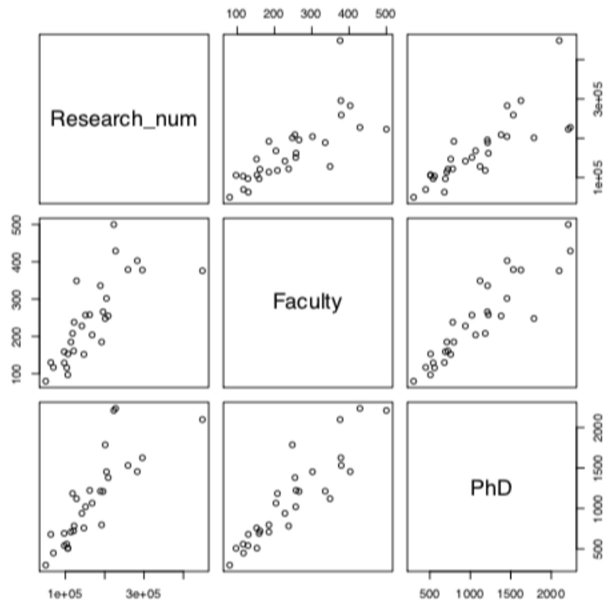
3.13 (Research expenditures data)

a. The correlation matrix and the scatter plot are shown below. We see that PhD has a higher correlation with Research than Faculty, but PhD and Faculty are very highly correlated with each other, so both may not be included in the prediction equation.

```
             Research_num   Faculty       PhD
Research_num    1.0000000 0.7648421 0.8174254
Faculty         0.7648421 1.0000000 0.9036829
PhD             0.8174254 0.9036829 1.0000000
```



b. The regression equation is given below. Only PhD is significant; Faculty is highly nonsignificant. Yet we cannot simply increase Research by increasing the number of PhD students since there must be enough faculty with grants to support those PhD students. It is the Faculty that bring in PhD students. The apparently anamolous result is because only a relatively small fraction of faculty have many grants. Thus increasing this fraction would increase the number of PhD students and increase Research. If faculty don't have grants then they can't bring PhD students.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23525.91   22034.47   1.068   0.2951
Faculty       107.13     193.39   0.554   0.5842
PhD           107.14      40.06   2.675   0.0125 *
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 49040 on 27 degrees of freedom
Multiple R-squared:  0.6719,    Adjusted R-squared:  0.6476
F-statistic: 27.65 on 2 and 27 DF,  p-value: 2.923e-07
```

c. To compute the partial correlations we find the SSE's for three models (full model, only Faculty and only PhD) by computing the corresponding ANOVA tables given below.

```
Analysis of Variance Table

Response: research$Research_num
          Df     Sum Sq    Mean Sq F value    Pr(>F)
Faculty    1 1.1576e+11 1.1576e+11 48.1414 1.862e-07 ***
PhD        1 1.7203e+10 1.7203e+10  7.1539   0.01255 *
Residuals 27 6.4925e+10 2.4046e+09

Analysis of Variance Table

Response: research$Research_num
          Df     Sum Sq    Mean Sq F value    Pr(>F)
Faculty    1 1.1576e+11 1.1576e+11  39.467 8.592e-07 ***
Residuals 28 8.2128e+10 2.9331e+09

Analysis of Variance Table

Response: research$Research_num
          Df     Sum Sq    Mean Sq F value    Pr(>F)
PhD        1 1.3223e+11 1.3223e+11  56.384 3.529e-08 ***
Residuals 28 6.5663e+10 2.3451e+09
```

Denoting Research = y, Faculty = x_1 and PhD = x_2, we have
SSE$(x_1, x_2) = 6.4925e{+}10$, SSE$(x_1) = 8.2128e{+}10$, SSE$(x_2) = 6.5663e{+}10$.
Hence

$$r^2_{yx_1|x_2} = 1 - \frac{6.4925}{6.5663} = 0.0112, r_{yx_2|x_1} = 1 - \frac{6.4925}{8.2128} = 0.2095.$$

Hence $r_{yx_1|x_2} = \sqrt{0.0112} = 0.1060$ and $r_{yx_2|x_1} = \sqrt{0.2095} = 0.4577$.

These partial correlations can be computed using the alternative formula

(3.30):

$$r_{yx_1|x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{(1 - r^2_{yx_2})(1 - r^2_{x_1x_2})}}.$$

Thus

$$r_{yx_1|x_2} = \frac{0.7648 - (0.8174)(0.9037)}{\sqrt{(1 - 0.8174^2)(1 - 0.9037^2)}} = 0.1059$$

and

$$r_{yx_2|x_1} = \frac{0.8174 - (0.7648)(0.9037)}{\sqrt{(1 - 0.7648^2)(1 - 0.9037^2)}} = 0.4577.$$

The t-statistics are given by

$$t_{yx_1|x_2} = \frac{0.1060\sqrt{27}}{\sqrt{1 - 0.1060^2}} = 0.554 \text{ and } t_{yx_2|x_1} = \frac{0.4577\sqrt{27}}{\sqrt{1 - 0.4577^2}} = 2.675.$$

Notice that these t-statistics are the same as the ones given for the corresponding regression coefficients in the regression output.

3.14 (Standardized regression): First calculate

$$\boldsymbol{R}^{-1} = \frac{1}{1 - 0.5^2}\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} = \frac{4}{3}\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}.$$

Hence

$$\widehat{\boldsymbol{\beta}}^* = \begin{bmatrix} \widehat{\beta}_1^* \\ \widehat{\beta}_2^* \end{bmatrix} = \frac{4}{3}\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}\begin{bmatrix} 0.4 \\ 0.8 \end{bmatrix} = \frac{4}{3}\begin{bmatrix} 0 \\ 0.8 \end{bmatrix}.$$

Thus $\widehat{\beta}_1^* = 0$ and $\widehat{\beta}_2^* = (4/3)0.8$. Hence

$$\widehat{\beta}_1 = 0, \widehat{\beta}_2 = \left(\frac{5}{4}\right)\left(\frac{4}{3}\right) 0.8 = \cancel{\frac{4}{3}}, \widehat{\beta}_0 = 10 - \left(\frac{4}{3}\right) 5 = \cancel{\frac{10}{3}}.$$

The formula are correct
but the answer is wrong.
We should have
\hat{\beta}^* = [0; 0.8].
\hat{\beta_1} = 0
\hat{\beta_2} = 1
\hat{\beta_0} = 5

3.15 (Sales data):

a. The correlation matrix between x_1, x_2 and y is as follows. $\boldsymbol{R}$ is the top left hand corner 2×2 submatrix and $\boldsymbol{r}$ is the right hand column subvector of the top two elements.

```
           x1        x2         y
x1 1.0000000 0.9132577 0.9708553
x2 0.9132577 1.0000000 0.9040219
y  0.9708553 0.9040219 1.0000000
```

We compute the partial correlations as follows.

$$r_{yx_1|x_2} = \frac{0.971 - 0.904 \times 0.913}{\sqrt{(1 - 0.904^2)(1 - 0.913^2)}} = \cancel{0.887}$$

0.834

and

$$r_{yx_2|x_1} = \frac{0.904 - 0.971 \times 0.913}{\sqrt{(1 - 0.971^2)(1 - 0.913^2)}} = \cancel{0.756}.$$

0.178

b. Now

$$\boldsymbol{R}^{-1} = \frac{1}{1-0.913^2}\begin{bmatrix} 1 & -0.913 \\ -0.913 & 1 \end{bmatrix} = \begin{bmatrix} 6.009 & -5.486 \\ -5.486 & 6.009 \end{bmatrix}.$$

Then

$$\begin{bmatrix} \widehat{\beta}_1^* \\ \widehat{\beta}_2^* \end{bmatrix} = \boldsymbol{R}^{-1}\boldsymbol{r} = \begin{bmatrix} 6.009 & -5.486 \\ -5.486 & 6.009 \end{bmatrix}\begin{bmatrix} 0.971 \\ 0.904 \end{bmatrix} = \begin{bmatrix} 0.875 \\ 0.105 \end{bmatrix}.$$

Thus $\widehat{\beta}_1^* = 0.875$ and $\widehat{\beta}_2^* = 0.105$. Note that $\widehat{\beta}_1^* > \widehat{\beta}_2^*$ corresponding to $r_{yx_1|x_2} > r_{yx_2|x_1}$.

c. The unstandardized regression equation is

$$\widehat{y} = -2.6062 + 0.1922x_1 + 0.3406x_2.$$

The standard deviations can be computed as:
$s_{x_1} = 6.830, s_{x_2} = 0.461, s_y = 1.501$. Hence

$$\widehat{\beta}_1^* = 0.1922\left(\frac{6.830}{1.501}\right) = 0.875 \quad \text{and} \quad \widehat{\beta}_2^* = 0.3406\left(\frac{0.461}{1.501}\right) = 0.105.$$

d. We see that although $\widehat{\beta}_1 = 0.1922 < \widehat{\beta}_2 = 0.3406$, $\widehat{\beta}_1^* = 0.875 > \widehat{\beta}_2^* = 0.105$. Thus x_2 is a better predictor of y.

3.16 (Salary data):

a. See the following output.

```
The regression equation is
log(Salary) = 4.43 + 0.00748 YrsEm + 0.00168 PriorYr
             + 0.0170 Education+ 0.000390 Super + 0.0231 Female
             - 0.0388 Advertising- 0.0057 Engineering
             - 0.0938 Sales


Predictor         Coef    SE Coef        T      P
Constant       4.42879    0.02134   207.54  0.000
YrsEm         0.007479   0.001193     6.27  0.000
PriorYr       0.001684   0.001957     0.86  0.395
Education     0.017034   0.003336     5.11  0.000
Super        0.0003901  0.0008056     0.48  0.631
Female         0.02307    0.01429     1.61  0.115
Advert        -0.03878    0.02491    -1.56  0.128
Engg          -0.00573    0.01977    -0.29  0.774
Sales         -0.09378    0.02257    -4.15  0.000
```

```
S = 0.0458640   R-Sq = 86.3%   R-Sq(adj) = 83.4%

Analysis of Variance

Source          DF        SS        MS      F      P
Regression       8  0.491757  0.061470  29.22  0.000
Residual Error  37  0.077830  0.002104
Total           45  0.569587
```

b. The coefficient of Male will be -0.0231. The coefficient of Purchase will be +0.0938, the coefficient of Advert will change to $-0.0388 + 0.0938 = 0.055$ and the coefficient of Engg will change to $-0.0573 + 0.0938 = 0.0881$. Also, the constant term will change to $4.429 + 0.0231 - 0.0938 = 4.3583$.

c. The coefficient of Engg is highly nonsignificant when Purchase is used as the reference category because there is not a significant difference between the salaries of employees in Engg and Purchase. But when the reference category is changed to Sales, there is a significant difference between the salaries of employees in Engg and Sales (the coefficient changes from -0.0573 to +0.0881). So if the coefficient of a dummy variable is nonsignificant, all it tells you is that that category is not significantly different from the reference category.

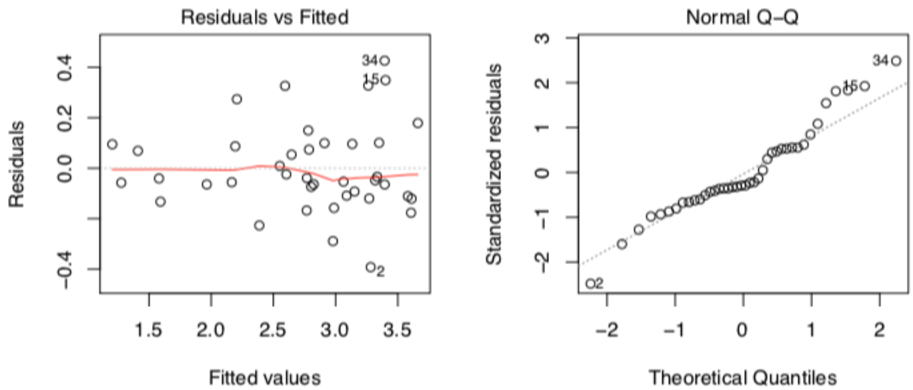
d. The regression equation is

```
log(Salary) = 4.439 + 0.00766 YrsEm + 0.0184 Education
              - 0.0365 Advert - 0.0025 Engineering
              - 0.0876 Sales.
```

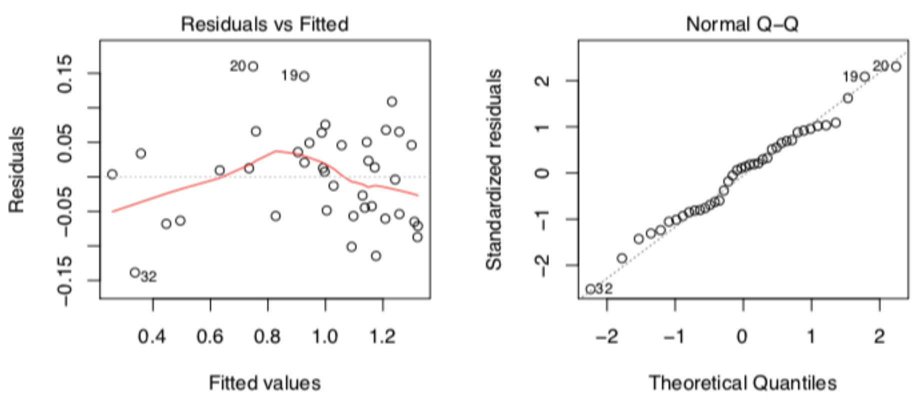
Both Advertising and Engineering turn out to be nonsignificant with P-values 0.157 and 0.901, respectively. We could drop them (i.e., pool them with Purchase) and refit the model.

4.4 (College GPA and entrance test scores: Checking normality and homoscedasticity):

a. The fitted values and the normal Q-Q plots are shown below. The fitted values plot is funnel-shaped indicating heteroscedasticity and the normal plot also shows non-normality. The funnel is expanding linearly suggesting logarithmic transformation.

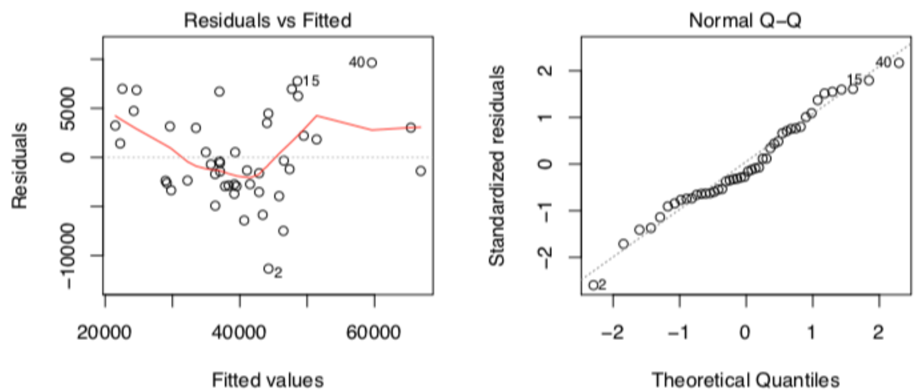


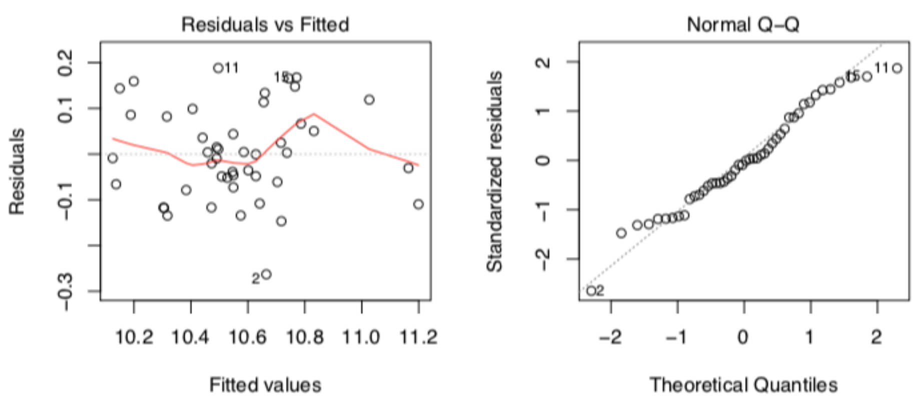
b. The same plots are shown when log(GPA) is used as the response variable. Residuals vs. fitted values plot does not indicate any heteroscedasticity although the plot appears quadratic. This is due to the quadratic terms being not included in the model which we know are significant from Example 3.8. The normal plot is slightly improved. Overall, the logarithmic transformation has a somewhat improved fit.



4.6 (Employee salaries: Checking normality and homoscedasticity):

a. The fitted values and the normal Q-Q plots for both models (using Salary and log(Salary) as response variables) are shown below. The normal plot does not exhibit much improvement for log(Salary) over Salary.





b. However, the fitted values plot shows some improvement (less funnel-shaped and more parallel pattern) indicating using log(Salary) instead of Salary makes the response variable more homoscedastic.