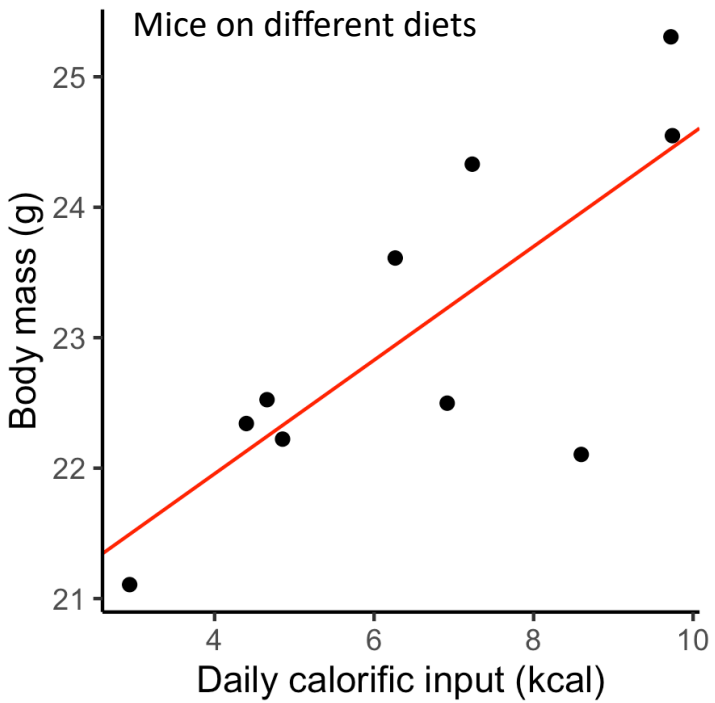
13. Linear models

"All models are wrong, but some are useful"

George Box

Simple quantitative model

	Daily calorific input (kcal)	Body mass (g)
1	8.6	22.1
2	2.9	21.1
3	4.4	22.3
4	4.9	22.2
5	9.7	25.3
6	9.7	24.5
7	6.3	23.6
8	4.7	22.5
9	7.2	24.3
10	6.9	22.5



Best-fitting linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Simple quantitative model

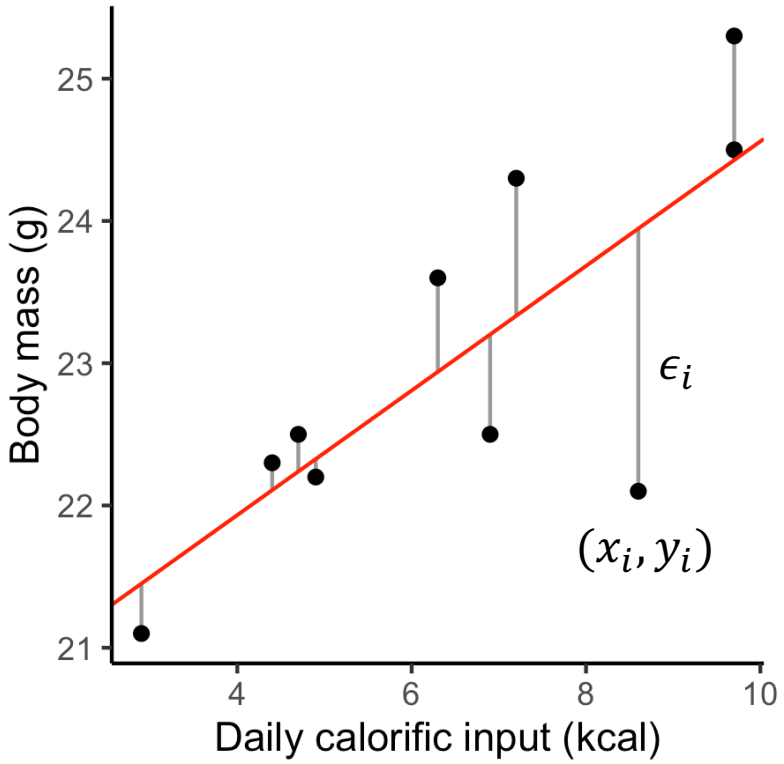
Coefficients

Residual (noise)

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Response

Predictor



We find coefficients by minimising

$$\sum_{i=1}^{n} \epsilon_i^2$$

Linear model in R

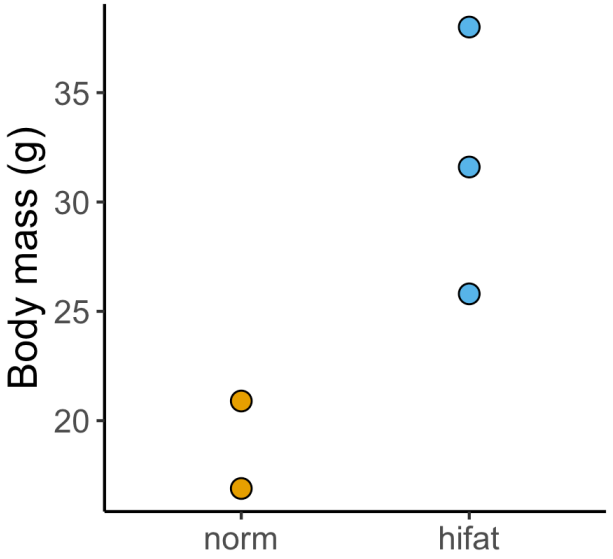
```
> ms <- data.frame(
    kcal = c(8.6, 2.9, 4.4, 4.9, 9.7, 9.7, 6.3, 4.7, 7.2, 6.9),
    mass = c(22.1, 21.1, 22.3, 22.2, 25.3, 24.5, 23.6, 22.5, 24.3, 22.5)
  )
```

```
> f <- lm(mass ~ kcal, data=ms)
> summary(f)

Call:
lm(formula = mass ~ kcal, data = ms)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8462 -0.2947  0.1323  0.5608  0.9667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.1813     0.8750  23.065 1.33e-08 ***
kcal          0.4378     0.1269   3.449  0.00871 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8862 on 8 degrees of freedom
Multiple R-squared:  0.5979,         Adjusted R-squared:  0.5476
F-statistic:  11.9 on 1 and 8 DF,  p-value: 0.008709
```

intercept: $\beta_0 = 20.2 \pm 0.9$
slope: $\beta_1 = 0.4 \pm 0.1$

Categorical (qualitative) predictor

	Body mass (g)	Diet
1	16.8	norm
2	20.9	norm
3	25.8	hifat
4	38.0	hifat
5	31.6	hifat



Linear model

$$y_1 = \mu_1 + \epsilon_1$$
$$y_2 = \mu_1 + \epsilon_2$$
$$y_3 = \mu_2 + \epsilon_3$$
$$y_4 = \mu_2 + \epsilon_4$$
$$y_5 = \mu_2 + \epsilon_5$$

μ_1, μ_2 - unknown true population means for diet and hifat

Linear model for qualitative predictors

$$
\begin{aligned}
y_1 &= \mu_1 + \epsilon_1 \\
y_2 &= \mu_1 + \epsilon_2 \\
y_3 &= \mu_2 + \epsilon_3 \\
y_4 &= \mu_2 + \epsilon_4 \\
y_5 &= \mu_2 + \epsilon_5
\end{aligned}
\quad \Rightarrow \quad
\begin{aligned}
y_1 &= 1\mu_1 + 0\mu_2 + \epsilon_1 \\
y_2 &= 1\mu_1 + 0\mu_2 + \epsilon_2 \\
y_3 &= 0\mu_1 + 1\mu_2 + \epsilon_3 \\
y_4 &= 0\mu_1 + 1\mu_2 + \epsilon_4 \\
y_5 &= 0\mu_1 + 1\mu_2 + \epsilon_5
\end{aligned}
\quad \Rightarrow \quad
\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{pmatrix}
$$

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\mu} + \boldsymbol{\epsilon}$$

Design matrix

$$\mathbf{W} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

Contrasts

$\beta_0 = \mu_1$
$\beta_1 = \mu_2 - \mu_1$

$$y_1 = \mu_1 + \epsilon_1 \quad y_1 = \beta_0 + \epsilon_1$$
$$y_2 = \mu_1 + \epsilon_2 \quad y_2 = \beta_0 + \epsilon_2$$
$$y_3 = \mu_2 + \epsilon_3 \Rightarrow y_3 = \beta_0 + \beta_1 + \epsilon_3$$
$$y_4 = \mu_2 + \epsilon_4 \quad y_4 = \beta_0 + \beta_1 + \epsilon_4$$
$$y_5 = \mu_2 + \epsilon_5 \quad y_5 = \beta_0 + \beta_1 + \epsilon_5$$

$$\Rightarrow \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Design matrix

```
> mdat <- data.frame(
    mass = c(16.9, 20.9, 25.8, 38, 31.6),
    diet = c("norm", "norm", "hifat", "hifat", "hifat")
  )
> mdat$diet <- factor(mdat$diet, levels=c("norm", "hifat"))
```

```
> model.matrix(mass ~ diet, data=mdat)
  (Intercept) diethifat
1           1         0
2           1         0
3           1         1
4           1         1
5           1         1
```

```
> model.matrix(mass ~ 0 + diet, data=mdat)
  dietnorm diethifat
1        1         0
2        1         0
3        0         1
4        0         1
5        0         1
```

Linear model fitting in R

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

```
> f1 <- lm(mass ~ diet, data = mdat)
> summary(f1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.900      3.708   5.098   0.0146 *
diethifat     12.900      4.787   2.695   0.0741 .
```

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$$

$\beta_0 = \mu_1 = 19 \pm 4$
$\beta_1 = \mu_2 - \mu_1 = 13 \pm 5$

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\mu} + \boldsymbol{\epsilon}$$

```
> f2 <- lm(mass ~ 0 + diet, data = mdat)
> summary(f2)
          Estimate Std. Error t value Pr(>|t|)
dietnorm    18.900      3.708   5.098  0.01460 *
diethifat   31.800      3.027  10.504  0.00184 **
```

$$\mathbf{W} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

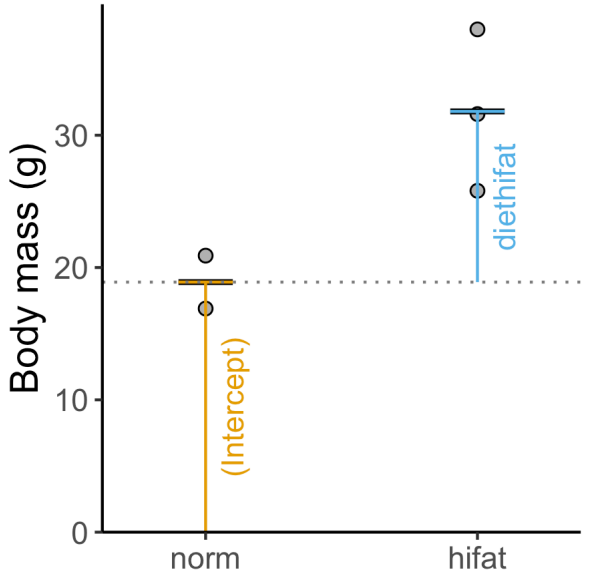
$\mu_1 = 19 \pm 4$
$\mu_2 = 32 \pm 3$

The meaning of coefficients

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

```
> f1 <- lm(mass ~ diet, data = mdat)
> summary(f1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.900      3.708   5.098   0.0146 *
diethifat     12.900      4.787   2.695   0.0741 .
```

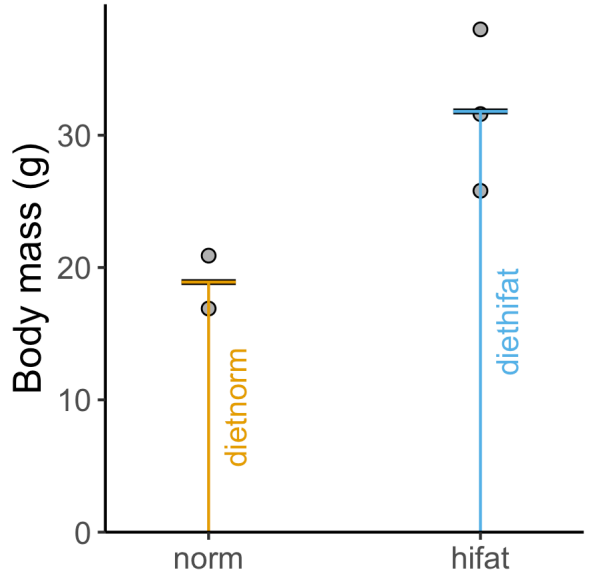
$\beta_0 = \mu_1 = 19 \pm 4$
$\beta_1 = \mu_2 - \mu_1 = 13 \pm 5$



$$\mathbf{Y} = \mathbf{W}\boldsymbol{\mu} + \boldsymbol{\epsilon}$$

```
> f2 <- lm(mass ~ 0 + diet, data = mdat)
> summary(f2)$coefficients
           Estimate Std. Error t value Pr(>|t|)
dietnorm     18.900      3.708   5.098  0.01460 *
diethifat    31.800      3.027  10.504  0.00184 **
```

$\mu_1 = 19 \pm 4$
$\mu_2 = 32 \pm 3$



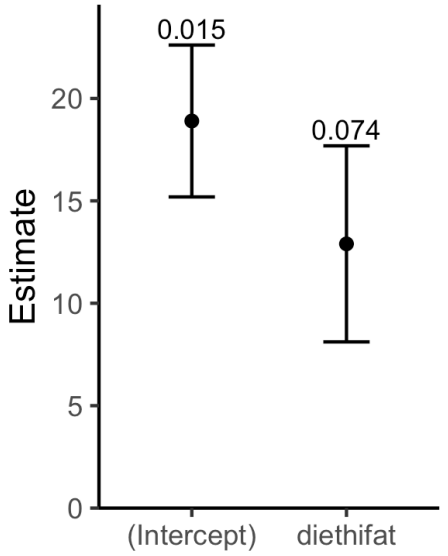
The meaning of p-values

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

```
> f1 <- lm(mass ~ diet, data = mdat)
> summary(f1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.900      3.708   5.098   0.0146 *
diethifat     12.900      4.787   2.695   0.0741 .
```

H_0: effect size is equal zero

dietfat = difference between normal and high-fat diet

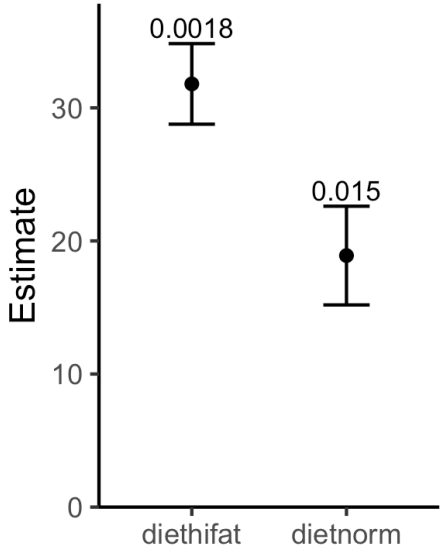


$$\mathbf{Y} = \mathbf{W}\boldsymbol{\mu} + \boldsymbol{\epsilon}$$

```
> f2 <- lm(mass ~ 0 + diet, data = mdat)
> summary(f2)$coefficients
           Estimate Std. Error t value Pr(>|t|)
dietnorm     18.900      3.708   5.098  0.01460 *
diethifat    31.800      3.027  10.504  0.00184 **
```

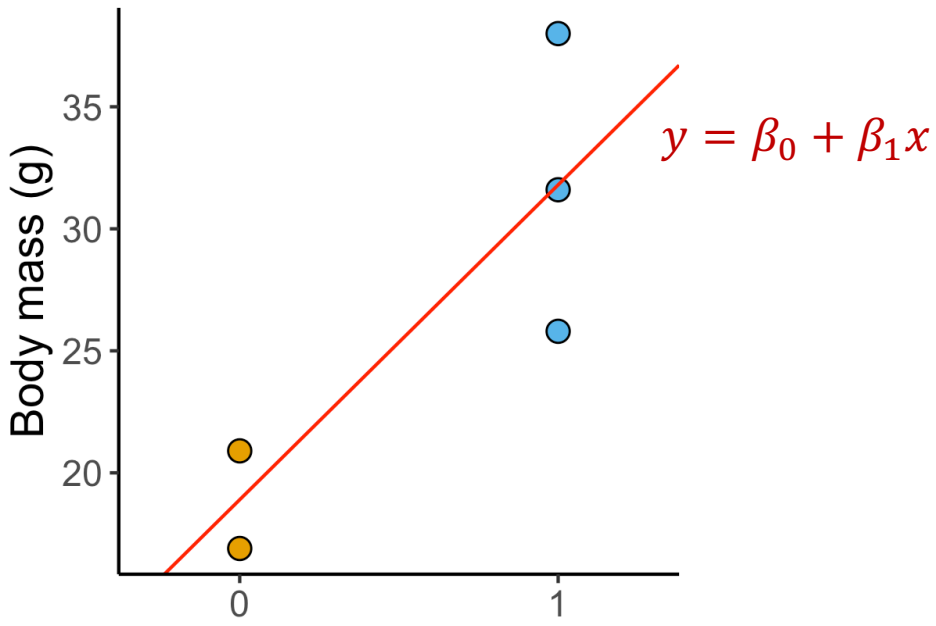
H_0: effect size is equal zero

dietfat = mean of the high-fat diet



It is a linear model

intercept: $\beta_0 = 19 \pm 4$
slope $\beta_1 = 13 \pm 5$



Warning: it does not matter what happens at any x other than 0 and 1.

It is a t-test

```
> t.test(mass ~ diet, data = mdat, var.equal=TRUE)

        Two Sample t-test

data:  mass by diet
t = -2.695, df = 3, p-value = 0.0741
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -28.132955   2.332955
sample estimates:
 mean in group norm mean in group hifat
              18.9                31.8
```

```
           Estimate Std. Error t value Pr(>|t|)
dietnorm     18.900      3.708   5.098  0.01460 *
diethifat    31.800      3.027  10.504  0.00184 **
```

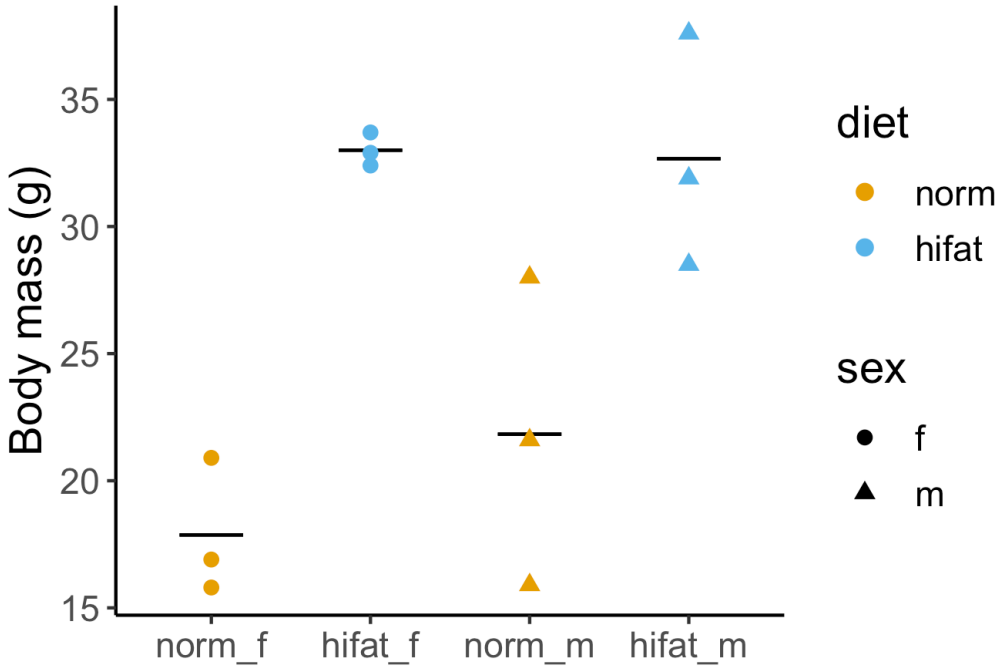
$$\mathbf{Y} = \mathbf{W}\boldsymbol{\mu} + \boldsymbol{\epsilon}$$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    18.900      3.708   5.098   0.0146 *
diethifat      12.900      4.787   2.695   0.0741 .
```

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Multiple variables

	Body mass (g)	Diet	Sex
1	16.9	norm	f
2	20.9	norm	f
3	15.8	norm	f
4	28.0	norm	m
5	21.6	norm	m
6	15.9	norm	m
7	32.4	hifat	f
8	33.7	hifat	f
9	32.9	hifat	f
10	28.5	hifat	m
11	37.6	hifat	m
12	31.9	hifat	m



Linear model with multiple variables

Model formula in R:

```
Y ~ X1 + X2
```

factors
categorical variables

Mathematical representation of R formula:

$$\mathbf{Y} = \beta_0 + \mathbf{X_1}\beta_1 + \mathbf{X_2}\beta_2$$

dummy variables
0 or 1
columns of design matrix

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Design matrix

```
> mds <- data.frame(
    mass = c(16.9, 20.9, 15.8, 28, 21.6, 15.9, 32.4, 33.7, 32.9, 28.5, 37.6, 31.9),
    diet = c(rep("norm", 6), rep("hifat", 6)),
    sex = c("f", "f", "f", "m", "m", "m", "f", "f", "f", "m", "m", "m")
  )
> mds$diet <- relevel(as.factor(mds$diet), "norm")
> mds$sex <- relevel(as.factor(mds$sex), "f")
```

```
> model.matrix(mass ~ diet + sex, data=mds)
   (Intercept) diethifat sexm
1            1         0    0
2            1         0    0
3            1         0    0
4            1         0    1
5            1         0    1
6            1         0    1
7            1         1    0
8            1         1    0
9            1         1    0
10           1         1    1
11           1         1    1
12           1         1    1
```

The baseline is diet = normal, sex = f

Linear model in R

```
> f <- lm(mass ~ diet + sex, data = mds)
> summary(f)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.942      2.005   9.448 5.73e-06 ***
diethifat     12.983      2.315   5.608 0.000331 ***
sexm           1.817      2.315   0.785 0.452780
```

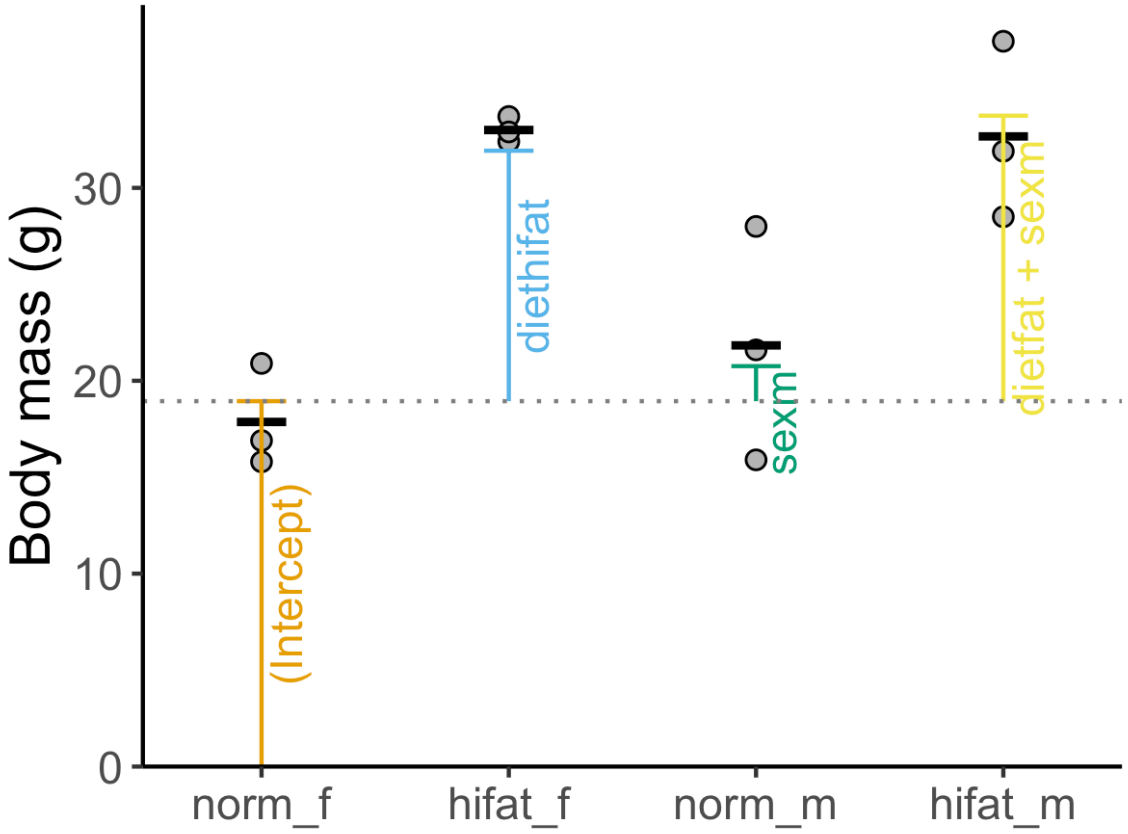
Sex not significant

R notation	Mathematical representation	Value	Description
(Intercept)	β_0	19 ± 2	Mean of the reference (diet = norm, sex = f)
diethifat	β_1	13 ± 2	Effect size for diet = hifat
sexm	β_2	2 ± 2	Effect size for sex = m

Warning: in this model we assume diet and sex are independent. They might not be!

Fit coefficients

R notation	Value	Description
`(Intercept)`	19 ± 2	Mean of the reference (diet = norm, sex = f)
`diethifat`	13 ± 2	Effect size for diet = hifat
`sexm`	2 ± 2	Effect size for sex = m



Model with interactions

```
> model.matrix(mass ~ diet + sex + diet:sex, data=mds)
   (Intercept) diethifat sexm diethifat:sexm
1            1         0    0              0
2            1         0    0              0
3            1         0    0              0
4            1         0    1              0
5            1         0    1              0
6            1         0    1              0
7            1         1    0              0
8            1         1    0              0
9            1         1    0              0
10           1         1    1              1
11           1         1    1              1
12           1         1    1              1
```

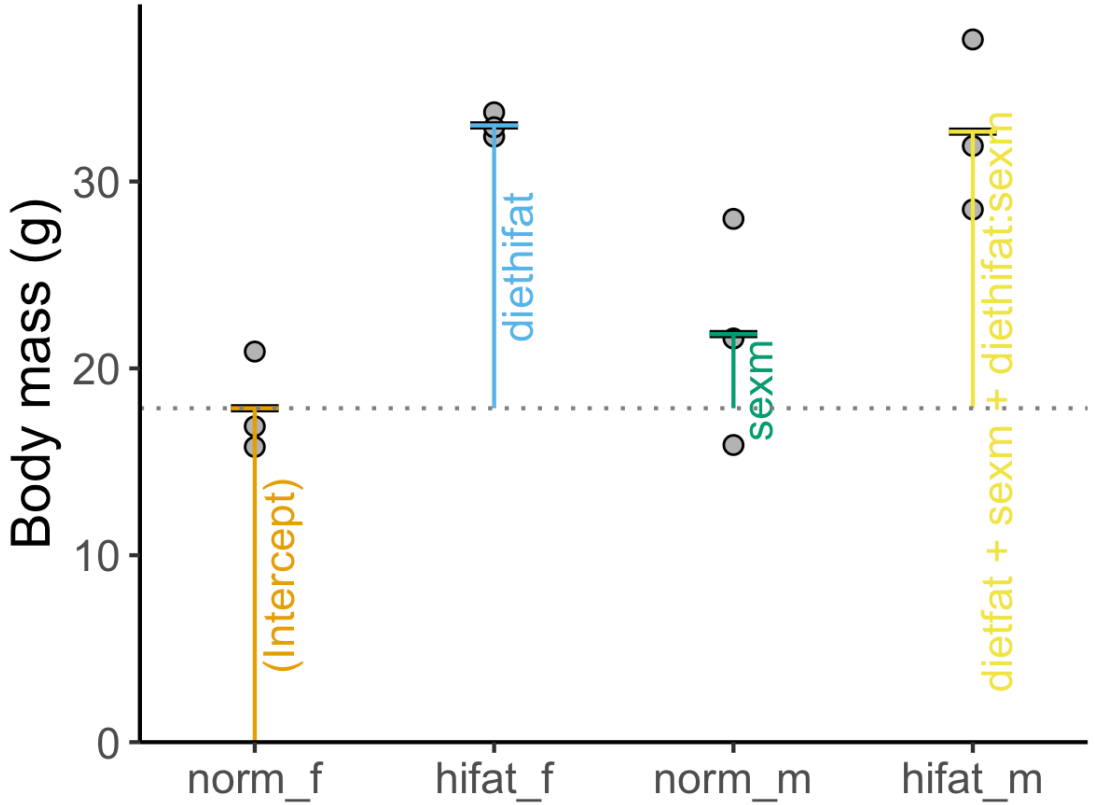
$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_{1:2} x_{1,i}\, x_{2,i} + \epsilon_i$$

Interaction term

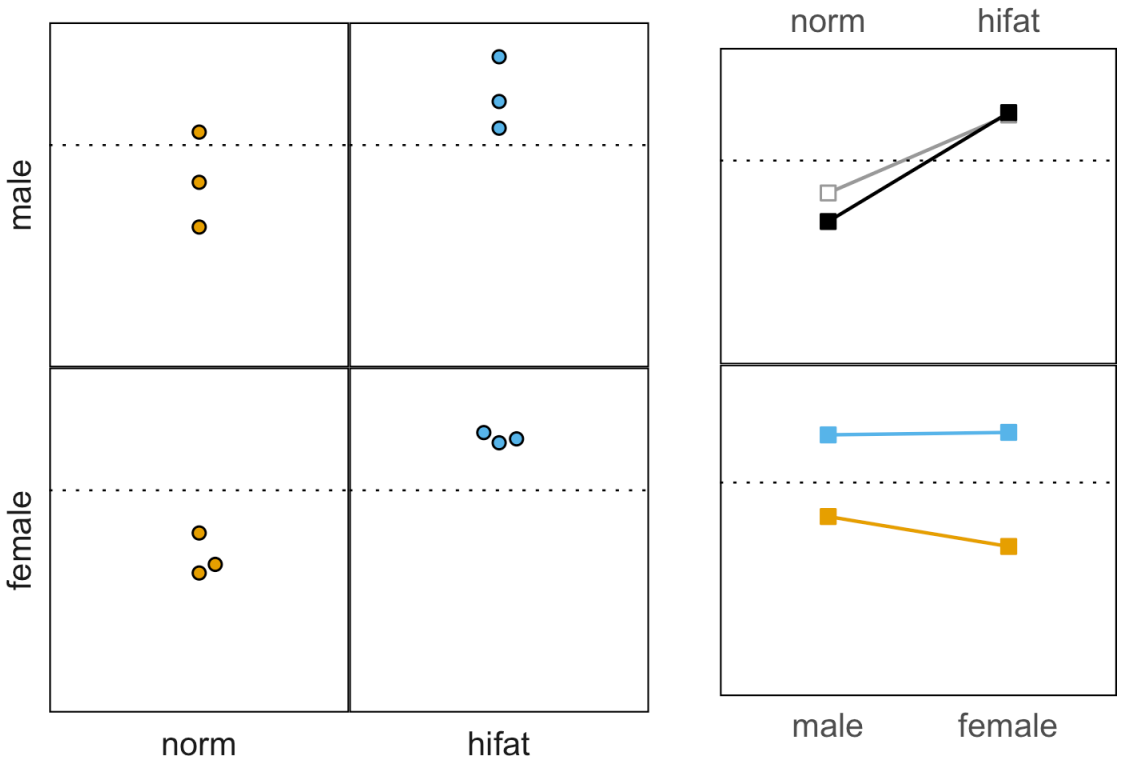
Coefficients with interactions

```
> f <- lm(mass ~ diet + sex + diet:sex, data = mds)
> summary(f)$coefficients
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       17.867      2.335   7.652 6.01e-05 ***
diethifat         15.133      3.302   4.583  0.00179 **
sexm               3.967      3.302   1.201  0.26400
diethifat:sexm    -4.300      4.670  -0.921  0.38407
```

Interaction not significant, we are overfitting



It's ANOVA

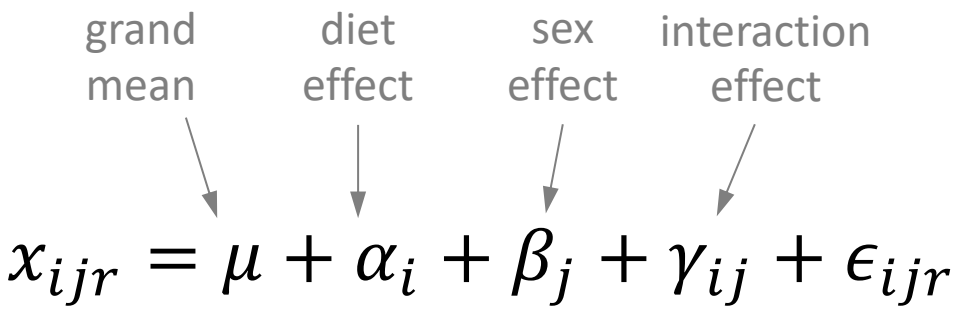


grand mean — column effect — row effect — interaction effect

$$x_{ijr} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijr}$$

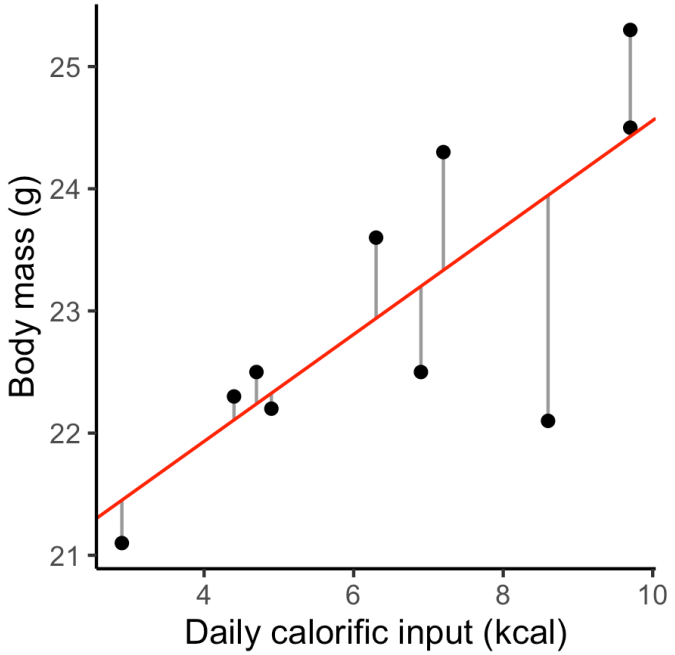
It's ANOVA

```
> anova(f)
Analysis of Variance Table

Response: mass
          Df Sum Sq Mean Sq F value    Pr(>F)
diet       1 505.70  505.70 30.9203 0.0005343 ***
sex        1   9.90    9.90  0.6054 0.4589268
diet:sex   1  13.87   13.87  0.8479 0.3840723
Residuals  8 130.84   16.35
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



R-squared

```
> f <- lm(mass ~ kcal, data=ms)
> summary(f)

Call:
lm(formula = mass ~ kcal, data = ms)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8462 -0.2947  0.1323  0.5608  0.9667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.1813     0.8750  23.065 1.33e-08 ***
kcal          0.4378     0.1269   3.449  0.00871 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8862 on 8 degrees of freedom
Multiple R-squared:  0.5979,         Adjusted R-squared:  0.5476
F-statistic:  11.9 on 1 and 8 DF,  p-value: 0.008709
```



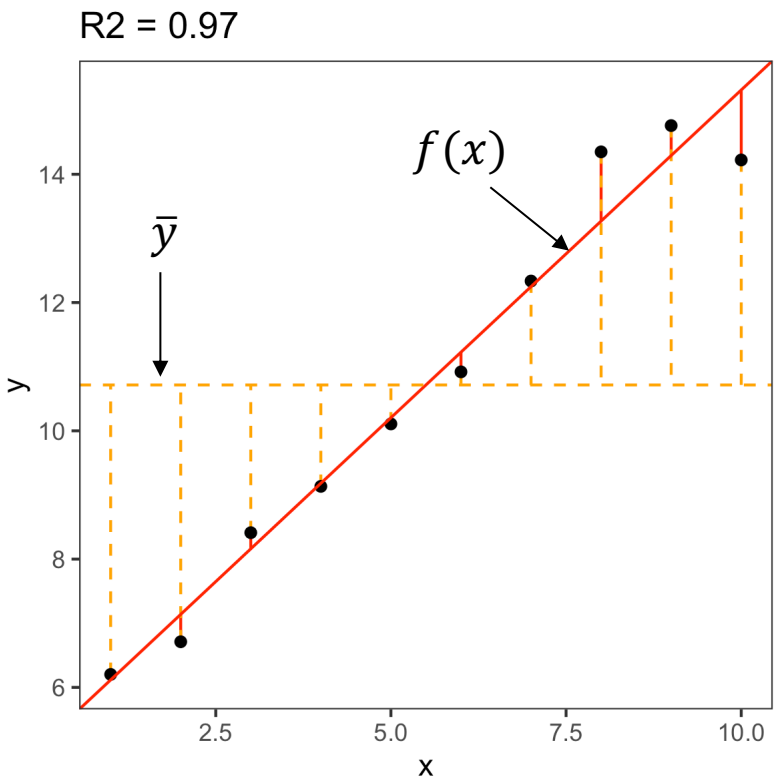
R-squared

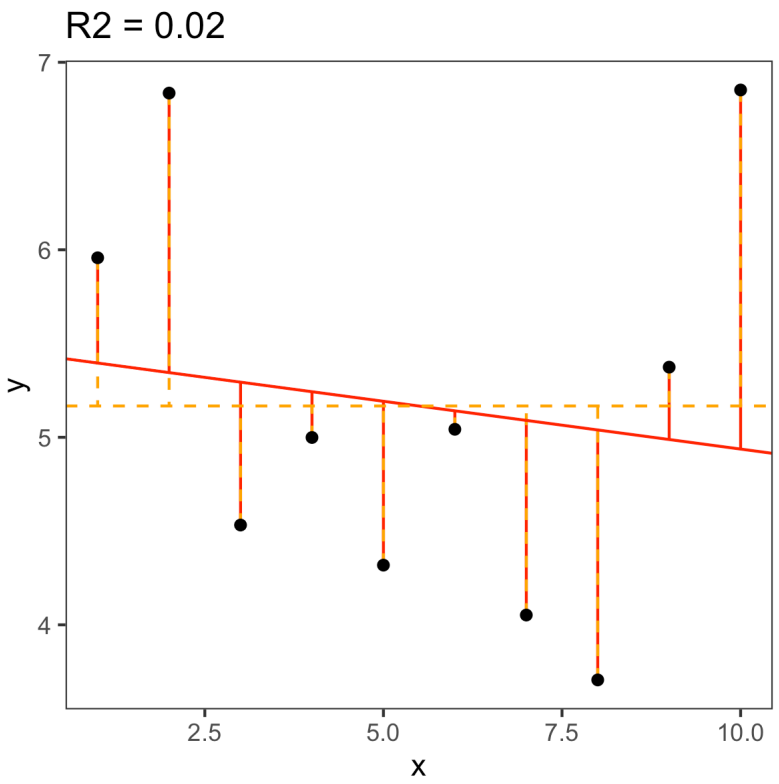
$$R^2 = 1 - \frac{\text{variance explained by the model}}{\text{total variance}}$$

$$R^2 = 1 - \frac{\sum\left(y_i - f(x_i)\right)^2}{\sum(y_i - \bar{y})^2}$$

It is a measure of fit quality. The higher R^2, the better fit.

Adjusted R^2 takes into account number of model parameters.



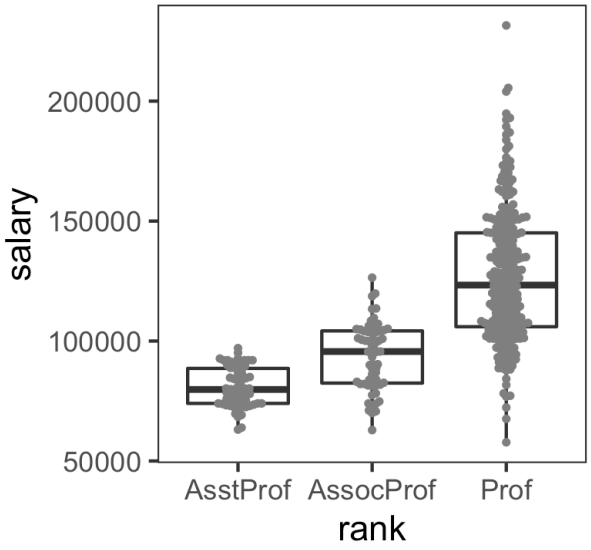


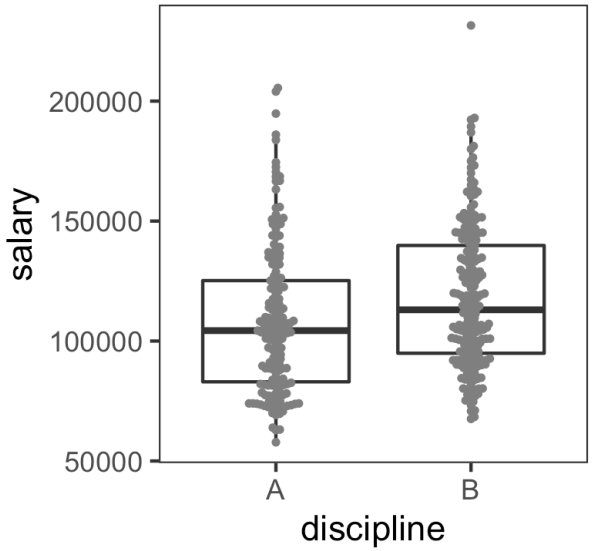
Example

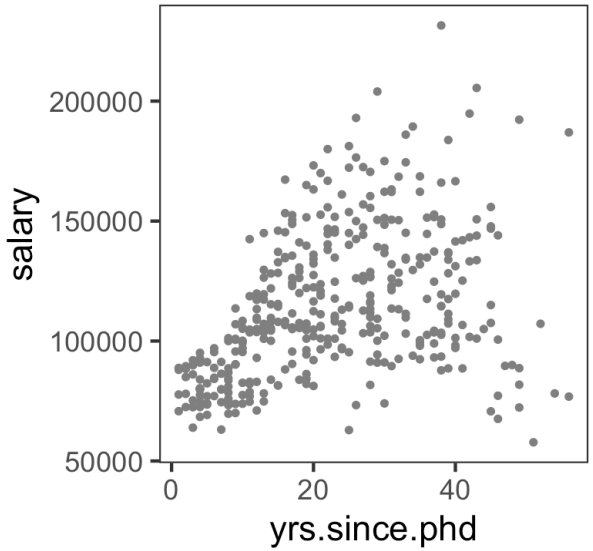
- The 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S.

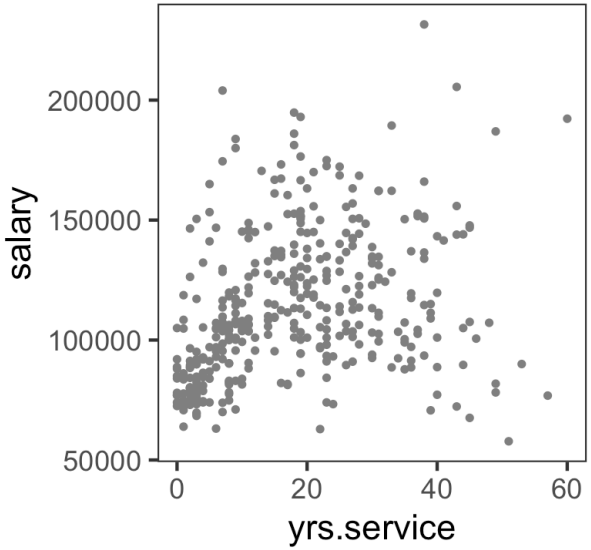
```
> install.packages("car")
> library(car)
> head(Salaries)
      rank discipline yrs.since.phd yrs.service  sex salary
1     Prof          B            19          18 Male 139750
2     Prof          B            20          16 Male 173200
3 AsstProf          B             4           3 Male  79750
4     Prof          B            45          39 Male 115000
5     Prof          B            40          41 Male 141500
6 AssocProf         B             6           6 Male  97000
```

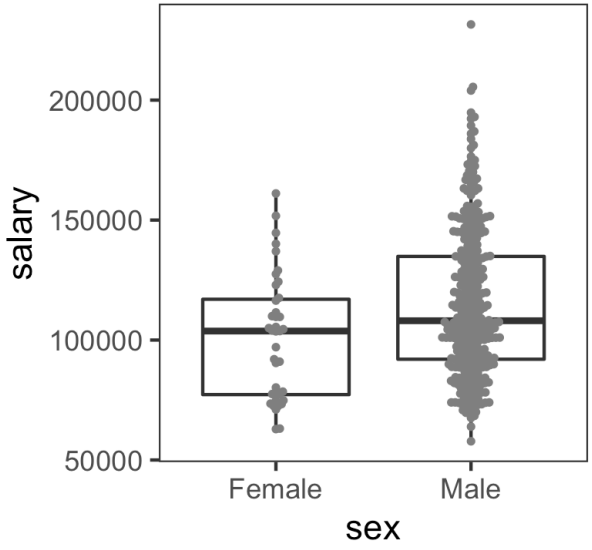
Overview


salary
200000
150000
100000
50000
AsstProf
AssocProf
Prof
rank


salary
200000
150000
100000
50000
A
B
discipline


salary
200000
150000
100000
50000
0
20
40
yrs.since.phd


salary
200000
150000
100000
50000
0
20
40
60
yrs.service


salary
200000
150000
100000
50000
Female
Male
sex

Is there a gender pay gap?

```
Call:
lm(salary ~ sex, data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max
-57290 -23502  -6828  19710 116455

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   101002       4809  21.001  < 2e-16 ***
sexMale        14088       5065   2.782  0.00567 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30030 on 395 degrees of freedom
Multiple R-squared:  0.01921,    Adjusted R-squared:  0.01673
F-statistic: 7.738 on 1 and 395 DF,  p-value: 0.005667
```

Multivariate model

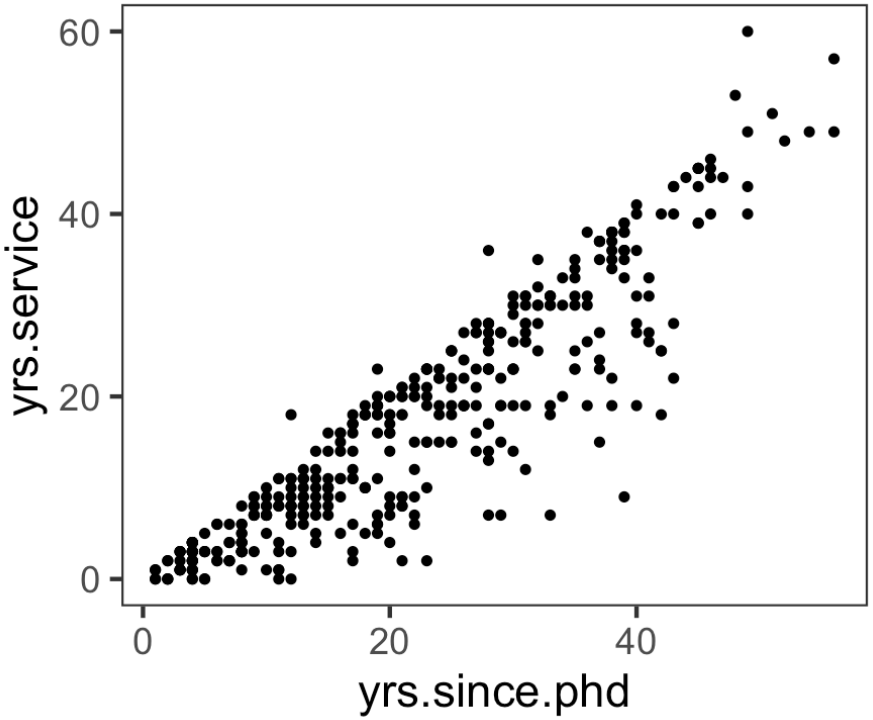
```
Call:
lm(salary ~ rank + discipline + yrs.since.phd + yrs.service + sex,
   data = Salaries)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    65955.2     4588.6  14.374  < 2e-16 ***
rankAssocProf  12907.6     4145.3   3.114  0.00198 **
rankProf       45066.0     4237.5  10.635  < 2e-16 ***
disciplineB    14417.6     2342.9   6.154 1.88e-09 ***
yrs.since.phd    535.1      241.0   2.220  0.02698 *
yrs.service     -489.5      211.9  -2.310  0.02143 *
sexMale         4783.5     3858.7   1.240  0.21584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22540 on 390 degrees of freedom
Multiple R-squared:  0.4547,     Adjusted R-squared:  0.4463
F-statistic:  54.2 on 6 and 390 DF,  p-value: < 2.2e-16
```

Correlated predictors

VIF – variance inflation factors, measure of how much a given variable is explained by others

```
> vif(full_model)
                  GVIF Df GVIF^(1/(2*Df))
rank          2.013193  2        1.191163
discipline    1.064105  1        1.031555
yrs.since.phd 7.518936  1        2.742068
yrs.service   5.923038  1        2.433729
sex           1.030805  1        1.015285
```

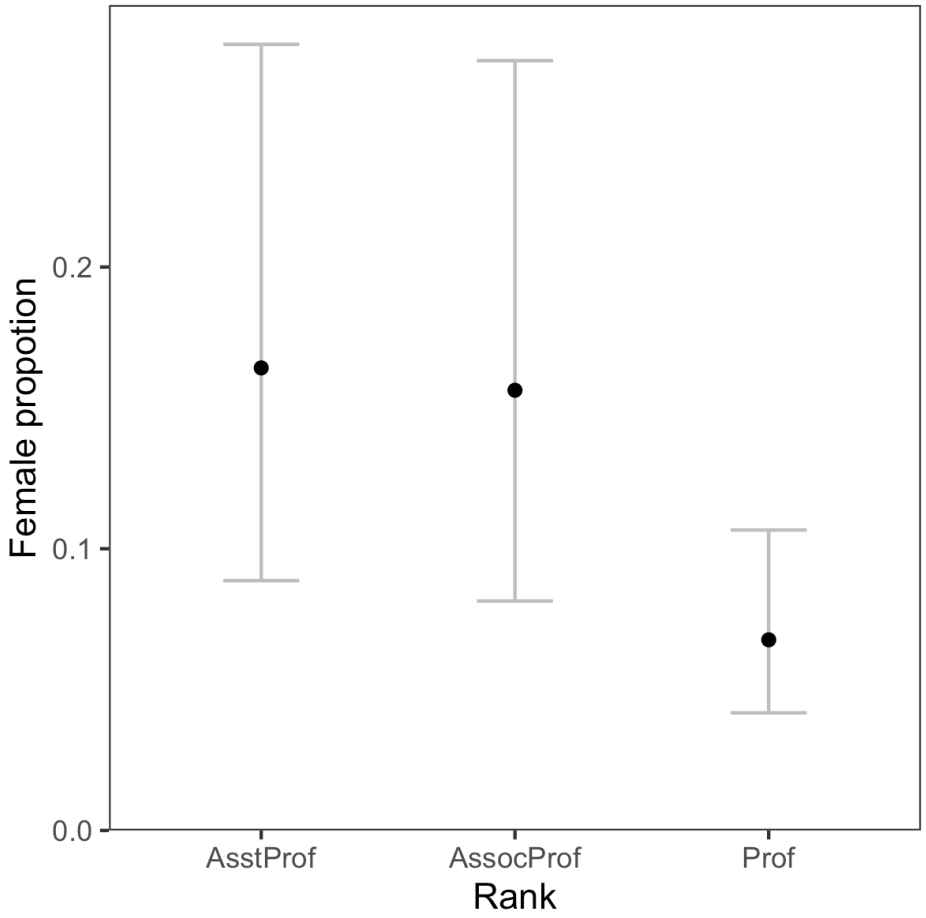
```
> vif(reduced_model)
                  GVIF Df GVIF^(1/(2*Df))
rank          1.987205  2        1.187301
discipline    1.055727  1        1.027486
yrs.since.phd 2.065517  1        1.437191
sex           1.028359  1        1.014080
```



Multivariate reduced model

```
Call:
lm(salary ~ rank + discipline + yrs.since.phd + sex, data = Salaries)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   67884.32    4536.89  14.963  < 2e-16 ***
rankAssocProf 13104.15    4167.31   3.145  0.00179 **
rankProf      46032.55    4240.12  10.856  < 2e-16 ***
disciplineB   13937.47    2346.53   5.940 6.32e-09 ***
yrs.since.phd    61.01     127.01   0.480  0.63124
sexMale        4349.37    3875.39   1.122  0.26242
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22660 on 391 degrees of freedom
Multiple R-squared:  0.4472,     Adjusted R-squared:  0.4401
F-statistic: 63.27 on 5 and 391 DF,  p-value: < 2.2e-16
```

Fewer female professors



Hand-outs available at
https://dag.compbio.dundee.ac.uk/training/Statistics_lectures.html