Assignment 1

630516am and 590049bm

15/3/2021

Assignment 1

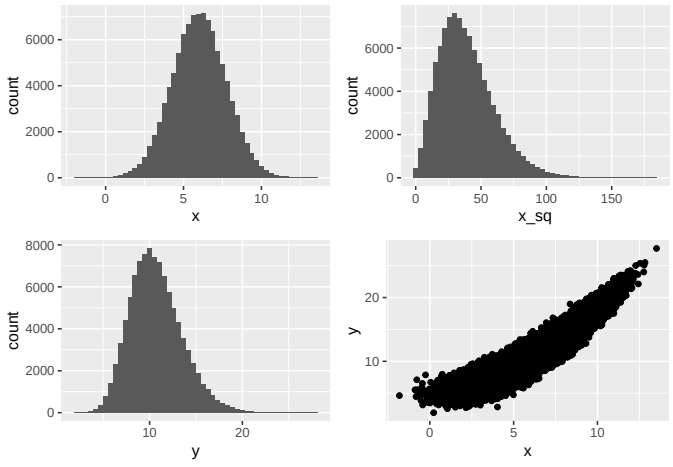
At the beginning of the code,set the random seed to 810 using set.seed(). Failure to do so will be penalised.

```
set.seed(810)
```

a. Simulate 100,000 observations from the DGP

```
x <- rnorm(100000, mean = 6, sd = sqrt(3))
alpha <- 5
e <- rnorm(100000, 0, 1)
y <- alpha + 0.3*x + 0.1*x^2 + e
dataset <- data.frame(x = x, x_sq = x^2, y = y)
```

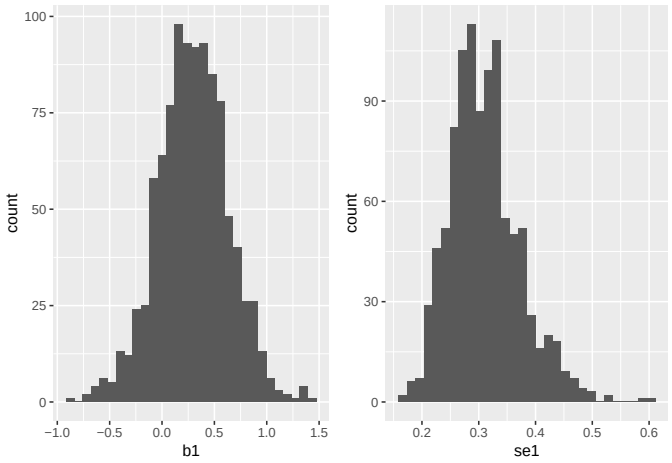
```
p1 <- dataset %>%
    ggplot(aes(x = x)) + geom_histogram(bins = 50)
p2 <- dataset %>%
    ggplot(aes(x = x_sq)) + geom_histogram(bins = 50)
p3 <- dataset %>%
    ggplot(aes(x = y)) + geom_histogram(bins = 50)
p4 <- dataset %>%
    ggplot(aes(x = x, y = y)) + geom_point()
cowplot::plot_grid(p1, p2, p3, p4, nrow = 2, ncol = 2)
```



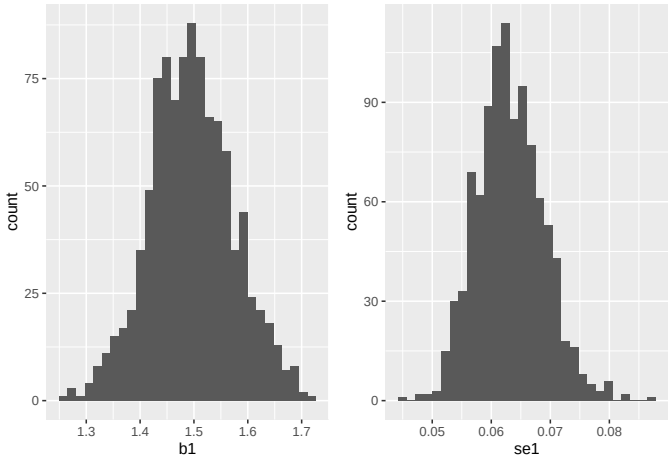
b. Break the data into 1000 datasets of 100 observations sequentially (i.e. dataset 1 comprises observations 1-100 from your simulation, dataset 2 comprises observations 101-200 and so on

```
hundred <- dataset %>%
   group_by((row_number()-1) %/% (n()/1000)) %>%
   nest %>% pull(data)
```

```
beta1 <- map_dbl(hundred, ~ lm(data = .x, formula = y ~ x + x_sq) %>%
    .$coefficients %>%
    .[2])
se1 <- map_dbl(hundred, ~ lm(data = .x, formula = y ~ x + x_sq) %>%
                  summary() %>%
                  .$coefficients %>%
                  .[2,2])
p1_een <- data.frame(b1 = beta1, se1 = se1) %>%
    ggplot(aes(x = b1)) + geom_histogram()
p2_een <- data.frame(b1 = beta1, se1 = se1) %>%
    ggplot(aes(x = se1)) + geom_histogram()
cowplot::plot_grid(p1_een, p2_een)
```




```
beta_omv <- map_dbl(hundred, ~ lm(data = .x, formula = y ~ x) %>%
    .$coefficients %>%
    .[2])
se_omv <- map_dbl(hundred, ~ lm(data = .x, formula = y ~ x) %>%
                  summary() %>%
                  .$coefficients %>%
                  .[2,2])
p1_omv <- data.frame(b1 = beta_omv, se1 = se_omv) %>%
    ggplot(aes(x = b1)) + geom_histogram()
p2_omv <- data.frame(b1 = beta_omv, se1 = se_omv) %>%
    ggplot(aes(x = se1)) + geom_histogram()
cowplot::plot_grid(p1_omv, p2_omv)
```



We see that the coefficients from the fully specified model (without omitted variable biased) are normally distributed around the true coefficient value from the DGP, whereas the coefficient in the wrongly specified model is noisy distributed around a wrong value.

c. Imagine we had instead generated $X_i = c$ for all x_i and tried to perform our simulation above. This would fail because we would violate one of the necessary conditions for computing the least squares estimator. Which one, and how?

We would violate assumption 1:

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 > 0$$

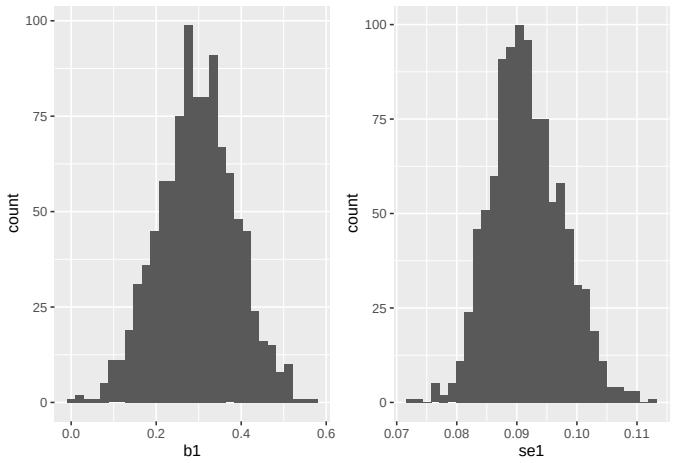
as the sum of the square of the difference between Xi and X bar would be zero if $X_i = c$. Xi = c would also mean that X is not full rank and therefore not invertible. Hence, we could not calculate the least squares estimator.

d. Simulate another dataset as in a), but with 1,000,000 observations. Repeat part b), but now using 1000 observations per model. Only fit the equation $Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \epsilon$ this time. Plot the histograms of the estimate of β_1 and $se(\beta_1)$ and compare them to the estimates from the same model in (b). What happens to the distribution of the coefficient estimates and standard errors when we increase the sample size per model?

```
x2 <- rnorm(1000000, mean = 6, sd = sqrt(3))
alpha2 <- 5
e2 <- rnorm(1000000, 0, 1)
y2 <- alpha2 + 0.3*x2 + 0.1*x2^2 + e2
dataset2 <- data.frame(x = x2, x_sq = x2^2, y = y2)
```

```
thousand <- dataset2 %>%
   group_by((row_number()-1) %/% (n()/1000)) %>%
   nest %>% pull(data)
```

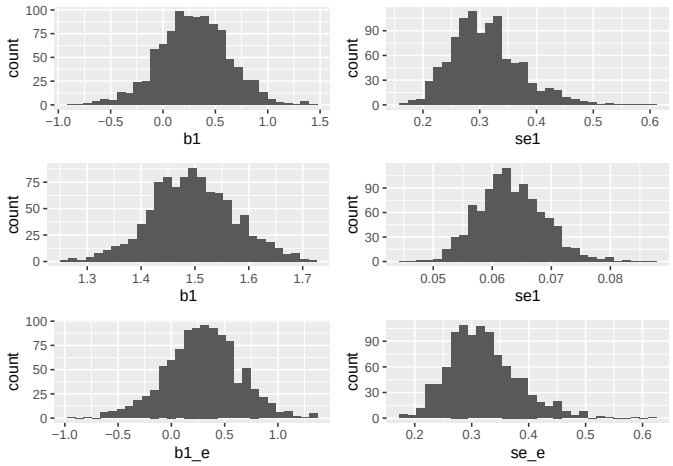
```
beta_ef <- map_dbl(thousand, ~ lm(data = .x, formula = y ~ x + x_sq) %>%
    .$coefficients %>%
    .[2])
se_ef <- map_dbl(thousand, ~ lm(data = .x, formula = y ~ x + x_sq) %>%
                  summary() %>%
                  .$coefficients %>%
                  .[2,2])
p1_ef <- data.frame(b1 = beta_ef, se1 = se_ef) %>%
    ggplot(aes(x = b1)) + geom_histogram()
p2_ef <- data.frame(b1 = beta_ef, se1 = se_ef) %>%
    ggplot(aes(x = se1)) + geom_histogram()
cowplot::plot_grid(p1_ef, p2_ef)
```



Comparing this histogram to thr histogram from b), we can see that the mean stays approximately the same. However, the spread and therefore the variance of the estimate from d) is smaller than from the estimate of b). This makes sense, as with an increasing number of observations the variance of the estimate decreases. The same can be observed with the standard error of the estimate. The mean stays approximately the same but the variance of the standard error decreases.

e. Now, generate a new variable C_i that is correlated with X_i. Do this by creating a vector of observations drawn from N(1,2) and adding them to X_i. Add C_i the dataset from the previous questions (i.e with 100,000 observations total).

```
c <- rnorm(100000, mean = 1, sd = sqrt(2)) + x
dataset <- data.frame(x = x, x_sq = x^2, y = y, c = c)
hundred <- dataset %>%
   group_by((row_number()-1) %/% (n()/1000)) %>%
   nest %>% pull(data)
beta_twee <- map_dbl(hundred, ~ lm(data = .x, formula = y ~ x + x_sq + c) %>%
    .$coefficients %>%
    .[2])
se_twee <- map_dbl(hundred, ~ lm(data = .x, formula = y ~ x + x_sq + c) %>%
                  summary() %>%
                  .$coefficients %>%
                  .[2,2])
p1_twee <- data.frame(b1_e = beta_twee,
           se_e = se_twee) %>%
    ggplot(aes(x = b1_e)) + geom_histogram()
p2_twee <- data.frame(b1_e = beta_twee,
           se_e = se_twee) %>%
    ggplot(aes(x = se_e)) + geom_histogram()
cowplot::plot_grid(p1_een, p2_een,
                   p1_omv, p2_omv,
        p1_twee, p2_twee,
        nrow = 3, ncol = 2)
```



Re-run the regressions as $Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 C_i + \epsilon_i$ and plot the distribution of these coefficient estimates and standard errors for β_1 next to the coefficient estimates from the second part. What happens to the coefficient estimates, and why?

While the first regression has a different mean, the estimators of regression 2 and 3 have approximately the same mean and distribution. Therefore, the inclusion of c does not add any additional information to our estimation and hence is a redundant variable while including the square of x changes the estimation of beta 1, indicating that it is a non-linear relationship! We also can observe this with the standard errors of the estimate. The standard error of the second and third estimator has approximately the same mean and distribution, indicating that c is a redundant variable.

The coefficients are not impacted, because there is no omitted variable bias. There is no relationship between Y and C, hence, the OLS coefficient for β_1 is still unbiased. We do pay a small penalty in terms of efficiency for adding an unnecessary variable, but with $N = 1000$, this is negligable.

Question 2

You are a labor economist trying to estimate the gender wage gap within occupations for women with children - that is, the effect of gender on wages given occupational choice. You have access to a dataset containing a set of wages,W_i, a gender dummy D_i, a set of occupational dummiesO_{ij} and hours spent on childcare C_i for a sample of men and women with children. Assume that there is a positive covariance between each of your regressors, some covariance between each of the regressors and wages, and that gender at least partially determines occupational choice and hours spent on childcare.

a. Derive the expected value of the least-squares estimator for the coefficient on the gender dummy without controlling for either occupation or hours spent on childcare

We suppose that the DGP is:

$$W_i = \alpha D_i + \beta O_i + \delta C_i + \epsilon_i$$

We have that $\alpha_{OMV} = (D^T D)^{-1} D^T W$ and

$$\mathbb{E}[\alpha_{OMV}] = \mathbb{E}[(D^T D)^{-1} D^T W] =$$

$$\mathbb{E}[(D^T D)^{-1} D^T (\alpha D + \beta O + \delta C + \epsilon)]$$

which simplifies to:

$$\alpha + \beta \cdot \mathbb{E}[(D^T D)^{-1} D^T O] + \delta \cdot \mathbb{E}[(D^T D)^{-1} D^T C] =$$

$$\alpha + \beta P_o + \delta P_c$$

using the terminology in Heij.

The variance of α_{OMV} is:

$$\text{Var}(\alpha_{OMV}) = (\mathbb{E}[\hat{\alpha}_{OMV} - \mathbb{E}[\hat{\alpha}_{OMV}]])\,(\mathbb{E}[\hat{\alpha}_{OMV} - \mathbb{E}[\hat{\alpha}_{OMV}]])^T =$$

$$(\mathbb{E}[(D^TD)^{-1}D^T\epsilon])(\mathbb{E}[(D^TD)^{-1}D^T\epsilon])^T = \sigma^2(D^TD)^{-1}$$

The variance of the true OLS estimator is:

$$\text{Var}(\alpha_{OLS}) = \sigma^2(\tilde{D}^T\tilde{D})^{-1}$$

where $\tilde{D}$ is obtained by applying the Frish-Lovell-Waugh theorem to partial out the effects of C and O: $\tilde{D} = (I - C(C^TC)^{-1}C^T)(I - O(O^TO)^{-1}O^T)$. (This is equal to $(I - Proj(C))(I - Proj(O))$) (if this is what the question means with "in terms of the projection matrices").

Then, using the result by Heij on p. 144 that the covariance between the estimators is zero, we can see that the difference between the variances boils down to:

$$\text{Var}(\alpha_{OLS}) - \text{Var}(\alpha_{OMV}) = \sigma^2(\tilde{D}^T\tilde{D})^{-1} - \sigma^2(D^TD)^{-1}$$

Now, since $\tilde{D}^T\tilde{D}$ and D^TD are scalars, we realize that the OLS variance is larger than the OMV variance if:

$$\text{Var}(D) > \text{Var}(\tilde{D})$$

As $\tilde{D}$ are the residuals after having regressed D on C and O, we know that the variation in $\tilde{D}$ is always weakly smaller than the variance in the pure D. The only case in which they are equal is when D and C, O are perfectly orthogonal, i.e. if there is no relationship. In all other cases, the variance of $\hat{\alpha}_{OMV} < \hat{\alpha}_{OLS}$.

b. A friend who has taken an undergraduate econometrics course suggests including both the occupational dummies and hours spent on childcare as control variables, to remove omitted variable bias. Now imagine we control for both of these in our least-squares regression. Derive the coefficient estimate and the variance of the estimator.

We estimate the model:

$$W = X\beta + \epsilon$$

where $X = [DOC]$ and $\beta = (\alpha, \beta, \delta)^T$. The OLS estimator is then equal to:

$$\beta = \begin{pmatrix}\alpha\\ \beta\\ \delta\end{pmatrix} = (X^TX)^{-1}X^TW = (X^TX)^{-1}X^T(X\beta) + (X^TX)^{-1}X^T\epsilon =$$

$$\beta + (X^TX)^{-1}X^T\epsilon$$

The expected value of the estimator is then:

$$\mathbb{E}[\beta] = \beta$$

because $\mathbb{E}[\epsilon] = 0$. The variance of this estimator is, according to the OLS-formula (which we coincedentally have used in the previous question) and A1-A6:

$$\text{Var}(\beta) = \sigma^2(X^TX)^{-1}$$

c. Think about what you are trying to estimate. Why would including hours spent on childcare as a control be incorrect, despite your result above? To try to sharpen your thinking you might want to draw a causal diagram for the data-generating process (if you do not know what these are, see https://mixtape.scunning.com/dag.html - you do not need to worry about the formalisms), thinking about how each variable determines each other.

Wages is an outcome variable, which consist partially of hours people have worked (at a paid job). Hours spent on childcare is another way you can use your time, hence, is also an outcome variable. Hence, part of the effect of gender on wages is happening through the effect of working on childcare. Hence, conditioning on childcare would imply 'conditioning on the outcome', and thus downwardly bias your coefficient of interest (i.e. gender). Therefore, it would generally not be a good idea to control for childcare in estimating the effect of gender on wages.

Question 3

a. Derive the log-likelihood function for the service life of n machines.

$$L(\alpha, x) = \Pi_{i=1}^{N} f(x_i) = \alpha^n \exp(-\alpha \sum_{i=1}^{N} x_i)$$

Taking the logs on both sides and applying the corresponding rules gives:

$$\log L(\alpha, x) = \sum_{i=1}^{N} f(x_i) = n\ln(\alpha) - \alpha \sum_{i=1}^{N} x_i$$

b. Derive $\hat{\alpha}$ the maximum likelihood estimator for α. Please check the second-order condition to ensure your result indeed maximizes the log-likelihood function.

Taking the first derivative and setting it equal to zero gives:

$$\alpha_{MLE} = \frac{n}{\sum_{i=1}^{N} x_i}$$

The second derivative of the log-likelihood function is:

$$\frac{\partial \log L}{\alpha} = -\frac{n}{\alpha^2}$$

which is always negative (given $\alpha \neq 0$) hence, a maximum is attained.

c. Derive the log-likelihood function (Weibull)

$$L(\beta, \gamma, x) = \Pi_{i=1}^{N} f(x_i) = \Pi_{i=1}^{N} \frac{\beta}{\lambda} \left(\frac{x_i}{\lambda}\right)^{\beta-1} \exp\{\left(-\frac{x_i}{\lambda}\right)^{\beta}\}$$

The log-likelihood is then obtained by taking the log on both sides:

$$n \log \beta - n \log \lambda + (\beta - 1) \sum_{i=1}^{N} \log x_i - (\beta - 1) n \log \lambda - \sum_{i=1}^{N} \left(\frac{x_i}{\lambda}\right)^{\beta}$$

d. The researcher further assumes that β is known but λ is unknown. Derive $\hat{\lambda}$, the maximum likelihood estimator for λ. You do not need to verify the second-order condition in this question.\

	Model 1	Model 2
(Intercept)	3258.239***	3442.130***
	(56.012)	(81.213)
age	6.401***	−2.273
	(2.009)	(2.308)
smoker	−237.315***	−185.032***
	(27.431)	(27.936)
alcohol	−33.934	−17.815
	(97.407)	(96.208)
drinks	−12.546	−7.230
	(19.427)	(19.187)
unmarried		−244.106***
		(28.539)
educ		7.276
		(5.593)
N	3000	3000
Adj. R2	0.03	0.06

* p < 0.1, ** p < 0.05, *** p < 0.01

Taking the first derivative of the log-likelihood with respect to λ and setting it to zero, gives us:

$$-\frac{n}{\lambda} + \beta \sum_{i=1}^{N} \left(\frac{x_i^{\beta}}{\lambda^{\beta+1}} \right) - n(\beta - 1) = 0$$

Reaaranging for λ, gives us the following maximum likelihood estimate for λ:

$$\hat{\lambda} = \sum_{i=1}^{N} \left(\frac{x_i}{\sqrt[\beta]{n}} \right)$$

e. Which distribution specification do you prefer and why?

The exponential distribution is a special case of the Weibull distribution with $\beta = 1$. Having only one unknown parameter, the exponential distribution is easier to use for the maximum likelihood estimation. The Weibull distribution on the other side is much more complex as it is not possible to solve the maximum likelihood estimation analytically but only numerically when we have to unknown parameters. Hence, if the memory ability of the Weibull distribution is not necessary, it is more convenient to use the exponential distribution.

Question 4

a. Estimate the two following models. Interpret the results of the two models and also compare the results.

```
dataas1 <- readr::read_csv("./DataAS1.csv")
```

```
model1 <- lm(data = dataas1, birthweight ~ age + smoker + alcohol + drinks)
model2 <- update(model1, . ~ . + unmarried + educ)
modelsummary(list(model1, model2),
             stars = c("*" = 0.10, "**" = 0.05, "***" = 0.01),
             gof_map = gm
             )
```

From the first model, the variables age and smoker are significant at a p value of 0.1 The age has a positive effect on the birth weight, whereas the mother being a smoker has a negative effect. In this second model,

the variable age is not significant anymore (and also negative), while smoker is still significant at a p value of 0.1. However, in comparison to the first model, its effect is smaller (although still very high). Of the added variables, the mother being unmarried is a significant variable and has a higher (negative) impact on the birth weight than the mother being a smoker. Education, on the other hand, has no significant effect. As we added two new independent variables, it is only natural that R^2 increases. Nonetheless, this does not mean that the new variables add to the explanation of the independent variable.

b. Test the null hypothesis H0: $\beta_{unmarried} = \beta_{educ} = 0$. Please include the name of the test, the value of the test statistics, the p-value and the conclusion in your answer.

The test is called the F-test. We first comptue it manually and then corrobate it, executing it using the `car` package:

```
RSS1 <- t(resid(model1))%*%resid(model1)
RSS1
```

```
##             [,1]
## [1,] 1017772843
```

```
RSS2 <- t(resid(model2))%*%resid(model2)
RSS2
```

```
##           [,1]
## [1,] 991102509
```

```
# ------------------------------------------------------------------------
# --- joint significance of minority and gender
# compute F-test: H0: b(unmarried)=b(educ)=0
# model 2 (model2): model under H0
g = 2 # number of restrictions
k = 6
n = nrow(dataas1) # number of observations
Ftest = ((RSS1-RSS2)/g)/(RSS1/(n-k))
Ftest
```

```
##          [,1]
## [1,] 39.22829
```

```
1-pf(Ftest, g, n-k) #p-value for F(g,n-k)
```

```
##      [,1]
## [1,]    0
```

```
# H0 is rejected at 5 % significance level
```

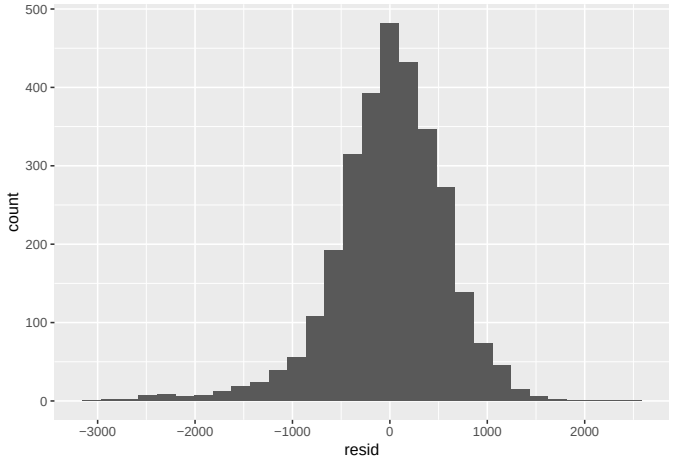
```
linearHypothesis(model2, c("unmarried=0", "educ=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## unmarried = 0
## educ = 0
##
## Model 1: restricted model
## Model 2: birthweight ~ age + smoker + alcohol + drinks + unmarried + educ
##
##   Res.Df        RSS Df Sum of Sq     F    Pr(>F)
## 1   2995 1017772843
## 2   2993  991102509  2  26670334 40.27 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. Are the residuals of model 2 normally distributed? Provide a histogram for the residuals and also

perform a formal test.

```
data.frame(resid = model2$residuals) %>%
    ggplot(aes(x = resid)) + geom_histogram()
```



To test whether the residuals are normally distributed, we can use the Shapiro Test. The null hypothesis of the Shapiro test is that the residuals are normally distributed. Hence, the alternative hypothesis is that the residuals are not normally distributed. As the p value is below 0.05, we reject the null hypothesis. Therefore, there is evidence that the residuals are not normally distributed.

```
residuals2 <- model2$residuals
shapiro.test(residuals2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals2
## W = 0.96175, p-value < 2.2e-16
```

d. There could be nonlinear relations between the dependent variable and the independent variables. Perform a test to investigate the possible non-linear relation. What is your conclusion?

The t-statistic of the non-linear model with squared fitted values gives us evidence that the model is non-linear as the null hypothesis is rejected. For the non-linear model with value^3, the F-statistics gives us evidence that the squared fitted values and values^3 are jointly non-zero (the t statistics on the other side, does not reject the null hypothesis for both estimators being zero independetly!).

```
# non-linear models
n = nobs(model2)
C <- rep(1, n)
X <- cbind(C,dataas1$age, dataas1$smoker, dataas1$alcohol, dataas1$drinks, dataas1$unmarried, dataas1$educ)
bhat <- coefficients(model2)
predval <- X%*%bhat
# (1) non-linear model with squared fitted value
predval2 <- as.matrix(predval^2)
model3 <- lm(birthweight ~ age + smoker + alcohol + drinks +
               unmarried + educ + predval2, data=dataas1)
# non-linear term significant (t-test for H0: b(predval2)=0)
# reject H0
# (2) non-linear model with fitted value^3
predval3<-as.matrix(predval^3)
model4 <- lm(birthweight ~ age + smoker + alcohol + drinks +
               unmarried + educ + predval2 + predval3, data=dataas1)
RSS3 <- t(resid(model4))%*%resid(model4)
# compute F-test: H0: b(predval2)=b(predval3)=0
```

```
k = 9
g = 2 # number of restrictions
Ftest = ((RSS2-RSS3)/g)/(RSS3/(n-k))
Ftest
```

```
##          [,1]
## [1,] 2.348508
```

```
1-pf(Ftest, g, n-k) #p-value
```

```
##           [,1]
## [1,] 0.0956877
```

```
# do not reject H0, there is evidence that predval2 and predval3 are jointly! not zero
```

e. Use the OLS method to estimate a new model (i.e. model 3) whose dependent variable is the log of birthweight and the independent variables are the same as those in model 2.

```
model_e <- lm(data = dataas1, log(birthweight) ~ age + smoker + alcohol +
                drinks + unmarried + educ)
```

Then use the maximum likelihood method to estimate model 3 again. Provide a table to show the coefficients of all regressors for two methods. Explain why the coefficients of regressors for the two methods are similar or why they are different.

```
y = log(dataas1$birthweight)
LL <- function(theta,y,X){
  n <- nrow(X)
  k <- ncol(X)
  beta <- theta[1:k]
  sigma2 <- theta[k+1]
  e <- y-X%*%beta
  logl <- -.5*n*log(2*pi)-.5*n*log(sigma2)-((t(e)%*%e)/(2*sigma2))
  return(-logl)
}
ML <- optim(c(1,1,1,1,1,1,1,1),LL,method="BFGS",hessian=T,y=y,X=X)
```

Variable	Estimate OLS	Estimate ML
(Intercept)	8.1336310	8.133631029
age	−0.0013215	−0.001321462
smoker	−0.0563769	−0.056376933
alcohol	−0.0130960	−0.013095958
drinks	−0.0020363	−0.002036251
unmarried	−0.0881805	−0.088180499
educ	0.0031101	0.003110116

The coefficients of the two regressors are very similar. Usually, the maximum likelihood estimation is used when the error term has a linear effect, i.e. when ϵ is non-normal and too far from normality. However, to find maximum likelihood estimators we have to make assumptions on the distribution of the disturbances. In this estimation we assume that ϵ is normally distributed. The same assumption is made in OLS which is why the coefficient estimations are almost the same.