STAT 331 notes: Applied Linear Models

Johnew Zhang

November 29, 2012

Contents

1 Introduction

Regression deals with the functional relationship between a response (or outcome) variable y and one or more explanatory variables (or predictor variables) $x_1, \cdots, x_p$. A general expression for a regression model is

$$y = f(x_1, \cdots, x_p) + \epsilon$$

where

- function $f(x_1, \cdots, x_p)$ represents the deterministic relationship between y and $x_1, \cdots, x_p$
- the extra ϵ term is an error term or called noise. It represents unexplained variation in y due to other factors

Applications

	y	xs
global climate	surface temperature	GHG
finance	stock price index	unemployment rate, CPI, etc.
Economics	Unemployment rate	interest rate

Regression Modelling can be used for

- Identify important factors (or explanatory variables)
- estimation
- prediction

,k. In dumb first statistics class, we saw only a simplest form of the regression model

$$y = \beta_0 + \beta_1 x + \epsilon$$

where we have only one explanatory variable x, and the form of $f(x)$ is assumed to be known as a linear function. For example, $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ is a linear; however, $y = \beta_0 + \beta_1 exp(\beta_2 x)$ is not linear.

In this class, we will extend discussion to p explanatory variable $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$ where $\beta_0, \cdots, \beta_p$ are constants in the linear function, we normally call them regression parameters (or coefficients). Note that β_s are unknown and to be estimated from the data.

2 Review of Simple Linear Regression Model

2.1 The Model

Let y be the response variable, and x be the only explanatory variable. The simple linear regression model is given by

$$y = \beta_0 + \beta_1 x + \epsilon$$

where $\beta_0 + \beta_1 x$ represent the systematic relationship, and ϵ is random error. β_0 and β_1 are unknown regression parameters. (Note y is always random, and x is non-random)

Suppose we observe n pairs of value $\{(y_i, x_i), i = 1, \cdots, n\}$ on y and x from a random sample of subjects. Then for the ith observation, we have $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Formally, we make a number of assumptions about $\epsilon, \cdots, \epsilon_n$. Gauss-Markov Assumption (Conditional on x_i)

1. $E[\epsilon_i] = 0 \implies E(y_i) = \beta_0 + \beta_1 x_i$
2. $\epsilon_1, \cdots, \epsilon_n$ are statistically independent $\implies y_1, \cdots, y_n$ are independent
3. $Var(\epsilon_i) = \sigma^2 \implies Var(y_i) = \sigma^2$
4. ϵ_i is normally distributed for $i = 1, \cdots, n, \epsilon \sim N(0, \sigma^2)$

These four assumptions are often summarized as saying that $\epsilon_1, \cdots, \epsilon_n$ are independent and identically (iid) $N(0, \sigma^2)$.

In particular, assumption (i) is needed to ensure that a linear relationship between y and x is appropriate. Assumptions (ii) - (iv) also translate to assumption about $y_1, \cdots, y_n$. Assumption (ii) implies that $y_1, \cdots, y_n$ are independent. Assumption (ii) implies $Var(y_i) = \sigma^2$ (constant over x_i). Assumption (iv) implies that y_i is normally distributed, . Equivalently, we can summarize that $y_1, \cdots, y_n$ are independent normal such that

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

The two unknown regression parameters β_0 and β_1:

- β_0 is the intercept
- β_1 is the slope and is of primary interest.
  1. $\beta_1 = E[y|x = a + 1) - E[y|x = a]$
  2. If $\beta_1 = 0$, $E[y|x] = \beta_0$

2.2 Least Square Estimation (LSE)

Suppose we let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the chosen estimators for β_0 and β_1, respectively, and the fitted value for y_i from the regression line is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Then the least squares criterian chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to make the residuals

$$r_i = y_i - \hat{y}_i$$

"small". Specifically, LSE of β_0 and β_1 are chosen to minimize the sum of squared residuals:

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

The LSE of β_0 and β_1 are

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

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

where $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$ and $S_{xx} = \sum(x_i - \bar{x})^2$

A note on notation

- In this course, we used some times y_i to denote the random variable for response from the ith subject of a sample; and some times for the value (number) actually observed.
- Similarly, $\hat{\beta}_0$ and $\hat{\beta}_1$ will be used both for the estimators which are random variables if y_i's are random; and for particular estimates (i.e. 0.77) calculated from a particular data.

2.3 The Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$

We have following properties of LSE

1. $E[\hat{\beta}_0] = \beta_0$, $E[\hat{\beta}_1] = \beta_1$
2. The theoretical variance of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$Var(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2})$$

$$Var(\hat{\beta}_1) = \sigma^2 \frac{1}{\sum(x_i - \bar{x})^2}$$

3.

$$cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \bar{x}}{\sum(x_i - \bar{x})^2}$$

The proof of the results related to $\hat{\beta}_1$

We write

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \sum \frac{(x_i - \bar{x})}{S_{xx}} y_i = \sum c_i y_i$$

Hence

$$E[\hat{\beta}_1] = \sum c_i E[y_i]$$

($y_1, \cdots, y_n$ are independent)

$$E[\hat{\beta}_1] = \sum c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum c_i + \beta_1 \sum c_i x_i = \beta_1$$

Similarly,

$$\begin{aligned} Var(\hat{\beta}_1) &= \sum c_i^2 Var(y_i) \\ &= \sum \frac{(x_i - \bar{x})^2}{S_{xx}^2} \sigma^2 \\ &= \frac{\sigma^2}{S_{xx}} \end{aligned}$$

Result:

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

Consequence of LS fitting

1. $\sum r_i = 0$ (r means residual)
2. $\sum r_i x_i = 0$
3. $\sum r_i \hat{y}_i = 0$
4. The point $(\bar{x}, \bar{y})$ is always on the fitted regression line.

2.4 The Estimation of σ^2

Note that we can rewrite the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ as

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

to emphasize the analogy with the residuals

$$r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

We could say that r_i (which can be calculated) estimate the unobservable ϵ_i. The basic idea is then to use sample variance of $r_1, \cdots, r_n$ to estimate the unknown $Var(\epsilon_i) = \sigma^2$. The sample variance of $r_1, \cdots, r_n$

$$\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})^2$$

this is actually not an unbiased.

$$E[\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})^2] \neq \sigma^2$$

The unbiased estimator of σ^2 is defined as

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(r_i - \bar{r})^2$$

(Homework: proof $E[s^2] = \sigma^2$)
Hint:

1. $\sum_{i=1}^{n} r_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$
2. $E[S_{yy}] = (n-1)\sigma^2 + \beta_1^2 S_{xx}$
3. $E[\frac{S_{xy}^2}{S_{xx}}] = \beta_1^2 S_{xx} + \sigma^2$

2.5 Confidence Interval and Hypothesis Testing

Recall that

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

so

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

By definition,

$$P(-1.96 < \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} < 1.96) = 0.95$$

$$P(\hat{\beta}_1 - 1.96\frac{\sigma}{\sqrt{S_{xx}}} < \beta_1 < \hat{\beta}_1 - 1.96\frac{\sigma}{\sqrt{S_{xx}}}) = 0.95$$

95% CI for β_1 is $(\hat{\beta}_1 \pm 1.96\frac{\sigma}{\sqrt{S_{xx}}})$ when σ^2 is known.

In most practice, σ^2 is unknown, when we replace σ^2 by s^2, then the unknown standard deviation of $\hat{\beta}_1$ is replaced by standard error

$$SE(\hat{\beta}_1) = \sqrt{\frac{S^2}{S_{xx}}}$$

where $S^2 = \frac{1}{n-2}\sum r_i^2$

The standardized $\hat{\beta}_1$ random variable becomes

$$\frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim t_{n-2}$$

which is no longer standard normal but has a t-distribution with $n-2$ degree of freedom

A $100(1-\alpha)\%$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} SE(\hat{\beta}_1)$$

Hypothesis tests are derived and computed in the similar way. To test

$$H_0 : \beta_1 = \beta_1^* \text{ v.s. } H_a : \beta_1 \neq \beta_1^*$$

We use the t-statistic

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{SE[\hat{\beta}_1]}$$

which as a t_{n-2} distribution when H_0 is true.

2.5.1 The t-test Statistic

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

where $H_0 : \beta_1 = \beta_1^*$

Formally, if

$$|t| = |\frac{\hat{\beta}_1 - \beta^*}{SE(\hat{\beta}_1)}| > t_{n-2,\alpha/2}$$

There is evidence to reject $H_0 : \beta_1 = \beta_1^*$ at significant level of α. Otherwise, we can not reject H_0.

2.6 Prediction for Future Values

The fitted value

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

refers to an x which is part of the sample data

1. Predict a single future value at a given $x = x_p$. The future value is given as

   $$y_p = \beta_0 + \beta_1 x_p + \epsilon_p$$

   where ϵ_p is the future error.

Naturally, we replace ϵ_p by its expectation and use

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

to predict y_p.

Some properties of $\hat{y}_p$.

(a) $E[y_p - \hat{y}_p] = 0$ is an unbiased prediction.

(b) $Var(y_p - \hat{y}_p) = [1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}]\sigma^2$
$y_p - \hat{y}_p = \beta_0 + \beta_1 x_p + \epsilon_p - \hat{\beta}_0 - \hat{\beta}_1 x_p$
Note that ϵ_p is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$ since it is a future error that is unrelated to the data that $\hat{\beta}_0$ and $\hat{\beta}_1$ are calculated.

$$Var(y_p - \hat{y}_p) = Var(\epsilon_p) + Var(\hat{\beta}_0 + \hat{\beta}_1 x_p)$$

(c) It can be shown that

$$\frac{y_p - \hat{y}_p}{SE(y_p - \hat{y}_p} \sim t_{n-2}$$

where $SE(y - \hat{y}_p) = [1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}]s^2$

2. Predict the mean of future response values at a given $x = x_p$. We will still use

$$\hat{\mu}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

as the predicted future mean

$$\mu = \beta_0 + \beta_1 x_p$$

The variance of the prediction error

$$Var(\mu_p - \hat{\mu})$$

is smaller than the variance of prediction error of y_p.

2.7 Analysis of Variance (ANOVA)

Testing $H_0 : \beta_1 = 0$.

The total variation among the y_i's is measured by

$$SST = \sum (y_i - \bar{y})^2$$

If there is no variation (all y_i's are same), the $SST = 0$. The bigger the SST, the more variation. If we rewrite SST as

$$\begin{aligned} SST &= \sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2 + 2\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\ &= SSE + SSR + 0 \end{aligned}$$

where

- SSE refers to the sum of squares of residual. It measures the variability of y_i's that is unexplained by the regression model
- SSR refers to the sum of squares of regression . It measures the variability of response that is accounted for by the regression model

If $H_0 : \beta_1 = 0$ is true, SSR should be relatively "small" compare to SSE. Our decision is to reject H_0 if the ratio of SSR and SSE is large.

2.7.1 Some distribution result (when H_0 is true)

- $$\frac{SST}{\sigma^2} \sim \chi^2(n-1)$$

  To show this, recall that $y_1, \cdots, y_n$ are independent $\sim N(\beta_0, \sigma^2)$ then

  $$\sum(\frac{y_i - \beta_0}{\sigma})^2 \sim \chi^2(n)$$

  $$(\frac{\bar{y} - \beta_0}{\sigma/\sqrt{n}})^2 \sim \chi^2(1)$$

  By rearrangement of SST

  $$\begin{aligned} SST &= \sum_{i=1}^{n}(y_i - \beta_0 + \beta_0 - \bar{y})^2 \\ &= \sum_{i=1}^{n}(y_i - \beta_0)^2 - \sum_{i=1}^{n}(\beta - \bar{y})^2 \\ &= \sum_{i=1}^{n}(y_i - \beta_0)^2 - n(\bar{y} - \beta_0)^2 \\ \frac{SST}{\sigma^2} &= \sum_{i=1}^{n}\frac{(y_i - \beta_0)^2}{\sigma^2} - \frac{n(\bar{y} - \beta_0)^2}{\sigma^2} \\ &\sim \chi^2(n) - \chi^2(1) \sim \chi^2(n-1) \end{aligned}$$

Form Cockran's Theorem:

$\frac{SST}{\sigma^2}$ is independent of $\frac{n(\bar{y}-\beta_0)^2}{\sigma^2}$ and $\frac{SST}{\sigma^2} \sim \chi^2(n-1)$

- $$\frac{SSR}{\sigma^2} \sim \chi^2(1)$$

$$\begin{aligned} SSR &= \sum(\hat{y}_i - \bar{y})^2 \\ &= \sum(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \sum(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \hat{\beta}_1^2 \sum(x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx} \end{aligned}$$

Recall

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

$$(\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}^2 \sim \chi^2(1)$$

Under

$$H_0 : \frac{\hat{\beta}_1^2}{\sigma^2/S_{xx}} = \frac{\hat{\beta}_1^2 S_{xx}}{\sigma^2} \sim \chi^2(1)$$

- $$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$$

$$\begin{aligned} \frac{SST}{\sigma^2} &= \frac{SSE}{\sigma^2} + \frac{SSR}{\sigma^2} \\ \chi^2(n-1) &= \chi^2(n-2) + \chi^2(1) \end{aligned}$$

Based on theses results, we derive F-statistic

$$F = \frac{(SSR/\sigma^2)/1}{(SSE/\sigma^2)/n-2} \sim F(1, n-2)$$

It can be used for testing $H_0 : \beta_1 = 0$, we reject H_0 at α-level if

$$F > F_\alpha(1, n-2)$$

Recall

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{S^2/S_{xx}}}$$

.

$$t^2 = \frac{\hat{\beta}_1^2 S_{xx}}{S^2}$$

and

$$F = \frac{\hat{\beta}_1^2 S_{xx}}{S^2}$$

also $t_{n-2}^2 = F(1, n-2)$

The t-test and F-test for $H_0 : \beta_1 = 0$ are equivalent for SLR.

2.7.2 Terminology of ANOVA

Sum of squares	Source of variation	degree of freedom	mean squares	F
SSR	Regression	1	$MSR = SSR/1$	$F = MSR/MSE$
SSE	residual	n-2	$MSE = SSE/(n-2)$	$F = MSR/MSE$
SST	total	n -1		

Coefficient of Determination:

$$R^2 = \frac{SSR}{SST}, 0 \leq R \leq 1$$

It is a measure of goodness-of-fit of the regression model to the data. In the case of SLR,

$$R^2 = \frac{SSR}{SST} = \frac{\hat{\beta}_1 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2$$

where r^2 is the sample correlation coefficient. R^2 is applicable to multiple regression, but r^2 is not.

3 Review of Random Vectors and Matrix Algebra

3.1 Definition

$$y = (y_1, \cdots, y_n)'$$

$$E(y) = \begin{pmatrix} E(g_1) \\ \vdots \\ E(g_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \mu$$

$$V(y) = (\sigma_{ij})_{n\times n} = \begin{pmatrix} Var(y_1) & Cov(y_1, y_2) & \cdots & Cov(y_1, y_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(y_n, y_1) & \cdots & \cdots & Var(y_n) \end{pmatrix}_{n\times n}$$

$$V(y) = E((y - E(y))(y - E(y))') = E((y_i - \mu_i)^2)_{n\times n}$$

If $y_1, \cdots, y_n$ are independent and identical distributed $Var(y) = \sigma^2 I$

3.2 Basic Properties

$A = (a_{ij})_{m\times n}$, $b = (b_1, \cdots, b_m)'$, $c = (c_1, \cdots, c_n)'$

1. $E(AY + b) = AE(Y) + b$
2. $Var(Y + c) = Var(Y)$
3. $V(AY) = AV(Y)A'$
4. $V(AY + b) = AV(Y)A'$

3.3 Differentiating Over Linear and Quadratic Forms

1. $f(y) = f(y_1, \cdots, y_n)$
$$\frac{d}{dy}f = (\frac{d}{dy_1}f, \cdots, \frac{d}{dy_n}f)'$$
2. $f = c'y = \sum_{i=1}^n c_i y_i$
$$\frac{d}{dy}f = C$$
3. $f = y'Ay = \sum_i \sum_j a_{ij} y_i y_j = \sum_i a_{ii} y_i^2 + 2\sum\sum_{i<j} a_{ij} y_i y_j$
$$\frac{d}{dy}f = 2Ay$$
For example, $\frac{d}{dy_1}f = 2a_{11}y_1 + 2\sum_{i<j} a_{1j}y_j = 2\sum_{j=1}^n a_{1j}y_j$

3.4 Some Useful Results on Matrix

1. Trace:
$$Tr(A_{m\times m}) = \sum_{i=1}^m a_{ii}$$
$$Tr(B_{m\times n}C_{n\times m}) = Tr(C_{n\times m}B_{m\times n})$$
2. Rank of a matrix
$$rank(A_{n\times n}) = \# \text{ of linearly independent columns of } A_{m\times n}$$
3. vectors $y_1, \cdots, y_m$ are linearly independent off $c_1 y_1 + \cdots + c_m y_m = 0$ off $c_1 = c_2 = \cdots = c_m = 0$
4. Orthogonal Vectors & Matrices: two vectors are orthogonal

(a) two vectors are orthogonal: $y'x = 0$

(b) $A_{m\times m}$ is orthogonal if

$$A'A = AA' = I = \Longrightarrow A' = A^{-1}$$

(c) Eigenvector & Eigenvalues: A vector v_i is called an eigenvector of $A_{m\times m}$ if $\exists \lambda_i$ such that

$$Av_i = \lambda_i v_i, i = 1, 2, \cdots, k$$

where λ_i is the eigenvalue.

(d) Decomposition of a Symmetric Matrix:

$$A' = A$$

For a systematic matrix $A_{m\times m}$, $\lambda_1, \cdots, \lambda_m$ are real and exists an orthogonal matrix P such that

$$A = P\Lambda P'$$

where $\Lambda = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \lambda_m \end{pmatrix}$ is diagonal matrix with eigenvalues on diagonal, $p = [v_1, \cdots, v_m]$ is a matrix with eigenvectors on columns.

(e) Idempotent Matrix: $A_{m\times m}$ is idempotent if $A^2 = A$. Results:

i. If $A_{m\times m}$ is idempotent, then all eigenvalues are either 0 or 1

Proof.

$$Av_i = \lambda_i v_i = A^2 v_i = \lambda_i (Av_i) = \lambda_i^2 v_i$$

Hence $\lambda_i = \lambda_i^2$. This implies $\lambda_i = 0$ or 1. □

ii. If $A_{m\times m}$ is idempotent, exist an orthogonal matrix P such that

$$A = P\Lambda P'$$

where $\Lambda = diagonal(1, \cdots, 1, 0, \cdots, 0)_{m\times m}$

iii. $Tr(A) = rank(A) = Tr(\Lambda) =$ # of 1 in the diagonal

4 Multiple Linear Regression

4.1 Multivariate Normal Distribution

The random vector $y = (y_1, \cdots, y_n)'$ follow a multivariate normal distribution with a joint p.d.f .

$$f(y) = [\frac{1}{2\pi}]^{\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(y-\mu)' \sum (y-\mu)\}$$

where $\mu = E(y)_{n\times 1} = (E(y_1), \cdots, E(y_n))'$ and $\Sigma = Var(y)_{n\times n} = (\sigma_{ij})_{n\times n}$. We can write

$$y \sim MVN(\mu, \Sigma)$$

1. Margin Normality: If $y \sim MVN(\mu, \Sigma)$, then $y_i \sim N(\mu_i, \sigma_{ii})$ where σ_{ii} is the (i,i)th element of Σ.

2. $y_1, \cdots, y_n$ are independent if and only if Σ is diagonal. (In general, if $y_1, \cdots, y_n$ are independent , then the $Cov(y_i, y_j) = 0 (i \neq j)$. However this does not implies from the other way around.)

3. If $y \sim MVN(\mu, \Sigma)$, then let $z = Ay$, $z \sim MVN(A\mu, A\Sigma A')$

4. If $\sim MVN(\mu, \Sigma)$, $y_1 = AU, y_2 = BU$ Then y_1 and y_2 are independent if and only if $Cov(y_1, y_2) = 0$, $Cov(AU, BU) = 0$, $(AV(U)B' = 0$, $A\Sigma B' = 0$.

5. $y_1, \cdots, y_n$ are iid $N(\mu, \sigma^2 I)$,
$$y \sim MVN(\mu, \sigma^2 I)$$

6. If $y \sim MVN(0, \sigma^2 I)$,
$$y'y/\sigma^2 \sim \chi(n)$$

4.2 The Regression Model

Suppose we are interested in the relationship between a type of air pollutant and lung function. $y : FEVI$; x_1 : a type air pollutant, x_2 : age, x_3 : gender.

The general model is in the form:

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

where $x_{i1}, \cdots, x_{ip}$ are p explanatory variables, and $\beta_1, \cdots, \beta_p$ are regression coefficients associated with these explanatory variables respectively, $i = 1, \cdots, n$.

Assumption:

1. $E[\epsilon] = 0 \implies E[y_i] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$

2. $V(\epsilon) = \sigma^2 \implies V(y_i) = \sigma^2$

3. $\epsilon_1, \cdots, \epsilon_n$ are independent $\implies y_1, \cdots, y_n$ are independent.

4. A stronger assumption
$$\epsilon_i \sim N(0, \sigma^2) \implies y_i \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2)$$

Regression coefficients: $\beta_1, \cdots, \beta_p$. β_j: the average amount of increase (or decrease) in response when the nth covariate x_j increase (or decreases) by 1 unit while holding all other covariate fixed.

$$H_0 : \beta_j = 0 \implies x_j$$

is not (linearly) related to y, given all the other explanatory variables in the model.

In matrix form:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$y_{n\times 1} = X_{n\times(p+1)}\beta_{p+1} + \epsilon_{n\times 1}$$

where

$$\epsilon \sim MVN(0, \sigma^2 I)$$

and

$$y \sim MVN(x\beta, \sigma^2 I)$$

4.3 LSE of β

Lease squares chose $\hat{\beta}$ to make the $n \times 1$ vector $\hat{y} = x\hat{\beta}$ “close” to y (or to make the residual vector $r = y - \hat{y}$ “small”). Specifically, we want to minimize

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = (y - x\beta)'(y - x\beta) \\ &= y'y - y'x\beta - \beta'x'y + \beta'x'x\beta \end{aligned}$$

Set $\frac{\partial}{\partial\beta} S(\beta) = 0$ to get

$$\frac{\partial}{\partial\beta} S(\beta) = -2x'y + 2x'x\beta = 0$$

$$\hat{\beta} = (x'x)^{-1}x'y$$

(we require $x'x$ to be full rank)

Properties of LSE $\hat{\beta}$

1. $\hat{\beta}$ is unbiased
$$E(\hat{\beta}) = E[(x'x)^{-1}x'y] = (x'x)^{-1}x'E[y] = \beta$$
2. $Var(\hat{\beta}) = \sigma^2(x'x)^{-1}$
$$V((x'x)^{-1}x'y) = (x'x)^{-1}x'V(y)x(x'x)^{-1} = \sigma^2(x'x)^{-1}x'x(x'x)^{-1} = \sigma^2(x'x)^{-1}$$

Some useful results

1. Fitted values
$$\hat{y} = x\hat{\beta} = x(x'x)^{-1}x'y$$
Let $H = x(x'x)^{-1}x' \leftarrow$ “hat” matrix
$$\hat{y} = Hy$$
where H is idempotent and symmetric $\implies$ H is a projection matrix which projects y to $R(x)$ (a $(p+1)$ diminutional subspace spanned by linear combination of $p+1$ columns of x)

2. Residuals
$$r = y - \hat{y} = y - x(x'x)^{-1}x'y = (I - x(x'x)^{-1}x')y = (I - H)y$$
(homework: $I - H$ is idempotent)

   - $\sum r_i = 0$, $\sum_{i=1}^{n} r_i x_{i1} = 0$, $\cdots$, $\sum_{i=1}^{n} r_i x_{ip} = 0$ $\sum r_i \hat{y}_i = \hat{y}'r = 0$
   $$x'r = x'(I - x(x'x)^{-1}x')y = x'y - x'y = 0$$
   - $E(r) = 0$
   $$E(r) = E((I - x(x'x)^{-1}x')y) = x\beta - x\beta = 0$$
   - $v(r) = \sigma^2(I - x(x'x)^{-1}x')$
   $$v(r) = v((I - x(x'x)^{-1}x')y) = (I - H)\sigma^2 I(I - H)' = \sigma^2(I - H)$$

4.4 An estimation of σ^2

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} r_i^2$$

is an unbiased estimator of σ^2.

Proof.

$$\begin{aligned}
E(\sum_{i=1}^{n} r_i^2) &= E(r'r) = E(tr(r'r)) = E(tr(rr')) = tr(E(rr')) = tr(V(r)) = tr(v((I-H)y)) \\
&= tr(\sigma^2(I-H)) = \sigma^2(n - tr(x(x'x)^{-1}x')) = \sigma^2(n - tr(x'x(x'x)^{-1}) \\
&= \sigma^2(n - tr(I_{(p+1)\times(p+1)}) = \sigma^2(n-p-1)
\end{aligned}$$

□

4.5 Sampling Distribution of $\hat{\beta}, \hat{\sigma}^2$ under Normality

We assume

$$Y \sim MVN(x\beta, \sigma^2 I)$$

Results:

1. $\hat{\beta} \sim MVN(\beta, \sigma^2(x'x)^{-1})$, $\hat{\beta} = (x'x)^{-1}x'y$
2. $\hat{\beta}$ and $\hat{\sigma}^2$ are independent

   $\hat{\sigma}^2 = \frac{1}{n-p-1}r'r$, to show r and $\hat{\beta}$ are independent.
3. $$\begin{aligned} cov(r, \hat{\beta}) &= cov((I - x(x'x)^{-1}x')y, (x'x)^{-1}x'y) = (I - x(x'x)^{-1}x')\sigma^2 Ix(x'x)^{-1} \\ &= (I - x(x'x)^{-1}x')\sigma^2 Ix(x'x)^{-1} \\ &= \sigma^2(x(x'x)^{-1} - x(x'x)^{-1} = 0 \end{aligned}$$
4. $(n-p-1)\hat{\sigma}^2\sigma^2 \sim \chi^2(n-p-1)$

Proof. Note that we can rewrite

$$(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} = (n-p-1)\frac{\frac{1}{n-p-1}\sum r_i^2}{\sigma^2} = \frac{\sum r_i^2}{\sigma^2} = \frac{r'r}{\sigma^2} = (\frac{r}{\sigma})'(\frac{r}{\sigma}) = r^{*'}r^*$$

Recall $Y \sim MVN(x\beta, \sigma^2 I)$, then $r^* = \frac{(I-H)Y}{\sigma} \sim MVN(0, I-H)$. Since $I-H$ is idempotent, then there is an orthogonal matrix P such that

$$I - H = P\Lambda P'$$

where $\Lambda = \begin{pmatrix} 1 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & 1 & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}$ and number of 1's $= n-p-1$ and $tr(I-H) = tr(n)$

Now, if we define a new r.v.

$$z = p'r^*$$

then

$$z \sim MVN(0, P'(I-H)P')$$

$$z \sim MVN(0, \Lambda)$$

$z = (z_1, \cdots, z_n)'$, the first $n-p-1$ z_i's has $N(0,1)$, the rest are 0's.

Therefore,

$$(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} = r^{*'}r^*$$

$(z = p'r^*, z = r^*)$ so

$$= (pz)'(pz) = z'p'pz = z'z = \sum_{i=1}^{n-p-1} z_i^2 \sim \chi^2(n-p-1)$$

From result 1,

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

when σ^2 is unknown, we use

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum r_i^2$$

ti estimate σ^2

From 1 and 2, we know that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent, and

$$(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$$

(Note if $x \sim N(0,1), y \sim \chi^2(q)$, then $\frac{x}{\sqrt{y/q}} = t_q$)

Then

$$\frac{\frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 v_{ii}}}}{\sqrt{\frac{(n-p-1)\hat{\sigma}^2/\sigma^2}{n-p-1}}} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{v_{ii}\sigma^2}} \sim t_{n-p-1}$$

this also implies that we use standard error $se(\hat{\beta}_i) = \sqrt{v_{ii}\hat{\sigma}^2}$ to estimate the standard deviation $\sqrt{v_{ii}\sigma^2}$. The quantity can be used to construct CI and test hypothesis $H_0 L\beta_i = \beta_i^*$. □

4.6 Prediction

Suppose we interested in predicting y for given values of the explanatory variables $x_1, \cdots, x_p$. For example, our multiple regression model:

$$y(\text{FEVI}) = \beta_0 + \beta_1 x_1(\text{level of certain air pollutant}) + \beta_2 x_2(\text{age}) + \beta_3 x_3(\text{weight}) + \epsilon$$

We want to predict FEVI for a new cease with an arbitrary vector of explanatory variable values a_p (e.g. $a_p = (1, 10, 52, 170)'$)

Be cautious when extrapolating outside the ranges of the explanatory variables in the fitting data.

$$y_P = \beta_0 + \beta_1 \times 10 + \beta_2 \times 52 + \beta_3 \times 170 + \epsilon_p$$

We can estimate y_p by using $\hat{\beta}$ (LSE) to replace β, and set $\epsilon_p = 0$.

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 52 + \hat{\beta}_3 \times 170 a_p' \hat{\beta}$$

To place a confidence interval around the single predicted value y_p, we need to know

$$\begin{aligned} V(y_p - \hat{y}_p) &= V(a_p'\beta + \epsilon_p - a_p'\hat{\beta}) \\ &= V(\epsilon_p - a_p'\hat{\beta}) = V(\epsilon_p) + V(a_p'\hat{\beta}) \\ &= \sigma^2 + a_p' V(\hat{\beta}) a_p = \sigma^2 + \sigma^2 a_p'(x'x)^{-1} a_p \\ &= \sigma^2(1 + a_p'(x'x)^{-1} a_p) \end{aligned}$$

As usual, we have to replace σ^2 by $\hat{\sigma}^2 = \frac{1}{n-p-1} \sum r_i^2$, which lead to the result that

$$\frac{y_p - \hat{y}_p}{\sqrt{\hat{\sigma}^2(1 + a_p'(x'x)^{-1} a_p)}} \sim t_{n-p-1}$$

and $100(1-\alpha)\%$ CI for y_p is

$$\hat{y}_p \pm t_{n-p-1.\alpha/2} \sqrt{\hat{\sigma}^2(1 + a_p'(x'x)^{-1} a_p)}$$

What if we want to predict the mean of the response oat a given vector of values for explanatory variable, a_p'?

$$\mu_p = E[y_p] = a_p'\beta$$

The estimate of μ_p, $\hat{\mu}_p = a_p'\hat{\beta} = \hat{y}_p$ however $V(\mu_p - \hat{\mu}_p) = V(\hat{\mu}_p)$ is smaller than $V(y_p - \hat{y}_p)$. (derive $V(\mu_p - \hat{\mu}_p)$ and 95% CI for μ_p)

4.7 ANOVA Table

Consider the general model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i (\star)$$

$\hat{\beta}$ is LSE

$$SSE = \sum_{i=1}^{n} r_i^2 = r'r = (y - x\hat{\beta})'(y - x\hat{\beta})$$

Now if we consider the hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

under H_0, the general model $(\star)$ reduces to $y_i = \beta_0 + \epsilon_i \leftarrow$ restricted model the LSE of β_0 is $\hat{\beta}_0 = \bar{y}$

$$SSE(\hat{\beta}_0)\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}(y_i - \hat{\beta}_0) = \sum_{i=1}^{n}(y_i - \bar{y})^2 = SST$$

The difference

$$SSE(\hat{\beta}_0) = SSE(\hat{\beta}) = SST - SSE(\hat{\beta}) = SSR$$

is the additional sum of squares from the p-explanatory variables. SSR tells us how much variability in response is explained by the full model, over and above the simple mean model.

$$\begin{aligned} SSR &= \sum(y_i - \bar{y})^2 - (y - x\hat{\beta})'(y - x\hat{\beta}) = y'y - n\bar{y}^2 - [(I-H)y]'[(I-H)y] \\ &= y'y - n\bar{y}^2 - y'(I-H)y = \hat{\beta}'x'x\hat{\beta} - n\bar{y}^2 \end{aligned}$$

The F-test statistics

$$F = \frac{SSR/P}{SSE/n-p-1} = \frac{\text{additional sum of squares/p}}{\text{sum of squares of errors from the full model}/n-p-1}$$

is used to test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

vs

$$H_a : \text{at least one of } \beta_i \text{should be non-zero}$$

What does a statistically significant F-ratio. Imply?

- It indicates that there is strong evidence against the claim that none of the explanatory variables have an influence on response.

4.7.1 ANOVA Table

Source	df	sum of square	mean square	F
Regression Residual	p $n-p-1$	$SSR = \hat{\beta}'xx'\hat{\beta} - n\bar{y}^2$ $SSE = (y - x\hat{\beta})'(y - x\hat{\beta})$	$MSR = SSR/p$ $MSE = SSE/n-p-1$	$F = \frac{MSR}{MSE}$ same
Total	$n-1$			

The $R^2 = \frac{SSR}{SST}$ is an overall measurement of the goodness of fit of the model.

Does a large R^2 always mean that a significant relationship has been discovered? R^2 cannot decrease as more terms are added to the model (even if they are not relevant or useful)

5 Model evaluation and residual analysis

Some clarification

1. $$\hat{\beta} \sim MVN(\beta, \sigma^2(x^Tx)^{-1})$$
   $$(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)^T \sim MVN((\beta_0\beta_1\cdots\beta_p)^T, \sigma^2(x^Tx)^{-1})$$
   $$\hat{\beta}_1 \sim N(\beta_i, \sigma^2 v_{ii})$$

   where ith element of β, $(i,i)th$ entry of $(x^Tx)^{-1}$. Note that the i subscript does not represent the number for β in the model setting but actually the ith element of the β vector, i.e. β_0 is the first element, so it corresponds to $i = 1$.

   $V(\hat{\beta}_1 + \hat{\beta}_2 - \hat{\beta}_3)$, let $a_p = (0, 1, 1, -1, 0, \cdots, 0)^T$. $V(\hat{\beta}_1 + \hat{\beta}_2 - \hat{\beta}_3) a_p^T v(\hat{\beta}) a_p$

2. R^2 usually will increase as we add more explanatory variables in the model (even if they are not relevant). Suppose $p + 1 = n$, then $R^2 = 1$. Adjusted $R^2 = 1 - \frac{n-1}{n-p-1}(1 - R^2)$ to penalize for a large number of parameters.

3. R output: residual standard error $= \sqrt{MSE} = \sqrt{\hat{\sigma}^2} = \sqrt{s^2}$

4. Formally, under $H_:\beta_0 = \beta_1 = \cdots = \beta_p = 0$, the F-ratio $F = (SSR/p)/(SSE/(n - p - 1)) = MSR/MSE \sim F(p, n-p-1)$. Therefore if $F > F\alpha(p, n-p-1)$ then F is large enough to reject H_0.

5. Model evaluation and residual analysis. Given a particular dataset, a specific model (with a set of assumptions)

   - Least squares fit
   - Construct hypothesis test and confidence intervals
   - estimation and prediction

In practice, a more difficult task is to find a reason that model for a set of data. We will focus on techniques based on analysis of residuals for model checking.

5.1 Model and Model Assumptions

What is a "good" model?

- A "good" model is the one which is complex enough to provides good fit to the data and yet simple enough to use (i.e. make prediction). Well beyond the data.

Basic Model Assumptions:

1. $E[\epsilon_i] = 0 \implies E[y_i] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. This implies linearity.

2. $V(\epsilon_i) = \sigma^2$, constant variance. This implies homoscedaticity.

3. $\epsilon_1, \cdots, \epsilon_n$ are independent.

4. $\epsilon_i \sim N(0, \sigma^2)$.

Of course, we can not observe or compute the errors in practice, so their properties cannot be evaluated. Rather, we look at the residuals $r_1, \cdots, r_n$ in the fitted model

$$r_i = y_i - \hat{y}_i$$

If the residuals estimate the errors well, any pattern found in the residuals suggest that a similar relationship exists in the random error.

5.2 Relationship between residuals and random errors

We can write

$$\begin{aligned} r &= y - x\hat{\beta} = y - x(x'x)^{-1}x'y \\ &= (I - H)y = (I - H)(x\beta + \epsilon) \\ &= (I - H)\epsilon \end{aligned}$$

The residuals will approximately equal the errors if H is small relative to I. Since H is a projection matrix, H = HH, and the ith diagonal element can be written as

$$h_{ii} = (HH)_{ii} = \sum_j h_{ij}h_{ji}$$

H is also symmetric, $h_{ij} = h_{ji}$

$$h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$$

$$h_{ii}(1 - h_{ii}) = \sum_{j \neq i} h_{ij}^2$$

The right hand side is sum of squares, hence non-negative, and we see that

$$0 < h_{ii} < 1$$

If the diagonal elements h_{ii} are small, the off diagonal elements are also small.

Note

$$tr(H) = \sum h_{ii} = p + 1$$

Therefore the average of diagonal elements is $\frac{p+1}{n}$. If we try to fit nearly as many parameters as there are observations, the h_{ii}'s can not all be small relative to 1, and the residuals are poor estimator of the errors.

5.2.1 Statistical Properties of r ($r = (I - H)\epsilon$)

- $E[\epsilon] = 0$
- $V(r) = V((I - H)\epsilon) = (I - H)V(\epsilon)(I - H)' = (I - H)\sigma^2$
- $V(r) \neq V(\epsilon)$, but if H is small, they are "close", if H is not small, there could be substantial correlations among the residuals, and patterns will be apparent even if the error assumption hold.

Since r is a linear combination of ϵ, and ϵ follow $MVN(0, \sigma^2 I)$. Therefore

$$r \sim MVN(0, \epsilon^2 (I - H))$$

$$\text{then } r_i \sim N(0, \sigma^2 (1 - h_{ii}))$$

In summary, if the assumptions about ϵ hold and H is small relative to I, then

- $r = (I - H)\epsilon \approx \epsilon$
- $E[r] = 0, V(r) = \sigma^2 (I - H) \dot{\approx} \sigma^2 I$

$$r \dot{\sim} MVN(0, \sigma^2 I)$$

The residual should look approximately like a sample from un-correlated, mean zero, constant variance normal distribution.

Standardized Residuals:

$$d_i = \frac{r_i}{\sqrt{\hat{\sigma}^2 (1 - h_{ii})}}, i = 1, \cdots, n$$

$d_1, \cdots, d_n$ are approximately i.i.d. $N(0, 1)$.

5.3 Residual plots for checking $E(\epsilon_i) = 0$

Potentially, the most important assumption for linear regression models is $E[\epsilon_i] = 0$. The likely causes for violation of this assumption are:

- Effect of explanatory variables on response variable is not in fact linear
- Omission of some important explanatory variables

We shall consider three types of plot for checking this assumption:

- Residual versus $x_j, j = 1, \cdots, P$
- Partial residuals versus $x_j, j = 1, \cdots, p$
- Added-variable plots

5.3.1 Residual versus x_j

Suppose we fit a multiple regression model and

$$r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots \hat{\beta}_p x_{ip})$$

the residuals have the linear effect of x's removed from y. If x_j does have a linear effect on y (in other words the model assumption $E[\epsilon_i] = 0$ is not violated), when we plot raw residuals $r_1, \cdots, r_n$ against the n values $x_{1j}, \cdots, x_{nj}$, we expoct to see a random scatter.

On the other hand, if we see any obvious non-random pattern, it suggests the non-linearity and we could adapt the way x_j is modelled.

5.3.2 Partial residuals versus x_j

Plots of the raw residuals are some times difficult to interpret because we have to decide whether the scatter looks random or not. For

For each x_j, the partial residuals $r_i^{(j)}$ is defined as

$$r_i^{(j)} = r_i + \hat{\beta}_j x_{ij}$$

$i = 1, \cdots, n$. The estimated linear effect of x_j is added back into the residuals.

For each x_j, when we plot, we expect to see a linear trend if the model with a linear term in x_j is adequate. Hence the typical pattern of partial residual plots when the assumption is not violated is pretty linear.

the partial residuals for x_j attempt to correct y for all other explanatory variables , so that the plot of $r_i^{(j)}$ against $x_{ij} (i = 1, \cdots, n)$ shows the marginal effect of x_j

In R, a function $crPlots()$ in the car package has been made available to you to produce partial residuals plots.

5.3.3 Added-variable plots

When deciding whether a new explanatory variable should be included, an added variable plot turns out to be a more powerful graph. To provide the added-variable plot for a new explanatory variable.

5.4 Residual plots for checking constant variance $V(\epsilon_i) = \sigma^2$

Recall Residual plots for checking Model assumptions

1. If $E[\epsilon] = 0$ is ok. This is called first order assumption.

    - Residual v.s. x_j

- Parial Residual v.s. x_j.

$$r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p)$$
$$r_i^{(j)} = r_i + \hat{\beta}_j x_{ij}$$

2. $V(\epsilon_i) = \sigma^2$

   - Residual v.s. $\hat{y}_i$ If there is a fan shape, it means there is a violation of the assumption.

     Suppose we have $n = 5$, r_1, r_2, r_3, r_4, r_5. Ordered Residual $r_{(1)} < r_{(2)} < r_{(3)} < r_{(4)} < r_{(5)}$. Then those $r_{(i)}$ separates the normal distribution as 6 equal areas. Consider the following

     $$a_i = \frac{i}{n+1}, i = 1, 2, \cdots, 5,\ Pr(Z < z_i) = a_i = \Phi(z_i)$$

     where $z_i = \Phi^{-1}(a_i)$. For a Q-Q plot, it plot the ordered residual and the quantile.

5.5 Residual plots for detecting correlation in c_i's

Consequence of correlation in ϵ_i?

$$E[\hat{\beta}] = \beta$$

$$\begin{aligned} Var(\hat{\beta}) &= Var((X'X)^{-1}X'Y) = (X'X)^{-1}X'V(Y)X(X'X)^{-1} \\ &= (X'X)^{-1}X'V(\epsilon)X(X'X)^{-1} \end{aligned}$$

$Var(\epsilon)$ may be under estimated if we simply assume independence. This will be possibly to be really dangerous for hypothesis testing result.

The Durbin-Watson test is a formal statistical test for the correlation structure mentioned above. It tests $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$. The Durbin-Watson test statistic is

$$d = \frac{\sum_{i=2}^{n}(r_i - r_{i-1})^2}{\sum_{i=1}^{n} r_i}$$

D-W test for autocorrelation: $H_0 : \rho = 0$ v.s. $H_a : \rho \neq 0$. We can do a one-sided test for the hypothesis test ($\rho > 0$). The test statistic

$$d = \frac{\sum(r_i - r_{i-1})^2}{\sum r_i^2} \sim \text{Distribution}$$

If the p-value $< \alpha \leftarrow$ significant level, then we reject H_0. We conclude that there is strong evidence that the random errors are negatively/positively correlated.

```
dwtest(fit, alternative="two.sided"/"greater"/"less")
```

If p-value $> \alpha$, we can not reject H_0. We conclude that there is not enough evidence that there is autocorrelation among random errors.

6 Model Evaluation: Data Transformations

6.1 The Box-Cox Transformation

For each λ_i transform y_i to

$$Z_i = \begin{cases} y_i^\lambda & (\lambda \neq 0) \\ \log(y_i) & (\lambda = 0) \end{cases}$$

Fit the regression

$$Z_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

and calculate

$$MSE_{adj}$$

Remarks on data transformation

1. Once the transformation is selected, all subsequent estimation and tests are performed in terms of transformed values.
2. Transformation complicates the interpretation. Some transformations are easier to explain than others in some context.
3. The graphic diagnostics do not provide a clear cut decision rule. A natural criterion for assessing the necessity for transformation is whether important substantive results differ qualitatively before and after.
4. In multiple regression, the best solution may require transforming x's
5. For this course we focus on Box-Cox transformation of Y. If log transformation is chosen, then we may consider same log transformation on all explanatory variables (log-log model) if the improvement is substantial.

6.2 Logarithmic Transformation

6.2.1 Logarithmic Transformation of y only

In general, suppose we fit the model

$$\log y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

On the original scale, this model becomes

$$y = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon} = e^{\beta_0} e^{\beta_1 x_1} \cdots e^{\epsilon_p x_p} e^{\epsilon}$$

where the explanatory variables have multiplicative effects on response variable, and each appears as an exponential relationship. The multiplicative error $e^\epsilon = \epsilon^*$ has a lognormal distribution.

Interpretation of β_j: assume $x_j = a$.

$$E[y|x_j = a] = e^{\beta_0} e^{\beta_1 x_1} \cdots e^{\beta_j a} e^{\beta_{j+1} x_{j+1}} \cdots e^{\beta_p x_p} e^{\epsilon}$$

Now if $x_j = a + 1$

$$E[y|x_j = a+1] = e^{\beta_0} e^{\beta_1 x_1} \cdots e^{\beta_j (a+1)} e^{\beta_{j+1} x_{j+1}} \cdots e^{\beta_p x_p} e^{\epsilon}$$

$$\frac{E[y|x_j = a+1]}{E[y|x_j = a]} - 1 = e^{\beta_j} - 1 \implies \frac{E[y|x_j = a+1] - E[y|x_j = a]}{E[y|x_j = a]} = e^{\beta_j} - 1$$

$100(e^{\hat{\beta}_j} - 1)$ is interpreted as percentage of change in the average value of response variable per unit increase in explanatory variable x_j, while holding all the other explanatory variable fixed.

6.2.2 Logrithmic transformation of all variables

Suppose in general, we fit model

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \cdots + \beta_p \log(x_p) + \epsilon$$

On the original scale of y

$$y = e^{\beta_0} e^{\beta_1 \log(x_1)} \cdots e^{\beta_p \log(x_p)} e^{\epsilon} = e^{\beta_0} x_1^{\beta_1} \cdots x_p^{\beta_p} e^{\epsilon}$$

Essentially, explanatory variables now have multiplicative effects rather than additive effects on y, and each appears as a power relationship.

Interpretation of β_j:

$$100(e^{\hat{\beta}_j \log 1.01} - 1)$$

percentage change in average value of response variable per 1% change (increase) in x_j.

6.2.3 Logarithmic transformation of y and some x's

Consider the model with two explanatory variables

$$\log y = \beta_0 + \beta_1 \log x_1 + \beta_2 x_2 + \epsilon$$

where x_1 is transformed, but x_2 is not. On the original scale of y

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

thus x_1 has a power relationship, while x_2 has an exponential effect. In general, we can obtain a mixture of power and exponential multiplication effects.

6.2.4 95% CI for transformed estimate

Consider log model, 95% CI for y_p for a given vector of values a_p for explanatory variables.

$$\log(\hat{y}_p) = a_p'\hat{\hat{\beta}}$$

$$\hat{y}_p = e^{a_p'\hat{\hat{\beta}}}$$

There are two ways:

1. find 95% CI for $a_p'\hat{\beta} = y_p^*$ then 95% CI for $y_p = e^{a_p'\hat{\beta}}$ is
$$[e^L, e^U]$$
2. find $SE(e^{a_p'\hat{\beta}})$ based on delta method, then 95% CI for $y_p = e^{a_p'\hat{\beta}}$ is
$$e^{a_p'\hat{\beta}} \pm t_{n-p-1,0.05/2} SE(e^{a_p'\hat{\beta}})$$

6.3 Transformation for Stabilizing Variance

Consider the general model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

$$y_i = \mu_i + \epsilon_i$$

where μ_i is the mean of response. Furthermore, suppose that y_i has non-constant variance

$$V(y_i) = \mu_i^{\alpha}\sigma^2$$

where σ^2 is a constant of proportionality between the variance of y_i and the mean of y_i. If $\alpha > 0$, then variance increases with the mean. If $\alpha < 0$, then variance decreases with the mean.

Now we want to find a transformation $g(y_i)$ of y_i such that $g(y_i)$ has a construct variance. For this, we approximate $g(y_i)$ by a first-order Taylor series.

$$g(y_i) \approx g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$$

Then

$$V(g(y_i)) \approx V(g(\mu_i) + (y_i - \mu_i)g'(\mu_i)) = V(y_i)g'(\mu_i)^2 = g'(\mu_i)^2\mu_i^{\alpha}\sigma^2$$

To stabilize the variance, we may choose transformation $g(\cdot)$ such that

$$g'(\mu_i)^2 = \frac{1}{\mu_i^{\alpha}} \implies g'(\mu_i) = \frac{1}{\mu_i^{\alpha/2}}$$

Then choosing $g(y_i) = \begin{cases} \frac{y_i^{1-\alpha/2}}{1-\alpha/2} & \alpha \neq 2 \\ \log(y_i) & \alpha = 2 \end{cases}$ does the trick and lead to $V(g(y_i)) = \sigma^2$. This analysis does not tell us which function $g(\cdot)$ to choose as we do not know α and the true form of $V(y_i)$. It does, however explain why Box-Cox often choose transformation

$$y_i^\lambda$$

with $\lambda < 0$ or $\log y_i$.

6.4 Some Remedies for non-linearity- Polynomial Regression

Fit: $y = \beta_0 + \beta_1 x + \epsilon$.

Plot: $rv.s.x \implies$ non-linearity

Include higher order terms:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

Rule 1 If x^2 is in, then x should be in too. In general, if a higher term is in, all lower order terms should also be in.

Rule 2 We include a higher order term only f the new model is much "better".

7 Model Evaluation - Outliers and Influential Case

7.1 Outlier

An outlier is a particular case with unusual (extreme) value in y or/ and in x's.

Consider the following cases:

1. Case A is outlying in covariate x, but not in y. The response is right in the model.
2. Case B is not unusual with respect to x, but it is an outlier for y.
3. Cae C represents an outlier in the x as well as in y.

How to detect outliers?

- Simple diagnostic tool-graphs of standardized residuals

  $$d_i = \frac{r_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii}}}$$

  where h_{ii} is the (i,i) entry of $H = X(X'X)^{-1}X'$, and approximately

  $$d_i \sim N(0,1)$$

  Large values of d_i (e.g. $|d_i| > 2.5$) $\implies$ outlier in y.

- The real issue is not whether a case is an outlier or not; it is whether a case has a major influence on a given statistical procedure, in other words, keeping or removing the case will result in dramatically different results of the regression model $\Longrightarrow$ on fitted line $\hat{y}_i$ and on estimate $\hat{\beta}$. How to detect influential cases?

7.2 Hat Matrix and Leverage

Recall:

$$H = X(X'X)^{-1}X = (h_{ij})_{n\times n}$$

$$\hat{y} = HY$$

and the ith fitted value $\hat{y}_i$

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j$$

The weight h_{ii} indicates influence of y_i to $\hat{y}_i$.

- h_{ii} is large $\Longrightarrow$ $h_{ii}y_i$ dominates $\hat{y}_i$
- $0 \leq h_{ii} \leq 1$, if $h_{ii} \to 1$, then $\hat{y}_i \to y_i$

This implies that when h_{ii} is large, the fitted line will be forced to pass very close to the ith observation $(y_i, x_{i1}, \cdots, x_{ip})$. We say that the case i exerts high leverage on the fitted line.

Definition: h_{ii} is called the leverage value of case i. large h_{ii} → high leverage → influential on fitted line.

- The leverage h_{ii} is a function of x's but not y.
- The leverage h_{ii} is small for cases with $(x_{i1}, \cdots x_{ip})$ near the centroid $(\bar{x}_1, \cdots, \bar{x}_p)$ that is determined by all cases. The leverage h_{ii} will be large if $(x_{i1}, \cdots, x_{ip})$ is for away from the centroid. (h_{ii} is used to assess whether a case is unusual with regards to its covariates -the x dimension). e.g. Simple linear regression

$$(X'X)^{-1} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix}^{-1} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}$$

$$
\begin{aligned}
h_{ii} &= (1, x_i)\frac{1}{S_{xx}}\begin{pmatrix}\frac{1}{n}\sum x_i^2 & -\bar{x}\\ -\bar{x} & 1\end{pmatrix}\begin{pmatrix}1\\ x_i\end{pmatrix}\\
&= \frac{1}{S_{xx}}(\frac{1}{n}\sum x_i^2 - \bar{x}x_i, x_i - \bar{x})\begin{pmatrix}1\\ x_i\end{pmatrix}\\
&= \frac{1}{S_{xx}}(\frac{1}{n}S_{xx} + (x_i - \bar{x})^2)\\
&= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}
\end{aligned}
$$

The leverage is the smallest when $x_i = \bar{x}$, and it is large if x_i is far from $\bar{x}$.

Rule: The average leverage in a model with $(p+1)$ regression parameters is $\bar{h} = \frac{p+1}{n}$. If a case for which $h_{ii} > 2\bar{h} = \frac{2(p+1)}{n}$ then it is considered a high-leverage case.

7.3 Cook's Distance

It is a measure of influential on estimate $\hat{\beta}$. Consider model

$$y = X\beta + \epsilon$$

and

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

Suppose delete the ith case and fit model.

$$Y_{(-i)} = X_{(-i)}\beta + \epsilon_{(-i)}$$

where

$$Y_{-i} = \begin{pmatrix} y_1\\ \vdots\\ y_{i-1}\\ y_{i+1}\\ \vdots\\ y_n\end{pmatrix}_{(n-1)\times 1}, X_{-i} = \begin{pmatrix}1 & x_{i1} & \cdots & x_{ip}\\ 1 & x_{i-1,1} & \cdots & x_{i-1,p}\\ 1 & x_{i+1,1} & \cdots & x_{i+1,p}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & \cdots & x_{np}\end{pmatrix}_{(n-1)\times(1+p)}$$

and $\hat{\beta}_{(-i)} = (X'_{(-i)}X_{(-i)})^{-1}X'_{(-i)}Y_{(-i)}$

If ith case is influence, we expect big change in the estimate of β. The change

$$\hat{\beta} - \hat{\beta}_{(-i)}$$

is then a good measure of influence of the ith case.

Note that

- $\hat{\beta} - \hat{\beta}_{(-i)}$ is a vector, any large value in any component implies that the ith case is influential

$$(\hat{\beta} - \hat{\beta}_{(-i)})'(\hat{\beta} - \hat{\beta}_{(-i)})$$

- The magnitude of $\hat{\beta} - \hat{\beta}_{(-i)}$ should be adjusted by the variance of $\hat{\beta}$

$$V(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$$

Cook's D Statistic

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

An identity

$$\hat{\beta} - \hat{\beta}_{(-i)} = \frac{r_i}{1 - h_{ii}}(X'X)^{-1}X_i$$

where $x_i' = (1, x_{i1}, \cdots, x_{ip})$ is the ith row of X. Substituting this into the expression

$$D_i = \frac{r_i^2 X_i'(X'X)^{-1}X_i}{(1-h_{ii})^2(p+1)\hat{\sigma}^2} = \frac{d_i^2 X_i'(X'X)^{-1}X_i}{(1-h_{ii})^2(p+1)} = \frac{d_i^2 h_{ii}}{(1-h_{ii})(p+1)}$$

D_i is large if both d_i and h_{ii} are large. This implies it is an overall measure of influence. If $D_i > 1$, we will be concerned.

- D_i measures the influence of the ith case on all fitted values and on the estimated β.
- If h_{ii} large and d_i small, then D_i is small. vice versa. D_i is an overall measure of influence.
- How large is large enough? The cut off: if $D_i > 1$ (sometimes $D_i > 0.5$). We will be concerned.

Outliers and influential cases: remove or keep?

- Correct for obvious error due to data processing
- A careful decision on whether keep or remove them before/after analysis. The target population may change due to inclusion/exclusion of certain cases.
- Most investigator would hesitate to report rejecting H_0 if the removal of a case results in the H_0 not being rejected.
- Robust method - weighted least squares.

In R: Suppose fit a model

- To get Cook's D (cook.distance(fit)
- To get leverage h_{ii} $(fitinf < -influence(fit)$, $fitinf\$hat$ is just h_{ii} (contains a vector of diagonal of the "hat" matrix H))
- To get studendized residual d_i $fitsummary < -summary(fit)$, $s < -fitsummary\$sig$, $studr < -residuals(fit)/(sqrt(1 - fitinf\$hat) * s)$

8 Model Building and Selection

8.1 More Hypothesis Testing

8.1.1 Testing some but not all β's

Consider the general model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

Partition

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} X_A & X_B \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \beta_A \\ \beta_B \end{pmatrix}$$

Example: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$ with $p + 1 = 5$ parameters partitioned as

$$\beta_A = (\beta_0, \beta_1, \beta_2)', \beta_B = (\beta_3, \beta_4)'$$

and

$$X_A = \begin{pmatrix} 1 & x_{11} & x_{12} \\ \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}, X_B = \begin{pmatrix} x_{13} & x_{14} \\ \vdots & \vdots \\ x_{n3} & x_{n4} \end{pmatrix}$$

Suppose we want to test

$$H_0 : \beta_B = 0 \text{ v.s. } H_a : \beta_B \neq 0$$

for example

$$H_0 : \beta_3 = \beta_4 = 0 \text{ v.s. } H_a : \text{a least one of them is not 0}$$

we are not restricted to the last β_B elements, these ideas apply to any β_B elements

8.2 Extra Sum of Square Principle

A test follows from the change in the sum of square of regression between fitting model (1): just β_A (reduced model) and model (2): both β_A and β_B.

8.2.1 ANOVA Table for Testing Some β's

Source	df	SSR
regression fitting $_A$	p_A	$SSR(\hat{\beta}_A)$
residual fitting β_B extra to β_A	p_B	$SSR(\hat{\beta}) - SSR(\hat{\beta}_A)$
residuals	$n-p-1$	$SSE(\hat{\beta}) = r'r$
total	$n-1$	$SST = y'y - n\bar{y}$

ANOVA Version 1

Source	df	sum of square	
regression	p	SSR	$(\hat{y}-\bar{y})'(\hat{y}-\bar{y})$
residual	$n-p-1$	SSE	$r'r$
total	$n-1$	SST	$(y-\bar{y})'(y-\bar{y}) = y'Y - n\bar{y}^2$

$$\hat{\beta} = (x'x)^{-1}x'y, \hat{\beta}_0 = \bar{y}$$

ANOVA Version 2

Source	df	sum of square	
regression	p+1	SSR	$\hat{y}'\hat{y}$
residual	$n-p-1$	SSE	$r'r$
total	n	SST	$y'y$

The idea is then if $H_0 : \beta_B = 0$ is not true, the extra sum of squares of regression contributed by including β_B in the model should be large (relative to MSE)

Formally if all model assumption hold

$$F = \frac{(SSR(\hat{\beta}) - SSR(\hat{\beta}_A))/p_B}{MSE} \sim^{H_0} F(p_b, n-p-1)$$

If $F > F_\alpha(p_B, n-p-1)$, then we reject H_0 with significance level α, O.W., H_0 is not rejected.

Note that

$$SST - SSR(\hat{\beta}) = SSE$$
$$SST - SSR(\hat{\beta}_A) = SSE_0$$

where SSE_0 is sum of squares of residuals leaving out X_B (fitting the model subject to H_0) Then the difference

$$SSE_0 - SSE = SSR(\hat{\beta}) - SSR(\hat{\beta}_A)$$

Thus if the extra sum of squares of regression is small:

- The two models have similar residual sum of squares $\implies$ the two models fit about the source

- we choose simpler model $\implies$ we do not reject H_0.

  Mathematically,

$$F = \frac{(SSR(\hat{\beta}) - SSR(\hat{\beta}_A))/p_B}{MSE} = \frac{(SSE_0 - SSE)/p_B}{MSE}$$

8.2.2 The general linear hypothesis

To test the very general hypothesis concerning the regression coefficients β

$$H_0 : T\beta = b$$

where T is a $c \times (p+1)$ matrix of constance, and b is a $c \times 1$ vector of constance.

For example,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

the null hypothesis

$$H_0 : \beta_0 = 0 \text{ and } \beta_1 = \beta_2$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

thus $H_0 : T\beta = b$.

To test $H_0 : T\beta = b$ in general

1. Fit regression with no constrains

2. Compute SSE

3. Fit regression model subject to constrains

4. Compute the new SSE_0

5. Compute F-ratio

$$F = \frac{(SSE_0 - SSE)/c}{SSE/(n-p-1)}$$

6. If $F > F_\alpha(c, n-p-1)$, then reject the null hypothesis; otherwise, not reject.

Consider $H_0 : \beta_2 = \beta_3 = 0$

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

8.3 Categorical Predictors and InteractionTerms

8.3.1 Binary predictor

Recall low both weight infant example:

$$y : \text{head circa}$$

$$x_1 : \text{best age}$$

$$x_2 : \text{toxaemia}, 1 = \text{“yes”}, 0 = \text{“No”}$$

Consider model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$\hat{y} = 1.496 + 0.874 x_1 - 1.412 x_2 + \epsilon$$

- testing $\beta_2 = 0$

It is often not reasonable to assume the effect of other explanatory variables are some across different groups

Interaction terms:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

$$\implies \begin{cases} y = \beta_0 + \beta_1 x_1 + \epsilon & \text{if } x_2 = 0 \\ y = \beta_0 + \beta_2 + (\beta_1 + \beta_3) x_1 + \epsilon & \text{if } x_2 = 1 \end{cases}$$

by adding interaction term, it allows X_1 to have a different effect on y depending on the value of x_2 .

8.3.2 Hypothesis Testing of Interaction Term

$$H_0 : \beta_3 = 0 \text{ v.s. } H_a : \beta_3 \neq 0$$

tells whether the effect is different or not between groups.

8.3.3 categorical predictor with more than 2 levels

Example

$$y : \text{prestige score of occupations}$$

$$\text{exp. var : education (in years), income}$$

$$\text{type of occupation : blue collar, white collar, professional}$$

$$\text{Dummy variable} : D_1 = \begin{cases} 1 & \text{professional} \\ 0 & \text{O. W.} \end{cases}, D_2 = \begin{cases} 1 & \text{white collar} \\ 0 & \text{O. W.} \end{cases}$$

type of occupation	D_1	D_2
professional	1	0
white collar	0	1
blue collar	0	0

The categorical exp var. with k levels can be represented by $k-1$ dummies.
The regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 D_1 + \beta_4 D_2 + \epsilon$$

$$\text{prof } y = (\beta_0 + \beta_3) + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$\text{w.c. } y = (\beta_0 + \beta_4) + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$\text{b.c. } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where β_3 represents the constant vertical distance between the paralleled regression planes for prof and b.c. occupation. β_4 represents the constant vertical distance between the paralleled regression planes for w.c. and b.c. occupation.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 D_{i1} + \beta_4 D_{i2} + \epsilon_i$$

Testing Individua Hypothesis (t-test)

$$H_0 : \beta_3 = 0 \text{ v.s. } H_a : \beta_3 \neq 0$$

or

$$H_0 : \beta_4 = 0 \text{ v.s. } H_4 : \beta_4 \neq 0$$

- Testing difference between: “experimental” (prof and bc) group and “reference” (bc) group
- Testing overall effect of a categorical predictor

$$H_0 : \beta_3 = \beta_4 = 0 \text{ v.s. } H_a : \text{at least one not 0}$$

Model	Terms	df	SSE
1 (F)	x_1, x_2, D_1, D_2	93	4681.28
2 (R)	x_1, x_2	95	5272.44

$$F = \frac{(SSE_0 - SSE)/2}{SSE/93} = 5.95 > F_{0.05}(2.93) = 3.07 \sim F(2,93)$$

Therefore we reject H_0 and conclude that occupation type is overall significantly related to prestige score (the dummy variables are significant).

8.3.4 Modeling Interaction with categorical predictors

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 D_{i1} + \beta_4 D_{i2} + \beta_5 x_{i1} D_{i1} + \beta_6 x_{i1} D_{i2} + \beta_{7i2} D_{i1} + \beta_8 x_{i2} D_{i2} + \epsilon_i$
where $\beta_3 D_{i1} + \beta_4 D_{i2}$ is the mean effect, $\beta_5 x_{i1} D_{i1} + \beta_6 x_{i1} D_{i2}$ is education × type, $\beta_{7i2} D_{i1} + \beta_8 x_{i2} D_{i2}$ is income × type.

This model also can be written as

$$\text{prof: } y = (\beta_0 + \beta_1) + (\beta_1 + \beta_5)x_1 + (\beta_2 + \beta_7)x_2 + \epsilon$$

$$\text{w.c.: } y = (\beta_2 + \beta_4) + (\beta_1 + \beta_6)x_1 + (\beta_2 + \beta_8)x_2 + \epsilon$$

$$\text{b.c.: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where β_5, β_6 represent effect of interaction between education and occupation type; β_7 and β_8 represent effect of interaction between income and occupation type.

8.3.5 To test the significance of the interaction

For example, $H_0 : \beta_7 = \beta_8 = 0$ v.s. $H_a :$ at least one is not 0

Model	Terms	df	SSE
1	$x_1, x_2, D_1, D_2, x_1D_1, x_1D_2, x_2D_1, x_2D_2$	89	3552.624
2	$x_1, x_2, D_1, D_2, x_1D_1, x_2D_2$	91	4504.982

We use F-test, $F = \frac{(SSE_0 - SSE)/2}{SSE/89} = 11.929 > F_{0.05}(2.89) = 3.099$. Hence, we reject H_0 and conclude that there is significant evidence that the relationship between income and prestige score is different across different occupation type.

8.4 Variable Selection

Often many explanatory variables are available, investigators need to choose which to use.

Introduce only important variables

- Model is simpler and easier to understand
- Cost of prediction is reduced (fewer variables to measure)
- Accuracy of predicting new y's may improve (unnecessary explanatory variables inflates variance)

Algorithms

- Forward selection
- Backward elimination
- Stepwise regression
- Criterion based all subsets regression

8.4.1 Examples and discussion of which method to use follows

We illustrate the variable selection methods on some data on the 50 states in the USA from the 1970s. We will take the life expectancy as the response and the remaining variables as predictors.

```
State
Population
Income
Illiteracy
Life_Exp
Murder
Hs_Grad
Frost
Area
Use all variables in model:
g<-lm(Life_Exp~., data=statedata)
```

8.4.2 Backward elimination

1. Start with all p potential explanatory variables

2. Calculate the p-value based on either t-test or F-test on null hypothesis all betas = 0

3. Remove the explanatory variable with the largest p-value or the smallest —t— value, in this case area is dropped first, if the p-value is greater than alpha. Stop if not.

4. Repeat the procedure until all p-values for remaining variables are less than the significance level alpha.

Note: t-test and F-test give different results

Note: t-test can't be used for categorical predictors, must use the F-test if you have categorical predictors.

```
g<-update(g, .~. -Area)
```

Note that the final model depends on the significance level alpha (alpha to drop), the larger the alpha is, the bigger the final model is.

Once a variable is removed, it is never reconsidered (could be a problem)

8.4.3 Forward Selection

1. Fit p simple linear models, each with only a single explanatory variable. Use p t-test statistics and p-values on our null hypothesis, and choose the most significant predictor which has the smallest p-value. If the smallest p-value ¡ alpha, stop the algorithm.

2. Start with the single predictor model and repeat to add additional variables.

```
null<-lm(Life.Exp~1, data=statedata)
fullmodel<-lm(Life.Exp~., data=statedata
newmodel<- addterm(null, scope=fullmodel, test="F")
newmodel<-lm(Life.Exp~murder, data=statedata)
addterm(newmodel, scope=fullmodel, test="F")
```

Once a variable is added, it stays in the model (could be a problem, as it may become insignificant).

8.4.4 Stepwise Regression

It is a combination of backward and forward method. Depends on two alphas: 1 to enter and 2 to drop. At each stage a variable may be added or removed and there are several variations

Example

1. Start as in forward selection using alpha 1

2. At each stage once a predictor entered the model, check all other predictors previously in the model for their significance. Drop the least significant predictor if its p-value is greater than alpha 2.

3. Continue (do forward, then backward, etc) until no predictors can be added or removed.

Remarks

"one at a time" nature makes it possible to miss optimal model. Procedures not directly linked to final objective. Variable selection amplifies statistical significance of remaining variables, but the other variables may still be correlated with the response. All automatic algorithms should be used with caution. When there is multicollinearity among explanatory variables, the three methods may lead to quite different results.

8.5 The Principle of Marginality

$y_i = \beta_0 + \beta_0 x_{i1} + \beta_2 x_{i2} + \beta_3 D_{i1} + \beta_4 D_{i2} + \epsilon_i + \beta_5 x_{i1} D_{i1} + \beta_6 x_{i1} D_{i2} + \beta_7 x_{i2} D_{i1} + \beta_8 x_{i2} D_{i2}$

- If a model include higher order term, then the lower order term should also be included
- exam higher oder term (interactions) first, then proceed to test,estimate and interpret main effects.

8.6 All Subsets Regressions

Suppose we start with a regression model with p explanatory variables. Thus there are 2^p possible regressions. In principle, we can fit each regression and choose the "best" model based on some "fit" criterion.

Numerical criteria for model comparison

- R-square. The bigger the better.
- Adjusted R-square

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

  where p is the number of explanatory variables in the model. A large model may have a smaller R^2_{adj}.

- Mallows' C_k

  $$C_k = \frac{SSE_k}{MSE_{full}} - (n - 2(k+1))$$

  If the subset model (or candidate model) is adequate, then we expect

  $$E[\frac{SSE_k}{n-k-1}] \approx \sigma^2$$

  $$E[SSE_k] \approx (n-k-1)\sigma^2$$

  We also know that

  $$E[SSE/n-p-1] = \sigma^2$$

  Therefore,

  $$E[C_k] = E[\frac{SSE_k}{MSE} - (n - 2(k+1)] \approx k+1$$

  A candidate model is good if $C_k \leq k+1$. Look for the simplest model for which C_k is close to $k+1$.

- AIC

$$AIC = -2(\text{max log likelihood} - (p+1)) = n\log(\frac{SSE}{n}) + 2(p+1))$$

Under linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

then we know

$$y_i \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2)$$

and y_i's are independent.

The likelihood function

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} f(y_i) = f(y_1, \cdots, y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2}{2\sigma^2}\}$$

$$l(\beta, \sigma^2) = \log L(\beta, \sigma^2) = \frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2}{2\sigma^2}]$$

The LSE $\hat{\beta}$ are same as MLE $\hat{\beta}$.

$$l(\hat{\beta}, \sigma^2) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2} SSE$$

$$\frac{\partial l(\hat{\beta}, \sigma^2)}{\partial \sigma^2} = \frac{-n}{2}\frac{2\pi}{2\pi\sigma^2} + \frac{1}{2\sigma^4} SSE = 0$$

$$\hat{\sigma}^2 = \frac{SSE}{n}$$

$$l(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\frac{SSE}{n} - \frac{n}{2} = \text{constant} - \frac{n}{2}\log\frac{SSE}{n}$$

  - For linear regression model, the maximum log-likelihood is

$$l(\hat{\beta}, \hat{\sigma}^2) = \text{constant} - \frac{n}{2}\log\frac{SSE}{n}$$

  - AIC is a penalized maximum log-likelihood
  - Small AIC means better model.

Note that for a model of a given size (here size refers to the number of explanatory variables included in the model), all the criterion above will select the model with the smallest sum of squares of errors (SSE).

9 Multicollinearity in Regression Models

9.1 Multicollinearity

An example: pizza sales data

$$y : \text{sales (\$1000's)}\}$$

$$x_1 : \text{number of advertisements}$$

$$x_2 : \text{cost of advertisements (\$100's)}$$

Suppose fit a model

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

and get following results

	Est.	S.E.	t-value	p-value
Intercept	24.82	5.46	4.39	0.0007
x_1	0.66	0.54	1.23	0.2404
x_2	1.23	0.70	1.77	0.1000

$R^2 = 0.7789$, F-statistic: 22.899 on 2 and BDF. p-value is 0001

What do we find?

- $R^2 = 0.7789, x_1$ and x_2 together explain a large part (78%) of the variability in sales
- F-statistic and p-value indicate that one of them is important.
- we cannot reject $H_0 : \beta_1 = 0$ when x_2 is in the model. Similarly, we cannot reject $H_0 : \beta_2 = 0$ when x_1 is in the model. In other words, if one of x_1 and x_2 is in the model, then the extra contribution of the other variable toward the regression is not important. The individual t-test indicates that you do not need one variable if you already included the other.

This is because variables x_1 and x_2 are highly correlated. The two variables appear to express the same information. So no point to include both.

Definition: Collinearity: linear relationship between x_1 and x_2. Multicollinearity: there is a linear relationship involving more than two x variables. for example $x_1 \approx x_2 + x_3$.

9.2 Consequence of Multicollinearity

If there is an exact linear dependence

$$X = [1, x - 1, \cdots, x_p]$$

where $x_k = (x_{1k}, \cdots, x_{nk})$ is the $k + 1$ th column of X.

If one of x_k is a linear combination of the other, say

$$x_1 = c_0 1 + c_2 x_2 + \cdots + c_p x_p$$

then

$$rank(X) < p+1 \implies rank(x'x) < p+1$$

Hence $|x'x| = 0$, and $(x'x)^{-1}$ does not exist, not be able to solve $\hat{\beta}$. Under multicollinearity

$$|x'x| \approx 0$$

It is computationally unstable for

$$\hat{\beta} = (x'x)^{-1}x'y$$

- insignificance of important predictors
- opposite sign of $\hat{\beta}$
- large s.e. and wide c.i.

9.3 Detection of Multicollinearity among $(x_1, \cdots, x_n)$

First look pairwise sample correlations

$$r_{l_m} = \frac{\sum_{r=1}^{n}(x_{il} - \bar{l})(x_{im} - \bar{x}_m)}{\sqrt{\sum_{i=1}^{n}(x_{il} - \bar{x}_l)^2 \sum_{i=1}^{n}(x_{im} - \bar{x}_m)^2}}$$

r_{lm} measures the linear association between any two x variables, x_l and x_m.

$$\begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ \vdots & \ddots & \ddots & \vdots \\ r_{p1} & \cdots & \cdots & 1 \end{pmatrix}$$

$|r_{lm}| = 1$, x$_l$ and x_m are strongly linearly related.

A formal check: variance inflation factor. x_k is regressed (x_k is used as response) on the remaining $p-1$ x's.

$$x_{ik} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{k-1}x_{ik-1} + \beta_{k+1}x_{ik+1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

for $k = 1, 2, \cdots, p$.

The resulting

$$R_k^2 = \frac{SSR_k}{SST}$$

is a measue of how strongly x_k is linearly related to the rest of x's.

$$R_k^2 = 1 \implies \text{perfect linear}$$

$$R_k^2 = 0 \implies \text{not linearly related}$$

Variance Inflation Factors (VIF)

$$VIF_k = \frac{1}{1 - R_k^2} (\geq 1)$$

where $k = 1, \cdots, p$

- $VIF_k > 10$, strong evidence of multicollinearity
- $VIF_k \in [5, 10]$, since evidence of multicollinearity

9.4 Ridge Regression

Ridge regression is used when the design matrix X is multi collinear and the usual least squares estimate of β appear to be unstable.

$$\text{LSE } \hat{\beta} \min (y - x\beta)'(y - x\beta)$$

For $|x'x| \approx 0$, ridge regression makes the assumption that the regression coefficients are not likely to be very large. Suppose we place some upper bound on β

$$\sum_{j=1}^{p} \beta_j^2 = \beta'\beta < c$$

Minimize subject to constrains (lagrange Multiplier Method)

$$\min (y - x\beta)'(y - x\beta) + \lambda c \sum_{j=1}^{p} \beta_j^2$$

the second term is penalty depending on $\sum_{i=1}^{p} \beta_j^2$.

Ridge regression minimize

$$(y' - x\beta)'(y - x\beta) + \lambda \sum_{j=1}^{p} \beta_j^2$$

- in statistics, this is called “shrinkage”; you are shrinking $\sum \beta_j^2$ toward 0.
- λ is a shrinkage parameter that you have to choose
- the ridge regression solution $\hat{\beta}^k$

$$\frac{\partial}{\partial \beta} [(y' - x\beta)'(y - x\beta) + \lambda \beta'\beta] = 0$$

$$x'x\beta - x'y + \lambda\beta = 0$$

Therefore $\hat{\beta}^k = (x'x + \lambda I)^{-1} x'y$. Ridge regression is not an unbiased estimator.

Note that

- $\hat{\beta}^k$ is biased for β (LSE $\hat{\beta}$ is unbiased)
- choose λ such that
  1. bias is small
  2. $|X'X + \lambda I| \neq 0$.
  3. variance not large

10 Final Instruction

Final exam is similar to the midterm.