Stat 206: Linear Models

Lecture 9

October 26, 2015

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 when finding its inversion.
- 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.

Example

A data set with a response variable Y in dollars, two predictor variables X_1 in thousand dollars and X_2 in cents. The fitted regression function:

$$\hat{Y} = 200 + 20,000X_1 + 0.2X_2.$$

- X_1 has a much larger fitted regression coefficient than X_2. However, this is caused by using different currency units.
- The effect of a \$1,000 increase in X_1 (corresponding to 1-unit increase) when X_2 is kept fixed would be an increase of \$20,000 in the mean response, while the effect of a \$1,000 increase in X_2 (corresponding to 100,000-unit increase) when X_1 is kept fixed is also a \$20,000 increase in the mean response.

Correlation Transformation

Define transformed variables:

$$
\begin{aligned}
Y_i^* &= \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \overline{Y}}{s_Y}\right), \quad i = 1, \cdots, n, \\
X_{ik}^* &= \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik} - \overline{X}_k}{s_{X_k}}\right), \quad k = 1, \cdots, p-1,
\end{aligned}
$$

where

$$
\begin{aligned}
\overline{Y} &= \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \overline{X}_k = \frac{1}{n}\sum_{i=1}^{n} X_{ik} \ \ (k = 1, \cdots, p-1), \\
s_Y &= \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}}, \quad s_{X_k} = \sqrt{\frac{\sum_{i=1}^{n}(X_{ik} - \overline{X}_k)^2}{n-1}},
\end{aligned}
$$

are sample means and sample standard deviations, respectively.

- The sample means of the transformed variables are
  .
- The sample standard deviations of the transformed variables are .
- So all variables are and are
  .
- Correlation transformation the pairwise (sample) correlations among the X variables, the (sample) correlations between the X variables and the response variable.

Standardized Regression Model

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

where

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

- The intercept parameter is omitted, as its LS estimator will always be . *Why?*
- Relationships between parameters in the standardized model and those in the original model:

Design Matrix of Standardized Model

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

This matrix only has $p - 1$ columns, since the regression intercept is omitted in the standardized model.

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

is the correlation matrix of the X variables.

Correlation Matrix

- Its (k, l)-element r_{kl} is the correlation coefficient between X_k, X_l:
- All its elements are numbers
  .
- Its diagonal elements are , since the correlation of a variable with itself is .
- Correlation matrix is a matrix:

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

Response vector of the standardized model:

$$\mathbf{Y}^* = \begin{bmatrix} Y_1^* \\ Y_2^* \\ \vdots \\ Y_n^* \end{bmatrix}.$$

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

where r_{Yk} is the 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 Estimators of Standardized Model

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

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

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

The relationships are the same as those between the true coefficients.

What are the relationships between sums of squares in standardized model and those in the original model?

Body Fat

Sample means and sample standard deviations:

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

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

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:

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\beta}_1^* \\ \hat{\beta}_2^* \\ \hat{\beta}_3^* \end{bmatrix} = \mathbf{r}_{XX}^{-1}\mathbf{r}_{XY} = \begin{bmatrix} 4.26 \\ -2.93 \\ -1.56 \end{bmatrix}.$$

Least-squares estimators of the original model:

Multicollinearity

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

- This term is often reserved for the situation when the inter-correlation/collinearity among the X variables is

  .
- X variables being nearly collinear 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

  .

Interpreting Regression Coefficients

- In presence of multicollinearity, the magnitude/sign of a coefficient

  interpreted as reflecting the comparative importance/direction-of-effect of the corresponding X variable.
- **A regression coefficient should be interpreted as reflecting the**

  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:

$$\mathbf{X}^{*\prime}\mathbf{X}^{*} = \mathbf{r}_{XX} = \qquad\qquad , \quad \mathbf{X}^{*\prime}\mathbf{Y}^{*} = \qquad\qquad .$$

- LS estimators:

$$\hat{\beta}_j^{*} = \qquad\qquad , \quad j = 1, \cdots, p-1,$$

are the between the response variable Y and individual X variables.

- Variance-covariance matrix:

$$\sigma^2\{\hat{\boldsymbol{\beta}}^{*}\} = \qquad\qquad .$$

- LS estimators under the original model:

When the X variables are uncorrelated, the effect of an X variable

depend on other X variables in the model.

- The LS fitted regression coefficient and its standard error are by which other X variables are in the model.
- The LS fitted regression coefficients are with each other.
- The marginal contribution of an X variable in reducing the error sum of squares is the
  other X variables in the model, i.e.
- 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 .

Consider the standardized models.

$$
\begin{aligned}
SSE(X_{\mathcal{I}}, X_j) &= \sum_{i=1}^{n}\left(Y_i^* - \sum_{k\in\mathcal{I}}\hat{\beta}_k^* X_{ik}^* - \hat{\beta}_j^* X_{ij}^*\right)^2 \\
&= \sum_{i=1}^{n}\left(Y_i^* - \sum_{k\in\mathcal{I}}\hat{\beta}_k^* X_{ik}^*\right)^2 - 2\sum_{i=1}^{n} Y_i^*\hat{\beta}_j^* X_{ij}^* \\
&+ \sum_{i=1}^{n}\left(\hat{\beta}_j^* X_{ij}^*\right)^2 \\
&= SSE(X_{\mathcal{I}}) + \sum_{i=1}^{n}\left(Y_i^* - \hat{\beta}_j^* X_{ij}^*\right)^2 - \sum_{i=1}^{n}(Y_i^*)^2 \\
&= SSE(X_{\mathcal{I}}) + SSE(X_j) - SSTO = SSE(X_{\mathcal{I}}) - SSR(X_j).
\end{aligned}
$$

Use the facts that the X variables are uncorrelated and the LS estimators do not depend on the model being fitted. The SSE and SSR of the original model are those of the standardized models multiplied by $(n-1)s_Y^2$*, so the above also holds for the original model.*

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 . So the matrix $\underset{p\times p}{\mathbf{X}'\mathbf{X}}$ is .
- LS estimators are because the least-squares equation

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

  has solutions.
- This means that there exist vectors $\mathbf{b}$ that minimize the least squares criterion:

$$Q(\mathbf{b}) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{i1} - \cdots - b_{p-1} X_{i,p-1})^2 .$$

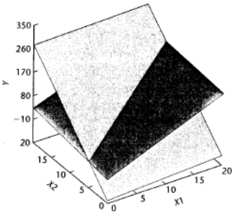
- If X variables are perfectly correlated, then there exists a nonzero vector $\underset{p\times 1}{\mathbf{c}}$ such that
$$\underset{n\times p}{\mathbf{X}}\underset{p\times 1}{\mathbf{c}} = \mathbf{0}_n.$$
- If $\mathbf{b}$ is a solution to the least-squares equation, i.e.,
$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y},$$
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 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 fit of the data.

- The least-squares fitted values $\widehat{\mathbf{Y}}$ is and is the

  of the response vector $\mathbf{Y}$ to the linear subspace of $\mathbb{R}^n$ generated by the columns of the design matrix $\mathbf{X}$.
- Estimation of mean responses and predictions of new observations are still possible if they

  What does this mean now?
- However, the regression coefficients are

  .

Estimation of Mean Response and Prediction

- Suppose $\mathbf{X}_h$ is a $p \times 1$ column vector satisfies:

$$\mathbf{X}_h' = \mathbf{c}'\mathbf{X} \quad (*)$$

  for some $n \times 1$ column vector $\mathbf{c}$, i.e., $\mathbf{X}_h$ is within the linear subspace of $\mathbb{R}^p$ generated by the rows of the design matrix $\mathbf{X}$ (the row space).

- The estimator of the mean response $E(Y_h)$ or the predicted value for Y_h is:

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

- Although the LS estimators $\widehat{\boldsymbol{\beta}}$ is not uniquely defined, the fitted values vector $\widehat{\mathbf{Y}}$ is and so is $\widehat{Y}_h$ as long as $\mathbf{X}_h$ satisfies $(*)$.

Example

```
> newX=data.frame(X1=3, X2=5+0.5*3)
> 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  35 25.2425 44.7575

$se.fit
[1] 2.267787

$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
```

```
> 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
```

Effects of Multicollinearity

- With multicollinearity, the estimated regression coefficients tend to have sampling variability (i.e., standard errors). This leads to:
  - confidence intervals.
  - It's possible that of the regression coefficients is statistically significant, but at the same time there is a regression relation between the response variable and the entire set of X variables.
- Multicollinearity does not prevent us from getting a of the data.
- Multicollinearity does not tend to severely affect inferences about mean response or prediction **if**

prediction

Interpretation of Regression Coefficients and ESS

In the presence of multicollinearity:

- The regression coefficient of an X variable which other X variables are also in the model.
- Therefore, a regression coefficient reflect any inherent effect of the corresponding X variable on the response variable, but only a given whatever other X variables are also in the model.
- Similarly, there is 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 given other X variables also included in the model.

Quantify Multicollinearity: Variance Inflation Factor

Under the standardized model:

$$\boldsymbol{\sigma}^2(\hat{\boldsymbol{\beta}}^*) = \qquad\qquad\qquad\qquad .$$

- 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 $\hat{\beta}_k^*$:

$$\sigma^2(\hat{\beta}_k^*) = \qquad\qquad\qquad\qquad .$$

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

$$\sigma^2(\hat{\beta}_k) = \qquad\qquad\qquad\qquad .$$

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

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 $R_k^2 =$ and $VIF_k =$.
- If $R_k^2 > 0$, then VIF_k , indicating an variance for $\hat{\beta}_k^*$ (eqv. $\hat{\beta}_k$) due to the between X_k and the other X variables.
- If X_k has a perfect linear association with the rest of the X variables, i.e., X_k is their , then $R_k^2 =$, $VIF_k =$ and so the variance of $\hat{\beta}_k^*$ (eqv. $\hat{\beta}_k$) is .
- 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,

Each predictor is with the rest of the predictors.

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) depending on which other X variables are included in the model.
- The standard errors of the fitted regression coefficients are becoming when more X variables are included into the model.
- MSE tends to 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 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 since X_2 contains much of the information as X_1 in terms of explaining Y.

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 test, testing whether the of an X variable is significant given X variables being included in the model.
- The three tests of the marginal effects of X_1, X_2, X_3 together are 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 X variables, whereas the reduced model for testing regression relation contains X variable.

Model 4

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
```

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

Body Fat

The precision of predictions within the region of the observations does not tend to be worsened with additional correlated X variables in the model.

```
> newX=data.frame(X1=25, X2=50, X3=29)
> predict.lm(fit1, newX, se.fit=TRUE)
$fit
       1
19.93356
$se.fit
[1] 0.6317416
...
> predict.lm(fit2, newX, se.fit=TRUE)
$fit
       1
19.19284
$se.fit
[1] 0.5758769
...
> predict.lm(fit3, newX, se.fit=TRUE)
$fit
       1
19.35566
$se.fit
[1] 0.6243083
...
> predict.lm(fit4, newX, se.fit=TRUE)
$fit
       1
19.19885
$se.fit
[1] 0.6194612
```