STAT 420 – Homework 5

1. Dating (without R)

a.

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

b. Method 1 begins by calculating the t-test statistic.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.87142\sqrt{6-2}}{\sqrt{1-0.87142^2}} \approx \mathbf{3.553}.$$

There are $n-2 = 4$ degrees of freedom. According to the t-distribution, the critical region is $|t| > t_{\alpha/2}(n-2) = t_{0.025}(4) = 2.776$. Since the test statistic does lie the critical region, we reject H_0 and conclude that there is a significant correlation between the variables. And since the t-test statistic falls between $t_{0.025}(4) = 2.776$ and $t_{0.01}(4) = 3.747$, the two-sided p-value would be **between 0.02 and 0.05**.

Method 2 begins by calculating the W-test statistic:

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

Under H_0, $\mu_W = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} = \frac{1}{2}\cdot\ln\left(\frac{1+0}{1-0}\right) = 0$, $\sigma_W^2 = \frac{1}{n-3} = \frac{1}{3}$.

We then standardize W under its distribution to create a Z-test statistic:

$$Z = \frac{W-\mu_W}{\sigma_W} = \frac{1.33895-0}{\sqrt{1/3}} \approx \mathbf{2.32}.$$

According to the Z-distribution, the critical region is $|z| > z_{\alpha/2} = z_{0.025} = 1.96$. Since the test statistic does lie the critical region, we reject H_0 and conclude that there is a significant correlation between the variables. The two-sided p-value would be $2 \times P(Z > 2.32) = 2 \times 0.0102 =$ **0.0204**.

c. Begin by calculating the W-test statistic, which is the same as in part b because it's not based on the null hypothesis:

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

Under this H_0, $\mu_W = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} = \frac{1}{2}\cdot\ln\left(\frac{1+0.30}{1-0.30}\right) \approx 0.30952$, $\sigma_W^2 = \frac{1}{n-3} = \frac{1}{3}$.

We then standardize W under its distribution to create a Z-test statistic:

$$Z = \frac{W-\mu_W}{\sigma_W} = \frac{1.33895-0.30952}{\sqrt{1/3}} \approx \mathbf{1.78}.$$

According to the Z-distribution, the critical region is $z > z_\alpha = z_{0.05} = 1.645$. Since the test statistic does lie the critical region, we reject H_0 and conclude that the correlation between the variables is greater than 0.3. The p-value would be $P(Z > 1.78) = \mathbf{0.0375}$.

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?

Begin by calculating the W-test statistic, which is the same as in part b because it's not based on the null hypothesis:

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

Under this H_0, $\mu_W = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} = \frac{1}{2}\cdot\ln\left(\frac{1+0.50}{1-0.50}\right) \approx 0.54931$, $\sigma_W^2 = \frac{1}{n-3} = \frac{1}{3}$.

We then standardize W under its distribution to create a Z-test statistic:

$$Z = \frac{W-\mu_W}{\sigma_W} = \frac{1.33895-0.54931}{\sqrt{1/3}} \approx \mathbf{1.37}.$$

According to the Z-distribution, the critical region is $|z| > z_{\alpha/2} = z_{0.025} = 1.96$. Since the test statistic does not lie the critical region, we fail to reject H_0 and conclude that the correlation between the variables is not significantly different than 0.5. The two-sided p-value would be $2 \times P(Z > 1.37) = 2 \times 0.0853 = \mathbf{0.1706}$.

e. The $100(1-\alpha)\%$ confidence interval for ρ is

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

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

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

Thus, we are 95% confident that the true value of the correlation coefficient is in the interval

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

f. The correlation coefficient is not affected by linear transformations, including adding (or subtracting) the same number to all values of one variable. So, r = **0.87142**.

g. These two variables have a perfect linear relationship of $y = x + 3$. So, r = **1**.

2. Prostate Data (with R)

a. 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
```

b. The 95% confidence interval is

```
> confint(fit, "age", level=0.95)
            2.5 %      97.5 %
age   -0.04184062   0.002566267
```

Since 0 falls in the interval as a plausible value for the coefficient of Age, we would fail to reject the associated null hypothesis meaning that the p-value must be greater than $1 – 0.95 = 0.05$.

The 90% confidence interval is

```
> confint(fit, "age", level=0.90)
              5 %          95 %
age   -0.0382102   -0.001064151
```

Since 0 does not fall in the interval as a plausible value for the coefficient of Age, we would reject the associated null hypothesis meaning that the *p*-value must be less than 1 – 0.90 = 0.10.

Note that the actual *p*-value is 0.08229.

c.

```
> newc = data.frame(lcavol=1.44692, lweight=3.62301,
+ age=65.00000, lbph=0.30010, svi=0.00000, lcp=-0.79851,
+ gleason=7.00000, pgg45=15.00000)

> predict.lm(fit,newc,interval=c("prediction"),level=0.95)
          fit       lwr      upr
[1,] 2.389053 0.9646584 3.813447
```

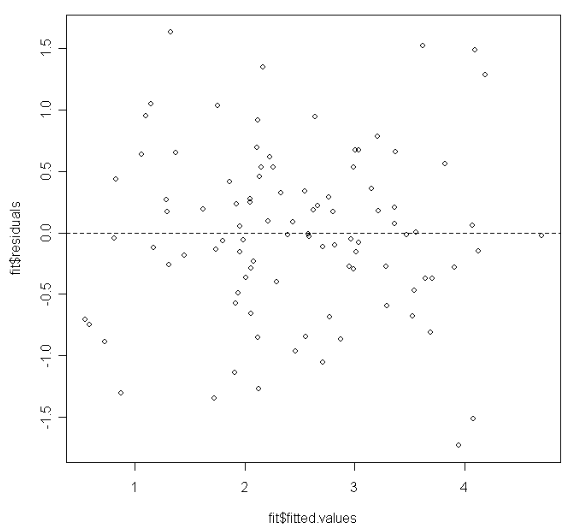
d.

```
> newd = data.frame(lcavol=1.44692, lweight=3.62301,
+ age=20.00000, lbph=0.30010, svi=0.00000, lcp=-0.79851,
+ gleason=7.00000, pgg45=15.00000)

> predict.lm(fit,newd,interval=c("prediction"),level=0.95)
          fit      lwr      upr
[1,] 3.272726 1.538744 5.006707
```

This interval is wider than the one in part c because the value Age = 20 is farther away from the average value of Age (63.87) than the value Age = 65.

e.

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



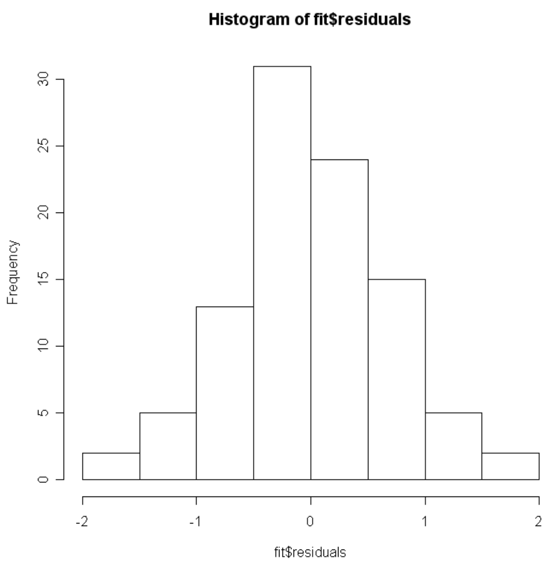
The residuals look quite random. There's no clear evidence for a non-constant variance.

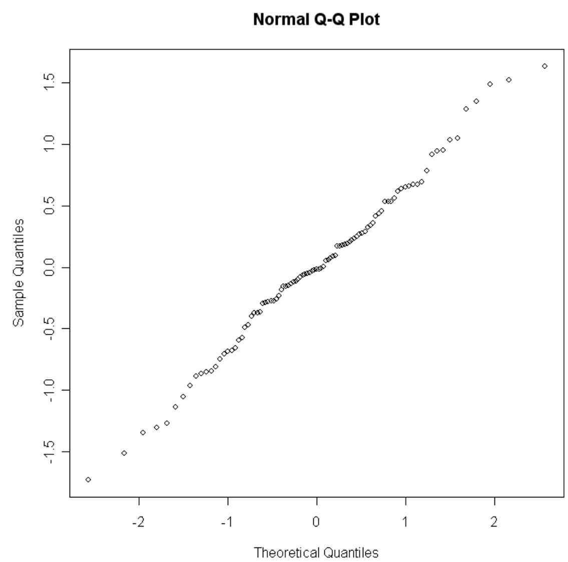
The Breusch-Pagan test confirms that constant variance in the residuals is a reasonable assumption:

```
> bptest(fit)

   studentized Breusch-Pagan test

data:  fit
BP = 10.0802, df = 8, p-value = 0.2594
```

f.

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





There's a little evidence for non-normality, but it's not outstanding. Normality assumption seems to be reasonable as well. The Shapiro-Wilk test supports the visuals.

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

        Shapiro-Wilk normality test

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

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

```
> summary(fit)

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

```
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
```

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

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

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
```

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

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 is rather large, we fail to reject H_0. Therefore, the smaller model (Model 1, H_0) is preferred.