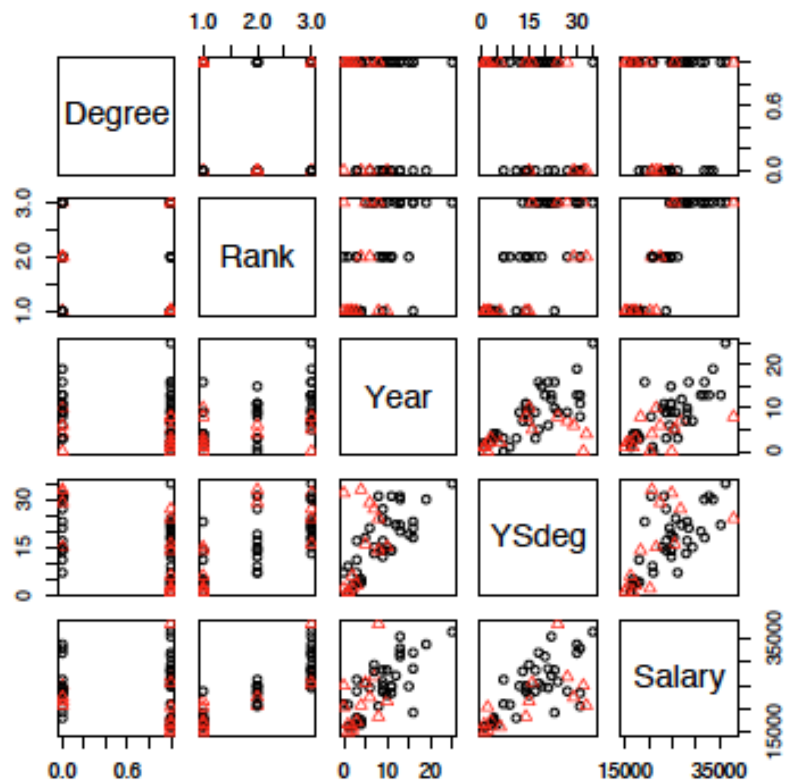
6.13 Sex discrimination The data in the file `salary.txt` concern salary and other characteristics of all faculty in a small Midwestern college collected in the early 1980s for presentation in legal proceedings for which discrimination against women in salary was at issue. All persons in the data hold tenured or tenure track positions; temporary faculty are not included. The data were collected from personnel files, and consist of the quantities described in Table 6.6.

6.13.1. Draw an appropriate graphical summary of the data, and comment of the graph.

Solution:



This scatterplot matrix uses the *Sex* indicator to mark points; females are the red triangles. A scatterplot matrix is less helpful with categorical predictors, and a sequence of plots might have been preferable here. Nevertheless, we see: (1) females are concentrated in the lowest rank; (2) females generally have lower *Years* of service; (3) the mean function for the regression of *Salary* on *YSdeg* will probably have a different slope for males and females. ∎

6.13.2. Test the hypothesis that the mean salary for men and women is the same. What alternative hypothesis do you think is appropriate?

Solution: This is simply a two-sample t-test, which can be computed using regression software by fitting an intercept and a dummy variable for *Sex*:

```
> summary(m0 <- lm(Salary ~ Sex, salary))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    24697        938   26.33   <2e-16
Sex            -3340       1808   -1.85     0.07

Residual standard error: 5780 on 50 degrees of freedom
Multiple R-Squared: 0.0639,
F-statistic: 3.41 on 1 and 50 DF,  p-value: 0.0706
```

The significance level is 0.07 two-sided, and about 0.035 for the one-sided test that women are paid less. The point estimate of the *Sex* effect is \$3340 in favor of men. ■

6.13.3. Obtain a test of the hypothesis that salary adjusted for years in current rank, highest degree, and years since highest degree is the same for each of the three ranks, versus the alternative that the salaries are not the same. Test to see if the sex differential in salary is the same in each rank.

Solution: This problem asks for two hypothesis tests. The first test is ambiguous, and is either asking to test that the main effect of *Rank* is zero, meaning that rank has no effect on (adjusted) salary, or a test that all the *Rank* by other term interactions are zero, meaning that the regressions are parallel. We do both tests:

```
> m1 <- lm(Salary ~ Year +YSdeg + Degree, salary)
> m2 <- update(m1, ~.+ factor(Rank))
> m3 <- update(m2, ~.+ factor(Rank):(Year+YSdeg+Degree))
> anova(m1,m2,m3)
Analysis of Variance Table

Model 1: Salary ~ Year + YSdeg + Degree
Model 2: Salary ~ Year + YSdeg + Degree + factor(Rank)
Model 3: Salary ~ Year + YSdeg + Degree + factor(Rank) + Year:factor(Rank) +
    YSdeg:factor(Rank) + Degree:factor(Rank)
  Res.Df      RSS Df Sum of Sq     F  Pr(>F)
1     48 6.72e+08
2     46 2.68e+08  2  4.04e+08 35.84 1.2e-09
3     40 2.25e+08  6  4.25e+07  1.26     0.3
```

The small p-value for comparing models 1 and 2 suggests that there is indeed a rank effect (as those of us at higher ranks would hope...). The small p-value for comparing model 2 to model 3 suggest that the effects of the other variables are the same in each rank, meaning that the effect of rank is to add an amount to salary for any values of the other terms.

The second test asks specifically about a *Sex* by *Rank* interaction.

```
> m4 <- update(m1, ~.+Sex)
> m5 <- update(m4, ~.+Sex:factor(Rank))
> anova(m1,m4,m5)
Analysis of Variance Table

Model 1: Salary ~ Year + YSdeg + Degree
Model 2: Salary ~ Year + YSdeg + Degree + Sex
Model 3: Salary ~ Year + YSdeg + Degree + Sex + Sex:factor(Rank)
  Res.Df      RSS Df Sum of Sq    F Pr(>F)
1     48 6.72e+08
2     47 6.59e+08  1  1.35e+07 1.07  0.306
3     45 5.65e+08  2  9.36e+07 3.73  0.032
```

These tests should be examined from bottom to top, so we first compare model 2, including a *Sex* effect, to model 3, which includes a *Sex* by *Rank* interaction.

There is some evidence ($p = .032$) that the *Sex* differential depends on rank. The other test of no *Sex* effect is made irrelevant by the significance of the first test: given an interaction, a test for a main effect is not meaningful. Model 2 seems most appropriate, we examine it in a non-standard parameterization.

```
> summary(
lm(formula = Salary ~ -1 + Year + YSdeg + Degree + factor(Rank) +
    Sex:factor(Rank), data = salary))

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
Year                  522.1      105.5    4.95  1.2e-05
YSdeg                -148.6       86.8   -1.71    0.094
Degree              -1501.5     1029.8   -1.46    0.152
factor(Rank)1       17504.7     1285.0   13.62  < 2e-16
factor(Rank)2       22623.7     1580.9   14.31  < 2e-16
factor(Rank)3       28044.0     2103.1   13.33  < 2e-16
factor(Rank)1:Sex     444.3     1153.5    0.39    0.702
factor(Rank)2:Sex     942.6     2194.9    0.43    0.670
factor(Rank)3:Sex    2954.5     1609.3    1.84    0.073

Residual standard error: 2400 on 43 degrees of freedom
```

The coefficients for the three *Rank* terms correspond to intercept for the three ranks for males. The *Rank* by *Sex* terms give the *Sex* differentials in each of the three ranks; in each rank the differential for females is *positive*, although relatively small, meaning that adjusting for *Rank*, *Year*, *Degree* and *YSdeg*, the women are better paid than the men by a small amount. ∎

6.13.4. Finkelstein (1980), in a discussion of the use of regression in discrimination cases, wrote, "... [a] variable may reflect a position or status bestowed by the employer, in which case if there is discrimination in the award of the position or status, the variable may be 'tainted'." Thus, for example, if discrimination is at work in promotion of faculty to higher ranks, using rank to adjust salaries before comparing the sexes may not be acceptable to the courts.

Fit two mean functions, one including *Sex*, *Year*, *YSdeg* and *Degree*, and the second adding *Rank*. Summarize and compare the results of leaving out rank effects on inferences concerning differential in pay by sex.

Solution:

```
> summary(m7 <- update(m3, ~Sex+Year+YSdeg+Degree))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13884.2     1639.8    8.47  5.2e-11
Sex          -1286.5     1313.1   -0.98  0.33221
Year           352.0      142.5    2.47  0.01719
YSdeg          339.4       80.6    4.21  0.00011
Degree        3299.3     1302.5    2.53  0.01470

Residual standard error: 3740 on 47 degrees of freedom
Multiple R-Squared: 0.631,
F-statistic: 20.1 on 4 and 47 DF,  p-value: 1.05e-09
```

If we ignore *Rank*, then the coefficient for *Sex* is again negative, indicating an advantage for males, but the p-value is .33 (or .165 for a one-sided test), indicating that the difference is not significant.

One could argue that other variables in this data set are tainted as well, so using data like these to resolve issues of discrimination will never satisfy everyone. ∎

6.14 Using the salary data in Problem 6.13, one fitted mean function is:

$$\mathrm{E}(Salary|Sex, Year) = 18223 - 571 Sex + 741\, Year + 169 Sex \times Year$$

6.14.1. Give the coefficients in the estimated mean function if *Sex* were coded so males had the value 2 and females had the value 1 (the coding given in the data file is 0 for males and 1 for females).

Solution: Changing the coding for the *Sex* indicator will change only the coefficient for *Sex* and the coefficient for the intercept. Suppose $\hat{\beta}_0$ and $\hat{\beta}_1$ are the intercept and estimate for *Sex* in the original parameterization, and let $\hat{\eta}_0$ and $\hat{\eta}_1$ be the corresponding estimates in the new coding for *Sex*. Then we must have:

$$\begin{aligned}\text{For males: } \hat\beta_0 + \hat\beta_1 \times 0 &= \hat\eta_0 + \hat\eta_1 \times 2 \\ \text{For females: } \hat\beta_0 + \hat\beta_1 \times 1 &= \hat\eta_0 + \hat\eta_1 \times 1\end{aligned}$$

Substituting for $\hat\beta_0$ and $\hat\beta_1$,

$$\begin{aligned}18223 &= \hat\eta_0 + 2\hat\eta_1 \\ \cancel{18823} - 571 &= \hat\eta_0 + \hat\eta_1\end{aligned}$$

These two equations in two unknowns are easily solved to give $\hat\eta_0 = 17681$, and $\hat\eta_1 = +571$. ∎

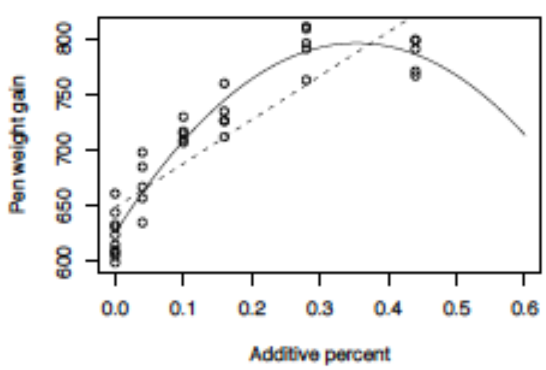
6.14.2. Give the coefficients if *Sex* is coded as -1 for males and $+1$ for females.

Solution: The intercept will change to ~~$18223 + 571/2 = 18508.5$~~. The *Sex* coefficient will become $-571/2 = -285.5$. ∎

6.15 Pens of turkeys were grown with an identical diet, except that each pen was supplemented with an amount A of an amino acid methionine as a percentage of the total diet of the birds. The data in the file `turk0.txt` gives the response average weight *Gain* in grams of all the turkeys in the pen for 35 pens of turkeys receiving various levels of A.

6.15.1. Draw the scatterplot of *Gain* versus A and summarize. In particular, does simple linear regression appear plausible?

Solution:



For larger values of A, the response appears to level off, or possibly decrease. Variability appears constant across the plot. The lines on the plot refer to Problem 6.15.3. ∎

6.15.2. Obtain a lack of fit test for the simple linear regression mean function, and summarize results. Repeat for the quadratic regression mean function.

Solution:

```
Response: Gain
            Df Sum Sq Mean Sq F value  Pr(>F)
A            1 124689  124689   368.1 < 2e-16
Lack of fit  4  25353    6338    18.7 1.1e-07
Pure error  29   9824     339
```

Quadratic mean function:

```
Response: Gain
            Df Sum Sq Mean Sq F value Pr(>F)
A            1 124689  124689  368.09 <2e-16
I(A^2)       1  23836   23836   70.37  3e-09
Lack of fit  3   1516     505    1.49   0.24
Pure error  29   9824     339
```

There is lack of fit for the simple linear regression model, but the quadratic model is adequate. ∎

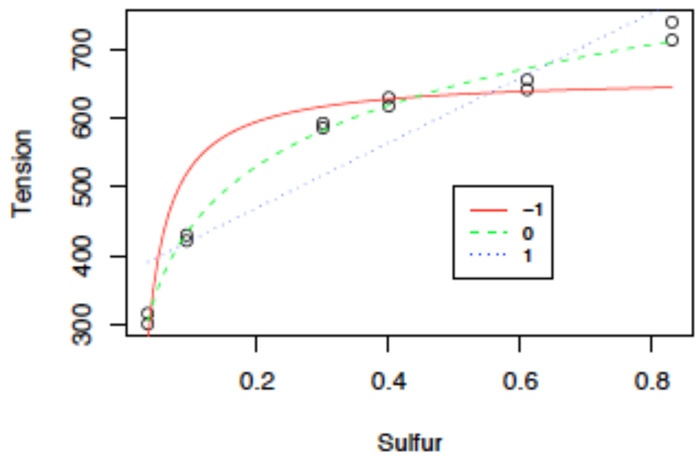
6.15.3. To the graph drawn in Problem 6.15.1 add the fitted mean functions based on both the simple linear regression mean function and the quadratic mean function, for values of A in the range from 0 to 0.60, and comment.

Solution: The straight line mean function does not match the data, and leads to the unlikely results that (1) *Gain* could be increased indefinitely as A is increased, and (2) the rate of increase is constant. The quadratic mean function is reasonable for the range of A observed in the data, but it implies that *Gain* actually decreases for $A > .4$ or so. This is probably also quite unrealistic. The conclusion is that the polynomial model is useful for interpolation here, but certainly not for extrapolation outside the range of the data. ∎

7.1 The data in the file `baeskel.txt` were collected in a study of the effect of dissolved sulfur on the surface tension of liquid copper (Baes and Kellogg, 1953). The predictor *Sulfur* is the weight percent sulfur, and the response is *Tension*, the decrease in surface tension in dynes per cm. Two replicate observations were taken at each value of *Sulfur*. These data were previously discussed by Sclove (1972).

7.1.1. Draw the plot of *Tension* versus *Sulfur* to verify that a transformation is required to achieve a straight-line mean function.

Solution:



∎

7.1.2. Set $\lambda = -1$, and fit the mean function

$$\mathrm{E}(\mathit{Tension}|\mathit{Sulfur}) = \beta_0 + \beta_1 \mathit{Sulfur}^{\lambda}$$

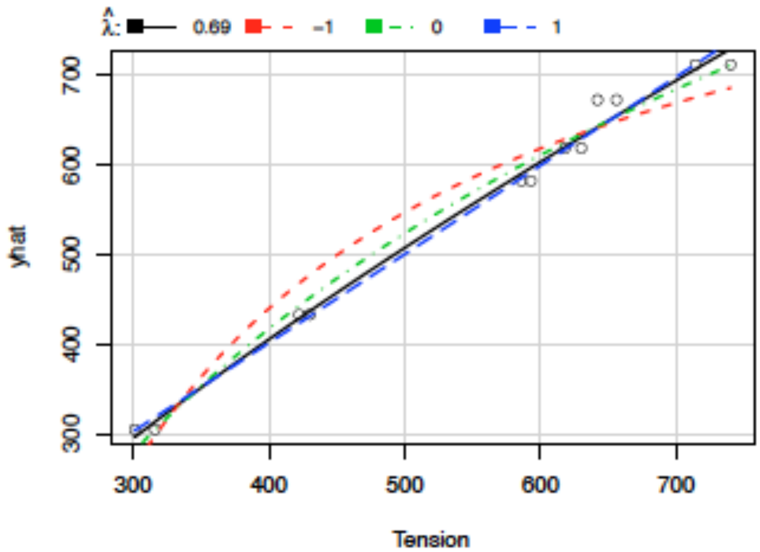
using OLS; that is, fit the OLS regression with *Tension* as the response and $1/\mathit{Sulfur}$ as the predictor. Let *new* be a vector of 100 equally spaced values between the minimum value of *Sulfur* and its maximum value. Compute the fitted values from the regression you just fit, given by $\mathit{Fit.new} = \beta_0 + \beta_1 \mathit{new}^{\lambda}$. Then, add to the graph you drew in Problem 7.1.1 the line joining the points $(\mathit{new}, \mathit{Fit.new})$. Repeat for $\lambda = 0, 1$. Which of these three choices of λ gives fitted values that match the data most closely?

Solution: From the above figure, only the log transformation closely matches the data. ∎

7.1.3. Replace *Sulfur* by its logarithm, and consider transforming the response *Tension*. To do this, draw the inverse response plot with the fitted values from the regression of *Tension* on $\log(\mathit{Sulphur})$ on the vertical axis and *Tension* on the horizontal axis. Repeat the methodology of Problem 7.1.2 to decide if further transformation of the response will be helpful.

Solution: As pointed out in the text, with a single predictor the inverse response plot is equivalent to a plot of the response on the horizontal axis and the predictor on the vertical axis. The plot can be drawn most easily with the `invResPlot` function

```
> invResPlot(lm(Tension ~ log(Sulfur), baeskel))
      lambda       RSS
1  0.6860853  2202.113
2 -1.0000000 10594.340
3  0.0000000  3658.171
4  1.0000000  2509.564
```



Untransformed, $\lambda = 1$, matches well, almost as well as the optimal valie of about $2/3$, suggesting no further need to transform. This could be verified by performing a lack of fit test from the regression of *Tension* on $\log(\mathit{Sulfur})$,

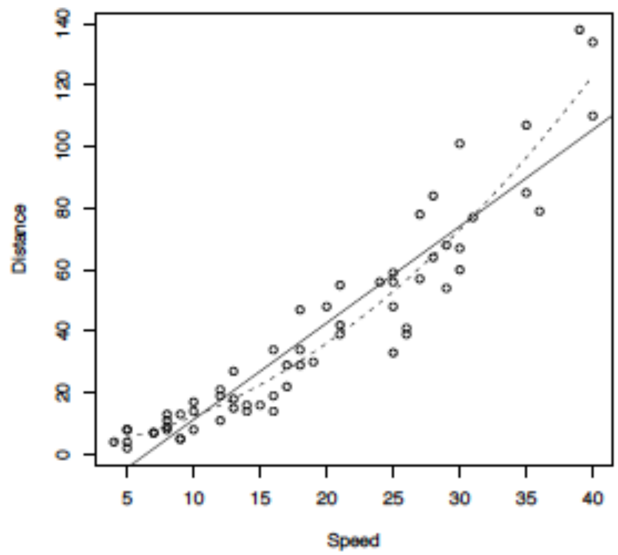
```
> m1 <- lm(Tension~log(Sulfur)+factor(Sulfur))
> anova(m1)
Analysis of Variance Table

Response: Tension
               Df Sum Sq Mean Sq F value  Pr(>F)
log(Sulfur)     1 241678  241678 2141.90 6.8e-09
factor(Sulfur)  4   1859     465    4.12   0.061
Residuals       6    677     113
```

The lack-of-fit test has p-value of 0.06. ∎

7.2 The (hypothetical) data in the file `stopping.txt` give stopping times for $n = 62$ trials of various automobiles traveling at *Speed* miles per hour and the resulting stopping *Distance* in feet (Ezekiel and Fox, 1959).

7.2.1. Draw the scatterplot of *Distance* versus *Speed*. Add the simple regression mean function to your plot. What problems are apparent? Compute a test for lack of fit, and summarize results.

Solution:



The solid line is for simple regression, and the dashed line is a quadratic fit. A lack of fit test can be done using a pure error analysis, since there are replications, or by comparing the quadratic mean function to the simple linear regression mean function.

```
> m1<-lm(Distance~Speed,stopping)
> m2 <- lm(Distance~Speed+I(Speed^2), data=stopping)
> pureErrorAnova(m1)
Analysis of Variance Table

Response: Distance
            Df Sum Sq Mean Sq F value Pr(>F)
Speed        1  59639   59639  625.95 <2e-16
Lack.of.Fit 26   5071     195    2.05  0.025
Residuals   34   3239      95
> anova(m2)
Analysis of Variance Table

Response: Distance
           Df Sum Sq Mean Sq F value  Pr(>F)
Speed       1  59639   59639   605.2 < 2e-16
I(Speed^2)  1   2496    2496    25.3 4.8e-06
Residuals  59   5814      99
```

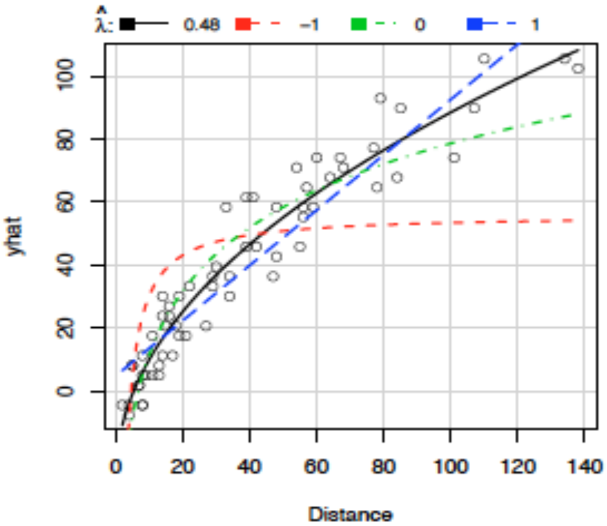
Both methods indicate that the simple regression mean function is not adequate.

7.2.2. Find an appropriate transformation for *Distance* that can linearize this regression.

Solution: Using the inverse response plot method:

```
> invResPlot(m1)  # suggests square root of Distance
```

```
      lambda       RSS
1  0.4849737  4463.944
2 -1.0000000 33149.061
3  0.0000000  7890.434
4  1.0000000  7293.835
```

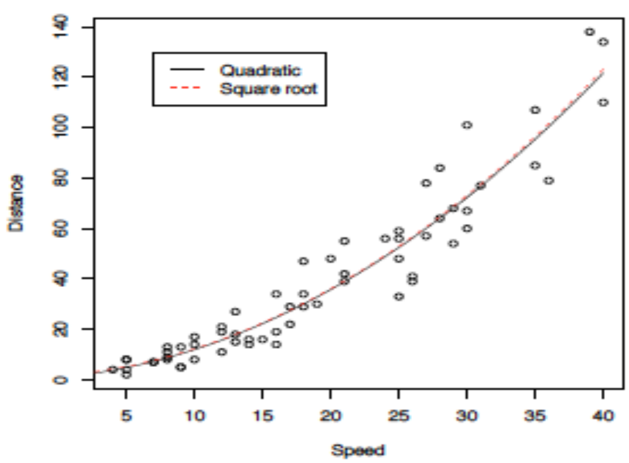


The optimal transformation is at about $\hat{\lambda} = .49$

This suggests using the square root scale for *Distance*.∎

7.2.3. Hald (1960) has suggested on the basis of a theoretical argument that the mean function $\mathrm{E}(Distance|Speed) = \beta_0 + \beta_1 Speed + \beta_2 Speed^2$, with $\mathrm{Var}(Distance|Speed) = \sigma^2 Speed^2$ is appropriate for data of this type. Compare the fit of this model to the model found in Problem 7.2.2. For *Speed* in the range 0 to 40 mph, draw the curves that give the predicted *Distance* from each model, and qualitatively compare them.

Solution:



The plot of fitted values from the weighted quadratic model and the squares of the fitted values of the unweighted analysis in square root scales are virtually identical. ∎