MLR: Collinearity and Model Selection

Keegan Korthauer

Department of Statistics

UW Madison

Basic MLR Model

- Dependent continuous variable **y**
- *p* independent continuous variables $\mathbf{x_1, x_2,\ldots, x_p}$
- *n* observations: ordered pairs $\mathbf{(y_i, x_{1i}, x_{2i},\ldots, x_{pi})}$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

- Predicted y_i for a set of $x_{1i}, x_{2i},\ldots, x_{pi}$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi}$$

Basic Question

How do we decide which variables to include in a multiple linear regression model?

Need to consider two main factors:

1. Relationship among predictors
2. Combined effect of predictors on response

CONFOUNDING AND COLLINEARITY

SLR vs MLR

- Fitting separate SLR models to each predictor variable is **not** the same as fitting a MLR model
- MLR models take into account how the predictors are related to one another
- As a result, the coefficient estimates for a predictor variable will almost always be different when used alone in a SLR model versus with other predictors in a MLR model

Example – Patient Satisfaction Survey

A hospital administrator wished to study the relationship between the following variables on a random sample of 46 patients:

- Y: patient satisfaction (percent)
- X_1: patient's age
- X_2: severity of illness (an index)
- X_3: patient's anxiety level (an index)

Let's look at **SLR models for each predictor separately**

Satisfaction = $\beta_0 + \beta_1$*Age + ε

```
> fit.Age <- lm(Satis~Age)
> summary(fit.Age)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 119.9432     7.0848  16.930  < 2e-16 ***
Age          -1.5206     0.1799  -8.455 9.06e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.76 on 44 degrees of freedom
Multiple R-squared: 0.619,    Adjusted R-squared: 0.6103
F-statistic: 71.48 on 1 and 44 DF,  p-value: 9.058e-11
```

Satisfaction = $\beta_0 + \beta_1$*Severity + ε

```
> fit.Sev <- lm(Satis~Sev)
> summary(fit.Sev)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 183.0770    24.3249   7.526 1.95e-09 ***
Sev          -2.4093     0.4806  -5.013 9.23e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.91 on 44 degrees of freedom
Multiple R-squared: 0.3635,   Adjusted R-squared: 0.3491
F-statistic: 25.13 on 1 and 44 DF,  p-value: 9.23e-06
```

Satisfaction = $\beta_0 + \beta_1$*Anxiety + ε

```
> fit.Anx <- lm(Satis~Anx)
> summary(fit.Anx)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  146.449     15.304   9.569 2.55e-12 ***
Anx          -37.117      6.637  -5.593 1.33e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.33 on 44 degrees of freedom
Multiple R-squared: 0.4155,   Adjusted R-squared: 0.4022
F-statistic: 31.28 on 1 and 44 DF,  p-value: 1.335e-06
```

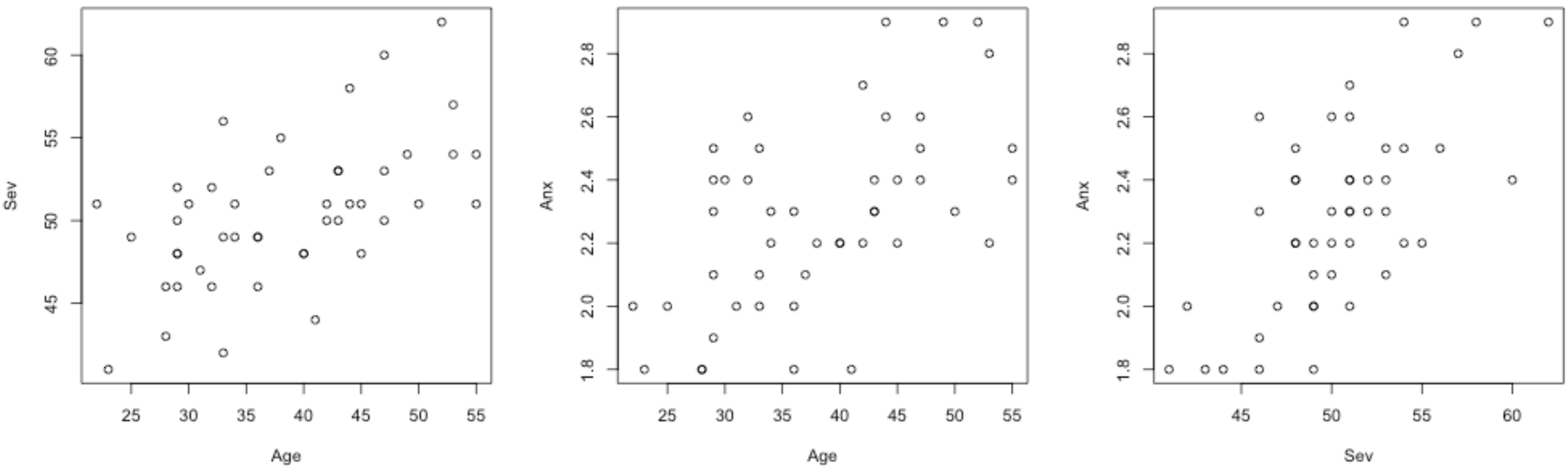
Conclusions?

- From the three SLR models, we see:
  - Age, severity and anxiety all have coefficients that are significantly different from zero
  - We might be tempted to conclude that increasing **either** age, severity, **or** anxiety will lead to decreased patient satisfaction

- First, be wary of saying that any of these three might be **causes** of decreased satisfaction

- Second, need to consider the possibility of confounding among the three predictors
  - What if patient satisfaction is really only dependent on one or two of these factors??

Relationships Among Predictors



Case in point: it appears that higher anxiety levels are observed in more severe cases.

- If the satisfaction truly only depends on anxiety levels, the significant severity coefficient in the SLR model is due to spurious associations
- In that case, we shouldn't be predicting satisfaction from severity

How do we know which predictor(s) to use?

The MLR Model Provides a Clue

```
> fit1 <- lm(Satis~Age+Sev+Anx)
> summary(fit1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
Age          -1.1416     0.2148  -5.315 3.81e-06 ***
Sev          -0.4420     0.4920  -0.898   0.3741
Anx         -13.4702     7.0997  -1.897   0.0647 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.06 on 42 degrees of freedom
Multiple R-squared: 0.6822,   Adjusted R-squared: 0.6595
F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

Collinearity

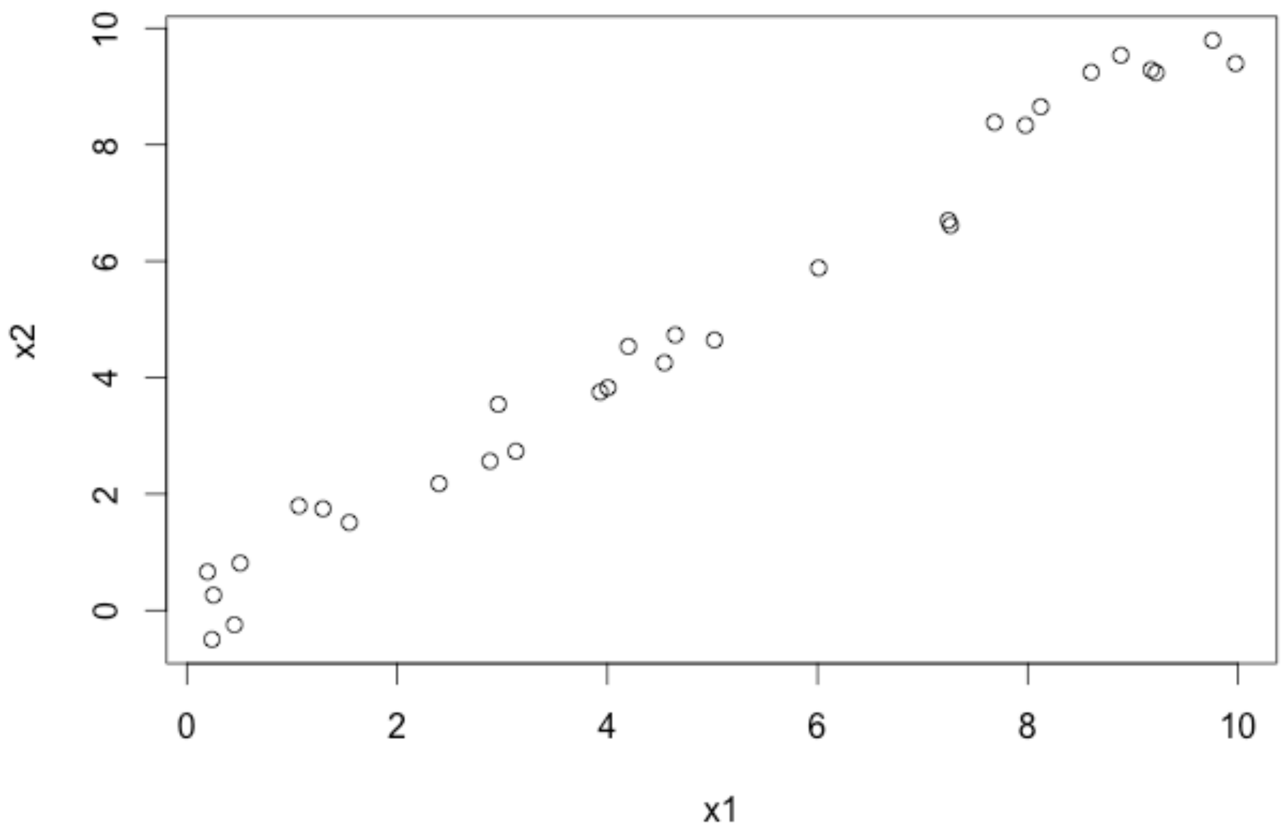
- **Caution:** when two predictor variables are very highly correlated with one another, MLR may not help determine which is the more important one

- Consider the following scenario:
  - We fit an SLR model with one predictor X_1
  - Let X_2 be another predictor that has correlation 0.99 with X_1
  - Would we want to fit a MLR that includes both predictors?

No: X_2 contains almost the same information as X_1 so it doesn't help us predict Y if we are already using X_1

Example of Collinear Predictors

```
> cor(x1, y)
[1] 0.9673696
> cor(x2, y)
[1] 0.9660594
> cor(x1,x2)
[1] 0.9911231
```



Separate SLRs

```
> summary(lm(y~x1))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.46657    0.55433   0.842    0.407
x1           1.93776    0.09591  20.203   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.712 on 28 degrees of freedom
Multiple R-squared: 0.9358, Adjusted R-squared: 0.9335
F-statistic: 408.2 on 1 and 28 DF,  p-value: < 2.2e-16

> summary(lm(y~x2))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.63781    0.55855   1.142    0.263
x2           1.89322    0.09567  19.789   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.745 on 28 degrees of freedom
Multiple R-squared: 0.9333, Adjusted R-squared: 0.9309
F-statistic: 391.6 on 1 and 28 DF,  p-value: < 2.2e-16
```

MLR with Both Predictors

```
> summary(lm(y~x1+x2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4993     0.5519   0.905    0.374
x1            1.1204     0.7173   1.562    0.130
x2            0.8069     0.7018   1.150    0.260

Residual standard error: 1.702 on 27 degrees of freedom
Multiple R-squared: 0.9388,  Adjusted R-squared: 0.9343
F-statistic: 207.1 on 2 and 27 DF,  p-value: < 2.2e-16
```

The relationship between X_1 and X_2 is so strong that it is impossible to determine which one is a better predictor of Y

Dealing with Collinearity

- **Diagnosis**: highly correlated predictors
  - rule of thumb $r > 0.80$

- **Remedies:**
  - collect more data points, if possible, where the two predictors are not correlated and re-fit the MLR model
  - remove one of the offenders using existing prior knowledge of the relationships

BASIC PRINCIPLES OF MODEL SELECTION

Basic Principles of Model Selection

- In many situations, a large number of variables might be related to the response – How do we decide which to include in the model?

- **Principle of Parsimony (Occam's Razor):** A model should contain the smallest number of variables necessary to fit the data

- Exceptions: unless physical theory dictates otherwise,
  1. Linear models should always contain an intercept
  2. If a power term x^n is included, also include all lower powers $x, x^2, \ldots, x^{n-1}$
  3. If an interaction term $x_i x_j$ is included, also include the two main effects for x_i and x_j

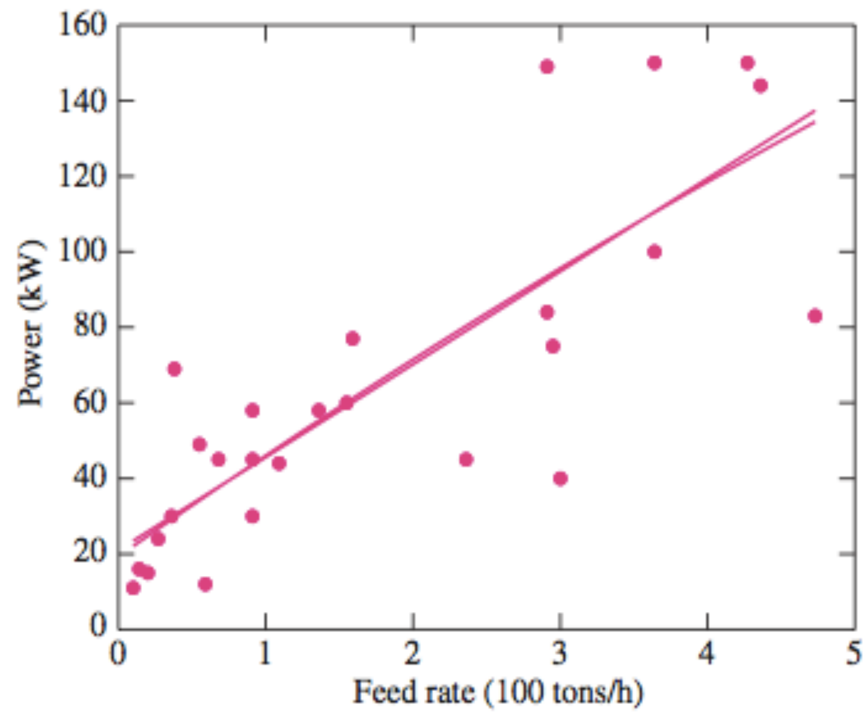
Example – Feed Rate of Industrial Jaw Crushers

Simple Linear Model:

```
The regression equation is
Power = 21.0 + 24.6 FeedRate

Predictor        Coef     SE Coef        T        P
Constant       21.028       8.038     2.62    0.015
FeedRate       24.595       3.338     7.37    0.000

S = 26.20        R-Sq = 68.5%     R-Sq(adj) = 67.2%
```

Should we add a quadratic term?

```
The regression equation is
Power = 19.3 + 27.5 FeedRate -0.64 FeedRate^2

Predictor        Coef     SE Coef        T        P
Constant        19.34       11.56     1.67    0.107
FeedRate        27.47       14.31     1.92    0.067
FeedRate^2    -0.6387       3.090    -0.21    0.838

S = 26.72        R-Sq = 68.5%     R-Sq(adj) = 65.9%
```



HT for Comparing Two Models

- Given some MLR model, how do we determine whether we can **drop** some variables?

- Full Model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \boxed{\beta_{k+1} x_{(k+1)i} + \ldots + \beta_p x_{pi}} + \varepsilon_i$$

- Potential Reduced Model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

- Perform Hypothesis test of the null:

$$H_0 : \beta_{k+1} = \beta_{k+2} = \ldots = \beta_p = 0$$

Test Statistic

- Notation:
  - SSE_{full}: sum of squared error for full model (df = n-p-1)
  - $SSE_{reduced}$: sum of squared error for reduced model (df =n-k-1)

- F Test statistic:

$$F = \frac{(SSE_{reduced} - SSE_{full})/(p-k)}{SSE_{full}/(n-p-1)} \sim F_{p-k,\, n-p-1}$$

- The idea: if the reduced model is as good as the full, the F statistic should be near 1
  - If H_0 is false, $SSE_{reduced}$ tends to be larger, leading to more extreme test statistic F

Testing Reduced vs Full Model

- F test on previous slide assumes (1) that full model is correct and (2) that the dropped variables are picked independently of the data
  - Strong assumptions; rarely true in practice
  - No way to check assumptions

- Used informally in practice (not a rigorous statistical method) to find a parsimonious model

- Process is illustrated in the Example on pages 623-626

Adjusted R^2

- We use R^2 to measure the goodness-of-fit of a model
- Issue when using it for MLR: R^2 increases as you add more variables to the model
  - Using R^2 as a criterion for model selection would always lead to the model that contains more predictors!
- Adjusted R^2:

$$\text{Adjusted } R^2 = R^2 - \left(\frac{k}{n-k-1}\right)(1-R^2)$$

where *k* is the number of variables in the model

- Strikes a balance between adding more predictors and increasing goodness-of-fit

Adjusted R^2

```
> fit1 <- lm(Satis~Age+Sev+Anx)
> summary(fit1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
Age          -1.1416     0.2148  -5.315 3.81e-06 ***
Sev          -0.4420     0.4920  -0.898   0.3741    
Anx         -13.4702     7.0997  -1.897   0.0647 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.06 on 42 degrees of freedom
Multiple R-squared: 0.6822,  Adjusted R-squared: 0.6595
F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

Best Subset Selection

- Idea: find the ‘best’ subset of predictors, where ‘best’ means that the model is optimal in terms of some statistic
- One possible (easy) choice – **maximize the adjusted R^2**
- For all possible combinations of predictors, calculate adjusted R^2 and choose the model with the largest value

Example – Patient Satisfaction

Here are the adjusted R^2 for all possible models (excluding interactions):

- Age: 0.6103
- Sev: 0.3491
- Anx: 0.4022
- Age, Sev: 0.6389
- Age, Anx: 0.661
- Anx, Sev: 0.4437
- Age, Anx, Sev: 0.6595

Stepwise Model Selection

- User chooses two threshold p-values: α_{in} and α_{out} (with $\alpha_{in} \leq \alpha_{out}$)
- Start with the null model (no covariates)
- In each step, do a **forward selection** and **backward elimination**
  - **Forward Selection:** add in the variable with the smallest p-value $< \alpha_{in}$
  - **Backward Elimination:** remove the variable with the largest p-value $> \alpha_{out}$
- Proceed until no more variables can be added or dropped
- Don't have to evaluate every single combination of covariates

Example – Patient Satisfaction

Let α_{in} = 0.15 and α_{out} = 0.15

1. Add in Age (smallest p-value in separate SLR models and < α_{in})

2. Starting from model with Age, check MLR models: Age + Sev and Age + Anx. Add in Anx since it has a smaller p-value than Sev when added to the model with Age (and < α_{in})

3. Check to see if Sev can be added to the model that includes Age + Anx. The p-value for Sev is 0.3741 > α_{in} so it cannot be added. Both Age and Anx have p-values < α_{out} so selection is complete

Final model: Satis ~ Age + Anx

Downsides of Automatic Procedures

- Will ignore the rules to include lower-order terms of polynomials or interactions

- Could be little practical difference between some models

- Statistics calculated from data are random, so the result of which model is best is also random

- Will always find a model, whether it should or not

Next

- Factorial Experiments
  - One-way Analysis of Variance (ANOVA)
  - Pairwise comparisons in ANOVA
  - Two-way ANOVA

- HW 11 due Friday, HW 12 due last day of class

- Final Exam (cumulative) Sunday, May 11 from 2:45-4:45 in B130 Van Vleck