Lecture 8

- Model selection
- Partial regression plots
- Regression diagnostics

Variable Selection

- We want to choose a model that includes a subset of the available explanatory variables
- Two separate problems
  - How many explanatory variables should we use (subset size)
  - Given the subset size, which variables should we choose

Example

- Y is survival time
- X's are
  - Blood clotting score
  - Prognostic index
  - Enzyme function test
  - Liver function test

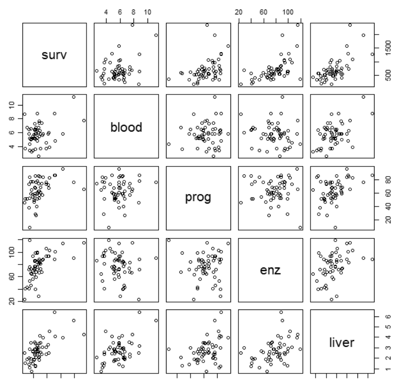
Example

- n = 54 patients
- Diagnostics suggest that Y should be transformed with a log
- Start with the usual plots and descriptive statistics

Data

```
survival<-
read.table('ch09ta01a.txt',
header=TRUE);
pairs(~surv+blood+prog+enz+live
r,survival);
```

Scatter Plot Matrix



The two problems in variable selection

- To determine an appropriate subset size you may use e.g. C_p, SBC or AIC
- For comparing models with the same number of variables, we use R^2

C_p

- The basic idea is to compare subset models with the full model
- A subset model is good if there is not substantial bias in the predicted values (relative to the full model)
- Bias - $E(\hat{Y}_i) - E(Y_i) = B_i$
- C_p is an estimator of $\sum_{i=1}^{n} B_i^2 / \sigma^2$

C_p

$$C_p = \frac{SSE_p}{MSE(F)} - (n - 2p)$$

Use of C_p

- p is the number of regression coefficients including the intercept (this is consistent with the notation we have been using)
- A model is good according to this criterion if C_p is close to or smaller than p
- Pick the smallest model for which
- C_p is close to or smaller than p or the one for which C_p is the smallest (minimize MSE for prediction)

SBC and AIC

Chose the model for which log(likelihood) - penalty for the dimension is maximal

AIC – minimize $n \log\left(\frac{SSE_p}{n}\right) + 2p$

- SBC – minimize $n \log\left(\frac{SSE_p}{n}\right) + p \log(n)$

Ordering models of the same subset size

- use R^2
- This approach can lead us to consider several models (subsets) that give us approximately the same predicted values
- We may need to apply knowledge of the subject matter to make a final selection

Proc reg

```
library("leaps");
b<-
regsubsets(lsurv~blood+prog+enz+
liver, nbest=3, survival);
u<-summary(b);
x<-cbind(u$bic,u$cp, u$rsq,
u$which)
```

				Int	blood	prog	enz	liver
1	-22.146376	66.488856	0.4275662	1	0	0	1	0
1	-21.581055	67.714773	0.4215420	1	0	0	0	1
1	-5.497592	108.555776	0.2208467	1	0	1	0	0
2	-46.813822	20.519679	0.6632899	1	0	1	1	0
2	-37.443097	33.504067	0.5994837	1	0	0	1	1
2	-30.988866	43.851738	0.5486346	1	1	0	1	0
3	-60.502425	3.390508	0.7572918	1	1	1	1	0
3	-52.364713	11.423673	0.7178164	1	0	1	1	1
3	-35.185709	32.931969	0.6121232	1	1	0	1	1
4	-56.942091	5.000000	0.7592108	1	1	1	1	1

Other approaches

- Maximize adjusted R^2
- PRESS (prediction SS)
  - For each case i
  - Delete the case and predict Y using a model based on the other n-1 cases
  - Look at the SS for observed minus predicted

Other approaches (2)

- Step type procedures
  - Forward selection (Step up)
  - Backward elimination (Step down)
  - Stepwise (forward selection with a backward glance)

Partial regression plots

- Also called added variable plots or adjusted variable plots
- One plot for each X_i

Partial regression plots (2)

- Consider X_1
  - Use the other X's to predict Y
  - Use the other X's to predict X_1
  - Plot the residuals from the first regression vs the residuals from the second regression

Partial regression plots (3)

- These plots show the strength of relatioship between Y and X_i in the full model. They can also detect
  - Nonlinear relationships
  - Heterogeneous variances
  - Outliers

Example

- Y is amount of life insurance
- X_1 is average annual income
- X_2 is a risk aversion score
- n = 18 managers

Create a data set

```
insurance<-read.table
('ch10ta01.txt', col.names=
c("income", "risk", "insurance"));
```

The partial option with proc reg

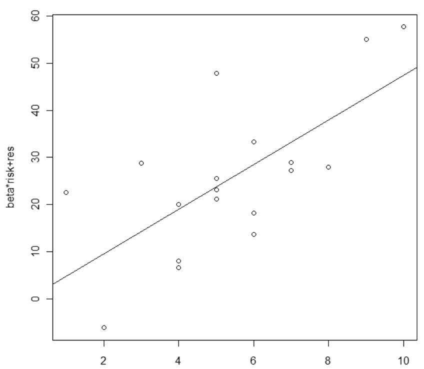
```
library("faraway");
reg1<-lm(insurance~income+risk,
insurance);
prplot(reg1,1);
prplot(reg1,2);
summary(reg1);
```

Output

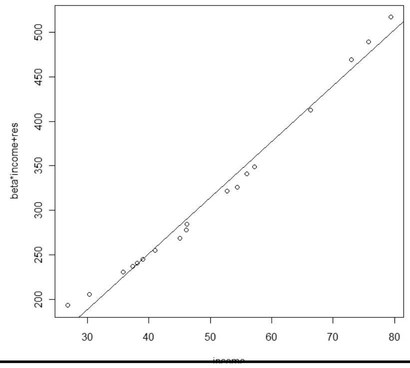
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -205.7187    11.3927 -18.057 1.38e-11 ***
income         6.2880     0.2041  30.801 5.63e-15 ***
risk           4.7376     1.3781   3.438  0.00366 **
---

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
```

The plot for risk



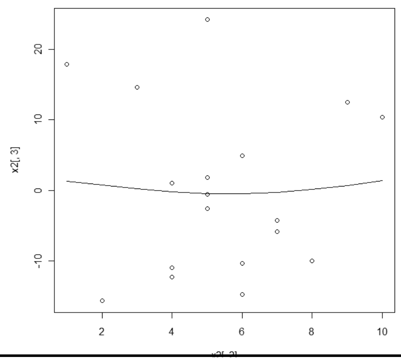
The plot for income



Plot the residuals vs risk

```
x<-cbind(insurance$income,
insurance$risk,reg1$residuals);
x2<-x[order(x[,2]),];
plot(x2[,3]~x2[,2]);
s<-smooth.spline(x2[,2],x2[,3],
spar=0.7);
lines(s);
```

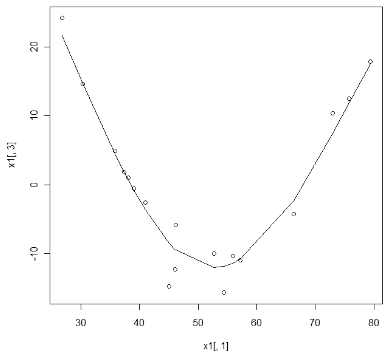
The graph



Plot residuals vs income

```
x1<-x[order(x[,1]),];
plot(x1[,3]~x1[,1]);
s<-smooth.spline(x1[,1],x1[,3], spar=0.7);
lines(s);
```

Plot residuals vs income



Regression Diagnostics

- **Studentized deleted residuals**
- **Hat matrix diagonals**
- **Dffits, Cook’s D, DFBETAS**
- **Variance inflation factor**
- **Tolerance**

Residuals

- There are several versions
  - Residuals
    - $(Y_i - \hat{Y}_i) = e_i$
  - Studentized residuals

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

    - Studentized means dividing by the standard error
    - These are like $t_{(n-p)}$

Residuals (2)

- Studentized deleted residuals
  - Deleted means delete case *i* when computing this residual for case i

Residuals (3)

- We use the notation (i) to indicate that case i has been deleted from the computations
- $Y_{(i)} = Y_i - \hat{Y}_{i(i)}$ is the deleted residual

$Y_{(i)} = e_i/(1-h_{ii})$

Var $Y_{(i)}$ =Var $e_i/(1-h_{ii})^2 = MSE_{(i)}/(1-h_{ii})$

$MSE_{(i)}$ is the MSE with case i deleted

- The studentized deleted residual is

$$\frac{Y_{(i)}\sqrt{(1-h_{ii})}}{\sqrt{MSE_{(i)}}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

Residuals (4)

- When we examine the residuals we are looking for
  - Outliers
  - Non normal error distributions
  - Influential observations

Studentized residuals

```
x1<-rstandard(reg1);
x2<-rstudent(reg1);
x<-cbind(x1,x2);
```

Output

```
        x1          x2
1  -1.20587814 -1.22592579
2  -0.91036231 -0.90484533
3   2.12082543  2.44867347
4  -0.36253288 -0.35178460
5  -0.20962843 -0.20281761
6   1.01288864  1.01382844
7   2.29272103  2.74826933
8  -0.84557683 -0.83709929
9  -0.84223637 -0.83362782
10  0.08793325  0.08497349
11  0.41506608  0.40331472
12  1.17680637  1.19332347
13  0.15004659  0.14506769
14 -1.39233371 -1.44149247
15 -0.48693378 -0.47418536
16 -1.01122970 -1.01204637
17  1.27145643  1.30041597
18 -0.04785973 -0.04624043
```

Hat matrix diagonals

- h_{ii} is a measure of how much Y_i is contributing to the prediction of $\hat{Y}_i$
- $\hat{Y}_1 = h_{11}Y_1 + h_{12} Y_2 + h_{13}Y_3 + \ldots$
- h_{ii} is sometimes called the leverage of the i^{th} observation

Hat matrix diagonals (2)

- $0 \leq h_{ii} \leq 1$
- $\text{Sum}(h_{ii}) = p$
- Large value of h_{ii} suggess that i – th case is distant from the center of all X's
- The average value is p/n
- Values far from this average point to cases that should be examined carefully

Hat diagonals

```
h<-matrix(hatvalues(reg1),18,1);
 [1,] 0.06928999
 [2,] 0.10064451
 [3,] 0.18901274
 [4,] 0.13157726
 [5,] 0.07559158
 [6,] 0.34985551
 [7,] 0.62250833
 [8,] 0.13187873
 [9,] 0.06575455
[10,] 0.10052380
[11,] 0.12011384
[12,] 0.29940207
[13,] 0.09441512
[14,] 0.20960495
[15,] 0.09569345
[16,] 0.07752426
[17,] 0.18175654
[18,] 0.08485276
```

DFFITS

- A measure of the influence of case i on $\hat{Y}_i$
- It is a standardized version of the difference between $\hat{Y}_i$ computed with and without case i
- It is closely related to h_{ii}
- (1 for small data sets $2\sqrt{p/n}$ for large)

Cook's Distance

- A measure of the influence of case i on all of the $\hat{Y}_i$'s
- It is a standardized version of the sum of squares of the differences between the predicted values computed with and without case i
- (median of F(p,n-p))

DFBETAS

- **A measure of the influence of case i on each of the regression coefficients**
- **It is a standardized version of the difference between the regression coefficient computed with and without case i**
- **(1 for small data sets $2/\sqrt{n}$ for large)**

Variance Inflation Factor

- **The VIF is related to the variance of the estimated regression coefficients**
- **$VIF_k=(1 - R^2_k)^{-1}$, where R^2_k is the squared multiple correlation obtained in a regression where all other explanatory variables are used to predict X_k**

VIF and Tolerance

- **We calculate it for each explanatory variable**
- **One suggested rule is that a value of 10(0) or more for VIF indicates excessive multicollinearity**
- **TOL = 1/VIF**

Full diagnostics

```
x1<-dffits(reg1);
x2<-cooks.distance(reg1);
x3<-dfbeta(reg1);

res<-cbind(x1,x2,x3);
library("HH");
v<-vif(reg1);
```

Output (influence)

	x1	x2	(Intercept)	income	risk
1	-0.33449	3.6086e-02	-1.3214	0.024999295	-0.1500880915
2	-0.30269	3.0914e-02	-0.4522	-0.030183342	0.2388595313
3	1.18214	3.4943e-01	9.4662	-0.174531639	0.1713647622
4	-0.13693	6.6377e-03	0.9042	-0.017267160	-0.0582497731
5	-0.05799	1.1978e-03	-0.4634	0.006031320	0.0015318152
6	0.74371	1.8402e-01	-6.0300	0.062174829	0.7056598554
7	3.52921	2.8894e+00	-3.4683	0.452994431	-3.0753786845
8	-0.32626	3.6205e-02	0.9387	0.005246229	-0.3413906605
9	-0.22115	1.6642e-02	0.3543	-0.013850595	-0.0508972758
10	0.02840	2.8804e-04	0.2811	-0.002907831	-0.0130668697
11	0.14901	7.8393e-03	1.0123	-0.022204864	0.0760740853
12	0.78010	1.9727e-01	-6.5387	0.090493328	0.5566674725
13	0.04684	7.8242e-04	0.4103	-0.006207116	0.0020624386
14	-0.74231	1.7136e-01	-2.9775	-0.052375248	0.8344029029
15	-0.15425	8.3634e-03	-0.1915	0.011154275	-0.1348746687
16	-0.29338	2.8645e-02	-2.0608	0.005270169	0.1960552402
17	0.61289	1.1969e-01	6.4636	-0.072016993	-0.3472612267
18	-0.01408	7.0793e-05	-0.1191	0.001697464	-0.0001870976

Output (tolerance)

- **income risk**
- **1.069249 1.069249**

Regression Diagnostics Summary

- Check normality of the residuals with a normal quantile plot
- Plot the residuals versus predicted values, versus each of the X's and (where appropriate) versus time
- Examine the partial regression plots
  - If there appears to be a curvilinear pattern, generate the graphics version with a smooth

Regression Diagnostics Recommendations (2)

- 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

Regression Diagnostics Recommendations (3)

- 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

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i, \quad \mathrm{Var}(\xi_i) = \sigma_i^2$$

$$Y_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma_i^2\right)$$

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

$$L = f_1 \cdot f_2 \cdot \ldots \cdot f_n - \text{likelihood function}$$

Weighted regression

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

$$\sum \frac{1}{\sigma_i^2}\left(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{ip-1}\right)^2$$

- Weights $w_i = 1/\sigma_i^2$

Weighted least squares

- Least squares problem is to minimize the sum of w_i times the squared residual for case i
- Computations are easy, use the weight statement in proc lm
- $b_w = (X'WX)^{-1}(X'WY)$
  - where W is a diagonal matrix with the weights
- The problem is to determine the weight

Determination of weights

- Find a relationship between the absolute residual and another variable and use this as a model for the standard deviation
- Similarly for the squared residual and the variance
- Use grouped data or approximately grouped data to estimate the variance

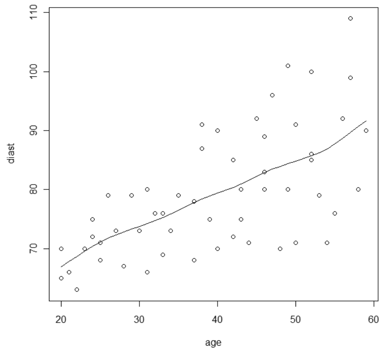
Example

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

Get the data

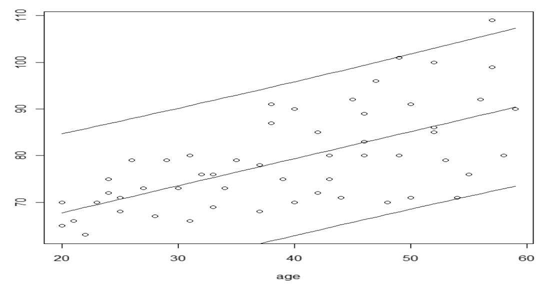
```
pressure<-read.table('ch10ta01.dat',
col.names=c("age", "diast"));
pressure<-pressure
[order(pressure$age),];
plot(diast~age, pressure);
s<-smooth.spline(pressure$age,
pressure$diast, spar=0.7);
lines(s);
```

Diastolic bp vs age



Prediction intervals (1)

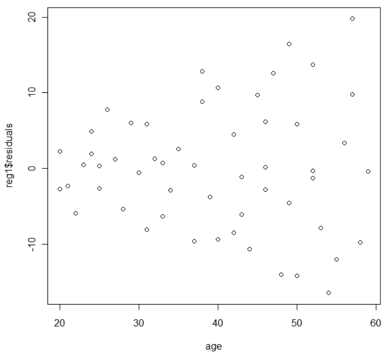
- reg1=lm(diast~age, pressure);
- c1<-predict.lm(reg1, se.fit=TRUE, interval='prediction');
- plot(diast~age, pressure)
- lines(c1$fit[,1]~age, pressure)
- lines(c1$fit[,2]~age,pressure)
- lines(c1$fit[,3]~age,pressure)



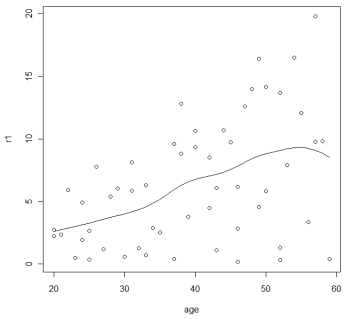
Calculate the absolute and squared residuals

```
r1<-abs(reg1$residuals);
r2<-reg1$residuals^2;
plot(reg1$residuals~age,pressure);
plot(r1~age,pressure);
s<-smooth.spline(pressure$age,r1,
spar=0.7);
lines(s);
plot(r2~age,pressure);
s<-smooth.spline(pressure$age,r2,
spar=0.7);
lines(s);
```

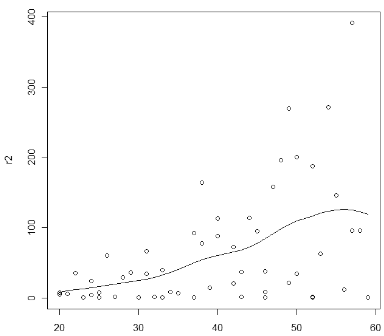
Residuals vs age



Absolute value of the residuals vs age

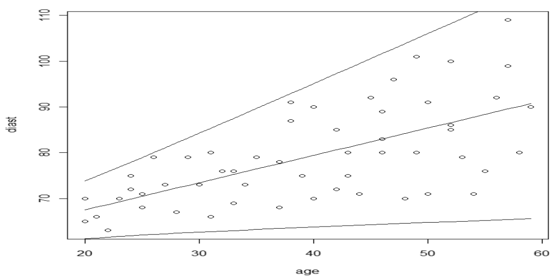


Squared residuals vs age



Calculate weights

```
reg2<-lm(r1~age, pressure);
c1<-predict.lm(reg2);
w<-1/(c1^2);
reg3<-lm(diast~age,
weights=w,pressure);
c1<-predict.lm(reg3,  se.fit=TRUE,
interval='prediction');
plot(diast~age, pressure)
lines(c1$fit[,1]~age, pressure)
lines(c1$fit[,2]~age,pressure)
lines(c1$fit[,3]~age,pressure)
```



Ridge regression

- Similar to a very old idea in numerical analysis
- If (X'X) is difficult to invert (near singular) then approximate by inverting (X'X+kI).
- Estimators of coefficients are biased but more stable.
- For some value of k ridge regression estimator has a smaller mean square error than ordinary least square estimator.
- Interesting but has not turned out to be a useful method in practice .
- Library(''MASS", lm.ridge)

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
  - Reiterated weighted linear regression
  - e.g. rlm function in library ''MASS"

Nonparametric regression

- Several versions
- We have used smoothed splines
- Interesting theory
- All versions have some smoothing parameter similar to the *par=0.7*
- Confidence intervals and significance tests not fully developed

Bootstrap

- Very important theoretical development that has a major impact on applied statistics
- Based on simulation
- Sample *with* replacement from the data or residuals and get the distribution of the quantity of interest
- CI based on quantiles of the sampling distribution

Model validation

- Three approaches to checking the validity of the model
  - Collect new data, does it fit the model
  - Compare with theory, other data, simulation
  - Use some of the data for the basic analysis and some for validity check

One qualitative explanatory variable

- Indicator (or dummy) variables have the value 0 when the quality is absent and 1 when the quality is present
- Examples include
  - Gender as an explanatory variable
  - Placebo versus control

Binary predictor

- X_1 has values 0 and 1 corresponding to two different groups
- X_2 is a continuous variable
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \xi$
- For $X_1 = 0$, $Y = \beta_0 + \beta_2 X_2 + \xi$
- For $X_1 = 1$, $Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2 + \xi$

Binary predictor

- For $X_1 = 0$, $Y = \beta_0 + \beta_2 X_2 + \xi$
- For $X_1 = 1$, $Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2 + \xi$
- H_0: $\beta_1 = \beta_3 = 0$ tests the hypothesis that the lines are the same
- H_0: $\beta_1 = 0$ tests equal intercepts
- H_0: $\beta_3 = 0$ tests equal slopes

More models

- If a categorical (qualitative) variable has several *k* possible values we need *k-1* indicator variables
- These can be defined in many different ways;
- We also can have several categorical explanatory variables, interactions, etc

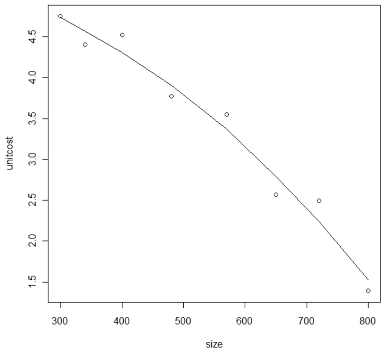
More models (2)

- Piecewise linear regression
- At some (known) point we allow the slope to change

Example

- NKNW p 476
- Y is unit cost
- X_1 is lot size
- The slope is allowed to change at a lot size of 500
- n = 8

Plot the data



Model

- Our model has
  - An intercept
  - A coefficient for lotsize (the slope)
  - An additional explanatory variable that will add a constant to the slope whenever lotsize is greater than 500

New variable

```
ind<-as.numeric(cost$size>500);
cost$cslope<-ind*(cost$size-500);
unitcost size cslope
1     2.57  650    150
2     4.40  340      0
3     4.52  400      0
4     1.39  800    300
5     4.75  300      0
6     3.55  570     70
7     2.49  720    220
8  3.77  480      0
reg3<-lm(unitcost~size+cslope,
 cost);
```

Results of regression

```
Coefficients:
             Est      Std      t  Pr(>|t|)
Int         5.895    0.604   9.757 0.0001 **
size       -0.003    0.001 -2.650 0.0454 *
cslope     -0.003    0.002 -1.685 0.1527

Residual standard error: 0.2449
on 5 degrees of freedom
Multiple R-squared: 0.9693,
F-statistic: 79.06 on 2 and 5 DF,
p-value: 0.0001645
```

Plot data with fit

```
cost<-cost[order(cost$size),];
reg3<-lm(unitcost~size+cslope,
cost);
x1<-predict.lm(reg3);
plot(unitcost~size, cost);
lines(x1~size, cost);
```

The plot