Week 3. Multiple/multivariate linear regression model

Section 8.1 and Section 8.6

`lm` function for multivariate linear regression.

Multiple/multivariate linear regression model (shortly, linear model) is compactly written in vector/matrix form as

$$\mathbf{y}^{n\times 1}= \mathbf{X}^{n\times m}\boldsymbol{\beta}^{m\times 1}+\boldsymbol{\varepsilon}^{n\times 1}.$$

Assumptions:

- The vector of dependent variables $\mathbf{y}$ and the vector of errors $\boldsymbol{\varepsilon}$ are random but the design matrix $\mathbf{X}$ is fixed/nonrandom.
- In order to estimate $\boldsymbol{\beta}$ uniquely, we assume the identifiability condition:

$$\text{rank}(\mathbf{X}) = m < n.$$

- We say that matrix $\mathbf{X}$ has full rank; $m < n$ means that the number of observations (sample size) is greater than the number of the beta parameters to estimate. The full rank condition means that the column vectors of matrix $\mathbf{X}$ are linearly independent: no column can be expressed as a linear combination of others.
- $cov(\boldsymbol{\varepsilon}) =\sigma^2\mathbf{I}$, that is, errors have zero mean, constant variance and the same variance (homoscedastistic). Since $\mathbf{X}\boldsymbol{\beta}$ is nonrandom we have $cov(\mathbf{y}) =\sigma^2\mathbf{I}$.
- In a special case when $\boldsymbol{\varepsilon}$ are normally distributed we write

$$\mathbf{y} \sim\mathcal{N}(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}).$$

Simple linear regression with the sample size $n = 5$ we have

$$\mathbf{X}^{5\times 2}= \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \end{bmatrix}, \quad \boldsymbol{\beta}^{2\times 1}= \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

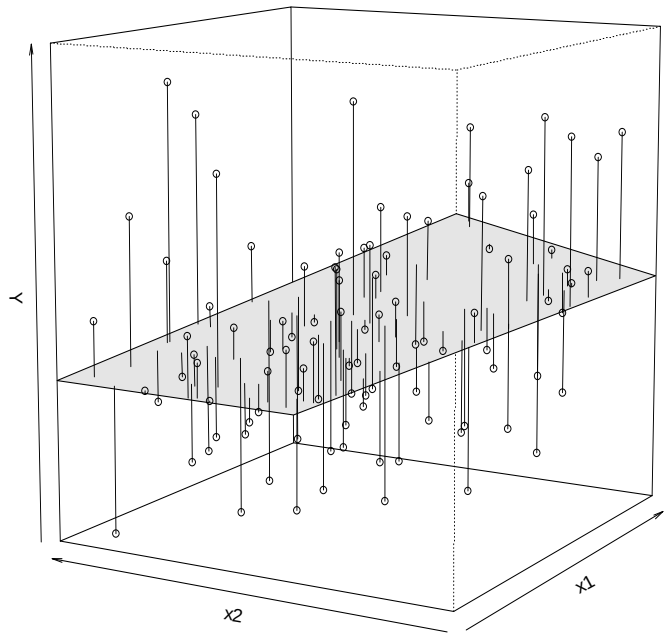
- How to express the full rank condition in layman terms?
- What happens when this condition fails?

Linear least squares

Least squares is used to estimate the beta coefficients as the solution of an (quadratic) optimization problem:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 .$$

The ordinary least squares (OLS) defined by this criterion finds the plane in R^{m+1} that minimizes the sum of squared vertical distances (along the y-axis). This geometric illustration will be referred to as the observation-space geometry.



Observation-space geometry of the OLS ($m = 2$). The least squares plane minimizes the sum of squared vertical distances between observation Y_i and the projected point (fitted value).

The function to minimize is quadratic in $\boldsymbol{\beta}$ and therefore admits a closed-form solution. Define the *residual sum of squares (RSS)* as a function of $\boldsymbol{\beta}$:

$$S(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2, \quad \boldsymbol{\beta} \in R^m.$$

The necessary condition for the minimum is that the derivative at the solution is zero,

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}.$$

The system of m equations, written in vector form as

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0},$$

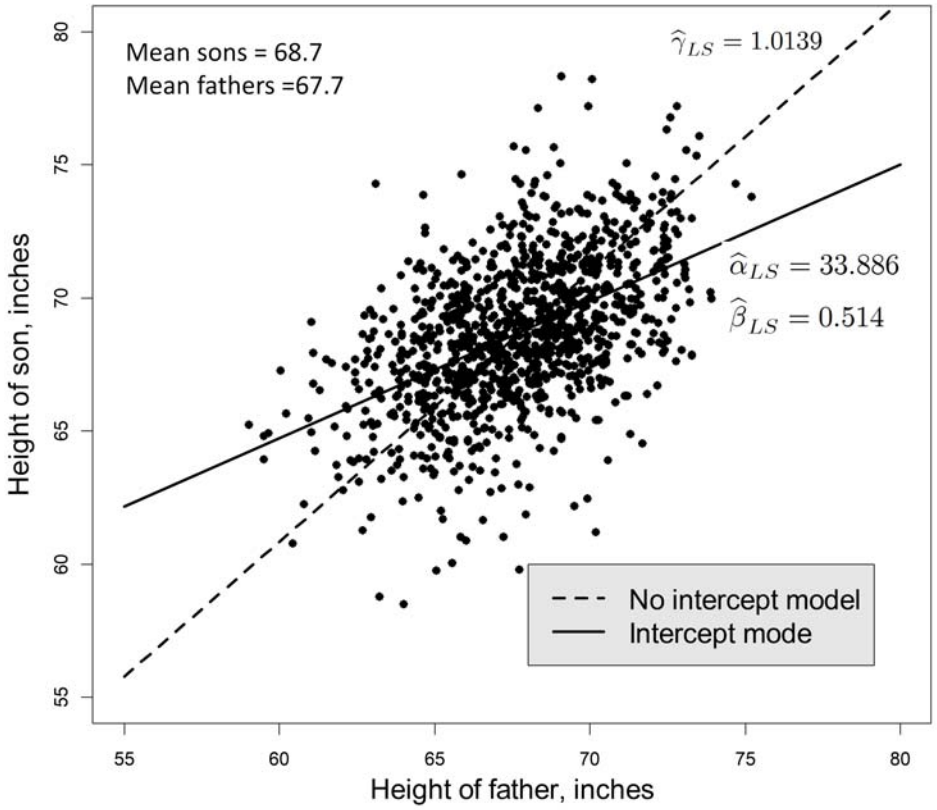
is called the *normal equation*. "Normal" here has nothing to do with the normal distribution, see the explanation below. In the geometric language, we write $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} \perp \mathbf{X}$: vector $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is orthogonal to each vector column of matrix $\mathbf{X}$. Rewrite the normal equation as $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}$ and come to the ordinary least squares (OLS) estimator as the solution to the optimization problem (**??**)

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Note that the condition on full rank is crucial here because it implies that matrix $\mathbf{X}'\mathbf{X}$ is nonsingular.

Components of vector $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ are called the *fitted (or predicted) values*. Vector $\mathbf{y} - \widehat{\mathbf{y}}$ is called the residual vector, and its components, $Y_i - \widehat{Y}_i$, are called least squares *residuals*. In the above figure $\widehat{Y}_i$ is the height of the projected point and the residual, $Y_i - \widehat{Y}_i$, is the difference between the observation and the fitted value.

Regression to the mean or why "regression?"



Galton height data: the height of sons versus height of fathers ($n = 1078$) is fitted with intercept (solid line) and without intercept (dotted line). This phenomenon is called "regression to the mean."

Regression to the mean: heights of fathers and sons. Sir Francis Galton (1822–1911) was a British scientist who studied the relationship between the heights of fathers and sons in his famous paper "Regression Towards Mediocrity in Hereditary Stature" published in 1885. Taller fathers have shorter sons and vice versa, or as Sir Galton put in "Towards Mediocrity."

`lm` function in R

See the R code `kidsdrink` from Section 8.6. with interpretation of the coefficients on the log scale.

```
Call:
lm(formula = logalcm ~drink + age + boy + race + alcbr + pared +
 inc + grade, data = d)
Residuals:
 Min 1Q Median 3Q Max
-10.1467 -0.3481 0.0987 0.4663 2.1806
Coefficients:
 Estimate Std.  Error t value Pr(>|t|)
(Intercept) -0.093682 0.133021 -0.704 0.481314
drink 0.432670 0.028335 15.270 < 2e-16 ***
age 0.137065 0.009459 14.490 < 2e-16 ***
boy 0.048303 0.024298 1.988 0.046895 *
race 0.266762 0.045818 5.822 6.29e-09 ***
alcbr 0.267472 0.040170 6.658 3.16e-11 ***
pared -0.022162 0.027022 -0.820 0.412198
inc 0.084440 0.037931 2.226 0.026063 *
grade -0.090740 0.025978 -3.493 0.000483 ***
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
Residual standard error:  0.7416 on 3796 degrees of freedom
Multiple R-squared:  0.1969, Adjusted R-squared:  0.1952
F-statistic:  116.3 on 8 and 3796 DF, p-value:  < 2.2e-16
```

The vector-space geometry and coefficient of determination

Section 8.1.2.

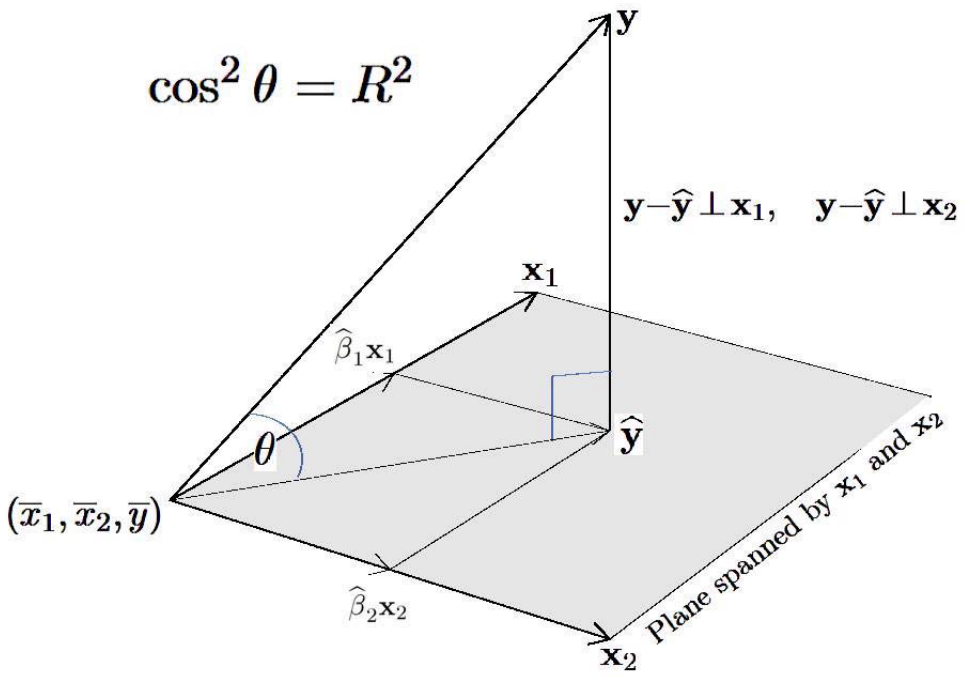
As was mentioned above, typically, a linear model contains an intercept term. Since we are mostly interested in the slope coefficients, the intercept term can be eliminated through averaging. Thus in this section, when referring to $\mathbf{y}$ or $\mathbf{x}_j$ we mean $\mathbf{y}-\overline{Y}\mathbf{1}$ and $\mathbf{x}_j - \overline{x}_j\mathbf{1}$.

θ is the angle between the vector of the dependent variable and the plane spanned by a set of predictors/independent variables.

R^2 is called the coefficient of determination

$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ is the vector of fitted values/projection on the plane spanned by $\mathbf{X}$.

$\mathbf{r} = \mathbf{y}-\widehat{\mathbf{y}}$ is the residual vector.



The vector-space geometry of the OLS. The least squares solution finds the point on the plane closest to $\mathbf{y}$. *This point,* $\widehat{\mathbf{y}}$, *is the base of the perpendicular (projection) dropped from* $\mathbf{y}$ *onto the plane. A smaller angle* θ *indicates a better regression.*

Pythagorean theorem of regression analysis. The following decomposition of the sum of squares holds.

$$\|\mathbf{y}\|^2 = \|\mathbf{y}-\widehat{\mathbf{y}}\|^2 + \|\widehat{\mathbf{y}}\|^2 .$$

Customary interpretation of the **coefficient of determination** (squared correlation coefficient between the dependent variable and fitted values): the proportion of variance of the dependent variable explained by predictors.

$$R^2 = 1 - \frac{\|\mathbf{r}\|^2}{\|\mathbf{y}-\overline{y}\mathbf{1}\|^2}.$$

Coefficient of determination typically increases with addition of a new predictor, but never decreases. Why?

Theorem 1 *The regression coefficients in multiple linear regression can be obtained from separate simple regressions if* ***predictors are uncorrelated (orthogonal).***

Demonstrate this fact geometrically.

Efficient estimation in statistics

Let $Y_1, Y_2, ..., Y_n$ is a random sample from a population with mean μ, that is, $E(Y_i) = \mu$, $var(Y_i) = \sigma^2$ and Y_i and Y_j are uncorrelated for every $i \neq j$. Then

$$\widehat{\mu} = \overline{Y}$$

is (a) unbiased and (b) has minimum variance among all unbiased linear estimators of μ, that is, if

$$\widetilde{\mu} = \sum_{i=1}^{n} \lambda_i Y_i$$

is an unbiased estimator then

$$var(\widetilde{\mu}) \geq var(\widehat{\mu}).$$

Proof. We have

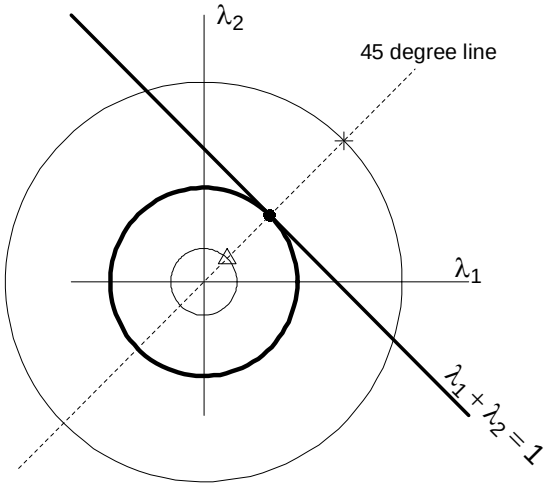
$$\begin{aligned} E(\widetilde{\mu}) &= \sum_{i=1}^{n} \lambda_i E(Y_i) = \mu \sum_{i=1}^{n} \lambda_i, \\ var(\widetilde{\mu}) &= \sum_{i=1}^{n} \lambda_i^2 var(Y_i) = \sigma^2 \sum_{i=1}^{n} \lambda_i^2. \end{aligned}$$

Finding optimal $\lambda_1, ..., \lambda_n$ is equivalent to solving the optimization problem

$$\min_{\sum_{i=1}^{n} \lambda_i = 1} \sum_{i=1}^{n} \lambda_i^2.$$

The answer: $\lambda_i = const$. Thus

$$\lambda_i = \frac{1}{n}\text{: arithmetic mean.}$$



Gauss-Markov theorem

Section 8.2

Theorem 2 *(a) The OLS estimator*

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

is unbiased, (b) with the covariance matrix of the OLS estimator is given by

$$\text{cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

(c) OLS estimator is efficient among all linear unbiased estimators of $\boldsymbol{\beta}$.

Proof. Prove that the OLS estimator is unbiased using the fact that $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$,

$$E(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Find the covariance matrix using the fact that $cov(\mathbf{y}) = \sigma^2\mathbf{I}$ and $cov(\mathbf{Au}) = \mathbf{A}cov(\mathbf{u})\mathbf{A}'$,

$$\begin{aligned} cov(\widehat{\boldsymbol{\beta}}) &= cov((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'cov(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Remarks:

- "Unbiased" means that $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$:
- "Efficient" means that if $\widetilde{\boldsymbol{\beta}} = \mathbf{Ly}$ is an unbiased linear estimator then

  $$cov(\widetilde{\boldsymbol{\beta}}) \geq cov(\widehat{\boldsymbol{\beta}}),$$

  where inequality means that the difference between the left- and right-hand side is a nonnegative definite matrix.

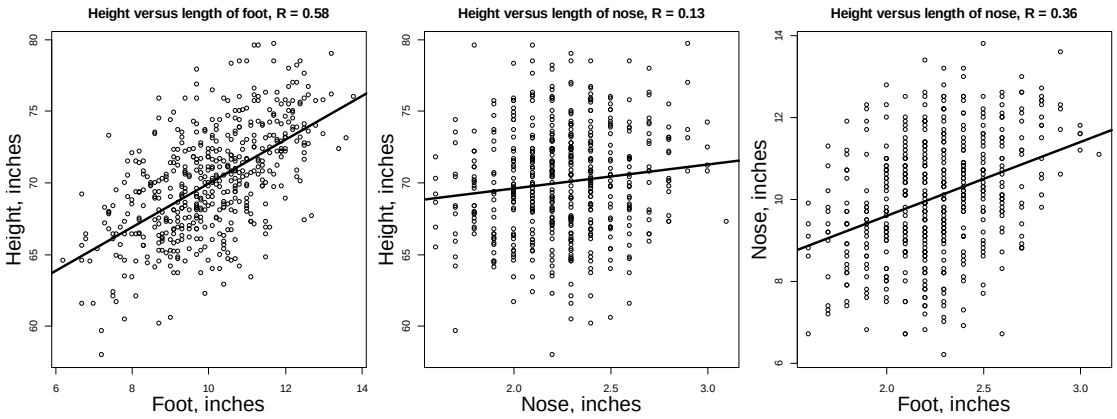
Example. How to find the most efficient estimator of the intercept where the slope is treated as a nuisance parameter (not important)?

Example. How to prove that $\widetilde{\beta} = y_k/x_k$ where $x_k = \max_{i=1,\ldots,n} x$ is not an good as the OLS estimator of the slope $\widehat{\beta} = \mathbf{y}'\mathbf{x}/\left\|\mathbf{x}\right\|^2$?

Interpretation of regression coefficients

Example 8.6.3.

File `HeightFootNose.csv` contains measurements of 497 people's height and length of foot and nose (in inches); the `R` code is found in file `hfn.r`. The scatterplots `Height` versus `Foot` and `Nose` are shown below.



Although the regressions of ***`Height`*** *on* ***`Foot`*** *and* ***`Nose`*** *have positive slopes (the first two plots) the slope at* ***`Nose`*** *in the birvariate regression is negative because* ***`Foot`*** *and* ***`Nose`*** *correlate (the third plot).*

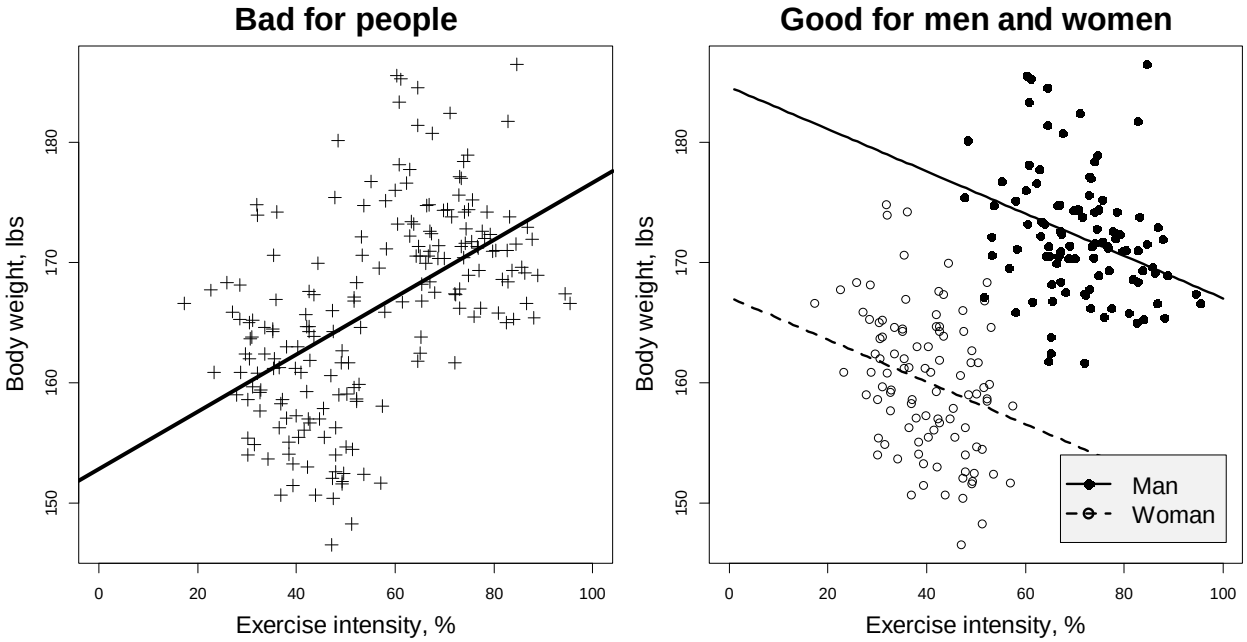
```
lm(formula = Height ~ Foot + Nose, data = da)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  56.6307     1.2436  45.537   <2e-16 ***
Foot          1.6140     0.1017  15.874   <2e-16 ***
Nose         -1.2499     0.5125  -2.439   0.0151 *
Residual standard error: 3.048 on 494 degrees of freedom
Multiple R-squared:  0.3486,    Adjusted R-squared:  0.346
F-statistic: 132.2 on 2 and 494 DF,  p-value: < 2.2e-16
```

Does it mean that shorter people have longer noses?

Simpson's paradox

Example 8.44. Good for men and women, bad for people

See R code `simpson`



Simpson paradox: good for men and women and bad for people. Overlooking clusters may lead to erroneous results. If gender is not taken into account exercise increases body weight (left). If gender is represented by a dummy variable exercising decreases body weight (right).

Modeling

$$B_i = \alpha + \alpha_m M_i + \beta E_i + \varepsilon_i$$

where

- B_i = body weight of the ith person
- M_i is a binary (dummy) variable: $M_i = 1$ if man and $M_i = 0$ if woman
- E_i = exercise
- $\alpha + \alpha_m$ = intercept for men
- α = intercept for women
- β is the common slope: 1% increase of exersize reduces β% of weight

Homework 3

1. (10 points). Display the time of watching alcohol scenes as a function of age for a black girl who drinks, has an alcohol related item, with high income and high parents' education, and has good grades as in function `kidsdrink(job=2)`. To contrast, display the same girl but who does not drink and does not have an alcohol related item. Compute, interpret, and display the effect of drinking and having an alcohol related item.

2. (10 points). How to find the most efficient estimator of the intercept in the multiple regression where the slopes are treated as a nuisance parameters (not a subject of interest)? Provide a proof.

3. (10 points). Provide a geometric interpretation of the following fact: the coefficient of determination does not change if the new predictor is orthogonal to the previous set of predictors and the residual vector from the previous regression.

4. (10 points). Plot residuals from regression of Height on Foot versus residuals from regression of Nose on Foot. Explain the result in connection to the negative slope in the multiple regression of Height on Foot and Nose.