Data cleaning

Regression assumptions

Influential observations

Multicollinearity

Data cleaning

Regression assumptions

Influential observations

Multicollinearity

"90% of statistics is data cleaning."

— Name missing

What to look for

- Look for cases with out-of-range values that may indicate errors in the data

What to look for

- Look for cases with out-of-range values that may indicate errors in the data
    - When data is clearly in error and it's not clear how to fix it, usually best to just omit those cases

What to look for

- Look for cases with out-of-range values that may indicate errors in the data
    - When data is clearly in error and it's not clear how to fix it, usually best to just omit those cases
- Look for cases with missing data

What to look for

- Look for cases with out-of-range values that may indicate errors in the data
    - When data is clearly in error and it's not clear how to fix it, usually best to just omit those cases
- Look for cases with missing data
    - Can simply omit cases when data is missing in the variable(s) you plan to use—but then you lose all data in the other variables too!

What to look for

- Look for cases with out-of-range values that may indicate errors in the data
  - When data is clearly in error and it's not clear how to fix it, usually best to just omit those cases
- Look for cases with missing data
  - Can simply omit cases when data is missing in the variable(s) you plan to use—but then you lose all data in the other variables too!
  - Can try to impute values for missing data by running a regression to predict the missing values from the other variables

What to look for

- Look for cases with out-of-range values that may indicate errors in the data
    - When data is clearly in error and it's not clear how to fix it, usually best to just omit those cases
- Look for cases with missing data
    - Can simply omit cases when data is missing in the variable(s) you plan to use—but then you lose all data in the other variables too!
    - Can try to impute values for missing data by running a regression to predict the missing values from the other variables
- If cases are not Missing Completely At Random (MCAR), then omitting cases with missing data may bias your conclusions

What to look for

- Look for cases with out-of-range values that may indicate errors in the data
    - When data is clearly in error and it's not clear how to fix it, usually best to just omit those cases
- Look for cases with missing data
    - Can simply omit cases when data is missing in the variable(s) you plan to use—but then you lose all data in the other variables too!
    - Can try to impute values for missing data by running a regression to predict the missing values from the other variables
- If cases are not Missing Completely At Random (MCAR), then omitting cases with missing data may bias your conclusions
- For categorical data with missing variables, can simply recode missing values as a special "unknown" category

What to look for

- Look for cases with out-of-range values that may indicate errors in the data
  - When data is clearly in error and it's not clear how to fix it, usually best to just omit those cases
- Look for cases with missing data
  - Can simply omit cases when data is missing in the variable(s) you plan to use—but then you lose all data in the other variables too!
  - Can try to impute values for missing data by running a regression to predict the missing values from the other variables
- If cases are not Missing Completely At Random (MCAR), then omitting cases with missing data may bias your conclusions
- For categorical data with missing variables, can simply recode missing values as a special "unknown" category
- No easy solutions!

We'll use the college data set to predict a school's graduation rate from various factors.

Take a look at the data; what potential issues do you see? We'll be using the following variables:

- Graduation rate
- Acceptance rate
- SAT score variables
- In-state tuition
- Out-of-state tuition

Many colleges have no SAT scores reported, so let's ignore those colleges (to enable a fair comparison) and also remove colleges with an obviously incorrect graduation rate of $> 100\%$:

```
my.sample <- colleges %>%
  filter(!is.na(Average.combined.SAT) &
           Graduation.rate <= 100)
```

Data cleaning

Regression assumptions

Influential observations

Multicollinearity

```
model <- lm(Graduation.rate ~ Average.combined.SAT + In.state.tuition,
            data=my.sample)
summary(model)
```

```
Call:
lm(formula = Graduation.rate ~ Average.combined.SAT + In.state.tuition,
    data = my.sample)

Residuals:
    Min      1Q  Median      3Q     Max
-45.526  -9.182   0.051   8.704  43.661

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -8.3246456  4.3708279  -1.905   0.0572 .
Average.combined.SAT  0.0611221  0.0048878  12.505   <2e-16 ***
In.state.tuition      0.0012486  0.0001111  11.237   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.75 on 709 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.4469,Adjusted R-squared:  0.4453
F-statistic: 286.4 on 2 and 709 DF,  p-value: < 2.2e-16
```

Multiple regression assumptions

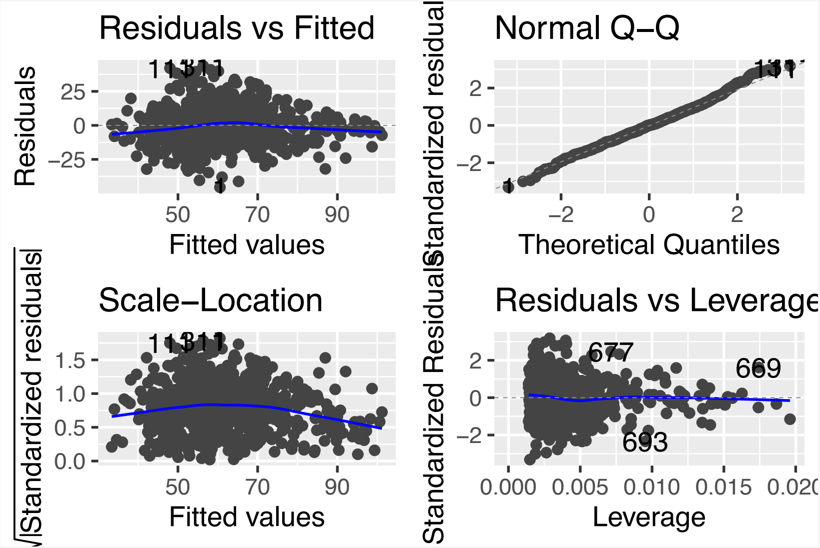
We need four things to be true for statistical inference (i.e., hypothesis tests, *p*-values, confidence intervals) to work for multiple regression:

Multiple regression assumptions

We need four things to be true for statistical inference (i.e., hypothesis tests, *p*-values, confidence intervals) to work for multiple regression:

1. The errors are independent.
2. Y is a linear function of the X's (except for the errors).
3. The errors are normally distributed.
4. The variance of Y is the same for any value of X ("homoscedasticity").

```r
library(ggfortify)
autoplot(model)
```



Assumption 1: Independence of errors

- Independence means that knowing the error (over-/under-prediction by the regression line) for one case doesn't tell you anything about the error for another case

Assumption 1: Independence of errors

- Independence means that knowing the error (over-/under-prediction by the regression line) for one case doesn't tell you anything about the error for another case
- Independence is usually most problematic with time series data (i.e., $X = \text{time}$)

Assumption 1: Independence of errors

- Independence means that knowing the error (over-/under-prediction by the regression line) for one case doesn't tell you anything about the error for another case
- Independence is usually most problematic with time series data (i.e., $X = \text{time}$)
- Since each college is completely separate, there is no reason to think the errors are not independent

Assumption 1: Independence of errors

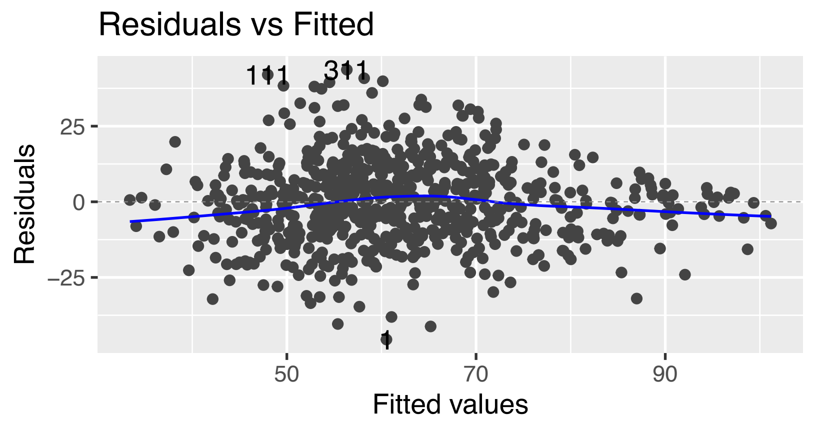
- Independence means that knowing the error (over-/under-prediction by the regression line) for one case doesn't tell you anything about the error for another case
- Independence is usually most problematic with time series data (i.e., $X = \text{time}$)
- Since each college is completely separate, there is no reason to think the errors are not independent
- However, we could have a violation of independence if e.g. all of the colleges in Texas (say) all implemented the same policies to try to improve graduation rates

Multiple regression assumptions

1. The errors are independent. ✓
2. Y is a linear function of the X's (except for the errors).
3. The errors are normally distributed.
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 2: Linearity

Look at the residual plot—there should be no trend (the blue line should be roughly horizontal):

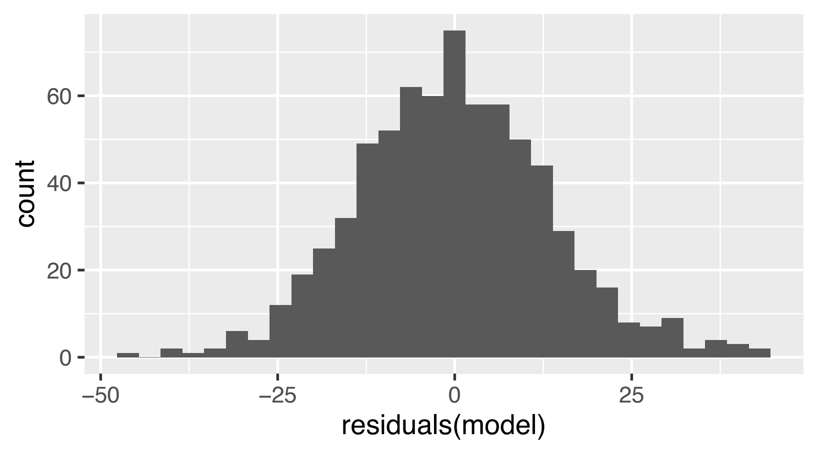


Multiple regression assumptions

1. The errors are independent. ✓
2. Y is a linear function of the X's (except for the errors). ✓
3. The errors are normally distributed.
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 3: Normality of residuals

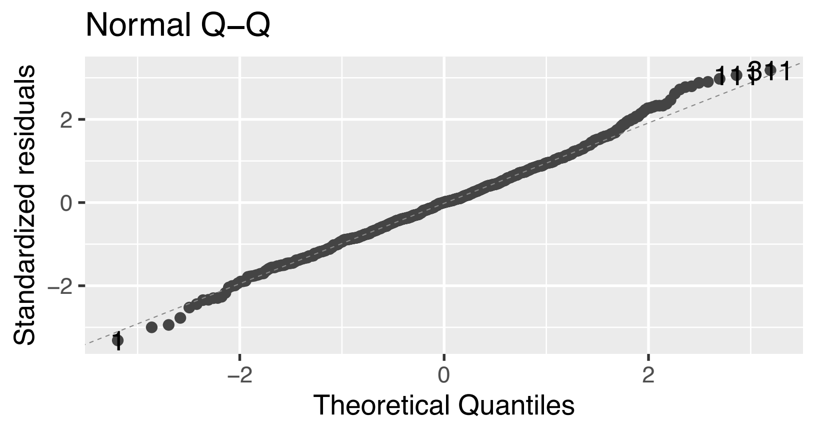
We need the histograms to be Normally distributed:



But it's hard to tell from a histogram!

Assumption 3: Normality of residuals

To be more careful we can use a Q-Q plot (a straight line indicates normality):

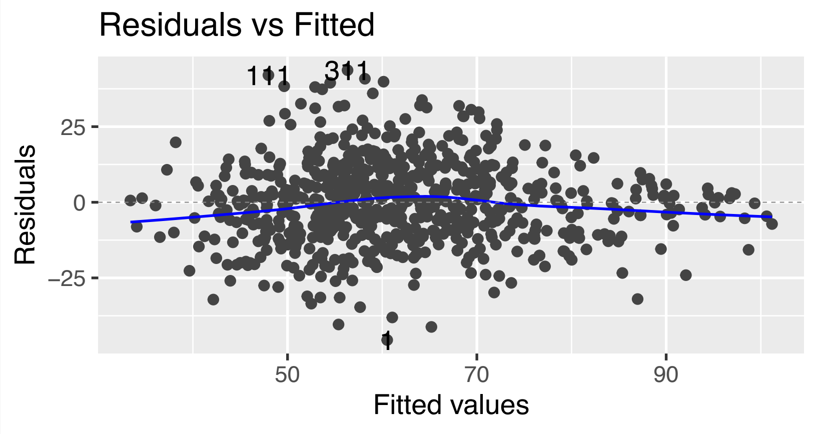


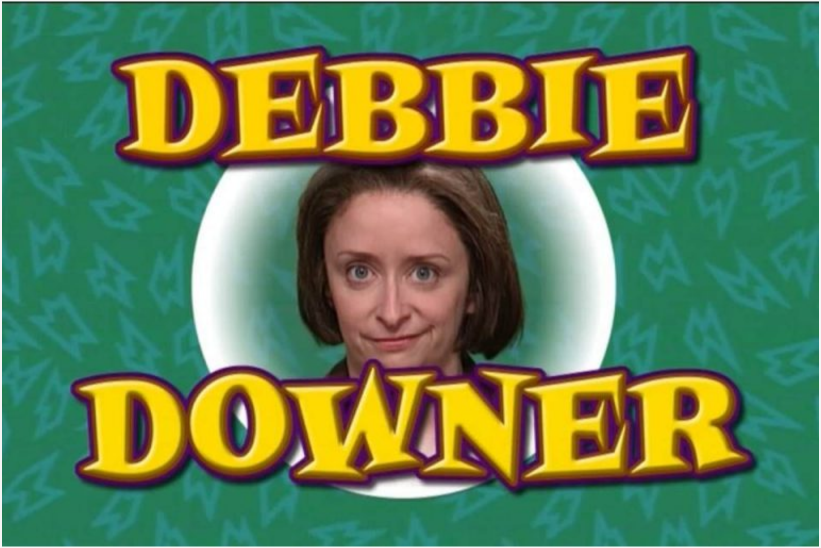
Multiple regression assumptions

1. The errors are independent. ✓
2. Y is a linear function of the X's (except for the errors). ✓
3. The errors are normally distributed. ✓
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 4: Homoscedasticity

Look at the residual plot—we want a roughly constant vertical "thickness":




DEBBIE
DOWNER

Multiple regression assumptions

1. The errors are independent. ✓
2. Y is a linear function of the X's (except for the errors). ✓
3. The errors are normally distributed. ✓
4. The variance of Y is the same for any value of X ("homoscedasticity"). ¯_(ツ)_/¯

What do I do if assumptions are violated?

- If any assumption is not satisfied, we should not trust the *p*-values or confidence intervals that come out of the model

What do I do if assumptions are violated?

- If any assumption is not satisfied, we should not trust the p-values or confidence intervals that come out of the model
- However, as long as linearity is satisfied, we can still use the model for making predictions (we just can't put reliable CIs on those predictions)

What do I do if assumptions are violated?

- If any assumption is not satisfied, we should not trust the *p*-values or confidence intervals that come out of the model
- However, as long as linearity is satisfied, we can still use the model for making predictions (we just can't put reliable CIs on those predictions)
- Is there anything else we can do? (Yes—stay tuned for next week!)

Data cleaning

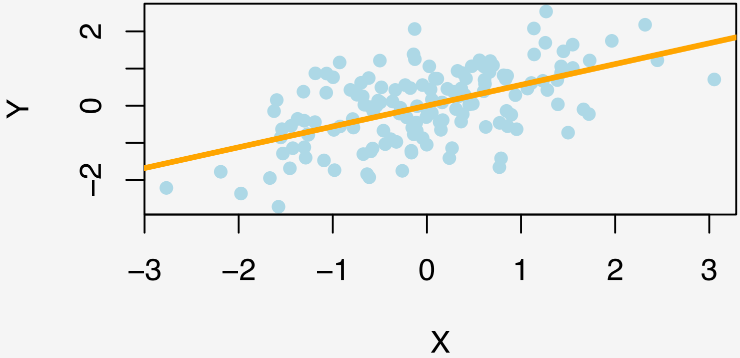
Regression assumptions

Influential observations

Multicollinearity

What a single case can do

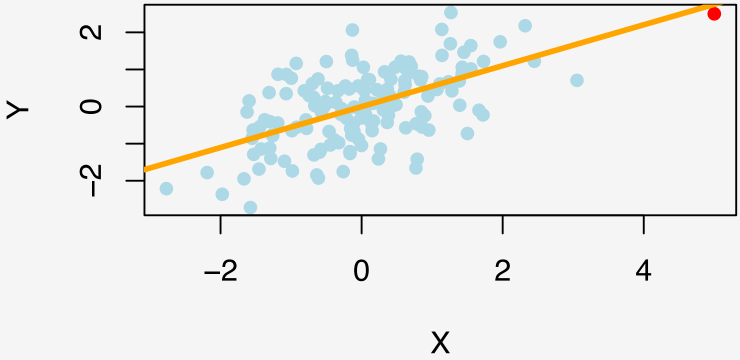
Let's take some hypothetical sample data:



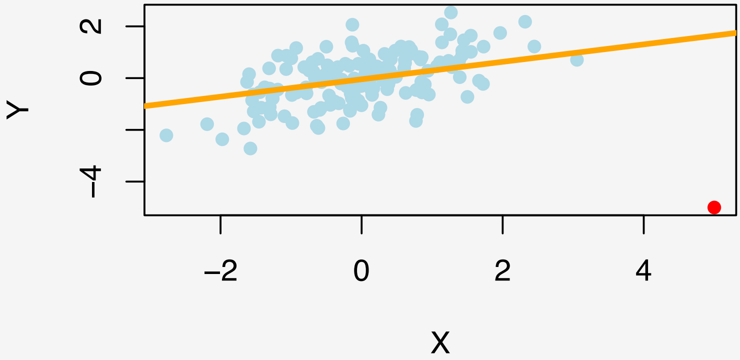
What a single case can do

Even a single case can wreak havoc on the regression line. Let's add one outlier, at $X = 5$, and see what happens with different Y values.

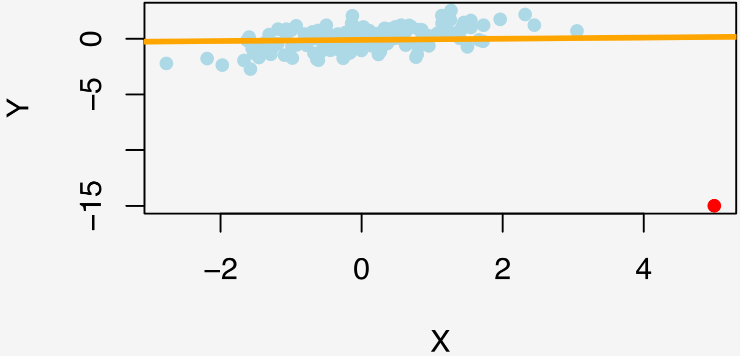
What a single case can do



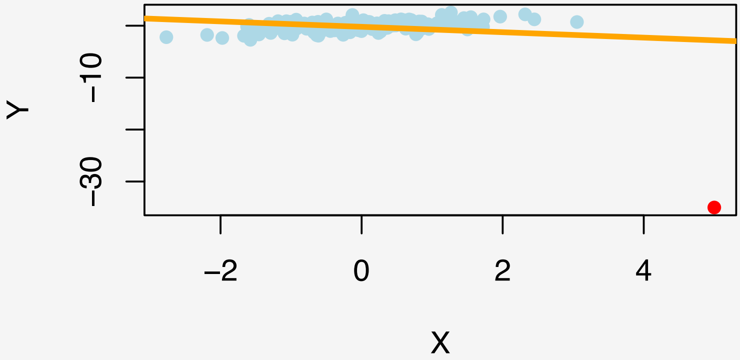
What a single case can do



What a single case can do



What a single case can do




BLACKMAIL

Regression is like blackmail

Blackmail:

- Compromising information gives a blackmailer leverage—the *potential* to have a big impact
- Once the blackmailer uses the information, that gives them influence

Regression is like blackmail

Blackmail:

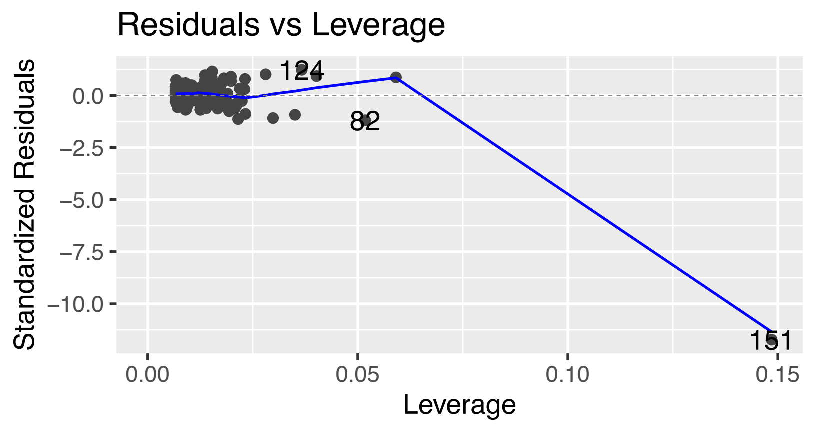
- Compromising information gives a blackmailer leverage—the *potential* to have a big impact
- Once the blackmailer uses the information, that gives them influence

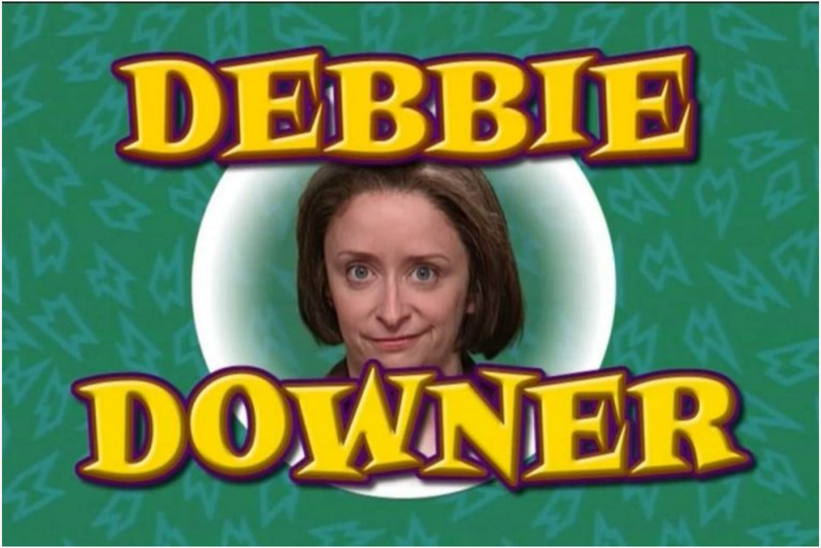
Regression:

- When a case has a very unusual X value (i.e., far from $\overline{X}$), it has leverage—the *potential* to have a big impact on the regression line
- When that case *also* has a Y value that is out of line with the general trend, it will pull the regression line towards it—giving it influence

How do I know what cases are influential?

Look for cases with high leverage and a large (positive or negative) residual:




DEBBIE
DOWNER

What can we do about influential observations?

- If removing the influential observation doesn't make a substantial difference in your analysis, don't worry about it

What can we do about influential observations?

- If removing the influential observation doesn't make a substantial difference in your analysis, don't worry about it
- If it does:

What can we do about influential observations?

- If removing the influential observation doesn't make a substantial difference in your analysis, don't worry about it
- If it does:
  - Consider whether it could be a mistake (happens more than you might think!); if it is, correct the error or drop the case

What can we do about influential observations?

- If removing the influential observation doesn't make a substantial difference in your analysis, don't worry about it
- If it does:
  - Consider whether it could be a mistake (happens more than you might think!); if it is, correct the error or drop the case
  - If not, hold out the influential observation(s) and report on them separately

What can we do about influential observations?

- If removing the influential observation doesn't make a substantial difference in your analysis, don't worry about it
- If it does:
  - Consider whether it could be a mistake (happens more than you might think!); if it is, correct the error or drop the case
  - If not, hold out the influential observation(s) and report on them separately
  - Do not just throw out and ignore influential observations!

Data cleaning

Regression assumptions

Influential observations

Multicollinearity

Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors X_1 and X_2 are highly correlated, it is hard to estimate the effect of changing X_1 while keeping X_2 constant.

Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors X_1 and X_2 are highly correlated, it is hard to estimate the effect of changing X_1 while keeping X_2 constant.
- This means we will have large standard errors, and large p-values, for X_1 and/or X_2.

Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors X_1 and X_2 are highly correlated, it is hard to estimate the effect of changing X_1 while keeping X_2 constant.
- This means we will have large standard errors, and large p-values, for X_1 and/or X_2.
- This does not mean there isn't a relationship between X_1 and Y, or X_2 and Y – it just means we can't pin down that relationship, because of the correlation!

Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors X_1 and X_2 are highly correlated, it is hard to estimate the effect of changing X_1 while keeping X_2 constant.
- This means we will have large standard errors, and large p-values, for X_1 and/or X_2.
- This does not mean there isn't a relationship between X_1 and Y, or X_2 and Y – it just means we can't pin down that relationship, because of the correlation!

Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors X_1 and X_2 are highly correlated, it is hard to estimate the effect of changing X_1 while keeping X_2 constant.
- This means we will have large standard errors, and large p-values, for X_1 and/or X_2.
- This does not mean there isn't a relationship between X_1 and Y, or X_2 and Y – it just means we can't pin down that relationship, because of the correlation!

Correlation between the response and the predictors is good, but correlation between the predictors is not!

The effect of multicollinearity

We want to avoid multicollinearity in our models (if we can)!

The effect of multicollinearity

We want to avoid multicollinearity in our models (if we can)!

- Any conclusions based on the p-values, coefficients, and confidence intervals of the highly correlated variables will be unreliable.

The effect of multicollinearity

We want to avoid multicollinearity in our models (if we can)!

- Any conclusions based on the p-values, coefficients, and confidence intervals of the highly correlated variables will be unreliable.
- These statistics will not be stable: adding new data or predictors to the model could drastically change them.

How can we detect multicollinearity?

One way to see if two variables are collinear is to check the correlation between the two:

```
cor(my.sample$Average.math.SAT,
    my.sample$Average.verbal.SAT,
    use="complete.obs")

[1] 0.9194207
```

Any large correlation is potentially problematic.

How can we detect multicollinearity?

One way to see if two variables are collinear is to check the correlation between the two:

```
cor(my.sample$Average.math.SAT,
    my.sample$Average.verbal.SAT,
    use="complete.obs")
```

```
[1] 0.9194207
```

Any large correlation is potentially problematic. But what if there is multicollinearity among 3+ predictors?

How can we detect multicollinearity?

A better way to check multicollinearity is using Variance Inflation Factors (VIF).

- The VIF is

$$\mathrm{VIF}(\beta_j) = \frac{1}{1 - R_j^2},$$

where R_j^2 is the R^2 in a regression predicting X variable j from the other X variables.

How can we detect multicollinearity?

A better way to check multicollinearity is using Variance Inflation Factors (VIF).

- The VIF is

$$\mathrm{VIF}(\beta_j) = \frac{1}{1 - R_j^2},$$

  where R_j^2 is the R^2 in a regression predicting X variable j from the other X variables.
- $\mathrm{VIF}(\beta_j) = 0$ when $R_j^2 = 0$; i.e., the jth predictor variable is completely independent from the others.

How can we detect multicollinearity?

A better way to check multicollinearity is using Variance Inflation Factors (VIF).

- The VIF is
$$\mathrm{VIF}(\beta_j) = \frac{1}{1 - R_j^2},$$
where R_j^2 is the R^2 in a regression predicting X variable j from the other X variables.
- $\mathrm{VIF}(\beta_j) = 0$ when $R_j^2 = 0$; i.e., the jth predictor variable is completely independent from the others.
- $\mathrm{VIF}(\beta_j)$ increases as R_j^2 does, and is ∞ when there is perfect multicollinearity; i.e., when X_j is perfectly predictable from the other X variables.

How can we detect multicollinearity?

A better way to check multicollinearity is using Variance Inflation Factors (VIF).

- The VIF is
$$\mathrm{VIF}(\beta_j) = \frac{1}{1 - R_j^2},$$
where R_j^2 is the R^2 in a regression predicting X variable j from the other X variables.
- $\mathrm{VIF}(\beta_j) = 0$ when $R_j^2 = 0$; i.e., the jth predictor variable is completely independent from the others.
- $\mathrm{VIF}(\beta_j)$ increases as R_j^2 does, and is ∞ when there is perfect multicollinearity; i.e., when X_j is perfectly predictable from the other X variables.
- The VIF measures the increase in standard error between a simple regression with the variable in question and the multiple regression under consideration

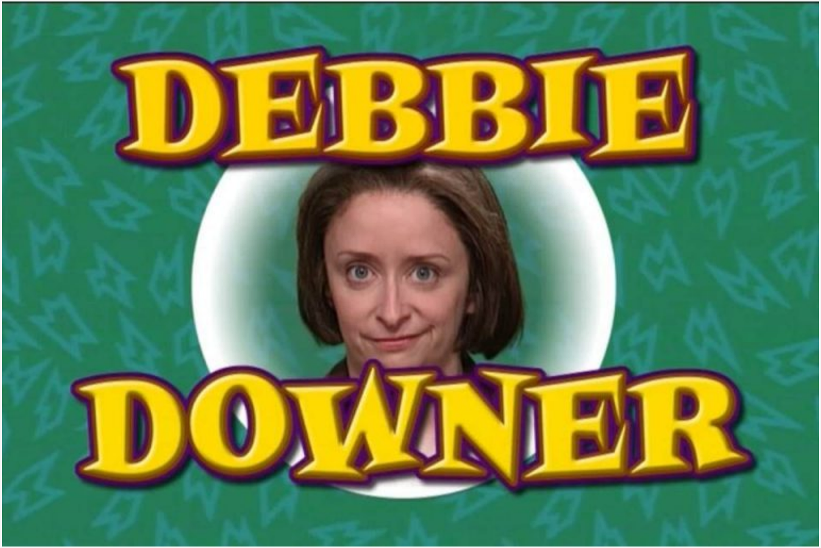
How can we detect multicollinearity?

To calculate VIF for each predictor, you would have to run one regression for each predictor. But you don't have to run all of these regressions by hand!

```
model <- lm(Graduation.rate ~ Average.math.SAT +
              Average.verbal.SAT + Acceptance.rate,
              data=my.sample)
library(car)
vif(model)

  Average.math.SAT Average.verbal.SAT    Acceptance.rate
          6.647073           6.454614           1.224408
```

Predictors with VIF > 5 are a cause for concern.


DEBBIE
DOWNER

Dealing with multicollinearity

There are two general strategies for dealing with multicollinearity:

- Drop one of the variables with a high VIF factor, and rerun to see if VIFs have improved. (Just like we drop one of the dummy variables when putting a categorical variable in the model!)

And a third option – Accept it!

Dealing with multicollinearity

There are two general strategies for dealing with multicollinearity:

- Drop one of the variables with a high VIF factor, and rerun to see if VIFs have improved. (Just like we drop one of the dummy variables when putting a categorical variable in the model!)
- Combine the variables that correlate into a composite variable. (Combined SAT score = Math + Verbal)

And a third option – Accept it!

```r
summary(lm(Graduation.rate ~ Average.combined.SAT + Acceptance.rate,
           data=my.sample))
```

```
Call:
lm(formula = Graduation.rate ~ Average.combined.SAT + Acceptance.rate,
    data = my.sample)

Residuals:
    Min      1Q  Median      3Q     Max
-46.962  -9.619  -0.668   8.877  47.632

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -27.632646   6.619264  -4.175 3.35e-05 ***
Average.combined.SAT   0.090296   0.004967  18.179  < 2e-16 ***
Acceptance.rate        0.024146   0.039224   0.616    0.538
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.91 on 725 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.3507,Adjusted R-squared:  0.3489
F-statistic: 195.8 on 2 and 725 DF,  p-value: < 2.2e-16
```

Multicollinearity and uncertainty

When is collinearity *not* an issue?

- When there is high collinearity in X's that are strictly for adjustment, not interpretation, and the coefficient we want to interpret corresponds to a variable with low multicollinearity
- When multicollinearity comes from how we construct X – e.g., adding polynomial terms (next week!), turning categories into dummy variables, or adding interactions.
- When we are just trying to predict using relatively few X's, and only predicting at "typical" X values.