Notes on Multiple Regression and Model Selection

Recap of Multiple Regression

The Model

The multiple linear regression model is given by:

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

where the error terms are assumed to be normally distributed:

$$\epsilon \sim N(0, \sigma^2 I_n)$$

We have n observations $(y_i, x_{i1}, \ldots, x_{ip})$ for $i = 1, \ldots, n$. The design matrix X is considered fixed (unknown). The parameter vector β is a constant but unknown parameter. Y is the observed response vector.

Assumptions & Diagnostics

Key assumptions to check with diagnostics on the residuals ($\hat{\epsilon}$) are:

- **Linearity**: The relationship between predictors and the response is linear.
- **Influential outliers**.
- **Homoscedasticity**: The variance of errors is constant (σ^2).
- **Normality of errors**.

Common issues to diagnose include outliers and influential points. Weighted Least Squares (WLS) can be used if homoscedasticity is violated, where we assume $E(\frac{\epsilon_i}{w_i})^2 = \sigma^2$.

Least Squares Estimation

The method of least squares minimizes the sum of squared errors:

$$\epsilon^T \epsilon = \sum_{i=1}^{n} \epsilon_i^2$$

This leads to the normal equations:

$$(X^T X)\beta = X^T Y$$

The least squares estimate for β is:

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

The residuals are calculated as $\hat{\epsilon} = y - X\hat{\beta}$.

Properties of the Estimator

The least squares estimator $\hat{\beta}$ is unbiased and its variance is:

- Expectation: $E[\hat{\beta}] = \beta$
- Variance: $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$

An unbiased estimator for the error variance σ^2 is the Mean Squared Error (MSE) or Mean Squared Residual (MSR):

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(p+1)} = \frac{SSR}{n-(p+1)}$$

where $E[s^2] = \sigma^2$.

ANOVA

The total variation in the response can be decomposed as:

- Total Sum of Squares (SST): $SST = \sum(y_i - \bar{y})^2$
- Sum of Squared Residuals (SSR): $SSR = \sum(y_i - \hat{y}_i)^2$
- Sum of Squares due to Regression (SSReg): $SSReg = SST - SSR$

The coefficient of determination, R^2, measures the proportion of variance in Y that is predictable from X.

$$R^2 = 1 - \frac{SSR}{SST}$$

Interpretation and Geometry

- **Coefficients**: β_j is the change in the mean of Y for a one-unit change in x_j, holding all other predictors constant. No causal relation is claimed in regression.
- **Collinearity**: If predictors are highly correlated, X^TX is near-singular, inflating the variance of $\hat{\beta}$.
- **Interaction**: An interaction between x_1 and x_2 means the effect of x_1 on Y depends on the level of x_2.
- **Geometry**: The fitted values $\hat{y} = X\hat{\beta}$ are the projection of the response vector y onto the subspace spanned by x. The residual vector $\hat{\epsilon}$ is orthogonal to this subspace, which implies $\epsilon^TX = 0$.

Distributional Properties under Normality

Assuming the error terms ϵ_i are independent and identically distributed as $N(0, \sigma^2)$, we have the following key results for the least squares estimators:

- The estimators $\hat{\beta}$ and s^2 (also denoted as $\hat{\sigma}^2$) are **independent**.
- The sampling distribution of the coefficient vector $\hat{\beta}$ is a multivariate normal distribution:
$$\hat{\beta} \sim \mathcal{N}_{p+1}\left(\beta, \sigma^2(X^T X)^{-1}\right)$$
- The sampling distribution for the variance estimator s^2 is related to the chi-squared distribution:
$$\frac{(n-(p+1))\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-(p+1)}$$
- These two results will help us in constructing **confidence intervals** and conducting **hypothesis tests** for the regression coefficients.
- **ANOVA**, $\frac{\text{MSReg}}{\text{MSE}} \sim F(distribution)$

Inference

Assuming errors are normally distributed, we can make inferences about the parameters. For any coefficient β_j:

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-(p+1)}$$

This is used for hypothesis tests and constructing confidence intervals for the coefficients.

Let $\theta = x_0^T \beta$ be the mean response for a given vector of predictors x_0.

- Point Estimate: $\hat{\theta} = x_0^T \hat{\beta}$
- Variance: $Var(\hat{\theta}) = \sigma^2 x_0^T (X^T X)^{-1} x_0$
- Standard Error: $se(\hat{\theta}) = \sqrt{s^2 x_0^T (X^T X)^{-1} x_0}$

A $(1-\alpha)100\%$ confidence interval for the mean response θ is:

$$\hat{\theta} \pm t_{\alpha/2, n-(p+1)} \cdot se(\hat{\theta})$$

A prediction interval is for a single future observation $y_0 = x_0^T \beta + \epsilon_0$. The variance of the prediction error $(\hat{y}_0 - y_0)$ is:

$$Var(\hat{y}_0 - y_0) = Var(\hat{y}_0) + Var(\epsilon_0) = \sigma^2(x_0^T (X^T X)^{-1} x_0 + 1)$$

The prediction interval is wider than the confidence interval:

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

Confidence vs. Prediction Intervals

- **Confidence Interval (CI)** for the **expected value** of Y for a given x_0. This is an interval for the mean response $E[Y|x_0]$.
- **Prediction Interval (PI)** for an **individual value** of Y for a given x_0. This is an interval for a single future observation, which includes both the uncertainty in the model and the inherent random error.

A scatter plot will show the confidence interval for the mean response at x_0 (narrower bar, for θ) and the prediction interval for an individual response (wider range of points).

Model Setup

Our model is based on a set of observations $(x_1, y_1), \ldots, (x_n, y_n)$. The general form of the model for an observation i is:

$$y_i = \underbrace{\beta_0 + \beta_1 x_i + \ldots}_{\theta} + \epsilon_i$$

$\theta_1 = \theta + \epsilon_i$ For a prediction, denoted $\hat{\theta}_1$, we have:

$$E[\hat{\theta}_1] = \theta$$
$$Var(\hat{\theta}_1) = Var(\hat{\theta}) + Var(\epsilon_0)$$

Derivation of the Prediction Interval

Given a set of observations $(x_1, y_1), \ldots, (x_n, y_n)$, our model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Derivation of the Prediction Interval

Here is the step-by-step derivation from your notes:

We want an interval for $\theta_1 = x_0^T \beta + \epsilon_0$

Point estimate $\implies \hat{\theta}_1 = x_0^T \hat{\beta}$ since $\boxed{E[\epsilon_0] = 0}$

$\text{Var}(\hat{\theta}_1 + \epsilon_0) = \text{Var}(\hat{\theta}_1) + \text{Var}(\epsilon_0)$ [since $\hat{\theta}_1$ and ϵ_0 are independent]

$= x_0 (X^T X)^{-1} x_0^T \sigma^2 + \sigma^2$

Estimate σ^2 by s^2 with $n - (p+1)$ d.f.

PI for θ_1 is: $x_0^T \hat{\beta} \pm t_{n-(p+1),1-\frac{\alpha}{2}} \cdot s \sqrt{x_0^T (X^T X)^{-1} x_0 + 1}$

Model Selection

Given several possible predictors $x_1, x_2, \ldots, x_{1000}$. Should you take all of them in your model? Why take a bigger (more predictors) model?

1. Full model (Correct)

Let the dimensions of the predictor matrices be $\dim(X) = (n, p)$ and $\dim(x) = (n, 1)$. The expected value for the full model is:

$$E[Y] = \begin{pmatrix} X & x \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix}$$

If we take the smaller model $Y = X\beta + \epsilon$, the fitted value is $\hat{y} = X\hat{\beta}$ from the smaller model.

Expectation of the Smaller Model's Fit

The expected value of the fitted $\hat{y}$ from the smaller model, under the assumption that the full model is true, is:

$$\begin{aligned} E[\hat{y}] &= X(X^T X)^{-1} X^T (X\beta + Xb) \\ &= X\beta + X(X^T X)^{-1} X^T xb \\ &\neq X\beta + xb \end{aligned}$$

This is a **biased estimate**.
[Intuitively it means we are making a mistake systematically].

$$\begin{aligned} m_{pg} &\sim (wt \cdot hp) + wt^2 + hp^2 \\ m_{pg} &\sim (wt \cdot hp) + wt^2 + hp^2 \end{aligned}$$

So it is better to include more predictors to avoid bias.

The Apparent Benefit of Adding Predictors

We have shown that $SSR < SST$. With a larger number of predictors, the Sum of Squared Residuals (SSR) decreases.

$$R^2 = 1 - \frac{SSR}{SST}$$

As SSR decreases, R^2 increases. Explaining more of variation in y by regressing on more predictors.

Why do we not add as many predictors as we can?

1. **Data Collection and Analysis:** It is often difficult and expensive to collect data for many variables. Analyzing a model with too many predictors can also be practically challenging.

2. **Difficulty in Interpretation:** Models with a large number of predictors become hard to interpret. The estimated coefficients can be unstable. Remember the example of $\hat{\beta}_1$ changing its sign after adding another predictor, x_2.

3. **High Variance of Coefficients:** The variance of the estimated coefficient vector $\hat{\beta}$ is given by:
$$\text{Var}(\hat{\beta}) = (X^T X)^{-1}\sigma^2$$
When there are a lot of predictors, particularly when $p \approx n$ or exceeds it ($p > n$), the matrix $(X^T X)$ becomes **near-singular** or **exactly singular**.

   As a result, its inverse is unstable or its elements become very large. This inflates the variance of the estimates, and consequently, we will have very large and unreliable **confidence intervals (CIs)**.

4. **High Variance of Predictions:** It's not just the coefficients that become unstable; the predictions themselves also suffer from high variance. The vector of fitted values and its variance-covariance matrix are:
$$\hat{y} = X(X^T X)^{-1}X^T Y$$
$$\implies \text{Var}(\hat{y}) = X(X^T X)^{-1}X^T\sigma^2$$
The confidence intervals (CI) and prediction intervals (PI) are built from this variance.
$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\text{Var}(\hat{y}_i) &= \frac{1}{n}\text{Tr}\left(\text{Var}(\hat{y})\right)\\ &= \frac{\sigma^2}{n}\text{Tr}\left(X(X^T X)^{-1}X^T\right)\\ &= \frac{p+1}{n}\sigma^2\end{aligned}$$
This result shows that the average variance is directly proportional to the number of parameters $(p+1)$. A **larger** p leads to **larger variance** in our predictions.

The Bias-Variance Trade-off

Clearly, we want a criterion that will balance the **bias** (a problem for smaller, overly simple models) with the **variance** (a problem for larger, overly complex models).

The Mean Squared Error (MSE) is one conceptual possibility for a criterion, as it can be decomposed into bias and variance:

$$\text{MSE} = (\text{Bias})^2 + \text{Variance}$$

In practice, we use several criteria that penalize model complexity to approximate this balance. Here are some of the most common ones:

1. **Adjusted** R^2: A modification of R^2 that adjusts for the number of predictors in the model. It penalizes the addition of useless variables.

$$R^2_{\text{adj}} = 1 - \frac{SSR/(n-(p+1))}{SST/(n-1)}$$

2. **AIC (Akaike Information Criterion)**: A criterion based on information theory. We seek to find the model with the lowest AIC.

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

   Here, $\hat{L}$ is the maximized value of the likelihood function for the model, and p is the number of estimated parameters. The term $-2\ln(\hat{L})$ is a measure of goodness of fit, while $2p$ is the penalty for complexity.

3. **BIC (Bayesian Information Criterion)**: Also known as the Schwarz Bayesian Information Criterion. Like AIC, the goal is to find the model with the lowest BIC.

$$\text{BIC} = -2\ln(\hat{L}) + p\ln(n)$$

   The penalty term in BIC, $p\ln(n)$, is stricter than that of AIC for any $n \geq 8$. This means BIC tends to favor simpler models more strongly than AIC does.

Model Selection

The standard Sum of Squared Residuals (**SSR**) and the coefficient of determination ($\mathbf{R^2}$) are **not appropriate** for comparing models with different numbers of predictors.

- As the number of predictors, p, increases ($\uparrow$), the SSR decreases ($\downarrow$) and R^2 increases ($\uparrow$).

This raises the critical question: How much does the variance increase for an additional predictor?

Let's consider the distributional properties. Under the assumption of normality for the error terms:

- A model with p predictors has SSR_1, for which $\frac{SSR_1}{\sigma^2} \sim \chi^2_{n-(p+1)}$.
- A model with $p+1$ predictors has SSR_2, for which $\frac{SSR_2}{\sigma^2} \sim \chi^2_{n-(p+2)}$.

The difference in the reduction of the sum of squares is $\sigma^2\chi^2$, provided the new predictor is not a linear combination of the existing predictors.

A key insight is that the Mean Squared Residuals (MSR) for different models should be comparable estimates of the true error variance, σ^2.

$$E\left[\frac{SSR_1}{n-(p+1)}\right] \approx E\left[\frac{SSR_2}{n-(p+2)}\right]$$

Claim: $SSR/df = MSR$ is comparable across models.

Model Selection Criteria

1. **Adjusted $\mathbf{R^2}$**: This criterion is based on the MSR.

   $$R^2_{\text{adj}} = 1 - \frac{MSR}{MST}$$

   where the Mean Square Total is defined as:

   $$MST = \frac{1}{n-1}\sum(y_i - \bar{y})^2$$

   The procedure is to compare the R^2_{adj} across models and pick the one with the **highest $\mathbf{R^2_{adj}}$**.

2. **Mallows's $\mathbf{C_p}$**: A criterion that estimates the mean squared prediction error. $\tilde{p} = p + 1 =$ total number of parameters

   $$C_p = \frac{SSR}{s^2} - (n - 2\tilde{p})$$

   Here, SSR is for the current model, s^2 is the MSE from the *full model*, and $\tilde{p} = p + 1$ is the total number of parameters. For a model with low bias, we expect $E[SSR_p] \approx \sigma^2(n - (p + 1))$, which implies that for a good model:

   $$C_p \approx \tilde{p}$$

   A **low $\mathbf{C_p}$ value is desired**. The procedure is to calculate C_p for every subset of predictors and pick the model where the value of $C_p - \tilde{p}$ is **lowest** or $\tilde{p} - C_p$ is **highest**.ex C_1=5 possibility, C_2=10 possibility.

   A plot of C_p vs. $\tilde{p}$. Good models will have points lying close to the reference line.

3. **AIC (Akaike Information Criterion)**: A widely used criterion based on information theory.

   $$\text{AIC} = -2\ln(\hat{L}) + 2\tilde{p}$$

   Adding more predictors leads to a **higher maximized likelihood ($\hat{L}$)**, which in turn makes the $-2\ln(\hat{L})$ term lower (indicating a better fit). The AIC **balances** this goodness-of-fit term by adding a penalty for the number of parameters, $2\tilde{p}$. The model with the **lowest AIC** is preferred. The model with the **lowest AIC** is preferred.

For a linear regression model with normally distributed errors, the maximized log-likelihood function is:

$$\ln(\hat{L}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\frac{SSR}{n}\right) - \frac{n}{2}$$

The likelihood function is evaluated at the maximum likelihood estimates (MLEs). The MLE for the error variance, $\hat{\sigma}^2$, is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} = \frac{SSR}{n}$$

The maximized likelihood function, $\hat{L}$, can be written by substituting this estimate:

$$\begin{aligned}\hat{L} &= \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}} \exp\left\{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right\} \\ &= \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}} \exp\left\{-\frac{n}{2}\right\}\end{aligned}$$

Taking the natural logarithm gives the maximized log-likelihood, which is used in the AIC and BIC formulas:

$$\ln(\hat{L}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\frac{SSR}{n}\right) - \frac{n}{2}$$

Substituting this into the general formula for AIC gives us a version that is directly comparable across different linear models:

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

4. **BIC (Bayesian Information Criterion)**: Also known as the **Schwarz Bayesian Information Criterion**, BIC is another popular metric that tends to favor simpler models more strongly than AIC does.

$$\text{BIC} = -2\ln(\hat{L}) + p\ln(n)$$

Using the same log-likelihood function, this can be expanded to:

$$\text{BIC} = n\ln\left(\frac{SSR}{n}\right) + (\ln n)p + n\ln(2\pi) + n$$

The selection rule is to look at all subsets of predictors and find the model with the **lowest BIC**. p predictors , approx number of submodels is 2^p

The order of predictors in a model does not matter (e.g., $y = ax_1 + bx_2$ is the same as $y = bx_2 + ax_1$). The goal is to fix a criterion (e.g., R^2_{adj}, C_p, AIC, BIC) and find the subset of predictors that gives the best(highest or lowest) value according to that criterion.

However, checking all subsets is **extremely time-consuming**, as it requires fitting and evaluating 2^p models. For this reason, automated search algorithms are often used.

Forward Selection Algorithm

Let's use AIC as our criterion for this example. Forward selection is a greedy algorithm that works as follows:

- **Step 0:** Start with the null model (intercept only, $p = 0$) and calculate its AIC, let's call it AIC_0.
- **Step 1:** Fit all p models of order 1 (i.e., with one predictor: $Y \sim x_1, Y \sim x_2, \dots$). Choose the one with the **lowest AIC**. Let's say the model $Y \sim x_5$ is selected, with an AIC value of AIC_1.
- **Step 2:** Now consider all models that contain x_5 and one other predictor (e.g., $Y \sim x_1 + x_5$, $Y \sim x_2 + x_5$, etc.). From this set of $p - 1$ models, take the one with the lowest AIC. Let's say this is $Y \sim x_2 + x_5$, with an AIC value of AIC_2.
- **Stopping Rule:** Compare the AIC from the current step to the previous one.
  - If $AIC_2 < AIC_1$: The model has improved. **Continue** the process.
  - If $AIC_2 \geq AIC_1$: The model did not improve. **Stop** the algorithm and choose the model from the previous step (in this case, $Y \sim x_5$).
- **Step 3 and beyond:** If we continue, we would now consider all models with x_2, x_5, and one other predictor, finding the one with the lowest AIC, say AIC_3. We would then check if $AIC_3 < AIC_2$ to decide whether to proceed or stop.

This process reduces the computational complexity from $O(2^p)$ to $O(p^2)$. However, a major drawback is that this algorithm **may not reach the "best" model** from all possible subsets. It is a greedy search and can miss optimal combinations of predictors.This is the price to pay for a faster algorithm.

0.1 Backward Elimination

This algorithm is the inverse of forward selection. It begins with the full model and iteratively removes the least consequential predictor.

1. **Start with the full model** containing all p predictors: $Y \sim x_1 + x_2 + \cdots + x_p$. Calculate its AIC, let's call it AIC_0.
2. **Step 1:** Fit all p models that have exactly one predictor removed. Compare the AIC values of these p models with AIC_0.
3. **Decision:** Find the model from Step 1 with the lowest AIC, say AIC_1. This corresponds to dropping the predictor x_i that most improves the model fit.
   - If $AIC_1 < AIC_0$, then permanently drop predictor x_i from the model. **Continue** to the next step with this new, smaller model as your baseline.
   - If $AIC_1 \geq AIC_0$, then no single predictor can be removed to improve the model. **Stop** the algorithm and select the current model.
4. **Continue...** Repeat the process of identifying and removing the least useful predictor until the stopping criterion is met.

0.1.1 Stepwise Regression

This is a hybrid approach. At each step, the algorithm considers both adding a variable (a forward step) and removing a variable (a backward step). This allows it to explore more of the 2^p space of all possible models compared to the simpler one-way paths of pure forward or backward selection.

Using Partial Correlation for Forward Selection

This concept can be used directly as a criterion in a forward selection algorithm. At each step, for all predictors not currently in the model, we would calculate their partial correlation with the response, given the predictors that are in the model.

Residuals from the model $y \sim \text{lm}(x_1 + x_4)$ are $\hat{\varepsilon}_{y\cdot 14}$.

Residuals from the model $x \sim \text{lm}(x_1 + x_4)$ are $\hat{\varepsilon}_{3\cdot 14}$.

For example, if x_1 and x_4 are in the model, we would calculate $r_{y2\cdot 14}$, $r_{y3\cdot 14}$, and $r_{y5\cdot 14}$.

The selection rule is: **Take the next predictor to enter the model as the one which has the highest absolute partial correlation**.

Correlation between $\hat{\epsilon}_{y3\cdot 14}$, and $\hat{\epsilon}_{y\cdot 14}$ is partial correlation of y and x3 given x1 and x4.

Calculating Partial Correlation Given the Rest

To find the partial correlation for a specific predictor, say x_1, given all other predictors currently in the model (e.g., $x_2, \ldots, x_5$), we follow a two-step regression process:

First, we find the residuals from regressing the response, y, on all predictors *except* for x_1:

$$y \sim \text{lm}(x_2 + \cdots + x_5) \implies \hat{\epsilon}_{y\cdot 2345}$$

Next, we find the residuals from regressing x_1 on those same predictors:

$$x_1 \sim \text{lm}(x_2 + \cdots + x_5) \implies \hat{\epsilon}_{1\cdot 2345}$$

The partial correlation is the simple correlation between these two sets of resulting residuals:

$$r_{y1\cdot 2345} = \text{corr}(\hat{\epsilon}_{y\cdot 2345}, \hat{\epsilon}_{1\cdot 2345})$$

Similarly, one could calculate $r_{y2\cdot 1345}$ and so on for each predictor in the model.

In R the reporting is partial correlation of y of each predictor given the rest.

After fitting the y on all other predictors what is the residual correlation with the particular predictor. eg. Taken the contribution of $x_2, .., x_5$, what is the additional contribution of x1.

Those predictor with low absolute partial correlation can be dropped.