STAT 331 - Applied Linear Models

Fantastic Models and How to Abuse Them

Jose Luis Avilez
Faculty of Mathematics
University of Waterloo

Contents

1 Preliminaries **2**

2 Simple Linear Regression **3**
2.1 Estimating Simple Linear Regression . . . 3
2.2 Inference in Simple Linear Regression . . . 5
2.3 ANOVA and R^2 . . . 7

3 Matrix Algebra and Random Vectors **9**
3.1 Elementary Linear Algebra . . . 9
3.2 Random vectors and random matrices . . . 11
3.3 The Simple Linear Model in Matrix Form . . . 11

4 Multiple Linear Regression **13**
4.1 Generalising the simple linear model . . . 14
4.2 ANOVA and Partial F-test . . . 17
4.3 Generalised Least Squares . . . 17
4.3.1 Non-constant noise . . . 17
4.3.2 Correlated Noise . . . 18

5 Specification Issues in Regression **20**
5.1 One and two sample problem . . . 20
5.2 Polynomial models . . . 20
5.3 Systems of straight lines . . . 21
5.4 Multicollinearity . . . 21
5.5 Orthogonal Parameters . . . 22

6 Model Checking **24**
6.1 Residual Analysis . . . 24
6.1.1 Diagnostic Plots . . . 24
6.1.2 Correlated Errors . . . 27
6.2 Effect of individual data-points . . . 28
6.2.1 Outliers . . . 28
6.3 Influential and Leverage Points . . . 29
6.4 Assessing the Adequacy of the Functional Form . . . 31

7 Model Selection **32**
7.1 Criterion-based methods . . . 32

8 Nonlinear regression models **34**

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 not 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

$$a_1 v_1 + \ldots + a_n v_n = 0$$

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 $Q(y) = B(y, y) = y^t H y$ 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 $y \in V$, $y^t A y > 0$. We say it is **semi-positive definite** if $y^t A y \geq 0$.

Definition 3.17. A square matrix A is **orthogonal** if $A^t A = I$. 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 $\langle \beta_0, X^*X\beta_0 - X^*y\rangle = 0$. 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, idempotent, and satisfies* $\frac{\partial \hat{\mu}_i}{\partial y_j} = H_{ij}$.

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

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

as required. ■

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

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

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

■

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,*

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

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

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*

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

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

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

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 $\frac{e^T e}{\sigma^2} \sim \chi^2(n-p-1)$. Hence $\mathrm{E}\left[\frac{e^T e}{\sigma^2}\right] = n-p-1$, so that $\mathrm{E}\left[e^T e\right] = \sigma^2(n-p-1)$. Thus, we have that $S^2 = \frac{e^T e}{n-p-1}$ 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 $C = (X^T X)^{-1} X^T + D$ for some matrix D. Note that

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

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,

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

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

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

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*

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

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,

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

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,

$$\begin{aligned}
\hat{\beta}_{GLS} &= \left((\Gamma^{-1}X^T)^T(\Gamma^{-1}X)\right)(\Gamma^{-1}X)^T y \\
&= (X^T(\Gamma^T)^{-1}\Gamma^{-1}X)^{-1}X^T(\Gamma^T)^{-1}\Gamma^{-1}y \\
&= (X^T(\Gamma\Gamma^T)^{-1}X^T)X^T(\Gamma\Gamma^T)^{-1}y \\
&= (X^T\Sigma^{-1}X^T)X^T\Sigma^{-1}y
\end{aligned}$$

as desired. ■

Chapter 5

Specification Issues in Regression

In this chapter, we look at specific cases that may arise in regression, or particular problems which lend themselves to the methods developed above. Since most of these are examples, I shall keep this chapter short.

5.1 One and two sample problem

In the **one-sample problem** we are interested whether the mean of a uniform process takes some particular value. This can be solved by performing a one-sample t-test. Observe that this can be posed as a regression problem in determining the value β for $y_i = \beta + \epsilon_i$. In this case our design matrix is a column vector of ones.

Suppose instead that we sample our dataset from two distinct set of conditions. We call this the **two-sample problem**. We can pose this as a regression problem by writing

$$y_i = \begin{cases} \beta_1 + \epsilon_i & 1 \leq i \leq m \\ \beta_2 + \epsilon_i & m+1 \leq i \leq n \end{cases}$$

Alternatively it can be written as $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ where x_{ij} for $j = 1, 2$ are the appropriate indicator variables. Note that $\mathrm{E}[y_i] = \beta_1 x_{i1} + \beta_2 x_{i2}$. Additionally, note that our design matrix looks as follows

$$X = \left(\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{array}\right)$$

We can now test the hypothesis that $\beta_1 = \beta_2$ using the usual methods. Extending this to the K-sample problem should be trivial by now.

5.2 Polynomial models

Suppose a straight line makes for a poor fit. We can then try to fit a polynomial (which is still linear in the coefficients) to our dataset. We explore the case where $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, with the usual assumptions for a linear model. Generalising this should be trivial. In this case, our expectation is $\mathrm{E}[y_i] = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$. We can treat x_i^2 as a new predictor to obtain the following design matrix:

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$$

We can now find the OLS estimate in the usual way. Extrapolation to higher powers or multiple predictors with higher powers is as expected. Two caveats, though:

1. We may visit the oracle and ask her what the true polynomial model is and attempt fitting our dataset to such parameters. If we are not devilish enough to have access to her, we may incur in the atrocity of choosing an arbitrary order for our polynomial. Obviously any p points completely determine a polynomial of degree $p-1$. However, oscillations may be wild so doing this would, probably be a bad idea as it tends to lead to overfitting.

2. Interpreting the parameters is not as straightforward. We can no longer say that they represent the expected change in y for a one-unit increase in a parameter x. Instead, careful case-by-case analysis would be necessary.

5.3 Systems of straight lines

We introduce indicator variables; there are two cases to consider. The first is where we deal with a contrast and expect that the presence of something (its indication) will yield a parallel regression surface. The model in this case takes the form

$$y = \beta_0 + \beta_B X_B + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

where X_B is a binary variable. That is, the effect of the binary variable is the same across the whole domain. In the second case, the presence of this new variable may change with changes in other predictors, so we write

$$y = \beta_0 + \beta_1 x_i + \beta_2 t_i + \beta_3 x_i t_i + \epsilon$$

IN this case we obtain two skew regression surfaces. For this case we add two new columns, one for the indicator and one for the values of the variable t_i when $x_i = 1$. With these constructions we can then test whether the presence of the new variable is significant to the model.

5.4 Multicollinearity

All throughout we assumed we were smart scientists and made sure to construct our design matrix so that the columns were linearly independent. However, this is not always possible. Recall that $\operatorname{rank}\left(X^T X\right) = \operatorname{rank}(X)$. Then if X is column-rank deficient, we will not be able to compute $(X^T X)^{-1}$. However, even if the matrix is invertible, a near-singular matrix (in the sense that its determinant is close to zero) may still be problematic. This is because the inverse depends on the reciprocal of the determinant of the original matrix and thus its entries will explode.

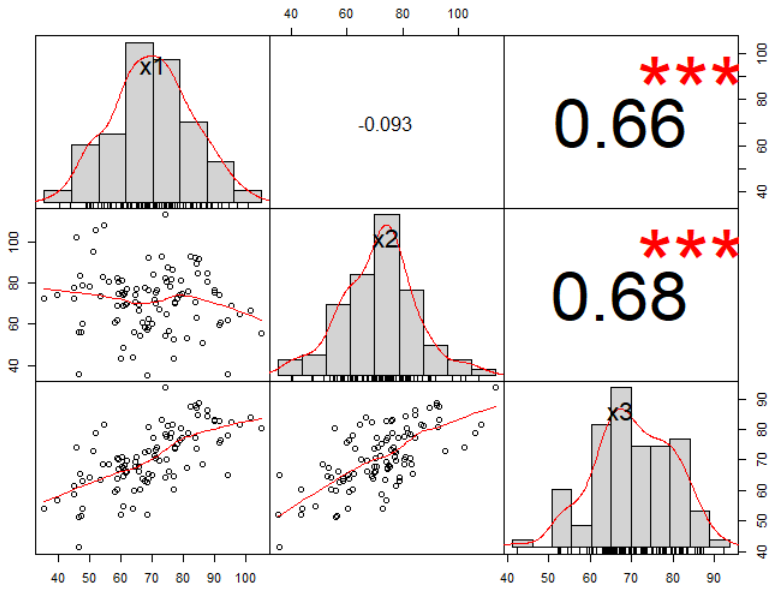
To detect pairwise collinearity, we may plot the correlation matrix and look for correlation values which are close to one. As a rule of thumb, we try to delete values of correlation over 0.8, as they indicate strong linear relationships between two variables. If these exist, it is strong indication that our matrix may be near singular. However, multicollinearity may be harder, as displayed in the example below.

Example 5.1. Suppose $X_1, X_2 \sim N(0, \sigma^2)$ independently. Define $X_3 = \frac{X_1 + X_2}{2}$. Evidently, the set $\{X_1, X_2, X_3\}$ is linearly dependent, but the correlation matrix may fail to reveal it. Note that

$$\begin{aligned}
\operatorname{Corr}(X_1, X_3) &= \frac{\operatorname{Cov}(X_1\,,\, X_3)}{\sqrt{\operatorname{Var}(X_1)\operatorname{Var}(X_3)}} \\
&= \frac{\operatorname{Cov}\left(X_1\,,\, \frac{X_1+X_3}{2}\right)}{\sqrt{\sigma^2 \cdot \frac{\sigma^2}{2}}} \\
&= \frac{\sigma^2/2}{\sigma^2/\sqrt{2}} \\
&= \frac{\sqrt{2}}{2} \\
&\approx 0.7
\end{aligned}$$

which is not easy to distinguish in a correlation matrix, as shown below.

Figure 5.1: A case of perfect multicollinearity showing weakly on the correlation matrix.




```
library("PerformanceAnalytics")
# Simulation for independent normal variables
x1 <- rnorm(100, mean=70, sd=15)
x2 <- rnorm(100, mean=70, sd=15)
# Add in a linear combination of X1 and X2
x3 <- (x1+x2)/2
df <- data.frame(x1,x2,x3)
chart.Correlation(df, histogram=TRUE, pch=19)
```

To help us diagnose this we may note that in the case of multicollinearity we may observe low p-values for the overall fit of the model but high p-values for the independent regressors. Furthermore, we may observe a high R^2 value, with a highly significant F-test with no significant t-statistics. We will also note that the parameter estimates may vary wildly when a different combination of predictors is chosen. We introduce a few extra statistics which provide a nice rule of thumb for testing for multicollinearity.

Definition 5.2. Suppose $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$ is a linear model. We define the j-th **tolerance** as $1 - R_j^2$ where R_j^2 is the R^2 value of the estimated model for

$$x_j = \beta_0^* + \beta_1^* x_1 + \ldots + \beta_{j-1}^* x_{j-1} + \beta_{j+1}^* x_{j+1} + \ldots + \beta_p^* x_p$$

The j-th **variable inflation factor** is

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

Remark. As a rule of thumb, we use $VIF_j > 10$ as an indication that β_j is subject to multicollinearity. We use variable inflation factors usually, rather than tolerance, as they indicate the factor by which standard errors for a particular coefficient have been inflated.

5.5 Orthogonal Parameters

Now, suppose that the vectors in the matrix X are orthogonal, then we will have that $X^T X$ is a diagonal matrix, whose inverse $(X^T X)^{-1}$ is simply the reciprocal of the diagonal and zero elsewhere. This makes the formulation of the model extremely nice. We state a few results below but do not prove them as they have been proven in assignments.

Theorem 5.3. *The estimates for the coefficients are the same as the ones obtained from estimating them in a simple linear model.*

Remark. A simple corollary of this is that $\hat{\beta}_1 = \frac{X_i^T y}{X_i^T X_i}$.

Theorem 5.4. *For a linear model with orthogonal parameters, the SSR is additive:*

$$SSR(x_1, \ldots, x_p) = SSR(x_1) + \ldots + SSR(x_p)$$

Remark. It follows that $SSR(x_i) = \hat{\beta}_i(X_i^T y)$.

Theorem 5.5. *Suppose in a linear model the independent variables are orthogonal, then* $\text{Cov}\left(\hat{\beta}_i\,,\,\hat{\beta}_j\right) = 0$.

Chapter 6

Model Checking

There are many reasons why models may be inadequate. For instance, the functional form of the model may be absolutely wrong and thus need some tweaking. On the other hand, the specification of the errors may be incorrect in that the errors are not heteroscedastic, the errors are not normally distributed, or they are not independent. We also want to determine whether our data has outliers.

In this chapter we study how to diagnose and treat such pitfalls.

6.1 Residual Analysis

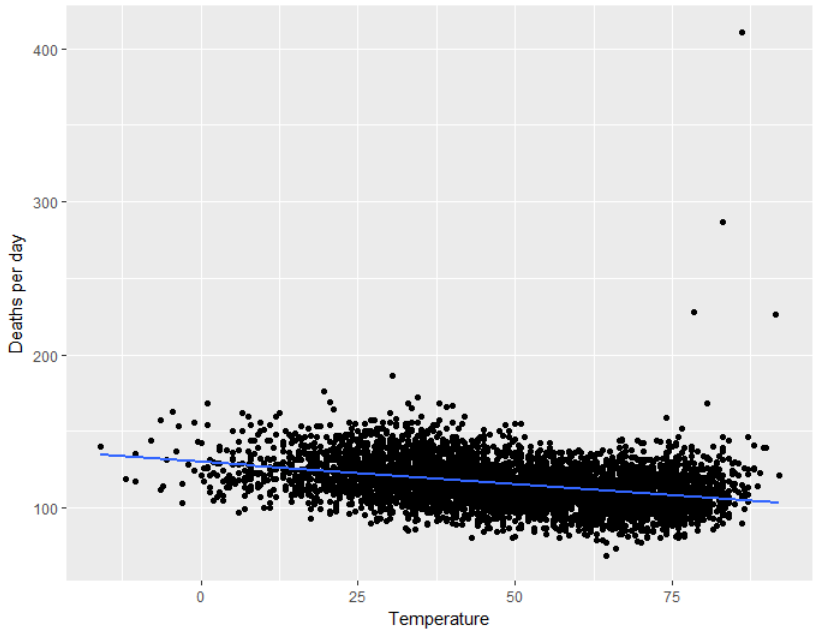
Since the residual vector e and the vector of fitted values $\hat{\mu}$ are independent, the latter carries no information about the former. Thus, plot of the former yield plenty of information which we did not originally have.

6.1.1 Diagnostic Plots

The first diagnostic plot is a plot of residuals vs fitted values. We should expect to see a random patters around zero, where the width of the pattern is constant when the regression assumptions hold. We showcase one example for a regression problem involving finding a relationship between mortality and temperature.

Example 6.1. Given data from Chicago, we attempt to find the coefficients for the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where x_i represents temperature and y_i represents the daily death rate. Using the base R package for Chicago data we obtain the following regression line:

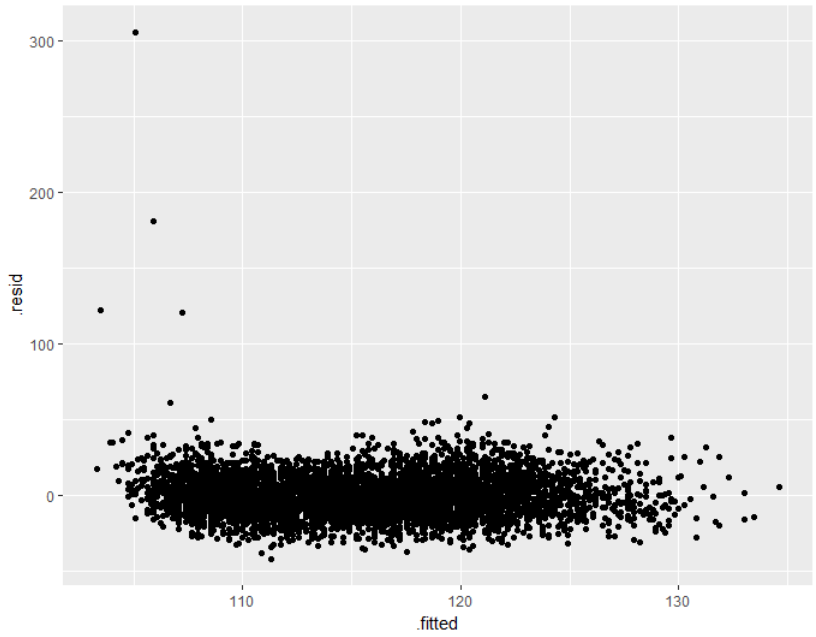
Figure 6.1: A plot of the regression line for Chicago data.




```
library(gamair)
library(ggplot2)
data(chicago)
# Plot deaths each day vs. temperature
plt <- ggplot( data=chicago , aes ( tmpd , death ) ) + geom_point()
plt <- plt + labs(x = "Temperature", y = "Deaths␣per␣day")
d.temp.m <- lm(death ~ tmpd , data=chicago)
plt + geom_smooth(method='lm', formula = y~x, se=FALSE)
```

The residuals of this model estimation can be seen in the diagnostic plot below:

Figure 6.2: A plot of the residuals for the model in the figure above.




```
res <- ggplot(lm(death ~ tmpd , data=chicago)) + geom_point(aes(x=.fitted, y=.resid))
res
```

We observe that for most of the range of fitted values, the residuals behave according to the assumptions of the errors. However, such a pattern breaks at lower values of our outcome variable, which merits further exploration.

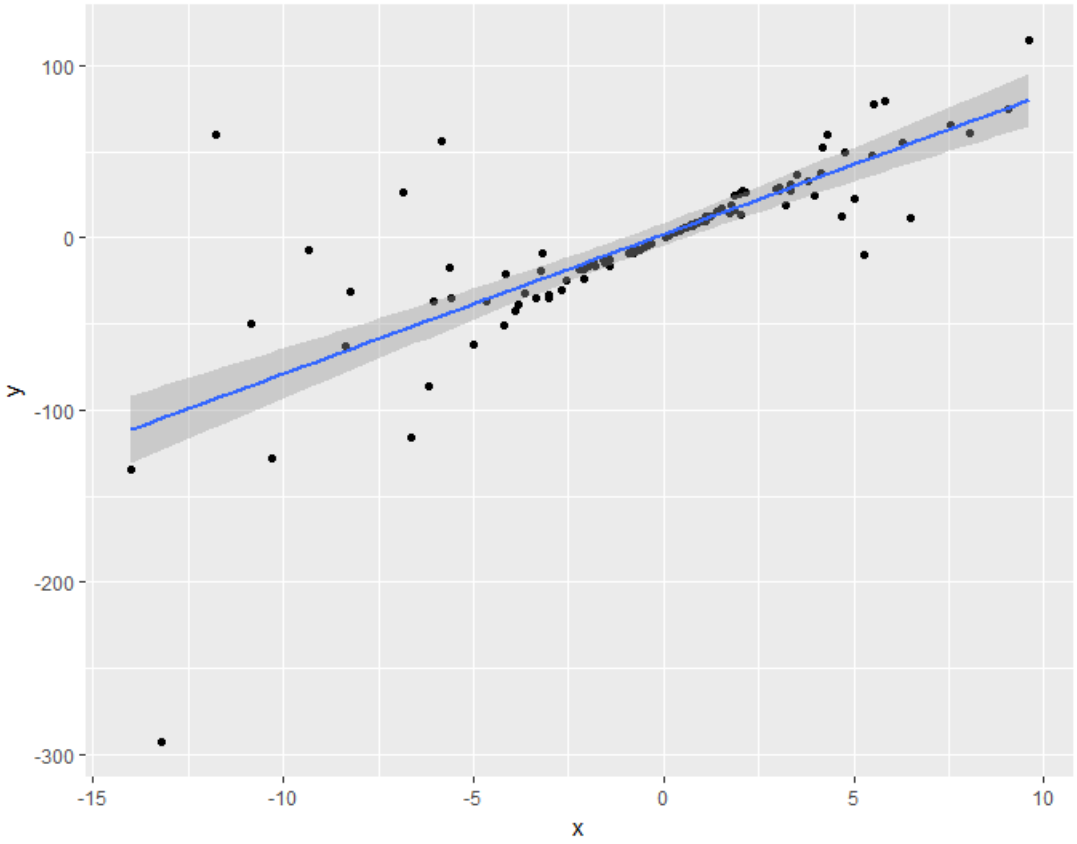
In the figure above, the errors were homoscedastic throughout most of the fitted values. We now look at an example where that assumption breaks down completely.

Example 6.2. We produce some dummy data with heteroscedastic errors from the code below:

```
x <- rnorm(mean=0, sd = 5, n = 100)
y <- 10 * x + rnorm(n=100, mean = 0, sd = x^2)
het.df <- data.frame(x,y)
het.plt <- ggplot(data = het.df, aes(x,y)) + geom_point()
het.plt + geom_smooth(method = 'lm', formula = y~x)
```

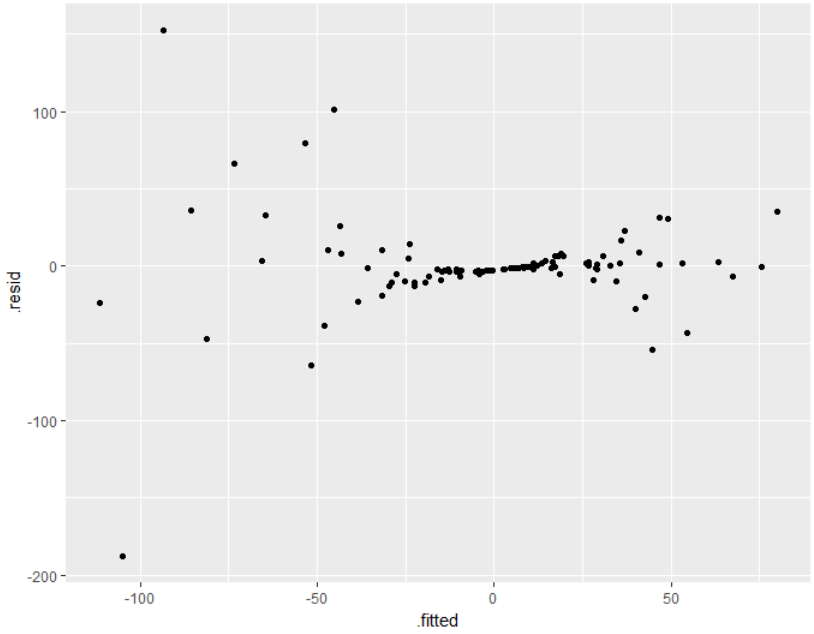
Which in turn produces the following plot with its corresponding fitted model:

Figure 6.3: A plot derived from data with a heteroscedastic errors, showing confidence intervals for the standard errors



whose heteroscedastic errors become evident in the residual plot below:

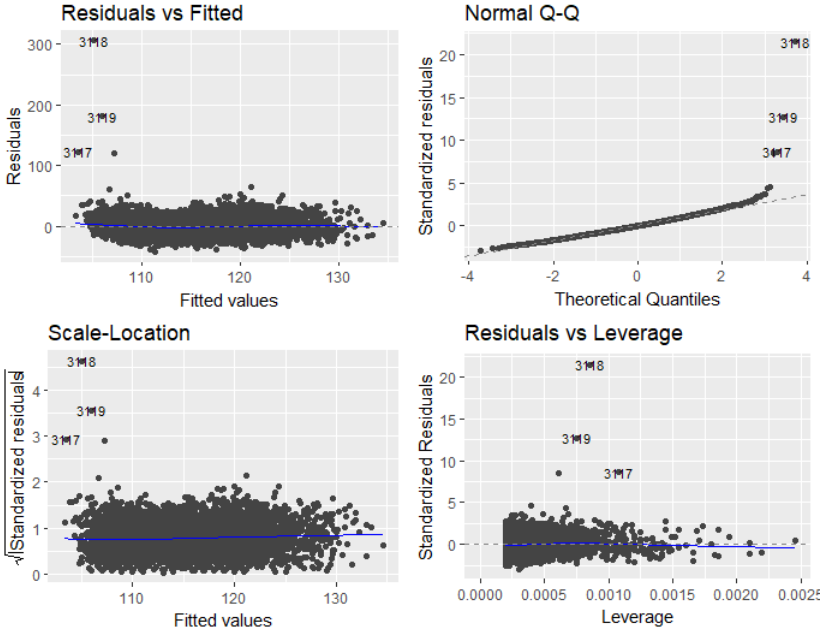
Figure 6.4: Residual plot showing heteroscedastic errors.



There are plenty more diagnostic plots that can be built. Amazingly, R has a package that can build all the diagnostic plots with a single line of code.

```
library(ggfortify)
autoplot(d.temp.m, label.size = 3)
```

Figure 6.5: Other diagnostic plots.



6.1.2 Correlated Errors

Definition 6.3. A **time series** is a dataset indexed by time. It is usually sampled at equal intervals of time.

Definition 6.4. A time series is said to be **weakly stationary** if it satisfies the following properties:

1. The mean $\mathrm{E}[x_t]$ is the same for all t.
2. The variance $\mathrm{Var}(x_t)$ is the same for all t.

3. The covariance between x_t and x_{t-k} is the same for all t.

Definition 6.5. A serially correlated regression model is one in which,

$$\begin{aligned} y_t &= X_t\beta + \epsilon_t \\ \epsilon_t &= \rho\epsilon_{t-1} + w_t \\ |\rho| < 1 &\text{ and } w_t \sim N(0, \sigma^2) \text{ (i.i.d.)} \end{aligned}$$

Definition 6.6. The **lag-k autocorrelation** of a time series model is defined as

$$r_k = \frac{\sum_{t=k+1}^{n} e_t e_{t-k}}{\sum_{t=1}^{n} e_t^2}$$

Remark. If the errors in the regression model are uncorrelated then we would expect $\mathrm{E}[r_k] \approx 0$ and $\mathrm{Var}(r_k) \approx \frac{1}{n}$.

As a diagnostic plot, we may try plotting r_k against the lag k and placing horizontal bands at twice the standard errors. Sample autocorrelation functions outside those ranges are evidence pointing towards autocorrelation.

Definition 6.7. The **Durbin-Watson** test statistic is defined as

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Its purpose is to detect the presence of autocorrelation in the residuals from a regression analysis.

Interpreting the Durbin-Watson statistic. It turns out that the D statistic is approximately equal to $2(1-r_1)$ where r_1 is the sample autocorrelation of the residuals. In that case, we note that D lies between 0 and 4. A value of $D = 2$ indicates no autocorrelation. A value of D substantially less than 2 indicates evidence of positive serial autocorrelation. A value larger than two indicates negative serial autocorrelation.

6.2 Effect of individual data-points

In this section we discuss the effect of outliers, leverage points, and influential points.

6.2.1 Outliers

Definition 6.8. We provide a pseudo-definition for outlier. An **outlier** is a data point whose y value is "unusual" given the value of its predictor variables. Such a point will have a high residual.

Definition 6.9. A **studentised residual** for outliers on y is defined as

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

which follow a $N(0, 1)$ distribution, provided that the model assumptions are satisfied.

Remark. We say that values $|d_i| > 2.5$ are unexpected in the y dimension.

Treatment of outliers. There are many reasons for why outliers may occur. In some cases, such as misrecordings, such cases can be dropped. Otherwise, we may attempt to fit the model with and without the outliers. If the conclusions do not vary wildly, we can keep the analysis with them. Otherwise, we should strive to look for more data.

6.3 Influential and Leverage Points

We introduce a few more pseudo-definitions in this section.

Definition 6.10. A point is said to have **high leverage** if the value on its predictor variables x_i is extreme. We say that **leverage** is a measure of how much independent variables deviate from their means.

It is important to note that leverage depends solely on the value of the predictors and not of the outcome variable.

Definition 6.11. More formally, we can define the **leverage** of the i-th observation as H_{ii}, the i-th diagonal entry in the influence matrix.

Theorem 6.12. *The following are true for h_{ii}:*

1. $\frac{1}{n} \leq h_{ii} \leq 1$
2. $\hat{\mu} = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j$
3. $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$
4. *For the case with a single explanatory variable,*

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

Proof. The first three are left as exercises. For the last one, we use a trick. Note that for any $p \in \mathbb{Z}^+$ we have that $H_{ij} = \frac{\partial \hat{\mu}_i}{\partial y_j}$ (this comes from the fact that $\hat{\mu} = Hy$ and simply observing the corresponding entry in the derivative matrix). For the case $p = 1$, we have

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

If we substitute in $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$ and take the required partial derivative we should get the desired result. ■

Remark. Note that the leverage is smallest when $x_i = \bar{x}$.

Theorem 6.13. *The mean leverage is $\bar{h} = \frac{p+1}{n}$.*

Proof. The sum of the leverages is the trace of the influence matrix. Since the trace is invariant under cyclic permutations of the products, we obtain:

$$\begin{aligned} \text{tr}(H) &= \text{tr}(X(X^TX)^{-1}X^T) \\ &= \text{tr}((X^TX)^{-1}(X^TX)) \\ &= \text{tr}(I_{p+1}) \\ &= p + 1 \end{aligned}$$

Thus, the mean leverage is $\bar{h} = \frac{p+1}{n}$. ■

Theorem 6.14. *The diagonal elements of h_{ii} are given by*

$$h_{ii} = X_i^T(X^TX)^{-1}X_i$$

Proof. Follows from simple matrix algebra. ■

Definition 6.15. With the pre-amble above, we can be more precise in our definition of **high leverage**. We say that a point has high leverage if

$$h_{ii} > 2\bar{h} = \frac{2(p+1)}{n}$$

Definition 6.16. *Pseudo-definition.* An **influential point** in one which, when removed from the model's estimation, would cause a marked change in the statistical analysis.

Remark. Observations which have large hat diagonals and large residuals are likely to be influential.

Definition 6.17. Cook's statistic is defined as

$$D_i = \frac{(X\hat{\beta} - X\hat{\beta}_{(i)})^T (X\hat{\beta} - X\hat{\beta}_{(i)})}{(p+1)s^2}$$

where $\hat{\beta}_{(i)}$ is the estimated parameter vector when removing the i-th observation.

Theorem 6.18. *The value for Cook's statistic can be re-written as:*

$$D_i = \frac{(\hat{\mu} - \hat{\mu}_{(i)})^T (\hat{\mu} - \hat{\mu}_{(i)})}{(p+1)s^2}$$
$$D_i = \frac{e_i^2 x_i^T (X^T X)^{-1} x_i}{(1-h_{ii})^2 (p+1)s^2}$$

Proof. Exercise from HW question. Some of the work might be simplified by the theorem below. ■

Theorem 6.19. *The estimated parameter vector when removing a prediction can be computed as follows:*

$$\hat{\beta}_{(i)} = \hat{\beta} - \left(\frac{e_i}{1-h_{ii}}\right)(X^T X)^{-1} X_i$$

Proof. Exercise from practice questions. ■

There is no broad agreement on how to interpret Cook's statistic. Some people argue that values above 0.5 should be examined; others argue that values over 1 should be further investigated. A separate camp believes that the cut off point to determine an influential point as $\frac{4}{n}$. Pick one, justify it, stick to it.

Definition 6.20. The following definitions are about **PRESS** statistics. The **prediction error** of the i-th observation is defined as

$$e_{(i)} = y_i - \hat{\mu}_{(i)} = y_i - x_i \hat{\beta}_{(1)}$$

The **predicted residual error sum of squares (PRESS)** is defined as

$$\sum_{i=1}^{n} e_{(i)}^2$$

The **PRESS statistic** is defined as

$$PRESS = \sum_{i=1}^{n} y_i - \hat{\mu}_{(i)}$$

Theorem 6.21. *The PRESS statistic satisfies*

$$PRESS = \sum_{i=1}^{n} y_i - \hat{\mu}_{(i)} = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2$$

Proof. By a theorem above,

$$\begin{aligned} e_{(i)} &= y_i - x_i \hat{\beta}_{(1)} \\ &= y_i - x_i \left[\hat{\beta} - \frac{e_i (X^T X)^{-1} x_i}{1-h_{ii}}\right] \\ &= (y_i - \hat{\mu}_i) + \frac{e_i x_i^T (X^T X)^{-1} x_i}{1-h_{ii}} \\ &= e_i + \frac{e_i h_{ii}}{1-h_{ii}} \\ &= \frac{e_i}{1-h_{ii}} \end{aligned}$$

This is the summand, so the result follows. ■

Remark. The PRESS statistic is usually regarded as a measure of how well the model will perform at predicting new data. The best model structures arise when the value of PRESS is lowest.

Definition 6.22. The **standardised difference between the fitted value (DEFITS)** is defined as

$$DEFITS_i = \frac{e_i h_{ii}}{(1-h_{ii})\sqrt{s^2_{(i)} h_{ii}}}$$

where

$$s^2_{(i)} = \frac{\sum_{i \neq j} y_j - x_j^T \hat{\beta}_{(i)}}{n-p-2}$$

is the unbiased estimate of σ^2 without the i-th observation. This quantity is equivalent to

$$s^2_{(i)} = \frac{(n-p-1)s^2 - \frac{e_i^2}{1-h_{ii}}}{n-p-2}$$

Remark. DEFITS measures how many standard deviations the fitted value changes if the i-th observation is removed. Values of DEFITS larger than $2\sqrt{\frac{p}{n}}$ are considered influential.

6.4 Assessing the Adequacy of the Functional Form

I'll post notes on the lack of fit hypothesis test later.

Chapter 7

Model Selection

In this chapter, we cover two ways to select models:

1. Criterion-based methods (e.g. AIC, C_p, PRESS)
2. Automatic methods (e.g. backward elimination, forward selection, stepwise regression)

7.1 Criterion-based methods

Suppose we fit the model $y = X\beta + \epsilon$ where X is the design matrix for a dataset with q explanatory variables. How many different regression models are possible assuming an additive first order functional form? Obviously, we can fit up to 2^q different regression models. Ideally we want to choose a model

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

where $p \leq q$ such that no important variable is left out of the model and no unimportant variable is included in it.

One method that can be used is maximising R^2 for regression. However, since R^2 increases monotonically with the number of parameters, we need to adjust it. We introduce the following definition.

Definition 7.1. For a linear model, the **adjusted R^2 value** is

$$R^2_{adj,p} = 1 - \frac{SSE_p/(n-p-1)}{SST/(n-1)} = 1 - \frac{S_p^2}{S_y^2} = 1 - \frac{n-1}{n-p-1}(1-R_p^2)$$

As usual, we are not big fans of R^2 values, so we leave that discussion there. Instead, we introduce a few fantastic and ingenious criteria.

Definition 7.2. The **Akaike Information Criterion (AIC)** is defined as

$$AIC_p = n \ln\left(\frac{SSE_p}{n}\right) + 2(p+1)$$

Definition 7.3. The **Bayes Information Criterion (BIC)** is defined as

$$BIC_p = n \ln\left(\frac{SSE_p}{n}\right) + (p+1)\ln(n)$$

In both cases, we prefer models which have low values of AIC and BIC. Both of them penalise models with more parameters, but the BIC increases the penalty for even larger models.

Definition 7.4. Mallow's C_p statistic is defined as

$$C_p = \frac{SSE_p}{S^2} - [n - 2(p+1)]$$

Remark. We can think of C_p as the sum of the mean squared error plus a penalty. If the model is correct, then $C_p \approx p + 1$. We can actually formalise this approximation.

Theorem 7.5. *Given a linear model which follows the usual assumptions, we have* $\mathrm{E}[C_p] \approx p + 1$.

Proof. Recall that $\frac{SSE_p}{\sigma^2} \sim \chi^2(n-p-1)$ so that $\mathrm{E}\left[\frac{SSE_p}{\sigma^2}\right] = n - p - 1$ and

$$\mathrm{E}\left[\frac{SSE_p}{n-p-1}\right] = \sigma^2 \qquad \mathrm{E}[SSE_p] = (n-p-1)\sigma^2$$

Given that S^2 is an unbiased estimator of σ^2, we then have

$$\mathrm{E}[C_p] \approx \frac{(n-p-1)\sigma^2}{\sigma^2} - [n - 2(p+1)] = p + 1$$

The approximation can actually be bounded with the help of some inequalities and some regularity conditions, but we leave that for a mathematical statistics course. ■

Remark. A suitable strategy is picking a subset of predictors such that C_p is close to $p + 1$.

Aside. Now recall the PRESS statistic in the chapter above. We will not go into a full derivation, but you can trust us when we say that models with small PRESS statistics are preferable.

Chapter 8

Nonlinear regression models

Previously in these notes we dealt with the problem of fitting the regression surface for a model of the form

$$y = \mu + \epsilon$$

where μ is a deterministic term which is additive and linear in its parameters. However, some kid at UWaterloo once said "linear models fail to approximate reality when reality fails to be linear". He's right, so in this chapter we explore different relationships that can be modelled in reality. Most things in this chapter are labelled as examples, as no new theory is introduced.

Example 8.1. A case of a **multiplicative model** is one of the form

$$\mu = \beta_0 x_1^{\beta_1} e^{\beta_2 x_2} \epsilon$$

Clearly, this model is not additive and thus not of the form we are used to. However, we can transform this model to a form where we can use our familiar tools to deal with it. If we take the natural logarithm of both sides, we obtain:

$$\ln \mu = \ln \beta_0 + \beta_1 \ln x_1 + \beta_2 x_2 + \ln \epsilon$$

which is an additive model which is linear in its parameters. We can now use our usual linear estimation methods, provided that the transformed errors $\ln \epsilon$ follow non-correlated Gaussian noise assumptions.

Example 8.2. We may discover that the process we are modelling can be described by a differential equation, which is the sometimes the case in the natural sciences. For example, the exponential growth in a bacterial population can be described by

$$\frac{dP}{dt} = kP$$

for suitable choices for the domain of t. This equation solves for $P = P_0 e^{kt}$, which can later be linearised and modelled linearly.

Example 8.3. The simplest growth model is the **linear trend model**. We can write

$$\mu_t = \alpha + \gamma t$$

where the independent variable is time. There is nothing new here, except that our usual diagnostic procedures have to tune their senses, since dealing with time-series data can be tricky.

Example 8.4. The **exponential trend model** can be used whenever the differential equation shown in Example 8.2 is suitable. We can model the deterministic component of our equation by

$$\mu_t = \beta \exp(\gamma t)$$

which then linearises by taking the natural logarithm to

$$\ln \mu_t = \ln \beta + \gamma t$$

The main caveat in this case is that this will yield unbounded growth. Thus, we must ensure that we restrict the domain in an intelligent manner before we make silly predictions outside our time range. An exponential decay model can be modelled similarly, simply by sticking a negative sign in the exponent's argument.

Example 8.5. A **modified exponential model** is one of the form

$$\mu_t = \alpha - \beta \exp(-\gamma t)$$

which corrects for the fact that the model in Example 8.4 is unbounded. The starting value is $\alpha - \beta$ and it approaches α as $t \to \infty$.

Example 8.6. A **logistic trend model** can be described by

$$\mu_t = \frac{\alpha}{1 + \beta \exp(-\gamma t)} \qquad \alpha > 0,\ \beta > 0,\ \gamma > 0$$

This model arises from a differential equation and is nice enough to estimate. We leave the details to the reader.

We now attempt to develop and formalise some of the heuristics involved in estimating such models.

Definition 8.7. A **nonlinear regression model** is given by

$$y_i = \mu_i + \epsilon_i = \mu(x_i, \beta) + \epsilon_i$$

where $\mu(x_i, \beta)$ is the nonlinear regression component and $\epsilon_i \sim N(0, \sigma^2)$ independently.

Theorem 8.8. *The log-likelihood function for a general non-linear model is*

$$\ln L(\beta, \sigma^2 | Y) = c - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu(x_i, \beta))^2$$

Proof. Trivial. ∎

Remark. Note that the maximum likelihood estimator will minimise the following sum of squares

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

which is usually not possible to perform analytically, so we have to use an algorithm to estimate the parameters. Two approaches we can take are using the Newton-Raphson[1] method or the Gauss-Newton algorithm.

[1] If you are wondering who Raphson is, don't worry, nobody knows. He replicated Newton's work and got his name stuck in the algorithm.