Multiple Linear Regression Review

Mateo Umaguing
March 7, 2021

1 Fitting the Model

1.1 Data

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

is our feature vector. This follows the joint distribution $P(X_1, X_2, ..., X_p, Y)$ or $P(X, Y)$. Our probability distribution for the model is

$$P(Y|X_1 = x_1, ..., X_p = x_p) = P(Y|X = x)$$

and our model is

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + ..i. + \beta_p x_p.$$

If we believe in our model,

$$y_i = E(Y_i|X = x_i) + \epsilon_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i.$$

1.2 Error Assumptions

The following assumptions regarding the errors must hold:

1. $E(\epsilon_i|X = x_i) = 0$
2. $Var(\epsilon_i|X = x_i) = \sigma^2$
3. ϵ_i's iid
4. Extra: ϵ_i's follow a normal distribution with mean 0 and variance σ^2

1.3 Model

Our model is

$$y = X\beta + \epsilon$$

This means

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} \end{pmatrix} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

A column of 1s should be added to the beginning of X to account for β_0. The least squares estimator is found by minimizing $RSS(\hat{\beta}) = (y - X\beta)^T(y - X\beta)$:

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

1.4 Least Squares Estimators and Variance

Properties of least squares estimators:

1. $\hat{\beta}$ is unbiased ($E(\hat{\beta}) = \beta$).
2. The variance-covariance of $\hat{\beta}$ is $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$
3. If $\epsilon \sim N(0, \sigma^2)$ then $\hat{\beta}_j \sim N(\beta_j, var(\hat{\beta}_j))$.

The variance-covariance matrix contains the variance of the j-th least squares estimator $\hat{\beta}_j$ along the diagonal and $\text{cov}(\hat{\beta}_j, \hat{\beta}_i)$ for all other non-diagonal values.

The residuals $\hat{\epsilon} := y - X\hat{\beta}$. Properties:

1. $\sum_{i=1}^n \hat{\epsilon}_i = 1_n^T\hat{\epsilon} = 0$ in which 1_n is an $n \times 1$ vector of 1s
2. $X^T\hat{\epsilon} = 0_n$

The variance is estimated by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^n \hat{\epsilon}_i^2 = \frac{\hat{\epsilon}^T\hat{\epsilon}}{n-p-1}$$

If the assumptions regarding the errors hold, then

- $E(\hat{\sigma}^2) = \sigma^2$
- $n-p-1$ is the residual D.O.F.

2 Parameter Inference

2.1 Distribution of $\hat{\beta}_j$

Under assumptions 1.-4. of the errors,

$$\hat{\beta}_j \sim N(\beta_j, v_{jj}) \implies \frac{\hat{\beta}_j - \beta_j}{\sqrt{v_{jj}}} \sim N(0,1)$$

When σ^2 is unknown,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{v_{jj}}} \sim t_{n-p-1}$$

2.2 Confidence Intervals

The $100(1-\alpha)\%$ CI for β_j is given by

$$\hat{\beta}_j - t_{\alpha/2,n-p-1}\sqrt{\hat{v}_{jj}} \le \hat{\beta}_j \le \hat{\beta}_j + t_{\alpha/2,n-p-1}\sqrt{\hat{v}_{jj}}$$

For fixed n and p, smaller $\sqrt{v_{jj}}$ means a smaller confidence interval. Therefore smaller $\sqrt{v_{jj}}$ means a more precise estimation of β_j.

2.3 Hypothesis Tests

The two-tailed test is:

$$H_0 : \beta_j = \beta_j^* \text{ vs. } H_1 : \beta_j \neq \beta_j^*$$

When σ^2 is unknown,

$$T = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{v_{jj}}}$$

Reject H_0 if:

1. $|T| \ge t_{\alpha/2,n-p-1}$ or
2. p-value $P(|t| \ge |T|) < \alpha$

2.4 ANOVA

2.4.1 "Overall" F-test

Test:

$$H_0 : E(Y|X = x) = \beta_0 \text{ or } \beta_1 = \beta_2 = ... = \beta_p = 0$$
$$H_1 : E(Y|X = x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \text{ or at least one } \beta_j \neq 0$$

The test statistic is

$$F = \frac{SS_{reg}/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$

The D.O.F. is p for the numerator and $n-p-1$ for the denominator.

$$SS_{total} = SS_{reg} + RSS$$

The total variation in Y is equal to the variation "explained by regression" plus the variation unexplained.

Source of variation	Degrees of freedom (df)	Sum of Squares (SS)	Mean square (MS)	F
Regression	p	SS_{reg}	SS_{reg}/p	$F = \frac{SS_{reg}/p}{RSS/(n-p-1)}$
Residual	$n-p-1$	RSS	$S^2 = \frac{RSS}{n-p-1}$	
Total	$n-1$	$SS_{total} = SYY$		

Aside. $SS_{total} = \frac{(n-p-1)\hat{\sigma}^2}{1-R^2}$

2.4.2 "Partial" F-test

Test the hypothesis that a subset of $k(< p)$ predictors is not significant:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \text{ vs. } H_1 : H_0 \text{ is not true}$$

Our statistic is

$$F = \frac{(RSS(reduced) - RSS(full))/k}{RSS(full)/(n-p-1)}$$

We reject H_0 if:

1. $|F| \geq F_{\alpha,k,n-p-1}$
2. p-value $P(|f| \geq |F|) < \alpha$

It is important to note that the partial F-test tests if a subset of variables is significant *given that the other variables are included in the model.* The extra sum of squares, or the numerator of the F-statistic of the partial F-test $RSS(reduced) - RSS(full)$ is the **reduction in the residual sum of squares** (RSS) when 1+ predictors are added to the model or the **increase in the regression sum of squares** (SS_{reg}) when 1+ predictors are added to the model.

$$RSS(reduced) - RSS(full) = SS_{reg}(full) - SS_{reg}(reduced)$$

2.4.3 Sequential ANOVA

First fit the null model, then a model with X_0 and X_1, then with X_0, X_1, and X_2, and so on. On each model, test the new variable X_j added with $SS_{reg}(X_j|X_{j-1}, X_{j-2}, ..., X_0) = SS_{reg}(X_0, ..., X_j) - SS_{reg}(X_0, ..., X_{j-1})$. The ANOVA function output in R *sequentially* lists SS_{reg} and the sequential F-test. In other words, order matters.

2.5 Multiple and Adjusted R^2 statistics

As a reminder, $R^2 = SS_{reg}/SS_{total}$ is the percentage of variation "explained by" the regression. This value will *always increase if you add another variable*, therefore you cannot use R^2 to determine if a new variable is significant. Adjusted R^2 is a better measure to decide whether or not to add a new variable.

$$R^2_{adj} = 1 - \frac{RSS/(n-k-1)}{SS_{tot}/(n-1)}$$

A variable is useful/important if adjusted R^2 increases. The non-adjusted R^2 value decreases with the addition of *any* variable, however the $n-k-1$ term reduces the reduction in RSS since k increases with more variables.

3 Modeling Categorical Predictors

3.1 Dummy Coding

For a categorical variable with k categories, $k-1$ dummy variables are required. β_0 is the mean response under the reference category. β_1 is the mean response when we change from the reference category to another category. For example, if the origin of a car model is American, Japanese, or European, we can choose American as the reference category and create 2 dummy variables.

$$d_1 = \begin{cases} 1 & \text{car is European} \\ 0 & \text{car is not European} \end{cases}$$

$$d_2 = \begin{cases} 1 & \text{car is Japanese} \\ 0 & \text{car is not Japanese} \end{cases}$$

β_0 is the mean response for American cars. β_1 is the amount of increase in the mean response when changing from an American to a European car. β_2 is the amount of increase in the mean response when changing from an American car to a Japanese car. The model can be written as

$$y_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{car is European} \\ \beta_0 + \beta_2 + \epsilon_i & \text{car is Japanese} \\ \beta_0 + \epsilon_i & \text{car is American} \end{cases}$$

β_1 here can be interpreted as the difference in mean response between American and European cars while β_2 can be interpreted as the difference in mean response between American and Japanese cars. Overall F-tests can be used to make inferences on the categorical predictors.

3.2 Analysis of Covariance (ANCOVA)

ANCOVA is the analysis of a model that combines categorical and numerical predictors.

3.2.1 Additive Model

Suppose we added a third numerical predictor to our model. An additive model does not include interactions between the 'Origin' and 'Weight' predictors.

$$y_i = \begin{cases} (\beta_0 + \beta_1) + \beta_3 x_i + \epsilon_i & \text{car is European} \\ (\beta_0 + \beta_2) + \beta_3 x_i + \epsilon_i & \text{car is Japanese} \\ \beta_0 + \beta_3 x_i + \epsilon_i & \text{car is American} \end{cases}$$

3.2.2 Non-additive Model

A non-additive model includes interactions between the predictors.

$$y_i = \begin{cases} (\beta_0 + \beta_1) + (\beta_3 + \beta_4)x_i + \epsilon_i & \text{car is European} \\ (\beta_0 + \beta_2) + (\beta_3 + \beta_5)x_i + \epsilon_i & \text{car is Japanese} \\ \beta_0 + \beta_3 x_i + \epsilon_i & \text{car is American} \end{cases}$$

A partial F-test can inform which model is appropriate.

$$H_0 : \beta_4 = \beta_5 = 0 \text{ vs. } H_1 : \beta_4 \text{ or } \beta_5 \neq 0$$

Under H_1, $y_i = \beta_0 + \beta_1 d_{1i} + \beta_2 d_{2i} + \beta_3 x_i + \beta_4 d_{1i} x_i + \beta_5 d_{2i} x_i + \epsilon_i$. The F-statistic is

$$F = \frac{(RSS(reduced) - RSS(full))/k}{RSS(full)/(n-p-1)}$$

in which $k = 2$ (number of coefficients in H_0), $p = 5$ (number of coefficients in the full model other than the intercept), and $n = 392$ (number of obs.).

4 Predictions

Predictions are made using the following formula:

$$\hat{y} = x^T \hat{\beta}$$

$\hat{E}(Y|X = x)$ is the mean response value given $X = x$. We can find the mean and individual values of the response given $X = x_0$ by using our formula and confidence intervals.

4.1 Mean response

For the mean response, we find

$$\hat{\mu}_0 = x_0^T \hat{\beta}$$

in which $\hat{\mu}$ is an estimate of $\mu_0 = E(Y|X = x)$. The variance of the estimator is $Var(\hat{\mu}_0) = \sigma^2 x_0^T (X^T X)^{-1} x_0$. We use confidence intervals for the mean response. The $100(1-\alpha)\%$ CI is given by

$$\hat{\mu}_0 - t_{\alpha/2,n-p-1}\sqrt{\hat{\sigma}^2 x_0^T (X^T X)^{-1} x_0} \leq \mu_0 \leq \hat{\mu}_0 + t_{\alpha/2,n-p-1}\sqrt{\hat{\sigma}^2 x_0^T (X^T X)^{-1} x_0}$$

4.2 Individual response

For the individual response, we find

$$\hat{y}_0 = x_0^T \hat{\beta}$$

in which $\hat{y}_0$ is a point estimate of Y_0. The variance of the estimator is $Var(\hat{y}_0) = \sigma^2(1 + x_0^T (X^T X)^{-1} x_0)$. We use prediction intervals for the individual response. The $100(1-\alpha)\%$ CI is given by

$$\hat{y}_0 - t_{\alpha/2,n-p-1}\sqrt{\hat{\sigma}^2(1 + x_0^T (X^T X)^{-1} x_0)} \leq \mu_0 \leq \hat{y}_0 + t_{\alpha/2,n-p-1}\sqrt{\hat{\sigma}^2(1 + x_0^T (X^T X)^{-1} x_0)}$$

4.3 Discussion

Prediction intervals are larger and rely on the fact that the errors are normally distributed.

5 Regression Diagnostics

We must make some assumptions regarding the residuals:

1. X and ϵ are independent
2. $E(\epsilon) = E(\epsilon|X = x) = 0$
3. $Var(\epsilon) = Var(\epsilon|X = x) = \sigma^2 = Var(Y|X = x)$

Some assumptions regarding the errors must also be satisfied:

1. $E(\epsilon_i|X = x_i) = 0$
2. $Var(\epsilon_i|X = x_i) = \sigma^2$
3. ϵ_i's are iid.
4. The ϵ_i's follow a normal dist. with mean 0 and variance σ^2

5.1 Regression Diagnostics

5.1.1 Residual Analysis

We must have:

1. Constant variance of residuals - use residuals vs. fitted values plot if there are no high leverage points. Use scale-location plot (standardized residuals vs. fitted values) if there are high leverage points.
2. Residuals follow a normal distribution - use normal qqplot.
3. Residuals are independent - use residuals vs. time (or predictor) plot.

Use **marginal model plots** to diagnose problems. The consequences of faulty assumptions are as follows:

1. If the model structure is incorrect, $\hat{\beta}_j$ and $\hat{y}_i$ are inaccurate.
2. If the residuals do not follow a normal distribution but if the sample size is large enough, we can get approximately unbiased estimates of coefficients and confidence intervals/p-values for t- and F- tests. However, confidence interval predictions are invalidated.
3. If the residuals do not have constant variance, all inference tools are invalidated.
4. If the residuals are dependent, then all inference tools are invalidated.

5.1.2 Pesky Points

Pesky points can be:

1. High leverage points - unusually large effect on estimated model
2. Outlier - does not follow pattern of data

The leverage score of the i'th observation **in simple linear regression** can be found using

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

This is the ii'th observation of the $n \times n$ hat matrix in which

$$h_{ij} = \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}\right)$$

A high leverage point is one in which $h_{ii} > 4/n$. In **multiple linear regression**, the hat matrix can be found with the formula

$$H = X(X^T X)^{-1} X^T$$

The properties follow:

- Of size $n \times n$
- Symmetric ($H^T = H$)
- The sum of the diagonals is $\sum_{i=1}^{n} h_{ii} = p$ in which p is the total number of coefficients in the model including the intercept.

A high leverage point is an observation with a high leverage score

$$h_{ii} > \frac{2p}{n}$$

A bad high leverage point is one in which the standardized residual is outside of (-2,2) for small to moderate-sized data sets, and (-4, 4) for large data sets. Outliers are observations in which the corresponding standardized residual is within these bounds.

5.2 Multicollinearity

This means that the predictors are correlated with each other. If one is a linear combination of other predictors, we will get an error in R. A symptom of multicollinearity is when one variable is removed or added and the other variables change and/or the p-value changes. We can use color maps or variance inflation factors to check for this.

5.2.1 Color Maps

These compute the correlation between X_j and $X_{j'}$ values. Dark blue indicates a correlation close to 1 while dark red indicates a correlation close to -1. Ideally, correlations are smaller than 0.6. If non-diagonal elements are dark-colored, then there is a correlation between the variables.

5.2.2 Variance Inflation Factor (VIF)

The VIF is obtained as follows:

1. Fit a model with X_j as the response
2. Calculate multiple R_j^2 of the model
3. $\text{VIF}(X_j) = \frac{1}{1-R_j^2}$

This value can be interpreted as the correlation between a linear combination of the other variables and X_j. It can be shown that:

$$Var(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \frac{\sigma^2}{(n-1)S_{xj}^2}, j = 1, ..., p$$

in which

$$S_{xj}^2 = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)/(n-1)$$

is the sample variance of x_j, n and σ^2 is the variance of the error. Larger R_j values imply larger $Var(\hat{\beta}_j)$. The smallest value for VIF is 1 and values greater than 5 generally mean collinearity problems with X_j.

6 Remedies for Faulty Assumptions

A model is incorrect if:

1. The residuals do not have constant variance
2. The residuals are not independent

The model can be corrected with certain remedies as detailed below.

6.1 Transformations

A transformation can assist if the model structure is incorrect (change the mean function to be linear), if the residuals do not have constant variance, or if the residuals do not follow a normal distribution. The strategy is as follows:

1. Transform only Y
2. Transform only predictors (X columns)
3. Transform both Y and X
4. Use diagnostic columns to determine best transformation

6.1.1 Inverse Response Plot

We must discover the function $g(z)$ such that $Y = g(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon)$ and then invert it to apply this function to Y. An inverse response plot considers transformations

$$Y^\lambda \text{ where } Y^0 = \ln Y.$$

The invResPlot() function gives the residual sums of squares for $\hat{E}(Y^\lambda|X)$ for various values of λ. The marginal model plots plot the data, model, and the data trend.

6.1.2 Box-Cox Transformation

7 Variable Selection

Variable selection is useful when we believe that only certain predictors have coefficients different than zero and/or $n < p$. **Sparsity** is the concept that only a few predictors will be important.

7.1 Subset Selection Algorithms

7.1.1 Best Subset Selection

Incrementing a model of size 0 to size $k \leq p$ and finding the best model out of all $\binom{p}{k}$ models in terms of RSS yields the best subset selection algorithm. Find the best $p+1$ models $\mathscr{M}_0, ..., \mathscr{M}_p$.

7.1.2 Forward Selection

Incrementing a model of size 0 to size $k < p$ and finding the best model out of all $p-k$ models in terms of RSS yields the forward selection algorithm. Pick the best among these $p-k$ models and label it $\mathscr{M}_{k+1}$.

7.1.3 Backward Selection

Do that but backwards.

7.2 Selection Criteria

Two popular model selection criteria are Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC)

$$AIC = n \ln\left(\frac{RSS}{n}\right)$$

The best model has the smallest AIC value.

$$BIC = n \ln\left(\frac{RSS}{n}\right) + \ln(n)k$$

The best model has the smallest BIC value. This is for $n > 8$ and $\ln n > 2$. BIC favors simpler model. The probability BIC chooses the correct model increases as n grows.

8 Additional Topics