清华大学统计学辅修课程

Linear Regression Analysis

Lecture 10-
Interaction Models &
Model Selection and Diagnostics

周在莹

清华大学统计学研究中心

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

Topic 1:

Interaction Models

Interaction Models

- With several explanatory variables, we need to consider the possibility that the effect of one variable depends on the value of another variable

- Let's consider two special cases
  - Case 1. One binary variable (Y/N) and one continuous variable
  - Case 2. Two continuous variables

Case 1: One Binary Variable and One Continuous Variable

- X_1 takes values 0 and 1 corresponding to the two different groups (Y/N)
- Variable X_2 is continuous
- Full model:

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

- When $X_1 = 0$ (Group 1), $Y = \beta_0 + \beta_2 X_2 + \varepsilon$
- When $X_1 = 1$(Group 2), $Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2 + \varepsilon$
  - β_0, β_2 are the intercept, slope for Group 1
  - $\beta_0 + \beta_1$, $\beta_2 + \beta_3$ are the intercept, slope for Group 2
- H_0: $\beta_1 = \beta_3 = 0$ tests the groups are the same, or the regression lines are the same
- H_0: $\beta_1 = 0$ tests the intercepts are the same for the two groups
- H_0: $\beta_3 = 0$ tests the slopes are the same for the two groups

Innovation Example KNNL p316

- Y is number of months for an insurance company to adopt an innovation
- X_1 is the size of the firm (a continuous variable)
- X_2 is the type of firm (a qualitative or categorical variable)
- X_2 takes the value 0 if it is a mutual fund firm and 1 if it is a stock fund firm
- The Question:
  - We ask whether or not stock firms adopt the innovation slower or faster than mutual firms
- We ask the question across all firms, regardless of size

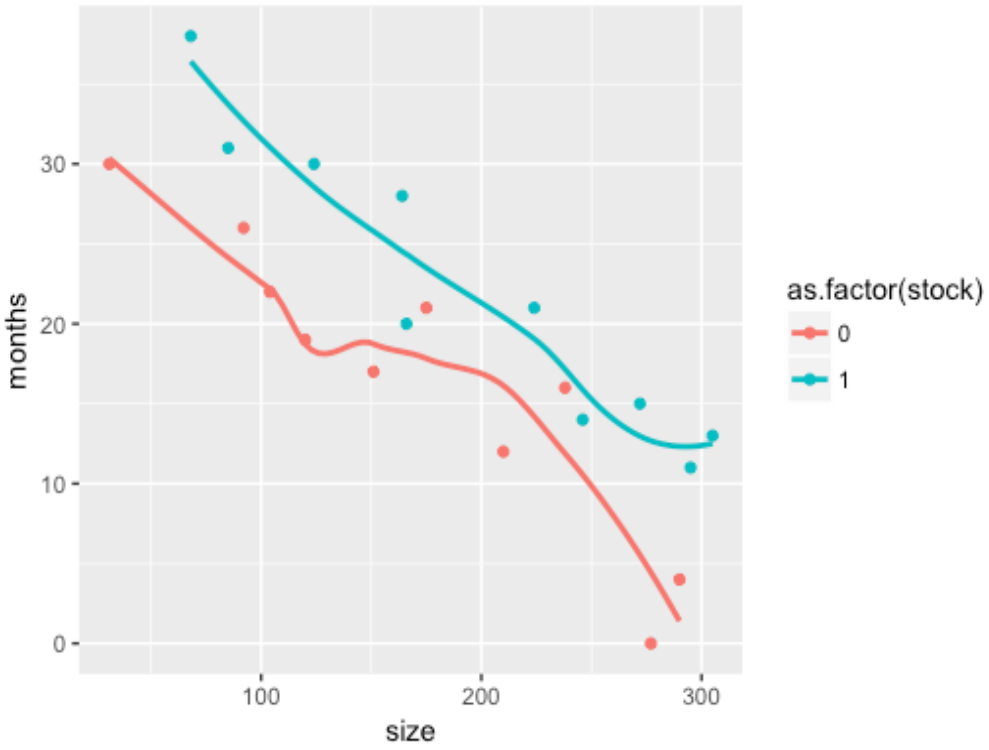
Firm i	(1) Number of Months Elapsed Y_i	(2) Size of Firm (million dollars) X_{i1}	(3) Type of Firm	(4) Indicator Code X_{i2}	(5) $X_{i1}X_{i2}$
1	17	151	Mutual	0	0
2	26	92	Mutual	0	0
3	21	175	Mutual	0	0
4	30	31	Mutual	0	0
5	22	104	Mutual	0	0
6	0	277	Mutual	0	0
7	12	210	Mutual	0	0
8	19	120	Mutual	0	0
9	4	290	Mutual	0	0
10	16	238	Mutual	0	0
11	28	164	Stock	1	164
12	15	272	Stock	1	272
13	11	295	Stock	1	295
14	38	68	Stock	1	68
15	31	85	Stock	1	85
16	21	224	Stock	1	224
17	20	166	Stock	1	166
18	13	305	Stock	1	305
19	30	124	Stock	1	124
20	14	246	Stock	1	246

Plot the Data

- Smooth lines are automatically fit to each group (defined by categorical aesthetics or the group aesthetic)

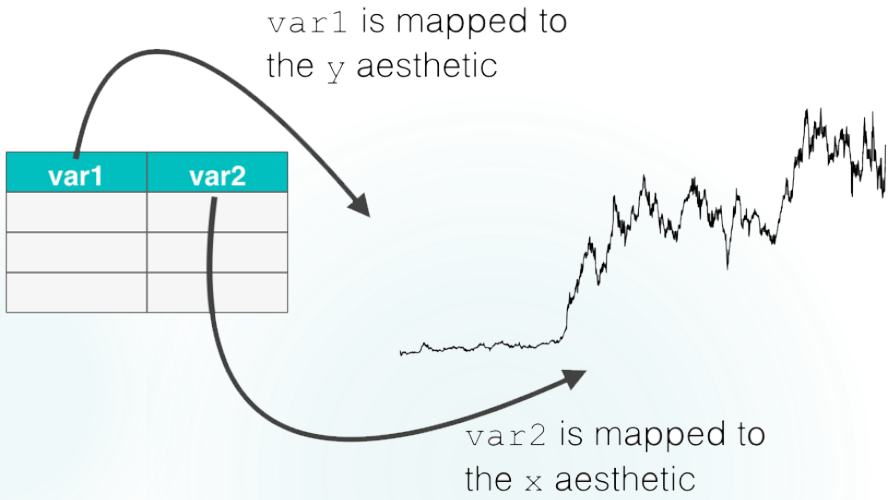
```
> library(ggplot2)
> ggplot(a1,
      aes(size, months, color = as.factor(stock))) +
  geom_point( ) +
  geom_smooth(se = FALSE, method='loess')
```

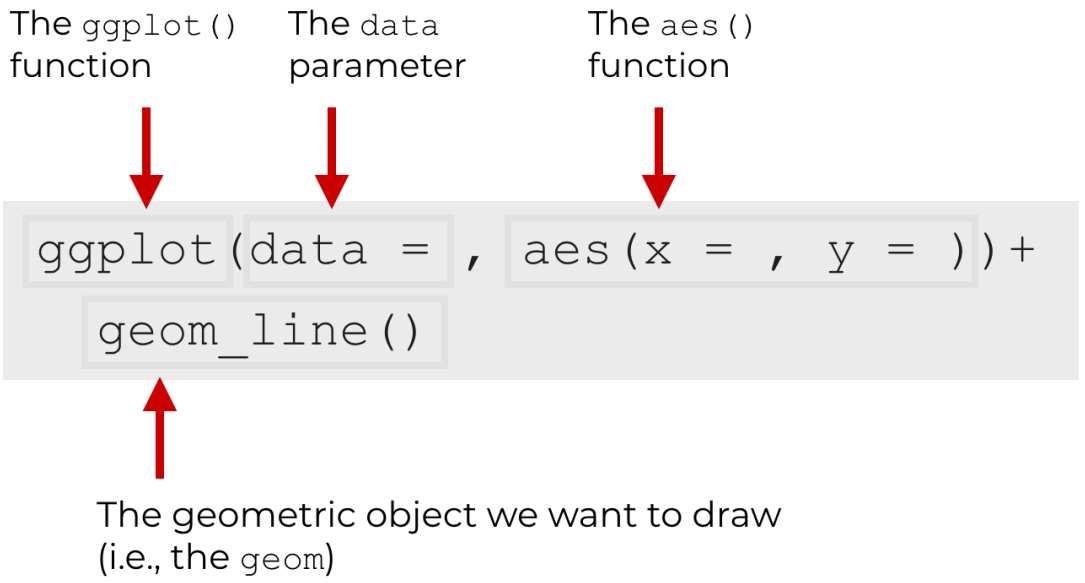
Same intercept?
Same slope?
Same line?



Note on the ggplot2 Syntax

- The aes() function specifies the *aesthetic mappings* from the data to the chart





- Geometric objects are things that we can draw: bars, points, lines, etc
- The type of geom you select dictates the type of chart you make

Interaction Effects

$stock = 0$ (Group 1), $Y = \beta_0 + \beta_1 size + \varepsilon$
$stock = 1$(Group 2), $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, size + \varepsilon$

- Interaction expresses the idea that the effect of one explanatory variable on the response depends on another explanatory variable
- Here this would mean that the slope of the line depends on the type of firm
- Are both lines the same?
- From scatterplot, looks like different intercepts but can use the test statement for formal assessment

```
> a1$sizestock = a1$size * a1$stock
>  reg1 <- lm(months ~ size + stock + sizestock, data=a1)
>  summary(reg1)
>  anova(reg1)
>  reg2 <- lm(months ~ size, data=a1)
>  anova(reg2, reg1)  # test for stock + sizestock
```

Analysis of Variance Table

Model 1: months ~ size
Model 2: months ~ size + stock + sizestock

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	18	492.63				
2	16	176.38	2	316.25	14.344	0.00027 ***

Reject H_0. There is a difference in the linear relationship across groups

How Are They Different?

1. No difference in slopes assuming different intercepts
2. Potentially different intercepts assuming different slopes

$$Y = \beta_0 + \beta_1 size + \varepsilon$$
$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, size + \varepsilon$$

- Note that size = 0 is outside range of the data used to fit model
- Can center size so comparison of "intercepts" made in middle of data set range → more precision

```
>a1c = a1
>a1c$size = scale(a1$size, center = TRUE, scale = FALSE)
>a1c$sizestock = a1c$size * a1c$stock
>reg3 <- lm(months ~ size + stock + sizestock, data=a1c)
>summary(reg3)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.8383695	2.4406498	13.864	2.47e-10 ***
size	-0.1015306	0.0130525	-7.779	7.97e-07 ***
stock	8.1312501	3.6540517	2.225	0.0408 *
sizestock	-0.0004171	0.0183312	-0.023	0.9821

1. No difference in slopes assuming different intercepts – same result
2. Significant different intercepts assuming different slopes

   BIG change in P-value

Coefficients: (Centered)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15.3750253	1.0636706	14.455	1.33e-10 ***
size	-0.1015306	0.0130525	-7.779	7.97e-07 ***
stock	8.0553930	1.5039911	5.356	6.43e-05 ***
sizestock	-0.0004171	0.0183312	-0.023	0.982

Two Parallel Lines?

```
> reg5 <- lm(months ~ size + stock, data=a1)
>  summary(reg5)
>  anova(reg5)
```

Analysis of Variance Table

Response: months

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
size	1	1188.17	1188.17	114.51	5.683e-09 ***
stock	1	316.25	316.25	30.48	3.742e-05 ***
Residuals	17	176.39	10.38		

- Can show general linear test same as *t*-test for H_0: $\beta_3=0$

Intercept for stock firms is

33.87+8.05 = 41.92

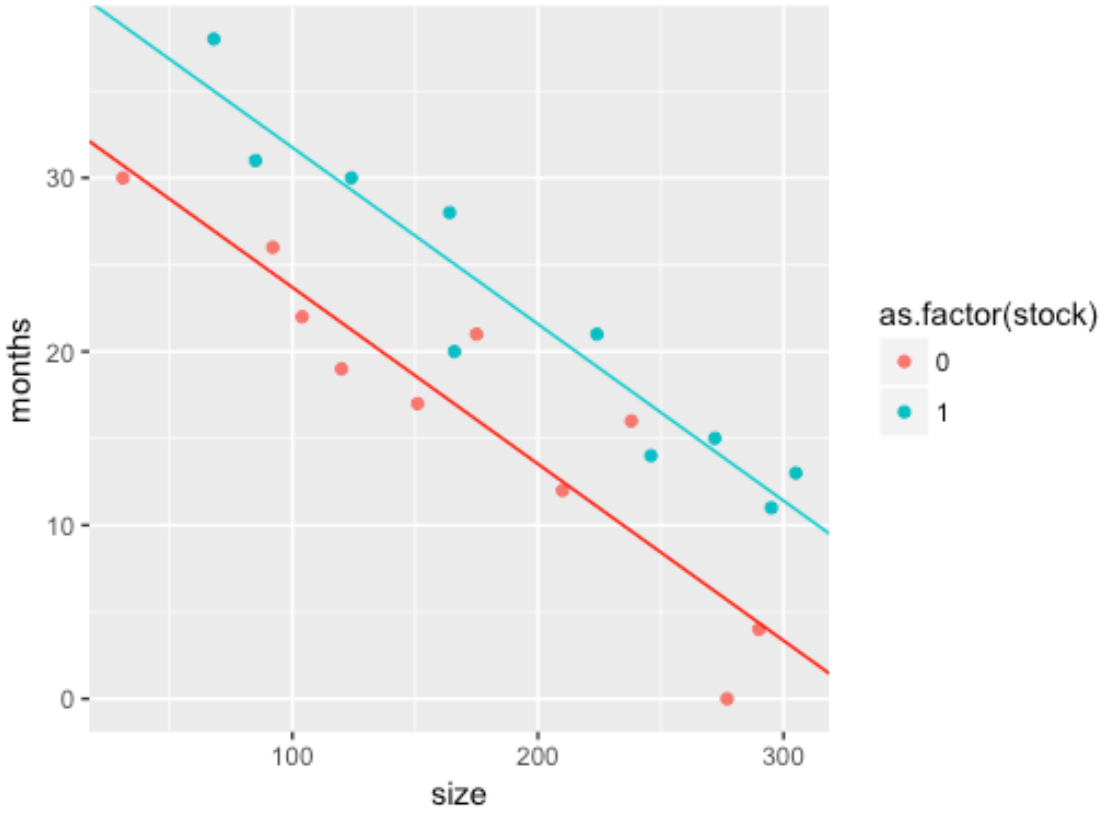
Common slope is –0.10

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.874069	1.813858	18.675	9.15e-13 ***
size	-0.101742	0.008891	-11.443	2.07e-09 ***
stock	8.055469	1.459106	5.521	3.74e-05 ***

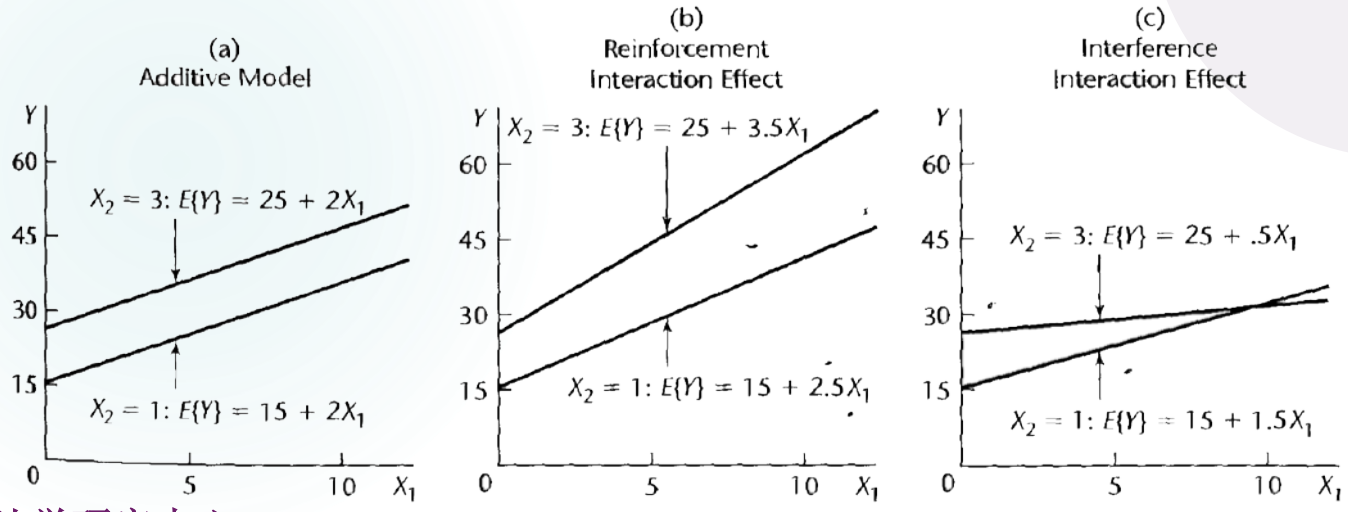
Plot the Two Fitted Lines


```
beta <- reg5$coefficients

ggplot(a1, aes(size, months, color =
                    as.factor(stock))) +
    geom_point() +
    geom_abline(intercept = beta[1],
                slope = beta[2], color='red') +
    geom_abline(intercept = beta[1]+beta[3],
                slope = beta[2], color='darkturquoise')
```



Case 2. Two Continuous Variables

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$
- Can be rewritten as follows
  - $Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon$
  - $Y = \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2 + \varepsilon$
- The coefficient of one explanatory variable depends on the value of the other explanatory variable



Last Slide

- We went over KNNL 8.2 – 8.7
- We used lec10_1.R to generate the output

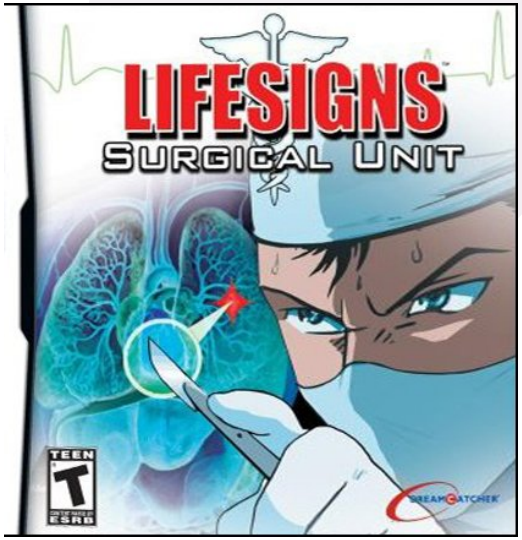
Topic 2:
Model Selection
&
Diagnostics

Variable/Model Selection

- We want to choose a "best" model that is a **subset** of the available explanatory variables
- Two aspects in variable selection
  - 1. How many explanatory variables should we use (i.e., subset size)
  - 2. Given the subset size, which variables should we choose

Surgical Unit Example

- KNNL Page 350, Section 9.2
- n = 54 patients / cases (of 108 patients)
- Y : survival time (liver operation)
- X's (explanatory variables) are
  - Blood clotting score(凝血评分)
  - Prognostic index(预后指数)
  - Enzyme function test(酶功能试验)
  - Liver function test (肝功能试验)
  - (There are 4 other explanatory variables)



EDA

- We start with the usual plots and descriptive statistics
- Note that time-to-event / survival data are often heavily skewed and typically transformed with a log prior to model fitting
- Log Transform of Y
- Recall that regression model requires $Y|X$ to be Normally distributed, not Y
- Better to look at residuals
- With data like these, transform reduces influence of long right tail and stabilizes the variance of the residuals

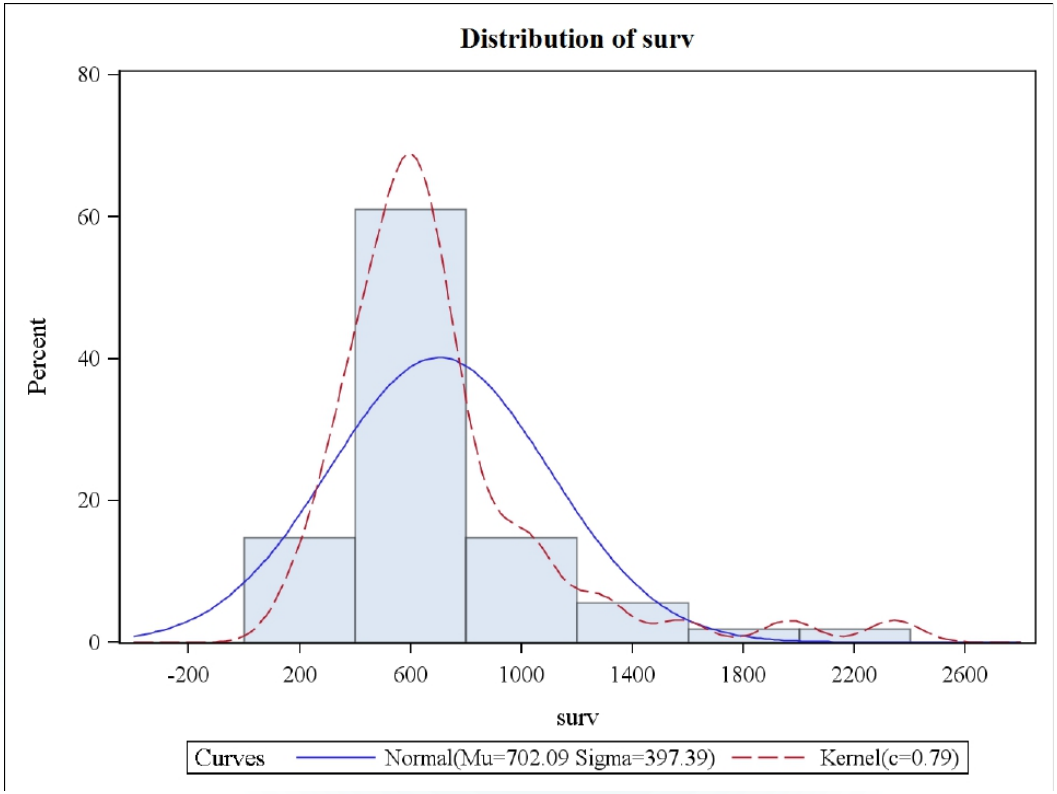
Read Data

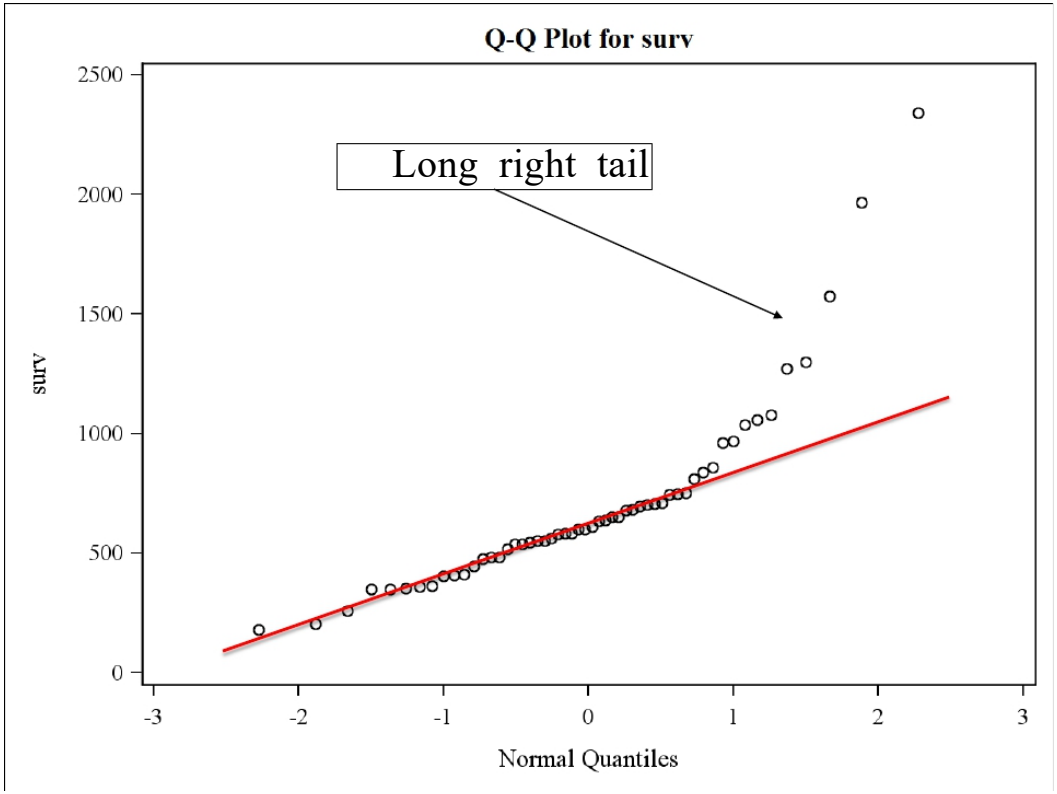
- a1 = read.table("CH09TA01.txt")
- colnames(a1) = c("blood", "prog", "enz", "liver", "age", "gender", "alcmod", "alcheavy", "surv", "lsurv") `log(surv)`
- View(a1) **`Dummy variables for alcohol use`**

Alcohol Use	alcmod	alcheavy
None	0	0
Moderate	1	0
Severe	0	1

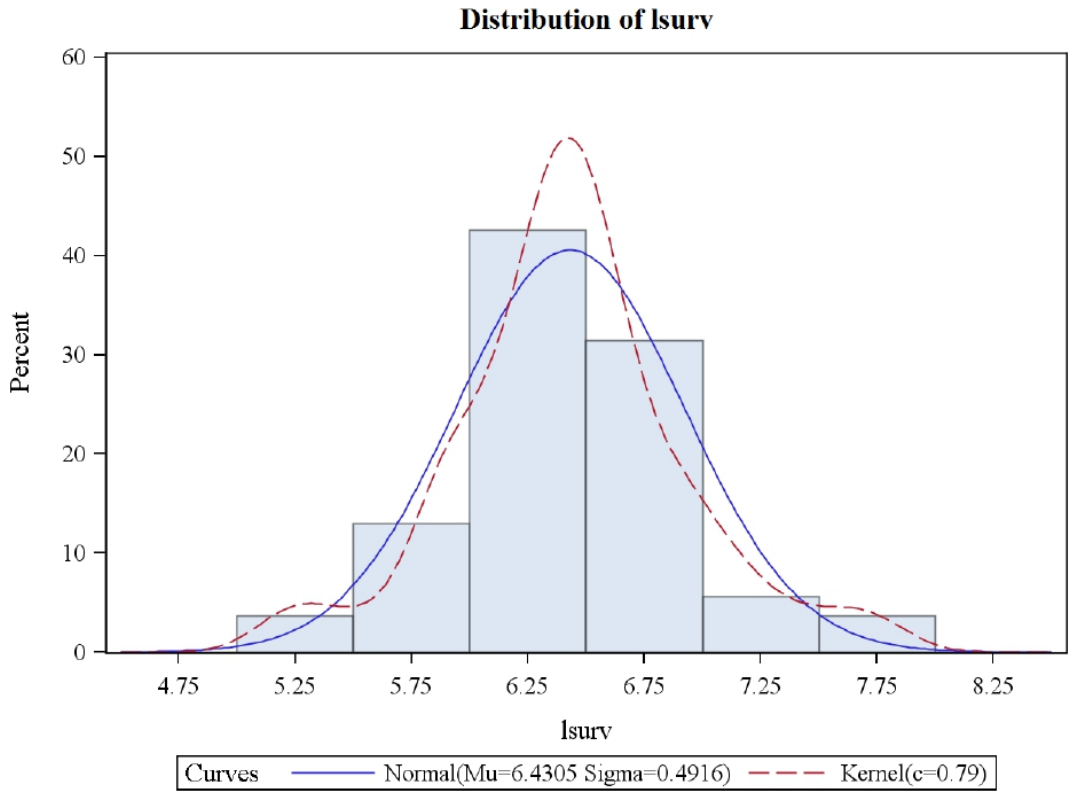
Obs	blood	prog	enz	liver	age	Gender	alcmod	alcheavy	surv	lsurv
1	6.7	62	81	2.59	50	0	1	0	695	6.544
2	5.1	59	66	1.70	39	0	0	0	403	5.999
3	7.4	57	83	2.16	55	0	0	0	710	6.565
4	6.5	73	41	2.01	48	0	0	0	349	5.854
5	7.8	65	115	4.30	45	0	0	1	2343	7.759
6	5.8	38	72	1.42	65	1	1	0	348	5.852

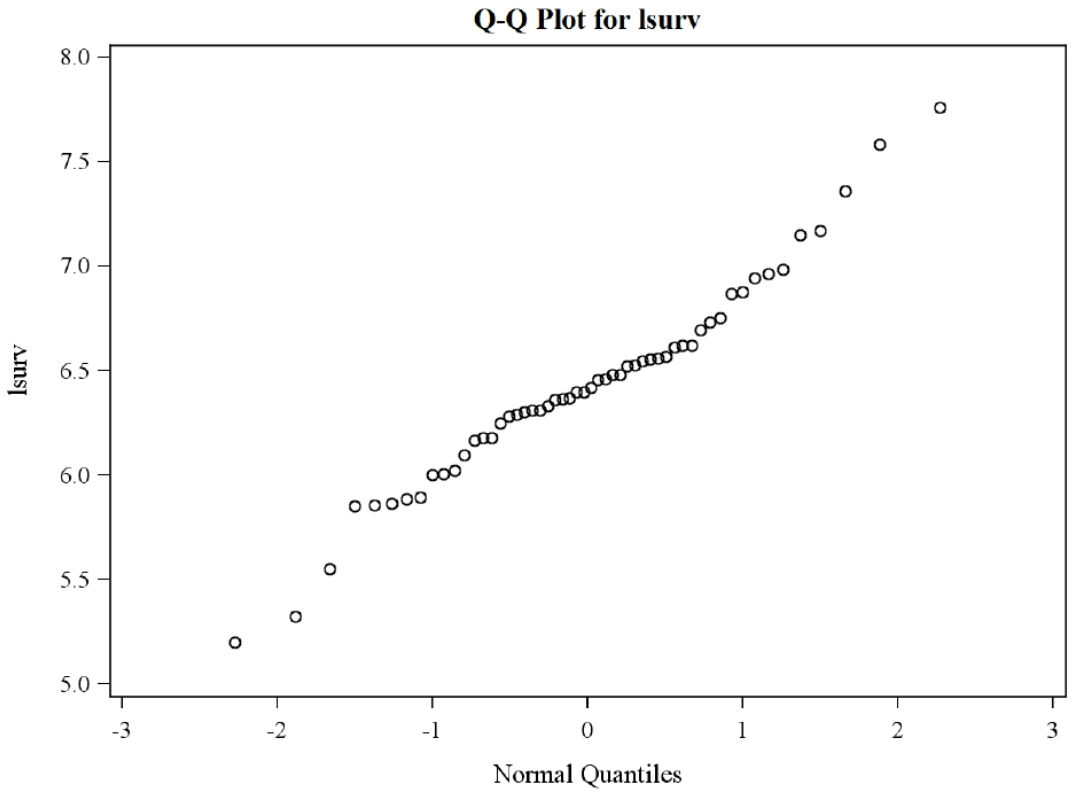
surv


Distribution of surv
Percent
80
60
40
20
0
-200
200
600
1000
1400
1800
2200
2600
surv
Curves
Normal(Mu=702.09 Sigma=397.39)
Kernel(c=0.79)


Q-Q Plot for surv
Long right tail
surv
2500
2000
1500
1000
500
0
-3
-2
-1
0
1
2
3
Normal Quantiles

log(surv)


Distribution of lsurv
Percent
60
50
40
30
20
10
0
4.75
5.25
5.75
6.25
6.75
7.25
7.75
8.25
lsurv
Curves
Normal(Mu=6.4305 Sigma=0.4916)
Kernel(c=0.79)


Q-Q Plot for lsurv
lsurv
8.0
7.5
7.0
6.5
6.0
5.5
5.0
-3
-2
-1
0
1
2
3
Normal Quantiles

Normality Tests


```
> shapiro.test(a1$surv)
```

```
        Shapiro-Wilk normality test

data:  a1$surv
W = 0.80239, p-value = 4.643e-07
```

```
> shapiro.test(a1$lsurv)
```

```
        Shapiro-Wilk normality test

data:  a1$lsurv
W = 0.97508, p-value = 0.3191
```

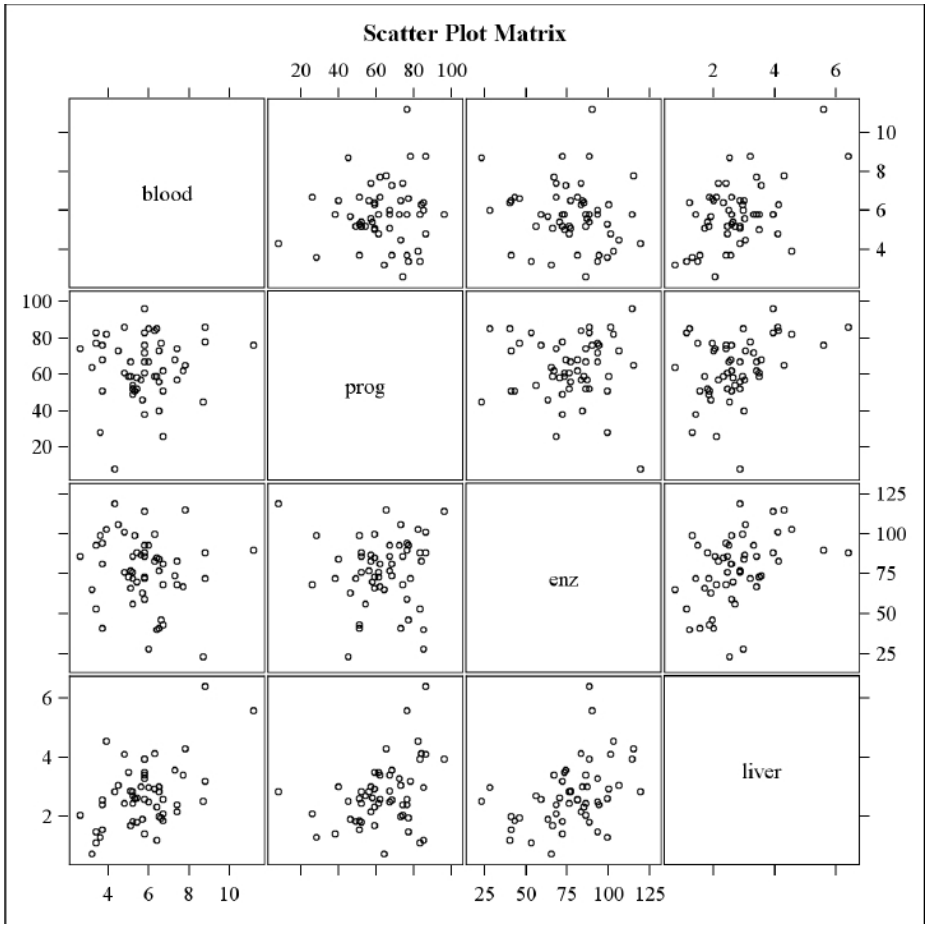
Generate Scatterplots

> pairs(a1[,c("blood", "prog", "enz", "liver")], main= 'Scatter Plot Matrix')

- Correlation Summary

	blood	prog	enz	liver
blood	1.00000000	0.09011973	-0.14963411	0.5024157
prog	0.09011973	1.00000000	-0.02360544	0.3690256
enz	-0.14963411	-0.02360544	1.00000000	0.4164245
liver	0.50241567	0.36902563	0.41642451	1.0000000

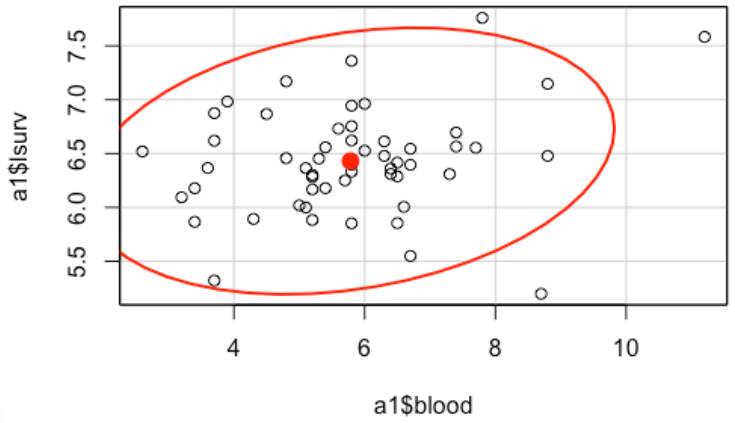
	blood	prog	enz	liver
lsurv	0.2461879	0.4699432	0.6538855	0.6492627

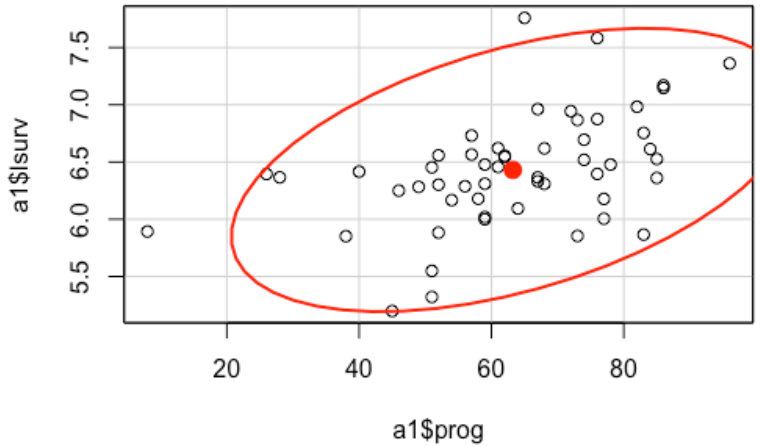


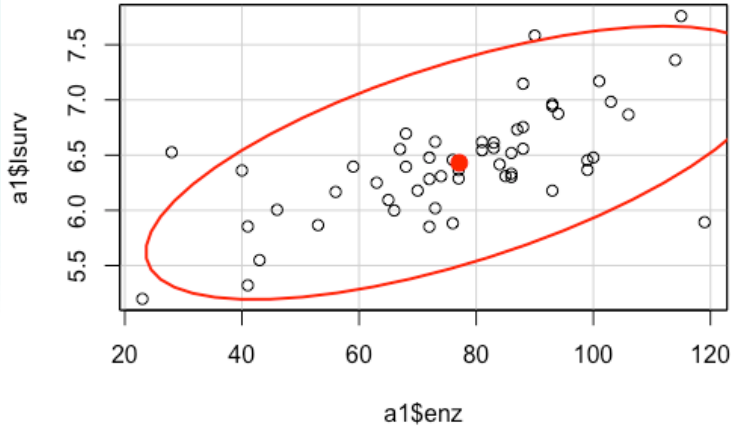
Prediction/Data Ellipse

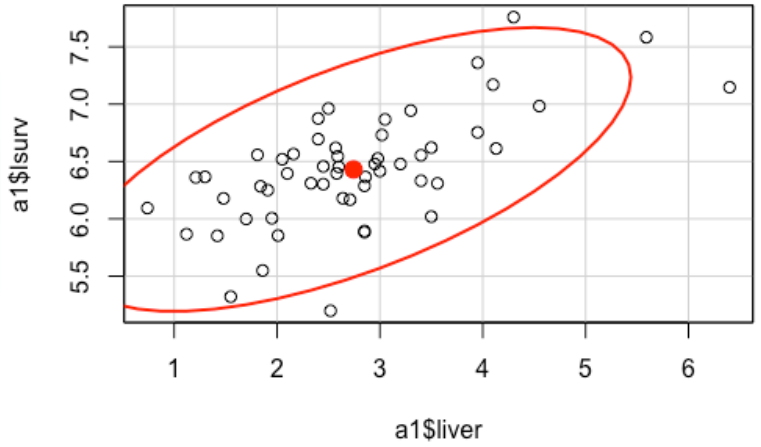
- Assume (blood, lsurv) are bivariate normal with their means, variances and correlation (0.25), and describe the ellipse in 2-dimensions centered on the means, which assigns probability 0.95 to the area under it
- We also call it a “prediction ellipse” for if we were to predict a new observation for that bivariate normal, it would land in the ellipse with probability 0.95

```
> library(car)
> dataEllipse(a1$blood, a1$lsurv, levels = 0.95)
> dataEllipse(a1$prog, a1$lsurv, levels = 0.95)
> dataEllipse(a1$enz, a1$lsurv, levels = 0.95)
> dataEllipse(a1$liver, a1$lsurv, levels = 0.95)
```


a1$lsurv
5.5
6.0
6.5
7.0
7.5
4
6
8
10
a1$blood


a1$lsurv
5.5
6.0
6.5
7.0
7.5
20
40
60
80
a1$prog


a1$lsurv
5.5
6.0
6.5
7.0
7.5
20
40
60
80
100
120
a1$enz


a1$lsurv
5.5
6.0
6.5
7.0
7.5
1
2
3
4
5
6
a1$liver

Different Approaches to VS

- Exhaustive search/comparison based on specific criteria/rules
  - R^2 (or SSE), Adjusted R^2 (or MSE), Mallow's C_p, AIC, SBC (or BIC), PRESS, Cross-Validation (CV), etc
  - can be used when total number of explanatory variables is small or no more than 30 or 40
- Automatic/greedy search in step/stage-wise fashion
  - Forward-selection, backward-selection, stepwise regression, etc
  - Can handle relatively large number of variables
- Penalized optimization approach
  - The Lasso: Least Absolute shrinkage and Selection Operator, by Tibshirani

Basic Setup

- There are P-1 potential explanatory variables, and n cases, where $n > P$
- The full model consists of the intercept and all P-1 variables; The number of parameters is P
- The total number of possible (sub)models that includes the intercept is 2^{p-1}(=1024 for p-1 = 10)
- For a sub-model of p-1variables, the number of parameters is p with $1 \leq p \leq P$

R_p^2 and Adjusted $R_{a,p}^2$

- Recall

$$R_p^2 = 1 - \frac{\text{SSE}(p)}{\text{SST}}$$

$$R_{a,p}^2 = 1 - \frac{n-1}{n-p}\frac{\text{SSE}(p)}{\text{SST}} = 1 - \frac{\text{MSE}(p)}{\text{MST}}$$

- Adding variables to a model will always increase R_p^2, or equivalently, decrease SSE(p). The rate of increase or decrease will slow down, indicating unimportant explanatory variables. The full model maximizes R_p^2 (or minimizes SSE(p))
- Adding variables may increase or decrease $R_{a,p}^2$, or equivalently, decrease/increase MSE(p). Therefore, we can choose a model that maximizes $R_{a,p}^2$ or minimizes MSE(p))

Mallow's C_p

- Consider a model of p-1 variables:

$$\hat{Y}^p = X_p\left(X_p'X_p\right)^{-1}X_p'Y = H_pY$$

  - $E(\hat{Y}^p) = H_pE(Y) = H_p\mu,\ Var(\hat{Y}^p) = \sigma^2H_p$

  where $\mu' = (\mu_1, \mu_2, \ldots, \mu_n)$ be the **true** mean responses at the X_i's

- Expected (mean) squared error:

$$E(\hat{Y}^p - \mu)^2 = (E(\hat{Y}^p) - \mu)^2 + Var(\hat{Y}^p) = Bias^2 + Variance$$

  - The Bias-Variance trade-off

- Total mean squared error:

$$\sum_{i=1}^{n} E(\hat{Y}_i^p - \mu_i)^2 = \sum_{i=1}^{n} (E(\hat{Y}_i^p) - \mu_i)^2 + \sum_{i=1}^{n} Var(\hat{Y}_i^p)$$

- The variance part:

$$\sum_{i=1}^{n} Var(\hat{Y}_i^p) = tr\left(Var\left(\hat{Y}^p\right)\right) = tr\left(\sigma^2 H_p\right) = p\sigma^2$$

- The bias part:

$$\begin{aligned}
&\left(E\left(\hat{Y}^p\right) - \mu\right)'\left(E\left(\hat{Y}^p\right) - \mu\right) \\
&= \left(H_p\mu - \mu\right)'\left(H_p\mu - \mu\right) = \mu'(I - H_p)\mu \\
&= E\left(Y'\left(I - H_p\right)Y\right) - \sigma^2 tr\left(I - H_p\right) \\
&= E\left(SSE(p)\right) - (n-p)\sigma^2
\end{aligned}$$

Notice that
$MSE(X_1,\ldots,X_{P\text{-}1}) = SSE(P)/(n\text{-}P)$
is an unbiased estimator of σ^2

- Total mean squared error:

$$\sum_{i=1}^{n} E(\hat{Y}_i^p - \mu_i)^2 = E\left(SSE(p)\right) - (n-2p)\sigma^2$$

- The model's performance measure:

$$\Gamma_p = \frac{\sum_{i=1}^{n} E(\hat{Y}_i^p - \mu_i)^2}{\sigma^2} = \frac{E\left(SSE(p)\right)}{\sigma^2} - (n-2p)$$

Mallow's C_p

- The model's performance measure:

$$\Gamma_p = \frac{\sum_{i=1}^{n} E(\hat{Y}_i^p - \mu_i)^2}{\sigma^2} = \frac{E(SSE(p))}{\sigma^2} - (n - 2p)$$

- Mallow's C_p is an estimator of Γ_p:

$$C_p = \frac{E(SSE(p))}{\hat{\sigma}^2} - (n - 2p)$$

where $\hat{\sigma}^2$=*MSE* of the full model, and *SSE*(p) is the error (residual) sum of squares of the model under consideration

- When a model is unbiased, $C_p \approx p$; Among all unbiased models, prefer (choose) model with small C_p. It can happen $C_p < p$, due to sampling variation; usually taken as evidence that the model is unbiased
- Use the plot of C_p versus p

Akaike Information Criterion (AIC)

- The general formula:

$$AIC(p) = -2\log(\hat{L}) + 2p$$

where $\hat{L}$ is the maximum likelihood under the model

- Under a linear regression model involving $p-1$ variables:

$$AIC(p) = n\log\left(\frac{SSE(p)}{n}\right) + 2p$$

- The AIC criterion:
  - Select the model that minimizes AIC(p)

Bayesian Information Criterion (BIC)

- Also known as Schwarz's Bayesian Criterion (SBC)
- The general formula:

$$BIC(p) = -2\log(\hat{L}) + [\log(n)]p$$

where $\hat{L}$ is the maximum likelihood under the model

- Under a linear regression model involving $p-1$ variables:

$$BIC(p) = n\log\left(\frac{SSE(p)}{n}\right) + [\log(n)]p$$

- The BIC criterion:
  - Select the model that minimizes BIC(p)

The PRESS Criterion

- PRESS stands for Prediction Sum of Squares

$$PRESS(p) = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(-i)})^2$$

$$\hat{Y}_{i(-i)} = (1, X_{i1}, X_{i2}, \dots, X_{i(p-1)})\hat{\beta}_{(-i)}$$

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

where $X_{(-i)}$ is the design matrix without the i^{th} case

- The PRESS criterion:
  - Select the model that minimizes PRESS(p)
- In fact:

$$Y_i - \hat{Y}_{i(-i)} = \frac{e_i}{1 - h_{ii}}$$

Lemma. If A and $A + B$ are invertible, and B has rank 1, then let $g = \text{trace}(BA^{-1})$. Then $g \neq -1$ and

$$(A + B)^{-1} = A^{-1} - \frac{1}{1+g}A^{-1}BA^{-1}.$$

$$(X'_{(-i)}X_{(-i)})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}X'_iX_i(X'X)^{-1}}{1 - h_{ii}}$$

Automatic/Greedy Search Procedures

- Forward selection (step up)
  - Start with the null model, select and add one variable to the model each step, using *t*- or *F*- test or their *P*-values (either the test statistic larger then a pre-specified value or the *P*-value is less than a pre-specified value. Stop when no additional variables can be added
- Backward elimination
  - Start with the Full model, select and eliminate one variable from the model each step, using *t*- or *F*-test or *P*-values. Stop when we cannot eliminate more variables
- Stepwise regression (forward selection with a backward glance)
  - Alternate the forward selection step and the backward eliminate step. Stop when we cannot add or eliminate variables

The Rule of Thumb

- When P is small, use exhaustive search/comparison based on various criteria
- When P is moderate, use "best" algorithms based on various criteria (only produces the top models)
- When P is large, use automatic/greedy search procedures or use penalized minimization procedures (Lasso)

Variable Selection in R

- Use C_p, R^2 or adjusted R^2

```
> library(leaps)
> leaps(x=, y=, wt=rep(1, NROW(x)), int=TRUE,
        method=c("Cp", "adjr2", "r2"), nbest=10)
```

nbest: limit the output to the best *n* models of each subset size

- Ordering models of same subset size
  - This approach can lead us to consider several models that give approximately the same predicted values
  - May need to apply knowledge of the subject matter to make a final selection
  - Not that important if prediction is the key goal

Ordering Models of Same Subset Size

- R Code

```
> fullres <-  lm(lsurv ~ blood + prog + enz + liver, data=a1)$residuals
> sigsqhat <- sum(fullres^2)/(n-5)
> rsquared <- summary(selectedMod)$r.squared
> aic <- extractAIC(selectedMod)[2]  #n*log(sum(fit$res^2)/n)+2*p
> bic <- extractAIC(selectedMod, k = log(n))[2]
> cp <- sum(selectedMod$residuals^2)/sigsqhat + 2 * selectedMod$rank - n
```

$$C_p = \frac{E(SSE(p))}{\hat{\sigma}^2} - (n - 2p)$$

Selection Results

#X's in Model	R-Square	C(p)	AIC	SBC
1	0.4276	66.4889	-103.8269	-99.84889
1	0.4215	67.7148	-103.2615	-99.28357
1	0.2208	108.5558	-87.1781	-83.20011
2	0.6633	20.5197	-130.4833	-124.51634
2	0.5995	33.5041	-121.1126	-115.14561
2	0.5486	43.8517	-114.6583	-108.69138
3	0.7573	3.3905	-146.1609	-138.20494
3	0.7178	11.4237	-138.0232	-130.06723
3	0.6121	32.9320	-120.8442	-112.88823
4	0.7592	5.0000	-144.5895	-134.64461

#X's in Model	Parameter Estimates				
	Intercept	blood	prog	enz	liver
1	5.26426	.	.	0.01512	.
1	5.61218	.	.	.	0.29819
1	5.56613	.	0.01367	.	.
2	4.35058	.	0.01412	0.01539	.
2	5.02818	.	.	0.01073	0.20945
2	4.54623	0.1079	.	0.01634	.
3	3.76618	0.0954	0.01334	0.01645	.
3	4.40582	.	0.01101	0.01261	0.12977
3	4.78168	0.0448	.	0.01220	0.16360
4	3.85195	0.0836	0.01266	0.01563	0.03216

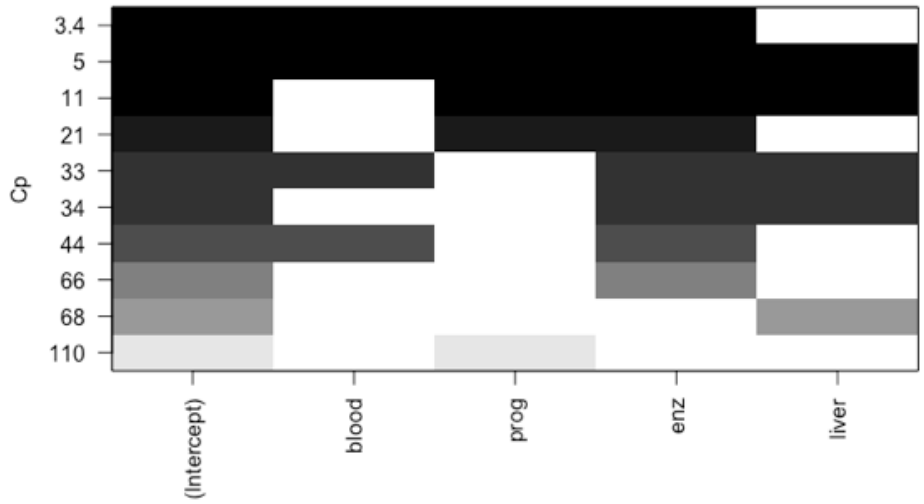
Plot of Subset Regression

- Regsubsets plot based on C_p

```
> regsubsetsObj <- regsubsets(x=predictors, y=response,
                              nbest = nmodels, really.big = T)
> plot(regsubsetsObj, scale = "Cp")
```

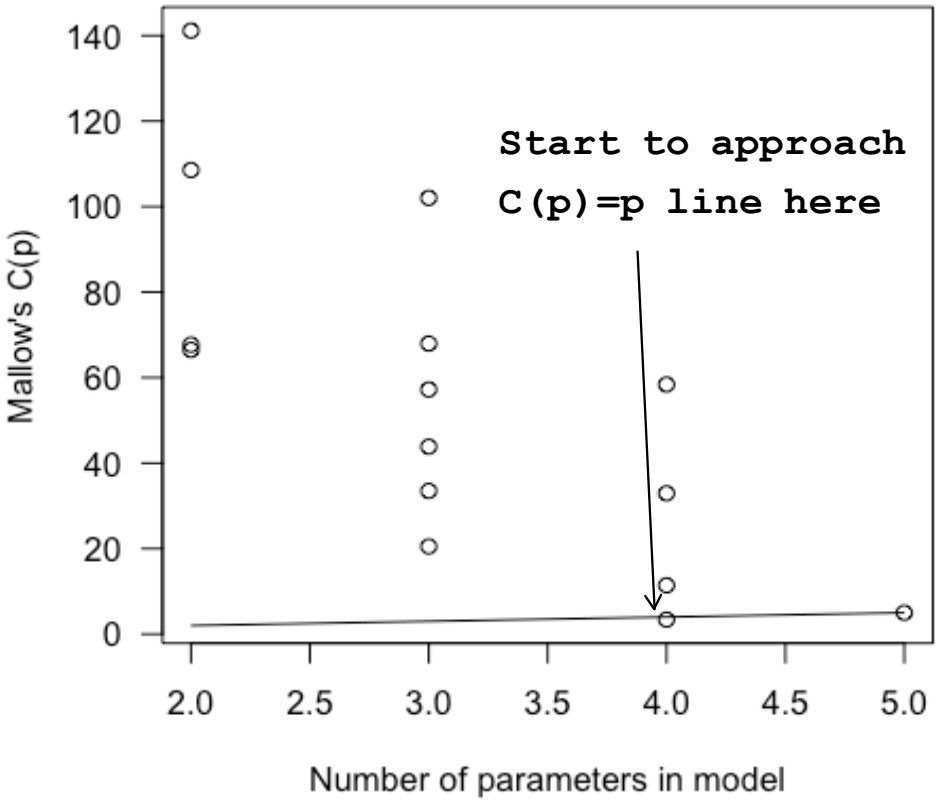
- WARNING: scale = "*Cp*" just lists models in order based on lowest *C*(*p*), regardless of whether it is good or not

#X's in Model	C(p)	R-Square	AIC	SBC
3	3.3905	0.7573	-146.1609	-138.20494
4	5.0000	0.7592	-144.5895	-134.64461
3	11.4237	0.7178	-138.0232	-130.06723



How to Choose with C_p

- 1. Want small $C(p)$
- 2. Want $C(p)$ near p
- In original paper, it was suggested to plot $C(p)$ versus p and consider the smallest model that satisfies these criteria
- Can be somewhat subjective when determining "near"



Model Validation

- Data used to fit/train a model and generate parameter estimates are training data
- In general, a separate data (test or new data) should be used for validate a fitted model (e.g. predictive ability, etc.)
- When new data are not available, various types cross-validation techniques (split data in to training/test, leave-one out, 10-fold) can be used

Additional Multiple Regression Diagnostics

Friedrich Wilhelm Herschel
Sir William Herschel
German-born British
astronomer, composer

> 'Most of the phenomena which nature presents are very complicated; and when the effects of all known causes are estimated with exactness and subducted, the residual facts are constantly appearing in the form of phenomena altogether new, and leading to the most important conclusions.'
>
> HERSCHEL, op. cit.

Residuals and Influence in Regression
R. Dennis Cook and Sanford Weisberg

Additional Multiple Regression Diagnostics

- Partial regression plots/AV plots
- Studentized deleted residuals
- Hat matrix diagonals
- Dffits, Cook's D, DFBETAS
- Variance inflation factor
- Tolerance

Insurance Example

- KNNL Page 386, Section 10.1
- Y is amount of life insurance
- X_1 is average annual income
- X_2 is a risk aversion score
- $n = 18$ managers

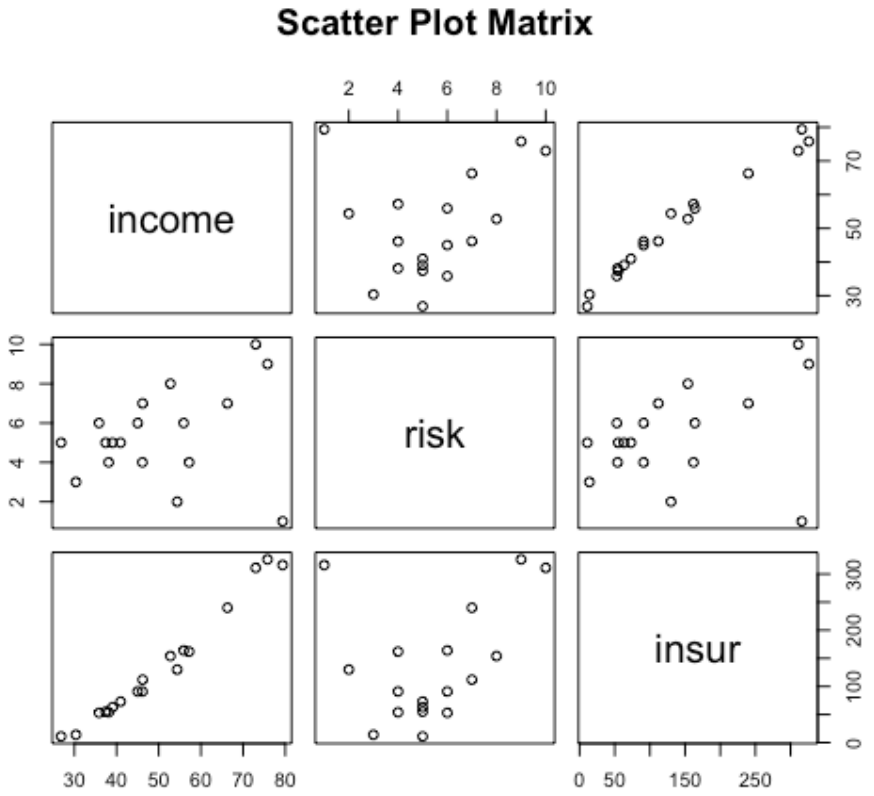
EDA

- Read in the data set

```
> a2 = read.table("CH10TA01.txt")
> colnames(a2) = c("income", "risk", "insur")
```

- Scatter plot matrix

```
> cor(a2)
> pairs(a2, main = 'Scatter Plot Matrix')
```



Partial Regression Plots(AV Plots)

- Also called added variable plots or adjusted variable plots
- One plot for each X_i
- These plots show the strength of the marginal relationship between Y and X_i in the full model (recall partial correlation)
- They can also detect
  - Nonlinear relationships
  - Heterogeneous variances
  - Outliers
- Consider the plot for X_i
  - Use the other X's to predict Y
  - Use the other X's to predict X_i
  - Plot the residuals from the first regression *vs* the residuals from the second regression

The Partial Regression Plot


```
> lmfit1 = lm(insur ~ income + risk, data=a2)
> summary(lmfit1)
> anova(lmfit1)
```

Curvilinear relationship

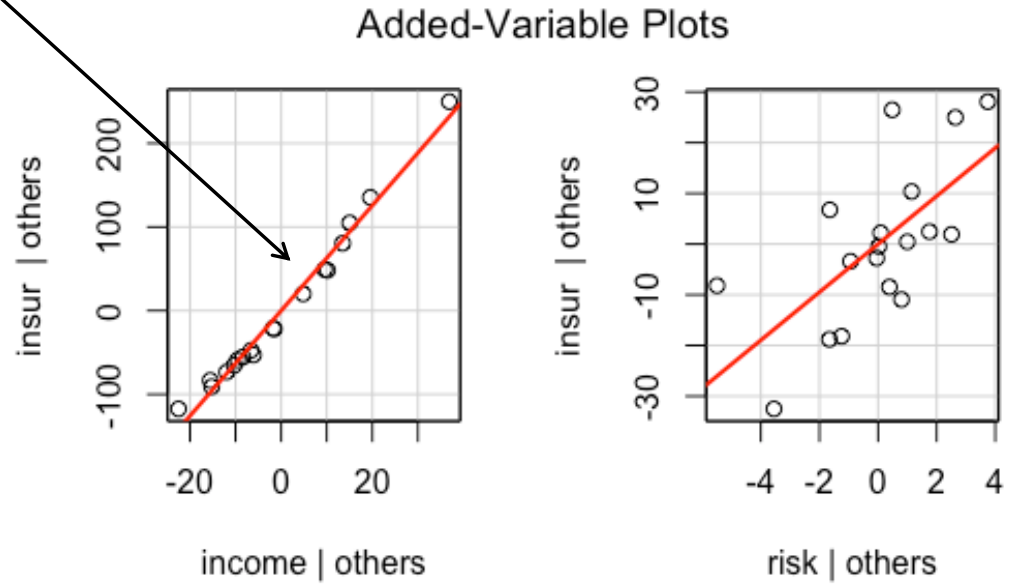
```
Coefficients:
             Estimate  Std. Error t value  Pr(>|t|)
(Intercept) -205.7187   11.3927  -18.057 1.38e-11 ***
income         6.2880    0.2041   30.801 5.63e-15 ***
risk           4.7376    1.3781    3.438  0.00366 **

Residual standard error: 12.66 on 15 degrees of freedom
Multiple R-squared:  0.9864,    Adjusted R-squared:  0.9845
F-statistic: 542.3 on 2 and 15 DF,  p-value: 1.026e-14
```

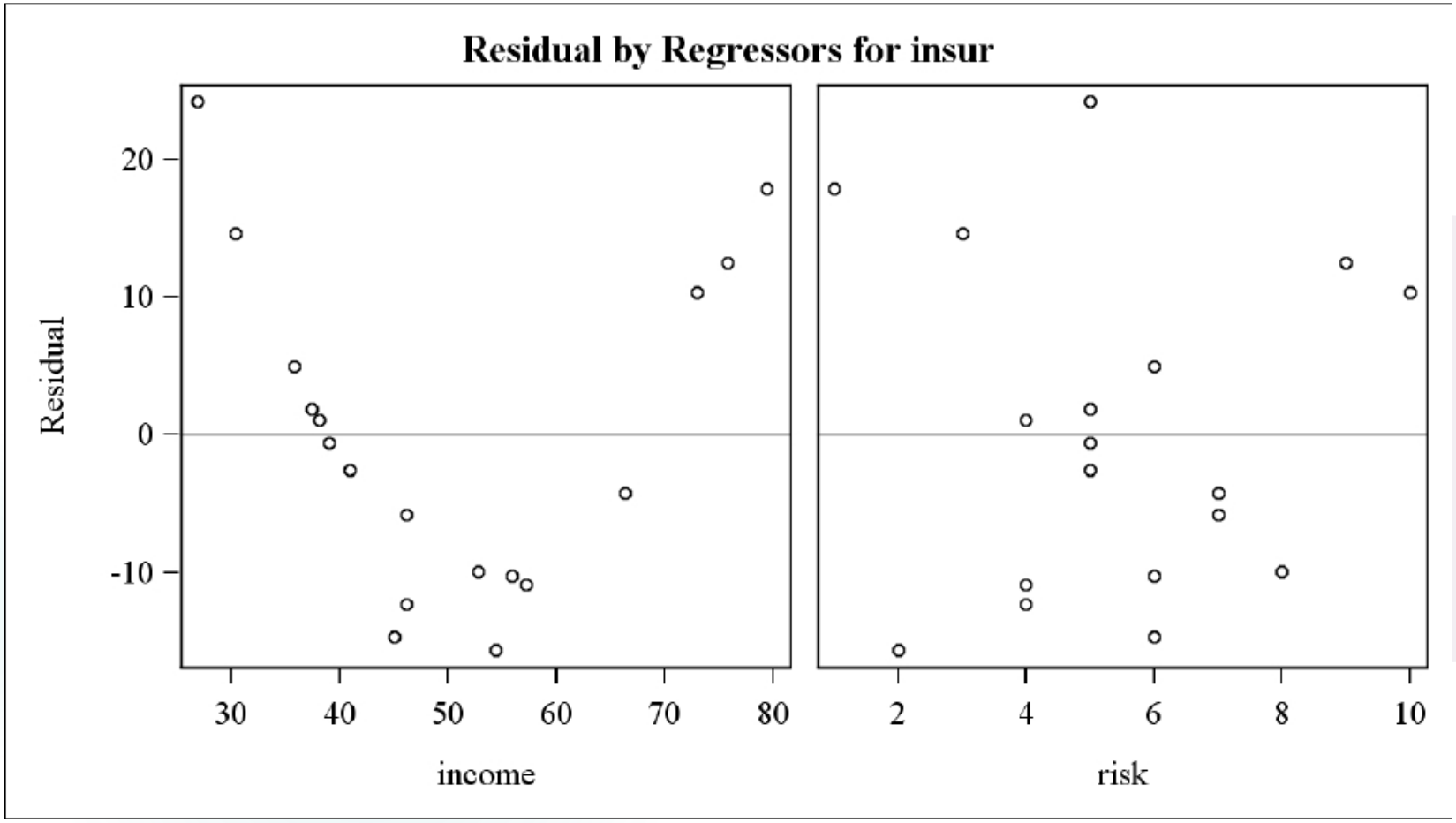
```
> library(car)
> avPlots(lmfit1)
```

Analysis of Variance Table

Response: insur

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
income	1	172024	172024	1072.851	2.268e-15 ***
risk	1	1895	1895	11.819	0.003662 **
Residuals	15	2405	160		





- Can also see the curvilinear relationship here

Residuals and Studentized Residuals

- (Ordinary) Residuals:

$$e = (e_1, e_2, \dots, e_n)' = Y - \hat{Y} = (I - H)Y$$

$$Var(e) = \sigma^2(I - H), Var(e_i) = \sigma^2(1 - h_{ii})$$

where h_{ii} is the i^{th} diagonal of H

- (Internally) Studentized residuals

$$e_i^* = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

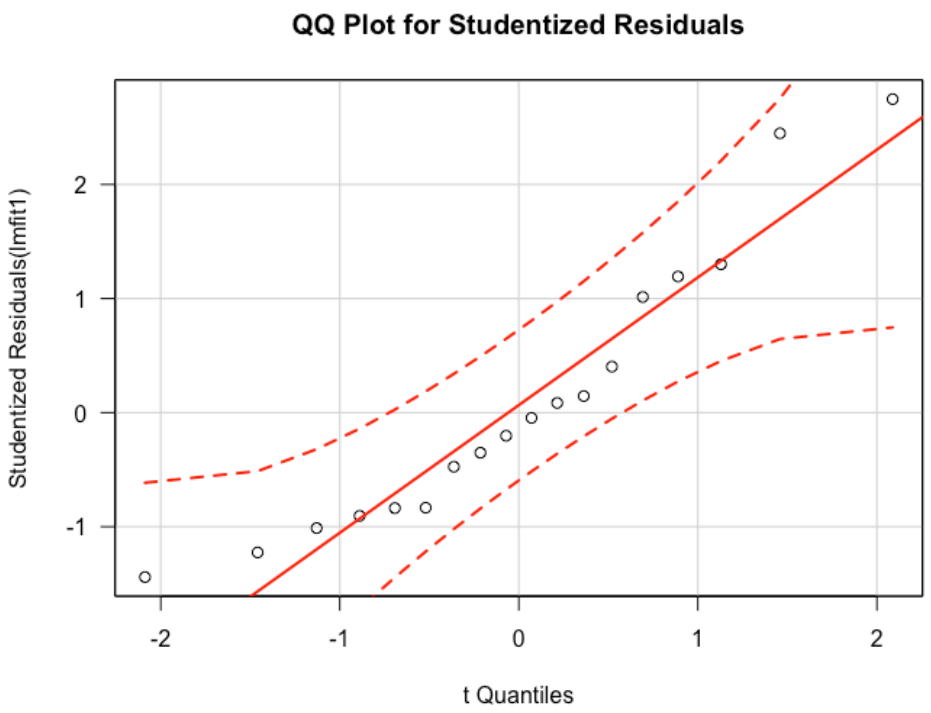
$e_1^*, e_2^*, \dots, e_n^*$ can be used to detect cases with mainly outlying Y observations

- Two types of outliers: cases with outlying Y, and cases with outlying X

The Studentized Residuals in R

- library(MASS)
- standresid = stdres(lmfit1)
- round(standresid, digits=4)

1	2	3	4	5	6
-1.2059	-0.9104	2.1208	-0.3625	-0.2096	1.0129
7	8	9	10	11	12
2.2927	-0.8456	-0.8422	0.0879	0.4151	1.1768
13	14	15	16	17	18
0.1500	-1.3923	-0.4869	-1.0112	1.2715	-0.0479



Studentized Deleted Residuals

- Deleted residuals:

$$d_i = Y_i - \hat{Y}_{i(-i)} = \frac{e_i}{1 - h_{ii}}$$

$$Var(d_i) = \frac{\sigma^2}{1-h_{ii}},\ s^2(d_i) = \frac{MSE_{(-i)}}{1-h_{ii}}$$

- Studentized deleted residuals/externally studentized residuals

$$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{MSE_{(-i)}(1-h_{ii})}} \sim t(n-p-1)$$

$MSE_{(-i)}$ is an explicit function of MSE, e_i and h_{ii}, can be easily calculated

$$(n-p)MSE = (n-p-1)\ MSE_{(-i)} + \frac{e_i^2}{1-h_{ii}}$$

- Test for outliers: declare case i has outlying Y observation if

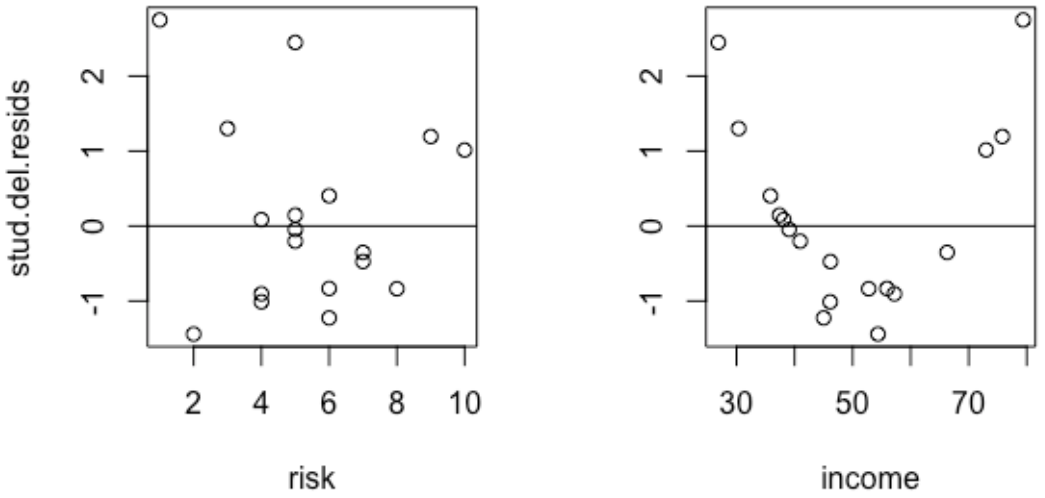
$|t_i| > t_{\alpha/2n,n-p-1}$ (Note: Bonferroni Adjustment)

The Studentized Deleted Residuals in R


```
> stud.del.resids = rstudent(lmfit1)
> round(stud.del.resids, digits=4)
```

1	2	3	4	5	6
-1.2259	-0.9048	2.4487	-0.3518	-0.2028	1.0138
7	8	9	10	11	12
-0.8371	-0.8336	0.0850	0.4033	1.1933	
13	14	15	16	17	18
0.1451	-1.4415	-0.4742	-1.0120	1.3004	-0.0462

- The studentized deleted residuals *vs* predictors



Assessing Outliers


```
> outlierTest(lmfit1)
```

```
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
     rstudent  unadjusted p-value Bonferonni p
7  2.748269          0.015698         0.28257
```

Hat Matrix and Leverage

- Hat matrix H:

$$\hat{Y} = HY, \hat{Y}_i = h_{i1}Y_1 + \cdots + h_{ii}Y_i + \cdots + h_{in}Y_n$$

- Leverage of Case i: h_{ii}

$$0 \le h_{ii} \le 1, \sum_{i=1}^{n} h_{ii} = tr(H) = rank(H) = p$$

- $\bar{h} = (\sum_{i=1}^{n} h_{ii})/n = p/n$
- h_{ii} is a measure of distance between the X values of Case i from the means of X values of all n cases; weight in prediction
- Usually declare case i has outlying X values (Outlier in X) when $h_{ii} > 2p/n$
- Moderate leverage if (0.2, 0.5); High leverage (>0.5)

The Leverage Values in R

> round(hatvalues(lmfit1), digits=4)

1	2	3	4	5	6
0.0693	0.1006	0.1890	0.1316	0.0756	0.3499
7	8	9	10	11	12
0.6225	0.1319	0.0658	0.1005	0.1201	0.2994
13	14	15	16	17	18
0.0944	0.2096	0.0957	0.0775	0.1818	0.0849

Influential Case and DIFFITS

- Influence of outliers in Y or X values needs to be carefully investigated
- Influence in the sense whether removal of an outlier can cause dramatic change in regression results (fitted model)
- An outlier that can cause big change is called an Influential case (data point)
- Suppose Case i is an outlying case
- DiFference caused to FITted values:

$$(DFFITS)_i = \frac{Y_i - \hat{Y}_{i(-i)}}{\sqrt{MSE_{(-i)} h_{ii}}} = t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

- Consider influential if DFFITS>1 (small to medium data) or

  DFFITS > $2\sqrt{p}/\sqrt{n}$ (large data)

Cook's Distances

- Cook's distance is a measure of aggregated influence of Case i:

$$D_i = \frac{\sum_{j=1}^{n}\left(Y_j - \hat{Y}_{j(-i)}\right)^2}{p \cdot MSE} = \frac{e_i^2}{p \cdot MSE}\frac{h_{ii}}{(1-h_{ii})^2}$$

- Index influence plot and also useful to check

$$v = P(F < D_i | F(p, n-p))$$

- If υ < 20%, little influence; if υ ≥ 50%, substantial influence, and Case i is influential
- Check if $D_i > 4/n$; Yes, implies that Case i is highly influential
- Note: In contrast to the DFFITS measure which considers the influence of the ith case on the fitted value Y_i for this case, Cook's distance measure is an aggregate influence measure, showing the effect of the ith case on all n fitted values

The Cook's Distances in R


```
> round(cooks.distance(lmfit1), digits=4)
```

1	2	3	4	5	6
0.0361	0.0309	0.3494	0.0066	0.0012	0.1840
[redacted]	8	9	10	11	12
	0.0362	0.0166	0.0003	0.0078	0.1973
13	14	15	16	17	18
0.0008	0.1714	0.0084	0.0286	0.1197	0.0001

DFBETAS

- Influence on the regression coefficients
- DiFference in BETA estimates:

$$(DFBETAS)_{k(-i)} = \frac{b_k - b_{k(-i)}}{\sqrt{MSE_{(-i)}c_{kk}}}, k = 0,1,\dots,p-1$$

$$Var(b_k) = \sigma^2\left((X'X)^{-1}\right)_{kk} = \sigma^2 c_{kk}$$

- A case considered influential if
  - DEBETAS > 1 for small to medium data or
  - DEBETAS > $2/\sqrt{n}$ for large data
- dfbetas(lmfit1)

	(Intercept)	income	risk
1	-0.11791502	0.124491661	-0.1107217037
2	-0.03945312	-0.146953233	0.1722774459
3	0.95935296	-0.987078887	0.1435731540
4	0.07701539	-0.082073331	-0.0410156333
5	-0.03935568	0.028583776	0.0010754435
6	-0.52978181	0.304838003	0.5125354924
7	-0.36492941	2.659822663	-2.6750533100
8	0.08157574	0.025440338	-0.2452456420
9	0.03078321	-0.067151914	-0.0365559869
10	0.02384654	-0.013764209	-0.0091627889
11	0.08634720	-0.105688246	0.0536400695
12	-0.58199873	0.449491490	0.4096139916
13	0.03482702	-0.029395861	0.0014469428
14	-0.27058334	-0.265611499	0.6268600751
15	-0.01637040	0.053207315	-0.0953091071
16	-0.18104226	0.025836093	0.1423819102
17	0.58027432	-0.360800840	-0.2577287527
18	-0.01010224	0.008033481	-0.0001311733

Variance Inflation Factor for Detecting Excessive Multicollinearity

- For standardized regression model, can show that

$$Var(b_k^*) = \sigma^{*2}(VIF)_k$$

- Variance Inflation Factor (VIF)

$$(VIF)_k = (1 - R_k^2)^{-1}$$

where R_k^2 is from regressing X_k against the other p-2 explanatory variables

- Average *VIF*:

$$\overline{VIF} = \sum_{k=1}^{p-1} (VIF)_k / (p-1)$$

- There is Excessive multicollinearity when the largest *VIF* exceeds 10 or the average *VIF* is considerably larger than 1

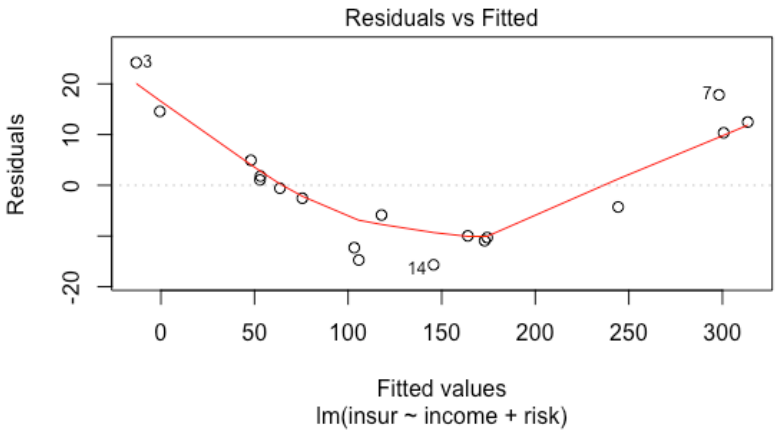
Tolerance

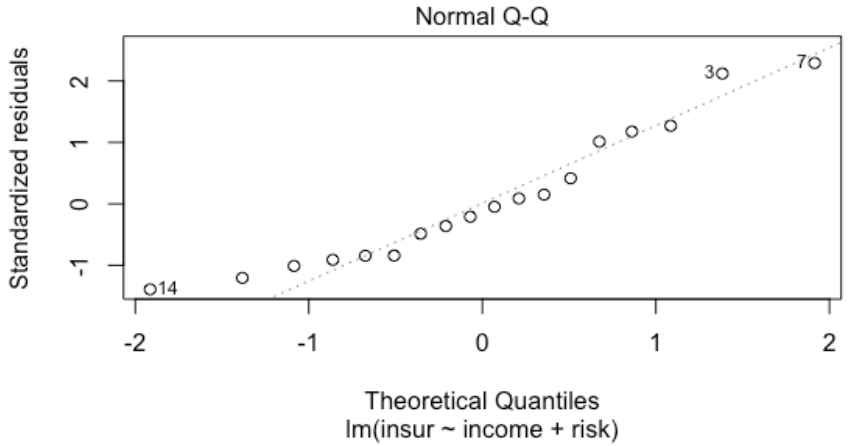
- $TOL = 1 - R_k^2$
- $TOL = 1/VIF$
- Described in comment on p 410

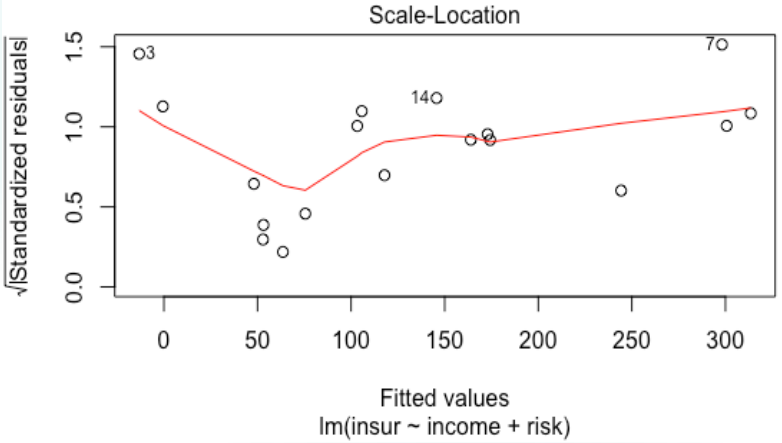
```
> library(car)
> vif(lmfit1)
```

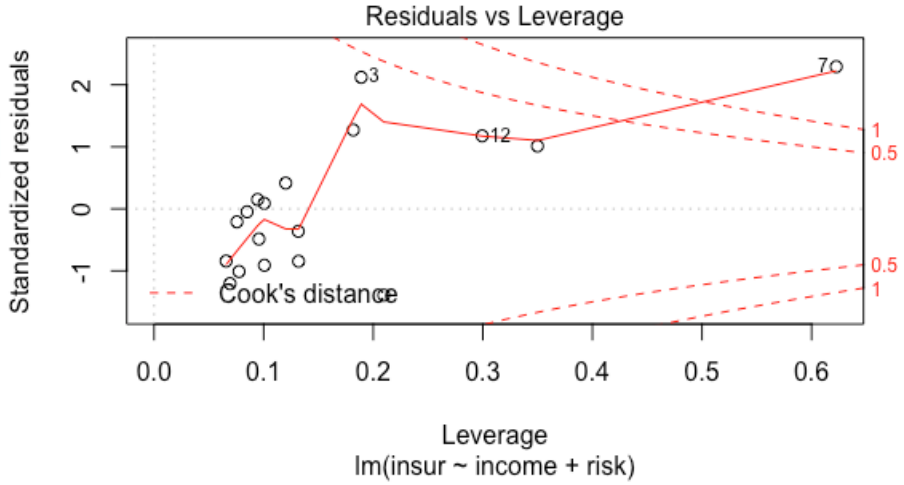
income	risk
1.069249	1.069249

Full Diagnostics

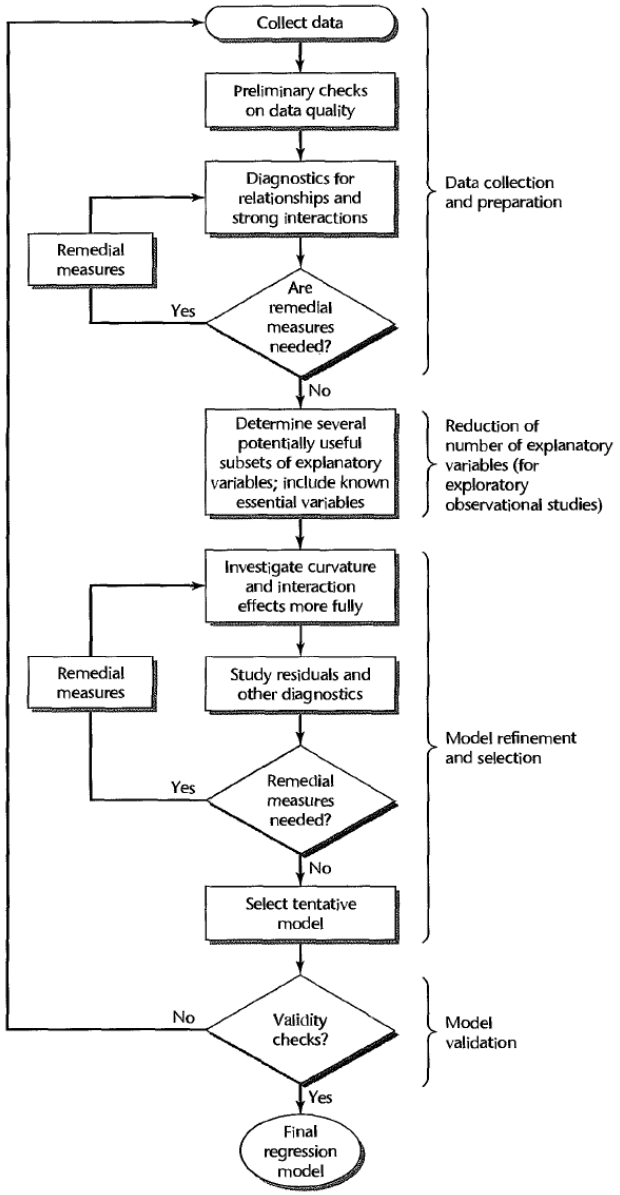








Strategy for Building a Regression Model P344



Last Slide

- We went over KNNL Chapters 9 and 10
- We used program lec10_2.R to generate the output