Polynomial Regression

BIOS 6611

CU Anschutz

Week 12

Polynomial Models

Polynomial Models with One Variable

A k^{th} order polynomial in one variable, x, is an expression of the following form:

$$y = c_0 + c_1 x + c_2 x^2 + \ldots + c_k x^k$$

in which the c's and the k (which much be a nonnegative whole number) are constants.

The statistical model is an expression of the form:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \epsilon, \text{ where } \epsilon \sim N\left(0, \sigma^2_{Y|X}\right)$$

The statistical model is a linear regression model because Y is a linear function of the β's.

Polynomial Models

Polynomial models are useful:

- In situations where the analyst knows that curvilinear effects are present in the true response function.
- As approximating functions to unknown and possibly very complex nonlinear relationships.

Important considerations when using polynomial models include:

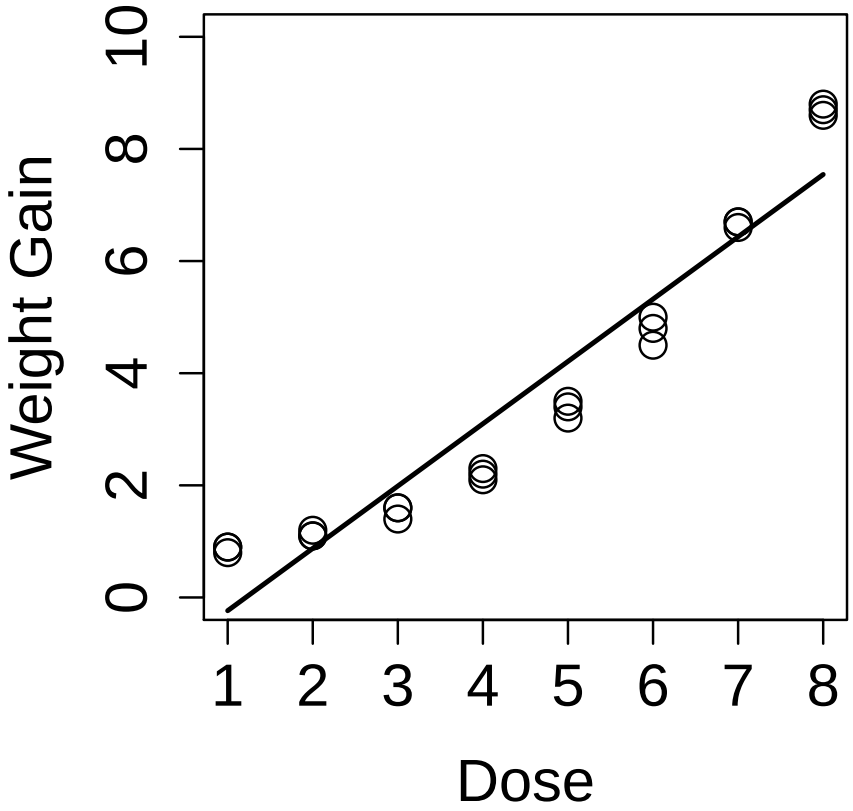
- Selecting the order of the model (model selection strategy)
- Extrapolation
- Ill-conditioning

Motivating Example

From our KKNR textbook, a laboratory study is undertaken to determine the relationship between equally spaced doses (X) of a certain drug and weight gain (Y). 24 laboratory animals of the same sex, age, and size are selected and 3 animals are randomly assigned to each dose group.

The results from the study were

	Wgt Gain (Y)		
X	1	2	3
1	0.9	0.9	0.8
2	1.1	1.1	1.2
3	1.6	1.6	1.4
4	2.3	2.1	2.2
5	3.5	3.4	3.2
6	5	4.5	4.8
7	6.6	6.7	6.7
8	8.7	8.6	8.8



Model 1: No Polynomial Terms

Starting with a simple linear regression model with no polynomial terms, we have:

```
wtgain <- data.frame( dose=rep(1:8, each=3),
    wgtgain=c(0.9,0.9,0.8,1.1,1.1,1.2,1.6,1.6,1.4,2.3,2.1,2.2,
              3.5,3.4,3.2,5,4.5,4.8,6.6,6.7,6.7,8.7,8.6,8.8) )
lm1 <- lm( wgtgain ~ dose, data=wtgain)
summary(lm1)$coefficients
```

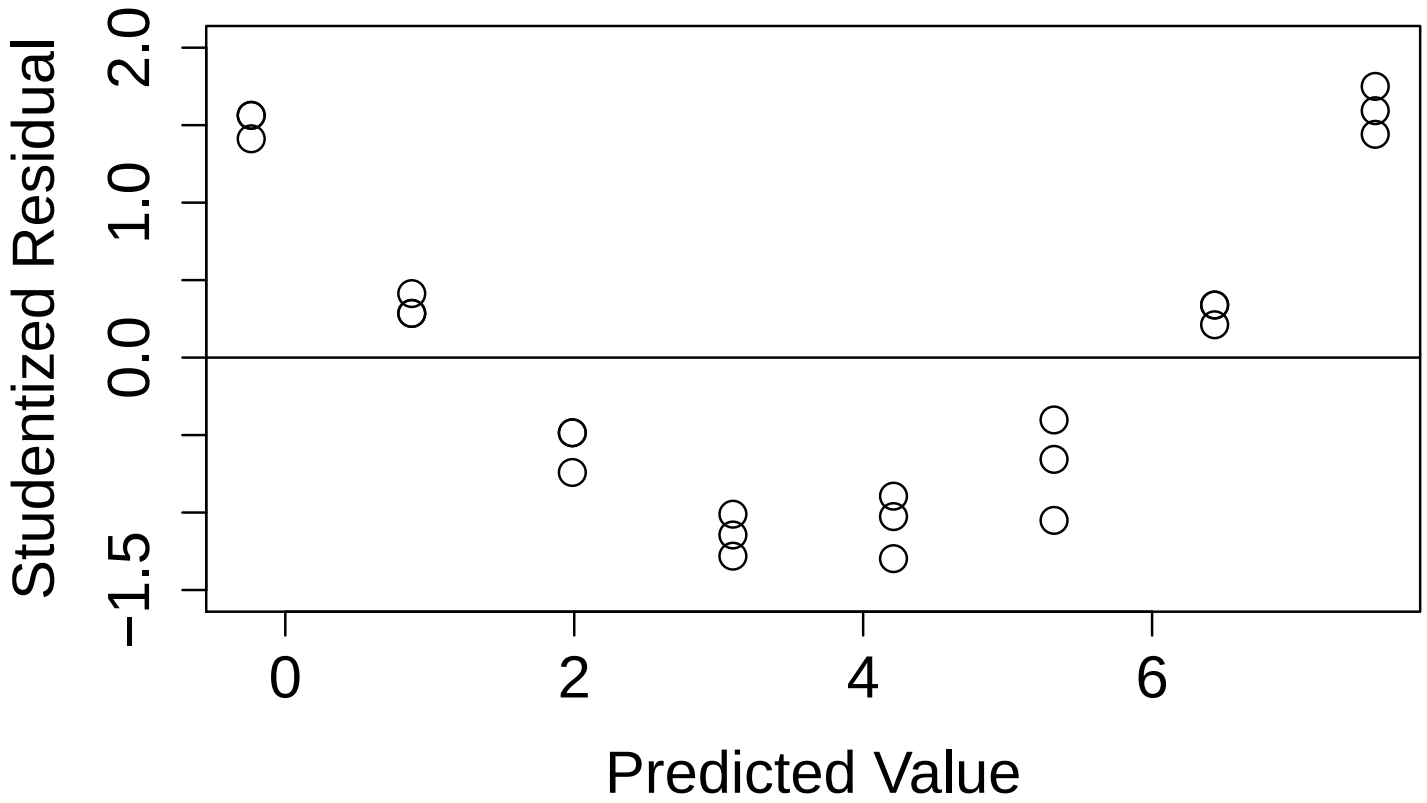
```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -1.347619 0.36361822 -3.706137 1.231652e-03
## dose         1.111508 0.07200714 15.436080 2.755834e-13
```

Intercept: The predicted weight gain with the dose is 0 is -1.35, however this is beyond the range of our data.

Slope: There is a significant association between dose and weight gain, where for everyone one unit increase in dose, weight gain increases by 1.11 units (95% CI: 0.96, 1.26), on average ($p<0.001$).

Model 1: Residual Plot

```
plot(x=predict(lm1) ,y=rstudent(lm1),
     xlab='Predicted Value', ylab='Studentized Residual',
     cex.lab=1.5, cex.axis=1.5, cex=1.5, ylim=c(-1.5,2))
abline( h=0 )
```

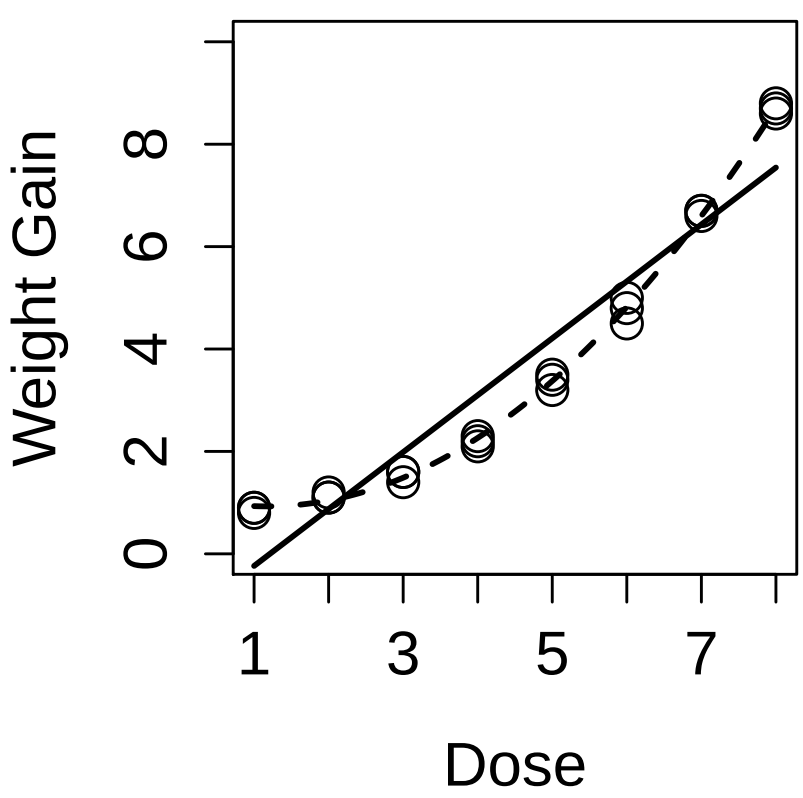


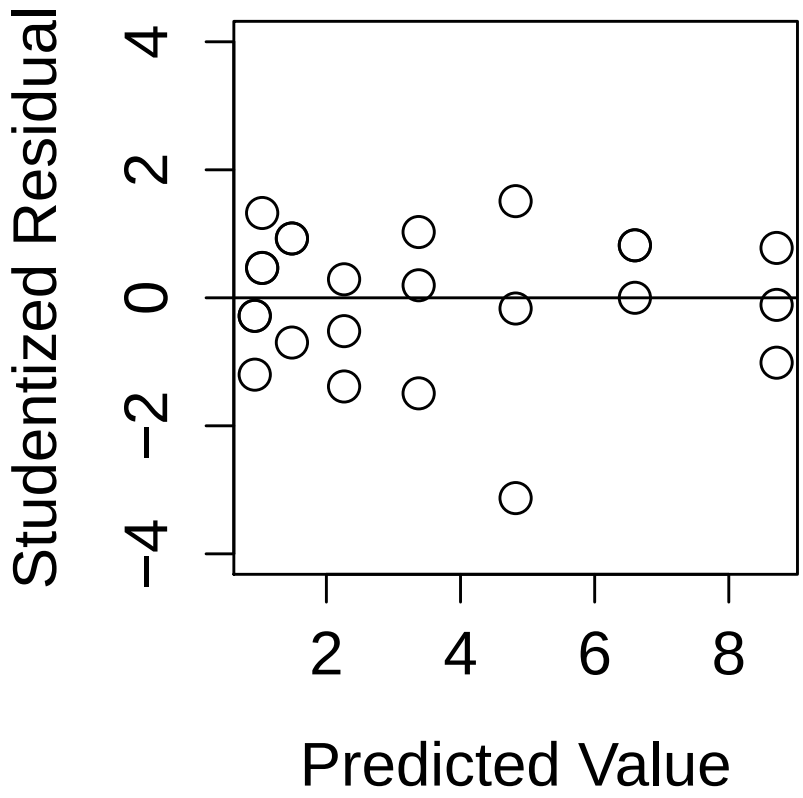
Does a straight-line model fit the data?

Model 2: Quadratic Model

What if we fit a polynomial model of order 2?

```
wtgain$dose2 <- wtgain$dose^2
lm2 <- lm( wgtgain ~ dose + dose2, data=wtgain)
lm2_alt <- lm( wgtgain ~ dose + I(dose^2), data=wtgain) #equivalent coding
```





Model 2: Quadratic Model

```
summary(lm2)
```

```
##
## Call:
## lm(formula = wgtgain ~ dose + dose2, data = wtgain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32083 -0.06964  0.01230  0.10020  0.17917
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.155357   0.102423  11.280 2.26e-10 ***
## dose        -0.390278   0.052219  -7.474 2.41e-07 ***
## dose2        0.166865   0.005664  29.461  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1272 on 21 degrees of freedom
## Multiple R-squared:  0.998,  Adjusted R-squared:  0.9978
## F-statistic:  5248 on 2 and 21 DF,  p-value: < 2.2e-16
```

Model 2: Quadratic Model

Is the overall regression model significant? That is, is more of the variation in Y explained by the second-order model than by ignoring X completely and just using $\bar{Y}$?

Does the second-order model provide significantly more predictive power than the straight-line model?

Model 1 vs. Model 2

Model 1 (SLR): $\hat{Y} = -1.35 + 1.11X_1$

Source	Sums of Squares	Degrees of Freedom	Mean Square	F-value	p-value
Model	155.667	1	155.667	238.273	$<$0.001
Error	14.373	22	0.653		
Total	170.040	23			

Model 2 (Quadratic Model): $\hat{Y} = 1.16 + -0.39X_1 + 0.17X_1^2$

Source	Sums of Squares	Degrees of Freedom	Mean Square	F-value	p-value
Model	169.70	2	84.85	5247.78	$<$0.001
Error	0.34	21	0.016		
Total	170.04	23			

What happened to our β's?

What happened to the MSE?

Should I Add Higher Order Terms?

Given that a second-order model is more appropriate than a straight-line model, should we add higher order terms to the second-order model?

It is possible adding higher order terms may be beneficial, but we must balance this with our consideration of identifying a parsimonious model.

As we saw with our example, we can evaluate the potential significance by using a partial F-test or a t-test. But we should not only rely on a low p-value, but also examine the residual plots and consider the context of our problem. In cases where we have replicates at each level, we have additional tools.

Testing Lack of Fit with Replicates

Estimating MSE and Lack of Fit

Recall that we use MSE to estimate $\sigma^2_{Y|X}$, which is calculated as

$$\hat{\sigma}^2_{Y|X} = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n-2} = \frac{SS_{Error}}{n-2} = MS_{Error}$$

The MSE will only provide an unbiased estimate of the error variance when the hypothesized model is correct, otherwise $\hat{\sigma}^2_{Y|X} > \sigma^2_{Y|X}$.

If the model is incorrect, then two factors contribute to the inflation of the SSE:

1. The true variability in Y (**pure error**)
2. Error due to fitting an incorrect model (**lack of fit error**)

With replicate observations we can formally test for lack of fit.

Estimating Pure Error By Hand

	Wgt Gain (Y)				SS_{PE}	
X	1	2	3	$\bar{Y}_x$	$\sum_m (Y_{mx} - \bar{Y}_x)^2$	df
1	0.9	0.9	0.8	0.866667	0.006667	2
2	1.1	1.1	1.2	1.133333	0.006667	2
3	1.6	1.6	1.4	1.533333	0.026667	2
4	2.3	2.1	2.2	2.200000	0.020000	2
5	3.5	3.4	3.2	3.366667	0.046667	2
6	5.0	4.5	4.8	4.766667	0.126667	2
7	6.6	6.7	6.7	6.666667	0.006667	2
8	8.7	8.6	8.8	8.700000	0.020000	2
					$\sum = 0.26$	$\sum = 16$
				$MS_{PE} = 0.26/16 = 0.01625$		

So, to test the linear trend using the "pure error" and Model 1 results:

$$t = \frac{\hat{\beta}_{dose}}{\sqrt{\frac{MSE(pure)}{MSE(pure+LOF)} \times (SE(\hat{\beta}_{dose}))^2}} = \frac{1.11151}{\sqrt{\frac{0.01625}{0.65331} \times (0.072012)^2}} = 97.87$$

Our previous $t = 15.44$ from Model 1. The variance was 40 times higher due to lack of fit error from the straight line model.

Estimating Pure Error via a Model

We can also obtain the "pure error" by fitting a *saturated model* (i.e., a model that includes a dummy code for each level k of X):

```
lm_pure <- lm( wgtgain ~ as.factor(dose), data=wtgain)
coef(lm_pure)
```

```
##      (Intercept) as.factor(dose)2 as.factor(dose)3 as.factor(dose)4
##        0.8666667        0.2666667        0.6666667        1.3333333
## as.factor(dose)5 as.factor(dose)6 as.factor(dose)7 as.factor(dose)8
##        2.5000000        3.9000000        5.8000000        7.8333333
```

```
linreg_anova_func(lm_pure, ndigits=3)
```

Source	Sums of Squares	Degrees of Freedom	Mean Square	F-value	p-value
Model	169.78	7	24.254	1492.568	<0.001
Error	0.26	16	0.016		
Total	170.04	23			

Our estimated MSE from the ANOVA table for this saturated model matches our "by hand" calculation of the pure error.

Lack of Fit

The difference in the regression sum of squares between the lower-order model being considered and the full model containing all higher-order terms is the **lack of fit sum of squares**.

The **lack of fit test statistic** is a partial F test for testing the addition of the higher-order terms (up to the highest order) to the polynomial model:

$$F = \frac{[SS_{model}(full) - SS_{model}(reduced)]/k}{MS_{error}(full)} \sim F_{k,n-p-k-1}$$

Like before, we will let n = number of observations, p = number of IVs in the *reduced model*, and k = number of IVs *removed* from the full model.

Note: The SSE of the highest-order polynomial model is equivalent to the SSE for a model including a dummy variable for each dose level.

Testing Lack of Fit for the Straight-Line Model

For our example, the highest order polynomial we could fit is 7 (i.e., number of dose levels minus 1). The lack of fit test for Model 1 (straight-line model) tests

$H_0 : \beta_{quad} = \beta_{cubic} = ... = \beta_{septic} = 0$, or equivalently

$H_0 : \beta_{x^2} = \beta_{x^3} = ...\beta_{x^7} = 0.$

The partial F-test is:

$$F = \frac{[SS_{model}(full) - SS_{model}(reduced)]/k}{MS_{error}(full)} = \frac{[169.78 - 155.667]/6}{0.01625} = \frac{14.113/6}{0.01625} = 144.75$$

We can then compare this to an $F_{6,16}$ distribution for our critical value and to calculate a p-value. The critical value is `qf(0.95,6,16)`$=2.741$ and p=`pf(144.75,6,16, lower.tail=F)`$=4.9947211 \times 10^{-13}$.

Since p$<$0.001 and F$>$2.741, we reject H_0. At least one higher order term is not equal to 0.

Testing Lack of Fit for the Quadratic Model

The lack of fit test for Model 2 (quadratic model) tests

$H_0 : \beta_{cubic} = ... = \beta_{septic} = 0$, or equivalently

$H_0 : \beta_{x^3} = ...\beta_{x^7} = 0.$

The partial F-test is:

$$F = \frac{[SS_{model}(full) - SS_{model}(reduced)]/k}{MS_{error}(full)} = \frac{[169.78 - 169.70]/5}{0.01625} = \frac{0.08/5}{0.01625} = 0.98$$

We can then compare this to an $F_{5,16}$ distribution for our critical value and to calculate a p-value. The critical value is `qf(0.95,5,16)`=2.852 and p=`pf(0.98,5,16, lower.tail=F)`=0.46.

Since p>0.05 and F<2.741, we fail to reject H_0. This suggests no higher order terms are needed, and the quadratic model has the "best fit".

Hierarchical Modeling and Collinearity

Hierarchical Models

Consider the polynomial model of order 2 (the quadratic model):

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

Suppose we fit this model and the coefficient $\hat{\beta}_1$ is not significant, but $\hat{\beta}_2$ is. If we removed the X term our reduced model becomes:

$$Y = \beta_0 + \beta_2 X^2 + \epsilon$$

But suppose we then made a location change in our predictor X, e.g., $X + z$, where z is some constant. The model would become:

$$Y = \beta_0 + \beta_2 X^2 + 2\beta_2 Xz + \beta_2 z^2 + \epsilon$$

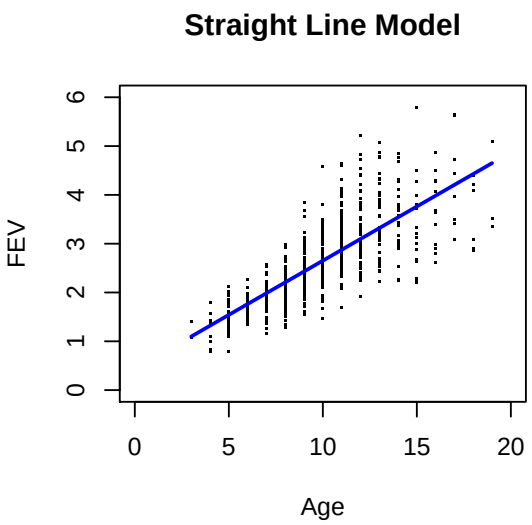
The first order X term has reappeared, so our model has effectively changed.

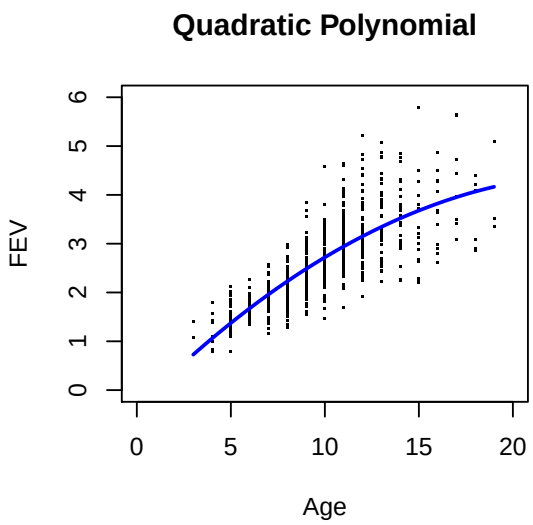
Hierarchical Models

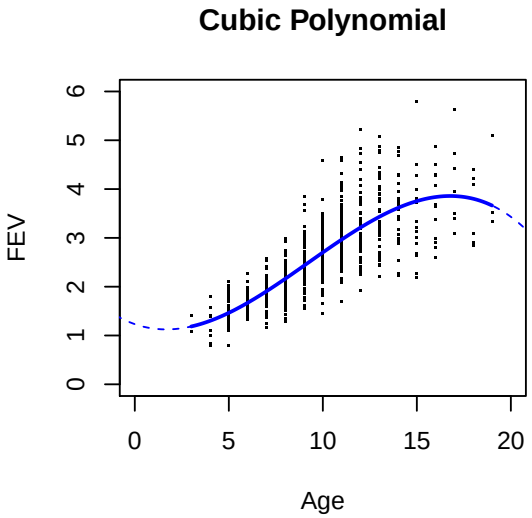
In general, location changes should *not* make any important changes to the model, but in this case an additional term has been added.

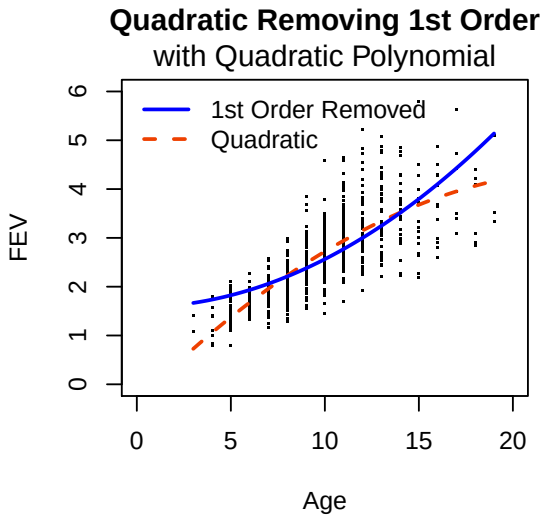
For this reason, we should not remove lower order terms in the presence of higher order terms (we do not want the conclusion to depend on the choice of location).

FEV Example - Polynomial Style









For our FEV data set, we have examples of different polynomial models, including one where we've removed the first order term.

The best statistical model is the cubic model (either via the pure error F-test or by using partial F-tests to compare the quadratic, cubic, and quartic models).

However, the behavior in the cubic model's tails may not make scientific sense. This is a good example of why plotting the regression equation can be a helpful step in identifying a model that is both statistically and scientifically meaningful.

Collinearity Problems

Collinearity problems can arise in polynomial models that may make evaluating the statistical significance challenging. Two possible solutions include:

- Centering the predictors if you are only interested in comparing a first-order and second-order polynomial model (removes the collinearity).
- Orthogonal polynomial contrasts can be used and can be extended to higher order comparisons. Supplemental lecture notes are provided on this topic for your information, but this will not be on the homework or exams.