San José State University

Math 261A: Regression Theory & Methods

Multiple Linear Regression

Dr. Guangliang Chen

This lecture is based on the following textbook sections:

- **Chapter 3: 3.1 - 3.5, 3.8 - 3.10**

Outline of this presentation:

- The multiple linear regression problem
- Least-square estimation
- Inference
- Some issues

The multiple linear regression problem

Consider the body data again. To construct a more accurate model for predicting the weight of an individual (y), we may want to add other body measurements, such as head and waist circumferences, as additional predictors besides height (x_1), leading to multiple linear regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad (1)$$

where

- y: response, $x_1, \ldots, x_k$: predictors
- $\beta_0, \beta_1, \ldots, \beta_k$: coefficients
- ϵ: error term

An example of a regression model with $k = 2$ predictors

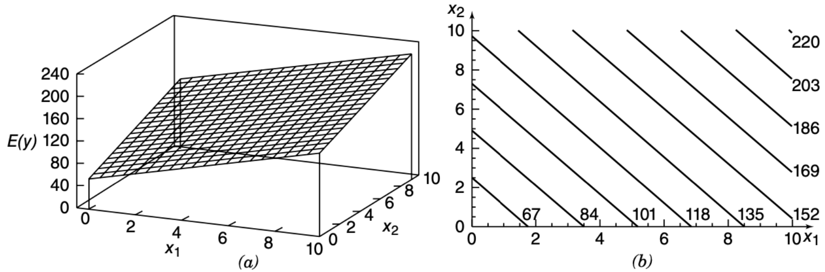


Figure 3.1 (*a*) The regression plane for the model $E(y) = 50 + 10x_1 + 7x_2$. (*b*) The contour plot.

Remark. Some of the new predictors in the model could be powers of the original ones

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$$

or interactions of them,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

or even a mixture of powers and interactions of them

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

These are still linear models (in terms of the regression coefficients).

An example of a full quadratic model

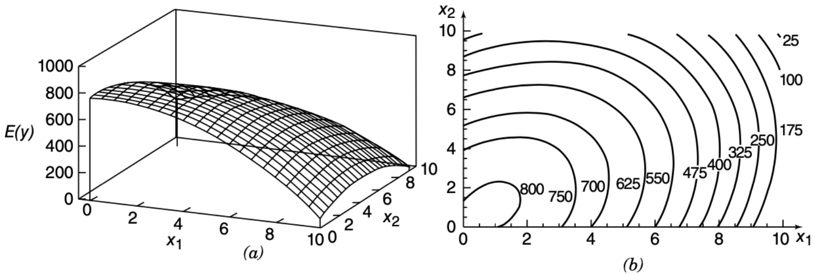


Figure 3.3 (*a*) Three-dimensional plot of the regression model $E(y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2$, (*b*) The contour plot.

The sample version of (1) is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad 1 \le i \le n \tag{2}$$

where the ϵ_i are assumed for now to be uncorrelated:

$$\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \quad i \neq j$$

and have the same mean zero and variance σ^2:

$$\mathrm{E}(\epsilon_i) = 0, \quad \mathrm{Var}(\epsilon_i) = \sigma^2, \quad \text{for all } i$$

(Like in simple linear regression, we will add the normality and independence assumptions when we get to the inference part)

Letting

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

we can rewrite the sample regression model in matrix form

$$\underbrace{\mathbf{y}}_{n\times 1} = \underbrace{\mathbf{X}}_{n\times p} \cdot \underbrace{\boldsymbol{\beta}}_{p\times 1} + \underbrace{\boldsymbol{\epsilon}}_{n\times 1} \tag{3}$$

where $p = k + 1$ represents the number of regression parameters (note that k is the number of predictors in the model).

Least squares (LS) estimation

The LS criterion can still be used to fit a multiple regression model

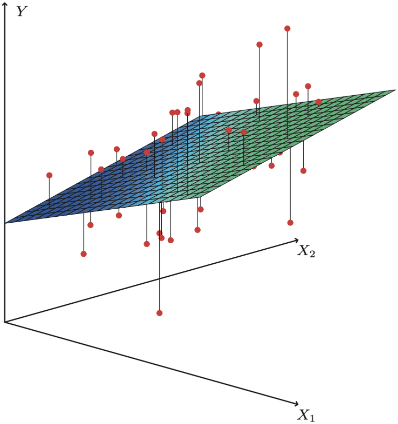
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

to the data as follows:

$$\min_{\hat{\boldsymbol{\beta}}} S(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

where for each $1 \le i \le n$,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$$



Let $\mathbf{e} = (e_i) \in \mathbb{R}^n$ and $\hat{\mathbf{y}} = (\hat{y}_i) = \mathbf{X}\hat{\boldsymbol{\beta}} \in \mathbb{R}^n$. Then $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$. Correspondingly the above problem becomes

$$\min_{\hat{\boldsymbol{\beta}}} S(\hat{\boldsymbol{\beta}}) = \|\mathbf{e}\|^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$$

Theorem 0.1. If $\mathbf{X}'\mathbf{X}$ is nonsingular, then the LS estimator of β is

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

Remark. The nonsingular condition holds true if and only if all the columns of $\mathbf{X}$ are linearly independent (i.e. $\mathbf{X}$ is of full column rank).

Remark. This is the same formula for $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)'$ in simple linear regression. To demonstrate it, consider the toy data set of 3 points: $(0,1),(1,0),(2,2)$ used before. The new formula gives that

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \left(\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \\
&= \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ 4 \end{bmatrix} \\
&= \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}
\end{aligned}
$$

Proof. We first need to derive some formulas about the gradient of a function of multiple variables:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{x}'\mathbf{a}\right) &= \frac{\partial}{\partial \mathbf{x}}\left(\mathbf{a}'\mathbf{x}\right) = \mathbf{a} \\
\frac{\partial}{\partial \mathbf{x}}\left(\|\mathbf{x}\|^2\right) &= \frac{\partial}{\partial \mathbf{x}}\left(\mathbf{x}'\mathbf{x}\right) = 2\mathbf{x} \\
\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{x}'\mathbf{A}\mathbf{x}\right) &= 2\mathbf{A}\mathbf{x} \\
\frac{\partial}{\partial \mathbf{x}}\left(\|\mathbf{B}\mathbf{x}\|^2\right) &= \frac{\partial}{\partial \mathbf{x}}\left(\mathbf{x}'\mathbf{B}'\mathbf{B}\mathbf{x}\right) = 2\mathbf{B}'\mathbf{B}\mathbf{x}
\end{aligned}$$

Using the identity $\|\mathbf{u}-\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\mathbf{u}'\mathbf{v}$, we write

$$\begin{aligned} S(\hat{\boldsymbol{\beta}}) &= \|\mathbf{y}\|^2 + \|\mathbf{X}\hat{\boldsymbol{\beta}}\|^2 - 2(\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{y} \\ &= \mathbf{y}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \end{aligned}$$

Applying the formulas on the preceding slide, we obtain

$$\frac{\partial S}{\partial \hat{\boldsymbol{\beta}}} = 0 + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - 2\mathbf{X}'\mathbf{y}$$

Setting the gradient equal to zero

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \;\longleftarrow \text{least squares normal equations}$$

and solving for $\hat{\boldsymbol{\beta}}$ will complete the proof. $\square$

Remark. The very first normal equation in the system

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

is

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum x_{i1} + \hat{\beta}_2 \sum x_{i2} + \cdots + \hat{\beta}_k \sum x_{ik} = \sum y_i$$

which simplifies to

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \cdots + \hat{\beta}_k \bar{x}_k = \bar{y}$$

This indicates that the centroid of the data, i.e., $(\bar{x}_1, \ldots, \bar{x}_k, \bar{y})$, is on the least squares regression plane.

Remark. The fitted values of the least squares model are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{H}}\mathbf{y} = \mathbf{H}\mathbf{y}$$

and the residuals are

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

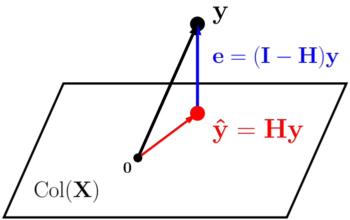
The matrix $\mathbf{H} \in \mathbb{R}^{n\times n}$ is called **the hat matrix**, satisfying

$$\mathbf{H}' = \mathbf{H} \text{ (symmetric)}, \quad \mathbf{H}^2 = \mathbf{H} \text{ (idempotent)}, \quad \mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{O}$$

Geometrically, it is the orthogonal projection matrix onto the column space of $\mathbf{X}$ (subspace spanned by the columns of $\mathbf{X}$):

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} = \mathbf{X}\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}_{\hat{\boldsymbol{\beta}}} \in \mathrm{Col}(\mathbf{X})$$

$$\hat{\mathbf{y}}'(\mathbf{y}-\hat{\mathbf{y}}) = (\mathbf{H}\mathbf{y})'(\mathbf{I}-\mathbf{H})\mathbf{y} = \mathbf{y}'\underbrace{\mathbf{H}(\mathbf{I}-\mathbf{H})}_{=\mathbf{O}}\mathbf{y} = 0.$$



Example 0.1 (body dimensions data[1])**.** Besides the predictor *Height*, we include *Waist Girth* as a second predictor to preform multiple linear regression for predicting *Weight*.

(R demonstration in class).

[1] `http://jse.amstat.org/v11n2/datasets.heinz.html`

Inference in multiple linear regression

- **Model parameters**: $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ (intercept and slopes), σ^2 (noise variance)

- **Inference tasks** (for the parameters above): point estimation, interval estimation*, hypothesis testing*

- **Inference of the mean response** at $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0k})'$:

$$\mathrm{E}(y \mid \mathbf{x}_0) = \beta_0 + \beta_1 x_{01} + \cdots + \beta_k x_{0k} = \mathbf{x}_0' \boldsymbol{\beta}$$

**To perform these two inference tasks, we will additionally assume that the model errors ϵ_i are normally and independently distributed with mean 0 and variance σ^2, i.e., $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$.*

Expectation and variance of a vector-valued random variable

Let $\vec{X} = (X_1, \ldots, X_n)' \in \mathbb{R}^n$ be a vector-valued random variable. Define

- **Expectation**: $\mathrm{E}(\vec{X}) = (\mathrm{E}(X_1, \ldots, \mathrm{E}(X_n))'$
- **Variance** (also called covariance matrix):

$$\mathrm{Var}(\vec{X}) = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Var}(X_n) \end{bmatrix}$$

Point estimation in multiple linear regression

First, like in simple linear regression, the least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased linear estimator for $\boldsymbol{\beta}$.

Theorem 0.2. Under the assumptions of multiple linear regression,

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$$

That is, $\hat{\boldsymbol{\beta}}$ is a (componentwise) unbiased estimator for $\boldsymbol{\beta}$:

$$\mathrm{E}(\hat{\beta}_i) = \beta_i, \quad \text{for all } \ i = 0, 1, \ldots, k$$

Proof. We have

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\cdot\mathbf{X}\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\cdot\boldsymbol{\epsilon} \\ &= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}.\end{aligned}$$

It follows that

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underbrace{\mathrm{E}(\boldsymbol{\epsilon})}_{=\mathbf{0}} = \boldsymbol{\beta}$$

$\square$

Next, we derive the variance of $\hat{\boldsymbol{\beta}}$:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j))_{0 \le i,j \le k}.$$

Theorem 0.3. Let $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} = (C_{ij})_{0 \le i,j \le k}$. Then

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{C}.$$

That is,

$$\mathrm{Var}(\hat{\beta}_i) = \sigma^2 C_{ii} \qquad \text{and} \quad \mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}.$$

Proof. Using the formula:

$$\mathrm{Var}(\mathbf{A}\mathbf{y}) = \mathbf{A} \cdot \mathrm{Var}(\mathbf{y}) \cdot \mathbf{A}',$$

we have

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}(\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{A}}\mathbf{y}) \\
&= \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{A}} \cdot \underbrace{\mathrm{Var}(\mathbf{y})}_{=\sigma^2\mathbf{I}} \cdot \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}}_{\mathbf{A}'} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}$$

Lastly, we can derive an estimator of σ^2 from the residual sum of squares

$$SS_{Res} = \sum e_i^2 = \|\mathbf{e}\|^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$$

Theorem 0.4. We have

$$\mathrm{E}(SS_{Res}) = (n-p)\sigma^2.$$

This implies that

$$MS_{Res} = \frac{SS_{Res}}{n-p}$$

is an unbiased estimator of σ^2.

Remark. The total and regression sums of squares are defined in the same way as before:

$$SS_R = \sum(\hat{y}_i - \bar{y})^2 = \sum \hat{y}_i^2 - n\bar{y}^2 = \|\hat{\mathbf{y}}\|^2 - n\bar{y}^2$$
$$SS_T = \sum(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = \|\mathbf{y}\|^2 - n\bar{y}^2$$

They can be used to assess the adequacy of the model through the coefficient of determination

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

The larger R^2 (i.e., the smaller SS_{Res}), the better the model.

Example 0.2 (Weight $\sim$ Height + Waist Girth). For this model,

$$MS_{Res} = 4.529^2 = 20.512$$

In contrast, for the simple linear regression model (Weight $\sim$ Height),

$$MS_{Res} = 9.308^2 = 86.639.$$

Therefore, the multiple linear regression model has a smaller total fitting error $SS_{Res} = (n-p)MS_{Res}$.

```
> mymodel2<-lm(weight~height+waist_girth, data=mydata )
> summary(mymodel2)

Call:
lm(formula = weight ~ height + waist_girth, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8643  -2.8947  -0.1823   2.5674  20.6156

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -75.07047    3.74259  -20.06   <2e-16 ***
height        0.44432    0.02569   17.30   <2e-16 ***
waist_girth   0.88563    0.02194   40.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 4.529 on 504 degrees of freedom
Multiple R-squared:  0.8853,    Adjusted R-squared:  0.8848
F-statistic:  1945 on 2 and 504 DF,  p-value: < 2.2e-16
```

The coefficient of determination of this model is $R^2 = 0.8853$, which is much higher than that of the smaller model.

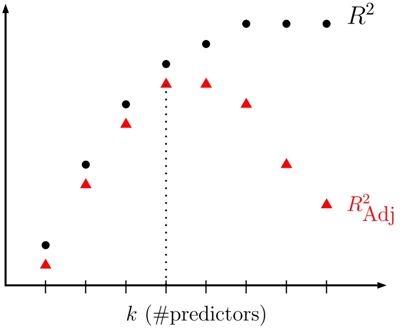
Adjusted R^2

R^2 measures the goodness of fit of a single model and is not a fair criterion for comparing models with different sizes k (e.g., nested models)

The adjusted R^2 criterion is more suitable for such comparisons:

$$R^2_{\text{Adj}} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)}$$

The larger the R^2_{Adj}, the better the model.



Remark.

- As p (i.e., k) increases, SS_{Res} will either decrease or stay the same:
  - If SS_{Res} does not change (or decreases by very little), then R^2_{Adj} will decrease. $\longleftarrow$ The smaller model is better
  - If SS_{Res} decreases relatively more than $n-p$ does, then R^2_{Adj} would increase. $\longleftarrow$ The larger model is better
- We can write instead

$$R^2_{\text{Adj}} = 1 - \frac{n-1}{n-p}(1-R^2)$$

This implies that $R^2_{\text{Adj}} < R^2$.

Summary: Point estimation in multiple linear regression

Model parameters	Point estimators	Properties: Bias	Properties: Variance
$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	unbiased	$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
σ^2	$MS_{Res} = \frac{SS_{Res}}{n-p}$	unbiased	

Remark. For the mean response at $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0k})'$:

$$\mathrm{E}(y \mid \mathbf{x}_0) = \beta_0 + \beta_1 x_{01} + \cdots + \beta_k x_{0k} = \mathbf{x}_0'\boldsymbol{\beta}$$

an unbiased point estimator is

$$\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_k x_{0k} = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$$

Next

We consider the following inference tasks in multiple linear regression:

- **Hypothesis testing**
- **Interval estimation**

For both tasks, we need to additionally assume that the model errors ϵ_i are iid $N(0, \sigma^2)$.

Hypothesis testing in multiple linear regression

Depending on how many regression coefficients are being tested together, we have

- ANOVA F Tests for Significance of Regression on All Regression Coefficients
- Partial F Tests on Subsets of Regression Coefficients
- Marginal t Tests on Individual Regression Coefficients

ANOVA for Testing Significance of Regression

In multiple linear regression, the significance of regression test is

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$
$$H_1 : \beta_j \neq 0 \text{ for at least one } j$$

The ANOVA test works very similarly: The test statistic is

$$F_0 = \frac{MS_R}{MS_{Res}} = \frac{SS_R/k}{SS_{Res}/(n-p)} \overset{H_0}{\sim} F_{k,n-p}$$

and we reject H_0 if

$$F_0 > F_{\alpha,k,n-p}$$

Example 0.3 (Weight $\sim$ Height + Waist Girth). For this multiple linear regression model, regression is significant because the ANOVA F statistic is

$$F_0 = 1945$$

and the p-value is less than 2.2e-16.

Note that the p-values of the individual coefficients can no longer be used for conducting the significance of regression test.

```
> mymodel2<-lm(weight~height+waist_girth, data=mydata )
> summary(mymodel2)

Call:
lm(formula = weight ~ height + waist_girth, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.8643  -2.8947  -0.1823   2.5674  20.6156 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -75.07047    3.74259  -20.06   <2e-16 ***
height        0.44432    0.02569   17.30   <2e-16 ***
waist_girth   0.88563    0.02194   40.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 4.529 on 504 degrees of freedom
Multiple R-squared:  0.8853,	Adjusted R-squared:  0.8848 
F-statistic:  1945 on 2 and 504 DF,  p-value: < 2.2e-16
```

Marginal Tests on Individual Regression Coefficients

The hypothesis for testing the significance of any individual predictor x_j, given all the other predictors, to the model is

$$H_0 : \beta_j = 0 \qquad \text{vs} \quad H_1 : \beta_j \neq 0$$

If H_0 is not rejected, then the regressor x_j is insignificant and can be deleted from the model (while preserving all other regressors).

To conduct the test, we need to use the point estimator $\hat{\beta}_j$ (which is linear, unbiased) and determine its distribution when H_0 is true:

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 C_{jj}), \quad j = 0, 1, \ldots, k$$

The test statistic is

$$t_0 = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \overset{H_0}{\sim} t_{n-p} \qquad (\hat{\sigma}^2 = MS_{Res})$$

and we reject H_0 if

$$|t_0| > t_{\alpha/2,\, n-p}$$

Example 0.4 (Weight $\sim$ Height $+$ Waist Girth)**.** Based on the previous R output, both predictors are significant when the other is already included in the model:

- Height: $t_0 = 17.30$, p-value $<$ 2e-16
- Waist Girth: $t_0 = 40.36$, p-value $<$ 2e-16

Partial F Tests on Subsets of Regression Coefficients

Consider the full regression model with k regressors

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

Suppose there is a partition of the regression coefficients in $\boldsymbol{\beta}$ into two groups (the last r and the preceding ones):

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \in \mathbb{R}^p, \quad \boldsymbol{\beta}_1 = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-r} \end{bmatrix} \in \mathbb{R}^{p-r}, \; \boldsymbol{\beta}_2 = \begin{bmatrix} \beta_{k-r+1} \\ \vdots \\ \beta_k \end{bmatrix} \in \mathbb{R}^r$$

We wish to test

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0} \ (\beta_{k-r+1} = \cdots = \beta_k = 0) \quad \text{vs} \quad H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$$

to determine if the last r predictors may be deleted from the model.

Corresponding to the partition of $\boldsymbol{\beta}$ we partition $\mathbf{X}$ in a conformal way:

$$\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2], \qquad \mathbf{X}_1 \in \mathbb{R}^{n\times(p-r)}, \quad \mathbf{X}_2 \in \mathbb{R}^{n\times r},$$

such that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = [\mathbf{X}_1 \ \mathbf{X}_2] \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

We compare two contrasting models:

$$\begin{aligned} &\text{(Full model)} \quad && \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &\text{(Reduced model)} \quad && \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon} \end{aligned}$$

The corresponding regression sums of squares are

$$\begin{aligned} (df = k) \quad & SS_R(\boldsymbol{\beta}) = \|\mathbf{X}\hat{\boldsymbol{\beta}}\|^2 - n\bar{y}^2, \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ (df = k - r) \quad & SS_R(\boldsymbol{\beta}_1) = \|\mathbf{X}_1\hat{\boldsymbol{\beta}}_1\|^2 - n\bar{y}^2, \quad \hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} \end{aligned}$$

Thus, the regression sum of squares due to $\boldsymbol{\beta}_2$ given that $\boldsymbol{\beta}_1$ is already in the model, called **extra sum of squares**, is

$$(df = r) \qquad SS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_1)$$

Note that with the residual sums of squares

$$SS_{Res}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2, \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$
$$SS_{Res}(\boldsymbol{\beta}_1) = \|\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1\|^2, \quad \hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$$

we also have

$$SS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) = SS_{Res}(\boldsymbol{\beta}_1) - SS_{Res}(\boldsymbol{\beta})$$

Finally, the (partial F) test statistic is

$$F_0 = \frac{SS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)/r}{SS_{Res}(\boldsymbol{\beta})/(n-p)} \overset{H_0}{\sim} F_{r,n-p}$$

and we reject H_0 if

$$F_0 > F_{\alpha,r,n-p}$$

Example 0.5 (Weight $\sim$ Height + Waist Girth). We use the extra sum of squares method to compare it with the reduced model (Weight $\sim$ Height):

```
> mymodel1<-lm(weight~height, data=mydata)
> mymodel2<-lm(weight~height+waist_girth, data=mydata )
> anova(mymodel1, mymodel2)
Analysis of Variance Table

Model 1: weight ~ height
Model 2: weight ~ height + waist_girth
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    505 43753
2    504 10337  1     33416 1629.2 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remark. The partial F test on a single predictor x_j, $\boldsymbol{\beta} = [\boldsymbol{\beta}_{(j)}; \beta_j]$ based on the extra sum of squares

$$SS_R(\beta_j \mid \boldsymbol{\beta}_{(j)}) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_{(j)})$$

can be shown to be equivalent to the marginal t test for β_j.

For example, for Waist Girth,

- marginal t test: $t_0 = 40.36$
- partial F test: $F_0 = 1629.2$

Note that $F_0 = t_0^2$ (thus same test).

```
> mymodel2<-lm(weight~height+waist_girth, data=mydata )
> summary(mymodel2)

Call:
lm(formula = weight ~ height + waist_girth, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.8643  -2.8947  -0.1823   2.5674  20.6156 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -75.07047    3.74259  -20.06   <2e-16 ***
height        0.44432    0.02569   17.30   <2e-16 ***
waist_girth   0.88563    0.02194   40.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 4.529 on 504 degrees of freedom
Multiple R-squared:  0.8853,	Adjusted R-squared:  0.8848 
F-statistic:  1945 on 2 and 504 DF,  p-value: < 2.2e-16
```

Remark. There is a decomposition of the regression sum of squares

$$SS_R \leftarrow SS_R(\beta_1, \ldots, \beta_k \mid \beta_0)$$

into a sequence of marginal extra sums of squares, each corresponding to a single predictor:

$$\begin{aligned} &SS_R(\beta_1, \ldots, \beta_k \mid \beta_0) \\ =&SS_R(\beta_1 \mid \beta_0) \\ &+ SS_R(\beta_2 \mid \beta_1, \beta_0) \\ &+ \cdots \\ &+ SS_R(\beta_k \mid \beta_{k-1}, \ldots, \beta_1, \beta_0) \end{aligned}$$

```
> mymodel2<-lm(weight~height+waist_girth, data=mydata )
> anova(mymodel2)
Analysis of Variance Table

Response: weight
             Df Sum Sq Mean Sq F value    Pr(>F)
height        1  46370   46370  2260.8 < 2.2e-16 ***
waist_girth   1  33416   33416  1629.2 < 2.2e-16 ***
Residuals   504  10337      21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
```

From the above output:

- $SS_R(\beta_1 \mid \beta_0) = 46370$, the predictor height is significant
- $SS_R(\beta_2 \mid \beta_1, \beta_0) = 33416$, waist girth is significant given that height is already in the model
- $SS_R(\beta_1, \beta_2 \mid \beta_0) = 79786$

Summary: hypothesis testing in regression

- ANOVA F test: $H_0 : \beta_1 = \cdots = \beta_k = 0$. Reject H_0 if

$$F_0 = \frac{MS_R}{MS_{Res}} = \frac{SS_R/k}{SS_{Res}/(n-p)} > F_{\alpha,k,n-p}$$

- Marginal t-tests: $H_0 : \beta_j = 0$. Reject H_0 if

$$|t_0| > t_{\alpha/2,\, n-p}, \quad t_0 = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$$

- Partial F test: $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$. Reject H_0 if

$$\frac{SS_R(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)/r}{SS_{Res}(\boldsymbol{\beta})/(n-p)} > F_{\alpha,r,n-p}$$

Interval estimation in multiple linear regression

We construct the following

- Confidence intervals for individual regression coefficients $\hat{\beta}_j$
- Confidence interval for the mean response
- Prediction interval

under the additional assumption that the errors ϵ_i are independently and normally distributed with zero mean and constant variance σ^2.

Confidence intervals for individual regression coefficients

Theorem 0.5. Under the normality assumption, a $1-\alpha$ confidence interval for the regression coefficient $\beta_j,\ 0 \le j \le k$ is

$$\hat{\beta}_j \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\, C_{jj}}$$

```
> confint(mymodel2, level=0.95)
                 2.5 %      97.5 %
(Intercept) -82.4234684 -67.7174691
height        0.3938544   0.4947853
waist_girth   0.8425226   0.9287393
```

Confidence interval for the mean response

In the setting of multiple linear regression, the mean response at a given point $\mathbf{x}_0 = (1, x_{01}, \dots, x_{0k})'$ is

$$\mathrm{E}(y \mid \mathbf{x}_0) = \mathbf{x}_0'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{01} + \cdots + \beta_k x_{0k}$$

A natural point estimator for $\mathrm{E}(y \mid \mathbf{x}_0)$ is the following:

$$\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_k x_{0k}.$$

Furthermore, we can construct a confidence interval for $\mathrm{E}(y \mid \mathbf{x}_0)$.

Since $\hat{y}_0$ is a linear combination of the responses, it is normally distributed with

$$\mathrm{E}(\hat{y}_0) = \mathbf{x}_0' \mathrm{E}(\hat{\boldsymbol{\beta}}) = \mathbf{x}_0' \boldsymbol{\beta}$$

and

$$\mathrm{Var}(\hat{y}_0) = \mathbf{x}_0' \mathrm{Var}(\hat{\boldsymbol{\beta}}) \mathbf{x}_0 = \sigma^2 \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0$$

We can thus obtain the following result.

Theorem 0.6. Under the normality assumption on the model errors, a $1-\alpha$ confidence interval on the mean response $\mathrm{E}(y \mid \mathbf{x}_0)$ is

$$\hat{y}_0 \pm t_{\alpha/2,\, n-p} \sqrt{\hat{\sigma}^2\, \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0}$$

Prediction intervals for new observations

Given a new location $\mathbf{x}_0$, we would like to form a prediction interval on the future observation of the response at that location

$$y_0 = \mathbf{x}_0'\boldsymbol{\beta} + \epsilon_0$$

where $\epsilon_0 \sim N(0, \sigma^2)$ is the error.

We have the following result.

Theorem 0.7. Under the normality assumption on the model errors, a $1-\alpha$ prediction interval for the future observation y_0 at the point $\mathbf{x}_0'$ is

$$\hat{y}_0 \pm t_{\alpha/2,\,n-p}\sqrt{\hat{\sigma}^2\left(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right)}$$

Proof. First, note that the mean of the response y_0 at $\mathbf{x}_0$, i.e., $\mathbf{x}_0'\boldsymbol{\beta}$, is estimated by $\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$.

Let $\Psi = y_0 - \hat{y}_0$ be the difference between the true response and the point estimator for its mean. Then Ψ (as a linear combination of $y_0, y_1, \ldots, y_n$) is normally distributed with mean

$$\Psi = \mathrm{E}(y_0) - \mathrm{E}(\hat{y}_0) = \mathbf{x}_0'\boldsymbol{\beta} - \mathbf{x}_0'\boldsymbol{\beta} = 0$$

and variance

$$\mathrm{Var}(\Psi) = \mathrm{Var}(y_0) + \mathrm{Var}(\hat{y}_0) = \sigma^2 + \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$$

It follows that

$$\frac{y_0 - \hat{y}_0}{\sqrt{\sigma^2 \left(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right)}} \sim N(0,1)$$

and correspondingly,

$$\frac{y_0 - \hat{y}_0}{\sqrt{MS_{Res} \left(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right)}} \sim t_{n-p}$$

Accordingly, a $1-\alpha$ prediction interval on a future observation y_0 at x_0 is

$$\hat{y}_0 \pm t_{\alpha/2,\, n-p}\sqrt{MS_{Res} \left(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right)}$$

Summary: interval estimation in regression

- β_j (for each $0 \le j \le k$): $\hat{\beta}_j \pm t_{\alpha/2,n-p}\sqrt{MS_{Res}\,C_{jj}}$
- σ^2: $\left(\frac{(n-p)MS_{Res}}{\chi^2_{\frac{\alpha}{2},n-p}}, \frac{(n-p)MS_{Res}}{\chi^2_{1-\frac{\alpha}{2},n-p}}\right)$
- $\mathrm{E}(y \mid \mathbf{x}_0)$: $\hat{y}_0 \pm t_{\alpha/2,\,n-p}\sqrt{MS_{Res}\,\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$
- y_0 (at $\mathbf{x}_0$): $\hat{y}_0 \pm t_{\alpha/2,\,n-p}\sqrt{MS_{Res}\,(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}$

Some issues in multiple linear regression

- Hidden extrapolation
- Units of measurements
- Multicollinearity

Hidden extrapolation

In multiple linear regression, extrapolation may occur even when all predictor values are within their ranges.

We can use the hat matrix

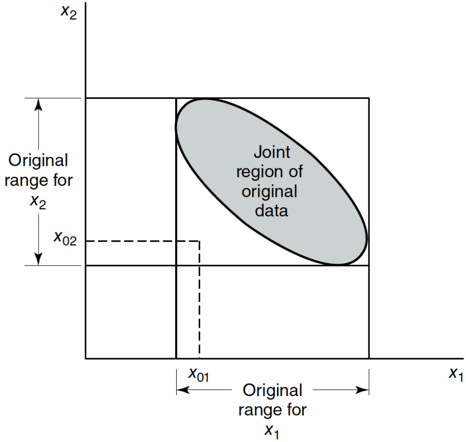
$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

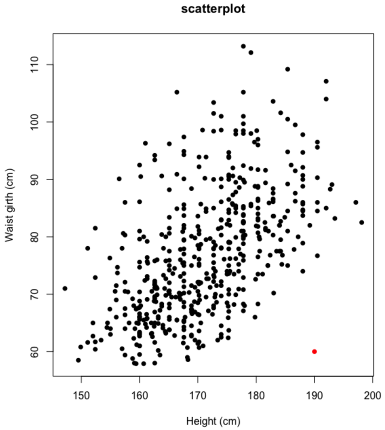
to detect hidden extrapolation: Let

$$h_{\max} = \max h_{ii}.$$

Then $\mathbf{x}_0$ is an extrapolation point if

$$\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 > h_{\max}$$






```
> hmax = max(hatvalues(mymodel2))
> hmax
[1] 0.02686508
> plot(mydata$height, mydata$waist_girth,
+       xlab="Height (cm)",
+       ylab="Waist girth (cm)",
+       pch=16, main="scatterplot")
> points(x=190,y=60, pch=16, col="red")
> X <- cbind(as.matrix(rep(1,507)),
+             mydata$height,
+             mydata$waist_girth)
> G = t(X)%*%X
> x0 <- as.matrix(c(1, 190, 60))
> t(x0)%*%solve(G)%*%x0
           [,1]
[1,] 0.02990657
```

Units of measurements

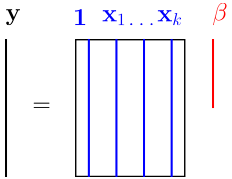
The choices of the units of the predictors in a linear model may cause their regression coefficients to have very different magnitudes, e.g.,

$$y = 3 - 20x_1 + 0.01x_2$$

In order to directly compare regression coefficients, we need to scale the regressors and the response to be on the same magnitude.

Two common scaling methods:

- **Unit Normal Scaling**
- **Unit Length Scaling**



Unit Normal Scaling: For each regressor x_j (and the response), rescale the observations of x_j (or $\mathbf{y}$) to have zero mean and unit variance.

Let

$$\bar{x}_j = \frac{1}{n}\sum_i x_{ij}, \quad s_j^2 = \frac{1}{n-1}\underbrace{\sum_i (x_{ij}-\bar{x}_j)^2}_{S_{jj}}, \quad s_y^2 = \frac{1}{n-1}\underbrace{\sum_i (y_i-\bar{y})^2}_{=SS_T}.$$

Then the normalized predictors and response are

$$z_{ij} = \frac{x_{ij}-\bar{x}_j}{s_j}, \quad y_i^* = \frac{y_i-\bar{y}}{s_y}$$

This leads to a linear regression model without intercept: $\mathbf{y}^* = \mathbf{Z}\hat{\mathbf{b}}$.

```
> # unit normal scaling
> mydata_std <- data.frame(scale(mydata))
> mynewmodel2 <- lm(weight~height+waist_girth, data=mydata_std)
> summary(mynewmodel2)

Call:
lm(formula = weight ~ height + waist_girth, data = mydata_std)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.11378 -0.21690 -0.01366  0.19238  1.54473 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -8.281e-16  1.507e-02    0.00        1    
height       3.132e-01  1.811e-02   17.30   <2e-16 ***
waist_girth  7.308e-01  1.811e-02   40.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3393 on 504 degrees of freedom
Multiple R-squared:  0.8853,	Adjusted R-squared:  0.8848 
F-statistic:  1945 on 2 and 504 DF,  p-value: < 2.2e-16
```

Unit Length Scaling: For each regressor x_j (and the response), rescale the observations of x_j (or $\mathbf{y}$) to have zero mean and unit length.

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{S_{jj}}} = \frac{z_{ij}}{\sqrt{n-1}}, \quad y_i^0 = \frac{y_i - \bar{y}}{\sqrt{SS_T}} = \frac{y_i^*}{\sqrt{n-1}}$$

This also leads to a linear regression model without intercept: $\mathbf{y}^0 = \mathbf{W}\hat{\mathbf{b}}$.

Remark.

- $\mathbf{W} = \frac{1}{\sqrt{n-1}}\mathbf{Z}$ and $\mathbf{y}^0 = \frac{1}{\sqrt{n-1}}\mathbf{y}^*$. Thus, the two scaling methods will yield the same standardized regression coefficients $\hat{\mathbf{b}}$.
- Entries of $\mathbf{W}'\mathbf{W}$ are correlations between the regressors.

Proof: We examine the (j,ℓ)-entry of $\mathbf{W}'\mathbf{W}$:

$$
\begin{aligned}
(\mathbf{W}'\mathbf{W})_{j\ell} &= \sum_{i=1}^{n} w_{ij} w_{i\ell} \\
&= \sum \frac{x_{ij}-\bar{x}_j}{\sqrt{S_{jj}}} \frac{x_{i\ell}-\bar{x}_\ell}{\sqrt{S_{\ell\ell}}} \\
&= \frac{\sum (x_{ij}-\bar{x}_j)(x_{i\ell}-\bar{x}_\ell)}{\sqrt{S_{jj}}\sqrt{S_{\ell\ell}}} \\
&= \frac{\frac{1}{n-1}\sum (x_{ij}-\bar{x}_j)(x_{i\ell}-\bar{x}_\ell)}{\sqrt{\frac{1}{n-1}\sum (x_{ij}-\bar{x}_j)^2}\sqrt{\frac{1}{n-1}\sum (x_{i\ell}-\bar{x}_\ell)^2}} \\
&= \mathrm{Corr}(x_j, x_\ell)
\end{aligned}
$$

Multicollinearity

A serious issue in multiple linear regression is multicolinearity, or near-linear dependence among the regression variables, e.g., $x_3 \approx 2x_1 + 5x_2$.

- $\mathbf{X}$ won't be of full rank, leading to a singular $\mathbf{X}'\mathbf{X}$.
- The redundant predictors contribute no new information about the response .
- The estimated slopes in the regression model will be arbitrary.

We will discuss in more detail how to diagnose (and fix) the issue of multicollinearity in Chapter 9.

Further learning

3.3.3 The Case of Orthogonal Columns in $\mathbf{X}$

3.3.4 Testing the General Linear Hypothesis $H_0 : \mathbf{T}\beta = \mathbf{0}$

- **Projection matrices**
  - Concepts
  - Computing via SVD