Biostatistics 201a - Lecture 22 - 11/14/11

Contents: - Transformations
- Centering
- Leadin to Regression Assumptions

pages = 11

11/14/11

- HW5 turn in up front
- HW6 is due next Wednesday
- Project info will be posted later today

Last time we were talking about curvilinear models / transformations

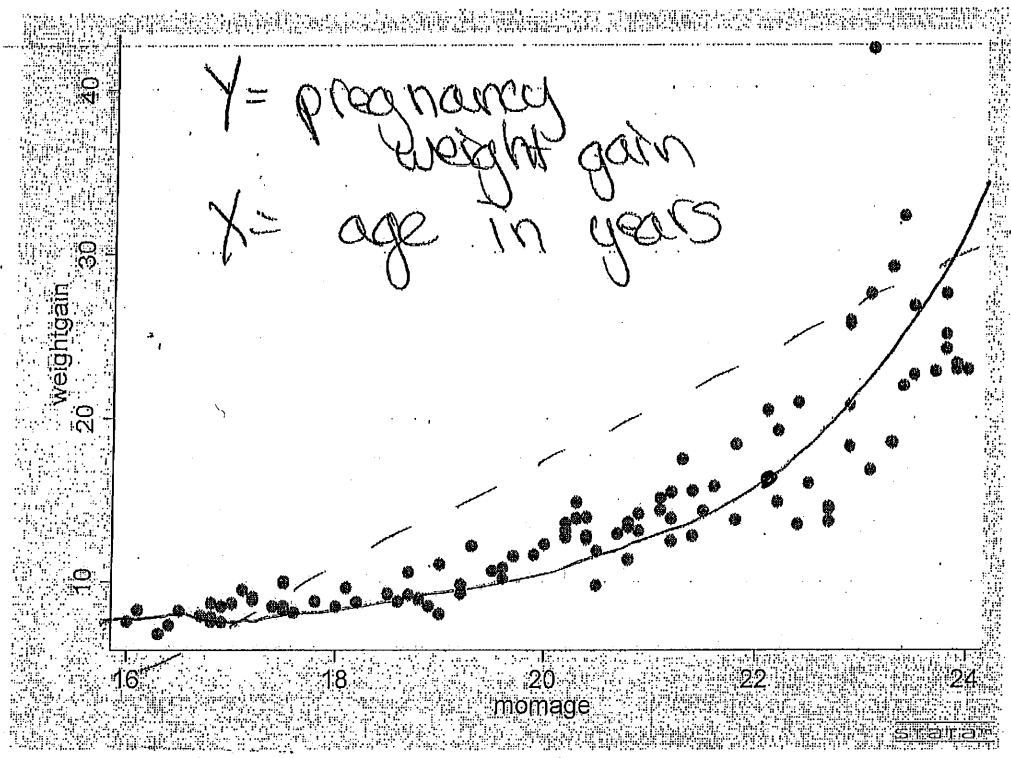
- Idea - create a new X' (or Y') based on functions of existing variables to better understand the shape of relationship between orginal Y and X.

Example (powers of X):

Y = pregnancy weight gain
X_1 = age of woman in years

Want to understand how X relates to Y.

Looks like curved (or maybe piecewise linear) relationship




```
reg weightgain momage momage2
```

Source	SS	df	MS
Model	3240.92603	2	1620.46302
Residual	975.903916	97	10.0608651
Total	4216.82995	99	42.5942419

Number of obs	= 100
F(2, 97)	= 161.07
Prob > F	= 0.0000
R-squared	= 0.7686
Adj R-squared	= 0.7638
Root MSE	= 3.1719

weightgain	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
momage	-12.32079	2.606697	-4.73	0.000	-17.49436	-7.147216
momage2	.363564	.0646568	5.62	0.000	.2352382	.4918897
_cons	112.6111	25.99637	4.33	0.000	61.01549	164.2067

What if we try fix this by adding $X_2 = X_1^2$?

Model would be

$$Y = \beta_0 + \beta_1 Age + \beta_2 Age^2 + \varepsilon$$

Fitted equation

$$\hat{Y} = 112.6 - 12.32 Age + .36 Age^2$$

equation of an up-opening parabola. - weight gain accelerates with aging.

Question: Is the quadratic model better than an SLR?

- Check R^2_{adj}, RMSE, etc. do need to take into account that the quadratic model uses 2 predictors (d.f.'s not same)
- Perform an hypothesis test on β_2, the coefficient of Age^2,

H_0: $\beta_2 = 0$ – Age^2 does not add any additional info about weightgain beyond what is explained by age alone – i.e. curved model is not better than linear model

H_A: $\beta_2 \neq 0$ – Age^2 is useful; the relationship between age and weightgain is significantly better described by the curvilinear shape.

Test statistic: $t_{obs} = \frac{b_2 - 0}{Sb_2}$

just like any other test

$t_{obs} = 5.62$ p-value $= 2P(t_{47} \geq 5.62) \approx 0$

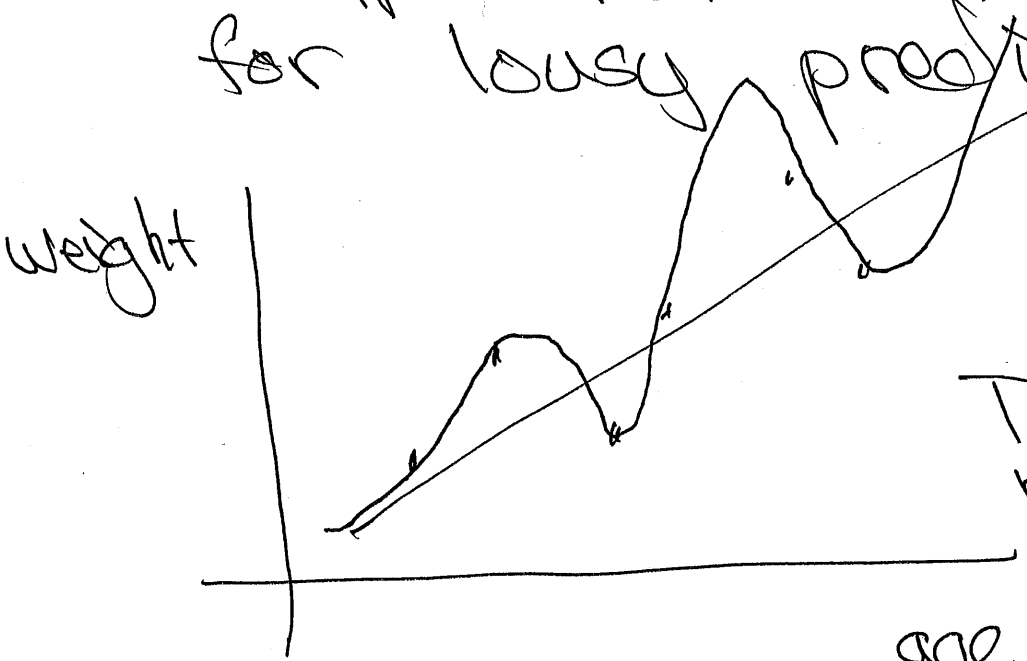
So we reject H_0 and conclude that weightgain/age relationship is not linear. I haven't proved that the quadratic model is best, just that it's better than an SLR.

We could try other shapes and compare their fits. An art, not a science.

Two notes:

(1) Anything (shape) can be arbitrarily well approximated by powers of X: $X^2, X^3, X^4, \ldots$ the higher you go, the "better" the fit will look. ($R^2 \uparrow$, RMSE $\downarrow$, ...)

(2) But beware of overfitting! If we use as many powers as data points we'll always get a perfect fit to our sample but it may make for lousy predictions

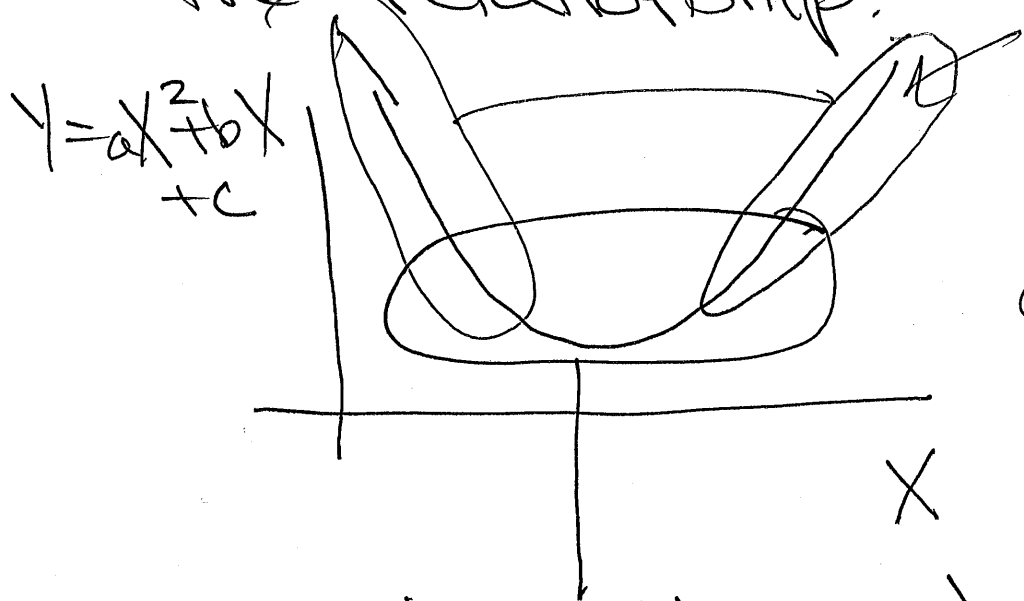


5th degree polynomial

This doesn't make any sense!

Usually aim for lowest power that fits well - don't add higher powers that aren't statistically significant

What about multicollinearity?
e.g. age^2 is completely dependent on age. Well, the relationships between X and X^2, X^3, $1/X$, etc. are not linear relationships. Multic. ⇒ problem if predictors are linearly related. So it may be ok. It matters how much of your data are near the "bend" in the relationship.



here curve looks pretty straight. so if all data are here we'll have a problem

here the relationship is very non-linear so if this is where data are we're probably ok.

What if we do have multicollinearity between X, X^2 say? This is an example of a "structural multicollinearity" - it's induced by the way we defined our variables. There's a trick called "centering" which we can use to fix this. Instead of using X, X^2 we define ($\bar{X}$ sample mean)

$$X' = X - \bar{X} \qquad (X')^2 = (X - \bar{X})^2$$

This gives a new model

$$Y = \beta_0' + \beta_1'(X - \bar{X}) + \beta_3'(X - \bar{X})^2$$

This puts the "bend" in the parabola at $\bar{X}$ where our data are but is also the point with least multic

So $X', (X')^2$ are less colinear than X and X^2

Does this change our predictions?
No! If we multiply out

$$Y = \beta_0' + \beta_1' X - \beta_1' \bar{X} + \beta_2' X^2 - 2\beta_2' \bar{X} X + \beta_2' \bar{X}^2$$

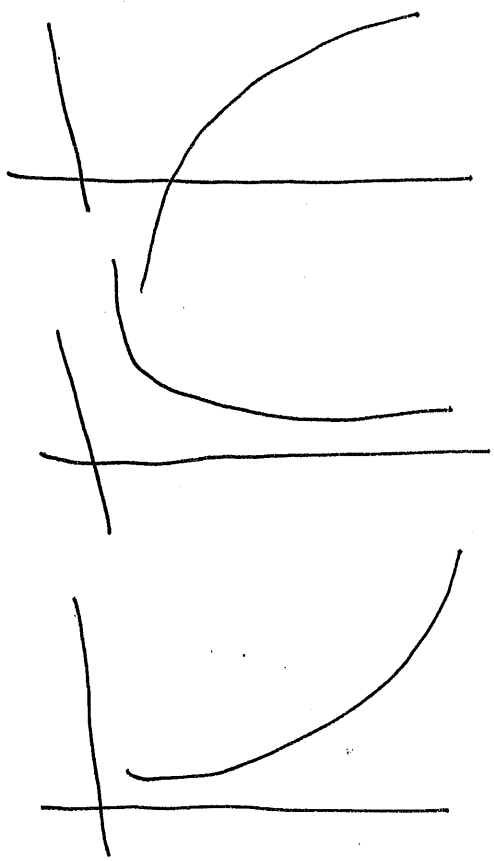
and if rearrange terms this just has the form $Y = aX^2 + bX + c$, as before - just grouped differently

Pros: - Same predictions
- more stable βs because less multic.

Con: Interpretation - everything is in terms of $X' = X - \bar{X}$ which may not be so meaningful to us and which depends on sample.

Other popular transformations:

For X:



$\ln X$, $\sqrt{X}$ } relationship increasing but not too rapidly

$1/X$, e^{-X} } rapid decay

X^2, e^X } increasing faster than linearly

For Y: You can do basically the same transformations

So.... when do you transform X and when do you transform Y?

Goal in any transformation is to arrive at a set of outcome/predictor variables

that have a linear relationship with well behaved errors

$$Y' = \beta_0 + \beta_1 X_1' + \beta_2 X_2' + \cdots$$

Transformations of X and Y have different effects on, both the error terms and also on the relationship of Y to the other variables.

To understand this better we need to look in detail at our assumptions about the error terms!