清华大学统计学辅修课程

Linear Regression Analysis

Lecture 11- Remedies & Single Factor Analysis of Variance

周在莹

清华大学统计学研究中心

http://www.stat.tsinghua.edu.cn

Topic 1: Remedies

Outline

- Review regression diagnostics
- Remedial measures
  - Weighted regression
  - Ridge regression
  - Robust regression
  - Bootstrapping

Regression Diagnostics Summary

- Check normality of the residuals with normal quantile plot or histogram
- Plot the residuals versus predicted values, versus each of the *X*'s and (when appropriate) versus time/space
- Examine the partial regression plots
  - Use the graphics smoother to see if there appears to be a curvilinear pattern
- Examine
  - The studentized deleted residuals
  - The hat matrix diagonals
  - DFFITS, Cook's D, and the DFBETAS
- Check observations that are extreme on these measures relative to the other observations
- Examine the tolerance for each *X*
- If there are variables with low tolerance, you need to do some model building
- Recode variables
  - Variable selection

Remedial Measures

- Weighted least squares
- Ridge regression
- Robust regression
- Nonparametric regression
- Bootstrapping

Maximum Likelihood

- $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, $f_i = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{1}{2}\left(\frac{Y_i-\beta_0-\beta_1 X_i}{\sigma}\right)^2\right\}$
- Likelihood function: $L = f_1 \cdot f_2 \ldots f_n$
- Find $\beta_0, \beta_1, \sigma^2$ which maximize L

- What if Y_i's have different but known variances, $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$?

$$f_i = \frac{1}{\sqrt{2\pi}\sigma_i} exp\left\{-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma_i}\right)^2\right\}$$

- Likelihood function: $L = f_1 \cdot f_2 \ldots f_n$
- Find β_0, β_1 which maximize L

Weighted Regression

- Maximization of L with respect to β's is equivalent to minimization of

$$\sum_{i=1}^{n}\left(\frac{Y_i-\beta_0-\beta_1X_{i,1}-\cdots-\beta_{p-1}X_{i,p-1}}{\sigma_i}\right)^2$$

- Weight of each case is $w_i=\frac{1}{\sigma_i^2}$
- Weighted Least Squares
- Least squares problem is to minimize $\sum w_i e_i^2$, the sum of w_i times the squared residual for case i
- Computation is easy…specify the *weights* statement in *lm*()

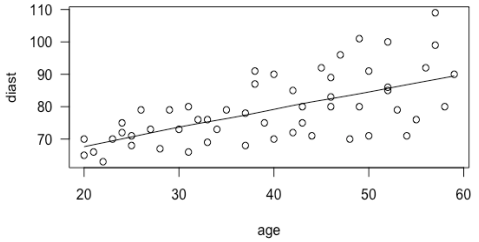
Unknown Variances? Determination of Weights

- $W^{\frac{1}{2}}Y = W^{\frac{1}{2}}X\beta + W^{\frac{1}{2}}\varepsilon,\ b_w = (X'WX)^{-1}(X'WY)$

  where W is a diagonal matrix of the weights, $W = \text{diag}(w_1, w_2, \ldots, w_n)$

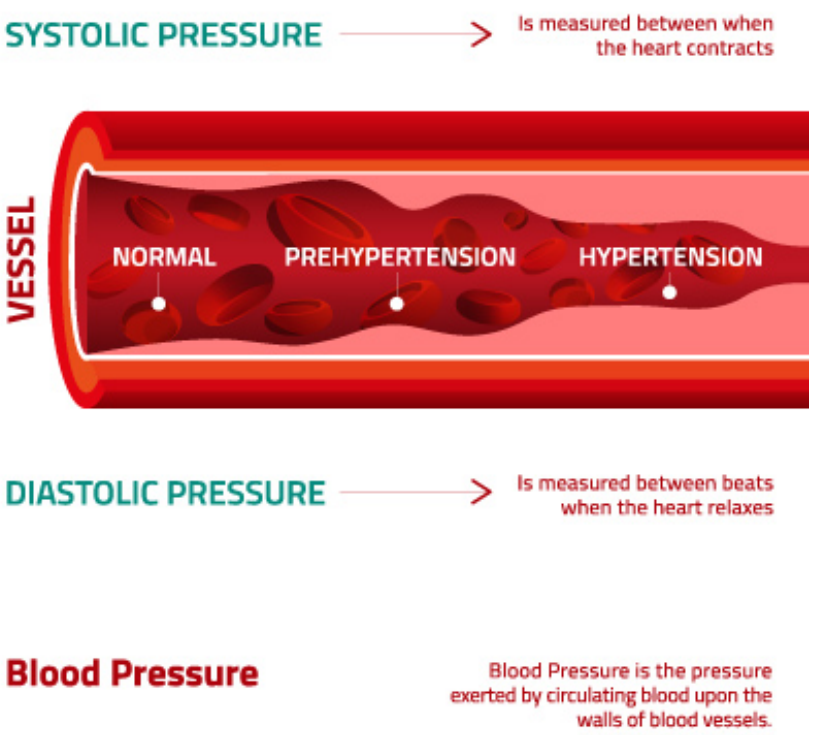
- The problem in practice now becomes determining the weights
- Find a relationship between the absolute residual and another variable and use this as a model for the standard deviation
  - Similar approach using the squared residual to model the variance
- Or use replicates or near replicates(grouped data or approximately grouped data) to estimate the variance for all cases in the group
- With a model for the standard deviation or the variance, we can approximate the optimal weights
- Optimal weights are proportional to the inverse of the variance

$$w_i = \frac{1}{\sigma_i^2}$$



Blood Example KNNL p 427

- Y is diastolic blood pressure
- X is age
- n = 54 healthy adult women aged 20 to 60 years old

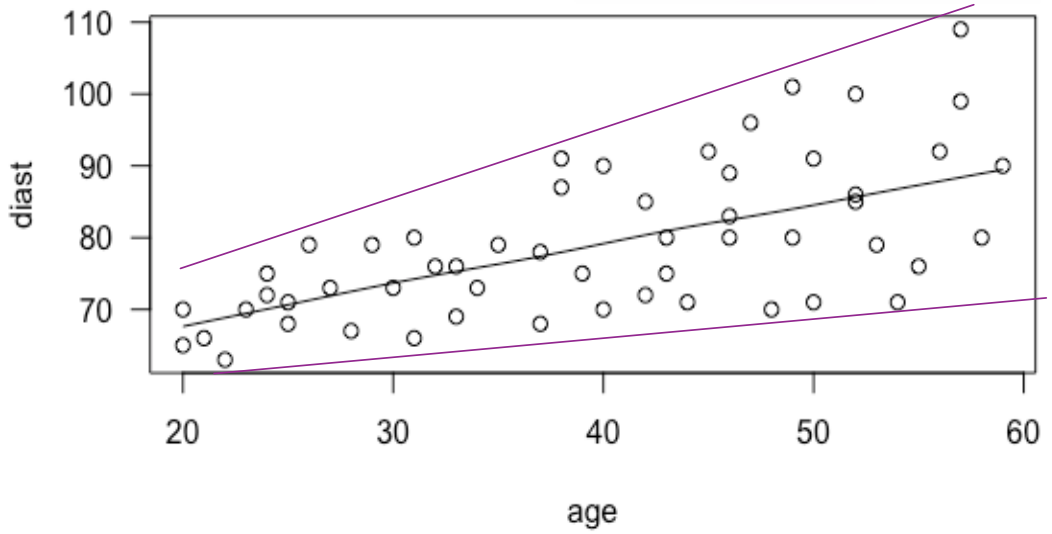


EDA

- Get the data and check it
- Plot the relationship
- Diastolic bp *vs* age

```
> a1 = read.table("CH11tA01.txt")
> colnames(a1) = c("age", "diast")
> View(a1)
```

```
> scatter.smooth(a1$age, a1$diast,
                 xlab = 'age', ylab = 'diast', las = 1)
```



- Strong linear relationship, no skewness but non-constant variance

Analysis

- Run the regression
- Estimators still unbiased but no longer have minimum variance

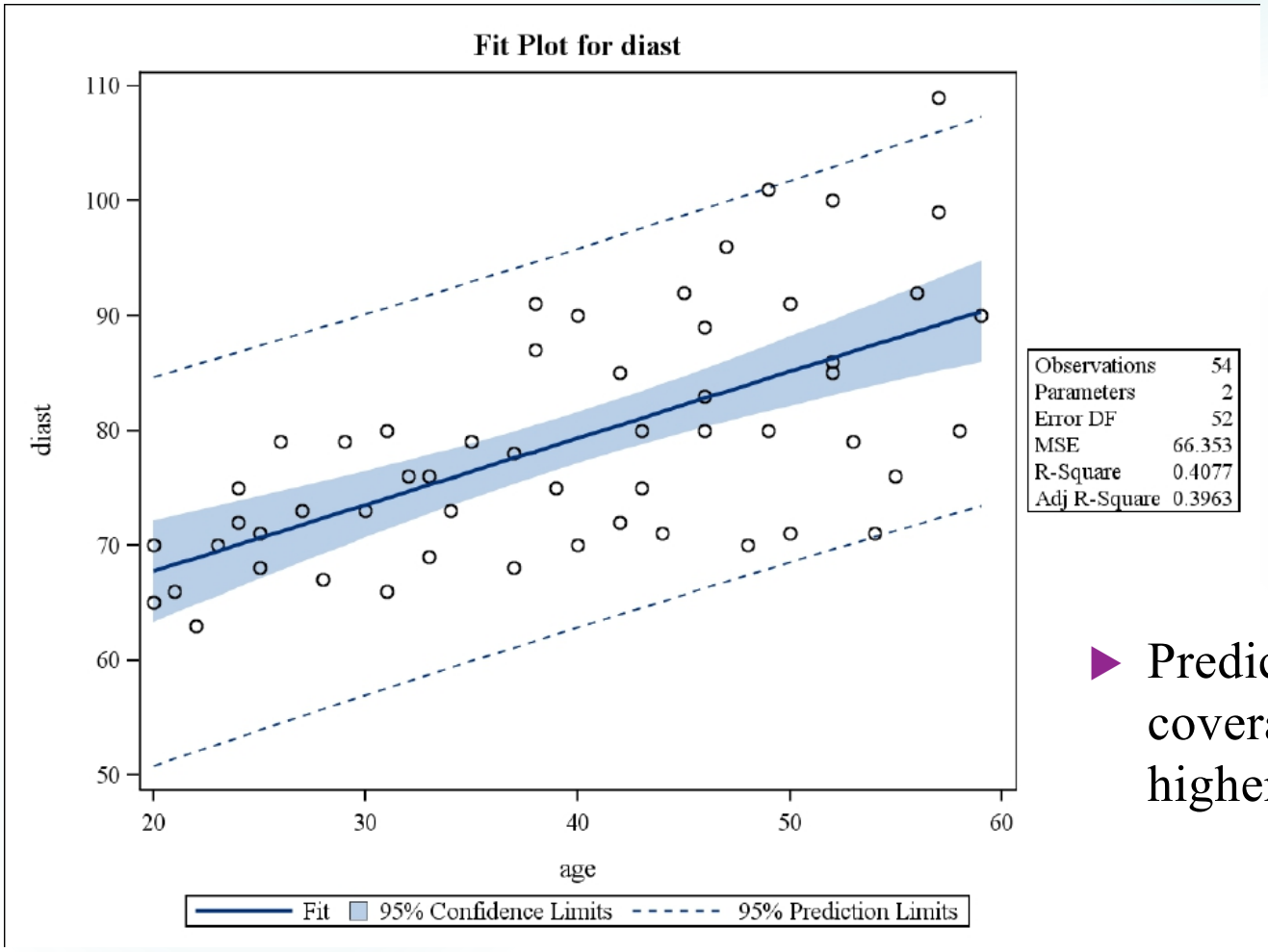
```
fit <- lm(diast ~ age, data = a1)
summary(fit)
plot(fit)
```

```
Coefficients:
             Estimate   Std. Error  t value Pr(>|t|)
(Intercept) 56.15693    3.99367  14.061  < 2e-16 ***
age          0.58003    0.09695   5.983   2.05e-07 ***

Residual standard error: 8.146 on 52 degrees of freedom
Multiple R-squared:  0.4077,  Adjusted R-squared:  0.3963
F-statistic: 35.79 on 1 and 52 DF,  p-value: 2.05e-07
```

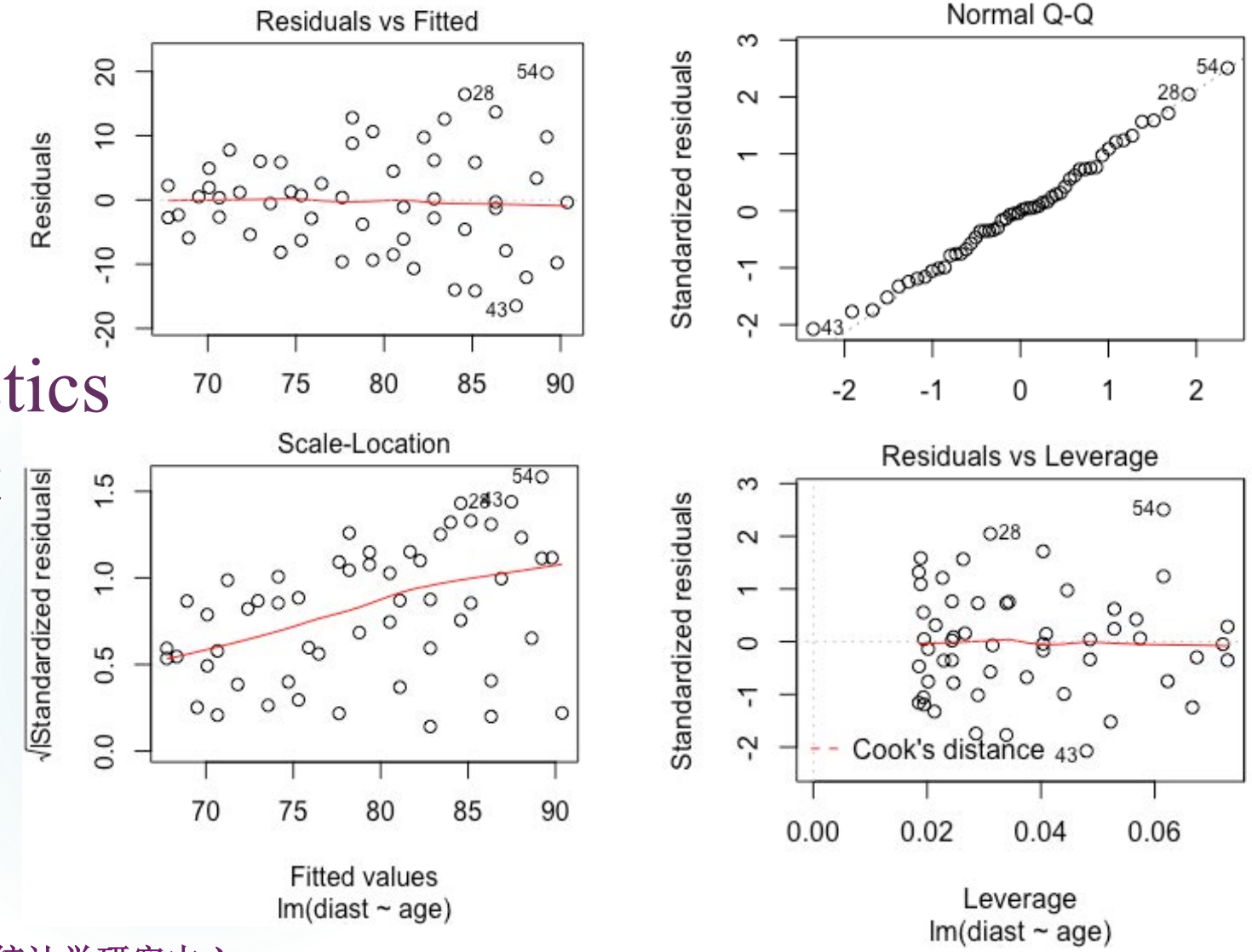
```
Analysis of Variance Table

Response: diast
             Df Sum Sq Mean Sq F value   Pr(>F)
age          1  2375.0  2374.97  35.793 2.05e-07 ***
Residuals 52  3450.4       66.35
```



- Prediction interval coverage often lower or higher than 95%

Fit Diagnostics for diast

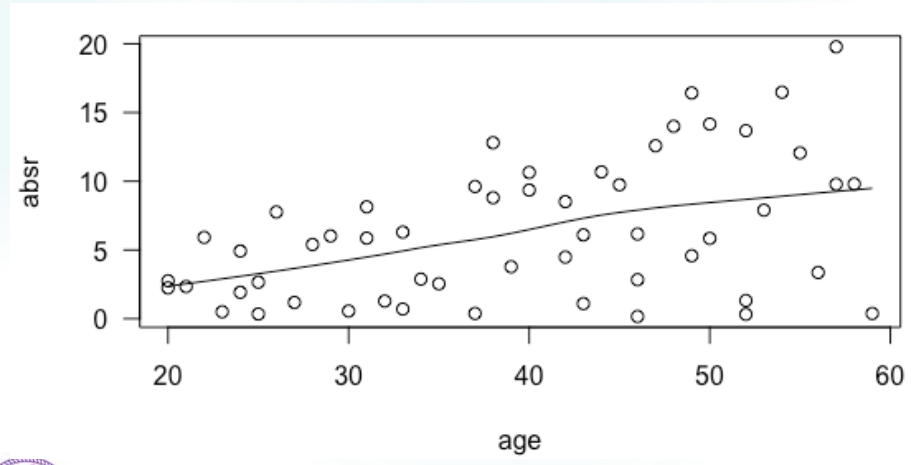


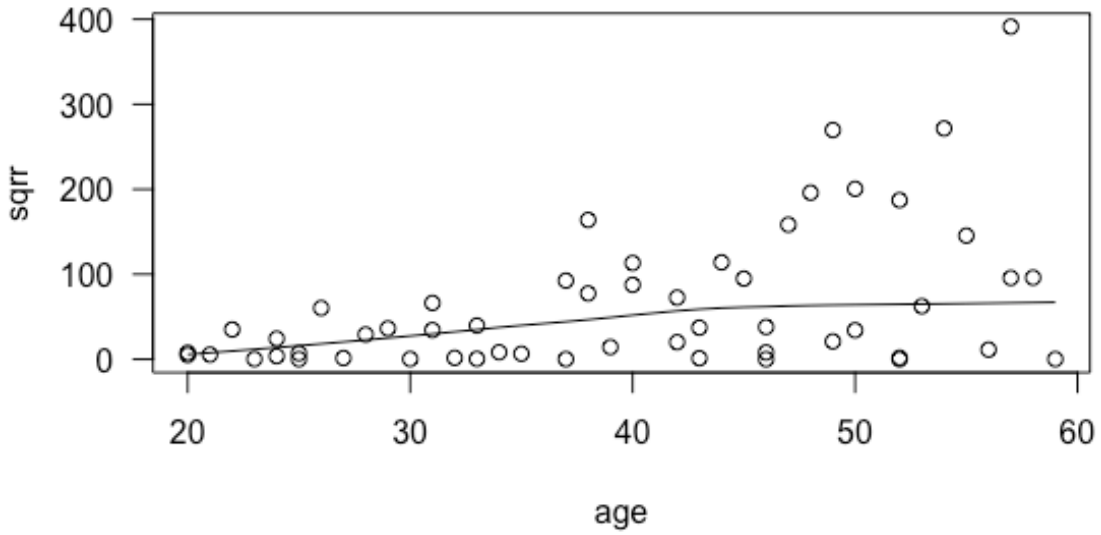
Residual Plots

- Use the output data set a2 to get the absolute and squared residuals
- Generate plots with smoothing
  - Absolute Residual *vs* Age
  - Squared Residuals *vs* Age

```
> a1$resids = fit$residuals
> a1$absr = abs(fit$residuals)
> a1$sqrr =  (fit$residuals)^2
```

```
> plot(absr ~ age, data=a1, las=1)
> lines(lowess(a1$age, a1$absr))
> plot(sqrr ~ age, data=a1, las=1)
> lines(lowess(a1$age, a1$sqrr))
```





Weighted Least Squares

- Model the std dev *vs* age (absolute value of the residual)

  估计值代替

  - Note that *shat* is the predicted/estimated expected standard deviations
- Compute the weights
- Regression with weights

```
> fit2 = lm(absr ~ age, data=a1)
> a1$shat = predict(fit2)
```

$$w_i = \frac{1}{\sigma_i^2}$$

```
> a1$wt = 1/(a1$shat)^2
```

```
> fit3 = lm(diast ~ age, weights = wt, data = a1)
> summary(fit3)
```

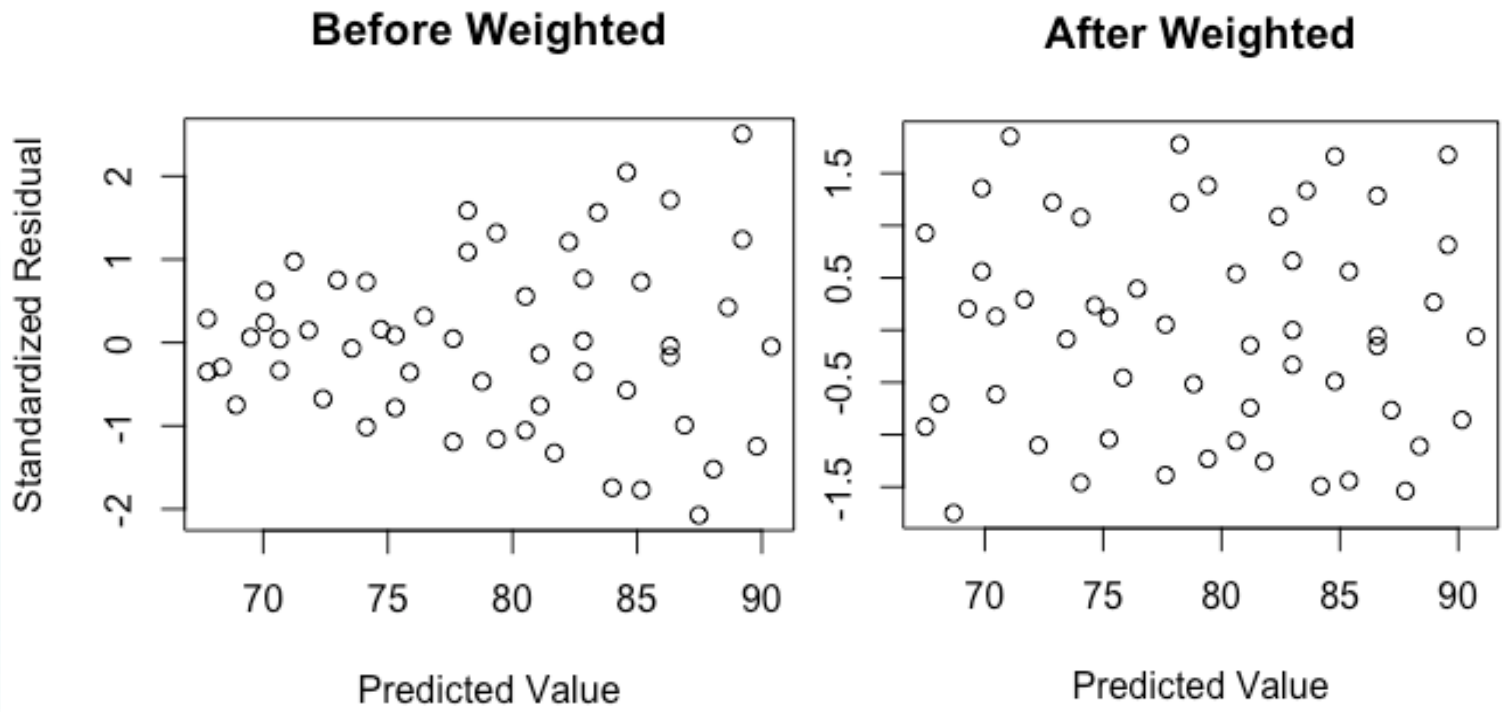
Output Comparison

Compare with before

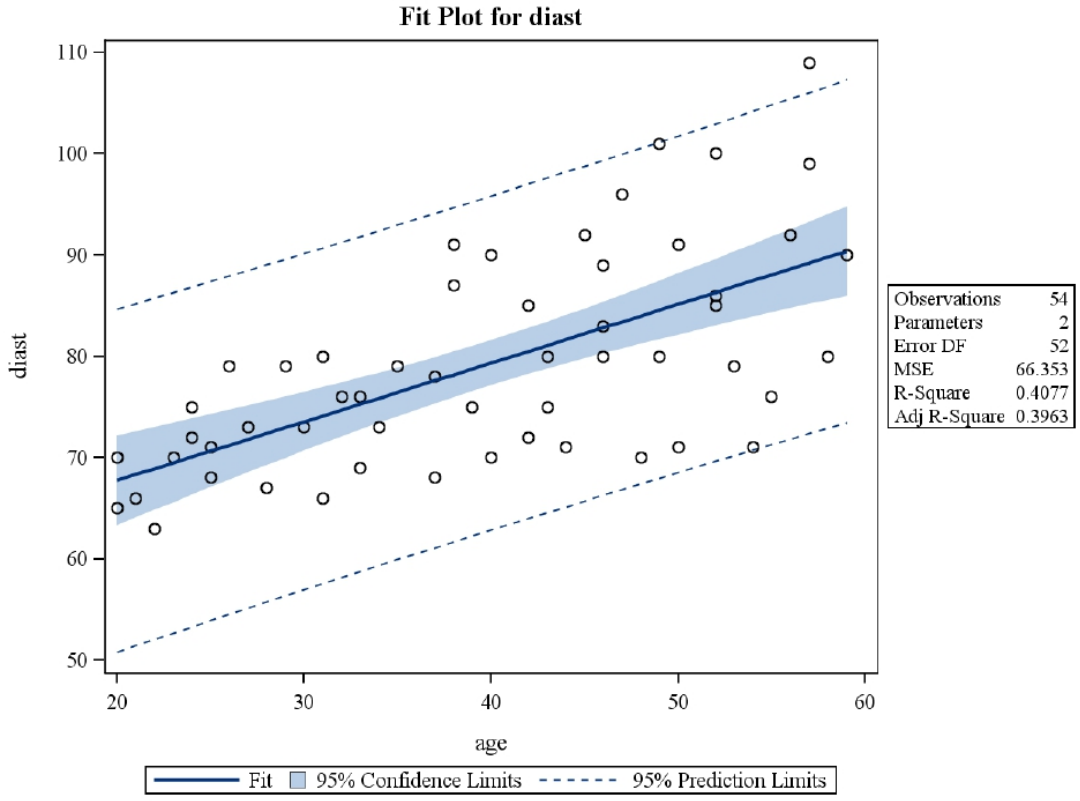
Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 55.56577 2.52092 22.042 < 2e-16 *** age 0.59634 0.07924 7.526 7.19e-10 *** Residual standard error: 1.213 on 52 degrees of freedom Multiple R-squared: 0.5214, Adjusted R-squared: 0.5122 F-statistic: 56.64 on 1 and 52 DF, p-value: 7.187e-10	Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 56.15693 3.99367 14.061 < 2e-16 *** age 0.58003 0.09695 5.983 2.05e-07 *** Residual standard error: 8.146 on 52 degrees of freedom Multiple R-squared: 0.4077, Adjusted R-squared: 0.3963 F-statistic: 35.79 on 1 and 52 DF, p-value: 2.05e-07

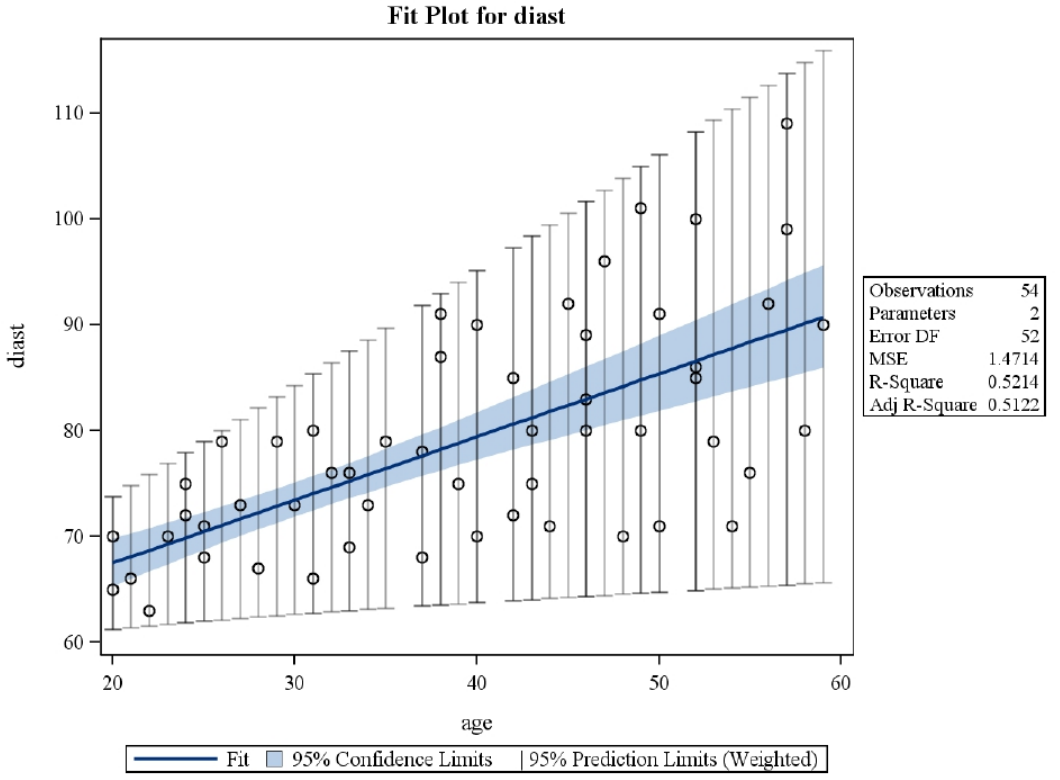
- Reduction in std err of the age coeff

Fit Diagnostics Comparison



Fit Plot Comparison





Ridge Regression

- If $(X'X)$ is difficult to invert (near singular) then approximate by inverting $(X'X + kI)$
- Estimators of coefficients are now biased but more stable
- For some value of k, ridge regression estimator has a smaller mean square error than ordinary least square estimator
- Cross-validation / ridge plots used to determine k

Ridge Regression

- Can express ridge constraint in terms of finding β to minimize:

$$(Y - Z\beta)'(Y - Z\beta) + \lambda \sum \beta_j^2$$

where Z is the standardized X

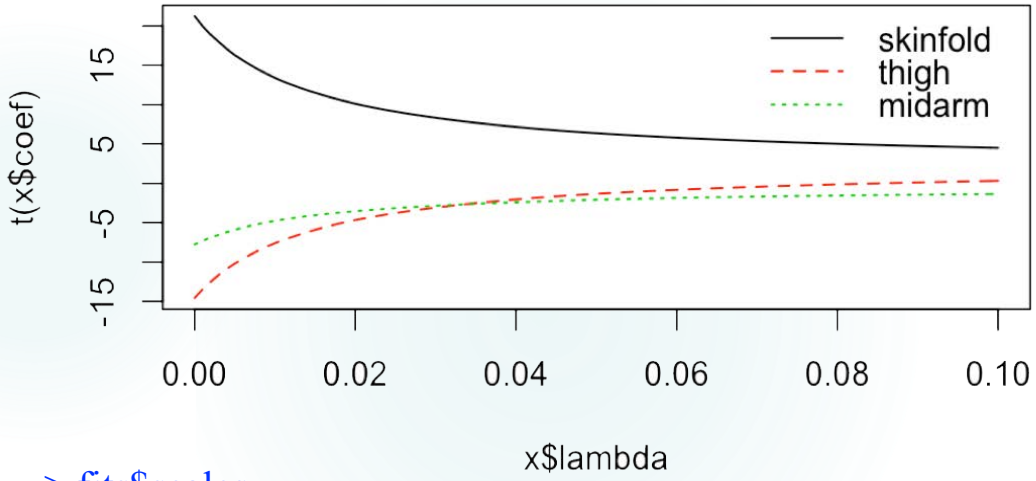
- Note that LASSO is a variation of this approach in which you minimize

$$(Y - Z\beta)'(Y - Z\beta) + \lambda \sum |\beta_j|$$

- Can specify 'lambda' in R

Ridge Regression in R

- Require Package ‘MASS’
- The ridge regression will penalize your coefficients, such that those who are the least efficient in your estimation will "shrink" the fastest




```
#Bodyfat example ”skinfold", "thigh", "midarm", "fat"
> library(MASS)
> fits = lm.ridge(fat ~ skinfold + thigh + midarm,
                  lambda=seq(0, 0.1, 0.001), data=b1)
> plot(fits)
> legend("topright", c("skinfold", "thigh", "midarm"),
                     col = 1:3, lty = 1:3, bty = 'n')
> whichIsBest = which.min(fits$GCV)
> coef(fits)[whichIsBest, ]
```

```
> fits$scales
 skinfold     thigh     midarm
4.896067 5.102068 3.554800
```

λ		skinfold	thigh	midarm
0.019	43.8401126	2.1174933	-0.9597309	-1.0180612
0	117.084695	4.3340920	-2.856847936	-2.1860603

Robust Regression

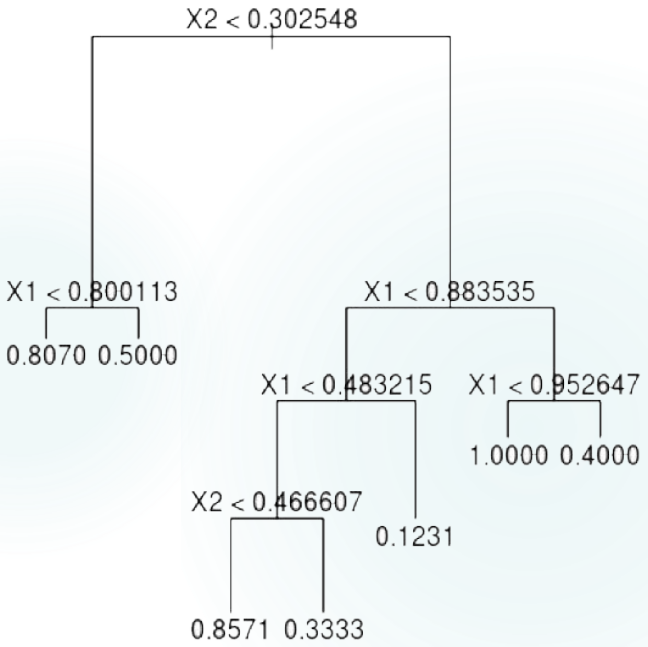
- Basic idea is to have a procedure that is not sensitive to outliers
- Alternatives to least squares, minimize
  - sum of absolute values of residuals
  - median of the squares of residuals
- Do weighted regression with weights based on residuals, and iterate

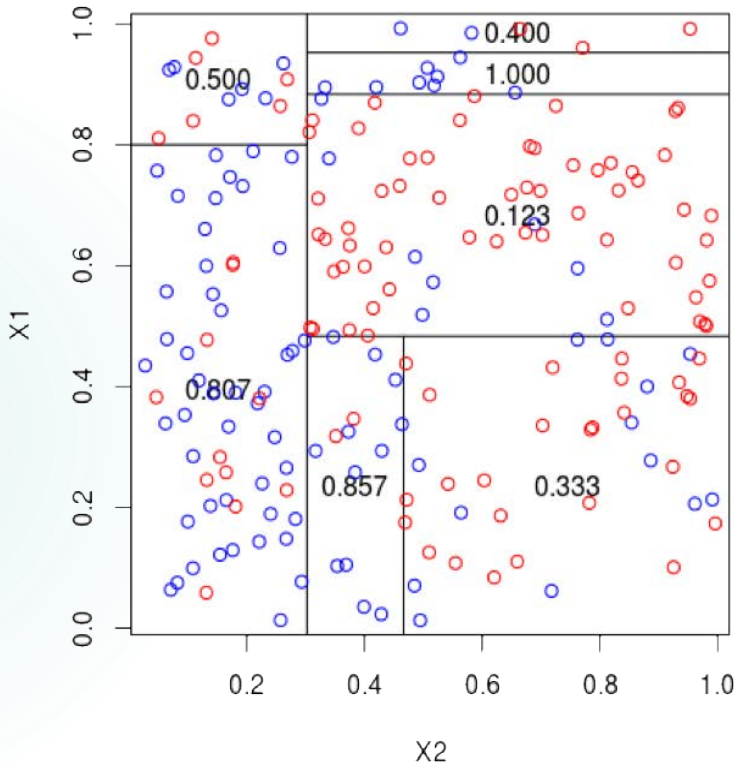
Nonparametric Regression

- Several versions
- Interesting theory
- All versions have some smoothing or penalty parameter
- Local polynomial regression is performed by the standard R functions *lowess*(locally weighted scatterplot smoother, for the simple-regression case) and *loess* (local regression, more generally).
- Simple-regression smoothing-spline estimation is performed by the standard R function *smooth.spline*

Regression Trees

- Standard approach in area of "data mining" replacing multiple regression

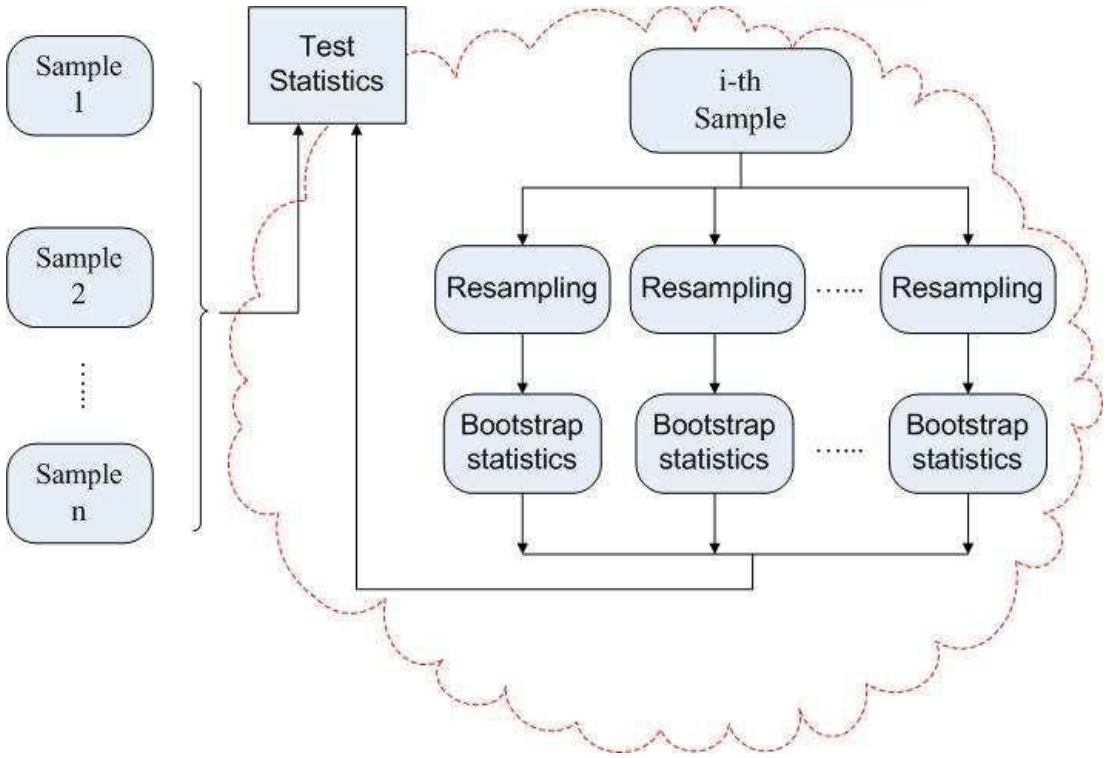




- Basically partition the X space into rectangles
  - Repeatedly split data two nodes based on a single predictor
- Predicted value is mean of responses in rectangle

Bootstrap

- Very important theoretical development that has had a major impact on applied statistics
- Uses resampling to approximate the sampling distribution
- Sample *with* replacement from the data or residuals and repeatedly refit model to get the distribution of the quantity of interest



Background Reading

- We used R program lec11_1.R
- This completes Chapter 11

Topic 2:
Single Factor
Analysis of Variance

Outline

- Single factor Analysis of Variance
  - One set of treatments
    - Cell means model
    - Factor effects model
- Link to linear regression using indicator explanatory variables

One-Way ANOVA

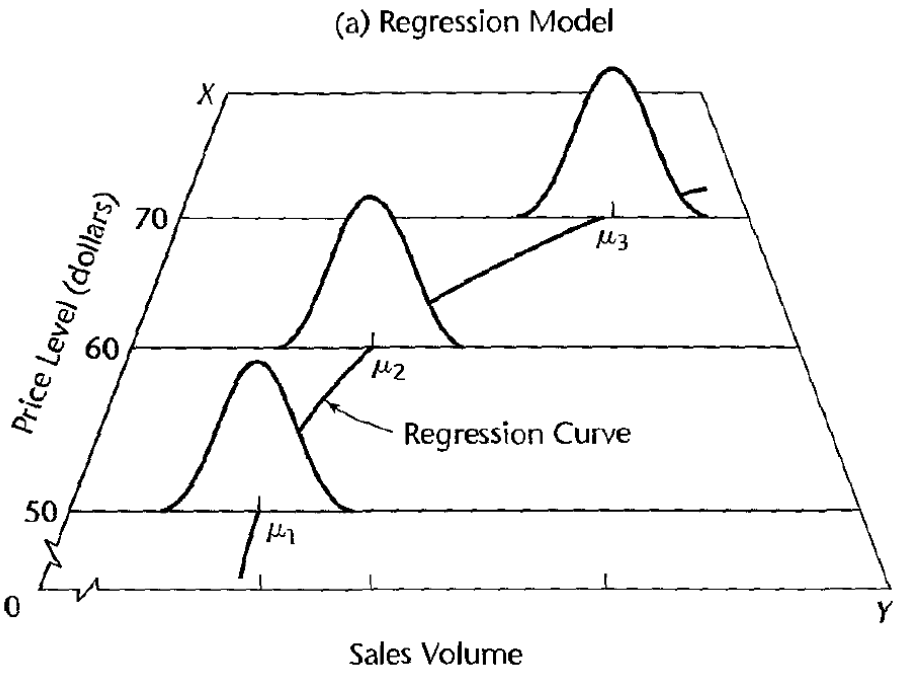
- The response variable Y is continuous
- The explanatory variable X is **categorical**
  - We call it a **factor**
  - The possible values are called **levels**
- This approach is a generalization of the **independent two-sample pooled t-test**
- In other words, it can be used when there are more than two treatments

Data for One-Way ANOVA

- Y is the response variable
- X is the factor (qualitative/discrete)
- r is the number of levels
  - often refer to these levels as groups or treatments

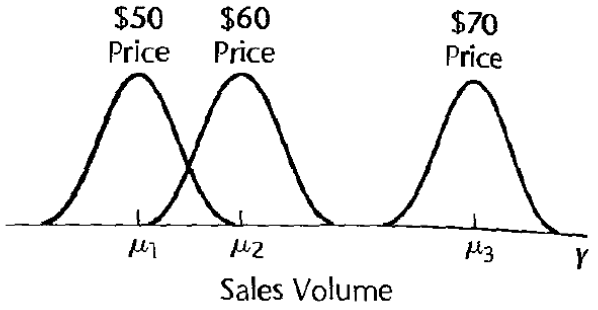
Illustrative Comparison

- ANOVA does not consider the quantitative differences in the three X levels or their statistical relation to expected Y



(b) Analysis of Variance Model

The three X levels are treated as separate populations, each leading to a probability distribution of Y



Model Assumptions & Goals

- ANOVA model assumes that:
  - 1. Each probability distribution is normal
  - 2. Each probability distribution has the same variance
  - 3. The responses for each factor level are random selections from the corresponding probability distribution and are independent of the responses for any other factor level
- ANOVA focuses on the mean responses for the different factor levels
- The analysis of the sample data from the factor level probability distributions usually proceeds in two steps:
  - 1. Determine whether or not the factor level means are the same
  - 2. If the factor level means differ, examine how they differ and what the implications of the differences are

Notation

- For Y_{ij} we use
  - i to denote the level of the factor
  - j to denote the j^{th} observation at factor level i
- $i = 1,..., r$ levels of factor
- $j = 1,..., n_i$ observations for level i of factor X
  - Note that n_i does not need to be the same for each level

Cereal Package Example KNNL P 685

- Y is the number of cases of cereal sold
- X is the design of the cereal package
  - there are four levels for X because there are four different designs
- i = 1 to 4 levels
- j = 1 to n_i stores with design i (n_i = 5, 5, 4, 5)
- Will use n if n_i the same across levels

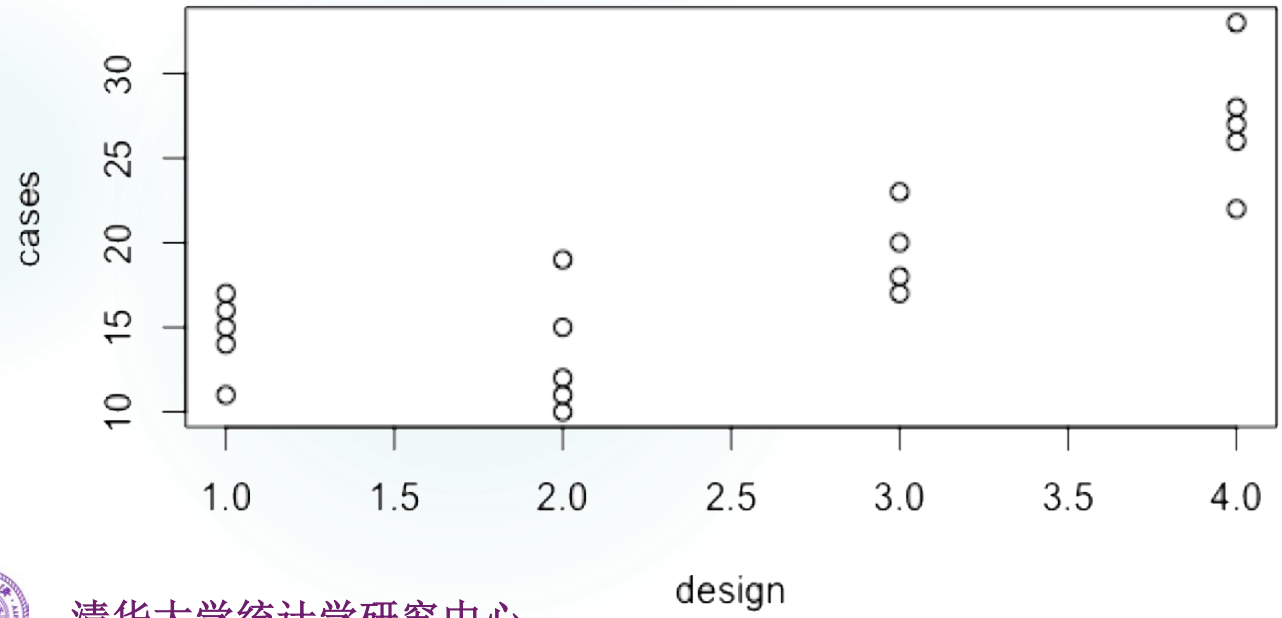
EDA

- Data for one-way ANOVA

```
a1 = read.table("CH16TA01.txt")
colnames(a1) = c("cases", "design", "store")
View(a1)
```

- Plot the data

```
plot(cases ~ design, data=a1)
```



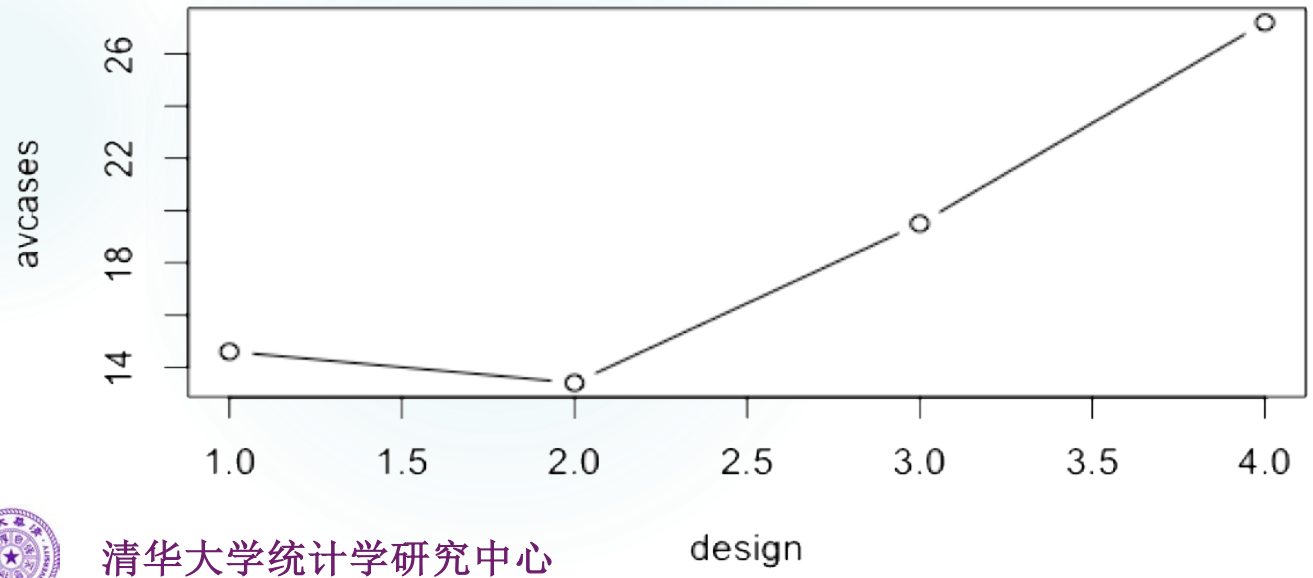
	cases	design	store
1	11	1	1
2	17	1	2
3	16	1	3
4	14	1	4
5	15	1	5
6	12	2	1
...	...	...	...
19	28	4	5

EDA

- Plot the means
- Also called the means plot

Specify the factor variables, need to be a list

```
> a2 = aggregate(a1$cases, list(a1$design), mean)
> colnames(a2) = c("design", "avcases")
> a2$freq = tapply(a1$cases, a1$design, length)
> plot(avcases ~ design, data=a2, type = 'b')
```



	design	avcases	freq
1	1	14.6	5
2	2	13.4	5
3	3	19.5	4
4	4	27.2	5

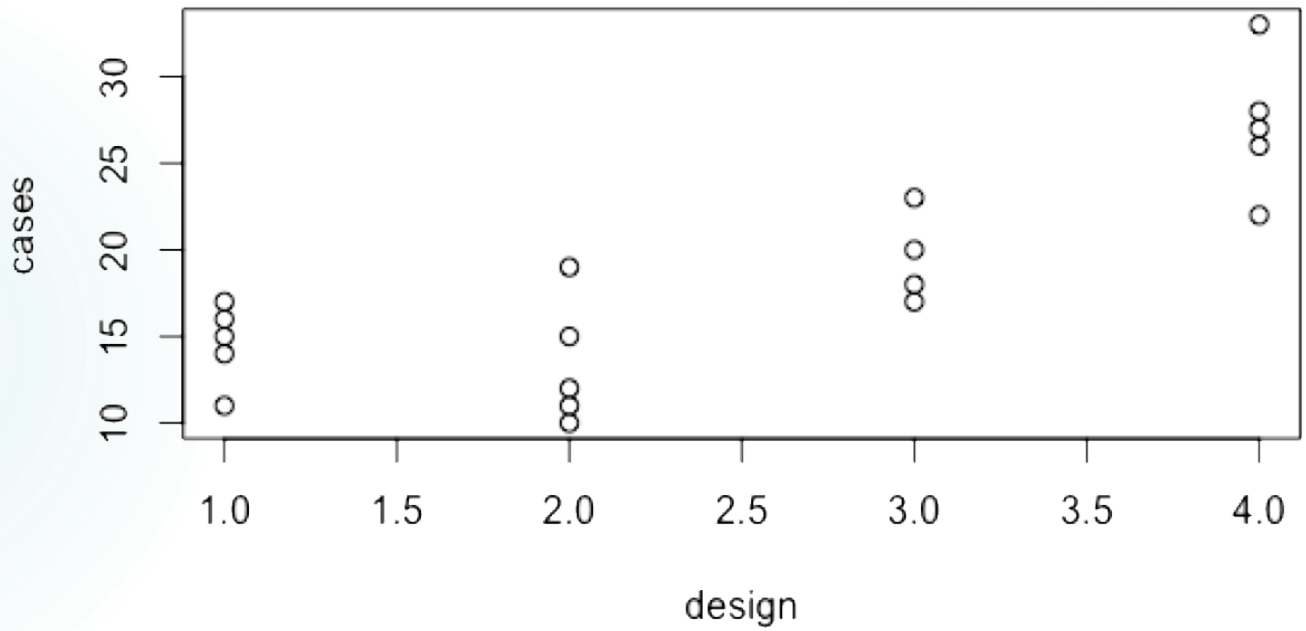
The Model

- We assume that the response variable is Normally distributed with a
  - 1. mean that may depend on the level of the factor
  - 2. constant variance
- All observations assumed independent
- NOTE: Same assumptions as linear regression except there is no assumed linear relationship between X and $\mathrm{E}(Y|X)$

The Scatterplot

Based on scatterplot and consider:

- Independence?
- Constant variance?
- Normally distributed?



Cell Means Model

- A "cell" refers to a level of the factor
- $Y_{ij} = \mu_i + \varepsilon_{ij}$

  where μ_i is the theoretical mean or expected value of all observations at level (or in cell) i

  - the ε_{ij} are iid $N(0, \sigma^2)$ which means $Y_{ij} \sim N(\mu_i, \sigma^2)$ and independent

Parameters

- The parameters of the model are
  - $\mu_1, \mu_2, \ldots, \mu_r$
  - σ^2
- Question (Version 1) – Does our explanatory variable help explain Y?
- Question (Version 2) – Do the μ_i's vary?
- H_0: $\mu_1 = \mu_2 = \ldots = \mu_r = \mu$ (a constant)
- H_a: not all μ_i's are the same

Estimates

- Estimate μ_i by the mean of the observations at level i, $\bar{Y}_{i.}$

$$\hat{\mu}_i = \bar{Y}_{i.} = \sum_j Y_{ij}/n_i \text{ (sample mean)}$$

- For each level i, also get an estimate of the variance s_i^2

$$s_i^2 = \sum_j \left(Y_{ij} - \bar{Y}_{i.}\right)^2 / \left(n_i - 1\right) \text{ (sample variance)}$$

- We combine these to get an overall estimate of σ^2
- Sample approach as pooled t-test

Pooled Estimate of σ^2

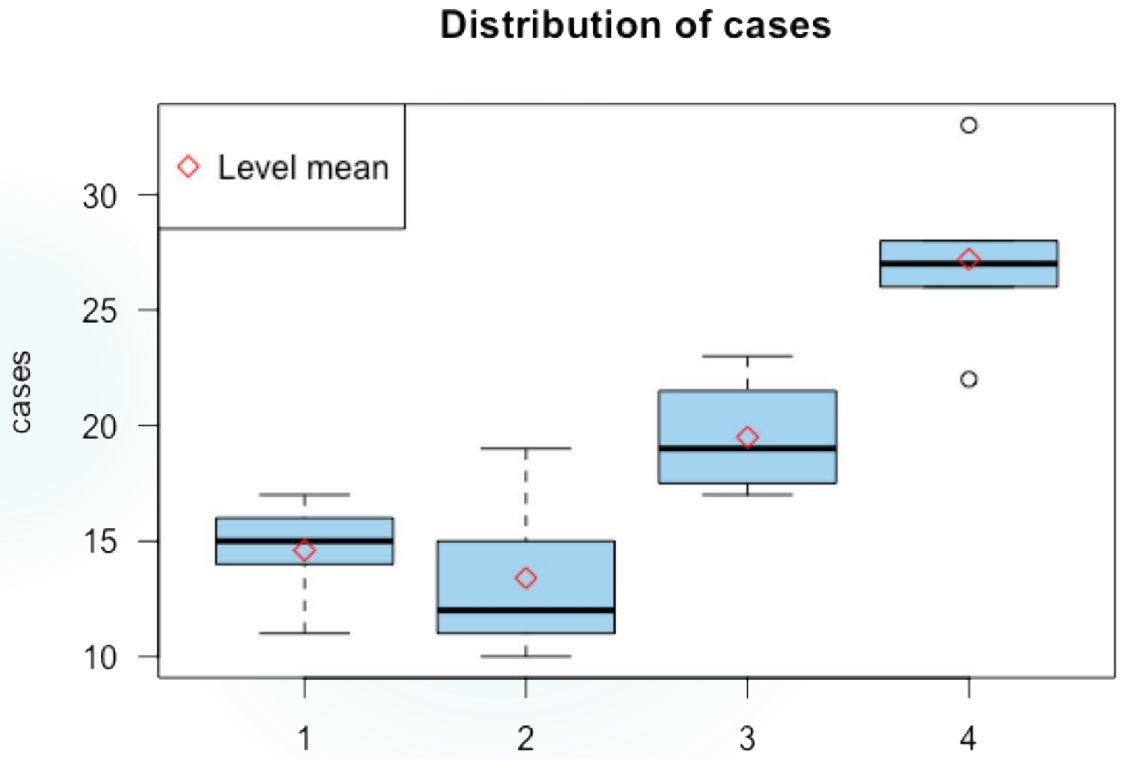
- If the n_i were all the same we would average the s_i^2
- Do **NOT** average the s_i
- In general we pool s_i^2, given weights proportional to the df, n_i - 1
- The pooled estimate is

$$s^2 = \sum\left((n_i - 1)s_i^2\right) \Big/ \left(\sum(n_i - 1)\right)$$

$$= \sum\left((n_i - 1)s_i^2\right)/(n_T - r)$$

Sample Means and Sample Standard Deviations


```
> a2$sd <- tapply(a1$cases, a1$design, sd)
```



Level of design	n_i	cases	
		Mean	StdDev
1	5	14.6000000	2.30217289
2	5	13.4000000	3.64691651
3	4	19.5000000	2.64575131
4	5	27.2000000	3.96232255

Number of Observations Read	19

Regression


```
> fit = lm(cases ~ design, data = a1)
> summary(fit)
> anova(fit)
```

- Important to inform R that 'design' is a factor!

```
> a1$design = factor(a1$design)
```

	Levels	Values
design	4	1 2 3 4

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.7000     2.1557    3.572  0.00235 **
design       4.4191     0.7913    5.584  3.29e-05 ***

Residual standard error: 3.936 on 17 degrees of freedom
Multiple R-squared:  0.6472,  Adjusted R-squared:  0.6264
F-statistic: 31.19 on 1 and 17 DF,  p-value: 3.289e-05
```

```
Analysis of Variance Table

Response: cases
           Df Sum Sq Mean Sq F value    Pr(>F)
design      1  483.08  483.08  31.186 3.289e-05 ***
Residuals 17 263.34   15.49
```

Regression

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
design1	14.600	1.452	10.053	4.66e-08	***
design2	13.400	1.452	9.226	1.43e-07	***
design3	19.500	1.624	12.009	4.28e-09	***
design4	27.200	1.452	18.728	8.16e-12	***

Residual standard error: 3.248 on 15 degrees of freedom
Multiple R-squared: 0.9785, Adjusted R-squared: 0.9727
F-statistic: 170.3 on 4 and 15 DF, p-value: 2.64e-12

Analysis of Variance Table

Response: cases

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
design	4	7183.8	1795.95	170.29	2.64e-12	***
Residuals	15	158.2	10.55			

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	14.600	1.452	10.053	4.66e-08	***
design2	-1.200	2.054	-0.584	0.5677	
design3	4.900	2.179	2.249	0.0399	*
design4	12.600	2.054	6.135	1.91e-05	***

Residual standard error: 3.248 on 15 degrees of freedom
Multiple R-squared: 0.7881, Adjusted R-squared: 0.7457
F-statistic: 18.59 on 3 and 15 DF, p-value: 2.585e-05

Analysis of Variance Table

Response: cases

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
design	3	588.22	196.074	18.591	2.585e-05	***
Residuals	15	158.20	10.547			

```
> fit1 = lm(cases ~ 0 + design, data = a1)
> summary(fit)
> anova(fit)
```

Model Comparison

- Compare the sample means and sample standard deviations
- Provides estimates based on model (i.e., constant variance)

Regression:

	Estimate	Std. Error
design1	14.600	1.452
design2	13.400	1.452
design3	19.500	1.624
design4	27.200	1.452

Cell means:

design	ni	mean	sd
1	5	14.6	2.302173
2	5	13.4	3.646917
3	4	19.5	2.645751
4	5	27.2	3.962323

ANOVA Table

Notation:

- treatment sample mean
$$\bar{Y}_{i\cdot} = \sum_j Y_{ij}/n_i$$
- overall sample mean
$$\bar{Y}_{\cdot\cdot} = \sum_i \sum_j Y_{ij}/n_T$$
- total number of observations
$$n_T = \sum_i n_i$$
- when $n_i = n$ for all i,
$$\bar{Y}_{\cdot\cdot} = \sum_i \bar{Y}_{i\cdot}/r$$

Source	df	SS	MS
Model	$r-1$	$\sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	SSR/df_R
Error	$n_T - r$	$\sum_i \sum_j (\bar{Y}_{ij} - \bar{Y}_{i\cdot})^2$	SSE/df_E
Total	$n_T - 1$	$\sum_i \sum_j (\bar{Y}_{ij} - \bar{Y}_{\cdot\cdot})^2$	SST/df_T

- model/ regression/ treatment
  error/ residual SS/ deviances

Expected Mean Squares

- $E(MSE) = \sigma^2$
- $E(MSR) = \sigma^2 + (\sum_i n_i (\mu_i - \mu.)^2)/(r-1)$

  where $\mu. = (\sum_i n_i \mu_i)/n_T$
- $E(MSR) > E(MSE)$ when the group means are different
- See KNNL p 694 – 698 for more details
- In more complicated models, the EMS tell us how to construct the F test

F Test

- $F^* = MSR/MSE$
- H_0: $\mu_1 = \mu_2 = \ldots = \mu_r = \mu$ (a constant)
- H_a: not all μ_i's are the same
- Under H_0, $F^* \sim F(r-1, n_T - r)$
- Reject H_0 when F^* is large
- Typically report the P-value

Maximum Likelihood Approach


```
> fit3 = glm(cases ~ 0 + design, data=a1,
                    family = gaussian(link = identity))
> summary(fit3)
```

```
Coefficients:
        Estimate Std. Error t value Pr(>|t|)
design1   14.600      1.452  10.053  4.66e-08 ***
design2   13.400      1.452   9.226  1.43e-07 ***
design3   19.500      1.624  12.009  4.28e-09 ***
design4   27.200      1.452  18.728  8.16e-12 ***

(Dispersion parameter for gaussian family taken to be 10.54667)

    Null deviance: 7342.0  on 19  degrees of freedom
Residual deviance:  158.2  on 15  degrees of freedom
AIC: 104.19

Number of Fisher Scoring iterations: 2
```

Factor Effects Model

- A reparameterization of the cell means model
- Useful way at looking at more complicated models
- Null hypotheses are easier to state
- $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$
  - the ε_{ij} are iid $N(0, \sigma^2)$

Parameters

- The parameters of the model are
  - $\mu, \tau_1, \tau_2, \ldots, \tau_r$
  - σ^2
- The cell means model had $r + 1$ parameters
  - r μ's and σ^2
- The factor effects model has $r + 2$ parameters
  - μ, the r τ's, and σ^2
  - Cannot uniquely estimate all parameters

Example

- Suppose $r = 3$; $\mu_1 = 10$, $\mu_2 = 20$, $\mu_3 = 30$
- What is an equivalent set of parameters for the factor effects model?

- We need to have $\mu + \tau_i = \mu_i$ so…
  - 1. $\mu = 0$, $\tau_1 = 10$, $\tau_2 = 20$, $\tau_3 = 30$
  - 2. $\mu = 20$, $\tau_1 = -10$, $\tau_2 = 0$, $\tau_3 = 10$
  - 3. $\mu = 5000$, $\tau_1 = -4990$, $\tau_2 = -4980$, $\tau_3 = -4970$

  all provide the same means

Problem with Factor Effects?

- These parameters are not *estimable* or not well defined (i.e., not unique)
  - There are many solutions to the least squares problem
  - There is an $X'X$ matrix for this parameterization that does not have an inverse (perfect multicollinearity)
- We addressed similar situation in multiple regression. Parameter estimators provided by R are *biased*

Factor Effects Solution

- Put a constraint on the τ_i
- Common to assume $\Sigma_i \tau_i = 0$
- This effectively reduces the number of parameters by one
- Numerous other constraints possible
  - $\Sigma_i \tau_i = 100$
  - $\tau_r = 0$

Consequences

- Regardless of constraint, we always have $\mu_i = \mu + \tau_i$
- The constraint $\Sigma_i \tau_i = 0$ implies
  - $\mu = (\Sigma_i \mu_i)/r$ (unweighted overall mean)
  - $\tau_i = \mu_i - \mu$ (group effect)
- The "unweighted" complicates things when the are not all equal; see KNNL p 702-708

Hypotheses

- H_0: $\mu_1 = \mu_2 = \ldots = \mu_r = \mu$ (a constant)
- H_a: not all μ_i's are the same

are translated into

- H_0: $\tau_1 = \tau_2 = \ldots = \tau_r = 0$
- H_a: at least one τ_i is not 0

Estimates of Parameters

- With the constraint $\Sigma_i \tau_i = 0$
- $\hat{\mu} = \sum_i \bar{Y}_{i.}/r$

  $= \bar{Y}_{..} \quad (if\ n_i = n)$
- $\hat{\tau}_i = \bar{Y}_{i.} - \hat{\mu}$

Cereal Package Example

- Y is the number of cases of cereal sold
- X is the design of the cereal package
- $i = 1$ to 4 levels
- $j = 1$ to n_i stores with design i ($n_i = 5, 5, 4, 5$)

```
> fit = lm(cases ~ design, data = a1)
```

Coding for X

- $r = 4$ explanatory variables
- The i^{th} explanatory variable is equal to 1 if the observation is from the i^{th} group
- In other words, the rows of X are

  1 1 0 0 0 for design=1

  1 0 1 0 0 for design=2

  1 0 0 1 0 for design=3

  1 0 0 0 1 for design=4

Solution Used by R

- Recall, $X'X$ does not have an inverse
- We use the first level as the baseline, so X is

  1 0 0 0 for design=1

  1 1 0 0 for design=2

  1 0 1 0 for design=3

  1 0 0 1 for design=4
- Dropping one column corresponds to the constraint $\tau_1 = 0$
- Recall that μ and the τ_i are not estimable
- But the linear combinations $\mu + \tau_i$ are estimable
- These are estimated by the cell means (i.e., sample means)

Interpretation

- If $\tau_r = 0$ (in our case, $\tau_1 = 0$), then the corresponding estimate should be zero
- The intercept is then estimated by the sample mean of Group/Level 1
- Since $\mu + \tau_i$ is the mean of group i, the τ_i are estimated as the differences between the sample mean of Group i and the sample mean of Group 1
- Recall the sample means

Cell means:

design	ni	mean	sd
1	5	14.6	2.302173
2	5	13.4	3.646917
3	4	19.5	2.645751
4	5	27.2	3.962323

Parameter Estimates based on means:

design	mean	μ=14.6
1	14.6	$\tau_1 = 14.6 - 14.6 = 0$
2	13.4	$\tau_2 = 13.4 - 14.6 = -1.2$
3	19.5	$\tau_3 = 19.5 - 14.6 = 4.9$
4	27.2	$\tau_4 = 27.2 - 14.6 = 12.6$

Relationship with Regression

- Analysis of variance models are a basic type of statistical model
- Like regression models…
  - They are concerned with the statistical relation between one or more predictor variables and a response variable
  - They are appropriate for both observational data and data based on formal experiments
  - The response variable for analysis of variance models is a quantitative variable
- Analysis of variance models differ from ordinary regression models in two key respects:
  - 1. The explanatory or predictor variables in analysis of variance models may be qualitative (gender, geographic location, plant shift, etc.)
  - 2. If the predictor variables are quantitative, no assumption is made in analysis of variance models about the nature of the statistical relation between them and the response variable. Thus, the need to specify the nature of the regression function encountered in ordinary regression analysis does not arise in analysis of variance models

Last Slide

- Read KNNL Chapter 16 up to 16.10
- We used programs lec11_2.R to generate the output for today
- Will focus more on the relationship between regression and one-way ANOVA in next topic