STAT 224 Lecture 18
Chapter 9 Multicollinearity

Yibi Huang

Predictors of an MLR Model Cannot Be Linearly Dependent

MLR requires predictors to be **linearly independent**, i.e., no predictor can be expressed as a linear combination of others

- Ex: $X_1 = \#$ of undergrads, $X_2 = \#$ of grads, $X_3 = \#$ of students, $\Rightarrow X_1, X_2, X_3$ are linearly dependent since $X_1 + X_2 = X_3$

No unique LS estimates for coefficients if predictors are linearly dependent

- Ex. if $X_1 + X_2 = X_3$, then in the model

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 + X_2) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3) X_1 + (\beta_2 + \beta_3) X_2 + \varepsilon \end{aligned}$$

the coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$ and $(\beta_0, \beta_1', \beta_2', \beta_3')$ give identical mean of Y if $\beta_1' = \beta_1 + \beta_3$, $\beta_2' = \beta_2 + \beta_3$, $\beta_3' = 0$.

The Problem of Multicollinearity (MC)

- *Multicollinearity (MC)* means predictors in an MLR model have a **close-to-exact linear relationship**, i.e., predictors are nearly linearly dependent
- When MC problem exists, the LS estimates for β_j's exists but would have large variability.
- Recall we interpret $\hat{\beta}_j$ as the mean response change when X_j increases by one unit **holding all other predictors fixed**. If predictors are strongly correlated, we might not be able to alter X_j while holding other predictors fixed.

Equal Educational Opportunity (EEO) Data

Ex: Equal Educational Opportunity (EEO) Data

Data: http://www.stat.uchicago.edu/~yibi/s224/data/P236.txt

To examine the existence (or lack) of equal educational opportunities in public educational institutions, the following variables were measured for 70 schools selected at random in 1965.

- `ACHV`: Student achievement index (higher values are better)
- `FAM`: Faculty credentials index
- `PEER`: the influence of their peer group in the school
- `SCHOOL`: School facility/resource index

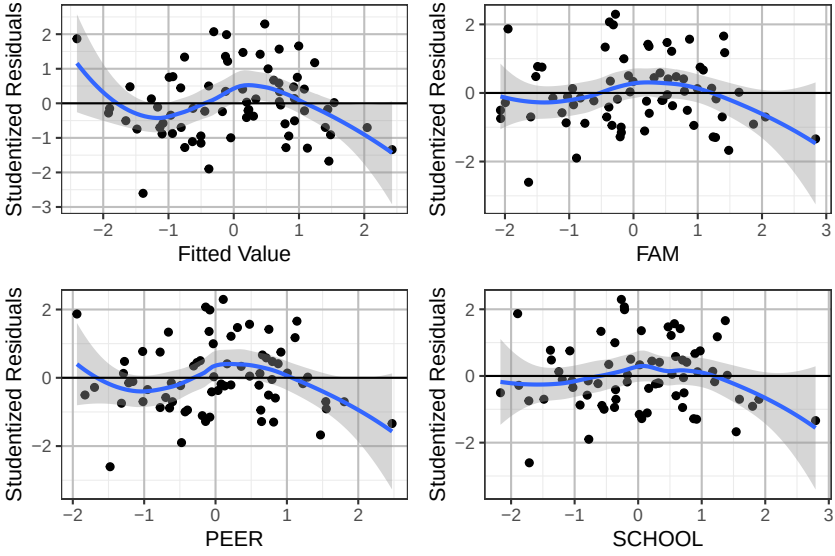
Goal: to identify important determinants of student achievement

```
EEO = read.table("P236.txt", h=T)
```

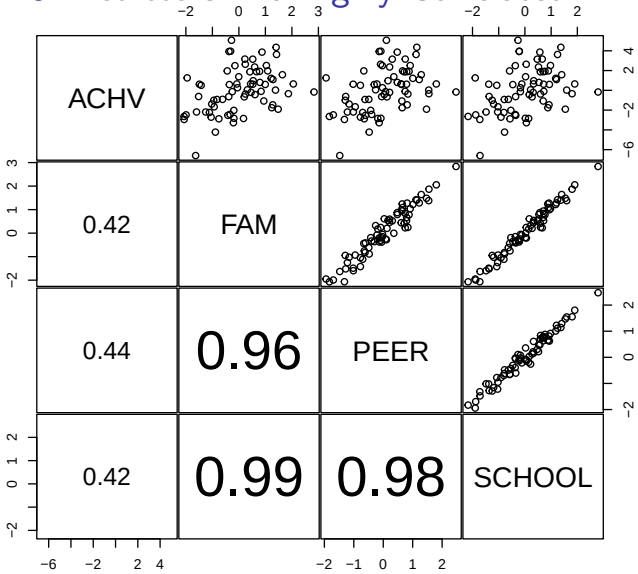
```
summary(lm(ACHV ~ FAM + PEER + SCHOOL, data=EEO))

Call:
lm(formula = ACHV ~ FAM + PEER + SCHOOL, data = EEO)

Residuals:
    Min      1Q  Median      3Q     Max
-5.210 -1.393 -0.295  1.142  4.588

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.070      0.251   -0.28     0.78
FAM            1.101      1.411    0.78     0.44
PEER           2.322      1.481    1.57     0.12
SCHOOL        -2.281      2.220   -1.03     0.31

Residual standard error: 2.07 on 66 degrees of freedom
Multiple R-squared:  0.206, Adjusted R-squared:  0.17
F-statistic: 5.72 on 3 and 66 DF,  p-value: 0.00153
```

- None of the 3 predictors is significant but the overall F-stat is significant
  (i.e., at least one of the 3 predictor is significant)
- SCHOOL has a negative coefficient! Normally, we expect higher student achievement if schools have more resources.

All Residual Plots Look Fine



All 3 Predictors Are Highly Correlated!



- Extremely high correlations between the 3 predictors!
- Any of 3 predictors can accurately predict the other two.
- So it's almost like we have only one predictor.
  Only 1 of the 3 predictors is really needed.
- However, multicollinearity prevents us from identifying the important predictors

Look! FAM, PEER, and SCHOOL are all significant if only one of them is included in the model.

```
summary(lm(ACHV ~ FAM, data=EEO))$coef
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -0.02427     0.2486 -0.09761 0.922526
FAM          0.88010     0.2310  3.81036 0.000301
summary(lm(ACHV ~ PEER, data=EEO))$coef
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03087     0.2460 -0.1255 0.900525
PEER         1.08090     0.2676  4.0387 0.000139
summary(lm(ACHV ~ SCHOOL, data=EEO))$coef
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept) -0.01043     0.2487 -0.04194 0.9666696
SCHOOL       0.92834     0.2446  3.79540 0.0003163
```

The multiple R^2 for the 3 models above are 0.1759, 0.1935, and 0.1748 respectively, all close to the multiple $R^2 \approx 0.2063$ of the model including all 3 predictors.

```
lmFPS = lm(ACHV ~ FAM+PEER+SCHOOL, data=EEO)
lmF = lm(ACHV ~ FAM, data=EEO)
lmP = lm(ACHV ~ PEER, data=EEO)
lmS = lm(ACHV ~ SCHOOL, data=EEO)
```

```
anova(lmF, lmFPS)
```

```
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1     68 294
2     66 283  2      10.8 1.26   0.29
```

```
anova(lmP, lmFPS)
```

```
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1     68 287
2     66 283  2      4.56 0.53   0.59
```

```
anova(lmS, lmFPS)
```

```
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1     68 294
2     66 283  2      11.2 1.31   0.28
```

All 3 single-predictor models fit the data nearly as well as the model including all 3 predictors. Cannot tell which predictor is more important.

Summary: Effects of Multicollinear Data

- Trouble in identifying important predictors
- Estimates are very sensitive to what other variables exist in the model,
  - the estimated coefficient of SCHOOL changes from -2.281 to 0.928 when FAM and PEER is removed from the model.

```
lm(ACHV ~ FAM+PEER+SCHOOL, data=EEO)$coef
(Intercept)          FAM         PEER       SCHOOL
   -0.06996      1.10126      2.32206     -2.28100
lm(ACHV ~ SCHOOL, data=EEO)$coef
(Intercept)       SCHOOL
   -0.01043      0.92834
```

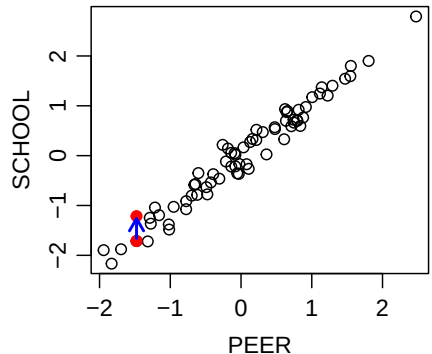
- Estimated coefficients are very sensitive to small changes in data.
  See the example on the next page.

If the value of SCHOOL of the 28th school is decreased by 0.5 from -1.713 to -2.213, observe the estimates for β's change drastically!

```
EEO.new = EEO
EEO.new$SCHOOL[28] = EEO$SCHOOL[28] - 0.5
lm(ACHV ~ FAM+PEER+SCHOOL, data=EEO.new)$coef
(Intercept)          FAM         PEER       SCHOOL
   -0.02081     -0.28845      0.94995      0.41394
```

compared to the original estimates

```
lm(ACHV ~ FAM+PEER+SCHOOL, data=EEO)$coef
(Intercept)          FAM         PEER       SCHOOL
   -0.06996      1.10126      2.32206     -2.28100
```



Why Multicollinearity Making Predictors Insignificant?

In Slides L04.pdf, we said the LS estimate $\hat{\beta}_1$ for FAM in the MLR model

$$\text{ACHV} = \beta_0 + \beta_1\,\text{FAM} + \beta_2\,\text{PEER} + \beta_3\,\text{SCHOOL} + \varepsilon$$

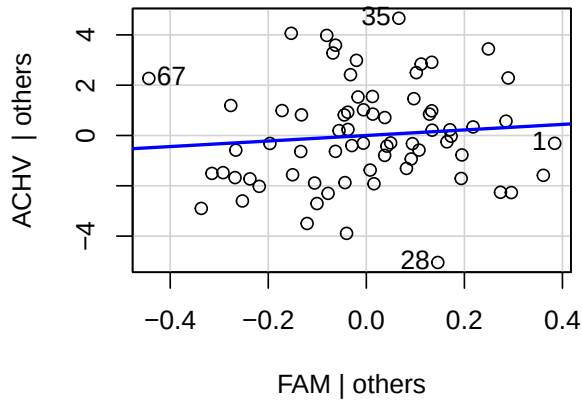
would be identical to the slope for the SLR model below

1. Regress ACHV on PEER and SCHOOL
2. Regress FAM on PEER and SCHOOL
3. Fit a SLR model using
   - the residuals from Step 1 as the response, and
   - the residuals from Step 2 as the predictor.

Recall in SLR, $s.e.(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\sum_i (x_i - \bar{x})^2}$ is inversely proportional to the SD of the predictor.

When FAM is highly-collinear with PEER and SCHOOL, the residuals in Step 2 would be nearly 0, and hence the SD ≈ 0 as well.
So $s.e.(\hat{\beta}_1)$ would be huge $\Rightarrow$ small t-value $\Rightarrow$ insignificant predictor

```r
library(car)
avPlots(lm(ACHV ~ FAM+PEER+SCHOOL, data=EEO), "FAM")
```



Missing Data Viewpoint of Multicollinearity

- To examine the effect of FAM on ACHV after accounting for PEER and SCHOOL, one need to fix the level of PEER and SCHOOL and the vary FAM, and see how ACHV changes.
- In other words, we need observations with similar values of PEER and SCHOOL but different FAM.
- However, as the 3 predictors are highly collinear, once PEER and SCHOOL are fixed, FAM is nearly completely determined.

Missing Data Viewpoint of Multicollinearity

We can divide the 3-dim space of our predictors into 8 regions, based on having high/low values for each variable.

Combination	FAM	PEER	SCHOOL
1	+	+	+
2	+	+	−
3	+	−	+
4	−	+	+
5	+	−	−
6	−	+	−
7	−	−	+
8	−	−	−

- To completely understand the process, we need data from all eight combinations.
- But we only have $(+,+,+)$ and $(-,-,-)$ in our data.

A Solution to Tackle Multicollinearity

Problem: We do not have the necessary diversity of data to separate the effects of `FAM`, `PEER`, and `SCHOOL`.

- One solution to tackle multicollinearity is to obtain more data with the missing combinations of predictors
- Better to taking to obtain more combinations when the data are collected, though not always possible
- Moreover, obtaining more data is not always possible due to limitations on time or money
- Sometimes such observations doesn't exist

Inherent Multicollinearity

- Sometimes, MC is an inherent characteristic of the variables being studied.
- `FAM`, `PEER`, and `SCHOOL` may only exist in the population in combinations $(-,-,-)$ and $(+,+,+)$.
- Hence, it may be impossible to sample from the other combinations.
- In this case, we may seek to explain what causes the correlation to discover more fundamental variables

Example: French Import Data

Example: French Import Data (p.241)

Goal: to understand the relation betw. French import and other economic variables.

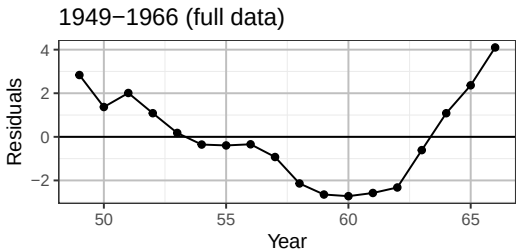
- `IMPORT`: Amount of Import
- `DOPROD`: Amount of Domestic Production
- `STOCK`: Amount of Stock Formation
- `CONSUM`: Amount of Domestic Consumption
- `YEAR`: Last 2 digits of year 1949-1966

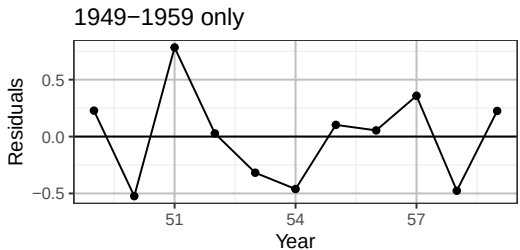
All measured in billions of French francs in 1949-1966.
Data: http://www.stat.uchicago.edu/~yibi/s224/data/P241.txt

Questions:

- If the relationship between `IMPORT` and the other variables is volatile, then our predictions will be unreliable.
- These are concerns for all forecasting models and are not a problem specific to multicollinearity.

```
p241 = read.table("P241.txt", h=T)
french1 = lm(IMPORT ~ DOPROD + STOCK + CONSUM, data=p241)
french2 = lm(IMPORT ~ DOPROD + STOCK + CONSUM, data=p241, subset=YEAR<60)
```





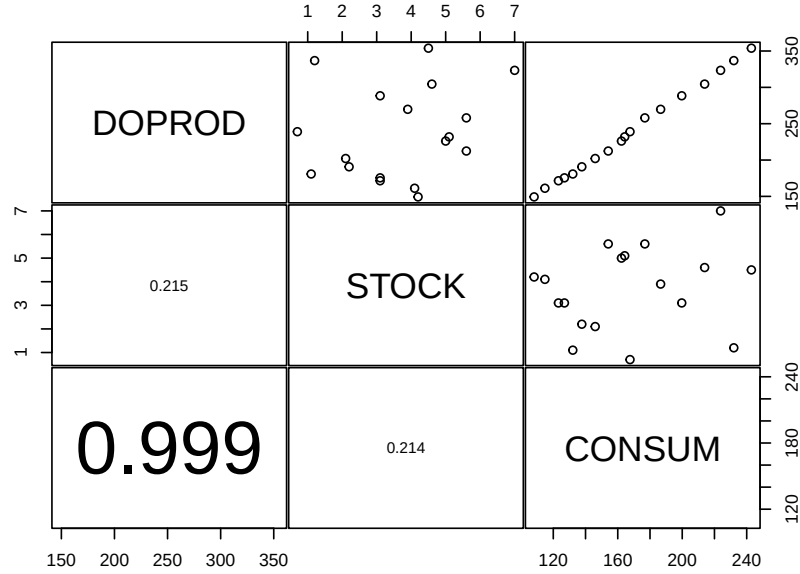
- Residual plot of the full data is not satisfactory.
- Inception of European Common Market in 1960 seemed to change the relation
- We focus only on Years 1949-1959 to simplify our discussion of multicollinearity. and the residual plot looks fine.

Fitted Model (1949-1959)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.12799    1.21216  -8.355  6.9e-05 ***
DOPROD       -0.05140    0.07028  -0.731 0.488344
STOCK         0.58695    0.09462   6.203 0.000444 ***
CONSUM        0.28685    0.10221   2.807 0.026277 *
---
Residual standard error: 0.4889 on 7 degrees of freedom
Multiple R-squared: 0.9919, Adjusted R-squared: 0.9884
F-statistic: 285.6 on 3 and 7 DF,  p-value: 1.112e-07
```

- R^2 and F-test indicate that the model is very significant.
- `STOCK` and `CONSUM` are significant.
- `DOPROD` and `IMPORT` should be positively correlated
  - Generally, higher Domestic Production requires importing more materials.
  - However, `DOPROD` has a insignificant negative slope! Why?

Pairwise Scatterplot of Predictors (French Economy Data)



Relationship between DOPROD and CONSUM

- `DOPROD` and `CONSUM` have a correlation of about 0.999!
  `CONSUM` explains $0.999^2 \approx 99.8\%$ of the variance in `DOPROD`.
- The least squares relationship is:

$$\text{CONSUM} = 6.259 + .686 \cdot \text{DOPROD}$$

- Hence, `CONSUM` consisted of about $69\% \approx 2/3$ of `DOPROD` each year.
- Intuitively, it makes sense that CONSUMption is correlated positively with DOmestic PRODuction.

Questions of Interest

1. How can MC be detected?
2. How does MC affect statistical inference and forecasting?
3. How can we resolve problems with MC?

Detection of Multicollinearity

Looking for Multicollinearity

Multicollinearity is...

- associated with unstable coefficient estimates.
- a result of linear relationships between predictors.
- not due to model mis-specification.

Therefore, we do not worry about fixing multicollinearity until the model diagnostics of other assumptions are satisfactory.

- Nevertheless, some indications of multicollinearity arise during the process of adding and removing variables and altering or removing observations.

Signs of Multicollinearity

While finding a good model, look for instability in estimated $\hat{\beta}_j$'s:

- Large changes in some $\hat{\beta}_j$'s when a variable is added or deleted.
- Large changes in some $\hat{\beta}_j$'s when a data point is altered or dropped.

Once the model fit is good, look for:

- Signs of some $\hat{\beta}_j$'s do not conform to prior expectations.
- Coefficients or variables that are expected to be important have large standard errors (small t values.)

Examples:

- Standard errors for all predictors are high in the EEO Data, resulting in small t-values. We would expect all three to be important.
- t-value for `DOPROD` was small and negative, when we would expect it to be positive and important.

Multicollinearity and Correlation

- Of course, the pairwise scatterplot is helpful for detecting multicollinearity.
- Unfortunately, it only helps to detect linear relationships between pairs of variables.
- There may be a higher-level relationship even with no pairwise correlations.
- The next example exhibits this type of sneaky multicollinearity.

Example: Tricky Multicollinearity — Sales Data

Data: http://www.stat.uchicago.edu/~yibi/s224/data/P248.txt

```
p248 = read.table("P248.txt", h=T)
```

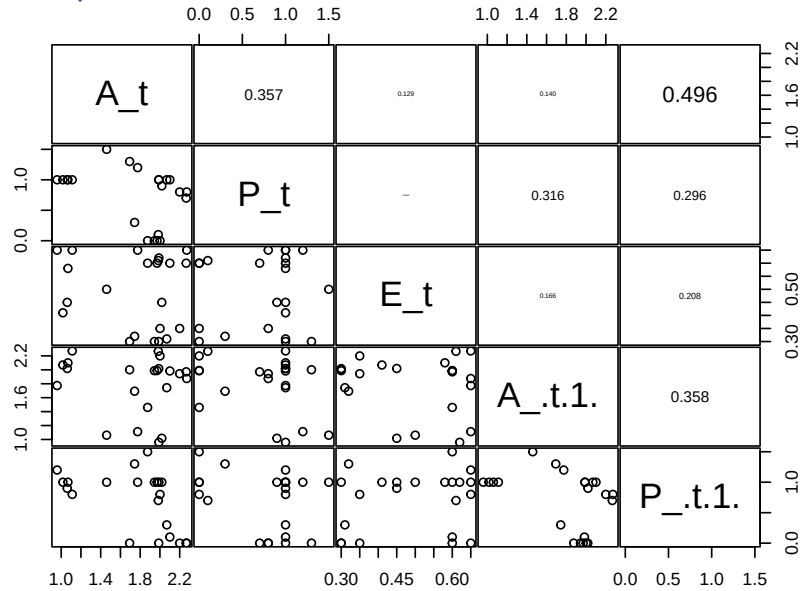
A_t	Advertising Expenditures
P_t	Promotion Expenditures
E_t	Sales Expense
S_t	Aggregate Sales (Response)

Proposed Model:

$$S_t = \beta_0 + \beta_1 A_t + \beta_2 P_t + \beta_3 E_t + \beta_4 A_{t-1} + \beta_5 P_{t-1} + \varepsilon_t$$

- Pairwise correlations are all small ($|\hat{\rho}_{ij}| < 0.5$).
- Volatility of estiamted coefficients indicates MC.

Pairwise Scatterplot — Sales Data



```
summary(lm(S_t ~ E_t+A_t+P_t+A_.t.1.+P_.t.1., data = p248))$coef
             Estimate Std. Error t value      Pr(>|t|)
(Intercept)  -14.194     18.715 -0.7584 0.45923402483
E_t           22.521      2.142 10.5123 0.00000001365
A_t            5.361      4.028  1.3310 0.20185348337
P_t            8.372      3.586  2.3345 0.03293209971
A_.t.1.        3.855      3.578  1.0774 0.29728619770
P_.t.1.        4.125      3.895  1.0590 0.30534041713
summary(lm(S_t ~ E_t + P_t + A_.t.1. + P_.t.1., data = p248))$coef
             Estimate Std. Error t value      Pr(>|t|)
(Intercept)  10.5094     2.4576  4.2763 0.000510400859
E_t          22.7942     2.1804 10.4544 0.000000008041
P_t           3.7018     0.7571  4.8893 0.000138233208
A_.t.1.      -0.7692     0.8746 -0.8795 0.391387968077
P_.t.1.      -0.9687     0.7423 -1.3050 0.209273037432
```

When we delete A_t, the estimates and s.e. for coefficients of P_t, A_{t-1}, and P_{t-1} change wildly! Sign of multicollinearity.

Higher-Level Correlation (Sales Data)

- Regression diagnostics for both models are satisfactory (not shown).
- It turns out that the firm exercised strict control over promotion and sales expenditures:

$$A_t + P_t + A_{t-1} + P_{t-1} \approx 5$$

- R^2 for regressing A_t on these 3 variables is .9727.
- Hence P_t, A_{t-1}, and P_{t-1} together explain 97% of the variability in A_t.
- This linear relationship is not seen in any pairwise correlations.

Variance Inflation Factor

The variance inflation factor for a predictor X_j in the MLR model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

is defined to be

$$\text{VIF}_j = \frac{1}{1 - R_j^2}, \quad j = 1, \ldots, p,$$

where R_j^2 is the multiple R^2 for regressing X_j on the other predictors in the model.

- If X_j is orthogonal (has 0 correlation) to all other predictors, then $R_j^2 = 0$ and $\text{VIF}_j = 1$.
- Increasing values of VIF_j indicate departure from orthogonality toward multicollinearity.
- A rule of thumb: VIFs > 10 suggest multicollinearity

Intrepeting VIF_j

- In simple linear regression, $\text{Var}(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2$.
- In multiple linear regression, the situation is complicated by the existence of correlation among the predictors. **This is what VIF measures**.
- In fact, one can show that

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2} \cdot \text{VIF}_j$$

  - We see that it looks like the SLR equation, but with an extra factor of VIF_j.
- VIF_j indicates the proportional increase in $\text{Var}(\hat{\beta}_j)$ due to its collinearity with other predictors, relative to the orthogonal case.

VIF in R

The `vif()` function in the `car` library can calculate the VIF_j for us.

```
library(car)
vif(lm(ACHV ~ FAM + PEER + SCHOOL, data=EEO))
   FAM   PEER SCHOOL
 37.58  30.21  83.16
```

VIF in R

The `vif()` function in the `car` library can calculate the VIF_j for us.

```
library(car)
vif(lm(ACHV ~ FAM + PEER + SCHOOL, data=EEO))
   FAM   PEER SCHOOL
 37.58  30.21  83.16
```

Observe the VIF for `SCHOOL` in the model above is

$$\frac{1}{1-R^2} = \frac{1}{1-0.987974} \approx 83.155$$

where $R^2 = 0.987974$ is the multiple R^2 of regressing `SCHOOL` on the other two predictors `FAM` and `PEER`.

```
summary(lm(SCHOOL ~ FAM + PEER, data=EEO))$r.squared
[1] 0.987974
```

Variance Inflation Due to Multicollinearity

Observe how much the `Std. Error` for `SCHOOL` is inflated when the `FAM` and `PEER` are included

```
summary(lm(ACHV ~ SCHOOL, data=EEO))$coef
            Estimate Std. Error  t value  Pr(>|t|)
(Intercept) -0.01043     0.2487 -0.04194 0.9666696
SCHOOL       0.92834     0.2446  3.79540 0.0003163
summary(lm(ACHV ~ FAM + PEER + SCHOOL, data=EEO))$coef
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06996     0.2506 -0.2791   0.7810
FAM          1.10126     1.4106  0.7807   0.4378
PEER         2.32206     1.4813  1.5676   0.1218
SCHOOL      -2.28100     2.2204 -1.0273   0.3080
```

The ratio of two s.e.'s of `SCHOOL` in the two models is $2.2204/0.2446 \approx 9.08$, close to $\sqrt{\text{VIF}_{\texttt{SCHOOL}}} = \sqrt{83.33} \approx 9.1285$.

VIF for Sales Data

$$S_t = \beta_0 + \beta_1 A_t + \beta_2 P_t + \beta_3 E_t + \beta_4 A_{t-1} + \beta_5 P_{t-1} + \varepsilon_t$$

```
vif(lm(S_t ~ E_t + A_t + P_t + A_.t.1. + P_.t.1., data = p248))
    E_t     A_t     P_t A_.t.1. P_.t.1.
  1.076  36.942  33.474  25.916  43.521
```

When A_t is excluded

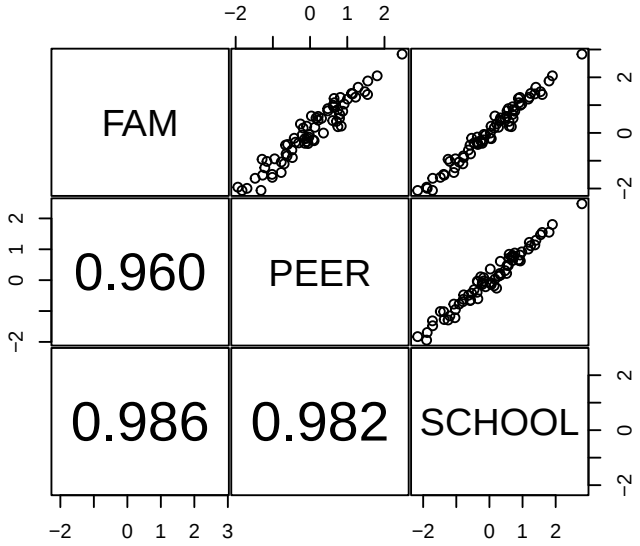
```
vif(lm(S_t ~ E_t + P_t + A_.t.1. + P_.t.1., data = p248))
    E_t     P_t A_.t.1. P_.t.1.
  1.066   1.427   1.481   1.512
```

Effects of MultiCollinearity on Prediction

Example: EEO Data

Recall all data points in the EEO data lie around the line

FAM $\approx$ PEER $\approx$ SCHOOL



Effects of MultiCollinearity on Prediction

The first model suffers from multicollinearity, but the second one doesn't.

```
lmFPS = lm(ACHV ~ FAM+PEER+SCHOOL, data=EEO)  # multicollinearity
lmP = lm(ACHV ~ PEER, data=EEO)               # no-multicollinearity
```

The two models give similar confidence intervals for prediction along the line where all data points lie.

```
predict(lmFPS, data.frame(FAM=1, PEER=1, SCHOOL=1), interval="confidence")
    fit    lwr   upr
1 1.072 0.3321 1.813
predict(lmP, data.frame(PEER=1), interval="confidence")
   fit   lwr   upr
1 1.05 0.343 1.757
predict(lmFPS, data.frame(FAM=2, PEER=2, SCHOOL=2), interval="confidence")
    fit    lwr   upr
1 2.215 0.9783 3.451
predict(lmP, data.frame(PEER=2), interval="confidence")
    fit    lwr   upr
1 2.131 0.9781 3.284
```

Effects of MultiCollinearity on Prediction

However, predictions off the space/line/hyperplane whether the data points lie based on the two models can be very different

```
predict(lmFPS, data.frame(FAM=1, PEER=1, SCHOOL=0), interval="confidence")
    fit    lwr   upr
1 3.353 -1.166 7.873
predict(lmP, data.frame(PEER=1), interval="confidence")
   fit   lwr   upr
1 1.05 0.343 1.757

predict(lmFPS, data.frame(FAM=0, PEER=1, SCHOOL=2), interval="confidence")
    fit    lwr   upr
1 -2.31 -9.291 4.671
predict(lmP, data.frame(PEER=1), interval="confidence")
   fit   lwr   upr
1 1.05 0.343 1.757
```

Example: Prediction of Sales Data

Same thing occurs for the Sales Data.

The model `sales1` below suffers from MC while the model `sales2` doesn't.

```
sales1 = lm(S_t ~ E_t +A_t+P_t+A_.t.1.+ P_.t.1., data = p248)
sales2 = lm(S_t ~ E_t + P_t + A_.t.1. + P_.t.1., data = p248)
```

Two models give similar confidence intervals for prediction around the hyperplane $A_t + P_t + A_{t-1} + P_{t-1} \approx 5$ where all data points lie around.

```
predict(sales1, data.frame(E_t=0.6, A_t=2, P_t=0, A_.t.1.=2, P_.t.1.=1),
        interval="confidence")
    fit   lwr   upr
1 21.87 20.56 23.19
predict(sales2, data.frame(E_t=0.6, P_t=0, A_.t.1.=2, P_.t.1.=1),
        interval="confidence")
    fit   lwr   upr
1 21.68 20.38 22.98
```

However, for predictions off the space/line/hyperplane of existing data points, the two models give very different confidence intervals for prediction.

```
predict(sales1, data.frame(E_t=0.6, A_t=2, P_t=0, A_.t.1.=2, P_.t.1.=0),
        interval="confidence")
    fit   lwr   upr
1 17.75 9.692 25.81
predict(sales2, data.frame(E_t=0.6, P_t=0, A_.t.1.=2, P_.t.1.=0),
        interval="confidence")
    fit  lwr  upr
1 22.65 20.6 24.7
predict(sales1, data.frame(E_t=0.6, A_t=2, P_t=1, A_.t.1.=1, P_.t.1.=0),
        interval="confidence")
    fit   lwr   upr
1 22.27 14.24 30.29
predict(sales2, data.frame(E_t=0.6, P_t=1, A_.t.1.=1, P_.t.1.=0),
        interval="confidence")
    fit   lwr   upr
1 27.12 24.91 29.33
```

Forecasting with Multicollinearity (French Import Data)

- If we wish to infer about the importance of the predictors, we are in trouble.
- Our goal is to use the model to forecast `IMPORT`, so we can get meaningful statements.
- Fitted Model:

$$\text{IMPORT} = -10.13 - 0.051 \cdot \text{DOPROD} + 0.587 \cdot \text{STOCK} + 0.287 \cdot \text{CONSUM}.$$

- Given accurate forecasts for `DOPROD`, `STOCK`, and `CONSUM`, we can still use this equation to predict `IMPORT`.
  - Assumption: The relationship stays consistent for the YEAR of prediction.
  - Note: This assumption is required of all forecasting models.

Forecasting with Multicollinearity (2)

Suppose we forecast an increased DOPROD of 10 units (1 unit = 1 billion of French francs) between 1959 (last year of data) and 1960.

Naive Forecast: $\text{IMPORT}_{1960} = \text{IMPORT}_{1959} - 0.051(10)$.

- This assumes STOCK and CONSUM stayed unchanged
- In reality, we believe CONSUM $= 6.259 + 0.686 \cdot$ DOPROD, therefore CONSUM would also be expected to increase by $0.686(10) = 6.86$ units.

Smart Forecast:

$$\begin{aligned} \text{IMPORT}_{1960} &= \text{IMPORT}_{1959} - 0.051(10) + 0.287(6.86) \\ &= \text{IMPORT}_{1959} + 1.46 \end{aligned}$$

- Taking the multicollinearity into account, we predict an increased IMPORT of 1.46 units.

The smart forecast would be nearly the same as the model using only DOPROD and STOCK but no CONSUM.

```
summary(lm(IMPORT ~ DOPROD + STOCK,data=p241, subset=YEAR<60))$coef
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  -8.4401    1.43518  -5.881 0.00036962781
DOPROD        0.1453    0.00703  20.672 0.00000003142
STOCK         0.6225    0.12787   4.868 0.00124285218
```