STA302 Notes

Ian Zhang

July 29, 2024

Contents

1 Simple Linear Regression Models

1.1 SLR Basics

Definition (Linear Regression)**.** Process of estimating a linear relationship between a dependent and independent variable(s)

1.1.1 Estimation of a Trend

Given

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Y is the random response variable
- X is the fixed predictor variable
- ε is the random error

We collect sample data

$$(x_1, y_1), \ldots, (x_n, y_n)$$

to estimate β_0 and β_1.
However, we don't know the true population errors

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

since we don't know the true β_0 and β_1, so we consider the estimated errors (called **residuals**)

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = y_i - \hat{y}_i$$

where $\hat{y}_i$ is an estimated mean. We measure total error around the estimated trend using Residual Sum of Squares (RSS)

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

1.2 Ordinary LSE

The line of best fit should fit snugly among the data points, so we want to *minimize* the amount of error. To do this, we minimize RSS.

- Find line of best fit by finding estimate for β_0 and β_1 that minimize RSS
- We square the residuals so cancellation of positive and negative residuals is prevented

To perform LSE, we take partial derivatives WRT $\hat{\beta}_0$ and $\hat{\beta}_1$ of the RSS and set them to equal 0 to obtain estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$. This yields

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

1.3 Interpreting SLR Estimate

With SLR, we are estimating mean responses to our predictor:

$$E(Y|X) = \beta_0 + \beta_1 X$$

- $\hat{\beta}_0$ is the mean response when the predictor is 0
- $\hat{\beta}_1$ is the change in mean response for a one unit increase in the value of the predictor

After SLR, $\hat{y} = \widehat{E(Y|X)}$ can be used to make predictions for specific x values

2 Multiple Linear Regression

Consider a model with multiple predictors

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Collect n sets of data:

$$(y_i, x_{i1}, \ldots, x_{ip})$$

2.1 Linear Regression in Matrix Form

2.1.1 SLR

Let

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Then we can rewrite the SLR equation as an equation of matrix operations:

$$Y = X\beta + \varepsilon$$

Each row of the above equation is equivalent to the ith algebraic form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$

for $i = 1, \ldots, n$

2.1.2 MLR

For multiple predictors, the only change is to β and X:

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

In the X matrix, each column represents a predictor, aside from the first column of 1s, which exists to facilitate matrix multiplication since β is a $(p+1) \times 1$ matrix.

2.1.3 Qualitative Predictor Variables

Suppose X_1 is qualitative with values "Yes", "No", "Maybe", while the rest are numeric. We can represent X_1 using indicator functions:

$$X_1 = \begin{cases} 1 & \text{if "Yes"} \\ 0 & o.w. \end{cases} \qquad X_2 = \begin{cases} 1 & \text{if "No"} \\ 0 & o.w. \end{cases}$$

Thus if both X_1 and X_2 are 0, then the entry is "Maybe". The matrix would look like:

$$X = \begin{bmatrix} 1 & 1 & 0 & \cdots \\ 1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \cdots \\ 1 & 0 & 0 & \cdots \end{bmatrix}$$

where the first row is "Yes", second is "Maybe", second last is "No", last is "Maybe".

2.1.4 Conditional Mean and Distribution of Responses

- SLR conditions on the value of 1 predictor
  - At each X value, we have a distribution of Y responses with mean $E(Y|X)$
- MLR conditions on the values of p predictors
  - At values (x_1, x_2) for (X_1, X_2), we have a distribution of responses Y with mean $E(Y|x_1, x_2)$

2.2 MLR LSE

We follow a similar procedure as in SLR. Given the MLR model

$$Y = X\beta + \varepsilon$$

we get a residual vector for our estimate $\hat{Y}$

$$\hat{e} = \begin{bmatrix} \hat{e}_1 \\ \vdots \\ \hat{e}_n \end{bmatrix}$$

so the RSS is

$$RSS = \hat{e}^T\hat{e} = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{1p}))^2$$

Theorem. Let $u = a^T x = x^T a$ where $a = (a_1, \ldots, a_p$ is a vector of constants. Then

$$\frac{\partial u}{\partial x} = \frac{\partial(a^T x)}{\partial x} = \frac{\partial(x^T a)}{\partial x} = a$$

Theorem. Let $u = x^T A x$ where A is a symmetric matric of constants. Thne

$$\frac{\partial u}{\partial x} = \frac{\partial(x^T A x)}{\partial x} = 2Ax$$

Since we want to find an expression for $\hat{\beta}$, we have

$$\begin{aligned} RSS &= \hat{e}^T\hat{e} \\ &= (Y - X\hat{\beta})^T(Y - X\hat{\beta}) \\ &= (Y^T - (X\hat{\beta})^T)(Y - X\hat{\beta}) \\ &= Y^TY - Y^TX\hat{\beta} - (X\hat{\beta})^TY + (X\hat{\beta}^T)(X\hat{\beta}) \\ &= Y^TY - Y^TX\hat{\beta} - Y^TX\hat{\beta} + \hat{\beta}^TX^TX\hat{\beta} \\ &= Y^TY - 2Y^TX\hat{\beta} + \hat{\beta}^TX^TX\hat{\beta} \end{aligned}$$

Thus

$$\frac{\partial RSS}{\partial \hat{\beta}} = -2X^TY + 2X^TX\hat{\beta} = 0 \implies \hat{\beta} = (X^TX)^{-1}X^TY$$

assuming X^TX is invertible. Thus our fitted values $\hat{Y}$ are

$$\hat{Y} = X(X^TX)^{-1}X^TY$$

Note that the matrix $X(X^TX)^{-1}X^T$ is called the **hat matrix**.

2.3 Interpreting MLR Estimate

The estimation and interpretation of coefficients in MLR conditions on all other predictors

- Consider only 1 fixed value of all the other predictors when interpreting the coefficient of interest

2.3.1 Interpreting the Coefficients in MLR

- $\hat{\beta}_0$ is the mean response when all the predictors are 0
- $\hat{\beta}_j$ is the mean change in response for a 1 unit increase in X_j when all other predictors are **held fixed**

2.3.2 Interpretation with Indicator Variables

Suppose we have a qualitative predictor:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 \mathbb{I}(A) + \hat{\beta}_4 \mathbb{I}(B)$$

This will give us **different intercepts but common slopes**(since indicator variables only take on 0 or 1 in value), thus our intercepts are either $\hat{\beta}_0$, $\hat{\beta}_0 + \hat{\beta}_3$, or $\hat{\beta}_0 + \hat{\beta}_4$. The interpretation of these coefficients must compare each level to the reference level (ie: when both $\mathbb{I}(A)$ and $\mathbb{I}(B)$ are 0)

2.3.3 Interpretation with Interaction Variables

An **interaction** term allows the relationship between response and one predictor vary according to the values of a second predictor

- Allows us to explore the joint effect of X_i and X_j on Y by multiplying them

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2$$

This will give us **(potentially) different intercepts and different slopes**. The potentially different intercepts depends on whether or not X_2 is categorical.

- Graphically, the difference between the two is a model with interaction term will showcase different slopes, a model without will showcase same slopes

3 Linear Regression Assumptions

In order to conduct linear regression, we must verify the following conditions hold:

1. Linearity of the relationship
    - Verifies that the relationship of the population is truly linear

    $$E(\varepsilon|X) = 0 \text{ or } E(Y|X) = X\beta \text{ or } Y = X\beta + \varepsilon$$

    - Ensures we estimate coefficients unbiasedly
2. Uncorrelated Errors

- Each data point in population must be uncorrelated with the others

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ or } \text{Cov}(y_i, y_j) = 0$$

- Ensures correct precision of estimates

3. Constant Error Variance

- Conditional on any value of X, the variance of errors is constant and the same for all X values

$$\text{Var}(\varepsilon|X) = \sigma^2 I \text{ or } \text{Var}(\varepsilon_i|X) = \text{Var}(y_i|X) = \sigma^2$$

- Ensures that reasonable estimates of variability for all conditional means are obtained

4. Normality

- All conditional distributions must have the same shape
- ε and Y's distributions must be the same shape for all values of X

$$\varepsilon|X \sim \mathcal{N}_n(0, \sigma^2 I) \text{ or } Y|X \sim \mathcal{N}_n(X\beta, \sigma^2 I) \text{ or } \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

  - Notice how regardless of value of X, the variance stays the same (constant variance assumption), while the expected value is consistent with those from the linearity assumption

- Allows to utilize properties of normal random variables for inferences (useful later)

3.1 Verifying Assumptions

The linear regression assumptions can be captured by

$$\varepsilon|X \sim \mathcal{N}_n(0, \sigma^2 I)$$

Since residuals are sample analogues of ε, they capture noise leftover after estimating a trend between Y and X

- Check unbiasedness:

$$\hat{e} = Y - \hat{Y} = X(\beta - \hat{\beta}) + \varepsilon \approx \varepsilon$$

- If the above does not hold, then we can conclude that a violation of assumption has occured

3.1.1 Verification through Residual Plots

We can also verify the assumptions through the following types of plots:

- Residual vs each predictor scatterplot
    - For linearity, uncorrelated errors and constant variance
- Residual vs response scatterplot
    - Also for linearity, uncorrelated errors and constant variance
- Normal Q-Q plots
    - For normality

For scatterplots, look for the following cues:

- Any systematic pattern, such as curves would indicate a violation of linearity
    - Note that vertical strips are allowed in residual predictor scatterplots since this represents the discrete behaviour of the data
- Large groups of points, such as clumps rather than than truly scattered points would indicate a violation of uncorrelated errors
- Any sort of fanning pattern, either increasing or decreasing, would indicate a violation of constant variance

For QQ plots, look to make sure that the line of scattered points mostly follows the normal line; if not then normality is violated.

3.1.2 Additional Conditions for MLR

If the relationship between predictors or predictor and response is too complex, then residual plots become unreliable since any relationship could appear as an assumption violation

- Given that in MLR there are multiple predictors, patterns in plots can't be used to identify specific violations and can give misleading information, although residual plots can always be used to verify that a valid model has been fit

Thus 2 extra conditions must be met in order to use residual plots in MLR:

1. Conditional mean response: the mean responses conditional on X taking on a value x_i are a single function of a linear combination involving the elements of beta

$$E(Y_i|X = x_i) = g(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

2. Conditional mean predictor: the relationship between any 2 predictors can only be at most linear:

$$E(X_i|X_j) = a_0 + a_1X_j$$

We check the first one using a scatterplot of response vs fitted values, the second using pairwise scatterplots of all predictors and observe the trends in all graphs (respectively, if randomly scattered or if not following a linear pattern, then we can't use residual plots)

3.2 Mitigating Violated Assumptions

3.2.1 Variance Stabilizing Transformations

- Specifically targets violated of constant variance
- **Transformation will only be applied to the response variable**

Let f be a function of the response, thus

$$\text{Var}[f(Y)] = [f'(E(Y))]^2\text{Var}(Y)$$

If $\text{Var}[f(Y)]$ is constant, then f is a **variance stabilizing transformation**.

- We say f stabilizes variance
- Since f makes variance constant, then the modified model with $f(Y)$ as the response satisfies the constant variance assumption
- If the data is right-skewed, consider a log transformation

3.3 Box-Cox Transformation

- Used to improve normality and linearity
- Is a power transformation
- Can be used on response, predictors, or both

The Box-Cox method uses MLE to estimate the power transformation. By the normality assumption, we know that

$$Y|X \sim \mathcal{N}(X\beta, \sigma^2 I)$$

thus our log-likelihood function is SLR is

$$\begin{aligned}\log(\mathcal{L}(\beta_0, \beta_1, \sigma^2|Y)) &= \log\left(\frac{1}{(\sigma\sqrt{2\pi})^n}\exp\left(-\sum_{i=1}^{n}\frac{(y_i - \beta_0 - \beta_1x_i)^2}{2\sigma^2}\right)\right) \\ &= -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{\sigma^2}\log\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1x_i)^2\right)\end{aligned}$$

$$= -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{\sigma^2}\log(RSS)$$

To perform a Box-Cox transformation on Y, modify the RSS so

$$RSS = \sum(\Psi_M(Y, \lambda) - \beta_0 - \beta_1 x_i)^2$$

where $\Psi_M(Y, \lambda)$ is a modified power transformation applied to Y. After finding the MLE of λ, we take Y^λ as the transformation

- The original Box-Cox is very complicated, so taking Y^λ allows for a more simple transformation that still gives a line of best fit that minimizes RSS and gets close to normality, effectively satisfying the normality assumption
- If λ is effectively 0, use $\ln(Y)$ instead

If modifying a predictor, do the same and take X_i^λ.
If performing a transformation simultaneously, define RSS in SLR as

$$RSS = \sum(\Psi_M(Y, \lambda_y) - \beta_0 - \beta_1 \Psi_S(X, \lambda_x))^2$$

The MLE will be an ordered pair (λ_y, λ_x), which we take Y^{λ_y} and X^{λ_x} as our transformations.
We can also pick even simpler λ values, for example, if $\lambda = 0.10102$ we can just take $\lambda = 0$ and use ln instead of taking the variable to the 0.010102 power. Pick λ values near easier values to work with, such as $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, -\frac{1}{2}, -1$, etc.

3.4 Impact of Violated Assumptions

For any estimate, we need a measure of error or variation in order to measure how different samples may vary in value.

- This requires the sampling distribution, whose properties are given by the assumptions like the mean, variance, or shape
- When the assumptions don't hold, then these properties no longer hold, which implies taking different samples could have a large impact

3.4.1 Sampling Distribution of the Estimated Coefficients

By the assumptions,

$$Y|X \sim \mathcal{N}(X\beta, \sigma^2 I)$$

Our estimator of β is

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

If the assumptions hold, we can necessarily find a sampling distribution of $\hat{\beta}$. Suppose they do. By linearity of Normal distributions,

$$AY \sim \mathcal{N}_n(A\mu_Y, A\Sigma A^T) \text{ or } \sum_i a_i y_i \sim \mathcal{N}\left(\sum_i a_i\mu_i, \sum_i a_i^2\sigma_i^2\right)$$

where A is a matrix of constants. Thus, the sampling distribution of $\hat{\beta}$ is

$$\mathcal{N}_n(\beta, \sigma^2(X^TX)^{-1})$$

Notice that this tells us that $\text{Bias}_\beta(\hat{\beta}) = 0$, which satisfies linearity, $\hat{\beta}$ are correlated, and $\hat{\beta}$ has the same constant σ^2 as the errors.
If there so happens to be violations, this distribution tells us that

- If linearity is violated, $\hat{\beta}$ is no longer an unbiased estimator
- If constant variance is violated, we no longer have a single σ^2 as part of the variance, thus we have an over-/under-estimation of error
- If uncorrelated errors is violated, we have an over-/under-estimation of variance
- If normality is violated, the estimator doesn't have a normal distribution

Properties:

- $E(\hat{\beta}|X) = \beta$
- $\text{Cov}(\hat{\beta}|X) = \sigma^2(X^TX)^{-1}$
- The standard error of $\hat{\beta}_j$ is the standard deviation of $\hat{\beta}_j = \sigma\sqrt{(X^TX)^{-1}_{(j+1,j+1)}}$ which is the square root of the $j+1$ element on the diagonal of the covariance matrix of $\hat{\beta}$

4 Inference on Regression Components

However, σ^2 is unknown, so we have to perform inference to estimate a value. Consider the estimator

$$s^2 = \frac{RSS}{n-p-1} = \frac{\hat{e}^T\hat{e}}{n-p-1}$$

which we can use in practice with

$$s^2(X^TX)^{-1}$$

as a covariance matrix. However, this adds sampling variation since s^2 is an estimated value and has its own sampling distribution satisfying

$$\frac{\hat{\beta}-\beta}{\sqrt{s^2(X^TX)^{-1}}} \sim T_{n-p-1}$$

where T_{n-p-1} is T distribution with $n-p-1$ degrees of freedom (n observations, $p+1$ estimated parameters).

4.1 Inference on Coefficients

With the estimator s^2, we've estimated the sampling distribution of $\hat{\beta}$. Consider the two forms of inferential processes:

- Confidence Intervals: estimate $-$ critical value $\times$ standard error, which represents a $(1-\alpha)\%$ chance that one of the $(1-\alpha)\%$ confidence intervals overlapped the truth
- Hypothesis Test: A test statistic $\frac{\text{estimate}-\text{truth}}{\text{standard error}}$ is built where we compare the estimated value to the proposed truth, and the more different the estimate is (i.e., the larger the test statistic), the more unlikely the proposed truth is true

Consider inference for individual coefficients β_j.
The $(1-\alpha)\%$ **confidence interval** for β_j is

$$\hat{\beta}_j \pm t_{\frac{\alpha}{2},n-p-1} s\sqrt{(X^TX)^{-1}_{(j+1,j+1)}}$$

where $t_{\frac{\alpha}{2},n-p-1}$ is the $\frac{\alpha}{2}$ quantile of the T distribution with $n-p-1$ degrees of freedom, $(X^TX)^{-1}_{(j+1,j+1)}$ is the $(j+1,j+1)$ element of $(X^TX)^{-1}$ (note how $s\sqrt{(X^TX)^{-1}_{(j+1,j+1)}}$ is approximately equal to the standard error of $\hat{\beta}_j$). This interval represents that $(1-\alpha)\%$ of all intervals computed using data repeatedly from the same population will contain the true value of β_j
For a hypothesis test on β_j, we consider a null hypothesis $H_0 : \beta_j = \beta_j^0$, where β_j^0 is the hypothesized true value (usually 0). Construct a test statistic

$$t^* = \frac{\hat{\beta}_j - \beta_j^0}{s\sqrt{(X^TX)^{-1}_{(j+1,j+1)}}}$$

Regardless of which alternate hypothesis we test, we keep the same test statistic. If

- $|t^*| > t_{\frac{\alpha}{2},n-p-1}$, reject the null
- If $P\left(|T_n - p - 1| \geq |t^*|\right) < \alpha$ (i.e.: the p value is lower than the significance level), reject the null

By rejecting the null hypothesis, we conclude that $\beta_j \neq 0$, so there must exist a linear relationship between Y and X_j.

4.2 Inference on Mean Response

Since we estimate $E(Y|X)$ using $\hat{Y} = X\hat{\beta}$, we can also construct confidence intervals for $\hat{Y}$ and perform hypothesis tests. Treat each mean response individually: for each $x_0^T = \begin{bmatrix} 1 & x_1 & \cdots & x_p \end{bmatrix}$ estimate $\hat{y}_0 = \hat{E}(Y|X = x_0^T) = x_0^T\hat{\beta}$. The sampling distribution of $\hat{y}_0$ is

$$\hat{y}_0|X, x_0 \sim \mathcal{N}(x_0^T\beta, \sigma^2 x_0^T(X^TX)^{-1}x_0)$$

As before, we estimate σ^2 using s^2, satisfying

$$\frac{\hat{y}_0 - x_0^T\beta}{\sqrt{s^2 x_0^T(X^TX)^{-1}x_0}} \sim T_{n-p-1}$$

$(1-\alpha)\%$ confidence interval for $y_0 = x_0^T\beta$ is

$$x_0^T\hat{\beta} \pm t_{\frac{\alpha}{2},n-p-1}s\sqrt{x_0^T(X^TX)^{-1}x_0}$$

For hypothesis test, we take $H_0 : y_0 = y_0^0$ and construct a test statistic

$$t^* = \frac{\hat{y}_0 - y_0^0}{s\sqrt{x_0^T(X^TX)^{-1}x_0}}$$

Reject the null hypothesis under the same conditions as for inference on $\hat{\beta}$.

4.3 Prediction Intervals

Since the regression model only predicts values for $E(Y|X = x_0)$, we want to make a prediction about the true y_0, which is likely not equivalent to $E(Y|X = x_0)$. We need to consider prediction error:

$$y_0 - \hat{y}_0 = (x_0^T\beta - \hat{y}_0) + \varepsilon_0$$

Consider the distribution of prediction error: by our assumptions

$$y_0 - \hat{y}_0|X, x_0 \sim \mathcal{N}(0, \sigma^2[1 + x_0^T(X^TX)^{-1}x_0])$$

Since σ^2 is unknown, estimate it with s^2.
Since we can't predict an actual value, instead consider a prediction interval: a range of possible values for an actual response.

$$x_0^T\hat{\beta} \pm t_{\frac{\alpha}{2},n-p-1}s\sqrt{1 + x_0^T(X^TX)^{-1}x_0}$$

The prediction interval is centered at $\hat{y}_0$ and is wider than the CIs for $E(Y|X = x_0)$

since the extra σ^2 term in the distribution

- Includes variation in estimating conditional mean and variation around conditional mean

The interpretation has the same idea: these are the most likely $(1-\alpha)\%$ values the random variable could take.

5 Decomposing Variance

5.1 Sum of Squares

By the assumptions, we know that

$$\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$$

Recall that fitting a linear model minimizes the RSS (aka the variation around the line). This means that if we see a plot with less variation around the trend, the linear trend is more prominent as the predictors are able to explain better the variation.

5.1.1 Sources of Variation in Regression

The variation observed in the response is quantified by the Total Sum of Squares:

$$SST = \sum (y_i - \bar{y})^2$$

Each predictor added to the model for y is able to explain a portion of this variation. Notice that SST is a function of the sample variance of y. As more predictors are added, more variation is able to be explained.
The overall variation *explained* by the predictors is the Regression Sum of Squares:

$$SS_{reg} = \sum (\hat{y}_i - \bar{y})^2$$

The leftover variation left unexplained after fitting the model is the Regression Sum of Squares:

$$RSS = \sum (y_i - \hat{y}_i)^2$$

Note that

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

By minimizing the RSS, we minimize the amount of variation unexplained by the predictors, so most of the variation makes sense and we can conclude a stronger relationship. By definition, each of the Sums of Squares has a certain number of degrees of freedom:

1. SST has $n-1$ degrees of freedom

2. SS_{reg} has p degrees of freedom

3. RSS has $n - p - 1$ degrees of freedom

5.2 ANOVA Test

Recall that in MLR, the T-test on the coefficients can only say if the predictor is linearly related to the response *in the presence of others*. The test is not able to give an overall notion on whether or not a linear relationship exists between the predictors and response in the first place.

A significant linear relationship means that a significant amount of variation is able to be explained by the model, so in theory we want SS_{reg} to be large relative to SST. However, given every dataset is different, then the definitions of SS_{reg} and SST do not allow us to simply observe the ratio of SS_{reg} to SST given each dataset will have a different SST. However, in decomposing SST, a model where SS_{reg}/SST is large necessarily requires RSS/SST to be small. Therefore, the test of overall significance checks if SS_{reg} is significantly larger than RSS.

We test

$$H_0 : \boldsymbol{\beta}_1 = 0 \qquad H_1 : \boldsymbol{\beta}_1 \neq 0$$

where we we split our vector of coefficients into

$$\beta^T = \begin{bmatrix} \beta_0 & \boldsymbol{\beta}_1 \end{bmatrix}$$

Under the assumptions of the null hypothesis, consider the test statistic

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

Dividing each SS quantity by its degrees of freedom standardizes the quantity and gives the Mean Squares Regression and Mean Squares Residual quantities respectively. Then following the procedure of the hypothesis test, if $F^* > F_{(1-\alpha),(p+1,n-p-1)}$, or in other words, the p value $< \alpha$, we reject H_0 at level α and conclude a statistically significant linear relationship exists for at least one predictor.

5.3 Partial F Test

Suppose a perform an ANOVA test and conclude that at least one predictor is related to Y, then after performing T tests on each predictor, we conclude that only a small subset of predictors of significantly related. The Partial F Test tests whether or not the simple model with only this subset is as good as the model with all the predictors (i.e.: tests if we can remove predictors or not). If the simple model is just as good, then the SS_{reg}

should be similar and equivalently, the RSS should be similar. If the smaller model has a much larger RSS, then we can't remove all insignificant predictors at once.
The Partial F Test compares 2 models:

1. Full model with p predictors
2. Reduced model with $p-k$ predictors

Each model uses the same data and response so the SST remains the same. However, by definition of RSS, a model with more predictors will *always* result in a smaller RSS, which would imply a larger SS_{reg} over the model with less predictors. As such, we adjust each quantity using its respective degrees of freedom.
We test

$$H_0 : \boldsymbol{\beta}_2 = 0 \qquad H_1 : \boldsymbol{\beta}_2 \neq 0$$

where

$$\beta^T = \begin{bmatrix} \beta_0 & \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 \end{bmatrix}$$

where $\boldsymbol{\beta}_2$ is a vector of k coefficients of the removed predictors, $\boldsymbol{\beta}_1$ is a vector of $p-k$ coefficients of the predictors we keep. Under the assumptions of the null hypothesis, consider the test statistic

$$F^* = \frac{\frac{RSS_{drop}}{k}}{\frac{RSS_{full}}{n-p-1}} \sim F(k, n-p-1)$$

In hypothesis fashion, if $F^* > F_{(1-\alpha),(k,n-p-1)}$, then we reject the null hypothesis and conclude that the additional SS_{reg} from the k removed predictors explains a lot of variation, so we want to keep these predictors. However, if the null hypothesis were to be accepted, then we conclude there does not exist a significant linear relationship between Y and any of the k predictors.
In general, the order of tests in an analysis is

$$\text{ANOVA test} \longrightarrow \text{T test on each individual coefficient} \longrightarrow \text{Partial F Test}$$

It's also worth noting that for a partial F test, we need to verify assumptions of both the full and reduced models.

5.4 Goodness of a Model

Since each dataset has a different SST, then working on different responses means a larger SS_{reg} could be a result of an overall larger SST. So, we standardize the variation explained by the SST so its no longer dependent on the starting variation. Define the **Coefficient of Determination**

$$R^2 = \frac{SS_{reg}}{SST} = 1 - \frac{RSS}{SST}$$

R^2 represents the proportion of variation in the response explained by the model. Remember that comparing two models with a different number of predictors will always result in the larger model having a larger SS_{reg}, so a larger model will also always have a larger R^2 even if the extra predictors are not significant. So, we consider the adjusted R^2 by adjusting using the degrees of freedom:

$$R^2_{adj} = 1 - \frac{\frac{RSS}{n-p-1}}{\frac{SST}{n-1}}$$

Note that this adjusted quantity does not have the interpretation of the proportion of variance that is explained by the model; it only says that a bigger model is better only if the SS_{reg} has increased enough to compensate for adding complexity.

5.5 Problems With Related Predictors

Recall that the conditional mean predictor condition specifies that the relationship between each predictor is at most linear. Ideally, we want no relationship between the predictors; a linear trend means the predictors are correlated (collinear).

5.5.1 Rank of Design Matrix and Correlation

If we have perfect correlation (i.e.: 1 or -1), then this means one predictor is perfectly related to another. This is an issue since we would have a column of the X matrix being a linear combination of others, thus X^TX is no longer invertible.

5.5.2 Multicollinearity

Multicollinearity is when more than 2 predictors are related. This is an issue since the model won't be able to distinguish how much variation is due to only X_1 vs only to X_2. Some issues arising from multicollinearity are

- Wrong estimated coefficients: coefficients might have the wrong sign compared to literature
- Contradictory significance: many predictors might be insignificant when the overall F test is highly significant
- Inflated variances: standard errors of estimated coefficients are much larger than they should be

To check if multicollinearity is present, we can look at the correlation matrix or conditional mean predictor pairwise scatterplots, although the issue with these is that we can only compare 2 predictors at a time and we can't consider the conditionality of the predictors

on one another nor with the response. Thus, we used a measure called the **variance inflation factor (VIF)**. Notice how for any model, we have

$$\text{Var}(\hat{\beta}_j) = \frac{1}{1-R_j^2} \times \frac{\sigma^2}{(n-1)s_{x_j}^2}$$

for all $j = 1, \ldots, p$. The VIF is the term

$$\frac{1}{1-R_j^2}$$

where R_j^2 is the R^2 from a model which includes X_j as a response. Notice how if R_j^2 grows towards 1, then the VIF increases in value, which indicates inflated variance. Since R_j^2 gives how much variation is is explained by the model including X_j as a response, then as this variation increases, it tells us that the relationship between X_j and other predictors is higher, thus the VIF increases. Generally, we use a cutoff of 5.
The are 2 main ways to deal with multicollinearity:

1. Collect more data
2. Respecify the model

6 Problematic Observations

6.1 Leverage Observations

Definition (Leverage observation)**.** An observation that is very distant from the center of the X space that may change $\hat{Y}$.

These points have the potential to shift the regression line but won't always do it. To find leverage points, we consider the hat matrix.
Notice that since $\hat{Y} = X\hat{\beta}$ and $\hat{\beta} = (X^TX)^{-1}X^TY$, then $\hat{Y} = HY$, so

$$\hat{y}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j$$

The $h_{ii}y_i$ term represents the effect that y_i has on its own fitted value. Thus, we call the diagonal elements h_{ii} the **leverage** of observation i. These values give a metric on the impact the value of y_i has on $\hat{y}_i$. Since H is idempotent and $\text{tr}(H) = p + 1$, then $0 \leq h_{ii} \leq 1$, so the fraction of $\hat{y}_i$ due to to y_i vs the other responses is given by h_{ii}. If h_{ii} is close to 1, then this implies the other h_{ij} are all 0, thus $\hat{y}_i \approx y_i$. This means that a different line may have been estimated if this observation wasn't used. Notice we use the interpretation of *may have* given leverage points only have the *potential* to shift the line. If $h_{ii} > \frac{2(p+1)}{n}$, then this point is a leverage point.

6.1.1 Outliers in Regression

A regressional outlier is a point that is very far from the trend/conditional mean. To determine these points, we consider the residuals. By definition,

$$\hat{e} = Y - \hat{Y} = Y - HY = (I - H)Y$$

thus we have

$$\hat{e}_i = (1 - h_{ii})y_i - \sum_{j \neq i} h_{ij} y_j$$

Since $0 \leq h_{ii} \leq 1$, then the impact the response y_i has on $\hat{e}_i$ depends on its distance in the X space h_{ii}, so the higher the leverage, the lower the weight of y_i on $\hat{e}_i$. Note that the residuals do not have constant variance:

$$\text{Cov}(\hat{e} \mid X) = \text{Cov}((I - H)Y \mid X) = \sigma^2(I - H)$$

since the $I - H$ matrix contains different values depending on the sample taken. Further, this shows that

$$\text{Var}(\hat{e}_i \mid X) = \sigma^2(1 - h_{ii})$$

To achieve constant variance, standardize each $\hat{e}_i$ by its variance

$$r_i = \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}}$$

If this value if outside of the interval $[-2, 2]$, then the point is a regressional outlier.

6.2 Influential Observations

Influential observations can influence the model estimation in 3 ways:

1. Affect all fitted values
2. Affect its own fitted value
3. Affect how at least one coefficient is estimated

We observe this through delete-one measures: deleting a single observation and fitting new models.

6.2.1 Influential on all Fitted Values

We consider a measure called Cook's Distance. First, fit a model with all n observations. Then, refit the model using $n - 1$ observations. The difference in estimated trend of the two models tells us the influence of the deleted observation on all fitted values. Instead

of fitting n different delete-one models, use

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

The above quantity incorporates effect due potentially to being distant from the X space or being far from the estimated trend.

6.2.2 Influential on own Fitted Value

Here we use a different measure called DFFITS. First, fit a model with all n observations, then refit using $n-1$ observations. The change in $\hat{y}_i$ values, accounted for variation, is how influential the point is on its own fitted value. Again, rather than fitting n models, we use

$$DFFITS_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}} \frac{\hat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

where $s^2_{(i)}$ is the sample variance from the model omitting observation i.

6.2.3 Influential on Estimated Coefficients

Here we use yet another measure called DFBETAS. We observe how each individual coefficient changes with and without each observation.

$$DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{s^2_{(i)}(X^TX)^{-1}_{j+1,j+1}}}$$

where $\hat{\beta}_{j(i)}$ is the coefficient of x_j from the model without point i.

6.3 Addressing Problematic Observations

Each measure introduced quantifies the extent of each potential issue, so we define cutoff values for each metric to know when the amount of leverage, outlying-ness, and influence is substantial.

Type of Point	Measure	Cutoff
Leverage	h_{ii}	$h_{ii} > 2\left({}^{(p+1)}/_{n}\right)$
Outlier	r_i	$r_i \notin [-2,2]$ if dataset "small" (e.g. $n < 50$) $r_i \notin [-4,4]$ if dataset "large" (e.g. $n \geq 50$)
Influence	D_i	$D_i >$ median of $F(p+1, n-p-1)$
	$DFFITS_i$	$\lvert DFFITS_i \rvert > 2\sqrt{{}^{(p+1)}/_{n}}$
	$DFBETAS_{j(i)}$	$\lvert DFBETAS_{j(i)} \rvert > {}^{2}/_{\sqrt{n}}$

It is important to note that unless there is a contextual reason, do not remove problematic observations and simply note them and their impact as a limitation of the model.

7 Model Selection

For model selection, we need to consider the purpose of the model: either prediction or description. If we're predicting using the model, then extra predictors will help explain more variation and have a better accuracy if not over-fitted. If the purpose is description, then too many predictors will hurt interpretability. Thus, we need to use model selection to select the best model for our purposes.

7.1 Numerical Measures of Goodness

We introduce 2 measures based on the log-likelihood function of σ^2. Because it's not as simple as a model is better if it has more predictors, we introduce penalty terms in order to control for added X's.

Akaike's Information Criteria (AIC)

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

where the penalty term is $2p$.

Bayesian Information Criteria (BIC)

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

where the penalty term is $(p+2)\ln(n)$. Note that the penalty term for BIC is smaller than that of AIC because the BIC includes sample size. As such, the BIC is more likely to favour simpler models.

- For both the AIC and BIC, a smaller number indicates a better model

7.2 All Possible Subsets Model Selection

There are 2 steps to performing this:

1. Compare models of each size using adjusted R^2
    - Choose model with highest R^2_{adj} in each group
2. Use R^2_{adj}, AIC, AIC corrected, BIC to pick the best of the best

There are pros and cons of this method. The pros include that all possible models are fit and compared, and there's a certain level of flexibility in our definition of best. However,

with a large number of predictors, this can be very impractical. Furthermore, the best of each subset does not consider model issues, so we can perform this without assumptions and the method does not account for multicollinearty or problematic observations. This means the result may not be reliable.

7.3 Automated Selection Methods

These methods use the AIC or BIC instead of adjusted R^2 to decide between models. There are 3 different types:

1. Forward selection: start with intercept model and add predictors
2. Backward selection: start at full model and remove predictors
3. Stepwise selection: iterate between forward and backward

At each step, the model from the previous step is taken and each possible predictor available is added or deleted. We compute the AIC or BIC for each and the smallest value is chosen. This chosen model is then the starting model for the next step, and once no smaller AIC or BIC is produced, we take the last chosen model as the final. In R, we do selection using the stepAIC function, specifying $k = 2$ for AIC, $k = log(n)$ for BIC. We also need to specify direction, either "forward", "backward", or "both".
Again, there are pros and cons to this. The pros are that its less intensive given we're not considering all possible subsets. It also gives an idea of the preferred model, though this may not be the best one. Stepwise also considers the conditional nature of regression. However, the cons are that all methods may not agree on the same preferred model. Additionally, this still runs in the presence of model violations or other issues, and also do not consider context of data (i.e.: indiscriminately removes predictors). While automation methods save time when p is large, the risk of not getting the best model exists since not every subset is considered, thus automation only gives an *idea* of the best model.