Linear Regression

I. The Linear Model and Least-Squares Regression

I.1 Simple Linear Regression

The Simple Linear Regression (SLR) model is given by:

$$y = \beta_0 + \beta_1 x + e$$

In R (and most other software packages) the default null hypothesis tested is $H_0 : \beta_k = 0$ against the alternative $H_1 : \beta_k \neq 0, k = 1, 2$. Therefore, if the respective estimated p-value is small ($p-$value $< \alpha$), we reject H_0 and conclude that the respective parameter is statistically significantly different from 0.

The Simple Linear Regression (SR) assumptions of the model are that:

1. The value of y, for each value of x, is $y = \beta_0 + \beta_1 x + e$. A linear relationship.
2. $E[e] = 0$. In the residual plot, see whether the sample average is around 0.
3. $var[e] = var[y] = \sigma^2$. The errors are *homoskedastic*)
4. $cov(y_i, y_j) = cov(e_i, e_j) = 0$ (even better if the errors are independent which is a stronger condition)
5. Optional: The errors are normally distributed

```r
# Example 1: Actual measured weight (y=weight) vs. reported weight (x=repwt)
library("car")
#summary(Davis)
davis.mod <- lm(weight ~ repwt, data=Davis)
S(davis.mod)
```

```
## Call: lm(formula = weight ~ repwt, data = Davis)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.3363     3.0369   1.757   0.0806 .
## repwt         0.9278     0.0453  20.484   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 8.419 on 181 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared: 0.6986
## F-statistic: 419.6 on 1 and 181 DF,  p-value: < 2.2e-16
##     AIC     BIC
## 1303.06 1312.69
```

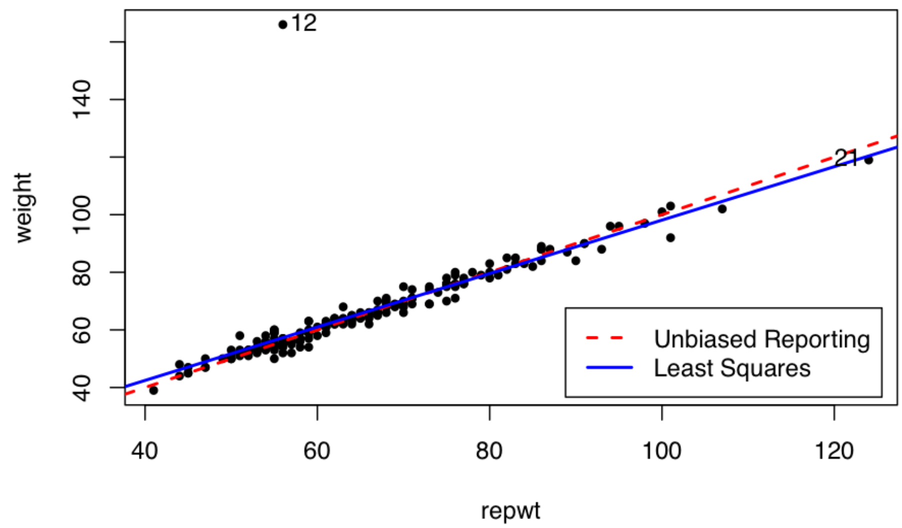
```r
# In this case we see that the y-intercept is not stat. significant but the slope is.
names(davis.mod)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "na.action"     "xlevels"       "call"          "terms"
## [13] "model"
```

```r
Confint(davis.mod)
```

	Estimate	2.5 %	97.5 %
(Intercept)	5.3362605	-0.6560394	11.328561
repwt	0.9278428	0.8384665	1.017219

```
#Note: When the interval crosses zero like in the case of the y-intercept, this suggests
#that the paramter estimate is not statistically significant.
plot(weight ~ repwt, data=Davis,pch=20)
abline(0, 1, lty="dashed", lwd=2,col="red")
abline(davis.mod, lwd=2, col="blue")
legend("bottomright", c("Unbiased Reporting", "Least Squares"),
       lty=c("dashed", "solid"), col=c("red","blue"),lwd=2, inset=0.02)
with(Davis, showLabels(repwt, weight, n=2, method="mahal"))
```



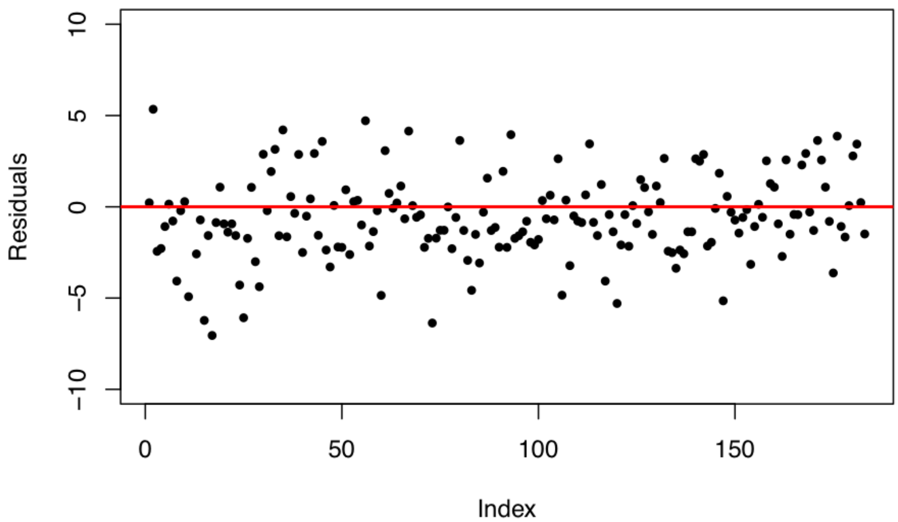

```
## [1] 12 21
```

```
# We can now try to test the SLR assumtions 1-6:

# SLR1: From the scatterplot, the linear relationship does appear to hold. Alos, the regression output
S(davis.mod)
```

```
## Call: lm(formula = weight ~ repwt, data = Davis)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.3363     3.0369   1.757   0.0806 .
## repwt         0.9278     0.0453  20.484   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 8.419 on 181 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared: 0.6986
```

```
## F-statistic: 419.6 on 1 and 181 DF,  p-value: < 2.2e-16
##      AIC     BIC
## 1303.06 1312.69
```

```
# SLR2: E[e] = 0
plot(davis.mod$residuals,pch=20, ylab="Residuals",ylim=c(-10,10))
abline(h=0, lwd=2, col="red")
```




```
mean(davis.mod$residuals)
```

```
## [1] -3.342068e-18
```

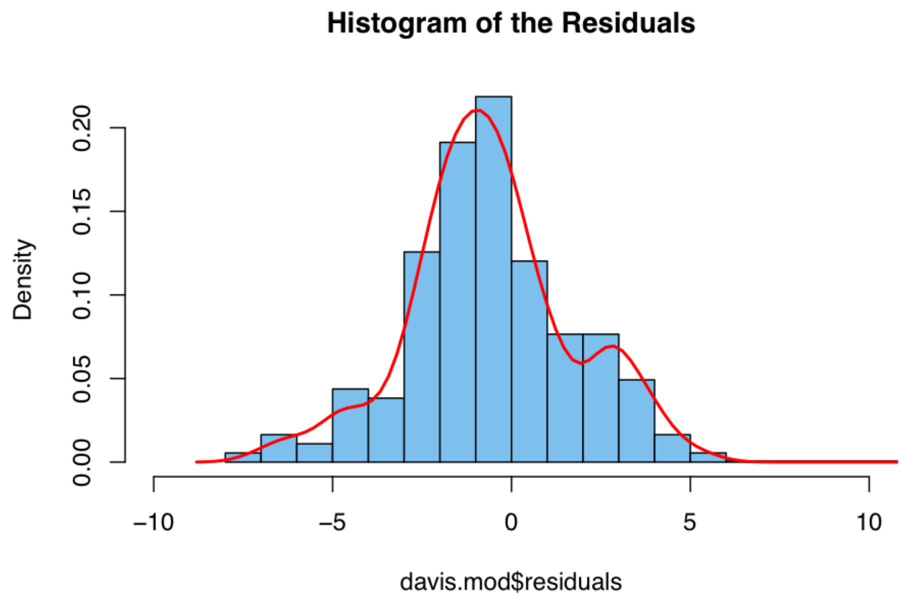
```
# SLR3: var[e] = var[y]
var(davis.mod$residuals)
```

```
## [1] 70.48366
```

```
var(Davis$weight)
```

```
## [1] 227.8593
```

```
#SLR4: cov(y_i, y_j) = cov(e_i, e_j) = 0
# This can be difficult to say

# SLR5: From the plot x takes on many different values

# SLR6: e ~ normal
hist(davis.mod$residuals,breaks ="FD",col="skyblue2", freq = FALSE, ylab = "Density",
     main = "Histogram of the Residuals",xlim=c(-10,10))
lines(density(davis.mod$residuals),lwd = 2, col ="red")
```




```
#Note: A Jarque-Bera Test is more objective but for now a visual inspection of the histogram will suffi
```

We can see that observation 12 is an *influential outlier*, i.e., can destabilize the regression fit. Therefore, we might consider removing it and re-estimating the regression fit.

```
davis.mod.2 = update(davis.mod, subset=-12)
S(davis.mod.2)
```

```
## Call: lm(formula = weight ~ repwt, data = Davis, subset = -12)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.73380    0.81479   3.355 0.000967 ***
## repwt        0.95837    0.01214  78.926  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 2.254 on 180 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared: 0.9719
## F-statistic:  6229 on 1 and 180 DF,  p-value: < 2.2e-16
##    AIC    BIC
## 816.27 825.88
```

```
# We can compare the estimates with and without the outlier:
cbind(Original=coef(davis.mod),NoCase12=coef(davis.mod.2))
```

	Original	NoCase12
(Intercept)	5.3362605	2.7338020
repwt	0.9278428	0.9583743

```
# Given that the slope did not change much, this outlier is considered a
# low-leverage point (i.e., its predicted value is near the centroid of the predictions).
```

When determining whether the linear regression is the correct model, we look at:

- t-stats and respective $p-$values of the estimated parameters

  $$SE = sb_1 = sqrt[\sum(y_i - \hat{y}_i)^2/(n-2)].sqrt[\sum(x_i - \bar{x})^2]$$

  Degrees of freedom = n-2.

  $$t = \frac{b_1}{SE}$$

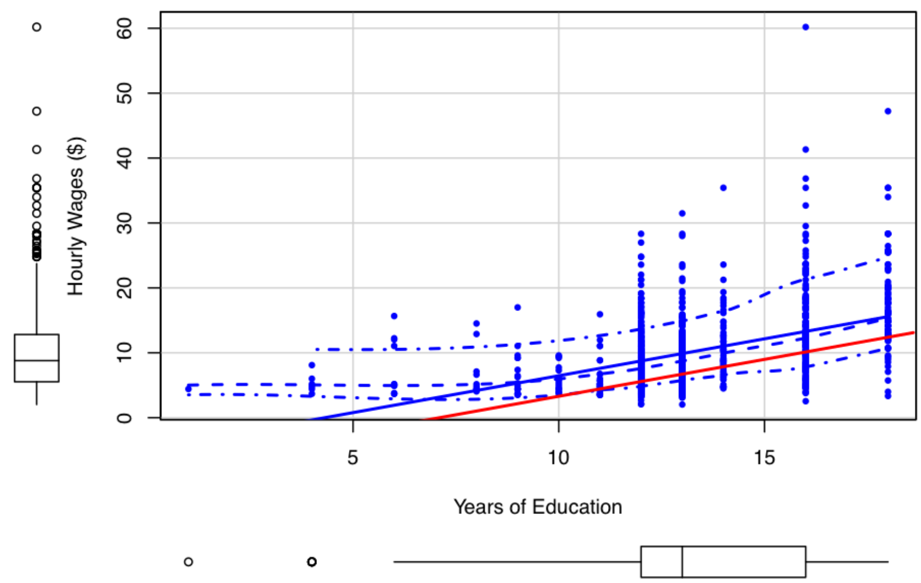
- P-value. The P_value is the probability of observing a sample statistic as extreme as the test statistic.
- Interpret results. Since the P-value (0.0242) is less than the significance level (0.05), we cannot accept the null hypothesis.
- Coefficient of Determination: R^2
- Plot of the residuals:
  - (1) randomly distributed about 0 and
  - (2) Constant Variance
- Interpretability of the estimates
- Robustness of the model estimates
- High correlation between actual and predicted values

The normality of the residuals can be helpful but is not required. We can perform a Jarque-Bera (JB) test on the residuals to determine if they are normally distributed. The JB statistic is given by:

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right),$$

where K = kurtosis and S = skeweness. For a normal distribution $K = 3$ and $S = 0$. We can show that $JB \sim \chi^2_{v=2}$, where the null hypothesis is $H_0 : JB = 0$.

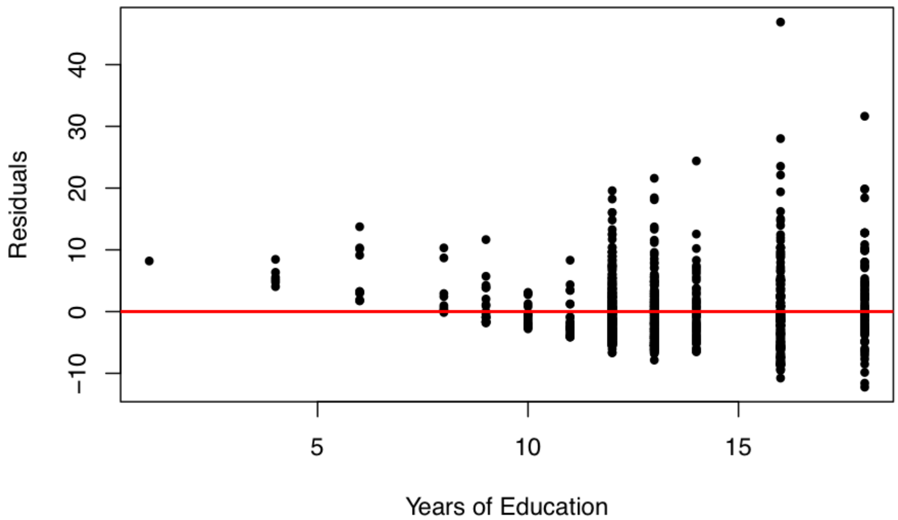
```
# Example 2: Wages (y=wage) vs. Years of Education (x=education)
library(PoEdata)
data("cps_small")
reg.mod = lm(cps_small$wage~cps_small$educ)
scatterplot(cps_small$educ, cps_small$wage, xlab="Years of Education",
            ylab="Hourly Wages ($)",pch=20)
abline(reg.mod, lwd=2, col="red")
```



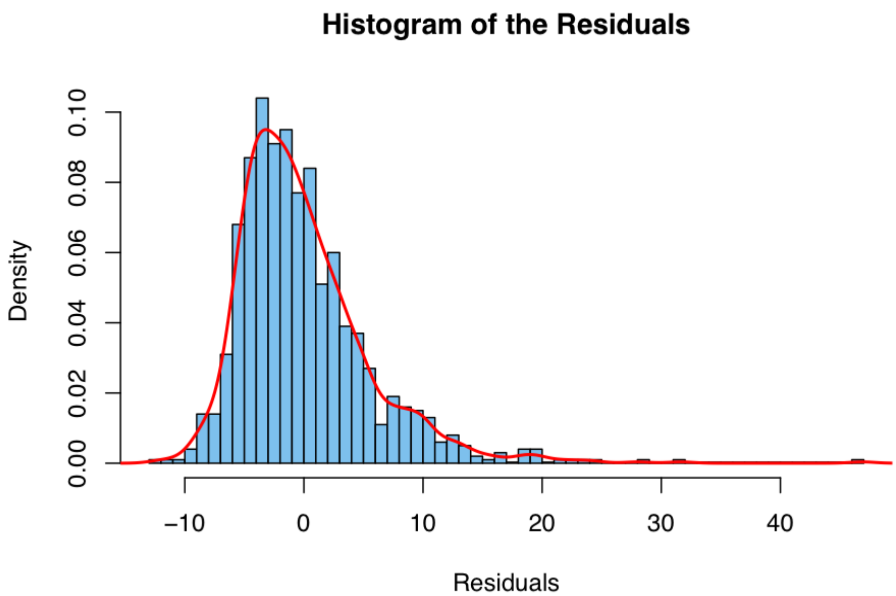

```r
S(reg.mod)
```

```
## Call: lm(formula = cps_small$wage ~ cps_small$educ)
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -4.91218    0.96679  -5.081 4.48e-07 ***
## cps_small$educ  1.13852    0.07155  15.912  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 5.582 on 998 degrees of freedom
## Multiple R-squared: 0.2024
## F-statistic: 253.2 on 1 and 998 DF,  p-value: < 2.2e-16
##     AIC     BIC
## 6280.86 6295.58
```

```r
# The linear fit does not appear to be adequate despite the stat. sig. of the parameter
# estimates.We can look at the Residuals plot to determine if the model is appropriate:
plot(cps_small$educ,reg.mod$residuals,pch=20, ylab="Residuals", xlab="Years of Education")
abline(h=0, lwd=2, col="red")
```




```r
# The plot shows that both of the residuals' conditions are not satisfied.
# Q1: Do you think a log transformation of wages might improve our model? Try it!
# Q2: Are the residuals normally distributed?
hist(reg.mod$residuals,breaks ="FD",col="skyblue2", freq = FALSE, ylab = "Density",
     xlab="Residuals",main = "Histogram of the Residuals")
lines(density(reg.mod$residuals),lwd = 2, col ="red")
```



```r
library(tseries)
jarque.bera.test(reg.mod$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  reg.mod$residuals
## X-squared = 3023.5, df = 2, p-value < 2.2e-16
```

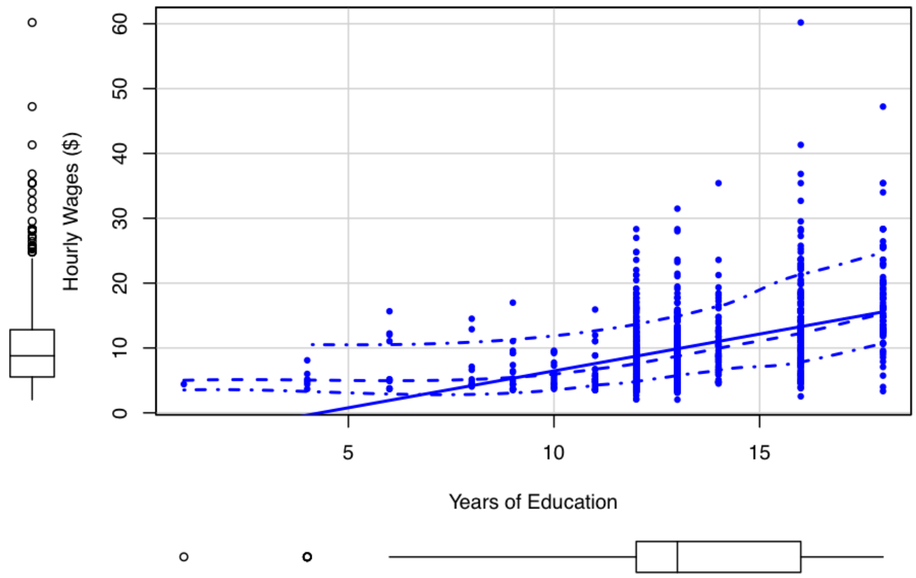
```r
# From the histogram and JB Test we can conclude that the residuals are not normally distributed.
# Another diagnostic is the Prediction Accuracy. A correlation between the actuals and predicted
# values can be used as a form of accuracy measure

actuals_preds <- data.frame(cbind(actuals=cps_small$wage, predicteds=reg.mod$fitted.values))
correlation_accuracy <- cor(actuals_preds)
print(correlation_accuracy )
```

```
##              actuals predicteds
## actuals    1.0000000  0.4498506
## predicteds 0.4498506  1.0000000
```

```r
head(actuals_preds)
```

actuals	predicteds
2.03	9.888543
2.07	8.750025
2.12	8.750025
2.54	13.304094
2.68	8.750025
3.09	9.888543

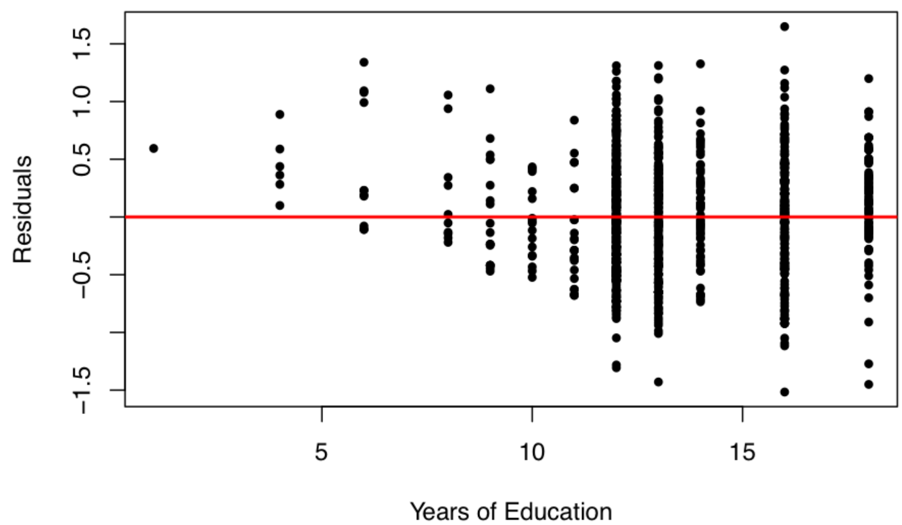
```r
# Example 4: Wages (y=wage) vs. Years of Education (x=education)
library(PoEdata)
data("cps_small")
reg.mod = lm(log(cps_small$wage)~cps_small$educ)
scatterplot(cps_small$educ, cps_small$wage, xlab="Years of Education",
            ylab="Hourly Wages ($)",pch=20)
abline(reg.mod, lwd=2, col="red")
```



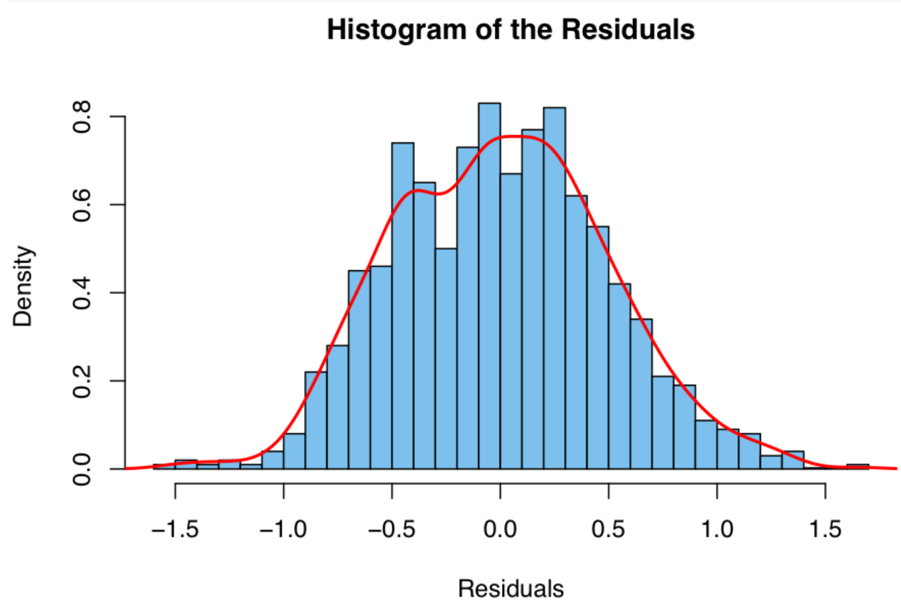

```r
S(reg.mod)
```

```
## Call: lm(formula = log(cps_small$wage) ~ cps_small$educ)
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.788374   0.084898   9.286   <2e-16 ***
## cps_small$educ 0.103761   0.006283  16.514   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 0.4902 on 998 degrees of freedom
## Multiple R-squared: 0.2146
## F-statistic: 272.7 on 1 and 998 DF,  p-value: < 2.2e-16
##     AIC     BIC
## 1415.79 1430.52
```

```r
# The linear fit does not appear to be adequate despite the stat. sig. of the parameter
# estimates.We can look at the Residuals plot to determine if the model is appropriate:
plot(cps_small$educ,reg.mod$residuals,pch=20, ylab="Residuals", xlab="Years of Education")
abline(h=0, lwd=2, col="red")
```




```r
# The plot shows that both of the residuals' conditions are not satisfied.
# Q1: Do you think a log transformation of wages might improve our model? Try it!
# Q2: Are the residuals normally distributed?
hist(reg.mod$residuals,breaks ="FD",col="skyblue2", freq = FALSE, ylab = "Density",
     xlab="Residuals",main = "Histogram of the Residuals")
lines(density(reg.mod$residuals),lwd = 2, col ="red")
```



```r
library(tseries)
jarque.bera.test(reg.mod$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  reg.mod$residuals
## X-squared = 3.4815, df = 2, p-value = 0.1754
```

```r
# From the histogram and JB Test we can conclude that the residuals are not normally distributed.
# Another diagnostic is the Prediction Accuracy. A correlation between the actuals and predicted
# values can be used as a form of accuracy measure

actuals_preds <- data.frame(cbind(actuals=cps_small$wage, predicteds=reg.mod$fitted.values))
correlation_accuracy <- cor(actuals_preds)
print(correlation_accuracy )
```

```
##              actuals predicteds
## actuals    1.0000000  0.4498506
## predicteds 0.4498506  1.0000000
```

```r
head(actuals_preds)
```

actuals	predicteds
2.03	2.137265
2.07	2.033504
2.12	2.033504
2.54	2.448547
2.68	2.033504
3.09	2.137265

I.3 Prediction vs. Confidence Intervals

Confidence interval: $\widehat{y} \pm t_{n-2} sy \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$ and $s_y = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$ The width of the CI for E(y) increases as x^* moves away from the center. We can be much more certain of our predictions at the center of the data than at the edges. Mathematically, as $(x^* - \bar{x})^2$ term increases, the margin of error of the confidence interval increases as well.

Prediction interval: $\widehat{y} \pm t_{n-2} sy \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$. If we repeat the study of obtaining a regression data set many times, each time forming a XX% prediction interval at x^*, and wait to see what the future value of y is at x^*, then roughly XX% of the prediction intervals will contain the corresponding actual value of y.

A prediction interval is similar in spirit to a confidence interval, except that the prediction interval is designed to cover a 'moving target', the random future value of y. While the confidence interval is designed to cover the 'fixed target', the average value of y, E(y), for a given x^*.

```r
# Example 7: Weekly Food Expenditure (y=food_exp in $) vs. Income (x=income in units of $100)
# Prediction vs, Confidence Intervals
library(PoEdata)
data(food)
```

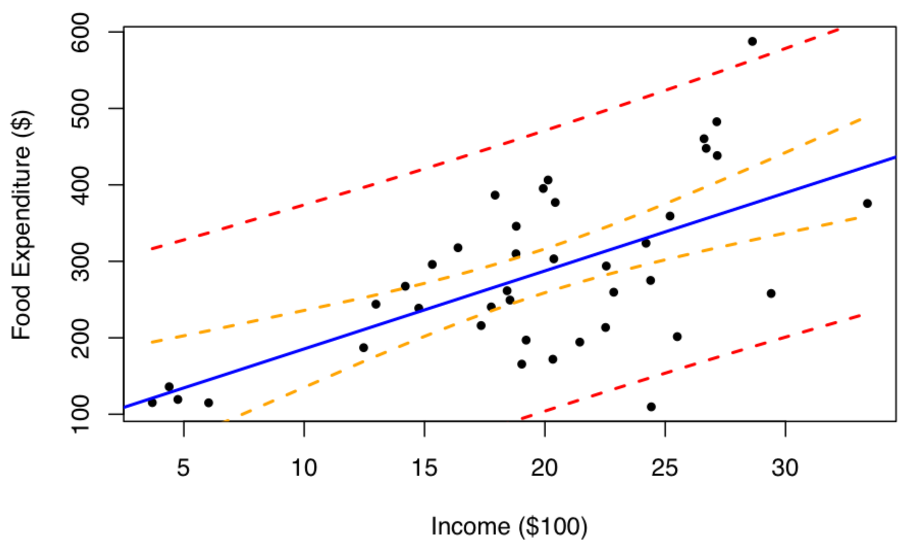
```r
attach(food)
reg.mod <- lm(food_exp~income, data=food)
predict(reg.mod, newdata=data.frame(income=20), interval = "confidence",level=0.95)
```

fit	lwr	upr
287.6089	258.9069	316.3108

```r
predict(reg.mod, newdata=data.frame(income=20), interval = "prediction",level=0.95)
```

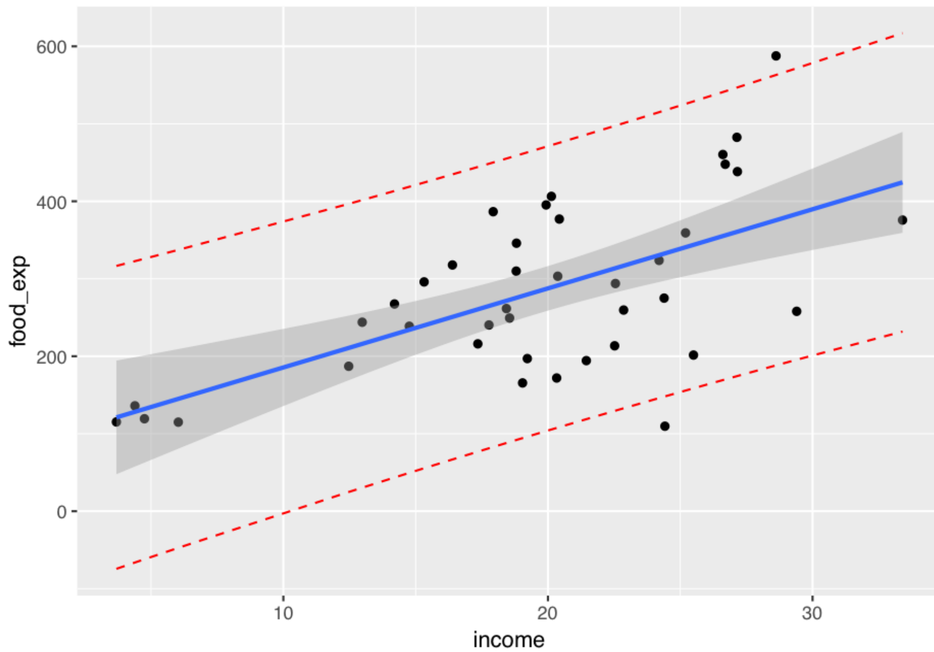
fit	lwr	upr
287.6089	104.1323	471.0854

```r
plot(income, food_exp,pch=20,xlab="Income ($100)", ylab="Food Expenditure ($)")
abline(reg.mod, col="blue",lwd=2)
pred_int = predict(reg.mod, newdata=data.frame(income), interval = "prediction",level=0.95)
conf_int = predict(reg.mod, newdata=data.frame(income), interval = "confidence",level=0.95)

# We can plot these for the entie dataset:
lines(income, conf_int[,2], col="orange", lty=2,lwd =2)
lines(income, conf_int[,3], col="orange", lty=2,lwd=2)
lines(income, pred_int[,2], col="red", lty=2,lwd =2)
lines(income, pred_int[,3], col="red", lty=2,lwd=2)
```




```r
#As expected, the prediction interval is wider than the confidence interval for the
# same level of confidence.

# A prettier version of this plot:
library(ggplot2)
```

```r
temp_var <- predict(reg.mod, interval="prediction")
new_df <- cbind(food, temp_var)
ggplot(new_df, aes(income, food_exp))+
    geom_point() +
    geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
    geom_line(aes(y=upr), color = "red", linetype = "dashed")+
    geom_smooth(method=lm, se=TRUE)
```




```r
#$$E(WAGE) = \beta_0 + \beta_1 FEMALE = \left\{ \begin{array}{ll}
#          \beta_0 + \beta_1 &  & \mbox{if $FEMALE=1$}\\
#         \beta_0  &  & \mbox{if $FEMALE=0$}.\end{array} \right.$$
```

I.4 Models with Factors (Indicator Variables)

Factor variables (also known as indicator or dummy variables) are used in regression to represent categorical variables, i.e., nonquantitative characteristics such as gender, race, drink preference, etc. The simplest example of a factor variable is a binary one, i.e., it can only take on two values (these are referred to as the *levels* of the factor). For example, we can create a binary indicator variable for gender, I, such that $I = 1$ (gender = female) and $I = 0$ (gender = male). For example, consider a regression of hourly wages (`WAGE`) on gender (`FEMALE` $= 1$(female)). In this case our model can be expressed as:

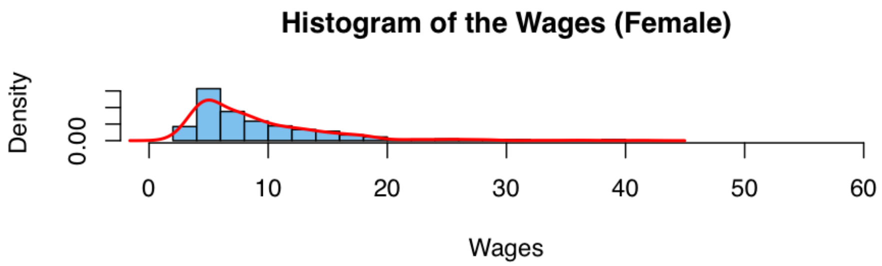
$$E(WAGE) = \beta_0 + \beta_1 FEMALE = \left\{ \begin{array}{ll} \beta_0 + \beta_1 & \text{if } FEMALE = 1 \\ \beta_0 & \text{if } FEMALE = 0. \end{array} \right.$$

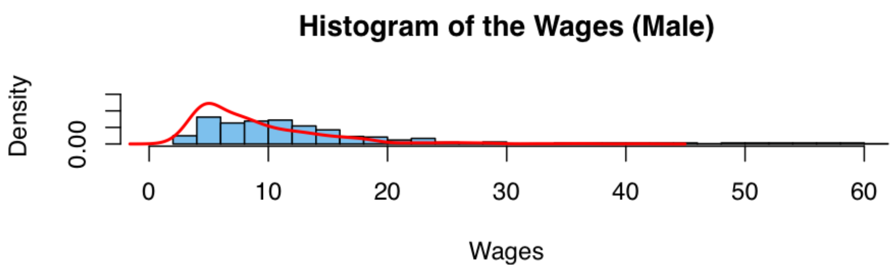
In this case, our interpretation of the model parameters is different from the previous examples since $\beta_0 = \overline{WAGE}_{Male}$ and $\beta_1 = \overline{WAGE}_{Female} - \overline{WAGE}_{Male}$.

```r
# Example 8: Wages (y=wage) vs. Gender (x=female)
library(PoEdata)
data("cps_small")
attach(cps_small)
# Note: When dealing with indicator variables it is important to check the number of
# observations in each category to see if the data are balanced or not.
S(cps_small)
```

wage	educ	exper	female	black	white	midwest	
Min. : 2.03	Min. : 1.00	Min. : 0.00	Min. :0.000	Min. :0.000	Min. :0.000	Min. :0.000	M
1st Qu.: 5.53	1st Qu.:12.00	1st Qu.:10.00	1st Qu.:0.000	1st Qu.:0.000	1st Qu.:1.000	1st Qu.:0.000	1st
Median : 8.79	Median :13.00	Median :18.00	Median :0.000	Median :0.000	Median :1.000	Median :0.000	Mec
Mean :10.21	Mean :13.29	Mean :18.78	Mean :0.494	Mean :0.088	Mean :0.912	Mean :0.237	Me
3rd Qu.:12.78	3rd Qu.:16.00	3rd Qu.:26.00	3rd Qu.:1.000	3rd Qu.:0.000	3rd Qu.:1.000	3rd Qu.:0.000	3rd
Max. :60.19	Max. :18.00	Max. :52.00	Max. :1.000	Max. :1.000	Max. :1.000	Max. :1.000	Ma

```r
# Did you find any indicator variables that are not well balanced?

wage_female = wage[which(female==1)]
wage_male = wage[which(female==0)]

# First we will look at the distribution of wages conditioned on gender
par(mfrow=c(2,1))
hist(wage_female,breaks ="FD",col="skyblue2", freq = FALSE, ylab = "Density",
     main = "Histogram of the Wages (Female)",xlab="Wages",ylim=c(0,0.17),xlim=c(0,60))
lines(density(wage_female),lwd = 2, col ="red")

hist(wage_male,breaks ="FD",col="skyblue2", freq = FALSE, ylab = "Density",
     main = "Histogram of the Wages (Male)",xlab="Wages",ylim=c(0,0.17),xlim=c(0,60))
lines(density(wage_female),lwd = 2, col ="red")
```






```r
# Now we can estimate the model and interpret the results
reg.mod = lm(wage~female)
S(reg.mod)
```

```
## Call: lm(formula = wage ~ female)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.5255     0.2715  42.455  < 2e-16 ***
## female       -2.6568     0.3862  -6.878 1.07e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard deviation: 6.107 on 998 degrees of freedom
## Multiple R-squared: 0.04526
## F-statistic: 47.31 on 1 and 998 DF,  p-value: 1.067e-11
##     AIC     BIC
## 6460.65 6475.37
```

```r
# We interpret the y-intercept as the mean hourly wage for men (beta_0 = $11.52), and
# the slope as the difference in hourly wages between women and men (beta_1 = $-2.66).
# On average, all other things being equal, women are predicted to earn $2.66/hr less
# than men.
```