

Advanced Quantitative Methods

PhD Business and Law

& Human Sciences

2009-2010

Dr. Patrick Murphy,
School of Mathematical Sciences.



©These notes are copyrighted property and were created by Dr. Patrick Murphy.

Contents

0	Essentials	iii
0.1	Provisional Course Outline for Semester 1	iii
0.2	Assessment	vi
0.3	Timetable	vi
0.4	Books	viii
0.5	Prerequisites	viii
0.6	Lecture Notes and Textbooks	viii
0.7	Software	ix
1	Revision of Introductory Quantitative Methods	1
2	Simple Linear Regression	2
2.1	Notation and Definitions	2
2.2	Are two variables related?	2
2.3	Simple Linear Regression Model	5
2.4	Least Squares Line	11
2.5	How well does our Model fit?	13
	2.5.1 The Coefficient of Determination	15
	2.5.2 The Coefficient of Variation	17
2.6	Hypothesis Testing and ANOVA	17

2.7	Confidence and Prediction Intervals	19
2.8	Analysis of Residuals	20
3	Multiple Linear Regression	27
3.1	Defining the Multiple Linear Regression Model	27
3.2	Similarities between Multiple Linear Regression Model and Simple Linear Regression Model	28
3.3	How To Programme Multiple Linear Regression in R	30
3.4	Hypothesis Testing and ANOVA	31
3.4.1	Testing Individual Parameters	31
3.4.2	Testing the Overall Model	31
3.4.3	Extra Sum of Squares Principle	32
3.5	Goodness of Fit and Choice of Best Model	34
3.6	Multicollinearity	34
3.7	Factors: Qualitative Predictor Variables	35
3.8	Residual Analysis and Influence Diagnostics	36
4	Reading Data into R	38

Chapter 0

Essentials

0.1 Provisional Course Outline for Semester 1

This course will cover four broad topics

- Using R The course will introduce you to the statistical package R which is freely available and allows students to simply analyse data and perform very powerful statistical modeling on data.
- Applied linear regression The classes in the first semester will focus on the concepts and issues in applied linear regression and analysis of variance. How do we estimate a linear regression line? How do we interpret the results? How do we know what is statistically significant? How do we perform diagnostic analysis on the regression results? Although this part of the course will be primarily theoretical, with limited application in R, the amount of mathematics will be low.

- Multinomial regression II To continue the analysis of multinomial regression models, the session will focus on hypothesis testing and fitting probabilities.
- Ordinal regression There are different techniques if the dependent variable is ordinal and has more than two categories.
- Nonlinear Regression This session on regression will introduce regression models than cannot be fitted to a linear function of the explanatory variables.
- Poisson Regression The module will then turn to regression models for which the dependent variable represents the frequency of some event (i.e. counts).

nt variable is dichotomous.

0.2 Assessment

A project will be assigned at the end of the first 6 weeks of this course which will be due for completion in the Second Semester (see syllabus).

0.3 Timetable

LECTURES:

3.00 to 5.00 Thursdays N302.

0.4 Books

New Cambridge Elementary Statistical Tables

Kutner, Nachtsheim, Neter *Applied Linear Regression Models***

Myers *Classical and Modern Regression with Applications***

Verzani *Using R for Introductory Statistics***

0.5 Prerequisites

Students should be familiar with the content of the course Introduction to Quantitative Methods.

- Principles of Inferential Statistics
- Hypothesis Testing: Principles, Errors, P-Values
- Confidence Intervals
- Correlation

0.6 Lecture Notes and Textbooks

- The lecture notes will only provide an outline of the material covered in the course.
- They are not intended to stand alone but instead should be supplemented by reading appropriate chapters in the text books.

- Extra material in the form of explanations, examples and demonstrations will be given in class and these form an integral part of the course.

0.7 Software

Although the course will focus on the theory and main concepts and issues in applied regression analysis, you will also learn how to use the statistical software package R, which is freely available at <http://www.r-project.org>. You should download and install this at home, so you can get as much hands-on practice as possible.

Chapter 1

Revision of Introductory Quantitative Methods

Chapter 2

Simple Linear Regression

2.1 Notation and Definitions

We define the following Sums of Squares:

$$SS_{XY} = \sum (X - \bar{X})(Y - \bar{Y}) \quad (2.1)$$

$$SS_{XX} = \sum (X - \bar{X})^2 \quad (2.2)$$

$$SS_{TOTAL} = SS_{YY} = \sum (Y - \bar{Y})^2 \quad (2.3)$$

2.2 Are two variables related?

Consider the problem of establishing whether a relationship exists between two variables X and Y . One way to establish if a relationship is present is to draw a scatter plot such as in figures 2.2 and 2.2. Clearly X and Y are related in Figure 2.2 but not related in Figure 2.2.

An alternative way to check for the presence of a linear relationship

Pearson's Sample Correlation Coefficient:

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}} \quad (2.4)$$

- Values of r close to $+1$ indicate a positive linear relationship between X and Y .
- Values close to -1 indicate a negative linear relationship between X and Y .
- Values close to 0 indicate no linear relationship between X and Y . There may however be a non-linear relationship between X and Y .

The sample correlation coefficient tells us information about the sample which we have chosen. If we wish to know something about the relationship between X and Y in the Population then we need to consider the True Population Correlation Coefficient ρ .

Although we don't know ρ we may use r as an estimator of ρ and so we may use r to conduct hypothesis tests and construct confidence intervals for ρ .

For instance we may test the hypotheses that there is a linear relationship between X and Y , namely:

$$H_0 : \quad \rho = 0 \quad (2.5)$$

$$H_A : \quad \rho \neq 0 \quad (2.6)$$

using the test statistic

$$t = r \sqrt{\frac{n-2}{1-r^2}} \quad (2.7)$$

which follows a t -distribution with $n - 2$ degrees of freedom.

QUESTION: Is there a correlation between Shoe Size and Waist Size in Human Beings?

2.3 Simple Linear Regression Model

Recall that Correlation does not imply Causation. However sometimes causation really is present. In this course we will be considering such situations where there is a relationship between two (or more) variables. In this case, changes in one variable CAUSE changes in the other variable and there is a definite causal direction to the process.

In our notation Y will represent our DEPENDENT (Response) variable and X will represent our INDEPENDENT (Predictor) variable. We may model the relationship between the observations (data) that we make of each variable, X_i and Y_i , using the general regression model:

$$Y_i = f(X_i) + \epsilon_i, \quad (2.8)$$

So we are saying that the error terms are random and if we have a good model, then the first part of the model:

$$Y_i = \beta_0 + \beta_1 X_i$$

explains all of the relationship between X and Y with just random fluctuations remaining to be explained by ϵ_i .

We also assume that the X values are NOT RANDOM. They are fixed known values and all of the randomness is in Y .

The Model describes a relationship which holds true for all pairs of observations (X_i, Y_i) in the entire population. The model contains two unknown POPULATION PARAMETERS β_0 and β_1 .

Given a data set consisting of a set of observations

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$$

we wish to find our best estimate for equation 2.9. Or more particularly we wish to find our best estimates for β_0 and β_1 . We may visualise the task involved using the following plots.

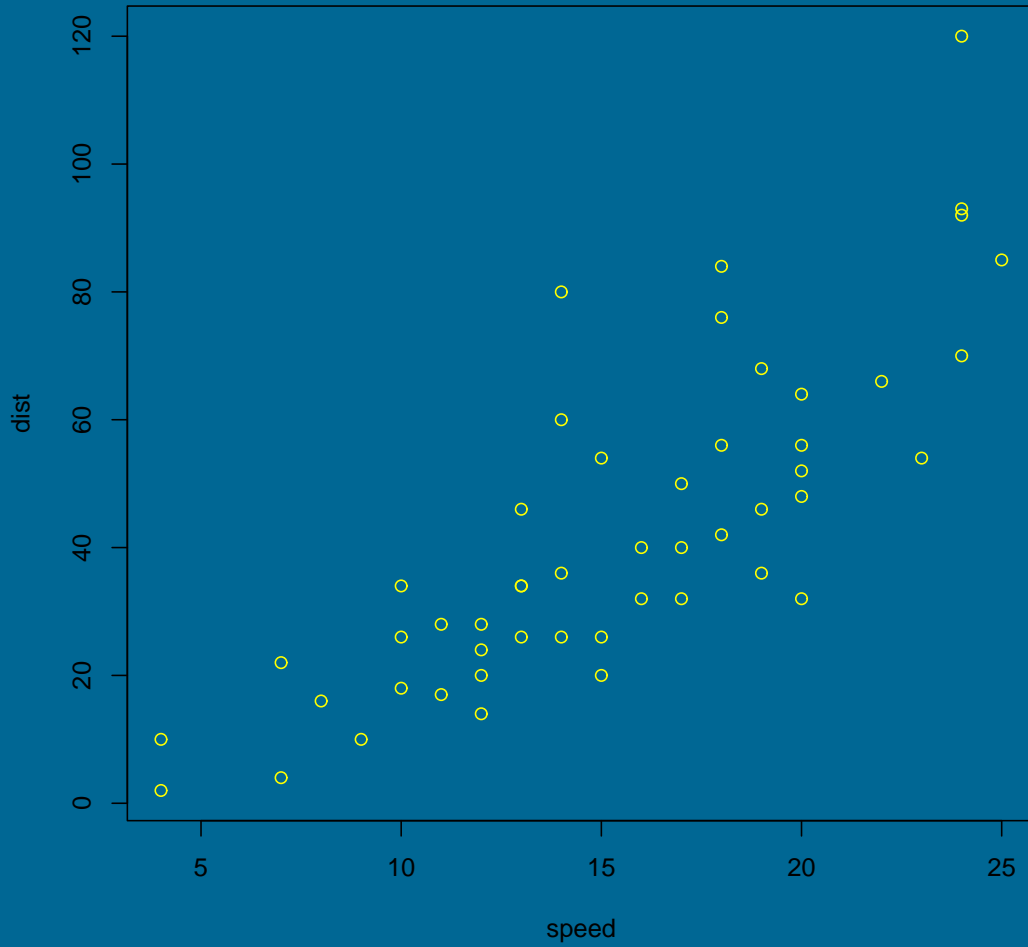


Figure 2.3: Plot of Stopping Distance versus Speed for Cars (1920s)

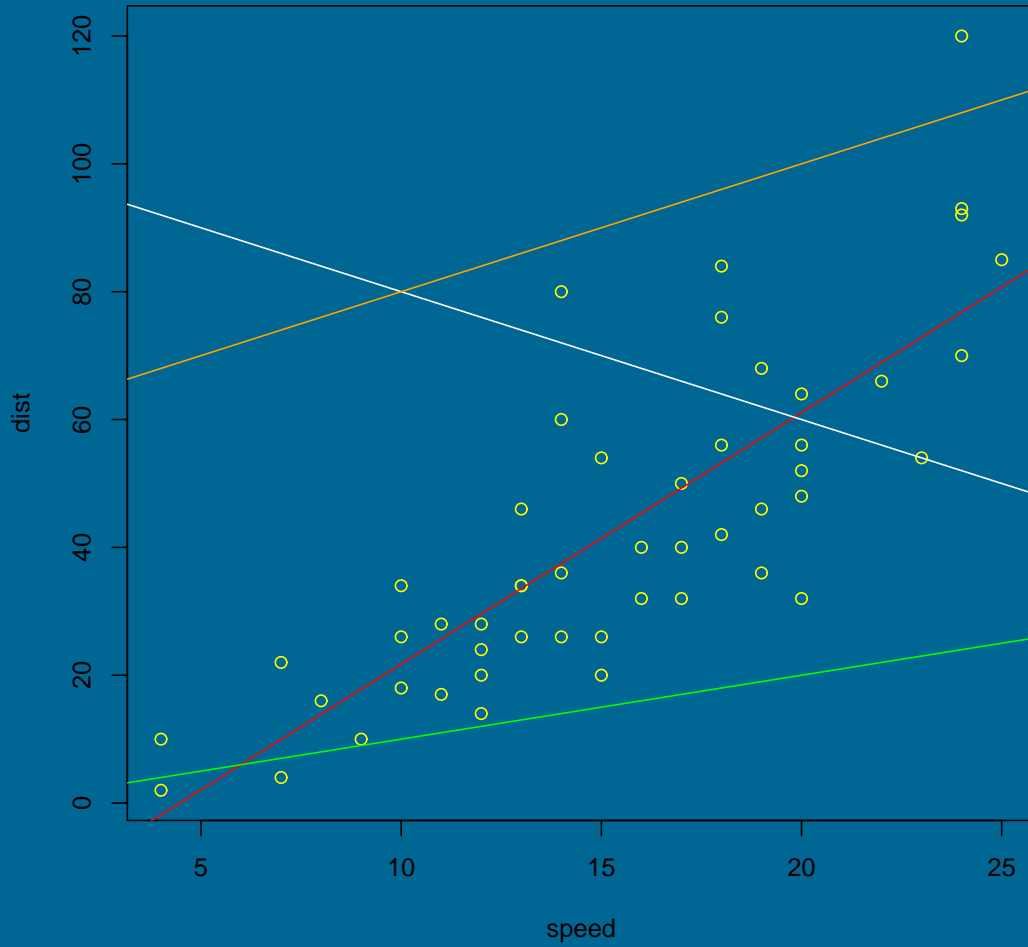


Figure 2.4: Plot of Stopping Distance versus Speed for Cars with Possible Fitted Lines (1920s)

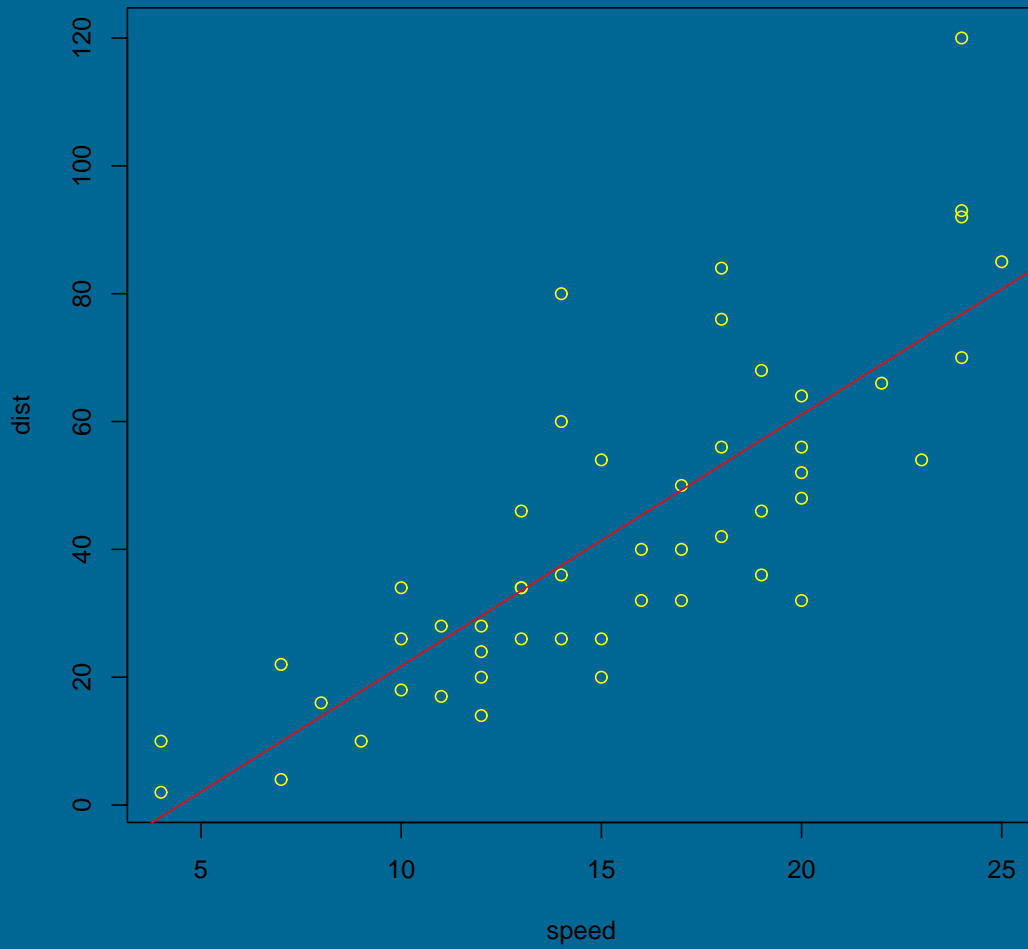


Figure 2.5: Plot of Stopping Distance versus Speed for Cars with Least Squares Line (1920s)

2.4 Least Squares Line

So to recap, we believe that there is a relationship (Equation 2.9) between X and Y which holds true for the entire population of X and Y values:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

If we could measure all of the values of X and Y in the population we could compute this equation exactly. However, usually we only have access to a sample of X and Y pairs. So our task is to estimate Equation 2.9 from that sample data. The method that we choose to use to estimate the "best fitting line" is called Least Squares.

Figure 2.3 displays the Least Squares Line. This line is calculated by choosing estimates for the parameter values β_0 and β_1 which minimise the Error Sum of Squares (sum of the squared distances from each point to the line):

$$SSE = \sum (\epsilon_i)^2 = \sum (Y_i - \beta_0 + \beta_1 X_i)^2. \quad (2.10)$$

SSE is also known as the Residual Sum of Squares or Deviance.

The values that we get for the parameters using this process are

called the Least Squares Estimates:

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}} \quad (2.11)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \quad (2.12)$$

These estimators are unbiased:

$$E(\hat{\beta}_0) = \beta_0 \quad (2.13)$$

$$E(\hat{\beta}_1) = \beta_1. \quad (2.14)$$

These equations tell us that if we compute estimates for the two parameters for all possible samples of X and Y pairs from the population and if we average these sample estimates then the averages will correspond to the true population parameters.

These allow us to compute an estimate a value for Y for a given value of X namely:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (2.15)$$

And the variances of the parameter estimates are given by:

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{SS_{XX}} \right) \quad (2.16)$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SS_{XX}}. \quad (2.17)$$

We also note that s^2 is an unbiased estimator of σ^2 (the variance of the ϵ_i error terms).

$$s^2 = \frac{SSE}{n - 2}. \quad (2.18)$$

2.5 How well does our Model fit?

Now that we have a model to describe the relationship between X and Y we might, after fitting the data to our Least Squares Line, ask ourselves the question: "How good is this Model at explaining the behaviour of the Y values?" In other words is our model a good model? How good is it?

According to our Linear Regression Model most of the variation in Y is caused by its relationship with X , and this knowledge of X should help us in predicting Y .

This is what we do in the regression model we estimate Y by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i. \quad (2.19)$$

Except in the case where all the points lie exactly on a straight line (ie where $r = +1$ or $r = -1$) the model does not explain all the variation in y . The amount that is left unexplained by the model is SSE .

Consider data where there is no relationship between Y and X :

So here the variation in Y is not caused by a relationship between

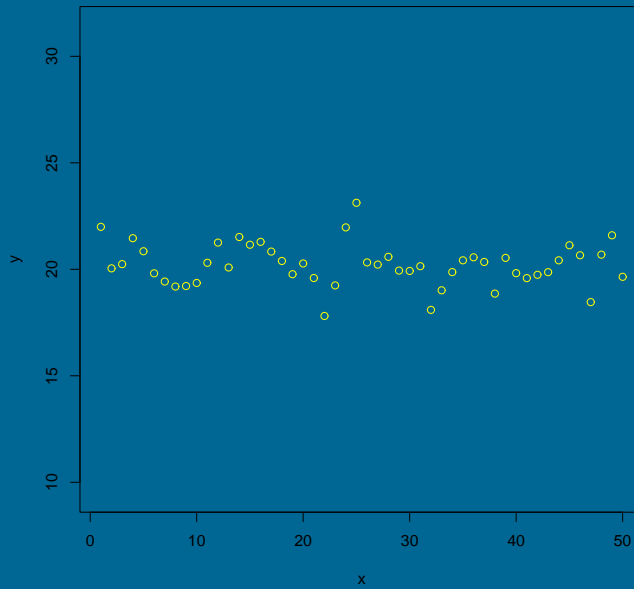


Figure 2.6: Plot of bivariate data in which Y does not depend on X . Here the best estimate for Y is \bar{Y} irrespective of what the value of X is.

X and Y . In this situation the best estimate of Y is the same for all values of X and is given by \bar{Y} .

And the Sum of Squared Deviations of the actual Y 's from this prediction would be $SS_{TOTAL} = SS_{YY} = \sum(Y - \bar{Y})^2$.

So in deciding whether there is a linear relationship between X and Y we could compare the performance of \hat{Y} against \bar{Y} as estimators of Y .

1. If little or none of the variation in Y is explained by the contribution of X then estimating Y using the regression model will be no better than estimating it using \bar{Y} and so SSE will be almost equal to SS_{TOTAL} .

2. If all of the variation in Y is explained by its relationship with X then SSE will be zero.

2.5.1 The Coefficient of Determination

Consider the following quantity, known as the Coefficient of Determination:

$$R^2 = \frac{SS_{TOTAL} - SSE}{SS_{TOTAL}}. \quad (2.20)$$

Clearly this measures the percentage of total variation that can be explained by the simple linear regression model. A value of R^2 equal to 1 indicates that $SSE = 0$ and so the model has no error and is a perfect straight line. If however $R^2 = 0$ then $SSE = SS_{TOTAL}$ and so the model is no good.

This can be seen more clearly when we consider that the Total Variability may be Partitioned as follows:

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \quad (2.21)$$

$$SS_{TOTAL} = SS_{REG} + SSE \quad (2.22)$$

Total Variability in Y =
Variability Explained by Regression Model +
Unexplained Residual Variability.

WARNING: EXTREME CARE MUST BE TAKEN IN INTERPRETING R^2 .

- R^2 depends on the variability in the X values and larger values of SS_{XX} will produce larger values of R^2 . This will happen even when the variance of the residuals in the model as measured by s^2 is large.
- R^2 can also appear artificially large if the slope of the regression line is large.
- When we come to consider Multiple Linear Regressions we must be aware that R^2 will always decrease if a model is over-fitted. This increase in R^2 does not necessarily mean the additional term should be included in the model.
- A version of R^2 called the Adjusted R^2 makes an attempt to compensate for the fact that R^2 will increase with the inclusion of extra terms in the model. The Adjusted R^2 is scaled by the degrees of freedom in the model.
- How large should R^2 be for a good model? It depends on the circumstances. Social Scientists will usually be happy with much smaller values of R^2 than will "Real" Scientists.

QUESTION: Can we demonstrate how increased variability in X may increase R^2 ?

2.5.2 The Coefficient of Variation

Another measure that is sometimes used to measure how good the model is at explaining the variability in Y is the Coefficient of Variation which is the ratio of the estimated standard deviation of the error terms ϵ to the mean value of Y :

$$CV = \frac{100 \times s}{\bar{Y}} \quad (2.23)$$

2.6 Hypothesis Testing and ANOVA

The two measures described in the last section allow us to measure how well the model fits the data in our sample. Consider how we might test whether the model is a good fit for the population of X and Y values not just for the sample. We may examine how well the model fits in the population by performing an Analysis of Variance:

Source	Sum of Squares	df	Mean Squares	F
Regression	SS_{REG}	1	$SS_{REG}/1$	$F = MS_{REG}/MS_{ERROR}$
Error	SSE	$n - 2$	$SSE/(n - 2)$	
Total	SS_{TOTAL}	$n - 1$		

Table 2.1: ANOVA TABLE FOR SIMPLE LINEAR REGRESSION

If we reject the F test in this ANOVA we conclude that the model is a good fit in the population.

Using the results from earlier (c.f. Equations 2.16 and 2.17) we may

also perform hypothesis tests on the parameters in the model.

To test

$$H_0 : \beta_1 = b \quad (2.24)$$

$$H_A : \beta_1 \neq b \quad (2.25)$$

we may use the test statistic

$$t = \frac{\beta_1 - b}{s/\sqrt{SS_{XX}}}, \quad (2.26)$$

which follows a t-distribution with $n - 2$ degrees of freedom.

To test

$$H_0 : \beta_0 = a \quad (2.27)$$

$$H_A : \beta_0 \neq a \quad (2.28)$$

we may use the test statistic

$$t = \frac{\beta_0 - a}{s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{SS_{XX}}}}, \quad (2.29)$$

which also follows a t-distribution with $n - 2$ degrees of freedom.

Performing the two tests above with $a = 0$ and $b = 0$ allow us to determine whether the terms β_0 or β_1 should be present in our model.

If we determine that β_1 is not significantly different from zero then we would be saying that Y does not depend on X (in a linear fashion). Consequently the model is useless for predicting Y in terms of its linear relationship with X . For this reason the hypothesis test:

$$H_0 : \beta_1 = 0 \quad (2.30)$$

$$H_A : \beta_1 \neq 0 \quad (2.31)$$

is known as the Model Utility Test.

2.7 Confidence and Prediction Intervals

Two different confidence intervals are often computed.

A confidence interval for the mean value of Y at a given particular value of X (say X_0) is given by

$$\hat{Y}(X_0) \pm t_{critical} s \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_{XX}}} \quad (2.32)$$

A prediction interval for a future single observation of Y at a given particular value of X (say X_0) is given by

$$\hat{Y}(X_0) \pm t_{critical} s \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{SS_{XX}}} \quad (2.33)$$

2.8 Analysis of Residuals

influence rstandard rstudent dffits dfbeta dfbetas covratio cooks.distance
hatvalues hat

1. Histogram of Residuals ($Y_i - \hat{Y}_i$)

We know that the residuals should follow the normal distribution. So one of the checks that we can do to examine if our model is a good fit is to plot a histogram of the residuals and see if it looks like a normal Histogram.

2. Normal Probability Plot (QQ Plot)

The QQ plot is a special plot that allows us to check if data follows a particular distribution. In our case we are interested in checking if the residuals follow a Normal distribution. So this plot provides us with an additional tool for checking the Normality of the residuals which we can use in conjunction with the Histogram mentioned above.

3. Plot of Residuals versus fitted values \hat{Y}_i

This plot allows us to test several other assumptions of the model. We know that our residuals should all have the same variance σ^2 . Consider the following situations:

- If this assumption holds true then the plot should look like Figure 3.

- If however the variance of the residuals is not constant but rather varies then we say the model contains Heteroscedasticity (Non-constant variance). Figure 3 shows a plot of residuals that are Heteroscedastic.
- Figure 3 shows a lot of residuals which indicates that the model that we have used is insufficient and should have included a curvature term such as X^2 .

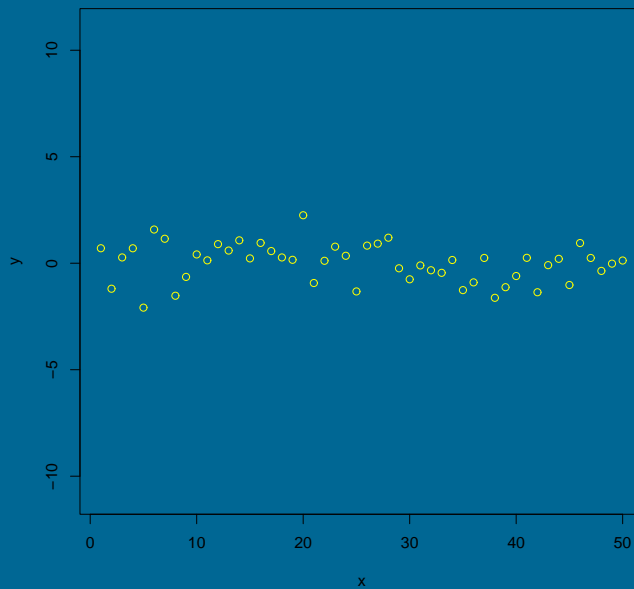


Figure 2.7: Plot of residuals versus fitted values showing well behaved residuals.

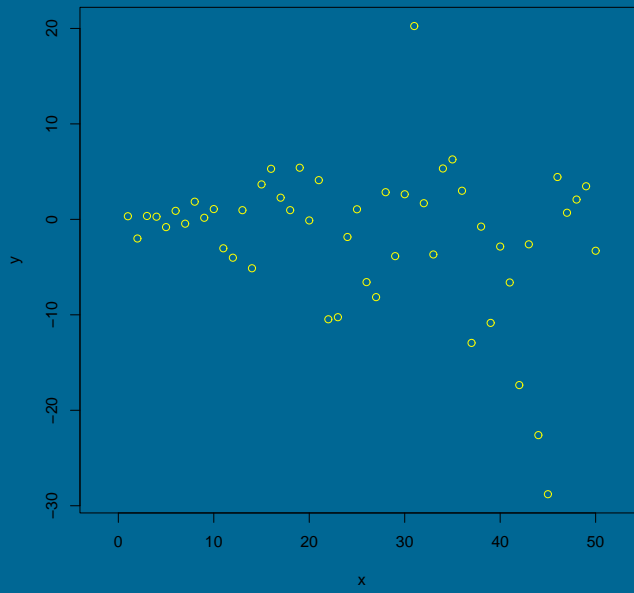


Figure 2.8: Plot of residuals versus fitted values showing Heteroscedasity in residuals.

Figure 2.9: Plot of residuals versus fitted values showing that the model should have included a higher order term to account for curvature such as X^2 .

INSTALL DATA SETS FROM TEXT BOOK

```
install.packages("UsingR")
```

SELECT IRELAND MIRROR OR ANY OTHER

```
library(UsingR)
```

```
galton
```

```
best.times
```

```
babies
```

Fit the Regression Model:

```
regression1=lm(dist speed, data=cars)
```

Regression Coefficients, β_0 and β_1 :

```
coef(regression1)
```

Output from the Regression Model:

```
summary(regression1)
```

ANOVA for Regression Model:

```
anova(regression1)
```

Individual Residuals from the Regression Model:

```
residuals(regression1)
```

Sum of Square for Error (SSE) from Regression Model:

```
deviance(regression1)
```

Histogram of residuals from Regression Model:

```
hist(residuals(regression1))
```

Other Plots of Residuals from Regression Model:

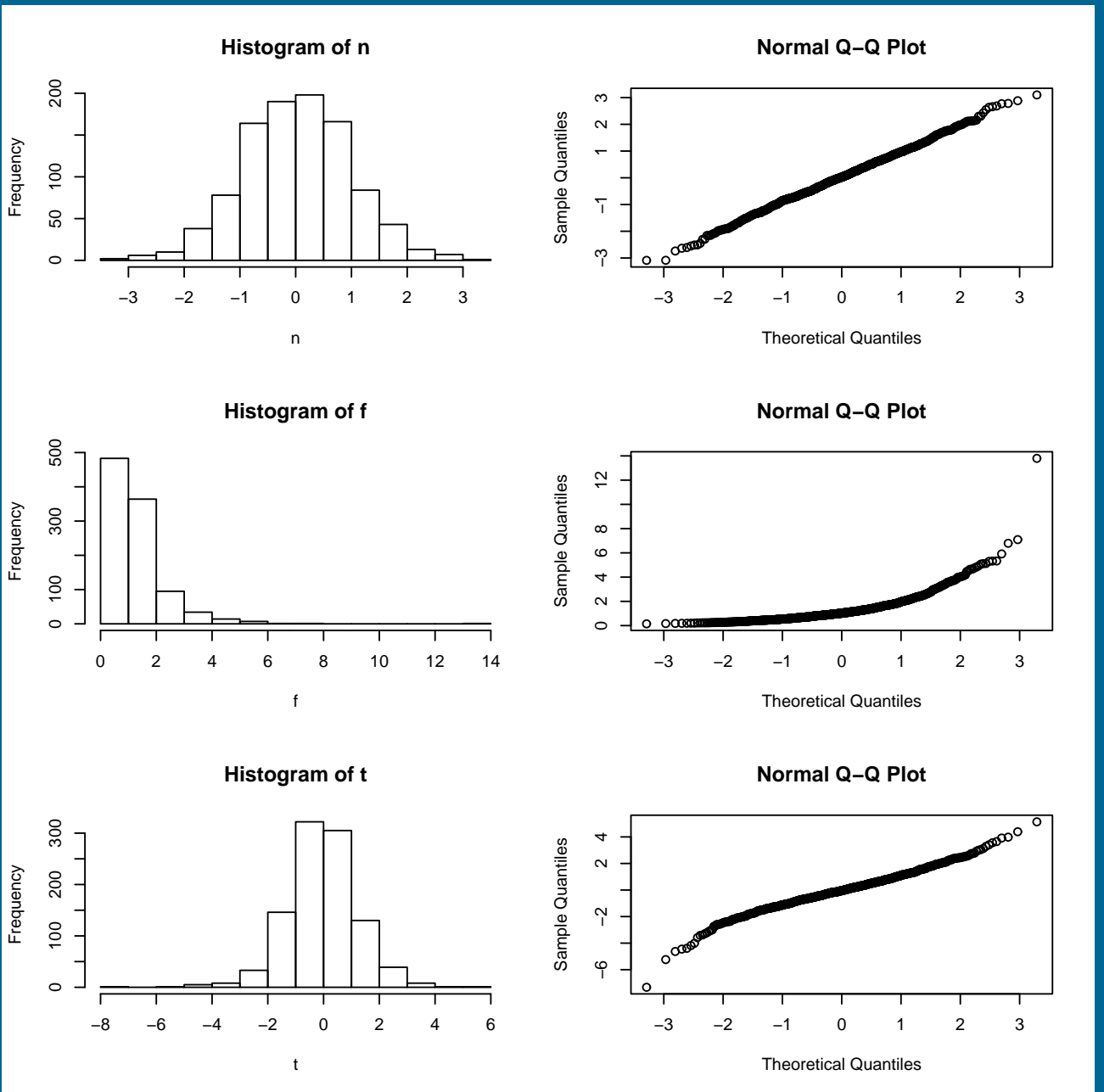


Figure 2.10: Normal Probability Plots for Normal, F and t data

Chapter 3

Multiple Linear Regression

3.1 Defining the Multiple Linear Regression Model

In the previous chapter we considered how to estimate relationships between two variables X and Y given a sample of data:

In this chapter we consider a situation where we have one dependent variable Y but several explanatory variables X_1, X_2, \dots, X_n

All of our models will be linear in the parameters which are contained in the model. That is the parameters will not under-go any transformations while the data may be transformed. The simplest form Multiple Linear Regression Model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon. \quad (3.1)$$

The errors ϵ_i are once again independent, identically distributed normal random variables with mean 0 and variance 1.

Other models we might consider are:

The Polynomial Regression Model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k + \epsilon. \quad (3.2)$$

or

A Regression Model with an Interaction Term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon. \quad (3.3)$$

Two Regression Models with Transformed Variables

$$Y = \beta_0 + \beta_1 \log(X_1) + \beta_2 \log(X_2) + \dots + \beta_k \log(X_k) + \epsilon. \quad (3.4)$$

$$\log(Y) = \beta_0 + \beta_1 \left(\frac{1}{X_1} \right) + \beta_2 \left(\frac{1}{X_2} \right) + \dots + \beta_k \left(\frac{1}{X_k} \right) + \epsilon. \quad (3.5)$$

3.2 Similarities between Multiple Linear Regression Model and Simple Linear Regression Model

The exact same least squares procedure as we used in the Simple Linear Regression Model is used to estimate the parameters in the Multiple Linear Regression Model. That is we are again trying to minimise the Sum of the Squared Error terms.

Just as in the Simple Linear Regression model we can check how

well our model fits to the sample data using the Coefficient of determination:

$$R^2 = \frac{SS_{TOTAL} - SSE}{SS_{TOTAL}}. \quad (3.6)$$

where again we have the following relationship between the different sums of squares:

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \quad (3.7)$$

$$SS_{TOTAL} = SS_{REG} + SSE \quad (3.8)$$

Note that in a multiple linear regression 3.1 with the estimator of Σ^2 (the variance of ϵ) becomes

$$\hat{\sigma}^2 = \frac{SSE}{n - k - 1} \quad (3.9)$$

which simplifies to the familiar equation:

$$\hat{\sigma}^2 = \frac{SSE}{n - 2} \quad (3.10)$$

in the case with one independent variable X_1 .

3.3 How To Programme Multiple Linear Regression in R

Try fitting the best regression model you can to the data sets *women* and *trees*.

1.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

```
regression=lm(Y~ X1+X2, data =PATRICK)
```

2.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

```
regression=lm(Y~ 1+ X1+ I(X1^2), data =PATRICK)
```

3.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

```
regression=lm(Y~ poly(X1,2), data =PATRICK)
```

4.

$$Y = \beta_0 + \beta_1 X_1 X_2$$

```
regression=lm(Y~ X1:X2, data =PATRICK)
```

5.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

```
regression=lm(Y~ X1*X2, data =PATRICK)
```

6.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1 X_2$$

```
regression=lm(Y~ (X1+X2)^2, data =PATRICK)
```

7.

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1 X_2$$

```
regression=lm(log(Y)~ (X1+X2)^2, data =PATRICK)
```

3.4 Hypothesis Testing and ANOVA

3.4.1 Testing Individual Parameters

As in the Simple Linear Regression Model we can perform t -tests to examine whether each individual β_i term should be in the model or not:

$$H_0 : \beta_i = 0 \quad (3.11)$$

$$H_A : \beta_i \neq 0. \quad (3.12)$$

3.4.2 Testing the Overall Model

We may also perform an overall F test to consider how well the model fits. In a model with k parameters β_1 to β_k plus and intercept

β_0 :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon. \quad (3.13)$$

the ANOVA table takes the following form:

Source	Sum of Squares	df	Mean Squares	F
Regression	SS_{REG}	k	SS_{REG}/k	$F = MS_{REG}/MS_{ERROR}$
Error	SSE	$n - k - 1$	$SSE/(n - k - 1)$	
Total	SS_{TOTAL}	$n - 1$		

Table 3.1: ANOVA TABLE FOR MULTIPLE LINEAR REGRESSION

3.4.3 Extra Sum of Squares Principle

In addition to these two hypothesis tests which are familiar from the Simple Linear Regression Model, in the Multiple Linear Regression Model we may also perform an additional test which compares **NESTED** models. Consider the two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon, \quad (3.14)$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \dots + \beta_{k+p} X_{k+p} + \epsilon. \quad (3.15)$$

By comparing the values of SSE for these two models we can construct a hypothesis test to see if the larger model is a **significantly** better fit.

If $SSE(k + p)$ is the Error Sum of Squares for the larger model (3.15) and if $SSE(k)$ is the Error Sum of Squares for the smaller model (3.14) then we call the difference

$$SSE(k) - SSE(k + p) \tag{3.16}$$

the **Extra Sum of Squares**.

Using this Extra Sum of Squares we can construct an F test statistic (3.19) to test that the bigger model is a better fit or more specifically to test

$$H_0 : \beta_{k+1} = \beta_{k+2} = \dots = \beta_{k+p} = 0 \tag{3.17}$$

$$H_A : \text{at least one of } \beta_i \neq 0 \text{ for } i \geq k + 1. \tag{3.18}$$

$$F = \frac{(RSS(k) - RSS(k + p))/p}{RSS(k + p)/(n - k - p - 1)} \tag{3.19}$$

To conduct this test in R we first fit the two models

$$\text{regression1} = \text{lm}(Y \sim X1 + X2 + \dots + X_k)$$

$$\text{regression2} = \text{lm}(Y \sim X1 + X2 + \dots + X_k + \dots + X_{k+p})$$

then we use the command:

anova(regression1, regression2)

to perform the F-test.

3.5 Goodness of Fit and Choice of Best Model

We may use the Coefficient of Determination, the ANOVA F-test and Extra Sum of Squares test to establish which of our models is the "best" model.

3.6 Multicollinearity

Multiple Regressions can perform badly if the predictor variables are not linearly independent. In a sense, if the predictor variables are correlated with each other then they stop acting as independent (and hence good) predictors. If this multicollinearity is present then the variances of the estimates of the coefficients ($\hat{\beta}_i$) are inflated beyond what they would be if this multicollinearity was not present. In other words the parameter estimates are much more uncertain than they should be.

3.7 Factors: Qualitative Predictor Variables

Consider trying to fit a model to predict a person's height at 21 years of age on the basis of their height as a 1 year old. Clearly for such a model the gender of the individual is important so that we would like a model that looks something like:

$$AdultHeight = \beta_0 + \beta_1 HeightAge1 + \beta_2 Gender + \epsilon$$

Or consider modelling the salary of an individual on the basis of the years that they have been employed and the type of job that they have.

$$salary = \beta_0 + \beta_1 YearsEmployed + \beta_2 TypeOfJob + \epsilon$$

In the examples above the two variables "Gender" and "Type of Job" are both Qualitative or Categorical variables.

When we try to run a regression involving these Qualitative variables we must be careful and treat them somewhat differently than Numerical Variables.

In particular when including a categorical variable in the *lm* expression in R we must enclose that variable in the expression "factor()". So if *Y* is our response variable and *X1* is a numerical predictor variable and *X2* is a categorical predictor variable then we use the R command:

```
lm(Y ~ X1 + factor(X2)).
```

Consider:

```
plot(wt ~ wt1, data = babies, pch = smoke, subset = wt1 < 800)
```

```
regression = lm(wt ~ wt1 + factor(smoke), data = babies, subset = wt1 < 800)
```

```
summary(regression)
```

3.8 Residual Analysis and Influence Diagnostics

To consider how well our model fits the data we need to examine how well the residuals from our model satisfy the assumptions that we placed on the error terms ϵ . To test these assumptions we may use the exact same plots and techniques that we used in Simple Linear Regression.

There are also other more sophisticated ways to examine the behaviour of residuals which will be considered later in this course.

These methods include:

Chapter 4

Reading Data into R

We can read data in directly from the command line:

```
x=c(2,4,6,8,10,12,14,16,18,20)
```

```
x[1]
```

```
x[3]
```

We can read bigger data sets in using the scan command:

```
x=scan('C:/OSPAR_TREND_DETECTION/AQM_data.txt',skip=1)
```

```
x= matrix(x,ncol=7,byrow=TRUE)
```

The scan command works with Tab-Space delimited text files. Always save Excel files as Tab-Space delimited text files and then use the scan command to read them into R.