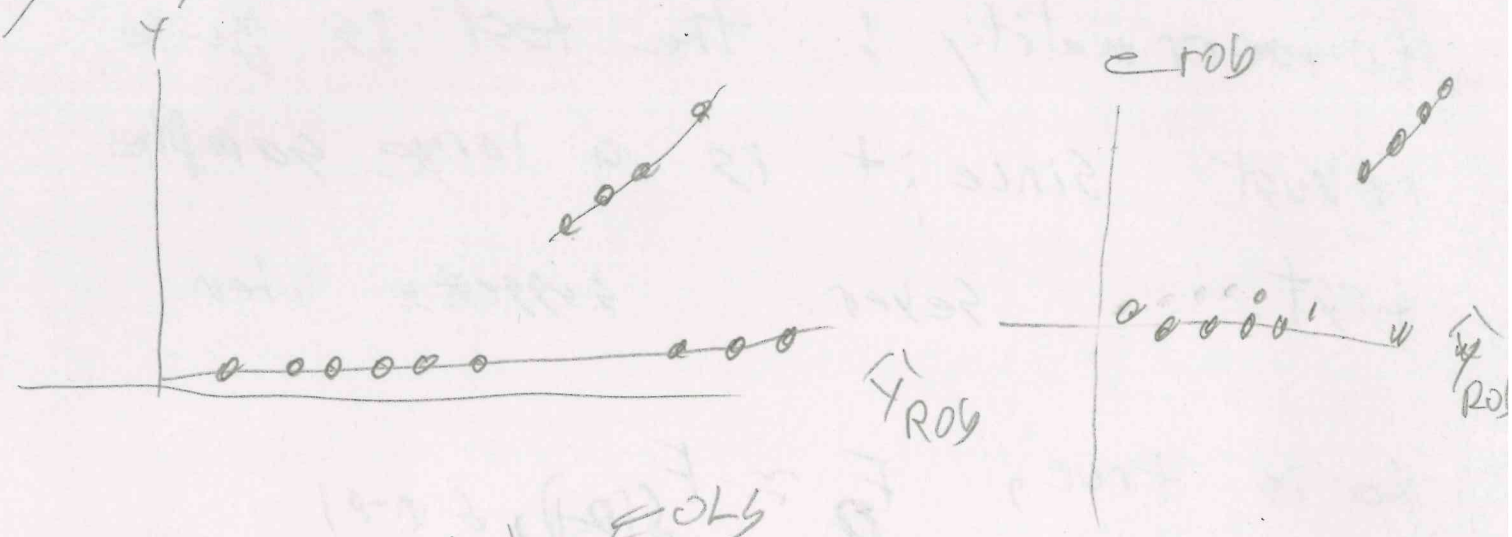
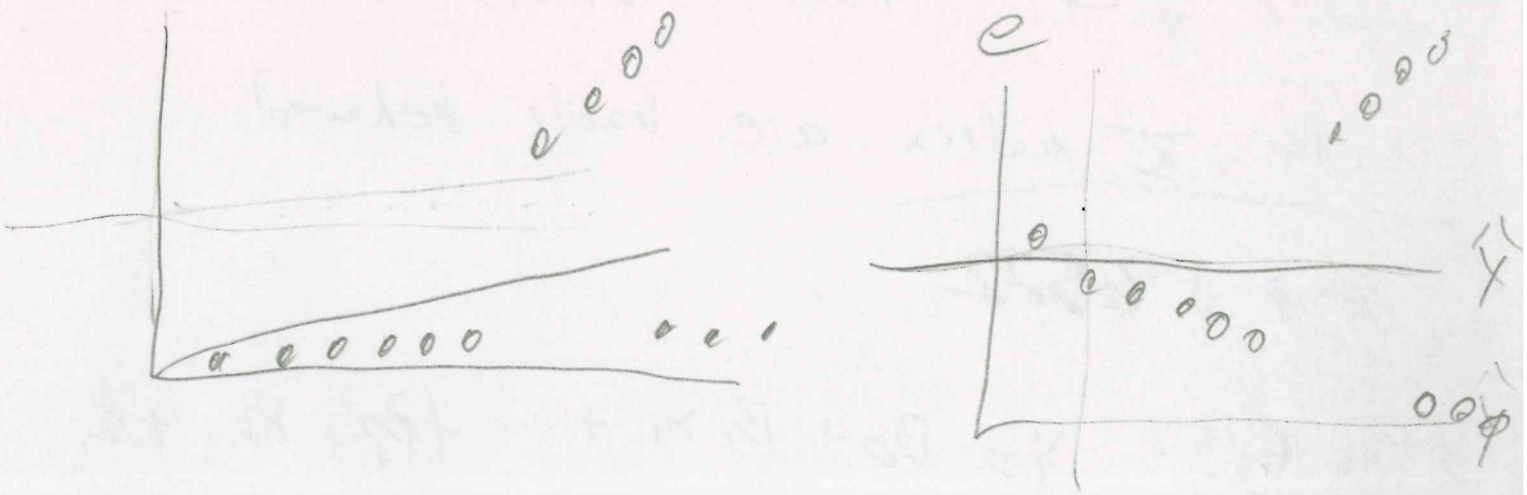
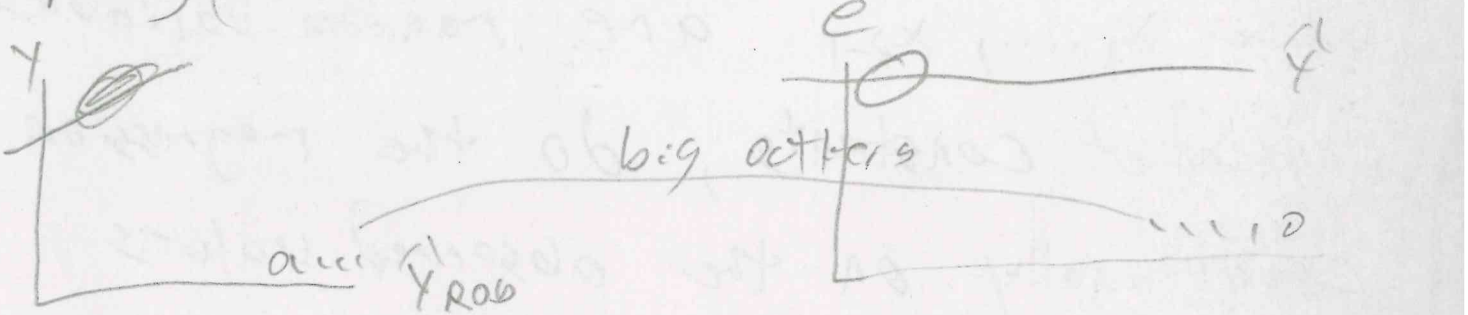
ex) rmreg2(belx, bely)



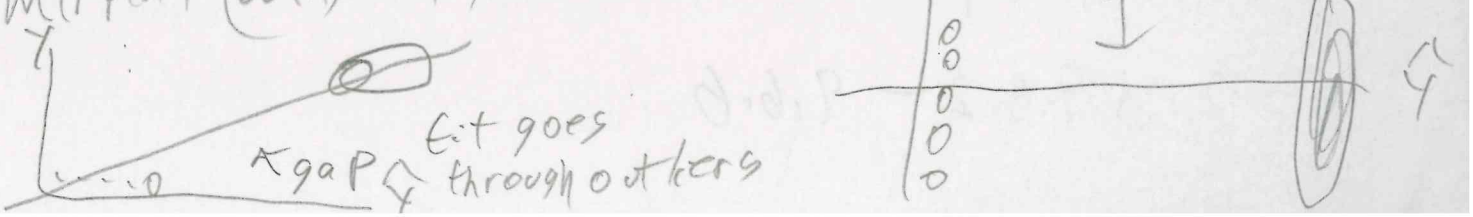
mlrplot4(belx, bely) ← OLS



ex rmreg2(boxx, boxY)



mlrplot4(butt, butY)



§9.5 (15) p235 Robustness of ANOVA F test to nonnormality: the test is quite robust since it is a large sample test. Scheffé suggests when H_0 is true, $F_0 \simeq F_{\delta(p-1),\delta(n-p)}$

where $\delta \to 1$ fast. unless the variables in the X matrix are badly behaved

---

Skip §9.5.2

§9.6 (16) If $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon_i$

where $x_1, \ldots, x_{p-1}$ are random variables instead of constants, do the regression conditionally on the observed values of $x_1, \ldots, x_{p-1}$.

Skip §9.6.2 — 9.6.6

§9.7 17) p249 Collinearity occurs when at least one column of $\underline{X} = [\underline{v}_0 \cdots \underline{v}_{p-1}]$ is highly correlated with a linear combination of the other columns: $\text{corr}\left(\underline{v}_j, \sum_{\substack{i=0 \\ i \neq j}}^{p-1} \gamma_i \underline{v}_i\right)$ for at least one j. is high. Collinearity makes $\underline{X}'\underline{X}$ nearly singular.

18) Regress $\underline{v}_j$ on $\underline{v}_0, \underline{v}_1, \ldots, \underline{v}_{j-1}, \underline{v}_{j+1}, \ldots, \underline{v}_{p-1}$ and let R_j^2 be the coefficient of determination from the regression. R_j^2 is a measure of the collinearity of $\underline{v}_j$ with the other predictors.

19) $1 - R_j^2 \approx 0$ means $\underline{v}_j \approx$ a linear combination of the other columns of $\underline{X}$ so $\underline{X}'\underline{X}$ is nearly singular.

20) p254-5 The jth variance inflation factor

$$VIF_j = \frac{1}{1 - R_j^2}.$$

Since $\text{Cov}(\hat{\underline{\beta}}) = \sigma^2 (\underline{X}'\underline{X})^{-1}$ ↑ p.249

$$SE(\hat{B}_j) = \sqrt{MSE}\, d_{jj} = \sqrt{MSE}\,(X'X)^{-1}_{j+1,j+1}$$

(see p119; p.42)

$$= \frac{\sqrt{MSE}}{SD(X_j)\sqrt{n-1}} \sqrt{\frac{1}{1-R_j^2}} = \frac{\sqrt{MSE}\,\sqrt{VIF_j}}{SD(X_j)\sqrt{n-1}}$$

$$SD(X_j) = \sqrt{\frac{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}{n-1}} = \sqrt{\hat{V}(X_j)}$$ ← sample variance. Hence $SE(\hat{B}_j)$

is large if $R_j^2 \approx 1$ or if VIF_j is large.

Then the CI for B_j will be long, and it is harder to reject $H_0: B_j = 0$ so the test has low power. Note that $VIF_j =$

$$\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2 (X'X)^{-1}_{j+1,j+1} = (n-1)\hat{V}(X_j)\, d_{jj}.$$

Skip §9.7.3, 9.7.4

2) (p261) Collinearity has little effect on prediction.

$$\hat{Y}_t = x_t^T \hat{B}, \quad V(\hat{Y}_t) = \sigma^2 x_t^T (X'X)^{-1} X' X (X'X)^{-1} x_t$$

$$= \sigma^2 x_t^T (X'X)^{-1} x_t = \sigma^2 h_t.$$ Still have

$\max h_i \to 0$. So $h_t \to 0$ if $h_t \le \max(h_i)$.

When collinearity is low, h_c = constant

$h_c = C$ when collinearity is is low

$h_c = C$ when collinearity is high

2 predictors $\underline{x}_1$, $\underline{x}_2$

Ch 10 More Departures From Assumptions

1) residual $e_i = y_i - \hat{y}_i = y_i - \underline{x}_i^T \hat{\underline{\beta}} =$

$$\underline{x}_i^T \underline{\beta} + \varepsilon_i - \underline{x}_i^T \hat{\underline{\beta}} = \varepsilon_i + x_i^T (\underline{\beta} - \hat{\underline{\beta}}) = \varepsilon_i + N_i.$$

But $\sqrt{n}(\hat{\underline{\beta}} - \underline{\beta}) \xrightarrow{D} N_p(\underline{0}, \sigma^2 W)$

So $\hat{\underline{\beta}} - \underline{\beta} \sim AN_p(\underline{0}, \sigma^2 \frac{W}{n}) \sim AN_p[\underline{0}, MSE(X'X)^{-1}]$

So $\underline{\beta} - \hat{\underline{\beta}} \sim AN_p[\underline{0}, MSE(X'X)^{-1}]$.

So $N_i = \underline{x}_i^T (\underline{\beta} - \hat{\underline{\beta}}) \sim AN_1(0, MSE\, x_i^T (X'X)^{-1} \underline{x}_i)$

$= AN_1(0, MSE\, h_i)$.

($\frac{X^TX}{n} \to W^{-1}$, $n(X'X)^{-1} \approx W$ so $\frac{W}{n} \simeq (X'X)^{-1}$)

2) So the residual $e_i = \varepsilon_i + N_i$ where

$N_i \sim AN(0, MSE\, h_i)$ and $h_i \to 0$ as $n \to \infty$. But for moderate n, the N_i term could dominate, and the e_i tend to be more normal than the ε_i if the ε_i are not iid $N(0, \sigma^2)$.

3) The residual plot $\hat{Y}$ vs. e should always be used. The plot of e vs. Y is not used since $corr(\underline{e}, \underline{Y}) = \sqrt{1 - R^2}$.

Want plots where $corr(\underline{a}, \underline{e}) = 0$.

If the predictors are no good, $R^2 \approx 0$, $\hat{Y}_i \approx \bar{Y}$, $e_i \approx Y_i - \underbrace{\bar{Y}}_{\text{constant}}$ and $corr(\underline{e}, \underline{Y}) \approx 1$.

Leadership

Prop $\text{corr}(\underset{\sim}{e}, \underset{\sim}{Y}) = 1 - R^2$:

$$\text{corr}(\underset{\sim}{e}, \underset{\sim}{Y}) = \frac{\sum (e_i - \bar{e})(y_i - \bar{y})}{\sqrt{\sum e_i^2 \; \sum (y_i - \bar{y})^2}}$$

But $\bar{e} = 0$ so top $= \sum e_i (y_i - \bar{y}) =$

$$\sum e_i y_i + \underbrace{\bar{y} \sum e_i}_{0} = \underline{e}^T \underline{Y}.$$

Now $\sum e_i^2 = \underline{e}^T \underline{e} = \underline{Y}^T \underbrace{(I - P)\underline{Y}}_{\underline{e}} = \underline{Y}^T \underline{e} = \underline{e}^T \underline{Y}$

So $\text{corr}(\underset{\sim}{e}, \underset{\sim}{Y}) = \dfrac{\sum e_i^2}{\sqrt{\sum e_i^2 \; \sum (y_i - \bar{y})^2}} = \sqrt{\dfrac{\sum e_i^2}{\sum (y_i - \bar{y})^2}}$

$$= \sqrt{\frac{SSE}{SSTO}}. \quad R^2 = 1 - \frac{SSE}{SSTO}$$

So $R^2 = 1 - \sqrt{\text{corr}(\underset{\sim}{e}, \underset{\sim}{Y})}$ or

Leadership $\text{corr}(\underset{\sim}{e}, \underset{\sim}{Y}) = \sqrt{1 - R^2}$.

4) p266-7 Use $\underline{R} = H$ in this chapter.

$$\underline{e} = (I-H)\underline{Y}, \quad \underline{\hat{Y}} = H\underline{Y}$$

$$E(\underline{e}) = \underline{0}$$

$$cov(\underline{e}) = \sigma^2(I-H)$$

$$cov(\underline{e}, \underline{\hat{Y}}) = \underset{\text{scalar}}{0}$$

$$cov(\underline{\hat{Y}}) = \sigma^2 H$$

5) Let $h_i = H_{ii}$ be the ith leverage

The internally studentized residual

$$r_i = \frac{e_i}{\sqrt{MSE}\sqrt{1-h_i}}$$ hand

externally studentized residual

$$t_i = \frac{e_i}{\sqrt{MSE_{(i)}}\sqrt{1-h_i}}$$ have $V(r_i) \approx 1$, $V(t_i) \approx 1$.

Here $MSE_{(i)}$ is the MSE from the regression model without the ith case.

6] P 269 $\sum_{i=1}^{n} h_i = p = tr(H)$

$\frac{1}{n} \leq h_i \leq 1$. if X has a column of 1's.

The average h_i value is $\frac{p}{n}$.

§10.3 →] elective $E(Y|x) = \beta' x$ with response and residual plots.

Skip §10.3.1

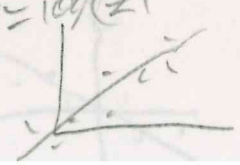
8] p276 Power transformation $t_\lambda(x) = \begin{cases} x^\lambda, & \lambda \neq 0 \\ \log(x) & \lambda = 0 \end{cases}$

are useful for removing strong nonlinearities from the predictors.

9] p297 If Z is the "response"

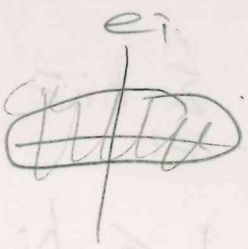
Plot $Y = t_\lambda(Z)$ vs $\hat{Y}_\lambda$ for $\lambda = -1, -\frac{1}{2}, -\frac{1}{3}, 0, \frac{1}{3}, \frac{1}{2}$

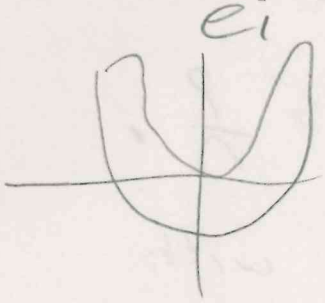
Choose λ so response plot is linear.



log transformation is good

10) plot x_i vs e_i





$\leftarrow$ add x_i^2 to model

11) p276 In a (generalized) additive model (GAM), $y_i = \alpha + \sum_{j=1}^{p-1} s_j(x_{ij}) + \varepsilon_i$

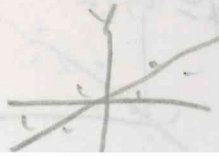
$= AP_i + \varepsilon_i$

where $s_j(x_{ij}) = B_j x_{ij}$ or s_j is unspecified. If s_j is unspecified the software will show a plot of $\hat{s}_j$.

If $\hat{s}_j$ is linear, $B_j x_{ij}$ may be good.

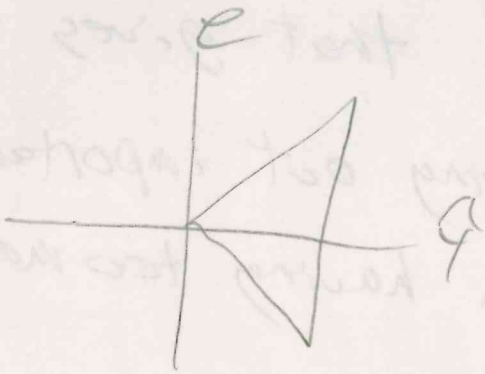
If $\hat{s}_j$ is a quadratic, add x_j^2 to the OLS model. (A GAM can also be used when you can't find an OLS model with good response and residual plots.

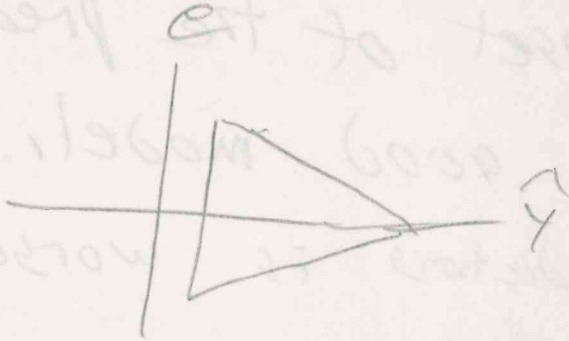
Let $EAP_i = \hat{\alpha} + \sum_{j=1}^{p-1} \hat{s}_j(x_{ij})$. The response plot of EAP vs Y should be linear for a good GAM.

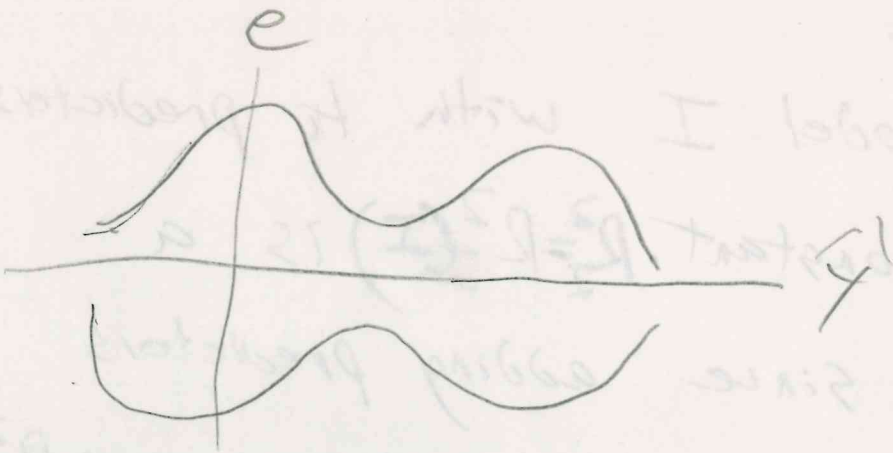


Stat 4 § 10.3

Stat 4 § 10.4 12] Detect non constant variance from residual plots



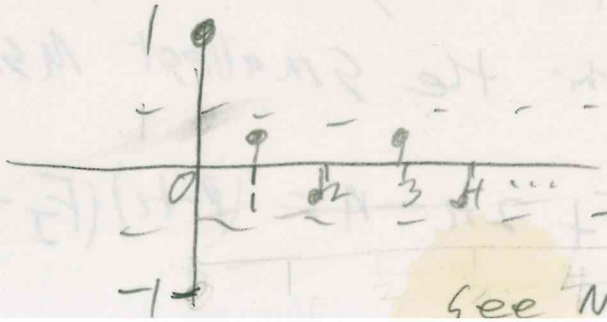




13) WLS, GLS are remedies

14) e_i could be correlated

ACF and PACF of e_i are useful



want plotted points between SE bars except at lag = 0

See Math 424 Time series

Step 6 10.5 - 10.7

Ch 12 1) Variable selection or subset selection tries to find a subset of the predictors that gives a good model. Leaving out important predictors is worse than having too many predictors.

2) p400-1 Consider model I with k predictors including a constant. $R_I^2 = R^2(I)$ is a poor measure since adding predictors does not decrease and usually increases R^2.

$MSE(I)$ and adjusted R^2 (not everyone uses this formula)

$$\bar{R}^2(I) = 1 - \{1 - R^2(I)\}\frac{n}{n-k} = 1 - MSE(I)\frac{n}{SSTO}$$

are better measures.

The model with the largest $\bar{R}^2(I)$ is also the model with the smallest $MSE(I)$.

3) $$C_p(I) = \frac{SSE(I)}{MSE(F)} + 2k - n = (p-k)(F_I - 1) + k$$

where F_I is the partial F test statistic when the reduced model is I and MSE(F) is the MSE of the full model that uses all p predictors, including a constant.

4) $corr(\underset{\uparrow}{e_F}, e_I) = \sqrt{\frac{SSE(F)}{SSE(I)}} = \sqrt{\frac{n-p}{C_p(I)+n-2k}}$

full

5) Models with $C_p(I) \leq \min(2k, p)$ are of interest.

6) Let I_{min} be the model (eg from forward selection) with the smallest $C_p(I)$. Find the submodel I_I with the fewest number of predictors such that

$C_p(I_I) \leq C_p(I_{min}) + 1$. Then model I_I is the initial model to be examined, It is possible that $I_I = I_{min}$ or $I_I = I_F$ (the full model.

7) a) models I with fewer predictors than I_I such that $C_p(I) \leq C_p(I_{min}) + 4$ should also be examined.

b) Suppose model I has k predictors including a constant. Models I with fewer predictors than I_I such that $C_p(I_{min}) + 4 < C_p(I) \leq \min(2k, p)$ should be checked but often underfit; important predictors are omitted.

8) P414-418 Forward selection and backward elimination; see E3 rev 113).

9) C_p needs a full model with $n \geq 10p$ or $5p$. Other criterion such as $\bar{R}^2(I)$ ($MSE(I)$) can be used.

If there are J potential predictors where $J > n$ (or $J > \frac{n}{5}$), use forward selection with ($\bar{R}^2(I)$ or EBIC) until there are $\approx \frac{n}{5}$ (or $\frac{n}{10}$) predictors in the model.

10) The full model should be good. Make the usual checks on the selected submodel. The response and residual plots of the submodel should be similar to those of the full model.

ex) backward elimination

current terms x1 x2 x3 x4 x5

		k	Cp
delete	x1	5	3.132
	x3	5	3.143
	x4	5	4.378
	x2	5	6.698
	x5	5	297.743

current terms x2 x3 x4 x5

		k	Cp
delete	x3	4	2.027
	x4	4	2.536
	x2	4	4.834
	x5	4	306.913

current terms x2 x4 x5

		k	Cp	
delete	x4	3	1.617	← Imin = I1
	x2	3	4.258	
	x5	3	305.047	

current terms x2 x5

		k	Cp	
delete	x2	2	4.456	← fewer predictors than I...
	x5	2	337.149	

what are the terms, including an intercept, in I_I?

Soln intercept, x_2, x_5

$k = 3$

Since x_4 was deleted

ex) forward selection

current terms

x_3, x_4, x_5

		k	C_p	
add	x_6	5	6.357	← best
	x_7	5	16.748	
	x_1	5	17.691	
	x_2	5	19.763	

"best model" has

intercept, x_3, x_4, x_5, and x_6

$k = 5$

since x_6 was added to the model

---

see HW 9 5), 6).