STAT 420 Spring 2014

Homework 5: Solutions

Exercise 1

(a) The correlation coefficient between men and women's heights is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{27}{\sqrt{24}\sqrt{40}} = \mathbf{0.8714213}$$

(b) To test $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ at $\alpha = 0.05$, use t test approach (because $\rho = 0$). The T test statistic is given by

$$T = \frac{\sqrt{n-2}r}{\sqrt{1-r^2}} = \frac{\sqrt{4}(0.8714213)}{\sqrt{1-(0.8714213)^2}} = 3.55294$$

which follows a t_4 distribution. The p-value is given by

$$P(|t_4| \geq |T|) = 2P(t_4 > T) = 2P(t_4 > 3.55294) = 2(0.01186794) = \mathbf{0.023736}$$

so we **Reject $\mathbf{H_0}$** at $\alpha = 0.05$.

(c) To test $H_0 : \rho = 0.3$ versus $H_1 : \rho > 0.3$ at $\alpha = 0.05$, use Fisher's r-to-z transformation (because $\rho \neq 0$). First, form the transformed variable

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0.8714213}{1-0.8714213}\right) = \frac{1}{2}\ln(14.55467) = 1.338956$$

Next form the expected value of z under H_0:

$$z_0 = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) = \frac{1}{2}\ln\left(\frac{1.3}{0.7}\right) = \frac{1}{2}\ln(1.857143) = 0.3095196$$

Now standardize z to be approximately $\mathrm{N}(0,1)$:

$$Z^* = \frac{z - z_0}{1/\sqrt{n-3}} = \frac{1.338956 - 0.3095196}{1/\sqrt{3}} = 1.783036$$

Finally, check $\mathrm{N}(0,1)$ CDF to get the p-value:

$$P(Z > Z^*) = P(Z > 1.783036) = \mathbf{0.03729022}$$

so we **Reject $\mathbf{H_0}$** at $\alpha = 0.05$.

(d) To test $H_0 : \rho = 0.5$ versus $H_1 : \rho \neq 0.5$ at $\alpha = 0.05$, use Fisher's r-to-z transformation (because $\rho \neq 0$). First, form the transformed variable (same as before)

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0.8714213}{1-0.8714213}\right) = \frac{1}{2}\ln(14.55467) = 1.338956$$

Next form the expected value of z under H_0 (different):

$$z_0 = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) = \frac{1}{2}\ln\left(\frac{1.5}{0.5}\right) = \frac{1}{2}\ln(3) = 0.5493061$$

Now standardize z to be approximately $\mathrm{N}(0,1)$:

$$Z^* = \frac{z - z_0}{1/\sqrt{n-3}} = \frac{1.338956 - 0.5493061}{1/\sqrt{3}} = 1.367714$$

Finally, check $\mathrm{N}(0,1)$ CDF to get the p-value:

$$P(|Z| > |Z^*|) = 2P(Z > Z^*) = 2P(Z > 1.367714) = 2(0.08570081) = \mathbf{0.1714016}$$

so we **Retain $\mathbf{H_0}$** at $\alpha = 0.05$.

(e) To form a 95% confidence interval for ρ, use Fisher's r-to-z transformation to form the CI on the transformed scale; then convert the CI back to the correlation scale. First, form the transformed variable (same as before)

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0.8714213}{1-0.8714213}\right) = \frac{1}{2}\ln(14.55467) = 1.338956$$

Next, obtain the critical values ($Z_{.975} = 1.959964$), and form the CI on Z scale:

$$1.338956 \pm (1.959964)(1/\sqrt{3}) = [0.2073703;\ 2.470542]$$

Finally, use the inverse Fisher transformation $r = \frac{\exp(2z)-1}{\exp(2z)+1}$ to get the 95% CI back on the original r scale:

$$\left[\frac{\exp\{2(0.2073703)\}-1}{\exp\{2(0.2073703)\}+1};\ \frac{\exp\{2(2.470542)\}-1}{\exp\{2(2.470542)\}+1}\right] = [0.204448;\ 0.9858077]$$

(f) Correlation would be the same, i.e., $r = 0.8714213$. To prove this, define $x_i^* = x_i - 2$ and note that $\bar{x}^* = \frac{1}{n}\sum_{i=1}^n x_i^* = \frac{1}{n}\sum_{i=1}^n (x_i - 2) = \bar{x} - 2$. Next, note that

$$x_i^* - \bar{x}^* = [x_i - 2] - [\bar{x} - 2] = x_i - \bar{x} \qquad \forall i \in \{1, \ldots, n\}$$

which implies that $r^* = \frac{\sum_{i=1}^n (x_i^* - \bar{x}^*)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i^* - \bar{x}^*)^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = r.$

(g) Perfect positive correlation, i.e., $r = 1$. To prove this, define $y_i^* = x_i + 3$, and note that $\bar{y}^* = \frac{1}{n}\sum_{i=1}^n y_i^* = \frac{1}{n}\sum_{i=1}^n (x_i + 3) = \bar{x} + 3$. Next, note that

$$y_i^* - \bar{y}^* = [x_i + 3] - [\bar{x} + 3] = x_i - \bar{x} \qquad \forall i \in \{1, \ldots, n\}$$

which implies that $r^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i^* - \bar{y}^*)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i^* - \bar{y}^*)^2}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = 1.$

Exercise 2

R code to load `Faraway` package, print first row of data, and fit additive regression model with all columns included as predictors. Note that putting a period after the tilde in `lm` includes all variables in `prostate` (except the response variable `lpsa`) as predictors.

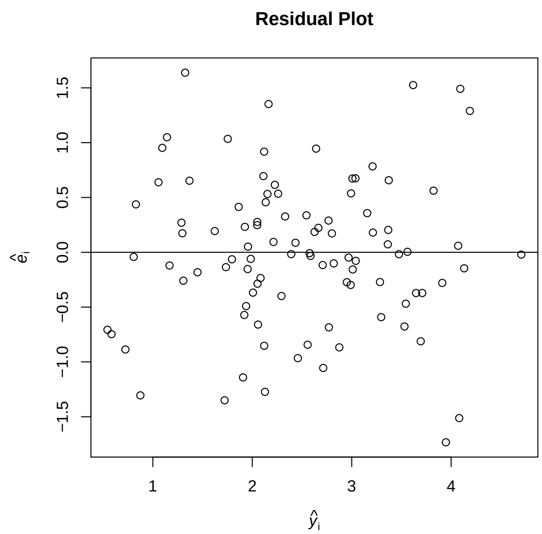
```
> library(faraway)
> prostate[1:3,]
      lcavol lweight age      lbph svi      lcp gleason pgg45      lpsa
1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
> pmod=lm(lpsa~.,data=prostate)
> summary(pmod)$coef
                Estimate  Std. Error    t value     Pr(>|t|)
(Intercept)  0.669336698 1.296387471  0.5163091 6.069335e-01
lcavol       0.587021826 0.087920303  6.6767493 2.110698e-09
lweight      0.454467424 0.170012435  2.6731423 8.955363e-03
age         -0.019637176 0.011172725 -1.7575995 8.229321e-02
lbph         0.107054031 0.058449214  1.8315735 7.039846e-02
svi          0.766157326 0.244309148  3.1360157 2.328749e-03
lcp         -0.105474263 0.091013487 -1.1588861 2.496377e-01
gleason      0.045141598 0.157464523  0.2866779 7.750328e-01
pgg45        0.004525231 0.004421179  1.0235350 3.088604e-01
```

(a) From the R code:

```
> confint(pmod,"age",level=.9)
             5 %         95 %
age -0.0382102 -0.001064151
> confint(pmod,"age",level=.95)
           2.5 %      97.5 %
age -0.04184062 0.002566267
```

We know that `age` is significant at $\alpha = 0.10$ (because 0 is not included in the 90% CI) but is not significant at $\alpha = 0.05$ (because 0 is included in the 95% CI).

(b) Plot of the residuals versus fitted values:

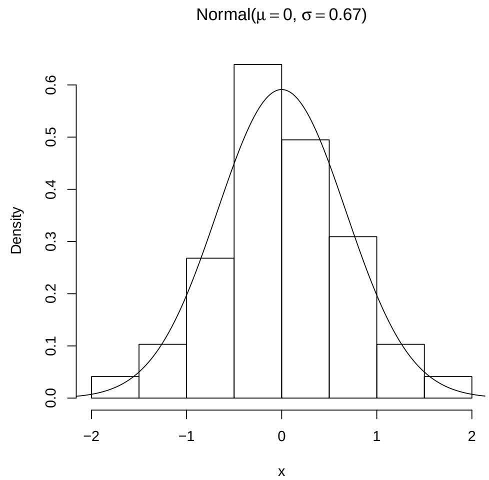


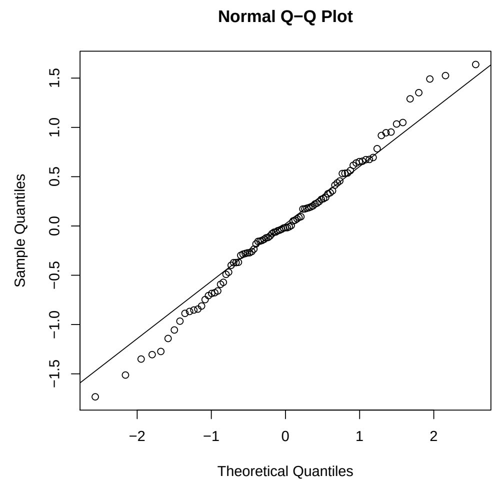
The assumption of constant error variance looks reasonable. To formally test the homogeneity of variance assumption, we could use the Breusch-Pagan test.

```
> BPtest=function(mymod){
+     mymod$model[,1]=(mymod$resid)^2
+     newmod=lm(formula(mymod),data=mymod$model)
+     modsum=summary(newmod)
+     Rsq=modsum$r.squared
+     BPstat=Rsq*(dim(mymod$model)[1])
+     pval=1-pchisq(BPstat,modsum$df[1]-1)
+     list(BP=BPstat,df=modsum$df[1]-1,pval=pval)
+ }
> BPtest(pmod)
$BP
[1] 10.08024

$df
[1] 8

$pval
[1] 0.2594394
```

The Breusch-Pagan test statistic is $\chi^2_{BP} = 10.08024 \sim \chi^2_8$ and has a p-value of $p = 0.2594$, so we retain the null hypothesis of constant error variance.

(c) Histogram using `hnorm` (see Notes 6A) and QQ-plot:





The normality assumption looks reasonable from the plots. To formally test the normality assumption, we could use the Shapiro-Wilk test.

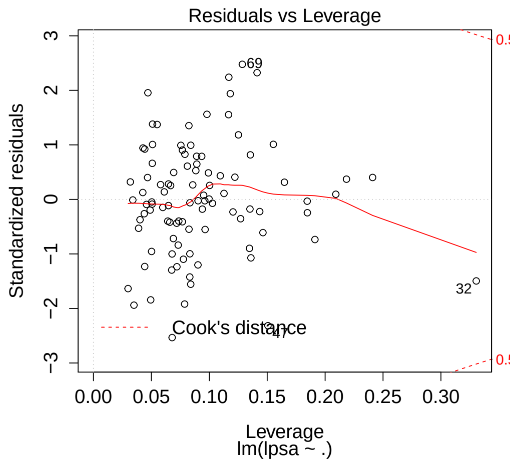
```
> shapiro.test(pmod$resid)

        Shapiro-Wilk normality test

data:  pmod$resid
W = 0.9911, p-value = 0.7721
```

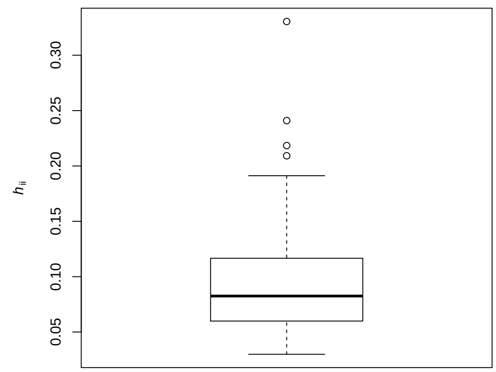
The Shapiro-Wilk test statistic is $W = 0.9911$ and has a p-value of $p = 0.7721$, so we retain the null hypothesis of normality. We could have also used the Looney-Gulledge correlation test (see Notes 6A). Below code uses 10,000 Monte Carlo samples:

```
> set.seed(1234)
> LGnormtest(pmod$resid,nsamp=10000)
$rho
[1] 0.9960252

$cval
       5%
0.9871162

$pval
[1] 0.7618
```

Like Shaprio-Wilk test, Looney-Gulledge test retains the null hypothesis of normality.

(d) Influence plot and box plot of residuals is given below:






```
> pdiag=influence(pmod)
> summary(pdiag$hat)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02990 0.05990 0.08251 0.09278 0.11670 0.33050
> idx=which(pdiag$hat>=(2*9/97))
> pdiag$hat[idx]
       32        37        41        74        92
0.3304757 0.2184392 0.2410079 0.1912109 0.2092421
> prostate[idx,]
      lcavol lweight age      lbph svi      lcp gleason pgg45    lpsa
32 0.1823216  6.1076  65  1.704748   0 -1.38629       6     0 2.00821
37 1.4231083  3.6571  73 -0.579818   0  1.65823       8    15 2.15756
41 0.6205765  3.1420  60 -1.386294   0 -1.38629       9    80 2.29757
74 1.8389611  3.2367  60  0.438255   1  1.17865       9    90 3.07501
92 2.5329028  3.6776  61  1.348073   1 -1.38629       7    15 4.12955
> range(prostate$lweight)
[1] 2.3749 6.1076
> sort(prostate$lweight,decreasing=TRUE)[1:5]
[1] 6.1076 4.7804 4.7181 4.5245 4.4338
```

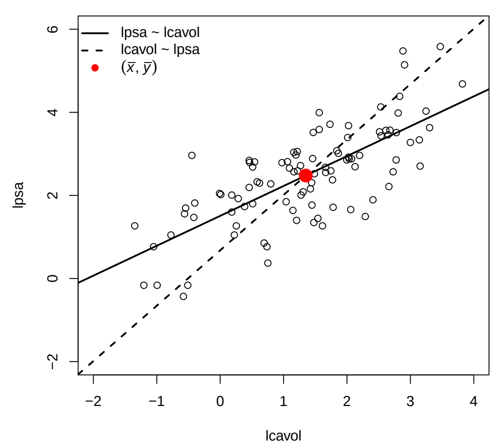
Note that the point with the largest leverage (# 32) has the largest `lweight` of 6.1076, whereas the next largest `lweight` is only 4.7804.

(e) We can fit the reduced model using the `update` function to update our old `lm` model. The dots (to the left and right of tilde) represent our old formula, and we use a minus sign ($-$) to denote which terms to drop.

```
> rmod=update(pmod,.~.-age-lbph-lcp-gleason-pgg45)
> anova(rmod,pmod)
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    3.6218 1.4434 0.2167
```

The F test is not significant, suggesting that the reduced model should be preferred.

(f)
```
plot(prostate$lcavol,prostate$lpsa,xlim=c(-2,4),
     ylim=c(-2,6),xlab="lcavol",ylab="lpsa")
m1mod=lm(lpsa~lcavol,data=prostate)
m2mod=lm(lcavol~lpsa,data=prostate)
abline(m1mod$coef[1],m1mod$coef[2],lty=1,lwd=2)
abline(-m2mod$coef[1]/m2mod$coef[2],1/m2mod$coef[2],lty=2,lwd=2)
points(mean(prostate$lcavol),mean(prostate$lpsa),col="red",cex=2,pch=19)
mybars=expression((bar(italic(x))*", "*bar(italic(y))))
legend("topleft",c("lpsa ~ lcavol","lcavol ~ lpsa",mybars),
       lty=c(1,2,NA),pch=c(NA,NA,19),lwd=c(2,2,NA),
       bty="n",col=c("black","black","red"))
```



Note that the two regression lines pass through the point $(\bar{x}, \bar{y})$.

Exercise 3

Remember that $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, so the diagonals are given by $h_{ii} = \begin{pmatrix} 1 & x_i \end{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} \begin{pmatrix} 1 \\ x_i \end{pmatrix}$.

Now, remember (from Notes 3) that the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ has the form

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

Putting this together and simplifying proves the result

$$\begin{aligned}
h_{ii} &= \begin{pmatrix} 1 & x_i \end{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} 1 & x_i \end{pmatrix} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} 1 & x_i \end{pmatrix} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}x_i \\ x_i - \bar{x} \end{pmatrix} \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \left(\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}x_i + x_i(x_i - \bar{x}) \right) \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \left(\frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x}x_i + x_i^2 \right) \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 + \bar{x}^2 - 2\bar{x}x_i + x_i^2 \right) \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 + (x_i - \bar{x})^2 \right) \\
&= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}
\end{aligned}$$