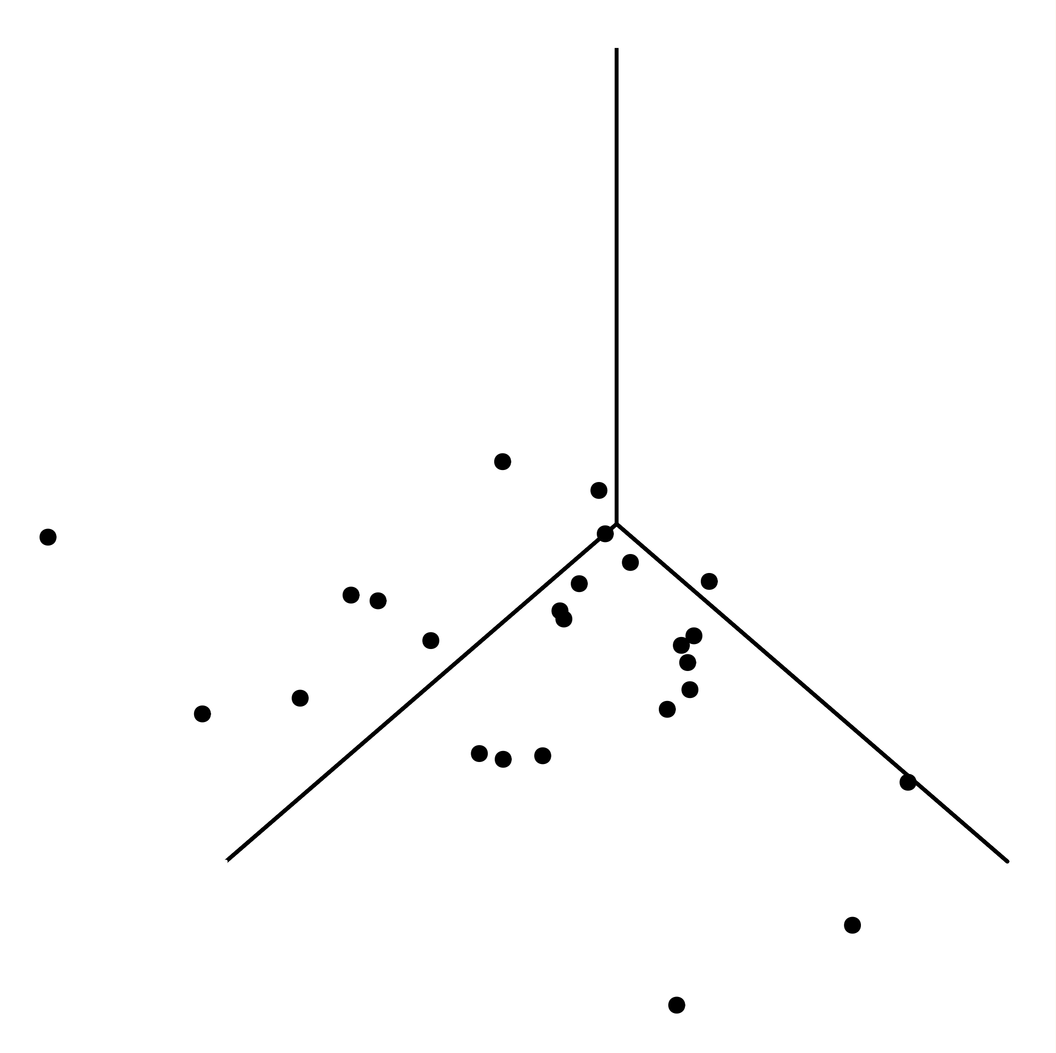
MLR Collinearity and Variable Selection

Stat 230

May 04 2022

Overview



Today:

- Issues with correlated predictors
- Collinearity
- Variance Inflation Factor
- Model selection strategies

Correlated predictors

It can be hard to "see" the MLR effect of a predictor in basic EDA

- e.g. effect of `advance` in the supervisor MLR (partial residual plot slides)

Interpretation of MLR effects can be challenging

- e.g. can we really "hold `learn` rating constant" when interpreting the effect of `advance` if they are correlated?

Parameter estimates have larger SE's than if predictors weren't correlated

- can make it harder to find "statistically significant" predictors

Collinearity

Two predictors x_1 and x_2 are (exactly) collinear if for all cases $i = 1, \ldots, n$:

$$c_1 x_{1,i} + c_2 x_{2,i} = c \quad \Rightarrow \quad x_{1,i} = \left(c - c_2 x_{2,i}\right) / c_1$$

where c_1, c_2, c are all known constants.

Example: x_1 = height in inches and x_2 = height in feet

- $x_{1,i} = 12 x_{2,i}$ so we could have $c = 0, c_2 = -12, c_1 = 1$

Collinearity

Can extend collinearity to more than two terms:

$$c_1 x_{1,i} + c_2 x_{2,i} + \cdots + c_p x_{p,i} = c$$

Example: $x_{\%\text{ frosh}} + x_{\%\text{ soph}} + x_{\% junior} + x_{\%\text{ senior}} = 100\%$

- Terms that are approximately collinear will have high correlation and will show a linear relationship in the scatterplot matrix.

Variance Inflation Factor (VIF)

VIF for predictor x_i is equal to

$$VIF_i = \frac{1}{1 - R_i^2}$$

R_i^2 is the R-squared value for the regression of x_i on all other model predictors

$R_i^2 \approx 0$ means $VIF_i \approx 1$

- Little collinearity between x_i and other terms

$R_i^2 \approx 0.5$ means $VIF_i \approx 2$

- moderate collinearity between x_i and other terms

$R_i^2 \approx 0.9$ means $VIF_i \approx 10$

- lots of collinearity between x_i and other terms

Why "Variance inflation" factor?

The variance (squared SE) of $\hat{\beta}_i$ is equal to

$$SE\left(\hat{\beta}_i\right)^2 = \frac{\hat{\sigma}^2}{(n-1)s_i^2}\frac{1}{1-R_i^2} = \frac{\hat{\sigma}^2}{(n-1)s_i^2}VIF_i$$

Implications:

No collinearity in x_i

$$\Rightarrow VIF_i = 1$$

$$\Rightarrow SE\left(\hat{\beta}_i\right) = \frac{\hat{\sigma}}{\sqrt{(n-1)s_i^2}}$$

More collinearity in x_i

$$\Rightarrow VIF_i > 1$$

$$\Rightarrow SE\left(\hat{\beta}_i\right) > \frac{\hat{\sigma}}{\sqrt{(n-1)s_i^2}}$$

Why "Variance inflation" factor?

$$SE\left(\hat{\beta}_i\right)^2 = \frac{\hat{\sigma}^2}{(n-1)s_i^2}\frac{1}{1-R_i^2} = \frac{\hat{\sigma}^2}{(n-1)s_i^2}VIF_i$$

- higher $SE\left(\hat{\beta}_i\right)$
- more uncertainty when estimating β_i
- larger confidence intervals
- smaller t-test stats and larger p-values

VIF in R using `car` package

- `vif(my_lm)`

Generalized VIF.

- If `my_lm` used a categorical predictor with more than 2 levels, R will report Generalized VIF.
- It's exactly the same thing as VIF for single terms but is a sort of cumulative VIF across all levels of the factor variable (across all indicator variables for the factor).

VIF is influenced by outliers.

- Make sure to check for outliers before putting too much "weight" on VIF values.

Supervisor satisfaction

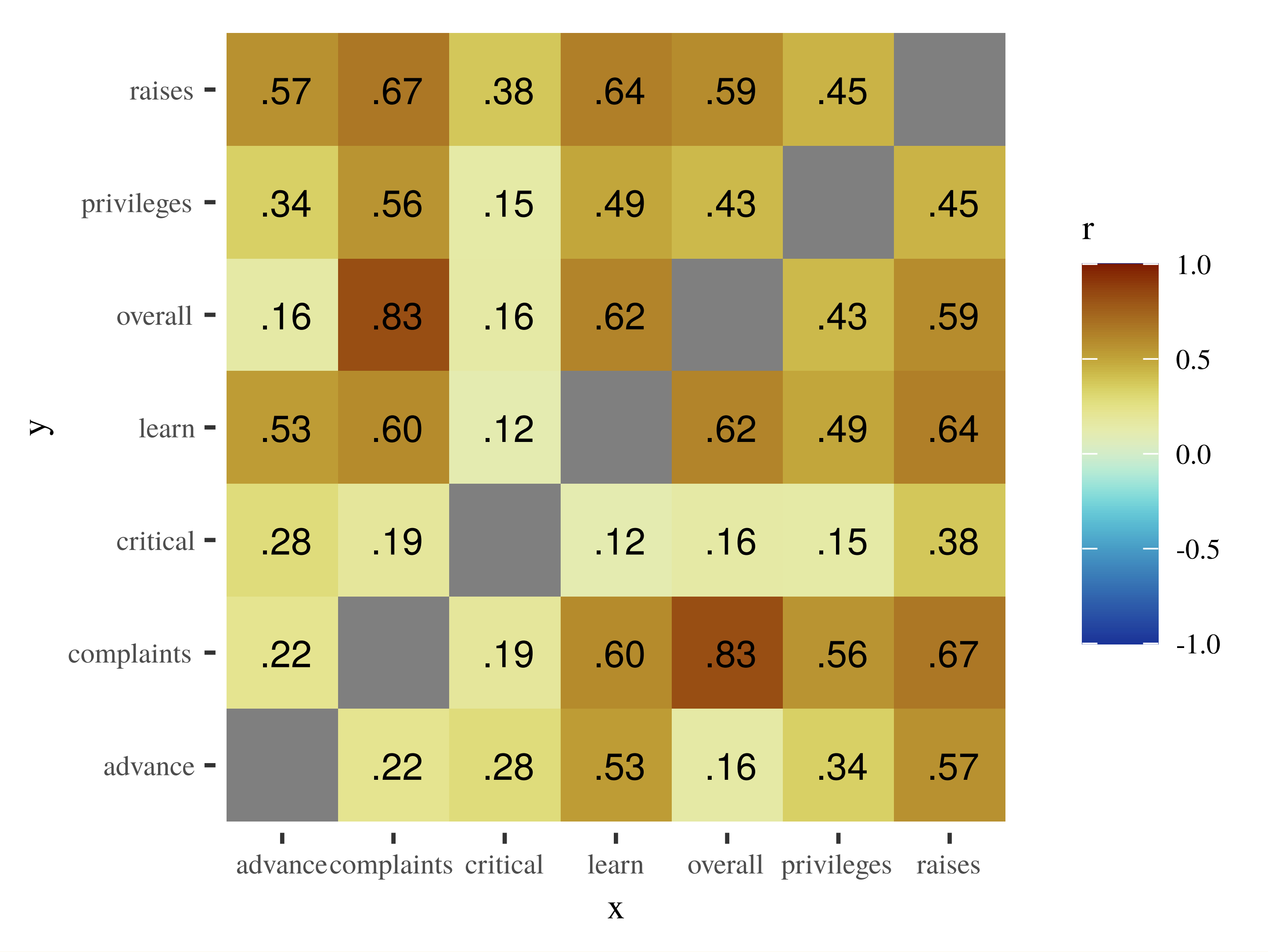
Question: What supervisor characteristics are most important to employees in a large company who were asked to rate their immediate supervisor

- **Response:** overall rating on a scale of 0 (bad) to 100 (good)
- **Predictors** from survey questions measured on an agreement scale ($0 =$ completely disagree to $100 =$ completely agree)

Predictors	Description
raises	"Your supervisor bases raises on performance."
learn	"Your supervisor provides opportunities to learn new things."
advance	"I am not satisfied with the rate I am advancing in the company."
complaints	"Your supervisor handles employee complaints appropriately."
privileges	"Your supervisor allows special privileges."
critical	"Your supervisor is too critical of poor performance."

Correlation heatmaps

```
# nice correlation heatmaps
library(corrr)
library(scico)
library(paletteer)
supervisor %>%  # only numerical variables
  correlate() %>%
  stretch() %>%
  ggplot(aes(x, y, fill = r)) +
  geom_tile() +
  theme(text = element_text(size = 8))+
  geom_text(aes(label = as.character(fashion(r)
  scale_fill_paletteer_c("scico::roma",
                         limits = c(-1, 1),
                         direction = -1)
```



Example: supervisor satisfaction

All pairwise relationships show some degree of positive association

- moderate/strong correlations between `overall` and the predictors `complaints`, `learn`, and `raises`
- `raises` has moderate/strong correlation with all but one other predictor

Fit model

Let's model `overall` rating as a function of all predictors

```
library(moderndive)
supervisor_lm <- lm(overall ~ complaints + privileges + learn + raises
                    + critical + advance, data = supervisor)
get_regression_table(supervisor_lm)
```

```
# A tibble: 7 × 7
  term       estimate std_error statistic p_value lower_ci upper_ci
  <chr>         <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept    10.8      11.6       0.931   0.362  -13.2     34.8
2 complaints    0.613     0.161     3.81    0.001    0.28     0.946
3 privileges   -0.073     0.136    -0.538   0.596   -0.354    0.208
4 learn         0.32      0.169     1.90    0.07    -0.028    0.669
5 raises        0.082     0.221     0.369   0.715   -0.376    0.54
6 critical      0.038     0.147     0.261   0.796   -0.266    0.342
7 advance      -0.217     0.178    -1.22    0.236   -0.586    0.152
```

Only complaints is statistically significant at the 5% level after accounting for other terms

Example: supervisor satisfaction

```
library(car)
vif(supervisor_lm)
```

```
complaints privileges      learn     raises   critical    advance 
  2.667060   1.600891   2.271043   3.078226   1.228109   1.951591 
```

If goal is to find the variables that are most predictive of `overall` rating:

- We need to reduce the number of insignificant terms and redundant terms.

Let's test whether `privileges`, `raises` and `critical` are needed (high p-values, high VIF for raises)

Example: supervisor satisfaction

```
supervisor_lm_red <- lm(overall ~ complaints + learn + advance, data = supervisor)
anova(supervisor_lm_red, supervisor_lm)
```

```
Analysis of Variance Table

Model 1: overall ~ complaints + learn + advance
Model 2: overall ~ complaints + privileges + learn + raises + critical +
    advance
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     26 1179.1
2     23 1149.0  3    30.109 0.2009 0.8947
```

With `complaints`, `learn` and `advance` already in the model, `raises`, `privileges` and `critical` are statistically insignificant (F = 2.01, df=3,23, p = 0.895).

Example: supervisor satisfaction

```
get_regression_table(supervisor_lm_red)
# A tibble: 4 × 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept   13.6       7.54       1.8    0.084   -1.93    29.1
2 complaints   0.623     0.118      5.27   0        0.38     0.866
3 learn        0.312     0.154      2.03   0.053   -0.005    0.629
4 advance     -0.187     0.145     -1.29   0.208   -0.485    0.111
vif(supervisor_lm_red)
complaints      learn    advance
  1.582432   2.094593   1.420359
```

`complaints` is the most statistically significant term and has the largest coefficient effect estimate:

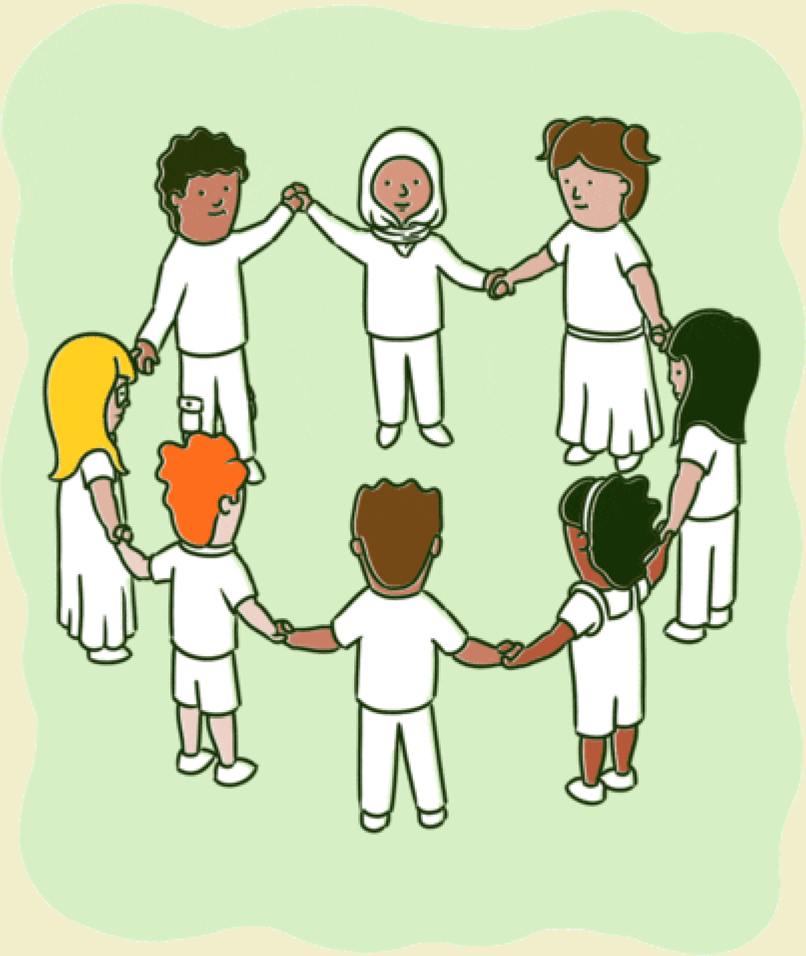
- a 10 point increase in how a supervisor handles `complaints` is associated with a 6.2 point increase in mean `overall` rating, holding other predictors fixed.

`learn` is somewhat statistically significant and it's effect is about half the size of `complaints`:

- a 10 point increase in how a supervisor provides opportunities to learn new things is associated with a 3.1 point increase in mean `overall` rating, holding other predictors fixed.

Your Turn 1

05:00



- Go over to the in class activity file
- Complete the activity in your group

SLR/MLR modeling strategy

1. Identify objectives of analysis
2. Understand your data: EDA, graphical/numerical summaries, clean up data, consider missing data issues
3. Consider if transformations are needed.
4. Fit "rich" model (with many terms), check assumptions, reconsider transformations, check outliers.
    - can be an iterative process (rethinking transformations, outliers)
5. Once a suitable large model is found, if needed, check for collinearity and use variable selection techniques to reduce the number of model terms.
6. Check assumptions/outliers, proceed with needed inference/interpretation (CI, PI, etc)

Consider your modeling objective

"The single most important tool in selecting a subset of variables is the analyst's knowledge of the area under study and of each of the variables."

Applied Linear Regression, Sandy Weisberg

Three common objectives (goals):

1. Adjusting for explanatory variables
2. Fishing expedition
3. Prediction (often, predictive analytics, supervised learning)

Consider your modeling objective

Adjusting (controlling) for the effects of a group of explanatory variables.

- (one approach) Find an appropriate model with only the controlling variables, then add the main explanatory variable of interest to the model.

HW Race and Wage Example: What, if any, is the effect of race on wages after accounting for region, SMSA, education and experience levels?

Fit a model with just region, SMSA, education and experience

- if race is collinear with other predictors: refine the model (find significant predictors) to reduce collinearity between race and other term
- if race is not collinear with other predictors: could refine or not refine

Add the variable of interest, race, to the model (and possible interactions) to determine its effect on wages.

Consider your modeling objective

Fishing for an explanation

- What to do a when you have a very large set of candidate predictors from which you wish to extract a few?
- No well-defined questions, may be start from scratch to add new predictors one step at a time, if feasible
- Often collinearity is an issue, interpretation of coefficients can be difficult
- If predictors are correlated, than can we really "hold all other terms constant" while changing the value of another term?

Consider your modeling objective

Prediction

- Usually don't care about interpreting model, just want a good model (low prediction error) with easy to measure explanatory variables.

Classification problem

- Are you going to default on your loan? Plug your employment, loan history, personal background in to a model and predict default (yes/no).
- We don't care about parameter interpretation!

Thoughts on Variable Selection

- Modeling goals 1 and 2 often involve determining the most "statistically significant" variables.
- A model with p terms will have 2^p possible models!.
- Usually, when faced with many variables, there is no one best model!
- But some models are better than others. You need to (correctly) justify the choice of your "best" model.

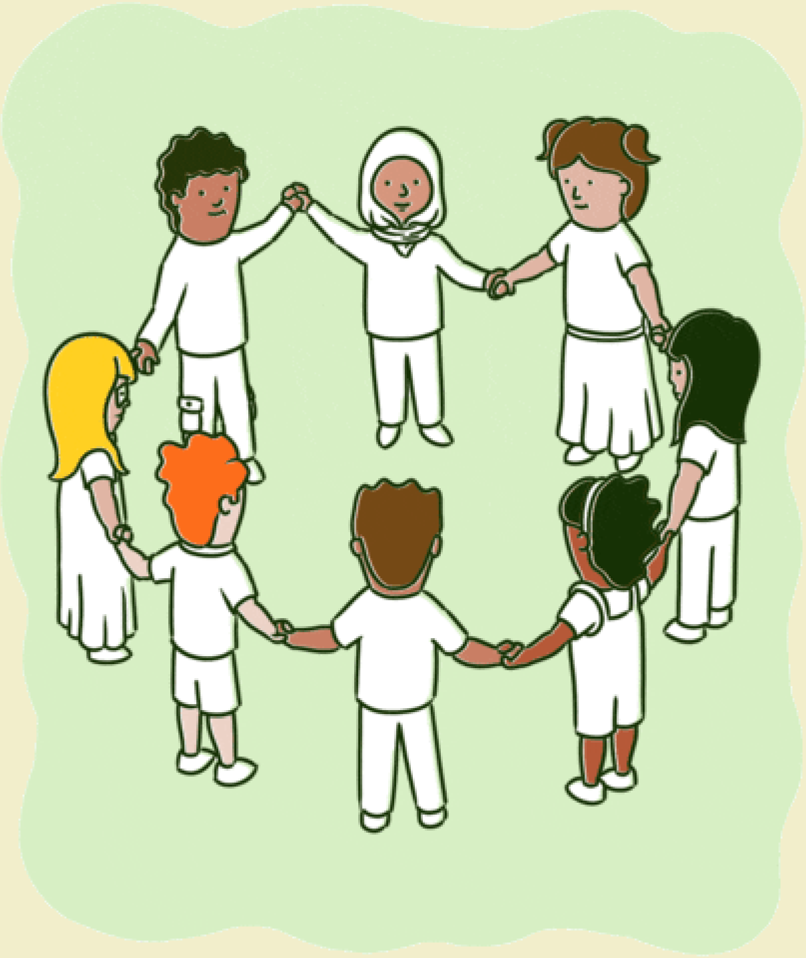
Thoughts on Variable Selection

Perhaps for an introductory class like STAT 230, **backwards selection** approach is easier:

- Start with a rich (large) model
- Take out one term at a time with t-tests and see if $R^2_{adjusted}$ increases
- OR take out multiple terms at a time with F-tests and verify with $R^2_{adjusted}$
- We will not deal with automatic model selection in this course (e.g. using criteria like AIC, BIC, Mallows C_p etc.)
- It's best to use our own judgment and intuition about our data

Your Turn 2

05:00



- Go over to the in class activity file
- Complete the activity in your group