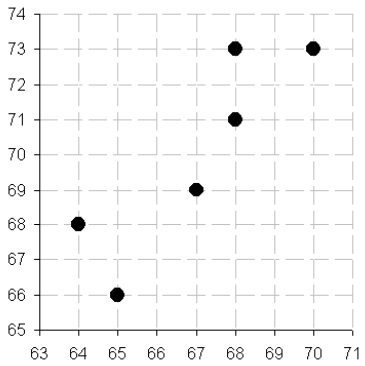
STAT 420 Fall 2012

Homework #5

(due Friday, October 12, by 3:00 p.m.)

1. A students wonders if people of similar heights tend to date each other. She measures herself, her dormitory roommate, and the women in the adjoining rooms; then she measures the next man each woman dates. Here are the data (heights in inches):



Women, x	68	64	68	67	70	65
Men, y	73	68	71	69	73	66

$\sum x = 402, \quad \sum y = 420, \quad \sum x^2 = 26{,}958, \quad \sum y^2 = 29{,}440, \quad \sum xy = 28{,}167,$

$\sum (x - \bar{x})^2 = 24, \quad \sum (y - \bar{y})^2 = 40, \quad \sum (x - \bar{x})(y - \bar{y}) = \sum (x - \bar{x})\, y = 27.$

Assume that (X, Y) have a bivariate normal distribution.

a) Find the sample correlation coefficient r between the heights of the women and men.

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}} = \frac{27}{\sqrt{24}\sqrt{40}} \approx \mathbf{0.87142}.$$

b) Test $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$ at $\alpha = 0.05$. What is the p-value of this test? (You may give a range for the p-value.)

Test Statistic: $$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.87142\sqrt{6-2}}{\sqrt{1-0.87142^2}} \approx \mathbf{3.553}.$$

Rejection Region: Reject H_0 if $T < -t_{0.025}$ or $T > t_{0.025}$

$n - 2 = 4$ degrees of freedom $\quad t_{0.025}(4) = \mathbf{2.776}.$

Reject H_0: $\rho = 0$ at $\alpha = 0.05$.

$t_{0.025}(4) = 2.776 < 3.553 < 3.747 = t_{0.01}(4)$

p-value = 2 tails. $\quad \mathbf{0.02} < \text{p-value} < \mathbf{0.05} \quad (\text{p-value} \approx 0.02556)$

c) Test $H_0 : \rho = 0.3$ vs. $H_1 : \rho > 0.3$ at $\alpha = 0.05$. What is the p-value of this test?

$$W = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\cdot\ln\left(\frac{1+0.87142}{1-0.87142}\right) = 1.33895.$$

Under H_0, $\mu_W = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} = \frac{1}{2}\cdot\ln\left(\frac{1+0.30}{1-0.30}\right) = 0.30952,$

$$\sigma_W^2 = \frac{1}{n-3} = \frac{1}{3}.$$

Test Statistic: $Z = \dfrac{W-\mu_W}{\sigma_W} = \dfrac{1.33895-0.30952}{\sqrt{1/3}} = \mathbf{1.783}.$

Rejection Region: Rejects H_0 if $z > z_{0.05}$

$z_{0.05} = 1.645.$ **Reject H_0.**

P-value = right tail = $P(Z > 1.783) = \mathbf{0.0375}$.

d) Test $H_0 : \rho = 0.5$ vs. $H_1 : \rho \neq 0.5$ at $\alpha = 0.05$. What is the p-value of this test?

$$W = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\cdot\ln\left(\frac{1+0.87142}{1-0.87142}\right) = 1.33895.$$

Under H_0, $\mu_W = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} = \frac{1}{2}\cdot\ln\left(\frac{1+0.50}{1-0.50}\right) = 0.54931,$

$$\sigma_W^2 = \frac{1}{n-3} = \frac{1}{3}.$$

Test Statistic: $Z = \dfrac{W-\mu_W}{\sigma_W} = \dfrac{1.33895-0.54931}{\sqrt{1/3}} = \mathbf{1.368}.$

Rejection Region: Rejects H_0 if $Z < -z_{0.025}$ or $Z > z_{0.025}$

$z_{0.025} = 1.96.$ **Do NOT Reject H_0.**

P-value = 2 tails = $2 \times P(Z > 1.368) = 2 \times 0.0853 = \mathbf{0.1706}$.

e) Construct a 95% confidence interval for ρ.

$100(1-\alpha)$ % confidence interval for ρ:

$$\left(\frac{e^{a}-1}{e^{a}+1}, \frac{e^{b}-1}{e^{b}+1} \right), \quad \text{where} \quad a = \ln\frac{1+r}{1-r} - \frac{2z_{\alpha/2}}{\sqrt{n-3}}, \quad b = \ln\frac{1+r}{1-r} + \frac{2z_{\alpha/2}}{\sqrt{n-3}}.$$

$$a = \ln\frac{1+r}{1-r} - \frac{2z_{\alpha/2}}{\sqrt{n-3}} = 2.6779 - \frac{2\cdot 1.96}{\sqrt{3}} = 0.4147.$$

$$b = \ln\frac{1+r}{1-r} + \frac{2z_{\alpha/2}}{\sqrt{n-3}} = 2.6779 + \frac{2\cdot 1.96}{\sqrt{3}} = 4.9411.$$

$$\left(\frac{e^{0.4147}-1}{e^{0.4147}+1}, \frac{e^{4.9411}-1}{e^{4.9411}+1} \right) \qquad \mathbf{(0.2044, 0.9858)}$$

f) If every woman wore 2-inch heels when she was measured, what is the correlation between the actual female and male heights? Justify your answer.

The correlation coefficient is not affected by adding (or subtracting) the same number to all the values of one variable.

$r \approx \mathbf{0.87142}$.

g) If every woman dated a man exactly 3 inches taller than herself, what would be the correlation between female and male heights? Justify your answer.

Perfect lineal relationship: $y = x + 3$ $\qquad r = \mathbf{1}$.

2. The dataset `prostate` comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. The data frame has 97 rows and 9 columns:

`lcavol`	log(cancer volume)
`lweight`	log(prostate weight)
`age`	age
`lbph`	log(benign prostatic hyperplasia amount)
`svi`	seminal vesicle invasion
`lcp`	log(capsular penetration)
`gleason`	Gleason score
`pgg45`	percentage Gleason scores 4 or 5
`lpsa`	log(prostate specific antigen)

(Source: Andrews D.F. and Herzberg A.M. (1985) *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer-Verlag, New York.)

```
> library(faraway)
> data(prostate)
> prostate[1:5,]        ### so we can see what the data set looks like
      lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
2 -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
3 -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
4 -1.2039728  3.2828  58 -1.386294   0 -1.38629       6     0 -0.16252
5  0.7514161  3.4324  62 -1.386294   0 -1.38629       6     0  0.37156
> attach(prostate)
```

The data set is also available Compass. `prostate.csv` is a comma-delimited value file. The first line contains the names for the variables.

Fit a model with `lpsa` as the response and the other variables as predictors.

```
> fit = lm(lpsa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45)
> summary(fit)

Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
    gleason + pgg45)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7331 -0.3713 -0.0170  0.4141  1.6381

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669337   1.296387   0.516  0.60693
lcavol       0.587022   0.087920   6.677 2.11e-09 ***
lweight      0.454467   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766157   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045142   0.157465   0.287  0.77503
pgg45        0.004525   0.004421   1.024  0.30886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7084 on 88 degrees of freedom
Multiple R-Squared: 0.6548,     Adjusted R-squared: 0.6234
F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

a) Compute 90 and 95% CIs for the parameter associated with `age`. Using just these intervals, what could we have deduced about the p-value for `age` in the regression summary? (3.1 (a))

```
> confint(fit,"age",level=0.95)
            2.5 %       97.5 %
age -0.04184062 0.002566267
> confint(fit,"age",level=0.90)
             5 %          95 %
age -0.0382102 -0.001064151
```

OR

95% CI for β_{age} :

```
> qt(0.975,88)
[1] 1.98729
```

$$t_{0.025}(88) = 1.98729$$

$$-0.019637 \pm 1.98729 \times 0.011173$$

$$-0.019637 \pm 0.022204$$

$$\mathbf{(-0.041841,\ 0.002567)}$$

90% CI for β_{age} :

```
> qt(0.95,88)
[1] 1.662354
```

$$t_{0.05}(88) = 1.662354$$

$$-0.019637 \pm 1.662354 \times 0.011173$$

$$-0.019637 \pm 0.018573$$

$$\mathbf{(-0.038210,\ -0.001064)}$$

Consider $H_0: \beta_{age} = 0$ vs. $H_1: \beta_{age} \neq 0$.

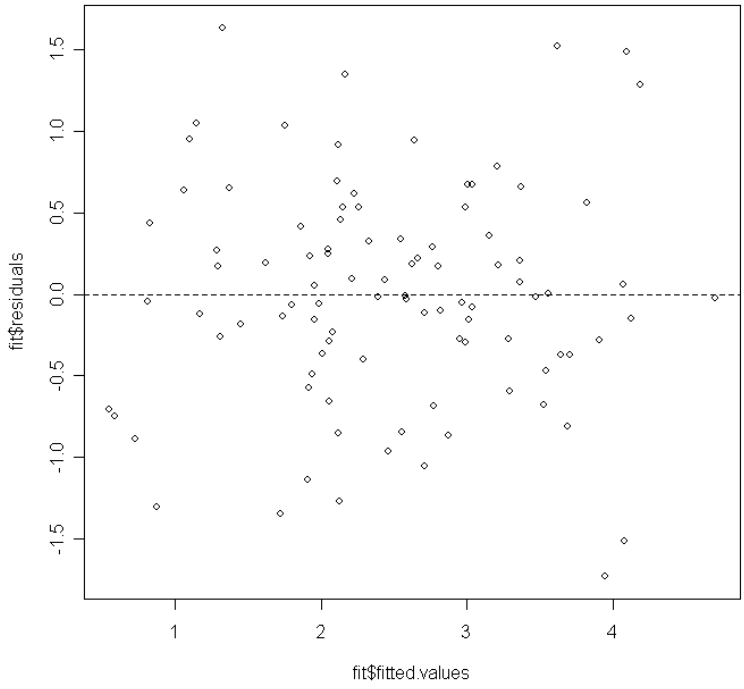
95% CI for β_{age} does cover 0 $\Leftrightarrow$ Do NOT Reject H_0 at $\alpha = 0.05$.

90% CI for β_{age} does NOT cover 0 $\Leftrightarrow$ Reject H_0 at $\alpha = 0.10$.

$\Rightarrow$ **0.05 <** p-value **< 0.10**. (Indeed, p-value = 0.08229.)

b) Plot the residuals vs. the fitted values. Check the constant variance assumption for the errors. (4.3 (a))

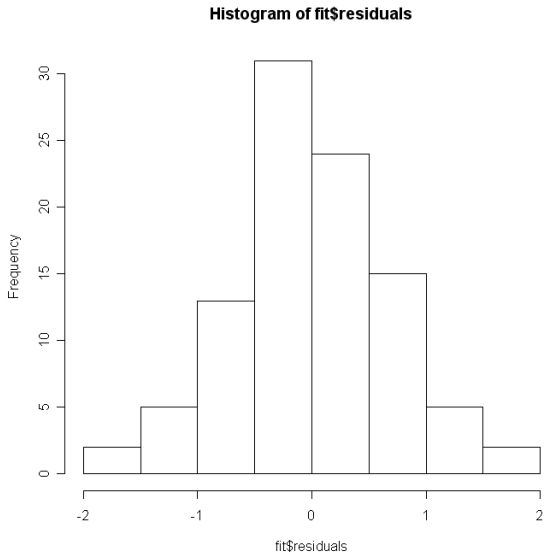
```
> plot(fit$fitted.values,fit$residuals)
> abline(h=0,lty=2)
```



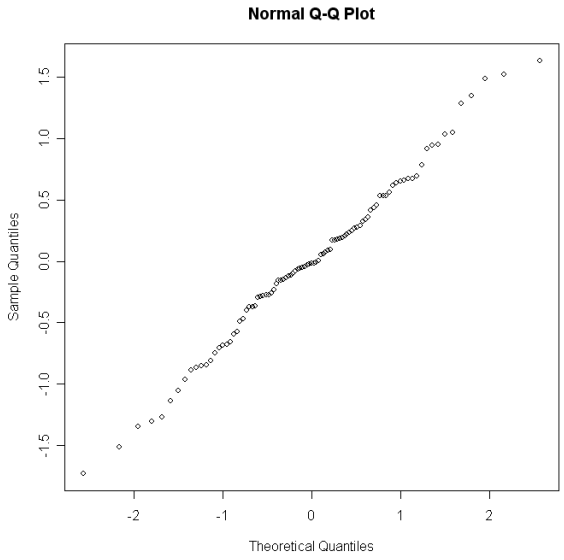
The residuals look quite random. There's no clear evidence for a non-constant variance.

c) Make a histogram and a Normal Q-Q plot for the residuals. Check the normality assumption for the errors. (4.3 (b))

```
> hist(fit$residuals)
```




```
> qqnorm(fit$residuals)
```



There's a little evidence for non-normality, but it's not outstanding. Normality assumption seems to be fine.

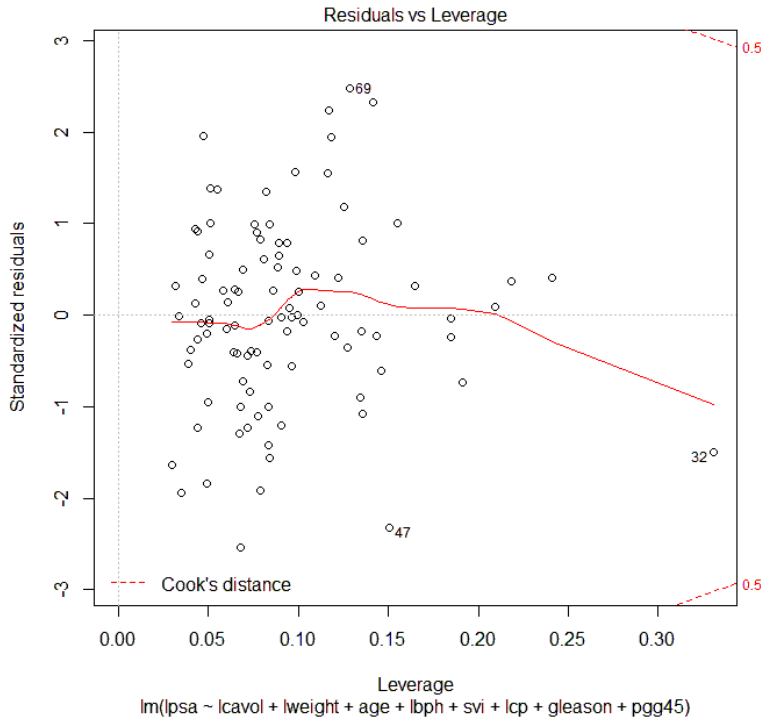
```
> shapiro.test(fit$residuals)

        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.9911, p-value = 0.7721
```

d) Check for large leverage points (that is, identify point(s) with large leverage). (4.3 (c))

```
> plot(fit)
```



Point **32** has large leverage:

```
32  0.1823216  6.1076  65  1.704748  0 -1.38629        6      0  2.00821
```

This is most likely due to the value of `lweight` (6.1076), the largest in the data set. The next (second) largest value of `lweight` is 4.7804.

OR

```
> X = model.matrix(fit)
> H = X %*% solve(t(X)%*%X) %*% t(X)
> lev = rep(0,97)
> for (i in 1:97) { lev[i] = H[i,i] }
```

```
> lev
 [1] 0.07873101 0.06758053 0.13596177 0.07766218 0.03499946 0.08331908
 [7] 0.02989838 0.04944610 0.09401490 0.04023404 0.04386826 0.08925939
[13] 0.04428928 0.07318519 0.05020755 0.06897432 0.06664413 0.08320122
[19] 0.12212111 0.04895576 0.03901634 0.08400872 0.04434074 0.07206303
[25] 0.04582684 0.06594655 0.12048487 0.06479337 0.12707056 0.14633177
[31] 0.05065029 0.33047574 0.09515819 0.04280678 0.05106283 0.06791041
[37] 0.21843920 0.09801067 0.06794996 0.08106758 0.24100789 0.06115256
[43] 0.04674467 0.09036588 0.04262527 0.05037151 0.15065950 0.03401242
[49] 0.13512286 0.05080725 0.09924342 0.06415518 0.09348895 0.07187492
[55] 0.13470108 0.05990394 0.11665631 0.08910835 0.05105674 0.05799578
[61] 0.07677022 0.08328592 0.18468066 0.09024807 0.06930978 0.03186343
[67] 0.10275238 0.06477415 0.12851989 0.10032173 0.07369386 0.08242713
[73] 0.10951482 0.19121086 0.09640539 0.08250756 0.08575379 0.11272985
[79] 0.09614805 0.08839341 0.04703294 0.13546482 0.09985996 0.16486479
[85] 0.05500489 0.07678173 0.08402812 0.07548214 0.14356635 0.12517373
[91] 0.15531867 0.20924207 0.07897648 0.18454695 0.14129097 0.11814056
[97] 0.11689127
>
> sum(lev)
[1] 9
>
> for (i in 1:97) { if (lev[i]>(2*9/97)) { print(i); print(lev[i]) } }
[1] 32
[1] 0.3304757
[1] 37
[1] 0.2184392
[1] 41
[1] 0.2410079
[1] 74
[1] 0.1912109
[1] 92
[1] 0.2092421
```

e) Remove all predictors that are not significant at a 5% level. Test this model against the full model question. Which model is preferred? (3.2 (b))

```
> summary(fit)

Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
    gleason + pgg45)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7331 -0.3713 -0.0170  0.4141  1.6381

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669337   1.296387   0.516  0.60693
lcavol       0.587022   0.087920   6.677 2.11e-09 ***
lweight      0.454467   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766157   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045142   0.157465   0.287  0.77503
pgg45        0.004525   0.004421   1.024  0.30886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7084 on 88 degrees of freedom
Multiple R-Squared: 0.6548,     Adjusted R-squared: 0.6234
F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

Only three predictors are significant at a 5% level – `lcavol`, `lweight`, and `svi`.

```
> fit2 = lm(lpsa~lcavol+lweight+svi)
> summary(fit2)

Call:
lm(formula = lpsa ~ lcavol + lweight + svi)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72964 -0.45764  0.02812  0.46403  1.57013
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.26809    0.54350  -0.493  0.62298
lcavol       0.55164    0.07467   7.388  6.3e-11 ***
lweight      0.50854    0.15017   3.386  0.00104 **
svi          0.66616    0.20978   3.176  0.00203 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7168 on 93 degrees of freedom
Multiple R-Squared: 0.6264,     Adjusted R-squared: 0.6144
F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

Consider $\mathbf{H_0: \beta_{age} = \beta_{lbph} = \beta_{lcp} = \beta_{gleason} = \beta_{pgg45} = 0}$

vs. H_1: at least one of β_{age}, β_{lbph}, β_{lcp}, $\beta_{gleason}$, and β_{pgg45} is not zero.

```
> anova(fit2,fit)
Analysis of Variance Table

Model 1: lpsa ~ lcavol + lweight + svi
Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
    gleason + pgg45
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     93 47.785
2     88 44.163  5     3.622 1.4434 0.2167
```

Since p-value = 0.2167 > 0.10, we **Do NOT Reject $\mathbf{H_0}$** at $\alpha = 0.10$ (or smaller α).

Therefore, the smaller model is simpler and almost as good as the larger model, and we prefer the smaller model:

$$\text{lpsa} = \beta_0 + \beta_{\text{lcavol}} \times \text{lcavol} + \beta_{\text{lweight}} \times \text{lweight} + \beta_{\text{svi}} \times \text{svi} + \varepsilon.$$

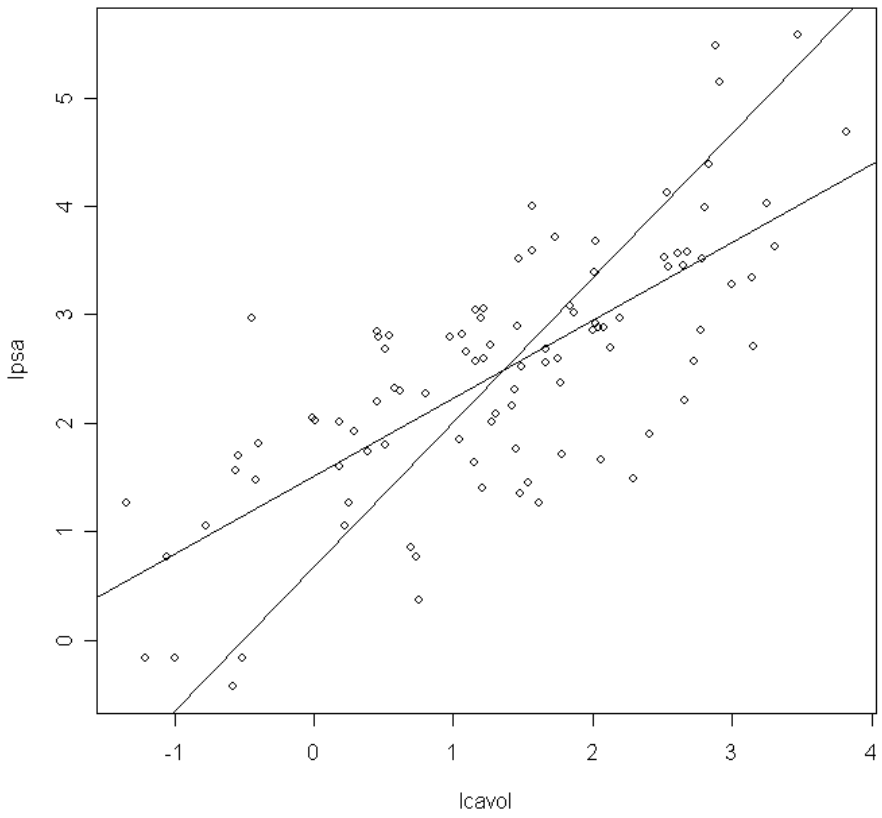
f) Using the `prostate` data, plot `lpsa` against `lcavol`. Fit the regressions of `lpsa` on `lpsa` on `lcavol` and `lcavol` on `lpsa`. Display both regression lines on the plot. At what point do the two lines intersect? (2.5)

Hint 1: If $x = m\,y + b,$ then $y = \frac{1}{m}\,x - \frac{b}{m}.$

Hint 2: `abline(` *y*-intercept `,` slope `)`

If x = `lcavol` and y = `lpsa`:

```
> fit1 = lm(lpsa ~ lcavol)
> fit2 = lm(lcavol ~ lpsa)
> plot(lcavol, lpsa)
> abline(fit1$coeff[1], fit1$coeff[2])
> abline(-fit2$coeff[1]/ fit2$coeff[2], 1/ fit2$coeff[2])
```



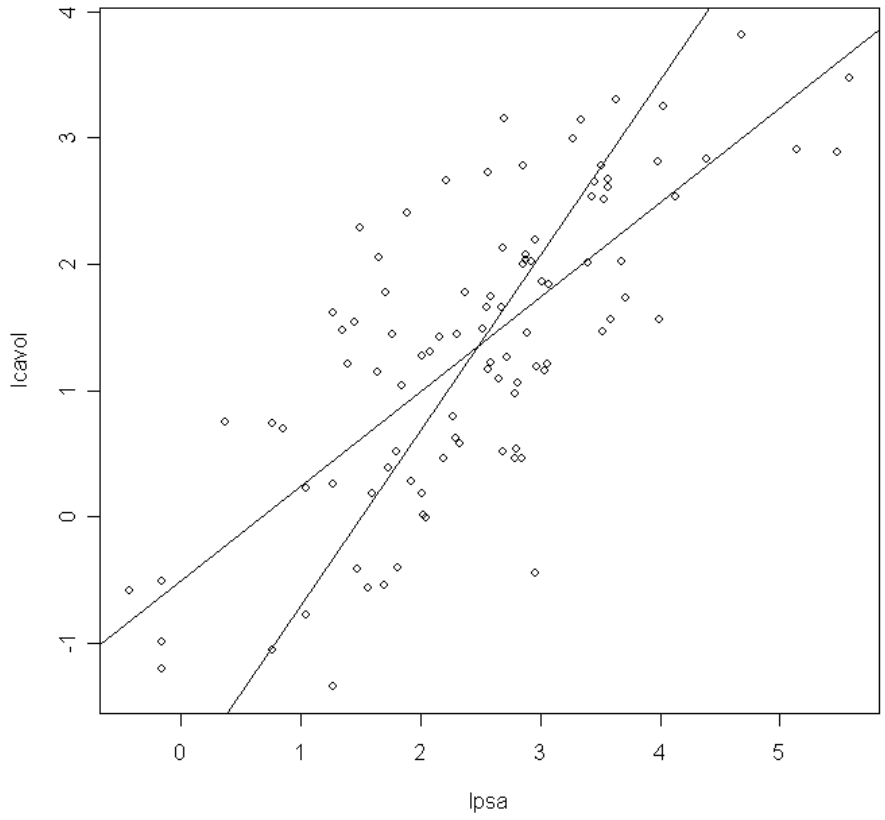
The least-squares regression line always goes through point $(\bar{x}, \bar{y})$.

Therefore, the two regression lines intersect at **the point of averages**.

```
> mean(lcavol)
[1] 1.350010
> mean(lpsa)
[1] 2.478387
```

If x = `lpsa` and y = `lcavol`:

```
> fit1 = lm(lcavol ~ lpsa)
> fit2 = lm(lpsa ~ lcavol)
> plot(lpsa, lcavol)
> abline(fit1$coeff[1], fit1$coeff[2])
> abline(-fit2$coeff[1]/ fit2$coeff[2], 1/ fit2$coeff[2])
```



3. Prove (show) that for simple linear regression model, the leverages are

$$h_{ii} = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{\text{SXX}}, \qquad i = 1, 2, \ldots, n.$$

$$\mathbb{H} = \mathbb{X}\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}.$$

$$\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right) = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}.$$

$$\det\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right) = n \cdot \sum x_i^2 - \left(\sum x_i\right)^2 = n\,\text{SXX}.$$

$$\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1} = \begin{bmatrix} \frac{\sum x_i^2}{n\,\text{SXX}} & -\frac{\sum x_i}{n\,\text{SXX}} \\ -\frac{\sum x_i}{n\,\text{SXX}} & \frac{n}{n\,\text{SXX}} \end{bmatrix} = \begin{bmatrix} \frac{\sum x_i^2}{n\,\text{SXX}} & -\frac{\bar{x}}{\text{SXX}} \\ -\frac{\bar{x}}{\text{SXX}} & \frac{1}{\text{SXX}} \end{bmatrix}.$$

Then

$$\mathbb{H} = \begin{bmatrix} 1 & x_1 \\ \ldots & \ldots \\ 1 & x_i \\ \ldots & \ldots \\ 1 & x_n \end{bmatrix} \cdot \begin{bmatrix} \frac{\sum x_i^2}{n\,\text{SXX}} & -\frac{\bar{x}}{\text{SXX}} \\ -\frac{\bar{x}}{\text{SXX}} & \frac{1}{\text{SXX}} \end{bmatrix} \cdot \begin{bmatrix} 1 & \ldots & 1 & \ldots & 1 \\ x_1 & \ldots & x_i & \ldots & x_n \end{bmatrix}.$$

Therefore,

$$h_i = \mathbb{H}_{ii} = \frac{\sum x_i^2}{n\,\text{SXX}} - 2 \cdot \frac{\bar{x} \cdot x_i}{\text{SXX}} + \frac{x_i^2}{\text{SXX}}$$

$$= \frac{\sum x_i^2}{n\,\text{SXX}} - \frac{n \cdot \bar{x}^2}{n\,\text{SXX}} + \frac{x_i^2}{\text{SXX}} - 2 \cdot \frac{\bar{x} \cdot x_i}{\text{SXX}} + \frac{\bar{x}^2}{\text{SXX}}$$

$$= \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{\text{SXX}}, \qquad \text{since} \quad \text{SXX} = \sum\left(x_i - \bar{x}\right)^2 = \sum x_i^2 - n \cdot \bar{x}^2.$$