General Linear Regression Models

For $i = 1, \cdots n$:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i. \tag{1}$$

- Y_i : value of the response variable Y in the *ith* case.
- $X_{i1}, \cdots, X_{i,p-1}$: values of the variables $X_1, \cdots, X_{p-1}$ in the *ith* case.
- $\beta_0, \beta_1, \cdots, \beta_{p-1}$: regression coefficients.
  - p: the number of regression coefficients.
  - In simple regression $p = 2$.
- ϵ_i: error terms where $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.
- Response function (surface)/ mean response:

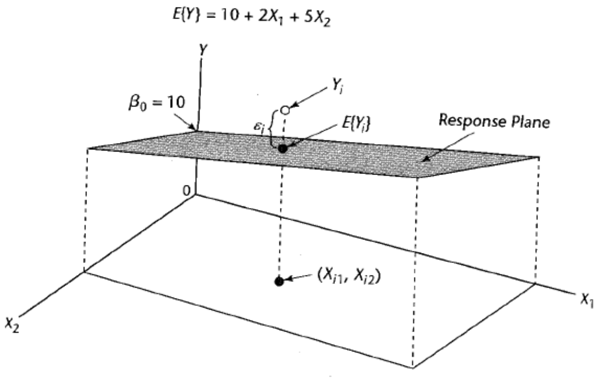
$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}. \tag{2}$$

First-Order Models

$X_1, \cdots, X_{p-1}$ represent $p - 1$ **distinct** predictor variables.

- Response function defines a **hyperplane** in $\mathbb{R}^p$.
- β_k indicates the change in mean response $E(Y)$ with a unit increase in the predictor X_k, when all other predictors are held constant. This change is the same irrespective of the levels at which other predictors are held.
- **The effects of the predictor variables are additive (without interactions).**

Figure : Response plane for a first-order model with two predictors.



From Applied Linear Statistical Models by Kutner, Nachtsheim, Neter and Li

Models with Interactions

Sometimes the effect of one predictor depends on the value(s) of the other predictor(s), i.e., the effects are **non-additive or interacting**.

- For example: How education level affects income may depend on gender.
- These models include the cross product terms.
- Example. Non-additive model with two predictors:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i, \quad i = 1, \cdots, n.$$

  - This model is in the form of the general linear model with $p - 1 = 3$ by defining $X_{i3} := X_{i1} X_{i2}$.
  - The mean response $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$ is linear in the parameters $\beta_0, \beta_1, \beta_2$, but is not linear in the original predictors X_1, X_2.

Example

Brand-liking (Y)	Moisture (X1)	Sweetness (X2)
64.0	4.0	2.0
73.0	4.0	4.0
61.0	4.0	2.0
76.0	4.0	4.0
...	...	...

Design matrix of a first-order model:

$$\mathbf{X} = \begin{bmatrix} 1 & 4.0 & 2.0 \\ 1 & 4.0 & 4.0 \\ 1 & 4.0 & 2.0 \\ 1 & 4.0 & 4.0 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

Design matrix of a non-additive model:

$$\mathbf{X} = \begin{bmatrix} 1 & 4.0 & 2.0 & 8.0 \\ 1 & 4.0 & 4.0 & 16.0 \\ 1 & 4.0 & 2.0 & 8.0 \\ 1 & 4.0 & 4.0 & 16.0 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Polynomial Regression Models

These models contain squared and/or higher-order terms of the predictor variable(s), making the response function curvilinear.

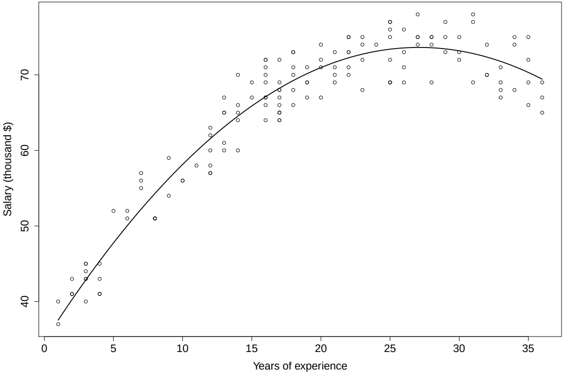
- 2nd-order polynomial regression model with one predictor:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i, \quad i = 1, \cdots, n.$$

  - By defining, $X_{i1} := X_i, X_{i2} := X_i^2$, this model is in the form of the general linear model with $p - 1 = 2$.

Example

Figure : Scatter plot of salary against years of experience



The regression relation appears to be quadratic.

Case	Salary	Experience
1	71	26
2	69	19
3	73	22
4	69	17
5	65	13
6	75	25
...	...	...

Design matrix of a 2nd-order polynomial regression model:

$$\mathbf{X} = \begin{bmatrix} 1 & 26 & 26^2 \\ model(1)1 & 19 & 19^2 \\ 1 & 22 & 22^2 \\ 1 & 17 & 17^2 \\ 1 & 13 & 13^2 \\ 1 & 25 & 25^2 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

Models with Transformed Variables

These models often have complex curvilinear response functions/surfaces.

- Example. Model with logarithm-transformed response variable:

$$\log Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \quad i = 1, \cdots n.$$

  - This model is in the form of the general linear model by defining $\tilde{Y}_i := \log Y_i$.

Key defining features of the general linear regression model: The response function is linear in the regression coefficients: $\beta_0, \beta_1, \cdots, \beta_{p-1}$. However, the response function does not need to be linear in the original predictors, i.e., the response surface could be nonlinear.

- In contrasts, **nonlinear regression models** are nonlinear in the parameters. For example:

$$Y_i = \beta_0 \exp(\beta_1 X_i) + \epsilon_i, \quad i = 1, \cdots n.$$

- The above model can not be expressed in the form of general linear regression model by taking transformations and/or introducing new X variables.

General Linear Regression Model in Matrix Form

Model equations:

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}},$$

where the design matrix $\mathbf{X}$ and the coefficients vector $\boldsymbol{\beta}$:

$$\underset{n\times p}{\mathbf{X}} := \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & X_{i1} & X_{i2} & \cdots & X_{i,p-1} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}, \quad \underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}.$$

Each row of $\mathbf{X}$ corresponds to a case and each column of X corresponds to the n observations of an X variable.

Model assumptions:

$$\mathbf{E}\{\boldsymbol{\epsilon}\} = \mathbf{0}_n, \quad \boldsymbol{\sigma}^{\mathbf{2}}\{\boldsymbol{\epsilon}\} = \sigma^2\mathbf{I}_n.$$

- The response vector has:

$$\mathbf{E}\{\mathbf{Y}\} = \mathbf{X}\boldsymbol{\beta}, \quad \boldsymbol{\sigma}^{\mathbf{2}}\{\mathbf{Y}\} = \sigma^2\mathbf{I}_n.$$

- Under the Normal error model, **Y** is a vector of independent normal random variables.

Least Squares Estimators

- Least squares criterion:

$$\begin{aligned} Q(\mathbf{b}) &= \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{i1} - \cdots - b_{p-1} X_{i,p-1})^2 \\ &= (\mathbf{Y} - \mathbf{X}b)' (\mathbf{Y} - \mathbf{X}b), \quad \underset{p\times 1}{\mathbf{b}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}. \end{aligned}$$

- Differentiate $Q(\cdot)$ and set the gradient to zero $\Longrightarrow$ normal equation:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}.$$

LS estimators are solutions of the normal equation:

$$\underset{p\times 1}{\hat{\boldsymbol{\beta}}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix} = \underset{p\times p}{(\mathbf{X}'\mathbf{X})^{-1}} \underset{p\times n}{\mathbf{X}'} \underset{n\times 1}{\mathbf{Y}}. \tag{3}$$

- $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$:

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

- Variance-covariance matrix of $\hat{\boldsymbol{\beta}}$:

$$\boldsymbol{\sigma}^{\mathbf{2}}\{\boldsymbol{\beta}\} = \sigma^2 \underset{p\times p}{(\mathbf{X}'\mathbf{X})^{-1}}.$$

Notes: hereafter, assume $\mathbf{X}'\mathbf{X}$ *is of full rank p (therefore, we must have* $p \leq n$*).*

Fitted Values and Residuals

$$\underset{n\times 1}{\widehat{\mathbf{Y}}} := \begin{bmatrix} \widehat{Y}_1 \\ \widehat{Y}_2 \\ \vdots \\ \widehat{Y}_n \end{bmatrix} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}, \quad \underset{n\times 1}{\mathbf{e}} := \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}.$$

- Both are linear transformations of the observations vector $\mathbf{Y}$.
- Under the Normal error model, both are normally distributed.
- Expectations and variance-covariance matrices of the fitted values vector and residuals vector:

$$\mathbf{E}\{\widehat{\mathbf{Y}}\} = \mathbf{X}\beta = \mathbf{E}\{\mathbf{Y}\}, \quad \boldsymbol{\sigma}^2\{\widehat{\mathbf{Y}}\} = \sigma^2\mathbf{H}.$$

$$\mathbf{E}\{\mathbf{e}\} = \mathbf{E}\{\mathbf{Y}\} - \mathbf{E}\{\widehat{\mathbf{Y}}\} = \mathbf{0}_n, \quad \boldsymbol{\sigma}^2\{\mathbf{e}\} = \sigma^2(\mathbf{I}_n - \mathbf{H}).$$

Hat Matrix

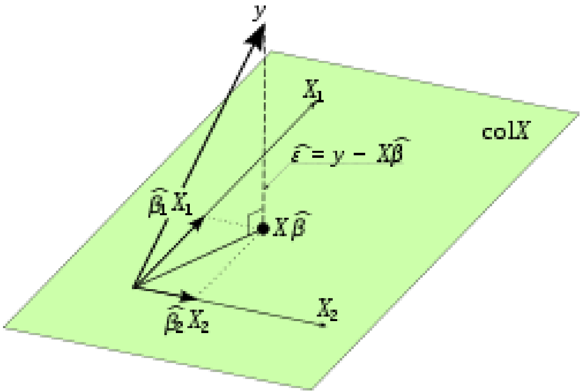
$$\underset{n\times n}{\mathbf{H}} := \underset{n\times p}{\mathbf{X}} \underset{p\times p}{(\mathbf{X}'\mathbf{X})^{-1}} \underset{p\times n}{\mathbf{X}'}.$$

- $\mathbf{H}$ and $\mathbf{I}_n - \mathbf{H}$ are projection matrices: symmetric and idempotent.
- $rank(\mathbf{H}) = p, \quad rank(\mathbf{I}_n - \mathbf{H}) = n - p.$
- $\mathbf{H}$ is the projection matrix to the column space $\langle X \rangle$ of the design matrix $\mathbf{X}$.
  - Fitted value vector $\widehat{\mathbf{Y}} = \mathbf{HY}$ is the projection of the response vector $\mathbf{Y}$ to $\langle X \rangle$.
  - Residual vector $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$ is orthogonal to $\langle X \rangle$.

What are the covariances between $\mathbf{e}$ and $\hat{\mathbf{Y}}$, $\mathbf{e}$ and $\overline{Y}$? What's the implication under the Normal error model?

Geometric Interpretation of Linear Regression

Figure : Orthogonal projection of response vector **Y** onto the linear subspace of $\mathbb{R}^n$ generated by the columns of the design matrix **X**.



Multiple Regression: Example

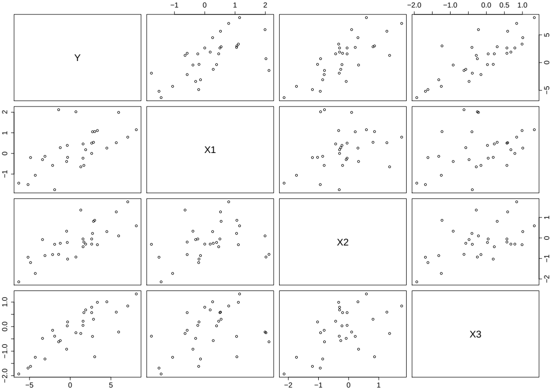
$n = 30$ cases, response variable Y, three predictor variables X_1, X_2, X_3.

case	Y	X1	X2	X3
1	3.01	1.06	0.86	-1.23
2	-3.40	-0.30	-0.08	-0.48
3	2.74	1.05	0.22	-0.40
...	...	...	...	...
30	-1.42	2.12	-0.8	-0.62

First examine each variable marginally: variable type, summary statistics, histogram, boxplot, pie chart, missing values?, outliers?, etc. Then explore their relationships through pairwise scatter plots.

Example: Scatter Plot Matrix

Figure : Pairwise scatter plots between response and predictors and among predictors



All variables appear to be positively correlated. No obvious nonlinearity.

Example: Model 1

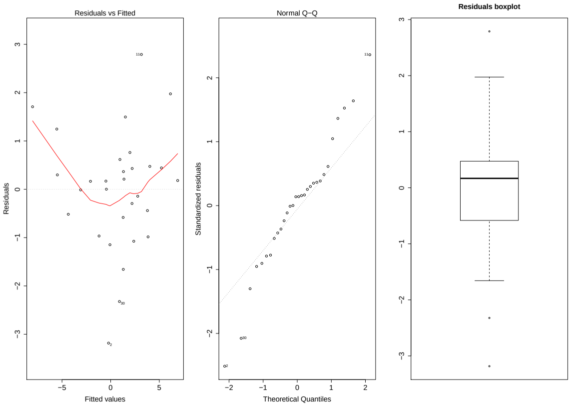
First-order model (only additive effects, a.k.a. *main effects*):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 30.$$

R summary output:

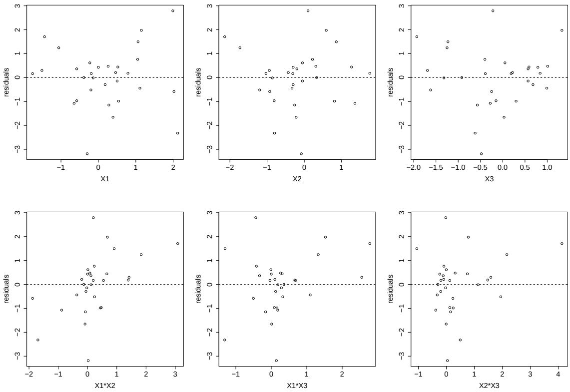
```
Call:
lm(formula = Y ~ X1 + X2 + X3, data = data)

Residuals:
Min      1Q  Median      3Q     Max
-3.1834 -0.5663  0.1673  0.4658  2.7901

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2010     0.2541   4.727 6.91e-05 ***
X1            1.1107     0.2672   4.156 0.000311 ***
X2            1.7978     0.3287   5.469 9.78e-06 ***
X3            1.9596     0.3362   5.829 3.83e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.299 on 26 degrees of freedom
Multiple R-squared: 0.8883,     Adjusted R-squared: 0.8754
F-statistic: 68.93 on 3 and 26 DF,  p-value: 1.667e-12
```

Figure : Model 1: Residual Plots



Residuals vs. fitted values plot shows nonlinearity. Residuals Q-Q plot shows heavy-tail. Residuals boxplot shows that most of residuals are in between 3, –3.

Figure : Model 1: Residuals vs. interaction term $X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3$ Plots



Residuals vs. the interaction term X_1X_2 shows a clear linear pattern. This term should be included in the model.

Example: Model 2

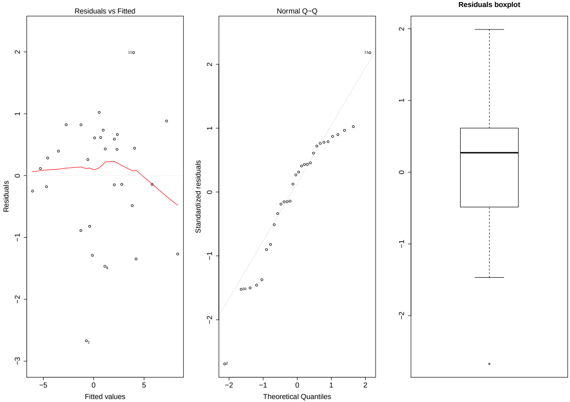
Nonadditive model with interaction between X_1 and X_2:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \epsilon_i, \quad i = 1, \cdots, 30.$$

$(p = 5)$

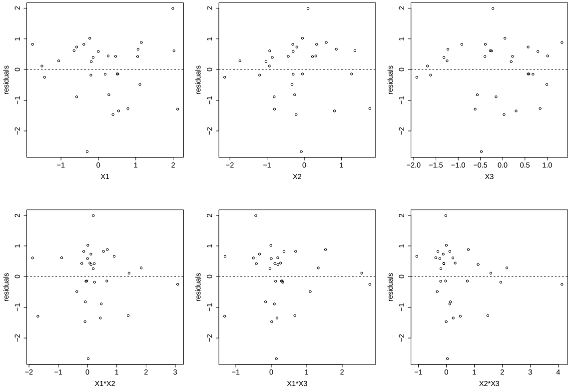
```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X1:X2, data = data)

Residuals:
Min      1Q  Median      3Q     Max
-2.6715 -0.4267  0.2715  0.6138  1.9901

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8832     0.2153   4.103  0.00038 ***
X1            1.5946     0.2421   6.587 6.69e-07 ***
X2            1.7091     0.2605   6.560 7.16e-07 ***
X3            2.1266     0.2687   7.916 2.85e-08 ***
X1:X2         1.0076     0.2467   4.084  0.00040 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.026 on 25 degrees of freedom
Multiple R-squared: 0.933,     Adjusted R-squared: 0.9223
F-statistic: 87.04 on 4 and 25 DF,  p-value: 2.681e-14
```

Figure : Model 2: Residual Plots



Residuals vs. fitted values plot shows no obvious nonlinearity.
Residuals Q-Q plot shows no severe deviation from Normality.
Residuals boxplot shows that most of residuals are in between 2, −2.

Figure : Model 2: Residuals vs. Each of $X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3$ Plots



None of these plots shows an obvious pattern. Model 2 seems to be adequate.

Example: Model 3

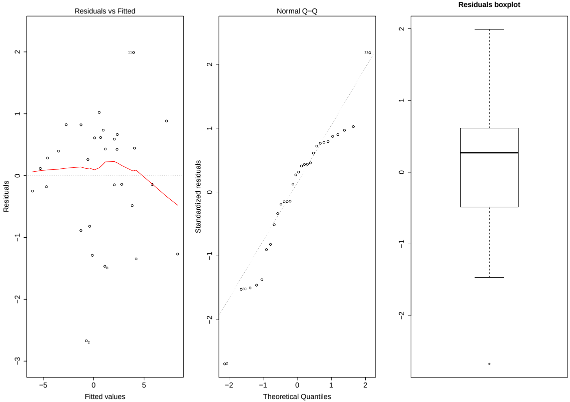
Nonadditive model with all three two-way interaction terms:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \beta_6 X_{i2} X_{i3} + \epsilon_i, \quad i = 1, \cdots, 30.$$

($p = 7$)

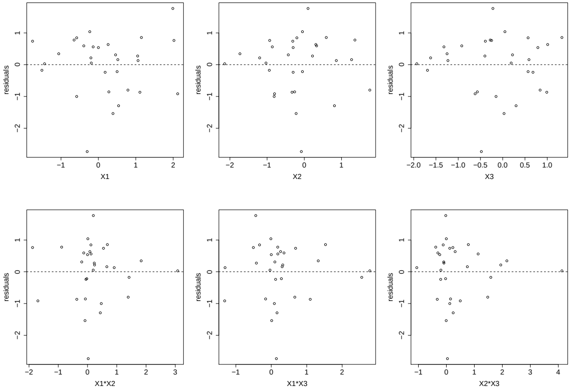
```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X1:X2 + X1:X3 + X2:X3, data = data)

Residuals:
Min      1Q  Median      3Q     Max
-2.7354 -0.6588  0.1868  0.6246  1.7705

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8927     0.2278    3.920 0.000687 ***
X1            1.7179     0.2819    6.095 3.24e-06 ***
X2            1.5828     0.2925    5.411 1.69e-05 ***
X3            1.9982     0.3041    6.571 1.05e-06 ***
X1:X2         1.1925     0.3368    3.541 0.001744 **
X1:X3         0.2227     0.4009    0.555 0.583989
X2:X3        -0.4403     0.3675   -1.198 0.243074
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.038 on 23 degrees of freedom
Multiple R-squared: 0.937,      Adjusted R-squared: 0.9205
F-statistic: 56.99 on 6 and 23 DF,  p-value: 1.172e-12
```

Figure : Model 3: Residual Plots



Residuals vs. fitted values plot shows no obvious nonlinearity.
Residuals Q-Q plot shows no severe deviation from Normality.
Residuals boxplot shows that most of residuals are in between 2, −2.

Figure : Model 3: Residuals vs. Each of $X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3$ Plots



None of these plots shows an obvious pattern. Model 3 seems to be adequate, but there is no obvious improvement over Model 2.

Analysis of Variance

$$\mathbf{SSTO} = \mathbf{SSE} + \mathbf{SSR}, \quad \mathbf{d.f.}(\mathbf{SSTO}) = \mathbf{d.f.}(\mathbf{SSE}) + \mathbf{d.f.}(\mathbf{SSR}).$$

- **Total sum of squares**:

$$SSTO = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \mathbf{Y}'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y}, \ \ d.f.(SSTO) = rank(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n) = n - 1.$$

- **Error sum of squares**:

$$SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y}, \ \ d.f.(SSE) = rank(\mathbf{I}_n - \mathbf{H}) = n - p.$$

- **Regression sum of squares**:

$$SSR = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = \mathbf{Y}'(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)\mathbf{Y}, \ \ d.f.(SSR) = rank(\mathbf{H} - \frac{1}{n}\mathbf{J}_n) = p - 1.$$

Sampling distributions of sums of squares (SS) under the Normal error model:

- *SSE* and *SSR* are independent.
  Notes: use the facts that $\mathbf{e}$ *are independent with* $\hat{\mathbf{Y}}$ *and* $\overline{Y}$*.*
  Why?
- $SSE \sim \sigma^2\chi^2_{(n-p)}$. *What is* $E(SSE)$*?*
- If $\beta_1 = \cdots = \beta_{p-1} = 0$, then $SSR \sim \sigma^2\chi^2_{(p-1)}$. *What is* $E(SSR)$ *in such a case? And what would be the sampling distribution of SSTO?*

Mean squares (MS): **MS** $=$ **SS**/**d.f.**(**SS**).

- MSE (mean squared error):

$$MSE = \frac{SSE}{n-p}, \quad E(MSE) = \sigma^2.$$

MSE is an unbiased estimator of the error variance σ^2.

- MSR:

$$MSR = \frac{SSR}{p-1}.$$

$$E(MSR) = \begin{cases} \sigma^2 & \text{if} \quad \beta_1 = \cdots = \beta_{p-1} = 0 \\ > \sigma^2 & \text{if} \quad \text{otherwise} \end{cases}$$

- $MSTO = \frac{SSTO}{n-1}$.

For n cases, up to how many X variables can be included in the model?

F Test of Regression Relation

Under the Normal error model

- Test **whether there is a regression relation between the response variable Y and the set of X variables**:

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0 \quad vs.$$

$$H_a : \text{not all } \beta_k \text{s equal zero.}$$

- F ratio and its null distribution:

$$F^* = \frac{MSR}{MSE}, \quad F^* \sim_{H_0} F_{p-1,n-p},$$

  where $F_{p-1,n-p}$ denotes the F distribution with $(p-1, n-p)$ degrees of freedom.

- Decision rule at level α: reject H_0 if $F^* > F(1-\alpha; p-1, n-p)$.

ANOVA Table

Source of Variation	SS	d.f.	MS	F^*
Regression	$SSR = \mathbf{Y}'(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)\mathbf{Y}$	$p-1$	$MSR = \frac{SSR}{p-1}$	$F^* = \frac{MSR}{MSE}$
Error	$SSE = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y}$	$n-p$	$MSE = \frac{SSE}{n-p}$	
Total	$SSTO = \mathbf{Y}'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y}$	$n-1$		

Example Model 2: $n = 30, p = 5$.

Source of Variation	SS	d.f.	MS	F^*
Regression	$SSR = 366.4846$	4	$MSR = 91.62116$	$F^* = 87.03703$
Error	$SSE = 26.31672$	25	$MSE = 1.052669$	
Total	$SSTO = 392.8013$	29		

Pvalue $= P(F_{4,25} > 87.037) \approx 0$, so there is a significant regression relation between Y and X_1, X_2, X_3, X_1X_2.

Coefficient of Multiple Determination

$$R^2 := \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}.$$

- R^2 is the proportional reduction of the total variation in Y by using the X variables to explain Y.
- $0 \leq R^2 \leq 1$.
  When $R^2 = 0$? When $R^2 = 1$?
- **Adding more X variables to the model will always increase** R^2 because:
  (i) $SSTO$ remains the same.
  (ii) SSE becomes smaller.

Since adding more X variables can only increase R^2, does this mean we should use as many X variables as possible?

- With more X variables, the model fits the observed data better due to smaller SSE.
- However, a model with many X variables that are unrelated to the response variable and/or are highly correlated with each other tends to
    - **overfit** the observed data and often do a poor job for prediction due to increased sampling variability.
    - make interpretation difficult.
    - make prediction more costly.
- We will discuss this in more details later.

Adjusted Coefficient of Multiple Determination

Adjust for the number of X variables in the model:

$$R_a^2 = 1 - \frac{MSE}{MSTO} = 1 - \frac{n-1}{n-p}\frac{SSE}{SSTO}.$$

- $R_a^2 \leq R^2$.
- R_a^2 **may become smaller when adding more** X **variables into the model** because:
  - the decrease in SSE may be more than offset by the loss of degrees of freedom in SSE.

Example

- Model 1: $Y \sim X_1, X_2, X_3$

$$R^2 = 0.8883, \quad R_a^2 = 0.8754$$

- Model 2 : $Y \sim X_1, X_2, X_3, X_1X_2$

$$R^2 = 0.933, \quad R_a^2 = 0.9223.$$

- Model 3: $Y \sim X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3$.

$$R^2 = 0.937, \quad R_a^2 = 0.9205.$$

(i) For each model, $R^2 > R_a^2$; (ii) Adding more X variable(s) increases R^2. The increase of R^2 is much more from Model 1 to Model 2 than from Model 2 to Model 3; (iii) Model 3 has a smaller R_a^2 than Model 2.

Inferences about Regression Coefficients

LS estimators:

$$\underset{p\times 1}{\hat{\boldsymbol{\beta}}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix} = \underset{p\times p}{(\mathbf{X}'\mathbf{X})^{-1}} \underset{p\times n}{\mathbf{X}'} \underset{n\times 1}{\mathbf{Y}}.$$

$$\underset{p\times 1}{\mathbf{E}\{\hat{\boldsymbol{\beta}}\}} = \boldsymbol{\beta}, \quad \underset{p\times p}{\boldsymbol{\sigma}^{\mathbf{2}}\{\hat{\boldsymbol{\beta}}\}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

The standard error of $\hat{\beta}_k$, $s(\hat{\beta}_k)$, is the positive square-root of the $(k+1)th$ diagonal element of $MSE(\mathbf{X}'\mathbf{X})^{-1}$.

- Studentized pivotal quantity:

$$\frac{\hat{\beta}_k - \beta_k}{s\{\hat{\beta}_k\}} \sim t_{(n-p)}.$$

- $(1-\alpha)$-Confidence interval for β_k:

$$\hat{\beta}_k \pm t(1-\alpha/2; (n-p))s\{\hat{\beta}_k\}.$$

- Two-sided T-Test: $H_0 : \beta_k = \beta_k^0 \;\; vs. \;\; H_a : \beta_k \neq \beta_k^0$.
- T statistic:

$$T^* = \frac{\hat{\beta}_k - \beta_k^0}{s\{\hat{\beta}_k\}} \underset{H_0}{\sim} t_{(n-p)}.$$

  At level α, the decision rule is to reject H_0 if and only if $|T^*| > t(1-\alpha/2; (n-p))$.

Example: Model 2

Nonadditive model with interaction between X_1 and X_2:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \epsilon_i, \quad i = 1, \cdots, 30.$$

$(p = 5)$

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X1:X2, data = data)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8832     0.2153    4.103  0.00038 ***
X1            1.5946     0.2421    6.587 6.69e-07 ***
X2            1.7091     0.2605    6.560 7.16e-07 ***
X3            2.1266     0.2687    7.916 2.85e-08 ***
X1:X2         1.0076     0.2467    4.084  0.00040 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.026 on 25 degrees of freedom
Multiple R-squared: 0.933,      Adjusted R-squared: 0.9223
F-statistic: 87.04 on 4 and 25 DF,  p-value: 2.681e-14
```

Model 3

Test whether there is an interaction between X_1 and X_2. Use $\alpha = 0.01$.

- $H_0 : \beta_4 = 0, \quad vs., \quad H_a : \beta_4 \neq 0.$
- $T^* = \frac{1.0076-0}{0.2467} = 4.084.$
- $n = 30, p = 5, t(0.995; 25) = 2.787.$
- Since $|4.084| > 2.787$, reject the null hypothesis and conclude that there is a significant interaction effect between X_1 and X_2.
- Alternatively, pvalue=$P(|t_{(25)}| > |4.084|) = 0.00040 < 0.01$, so reject H_0.

Notes: pvalue for the two-sided alternative is in the R output.

Estimation of the Mean Response

- For a given set of values of the X variables:

$$\underset{p\times 1}{\mathbf{X}_h} = \begin{bmatrix} 1 \\ X_{h1} \\ \vdots \\ X_{h,p-1} \end{bmatrix}$$

- Corresponding mean response:

$$E(Y_h) = \mathbf{X}_h'\boldsymbol{\beta} = \beta_0 + \beta_1 X_{h1} + \cdots + \beta_{p-1} X_{h,p-1}.$$

- $\widehat{Y}_h := \mathbf{X}'_h\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $E(Y_h)$:

$$E(\widehat{Y}_h) = E(\mathbf{X}'_h\hat{\boldsymbol{\beta}}) = \mathbf{X}'_h\mathbf{E}\{\hat{\boldsymbol{\beta}}\} = \mathbf{X}'_h\boldsymbol{\beta} = E(Y_h).$$

$$\sigma^2(\widehat{Y}_h) = \mathbf{X}'_h\boldsymbol{\sigma}^{\mathbf{2}}\{\hat{\boldsymbol{\beta}}\}\mathbf{X}_h = \sigma^2\left(\mathbf{X}'_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h\right).$$

- Standard error of $\widehat{Y}_h$:

$$s(\widehat{Y}_h) = \sqrt{MSE\left(\mathbf{X}'_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h\right).}$$

- $(1-\alpha)$-confidence interval for $E(Y_h)$:

$$\widehat{Y}_h \pm t(1-\alpha/2; n-p)s(\widehat{Y}_h).$$

Prediction of a New Observation

- $Y_{h(new)} = \mathbf{X}_h' \boldsymbol{\beta} + \epsilon_h$: independent with the observations Y_is.
- Predicted value: $\widehat{Y}_h := \mathbf{X}_h' \hat{\boldsymbol{\beta}}$

$$\sigma^2(pred_h) := Var(\widehat{Y}_h - Y_{h(new)}) = \sigma^2(\widehat{Y}_h) + \sigma^2(Y_{h(new)}) = \sigma^2 \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h + \sigma^2.$$

- Standard error for prediction:

$$s(pred_h) = \sqrt{MSE\left[1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h\right]}.$$

- $(1-\alpha)$-prediction interval for $Y_{h(new)}$:

$$\widehat{Y}_h \pm t(1-\alpha/2; n-p)s(pred_h).$$

Example

Estimate the mean response when $X_1 = 0.8, X_2 = 0.5, X_3 = -1$ under Model 2.

- $\mathbf{X}_h' = \begin{bmatrix} 1 & 0.8 & 0.5 & -1 & 0.8 \times 0.5 \end{bmatrix}$
- $n = 30, p = 5$:
$$\widehat{Y}_h := \mathbf{X}_h'\hat{\beta} = 1.290,$$
$$\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h = 0.170, \quad MSE = 1.053$$
$$s(\widehat{Y}_h) = \sqrt{1.053 \times 0.170} = 0.423.$$
- A 99%-confidence interval for $E(Y_h)$: $t(0.995; 25) = 2.787$
$$1.290 \pm 2.787 \times 0.423 = [0.111, 2.469].$$

Predict a new observation when $X_1 = 0.8, X_2 = 0.5, X_3 = -1$ under Model 2.

- Standard error for prediction:

$$s(pred) = \sqrt{1.053 \times (1 + 0.170)} = 1.1098.$$

- A 99%-prediction interval for Y_{hnew}:

$$1.290 \pm 2.787 \times 1.1098 = [-1.803, 4.383].$$

- R codes.

```
> newX=data.frame(X1=0.8, X2=0.5, X3=-1)
> predict.lm(fit2, newX, interval="confidence",
+ level=0.99, se.fit=TRUE)

> predict.lm(fit2, newX, interval="prediction",
+ level=0.99, se.fit=TRUE)
```

Extra Sum of Squares

$\mathcal{I}$ and $\mathcal{J}$ are two **non-overlapping** index sets.

- **Extra sum of squares (ESS)**:

$$SSR(X_{\mathcal{J}}|X_{\mathcal{I}}) := SSE(X_{\mathcal{I}}) - SSE(X_{\mathcal{I}}, X_{\mathcal{J}}).$$

- It indicates the **reduction in error sum of squares by adding $X_{\mathcal{J}}$ to the model where $X_{\mathcal{I}}$ is the set of X variables.**
- Degrees of freedom: $d.f.(SSR(X_{\mathcal{J}}|X_{\mathcal{I}})) = |\mathcal{J}|$.
- Mean squares: $MSR(X_{\mathcal{J}}|X_{\mathcal{I}}) := \frac{SSR(X_{\mathcal{J}}|X_{\mathcal{I}})}{d.f.(SSR(X_{\mathcal{J}}|X_{\mathcal{I}}))}$.

Notations.

- $\mathcal{I}$: an index set; $X_{\mathcal{I}} := \{X_i : i \in \mathcal{I}\}$.
    - E.g. $\mathcal{I} = \{2, 3\}$, $X_{\mathcal{I}} = \{X_2, X_3\}$.
- $SSE(X_{\mathcal{I}})$ and $SSR(X_{\mathcal{I}})$ denote the error sum of squares and regression sum of squares, respectively, under the regression model with $X_{\mathcal{I}} := \{X_i : i \in \mathcal{I}\}$ being the X variables.
    - E.g., $SSE(X_2, X_3)$ is the error sum of squares of the model with X_2 and X_3.

Body Fat

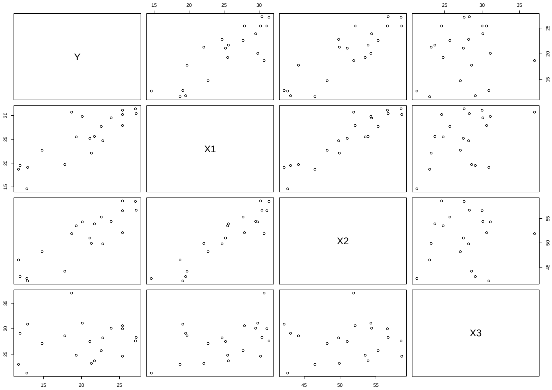
A researcher measured the amount of body fat (Y) of 20 healthy females 25 to 34 years old, together with three (potential) predictor variables, triceps skinfolds thickness (X_1), thigh circumference (X_2), and midarm circumference (X_3). The amount of body fat was obtained by a cumbersome and expensive procedure requiring immersion of the person in water. Thus it would be helpful if a regression model with some or all of these predictors could provide reliable estimates of body fat as these predictors are easy to measure.

A snapshot of the data.

case	X1 Triceps	X2 Thigh	X3 MidArm	Y BodyFat
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
3	30.7	51.9	37.0	18.7
4	29.8	54.3	31.1	20.1
5	19.1	42.2	30.9	12.9
6	25.6	53.9	23.7	21.7
...	...	...	...	...

First check the variable type, distribution, etc.,of each variable.

Scatter plot matrix.



No obvious nonlinearity.

Correlation matrix.

	X1	X2	X3	Y
X1	1.0000000	0.9238425	0.4577772	0.8432654
X2	0.9238425	1.0000000	0.0846675	0.8780896
X3	0.4577772	0.0846675	1.0000000	0.1424440
Y	0.8432654	0.8780896	0.1424440	1.0000000

X_1 and X_2 are strongly correlated, X_1 and X_3 are moderately correlated, X_2 and X_3 are weakly correlated. Moreover, X_1, X_2 are strongly correlated with Y and X_3 is weakly correlated with Y.

Consider the following 4 models.

- Model 1: regression of Y on X_1

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- Model 2: regression of Y on X_2

$$Y_i = \beta_0 + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- Model 3: regression of Y on X_1 and X_2

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- Model 4: regression of Y on X_1, X_2 and X_3.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 20.$$

Boy Fat: Model 1

```
> summary(fit1)

Call:
lm(formula = Y ~ X1, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4961     3.3192  -0.451    0.658
X1            0.8572     0.1288   6.656 3.02e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.82 on 18 degrees of freedom
Multiple R-squared: 0.7111,     Adjusted R-squared: 0.695
F-statistic:  44.3 on 1 and 18 DF,  p-value: 3.024e-06

> anova(fit1)
Analysis of Variance Table

Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1          1 352.27  352.27  44.305 3.024e-06 ***
Residuals 18 143.12    7.95
```

GLT

Boy Fat: Model 2

```
> summary(fit2)

Call:
lm(formula = Y ~ X2, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.6345     5.6574  -4.178 0.000566 ***
X2            0.8565     0.1100   7.786  3.6e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.51 on 18 degrees of freedom
Multiple R-squared: 0.771,      Adjusted R-squared: 0.7583
F-statistic: 60.62 on 1 and 18 DF,  p-value: 3.6e-07

> anova(fit2)
Analysis of Variance Table

Response: Y
Df Sum Sq Mean Sq F value  Pr(>F)
X2         1 381.97  381.97  60.617 3.6e-07 ***
Residuals 18 113.42    6.30
```

Boy Fat: Model 3

```
> summary(fit3)

Call:
lm(formula = Y ~ X1 + X2, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
X1            0.2224     0.3034   0.733   0.4737
X2            0.6594     0.2912   2.265   0.0369 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.543 on 17 degrees of freedom
Multiple R-squared: 0.7781,     Adjusted R-squared: 0.7519
F-statistic:  29.8 on 2 and 17 DF,  p-value: 2.774e-06

> anova(fit3)
Analysis of Variance Table

Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27 54.4661 1.075e-06 ***
X2         1  33.17   33.17  5.1284    0.0369 *
Residuals 17 109.95    6.47
```

GLT

Boy Fat: Model 4

```
> summary(fit4)
Call:
lm(formula = Y ~ X1 + X2 + X3, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
X1             4.334      3.016   1.437    0.170
X2            -2.857      2.582  -1.106    0.285
X3            -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared: 0.8014,     Adjusted R-squared: 0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06

> anova(fit4)
Analysis of Variance Table

Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27 57.2768 1.131e-06 ***
X2         1  33.17   33.17  5.3931   0.03373 *
X3         1  11.55   11.55  1.8773   0.18956
Residuals 16  98.40    6.15
```

GLT

Body Fat: ESS

From the R outputs, we can derive a number of extra sums of squares. For example:

- From Model 1, $SSE(X_1) = 143.12$ and from Model 3, $SSE(X_1, X_2) = 109.95$. So

  $$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2) = 143.12 - 109.95 = 33.17.$$

- From Model 2, $SSE(X_2) = 113.42$, so

  $$SSR(X_1|X_2) = SSE(X_2) - SSE(X_1, X_2) = 113.42 - 109.95 = 3.47.$$

- Both extra sums of squares have degrees of freedom 1, so $MSR(X_2|X_1) = 33.17$ and $MSR(X_1|X_2) = 3.47$.
- The reduction of SSE by adding X_2 to a model with X_1 is much more than the reduction of SSE by adding X_1 to a model with X_2.

- From Model 4, $SSE(X_1, X_2, X_3) = 98.40$, so

$$\begin{aligned} SSR(X_3|X_1, X_2) &= SSE(X_1, X_2) - SSE(X_1, X_2, X_3) \\ &= 109.95 - 98.40 = 11.55. \end{aligned}$$

  This extra sum of squares has degrees of freedom 1, so $MSR(X_3|X_1, X_2) = 11.55$.
- Moreover,

  $$SSR(X_2, X_3|X_1) = SSE(X_1) - SSE(X_1, X_2, X_3) = 143.12 - 98.40 = 44.72,$$

  $$SSR(X_1, X_3|X_2) = SSE(X_2) - SSE(X_1, X_2, X_3) = 113.42 - 98.40 = 15.02.$$

  These two extra sums of squares have degrees of freedom 2, so $MSR(X_2, X_3|X_1) = 44.72/2 = 22.36$, $MSR(X_1, X_3|X_2) = 15.02/2 = 7.51$.

Are there other ESS that can be derived from the R outputs?

Decomposition of SSR into ESS

For a model with multiple X variables, the regression sum of squares (SSR) can be expressed as the sum of several extra sums of squares.

- For example:

$$SSR(X_1, X_2) = SSR(X_1) + SSR(X_2|X_1).$$

  $SSR(X_1)$ measures the contribution by having X_1 alone in the model, whereas $SSR(X_2|X_1)$ measures the additional contribution when X_2 is added, given that X_1 is already in the model.

- However, such decomposition is usually not unique. For example,

$$SSR(X_1, X_2) = SSR(X_2) + SSR(X_1|X_2).$$

Read *anova()* output

It provides decomposition of *SSR* into single d.f. ESS, **in the order of the *X* variables entering the model.**

```
Call:
lm(formula = Y ~ X1 + X2 + X3, data = fat)
> anova(fit4)
Analysis of Variance Table
Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1          1 352.27  352.27 57.2768 1.131e-06 ***
X2          1  33.17   33.17  5.3931   0.03373 *
X3          1  11.55   11.55  1.8773   0.18956
Residuals 16  98.40    6.15
```

Source of Variation	SS	d.f.	MS
Regression	396.99	3	132.33
X_1	352.27	1	352.27
$X_2 \mid X_1$	33.17	1	33.17
$X_3 \mid X_1, X_2$	11.55	1	11.55
Error	98.40	16	6.15
Total	495.39	19	

For example: $SSR(X_2, X_3 \mid X_1) = SSR(X_2 \mid X_1) + SSR(X_3 \mid X_1, X_2) = 33.17 + 11.55 = 44.72$.

How to get $SSR(X_2|X_1, X_3)$ from the R output of Model 4? We need to enter the X variables in the following order: X_1, X_3, X_2.

```
Call:
lm(formula = Y ~ X1 + X3 + X2, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
X1             4.334      3.016   1.437    0.170
X3            -2.186      1.595  -1.370    0.190
X2            -2.857      2.582  -1.106    0.285

> anova(fit4.alt2)
Analysis of Variance Table
Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27 57.2768 1.131e-06 ***
X3         1  37.19   37.19  6.0461   0.02571 *
X2         1   7.53    7.53  1.2242   0.28489
Residuals 16  98.40    6.15
```

Then we can get $SSR(X_2|X_1, X_3) = 7.53$.

General Linear Tests

$\mathcal{I}$ and $\mathcal{J}$ are two non-overlapping index sets.

- **Full model**: Contain both $X_{\mathcal{I}}$ and $X_{\mathcal{J}}$.
- Test whether $X_{\mathcal{J}}$ may be dropped out of the full model:

$$H_0 : \beta_j = 0, \ \text{ for } \textbf{all } j \in \mathcal{J}$$

vs.

$$H_a : \text{some } \beta_j : j \in \mathcal{J} \text{ are nonzero.}$$

- H_0 corresponds to a **reduced model** with only $X_{\mathcal{I}}$.

Basic idea: Compare *SSE* under the full model with *SSE* under the reduced model by an F ratio:

$$F^* = \frac{\frac{SSE(R)-SSE(F)}{df_R-df_F}}{\frac{SSE(F)}{df_F}} = \frac{MSR(X_{\mathcal{J}}|X_{\mathcal{I}})}{MSE(F)}.$$

- Under H_0 (i.e., the reduced model):

$$F^* \sim_{H_0} F_{df_R-df_F,df_F}.$$

- Reject H_0 at level α if the observed $F^* > F(1-\alpha; df_R-df_F, df_F)$.

Rationale behind the general linear tests.

- If $SSE(F)$ is close to $SSE(R)$, then the additional X variables in the full model do not contribute much to explain the variation in the observations.
  Thus a small $SSE(R) - SSE(F)$ is evidence for H_0, i.e., the reduced model.
- On the other hand, a large $SSE(R) - SSE(F)$ means that the additional X variables in the full model substantially reduce the deviation of the observations around the fitted regression surface, and thus serves as evidence for H_a, i.e., the full model.

F-test for Regression Relation

- Full model with $X_1, \cdots, X_{p-1}$:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \quad i = 1, \cdots n.$$

- Reduced model with no X variable:

$$Y_i = \beta_0 + \epsilon_i, \;\; i = 1, \cdots, n.$$

  So $SSE(R) = SSTO$ and $df_R = n - 1$.
- $SSE(R) - SSE(F) = SSTO - SSE(F) = SSR(F)$, and $df_R - df_F = (n-1) - (n-p) = p - 1 = d.f.(SSR(F))$.
- F ratio

$$F^* = \frac{SSR(F)/(p-1)}{SSE(F)/(n-p)} = \frac{MSR(F)}{MSE(F)}.$$

Test whether a Single $\beta_k = 0$

Body fat: Test for the model with all three predictors whether the midarm circumference (X_3) can be dropped.

- Full model: $SSE(F) = 98.40$ with d.f. 16.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- Null and alternative hypotheses:

$$H_0 : \beta_3 = 0 \;\; vs. \quad H_a : \beta_3 \neq 0.$$

- Reduced model: $SSE(R) = 109.95$ with d.f. 17.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- $F^* = \frac{11.55/1}{98.40/16} = 1.88.$
- Pvalue=$P(F_{1,16} > 1.88) = 0.189$. So we can drop X_3 from the full model.

Equivalence between F-test and T-test

- Test whether X_k can be dropped from a regression model with $p-1$ X variables:

$$H_0 : \beta_k = 0 \;\; vs. \;\; H_a : \beta_k \neq 0.$$

- We can use an F-test: $F^* \underset{H_0}{\sim} F_{1,n-p}$.

- Alternatively, we may use a T-test:

$$T^* = \frac{\hat{\beta}_k}{s\{\hat{\beta}_k\}} \underset{H_0}{\sim} t_{(n-p)},$$

  where $\hat{\beta}_k$ is the LS estimator of β_k and $s\{\hat{\beta}_k\}$ is its standard error under the full model.

- It can be show that $F^* = (T^*)^2$ and $F(1-\alpha; 1, n-p) = (t(1-\alpha/2; n-p))^2$. So in this case F-test and T-test are equivalent.

Notes: for one one-sided alternatives, we still need the T-tests.

Test whether Several $\beta_k = 0$

Body fat: Test whether both thigh circumference (X_2) and midarm circumference (X_3) can be dropped from the model with all three predictors.

- Full model: $SSE(F) = 98.40$ with d.f. 16.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- Null and alternative hypotheses:

$$H_0 : \beta_2 = \beta_3 = 0 \;\; vs. \quad H_a : \text{not both } \beta_2 \text{ and } \beta_3 \text{ equal zero.}$$

- Reduced model: $SSE(R) = 143.12$ with d.f. 18.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- $F^* = \frac{44.72/2}{98.40/16} = 3.635.$
- Pvalue$= P(F_{2,16} > 3.635) = 0.0499$. The result is barely significant at $\alpha = 0.05$.

Standardization

Different X variables often have different units which could make their values vastly different.

- Regression coefficients are not in the same scale and thus are hard to interpret.
- Elements of $\mathbf{X}'\mathbf{X}$ differ substantially in order of magnitude, causing numerical instability while solving for its inverse.
- A regression model can be reparametrized into a standardized regression model through centering and rescaling.
- This process is called **standardization**, a.k.a. **correlation transformation**. It also helps with the understanding of regression model.

Correlation Transformation

Define transformed variables:

$$X_{ik}^{*} = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik} - \overline{X}_k}{s_{X_k}}\right), \quad k = 1, \cdots, p-1,$$

where

$$\overline{X}_k = \frac{1}{n}\sum_{i=1}^{n} X_{ik}, \quad s_{X_k} = \sqrt{\frac{\sum_{i=1}^{n}(X_{ik} - \overline{X}_k)^2}{n-1}}, \quad (k = 1, \cdots, p-1).$$

are sample means and sample standard deviations, respectively.

- The sample means of the transformed variables are all zero.
- The sample standard deviations of the transformed variables are all $\frac{1}{\sqrt{n-1}}$.
- So all variables are centered and are on the same scale.
- Correlation transformation does not change the pairwise (sample) correlations among the X variables, nor does it change the (sample) correlations between the X variables and the response variable.

Standardized Regression Model

Rewrite the regression model in terms of standardized variables:

$$Y_i = \beta_0^* + \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \cdots + \beta_{p-1}^* X_{i,p-1}^* + \epsilon_i, \quad i = 1, \cdots n,$$

where

$$\beta_k^* = \sqrt{n-1} s_{X_k} \beta_k \;\; (k = 1, \cdots, p-1), \quad \beta_0^* = \beta_0 + \sum_{k=1}^{p-1} \beta_k \bar{X}_k$$

is a "reparametrization" of the original model.

Design Matrix of Standardized Model

$$\underset{n\times p}{\mathbf{X}^*} = \begin{bmatrix} 1 & X_{11}^* & \cdots & X_{1,p-1}^* \\ 1 & X_{21}^* & \cdots & X_{2,p-1}^* \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1}^* & \cdots & X_{n,p-1}^* \end{bmatrix}.$$

$$\underset{p\times p}{\mathbf{X}^{*\prime}\mathbf{X}^*} = \begin{bmatrix} n & 0 & 0 & \cdots & 0 \\ 0 & 1 & r_{12} & \cdots & r_{1,p-1} \\ 0 & r_{21} & 1 & \cdots & r_{2,p-1} \\ 0 & \vdots & \cdots & \vdots & \\ 0 & r_{p-1,1} & r_{p-1,2} & \cdots & 1 \end{bmatrix} = \begin{bmatrix} n & \mathbf{0}^T \\ \mathbf{0} & \underset{(p-1)\times(p-1)}{\mathbf{r}_{XX}} \end{bmatrix},$$

where $\mathbf{r}_{XX}$ is the sample correlation matrix of the X variables.

Correlation Matrix

- Its (k, l)-element r_{kl} is the sample correlation coefficient between X_k, X_l:

$$r_{kl} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_{ik} - \overline{X}_k)(X_{il} - \overline{X}_l)}{s_{X_k} s_{X_l}}, \quad 1 \leq k, l \leq p - 1.$$

- All its elements are unit-less numbers in between -1 and 1.
- Its diagonal elements are all one, since the correlation of a variable with itself is one, i.e., $r_{kk} \equiv 1$ for $k = 1, \cdots, p - 1$.
- Correlation matrix is a symmetric matrix: $r_{kl} = r_{lk}$.

$\mathbf{X}'\mathbf{Y}$ Matrix of Standardized Model

$$\underset{p\times 1}{\mathbf{X}^{*\prime}\mathbf{Y}} = \begin{bmatrix} n\bar{Y} \\ \sqrt{n-1}s_Y r_{Y1} \\ \sqrt{n-1}s_Y r_{Y2} \\ \vdots \\ \sqrt{n-1}s_Y r_{Y,p-1} \end{bmatrix} = \sqrt{n-1}s_Y \begin{bmatrix} \frac{n}{\sqrt{n-1}s_Y}\bar{Y} \\ \underset{(p-1)\times 1}{\mathbf{r}_{XY}} \end{bmatrix}$$

where r_{Yk} is the sample correlation coefficient between Y and X_k:

$$r_{Yk} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_{ik} - \overline{X}_k)(Y_i - \overline{Y})}{s_{X_k}s_Y}, \quad k = 1, \cdots p-1.$$

LS Fit of Standardized Model

$$\underset{p\times 1}{\hat{\boldsymbol{\beta}}^*} = \begin{bmatrix} \hat{\beta}_0^* \\ \hat{\beta}_1^* \\ \hat{\beta}_2^* \\ \vdots \\ \hat{\beta}_{p-1}^* \end{bmatrix} = \begin{bmatrix} \bar{Y} \\ \sqrt{n-1}\, s_Y \underset{(p-1)\times 1}{\mathbf{r}_{XX}^{-1}\mathbf{r}_{XY}}. \end{bmatrix}$$

- These are called *fitted standardized regression coefficients*.
- Relationships with the LS estimators of the original model:

$$\begin{aligned} \hat{\beta}_k &= \frac{1}{\sqrt{n-1}\, s_{X_k}} \hat{\beta}_k^*, \quad k = 1, \cdots, p-1 \\ \hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \cdots - \hat{\beta}_{p-1} \overline{X}_{p-1}. \end{aligned}$$

Body Fat

Sample means and sample standard deviations ($n = 30$):

$$\overline{Y} = 20.20, \quad \overline{X}_1 = 25.30, \quad \overline{X}_2 = 51.17, \quad \overline{X}_3 = 27.62;$$

$$s_Y = 5.11. \quad s_{X_1} = 5.02, \quad s_{X_2} = 5.23, \quad s_{X_3} = 3.65.$$

Correlation matrices:

$$\mathbf{r}_{XX} = \begin{bmatrix} 1.00 & 0.92 & 0.46 \\ 0.92 & 1.00 & 0.08 \\ 0.46 & 0.08 & 1.00 \end{bmatrix}, \quad \mathbf{r}_{XY} = \begin{bmatrix} 0.84 \\ 0.88 \\ 0.14 \end{bmatrix}.$$

Least-squares estimators of the standardized model:

$$\hat{\beta}_0^* = \overline{Y} = 20.20, \quad \begin{bmatrix} \hat{\beta}_1^* \\ \hat{\beta}_2^* \\ \hat{\beta}_3^* \end{bmatrix} = \sqrt{n-1}\, s_Y \mathbf{r}_{XX}^{-1} \mathbf{r}_{XY} = 27.5 \times \begin{bmatrix} 4.26 \\ -2.93 \\ -1.56 \end{bmatrix}.$$

Least-squares estimators of the original model:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 4.33 \\ -2.86 \\ -2.18 \end{bmatrix} = \begin{bmatrix} \frac{5.11}{5.02} \times 4.26 \\ \frac{5.11}{5.23} \times (-2.93) \\ \frac{5.11}{3.65} \times (-1.56) \end{bmatrix}.$$

Multicollinearity

Multicollinearity refers to the situation when the X variables are **intercorrelated** among themselves.

- This term is often reserved for the situation when the inter-correlation/collinearity among the X variables is **very high**.
- X variables being nearly collinear/highly intercorrelated means that there exist constants $c_0, c_1, \cdots, c_{p-1}$ not all zero such that

$$c_0 + c_1 X_{i1} + \cdots + c_{p-1} X_{i,p-1} \approx 0, \quad i = 1, \cdots, n.$$

i.e., there exists a nonzero vector $\mathbf{c}$ such that $\underset{n\times p}{\mathbf{X}} \underset{p\times 1}{\mathbf{c}} \approx \mathbf{0}_n$.

Interpreting Regression Coefficients

- In presence of multicollinearity, **a regression coefficient should be interpreted as reflecting the marginal/partial effect of the corresponding X variable, given whatever other X variables also in the model.**
- To understand the effects of multicollinearity, we consider two extreme situations: (i) When the X variables are not correlated with each other at all; (ii) When they are perfectly intercorrelated.
- In practice, it is usually somewhere in between (i) and (ii).

Uncorrelated X Variables

- Under standardized model: $\mathbf{r}_{XX} = \mathbf{I}_{p-1}$
- Fitted standardized regression coefficients:

$$\hat{\beta}_k^* = \sqrt{n-1} s_Y \times r_{YX_k}, \quad k = 1, \cdots, p-1$$

  are the sample correlation coefficients (up to a scaling factor) between the response variable Y and the respective X variables.
- Variance-covariance matrix:

$$\sigma^2\{\begin{bmatrix} \hat{\beta}_0^* \\ \hat{\beta}_1^* \\ \hat{\beta}_2^* \\ \vdots \\ \hat{\beta}_{p-1}^* \end{bmatrix}\} = \sigma^2(X^{*,T}X^*)^{-1} = \sigma^2 \begin{bmatrix} \frac{1}{n} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{I}_{p-1} \end{bmatrix}.$$

  So the LS estimators of the standardized model are uncorrelated. *How about the LS estimators of the original model?*

When the X variables are uncorrelated, the effect of an X variable does **not** depend on other X variables in the model.

- The LS fitted regression coefficient of an X variable is **not** affected by which other (uncorrelated) X variables are in the model.
- The LS fitted regression coefficients of the X variables are uncorrelated with each other.
- The contribution of an X variable in reducing the error sum of squares is the **same** with or without other (uncorrelated) X variables in the model, i.e.

$$SSR(X_j|X_{\mathcal{I}}) = SSR(X_j).$$

This is a strong advocate for controlled experiments, since there it may be possible to use an *orthogonal design* where the levels of the X variables are chosen such that their sample correlations are (nearly) zero.

Crew Productivity

A study on the effect of work crew size (X_1) and level of bonus pay (X_2) on productivity (Y).

case	X1 crew-size	X2 bonus-pay	Y productivity
1	4	2	42
2	4	2	39
3	4	3	48
4	4	3	51
5	6	2	49
6	6	2	53
7	6	3	61
8	6	3	60

Pairwise correlation matrix.

	X1	X2	Y
X1	1.00	0.00	0.74
X2	0.00	1.00	0.64
Y	0.74	0.64	1.00

X_1 and X_2 are uncorrelated.

Crew Productivity: Model 1

```
Call:
lm(formula = Y ~ X1, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-6.750 -3.750  0.125  4.500  6.000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.500     10.111   2.324   0.0591 .
X1             5.375      1.983   2.711   0.0351 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.609 on 6 degrees of freedom
Multiple R-squared: 0.5505,     Adjusted R-squared: 0.4755
F-statistic: 7.347 on 1 and 6 DF,  p-value: 0.03508

> anova(fit1)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
X1         1 231.12 231.125   7.347 0.03508 *
Residuals  6 188.75  31.458
```

Crew Productivity: Model 2

```
Call:
lm(formula = Y ~ X2, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-7.000 -4.688 -0.250  5.250  7.250

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   27.250     11.608   2.348   0.0572 .
X2             9.250      4.553   2.032   0.0885 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.439 on 6 degrees of freedom
Multiple R-squared: 0.4076,     Adjusted R-squared: 0.3088
F-statistic: 4.128 on 1 and 6 DF,  p-value: 0.08846

> anova(fit2)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
X2         1 171.12 171.125  4.1276 0.08846 .
Residuals  6 248.75  41.458
```

Crew Productivity: Model 3

```
Call:
lm(formula = Y ~ X1 + X2, data = data)

Residuals:
     1      2      3      4      5      6      7      8
 1.625 -1.375 -1.625  1.375 -2.125  1.875  0.625 -0.375

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3750     4.7405   0.079 0.940016
X1            5.3750     0.6638   8.097 0.000466 ***
X2            9.2500     1.3276   6.968 0.000937 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.877 on 5 degrees of freedom
Multiple R-squared: 0.958,      Adjusted R-squared: 0.9412
F-statistic: 57.06 on 2 and 5 DF,  p-value: 0.000361

> anova(fit3)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X1         1 231.125 231.125  65.567 0.0004657 ***
X2         1 171.125 171.125  48.546 0.0009366 ***
Residuals  5  17.625   3.525
```

Perfectly Correlated X variables

A set of X variables is said to be *collinear* if one or several of them may be expressed as a linear combination of the other X variables (including $\mathbf{1}_n$).

- The design matrix $\underset{n\times p}{\mathbf{X}}$ is not of full column rank: $rank(\mathbf{X}) < p$. So the matrix $\underset{p\times p}{\mathbf{X}'\mathbf{X}}$ is not invertible.
- LS estimators are not well-defined because the least-squares equation
$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$$
has many solutions.
- This means that there exist many vectors $\mathbf{b}$ that minimize the least squares criterion:
$$Q(\mathbf{b}) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{i1} - \cdots - b_{p-1} X_{i,p-1})^2 .$$

- If X variables are perfectly correlated, then there exists a nonzero vector $\underset{p\times 1}{\mathbf{c}}$ such that

$$\underset{n\times p}{\mathbf{X}}\underset{p\times 1}{\mathbf{c}} = \mathbf{0}_n.$$

- If $\mathbf{b}$ is a solution to the least-squares equation, i.e.,

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y},$$

then $\mathbf{b} + k\mathbf{c}$ is also a solution where $k \in \mathbb{R}$ is an arbitrary scalar since

$$\begin{aligned} \mathbf{X}'\mathbf{X}(\mathbf{b} + k\mathbf{c}) &= \mathbf{X}'\mathbf{X}\mathbf{b} + k\mathbf{X}'\mathbf{X}\mathbf{c} \\ &= \mathbf{X}'\mathbf{Y} + k\mathbf{X}'\mathbf{0}_n = \mathbf{X}'\mathbf{Y}. \end{aligned}$$

- Similarly, if $\mathbf{b}$ minimizes the least-squares criterion function $Q(\cdot)$, then $\mathbf{b} + k\mathbf{c}$ also minimizes $Q(\cdot)$ since

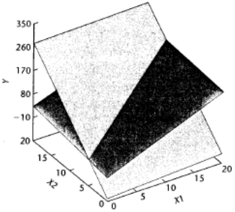
$$\begin{aligned} Q(\mathbf{b}) &= (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b}) \\ &= (\mathbf{Y} - \mathbf{X}(\mathbf{b} + k\mathbf{c}))'(\mathbf{Y} - \mathbf{X}(\mathbf{b} + k\mathbf{c})) = Q(\mathbf{b} + k\mathbf{c}). \end{aligned}$$

Example

case	X1	X2	Y
1	2	6	24
2	8	9	82
3	6	8	66
4	10	10	98

- X variables (including the column of 1) are perfectly correlated since $X_2 = 5 + 0.5X_1$.
- There are infinitely many response functions that fit this data equally "best" (with $SSE = 17.14$).

FIGURE 7.2 Two Response Planes That Intersect when $X_2 = 5 + .5X_1$.



- The two response surfaces in the figure are completely different, but they have the same y values on $X_2 = 5 + 0.5X_1$: $y = 7.14 + 9.29X_1$.
- Actually, any response surface that passes the intersecting line will fit the data equally well as these two, e.g.,

$$\widehat{Y} = 7.14 + 9.29X_1, \quad \widehat{Y} = -85.71 + 18.57X_2.$$

Can you think about some others?

```
Call:
lm(formula = Y ~ X1, data = data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.1429     3.5341   2.021  0.18066
X1            9.2857     0.4949  18.764  0.00283 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.928 on 2 degrees of freedom
Multiple R-squared: 0.9944,     Adjusted R-squared: 0.9915
F-statistic: 352.1 on 1 and 2 DF,  p-value: 0.002828

Call:
lm(formula = Y ~ X2, data = data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -85.7143     8.2956  -10.33  0.00924 **
X2           18.5714     0.9897   18.76  0.00283 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.928 on 2 degrees of freedom
Multiple R-squared: 0.9944,     Adjusted R-squared: 0.9915
F-statistic: 352.1 on 1 and 2 DF,  p-value: 0.002828
```

```
Call:
lm(formula = Y ~ X1 + X2, data = data)

Residuals:
      1       2       3       4 
-1.7143  0.5714  3.1429 -2.0000 

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   7.1429     3.5341   2.021  0.18066   
X1            9.2857     0.4949  18.764  0.00283 **
X2                NA         NA      NA       NA   
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.928 on 2 degrees of freedom
Multiple R-squared: 0.9944,     Adjusted R-squared: 0.9915 
F-statistic: 352.1 on 1 and 2 DF,  p-value: 0.002828
```

Here, R discards X_2 and fits a model only using X_1.

When X variables are perfectly correlated, we may still get a good fit of the data.

- The least-squares fitted values $\widehat{\mathbf{Y}}$ is uniquely defined and is the orthogonal projection of the response vector $\mathbf{Y}$ to the linear subspace of $\mathbb{R}^n$ generated by the columns of the design matrix $\mathbf{X}$ (the column space).
- Estimation of mean responses and predictions of new observations are still possible if they are done within the **row space** of the design matrix. (Read the next slides)
- However, the regression coefficients are not meaningful anymore without additional constraints.

```
> newX=data.frame(X1=3, X2=5)
> predict.lm(fit1, newX, interval="confidence",se.fit=TRUE)
$fit
  fit     lwr     upr
1  35 25.2425 44.7575

$se.fit
[1] 2.267787

$df
[1] 2

$residual.scale
[1] 2.9277

> predict.lm(fit2, newX,interval="confidence", se.fit=TRUE)
$fit
       fit       lwr      upr
1 7.142857 -8.063107 22.34882

$se.fit
[1] 3.534091

$df
[1] 2

$residual.scale
[1] 2.9277
```

```
> predict.lm(fit3, newX,interval="confidence",se.fit=TRUE)
$fit
  fit     lwr     upr
1  35 25.2425 44.7575

$se.fit
[1] 2.267787

$df
[1] 2

$residual.scale
[1] 2.9277

Warning message:
In predict.lm(fit3, newX, interval = "confidence", se.fit = TRUE) :
  prediction from a rank-deficient fit may be misleading
```

Body Fat

Correlation matrices.

$$\mathbf{r}_{XX} = \begin{bmatrix} 1.00 & 0.92 & 0.46 \\ 0.92 & 1.00 & 0.08 \\ 0.46 & 0.08 & 1.00 \end{bmatrix}, \quad \mathbf{r}_{XY} = \begin{bmatrix} 0.84 \\ 0.88 \\ 0.14 \end{bmatrix}.$$

X_1 and X_2 are highly correlated, X_1 and X_3 are moderately correlated, X_2 and X_3 are not much correlated.

Variables in Model	$\hat{\beta}_1$	$\hat{\beta}_2$	$s\{\hat{\beta}_1\}$	$s\{\hat{\beta}_2\}$	MSE
Model 1: X_1	0.8572	-	0.1288	-	7.95
Model 2: X_2	-	0.8565	-	0.1100	6.3
Model 3: X_1, X_2	0.2224	0.6594	0.3034	0.2912	6.47
Model 4: X_1, X_2, X_3	4.334	-2.857	3.016	2.582	6.15

- The regression coefficient for X_1 (X_2) varies drastically depending on which other X variables are included in the model.
- The standard errors of the fitted regression coefficients are becoming inflated when more X variables are included into the model.
- MSE tends to decrease as additional X variables are added into the model.

◂ R outputs

- $SSR(X_1) = 352.27$, $SSR(X_1|X_2) = 3.47$.
- The reason why $SSR(X_1|X_2)$ is so small compared to $SSR(X_1)$ is that X_1 and X_2 are highly correlated with each other **and with the response variable Y.**
    - When X_2 is already in the model, the marginal contribution from X_1 in explaining Y is small since X_2 contains much of the same information as X_1 in terms of explaining Y.

What would happen if X_1 and X_2 were not correlated with Y, but were highly correlated among themselves?

Body Fat: Model 1

```
> summary(fit1)

Call:
lm(formula = Y ~ X1, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4961     3.3192  -0.451    0.658
X1            0.8572     0.1288   6.656 3.02e-06 ***
---

Residual standard error: 2.82 on 18 degrees of freedom
Multiple R-squared: 0.7111,     Adjusted R-squared: 0.695
F-statistic:  44.3 on 1 and 18 DF,  p-value: 3.024e-06

> anova(fit1)
Analysis of Variance Table

Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27  44.305 3.024e-06 ***
Residuals 18 143.12    7.95
```

back

Body Fat: Model 2

```
> summary(fit2)

Call:
lm(formula = Y ~ X2, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.6345     5.6574  -4.178 0.000566 ***
X2            0.8565     0.1100   7.786  3.6e-07 ***

Residual standard error: 2.51 on 18 degrees of freedom
Multiple R-squared: 0.771,      Adjusted R-squared: 0.7583
F-statistic: 60.62 on 1 and 18 DF,  p-value: 3.6e-07

> anova(fit2)
Analysis of Variance Table

Response: Y
Df Sum Sq Mean Sq F value  Pr(>F)
X2         1 381.97  381.97  60.617 3.6e-07 ***
Residuals 18 113.42    6.30
```

back

Body Fat: Model 3

```
> summary(fit3)

Call:
lm(formula = Y ~ X1 + X2, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
X1            0.2224     0.3034   0.733   0.4737
X2            0.6594     0.2912   2.265   0.0369 *
---

Residual standard error: 2.543 on 17 degrees of freedom
Multiple R-squared: 0.7781,     Adjusted R-squared: 0.7519
F-statistic:  29.8 on 2 and 17 DF,  p-value: 2.774e-06

> anova(fit3)
Analysis of Variance Table

Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27 54.4661 1.075e-06 ***
X2         1  33.17   33.17  5.1284    0.0369 *
Residuals 17 109.95    6.47
```

back

Body Fat: Model 4

```
> summary(fit4)

Call:
lm(formula = Y ~ X1 + X2 + X3, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
X1             4.334      3.016   1.437    0.170
X2            -2.857      2.582  -1.106    0.285
X3            -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared: 0.8014,     Adjusted R-squared: 0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06

> anova(fit4)
Analysis of Variance Table

Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27 57.2768 1.131e-06 ***
X2         1  33.17   33.17  5.3931   0.03373 *
X3         1  11.55   11.55  1.8773   0.18956
Residuals 16  98.40    6.15
```

◂ partial

◂ multicollinearity

Effects of Multicollinearity: Summary

- With multicollinearity, the estimated regression coefficients tend to have large sampling variability (i.e., large standard errors). This leads to:
    - Wide confidence intervals.
    - It's possible that none of the regression coefficients is statistically significant, but at the same time there is a significant regression relation between the response variable and the entire set of X variables.
- Multicollinearity does not prevent us from getting a good fit of the data.

◂ prediction

Interpretation of Regression Coefficients and ESS

In the presence of multicollinearity:

- The regression coefficient of an X variable depends on which other X variables are also in the model.
- Therefore, a regression coefficient does **not** reflect any inherent effect of the corresponding X variable on the response variable, but only a marginal effect given whatever other X variables are also in the model.
- Similarly, there is **no** unique sum of squares that can be ascribed to any one X variable.
    - The reduction in the total variation in Y ascribed to an X variable must be interpreted as a margin reduction given other X variables also included in the model.

Quantify Multicollinearity: Variance Inflation Factor

Under the standardized model:

$$\boldsymbol{\sigma}^{\mathbf{2}}(\hat{\boldsymbol{\beta}}^{*}) = \sigma^2 \begin{bmatrix} \frac{1}{n} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{r}_{XX}^{-1} \end{bmatrix}$$

- The kth diagonal element of the inverse correlation matrix $\mathbf{r}_{XX}^{-1}$ is called the **variance inflation factor (VIF)** for $\hat{\beta}_k^*$, denoted by VIF_k.
- The variance of the estimated regression coefficient $\hat{\beta}_k^*$:

$$\sigma^2(\hat{\beta}_k^*) = VIF_k \sigma^2, \quad k = 1, \cdots, p-1.$$

- The variance of the estimated regression coefficient $\hat{\beta}_k$ in the original model:

$$\sigma^2(\hat{\beta}_k) = VIF_k \times \frac{\sigma^2}{\sum_{i=1}^n (X_{ik} - \bar{X}_k)^2}, \quad k = 1, \cdots, p-1.$$

It can be shown that

$$VIF_k = \frac{1}{1 - R_k^2} (\geq 1), \qquad k = 1, \cdots, p - 1,$$

where R_k^2 is the coefficient of multiple determination when X_k is regressed on the rest of X variables $\{X_j : 1 \leq j \neq k \leq p - 1\}$.

- If X_k is uncorrelated with the rest of the X variables, then $R_k^2 = 0$ and $VIF_k = 1$ (no inflation).
- If $R_k^2 > 0$, then $VIF_k > 1$, indicating an inflated variance for $\hat{\beta}_k^*$ (eqv. $\hat{\beta}_k$) due to the intercorrelation between X_k and the other X variables.
- If X_k has a perfect linear association with the rest of the X variables, then $R_k^2 = 1$, $VIF_k = \infty$ and so the variance of $\hat{\beta}_k^*$ (eqv. $\hat{\beta}_k$) is infinity (ill-defined).
- In practice, $\max_k VIF_k > 10$ is often taken as an indication that multicollinearity is high.

Body Fat

Correlation matrices.

$$\mathbf{r}_{XX} = \begin{bmatrix} 1.00 & 0.92 & 0.46 \\ 0.92 & 1.00 & 0.08 \\ 0.46 & 0.08 & 1.00 \end{bmatrix}, \quad \mathbf{r}_{XY} = \begin{bmatrix} 0.84 \\ 0.88 \\ 0.14 \end{bmatrix}.$$

X_1 and X_2 are highly correlated, X_1 and X_3 are moderately correlated, X_2 and X_3 are not much correlated. Moreover,

$$\mathbf{r}_{XX}^{-1} = \begin{bmatrix} 708.84 & -631.92 & -270.99 \\ -631.92 & 564.34 & 241.49 \\ -270.99 & 241.49 & 104.61 \end{bmatrix}$$

So,

$$R_1^2 = 0.9986, \quad R_2^2 = 0.9982, \quad R_3^2 = 0.9904.$$

Each predictor is highly intercorrelated with the rest of the predictors.

In Model 4, none of the three X variables is statistically significant by the T-tests. However, the F-test for regression relation is highly significant. Is there a paradox?

- From the general linear test perspective, each T-test is a marginal test, testing whether the marginal effect of an X variable is significant given **all other** X variables being included in the model.
- The three tests of the marginal effects of X_1, X_2, X_3 together are not equivalent to testing whether there is a regression relation between Y and (X_1, X_2, X_3).
- The reduced model for each individual test contains **all other** X variables and thus may lead to non-significant results due to multicollinearity.
- On the other hand, the reduced model for testing regression relation contains no X variable.

Model 4