Linear Regression

Matthew Chen

September 8, 2025

Contents

In linear regression, we explore some linear relationship between a continuous response variable and one or more dependent covariates. As a foundational model, linear regression is the gateway to statistical modeling and provides a versatile method to study the strength of relationships and to make predictions. In these notes, we discuss estimation and inference of linear models, how to understand and decompose sources of variance, and how to address common problems.

1 Ordinary Least Squares

1.1 Least Squares as a Gaussian MLE

Consider the following statistical model where the response variable Y is modelled as a function of a fixed feature variable X with parameters θ, and an additive random error term.

$$Y = f(X;\theta) + \epsilon$$

Further, if we make the assumption that the additive errors are independent Gaussian, with zero mean and constant variance σ^2, we have that

$$Y|X \sim \mathcal{N}(f(X;\theta), \sigma^2)$$

We are interested in estimating θ using n observed data pairs $\{x_i, y_i\}_{i=1}^n$ via maximum likelihood estimation (MLE). Let $\phi(f(x_i), \sigma^2)$ denote the normal probability density function with mean $f(x_i)$ and variance σ^2. Writing out the joint likelihood of the data, assuming independence between observations, we have

$$\begin{aligned} L\left(y_1, ..., y_n | x_1, ..., x_n; \theta\right) &= \prod_{i=1}^{n} \phi(y_i | f(x_i;\theta), \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(y_i - f(x_i;\theta))^2\right) \end{aligned}$$

Taking the log of the expression, we arrive at the log-likelihood

$$\ell(\theta) = \sum_{i=1}^{n} log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(y_i - f(x_i;\theta))^2$$

Writing out the maximum likelihood estimator we have

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(y_i - f(x_i;\theta))^2$$

Removing terms that do not depend on θ

$$\begin{aligned} \hat{\theta}_{MLE} &= \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} -(y_i - f(x_i;\theta))^2 \\ &= \underset{\theta}{\text{argmin}} \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2 \end{aligned}$$

We arrive at the conclusion that minimizing square error deviations will provide the MLE estimate for the model given Gaussian assumptions.

1.2 Simple Linear Regression

In Section 1.1 we demonstrated the principle of least squares for an arbitrary function $f(X)$ and its relationship with the Gaussian distribution. Here, we consider the simple case where $X \in \mathbb{R}$ (only one predictor variable) and $f(X)$ is a linear function of X with coefficient β_1 and an intercept term β_0. That is,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where the subscript i represents the ith case. Here, ϵ_i and subsequently Y_i are considered random variables while X is considered as a fixed non-random variable. We make the key assumptions that the error terms have zero mean $\mathbb{E}\epsilon_i = 0 \forall i$, constant variance $Var(\epsilon) = \sigma^2 \forall i$, and are uncorrelated with each other $cov(\epsilon_i, \epsilon_j) \forall i \neq j$. We do not make any distributional assumptions at this point.

Fitting the model based on the least squares criterion Q using observed data pairs $\{x_i, y_i\}_{i=1}^n$

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

by setting $\partial Q/\partial \beta_0$ and $\partial Q/\partial \beta_1$ to zero we have the solution

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

where S_x and S_y denotes the (unbiased) sample standard deviation of x and y, respectively, and r_{xy} denotes the correlation coefficient between x and y.

We define the *fitted values* $\hat{y}_i$, to be the values along the fitted regression line

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

and the *residuals* e_i to be the differences between the observed y_i and their respective fitted values.

$$e_i = y_i - \hat{y}_i$$

From the equation for the fitted values, note that the fitted regression line passes through the center of the data $(\bar{x}, \bar{y})$.

1.2.1 Properties of Residuals

Note that the following properties are true for the residuals,

1. $\sum_{i=1}^n e_i = 0$
2. $\sum_{i=1}^n x_i e_i = 0$
3. $\sum_{i=1}^n \hat{y}_i e_i = 0$

where the third property can be derived using the first two.

Example. Show that the above 3 properties are true.

To show property 1, we only need to evaluate for $\hat{\beta}_0$ and simplify

$$\begin{aligned}
\sum_{i=1}^n e_i &= \sum_{i=1}^n y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= \sum_{i=1}^n y_i - (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i) \\
&= n\bar{y} - n\bar{y} + \hat{\beta}_1 n\bar{x} - \hat{\beta}_1 n\bar{x} = 0
\end{aligned}$$

Note that it can be shown algebraically that

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

and

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}$$

Then we have that

$$\begin{aligned}
\sum_{i=1}^{n} x_i e_i &= \sum_{i=1}^{n} x_i \left(y_i - (\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i)\right) \\
&= \sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y} - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})x_i \\
&= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sum_{i=1}^{n}(x_i - \bar{x})^2 = 0
\end{aligned}$$

And finally

$$\begin{aligned}
\sum_{i=1}^{n} \hat{y}_i e_i &= \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i)e_i \\
&= \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i e_i = 0
\end{aligned}$$

1.2.2 Estimation of Error Variance

Intuitively, we can estimate the parameter σ^2 using some measure of sample variance of the residuals. We define the *Error Sum of Squares* (SSE) as the following

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Note that SSE has $n - 2$ degrees of freedom, since there are n data observations and 2 constraints on e_i as discussed above.

We also define the *Mean Square Error* (MSE) as

$$MSE = \frac{SSE}{n-2}$$

It can be shown that (see Section 2.3)

$$\mathbb{E}SSE = (n-2)\sigma^2$$

that is, MSE is an unbiased estimator for σ^2.

1.2.3 Properties of the Least-Squares Estimators

First, note that the least squares estimators are unbiased, that is $\mathbb{E}\hat{\beta}_0 = \beta_0$ and $\mathbb{E}\hat{\beta}_1 = \beta_1$.

Example. Show that the least squares estimators are unbiased in simple regression.
Let $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$. Recall that

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{S_{xx}} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon_i)}{S_{xx}} \\
&= \frac{\beta_0}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) + \frac{\beta_1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) x_i + \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) \epsilon_i \\
&= 0 + \frac{\beta_1 S_{xx}}{S_{xx}} + \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) \epsilon_i
\end{aligned}$$

Finally, taking the expectation, we have

$$\mathbb{E}\hat{\beta}_1 = \beta_1 + \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}\epsilon_i = \beta_1 + 0 = \beta_1$$

Recall that

$$\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\
\mathbb{E}Y_i &= \beta_0 + \beta_1 X_i
\end{aligned}$$

so we also have that

$$\begin{aligned}
\mathbb{E}\bar{y} &= \beta_0 + \beta_1 \bar{x} \\
\mathbb{E}\hat{\beta}_0 &= \mathbb{E}(\bar{y} - \bar{x}\hat{\beta}_1) = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0
\end{aligned}$$

Further, we have that the population variances of the least squares estimators are the following

$$\begin{aligned}
Var(\hat{\beta}_0) &:= \sigma^2\{\hat{\beta}_0\} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
Var(\hat{\beta}_1) &:= \sigma^2\{\hat{\beta}_1\} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}
\end{aligned}$$

Since σ^2 is unknown, we evaluate for *standard error* by replacing σ^2 with its unbiased estimator, MSE.

$$\begin{aligned}
s\{\hat{\beta}_0\} &= \sqrt{MSE \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)} \\
s\{\hat{\beta}_1\} &= \sqrt{\frac{MSE}{\sum_{i=1}^n (x_i - \bar{x})^2}}
\end{aligned}$$

Note how the variance decrease when $\sum_{i=1}^n (x_i - \bar{x})^2$ increases. That is, variance decreases when either the sample size is larger or if the spread of the data in x is greater.

1.3 Normal Error Model

Up until this point, we are unable to make statistical inferences on our fitted regression line since we have no assumption of distribution. However, we know that by minimizing the least squares, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the MLE estimators for β_0 and β_1 under the normal distribution. In this section, we discuss sampling distributions under *normal error model*.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

Equivalently, we have that $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$ and are independent with each other.

Examining the solution for $\hat{\beta}_0$ and $\hat{\beta}_1$, we see that they are linear functions of Y_i, that is, both least squares estimators are also normally distributed. We have that

$$\hat{\beta}_0 \sim \mathcal{N}(\beta_0, \sigma^2\{\hat{\beta}_0\}), \hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma^2\{\hat{\beta}_1\})$$

It can be shown that

$$\frac{\hat{\beta}_1 - \beta_1}{s\{\hat{\beta}_1\}} \sim t_{(n-2)}$$

is a pivotal quantity that follows the t-distribution with $n-2$ degrees of freedom. Intuitively, we know that

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \mathcal{N}(0, 1)$$

Further, SSE is a sum of squares of zero mean normal random variables, which can be shown to follow $\sim \sigma^2 \chi^2_{(n-2)}$. Then,

$$SSE/\sigma^2 \sim \chi^2_{(n-2)}$$

and

$$\sqrt{SSE/\sigma^2(n-2)} = \sqrt{MSE}/\sigma \sim \sqrt{\chi^2_{(n-2)}/(n-2)}$$

so we have that

$$\frac{Z}{\sqrt{MSE}/\sigma} = \frac{\hat{\beta}_1 - \beta_1}{s\{\hat{\beta}_1\}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{(n-2)}/(n-2)}} \sim t_{(n-2)}$$

Rearranging the pivotal quantity, we can create exact α level confidence intervals for the regression coefficients, i.e.

$$\mathbb{P}\left(\hat{\beta}_1 - t(1-\alpha/2; n-2) * s\{\hat{\beta}_1\} < \beta_1 < \hat{\beta}_1 + t(1-\alpha/2; n-2) * s\{\hat{\beta}_1\}\right) = 1 - \alpha$$

or similarly, test the significance of a regression coefficient where the true population coefficient β_1 is defined under the null hypothesis. For example, we can test the hypothesis

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

using the statistic

$$T^* = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{s\{\hat{\beta}_1\}} = \frac{\hat{\beta}_1 - 0}{s\{\hat{\beta}_1\}}$$

rejecting the null hypothesis at the α level if

$$p = \mathbb{P}\left(|t_{(n-2)}| > |T^*|\right) < \alpha$$

1.4 Estimation and Prediction

1.4.1 Estimating Mean Response

Suppose that we are interested in the response variable at $X = X_h$ after fitting the regression line. Specifically, we want to estimate the quantity $\mathbb{E}Y_h$ using fitted value $\hat{Y}_h$ as an unbiased estimator.

Example. Find the variance of the estimator $\hat{Y}_h$.
We have that $\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = \bar{y} + \hat{\beta}_1(X_h - \bar{x})$. Then

$$\begin{aligned} Var(\hat{Y}_h) &= Var(\bar{y}) + Var(\hat{\beta}_1(X_h - \bar{x})) + 2Cov(\bar{y}, \hat{\beta}_1(X_h - \bar{x})) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) + (X_h - \bar{x})^2 Var(\hat{\beta}_1) + 2Cov(\bar{y}, \hat{\beta}_1(X_h - \bar{x})) \\ &= \frac{\sigma^2}{n} + (X_h - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + 2Cov(\bar{y}, \hat{\beta}_1(X_h - \bar{x})) \end{aligned}$$

Solving for $Cov(\bar{y}, \hat{\beta}_1(X_h - \bar{x}))$. Let $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

$$\begin{aligned} Cov(\bar{y}, \hat{\beta}_1(X_h - \bar{x})) &= Cov\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \frac{X_h - \bar{x}}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})y_i\right) \\ &= \frac{X_h - \bar{x}}{nS_{xx}} Cov\left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n}(x_i - \bar{x})y_i\right) \\ &= \frac{X_h - \bar{x}}{nS_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})Var(y_i) \ (since\ Cov(y_i, y_j) = 0\ \forall i \neq j) \\ &= \frac{(X_h - \bar{x})\sigma^2}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \end{aligned}$$

Then finally we have that

$$Var(\hat{Y}_h) = \frac{\sigma^2}{n} + (X_h - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

As before, we can use MSE to evaluate for the standard error

$$s\{\hat{Y}_h\} = \sqrt{MSE\left(\frac{1}{n} + \frac{(X_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

We now have all the tools to conduct statistical inference on the estimation of mean response. Under the normal error model, we have

$$\hat{Y}_h \sim \mathcal{N}(\mathbb{E}Y_h, Var(\hat{Y}_h))$$

so that

$$\frac{\hat{Y}_h - \mathbb{E}Y_h}{s\{\hat{Y}_h\}} \sim t_{(n-2)}$$

1.4.2 Predicting an Outcome

The difference between estimation and prediction is that we predict random variables while we estimate fixed quantities. such as $\mathbb{E}Y_h$. Predicting Y_h has two sources of variability, first, sampling variability in $\hat{Y}_h$, and second, the natural variability in the target, Y_h itself.

Under the normal error model, we have that the quantity $\hat{Y}_h - Y_h$ is normally distributed, and it can be shown that

$$\frac{\hat{Y}_h - Y_h}{s\{pred_h\}} \sim t_{(n-2)}$$

where the standard error of the prediction is

$$s\{pred_h\} = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(X_h - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)}$$

Then we can create an α level prediction interval for Y_h

$$\hat{Y}_h \pm t(1 - \alpha/2; n-2) * s\{pred_h\}$$

where $t(.; n-2)$ is the inverse cdf for t_{n-2}. Note that $s\{pred_h\} > s\{\hat{Y}_h\}$ due to the increased variability of the prediction task. See Section 2.4.3.

1.5 Analysis of Variance (ANOVA)

We may be interested in analyzing the proportion of variance in Y that our predictor variable X can explain. To do so, we can attribute the variance in the observations to either the variation in the error term, or variation that can be linearly explained by X.

Example. Show the decomposition of variance into the sum of squares

$$\begin{aligned}
\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n \left((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right)^2 \\
&= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^n e_i(\hat{y}_i - \bar{y}) \\
&= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^n e_i\hat{y}_i - 2\bar{y}\sum_{i=1}^n e_i \\
&= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2
\end{aligned}$$

We define the total variation as the *total sum of squares* (SSTO). The first term, which represents variation attributable to the error term, is the *error sum of squares* (SSE). The second term, which represents variation attributable linearly to X, is the *regression sum of squares* (SSR). We just showed that

$$SSTO = SSE + SSR$$

Example. Show that SSR in simple regression is $\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$

$$\begin{aligned}
SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^n (\bar{y} + \hat{\beta}_1(x_i - \bar{x}) - \bar{y})^2 \\
&= \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2
\end{aligned}$$

SSTO has $n-1$ degrees of freedom, coming from n different observations and 1 constraint, $\sum_{i=1}^n y_i - \bar{y} = 0$. SSE has $n-2$ degrees of freedom, coming from n different observations and 2 constraints on the residuals, $\sum_{i=1}^n e_i = 0$ and $\sum_{i=1}^n x_i e_i = 0$ as discussed in Section 1.2.1. Finally, SSR has only 1 degree of freedom in

simple regression, since there is only one predictor variable $\hat{\beta}_1$.

By normalizing the sum of squares by their respective degrees of freedom, we define the mean squares (MS). Specifically we have *mean square error* $MSE = SSE/n-2$ and *regression mean square* $MSR = SSR/1$.

1.5.1 ANOVA F-test

Suppose we want to test for the significance of the regression relationship by using the variance explained by the predictor variables. For example, we can test the hypothesis

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

using the statistic

$$F^* = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)} \sim F_{1,n-2}$$

Under the null distribution, it can be shown that $SSR/\sigma^2 \sim \chi^2_1$ and we know that $SSE/\sigma^2 \sim \chi^2_{n-2}$. Crucially, SSE and SSR are independent (see Section 2.3) since they are functions of uncorrelated normal random variables. Then the null distribution follows from the fact that the ratio of two independent chi-squared random variables divided by their respective degrees of freedom follows an F-distribution. That is, we can reject the null hypothesis at the α level if $F^* > F(1-\alpha; 1; n-2)$, where $F(.; 1, n-2)$ describes the inverse CDF function for the $F_{1,n-2}$ distribution.. Note that in the case of simple linear regression, the F-test is equivalent to the two-sided t-test for $\beta_1 = 0$.

The results of the ANOVA can be represented in an ANOVA table, as shown in Table 1.

Table 1: ANOVA table for simple regression

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F^*
Regression	$SSR = \sum_{i-1}^{n}(\hat{y}_i - \bar{y})^2$	1	$MSR = SSR/1$	MSR/MSE
Error	$SSE = \sum_{i-1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	$MSE = SSE/(n-2)$	
Total	$SSTO = \sum_{i-1}^{n}(y_i - \bar{y})^2$	$n-1$		

We can also define the *coefficient of determination*, R^2, as the proportion of variation in Y explained by variation in X.

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

R^2 is between 0 and 1 for linear regression where X accounts for all variation when $R^2 = 1$ and there is no evidence of linear association (horizontal regression line) when $R^2 = 0$.

1.6 Assumptions and Remedial Measures

So far, we have made the following key assumptions for our model

1. Linearity of the regression relationship
2. Normality of the error terms
3. Constant variance of the error terms
4. Independence of the error terms

Violation of the linearity assumption, as well as the omission of important predictors may lead to serious consequences. On the other hand, non-constant variance (heteroscedasticity) and non-independence may lead to invalid variance estimation or inference, but are generally less serious. Small departures from normality are not serious but severe deviations may adversely affect inference results. Finally, outliers could be serious for small datasets.

To fix a nonlinear regression relationship, one could consider a nonlinear transformation in Y or X. For example, we can regress $\sqrt{X}$, $log(X)$, or $1/X$ instead of the original X. Similarly, the error distribution may be fixed by transformations on Y. And finally, outliers may be removed for small datasets.

Plotting the residuals against the fitted values are a practical way to check regression assumptions. For example, if there are nonlinear terms they will be mistakenly included in the residuals, which will appear as a systematic effect in the residual plot. If there is evidence of heteroscedasticity, the spread of the residuals will be unequal in the plot. Similarly, normality assumptions can be practically checked using a normal quantile-quantile (QQ) plot.

1.6.1 Box-Cox Procedure

The Box-Cox procedure automatically chooses a power transformation in Y by choosing λ that maximizes the loglikelihood (minimizes SSE). The transformation is done as the following

$$y_i^* = \begin{cases} K_1 \frac{y_i^\lambda - 1}{\lambda}, \lambda \neq 0 \\ K_2 log(y_i), \lambda = 0 \end{cases}$$

where $K_2 = \left(\prod_{j=1}^n y_j\right)^{\frac{1}{n}}$ and $K_1 = 1/K_2^{\lambda-1}$.

2 Multiple Regression

2.1 Types of Multiple Regression

In multiple regression, we discuss linear regression with more than one predictor X variables. That is, we model

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{ip-1} + \epsilon_i$$

where there are p coefficients and $p-1$ X variables. As before, we assume

$$\mathbb{E}\epsilon_i = 0 \forall i, Var(\epsilon_i) = \sigma^2 \forall i, Cov(\epsilon_i, \epsilon_j) = 0 \forall i \neq j$$

so that the mean response is

$$\mathbb{E}Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{ip-1}$$

Below are some examples of different variants of multiple regression.

2.1.1 First Order Models

First order models only consider the additive effects of distinct predictor variables. That is, we regress Y against $X_1, ..., X_{p-1}$ with no consideration for interactions. In this model, β_k represents the change in mean response $\mathbb{E}Y$ per unit increase of X_k, with the other predictor variables held constant.

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + \epsilon$$

2.1.2 Models with Interactions

In this model, we additionally capture interactive effects, that is, where the effect of one predictor depends on values from other predictors. For example, when regressing Y against X_1 and X_2, we can capture the second order interaction as follows

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

2.1.3 Polynomial Regression

In this model, we capture higher order powers of different predictor variables. For example,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \epsilon$$

2.1.4 Model with Transformed Variables

We can also model other nonlinear transformations of our original predictors. As long as the mean response is linear in its coefficients, we can continue to use the linear regression model. For example, we can model

$$log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

2.2 Matrix Form for Multiple Regression

We can write the multiple regression model in matrix form as follows

$$Y = X\beta + \epsilon$$

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n,p}$, $\beta \in \mathbb{R}^p$, and $\epsilon \in \mathbb{R}^n$. X, known as the *design matrix* is a matrix where the n rows represent each data observation, and each column represents a different predictor variable. By convention, we include a column of ones as the first column of X, lining up with the intercept coefficient β_0.

As usual, we assume that

$$\mathbb{E}\epsilon = \vec{0}, \mathbb{E}Y = X\beta$$

with covariance matrices

$$Var(\epsilon) = \sigma^2 I_n, Var(Y) = \sigma^2 I_n$$

Further, under the normal error model, ϵ (or Y) is a multivariate normal random vector.

2.2.1 The Least Squares Estimators

As before, we solve for $\hat{\beta}$ as to minimize the least squares criterion

$$\hat{\beta} = \underset{\beta}{argmin} ||Y - X\beta||_2^2$$

Example. Find a closed-form solution for $\hat{\beta}$
Let the least squares criterion be

$$Q(\beta) = ||Y - X\beta||_2^2 = (Y - X\beta)^T (Y - X\beta)$$

Our goal is to solve for $\nabla_\beta Q(\beta) = 0$. When a multivariate function is composite of an affine function, i.e. $g(x) = f(Ax + b)$, we can use the Jacobian chain rule arrive at

$$\nabla_x g(x) = A^T \nabla_{Ax+b} f(Ax + b)$$

Applying this to the least squares criterion, we have that

$$\nabla_\beta Q(\beta) = -X^T \nabla_{Y - X\beta} ||Y - X\beta||_2^2 = -2X^T (Y - X\beta) = 0$$

so we have that

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

so finally we have

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

Note that for a solution to exist, $X^T X \in \mathbb{R}^{p,p}$ must be invertible. That is, it must be full rank, $Rank(X^T X) = p$ which requires $Rank(X) = p$. This requires that $n \geq p$ (if $n < p$, $Rank(X) \leq n < p$ which would make

X^TX rank deficient and invertible).

Example. Show that $\hat{\beta}$ is an unbiased estimator for β.

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

Example. Find the covariance matrix for $\hat{\beta}$.

For a linear transformation of a random vector, $Y = AX + b$, we have that $Cov(Y) = AVar(X)A^T$. Here, we note that $\hat{\beta}$ is a linear function of Y. Then we have that

$$\begin{aligned} Var(\hat{\beta}) &= (X^TX)^{-1}X^TVar(Y)X(X^TX)^{-1} \ (X^TX, (X^TX)^{-1} \textit{ are square and symmetric}) \\ &= (X^TX)^{-1}X^T(\sigma^2 I_n)X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1} \end{aligned}$$

2.2.2 The Hat Matrix

We define the *hat matrix* H as the following

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

Note that H is a projection matrix to the column space of X. This implies that H is idempotent (i.e. $HH = H$) and symmetric (i.e. $H^T = H$).

Example. Show that H orthogonally projects to the column space of X

First, we want to show that for any vector w, $Hw \in Col(X)$. Let $c = (X^TX)^{-1}X^Tw$

$$Hw = X(X^TX)^{-1}X^Tw = Xc \in Col(X)$$

Next, to show that the projection is orthogonal, we want to show that for any $v \in Col(X)$, $(w - Hw)^Tv = 0$. We make the observation that for any v, there exists c such that $v = Xc$ by definition of the column space. Then we have

$$(w - Hw)^Tv = w^T(I_n - H)Xc = w^T(X - HX)c = 0$$

since

$$X - HX = X - X(X^TX)^{-1}X^TX = X - X = 0$$

Example. Show that $Rank(H) = p$ and $Rank(I_n - H) = n - p$.
H is a projection matrix (symmetric and idempotent) onto $Col(X)$ so the rank is equal to its trace (see linear algebra review in the Appendix, A.6).

$$Rank(H) = Tr(H) = Tr(X(X^TX)^{-1}X^T) = Tr((X^TX)^{-1}X^TX) = Tr(I_p) = p$$

Similarly, $I_n - H$ is a projection matrix to the space orthogonal to $Col(X)$ so

$$Rank(I_n - H) = Tr(I_n - H) = Tr(I_n) - Tr(H) = n - p$$

2.2.3 Fitted Values and Residuals

With the hat matrix, we can write the fitted values as

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

with the residuals

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

As usual, we have that $\hat{Y}$ is an unbiased estimator for $\mathbb{E}Y$

$$\mathbb{E}\hat{Y} = \mathbb{E}HY = H\mathbb{E}Y = HX\beta = X\beta = \mathbb{E}Y$$
$$\mathbb{E}e = \mathbb{E}Y - \mathbb{E}\hat{Y} = 0$$

Example. Find the covariance matrix of $\hat{Y}$ and the covariance matrix of e.

$$Var(\hat{Y}) = H^T(\sigma^2 I_n)H = \sigma^2 H \ (since\ H\ is\ symmetric,\ idempotent)$$

Similarly,

$$Var(e) = \sigma^2(I_n - H)$$

since $I_n - H$ is a projection matrix to the space orthogonal to $Col(X)$, so it is also symmetric and idempotent.

2.3 ANOVA for Multiple Regression

Recall the sum of squares decomposition for variance, i.e. $SSTO = SSE + SSR$. Below, we express the sum of squares in matrix form. Note that $\bar{y} = \frac{1}{n}J_nY$ where J_n is a square matrix of ones, and that the matrices $I_n - \frac{1}{n}J_n$, $I_n - H$, and $H - \frac{1}{n}J_n$ are all projection matrices, that is, are all symmetric and idempotent.

$$SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \left((I_n - \frac{1}{n}J_n)Y\right)^T \left(Y(I_n - \frac{1}{n}J_n)\right) = Y^T(I_n - \frac{1}{n}J_n)Y$$
$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \left((I_n - H)Y\right)^T \left((I_n - H)Y\right) = Y^T(I_n - H)Y$$
$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \left((H - \frac{1}{n}J_n)Y\right)^T \left((H - \frac{1}{n}J_n)Y\right) = Y^T(H - \frac{1}{n}J_n)Y$$

The sum of squares, SSTO, SSE, and SSR have degrees of freedom equal to the ranks of the matrices $I_n - \frac{1}{n}J_n$, $I_n - H$, and $H - \frac{1}{n}J_n$, respectively. That is, SSTO has degrees of freedom $n-1$, SSE has degrees of freedom $n-p$, and SSR has degrees of freedom $p-1$.

Example. Show that $\mathbb{E}SSE = \sigma^2(n-p)$
Since $SSE \in \mathbb{R}$, we have that $SSE = Tr(SSE)$ so

$$\begin{aligned}\mathbb{E}SSE &= \mathbb{E}Tr(SSE) = \mathbb{E}Tr(Y^T(I_n - H)Y) \\ &= \mathbb{E}Tr((I_n - H)YY^T) \\ &= Tr((I_n - H)\mathbb{E}YY^T)\end{aligned}$$

We know that

$$Var(Y) = \mathbb{E}YY^T - \mathbb{E}Y\mathbb{E}Y^T$$
$$\mathbb{E}YY^T = Var(Y) + \mathbb{E}Y\mathbb{E}Y^T = \sigma^2 I_n + (X\beta)(X\beta)^T = \sigma^2 I_n + X\beta\beta^T X^T$$

Additionally, note that $HX = X$ since H projects to $Col(X)$ so that $(I_n - H)X = 0$. Putting it together, we have

$$\begin{aligned}\mathbb{E}SSE &= Tr(\sigma^2(I_n - H) + (I_n - H)X\beta\beta^T X^T) = \sigma^2 Tr(I_n - H) \\ &= \sigma^2(n-p)\end{aligned}$$

This implies that $\mathbb{E}MSE = \mathbb{E}SSE/(n-p) = \sigma^2$, so MSE is an unbiased estimator for σ^2.

Example. Show that SSE and SSR are independent.

We can begin by showing that the residuals e are independent of the fitted values $\hat{Y}$. Since both e and $\hat{Y}$ are multivariate normal random variables, it suffices to show that $Cov(e, \hat{Y}) = 0$ since uncorrelated jointly normal random variables are independent. Additionally, we use the fact that for deterministic matrices A, B and random vector Y, $Cov(AY, BY) = AVar(Y)B^T$. We have that

$$\begin{aligned} Cov(e, \hat{Y}) &= Cov((I_n - H)Y, HY) \\ &= (I_n - H)Var(Y)H^T \\ &= \sigma^2(I_n - H)H^T \\ &= \sigma^2(H - HH^T) \\ &= \sigma^2(H - H) = 0 \ (H \ is \ idemopotent, \ symmetric) \end{aligned}$$

We also note that SSE is a function of the residuals e since $SSE = e^T e$. Further, SSR is a function of the fitted values $\hat{Y}$. Noting that $\frac{1}{n}\sum_{i=1}^n \hat{Y}_i = \frac{1}{n}\sum_{i=1}^n Y_i + e_i = \frac{1}{n}\sum_{i=1}^n Y_i = \bar{Y}$ we have

$$SSR = ||\hat{Y} - \bar{Y}||_2^2 = ||\hat{Y} - \frac{1}{n}J_n\hat{Y}||_2^2$$

Using the fact that (non-random) functions of independent random variables are independent, we have that SSE is independent with SSR.

2.3.1 F-test for Regression Relationship

As for simple regression, we would like to test for the significance of the regression relationship. That is, we can test

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = ... = \beta_{p-1} = 0 \\ H_1 &: \vec{\beta} \neq 0 \end{aligned}$$

Under the normal error model, $SSE \sim \sigma^2\chi^2_{n-p}$. Under the null hypothesis, $SSR \sim \sigma^2\chi^2_{p-1}$. We also know that SSR and SSE are independent, so the ratio of the two, divided by their respective degrees of freedom, should follow an F distribution. That is, under the null hypothesis,

$$F^* = \frac{SSR/\sigma^2(p-1)}{SSE/\sigma^2(n-p)} = \frac{MSR}{MSE} \sim F_{p-1,n-p}$$

We reject the null hypothesis at the α level when

$$F^* > F(1-\alpha; p-1, n-p)$$

where $F(.; p-1, n-p)$ describes the inverse CDF function for the $F_{p-1,n-p}$ distribution.

The results of the ANOVA can be summarized by an ANOVA table (Table 2). Note that the degrees of freedom have been updated for the sum of squares compared to simple regression.

Table 2: ANOVA table for multiple regression

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F^*
Regression	$SSR = Y^T(H - \frac{1}{n}J_n)Y$	$p-1$	$MSR = SSR/(p-1)$	MSR/MSE
Error	$SSE = Y^T(I_n - H)Y$	$n-p$	$MSE = SSE/(n-p)$	
Total	$SSTO = Y^T(I_n - \frac{1}{n}J_n)Y$	$n-1$		

2.3.2 Coefficient of Determination

As before in simple regression, we can quantify the proportion of variation in Y explained by the Xvariables using the coefficient of determination

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

However, adding more X variables to the model will never decrease R^2. $SSTO$ is constant and remains the same, while SSE will not increase (SSR will not decrease). To understand this, we can observe that at the very least the regression coefficients for the new variables can be 0, so that SSE will remain the same as before.

A model that has too many X variables that are unrelated to the response variable and/or highly correlated with each other will risk overfitting (increase of sampling variability which leads to poor prediction results), be more difficult to interpret, and may incur costly model maintenance. Thus, we introduce the *adjusted coefficient of determination*, R_a^2, where $R_a^2 \leq R^2$.

$$R_a^2 = 1 - \frac{MSE}{MSTO} = 1 - \frac{n-1}{n-p}\frac{SSE}{SSTO}$$

Note that unlike R^2, it is possible for R_a^2 to decrease when adding more X variables since any decrease in SSE may be offset by its loss of degrees of freedom.

2.4 Estimation and Prediction

2.4.1 Inference for Regression Coefficients

Recall that $\mathbb{E}\hat{\beta} = \beta$ and $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$. For any particular regression coefficient $\hat{\beta}_k$ where $k = 0, ..., 1-p$, we can evaluate for the standard error $s\{\hat{\beta}_k\}$ using the square root of the $k+1$ diagonal element of $MSE(X^TX)^{-1}$. Then

$$\frac{\hat{\beta}_k - \beta_k}{s\{\hat{\beta}_k\}} \sim t_{(n-p)}$$

is a pivotal quantity. It follows that an α level confidence interval for β_k is

$$\mathbb{P}\left(\hat{\beta}_k - t(1-\alpha/2; n-p) * s\{\hat{\beta}_k\} < \beta_k < \hat{\beta}_k + t(1-\alpha/2; n-p) * s\{\hat{\beta}_k\}\right) = 1 - \alpha$$

As before, we can also test for the significance of a single regression coefficient

$$H_0 : \beta_k = 0$$
$$H_1 : \beta_k \neq 0$$

using the statistic

$$T^* = \frac{\hat{\beta}_k - \beta_k^{(0)}}{s\{\hat{\beta}_k\}} = \frac{\hat{\beta}_k - 0}{s\{\hat{\beta}_k\}}$$

rejecting the null hypothesis at the α level if

$$p = \mathbb{P}\left(|t_{(n-p)}| > |T^*|\right) < \alpha$$

2.4.2 Estimation of Mean Response

At a new X_h, $\hat{Y}_h$ is an unbiased estimator of $\mathbb{E}Y_h$ since

$$\mathbb{E}\hat{Y}_h = X_h^T\mathbb{E}\hat{\beta} = X_h^T\beta = \mathbb{E}Y_h$$

with variance

$$Var(\hat{Y}_h) = Var(X_h\hat{\beta}) = X_h^T Var(\hat{\beta}) X_h = \sigma^2 X_h^T (X^T X)^{-1} X_h$$

Then the standard error for $\hat{Y}_h$ is

$$s\{\hat{Y}_h\} = \sqrt{MSE * X_h^T (X^T X)^{-1} X_h}$$

and finally we have the pivotal quantity

$$\frac{\hat{Y}_h - \mathbb{E}Y_h}{s\{\hat{Y}_h\}} \sim t_{(n-p)}$$

with which we can create an α level confidence interval for $\mathbb{E}Y_h$

2.4.3 Prediction of a New Observation

At the new X_h, we model

$$Y_h = X_h^T \beta + \epsilon_h$$

where Y_h is independent with previous observations Y_i (by the same reasoning, Y_h is independent of $\hat{\beta}$ since $\hat{\beta}$ is a linear combination of the previous Y_i). Under the normal error model, we predict the new outcome using $\hat{Y}_h = X_h^T \hat{\beta}$. The variance of the prediction is then

$$Var(\hat{Y}_h - Y_h) = Var(\hat{Y}_h) - Var(Y_h) = \sigma^2 X_h^T (X^T X)^{-1} X_h + \sigma^2 = \sigma^2 (1 + X_h^T (X^T X)^{-1} X_h)$$

so the standard error is

$$s(pred_h) = \sqrt{MSE\left(1 + X_h^T (X^T X)^{-1} X_h\right)}$$

We can then build an α level prediction interval as follows

$$\hat{Y}_h \pm t(1 - \alpha/2; n - p) * s(pred_h)$$

where $t(.; n-p)$ is the inverse CDF for t_{n-p}.

3 Extra Sum of Squares

3.1 Defining the Extra Sum of Squares

This section address the question of comparing two different linear regression models by analyzing how adding variables affects the sum of squares. We first define the *Extra Sum of Squares* (ESS). Let I and J be (non-overlapping) sets of indices, i.e. $X_I = \{X_i : i \in I\}$ and $X_J = \{X_i : i \in J\}$. Let $SSE(X_I)$ and $SSE(X_I)$ denote the error sum of squares and regression sum of squares, respectively, under a regression model fitted with X_I. Then we can define the reduction in error sum of squares (SSE) by adding X_J to the model with X_I as the following

$$SSR(X_J|X_I) := SSR(X_I, X_J) - SSR(X_I) = SSE(X_I) - SSE(X_I, X_J)$$

Note that the degrees of freedom of $SSR(X_J|X_I)$ is simply $|J|$, or the number of additional X variables being added to the model with X_I. Naturally, the mean square is then

$$MSR(X_J|X_I) := \frac{SSR(X_J|X_I)}{|J|}$$

Since adding X variables never decreases SSR, $SSR(X_J|X_I) \geq 0$. Also note that the extra sum of squares are not commutative, that is, $SSR(X_J|X_I) \neq SSR(X_I|X_J)$.

We can decompose SSR into extra sums of squares. For example, by definition we see that

$$SSR(X_1, X_2) = SSR(X_1) + SSR(X_2|X_1)$$

However, the decomposition is not unique. For example,

$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2)$$
$$SSR(X_1, X_2, X_3) = SSR(X_2) + SSR(X_1|X_2) + SSR(X_3|X_1, X_2)$$

3.2 Generalized Linear Tests

Let I and J be non-overlapping index sets as before. Let the *full model* include both X_I and X_J. Let the *reduced model* include only X_I. We are interested in testing whether or not X_J can be dropped out of the full model. That is, we test if the reduction in SSE by adding X_J to the model with only X_I is significant.

$$H_0 : \beta_j = 0 \ \forall j \in J$$
$$H_1 : otherwise$$

We use the test statistic

$$F^* = \frac{SSR(X_J|X_I)/df(SSR(X_J|X_I))}{SSE(X_I, X_J)/df(SSE(X_I, X_J))} = \frac{MSR(X_J|X_I)}{MSE(X_I, X_J)}$$

Note that the numerator captures the reduction in SSE by adding X_J to the reduced model with X_I while the denominator is the MSE of the full model with both X_I and X_J. We can also write the test statistic as

$$F^* = \frac{(SSE_R - SSE_F)/(df(SSE_R) - df(SSE_F))}{SSE_F/df(SSE_F)}$$

where the subscripts R and F correspond to "reduced model" and "full model", respectively.
Under the null hypothesis, we have that

$$F^* \sim F_{df(SSE_R)-df(SSE_F), df(SSE_F)}$$

so we reject the null at the α level if

$$F^* > F(1-\alpha; df(SSE_R) - df(SSE_F), df(SSE_F))$$

where $F(.; df(SSE_R)-df(SSE_F), df(SSE_F))$ is the inverse cdf of an F distribution with $df(SSE_R)-df(SSE_F)$ and $df(SSE_F)$ degrees of freedom.

3.2.1 F-test for Regression Relationship

Suppose we have the full model with all $X_1, ..., X_{p-1}$ predictors and are comparing it to a reduced model with no X predictors (intercept only). For the full model, we have that $df(SSE_F) = n - p$ and for the reduced model we have $df(SSE_R) = n - 1 = df(SSTO)$. Also note that when there are no X predictors, the intercept (and thus all the fitted values) are equal to the mean of the response. Then we have under the reduced model

$$\hat{\beta}_0 = \hat{Y}_i = \bar{Y}$$
$$SSE_R = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = SSTO$$

Then the test statistic is

$$\begin{aligned} F^* &= \frac{\left(SSE_R - SSE_F\right)/\left(df(SSE_R) - df(SSE_F)\right)}{SSE_F/df(SSE_F)} \\ &= \frac{\left(SSTO - SSE_F\right)/\left((n-1) - (n-p)\right)}{SSE_F/(n-p)} \\ &= \frac{\left(SSR_F\right)/(p-1)}{SSE_F/(n-p)} = \frac{MSR_F}{MSE_F} \end{aligned}$$

which is exactly the same test statistic that we used to test the regression relationship in Section 2.3.

3.2.2 Test a Single Coefficient

Suppose we want to test if a single $\beta_k = 0$

$$\begin{aligned} H_0 &: \beta_k = 0 \\ H_1 &: \beta_k \neq 0 \end{aligned}$$

with the test statistic

$$F^* = \frac{(SSE_R - SSE_F)/1}{MSE_F}$$

which follows $F_{1,n-p}$ under the null distribution. This is equivalent to the two-sided T-test for a single regression coefficient.

3.2.3 Test Several Coefficients

Suppose we have a model with four different predictors ($p = 5$), X_1, X_2, X_3 , and X_4 and we want to test

$$\begin{aligned} H_0 &: \beta_2 = \beta_3 = 0 \\ H_1 &: otherwise \end{aligned}$$

Then we would use the test statistic

$$F^* = \frac{(SSE(X_1, X_4) - SSE(X_1, X_2, X_3, X_4))/2}{SSE(X_1, X_2, X_3, X_4)/n - 5}$$

which follows $F_{2,n-5}$ under the null distribution.

3.2.4 Test Equality of Coefficients

We can also test the equality of several regression coefficients.

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \beta_3 \\ H_1 &: otherwise \end{aligned}$$

To do so, we can set up the reduced model as follows

$$Y = \beta_0 + \beta_1(X_1 + X_2 + X_3) + \beta_4(X_4) + \epsilon$$

and compare it to the full model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

We have that the reduced model has 3 regression coefficients so $df(SSE_R) = n - 3$. The full model has 5 regression coefficients so $df(SSE_F) = n - 5$. Overall, we use the test statistic

$$F^* = \frac{(SSE_R - SSE_F)/((n-3) - (n-5))}{SSE_F/(n-5)} = \frac{(SSE_R - SSE_F)/2}{SSE_F/(n-5)}$$

which follows $F_{2,n-5}$ under the null distribution.

3.3 Regression Coefficients as Partial Coefficients

We first quantify the proportional reduction in SSE by adding a single X_j into the model where $j \notin I$. As before, let I be the set of indices in the reduced model.

$$R^2_{Y,j|I} := \frac{SSE(X_I) - SSE(X_j, X_I)}{SSE(X_I)} = \frac{SSR(X_j|X_I)}{SSE(X_I)}$$

which is a quantity between 0 and 1.

3.3.1 Extra Sum of Squares as SSR

It can be shown that $SSR(X_j|X_I)$ is the SSR of the simple regression between two sets of residuals, specifically, the residuals of the reduced model

$$e(Y|X_I) = Y - \hat{Y}(X_I)$$

and the residuals of regression X_j against X_I

$$e(X_j|X_I) = X_j - \hat{X}_j(X_I)$$

We can think of the first set of residuals as the portion of Y that is not in the column space of X_I (recall that the fitted values represent an orthogonal projection of Y onto the column space of the X; they are the portion of Y best explained by or closest to that column space). Similarly, the second set of residuals represent the portion of X_j that is not in the column space of X_I. Ultimately, we are interested in the *linear association between Y and X_j after the linear effects of the variables already in the model, X_I, have been removed*. Intuitively, the first set of residuals represent what we want to explain given that some portion of Y we have already explained with X_I. Similarly, the second set of residuals represent the portion of X_j, or new information from X_j, that we have not already accounted for with X_I.

3.3.2 Connection Between Simple and Multiple Regression

Example. Show that $R^2_{Y,j|I}$ is the coefficient of determination R^2 of the simple regression between the two sets of residuals $e(Y|X_I)$ and $e(X_j|X_I)$.

We know that $SSR(X_j|X_I)$ is SSR in the simple regression. Further we have that $SSTO$ in this context is

$$SSTO = \sum_{i=1}^{n}(e(Y|X_I)_i - \bar{e}(Y|X_I))^2 = \sum_{i=1}^{n}(e(Y|X_I)_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i(X_I))^2 = SSE(X_I)$$

so from definition we have

$$R^2_{Y,j|I} := \frac{SSR(X_j|X_I)}{SSE(X_I)} = \frac{SSR}{SSTO} = R^2$$

Lastly, consider the full model with all X variables, and let $\hat{\beta}_j$ be the least-squares fitted regression coefficient for X_j. Let X_{-j} denote all X variables except X_j (in the previous notation, we have $X_I = X_{-j}$). It can be shown that $\hat{\beta}_j$ is precisely the regression slope when simple regression the sets of residuals $e(Y|X_{-j})$ and $e(X_j|X_{-j})$. This means that we should interpret the regression coefficients in multiple regression as the variables individual contribution in linearly explaining the response, *accounting for the linear effects of all the other variables in the model*. We should interpret regression coefficients within the context of other variables in the model; they should not be interpreted independently unless the X variables are uncorrelated (see Section 5).

4 Standardization

4.1 Standardized Regression Model

In some cases, X variables could differ substantially in order of magnitude, which means that the regression coefficients are not comparable with each other. Differences in scale could also cause numerical instability when inverting X^TX. As a solution, we transform each of the X variables so that they have 0 mean and equal standard deviation. Specifically, for each $k = 1, ...p-1$ we compute

$$X_k^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_k - \bar{X}_k}{S_k}\right)$$

where $\bar{X}_k = \frac{1}{n}\sum_{i=1}^n X_{ik}$ is that sample mean and $S_k = \sqrt{\frac{\sum_{i=1}^n (X_{ik}-\bar{X}_k)^2}{n-1}}$ is the sample standard deviation for X_k. Note that as a result of this transformation, we have that $\bar{X}_k^* = 0$ and $S_k^* = \frac{1}{\sqrt{n-1}}$.

To go back to the original X_k we have $X_k = X_k^*\sqrt{n-1}S_k + \bar{X}_k$. Then

$$\beta_k X_k = \beta_k(X_k^*\sqrt{n-1}S_k + \bar{X}_k) = \beta_k\sqrt{n-1}S_k X_k^* + \beta_k\bar{X}_k$$

From here, it is natural to define

$$\beta_k^* = \sqrt{n-1}S_k\beta_k$$

where the second terms $\beta_k\bar{X}_k$ are grouped into β_0^*

$$\beta_0^* = \beta_0 + \sum_{k=1}^{p-1}\beta_k\bar{X}_k$$

Putting it together, we have the standardized regression model as follows

$$Y = \beta_0^* + \beta_1^*X_1^* + ... + \beta_{p-1}^*X_{p-1}^* + \epsilon$$

4.2 Least Squares Estimator for Standardized Model

Example. Find the least squares estimator for the standardized model.

It can be verified that $\sum_{i=1}^n X_{ik}^* = 1$ for any $k = 1, ..., p-1$, and more generally, $\sum_{i=1}^n X_{ik}^*X_{il}^* = r_{kl}$ where r_{kl} is the sample correlation between X_k and X_l.

Given these facts, we can write out $X^{*T}X^*$ using the standardized design matrix $X^* \in \mathbb{R}^{n,p}$

$$X^{*T}X^* = \begin{pmatrix} n & 0^T \\ 0 & r_{xx} \end{pmatrix} \in \mathbb{R}^{p,p}$$

where $r_{xx} \in \mathbb{R}^{(p-1),(p-1)}$ is the sample correlation matrix for the X variables. Taking the inverse we have

$$(X^{*T}X^*)^{-1} = \begin{pmatrix} 1/n & 0^T \\ 0 & r_{xx}^{-1} \end{pmatrix}$$

Similarly, we can write an expression for $X^{*T}Y$

$$X^{*T}Y = \sqrt{n-1}S_y\begin{pmatrix} \frac{n\bar{Y}}{\sqrt{n-1}S_y} \\ r_{xy} \end{pmatrix} \in \mathbb{R}^{p,1}$$

Finally, evaluating for $\hat{\beta}^*$

$$\hat{\beta}^* = (X^{*T}X^*)^{-1}X^{*T}Y = \begin{pmatrix} \bar{Y} \\ \sqrt{n-1}S_y r_{xx}^{-1}r_{xy} \end{pmatrix} \in \mathbb{R}^{p,1}$$

As before, we have that the estimators are unbiased

$$\mathbb{E}\hat{\beta}^* = \beta^*$$

with variance

$$Var(\hat{\beta}^*) = \sigma^2(X^{*T}X^*)^{-1} = \sigma^2 \begin{pmatrix} 1/n & 0^T \\ 0 & (r_{xx})^{-1} \end{pmatrix}$$

Importantly, following this standardization, the fitted values, residuals, and sums of squares are the same as the original model. The following is the relationship between the original and standardized least squares estimators.

$$\hat{\beta}_k^* = \sqrt{n-1}S_k\hat{\beta}_k, \ k = 1, ..., p-1$$
$$\hat{\beta}_0^* = \hat{\beta}_0 + \sum_{k=1}^{p-1} \hat{\beta}_k \bar{X}_k$$

5 Multicollinearity

5.1 Uncorrelated Predictor Variables

Consider the case where the X variables $X_1, X_2, ..., X_{p-1}$ are uncorrelated with each other. Then we have that $r_{xx} = I_{p-1}$ and under the standardized model (we always can convert back to the original model),

$$\hat{\beta}^* = \begin{pmatrix} \bar{Y} \\ \sqrt{n-1}S_y r_{xx}^{-1} r_{xy} \end{pmatrix} = \begin{pmatrix} \bar{Y} \\ \sqrt{n-1}S_y r_{xy} \end{pmatrix}$$

with covariance matrix

$$Var(\hat{\beta}^*) = \sigma^2 \begin{pmatrix} 1/n & 0^T \\ 0 & I_{p-1} \end{pmatrix}$$

Notably, we have that for any $k = 1, ..., p-1$

$$\hat{\beta}_k^* = \sqrt{n-1}S_y r_{x_k y}$$

and that the fitted least squares coefficients are uncorrelated with each other (under the normal error model, they are independent of each other). That is, the contribution of any X variable in reducing the error sum of squares does not depend on the other X variables in the model. In other words, the contribution of the X variable in reducing the error sum of squares is equal to its marginal effect.

$$SSR(X_j|X_{-j}) = SSR(X_j)$$

Example. Show the fitted regression coefficients are the same when regressing Y on $X_1, .., X_{p-1}$ and when regressing Y on each individual X_j alone, given the X variables are uncorrelated.

We will show this for the standardized model, since we can always can convert back to the original model. We know that when conducting the multiple regression with uncorrelated X, we have that for any $k = 1, .., .p-1$

$$\hat{\beta}_k^* = \sqrt{n-1}S_y r_{x_k y}$$

For regressing Y on an individual X_j, we have

$$\hat{\beta}_j^* = r_{x_j y}\frac{S_y}{S_{x_j}}$$

Since in the standardized model, $S_{x_j} = 1/\sqrt{n-1}$ then

$$\hat{\beta}_j^* = \sqrt{n-1}S_y r_{x_j y}$$

which is the same result as the multiple regression model.

5.2 Correlated Predictor Variables

Now consider the case where X variables are inter-correlated. That is, there exists a nonzero vector c such that $Xc \approx 0$ and X^TX is nearly singular.

In this case, the fitted regression coefficient varies significantly depending on which other X variables are in the model. Thus, the standard errors of the fitted regression coefficients are inflated when more X variables are added to the model. For example, consider if X_1 and X_2 are highly correlated with each other and with the response Y. When X_2 is already in the model, the additional contribution from X_1 in explaining Y is small since X_2 contains much of the same information as X_1 in terms of linearly explaining Y.

With multicollinearity, the estimated regression coefficients tend to have large sampling variability (inflated variance, large standard errors) which results in wide confidence intervals. It is also possible that none of the regression coefficients are significant when tested individually (i.e. T-test), but there is a significant regression relationship on the entire set of predictors (i.e. F-test). This is because the T-test for a single regression coefficient is the same as testing the coefficient against the full model, where much of the information is already contained.

5.2.1 Variance Inflation Factor

Recall that the variance of the least squares estimator $\hat{\beta}_k^*$ under the standardized model is

$$Var(\hat{\beta}_k^*) = \sigma^2(r_{xx}^{-1}[k,k])$$

where $r_{xx}^{-1}[k,k]$ is the kth diagonal of r_{xx}^{-1}. We define the *Variance Inflation Factor* (VIF) as the following

$$VIF(\hat{\beta}_k^*) := r_{xx}^{-1}[k,k]$$

It can be shown that VIF can also be calculated as the following

$$VIF_k = \frac{1}{1-R_k^2}$$

where R_k^2 is the coefficient of multiple determination when regressing X_k against X_{-k}. If X_k is uncorrelated with other predictors, then $R_k^2 = 0$ and $VIF_k = 1$ and there is no variance inflation. If $R_k^2 > 0$ then $VIF_k > 1$ and the variance of the estimator is inflated due to inter-correlation between X_k and the other predictors. If X_k has perfect linear association with the other predictors, then R_k^2 and $VIF_k = \infty$. In this case, the least squares estimators are poorly defined (X^TX is not invertible).

In practice, $\underset{k}{max} VIF_k > 10$ is a rule of thumb indicating high multicollinearity.

6 Model Selection

6.1 The Bias-Variance Trade-off

We define a *correct model* as a model that contains all important predictor X variables. In this case, there is very little bias in the model, however, the correct model may not be a good model. That is, there can be too many nuisance variables which causes large sampling variability due to high multicollinearity and ultimately leads to overfitting. A good model should contain all important X variables, yet at the same time it should have few nuisance variables. Such models are said to achieve the bias-variance tradeoff - we omit some nuisance variables (potentially introducing bias) to achieve lower variance.

As before, let $Var(Y) = \sigma^2 I_n$ and let M denote an arbitrary model (not necessarily a correct model). If M is a correct model, then it is unbiased and there exists a vector β such that $\mu = EY = X\beta$ i.e. $\mu \in Col(X)$. Up until now, we have assumed that our models were correct.

6.1.1 Model Variance

The in-sample variance of M is defined as the overall variance of the fitted values, that is, the sum of the variances of each of the fitted values

$$Var_{in}(M) := \sum_{i=1}^{n} Var(\hat{Y}_i) = Tr(\sigma^2 H_M) = \sigma^2 p$$

That is, larger models always have larger variances whether they are correct or not.

6.1.2 Model Bias

We define the in-sample bias of M that represents the overall biases of the fitted values by the Euclidean norm of the bias vector

$$bias_{in}(M) := ||\mathbb{E}\hat{Y} - \mathbb{E}Y||_2 = ||(H - I_n)\mu||_2$$

The model bias depends on how well the column space of X approximates the response mean vector μ; $bias_{in}(M)$ is L2-norm of the portion of μ that is not in $Col(X)$.

Example. Find $\mathbb{E}SSE$ under a biased model.

$$\begin{aligned}
\mathbb{E}SSE &= \mathbb{E}Tr(SSE) = \mathbb{E}Tr(Y^T(I_n - H)Y) \\
&= Tr((I_n - H)\mathbb{E}YY^T) \\
&= Tr((I_n - H)(Var(Y) + \mathbb{E}Y\mathbb{E}Y^T) \\
&= Tr((I_n - H)(\sigma^2 I_n + \mu\mu^T) \\
&= \sigma^2 Tr(I_n - H) + Tr(\mu^T(I_n - H)\mu) \\
&= \sigma^2(n - p) + ||(I_n - H)\mu||_2^2 \\
&= \sigma^2(n - p) + bias_{in}^2
\end{aligned}$$

6.1.3 Mean Squared Estimation Error (MSEE)

When estimating the parameter θ with $\hat{\theta}$ we have that the mean squared estimation error (theoretical quantity) is

$$\begin{aligned}
\mathbb{E}(\theta - \hat{\theta})^2 &= \mathbb{E}(\theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta})^2 \\
&= (\theta - \mathbb{E}\hat{\theta})^2 + \mathbb{E}(\mathbb{E}\hat{\theta} - \hat{\theta})^2 + 2\mathbb{E}(\theta - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \hat{\theta}) \\
&= bias^2(\hat{\theta}) + Var(\hat{\theta})
\end{aligned}$$

The cross-product term vanishes since $\mathbb{E}(\mathbb{E}\hat{\theta} - \hat{\theta}) = 0$. Applying this decomposition to a single fitted value $\hat{Y}_i$ for model M,

$$MSEE_i(M) := \mathbb{E}(\hat{Y}_i - \mu_i)^2 = Var(\hat{Y}_i) + (\mathbb{E}\hat{Y}_i - \mu_i)^2$$

We have that the in-sample MSEE (representing all data observations) is

$$\begin{aligned}
MSEE_{in}(M) &:= \sum_{i=1}^{n} MSEE_i(M) \\
&= \sum_{i=1}^{n} Var(\hat{Y}_i) + ||\mathbb{E}\hat{Y} - \mu||_2^2 \\
&= Var_{in}(M) + bias_{in}^2(M)
\end{aligned}$$

Larger models have larger variances and tend to "overfit" the data (i.e. model parameters and fitted values are more sensitive to the error noise) but have lower bias since $Col(X)$ is larger and may be able to better approximate μ. On the other hand, smaller models may have too much bias and may "underfit" the data.

6.2 Selection Criterion

In model selection, the *full model* is said to contain all $p-1$ potential X variables and is assumed to be correct. The *candidate model* contains a subset of the potential X variables, that is, there are 2^{p-1} potential candidate models. The goal of model selection is to find a candidate that balances bias and variance. There are various criterion we can use to evaluate and compare candidate models, including R_a^2, AIC, BIC, $Press_p$, etc...

6.2.1 Mallow's C_p Criterion

Mallow's C_p is defined as follows

$$C_p := \frac{SSE_p}{\hat{\sigma}^2} - (n-2p)$$

where $\hat{\sigma}^2$ is the MSE of the full model.

To understand Mallow's C_p we make the very rough approximation

$$\begin{aligned}
\mathbb{E}C_p(M) &\approx \frac{\mathbb{E}SSE(M)}{\sigma^2} - (n-2p) \\
&= \frac{(n-p)\sigma^2 + bias_{in}^2(M)}{\sigma^2} - (n-2p) \\
&= \frac{p\sigma^2 + bias_{in}^2(M)}{\sigma^2} \\
&= \frac{Var_{in}(M) + bias_{in}^2(M)}{\sigma^2} \\
&= \frac{MSEE_{in}(M)}{\sigma^2}
\end{aligned}$$

In other words, Mallow's C_p can be seen as a proxy for $MSEE$ of a candidate model. If the candidate is unbiased, then $C_p \approx p$. In general, we look for models where $C_p \approx p$ and have fewer X variables (smaller variance).

6.2.2 Information Criterion

Founded in information theory, information criterion estimate the amount of information "lost" when using a statistical model. They contain two terms, a term that rewards goodness of fit and another that penalizes model complexity (model size). That is, information criterion deal with the bias-variance tradeoff.

Akaike Information Criterion (AIC) is defined as follows

$$AIC_p := nlog\left(\frac{SSE_p}{n}\right) + 2p$$

Similarly, *Bayesian Information Criterion* (BIC) is defined as

$$BIC_p := nlog\left(\frac{SSE_p}{n}\right) + plog(n)$$

When $n \geq 8$ then $log(n) > 2$ and BIC puts more penalty on model complexity than AIC. That is, BIC should be used over AIC when the goal is to choose smaller models.

6.2.3 $Press_p$ Criterion

The $Press_p$ criterion, also known as leave-one-out cross validation, evaluates the (out of sample) predictive performance of the model. Models that balance overfitting and underfitting will have the highest out of sample performance.

Let Y_i be the ith observed response value. Let $\hat{Y}_{i(i)}$ be the predicted values for the ith case after fitting a model using all data points except the ith one. Define the *deleted residual* d_i as

$$d_i := Y_i - \hat{Y}_{i(i)}$$

The $Press_p$ is defined as follows

$$Press_p := \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} Y_i - \hat{Y}_{i(i)}$$

Example. Show that the deleted residual is equivalent to $d_i = \frac{e_i}{1-h_{ii}}$ where e_i is the residual fitted on all datapoints and $h_{ii} = H[i,i]$ is the ith diagonal of the hat matrix.

Let $\hat{\beta}_{(i)}$ denote the leave-i-out regression coefficients, which minimizes

$$Q_{(i)} = \sum_{j \neq i} (Y_j - x_j^T \beta)^2$$

where x_j are the X values for the jth observation.

Define $\tilde{Y} := (Y_1, ...\hat{Y}_{i(i)}...Y_n)$ that replaces Y_i with $\hat{Y}_{i(i)}$. Minimizing the following least squares criterion also gives $\hat{\beta}_{(i)}$, that is

$$\hat{\beta}_{(i)} = \underset{\beta}{argmin} ||\tilde{Y} - X\beta||_2^2$$

This can be easily seen since

$$||\tilde{Y} - X\beta||_2^2 = \sum_{j \neq i} (Y_j - x_j^T \beta)^2 + (\hat{Y}_{i(i)} - x_i^T \beta)^2$$

However, when $\beta = \hat{\beta}_{(i)}$ the second term vanishes and $||\tilde{Y} - X\hat{\beta}_{(i)}||_2^2 = Q_{(i)}(\hat{\beta})$. That is, $\hat{\beta}_{(i)}$ is the fitted coefficient using $\tilde{Y}$ as the response vector instead of Y. Then

$$\begin{aligned}
\hat{\beta}_{(i)} &= (X^T X)^{-1} X^T \tilde{Y} \\
\hat{Y}_{(i)} &= X\hat{\beta}_{(i)} = X(X^T X)^{-1} X^T \tilde{Y} = H\tilde{Y}
\end{aligned}$$

That is, $\hat{Y}_{i(i)}$ is the ith element of $H\tilde{Y}$. Then we have

$$\begin{aligned}
\hat{Y}_{i(i)} = (H\tilde{Y})_i &= \sum_{j=1}^{n} h_{ij} \tilde{Y}_j = \sum_{j \neq i}^{n} h_{ij} Y_j + h_{ii} \hat{Y}_{i(i)} \\
&= \sum_{j \neq i}^{n} h_{ij} Y_j + h_{ii} \hat{Y}_{i(i)} + h_{ii} Y_i - h_{ii} Y_i \\
&= \sum_{j=1}^{n} h_{ij} Y_j + h_{ii} (\hat{Y}_{i(i)} - Y_i) \\
&= \hat{Y}_i - h_{ii} d_i
\end{aligned}$$

Finally evaluating,

$$d_i = Y_i - \hat{Y}_{i(i)} = Y_i - \hat{Y}_i + h_{ii} d_i = e_i + h_{ii} d_i$$

And rearranging

$$d_i = \frac{e_i}{1 - h_{ii}}$$

In other words, we can evaluate for the $Press_p$ criterion using the following

$$Press_p = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{1 - h_{ii}} \right)^2$$

6.3 Stepwise Procedures

The number of possible candidate models, or subsets of the full model, is 2^{p-1} which grows very fast with the number of potential X variables $(p-1)$. That is, best subsets algorithms (exhaustive search) are infeasible to use when p is large. Instead, we consider stepwise regression which is a greedy algorithm that adds or deletes a single X at each step. Note that stepwise procedures may find a local optimum rather than the global best model.

6.3.1 Forward Stepwise Procedure

Given the model selection criteria (i.e. AIC, BIC, etc...), the initial model M_0 (i.e. the null model with only the intercept), and the pool of potential X variables, we start from M_0 and consider adding a new variable or deleting a current one at each step based on the change in the selection criteria. That is, starting from M_0 we repeat the following steps until the selection criterion no longer improves

- Consider the change in the selection criterion by adding a variable from the pool of variables not already in the model
- Consider the change in the selection criterion by deleting a variable already in the model
- Choose the operation that improves the selection criterion the most

6.3.2 Forward Selection and Backward Elimination

As opposed to the forward stepwise procedure, forward selection only considers adding variables starting from the null model and backward elimination only considers deleting variables starting from the full model. However, backward methods cannot be used when $p > n$ and the full model cannot be fitted. Overall, forward stepwise regression is optimal with high multicollinearity between variables, since it considers both deleting and adding variables. Another common strategy in practice is to perform one pass of forward selection, and then one pass of backward elimination.

7 Outliers

In this section, we discuss the case when the data contains observations that are outlying or extreme.

7.1 Outlying in Y

Recall the base assumption that $Var(Y) = \sigma^2 I_n$ which can be used to show that the variance of the residuals is

$$Var(e) = Var((I_n - H)Y) = (I_n - H)(\sigma^2 I_n)(I_n - H)^T = \sigma^2(I_n - H)$$

with standard error

$$s^2\{e\} = MSE(I_n - H)$$

That is, the variance of the ith residual is $\sigma^2(1 - h_{ii})$ where h_{ii} is the ith diagonal of the hat matrix, $H[i, i]$.

Example. Show that the variance of the ith residual ranges from 0 to σ^2.

To show this, we first need to show that the diagonals of the hat matrix are between 0 and 1. Since we know that the $HH = H$ and $H^T = H$ we have that

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2$$

which shows that $h_{ii} \geq 0$. Further, we see that

$$\begin{aligned} h_{ii} &= \sum_{j \neq i} h_{ij}^2 + h_{ii}^2 \\ h_{ii} - h_{ii}^2 &= \sum_{j \neq i} h_{ij}^2 \\ h_{ii}(1 - h_{ii}) &= \sum_{j \neq i} h_{ij}^2 \geq 0 \end{aligned}$$

which implies that $h_{ii} \leq 1$. Thus, $0 \leq h_{ii} \leq 1$.Since the residual variance is $\sigma^2(1 - h_{ii})$, we have that it ranges between 0 and σ^2.

As an aside, note that studentized residuals (denoted r_i) have roughly constant variance and thus are more comparable to each other (residual plots in R use studentized residuals by default).

$$r_i = \frac{e_i}{s\{e_i\}} = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

However, when detecting outliers, deleted residuals are more useful since the outlying points themselves does not affect the fitted values (see Figure 1). Recall that deleted residuals can be written as the following (see Section 6.2.3).

$$d_i := Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1 - h_{ii}}$$

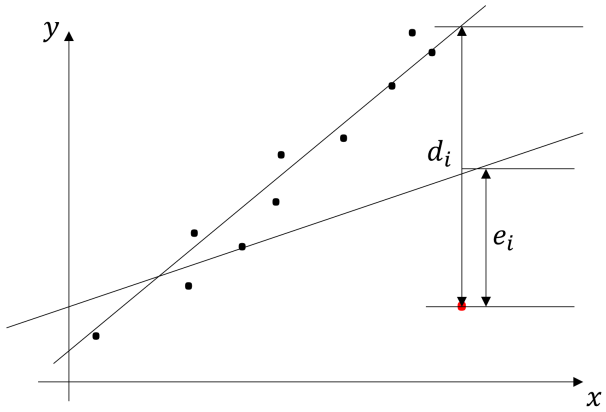

Figure 1: Residual vs deleted residual for an outlying point (highlighted in red)

We can also calculate studentized deleted residuals (also known as externally studentized residuals)

$$t_i := \frac{d_i}{s\{d_i\}} = \frac{d_i}{\sqrt{MSE_{(i)}/(1 - h_{ii})}}$$

where $MSE_{(i)}$ is the MSE of the regression model fitted excluding case i. As with calculating the deleted residuals, a "shortcut" can be used to calculate the studentized deleted residuals using all observations

$$t_i = e_i \sqrt{\frac{n - p - 1}{SSE(1 - h_{ii}) - e_i^2}}$$

We can use multiple testing to identify outlying cases in Y. We test the hypothesis(es)

$$H_0 : model\ is\ correct\ and\ all\ cases\ follow\ the\ model$$
$$H_{1i} : ith\ case\ is\ outlying\ in\ Y$$

Under the null hypothesis, we have that the model is correct so $\mathbb{E}t_i = \mathbb{E}d_i = \mathbb{E}e_i = 0$ and

$$t_i \sim t_{(n-p-1)}$$

At level α, the Bonferroni correction identifies cases with

$$|t_i| > t(1 - \alpha/(2n); n - p - 1)$$

as outlying in Y, where $t(.; n - p - 1)$ is the inverse cdf for the $t_{(n-p-1)}$ distribution.

7.2 Leverage and Outlying in X

The ith diagonal of H, h_{ii} is called the *leverage* of the ith observation. Since $\hat{Y} = HY$ then

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij}Y_j = h_{ii}Y_i + \sum_{j \neq i} h_{ij}Y_j$$

The leverage h_{ii} measures the role of the X values in determining the fitted value $\hat{Y}_i$. We also see that the larger h_{ii} is, the more important Y_i is in determining $\hat{Y}_i$.

We can write

$$h_{ii} = x_i^T(X^TX)^{-1}x_i = \frac{1}{n} + x_i^{*T}(r_{xx})^{-1}x_i^*$$

where x_i represents the ith observation vector and x_i^* is the standardized ith observation vector. Recall that

$$x_i^{*T} = \frac{1}{\sqrt{n-1}}\left(X_{i1} - \bar{X}_1, ..., X_{i,p-1} - \bar{X}_{p-1}\right)$$

That is, h_{ii} reflects the Mahalanobis distance between $\vec{x}_i$ and the sample mean of the X values. Mahalanobis distance is a distance measure that is weighted based on the correlation structure of the data.

The mean leverage is

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = Tr(H)/n = p/n$$

In practice, values where $h_{ii} > 2p/n$ are identified as outlying in X when the sample size is moderately large, as a rule of thumb.

7.3 Influential Cases

Ultimately, we are interested in outlying cases in either X or Y are influential in determining the fitted regression function. To do this, we look at *Cook's Distance*, which aggregates the influence on all fitted values by the omission of a single case in the fitting process. We define

$$D_i := \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p(MSE)}$$

Similar to the calculation of deleted residuals, we can calculate D_i using all cases as follows

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

where r_i is the ith studentized residual. Note that the magnitude in D depends on both r_i and h_{ii}. Outlying in X or Y alone does not make a case influential. In practice, a rule of thumb of $D_i > \frac{4}{n-p}$ is often used as an indication of high influence.
.

A Appendix: Linear Algebra Review

A.1 Fundamental Subspaces of a Matrix

Let A be a mxn matrix. The four fundamental subspaces of A include the column space (also known as range), row space, null space, and left null space.

The column space is defined as the span of the columns of A, that is

$$Col(A) = \{v : v = Ax, x \in \mathbb{R}^n\}$$

The row space is defined as the span of the rows of A, that is

$$Row(A) = Col(A^T) = \{v : v = A^T x, x \in \mathbb{R}^m\}$$

The null space is the vector subspace defined by

$$Null(A) = \{v : Av = 0\}$$

Recall that for a matrix to be invertible, the null space must only contain the trivial solution $v = 0$.

Finally, the left null space is defined as follows

$$Null(A^T) = \{v : A^T v = 0\}$$

The Fundamental Theorem of Linear Algebra also states that $Col(A)$ and $Null(A^T)$ are orthogonal to each other. Similarly, $Col(A^T)$ and $Null(A)$ are orthogonal to each other. To understand this, let $x \in Null(A)$ so that $Ax = 0$.

$$Ax = \begin{pmatrix} -a_1^T- \\ -a_2^T- \\ \dots \\ -a_m^T- \end{pmatrix} x = \begin{pmatrix} a_1^T x \\ a_2^T x \\ \dots \\ a_m^T x \end{pmatrix} = \vec{0}$$

The inner product of any row of A with x is 0, so it follows that the innder product of x with any linear combination of rows is also 0. In other words, x and any vector in $Row(A)$ are orthogonal. That is, $Null(A)$ and $Row(A)$ are orthogonal spaces.

A.2 Eigenvalues and Eigenvectors

An eigenvector of an nxn square matrix A is a non-zero vector $\vec{e} \in \mathbb{R}^n$ such that

$$Ae = \lambda e$$

for some scalar λ called an eigenvalue of A. We say that e is an eigenvector corresponding to λ.
Note that $e = 0$ is not a valid eigenvector (since the above relationship will be trivial) but $\lambda = 0$ is a valid eigenvalue. In this case,

$$Ae = 0e$$
$$Ae = 0$$

which only has a solution when A is not invertible (i.e. A is singular). To understand intuitively why this is the case, consider Ax as a linear transformation of the vector x. If x is mapped to the origin (the zero vector), there is no way to recover information from the transformation, and therefore undo it.

To find the eigenvalues of A, we can simply write

$$Ae - \lambda e = 0$$
$$(A - \lambda I)e = 0$$

which has a solution when $A - \lambda I$ is singular, which also indicates that the determinant of $A - \lambda I$ is 0. Thus, we want to solve for λ such that

$$det(A - \lambda I) = 0$$

This is known as the characteristic equation. To solve for the corresponding eigenvectors, we only need to solve for e after evaluating $(A - \lambda I)e = 0$ for λ.

Additionally note that the eigenvalues of an diagonal matrix are simply the diagonal elements themselves with eigenvectors corresponding to the standard basis vectors.

A.3 Diagonalization

Suppose the nxn matrix A has n real eigenvalues $\lambda_1, ..., \lambda_n$ and n associated eigenvectors $e_1, .., e_n$. Then note that all eigenvector/eigenvalue equations can be expressed with

$$AQ = Q\Lambda$$

where $Q = [e_1; , ..; e_n] \in \mathbb{R}^{nxn}$ consists of the eigenvectors of A in its columns and $\Lambda = diag(\lambda_1, ..., \lambda_n)$. Then we can write

$$A = AQQ^{-1} = Q\Lambda Q^{-1}$$

We call such A as diagonalizable, and A has the following properties

1. $tr(A) = tr(Q\Lambda Q^{-1}) = tr(QQ^{-1}\Lambda) = \sum_{i=1}^{n} \lambda_i$
2. $det(A) = \prod_{i=1}^{n} \lambda_i$
3. if A is nonsingular (A^{-1} exists), then $1/\lambda_i$ is an eigenvalue of A^{-1} with associated eigenvector e_i

Note that here, we assume that the inverse of Q exists. Thus, for the matrix A to be diagonalizable, A must have n linearly independent eigenvectors. That is, the n eigenvectors form a basis in $\mathbb{R}^n$ for a diagonalizable A. This is not to be confused with A being nonsingular, when A is singular there are eigenvectors that correspond to $\lambda = 0$ which represent the null space, where $Ae = 0$.

A.4 Diagonalization of Symmetric Matrices

Now suppose that the matrix A is symmetric (that is $A^T = A$). It can be shown that any two eigenvectors (corresponding to different eigenvalues/ different eigenspaces) are orthogonal to each other for a symmetric matrix. The proof of this is simple. Let v_1 and v_2 be eigenvectors that correspond to distinct eigenvalues λ_1 and λ_2. Then

$$\lambda_1 v_1 \cdot v_2 = (\lambda_1 v_1)^T v_2$$

since v_1 is an eigenvector, we can write the above as

$$(Av_1)^T v_2 = v_1^T A^T v_2 = v_1^T (Av_2)$$

since v_2 is an eigenvector, we can write the above as

$$v_1^T (\lambda_2 v_2)$$

This implies that

$$\lambda_1 v_1 \cdot v_2 = \lambda_2 v_1 \cdot v_2$$
$$(\lambda_1 - \lambda_2) v_1 \cdot v_2 = 0$$

Since we have that $\lambda_1 \neq \lambda_2$, we have shown that $v_1 \cdot v_2 = 0$ so v_1 and v_2 are orthogonal. Thus, distinct eigenspaces are mutually orthogonal.

A symmetric nxn matrix A is special because we can show that

1. A has n real eigenvalues (counting multiplicities)
2. The geometric multiplicity of λ, $dim(Null(A - \lambda I))$, that is, the dimension or number of basis vectors in the eigenspace of λ, is equal to the algebraic multiplicity of λ (the order corresponding to λ in the characteristic equation)
3. The eigenspaces are mutually orthogonal (as we have shown above)

Note that it is possible that an eigenvalue may have larger multiplicity (put plainly, the case where multiple eigenvectors corresponding to the same eigenvalue). Utilizing fact 2 above, we can simply choose for a specific eigenvalue λ an orthonormal basis for the eigenspace $\{v : Av = \lambda v\}$, for example, using the Gram-Schmidt process. We do this so that all chosen eigenvectors $e_1, ..., e_n$ are mutually orthogonal, whether they correspond to the same eigenspace or not. Such a diagonalization of a symmetric A is called an *orthogonal* diagonalization. In fact, A is orthogonally diagonalizable *if and only if* it is symmetric.

Specifically, we can now take an orthonormal matrix Q (we have chosen eigenvectors such that every column of Q has unit length and are mutually orthogonal). Since we now have $QQ^T = Q^TQ = I$, we see that $Q^T = Q^{-1}$. Then we can write the diagonalization of A as

$$A = Q\Lambda Q^{-1} = Q\Lambda Q^T$$
$$A = \lambda_1 e_1 e_1^T + ... + \lambda_n e_n e_n^T$$

This is known as the spectral decomposition of the symmetric matrix A.

A.5 Quadratic Forms

A quadratic form on $\mathbb{R}^n$ is a function written as

$$f(x) = x^T Ax$$

where A is a nxn *symmetric* matrix. For example, in $\mathbb{R}^2$

$$\begin{aligned} x^T Ax &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 3 & -2 \\ -2 & 7 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= 3x_1^2 - 2x_1x_2 - 2x_1x_2 + 7x_2^2 \end{aligned}$$

Note that the diagonals of A represent the coefficients for the quadratic terms, and the off diagonals represent the cross product terms.

We say that a matrix A is *positive definite* if for all $x \in \mathbb{R}^n$

$$x^T Ax > 0$$

and *positive semdefinite* if for all $x \in \mathbb{R}^n$

$$x^T Ax \geq 0$$

To check if a symmetric matrix is positive semidefinite, it is sufficient to check if all the eigenvalues of A are non-negative, that is

$$\lambda_1, .., \lambda_n \geq 0$$

To understand this, we can substitute the spectral decomposition of A into the definition of positive semidefinite

$$\begin{aligned} x^T Ax &= x^T(\lambda_1 e_1 e_1^T + ... + \lambda_n e_n e_n^T)x \\ &= \lambda_1 x^T e_1 e_1^T x + ... + \lambda_n x^T e_n e_n^T x \\ &= \lambda_1 (x^T e_1)^2 + ... + \lambda_n (x^T e_n)^2 \end{aligned}$$

which is non-negative for all x if and only if all the eigenvalues $\lambda_1, .., \lambda_n \geq 0$. Additionally, using a similar argument, we can show that if a matrix A is positive definite, we have that its eigenvalues $\lambda_1, .., \lambda_n > 0$.

We write

$$A \succeq 0$$

denoting A as positive semidefinite. If A is positive definite we write

$$A \succ 0$$

If we write

$$A \succeq B$$

it denotes that $A - B$ is positive semidefinite.

A.6 Projection Matrix

A projection matrix P onto a vector subspace defines a mapping from an arbitrary vector onto the closest point in that vector subspace. Below, we show that a matrix of the form $P = A(A^T A)^{-1} A^T \in \mathbb{R}^{n,n}$ defines an orthogonal projection onto $Col(A)$ where $A \in \mathbb{R}^{n,m}$.

Let y be an arbitrary vector and Ax be its respective projection onto $Col(A)$. By construction of an orthogonal projection, we have that $b - Ax$ is orthogonal to $Col(A)$. Since a vector orthogonal to the column space is in the left null space, we can write

$$\begin{aligned} A^T(b - Ax) &= 0 \\ A^T b - A^T Ax &= 0 \\ A^T b &= A^T Ax \\ (A^T A)^{-1} A^T b &= x \\ A(A^T A)^{-1} A^T b &= Ax \\ Pb &= Ax \end{aligned}$$

so P is a projection onto $Col(A)$.

Example. Show that P is symmetric and idempotent.

$$P^T = \left(A(A^T A)^{-1} A^T\right)^T = A(A^T A)^{-1} A^T = P$$

so P is symmetric.

$$P^2 = PP = A(A^T A)^{-1} A^T A(A^T A)^{-1} A^T = A(A^T A)^{-1} A^T = P$$

so P is idempotent.

Example. Show that $Rank(P) = Tr(P) = m$.
We first need to show that a projection matrix only has 0 or 1 eigenvalues. Let v be a eigenvector of P corresponding to eigenvalue λ. Since P is idempotent,

$$\lambda v = Av = AAv = \lambda Av = \lambda^2 v$$

Noting that $v \neq 0$, $\lambda^2 = \lambda$ so λ can only be 0 or 1.

Let $P = Q\Lambda Q^T$ be the spectral decomposition of P. Then we have

$$Tr(P) = Tr(Q\Lambda Q^T) = Tr(\Lambda Q^T Q) = Tr(\Lambda) = \sum_{i=1}^{n} \lambda_i$$

By the Rank-Nullity theorem, we know that $Rank(P) + Nullity(P) = n$, the number of columns in P. The nullity of P is simply the dimension of the null space P, which is equal to the number of zero eigenvalues (recall that the algebraic and geometric multiplicity of $\lambda = 0$ are the same since P is square and symmetric). Conversely, we have that $Rank(P) = n - Nullity(P)$, which is the number of non-zero eigenvalues since P has n real eigenvalues. Since the number of nonzero eigenvalues of P is precisely $Tr(\Lambda)$ because all the eigenvalues are either 0 or 1, we have that

$$Rank(P) = Tr(\Lambda) = Tr(P)$$

Finally, we have that

$$Rank(P) = Tr(P) = Tr(A(A^T A)^{-1} A^T) = Tr(I_m) = m$$