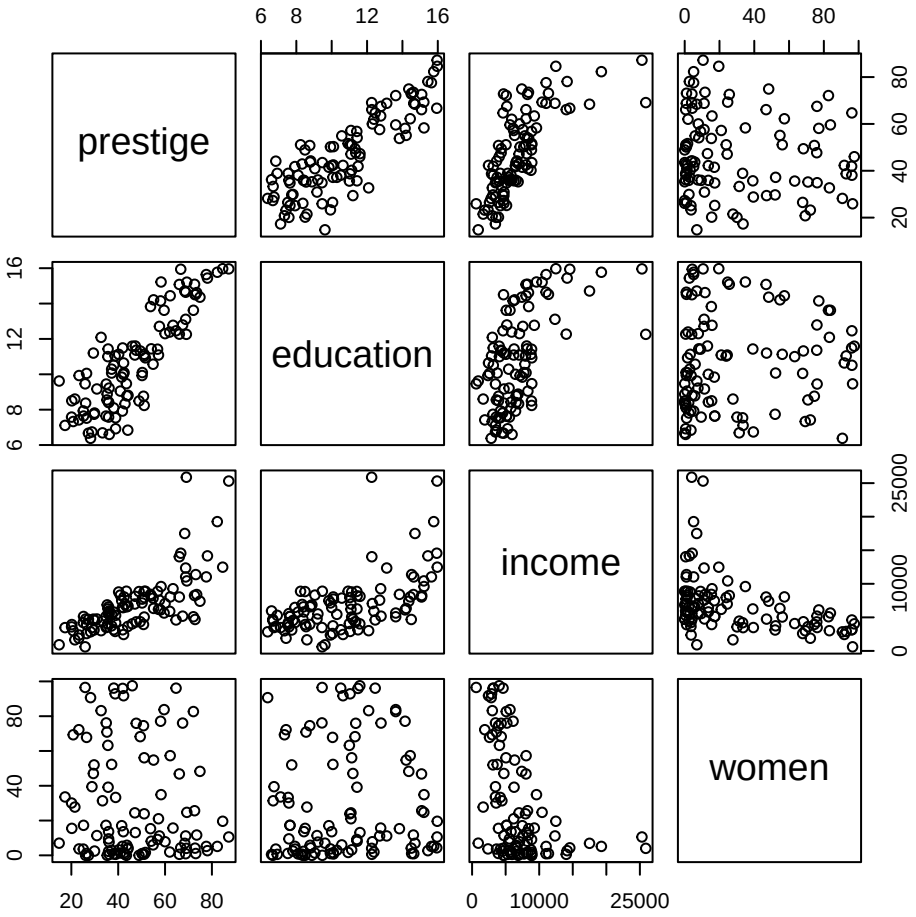
Multiple regression activity: solutions

STAT3530 Applied Linear Models

Use these to check your work on the `YOUR TURN` section of `scripts/09-multiple-regression.R`. The goal was to decide which of education, income, and percentage women are associated with prestige, and say how much each is worth.

1. Look at the data

```
pairs(prestige[, c('prestige', 'education', 'income', 'women')])
```



Prestige rises steadily with education, so that relationship looks straight. Income is strongly right-skewed, and its relationship with prestige rises quickly and then levels off — not a straight line. Percentage women shows no obvious relationship with prestige on its own.

2. Fit the model

```
naive <- lm(prestige ~ education + income + women, data = prestige)
round(summary(naive)$coefficients, 4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.7943     3.2391 -2.0976   0.0385
education     4.1866     0.3887 10.7708   0.0000
income        0.0013     0.0003  4.7288   0.0000
women        -0.0089     0.0304 -0.2929   0.7702
```

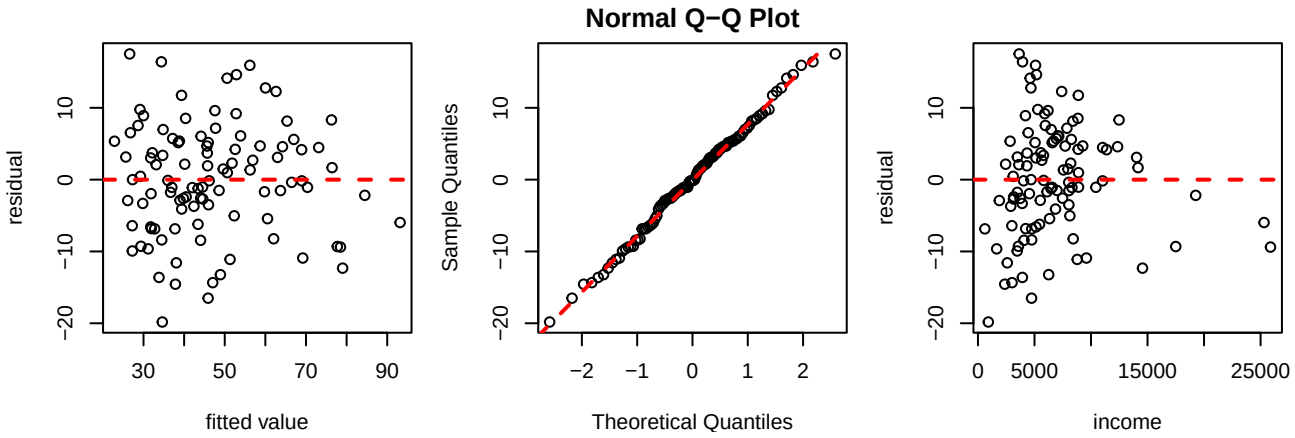
Education and income are significant, women is not ($p = 0.77$). $R^2 = 0.798$.

3. Check the assumptions

```
plot(fitted(naive), resid(naive), xlab = 'fitted value', ylab = 'residual')
abline(h = 0, col = 'red', lty = 2, lwd = 2)

qqnorm(resid(naive)); qqline(resid(naive), col = 'red', lty = 2, lwd = 2)

plot(prestige$income, resid(naive), xlab = 'income', ylab = 'residual')
abline(h = 0, col = 'red', lty = 2, lwd = 2)
```



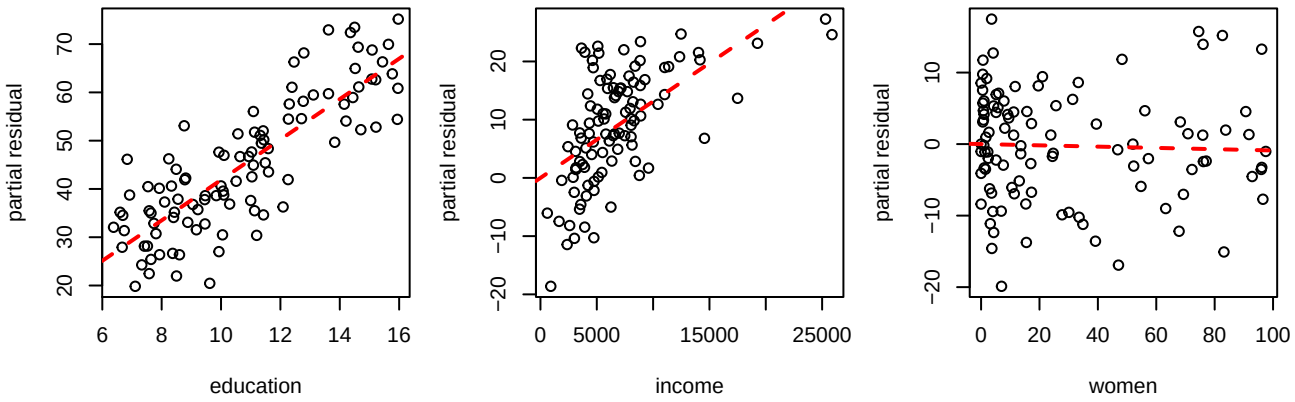
The QQ plot is acceptable. The residuals against fitted values bend, and the residual plot against income shows why: the residuals start negative, rise, and fall away again at high income. Income is on the wrong scale. The residual plots against education and women (not shown) are shapeless.

4. Partial residual plots

```
pr <- resid(naive) + coef(naive)['education'] * prestige$education
plot(prestige$education, pr, xlab = 'education', ylab = 'partial residual')
abline(lm(pr ~ prestige$education), col = 'red', lty = 2, lwd = 2)

pr <- resid(naive) + coef(naive)['income'] * prestige$income
plot(prestige$income, pr, xlab = 'income', ylab = 'partial residual')
abline(lm(pr ~ prestige$income), col = 'red', lty = 2, lwd = 2)

pr <- resid(naive) + coef(naive)['women'] * prestige$women
plot(prestige$women, pr, xlab = 'women', ylab = 'partial residual')
abline(lm(pr ~ prestige$women), col = 'red', lty = 2, lwd = 2)
```



Education and women are straight. Income climbs steeply and then flattens — the shape of a logarithm — so refit with `log(income)`.

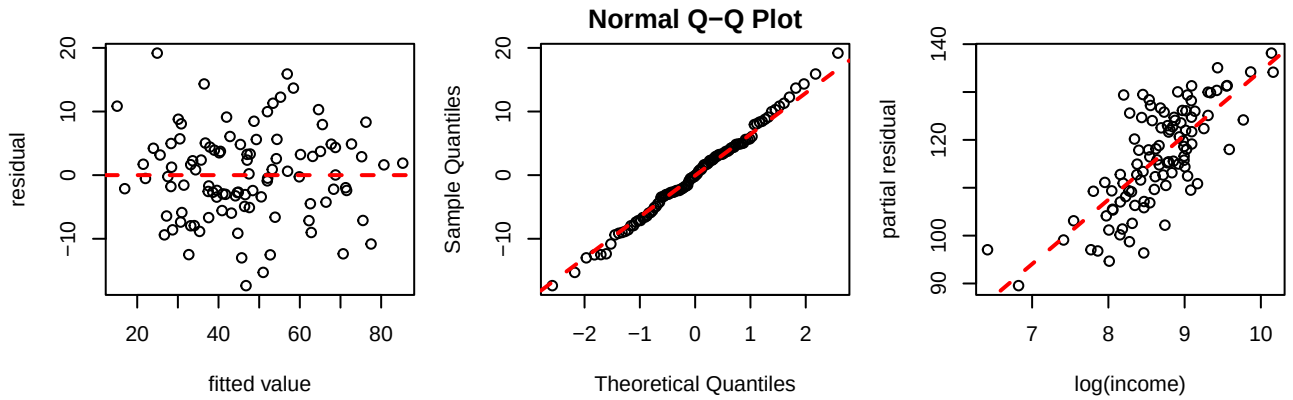
```
fit <- lm(prestige ~ education + log(income) + women, data = prestige)
round(summary(fit)$coefficients, 4)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -110.9658    14.8429 -7.4760     0.00
education      3.7305     0.3544 10.5268     0.00
log(income)   13.4382     1.9138  7.0219     0.00
women          0.0469     0.0299  1.5685     0.12
```

```r
plot(fitted(fit), resid(fit), xlab = 'fitted value', ylab = 'residual')
abline(h = 0, col = 'red', lty = 2, lwd = 2)

qqnorm(resid(fit)); qqline(resid(fit), col = 'red', lty = 2, lwd = 2)

pr <- resid(fit) + coef(fit)['log(income)'] * log(prestige$income)
plot(log(prestige$income), pr, xlab = 'log(income)', ylab = 'partial residual')
abline(lm(pr ~ log(prestige$income)), col = 'red', lty = 2, lwd = 2)
```



The bend is gone and the partial residual plot for `log(income)` is straight. R^2 rises from 0.798 to 0.835.

5. Answer the question

```r
round(confint(fit), 3)
```

```
                2.5 %  97.5 %
(Intercept) -140.421 -81.511
education      3.027   4.434
log(income)    9.640  17.236
women         -0.012   0.106
```

Education. Each additional year of average education is associated with an increase in mean prestige of between 3.0 and 4.4 points, holding income and percentage women fixed.

Income. The coefficient is on the log scale, so it is read per *doubling* rather than per dollar: doubling average income is associated with an increase in mean prestige of $\hat{\beta} \log 2 = 13.438 \times \log 2 = 9.31$ points, with a 95% interval of 6.68 to 11.95. An extra dollar is worth much more to a low-paid occupation than to a well-paid one, which is what the curvature in step 3 was telling us.

Women. The interval for the percentage women coefficient is -0.012 to 0.106, which contains zero. Once education and income are accounted for, these data do not show an association between the percentage of women in an occupation and its prestige.

Note the sign changed when income was transformed, from -0.0089 to 0.0469. Neither is significant, but a coefficient means nothing until the rest of the model is correctly specified.

One loose end. Cook's distance for the final model singles out a couple of occupations. What happens when they are set aside is a homework problem, so it is left alone here.