Homework 2
SOLUTIONS

2. RPD Ex. 1.3(30)

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\ \varepsilon_i &\sim NID(0, \alpha^2) \end{aligned}$$

$$\begin{array}{llcl} a. & E(Y_i) & = & E(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 + \beta_1 X_i + E(\varepsilon_i) = \beta_0 + \beta_1 X_i \\ b. & \sigma^2(Y_i) & = & E\left[(\beta_0 + \beta_1 X_i + \varepsilon_i)^2\right] - E^2(\beta_0 + \beta_1 X_i + \varepsilon_i) \\ & & = & E\left[(\beta_0 + \beta_1 X_i)^2 + 2(\beta_0 + \beta_i X_i)\varepsilon_1 + E_i^2\right] - (\beta_0 + \beta_1 X_i)^2 \\ & & = & (\beta_0 + \beta_1 X_i)^2 + 0 + E(\varepsilon_i^2) - (\beta_0 + \beta_i X_i) \\ & & = & \sigma^2 \\ c. & Cov(Y_i, Y_{i'}) & = & E\left[(\beta_0 + \beta_1 X_i + \varepsilon_i - \beta_0 - \beta_1 X_i)(\beta_0 + \beta_1 X_{i'} + \varepsilon_{i'} - \beta_0 - \beta_i X_{i'})\right] \\ & & = & E(\varepsilon_{i'}\varepsilon_{i'}) = 0 \quad , \quad i \neq i' \end{array}$$

3.

$a.$ Show: $\sum \hat{Y}_i e_i = 0$

$$\begin{aligned} \sum \hat{Y}_i e_i &= \sum \hat{Y}_i (Y_i - \hat{Y}_i) = \sum (\hat{\beta}_0 + \hat{\beta}_i X_i)(Y_i - \hat{\beta} - \hat{\beta}_i X_i) \\ &= \hat{\beta}_0 \sum Y_i + \hat{\beta}_i \sum X_i Y_i - n\hat{\beta}_0^2 - \hat{\beta}_1 \hat{\beta}_0 \sum X_i - \hat{\beta}_1 \hat{\beta}_0 \sum X_i - \hat{\beta}_1^2 \sum X_i^2 \\ &= \hat{\beta}_i \underbrace{\left(\sum Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum X_i\right)}_{\text{Normal Equation 1}} + \hat{\beta}_1 \underbrace{\left(\sum X_i Y_i - \hat{\beta}_0 \sum X_i - \hat{\beta}_1 \sum X_i^2\right)}_{\text{Normal Equation 2}} \\ &= \hat{\beta}_0 \cdot 0 + \hat{\beta}_1 \cdot 0 = 0 \end{aligned}$$

$b.$ Show: $\sum Y_i = \sum \hat{Y}_i$

$\sum Y_i = \sum(\overline{Y} - \beta_1 \overline{X} + \beta_1 X_i) = n\overline{Y} - \beta_1 \sum(\overline{X} - X_i) = n\overline{Y}$ QED

$c.$ Show: $S_{YY} = SS_{\text{Reg}} + RSS$

$$\begin{array}{lccl} \iff & \sum(Y_i - \overline{Y})^2 = \sum(\hat{Y}_i - \overline{Y})^2 & & + \sum(Y_i - \hat{Y}_i)^2 \\ \iff & \sum Y_i^2 - 2\overline{Y}\sum Y_i + n\overline{Y}^2 & = & \sum \hat{Y}_i^2 - 2\overline{Y}\sum \hat{Y}_i + n\overline{Y}^2 + \sum Y_i^2 \\ \iff & 0 & = & 2\left[\sum \hat{Y}_i^2 - \sum Y_i \hat{Y}_i\right] \\ \iff & 0 & = & 2\left[\sum \hat{Y}_i^2 - \sum(\hat{Y}_i + e_i)\hat{Y}_i\right] \\ \iff & 0 & = & 2\left[\sum \hat{Y}_i^2 - \sum \hat{Y}_i^2 - \sum \hat{Y}_i e_i\right] \\ \iff & 0 & = & -2\sum \hat{Y}_i e_i \\ \iff & 0 & = & 0 \qquad \text{QED} \end{array}$$

Homework 2

4.

$$
\begin{aligned}
&a. \quad \text{Show: } \hat{\beta}_1 = r_{xy}\cdot\tfrac{S_Y}{S_X}\\
&\hat{\beta}_1 = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2} = \frac{S_Y}{S_X}\cdot\frac{\sum\left(\frac{X_i-\overline{X}}{S_X}\right)\left(\frac{Y_i-\overline{Y}}{S_Y}\right)}{\sum\frac{\left(X_i-\overline{X}\right)^2}{S_X^2}} = \frac{S_Y}{S_X}\cdot r_{XY}
\end{aligned}
$$

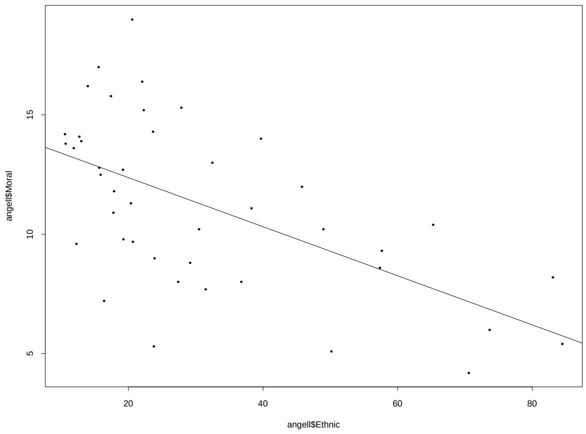
$$
\begin{aligned}
b. \qquad R^2 &= \frac{S_{XY}^2}{S_{XX}S_{YY}} = \frac{\left[\sum(X_i-\overline{X})(Y_i-\overline{Y})\right]^2}{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2} = \frac{\frac{1}{S_X^2 S_Y^2}\cdot\left[\sum(X_i-\overline{X})(Y_i-\overline{Y})\right]^2}{\frac{1}{S_X^2 S_Y^2}\sum(X_i-\overline{X})\sum(Y_i-\overline{Y})^2}\\
&= \frac{\left[\sum\frac{X_i-\overline{X}}{S_X}\cdot\frac{Y_i-\overline{Y}}{S_Y}\right]^2}{\left[\sum\left(\frac{X_i-\overline{X}}{S_x}\right)^2\sum\left(\frac{Y_i-\overline{Y}}{S_Y}\right)^2\right]} = R_{XY}^2
\end{aligned}
$$

* Other approach:

$$
\begin{aligned}
SS_{\text{Reg}} &= \sum(\hat{Y}_i-\overline{Y})^2 = \sum(\hat{\beta}_0+\hat{\beta}_i X_i-\overline{Y})^2 = \sum(\overline{Y}-\hat{\beta}_1\overline{X}+\hat{\beta}_1 X_i-\overline{Y})^2\\
&= \hat{\beta}_1^2\sum(X_i-\overline{X})^2\\
=> R^2 &= \frac{SS_{\text{Reg}}}{S_{YY}} = \frac{\hat{\beta}_1^2\sum(X_i-\overline{X})^2}{S_{YY}} = \frac{r_{XY}^2\cdot\frac{S_Y^2}{S_X^2}\cdot\sum(X_i-\overline{X})^2}{S_{YY}} = r_{XY}^2
\end{aligned}
$$

5. Regress `Moral` on `Ethnic`.

(a)
```
> myfit _ lm(Moral ~ Ethnic, data = angell)
> summary(myfit)

Call: lm(formula = Moral ~ Ethnic, data = angell)
Residuals:
    Min     1Q Median     3Q    Max
 -6.678 -2.61 0.2493 2.297 6.693

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  14.4236   0.8251    17.4815   0.0000
     Ethnic  -0.1028   0.0221    -4.6450   0.0000

Residual standard error: 2.926 on 41 degrees of freedom
Multiple R-Squared: 0.3448
F-statistic: 21.58 on 1 and 41 degrees of freedom, the p-value is 3.486e-05

Correlation of Coefficients:
       (Intercept)
Ethnic -0.8411

> plot(angell$Ethnic,angell$Moral)
> abline(myfit)
```



From the summary of the regression and from the scatter plot we see that `Moral` has a generally negative relationship with `Ethnic`. This is borne out by the estimated regression equation

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ & 14.4236 - 0.1028 \cdot x \end{aligned}$$

The residual standard error is listed directly in the `summary` printout as well,

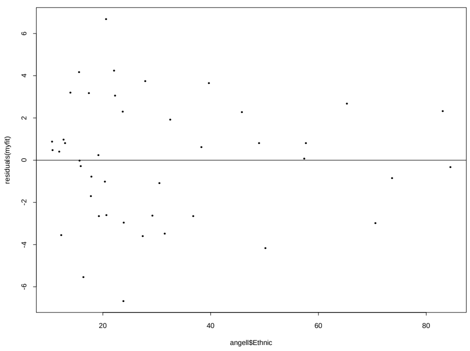
$$s = \sqrt{RSS/(n-2)} = 2.926$$

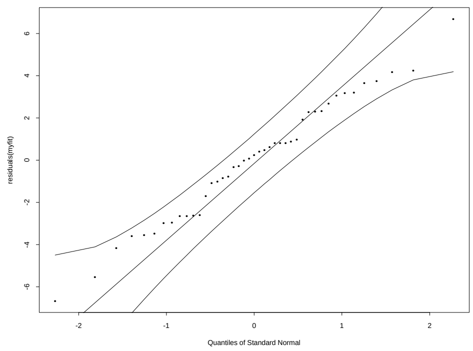
and the correlation between `Ethnic` and `Moral` is

$$r_{xy} = \sqrt{R^2} = \sqrt{0.3448} = -0.5872$$

(with the negative sign corresponding to the negative value of $\hat{\beta}_1$).

(b)
```
> plot(angell$Ethnic,residuals(myfit))
> abline(h=0) # or abline(0,0)
> qqenv(residuals(myfit))
```

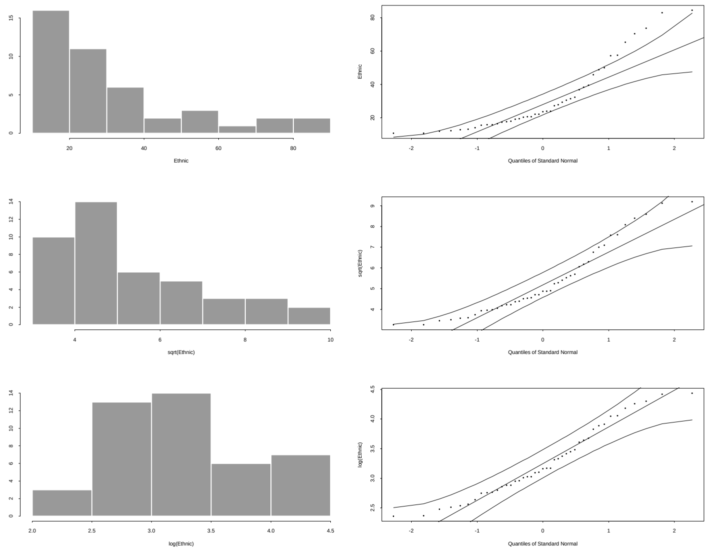




The quantile normal plot (on the right) doesn't show any severe violation of the assumption of normality for the residuals from the regression, and the raw residual plot (on the left) doesn't show any particular pattern of positive and negative residuals, and no unusual features (outliers and the like). On this basis, the regression looks pretty good. However, the raw residual plot does show that the `Ethnic` variable is bunched up in lower values, with some right skewing.

6.

```
> attach(angell)
> par(mfrow=c(3,2))
> hist(Ethnic)
> qqenv(Ethnic)
> hist(sqrt(Ethnic))
> qqenv(sqrt(Ethnic))
> hist(log(Ethnic))
> qqenv(log(Ethnic))
```

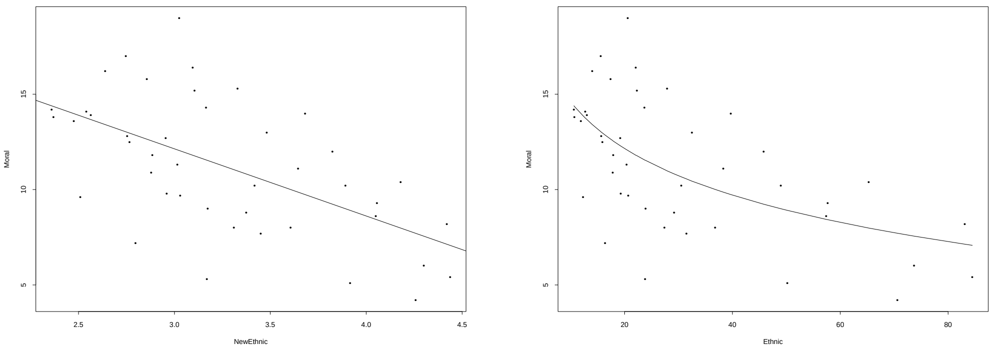


`Ethnic` has quite a lot of right skew, as we suspected from the residual plot above. `sqrtEthnic` still has some right skew, though there are relatively few violations in the quantile-normal plot. `log(Ethnic)` mostly removes the right skew (and maybe even adds a little left skew), though we have not quite got a symmetric distribution yet. Still, among these three, `log(Ethnic)` is best.

7.

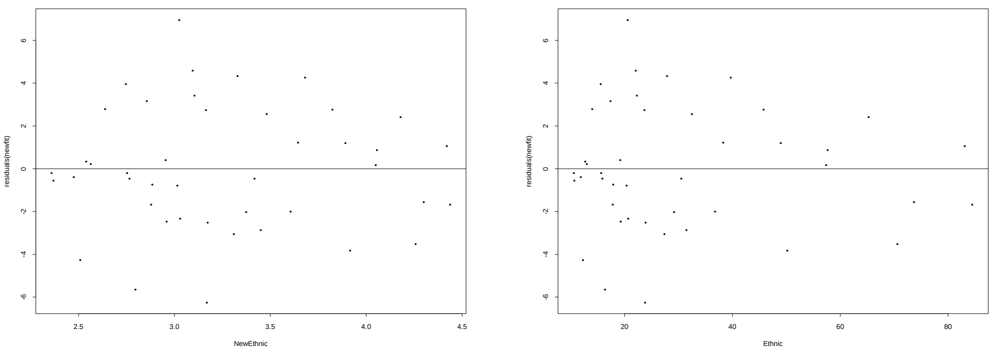
```
> NewEthnic _ log(Ethnic)
> newfit _ lm(Moral ~ NewEthnic)
> summary(newfit)

Call: lm(formula = Moral ~ NewEthnic)
Residuals:
    Min     1Q  Median    3Q   Max
 -6.242 -2.014 -0.2082 2.486 6.949

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  22.7138   2.5478     8.9150   0.0000
  NewEthnic  -3.5246   0.7678    -4.5906   0.0000

Residual standard error: 2.938 on 41 degrees of freedom
Multiple R-Squared: 0.3395
F-statistic: 21.07 on 1 and 41 degrees of freedom, the p-value is 4.141e-05
```

```
Correlation of Coefficients:
           (Intercept)
NewEthnic -0.9844
> plot(NewEthnic,Moral)
> abline(lm(Moral ~ NewEthnic))
> plot(Ethnic,Moral)
> lines(sort(Ethnic),fitted(lm(Moral ~ NewEthnic))[order(Ethnic)])
```



I like the plot on the left (with the x-axis in the `log(Ethnic)` scale) for examining the regression, since I have some practice "eyeballing" a straight-line fit to data. The plot on the right, with the x-axis in the `Ethnic` scale, is more useful for discussion with a colleague, say, in sociology, since `Ethnic` is the variable of interest, not `log(Ethnic)`. Moreover, the plot on the right clearly shows the steeper slope in the relationship between `Ethnic` and `Moral` at low values of `Ethnic`, which is totally obscured in the `log(Ethnic)` plot; this increase in slope for lower `Ethnic` values might be scientifically interesting to the sociologist.

Here are the corresponding residual plots:



Since residual plots are mostly for assessing the goodness of the regression, I would be inclined to stick with the plot on the left–again, since it corresponds to a completely linear model, which I am happier using my intuition with.