

Regression III

Lecture 5: Linearity

Dave Armstrong

University of Wisconsin – Milwaukee
Department of Political Science

e: armstrod@uwm.edu
w: www.quantoid.net/ICPSR.php

1 / 70

The Linearity Assumption

Testing in GLMs
Local Polynomial Regression

Non-linearity

Assessing Non-linearity
Fixing Non-linearity
Fixing Non-linearity: Polynomials
Maximum Likelihood Transformations

Example

2 / 70

The Linearity Assumption

Perhaps the most important assumption of the linear model is that the relationship between y and x is accurately described by a line.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

This allows us to:

1. Characterize the relationship between y and x with a single (or small set of) numbers.
2. Easily interpret the marginal effect of x .
3. Easily present the results of the modeling enterprise.

3 / 70

Diagnosing Non-Linearity

We are often interested in the extent to which data we observe follow the assumption of linearity.

- Binary variables are always linearly related to the observed variables (two points define a line)
- Multi-category and continuous variables are not always linearly related to the response.
- We want to know the extent to which these variables exhibit linear relationships.

4 / 70

Linearity and Multi-Category Variables

Multi-category variables are often not problematic because we code them as a series of dummy variables. Thus, we are not imposing any functional form on the relationship between the categories and the response variable.

The waters are a bit murkier for ordinal variables (e.g., state repression or political ideology).

- These variables are often operationalized with relatively few categories.
- However, we often have a strong suspicion that the relationship between these variables and the response is “roughly linear”.
 - If the relationship is *not* linear and we represent it with a line, then we are getting a *biased* estimate of the relationship.
 - If the relationship could be represented linearly, and we represent it with a series of dummy variables, we are getting estimates that are *inefficient*

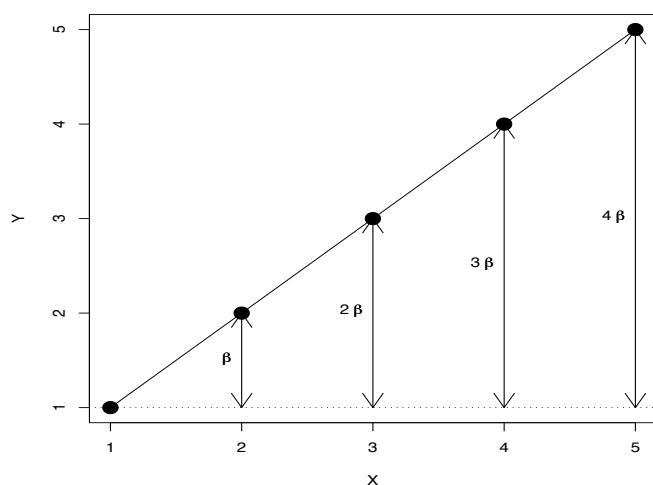
5 / 70

Testing the Hypothesis

- There are a couple of different ways we can test this hypothesis. One is a bit easier, so I will advocate that method in general, but one is a bit more explicit about what the exercise entails.
- We want to see if the unrestricted model (the one including the dummy variables) is statistically different from the one including only one linear term.
- What would we expect if response were perfectly linearly related to the explanatory variable?

6 / 70

Expectations



7 / 70

An Example

I generated data with the following such that $x_i \in \{1, 2, 3, 4, 5\}$ and

$$y_i = 2 + x + \varepsilon_i$$

where $\varepsilon_i \sim N(0, 2)$.

We can use a general linear hypothesis test to get the desired result. To accomplish this, we need to do:

1. Run the model by creating dummy variables for all but the smallest category of the variable in question.
2. Use the test of general linear hypotheses with the following matrices:

$$\mathbf{L} = \begin{bmatrix} 0 & 2 & -1 & 0 & 0 \\ 0 & 3 & 0 & -1 & 0 \\ 0 & 4 & 0 & 0 & -1 \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

8 / 70

Example Continued

Here is the model output:

Table:

(Intercept)	3.18*
	(0.19)
x2	0.60*
	(0.28)
x3	2.06*
	(0.28)
x4	2.75*
	(0.28)
x5	4.03*
	(0.28)
N	500
R ²	0.36
adj. R ²	0.36
Resid. sd	1.94

Standard errors in parentheses
* indicates significance at $p < 0.05$

9 / 70

Hypothesis Test

We can also perform a hypothesis test using the general linear hypothesis testing:

```
> L <- matrix(c(
+ 0, 2, -1, 0, 0,
+ 0, 3, 0, -1, 0,
+ 0, 4, 0, 0, -1), ncol=5, byrow=T)
> c <- rep(0, 3)
> e <- matrix(residuals(mod), ncol=1)
> s2e <- (t(e)%*% e)/with(mod, df.residual)
> b <- coef(mod)
> X <- model.matrix(mod)
> F0 <- (t(L)%*%b-c) %*% solve(L%*%solve(t(X) %*% X)%*%
+ t(L))%*%(L%*%b - c)/(3*s2e)
> cat("F = ",F0, "\n", sep="")

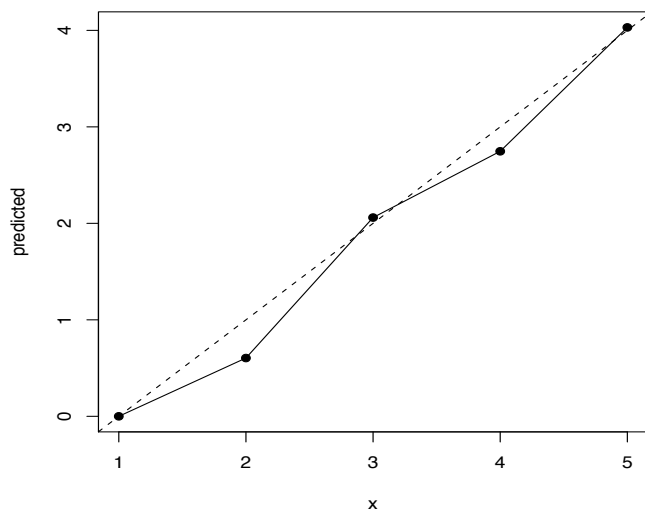
F = 1.400767

> cat("Pr(>F) = ", 1-pf(F0, 3, with(mod, df.residual)), "\n", sep="")

Pr(>F) = 0.2417872
```

10 / 70

Linear vs. Non-linear effect



11 / 70

Results

The results of the F -test suggest that the dummy variable model is not significantly better than the model with one linear term (i.e., $p > 0.05$).

There is another, equivalent way to do this test:

```
> library(xtable)
> restricted.mod <- lm(y ~ as.numeric(x))
> unrestricted.mod <- lm(y ~ x)
> xtable(anova(restricted.mod, unrestricted.mod, test="F"))
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	498	1888.36				
2	495	1872.46	3	15.90	1.40	0.2418

12 / 70

Linearity of Factors in GLMs

```
> anes <- read.dta("anes1992.dta")
> anes[["pidfac"]] <- as.factor(anes[["pid"]])
> unrestricted.mod <- glm(votedem ~ retnat + pidfac + age + male + educ +
+ black + south, data=anes, family=binomial)
> restricted.mod <- glm(votedem ~ retnat + pid + age + male + educ +
+ black + south, data=anes, family=binomial)
```

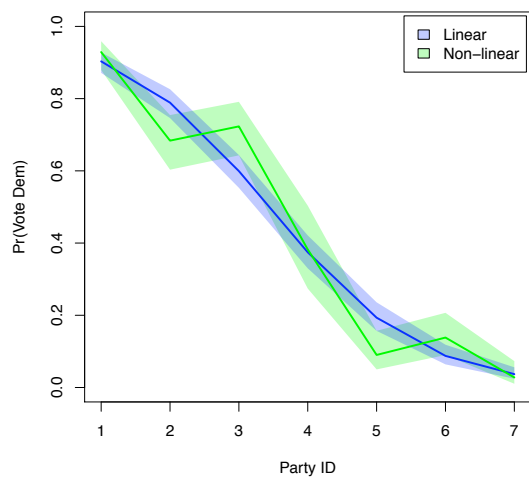
13 / 70

	Model 2	
(Intercept)	-0.10 (0.72)	0.51 (0.67)
retnatsame	1.24* (0.61)	1.33* (0.59)
retnatworse	1.57* (0.58)	1.65* (0.57)
pidfac2	-1.80* (0.35)	
pidfac3	-1.61* (0.35)	
pidfac4	-3.05* (0.39)	
pidfac5	-4.88* (0.44)	
pidfac6	-4.40* (0.39)	
pidfac7	-6.12* (0.60)	
age	0.01* (0.01)	0.01* (0.01)
male	-0.29 (0.18)	-0.29 (0.18)
educ	0.28* (0.11)	0.25* (0.10)
black	2.20* (0.49)	2.15* (0.48)
south	0.08 (0.23)	0.01 (0.22)
pid		-0.92* (0.06)
<i>N</i>	1039	1039
AIC	796.00	820.09
BIC	1072.98	998.15
log L	-342.00	-374.05

Standard errors in parentheses
* indicates significance at $p < 0.05$

14 / 70

Plot of effects



15 / 70

Diagnosing Non-Linearity

Diagnosing non-linearity in relationships between continuous predictors is a bit more tricky.

We will use an analysis of the residuals to diagnose whether the relationship between X and y is well-characterized by a line.

We will also need to figure out a flexible way to model the dependencies between X and the residuals.

- To do this, we will need to learn something about non-parametric regression

16 / 70

The Linearity Assumption

Testing in GLMs

Local Polynomial Regression

Non-linearity

Assessing Non-linearity

Fixing Non-linearity

Fixing Non-linearity: Polynomials

Maximum Likelihood Transformations

Example

17 / 70

Parametric vs. Non-parametric

Our goal is to trace the dependence of y on x . Specifically, we usually want to get something like:

$$y_i|x_i = f(x_i) + e_i$$

We usually define $f(\cdot)$ to be “smooth”.

- The linear functional form ($f(x_i) = \alpha + \beta x_i$) is the “smoothest” of smooth function.

The above model is parametric, because we are estimating *parameters* that describe relationship between y and x .

It is possible to characterize the relationship without estimating global parameters (i.e., parameters that apply to all of the observations equally) - what we call *non-parametric* models.

18 / 70

Global vs. Local Parametric Models

All of the models we will talk about below are *locally* parametric.

- They fit a parametric model to a relatively small subset of the data.
- The sum total of these many local parametric fits is a non-parametric fit - one that does not impose the same functional form for all of the data.

Because these models remain locally parametric, we can usually use information from the many local models to derive standard errors for the fit. (More on this later)

19 / 70

Local Polynomial Regression

The steps to local polynomial regression are as follows:

1. Fit the local regressions using weights w_i
2. Calculate the residuals $\hat{\epsilon}_i = y_i - \hat{y}_i$
3. Determine the median of the absolute values of the residuals $\hat{q}_{.5}$
4. Find the robustness weights:

$$r_i = B\left(\frac{\hat{\epsilon}_i}{6\hat{q}_{.5}}\right)$$

where:

$$B(u) = \begin{cases} (1 - u^2)^2, & \text{if } |u| < 1; \\ 0, & \text{otherwise.} \end{cases}$$

5. Repeat the loess procedure using weights $r_i w_i$
6. Repeat steps 2-5 until the loess model converges.

20 / 70

Local Polynomial Regression (2)

To estimate the local polynomial regression between y and x , you would estimate:

$$\frac{y_i}{w_i} = \beta_0 + \beta_1 \frac{x_i}{w_i} + \beta_2 \frac{x_i^2}{w_i} + \frac{\varepsilon_i}{w_i}$$

- There are two different versions of this type of regression: Loess and Lowess.
- In **R**, The important difference between these two is that Loess can take multiple predictors (i.e., multiple nonparametric regression) whereas Lowess only takes 1. Further, the user has much more control over loess than lowess, so we spend time on the former.

21 / 70

Choosing the Span

The choice of *span* (i.e., the number of points included in each local model) - this encapsulates the bias-variance tradeoff.

- A bigger span can induce bias which results in a non-parametric estimate that is not faithful to the local patterns in the data
- A smaller span can exhibit considerable variability while sticking very closely to the local pattern in the data. Overfitting is a potential problem here.

Overfitting is not necessarily a problem if we *only* care about the relationship in this sample. However, if we are (either explicitly or implicitly) trying to say something about a population with the sample, then overfitting can be a real problem.

22 / 70

Choosing Polynomial Degree and Weight Function

Polynomial Degree:

- Higher degree polynomials are more likely to overfit the data.
- The most common advice is to set the polynomial degree to 2 and adjust the span to generate the required smoothness of fit.

Weight Function:

- The default in **R** is the *tricube* weight function.
- There is little reason to change this as it generally has a relatively small effect on the overall estimate.

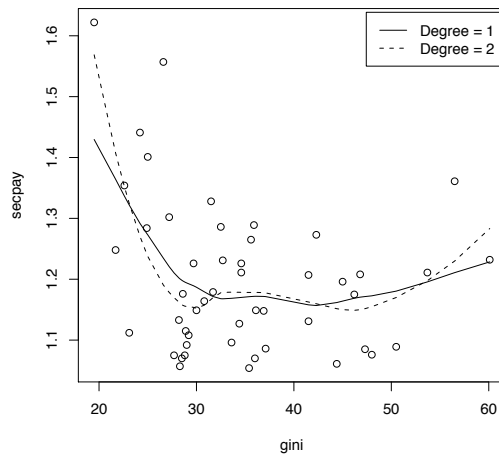
23 / 70

Making the Loess Graph

```
> dat <- read.table("weakliem.txt", header=T)
> out.loess <- loess(secpay ~ gini, data=dat, span=.75, degree=1,
+   family="symmetric")
> plot(secpay ~ gini, data=dat)
> lines(out.loess$fitted[order(dat$gini)] ~
+   dat$gini[order(dat$gini)])
> out.loess <- loess(secpay ~ gini, data=dat, span=.75, degree=2,
+   family="symmetric")
> lines(out.loess$fitted[order(dat$gini)] ~
+   dat$gini[order(dat$gini)], lty=2)
> legend("topright", c("Degree = 1", "Degree = 2"),
+   lty=c(1,2), inset=.01)
```

24 / 70

Loess Graph



25 / 70

Interpretation of Non-Parametric Fits

- Often, we are tempted to impose some meaning on small bumps and dips in the local fit. As Keele (2007) suggests - "it is a temptation analysis should resist."
- It is often useful to consider the overall general pattern in the data and if there appears to be a pattern that can be modeled parametrically - impose that fit and assess the difference between the parametric and non-parametric models (more on this later).

26 / 70

The Linearity Assumption
Testing in GLMs
Local Polynomial Regression

Non-linearity

Assessing Non-linearity
Fixing Non-linearity
Fixing Non-linearity: Polynomials
Maximum Likelihood Transformations

Example

27 / 70

Non-linearity

- The assumption that the average error $E(\varepsilon)$ is everywhere zero implies that the regression surface accurately reflects the dependency of Y on the X 's
- We can see this as linearity in the broad sense
 - *i.e.*, non-linearity refers to a partial relationship between two variables that is not summarized by a straight line, but it could also refer to situations when two variables specified to have additive effects actually interact.
- Violating this assumption implies that the model fails to account for a systematic pattern between Y and the X 's
 - Often models characterized by this violation will still provide a useful approximation of the pattern in the data, but they can also be misleading
- It is impossible to directly view the regression surface when more than two predictors are specified, but we can employ *partial residual plots* to assess non-linearity.

28 / 70

The Linearity Assumption
 Testing in GLMs
 Local Polynomial Regression

Non-linearity
 Assessing Non-linearity
 Fixing Non-linearity
 Fixing Non-linearity: Polynomials
 Maximum Likelihood Transformations

Example

Plots for assessing non-linearity

- Scatterplot matrices are useful for preliminary assessments of the relationship between several variables in a multiple regression model, but can be misleading because they plot the the marginal rather than partial relationships between Y and each X (i.e., they do not control for the other X 's)
- Conditioning plots are better, but still have trouble if there are too many X 's
- Partial-regression plots (or added-variable plots) are not very useful either because they are unable to distinguish between monotone linearity (which can often be corrected with a simple transformation) and non-monotone non-linearity (which cannot be corrected with a transformation)
- Partial-residual plots, however, can reveal both monotone and non-monotone linearity

Failure of partial regression plots (AV plots)

- The two plots in column (a) represent the scatterplot of Y and X and the plot of the residuals E and X for one regression; column (b) gives the same for another regression.
- Notice that although column (a) is characterized by non-monotone non-linearity, this was not picked up in the simple residual plot, where the pattern is identical to that in (b) which is characterized by monotone non-linearity.
- Only (b) can be transformed to satisfy the linearity requirement

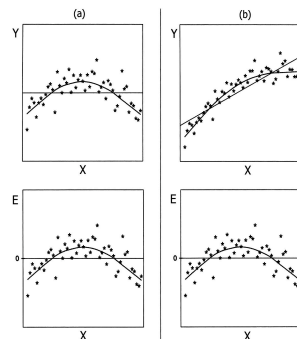


Figure 12.5 from Fox (1997)

Partial-Residual Plots (Component-plus-residual plots)

- The partial residual for the j^{th} explanatory variable from a multiple regression is

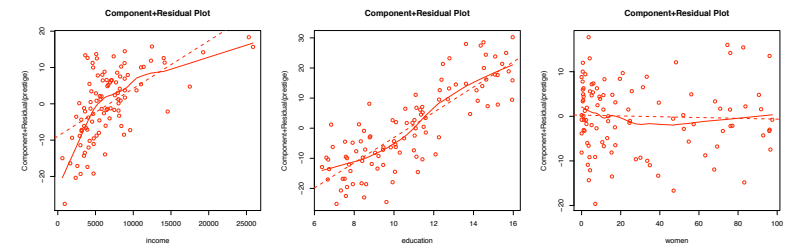
$$E_i^{(j)} = E_i + B_j X_{ij}$$
 - This simply adds the linear component of the partial regression between Y and X_j (which may be characterized by a non-linear component) to the least squares residuals
- The “partial residuals” $E^{(j)}$ are plotted versus X_j , meaning that B_j is the slope of the multiple simple regression of $E^{(j)}$ on X_j
 - A non-parametric smooth helps assess whether the linear trend adequately captures the partial relationship between Y and X .

Example of partial residual plots (1): The Canadian Prestige Data

```
> data(Prestige)
> Prestige.model<-lm(prestige ~ income + education +
+ women, data=Prestige)
> library(car)
> cr.plot(Prestige.model, "income")
> cr.plot(Prestige.model, "education")
> cr.plot(Prestige.model, "women")
```

33 / 70

Example of partial residual plots (2): The Canadian Prestige Data



- The plot for income suggests a power transformation down the ladder of powers; for education the departure from linearity isn't problematic; for % women, there appears to be no relationship

34 / 70

The Linearity Assumption
Testing in GLMs
Local Polynomial Regression

Non-linearity
Assessing Non-linearity
Fixing Non-linearity
Fixing Non-linearity: Polynomials
Maximum Likelihood Transformations

Example

35 / 70

Handling Non-linearity: Common Strategies

Simple, monotone

- Transformations of Y and/or X

Complicated Non-linearity

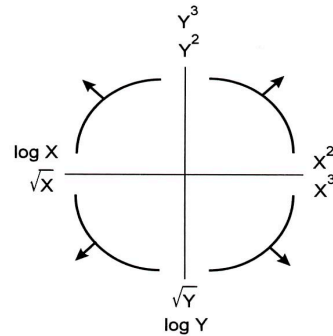
1. Divide X into categories, and employ a set of dummy regressors
 - Will capture the large differences between groups, but could miss substantial differences within them
2. Polynomial Regression
 - If pattern has too many turns, polynomials tend to oversmooth peaks
 - Also, problem with collinearity, but can be countered with orthogonal polynomials
3. Regression Splines
4. Nonparametric regression (including generalized additive models)

36 / 70

Transformable Non-linearity: Bulging rule revisited

- The direction of the bulge indicates the appropriate type of power transformation for Y and/or X
- A bulge to the top left of the scatterplot suggests transforming Y up the ladder of powers and/or X down the ladder of powers will straighten the relationship

Figure 4.6 from Fox (1997)



37 / 70

The Linearity Assumption Testing in GLMs Local Polynomial Regression

Non-linearity

- Assessing Non-linearity
- Fixing Non-linearity
- Fixing Non-linearity: Polynomials
- Maximum Likelihood Transformations

Example

38 / 70

Polynomial Regression

- Two or more regressors of ascending power (i.e., linear, quadratic and cubic terms) are used to capture the effects of a single variable
 - For every bend in the curve, we add another term to the model, going up in power each time
- The terms fit a non-linear function of the explanatory variable X , but the parameters enter the formula in a linear fashion - Y is predicted by a linear combination of parameter estimates times the values of X
 - In other words, polynomial models are linear in the parameters even though they are non-linear in the variables

Order	Equation
First	$Y = \alpha + \beta_1 X$
Second	$Y = \alpha + \beta_1 X + \beta_2 X^2$
Third	$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$

39 / 70

Polynomial equations: How to choose the order

- It is initially useful to look at the bends in a smooth of the scatterplot or partial residual plot
 - If there is only one, a second order polynomial should be tried. For each extra bulge, we go up one in order
- A good strategy is to start with one more than you think the model needs and drop the term if it is not statistically significant
- Incremental F -tests (or analysis of deviance) can be used to help pick the "right" order to use in the equation
 - If the term is not statistically significant, it is usually advisable to delete the term from the model - we want as few order terms as possible
 - For orthogonal polynomials, t -tests can be used
- If the order is too high, however, the results will not be easy to interpret (higher than third order is rarely used)

40 / 70

Orthogonal Polynomials: Removing Collinearity

- Polynomial regressors $X^1, X^2, X^3, \dots, X^{m-1}$ are usually highly correlated, making t -tests for individual terms invalid
- It is not necessary to remove the correlation, but it has 2 benefits:
 - It stabilizes the estimates (recall that the VIF $(1 - R_j^2)^{-1}$ tells us how much the standard errors are affected by multicollinearity)
 - Determining the appropriate order is simple because individual t -tests can be used
- It is possible to orthogonalize the power regressors before fitting the model
 - Regress each of the higher order terms separately on X
 - Let X^{2*} represent the residual from the X^2 model, in general, let X^{m*} represent the residual from the X^m model
 - In the polynomial regression model, replace the original terms X^2, X^3, \dots with the new variables, X^{2*}, X^{3*}, \dots as the set of regressors

41 / 70

Orthogonal Polynomials in R Example: Prestige Data

- One can fit a polynomial regression by calculating the regressors individually and adding them to the regression equation - i.e., calculate and add a quadratic term X^2 and a cubic term X^3 manually.
- Orthogonal Polynomials* can be added in a much more simple - and better - way in **R**, however, by specifying a `poly` argument to the variable
 - The order of the polynomial is specified after the variable name

42 / 70

Regression Output

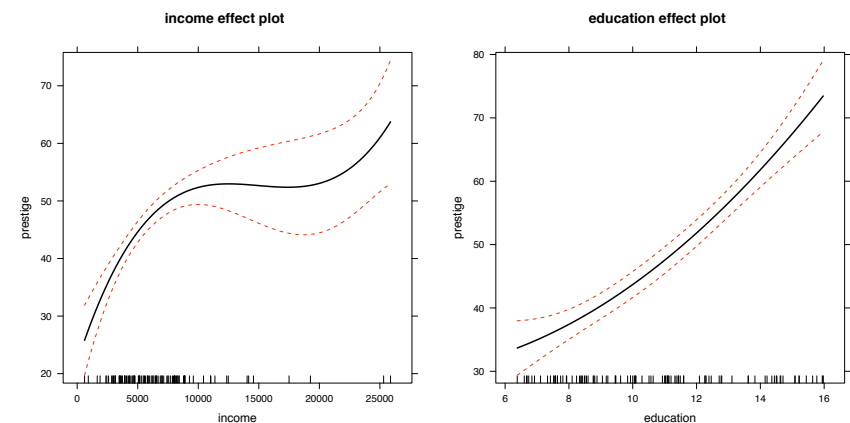
(Intercept)	46.83*
	(0.70)
poly(income, 3)1	54.22*
	(9.30)
poly(income, 3)2	-26.64*
	(7.34)
poly(income, 3)3	21.33*
	(7.52)
poly(education, 2)1	113.85*
	(9.45)
poly(education, 2)2	16.34*
	(7.60)
<i>N</i>	102
R^2	0.84
adj. R^2	0.83
Resid. sd	7.07

Standard errors in parentheses
* indicates significance at $p < 0.05$

- Since orthogonal polynomials were used, the t -test for the individual parameters is all that is needed. An F -test will show nothing different
- Nonlinear effects are difficult to comprehend in numerical form. Graphing the fitted values provides a much better alternative.

43 / 70

Effect Displays for Income and Education



44 / 70

The Linearity Assumption
Testing in GLMs
Local Polynomial Regression

Non-linearity
Assessing Non-linearity
Fixing Non-linearity
Fixing Non-linearity: Polynomials
Maximum Likelihood Transformations

Example

45 / 70

Maximum Likelihood Transformation Methods

- Although the *ad hoc* methods for assessing non-linearity are usually effective, there are more sophisticated techniques based on maximum likelihood estimation
- These techniques embed the usual multiple-regression model in a more general non-linear model that contains (a) parameter(s) for the transformation(s)
 - The transformation parameter λ is estimated simultaneously with the usual regression parameters by maximizing the likelihood and this obtaining MLEs: $\mathcal{L}(\lambda, \alpha, \beta_1, \dots, \beta_k, \sigma_\varepsilon^2)$
 - If $\lambda = \lambda_0$ (i.e., there is no transformation), a likelihood ratio test, Wald test, or score test of $H_0 : \lambda = \lambda_0$ can assess whether the transformation is required
- If several variables need to be transformed, several such parameters need to be included

46 / 70

Maximum Likelihood Methods: Box-Cox Transformation of Y

- The Box-Cox transformation of Y functions to *normalize the error distribution, stabilize the error variance and straighten the relationship* of Y to the X 's
- The general Box-Cox model is:

$$Y_i^\lambda = \alpha + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ and

$$Y_i^\lambda = \begin{cases} \frac{Y_i^\lambda - 1}{\lambda}, & \text{for } \lambda \neq 0 \\ \log_e Y_i, & \text{for } \lambda = 0 \end{cases}$$

- If $\lambda=1$, no transformation is necessary
- Note that all of the Y_i *must* be positive

47 / 70

More Box-Cox Transformation

- A simple way to estimate the Box-Cox transformation is to use a constructed variable regression model, where the constructed variable is:

$$G_i = Y_i \left[\log_e \left(\frac{Y_i}{\tilde{Y}} \right) - 1 \right]$$

where \tilde{Y} is the geometric mean of Y_i :

$$\tilde{Y} \equiv \left(\prod_{i=1}^n Y_i \right)^{\frac{1}{n}}$$

- The augmented regression, include the constructed variable is:

$$Y_i = \alpha + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \phi G_i + \varepsilon_i$$

- The suggested transformation from this model is simply $\lambda = 1 - \phi$
- A t -test for the $H_0 : \phi = 0$, namely $t_0 = \frac{\hat{\phi}}{SE(\hat{\phi})}$ assesses the need for a transformation

48 / 70

Box-Cox Transformation Example: Ornstein Data (1)

```

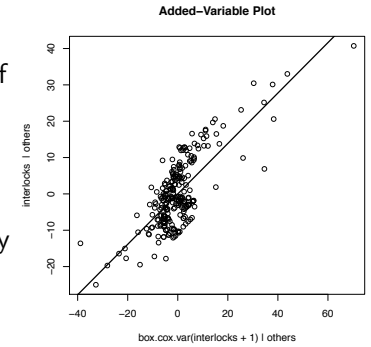
> data(Ornstein)
> mod <- lm(interlocks ~ box.cox.var(interlocks+1) + assets +
+ sector + nation, data=Ornstein)
> summary(mod)
Call:
lm(formula = interlocks ~ box.cox.var(interlocks + 1) + assets +
    sector + nation, data = Ornstein)

Residuals:
    Min       1Q   Median       3Q      Max
-17.1411  -4.6491  -0.1669   4.9191  13.4423

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.312e+01  1.034e+00  12.693 < 2e-16 ***
box.cox.var(interlocks + 1)  6.938e-01  3.914e-02  17.727 < 2e-16 ***
assets         -3.565e-05  6.224e-05  -0.573  0.56734
sectorBNK      -2.022e+00  3.964e+00  -0.510  0.61039
sectorCDN     -3.964e+00  3.092e+00  -1.282  0.20116
sectorFIN       8.043e+00  1.738e+00  4.628 6.13e-06 ***
sectorHLD     -8.327e-03  2.619e+00  -0.003  0.99747
sectorMAN     -1.621e-01  1.352e+00  -0.120  0.90466
sectorMER       1.947e+00  1.724e+00  1.129  0.25995
sectorMIN       4.285e+00  1.352e+00  3.169  0.00173 **
sectorTRN       5.354e+00  1.805e+00  2.966  0.00333 **
sectorWOD       3.827e+00  1.769e+00  2.163  0.03154 *
nationOTH      -2.380e-01  1.763e+00  -0.135  0.89274
    
```

Box-Cox Transformation Example: Ornstein Data (2)

- The coefficient for the Box-Cox variable in the model is 0.69, suggesting that a transformation of Y_i to the power of $1-0.69=0.31$ is needed
- The added-variable plot allows us to see that the choice of transformation was not influence by only a few cases - it seems to be needed throughout most of the data



Box-Tidwell Transformation of the X's (1)

- Maximum likelihood can also be used to find an appropriate linearizing transformation for the X variables
- The Box-Tidwell model is a non-linear model that estimates transformation parameters for the X 's simultaneously with the regular parameters

$$Y_i = \alpha + \beta_1 X_{i1}^{\gamma_1} + \dots + \beta_k X_{ik}^{\gamma_k} + \varepsilon_i$$

where the errors are iid: $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I}_n)$ and the X_{ij} are positive

- Explicit in this model is a power transformation of each of the X 's
 - Of course, we would not want to transform dummy variables and the like, so we should not attempt to estimate transformation parameters for them

Box-Tidwell Transformation of the X's (2)

The Box and Tidwell procedure yields a constructed variable diagnostic in the following way:

1. Regress Y on the X 's and obtain A, B_1, \dots, B_k .
2. Regress Y on the X 's and the constructed variables $X_1 \log_e X_1, \dots, X_k \log_e X_k$ to obtain $A', B'_1, \dots, B'_k, D_1, \dots, D_k$
3. The constructed variables are used to assess the need for a transformation of X_j by testing the null hypothesis $H_0 : \delta_j = 0$ where $D_j = \hat{\delta}_j$
4. A preliminary estimate of the transformation parameter γ_j is given by

$$\tilde{\gamma}_j = 1 + \frac{D_j}{B_j}$$

where B_j is the coefficient on X_j from the original equation in step 1

5. Steps 1,2, and 4 are iterated until the transformation parameters converge

Box-Tidwell transformation Example: Prestige Data

```
> box.tidwell(prestige ~ income + education,
+             ~poly(women, 2), data=Prestige)
```

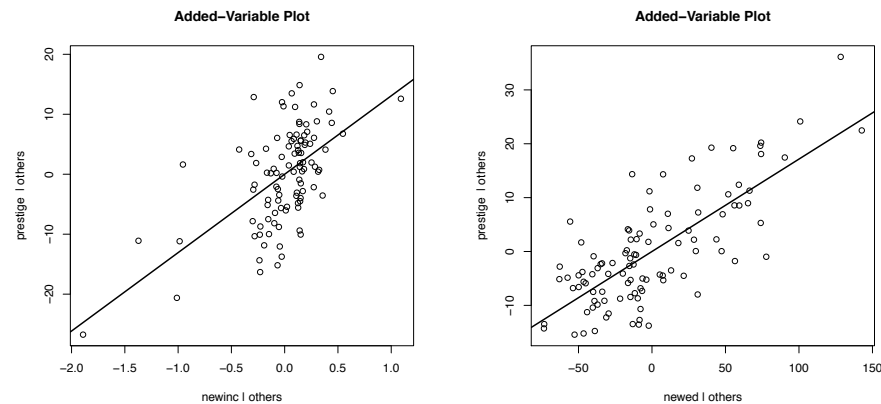
```
           income education
Initial Power -0.91030  2.24354
Score Statistic -5.30129  2.40556
p-value      0.00000  0.01615
MLE of Power  -0.03777  2.19283
```

iterations = 12

- A quadratic partial regression is included for women because we saw earlier that this was needed
- The statistically significant score tests indicate that transformations are needed for both variables
- The MLE of Power suggests that income should be transformed by a power of -0.037 (suggesting the log would work well) and education by a power of 2.19, suggesting that education² would suffice

53 / 70

Added Variable Plots for the Box-Tidwell constructed Variables (1)



- The graphs here both provide general support for the transformations found from the Box-Tidwell Model

54 / 70

Working through the Duncan example

Remember from yesterday that we saw the following in the Duncan model:

- The non-linearity between income and prestige was simple - monotone, so it can likely be transformed
- If any non-linearity exists between education and prestige, it looks like a cubic function, though the relationship may not be significantly non-linear.
- It looks like the variable women has no effect on prestige; however, if any non-linearity exists, it looks like a quadratic form.

55 / 70

Income

If we simply want to include the logged version of income, we can do so as follows:

(Intercept)	-110.97*
	(14.84)
log(income)	13.44*
	(1.91)
education	3.73*
	(0.35)
women	0.05
	(0.03)
N	102
R ²	0.84
adj. R ²	0.83
Resid. sd	7.09

Standard errors in parentheses

* indicates significance at $p < 0.05$

56 / 70

Transform of Income

We could also use the Box-Tidwell transformation of income to see what transformation parameter it suggests.

```
> box.tidwell
```

```
      1
Initial Power  -0.74533
Score Statistic -4.83382
p-value        0.00000
MLE of Power   0.08073
```

```
iterations = 10
```

This suggests that a power transformation of 0.08073 is the right number (rather than 0, which we used above). We can then use this to estimate the model:

57 / 70

Transformation of Income

(Intercept)	-165.67*
	(22.44)
newinc	85.26*
	(12.09)
education	3.66*
	(0.36)
women	0.05
	(0.03)
<i>N</i>	102
<i>R</i> ²	0.84
adj. <i>R</i> ²	0.83
Resid. sd	7.08

Standard errors in parentheses

* indicates significance at $p < 0.05$

58 / 70

Diagnostics for Education

We argued yesterday that the relationship is very likely linear, but there was a little evidence that it might be polynomial. Let's see what it looks like after we fixed the income relationship.

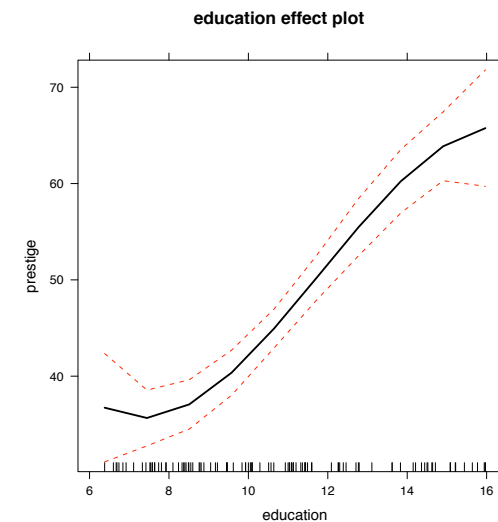
(Intercept)	-116.71*
	(24.68)
newinc	80.63*
	(11.95)
poly(education, 3)1	103.09*
	(9.71)
poly(education, 3)2	11.62
	(7.11)
poly(education, 3)3	-13.76
	(7.18)
women	0.04
	(0.03)
<i>N</i>	102
<i>R</i> ²	0.85
adj. <i>R</i> ²	0.84
Resid. sd	6.92

Standard errors in parentheses

* indicates significance at $p < 0.05$

59 / 70

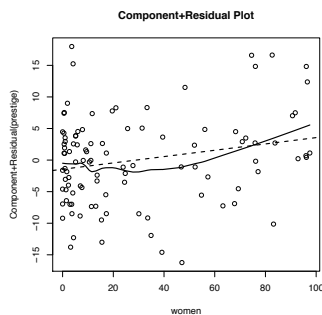
Polynomial Effect of Education



60 / 70

Women and Prestige

Now that we've fixed the other two relationships, we can look at the women component+residual plot.



61 / 70

Polynomial for Women

It looks as if anything non-linear is going on here, it is quadratic, but it is much more likely that nothing is going on.

(Intercept)	-164.74*
	(21.45)
newinc	85.28*
	(11.87)
education	3.71*
	(0.35)
poly(women, 2)1	15.91
	(9.43)
poly(women, 2)2	15.14*
	(6.98)
N	102
R ²	0.84
adj. R ²	0.84
Resid. sd	6.95

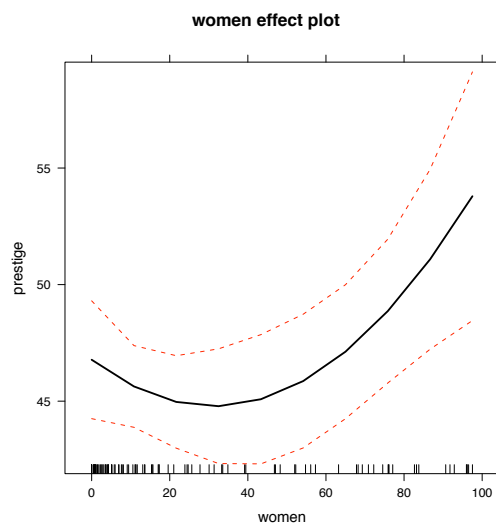
Standard errors in parentheses

* indicates significance at $p < 0.05$

It looks like there is a significant quadratic relationship between income and women.

62 / 70

Polynomial Effect of Women



63 / 70

Re-checking Income

Since we've made some changes, we can re-check the relationship with income:

```
> box.tidwell(prestige ~ income, ~education +  
+ poly(women, 2), data=Prestige)
```

```
1  
Initial Power -0.85314  
Score Statistic -5.03000  
p-value 0.00000  
MLE of Power 0.01611
```

```
iterations = 10
```

Now, the transformation power is very close to zero. We could just use the log-transform, which will be very similar and will be easier to interpret.

64 / 70

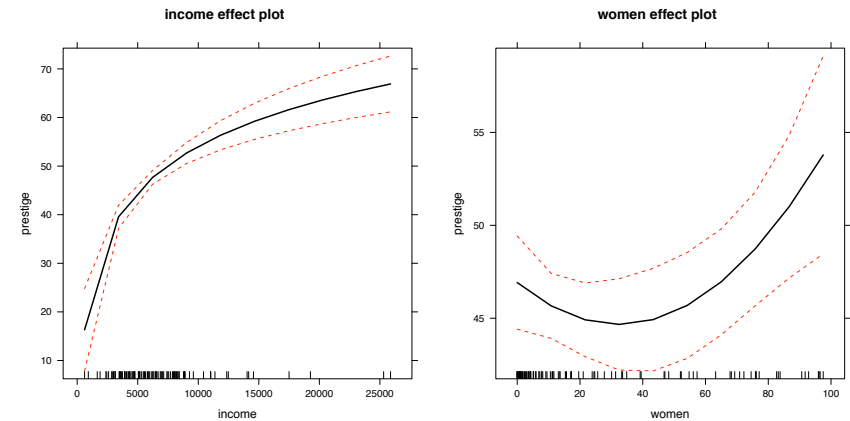
Final Model

(Intercept)	-110.60*
log(income)	13.50*
education	3.77*
poly(women, 2)1	15.09
poly(women, 2)2	15.87*
<i>N</i>	102
<i>R</i> ²	0.84
adj. <i>R</i> ²	0.84
Resid. sd	6.95

Standard errors in parentheses
* indicates significance at $p < 0.05$

65 / 70

Effects from Final Model



66 / 70

Transformations and Interpretation

Often as with income (above) we have to transform a variable to linearize its relationship with y . However, it is worth putting your interpretations back in the metric of y and x . This is easy for log transformations:

- $\widehat{\log(y)} = \alpha + \beta x$: when x increases by one *unit*, on average y increases by $100 \times \beta$ percent.
- $\hat{y} = \alpha + \beta \log(x)$: when x increases by 1 percent, on average y increases by $\frac{\beta}{100}$ units.
- $\widehat{\log(y)} = \alpha + \beta \log(x)$: when x increases by 1 percent, on average y increases by β percent

67 / 70

The Linearity Assumption

Testing in GLMs

Local Polynomial Regression

Non-linearity

Assessing Non-linearity

Fixing Non-linearity

Fixing Non-linearity: Polynomials

Maximum Likelihood Transformations

Example

68 / 70

Readings

Today: Linearity Diagnostics

- * Bowman and Azzalini (1997) Chapter 1
- * Fox (2008) Chapters 4 & 12 (Sections 12.3-12.5)
- * Fox (2002) Chapter 3
- * Jacoby (1999)
 - Cook and Weisberg (1999) Chapter 16
 - Box and Tidwell (1962)
 - Breiman and Friedman (1985a,b), Pregibon and Vardi (1985), Buja and Kass (1985), Fowlkes and Kettering (1985)

Tomorrow: Inference for Non-Parametric Models, Splines and GAM's

- * Fox (2008) Chapters 17 & 18
 - Fox (2000b,a)
- * Keele (2008) Chapters 2 & 3
- * Fox (2000a)
- * Keele (2008) Chapters 4-6
 - Wood (2006)
 - Hastie and Tibshirani (1990)

- Bowman, Adrian and Adelchi Azzalini. 1997. *Applied Smoothing Techniques for Data Analysis*. Oxford University Press.
- Box, George and P.W. Tidwell. 1962. "Transformation of the Independent Variables." *Technometrics* 4:531–550.
- Breiman, Leo and Jerome H. Friedman. 1985a. "Estimating Optimal Transformations for Multiple Regression and Correlation." *Journal of the American Statistical Association* 80(391):580–598.
- Breiman, Leo and Jerome H. Friedman. 1985b. "Estimating Optimal Transformations for Multiple Regression and Correlation: Rejoinder." *Journal of the American Statistical Association* 80(391):614–619.
- Buja, Andreas and Robert E. Kass. 1985. "Estimating Optimal Transformations for Multiple Regression and Correlation: Comment." *Journal of the American Statistical Association* 80(391):602–607.
- Cook, R. Dennis and Sanford Weisberg. 1999. *Applied Regression Including Computing and Graphics*. New York: Wiley & Sons, Inc.
- Fowlkes, E.B. and J.R. Kettering. 1985. "Estimating Optimal Transformations for Multiple Regression and Correlation: Comment." *Journal of the American Statistical Association* 80(391):607–613.
- Fox, John. 2000a. *Multiple and Generalized Nonparametric Regression*. Thousand Oaks: Sage.
- Fox, John. 2000b. *Nonparametric Simple Regression*. Thousand Oaks: Sage.
- Fox, John. 2002. *An R and S-Plus Companion to Applied Regression*. Thousand Oaks: Sage Publications.
- Fox, John. 2008. *Applied Regression Analysis and Generalized Linear Models, 2nd edition*. Thousand Oaks, CA: Sage, Inc.
- Hastie, Trevor J. and Robert J. Tibshirani. 1990. *Generalized Additive Models*. Boca Raton, FL: Chapman & Hall/CRC.
- Jacoby, William G. 1999. "Levels of Measurement and Political Research: An Optimistic View." *American Journal of Political Science* 43(1):271–301.
- Keele, Luke J. 2008. *Semiparametric Regression for the Social Sciences*. New York: Wiley & Sons, Inc.
- Pregibon, Daryl and Yehuda Vardi. 1985. "Estimating Optimal Transformations for Multiple Regression and Correlation: Comment." *Journal of the American Statistical Association* 80(391):598–601.
- Wood, Simon. 2006. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.