ENV255/STA242, March 2, 2004

1 Outline

- R^2 in regression, definition of $R^2_{adjusted}$
- Polynomial regresion
- Leverage, residuals and influence in regression

2 R^2

R^2 cannot help with model goodness of fit, model adequacy, statistical significance of a regression, or need for transformation. Can provide a summary of tightness of fit, and sometimes, clarify *practical significance*.

3 Polynomial regression overview, Ch. 9

- Example of Koehn, 1978 regarding allele frequency as a function of distance from Southport, Conn.
- Research question: What are the genetic differences between lagoonal and oceanic populations? How are genetic differences affected by environmental variation? Theory that elevated temperatures or hypersalinity of a lagoon region might cause selective extinction of particular alleles with age and maintain a genetic gradient. (Planes et al., 1998, *Coral Reefs*)
- Used when (1) true response Y is a polynomial function or (2) when the true response function is unknown (or complex) and a polynomial function is a good approximation to the true function.
- As higher order terms are added, the curve becomes more complex and can fit a set of data increasingly well. At the same time, the residual mean square loses a degree of freedom.
- Retain lower powers of X up to the highest power considered. Higher order terms are viewed as providing refinements in the specification of the response function. If there is evidence that only a higher power of X relates meaningfully to Y, while lower powers have no effect and no biological meaning, lower powers can be omitted.
- Often data are centered $(X - \overline{X})$ to reduce problems of multicollinearity among X terms of different powers.

4 Influence, Leverage and Residuals

1. Examination of bee pollen example. Parallel lines model regressing logit of proportion removed on log of time of duration and bee type (worker, queen).
2. *Identifying outlying X observations: Leverage.* How far is a given X from the other X's? Plots identify points 1 and 36 as worthy of further examination.
3. *Identifying outlying Y observations: Residuals.*
   (a) Standardized residuals (studentized or "internally studentized")
   (b) Externally studentized residuals
4. *Identifying influential cases: Cook's Distance, DFFITS, DFBETAS*

(a) Cook's Distance: What is the influence of the i^{th} case on the set of all fitted values?

(b) DFFITS: What is the influence of the i^{th} case on an individual $\hat{Y}_i$? (function of externally studentized residuals)

$$
\begin{aligned}
(\text{DFFITS})_i &= \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\sigma^2_{(i)} \times h_i}} &(1)\\
&= (\text{ext. stud. res.})_i \times \sqrt{\frac{h_i}{1-h_i}} &(2)
\end{aligned}
$$

- An approximate number of standard deviations that $\hat{Y}_i$ changes when the i^{th} case is removed.
- *Rough rules:* DFFITS > 1 of concern for small-medium datasets; DFFITS $> 2\ sqrt\frac{p}{n}$ for large datasets.

(c) DFBETAS: What is the influence of the i^{th} case on each regression coefficient β_k?