Medical Statistics Summary Sheet

Shravan Vasishth (vasishth@uni-potsdam.de)

February 24, 2015

Contents

Medical Statistics Summary Sheet
Compiled by: Shravan Vasishth (vasishth@uni-potsdam.de)
Version dated: February 24, 2015

Types of expt designs

Parallel Group Designs

: To compare k treatments, divide patients, at random, into k groups, the n_i patients in group i receive treatment i. Each patient receives just one treatment. Comparisons are between patients. n_i not necessarily the same across groups.

In Series Designs

Each patient receives all k treatments in the same order. Comparisons made within patients.
Problems: Patients enter when disease is bad, hence likely to improvement regardless of treatment, so later treatments appear better. Reverse occurs for a progressive disease, i.e. problems occur if underlying disease is not stable.
Advantages

1. Need fewer patients than parallel designs
2. Patients can state ?preferences? between treatments
3. Might be able to allocate treatments simultaneously e.g. skin cream on left and right hands

Disadvantages

1. Treatment effect might depend on when it is given
2. Treatment effect may persist into subsequent periods and mask/modify effects of later treatments
3. Withdrawals cause problems (i.e. if a patient leaves before trying all treatments)
4. Not universally applicable, e.g. drug treatment compared with surgery
5. Can only use for short term effects

Crossover trials

Period, Treatment and Period:Treatment (Carrover) effects

1. All patients get all treatments but in different orders.
2. Period and Carryover effects are nuisance variables, the main interest is in Treatment effects.

Factorial designs

Sequential designs

Advantages

1. Detect large differences quickly
2. Avoids ethical problem of fixed size designs (no patient should receive treatment known to be inferior)

Disadvantages

1. Responses needed quickly (before next pair arrive)
2. Drop-outs cause difficulties
3. Constant surveillance necessary
4. Requires pairing of patients
5. Calculation of boundaries highly complex

Per protocol vs intention to treat analyses

Per protocol: only analyze patients who conform to original protocol
Intention to treat: analyze all data, including dropouts etc. (This has lower risk of bias).

Randomization

Randomization protects against accidental and selection bias, and provides a basis for statistical tests
Types of randomization include

1. simple (but may be unbalanced over treatments)
2. blocked (but small blocks may be decoded)
3. stratified (but may require small blocks)
4. minimization (but lessens randomness)

Adjustment of p-value under repeated analyses as data comes in

Pocock 1983

Crossover trials

carryover, treatment, period

Randomization

Historical/database controls: use previous data as control and assign all patients to treatment. As a compromise, one could have a small number of controls to compare with historical controls.

Unequal allocation (4.3.2)

May decide we need most information on B to get more accurate estimates of the B effect; variation A is probably known reasonably well already if it is the standard.

Stratified randomization: Suppose we have patients in different age ranges and either M or F. Then find all possible combinations of each level, and then produce separate randomization lists for each level.

Adaptive randomization/minimization:

Crossover trials

Combining trials

Manten-Haenszel Test

Here, we assume a data format as follows:

For each study we can compute the mean and variance using the formulas from the lecture notes:

$$E[Y_1] = \frac{n_1 t}{n} \quad Var(Y_1) = \frac{n_1 n_2 t(n-t)}{n^2(n-1)} \tag{1}$$

Then, the statistic is

$$T_{MH} = (Y_i - E[Y_i])^2 / Var(Y_i) \sim \chi_1^2 \tag{2}$$

If we have multiple studies i, we use the above procedure to get all the $E[Y_i]$, and $Var(Y_i)$, and compute $W = \sum Y_i, E[W] = \sum E[Y_i], Var(W) = \sum Var(Y_i)$, and then use the above test again on the W's.

$$T_{MH} = (W - E[W])^2 / Var(W) \sim \chi_1^2 \tag{3}$$

This summing up procedure avoids Simpson's paradox (combining studies can give different results than separate analyses, due (inter alia) to sample size differences), but not sure why.

Some notes on MH test:

1. This test is appropriate when treatment differences are consistent across tables.
2. Logistic regression gives you the same results.
3. "The Mantel-Haenszel test is simpler if one has just two qualitative prognostic factors to adjust for and wishes only to assess significance, not magnitude, of a treatment difference." (p. 97)
4. "The logistic approach is more general and can include other covariates, further, it can test whether treatment differ- ences are consistent across tables." (p. 97)
5. If treatments across trials is inconsistent or if success rates differ markedly.
6. Logistic regression can solve Simpson's paradox because we can include trial effect.

Survival analysis

```
                      right censoring:
t0 ----------o t1     t0 ---------- c
start       dead      start       lost
```

Survival time $t \geq 0$. Define a random variable $T \sim f(x)$ where the cdf is

$$F(x) = P(T < t) = \int_0^t f(u)\, du \tag{4}$$

The probability that survival is $\geq$ t is

$$S(t) = 1 - F(t) = P(T \geq t) \tag{5}$$

Hazard function: Consider $P(t \leq T < t + \delta t \mid T \geq t)$. Divide by δt to get probability *per unit time* = rate.

Hazard rate definition:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t \mid T \geq t)}{\delta t} \right\} \tag{6}$$

If we rearrange terms, we get:

$$h(t)\delta t = \lim_{\delta t \to 0} \{P(t \leq T < t + \delta t \mid T \geq t)\} \tag{7}$$

This is the probability of dying during $t + \delta t$; or the risk of death *at time t*.

Focusing on the right-hand side $P(t \leq T < t+\delta t \mid T \geq t)$, we can use the conditional probability rule to determine that:

$$\begin{aligned} P(t \leq T < t + \delta t \mid T \geq t) &= \frac{P(t \leq T < t + \delta t)}{P(T \geq t)} \\ &= \frac{F(t + \delta t) - F(t)}{P(T \geq t)} \end{aligned} \tag{8}$$

From equation 5 we know that $P(T \geq t) = S(t)$. So we can restate $h(t)$ as:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t) - F(t)}{\delta t} \frac{1}{S(t)} \right\} \tag{9}$$

Now,

$$\lim_{\delta t \to 0} \left\{ \frac{F(t+\delta t) - F(t)}{\delta t} \right\} = \frac{d(F(t))}{dt} = f(t) \tag{10}$$

Therefore:

$$\boxed{h(t) = \frac{f(t)}{S(t)}} \tag{11}$$

Now,

$$\frac{d(\log(S(t)))}{dt} = -\frac{f(t)}{S(t)} \tag{12}$$

This is because

$$S(t) = 1 - F(t) = 1 - \int_0^t f(t)\, dt) \tag{13}$$

Taking logs:

$$\log S(t) = \log(1 - F(t)) = \log(1 - \int_0^t f(t)\, dt) \tag{14}$$

If we now take the derivative of $\log S(t)$ with respect to t:

$$\frac{d(\log S(t))}{dt} = \frac{d(\log(1 - F(t)))}{dt} = \frac{d(\log(1 - \int_0^t f(t)\, dt))}{dt} \tag{15}$$

We use the chain rule to solve this derivative: Let $u = 1 - \int_0^t f(t)\, dt = S(t)$. We can write:

$$\frac{d(\log(1 - \int_0^t f(t)\, dt))}{dt} = \frac{d(\log(u))}{dt} \tag{16}$$

Now, $\frac{du}{dt} = -f(t)$; also, $\frac{d(\log u)}{du} = \frac{1}{u} = \frac{1}{S(t)}$. So, by the chain rule:

$$\frac{d(\log u)}{du}\frac{du}{dt} = \frac{d(\log u)}{dt} = -\frac{f(t)}{S(t)} \tag{17}$$

Therefore:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d(\log(S(t)))}{dt} \tag{18}$$

The cumulative distribution function H(t):

$$H(t) = \int_0^t h(u)\, du = -\log(S(t)) \tag{19}$$

Since $\log(S(t)) = -H(t)$, if we take exponents on both sides:

$$\boxed{S(t) = \exp(-H(t))} \tag{20}$$

Since $f(t) = h(t)S(t)$, replacing S(t), we get:

$$\boxed{f(t) = h(t)S(t) = h(t)\exp(-H(t))} \tag{21}$$

In summary, for any random variable T, we will define, f(t), S(t), and h(t).

	Exponential	*Weibull*
$f(t)$	$\lambda \exp(-\lambda t)$	$\lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma})$
$S(t)$	$\exp(-\lambda t)$	$\exp(-\lambda t^{\gamma})$
$h(t)$	λ	$\lambda\gamma t^{\gamma-1}$
		$Alternative: \lambda\gamma(\lambda t)^{\gamma-1} = \lambda^{\gamma}\gamma t^{\gamma-1}$

In Weibull, if $\gamma > 1$ hazard is increasing, and if $\gamma < 1$ hazard is decreasing. If $\gamma = 1$, then the Weibull reduces to the exponential.

Life tables

: to-do

Kaplan-Meier product estimates of S(t)

Here, we estimate the survival distribution without making any assumptions. The estimator is a **non-parametric MLE**.

1. k: number of failures
2. $t_1, \ldots, t_k$ unique event times (ordered)
3. d_i: deaths at t_i
4. n_i: number at risk (still alive) at t_i

$$\hat{S}(t) = \prod_{i:t_i<t} (1 - \frac{d_i}{n_i}) = \prod_{i:t_i<t} (n_i - \frac{d_i}{n_i}) \tag{22}$$

$$\hat{H}(t) = -\log \hat{S}(t) \tag{23}$$

With censoring:

1. j is the time index: $j = 1, \ldots, k$, i.e., $t_1, \ldots, t_k$.
2. d_j is the failure at time index j.
3. n is total number of observations $\sum d_j$
4. I_j: number of individuals censored during $t_{j-1} < t < t_j$.
5. r_j is the number at risk (the number alive) just before time t_j.
6. Note that $r_1 = n - I_1$.
7. For $j \geq 2, r_j = r_j - d_j - I_{j+1} = n - (d_1 + \cdots + d_{j-i}) - (I_1 + \cdots + I_j)$.

$$\hat{S}(t) = \prod_{j=1}^{s}(1 - \frac{d_j}{r_j}) \text{ for } t_{(s)} < t < t_{(s+1)} \tag{24}$$

Note:

1. This assumes that I_j censoring survive up to the preceding time period t_{j-1} and then are removed immediately; this is different from life tables.
2. KM estimates are used when the intervals between events are quite short and the number of withdrawals in any interval is therefore quite small.

Greenwood's variance formula:

$$Var(\hat{S}(t)) = (\hat{S}(t))^2 \sum_{j=1}^{s} \frac{d_j}{r_j(r_j - d_j)} \text{ for } t_{(s)} < t < t_{(s+1)} \tag{25}$$

$$\hat{H}(t) = \sum_{j=1}^{s} \frac{d_j}{r_j} \text{ for } t_{(s)} < t < t_{(s+1)} \tag{26}$$

Simple example from Harrell's Regression Modeling Strategies (p. 402)
Given failure times (+ denotes censoring): $1, 3, 3, 6^+, 8^+, 9, 10^+$.

j	t_i	I_j	r_i	d_i	$(r_i - d_i)/r_i$	S(t)	interval
0	0	0				1	$0 \leq t < 1$
1	1	0	7	1	6/7	6/7=0.85	$1 \leq t < 3$
2	3	0	6	2	4/6	$(6/7) \times (4/6) = 0.57$	$3 \leq t < 9$
3	9	2	2	1	1/2	$(6/7) \times (4/6) \times (1/2) = 0.29$	$9 \leq t < 10$

In R:

```
> library(survival)
> time<-c(1,3,3,6,8,9,10)
> censor<-c(1,1,1,0,0,1,0)
> df<-data.frame(time=time,censor=censor)
> harrell_sv<-with(df,Surv(time,censor,
                           type="right"))
> harrell_sv
```

```
[1]  1   3   3   6+  8+  9  10+
```

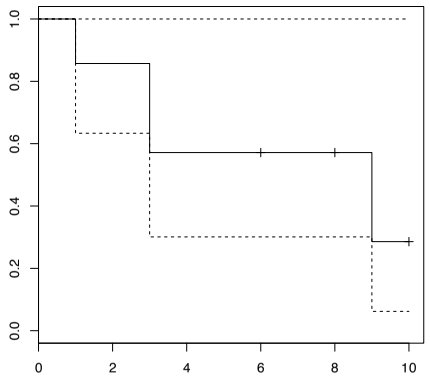
```
> harrell_fit<-survfit(harrell_sv~1,data=df)
> summary(harrell_fit)
```

```
Call: survfit(formula = harrell_sv ~ 1, data = df)

 time n.risk n.event survival std.err lower 95% CI
    1      7       1    0.857   0.132        0.633
    3      6       2    0.571   0.187        0.301
    9      2       1    0.286   0.223        0.062
 upper 95% CI
            1
            1
            1
```

Figure 1: Kaplan-Meier plot for Harrell example.

```
> plot(harrell_fit)
```



Note that the estimate of $S(t)$ is undefined for $t > 10$.
Example from lecture notes (p. 19):

```
> library(survival)
> load("data/tumour.Rdata")
> ## Note:
> ## censor: 0=censored, 1=complete
> head(tumour)
```

```
  time censor
1  3.0      1
2  4.0      0
3  5.7      0
4  6.5      1
5  6.5      1
6  8.4      0
```

```
> tumour_sv<-with(tumour,Surv(time,censor,
                              type="right"))
> tumour_sv
```

```
 [1]  3.0   4.0+  5.7+  6.5   6.5   8.4+ 10.0  10.1+ 12.0
[10] 15.0
```

```
> tumour_fit<-survfit(tumour_sv~1,data=tumour)
> summary(tumour_fit)
```

```
Call: survfit(formula = tumour_sv ~ 1, data = tumour)

 time n.risk n.event survival std.err lower 95% CI
  3.0     10       1    0.900  0.0949        0.7320
  6.5      7       2    0.643  0.1679        0.3852
 10.0      4       1    0.482  0.1877        0.2248
 12.0      2       1    0.241  0.1946        0.0496
 15.0      1       1    0.000     NaN            NA
 upper 95% CI
            1
            1
            1
            1
           NA
```

to-do? (not sure if needed): Computing CIs by hand. See Harrell.

R code: Representative examples

The `survival` package has the following key functions:

1. `Surv`
2. `survfit`
3. `survreg`
4. `survdiff` (log rank two sample test)

The package `coin` does conditional tests (`surv_test`): to-do. See ??survival for a reference to an excellent vignette by Hothorn and Everett.

Parametric models (single-sample data)

Given non-negative failure times $T \sim f(t)$, cdf, $F(t)$, and $S(t) = 1 - F(t), h(t) = \frac{f(t)}{S(t)}$. The pdf $f(t)$ depends on some parameter θ; we use MLE to estimate θ and get its variance and therefore get CIs for the parameter.

Exponential Recall that for the exponential distribution:

$$f(t) = \lambda \exp(-\lambda t) \quad S(t) = \exp(-\lambda t) \quad h(t) = \lambda \tag{27}$$

If the data are **uncensored**:

$$\hat{\lambda} = \frac{n}{\sum t} \quad 95\%CI : \left[\frac{\chi^2_{2n,0.025}}{2\sum t}, \frac{\chi^2_{2n,0.975}}{2\sum t} \right] \tag{28}$$

Also:

$$\hat{S}(t) = \exp(-\hat{\lambda} t) \tag{29}$$

How the above comes about:

$$L(\lambda; t_1, t_2, \ldots, t_n) = \prod f(t_i) = \lambda^n e^{-\lambda \sum t_i} \tag{30}$$

$$\ell(\lambda) = n \log(\lambda) - \lambda \sum t_i \Rightarrow \hat{\lambda} = \frac{n}{\sum t} \tag{31}$$

Confidence intervals:
Recall two facts: $Y = \sum T_i \sim Gamma(n, \lambda)$ and $Z = 2\lambda Y \sim \chi^2_{2n}$.
$P(\chi^2_{2n,0.025} < 2\lambda \sum T_i < \chi^2_{2n,0.975} = 0.95$, and so a 95% CI CI is $\left[\frac{\chi^2_{2n,0.025}}{2\sum t}, \frac{\chi^2_{2n,0.975}}{2\sum t} \right]$

If the data are **censored**:
There are two cases. We either have complete observations, in which case
$f(x) = \lambda \exp(-\lambda t)$
Or we have censored observations, in which case
$f(x) = \exp(-\lambda c_i)$ (not sure why this is so)
The above assume c_i are fixed and are given for all individuals (i.e., non-random). I.e., for complete observations, we have $t_i \leq c_i$ and for the censored ones we have $t_i > c_i$.
To define the likelihood, we define a censoring indicator $\delta_i = 1$ if we have a complete observation, and $\delta_i = 0$ if censored. Then:

$$L(\lambda) = \prod [\exp(-\lambda t_i)]^{\delta_i} [\exp(-\lambda c_i)]^{1-\delta_i} \tag{32}$$

taking the log likelihood:

$$\ell(\lambda) = \log \lambda \sum \delta_i - \lambda \sum t_i \delta_i - \lambda \sum (1 - \delta_i) c_i \tag{33}$$

Taking the derivative:

$$\frac{d\ell}{d\lambda} = \frac{\sum \delta_i}{\lambda} - \sum (t_i \delta_i + (1 - \delta_i) c_i) \Rightarrow \hat{\lambda} = \frac{\sum \delta_i}{\sum (t_i \delta_i + (1 - \delta_i) c_i)} \tag{34}$$

Note that $\frac{d^2\ell}{d\lambda^2} = -\frac{1}{\lambda^2} \sum \delta_i$, and therefore
$-\frac{d^2\ell}{d\lambda^2} = \frac{1}{\lambda^2} \sum \delta_i$.
We can use the asymptotic properties of MLEs to get:

$$\hat{\lambda} \xrightarrow{d} N(\lambda, I^{-1}) \quad I = E[-\frac{\delta^2 \ell}{\delta \lambda^2}] = E[\frac{1}{\lambda^2} \sum \delta_i] \tag{35}$$

To find $E[-\frac{\delta^2\ell}{\delta\lambda^2}]$ we have to find the expectation of $\sum \delta_i$. Now:

$$\begin{aligned} E[\delta_i] =& 1 \times P(T_i < c_i) + 0 \times P(T > c_i) \\ =& 1 - \exp(-\hat{\lambda} c_i) \end{aligned} \tag{36}$$

It follows that $\sum \delta_i = \sum (1 - \exp(-\hat{\lambda} c_i))$.
Therefore:

$$Var(\hat{\lambda}) = I^{-1} = \frac{1}{E[-\frac{\delta^2\ell}{\delta\lambda^2}]} = \frac{\hat{\lambda}^2}{\sum (1 - \exp(-\hat{\lambda} c_i))} \tag{37}$$

Estimating the mean
For the exponential, $\hat{\mu} = \frac{1}{\hat{\lambda}}$.
Next, we compute the variance. Recall: $Var(g(\hat{\lambda})) = [g'(\lambda)^2 var(\lambda)]_{\lambda=\hat{\lambda}}$.

$$Var(\hat{\mu}) = var(\frac{1}{\hat{\lambda}}) = \frac{\hat{\mu}^2}{\sum (1 - \exp(-\hat{\lambda} c_i))} \tag{38}$$

[Is the above a mistake? Need to check this.]
Estimating the median
To estimate the median S_α, note that there is some value S_α such that $\alpha = P(T \geq S_\alpha) = S(S_\alpha) = \exp(-\lambda S_\alpha)$.
It follows that

$$\begin{aligned} \alpha &= \exp(-\lambda S_\alpha) \\ \leftrightarrow \log \alpha &= -\lambda S_\alpha \\ \therefore S_\alpha &= -\frac{\log \alpha}{\lambda} \end{aligned} \tag{39}$$

So,

$$\begin{aligned} Var(S_\alpha) =& Var(-\frac{\log \alpha}{\lambda}) \\ =& (-\log \alpha)^2 Var(\frac{1}{\lambda}) \\ =& (-\log \alpha)^2 \frac{\hat{\mu}^2}{\sum \delta_i} \end{aligned} \tag{40}$$

I didn't understand how we got the variance of $\hat{\mu} = 1/\hat{\lambda}$ to be $\hat{\mu}^2/\sum \delta_i$.
Since $\hat{\mu} = g(\hat{\lambda}) = 1/\hat{\lambda}$, it follows that $g''(\lambda) = 1/\lambda^3$.
So, $Var(g(\lambda)) = g''(\lambda)var(\lambda) = (1/\lambda^3)(\lambda^2/\sum \delta_i) = (1/\lambda^3)(1/\sum \delta_i) = \mu/\sum \delta_i$ and not $\mu^2/\sum \delta_i$.

Example: Lung cancer data.

Sampling Theory and Design of Experiments

Review of General Linear Models

[Also see LinearModelsSummary.pdf]
A deterministic model would be $y = f\phi(x, \beta) = \beta_0 + \beta_1 x$. Cf. a non-deterministic model: $y = f\phi(x, \beta, \epsilon) = \beta_0 + \beta_1 x + \epsilon$. The general linear model is:

$$Y = \sum_{i=1} f_i(x_i)\beta_i + \epsilon \quad E[Y] = \sum \mathbf{f}(\mathbf{x})\beta \tag{41}$$

The matrix formulation:

$$Y = X\beta + \epsilon \Leftrightarrow y_j = f(x_j)^T\beta + \epsilon_j, i = 1, \ldots, n \tag{42}$$

$E[Y] = X\beta$. X is the **design matrix**.
Example: $y = \beta_0 + \beta_1 x + \epsilon$. Here, $f(x) = (1x)$.

Least squares estimation: Geometric argument When we have a deterministic model $y = f\phi(x, \beta) = \beta_0 + \beta_1 x = X\beta$, this implies a perfect fit to all data points. This is like solving the equation $Ax = b$ in linear algebra: $X\beta = y$.
When we have a non-deterministic model $y = f\phi(x, \beta, \epsilon) = \beta_0 + \beta_1 x + \epsilon$, there is no unique solution. Now, the equation Ax is an approximation to b in $Ax = b$. We try to get Ax as close to b as possible, i.e., $\mid b - Ax \mid$ is minimized. The problem now becomes finding $\hat{x}$ such that $A\hat{x} = \hat{b}$.
Now, notice that $(Y - X\hat{\beta})$ and $X\beta$ are perpendicular to each other, i.e.,

$$(Y - X\hat{\beta})^T X\beta = 0 \Leftrightarrow (Y - X\hat{\beta})^T X = 0 \tag{43}$$

Multiplying out the terms:

$$\begin{aligned} &(Y - X\hat{\beta})^T X = 0 \\ \Leftrightarrow &Y^T X - \hat{\beta}^T X^T X = 0 \\ \Leftrightarrow &Y^T X = \hat{\beta}^T X^T X \\ \Leftrightarrow &(Y^T X)^T = (\hat{\beta}^T X^T X)^T \\ \Leftrightarrow &X^T Y = XX^T\hat{\beta} \end{aligned} \tag{44}$$

This gives us the important result: $\hat{\beta} = (XX^T)^{-1}X^T Y$.
X is of full rank, therefore $X^T X$ is positive definite symmetric $p \times p$ and invertible.
[to-do: summarize ch 6 of Lay in matrix algebra notes]

Statistical properties of LSEs

$$E[\hat{\beta}] = (XX^T)^{-1}X^T Y = (XX^T)^{-1}X^T X\beta = \beta \tag{45}$$

$$\begin{aligned} Cov(\hat{\beta}) = Var(\hat{\beta}) = Var([(XX^T)^{-1}X^T]Y) &= [(XX^T)^{-1}X^T]\sigma^2 I[(XX^T)^{-1}X^T]^T \\ &= [(XX^T)^{-1}X^T]\sigma^2 IX[(XX^T)^{-1}]^T \\ &= \sigma^2(XX^T)^{-1}X^T X[(XX^T)^{-1}]^T \\ &= \sigma^2(XX^T)^{-1}X^T X(XX^T)^{-1} \\ &= \sigma^2(XX^T)^{-1} \end{aligned} \tag{46}$$

Note that $[(XX^T)^{-1}]^T = (XX^T)^{-1}$ because $(XX^T)^{-1}$ is symmetric.

Overparameterization and contrast coding

Suppose there are three groups, so our model is

$$Y_i = \alpha + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i \tag{47}$$

A typical thing we do is **dummy coding**:

Let μ_i be the i-th group. Taking expectations:

$$\begin{aligned} \mu_1 &= \alpha + \gamma_1 \times 1 + \gamma_2 \times 0 = \alpha + \gamma_1 \\ \mu_2 &= \alpha + \gamma_1 \times 0 + \gamma_2 \times 1 = \alpha + \gamma_2 \\ \mu_3 &= \alpha + \gamma_1 \times 0 + \gamma_2 \times 0 = \alpha \end{aligned} \tag{48}$$

There are three parameters, and three equations:

$$\mu_1 = \alpha + \gamma_1 \quad \mu_2 = \alpha + \gamma_2 \quad \mu_3 = \alpha \tag{49}$$

Overparameterization occurs in the following situation: Let j index the groups. Then:

$$y_{ij} = \mu + \alpha_j + \epsilon_{ij} \tag{50}$$

Taking expectations: $\mu_j = \mu + \alpha_j$. Now we have the equations

$$\begin{aligned}\mu_1 =& \mu + \alpha_1 \\ \mu_2 =& \mu + \alpha_2 \\ \mu_3 =& \mu + \alpha_3\end{aligned} \tag{51}$$

There are four parameters, and three equations:
These equations can't be solved (don't have a unique solution). The model is said to be overparameterized or underdetermined.
The solution is to place a restriction on the parameters: express one parameter in terms of the others. An example is sum contrast coding: if there are p parameters, then, just stipulate that $\alpha_1 + \cdots + \alpha_n = \sum_{i=1}^{p} \alpha_i = 0$.
Another example is deviation regressors or **effects coding**: Let m be the maximum number of groups. For each of the j groups,

$$D_j = \begin{cases} 1 & groupj \\ -1 & groupm \\ 0 & \text{otherwise} \end{cases} \tag{52}$$

Here, we have constrained the parameters so that $\sum \alpha_i = 0$, i.e., $\alpha_3 = -\alpha_1 - \alpha_2$. Now we have three equations and three parameters, which has a unique solution.

$$\begin{aligned}\mu_1 =& \mu + \alpha_1 \\ \mu_2 =& \mu + \alpha_2 \\ \mu_3 =& \mu + \alpha_3 = \mu - \alpha_1 - \alpha_2\end{aligned} \tag{53}$$

Note that two of the parameters are correlated when we use effects coding:
Example:

```
> m<-matrix(c(c(rep(1,9)),
            c(rep(1,3),rep(0,3),rep(-1,3)),
            c(rep(0,3),rep(1,3),rep(-1,3))),byrow=FALSE,nrow=9)
> cov(m)

     [,1]  [,2]  [,3]
[1,]    0 0.000 0.000
[2,]    0 0.750 0.375
[3,]    0 0.375 0.750
```

Polynomial regression

$$E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{54}$$

If the design matrix is:

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ 1 & x_{13} & x_{23} \end{pmatrix} \tag{55}$$

This matrix full rank (i.e., non-singular) iff there exists no linear relationship like:

$$\lambda_0 + \lambda_1 x_{1j} + \lambda_2 x_{2j} = 0 \quad \text{for } j = 1, 2, 3 \tag{56}$$

λ_i are not all zero.
If X has full rank, the three points are not collinear. Example of collinearity: each triple of rows is for each group α_i.

$$X = \left(\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{array}\right) \tag{57}$$

The columns c2-c4 in this matrix are collinear: c2-c3-c4=0. Therefore the matrix is not full rank, therefore not invertible.
This motivates the use of the corner-point constraint (dummy coding) or effects coding (depending on what the research question is). Now, if we remove the final column, we have full rank and an invertible matrix. See below:

$$\left(\begin{array}{ccc} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ \hline 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{array}\right) \tag{58}$$

The first two parameters as α_1 and α_2.
Example: Let $y = \beta_0 + \beta_1 x$. How to make the design matrix orthogonal? Centering achieves that. (to-do: discussion based on Draper and Smith)

Orthogonality

Let

$$\beta_p = \begin{pmatrix} \gamma_q \\ \delta p - q \end{pmatrix} \quad X_{n\times p} = V_{n\times q} W_{n\times(p-q)} \tag{59}$$

V and W are orthogonal, i.e., $V^T W = 0$.
Consequence of orthogonality:

$$Cov(\hat{\beta}) = \sigma^2 \begin{pmatrix} (V^T V)^{-1} & 0 \\ 0 & (W^T W)^{-1} \end{pmatrix} \tag{60}$$

γ and δ are independent in the statistical sense. Excluding δ will not affect estimate of γ's sampling distribution.

Prediction

If we want to predict a new value given a new data point x_0.

$$E[Y[x_0]] = y(x_0) = f(x_0)^T \beta \tag{61}$$

The estimate $\hat{y}(x_0) = f(x_0)^T \hat{\beta}$ is unbiased. The variance is

$$\begin{aligned} Var(\hat{y}(x_0)) =& f(x_0)^T Cov\hat{\beta} f(x_0) \\ =& \sigma^2 f(x_0)^T f(x_0) \end{aligned} \tag{62}$$

So, variance (accuracy) depends on depends on X and x_0.

Example: Consider simple linear regression. We know (LinearModelsSummary.pdf) that

$$(X^T X)^{-1} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum x^2 & -\sum x \\ -\sum x & n \end{pmatrix} \tag{63}$$

$$\begin{aligned} Var(\hat{y}(x_0)) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) =& \frac{\sigma^2}{nS_{xx}} \begin{pmatrix} 1 & x_0 \end{pmatrix} \begin{pmatrix} \sum x^2 & -\sum x \\ -\sum x & n \end{pmatrix} \begin{pmatrix} 1 \\ x_0 \end{pmatrix} \\ =& \sigma^2 (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}) \end{aligned} \tag{64}$$

where $\bar{x} = \frac{\sum x}{n}$, and $S_{xx} = (\sum x_2) - n\bar{x}^2$.
One should avoid predicting outside the region containing the design points, because we usually have no idea whether the model holds outside the range of the design points.

Standardized Information Matrix (SIM)

$M = \frac{1}{n} X^T X$.
Standard variance at point x:

$$\begin{aligned} d(x) =& n f(x)^T (X^T X)^{-1} f(x) \\ =& f(x)^T M^{-1} f(x) \end{aligned} \tag{65}$$

This is a very important equation for the next section.
The SIM remains unchanged for different sample sizes.