Review: Second Half of Course
Stat 704: Data Analysis I, Fall 2014

Tim Hanson, Ph.D.

University of South Carolina

Higher order models

* (8.1 & 8.2) Polynomials & interactions in regression. Examples:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \epsilon_i,$$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_{11} x_{i1}^2 + \beta_2 x_{i2} + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \epsilon_i.$$

* Interpretation. First-order & second-order Taylor's approximation to general surface $Y_i = g(x_{i1}, x_{i2}) + \epsilon_i$.
* Extrapolation. Centering.
* Use of higher order models to test simpler (e.g. first-order) models.
* Know how to interpret interaction models! (pp. 306–308).

In the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_{11} x_{i1}^2 + \beta_2 x_{i2} + \epsilon_i,$$

how would you find when $E(Y_i)$ is either the largest or smallest w.r.t. x_{i1}?

Is this possible to do when there is an interaction? For example,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_{11} x_{i1}^2 + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \epsilon_i.$$

* (8.3) Categorical predictors with two or more levels. Dummy variables, baseline group. We primarily used zero-one dummy variables.
* (8.4) Interactions between continuous and categorical predictors. Example where $x_{i2} = 0$ or 1:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 I\{x_{i2} = 1\} + \beta_{12} x_{i1} I\{x_{i2} = 1\} + \epsilon_i.$$

  Interpretation! What is β_{12} here?
* Nice example is "soap production" pp. 330–334.

Model & variable selection

* (9.1) Model building, pp. 343–349.
* Preliminaries: functional form of response & predictors and possible interactions. *Use prior knowledge*.
* Reduction of predictors. Whether or not this is done depends on the type of study: controlled experiments (no), controlled experiment with "adjusters" (yes), confirmatory observational studies (no), exploratory studies – i.e. "data dredging" – YES!
* Model refinement and selection: 9.3 & 9.4.
* Model validation: 9.6, chapter 10???

* (9.3) Model selection:
* Maximize $R_a^2 = 1 - \frac{MSE_p}{SSTO/(n-1)}$. Penalizes for adding predictors. Same as minimizing MSE_p.
* Minimize $C_p = \frac{SSE_p}{MSE_P} - (n - 2p)$. P is number of predictors *including intercept* in large model that provides good estimate of σ^2 & p is number of predictors *including intercept* in smaller model. $E(C_p) \approx p$ (or $< p$) have little "bias." Also penalizes for adding too many predictors.
* AIC$_p$ & BIC$_p$ (BIC also called SBC). $AIC = -2\log \mathcal{L}(\hat{\beta}) + 2p$.
* PRESS$= \sum_{i=1}^n (Y_i - \hat{Y}_{i(i)})^2$. Leave-one-out cross-validated measure of model's predictive ability.
* (9.4) Backwards elim., forward selection, stepwise variable selection.
* (9.6) Validation: we did not cover this much. Use of "validation sample." PRESS gets at same thing – model with smallest PRESS has best "n-fold out of sample prediction."

Chapter 10: Diagnostics

* *Can we trust the model?*
* (10.1) Added variable (partial regression) plots: regress Y on all predictors except x_j and regress x_j on remaining predictors, plot residuals versus residuals. Flat pattern $\Rightarrow x_j$ not needed, linear pattern $\Rightarrow x_j$ needed as linear term, curved pattern $\Rightarrow x_j$ needed, but transformed.
* Added variable plots assume remaining predictors *are correctly specified*. Can lead to lots of model "tweaking."
* (10.2) $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is $n \times n$ "hat matrix." h_{ii} are diagonal leverages.
* $r_i = \frac{Y_i - \hat{Y}_i}{\sqrt{MSE(1-h_{ii})}}$ is (internally) studentized residual.
* $t_i \frac{Y_i - \hat{Y}_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$ is (externally) studentized deleted residuals. Has t_{n-p-1} distribution if model is true: Bonferroni check for observations outlying *with respect to the model*.

* r_i or t_i versus predictors $x_1, \ldots, x_k$ and/or versus fitted $\hat{Y}_i$ confirm or invalidate modeling assumptions...also get at "adequacy." Checking for linearity & constant variance. Residuals should be approximately normal if normality holds and model is correct. Non-normality affects hypothesis tests in small samples.
* (10.3) h_{ii} are leverage points: *potentially influential* values of $\mathbf{x}_i$. Rule of thumb $h_{ii} > \frac{2p}{n}$ outlying with respect to the rest of the predictor vectors. $h = \mathbf{x}_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h$ used to check for hidden extrapolation with high dimensional predictors.
* (10.4) $\text{DFFIT}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$ is number of standard deviations fitted $\hat{Y}_i$ changes when leaving out $(\mathbf{x}_i, Y_i)$. $\text{DFFIT}_i > 1$ or $2\sqrt{p/n}$ indicates influential observation.

* Cook's distance $D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\ MSE} = \frac{r_i^2 h_{ii}}{(1-h_{ii})p}$. When is D_i large? Aggregate measure of how fitted surface changes when leaving out i. Look for values substantially larger than other values.
* What do large values of r_i indicate? Large values of D_i? Large values of both?
* (10.5) $\text{VIF}_j = 1/(1 - R_j^2)$ where R_j^2 is R^2 regressing x_j onto the rest of the predictors. Check for $\text{VIF}_j > 10$. VIF's are for *predictors* not *observations*.

Chapter 11: Other types of regression

* (11.1) Weighted least squares: fix for non-constant variance that *does not involve transforming the response*.
* Recipe on page 425 to estimate weights w_i.
* $\mathbf{b}_w = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$.
* (11.2) Ridge regression: fix for multicollinearity.
* Adds bias, reduces variance by adding a biasing constant c to OLS estimate $\mathbf{b}_r = ((\mathbf{X}^*)'\mathbf{X}^* + c\mathbf{I})^{-1}(\mathbf{X}^*)'\mathbf{Y}^*$. Pick c from "ridge trace" of standardized regression coefficients.
* Increasing c *reduces variance* of estimator but biases coefficients toward zero.
* LASSO is same idea, but regression parameter estimates can be zero: useful for multicollinearity *and* variable selection.

* (11.3) Least absolute deviations (LAD) or L_1 regression minimizes (p. 438)

$$Q(\beta) = \sum_{i=1}^{n} |Y_i - \mathbf{x}_i'\beta|.$$

* Dampens effect of outliers (w.r.t. the model). Also useful for non-normal errors.
* Fit in PROC QUANTREG.
* IRLS (iteratively reweighted least squares) regression. Special case is Huber regression which is between OLS and LAD. Fit in PROC ROBUSTREG.

Additive model

* Fit in `proc gam` or `proc transreg` (also `gam` package in R). Will come back to this in STAT 705. Most general version:

$$h(Y_i) = \beta_0 + \beta_1 x_{i1} + \tilde{g}_1(x_{i1}) + \cdots + \beta_k x_{ik} + \tilde{g}_k(x_{ik}) + \epsilon.$$

* The functions $h(\cdot), \tilde{g}_1(\cdot), \ldots, \tilde{g}_k(\cdot)$ fit via splines.
* Obviates the use of "added variable" plots. All k transformations estimated simultaneously.
* `proc gam` provides tests of $H_0 : \tilde{g}_j(\mathbf{x}_j) = 0$ all x_j but does not estimate $h(\cdot)$.
* `proc transreg` provides estimates of $g_j(x_j) = \beta_j x_j + \tilde{g}_j(x)$ as well as $h(\cdot)$, but does not test $H_0 : \tilde{g}_j(\mathbf{x}_j) = 0$ all x_j (that I know of).
* Can be used to quickly find suitable transformations of predictors in normal-errors regression.
* Possible to include interactions & interaction surfaces in `proc gam`.

Exam II

* Will take place in LC 205, a few doors down, Thursday, 12/4 at the usual time. This is a STAT 205 lab.
* Exam II will be one data analysis project and a few short answer. Hopefully shorter than midterm.
* Closed book, closed notes, you can use SAS & R documentation, not internet.
* You will need to build and validate a good, predictive regression model for a given data set. Be complete as you can. Look at the applied parts of old qualifying exams to get an idea.
* Either use PC's in lab (you need a login), or else bring your own laptop. You will need to print out your exam in the lab (preferred), or else email it to me and I'll print it out (not preferred).
* Questions?