Leverage

- Some cases have high *leverage*, the potential to greatly affect the fit.
- These cases are outliers in the space of predictors.
- Often the residuals for these cases are not large because the response is in line with the other values, or the high leverage has caused the fitted model to be pulled toward the observed response.
- The **leverage** exerted by the i'th case is h_{ii}, the i'th diagonal element of the hat matrix.
- a **rule of thumb** is to flag cases where $h_{ii} > 2p/n$, where p is the number of columns of $\boldsymbol{X}$, equal to $k + 1$ in a multiple regression with k predictors and an intercept.

Leverage, continued

- In simple linear regressionm

  $$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

  so the minimum is $1/n$ at $\bar{x}$ and the maximum occurs when x is furthest from $\bar{x}$.
- More generally, h_{ii} measures the distance of the predictors from their centroid.
- The sum of the h_{ii} is $tr(\boldsymbol{H}) = k + 1 = p$, so their average is $\bar{h} = (k+1)/n = \mathrm{p/n}$.

A property of leverage - $0 \leq h_{ii} \leq 1$

Because $\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}$ and $\boldsymbol{H} = \boldsymbol{H}^T$,

$$h_{ii} = \sum_{j=1}^{n} h_{ij} h_{ji} = h_{ii}^2 + \sum_{j \neq i}^{n} h_{ij}^2 \tag{1}$$

so

$$h_{ii}(1 - h_{ii}) \geq 0$$

and

$$0 \leq h_{ii} \leq 1.$$

Leverage, continued

- The fitted value at case i is

$$\hat{y}_i = (\boldsymbol{Hy})_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i}^{n} h_{ij} y_j$$

  a linear combination of all the responses.

- Ideally all cases contribute, with those at and closest to $\boldsymbol{x}_i$ dominating.
- In influential cases h_{ii} approaches 1, and h_{ij} approaches 0, for $j \neq i$.

Multicolinearity

- Multicollinearity between the predictor variables means that the columns of $\boldsymbol{X}$ are nearly linearly dependent.
- In this case the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is ill conditioned, in which case
  - $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ will typically have large diagonal elements
  - meaning that the standard errors of the $\hat{\beta}_j$ are large
  - and the least squares estimates of the β_j's will often be excessively large in absolute value.
- **variance inflation factors** measure the linear relationships among columns of $\boldsymbol{X}$.
- The VIF for the j'th regesson coefficient can be written as

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination from regressing X_j on the other predictor variables.

Multicolinearity, continued

- **Rule of thumb:** A VIF larger than 5 implies serious problems with multicollinearity.
- When there is multicollinearity the estimated β's are very sensitive to minor changes in the data, as are the predicted values of future y's.
- What to do in the presence of multicolinearity?
  - Try some new combinations of the predictors which might be closer to orthogonal. **It is always best to centre predictor variables by removing their means, as the centred variables will have less correlation than the uncentred variables.**
  - Ridge regression - replaces $\boldsymbol{X}^T\boldsymbol{X}$ by $\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I}$. Gives reduced variance to the resulting estimator, but generates a biased estimator.
  - Principal components regression - uses new variables (principal components) which are transformed versions of the predictor variables, and are orthogonal. Principal components are a topic in Stat4350.
  - Remove some predictor variables from the regression.

Standardized residuals

- We have seen that the usual residuals typically do not have the same variances.
- Therefore, when doing residual analysis, it is best to use the externally standardized residuals.

  $$t_i = \frac{e_i/(1-h_{ii})}{s_{(i)}/\sqrt{1-h_{ii}}} = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

  where

Standardized residuals - continued

- e_i is the usual residual $y_i - \hat{y}_i$
- h_{ii} is the leverage of the i'th case
- $s^2_{(i)} = \frac{SSE_{(i)}}{n-k-2}$ is called the deleted variance estimate, and
- $SSE_{(i)} = SSE - \frac{e_i^2}{1-h_{ii}}$ is called the deleted residual sum of squares.
- **Rule of thumb** - it is recommended that cases with standardized residual greater than 2 in absolute value should be examined by the data analyst.

Case Deletion Diagnostics

- the i'th **deleted estimate of β** is the least squares estimate when the i'th case is deleted, and is given by

$$\hat{\beta}_{(i)} = \hat{\beta} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i \frac{e_i}{1-h_{ii}}.$$

- **Cook's distance** for the i'th case is defined as

$$D_i = \frac{1}{(k+1)s^2}(\hat{\beta}-\hat{\beta}_{(i)})^T\boldsymbol{X}^T\boldsymbol{X}(\hat{\beta}-\hat{\beta}_{(i)}) = \frac{(\hat{\boldsymbol{y}}_{(i)}-\hat{\boldsymbol{y}})'(\hat{\boldsymbol{y}}_{(i)}-\hat{\boldsymbol{y}})}{(k+1)s^2}$$

  - It measures the change in the estimate of β when the ith case is deleted
  - and where $\hat{\boldsymbol{y}}_{(i)}$ are the predicted values based on all of the observations but the i'th, Cook's distance also measures the change in predicted values
  - A **rule of thumb** is to carefully exam cases for which $D_i > F(.5, p, n-p) \approx 1$.

Transformations

- A **variance stabilizing transformation** may be useful when the variance of y appears to depend on the value of the regressor variables, or on the mean of y. In general, if y has mean μ_y and variance σ_y^2, then a function $h(y)$ has approximate mean $h(\mu_y)$ and variance approximately equal to $(h'(\mu_y))^2\sigma_y^2$. This can be used to calculate a variance stabilizing transform.
- For example, suppose Y has a Poisson distribution with mean $\mu_Y = \mu$ and variance $\sigma_Y^2 = \mu$. let $Z = h(Y) = \sqrt{Y}$. Then $h'(Y) = .5Y^{-1/2}$, so that $h'(\mu) = .5\mu^{-1/2}$, and the variance of Z is approximately $(h'(\mu_y))^2\sigma_y^2 = (.5\mu^{-1/2})^2\mu = .25$. The variance stabilizing transformation for the Poisson distribution is the $\surd$ transform.

Generalized least squares

- Generalized (or weighted) least squares gives an estimate of $\boldsymbol{\beta}$ in the model
  - $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $E(\boldsymbol{\epsilon}) = \mathbf{0}$ $Cov(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{V}$,
  - with $\boldsymbol{V}$ known.
- the estimate turns out to be

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y}$$

- The **Gauss-Markov theorem** states that $\hat{\boldsymbol{\beta}}$ is the minimum variance unbiased estimator of $\boldsymbol{\beta}$, so the best that one can do in terms of minimizing the expected mean squared error.

Intrinsically linear models

- Some models are intrinsically linear, and can be appropriately transformed to give a linear relation.
- For example, in Economics, the Cobb-Douglas production function is $P = kL^{\alpha}C^{\gamma}\epsilon$,
- Taking logarithms gives $log(P) = \beta_0 + \beta_1 log(L) + \beta_2 log(C) + log(\epsilon)$
- Which is a linear function of $log(L)$ and $log(C)$ with parameters $\beta_0 = log(k)$, $\beta_1 = \alpha$ and $\beta_2 = \gamma$.

another intrinsically linear model

- In biochemistry, where y is reaction rate and x is substrate concentration, the Michaelis-Menten equation states that

$$y = \frac{V_{max} x}{K_m + x}$$

- V_{max} and K_m are parameters to be estimated.
- Note that as $x \to \infty$, $y \to V_{max}$.
- The Lineweaver-Burk plot, or double reciprocal plot, is a plot of $1/y$ vs $1/x$, provides a convenient means of estimating the two model parameters.

$$\frac{1}{y} = \frac{K_m}{V_{max}} \frac{1}{x} + \frac{1}{V_{max}}$$

- or

$$\frac{1}{y} = \beta_0 + \beta_1 \frac{1}{x}$$

- with $\beta_0 = \frac{1}{V_{max}}$ and $\beta_1 = \frac{K_m}{V_{max}}$
- Find the least squares estimators of β_0 and β_1 and transform

How to choose a model

- If candidate models are restricted to multiple regression models, one strategy is to choose a model which maximizes adjusted R^2 among the models under consideration.
- For a model with p parameters,

$$R^2_{Adj,p} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2_p)$$

  where R^2_p is the usual R^2 for that model.
- R^2 is not a good criterion to use, as it will always be maximized for the largest model considered.
- It turns out that choosing the candidate model to maximize $R^2_{Adj,p}$ is equivalent to choosing the model which minimizes $MSE(p)$.