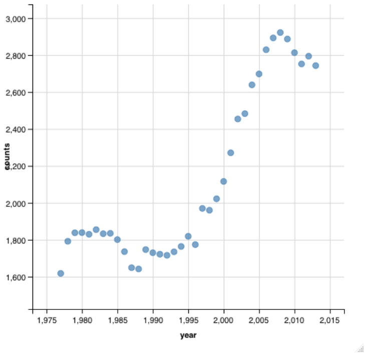
Assignment #12: Advanced Regression Topics

Scenario: 1,618 undergraduate students were enrolled at St. Ambrose during the Fall semester of 1977. By 2013, that number increased to 2,743. The plot to the right shows the trend in enrollment over time:

Source: 2013 StatPak

Data: http://www.bradthiessen.com/html5/data/sauenroll.csv



1. Obviously, it doesn't look like we have a linear trend here, but we can find the best-fitting linear model:

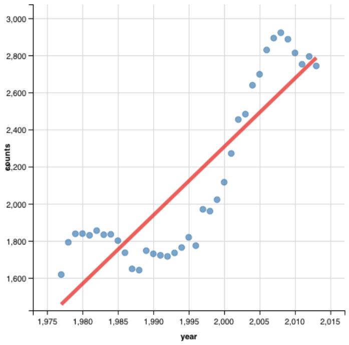
Model: # of students = $b_0 + b_1$(year)

Least-squares line: y = -71557.5 – 36.932x

$R^2 = 0.7556$

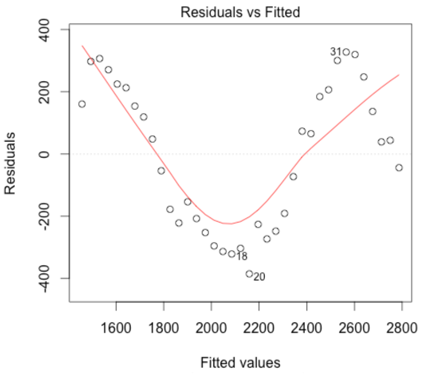
RMSE = 230.6

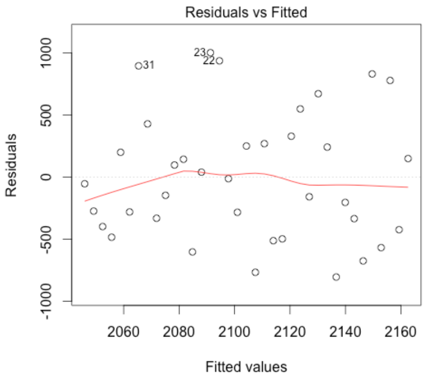
F = 108.2 (p = 2.998e-12)

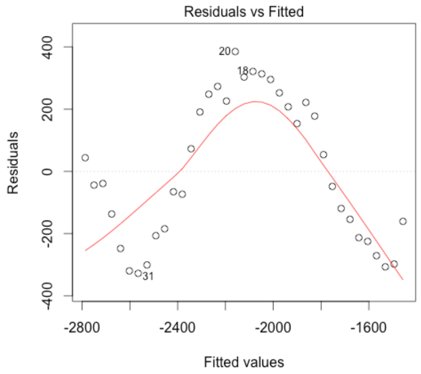


The F-statistic indicates this linear model is better than a model with no predictors and the R-squared value shows the fit isn't too bad.

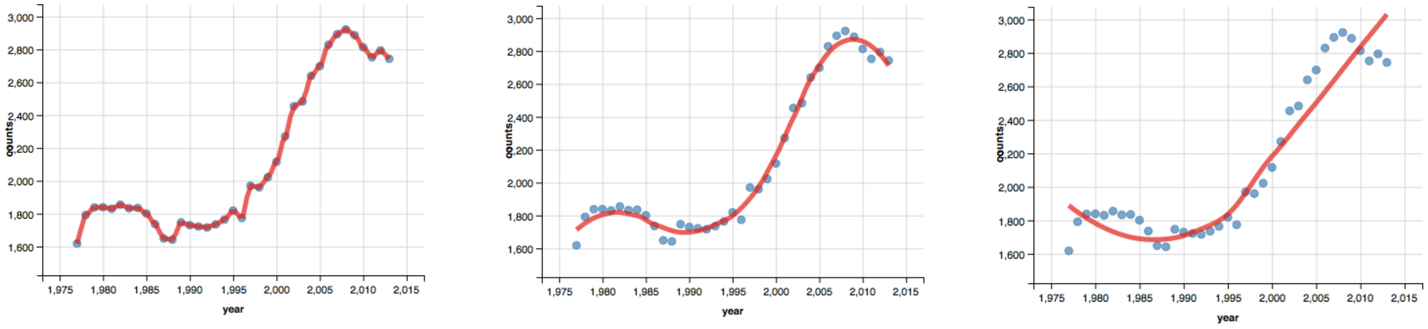
Suppose we plot our residuals from this linear model versus the fitted values. What would that graph look like? One of the following 3 plots is the actual residual plot. The other two are from simulated datasets. Circle the plot that was generated from the linear model displayed above.







2. Below, I've sketch lowess curves with bandwidth (spans) of 0.1 (left), 0.5 (middle), and 1.0 (right).



The curve on the left fits the data perfectly (R-squared = 1.0). Explain why we would not want to use this to model enrollment over time.

____

____

What bandwidth (span) would you choose to use to best model enrollment over time? Briefly justify your choice.

I would choose a bandwidth between… (pick one to justify)

0.0 - 0.1 because ____

0.1 - 0.5 because ____

0.5 - 1.0 because ____

1.0 + because ____

3. Let's improve upon our linear model by including some higher-powered terms. On the next page, I fit a model that includes a quadratic term and a model that also includes a cubic term.

I then compared these models using the omnibus F-test.

From all this output, which model (if any) would you choose? Briefly justify your choice.

____

____

____

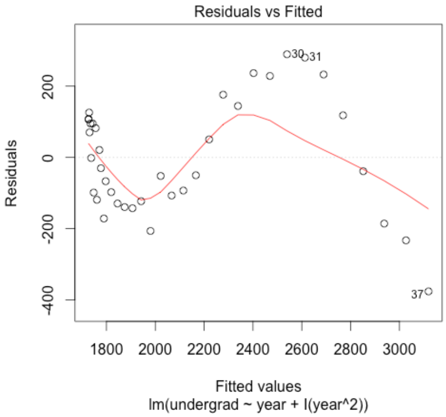
Model: $y = b_0 + b_1(x) + b_2(x)^2$

$R^2 = 0.8816$

RMSE = 162.8

Compared to null model:

F = 126.6 ($p < 2.2e-16$)



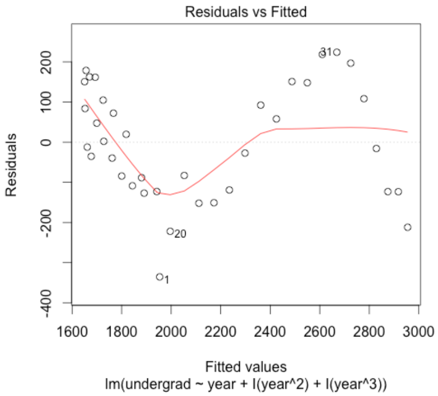
Model: $y = b_0 + b_1(x) + b_2(x)^2 + b_3(x)^3$

$R^2 = 0.9075$

RMSE = 146.1

Compared to null model:

F = 107.9 ($p < 2.2e-16$)



```
Analysis of Variance Table

Model 1: undergrad ~ 1
Model 2: undergrad ~ year
Model 3: undergrad ~ year + I(year^2)
Model 4: undergrad ~ year + I(year^2) + I(year^3)

  Res.Df     RSS Df Sum of Sq        F    Pr(>F)    
1     36 7614556                                    
2     35 1861258  1   5753297 269.4934 < 2.2e-16 ***
3     34  901452  1    959806  44.9588  1.23e-07 ***
4     33  704503  1    196949   9.2254  0.004639 ** 
```

4. We can use cross-validation to help choose the best model

Average cross-validated mean square error:	Model
55020	enrollment = f(year)
29083	enrollment = f(year + year^2)
27059	enrollment = f(year + year^2 + year^3)

From this, which model would you choose?

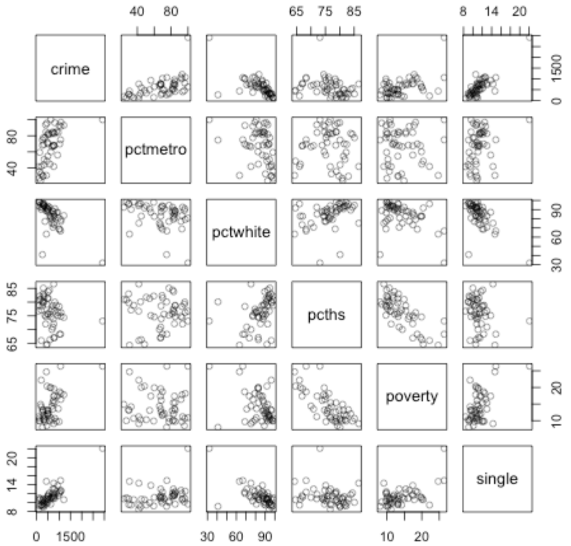
5. If you're at all familiar with desmos.com, you can see how to fit linear, polynomial, and nonlinear models using this example dataset at: https://www.desmos.com/calculator/6wzg3nqr3h. As of November 2014, regression models weren't fully implemented into desmos.com, so they still had some bugs to work out.

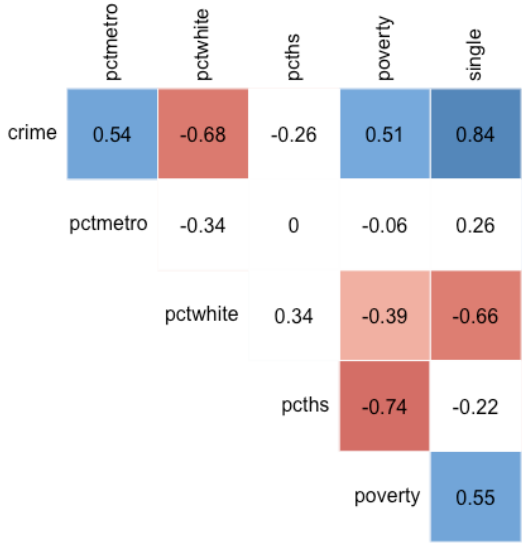
6. Let's investigate some robust regression methods with a new dataset. I'll use a crime dataset from an old edition of a statistics book I had in my office. This dataset contains the following variables for 51 states (counting D.C.):

   - crime = violent crimes per 100,000 people
   - pctmetro = percent of population living in metropolitan areas
   - pctwhite = percent of population that is white
   - pcths = percent of population with at least a high school education
   - poverty = percent of population living under poverty
   - single = percent of population that are single parents

   Source: Agresti, A. & Finlay, B. (1997). Statistical Methods for Social Sciences, 3rd edition. Prentice Hall.

   Below, I've pasted a scatterplot matrix and correlogram for this dataset.





	pctmetro	pctwhite	pcths	poverty	single
crime	0.54	-0.68	-0.26	0.51	0.84
pctmetro		-0.34	0	-0.06	0.26
pctwhite			0.34	-0.39	-0.66
pcths				-0.74	-0.22
poverty					0.55

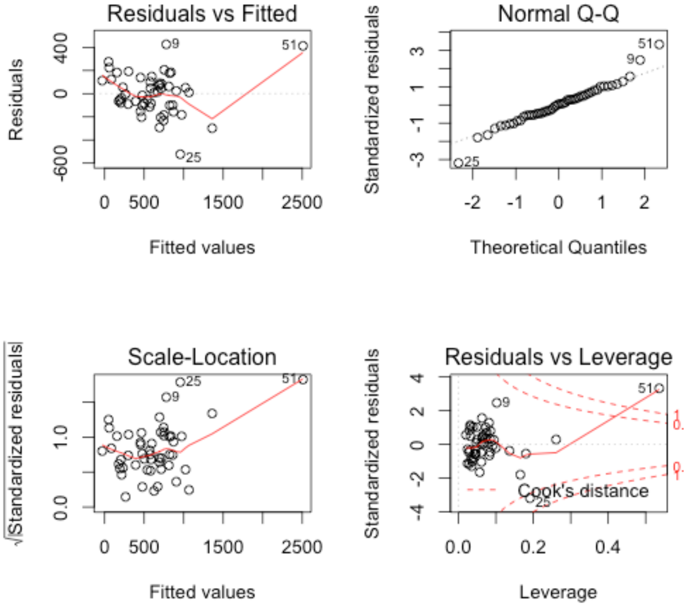
I fit the following model to this data:

Model: crime = $b_0 + b_1$(pctmetro) + b_2(poverty) + b_3(single)

Formula: y = -1666.4 + 7.829(pctmetro) +17.68(poverty) +132.4(single)

$R^2 = 0.8399$
RMSE = 182.1
F = 82.16 ($p < 2.2e-16$)

The residuals plots are displayed to the right.
(question on next page)



Based on those residual plots, which assumptions appear to be violated?

____

7. If you look closely at those residual plots, you'll notice a few data points have been labeled (observations #9, 25, 51). These observations (which represent: 9 = Florida, 25 = Mississippi, 51 = Washington DC) have relatively large residuals. To potentially lessen the impact of those outliers, we could run a robust regression analysis.

   Here are the results from a bootstrap-residuals method (like the one on page 8 of activity #12):

Bootstrap Residuals Method

	Linear Model		**Bootstrap Residuals**	
	Coefficient	Standard Error	Coefficient	Standard Error
(intercept)	-1666.4359	147.852	-1664.33	159.7031
pctmetro	7.8289	1.255	7.797	1.3387
poverty	17.6802	6.941	17.661	7.4232
single	132.4081	15.503	132.476	16.9180

   Here are the results from a bootstrap-cases method (like the one on page 9 of activity #12):

Bootstrap Cases Method

	Linear Model		**Bootstrap Cases**	
	Coefficient	Standard Error	Coefficient	Standard Error
(intercept)	-1666.4359	147.852	-1560.263	303.3921
pctmetro	7.8289	1.255	7.68378	1.4298
poverty	17.6802	6.941	18.9687	7.5617
single	132.4081	15.503	121.9891	29.9935

   As we've seen, we can use a t-test or confidence interval to test the statistical significance of each coefficient of our regression model. Roughly, a confidence interval is obtained by taking the coefficient +/- 2 standard errors.

   If we were to construct confidence intervals for each coefficient, which (if any) would be significantly different from zero? Circle all the significant coefficients in each row:

Linear model:	**intercept**	**pctmetro**	**poverty**	**single**
Bootstrap-residuals:	**intercept**	**pctmetro**	**poverty**	**single**
Bootstrap-cases:	**intercept**	**pctmetro**	**poverty**	**single**

8. Let's stick with this crime dataset and use a quantile regression model. First, let's run a quantile regression model for the 50th percentile.

   Here's the output I got from fitting a model with the predictors pctmetro, poverty, and single:

```
Call: rq(formula = crime ~ pctmetro + poverty + single, tau = 0.5, data = crime2)

Coefficients:
            coefficients lower bd    upper bd  
(Intercept) -1704.31718  -2132.91639  -901.89605
pctmetro        7.67461      4.01853     9.24265
poverty        17.75926     15.50658    37.62865
single        137.35707     61.58844   175.70516
```

   Interpret the coefficient for single. Keep in mind that our dependent variable is the number of crimes per 100,000 people and the predictor is the percent of the population that are single parents:

   137.36 represents: ____

   ____

   This time, let's transform all our predictors to z-scores (and <u>not</u> transform our dependent variable). I run the model again and obtain:

```
Call: rq(formula = crime ~ scale(pctmetro) + scale(poverty) + scale(single),
                   tau = 0.5, data = crime2)

Coefficients:
                coefficients lower bd  upper bd 
(Intercept)     621.73895    540.39079 657.86311
scale(pctmetro) 168.51253     88.23537 202.94199
scale(poverty)   81.41275     71.08589 172.49882
scale(single)   291.40222    130.65952 372.75748
```

   Interpret the intercept

   621.74 represents: ____

   ____

   Predict the # of crimes (per 100,000) for a state that is **+1** standard deviation from the mean on each predictor:

   Prediction = ____

   Predict the # of crimes (per 100,000) for a state that is **–1** standard deviation from the mean on each predictor:

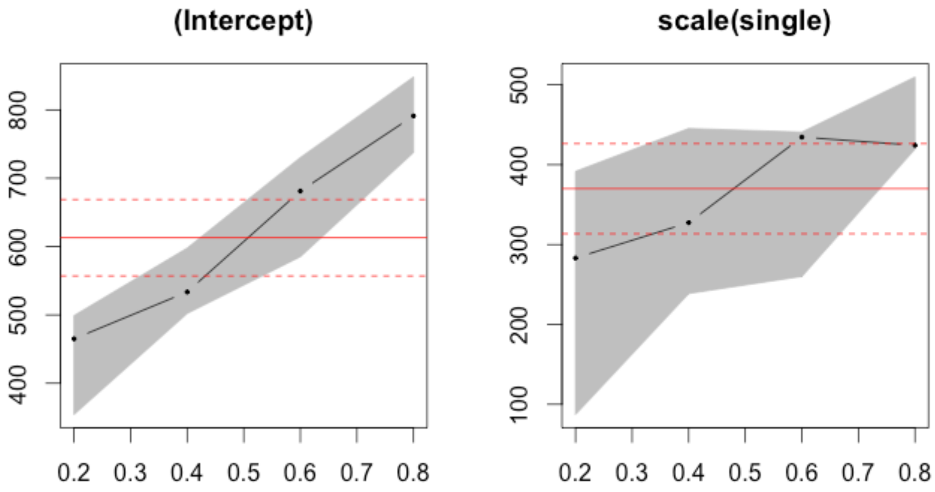
   Prediction = ____

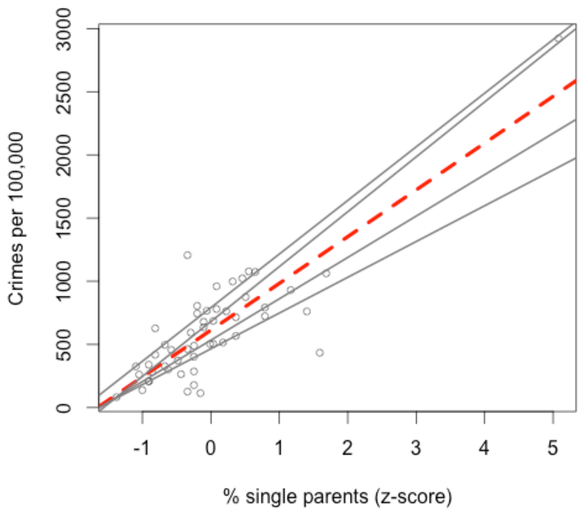
9. Let's simplify our model to predict crimes based only on the percent of the population that are single parents. Here's a simple linear regression model (notice that I converted the predictor to a z-score):

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      612.84      33.96   18.05  < 2e-16 ***
scale(single)    370.03      34.30   10.79 1.53e-14 ***

Residual standard error: 242.5 on 49 degrees of freedom
Multiple R-squared:  0.7037,  Adjusted R-squared:  0.6977
F-statistic: 116.4 on 1 and 49 DF,  p-value: 1.529e-14
```

Now, let's conduct a quantile regression for the 20th, 40th, 60th, and 80th percentiles:





From these graphs, what conclusions can we make?

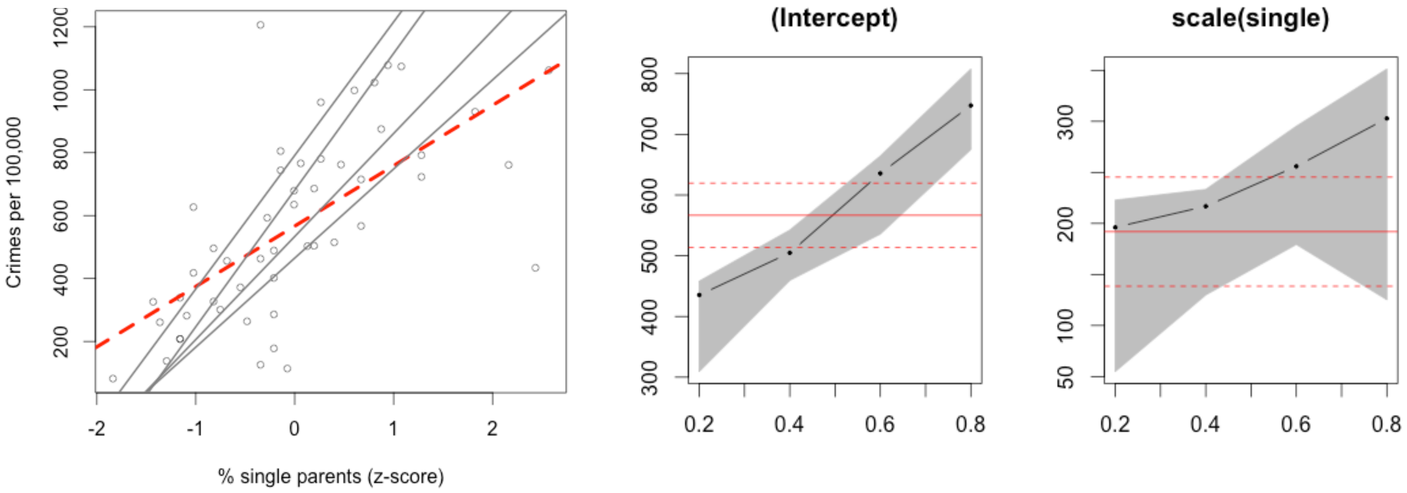
10. Notice that outlier in the top-right corner of the previous graph? It represents Washington DC (a "state" in which 22.1% of its population are single parents). Let's eliminate this outlier (with the justification that it's not a state) and see what impact it has on our linear and quantile regression analyses:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      566.66      32.18  17.612  < 2e-16 ***
scale(single)    191.93      32.50   5.905  3.5e-07 ***

Residual standard error: 227.5 on 48 degrees of freedom
Multiple R-squared:  0.4208,   Adjusted R-squared:  0.4087
F-statistic: 34.87 on 1 and 48 DF,  p-value: 3.499e-07
```

Compare this to the linear regression model on the previous page. What impact did removing this outlier have on our results?

Now let's again run our quantile regression models:



From this, what can we conclude about the relationship between the percent of single parents in a state and that state's crime rate?