параметра mpg – пробег автомобиля (в милях) в расчете на один галлон топлива:

```
# Арифметическая средняя:
mean(mtcars$mpg)
[1] 20.1
# Медиана
median(mtcars$mpg)
[1] 19.2
# Дисперсия:
var(mtcars$mpg)
[1] 36.3
# Стандартное отклонение:
sd(mtcars$mpg)
[1] 6.0
# Минимальное значение:
min(mtcars$mpg)
[1] 10.4
# Максимальное значение:
max(mtcars$mpg)
[1] 33.9
```

Квантили рассчитываются в R при помощи функции quantile():

```
quantile(mtcars$mpg)
```

```
0%    25%   50%   75%   100%  
10.400 15.425 19.200 22.800 33.900
```

При настройках, заданных по умолчанию, выполнение указанной команды приведет к расчету минимального (10.4) и максимального (33.9) значений, а также трех *квартилей*, т.е. значений, которые делят совокупность на четыре равные части – 15.4, 19.2 и 22.8.

Разница между первым и третьим квартилями носит название *интерквартильный размах* (*ИКР*; англ. *interquartile range*). ИКР является *робастным* аналогом дисперсии и может быть рассчитан в R при помощи функции **IQR()**:

```
IQR(mtcars$mpg)
```

```
[1] 7.375
```

Отсутствующие значения в данных могут несколько усложнить вычисления. В целях демонстрации заменим 3-е значение переменной `mpg` на **NA**, т.е. на обозначение, используемое в R для отсутствующих наблюдений (от *not available* – не доступно), а затем попытаемся вычислить среднее значение:

```
mtcars$mpg[3] <- NA  
# Просмотрим результат:  
head(mtcars$mpg)  
[1] 21.0 21.0 NA 21.4 18.7 18.1  
# Попробуем рассчитать среднее значение для mpg:  
mean(mtcars$mpg)  
[1] NA
```

Получили ошибку, так как R не пропускает отсутствующие значения автоматически, если мы не включим соответствующую опцию – `na.rm` (от *not available* и *remove* – удалить):

```
mean(mtcars$mpg, na.rm = TRUE)  
[1] 20.0
```

В системе R имеется возможность и более быстрого расчета основных параметров описательной статистики. Для этого, в частности, служит *функция общего назначения* `summary()`:

```
summary(mtcars$mpg)  
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's  
10.40   15.35   19.20   20.00   22.15   33.90    1.00
```

Подобную сводку мы можем получить сразу для всей таблицы данных.

Протокол разведочного анализа данных

В публикации ([Zuur et al., 2010](#)) предложен протокол разведочного анализа данных, который состоит в следующем.

- Выявление точек-выбросов
- Проверка однородности дисперсий
- Проверка нормальности распределения данных
- Выявление избыточного количества нулевых значений
- Выявление коллинеарных переменных

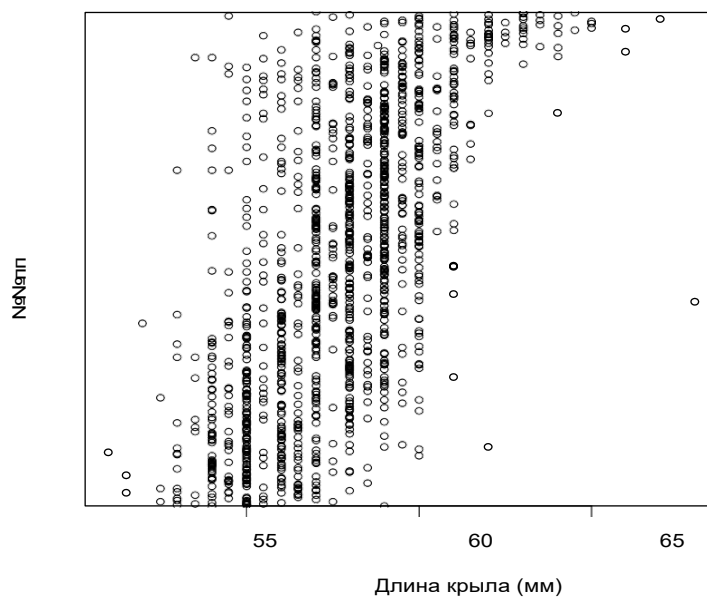
- Выявление характера связи между анализируемыми переменными
- Выявление взаимодействий между переменными-предикторами
- Выявление пространственно-временных корреляций между значениями зависимой переменной

Анализ выбросов

Под "выбросом" мы будем понимать наблюдение, которое "слишком" велико или "слишком" мало по сравнению с большинством других имеющихся наблюдений.

Обычно для выявления выбросов используют диаграмму размахов или точечную диаграмму Кливленда.

Например, на следующей диаграмме Кливленда, представлены данные о длине крыла у 1295 воробьев ([Zuur et al. 2010](#)). Здесь таблица была предварительно упорядочена в соответствии с весом птиц, и поэтому облако точек имеет примерно S-образную форму.



На рисунке хорошо выделяется точка, соответствующая длине крыла 68 мм. Однако это значение длины крыла не следует рассматривать в качестве выброса, поскольку оно лишь незначительно отличается от других значений длины. Эта точка выделяется на общем фоне лишь потому, что исходные значения длины крыла были упорядочены по весу птиц. Соответственно, выброс скорее стоит искать среди значений веса (т.е. очень высокое значение длины крыла (68 мм) было отмечено у воробья, необычно мало весящего для этого).

Более строгий подход к определению выбросов состоит в оценке того, какое влияние эти необычные наблюдения оказывают на результаты анализа. При этом следует делать различие между необычными наблюдениями для зависимых и независимых переменных (предикторов).

Альтернативой удалению необычных значений предиктора является нормализующее преобразование (чаще всего, логарифмирование).

В конечном счете, решение об удалении необычных значений из анализа принимает сам исследователь. При этом он должен помнить о том, что причины для возникновения таких наблюдений могут быть разными.

Заполнение пропущенных значений в таблицах данных

Большинство статистических методов предполагает, что в ходе наблюдений были

получены полностью укомплектованные матрицы, векторы и другие информационные структуры эксперимента. Поскольку на практике пропуски в данных являются повсеместным явлением, прежде чем начать аналитические изыскания, необходимо привести обрабатываемые таблицы к "каноническому" виду, т.е. либо удалить фрагменты объектов с недостающими элементами, либо заменить имеющиеся пропуски на некоторые разумные значения.

На практике процедура "борьбы с пропусками" обычно включает следующие шаги:

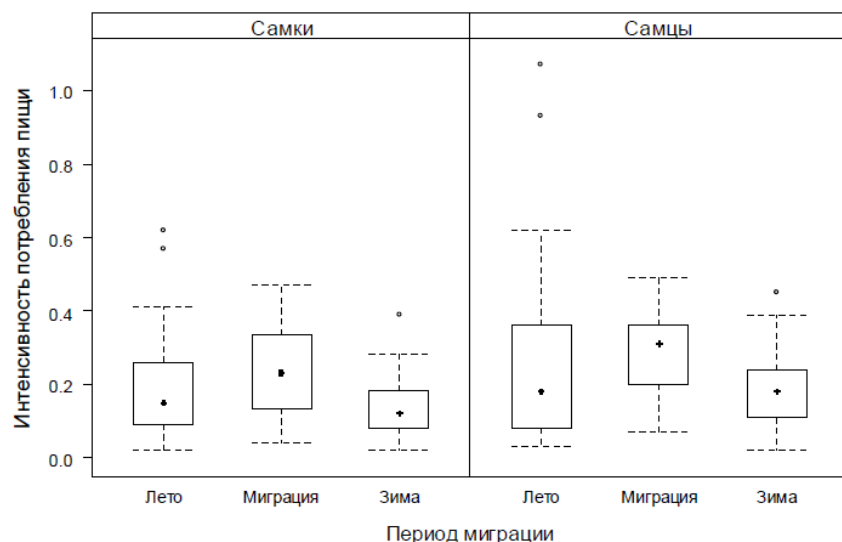
1. Идентификация недостающих данных.
2. Исследование закономерностей появления отсутствующих значений.
3. Формирование наборов данных, не содержащих пропуски (в результате удаления или замены соответствующих фрагментов).

Необходимо признать, что идентификация недостающих данных является здесь единственным однозначным шагом. Анализ, почему данные отсутствуют, зависит от вашего понимания процессов, которые воспроизводят экспериментальную информацию. Решение о способе устранения пропущенных значений также будет зависеть от вашей оценки того, какие процедуры приведут к самым надежным и точным результатам. Можно порекомендовать использование пакета **mice**, который является хорошим средством реализации всех шагов упомянутой процедуры

Проверка однородности групповых дисперсий

С проблемой выбросов тесно связан второй пункт протокола разведочного анализа – проверка условия однородности дисперсии (англ. вариант термина "*homogeneity of variance*"). Однородность групповых дисперсий постулируется как важное условие применимости дисперсионного анализа (ANOVA) и других линейных моделей регрессионного типа, а также ряда методов многомерной статистики (например, дискриминантного анализа).

На рисунке ниже приведены категоризованные диаграммы размахов для значений интенсивности потребления пищи канадским веретенником – птицы из семейства бекасовых ([Zuur et al., 2010](#)).



Если бы стояла задача применить параметрический дисперсионный анализ для установления эффектов пола и периода наблюдений на интенсивность потребления пищи веретенником (а также взаимодействия между этими двумя факторами), то должны были бы выполняться следующие условия:

- дисперсии не различаются у самцов и самок;
- дисперсии не различаются между периодами наблюдений;

- дисперсии не различаются между периодами наблюдения у каждого из полов.

Как можно увидеть из графика, дисперсия интенсивности потребления пищи у самцов низка зимой и несколько повышается в летний период, но в целом в приведенном примере различия групповых дисперсий невелики и не потребуют особых действий со стороны аналитика.

При выявлении существенной неоднородности дисперсии возможны два решения: определенное преобразование исходных данных (например, логарифмирование) или использование моделей, основанных на обобщенном методе наименьших квадратов (*Generalized Least Squares Models*) и допускающих неоднородность дисперсии.

Выявление коллинеарности

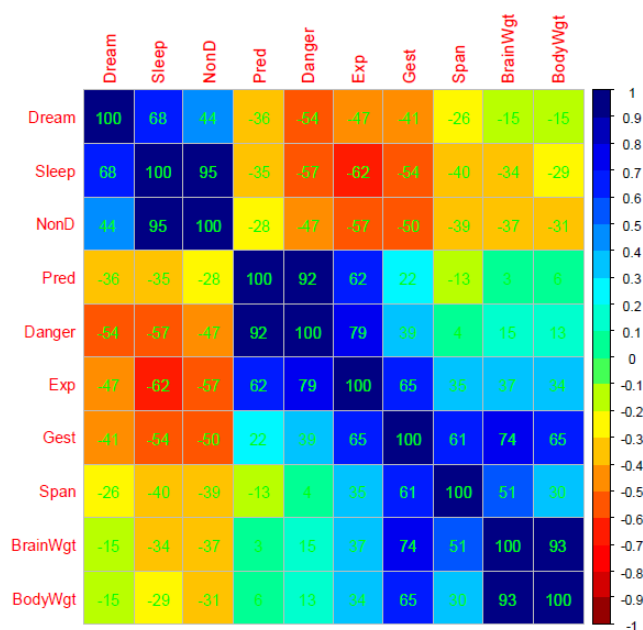
Когда цель анализа заключается в нахождении переменных (*предикторов*), связанных со значениями *зависимой переменной*, важным этапом разведочного анализа данных является обнаружение коллинеарности. Под *коллинеарностью* (англ. *collinearity*) понимают наличие линейной зависимости между двумя предикторами. В задачах с несколькими предикторами (например, при выполнении *множественного регрессионного анализа*) говорят также о *мультиколлинеарности* (англ. *multicollinearity*), т.е. наличии линейной зависимости сразу между несколькими переменными.

Проблема *мультиколлинеарности* приводит к проблемам при вычислении регрессионных коэффициентов: оценки параметров модели оказываются неустойчивыми и становится трудно анализировать вклад каждого отдельного фактора в прогнозируемую величину. Как результат, исследователь может столкнуться с "парадоксальной" ситуацией, когда, например, все коэффициенты множественной регрессионной модели статистически незначимы, тогда как сама модель оказывается значимой (т.е. проверяемая при помощи *F*-теста гипотеза о равенстве *всех* коэффициентов нулю отвергается).

Традиционным способом оценки мультиколлинеарности является анализ корреляционной матрицы. Хорошим способом повысить наглядность отображения уровня взаимных зависимостей переменных является использование раскрашенных матриц, создаваемых функцией **corrplot()** из одноименного пакета. Рассмотрим пример с изучением потребности в сне у животных:

```
load(file="sample-data/sleep_imp.Rdata") M <- cor(sleep_imp3)
library(corrplot)
col4 <- colorRampPalette(c("#7F0000", "red", "#FF7F00", "yellow",
"#7FFF7F", "cyan", "#007FFF", "blue", "#00007F"))
corrplot(M, method="color", col=col4(20), cl.length=21, order =
"AOE",
addCoef.col="green")
```

Нетрудно выделить на рисунке два блока предикторов: таксономические (BodyWgt + BrainWgt) и экологические (Pred + Danger) показатели, которые высоко коррелируют между собой.



Другим способом оценить степень мультиколлинеарности m -мерного комплекса переменных является вычисление фактора инфляции дисперсии (*Variance Inflation Factor*, VIF), который для каждого j -го предиктора из m тем выше, чем теснее его линейная связь со всеми остальными ($m - 1$) независимыми переменными. Значения VIF находятся следующим образом:

- строится регрессионная модель зависимости переменной X_j от всех остальных предикторов: $X_j = \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_m X_m + \beta_0 + \varepsilon$;
- рассчитывается коэффициент детерминации R_j^2 для полученной регрессионной модели и оценивается доля увеличения дисперсии коэффициента регрессии β_j вследствие эффекта корреляции данных по формуле: $VIF(\beta_j) = 1/(1 - R_j^2)$;
- эти действия выполняются последовательно для всех переменных $j = 1, 2, \dots, m$.
- Значение фактора инфляции дисперсии, равное 1, говорит об ортогональности вектора реализаций признака по отношению ко всем остальным. Принято считать, что значения VIF_j от 1 до 2 (что соответствует R_j^2 от 0 до 0.5) означают отсутствие проблемы мультиколлинеарности для X_j , т.е. добавление в модель или удаление из нее любых других независимых переменных не изменяет ни самой оценки коэффициента β_j , ни критериев его статистической значимости. Если $VIF_j = 4$, то это означает, что ошибка выборочной оценки параметра β_j в $4^{1/2} = 2$ раза превышает эту оценку, но полученную при полном отсутствии мультиколлинеарности.

Функции для автоматического расчета VIF и выполнения перечисленных выше шагов реализованы в нескольких пакетах для R. Одним из примеров таких функций является `vif()` из пакета `car`:

```
library(car)
vif(lm(Sleep~BodyWgt+BrainWgt+Span+Gest+Pred+Exp+Danger, data=sleep_imp3))
      BodyWgt BrainWgt      Span      Gest      Pred      Exp      Danger
      12.363      15.895      2.659      4.124      8.211      4.921      12.879
```

Представленные оценки вектора VIF указывают на высокую мультиколлинеарность предикторных переменных рассматриваемого примера.

Предложена итерационная процедура построения моделей с использованием фактора инфляции дисперсии: значения VIF_j сравниваются с выбранным пороговым значением и предиктор с максимальным VIF , превышающим пороговое значение, исключается из анализа (эти шаги повторяются, пока в модели не останутся предикторы с низкими VIF). При этом к формальному исключению предикторов стоит относиться

критически. Так, например, нет единого мнения в отношении того, какое значение VIF следует считать пороговым (обычно $VIF = 5$, хотя реже предлагается и $VIF = 10$).

Выявление характера связи между переменными

Существующие в природе количественные соотношения между переменными являются, как правило, нелинейными, и часто выполняемое приведение их к линейной форме – лишь одна из возможностей удобной аппроксимации, которая не обязательно будет удовлетворительно моделировать изучаемые процессы по наблюдаемым данным. Поэтому на этапе разведочного анализа данных важно исследовать характер взаимодействия между анализируемыми переменными. Обнаруженные на этом этапе закономерности будут определять выбор статистической модели для описания данных, необходимость преобразования нелинейно связанных переменных и т.д.

Характер связи между переменными проще всего выявить, используя соответствующие графические средства. Так, при анализе нескольких количественных переменных очень удобным инструментом являются *матричные диаграммы рассеяния*. Матричные диаграммы рассеяния в R можно построить при помощи нескольких функций. Проще всего воспользоваться базовой R-функцией `pairs()`, которая имеет ряд аргументов для тонкой настройки графика.

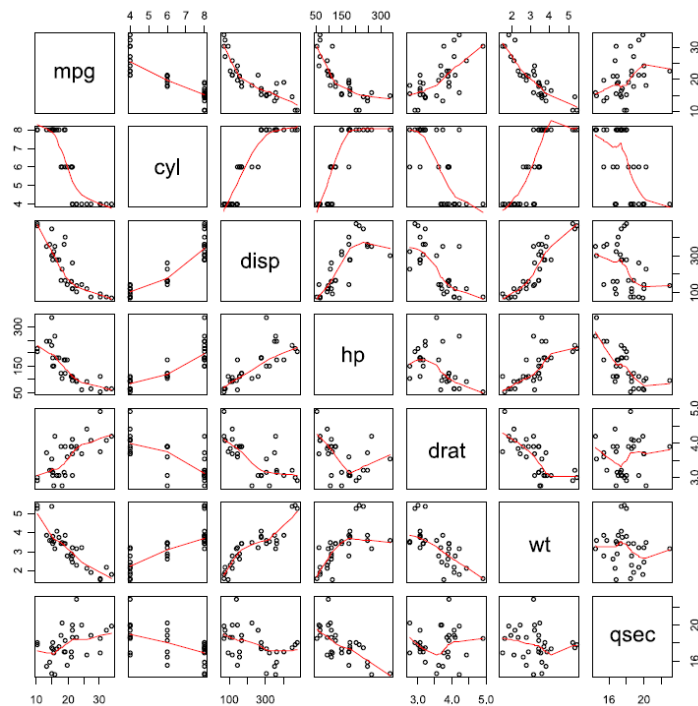
В качестве примера рассмотрим первые семь количественных параметров, описывающих разные модели автомобилей, из таблицы `mtcars`. Для облегчения интерпретации характера связи между анализируемыми переменными мы можем добавить сглаживающую кривую к каждой диаграмме рассеяния (аргумент `panel` со значением `panel.smooth`):

```
cars <- mtcars[, 1:7]
pairs(cars, panel = panel.smooth)
```

Аргументы `lower.panel` и `upper.panel` позволяют назначить практически любую функцию для преобразования исходных данных и последующего отображения результатов работы этой функции на графике (ниже или выше центральной диагонали матрицы соответственно). Например, вот так мы можем добавить к графику значения коэффициентов корреляции Спирмена, определив собственную функцию заполнения панели:

```
# Функция для оформления панелей с коэффициентом корреляции
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y, method = "spearman"))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep = "")
  # эта команда позволяет изменять размер шрифта
  # в соответствии со значением коэффициента корреляции:
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}

# Строим сам график:
pairs(cars, , panel = panel.smooth, lower.panel = panel.cor)
```



Выявление взаимодействий между предикторами

Как было показано выше, вопрос о силе и характере взаимодействия между предикторами, измеренными в номинальных шкалах, решается с использованием корреляционного анализа или функций сглаживания. В отношении категориальных переменных отличным инструментом для выявления взаимодействий между анализируемыми предикторами являются категоризованные графики R.

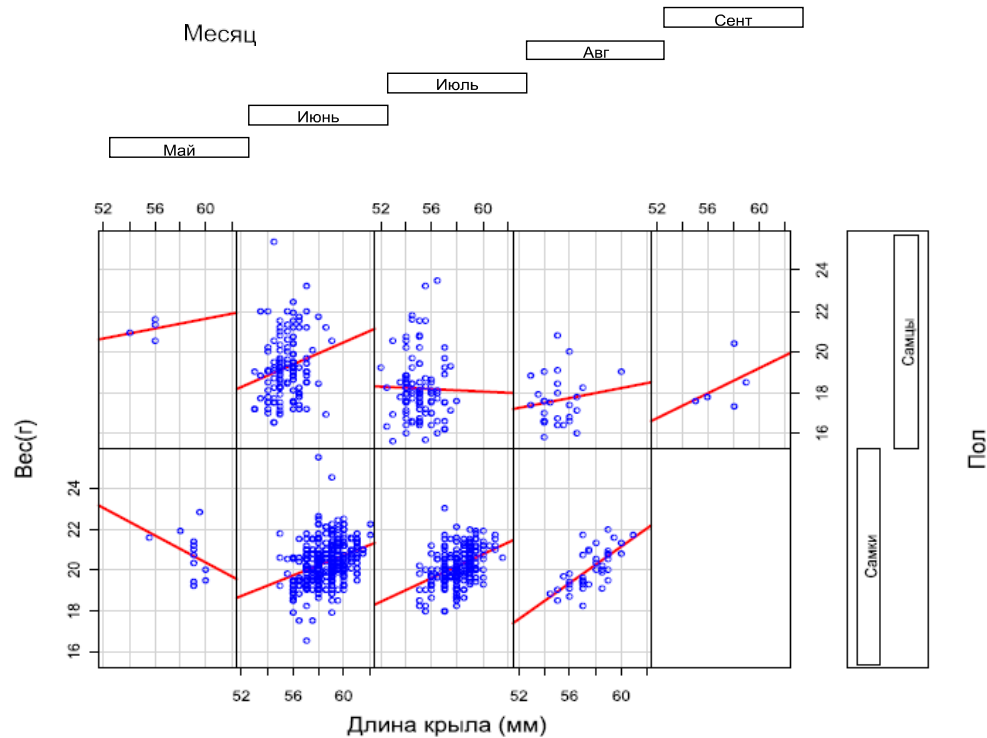
Обратимся к орнитологическому примеру. Исследователи задались вопросом о существовании *взаимодействия между* весом воробьев (weight), длиной их крыльев (length), полом (sex) и временем года (month).

```
# Загрузка исходных данных из файла
Sparrows <- read.table(file = "sample-data/SparrowsElphick.txt",
header = TRUE)
# Отбираем необходимый комплект данных: воробьи с длиной крыла >= 65
I1 <- Sparrows$SpeciesCode==1 & Sparrows$Sex != "0" &
Sparrows$wingcrd<65
Wing1<- Sparrows$wingcrd[I1]
Wt1 <- Sparrows$wt[I1]
Mon1 <- factor(Sparrows$Month[I1])
Sex1<- factor(Sparrows$Sex[I1])
# Определим месяц и пол как категориальные переменные
fMonth1 <- factor(Mon1, levels=c(5,6,7,8,9),
labels=c("May", "Jun", "Jul", "Aug", "Sep"))
fSex1 <- factor(Sex1, levels =c(4,5),
labels=c("Male", "Female"))
```

Силу и направление связи между весом воробьев и длиной крыла в зависимости от их пола и времени проведения измерений наглядно и во всех деталях можно проанализировать на категоризованных диаграммах рассеяния. Для удобства интерпретации этой связи к каждому графику добавим линии регрессии. Если прямые на всех приведенных диаграммах окажутся параллельными, то можно сделать вывод об отсутствии влияния пола воробьев и времени года на связь между длиной крыла и весом

ПТИЦ.

```
# Вывод категориальной диаграммы
coplot(Weil ~ Wingl | fMonth1 * fSex1, ylab = "Weight (g)",
xlab = "Wing length (mm)",
panel = function(x, y, ...) {
  tmp <- lm(y ~ x, na.action = na.omit)
  abline(tmp)
  points(x, y) })
```



Поскольку линии регрессии на диаграмме не параллельны, это указывает на необходимость включения в статистическую модель соответствующих взаимодействий. На языке R эту модель можно было бы записать в виде `weight ~ length*sex*month`. Подобная запись предполагает наличие в модели эффектов всех трех индивидуальных предикторов, а также всех парных и одного тройного взаимодействия.

Средства визуализации позволяют также наглядно установить, что для некоторых сочетаний "месяц/пол" мы имеем недостаточно репрезентативное число наблюдений (например, в сентябре данные для самцов не были получены вовсе), что может подвергнуть сомнению результаты последующего анализа.

Задание.

Провести разведочный анализ данных по своему варианту (файлы задания находятся в папке «задания» в подпапках с номером варианта):

1. Проанализировать описательную статистику по ключевым атрибутам.
2. Предложить методы анализа выбросов, учитывая особенности данных. Сделать анализ выбросов, удалить выбросы.
3. Выявить наличие или отсутствие коллинеарности между предикторами.
4. Исследовать взаимодействие между предикторами.

Контрольные вопросы

1. Опишите кратко основные методы описательной статистики.
2. Как связаны медиана и квартили?

3. Что такое выбросы?
4. С помощью каких графических методов можно определить выбросы?
5. В чем заключается проблема мультиколлинеарности?
6. Для чего используются матричные диаграммы рассеяния?