

# Linear Regression

Professor Jie Peng, PhD

Department of Statistics

University of California, Davis

# Multiple Regression: General Linear Tests

# General Linear Tests

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

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

$$H_0 : \beta_j = 0, \text{ for all } j \in \mathcal{J} \quad \text{vs.} \quad \underline{H_a : \text{not all } \beta_j : j \in \mathcal{J} \text{ is zero}}$$

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

# F Test

Compare SSE under the full model with SSE under the reduced model by an F ratio: Full model's SSE is certainly no larger than reduced model's SSE

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} = \frac{MSR(X_J | X_I)}{MSE(F)}$$

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

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

$$\frac{SSE(-X_i) - SSE(F)}{1} / \frac{SSE(F)/n-p}{1}$$

$F_{1, n-p}$

$$pF(x, 1, n-p) < pF(x, p-1, n-p)$$

- Reject  $H_0$  at level  $\alpha$  iff the observed

$$F^* > F(1 - \alpha; df_R - df_F, df_F).$$

$qF$

$$\frac{SSTO - SSE(F)}{p-1} / \frac{SSE(F)/n-p}{1}$$

$F_{p-1, n-p}$

# **Multiple Regression:**

## **General Linear Tests**

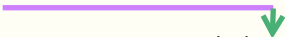
### **Examples**

# F-test for Regression Relation

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

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

- ▶ Reduced model with no  $X$  variable:

$b_0 = \text{mean}(Y)$  

$$Y_i = \beta_0 + \epsilon_i, \quad i = 1, \dots, n, \quad SSE(R) = SSTO, \quad df_R = n - 1$$

- ▶  $SSE(R) - SSE(F) = SSTO - SSE(F) = SSR(F)$ , and

$$df_R - df_F = (n - 1) - (n - p) = p - 1 = d.f.(SSR(F))$$

- ▶  $F^* = \frac{SSR(F)/(p-1)}{SSE(F)/(n-p)} = \frac{MSR(F)}{MSE(F)} \quad \sim H_0 \quad F(p-1, n-p) \quad H_0: b_1 = \dots = b_{p-1} = 0$

## Test whether a Single $\beta_k = 0$

H0:  $\beta_3=0$

Body Fat: for the model with all three predictors, test whether the midarm circumference ( $X_3$ ) can be dropped.

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

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

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

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

- ▶  $F^* = \frac{11.55/1}{98.40/16} = 1.88$ ; Pvalue= $P(F_{1,16} > 1.88) = 0.189$ , so  $X_3$  can be dropped. we cannot reject H0, which means  $\beta_3=0$  can not be rejected.

# Equivalence between F-test and T-test

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

► 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  $\beta_k$  and  $s\{\hat{\beta}_k\}$  is its standard error. At level  $\alpha$ , reject  $H_0$  when  $|T^*| > t(1 - \alpha/2; n - p)$ .

►  $F^* = (T^*)^2$  and  $F(1 - \alpha; 1, n - p) = (t(1 - \alpha/2; n - p))^2 \rightarrow$  F-test and two-sided T-test are equivalent.

*For one-sided alternatives, we still need the T-tests.*



## Test whether Several $\beta_k = 0$

Body Fat: Test whether both  $X_2$  and  $X_3$  can be dropped from the model with all three predictors:

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

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

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

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

- ▶  $F^* = \frac{44.72/2}{98.40/16} = 3.635$ ; Pvalue =  $P(F_{2,16} > 3.635) = 0.0499$

reject  $H_0$  at sig.lev=0.05 we cannot drop (x2,x3) together

## Test Equality of Several $\beta_k$ s

- ▶ Full model:  $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$
- ▶ For  $q \leq p-1$ :  $H_0 : \beta_1 = \cdots = \beta_q$  vs.  $H_a : \beta_1, \cdots, \beta_q$  are not all equal
- ▶ Reduced model:  $Y_i = \beta_0 + \beta_c (X_{i1} + \cdots + X_{iq}) + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$
- ▶  $\beta_c$  denotes the common value of  $\beta_1, \cdots, \beta_q$  under  $H_0$ , and  $X_1 + \cdots + X_q$  is the corresponding (new)  $X$  variable.  $SSE(R)$  has d.f.  $n - (p - q + 1)$ .
- ▶  $F^* = \frac{(SSE(R) - SSE(F))/(q-1)}{SSE(F)/(n-p)} \underset{H_0}{\sim} F_{q-1, n-p}$

# **Multiple Regression: Regression Coefficients as Partial Coefficients**

# Coefficient of Partial Determination

只加一个

Proportional reduction in SSE by adding one  $X$  variable into a model: ( $j \notin I$ )

$$R_{Y,j|I}^2 := \frac{SSE(X_I) - SSE(X_{j \cup I})}{SSE(X_I)} = \frac{SSR(X_j|X_I)}{SSE(X_I)}$$

- ▶ Between 0 and 1
- ▶ Example:  $R_{Y,1|2}^2 = \frac{SSR(X_1|X_2)}{SSE(X_2)}$  is the proportional reduction in SSE by including  $X_1$  into the model with  $X_2$ .

# 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.

## Boy Fat: Model 3

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

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

# Boy Fat: Model 4

```
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

Analysis of Variance Table

Response: Y

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

SSR(x1)

SSR(x2 | x1)

SSR(x3 | x1, x2)

SSE(all)

## Body Fat

$$SSE(x) = SSTO - SSR(x) = SSE(\text{all}) + SSR(\text{not } x)$$

$$R_{Y,2|1}^2 = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{33.17}{33.17 + 11.55 + 98.40} = 23.2\%.$$


$$R_{Y,3|12}^2 = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} = \frac{11.55}{11.55 + 98.40} = 10.5\%.$$

When  $X_2$  is added to the model containing  $X_1$ , SSE is reduced by 23.2%; When  $X_3$  is added to the model containing  $X_1, X_2$ , SSE is reduced by 10.5%.



# Extra Sum of Squares as SSR and Interpretation of Coefficient of Partial Determination

把新的一列X  
当成Y  
用别的列  
回归它  
得到 $\hat{X}_j$



It can be shown that:

- ▶  $SSR(X_j|X_I)$  is the SSR when regressing the residuals  $e(Y|X_I) = Y - \hat{Y}(X_I)$  to the residuals  $e(X_j|X_I) = X_j - \hat{X}_j(X_I)$ .  
the same as "... to  $X_j$ "
- ▶  $R^2_{Y, \cdot | I}$  is the coefficient of simple determination between the two sets of residuals. = squared correlation coefficient between the two sets of residuals.
- ▶  $R^2_{Y, \cdot | I}$  thus measures linear association between  $Y$  and  $X_j$  after the linear effects of  $X_I$  have been adjusted for.

Example:  $R_{Y,1|2}^2$

- ▶ Regress  $Y$  on  $X_2$ :  $e_i(Y|X_2) = Y_i - \widehat{Y}_i(X_2)$ ,  $i = 1, \dots, n$ .
- ▶ Regress  $X_1$  on  $X_2$ :  $e_i(X_1|X_2) = X_{i1} - \widehat{X}_{i1}(X_2)$ ,  $i = 1, \dots, n$ .
- ▶  $R_{Y1|2}^2$  equals to the coefficient of simple determination between  $e_i(Y|X_2)$  and  $e_i(X_1|X_2)$ .
- ▶ It measures the linear association between  $Y$  and  $X_1$  after the linear effects of  $X_2$  have been adjusted for.

# Partial Correlations

The **signed** square-root of a coefficient of partial determination is called a partial correlation.

- ▶ The sign is the same as the sign of the corresponding fitted regression coefficient (in the larger model).
- ▶ Partial correlation is the correlation coefficient between the two respective sets of residuals.

# Body Fat

$$r_{Y2|1} = \sqrt{R^2_{Y2|1}} * \text{sign}(\hat{\beta}_2)$$

- ▶  $r_{Y2|1} = \sqrt{0.232} = 0.482$ , since in Model 3,  $\hat{\beta}_2 > 0$ .
- ▶  $r_{Y3|12} = -\sqrt{0.105} = -0.324$ , since in Model 4,  $\hat{\beta}_3 < 0$ .

# LS Fitted Regression Coefficients as Partial Coefficients

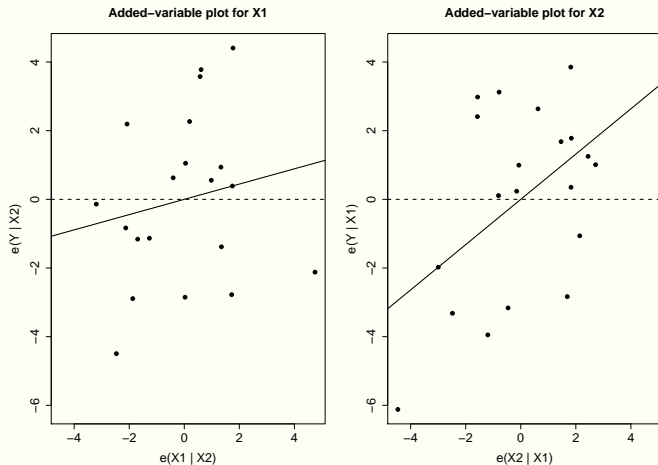
The LS fitted regression coefficients  $\hat{\beta}$  are indeed partial coefficients.

- ▶ Consider  $p - 1$   $X$  variables in the model. Let  $\hat{\beta}_j$  be the LS fitted regression coefficient for  $X_j$ .
- ▶ Then  $\hat{\beta}_j$  equals to the LS fitted regression coefficient when regressing the residuals  $e(Y|X_{-(j)}) = Y - \hat{Y}(X_{-(j)})$  to the residuals  $e(X_j|X_{-(j)}) = X_j - \hat{X}_j(X_{-(j)})$ , where  $X_{-(j)} = \{X_l : 1 \leq l \neq j \leq p\}$ .

## Added-Variable Plots

- ▶ Both the response variable  $Y$  and  $X_j$  are regressed onto the rest of the  $X$  variables, denoted by  $X_{-(j)}$ , in the model.
- ▶ The residuals reflect the part of  $Y$  ( $X_j$ ) that is not linearly associated with the rest of the  $X$  variables.
- ▶ The plot of these two sets of residuals against each other:
  - ▶ shows the marginal importance of  $X_j$  in reducing the residual variability in  $Y$  after accounting for the linear effects in the rest of the  $X$  variables.
  - ▶ provides information about the nature of the marginal effect of  $X_j$  on  $Y$ , e.g., linear or curvilinear.

Figure: Body Fat  $Y \sim X_1, X_2$ : Added-variable plots



- ▶ Added-variable plot for  $X_1$  (given  $X_2$ ) implies that  $X_1$  is of not much additional help in explaining  $Y$  when  $X_2$  is already in the model. This is consistent with  $R^2_{Y1|2} = 3.1\%$ .
- ▶ Added-variable plot for  $X_2$  (given  $X_1$ ) shows that  $X_2$  is of some help in explaining  $Y$  when  $X_1$  is already in the model. From previous slides,  $R^2_{Y2|1} = 23.2\%$ . It also shows that a linear term of  $X_2$  in the model is adequate.



# Standardization

# Standardization

$X$  values could differ substantially in order of magnitude. This could lead to:

- ▶ Regression coefficients not comparable
- ▶ Numerical instability in inverting  $\mathbf{X}'\mathbf{X}$

A regression model can be *reparametrized* into a *standardized regression model* through centering and rescaling.

## Transformed X Variables

$X^* =$

$1/\sqrt{n-1} * \text{scale}(X1..p-1)$

$X^*_{ik} = \frac{1}{\sqrt{n-1}} \left( \frac{X_{ik} - \bar{X}_k}{s_k} \right), \quad i = 1, \dots, n, \quad k = 1, \dots, p-1,$

where

$$\bar{X}_k = \frac{1}{n} \sum_{i=1}^n X_{ik}, \quad s_k = \sqrt{\frac{\sum_{i=1}^n (X_{ik} - \bar{X}_k)^2}{n-1}}$$

are sample mean and sample standard deviation of  $X_k$ , respectively.

The transformed X variables are centered and are on the same scale:

- ▶ Their sample means equal zero. 为了要使得 *sum of  $X_i$*  的 *r* 是 1，这边就是  $1/\text{sqrt}(n-1)$  了！
- ▶ Their sample standard deviations equal  $\frac{1}{\sqrt{n-1}}$ . 不是 1！

Moreover, standardization does not change pairwise sample correlations.

# Standardized Regression Model

# 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

$$\underline{\beta_k^* = \sqrt{n-1} s_k \beta_k}, \quad k = 1, \cdots, p-1,$$

$$\underline{\beta_0^* = \beta_0 + \sum_{k=1}^{p-1} \beta_k \bar{X}_k}$$

is a *reparametrization* of the original model.

## Standardized Model: Design Matrix

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

## Standardized Model: $\mathbf{X}'\mathbf{X}$

$$\boxed{\mathbf{X}^{*'}\mathbf{X}^*_{p \times p}} = \begin{bmatrix} n & 0 & 0 & \cdots & 0 \\ 0 & 1 & r_{12} & \cdots & r_{1,p-1} \\ 0 & r_{21} & 1 & \cdots & r_{2,p-1} \\ 0 & \vdots & \cdots & \vdots & \\ 0 & r_{p-1,1} & r_{p-1,2} & \cdots & 1 \end{bmatrix} = \boxed{\begin{bmatrix} n & \mathbf{0}^T \\ \mathbf{0} & \mathbf{r}_{XX} \\ & (p-1) \times (p-1) \end{bmatrix}},$$

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

$$(\mathbf{X}^{*'}\mathbf{X}^*)^{-1} = \begin{bmatrix} \frac{1}{n} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{r}_{XX}^{-1} \end{bmatrix}$$



## Sample Correlation Matrix $\mathbf{r}_{XX}$

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

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

---

- ▶ All elements are unit-less numbers between  $-1$  and  $1$ .
- ▶ All diagonal elements are one.
- ▶ Symmetric:  $r_{kl} = r_{lk}$

## Standardized Model: $\mathbf{X}'\mathbf{Y}$

$$\mathbf{X}^{*'}\mathbf{Y} = \begin{bmatrix} n\bar{Y} \\ \sqrt{n-1}s_Y r_1 \\ \sqrt{n-1}s_Y r_2 \\ \vdots \\ \sqrt{n-1}s_Y r_{p-1} \end{bmatrix} = \sqrt{n-1}s_Y \begin{bmatrix} \frac{n}{\sqrt{n-1}s_Y} \bar{Y} \\ \mathbf{r}_{XY} \\ (p-1) \times 1 \end{bmatrix},$$

where  $r_k$  is the sample correlation coefficient between  $Y$  and  $X_k$ :

$$r_k = \frac{\frac{1}{n-1} \sum_{i=1}^n (X_{ik} - \bar{X}_k)(Y_i - \bar{Y})}{s_k s_Y}, \quad k = 1, \dots, p-1$$

## Standardized Model: Least Squares Estimator

$$\hat{\boldsymbol{\beta}}^*_{p \times 1} = \begin{bmatrix} \hat{\beta}_0^* \\ \hat{\beta}_1^* \\ \hat{\beta}_2^* \\ \vdots \\ \hat{\beta}_{p-1}^* \end{bmatrix} = \begin{bmatrix} \bar{Y} \\ \sqrt{n-1} \mathbf{s}_Y \mathbf{r}_{XX}^{-1} \mathbf{r}_{XY} \end{bmatrix}$$

These are called the *fitted standardized regression coefficients*.

$$E(\hat{\boldsymbol{\beta}}^*) = \boldsymbol{\beta}^*, \quad \sigma^2\{\hat{\boldsymbol{\beta}}^*\} = \sigma^2(\mathbf{X}^{*'}\mathbf{X}^*)^{-1} = \sigma^2 \begin{bmatrix} \frac{1}{n} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{r}_{XX}^{-1} \end{bmatrix}$$

# Relationships with the Original Model

- ▶ Fitted regression coefficients:

$$\hat{\beta}_k^* = \sqrt{n-1} s_k \hat{\beta}_k, \quad k = 1, \dots, p-1$$

$$\hat{\beta}_0^* = \hat{\beta}_0 + \sum_{k=1}^{p-1} \hat{\beta}_k \bar{X}_k (= \bar{Y})$$

- ▶ Fitted values, residuals, and sums of squares are the same as the original model.

# Uncorrelated X Variables

# Uncorrelated X Variables

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

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

- ▶ Variance-covariance matrix:

$$\sigma^2\{\hat{\beta}^*\} = \sigma^2 \begin{bmatrix} \frac{1}{n} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{I}_{p-1} \end{bmatrix}$$

If the  $X$  variables in the model are uncorrelated with each other, then the effect of one  $X$  variable does not depend on other  $X$  variables:

- ▶ The fitted regression coefficient of an  $X$  variable is not affected by other  $X$  variables.
- ▶ The fitted regression coefficients are uncorrelated with each other.
- ▶ The contribution of an  $X$  variable in reducing the error sum of squares equals its marginal effect:

$$SSR(X_j|X_{-(j)}) = SSR(X_j).$$

## Example: Crew Productivity

A study on the effect of work crew size ( $X_1$ ) and level of bonus pay ( $X_2$ ) on productivity ( $Y$ ). The levels of  $X_1$  and  $X_2$  are chosen such that they are uncorrelated (this is called an *orthogonal design*).

case	X1	X2	Y
crew-size	bonus-pay	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



# Crew Productivity: Model 1

Call:

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

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 \*

---

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)
```

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 .

---

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)
```

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 ***

---

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

©Jie Peng 2020. This content is protected and may not be shared, uploaded, or distributed.

# Multicollinearity

# 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 high.
- ▶ This means there exists a nonzero vector  $\mathbf{c}$  such that

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

- ▶ Consequently, the matrix  $\underset{p \times p}{\mathbf{X}'\mathbf{X}}$  would be nearly singular.

$$\text{det}(\text{Singular})=0$$

# Body Fat

Variables:  $Y$  : body fat,  $X_1$ : triceps skinfolds thickness,  $X_2$ : thigh circumference,  $X_3$ : midarm circumference

Sample correlation matrix:

	X1	X2	X3	Y
X1	1.00000000	0.9238425	0.4577772	0.8432654
X2	0.9238425	1.00000000	0.0846675	0.8780896
X3	0.4577772	0.0846675	1.00000000	0.1424440
Y	0.8432654	0.8780896	0.1424440	1.00000000

## Compare Models

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 fitted regression coefficient for  $X_1$  ( $X_2$ ) varies drastically depending on which other  $X$  variables are in the model.
- ▶ The standard errors of the fitted regression coefficients are inflated when more  $X$  variables are added to the model.

- ▶  $X_1$  and  $X_2$  are highly correlated with each other **and** with the response variable  $Y$ .
- ▶ When  $X_2$  is already in the model, the additional contribution from  $X_1$  in explaining  $Y$  is small since  $X_2$  contains much of the same information in terms of explaining  $Y$ :

$$SSR(X_1) = 352.27, \quad SSR(X_1|X_2) = 3.47$$



# Boy Fat: Model 4

```
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

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		

*F-test significant* = 不能删除  $H_0$  中等于 0 的变量 = 有相关

None of the X variables is statistically significant by the T test.

However, the F-test of regression relation is highly significant.

- ▶ The reduced model for each individual T test contains all other X variables and thus may be non-significant due to

multicollinearity.  $b_i = ((X'X)^{-1}X'Y)_i$ ,  $s(b_i) = (SSE/(n-p)(X'X)^{-1})_{i,i}$   
 $t_i = (b_i - 0)/s(b_i)$

- ▶ The reduced model for the F test contains no X variable.  
*reduced model*  $Y \sim \epsilon$  整个 F test 里还是有所有的 X 变量的，  
只是 F test 里的 reduced model 里没有 X 变量
- ▶ The three T tests together are not equivalent to testing

whether there is a regression relation between Y and the set of X variables (i.e., F test).

一个对 T test 显著的 X 不能保证 all Xs as a whole 对与 F test 都显著

*F test:  $(SSE(R) - SSE(F))(df_{EF}) / [SSE(F)(df_{ER} - df_{EF})] = (SSR(F))(df_{EF}) / [SSE(F)(df_{ER} - df_{EF})]$*

# Effects of Multicollinearity

With multicollinearity, the estimated regression coefficients tend to have large sampling variability (i.e., large standard errors)  $\implies$

- ▶ Wide confidence intervals
- ▶ It's possible that none of the regression coefficients is statistically significant, but there is a significant regression relation between the response variable and the entire set of  $X$  variables.

However, multicollinearity does not prevent us from getting a good fit of the data.

## With multicollinearity:

- ▶ The regression coefficient of an  $X$  variable depends on which other  $X$  variables also in the model.
- ▶ So regression coefficient does not reflect any inherent effect of the corresponding  $X$  variable, but reflects only a marginal effect given whatever other  $X$  variables also in the model.
- ▶ Similarly, 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 in the model.

# Variance Inflation Factor

## Quantify Multicollinearity

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

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

$$\sigma^2\{\hat{\beta}_k^*\} = VIF_k \sigma^2, \quad k = 1, \dots, p-1$$

It can be shown that  $VIF_k = \frac{1}{1-R_k^2}$ , where  $R_k^2$  is the coefficient of multiple determination when  $X_k$  is regressed onto the rest  $X$  variables:

- ▶  $VIF_k \geq 1$
- ▶ If  $X_k$  is uncorrelated with the rest  $X$  variables, then  $R_k^2 = 0$  and  $VIF_k = 1 \implies$  no variance inflation
- ▶ If  $R_k^2 > 0$ , then  $VIF_k > 1 \implies$  an inflated variance of  $\hat{\beta}_k^*$  due to intercorrelation between  $X_k$  and the rest  $X$  variables
- ▶ If  $X_k$  has a perfect linear association with the rest  $X$  variables, then  $R_k^2 = 1$  and  $VIF_k = \infty \implies$  LS estimator not well defined

# Diagnostic of Multicollinearity by VIF

In practice, the largest VIF,  $\max_k VIF_k$ , greater than 10 is often taken as an indication of high multicollinearity.



## Body Fat

$$\mathbf{r}_{XX} = \begin{bmatrix} 1.00 & 0.92 & 0.46 \\ 0.92 & 1.00 & 0.08 \\ 0.46 & 0.08 & 1.00 \end{bmatrix}, \quad \mathbf{r}_{XX}^{-1} = \begin{bmatrix} 708.84 & -631.92 & -270.99 \\ -631.92 & 564.34 & 241.49 \\ -270.99 & 241.49 & 104.61 \end{bmatrix}$$

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

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

Each X variable is highly intercorrelated with the rest X variables.

# Identifiability

# Unidentifiability

A model is *unidentifiable* if its parameters can not be uniquely estimated. For regression models, unidentifiability occurs when

- ▶ columns of the design matrix  $\mathbf{X}_{n \times p}$  are linearly dependent (i.e., perfect collinearity)  $\implies \text{rank}(\mathbf{X}) < p$
- ▶  $\implies \mathbf{X}'\mathbf{X}_{p \times p}$  is not invertible
- ▶  $\implies$  LS estimator is not well defined because the normal equation  $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$  has many solutions
- ▶  $\implies$  there exist many vectors that minimize the least squares criterion.

# What Causes Unidentifiability?

Possible causes include:

- ▶ More variables than cases, i.e.,  $p > n \implies$  select a smaller subset of variables
- ▶ A feature is recorded by two different units and both are included in the model  $\implies$  remove redundancy
- ▶ Some linear combinations of variables are included in the model  $\implies$  eliminate them

By default, R will fit the largest identifiable model by removing variables in the reverse order of appearance in the model formula.

## 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”.

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 \*\*

---

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 \*\*

---

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

©Jie Peng 2020. This content is protected and may not be shared, uploaded, or distributed.

Call:

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

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

---

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

R discards  $X_2$  and fits a model only using  $X_1$ .

# Polynomial Regression



# Polynomial Regression

One of the most commonly used models to describe a curvilinear regression relation:

- ▶ very flexible and easy to fit
- ▶ higher than third-order terms are rarely employed in practice because of
  - ▶ high sampling variability
  - ▶ *overfitting*: fit the observed data well, but do not generalize well to new observations

# Centering

In practice, centered X variables  $\tilde{X}_k = X_k - \bar{X}_k$  are often used in polynomial regression models:

- ▶ Centering reduces the correlation between the linear term and the quadratic term substantially and thus improves numerical accuracy.
- ▶ Centering does not change the fitted regression function.

## Second-Order Model with One Predictor

$$\begin{aligned}Y_i &= \beta_0 + \beta_1(X_i - \bar{X}) + \beta_2(X_i - \bar{X})^2 + \epsilon_i, \quad i = 1, \dots, n \\&= \beta_0 + \beta_1\tilde{X}_i + \beta_2\tilde{X}_i^2 + \epsilon_i, \quad \tilde{X}_i = X_i - \bar{X}\end{aligned}$$

The response function is a *parabola*:

$$y = \beta_0 + \beta_1\tilde{x} + \beta_2\tilde{x}^2$$

- ▶  $\beta_0$  is the mean response when  $\tilde{x} = 0$  (i.e.  $x = \bar{X}$ ).
- ▶  $\beta_1$  is called the *linear effect coefficient* and  $\beta_2$  is called the *quadratic effect coefficient*.

## Second-Order Model with Two Predictors

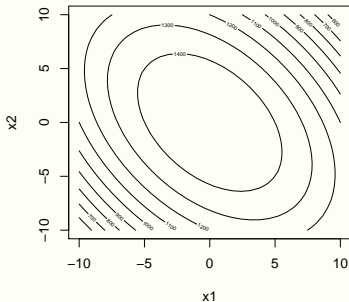
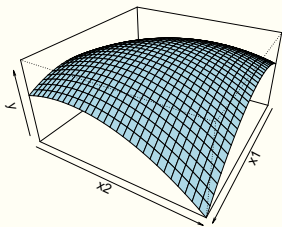
$$Y_i = \beta_0 + \beta_1 \tilde{X}_{i1} + \beta_2 \tilde{X}_{i2} + \beta_{11} \tilde{X}_{i1}^2 + \beta_{22} \tilde{X}_{i2}^2 + \beta_{12} \tilde{X}_{i1} \tilde{X}_{i2} + \epsilon_i, i = 1, \dots, n$$

- ▶ response surface is a *conic section*:

$$y = \beta_0 + \beta_1 \tilde{x}_1 + \beta_2 \tilde{x}_2 + \beta_{11} \tilde{x}_1^2 + \beta_{22} \tilde{x}_2^2 + \beta_{12} \tilde{x}_1 \tilde{x}_2$$

- ▶ separate *linear and quadratic terms* for each predictor
- ▶ a cross-product term representing the interaction between the two predictors
- ▶  $\beta_{12}$  is called the *interaction effect coefficient*

A quadratic response surface:  $y = 1500 - 4x_1^2 - 3x_2^2 - 3x_1x_2$



The contour plot shows combinations of  $(x_1, x_2)$  that yield the same value of  $y$ .

## Extension: Second-Order Model with $K$ Predictors

$$Y_i = \beta_0 + \sum_{k=1}^K \beta_k \tilde{X}_{ik} + \sum_{k=1}^K \beta_{kk} \tilde{X}_{ik}^2 + \sum_{1 \leq k < k' \leq K} \beta_{kk'} \tilde{X}_{ik} \tilde{X}_{ik'} + \epsilon_i, i = 1, \dots, n$$

- ▶ response function:

$$y = \beta_0 + \sum_{k=1}^K \beta_k \tilde{x}_k + \sum_{k=1}^K \beta_{kk} \tilde{x}_k^2 + \sum_{1 \leq k < k' \leq K} \beta_{kk'} \tilde{x}_k \tilde{x}_{k'}$$

- ▶  $\beta_k$ s are *linear effect coefficients*;  $\beta_{kk}$ s are *quadratic effect coefficients*.
- ▶  $\{\beta_{kk'} : 1 \leq k < k' \leq K\}$  are *interaction effect coefficients* between the respective pairs of predictors.

## Third-Order Model with One Predictor

$$\begin{aligned}Y_i &= \beta_0 + \beta_1(X_i - \bar{X}) + \beta_2(X_i - \bar{X})^2 + \beta_3(X_i - \bar{X})^3 + \epsilon_i, \quad i = 1, \dots, n \\&= \beta_0 + \beta_1\tilde{X}_i + \beta_2\tilde{X}_i^2 + \beta_3\tilde{X}_i^3 + \epsilon_i, \quad \tilde{X}_i = X_i - \bar{X}\end{aligned}$$

The response function is a cubic polynomial:

$$y = \beta_0 + \beta_1\tilde{x} + \beta_2\tilde{x}^2 + \beta_3\tilde{x}^3$$

# Polynomial Regression:

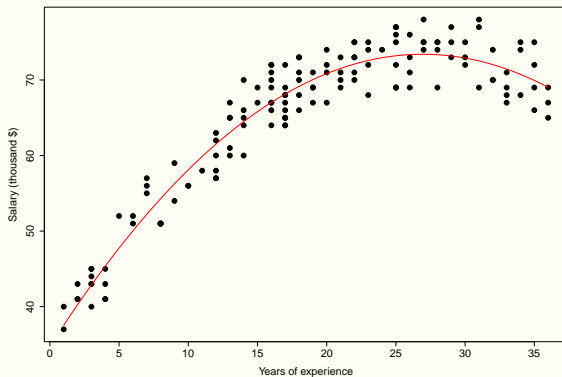
## Example



# Salary

Professional organizations regularly survey their members for information concerning salaries, pensions, and conditions of employment. One goal is to relate salary to years of experience. This data has years of experience ( $X$ ) and salary ( $Y$ ) on 143 cases.

# Salary: Scatter Plot



# Salary: Second-Order Model

```
> salary.c=salary
# Correlation coefficient between X and X^2 is 0.965 for the original variable "year of experience",
# and is -0.0414 for the centered variable
> salary.c[, "Experience"] = salary[, "Experience"] - mean(salary[, "Experience"]) ## center the X variable
> fitc = lm(Salary ~ Experience + I(Experience^2), data = salary.c) ## fit a second-order model
> summary(fitc)
```

Call:

```
lm(formula = Salary ~ Experience + I(Experience^2), data = salary.c)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	69.927208	0.323090	216.43	<2e-16 ***
Experience	0.861177	0.024957	34.51	<2e-16 ***
I(Experience^2)	-0.053316	0.002477	-21.53	<2e-16 ***

---

Residual standard error: 2.817 on 140 degrees of freedom

Multiple R-squared: 0.9247, Adjusted R-squared: 0.9236

F-statistic: 859.3 on 2 and 140 DF, p-value: < 2.2e-16

# Salary: Third-Order Model

Call:

```
lm(formula = Salary ~ Experience + I(Experience^2) + I(Experience^3),  
data = salary.c)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	69.9484745	0.3224575	216.92	<2e-16 ***
Experience	0.9364986	0.0603531	15.52	<2e-16 ***
I(Experience^2)	-0.0537196	0.0024866	-21.60	<2e-16 ***
I(Experience^3)	-0.0003957	0.0002888	-1.37	0.173

---

```
> anova(fit3)
```

Analysis of Variance Table

Response: Salary

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Experience	1	9962.9	9962.9	1263.1043	<2e-16 ***
I(Experience^2)	1	3677.9	3677.9	466.2810	<2e-16 ***
I(Experience^3)	1	14.8	14.8	1.8764	0.173
Residuals	139	1096.4	7.9		

First test whether the third-order term may be dropped.

- ▶ full model: third-order model vs. reduced model:  
second-order model
- ▶  $SSR(X^3|X, X^2) = 14.8$  with d.f. 1
- ▶  $SSE(X, X^2, X^3) = 1096.4$  with d.f. 139
- ▶ F-statistic = 1.876
- ▶ pvalue = 0.173
- ▶ Therefore, the third-order term is not significant and may be dropped

Then test whether the second-order term may be dropped.

- ▶ full model: second-order model vs. reduced model: first-order model
- ▶  $SSR(X^2|X) = 3677.9$  with d.f. 1
- ▶  $SSE(X, X^2) = SSE(X, X^2, X^3) + SSR(X^3|X, X^2) = 1111.2$   
with d.f. 140
- ▶ F-statistic = 466.28
- ▶ pvalue <  $2e - 16$
- ▶ Therefore, the second-order term is very significant and should be retained.