STAT 331 - Applied Linear Models

Fantastic Models and How to Abuse Them

Jose Luis Avilez
Faculty of Mathematics
University of Waterloo

Contents

Chapter 1

Preliminaries

Definition 1.1. We define a **statistical model** as an equation

$$y = \mu + \epsilon$$

where μ is a **deterministic** component and ϵ is a **stochastic** component (or noise).

Definition 1.2. A **response** variable is denoted Y and its values are $(y_1, \ldots, y_n)$; an **independent** variable is denoted X and its values are $(x_1, \ldots, x_n)$; the **regression slope** is denoted β; the **noise** term is denoted ϵ; the regression equation is then given by

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

Definition 1.3. To emphasise that the model applies to each potential experiment, we index using our dataset (i.e. $\{(x_i, y_i)\}_{i=1,\ldots,n}$ are data points) to say

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Definition 1.4. We say that the noise is exhibits **homoscedasticity** if each ϵ_i has equal variance. **Heteroscedasticity** means they have unequal variances.

Definition 1.5. In a **simple linear model** there is only one explanatory variable and we make the following assumptions for the error term ϵ:

1. ϵ_i is normally distributed for each i
2. $\mathrm{E}(\epsilon_i) = 0$, for $i = 1, 2, \ldots, n$
3. $\mathrm{Var}(\epsilon_i) = \sigma^2$
4. ϵ_i and ϵ_j are independent random variables for $i \neq j$

Theorem 1.6. *In a simple linear model, if we take x_i to be deterministic and each y_i as a random variable, $\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$.*

Proof. $\mathrm{E}[y_i] = \mathrm{E}[\beta_0 + \beta_1 x_i + \epsilon_i] = \beta_0 + \beta_1 x_i + \mathrm{E}[\epsilon] = \beta_0 + \beta_1 x_i$. ■

Definition 1.7. We define a **general linear model**[1] as

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Note that it has multiple independent variables. A more efficient way to write this is in matrix form

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

Except, no sane person puts those funny hats on top of their vectors, so we shall simply write $y = X\beta + \epsilon$ where X is the design matrix. Note it has a column of 1s to multiply out the constant β_0 term.

Definition 1.8. We say that a model is **"parsimonious"** if it is "economic" and has "low complexity". We use inverted commas since these are not well-defined mathematical constructs.

[1] Not to be confused with **generalised**.

Chapter 2

Simple Linear Regression

For this chapter, we explore the consequences of Definition 1.5 and how to test their assumptions.

To obtain estimates of the parameters in a simple linear model we have two available methods: (i) **maximum likelihood estimation**, and (ii) **least squares estimate**. The former requires distributional assumptions; the latter does not.

2.1 Estimating Simple Linear Regression

Theorem 2.1. *For a simple linear model, the maximum likelihood estimators are given by* $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ *and* $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Proof. Given that the y_i are independent, we have that the likelihood function is

$$L(\beta_0, \beta_1, \sigma^2) = f(y_1, \ldots, y_n | \beta_0, \beta_1, \sigma^2) = \prod_i^n f(y_i | \beta_0, \beta_1, \sigma^2)$$

Under the normality assumption for y_i, we then have

$$f(y_i | \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_1)^2\right)$$

Thus, the log-likelihood function is given by

$$l(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

The remainder of the result follows from maximising the log-likelihood for the parameters. We show the computation in an upcoming Theorem. ■

Definition 2.2. We say that $\hat{\beta}_0$ and $\hat{\beta}_1$ are least squares estimates if they minimise the equation

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

Theorem 2.3. *The least squares estimates are equal to the maximum likelihood estimates*[1].

Proof. Taking partial derivatives with respect to the parameters, we obtain,

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \tag{2.1}$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^n x_i(y_i - \beta_0 - \beta_1 x_i) \tag{2.2}$$

[1] Proofs for this theorem can be seen in Lectures 1 and 4 of Shalizi's notes

To maximise the parameters, we set the partial derivatives to zero. It is easy to see that the first expression is minimised when $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. Minimising the second expression requires a bit more algebraic mumbo-jumbo.

$$\begin{aligned}
0 &= \sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) \\
&= \sum_{i=1}^{n}(x_i y_i) - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n} x_i y_i - n\bar{x}(\bar{y} - \beta_1\bar{x}) - \beta_1 \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\beta_1\bar{x}^2 - \beta_1 \sum_{i=1}^{n} x_i^2 \\
&\iff \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 + n\bar{x}^2} \\
&= \frac{S_{xy}}{S_{xx}}
\end{aligned}$$

Ta-da! ■

Definition 2.4. The following two equations are called **normal equations**:

$$n\hat{\beta}_0 + \left(\sum x_i\right)\hat{\beta}_1 = \sum y_i \tag{2.3}$$

$$\left(\sum x_i\right)\hat{\beta}_0 + \left(\sum x_i^2\right)\hat{\beta}_1 = \sum x_i y_i \tag{2.4}$$

Definition 2.5. The **residual**, e_i, of the fitted value at x_i is $e_i = y_i - \hat{\mu}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$.

Theorem 2.6. *In a regression line fitted by the least squares estimate procedure, the following are facts about residuals:*

1. $\sum e_i = 0$
2. $\sum e_i x_i = 0$
3. $\sum \hat{\mu}_i e_i = 0$

Proof. Follows from the minimisation procedure used in Theorem 2.3. In particular, the sum of residuals is zero because we set equation 2.1 to 0. Furthermore, the second equation is zero because this is exactly equation 2.2. The third equality follows from the above two,

$$\sum \hat{\mu}_i e_i = \hat{\beta}_0 \sum e_i + \hat{\beta}_1 \sum e_i x_i = 0$$

■

Theorem 2.7. *The maximum likelihood estimate of σ^2 is $\hat{\sigma^2} = \frac{S(\hat{\beta}_0, \hat{\beta}_1)}{n}$.*

Proof. We take the partial derivative of the log-likelihood function with respect to σ^2. Recall that the log-likelihood function is given by

$$l(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Taking the desired partial derivative, we obtain,

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Setting this to zero yields,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = \frac{S(\hat{\beta}_0, \hat{\beta}_1)}{n}$$

■

Theorem 2.8. *The estimated value of σ^2 using the least squares estimate method is*

$$S^2 = \frac{S(\hat{\beta}_0, \hat{\beta}_1)}{n-2}$$

We call this the least square error and it has $n-2$ degrees of freedom. In R, the summary output for a linear model is the ***residual standard error****, which is simply $S = \sqrt{S^2}$.*

Proof. We switch the denominator to the denominator corresponding to the appropriate degrees of freedom. ■

2.2 Inference in Simple Linear Regression

Theorem 2.9. *The mean squared error, S^2 is an unbiased estimate for σ^2. That is, $\mathrm{E}\left(S^2\right) = \sigma^2$.*

Proof. A "cheat" proof requires noting that

$$\frac{S(\hat{\beta}_0, \hat{\beta}_1)}{\sigma} \sim \chi^2(n-2)$$

and that the expected value of a random variable following a $\chi^2(n-2)$ distribution is $n-2$. From this, it follows that S^2 is unbiased.

Other proofs follow from manipulating the expression enough into one whose expectations we already know. However, we spare the reader of such algebraic monstrosities.

■

Theorem 2.10. *The estimators $\hat{\beta}_0, \hat{\beta}_1$ are unbiased; that is $\mathrm{E}\left[\hat{\beta}_{0,1}\right] = \beta_{0,1}$. The estimator $\hat{\mu}_0$ is also unbiased.*

Proof. We can write

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

where $c_i = \frac{x_i - \bar{x}}{S_{xx}}$. Thus,

$$\mathrm{E}\left[\hat{\beta}_1\right] = \mathrm{E}\left[\sum c_i y_i\right] = \sum c_i \mathrm{E}\left[y_i\right] = \sum c_i \mathrm{E}\left[\beta_0 + \beta_1 x_i\right] = E[\beta_0]\sum c_i + \beta_1 \sum c_i \mathrm{E}\left[x_i\right] = \beta_1 \frac{S_{xx}}{S_{xx}} = \beta_1$$

Likewise,

$$\mathrm{E}\left[\hat{\beta}_0\right] = \mathrm{E}\left[y_i - \hat{\beta}_1 x_i\right] = \bar{y} - \beta_1 \bar{x} = \beta_0$$

■

Theorem 2.11. *The estimator $\hat{\mu}$ is an unbiased estimate for μ and S^2 is an unbiased estimator for σ^2.*

Proof. The first follows easily from Theorem 2.10. The second estimator requires finding a pivotal quantity which follows a chi-squared distribution with $n-2$ degrees of freedom. I'll provide details later. ■

Theorem 2.12. *The following are the variances for the estimators:*

1. $\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}}$
2. $\mathrm{Var}\left(\hat{\beta}_0\right) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$

3. $\text{Var}\left(\hat{\mu_0}\right) = \sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]$

Proof. The first two follow by our usual variance formulas. The third point requires writing

$$\text{Var}\left(\hat{\mu_0}\right) = \text{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) = \text{Var}\left(\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_0\right) = \text{Var}\left(\bar{y} + \hat{\beta}_1(x_0-\bar{x})\right) = \text{Var}\left(\overline{Y}\right) + (x_0-\bar{x})^2\text{Var}\left(\hat{\beta}_1\right)$$

which simplifies to the desired expression. ■

Theorem 2.13. *If $Z \sim N(0,1)$ and $S \sim \chi_d$ where Z and S are independent, then $\frac{Z}{\sqrt{S/d}} \sim t_d$.*

Proof. The proof of this fact is left for a Mathematical Statistics class. However, we will use this result extensively when deriving confidence intervals for our parameters. ■

Theorem 2.14. $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$

Proof. Follows from the fact that it is a linear combination of y_i, each of which is normally distributed. ■

Theorem 2.15. $\frac{\hat{\beta}_1-\beta_1}{\frac{s}{\sqrt{S_{xx}}}} \sim t(n-2)$

Proof. Follows from Theorem 2.13 and 2.14. ■

Theorem 2.16. $\frac{\hat{\mu}_0-\mu_0}{\sigma\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]^{\frac{1}{2}}} \sim N(0,1)$*.*

Proof. Since $\hat{\mu}$ is the linear combination of normally distributed random variables, it follows a normal distribution. The variance for $\hat{mu}$ was derived in Theorem 2.12, so normalising by subtracting by the mean and dividing by its standard deviation yields the standard normal distribution. ■

An immediate corollary of the theorem above is the useful result used to compute confidence intervals for the mean response.

Theorem 2.17. $\frac{\hat{\mu}_0-\mu_0}{s\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]^{\frac{1}{2}}} \sim t(n-2)$*.*

Proof. Follows from Theorem 2.13 and Theorem 2.16. ■

Example 2.18. I might post an example here, at a later date, showing how to build the confidence intervals for the estimated quantities above.

Now, suppose we want to predict where a new observation will lie, given that the parameters in our model are unknown. To do so, we need to quantify the error of obtaining a new prediction. We describe the procedure in the upcoming theorem.

Theorem 2.19. *The pivotal quantity related to a new observation at the level x_0 is*

$$\frac{\hat{y}-y}{s\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \sim t(n-2)$$

Proof. Assume the linear model holds. Then

$$\text{Var}\left(y-\hat{y}\right) = \text{Var}\left(y-\hat{\mu}\right) = \text{Var}\left(y\right) + \text{Var}\left(\hat{\mu}\right) = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right] = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]$$

Note then that $y-\hat{y} \sim N(0,\ell^2)$, where ℓ^2 is simply the expression above. Using Theorem 2.13, the result follows. ■

2.3 ANOVA and R^2

As much as it pains me, I have to include this section here for completion. To find out why it is painful, check out Lecture 10 in the Cosma Shalizi notes[2]. For that very reason, we include how to compute an ANOVA table and leave the reader to figure out why it's a waste of time.

Definition 2.20. For the purpose of ANOVA, we define three terms:

1. The **total sum of squares (SST)** is
$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
2. The **regression sum of squares (SSR)** is given by
$$SSR = \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2$$
it is usually interpreted as the "explained" sum of squares. Whatever that may mean.
3. The **error sum of squares (SSE)** is given by
$$SSE = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)$$
and is usually interpreted as "unexplained" sum of squares.

Theorem 2.21. *For a simple linear regression model, $SST = SSR + SSE$.*

Proof. Observe that
$$y_i - \bar{y} = (\hat{\mu}_i - \bar{y}) + (y_i - \hat{\mu}_i)$$
Taking squares and adding over $1 \leq i \leq n$ yields the required result. ∎

Theorem 2.22. *For the sum of squared errors, we have the following identity:*
$$SSE = \sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i$$

Proof. The proof is by direct algebraic manipulation.
$$\begin{aligned}
SSE &= \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 \\
&= \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \hat{\mu}_i - \sum_{i=1}^{n} y_i \hat{\mu}_i + \sum_{i=1}^{n} \hat{\mu}_i^2 \\
&= \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \hat{\mu}_i - \sum_{i=1}^{n} \hat{\mu}_i (y_i - \hat{\mu}_i) \\
&= \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i (\hat{\beta}_0 - \hat{\beta}_1 x_i) - \sum_{i=1}^{n} \hat{\mu}_i e_i \\
&= \sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i
\end{aligned}$$
as required. ∎

[2] Which can be found here: `http://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/10/lecture-10.pdf`

Theorem 2.23. *For the SSR, we have the following identity,*

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Proof. By direct computation,

$$\begin{aligned} SSR &= \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \end{aligned}$$

as required. ■

Below, we present an ANOVA table. Look at it, and forget it for the rest of your life (modulo the final exam).

Source of Variation	**DF**	**Sum of Squares**	**Mean Squares**	**F**
Regression Model	1	$SSR = \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2$	$MSR = SSR/1$	$F = MSR/MSE$
Error	$n-2$	$SSE = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$	$MSE = \frac{SSE}{n-2}$	
Total	$n-1$	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$		

Definition 2.24. For a simple linear regression model, the F statistic is defined as

$$F = \frac{MSR}{MSE}$$

For a simple linear model, the F statistic is tested against an F distribution with 1 numerator degree of freedom and $n-2$ denominator degrees of freedom.

Definition 2.25. The **coefficient of determination**, R^2 is defined as

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

and it is usually, incorrectly, interpreted as the proportion of the variance "explained" by the model.

Chapter 3

Matrix Algebra and Random Vectors

This first part of this chapter is a review from some facts from MATH 146 and MATH 245. I will state the theorems and definitions without proof. If you wish to see proofs of these statements, please find the set of notes titles "MATH 245 - Fantastic Theorems and How to Prove Them"[1]. Some of the notes here are extracted verbatim from the aforementioned notes.

3.1 Elementary Linear Algebra

Definition 3.1. A **matrix** $A \in M_{m\times n}(\mathbb{F})$ is the rectangular array:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

Yes, I know, it's embarrassing not to remember the order m and n come in. Oops.

Definition 3.2. Let $A \in M_{m\times n}(\mathbb{F})$ and $B \in M_{n\times p}(\mathbb{F})$. We define the **product** of A and B as

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj} \qquad \text{for } 1 \le i \le m, \quad 1 \le j \le p$$

Note that, in general, matrix multiplication is commutative.

Definition 3.3. The **trace** of a matrix is the linear transformation $tr : M_{n\times n}(\mathbb{F}) \to \mathbb{F}$ defined as

$$tr(A) = \sum_{i=1}^{n} A_{ii}$$

Definition 3.4. The **transpose** of a matrix, denoted A^t or A', is its reflection across the main diagonal. Two matrices are **symmetric** if $A^t = A$.

Definition 3.5. Let V be a vector space over $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$. An **inner product** on V is a function that assigns to every ordered pair of vectors $x, y \in V$ a scalar, denoted $\langle x, y\rangle$, such that the following hold:

1. $\langle x + z, y\rangle = \langle x, y\rangle + \langle z, y\rangle$
2. $\langle cx, y\rangle = c\langle x, y\rangle$
3. $\overline{\langle x, y\rangle} = \langle y, x\rangle$

[1] They may be found here: `https://github.com/jlavileze/Fantastic-Theorems`.

4. $\langle x, x\rangle > 0$ if $x > 0$

Definition 3.6. Let V be an inner product space. We say that $v, w \in V$ are **orthogonal** if $\langle v, w\rangle = 0$.

Definition 3.7. Let V be an inner product space. The **norm** of $v \in V$ is the non-negative real number $\|v\| = \sqrt{\langle v, v\rangle}$.

Definition 3.8. A set of vectors $\{v_1, \ldots, v_n\}$ in V is said to be **linearly dependent** if there exists scalars $a_1, \ldots, a_n \in \mathbb{F}$, not all zero, such that

$$a_1v_1 + \ldots + a_nv_n = 0$$

If a set is not linearly dependent, it is said to be **linearly independent**.

Definition 3.9. The **rank** of a matrix $A \in M_{r\times c}(\mathbb{F})$ is the largest number of linearly independent rows or columns[2].

Definition 3.10. We say that a matrix $A \in M_{m\times m}(\mathbb{F})$ is **nonsingular** if its rank is m. The matrix is **singular** otherwise.

Definition 3.11. Let $A \in M_{n\times n}(\mathbb{F})$. If $n = 1$ so that $A = (A_{11})$ we define the determinant of A to be $\det(A) = A_{11}$. For $n \geq 2$, we define the determinant recursively as:

$$\det(A) = \sum_{j=1}^{n} (-1)^{1+j} A_{1j} \cdot \det(\widetilde{A}_{1j})$$

In fact, the determinant can be computed by cofactor expansion along any row or column.

Theorem 3.12. *A matrix is nonsingular if and only if its determinant is nonzero*[3].

Theorem 3.13. *Here are some awesome facts about determinants:*

1. $\det(AB) = \det A \det B$
2. $\det(A^t) = \det(A)$
3. *Determinants are invariant under Type III elementary row operations.*
4. *Multiplying a row or column by a scalar $c \in \mathbb{F}$ scales the determinant by c.*

Theorem 3.14. *If A is invertible then $(A^t)^{-1} = (A^{-1})^t$.*

Definition 3.15. Let V be a finite-dimensional vector space over $\mathbb{F}$. We say a function $Q : V \to \mathbb{F}$ is a **quadratic form** if there exists a symmetric bilinear form $B \in \mathcal{B}(V)$ such that $Q(y) = B(y, y) = y^tHy$ where H is the matrix representation with respect to some basis of B.

Definition 3.16. We say that a symmetric matrix A is **positive definite** if for all non-zero $y \in V$, $y^tAy > 0$. We say it is **semi-positive definite** if $y^tAy \geq 0$.

Definition 3.17. A square matrix A is **orthogonal** if $A^tA = I$. It follows by Theorem 3.13 that $\det(A) = \pm 1$.

Definition 3.18. A square matrix A is said to be **idempotent** if $A^2 = A$.

Theorem 3.19. *The determinant of an idempotent matrix is either zero or one*[4]*. The rank of an idempotent matrix is its trace*[5]*.*

[2] It is an awful idea to define the rank of a linear transformation in terms of a matrix. Apologies to the reader for this heinous crime.

[3] In fact, there is a big equivalence theorem between ranks, reduced row echelon forms, existence of inverses, factorisation into elementary row and column operations, and determinants. The proof of the equivalence is a beautiful exercise in a first course in Linear Algebra. We refer you to Ross Willard's MATH 146 Winter 2017 notes for a statement and a proof of it.

[4] This is trivial.

[5] This is non-trivial; I might post a proof of it at a later date.

3.2 Random vectors and random matrices

Gradients with respect to vectors: $\nabla_x(\boldsymbol{x}^T\boldsymbol{a}) = \boldsymbol{a}$ and $\nabla_x(\boldsymbol{b}\boldsymbol{x}) = \boldsymbol{b}^T$; for a bilinear form $\nabla_x(\boldsymbol{x}^T\boldsymbol{c}\boldsymbol{x}) = (\boldsymbol{c}+\boldsymbol{c}^T)\boldsymbol{x}$, where $\boldsymbol{c} \in M_{p\times p}(\mathbb{F})$.
Expectations of vectors: Let $\boldsymbol{Z}$ be a random vector, $\boldsymbol{a}$ a non-random vector, and $\boldsymbol{c}$ a non-random matrix. Then,

$$\mathrm{E}\left[\boldsymbol{a}\boldsymbol{Z}\right] = \boldsymbol{a}\mathrm{E}\left[\boldsymbol{Z}\right]; \qquad \mathrm{Var}\left(\boldsymbol{Z}\right) = \mathrm{E}\left[\boldsymbol{Z}\boldsymbol{Z}\right]^T - \mathrm{E}\left[\boldsymbol{Z}\right]\left(\mathrm{E}\left[\boldsymbol{Z}\right]\right)^T$$
$$\mathrm{Var}\left(\boldsymbol{c}\boldsymbol{Z}\right) = \boldsymbol{c}\mathrm{Var}\left(Z\right)\boldsymbol{c}^T; \qquad \mathrm{E}\left[\boldsymbol{Z}^T\boldsymbol{c}\boldsymbol{Z}\right] = \mathrm{E}\left[\boldsymbol{Z}\right]^T\boldsymbol{c}\mathrm{E}\left[\boldsymbol{Z}\right] + \mathrm{tr}\left(\boldsymbol{c}\mathrm{Var}\left(\boldsymbol{Z}\right)\right)$$

Facts of life: We stop boldfacing our vectors and matrices now[6]. $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ (furthermore, trace is invariant under cyclic permutations); x, y are orthogonal $\iff \langle x, y\rangle = 0$; $\langle T(x), y\rangle = \langle x, T^*(y)\rangle$, where T^* is the adjoint of T. $\mathrm{rank}(T^*T) = \mathrm{rank}(T)$. Crazy inverse identity: $(A + BCB^T)^{-1} = A^{-1} - A^{-1}B(C^{-1} + B^TA^{-1}B)^{-1}B^TA^{-1}$.

Now we introduce some results about regression.

3.3 The Simple Linear Model in Matrix Form

Theorem 3.20. *Suppose we have n paired observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. If we let $Y = (y_1, y_2, \ldots, y_n)^t$, $\beta = (\beta_0, \beta_1)^t$ and*

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Then the simple linear model can be expressed as the matrix equation

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

where $\epsilon = (\epsilon_i)$ with each $\epsilon_i \sim N(0, \sigma^2)$.

Proof. This is a simple exercise in book-keeping. The vector ϵ is said to be a random vector. ■

Theorem 3.21. *The normal equations for least squares optimisation in matrix form are given by*

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Proof. This should be a simple book-keeping exercise, again. ■

Theorem 3.22. *The least squares estimate for regression is given by*

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

Proof. One way to do so is by noting that the expression given above is the solution to the normal equation when inverting the matrix in the left hand side. Alternatively, we have,

$$\begin{aligned} M = \|e\|^2 &= \|y - X\beta\|^2 \\ &= (y - X\beta)^T(y - X\beta) \\ &= (y^T - \beta^TX^T)(y - X\beta) \\ &= y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta \\ &= y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta \end{aligned}$$

[6]Cool kids don't put funny arrows or fancy fonts on their vectors.

where the last line holds since the matrices we are working with are 1×1 at this stage, so they are all symmetric. Taking the matrix derivative with respect to β, we obtain,

$$\nabla_\beta M = -2X^T y + 2X^T X\beta$$

Setting this to zero yields

$$X^T X\hat{\beta} = X^T y$$

Note that $\text{rank}(X^T X) = \text{rank}(X)$. For simple linear regression, having a rank-deficient design matrix would imply taking observations only at one point. Since we are good scientists and do not do silly things like those, we get for free that $X^T X$ is invertible. Thus,

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

■

Chapter 4

Multiple Linear Regression

Definition 4.1. Let $\{(y_i, x_{i1}, x_{i1}, \ldots, x_{ip})\}_{i=1}^n = \{y_i, \vec{x}_i\}_{i=1}^n$ be a dataset of n grouped data points (paired y_i with a p-dimensional vector $\vec{x}$)[1]. Then, the **general linear model** follows the equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

Definition 4.2. Let $\{(y_i, x_{i1}, x_{i1}, \ldots, x_{ip})\}_{i=1}^n$ be a dataset. Then, the matrix $X \in M_{n\times(p+1)}(\mathbb{R})$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

is called the **design matrix**.

Theorem 4.3. *Let X be a design matrix, $Y = (y_1 \ldots y_n)^t$ and $\beta = (\beta_0 \beta_1 \ldots \beta_p)^t$, then the general linear model can be expressed as the matrix equation*

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

Remark. The $X\beta$ term is called the deterministic component.

Proof. The proof is, again, a trivial book-keeping exercise. ■

Definition 4.4. The **mean squared error** is

$$MSE(\beta) = \frac{1}{n} e^t e$$

where $e = Y - X\beta$.

Theorem 4.5. *The estimate for the parameter vector $\hat{\beta}$ is given by $\hat{\beta} = (X^*X)X^*y$.*

Proof 1. We give a linear algebraic proof first. Note that we want to minimise the quantity $\|y - X\beta\|$. Now, we want to find $\hat{\beta}$ such that

$$\left\|y - X\hat{\beta}\right\| \leq \|y - X\beta\|$$

for all $\beta \in \mathbb{F}^{p+1}$.

Note that $\text{rank}(X^*X) = \text{rank}(X)$. Then if $\text{rank}(X) = n$, then X^*X is invertible. Now let us define the vector space $W = \{X\beta : \beta \in \mathbb{F}^{p+1}\}$; namely $W = R(L_X)$. By the existence of the orthogonal projection (since this is a finite dimensional inner product space), there exists a unique vector which is closest to y, our vector of outcomes. Call this vector $X\beta_0$. Then, we want to solve

$$\|X\beta_0 - y\| \leq \|X\beta - y\|$$

[1] Normal people do not put funny arrow hats on top of their vectors. I'm normal, so I'll only use that once here and never again.

for all $\beta \in \mathbb{F}^{p+1}$. Thus, from our linear algebra toolbox, we observe that $X\beta_0 - y \in W^{\perp}$, so that $\langle X\beta_0, X\beta_0 - y\rangle = 0$ and $\langle x_0, X^*X\beta_0 - X^*y\rangle = 0$. Since $\beta_0 \neq 0$, we obtain

$$X^*X\beta_0 - X^*y = 0$$

which solves for, under the assumption that we are good scientists and X^*X is invertible,

$$\beta_0 = (X^*X)^{-1}X^*y$$

which is our usual regression coefficient vector. Note that we have derived linear regression for arbitrary inner product spaces. That means, you can now perform regression over the vector space of polynomials with your favourite L_p product, over matrices with the Frobenius inner product, or over finite-dimensional subspaces of continuous functions with, say, their standard inner product. Woo-hoo! Viva Linear Algebra! ■

It turns out that the usual calculus proof for the model when working with the real numbers with the standard dot product is exactly the one shown in Chapter 3, so we skip it here and assume the reader is familiar with it. Instead, we spend most of this section generalising the results from Chapter 2.

4.1 Generalising the simple linear model

Theorem 4.6. *The fitted values for a regression model are given by*

$$\hat{\mu} = X\hat{\beta}$$

Proof. From definition. Observe that $\hat{\mu} = X\hat{\beta} = X(X^TX)^{-1}X^Ty = Hy$. The matrix pre-multiplying y has a special name, which we state below. ■

Definition 4.7. The **influence matrix** or **hat matrix** is

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

Theorem 4.8. *The influence matrix is symmetric and idempotent.*

Proof. Trivial. ■

Theorem 4.9. *The residual vector satisfies $e = (I - H)y$.*

Proof. A simple one-liner:

$$e = y - \hat{\mu} = y - X\hat{\beta} = y - Hy = (I - H)y$$

■

Theorem 4.10. *The vector of residuals and the columns of X are orthogonal.*

Proof. A simple calculation:

$$X^Te = X^T(y - X\hat{\beta}) = X^Ty - X^TX(X^TX)^{-1}X^Ty = X^Ty - X^Ty = 0$$

■

Theorem 4.11. *The vector of fitted values, $\hat{\mu}$, and the vector of residuals are orthogonal.*

Proof. Using the inner product,

$$\langle \hat{\mu}, e\rangle = \langle Hy, (I - H)y\rangle = \langle y, H^T(I - H)y\rangle = \langle y, (H - H)y\rangle = \langle y, 0\rangle = 0$$

Hence $\hat{\mu}$ and e are orthogonal. ■

Theorem 4.12. *The least square estimator is unbiased.*

Proof. Turns out that the proofs for these properties become easier with linear algebra. We have that

$$\mathrm{E}\left[\hat{\beta}\right] = (X^TX)^{-1}X^T\mathrm{E}\left[y\right] = (X^TX)^{-1}X^TX\beta = \beta$$

as required. ∎

Theorem 4.13. *The variance of the least squares estimate is given by* $\mathrm{Var}\left(\hat{\beta}\right) = \sigma^2(X^TX)^{-1}$.

Proof. Using the properties of variances of vectors in the previous section, we obtain,

$$\begin{aligned}\mathrm{Var}\left(\hat{\beta}\right) &= \mathrm{Var}\left((X^TX)X^Ty\right) \\ &= (X^TX)^{-1}X^T\mathrm{Var}\left(y\right)X(X^TX)^{-1} \\ &= (X^TX)^{-1}X^T(\sigma^2I)X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}\end{aligned}$$

∎

Remark. In the variance vector defined above, the diagonal entries represent the variances of each individual entry of the parameter vector. The off-diagonal entries represent the corresponding covariances.

Theorem 4.14. *The distribution of the estimator for the parameters follows a multivariate normal distribution. That is,*

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

Proof. The parameters of the normal distribution are derived in the two theorems above. The distribution of the vector is normal since, by assumption, the errors are Gaussian. ∎

Theorem 4.15. *For linear combinations of the parameter vector, which are determined by a vector a, we have:*

1. $\mathrm{E}\left[a^T\hat{\beta}\right] = a^T\beta$
2. $\mathrm{Var}\left(a^T\hat{\beta}\right) = \sigma^2a^T(X^TX)^{-1}a$
3. $a^T\hat{\beta} \sim N(a^T, \sigma^2a^t(X^TX)^{-1}a)$

Proof. Trivial. ∎

Theorem 4.16. *The fitted values* $\hat{\mu}$ *satisfy:*

1. *Unbiasedness.*
2. $\mathrm{Var}\left(\hat{\mu}\right) = H\sigma^2$

Proof. Trivial. ∎

Theorem 4.17. *The distribution of a new prediction satisfies*

$$y_p - \hat{\mu}_p \sim N(0, \sigma^2(1 + a^T(X^TX)^{-1}a))$$

Proof. We write $y_p - \hat{\mu}_p = \mu_p - \hat{\mu}_p + \epsilon_p$. Hence, it is clear that the expectation is zero. The variance is given by

$$\mathrm{Var}\left(y_p - \hat{\mu}_p\right) = \mathrm{Var}\left(\mu_p - \hat{\mu}_p + \epsilon_p\right) = \mathrm{Var}\left(\hat{\mu}_p\right) + \mathrm{Var}\left(\epsilon_p\right) = \sigma^2a^T(X^TX)^{-1}a + \sigma^2$$

as required. Normality arises from the fact that this is a linear combination of random variables. ∎

Remark. From the above statements, we expect the reader to be able to fill in the details for computing confidence intervals, taking care when choosing the number of degrees of freedom for t-distributions (usually $n-p-1$).

Theorem 4.18. *The residual vector satisfy the following distribution*

$$e \sim N(0, (I - H)\sigma^2)$$

Proof. Trivial. ∎

Theorem 4.19. *The vectors $\hat{\beta}$ and e are statistically independent.*

Proof. The proof of this fact requires a little trick using block-matrices. Typesetting it is a pain, so I might include it later. ∎

Theorem 4.20. *The estimate S^2 is an unbiased estimate of σ^2.*

Proof. We use a trick. Note that $\frac{e^Te}{\sigma^2} \sim \chi^2(n-p-1)$. Hence $\mathrm{E}\left[\frac{e^Te}{\sigma^2}\right] = n-p-1$, so that $\mathrm{E}\left[e^Te\right] = \sigma^2(n-p-1)$. Thus, we have that $S^2 = \frac{e^Te}{n-p-1}$ whose expectation is σ^2, using the fact above. ∎

Theorem 4.21. *The residuals e and the fitted values $\hat{\mu}$ are statistically independent.*

Proof. From the above we have that e and $\hat{\beta}$ are statistically independent. Since $\hat{\mu} = X\hat{\beta}$, where X is a non-random matrix, the result follows. ∎

Theorem 4.22. ***Gauss-Markov Theorem.*** *Suppose that the errors in a linear regression model satisfy:*

1. *Zero expectation*

2. *Uncorrelated*

3. *Homoscedastic with finite variance*

then, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares estimator.

Proof. Suppose $\hat{\beta}$ is the OLS estimator and let $\tilde{\beta} = Cy$ be another estimator where we write $C = (X^TX)^{-1}X^T + D$ for some matrix D. Note that

$$\begin{aligned}
\mathrm{E}\left[\tilde{\beta}\right] &= \mathrm{E}\left[Cy\right] \\
&= C\mathrm{E}\left[y\right] \\
&= ((X^TX)^{-1}X^T + D)X\beta \\
&= (X^TX)^{-1}X^TX\beta + DX\beta \\
&= \beta + DX\beta
\end{aligned}$$

Since, by assumption, the estimator is unbiased, we must have that $\beta \in \mathrm{Ker}(DX)$ for all estimates of β. The only matrix that satisfies this is $DX = 0$. Now, we look at the variance of our estimator,

$$\begin{aligned}
\mathrm{Var}\left(\tilde{\beta}\right) &= C\mathrm{Var}\left(y\right)C^T \\
&= C\sigma^2IC^T \\
&= \sigma^2((X^TX)^{-1}X^T + D)((X^TX)^{-1}X^T + D)^T \\
&= \sigma^2((X^TX)^{-1}X^T + D)(X(X^TX)^{-1} + D^T) \\
&= \sigma^2\left[(X^TX)^{-1}X^TX(X^TX)^{-1} + (X^TX)^{-1}X^TD^T + DX(X^TX)^{-1} + DD^T\right] \\
&= \sigma^2\left[(X^TX)^{-1}(X^TX)^{-1}(DX)^T + DX(X^TX)^{-1} + DD^T\right] \\
&= \sigma^2\left[(X^TX)^{-1} + DD^T\right]
\end{aligned}$$

since $DX = 0$. Since DD^T is a positive semi-definite matrix, and $\mathrm{Var}\left(\hat{\beta}\right) = \sigma^2(X^TX)^{-1}$, we have that

$$\mathrm{Var}\left(\tilde{\beta}\right) = \mathrm{Var}\left(\hat{\beta}\right) + DD^T$$

That is, the variance of any other estimator differs from the variance of the OLS by a positive semi-definite matrix, which makes the OLS the unbiased estimator with least variance. ∎

4.2 ANOVA and Partial F-test

The ANOVA table introduced in the Chapter before this is generalised as follows:

Source of Variation	**DF**	**Sum of Squares**	**Mean Squares**	**F**
Regression Model	p	$SSR = \hat{\beta}^T X^T y - n\bar{y}^2$	$MSR = SSR/p$	$F = MSR/MSE$
Error	$n-p-1$	$SSE = y^T y - \hat{\beta}^T X^T y$	$MSE = \frac{SSE}{n-p-1}$	
Total	$n-1$	$SST = y^T y - n\bar{y}^2$		

The remainder of the section is a bit of a pain. I'll post notes later.

4.3 Generalised Least Squares

For a multiple linear regression model, we made pretty strong assumptions on the distribution of the errors. Namely, we assumed they were independent and homoscedastic. In practice, however, these assumptions are not true with probability 1. So we need to find a way to fix our models when such calamities occur.

In class, we looked first at Generalised Least Squares and then discussed a special case: Weighted Least Squares. In this section, we introduce the latter first, before developing the theory in full generality.

4.3.1 Non-constant noise

Suppose the errors are not homoscedastic. That is, they are heteroscedastic. Then, instead of minimising the mean square error, we attempt to minimise the weighted mean square error.

Theorem 4.23. *For a linear model, the weighted least squares estimate is given by*

$$\hat{\beta}_{WLS} = (X^T W X)^{-1} X^T W y$$

where W is a matrix with weights on its diagonal and zeroes everywhere else.

Proof. Instead of minimising the quantity

$$M = \sum_{i=1}^{n} (y_i - \text{Row}_i(X)\beta)$$

we instead minimise the weighted mean

$$S = \sum_{i=1}^{n} w_i (y_i - \text{Row}_i(X)\beta)$$

If we write the matrix $W = (\delta_{ij} w_i)_{ij}$ where δ_{ij} is the Kroenecker delta, then this is equivalent to minimising

$$S = (y - X\beta)^T W (y - X\beta) = y^T W y - y^T W X \beta - \beta^T X^T W y + \beta^T X^T W X \beta$$

which simplifies to

$$S = y^T W y - 2\beta^T X^T W y + \beta^T X^T W X \beta$$

since these quantities are single-entry matrices. Indeed, we take the gradient with respect to β to obtain,

$$\nabla_\beta S = -2X^T W y + 2X^T W X \beta$$

Setting this quantity to zero, we obtain

$$X^T W X \hat{\beta}_{WLS} = X^T W y$$

And if we are smart cookies, we have ensured that X^TWX is invertible, so that the estimate yields

$$\hat{\beta}_{WLS} = (X^TWX)^{-1}X^TWy$$

■

A weak remark. As a sanity check, we note that if $W = I$, the equation above returns our OLS estimate.

A strong remark. Why in the world would we want to do this? Clearly, this is computationally more expensive and the choice of the matrix W needs some wizardry. Suppose we are wizards and can choose W perfectly, whatever that may mean in the real world. Then, there are three reasons for taking a least squares estimate[2]:

1. *Focusing estimation accuracy to subset of domain:* Some values of our predictors may occur more often, may be expensive to predict, or may be costly if predictions are mistaken. Setting high weights in such regions will ensure that estimates dominate in such region.

2. *Decreasing imprecision:* Under the homoscedastic assumption, the Gaus-Markov Theorem holds, but not so under the heteroscedastic mayhem. In such situation, in fact, we can set $w_i = \frac{1}{\sigma_i}$, provided that we visited the oracle and she gave us the true noises. It is, furthermore, a waste of time to treat every possible point equally. Thus, in estimation, we should give more attention to the case where the noise is small.

3. *Sampling corrections:* Sampling biases happen; deal with it. We may be interested in giving more weight to those values we have undersampled and less weight to those we have oversampled. We can, in fact, set weights to be the reciprocal of the sampling probability (provided that the oracle has given us the exact estimate of the bias), then we might achieve a more parsimonious survey weighting.

There is much more that can be said about weighted regression. We refer you to the reference in this page's footnote if you are interested. Instead, we move on to generalised least squares regression.

4.3.2 Correlated Noise

Suppose we have strong reason to believe that the model

$$y = X\beta + \epsilon \qquad \mathrm{E}\left[\epsilon\right] = 0 \qquad \mathrm{Var}\left(\epsilon\right) = \Sigma$$

is the right model, where Σ is *not* a diagonal matrix; that is, the errors are correlated (the off-diagonal entries in the matrix Σ are actually the covariances between the errors). This is in fact the most common case in real life, as there is no heuristic (let alone, mathematical) reason to believe that the errors are uncorrelated. Either way, we want to estimate β in the model above.

Theorem 4.24. *Suppose a process can be modelled under the assumption that*

$$y = X\beta + \epsilon \qquad \mathrm{E}\left[\epsilon\right] = 0 \qquad \mathrm{Var}\left(\epsilon\right) = \Sigma$$

where Σ is a variance matrix (that is, it is symmetric and positive-definite). Then the generalised least squares estimate is

$$\hat{\beta}_{GLS} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$$

Proof. We begin with a lemma.

Lemma. The variance matrix Σ has a square root. That is, there exists a matrix Γ with $\Gamma^2 = \Sigma$.

Proof of Lemma. Since Σ is a self-adjoint matrix over a real field, it is orthogonally diagonalisable. That is, $\Sigma = P^TDP$, where D is diagonal. Since Σ is positive-definite, all its eigenvalues are positive, so that the entries in D are zero in the off-diagonal and strictly positive in the diagonal. Thus, we can write $U = (\delta_{ij}\sqrt{\lambda_i})$, and observe that $U^2 = D$. Thus, $\Sigma = P^TU^2P = (P^TUP)^T(P^TUP)$. Let $\Gamma = P^TUP$ and we have that $\Sigma = \Gamma\Gamma^T$, and we are

[2]From C. Shalizi. Modern Regression, Lecture 24-25

done. ■

Now we are ready to prove the theorem. Observe that in our lemma, the matrix Γ is invertible, since it is a matrix with positive eigenvalues (the determinant is the product of the eigenvalue; since all are positive, the determinant is positive). We left-multiply the model by Γ^{-1} to obtain

$$\Gamma^{-1}y = \Gamma X\beta + \Gamma^{-1}\epsilon \qquad (*)$$

We now wonder what happened to the distribution of our errors? Our linear transformation yields

$$\mathrm{E}\left[\Gamma^{-1}\epsilon\right] = \Gamma^{-1}\mathrm{E}\left[\epsilon\right] = 0$$

and,

$$\mathrm{Var}\left(\Gamma^{-1}\epsilon\right) = \Gamma^{-1}\mathrm{Var}\left(\epsilon\right)(\Gamma^{-1})^T = \Gamma^{-1}\Sigma(\Gamma^{-1})^T = \Gamma^{-1}(\Gamma\Gamma^T)(\Gamma^{-1})^T = I$$

That is, the transformed model $(*)$ satisfies the assumptions of homoscedastic Gaussian uncorrelated errors, which we had for the original model. Good-ie! We can now perform OLS as usual to obtain,

$$\hat{\beta}_{GLS} = \left((\Gamma^{-1}X^T)^T\right)$$

Chapter 5

Specification Issues in Regression

5.1 One and two sample problem

5.2 Polynomial models

5.3 Systems of straight lines

5.4 K-sample problem

5.5 Multicollinearity

5.6 Orthogonal Parameters

Chapter 6

Model Checking

6.1 Residual Analysis

6.1.1 Diagnostic Plots

6.1.2 Correlated Errors