STAT 425

# Collinearity

# Collinearity

Consider a MLR model with a design matrix $\mathbf{X}_{n\times p}$ including the intercept. If the columns of $\mathbf{X}$ are **orthogonal** to each other (i.e., the sample correlation of any two predictors is equal to $0$), then the LS problem is greatly simplified:

$$\hat{\beta}_j = [(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{y}]_{\mathbf{j}} = \frac{\mathbf{X}_{.j}^\top\mathbf{y}}{||\mathbf{X}_{.j}||^2}$$

where $\mathbf{X}_{.j}$ denotes the $j$-th column of $\mathbf{X}$.

In other words, in this case (only) the LS regression coefficient for the $j$-th predictor does not depend on whether other predictors are included in the model or not.

# Collinearity

- In practice, we often encounter problems in which many of the predictors are highly correlated.
- In such cases, the values and sampling variance of regression coefficients can be highly dependent on the particular predictors chosen for the model.

# Exact Collinearity

- If there exists a set of constants $c_1, c_2, \dots, c_p$ (at least one of them is non-zero), such that the corresponding linear combination of the columns of $\mathbf{X}$ is zero, i.e.:

$$\sum_{j=1}^{p} c_j \mathbf{X}_{.j} = \mathbf{0}$$

  then the columns of $\mathbf{X}$ are called linearly dependent and there is exact collinearity. That is, at least one column in the design matrix X can be expressed as a linear combination of other columns.

## What happens when the columns of $\mathbf{X}$ are collinear?

1. $(\mathbf{X}^\top \mathbf{X})^{-1}$ does not exist,
2. The LS estimate $\hat{\boldsymbol{\beta}}$ is not unique, and
3. The coefficients of the linear model are not identifiable.

**Example:** Suppose the 1st column of $\mathbf{X}$ is the intercept, and the 2nd column of $\mathbf{X}$ is the vector $(2, 2, \ldots, 2)^\top$. Then if $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots)^\top$ is one LS estimate of $\boldsymbol{\beta}$, the vector $(\hat{\beta}_1 - c, \hat{\beta}_2 + c/2, \hat{\beta}_3, \ldots)^\top$ is also an estimate of $\boldsymbol{\beta}$, where $c$ is any real number.

**Note:** In case of exact collinearity the column space of $\mathbf{X}$ has dimension $< p$. In this case we can often fit an equivalent model by eliminating one or more redundant variables.

# Approximate Collinearity

- We generally do not need to worry about exact collinearity [1], but approximate collinearity. That is, at least one column $\mathbf{X}_{,j}$ can be approximated by the others:

$$\mathbf{X}_{.k} \approx -\sum_{j \neq k} c_j \mathbf{X}_{.j} / c_k$$

A simple diagnostic for this is to obtain the regression of $\mathbf{X}_{.k}$ on the remaining predictors, and if the corresponding $R_k^2$ is close to 1, we would diagnose approximate collinearity.

[1] **R** can detect it and fix it automatically

## Why approximate collinearity is a problem?

- In a multiple regression $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + e$, the LS estimate $\hat{\beta}_k$ is unbiased with variance:

$$var(\hat{\beta}_k) = \sigma^2 \left( \frac{1}{1 - R_k^2} \right) \left( \frac{1}{\sum_{i=1}^{n} (x_{ik} - \bar{x}_{.k})^2} \right)$$

  where $R_k^2$ is the R-square from the regression of $\mathbf{X}_{.k}$ on the remaining predictors. When $R_k^2$ is close to 1, the variance of $\hat{\beta}_k$ is large. Consequently we will have:

  1. large Mean Square Error
  2. large (inflated) p-value to the corresponding t-test, i.e, we could miss a significant predictor.

- The quantity $\left( \frac{1}{1-R_k^2} \right)$ is the variance inflation factor (VIF) for the $k$-th coefficient of the model

# Example: Car position data

Data on 38 drivers:

- Age: Drivers age in years
- Weight: Drivers weight in lbs
- HtShoes: height with shoes in cm
- Ht: height without shoes in cm
- Seated: seated height in cm
- Arm: lower arm length in cm
- Thigh: thigh length in cm
- Leg: lower leg length in cm
- hipcenter: horizontal distance of the midpoint of the hips from a fixed location in the car in mm

```
library(faraway)
data(seatpos)
attach(seatpos)
g=lm(hipcenter ~ ., seatpos)
summary(g)
```

# Example: Car position data

Collinearity Symptoms: None of the individual variables is significant. Large standard errors. High correlation among variables

```
##
## Call:
## lm(formula = hipcenter ~ ., data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.827 -22.833  -3.678  25.017  62.337
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 436.43213  166.57162   2.620   0.0138 *
## Age           0.77572    0.57033   1.360   0.1843
## Weight        0.02631    0.33097   0.080   0.9372
## HtShoes      -2.69241    9.75304  -0.276   0.7845
## Ht            0.60134   10.12987   0.059   0.9531
## Seated        0.53375    3.76189   0.142   0.8882
## Arm          -1.32807    3.90020  -0.341   0.7359
## Thigh        -1.14312    2.66002  -0.430   0.6706
## Leg          -6.43905    4.71386  -1.366   0.1824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.72 on 29 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6001
## F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

Calculate Variance Inflation Factor of model matrix $X$ (after removing the first column) using function **vif(.)**.

```
# Variance Inflation Factor (VIF)
round(vif(x), dig=2)
```

```
##     Age  Weight HtShoes      Ht  Seated     Arm   Thigh     Leg
##    2.00    3.65  307.43  333.14    8.95    4.50    2.76    6.69
```

```
sqrt(307.43)
```

```
## [1] 17.53368
```

Standard error of the estimated predictor $\hat{\beta}_{HtShoes}$ is approximately 17 times larger than it would have been without collinearity.

# A global measure of collinearity

- A global measure of collinearity is given by examining the eigenvalues of $\mathbf{X}^\top\mathbf{X}$. A popular measure is the condition number of $\mathbf{X}^\top\mathbf{X}$, denoted by:

$$\kappa = (\text{largest eigenvalue}/\text{smallest eigenvalue})^{1/2}$$

  An empirical rule for declaring collinearity is $\kappa \geq 30$

- Note that $\kappa$ is not scale-invariant, so we should standardize each column of $\mathbf{X}$ (i.e. each column should have zero mean and sample variance equal to 1, before calculating the condition number).

# Example: Car Seat Position data

```
# Standardize matrix
x = model.matrix(g)[,-1]
x = x - matrix(apply(x,2, mean), 38,8, byrow=TRUE)
x = x / matrix(apply(x, 2, sd), 38,8, byrow=TRUE)
apply(x,2,mean)
```

```
##           Age        Weight       HtShoes            Ht        Seated
## -2.193512e-17  2.810252e-16  9.566280e-16  1.941574e-16 -1.073010e-15
##           Arm         Thigh           Leg
## -1.070022e-16  8.909895e-17 -9.114182e-17
```

```
apply(x,2,var)
```

```
##     Age  Weight HtShoes      Ht  Seated     Arm   Thigh     Leg
##       1       1       1       1       1       1       1       1
```

```
e = eigen(t(x) %*% x)
sqrt(e$val[1]/e$val)
```

```
## [1]  1.000000  2.141737  3.497636  4.852243  5.404643  6.384606 10.615424
## [8] 59.766197
```

# Symptoms and Remedies of Collinearity

- Possible symptoms of collinearity:
  1. high pair-wise (sample) correlation between predictors
  2. high VIF
  3. high condition number
  4. $R^2$ is relatively large but none of the predictor is significant.
- What to do with collinearity?

  Remove some predictors from highly correlated groups of predictors.

  Another method we study later: regularize the model using penalized Least Squares estimation