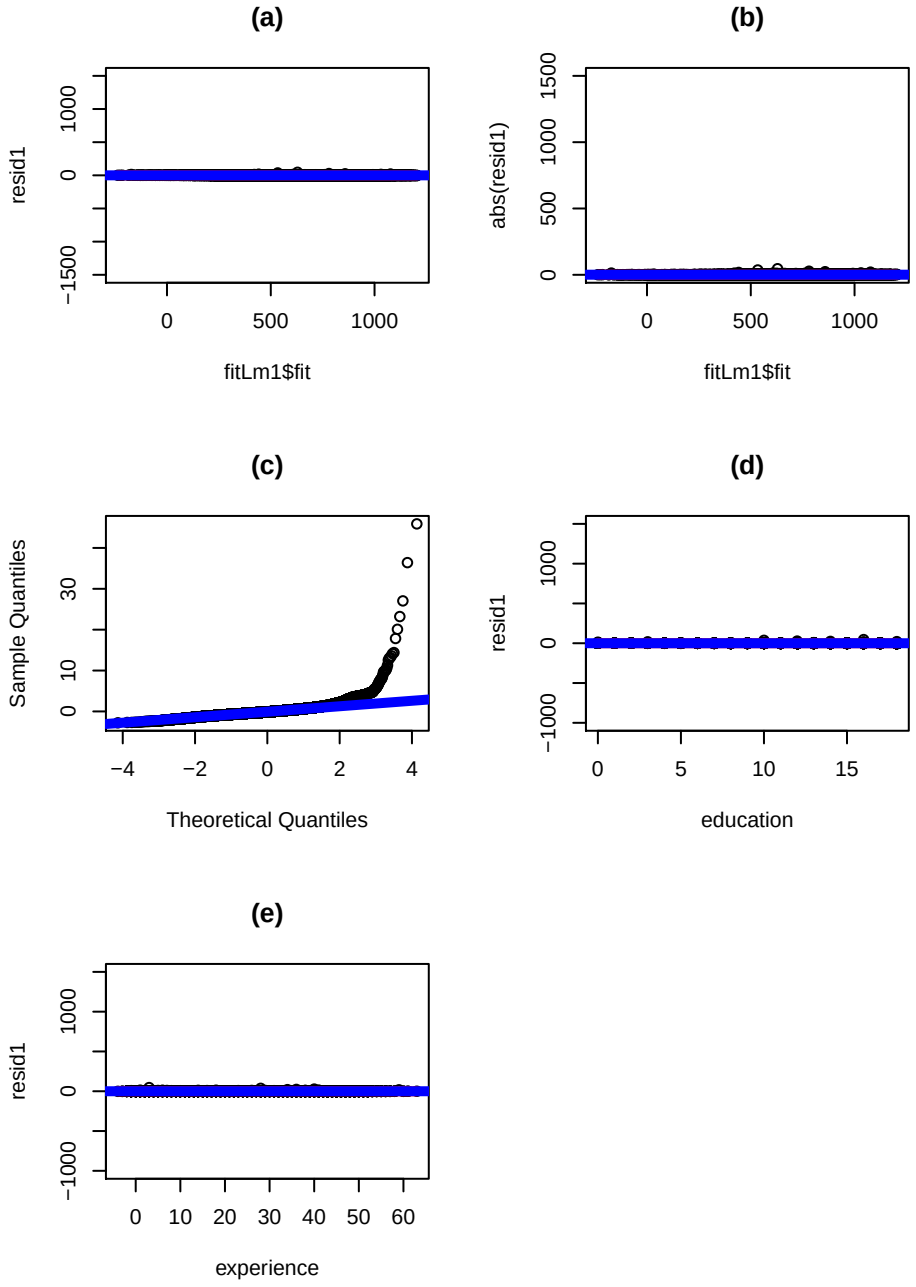
Solutions to Selected Computer Lab Problems and Exercises in Chapter 10 of *Statistics and Data Analysis for Financial Engineering, 2nd ed.* by David Ruppert and David S. Matteson



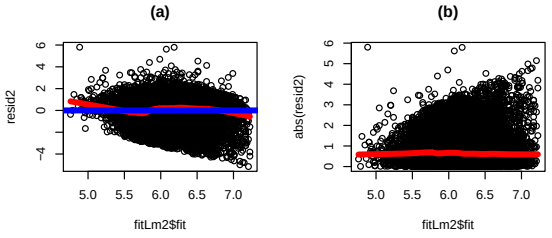
Problem 1. Panel (a) is a plot of the residuals versus the fitted values. We can see substantial heteroskedasticity since the scatter in the residuals increases as the fitted values increase. Comparing the red lowess smooth with the blue horizontal reference

line through 0, we see that there is also evidence of bias since all the residuals are positive at the smallest fitted values and the lowess curve is negative at the largest fitted values. The heteroskedasticity suggests that `wage` should be transformed, perhaps with the log transformation. Whether a transformation will fix the bias remains to be seen.

Panel (b) is a plot of the absolute residuals versus the fitted values. The lowess smooth has a nonmonotonic shape being higher at the ends than in the middle. What we see here is the combination of two effects already seen in panel (a), the bias which causes the residuals to be large and positive at small fitted values and the heteroskedasticity which causes the absolute residuals to be large at large fitted values. Possibly remedies to these problems were discussed in the previous paragraph.

Panel (c) is a normal QQ plot and shows severe right skewness. This could be alleviated by a log transformation used to alleviate heteroskedasticity.

Panels (d) and (e) are plots of the residuals versus the two predictors. There is evidence of nonlinear effects, especially for `experience` because of the differences between the lowess smooths and the horizontal reference lines.



Problem 2. The log transformation has alleviated several problems. The scatterplots in panels (a) and (b) suggest that there is still some heteroskedasticity, but a more careful analysis shows that this is an artifact of changing data densities.

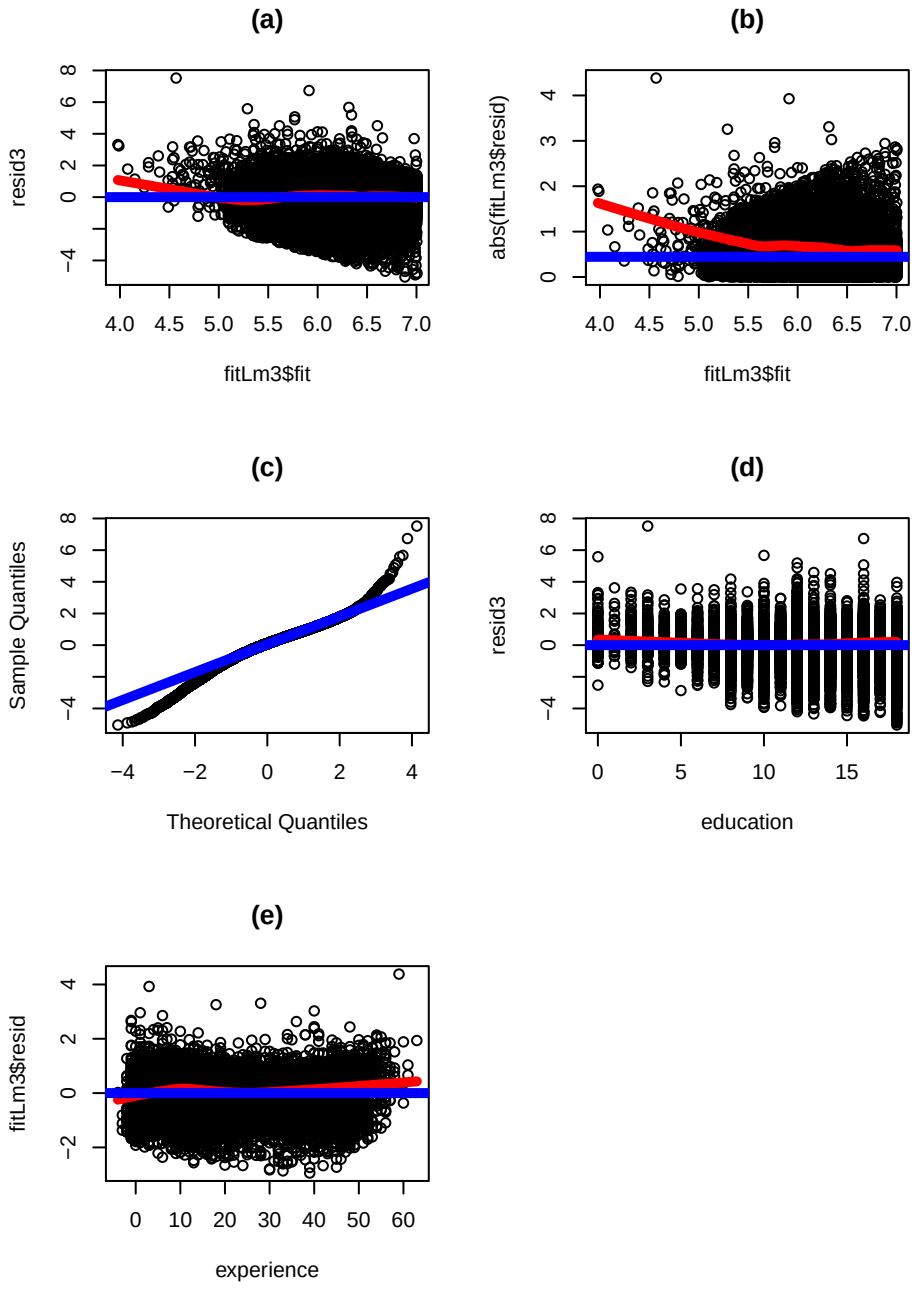
There are much more data on the right sides of these plots compared to the left sides. The lowess smooth in panel (b) is horizontal, which shows that the average absolute residual is constant so, in fact, there is no heteroskedasticity.

The normal QQ plot in panel (c) shows little skewness but the tails are somewhat heavier than for a normal distribution. However, this problem does not seem serious and can probably be ignored.

Panel (d) shows that the effect of `education` on the log of `wage` is nearly linear, but panel (e) shows that the effect of `experience` remains nonlinear.

The next thing to try would be using a nonlinear effect for `experience`, perhaps a quadratic polynomial.

Problem 3. Fitting a quadratic polynomial in `experience` reduces bias, but the plot of residuals versus fitted values still shows that all residuals are positive at small fitted values. This can be seen in the plot below.

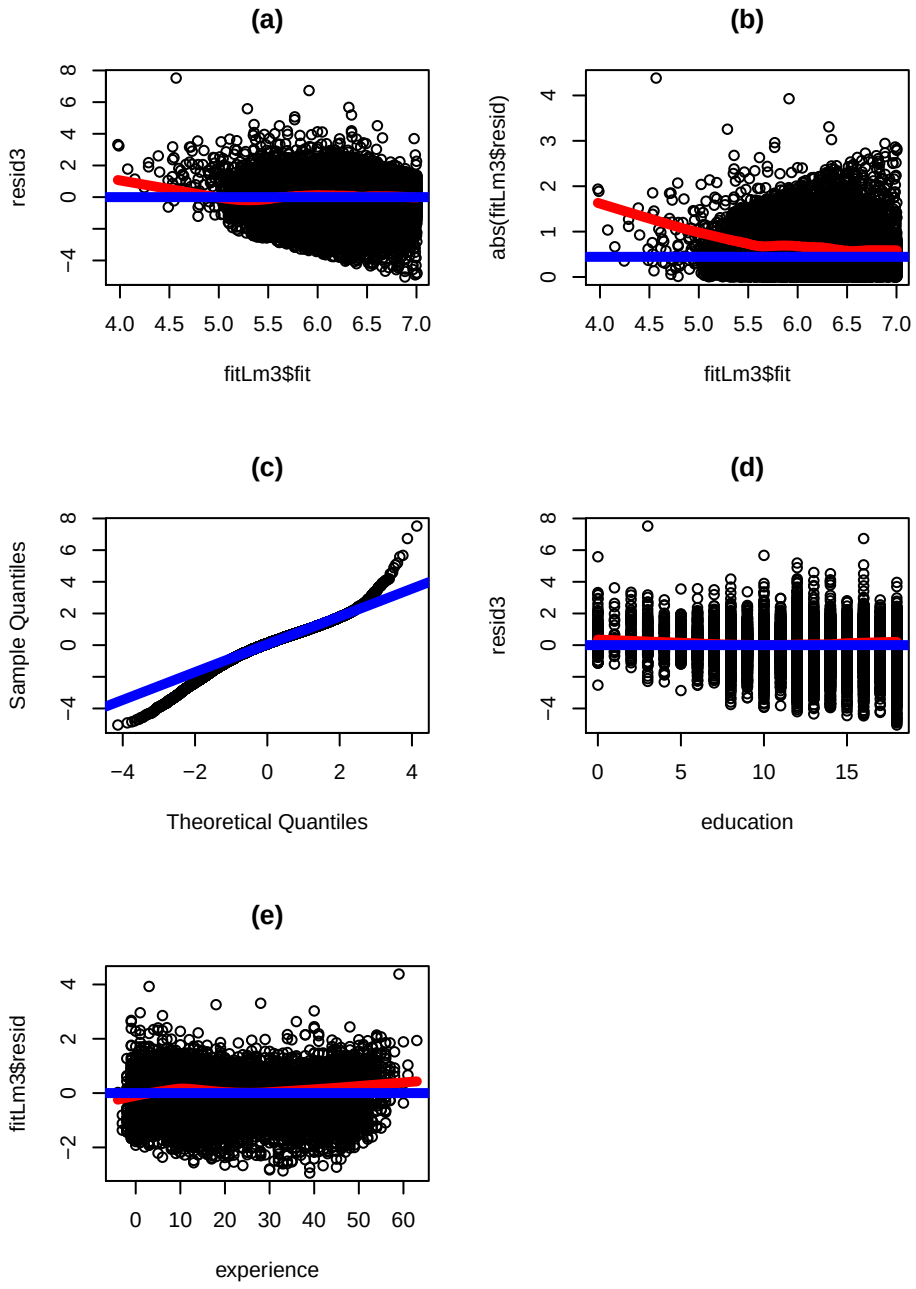


Although the model seems reasonably satisfactory, one might also use a quadratic effect for `education`. The summary below shows that the quadratic effect is highly significant.

```
Call:
lm(formula = log(wage) ~ poly(education, 2) + poly(experience,
    2) + ethnicity)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-2.99484 -0.31713  0.05801  0.37694  4.21137

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            6.189552   0.003618 1710.59   <2e-16 ***
poly(education, 2)1   41.289117   0.618231   66.79   <2e-16 ***
poly(education, 2)2    7.162751   0.590998   12.12   <2e-16 ***
poly(experience, 2)1  39.912698   0.615929   64.80   <2e-16 ***
poly(experience, 2)2 -41.522915   0.591486  -70.20   <2e-16 ***
ethnicityafam         -0.238888   0.012890  -18.53   <2e-16 ***
---

Residual standard error: 0.5824 on 28149 degrees of freedom
Multiple R-squared: 0.3382,     Adjusted R-squared: 0.3381
F-statistic:  2877 on 5 and 28149 DF,  p-value: < 2.2e-16
```

The residual plots for this model (below) seem highly satisfactory.

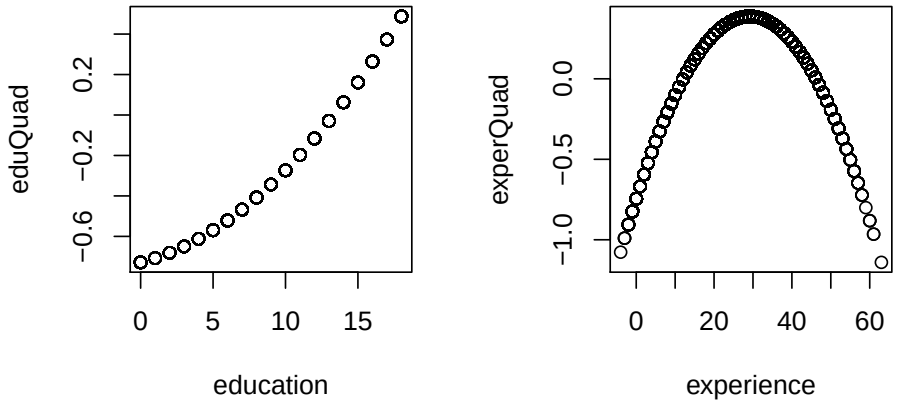


Problem 4. The effect of ethnicity can be described most easily. The regression coefficient is -0.239. This is the expected difference between log of `wage` for a Caucasian and an African-American with the same levels of eduction. Thus, the wages of the African-American will be only $\exp(-0.239) \times 100\% = 79\%$ of the wages of the Caucasian.

The effects of `education` and `experience` are nonlinear and are best described by graphs. These are produced by the following code and shown below the code. The effect of `education` on the log of `wage` is monotonically increasing

and convex. The effect of `experience` is nonmonotonic and concave. The shapes of these curves are constrained by the use of quadratic models, and it might be preferable to use nonparametric models.

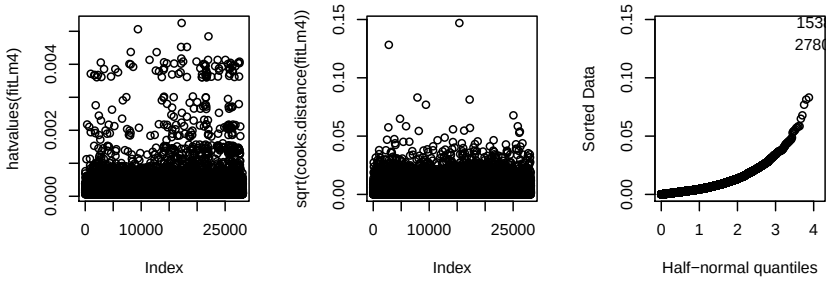
```
eduSpline = poly(education,2) %*% fitLm4$coef[2:3]
experSpline = poly(experience,2) %*% fitLm4$coef[4:5]
par(mfrow=c(1,2))
plot(education,eduSpline)
plot(experience,experSpline)
```



Problem 5. The plots are below. None of the hat diagonals are unusually large. A few are a bit larger than $2p/n$ which is 0.00043 here ($p = 6$), but a few values greater than $2p/n$ is to be expected since there are over 28,000 observations.

There is one unusually high value of Cook's D and it is observation 15,387. (This number is somewhat obscured on the half-normal plot but can be checked by printing the square root of Cook's D for observations 15,380–15,389.) We should examine this individual. As seen in the printout below, this person has a third grade education, 59 years of potential work experience and earns \$7716/week or over \$400,000 dollars per year. Clearly, this is an exceptional individual, but there is no reason to suspect that this information is incorrect. Since `wage` is log transformed in our model, this individual should not have a unduly large effect on the fit. He or she might be at least part of the reason why the effect of `experience` on the log of `wage` increases when `experience` exceeds 50.

```
> CPS1988[15387,]
         wage education experience ethnicity smsa region parttime
15387 7716.05         3         59      cauc  yes  south      yes
```



Exercise 1. Panel (a) shows clearly that the effect of X on Y is nonlinear. Because of the strong bias caused by the nonlinearity, the residuals are biased estimates of the noise and examination of the remaining plots is not useful. The remedy to this problem is to fit a nonlinear effect. A model that is quadratic in X seems like a good choice. After this model has been fit, the other diagnostic plots should be examined.

Exercise 4. Panel (a) shows some nonlinearity and right skewness is evident in several of the plots, especially the QQ plot in panel (c). The skewness suggests using a transformation of Y, e.g., a log or square root transformation. After the transformation, the residuals should be re-plotted to see if the nonlinearity was also removed by the transformation.

Exercise 5. Leverages only depend on the predictor variables. A high leverage means that this observation had unusual values of the predictors. Leverage measures the *potential* to be influential. A high-leverage point will have a large Cook's D only if it is also a residual outlier.