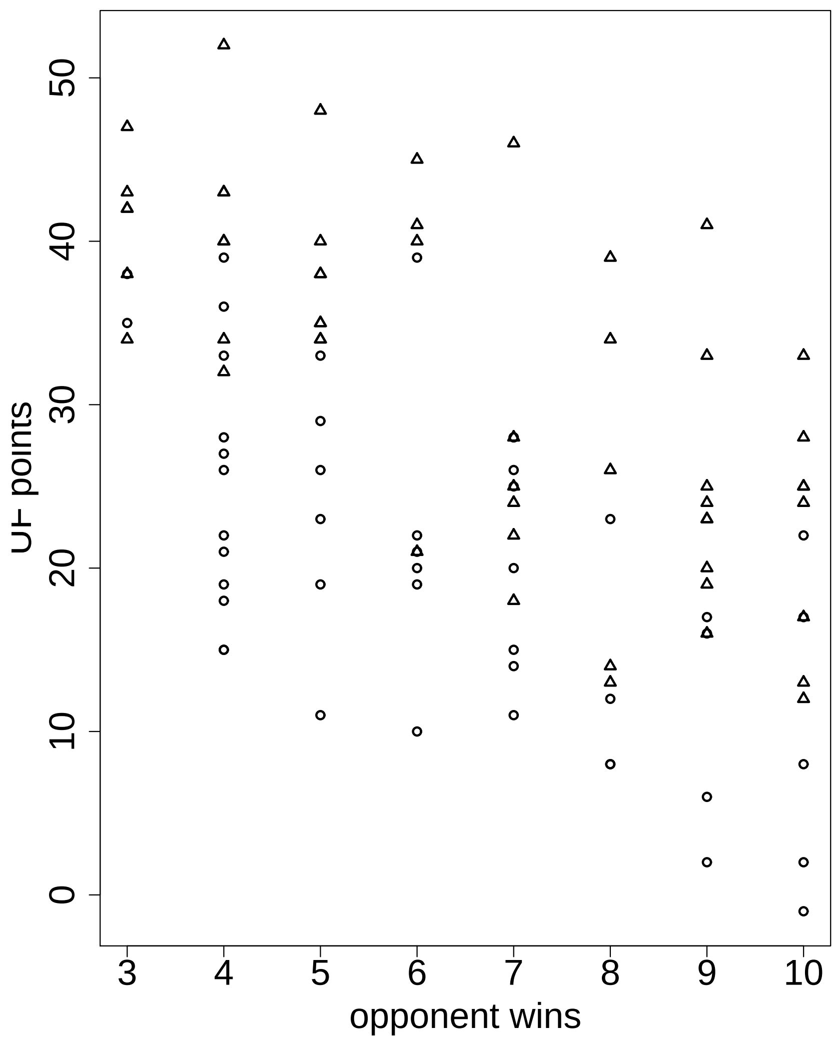
# 11. Qualitative Predictor Variables

**Example:** For the last 100 UF football games we have:

$Y_i$ = #points scored by UF football team in game $i$
$X_{i1}$ = #games won by opponent in their last 10 games

Distinguish between home ($\triangle$) and away ($\circ$) games.




Q: How can we incorporate "home" and "away" into the SLR ?

A: An **indicator variable**:

$$X_{i2} = \begin{cases} 1 & \text{home game} \\ 0 & \text{otherwise} \end{cases}$$

New model

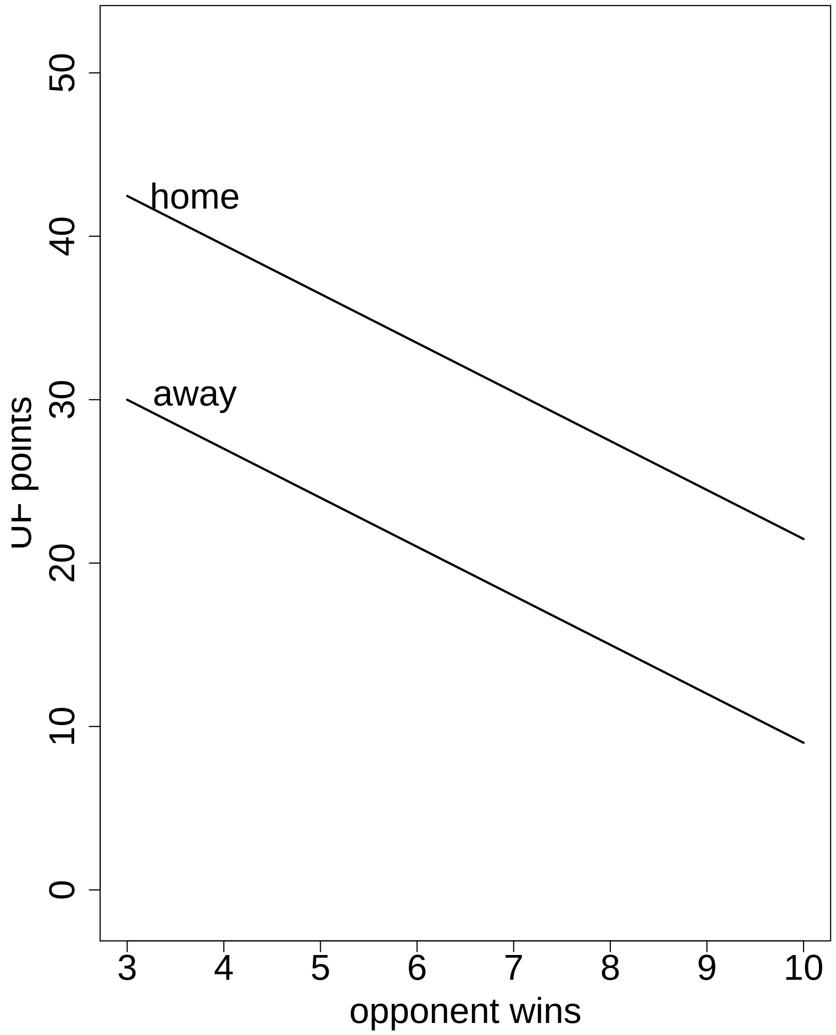
$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

For home games:

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 X_{i1}$$

For away games:

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2(0) = \beta_0 + \beta_1 X_{i1}$$




same slope $\beta_1$ but

different intercepts
$\beta_0 + \beta_2$ and $\beta_0$

How would you decide if a different intercept is necessary?
Test: $H_0 : \beta_2 = 0$ vs. $H_A$ : not $H_0$
t-test:
$t^* = b_2 / \sqrt{\text{MSE} \cdot [(\mathbf{X}'\mathbf{X})^{-1}]_{3,3}}$
F-test:
$F^* = \text{SSR}(X_2|X_1)/\text{MSE}(X_1, X_2)$

**Why not using two indicators ?**

$$X_{i2}^* = \begin{cases} 1 & \text{home game} \\ 0 & \text{otherwise} \end{cases} \qquad X_{i3}^* = \begin{cases} 1 & \text{away game} \\ 0 & \text{otherwise} \end{cases}$$

and considering the model

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2^* X_{i2}^* + \beta_3^* X_{i3}^*$$

Note, $X_{i2}^* + X_{i3}^* = 1$, the respective intercept in the $i$th row of $\mathbf{X}$. Hence, the columns of $\mathbf{X}$ are no longer linearly independent.

**General Rule:** A qualitative variable with $c$ classes will be represented by $c-1$ indicator variables, each taking on the values 0 and 1.

**Question:** How realistic are parallel lines ?

That is, how realistic is it to assume that "UF will score $\beta_2$ more points at home than away, regardless of the strength of the opponent"?

How can we make the model more flexible ?

**Answer:** Add the interaction term

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}$$

For home games: $\mathsf{E}(Y_i) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{i1}$

For away games: $\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1}$

**Q:** How would you answer the question "Is a single line sufficient"?

**A:** Test: $H_0 : \beta_2 = \beta_3 = 0$ vs. $H_A$ : not $H_0$

Test Statistic:

$$F^* = \frac{\mathsf{SSR}(X_1 X_2, X_2 | X_1)/2}{\mathsf{MSE}(X_1, X_2, X_1 X_2)}$$

Rejection rule: reject $H_0$, if $F^* > F(1 - \alpha; 2, n - p)$.

**Q:** How would you make sure this extra sum of squares is available in R?

**A:** Fit the model with the interaction term last !

## More Complex Models

**More than two classes**

Example: $Y_i =$ gas mileage
$X_{i1} =$ age of vehicle
we further have domestic, foreign, and trucks

**Remember General Rule:** The number of indicators that you need is one fewer than the number of levels.

Here we need two such indicators:

$$X_{i2} = \begin{cases} 1 & \text{domestic} \\ 0 & \text{otherwise} \end{cases} \qquad X_{i3} = \begin{cases} 1 & \text{foreign} \\ 0 & \text{otherwise} \end{cases}$$

Model:

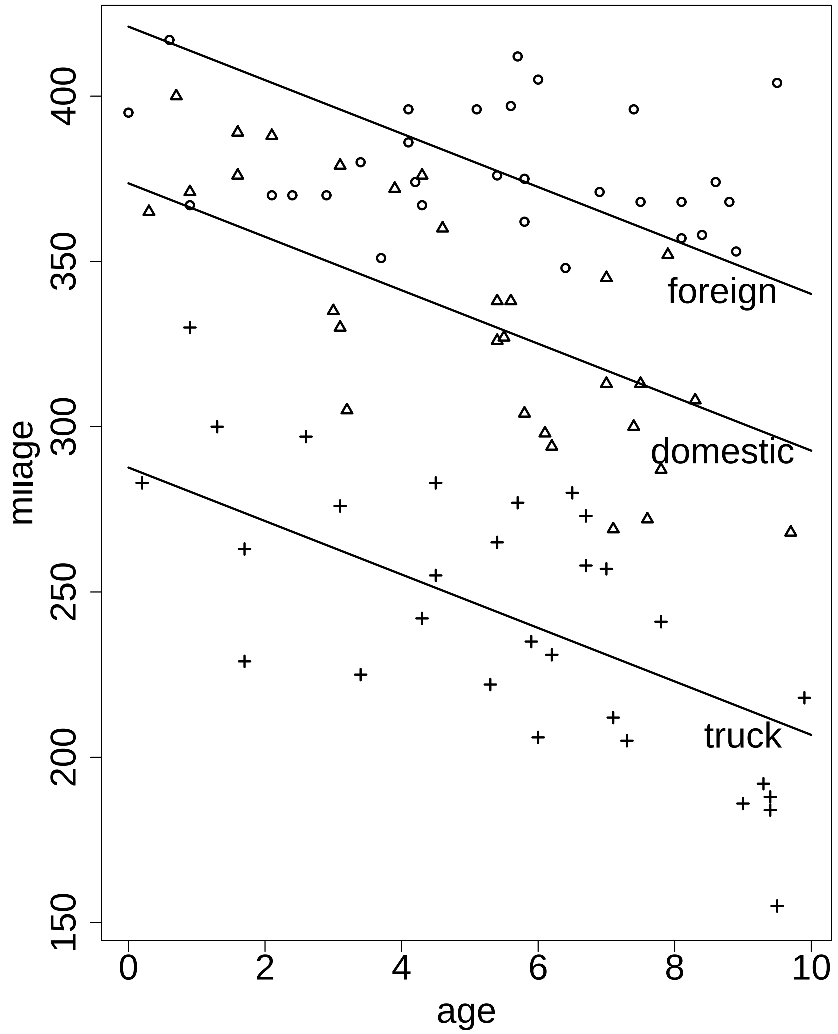
$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$$

$$X_{i2} = \begin{cases} 1 & \text{domestic} \\ 0 & \text{otherwise} \end{cases} \qquad X_{i3} = \begin{cases} 1 & \text{foreign} \\ 0 & \text{otherwise} \end{cases}$$

Model: $\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$

domestic: $\mathsf{E}(Y_i) = (\beta_0 + \beta_2) + \beta_1 X_{i1}$
foreign: $\mathsf{E}(Y_i) = (\beta_0 + \beta_3) + \beta_1 X_{i1}$
trucks: $\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1}$

```
> attach(car); car
   milage age     type
1     388 2.1 domestic
:
90    277 5.7    truck
> x2 <- rep(0, 90) + (type=="domestic")
> x3 <- rep(0, 90) + (type=="foreign")
> lm(milage ~ age + x2 + x3, data=car)
 (Intercept)          age           x2           x3
     287.638       -8.088       85.986      133.384
```



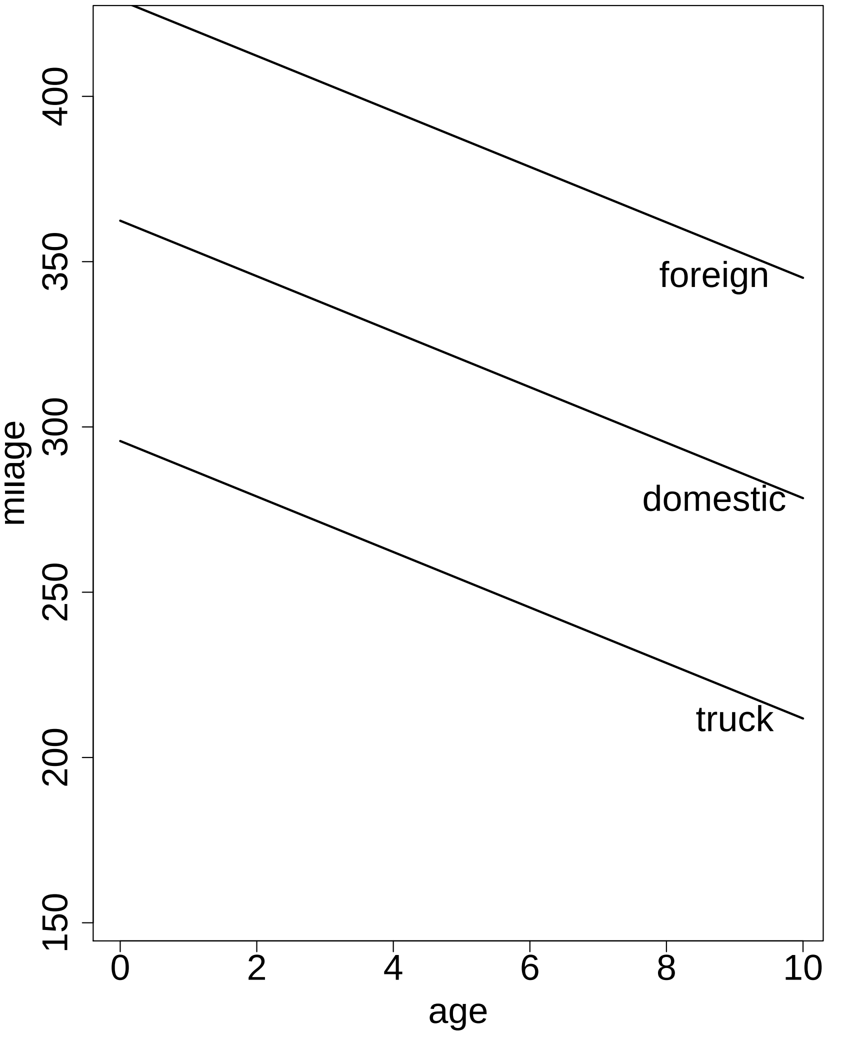
foreign
domestic
truck
mliage
age
150
200
250
300
350
400
0
2
4
6
8
10

**FAQ:** Why couldn't we use 1 indicator with 3 values:

$$X_{i2}^* = \begin{cases} 0 & \text{trucks} \\ 1 & \text{domestic} \\ 2 & \text{foreign} \end{cases}$$

Model: $\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2^* X_{i2}^*$

```
> x2star <- x2 + 2*x3
> lm(milage ~ age + x2star, data=car)
(Intercept)          age       x2star
    295.737       -8.394       66.653
```

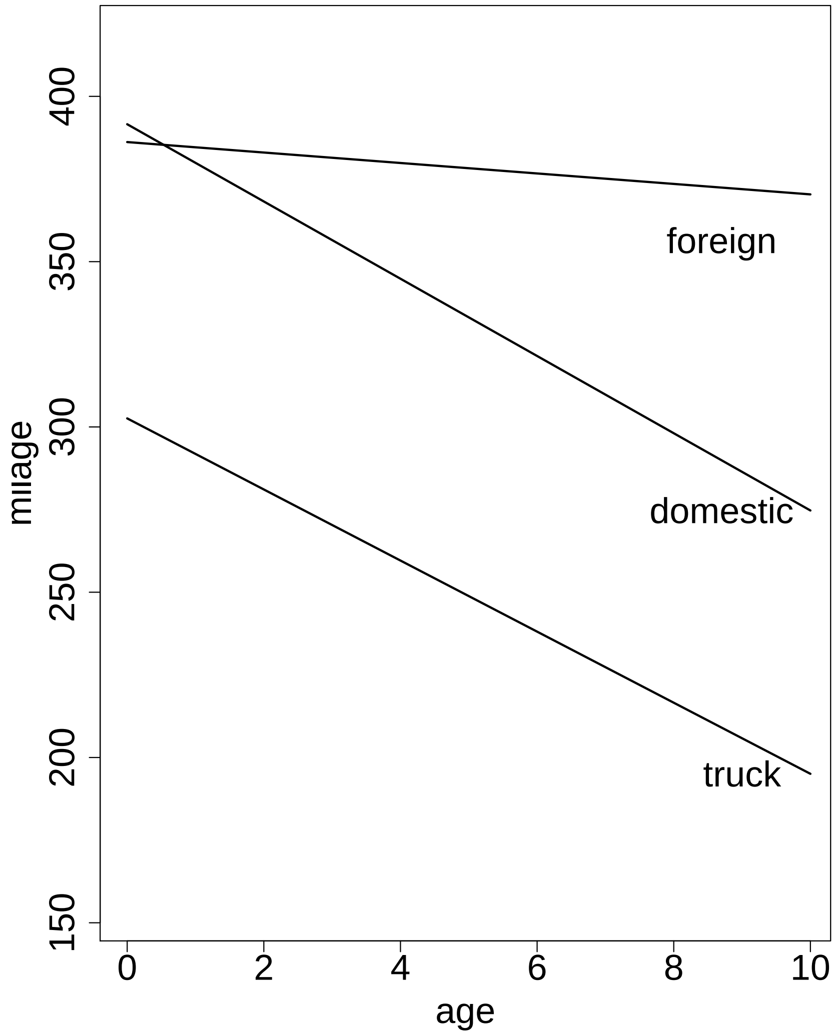

foreign
domestic
truck
milage
age
400
350
300
250
200
150
0
2
4
6
8
10

**Q:** How would we allow each type of vehicle to have its own intercept and slope?

**A:** Add Interactions!

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3}$$

```
> lm(milage ~ age + x2 + x3 + x2:age + x3:age)
Coefficients:
(Intercept)    age      x2      x3   age:x2   age:x3
   302.58   -10.75   88.99   83.60    -0.93     9.17
```




foreign:
$$E(Y_i) = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)X_1$$

domestic:
$$E(Y_i) = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)X_1$$

truck:
$$E(Y_i) = \beta_0 + \beta_1 X_1$$

**More than 1 Qualitative Predictor Variable:**

**Example:** 100 UF football games

$Y_i$ = #points scored by UF football team in game $i$
$X_{i1}$ = #games won by opponent in their last 10 games

Distinguish between home/away and day/night games.

$$X_{i2} = \begin{cases} 1 & \text{home} \\ 0 & \text{away} \end{cases} \qquad X_{i3} = \begin{cases} 1 & \text{day} \\ 0 & \text{night} \end{cases}$$

Model: $\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$

away/day: $\mathsf{E}(Y_i) = (\beta_0 + \beta_3) + \beta_1 X_{i1}$
away/night: $\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1}$

We score $\beta_3$ more points during the day than at night for away games.

$$
\begin{array}{ll}
\text{home/day:} & \mathsf{E}(Y_i) = (\beta_0 + \beta_2 + \beta_3) + \beta_1 X_{i1} \\
\text{home/night:} & \mathsf{E}(Y_i) = (\beta_0 + \beta_2) + \beta_1 X_{i1}
\end{array}
$$

We also score $\beta_3$ more points during the day than at night for home games.

Additional interactions are also possible!

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \beta_6 X_{i2} X_{i3}$$

**Example – House Data:**

$Y_i$ = price/1000
$X_{i1}$ = square feet/1000
$X_{i2} = \begin{cases} 1 & \text{new} \\ 0 & \text{used} \end{cases}$

A model that allows new and used houses to have their own slope and intercept is

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}$$

Submodels:

New: $\mathsf{E}(Y_i) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{i1}$
Used: $\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1}$

How would you test that the regression lines have the same slope?

$H_0 : \beta_3 = 0$ vs. $H_A : \beta_3 \neq 0$

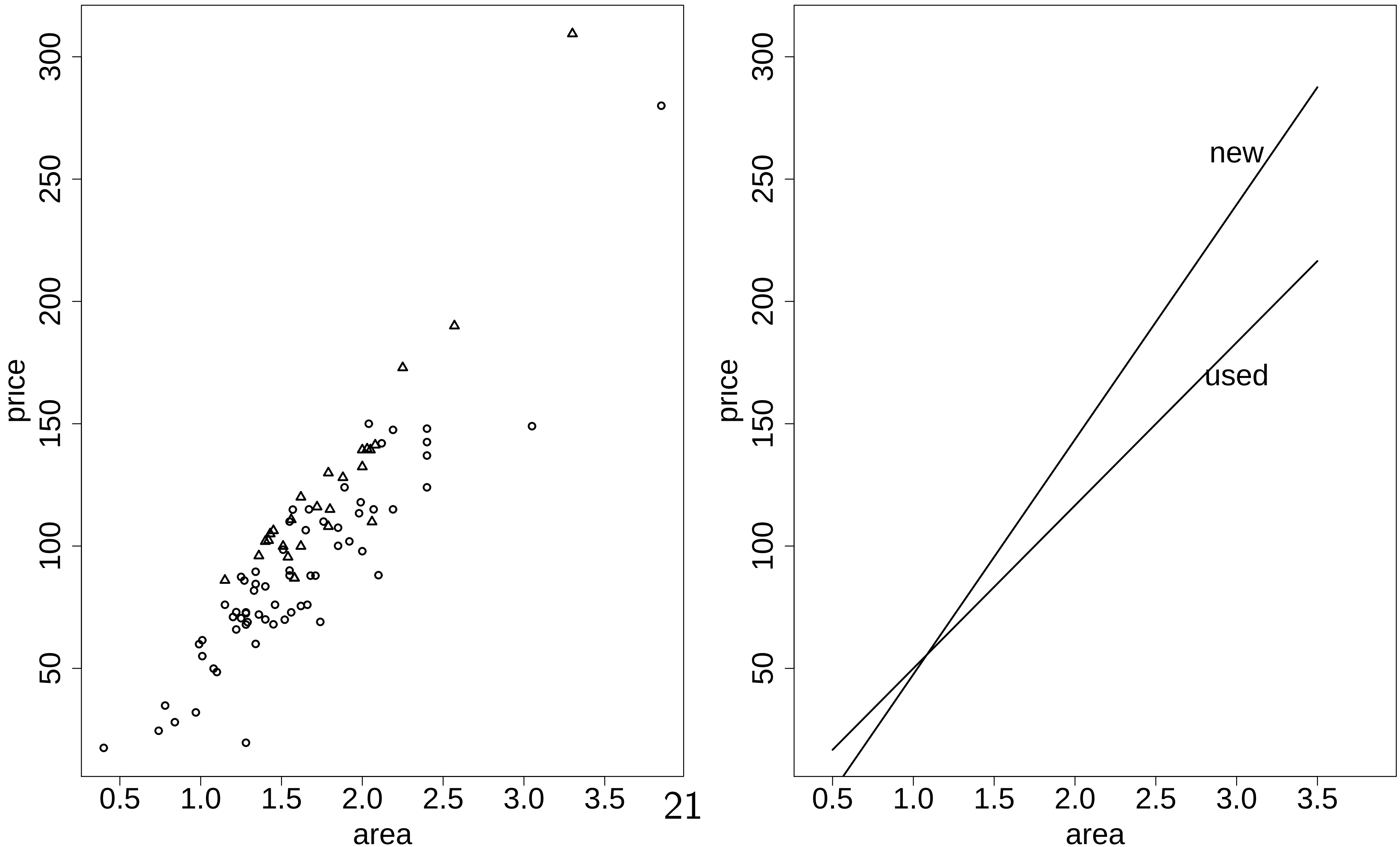
$$
\begin{aligned}
F^* &= \frac{\text{SSR(area*new|area, new)}/1}{\text{MSE(area, new, area*new)}} \\
t^* &= \frac{b_3}{\sqrt{\text{MSE} \cdot [(\mathbf{X}'\mathbf{X})^{-1}]_{4,4}}}
\end{aligned}
$$

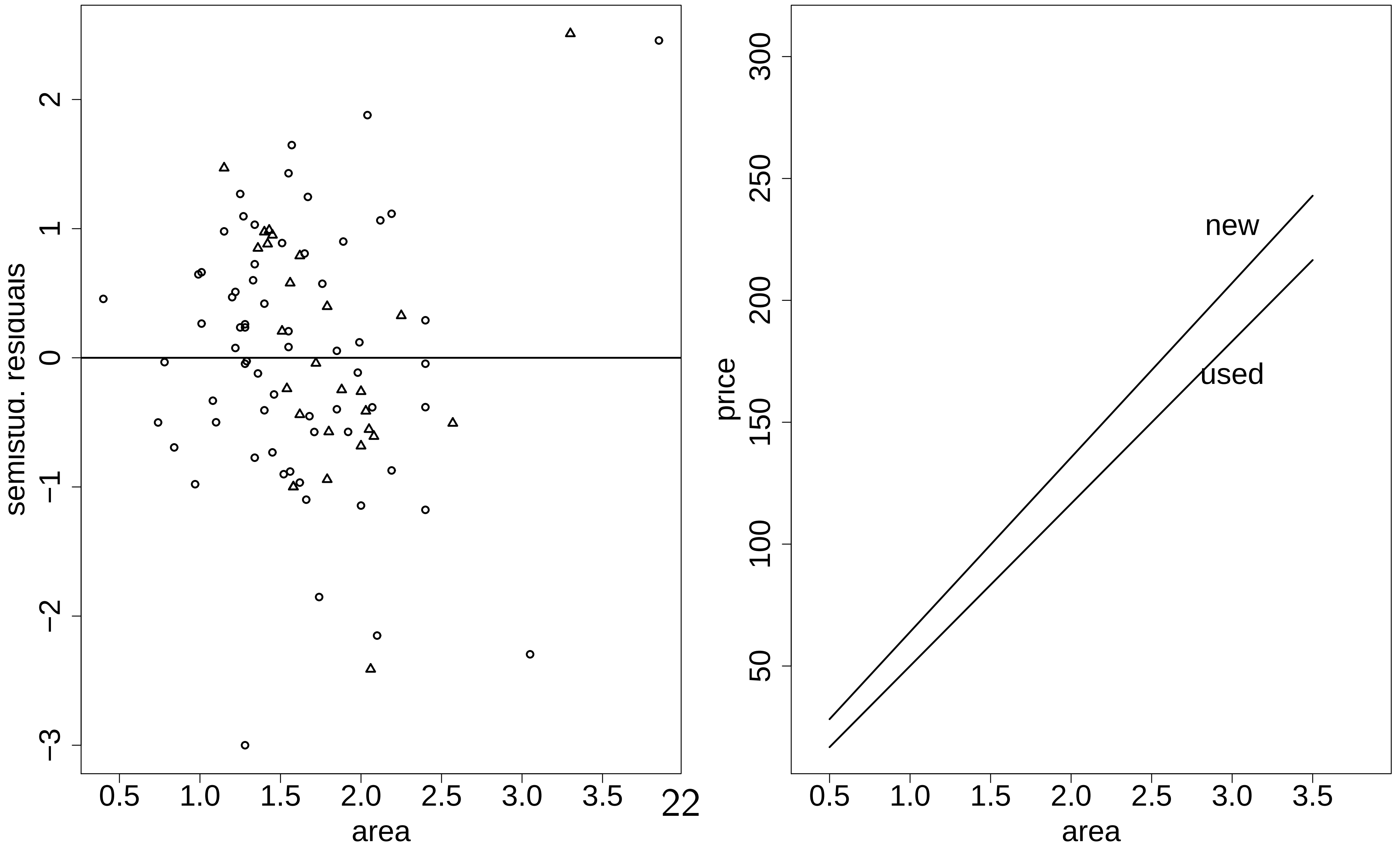
```
> attach(houses)
> hm <- lm(price ~ area+new+area:new); summary(hm)
Coefficients:
           Estimate Std.Error t value Pr(>|t|)
(Intercept) -16.600     6.210  -2.673 0.008944 **
area         66.604     3.694  18.033  < 2e-16 ***
new         -31.826    14.818  -2.148 0.034446 *
area:new     29.392     8.195   3.587 0.000547 ***
---
Sig.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual std. error: 16.35 on 89 degrees of freedom
Mult.R-Squared: 0.8675, Adjusted R-squared: 0.8631
F-stat: 194.3 on 3 and 89 df, p-value:     0
```

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

Response: price
          Df Sum Sq Mean Sq F value    Pr(>F)
area       1 145097  145097 542.722 < 2.2e-16 ***
new        1   7275    7275  27.210 1.178e-06 ***
area:new   1   3439    3439  12.865 0.0005467 ***
Residuals 89  23794     267
---
Sig.codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```


price
area
0.5 1.0 1.5 2.0 2.5 3.0 3.5
50 100 150 200 250 300
new
used


semistud. residuals
area
price
area
new
used

Let's compare two models:

Model 1: $\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}$

where $X_{i2} = \begin{cases} 1 & \text{new} \\ 0 & \text{used} \end{cases}$

Model 2: $\mathsf{E}(Y_i) = \beta_0^* + \beta_1^* X_{i1} + \beta_2^* X_{i2}^* + \beta_3^* X_{i1} X_{i2}^*$

where $X_{i2}^* = \begin{cases} 1 & \text{used} \\ 0 & \text{new} \end{cases}$

| parameter | model 1 | model 2 |
|---|---|---|
| intercept for new | $\beta_0 + \beta_2$ | $\beta_0^*$ |
| intercept for used | $\beta_0$ | $\beta_0^* + \beta_2^*$ |
| slope for new | $\beta_1 + \beta_3$ | $\beta_1^*$ |
| slope for used | $\beta_1$ | $\beta_1^* + \beta_3^*$ |

Thus, we should have

$$\begin{aligned} b_0^* &= b_0 + b_2 \\ b_1^* &= b_1 + b_3 \\ b_2^* &= -b_2 \\ b_3^* &= -b_3 \end{aligned}$$

Let's show that this is indeed the case:

$\mathbf{X}_{n\times 4}$ = design matrix for model 1
$\mathbf{X}^*_{n\times 4}$ = design matrix for model 2

We want to find $\mathbf{M}_{4\times 4}$, such that $\mathbf{X}^* = \mathbf{X}\mathbf{M}$

$$\begin{bmatrix} 1 & X_{11} & 0 & 0 \\ 1 & X_{21} & 1 & X_{21} \\ 1 & X_{31} & 1 & X_{31} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & 1 & X_{11} \\ 1 & X_{21} & 0 & 0 \\ 1 & X_{31} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & 1 & X_{n1} \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

$$\begin{aligned} \mathbf{b}^* &= (\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{Y} \\ &= \left((\mathbf{XM})'(\mathbf{XM})\right)^{-1}(\mathbf{XM})'\,\mathbf{Y} \\ &= \left(\mathbf{M}'\mathbf{X}'\mathbf{XM}\right)^{-1}\mathbf{M}'\mathbf{X}'\mathbf{Y} \\ &= \left(\mathbf{M}^{-1}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{M}')^{-1}\right)\mathbf{M}'\mathbf{X}'\mathbf{Y} \\ &= \mathbf{M}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{M}^{-1}\mathbf{b} \end{aligned}$$

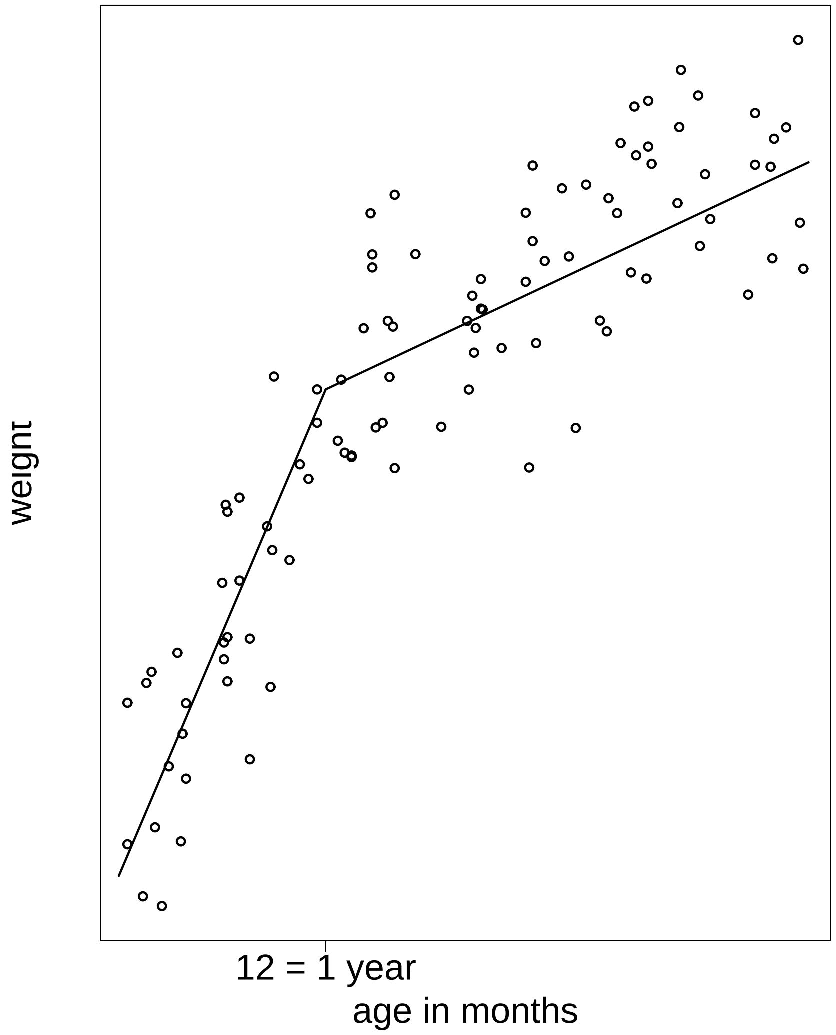
It's easy to show that $\mathbf{M} = \mathbf{M}^{-1}$, so

$$\begin{bmatrix} b_0^* \\ b_1^* \\ b_2^* \\ b_3^* \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} b_0 + b_2 \\ b_1 + b_3 \\ -b_2 \\ -b_3 \end{bmatrix}$$

**Piecewise Linear Regressions**

**Example:**
$Y_i =$ weight of a dog
$X_{i1} =$ age in months

We expect a different weight gain when the dog is a puppy and when it's fully grown. A scatter plot would look like




How would we model this type of data ?

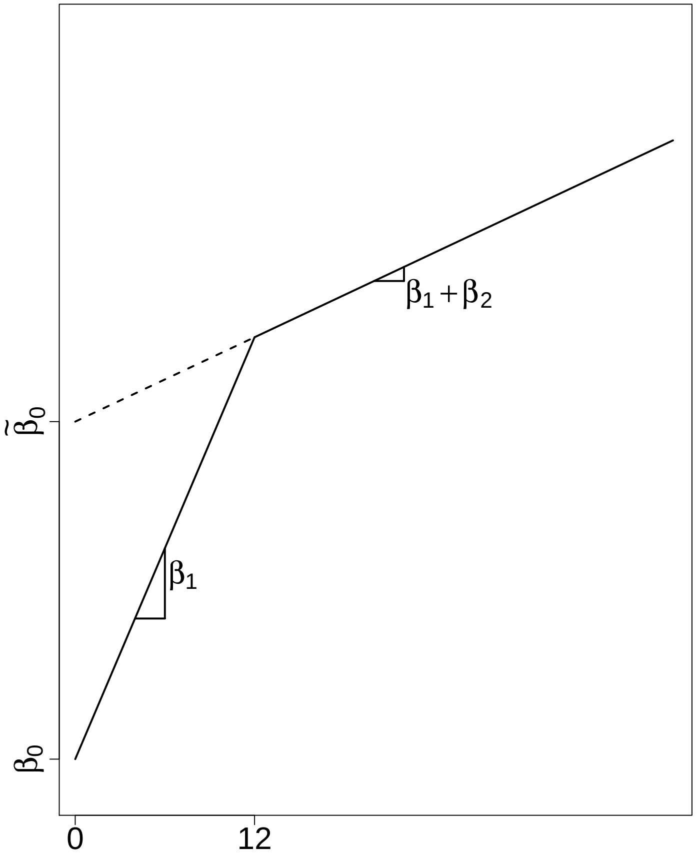
$$
\begin{aligned}
\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} \\
+\beta_2 (X_{i1} - 12) X_{i2}
\end{aligned}
$$

where

$$
X_{i2} = \begin{cases} 1 & X_{i1} > 12 \\ 0 & X_{i1} < 12 \end{cases}
$$

The age of 12 months is called **change-point**.

Derivation: We want




$X_{i1} < 12$:

$\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1}$

$X_{i1} \geq 12$:

$\mathsf{E}(Y_i) = \tilde{\beta}_0 + (\beta_1 + \beta_2) X_{i1}$

But, has to be the same at the changepoint:

$$\begin{aligned}\beta_0 + \beta_1(12) &= \tilde{\beta}_0 + (\beta_1 + \beta_2)(12) \\ \tilde{\beta}_0 &= \beta_0 - 12\beta_2\end{aligned}$$

Thus we want:

For $X_{i1} < 12$: $\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1}$

For $X_{i1} \geq 12$: $\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1} - 12\beta_2$