Summary

sagisk

December 12, 2021

1 Chapter 3: Linear Regression

We are given some relationship between response and predictors:
$y = f(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon$

- Where f-is the true, unknown function that characterise the distribution.
- $\epsilon \sim N(0, \sigma^2)$ -is the random noise in the model.
- $\beta_0 = E[Y|X = 0]$ is the intercept term.
- β_1 is the slope-the average increase in y associated with a one-unit increase in X.

Setup: $i = 1, ..., n$ observations of the (x_i, y_i). Goal: Obtain the estimates $\hat{b_0}, \hat{b_1}$ such that the predicted value $\hat{y} = \hat{b}_0 + \hat{b}_1 x_i$ for $i = 1, ...n$ is close to the true data response. Measure of closeness: The residual sum of squares. Where ith residual is given as $e_i = y_i - \hat{y}_i$.

Estimated coefficients:

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Properties: **Unbiased:** In the one set of data the estimated coefficients can over/under-estimate the true coefficients. However, the expected value of the estimates over multiple data sets will be equal to the true coefficients. **Standard Error (SE):** Tells the average amount that the estimates differ from the actual values. Assuming uncorrelated errors ϵ_i with common variance σ^2 we have:
$SE(\hat{b}_0)^2 = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]$
$SE(\hat{b}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

In general σ^2 is unknown so one uses its estimate: $\hat{\sigma} = RSE = \sqrt{RSS/(n-2)}$. -Confidence intervals: SE-s can be used to compute confidence intervals. For example, 95 percent of the intervals $[\hat{b}_1 - 2SE(\hat{b}_1), \hat{b}_1 + 2SE(\hat{b}_1)]$ will contain the true, unknown value of β_1 if we generate repeated samples and construct such interval for each one. -Hypothesis Test: SE-s can be used to perform hypothesis test:

$$H_0 : \beta_1 = 0$$

vs.

$$H_1 : \beta_1 \neq 0$$

To perform the test one should determine if the $\hat{b}_1$ is far enough from 0. Whether the distance of the estimate is far enough from 0 depends on how spread is the estimate i.e. on $SE(\hat{b}_1)$. In practise one computes t-test: $t_{n-2} = \frac{\hat{b}_1 - 0}{SE(\hat{b}_1)}$ and the corresponding p-value. p-value is the probability of observing by chance any number equal to $|t|$ or larger in absolute value, assuming $\beta_1 = 0$.

2 Model accuracy

Residual Standard Error (RSE) $RSE = \sqrt{RSS/(n-2)}$: Is the estimate of the standard deviation of ϵ i.e. estimates the irreducible noise of the true model. For example, if $RSE = 3.26$ than even if we know and use the true coefficients for the model we would still be off in the response by 3.26 units on average. Note: Whether it is a big/small deviation depends on the problem context. R^2 statistic $R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$: Is the variance of the response explained by the model. Here $TSS = \sum(y_i - \bar{y})^2$ is the total variance in Y and RSS is the variability that is left unexplained after performing the regression. Note: $R^2 \in [0,1]$ and whether its value is big or not depends on context. F-statistic $F = \frac{(TSS-RSS/p)}{RSS/(n-p-1)}$: Is used to perform the following hypothesis test:

$$H_0 : \beta_1 = \beta_2 = ...\beta_p = 0$$

vs.

$$H_a : \text{at least one } \beta_j \text{ is non-zero}$$

If no relationship: F statistic take on a value close to 1. If there is a relationship: F statistic take on a value bigger than 1. Note: However, the closeness depends on the problem context. For big n even F-statistic that is a bit larger than 1 might be an evidence against H_0. If n-is small a bigger F-statistic is needed.

F-statistic vs t-test: For one variable they are the same. However, only t-test is not sufficient to answer the question whether there is any relationship between any of the predictors and response. This happens since at 5 percent with 100 prediction 5 of the predictors will haave a small p-value even if there is no relationship. Meanwhile, F-statistic adjusts for the number of predictors.

3 Multiple Linear Regression (MLR):

Similar setup, goal, measure of closeness except the model is given with more parameters (coefficients): $y = f(X) + \epsilon = \beta_0 + \beta_1 X + ... + \beta_p X_p + \epsilon$

Interpretation: One important difference of MLR from simple linear regression (SLR) is that SLR just ignores all other coefficients except b_0, b_1 while in MLR to interpret b_i we don't ignore but assume that the other coefficients are held fixed.

Important question to answer in linear regression:

1. Is at least one of the predictors useful in predicting the response? Answer: Use F-statistic

2. Do all the predictors help to explain Y , or is only a subset of the predictors useful? Answer: Again, using p-values is not the best strategy. Instead one can do

   - Subset selection: Construct all possible models and determine which one is the best. Only feasible if p is small.
   - Forward selection: Begin with null model-in each step add the predictor that results in the lowest RSS (for one predictor model, for two predictor model, etc.)
   - Backward selection: Start with the full model. Remove the predictor with highest p-value. Fit the new $(p-1)$-variable model and remove the predictor with highest p-value and etc.
   - Mixed selection: We start with no variables in the model, and as with forward selection, we add the variable that provides the best fit. We continue to add variables one-by-one. When the p-value for one of the variables in the model rises above a certain threshold, then we remove that variable from the model.

3. How well does the model fit the data? Answer: R^2, RSE.

4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction? Answer: Compute the prediction interval for the response Y. Compute the confidence interval for the $\hat{Y}$.

4 Potential Problems:

1. Non-linearity of the data:
   - Assumption violated: Straight-line (linear) relationship between he predictors and the response.
   - How to check: Residual plots i.e. plot of the residuals vs the fitted values $\hat{y}$. The presence of a pattern may indicate a problem with some aspect of the linear model.
   - Solution: Use non-linear transformations of the predictors such as $\log X, \sqrt{X}, X^2$.

2. Correlation of Error Terms:
   - Assumption violated: The error terms $\epsilon_1, ..., \epsilon_n$ are uncorrelated. The computation of SE of coefficients is based on it.
   - How to check: Plot the residuals from our model as a function of time. If the errors are uncorrelated, then there should be no discernible pattern.
   - Solution:

3. Non-constant Variance of Error Terms:
   - Assumption violated: Error terms have a constant variance, $Var(\epsilon_i) = \sigma^2$. SE, confidence intervals and hypothesis test rely on it.
   - How to check: Plot residuals vs fitted values. The funnel shape is an indicator of heteroscedasticity.
   - Solutions: Transform Y to a concave function such as $\log Y, \sqrt{Y}$.

4. Outliers:
   - Nature: An outlier is a point for which y_i is far from the value predicted by the model.
   - How to check: Residual vs Fitted values plots. Even better option is Studentized residuals (residuals divided by their standard errors) vs Fitted values plot.
   - Solutions: Delete if they are a result of error. Otherwise, they can indicate for some deficiency with the model.

5. High Leverage Points:
   - Nature: An unusual value of x_i instead if y_i (as in outlier).
   - How to check: In simple linear regression can be identified by plots x_1 vs. x_2 predictors and eye-inspecting for the unusual values. In more general settings a *leverage statistic* can be used $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ where, $h_i \in [1/n, 1]$ and big h_i indicates for high leverage. (note: again the formula is for simple regression).
   - Solutions: Delete if they don't indicate some data peculiarity.

6. Collinearity:
   - Assumption violated: The inputs are i.i.d. in particular the independence is voialted.
   - Nature: One of the problems is that collinearity causes the dropoff of the accuracy of the estimates of the regression coefficients and hence, an increase in their standard errors. So, we may fail to reject $H_0 : \beta_j = 0$.
   - How to check: Scatter plots of qunatitative predictors. Look at the correlation matrix. However, a better way is to compute *variance inflation factor* ($VIF \geq 1$). For $VIF \geq 5, 10$ we have a problematic amount of collinearity $.VIF(\hat{\beta}_j)$ is computed by dividing the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own.if fit on its own.
   - Solutions: The first solution is to drop one of the problematic variables. The second solution is to combine the collinear variables together into a single predictor.

References