Miscellaneous Notes on Regression

Based on SJS and KNN

© by Xia, Tingfeng

Fall 2019, modified on Thursday 5th December, 2019

Preface

Notes for STA302H1F fall offering, 2019 with Prof. Shivon Sue-Chee. These notes are based on the KNN and SJS text, in an aim for better understanding of the course material.

Contents

1 Preliminaries

1.1 Distribution Theories

- Suppose $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, and consider $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$. Then,

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(df=n-1)}$$

- Under the Normal Error SLR model, where $e_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$, and $S^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\mathrm{e}}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ (Different from above!). Then,

$$\frac{(n-2)S^2}{\sigma^2} = \frac{(n-2)S^2/SXX}{\sigma^2/SXX} = \sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{\sigma}\right)^2 \sim \chi^2_{(df=n-2)}$$

- Let $Z \sim N(0,1)$, and $V \sim \chi^2_{(df=v)}$. Assume further that $Z \perp\!\!\!\perp V$, then

$$\frac{Z}{\sqrt{V/v}} \sim t_{(df=v)}$$

- Under the Normal Error SLR model,

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{S^2}{SXX}}} \sim t_{(df=n-2)}$$

- Suppose $V \sim \chi^2_{(df=v)}$, $W \sim \chi^2_{(df=w)}$ and $V \perp\!\!\!\perp W$. Then,

$$\frac{V/v}{W/w} \sim F_{(v,w)}$$

- Suppose $Q \sim t_{(df=v)}$, then

$$Q^2 \sim F_{(1,v)}$$

1.2 Matrix Calculus

Manuel to matrix calculus provided by professor: `http://biostat.mc.vanderbilt.edu/wiki/pub/Main/CourseBios312/chap2.pdf`

1.2.1 Lemma I (Real Valued Fcn Matrix Differentiation)

If $\boldsymbol{\theta}' = (\theta_1, \theta_2, \ldots, \theta_k)$ and $c' = (c_1, c_2, \ldots, c_k)$ is a vector of constant, such that

$$f(\boldsymbol{\theta}) = \mathbf{c}'\boldsymbol{\theta} = \boldsymbol{\theta}'\mathbf{c} = \sum_i c_i\theta_i \in \mathbb{R}$$

is a scalar, then

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{c}$$

1.2.2 Lemma II (Symmetric Quadratic Form)

Let $\mathbf{A}$ be a $k \times k$ ***symmetric*** square matrix. Suppose $f(\boldsymbol{\theta}) = \boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta}$. Then,

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 2\mathbf{A}\boldsymbol{\theta}$$

1.2.3 Lemma II (General Quadratic Form)

Let $\mathbf{A}$ be a $k \times k$ square matrix. Suppose $f(\boldsymbol{\theta}) = \boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta}$. Then, as the general case for the above,

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{\theta}'(\mathbf{A}' + \mathbf{A})$$

1.2.4 Matrix Idempotency

We say that a matrix A is idempotent if and only if $A^2 = AA = A$. It is worth mentioning that projection matrices are idempotent, i.e. you can only project once and projecting the second time makes no change to the already projected result.

2 Simple Linear Regression

2.1 Ordinary Least Square

2.1.1 Simple Linear Regression Models

The cost function to use in this case is the RSS, defined as

$$\text{RSS} = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

We will now derive the OLS estimators as follows.

Derivatives

$$\frac{\partial \,\text{RSS}}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial \,\text{RSS}}{\partial b_1} = -2 \sum_{i=1}^{n} x_i \, (y_i - b_0 - b_1 x_i) = 0$$

Normal Equations are obtained by rearranging

$$\sum_{i=1}^{n} y_i = b_0 n + b_1 \sum_{i=1}^{n} x_i \tag{1}$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 \tag{2}$$

OLS Regressor

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

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

Estimating Variance of Error Term (Using Residuals)

$$\text{Unbiased Estimator} \quad \hat{\sigma}^2 = S^2 = \frac{\text{RSS}}{n-2} = \frac{1}{n-2}\sum_{i=1}^n \hat{e}_i^2$$

Notes

- $\bar{\hat{e}} = 0$, since $\sum \hat{e}_i = 0$ as the least square estimates minimizes RSS. (This is like a minimization goal where derivatives are taken w.r.t $\hat{e}_i$.)
- S^2 has $n-2$ degrees of freedom since we have estimated two parameters, namely β_0 and β_1.

2.2 Inferences on Slope and Intercept

2.2.1 Inference Assumptions

The following assumptions need to be made in order to preform inference

- Y is explained by x through a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + e_i (i = 1, \ldots, n), \text{ i.e., } \mathrm{E}(Y|X = x_i) = \beta_0 + \beta_1 x_i$$

- Independent Errors, $e_i \perp\!\!\!\perp e_j, \forall i \neq j$
- Homoscedasticity, $\mathrm{Var}(e_i) = \sigma^2, \forall i$
- Normal Error: $e|X \sim N\left(0, \sigma^2\right)$

2.2.2 Inference of Slope

Distribution

$$\hat{\beta}_1 | X \sim N\left(\beta_1, \frac{\sigma^2}{SXX}\right)$$

Standardized Test Statistic (Var Known)

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{SXX}} \sim N(0,1)$$

Test Statistic (Var Unknown) Recall that degrees of freedom = sample size - number of mean parameters estimated. Then,

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{SXX}} = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}\left(\hat{\beta}_1\right)} \sim t_{(df=n-2)} \quad \text{where} \quad S^2 = \frac{\sum_i \hat{e}_i^2}{n-2}$$

Confidence Interval (Var Unknown) The $100(1-\alpha)\%$ CI is

$$CI \leftarrow \hat{\beta}_1 \pm t^*_{(\alpha/2, df=n-2)} \times SE(\hat{\beta}_1) \equiv \hat{\beta}_1 \pm t^*_{(\alpha/2, df=n-2)} \frac{S}{\sqrt{SXX}}$$

Distribution Proof Recall OLS regressor for β_1 is

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i \quad \text{where} \quad c_i = \frac{x_i \bar{x}}{SXX}$$

The expectation could be derived as[1]

$$\begin{aligned}
\mathrm{E}\left(\hat{\beta}_1 | X\right) &= \mathrm{E}\left[\sum_{i=1}^{n} c_i y_i | X = x_i\right] \\
&= \sum_{i=1}^{n} c_i E\left[y_i | X = x_i\right] \\
&= \sum_{i=1}^{n} c_i \left(\beta_0 + \beta_1 x_i\right) \\
&= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \\
&= \beta_0 \sum_{i=1}^{n} \left\{\frac{x_i - \bar{x}}{SXX}\right\} + \beta_1 \sum_{i=1}^{n} \left\{\frac{x_i - \bar{x}}{SXX}\right\} x_i \\
&= \beta_1
\end{aligned}$$

and the variance

$$\begin{aligned}
\mathrm{Var}\left(\hat{\beta}_1 | X\right) &= \mathrm{Var}\left[\sum_{i=1}^{n} c_i y_i | X = x_i\right] \\
&= \sum_{i=1}^{n} c_i^2 \,\mathrm{Var}\left(y_i | X = x_i\right) \\
&= \sigma^2 \sum_{i=1}^{n} c_i^2 \\
&= \sigma^2 \sum_{i=1}^{n} \left\{\frac{x_i - \bar{x}}{SXX}\right\}^2 \\
&= \frac{\sigma^2}{SXX}
\end{aligned}$$

Then, since $e_i|X$ are normally distributed, then $y_i = \beta_0 + \beta_1 x_i + e_i, Y_i|X$ is normally distributed. Since $\hat{\beta}_1|X$ is a linear combination of y_i's , $\hat{\beta}_1|X$ is normally distributed. $\mathcal{Q.E.D.}$†

[1] using the fact that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ and $\sum_{i=1}^{n}(x_i - \bar{x})\, x_i = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = SXX$

t-test using PMCC on Bi-variate Normal Recall that

$$\hat{\rho}_{MLE} = r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{SXY}{\sqrt{SXX\ SYY}}$$

Use the null hypothesis that $H_0 : \rho_{XY} = 0$ and alternative $H_1 : \rho_{XY} \neq 0$, then

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{b_1}{\sqrt{\frac{S^2}{SXX}}} \sim t_{(df=n-2)}$$

2.2.3 Inference of Intercept

Distribution

$$\hat{\beta}_0 | X \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)\right)$$

Standardized Test Statistic (Var Known)

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SXX}}} \sim N(0,1)$$

Test Statistic (Var Unknown)

$$Z = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SXX}}} = \frac{\hat{\beta}_0 - \beta_0}{\mathrm{SE}\left(\hat{\beta}_0\right)} \sim t_{(df=n-2)} \quad \text{where} \quad S^2 = \frac{\sum_i \hat{e}_i^2}{n-2}$$

Confidence Interval (Var Unknown)

$$CI \leftarrow \hat{\beta}_0 \pm t^*_{()\alpha/2, df=n-2)} \times SE(\hat{\beta}_0) \equiv \hat{\beta}_0 \pm t^*_{(\alpha/2, df=n-2)} S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SXX}}$$

Distribution Proof Recall that the OLS regressor of β_0 is given by

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

The expectation,

$$\begin{aligned}
\mathrm{E}\left(\hat{\beta}_0 | X\right) &= \mathrm{E}(\bar{y}|X) - \mathrm{E}\left(\hat{\beta}_1 | X\right)\bar{x} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(y_i | X = x_i\right) - \beta_1\bar{x} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(\beta_0 + \beta_1 x_i + e_i\right) - \beta_1\bar{x} \\
&= \beta_0 + \beta_1\frac{1}{n}\sum_{i=1}^{n}x_i - \beta_1\bar{x} \\
&= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0
\end{aligned}$$

and the variance,

$$\begin{aligned}\operatorname{Var}\left(\hat{\beta}_0|X\right) &= \operatorname{Var}\left(\bar{y} - \hat{\beta}_1\bar{x}|X\right) \\ &= \operatorname{Var}(\bar{y}|X) + \bar{x}^2\operatorname{Var}\left(\hat{\beta}_1|X\right) - 2\bar{x}\operatorname{Cov}\left(\bar{y}, \hat{\beta}_1|X\right)\end{aligned}$$

where

$$\operatorname{Var}(\bar{y}|X) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i|X = x_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\operatorname{Var}\left(\hat{\beta}_1|X\right) = \frac{\sigma^2}{SXX}$$

$$\operatorname{Cov}\left(\bar{y}, \hat{\beta}_1|X\right) = \operatorname{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} c_i y_i\right) = \frac{1}{n}\sum_{i=1}^{n} c_i \operatorname{Cov}(y_i, y_i) = \frac{\sigma^2}{n}\sum_{i=1}^{n} c_i = 0$$

Thus,

$$\operatorname{Var}\left(\hat{\beta}_0|X\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)$$

$\mathcal{Q.E.D.}$†

2.3 CI for *Unknown* Population Regression Line

Goal: Find a confidence interval for a unkown population regression line at $X = x^*$. The population regression line is given by

$$\operatorname{E}(Y|X = x^*) = \beta_0 + \beta_1 x^*$$

Distribution To get an estimate of the y value at $X = x^*$, we can use the regression output (evaluate the estimated regression line at $X = x^*$)

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

where we claim it follows the distribution

$$\hat{y}^* = \hat{y}|X = x^* \sim N\left(\beta_0 + \beta_1 x^*, \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX}\right)\right)$$

Proof of Distribution The expectation follows directly from definition, and we will now show that the variance has the claimed value. Notice that $\operatorname{Var}(\hat{\beta}_0|X) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)$, $\operatorname{Var}(\hat{\beta}_1|X) = \frac{\sigma^2}{SXX}$ and $\operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1|X) = \frac{-\bar{x}\sigma^2}{SXX}$, then

$$\begin{aligned}\operatorname{Var}(\hat{y}|X = x^*) &= \operatorname{Var}(\hat{\beta}_0 + \hat{\beta}_1 x^*|X = x^*) \\ &= \operatorname{Var}(\hat{\beta}_0|X = x^*) + \operatorname{Var}(\hat{\beta}_1 x^*|X = x^*) + 2\operatorname{Cov}(\hat{\beta}_1, \hat{\beta}_1 x^*|X = x^*) \\ &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right) + (x^*)^2\frac{\sigma^2}{SXX} + 2x^*\left(\frac{-\bar{x}\sigma^2}{SXX}\right) \\ &= \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX}\right)\end{aligned}$$

$\mathcal{Q.E.D.}$†

Standardized Test Statistic (Var Known)

$$Z = \frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{\sigma\sqrt{\left(\frac{1}{n} + \frac{(x^*-\bar{x})^2}{SXX}\right)}} \sim N(0,1)$$

Test Statistic (Var Known)

$$T = \frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{S\sqrt{\left(\frac{1}{n} + \frac{(x^*-\bar{x})^2}{SXX}\right)}} \sim t_{(df=n-2)} \quad \text{where} \quad S^2 = \frac{\sum_i \hat{e}_i^2}{n-2}$$

Confidence Interval A $100(1-\alpha)\%$ CI for $\mathrm{E}(Y|X=x^*) = \beta_0 + \beta_1 x^*$ is given by

$$\begin{aligned} CI &\leftarrow \hat{y}^* \pm t^*_{(\alpha/2, df=n-2)} S\sqrt{\left(\frac{1}{n} + \frac{(x^*-\bar{x})^2}{SXX}\right)} \\ &\equiv \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t^*_{(\alpha/2, df=n-2)} S\sqrt{\left(\frac{1}{n} + \frac{(x^*-\bar{x})^2}{SXX}\right)} \end{aligned}$$

Do note that this is only valid for x^* values in the range of the original data values of X. Avoid extrapolation.

2.4 Prediction Intervals for Actual Value of Y

Goal Find a prediction interval for the actual value of Y at x^*, a given value of X.

Important Notes

- $E(Y|X=x^*)$, the expected value or average value of Y for a given value x^* of X, is what one would expect Y to be in the long run when $X = x^*$. $E(Y|X=x^*)$ is therefore a fixed but unknown quantity whereas Y can take a number of values when $X = x^*$.
- $E(Y|X=x^*)$, the value of the regression line at $X = x^*$, is entirely different from Y^*, a single value of Y when $X = x^*$. In particular, Y^* need not lie on the population regression line.
- A confidence interval is always reported for a parameter (e.g., $E(Y|X=x^*) = b_0 + b_1 x^*$) and a prediction interval is reported for the value of a random variable (e.g., Y^*).

Difference Between CI and PI (My Thoughts) The intrinsic difference is that: The CI we found above for $E(Y|X=x^*)$ is a CI for a fixed value. We are trying to find, in the long run where can we expect the regression line to lie given infinite samples. However, PI is trying to report for a specific value, possibly a not-already-observed new value, what is the range that it may appear in. PI captures more variability.

Distribution

$$Y^* - \hat{y}^* \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX}\right]\right)$$

Test Statistic (Var Unknown)

$$T = \frac{Y^* - \hat{y}^*}{S\sqrt{\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX}\right)}} \sim t_{(df=n-2)}$$

Derivation of Distribution We base our prediction of Y at $X = x^*$ (which is Y^*) on

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

The error (deviation, to be precise) of our prediction is

$$Y^* - \hat{y}^* = \beta_0 + \beta_1 x^* + e^* - \hat{y}^* = (\mathrm{E}(Y|X = x^*) - \hat{y}^*) + e^*$$

that is, the deviation between $E(Y|X = x^*)$ and y^* plus the random fluctuation e^* (which represents the deviation of Y^* from $E(Y|X = x^*)$). Thus the variability in the error for predicting a single value of Y will exceed the variability for estimating the expected value of Y (because of the random error e^*). We have

$$\mathrm{E}(Y^* - \hat{y}^*) = \mathrm{E}(Y - \hat{y}|X = x^*) = 0$$

and

$$\mathrm{Var}(Y^* - \hat{y}^*) = \mathrm{Var}(Y - \hat{y}|X = x^*) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX}\right]$$

Prediction Interval A $100(1-\alpha)\%$ prediction interval for Y^* (the value of Y at $X = x^*$ is given by)

$$\begin{aligned} PI &\leftarrow \hat{y}^* \pm t_{(\alpha/2, df=n-2)} S\sqrt{\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX}\right)} \\ &\equiv \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{(\alpha/2, df=n-2)} S\sqrt{\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX}\right)} \end{aligned}$$

2.5 Analysis of Variance (ANOVA)

2.5.1 Sum of Squares Decomposition

Define Total Sample Variability

$$\mathrm{SST} = SYY = \sum_i^n (y_i - \bar{y})^2$$

and recall the familiar residual squared sum (Unexplained (or error) variability)

$$\text{RSS} = \sum_i^n (y_i - \hat{y}_i)^2$$

and we define (Sum of Squares explained by the regression model)

$$\text{SSreg } = \sum_i^n (\hat{y}_i - \bar{y})^2$$

Then, the decomposition is

$$\text{SST} = \text{RSS} + \text{SSreg}$$

2.5.2 Test for Zero Slope

***t*-test** Note that for SLR, this is equivalent to the F-test outlined below. Consider the null $H_0 : \beta_1 = 0$ against alternative $H_1 : \beta_1 \neq 0$, then

$$T = \frac{\hat{\beta}_1 - 0}{\text{se}\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1 - 0}{S/\sqrt{SXX}} \overset{H_0}{\sim} t_{(df=n-2)}$$

This is just a specific application of the general t-test that we mentioned earlier, not very interesting.

***F*-test** Assume that $e_i \perp\!\!\!\perp e_j, \forall i \neq j \wedge e_i \sim N(0, \sigma^2), \forall i$. Consider the null $H_0 : \beta_1 = 0$ against alternative $H_1 : \beta_1 \neq 0$, then

$$F = \frac{\text{SSreg}\,/1}{\text{RSS}\,/(n-2)} \overset{H_0}{\sim} F_{1,n-2}$$

2.5.3 Coefficient of Determination

The Coefficient of Determination (R^2) of a regression line is defined as the proportion of the total sample variability in the Y's explained by the regression model, that is

$$R^2 = \frac{\text{SSreg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}}$$

2.5.4 The ANOVA Table

The above F-test, as well as the sum of squares decomposition, could be summarized using the following handy table.

Source of Variation	df	SS	MS = SS/df	F
Regression	1	SSreg	SSreg/1	$F = \frac{SSreg/1}{\text{RSS}/(n-2)}$
Residual	$n-2$	RSS	$\text{RSS}/(n-2)$	
Total	$n-1$	SST		

3 Diagnostics and Transformations for SLR

3.1 Valid and Invalid Data

3.1.1 Residuals

One tool that we can use to validate a regression model is one or more plots of residuals (or standardized residuals). These plots will enable us to assess visually whether an appropriate model has been fit to the data no matter how many predictor variables are used.

Expected Behaviour We expect that the residual graph to have no discern-able pattern and centred at some value (0 in the case of standardized residual). Patterns such as curves, skewness et cetra indicates non-normal residuals. More on this in the below section.

3.1.2 Reading Residual Plots

Criterion One way of checking whether a valid simple linear regression model has been fit is to plot residuals versus x and look for patterns. If no pattern is found then this indicates that the model provides an adequate summary of the data, i.e., is a valid model. If a pattern is found then the shape of the pattern provides information on the function of x that is missing from the model.

Rationale Suppose that the true model is a straight line (which we never know) defined as

$$Y_i = \mathrm{E}\left(Y_\mathrm{i}|X_i = x_i\right) + e_i = \beta_0 + \beta_1 x_i + e_i \tag{3}$$

where

$$e_i = \text{Random error on } Y_i \quad \text{and } E(e_i) = 0$$

and we fit a regression line

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Under the assumption that our regression line is very close to the true model, i.e. $\beta_0 \approx b_0$ and $\beta_1 \approx b_1$), we see

$$\begin{aligned}
\hat{e}_i &= y_i - \hat{y}_i \\
&= \beta_0 + \beta_1 x_i + e_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\
&= \left(\beta_0 - \hat{\beta}_0\right) + \left(\beta_1 - \hat{\beta}_1\right) x_i + e_i \\
&\approx e_i
\end{aligned}$$

which means that our residuals resembles the random error!

3.2 Regression Diagnostics

Categorization

- **X-Direction Outlier, i.e. Leverage Point**: Away from the bulk of data in the x-direction.

  - **Good:** Not much change after removing the data point, i.e. the data point originally was quite close to the regression line although away from the bulk of data in the x direction. *"A good leverage point is a leverage point which is NOT also an outlier."*
  - **Bad, Influential Point:** If its Y-value does not follow the pattern set by the other data points, i.e. a bad leverage point is a leverage point which is also an outlier.

- **Y-Direction Outlier** Trait: large residuals

3.2.1 Leverage Point

Defining The Hat The hat came from yet another representation of the $\hat{y}_i$. Recall that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, where $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, $\hat{\beta}_1 = \sum_{j=1}^n c_j y_j$ and $c_j = \frac{x_j - \bar{x}}{SXX}$. Then we have

$$\begin{aligned}\hat{y}_i &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \\ &= \frac{1}{n}\sum_{j=1}^n y_j + \sum_{j=1}^n \frac{(x_j - \bar{x})}{SXX} y_j (x_i - \bar{x}) \\ &= \sum_{j=1}^n \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX}\right] y_j = \sum_{j=1}^n h_{ij} y_j\end{aligned}$$

where we define

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

Property of The Hat Recall that $\sum_{j=1}^n [x_j - \bar{x}] = n\bar{x} - n\bar{x} = 0$, then

$$\sum_{j=1}^n h_{ij} = \sum_{j=1}^n \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX}\right] = \frac{n}{n} + \frac{(x_i - \bar{x})}{SXX}\sum_{j=1}^n [x_j - \bar{x}] = 1$$

Thus,

$$\hat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j \quad \text{where } h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

Defining Leverage The term $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$ above is commonly known as the leverage of the ith data point. Notice the following in the definition of the leverage h_{ii}

- The second term measures the proportion, in terms of squared deviation in x-direction over sum of square of total deviation in x-direction, of the i-th data point's deviation. When the second term tends to 1, meaning that i-th data point is some extreme outlier in the x-direction, then h_{ii} would close to one, signifying the 'leverage'-ness.
- Recall that $\sum_{j=1}^n h_{ij} = 1$, then when $h_{ii} \cong 1$, $h_{ij} \to 0$ and

  $$\hat{y}_i = 1 \times y_i + \text{ other terms } \cong y_i$$

  which means $\hat{y}_i$ will be very close to y_i, regardless of the rest dataset.
- A point of high leverage (or a leverage point) can be found by looking at just the values of the x's and not at the values of the y's

Average of Leverage For simple linear regression,

$$\text{average}\left(h_{ii}\right) = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{2}{n}$$

Identifying Leverage Rule: x_i is a high leverage (i.e., a leverage point) in a SLR model if

$$h_{ii} > 2 \times \text{average}\left(h_{ii}\right) = 2 \times 2/n = 4/n$$

Dealing with 'Bad' Leverage

- **Remove invalid data points;** Question the validity of the data points corresponding to bad leverage points, that is: Are these data points unusual or different in some way from the rest of the data? If so, consider removing these points and refitting the model without them.
- **Fit a different regression model;** Question the validity of the regression model that has been fitted, that is: Has an incorrect model been fitted to the data? If so, consider trying a different model by including extra predictor variables (e.g., polynomial terms) or by transforming Y and/or x (which is considered later in this chapter).

3.2.2 Standardized Residuals

Problem of Non-constant Variance Recall that

$$h_{ij} = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2} = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{SXX}$$

and (we will show this later)

$$\text{Var}\left(\hat{e}_i\right) = \sigma^2\left[1 - h_{ii}\right]$$

which is indeed non-constant for different data points. When $h_{ii} \cong 1$ (h_{ii} is very close to 1), the i-th data point is a leverage point and

$$\text{Var}(\hat{e}_i) = \sigma^2\left[1 - h_{ii}\right] \approx 0 \quad \text{and} \quad \hat{y}_i \cong y_i$$

The above results intuitively makes sense: When i-th data point is a leverage, $\hat{e}_i$ will be small and it does not vary much (data point close to the estimated regression line).

Derivation of Residual Variance (Not Important) Recall that

$$\hat{y}_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j \quad \text{where } h_{ij} = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2} = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{SXX}$$

Then,

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

Hence,

$$\begin{aligned}\operatorname{Var}(\hat{e}_i) &= \operatorname{Var}\left((1-h_{ii})\,y_i - \sum_{j\neq i} h_{ij}y_j\right)\\ &= (1-h_{ii})^2\,\sigma^2 + \sum_{j\neq i} h_{ij}^2\sigma^2\\ &= \sigma^2\left[1-2h_{ii}+h_{ii}^2+\sum_{j\neq i}h_{ij}^2\right]\\ &= \sigma^2\left[1-2h_{ii}+\sum_{j}h_{ij}^2\right]\end{aligned}$$

Notice that

$$\begin{aligned}\sum_{j=1}^n h_{ij}^2 &= \sum_{j=1}^n\left[\frac{1}{n}+\frac{(x_i-\bar{x})(x_j-\bar{x})}{SXX}\right]^2\\ &= \frac{1}{n}+2\sum_{j=1}^n\frac{1}{n}\times\frac{(x_i-\bar{x})(x_j-\bar{x})}{SXX}+\sum_{j=1}^n\frac{(x_i-\bar{x})^2(x_j-\bar{x})^2}{SXX^2}\\ &= \frac{1}{n}+0+\frac{(x_i-\bar{x})^2}{SXX}\\ &= h_{ii}\end{aligned}$$

So,

$$\operatorname{Var}(\hat{e}_i) = \sigma^2\left[1-2h_{ii}+h_{ii}\right] = \sigma^2\left[1-h_{ii}\right]$$

and

$$\operatorname{Var}(\hat{y}_i) = \operatorname{Var}\left(\sum_{j=1}^n h_{ij}y_j\right) = \sum_{j\neq i}h_{ij}^2\operatorname{Var}(y_j) = \sigma^2\sum_j h_{ij}^2 = \sigma^2 h_{ii}$$

Overcome with Standardization The above problem of each $\hat{e}_i$ having different variances could be overcame by standardizing the residuals. The i-th standardized residual is defined as (notice that the $s=\hat{\sigma}$ is the estimated variance in the SLR settings)

$$r_i = \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}} \quad \text{where } s = \sqrt{\frac{1}{n-2}\sum_{j=1}^n \hat{e}_j^2}$$

Advantages of Standardization

- When points of high leverage exist, instead of looking at residual plots, it is generally more informative to look at plots of standardized residuals since plots of the residuals will have non-constant variance even if the errors have constant variance.

- When points of high leverage do not exist, there is generally little difference in the patterns seen in plots of residuals when compared with those in plots of standardized residuals.
- The other advantage of standardized residuals is that they immediately tell us how many estimated standard deviations any point is away from the fitted regression model.

⋆ **Recognizing Outliers Using Standardized Residuals**

- An **outlier** is a point whose standardized residual falls outside the interval from -2 to 2, i.e. $|r_i| > 2$
- A **Bad Leverage Point** is a leverage point whose standardized residual falls outside the interval from -2 to 2, i.e. $|r_i| > 2 \wedge h_{ii} > \frac{4}{n}$
- A **Good Leverage Point** is a leverage point whose standardized residual falls inside the interval from -2 to 2, i.e. $|r_i| \leq 2 \wedge h_{ii} > \frac{4}{n}$
- **Dealing with large datasets:** In this case, we should change the above criterion to $|r_i| > 4$ and $|r_i| \leq 4$ respectively. This is to give allowance for more occurrence of rare events in a large data set.

Correlation Between Residuals Even if the errors are independent (homogeneous), i.e. $e_i \perp\!\!\!\perp e_j$ $(i \neq j)$, the residuals are still correlated. It can be shown that the covariance and the correlation is given by

$$\operatorname{Cov}(\hat{e}_i, \hat{e}_j) = -h_{ij}\sigma^2 (i \neq j)$$

$$\operatorname{Corr}(\hat{e}_i, \hat{e}_j) = \frac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{ij})}} (i \neq j)$$

Such correlation could be safely ignored in practice. They are usually given raise by inherent correlation such as data collected over time.

Variance of Residuals Above we discussed 'inter-correlation' of residuals. The variance of a single residual is

$$\operatorname{Var}(\hat{e}_i) = (1 - h_{ii})\,\sigma^2$$

3.2.3 Recommendations for Handling Outliers & Leverage

We have discussed multiple ways of assessing outliers and talked about the way to deal with them by removing them. However, it is not always a good idea to delete them for the following reasons:

- Points should not be routinely deleted from an analysis just because they do not fit the model. Outliers and bad leverage points are signals, flagging potential problems with the model.
- Outliers often point out an important feature of the problem not considered before. They may point to an alternative model in which the points are not an outlier. In this case it is then worth considering fitting an alternative model.

3.2.4 Influence of Certain Cases

It can sometimes be the case that certain data points in a data set are drastically controlling the entire regression model (the model has payed too much attention to them). We now develop methods where we measure the "importance" of a data point.

Cook's Distance First, define (recall if already defined) the following notation

- $\hat{y}_{j(i)}$ means the the fitted value of the j-th data point on the regression line obtained by removing the i-th case.
- $S^2 = \frac{1}{n-2}\sum_{j=1}^{n}\hat{e}_j^2$ is the variance (Original MSE) of the total regression model.
- $r_i = \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}}$ where $s = \sqrt{\frac{1}{n-2}\sum_{j=1}^{n}\hat{e}_j^2}$

Then, the Cook's Distance of the i-th data point is given by

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_{j(i)} - \hat{y}_j\right)^2}{2S^2} = \frac{r_i^2}{2}\frac{h_{ii}}{1-h_{ii}}$$

where we should note that D_i may be large due to large r_i, or large h_{ii} or both.

Rule: Cook's Distance

- A point is noteworthy if
$$D_i > \frac{4}{n-2}$$
- In practice, look for gaps in the values of Cook's Distance and not just whether one value exceeds tyhe suggested cut off.

3.2.5 Normality of the Errors

The assumption of normal errors is (especially) needed in small samples for the validity of t-dist based tests and inferences. This assumption is generally checked by looking at the distribution of the residuals or standardized residuals. Recall that the i-th least squares residuals is given by $\hat{e}_i = y_i - \hat{y}_i$. We will now show $\hat{e}_i = e_i - \sum_{j=1}^{n} h_{ij}e_j$. First, in the derivation we will need these two facts

$$\sum_{i=1}^{n} h_{ij} = 1$$

and

$$\sum_{j=1}^{n} x_j h_{ij} = \sum_{j=1}^{n}\left[\frac{x_j}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})x_j}{SXX}\right] = \bar{x} + \frac{(x_i-\bar{x})SXX}{SXX} = x_i$$

We then proceed as follows

$$\begin{aligned}
\hat{e}_i = y_i - \hat{y}_i &= y_i - h_{ii}y_i - \sum_{j\neq i} h_{ij}y_j \\
&= y_i - \sum_{j=1}^{n} h_{ij}y_j \\
&= \beta_0 + \beta_1 x_i + e_i - \sum_{j=1}^{n} h_{ij}\left(\beta_0 + \beta_1 x_j + e_j\right) \\
&= \beta_0 + \beta_1 x_i + e_i - \beta_0 - \beta_1 x_i - \sum_{j=1}^{n} h_{ij}e_j \\
&= e_i - \sum_{j=1}^{n} h_{ij}e_j
\end{aligned}$$

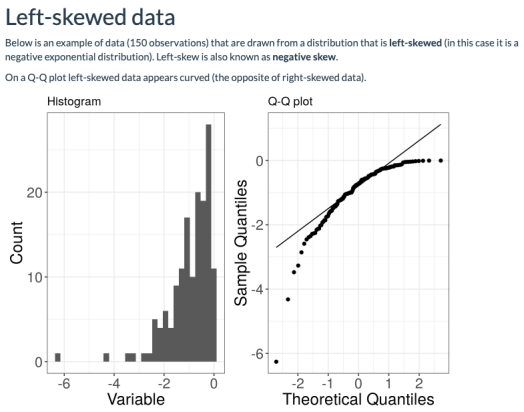
$\mathcal{Q.E.D.}$†

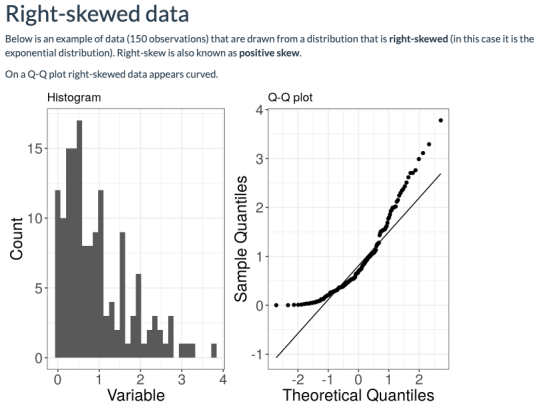
The above result showed that the i-th least squares residual is equal to e_i minus a weighted sum of all the e's. There are two cases to consider,

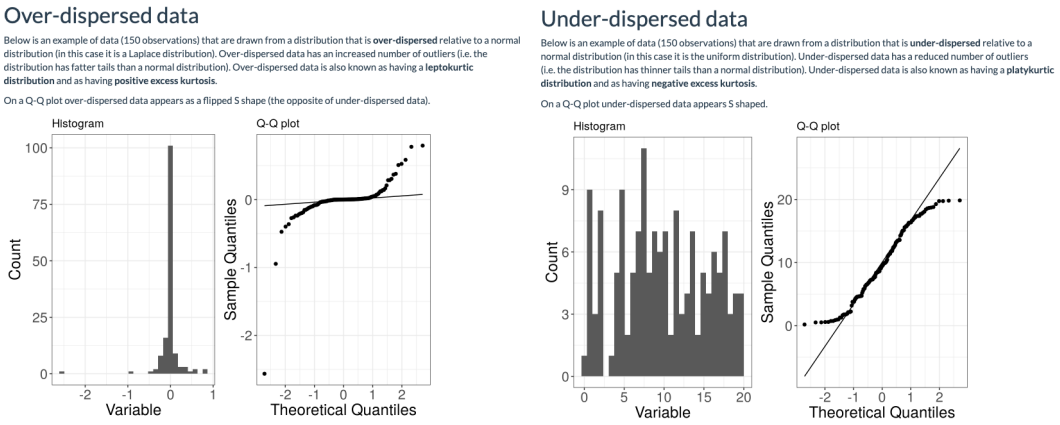
- In small to moderate samples, the second term could dominate the first and first and the residuals can look like they come from a normal distribution even if the errors do not.
- When n is large, the second term in the derived result (thistle coloured) has a much smaller variance than that of the first term and as such the first term dominates the last equation.

Conclusion: For large samples, the residuals can be used to assess normality of the errors.

Assessment Using Normal Q-Q A normal probability plot of the standardized residuals is obtained by plotting the ordered standardized residuals on the vertical axis against the expected order statistics from a standard normal distribution on the horizontal axes. If the resulting plot produces points "close" to a straight line then the data are said to be consistent with that from a normal distribution. On the other hand, departures from linearity provide evidence of non-normality.







The above four are figures that I borrowed from `http://www.ucd.ie/ecomodel/Resources/QQplots_WebVersion.html` which illustrates how to interpret QQ plots with non-normal behaviour.

3.2.6 Constant Variance (Homoscedasticity)

A crucial assumption in any regression analysis is that errors have constant variance. Notice the difference between error and residual, we have demonstrated in **(3.2.2 Standardized Residuals)** that residuals are not of constant variance. There are two general methods that we can adopt to overcome this issue, namely (both of which will be discussed later)

- Transformations
- Weighted Least Squares

Important: *Ignoring non-constant variance when it exists invalidates all inferential tools, including p-values, CI, PI, et cetra!*

Behaviour of Non-Homoscedasticity For example, on the plot explanatory var against standardized residuals, we might see that as x increases, the residuals are more spread out, indicating an increasing trend in the variance.

Checking for Constant Variance To check this, check the plot of

$$|\text{ Residuals }|^{0.5} \text{ against } x \quad \text{or} \quad |\text{ Standardized Residuals }|^{0.5} \text{ against } x$$

The power of 0.5 here is used to reduce skewness in the absolute values. In the above mentioned example where the residuals become more spread out as x increases, the plot $|$ Std Residuals $|^{0.5}$ against x will have an overall increasing trend! This is essentially mirroring all the points to the positive side (and de-skew) to observe a general trend.

3.3 Transformation

3.3.1 Variance Stabilizing Transformations

Goal When non-constant variance exists, it is often possible to transform one or both of the regression variables to produce a model in which the error variance is constant.

Delta Method, Poisson Suppose that $Y \sim \text{Poi}(\mu = \lambda)$ and we want to find the appropriate transformation of Y for stabilizing variance. In this case, square root is the appropriate transformation to apply. We will now justify this choice. Consider the McLauren Series expansion

$$f(Y) = f(\text{E}(Y)) + f'(\text{E}(Y))(Y - \text{E}(Y)) + \dots$$

According to the delta rule, the first order variance term is obtained by taking variance on both sides of the above equation, which yields

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

Using the proposed transformation $f(Y) = Y^{0.5}$ and recall from properties of Poisson Random Variable that $\text{Var}(Y) = \lambda = \text{E}(Y)$, then

$$\text{Var}\left(Y^{0.5}\right) \simeq \left[0.5(\text{E}(Y))^{-0.5}\right]^2 \text{Var}(Y) = \left[0.5\lambda^{-0.5}\right]^2 \lambda = \text{constant}$$

Rule of Thumb: When both Y and X are measured in the same units then it is often natural to consider the same transformation for both X and Y

Hence in this case our regression model would be

$$Y = \beta_0 + \beta_1 x + e$$

where

$$Y \leftarrow \sqrt{Y} \quad \text{and} \quad x \leftarrow \sqrt{x}$$

3.3.2 Logarithms to Estimate Percentage Effects

Consider the regression model

$$\log(Y) = \beta_0 + \beta_1 \log(x) + e$$

The slope,[2]

$$\begin{aligned}
\beta_1 = \frac{\Delta \log(Y)}{\Delta \log(x)} &= \frac{\log(Y_2) - \log(Y_1)}{\log(x_2) - \log(x_1)} = \frac{\log(Y_2/Y_1)}{\log(x_2/x_1)} \\
&\cong \frac{Y_2/Y_1 - 1}{x_2/x_1 - 1} \; (\text{ using } \log(1+z) \cong z \text{ and assuming } \beta_1 \text{ is small }) \\
&= \frac{100\,(Y_2/Y_1 - 1)}{100\,(x_2/x_1 - 1)} = \frac{\%\Delta Y}{\%\Delta x}
\end{aligned}$$

Interpretation We showed above that $\%\Delta Y \simeq \beta_1 \times \%\Delta x$. Thus for every 1% increase in x, the model predicts a $\beta_1\%$ increase in Y (provided β_1 is small).

[2]Notice that the first step is possible since here we are considering the regression straight line

4 Weighted Least Square Regression

4.1 Motivation and Set-Up

Consider the straight line (simple) linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{where} \quad e_i \sim N\left(0, \frac{\sigma^2}{w_i}\right)$$

For the weight w_i, we should note the following

- $w_i \to \infty \implies Var(e_i) \to 0$. In this case, the estimates of the regression parameters β_0, β_1 should be such that the fitted line at x_i should be very close to y_i. (Small variance means more strict in terms of deviation from the regression line, corresponding to a larger emphasis on the i-th data point.)
- If w_i is some small value, then the variance of the i-th data point would be quite large. In this case, we have a loose restriction of the deviation of the i-th data point from the regression line meaning that little emphasis is taken for this data point.
- $w_i \to 0 \implies Var(e_i) \to \infty$. In this case, we have the variance tending to infinity. Meaning that there is absolutely no restriction/emphasis on the i-th data point and it could be simply removed from the set.

We define the cost function, WRSS as

$$\text{WRSS} = \sum_{i=1}^{n} w_i \left(y_i - \hat{y}_{W_i}\right)^2 = \sum_{i=1}^{n} w_i \left(y_i - b_0 - b_1 x_i\right)^2$$

and the estimators $\mathbf{b} = [b_0, b_1]^T$ are derived using MLE.

Intuition behind WRSS This cost function may seem wierd at first glance, but it intuitively makes sense. Notice that when w_i is large, the i-th lost term $w_i \left(y_i - \hat{y}_{W_i}\right)^2$ is payed more emphasis on. On the contrary, when $w_0 \to 0$, the term $\to 0$. (Indeed, when Variance of the term $\to \infty$ we just neglect it.)

4.2 Deriving LS Regressors

Derivatives

$$\frac{\partial \text{WRSS}}{\partial b_0} = -2 \sum_{i=1}^{n} w_i \left(y_i - b_0 - b_1 x_i\right) = 0 \tag{4}$$

$$\frac{\partial \text{WRSS}}{\partial b_1} = -2 \sum_{i=1}^{n} w_i x_i \left(y_i - b_0 - b_1 x_i\right) = 0 \tag{5}$$

Normal Equations Obtained from rearranging the above equations, we will call them Normal Eq1 and Normal Eq2 respectively for later reference.

$$\sum_{i=1}^{n} w_i y_i = b_0 \sum_{i=1}^{n} w_i + b_1 \sum_{i=1}^{n} w_i x_i \tag{6}$$

$$\sum_{i=1}^{n} w_i x_i y_i = b_0 \sum_{i=1}^{n} w_i x_i + b_1 \sum_{i=1}^{n} w_i x_i^2 \tag{7}$$

Rearranging Use Normal Eq1 $\times \sum_{i=1}^{n} w_i x_i$ and Normal Eq2 $\times \sum_{i=1}^{n} w_i$

$$\sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i y_i = b_0 \sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i + b_1 \left(\sum_{i=1}^{n} w_i x_i \right)^2 \tag{8}$$

$$\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i y_i = b_0 \sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i + b_1 \sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i^2 \tag{9}$$

WLS Slope Regressor [3]

$$\hat{\beta}_{1W} = \frac{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i y_i - \sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} \sum_{i=1}^{n} w_i x_i^2 - \left(\sum_{i=1}^{n} w_i x_i \right)^2} \tag{10}$$

$$= \frac{\sum_{i=11}^{n} x_i \left(x_i - \bar{x}_W \right) \left(y_i - \bar{y}_W \right)}{\sum_{i=1}^{n} w_i \left(x_i - \bar{x}_W \right)^2} \tag{11}$$

WLS Intercept Regressor

$$\hat{\beta}_{0W} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} - \hat{\beta}_{1W} \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \bar{y}_w - \hat{\beta}_{1W} \bar{x}_W \tag{12}$$

5 Multiple Linear Regression (Under Construction)

5.1 SLR in Matrix Form

5.1.1 Set-Up

The simple linear regression model is

$$Y = X\beta + e$$

where $Y \in M_{n\times 1}(\mathbb{R}), X \in M_{n\times 2}(\mathbb{R}), \beta \in M_{2\times 1}(\mathbb{R}), e \in M_{n\times 1}(\mathbb{R})$.

5.1.2 The Design Matrix

$$X_{n\times 2} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \implies X\beta(n\times 1) \begin{bmatrix} \beta_0 + X_1\beta_1 \\ \beta_0 + X_2\beta_1 \\ \vdots \\ \beta_0 + X_n\beta_1 \end{bmatrix}$$

[3] Note that $\bar{x}_W = \sum_{i=1}^{n} w_i x_i / \sum_{i=1}^{n} w_i$ and $\bar{y}_W = \sum_{i=1}^{n} w_i y_i / \sum_{i=1}^{n} w_i$

5.1.3 Normal Error Regression Model

Gauss - Markov Conditions

- The errors have zero mean, $E(\mathbf{e}) = \mathbf{0}$
- The errors have constant variance, σ^2
- The errors are uncorrelated, $V(\mathbf{e}) = \sigma^2\mathbf{I}$

Jointly Normal The error terms follow a multivariate normal,

$$\mathbf{e} \sim N_n(\Sigma = \mathbf{0}, \sigma^2 I)$$

5.1.4 OLS in Matrix Form

Consider $\beta = [\beta_0, \beta_1]'$, and the cost function

$$\begin{aligned} RSS(\beta) &= (Y - X\beta)'(Y - X\beta) \\ &= Y'Y + (X\beta)'X\beta - Y'X\beta - (X\beta)'Y \\ &= Y'Y + \beta'\left(X'X\right)\beta - 2Y'X\beta \end{aligned}$$

Derivative

$$\frac{\partial RSS(\beta)}{\partial \beta} = 0 - 2X'Y + 2X'X\beta$$

Normal Equation obtained by setting derivate to zero and re-arrange

$$2X'X\hat{\beta} = 2X'Y$$

OLS Regressor

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

Reduction From Matrix Notation to Scaler Notation We will now show that the matrix form we just derived is equivalent to the form that we discussed/derived earlier in the chapter, where we computed the two estimators separately. First, Let's show some identities that will be used in the derivation.

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots \\ x_1 & x_2 & \cdots x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} = n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{n}\sum_{i=1}^n x_i^2 \end{pmatrix}$$

$$\implies \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \frac{1}{n\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - (\bar{x})^2\right)} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} = \frac{1}{SXX} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

and

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

Thus, the regressor could be break down into

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= \frac{1}{SXX}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}\begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix} \\
&= \frac{1}{SXX}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i y_i - \bar{x}\sum_{i=1}^n y_i \end{pmatrix} \\
&= \begin{pmatrix} \frac{\bar{y}\left\{\sum_{i=1}^n x_i^2 - n\bar{x}^2\right\} - \bar{x}\left\{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}\right\}}{SXX} \\ SXY/SXX \end{pmatrix} \\
&= \begin{pmatrix} \bar{y} - \frac{SXY}{SXX}\bar{x} \\ \frac{SXY}{SXX} \end{pmatrix}
\end{aligned}$$

Indeed, the results that we get confirms that our matrix form is equivalent. $\mathcal{Q}.\mathcal{E}.\mathcal{D}.\dagger$

5.1.5 Properties of OLS Regressors

Expectation: Unbiased Estimator

$$\begin{aligned}
E(\hat{\beta}) &= E((X'X)^{-1}X'Y) \\
&= E(Y)(X'X)^{-1}X' \\
&= (X'X)^{-1}X'E(X\beta + e) \\
&= (X'X)^{-1}X'(X\beta + 0) \\
&= I\beta = \beta
\end{aligned}$$

$\mathcal{Q}.\mathcal{E}.\mathcal{D}.\dagger$

Variance - Covariance Matrix

$$\begin{aligned}
VAR(\hat{\beta}) &= VAR(\underbrace{(X'X)^{-1}X'}_{\text{constant}}Y) \\
&= [(X'X)^{-1}X'] \underbrace{\sigma^2 I}_{VAR(Y)} [(X'X)^{-1}X']' \\
&= \sigma^2 I(X'X)^{-1}\underbrace{X'X(X'X)^{-1}}_{\equiv I} \\
&= \sigma^2(X'X)^{-1} \\
&= \begin{bmatrix} VAR(\hat{\beta}_0) = \frac{\sigma^2\sum x_i^2}{nSXX} & COV(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2\bar{x}}{SXX} \\ COV(\hat{\beta}_1, \hat{\beta}_0) = \frac{-\sigma^2\bar{x}}{SXX} & VAR(\hat{\beta}_1) = \frac{\sigma^2}{SXX} \end{bmatrix}
\end{aligned}$$

5.1.6 The Hat Matrix

Defining the Hat We have

$$\widehat{\mathbf{e}} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \quad \text{and} \quad \widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

where

$$\text{HAT Matrix} \quad \mathbf{H} := \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

Thus, $\widehat{\mathbf{e}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$

Facts About the Hat Matrix

- **H** is symmetric

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

- **H** is idempotent

$$\begin{aligned} H^2 = HH &= X(X'X)^{-1}\underbrace{X'X(X'X)^{-1}}_{\equiv I}X' \\ &= X(X'X)^{-1}X' = H \end{aligned}$$

5.1.7 Properties of the Residuals in Matrix Form

$$\begin{aligned} \hat{e} &= (I-H)Y = (I-H)(X\beta + e) \\ &= (I-H)X\beta + (I-H)e \\ &= IX\beta - HX\beta + (I-H)e \\ &= X\beta - X\underbrace{X(X'X)^{-1}X}_{\equiv I}\beta + (I-H)e \\ &= X\beta - X\beta + (I-H)e \\ &= (I-H)e \end{aligned}$$

Expectation

$$E(\hat{e}|X) = E((I-H)Y|X) = E((I-H)e|X) = \underbrace{E(e|X)}_{\equiv 0}(I-H) = 0$$

Variance - Covariance

$$\begin{aligned} VAR(\hat{e}|X) &= VAR((I-H)e|X) \\ &= (I-H)VAR(e|X)(I-H)' \\ &= (I-H)\sigma^2 I(I-H)' \\ &= (I-H)\sigma^2 I(I-H) \quad \text{since } (I-H) \text{ is symmetric} \\ &= \sigma^2(I-H) \quad \text{since } (I-H) \text{ is idempotent} \end{aligned}$$

5.1.8 ANOVA in Matrix Form

Total Sum Squared (SST)

$$\begin{aligned} SST = \sum (y_i - \bar{y})^2 &= \sum y_i^2 - n\bar{y}^2 \\ &= Y'Y - \frac{1}{n}Y'JY \\ &= Y' \underbrace{\underbrace{(I - \frac{1}{n}J)}_{\text{symmetric square matrix}}}_{\text{quadratic form}} Y \end{aligned}$$

To find the corresponding df, we first check the idempotency of $I - \frac{1}{n}J$. Indeed,

$$(I - \frac{1}{n}J)(I - \frac{1}{n}J) = I^2 - \frac{2}{n}J + \frac{1}{n^2}J^2 = I - \frac{1}{n}J$$

Then, applying $Rank$(idempotent mat) $= Tr$(idempotent mat), we have

$$Rank(I - \frac{1}{n}J) = n - 1 = \text{degrees of freedom of } SST$$

Residual Sum Squared (RSS) Notice that $(I - H)$ is symmetric and idempotent,

$$RSS = \sum \hat{e}_i^2 = \hat{e}'\hat{e} = Y'(I - H)(I - H)Y = Y'(I - H)Y$$

and the corresponding degrees of freedom is

$$Rank(I - H) = Rank(I) - Rank(H) = Tr(I) - Tr(H) = n - 2 = \text{df of MSE}$$

Regression Sum Squared (SSReg)

$$\begin{aligned} SSReg &= SST - RSS \\ &= Y'(I - 1/nJ)Y - Y'(I - H)Y \\ &= Y'IY - Y'1/nJY - Y'IY + Y'HY \\ &= Y'(H - \frac{1}{n}J)Y \end{aligned}$$

and the corresponding degrees of freedom is

$$Rank(H - \frac{1}{n}J) = Rank(H) - Rank(\frac{1}{n}J) = \sum h_{ii} - \sum \frac{1}{n} = 2 - 1 = 1 = \text{df of MSReg}$$

5.1.9 ANOVA Table in Matrix Form

Source	SS	*df*
Regression	$\mathbf{Y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}$	1
Error	$\mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$	$n-2$
Total	$\mathbf{Y}'\left(\mathbf{I} - \frac{1}{\mathbf{n}}\mathbf{J}\right)\mathbf{Y}$	$n-1$

5.2 Estimation and Inference in MLR

5.2.1 The MLR Model

In the familiar SLR settings, we have that $Y = \beta_0 + \beta_1 x + error$, where we had one predictor variable (so called 'simple'). In the MLR settings, we want to add multiple predictors. This leads to the formulation of expectation

$$\mathrm{E}\left(Y|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Thus,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + e_i$$

where e_i is the random fluctuation (or error) in Y_i such that $E(e_i|X) = 0$. In this case the response variable Y is predicted from p predictor variables $X_1, X_2, ..., X_p$ and the relationship between Y and $X_1, X_2, ..., X_p$ is *linear* in the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

5.2.2 OLS Regressors - Expanded Scaler Form

The RSS Cost The least squares regressors of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the value of $b_0, b_1, b_2, \ldots, b_p$ for which the sum of the squared residuals, we define our cost function RSS

$$\mathrm{RSS} = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_p x_{pi}\right)^2$$

Derivatives

$$\frac{\partial \mathrm{RSS}}{\partial b_0} = -2 \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_p x_{pi}\right) = 0$$

$$\frac{\partial \mathrm{RSS}}{\partial b_1} = -2 \sum_{i=1}^{n} x_{1i} \left(y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_p x_{pi}\right) = 0$$

$$\ldots$$

$$\frac{\partial \mathrm{RSS}}{\partial b_p} = -2 \sum_{i=1}^{n} x_{pi} \left(y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_p x_{pi}\right) = 0$$

which will give us a system of $p+1$ equations and $p+1$ variables to optimize on. Notice that this usually requires computers to help in the actual optimization calculations. Hence we should see the much more concise matrix formulation.

5.2.3 OLS Regressors - Matrix Form

Let $Y \in M_{n\times(p+1)}(\mathbb{R}), X \in M_{n\times(p+1)}(\mathbb{R}), \beta \in M_{(p+1)\times 1}(\mathbb{R})$ and $e \in M_{n\times 1}(\mathbb{R})$ given by

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots x_{1p} \\ 1 & x_{21} & \cdots x_{2p} \\ \vdots & \vdots & \\ 1 & x_{n1} & \cdots x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Then the multiple linear regression model in matrix notation could be written as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

Notation Let $\mathbf{x}_i'$ denote the i-th row of the design matrix $\mathbf{X}$, then

$$\mathbf{x}_i' := \begin{bmatrix} 1 & x_{i1} & x_{i2} & \dots & x_{ip} \end{bmatrix} \in M_{1\times(p+1)}(\mathbb{R})$$

This enables us to write

$$\mathrm{E}(Y|X=x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = \mathbf{x}_i'\beta$$

RSS Cost (Matrix Notation) The residual sum of squares as a function of β can be written in matrix form as

$$\begin{aligned} \mathrm{RSS}(\beta) &= (\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta) \\ &= \mathbf{Y}'\mathbf{Y} + (\mathbf{X}\beta)'\mathbf{X}\boldsymbol{\beta} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - (\mathbf{X}\boldsymbol{\beta})'\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} + \beta'\left(\mathbf{X}'\mathbf{X}\right)\beta - 2\mathbf{Y}'\mathbf{X}\beta \end{aligned}$$

Normal Equation is obtained by setting the derivative to zero

$$\left(\mathbf{X}'\mathbf{X}\right)\beta = \mathbf{X}'\mathbf{Y}$$

OLS Regressor

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

Hence, our fitted line is given by

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$$

and the residuals are

$$\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$$

5.2.4 Properties of OLS Regressors

Expectation: Unbiased Estimator

$$\begin{aligned} E(\hat{\beta}) &= E((X'X)^{-1}X'Y) \\ &= E(Y)(X'X)^{-1}X' \\ &= (X'X)^{-1}X'E(X\beta+e) \\ &= (X'X)^{-1}X'(X\beta+0) \\ &= I\beta = \beta \end{aligned}$$

$\mathcal{Q.E.D.}$†

Variance - Covariance Matrix

$$\begin{aligned} VAR(\hat{\beta}) &= VAR(\underbrace{(X'X)^{-1}X'}_{\text{constant}}Y) \\ &= [(X'X)^{-1}X'] \underbrace{\sigma^2 I}_{VAR(Y)} [(X'X)^{-1}X']' \\ &= \sigma^2 I(X'X)^{-1}\underbrace{X'X(X'X)^{-1}}_{\equiv I} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

5.2.5 The Hat Matrix

Defining the Hat We have

$$\widehat{\mathbf{e}} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \quad \text{and} \quad \widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

where

$$\text{HAT Matrix} \quad \mathbf{H} := \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

Thus, $\widehat{\mathbf{e}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$

Facts About the Hat Matrix

- **H** is symmetric

$$H' = (X(X'T)^{-1}X')' = X(X'X)^{-1}X' = H$$

- **H** is idempotent

$$\begin{aligned} H^2 = HH &= X(X'X)^{-1}\underbrace{X'X(X'X)^{-1}}_{\equiv I}X' \\ &= X(X'X)^{-1}X' = H \end{aligned}$$

5.2.6 Properties of the Residuals in Matrix Form

$$\begin{aligned} \hat{e} &= (I-H)Y = (I-H)(X\beta + e) \\ &= (I-H)X\beta + (I-H)e \\ &= IX\beta - HX\beta + (I-H)e \\ &= X\beta - X\underbrace{X(X'X)^{-1}X}_{\equiv I}\beta + (I-H)e \\ &= X\beta - X\beta + (I-H)e \\ &= (I-H)e \end{aligned}$$

Expectation

$$E(\hat{e}|X) = E((I-H)Y|X) = E((I-H)e|X) = \underbrace{E(e|X)}_{\equiv 0}(I-H) = 0$$

Variance - Covariance

$$\begin{aligned} VAR(\hat{e}|X) &= VAR((I-H)e|X) \\ &= (I-H)VAR(e|X)(I-H)' \\ &= (I-H)\sigma^2 I(I-H)' \\ &= (I-H)\sigma^2 I(I-H) \quad \text{since } (I-H) \text{ is symmetric} \\ &= \sigma^2(I-H) \quad \text{since } (I-H) \text{ is idempotent} \end{aligned}$$

5.2.7 Residual Sum of Squares (RSS)

The residual sum of squares as a function of the OLS regressors $\hat{\beta}$ can be written in matrix form as follows

$$\text{RSS} = \text{RSS}(\hat{\beta}) = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \hat{\mathbf{e}}'\hat{\mathbf{e}} = \sum_{i=1}^{n} \hat{e}_i^2$$

5.2.8 Estimating Error Variance

$$\text{Unbiased Estimator} \quad \hat{\sigma}^2 = S^2 = \frac{\text{RSS}}{n-p-1} = \frac{1}{n-p-1}\sum_{i=1}^{n} \hat{e}_i^2$$

5.2.9 CIs and Significance Tests

[4] Assuming that the errors are normally distributed[5] with constant variance, then for each $i = 0, 1, ..., p$ we have the test statistic

$$T_i = \frac{\hat{\beta}_i - \beta_i}{\text{se}(\hat{\beta}_i)} \sim t_{df=(n-p-1)}$$

where we recall that the variance-covariance matrix is $VAR(\hat{\beta}) = \sigma^2(X'X)^{-1}$, and thus

$$\text{se}(\hat{\beta}_i) = \hat{\sigma}^2 \left[(X'X)^{-1}\right]_{ii} = \frac{\left[(X'X)^{-1}\right]_{ii}}{n-p-1}\sum_{i=1}^{n} \hat{e}_i^2$$

5.2.10 ANOVA and Global F-Test

Hypothesis The goal is to test if there is a linear association between the explanatory variables and the explained variable. So the null hypothesis we will use is $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ against the alternative hypothesis that $H_A :$ at least some of the $\beta_i \neq 0$.

Test Statistic Under the assumption that $e_1, ..., e_n$ are independent and normally distributed, it can be shown that the F test statistic is defined as follows and follows a $\mathcal{F}$ distribution.

$$F = \frac{\text{SSreg } /p}{\text{RSS } /(n-p-1)} \sim \mathcal{F}_{(p,n-p-1)}$$

MLR ANOVA Table

Source of Variation	df	SS	MS = SS/df	F
Regression	p	SSReg	SSReg/p	$F = \frac{\text{SSreg } /p}{\text{RSS } /(n-p-1)}$
Residual	$n-p-1$	RSS	$S^2 = \frac{\text{RSS}}{n-p-1}$	
Total	$n-1$	SST = SYY		

[4] **Q:** Why was this not included in the class slides?

[5] It is worth noticing that errors are almost always assumed to be normal when we are trying to do inferences on the regression results (Estimates).

Notes

- R^2, the coefficient of determination of the regression line, is defined as the proportion of the total sample variability in the Y's explained by the regression model, that is

$$R^2 = \frac{\text{SSreg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}}$$

- The above formulation doesn't capture the problem where we have irrelevant predictor variables in the regression equation. To compensate, we define the **adjusted coefficient of determination**, R^2_{adj}

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

  Here $S^2 = \frac{\text{RSS}}{n-p-1}$ is an unbiased estimate of $\sigma^2 = \text{Var}(Y_i) = \text{Var}(e_i)$ while $SST/(n-1)$ is an unbiased estimate of $\sigma^2 = \text{Var}(Y_i)$ when $\beta_1 = \beta_2 = ... = \beta_p = 0$. Thus, when comparing models with different numbers of predictors one should use R^2_{adj} rather than R^2.

- The F-test is always used first to test for the existence of a linear association between Y and ANY of the p x-variables. If the F-test is significant then a natural question to ask is

  > For which of the p x-variables is there evidence of a linear association with Y? To answer this question we could perform p separate t-tests of $H_0 : b_1 = 0$. However, as we shall see later there are problems with interpreting these t-tests when the predictor variables are highly correlated.

  The above problem could be addressed by the **partial F-test**[6] which is defined below.

Extra Notes on R-Squared & Adjusted R-Squared

- As $p \uparrow$, we know
  - SST remains the same
  - SSReg remains the same or increases
  - RSS remains the same or decreases

  Thus, $R^2 \uparrow$ regardless.

- Notice that if we use the normal R-Squared (R^2), then the value of R^2 will always increase regardless of the usefulness of extra predictors added when we increase p. That is, when p gets larger, R^2 will *always* increase, which makes this statistics useless. (Not helpful in telling whether additional predictors are useful for explaining the response)

- As $p \uparrow$, we know $(n-p-1) \downarrow$. Then $\frac{n-1}{n-p-1} \uparrow$ and thus $R^2_{adj} \downarrow$.

- It is always the case that $R^2_{adj} < R^2$

[6] Notice that here the partial means we are not testing all the β_1 up to β_p's but only a subset of them.

5.2.11 Partial F-Test

Goal Test whether a specified subset of the predictors have regression coefficients equal to zero.

Hypothesis Suppose that we are interested in testing

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \text{ where } k < p$$
$$\text{i.e., } Y = \beta_0 + \beta_{k+1}x_{k+1} + \ldots + \beta_p x_p + e \text{ (reduced model)}$$

against

$$H_A : H_0 \text{ is not true}$$
$$\text{i.e., } Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{k+1}x_{k+1} + \ldots + \beta_p x_p + e \text{ (full model)}$$

This can be done using the F-Test. Define

$$\text{RSS(Full)} := \text{RSS under the full model}$$
$$\text{RSS(reduced)} := \text{RSS under the reduced model}$$

Test Statistics Then, the F-statistic is given by

$$\begin{aligned} F &= \frac{(\text{RSS(reduced)} - \text{RSS(full)})/\left(df_{\text{reduced}} - df_{\text{full}}\right)}{\text{RSS(full)}/df_{\text{full}}} \\ &= \frac{(\text{RSS(reduced)} - \text{RSS(full)})/k}{\text{RSS(full)}/(n-p-1)} \overset{H_0}{\sim} \mathcal{F}_{(k,n-p-1)} \end{aligned}$$

The reduction on the second line above is done by noticing that the reduced model has $p+1-k$ predictors and thus $(df_{\text{reduced}} - df_{\text{full}}) = [n-(p+1-k)] - [n-(p+1)] = k$.

Using R Output To do the above test, we need the ANOVA tables from the reduced and full models in R.

Partial F-Test Interpretation The above partial F-Test tries to capture the idea of "Is the unexplained variation reduced by a significant amount when the predictors are added to the model?".It is important to remember which number is bigger than which in the above calculations. We have, in general,

- SST(full) = SST(Reduced)
- SSReg(full) $\geq$ SSReg(reduced)
- RSS(full) $\leq$ RSS(reduced)

Notice that the third inequality could be simply remembered as "denominator of the F statistics need to be positive", i.e. $(\text{RSS(reduced)} - \text{RSS(full)}) \geq 0 \implies \text{RSS(reduced)} \geq \text{RSS(full)}$

Rationale for Above Relationship

- $\text{RSS(full)} = \min_{\beta \in \mathbb{R}^{p+1}} (Y - X\beta)'(Y - X\beta)$
- $\text{RSS(full)} = \min_{\beta \in \mathbb{R}^{k+1}} (Y - X\beta)'(Y - X\beta)$
- The minimum in a higher dimensional space is always the same or less.

Special Case of $k = 1$ In this special case of the Partial F-test, we are testing the whether some specific β_i has the value of one, i.e. is irrelevant to the explained variable. A catch here is that the order in which individual predictors are added to the model in R is important! We can just use a single summary output of the full model if the variable whose coefficient is β_i was listed last in the `lm()` call.

5.2.12 Combining Global F-test with *t*-tests

Case A: If the global F-test is significant, then

- If all or some f the t-tests are significant, then there exists some useful explanatory variable for the predicted variable.
- If all t-test are *not* significant, then there is an indication of multicollinearity[7], i.e. strongly correlated X's. This implies that individual X do not contribute to the prediction of Y over and above other X's.

Case B: If the Global F-Test is NOT significant, then

- If all t-tests are not significant, then none of the listed predictor variable contribute to the prediction of Y.
- If some of the t-tests are significant, then
  1. The model has no predictive ability. Likely, if there are many predictors, there are type I errors[8] in the t-tests.
  2. The predictors are poorly chosen. The contribution of one useful predictor among many poor ones may not be enough for the model (Global F-test) to be significant.

5.3 Analysis of Covariance (ANCOVA)

We shall first discuss the setup of ANCOVA, which is stated as follows. Consider the situation in which we want to model a response variable, Y based on a continuous predictor, x and a dummy variable, d. Suppose that the effect of x on Y is linear. This situation is the simplest version of what is commonly referred as Analysis of Covariance, since the predictors include both quantitative variables (i.e., x) and qualitative variables (i.e., d).

5.3.1 Coincident Regression Lines

This is the simplest case of all, which happens when the categorical dummy variable has no effect on Y, that is

$$Y = \beta_0 + \beta_1 x + e \quad \text{for } b \in \{0, 1\}$$

[7] Discussed below in 6.3

[8] Type I errors refers to the error caused when we rejected hypotheses that are actually true.

5.3.2 Parallel Regression Lines

In this case, the dummy variable produces so called *additive changes* to the regression model. The model takes the form

$$Y = \beta_0 + \beta_1 x + \beta_2 d + e = \begin{cases} Y = \beta_0 + \beta_1 x + e & \text{when } d = 0 \\ Y = \beta_0 + \beta_2 + \beta_1 x + e & \text{when } d = 1 \end{cases}$$

Here the regression coefficient β_2 measures the additive change in Y due to the dummy variable.

5.3.3 Regression Lines w/ Equal Intercepts & Different Slopes

In this model, the dummy variable only changes the size of the effect of x on Y. This is described by the following formulation,

$$Y = \beta_0 + \beta_1 x + \beta_3 d \times x + e = \begin{cases} Y = \beta_0 + \beta_1 x + e & \text{when } d = 0 \\ Y = \beta_0 + (\beta_1 + \beta_3) x + e & \text{when } d = 1 \end{cases}$$

5.3.4 Unrelated Regression Lines

This is the most general case of all, where the dummy variable produces an additive change in Y and also changes the size of the effect of x on Y. The formulation breaks into

$$Y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 d \times x + e = \begin{cases} Y = \beta_0 + \beta_1 x + e & \text{when } d = 0 \\ Y = \beta_0 + \beta_2 + (\beta_1 + \beta_3) x + e & \text{when } d = 1 \end{cases}$$

In this formulation, the regression coefficient

- β_2 measures the additive change in Y due to the dummy variable.
- β_3 measures the change in size of the effect of x on Y due to the dummy variable.

6 Diagnostics and Transformations for MLR

6.1 Regression Diagnostics for Multiple Regression

6.1.1 Leverage Points in Multiple Regression

Data points which exercise considerable influence on the fitted mode are called leverage points. Leverage is a measurement of the extent to which the fitted regression model is attracted by the given data point. We are interested in the relationship of the fitted values $\hat{\mathbf{Y}}$ and $\mathbf{Y}$. We should recall that from the previous section $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$.

Popular Rule For Identifying Leverage Points We say the i-th point is a point of high leverage in a MLR model with p predictors if

$$h_{ii} > 2 \times \text{average}\left(h_{ii}\right) = 2 \times \frac{(p+1)}{n}$$

6.2 Properties of Residuals in MLR

Recall that $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, $\mathrm{E}(\hat{\mathbf{e}}|\mathbf{X}) = 0$, and $\mathrm{var}(\hat{\mathbf{e}}|\mathbf{x}) = \sigma^2(\mathbf{I} - \mathbf{H})$.

Standardized Residuals The i-th least squares residual have variance given by

$$\mathrm{Var}\left(\hat{e}_i\right) = \sigma^2\left[1 - h_{ii}\right]$$

and thus the standardized residual is

$$r_i = \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}} \quad \text{where} \quad s = \hat{\sigma} = \sqrt{\frac{1}{n-(p+1)}\sum_{j=1}^{n}\hat{e}_j^2}$$

Fences As a common practice of labelling points as outliers is to check if the standardized residuals, as calculated above, is in the range $(-\infty, -2] \cup [2, \infty)$. Also notice when we have a large set of data, the fences should be changed to ± 4.

Residual Plot When a *valid* model has been fit to the data, the residuals r_i's against *any* predictor or *linear combination* of predictors (such as the fitted values) must be

- A random scatter of points around the horizontal axis.
- Constant variability as we look along the horizontal axis. **Note:** If we observe a pattern in the residual plot, then we should consider adding extra predictor variables, since clearly not all variation has been explained.

6.2.1 Classification

- **Outlier y:** A point x_i is an outlier in the y direction if $|r_i| > 2$. It is not an outlier otherwise.
- **Leverage Points:** A point x_i is a leverage point if $h_{ii} > \frac{2\times(p+1)}{n}$
  1. **Bad, Influential:** means Leverage + Outlier
  2. **Good:** means Leverage but not Outlier

6.3 Box-Cox Transformation

- **Note:** Details not needed.
- One of the most cited papers in Statistics - Box and Cox (1964)
- A general method for transforming a strictly positive response variable
- Aims to find transformation that makes the transformed variable close to normally distributed
- Considers a family of power transformation
- Based on maximizing a likelihood function

6.4 Added Variable Plots

Suppose that our current model is

$$Y = X\beta + e \qquad \text{(modelYX)}$$

and we are considering the introduction of an additional predictor variable, Z, that is, our new model is

$$Y = X\beta + Z\alpha + e \qquad \text{(modelYXZ)}$$

and the added-variable plot is obtained by plotting the residuals from (modelYX) against the residuals from the model

$$Z = X\delta + e \qquad \text{(modelZX)}$$

Rationale for using such method

- To visually assess the effect of each predictor, having adjusted for the effects of the other predictors
- To visually estimate α
- Can be used to identify points which have undue influence on the least squares estimate of α

6.5 Multi-Collinearity

Multicollinearity occurs when explanatory variables are highly correlated. In such case,

1. It is difficult to measure the individual influence of one of the predictors on the response.
2. The fitted equation is unstable
3. The estimated regression coefficients vary widely from data set to data set, even if the data sets are similar, and depending on which predictor is included in the model.
4. The estimated regression coefficients may even have opposite sign than what is expected.
5. When some X's are perfectly correlated, we can't acquire the estimate $\hat{\beta}$ since $X'X$ is singular.
6. Even when in the case where $X'X$ is close to singular, its determinant will be close to zero and the standard errors of estimated coefficients will be large.

6.5.1 Variance Inflation Factors (VIFs)

For the general MLR model,

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + e$$

Then, we have

$$\text{VAR}(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma^2}{(n-1)S_{x_j}^2} \quad \text{where } j = 1, ..., p$$

The j-th Variance Inflation Factor is defined as $\text{VIF}_j := \frac{1}{1-R_j^2}$. Common identification cut-off is 5.

7 Variable Selection

- Variable selection methods aim to choose the subset of the explanatory variables that is the best in a given sense
- **Overfitting:** occurs when too many predictors are in the final regression model. Essentially, we have learnt too much information about some data set and may not be able to generalize in future predictions. The term **under-fitting** refers to the opposite of above.
- In general, there is a bias variance trade-off: when we add more predictors to a valid model,
  - the bias of the predictions gets smaller,
  - but the variance of the estimated coefficients gets larger.
- **Thoughts:** Might be helpful to think of this in the following way: when we add more and more predictors to the regression model, we are capturing more and more information from the data set. To some extent, we will also capture inherent variability in the data which means we will have high variance. However, indeed we learnt the training data, at least, well so the bias is low. The reverse of the argument says about having a model that has too few predictors that fails to explain pattern in the data.

7.1 Information Criterion

7.1.1 Likelihood-based Criteria

Defining Likelihood Suppose that $y_i, x_{1i}, \ldots, x_{pi}, i = 1, \ldots, n$ are observed values of normal random variables and

$$y_i | x_{1i}, \ldots, x_{pi} \sim \mathcal{N}\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}, \sigma^2\right)$$

Thus, the *conditional density* of y_i given $x_{1i}, \ldots, x_{pi}$ is given by

$$f\left(y_i | x_{1i}, \ldots, x_{pi}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y_i - \{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\}\right)^2}{2\sigma^2}\right)$$

Assuming that the n observations are independent, then the likelihood function of the unknown parameters $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$ given Y is given by

$$\begin{aligned} L\left(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2 | Y\right) &= \prod_{i=1}^{n} f\left(y_i | x_i\right) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y_i - \{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\}\right)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\}\right)^2\right) \end{aligned}$$

Log-Likelihood Function is given by

$$\log L\left(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2 | Y\right) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\}\right)^2$$

Equivalence of MLE and OLS We notice that above only the the third term contains the regression parameters $\beta_0, \beta_1, \ldots, \beta_p$. The MLEs of $\beta_0, \beta_1, \ldots, \beta_p$ can be obtained by minimizing the third term only, which is equivalent to minimizing the RSS. Hence, MLEs, of $\beta_0, \beta_1, \ldots, \beta_p$ are equal to the least squares estimates .

Log-Like w/ MLR Substituting the least square estimates of $\beta_0, \beta_1, \ldots, \beta_p$, the log-likelihood function is rewritten as

$$\log L\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \sigma^2 | Y\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}RSS$$

where RSS is defined as $RSS := \sum_{i=1}^{n}\left(y_i - \{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\}\right)^2$. Solving for the MLE of σ^2, we get

$$\sigma^2_{MLE} = \frac{RSS}{n}$$

which differs slightly from the unbiased estimate of σ^2, namely, $S^2 = RSS/(n-p-1)$. Substituting the MLE of σ^2 into the expression for the log-like, we find that the likelihood associated with the maximum likelihood estimates is given by

$$\log L\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2 | Y\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{RSS}{n}\right) - \frac{n}{2}$$

7.1.2 Akaike's Information Criterion (AIC)

Goal Balance goodness-of-fit of the model and the complexity of the model, in terms of number of predictors.

AIC Defined

$$\text{AIC} = 2\left[-\log\left(L\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2 | Y\right)\right) + K\right] \qquad \text{where } K = p + 2$$

One can think of the $K = p + 2$ as a 'degree of freedom', since here we have $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$ in total of $p + 2$ estimated values in the model.

AIC in R

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

Rule The smaller the value of AIC the better the model.

7.1.3 AIC - Corrected

Goal When the sample size is small, or when the number of parameters estimated is a moderate to large fraction of the sample size, it is well-known that AIC has a tendency for over-fitting since the penalty for model complexity is not strong enough. As such, the AIC_C is developed to address this issue.

AIC_C Defined

$$\text{AIC}_\text{C} = -2\log(L(\hat{\theta}|Y)) + 2K + \frac{2K(K+1)}{n-K+1} = \text{AIC} + \frac{2K(K+1)}{n-K+1}$$

when $K = p+2$, we have the equivalent form

$$\text{AIC}_\text{C} = \text{AIC} + \frac{2(p+2)(p+3)}{n-p-1}$$

When To Use? In general, use AIC_C over AIC since when $n \to \infty$, AIC_C converges to AIC. It is recommended that AIC_C should be used over AIC if $\frac{n}{K} > 40$.

7.1.4 Bayesian Information Criterion (BIC)

BIC Defined The BIC is defined as,

$$\text{BIC} = -2\log\left(L\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2|Y\right)\right) + K\log(n) \qquad \text{where } K = p+2$$

The BIC is formulated quite similarly o the AIC we defined above, notice that we replaced the complexity penalty of $2K$ in AIC with $K\log n$ here in BIC. (Notice that the factors here are 2 and $\log n$ respectively.) We see that when $n \geq 8$, $\log n \geq 2$ and hence when sample size is larger than 8, BIC penalizes more to the model complexity. This results in BIC favouring simpler models than AIC

BIC in R

$$\text{BIC} = n\log\left(\frac{RSS}{n}\right) + p\log(n)$$

7.1.5 Data Strategy

A popular data strategy is to

1. Calculate $R^2_{\text{adj}}, \text{AIC}, \text{AIC}_C,$ and BIC, and then
2. compare the models which minimize AIC, AIC_C, and BIC with the model that maximizes R^2_{adj}

7.2 Stepwise Regression

7.2.1 Terminologies

If there are k terms that can be added to the mean function apart from the intercept, then there are 2^k possible regression equations. In general, we have the following terminologies

- **Backward Elimination:** starts with all the potential terms in the model, then removes the term with the largest p-value each time to give a smaller information criterion.
- **Forward Selection:** starts with no term in the model, then adds one term at a time (with the smallest p-value) until no further terms can be added to produce a smaller information criterion.
- **Stepwise regression:** alternates forward steps with backward steps.

7.2.2 Interpretation

- Backward elimination and forward selection considers at most $k + (k-1) + ... + 1 = \frac{k(k+1)}{2}$ of the 2^k possible predictor subsets.
- Stepwise regression can consider more subsets than the backward or forward methods.
- The idea is to end up with a model where no variables are redundant given the other variables in the model. We have a term for this: "parsimonious".
- Often, backward elimination and forward selection will produce the same final model.
- **Selection overstates significance**
  - estimates of regression coefficients are biased
  - p-values from F and t-tests are generally smaller than their true values.

7.3 Penalized Regression

Penalized Linear Regression performs vriable selection and regression coefficient estimation simultaneously. It can be formulated as a constrained OLS optimization problem, with the cost function

$$\mathcal{J} := \sum_{i=1}^{n} \left(Y_i - \boldsymbol{\beta}'\mathbf{x}_i\right)^2 + \sum_{j=1}^{p} p_\lambda(.) \qquad \text{optimized at } \min_{\beta_p} \mathcal{J}$$

In the above formulation, the $p(\cdot)$ is the penalty function and $\lambda \geq 0$ is the penalty (hyper-)parameter which we have to tune. When $\lambda = 0$, this is just the familiar OLS. We have two common choices for the penalty function,

- **(L2-Norm Penalty) Ridge:** $p_\lambda = \lambda\beta_j^2$
- **(L1-Norm Penalty) Lasso:** $p_\lambda = \lambda\left|\beta_j\right|$

8 Selected Properties, Formulae, and Theorems

This section contains various properties mentioned in the slides/book. They may or may not have appeared in precious sections.

8.1 Properties of Fitted Regression Line

- $\sum_{i=1}^{n} \hat{e}_i = 0$
- $RSS = \sum_{i=1}^{n} \hat{e}_i^2 \neq 0$, generally. Except for when we have perfect fit.
- $\sum_{i=1}^{n} \hat{e}_i x_i = 0$
- $\sum_{i=1}^{n} \hat{e}_i \hat{y}_i = 0$
- $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$

8.2 Rules of Expectation

- $E(a) = a, \forall a \in \mathbb{R})$
- $E(aY) = aE(Y)$
- $E(X \pm Y) = E(X) \pm E(Y)$
- $X \perp\!\!\!\perp Y \implies E(EY) = E(X)E(Y)$
- Tower Rule: $E(Y) = E(E(Y|X))$

8.3 Variance and Covariance

- $V(a) = 0, \forall a \in \mathbb{R}$
- $V(aY) = a^2 V(Y)$
- $\mathrm{Cov}(X, Y) = \mathrm{E}\{(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\} = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$
- $\mathrm{Cov}(Y, Y) = V(Y)$
- $\mathrm{V}(Y) = \mathrm{V}[\mathrm{E}(Y|X)] + \mathrm{E}[\mathrm{V}(Y|X)]$
- $\mathrm{V}(X \pm Y) = \mathrm{V}(X) + \mathrm{V}(Y) \pm 2\,\mathrm{Cov}(X, Y)$
- $\mathrm{Cov}(X, Y) = 0$, if X and Y are independent
- $\mathrm{Cov}(aX + bY, cU + dW) = ac\,\mathrm{Cov}(X, U) + ad\,\mathrm{Cov}(X, W) + bc\,\mathrm{Cov}(Y, U) + bd\,\mathrm{Cov}(Y, W)$
- Correlation: $\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{V}(X)\mathrm{V}(Y)}}$

8.4 The Theorem of Gauss-Markov

Under the conditions of the simple linear regression model, the OLS regressors are BLUE ("Best Linear Unbiased Estimators").

- Best - obtains the minimum variance among all unbiased linear estimators.
- Linear - Linear in the parameter space. That is, feature maps are linear, although the actual curve of regression is 'non-linear'.
- Unbiased - The estimators are unbiased, namely $\hat{\beta}_0, \hat{\beta}_1$.
- Estimator - Estimators $\hat{\beta}_0, \hat{\beta}_1$ for β_0 and β_1 respectively.

8.5 Matrix Form Rules

8.5.1 Summations

Consider $\mathbf{A}, \mathbf{B}$ as compatible matrices where appropriate and $k \in \mathbb{R}$ then

- $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
- $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- $\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB}$
- $k(\mathbf{A} + \mathbf{B}) = k\mathbf{A} + k\mathbf{B}$

8.5.2 Transpositions

- $(\mathbf{A}')' = \mathbf{A}$
- $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$
- $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$
- $(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$

8.5.3 Inversions

- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- $(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$
- $\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$
- $\left(\mathbf{A}^{T}\right)^{-1} = \left(\mathbf{A}^{-1}\right)^{T}$
- $[(\mathbf{X}'\mathbf{X})^{-1}]' = [(\mathbf{X}'\mathbf{X})']^{-1}$ (**)

8.5.4 Idempotency

- If A is an ***idempotent matrix***, i.e. $A^2 \equiv A$, then $Tr(A) = Rk(A)$.
- $A \in M_{n\times n}(\mathbb{R})$ is idempotent *if and only if*

$$Rk(A) + Rk(I_{n\times n} - A) = n$$

8.5.5 Other Misc

- The trace of a composition of linear operators has a cyclic property, $Tr(ABC) = Tr(CAB) = Tr(BCA)$. Notice that arbitrary permutation is not valid for this rule.

8.5.6 Covariance Matrix

- The variance-covariance matrix of a random vector $\mathbf{Y}$ is a symmetric, positive semi-definite matrix, defined as

$$\begin{aligned}\operatorname{Var}(\mathbf{Y}) &= \mathrm{E}\left[(\mathbf{Y} - \mathbf{E}(\mathbf{Y}))(\mathbf{Y} - \mathbf{E}(\mathbf{Y}))'\right] \\ &= E\begin{pmatrix} (Y_1 - E(Y_1))^2 & (Y_1 - E(Y_1))(Y_2 - E(Y_2)) & \cdots \\ (Y_2 - E(Y_2))(Y_1 - E(Y_1)) & (Y_2 - E(Y_2))^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}\end{aligned}$$

- Let $\mathbf{A}$ be a square matrix of constants

$$\begin{aligned}\operatorname{Var}(\mathbf{AY}) &= \mathrm{E}\left[(\mathbf{AY} - \mathbf{E}(\mathbf{AY}))(\mathbf{AY} - \mathbf{E}(\mathbf{AY}))'\right] \\ &= \mathbf{E}\left[\mathbf{A}(\mathbf{Y} - \mathbf{E}(\mathbf{Y}))(\mathbf{Y} - \mathbf{E}(\mathbf{Y}))'\mathbf{A}'\right] \\ &= \mathbf{AE}\left[(\mathbf{Y} - \mathbf{E}(\mathbf{Y}))(\mathbf{Y} - \mathbf{E}(\mathbf{Y}))'\right]\mathbf{A}' \\ &= \mathbf{AVar}(\mathbf{V})\mathbf{A}'\end{aligned}$$