Non-linear Relationships, Polynomial Regression & Potential Problems

Links: MATH 4322

Recall:

> ### Two Important Assumptions
>
> The **additive** assumption means that the effect of changes in a predictor X_j on the response Y is independent of the values of the other predictors. (covered in this note)
>
> The **linear** assumption means that the change in the response Y due to a one-unit change in X_j is constant, regardless of the value of X_j. (Linearity will be covered in the next note)

Non-linear Relationships

The linear regression model assumes a linear relationship between the response and predictors. The true relationship between the response and the predictors may be non-linear.

The *polynomial regression* is a very simple way to directly extend the linear model to accommodate non-linear relationships.

Polynomial Regression

Polynomial regression is a form of regression analysis in which the relationship between the predictor *x* and the response *y* is modeled as

the n^{th} degree polynomial *x*.

Model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_m x_i^m + \epsilon_i$$

for $i = 1, 2, \ldots, n$. We need to keep $m < n$. (We want to keep our terms, our powers, less than the number of observations.)
In R we use `lm(y~poly(x,m))`.

As an example of a non-linear relationship we can take the following example: In 1981, n = 78 bluegills were randomly sampled from Lake Mary in Minnesota. The researchers (Cook and Weisberg, 1999) measured and recorded some data.

- Response (y): length (in mm) of the fish
- Potential predictor (x1): age (in years) of the fish

The researchers were primarily interested in learning how the length of a bluegill fish is related to it age.
(This is a regression problem, and is more inference "This will be an exam question" - Prof. Poliak)

Linear Summary

fish.lm = lm(length~age,data = index)

```
> summary(fish.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.649      5.755   10.89   <2e-16 ***
age           22.312      1.537   14.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.51 on 76 degrees of freedom
Multiple R-squared:  0.7349,Adjusted R-squared:  0.7314
F-statistic: 210.7 on 1 and 76 DF,  p-value: < 2.2e-16
```

$H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$

$H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$

Equation:

$\hat{\text{length}} = 62.649 + 22.312 * \text{age}$

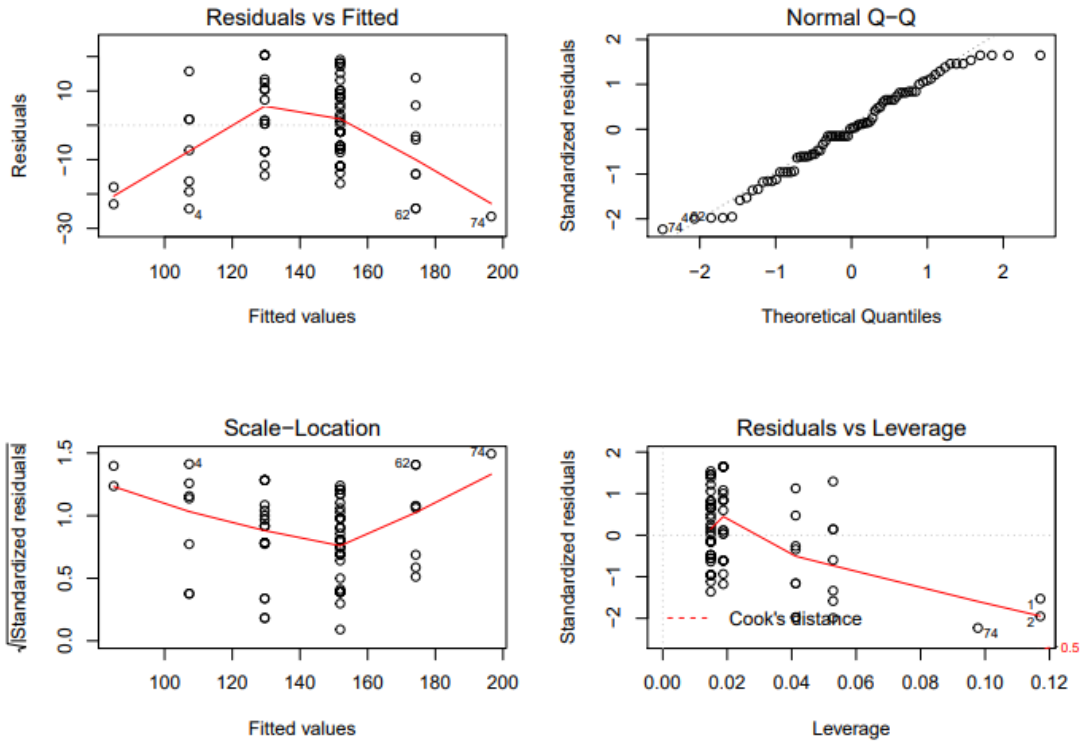
Is this a "good" equation to determine the relationship between length & age?

- Diagnostic plots. RSE = 12.51 adj. $R^2 = 0.7314$

We can see from this summary that the p-values shows that at the predictor is significant, and that its R-squared is also high.

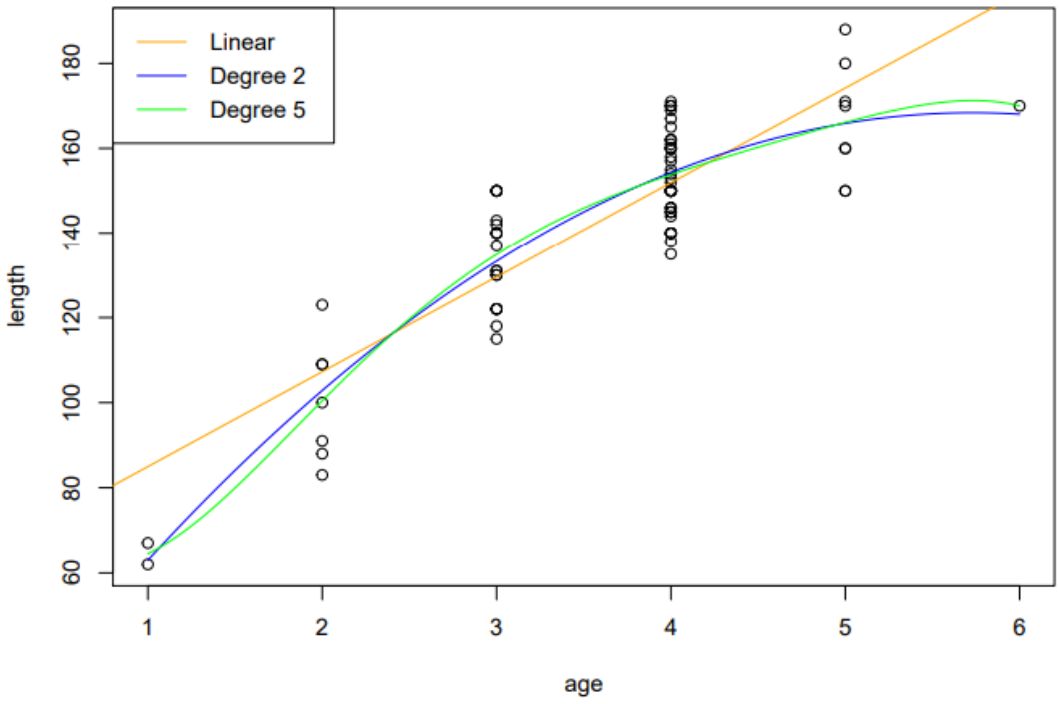
We can also look at the diagnostic plots:

Linear?



These plots show that the relationship is not linear, maybe to get a better fit we could add a higher degree to the polynomial.

Here are some higher degree fits on the data's scatter plot:



second degree regression model of the above example:

```
> fish.lm2 = lm(length~poly(age,2),data = index)
> summary(fish.lm2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   143.603      1.235 116.290  < 2e-16 ***
poly(age, 2)1 181.565     10.906  16.648  < 2e-16 ***
poly(age, 2)2 -54.517     10.906  -4.999 3.67e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.91 on 75 degrees of freedom
Multiple R-squared:  0.8011,Adjusted R-squared:  0.7958
F-statistic: 151.1 on 2 and 75 DF,  p-value: < 2.2e-16
```

equation: $\hat{length} = 143.603 + 181.565 * age - 54.517 * age^2$

Is this equation better? RSE = 10.91 adjR^2 = 0.7958

notice the RSE is smaller and the adjusted R^2 is bigger! If you do the diagnostics plot (see slide 12 of lecture 8 slides) you will see that the

residuals vs fitted plot and scale-location plots look much better. So this is not too bad.

If you attempt to do a cubic regression (3rd degree) you will see that the p value for the t-test of the 3rd degree predictor is not significant, which indicates that we don't want a 3rd degree and we can stop at the second degree (see slide 13 of lecture 8 slides).

"We don't want too many terms... it gets hectic with too many" ~ Prof. Poliak

Warnings About Polynomial Models

(source of this list)

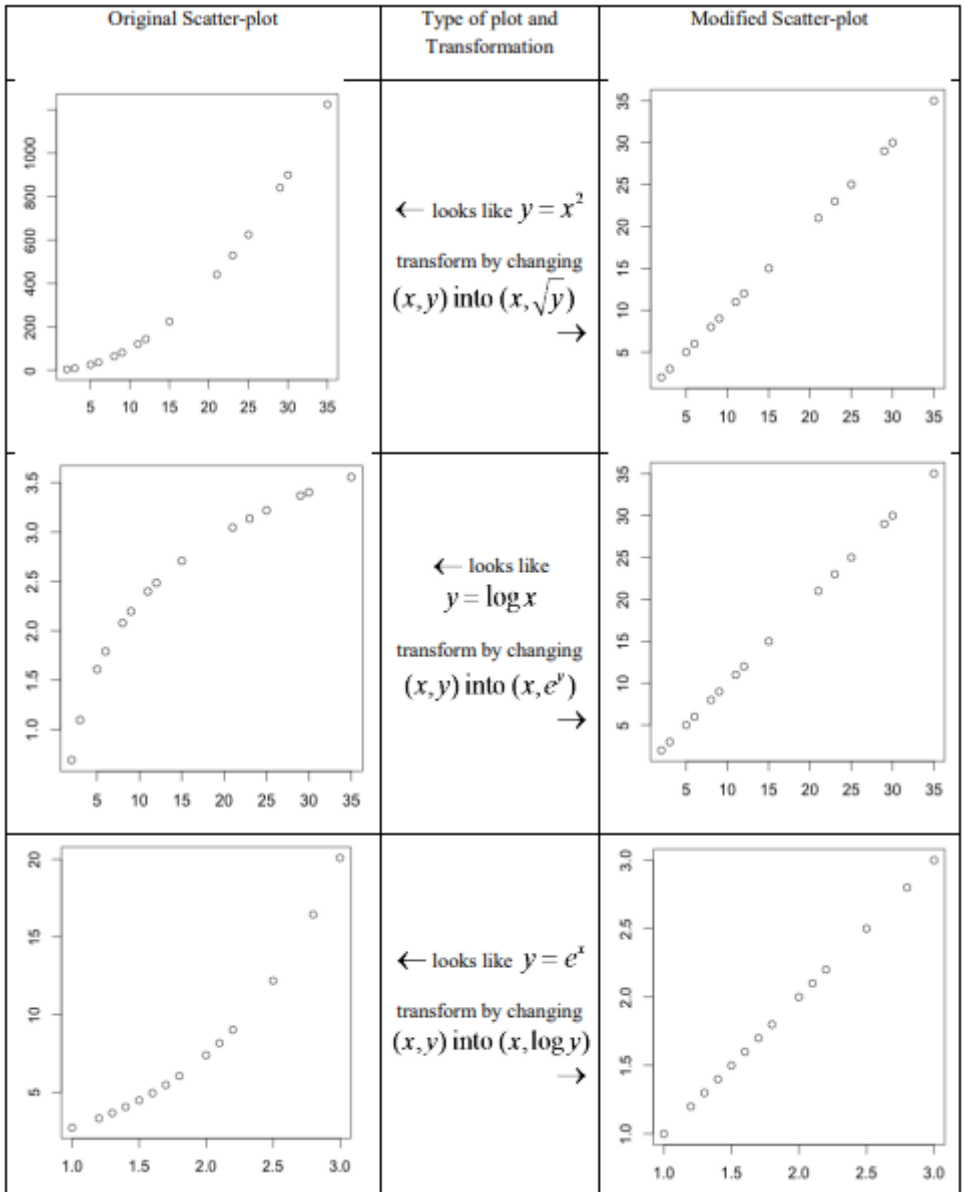
- The fitted model is more reliable when built on a larger sample size.
- Consider how large the size of the predictor(s) will be when incorporating higher degree terms as this may cause numerical overflow for the statistical software being used.
- Do not go strictly by low *p-values* to incorporate a higher degree term, but rather just uses these to support your model only if the resulting residual plots look reasonable. We want to look at the whole picture before making the decision (so look at R^2, RSE, AIC, BIC and Diagnostics Plots).
- As a standard practice if you have an n^{th} degree polynomial, always include each X^j such that $j < n$.

Potential Problems in Linear Regression

1. Non-linearity of the response-predictor relationships (see everything above)
2. Correlation of error terms
3. Non-constant variance of error terms

4. Outliers
5. High-leverage points
6. Collinearity

We could also transform the data to try and make it look linear:



Correlation of Error Terms

has to do with if the observations are dependent. "We will not be going over this in this course, it is explained more in detail in the masters program." ~ Prof. Poliak

- Assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are uncorrelated.
- For example if ϵ_i is positive provides little or no information about the sign of ϵ_{i+1}.
- If there is correlation among the error terms then the estimated standard errors will tend to underestimate the true standard errors. This results in narrower confidence and prediction intervals.
- Time series data is an example of correlation among the error terms.
- Residual plot is best way to tell if there is correlation.

"What happens [with the projects] is that sometimes the residual plots look really weird and that's because there's dependency" ~ Prof. Poliak

Non-constant Variance of Error Terms

Heterscedasticity is where the variances of the error terms increase with the value of the response. This will appear as a *funnel shape* in the residual plot.
Possible solution to get rid of that is to transform the *Y* using $log(Y)$ or $\sqrt{Y}$.

Outliers

An **outlier** is a point for which y_i is far from the value predicted by the model. Outliers can have an effect on the estimated regression parameters, RSE and R^2.
Scatterplot and residual plot would be the best detect outliers.

See baseball example from lecture 8 video (at around 43 minutes). But basically don't always ignore the extreme values.

High Leverage Points

High leverage points have an unusual values for *X* (while extreme values look at unusual *Y* values). High leverage observations tend to have a sizable impact on the estimated regression line.

Can be determined by scatterplots for a simple linear regression. In order to quantify an observation's leverage we compute the **leverage statistic**.

In `R` we can use the *Residuals vs Leverage* to see if there are high leverage observations (this is the bottom right plot in the diagnostics plots).

Collinearity

Collinearity refers to the situation in which two or more predictor variables are closely related to one another. In regression this will cause difficulty to separate out the individual effects of collinear variables on the response.

This reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\hat{\beta}_j$ to grow. The **power** of the hypothesis test for $H_0 : \beta_i = 0$ - probability of correctly detecting a nonzero coefficient - is reduced by collinearity.

"Collinearity... this refers to a situation in which two or more predictor variables are closely related to one another, if there's collinearity what happens is the power of the hypothesis of correctly detecting a non-zero coefficient is reduced by collinearity... because if there's a relationship between the two predictors we don't know in which way that relationship is" ~ Prof. Poliak

Detecting Multicollinearity

One way we look at this collinearity it by looking at the correlation matrix. In R: `cor()`. If the correlations are high we say there is collinearity.

The other thing we can check is the *variance inflation factor (VIF)*. The VIF is the ratio of the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own.

- The smallest possible value for *VIF* is 1, this means there is no correlation.
- If a *VIF* exceeds 4, further investigation is needed.
- If *VIF* is more than 10, then there is a sign of serious multicollinearity and requires correcting.
- The *VIF* for each variable can be computed using the formula:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

  Where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

In R this can be done with: `vif(name.lm)`