to be determined. Adaptively chosen basis functions are known as dictionary methods, where one has available possibly infinite set or dictionary $\mathcal{D}$ of candidate basis functions to choose from. The models are then built and is equivalent to the search problem.

5.7 Model Selection, Bias-Variance Tradeoff

The models described in general have either a smoothing or complexity parameter that determines the restriction of the model, such as the (i) λ penalty scalar, (ii) kernel width and (iii) number of basis functions.

Consider the KNN case again, with $Y = f(X) + \epsilon, \mathbb{E}[\epsilon] = 0, \text{Var}[\epsilon] = \sigma^2$. Assume for simplicity the values of x_i in sample are fixed and nonrandom. Then the expected prediction error at x_0, known as the *test/generalization* error can be decomposed into

$$\begin{aligned} EPE_k(x_0) &= \mathbb{E}[(Y - \hat{f}_k(x_0))^2 | X = x_0] && (5.27) \\ &= \sigma^2 + \text{Bias}^2(\hat{f}_k(x_0)) + \text{Var}_{\mathcal{T}}(\hat{f}_k(x_0)) && (5.28) \\ &= \sigma^2 + [f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})]^2 + \frac{\sigma^2}{k}. && (5.29) \end{aligned}$$

Note the parenthesis indicates the order statistic for the KNN selection. The first term, σ^2 is known as the irreducible error, and is present even if we knew the true f. We are able to control the other two terms, which constitute the *mean squared error* of $\hat{f}_k(x_0)$ in estimation of $f(x_0)$, where we have the bias and variance decomposition.

Definition 36 (Bias of Estimate)**.** *The bias is the difference between the true mean value of a random variable and the expectation of our estimate. In the MSE decomposition of the generalization error, this can be formulated* $\mathbb{E}_{\mathcal{T}}[\hat{f}_k(x_0) - f(x_0)]$*, where* $f(x_0)$ *is the non-random conditional mean at* x_0*.*

If the true function is reasonably smooth, the bias term likely increases with K in the nearest neighbour method.

The variance term on the other hand, decreases as the *inverse of* K in nearest neighbours. In the general function approximation paradigm, as the model complexity of our procedure increases, the variance tends to increase and (squared) bias tends to decrease. We want to choose the model complexity so that there is an attractive tradeoff between the bias and variance, so that the *test error* is minimized. Although we can estimate test error with the *training error* $(1/N)\sum_i (y_i - \hat{y}_i)$, this is often not a good estimate as it does not account for model complexity. The generalization issues and bias-variance tradeoff leads to the classical U-shaped test error in relation to model complexity.

5.8 Least Squares Methods

5.8.1 Simple Least Squares

The simple linear regression model assumes data follows relationship

$$y = \beta_0 + \beta_1 x + \epsilon \tag{5.30}$$

, where ϵ is the random error assumed $\mathbb{E}\epsilon = 0, \text{Var}\epsilon = \sigma^2$ (constant). The conditional mean response at x shall then be $\mathbb{E}(Y|X = x) = \mu_{y|x} = \beta_0 + \beta_1 x$, and conditional variance at x equivalent to $\text{Var}(Y|X =$

$x) = \sigma^2_{y|x} = \text{Var}(\beta_0 + \beta_1 x + \epsilon) = \sigma^2$. Note then how the variance is assumed constant in this model - we shall see how *heteroscedasticity* affects our model interpretations later on. Our least squares problem reduces to the problem of finding coefficient estimates and ensuring their adequacy. For samples taken from the population regression model, we assume $i \neq j \implies \epsilon_i \perp \epsilon_j$, that errors (and hence responses) are uncorrelated. In the simple linear regression equation (Equation 5.30), the β_1 slope specifies the change in the mean distribution of y under unit change in x.

5.8.1.1 Assumptions of the Simple Linear Equation

1. $\forall i, j, i \neq j \implies \epsilon_i \perp \epsilon_j$
2. $\sigma^2 = \text{Var}(\epsilon) = \text{Var}(\epsilon | X = x) = k \in \mathbb{R}$.

5.8.1.2 Model Fitting

The hypothesis space of our simple linear regression problem is the set of all candidate lines specified by $\{(\beta_0, \beta_1) : (\beta_0, \beta_1) \in \mathbb{R}^2\}$. From the sample regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i \in [n]$, we choose the model that minimises the sum of squared errors $S(\beta_0, \beta_1) = \sum_i \epsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$. We can easily derive by standard calculus the estimators $\hat{\beta}_0, \hat{\beta}_1$ by solving the linear equations

$$\frac{\delta S}{\delta \beta_0} = -2\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \overset{!}{=} 0 \tag{5.31}$$

$$\frac{\delta S}{\delta \beta_1} = -2\sum_i x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \overset{!}{=} 0 \qquad \text{with solutions} \tag{5.32}$$

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i = \sum_i y_i \tag{5.33}$$

$$\hat{\beta}_0 \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2 = \sum_i y_i x_i \tag{5.34}$$

From Equation 5.33 we have $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, and substituting into Equation 5.34 we obtain

$$(\bar{y} - \hat{\beta}_1 \bar{x}) \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2 = \sum_i y_i x_i, \tag{5.35}$$

$$\hat{\beta}_1 (\sum_i x_i^2 - \bar{x} \sum x_i) = \sum_i y_i x_i - \bar{y} \sum_i x_i \tag{5.36}$$

This gives solution

$$\hat{\beta}_1 = \frac{\sum_i y_i x_i - \frac{\sum_i y_i \sum_i x_i}{n}}{\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}} \tag{5.37}$$

$$= \frac{S_{xy}}{S_{xx}} \tag{5.38}$$

where

$$S_{xx} = \left(\sum_i x_i^2\right) - \frac{(\sum_i x_i)^2}{n} = \left(\sum_i x_i^2\right) - \frac{(n\bar{x})^2}{n} \tag{5.39}$$

$$= \left(\sum_i x_i^2\right) - n\bar{x}^2 = \left(\sum_i x_i^2\right) + n\bar{x}^2 - 2n\bar{x}^2 \tag{5.40}$$

$$= \sum_i (x_i^2 + \bar{x}^2 - 2x_i\bar{x}) = \sum_i (x_i - \bar{x})^2 \tag{5.41}$$

$$= (n-1)s^2 \tag{5.42}$$

and

$$\begin{aligned} S_{xy} &= \sum y_i x_i - \frac{\sum_i y_i \sum_i x_i}{n} = \left(\sum_i y_i x_i\right) - \frac{n\bar{y}n\bar{x}}{n} = \left(\sum_i y_i x_i\right) - n\bar{y}\bar{x} && (5.43) \\ &= \sum_i y_i(x_i - \bar{x}) && (5.44) \end{aligned}$$

and we obtain fitted model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and residuals on sample $i = e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. The least squares can be computed by the formula:

```
numpy.linalg.lstsq(a, b, rcond='warn')
```

that computes the vector x that approximately solves the equation a @ x = b.

5.8.1.3 Model Properties and Variance of Estimates

Recalling that $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i y_i(x-\bar{x})}{S_{xx}}$, we may express in form $\sum_i c_i y_i$ where $c_i = \frac{(x_i-\bar{x})}{S_{xx}}$. We see easily that $\sum_i c_i = 0$, and $\sum_i c_i x_i = \sum_i x_i \frac{x_i-\bar{x}}{\sum_i (x_i-\bar{x})^2} = \frac{\sum_i (x_i^2 - x_i\bar{x})}{\sum_i (x_i^2+\bar{x}^2-2x_i\bar{x})} = \frac{(\sum_i x_i^2)-n\bar{x}^2}{(\sum_i x_i^2)+n\bar{x}^2-2n\bar{x}^2} = 1$. Finally, $\sum_i c_i^2 = \sum \frac{(x_i-\bar{x})^2}{S_{xx}^2} = \frac{1}{S_{xx}}$. This is, the slope estimator $\hat{\beta}_1$ is a linear combinations of the observations y_i.

The least squares estimates are unbiased, in that $\mathbb{E}\hat{\beta}_1 = \beta_1, \mathbb{E}\hat{\beta}_0 = \beta_0$. The variance $\text{Var}(\hat{\beta}_1) = \text{Var}(\sum_{i=1}^n c_i y_i) = \sum_i c_i^2 \text{Var}(y_i) = \frac{\sigma^2}{S_{xx}}$, since we showed $\sum_i c_i^2 = \frac{1}{S_{xx}}$ and we assumed the errors (and accordingly response) are uncorrelated. The intercept variance follows $\text{Var}(\hat{\beta}_0) = \text{Var}(\bar{y}-\hat{\beta}_1\bar{x}) = \text{Var}(\bar{y}) + \bar{x}^2\text{Var}(\hat{\beta}_1) - 2\bar{x}Cov(\bar{y}, \hat{\beta}_1) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$ where the covariance term drops off since $Cov(\frac{1}{n}y_i, \sum_i c_i y_i) = \frac{1}{n}\sum_i c_i \text{Var}(y_i)$ and $\sum c_i = 0$.

The ordinary least squares estimators $(\hat{\beta}_i), i \in [p+1]$ are the best linear, unbiased estimators, in that it has the smallest variance compared to the other unbiased estimators formed from the linear combinations of y_i. Some useful results we arrive from the simple least squares method is that $\sum_i (y_i - \hat{y}_i) = \sum_i e_i = 0$, which implies $\sum y_i = \sum \hat{y}_i$. Not only are the sum of residuals zero, the regressor weighted errors and fitted value weighted errors are zero. such that $\sum_i x_i e_i = 0, \sum_i \hat{y}_i e_i = 0$. The fitted line always passes through $(\bar{x}, \bar{y})$.

We may obtain a point estimate of conditional variance of y given x using the residual sum of squares $SS_{res} = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2$. The residual sum of squares is a component of the total sum of squares, which is relevant to the unconditional variance. Writing sum of total squares $SS_T = \sum_i (y_i - \bar{y})^2$ and noting $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, then

$$\begin{aligned} \sum_i (y_i - \bar{y})^2 &= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 && (5.45) \\ &= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i [(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})] && (5.46) \\ &= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i \hat{y}_i(y_i - \hat{y}_i) - 2\sum_i \bar{y}(y_i - \hat{y}_i) && (5.47) \\ &= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i \hat{y}_i e_i - 2\sum_i \bar{y} e_i && (5.48) \\ &= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i \hat{y}_i e_i - 2\bar{y}\sum_i e_i && (5.49) \\ &= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 && (5.50) \\ &= SS_{res} + \sum_i (\hat{y}_i - \bar{y})^2 && (5.51) \end{aligned}$$

we see that the total unconditional dispersion can be broken down into the components explained by the model and unexplained by our model. That is $SS_T = SS_{res} + SS_{reg}$, where SS_{reg} is the regression sum of squares defined $\sum_i (\hat{y}_i - \bar{y})^2$.

We may show (verify this) that $\mathbb{E}SS_{res} = (n-2)\sigma^2$. An unbiased estimator $\hat{\sigma}^2$ for σ^2 is $\frac{SS_{res}}{n-2}$.

Definition 37 (Residual Mean Square)**.** *The value $\frac{SS_{res}}{n-2}$ is called the residual mean square and is an unbiased estimator of σ^2, the conditional variance of response at a given input. The value $\hat{\sigma} = \sqrt{MS_{res}}$ shall be called the residual standard error, or equivalently, the standard error of regression. $n-2$ indicates the degrees of freedom in residual sum of squares (SS_{res}), attributed to the loss of freedom from estimating $\hat{\beta}_0, \hat{\beta}_1$. The estimate $\hat{\sigma}$ depends on SS_{res}, and requires that model assumptions of independent errors and constant variance be satisfied.*

5.8.1.4 Assumptions of the Analysis of Model on Simple Linear Equations

1. The assumptions of the model also apply, for obvious reasons, as assumptions in the analysis of model. That is, we assume uncorrelated errors with constant variance and mean zero.

2. Additionally, we assume that $\epsilon_i \sim \Phi(0, \sigma^2)$. In fact, they are *identical and independent* normal random variables, implying that (i) for each value/level of regressor variable, the sub-population of responses follow normal distribution and (ii) each such sub-populations share constant variance σ^2.

5.8.1.5 Test of Significance on Regression Coefficients

We may perform hypothesis testing for the significance of regression coefficients, for instance under settings as follows:

$$H_0 : \beta_1 = \beta_{1_0} \qquad H_1 : \beta_1 \neq \beta_{1_0}$$

with test statistic

$$Z_o = \frac{\hat{\beta}_1 - \beta_{1_0}}{sd(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_{1_0}}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1).$$

Since $\hat{\beta}_1$ is a linear combination of y_i and $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, $\hat{\beta}_1$ must follow normal distribution. However, most often σ^2 is unknown and we use the estimator $\hat{\sigma}^2$, using test statistic

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1_0}}{\sqrt{\frac{MS_{res}}{S_{xx}}}} \sim t_{n-1},$$

which rejects the null hypothesis in a two-sided test under conditions $|t_o| > t_{n-2}(\alpha/2)$, where $t_{n-2}(\alpha/2)$ indicates the percentile point of a t-distribution of degrees of freedom $(n-2)$ with $\frac{\alpha}{2}$ right-tail probability. It is obvious from the test statistic that $SE(\hat{\beta}_1) = \sqrt{\frac{MS_{res}}{S_{xx}}}$. For test of intercept, our equivalent (and abbreviated) steps would follow $H_0 : \beta_0 = \beta_{0_0}$, $SE(\hat{\beta}_0) = \sqrt{MS_{res}(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}$ with test statistic following (n-2) degrees of freedom t-distributions. The $H_0 : \hat{\beta}_1 = 0$ implies there is no linear relationship between y and x supported by the data, and the rejection implies that x helps explain variability of the response.

5.8.1.6 Test of Significance on Regression Model and ANOVA Methods

We may test for the significance of the regression model by testing if any of the β coefficients are unlikely to be zero. In the case of the simple linear model, this turns out to be equivalent to the t-test on regression coefficients, since we only have one. We re-iterate here in short: $H_0 : \beta_1 = 0$, $t_0 = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$. The more general method is known as the Analysis of Variance (ANOVA) method. In Equation 5.45 we demonstrated that the sum of total squares may be decomposed into the sum of squared residuals and the regression/model sum of squares. Re-iterating:

$$SS_T = \sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y}_i)^2 + \sum_i (y_i - \hat{y}_i)^2 = SS_{reg} + SS_{res}.$$

Since SS_T has the constraint that $\sum_i (y_i - \bar{y}) = 0$, it has a degree of freedom of $(n-1)$. SS_{res} has degrees of freedom $(n-2)$, and since $SS_{reg} = \hat{\beta}_1 \cdot S_{xy}$ (verify this) is determined once $\hat{\beta}_1$ is decided, it has degrees of freedom one. Both sides match.

To test the hypothesis $H_0 : \beta_1 = 0$, we arrive at the following conclusions, conditional on null:

$$\frac{SS_{res}}{\sigma^2} = (n-2)\frac{MS_{res}}{\sigma^2} \sim \chi^2_{n-2} \tag{5.52}$$

$$\frac{SS_{reg}}{\sigma^2} = \chi^2_1 \tag{5.53}$$

$$SS_{res} \perp SS_{reg}. \tag{5.54}$$

Under H_0, this amounts to the test statistic

$$F_0 = \frac{SS_{reg}/1}{SS_{res}/(n-2)} \sim F_{1,(n-2)},$$

rejecting the null hypothesis when $F_0 > F_{1,(n-2)}(\alpha)$. We often call the terms $SS_{reg}/1 = MS_{reg}$ the regression mean square and $SS_{res}/(n-2) = MS_{res}$ the residual mean square. It can be shown the t-test and F-test in the simple linear model are identical, since

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{MS_{res}/S_{xx}}}$$

$$t_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MS_{res}} = \frac{\hat{\beta}_1 S_{xy}}{MS_{res}} = F_0.$$

The ANOVA tables may be generated in code using the following:

```
from statsmodels.formula.api import ols
model = ols('y ~ x', data=data).fit()
anova = statsmodels.stats.anova_lm(model, typ=2)
```

5.8.1.7 Confidence Intervals on Parameters and Variance Estimates

Assuming the same assumptions for the hypothesis (see Section 5.8.1.4) are satisfied, we may derive confidence intervals on the β coefficients. Assuming the IID errors, the sampling distribution of $\beta_i, i \in \{0,1\} = \frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)} \sim t_{n-2}$. The discussion on the standard errors of these estimates were employed in Section 5.8.1.5, and their $100(1-\alpha)\%$ confidence intervals are constructed $\beta_i \in \left[\hat{\beta}_i \pm t_{n-2}(\frac{\alpha}{2}) \cdot SE(\hat{\beta}_i)\right]$. The confidence interval for σ^2 corresponds to the interval $\left[\frac{(n-2)MS_{res}}{\chi^2_{n-2}(\frac{\alpha}{2})}, \frac{(n-2)MS_{res}}{\chi^2_{n-2}(1-\frac{\alpha}{2})}\right]$.

5.8.1.8 Confidence Intervals and Prediction Intervals on Response

Confidence Intervals

The regression function $\mathbb{E}(Y|X)$ gives point estimates for the conditional mean on response, given inputs regressor variables. The point estimator for $\mathbb{E}(y|x_0) = \widehat{\mathbb{E}(y|x_0)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is an unbiased estimator, and we can derive its variance. Recall that $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, then $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and

$$\text{Var}(\hat{\mu}_{y|x_0}) = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \text{Var}(\bar{y} + \hat{\beta}_1(x_0 - \bar{x})) = \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{S_{xx}}.$$

We therefore arrive at a confidence interval for conditional mean responses, with the following test statistic:

$$\frac{\hat{\mu}_{y|x_0} - \mathbb{E}(y|x_0)}{\sqrt{MS_{res} \cdot \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2} \tag{5.55}$$

and confidence intervals at $100(1-\alpha)\%$:

$$\hat{\mu}_{y|x_0} \pm t_{n-2,\frac{\alpha}{2}} \sqrt{MS_{res} \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \tag{5.56}$$

which has minimal interval width at $x_0 = \bar{x}$ and widens with $|x_0 - \bar{x}|$. *The standard error in the computation takes into account the sampling error. We interpret that if we keep repeating the sampling N times, then approximately $(1-\alpha)\%$ of the N confidence intervals constructed contain the true sub-mean.*

Prediction Intervals

Another application is the prediction of a new observation y given some specified $x = x_0$. The point estimate for this response is the same as the point estimate for the conditional mean response. In this part we are interested in making statistical conclusions about $\hat{y}_0$ instead of $\widehat{\mathbb{E}(y|x_0)}$. Consider the random variable $\psi = y_0 - \hat{y}_0$, which takes normal distribution $\Phi(0, \text{Var}(\psi))$ and $\text{Var}(\psi) = \text{Var}(y_0 - \hat{y}_0) = \text{Var}(y_0) + \text{Var}(\hat{y}_0) - 2Cov(y_0, \hat{y}_0) = \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$, since future observations are necessary independent of $\hat{y}_0$. This results in the prediction interval

$$\hat{y}_0 \pm t_{n-2,\frac{\alpha}{2}} \sqrt{MS_{res} \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}, \tag{5.57}$$

which takes into account both the sampling error but also the variability of individuals around the predicted conditional mean. It follows that the prediction interval (also a confidence interval, technically) is a superset of the confidence interval relating to the conditional mean.

5.8.1.9 No Intercept Models

The no-intercept model specified $y = \beta_1 x + \epsilon$ can be specified where the origin intersection shall be included as part of the problem specification. The estimators take different forms, although the principle is the same. We have $S(\beta_1) = \sum_i (y_i - \beta_1 x_i)^2$, least-squares equation $\hat{\beta}_1 \sum_i x_i^2 = \sum y_i x_i$, giving unbiased estimator

$$\hat{\beta}_1 = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

and fitted regression model $\hat{y} = \hat{\beta}_1 x$. We have an estimator $\hat{\sigma}^2$ for conditional variance $\hat{\sigma}^2 = MS_{res} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-1} = \frac{(\sum_i y_i^2) - \hat{\beta}_1 \sum_i y_i x_i}{n-1}$.

5.8.1.10 Coefficient of Determination, R^2

The quantity $R^2 = \frac{SS_{reg}}{SS_T} = 1 - \frac{SS_{res}}{SS_T}$ is known as the coefficient of determination, measuring the proportion of variability in response explained by the regressors and our model. We have $0 \leq SS_{res} \leq SS_T \implies R^2 \in [0,1]$ - but note that adding more terms monotonically increases the coefficient of determination. This does not indicate the appropriateness or complexity of our model!

5.8.1.11 Maximum Likelihood Estimators vs Simple Least Squares

It can be shown that the method of least squares in parameter estimation is identical when errors are assumed normal. Referring to the maximum likelihood method (see Definition 10), we have $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ when errors are assumed IID $N(0, \sigma^2)$. Using the Gaussian log-likelihood function in Equation (3.79):

$$\begin{aligned} L(\theta) &= -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - f_\theta(x_i))^2 && (5.58) \\ &= -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2, && (5.59) \end{aligned}$$

which by taking derivatives results in the linear equations

$$\begin{aligned} \frac{\delta \log L}{\delta \beta_0} &= \frac{1}{\tilde{\sigma}^2}\sum_i \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right) \overset{!}{=} 0 && (5.60) \\ \frac{\delta \log L}{\delta \beta_1} &= \frac{1}{\tilde{\sigma}^2}\sum_i \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right) x_i \overset{!}{=} 0 && (5.61) \\ \frac{\delta \log L}{\delta \sigma^2} &= -\frac{n}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4}\sum_i \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right)^2 \overset{!}{=} 0 && (5.62) \\ & && (5.63) \end{aligned}$$

and solutions

$$\begin{aligned} \tilde{\beta}_0 &= \tilde{y} - \tilde{\beta}_1 \bar{x} && (5.64) \\ \tilde{\beta}_1 &= \frac{\sum_i y_i(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} && (5.65) \\ \tilde{\sigma}^2 &= \frac{\sum_i (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{n} && (5.66) \end{aligned}$$

which give unbiased estimators of β_0, β_1 and biased estimator of σ. The relationship betweeen the least squares estimator and MLE estimator follows the equation

$$\tilde{\sigma}^2 = \left[\frac{n-2}{n}\right]\hat{\sigma}^2,$$

and we see that the bias is small when n large. In general, the MLE estimators have better statistical properties than the least-squares estimators, exhibiting consistency (see 9), asymptotic efficiency (see 12) and sufficiency (see 14). The downside of MLE is that it requires distributional assumptions about the errors - that they are IID $\sim \Phi(0, \sigma^2)$.

5.8.2 Multiple Least Squares

We may generalize the simple least squares model to multiple regressors, say k of them. Then our population regression model gives

$$y = \beta_0 + \left(\sum_i^k \beta_i x_i\right) + \epsilon \qquad \beta_i, i \in [k].$$

The $\beta_i, i \in [k]$ are known as regression coefficients and reflect the expected change in response given unit change in corresponding regressor, ceteris paribus. We note that 'linearity' is in coefficients, and the model may take general forms such as higher order polynomials.

5.8.2.1 Interaction Effects

The regressors may contain interaction effects, which are defined as models that contain a function on two 'atomic' regressors, which should already be included in the model. An example of such formulaic relationships could take form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$, and we shall note that in this model the level of change expected in y given unit change in x_1 depends on the level of x_2! Again, here we have a non-linear surface generated by the regression equation, but we maintain linearity in the regression coefficients and all conclusions apply.

5.8.2.2 Assumptions and Model Notations

Let n denote sample size as usual, and k take number of regressors. $y_i, i \in [n]$ is the i-th response, with x_{ij} taking the i-th observation of the regressor x_j.

We assume that $\mathbb{E}(\epsilon) = 0, \mathrm{Var}(\epsilon) = \sigma^2$ - particularly that errors are centered at zero and have constant variance. The sample regression model takes form $y_i = \beta_0 + \sum_j^k \beta_j x_{ij} + \epsilon_i, i \in [n]$.

5.8.2.3 Model Fitting

Taking least squares function $\sum_i \epsilon_i^2 = \sum_i^n (y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij})^2$ and computing minima, we obtain the linear equations

$$\frac{\delta S}{\delta \beta_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij}) \overset{!}{=} 0. \tag{5.67}$$

$$\frac{\delta S}{\delta \beta_{j \in [k]}} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij}) x_{ij} \overset{!}{=} 0. \tag{5.68}$$

resulting in $p = k+1$ normal equations. Encoding the response vector in $\mathbf{y} \in \mathbb{R}^{n \cdot 1}$, regressors in $\mathbf{X} \in \mathbb{R}^{n \cdot p}$ and $\beta \in \mathbb{R}^{p \cdot 1}$, $\epsilon \in \mathbb{R}^{n \cdot 1}$, we may rewrite the sample regression model in matrix form $\mathbf{y} = \mathbf{X}\beta + \epsilon$. Note that the matrix $\mathbf{X}$ is the matrix of $\mathbf{1}$'s in the leftmost column and n rows of regressor data row-stacked. We can equivalently express

$$\begin{aligned}
\mathbf{S}(\beta) = \sum_i \epsilon_i^2 &= \epsilon^T \epsilon = (\mathbf{y} - \mathbf{X}\beta)^{\mathbf{T}} (\mathbf{y} - \mathbf{X}\beta) && (5.69)\\
&= (\mathbf{y}^{\mathbf{T}} - \beta^{\mathbf{T}} \mathbf{X}^{\mathbf{T}})(\mathbf{y} - \mathbf{X}\beta) && (5.70)\\
&= \mathbf{y}^{\mathbf{T}}\mathbf{y} - \mathbf{y}^{\mathbf{T}}\mathbf{X}\beta - \beta^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{y} + \beta^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\beta && (5.71)\\
&= \mathbf{y}^{\mathbf{T}}\mathbf{y} + \beta^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\beta - \mathbf{2}\beta^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{y} && (5.72)
\end{aligned}$$

where in the last step we use the fact that transpose on scalars are identity transforms. By matrix calculus (see Section 2.1) we arrive at

$$\frac{\delta S}{\delta \beta} = (\mathbf{X^T X}\beta)^{\mathbf{T}} + (\beta^{\mathbf{T}}\mathbf{X^T X}) - \mathbf{2}(\mathbf{X^T y})^{\mathbf{T}} = \mathbf{2}(\beta^{\mathbf{T}}\mathbf{X^T X}) - \mathbf{2y^T X}, \tag{5.73}$$

giving solutions $\mathbf{X^T X}\beta = \mathbf{X^T y}$ and regression coefficient

$$\hat{\beta} = (\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T y} \tag{5.74}$$

. This gives solution if $\exists(\mathbf{X^T X})^{-\mathbf{1}}$. This exists if the regressors (columns) are linearly independent and our fitted regression model is written $\hat{y} = x\hat{\beta}$, where x is row vector of $[1, x_1, x_2 \cdots x_k]$. Substituting regression coefficient estimates (Equation 5.74), we obtain

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T y} = \mathbf{Hy} \tag{5.75}$$

where

Definition 38 (Hat Matrix)**.** $\mathbf{H} = \mathbf{X}(\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T} \in \mathbb{R}^{n,n}$ *is known as the hat matrix, for putting a 'hat' on the response vector.*

Corollary 7 (Residuals in Terms of Hat Matrix)**.** *Defining the residual terms $e_i = y_i - \hat{y}_i$, the n residuals may be written in matrix form:*

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - \mathbf{Hy} = (\mathbb{1} - \mathbf{H})\mathbf{y}$$

Note that the hat matrix H is symmetric and idempotent (that is $\mathbf{HH} = \mathbf{H}$), and is also often called the *projection matrix* (projections have the property $P^2 = P$). The matrix $(\mathbb{1} - \mathbf{H})$ is also symmetric and idempotent.

5.8.2.4 Model Properties and Variance of Estimates

Note that

$$\begin{aligned}
\mathbb{E}\hat{\beta} &= \mathbb{E}\left[(\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T y}\right] && (5.76)\\
&= \mathbb{E}\left[(\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T}(\mathbf{X}\beta + \epsilon)\right] && (5.77)\\
&= \mathbb{E}\left[(\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T X}\beta + (\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T}\epsilon)\right] && (5.78)\\
&= \mathbb{E}\left[\mathbb{1}\beta + (\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T}\epsilon)\right] && (5.79)\\
&= \beta && (5.80)
\end{aligned}$$

where the last two steps used the assumption of uncorrelated errors to arrive at unbiased regression coefficient estimates. Furthermore, we have

$$Cov(\hat{\beta}) = \mathbb{E}\left[(\hat{\beta} - \mathbb{E}\hat{\beta})(\hat{\beta} - \mathbb{E}\hat{\beta})^T\right], \tag{5.81}$$

which is a positive and symmetric semi-definite matrix in $\mathbb{R}^{p,p}$. The diagonals are the variance of the coefficient estimates. We have

$$\begin{aligned}
Cov(\hat{\beta}) &= Cov((\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T y}) && (5.82)\\
&= \left((\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T}\right) Cov(y) \left((\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T}\right)^T && (5.83)\\
&= \sigma^2 \left((\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T}\right) \left((\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T}\right)^T && (5.84)\\
&= \sigma^2(\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T X}(\mathbf{X^T X})^{-\mathbf{1}} && (5.85)\\
&= \sigma^2\mathbb{1}(\mathbf{X^T X})^{-\mathbf{1}} && (5.86)\\
&= \sigma^2(\mathbf{X^T X})^{-\mathbf{1}} && (5.87)
\end{aligned}$$

Writing $\mathbf{C} = (\mathbf{X^T X})^{-1}$, and the $Cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}$. Similar to the simple least squares approach, in the estimation of p parameters we arrive at SS_{res} of $(n-p)$ degrees of freedom. As before, we obtain the definition

Definition 39 (Residual Mean Square)**.** *The residual mean square is defined*

$$MS_{res} = \frac{SS_{res}}{n-p} \tag{5.88}$$

is unbiased estimator of σ^2*.*

which is model dependent.

5.8.2.5 Assumptions for Analysis of Multiple Least Squares Regression

In addition to the assumptions specified for the model in Section (5.8.2.2), we need to assume here that the random errors

$$\epsilon_i \overset{IID}{\sim} \Phi(0, \sigma^2)$$

5.8.2.6 Significance Tests for Regression Coefficients by t-tests and Partial Sum of Squares Method

We may test separately the contribution of a particular regressor in explaining the variability of the response, or any subset of them.

T-Test

To test the significance of any individual regression coefficient β_j we perform test (abbreviated) $H_0 : \beta_j = 0$, $t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{n-p}$, where C_{jj} is the j-th diagonal of $(\mathbf{X^T X})^{-1}$. This test is called the *partial/marginal* test of significance of $\hat{\beta}_j$ given all other regressors already present in the model.

Partial/Extra Sum-of-Squares Method

To determine the contribution to SS_{reg} of a particular set of regressors given other regressors already in the model, consider the following:

Definition 40 (Partial Sum-of-Squares Method)**.** *Consider again the population regression model written* $\mathbf{y} = \mathbf{X}\beta + \epsilon$ *where the* β *is re-arranged to form the block matrix*

$$\beta = [\beta_\mathbf{1} \beta_\mathbf{2}]^\mathbf{T},$$

where $\beta_\mathbf{1} \in \mathbb{R}^{p-r,1}$ *and* $\beta_\mathbf{2} \in \mathbb{R}^{r,1}$ *for* $r \leq p$ *partitions the regression coefficients of interest. Then we can rewrite such that* $\mathbf{y} = \mathbf{X_1}\beta_\mathbf{1} + \mathbf{X_2}\beta_\mathbf{2} + \epsilon$ *and we are interested in* $H_0 : \beta_\mathbf{2} = 0$ *vs* $H_1 : \beta_\mathbf{2} \neq 0$*. Now recall that in the full model we have* $\hat{\beta} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$*, and we define their regression sum of squares* $SS_{reg}(\beta)$*. Under the reduced model of the null hypothesis we are left with the regression equation* $\mathbf{y} = \mathbf{X_1}\beta_\mathbf{1} + \epsilon$*, and we can write their regression coefficient estimates as* $\hat{\beta}_\mathbf{1} = (\mathbf{X_1^T X_1})^{-1}\mathbf{X_1^T y}$ *with regression sum of squares* $SS_{reg}(\beta_1)$*. Then we define the partial sum of squares of* $\beta_\mathbf{2}$ *to be*

$$SS_{reg}(\beta_2|\beta_1) = SS_{reg}(\beta) - SS_{reg}(\beta_1),$$

and this has the number of degrees of freedom r*. Note that* $SS_{reg}(\beta_2|\beta_1) \perp MS_{res}$ *and for the null hypothesis* $H_0 : \beta_2 = 0$ *we construct the test statistic*

$$F_0 = \frac{SS_{reg}(\beta_2|\beta_1)/r}{MS_{res}} \sim F_{r,n-p}$$

under the null hypothesis, which rejects the null hypothesis if it turns out that $F_0 > F_{r,n-p}(\alpha)$*, the implication of which states that* $\exists j \in [k-r+1, k]$ *s.t* x_j *is a statistically significant regressor.*

Note that the result of the partial sum of squares in the case $r = 1$ yields the same conclusion and p-values as would a t-test. Also, if the columns in $\mathbf{X_1}$ are orthogonal to the columns in $\mathbf{X_2}$, then we must have $SS_{reg}(\beta_2) = SS_{reg}(\beta_2|\beta_1)$. When we do significance tests and decide to remove a variable, *in general* the regression coefficients of the refitted model are not the same. In the special case of orthogonal sets, no refitting is required.

5.8.2.7 Confidence Interval for Regression Coefficient Estimates

Recall in Section 5.82, we demonstrated that $Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$. Furthermore, since we know that $\hat{\beta}$ is just a linear combination of the observations, and that we have the assumptions that $\epsilon_i \overset{IID}{\sim} \Phi(0, \sigma^2)$ (and therefore that $y_i \sim \Phi(\beta_0 + \sum_j \beta_j x_{ij}, \sigma^2)$), we can then conclude

$$\hat{\beta} \sim \Phi(\beta, \sigma^2(X^TX)^{-1}) \tag{5.89}$$

where the *marginal distribution of a regression coefficient* $\hat{\beta}_{j\in[p]}$ is $\sim \Phi(\beta_j, \sigma^2 C_{jj})$ where C_{jj} is j-th diagonal of $(X^TX)^{-1}$. The (abbreviated) hypothesis test would have test statistic of form

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{n-p}, \qquad j \in [k]$$

and confidence interval

$$\hat{\beta}_j \pm t_{n-p}(\frac{\alpha}{2})\sqrt{\hat{\sigma}^2 C_{jj}}$$

with $SE(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$.
Joint Confidence Intervals on Regression Coefficients If we have a least squares model with predictions on $(\hat{\beta}_0, \hat{\beta}_1)$, we each have a $100(1-\alpha)\%$ confidence interval on the coefficients. Say we want a 95% confidence interval on both parameters, and assuming independence (they are not), then their 'joint correctness' is only 0.95^2. We need to define their joint distributions. It can be shown that (verify):

$$F_0 = \frac{(\hat{\beta}^{\mathbf{T}} - \beta)^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}(\hat{\beta}^{\mathbf{T}} - \beta)}{pMS_{res}} \sim F_{p,n-p} \tag{5.90}$$

which implies $\mathbb{P}\{F_0 \leq F_{p,n-p}(\alpha)\} = 1 - \alpha$. Constructing a $100(1-\alpha)\%$ joint confidence region for all parameters in β we obtain the values for which it results in test statistic $F_0 \leq F_{p,n-p}(\alpha)$.
Bonferroni Correction: we may also choose the family-wise error rate such that we instead use confidence intervals:

$$\hat{\beta}_j \pm t_{\frac{\alpha}{2p},n-p} \cdot SE(\hat{\beta}_j), \qquad j \in [k]. \tag{5.91}$$

5.8.2.8 Confidence Interval and Prediction Intervals of Estimates on Mean Response and Response Variables

We may want to make certain point estimates and give statistical comments about the relevancy of our estimates on conditional mean response, or just response.
Confidence for Prediction of Conditional Mean Response For some point $\mathbf{x_0}$ we want to construct (conditional) mean response intervals. Recall from Equation 5.82 $Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$. Further suppose we select the estimator $\hat{\mathbf{y}}_\mathbf{0} = \mathbf{x_0^T}\hat{\beta}$, which turns out to be an unbiased estimator of $\mathbb{E}(y|X = x_0)$. The variance of shall be computed

$$\begin{aligned} \text{Var}(\hat{y}_0) &= \text{Var}(x_0^T\hat{\beta}) \qquad (5.92)\\ &= x_0^T Cov(\hat{\beta})x_0 \qquad (5.93)\\ &= \sigma^2 x_0^T(X^TX)^{-1}x_0. \qquad (5.94) \end{aligned}$$

We can therefore define the confidence interval for mean response given:

$$\hat{y}_0 + t_{n-p}(\frac{\alpha}{2})\sqrt{\hat{\sigma}^2\mathbf{x_0^T}(\mathbf{X^TX})^{-1}\mathbf{x_0}} \tag{5.95}$$

Prediction Interval for Prediction of Response

For a particular point $\mathbf{x_0}$, we are interested in making the $100(1-\alpha)\%$ prediction interval for response given

$$\hat{y}_0 \pm t_{n-p}(\frac{\alpha}{2})\sqrt{\hat{\sigma}^2(1+\mathbf{x_0^T}(\mathbf{X^TX})^{-1}\mathbf{x_0})}. \tag{5.96}$$

5.8.2.9 Significance Tests for Regression Model

To test if there is a linear relationship between our response variable and any regressors (if our model is significant at all), we may construct hypothesis:

$$H_0 : \forall j \in [k], \beta_j = 0 \qquad H_1 : \exists j \in [k] \text{ s.t } \beta_j \neq 0.$$

Recall that we have the decomposition $SS_T = SS_{res} + SS_{reg}$. We construct the test-statistic:

$$F_0 = \frac{SS_{reg}/k}{SS_{res}/(n-k-1)} = \frac{MS_{reg}}{MS_{res}} \sim F_{k,n-p} \tag{5.97}$$

In particular, we have (verify this):

$$SS_{reg} = \hat{\beta}^T X^T y - \frac{(\sum y_i)^2}{n} \tag{5.98}$$

$$SS_{res} = y^T y - \hat{\beta}^T X^T y \tag{5.99}$$

$$SS_T = y^T y - \frac{(\sum y_i)^2}{n} \tag{5.100}$$

and the random variables follow the distribution (assuming null is true): $\frac{SS_{reg}}{\sigma^2} \sim \chi^2_k$, $\frac{SS_{res}}{\sigma^2} \sim \chi^2_{n-k-1}$ and $SS_{reg} \perp SS_{res}$. We reject the null hypothesis when $F_0 > F_{k,n-p}(\alpha)$.

5.8.2.10 Coefficients of Determination and Adjustments

Similar to the simple least squares coefficient of determination as in Section 5.8.1.10, we have $R^2 = \frac{SS_{reg}}{SS_T}$. We define them formally here.

Definition 41 (Coefficient of Determination)**.** *The 'R² value', also known as the coefficient of determination or proportion of variation explained by regressors, is defined*

$$R^2 = \frac{SS_{reg}}{SS_T}$$

To penalize model complexity, we have the adjusted coefficient of determination, defined:

Definition 42 (Adjusted Coefficient of Determination)**.**

$$R^2_{adj} = 1 - \frac{SS_{res}/(n-p)}{SS_T/(n-1)} \tag{5.101}$$

5.8.2.11 Interpretation of Model and Coefficients

In the case of a vanilla model with no interaction effects, the interpretation is obvious in that regression coefficients are mean change in response per unit regressor change, ceteris paribus. However, we can have a slightly more nuanced discussion. We can see the coefficients as the *contribution of x_j to response y after BOTH y, x_j have been linearly adjusted for all other regressors.* In particular, consider model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, and suppose we want to interpret the effect of x_2 on y. Then, let the steps follow:

$$\begin{aligned} \text{Model 1} : y \sim x_1 : \hat{y} &= \hat{\alpha}_0 + \hat{\alpha}_1 x_1, \text{ with residual } y - \hat{y} = e_{y \cdot x_1} && (5.102) \\ \text{Model 2} : x_2 \sim x_1 : \hat{x}_2 &= \hat{\gamma}_0 + \hat{\gamma}_1 x_1, \text{ with residual } x_2 - \hat{x}_2 = e_{x_2 \cdot x_1} && (5.103) \\ \text{Model 3} : e_{y \cdot x_1} \sim e_{x_2 \cdot x_1} : \hat{e}_{y \cdot x_1} &= \hat{\lambda}_0 + \hat{\lambda}_1 e_{x_2 \cdot x_1}, \text{ with residual } e_{y \cdot x_1} - \hat{e}_{y \cdot x_1} && (5.104) \end{aligned}$$

From Model 2 it follows that $x_2 = \gamma_0 + \gamma_1 x_1 + e_{x_2 \cdot x_1}$ and we can rewrite

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(\gamma_0 + \gamma_1 x_1 + e_{x_2 \cdot x_1}) + \epsilon && (5.105) \\ &= (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) x_1 + (\beta_2 e_{x2 \cdot x_1} + \epsilon) && (5.106) \end{aligned}$$

Relating to Model 1 we can map $\alpha_0 = \beta_0 + \beta_2\gamma_0, \alpha_1 = \beta_1 + \beta_2\gamma_1, e_{y \cdot x_1} = \beta_2 e_{x_2 \cdot x_1} + \epsilon$. We may perform a similar exercise to obtain $e_{y \cdot x_2} = \beta_1 e_{x_1 \cdot x_2} + \epsilon$. The general relationship for a multiple linear regression model can be summarized:

$$e_{y \cdot x_1 x_2 \cdots x_{j-1} x_{j+1} \cdots x_k} = \beta_j e_{x_j \cdot x_1 x_2 \cdots x_{j-1} x_{j+1} \cdots x_k} + \epsilon$$

where β_j is the contribution of x_j to y after both y, x_j have been linearly adjusted for all other regressors.

5.8.2.12 Regressor Variable Hull and Extrapolation of the Input Space

Consider that in the case of higher dimensions, an unseen input vector can be elementwise in the range of the regressors but lie outside the region of the original data. There is hence a hidden extrapolation.

Definition 43 (Regressor Variable Hull)**.** *Consider n data points and training data $(x_i)_{i \in [n]}$, where x_i is k-dimensional vector of input. The smallest convex set containing all of these data points shall be called the regressor variable hull. The set of points x satisfying*

$$\mathbf{x^T (X^T X)^{-1} x} \leq \max_{\mathbf{i}} \mathbf{diag(H)} = h_{max}$$

where $\mathbf{H}$ is the hat matrix (see Definition 38) $\mathbf{X(X^T X)^{-1} X^T}$ forming an ellipsoid enclosing all points inside the regressor variable hull.

For new input point of p-vector $x_0^T = [1, (x_{0j})_{j \in [k]}]$, we say that using the model to fit x_0 is extrapolation if $\mathbf{x_0^T (X^T X)^{-1} x_0} > h_{max}$.

5.8.2.13 Standardization of Model

In general, the units of $\hat{\beta}_j$ are in terms of the change in units of y per change in units of x_j. To make the model dimensionless, we can do standardization to yield *standardized regression coefficients.*

Unit Normal Scaling: to conduct unit scaling for regressors we take $z_{ij} = \frac{x_{ij} - \bar{x}}{s_j}, i \in [n], j \in [k]$ where $\bar{x}_j = \frac{1}{n} \sum_i x_{ij}$ and $s_j^2 = \frac{1}{n-1} \sum_i (x_{ij} - \bar{x}_j)^2$. For regressors, we can also perform $y_i^* = \frac{y_i - \bar{y}}{s_y}$. Using the new scaled response and regressors, we can construct

$$y^* = \sum_{j=1}^{k} b_j z_j,$$

which after standardization has no intercept!

Unit Length Scaling: for unit length scaling of regressors we form $w_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{S_{jj}}}, i \in [n], j \in [k]$ where S_{jj} is the *corrected sum of squares for x_j* with definition

Definition 44 (Corrected Sum of Squares)**.**

$$S_{jj} = \sum_{i}^{n} (x_{ij} - \bar{x}_j)^2. \tag{5.107}$$

The name follows since each new regressor $\bar{w}_j$ has mean zero and length $\sqrt{\sum_i^n (w_{ij} - \bar{w}_j)^2} = 1$. Using unit length scaling for response $y_i^0 = \frac{y_i - \bar{y}}{\sqrt{SS_T}}$ and fitting the model

$$\hat{y}^0 = \sum_{j}^{k} b_j w_j,$$

we arrive at some interesting properties. In particular, the matrix $\mathbf{W^T W} = \rho(\mathbf{X})$, the correlation matrix and $\mathbf{Z^T Z} = (\mathbf{n} - \mathbf{1})\mathbf{W^T W}$ - that is the estimates of regression coefficients from norm scaling and length scaling are the same.

5.8.2.14 Indicator Functions

We may also encode qualitative/categorical variables of k levels by a $k-1$ indicator function set - in fact any binary function that maps to discrete values shall suffice, but indicators are most used. Here the assumption that variance is constant in all levels of category is used, instead of across continuous axis. In the vanilla model we can easily see that the indicator function indicate change of intercept (see by substitution of $\{0, 1\}$), with the interpretation as difference of means. In the case of interaction terms with continuous variables this can also mean the change in both slope and intercept, which we may also verify simply by specifying such a model and substituting our own values! Note that interaction term requires that the atomic regressor is already included in the model, and we may test for the significance of the categorical variable using the partial sum of squares (see Section 40) method. Suppose the model follows $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$ where x_2 is the indicator function affecting intercept and slope of x_1, then we perform the (abbreviated) test $H_0 : \beta_2 = \beta_3 = 0$ with test statistic

$$F_0 = \frac{SS_{reg}(\beta_2, \beta_3 | \beta_1, \beta_0)/2}{MS_{res}} \sim F_{2,n-4}.$$

Note that the ‘partial sum’ comes from the fact that we may do something like $SS_{reg}(\beta_2, \beta_3 | \beta_1, \beta_0) = SS_{reg}(\beta_2 | \beta_1, \beta_0) + SS_{reg}(\beta_3 | \beta_2, \beta_1, \beta_0)$.

5.8.3 Adequacy of The Least Squares Method

We made assumptions in the phase of model fitting and during the phase of model analysis and interpretation (see Section 5.8.1.4 and Section 5.8.2.2). Obviously, ex-ante we made the assumption on the form of the model, in that it is parametric and the relationship between response and regressors be *approximately* linear, known as the *linearity assumption.* We also made assumptions about the errors - mostly that the errors ϵ has expectation zero and constant variance, known as the *constant variance/homogeneity assumption.* In the analysis, we further assumed they are $\epsilon \overset{IID}{\sim} \Phi(0, \sigma^2)$, which is our *normality assumption* and *independent error assumption.*

A possible method would be the study of residuals, either graphically or in numerical forms. Types of graphs include one-dimensional graphs (such as histograms, stem-and-leaf plots, dot plots, box plots et cetera), two-dimensional graphs (such as the scatter plot). However, in the case of multiple linear regression, even the pairwise correlation matrix (and therefore the scatter plot) may not indicate the presence of linear relationships between regressors. Ideally, pairwise plot of regressors show little to no correlation (linear patterns), but even then there may exist linear multivariate relationships that involve more than two variables. These plots can be performed ex-ante. An ex-post analysis may also be performed, after the model has been fit.

5.8.3.1 Residual Analysis

After fitting the model we may perform residual analysis of $e_i = \hat{y}_i - y_i, i \in [n]$, which have zero mean and variance approximated

$$\frac{\sum_i (e_i - \bar{e})^2}{n-p} = \frac{\sum_i e_i^2}{n-p} = \frac{SS_{res}}{n-p} = MS_{res}.$$

When $n >> p$ then the dependence between residuals e_i is weak, but otherwise they are not independent.

5.8.3.1.1 Leverage Values, Influential Values and the Variance of Residuals

Definition 45 (Leverage Values)**.** *Recall that we defined the hat matrix 38* $\mathbf{X}(\mathbf{X^TX})^{-1}\mathbf{X^T}$ *and obtained* $\hat{\mathbf{y}} = \mathbf{Hy}$*, and we can rewrite* $\hat{y}_i = \sum_j^n h_{ij} y_j$*, which shows that the response prediction is a weighted sum of all observations. In particular we shall then call* h_{ii} *the leverage value for the i-th observation, indicating the weight of* y_i *in* $\hat{y}_i$*.*

We also wrote the residuals to take $\mathbf{e} = (\mathbb{1} - \mathbf{H})\mathbf{y}$. By substitution of $\mathbf{y} = \mathbf{X}\beta + \epsilon$ we obtain

$$\begin{aligned}
\mathbf{e} &= (\mathbb{1} - \mathbf{H})(\mathbf{X}\beta + \epsilon) && (5.108)\\
&= \mathbf{X}\beta + \epsilon - \mathbf{HX}\beta - \mathbf{H}\epsilon && (5.109)\\
&= \mathbf{X}\beta + \epsilon - \mathbf{X}(\mathbf{X^TX})^{-1}\mathbf{X^TX}\beta - \mathbf{H}\epsilon && (5.110)\\
&= \mathbf{X}\beta + \epsilon - \mathbf{X}\beta - \mathbf{H}\epsilon && (5.111)\\
&= \epsilon - \mathbf{H}\epsilon && (5.112)\\
&= (\mathbb{1} - \mathbf{H})\epsilon && (5.113)\\
& && (5.114)
\end{aligned}$$

and using $\text{Var}(\epsilon) = \sigma^2 \mathbb{1}$ and the fact that $\mathbb{1} - \mathbf{H}$ is idempotent symmetric,

$$\text{Var}(\mathbf{e}) = \text{Var}((\mathbb{1} - H)\epsilon) = (\mathbb{1} - H)\text{Var}(\epsilon)(\mathbb{1} - H)^T = \sigma^2(\mathbb{1} - H) \tag{5.115}$$

Finally we arrive at the results

$$\text{Var}(\epsilon_i) = \sigma^2(1 - h_{ii}), \qquad Cov(e_i, e_j) = -\sigma^2 h_{ij} \tag{5.116}$$

where in the covariance term we use the fact that identity matrices have zero non-diagonals.

5.8.3.2 Standardization of Residuals

Following the variance study of residuals in Equations 5.116 we can form standardized residuals. In particular we define:

Definition 46 (Standardized Residuals)**.** *Standardized residuals are residuals defined as*

$$\frac{e_i}{\sigma\sqrt{1-h_{ii}}}, \qquad i \in [n] \tag{5.117}$$

but again recall that our estimator of σ is $\sqrt{MS_{res}}$. Another alternative is $\hat{\sigma}_{(i)}$ where

$$\hat{\sigma}^2_{(i)} = \frac{SS_{res(i)}}{n-k-2} = \frac{SS_{res(i)}}{n-p-1} \tag{5.118}$$

where $SS_{res(i)}$ is the sum squared residuals when model is fitted without the i-th observation. Then both $MS_{res}, \hat{\sigma}^2_{(i)}$ form unbiased estimator of σ^2. Then we may estimate the standardized residuals by substitution with the formulations:

Definition 47 (Internally Studentized Residuals)**.** *Internally Studentized Residuals are defined*

$$r_i - \frac{e_i}{\sqrt{MS_{res}(1-h_{ii})}}, \qquad i \in [n]$$

and

Definition 48 (Externally Studentized Residuals)**.** *Externally Studentized Residuals are defined*

$$r_i^* - \frac{e_i}{\sqrt{\hat{\sigma}_{(i)}(1-h_{ii})}}, \qquad i \in [n]$$

which can be shown to be related via the monotonic (and hence similar) transformation $r_i^* = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}}$.

5.8.3.3 Checking Normality Assumptions

We can also check for the normality assumptions on residuals using quantile plots such as the QQ plot (see Section 4.3.1). Plotting against normal scores the ordered standardized residuals, we can look out for the presence of tails / skewness, as well as large residuals representing potential outliers. Additionally, a scatter plot of standardized residuals against either the response or any of the regressors should yield no pattern. A funnel, cone or bow shape can indicate heteroscedasticity, curves and quadratic pattern can indicate that the linearity assumption is violated and so on. Additionally, under the normality assumption we should expect (most) standardized residuals to fall within an absolute value of three. For instance, a 'double bow' (imagine the shape of one's lips) often occurs when the response is a proportion between zero and one, since the variance of binomial random variable near 0.5 is greater than when it is near the extreme values of proportion (zero, one). Note that in the simple regression, scatter plot $r_i \sim X$ is *visually* identical to $r_i \sim \hat{y}$.

5.8.3.4 Note on Time Series Data

Additionally, if the data is serially sampled, it might be worth looking at the standardized residuals plot against the time order. Ideally, we obtain the 'no pattern' pattern, the violation of which might suggest we have variance as function of time $\sigma(t)$ or even autocorrelation $\rho(\epsilon_t, \epsilon_{t-1}) \neq 0$ which is a serious violation of the independence assumption of errors.

5.8.3.5 Outliers and Influential Data

Residuals that are considerably (absolutely) larger than the others may indicate potential outliers in the output space. QQ plots (see Section 4.3.1), scatter plots and normality tests are good ways of identifying outliers. Although outliers can be 'bad' and corrected/removed in the case of sampling faults, sometimes it just reflects a black swan event that is perfectly plausible - in this case deleting the data point can lead to dangerous conclusions and false sense of model accuracy. The removal of (bad) outliers can affect regression coefficient estimates, residual sum of squares, coefficients of determination and error variances. This in turn affects interval estimates and so on.

Recalling the idempotent property, again we have $\text{Var}(\hat{y}) = \text{Var}(Hy) = H\text{Var}(y) = H\sigma^2$. The hat matrix determines the covariance of $\hat{y}$ and e, the fitted value and error matrices. h_{ij} is the *amount of leverage exerted by the j-th observation y_j on i-th fitted value $\hat{y}_i$.* Recall from Definition 45 that $h_{ii} = x_i(X^TX)^{-1}x_i$ is a standardized measure of the distance of the i-th observation from the center of the x space. Large values of h_{ii} indicate potentially influential points since they are different from the other points in the input space! We have $\sum diag(\mathbf{H}) = rank(\mathbf{H}) = p$, therefore the average size of a hat diagonal should be $\frac{p}{n}$. We can say that for any hat diagonal exceeding $\frac{2p}{n}$, the point is a *leverage point* (assuming $\frac{2p}{n} < 1$). Points with large residuals and large diagonals are likely to be *influential points*, in that they affect model summary statistics and regression coefficient estimates to a more significant degree.

To measure the influence of data points, we can use the squared distance between the least square estimates based on all n data and without the data point. Letting the estimate for regression coefficients excluding point i be $\hat{\beta}_{(i)}$, then define

Definition 49 (Cook's Distance)**.** *Cook's Distance is defined*

$$D_i = (\mathbf{M}, c) = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T\mathbf{M}(\hat{\beta}_{(i)} - \hat{\beta})}{c} \qquad i \in [n] \tag{5.119}$$

where $\mathbf{M}, c$ usually takes form $\mathbf{X^TX}, pMS_{res}$ respectively. With these forms of $\mathbf{M}, c$, then $D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}$ where r_i is the internally studentized residuals defined as in Definition 47.

Points with large Cook's Distance are said to have considerable influence. We can interpret the value as follows: if $D_i = F_{p,n-p}(0.5) \approx 1$, then deleting the point i would move $\hat{\beta}_{(i)}$ to the boundary of an approximately 50% confidence region for β. Therefore, the cutoff $D_i > 1$ is often used.

Other useful measure of influence are *DFFITS* and *DFBETAS*. While invalid data may be removed, if there is no justification for removal we shall not do so. A common method called the 'robust estimation technique' would be to down-weight the influence of the point in the model estimates.

5.8.3.6 Lack of Fit

A lack-of-fit test requires that for a single level of response, we have replicate observations on response. Suppose in the simple linear model we have $y \sim x$ and in the training data we have $x = \{x_1, x_2, \cdots, x_m\}$ discrete, $m \leq n$ levels. Now assume that for each level $i \in [m]$ that we have n_i observations, and denote y_{ij} to be observation j for i-th level regressor. Note that we have $n = \sum_i n_i$. The ij-th residual is $y_{ij} - \hat{y}_i = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$. Squaring both sides and summing over all inputs we obtain

$$\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2}_{SS_{res}} = \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}_{SS_{PE}} + \underbrace{\sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2}_{SS_{LOF}} \tag{5.120}$$

since the cross terms evaluate to zero. We decomposed the sum of squared residuals into two components, namely the pure error and the lack of fit term. Compare this to the sum of squares in the general ANOVA problem discussed in Definition 54.

Definition 50 (Lack of Fit Sum of Squares)**.** *We define the lack of fit sum of squares to be*

$$SS_{LOF} = \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$$

with the terminologies as specified prior, and is the weighted sum of squared deviations of fitted and mean response values. If the fitted values are close to the mean, then the lack of fit value is small and strongly indicates a linear relationship of in the regression function.

The lack of fit test statistic follows distribution

$$F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}} \sim F_{m-2,n-m} \tag{5.121}$$

under the null hypothesis of H_0 : true regression function is linear .In a simple model, the $H_0 : \beta_1 \neq 0$ is the equivalency of 'model is linear' and we reject the null hypothesis if $F_0 > F_{m-2,n-m}(\alpha)$. Unfortunately, this is not very useful in many cases, especially in multiple regression models - repeat observations do not occur often in higher dimensionality.

5.8.3.7 Multicollinearity

When there is no linear relationship between regressor variables, we say that they are orthogonal. In most applications however, this is not the case. In the discussions that follow, we will assume all regressors have been centered and unit scaled. In many cases the regressors can be linearly dependent, affecting the inferences based on the regression model. The presence of near-linear dependencies between regressors is called the *multicollinearity* problem, and affects regression coefficient estimates. Define $\mathbf{X_j}$ to be the j-th column of input variable matrix $\mathbf{X}$. Then if there exists nonzero solutions to $\mathbf{t}$ where $\sum_j t_j \cdot X_j = 0$ then $(\mathbf{X^TX})^{\mathbf{-1}}$ does not exist. Multicollinearity can be introduced in different stages of the model building process, such as (i) data collection, (ii) model/population constraints, (iii) model specifications and (iv) over-defined modelling. If the sampling method is not a pure *random sampling* and involves the sampling of a sub-sample of the entire sample space - such that there exists correlation between variables - we can run into multicollinearity issues. In the case of problem specific constraints, we might have a situation where we are modelling $y \sim x_1 + x_2$ and it turns out x_1 and x_2 are related intrinsically (think income and housing size, et cetera). Model specifications such as higher order polynomial term might introduce multicollinearity, especially so when the range of a regressor variable is small. Over-defined models are situations in which $p > n$. Recall that the variance of regression estimates were given $Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$ (see Section 5.82), then the ℓ_2 error of our regression estimates can be computed as the squared distance: $L_1^2 = (\hat{\beta} - \beta)^T(\hat{\beta} - \beta)$ and we have

$$\mathbb{E}L_1^2 = \mathbb{E}[(\hat{\beta} - \beta)^T(\hat{\beta} - \beta)] = \sum_{j=1}^{j} \mathbb{E}(\hat{\beta}_j - \beta_j)^2 = \sum_{j=1}^{k} \text{Var}(\hat{\beta}_j) = \sigma^2 \text{trace}(\mathbf{X^TX})^{-1} \tag{5.122}$$

where the trace is the sum diagonals. When multicollinearity exists, some of the eigenvalues of $\mathbf{X^TX}$ will turn out to be small. Letting λ_j denote the j-th eigenvalue of X^TX, then $\mathbb{E}L_1^2 = \sigma^2\text{trace}(\mathbf{X^TX})^{-1} = \sigma^2 \sum_{j=1}^{k} \lambda_j^{-1}$. At least one of the eigenvalues being involved in a multicollinear relationship inflates the $\frac{1}{\lambda_j}$ and the expected error loss in estimation of $\hat{\beta}$ becomes large. Since $\text{Var}(\hat{\beta}_j) = \sigma^2 C_{jj}, j \in [k]$, we can show that *if* $\mathbf{X}$ *is unit-length scaled (verify this)* then $\rho(\mathbf{X}) = \mathbf{X^TX}$. We can show that j-th diagonal of $[\rho(X)]^{-1} = \frac{1}{1-R_j^2}$ where R_j^2 is the coefficient of determination from regression of $x_j \sim \{x_{i\in[k]\wedge j\neq i}\}$. We can see directly from this that strong multicollinearity effects between regressors inflate both the variance and covariance of our least squares estimates! Often, when $\mathbf{X}$ is unit scaled we just call $\mathbf{X^TX}$ as the correlation matrix of $\mathbf{X}$ and call it the correlation form of $\mathbf{X^TX}$. Multicollinearity decreases the generalisability of the model - although it fits well to training data, the poor regression coefficient estimates makes the model poor for prediction for inputs outside the observed input space.

5.8.3.7.1 Detection of multicollinearity - Plots, Variance Inflation Factors and Eigensystem Analysis :

Correlation Plots A simple (but non-sure) way of identifying potentially multicollinear variables is to look at the off-diagonal elements in $\rho(\mathbf{X})$. However, there are many instances when pairwise correlation do not reflect the presence of multicollinearity, since near-linear dependence involving more than two regressors are not reflected.

Variance Inflation Factors Recall that the diagonals of $\rho(X)^{-1}$ correlation matrix can be useful in detecting multicollinearity. In particular, we have the values:

$$\frac{1}{1-R_j^2}, \qquad j \in [k]. \tag{5.123}$$

Since the variance of the j-th regressor equals $\sigma^2 C_{jj}$ and C_{jj} increases with $\frac{1}{1-R_j^2}$, then we can view this value as the factor by which the variance of $\hat{\beta}_j$ increases due to non-linear dependence among regressors.

Definition 51 (Variance Inflation Factors)**.** *The variance inflation factor for the j-th regressor is the value*

$$VIF_j = \frac{1}{1-R_j^2}, \tag{5.124}$$

for which the presence of a large value indicates multicollinearity. [1]

Note that scaling data can help decrease the variance inflation factors and improve the fit.

Eigensystem Analysis Since the eigenvalues/characteristic values of a square matrix $\mathbf{A} \in \mathbb{R}^{p,p}$ are the k roots to the system $\|\mathbf{A} - \lambda\mathbb{1}\| = 0$, then the eigenvalues of $\rho(X)$, that is $\{\lambda_i, i \in [k]\}$, can be used to measure collinearity in data. Small eigenvalues indicate collinearity issues. Let the *condition number* of $\rho(X)$ be $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$, then if $\kappa > 100$ we say that collinearity issues exist. This does not tell us however, the number of regressors involved in the collinearity relationship. The number of condition indices, $\kappa_j = \frac{\lambda_{max}}{\lambda_j}, j \in [k]$ gives a measure of the number of such near-linear dependencies. The method of eigensystem analysis can also be used to identify the nature of this near-linear dependence.

5.8.4 Correction of Inadequacies

5.8.4.1 Transformation of Response-Regressor

A violation of the linearity assumption may be detected via the analysis of standardized residuals or in the lack of fit test, as we saw in Section 5.8.3. In these cases, the originally nonlinear function may be *transformably linear.* For instance, consider the following linearizable, non-linear functions and their linear transformations:

[1] A VIF value exceeding 5 or 10 can indicate that the associated coefficient estimates are poorly estimated. The VIFs not only detect collinearity but suggest which regressors are involved in the collinear relationship.

$y \sim x$	Transform	$y' \sim x'$
$y = \beta_0 x_1^{\beta}$	$y \to \log y, x \to logx$	$y' = \log \beta_0 + \beta_1 x'$
$y = \beta_0 \exp\{\beta_1 x\}$	$y \to \ln y$	$y' = \ln \beta_0 + \beta_1 x$
$y = \beta_0 + \beta_1 \log x$	$x \to \log x$	$y' = \beta_0 + \beta_1 x'$
$y = \frac{x}{\beta_0 x - \beta_1}$	$y \to \frac{1}{y}, x \to \frac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

When performing transformations, the problem domain needs to be taken into account. For instance, two equally feasible transformations $x \to \frac{1}{x}$ and adding a non-linear higher order term x^2 may be presented - then we should question if the quadratically U-shaped relationship between response and regressor is an intuitively reasonable one.

We may also rely on analytical methods to specify appropriate transformations, such as the Box-Cox method.

5.8.4.1.1 Box-Cox Method We may transform the response y to correct for non-normality and heteroscedasticity, using the power transform y^λ, and then estimating β, λ using maximum likelihood methods (see Section 10). We use the transformation

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda y \tilde{y}^{\lambda-1}} & \text{when } \lambda \neq 0 \\ \tilde{y} \log y & \text{when } \lambda = 0, \end{cases} \tag{5.125}$$

where $\tilde{y} = \exp\{\frac{1}{n}\sum_i^n \log y_i\}$ represents geometric mean of observations. The regression coefficients are obtained by fitting the model $\mathbf{y}^{(\lambda)} = \mathbf{X}\beta + \epsilon$ using the least squares (see 5.74) method or the maximum likelihood estimation method (see 10).

Note that when comparing between different λ transformations, we may not use SS_{res} since each of the computations are scaled differently. The MLE estimate corresponds to the λ value for which the SS_{res} from the *fitted model* is minimum. If $\lambda = 0$, then we choose $\log y$ as response, else our new response variable is y^λ.

5.8.4.1.2 Box-Tidwell Method Another transformation is on the regressor variables instead of the response. Assuming that $\epsilon \overset{IID}{\sim} \Phi(0, \sigma^2)$ is at least approximately satisfied, then we may apply the Box-Tidwell method as follows. For the simple model, consider transformation

$$\xi = \begin{cases} x^\alpha & \alpha \neq 0 \\ \log x & \alpha = 0 \end{cases} \tag{5.126}$$

and with least squares fit $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Fit a new model with addition of regressor $w = x \log x$, such that we have $\hat{y} = \hat{\beta}_0' + \hat{\beta}_1' x + \hat{\gamma} w$ and take $\alpha_1 = \frac{\hat{\gamma}}{\hat{\beta}_1} + 1$. Repeat the procedures using a new regressor $x \to x^{\alpha_1}$, and this converges rapidly to a satisfactory result of α.

Another method of correcting the inadequacies are to adjust weights of points such that they are approximately equal in importance to the model fits. This is discussed in the section on the weighted least squares method (see Section 5.8.5).

5.8.4.2 Ridge Regression

$$MSE(\hat{\beta}') = \mathbb{E}(\hat{\beta}' - \beta)^2 = \text{Var}(\hat{\beta}') + Bias(\hat{\beta}')^2. \tag{5.127}$$

The measurement model relating to this decomposition is given in Equation 3.2. If we are able to increase bias to smaller extent than the decrease in variance, our estimate has lower expected errors. From here we refer to X^TX in its correlation form.

Definition 52 (Ridge Estimator)**.** *Define the ridge estimator $\hat{\beta}_R$ as solution to $(X^TX + \lambda\mathbb{1})\hat{\beta}_R = X^Ty$, therefore we have*

$$\begin{aligned} \hat{\beta}_R &= (X^TX + \lambda\mathbb{1})^{-1}X^Ty & (5.128)\\ &= (X^TX + \lambda\mathbb{1})^{-1}X^TX\hat{\beta} & (5.129)\\ &= Z_\lambda\hat{\beta} & (5.130)\end{aligned}$$

where $\lambda \in [0,1]$ is called the biasing parameter.

Then we have (verify this)

$$\begin{aligned} MSE(\hat{\beta}_R) &= \mathrm{Var}(\hat{\beta}_R) + Bias(\hat{\beta}_R)^2 & (5.131)\\ &= \sigma^2\sum_{j=1}^{k}\frac{\lambda_j}{(\lambda_j+\lambda)^2} + \lambda^2\beta^T(X^TX + \lambda\mathbb{1})^{-2}\beta & (5.132)\end{aligned}$$

where each λ_j is the j-th eigenvalue of X^TX. As the biasing parameter increases, the variance decreases and bias increases. There exists $\lambda \neq 0$ such that $MSE(\hat{\beta}_R) < \mathrm{Var}(\hat{\beta})$, provided $\hat{\beta}^T\hat{\beta}$ is bounded. Note that since (verify this)

$$\begin{aligned} SS_{res} &= (y - X\hat{\beta}_R)^T(y - X\hat{\beta}_R) & (5.133)\\ &= (y - X\hat{\beta})^T(y - X\hat{\beta}) + (\hat{\beta}_R - \hat{\beta})^TX^TX(\hat{\beta}_R - \hat{\beta}), & (5.134)\end{aligned}$$

where the first term is the SS_{res} of the unbiased least squares, the ridge regression results in poorer fit and lower coefficients of determination as a tradeoff to lower mean squared error of regression estimates. The *ridge trace* is a useful plot to observe the elements of $\hat{\beta}_R$ against the values of λ. In the case of multicollinearity issues, the instability of regression coefficients can be observed from the ridge trace. The ridge regression method is often a technique employed to deal with multicollinear effects.

5.8.4.3 Principal Component Regression

5.8.5 Weighted Least Squares

When faced with data with heteroscedastic variance, we can fit by weighted least squares. The residual terms $(y_i - \hat{y}_i)$ are scaled by weight inversely proportional to $\mathrm{Var}(y_i)$. Recall that in the simple least squares 5.31 we took $S(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$, and we replace this with the weighted sum of squares objective function:

$$S(\beta_0, \beta_1) = \sum_i w_i(y_i - \beta_0 - \beta_1 x_i)^2 \qquad (5.135)$$

with normal equations

$$\hat{\beta}_0\sum_{i=1}^{n} w_i + \hat{\beta}_1\sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} w_i y_i \qquad (5.136)$$

$$\hat{\beta}_0\sum_{i=1}^{n} w_i x_i + \hat{\beta}_1\sum_{i=1}^{n} w_i x_i^2 = \sum_{i=1}^{n} w_i y_i x_i \qquad (5.137)$$

which by solving yields $\hat{\beta}_0, \hat{\beta}_1$ as regression coefficient estimates. The w_i here are not variable - they are set inversely proportional to their variance (such as $w_i = \frac{1}{\sigma_i^2}$) where the error i term is determined variance σ_i^2. These variances are unknown a priori but can be estimated and are to be discussed later. Writing the weighted means $\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$ and $\bar{y}_w = \frac{\sum w_i y_i}{\sum w_i}$, we obtain (verify this) regression estimates

$$\hat{\beta}_0 = \bar{y}_w - \hat{\beta}_1 \bar{x}_w \tag{5.138}$$

$$\hat{\beta}_1 = \frac{\sum_i w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_i w_i (x_i - \bar{x}_w)^2} \tag{5.139}$$

that yields unbiased estimates. The weighted mean squared residuals $MS_{(w)Res}$ is an unbiased estimator of σ^2. Interpretation of the models are the same as when weights are uniform. We may remove point i by setting $w_i = 0$. Additionally, outliers or influential points may be set $w_i < \bar{w}$ to down weight impact on coefficient estimates relative to others. Since we need conditional variance σ_i^2, this can be difficult to obtain. Suppose we already know the relationship such that $\text{Var}(y|x) = f(x)$ then we may encode it in a functional form. However, in many cases the underlying distributions and relationships are not known - and we may rely on methods such as estimation on the *multiple (nearly) repeated values of the regressor.* Althought we ideally want multiple response values at each level of the regressor value, we often do not have sufficient data. Instead, we bin the regressor axis and group them. Let this group formed from the neighbourhood of x_i have average $\bar{x}$ and sample variance s_y^2. We may perform a least squares for $s^y \sim \bar{x}$, and substitute $x_i \to \bar{x}$ to obtain an estimate of σ_i^2.

5.8.6 Generalized Least Squares

In the generalized least squares we want to fit

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

with the weaker assumptions that $\mathbb{E}\epsilon = 0, \text{Var}(\epsilon) = \sigma^2 V$. In the multiple least squares assumptions (see Section 5.8.2.2) our assumption of constant and independent assumption corresponds to $\mathbf{V} = \mathbb{1}$. When V is diagonal matrix then our problem set is uncorrelated but non-constant variance, and in the general case we have both correlated errors and non-constant variance.

We may no longer use the estimates $\hat{\beta} = (X^T X)^{-1} X^T y$ when $\mathbf{V} \neq \mathbb{1}$. What we want to do is then perform transformation of the model into a new set of observations satisfying the ordinary least squares assumptions and utilize that machinery. Let $\sigma^2 V$ be represent matrix $Cov(\epsilon)$, then the generalized least squares normal equations become (verify this):

$$(\mathbf{X^T V^{-1} X})\hat{\beta} = \mathbf{X^T V^{-1} y} \tag{5.140}$$

with solution

$$\hat{\beta} = (\mathbf{X^T V^{-1} X})^{-1} \mathbf{X^T V^{-1} y}, \tag{5.141}$$

the *generalized least square estimator of* β. Note that in the special case where $\sigma^2 V$ is diagonal matrix, let $\mathbf{W} = \mathbf{V^{-1}}$ and we derive the weighted least squares equation (verify this)

$$(\mathbf{X^T W X})\hat{\beta} = \mathbf{X^T W y} \tag{5.142}$$

with solution

$$\hat{\beta} = (\mathbf{X^T W X})^{-1} \mathbf{X^T W y}. \tag{5.143}$$

5.8.7 Variable Selection Methods

We discuss the tradeoff between model complexity and predictive power in this section. The basic strategy can be modelled as follows: 1) fit a full model, 2) perform analysis and validity studies, 3) determine statistical relevance and significance by tests, 4) edit model and repeat.

Assume a k regressor candidate problem with regressors $x_i, i \in [k]$ and response y. Then our model is specified $y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon, i \in [n]$. Suppose r regressors should be deleted from the model, then we can decompose $\mathbf{X_q}, \mathbf{X_r}$ such that $\mathbf{X}_q$ is the $q = k - r + 1$ columns of intercept and significant regressors, while the $\mathbf{X_r}$ are the deleted ones. We rewrite $y = X_q\beta_q + X_r\beta_r + \epsilon$. Recall the full model 5.74 $\hat{\beta}^* = (X^TX)^{-1}X^Ty$ where $\hat{\beta}^*$ consists of $\hat{\beta}_q^*$ and $\hat{\beta}_r^*$ and

$$\hat{\sigma}_*^2 = \frac{y^Ty - \hat{\beta}^{*T}X^Ty}{n-k-1} = \frac{y^T[\mathbb{1} - X(X^TX)^{-1}X]y}{n-k-1} \tag{5.144}$$

with fitted values $\hat{y}_i^*$. The subset model takes form $\hat{\beta}_q = (X_q^TX_q)^{-1}X_q^Ty$ and

$$\hat{\sigma}^2 = \frac{y^Ty - \hat{\beta}_q^TX_q^Ty}{n-q} = \frac{y^T[\mathbb{1} - X_q(X_q^TX_q)^{-1}X_q]y}{n-q} \tag{5.145}$$

and fitted values $\hat{y}_i$. Then a question we may ask is with regards to the consequence of mis-specifying our model by the r regressors. It turns out that (verify this) $\mathbb{E}\hat{\beta}_q = \beta + A\beta_r$ where $A = (X_q^TX_q)^{-1}X_q^TX_r$, which results in a biased estimator unless β_r is zero vector or the regressors are orthogonal to the retained variables, in that $X_q^TX_r = 0$. We also have $\text{Var}(\hat{\beta}_q) = \sigma^2(X_q^TX_q)^{-1}, \text{Var}(\hat{\beta}^*) = \sigma^2(X^TX)^{-1}$ by definition and writing $\text{Var}(\hat{\beta}_q^*) - \text{Var}(\hat{\beta}_q)$ we obtain a matrix such that all variances of regression coefficients in the full model are $\geq$ to variances of coefficients in the reduced model. *Removal of unnecessary variables will not increase the variance of remaining coefficients.* We also have the $MSE(\hat{\beta}_q) < MSE(\hat{\beta}_q^*)$, where the subset model has smaller mean squared errors.

Note that in the full model, the $\hat{\sigma}_*^2$ is unbiased estimator of σ^2, while the $\hat{\sigma}^2$ from the subset model is a biased upward estimate of σ^2. When predicting response at point $x^T = (x_q^T, x_r^T)$, the predicted response from full model is $\hat{y}^* = x^T\hat{\beta}^*$ with mean $x^T\beta$ and prediction variance $\text{Var}(\hat{y}^*) = \sigma^2[1 + x^T(X^TX)^{-1}x]$, in comparison to the predicted response of $\hat{y} = x_q^T\hat{\beta}_q$ with mean $\mathbb{E}\hat{y} = x_q^T\beta_q + x_q^TA\beta_r$ and mean of squared error

$$MSE(\hat{y}) = \sigma^2[1 + x_q^T(X_q^TX_q)^{-1}x_q] + (x_q^TA\beta_r - x_r^T\beta_r)^2. \tag{5.146}$$

Although (in general) we have a biased estimate of y, the variance of $\hat{y}^*$ from the full model is not less that the variance of $\hat{y}$ from the subset model.

We may show (verify this):

$$MSE(\hat{y}^*) = \text{Var}(\hat{y}^*) \geq MSE(\hat{y}) \tag{5.147}$$

in the misspecification of the model. Hence, deleting variables improves the precision of the regression estimates of the retained models, reduce variance of predicted response. If these deleted variables are not significant then the bias-variance tradeoff is favourable and we earn lower mean squared errors.

5.8.7.1 Selection Criterion

Some of the metrics to select different models can be named as follows: i) the coefficient of (multiple) determination ($R_q^2 = \frac{SS_{reg}(q)}{SS_T} = 1 - \frac{SS_{res}(q)}{SS_T}$), adjusted coefficient of multiple determination ($R_{adj}^2 =$

$1-\frac{n-1}{n-p}(1-R_p^2))$, Residual Mean Squares (MS_{res}), Akaike Information Criteria and Bayesian Information Criteria.

For the $MS_{res} = \frac{SS_{res}}{n-p}$, there is a tradeoff between the loss of degrees of freedom and a decrease in the SS_{res} values - ideally, we want a model with the minimum MS_{res}. We may desire to choose a model with the number of regressors such that this value is minimized, or a nearby model.

5.8.7.1.1 Akaike Information Criteria The AIC method is based on maximising the expected entropy of them model, trading off goodness-of-fit for simplicity of model.

Definition 53 (AIC)**.** *We define the AIC to be the value*

$$AIC = -2\log(L) + 2p, \qquad p = k + 1 \tag{5.148}$$

where L is defined the likelihood function of the model.

In the ordinary least squares setting, we have (verify this): $AIC = n\log(\frac{SS_{res}}{n}) + 2p$ and our goal may be to select the model with smallest AIC value.

5.8.7.1.2 Bayesian Information Criteria The BIC is an extension of AIC[2] (see Section 5.8.7.1.1) with several variants. The Schwartz and Sawa variant is given

$$BIC_s = -2\log(L) + p\log(n) \tag{5.149}$$

which in the ordinary least squares yields $BIC_s = n\log(\frac{SS_{res}}{n}) + p\log n$. We would like a model with the lowest BIC.

5.8.7.2 Computational Methods - Brute Force and Stepwise Greedy Solutions

5.8.7.2.1 Brute Force Method : With a total of k candidate regressors, we may fit a model over the power set and compare 2^k such models to compare based on agent-defined utility criterion (see 5.8.7.1). Note that since each model gives regression coefficient estimates, a useful side-effect is that we may be able to detect multicollinearity issues by observing instability of the regression estimates. If this is computationally unviable we may opt for greedy methods, although no global optima is guaranteed.

5.8.7.2.2 Forward Selection Method : The steps may be enumerated as follows (1) begin with the intercept model, and pick a regressor that has the highest correlation with the response. (2) If the F statistic of the model is significant beyond some threshold, greedily select another regressor that has the largest correlation with response after adjusting for the effect of the first regressor on response. This is known as *partial correlation.* (3) Compute the partial F statistic, and if this exceeds threshold add the regressor, repeat and otherwise terminate.

The partial correlation can be determined as follows:

1. First derive $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$

2. Then fit $\hat{x}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j} x_1, \text{ for all } j \in [2, k]$

3. Derive $\rho(\hat{y} - y, \hat{x}_j - x_j), \text{ for all } j \in [2, k]$

[2] note that when using AIC and BIC, the comparison of models are only valid when the response is the same. Additionally, the model should be fit on the same size of data, since the input n is part of the criterion.

The value in step 3 is known as the partial correlation value of x_j with y, which also yields the largest partial F statistic

$$\frac{SS_{reg}(x_2|x_1)}{MS_{res}(x_1, x_2)}$$

which we compare against the threshold to see if we shall continue.

5.8.7.2.3 Backward Elimination Method : The method may be described in a similar format to the forward selection, but in reverse order. First begin with the full k regressor model, and examine the partial F statistic (recall that this is equivalent to the t-test statistic) as if it were the last variable to be added in the model. If the smallest partial F statistic is lower than some threshold, remove the regressor and repeat or terminate.

5.8.7.2.4 Stepwise Regression Method : The stepwise method is a combination of the forward and backward elimination methods. Beginning with the intercept model, we initiate forward selection. The modification is that after the addition of each variable we conduct partial F tests for all variables in the model to see if any regressors previously added have become redundant due to relationships between it and the incoming regressor. The removal or addition of more regressors then follows from the comparison against two thresholds, which are generally distinct.

5.9 Analysis of Variance Methods - ANOVA/F-Test

The question we may have is whether the differences in population parameter estimates are considered significant or simply due to random variance.

5.9.1 F Test

A one-way layout is an experimental design in which *independent measurements are made under several treatments.* As such, the F test is a generalization of the two sample/treatment problem. Let there be I groups and J measurements in each group. We also discuss the case when $J_i \neq J_j$ for some $i \neq j$ group. Further denote Y_{ij} the j-th measurement in group i, and suppose the sampling is generated from the function

$$Y_{ij} = \mu + \alpha_i + e_{ij}, \qquad e_{ij} \overset{IID}{\sim} \Phi(0, \sigma^2) \tag{5.150}$$

and α_i are normalized such that $\sum_i \alpha_i = 0$.

The null hypothesis can be specified as follows: $H_0 : \forall i, \alpha_i = 0$, that there is no difference between the expected values under selected treatment.

Definition 54 (ANOVA Sum of Squares). *Let $\bar{Y}_i = \frac{1}{J}\sum Y_{ij}$ and $\bar{\bar{Y}} = \frac{1}{IJ}\sum^I \sum^J Y_{ij}$, then we can decompose the total sum of squared errors into the group sum of squares and sum squares between groups. In particular, we have*

$$\underbrace{\sum^I \sum^J (Y_{ij} - \bar{\bar{Y}})^2}_{SS_T} = \underbrace{\sum^I \sum^J (Y_{ij} - \bar{Y}_i)^2}_{SS_W} + \underbrace{J \sum^I (\bar{Y}_i - \bar{\bar{Y}})^2}_{SS_B} \tag{5.151}$$

where the second and third term attribute sum squares to within and between groups respectively.