THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Polynomial regression and data cleaning

Lecture 16

STA 371G

1. Model selection

2. Polynomial regression

3. Data cleaning

R^2

- R^2 has a similar meaning as in simple regression: how much of the variation in the response variable (Y) are explained by the predictor variables (X's) together?

R^2

- R^2 has a similar meaning as in simple regression: how much of the variation in the response variable (Y) are explained by the predictor variables (X's) together?
- Another way to think about R^2 is that

$$R^2 = \frac{\mathrm{Var}(\hat{Y})}{\mathrm{Var}(Y)},$$

i.e., it represents how much variance in Y the model predicts.

R^2

- R^2 has a similar meaning as in simple regression: how much of the variation in the response variable (Y) are explained by the predictor variables (X's) together?
- Another way to think about R^2 is that

$$R^2 = \frac{\mathrm{Var}(\hat{Y})}{\mathrm{Var}(Y)},$$

  i.e., it represents how much variance in Y the model predicts.
- R^2 always increases when you add more variables, even if you add variables that have no real relationship with Y.

```r
my.sample <- subset(colleges,
  !is.na(Average.combined.SAT) & Graduation.rate <= 100)
model1 <- lm(Graduation.rate ~ Average.combined.SAT + In.state.tuition,
             data=my.sample)
summary(model1)
```

```

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

Residuals:
      Min        1Q    Median        3Q       Max
-45.52572  -9.18156   0.05085   8.70420  43.66097

Coefficients:
                         Estimate   Std. Error  t value Pr(>|t|)
(Intercept)          -8.324645625  4.370827909 -1.90459 0.057238 .
Average.combined.SAT  0.061122082  0.004887825 12.50496  < 2e-16 ***
In.state.tuition      0.001248638  0.000111119 11.23692  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.7466 on 709 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.44686,Adjusted R-squared:   0.4453
F-statistic: 286.387 on 2 and 709 DF,  p-value: < 2.22e-16
```

```r
Random.numbers <- rnorm(nrow(my.sample))
model2 <- lm(Graduation.rate ~ Average.combined.SAT + In.state.tuition
               + Random.numbers, data=my.sample)
summary(model2)
```

```

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

Residuals:
      Min         1Q    Median        3Q       Max
-45.59477  -9.13473   0.06836   8.75583  43.74968

Coefficients:
                        Estimate   Std. Error  t value Pr(>|t|)
(Intercept)          -8.433559857  4.378188630 -1.92627 0.054471 .
Average.combined.SAT  0.061244215  0.004896088 12.50881  < 2e-16 ***
In.state.tuition      0.001248531  0.000111177 11.23012  < 2e-16 ***
Random.numbers        0.277098090  0.537299499  0.51572 0.606208
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.7537 on 708 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.447068,Adjusted R-squared:  0.444725
F-statistic: 190.816 on 3 and 708 DF,  p-value: < 2.22e-16
```

```r
Random.numbers <- rnorm(nrow(my.sample))
model2 <- lm(Graduation.rate ~ Average.combined.SAT + In.state.tuition
               + Average.math.SAT, data=my.sample)
summary(model2)
```

```

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

Residuals:
      Min        1Q    Median        3Q       Max
-45.27189  -9.06503   0.03009   8.64981  43.89591

Coefficients:
                        Estimate   Std. Error  t value Pr(>|t|)
(Intercept)          -8.144252350  4.434188943 -1.83669 0.066675 .
Average.combined.SAT  0.054229967  0.023519872  2.30571 0.021416 *
In.state.tuition      0.001256312  0.000115918 10.83790  < 2e-16 ***
Average.math.SAT      0.012667133  0.041953872  0.30193 0.762794
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.7492 on 706 degrees of freedom
  (21 observations deleted due to missingness)
Multiple R-squared:  0.447693,Adjusted R-squared:  0.445346
F-statistic: 190.758 on 3 and 706 DF,  p-value: < 2.22e-16
```

Adjusted R^2

- There are many, many possible models (think of how many combinations of predictors there are!) so we need some criterion to determine which model is best.

Adjusted R^2

- There are many, many possible models (think of how many combinations of predictors there are!) so we need some criterion to determine which model is best.
- R^2 is not good because adding even a variable of random numbers increases R^2.

Adjusted R^2

- There are many, many possible models (think of how many combinations of predictors there are!) so we need some criterion to determine which model is best.
- R^2 is not good because adding even a variable of random numbers increases R^2.
- Adjusted R^2 makes an adjustment to R^2 by adding a penalty for each variable added (in this example, adjusted R^2 went down even though R^2 increased).

```r
model1 <- lm(Graduation.rate ~ Average.combined.SAT + In.state.tuition,
             data=my.sample)
summary(model1)
```

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

Residuals:
      Min        1Q    Median        3Q       Max
-45.52572  -9.18156   0.05085   8.70420  43.66097

Coefficients:
                        Estimate   Std. Error  t value Pr(>|t|)
(Intercept)          -8.324645625  4.370827909 -1.90459 0.057238 .
Average.combined.SAT  0.061122082  0.004887825 12.50496  < 2e-16 ***
In.state.tuition      0.001248638  0.000111119 11.23692  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.7466 on 709 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.44686,Adjusted R-squared:   0.4453
F-statistic: 286.387 on 2 and 709 DF,  p-value: < 2.22e-16
```

```r
Random.numbers <- rnorm(nrow(my.sample))
model2 <- lm(Graduation.rate ~ Average.combined.SAT + In.state.tuition
               + Random.numbers, data=my.sample)
summary(model2)
```

```

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

Residuals:
      Min        1Q    Median        3Q       Max
-45.59477  -9.13473   0.06836   8.75583  43.74968

Coefficients:
                         Estimate   Std. Error  t value Pr(>|t|)
(Intercept)          -8.433559857  4.378188630 -1.92627 0.054471 .
Average.combined.SAT  0.061244215  0.004896088 12.50881  < 2e-16 ***
In.state.tuition      0.001248531  0.000111177 11.23012  < 2e-16 ***
Random.numbers        0.277098090  0.537299499  0.51572 0.606208
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.7537 on 708 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.447068,Adjusted R-squared:  0.444725
F-statistic: 190.816 on 3 and 708 DF,  p-value: < 2.22e-16
```

Which model is the best?

- In general, we want to select the model that is the most parsimonious, that is, the model that has the best combination of being simple with a high R^2.

Which model is the best?

- In general, we want to select the model that is the most parsimonious, that is, the model that has the best combination of being simple with a high R^2.
- This is easier said than done—using Adjusted R^2 is not enough. We'll come back to this next week!

1. Model selection

2. Polynomial regression

3. Data cleaning

Managing Yourself

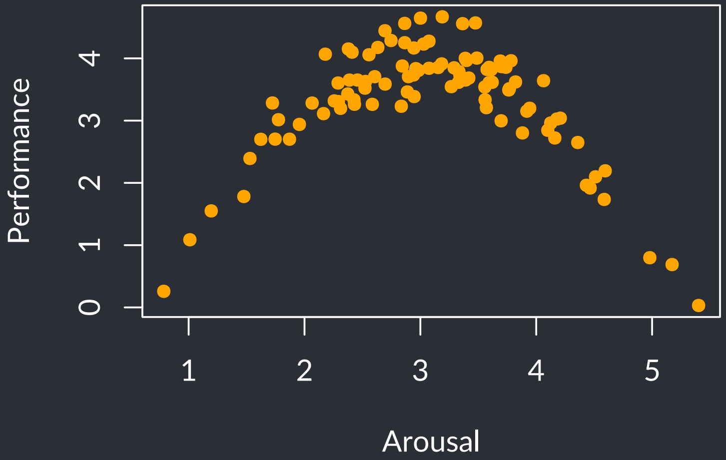
Are You Too Stressed to Be Productive? Or Not Stressed Enough?

by Francesca Gino

April 14, 2016

If you're like me, you often ask yourself how you can get more work done in a day. How can you best boost your productivity? I always assumed that if I could just reduce any stress I was facing, my productivity would rise. But my intuition was, in fact, wrong. It's true that stress can be a health risk, and that we're often encouraged to avoid it if we want to live happy, productive, and long lives. But research suggests that some stress can actually be beneficial to performance.

Let's look at some simulated Yerkes-Dodson data:



The correlation is almost 0, but there is a pretty strong relationship here—it's just not linear!

Polynomial regression

- Polynomial regression allows us to fit a polynomial (like $4x^2 - 6x + 3$ or $x^3 - 2x^2 + x - 7$) to the data

Polynomial regression

- Polynomial regression allows us to fit a polynomial (like $4x^2 - 6x + 3$ or $x^3 - 2x^2 + x - 7$) to the data
- To do this, we create variables X^2, X^3, etc. (up to as high as we want) and add them to a simple regression model to create a multiple regression model

Polynomial regression

- Polynomial regression allows us to fit a polynomial (like $4x^2 - 6x + 3$ or $x^3 - 2x^2 + x - 7$) to the data
- To do this, we create variables X^2, X^3, etc. (up to as high as we want) and add them to a simple regression model to create a multiple regression model
- For example, to fit a parabola (quadratic polynomial) to this data, we would build a model where the explanatory variables are X and X^2

```r
model1 <- lm(Performance ~ Arousal)
summary(model1)
```

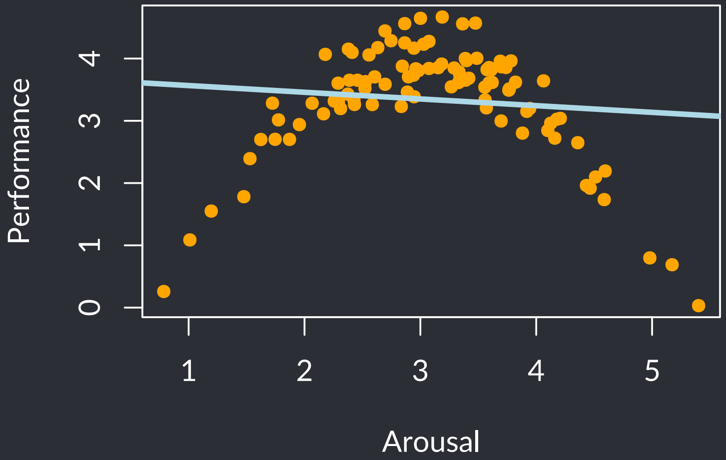
```

Call:
lm(formula = Performance ~ Arousal)

Residuals:
      Min        1Q    Median        3Q       Max 
-3.330631 -0.238013  0.238412  0.542176  1.336404 

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)    
(Intercept)  3.673979   0.329398 11.15363  < 2e-16 ***
Arousal     -0.107626   0.101830 -1.05692  0.29315    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.910048 on 98 degrees of freedom
Multiple R-squared:  0.0112704,Adjusted R-squared:  0.00118129 
F-statistic: 1.11709 on 1 and 98 DF,  p-value: 0.293145
```

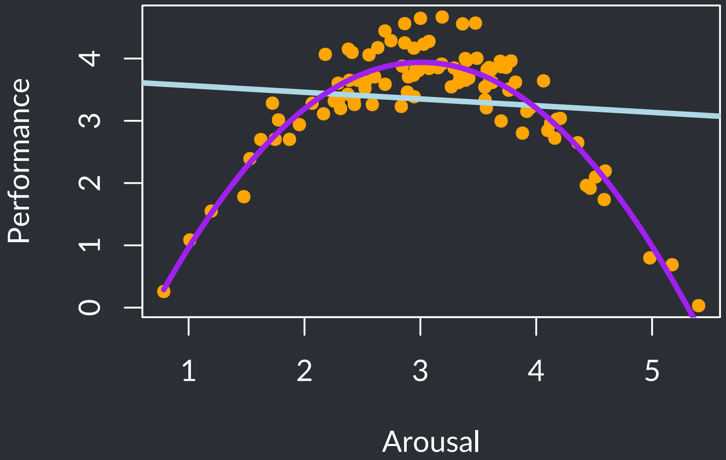
```r
Arousal2 <- Arousal^2
model2 <- lm(Performance ~ Arousal + Arousal2)
summary(model2)
```

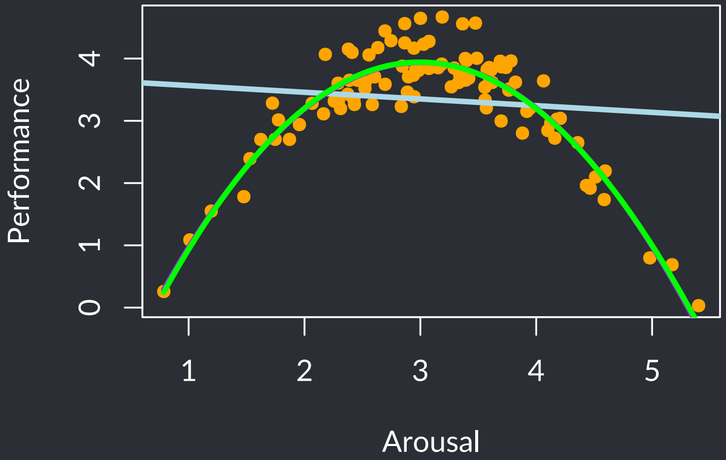
```

Call:
lm(formula = Performance ~ Arousal + Arousal2)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.687761 -0.218888 -0.045085  0.203100  0.794497 

Coefficients:
             Estimate Std. Error   t value   Pr(>|t|)    
(Intercept) -2.752770   0.284304  -9.68248  6.503e-16 ***
Arousal      4.455741   0.186369  23.90810 < 2.22e-16 ***
Arousal2    -0.741622   0.029668 -24.99739 < 2.22e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.335311 on 97 degrees of freedom
Multiple R-squared:  0.867141,Adjusted R-squared:  0.864402 
F-statistic: 316.549 on 2 and 97 DF,  p-value: < 2.22e-16
```

```r
Arousal3 <- Arousal^3
model3 <- lm(Performance ~ Arousal + Arousal2 + Arousal3)
summary(model3)
```

```

Call:
lm(formula = Performance ~ Arousal + Arousal2 + Arousal3)

Residuals:
       Min          1Q      Median          3Q         Max 
-0.6917765 -0.2205673 -0.0429428  0.1940644  0.7995118 

Coefficients:
              Estimate Std. Error  t value   Pr(>|t|)    
(Intercept) -2.8891429  0.5733027 -5.03947 2.1935e-06 ***
Arousal      4.6179937  0.6203317  7.44439 4.1776e-11 ***
Arousal2    -0.7988884  0.2108451 -3.78898 0.00026392 ***
Arousal3     0.0061735  0.0225016  0.27436 0.78439904    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.336921 on 96 degrees of freedom
Multiple R-squared:  0.867245,Adjusted R-squared:  0.863097 
F-statistic: 209.046 on 3 and 96 DF,  p-value: < 2.22e-16
```


Performance
0
1
2
3
4
1
2
3
4
5
Arousal


Performance
0
1
2
3
4
1
2
3
4
5
Arousal



Performance
0
1
2
3
4
1
2
3
4
5
Arousal

Important considerations with polynomial regression

- Use changes in Adjusted R^2 and the significant of the highest-order term to help you decide how many higher-order terms to add

Important considerations with polynomial regression

- Use changes in Adjusted R^2 and the significant of the highest-order term to help you decide how many higher-order terms to add
- If you include the term X^k (for any k), you should also include all lower-order terms, even if they are not significant

Important considerations with polynomial regression

- Use changes in Adjusted R^2 and the significant of the highest-order term to help you decide how many higher-order terms to add
- If you include the term X^k (for any k), you should also include all lower-order terms, even if they are not significant
- Be very careful with extrapolation when using models with polynomial terms!

1. Model selection
2. Polynomial regression
3. Data cleaning

Data set

We’re going to look at a data set of newly hired managers:

- Salary (response)
- Manager rating
- Years of experience
- Years since graduation
- Origin (internal or external hire)

Data issues

Data scientists report that they spend 70% of their time on obtaining and cleaning the data. Only 30% is for statistical analysis.

Data issues

Data scientists report that they spend 70% of their time on obtaining and cleaning the data. Only 30% is for statistical analysis.

Never run a regression without exploring and cleaning the data first!

Exploring the data: Outliers

Boxplots are commonly used to find cases that might be outliers.
Let's start by looking at the Salary column.

```
boxplot(manager$Salary, xlab="Salary", horizontal=T)
```

Exploring the data: outliers

If a case is shown as an outlier on the boxplot (i.e., 1.5 IQR above Q3 or 1.5 IQR below Q1):

- It might be an error.
- It might represent a missing value or other situation. (Consult the codebook for the data set, if there is one!)
- It might be a true outlier.

Exploring the data: Outliers

```
subset(manager, Salary > 200)

    Salary MngrRating YearsExp YrsSinceGrad   Origin
146    511        6.1        2            2 Internal

subset(manager, Salary < 0)

    Salary MngrRating YearsExp YrsSinceGrad   Origin
121    -66        5.7        1            2 Internal
```

Exploring the data: Outliers

```
subset(manager, Salary > 200)

    Salary MngrRating YearsExp YrsSinceGrad   Origin
146    511        6.1        2            2 Internal

subset(manager, Salary < 0)

    Salary MngrRating YearsExp YrsSinceGrad   Origin
121    -66        5.7        1            2 Internal
```

We can deal with outliers in two ways.

- If the result of **errors in the data**, we can try to correct or omit.

Exploring the data: Outliers

```
subset(manager, Salary > 200)

    Salary MngrRating YearsExp YrsSinceGrad   Origin
146    511        6.1        2            2 Internal

subset(manager, Salary < 0)

    Salary MngrRating YearsExp YrsSinceGrad   Origin
121    -66        5.7        1            2 Internal
```

We can deal with outliers in two ways.

- If the result of **errors in the data**, we can try to correct or omit.
- If not, consider omitting, but report on them separately.

Exploring the data: Outliers

Let's omit the outliers by creating a new data set `mclean` that consists of the subset of the data where the salary is between \$0 and \$200,000.

```
mclean <- subset(manager, Salary > 0 & Salary < 200)
```

We'll use `mclean` for our analysis, but we won't destroy the original data set!

Exploring the data: Outliers

```
boxplot(mclean$YearsExp, xlab="Years of Experience",
  horizontal=T)
```

Exploring the data: Outliers

99 must be a code for missing entry in the Years of Experience variable!

Exploring the data: Outliers

99 must be a code for missing entry in the Years of Experience variable!

Let’s label all 99s as `NA` (“not available” — R’s code for missing data).

Exploring the data: Outliers

99 must be a code for missing entry in the Years of Experience variable!

Let's label all 99s as NA ("not available" — R's code for missing data).

```
mclean$YearsExp[mclean$YearsExp == 99] <- NA
```

Exploring the data: Missing entries

Let's see if we have other missing data.

```
mclean[!complete.cases(mclean),]

    Salary MngrRating YearsExp YrsSinceGrad   Origin
103     75         NA        8            8 Internal
110     81         NA        9            9 External
124     73        5.9       NA            7 External
154     49        8.0        1            1     <NA>
```

Exploring the data: Missing entries

Let's see if we have other missing data.

```
mclean[!complete.cases(mclean),]

    Salary MngrRating YearsExp YrsSinceGrad   Origin
103     75         NA        8            8 Internal
110     81         NA        9            9 External
124     73        5.9       NA            7 External
154     49        8.0        1            1     <NA>
```

This isn't surprising—it is very common to have missing entries in your data. (The comma is needed so that we capture the full row.)

Exploring the data: Missing entries

There are two ways of dealing with missing data:

- Omit the rows that have missing entries in it.
- Try to predict values to fill the missing entries.

Omitting data is the easiest, but often not the best way, because you lose all the other information available in the same row.

Exploring the data: Missing entries

What should we replace the `NA`s in the Manager Rating and Years of Experience columns with?

Exploring the data: Missing entries

What should we replace the NAs in the Manager Rating and Years of Experience columns with?

The simplest way would be to use the averages in the respective columns.

```
mclean$MngrRating[is.na(mclean$MngrRating)] <-
  mean(mclean$MngrRating, na.rm=T)

mclean$YearsExp[is.na(mclean$YearsExp)] <-
  mean(mclean$YearsExp, na.rm=T)
```

Exploring the data: Missing entries

What should we replace the NAs in the Manager Rating and Years of Experience columns with?

The simplest way would be to use the averages in the respective columns.

```
mclean$MngrRating[is.na(mclean$MngrRating)] <-
  mean(mclean$MngrRating, na.rm=T)

mclean$YearsExp[is.na(mclean$YearsExp)] <-
  mean(mclean$YearsExp, na.rm=T)
```

A smarter and more advanced way is to predict the missing data from the other data (using regression!).

Exploring the data: Missing entries

What about the missing data for categorical variables?

Exploring the data: Missing entries

What about the missing data for categorical variables? Let's choose the easy way and omit them.

```
mclean <- na.omit(mclean)
```

Exploring the data: Missing entries

What about the missing data for categorical variables? Let's choose the easy way and omit them.

```
mclean <- na.omit(mclean)
```

This removes all the rows that contain missing entries (only the Origin column has missing entries in this case).

Exploring the data: Missing entries

What about the missing data for categorical variables? Let's choose the easy way and omit them.

```
mclean <- na.omit(mclean)
```

This removes all the rows that contain missing entries (only the Origin column has missing entries in this case).

We could also predict the missing entries, or treat the missing entries as a seperate level (e.g. "Unknown").

Exploring the data: Missing entries

- While dealing with the missing data, we assume that the data is "Missing Completely at Random" (MCAR).

Exploring the data: Missing entries

- While dealing with the missing data, we assume that the data is “Missing Completely at Random” (MCAR).
- If this assumption does not hold (e.g. if the missing data mostly belongs to external hires), the model will be biased.

Exploring the data: Missing entries

- While dealing with the missing data, we assume that the data is "Missing Completely at Random" (MCAR).
- If this assumption does not hold (e.g. if the missing data mostly belongs to external hires), the model will be biased.
- Making predictions for missing data based on available data reinforces the existing relationships between variables, so impacts the standard error.

Exploring the data: Missing entries

- While dealing with the missing data, we assume that the data is "Missing Completely at Random" (MCAR).
- If this assumption does not hold (e.g. if the missing data mostly belongs to external hires), the model will be biased.
- Making predictions for missing data based on available data reinforces the existing relationships between variables, so impacts the standard error.
- If a lot of data is missing (e.g. more than 5%) for a particular variable, you may have to discard the whole column.