STAT 601 Lec 2

Lectures: MW 2:26–3:45 PM, 1313 Sterling
Disc 321: F 9:30–10:45 AM, 382 Van Hise
Disc 322: F 8:00–9:15 AM, 114 Social Work

Instructor: Wei-Yin Loh
Office hours: M 4:00–5:00 PM, 1217C MSC

TA: Siyu Wang
In-person office hours: T 2:30–3:30 PM, 6173C MSC
Zoom office hours: R 11 AM–12:00 PM

Class email list: stat601-2-f23@g-groups.wisc.edu

Topics

1. Linear regression
    - Concepts
    - Theory
    - Practice
2. Generalized linear models — logistic, Poisson, proportional hazards
3. Design and analysis of statistical experiments
    - One and two-way factorial designs
    - Randomized block designs, including Latin square designs
    - Incomplete block designs

Recommended texts

1. Rencher (2000), *Linear Models in Statistics*, Wiley
2. Box, Hunter & Hunter (2005), *Statistics for Experimenters*, 1st or 2nd edition, Wiley
3. Mead, Gilmour & Mead (2012), *Statistical Principles for the Design of Experiments*, Cambridge University Press

Grading

Homework. 10% of average z-score, excluding 2 lowest

Quizzes. 10% of average z-score

Midterm (Wed Oct 25). 30% of z-score

Final (Fri Dec 15). 50% of z-score

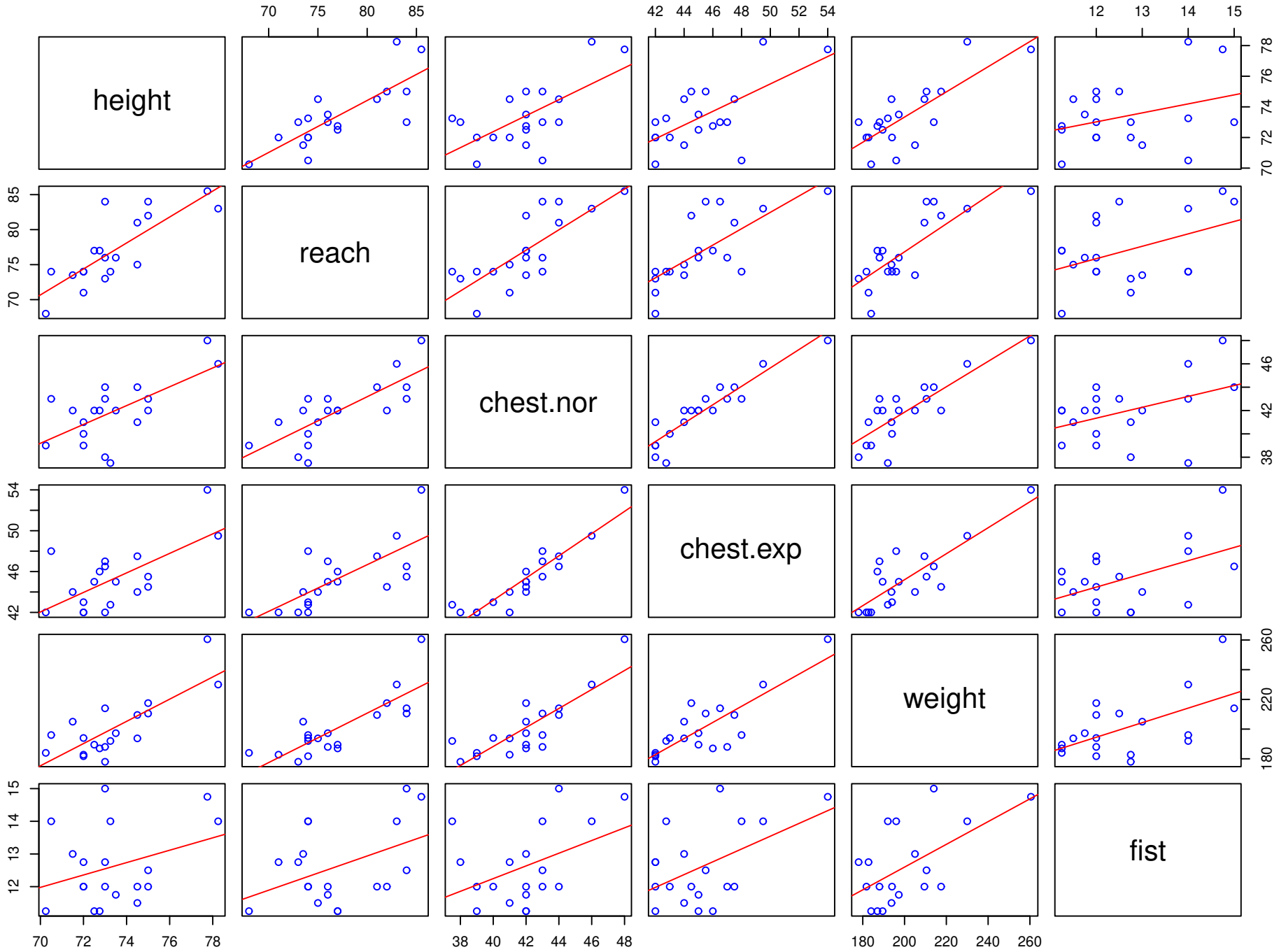
Linear regression example:
Height, reach and chest (normal & expanded), weight and fist measurements of 19 heavyweight boxing champions

https://www.boxingscene.com/forums/showthread.php?t=286210

Name	height	reach	chest.nor	chest.exp	weight	fist
Ali	75	84	43	45.5	210.5	12.5
Baer	74.5	81	44	47.5	209.5	12
Braddock	74.5	75	41	44	193.75	11.5
Carnera	77.75	85.5	48	54	260.5	14.75
Charles	72	74	39	42	181.75	12
Corbett	73	73	38	42	178	12.75
Dempsey	72.75	77	42	46	187	11.25
Foreman	75	82	42	44.5	217.5	12
Frazier	71.5	73.5	42	44	205	13
Johnson	73.25	74	37.5	42.75	192	14
Liston	73	84	44	46.5	214	15
Louis	73.5	76	42	45	197.25	11.75
Marciano	70.25	68	39	42	184	11.25
Patterson	72	71	41	42	182.75	12.75
Schmeling	73	76	43	47	188	12
Sullivan	70.5	74	43	48	196	14
Tunney	72.5	77	42	45	189.5	11.25
Walcott	72	74	40	43	194	12
Willard	78.25	83	46	49.5	230	14

R code for pairwise plots with regression lines

```
> z <- read.csv("boxers.csv")
> names(z)
[1] "Name"  "height"  "reach"  "chest.nor" "chest.exp" "weight"
[7] "fist"
> pairs(z[,-1])  ## plain
> pairs(z[,-1],
      panel=function(x,y,...){
        points(x,y,...)
        abline(lm(y ~ x),col="red")
      },col="blue")
```


height
reach
chest.nor
chest.exp
weight
fist
70 75 80 85
42 44 46 48 50 52 54
12 13 14 15
70 72 74 76 78
70 75 80 85
38 42 46
42 46 50 54
180 220 260
12 13 14 15
70 72 74 76 78
38 40 42 44 46 48
180 200 220 240 260

Correlations

```
> cor(z[,-1])
```

	height	reach	chest.nor	chest.exp	weight	fist
height	1.00	0.79	0.65	0.65	0.78	0.33
reach	0.79	1.00	0.78	0.73	0.82	0.43
chest.nor	0.65	0.78	1.00	0.93	0.83	0.43
chest.exp	0.65	0.73	0.93	1.00	0.83	0.50
weight	0.78	0.82	0.83	0.83	1.00	0.58
fist	0.33	0.43	0.43	0.50	0.58	1.00

Least squares linear regression

- Given data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, let $\bar{x} = n^{-1} \sum_i x_i$, $\bar{y} = n^{-1} \sum_i y_i$
- Find the line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ that minimizes the sum of squared errors

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

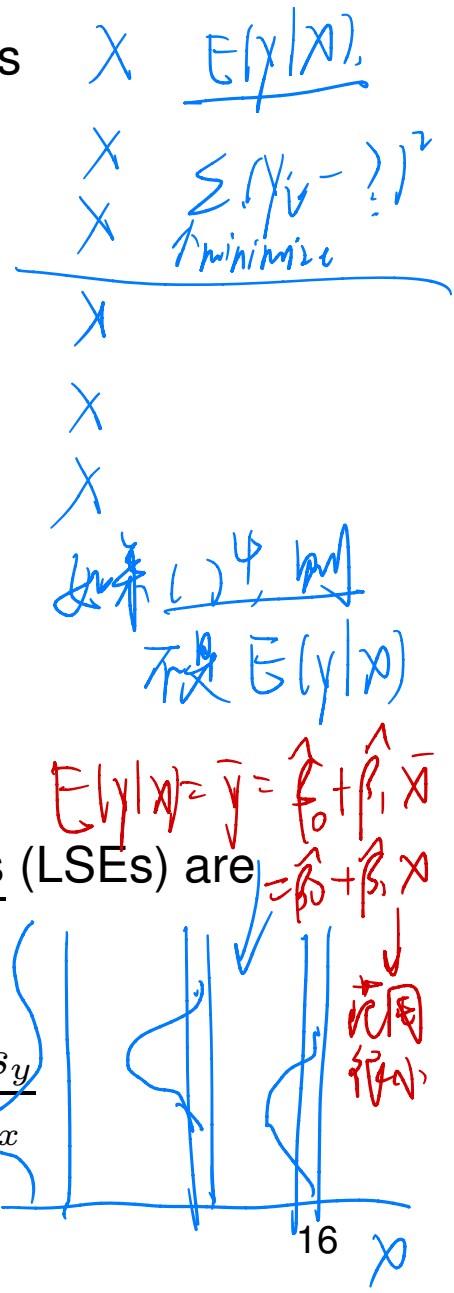
sufficient statistic

- Solution: differentiate SSE w.r.t. $\hat{\beta}_0$ and $\hat{\beta}_1$ and solve for them

$$\begin{aligned} \partial \text{SSE}/\partial \hat{\beta}_0 &= -2 \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \partial \text{SSE}/\partial \hat{\beta}_1 &= -2 \sum_i x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{aligned}$$

- Let $s_x^2 = (n-1)^{-1} \sum_i (x_i - \bar{x})^2$, $s_y^2 = (n-1)^{-1} \sum_i (y_i - \bar{y})^2$, and $r = (n-1)^{-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})/(s_x s_y)$. Least-squares estimates (LSEs) are

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i x_i^2 - n\bar{x}^2} = \frac{r s_y}{s_x} \end{aligned}$$



```
> model.nor <- lm(height ~ chest.nor, data=z)
> summary(model.nor)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.5251     6.2310    8.269 2.32e-07 ***
chest.nor     0.5214     0.1484    3.514  0.00266 **
Residual standard error: 1.646 on 17 degrees of freedom
Multiple R-squared:  0.4208,        Adjusted R-squared:  0.3867
F-statistic: 12.35 on 1 and 17 DF,  p-value: 0.002661

> model.exp <- lm(height ~ chest.exp, data=z)
> summary(model.exp)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.1082     5.6950   9.325 4.27e-08 ***
chest.exp     0.4478     0.1255   3.568  0.00237 **
Residual standard error: 1.636 on 17 degrees of freedom
Multiple R-squared:  0.4281,        Adjusted R-squared:  0.3945
F-statistic: 12.73 on 1 and 17 DF,  p-value: 0.00237
```

为什么系数会变，怎么会不变

↓ uncorrelated.

系数

statistically significant from 0

$H_0: \beta_1 = 0$

p-value = P(A statistic is as large or larger than the observed value)

$H_0: \beta_1 = 0$, $\hat{\beta}_1 = 0.5214$ or larger

p-value = $P_{H_0}(|\hat{\beta}_1| \geq 0.5214)$

Assumptions 1. random sample

2. linear model

Where do the p-values come from?

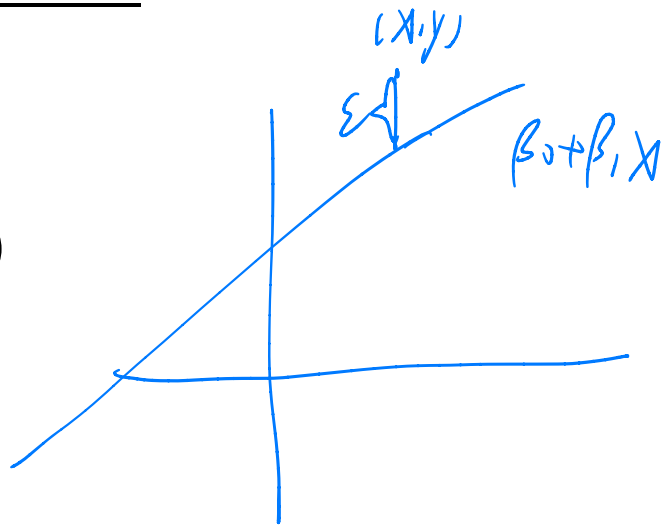
- Assume that, for each x,

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

  where ϵ is independent, identically distributed normal (IIDN) with mean 0 and variance σ^2

- Equivalent to writing

$$y \sim N(\beta_0 + \beta_1 x, \sigma^2)$$



Hypothesis testing terms

Null hypothesis $H_0: \mu = \mu_0$

Alternative hypothesis $H_1: \mu \neq \mu_0$

Significance level $\alpha = P_{\mu_0}(\text{Type I error})$

Test statistic

Null distribution of test statistic

Rejection region

Critical value

Type I error : H_0 is true but reject H_0

Type II error

Power

P-value

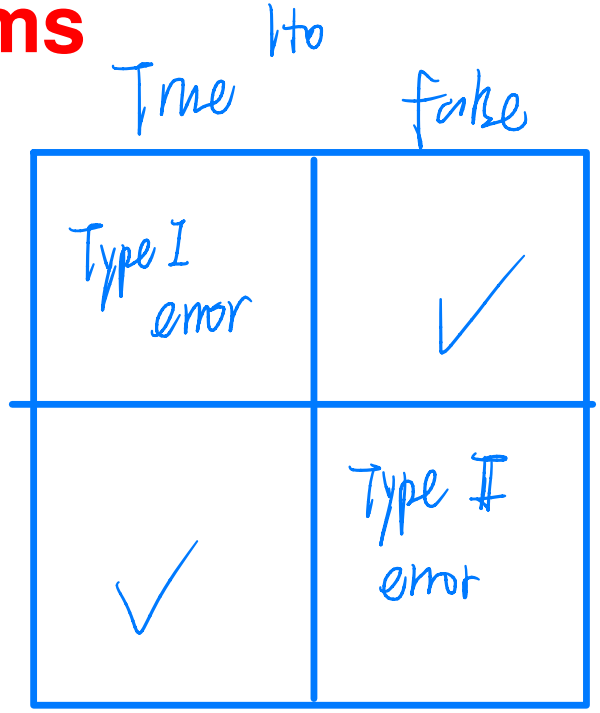


Illustration: one-sample t-test

- Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$
- Let $\bar{X} = n^{-1} \sum_i X_i$ and $S^2 = (n-1)^{-1} \sum_i (X_i - \bar{X})^2$
- Then $\bar{X} \sim N(\mu, n^{-1}\sigma^2)$, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, and $\bar{X}$ is independent of S^2

Null hypothesis. $H_0 : \mu = \mu_0$

Alternative hypothesis. $H_1 : \mu \neq \mu_0$

Significance level. α (specified)

Test statistic. $T = \sqrt{n}(\bar{X} - \mu_0)/S$

Null distribution of statistic. t_{n-1} (t distribution with $n-1$ df)

Rejection region. $|T| > t_{n-1,\alpha/2}$

Critical value. $t_{n-1,\alpha/2}$

P(**Type I error**).

$$P_{\mu_0}(\text{Reject } H_0) = P_{\mu_0}(|T| > t_{n-1,\alpha/2}) = P(|t_{n-1}| > t_{n-1,\alpha/2}) = \alpha$$

P(**Type II error**). For any $\mu \neq \mu_0$ and as $n \to \infty$,

$$\begin{aligned}
P_\mu(\text{Type II error}) &= P_\mu(\text{Do not reject } H_0) \\
&= P_\mu(|T| \leq t_{n-1,\alpha/2}) \\
&= P_\mu(-t_{n-1,\alpha/2} \leq \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \leq t_{n-1,\alpha/2}) \\
&= P_\mu\left(-t_{n-1,\alpha/2} - \frac{\mu - \mu_0}{S/\sqrt{n}} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1,\alpha/2} - \frac{\mu - \mu_0}{S/\sqrt{n}}\right) \\
&\approx P\left(-t_{n-1,\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \leq t_{n-1} \leq t_{n-1,\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) \to 0
\end{aligned}$$

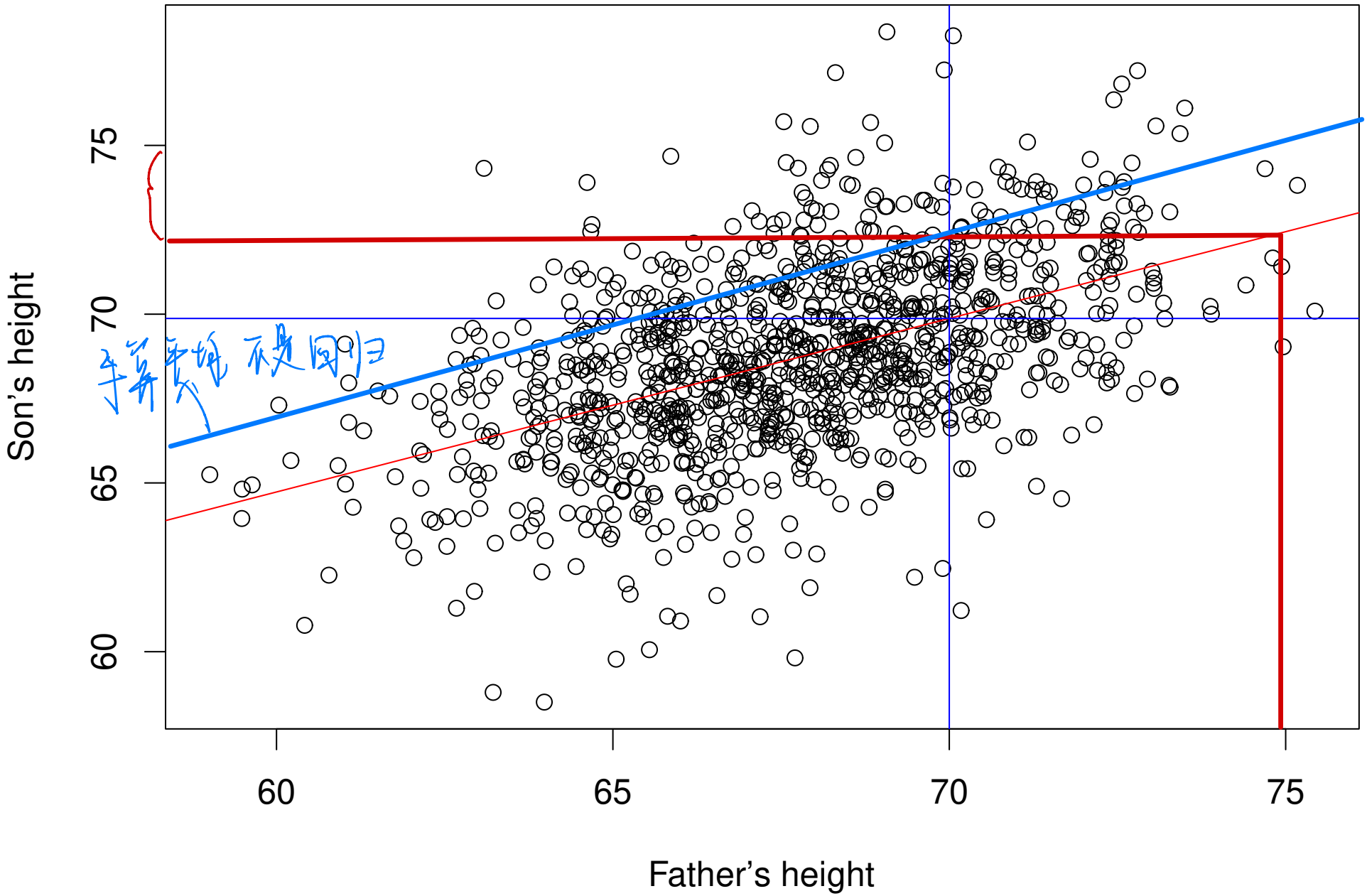
Since $P(-t_{n-1,\alpha/2} - \delta \leq t_{n-1} \leq t_{n-1,\alpha/2} - \delta) < 1 - \alpha$ for every $\delta \neq 0$, it follows that $P_\mu(\text{Type II error}) < 1 - \alpha$ for every $\mu \neq \mu_0$

Power. $P_\mu(\text{Reject } H_0) = 1 - P_\mu(\text{Type II error}) \to 1$ for $\mu \neq \mu_0$ ($\geq \alpha$ for all μ)

P-value. $P_{\mu_0}(|T| > T_{\text{obs}}) = P(|t_{n-1}| > \sqrt{n}(\bar{X}_{\text{obs}} - \mu_0)/S_{\text{obs}})$

Plotting Pearson's father-son data (requires UsingR and HistData packages)

```
> library("UsingR")
> data(package="UsingR") ## print names of datasets in library
> df <- as.data.frame(father.son)
> names(df)
[1] "fheight" "sheight"
> plot(df$sheight ~ df$fheight,xlab="Father's height",
     ylab="Son's height")
> model <- lm(df$sheight ~ df$fheight)
> abline(model,col="red")
> abline(v=70,col="blue")
> mean.son.height <- model$coef[1]+model$coef[2]*70
> abline(h=mean.son.height,col="blue")
```


Son's height
75
70
65
60
60
65
70
75
Father's height

Maximum likelihood estimates (MLEs)

- Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$
- Joint density of $\mathbf{y}$ given $\mathbf{x}$ is

$$f(\mathbf{y} \mid \mathbf{x}, \beta_0, \beta_1, \sigma) = (2\pi\sigma^2)^{-n/2} \exp[-(2\sigma^2)^{-1} \sum_i \{y_i - (\beta_0 + \beta_1 x_i)\}^2]$$

- Likelihood function is

$$\ell(\beta_0, \beta_1, \sigma \mid \mathbf{x}, \mathbf{y}) = (2\pi\sigma^2)^{-n/2} \exp[-(2\sigma^2)^{-1} \sum_i \{y_i - (\beta_0 + \beta_1 x_i)\}^2]$$

- Log-likelihood function is ↙ do not depend on σ

$$\log \ell(\beta_0, \beta_1, \sigma) = -(n/2) \log \sigma^2 - (2\sigma^2)^{-1} \sum_i \{y_i - (\beta_0 + \beta_1 x_i)\}^2 + \text{constant}$$

- For any σ^2, $\log \ell$ is maximum when $\sum_i \{y_i - (\beta_0 + \beta_1 x_i)\}^2$ is minimum
- Hence MLEs $(\hat{\beta}_0, \hat{\beta}_1)$ are same as LSEs

Maximum likelihood estimate of σ^2

- MLE $\hat{\sigma}^2$ is the solution of

$$\partial \log \ell / \partial \sigma^2 = -n/(2\sigma^2) + (2\sigma^4)^{-1} \sum_i \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\}^2 = 0$$

- Therefore

$$\hat{\sigma}^2 = n^{-1} \sum_i \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\}^2 = \text{SSE}/n$$

not unbiased

$S^2 = \frac{SSE}{n-2}$ unbiased

↑ 2 parameter

$S \to$ SD (sample)

$\sigma \to$ SD (population)

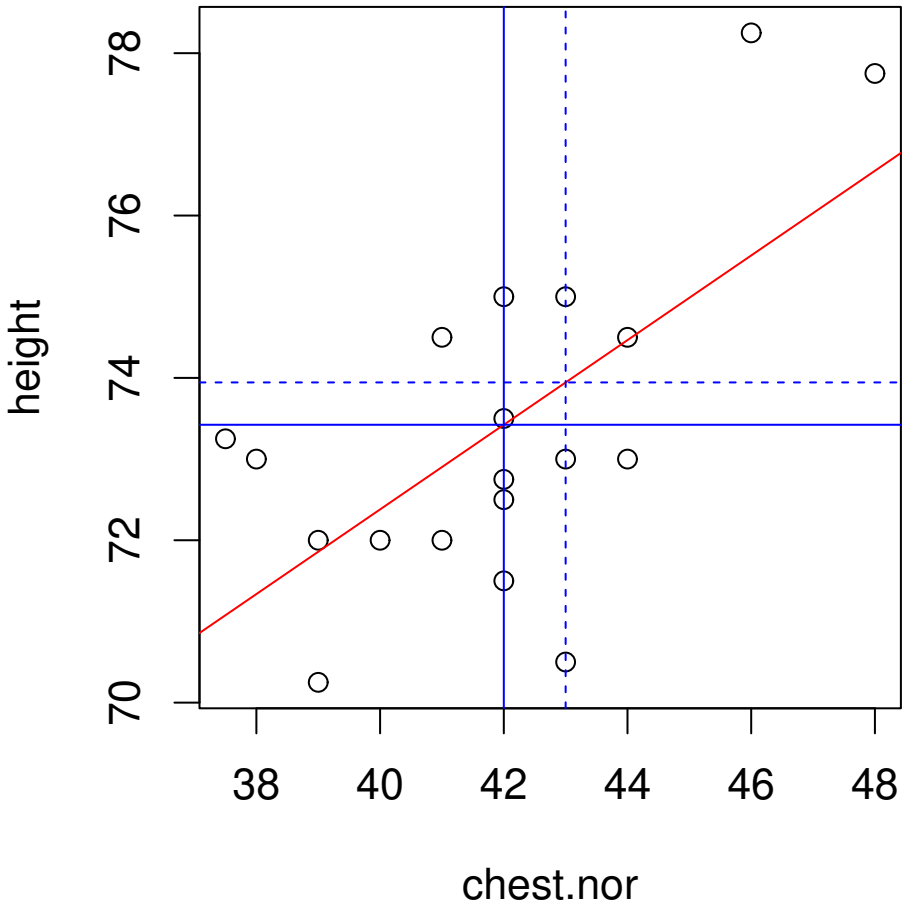
$E S^2 = \sigma^2 \not\Rightarrow E(S) = \sigma$

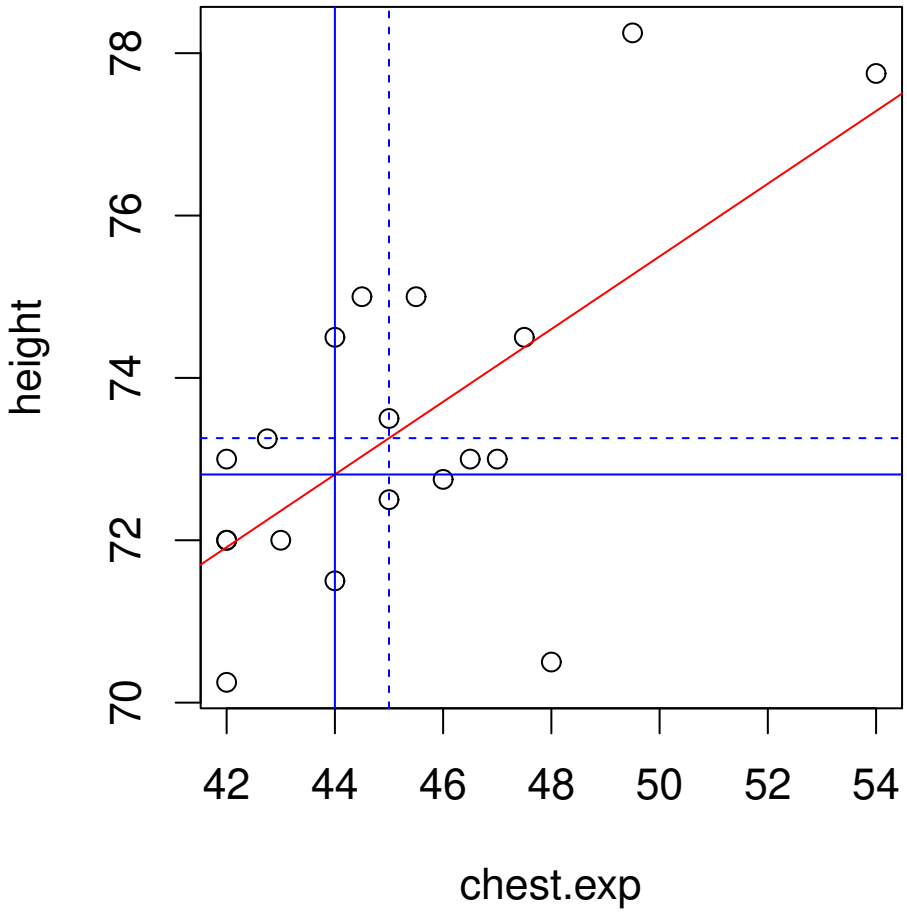
How to interpret regression coefficients and how to select among models

height vs chest.nor & height vs chest.exp

```
> model.nor <- lm(height ~ chest.nor, data=z)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.5251     6.2310   8.269 2.32e-07 ***
chest.nor     0.5214     0.1484   3.514  0.00266 **
Residual standard error: 1.646 on 17 degrees of freedom
Multiple R-squared:  0.4208,        Adjusted R-squared:  0.3867
F-statistic: 12.35 on 1 and 17 DF,  p-value: 0.002661
```

```
> model.exp <- lm(height ~ chest.exp, data=z)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.1082     5.6950   9.325 4.27e-08 ***
chest.exp     0.4478     0.1255   3.568  0.00237 **
Residual standard error: 1.636 on 17 degrees of freedom
Multiple R-squared:  0.4281,        Adjusted R-squared:  0.3945
F-statistic: 12.73 on 1 and 17 DF,  p-value: 0.00237
```


height
70
72
74
76
78
38
40
42
44
46
48
chest.nor


height
70
72
74
76
78
42
44
46
48
50
52
54
chest.exp

height vs chest.nor & chest.exp

```
> model.both <- lm(height ~ chest.nor+chest.exp, data=z)
> summary(model.both)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.6175     6.3120   8.178 4.17e-07 ***
chest.nor     0.2407     0.4002   0.601    0.556
chest.exp     0.2578     0.3407   0.757    0.460
Residual standard error: 1.667 on 16 degrees of freedom
Multiple R-squared:  0.4408,        Adjusted R-squared:  0.3709
F-statistic: 6.306 on 2 and 16 DF,  p-value: 0.009564
```

檢驗兩個模型比較

Y_1 = intercept + chest.nor + chest.exp

Y_2 = intercept + chest.exp

conditional 的 p 值

$H_0: \beta_1 = 0$

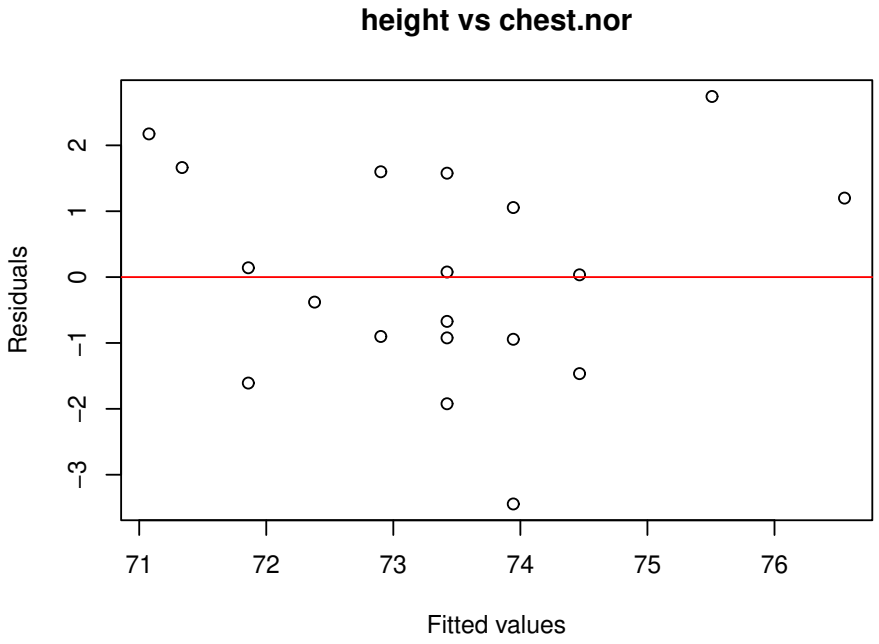
$H_0: \beta_2 = 0$

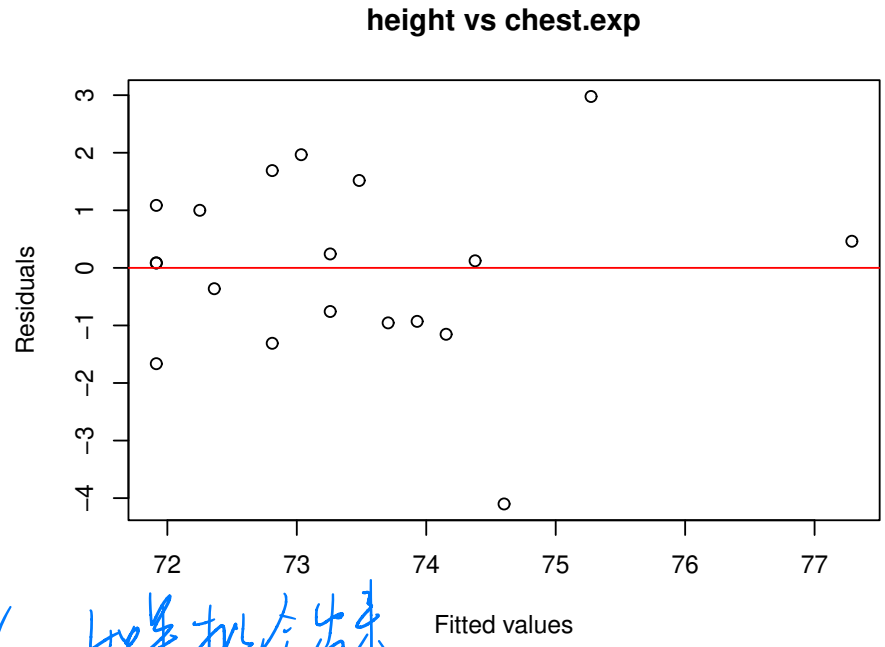
↓ 回歸平方和 / 殘差平方和

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

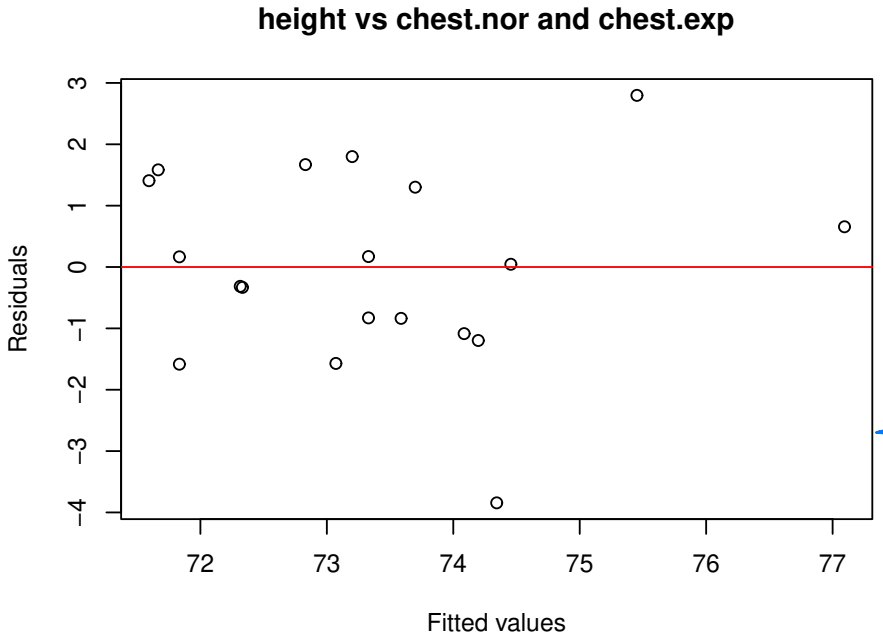
Diagnostic plots to check for violations of model structure and IIDN assumptions

1. Plot of residuals versus fitted values—may indicate nonconstant variance
2. Normal probability plot of residuals—may indicate nonnormal ϵ distribution





如果拟合出来是曲线，需要加平方项

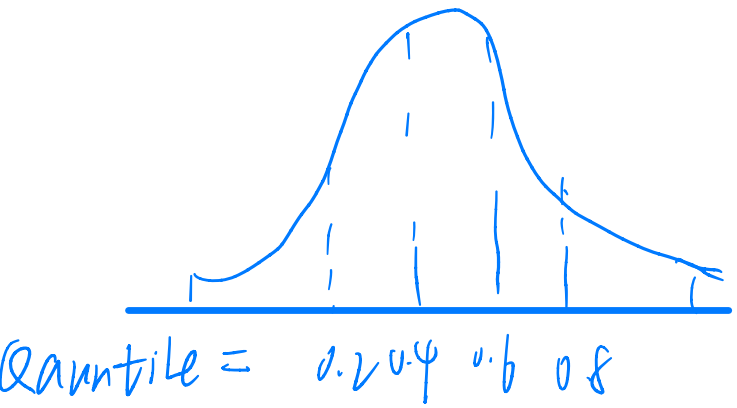


模型好，R^2 变大

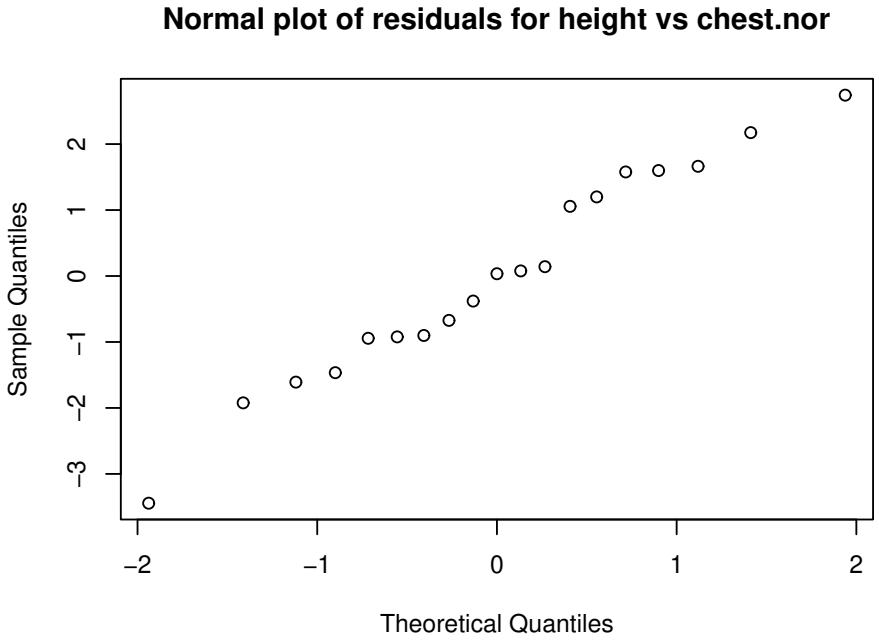
residual 加起来是0。

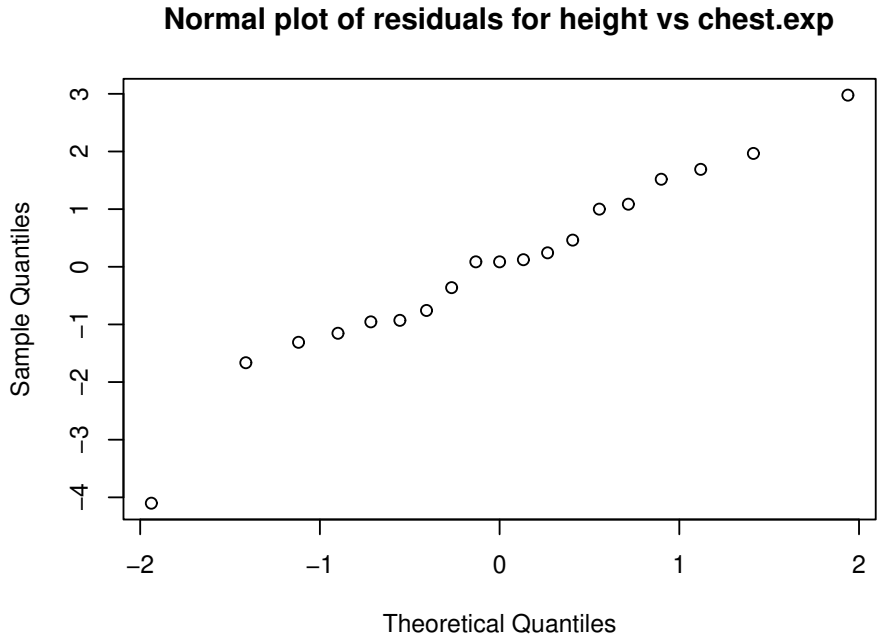
How is a normal probability plot constructed?

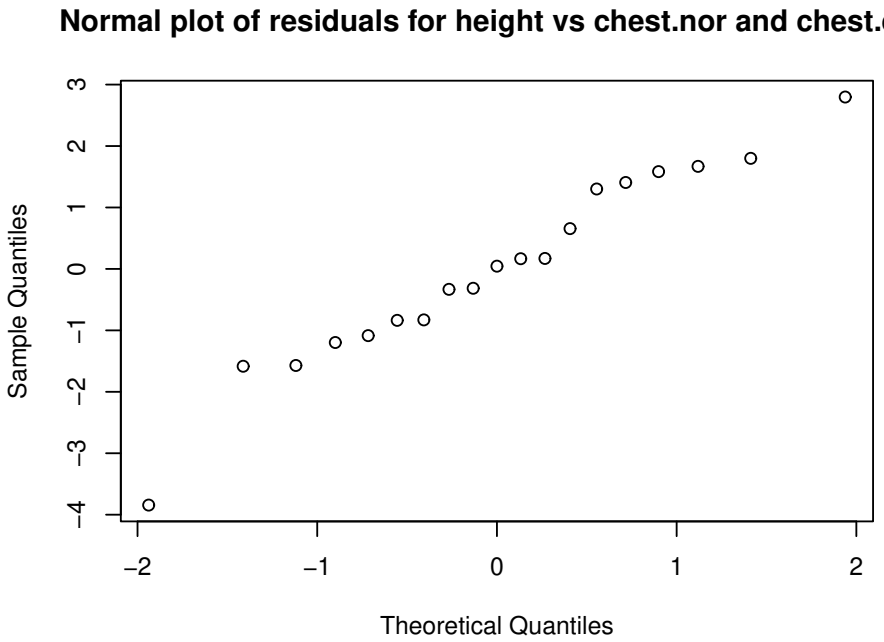
R function: qqnorm(x)



every region
same 面积.



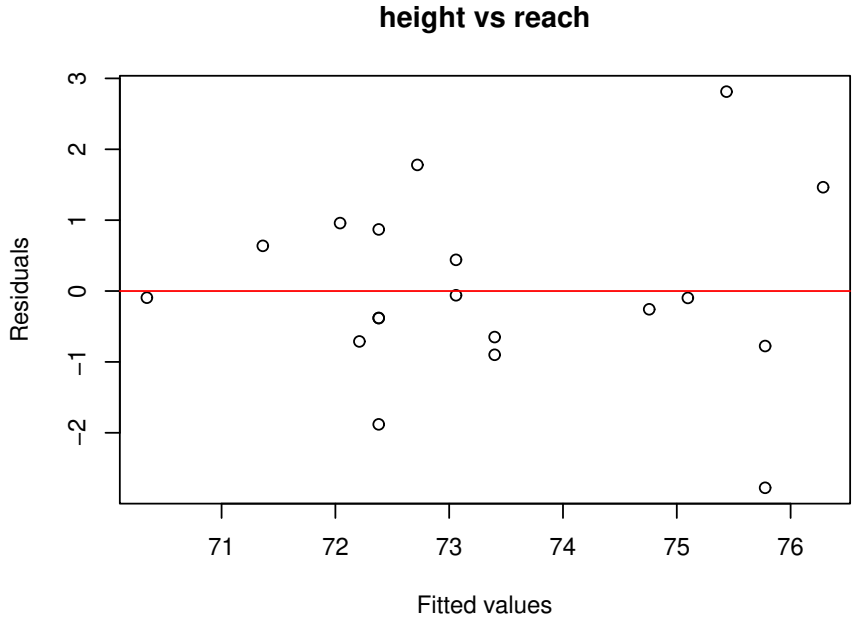


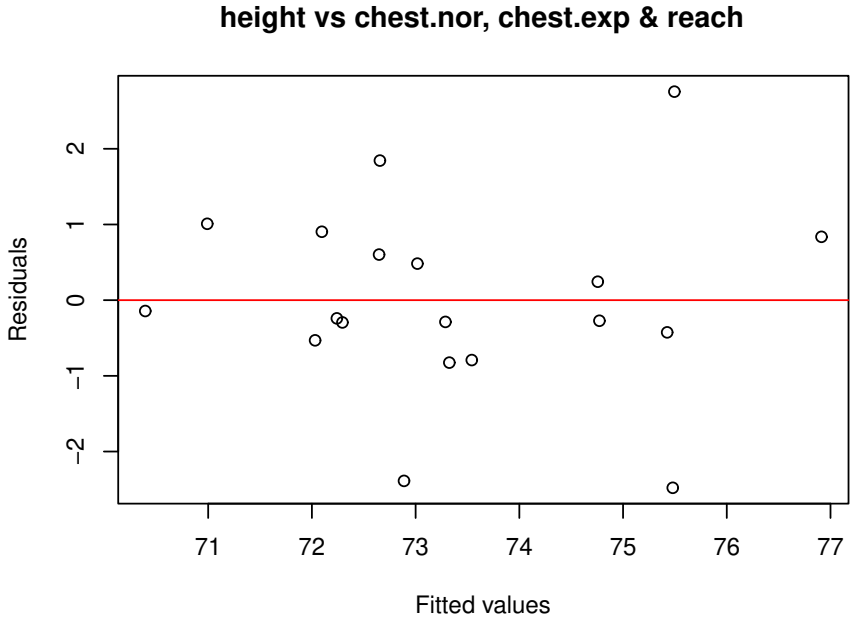


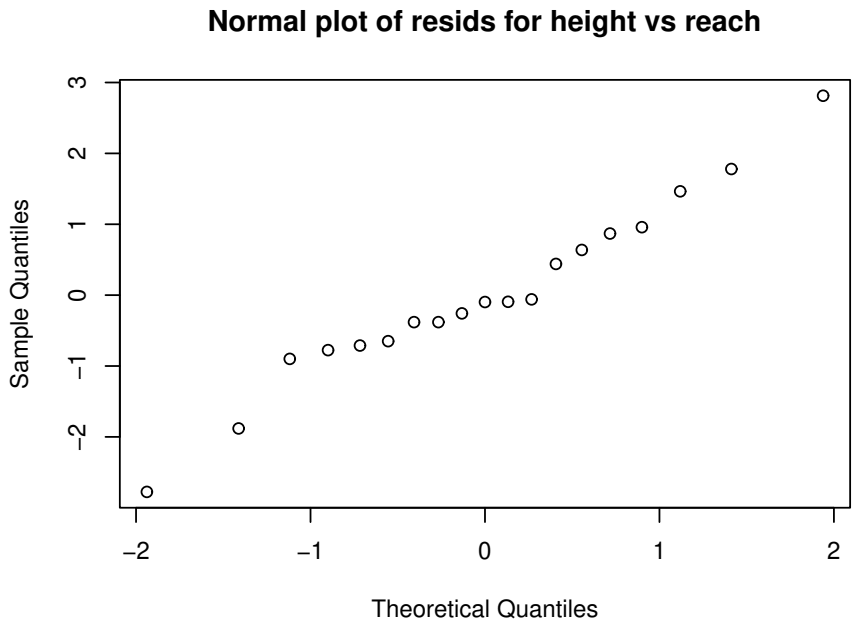
测试正态分布，不能判断是否显著

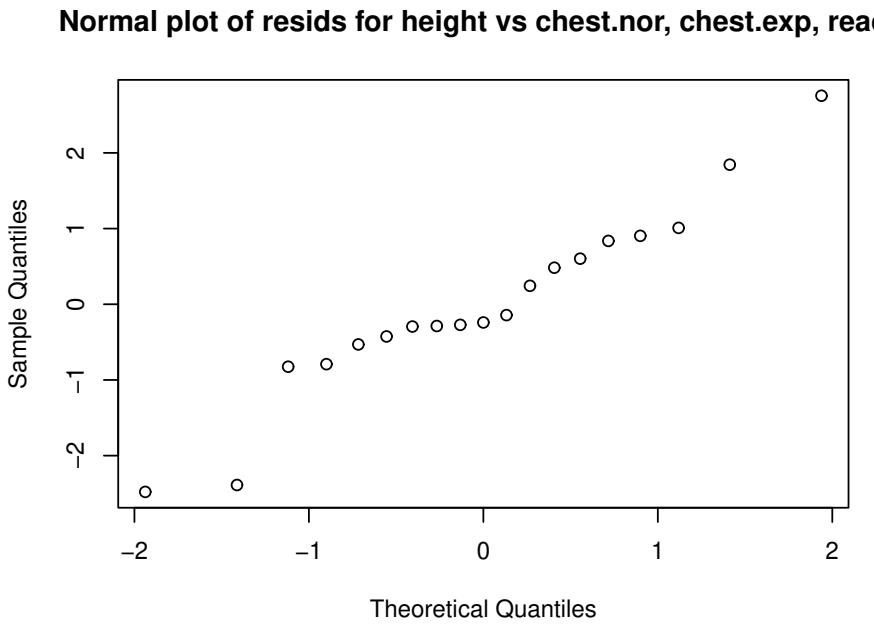
```
> model.reach <- lm(height ~ reach, data=z)
> summary(model.reach)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.25588    4.86946   9.705 2.40e-08 ***
reach        0.33953    0.06316   5.376 5.03e-05 ***
Residual standard error: 1.316 on 17 degrees of freedom
Multiple R-squared:  0.6296,        Adjusted R-squared:  0.6078
F-statistic:  28.9 on 1 and 17 DF,  p-value: 5.033e-05

model.3 <- lm(height ~ chest.nor+chest.exp+reach, data=z)
> summary(model.3)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  46.7189     5.4423   8.584 3.58e-07 ***
chest.nor    -0.1630     0.3559  -0.458  0.65352
chest.exp     0.2166     0.2802   0.773  0.45153
reach         0.3079     0.1042   2.954  0.00986 **
Residual standard error: 1.369 on 15 degrees of freedom
Multiple R-squared:  0.6464,        Adjusted R-squared:  0.5757
F-statistic: 9.142 on 3 and 15 DF,  p-value: 0.001104
```


height vs reach
Residuals
Fitted values
71
72
73
74
75
76
−2
−1
0
1
2
3


height vs chest.nor, chest.exp & reach
Residuals
Fitted values
71
72
73
74
75
76
77
−2
−1
0
1
2


Normal plot of resids for height vs reach
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
3


Normal plot of resids for height vs chest.nor, chest.exp, reach
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2

Partial residuals

```
> lm(height ~ chest.nor+chest.exp+reach, data=z)
(Intercept)    chest.nor    chest.exp        reach
    46.7189      -0.1630       0.2166       0.3079

> model1 <- lm(height ~ chest.exp+reach, data=z)
> model2 <- lm(chest.nor ~ chest.exp+reach, data=z)
> lm(model1$res ~ model2$res)
(Intercept)   model2$res
  2.339e-17   -1.630e-01
```

Homework 1
(due 11:59 PM Tue Sep 26 in Canvas)

Compute the leave-one-out cross-validation estimates of root mean squared prediction error (RMSPE) for the model

```
height ~ chest.nor + chest.exp + reach + weight + fist
```

and all its submodels

Which model has the lowest RMSPE?

Note: Let $\mathcal{D}$ denote the full data set. Let (x_i, y_i) denote the ith observation and $\mathcal{D}_{-i}$ the data subset without (x_i, y_i). Let $\mathcal{M}_{-i}$ denote the model fitted to $\mathcal{D}_{-i}$ and $\hat{y}_i^{(-i)}$ be the predicted value of y_i based on $\mathcal{M}_{-i}$. The leave-one-out RMSPE is

$$\sqrt{n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i^{(-i)})^2}$$

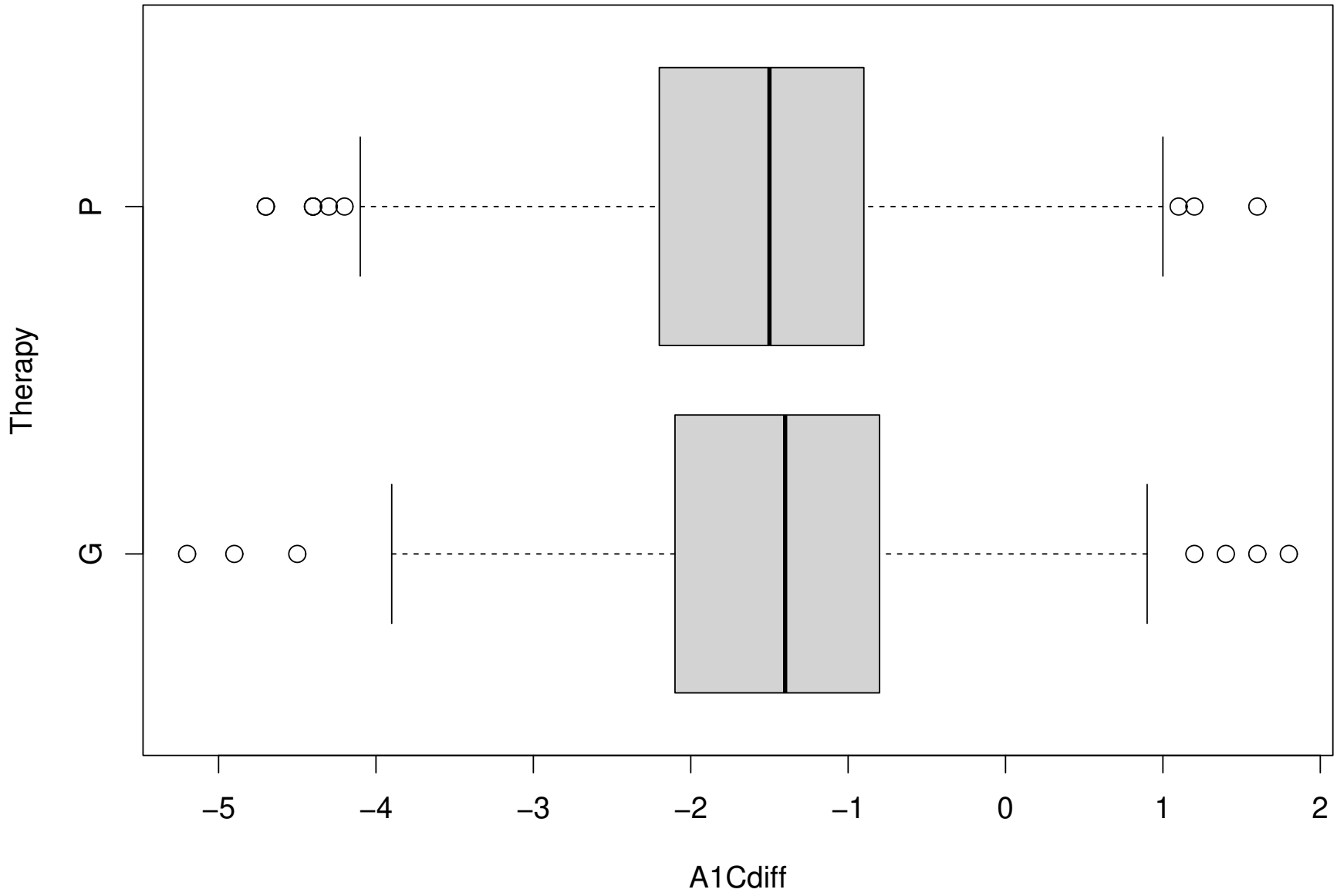
Diabetes data

- 802 subjects from a multi-center, randomized double-blind clinical trial
- Subjects are 35–75 years old with HbA1c (blood sugar level) between 7.5% and 11.0%, for whom diet was prescribed for at least 3 months (non-diabetics have HbA1c below 5.7%; prediabetics 5.7–6.4%)
- Subjects randomized to a 52-week treatment period of drug G *(Gliclazide)* or P *(Pioglitazone)*
- *Gliclazide* increases amount of insulin produced by the pancreas
- *Pioglitazone* improves how body uses insulin ("insulin sensitizer")
- 23 baseline (time 0) variables measured for each subject as well as their HbA1c at 52 weeks
- Interest is in whether there is a statistically significant difference between the effects of the two drugs on reduction of HbA1c

Diabetes variables

Variable	Meaning	Variable	Meaning
HDL	high-density cholesterol	Age	age in years
LDL	low-density cholesterol	Weight	weight in kg
TotalChol	total cholesterol	BMI	body mass index
Triglycerides	type of fat (lipid)	Waist	waist in cm
Creatinine	kidney function test	A1CBase	HbA1C at week 0
GGT	gamma-glutamyl transferase	HomaS	measure of insulin sensitivity
ALT	alanine aminotransferase	HomaIR	measure of insulin resistance
AST	aspartate aminotransferase	HomaB	measure of beta-cell function
FastInsulin	fasting insulin	Diastolic	Diastolic blood pressure
C-peptide	connecting peptide	Systolic	Systolic blood pressure
DDuration	Diabetes duration	Pulse	pulse rate
FastBG	Fasting blood glucose	A1C52	HbA1C at week 52

Boxplots of A1Cdiff = A1C52 - A1CBase



Two-sample t-tes

Proof

- $\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$, $s_{\widehat{\beta}} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$
- $t = \frac{\bar{y}_A-\bar{y}_B}{\hat{\sigma}\cdot\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}}$

Interpretation

Two tests are testing the same hypothesis: $H_0 : \mu_A = \mu_B$.

$\hat{\beta} / s_{\hat{\beta}} = t$

```
> z <- read.csv("diabetes.csv")
> names(z)
 [1] "HDL"           "LDL"           "TotalChol"     "Triglycerides"
 [5] "Creatinine"    "FastInsulin"   "ALT"           "AST"
 [9] "GGT"           "DDuration"     "Age"           "Weight"
[13] "BMI"           "Waist"         "A1CBase"       "FastBG"
[17] "HomaS"         "HomaIR"        "HomaB"         "Diastolic"
[21] "Systolic"      "Pulse"         "Therapy"       "A1C52"
[25] "A1Cdiff"
> t.test(A1Cdiff ~ Therapy, data=z,var.equal=TRUE,alternative="two.sided")
t = 1.4089, df = 800, p-value = 0.1592
```

无法拒绝零假设

因为是0-1分布

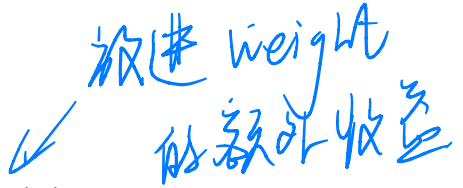
$y = \beta_0 + \beta_1 \cdot Treat + \beta_2 \cdot salary$

计算 β_1，固定 salary 为临界 Treat 看 y 的改变

```
> summary(lm(A1Cdiff ~ Therapy, data=z))
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.46445    0.05317 -27.542   <2e-16 ***
TherapyP    -0.10465    0.07428  -1.409    0.159
```

lm(A1Cdiff ~ . – A1CBase – A1C52, data=z)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.3845825	0.7412652	1.868	0.06216	.
HDL	7.0921082	7.4252863	0.955	0.33981	
LDL	6.7457429	7.4299154	0.908	0.36420	
TotalChol	-6.6762442	7.4283109	-0.899	0.36906	
Triglycerides	3.1720212	3.4028299	0.932	0.35154	
Creatinine	-0.0014036	0.0025733	-0.545	0.58559	
FastInsulin	-0.0042363	0.0022309	-1.899	0.05794	.
ALT	-0.0002606	0.0038294	-0.068	0.94576	
AST	-0.0038380	0.0056152	-0.684	0.49449	
GGT	0.0002968	0.0010794	0.275	0.78339	
DDuration	0.0134485	0.0086608	1.553	0.12088	
Age	0.0022026	0.0046120	0.478	0.63308	
Weight	0.0128551	0.0049309	2.607	0.00931	**
BMI	-0.0355874	0.0136901	-2.599	0.00951	**
Waist	-0.0111150	0.0056152	-1.979	0.04812	*



lm(A1Cdiff $\sim$. – A1CBase – A1C52, data=z)

FastBG	-0.1690230	0.0237942	-7.104	2.75e-12	***
HomaS	-0.0032221	0.0014140	-2.279	0.02295	*
HomaIR	0.2318614	0.0806527	2.875	0.00415	**
HomaB	-0.0039261	0.0028481	-1.379	0.16844	
Diastolic	0.0001408	0.0051229	0.027	0.97809	
Systolic	0.0005319	0.0027948	0.190	0.84912	
Pulse	-0.0080700	0.0038454	-2.099	0.03617	*
TherapyP	-0.0887768	0.0716958	-1.238	0.21600	

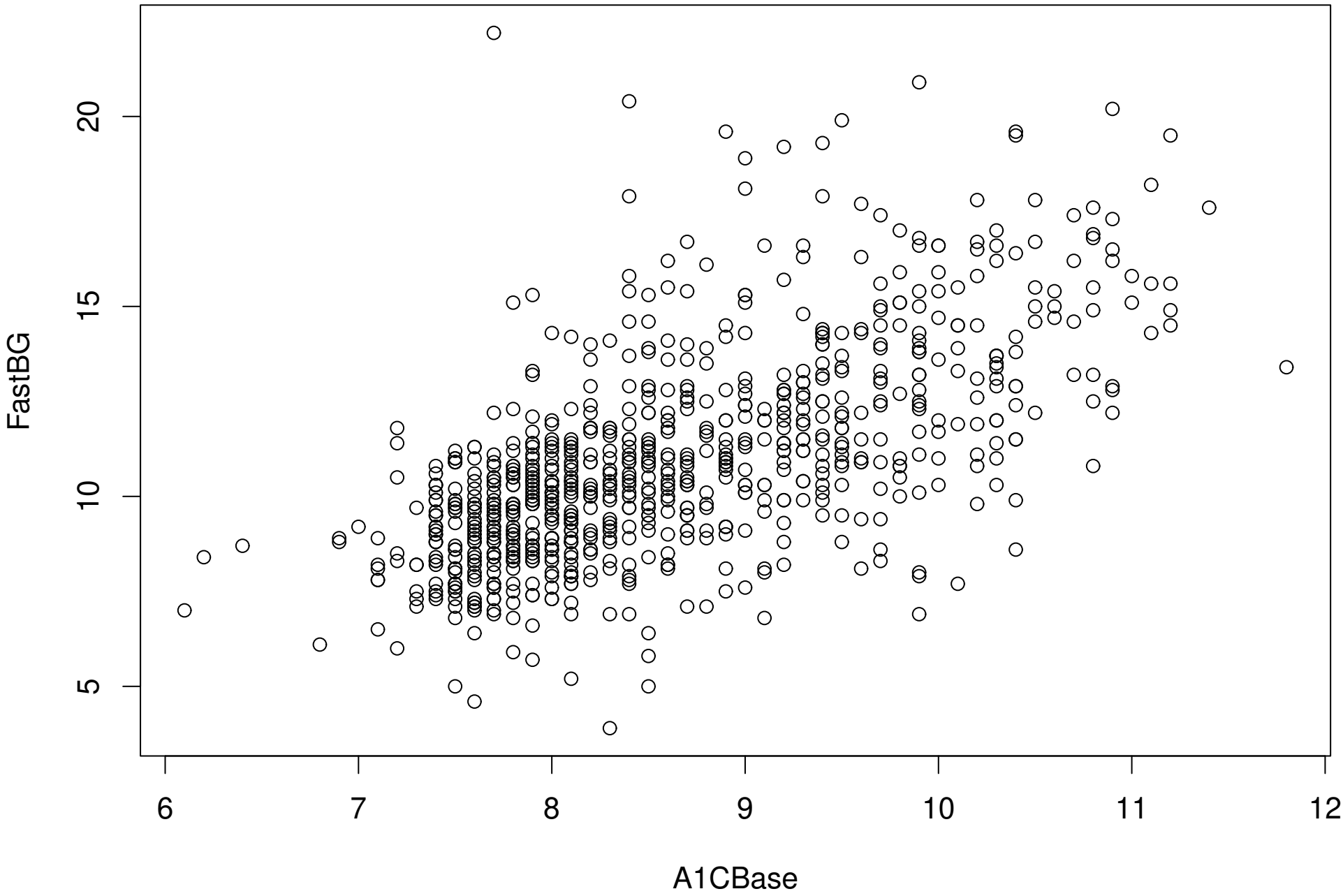
lm(A1Cdiff ~ . – A1C52, data=z)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.950e+00	6.685e-01	7.404	3.44e-13	***
HDL	7.042e+00	6.355e+00	1.108	0.2681	
LDL	6.887e+00	6.359e+00	1.083	0.2791	
TotalChol	-6.829e+00	6.357e+00	-1.074	0.2831	
Triglycerides	3.217e+00	2.912e+00	1.105	0.2697	
Creatinine	-1.312e-03	2.202e-03	-0.596	0.5517	
FastInsulin	6.488e-04	1.931e-03	0.336	0.7370	
ALT	1.283e-03	3.279e-03	0.391	0.6957	
AST	-6.220e-03	4.808e-03	-1.294	0.1961	
GGT	1.847e-05	9.239e-04	0.020	0.9841	
DDuration	6.068e-03	7.425e-03	0.817	0.4140	
Age	9.895e-04	3.948e-03	0.251	0.8021	
Weight	4.659e-03	4.248e-03	1.097	0.2731	
BMI	-1.508e-02	1.178e-02	-1.280	0.2009	
Waist	-6.533e-03	4.813e-03	-1.357	0.1751	

lm(A1Cdiff ~ . – A1C52, data=z)

A1CBase	-7.005e-01	4.145e-02	-16.899	< 2e-16	***
FastBG	5.209e-02	2.420e-02	2.152	0.0317	*
HomaS	-1.227e-03	1.216e-03	-1.009	0.3132	
HomaIR	-3.275e-02	7.078e-02	-0.463	0.6437	
HomaB	-3.039e-04	2.447e-03	-0.124	0.9012	
Diastolic	1.569e-04	4.384e-03	0.036	0.9715	
Systolic	-3.867e-04	2.392e-03	-0.162	0.8716	
Pulse	-7.553e-03	3.291e-03	-2.295	0.0220	*
TherapyP	-7.457e-02	6.136e-02	-1.215	0.2247	

FastBG vs A1CBase



Linear regression with p predictors

- Denote the sample by $\{(y_i, x_{i1}, x_{i2}, \ldots, x_{ip}), i = 1, 2, \ldots, n\}$
- The linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

may be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Vector differentiation

Theorem 1 *Let $\partial/\partial\boldsymbol{\beta}$ denote a column vector with ith component $\partial/\partial\beta_i$. For any constant vector $\mathbf{a}$ and symmetric matrix $\mathbf{A}$,*

$$\begin{aligned}\frac{\partial(\boldsymbol{\beta}'\mathbf{a})}{\partial\boldsymbol{\beta}} &= \mathbf{a}\\ \frac{\partial(\boldsymbol{\beta}'\mathbf{A}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} &= 2\mathbf{A}\boldsymbol{\beta}\end{aligned}$$

Proof. 1st equation is due to $\partial(\boldsymbol{\beta}'\mathbf{a})/\partial\beta_i = \partial(\sum_j \beta_j a_j)/\partial\beta_i = a_i$. For the 2nd,

$$\begin{aligned}\frac{\partial(\boldsymbol{\beta}'\mathbf{A}\boldsymbol{\beta})}{\partial\beta_i} &= \frac{\partial}{\partial\beta_i}\Big(\sum_j\sum_k a_{jk}\beta_j\beta_k\Big)\\ &= 2a_{ii}\beta_i + 2\sum_{i\neq j} a_{ij}\beta_j\\ &= 2\sum_j a_{ij}\beta_j\\ &= 2(\mathbf{A}\boldsymbol{\beta})_i\end{aligned}$$

Least squares theory

Theorem 2 *Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is an $n \times 1$, $\mathbf{X}$ is a $n \times k$ matrix of rank k, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of independent errors with mean 0 and constant variance σ^2. Let $\hat{\boldsymbol{\beta}}$ minimize the sum of squares*

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

and let SSE $= \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$ be the sum of squared errors. Then

1. $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
2. $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
3. $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
4. $s^2 = SSE/(n-k)$ *is an unbiased estimate of* σ^2
5. *For any vector* $\mathbf{c}$, $\mathbf{c}'\hat{\boldsymbol{\beta}}$ *is the best linear unbiased estimate (BLUE) of* $\mathbf{c}'\boldsymbol{\beta}$

Theorem 3 *Suppose in addition that* $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ *are normally distributed. Then*

1. $\hat{\boldsymbol{\beta}}$ *is the uniformly minimum variance unbiased (UMVU) estimate of* $\boldsymbol{\beta}$*, i.e., it minimizes the variance among all unbiased estimates for all values of* $\boldsymbol{\beta}$
2. $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
3. $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_k$
4. $\hat{\boldsymbol{\beta}}$ *is independent of* s^2
5. $(n-k)s^2/\sigma^2 \sim \chi^2_{n-k}$

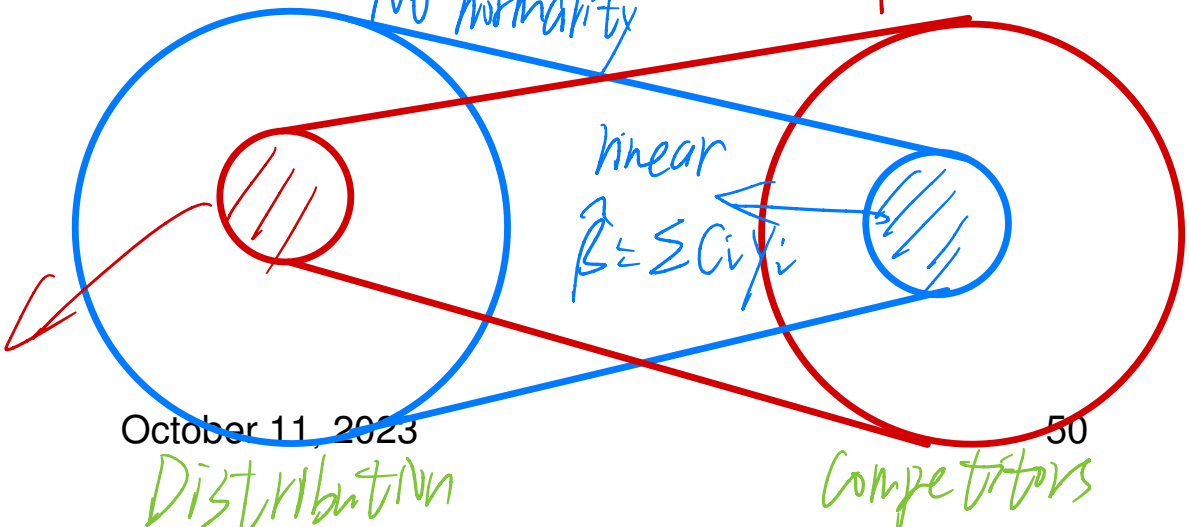
Why σ^2? Cause it depends on scale

No limit except for unbiased

Compare Thm 2, Thm 3

$\frac{SSE}{\sigma^2} = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{\sigma^2}$

normality of ϵ



Model in centered form

- Write $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$
- The model

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

can be written in terms of centered x's as

$$y_i = \alpha + \beta_1 (x_{i1} - \bar{x}_1) + \beta_2 (x_{i2} - \bar{x}_2) + \ldots + \beta_p (x_{ip} - \bar{x}_p) + \epsilon_i$$

where $\bar{x}_j = n^{-1} \sum_{i=1}^{n} x_{ij}$, $j = 1, 2, \ldots, p$, and

$$\alpha = \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \ldots + \beta_p \bar{x}_p$$

- Let $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_p)'$ and define

$$\mathbf{X}_1 = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- Let $\mathbf{J}$ a $n \times n$ matrix of 1's and $\mathbf{I}$ be the $n \times n$ identity matrix
- Define the *centered design matrix*

$$\mathbf{X}_c = (\mathbf{I} - n^{-1}\mathbf{J})\mathbf{X}_1 = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{np} - \bar{x}_p \end{pmatrix}$$

- Let $\mathbf{j} = (1, 1, \ldots, 1)'$ be a $n \times 1$ vector of 1's
- The centered model can be written as

$$\mathbf{y} = (\mathbf{j}, \mathbf{X}_c) \begin{pmatrix} \alpha \\ \boldsymbol{\beta}_1 \end{pmatrix} + \boldsymbol{\epsilon}$$

with least-squares estimates (LSEs)

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix} = [(\mathbf{j}, \mathbf{X}_c)'(\mathbf{j}, \mathbf{X}_c)]^{-1}(\mathbf{j}, \mathbf{X}_c)'\mathbf{y} = \begin{pmatrix} \bar{y} \\ (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y} \end{pmatrix}$$

- For the *simple linear regression model* $y = \alpha + \beta(x - \bar{x}) + \epsilon$, define $s_x^2 = (n-1)^{-1}\sum_{i=1}^n (x_i - \bar{x})^2$, $s_y^2 = (n-1)^{-1}\sum_{i=1}^n (y_i - \bar{y})^2$, and $r = (n-1)^{-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})/(s_x s_y)$. Then

$$\hat{\alpha} = \bar{y} \quad \hat{\beta} = r s_y / s_x$$

$\frac{L_{xy}}{L_{xx}}$

Sums of squares

- Define the corrected total sum of squares SST $= \sum_i (y_i - \bar{y})^2$ and the regression sum of squares

$$\text{SSR} = \|\mathbf{X}_c\hat{\boldsymbol{\beta}}_1\|^2 = (\mathbf{X}_c\hat{\boldsymbol{\beta}}_1)'(\mathbf{X}_c\hat{\boldsymbol{\beta}}_1) = \hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y}$$

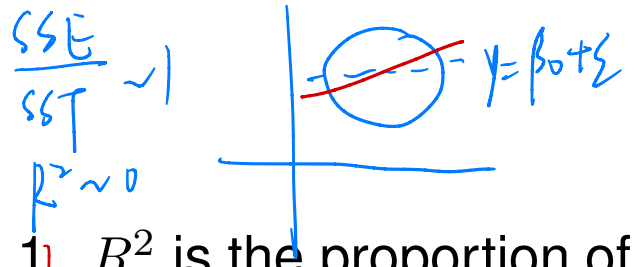
(recall that $\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y}$)

- The error sum of squared errors can be expressed as

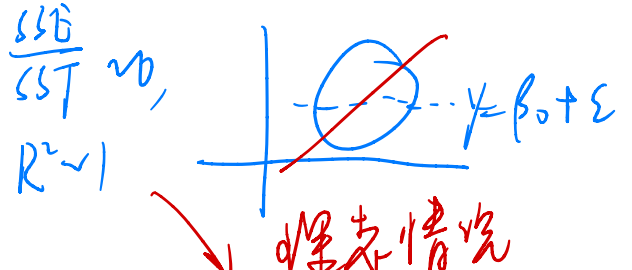
$$\begin{aligned}
\text{SSE} &= \|(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\|^2 = \|(\mathbf{y} - \bar{y}\mathbf{j} - \mathbf{X}_c\hat{\boldsymbol{\beta}}_1)\|^2 \\
&= \sum_i (y_i - \bar{y})^2 + \|\mathbf{X}_c\hat{\boldsymbol{\beta}}_1\|^2 - 2(\mathbf{y} - \bar{y}\mathbf{j})'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1 \\
&= \text{SST} + \text{SSR} - 2(\mathbf{y}'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1 + \bar{y}\mathbf{j}'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1) \\
&= \text{SST} + \text{SSR} - 2\hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y} = \text{SST} - \text{SSR}
\end{aligned}$$

because SSR $= \hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y}$ and $\mathbf{j}'\mathbf{X}_c = \mathbf{0}'$

- Therefore SST $=$ SSR $+$ SSE



Properties of R^2



1. R^2 is the proportion of SST due to regression

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

2. Adjusted R^2 is $R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-1-p)}{\text{SST}/(n-1)}$

3. $0 \le R^2 \le 1$ and $R^2 = 0$ if $\hat{\beta}_1 = \hat{\beta}_2 = \ldots = \hat{\beta}_p = 0$ ($\hat{\beta}_0$ need not be 0)

4. Let $\bar{\hat{y}} = n^{-1}\sum_i \hat{y}_i$. The positive square root $R = \sqrt{R^2}$ is called the *multiple correlation coefficient* and equals the correlation between the y_i and $\hat{y}_i$

$$R = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2 \sum_i (\hat{y}_i - \bar{\hat{y}})^2}}$$

5. R^2 increases if a variable is added to the model but R^2_{adj} can decrease

6. $E(R^2) = p(n-1)^{-1}$ if $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

Expectation and covariance operators

Definition 1

- *Let* $\mathbf{Z}$ *be an* $m \times n$ *matrix of random variables* Z_{ij}
- *Then* $E(\mathbf{Z})$ *is a* $m \times n$ *matrix of expected values* $E(Z_{ij})$

Theorem 4 *If* $\mathbf{A}$, $\mathbf{B}$, *and* $\mathbf{C}$ *are* $l \times m$, $n \times p$, *and* $l \times p$ *matrices of constants,*

$$E(\mathbf{AZB} + \mathbf{C}) = \mathbf{A}E(\mathbf{Z})\mathbf{B} + \mathbf{C}$$

Theorem 5 *If* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $m \times n$ *matrices of constants and* $\mathbf{X}$ *and* $\mathbf{Y}$ *are* $n \times 1$ *random variables, then*

$$E(\mathbf{AX} + \mathbf{BY}) = \mathbf{A}E(\mathbf{X}) + \mathbf{B}E(\mathbf{Y})$$

Definition 2

1. *If* $\mathbf{X}$ *and* $\mathbf{Y}$ *are* $m \times 1$ *and* $n \times 1$ *random vectors,* $Cov(\mathbf{X}, \mathbf{Y})$ *is a* $m \times n$ *matrix with* (i, j)*th element* $Cov(X_i, Y_j) = E\{(X_i - EX_i)(Y_j - EY_j)\}$
2. $Var(\mathbf{X}) = Cov(\mathbf{X}, \mathbf{X})$ *with* (i, j)*th element* $E\{(X_i - EX_i)(X_j - EX_j)\}$

Theorem 6 $Cov(\mathbf{X}, \mathbf{Y}) = E\{(\mathbf{X} - E\mathbf{X})(\mathbf{Y} - E\mathbf{Y})'\}$

Theorem 7 *If* $\mathbf{X}$ *and* $\mathbf{Y}$ *are* $m \times 1$ *and* $n \times 1$ *random vectors and* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $l \times m$ *and* $p \times n$ *matrices of constants,*

$$Cov(\mathbf{AX}, \mathbf{BY}) = \mathbf{A}Cov(\mathbf{X}, \mathbf{Y})\mathbf{B}'$$

Theorem 8 *If* $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{U}$, *and* $\mathbf{V}$ *are random vectors of the same dimension, then for all real numbers* a, b, c, *and* d

$$\begin{aligned} Cov(a\mathbf{X} + b\mathbf{Y}, c\mathbf{U} + d\mathbf{V}) &= acCov(\mathbf{X}, \mathbf{U}) + adCov(\mathbf{X}, \mathbf{V}) \\ &\quad + bcCov(\mathbf{Y}, \mathbf{U}) + bdCov(\mathbf{Y}, \mathbf{V}) \end{aligned}$$

Theorem 9 *Let* $\mathbf{X}$ *be an* $n \times 1$ *random vector and let* $\mathbf{A}$ *be an* $n \times n$ *symmetric matrix of constants. Then*

$$E(\mathbf{X}'\mathbf{A}\mathbf{X}) = tr(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta}$$

where $\boldsymbol{\theta} = E(\mathbf{X})$ *and* $\boldsymbol{\Sigma} = Var(\mathbf{X})$

difference $\hat{\sigma}_1^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$ $\bar{X}$ always known $\Leftarrow$ unbiased

$\hat{\sigma}_2^2 = \frac{\sum (X_i - \mu)^2}{n}$ (μ unknown), $E(\hat{\sigma}_2^2) = \frac{1}{n}\sum E(X_i - \mu)^2 = \frac{1}{n}\sum \mathrm{Var}(X_i)$, still unbiased

Noncentral chi-square distribution

Definition 3 *Let Z_i be independent $N(\theta_i, 1)$, $i = 1, 2, \ldots, p$. Then*

1. *$X = \sum_i Z_i^2$ has a noncentral $\chi_p^2(\lambda)$ with p df and noncentrality parameter $\lambda = \sum_i \theta_i^2$*
2. *The mean and variance of X are $(p + \lambda)$ and $2(p + 2\lambda)$, respectively*
3. *The characteristic function of X is*

$$Ee^{itX} = \frac{\exp\{i\lambda t/(1 - 2it)\}}{(1 - 2it)^{\nu/2}}$$

If $U_1 \sim \chi^2_{p_1}$, $U_2 \sim \chi^2_{p_2}(\lambda)$.

$\frac{U_2/p_2}{U_1/p_1} \sim F_{p_2,p_1}(\lambda)$, $\frac{U_1/p_1}{U_2/p_2} \sim \frac{1}{F_{p_2,p_1}(\lambda)}$

Noncentral F distribution

non centred

Definition 4 *Let* $U_1 \sim \chi^2_{p_1}(\lambda)$ *be independent of* $U_2 \sim \chi^2_{p_2}$. *Then*

central χ^2

$$\frac{U_1/p_1}{U_2/p_2} \sim F_{p_1,p_2}(\lambda)$$

a noncentral F *distribution with* p_1, p_2 *degrees of freedom and noncentraility parameter* λ

Theorem 10 *Suppose* $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$*, where* $\mathbf{X}$ *is* $n \times (p+1)$ *of rank* $p+1 < n$ *and let*

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

and $\lambda = \boldsymbol{\beta}_1'\mathbf{X}_c'\mathbf{X}_c\boldsymbol{\beta}_1/\sigma^2$*. Then*

1. $SSR/\sigma^2 \sim \chi^2_p(\lambda)$
2. $SSE/\sigma^2 \sim \chi^2_{n-p-1}$
3. *SSR and SSE are independent*
4. $F \sim F_{p,n-p-1}(\lambda)$ *(noncentral* F *distribution)*

Testing a subset of βs

- Consider testing $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$ in the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

  where $\boldsymbol{\beta}_2$ is $h \times 1$ and the intercept term β_0 is included in $\boldsymbol{\beta}_1$

- Let SSE_0 and SSE_1 be the residual sums of squares under H_0 and H_1, respectively

- Then

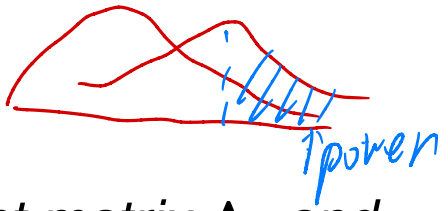
$$F = \frac{(\text{SSE}_0 - \text{SSE}_1)/h}{\text{SSE}_1/(n-p-1)} \sim F_{h,n-p-1}(\lambda)$$

  where

$$\lambda = \frac{\boldsymbol{\beta}_2'\{\mathbf{X}_2'\mathbf{X}_2 - \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\}\boldsymbol{\beta}_2}{\sigma^2}$$

Power = P(Reject H_0 | H_0 is false)
↪ H_0 should be rejected
= P($F > F_{h,n-p-1;\alpha}$ | H_0 is false)

↑ power ↑ more data, n
↑ what happens to H_1
λ ↑



Theorem 11 *Let* $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$, r *be the rank of a constant matrix* $\mathbf{A}$, *and* $\lambda = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$. *Then*

$1\times n \quad n\times n \quad n\times 1$

$$\mathbf{y}'\mathbf{A}\mathbf{y} \sim \chi^2_r(\lambda)$$

$F_{h,n-p-1;\alpha}$
n double
what will happen

if and only if $\mathbf{AV}$ *is idempotent (i.e.,* $\mathbf{AVAV} = \mathbf{AV}$*)*

Corollary 1 *Let* $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$ *with* $\mathbf{V}$ *of rank* n. *Then* $\mathbf{y}'\mathbf{V}^{-1}\mathbf{y} \sim \chi^2_n(\lambda)$ *with* $\lambda = \boldsymbol{\mu}'\mathbf{V}^{-1}\boldsymbol{\mu}$.

Test statistic starts from $C\hat{\beta}$

Testing $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$

- Suppose $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and $\mathbf{C}$ is $q \times (p+1)$ of rank $q \leq p+1$. Then

$$\mathbf{C}\hat{\boldsymbol{\beta}} \sim N_q[\mathbf{C}\boldsymbol{\beta}, \sigma^2\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'],$$

- Define

$$\begin{aligned} \text{SSH} &= (\mathbf{C}\hat{\boldsymbol{\beta}})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\hat{\boldsymbol{\beta}} \\ \lambda &= (\mathbf{C}\boldsymbol{\beta})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\boldsymbol{\beta}/\sigma^2 \end{aligned}$$

using Thm 11 up page Corollary 1

- Then
  1. $\text{SSH}/\sigma^2 \sim \chi^2_q(\lambda)$
  2. $\text{SSE}/\sigma^2 \sim \chi^2_{n-p-1}$
  3. SSH and SSE are independent and

$$\text{SSH} = \frac{\|C\hat{\beta}\|^2}{\text{Var}(C\hat{\beta})}$$

$$F = \frac{\text{SSH}/q}{\text{SSE}/(n-p-1)} \sim F_{q,n-p-1}(\lambda)$$

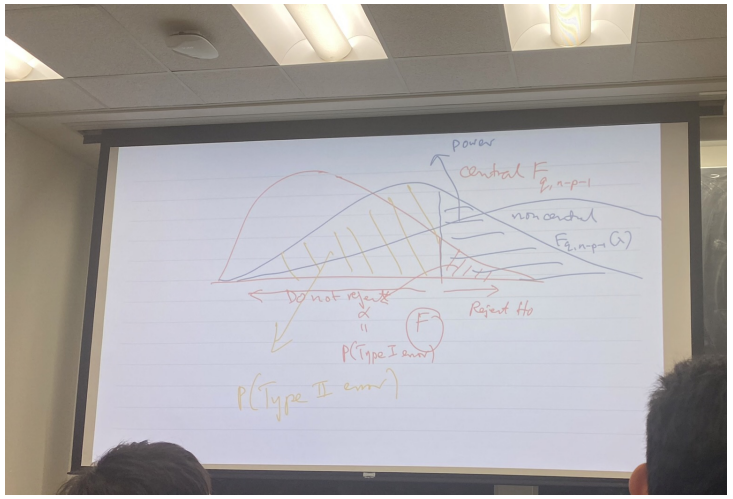
Testing $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$

Suppose $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and $\mathbf{C}$ is $q \times (p+1)$ of rank $q \leq p+1$. Then $\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{t} \sim N_q[\mathbf{C}\boldsymbol{\beta} - \mathbf{t}, \sigma^2\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']$. Define

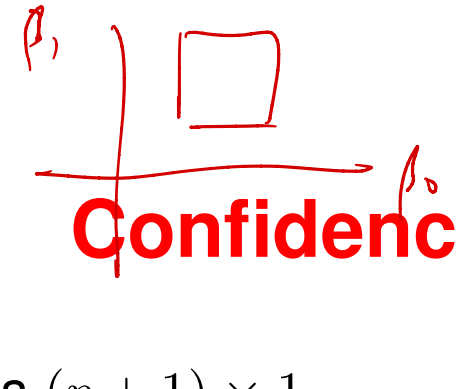
$$\begin{aligned} \text{SSH} &= (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{t})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{t}) \qquad (1) \\ \lambda &= (\mathbf{C}\boldsymbol{\beta} - \mathbf{t})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C}\boldsymbol{\beta} - \mathbf{t})/\sigma^2 \end{aligned}$$

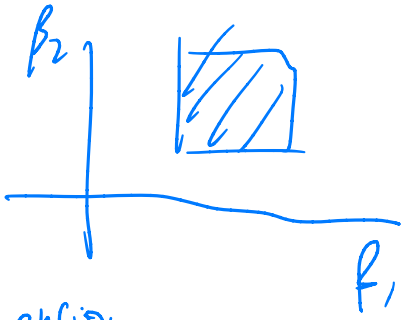
Then

1. $\text{SSH}/\sigma^2 \sim \chi^2_q(\lambda)$
2. $\text{SSE}/\sigma^2 \sim \chi^2_{n-p-1}$
3. SSH and SSE are independent and



$$F = \frac{\text{SSH}/q}{\text{SSE}/(n-p-1)} = \frac{\text{SSH}/q}{s^2} \sim F_{q,n-p-1}(\lambda) \qquad (2)$$





Confidence region for $\boldsymbol{\beta}$

because β is not 1 dimension

- Recall that $\boldsymbol{\beta}$ is $(p+1) \times 1$
- Put $\mathbf{C} = \mathbf{I}$ and $\mathbf{t} = \boldsymbol{\beta}$ in (1), so that $q = p+1$
- It follows from (2) that

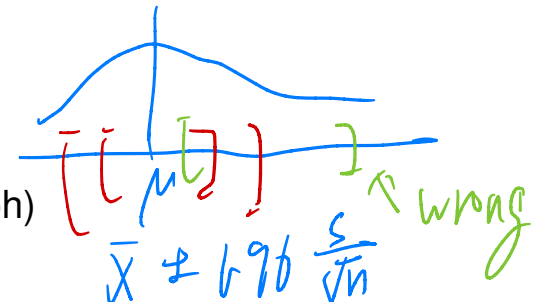
$$W = s^{-2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/(p+1) \sim F_{p+1,n-p-1}$$

- Therefore

$$P\{s^{-2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/(p+1) \leq F_{p+1,n-p-1;\alpha}\} = 1 - \alpha$$

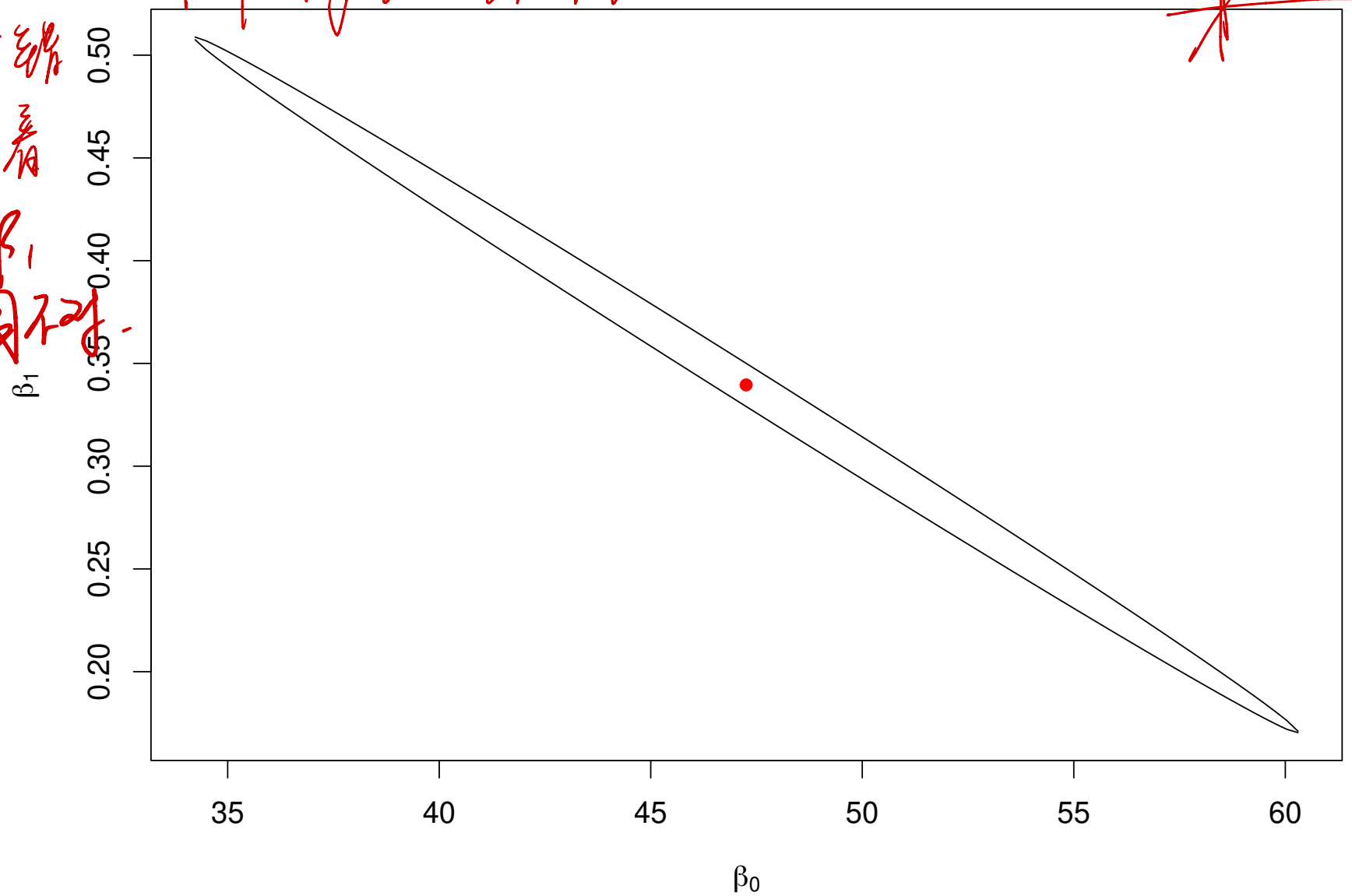
and a $100(1-\alpha)$% confidence region for $\boldsymbol{\beta}$ is

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq (p+1)s^2 F_{p+1,n-p-1;\alpha}$$



check $|\bar{X} - \mu| \leq 1.96 \frac{\sigma}{\sqrt{n}}$

95% confidence region for (β_0, β_1) from fitting height vs reach in boxer data



Confidence intervals for β_j and $\mathbf{a}'\boldsymbol{\beta}$

- Recall from Theorem 3 that $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
- Let $G = (\mathbf{X}'\mathbf{X})^{-1}$ and g_{ij} be the (i,j)th element of G
- Then $\hat{\beta}_j \sim N(\beta_j, \sigma^2 g_{jj})$ and $(\hat{\beta}_j - \beta_j)s^{-1}/\sqrt{g_{jj}} \sim t_{n-p-1}$
- A $100(1-\alpha)$% confidence interval for β_j is

$$\hat{\beta}_j \pm s\sqrt{g_{jj}}\, t_{n-p-1;\alpha/2}$$

- Similarly a $100(1-\alpha)$% confidence interval for $\mathbf{a}'\boldsymbol{\beta}$ is

$$\mathbf{a}'\hat{\boldsymbol{\beta}} \pm s\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}\, t_{n-p-1;\alpha/2} \tag{3}$$

Confidence interval for $E(y \,|\, \mathbf{x} = \mathbf{x}_0)$

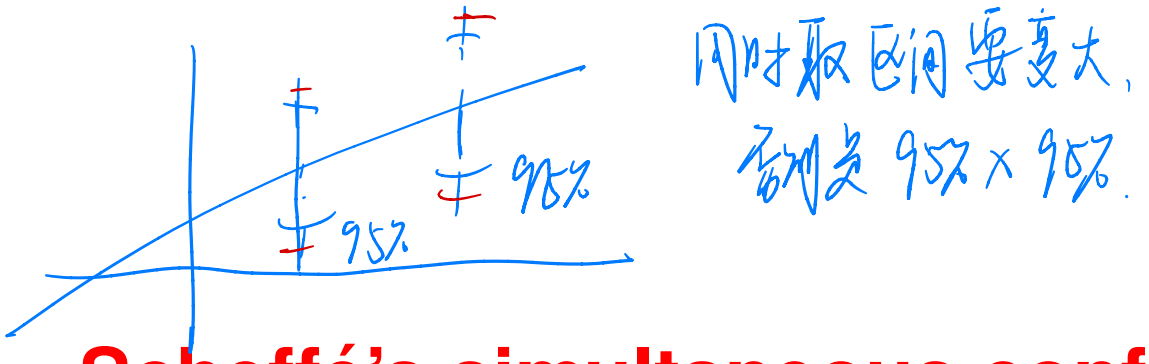
- Since $E(y \,|\, \mathbf{x} = \mathbf{x}_0) = \mathbf{x}_0'\boldsymbol{\beta}$, an estimate of $E(y \,|\, \mathbf{x} = \mathbf{x}_0)$ is

$$\widehat{E(y \,|\, \mathbf{x}_0)} = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$$

- Putting $\mathbf{a} = \mathbf{x}_0$ in (3) yields the $100(1-\alpha)\%$ interval for $E(y \,|\, \mathbf{x} = \mathbf{x}_0)$

$$\mathbf{x}_0'\hat{\boldsymbol{\beta}} \pm s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\; t_{n-p-1;\alpha/2}$$

$P($ $\downarrow$ $\ni \mathbf{x}_0'\boldsymbol{\beta}) = 1-\alpha$



Scheffé's simultaneous confidence band for $E(y \mid \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ for all $\mathbf{x}$

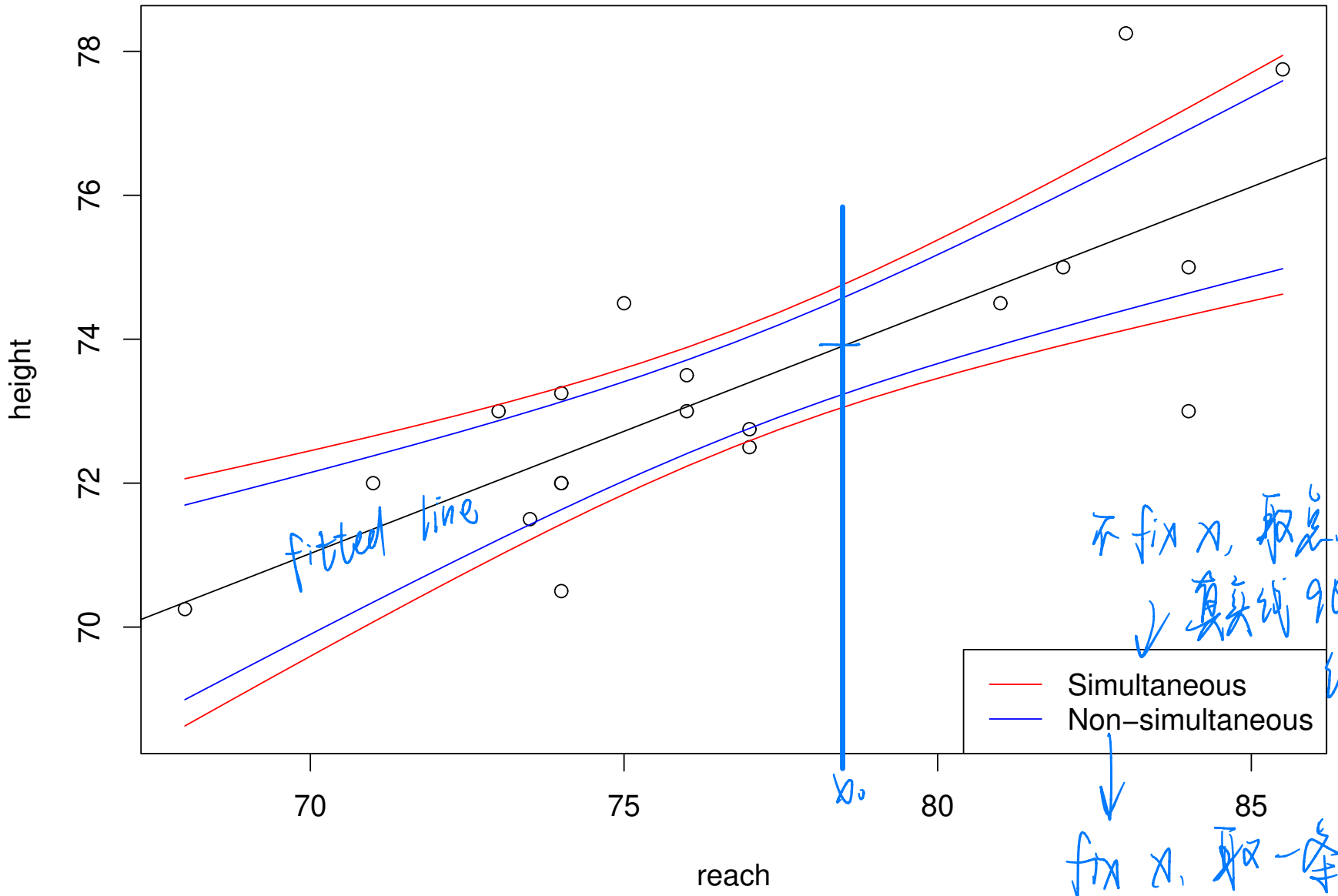
- Scheffé showed that

$$P\left(\mathbf{x}'\boldsymbol{\beta} \in \mathbf{x}'\hat{\boldsymbol{\beta}} \pm s\sqrt{(p+1)\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\, F_{p+1,n-p-1;\alpha}} \quad \text{for all } \mathbf{x}\right) = 1-\alpha$$

- Hence a $100(1-\alpha)$% simultaneous confidence band for $E(y \mid \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ is

$$\mathbf{x}'\hat{\boldsymbol{\beta}} \pm s\sqrt{(p+1)\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\, F_{p+1,n-p-1;\alpha}}$$

95% simultaneous and non-simultaneous confidence bands for boxer data



不fix x_0，取点随机，真实线95%在红线内

fix x_0，取一条线上的点，真实线在x_0处的点在蓝线内

Homework 2 (due 11:59 pm Tue Oct 10 in Canvas)

1. Reproduce the plot of 95% simultaneous confidence band for $E(y \mid \mathbf{x})$.
2. Show that for testing $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ vs $H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$ on page 62, the value of

$$F = \frac{(\text{SSE}_0 - \text{SSE}_1)/h}{\text{SSE}_1/(n-p-1)}$$

   is the same as the value of the statistic

$$F = \frac{\text{SSH}/h}{\text{SSE}/(n-p-1)}$$

   on page 64

Hints for HW 2 Problem #2

Let $\hat{\boldsymbol{\beta}}_r$ and $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ denote the LSEs of $\boldsymbol{\beta}$ under H_0 (reduced model) and H_1 (full model), respectively. Then

$$\text{SSE}_0 - \text{SSE}_1 = \text{SSR}_1 - \text{SSR}_0 = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \hat{\boldsymbol{\beta}}_r'\mathbf{X}'\mathbf{y} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_r)'\mathbf{X}'\mathbf{y}$$

Prove that this is equal to $\text{SSH} = (\mathbf{C}\hat{\boldsymbol{\beta}})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}$ by showing that

$$\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}$$

which can be done using Lagrange multipliers to minimize

$$f = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\lambda}'(\mathbf{C}\boldsymbol{\beta} - \mathbf{0})$$

with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ as follows:

1. Differentiate f with respect to $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ and set the results to $\mathbf{0}$ to obtain two equations
2. Eliminate $\boldsymbol{\lambda}$ from the equations and solve for $\hat{\boldsymbol{\beta}}_r$

Prediction interval for a future observation

- Let y_0 be a future value of y at $\mathbf{x} = \mathbf{x}_0$ and its predicted value be $\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$
- By independence of y_0 and $\hat{y}_0$,

$$\begin{aligned}
\text{var}(y_0 - \hat{y}_0) &= \text{var}(y_0) + \text{var}(\hat{y}_0) = \text{var}(\epsilon) + \text{var}(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) \\
&= \sigma^2 + \sigma^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 \\
&= \sigma^2 \left\{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right\}
\end{aligned}$$

- Hence, because $y_0 - \hat{y}_0$ is normally distributed with mean 0,

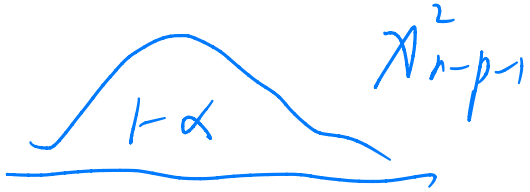
$$\frac{y_0 - \hat{y}_0}{s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \sim t_{n-p-1}$$

yielding $P(|y_0 - \hat{y}_0| \le s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\, t_{n-p-1;\alpha/2}) = 1 - \alpha$

- Therefore a $100(1-\alpha)$% prediction interval for y_0 is

$$\mathbf{x}_0'\hat{\boldsymbol{\beta}} \pm s\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}\, t_{n-p-1;\alpha/2}$$

compare to page 69



Confidence interval for σ^2

$s^2 = \frac{SSE}{n-p-1}$

- From Theorem 3, $(n-p-1)s^2/\sigma^2 = \text{SSE}/\sigma^2 \sim \chi^2_{n-p-1}$
- So

$$P\{\chi^2_{n-p-1;1-\alpha/2} \leq (n-p-1)s^2/\sigma^2 \leq \chi^2_{n-p-1;\alpha/2}\} = 1-\alpha$$

and a $100(1-\alpha)$% interval for σ^2 is

$$\frac{(n-p-1)s^2}{\chi^2_{n-p-1;\alpha/2}} \leq \sigma^2 \leq \frac{(n-p-1)s^2}{\chi^2_{n-p-1;1-\alpha/2}}$$

$\text{cov}(\mathbf{y}) = \begin{pmatrix} \sigma^2 & 2\sigma^2 \\ 2\sigma^2 & \sigma^2 \end{pmatrix}$

Generalized least squares

- Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{V}$, where $\mathbf{X}$ is a $n \times k$ matrix of rank k and $\mathbf{V}$ is a known positive definite matrix
- The generalized least squares estimate $\hat{\boldsymbol{\beta}}$ minimizes

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Definition

Theorem 12 *Let* $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ *and* $cov(\mathbf{y}) = \sigma^2\mathbf{V}$, *where* $\mathbf{X}$ *is a* $n \times k$ *matrix of rank* k *and* $\mathbf{V}$ *is a known positive definite matrix. Then*

1. *Best linear unbiased estimate (BLUE) of* $\boldsymbol{\beta}$ *is*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

   and

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

2. *An unbiased estimate of* σ^2 *is*

$$\begin{aligned} s^2 &= \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-k} \\ &= \frac{\mathbf{y}'\{\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\}\mathbf{y}}{n-k} \end{aligned}$$

3. *If* $\mathbf{y}$ *is* $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$, $\hat{\boldsymbol{\beta}}$ *is also the MLE*

Two proofs of BLUE

1. Differentiate $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$

2. (a) Since $\mathbf{V}$ is positive definite, $\mathbf{V} = \mathbf{P}\mathbf{P}'$ for a $n \times n$ nonsingular matrix $\mathbf{P}$

   (b) Multiply the model equation to get the model

$$\mathbf{P}^{-1}\mathbf{y} = \mathbf{P}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}^{-1}\boldsymbol{\epsilon}$$

   (c) Then $E(\mathbf{P}^{-1}\boldsymbol{\epsilon}) = \mathbf{0}$ and

$$\begin{aligned} \text{Cov}(\mathbf{P}^{-1}\boldsymbol{\epsilon}) &= \mathbf{P}^{-1}\text{Cov}(\boldsymbol{\epsilon})(\mathbf{P}^{-1})' = \mathbf{P}^{-1}\sigma^2\mathbf{V}(\mathbf{P}^{-1})' \\ &= \sigma^2\mathbf{P}^{-1}\mathbf{P}\mathbf{P}'(\mathbf{P}')^{-1} = \sigma^2\mathbf{I} \end{aligned}$$

   (d) From Theorem 2, the BLUE is

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= [(\mathbf{P}^{-1}\mathbf{X})'(\mathbf{P}^{-1}\mathbf{X})]^{-1}(\mathbf{P}^{-1}\mathbf{X})'\mathbf{P}^{-1}\mathbf{y} \\ &= [\mathbf{X}'(\mathbf{P}^{-1})'\mathbf{P}^{-1}\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{P}^{-1})'\mathbf{P}^{-1}\mathbf{y} \\ &= [\mathbf{X}'(\mathbf{P}')^{-1}\mathbf{P}^{-1}\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{P}')^{-1}\mathbf{P}^{-1}\mathbf{y} \\ &= [\mathbf{X}'(\mathbf{P}\mathbf{P}')^{-1}\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{P}\mathbf{P}')^{-1}\mathbf{y} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \end{aligned}$$

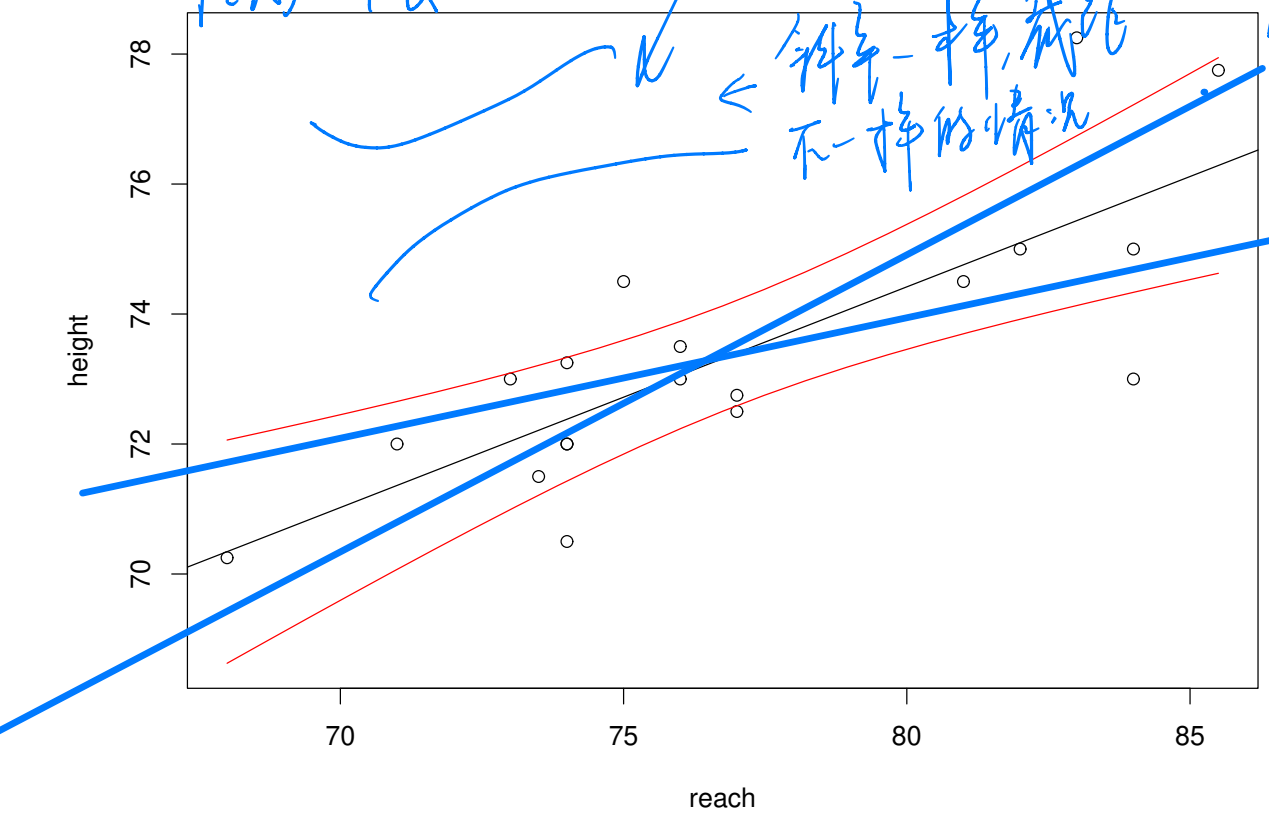
$$\mathbf{x}'\hat{\boldsymbol{\beta}} \pm s\sqrt{(p+1)\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\, F_{p+1,n-p-1;\alpha}}$$

A: the line in band

B: $\beta_{min} < \beta < \beta_{max}$

$P(A) = 1-\alpha$

A ⊄ B ⇒ $1-\alpha = P(A) < P(B)$



slop = $\hat{\beta}_{max}$

slop = $\hat{\beta}_{min}$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 19 & 1462.0 \\ 1462 & 112931.5 \end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 13.681 & -0.177 \\ -0.177 & 0.002 \end{pmatrix}$$

$F_{2,17;0.05} = 3.591531$

$t_{17;0.025} = 2.109816$

1° $P(\hat{\beta}_{min} < \beta < \hat{\beta}_{max}) > 1-\alpha$

2° $(\hat{\beta}_{min}, \hat{\beta}_{max})$.

$> \hat{\beta} \pm t_{n-2,\frac{\alpha}{2}} SD(\hat{\beta})$

↓

is BLUE

& smaller

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.25588    4.86946   9.705 2.40e-08 ***
reach        0.33953    0.06316   5.376 5.03e-05 ***
Residual standard error: 1.316 on 17 degrees of freedom
Multiple R-squared:  0.6296,Adjusted R-squared:  0.6078
```

I_L 95% t- interval

$I_1 = \begin{cases} \emptyset & \text{with prob } 4\% \\ (-\infty, +\infty) & 96\% \end{cases}$ ⇒ 有分数越大，不等待着期望越大

Model error misspecification

- Suppose model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$ with $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{V}$ for some $\mathbf{V} \neq \mathbf{I}$ and we assume that $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$
- Denoting the ordinary least squares (OLS) estimate by $\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$,

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}^*) &= \boldsymbol{\beta} \\ \text{cov}(\hat{\boldsymbol{\beta}}^*) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

$\text{cov}(\hat{\beta}^*) = A\,\text{cov}(y)A'$

- Thus OLS estimates remains unbiased but their variances tend to be larger

Model structure misspecification

- Suppose the predictors can be partitioned into two sets

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = (\mathbf{X}_1, \mathbf{X}_2) \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon} \\ &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \end{aligned}$$

- If we leave out $\mathbf{X}_2\boldsymbol{\beta}_2$ when it should be included (i.e., when $\boldsymbol{\beta}_2 \neq \mathbf{0}$), it is called underfitting
- If we include $\mathbf{X}_2\boldsymbol{\beta}_2$ when it should be excluded (i.e., when $\boldsymbol{\beta}_2 = \mathbf{0}$), it is called overfitting

Theorem 13 *(**Underfitting**). If we fit the reduced model* $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1^* + \boldsymbol{\epsilon}^*$, *where* $\mathbf{X}_1$ *is* $n \times (p+1)$, *when the true model is* $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$ *with* $cov(\mathbf{y}) = \sigma^2\mathbf{I}$, *then the mean and covariance of OLS estimate* $\hat{\boldsymbol{\beta}}_1^* = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$ *are*

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_1^*) &= \boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2 \\ cov(\hat{\boldsymbol{\beta}}_1^*) &= \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1} \end{aligned}$$

where $\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ *and the variance estimate*

$$s_1^2 = \frac{(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^*)'(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^*)}{n - p - 1}$$

has expected value

$$E(s_1^2) = \sigma^2 + \frac{\boldsymbol{\beta}_2'\mathbf{X}_2'\{\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\}\mathbf{X}_2\boldsymbol{\beta}_2}{n - p - 1} \geq \sigma^2$$

Corollary 2 *If* $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$, *i.e., if the columns of* $\mathbf{X}_1$ *are orthogonal to the columns of* $\mathbf{X}_2$, *then* $\hat{\boldsymbol{\beta}}_1^*$ *is unbiased*

overfit

等回来图是有无 outlier

Box-Cox transformation

Box-Cox method assumes that $\{y_1^\lambda, y_2^\lambda, \ldots, y_n^\lambda\}$ for some $\lambda \neq 0$, or $\{\log y_1, \log y_2, \ldots, \log y_n\}$ (corresponding to $\lambda = 0$)

1. are normally distributed
2. are mutually independent
3. have constant variance
4. satisfy a chosen model

$y^\lambda \sim N(X\beta, \sigma^2 I)$

解决另外三张图的问题

Maximum likelihood estimation of Box-Cox λ

- Define

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log y, & \lambda = 0 \end{cases}$$

- Assume that there is λ such that $\mathbf{y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + N(0, \sigma^2)$
- Likelihood function is

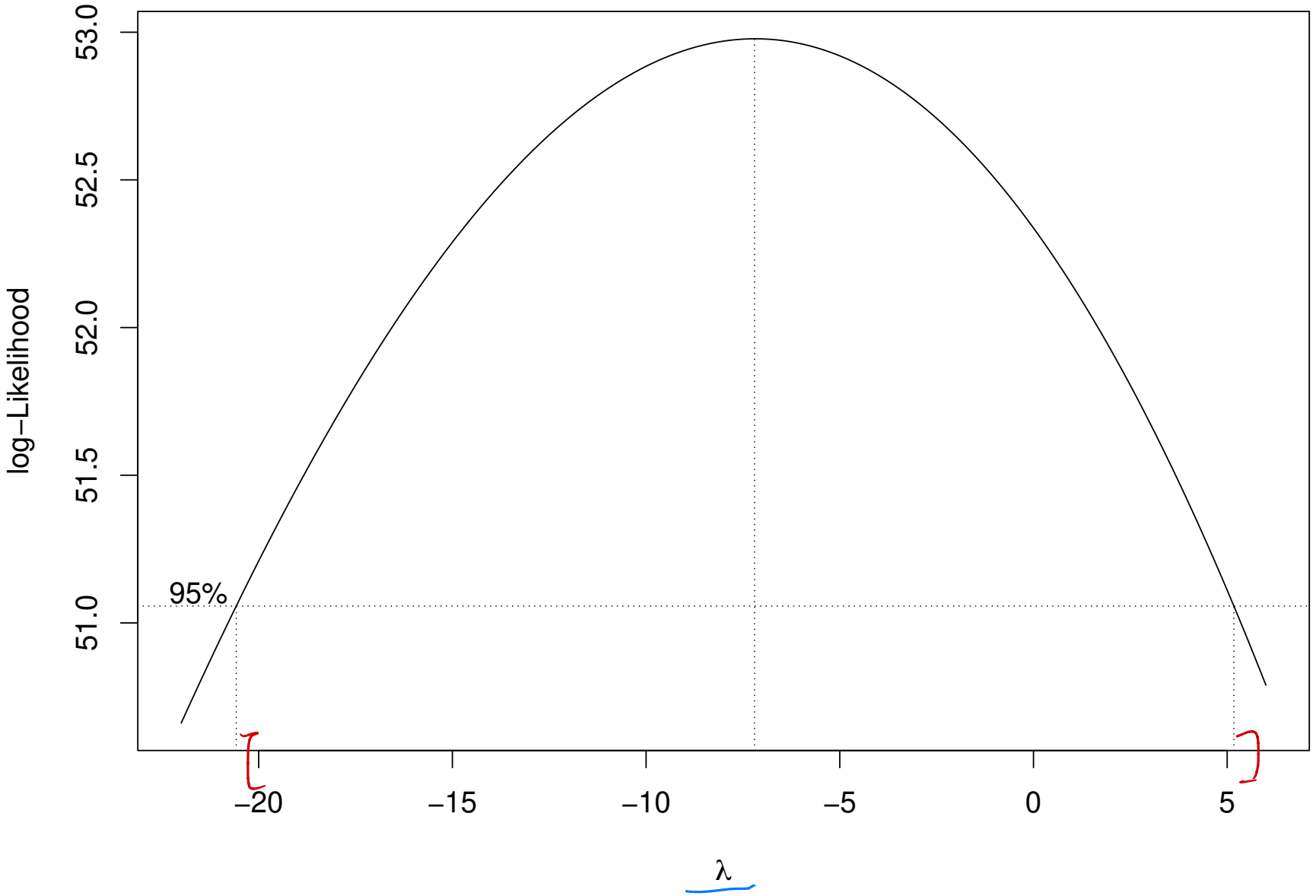
$$\frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{(\mathbf{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y}^{(\lambda)} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right\} J(\lambda, \mathbf{y})$$

with Jacobian

$$J(\lambda, \mathbf{y}) = \prod_{i=1}^{n} \frac{\mathrm{d}y_i^{(\lambda)}}{\mathrm{d}y_i} = \prod_{i=1}^{n} y_i^{\lambda - 1}$$

- Estimated λ is the MLE
- R function is called `boxcox` in the `MASS` package

`boxcox` plot for boxer data using all predictors



Solution without using boxcox

- Given λ, let $(\hat{\boldsymbol{\beta}}_\lambda, \hat{\sigma}_\lambda)$ maximize $\ell(\lambda, \boldsymbol{\beta}, \sigma)$
- Then $(\hat{\boldsymbol{\beta}}_\lambda, \hat{\sigma}_\lambda)$ are the LSEs for the $\mathbf{y}^{(\lambda)}$ data, i.e.,

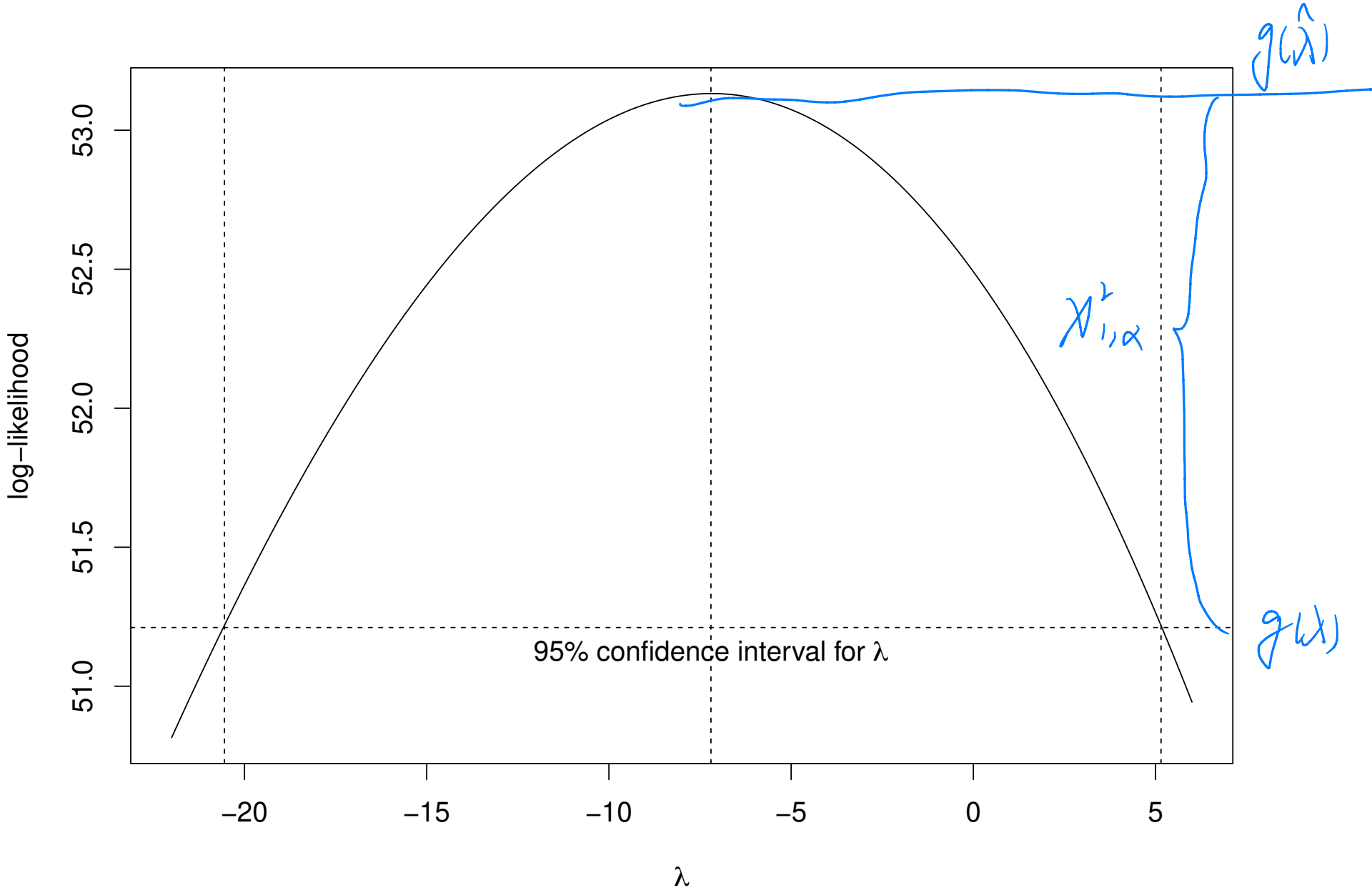
$$\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{(\lambda)}, \quad \hat{\sigma}^2_\lambda = \frac{(\mathbf{y}^{(\lambda)} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)'(\mathbf{y}^{(\lambda)} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)}{n-p-1}$$

and

$$\begin{aligned}
\ell(\lambda, \hat{\boldsymbol{\beta}}_\lambda, \hat{\sigma}_\lambda) &= \frac{\exp(-(n-p-1)/2)}{(2\pi)^{n/2}\hat{\sigma}^n_\lambda} J(\lambda, \mathbf{y}) \\
&= \hat{\sigma}_\lambda^{-n} \prod_{i=1}^n y_i^{\lambda-1} \frac{\exp(-(n-p-1)/2)}{(2\pi)^{n/2}}
\end{aligned}$$

- Let $g(\lambda) = \log \ell(\lambda, \hat{\boldsymbol{\beta}}_\lambda, \hat{\sigma}_\lambda) =$
  $-n\log(\hat{\sigma}_\lambda) + (\lambda-1)\sum_{i=1}^n \log(y_i) - \{(n-p-1) + n\log(2\pi)\}/2$
- MLE $\hat{\lambda}$ maximizes $g(\lambda)$
- $100(1-\alpha)$% confidence limits for λ are roots of $g(\lambda) = g(\hat{\lambda}) - \chi^2_{1;\alpha}/2$

Manual plot for boxer data using all predictors



Homework 3
(due 11:59 PM Tue Oct 17)

1. Reproduce the confidence region plot on page 67
2. Show that the Box-Cox λ can be computed as follows:
   (a) Let $\dot{y} = \sqrt[n]{y_1 y_2 \dots y_n}$ denote the geometric mean of the y observations
   (b) For each λ, transform the y_i to

$$z_i(\lambda) = \begin{cases} \frac{y_i^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{y} \log y_i, & \lambda = 0 \end{cases}$$

   (c) Fit the desired linear regression model to $z_1(\lambda), z_2(\lambda), \dots, z_n(\lambda)$ and compute its residual sum of squares $S(\lambda)$
   (d) The Box-Cox λ minimizes $S(\lambda)$

Generalized linear models

- Linear regression with normal errors assumes that $\mathbf{Y}$ is normally distributed with constant variance and mean $\boldsymbol{\mu} = E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$
- Advantage: $\boldsymbol{\beta}$ gives effects of each predictor, controlling for the others
- A generalized linear model extends this idea to other distributions by introducing a **link function**

$$\eta(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$$

- Three useful link functions:

  Normal: $\eta(\boldsymbol{\mu}) = \boldsymbol{\mu}$ (identity function)

  Bernoulli with $\boldsymbol{\mu} = \mathbf{p}$: $\eta(\mathbf{p}) = \log(\mathbf{p}/(1-\mathbf{p}))$ (logit function)

  Poisson: $\eta(\boldsymbol{\mu}) = \log(\boldsymbol{\mu})$ (log function)

Logistic regression example: Blowdown data

- A severe thunderstorm occurred in July 4, 1999, over 477,000 acres of the Boundary Waters Canoe Area Wilderness in northeastern Minnesota (data in `blowdown.csv` on Canvas; also in R package `alr4`)
- Observations from 3666 trees were collected on the following variables on each tree:
  1. whether it was blown down ($Y = 1$) or not ($Y = 0$)
  2. its trunk diameter *D* in centimeters
  3. its species *S* (A-aspen, BA-black ash, BF-balsam fir, BS-black spruce, C-cedar, JP-jack pine, PB-paper birch, RM-red maple, RP-red pine)
  4. local intensity *L* of the storm (measured by fraction of damaged trees in its vicinity)

Logistic regression theory

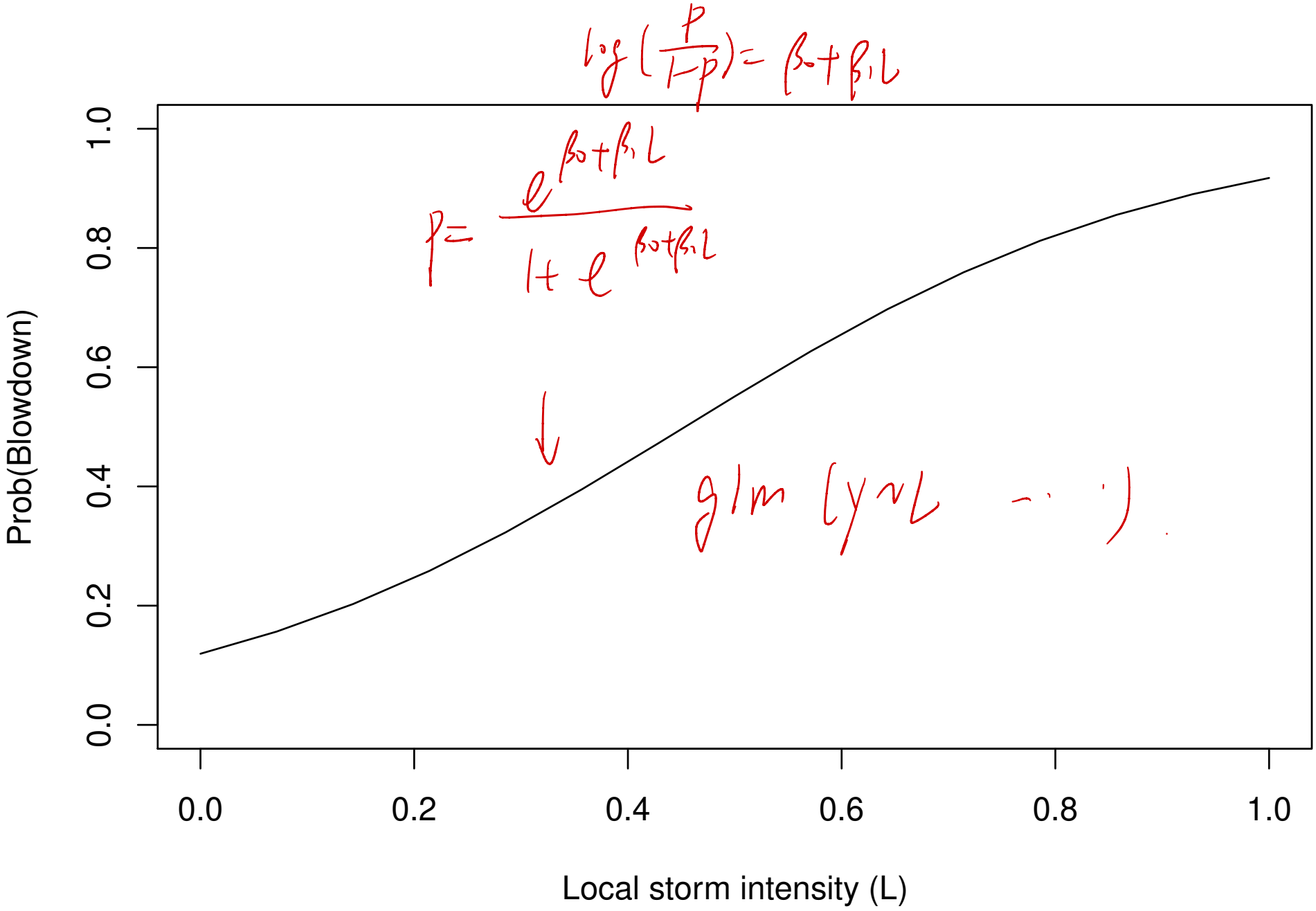
- Y be Bernoulli with $p = P(Y = 1)$, let $\boldsymbol{X} = (X_1, \ldots, X_p)$ be the predictor variables, and define the logit link function $\text{logit}(p) = \log(p/(1-p))$
- The *multiple linear logistic* regression model assumes that

$$\text{logit}(p) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

  or equivalently, $p = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_j)/\{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_j)\}$
- Let n be the sample size, let $(x_{i1}, \ldots, x_{ip}, y_i)$ be the values of $(X_1, \ldots, X_p, Y)$ for the ith observation and let $p_i = P(Y = y_i \mid \boldsymbol{X} = \boldsymbol{x}_i)$
- $\beta_0, \beta_1, \ldots, \beta_k$ are estimated by maximizing the likelihood function

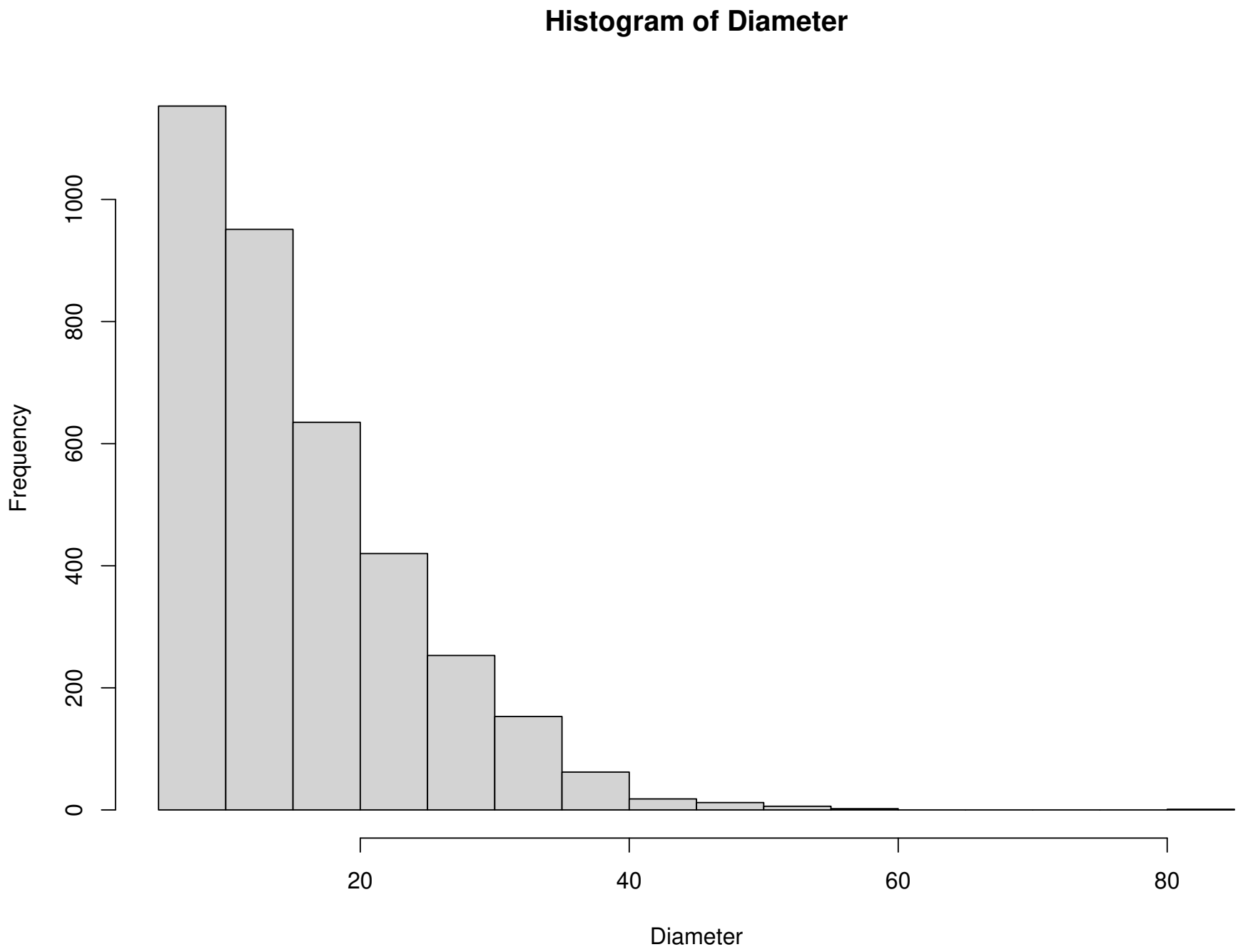
$$\prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i} = \frac{\exp\{\sum_{i=1}^{n} y_i(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})\}}{\prod_i \{1 + \exp(\beta_0 + \sum_j \beta_j x_{ij})\}}$$

Logistic regression model using only *L*



```
> z <- read.csv("blowdown.csv")
> glm(Y ~ ., data=z, family='binomial')
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.887322	0.213111	-13.548	< 2e-16	***
Diameter	0.075279	0.006872	10.955	< 2e-16	***
Intensity	4.632263	0.209640	22.096	< 2e-16	***
SpeciesBA	-2.337078	0.485614	-4.813	1.49e-06	***
SpeciesBF	-0.360890	0.173417	-2.081	0.0374	*
SpeciesBS	0.057240	0.151798	0.377	0.7061	
SpeciesC	-2.130241	0.215248	-9.897	< 2e-16	***
SpeciesJP	1.076664	0.178150	6.044	1.51e-09	***
SpeciesPB	-1.834158	0.185793	-9.872	< 2e-16	***
SpeciesRM	-1.912548	0.300052	-6.374	1.84e-10	***
SpeciesRP	-0.145442	0.424336	-0.343	0.7318	


Histogram of Diameter
Frequency
0
200
400
600
800
1000
20
40
60
80
Diameter

Using log(Diameter) in place of Diameter

```
> z$logDiameter <- log(z$Diameter)
> glm(Y ~ . - Diameter, data=z, family='binomial')
```

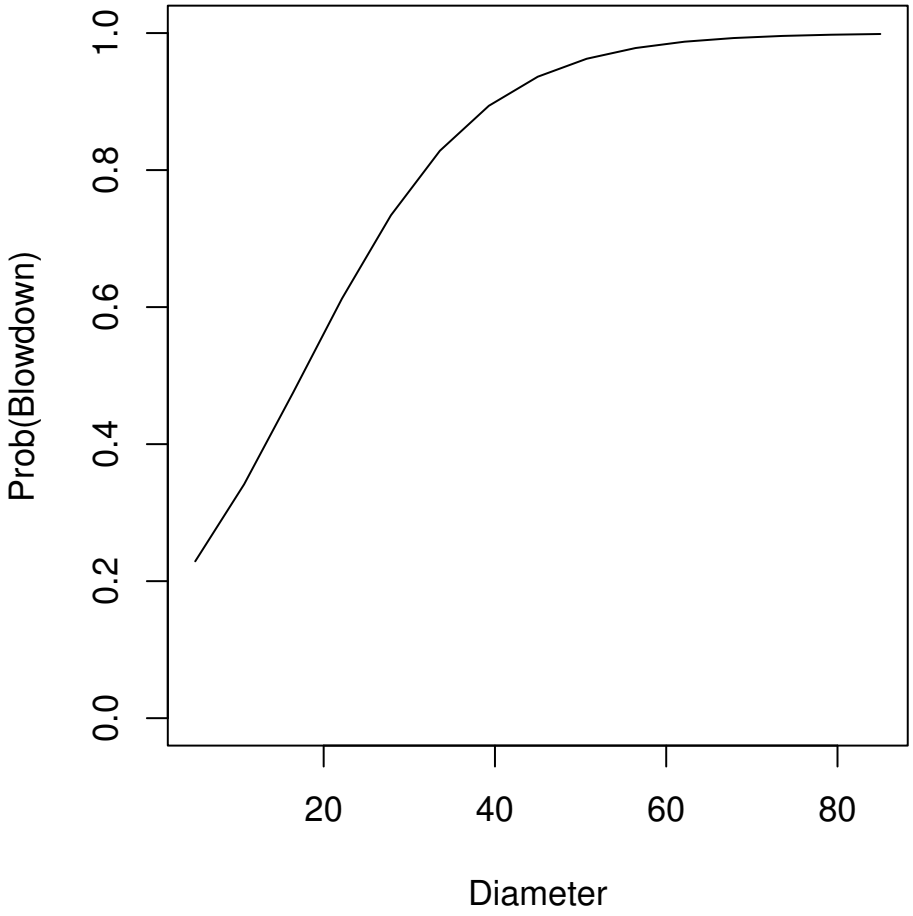
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-5.9971951	0.3748409	-15.999	< 2e-16	***
Intensity	4.6288861	0.2128451	21.748	< 2e-16	***
SpeciesBA	-2.2427869	0.4935773	-4.544	5.52e-06	***
SpeciesBF	0.0002284	0.1789331	0.001	0.999	
SpeciesBS	0.1672262	0.1517509	1.102	0.270	
SpeciesC	-2.0765125	0.2162343	-9.603	< 2e-16	***
SpeciesJP	1.0399651	0.1787626	5.818	5.97e-09	***
SpeciesPB	-1.7235679	0.1864623	-9.244	< 2e-16	***
SpeciesRM	-1.7956738	0.3019339	-5.947	2.73e-09	***
SpeciesRP	0.0031381	0.4131721	0.008	0.994	
logDiameter	1.5813423	0.1114597	14.188	< 2e-16	***

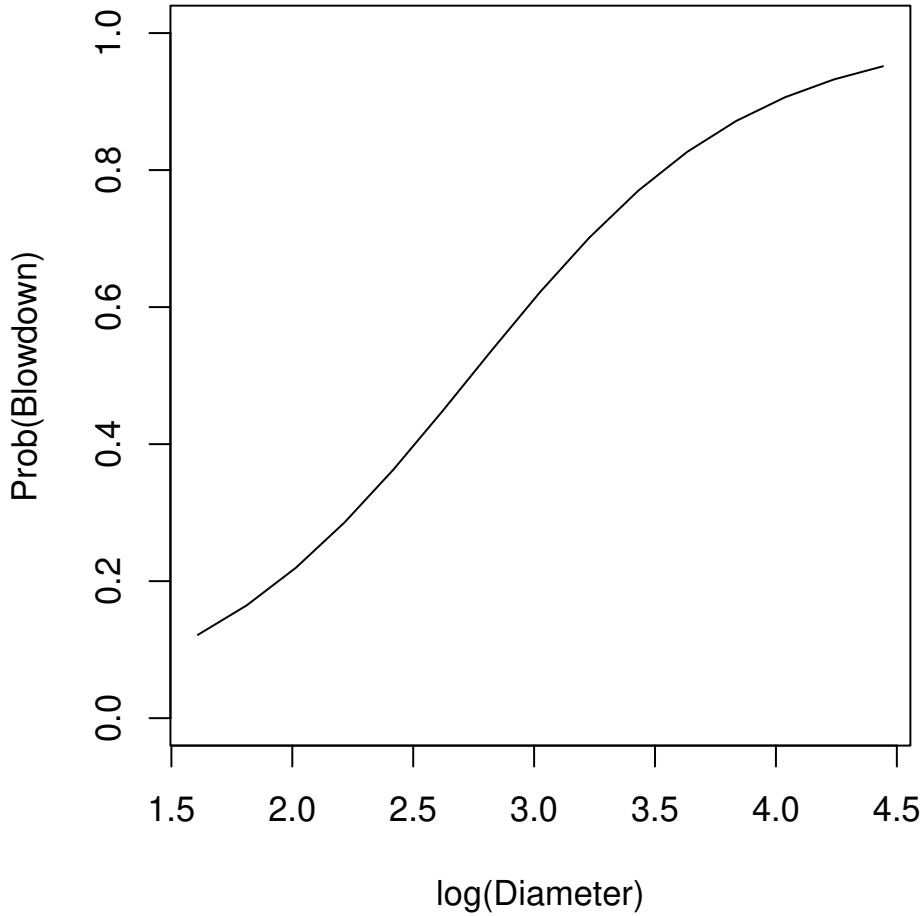
$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 x$

$\log\left(\frac{p_2}{1-p_2}\right) = \beta_0 + \beta_1 (x+1)$

$\log\left(\frac{p_2}{1-p_2} \Big/ \frac{p_1}{1-p_1}\right) = \beta_1$

Logistic regression on Diameter and log(Diameter)





Homework 4
(due 11:59 PM Tue Oct 24)

1. Show that the likelihood function for the Box-Cox transformation on page 86 is invariant to scale multiplication (e.g., lbs $\rightarrow$ kg).

2. Follow the steps on page 86 to reproduce the plot on page 87. Submit your R code.

UMASS aids study

- Data from randomized trials on treatment for drug abuse at University of Massachusetts (Hosmer et al., 2013); data in `uis.csv` on Canvas
- Main goal to compare treatment programs of two different durations in reduction of drug abuse and in prevention of high-risk HIV behavior
- Variables are
  1. Subject ID
  2. Age at enrollment
  3. BECK depression score at admission
  4. IV drug use history at admission (1=never, 2=previous, 3=recent)
  5. Number of prior drug treatments
  6. Subject's race (0=white, 1=other)
  7. Treatment assignment (0=short, 1=long)
  8. Treatment site (0=A, 1=B)
  9. Patient's status at end of treatment program (1=remained drug free, 0=otherwise)

R code

```
> z <- read.csv("uis.csv")
> z$SITE <- as.factor(z$SITE)
> z$RACE <- as.factor(z$RACE)
> z$TREAT <- as.factor(z$TREAT)
> z$IVHX <- as.factor(z$IVHX)
> glm(DFREE ~ . - ID,data=z, family='binomial')
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.4111283  0.5983427  -4.030 5.59e-05 ***
AGE          0.0504143  0.0174057   2.896  0.00377 **
BECK         0.0002759  0.0107982   0.026  0.97961
IVHX2       -0.6036962  0.2875974  -2.099  0.03581 *
IVHX3       -0.7336591  0.2549893  -2.877  0.00401 **
NDRUGTX     -0.0615329  0.0256441  -2.399  0.01642 *
RACE1        0.2260262  0.2233685   1.012  0.31159
TREAT1      -0.4424802  0.1992922   2.220  0.02640 *
SITE1        0.1489209  0.2176062   0.684  0.49375
```

Logistic regression with aggregated data: lung cancer example

- File `lungcancer.csv` gives number of lung cancer cases by age group and their population sizes in four Danish cities (data also from `eba1977` in R package `ISwR`)

```
          city       age          pop              cases
 Fredericia:6   40-54:4   Min.   : 509.0   Min.   : 2.000
 Horsens    :6   55-59:4   1st Qu.: 628.0   1st Qu.: 7.000
 Kolding    :6   60-64:4   Median : 791.0   Median :10.000
 Vejle      :6   65-69:4   Mean   :1100.3   Mean   : 9.333
                 70-74:4   3rd Qu.: 954.8   3rd Qu.:11.000
                 75+  :4   Max.   :3142.0   Max.   :15.000
```

- `cases/pop` is rate of lung cancer per capita by age group in each city

lungcancer.csv

```
"city","age","pop","cases"
"Fredericia","40-54",3059,11
"Horsens","40-54",2879,13
"Kolding","40-54",3142,4
"Vejle","40-54",2520,5
"Fredericia","55-59",800,11
"Horsens","55-59",1083,6
"Kolding","55-59",1050,8
"Vejle","55-59",878,7
"Fredericia","60-64",710,11
"Horsens","60-64",923,15
"Kolding","60-64",895,7
"Vejle","60-64",839,10
"Fredericia","65-69",581,10
"Horsens","65-69",834,10
"Kolding","65-69",702,11
:
```

Logistic regression R code for lung cancer data

```
> z <- read.csv("lungcancer.csv")
> z$city <- as.factor(z$city); z$age <- as.factor(z$age)
> glm(cases/pop ~ city + age, weight=pop, family='binomial', data=z)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.6262     0.2008 -28.021  < 2e-16 ***
cityHorsens  -0.3345     0.1827  -1.830   0.0672 .
cityKolding  -0.3764     0.1890  -1.991   0.0465 *
cityVejle    -0.2760     0.1891  -1.459   0.1444
age55-59      1.1070     0.2490   4.445 8.77e-06 ***
age60-64      1.5291     0.2325   6.577 4.81e-11 ***
age65-69      1.7819     0.2305   7.732 1.06e-14 ***
age70-74      1.8727     0.2365   7.918 2.42e-15 ***
age75+        1.4289     0.2512   5.688 1.29e-08 ***
```

$$\log(\texttt{rate}/(1-\texttt{rate})) = -5.6262 - 0.3345 \times I(\texttt{Horsens}) + \ldots + 1.4289 \times I(\texttt{age75+})$$

Poisson regression

- Recall that a Poisson variable Y with mean μ has density

$$P(Y = y) = \exp(-\mu)\mu^y/y!$$

- Poisson regression assumes that $\log \mu = \mathbf{X}\boldsymbol{\beta}$, or equivalently, $\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
- Likelihhod function is

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\exp\{-\exp(\mathbf{X}_i\boldsymbol{\beta})\}\exp(y_i\mathbf{X}_i\boldsymbol{\beta})}{y_i!}$$

- R function for fitting Poisson regression is `glm` with `family='poisson'`

Crab data

The data file `crab.csv` contains observations on the following variables for 173 female crabs:

color. color of crab (2=light medium, 3=medium, 4=dark medium, 5=dark)

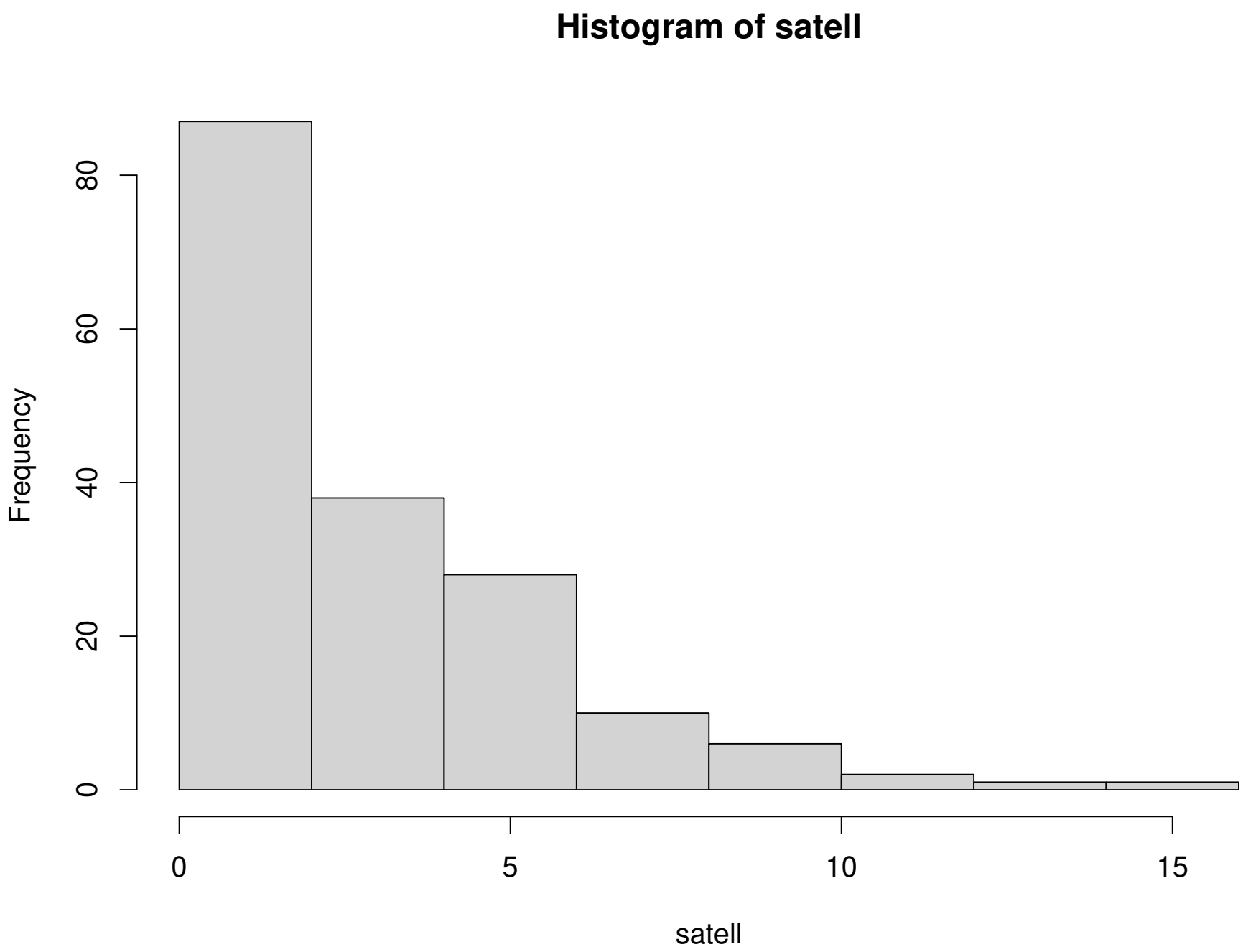
spine. spine condition (1=both good, 2=one worn or broken, 3=both worn or broken)

width. crab carapace width in cm

satell. number of satellites

weight. crab weight in in grams

y. whether crab has satellites (1=yes, 0=no)


Histogram of satell
Frequency
0
20
40
60
80
0
5
10
15
satell

Regression of `satell` on covariates ignoring `y`

```
> z <- read.csv("crab.csv")
> z$color <- as.factor(z$color)
> z$spine <- as.factor(z$spine)
> glm(satell ~ . - y, data=z, family='poisson')
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.3618003  0.9665506  -0.374  0.70817
color3      -0.2648512  0.1681107  -1.575  0.11515
color4      -0.5137051  0.1953624  -2.629  0.00855 **
color5      -0.5308601  0.2269157  -2.339  0.01931 *
spine2      -0.1503718  0.2135754  -0.704  0.48139
spine3       0.0872826  0.1199287   0.728  0.46674
width        0.0167487  0.0489197   0.342  0.73207
weight       0.0004965  0.0001663   2.986  0.00283 **
---
    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 549.59  on 165  degrees of freedom
```

highly correlated
所以一个 significant
另一个 not
weight 比 width
还要紧着，算时
p值小三个数量级

Poisson regression with offset

- Let Y be the number of cases in a population of size n and let $\mu = E(Y)$
- Y is binomial but it is approximately Poisson if n is large
- The model $\log(\mu/n) = \mathbf{X}\boldsymbol{\beta}$ can be written as a Poisson regression model

$$\log(\mu) = \mathbf{X}\boldsymbol{\beta} + \log(n)$$

- $\log(n)$ is called an offset variable; it acts as a predictor variable with regression coefficient fixed at 1

Insurance data (from `MASS` package)

District. district of residence of policyholder (1–4): 4 corresponds to major cities

Group. ordered factor: group of car with levels $<$1 litre, 1–1.5 litre, 1.5–2 litre, $>$2 litre

Age. ordered factor: age of insured in 4 groups labelled $<$25, 25–29, 30–35, $>$35

Holders. numbers of policy holders

Claims. numbers of claims

Insurance data

	District	Group	Age	Holders	Claims
1	1	<1l	<25	197	38
2	1	<1l	25-29	264	35
3	1	<1l	30-35	246	20
4	1	<1l	>35	1680	156
5	1	1-1.5l	<25	284	63
6	1	1-1.5l	25-29	536	84
7	1	1-1.5l	30-35	696	89
8	1	1-1.5l	>35	3582	400
9	1	1.5-2l	<25	133	19
10	1	1.5-2l	25-29	286	52
11	1	1.5-2l	30-35	355	74
12	1	1.5-2l	>35	1640	233
13	1	>2l	<25	24	4
14	1	>2l	25-29	71	18
15	1	>2l	30-35	99	19
16	1	>2l	>35	452	77
:					

$\log\left(\frac{\mu}{n}\right) = X\beta$

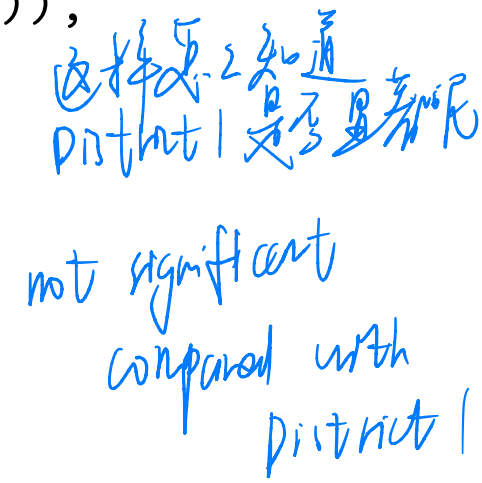
$\frac{\mu}{n} = e^{X\beta}$

$\log\left(\frac{E(\text{Claim})}{\text{Holders}}\right)$

$= \log(\mu) + \log(\text{Holders})$

R code for insurance data

```
> library('MASS')
> Insurance$Group <- factor(Insurance$Group,ordered=FALSE)
> Insurance$Age <- factor(Insurance$Age,ordered=FALSE)
> glm(Claims ~ District+Group+Age+offset(log(Holders)),
              data=Insurance,family=poisson)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.82174    0.07679 -23.724  < 2e-16 ***
District2    0.02587    0.04302   0.601 0.547597
District3    0.03852    0.05051   0.763 0.445657
District4    0.23421    0.06167   3.798 0.000146 ***
Group1-1.5l  0.16134    0.05053   3.193 0.001409 **
Group1.5-2l  0.39281    0.05500   7.142 9.18e-13 ***
Group>2l     0.56341    0.07232   7.791 6.65e-15 ***
Age25-29    -0.19101    0.08286  -2.305 0.021149 *
Age30-35    -0.34495    0.08137  -4.239 2.24e-05 ***
Age>35      -0.53667    0.06996  -7.672 1.70e-14 ***
```



Fitting lung cancer data with Poisson regression (Poisson approximation to binomial)

```
> z <- read.csv("lungcancer.csv")
> z$city <- as.factor(z$city); z$age <- as.factor(z$age)
> glm(cases ~ city + age, offset=log(pop), family='poisson', data=z)
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-5.6321	0.2003	-28.125	< 2e-16	***
cityHorsens	-0.3301	0.1815	-1.818	0.0690	.
cityKolding	-0.3715	0.1878	-1.978	0.0479	*
cityVejle	-0.2723	0.1879	-1.450	0.1472	
age55-59	1.1010	0.2483	4.434	9.23e-06	***
age60-64	1.5186	0.2316	6.556	5.53e-11	***
age65-69	1.7677	0.2294	7.704	1.31e-14	***
age70-74	1.8569	0.2353	7.891	3.00e-15	***
age75+	1.4197	0.2503	5.672	1.41e-08	***

rate in logistic regression

$$\log(\texttt{cases}/\texttt{pop}) = -5.6321 - 0.3301 \times I(\texttt{Horsens}) + \ldots + 1.4197 \times I(\texttt{age75+})$$

Linear: Linear in β_j $y = \beta_0 + \beta_1 e^{\beta_2 x} + \varepsilon$ not linear $y = \beta_0 + \beta_1 \log x + \varepsilon$ ✓

Survival and hazard functions

- Let U be survival time of a subject with density $f(u)$ and cdf

$$F(u) = P(U \leq u) = \int_{-\infty}^{u} f(x)\, dx$$

- **Survival probability function** is $S(u) = P(U > u) = 1 - F(u)$
- **Hazard rate** (instantaneous rate of death) at time u is

$$\lambda(u) = f(u)/S(u) \tag{4}$$

- **Cumulative hazard** function at time u is

$$\Lambda(u) = \int_{-\infty}^{u} \lambda(x)\, dx = \int_{-\infty}^{u} f(x)/S(x)\, dx = -\log(S(u))$$

and

$$S(u) = \exp(-\Lambda(u)) \tag{5}$$

Cox proportional hazards (PH) regression

- Assume that given covariate vector $\mathbf{x}$, $\lambda(u, \mathbf{x}) = \lambda_0(u) \exp(\mathbf{x}'\boldsymbol{\beta})$, where $\lambda_0(u)$ is an unknown **baseline hazard** function
- **log hazard ratio** (also called log relative risk) does not depend on u:

$$\log\{\lambda(u, \mathbf{x})/\lambda_0(u)\} = \mathbf{x}'\boldsymbol{\beta}$$

- Hazards of two subjects with covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ are **proportional**:

$$\lambda(u, \mathbf{x}_2)/\lambda(u, \mathbf{x}_1) = \exp\{(\mathbf{x}_2 - \mathbf{x}_1)'\boldsymbol{\beta}\}$$

- Let $\mathbf{x}_1 = (a_1, a_2, \ldots, a_k)$ and $\mathbf{x}_2 = (b_1, b_2, \ldots, b_k)$, such that $b_j = a_j + 1$ and $b_i = a_i$ for $i \neq j$. Then

$$\beta_j = \log\left\{\frac{\lambda(u, \mathbf{x}_2)}{\lambda(u, \mathbf{x}_1)}\right\}$$

  i.e., β_j = log relative risk due to increasing jth element of $\mathbf{x}$ by 1

- Let $S(u, \mathbf{x})$, $\lambda(u, \mathbf{x})$ denote survival probability and hazard functions at $\mathbf{x}$
- PH model assumes $\lambda(u, \mathbf{x}) = \lambda_0(u) \exp(\mathbf{x}'\boldsymbol{\beta})$
- Let $\Lambda_0(u) = \int_{-\infty}^{u} \lambda_0(z)\, dz$ be the **baseline cumulative hazard** and let $\Lambda(u, \mathbf{x}) = \int_{-\infty}^{u} \lambda(z, \mathbf{x})\, dz = \Lambda_0(u) \exp(\mathbf{x}'\boldsymbol{\beta})$ denote the **cumulative hazard**
- Then by (4) and (5),

$$\begin{aligned} S(u, \mathbf{x}) &= \exp\{-\Lambda(u, \mathbf{x})\} \\ &= \exp\{-\Lambda_0(u) \exp(\mathbf{x}'\boldsymbol{\beta})\} \\ f(u, \mathbf{x}) &= \lambda(u, \mathbf{x}) S(u, \mathbf{x}) \\ &= \lambda_0(u) \exp\{\mathbf{x}'\boldsymbol{\beta} - \Lambda_0(u) \exp(\mathbf{x}'\boldsymbol{\beta})\} \end{aligned}$$

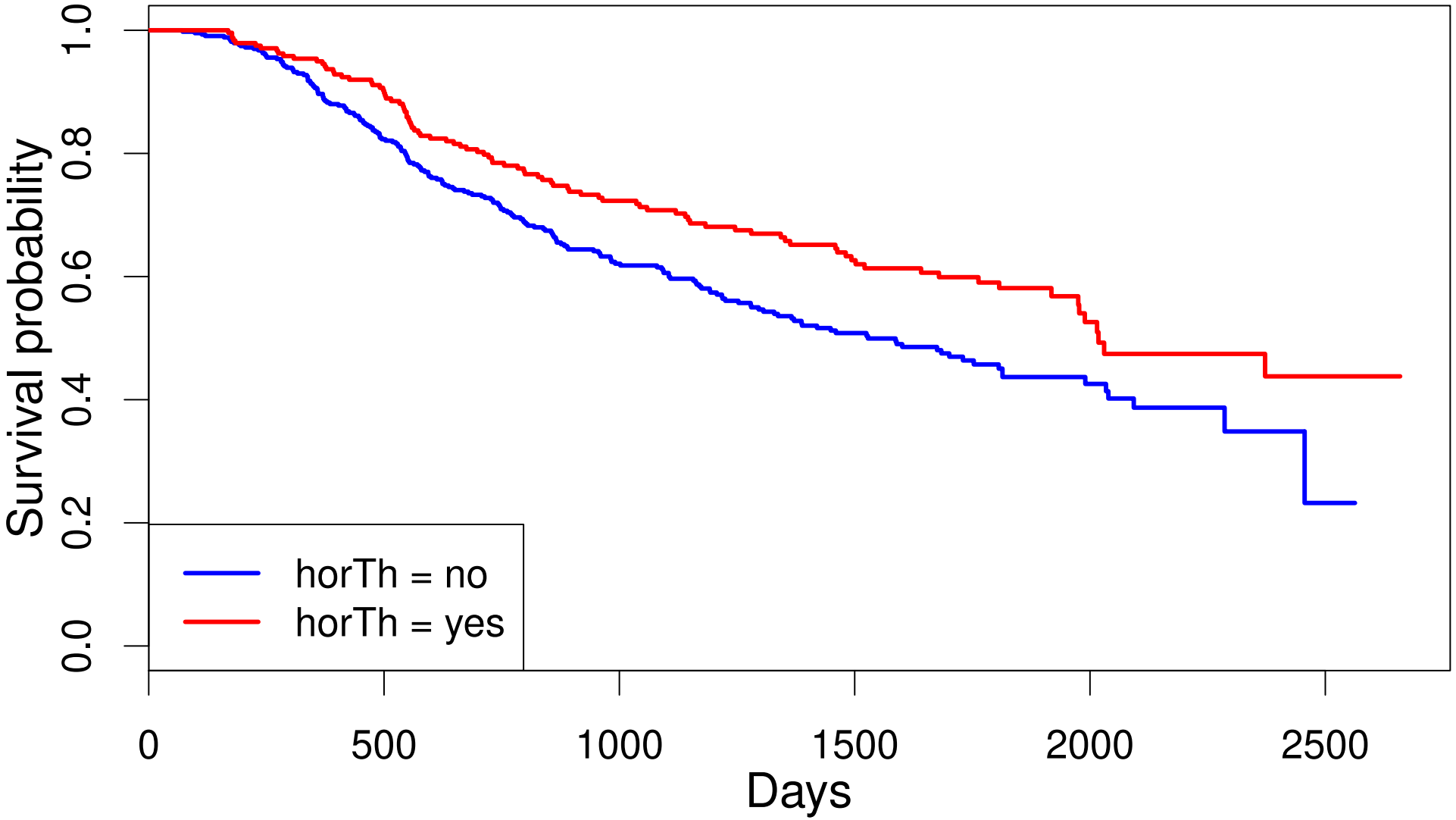
- Let u_i and $\mathbf{x}_i$ denote the survival time and covariate vector of subject i
- Let s_i be an independent observation from some censoring distribution and let $\delta_i = I(u_i < s_i)$ be the event (e.g., death) indicator
- Observed data for subject i is $(y_i, \delta_i, \mathbf{x}_i)$, where $y_i = \min(u_i, s_i)$
- $\hat{\boldsymbol{\beta}}$ maximizes the log-likelihood

$$\begin{aligned}
&\log\left\{\prod_{i=1}^{n} f(y_i, \mathbf{x}_i)^{\delta_i} S(y_i, \mathbf{x}_i)^{1-\delta_i}\right\} \\
&= \sum_{i=1}^{n} \delta_i \log\{f(y_i, \mathbf{x}_i)/S(y_i, \mathbf{x}_i)\} + \sum_{i=1}^{n} \log S(y_i, \mathbf{x}_i) \\
&= \sum_{i=1}^{n} \delta_i \{\log \lambda_0(y_i) + \mathbf{x}_i'\boldsymbol{\beta}\} - \sum_{i=1}^{n} \Lambda_0(y_i) \exp(\mathbf{x}_i'\boldsymbol{\beta}) \\
&= \sum_{i=1}^{n} \delta_i \{\log \Lambda_0(y_i) + \mathbf{x}_i'\boldsymbol{\beta}\} - \sum_{i=1}^{n} \Lambda_0(y_i) \exp(\mathbf{x}_i'\boldsymbol{\beta}) \\
&\quad + \sum_{i=1}^{n} \delta_i \log\{\lambda_0(y_i)/\Lambda_0(y_i)\}
\end{aligned}$$

Breast cancer example

- Randomized clinical trial of 672 subjects with primary node positive breast cancer (Schumacher et al., 1994); 14 subjects with censored times less than smallest uncensored time excluded; data from TH.data R package
- Response is recurrence-free survival time (8–2659 days, 299 uncensored, 387 censored)
- Eight predictor variables:
  1. **horTh** (hormone therapy, yes/no)
  2. **age** (21–80 years)
  3. **tsize** (tumor size, 3–120 mm)
  4. **pnodes** (number of positive lymph nodes, 1–51)
  5. **progrec** (progesterone receptor status, 0–2380 fmol)
  6. **estrec** (estrogen receptor status, 0–1144 fmol)
  7. **menostat** (menopausal status, pre/post)
  8. **tgrade** (tumor grade, 1, 2, 3)

Kaplan-Meier survival curves



R code for Kaplan-Meier plot

```
library(survival)
library(MASS)
z <- read.csv("cancerdata.csv")
leg.txt <- c("horTh = no","horTh = yes")
leg.col <- c("blue","red")
leg.lty <- rep(1,2)
y <- z$time
stat <- z$death
treat <- z$horTh
fit.bytreat <- survfit(Surv(y,stat) ~ treat, conf.type="none")
plot(fit.bytreat,conf.int=FALSE,col=leg.col,mark.time=FALSE,lwd=3)
mtext("Survival probability",side=2,line=2,cex=1.7)
mtext("Days",side=1,line=2,cex=1.7)
legend("bottomleft",legend=leg.txt,lty=leg.lty,col=leg.col,lwd=3)
```

0 → 1, X → X+1

R code for Cox PH model

```
> summary(coxph(Surv(time, death)~ ., data=z))
```

horThyes = 1. less likely die

still significant

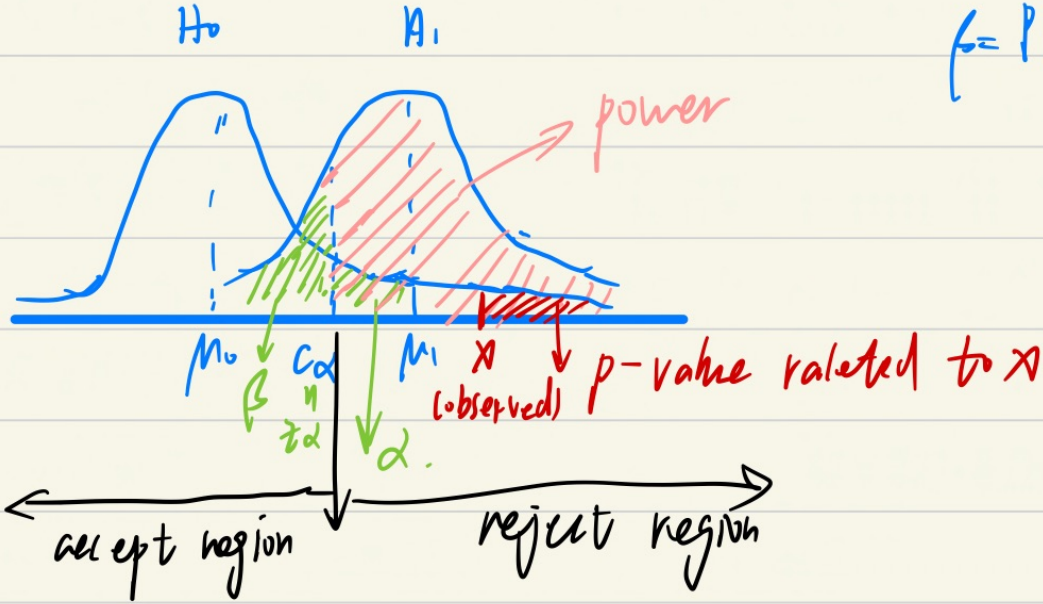
	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
horThyes	-0.3372029	0.7137640	0.1289618	-2.615	0.008929	**
age	-0.0093924	0.9906516	0.0092733	-1.013	0.311136	
menostatPre	-0.2672772	0.7654609	0.1833366	-1.458	0.144882	
tsize	0.0077164	1.0077463	0.0039497	1.954	0.050739	.
tgrade	0.2802894	1.3235128	0.1060553	2.643	0.008221	**
pnodes	0.0498939	1.0511596	0.0074094	6.734	1.65e-11	***
progrec	-0.0022378	0.9977647	0.0005758	-3.887	0.000102	***
estrec	0.0001674	1.0001674	0.0004477	0.374	0.708431	

similar pattern
correlated

partial effect

Consider $X \sim (\mu, 1^2)$, $H_0: \mu = 0$, $H_1: \mu = 1$, $\alpha = P[\text{Type I error}]$

$\beta = P[\text{Type II error}]$



若 σ^2 变的情况.