

Множественная регрессия

Линейные модели...

Марина Варфоломеева, Вадим Хайтов

Кафедра Зоологии беспозвоночных, Биологический факультет, СПбГУ



Множественная регрессия

- ▶ Техника подгонки множественных регрессионных моделей
- ▶ Проверка условий применимости множественных регрессионных моделей

Вы сможете

- ▶ Подобрать множественную линейную модель
- ▶ Протестировать ее статистическую значимость и валидность



Пример: Птицы в лесах Австралии

Фрагментация лесных местообитаний - одна из важнейших проблем Австралии.

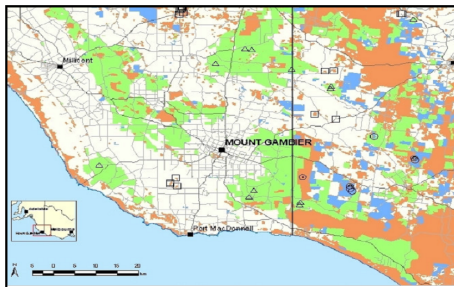
Вопрос: от каких факторов зависит обилие птиц во фрагментированных лесных массивах? (Loyn, 1987)

Зависимая переменная

- ▶ ABUND - Обилие птиц на стандартном маршруте

Предикторы

- ▶ AREA - площадь лесного массива (Га)
- ▶ YRISOL - год, в котором произошла изоляция лесного массива
- ▶ DIST - расстояние до ближайшего лесного массива (км)
- ▶ LDIST - расстояние до ближайшего более крупного массива (км)
- ▶ GRAZE - качественная оценка уровня выпаса скота (1 - низкий уровень, 5 - высокий уровень)
- ▶ ALT - высота над уровнем моря



Скачиваем данные

Не забудьте войти в вашу директорию для матметодов при помощи `setwd()`

```
# Данные можно загрузить с сайта  
library(downloader)  
# в рабочем каталоге создаем суб-директорию для данных  
if(!dir.exists("data")) dir.create("data")  
# скачиваем файл  
download(  
  url = "https://varmara.github.io/linmodr-course/data/loyn.csv",  
  destfile = "data/loyn.csv")
```



Читаем данные

```
bird <- read.csv("data/loyn.csv")
```

Проверяем, все ли правильно открылось

```
str(bird)
```

```
# 'data.frame': 56 obs. of 7 variables:
# $ ABUND : num 5.3 2 1.5 17.1 13.8 14.1 3.8 2.2 3.3 3 ...
# $ AREA : num 0.1 0.5 0.5 1 1 1 1 1 1 1 ...
# $ YRISOL: int 1968 1920 1900 1966 1918 1965 1955 1920 1965 1900 ...
# $ DIST : int 39 234 104 66 246 234 467 284 156 311 ...
# $ LDIST : int 39 234 311 66 246 285 467 1829 156 571 ...
# $ GRAZE : int 2 5 5 3 5 3 5 5 4 5 ...
# $ ALT : int 160 60 140 160 140 130 90 60 130 130 ...
```

Есть ли пропущенные значения?

```
sapply(bird, function(x)sum(is.na(x)))
```

Можно ли ответить на вопрос таким методом?

```
cor(bird)
```

#	ABUND	AREA	YRISOL	DIST	LDIST	GRAZE	ALT
# ABUND	1.0000	0.25597	0.50336	0.236	0.0872	-0.683	0.386
# AREA	0.2560	1.00000	-0.00149	0.108	0.0346	-0.310	0.388
# YRISOL	0.5034	-0.00149	1.00000	0.113	-0.0833	-0.636	0.233
# DIST	0.2361	0.10834	0.11322	1.000	0.3172	-0.256	-0.110
# LDIST	0.0872	0.03458	-0.08332	0.317	1.0000	-0.028	-0.306
# GRAZE	-0.6825	-0.31040	-0.63557	-0.256	-0.0280	1.000	-0.407
# ALT	0.3858	0.38775	0.23272	-0.110	-0.3060	-0.407	1.000



Можно ли ответить на вопрос таким методом?

```
cor(bird)
```

#	ABUND	AREA	YRISOL	DIST	LDIST	GRAZE	ALT
# ABUND	1.0000	0.25597	0.50336	0.236	0.0872	-0.683	0.386
# AREA	0.2560	1.00000	-0.00149	0.108	0.0346	-0.310	0.388
# YRISOL	0.5034	-0.00149	1.00000	0.113	-0.0833	-0.636	0.233
# DIST	0.2361	0.10834	0.11322	1.000	0.3172	-0.256	-0.110
# LDIST	0.0872	0.03458	-0.08332	0.317	1.0000	-0.028	-0.306
# GRAZE	-0.6825	-0.31040	-0.63557	-0.256	-0.0280	1.000	-0.407
# ALT	0.3858	0.38775	0.23272	-0.110	-0.3060	-0.407	1.000

Нет

- ▶ Обычная корреляция не учитывает, что взаимосвязь между переменными может находиться под контролем других переменных и их взаимодействий.
- ▶ Множественные тесты. При тестировании значимости множества коэффициентов корреляции нужно вводить поправку для уровня значимости. Лучше было бы учесть все в одном анализе.



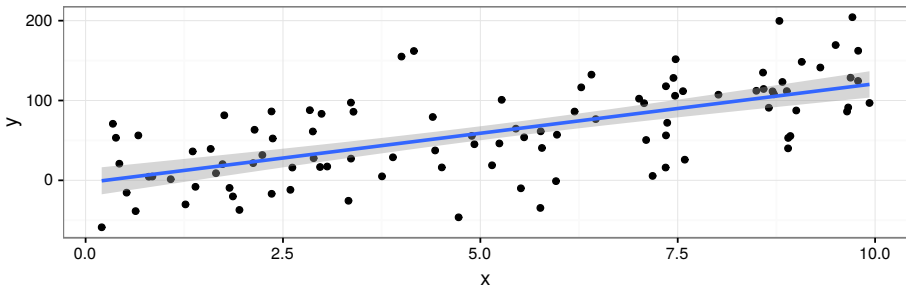
Нам предстоит построить множественную регрессионную модель

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_p x_{ip} + \varepsilon_i$$

- ▶ y_i - значение зависимой переменной для i -того наблюдения
- ▶ β_0 - свободный член (intercept). Значение Y при $X_1 = X_2 = X_3 = \dots = X_p = 0$
- ▶ β_1 - частный угловой коэффициент для зависимости Y от X_1 . Показывает насколько единиц изменяется Y при изменении X_1 на одну единицу и при условии, что все остальные предикторы не изменяются.
 $\beta_2, \beta_3, \dots, \beta_p$ - аналогично
- ▶ ε_i - варьирование Y , не объясняемое данной моделью

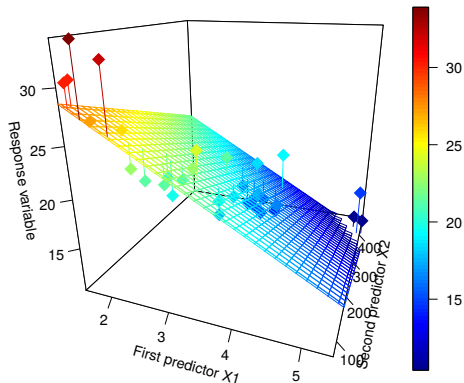
Геометрическая интерпретация множественной линейной модели

Для случая с одним предиктором $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ — линия регрессии



Геометрическая интерпретация множественной линейной модели

Для случая с двумя предикторами $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ — плоскость в трехмерном пространстве



Геометрическая интерпретация множественной линейной модели

Для случая с большим количеством предикторов

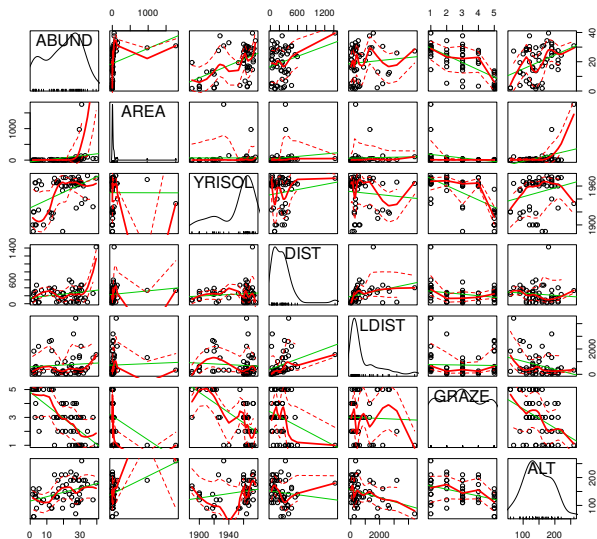
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_p x_{ip} + \varepsilon_i$$

Плоскость в n -мерном пространстве, оси которого образованы значениями предикторов

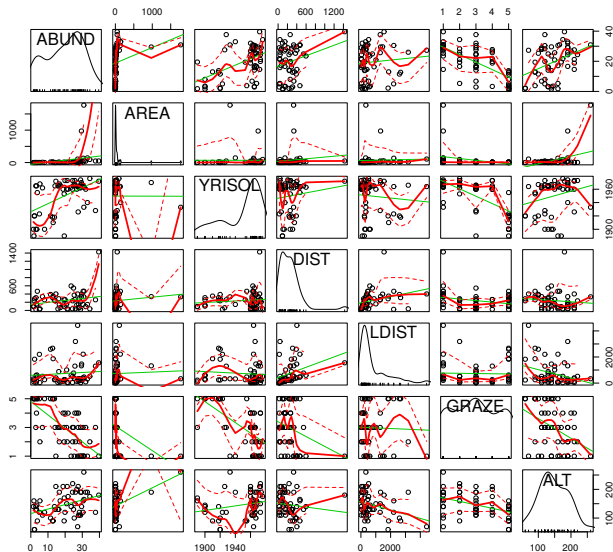


Исследование данных (Data Exploration)

```
library(car)  
scatterplotMatrix(bird)
```



Явные проблемы — есть сильные корреляции между некоторым предикторами



Задание

- ▶ Постройте множественную линейную регрессию для зависимости обилия птиц (ABUND) от других переменных (AREA, YRISOL, DIST, LDIST, GRAZE, ALT)

Решение

```
mod1 <- lm(ABUND ~ AREA + YRISOL + DIST + LDIST + GRAZE + ALT, data = bird)
mod1 <- lm(ABUND ~ ., data = bird) # то же самое
```

```
summary(mod1)
```

```
#
# Call:
# lm(formula = ABUND ~ ., data = bird)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -17.664  -4.641  -0.088   4.286  20.104
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) -109.696233   113.349038  -0.97   0.3379
# AREA         0.000887    0.004657   0.19   0.8498
# YRISOL       0.066928    0.056843   1.18   0.2447
# DIST        0.003811    0.005418   0.70   0.4851
# LDIST       0.001418    0.001310   1.08   0.2845
# GRAZE      -3.446640    1.106683  -3.11   0.0031 **
# ALT         0.047722    0.030888   1.54   0.1288
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 7.95 on 49 degrees of freedom
# Multiple R-squared:  0.512, Adjusted R-squared:  0.452
# F-statistic: 8.56 on 6 and 49 DF, p-value: 0.00000224
```



Проверка валидности модели

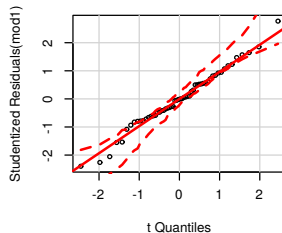
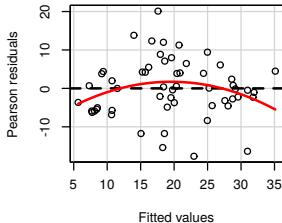
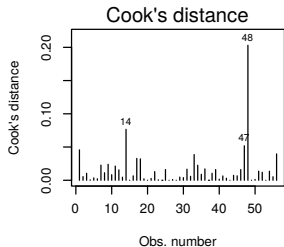
Вспомним условия применимости линейных моделей

- ▶ Линейная связь между зависимой переменной (Y) и предикторами (X)
- ▶ Независимость значений Y друг от друга
- ▶ Нормальное распределение Y для каждого уровня значений X
- ▶ Гомогенность дисперсий Y для каждого уровня значений X
- ▶ Отсутствие коллинеарности предикторов (для множественной регрессии)

- ▶ Проверьте условия применимости модели обилия птиц

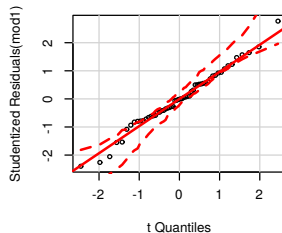
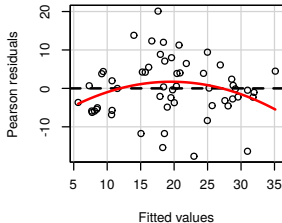
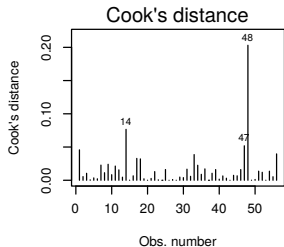
Решение средствами базовой графики

```
op <- par(mfrow = c(1, 3))  
plot(mod1, which = 4)  
residualPlot(mod1)  
qqPlot(mod1)  
par(op)
```



Решение средствами базовой графики

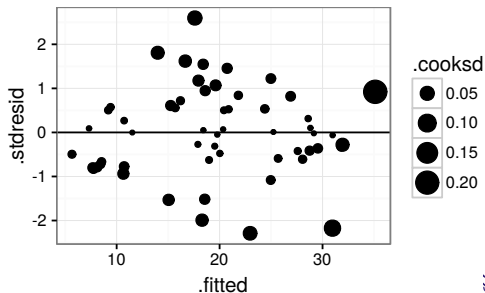
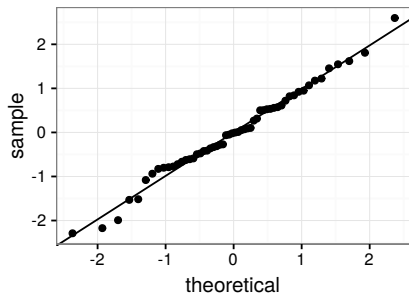
```
op <- par(mfrow = c(1, 3))  
plot(mod1, which = 4)  
residualPlot(mod1)  
qqPlot(mod1)  
par(op)
```



- ▶ Выбросов нет
- ▶ Гетерогенность дисперсий
- ▶ Отклонения от нормального распределения?

Решение в ggplot2

```
bird_diag <- fortify(mod1)
# квантильный график
mean_val <- mean(bird_diag$.stdresid)
sd_val <- sd(bird_diag$.stdresid)
gg_qq <- ggplot(bird_diag, aes(sample = .stdresid)) + geom_point(stat = "qq")
# остатки и расстояние Кука
gg_res <- ggplot(data = bird_diag, aes(x = .fitted, y = .stdresid, size = .cooksdi))
# вместе
library(gridExtra)
grid.arrange(gg_qq, gg_res, nrow = 1, widths = c(0.45, 0.55))
```



Мультиколлинеарность

Мультиколлинеарность

Мультиколлинеарность — наличие линейной зависимости между независимыми переменными (факторами) регрессионной модели.

При наличии мультиколлинеарности оценки параметров получаются неточными, а значит сложно будет дать интерпретацию влияния предикторов на отклик

Косвенные признаки мультиколлинеарности:

- ▶ Большие ошибки оценок параметров
- ▶ Большинство параметров модели недостоверно отличаются от нуля, но F критерий говорит, что вся модель значима

Проверка на мультиколлинеарность

- ▶ Фактор инфляции дисперсии (Variance inflation factor, VIF)

Как рассчитывается VIF

Мы должны оценить какую долю изменчивости конкретного предиктора могут объяснить другие предикторы (т.е. насколько предикторы независимы)

Для каждого предиктора:

1. Строим регрессионную модель данного предиктора от всех остальных

$$x_1 = c_0 + c_2x_2 + c_3x_3 + \dots + c_px_p$$

2. Находим R^2 модели
3. Вычисляем фактор инфляции дисперсии

$$VIF = \frac{1}{1 - R^2}$$



Что делать, если мультиколлинеарность выявлена?

- ▶ Можно последовательно удалить из модели избыточные предикторы с $VIF > 3$ (иногда $VIF > 2$)
 1. подбираем модель
 2. считаем VIF
 3. удаляем предиктор с самым большим VIF
 4. повторяем 1-3
- ▶ Можно заменить исходные предикторы новыми независимыми друг от друга переменными, полученными с помощью метода главных компонент

Проверяем отсутствие мультиколлинеарности

Функция `vif()` из пакета `car`

```
vif(mod1)
```

#	AREA	YRISOL	DIST	LDIST	GRAZE	ALT
#	1.34	1.84	1.23	1.26	2.31	1.57

Проверяем отсутствие мультиколлинеарности

Функция `vif()` из пакета `car`

```
vif(mod1)
```

#	AREA	YRISOL	DIST	LDIST	GRAZE	ALT
#	1.34	1.84	1.23	1.26	2.31	1.57

В нашей модели явной мультиколлинеарности нет

Однако, возможно, что `GRAZE` - избыточный предиктор

Удалим из модели избыточный предиктор

```
mod2 <- update(mod1, ~ . -GRAZE)  
vif(mod2)
```

```
#   AREA YRISOL   DIST  LDIST    ALT  
#   1.25   1.10   1.16   1.24   1.43
```

Удалим из модели избыточный предиктор

```
mod2 <- update(mod1, ~ . -GRAZE)  
vif(mod2)
```

```
#   AREA YRISOL   DIST  LDIST    ALT  
#   1.25   1.10   1.16   1.24   1.43
```

Теперь мультиколлинеарности нет

В этой модели осталось много незначимых предикторов

```
coef(summary(mod2))
```

#	Estimate	Std. Error	t value	Pr(> t)
# (Intercept)	-344.79073	91.61744	-3.763	0.000441
# AREA	0.00459	0.00488	0.941	0.351324
# YRISOL	0.17928	0.04760	3.767	0.000437
# DIST	0.00772	0.00571	1.352	0.182332
# LDIST	0.00192	0.00141	1.360	0.179784
# ALT	0.07638	0.03195	2.391	0.020622

Что дальше?

Два варианта действий:

- ▶ Оставить все как есть. Если значение коэффициента при предикторе не значимо отличается от нуля, значит, этот предиктор не влияет на обилие птиц
- ▶ Провести пошаговый подбор оптимальной модели (Об этом на следующей лекции)

Сегодня мы оставим все как есть и попытаемся выяснить, какие предикторы влияют сильнее всего



Какой из предикторов оказывает наиболее сильное влияние?

Для ответа на этот вопрос надо “уравнять” шкалы, всех предикторов, то есть стандартизовать их.

Коэффициенты при стандартизованных предикторах покажут, насколько сильно меняется отклик при изменении предиктора на одно стандартное отклонение.

Для стандартизации используем функцию `scale()`

```
mod2_scaled <- lm(ABUND ~ scale(AREA) + scale(YRISOL) + scale(DIST) + scale(L  
  scale(ALT), data = bird)
```



Какой из предиктов оказывает наиболее сильное влияние на усилие дыхательных мышц?

```
coef(summary(mod2_scaled))
```

#	Estimate	Std. Error	t value	Pr(> t)
# (Intercept)	19.51	1.15	16.960	2.22e-22
# scale(AREA)	1.22	1.30	0.941	3.51e-01
# scale(YRISOL)	4.59	1.22	3.767	4.37e-04
# scale(DIST)	1.69	1.25	1.352	1.82e-01
# scale(LDIST)	1.76	1.29	1.360	1.80e-01
# scale(ALT)	3.32	1.39	2.391	2.06e-02

Какой из предиктов оказывает наиболее сильное влияние на усилие дыхательных мышц?

```
coef(summary(mod2_scaled))
```

#	Estimate	Std. Error	t value	Pr(> t)
# (Intercept)	19.51	1.15	16.960	2.22e-22
# scale(AREA)	1.22	1.30	0.941	3.51e-01
# scale(YRISOL)	4.59	1.22	3.767	4.37e-04
# scale(DIST)	1.69	1.25	1.352	1.82e-01
# scale(LDIST)	1.76	1.29	1.360	1.80e-01
# scale(ALT)	3.32	1.39	2.391	2.06e-02

- ▶ Сильнее всего на обилие птиц влияют продолжительность изоляции и высота, на которой расположен лес
- ▶ При изменении продолжительности изоляции на 1 стандартное отклонение, обилие птиц изменяется на 4.59
- ▶ При изменении высоты на 1 стандартное отклонение, обилие птиц изменяется на 3.32

Задание

Постройте модель описывающую связь между усилием мышц, осуществляющих выдох (pmax) и следующими переменными:

- ▶ age - Возраст
- ▶ sex - Пол (0: male, 1:female)
- ▶ height - Рост (cm)
- ▶ weight - Вес (kg)
- ▶ bmp - Отклонения в весе от нормы (% of normal)
- ▶ fev1 - Объем наполненных легких
- ▶ rv - Остаточный объем легких
- ▶ frc - Функциональная остаточная емкость легких
- ▶ tlc - Общая емкость легких

Исключите из модели коллинеарные предикторы.

Для получения данных выполните следующий код:

```
library(ISwR)  
data(cystfibr)
```

Решение

```
M1 <- lm(pemax ~ ., data = cystfibr)
summary(M1)
```

```
#
# Call:
# lm(formula = pemax ~ ., data = cystfibr)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -37.34 -11.53   1.08  13.39  33.41
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  176.058    225.891   0.78    0.45
# age          -2.542     4.802   -0.53    0.60
# sex          -3.737    15.460   -0.24    0.81
# height       -0.446     0.903   -0.49    0.63
# weight        2.993     2.008    1.49    0.16
# bmp          -1.745     1.155   -1.51    0.15
# fev1          1.081     1.081    1.00    0.33
# rv            0.197     0.196    1.00    0.33
# frc          -0.308     0.492   -0.63    0.54
# tlc           0.189     0.500    0.38    0.71
#
# Residual standard error: 25.5 on 15 degrees of freedom
# Multiple R-squared:  0.637,    Adjusted R-squared:  0.42
# F-statistic: 2.93 on 9 and 15 DF,  p-value: 0.032
```



Проверяем на коллинеарность

```
vif(M1)
```

```
#   age    sex height weight    bmp   fev1    rv   frc    tlc  
# 21.83  2.27 13.95  47.78   7.12   5.42 10.54 17.14  2.66
```

Удаляем избыточные предикторы

```
M2 <- update(M1, .~.-weight)  
vif(M2)
```

```
#   age    sex height    bmp   fev1    rv   frc    tlc  
#  8.10  2.03   7.60   2.73   4.21 10.33 15.81  2.18
```

Продолжаем удалять избыточные предикторы

```
M3 <- update(M2, . ~. - frc)
vif(M3)
```

```
#   age    sex height    bmp   fev1    rv   tlc
#  7.34  1.61   7.60   1.79   2.87   2.84  1.77
```

Продолжаем удалять избыточные предикторы

```
M4 <- update(M3, . ~. - height)
vif(M4)
```

```
# age sex bmp fev1  rv  tlc
# 1.61 1.61 1.72 2.86 2.81 1.77
```

Наконец-то коллинеарности нет



Смотрим на полученную модель

summary(M4)

```
#
# Call:
# lm(formula = pemax ~ age + sex + bmp + fev1 + rv + tlc, data = cystfibr)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -50.45 -19.11   3.97   17.40  31.33
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  -83.525     82.808   -1.01   0.3265
# age           5.038      1.296    3.89   0.0011 **
# sex           4.991     12.913    0.39   0.7037
# bmp          -0.403      0.564   -0.71   0.4840
# fev1          1.931      0.780    2.48   0.0234 *
# rv            0.114      0.101    1.13   0.2745
# tlc           0.465      0.405    1.15   0.2654
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 25.3 on 18 degrees of freedom
# Multiple R-squared:  0.571,    Adjusted R-squared:  0.428
# F-statistic: 3.99 on 6 and 18 DF,  p-value: 0.0103
```



Взаимодействия предикторов

Взаимодействие предикторов

В регрессионные модели можно включать не только предикторы сами по себе, но и их взаимодействия.

Взаимодействие предикторов показывает, что угловой коэффициент одного предиктора зависит от значения другого предиктора.

Переменные, участвующие во взаимодействиях, иногда называют переменными-модераторами — они регулируют связь между предиктором и откликом.

Модели, в которых есть достоверное взаимодействие непрерывных предикторов сложно интерпретировать.

Для визуализации взаимодействия можно построить график отклика при нескольких значениях предиктора (например, при среднем значении модератора или $\pm 1 \cdot SD$)



Вернемся к данным по обилию птиц и построим модель для двух предикторов

Формулу модели со взаимодействием можно записать двумя способами

- ▶ $ABUND \sim YRISOL + GRAZE + YRISOL:GRAZE$
- ▶ $ABUND \sim YRISOL * GRAZE$

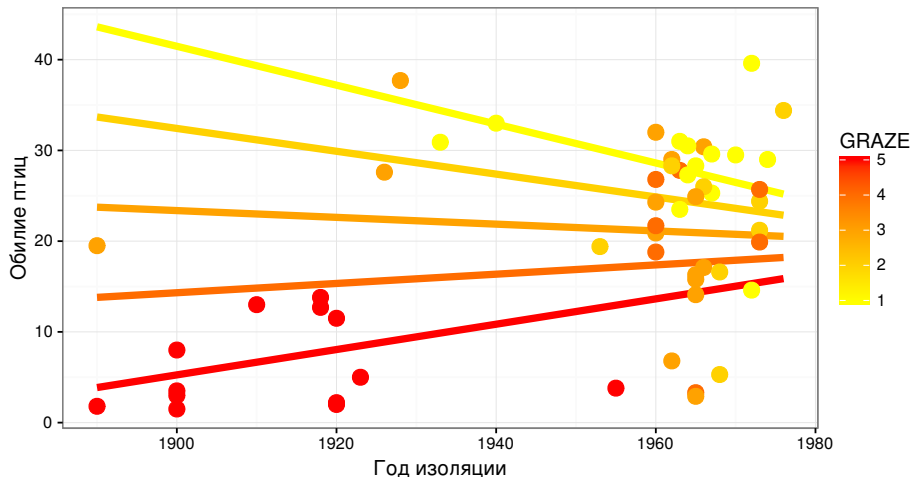
```
mod3 <- lm(ABUND ~ YRISOL * GRAZE, data = bird)
summary(mod3)
```

```
#
# Call:
# lm(formula = ABUND ~ YRISOL * GRAZE, data = bird)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -18.564  -4.300   0.961   4.050  15.378
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)   625.8955    304.2770     2.06   0.045 *
# YRISOL         -0.3028     0.1552    -1.95   0.056 .
# GRAZE        -177.1773    71.8370    -2.47   0.017 *
# YRISOL:GRAZE    0.0885     0.0368     2.40   0.020 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 7.6 on 52 degrees of freedom
# Multiple R-squared:  0.527,    Adjusted R-squared:  0.499
# F-statistic: 19.3 on 3 and 52 DF,  p-value: 0.0000000155
```



Как интерпретировать взаимодействия?

График отклика при нескольких значениях предиктора



НО! Не всегда все так просто трактуется...

Часто трактовка взаимодействий затруднительна, особенно если предикторов много.



График на предыдущем слайде получен с помощью кода

```
MyData <- expand.grid(YRISOL = seq(1890, 1976, 1),  
                     GRAZE = seq(1, 5, 1))  
MyData$Predicted <- predict(mod3, newdata = MyData)  
ggplot(MyData, aes(x = YRISOL, y = Predicted, group = GRAZE)) +  
  geom_line(aes(color = GRAZE), size = 2) +  
  geom_point(data = bird, aes(x = YRISOL, y = ABUND, color = GRAZE), size = 4)  
  scale_colour_continuous(low = "yellow", high = "red") +  
  xlab("Год изоляции") + ylab("Обилие птиц")
```



Include or Don't include? That is the question...

Вопрос о включении в модель взаимодействия предикторов совсем непростой

Существует несколько подходов:

1. Не включать взаимодействия в модель. Но если при валидации модели в остатках появляется явный паттерн, то это может быть следствием наличия взаимодействия предикторов
2. Основываясь на априорных знаниях свойств объектов включить только те взаимоотношения, которые имеют биологический смысл, либо взаимодействия с наиболее важными переменными (теми, ради которых была затеяна работа)
3. Включать в модель все взаимодействия, потом пошагово выбросить недостоверные (Model selection - на следующей лекции)



Include or Don't include? That is the question...

Вопрос о включении в модель взаимодействия предикторов совсем непростой

Существует несколько подходов:

1. Не включать взаимодействия в модель. Но если при валидации модели в остатках появляется явный паттерн, то это может быть следствием наличия взаимодействия предикторов
2. Основываясь на априорных знаниях свойств объектов включить только те взаимоотношения, которые имеют биологический смысл, либо взаимодействия с наиболее важными переменными (теми, ради которых была затеяна работа)
3. Включать в модель все взаимодействия, потом пошагово выбросить недостоверные (Model selection - на следующей лекции)

Включать в анализ и обсуждать все взаимодействия “дорого” (неудобно):

- ▶ Взаимодействия высоких порядков сложно интерпретировать
- ▶ Каждое взаимодействие — это коэффициент в модели, или несколько, если это взаимодействие с дискретной переменной. Чтобы подобрать модель нужно много данных — по 20-40 наблюдений в расчете на каждый коэффициент.



Summary

- ▶ При построении множественной регрессии важно, помимо других условий, проверить модель на наличие мультиколлинеарности
- ▶ Если модель построена на основе стандартизированных значений предикторов, то можно сравнивать влияние этих предикторов
- ▶ В модель можно (а иногда и нужно) включать взаимодействия предикторов



- ▶ Кабаков Р.И. R в действии. Анализ и визуализация данных на языке R. М.: ДМК Пресс, 2014.
- ▶ Quinn G.P., Keough M.J. (2002) Experimental design and data analysis for biologists, pp. 92-98, 111-130
- ▶ Diez D. M., Barr C. D., Cetinkaya-Rundel M. (2014) Open Intro to Statistics., pp. 354-367.
- ▶ Logan M. (2010) Biostatistical Design and Analysis Using R. A Practical Guide, pp. 170-173, 208-211
- ▶ Zuur, A.F. et al. 2009. Mixed effects models and extensions in ecology with R. - Statistics for biology and health. Springer, New York, NY. pp. 538-552.