Chapter 8: Additional Topics of Multiple Regression Model

CZ

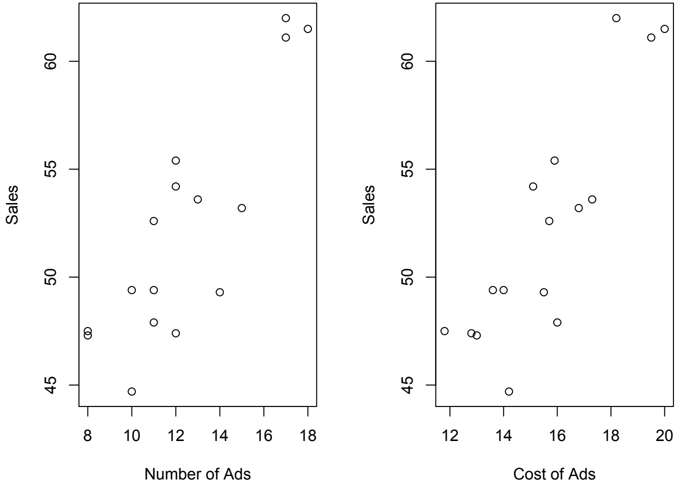
Fall, 2014

Multicollinearity: An Example

An example: Pizza Sales Data. A manager of a pizza outlet has collected monthly sales data over a 16-month period.

$$
\begin{aligned}
y &= \text{Sales (in thousands of dollars)} \\
x_1 &= \text{Number of advertisements} \\
x_2 &= \text{Cost of advertisements (in hundreds of dollars)}
\end{aligned}
$$

An Example


Sales
60
55
50
45
8
10
12
14
16
18
Number of Ads
Sales
60
55
50
45
12
14
16
18
20
Cost of Ads

An Example

- Fit the following model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

- The output:

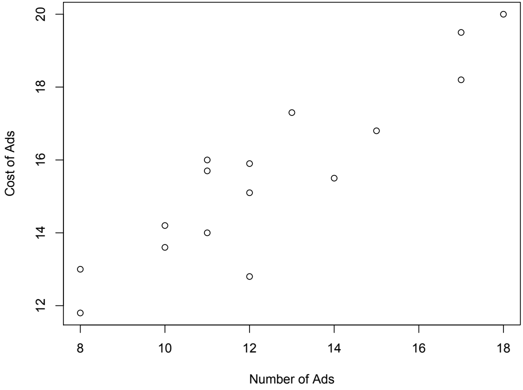
```
lm(formula = y ~ x1 + x2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.8231     5.6611   4.385 0.000738 ***
x1            0.6626     0.5386   1.230 0.240393
x2            1.2329     0.6962   1.771 0.100011
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.757 on 13 degrees of freedom
Multiple R-squared: 0.7789, Adjusted R-squared: 0.7449
F-statistic:  22.9 on 2 and 13 DF,  p-value: 5.492e-05
```

An Example

- $R^2 = 0.7789$ indicates x_1 and x_2 together explain a large part of the variability in sales.
- p-value for F-statistic indicate there is a strong evidence to reject $H_0 : \beta_1 = \beta_2 = 0$. At least, one of x_1 and x_2 is important.
- But we cannot reject $H_0 : \beta_1 = 0$ when x_2 is in the model. Similarly, we can not reject $H_0 : \beta_2 = 0$ when x_1 is in the model.
- If we consider just one x variable,
    - regress on x_1 alone, $R^2 = 0.7256$
    - regress on x_2 alone, $R^2 = 0.7532$

An Example

This is because variables x_1 and x_2 are highly correlated. The two variables express the same information. No point to include both.



Linear Dependency / Multicollinearity: Definition

- The columns of design matrix (the predictors) $\mathbf{1}, \mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_p}$ are linearly dependent, or have perfect multicollinearity if one column can be expressed as a linear combination of the other columns. That is, if there exist constants $c_i, \ i = 0, \ldots, p$, not all 0, such that $c_0\mathbf{1} + c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + ... + c_p\mathbf{X}_p = 0$.
- If there exist constants $c_i, \ i = 0, \ldots, p$, not all 0, such that $c_0\mathbf{1} + c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + ... + c_p\mathbf{X}_p \approx 0$. but may not be exactly linearly dependent, we call this situation **multicollinearity**.

Linear Dependency / Multicollinearity: Consequences

- If there is **perfect** multicollinearity among the columns of $\mathbf{X}$, then $|\mathbf{X}^\mathbf{T}\mathbf{X}| = \mathbf{0}$ and the inverse $(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}$ does not exist, thus $\hat{\beta} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}$ does not exists.
- If there is **multicollinearity**, $|\mathbf{X}^\mathbf{T}\mathbf{X}| \approx \mathbf{0}$ and the diagnoal elements of $(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}$ are large. Consequently, the variances of the estimated regression coefficients, $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ are large.

Linear Dependency: Example

Statistically, if one column can be written as (nearly) a linear combination of other columns, then it's redundant in the model. Consider the predictors $x_1 = (1, 2, 6, 10)$, $x_2 = (3, 4, 6, 8)$, and $x_3 = (5, 8, 18, 28)$. We could write the model as this:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Notice, $x_3 = 2 \cdot x_1 + x_2$, so that an equivalent model expression is

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (2 \cdot x_1 + x_2) + \epsilon \\ &= \beta_0 + (\beta_1 + 2\beta_3) x_1 + (\beta_2 + \beta_3) x_2 + \epsilon \end{aligned}$$

The point is **we don't need all three predictors in the model.**

Multicollinearity

In summary, if multicollinearity exists in the data:

- The variance of $\widehat{\beta}$ is large.
- Important predictors become insignificant in the model.
- It is hard to distinguish the effects, and the interpretation of the coefficients are problematic.

Detection of Multicollinearity

First check: Pairwise sample correlation coefficient

$$r_{lm} = \frac{\sum_{i=1}^{n}(x_{il} - \bar{x}_l)(x_{im} - \bar{x}_m)}{\sqrt{\sum_{i=1}^{n}(x_{il} - \bar{x}_l)^2 \sum_{i=1}^{n}(x_{im} - \bar{x}_m)^2}}$$

If $|r_{lm}| \approx 1$, x_l and x_m are strongly linearly related. No need to have both in the model. In the Pizza example, $r_{12} = 0.9015$.

A formal check: variance inflation factors (VIF)

- x_k is regressed on the remaining $p-1$ x's:

$$x_{ik} = \beta_0^* + \beta_1^* x_{i1} + \cdots + \beta_{k-1}^* x_{i,k-1} + \beta_{k+1}^* x_{i,k+1} + \cdots + \beta_p^* x_{ip} + \epsilon_i$$

- The resulting

$$R_k^2 = \frac{SSR}{SST}$$

  is a measure of how strongly x_k is linearly related to the rest of the x's.

$$VIF_k = \frac{1}{1 - R_k^2}$$

- If $VIF_k > 10$, strong evidence of multicollinearity.
- If $VIF_k > 5$, some evidence of multicollinearity.
- Actually, $Var(\hat{\beta}_k) = \sigma^2 \frac{VIF_k}{(n-1)\widehat{var}(x_k)}$ where $\widehat{var}(x_k)$ is the sample variance of x_k in the dataset.

Variance Inflation Factor

In the pizza example,

```
> library(car)
> vif(pizza)
      x1       x2
5.339243 5.339243
```

Linear Dependency and Sequential Sum of Squares

$SSR(x_i|x_j)$ can be quite low when x_i and x_j are highly correlated, even when x_i individually is a good predictor.

```
> anova(lm(y~x2))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value    Pr(>F)
x2         1 336.64  336.64  42.718 1.321e-05 ***
Residuals 14 110.33    7.88

> anova(lm(y~x1+x2))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 324.30  324.30  42.662 1.907e-05 ***
x2         1  23.84   23.84   3.136       0.1
Residuals 13  98.82    7.60
```

Linear Independency

At the other extreme is when correlations between each pair of predictors is 0. When this happens, a predictor's contribution SSR is fixed and doesn't depend on other predictors that are already in the model.

Linear Independency: Example

$$
\begin{aligned}
y &= \text{Shrinkage of parts produced by a molding operation} \\
x_1 &= \text{mold temperature} \\
x_2 &= \text{hold pressure} \\
x_3 &= \text{screw speed}
\end{aligned}
$$

It was decided to study the predictors at two levels each with coding -1 (low) and 1 (high). Eight experiments are taken under different combinations of the predictors.

Example

Shrinkage Data:

Run	x_1	x_2	x_3	y
1	-1	-1	-1	19.7
2	+1	-1	-1	19.1
3	-1	+1	-1	20
4	+1	+1	-1	19.5
5	-1	-1	+1	15.9
6	+1	-1	+1	15.3
7	-1	+1	+1	25.5
8	+1	+1	+1	24.9

Example

- Fit the following regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon$$

- The design matrix is:

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Example

- The pairwise sample correlations among x_1, x_2 and x_3 are all zero.
- The columns of $\mathbf{X}$ are orthogonal.
- $\hat{\beta} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathbf{T}\mathbf{Y} = (19.99, -0.29, 2.49, 0.41)^T$.
- $Var(\hat{\beta}) = \sigma^2(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}$ and

$$\sigma^2(\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}} = \sigma^2 \begin{pmatrix} 8 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 8 \end{pmatrix}^{-1}$$

Linear Independency

```
> anova(lm(y~x1+x2+x3))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value  Pr(>F)
x1         1  0.661   0.661  0.0618 0.81589
x2         1 49.501  49.501  4.6279 0.09784 .
x3         1  1.361   1.361  0.1273 0.73931
Residuals  4 42.785  10.696
> anova(lm(y~x1))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value Pr(>F)
x1         1  0.661  0.6612  0.0424 0.8437
Residuals  6 93.647 15.6079
> anova(lm(y~x2))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value  Pr(>F)
x2         1 49.501  49.501  6.6285 0.04207 *
Residuals  6 44.807   7.468
> anova(lm(y~x3))
Analysis of Variance Table
          Df Sum Sq Mean Sq F value Pr(>F)
x3         1  1.361  1.3613  0.0879 0.7769
Residuals  6 92.947 15.4912
```

Linear Independency and Sequential Sum of Squares

- Also we can check that in this case we have $SSR(x_2|x_1) = SSR(x_2)$ and $SSR(x_3|x_1, x_2) = SSR(x_3)$.
- $SSR(x_1, x_2, x_3) = SSR(x_1) + SSR(x_2) + SSR(x_3)$. That is, there's no overlap in the response variability explained by each of the three predictors.

Regression Model and Assumptions

- In the matrix form, the regression model is:

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

- Previously, we assume:

$$\boldsymbol{\epsilon} \sim MVN(0, \sigma^2 \mathbf{I})$$

- A more general assumption:

$$\boldsymbol{\epsilon} \sim MVN(0, \sigma^2 \mathbf{V})$$

  $\mathbf{V}$ is not necessarily an identity matrix.

Heteroscedasticity

- For example, in a multiple regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i.$$

  We assume: $\epsilon_i \sim N(0, \sigma_i^2)$. We call it "heteroscedasticity".

- To obtain "good" estimates of the regression coefficients, the objective function is redefined as the weighted error sum of squares:

$$\sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_p)^2$$

  where $w_i = 1/\sigma_i^2$.

Weighted Least Squares

- Assume $\mathbf{Y} \sim MVN(\mathbf{X}\beta, \sigma^2\mathbf{V})$. The weighted least squares (WLS) estimator is obtained by minimizing:

$$(\mathbf{Y}-\mathbf{X}\beta)^{\mathbf{T}}\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\beta)$$

- We obtain

$$\widehat{\beta}^{WLS} = (\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{Y}.$$

- Properties of $\widehat{\beta}^{WLS}$:
  - $E(\widehat{\beta}^{WLS}) = \beta$
  - $Var(\widehat{\beta}^{WLS}) = \sigma^2(\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
  - $\widehat{\beta}^{WLS} \sim N(\beta, \sigma^2(\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1})$

An Equivalent Representation

- Assume $\mathbf{Y} \sim MVN(\mathbf{X}\beta, \sigma^2\mathbf{V})$. Let $\mathbf{Y}^* = \mathbf{V}^{-\frac{1}{2}}\mathbf{Y}$ and $\mathbf{X}^* = \mathbf{V}^{-\frac{1}{2}}\mathbf{X}$, we have:

$$\mathbf{Y}^* \sim MVN(\boldsymbol{X}^*\beta, \sigma^2\mathbf{I})$$

- Therefore, WLS is equivalent to the ordinary least squares (OLS) applied to transformed data $(\mathbf{X}^*, \mathbf{Y}^*)$

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{WLS} &= (\mathbf{X}^{*\mathbf{T}}\mathbf{X}^*)^{-1}\mathbf{X}^{*\mathbf{T}}\mathbf{Y}^* \\ &= (\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{Y} \end{aligned}$$

An Unbiased Estimator of σ^2

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{(\mathbf{Y}^* - \mathbf{X}^*\hat{\beta}^{WLS})^T(\mathbf{Y}^* - \mathbf{X}^*\hat{\beta}^{WLS})}{n-p-1} \\
&= \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta}^{WLS})^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta}^{WLS})}{n-p-1}
\end{aligned}
$$

Compare with Ordinary Least Squares

If we ignore heteroscedasticity and perform an ordinary least squares (OLS), we get

$$\widehat{\boldsymbol{\beta}}^{OLS} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}.$$

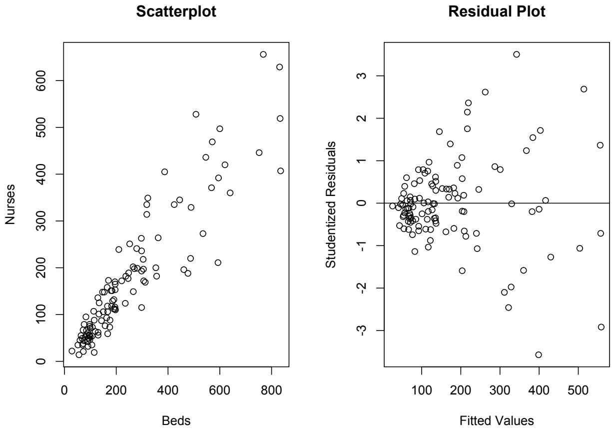
- $E(\widehat{\boldsymbol{\beta}}^{OLS}) = \boldsymbol{\beta}$
- $Var(\widehat{\boldsymbol{\beta}}^{OLS}) = \sigma^2(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{V}\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}$
- Note: one can show, $Var(\widehat{\boldsymbol{\beta}}^{OLS})$ exceeds $Var(\widehat{\boldsymbol{\beta}}^{WLS})$ by a positive semidefinite matrix, i.e., $Var(\widehat{\boldsymbol{\beta}}^{WLS})$ is usually smaller than $Var(\widehat{\boldsymbol{\beta}}^{OLS})$.

Example

The dataset “Senic”. The response variable is the number of nurses and the predictor is the number of beds in 113 different hospitals in US.

Example

We first fit a simple linear regression of Nurses vs. Beds.

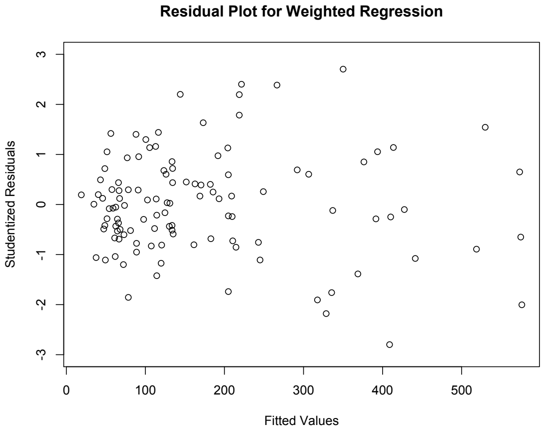


Example

- Due to heteroscedasticity, we assume
$$Var(\epsilon_i) = \sigma^2 Beds_i$$
for $i = 1, \ldots, n$.
- We perform a weighted least squares in R:

```
> mymodel=lm(Nurses~Beds,weights=1/Beds,data=senic)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.06824    5.52424  -0.193    0.847
Beds         0.69127    0.02836  24.379   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.039 on 111 degrees of freedom
Multiple R-squared: 0.8426, Adjusted R-squared: 0.8412
F-statistic: 594.3 on 1 and 111 DF,  p-value: < 2.2e-16
```

Example



Residual Plot for Weighted Regression
Studentized Residuals
Fitted Values
3
2
1
0
-1
-2
-3
0
100
200
300
400
500