Lecture 8

Diagnostics and Model Building

Last week we learned

Decision making and the hat matrix

F-test for significance of regression

t-test for individual regression coefficients

F-test for a group of predictors

Coefficient of multiple determination, R^2

The overall F-test and individual t-tests

- The overall F-test tests if there is a **linear relationship** between the response Y and **any** of the predictors $X_1, X_2, \ldots, X_p$:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_j \neq 0$$

- The test statistic for H_0 is

$$F = \frac{SS_R/p}{SS_E/(n-p-1)} = \frac{MS_R}{MS_E} \sim F_{p,n-p-1}$$

where

$$SS_R = \boldsymbol{y}^T H \boldsymbol{y} - n\bar{y}^2 \qquad SS_E = \boldsymbol{y}^T (I - H)\boldsymbol{y} \qquad H = X(X^TX)^{-1}X^T$$

- An individual t-test tests the significance of **any** individual regression coefficient:

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0$$

- The test statistic for H_0 is

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 c_{jj}}} \sim t_{n-p-1}$$

where

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-p-1} \qquad c_{jj} = (C)_{jj} = ((X^TX)^{-1})_{jj}$$

A partial F-test and R^2

- A partial F-test tests if a group of r predictors can be removed from the model:

$$H_0 : \beta_{p-r+1} = \ldots = \beta_p = 0 \quad \text{vs.} \quad H_1 : \text{ at least one } \beta_j \neq 0$$

- The test statistic for H_0 is

$$F = \frac{SS_R^{extra}/r}{SS_E^{full}/(n-p-1)} = \frac{MS_R^{extra}}{MS_E^{full}} \sim F_{r,n-p-1}$$

where SS_R^{extra} is the extra sum of squares

$$SS_R^{extra} = SS_R^{full} - SS_R^{reduced} = SS_E^{reduced} - SS_E^{full}$$

- The coefficient of multiple determination

$$R^2 = 1 - \frac{SS_E}{SS_T} \in [0,1]$$

- The adjusted coefficient of multiple determination

$$R_{adj}^2 = 1 - \frac{MS_E}{MS_T} = 1 - \frac{SS_E/(n-p-1)}{SS_T/(n-1)}$$

Lecture 8

Diagnostics and model building

Aim: to learn how to find the "best" MLR model

1. Regressions diagnostics
2. Factor variables
3. Variable selection
4. Assessing the predictive ability

Regressions diagnostics

- Determine if the proposed regression model is a valid model by examining regression plots and transforming response and/predictor variables (if needed).
- Determine which (if any) of the data points are leverage and/or outlier points.
- Assess the extent of collinearity among the predictor variables using variance inflation factors.
- Examine if the assumption of constant error variance is reasonable. If not, decide how to overcome this problem.

Leverage points

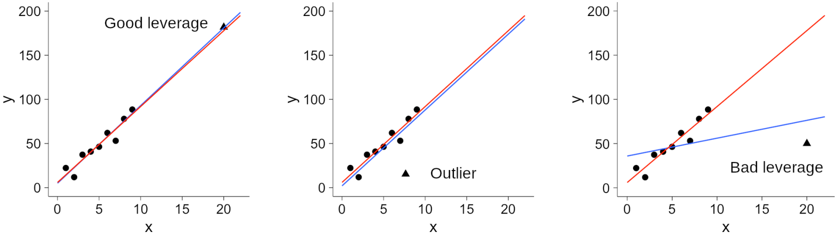
- Points having a considerable influence on the fitted model are called **leverage points**:
  - We have that

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

    where h_{ii} measures the extent to which $\hat{y}_i$ is attracted by y_i: if $h_{ii} \approx 1$, then $\hat{y}_i \approx y_i$.

  - The i-th point is a **leverage point** if

$$h_{ii} > 2 \times \text{average}(h_{ii}) = 2 \times \frac{p+1}{n}$$



Standardised residuals

- The hat matrix H allows us to write the vector of residuals as

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (I - H)\,\boldsymbol{y}$$

Then the least squares estimator $\boldsymbol{E}$ of $\boldsymbol{e}$ is

$$\boldsymbol{E} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (I - H)\boldsymbol{Y}$$

Claim. The sampling distribution of $\boldsymbol{E}$ is

$$\boldsymbol{E} \sim N_n(\mathbf{0}, \sigma^2(I - H))$$

Proof. Since $\boldsymbol{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$, we only need to find the mean and variance of $\boldsymbol{E}$:

$$\mathbb{E}(\boldsymbol{E}) = (I - H)\,\mathbb{E}(\boldsymbol{Y}) = (I - H)X\boldsymbol{\beta} = \mathbf{0}$$

since $HX = X$, and

$$\mathrm{Var}(\boldsymbol{E}) = (I - H)\,\mathrm{Var}(\boldsymbol{Y})\,(I - H)^T = (I - H)\,\sigma^2\,(I - H)^T = \sigma^2(I - H)$$

since $(I - H)(I - H)^T = (I - H)^2 = (I - H)$.

Standardised residuals

- We have shown that

$$\boldsymbol{E} \sim N_n(\mathbf{0}, \sigma^2(I-H)) \quad \Longrightarrow \quad \frac{E_i}{\sqrt{\sigma^2(1-h_{ii})}} \sim N(0,1)$$

- Consequently, we introduce **standardised residuals**, r_i, by

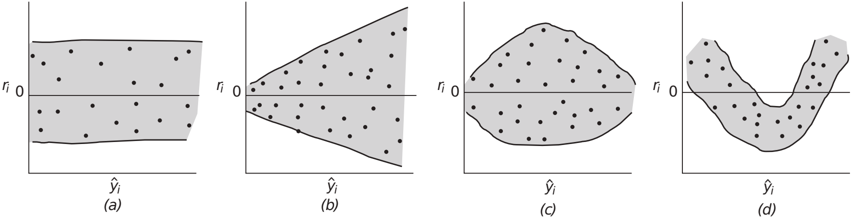
$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$$

- We shall follow the common practice of labelling points as **outliers** if
  - $|r_i| > 2$ in small to moderate size data sets, and
  - $|r_i| > 4$ in large data sets

Standardised residuals

- When a **valid model** has been fit, plots of standardised residuals should have the following features:
  - A random scatter of points around the horizontal axis, since the mean function of the r_i is zero when a correct model has been fit.
  - Constant variability as we look along the horizontal axis.

- A pattern in a residual plot indicates that an incorrect model has been fit, but the pattern itself does not provide direct information on how the model is misspecified.



Cook's distance

- The Cook's distance for a multiple linear regression model is defined by

$$D_i = \frac{(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})^T (\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})}{(p+1)\hat{\sigma}^2}$$

  where $\hat{\boldsymbol{y}}_{(i)}$ denotes the vector of fitted values based on the fit obtained when the ith case has been deleted from the fit.

- It can be shown that

$$D_i = \frac{r_i^2}{p+1} \cdot \frac{h_{ii}}{1-h_{ii}}$$

- A recommended rough cut-off for noteworthy values is $D_i \geq \dfrac{4}{n-p-1}$

Multicollinearity

- How do we choose predictors?

  – In theory, we would like to have individual predictors independent from each other.

  – In practice, predictors are often correlated among themselves.

- **Multicollinearity** is the prevalence of near-linear dependence among the regressors:

  – A predictor that can be expressed in terms of the remaining predictors **does not** contribute any new information about the response.

  – Linear dependence of predictors makes the estimated slopes arbitrary.

  – Linear dependence makes the matrix $X^T X$ non-invertible.

Multicolinearity

- **Example.** Suppose a person's height (in inches) is expressed as a function of arm-span (in inches)
$$\hat{y} = 12 + 1.1\,x$$
- Suppose two people took independent measurements in inches (x_1) and in centimetres (x_2) of the same subjects and both variables have erroneously been included in the same linear regression model
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$
- Since $x_2 = 0.394\,x_1$ we should have $\beta_1 + 0.394\beta_2 = 1.1$
- There are infinitely many possible solutions to $\beta_1 + 0.394\beta_2 = 1.1$, say $\beta_1 = 100$ and $\beta_2 = -251.015$ or $\beta_1 = -50$ and $\beta_2 = 129.695$.
- The computer used to fit the data and to compute parameter estimates cannot distinguish between sensible and nonsensical estimates.

Multicolinearity

- Multicollinearity can be estimated using **variance inflation factors**

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2}$$

  where R_j^2 is the coefficient of multiple determination obtained from regressing x_j on the remaining predictors.

- For example, if $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ then
  - R_1^2 is that of the model $x_{i1} = \beta_0 + \beta_2 x_{i2} + \varepsilon_i$
  - R_2^2 is that of the model $x_{i2} = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$

- Values of VIF greater than 10 indicate multicollinearity problems.

- A few methods of dealing with multicollinearity include:
  - increasing the number of observations
  - combining the linearly dependent predictor variables into one variable
  - eliminating unimportant variables

Lecture 8

Diagnostics and model building

1. Regressions diagnostics
2. Factor variables
3. Variable selection
4. Assessing the predictive ability

Factor variables

- Factor variables in a regression model can be encoded using **indicator variables** taking values 0 and 1.
- For instance, a gender factor can be encoded as

$$X_1 = \begin{cases} 0 & \text{Female} \\ 1 & \text{Male} \end{cases}$$

- In general, a factor variable with n levels requires $n-1$ indicators.
- For instance, a factor variable with levels A, B, and C can be encoded as

$$\begin{array}{lll} \text{A}: & X_1 = 1 & X_2 = 0 \\ \text{B}: & X_1 = 0 & X_2 = 1 \\ \text{C}: & X_1 = 0 & X_2 = 0 \end{array}$$

  or

$$\begin{array}{lll} \text{C}: & X_1 = -1 & X_2 = -1 \end{array}$$

- Indicator variables are also referred to as **dummy variables** or **binary variables**.

Lecture 8

Diagnostics and model building

1. Regressions diagnostics
2. Factor variables
3. Variable selection
4. Assessing the predictive ability

Variable selection

- Our goal is to find the "best" model from a class of multiple regression models using what are called **variable selection methods**.

- The **full model** is the following multiple regression model containing all p potential predictors:

  $$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

  Variable selection methods aim to choose the subset of the predictors that is "best" in a given sense.

- In general, including more predictors lowers the bias of predictions, but increases the variance:
  - Including too many predictors is commonly called **over-fitting**
  - The opposite is called **under-fitting**

- The two key aspects of variable selection methods are:
  - Evaluating each potential subset of p predictors
  - Deciding on the collection of potential subsets

Coefficient of multiple determination

- Recall that R^2 is defined as the proportion of the total sample variability in the response variable, Y, explained by the regression model, that is,

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

- Adding irrelevant predictor variables to the regression equation often increases R^2. To compensate for this an adjusted coefficient of determination, R^2_{adj}, is introduced,

$$R^2_{adj} = 1 - \frac{MS_E}{MS_T} = 1 - \frac{SS_E/(n-p-1)}{SS_T/(n-1)}$$

- It can be shown that adding predictor variables to the current model only leads to an increase in R^2_{adj} if the corresponding partial F-test statistic exceeds 1.

- The usual practice is to choose the subset of predictors that has the **highest** R^2_{adj}. This is equivalent to choosing the subset of predictors with the **lowest** $\hat{\sigma}^2$.

The Akaike's Information Criterion

- The Akaike's Information Criterion (AIC) is a measure of "goodness" of a regression model based on maximising the log-likelihood $L = L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2 | Y)$ of the model and penalising the complexity:

$$\text{AIC} = -2\ln(L) + 2K$$

  where K is the number of estimated parameters in the model, which in our case is $K = p+2$, since $\beta_0, \beta_1, \dots, \beta_p$, and σ^2 are estimated in the fitted model.

- The measure of complexity is necessary since adding irrelevant predictor variables to the regression equation can increase the log-likelihood.

- In case of the ordinary least squares the AIC is written as

$$\text{AIC} = n\ln\left(\frac{SS_E}{n}\right) + 2p$$

  This is the way R calculates the AIC. In general, the **smaller** is the AIC, the better is the model.

Adjusted Akaike's Information Criterion

- When the sample size is small, or when the number of parameters estimated is a moderate to large fraction of the sample size, the AIC has a tendency for over-fitting since the penalty for model complexity is not strong enough.

- In such cases a corrected version of AIC should be used,

$$\mathrm{AIC_C} = \mathrm{AIC} + \frac{2(p+2)(p+3)}{n-p-1}$$

- The general rule is that $\mathrm{AIC_C}$ should be used instead of AIC unless $n/K > 40$.
- When n gets large, $\mathrm{AIC_C}$ converges to AIC, and thus for large samples there is (almost) no difference between $\mathrm{AIC_C}$ and AIC.

Bayesian information criterion

- Another extension of AIC is the Bayesian information criterion defined by

$$\mathrm{BIS} = -2\ln(L) + K\ln(n)$$

- The BIC is similar to AIC except that the penalty $2K$ is replaced by $K\ln(n)$.

- When $n \geq 8$, $\ln(n) > 2$ and so the penalty term in BIC is greater than the penalty term in AIC. Thus, in these circumstances, BIC penalizes complex models more heavily than AIC, thus favouring simpler models than AIC.

- For model selection purposes, there is no clear choice between AIC and BIC:
  - AIC tends to choose models that are too complex due to a small penalty;
  - BIC chooses models that are too simple due to a large penalty.

- A popular data analysis strategy is to calculate R^2_{adj}, AIC, AIC$_\mathrm{C}$, and BIC, and compare the models which minimize AIC, AIC$_\mathrm{C}$, and BIC with the model that maximizes R^2_{adj}.

Finding the “best” model

- There are two distinctly different approaches to choosing the potential subsets of predictor variables: **all possible subsets** and **stepwise methods**.
- The first approach is based on considering all 2^p possible regression models and identifying the subset of the predictors that maximises a measure of fit or minimises an information criterion.
- This second approach is based on examining just a sequential subset of the 2^p possible regression models. Arguably, the two most popular variations on this approach are **backward elimination** and **forward selection**.
- Backward elimination and forward selection consider at most

$$p+(p-1)+(p-2)+\ldots+1=p(p+1)/2$$

models, but in practice both approaches produce the same model in many different situations.

Lecture 8

Diagnostics and model building

1. Regressions diagnostics
2. Factor variables
3. Variable selection
4. Assessing the predictive ability

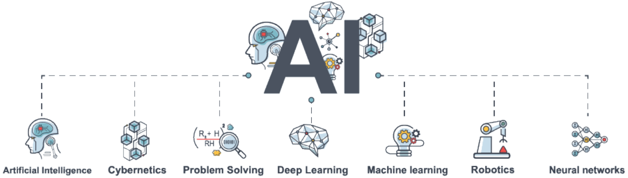
Assessing the predictive ability

- Given that the model selection process changes the properties of the standard inferential procedures, a standard approach to assessing the predictive ability of different regression models is to evaluate their performance on a new data set, i.e. one not used in the development of the models.

- In practice, this is often achieved by randomly splitting the data into a **training** data set, a **validation** data set and a **test** data set.

- The training and validation data sets are used to develop a number of regression models, while the test data set is used to evaluate the performance of these models.

- Then the “best“ model is the one having the smallest MS_E when evaluated on the test data set.



Next week

ANOVA: Single-Factor Experiments