Multiple Linear Regression

Module 4

DATA 5600
Introduction to Regression and Machine Learning for Analytics
Marc Dotson

Module Overview

- Model formulation
- Visualization techniques
- Geometry of the model
- Exploratory data analysis methods
- Interpretation of the coefficients and model output
- Multiple regression model assumptions

Supervisor Data

Complaints: handles appropriately
Privileges: allows special

- What makes a good or bad supervisor $(n = 30)$?

Supervisor	Rating	Complaints	Privileges	Learn	Raises	Critical	Advance
1	43	51	30	39	61	92	45
2	63	64	51	54	63	73	47
			…				
30	82	82	39	59	64	78	39

Learn: provides opportunities
Raises: based on performance?

Critical: too much on poor performance
Advance: not satisfied with rate of

- What is the response variable Y?

- What are the predictors (covariates, Xs)?

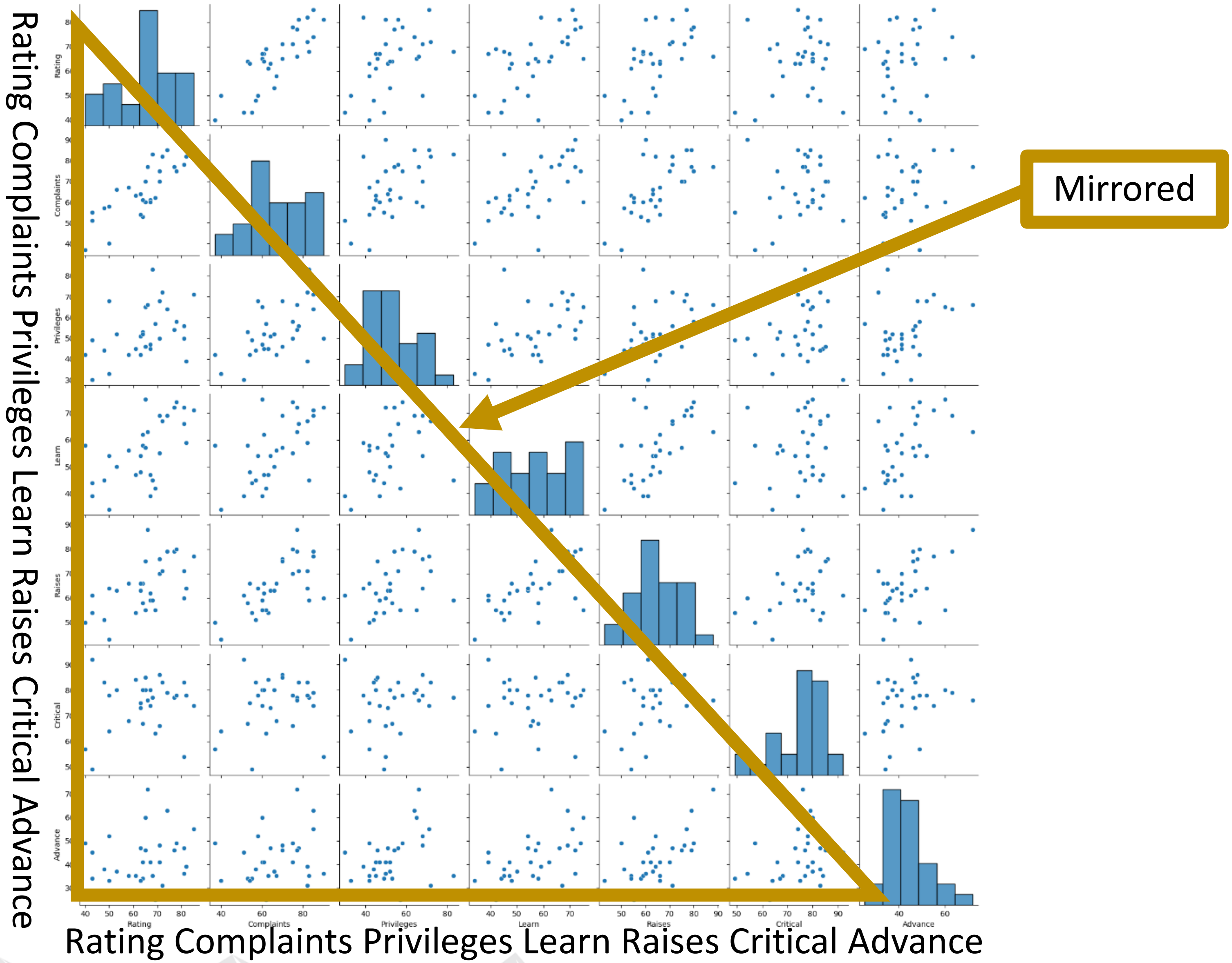
Changes from Simple Linear Regression

With multiple linear regression, a few things change:

1. Exploratory data analysis (EDA) methods
2. Geometry of the “line”
3. Interpretation of coefficients
4. Interpretation of confidence and prediction intervals
5. Interpretation of hypothesis tests
6. Interpretation of model evaluation metrics
7. Additional model assumption and diagnostics
8. Can include other variable types (interactions, higher order, etc.) (will discuss in a future module)

1. Exploratory Data Analysis (EDA) Methods

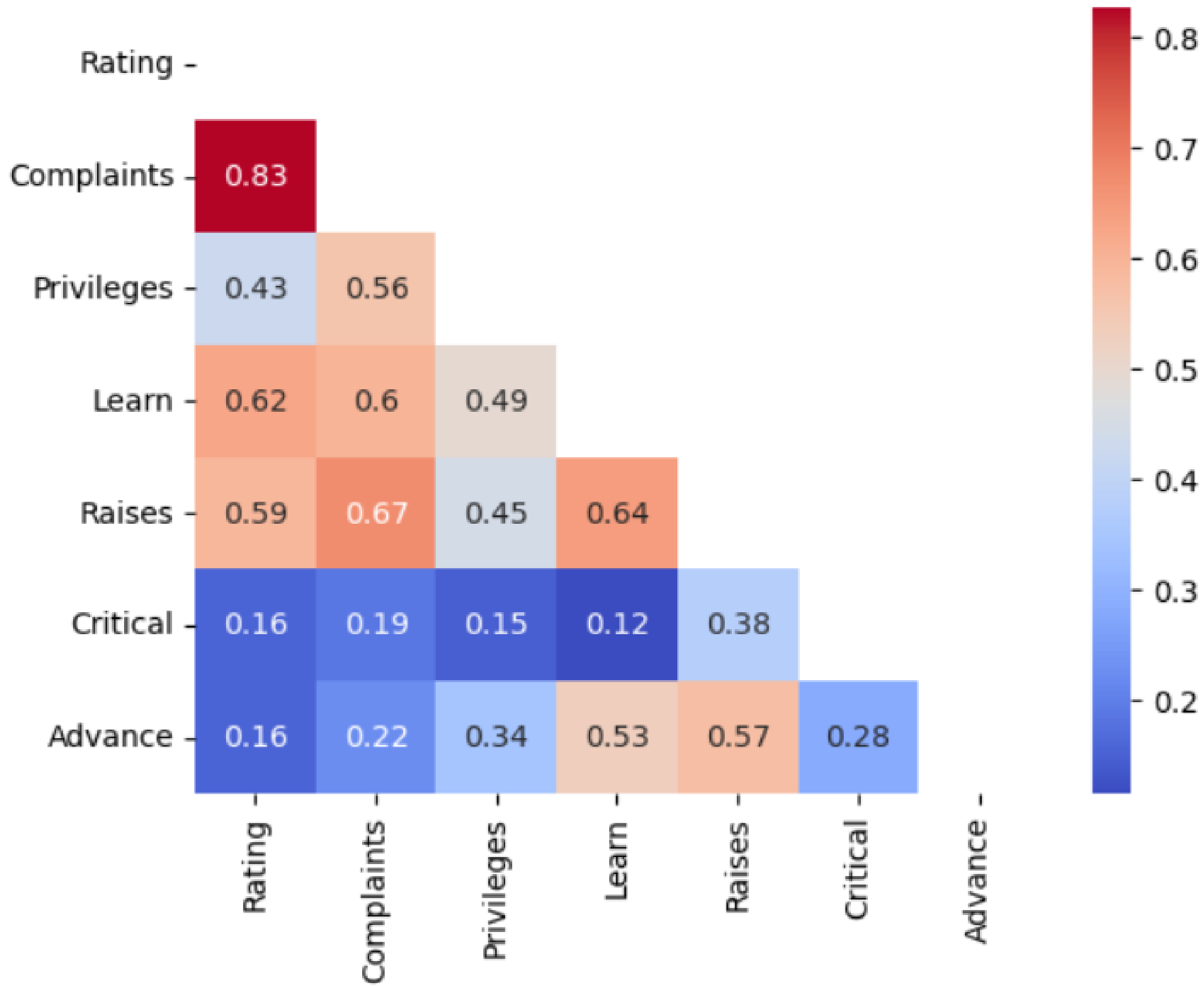
Scatterplot Matrix



Correlation Matrix

	Rating	Complaints	Privileges	Learn	Raises	Critical	Advance
Rating	1.000000	0.825418	0.426117	0.623678	0.590139	0.156439	0.155086
Complaints	0.825418	1.000000	0.558288	0.596736	0.669197	0.187714	0.224580
Privileges	0.426117	0.558288	1.000000	0.493331	0.445478	0.147233	0.343293
Learn	0.623678	0.596736	0.493331	1.000000	0.640314	0.115965	0.531620
Raises	0.590139	0.669197	0.445478	0.640314	1.000000	0.376883	0.574186
Critical	0.156439	0.187714	0.147233	0.115965	0.376883	1.000000	0.283343
Advance	0.155086	0.224580	0.343293	0.531620	0.574186	0.283343	1.000000

Color-Coded Correlation Matrix (Lower Half)

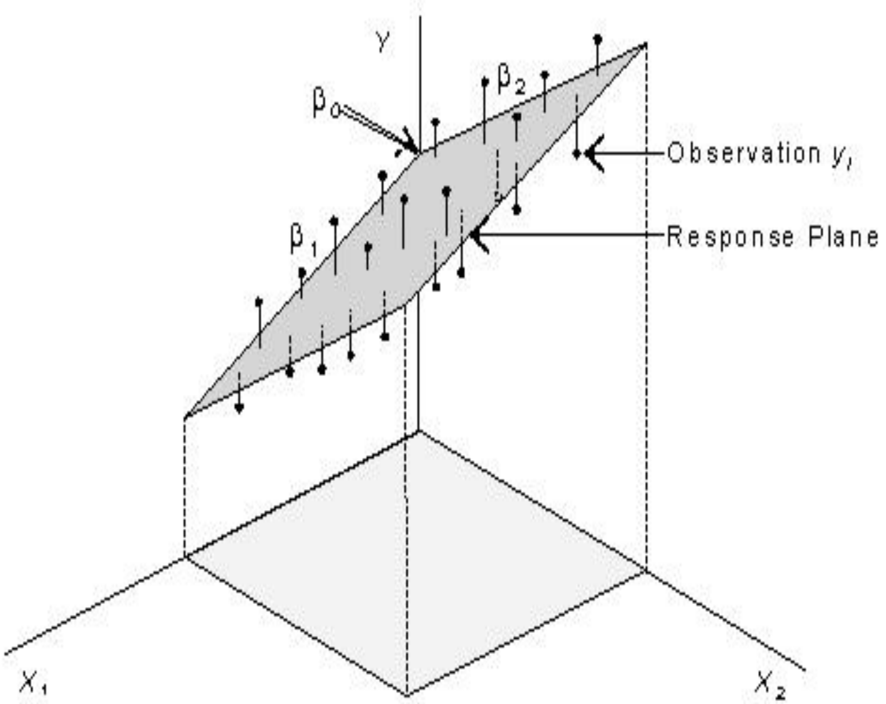


Code!

2. Geometry of the "Line"

Regression “Surface” Instead of “Line”

- For two predictors:



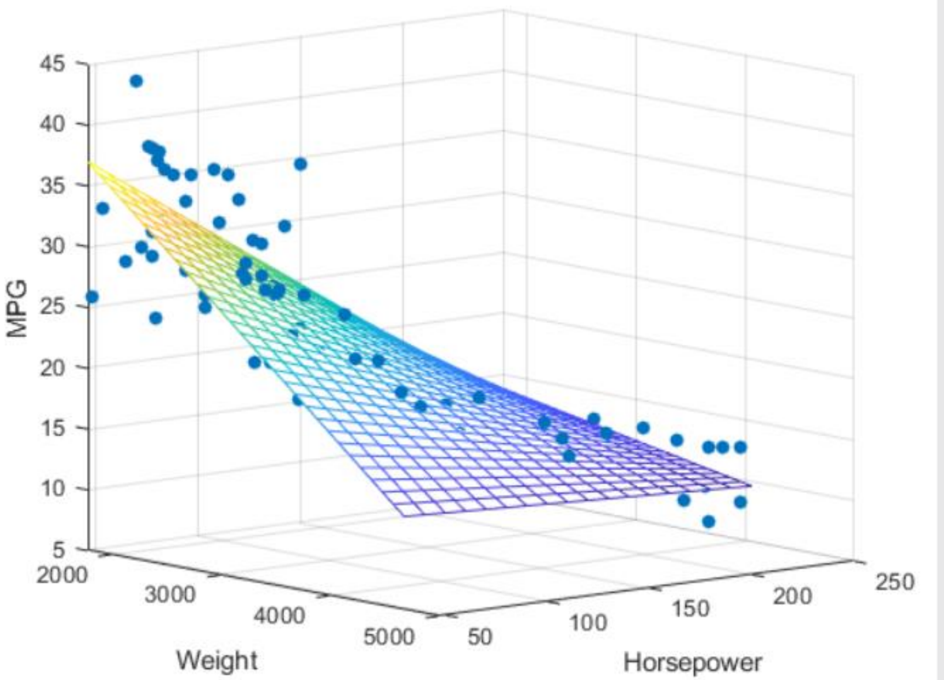


Figure from: https://towardsdatascience.com/graphs-and-ml-multiple-linear-regression-c6920a1f2e70

3. Interpretation of Coefficients

Code!

Supervisor Data Model Output 3. Coefficients

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.7871	11.589	0.931	0.362	-13.187	34.761
Complaints	0.6132	0.161	3.809	0.001	0.280	0.946
Privileges	-0.0731	0.136	-0.538	0.596	-0.354	0.208
Learn	0.3203	0.169	1.901	0.070	-0.028	0.669
Raises	0.0817	0.221	0.369	0.715	-0.376	0.540
Critical	0.0384	0.147	0.261	0.796	-0.266	0.342
Advance	-0.2171	0.178	-1.218	0.236	-0.586	0.152

Fitted Model:

$$\widehat{\text{Rating}}_i = 10.787 + 0.6132 \times \text{Complaints}_i - 0.073 \times \text{Privileges}_i + 0.3203$$
$$+0.0817 \times \text{Raises}_i + 0.0384 \times \text{Critical}_i - 0.217 \times \text{Advance}_i$$

Supervisor Data Model Output 3. Coefficients

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.7871	11.589	0.931	0.362	-13.187	34.761
Complaints	0.6132	0.161	3.809	0.001	0.280	0.946
Privileges	-0.0731	0.136	-0.538	0.596	-0.354	0.208
Learn	0.3203	0.169	1.901	0.070	-0.028	0.669
Raises	0.0817	0.221	0.369	0.715	-0.376	0.540
Critical	0.0384	0.147	0.261	0.796	-0.266	0.342
Advance	-0.2171	0.178	-1.218	0.236	-0.586	0.152

Note: just insert values for all x's in the model to make predictions (extrapolation happens when you predict outside the range of at least one covariate)

Fitted Model:

$$\widehat{\text{Rating}}_i = 10.787 + 0.6132 \times \text{Complaints}_i - 0.073 \times \text{Privileges}_i + 0.3203$$
$$+0.0817 \times \text{Raises}_i + 0.0384 \times \text{Critical}_i - 0.217 \times \text{Advance}_i$$

Supervisor Data Interpretations

$$\widehat{\text{Rating}}_i = 10.787 + 0.6132 \times \text{Complaints}_i - 0.073 \times \text{Privileges}_i + 0.3203 \times \text{Learn}_i$$
$$+0.0817 \times \text{Raises}_i + 0.0384 \times \text{Critical}_i - 0.217 \times \text{Advance}_i$$

- What is the interpretation of $\hat{\beta}_0 = 10.787$?

- What is the interpretation of $\hat{\beta}_1 = 0.6132$?

"Holding All Else Constant"

$$\widehat{\text{Rating}}_i = 10.787 + 0.6132 \times \text{Complaints}_i - 0.073 \times \text{Privileges}_i + 0.3203 \times \text{Learn}_i$$
$$+0.0817 \times \text{Raises}_i + 0.0384 \times \text{Critical}_i - 0.217 \times \text{Advance}_i$$

- The coefficients for Complaints represents the average change in Rating for every one unit increase in Complaints, <u>while all other predictors are held constant</u>.

- Coefficients are sometimes called "partial regression coefficients" because they reflects the partial effect of a predictor on the average of Y <u>after accounting for effects of other predictors</u>.

“Holding All Else Constant”

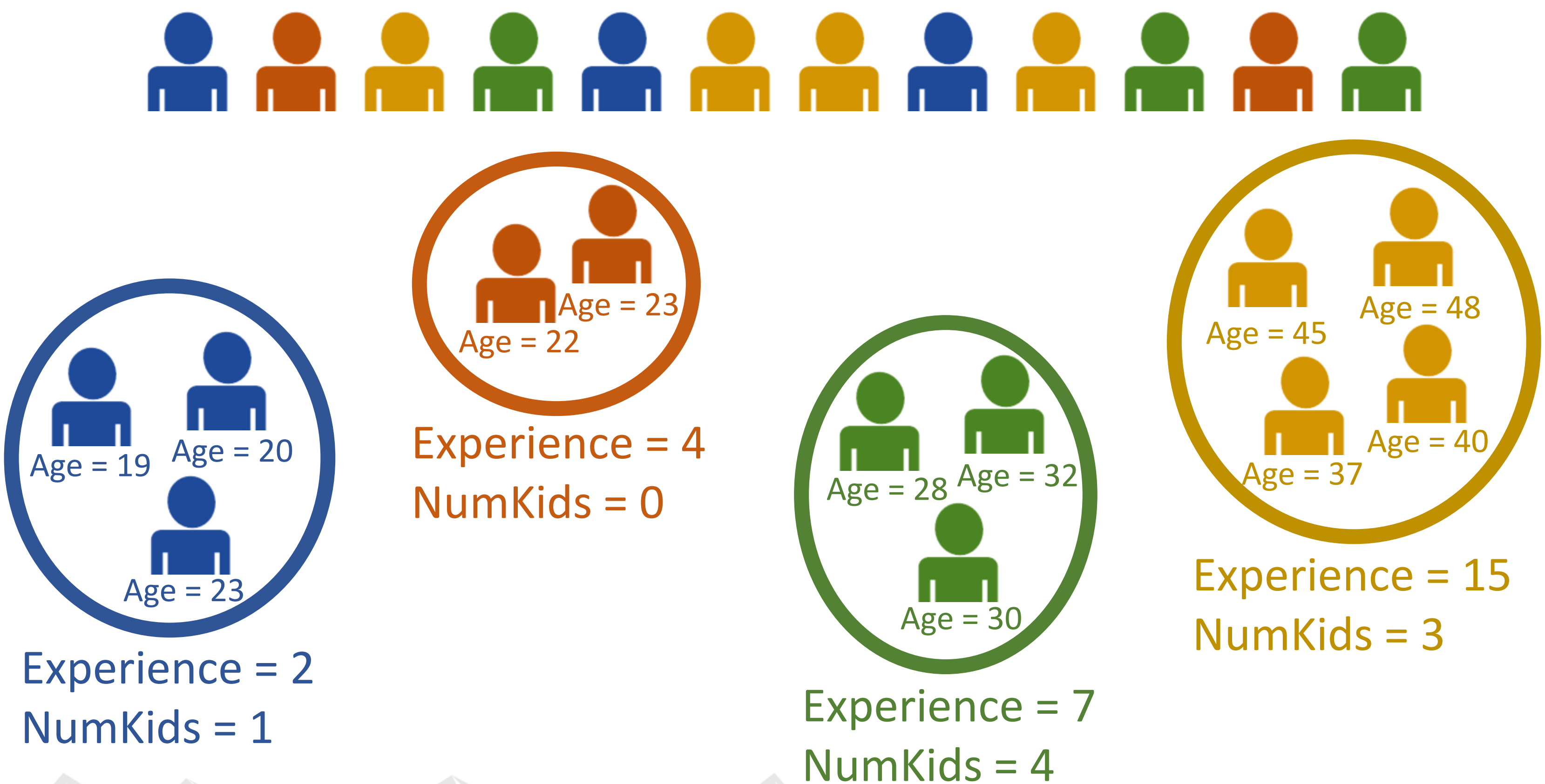
- It often does not make sense to “hold everything else constant” for a single individual. For example:
  - $\widehat{\text{Income}}_i = 9.7 + 2.2\ \text{Experience}_i + 1.5\ \text{NumKids}_i + 3.4\ \text{Age}_i$
  - We cannot gain a year of experience or have another child without getting a year older, so it doesn’t make sense to “hold age constant” when interpreting the coefficient for the number of kids, for instance.

- Instead, we can think about all those who fit given criteria on some predictors and think about the conditional relationship between Y and one X for those individuals.
  - Among people with the same years of experience and the same number of children, average income increases by 3.4 for every additional year in age.

“Holding All Else Constant”

$$\widehat{\text{Income}}_i = 9.7 + 2.2\ \text{Experience}_i + 1.5\ \text{NumKids}_i + 3.4\ \text{Age}_i$$

Among people with the same years of experience and the same number of children, average income increases by 3.4 for every additional year in age.

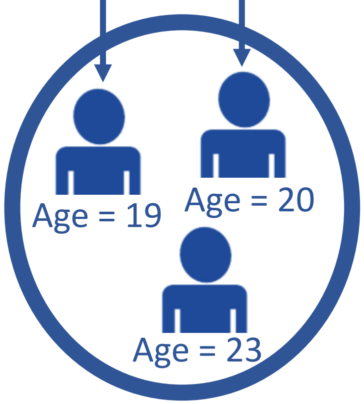


"Holding All Else Constant"

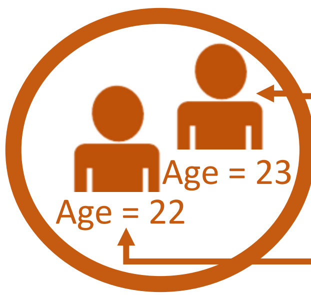
$$\widehat{\text{Income}}_i = 9.7 + 2.2\ \text{Experience}_i + 1.5\ \text{NumKids}_i + 3.4\ \text{Age}_i$$

Among people with the same years of experience and the same number of children, average income increases by 3.4 for every additional year in age.

$$\left.\begin{aligned}\widehat{\text{Income}}_i &= 9.7 + 2.2\,(2) + 1.5\,(1) + 3.4\,(19) = 80.2\\ \widehat{\text{Income}}_i &= 9.7 + 2.2\,(2) + 1.5\,(1) + 3.4\,(20) = 83.6\end{aligned}\right\}\ \text{Difference} = 3.4$$



Experience = 2
NumKids = 1



Experience = 4
NumKids = 0

$$\widehat{\text{Income}}_i = 9.7 + 2.2\,(4) + 1.5\,(0) + 3.4\,(23) = 96.7$$

$$\widehat{\text{Income}}_i = 9.7 + 2.2\,(2) + 1.5\,(1) + 3.4\,(22) = 93.3$$

Difference = 3.4

“Holding All Else Constant”

$$\widehat{\text{Income}}_i = 9.7 + 2.2\ \text{Experience}_i + 1.5\ \text{NumKids}_i + 3.4\ \text{Age}_i$$

These two statements are equivalent:

- “Holding all else constant, average income increases by 3.4 for every additional year in age.”
- “Among people with the same years of experience and the same number of children, average income increases by 3.4 for every additional year in age.”
  - This interpretation for the 3.4 coefficient applies for **any** value of Experience and NumKids!

Interpretation: Example

- Simple Linear Regression:

$$\widehat{\text{Fuel Efficiency}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Price}_i$$

  - $\hat{\beta}_1$ is negative and significant

- Multiple Linear Regression:

$$\widehat{\text{Fuel Efficiency}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Price}_i + \hat{\beta}_2 \text{Weight}_i$$

  - $\hat{\beta}_1$ is **positive** and significant
  - $\hat{\beta}_2$ is negative and significant

Interpretation: Notes

- If a variable we think should be important is not significant, that doesn't mean that the variable is not strongly linearly associated with the response. For example:
  - $\text{Horsepower}_i = \beta_0 + \beta_1 \text{Weight}_i + \beta_2 \text{EngineSize}_i$
  - If the coefficient for engine size is nearly 0, that does not mean that engine size is not important for understanding horsepower. It simply means that after allowing for the weight of the car, the engine size does not give much additional information.

Coefficient Confidence Intervals

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.7871	11.589	0.931	0.362	-13.187	34.761
Complaints	0.6132	0.161	3.809	0.001	0.280	0.946
Privileges	-0.0731	0.136	-0.538	0.596	-0.354	0.208
Learn	0.3203	0.169	1.901	0.070	-0.028	0.669
Raises	0.0817	0.221	0.369	0.715	-0.376	0.540
Critical	0.0384	0.147	0.261	0.796	-0.266	0.342
Advance	-0.2171	0.178	-1.218	0.236	-0.586	0.152

Note: $t_{\alpha/2}$ represents the upper $\alpha/2$ critical value from the t distribution with $n - p$ degrees of freedom (where p represents the number of βs in the model, including the intercept).

- A 95% confidence interval for β_1 (coefficient for Complaints) is

$$0.6132 \pm 2.0687 \times 0.1610 = (0.280, 0.946)$$

- How do you interpret this interval?

4. Interpretation of Confidence and Prediction Intervals

Confidence and Prediction Intervals

- Confidence intervals for the mean of Y

$$\hat{y}_i \pm t_{\alpha/2,n-p} SE_{CI}(\hat{y}_i)$$

- Prediction intervals for individual observations

$$\hat{y}_i \pm t_{\alpha/2,n-p} SE_{PI}(\hat{y}_i)$$

Note:

- $SE_{CI}(\hat{y}_i) < SE_{PI}(\hat{y}_i)$, so prediction intervals will be wider than confidence intervals

Confidence Interval: Supervisor Data

$$\widehat{\text{Rating}}_i = 10.787 + 0.6132 \times \text{Complaints}_i - 0.073 \times \text{Privileges}_i + 0.3203 \times \text{Learn}_i + 0.0817 \times \text{Raises}_i + 0.0384 \times \text{Critical}_i - 0.217 \times \text{Advance}_i$$

- The average rating for a supervisor with these scores [Complaints = 60, Privileges = 50, Learn = 56, Raises = 63, Critical = 76, Advance = 40] is

$$\begin{aligned}\widehat{\text{Rating}}_i &= 10.787 + 0.6132 \times 60 - 0.073 \times 50 + 0.3203 \times 56 + 0.0817 \times 63 \\ &\quad + 0.0384 \times 76 - 0.217 \times 40 \\ &= 61.25\end{aligned}$$

- 95% <u>confidence interval</u> for the average rating for a supervisor with the above scores is

$$(57.73, 64.76)$$

- Interpretation:

Prediction Interval: Supervisor Data

$$\widehat{\text{Rating}}_i = 10.787 + 0.6132 \times \text{Complaints}_i - 0.073 \times \text{Privileges}_i + 0.3203 \times \text{Learn}_i$$
$$+0.0817 \times \text{Raises}_i + 0.0384 \times \text{Critical}_i - 0.217 \times \text{Advance}_i$$

- The average rating for a supervisor with these scores [Complaints = 60, Privileges = 50, Learn = 56, Raises = 63, Critical = 76, Advance = 40] is

$$\widehat{\text{Rating}}_i = 10.787 + 0.6132 \times 60 - 0.073 \times 50 + 0.3203 \times 56 + 0.0817 \times 63$$
$$+0.0384 \times 76 - 0.217 \times 40$$
$$= 61.25$$

- 95% <u>prediction interval</u> for the rating of a supervisor with the above scores is

$$(46.21, 76.29)$$

- Interpretation:

5. Interpretation of Hypothesis Tests

Tests for Individual Variables

OLS Regression Results

Dep. Variable:	Rating	**R-squared:**	0.733
Model:	OLS	**Adj. R-squared:**	0.663
Method:	Least Squares	**F-statistic:**	10.50
Date:	Wed, 20 Sep 2023	**Prob (F-statistic):**	1.24e-05
Time:	10:37:29	**Log-Likelihood:**	-97.250
No. Observations:	30	**AIC:**	208.5
Df Residuals:	23	**BIC:**	218.3
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.7871	11.589	0.931	0.362	-13.187	34.761
Complaints	0.6132	0.161	3.809	0.001	0.280	0.946
Privileges	-0.0731	0.136	-0.538	0.596	-0.354	0.208
Learn	0.3203	0.169	1.901	0.070	-0.028	0.669
Raises	0.0817	0.221	0.369	0.715	-0.376	0.540
Critical	0.0384	0.147	0.261	0.796	-0.266	0.342
Advance	-0.2171	0.178	-1.218	0.236	-0.586	0.152

What do we learn about Rating from these tests?

Can we believe these results?

Note: If a coefficient's *t*-statistic is not significant, don't interpret the coefficient.

- You cannot be sure that the true parameter value is not really zero.

Test for Overall Model

OLS Regression Results

Dep. Variable:	Rating	**R-squared:**	0.733
Model:	OLS	**Adj. R-squared:**	0.663
Method:	Least Squares	**F-statistic:**	10.50
Date:	Wed, 20 Sep 2023	**Prob (F-statistic):**	1.24e-05
Time:	10:37:29	**Log-Likelihood:**	-97.250
No. Observations:	30	**AIC:**	208.5
Df Residuals:	23	**BIC:**	218.3
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	10.7871	11.589	0.931	0.362	-13.187	34.761
Complaints	0.6132	0.161	3.809	0.001	0.280	0.946
Privileges	-0.0731	0.136	-0.538	0.596	-0.354	0.208
Learn	0.3203	0.169	1.901	0.070	-0.028	0.669
Raises	0.0817	0.221	0.369	0.715	-0.376	0.540
Critical	0.0384	0.147	0.261	0.796	-0.266	0.342
Advance	-0.2171	0.178	-1.218	0.236	-0.586	0.152

What is this testing?

H_0: All coefficients equal zero

H_a: At least one coefficient is non–zero

What do we learn from this test?

Hypothesis Tests Comparison

We just talked about these:

- Hypothesis tests for individual variables
- Hypothesis test for overall model (testing all variables simultaneously)*

*Why test coefficients simultaneously?

- We want to avoid the multiple comparisons problem
  - Multiplicity (or multiple comparisons problem): if you do lots of tests then you are likely to commit errors (your overall type I error rate is inflated)

6. Interpretation of Model Evaluation Metrics

Supervisor Data Model Output

OLS Regression Results

Dep. Variable:	Rating	**R-squared:**	0.733
Model:	OLS	**Adj. R-squared:**	0.663
Method:	Least Squares	**F-statistic:**	10.50
Date:	Wed, 20 Sep 2023	**Prob (F-statistic):**	1.24e-05
Time:	10:37:29	**Log-Likelihood:**	-97.250
No. Observations:	30	**AIC:**	208.5
Df Residuals:	23	**BIC:**	218.3
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	10.7871	11.589	0.931	0.362	-13.187	34.761
Complaints	0.6132	0.161	3.809	0.001	0.280	0.946
Privileges	-0.0731	0.136	-0.538	0.596	-0.354	0.208
Learn	0.3203	0.169	1.901	0.070	-0.028	0.669
Raises	0.0817	0.221	0.369	0.715	-0.376	0.540
Critical	0.0384	0.147	0.261	0.796	-0.266	0.342
Advance	-0.2171	0.178	-1.218	0.236	-0.586	0.152

How do we interpret the R^2 value?

How do we interpret the adjusted R^2 value?

7. Additional Model Assumption and Diagnostics

Multiple Linear Regression Model Assumptions

Multiple linear regression inherits the assumptions of simple linear regression:

1. L – The X's vs Y are linear
2. I – The residuals are independent
3. N – The residuals are normally distributed and centered at zero
4. E – The residuals have equal (constant) variance σ^2 across all values of the X's (homoscedastic)
5. A – The model describes all observations (i.e., there are no influential points)
6. R – Additional predictor variables are not required

Additional assumption:

7. **No Multicollinearity**

Multiple Linear Regression Diagnostics

1. L – X vs Y is linear
   - Scatterplot Matrix
   - Residuals vs. Fitted Values Plot
   - Residuals vs. Predictor Plot
   - [best] Partial Regression Plots

2. I – The residuals are independent
   - [best] Think about how the data was collected

3. N – The residuals are normally distributed and centered at zero
   - Boxplot
   - Histogram
   - [best] Q-Q (Normal Probability) Plot
   - Shapiro-Wilk Test

4. E – The residuals have equal (constant) variance σ^2 across all values of X (homoscedastic)
   - Residuals vs. Fitted Values Plot

5. A – The model describes all observations (i.e., there are no influential points)
   - DFBETAS
   - DFFITS

6. R – Additional predictor variables are not required
   - Think about it

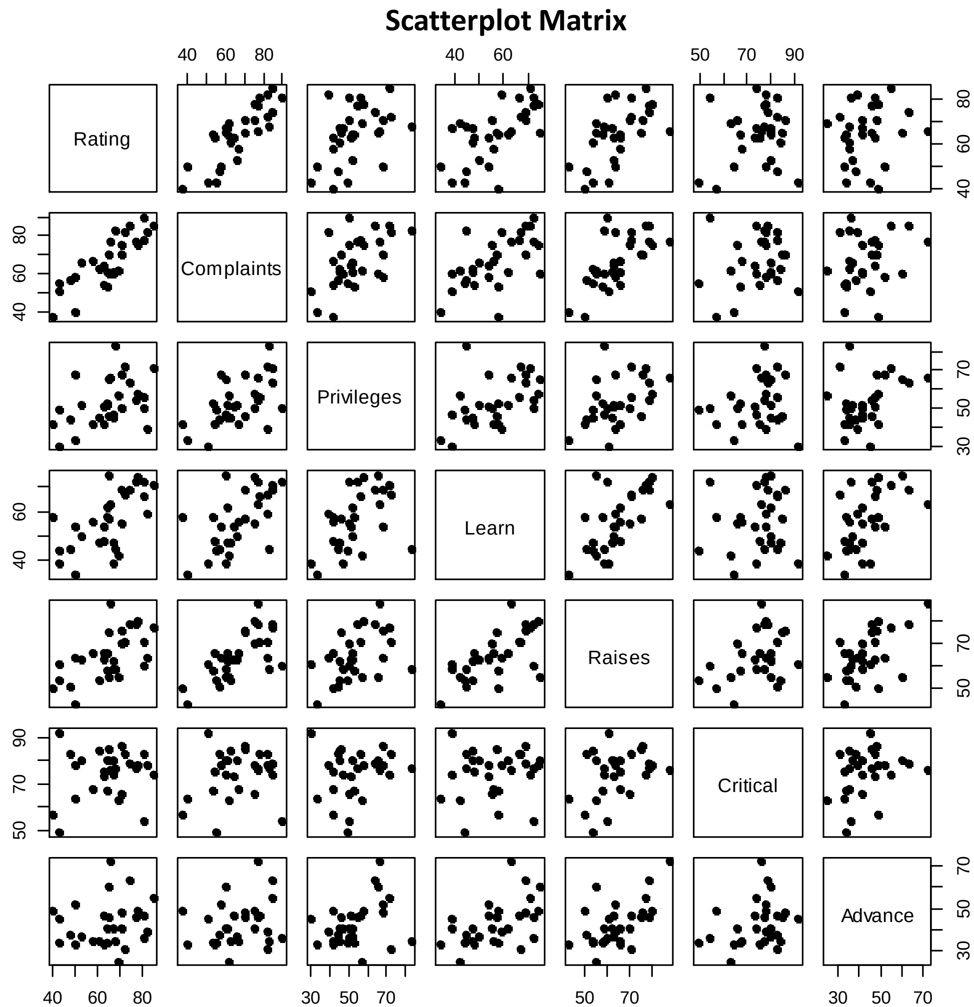
7. No Multicollinearity (discuss in coming slides)
   - Scatterplot Matrix
   - Correlation Matrix
   - Variance Inflation Factors

Partial Regression Plots

How to Visualize?

- Once we have our slope estimates, how can we plot our fitted regression line?
  - We could look at the scatterplot matrix, but it does not take into account the effect of the other independent variables in the model

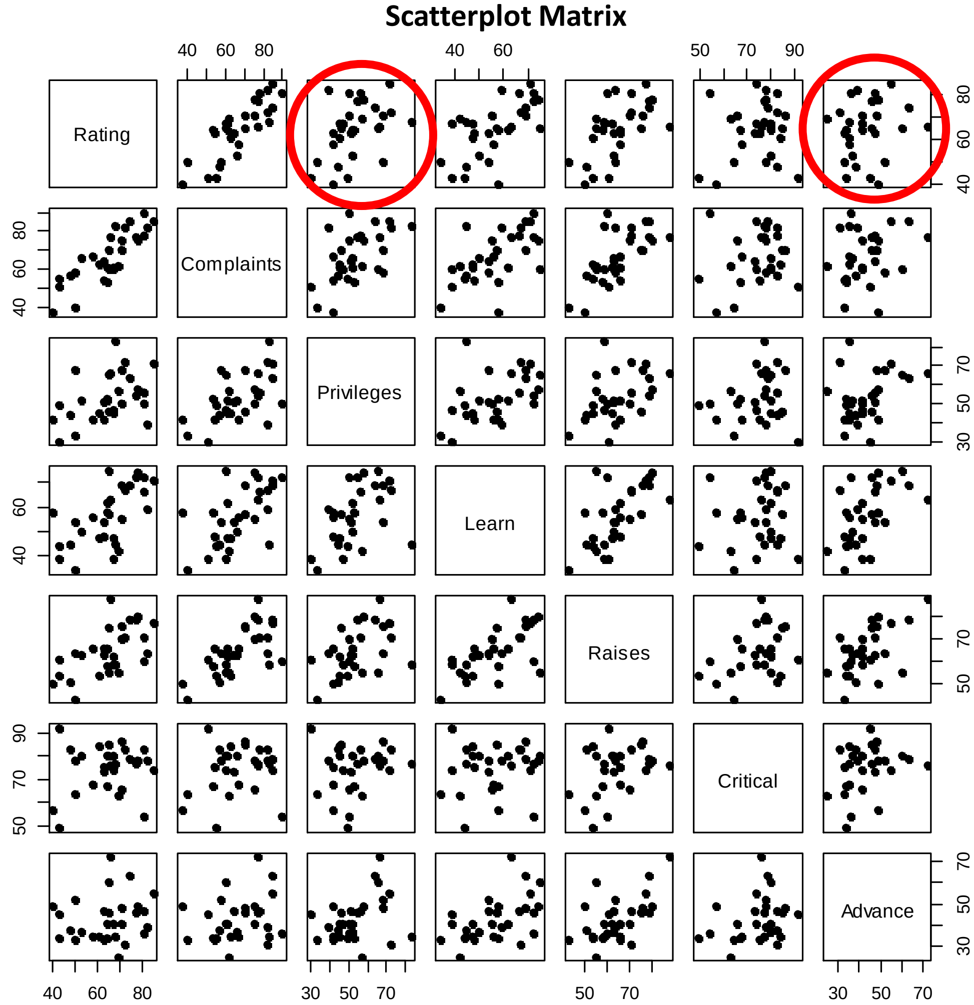
Complaints	$\hat{\beta}_1 =$	0.6132
Privileges	$\hat{\beta}_2 =$	-0.0731
Learn	$\hat{\beta}_3 =$	0.3203
Raises	$\hat{\beta}_4 =$	0.0817
Critical	$\hat{\beta}_5 =$	0.0384
Advance	$\hat{\beta}_6 =$	-0.2171



How to Visualize?

- Once we have our slope estimates, how can we plot our fitted regression line?
  - We could look at the scatterplot matrix, but it does not take into account the effect of the other independent variables in the model

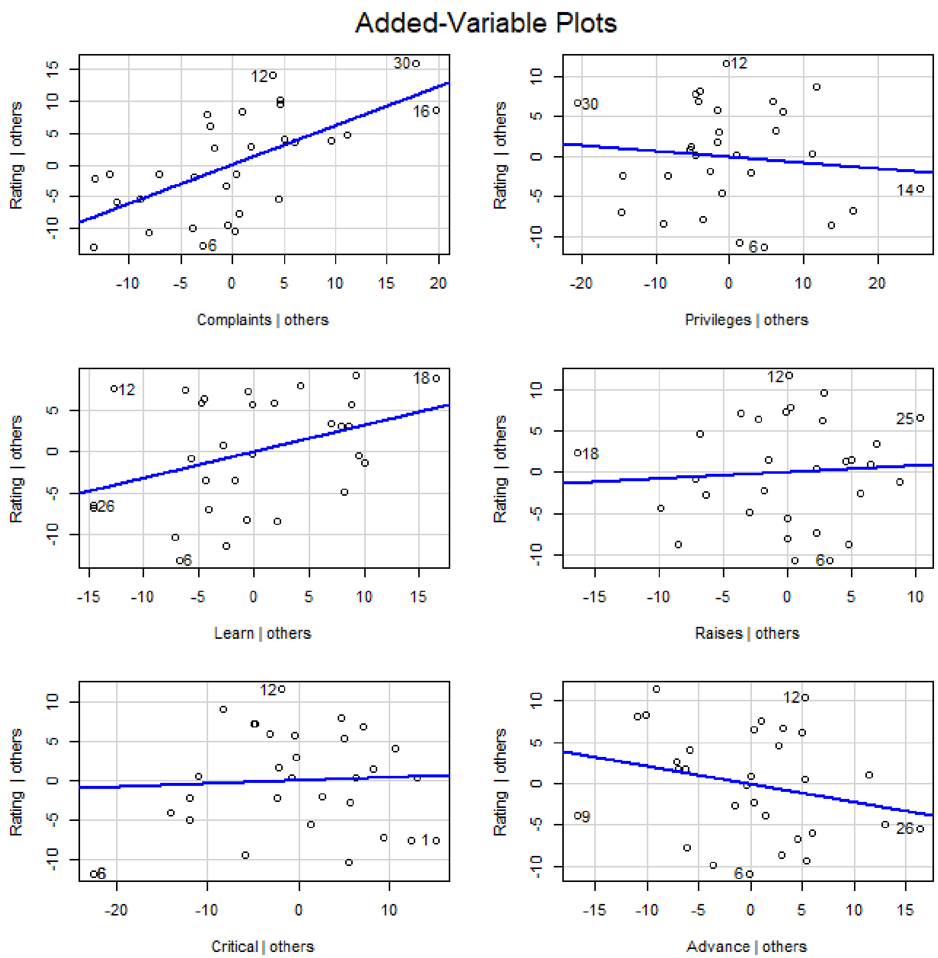
Complaints	$\hat{\beta}_1 =$	0.6132
Privileges	$\hat{\beta}_2 =$	−0.0731
Learn	$\hat{\beta}_3 =$	0.3203
Raises	$\hat{\beta}_4 =$	0.0817
Critical	$\hat{\beta}_5 =$	0.0384
Advance	$\hat{\beta}_6 =$	−0.2171



Partial Regression Plots

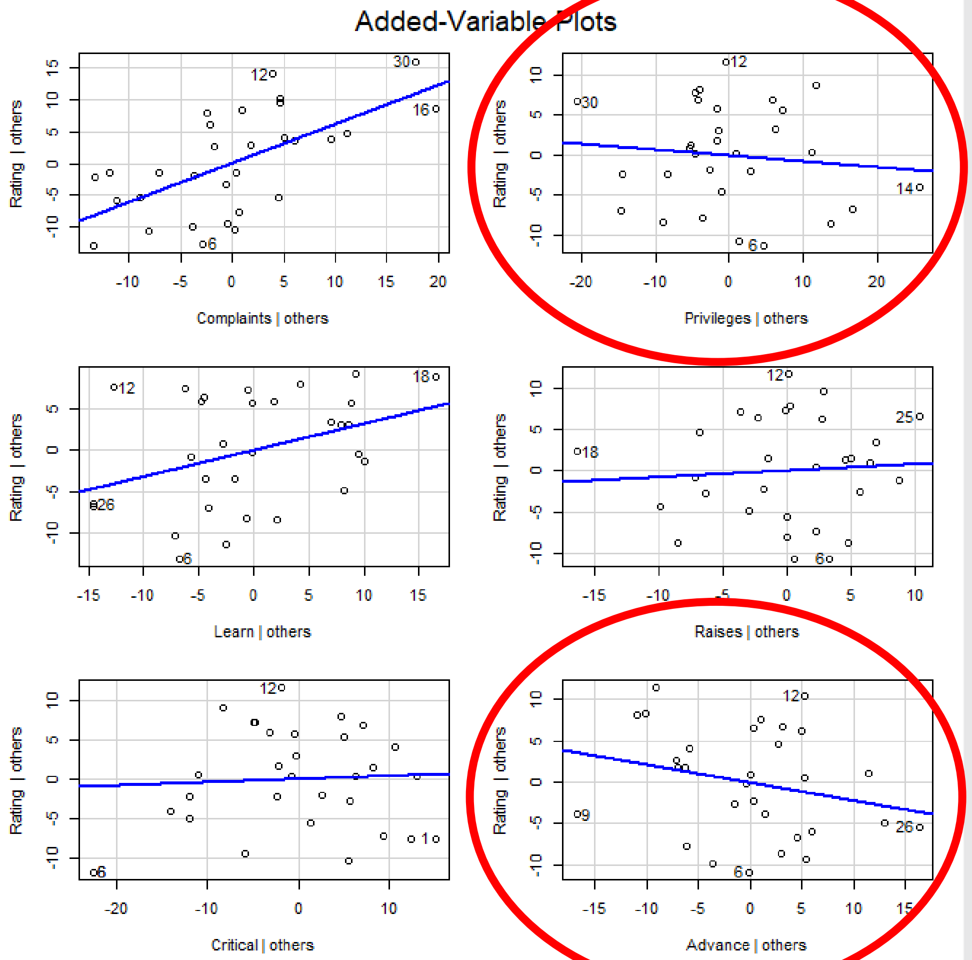
- Each of the plots should be fairly linear (or cloud-shaped)
- The first plot shows the relationship between Rating and Complaints after removing the linear effects of the other predictors

Complaints	$\hat{\beta}_1 =$	0.6132
Privileges	$\hat{\beta}_2 =$	-0.0731
Learn	$\hat{\beta}_3 =$	0.3203
Raises	$\hat{\beta}_4 =$	0.0817
Critical	$\hat{\beta}_5 =$	0.0384
Advance	$\hat{\beta}_6 =$	-0.2171



Partial Regression Plots

- Each of the plots should be fairly linear (or cloud-shaped)
- The first plot shows the relationship between Rating and Complaints after removing the linear effects of the other predictors



Complaints	$\hat{\beta}_1 =$	0.6132
Privileges	$\hat{\beta}_2 =$	-0.0731
Learn	$\hat{\beta}_3 =$	0.3203
Raises	$\hat{\beta}_4 =$	0.0817
Critical	$\hat{\beta}_5 =$	0.0384
Advance	$\hat{\beta}_6 =$	-0.2171

Partial Regression Plots

- A partial regression plot for a specific predictor has a slope that is the same as the multiple regression coefficient for that predictor.

- Partial regression plots attempt to show the effect of adding an additional variable to the model (given that one or more predictors are already in the model).

- For a specific predictor (Complaints), we take whatever is leftover after the fit of the other variables and look to see if what remains is correlated with the part of the predictor (Complaints) that is not already accounted for by the other variables.

Partial Regression Plots

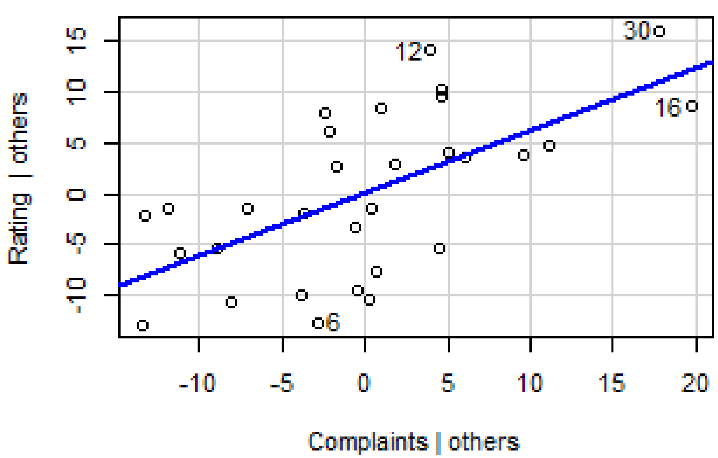
- Partial regression plots plot Y-residuals against X-residuals
- For the plot for Complaints:

1. The Y-axis is the residuals from the following model

$$\widehat{\text{Rating}}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Privileges}_i + \hat{\beta}_2 \times \text{Learn}_i + \hat{\beta}_3 \times \text{Raises}_i + \hat{\beta}_4 \times \text{Critical}_i + \hat{\beta}_5 \times \text{Advance}_i$$

2. The X-axis is the residuals from the following model

$$\widehat{\text{Complaints}}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Privileges}_i + \hat{\beta}_2 \times \text{Learn}_i + \hat{\beta}_3 \times \text{Raises}_i + \hat{\beta}_4 \times \text{Critical}_i + \hat{\beta}_5 \times \text{Advance}_i$$



We take whatever is leftover after fitting a model with the other variables (Y-axis) and look to see if what remains is correlated with the part of Complaints that is not already accounted for by the other variables (X-axis).

Code!

Multicollinearity

Multicollinearity

- Assumptions 6 and 7 go hand-in-hand: all important predictors are included and there is no multicollinearity

- **Underfitting**: not including important covariates (Xs) in the model that affect the response
  - Coefficient estimates and predictions are biased (bad)

- **Overfitting**: including covariates (Xs) in the model that don't affect the response
  - Coefficient estimates are still unbiased (good)
  - Standard error estimates are inflated due to collinearity (bad)

- **Multicollinearity**: When two or more predictor variables are highly correlated with each other

Multicollinearity

- In the case where one variable is a *perfect* linear combination of other variables, then the we cannot compute our estimates of the βs.
  - $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\mathbf{X}'\mathbf{Y}$, and $\boldsymbol{X}'\boldsymbol{X}$ would be singular (non-invertible).
- In other cases, one variable may not be a *perfect* linear combination of other variables, but there is still a high degree of correlation. In these cases, our estimates of the βs can still be computed, but there can be consequences.
  - Standard errors can be inflated making "significance" hard to detect even for useful variables
    - Significance tests on individual predictors could contradict the model F-test (F-test could say at least one of the Xs has some effect, while none of the individual predictors are significant)
  - Estimates can even have the wrong signs – completely opposite direction of effect

Multicollinearity Diagnostics

- Scatterplot Matrix (useful if collinearity involves only two predictors)
  - Look for two *predictors* whose scatterplot is strongly linearly correlated

- Correlation Matrix (useful if collinearity involves only two predictors)
  - *Rough* rule of thumb: coefficients should be less than 0.80

- Variance Inflation Factors (VIF)

Variance Inflation Factors

- Variance Inflation Factors (VIF) $R_k^2 = 0$ if no correlation
  - Let R_k^2 be the R^2 value when predictor X_k is regressed on the other predictors
  - $VIF\left(\hat{\beta}_k\right) = \frac{1}{1-R_k^2}$, for $k = 1, \ldots, p-1$
  - *A rough rule of thumb* is that multicollinearity is problematic if at least one of these occur:
    - The largest VIF is much more than 10
    - The mean VIF is much more than 1

	VIF
Complaints	2.67
Privileges	1.60
Learn	2.27
Raises	3.08
Critical	1.23
Advance	1.95
Mean	2.13

$1 \le VIF \le \infty$

$VIF = 1$ means no correlation

$VIF = 1.60$ means variance of Privileges is 60% bigger than what you would expect if there was no multicollinearity.

Code!

How Do We Deal with Multicollinearity?

1. Remove some variables from the model – choose a subset of predictor variables

2. Apply a shrinkage method (ridge regression, lasso, elastic net, etc.)

3. Combine the correlated variables somehow, like with principal component regression

Ultimately, we want model *parsimony*: we want the simplest model that will get the job done

Models Notes

Model

General/Theoretical Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i \,, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Fitted model (computed from data)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_{p-1} x_{i,p-1}$$

Notes:

- p represents the number of βs in the model (including the intercept)
- $x_{i,k}$ represents the value of variable X_k for observation i