Ec 122 - HW4

Ben Juarez

10/23/2021

HGL 4.6

The fitted least squares line is $\hat{y}_i = b_1 + b_2 x_i$.

Part a

Let us algebraically show that the fitted line passes through the point of the means, $(\bar{x}, \bar{y})$. Since we have that $b_1 = \bar{y} - b_2 \cdot \bar{x}$ for the least squares estimator for β_1, we can clearly see that $(\bar{x}, \bar{y})$ must be on the line because we can simply plug in for $x_i = \bar{x}$ as follows:

$$\hat{y}_i = \bar{y} - b_2\bar{x} + b_2 x_i \implies \hat{y}_i = \bar{y}$$

Part b

Let us algebraically show that the average value of $\hat{y}_i$ equals the sample average of y. In other words, we will show that $\bar{\hat{y}} = \bar{y}$ where $\bar{\hat{y}} = \sum \hat{y}_i / N$:

$$\bar{\hat{y}} = \sum \hat{y}_i / N = (\frac{1}{N}) \sum (b_1 + b_2 x_i) = (\frac{1}{N})(\sum b_1 + \sum b_2 x_i)$$

$$\implies \bar{\hat{y}} = (\frac{1}{N})(N \cdot b_1 + b_2 \cdot \sum x_i) = (b_1 + b_2 \cdot \frac{\sum x_i}{N})$$

We know $\bar{x} = \sum x_i / N$, so we can arrive at the following:

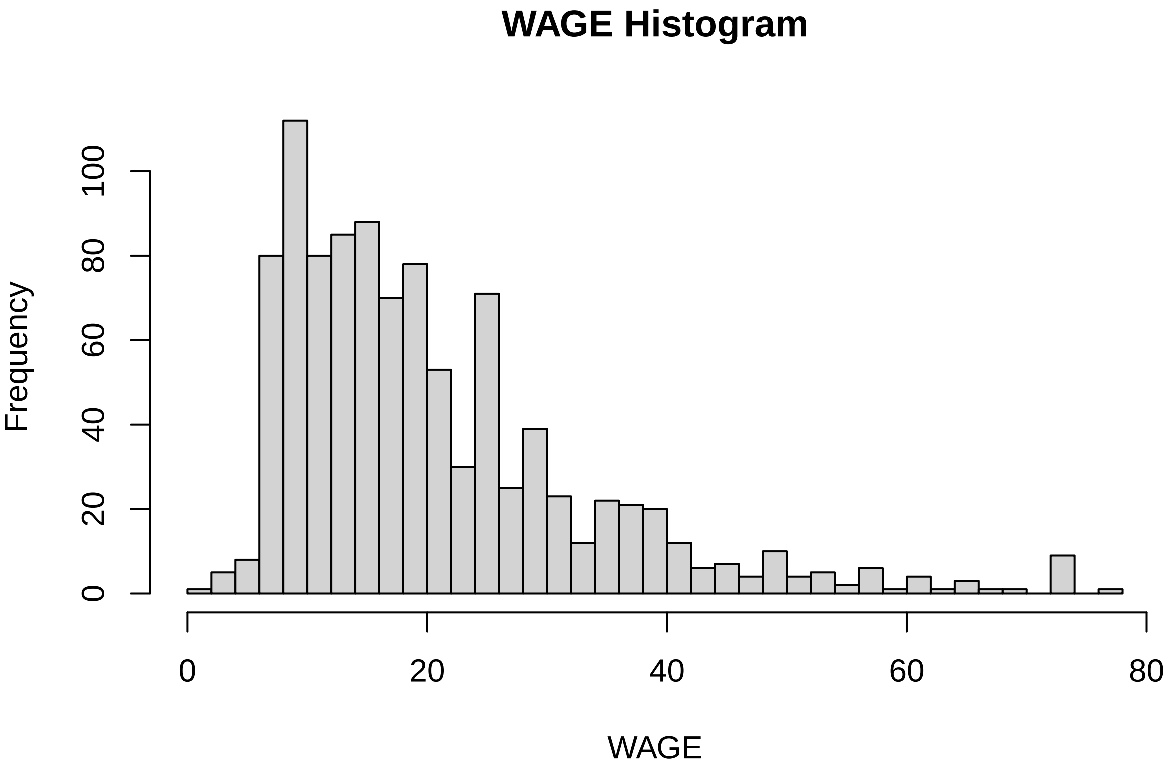
$$\implies \bar{\hat{y}} = b_1 + b_2 \cdot \bar{x} = \bar{y}$$

HGL 4.14

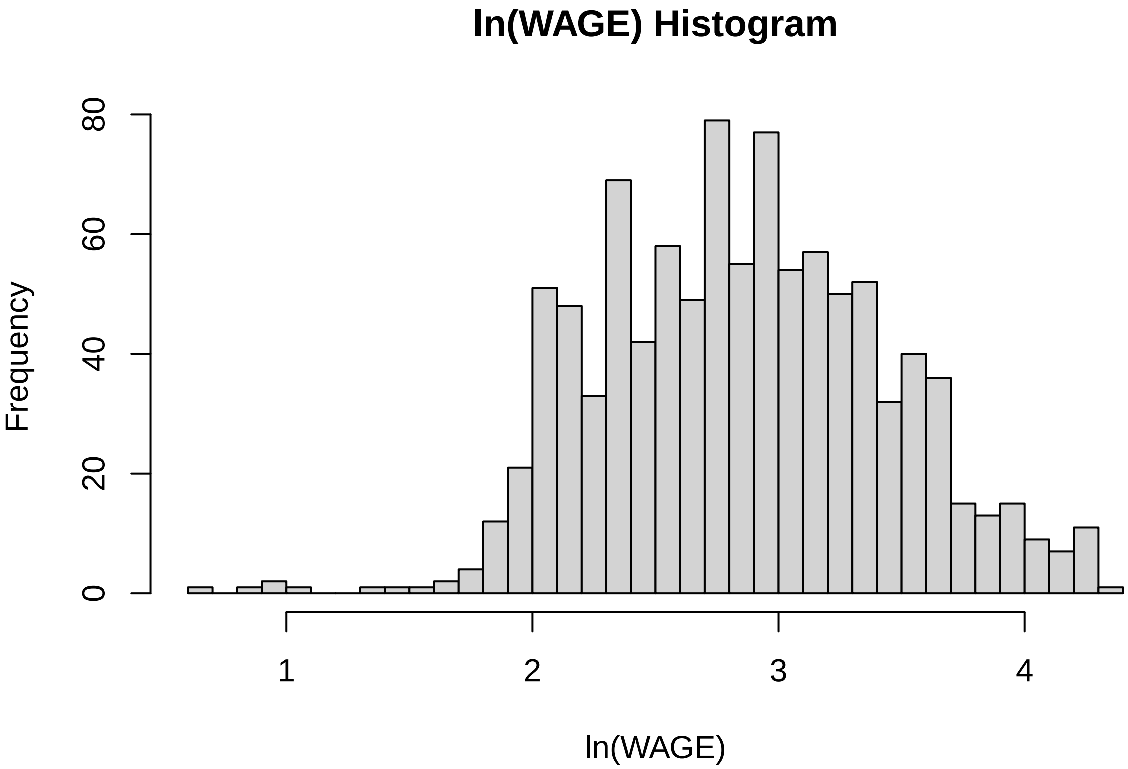
```r
setwd("~/Desktop/R")
data1 = read.table("dat/cps4_small.dat")
colnames(data1) = c("wage", "educ", "exper", "hrswk", "married", "female", "metro", "midwest", "south",
```

Part a

```r
hist1 = hist(data1$wage, breaks = 30, xlab = "WAGE", main = "WAGE Histogram")
```



```
hist2 = hist(log(data1$wage), breaks = 30, xlab = "ln(WAGE)", main = "ln(WAGE) Histogram")
```



It appears that the histogram for ln(WAGE) is more normally distributed since the histogram for wage is skewed right.

Part b

```r
educ = data1$educ
wage = data1$wage

wage_lm = lm(wage ~ educ)
wage_llr = lm(log(wage) ~ educ)
summary(wage_lm)
```

```
##
## Call:
## lm(formula = wage ~ educ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.626  -7.816  -2.623   5.019  55.376
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.7103     1.9142  -3.506 0.000476 ***
## educ          1.9803     0.1361  14.548  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.66 on 998 degrees of freedom
## Multiple R-squared:  0.175,  Adjusted R-squared:  0.1741
## F-statistic: 211.7 on 1 and 998 DF,  p-value: < 2.2e-16
```

```r
summary(wage_llr)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.55876 -0.39176  0.00699  0.36057  1.58413
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.609444   0.086423   18.62   <2e-16 ***
## educ        0.090408   0.006146   14.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5266 on 998 degrees of freedom
## Multiple R-squared:  0.1782, Adjusted R-squared:  0.1774
## F-statistic: 216.4 on 1 and 998 DF,  p-value: < 2.2e-16
```

```r
mean(wage)
```

```
## [1] 20.61566
```

```r
unname(coef(wage_lm)[2]/mean(wage) * 100)
```

```
## [1] 9.605744
```

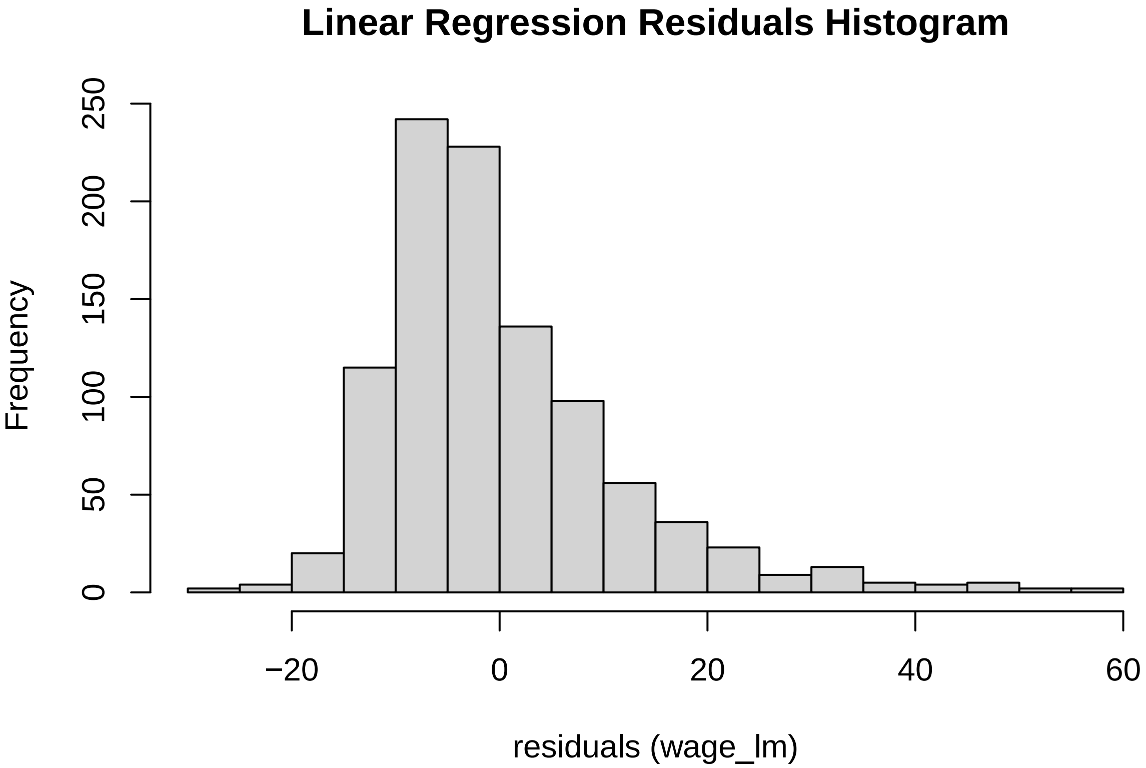
```
unname(coef(wage_llr)[2] * 100)
```

```
## [1] 9.040825
```

For an additional year of education, the percentage increase in wages for the average worker (20.6 dollars/hr) is either about 9.6% (linear model) or 9.04% (log-linear model).

Part c

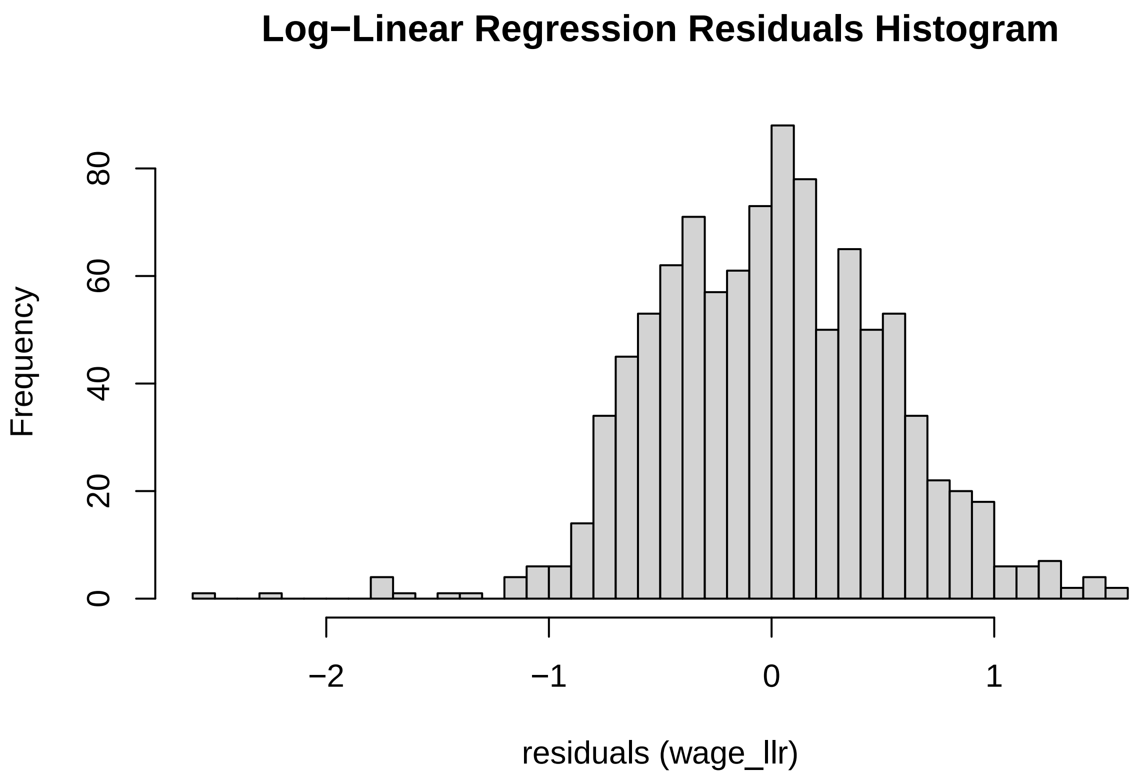
```
hist3 = hist(resid(wage_lm), breaks=30, xlab = "residuals (wage_lm)", main = "Linear Regression Residual
```



```
hist3 = hist(resid(wage_llr), breaks=30, xlab = "residuals (wage_llr)", main = "Log-Linear Regression Re

library("tseries")
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```



```
jarque.bera.test(resid(wage_lm))
```

```
##
##  Jarque Bera Test
##
## data:  resid(wage_lm)
## X-squared = 839.82, df = 2, p-value < 2.2e-16
```

```
jarque.bera.test(resid(wage_llr))
```

```
##
##  Jarque Bera Test
##
## data:  resid(wage_llr)
## X-squared = 27.528, df = 2, p-value = 1.053e-06
```

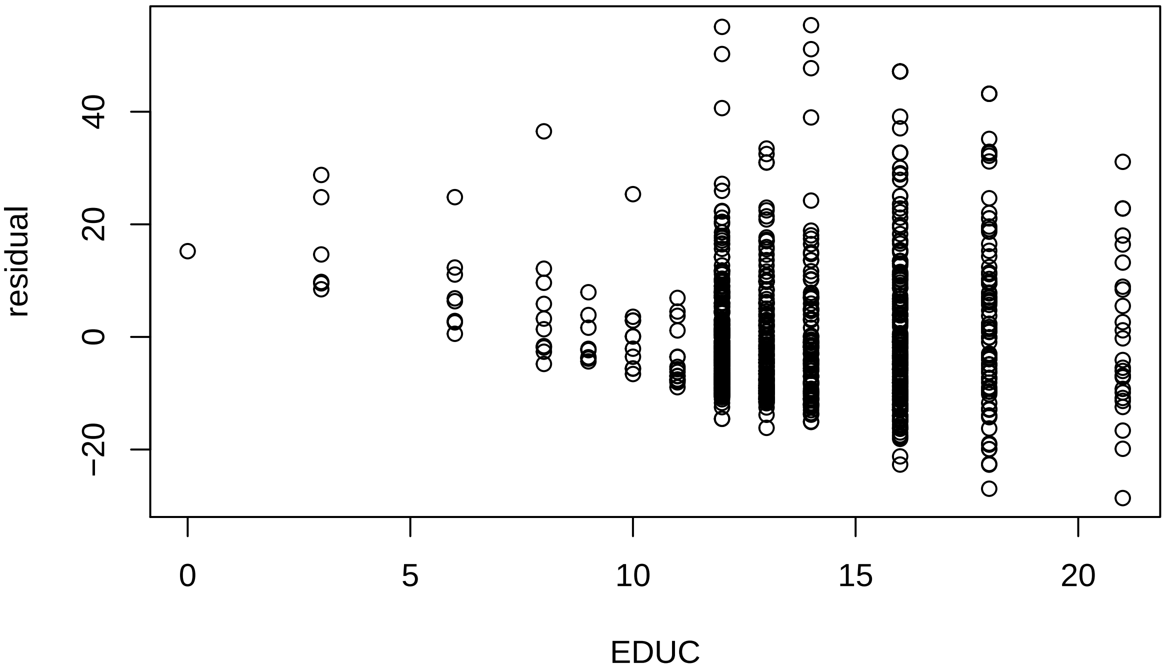
The histograms and the results from the Jarque Bera Test imply that the residuals from the log-linear model appear more compatible with normality, relative to the linear model.

Part d

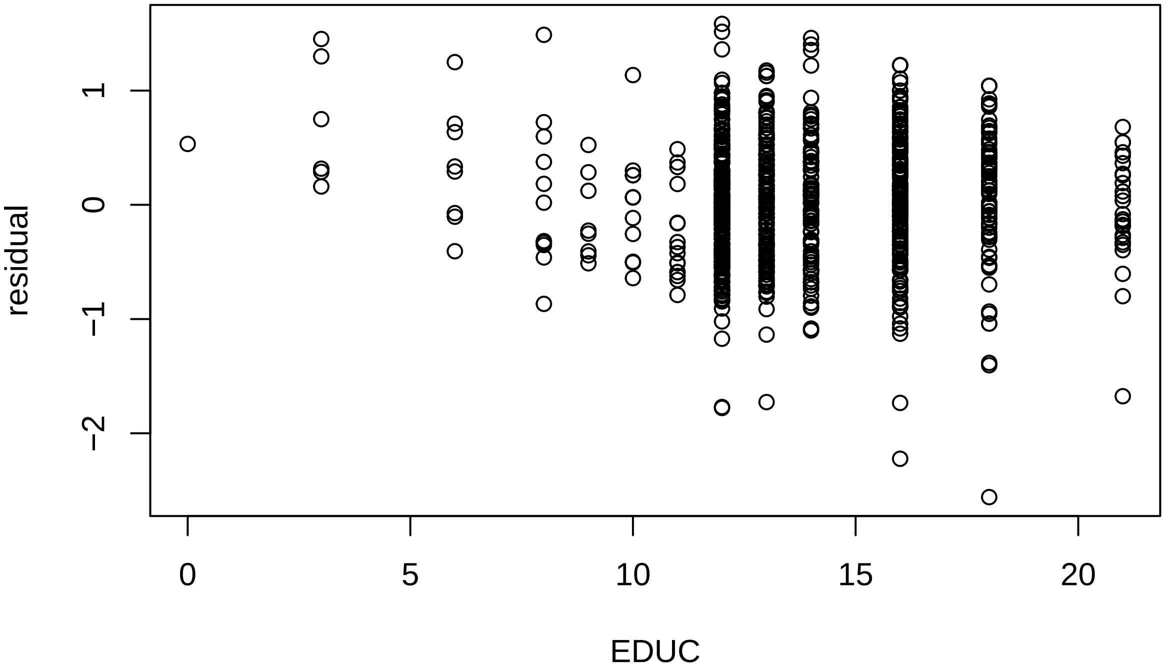
For the linear model, the R^2 is 0.175, and for the log-linear model, the R^2 is 0.1782. Although these values are similar, since $0.1782 > 0.175$, the log-linear model fits the data better.

Part e

```
plot(educ, resid(wage_lm), xlab="EDUC", ylab="residual")
```



```
plot(educ, resid(wage_llr), xlab="EDUC", ylab="residual")
```



We can see a pattern such that as EDUC increases, the residuals also increase in magnitude. Also, the residuals are essentially positive in the first quarter of the range for EDUC which could tell us something about the effect of EDUC up to a certain point (about educ = 12) when the residuals begin to change.

Part f

```
pred_lm = coef(wage_lm)[1] + coef(wage_lm)[2] * 16
pred_llr = exp(coef(wage_llr)[1] + coef(wage_llr)[2] * 16)
unname(pred_lm)
```

```
## [1] 24.97427
```

```
unname(pred_llr)
```

```
## [1] 21.24192
```

```
data_16 = as.data.frame(filter(data1, educ == 16))
mean(data_16$wage)
```

```
## Warning in mean.default(data_16$wage): argument is not numeric or logical:
## returning NA

## [1] NA
```

So, the predictions for the wage of a worker with 16 yrs. of education are ≈ 25 (linear model) and ≈ 21 (log-linear model). With the actual average wage of all workers with 16 yrs. of experience being 25.5, we can see that the linear model has a more accurate prediction.

Part g

I would use the log-linear model because it has a better R^2 value while also having residuals that more normally distributed. It is important to recognize that the linear model did have a more accurate prediction in the previous problem.

HGL 5.3

Part a

Using Table 5.6:

Part i

t-statistic $= \frac{b_1}{se(b_1)} = \frac{0.0091}{0.0190} = 0.479$

Part ii

Std. Error $= se(b_2) = \frac{0.0276}{6.6086} = 0.00418$

Part iii

Coefficient $= 0.0002 \cdot (-6.9624) = -0.00139$

Part iv

We know $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$:

$$SSE = 5.752896 \quad SST = (N-1) \cdot s_y^2 = (1519-1) \cdot (0.0633)^2 = 6.082459$$

$$\implies R^2 = 1 - \frac{5.75286}{6.082459} = 0.0542$$

Part v

$\hat{\sigma}^2 = \frac{SSE}{N-K} \implies \hat{\sigma} = \sqrt{\frac{SSE}{N-K}} = \sqrt{\frac{5.752896}{1519-4}} = 0.0616$

Part b

We can interpret $b_2 = 0.0276$ as the increase in WALC (household's budget on alcohol) if ln(TOTEXP) increases by 1. We can interpret $b_3 = -0.00139$ as the decrease in WALC (proportion of household's budget on alcohol) if AGE (age of household head) increases by 1 year. We can interpret $b_4 = -0.0133$ as the decrease in WALC (proportion of household's budget on alcohol) if NK increases by 1 (household gains one more child).

Part c

Let us compute a 95% interval estimate for β_3 as follows:

$$b_3 \pm t_{(1-\frac{\alpha}{2},N-K)=(0.975,1515)} \cdot se(b_3) = -0.00139 \pm 1.9615 \cdot 0.0002 = (-.00178, -0.000998)$$

This interval estimate tells us that if AGE (age of household head) increases by 1 year, then we are 95% confident that WALC (proportion of household's budget on alcohol) decreases by a value in this range.

Part d

Let us test the hypothesis (5% level of significance)that the budget proportion for alcohol does not depend on the number of children in the household: $H_0 : B_0 = 0$ vs. $H_1 : B_4 \neq 0$. So, we know $t_c = t_{(0.975,1515)} = 1.9615$ and we reject if $t > t_c$ or $t < -t_c$. We calculate t as follows:

$$t = \frac{b_4}{se(b_4)} = \frac{-0.0133}{0.0033} = -4.03$$

Since $t = -4.03 < -t_c = 1.9615$, we reject H_0 and accept H_1. This implies that NK (number of children in the household) does have an impact on the budget proportion for alcohol. The reason for this test outcome could be that having one or more children takes up a lot of the household's budget, meaning there is less money to spend on alcohol.

HGL 5.12

```r
setwd("~/Desktop/R")
data2 = read.table("dat/cocaine.dat")
colnames(data2) = c("price", "quant", "qual", "trend")
```

Part a

For B_2, I would expect the sign to be negative because it seems to make that than with a larger sale of cocaine, there would be better prices per gram. For B_3, I would expect the sign to be positive because it seems clear that higher quality cocaine would be more expensive. For B_4, I would expect the sign to be positive because it might make sense that there would be more demand as time went on (unsure about supply). I am not sure about the popularity of cocaine between 1984 and 1991, so I am unsure about this one.

Part b

```r
PRICE = data2$price
QUANT = data2$quant
QUAL = data2$qual
TREND = data2$trend
coke_lm = lm(PRICE ~ QUANT + QUAL + TREND)
summary(coke_lm)
```

```
##
## Call:
## lm(formula = PRICE ~ QUANT + QUAL + TREND)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.479 -12.014  -3.743  13.969  43.753
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 90.84669    8.58025  10.588 1.39e-14 ***
## QUANT       -0.05997    0.01018  -5.892 2.85e-07 ***
## QUAL         0.11621    0.20326   0.572   0.5700
## TREND       -2.35458    1.38612  -1.699   0.0954 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.06 on 52 degrees of freedom
## Multiple R-squared:  0.5097, Adjusted R-squared:  0.4814
## F-statistic: 18.02 on 3 and 52 DF,  p-value: 3.806e-08
```

It appears that my expectations for B_2 and B_3 were correct, but I was wrong for B_4. Perhaps demand did not match with increases in supply.

Part c

Knowing that $R^2 = 0.5097$, we can say that 50.97% of the variation in cocaine price is explained jointly by variation in quantity, quality, and time.

Part d

Let us test $H_0 : \beta_2 \geq 0$ vs $H_1 : beta_2 < 0$. We set up H_0, H_1 in this way because it makes sense that if sellers are not wiling to accept a lower price if they can make more sales in larger quantities, then B_2 would be positive. Let us perform this test with 5% significance as follows:

$$t_c = t_{(0.95,52)} = -1.675$$

$$t = \frac{b_2}{se(b_2)} = \frac{-0.05997}{0.01018} = -5.89 < t_c = -1.675$$

So, we reject H_0 which means that it is true that sellers will accept a lower price if they can make sales in larger quantities.

Part e

Let us test the hypothesis that the quality of cocaine has no influence on price ($\beta_3 = 0$) against the alternative that a premium is paid for better-quality cocaine ($\beta_3 > 0$):

$$H_0 : \beta_3 = 0 \text{ vs } H_1 : \beta_3 > 0 \implies t_c = t_{(0.05,52)} = 1.675$$

$$t = \frac{b_3}{se(b_3)} = \frac{0.1162}{0.20326} = 0.057 < t_c = 1.675$$

So, we do not reject H_0, so we cannot conclude that a premium is paid for better-quality cocaine.

Part f

We have that the average annual change in the cocaine price is -2.35. This could potentially be explained by decreasing demand and/or oversupply. There are a variety of factors that could play into these possible explanations. Perhaps the risk of getting caught selling cocaine went down or perhaps those who frequently purchased cocaine suffered too much from the negative effects of the substance.