Statistical Techniques II

EXST7015

Curvilinear Regression

Curvilinear Regression

- **As the name implies, these are regressions that fit curves.**
- **However, the regressions we will discuss are also linear models, so most of the techniques and SAS procedures we have discussed will still be relevant.**

Curvilinear Regression *(continued)*

- **We will discuss two basic types of curvilinear model.**
  - **The first are polynomial regressions. These are an extraordinarily flexible family of curves that will fit almost anything. Unfortunately, they rarely have a good, interpretation of the parameter estimates.**
  - **The second are models that are not linear, but that can be "linearized" by transformation. These models are referred to as "intrinsically linear", because after transformation they are linear, often SLR.**

Polynomial Regression

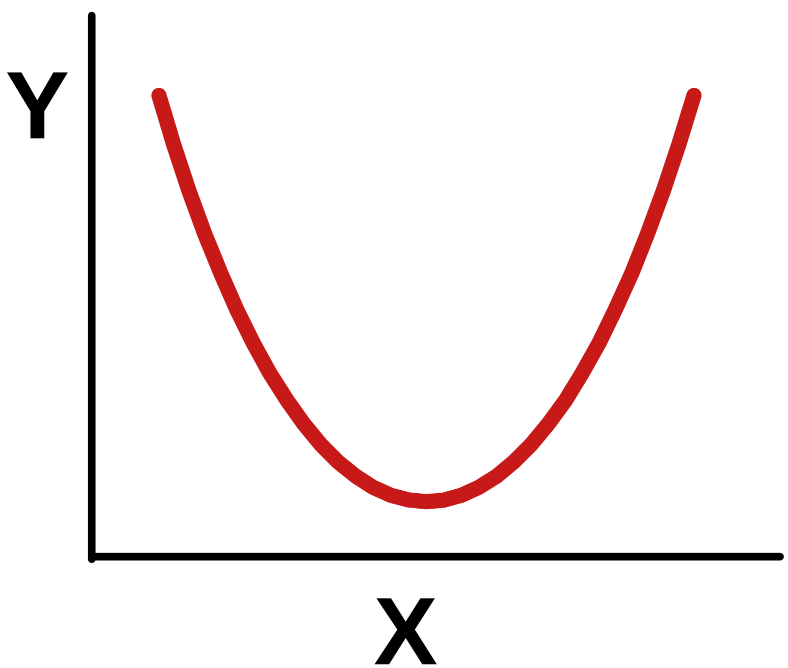
- **Polynomial regressions are multiple regressions that use power terms of the X_i variable to fit curves. As long as the value of the power is known, the model is linear.**
- **Only a single X_i is needed (though more can be used).**
- **The assumptions are the same as for any other multiple regression.**

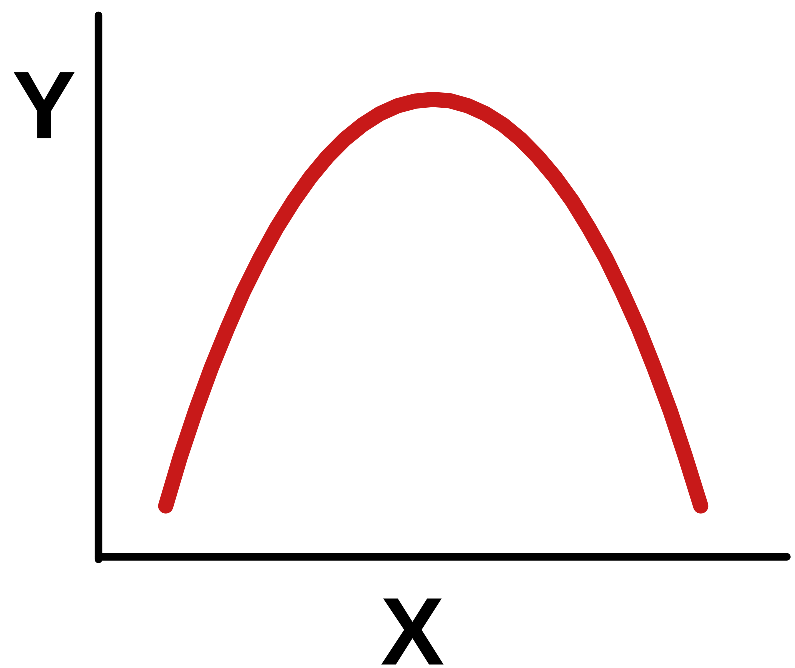
Polynomial Regression *(continued)*

- **Polynomial regressions are of the form**
  - $Y_i = b_0 + b_1X_i + b_2X^2_i + b_3X^3_i + ... + b_kX^k_i + e_i$
- **The simplest in this family of models is the "linear", which is just a simple linear regression. Polynomials proceed,**
  - **Quadratic** $Y_i = b_0 + b_1X_i + b_2X^2_i + e_i$
  - **Cubic** $Y_i = b_0 + b_1X_i + b_2X^2_i + b_3X^3_i + e_i$
  - **Quartic** $Y_i = b_0 + b_1X_i + b_2X^2_i + b_3X^3_i + b_4X^4_i + e_i$
  - **Quintic, etc.**

Polynomial Regression *(continued)*

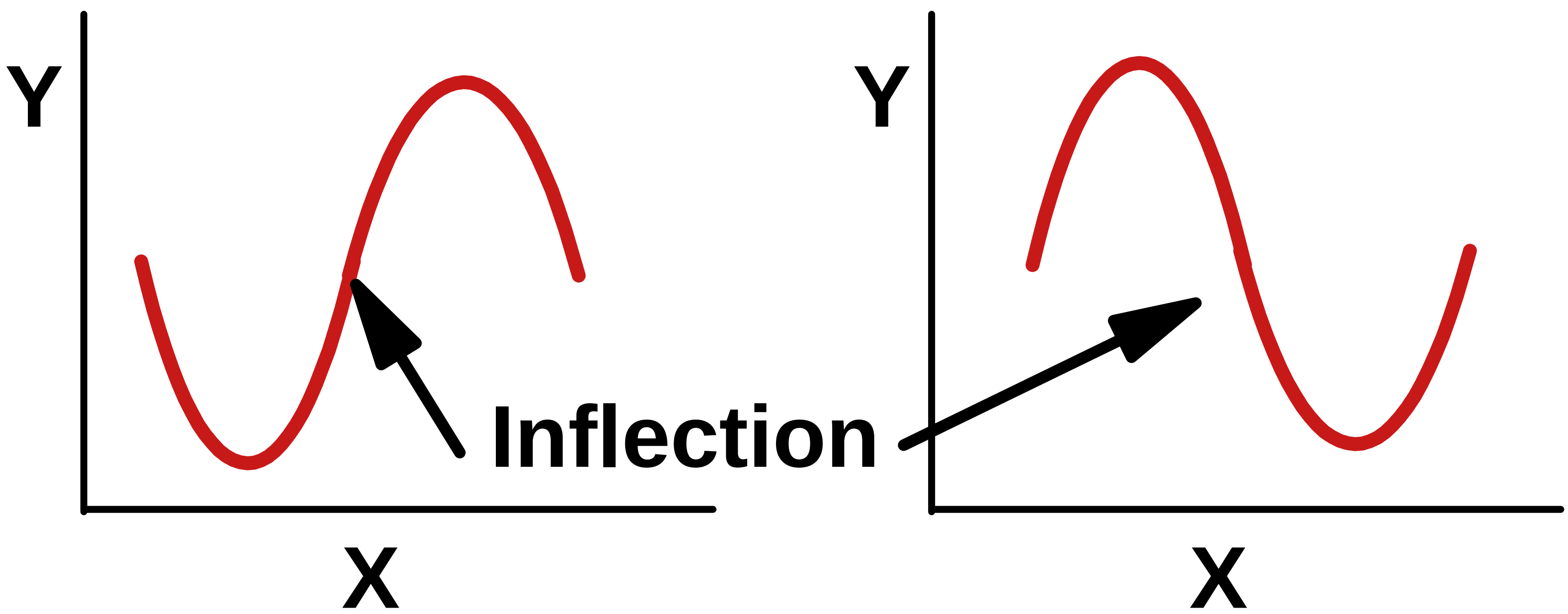
- **The quadratic fits a simple parabolic curve. Either concave or convex, depending on the sign on the regression coefficient.**





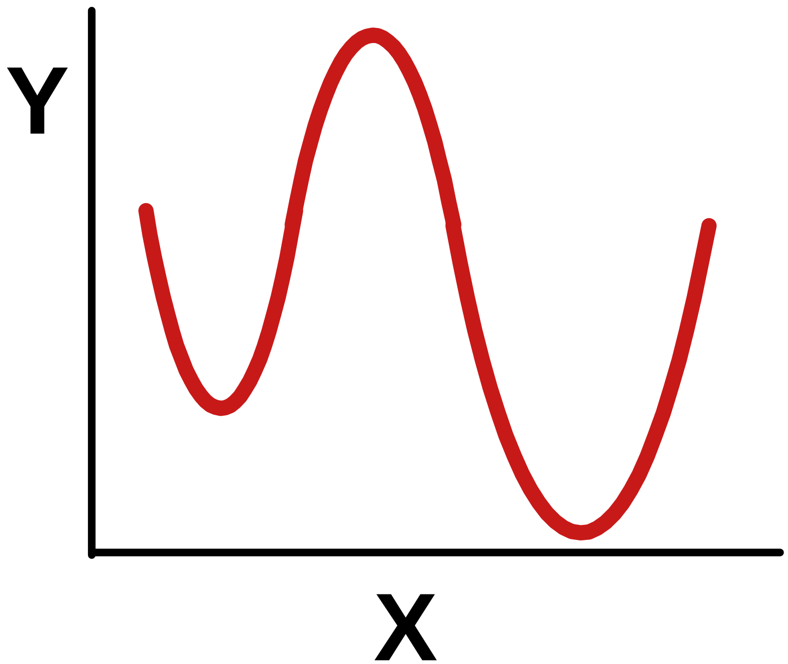
Polynomial Regression *(continued)*

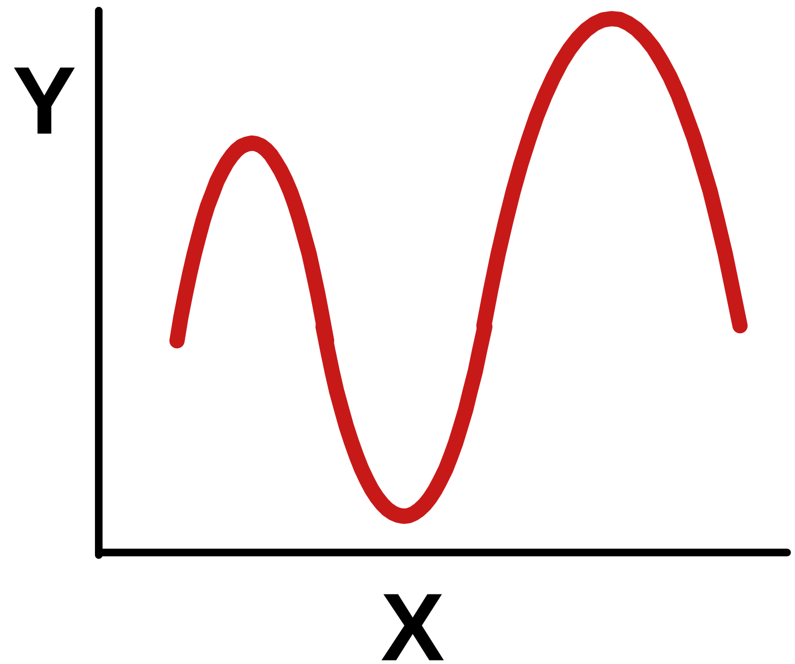
- **The cubic fits parabolic curves with an inflection. The inflection does not always occur within the range of the data.**



Polynomial Regression *(continued)*

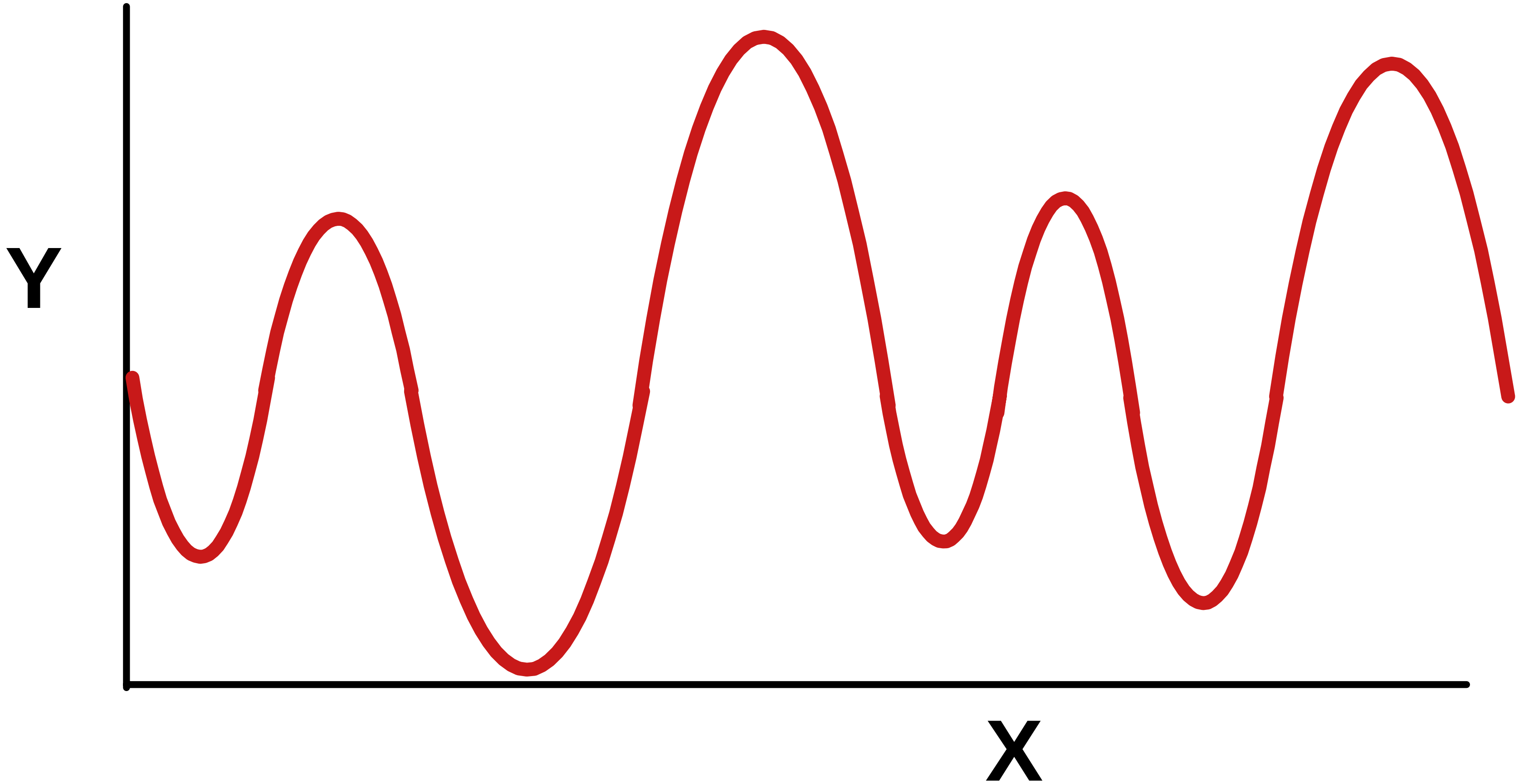
- **The quartic polynomial adds another inflection, and another peak or valley (maximum or minimum point). These are not usually symmetric.**





Polynomial Regression *(continued)*

- **The same pattern continues for larger models.**

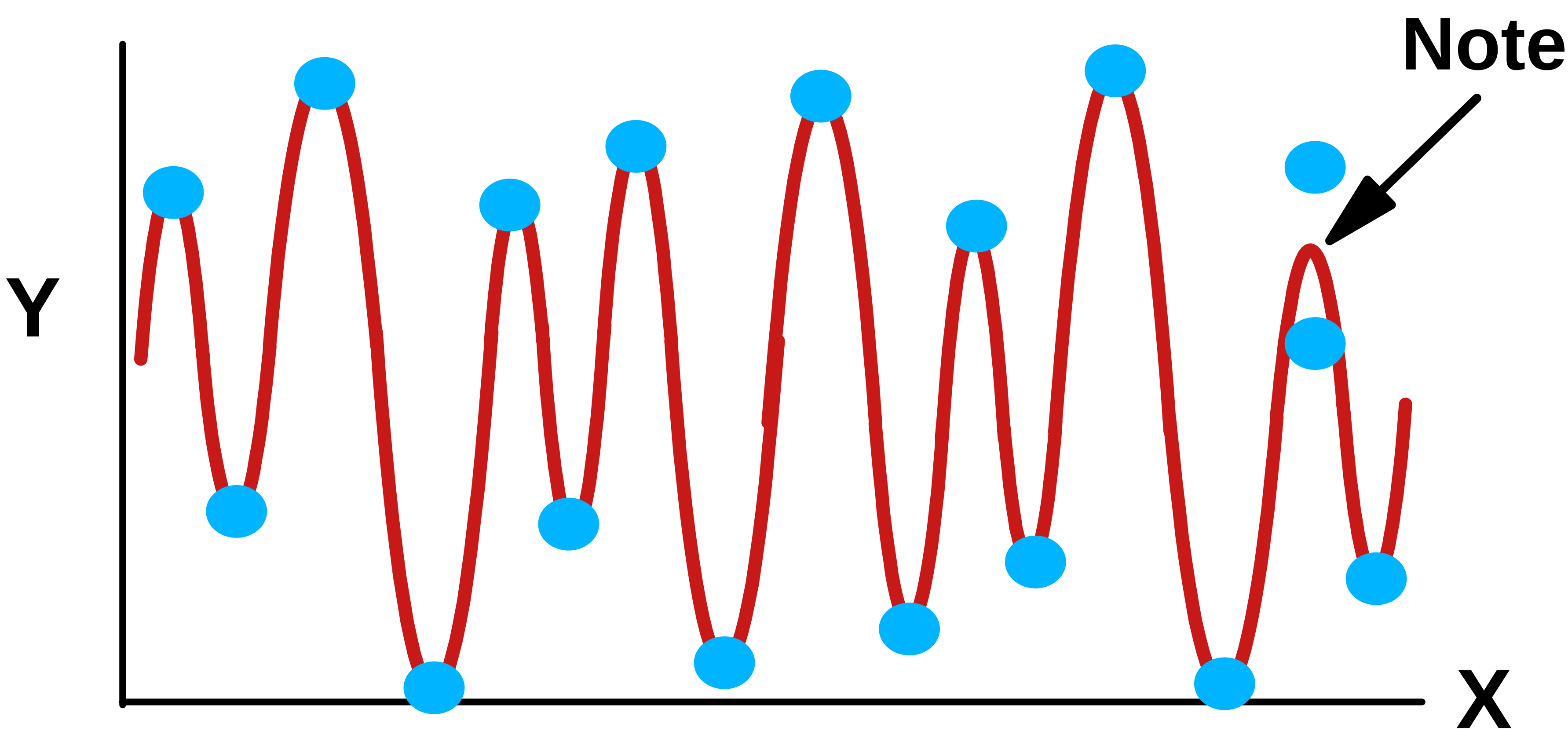


Polynomial Regression *(continued)*

- **What good are polynomials? They will fit anything. In fact, if no two X values are repeated, then a large enough polynomial will go through every observation.**
  - **A SLR exactly fits 2 points**
  - **A quadratic polynomial will exactly fit 3 points**
  - **A cubic will pass through each of 4 points**
  - **For n points, n-1 polynomial terms will pass through every point.**

Polynomial Regression *(continued)*

- **Sounds like a good thing? Only if you want to fit random scatter. How would you interpret the graph below?**

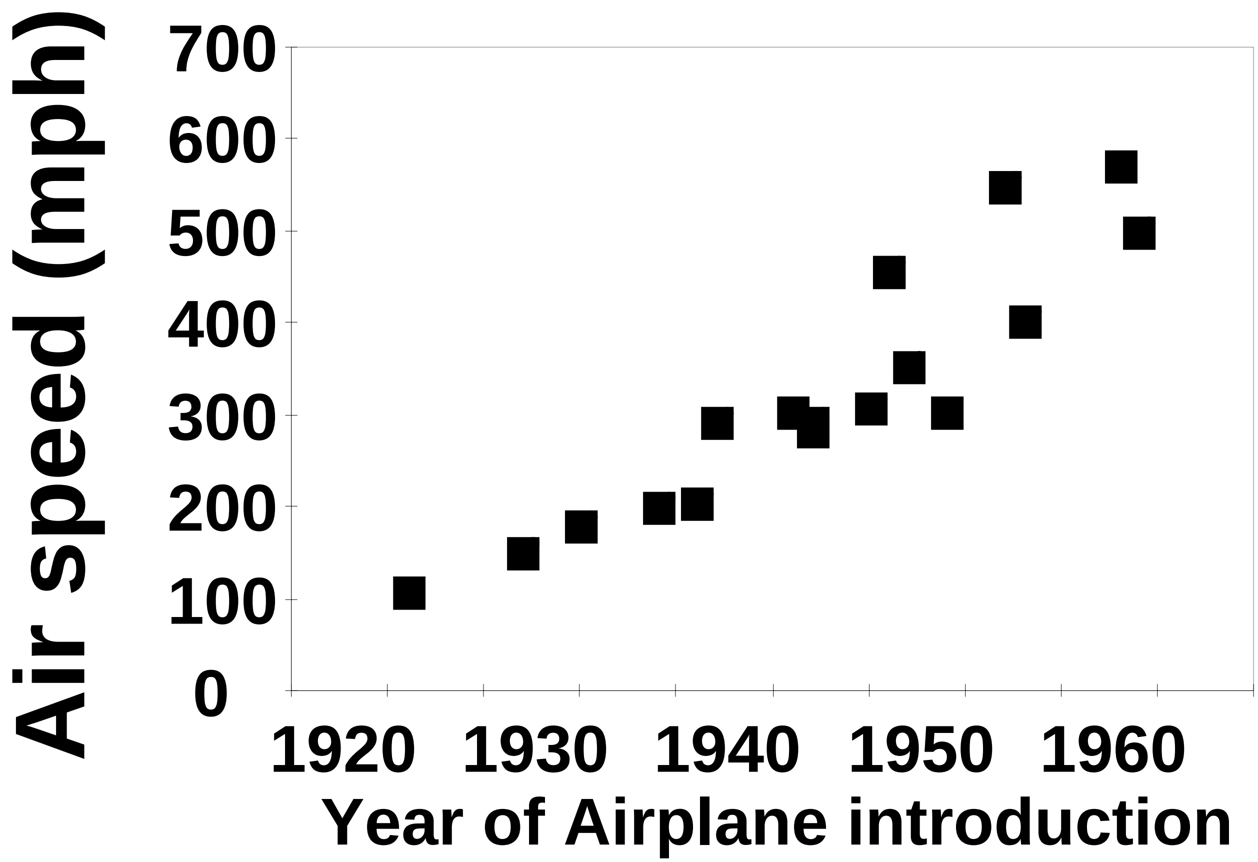


Polynomial Regression *(continued)*

- **Recall the air speed example from The Science of Flight by Peter P. Wagener, Am Sci, volume 74,(3),May-June 1986, page 274.**
- **We previously fitted an exponential growth curve with good results.**
- **Air speed example. I digitized the following data from a graph and omitted values after 1963.**

Polynomial Reg. - Example 1

YEAR	SPEED	AIRCRAFT
1926	108	Ford 5-AT
1932	150	247D
1935	179	DC-3
1939	200	307 Strat
1941	204	DC-4
1942	292	L-749
1946	304	DC-6
1947	283	Convair 2
1947	292	377 strat
1950	308	DC-6B
1952	354	DC-7
1954	304	Viscount
1951	458	Comet
1958	404	L188A Ele
1957	550	707/DC-8
1964	500	BAC1-11-2
1963	571	727



Polynomial Reg. - Example 1 *(continued)*

- **We will now proceed to fit a polynomial model (quadratic) to the data. This was the model chosen by the author.**

Polynomial Reg. - Example 1 *(continued)*

- The SAS statements are,
  - PROC GLM DATA=ONE;
  - MODEL SPEED = YEAR YEAR*YEAR;
  - RUN;
- Note that I used Year*Year to fit YEAR squared. You can do this in GLM, but not in PROC REG.
- The GLM output follows.

Polynomial Reg. - Example 1 *(continued)*

- **PROC GLM on airspeed example.**

Dependent Variable: SPEED

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	2	405441.6302	202720.8151	74.75	<.0001
Error	17	46104.9198	2712.0541		
Corrected Total	19	451546.5500			

R-Square	Coeff Var	Root MSE	SPEED Mean
0.897896	14.56099	52.07739	357.6500

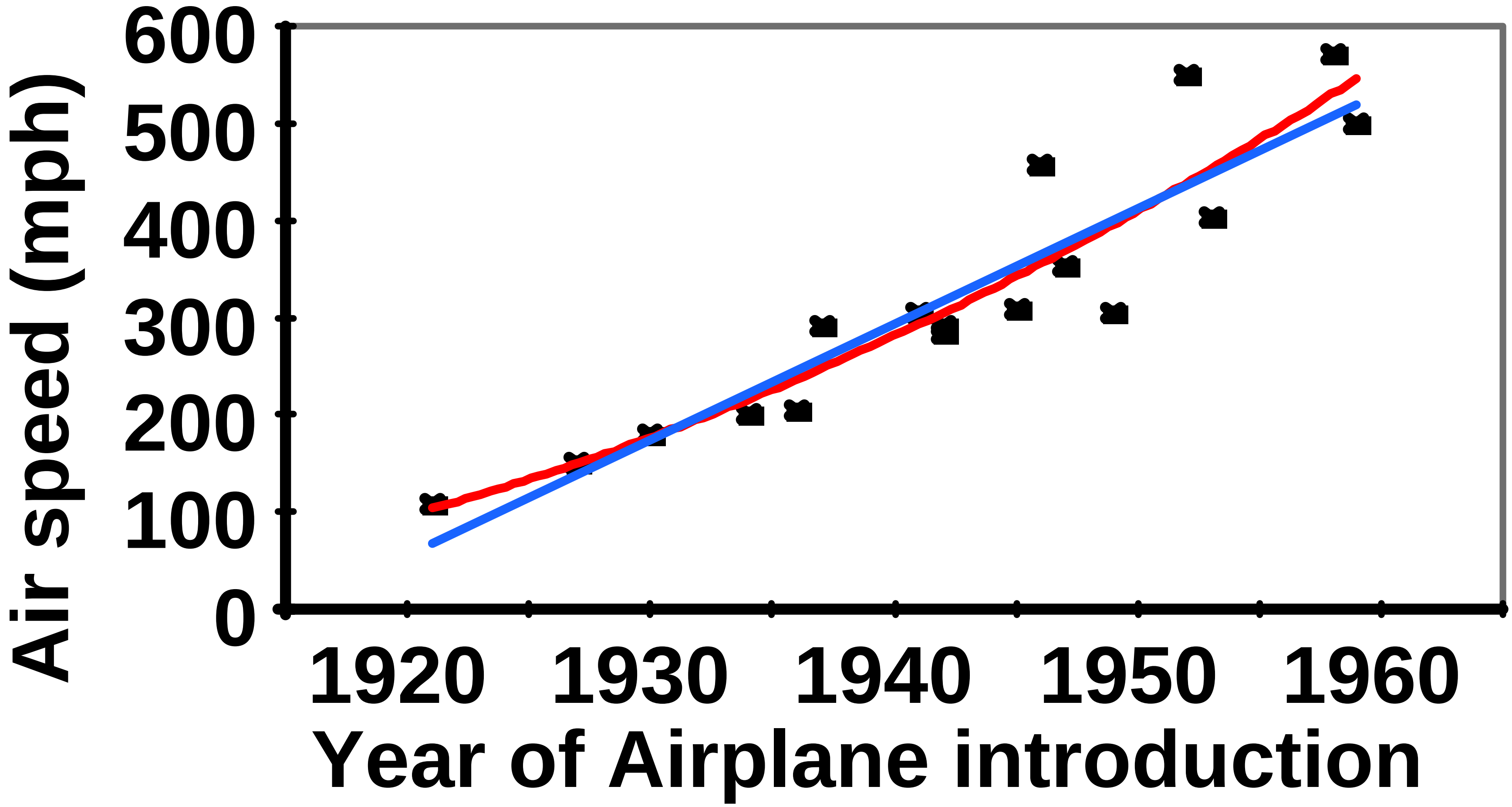
Source	DF	Type I SS	Mean Square	F Value	Pr > F
YEAR	1	404010.2946	404010.2946	148.97	<.0001
YEAR*YEAR	1	1431.3356	1431.3356	0.53	0.4774

Polynomial Reg. - Example 1 *(continued)*

- There is clearly a linear increasing trend over time (P>F)<0.0001. However, the additional term for quadratic curvature is not significant. There is not apparently any significant curvature in this example. At least not of a parabolic shape.
-

Poly. Reg. - Ex 1 *(continued)*

- **Airspeed example with linear (blue) and quadratic (red) models fitted.**



Polynomial Reg. - Example 1 *(continued)*

- **There is not a good test between the two models (exponential fitted earlier and quadratic/linear here). However, the exponential fitted well, adjusted for possible nonhomogeneous variance and was readily interpretable.**
- **The quadratic is not justified, and would reduce to a SLR. Also simple and interpretable.**

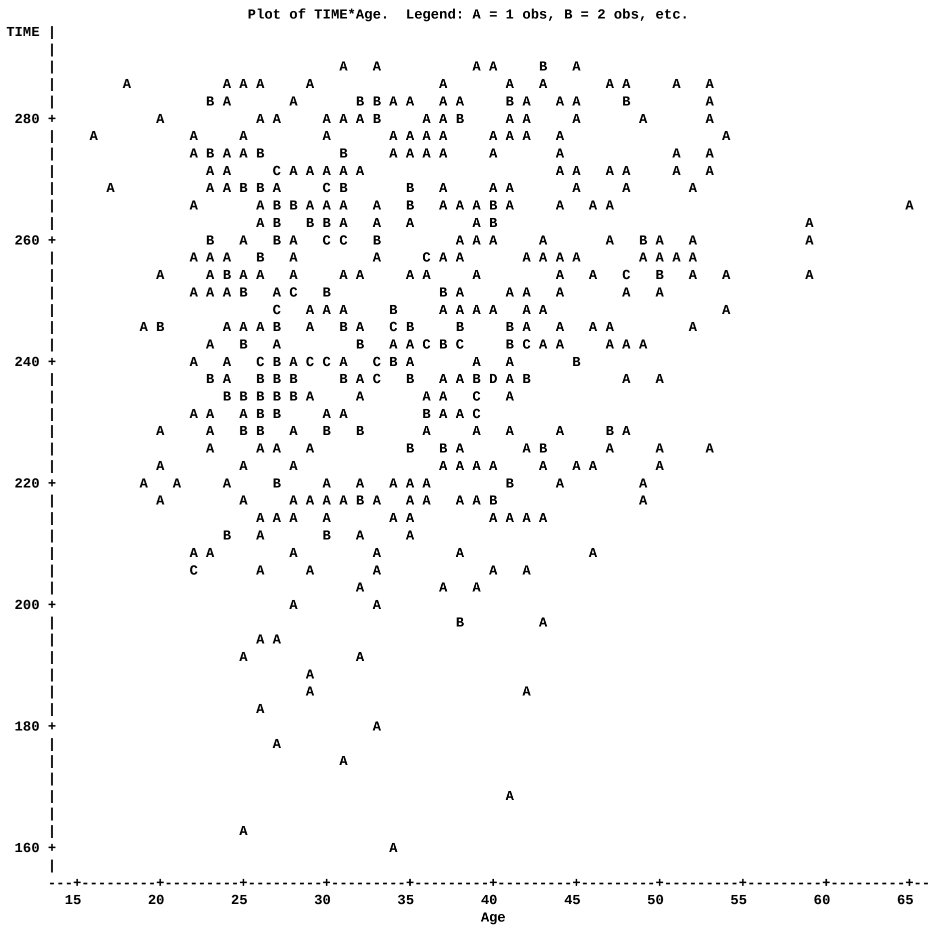
Polynomial Reg. - Example 2

- **10 K Race Results - Vermont.**
  - **Separate race results for 527 Women & 963 Men**
  - **Hypothesize that fastest runners will be neither the oldest nor the youngest.**
  - **This can be fitted with a polynomial.**
  - **Scatter plots for the two sexes, and the regression were run in SAS (below).**

Polynomial Reg. - Example 2 *(continued)*

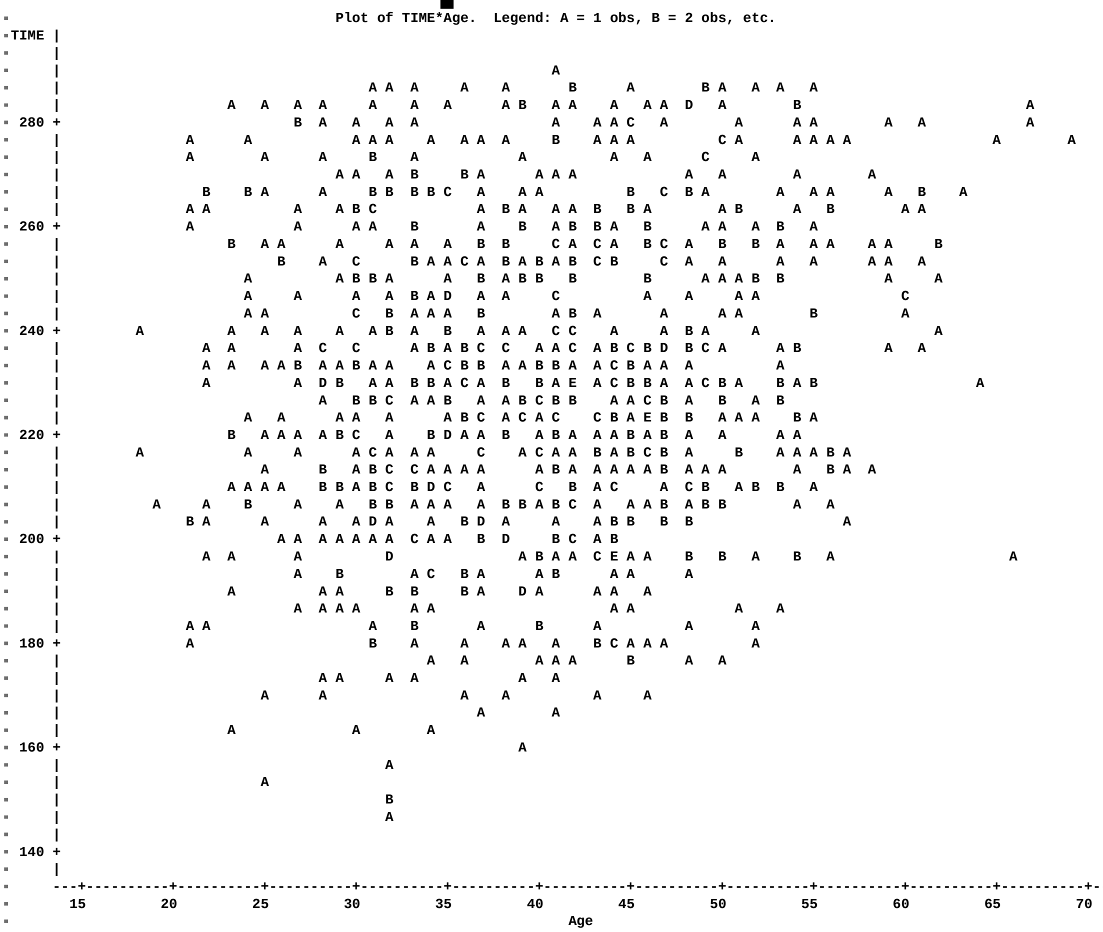
- **Scatter plot Sex=F**

sex=F



Polynomial Reg. - Example 2 *(continued)*

- **Scatter plot Sex=M**



Polynomial Reg. - Example 2 *(continued)*

- **The GLM Procedure - Sex=F**

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	270.94	15.3541	524	17.65	<.0001
Age	-1.7668	0.8679	524	-2.04	0.0423
Age*Age	0.02906	0.01179	524	2.46	0.0140

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Age	1	524	8.37	0.0040
Age*Age	1	524	6.08	0.0140

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Age	1	524	4.14	0.0423
Age*Age	1	524	6.08	0.0140

Polynomial Reg. - Example 2 *(continued)*

- The GLM Procedure - Sex=M

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	265.60	13.9782	960	19.00	<.0001
Age	-2.3003	0.6995	960	-3.29	0.0010
Age*Age	0.03392	0.008488	960	4.00	<.0001

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Age	1	960	22.10	<.0001
Age*Age	1	960	15.97	<.0001

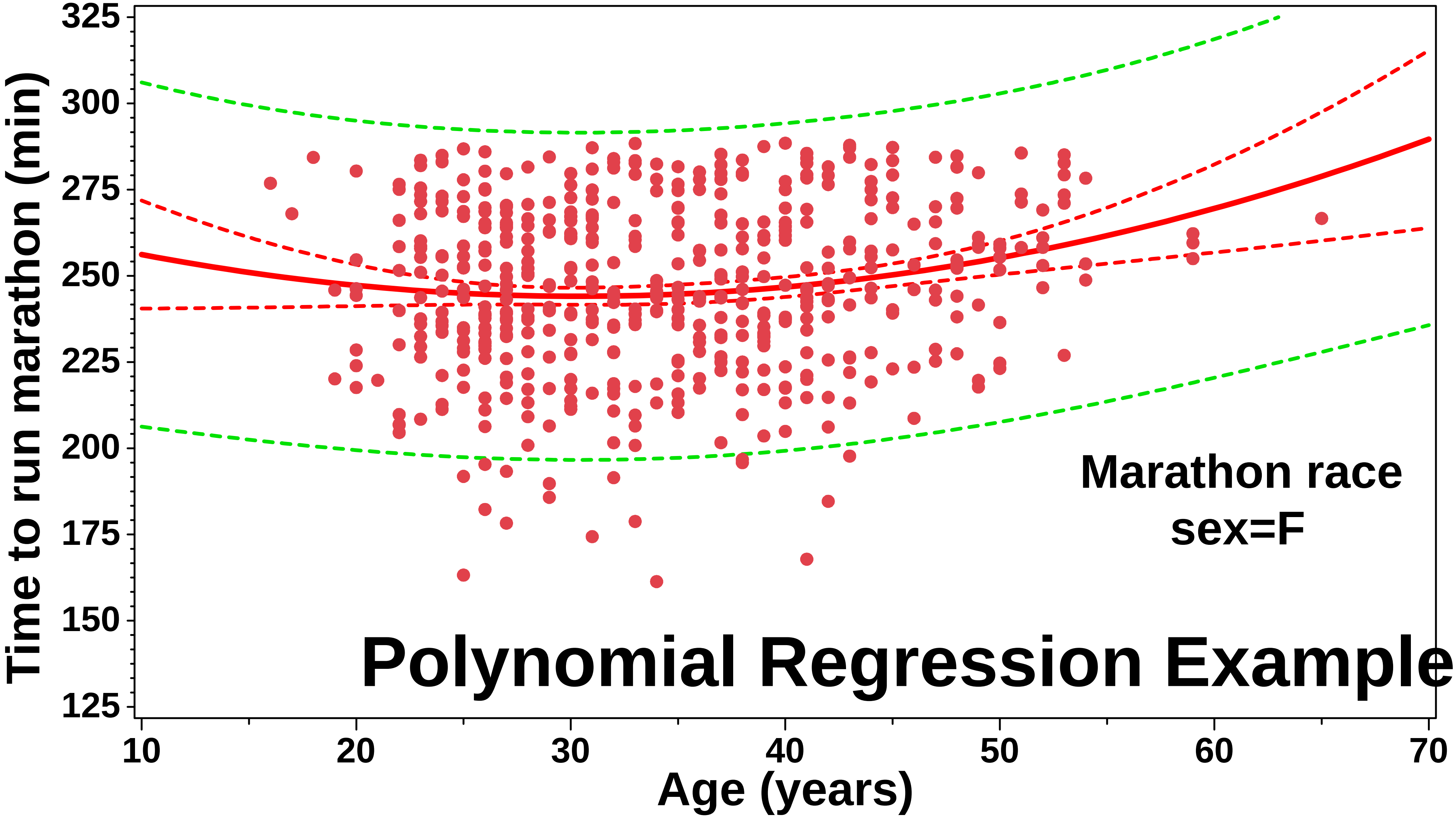
Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Age	1	960	10.81	0.0010
Age*Age	1	960	15.97	<.0001

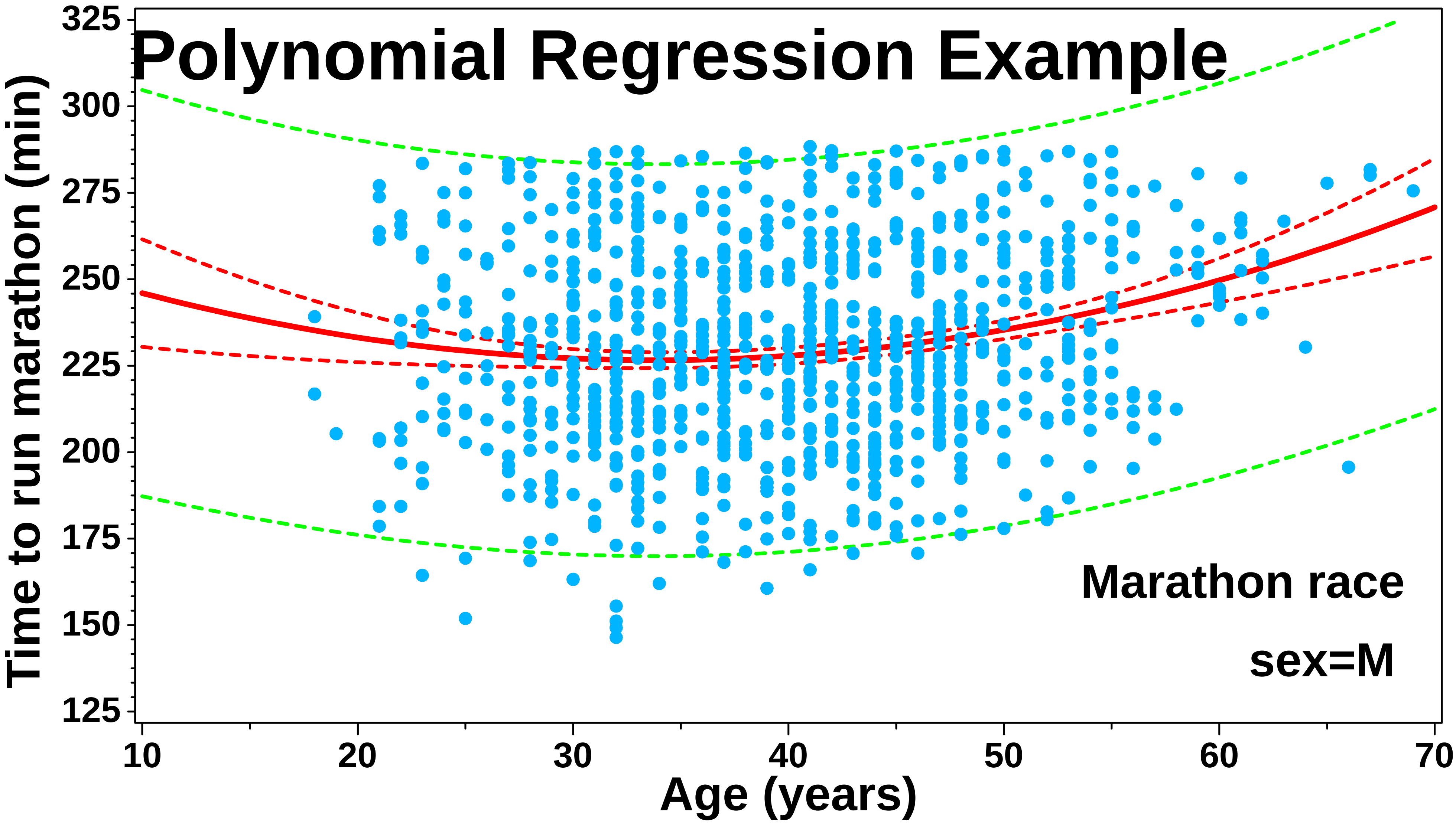
Polynomial Reg. - Example 2 *(continued)*

- **High resolution graphics were prepared in SAS and processed in Freelance. The statements to run the plots were:**

```
137         PROC GPLOT DATA=ONE; BY SEX;
138            TITLE1 font='TimesRoman' H=1 'Polynomial Regression Example';
139            TITLE2 font='TimesRoman' H=1 'Marathon race';
140          PLOT TIME*AGE=1 TIME*AGE=2 / overlay HAXIS=AXIS1 VAXIS=AXIS2;
141          AXIS1 LABEL=(font='TimesRoman' H=1 'Age (years)') WIDTH=1  MINOR=(N=1)
142                VALUE=(font='TimesRoman' H=1) color=black ORDER=10 TO 80 BY 10;
143          AXIS2 LABEL=(ANGLE=90 font='TimesRoman' H=1 'Time to run marathon (min)')
144                   WIDTH=1 VALUE=(font='TimesRoman' H=1) MINOR=(N=5) color=black
144                   ORDER=125 TO 450 BY 25;
145             SYMBOL1 color=red  V=None I=RQclI99  L=1 MODE=INCLUDE;
146             SYMBOL2 color=blue V=DOT  I=None     L=1 MODE=INCLUDE; RUN;
**** V = "dot" would place a dot for each point;
**** I = for regression: R requests fitted regression line, L, Q or C requests Linear,
     Quadratic or cubic, CLM or CLI requests corresponding confidence interval and
     95 specifies alpha level for CI (any value from 50 to 99);
**** I = for categories" requests STD (std dev) 1 (1 width, 2 or 3) M (of mean=stderr)
     J (join means of bars) t (add top & bottom hash) p (use pooled variance);
**** Other options for categories: omit M=std dev, use B to get bar for min/max;
```



Polynomial Regression Example


Polynomial Regression Example
Marathon race
sex=M
Time to run marathon (min)
Age (years)
325
300
275
250
225
200
175
150
125
10
20
30
40
50
60
70

Polynomial Reg. - Example 2 *(continued)*

- **So there is an intermediate age that runs the 10 K race fastest, and younger and older individuals take longer. What is that age?**
- **The fitted model for females is $\text{Time}=270.94-1.7668\text{Age}+0.02906\text{Age}^2$**
- **The fitted model for males is $\text{Time}=265.60-2.3003\text{Age}+0.03392\text{Age}^2$**

Polynomial Reg. - Example 2 *(continued)*

- If we take the first derivative and set this equal to zero, and solve for Age we get:
  - Age at minimum time = 1.7668 / 2(0.02906) = 30.4
  - Using the equation to solve for the average time at age = 30.4 we get 244 minutes for women, the best <u>average</u> time for any age.
  - Men had a minimum at 33.9 and had a time of 226.6 minutes at that age.

Polynomial Reg. Summary

- Polynomial regressions are treated like any other multiple regression, except that we use Type I SS for testing hypotheses.
- Note that the FULLY ADJUSTED regression coefficients are still used to fit the model.
- The ability to determine a minimum or maximum point is a useful application of polynomials (optimum performance @ age, optimum yield @ fertilizer level, etc).

Polynomial Reg. Summary *(continued)*

- **We have some new options as far as what we can do with regression.**
  - **Test if there a curvilinear relationship between the Y and X.**
  - **Test if the curvature is Quadratic? Cubic? Quartic? ...**
  - **We can now obtain a curvilinear predictive equation for Y on X.**
  -

Polynomial Reg. Summary *(continued)*

- **NOTES on POLYNOMIAL REGRESSION**
  - **Polynomial regressions are fitted successively starting with the linear term (a first order polynomial). These are tested in order, so Sequential SS are appropriate.**
  - **When the highest order term is determined, then <u>all lower order terms are also included</u>.**

Polynomial Reg. Summary *(continued)*

- **For example, we fit a fifth order polynomial, and only the CUBIC term is significant, then we would OMIT THE HIGHER ORDER NON-SIGNIFICANT TERMS, BUT RETAIN THOSE TERMS OF SMALLER ORDER THAN THE CUBIC.**
- **This does not mean that $Y_i = b_0 + b_1X^3_i + e_i$ is not a potentially useful model, only that this is not a "polynomial" model.**

Polynomial Reg. Summary *(continued)*

- If there are "s" different values of X_i, then s-1 polynomial terms (plus the intercept) will pass through every point (or the mean of every point if there are more than one observation per X_i value.
- It is often recommended that not more than 1/3 of the total number of points (different X_i values) be tied up in polynomial terms. For example, if we are fitting a polynomial to the 12 months of the year, don't use more than 4 polynomial terms (quartic).

Polynomial Reg. Summary *(continued)*

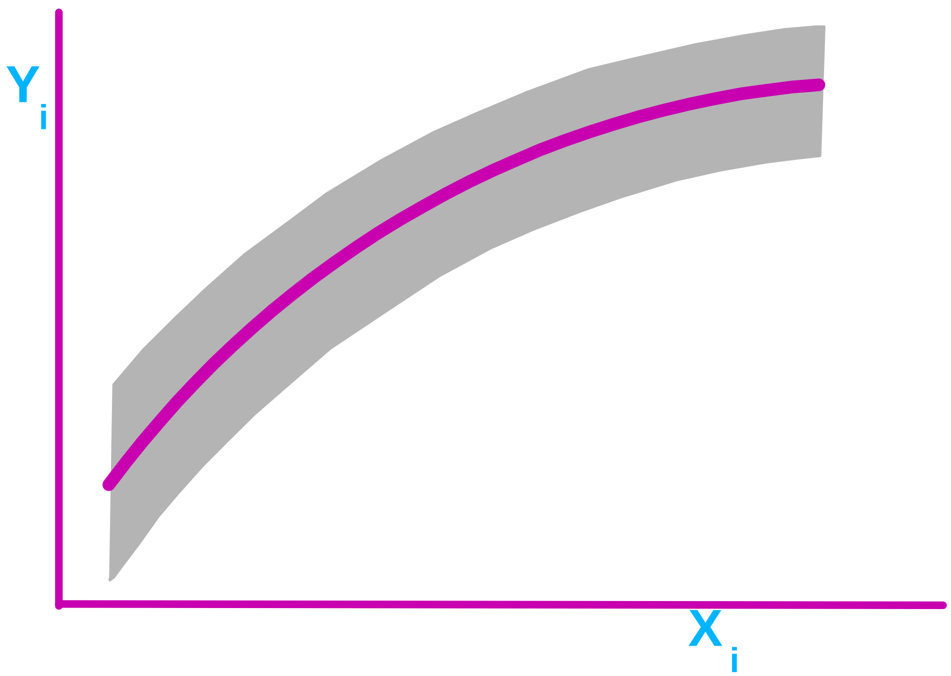
- **All of the assumptions for regression apply to polynomials.**
- **Polynomials are WORTHLESS outside the range of observed data!!! Do NOT try to extend predictions beyond the range of data.**
- **Polynomials generally do not have " biologically interpretable" regression coefficients.**

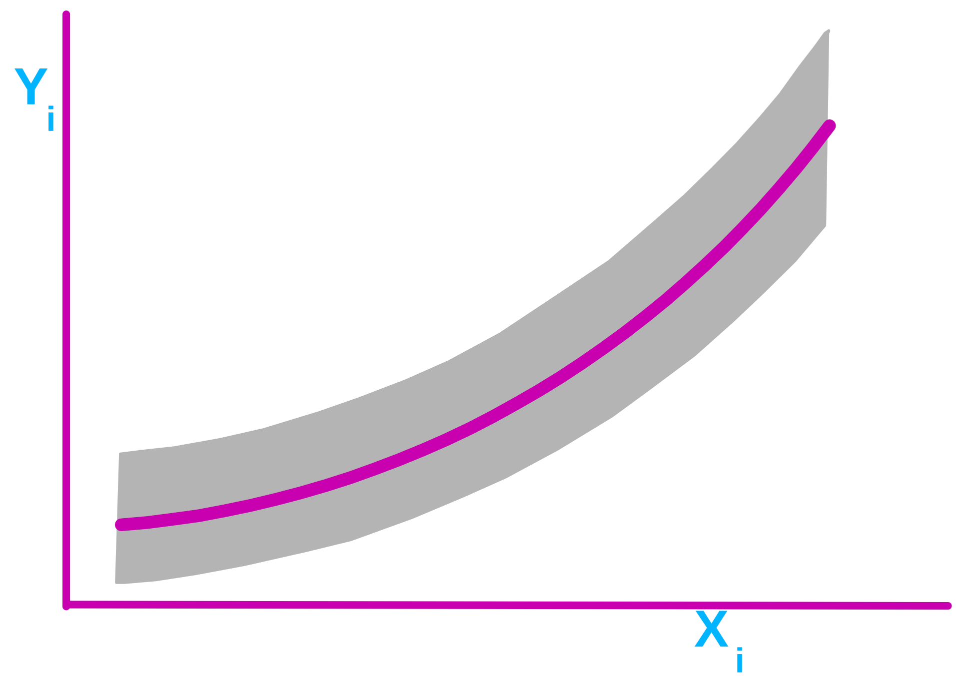
Polynomial Reg. Summary (continued)

- Since the power terms are correlated, multicollinearity could be an issue, but for two facts.
  - Using sequential SS gives exactly the needed tests, collinearity is not an issue.
  - Regression coefficients may be affected and variances inflated, but we are unlikely to be interested in the regression coefficients for polynomials anyway.
-

Polynomial Reg. Summary (continued)

- Recall that transformations of X_i will not influence variance. This is true for polynomials.





Response surfaces

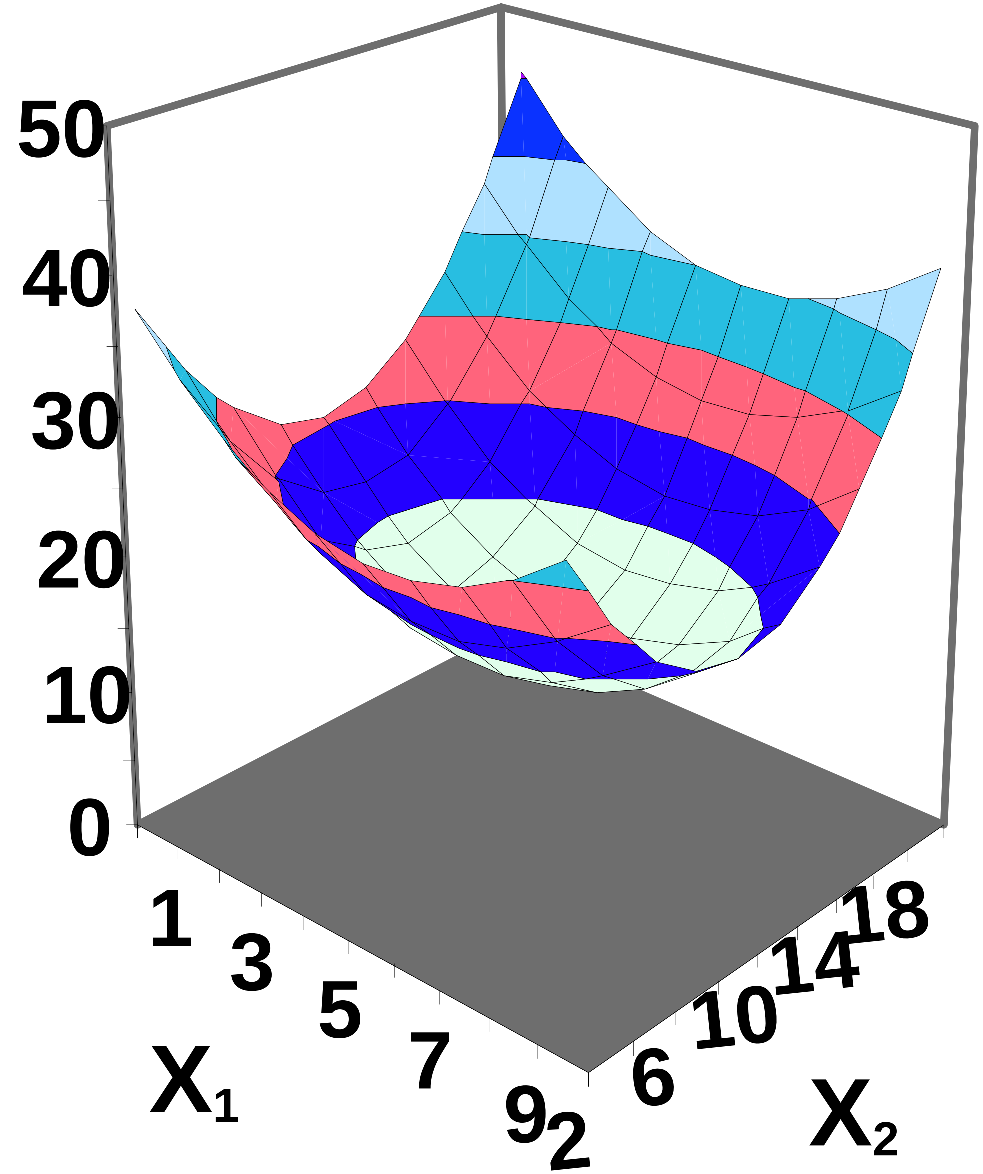
- Polynomials are of interest as an extremely flexible method of curve fitting, even though there are some severe restrictions on the interpretation and predictive ability (outside range) of the model.
- The ones we have looked at employed only one independent variable. Can you have several?
- YES, this is a "response surface".

Response surfaces *(continued)*

- For example, take 2 independent variables. We would include not only quadratic terms (and maybe cubic, etc), but also INTERACTIONS.
- $Y_i = b_0 + b_1 X_{1i} + b_2 X^2_{1i} + b_3 X_{2i} + b_4 X^2_{2i} + b_5 X_{1i} X_{2i} + e_i$
- Shape varies with size & sign of b_i
  - both concave produces a "bowl" shape
  - both convex yields a "hill" shape
  - one of each gives a "saddle" shape
- The interaction term allows for "twisting" of the surface.

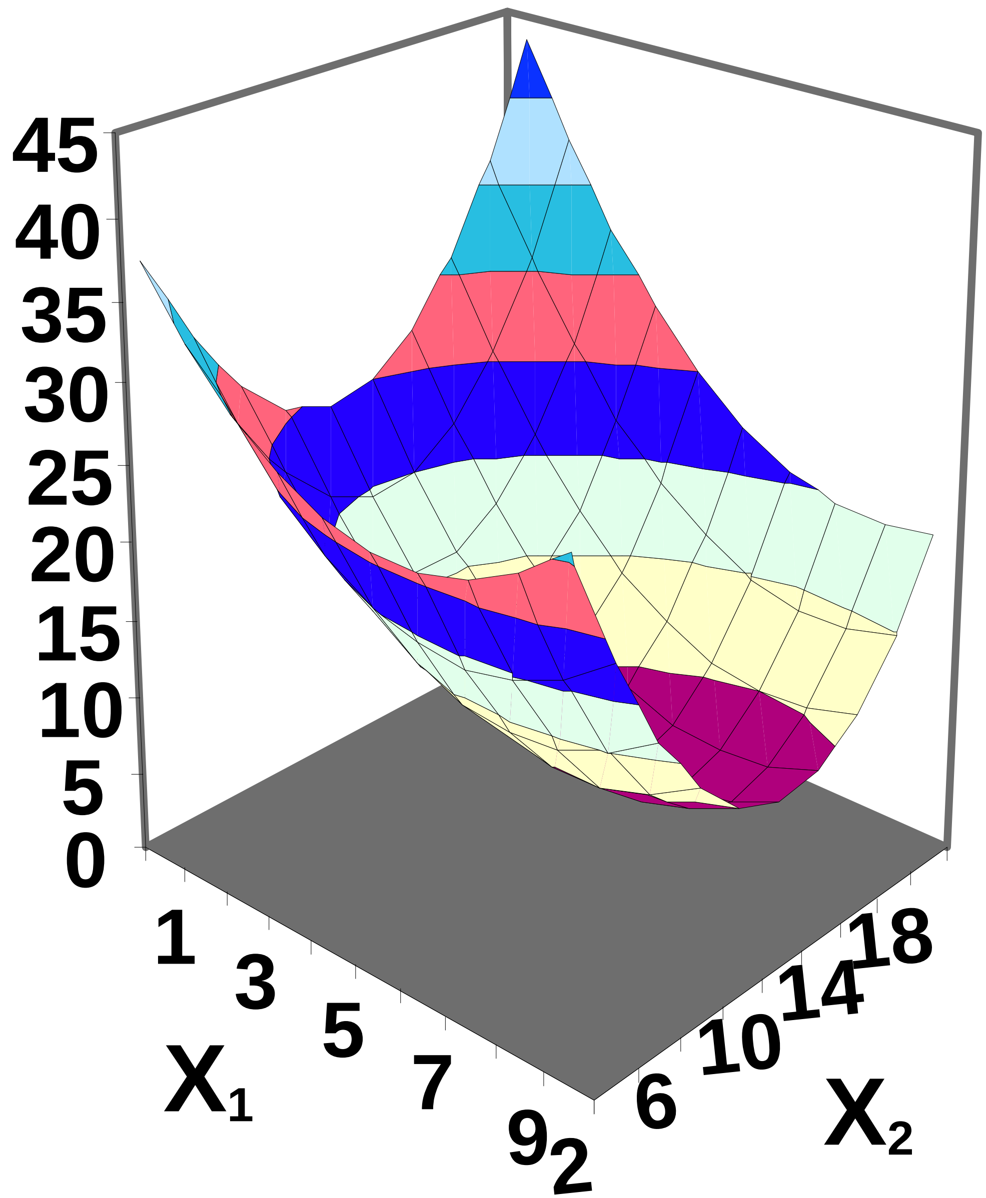
Response surfaces (*continued*)

- Both dimensions Concave
- Symmetric
  - no interaction
  -



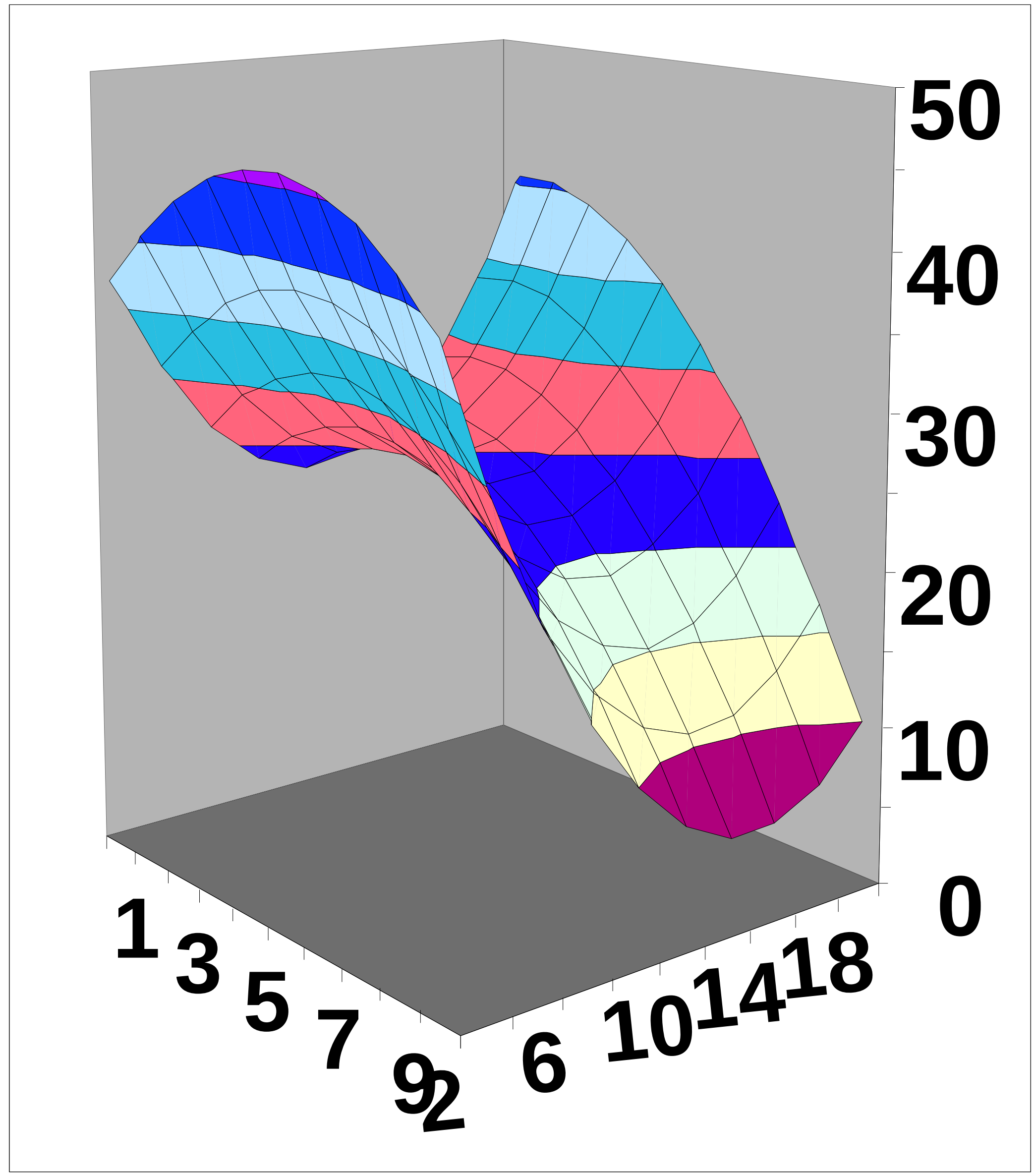
Response surfaces (*continued*)

- Both dimensions Concave
- Asymmetric
  - interaction present
  - Surface "twisted"



Response surfaces (*continued*)

- One dimension Concave and one convex
- Asymmetric
  - interaction present
  -



Curvilinear Regression Revisited

- **Remember the Air speed example. The author fitted a quadratic model to this data. However, many examples of technological development over time follow an "exponential" model. We will fit an exponential model to this example and compare.**

Curvilinear Example 1 *(continued)*

- First we fit the quadratic.

```
Dependent Variable: SPEED
                               Sum of          Mean
Source              DF        Squares        Square    F Value      Pr>F
Model                2    257583.3555   128791.6778      41.81    0.0001
Error               14     43121.7033     3080.1217
Corrected Total     16    300705.0588

       R-Square            C.V.       Root MSE           SPEED Mean
       0.856598        17.27670       55.49884             321.2353

Source       DF        Type I SS    Mean Square    F Value      Pr > F
YR            1      253866.5549    253866.5549      82.42      0.0001
YR*YR         1        3716.8006      3716.8006       1.21      0.2905

                                  T for H0:     Pr > |T|    Std Error of
Parameter        Estimate    Parameter=0                    Estimate
INTERCEPT    96.73022211           1.98       0.0672     48.75733514
YR            6.77213187           1.40       0.1825      4.82816387
YR*YR         0.12195807           1.10       0.2905      0.11102220
```

Curvilinear Example 1 *(continued)*

- **Then the exponential for comparison.**

```
Dependent Variable: LOGSPEED
                               Sum of           Mean
Source                 DF     Squares         Square    F Value      Pr>F
Model                   1   3.11456238    3.11456238     145.18    0.0001
Error                  15   0.32179173    0.02145278
Corrected Total        16   3.43635410

      R-Square              C.V.        Root MSE        LOGSPEED Mean
      0.906357          2.579479        0.146468             5.678188

Source     DF        Type I SS     Mean Square    F Value      Pr > F
YR          1       3.11456238      3.11456238     145.18      0.0001
Source     DF      Type III SS     Mean Square    F Value      Pr > F
YR          1       3.11456238      3.11456238     145.18      0.0001
                                  T for H0:    Pr > |T|   Std Error of
Parameter        Estimate    Parameter=0                   Estimate
INTERCEPT    4.750697794           56.04      0.0001     0.08477711
YR           0.041602463           12.05      0.0001     0.00345273
```

Curvilinear Example 1 *(continued)*

- **Notice that**
  - **that the TYPE I SS and TYPE II SS are the same (Simple Linear Regression),**
  - **the model has one fewer degrees of freedom (only one slope),**
  - **and that the model may actually fit better (as judged by the R^2, but this is not a definitive assessment since the variables are scaled differently).**

Curvilinear Example 1 *(continued)*

- **Recall the exponential models where the slope was interpreted as a "proportional" or percentage increase per X variable unit.**
- **Recall the percentage and the average annual increase in speed of 4.25%.**
- **Recall speed doubled every 16.67 years**
- **Polynomials usually DO NOT have a good interpretation of the regression coefficients.**

Curvilinear Example 1 *(continued)*

- **Exponential models. What can I say?**
  - **Better fit,**
  - **fewer d.f.,**
  - **clearer interpretation.**
- **I like them!**
- **A note on logarithms. This model requires natural logs. In SAS the function "LOG()" gives natural logs (LOG10 gives log base 10). In EXCEL the natural log function is "LN()".**