STAT 8051 HW 4

Jingxiang Li

October 4, 2014

Problem 7.8

Sue fits a wls regression with all weights equal to 2. Joe fits a wls regression to the same data with all weights equal to 1. What are the differences in estimates of coefficients, standard errors, σ^2 F-tests between Sues and Joes analyses?

Solution

Let `m1` be the regression model fitted by Sue with weights all equal to 2; `m2` be the regression model fitted by Joe, where all weights equal to 1. Since they both use constant weights, the estimates of coefficients and the F-tests should be the same, but the standard errors and σ^2 are different between these two models, but it's only a scale difference.

Claim that W_1 and W_2 are the weighting matrix respectively for `m1` and `m2`, so that we have $W_1 = 2I$, $W_2 = I$, where I is the $n \times n$ identity matrix. For the coefficient estimator, we have

$$\hat{\beta} = (X'WX)^{-1}X'WY$$

Let $\hat{\beta}_1$ be the estimator obtained from `m1`, $\hat{\beta}_2$ be the estimator obtained from `m2`, so that

$$\hat{\beta}_1 = (X' \times 2I \times X)^{-1}X' \times 2I \times Y = (X'X)^{-1}X'Y = \hat{\beta}_2$$

so that the estimates of coefficients between two models are the same.

Let e_i be the residual vectors from model `mi`, RSS_i be the RSS derived from model `mi`, σ_i^2 be the σ^2 from model `mi`, where $i = 1, 2.$ and let df be the degree of freedom. we have

$$RSS_1 = e_1'W_1e_1 = 2e_1'e_1$$

$$RSS_2 = e_2'W_2e_2 = e_1'e_1$$

Note that the estimates of coefficients between two models are the same, so that $e_1 = e_2$, and $RSS_1 = 2RSS_2$, and

$$\hat{\sigma}_1^2 = \frac{RSS_1}{df} = \frac{2RSS_2}{df} = 2\hat{\sigma}_2^2$$

then we have

$$\hat{\sigma}_1 = \sqrt{2}\hat{\sigma}_2$$

For σ^2, based on the error assumption of WLS, we have $Var(e_{1i}) = \sigma_1^2/2$, $Var(e_{2i}) = \sigma_2^2$, and note that $e_{1i} = e_{2i}, \forall i$, so that

$$\sigma_1^2 = 2\sigma_2^2$$

Let F_i be the F statistic for model `mi`, where $i = 1, 2$, and let n be the sample size.

$$F_1 = \frac{(\sum 2(y_i - \bar{y})^2 - RSS_1)/(n - 1 - df)}{RSS_1/df}$$

$$F_1 = \frac{(\sum 2(y_i - \bar{y})^2 - 2RSS_2)/(n - 1 - df)}{2RSS_2/df}$$

$$F_1 = \frac{(\sum (y_i - \bar{y})^2 - RSS_2)/(n - 1 - df)}{RSS_2/df}$$

$$F_1 = F_2$$

In summary, we have

$$\hat{\beta}_1 = \hat{\beta}_2$$
$$\hat{\sigma}_1 = \sqrt{2}\hat{\sigma}_2$$
$$\sigma_1^2 = 2\sigma_2^2$$
$$F_1 = F_2$$

Q.E.D.

Problem 7.7

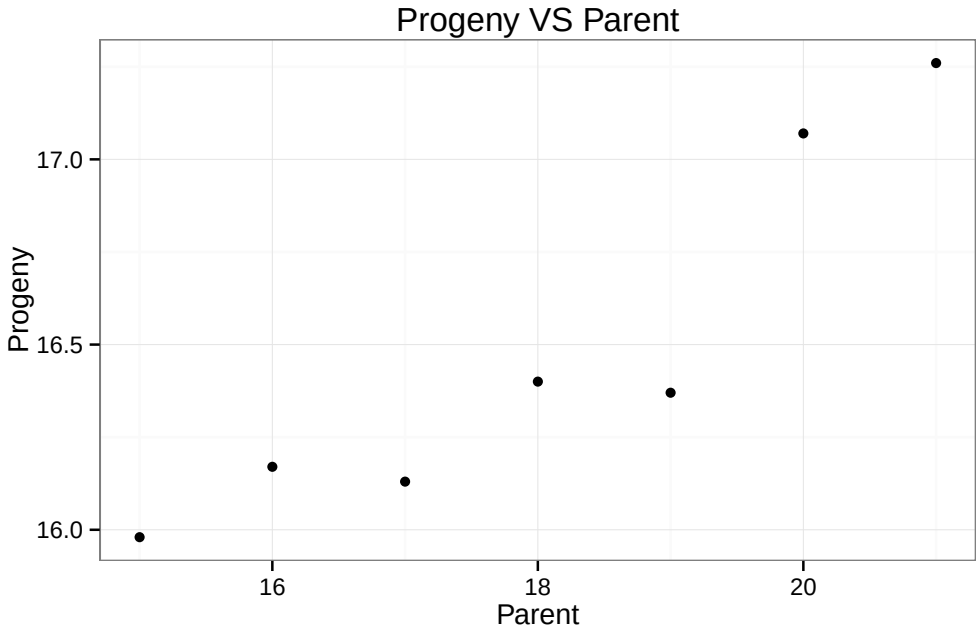
Galtons sweet peas (Data file: `galtonpeas`) Many of the ideas of regression first appeared in the work of Sir Francis Galton (18221911) on the inheritance of characteristics from one generation to the next. In Galton (1877), he discussed experiments on sweet peas. By comparing the sweet peas produced by parent plants to those produced by offspring plants, he could observe inheritance from one generation to the next. Galton categorized parent plants according to the typical diameter of the peas they produced. For seven size classes from 0.15 to 0.21 inches, he arranged for each of nine of his friends to grow 10 plants from seed in each size class; however, two of the crops were total failures. A summary of Galtons data were later published in Pearson (1930). The data file includes `Parent` diameter, `Progeny` diameter, and `SD` the standard deviation of the progeny diameters. Sample sizes are unknown but are probably large.

Problem 7.7.1

Draw the scatterplot of `Progeny` versus `Parent`.

Solution

```
require(alr4)
require(ggplot2)
data <- galtonpeas
p <- ggplot(data = data, aes(x = Parent, y = Progeny))
p <- p + geom_point()
p <- p + theme_bw() + theme(text = element_text(size = 12))
p <- p + ggtitle("Progeny VS Parent")
p
```

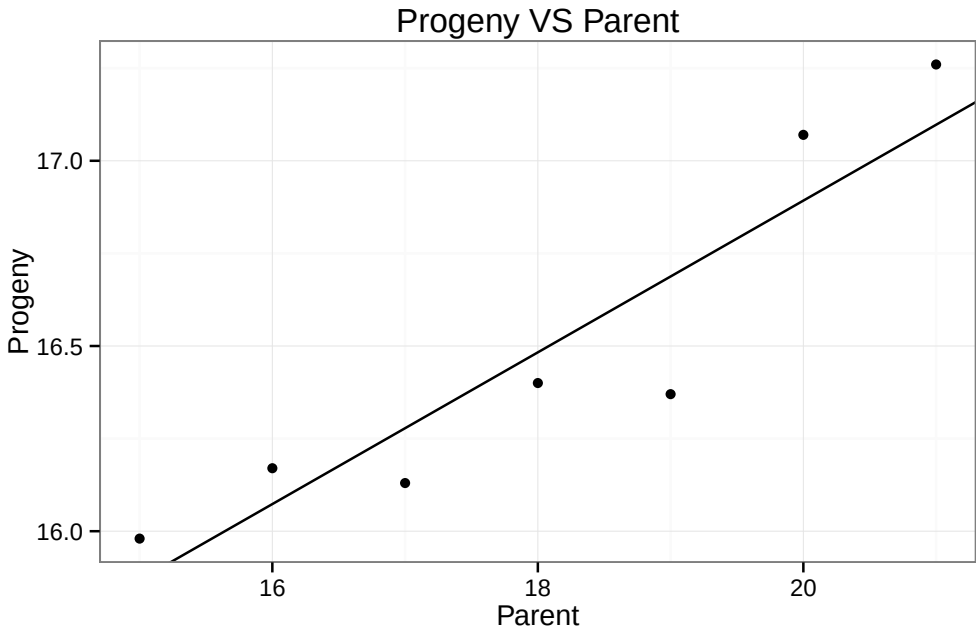


Problem 7.7.2

Assuming that the standard deviations given are population values, compute the weighted regression of `Progeny` on `Parent`. Draw the fitted mean function on your scatterplot.

Solution

```
model <- lm(Progeny ~ Parent, data = data, weights = 1 / data$SD^2)
coef <- model$coefficients

p <- p + geom_abline(intercept = coef[1], slope = coef[2])
p
```



Problem 7.7.3

Galton took the average size of all peas produced by a plant to determine the size class of the parental plant. Yet for seeds to represent that plant and produce offspring, Galton chose seeds that were as close to the overall average size as possible. Thus, for a small plant, the exceptional large seed was chosen as a representative, while larger, more robust plants were represented by relatively smaller seeds. What effects would you expect these experimental biases to have on (1) estimation of the intercept and slope and (2) estimates of error?

Solution

Following the way Galton determine the seeds that represent the parent plant, the estimation of the slope will be lower than the true nature, and thus the intercept term will increase, and the variance will increase too.

Due to the fact that larger seeds are chosen for small plant, smaller seeds are chosen for large plant, the offspring's size will be centralized. It means that for each level of `Parent`, `Progeny` are more closer to each other compared with the true nature, so that the slope coefficient will decrease. Then, since the central point for this regression model won't change too much and the slope will decrease, the intercept will increase. Then for the variance of model, since Galton choose exceptional large or small seeds to represent parent plant, the variance of `Progeny`'s size for each level of `Parent` will definitely increase.

Problem 7.8

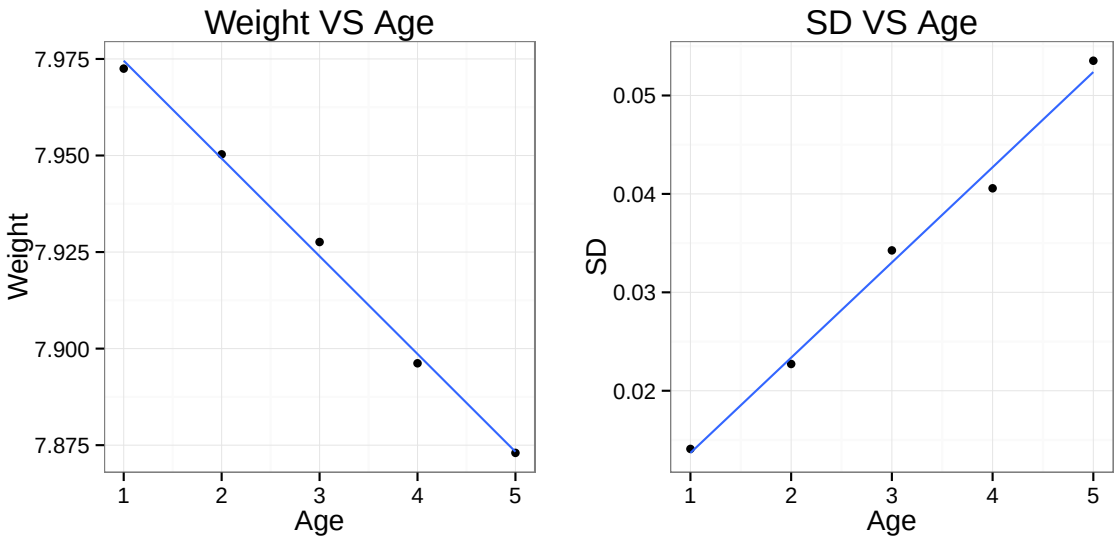
Jevonss gold coins (Data file: `jevons`) The data in this example are deduced from a diagram in Jevons (1868) and provided by Stephen M. Stigler. In a study of coinage, Jevons weighed 274 gold sovereigns that he had collected from circulation in Manchester, England. For each coin, he recorded the weight after cleaning to the nearest 0.001 g, and the date of issue. The data file includes `Age`, the age of the coin in decades, `n`, the number of coins in the age class, `Weight`, the average weight of the coins in the age class, `SD`, the standard deviation of the weights. The minimum `Min` and maximum `Max` of the weights are also given. The standard weight of a gold sovereign was 7.9876 g; the minimum legal weight was 7.9379 g.

Problem 7.8.1

Draw a scatterplot of `Weight` versus `Age`, and comment on the applicability of the usual assumptions of the linear regression model. Also draw a scatterplot of `SD` versus `Age`, and summarize the information in this plot.

Solution

```
require(gridExtra)
data <- jevons

p1 <- ggplot(data = data, aes(x = Age, y = Weight))
p1 <- p1 + geom_point()
p1 <- p1 + geom_smooth(formula = y ~ x, method = "lm", se = FALSE)
p1 <- p1 + theme_bw() + theme(text = element_text(size = 14))
p1 <- p1 + ggtitle("Weight VS Age")

p2 <- ggplot(data = data, aes(x = Age, y = SD))
p2 <- p2 + geom_point()
p2 <- p2 + geom_smooth(formula = y ~ x, method = "lm", se = FALSE)
p2 <- p2 + theme_bw() + theme(text = element_text(size = 14))
p2 <- p2 + ggtitle("SD VS Age")

grid.arrange(p1, p2, ncol=2)
```



For the regression model `Weight` versus `Age`, following the information provided by the scatterplot, the linear form is appropriate and there is no specific information shows that the variance of each case in this dataset is different. However, as the standard deviation for each point is given, we should consider WLS as our primary choice.

For the `SD` versus `Age` scatterplot, we can see that as `Age` goes on, the standard deviation of the weights increase linearly against `Age`.

Problem 7.8.2

To fit a simple linear regression model with Weight as the response, wls should be used with variance function $Var(\texttt{Weight}|\texttt{Age}) = \texttt{n}\sigma^2/\texttt{SD}^2$. Sample sizes are large enough to assume the `SD` are population values. Fit the wls model.

Solution

```
w <- data$SD^2 / data$n
m <- lm(Weight ~ Age, data = data, weight = w)
summary(m)
```

```
##
## Call:
## lm(formula = Weight ~ Age, data = data, weights = w)
##
## Weighted Residuals:
##         1         2         3         4         5
## -5.21e-06 -5.23e-07  1.94e-05 -2.07e-05  8.66e-06
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.00270    0.00559  1431.3  7.5e-10 ***
```

```
## Age         -0.02610    0.00129   -20.2  0.00027 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.74e-05 on 3 degrees of freedom
## Multiple R-squared:  0.993,Adjusted R-squared:  0.99
## F-statistic:  408 on 1 and 3 DF,  p-value: 0.000265
```

Problem 7.8.3

Is the fitted regression consistent with the known standard weight for a new coin?

Solution

To obtain the average weight for a new coin, we only need refer to the intercept term.

```
coef(m)
```

```
## (Intercept)         Age
##      8.0027     -0.0261
```

Note that the fitted intercept is 8.0027, which is larger than the standard weight 7.9876, suggesting that it is consistent with the known standard weight for a new coin.

Problem 7.8.4

For previously unsampled coins of Age $= 1, 2, 3, 4, 5$, estimate the probability that the weight of the coin is less than the legal minimum.

(*Hints:* The standard error of prediction is the square root of the sum of two terms, the assumed known variance of an unsampled coin of known Age, which is different for each age, and the estimated variance of the fitted value for that Age; the latter is computed from the formula for the variance of a fitted value. You should use the normal distribution rather than a t to get the probabilities.)

Solution

```
m_var <- vcov(m)
m_var <- m_var[1,1] + data$Age^2 * m_var[2,2] + 2 * data$Age * m_var[1,2]
m_sd <- sqrt(m_var + data$SD^2)
y_hat <- predict(object = m, newdata = data.frame(Age = 1:5))
pnorm((7.9379 - y_hat) / m_sd)
```

```
##        1        2        3        4        5
## 0.004325 0.291282 0.652945 0.835362 0.890044
```

Problem 7.8.5

Determine the Age at which the predicted weight of coins is equal to the legal minimum, and use the delta method to get a standard error for the estimated age. This problem is called inverse regression, and is discussed by Brown (1993).

Solution

To solve this problem, we construct estimator

$$g(\hat{\beta}) = \frac{7.9379 - \hat{\beta}_0}{\hat{\beta}_1}$$

So that the estimated Age should be

$$g^*(\hat{\beta}) = \frac{7.9379 - 8.0027}{-0.0261} = 2.4828$$

Using delta method, we can get a standard error for the estimated age.

$$g(\hat{\beta})' = (-\frac{1}{\hat{\beta}_1}, -\frac{7.9379 - \hat{\beta}_0}{\hat{\beta}_1^2}) = (38.3141, 95.1249)$$

So that

$$Var[g(\hat{\beta})] = g(\hat{\beta})' Var(\hat{\beta})(g(\hat{\beta})')^T$$

```
grad <- as.matrix(c(38.3141, 95.1249))
sqrt(t(grad) %*% vcov(m) %*% grad)
```

```
##          [,1]
## [1,] 0.09657
```

So that the standard error for the estimated age is 0.09657.

Problem 7.10

(Data file: `fuel2001`)

Problem 7.10.1

Use the bootstrap to estimate confidence intervals for the coefficients in the fuel data, and compare the results with the usual large sample ols estimates.

Solution

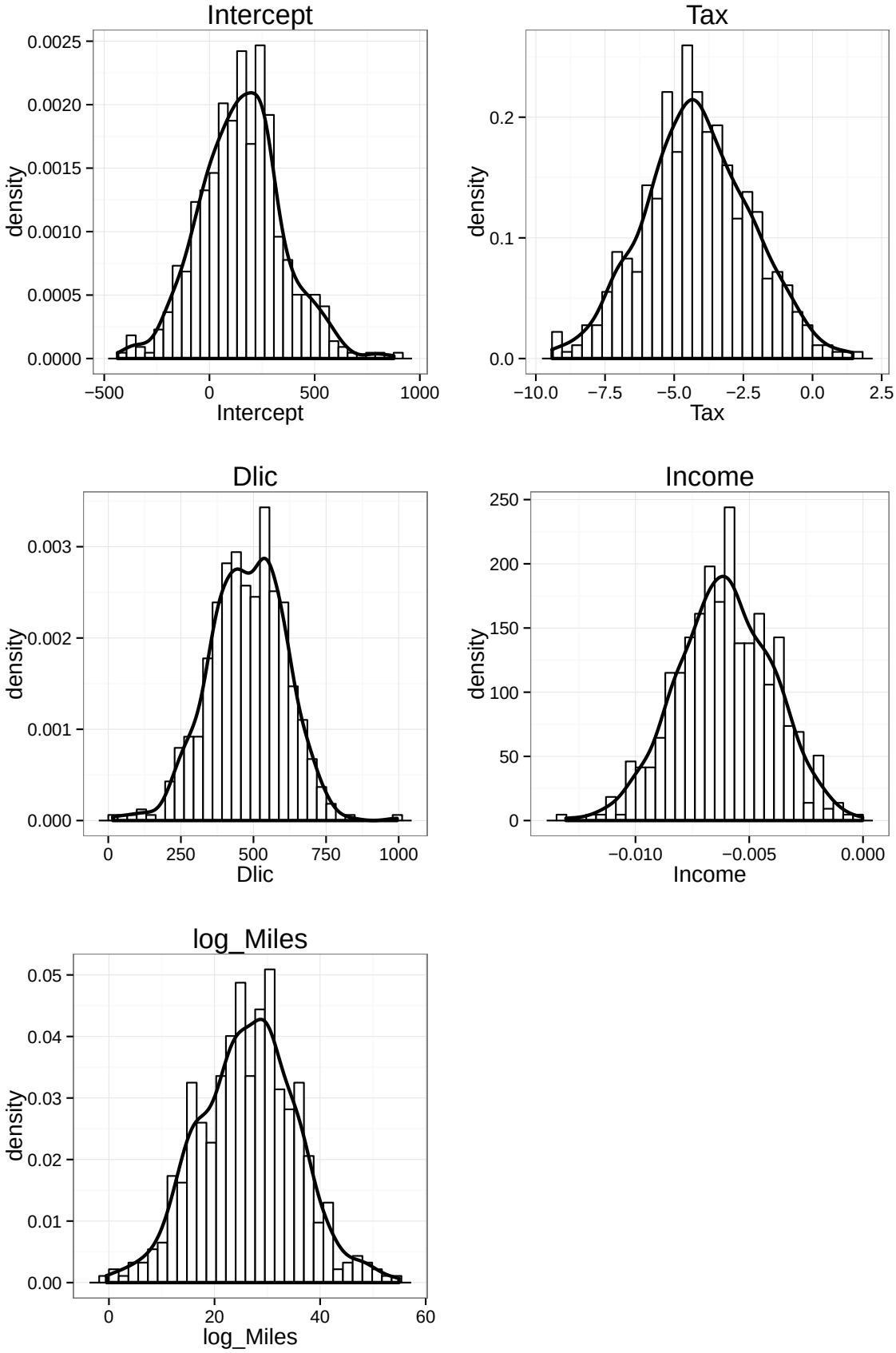
```
data <- fuel2001
data$Dlic <- data$Drivers / data$Pop
data$Fuel <- 1000 * data$FuelC / data$Pop
m <- lm(Fuel ~ Tax + Dlic + Income + log(Miles), data = data)
data_ori <- data

library(boot)
confint_bootstrap <- function(data, indices) {
    d <- data[indices, ] # allows boot to select sample
    data_tmp <- data_ori
    data_tmp$Fuel <- data_tmp$Fuel + d
    fit <- lm(Fuel ~ Tax + Dlic + Income + log(Miles), data_tmp)
    return(coef(fit))
}
set.seed(123)
results <- boot(data = as.matrix(m$residuals),
                statistic = confint_bootstrap, R = 500)
result <- results$t
colnames(result) <- names(coef(m))

## Bootstrap confidence interval
apply(X = result, MARGIN = 2, FUN = quantile, probs = c(0.025, 0.975))

##       (Intercept)    Tax  Dlic    Income log(Miles)
## 2.5%       -219.1 -7.870 231.1 -0.010151      8.587
## 97.5%       560.6 -0.405 705.6 -0.002059     44.601

## OLS confidence interval
t(confint(m))

##         (Intercept)     Tax  Dlic   Income log(Miles)
## 2.5 %       -238.1 -8.3144 213.2 -0.01055       7.96
## 97.5 %       546.5 -0.1416 730.6 -0.00172      45.55
```

In this problem, since the linear form of model is appropriate, we apply residual bootstrap to obtain the bootstrap confidence interval of coefficients. As the result displayed in the code chunk above, we see that the bootstrap confidence interval is more or less different from the interval derived from usual OLS, especially for the coefficient of `Tax`.

Problem 7.10.2

Examine the histograms of the bootstrap replications for each of the coefficients. Are the histograms symmetric or skewed? Do they look like normally distributed data, as they would if the large sample normal theory applied to these data? Do the histograms support or refute the differences between the bootstrap and large sample confidence intervals found in Problem 7.10.1?

```
colnames(result)[1] <- "Intercept"
colnames(result)[5] <- "log_Miles"

p1 <- ggplot(data = as.data.frame(result), aes(x = Intercept)) +
    geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
    geom_density(size = 1) + theme_bw() + theme(text = element_text(size = 14)) +
    ggtitle("Intercept")

p2 <- ggplot(data = as.data.frame(result), aes(x = Tax)) +
    geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
    geom_density(size = 1) + theme_bw() + theme(text = element_text(size = 14)) +
    ggtitle("Tax")

p3 <- ggplot(data = as.data.frame(result), aes(x = Dlic)) +
    geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
    geom_density(size = 1) + theme_bw() + theme(text = element_text(size = 14)) +
    ggtitle("Dlic")

p4 <- ggplot(data = as.data.frame(result), aes(x = Income)) +
    geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
    geom_density(size = 1) + theme_bw() + theme(text = element_text(size = 14)) +
    ggtitle("Income")

p5 <- ggplot(data = as.data.frame(result), aes(x = log_Miles)) +
    geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
    geom_density(size = 1) + theme_bw() + theme(text = element_text(size = 14)) +
    ggtitle("log_Miles")

grid.arrange(p1, p2, p3, p4, p5, ncol=2)
```



According to the histograms above, we can say that the histograms are not all symmetric, and not all of them look like normally distributed data. For example, the coefficient of the intercept seems like skewed, and the distribution of coefficient of `Dlic` seems has two peaks, suggesting that they may not be normally distributed, so that the large sample

normal theory may not be appropriate to these data.

However, it's kind of hard to say that the histograms support or refuse the differences between the bootstrap and large sample confidence intervals, because the shape of a histogram is largely affected by the binwitdh we choose, and it's risky to make decision based on histograms. Maybe we need to do some non-parametric test to determine whether they are normally distributed, and then we can say that if the differences are supported.