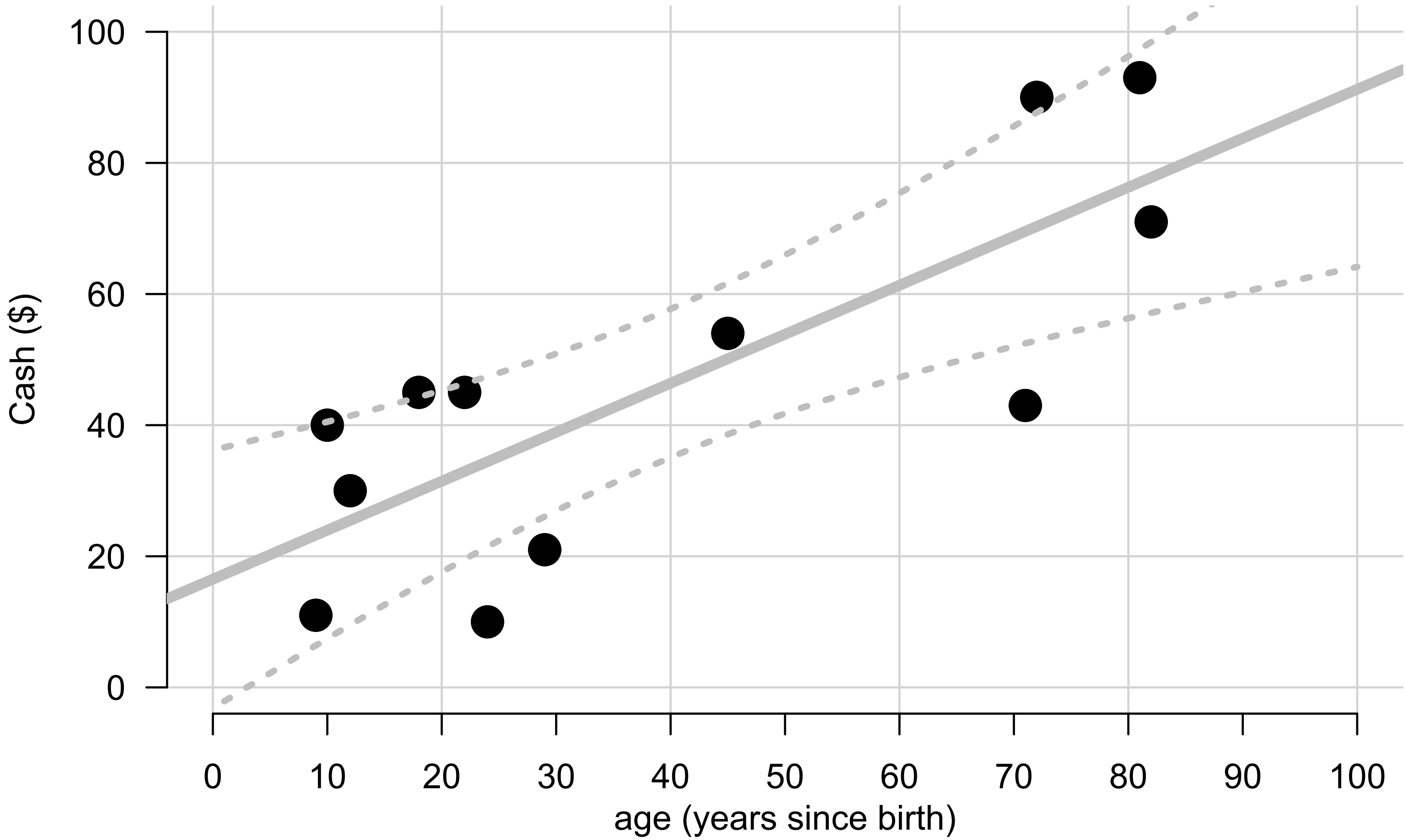
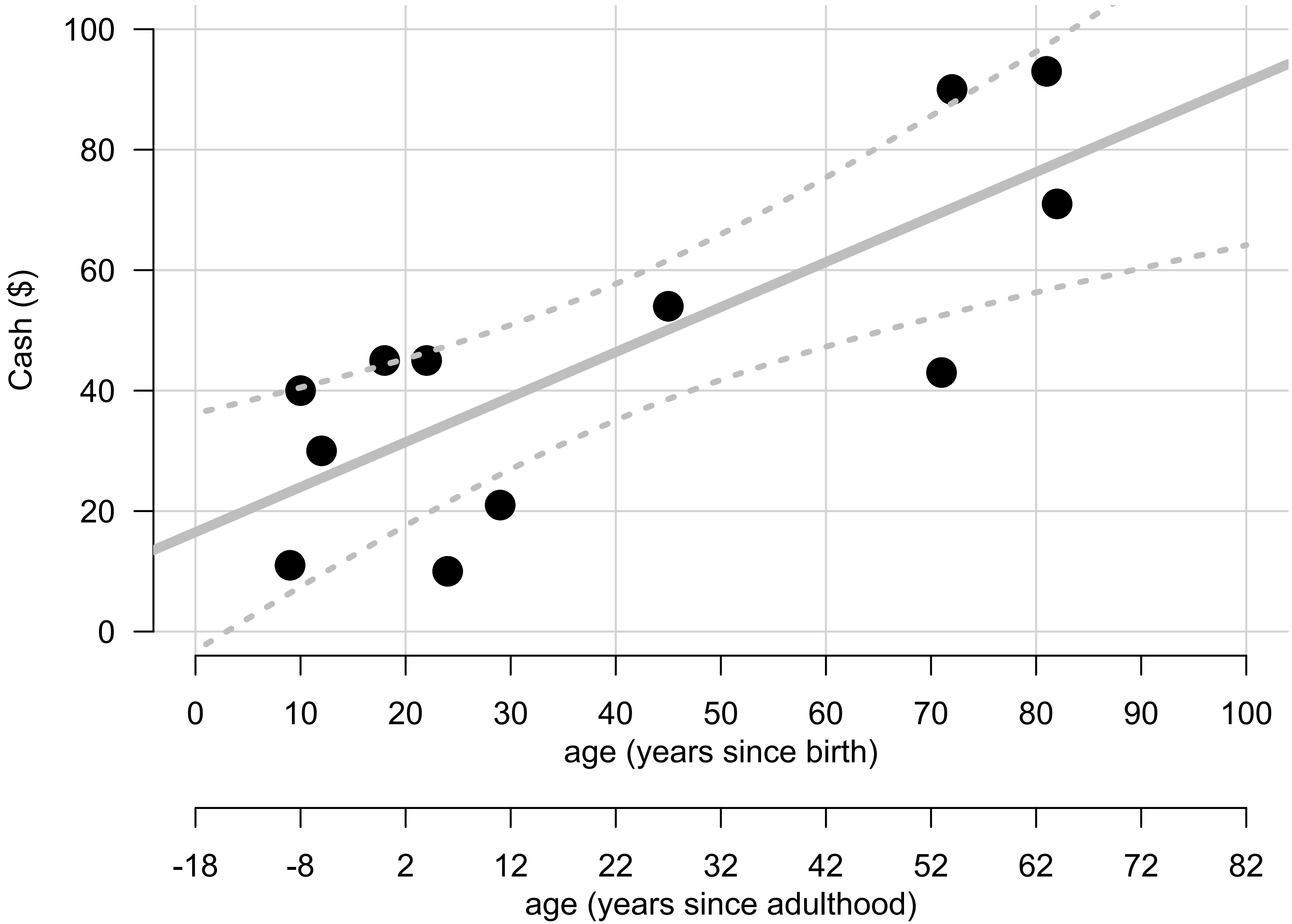
Stat 306: Finding Relationships in Data.

Lecture 12

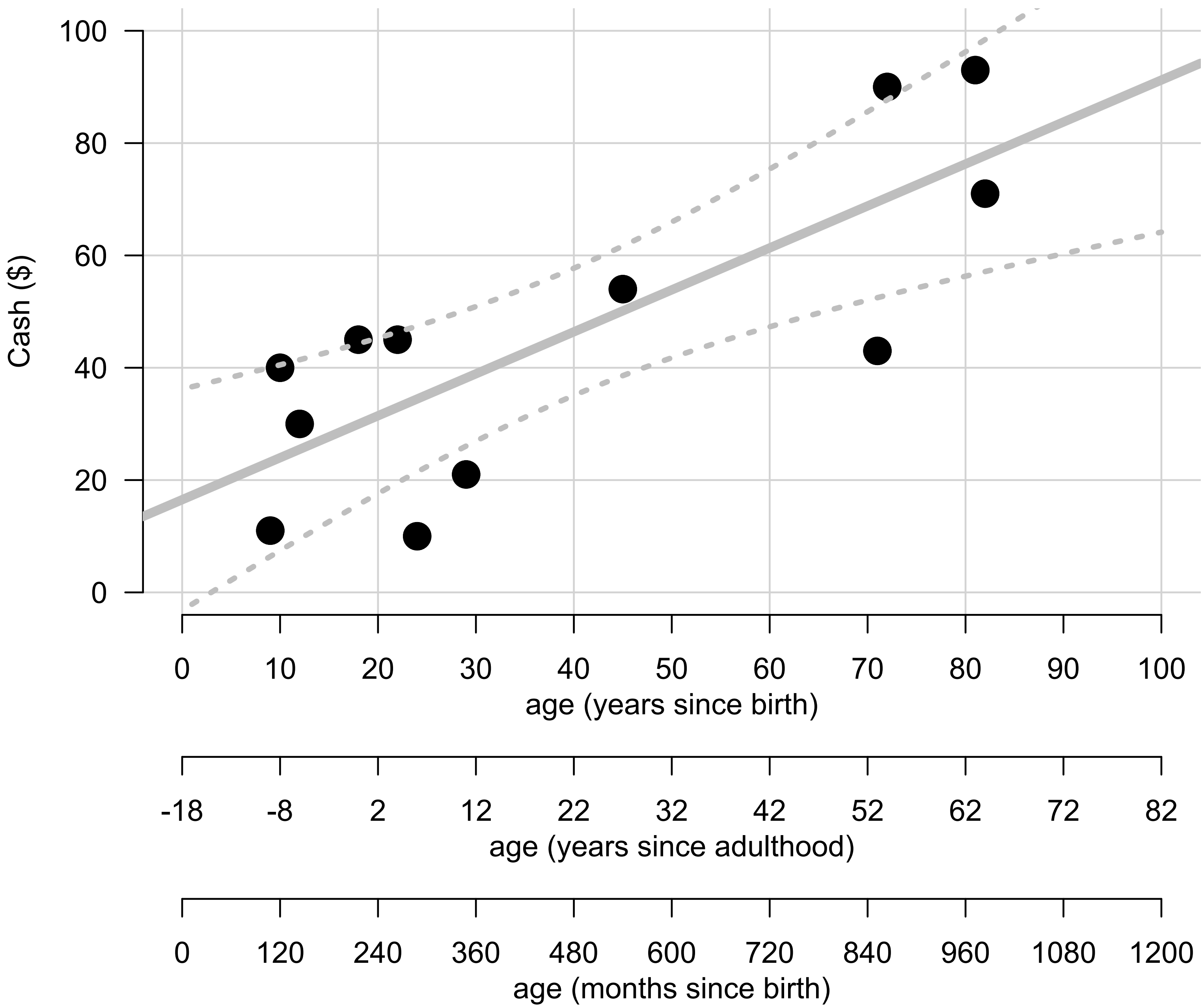
Section 3.11 Multicollinearity



Cash ($)
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80
90
100
age (years since birth)



Cash ($)
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80
90
100
age (years since birth)
-18
-8
2
12
22
32
42
52
62
72
82
age (years since adulthood)



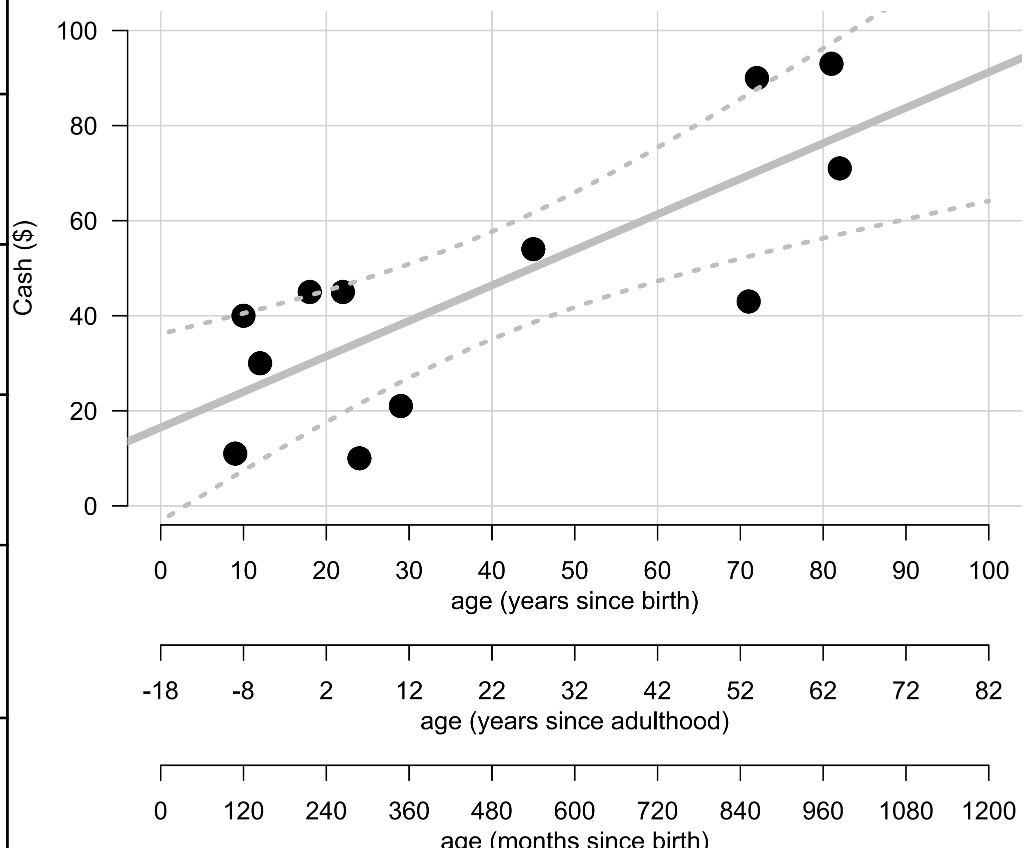
Cash ($)
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80
90
100
age (years since birth)
-18
-8
2
12
22
32
42
52
62
72
82
age (years since adulthood)
0
120
240
360
480
600
720
840
960
1080
1200
age (months since birth)

Changed?	Location shift to X_1	Scale change to X_1
b_1	✗	✓
$SE(b_1)$	✗	✓
Confidence Interval for β_1	✗	✓
p-value $H_0 : \beta_1 = 0$	✗	✗
MS(Res)	✗	✗
R-squared	✗	✗
Adjusted R-squared	✗	✗
F-test	✗	✗

Model:

$$Y = \beta_0 + \beta_1 X_1$$

Example:



```r
# We define y as "cash on hand ($)"
y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10,  90, 40,93)

# We also consider "age (years)" as a continuous variable, x1
# and also "income (K$)", as a continuous variable, x2
x1 <- c(82, 45, 71, 22, 29, 9, 12, 18, 24, 72, 10, 81)

# x1 as measured originally:
years_since_birth <- x1
# Location shift and scale change to x1:
years_since_adult <- x1-18
# Scale change to x1:
months_since_birth <- x1*24

# using "handmade" linear regression function:
linear_reg(y, X=cbind(1, years_since_birth))
linear_reg(y, X=cbind(1, years_since_adult))
linear_reg(y, X=cbind(1, months_since_birth))

# or using lm function:
 summary(lm(y~ years_since_birth))
 summary(lm(y~ years_since_adult))
 summary(lm(y~ months_since_birth))
```

What is different in the output between the 3 models?
What is the same ?

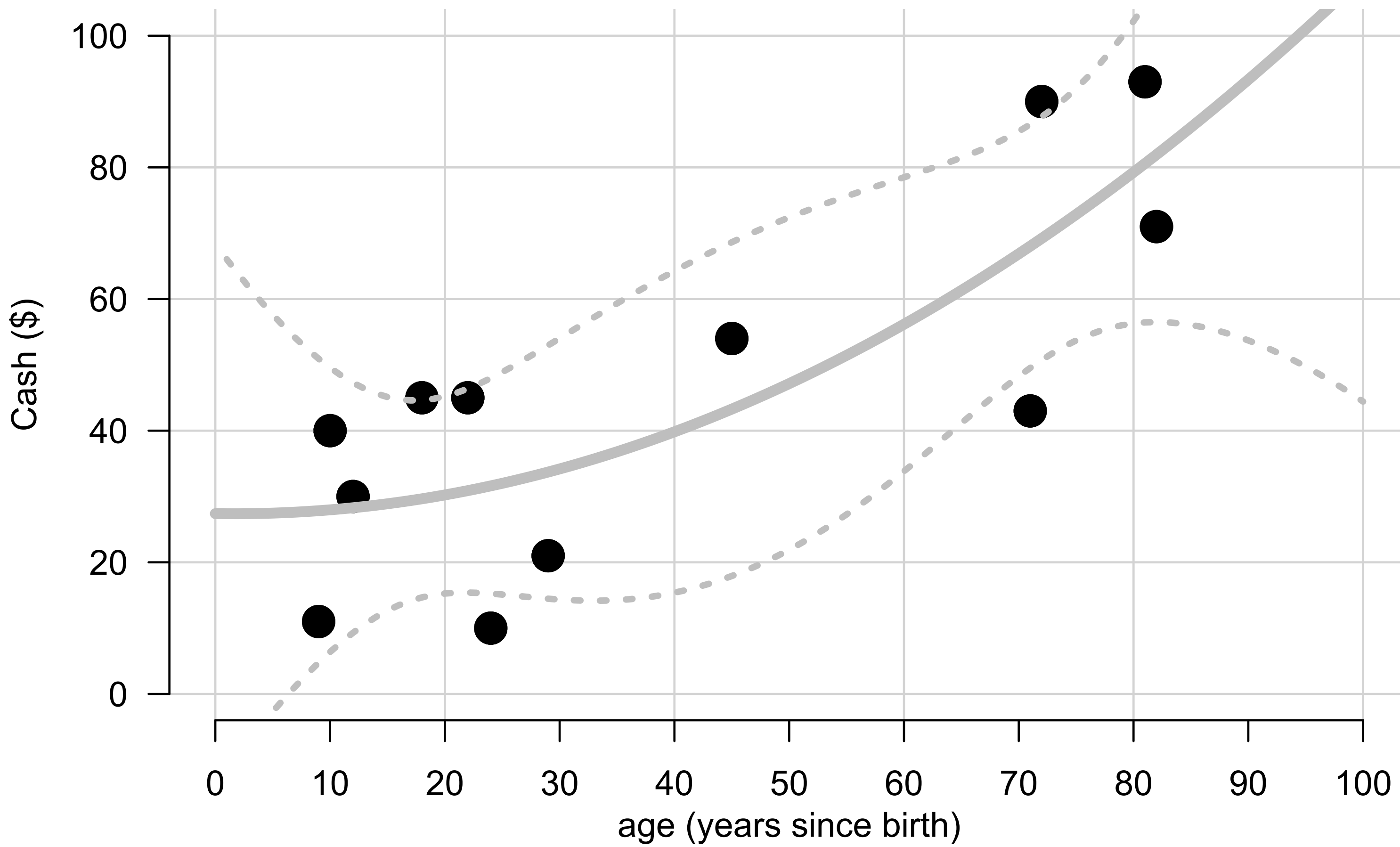
Changed?	Location shift to X_1	Scale change to X_1
b_1	✗	✓
$SE(b_1)$	✗	✓
Confidence Interval for β_1	✗	✓
p-value $H_0 : \beta_1 = 0$	✗	✗
MS(Res)	✗	✗
R-squared	✗	✗
Adjusted R-squared	✗	✗
F-test	✗	✗

Changed?	Location shift to X_1	Scale change to X_1
b_0	✓	✗
$SE(b_0)$	✓	✗
Confidence Interval for β_0	✓	✗
p-value $H_0 : \beta_0 = 0$	✓	✗

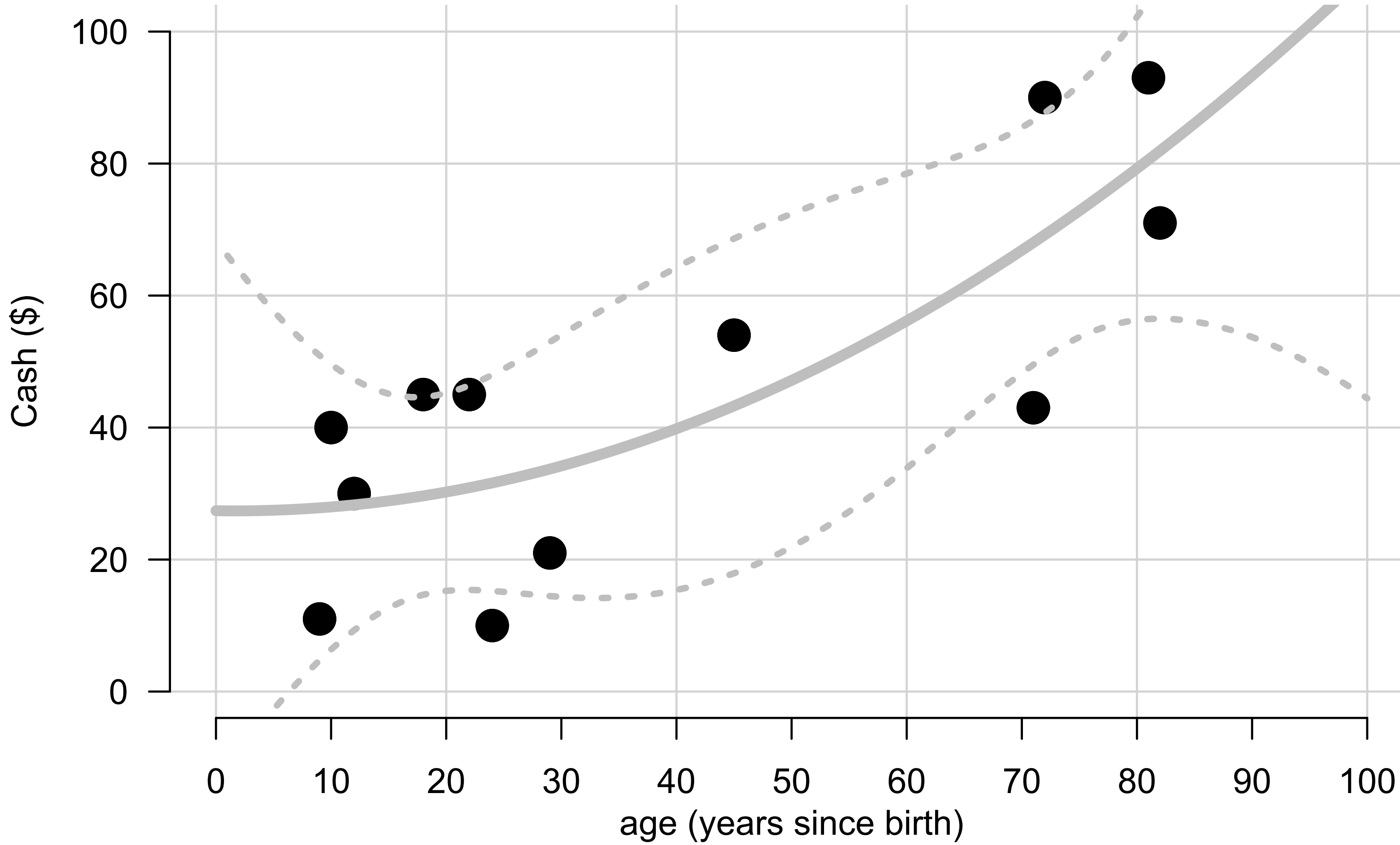
Model:

$$Y = \beta_0 + \beta_1 X_1$$

units



Cash ($)
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
age (years since birth)

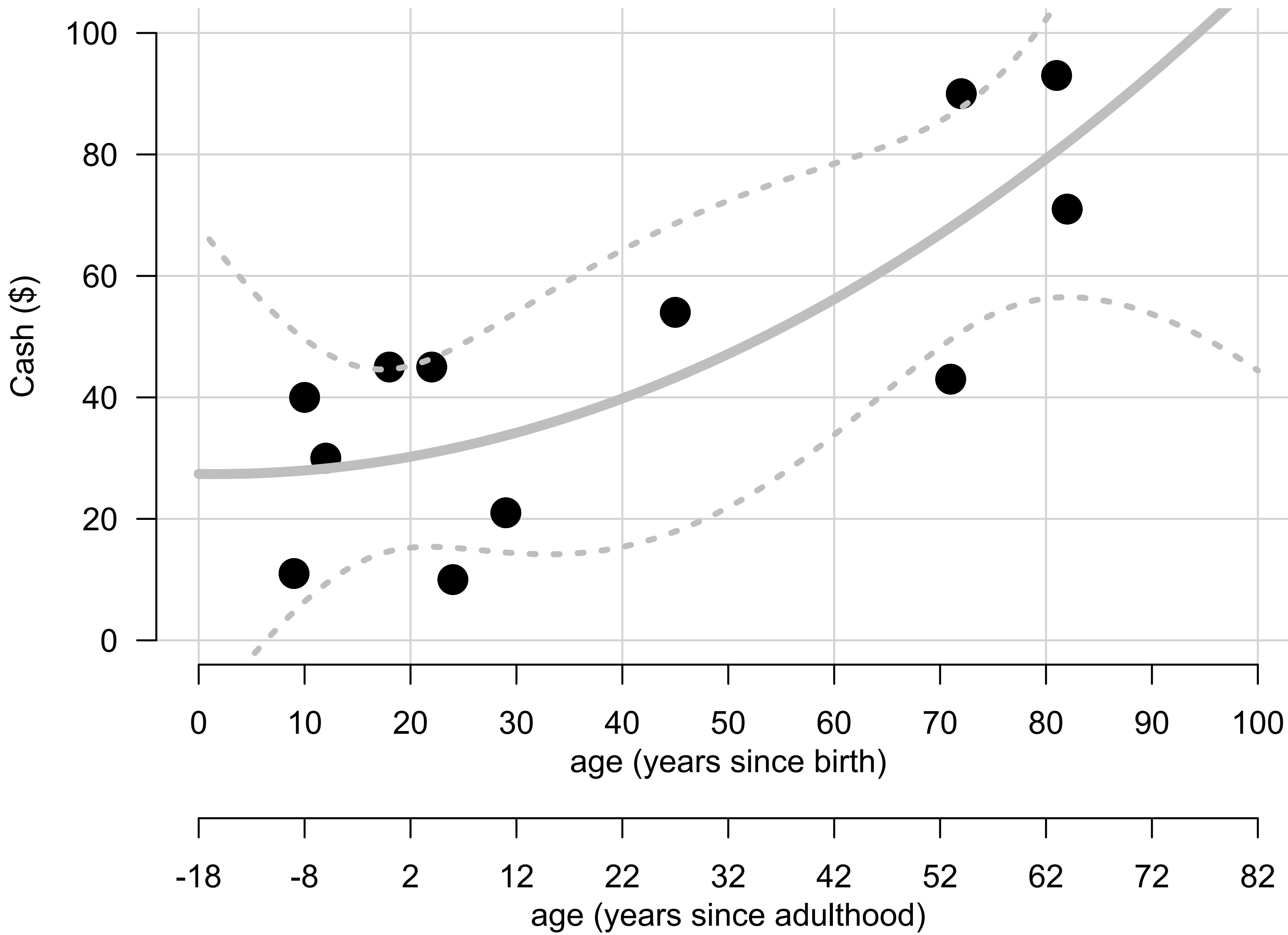


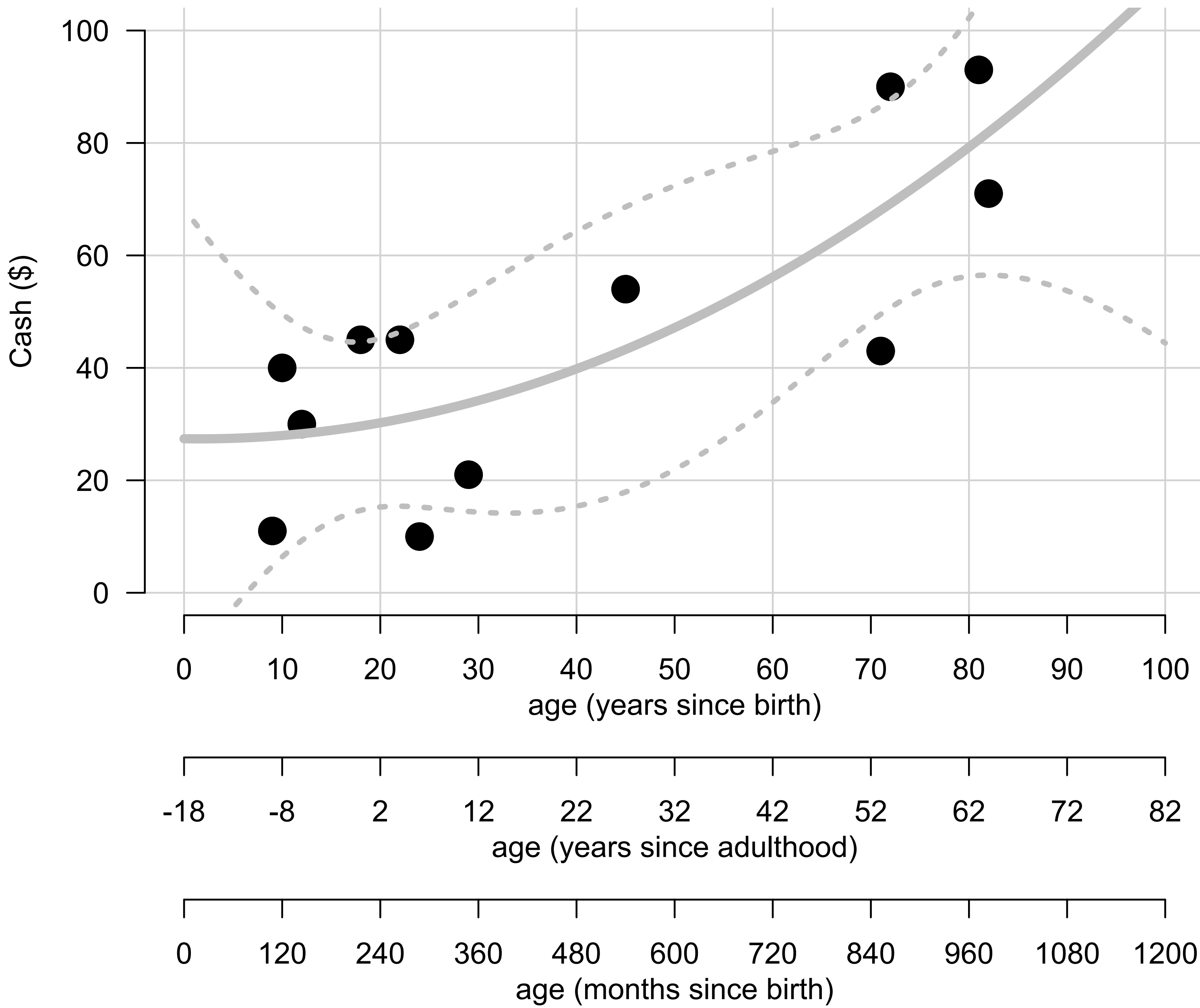
95% **Confidence Interval** for the **subpopulation mean**:

$$\hat{\mu}_Y(\mathbf{x}^*) \quad +/- \quad t_{n-k,0.975}\, se[\hat{\mu}_Y(\mathbf{x}^*)]$$

where: $se[\hat{\mu}_Y(\mathbf{x}^*)] = \hat{\sigma}\sqrt{\mathbf{x}^{*T}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^*}\,.$

Even


Cash ($)
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80
90
100
age (years since birth)
-18
-8
2
12
22
32
42
52
62
72
82
age (years since adulthood)



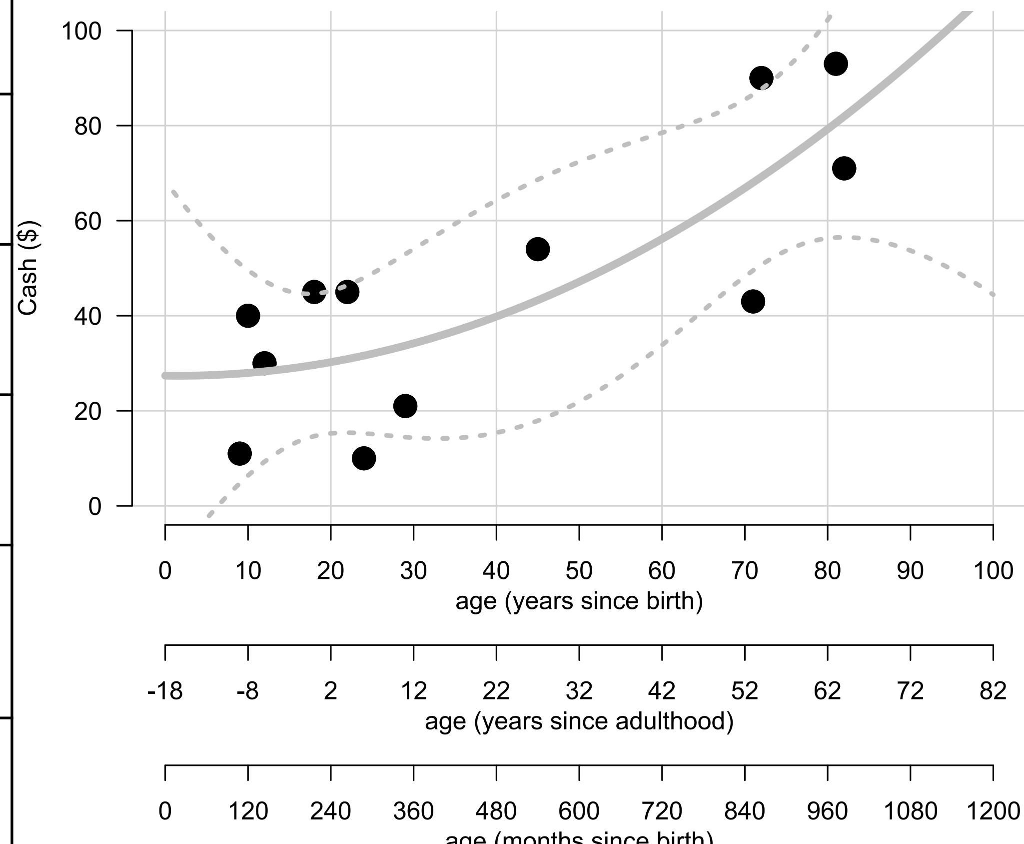
Cash ($)
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80
90
100
age (years since birth)
-18
-8
2
12
22
32
42
52
62
72
82
age (years since adulthood)
0
120
240
360
480
600
720
840
960
1080
1200
age (months since birth)

Changed?	Location shift to X_1	Scale change to X_1
b_1	✔	✔
$SE(b_1)$	✔	✔
Confidence Interval for β_1	✔	✔
p-value $H_0 : \beta_1 = 0$	✔	✔
MS(Res)	✗	✗
R-squared	✗	✗
Adjusted R-squared	✗	✗
F-test	✗	✗

Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2(X_1^2)$$

Example:



```r
# We define y as "cash on hand ($)"
y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10,  90, 40,93)

# We also consider "age (years)" as a continuous variable, x1
# and also "income (K$)", as a continuous variable, x2
x1 <- c(82, 45, 71, 22, 29, 9, 12, 18, 24, 72, 10, 81)

# x1 as measured originally:
years_since_birth <- x1
# Location shift and scale change to x1:
years_since_adult <- x1-18
# Scale change to x1:
months_since_birth <- x1*24

# using "handmade" linear regression function:
linear_reg(y, X=cbind(1, years_since_birth, (years_since_birth)^2))
linear_reg(y, X=cbind(1, years_since_adult, (years_since_adult)^2))
linear_reg(y, X=cbind(1, months_since_birth, (months_since_birth)^2))

# or using lm function:
summary(lm(y~ years_since_birth + I(years_since_birth^2)))
summary(lm(y~ years_since_adult + I(years_since_adult^2)))
summary(lm(y~ months_since_birth + I(months_since_birth^2)))
```

What is different in the output between the 3 models?
What is the same ?

Changed?	Location shift to X_1	Scale change to X_1
b_1	✔	✔
$SE(b_1)$	✔	✔
Confidence Interval for β_1	✔	✔
p-value $H_0 : \beta_1 = 0$	✔	✔
MS(Res)	✗	✗
R-squared	✗	✗
Adjusted R-squared	✗	✗
F-test	✗	✗

Changed?	Location shift to X_1	Scale change to X_1
b_0	✔	✗
$SE(b_0)$	✔	✗
Confidence Interval for β_0	✔	✗
p-value $H_0 : \beta_0 = 0$	✔	✗
b_2	✗	✔
$SE(b_2)$	✗	✔
Confidence Interval for β_2	✗	✔
p-value $H_0 : \beta_2 = 0$	✗	✗

Multicollinearity

Let's remember:

$$\mathrm{Var}[\mathbf{B}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Since a standard error is defined as an estimated square root of the variance of an estimator,

$$se(\hat{\beta}_j) = \hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}, \quad j = 0, 1, \ldots, p. \tag{3.77}$$

	Adding X_2, $Cor(X_2, X_1) = 0$
b_1	✗
$SE(b_1)$	✔
Confidence Interval for β_1	✔
p-value $H_0: \beta_1 = 0$	✔
MS(Res)	✔
R-squared	✔
Adjusted R-squared	✔

$$Y = \beta_0 + \beta_1 X_1$$

vs.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

```r
y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10,  90, 40,93)
x1 <- c(82, 45, 71, 22, 29, 9, 12, 18, 24, 72, 10, 81)
x2 <- c(-15.35, 6.66, -54.29, 23.52, -33.99, -24.88, 6.48, 28.91, -48.96,
 38.34, 31.52, 32.43)

# X1 and X2 are not correlated:
cor(x1, x2)

linear_reg(y, X=cbind(1, x1))
linear_reg(y, X=cbind(1, x1, x2))

summary(lm(y ~ x1))
summary(lm(y ~ x1 + x2))
```

	Adding X_2, $Cor(X_2, X_1) = 0$	Adding X_2, $Cor(X_2, X_1) \neq 0$
b_1	✗	✔
$SE(b_1)$	✔	✔
Confidence Interval for β_1	✔	✔
p-value $H_0 : \beta_1 = 0$	✔	✔
MS(Res)	✔	✔
R-squared	✔	✔
Adjusted R-squared	✔	✔

$$Y = \beta_0 + \beta_1 X_1$$

vs.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

```r
y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10,  90, 40,93)
x1 <- c(82, 45, 71, 22, 29, 9, 12, 18, 24, 72, 10, 81)
x2<-c(60,  55,  26,  21, 0,  15,  17,  31,  0, 112,  24,  92)

# X1 and X2 are somewhat correlated:
cor(x1, x2)

linear_reg(y, X=cbind(1, x1))
linear_reg(y, X=cbind(1, x1, x2))

summary(lm(y ~ x1))
summary(lm(y ~ x1 + x2))
```

	Adding X_2, $Cor(X_2, X_1) = 0$	Adding X_2, $Cor(X_2, X_1) \neq 0$	Adding X_2, $\lvert Cor(X_2, X_1) \rvert \approx 1$
b_1	✗	✔	!!
$SE(b_1)$	✔	✔	!!
Confidence Interval for β_1	✔	✔	!!
p-value $H_0 : \beta_1 = 0$	✔	✔	!!
MS(Res)	✔	✔	!!
R-squared	✔	✔	!!
Adjusted R-squared	✔	✔	!!

```r
y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10,  90, 40, 93)
x1 <- c(82, 45, 71, 22, 29, 9, 12, 18, 24, 72, 10, 81)
x2 <- c(101,  63,  91,  36,  43,  24,  31,  37,  36,  87,  34,  98)

# X1 and X2 are highly highly correlated:
cor(x1,x2)

linear_reg(y, X=cbind(1, x1))
linear_reg(y, X=cbind(1, x1, x2))

summary(lm(y ~ x1))
summary(lm(y ~ x1 + x2))
```

```r
y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10,  90, 40, 93)
x1 <- c(82, 45, 71, 22, 29, 9, 12, 18, 24, 72, 10, 81)
x2 <- c(41.0, 22.5, 35.5, 11.0, 14.5, 4.5, 6.0, 9.0, 12.0, 36.0, 5.0, 40.5)

# X1 and X2 are perfectly correlated:
cor(x1,x2)

linear_reg(y, X=cbind(1, x1))
linear_reg(y, X=cbind(1, x1, x2))

summary(lm(y ~ x1))
summary(lm(y ~ x1 + x2))
```

Let's remember:

$$\text{Var}[\mathbf{B}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Since a standard error is defined as an estimated square root of the variance of an estimator,

$$se(\hat{\beta}_j) = \hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}\,, \quad j = 0, 1, \ldots, p. \tag{3.77}$$

$$(\mathbf{X}^T\mathbf{X})\hat{\mathbf{b}} = \mathbf{X}^T\mathbf{y} \tag{3.10}$$

$$\text{or} \quad \hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.11}$$

Stat 306. Multicollinearity, condition for non-singular $\mathbf{X}^T\mathbf{X}$

1. $\mathbf{X}^T\mathbf{X}$ is non-singular or invertible if and only if $\mathbf{X}$ has linearly independent columns. [$\mathbf{X}^T\mathbf{X}$ is singular iff $\mathbf{X}$ has linearly dependent columns.

2. If $\mathbf{X}$ has columns that are nearly linearly dependent, then $(\mathbf{X}^T\mathbf{X})^{-1}$ exists but has diagonal entries that are large (hence SEs of some $\hat{\beta}$'s can be large). When this happens, the variables are said to be nearly multicollinear.

Mathematical result about $\mathbf{X}^T\mathbf{X}$ (Section 3.11).

(1) $\mathbf{X}^T\mathbf{X}$ is invertible $\iff$ $\mathbf{X}$ has full column rank ($\iff$ the columns of $\mathbf{X}$ are linearly independent)

$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$ is equation that least squares estimate $\hat{\boldsymbol{\beta}}$ satisfies.

If $\mathbf{X}^T\mathbf{X}$ is non-singular, solution $\hat{\boldsymbol{\beta}}$ is unique.

If $\mathbf{X}^T\mathbf{X}$ is singular, solution $\hat{\boldsymbol{\beta}}$ is non-unique (there exists $\hat{\boldsymbol{\beta}}$ from the geometry of least squares).

If $\text{nrow}(\mathbf{X}) = n < k = \text{ncol}(\mathbf{X})$, $\mathbf{X}^T\mathbf{X}$ is singular.

The opposite of statement (1) is:

(2) $\mathbf{X}^T\mathbf{X}$ is singular $\iff$ $\mathbf{X} = (\mathtt{X}_1, \ldots, \mathtt{X}_k)$ has column rank $< k = \text{ncol}(\mathbf{X})$ $\iff$ there is at least one non-trivial linear combination of columns of $\mathbf{X}$ that is linearly dependent; i.e., there are real numbers $a_1, \ldots, a_k$ (not all zero) such that the linear combination $a_1\mathtt{X}_1 + \cdots + a_k\mathtt{X}_k = \mathbf{0}$ ($\mathtt{X}_1, \ldots, \mathtt{X}_k, \mathbf{0}$ here are n-vectors).

Only one such linear combination (up to scaling) implies column rank=$k-1$. Two such unrelated linear combinations (up to scaling) implies column rank=$k-2$ etc.

Multicollinearity is due to:

- Poorly designed study
- Similar problem to having “no control group”

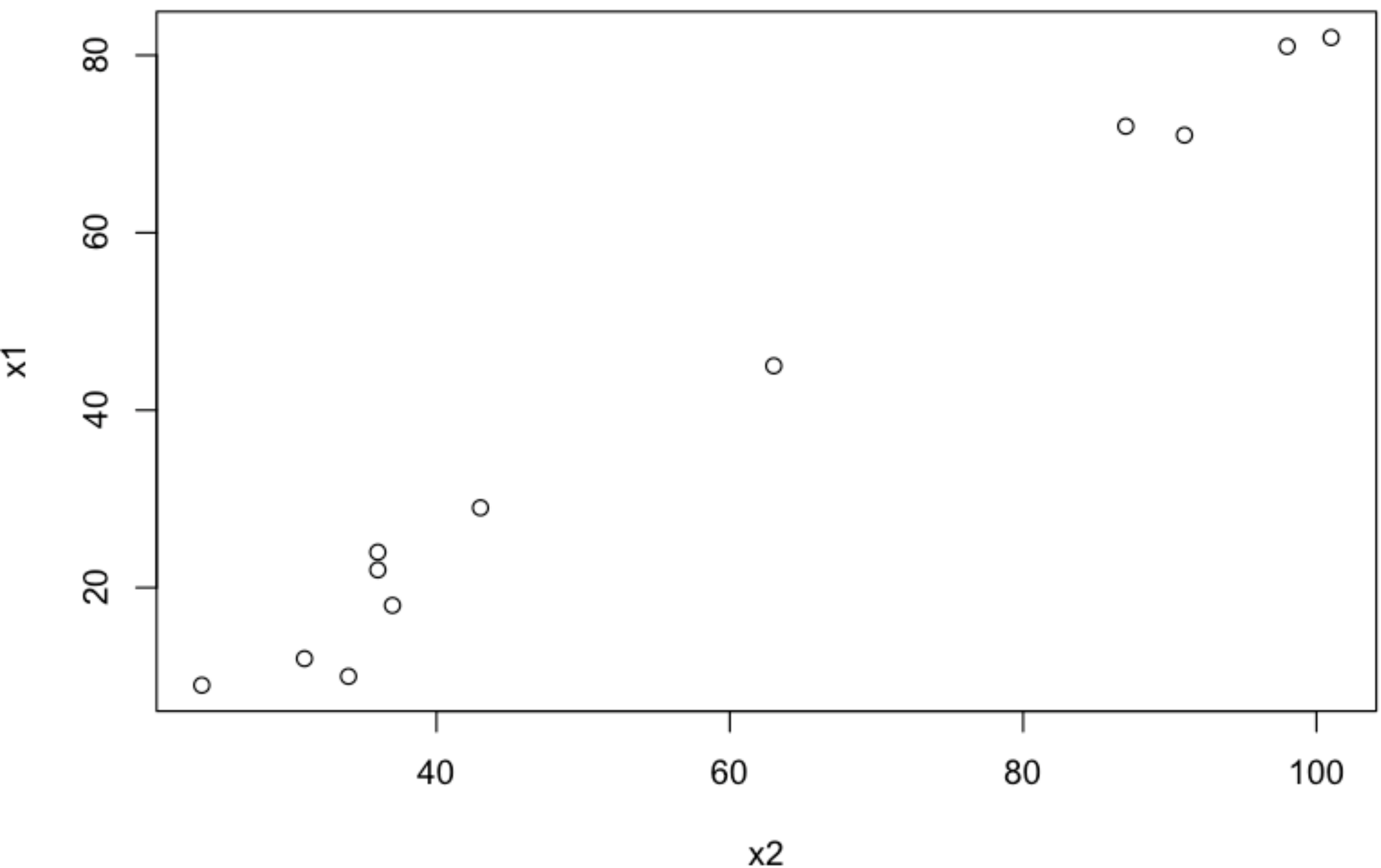
“difficult to disentangle the effect of x1 and x2”

We interpret:

- β_1 as the expected change in **y** due to x_1 , given x_2 is already in the model.
- β_2 as the expected change in **y** due to x_2 , given x_1 is already in the model.

However:

- x_1 and x_2 contribute redundant information about **y**.



80
60
40
20
x1
40
60
80
100
x2

Multicollinearity can be detected using VIF

Another way to check for highly correlated explanatory variables is through regressions of one x variable on the remaining explanatory variables. Let $R^2_{x_j \cdot \mathbf{x}_{-j}}$ denote the R^2 value when x_j is regressed on the other explanatory variables in $\mathbf{X}$. If $R^2_{x_j \cdot \mathbf{x}_{-j}}$ is close to 1, then there is a strong linear relationship between x_j and one or more of the other explanatory variables. Multicollinearity can be measured through *variance inflation factors*

$$VIF_j = \frac{1}{1 - R^2_{x_j \cdot \mathbf{x}_{-j}}}, \quad j = 1, \ldots, p. \tag{3.148}$$

If $\mathrm{VIF}_j \gg 1$, there is multicollinearity involving x_j in the data, and may explain why $SE(\hat{\beta}_j)$ is large.