LM HW 9

Joshua Ingram

5/2/2020

Problem 1

1.

```
prestige_1 <- lm(prestige ~ education + income + type + women, data = Prestige)
# looking at the summary to see the p-values for each variables... overall model is significant
summary(prestige_1)
```

```
##
## Call:
## lm(formula = prestige ~ education + income + type + women, data = Prestige)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7485  -4.4817   0.3119   5.2478  18.4978
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.8139032  5.3311558  -0.153 0.878994
## education    3.6623557  0.6458300   5.671 1.63e-07 ***
## income       0.0010428  0.0002623   3.976 0.000139 ***
## typeprof     5.9051970  3.9377001   1.500 0.137127
## typewc      -2.9170720  2.6653961  -1.094 0.276626
## women        0.0064434  0.0303781   0.212 0.832494
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.132 on 92 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.8349, Adjusted R-squared:  0.826
## F-statistic: 93.07 on 5 and 92 DF,  p-value: < 2.2e-16
```

```
# let's try an incremental F-test comparing a model with all variables to a model without women (highes
prestige_2 <- lm(prestige ~ education + income + type, data = Prestige)
# incremental F-test
anova(prestige_2, prestige_1)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: prestige ~ education + income + type
## Model 2: prestige ~ education + income + type + women
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     93 4681.3
## 2     92 4679.0  1    2.2881 0.045 0.8325
```

```r
# incremental F-test does not report a significant p-value... so there is not significant evidence that

# incremental F-test (line for "type") to determine if type is significant
anova(prestige_2)
```

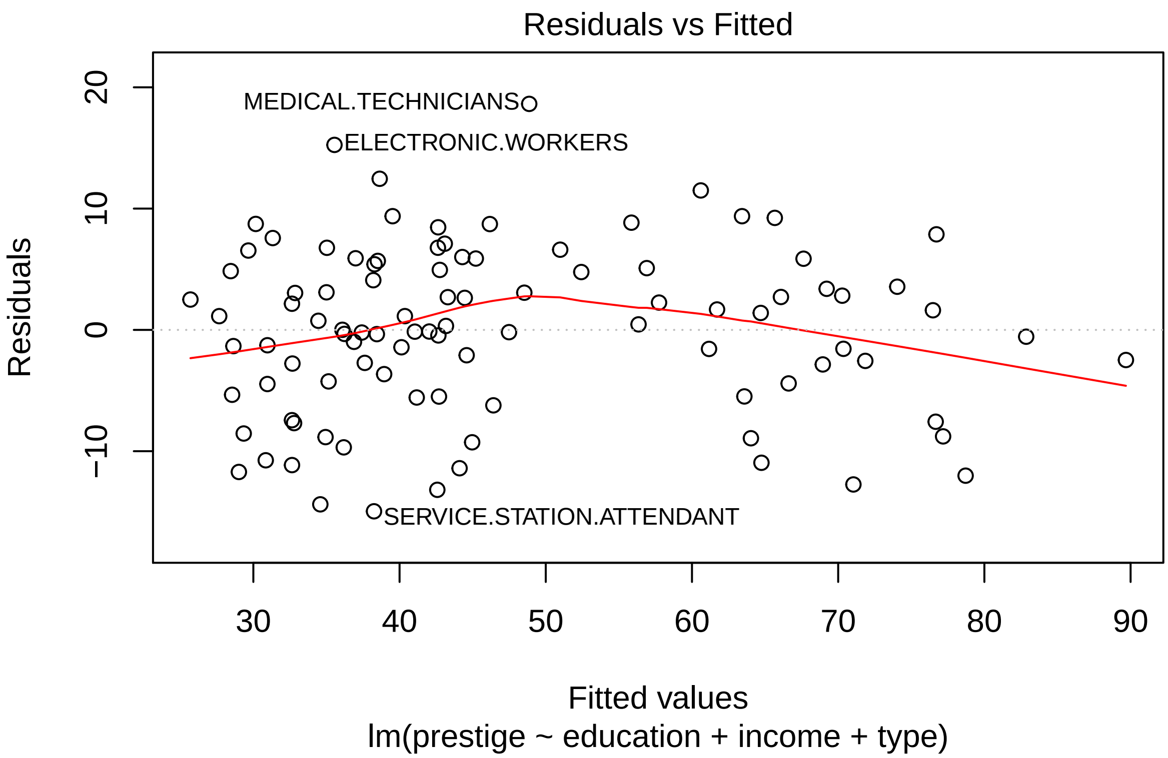
```
## Analysis of Variance Table
##
## Response: prestige
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## education  1 21282.5 21282.5 422.8056 < 2.2e-16 ***
## income     1  1792.0  1792.0  35.5999 4.355e-08 ***
## type       2   591.2   295.6   5.8721  0.003966 **
## Residuals 93  4681.3    50.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# final model is prestige_2 since type is significant
```

After conducting the incremental F-test, we've dropped women but retained all other predictors.

2.

```r
plot(prestige_2, 1)
```

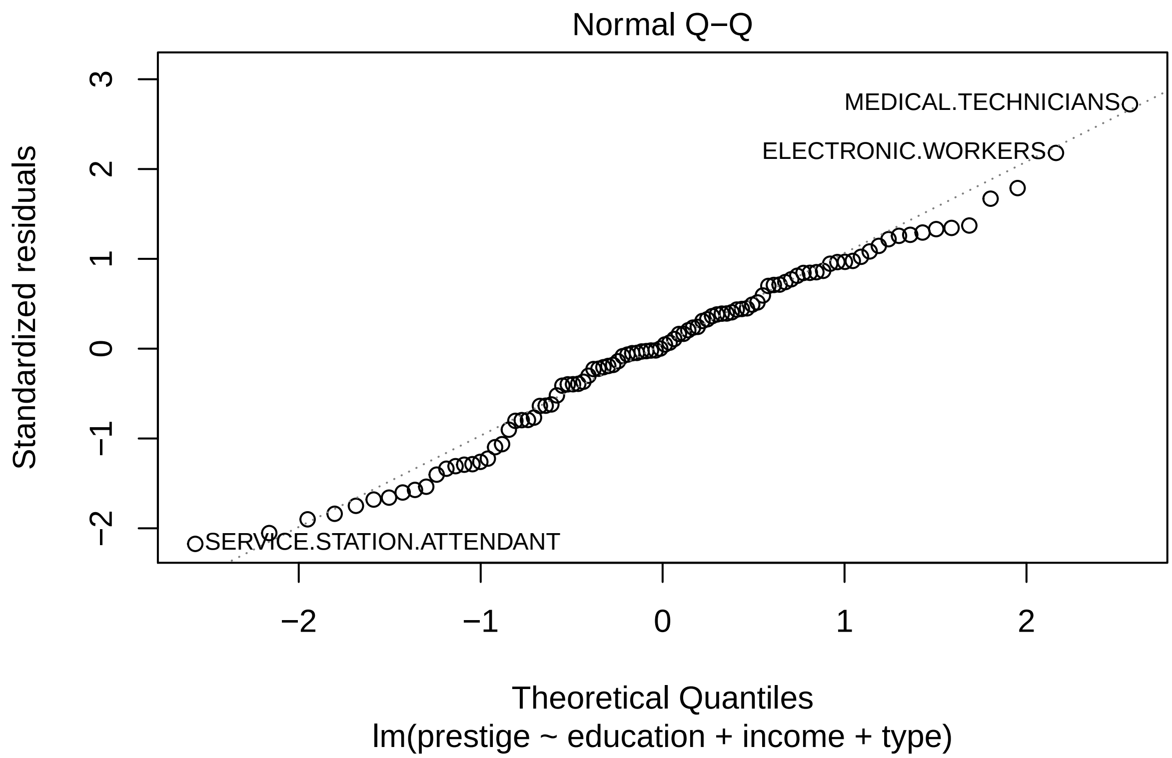


This plot allows us to check the assumption of homoscedasticity. This plot is a bit odd since there is almost a gap around 50-60 in the fitted values, but it seems that there is mostly homoscedasticity. (if we needed to adjust the model if the assumption is broken, we could perform a log transformation or sqrt on the response)

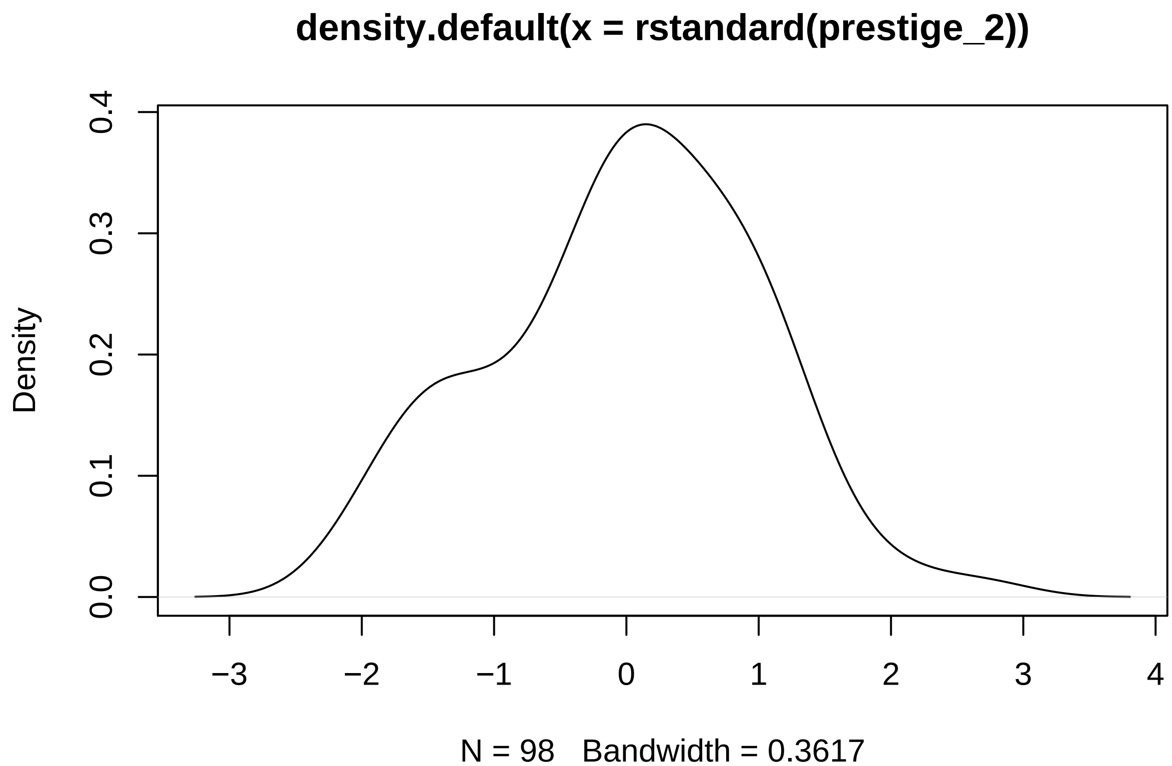
3.

We could use a normal q-q plot and a smooth density plot to check the normality assumption.

```
plot(prestige_2, 2)
```




```
plot(density(rstandard(prestige_2)))
```



In this case, the normality assumption seems to be noticeably broken. There is a subtle right-skew, as well as a non-smooth density plot... not necessarily bimodal, but there is a "hump" in the curve. It's not necessarily a big deal for inference to be conducted since we have a rather large sample and the Central Limit Theorem covers us.

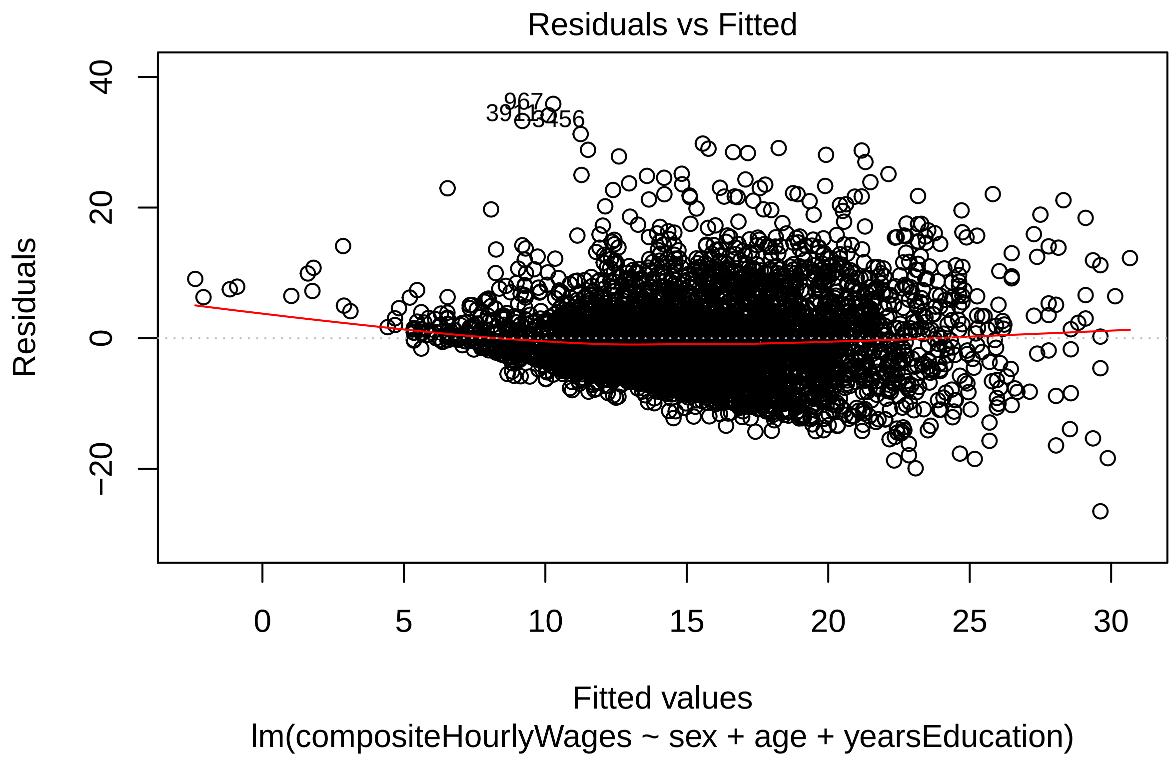
Problem 2

1.

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon_i, \epsilon_i \sim_{i.i.d} N(0, \sigma^2)$$

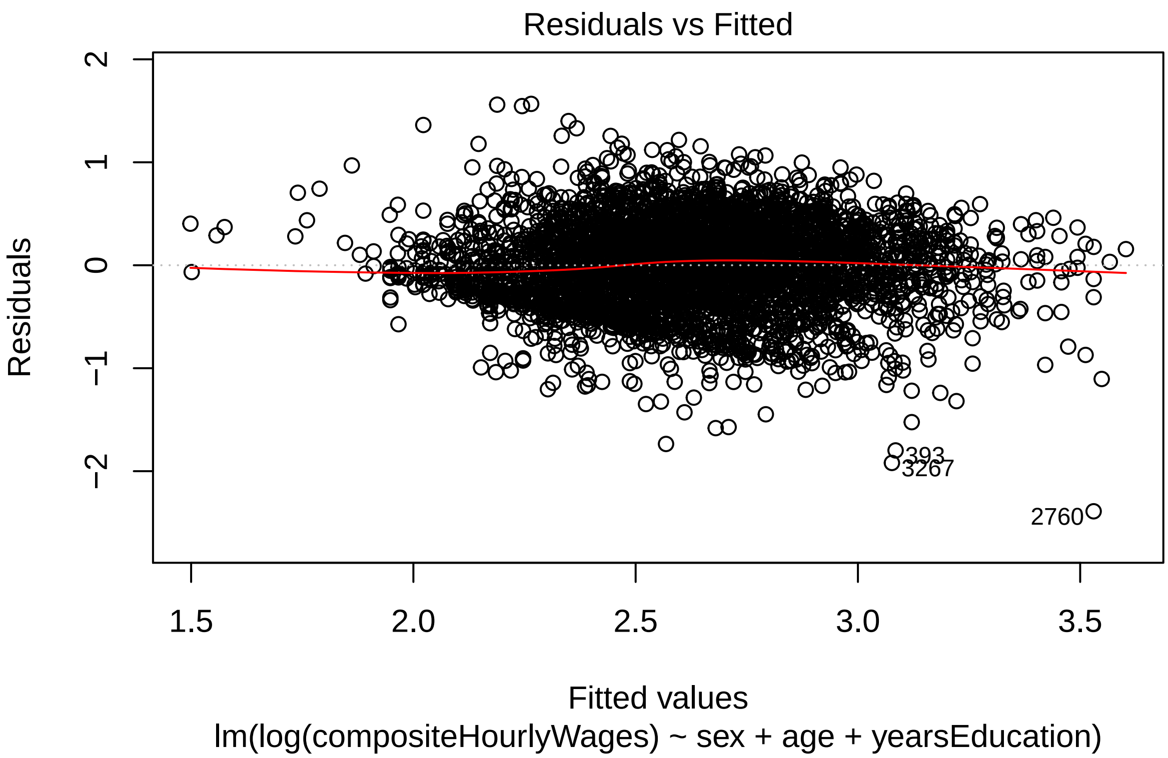
Where Y_i is *compositeHourlyWages*, $x_{1,i}$ is the dummy variable for *sex* (1 if male, 0 if female), $x_{2,i}$ is the variable for *age*, and $x_{3,i}$ is the variable for *yearsEducation*. ## 2.

```
ontario_1 <- lm(compositeHourlyWages ~ sex + age + yearsEducation, data = Ontario)
plot(ontario_1, 1)
```



There is obviously heteroscedasticity. Let's try a log transformation.

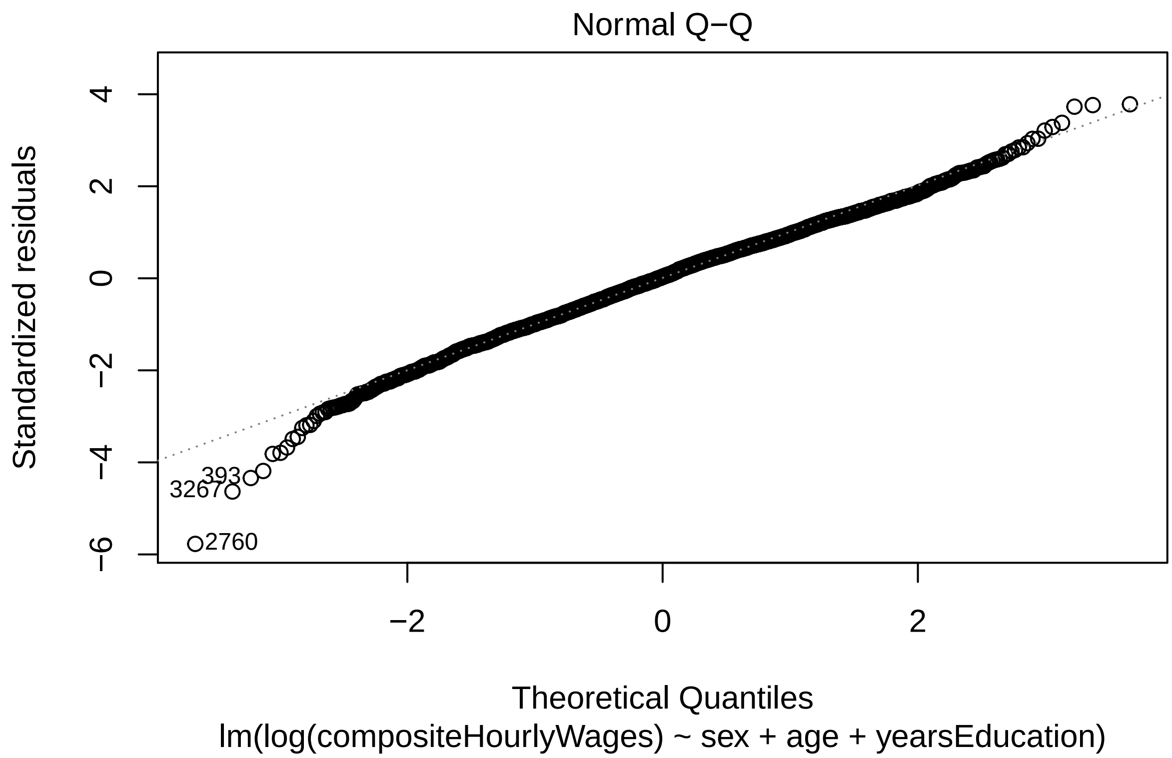
```
ontario_2 <- lm(log(compositeHourlyWages) ~ sex + age + yearsEducation, data = Ontario)
plot(ontario_2, 1)
```



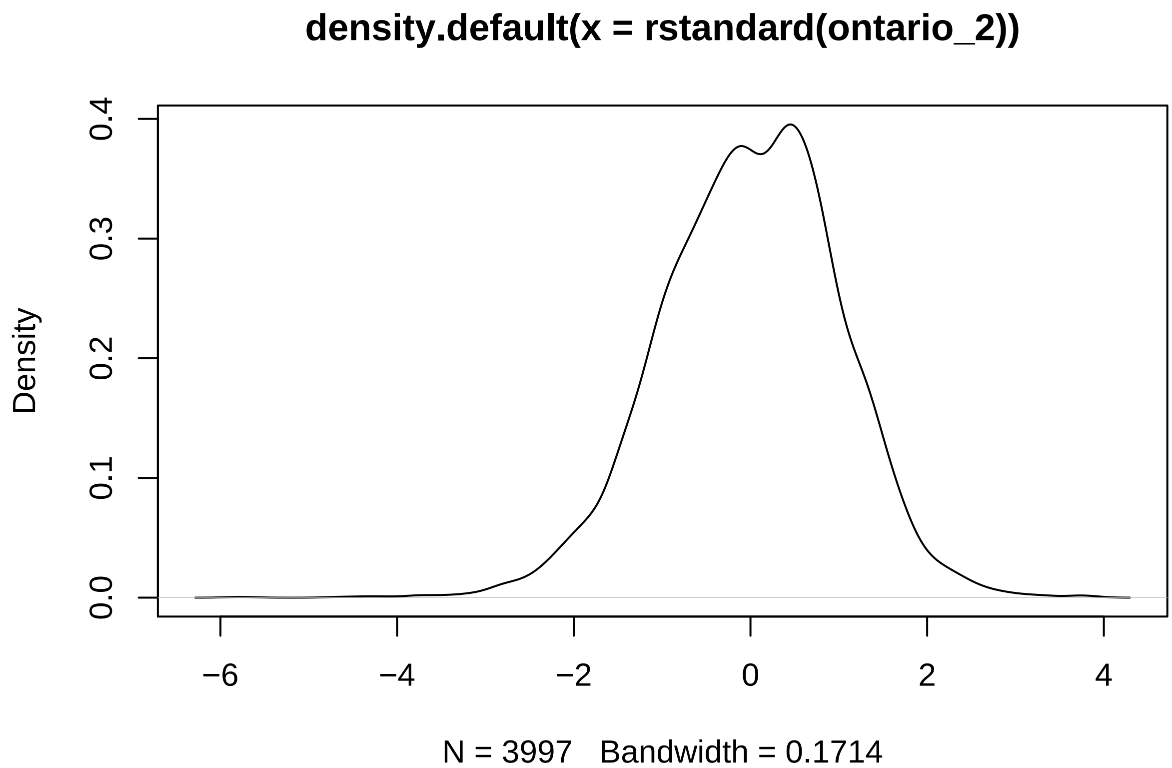
A log transformation on our response seems to fix the non-constant variance problem.

3.

```r
plot(ontario_2, 2)
```




```
plot(density(rstandard(ontario_2)))
```



The normality assumption isn't being followed that well in this example (some left-skewness), but it's not the worst. It's not a big deal if it's broken since we have a large sample. If we needed to fix this, we could use bootstrap sampling to estimate the coefficients or we could add a transformation to our variables.

4.

```r
summary(ontario_2)
```

```
##
## Call:
## lm(formula = log(compositeHourlyWages) ~ sex + age + yearsEducation,
##     data = Ontario)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.38930 -0.27670  0.01312  0.28413  1.56696
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.0990176  0.0379649   28.95   <2e-16 ***
## sexMale        0.2244959  0.0131208   17.11   <2e-16 ***
## age            0.0181548  0.0005491   33.06   <2e-16 ***
## yearsEducation 0.0558764  0.0021713   25.73   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4146 on 3993 degrees of freedom
## Multiple R-squared:  0.3212, Adjusted R-squared:  0.3207
## F-statistic: 629.7 on 3 and 3993 DF,  p-value: < 2.2e-16
```

$$log(composite\hat{H}ourlyWages) = 1.099 + 0.2245(sex_{male}) + 0.0182(age) + 0.0559(yearsEducation)$$

Interpretations:

sex - On average, we expect males to have $e^{0.2245}$ times more dollars in composite hourly wages than females, holding all else constant.

age - On average, for every one year increase in wage, we expect the composite hourly wages to multiply by $e^{0.0182}$, holding all else constant.

yearsEducation - On average, for every one year increase in the number of years of completed education, we expect the composite hourly wages to multiply by $e^{0.0559}$, holding all else constant.

Problem 3

1.

a.

```
Chile_subset <- Chile[which(Chile$vote == "Y" | Chile$vote == "N"),]
Chile_subset <- Chile_subset[which(is.na(Chile_subset$statusquo)==FALSE),]
```

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \epsilon_i, \epsilon_i \sim_{i.i.d} N(0, \sigma^2)$$

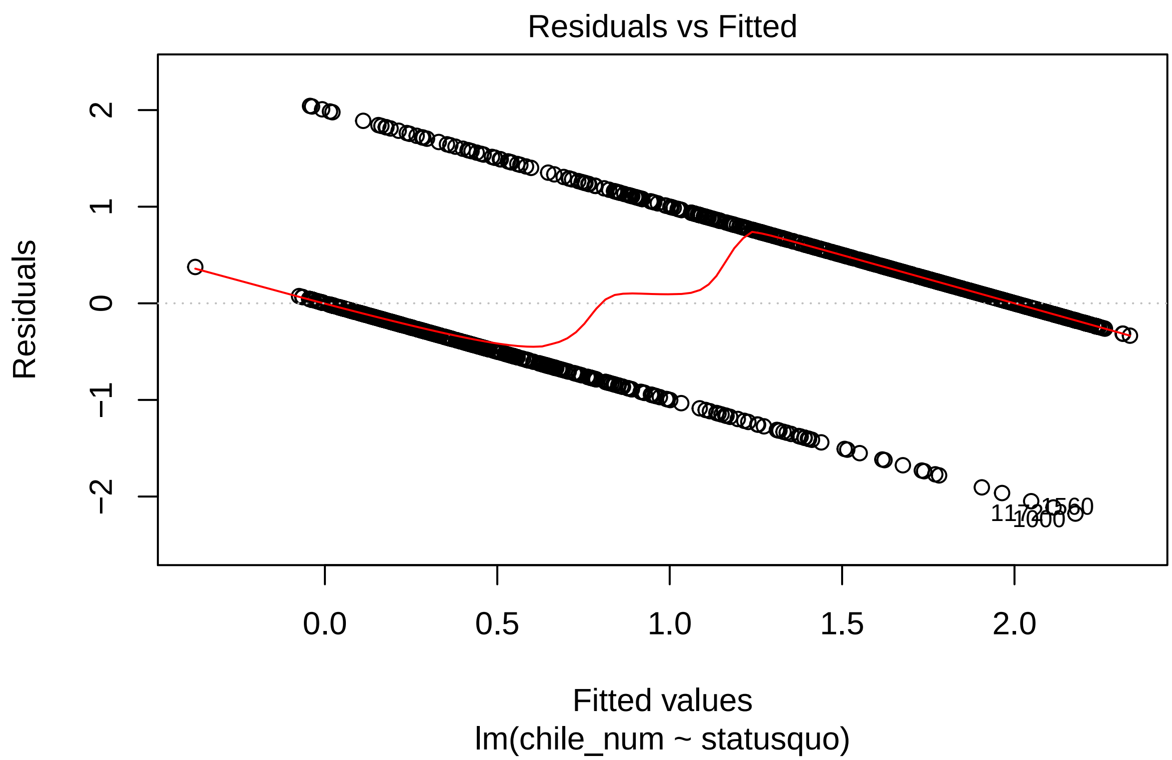
Where Y_i is vote (Y = 1, N = o) and $x_{1,i}$ is statusquo.

```
chile_num <- as.numeric(Chile_subset$vote)-2
chile_1 <- lm(chile_num ~ statusquo,
             data=Chile_subset)
summary(chile_1)
```

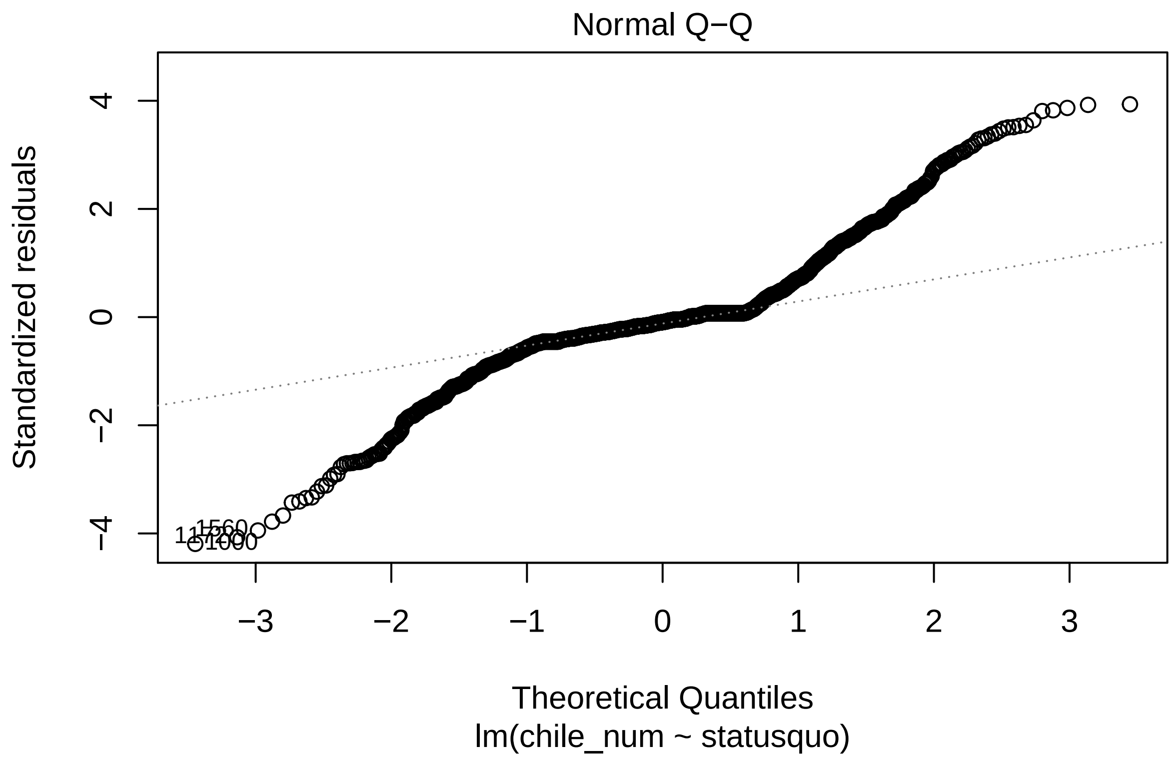
```
##
## Call:
## lm(formula = chile_num ~ statusquo, data = Chile_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.17669 -0.20447 -0.04835  0.08110  2.04304
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.98433    0.01241   79.35   <2e-16 ***
## statusquo    0.78816    0.01144   68.89   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5195 on 1752 degrees of freedom
## Multiple R-squared:  0.7303, Adjusted R-squared:  0.7302
## F-statistic:  4745 on 1 and 1752 DF,  p-value: < 2.2e-16
```

```r
plot(chile_1, 1)
```




```r
plot(chile_1,2)
```



we can already see that vote being a binary response is an issue and we should not us a classic linear regression model. We need to predict the probability of vote being Y or N, meaning our values should only be between 1 and 0. A classical linear regression model would output values betweeen negative and positive infinity. Also, our residual vs. fitted plot shows us that it the variance doesn't follow our assumption of a random "cloud" with no form and the normal q-q plot shows us that the normality assumption is completely broken... Overall, a binary response is just not appropriate for a classical linear regression model.

b.

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_{1,i}, Y_i \sim_{ind} bin(1, \pi)$$

Where Y_i is vote (Y = 1, N = o) and $x_{1,i}$ is statusquo. $\pi = P(Y_i = Y|x_{1,i})$.

```
chile_2 <- glm(vote ~ statusquo, data = Chile_subset, family = "binomial")
summary(chile_2)
```

```
##
## Call:
## glm(formula = vote ~ statusquo, family = "binomial", data = Chile_subset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1847  -0.2806  -0.1952   0.1879   2.8220
##
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.21531    0.09964   2.161   0.0307 *
## statusquo    3.20554    0.14310  22.401   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2431.28  on 1753  degrees of freedom
## Residual deviance:  752.59  on 1752  degrees of freedom
## AIC: 756.59
##
## Number of Fisher Scoring iterations: 6
```

$$log(\frac{\pi}{1-\pi}) = 0.21531 + 3.20554(statusquo)$$

The relationship between statusquo and vote is statistically significant, as we receive a z-value of 22.4 and a p-value of basically 0.

Interpretations:

odds - For every one unit increase in the scale support of the status-quo, we predict that the odds of voting yes to multiply by 3.21.

simple - As statusquo increases, we expect the probability of voting yes to increase.

c.

```
prob <- predict(chile_2, newdata=data.frame(statusquo=0),
        type="response")
prob
```

```
##         1
## 0.5536198
```

```
odds <- prob/(1-prob)
odds
```

```
##        1
## 1.240243
```

Odds of yes given statusquo = 0: 1.24

probability of yes given statusquo = 0: 0.554

d.

```
chile_prob <- predict(chile_2, type='response')
chile_predict <- ifelse(chile_prob > 0.50, "Y","N")
mean(chile_predict == Chile_subset$vote)
```

```
## [1] 0.9230331
```

We have 92.3% sample prediction accuracy for our sample.

2.

a.

```r
Chile_subset <- Chile_subset[complete.cases(Chile_subset),]
chile_3 <- glm(vote ~ ., data = Chile_subset, family = "binomial")
summary(chile_3)
```

```
##
## Call:
## glm(formula = vote ~ ., family = "binomial", data = Chile_subset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2009  -0.2753  -0.1344   0.2031   2.8616
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.046e+00  4.592e-01   2.279  0.02269 *
## regionM      7.072e-01  6.023e-01   1.174  0.24030
## regionN     -9.958e-02  3.587e-01  -0.278  0.78134
## regionS     -3.044e-01  2.928e-01  -1.040  0.29847
## regionSA    -3.012e-01  3.404e-01  -0.885  0.37619
## population   1.276e-06  1.414e-06   0.902  0.36714
## sexM        -5.515e-01  2.041e-01  -2.702  0.00689 **
## age          7.108e-04  7.472e-03   0.095  0.92422
## educationPS -9.676e-01  3.461e-01  -2.795  0.00518 **
## educationS  -6.575e-01  2.440e-01  -2.695  0.00705 **
## income      -2.972e-06  2.856e-06  -1.041  0.29807
## statusquo    3.229e+00  1.524e-01  21.184  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2360.29  on 1702  degrees of freedom
## Residual deviance:  703.48  on 1691  degrees of freedom
## AIC: 727.48
##
## Number of Fisher Scoring iterations: 6
```

```r
chile_null <- glm(vote ~ 1, data=Chile_subset, family="binomial")

# Test for overall significance of the model
anova(chile_null, chile_3, test = "LRT")
```

```
## Analysis of Deviance Table
```

```
##
## Model 1: vote ~ 1
## Model 2: vote ~ region + population + sex + age + education + income +
##     statusquo
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1702    2360.29
## 2      1691     703.48 11   1656.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We received a p-value of basically 0 from our likelihood ratio test, meaning our model with all variables is statistically significant.

b.

```r
# realized I could've used this function for question 1
step(chile_3, trace=F)
```

```
##
## Call:  glm(formula = vote ~ sex + education + statusquo, family = "binomial",
##     data = Chile_subset)
##
## Coefficients:
## (Intercept)         sexM  educationPS   educationS    statusquo
##      1.0153      -0.5742      -1.1074      -0.6828       3.1689
##
## Degrees of Freedom: 1702 Total (i.e. Null);  1698 Residual
## Null Deviance:       2360
## Residual Deviance: 708.2     AIC: 718.2
```

We ended up dropping region, income, age, and population.

```r
chile_4 <- step(chile_3, trace=F)
summary(chile_4)
```

```
##
## Call:
## glm(formula = vote ~ sex + education + statusquo, family = "binomial",
##     data = Chile_subset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2553  -0.2845  -0.1297   0.2009   2.9614
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0153     0.1890   5.373 7.75e-08 ***
## sexM         -0.5742     0.2022  -2.840 0.004518 **
## educationPS  -1.1074     0.2914  -3.800 0.000145 ***
## educationS   -0.6828     0.2217  -3.079 0.002077 **
## statusquo     3.1689     0.1448  21.886  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2360.29  on 1702  degrees of freedom
## Residual deviance:  708.24  on 1698  degrees of freedom
## AIC: 718.24
##
## Number of Fisher Scoring iterations: 6
```

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{male,i} + \beta_3 x_{PS,i} + \beta_3 x_{S,i}, Y_i \sim_{ind} bin(1, \pi)$$

Where Y_i is vote (Y = 1, N = o) and $x_{1,i}$ is statusquo, $x_{male,i}$ is the dummy variable for sex (male = 1, female = 0), $\beta_3 x_{PS,i}$ is the dummy variable for when education = PS, $\beta_3 x_{S,i}$ is the dummy variable for when education = S. $\pi = P(Y_i = Y | x_{1,i}, x_{male,i}, x_{PS,i}, x_{S,i})$.

$$log(\frac{\pi}{1-\pi}) = 1.0153 + 3.1689 x_{1,i} - 0.5742 x_{male,i} - 1.1074 x_{PS,i} - 0.06828 x_{S,i}$$

c.

educationPS and statusquo are the most statistic predictors.

Interpretations:

Odds:

EducationPS - We expect the odds of someone to vote Yes with a Post-secondary education to be $e^{-1.1074}$ times lower than that of people with a Primary education, holding all else constant.

statusquo - We expect the odds of voting yes to multiply by $e^{3.1689}$ for every one unit increase in the scale of support for the status-quo, holding all else constant.

Simple:

EducationPS - We expect the probability of voting yes to be less for someone with post-secondary education than someone with Primary education, holding all else constant.

statusquo - We expect the probability of voting yes to increase as statusquo increases, holding all else constant.

d.

```r
summary(Chile_subset)
```

```
##  region     population      sex          age         education      income
##  C :374   Min.   :  3750   F:814   Min.   :18.00   P :671    Min.   :  2500
##  M : 54   1st Qu.: 25000   M:889   1st Qu.:25.00   PS:343    1st Qu.: 15000
##  N :230   Median :175000           Median :36.00   S :689    Median : 15000
##  S :476   Mean   :150716           Mean   :38.06             Mean   : 36838
##  SA:569   3rd Qu.:250000           3rd Qu.:49.00             3rd Qu.: 35000
##           Max.   :250000           Max.   :70.00             Max.   :200000
```

```
##    statusquo        vote
##  Min.   :-1.72594   A:  0
##  1st Qu.:-1.09671   N:867
##  Median :-0.18511   U:  0
##  Mean   :-0.00467   Y:836
##  3rd Qu.: 1.16602
##  Max.   : 1.71355
```

```
prob <- predict(chile_4, newdata=data.frame(statusquo=-0.18511, sex="F", education = "S"),
        type="response")
prob
```

```
##         1
## 0.4368183
```

```
odds <- prob/(1-prob)
odds
```

```
##         1
## 0.7756257
```

Odds of yes: 0.776

probability of yes: 0.427

e.

```
chile_prob <- predict(chile_4, type='response')
chile_predict <- ifelse(chile_prob > 0.50, "Y","N")
mean(chile_predict == Chile_subset$vote)
```

```
## [1] 0.9283617
```

```
chile_prob <- predict(chile_3, type='response')
chile_predict <- ifelse(chile_prob > 0.50, "Y","N")
mean(chile_predict == Chile_subset$vote)
```

```
## [1] 0.9271873
```

We have 92.83% sample prediction accuracy for our sample for our reduced model and 92.71% for the full model. It seems that the reduced model is better at predictions of our sample data.

Problem 4

1.

```r
women_1 <- glm(working ~ ., data = Women, family = "binomial")
summary(women_1)
```

```
##
## Call:
## glm(formula = working ~ ., family = "binomial", data = Women)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7453   0.3107   0.5371   0.7240   1.8752
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.306811   0.361278  -0.849  0.39575
## regionBC        0.371359   0.257804   1.440  0.14973
## regionOntario   0.099649   0.167199   0.596  0.55118
## regionPrairies  0.071285   0.169580   0.420  0.67422
## regionQuebec   -0.549447   0.189921  -2.893  0.00382 **
## kids0004Yes    -0.994696   0.132940  -7.482  7.3e-14 ***
## kids0509Yes    -0.389064   0.119542  -3.255  0.00114 **
## kids1014Yes    -0.087132   0.153963  -0.566  0.57145
## familyIncome   -0.012601   0.004159  -3.030  0.00245 **
## education       0.216728   0.025528   8.490  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1988.1  on 1934  degrees of freedom
## Residual deviance: 1810.1  on 1925  degrees of freedom
## AIC: 1830.1
##
## Number of Fisher Scoring iterations: 5
```

```r
women_null <- glm(working ~ 1, data=Women, family="binomial")

# Test for overall significance of the model
anova(women_null, women_1, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: working ~ 1
## Model 2: working ~ region + kids0004 + kids0509 + kids1014 + familyIncome +
##     education
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1934     1988.1
## 2      1925     1810.1  9   177.96 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Given a p-value of basically 0 from the likelihood ratio test, we have evidence that our mode is statistically significant.

2.

Kids0509Yes and education are the two most significant predictors.

Interpretation:

Odds:

Kids0509Yes - We expect the odds of a women being in the labor force with kids from 5 to 9 years old to be $e^{-0.389064}$ time less than women without kids from 5 to 9 years old, holding all else constant.

Education - For every one year increase in the number of years of education, we expect the odds of a woman being in the labor force to multiply by $e^{0.2167}$, holding all else constant.

Simple:

Kids0509Yes - We expect the probability of a woman being in the labor force with kids from 5 to 9 years old to be less than that of a woman without kids from 5 to 9 years old, holding all else constant.

Education - We expect the probability of a woman being in the labor force to increase as the number of years of her education increase, holding all else constant.

3.

```
women_prob <- predict(women_1, type='response')
women_predict <- ifelse(women_prob > 0.50, TRUE,FALSE)
mean(women_predict == Women$working)
```

```
## [1] 0.7912145
```

Our model has a sample accuracy of 79.12%