Contents

I Weisberg 3

1 Scatter Plots and Regression 4

2 Simple Linear Regression 5
2.1 Ordinary Least Squares Estimation 5
2.2 Estimating the Variance 6
2.3 Confidence Intervals and t-Tests 6
2.4 Prediction 6
2.5 Fitted Values 7
2.6 Coefficient of Determination 7
2.7 Residuals 7

3 Multiple Regression 8
3.1 Adding a Regressor to an SLR Model 8
3.2 The MLR Model 8
3.3 Predictors and Regressors 8
3.4 Ordinary Least Squares 9
3.4.1 The Errors e 9
3.4.2 OLS Estimators 10

II Class Notes 11

4 Notes 2 Matrix Algebra 12
4.1 Orthogonal 12
4.2 Rules of Matrix Operation 12
4.3 Linear Dependence and Rank 13
4.4 Determinants 13
4.5 Positive Definite and Semidefinite Matrices 13
4.6 Inverses 14
4.7 Eigenvalues and Eigenvectors 14
4.8 Finding Eigenvectors and Eigenvalues 15
4.9 Properties of Eigenvalues and Eigenvectors 15
4.10 Random Vectors and Matrices 16

4.11 Multivariate Normal Distribution 16
4.12 Facts about the multivariate normal distribution 17

5 GLM Estimation and Testing 18
5.1 GLH 18
5.2 Estimability of a Parameter 19
5.3 Testability of a Hypothesis 19
5.4 Computation of Test Statistic and p-value 20
5.5 Wald Tests 20
5.6 F-Tests 21

6 Some Distributional Results for the GLM 22
6.1 A Full Rank Basis For Less Than Full Rank Models 22

7 Multiple Regression General Consideration 24

Part I

Weisberg

Chapter 1

Scatter Plots and Regression

Chapter 2

Simple Linear Regression

2.1 Ordinary Least Squares Estimation

2.1 ORDINARY LEAST SQUARES ESTIMATION **23**

Table 2.1 Definitions of Symbols[a]

Quantity	Definition	Description
$\bar{x}$	$\sum x_i/n$	Sample average of x
$\bar{y}$	$\sum y_i/n$	Sample average of y
SXX	$\sum(x_i-\bar{x})^2=\sum(x_i-\bar{x})x_i$	Sum of squares for the xs
SD_x^2	$\mathrm{SXX}/(n-1)$	Sample variance of the xs
SD_x	$\sqrt{\mathrm{SXX}/(n-1)}$	Sample standard deviation of the xs
SYY	$\sum(y_i-\bar{y})^2=\sum(y_i-\bar{y})y_i$	Sum of squares for the ys
SD_y^2	$\mathrm{SYY}/(n-1)$	Sample variance of the ys
SD_y	$\sqrt{\mathrm{SYY}/(n-1)}$	Sample standard deviation of the ys
SXY	$\sum(x_i-\bar{x})(y_i-\bar{y})=\sum(x_i-\bar{x})y_i$	Sum of cross-products
s_{xy}	$\mathrm{SXY}/(n-1)$	Sample covariance
r_{xy}	$s_{xy}/(\mathrm{SD}_x\mathrm{SD}_y)$	Sample correlation

[a]In each equation, the symbol Σ means to add over all n values or pairs of values in the data.

Minimize SSR
5 Assumptions: HILE Gauss

$$\hat{\beta}_1=\frac{SXY}{SXX}=r_{xy}\frac{SD_y}{SD_x}=r_{xy}\left(\frac{SYY}{SXX}\right)^{1/2}$$

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

2.2 Estimating the Variance

Residual Mean Square $\hat{\sigma^2} = \frac{RSS}{df}$ ($df = n - 2$ for SLR)
$RSS = SYY - \frac{SXY^2}{SXX} = SYY - \hat{\beta}_1^2 SXX$
$sig\hat{m}a^2 \sim \frac{\sigma^2}{n-2}\chi^2(n-2)$
$V\hat{a}r(\hat{\beta}_1|X) = \hat{\sigma^2}\frac{1}{SXX}$ $V\hat{a}r(\hat{\beta}_0|X) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)$
$se(\hat{\beta}_1|X) = \sqrt{V\hat{a}r(\hat{\beta}_1|X)}$

2.3 Confidence Intervals and t-Tests

Intercept
95% CI:
$\hat{\beta}_0 - t(\alpha/2, n-2)se(\hat{\beta}_0|X) \leq \beta_0 \leq \hat{\beta}_0 + t(\alpha/2, n-2)se(\hat{\beta}_0|X)$
$H_0 : \beta_0 = B_0^*$ β_1 arbitrary
$H_A : \beta_0 \neq B_0^*$ β_1 arbitrary
t-statistic $t = \frac{\hat{\beta}_0 - \beta_0^*}{se(\hat{\beta}_0|X)}$
Slope:
95% CI:
$\hat{\beta}_1 - t(\alpha/2, df)se(\hat{\beta}_1|X) \leq \beta_1 \leq \hat{\beta}_1 + t(\alpha/2, df)se(\hat{\beta}_1|X)$
$H_0 : \beta_1 = 0$
$H_A : \beta_1 \neq 0$

2.4 Prediction

Prediction point of y_* $\tilde{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$
$\tilde{y}_*$ predicts the yet unobserved y_*
$y_* = \beta_0 + \beta_1 x_* + e_*$
e_* is the random error attached to the future value, with variance σ^2
The prediction error variability has second component from the uncertainty in the estimates of the coefficients.
Combining the two sources of variability we have:
$Var(\tilde{y}_*|x_*) = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}\right)$
First term is var due to e_* second is error for estimating coefficients
$sepred(\tilde{y}_*|x_*) = \sigma\left(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}\right)^{1/2}$
A prediction interval uses the t-distribution with df equal to the df in estimating σ^2

2.5 Fitted Values

Obtaining an estimate of $E(Y|X = x_*)$
Quantity is estimated by $\hat{y} = \beta_0 + \beta_1 x_*$
$sefit(\hat{y}|x_*) = \hat{\sigma}\left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}\right)^{1/2}$
CI:
$(\hat{\beta}_0 + \hat{\beta}_1 x) - sefit(\hat{y}|x)[2F(\alpha; 2, n-2)]^{1/2} \leq y \leq$
$(\hat{\beta}_0 + \hat{\beta}_1) + sefit(\hat{y}|x)[2F(\alpha; 2, n-2)]^{1/2}$

2.6 Coefficient of Determination

SYY Total sum of squares
$SSreg = SYY - RSS$
$SSreg = SYY - \left(SYY - \frac{(SYY)^2}{SXX}\right) = \frac{(SXY)^2}{SXX}$
$R^2 = \frac{SSreg}{SYY} = 1 - \frac{RSS}{SYY}$
The left side is the proportion of variability in the response explained by regression on the predictor. The right side is 1 minus the remaining unexplained variability.

- R^2 is a scale-free one-number summary of the strength of the relationship between the x_i and the y_i in the data.
- Interpret as: about R^2% of the variablity in the observed values of the model are explained by the predictor variable.
- In SLR R^2 is the same as the square of the sample correlation between predictor and response

$R^2_{adj} = 1 - \frac{RSS/df}{SYY/(n-1)}$
adding a correction for df of the sums of squares that can facilitate comparing models in multiple regression

2.7 Residuals

Most common plot in SLR is resid vs fitted values

- A null plot would indicate no failure of assumptions
- Curvature might indicate that the fitted mean function is inappropriate
- Residuals that seem to increase or decrease in average magnitude with the fitted values might indicate nonconstant residual variance
- A few relatively large residuals may be indicative of outliers, cases for which the model is somehow inappropriate.

Chapter 3

Multiple Regression

3.1 Adding a Regressor to an SLR Model

Start with $E(Y|X_1 = x_1) = \beta_0 + \beta_1 x_1$
adding X_2 $E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
The main idea in adding X_2 is to explain the part of Y that has not already been explained by X_1

3.2 The MLR Model

$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ Conditioning on all regressors on the right side
$E(Y|X = x) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ Conditioning on specific values for the predictors
When the number of regressors, $p = 2$ the mean function corresponds to a plane in 3 dimensions
When $p > 2$, the fitted mean function is a hyperplane

3.3 Predictors and Regressors

- **Predictors-** the simplest type of regressor is equal to a predictor

- **Transformations of predictors-** sometimes original predictors need to be transformed to make the general MLR model hold to a reasonable approximation

- **Polynomials-** Problems with curved mean functions can sometimes be accommodated in the MLR model by including polynomial regressors in the predictor variables. Could include both a predictor X_1 and its square X_1^2 to fit a quadratic polynomial in that predictor.

- **Interactions and other combinations of predictors-** Combining several predictors is often useful. ex: bmi which is a function of both weight and height in place of both height and weight.
- **Interactions-** Products of regressors are often included in a mean function along with the base regressors to allow for joint effects.
- **Factors-** A categorical predictor with two or more levels. Factors are included in MLR using **dummy variables** which are typically regressors that have only two values, 0 and 1 indicating which category is present for a particular observation.

The marginal relationships between the response and each of the variables is not sufficient to understand the joint relationship between the response and the regressors.
The interrelationships among the regressors are also important.
The pairwise relationships between the regressors can be viewed in the remaining cells of the scatterplot matrix.
A more traditional, and less informative, summary of the two-variable relationships is the matrix of sample correlations.

3.4 Ordinary Least Squares

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{'}$ $(p+1) \times 1$ vector of unknown regression coefficients
Equation for the mean function evaluated at $\boldsymbol{x}_i$ is:
$E(Y|X = \boldsymbol{x}_i) = \boldsymbol{x}_i^{'}\boldsymbol{\beta}$
Mean function in matrix terms:
$E(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$
$\boldsymbol{Y}$ is vector of responses, $\boldsymbol{X}$ is the $nx(p+1)$ matrix whose ith row is $\boldsymbol{x}_i^{'}$

3.4.1 The Errors e

Define the unobservable random vector of errors $\boldsymbol{e}$ elementwise by:
$e_i = y_i = E(Y|X = \boldsymbol{x}_i) = y_i - \boldsymbol{x}_i^{'}\boldsymbol{\beta}$
$\boldsymbol{e} = (e_i, \ldots, e_n)^{'}$
$E(\boldsymbol{e}|X) = 0 \quad Var(\boldsymbol{e}|X) = \sigma^2\boldsymbol{I}_n$
where $Var(e|X)$ means the covariance matrix of $\boldsymbol{e}$ for a fixed value of X, $\boldsymbol{I}_n$ is the nxn matrix with ones on the diagonal and zeroes everywhere else.

Adding the assumption of normality we can write:
$(\boldsymbol{e}|X) \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$

3.4.2 OLS Estimators

The least squares estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is chosen to minimize the residual sum of squares function:

$$RSS(\boldsymbol{\beta}) = \sum (y_i - \boldsymbol{x}_i^{'})^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{'}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

OLS estimate $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}$ provided the inverse $\boldsymbol{X}^{'}\boldsymbol{X}^{-1}$ exists
Do not use the above equation to compute least squares estimates because of potential rounding errors
$\hat{\boldsymbol{\beta}}$ depends only on the sufficient statistics $\boldsymbol{X}^{'}\boldsymbol{X}$ and $\boldsymbol{X}^{'}\boldsymbol{Y}$ which are matrices of uncorrected sums of squares and cross products.

At the very least, computations should be based on *corrected* sums of squares and cross products. Suppose we define $\mathcal{X}$ to be the $n \times p$ matrix

$$\mathcal{X} = \begin{pmatrix} (x_{11} - \overline{\mathbf{x}}_1) & \cdots & (x_{1p} - \overline{\mathbf{x}}_p) \\ (x_{21} - \overline{\mathbf{x}}_1) & \cdots & (x_{2p} - \overline{\mathbf{x}}_p) \\ \vdots & \vdots & \vdots \\ (x_{n1} - \overline{\mathbf{x}}_1) & \cdots & (x_{np} - \overline{\mathbf{x}}_p) \end{pmatrix}$$

This matrix consists of the original $\mathbf{X}$ matrix, but with the first column removed and the column mean subtracted from each of the remaining columns. Similarly, $\mathcal{Y}$ is the vector with typical elements $y_i - \overline{y}$. Then

$$\mathcal{C} = \frac{1}{n-1}\begin{pmatrix} \mathcal{X}'\mathcal{X} & \mathcal{X}'\mathcal{Y} \\ \mathcal{Y}'\mathcal{X} & \mathcal{Y}'\mathcal{Y} \end{pmatrix}$$

is the matrix of sample variances and covariances

Part II

Class Notes

Chapter 4

Notes 2 Matrix Algebra

Symmetric Matrix- $\boldsymbol{A}$ where $a_{ij} = a_{ji} \quad \forall\, i, j$
Trace- sum of diagonal elements of a square matrix
Matrix Multiplication $\boldsymbol{A}_{r\times s}\boldsymbol{B}_{s\times t} = \{\sum_{k=1}^{s} a_{ik}b_{kj}\} = \boldsymbol{C}_{r\times t}$

4.1 Orthogonal

Orthogonal Matrix- a square matrix with $\boldsymbol{A}^{'} = \boldsymbol{A}^{-1}$
A square matrix $\boldsymbol{A}$ is orthogonal if $\boldsymbol{A}^{'}\boldsymbol{A} = \boldsymbol{I} = \boldsymbol{A}\boldsymbol{A}^{'}$
Two vectors x and y are orthogonal if $x^{'}y = 0$
x and y are **orthonormal** if they are orthogonal and are normalized: $x^{'}y = 0$, $x^{'}x = 1$, $y^{'}y = 1$
An orthogonal matrix has orthonormal columns.

4.2 Rules of Matrix Operation

Distributive Laws

- $A(B + C) = AB + AC$
- $(B + C)D = BD + CD$

Associative Laws
$(AB)C = A(BC)$ **Transpose Operations**

- $(A + B)^{'} = A^{'} + B^{'}$
- $(AB)^{'} = B^{'}A^{'}$

4.3 Linear Dependence and Rank

The columns of $\boldsymbol{A}$ are **linearly dependent** if they contain redundant information
If we can find two distinct vectors λ and γ such that $\boldsymbol{A\lambda} = \boldsymbol{A\gamma} = \boldsymbol{x}$ then the columns of $\boldsymbol{A}$ are linearly dependent
Equivalently let $\delta = \lambda - \gamma$
The columns of $\boldsymbol{A}$ are linearly dependent if there exists a vector $\delta \neq 0$ such that $\boldsymbol{A\delta} = \boldsymbol{0}$
Rank of $\boldsymbol{A}$ is the number of linearly independent columns in $\boldsymbol{A}$
If $\boldsymbol{A}$ is an $(r \times c)$ matrix with $r \geq c$, $\boldsymbol{A}$ is **full rank** if $rank(\boldsymbol{A}) = c$
If $rank(\boldsymbol{A}) < c$ $\boldsymbol{A}$ is less than full rank
In linear regression, the matrix of covariates $\boldsymbol{X}$ must have full rank in order for the parameter estimates $\hat{\boldsymbol{\beta}}$ to be unique
A square matrix less than full rank is called **singular**, if full rank called **nonsingular**
Elementary Row Operations:

1. multiplying a row by a nonzero constant
2. adding one row to another
3. exchanging two rows

4.4 Determinants

Determinant- a single number summary of a square matrix that gives us information about the rank of the matrix
The determinant of a diagonal or triangular matrix is the product of the diagonal values
If determinant=0 then the matrix is less than full rank and the inverse does not exist
If $det \neq 0$ then the inverse exists and full rank
For full rank matrices that conform, $|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|$
Also $|\boldsymbol{A}^{'}| = |\boldsymbol{A}|$

4.5 Positive Definite and Semidefinite Matrices

Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix. $\boldsymbol{A}$ is **positive definite** iff:

1. $a_{ii} > 0 \quad \forall\, i = 1, \ldots, n$
2. The determinant of every square submatrix of upper-left corner of $\boldsymbol{A}$ is positive

Positive semidefinite if we replace > 0 with ≥ 0
Nonnegative definite- positive definite or positive semidefinite
Covariance matrices are nonnegative definite

4.6 Inverses

Normal equations for the linear model:
$(\boldsymbol{X}_{n\times p})^{'}(\boldsymbol{X}_{n\times p})\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_{n\times p})^{'}\boldsymbol{y}_{n\times 1}$
If $\boldsymbol{A}$ is full rank then there exists a unique matrix $\boldsymbol{A}^{-1}$, the inverse of $\boldsymbol{A}$ where:
$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$
Properties of Inverses

1. For a scalar $\boldsymbol{A}_{1\times 1} = a,\ \boldsymbol{A}^{-1} = \frac{1}{a}$

2. The inverse of a diagonal matrix is the diagonal matrix of reciprocals of the diagonal elements

3. For conforming full rank matrices, $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$

4. A symmetric matrix has a symmetric inverse

5. $(\boldsymbol{A}^{'})^{-1} = (A^{-1})^{'}$

6. The determinant of the inverse is the inverse of the determinant.
$|\boldsymbol{A}^{-1}| = \frac{1}{|\boldsymbol{A}|}$

7. Inverse of a 2×2 matrix

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

$$\boldsymbol{A}^{-1} = \frac{1}{|\boldsymbol{A}|}\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

4.7 Eigenvalues and Eigenvectors

Eigenanalysis is defined only for square matrices
$\boldsymbol{A}$ is an $n \times n$ matrix
Right eigenvector of $\boldsymbol{A}$ is any nonzero $n \times 1$ vector x satisfying $\boldsymbol{Ax} = \lambda\boldsymbol{x}$
λ is the **eigenvalue** corresponding to x
eigen is German for characteristic
Eigenvectors are not unique, convention is to scale the eigenvector x so that $x^{'}x = 1$, normalizing it to unit length

4.8 Finding Eigenvectors and Eigenvalues

An $n \times n$ matrix has n eigenvalues
Definition of eigenvectors: $\boldsymbol{Ax} = \lambda \boldsymbol{x}$
Characteristic equation: $|\boldsymbol{A} - \lambda \boldsymbol{I}| = \boldsymbol{0}$

$$| \mathbf{A} - \lambda \mathbf{I} | = \left| \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\left| \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = \left| \begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} \right| = 0$$

$$(a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0$$

$$\lambda^2 - \lambda(a_{11} + a_{22}) + a_{11}a_{22} - a_{12}a_{21} = 0.$$

eigenvectors corresponding to these eigenvalues can be found using the equation $\boldsymbol{Ax} = \lambda \boldsymbol{x}$
Normalize an eigenvector $x = (x_1, x_2)'$: $\frac{1}{\sqrt{x_1^2 + x_2^2}}(x_1, x_2)'$

4.9 Properties of Eigenvalues and Eigenvectors

- For $\boldsymbol{A}_{n \times n}$ number of distinct eigenvalues ranges from 1 to n
- $trace(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i$
- $|\boldsymbol{A}| = \prod_{i=1}^{n} \lambda_i$
- $\boldsymbol{A}$ full rank $\Leftrightarrow$ $\boldsymbol{A}$ has no zero eigenvalues
- $|\boldsymbol{A}| = 0 \Leftrightarrow$ at least one eigenvalue is zero $\Leftrightarrow$ $\boldsymbol{A}$ is not full rank
- The number of nonzero eigenvalues of $\boldsymbol{A}$ is $rank(\boldsymbol{A})$
- Small eigenvalues imply that there are near-linear dependencies in the columns of $\boldsymbol{A}$
- $\boldsymbol{A}$ is positive definite if $min(\lambda_i) > 0$
- $\boldsymbol{A}$ is positive semidefinite if $min(\lambda_i) \geq 0$

4.10 Random Vectors and Matrices

$\boldsymbol{Z}$ is an $(n \times p)$ matrix of random variables

$$E(\boldsymbol{Z}) = \begin{pmatrix} E(Z_{11}) & \dots & E(Z_{1p}) \\ \vdots & \dots & \vdots \\ E(Z_{n1}) & \dots & E(Z_{np}) \end{pmatrix}$$

The expectation of a random matrix is the matrix of the expectations
For $\boldsymbol{Y}$ an $(n \times 1)$ random vector, the **covariance matrix** is:

$$Cov(\boldsymbol{Y}) = E[(\boldsymbol{Y}-\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})^{'}]$$
$$= \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \cdots & \cdots & \sigma_{nn} \end{pmatrix}$$
$$\text{where } \sigma_{ij} = E[(Y_i - \mu_i)(Y_j - \mu_j)^{'}] \quad i,j = 1,\dots,n$$

Let $\boldsymbol{\mu} = E(\boldsymbol{Y})$ and $\boldsymbol{\Sigma} = Cov(\boldsymbol{Y})$
Suppose $\boldsymbol{A}_{r\times n}$ is a matrix of constants and $\boldsymbol{b}_{r\times 1}$ is a vector of constants. Then:
$E(\boldsymbol{AY}+\boldsymbol{b}) = \boldsymbol{A}E(\boldsymbol{Y}) + \boldsymbol{b} = \boldsymbol{A\mu} + \boldsymbol{b}$
$Cov(\boldsymbol{AY}+\boldsymbol{b}) = \boldsymbol{A}Cov(\boldsymbol{Y})\boldsymbol{A}^{'} = \boldsymbol{A\Sigma A}^{'}$ Let $\boldsymbol{W}_{r\times 1}$ be a random vector with $E(\boldsymbol{W}) = \boldsymbol{\gamma}$ Then:
$Cov(\boldsymbol{W},\boldsymbol{Y}) = E[(\boldsymbol{W}-\boldsymbol{\gamma})(\boldsymbol{Y}-\boldsymbol{\mu})^{'}]$
Where $Cov(\boldsymbol{W},\boldsymbol{Y})$ is an $(r \times n)$ matrix of covariances with ij^{th} element equal to $Cov(W_i, Y_j)$

4.11 Multivariate Normal Distribution

Suppose $X = (X_1, \dots, X_n)^{'}$ Then X has an n dimensional multivariate normal distribution with mean μ and covariance matrix Σ if X has density:

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^{'}\Sigma^{-1}(x-\mu)\right\}$$

$X \sim N_n(\mu, \Sigma)$
Σ must be positive definite

4.12 Facts about the multivariate normal distribution

1. A linear transformation of a multivariate normal distribution (mvn) yields another mvn.
   If $X \sim N_n(\mu, \Sigma)$ and $Y = AX + b$
   with $\boldsymbol{A}_{r\times n}$ matrix of constants and $\boldsymbol{b}_{r\times 1}$ vector of constants.
   Then $Y \sim N_r(A\mu + b, A\Sigma A^{'})$

2. A linear combination of independent mvn distributions is an mvn distribution.
   Suppose $X_1, \ldots, X_k$ are independent with $X_i \sim N_n(\mu_i, \Sigma_i) \quad i = 1, \ldots, k$ and
   $a_1, \ldots, a_k$ are scalars
   Define $Y = a_1X_1 + \cdots + a_kX_k$
   Then $Y \sim N(\mu^*, \Sigma^*)$ where $\mu^* = \sum_{i=1}^k a_i\mu_i$ and $\Sigma^* = \sum_{i=1}^k a_i^2\Sigma_i$

3. Marginal distributions of mvn are also mvn
   Suppose $X \sim N_n(\mu, \Sigma)$
   Partition X into $X = \binom{X_1}{X_2}$ where X_1 is $r \times 1$ and X_2 is $(n-r) \times 1$
   Partition μ as $\mu = \binom{\mu_1}{\mu_2}$ where μ_1 is $r \times 1$ and μ_2 is $(n-r) \times 1$
   Partition Σ as:
   $$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$
   where Σ_{11} is $r \times r$, Σ_{21} is $(n-r) \times r$ and Σ_{22} is $(n-r) \times (n-r)$
   Then marginal distribution of $X_1 \sim N_r(\mu_1, \Sigma_{11})$
   Marginal distribution of $X_2 \sim N_{(n-r)}(\mu_2, \Sigma_{22})$

4. Conditional distributions of mvn are mvn.
   Suppose $X \sim N_n(\mu, \Sigma)$
   Using same partition as above, we have:
   $X_1|X_2 = x_2 \sim N_r(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma^*)$
   where $\Sigma^* = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

Chapter 5

GLM Estimation and Testing

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \text{ if X is full rank}$$
$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \text{ hat matrix, same rank as X}$$
$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{y} \text{ predicted values}$$
$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{1} - \boldsymbol{H}\boldsymbol{y})$$

5.1 GLH

assume iid Gaussian errors.
$\boldsymbol{\beta}$ is the matrix of primary parameters
$\boldsymbol{\theta}_{ax1} = \boldsymbol{C}_{axp}\boldsymbol{\beta}_{px1}$ is a matrix of secondary parameters defined by $\boldsymbol{C}$
Each row of $\boldsymbol{C}$ defines a new scalar parameter in terms of the $\boldsymbol{\beta}$'s ex: $\beta_1 - \beta_2$
Let $\boldsymbol{\theta}_0$ be matrix of known constants (hypothesized values) usually zero matrix

The (Univariate) General Linear Hypothesis:

$$H_0 : \boldsymbol{\theta}_{ax1} = \boldsymbol{\theta}_0$$
$$H_A : \boldsymbol{\theta}_{ax1} \neq \boldsymbol{\theta}_0$$

5.2 Estimability of a Parameter

A (linear) function of the parameters is defined to be **estimable** if it is identically equal to some linear function of the expected value of the vector of observations, $\boldsymbol{y}$

- A scalar parameter, $\theta_i = \boldsymbol{C}_{1xp}\boldsymbol{\beta}_{px1}$ is estimable
  $\Leftrightarrow \boldsymbol{C}_{1xp}\boldsymbol{\beta}_{px1} = \boldsymbol{t}'_{1xn}E(\boldsymbol{y}_{nx1})$
  for $\boldsymbol{t}$ a vector of constants
- For a vector we need: $\boldsymbol{\theta}_{ax1} = \boldsymbol{T}_{axn}E(\boldsymbol{y}_{nx1})$

There always exists $r = rank(\boldsymbol{X})$ distinct and estimable parameters
These are not necessarily elements of $\boldsymbol{\beta}$ but may be linear combinations of elements

- If $rank(\boldsymbol{X}) = r = p$, then $\hat{\boldsymbol{\beta}}$ exists (uniquely), $\boldsymbol{\beta}$ is estimable and any (nonzero) $\boldsymbol{C}$ gives estimable $\boldsymbol{\theta}$
  This is usually the case with continuous predictors unless some predictors are collinear
- If $rank(\boldsymbol{X}) = r < p$, $\boldsymbol{\beta}$ is not estimable (although as many as r elements may be), and for $\hat{\boldsymbol{\theta}} = \boldsymbol{C}\boldsymbol{\beta}$, we must check estimability.

To show a set of parameters: $\boldsymbol{\theta}_{a\times 1} = \boldsymbol{C}_{a\times p}\boldsymbol{\beta}_{p\times 1} = \boldsymbol{T}_{a\times n}E(\boldsymbol{y}_{n\times 1})$ is estimable: show that $\boldsymbol{C}_{a\times p} = \boldsymbol{T}_{a\times n}\boldsymbol{X}_{n\times p}$
Estimable $\hat{\boldsymbol{\theta}}$ shares the optimality of $\hat{\boldsymbol{\beta}}$

5.3 Testability of a Hypothesis

Likelihood Ratio (LR) Test - used for comparing the goodness of fit of two statistical models (null and alternative)
The test is based on the **likelihood ratio**, which expresses how many times more likely the data are under one model than the other.

Let $\boldsymbol{M}_{a\times a} = \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'$
Define GLH testability as the (unique) existence of the LR test
$\boldsymbol{\theta}$ is testable $\Leftrightarrow$

- $\boldsymbol{C}$ is full rank a (no redundancies) and
- $\boldsymbol{\theta}$ is estimable

Or equivalently

- $\boldsymbol{M}$ is full rank a and
- $\boldsymbol{\theta}$ is estimable

If $\boldsymbol{X}$ is full rank then $\boldsymbol{\theta}$ is testable $\Leftrightarrow$
$\boldsymbol{C}$ is full rank a or $\boldsymbol{M}$ is full rank a (because any $\boldsymbol{\theta}$ is estimable)

5.4 Computation of Test Statistic and p-value

Define the sums of squares hypothesis as
$SSH_{1\times 1} = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\boldsymbol{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

With HILE Gauss, the likelihood ratio statistic equals

$$\begin{aligned} F_{obs} &= \frac{SSH/a}{SSE/(n-r)} = \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\boldsymbol{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/a}{\widehat{\sigma}^2} \\ &= \frac{MSH}{MSE} \end{aligned}$$

Under H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, SSH and SSE are scaled χ^2 random variables, with $SSH/\sigma^2 \sim \chi^2(a)$, independently of $SSE/\sigma^2 \sim \chi^2(n-r)$. It can be shown that if $z_1 \sim \chi^2_{d_1}$, $z_2 \sim \chi^2_{d_2}$, and $z_1 \perp z_2$, then $\frac{z_1/d_1}{z_2/d_2}$ follows an $F_{d1,d2}$ distribution. Thus

$$F_{obs} = \frac{[SSH/\sigma^2]/a}{[SSE/\sigma^2]/(n-r)} = \frac{SSH/a}{SSE/(n-r)} \sim F(a, n-r).$$

The p-value equals the probability of observed or more extreme data arising under the null:
p-value= $P\{F(a, n-r) \geq F_{obs}\} = 1 - P\{F(a, n-r) < F_{obs}\}$
Reject H_0 if $F_{obs} > f_{crit} = F^{-1}(1-\alpha, a, n-r)$
$qf(prob, df_1, df_2)$ (F statistic)
$1 - pf(crit, df_1, df_2)$ (p-value)

All linear model GLH tests correspond to comparing two models, the "full" model, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ and a reduced model defined by constraints

5.5 Wald Tests

For a single coefficient β_j we can test $H_0 : \beta_j = 0$ if β_j is estimable
Using properties of the standard normal distribution we can base our test on the ratio:

$$t = \frac{\hat{\beta}_j - 0}{\sqrt{var(\hat{\beta}_j)}}$$

Obtain estimate $\hat{\sigma}^2 = \dfrac{SSE}{dfE}$
If we know σ^2 exactly then $t \sim N(0,1)$
If we estimate σ^2 from the data then $t \sim t_{dfE}$

5.6 F-Tests

two-sided F test uses α critical value
one-sided F test uses 2α critical value

Chapter 6

Some Distributional Results for the GLM

In analysis with the GLM, we use three kinds of distributions: multivariate Gaussian, χ^2 and F
Assume HILE Gauss assumptions hold

6.1 A Full Rank Basis For Less Than Full Rank Models

If $\boldsymbol{X}$ is less than full rank $(r < p)$, then *collinearity* exists among columns of $\boldsymbol{X}$. If $\mathbf{X}$ is less than full rank, then we say the model is also less than full rank. Also, $r = \text{rank}(\boldsymbol{X}) = \text{rank}(\mathbf{X}'\mathbf{X})$ is the # of estimable parameters.

For every less than full rank model, there exists a corresponding full rank model with r estimable parameters.
That is: for less than full rank $\boldsymbol{X}$ there exists a $p \times r$ matrix $\boldsymbol{V}_+$ such that:
$\boldsymbol{X}_{n\times p} = \boldsymbol{X}_{*,(n\times r)}\boldsymbol{V}'_{+,(r\times p)}$
with $rank(\boldsymbol{X}_*) = rank(\boldsymbol{V}_+) = r < p$
$\boldsymbol{X}_*$ provides a **full rank basis** for $\boldsymbol{X}$

Suppose that we have the model

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\ \boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1},$$

where rank$(\mathbf{X}) = r < p$.

Then, defining $\mathbf{X}_{*,(n\times r)} = \mathbf{X}_{n\times p}\ \mathbf{V}_{+,(p\times r)}$ with corresponding parameter vector $\boldsymbol{\beta}_{*,(r\times 1)}$, an equivalent full-rank model is given by

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{*,(n\times r)}\ \boldsymbol{\beta}_{*,(r\times 1)} + \boldsymbol{\varepsilon}_{n\times 1},$$

with $\widehat{\boldsymbol{\beta}_*} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}$.

Many possible choices of the matrix $\boldsymbol{V}_+$ exist, such as the set of eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ corresponding to non-zero eigenvalues.

Every parameter estimable in the original (less than full rank) model is also estimable in the full rank model, and any estimable parameter is expressible as a linear combination of the β_*'s.

Chapter 7

Multiple Regression General Consideration

Definitions of Basic Sums of Squares

For the model $\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\,\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$, we have

- $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$,
- $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, and
- $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \widehat{\mathbf{y}}$.

We are already familiar with the sum of squares for error, given by

$$\begin{aligned} SSE &= \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} = (\mathbf{y} - \widehat{\mathbf{y}})'(\mathbf{y} - \widehat{\mathbf{y}}) \\ &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}\widehat{\varepsilon}_i^2. \end{aligned}$$