STAT 420 – Homework 7

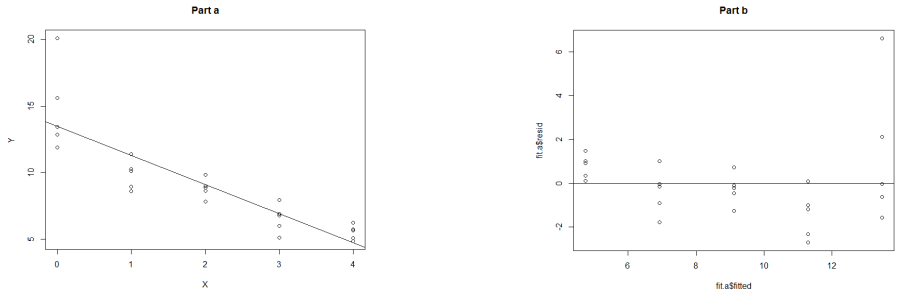
1. Plasma Data

The dataset `plasma` comes from a study on 25 healthy children to check what normal plasma levels of a polyamine in those children should look like. The explanatory variable (X) is the child's age in years, and the response variable (Y) is the plasma level. Note that $x = 0$ represents a newborn child.

a. The points initially seem to be linear, but overlaying the line suggests otherwise. The plasma levels for 1, 2, and 3-year olds generally fall below the line, while plasma levels for 4-year olds all lie above it. Perhaps more of a "bend" in the line with a curvilinear model would be better.

```
> summary(fit.a)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.4752     0.6379  21.126  < 2e-16 ***
X            -2.1820     0.2604  -8.379 1.92e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.841 on 23 degrees of freedom
Multiple R-squared:  0.7532,     Adjusted R-squared:  0.7425
F-statistic: 70.21 on 1 and 23 DF,  p-value: 1.92e-08
```

b. As noted in part a, many of the smaller fitted values have positive residuals, followed by many negative residuals. The variance also seems to increase as the fitted values increase suggesting non-constant variance.

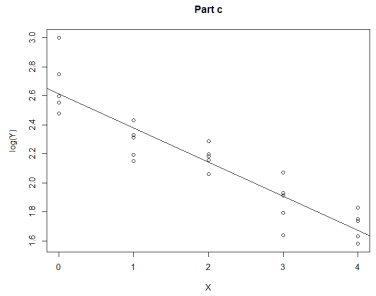


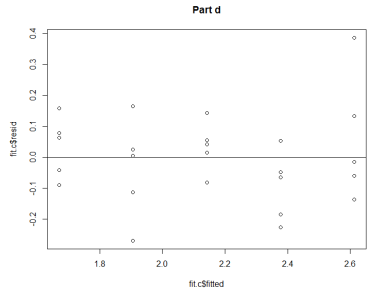
c. The points balance this least-squares regression line much better under the log-transform of the response.

```
> summary(fit.c)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.61302    0.04983   52.44  < 2e-16 ***
X           -0.23552    0.02034  -11.58 4.51e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

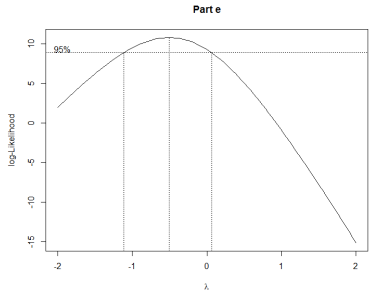
```
Residual standard error: 0.1439 on 23 degrees of freedom
Multiple R-squared:  0.8535,     Adjusted R-squared:  0.8472
F-statistic:   134 on 1 and 23 DF,  p-value: 4.509e-11
```

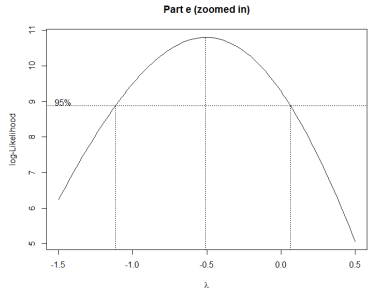
d. The range of residual values is much smaller than in part b. The points are also more randomly scattered about $y = 0$ than the earlier model.





e. The value of $\lambda = 0$ is among the recommended transformations in the Box-Cox plot because 0 falls within the 95% confidence interval.

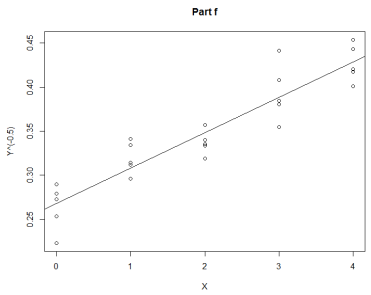


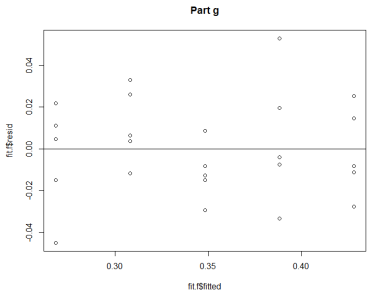


f. The value of λ "most recommended" by the Box-Cox method is –0.5. The model is highly significant and does not differ much from the log-transform model except for one thing. The RSE is significantly smaller.

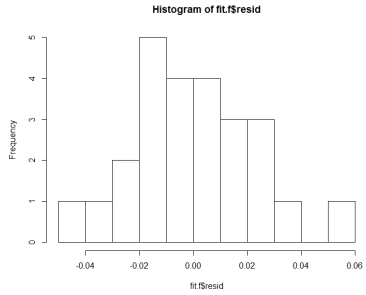
```
> summary(fit.f)

Call:
lm(formula = Y^(-0.5) ~ X)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.268026   0.008033   33.36  < 2e-16 ***
X           0.040062   0.003280   12.22 1.55e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02319 on 23 degrees of freedom
Multiple R-squared:  0.8665,     Adjusted R-squared:  0.8606
F-statistic: 149.2 on 1 and 23 DF,  p-value: 1.548e-11
```

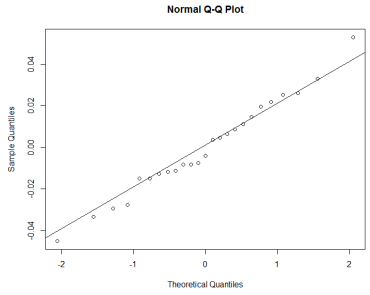
g. The residual plot is a little more random that the one in part d, but only a little bit. Including our response to part f, since there aren't major differences to the Box-Cox model and the previous one, we might just choose the model in part c since the log-transformation is easier to explain.





h. The histogram is mostly symmetric and bell-shaped . (Don't freak out over the tallest peak not being exactly in the middle because its height is actually only one more that those to its right which are centered at 0.) The Normal Q-Q plot looks pretty good, and the Shapiro-Wilk test fails to reject its null hypothesis. Thus, the normality assumption seems to hold very well.






```
> shapiro.test(fit.f$resid)

   Shapiro-Wilk normality test

data:  fit.f$resid
W = 0.9863, p-value = 0.9757
```

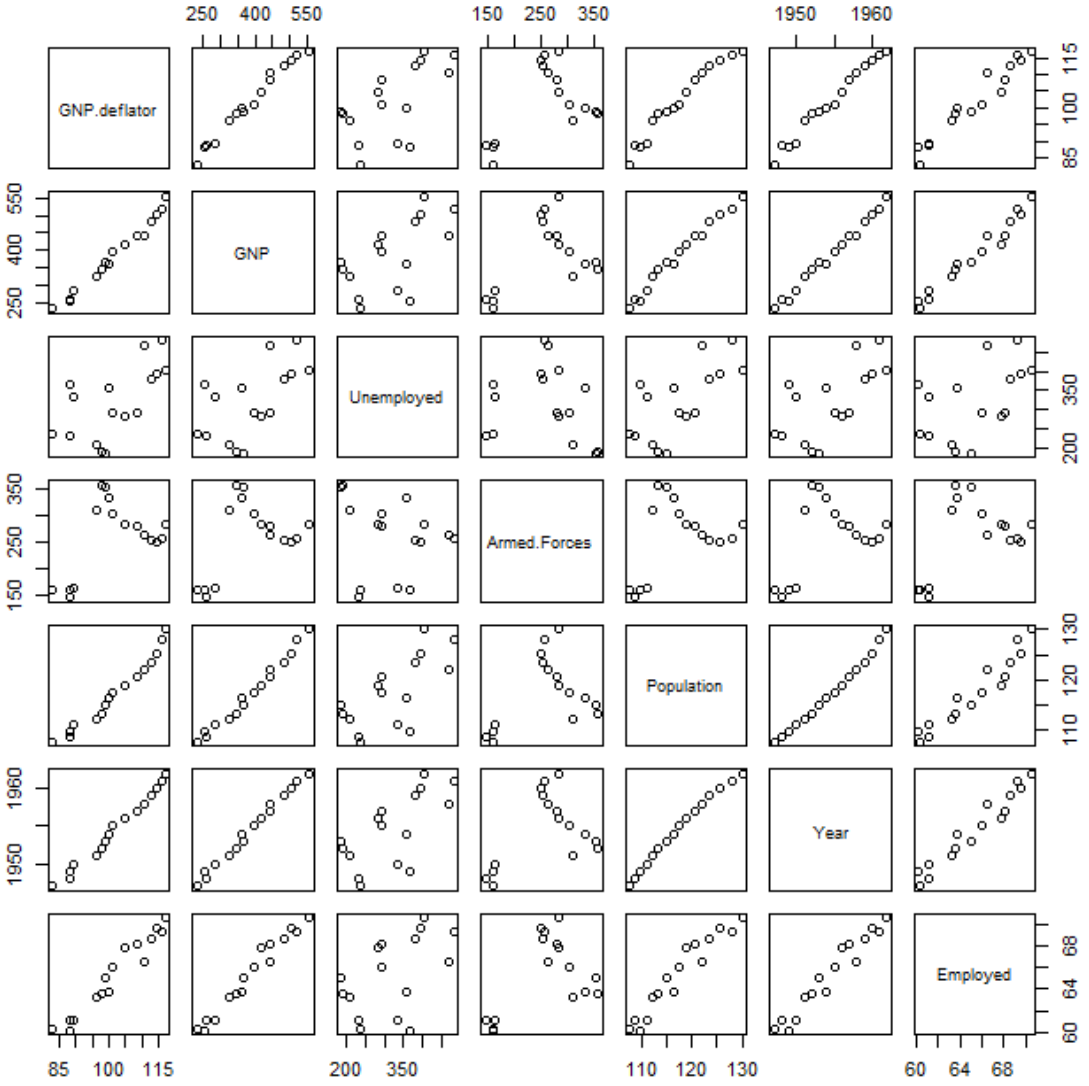
2. Longley Data

a.

```
> round(cor(longley),2)
             GNP.deflator  GNP Unemployed Armed.Forces Population Year Employed
GNP.deflator         1.00 0.99       0.62          0.46       0.98 0.99     0.97
GNP                  0.99 1.00       0.60          0.45       0.99 1.00     0.98
Unemployed           0.62 0.60       1.00         -0.18       0.69 0.67     0.50
Armed.Forces         0.46 0.45      -0.18          1.00       0.36 0.42     0.46
Population           0.98 0.99       0.69          0.36       1.00 0.99     0.96
Year                 0.99 1.00       0.67          0.42       0.99 1.00     0.97
Employed             0.97 0.98       0.50          0.46       0.96 0.97     1.00
```

b.

```
> plot(longley)
```



There seems to be quite a bit of strong correlation among several of the predictors including GNP.deflator, GNP, Population, and Year. This is likely to introduce multicollinearity if all variables are in a model predicting Employed.

c.

```
> fit <- lm(Employed ~ ., data = longley)
> vif(fit)
GNP.deflator       GNP   Unemployed Armed.Forces   Population         Year
      135.53   1788.51        33.62         3.59       399.15       758.98
```

All of the VIF values are extremely large, except for Armed.Forces.

d. The question is essentially asking for the coefficient of determination R^2.

```
> Popn.fit <- lm(Population ~ . - Employed, data = longley)
> summary(Popn.fit)$r.squared
[1] 0.9975
```

99.75% of the variation in Population is explained by a linear relationship with the other predictors.

e.

```
> Empl.fit <- lm(Employed ~ . - Population, data = longley)
> cor(resid(Popn.fit), resid(Empl.fit))
[1] -0.07514
```

Once you remove the effect of all the other predictors on the two variables in question, then the partial correlation coefficient between Population and Employed is only –0.075. This is nearly 0 and way down from the original correlation of 0.96 in part a.

f. Unemployed, Armed.Forces, and Year are the only variables that showed as individually significant in the full model.

```
> summary(fit)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.48e+03   8.90e+02   -3.91  0.00356 **
GNP.deflator  1.51e-02   8.49e-02    0.18  0.86314
GNP          -3.58e-02   3.35e-02   -1.07  0.31268
Unemployed   -2.02e-02   4.88e-03   -4.14  0.00254 **
Armed.Forces -1.03e-02   2.14e-03   -4.82  0.00094 ***
Population   -5.11e-02   2.26e-01   -0.23  0.82621
Year          1.83e+00   4.55e-01    4.02  0.00304 **
---

> fit2 <- lm(Employed ~ Year + Armed.Forces + Unemployed, data = longley)
> summary(fit2)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.80e+03   6.86e+01  -26.18  5.9e-12 ***
Year          9.56e-01   3.55e-02   26.92  4.2e-12 ***
Armed.Forces -7.72e-03   1.84e-03   -4.20   0.0012 **
Unemployed   -1.47e-02   1.67e-03   -8.79  1.4e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.332 on 12 degrees of freedom
Multiple R-squared:  0.993,      Adjusted R-squared:  0.991
F-statistic:  555 on 3 and 12 DF,  p-value: 3.92e-13
```

```
> vif(fit2)
        Year Armed.Forces   Unemployed
       3.891        2.223        3.318
```

The updated model is still highly significant, likely because the three removed variables were suspected for multicollinearity. The results VIF values are all considerably small and under 5, so it does not seem an issue exists any longer.

g. The F-test results in a fairly large *p*-value, leading us to fail to reject the null hypothesis and prefer the smaller model (the one found in part f).

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

Model 1: Employed ~ Year + Armed.Forces + Unemployed
Model 2: Employed ~ GNP.deflator + GNP + Unemployed + Armed.Forces +
Population +
    Year
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     12 1.3234
2      9 0.8364  3    0.4869 1.746  0.227
```