


Applied model specification

Katie Schuler

2024-10-08

 Under construction

Still working on this!

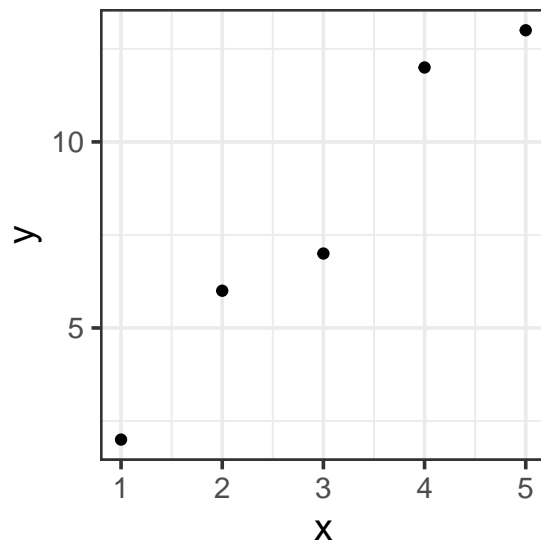
Now that we've covered the terminology and concepts, let's apply model specification to some real models.

```
library(tidyverse)
library(mosaic)
```

1 “Toy” data

Let's start with the simplest possible example, a dataset with two data points. Suppose you record how many days you study over the next two days. On day 1, you study for 2 hours. On day 2, you study for 3 hours. Your dataset might look something like this.

2 Plot



3 Data

```
function (object, ...)  
{  
  UseMethod("model")  
}  
<bytecode: 0x11ac97080>  
<environment: namespace:mosaic>
```

4 Code

```
toy_data <- tibble(  
  x = c(1, 2, 3, 4, 5),  
  y = c(2, 6, 7, 12, 13)  
)  
  
toy_data %>%  
  ggplot(aes(x = x, y = y)) +  
  geom_point() +  
  theme_bw(base_size = 14)
```

1. **Specify our response variable, y :** the response variable (data, output, prediction) is the variable you are trying to predict or explain with your model.

- y

2. **Specify explanatory variables, x_i :** the explanatory variables (regressors, inputs, predictors) are the predictors in your data that could help explain the response variable. Our data has only one possible:

- x

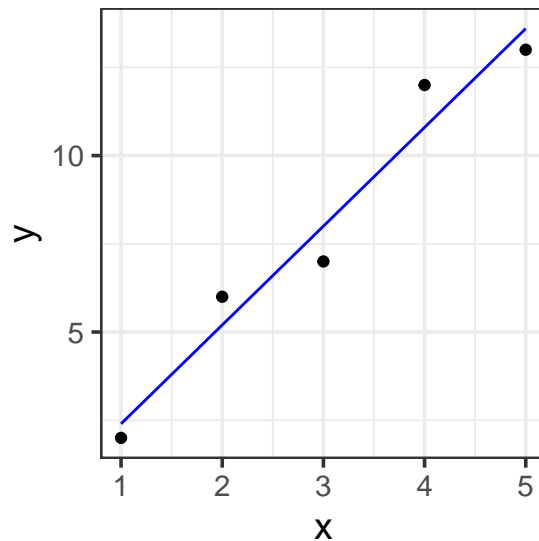
3. **Specify the functional form:** the functional form describes the relationship between the response and explanatory variables with a mathematical expression. In a linear model, we express this relationship as a weighted sum of inputs:

- $y = \sum_{i=1}^n w_i x_i$

4. **Specify model terms:** here we need to specify exactly *how* to express our explanatory variables in our functional form. The actual variables and constants that will be included in the model. There are four kinds of terms: (1) intercept, (2) main, (3) interaction, and (4) transformation. Here we have the simplest case of an intercept and one main term (no interactions or transformations necessary)

- $y = w_1 \mathbf{1} + w_2 x_2$
- in R: `y ~ 1 + x`

5 Plot



Model specification: $y = w_1 \mathbf{1} + w_2 \mathbf{x}$

Call:
lm(formula = y ~ 1 + x, data = toy_data)

Coefficients:
(Intercept) x
 -0.4 2.8

Fitted model: $y = (-0.4)1 + (2.8)x$

6 Data

Call:
lm(formula = y ~ 1 + x, data = toy_data)

Coefficients:
(Intercept) x
 -0.4 2.8

7 Code

```
model <- lm(y ~ 1 + x, data = toy_data)

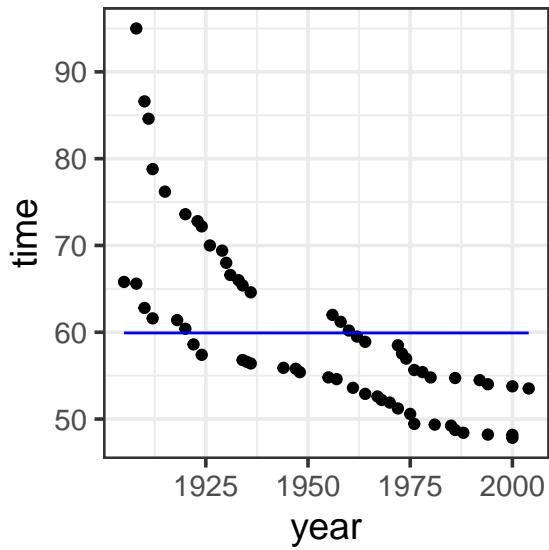
toy_data <- toy_data %>%
  mutate(with_formula = -0.4*1 + 2.8*x) %>%
  mutate(with_predict= predict(model, toy_data))

toy_data %>%
  ggplot(aes(x = x, y = y)) +
  geom_point() +
  geom_line(aes(y = with_predict), color = "blue") +
  theme_bw(base_size = 14)
```

8 Swim records

8.1 One input

9 Plot



Model specification: $y = w_1 \mathbf{1}$

Call:

```
lm(formula = time ~ 1, data = SwimRecords)
```

Coefficients:

(Intercept)
59.92

Fitted model: $y = (59.92)\mathbf{1}$

10 Data

year	time	sex	with_formula	with_predict
1905	65.80	M	59.92	59.92419
1908	65.60	M	59.92	59.92419
1910	62.80	M	59.92	59.92419
1912	61.60	M	59.92	59.92419
1918	61.40	M	59.92	59.92419
1920	60.40	M	59.92	59.92419
1922	58.60	M	59.92	59.92419
1924	57.40	M	59.92	59.92419
1934	56.80	M	59.92	59.92419
1935	56.60	M	59.92	59.92419
1936	56.40	M	59.92	59.92419
1944	55.90	M	59.92	59.92419
1947	55.80	M	59.92	59.92419
1948	55.40	M	59.92	59.92419
1955	54.80	M	59.92	59.92419
1957	54.60	M	59.92	59.92419
1961	53.60	M	59.92	59.92419
1964	52.90	M	59.92	59.92419
1967	52.60	M	59.92	59.92419
1968	52.20	M	59.92	59.92419
1970	51.90	M	59.92	59.92419
1972	51.22	M	59.92	59.92419
1975	50.59	M	59.92	59.92419
1976	49.44	M	59.92	59.92419
1981	49.36	M	59.92	59.92419
1985	49.24	M	59.92	59.92419
1986	48.74	M	59.92	59.92419
1988	48.42	M	59.92	59.92419
1994	48.21	M	59.92	59.92419
2000	48.18	M	59.92	59.92419
2000	47.84	M	59.92	59.92419
1908	95.00	F	59.92	59.92419
1910	86.60	F	59.92	59.92419
1911	84.60	F	59.92	59.92419
1912	78.80	F	59.92	59.92419
1915	76.20	F	59.92	59.92419
1920	73.60	F	59.92	59.92419
1923	72.80	F	59.92	59.92419
1924	72.20	F	59.92	59.92419
1926	70.00	F	59.92	59.92419
1929	69.40	F	59.92	59.92419
1930	68.00	F	59.92	59.92419
1931	66.60	F	59.92	59.92419
1933	66.00	F	59.92	59.92419
1934	65.40	F	59.92	59.92419
1936	64.60	F	59.92	59.92419
1956	62.00	F	59.92	59.92419
1958	61.20	F	59.92	59.92419
1960	60.20	F	59.92	59.92419
1962	59.50	F	59.92	59.92419
1964	58.90	F	59.92	59.92419

11 Code

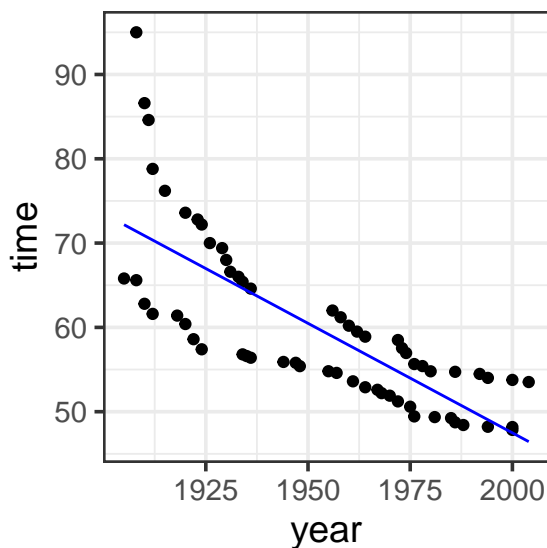
```
model <- lm(time ~ 1, data = SwimRecords)

SwimRecords_predict <- SwimRecords %>%
  mutate(with_formula = 59.92*1) %>%
  mutate(with_predict= predict(model, SwimRecords))

SwimRecords_predict %>%
  ggplot(aes(x = year, y = time)) +
  geom_point() +
  geom_line(aes(y = with_predict), color = "blue") +
  theme_bw(base_size = 14)
```

11.1 Two inputs

12 Plot



Model specification: $y = w_1 \mathbf{1} + w_2 \text{year}$

Call:

```
lm(formula = time ~ 1 + year, data = SwimRecords)
```

Coefficients:

(Intercept)	year
567.2420	-0.2599

Fitted model: $y = (567.2420)1 + (-0.2599)\text{year}$

13 Data

14 Code

```
model <- lm(time ~ 1 + year, data = SwimRecords)

SwimRecords_predict <- SwimRecords %>%
  mutate(with_formula = 567.2420*1 + -0.2599*year) %>%
  mutate(with_predict= predict(model, SwimRecords))

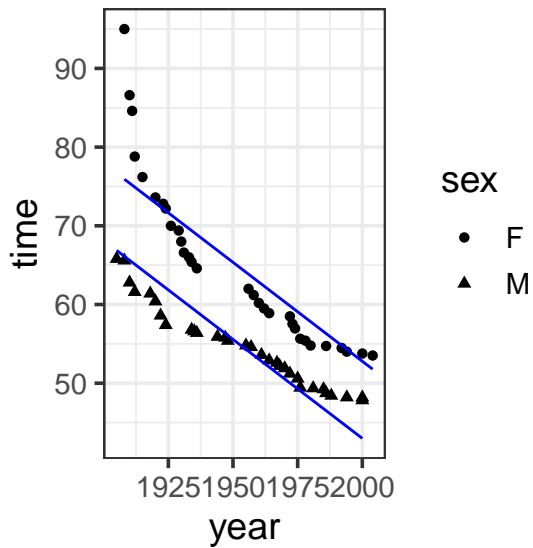
SwimRecords_predict %>%
  ggplot(aes(x = year, y = time)) +
  geom_point() +
  geom_line(aes(y = with_predict), color = "blue") +
  theme_bw(base_size = 14)
```

14.1 Three inputs

15 Plot

```
Warning: There was 1 warning in `mutate()`.
i In argument: `with_formula = 555.7168 * 1 + -0.2515 * year + -9.798 * sex`.
Caused by warning in `Ops.factor()`:
! '*' not meaningful for factors
```


year	time	sex	with_formula	with_predict
1905	65.80	M	72.1325	72.17614
1908	65.60	M	71.3528	71.39651
1910	62.80	M	70.8330	70.87676
1912	61.60	M	70.3132	70.35700
1918	61.40	M	68.7538	68.79774
1920	60.40	M	68.2340	68.27798
1922	58.60	M	67.7142	67.75823
1924	57.40	M	67.1944	67.23848
1934	56.80	M	64.5954	64.63971
1935	56.60	M	64.3355	64.37983
1936	56.40	M	64.0756	64.11995
1944	55.90	M	61.9964	62.04093
1947	55.80	M	61.2167	61.26130
1948	55.40	M	60.9568	61.00143
1955	54.80	M	59.1375	59.18229
1957	54.60	M	58.6177	58.66253
1961	53.60	M	57.5781	57.62302
1964	52.90	M	56.7984	56.84339
1967	52.60	M	56.0187	56.06376
1968	52.20	M	55.7588	55.80388
1970	51.90	M	55.2390	55.28413
1972	51.22	M	54.7192	54.76438
1975	50.59	M	53.9395	53.98474
1976	49.44	M	53.6796	53.72487
1981	49.36	M	52.3801	52.42548
1985	49.24	M	51.3405	51.38597
1986	48.74	M	51.0806	51.12610
1988	48.42	M	50.5608	50.60634
1994	48.21	M	49.0014	49.04708
2000	48.18	M	47.4420	47.48782
2000	47.84	M	47.4420	47.48782
1908	95.00	F	71.3528	71.39651
1910	86.60	F	70.8330	70.87676
1911	84.60	F	70.5731	70.61688
1912	78.80	F	70.3132	70.35700
1915	76.20	F	69.5335	69.57737
1920	73.60	F	68.2340	68.27798
1923	72.80	F	67.4543	67.49835
1924	72.20	F	67.1944	67.23848
1926	70.00	F	66.6746	66.71872
1929	69.40	F	65.8949	65.93909
1930	68.00	F	65.6350	65.67921
1931	66.60	F	65.3751	65.41934
1933	66.00	F	64.8553	64.89958
1934	65.40	F	64.5954	64.63971
1936	64.60	F	64.0756	64.11995
1956	62.00	F	58.8776	58.92241
1958	61.20	F	58.3578	58.40266
1960	60.20	F	57.8380	57.88290
1962	59.50	F	57.3182	57.36315
1964	58.90	F	56.7984	56.84339



Model specification: $y = w_1 \mathbf{1} + w_2 \mathbf{year} + w_3 \mathbf{sex}$

Call:

```
lm(formula = time ~ 1 + year + sex, data = SwimRecords)
```

Coefficients:

(Intercept)	year	sexM
555.7168	-0.2515	-9.7980

Fitted model: $y = (555.7168)\mathbf{1} + (-0.2515)\mathbf{year} + (-9.7980)\mathbf{sex}$

16 Data

17 Code

```
model <- lm(time ~ 1 + year, data = SwimRecords)
```

```
SwimRecords_predict <- SwimRecords %>%
  mutate(with_formula = 555.7168*1 + -0.2515*year + -9.7980 *sex) %>%
  mutate(with_predict= predict(model, SwimRecords))
```

year	time	sex	with_formula	with_predict
1905	65.80	M	NA	66.88051
1908	65.60	M	NA	66.12612
1910	62.80	M	NA	65.62319
1912	61.60	M	NA	65.12026
1918	61.40	M	NA	63.61148
1920	60.40	M	NA	63.10855
1922	58.60	M	NA	62.60563
1924	57.40	M	NA	62.10270
1934	56.80	M	NA	59.58806
1935	56.60	M	NA	59.33660
1936	56.40	M	NA	59.08513
1944	55.90	M	NA	57.07343
1947	55.80	M	NA	56.31903
1948	55.40	M	NA	56.06757
1955	54.80	M	NA	54.30732
1957	54.60	M	NA	53.80440
1961	53.60	M	NA	52.79854
1964	52.90	M	NA	52.04415
1967	52.60	M	NA	51.28976
1968	52.20	M	NA	51.03830
1970	51.90	M	NA	50.53537
1972	51.22	M	NA	50.03244
1975	50.59	M	NA	49.27805
1976	49.44	M	NA	49.02659
1981	49.36	M	NA	47.76927
1985	49.24	M	NA	46.76341
1986	48.74	M	NA	46.51195
1988	48.42	M	NA	46.00902
1994	48.21	M	NA	44.50024
2000	48.18	M	NA	42.99146
2000	47.84	M	NA	42.99146
1908	95.00	F	NA	75.92408
1910	86.60	F	NA	75.42115
1911	84.60	F	NA	75.16969
1912	78.80	F	NA	74.91822
1915	76.20	F	NA	74.16383
1920	73.60	F	NA	72.90651
1923	72.80	F	NA	72.15212
1924	72.20	F	NA	71.90066
1926	70.00	F	NA	71.39773
1929	69.40	F	NA	70.64334
1930	68.00	F	NA	70.39188
1931	66.60	F	NA	70.14041
1933	66.00	F	NA	69.63749
1934	65.40	F	NA	69.38602
1936	64.60	F	NA	68.88310
1956	62.00	F	NA	63.85382
1958	61.20	F	NA	63.35090
1960	60.20	F	NA	62.84797
1962	59.50	F	NA	62.34504
1964	58.90	F	NA	61.84211

```
SwimRecords_predict %>%
  ggplot(aes(x = year, y = time)) +
  geom_point() +
  geom_line(aes(y = with_predict), color = "blue") +
  theme_bw(base_size = 14)
```

17.1 Interaction

18 Plot

Warning: There were 2 warnings in `mutate()`.

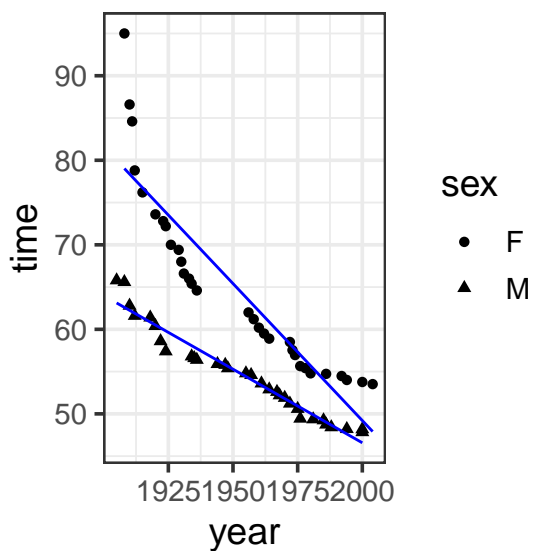
The first warning was:

i In argument: `with_formula = +...`.

Caused by warning in `Ops.factor()`:

! '*' not meaningful for factors

i Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.



Model specification: $y = w_1 \mathbf{1} + w_2 \mathbf{year} + w_3 \mathbf{sex} + w_4 \mathbf{year} * \mathbf{sex}$

Call:

```
lm(formula = time ~ 1 + year * sex, data = SwimRecords)
```

Coefficients:

(Intercept)	year	sexM	year:sexM
697.3012	-0.3240	-302.4638	0.1499

Fitted model: $y = (697.3012)1 + (-0.3240)\text{year} + (-302.4638)\text{sex} + (0.1499)\text{year} \times \text{sex}$

19 Data

20 Code

```
model <- lm(time ~ 1 + year, data = SwimRecords)

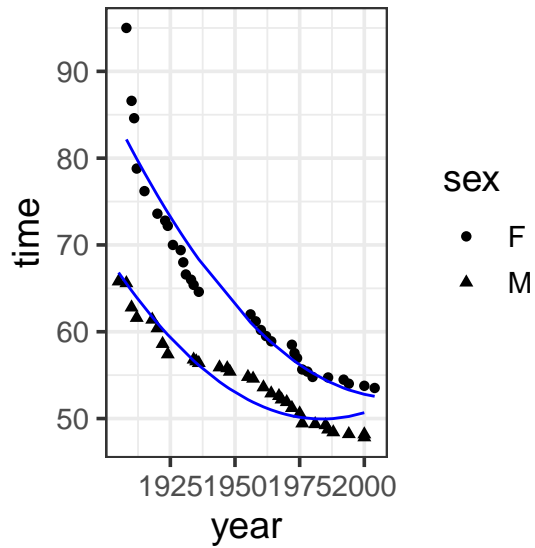
SwimRecords_predict <- SwimRecords %>%
  mutate(with_formula = 697.3012*1 + -0.3240*year +
    -302.4638*sex + 0.1499*year*sex
  ) %>%
  mutate(with_predict= predict(model, SwimRecords))

SwimRecords_predict %>%
  ggplot(aes(x = year, y = time)) +
  geom_point() +
  geom_line(aes(y = with_predict), color = "blue") +
  theme_bw(base_size = 14)
```

year	time	sex	with_formula	with_predict
1905	65.80	M	NA	63.12106
1908	65.60	M	NA	62.59867
1910	62.80	M	NA	62.25041
1912	61.60	M	NA	61.90215
1918	61.40	M	NA	60.85738
1920	60.40	M	NA	60.50912
1922	58.60	M	NA	60.16086
1924	57.40	M	NA	59.81260
1934	56.80	M	NA	58.07131
1935	56.60	M	NA	57.89718
1936	56.40	M	NA	57.72305
1944	55.90	M	NA	56.33002
1947	55.80	M	NA	55.80763
1948	55.40	M	NA	55.63350
1955	54.80	M	NA	54.41459
1957	54.60	M	NA	54.06634
1961	53.60	M	NA	53.36982
1964	52.90	M	NA	52.84743
1967	52.60	M	NA	52.32504
1968	52.20	M	NA	52.15091
1970	51.90	M	NA	51.80266
1972	51.22	M	NA	51.45440
1975	50.59	M	NA	50.93201
1976	49.44	M	NA	50.75788
1981	49.36	M	NA	49.88723
1985	49.24	M	NA	49.19072
1986	48.74	M	NA	49.01659
1988	48.42	M	NA	48.66833
1994	48.21	M	NA	47.62355
2000	48.18	M	NA	46.57878
2000	47.84	M	NA	46.57878
1908	95.00	F	NA	79.02170
1910	86.60	F	NA	78.37361
1911	84.60	F	NA	78.04956
1912	78.80	F	NA	77.72552
1915	76.20	F	NA	76.75338
1920	73.60	F	NA	75.13315
1923	72.80	F	NA	74.16101
1924	72.20	F	NA	73.83697
1926	70.00	F	NA	73.18887
1929	69.40	F	NA	72.21674
1930	68.00	F	NA	71.89269
1931	66.60	F	NA	71.56864
1933	66.00	F	NA	70.92455
1934	65.40	F	NA	70.59651
1936	64.60	F	NA	69.94842
1956	62.00	F	NA	63.46750
1958	61.20	F	NA	62.81941
1960	60.20	F	NA	62.17131
1962	59.50	F	NA	61.52322
1964	58.90	F	NA	60.87513

20.1 Transformation

21 Plot



Model specification:

$$y = w_1 \mathbf{1} + w_2 \mathbf{year} + w_3 \mathbf{sex} + w_4 \mathbf{year} \times \mathbf{sex} + w_5 \mathbf{year}^2$$

Call:

```
lm(formula = time ~ 1 + year * sex + I(year^2), data = SwimRecords)
```

Coefficients:

(Intercept)	year	sexM	I(year^2)	year:sexM
1.110e+04	-1.098e+01	-3.171e+02	2.729e-03	1.575e-01

Fitted model:

$$y = (697.3012)\mathbf{1} + (-0.3240)\mathbf{year} + (-302.4638)\mathbf{sex} + (0.1499)\mathbf{year} \times \mathbf{sex} \quad (1)$$

22 Data

year	time	sex	with_formula	with_predict
1905	65.80	M	72.1325	66.81874
1908	65.60	M	71.3528	65.55576
1910	62.80	M	70.8330	64.74106
1912	61.60	M	70.3132	63.94819
1918	61.40	M	68.7538	61.70057
1920	60.40	M	68.2340	60.99502
1922	58.60	M	67.7142	60.31130
1924	57.40	M	67.1944	59.64941
1934	56.80	M	64.5954	56.66741
1935	56.60	M	64.3355	56.39922
1936	56.40	M	64.0756	56.13650
1944	55.90	M	61.9964	54.23115
1947	55.80	M	61.2167	53.60669
1948	55.40	M	60.9568	53.40946
1955	54.80	M	59.1375	52.18160
1957	54.60	M	58.6177	51.87991
1961	53.60	M	57.5781	51.34200
1964	52.90	M	56.7984	50.99587
1967	52.60	M	56.0187	50.69886
1968	52.20	M	55.7588	50.61078
1970	51.90	M	55.2390	50.45097
1972	51.22	M	54.7192	50.31300
1975	50.59	M	53.9395	50.14697
1976	49.44	M	53.6796	50.10254
1981	49.36	M	52.3801	49.96226
1985	49.24	M	51.3405	49.94827
1986	48.74	M	51.0806	49.95841
1988	48.42	M	50.5608	49.99508
1994	48.21	M	49.0014	50.23605
2000	48.18	M	47.4420	50.67349
2000	47.84	M	47.4420	50.67349
1908	95.00	F	71.3528	82.16082
1910	86.60	F	70.8330	81.03116
1911	84.60	F	70.5731	80.47451
1912	78.80	F	70.3132	79.92332
1915	76.20	F	69.5335	78.30250
1920	73.60	F	68.2340	75.71028
1923	72.80	F	67.4543	74.22044
1924	72.20	F	67.1944	73.73474
1926	70.00	F	66.6746	72.77971
1929	69.40	F	65.8949	71.38810
1930	68.00	F	65.6350	70.93515
1931	66.60	F	65.3751	70.48765
1933	66.00	F	64.8553	69.60003
1934	65.40	F	64.5954	69.17790
1936	64.60	F	64.0756	68.33203
1956	62.00	F	58.8776	61.07389
1958	61.20	F	58.3578	60.46814
1960	60.20	F	57.8380	59.88422
1962	59.50	F	57.3182	59.32213
1964	58.90	F	56.7984	58.78187

23 Code

```
model <- lm(time ~ 1 + year, data = SwimRecords)

SwimRecords_predict <- SwimRecords %>%
  mutate(with_formula = 567.2420*1 + -0.2599*year) %>%
  mutate(with_predict= predict(model, SwimRecords))

SwimRecords_predict %>%
  ggplot(aes(x = year, y = time)) +
  geom_point() +
  geom_line(aes(y = with_predict), color = "blue") +
  theme_bw(base_size = 14)
```

24 Brain size (log)

25 Plant heights (polynomials)

26 Further reading

- [Ch 6: Language of models](#) in Statistical Modeling