Diagnostics

* All statistical modeling is based on some assumptions.

* It is important to check those assumptions before making any conclusions.

* This is the process of model checking and is based on diagnostics.

* Diagnostics may be graphical or numerical.

Assumptions of the Linear Model

Model Assumption We require that the true model can be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

* Implicit in this assumption is that all of the covariates which affect Y have been included in the model.

* This is generally not true!

* The model results, however, will still hold as long as any unobserved covariates can be considered part of the error term ε.

* This requires that unobserved variables are independent of the observed variables.

Assumptions of the Linear Model

Error Assumptions The validity of inferences in the linear model is based on certain assumptions about the error terms $\varepsilon_1, \ldots, \varepsilon_n$.

* The distribution of the error terms has mean 0.
* The variance of the error terms is σ^2 for any values of the covariates. This is known as the Homogeneity of Variance assumption.
* The error terms are normally distributed.
* The error terms $\varepsilon_1, \ldots, \varepsilon_n$ are independent.

Assumptions of the linear model

Covariate Assumptions

* The covariates are linearly independent.
  - Without his assumption, the least squares estimates cannot be found.
  - When the assumption does not hold we have collinearity.
  - Can also be caused by numerical issues.

* The covariates are assumed to be measured without error.
  - Almost never true in practice!
  - Measurement error is very hard to measure but can cause bias in the estimates.

Residuals

* Residuals are very important in checking model assumptions.

* Although some numerical examinations of residuals may be useful, the most important things are generally graphical diagnostics.

* Patterns in residual plots may often point to such issues as non-linearity, non-homogenous variance, non-normality or non-independence.

* No linear regression analysis is complete without a careful study of residual diagnostic plots.

Hat Matrix

* Recall that the least squares estimates of the parameters are

$$\hat{\boldsymbol{\beta}} \;=\; (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y}.$$

* Hence the fitted values can be written as

$$\hat{\boldsymbol{Y}} \;=\; \boldsymbol{X}\hat{\boldsymbol{\beta}} \;=\; \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y}.$$

* The $n \times n$ matrix $\boldsymbol{P} \;=\; \boldsymbol{X}(\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t$ is called the Projection Matrix or Hat Matrix.

* In other words the i^{th} fitted value can be written as a linear combination of all the original y values

$$\hat{y}_i \;=\; p_{i1}y_1 + p_{i2}y_2 + \cdots + p_{ii}y_i + \cdots + p_{in}y_n$$

Properties of the Hat Matrix

* $\boldsymbol{P}$ is symmetric ($\boldsymbol{P}^t = \boldsymbol{P}$).
* $\boldsymbol{P}$ is idempotent ($\boldsymbol{P}^2 = \boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$)
* Trace($\boldsymbol{P}$)=p+1.
* $\sum_{i=1}^{n} p_{ij} = \sum_{j=1}^{n} p_{ij} = 1.$

Hat Matrix in Simple Regression

* In the simple linear model we have

$$p_{ij} = \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{\sum_k (x_k - \overline{x})^2}.$$

* Of particular importance are the diagonal elements of the matrix which in the simple linear model are

$$p_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_j (x_j - \overline{x})^2}.$$

* This is the weight given the y_i in determining the fitted value $\widehat{y}_i$.

* These important quantities are called the Leverages.

Standardized Residuals

* Recall the raw residuals are given by

$$\boldsymbol{e} \;=\; \boldsymbol{Y} - \widehat{\boldsymbol{Y}} \;=\; (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y}.$$

* Hence the variance of the residuals is

$$\begin{aligned} \mathrm{Var}(\boldsymbol{e}) &= (\boldsymbol{I}_n - \boldsymbol{P})\,\mathrm{Var}(\boldsymbol{Y})\,(\boldsymbol{I}_n - \boldsymbol{P})^t \\ &= (\boldsymbol{I}_n - \boldsymbol{P})\left(\sigma^2 \boldsymbol{I}_n\right)(\boldsymbol{I}_n - \boldsymbol{P})^t \\ &= (\boldsymbol{I}_n - \boldsymbol{P})\,\sigma^2 \end{aligned}$$

* The standardized residuals are then

$$z_i \;=\; \frac{e_i}{\sigma\sqrt{1 - p_{ii}}}.$$

Internally Studentized Residuals

* The standardized residuals cannot be calculated since we do not know σ.

* We can use an unbiased estimator of σ^2, however.

* The most common method is to use the usual estimator of σ^2 from the linear model. This results in the Internally Studentized Residual

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-p_{ii}}}.$$

Externally Studentized Residuals

* Since the use of the studentized residuals is to examine the effect of the i^{th} observation, it may be beneficial to omit that observation when estimating σ^2.

* Hence we get the estimator of σ^2

$$\sigma^2_{(i)} = \frac{\text{SSE}_{(i)}}{n-p-2}$$

where $\text{SSE}_{(i)}$ is the error sum of squares if we omit $(y_i, \boldsymbol{x}_i)$ from the dataset and fit the same model.

* Using this estimator gives rise to the Externally Studentized Residual

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-p_{ii}}}.$$

Univariate Diagnostics and Graphs

* The first step in a regression analysis is generally to examine all of the variables in the model.

* Getting means, standard deviations and ranges of the response and predictor variables is part of this.

* One dimensional graphs such as histograms or boxplots are also very useful to see if there are any outliers in the covariates or response.

Univariate Outliers

* Outliers in a covariate can often indicate a point that will have very high influence on the fitted line.

* Outliers in the response are often points for which the model will not fit well.

* Outliers should always be examined for data-entry errors or other explanations. In some cases it may be wise to omit them from the analysis.

* It is important, however, to not routinely remove outliers unless they are truly errors. Sometimes outliers may really be a sign that something is wrong with the model or they can give important information about the underlying scientific question.

Bivariate Graphs

* In simple regression, one of the most useful plots is a scatterplot of the covariate against the response.

* This can be useful in detecting non-linearity in the model which needs to be corrected.

* It can also show points which are very influential or deviate significantly from the model.

Scatterplot Matrices

* In multiple regression we will usually employ a scatterplot matrix of the response and all covariates.

* Non-linearity in the plots of the response against the covariates can be examined.

* Also we can see if there are strong pairwise correlations between covariates which can sometimes indicate collinearity issues.

Residual Plots

Normal Q-Q Plot of Residuals Plot of the ordered studentized residuals against the N(0,1) quantiles.

* It should be close to the line $y = x$ if normality holds.
* Curvature in the tails indicates a violation of the normality assumption.

Residual-Fitted Value Plot A plot of studentized residuals (y-axis) against the fitted values (x-axis).

* The plot should look like a random scatter about the line $y = 0$ with constant variance.
* A pattern in the plot may indicate violation of one or more assumptions.

Residual Plots

Residual-Covariate Plots Plots of the studentized residuals against each of the covariates.

* Again this should look like a random scatter.
* In simple regression it is identical (in shape) to the residual-fitted value plot.

Residual-Index plot Plot of the studentized residuals against the observation number.

* Observations are usually stored in the order they were collected.
* This plot can detect auto-correlation in the errors violating the assumption of independence.

Outliers

* An outlier is a point which deviates from the model.

* While it is relatively easy to find outliers in univariate datasets and in simple regression, it is harder in multiple regression.

* An outlier in a regression setting may not be an outlier in any of the individual variables.

* Generally outliers will have a large studentized residual in absolute value.

* Typically we will examine a point with $|r_i| > 2$ as a possible outlier.

Influential Points

* Another common problem in regression is a point of high influence.

* Such points are ones that have a substantial impact on the regression parameter estimates.

* In many cases they do not have large residuals.

* One way to find such points is by refitting the model without each observation in turn and looking at the effect that has on the estimates.

* Obviously, for large datasets this is not really feasible!

High Leverage Points

* Outliers in one or more covariates can cause points to be highly influential.

* Recall that in simple regression, the leverages are given by

$$p_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_j (x_j - \overline{x})^2}.$$

* Outliers in the covariate will then have large values of p_{ii}.

* This is true in multiple regression also.

High Leverage Points

* A plot of the leverages p_{ii} against the index i is a useful way to identify points of high leverage.

* A widely used cutoff is to examine any points that have

$$p_{ii} > \frac{2(p+1)}{n}$$

* Note that being a high leverage point alone does not mean that a point will be influential but all high leverage points should be examined to see if they are influential points also.

* Note that high leverage points tend to have small residuals! This occurs because it can be shown that

$$p_{ii} + \frac{e_i^2}{\text{SSE}} \leqslant 1$$

Masking And Swamping

* When there is a single outlier, it is usually easy to spot.

* Masking occurs, however, when we fail to spot an outlier because it is hidden by other outliers in the dataset.

* On the other hand Swamping occurs when we wrongly describe points as outliers.

* This can happen because points of high influence have pulled the fitted model towards themselves making legitimate data points appear to have high residuals.

* It is always important to consider both residuals and leverages to find problematic points in the dataset.

Jackknifing

* The jackknife is a general method to look at the influence of individual points in any statistical modelling.

* It corresponds to removing each observation in turn and re-fitting the statistical model to the rest of the data.

* Deleting a highly influential observation will result in dramatically different results when the observation is included than when it isn't.

* Many measures of influence in regression are based on the idea of jackknifing.

Notation

* In what follows let $\widehat{\beta}_{0(i)}, \widehat{\beta}_{1(i)}, \ldots, \widehat{\beta}_{p(i)}$ be the least squares estimates of the coefficients in the model when the observation $(y_i, x_{1i}, \ldots, x_{pi})$ is removed from the dataset.

* Let $\widehat{\sigma}^2_{(i)} = \text{SSE}_{(i)}/(n-p-2)$ be the unbiased estimator of σ^2 when the i^{th} observation is deleted.

* Finally let $\widehat{y}_{1(i)}, \ldots, \widehat{y_{n(i)}}$ be the fitted values for the n observations form the fitted model without the i^{th} observation.

* Note that we do this prediction for the deleted observation also

$$\widehat{y}_{i(i)} = \widehat{\beta}_{0(i)} + \widehat{\beta}_{1(i)} x_{1i} + \cdots \widehat{\beta}_{p(i)} x_{pi}$$

Cook's Distance

* One of the most widely used measures of influence is due to D. R. Cook in 1977.

* It is a measure of the difference between the fitted values with and without the i^{th} observation

$$C_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{\hat{\sigma}^2(p+1)} \quad i = 1, \ldots, n.$$

* It can be shown that an alternative expression is

$$C_i = \frac{r_i^2}{p+1} \times \frac{p_{ii}}{1-p_{ii}} \quad i = 1, \ldots, n.$$

and so there is no need to actually fit the models omitting each observation.

Cook's Distance

* A practical cut-off that has been suggested is to examine points with $C_i > 1$.

* Plots of C_i against i are often useful as influential points will stand out.

* If all values of C_i are similar then there is likely no influential point and so no further action need be taken.

Welsch & Kuh Statistic

* An alternative measure of influence which just looks at the i^{th} fitted value from the model with and without the i^{th} observation introduced by Welsch & Kuh in 1977.

$$\text{DFITS}_i = \frac{\widehat{y}_i - \widehat{y}_{i(i)}}{\widehat{\sigma}_{(i)}\sqrt{p_{ii}}} \quad i = 1, \ldots, n.$$

* Again this can be calculated without fitting the jackknife models by expressing it as

$$\text{DFITS}_i = r_i^* \sqrt{\frac{p_{ii}}{1 - p_{ii}}} \quad i = 1, \ldots, n.$$

where r_i^* is the externally studentized residual described earlier.

Welsch & Kuh Statistic

* A common procedure for using the statistic to flag influential points is

$$|\text{DFITS}_i| \; > \; 2\sqrt{\frac{p+1}{n-p-1}}$$

* A less formal, but often more useful, diagnostic is again to plot the statistic against i and look for points which are abnormally large in absolute value.

* Since $r_i^* \approx r_i$ in most cases, $|\text{DFITS}_i| \; \approx \; \sqrt{(p+1)C_i}$ and so there is generally no need to look at both statistics.

Hadi's Influence Measure

* Another measure of influence introduced by Hadi in 1992.
* Define the normalized residual $d_i = e_i/\sqrt{SSE}$.
* The Hadi measure is then given by

$$H_i = \frac{p_{ii}}{1-p_{ii}} + \frac{p+1}{1-p_{ii}} \times \frac{d_i^2}{1-d_i^2} \quad i = 1, \ldots, n.$$

* Large abnormal values of H_i indicate an influential value.
* The Hadi measure is best examined graphically by plotting H_i against i.

The Potential Function

* Looking at all of these influence measures, we see that they can be broken into two terms.

* One term is related to the residuals and will be large if the point has a large residual in absolute value.

* The other term is related to $p_{ii}/(1 - p_{ii})$.

* This is called the potential function.

* The potential function is large for points of high leverage (usually outliers in the covariate space).

The Potential-Residual Plot

* Introduced by Hadi in 1992 to classify points as outliers or high leverage points.
* It is a plot of the potential function on the y-axis against the residual function on the x-axis. That is a scatterplot of

$$\frac{p_{ii}}{1-p_{ii}} \quad \text{against} \quad \frac{p+1}{1-p_{ii}} \times \frac{d_i^2}{1-d_i^2}$$

* Most points will be in the lower left corner of this plot.
* Points towards the lower right corner would be classed as outliers, those towards the upper left would be high leverage points and those towards the centre or upper right would be classed as a combination of both.

What to do with Influential Points

* A difficult question is *What should we do with points found to be influential?*

* A common mistake is to automatically discard them.

* This is fine if they are genuine errors but this is often not the case.

* Often outliers, point to incorrectness of the linear model.

* In this case, maybe using a different model, transforming the data, or getting more data may be the correct approach.

The Added Variable Plot

* Recall when we first looked at multiple regression, we considered the effect of adding one extra variable to a simple regression model.

* We can think of multiple regression as adding variables to the model in turn.

* An obvious question is then, whether the new variable improves the fit of the model or not.

* The Added Variable Plot (Mosteller & Tukey, 1977) is a graphical way to examine the effect of adding a new variable to a regression model.

The Added Variable Plot

* The Added Variable Plot for the variable X_{p+1} is a plot of two sets of residuals.

* The first set of residuals (called the Y-residuals) are the residuals from the model to predict Y from the covariates $X_1, \ldots, X_p$.

* The second set of residuals (called the X_{p+1}-residuals) are the residuals from treating X_{p+1} as the response and $X_1, \ldots, X_p$ as the covariates.

* A scatterplot of these two sets of residuals (Y-residuals on the y-axis, X_{p+1}-residuals on the x-axis) is the Added Variable Plot.

The Added Variable Plot

* Since the two sets of residuals in the Added Variable Plot have 0 mean, we can fit a simple regression model without an intercept to the plot.

* The slope of this plot will be exactly the coefficient of X_{p+1} in the multiple regression of Y with covariates $X_1, \ldots, X_{p+1}$.

* If the added variable plot is strongly linear then adding X_{p+1} will improve the fit of the model.

* From the Added Variable Plot, we can also see if there are points which are very influential in the added effect of X_{p+1} in the model.

Residual Plus Component Plot

* Another plot to look at the effect of a single variable X_j in multiple regression.

* It is a scatterplot of $(e + \widehat{\beta}_j X_j)$ on the y-axis against X_j on the x-axis.

* The slope of a simple linear model of the plot is equal to $\widehat{\beta}_j$.

* This plot is more sensitive to departures from the assumption of linearity of the relationship between Y and X_j adjusting for the other covariates.

* It is not, however, as sensitive to influential observations in the regression.