Linear Regression

S520

Arturo Valdivia

(Modified on April 17, 2023. Subject to change.)

Contents

1 The Linear Model

From the outset, we assume that there is a causal relationship,[1] i.e., a random variable, called the response, is affected by changes of another (or many other) variable(s), called the predictor(s).[2]

1.1 Conditional Random Variable

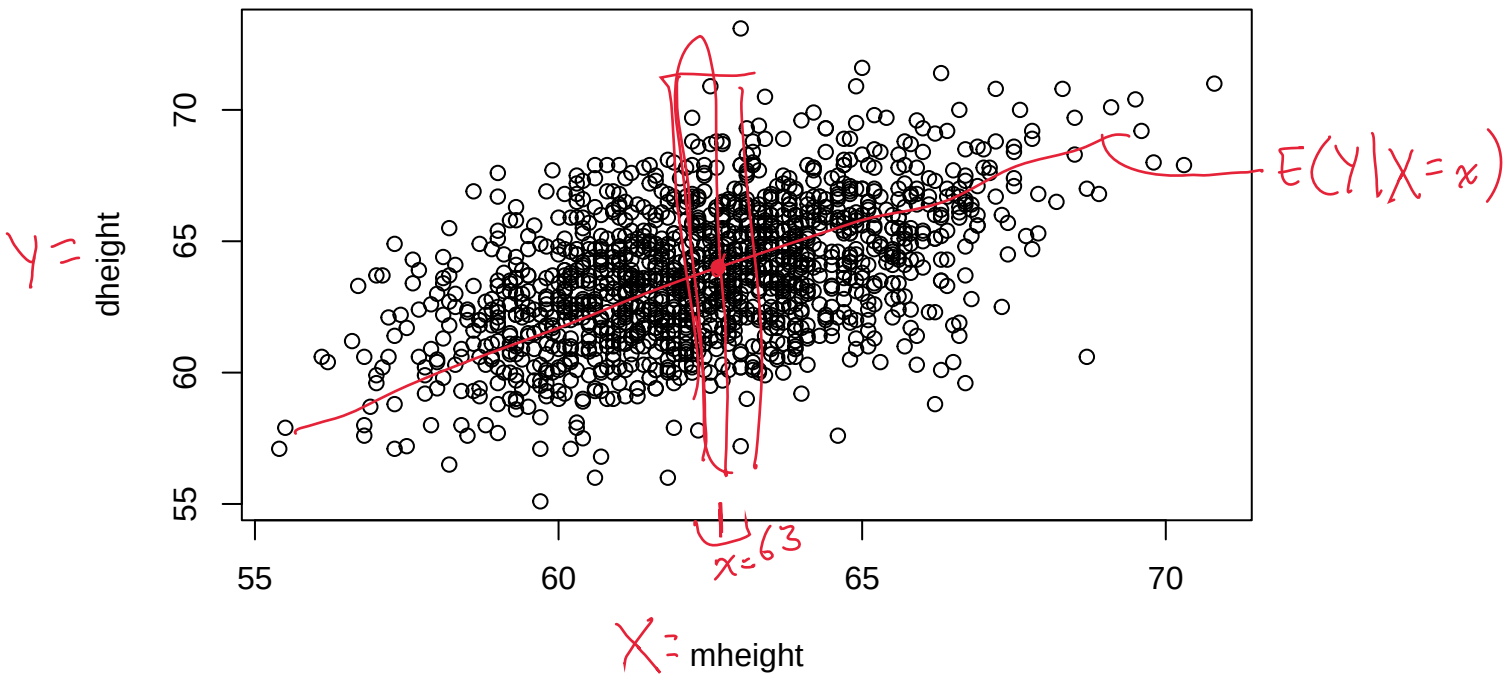
We account for a relationship that is not deterministic but of statistical nature. This means that the response, conditional on given values of the predictor(s), is itself a random variable.

- In mathematical terms, if Y is the response and X is the predictor taking an arbitrary but fixed value x, then Y given that $X = x$ is called a conditional random variable, and we write $(Y|X = x)$.[3]
- Observe that once x is given, $(Y|X = x)$ is just a random variable and we can find, for example, its expected value, $E(Y|X = x)$, and its variance, $Var(Y|X = x)$. These are key building blocks to be used for linear regression.

1.1.1 Example: Height of Mothers and Daughters

Let's use the dataframe `Heights`. The dataframe contains information of height in inches for mothers and corresponding daughters from a study performed by Karl Pearson between 1893 and 1898. We let daughter's height, `dheight`, be the response and mother's height, `mheight`, the predictor. A scatterplot is useful to plot this relationship.

```
Heights = read.table("Heights_Pearson.txt",header = T)
plot(dheight ~ mheight, data = Heights)
```



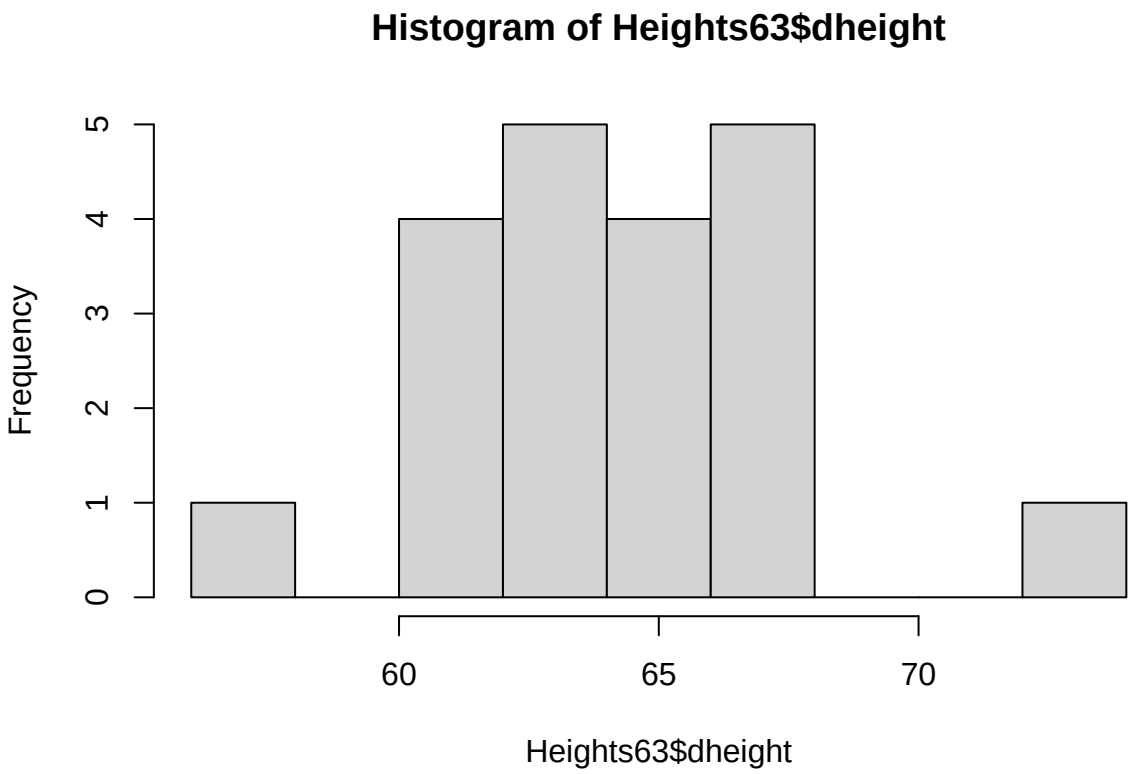
[1] Note that the problem at hand could also be built instead as many random variables being dependent of each other, so taken together they form a multivariate distribution, but we won't pursue that treatment here.

[2] The question of the existence of causation is a fundamental question in statistics and data analysis; however, it is not a question answered by the techniques/methods introduced in this section; here, we simply assume that this relationship exists.

[3] With three predictors, for example, the conditional random variable could be expressed as $(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$.

Observe that for any given value of `mheight`, there is a sub-sample of possible values for `dheight`. For example, if `mheight = 63`, then (`dheight|mheight = 63`) is presented in the following plot

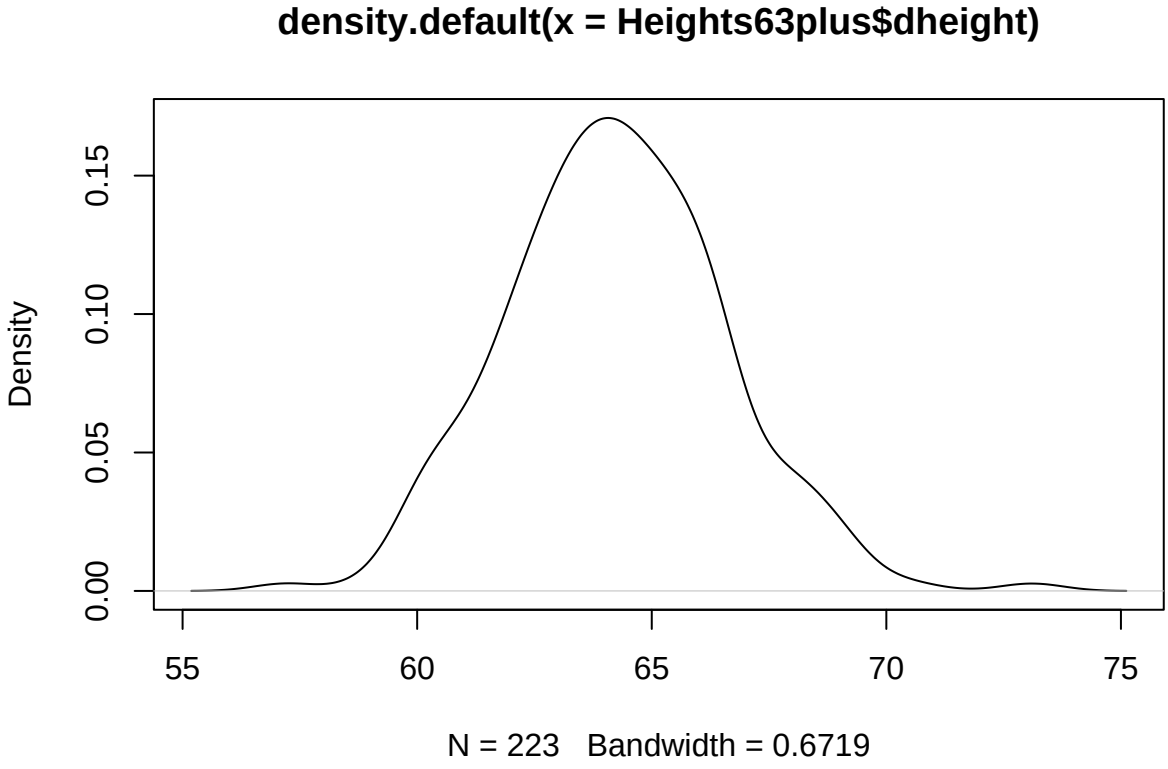
```
Heights63 <- subset(x = Heights, subset = (mheight == 63))
hist(Heights63$dheight)
```



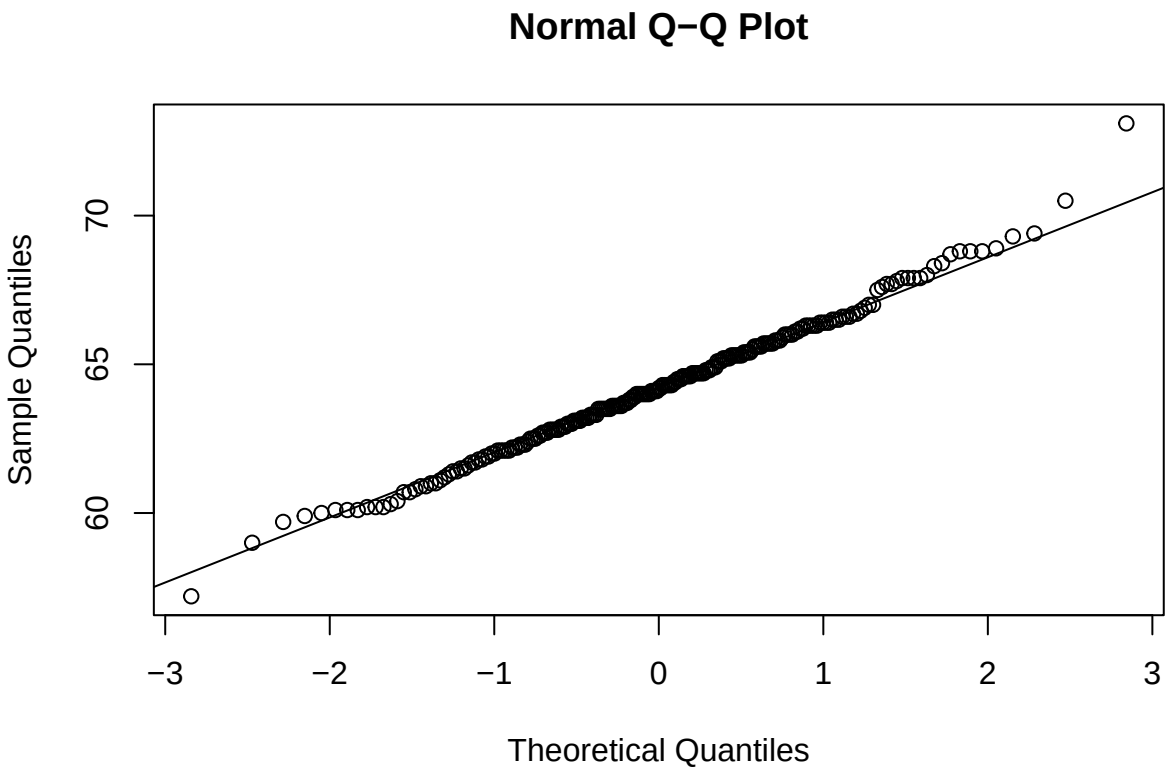
This histogram is not too useful because the number of observations with `mheight = 63` is very small. To better visualize the distribution of (`dheight|mheight = 63`) we could, for example, use a slightly larger group by considering all the data within a small interval around `mheight = 63`,

```
Heights63plus <- subset(Heights, subset = ((62.5 < mheight) & (mheight <= 63.5)))
hist(Heights63plus$dheight)
```


```
plot(density(Heights63plus$dheight))
```



```
qqnorm(Heights63plus$dheight)
qqline(Heights63plus$dheight)
```



Here (`dheight|mheight=63`) seems unimodal, approximately symmetric, not too far off from a normal distribution. The key point here is that `dheight` is a (potentially different) random variable for each given value of `mheight`.

2 Linear Regression

We say that there is a linear relationship between Y and X if the expected value of Y conditional on $X = x$, for some real number x, can be represented by a line, that is

$$E(Y|X = x) = \beta_0 + \beta_1 x \quad \text{→ Mean Function} \tag{1}$$

where β_0 (intercept) and β_1 (slope) are some scalars. We call (1) the mean function, and it provides the relationship between Y and X.

Don't forget that $(Y|X = x)$ is a random variable

While the expected value is fully defined by the value of X, the random variable $(Y|X = x)$ is not as there is some variation of potential values that Y could take. We assume, however, that this variation is constant for any value of X and take this into account by introducing the variance function:

$$Var(Y|X = x) = \sigma^2 \quad \text{→ Variance Function} \tag{2}$$

where σ^2 is a scalar that represent this constant variance for any value x in the range of X. When we focus in a single predictor as in , we refer to (1) and (2) as the **simple linear regression** model because only one predictor X is included.

Equivalently, a **multiple linear regression** model relates Y to $p \geq 2$ predictors, with mean function

$$E(Y|X_1 = x_1, \ldots, X_p = x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

and variance function

$$Var(Y|X_1 = x_1, \ldots, X_p = x_p) = \sigma^2.$$

Many key characteristics apply to both simple and multiple linear regression. These notes will focus mainly on simple linear regression, but an example of multiple linear regression will be presented at the end for your reference.

2.1 Error term and the linear model for y_i

As we did in previous chapters, we want to use a sample to make inferences about the population. Observe that now each observation in our sample is composed by a pair (y_i, x_i) for $i = 1, \ldots, n$ independent observations[4].

For simplicity, we'll also use the notation y_i to represent the conditional response $(Y|X = x_i)$ for the ith observation. So, depending on the context y_i may represent the observed value of the response or the conditional random variable $(Y|X = x_i)$. In the latter, we can study its distribution and all related properties.

As a random variable, the values of y_i don't have to be exactly equal to $E(Y|X = x_i)$. Instead,

$$y_i = E(Y|X = x_i) + e_i = \beta_0 + \beta_1 x_i + e_i$$

where e_i is a random variable called the error term. We assume that the mean $E(e_i) = 0$ and the variance $Var(e_i) = \sigma^2$ for $i = 1, \ldots, n$.[5] The error term is the difference between the observed y_i and the expected value $E(Y|X = x)$. To summarize, we assume that

1. The mean function is linear in the parameters, β_0 and β_1.
2. The variance is assumed constant for any values $X = x$.
3. e_i, e_j are pairwise independent and in turn, y_i, y_j are pairwise independent for all $i \neq j$

2.2 Simulation: Constructing samples

When dealing with real data, we assume that our linear model is a good representation of the relationship between Y and X and that there exist parameters β_0, β_1, and σ^2, even though there are unknown to us.

Let's simulate some data where we construct a linear relationship between the predictor X and the response Y. In this simulation we assume the parameters (true values) are $\beta_0 = 10, \beta_1 = 1.5$, and $\sigma^2 = 100$. We take a sample for values for the predictor (X) and use them alongside error terms to obtain values for the response (Y). Let's simulate a sample of 50 observations:

```
beta0 = 10
beta1 = 1.5
sigma_2 = 100
n = 50
x = sample(50:100, n, replace = TRUE)
e = rnorm(n, mean = 0, sd = sqrt(sigma_2))
y = beta0 + beta1 * x + e
data.sim <- data.frame(x = x, y = y)
```

Here are a few rows of our sample

```
head(data.sim)
```

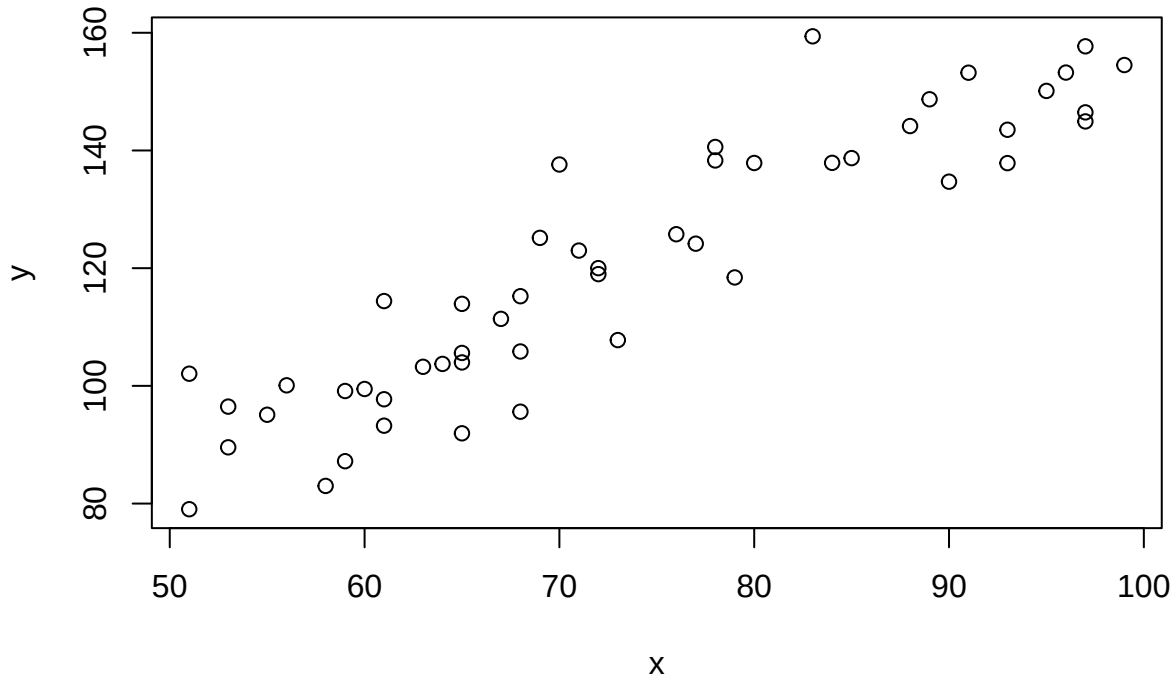
[4] or for multiple regression each observation is given by $(y_i, x_{i1}, \ldots, x_{ip})$

[5] Observe that, in principle, e_i is also dependent of $X = x_i$, but we assume that the expected value is zero regardless the value of X.

```
   x       y
1 90 134.699
2 51 102.067
3 60  99.471
4 93 143.521
5 68 115.236
6 67 111.385
```

The scatterlot for the entire sample is

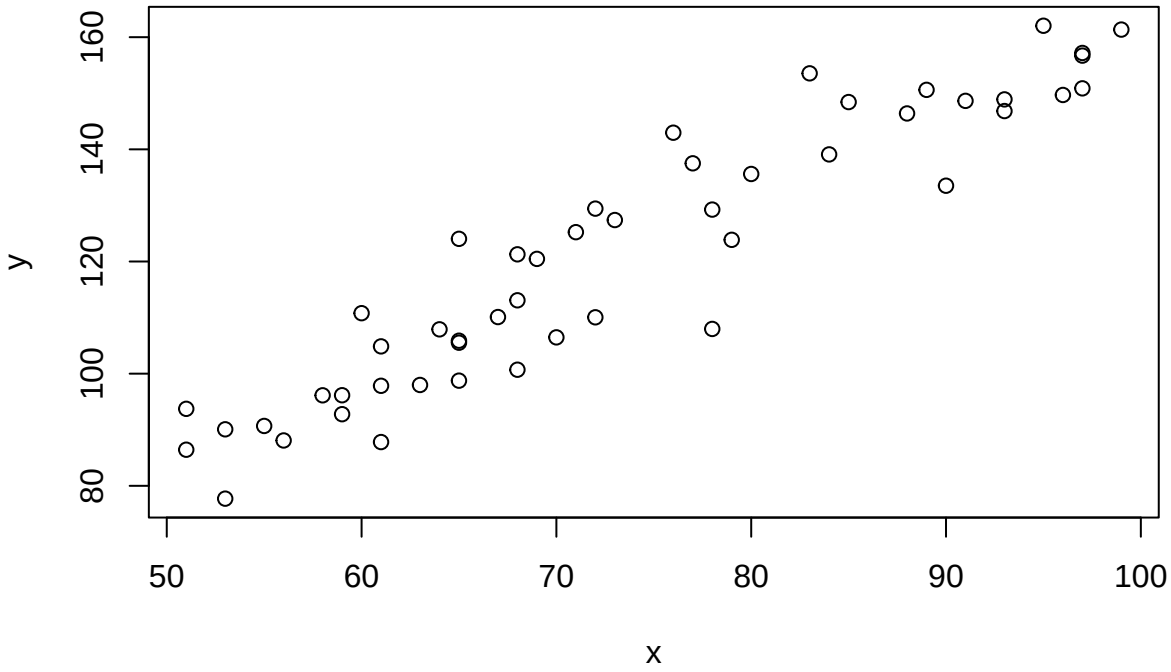
```
plot(y ~ x)
```



We can also simulate a new sample, where the predictor values are the same, but new responses are generated:

```
e = rnorm(n, mean = 0, sd = sqrt(sigma_2))
y = beta0+beta1*x + e
data.sim1 <- data.frame(x = x, y = y)
plot(y ~ x)
```

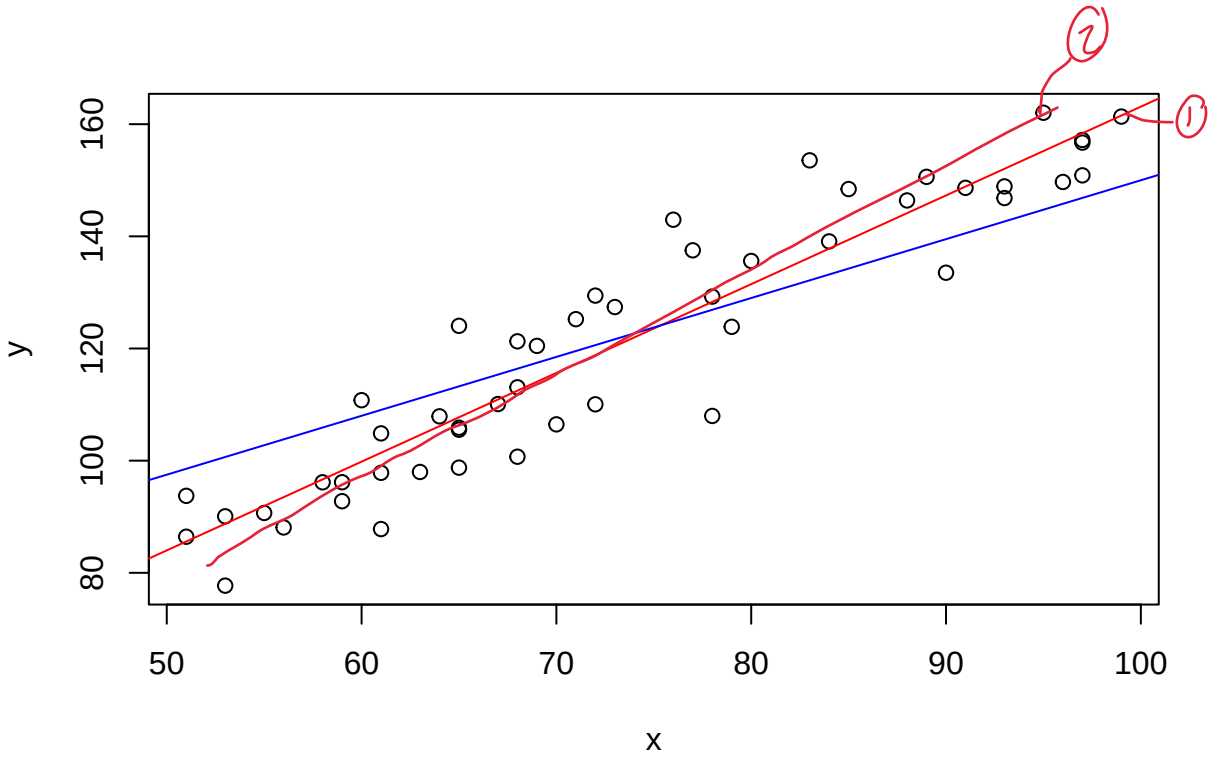
fixed.



Observe that the scatterplots are somewhat different, but they preserve the same positive linear trend showing the relationship between X and Y.

3 Least Squares Estimation

The first goal of linear regression is to, based on a sample, produce estimators for the unknown coefficients: intercept (β_0) and slope (β_1). The idea is to come up with these values such that, based on some criterion, provides the best representation of the data observed. For example, which line is more appropriate for the data at hand. The blue or the red line?

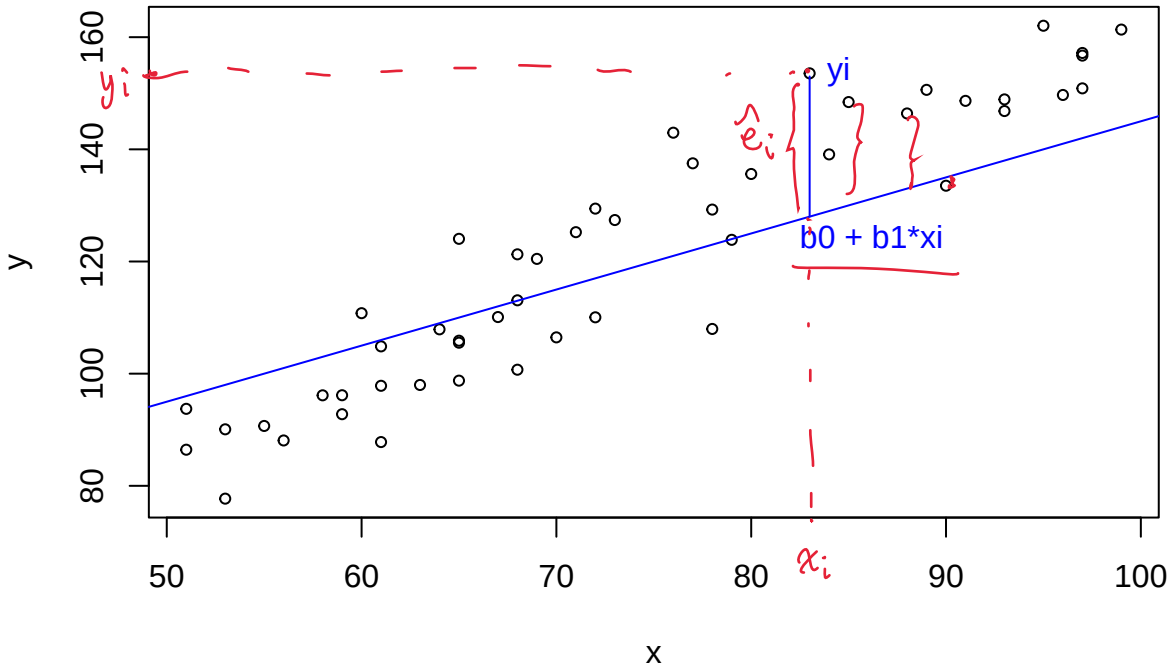


3.1 Ordinary Least Squares (OLS) Coefficient Estimators

We define a residual for the ith observation, $\hat{e}_i$, as the vertical distance between the observed response, y_i, and the y-axis value of the line for x_i, or

$$\hat{e}_i = y_i - b_0 - b_1 x_i$$

where b_0 and b_1 are the intercept and slope of any given line (such as the blue or red lines), as shown in the following plot for one possible observation:



We obtain residuals for all observations, to avoid positive or negative differences we square the residuals, and add them up. The resulting value is called the Residual Sum of Squares (RSS) for the line with intercept b_0 and slope b_1, or

$$RSS(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

To determine whether the red line or the blue line is more appropriate in the plot above, we compare the RSS value for each and pick the line with the smallest value. Of course, there is no reason to believe either of these lines is the best we can get.

The criterion of Ordinary Least Squares (OLS) provides the solution as the line with smallest RSS. We can express this optimization problem in terms of the intercept and slope as

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{b_0, b_1}{\text{Arg min}} \quad RSS(b_0, b_1) = \underset{b_0, b_1}{\text{Arg min}} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

that is $RSS(\hat{\beta}_0, \hat{\beta}_1)$ is the smallest RSS value that we can obtain and the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the OLS estimators of β_0 and β_1 in (1).

This is an optimization problem that can be solved using calculus, we can find the partial derivative of $RSS(b_0, b_1)$ with respect to b_0 and b_1, equate each equation to zero, and find the solutions for b_0 and b_1. If we do this, the solution is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{and} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{and} \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

The variance estimator, called the residual mean square, is defined as

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

where $RSS = RSS(\hat{\beta}_0, \hat{\beta}_1)$ is the RSS obtained by using the OLS coefficient estimators.

For multiple linear regression, when we have two or more predictors, we still minimize RSS. In this case, the equations for the OLS estimators, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, become more convoluted using basic algebra but are easier to solve and represent them using matrix algebra.[6]

3.1.1 Example: Height of Mothers and Daughters (continued)

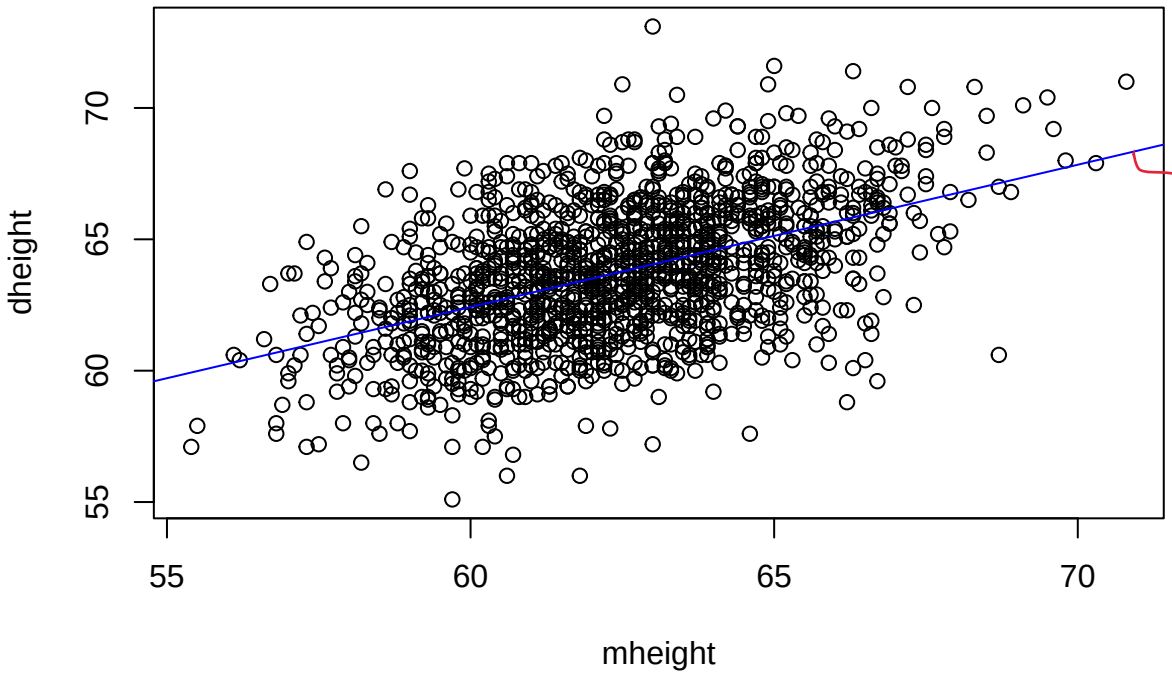
The function `lm()` in R produces the OLS estimates. For example, using the `Heights` data frame we get

```
mod1 = lm(dheight ~ mheight, data = Heights)
coef(mod1)
```

```
(Intercept)     mheight
  29.91744      0.54175
```

So $\hat{\beta}_0 = 29.92$ and $\hat{\beta}_1 = 0.54$. The OLS coefficient estimates define the estimated regression line. Graphically, we get

```
plot(dheight ~ mheight, Heights)
abline(mod1, col="blue")
```



Similarly, the `lm()` object produces the square root of the residual mean square, $\hat{\sigma}$, also called the standard error of regression

[6] The treatment of matrix algebra is beyond the scope of this course, but it's very useful when dealing with multiple linear regression.

```
sigma(mod1)
```

```
[1] 2.2663
```

3.1.2 Simulation: Comparing estimated regression line with true line

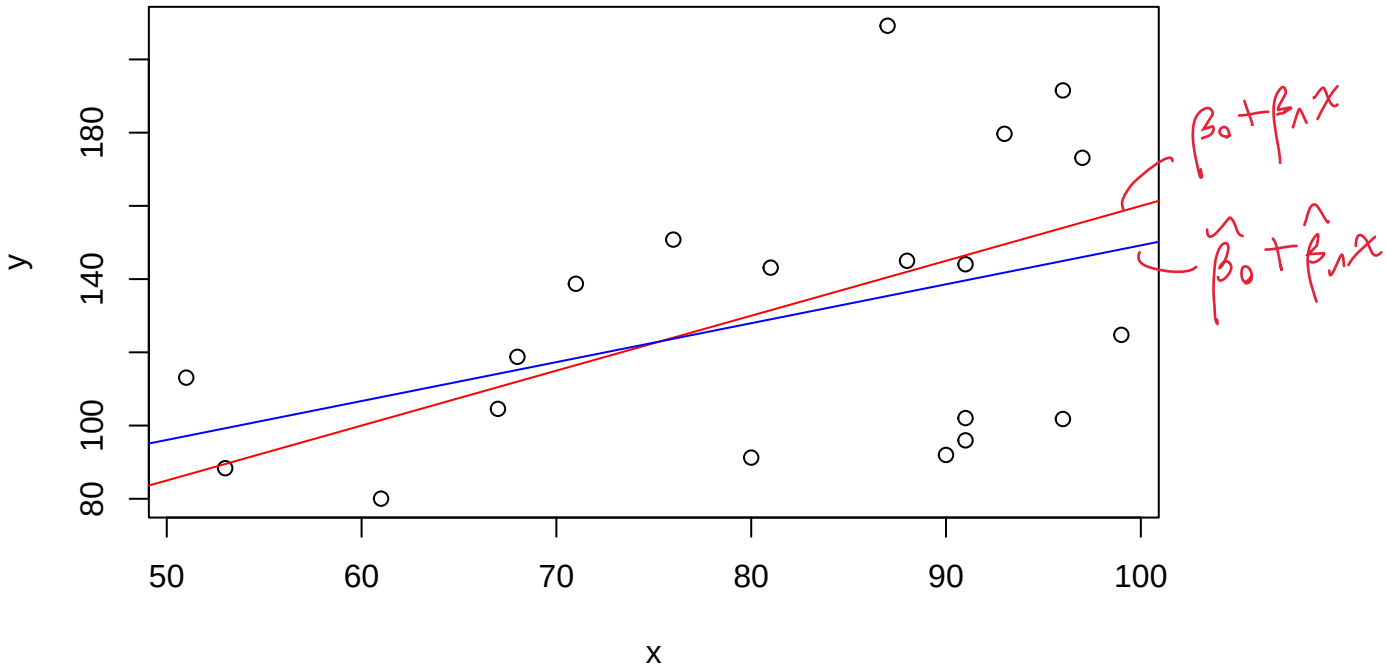
It is important to understand the key difference between the regression line, obtained based on data using the OLS criterion, $(\beta_0, \beta_1, \dots)$, and the line given by the mean function, which is based on the true parameters $\beta_0, \beta_1, \dots$. Let's use a simulation again to observe the difference.

```
beta0 = 10 # true parameter
beta1 = 1.5 # true parameter
sigma_2 = 30^2 # true parameter
n = 20
set.seed(500)
x = sample(50:100, n, replace = TRUE) #some given sample of values for X
e = rnorm(n, mean = 0, sd = sqrt(sigma_2)) # the random errors
y = beta0 + beta1 * x + e # the response values
data.sim <- data.frame(x = x, y = y)
m.sim <- lm(y ~ x, data = data.sim) # The regression line based on the data
betahat0 <- coef(m.sim)[1] # OLS estimator
betahat1 <- coef(m.sim)[2] # OLS estimator
c(beta0 = beta0, betahat0 = betahat0, beta1 = beta1, betahat1 = betahat1)
```

```
      beta0 betahat0.(Intercept)               beta1           betahat1.x
    10.0000              42.9290               1.5000               1.0628
```

Let's now construct a scatterplot that includes the mean function (true line) and the regression line

```
plot(y ~ x)
abline(a = beta0, b = beta1, col = "red") # mean function
abline(m.sim, col = "blue") # regression line
```



While the lines are fairly close in the graph, it is clear that both have different slopes and intercepts. In real life situations, the mean function (red) is unknown as we can only obtain the regression line (blue). Note also that the regression line depends on the data; if we were to take a new random sample, the regression line would be (slightly) different.

3.2 Interpretation of OLS Estimators

- The intercept in the mean function, $\hat{\beta}_0$, is the estimated value of the mean response when the predictor(s) is(are) zero only when the range of values for the predictor includes zero. If the range does not include zero, there is no meaningful interpretation of the intercept and it is used only for mathematical reasons (to be able to obtain the best line to represent the mean function). You can always check the range of the predictor (X) to determine if an interpretation for the intercept is needed:

```
summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   51.0    70.2    87.5    81.3    91.5    99.0
```

- In simple linear regression, the estimated slope in the mean function, $\hat{\beta}_1$, is the amount of change in the mean response, $E(y_i)$, when the predictor increases by one unit.
- In multiple linear regression, $\hat{\beta}_j$, $j = 1, \ldots, p$, are partial slopes. Each one is the estimated amount of change in the mean response when the regressor increases by one unit, holding all the other regressors fixed to any given value.

For example, for the `Heights` example, let's first look at the range of the predictor `mheight`:

```
summary(Heights)
```

```
    mheight         dheight
 Min.   :55.4   Min.   :55.1
 1st Qu.:60.8   1st Qu.:62.0
 Median :62.4   Median :63.6
 Mean   :62.5   Mean   :63.8
 3rd Qu.:63.9   3rd Qu.:65.6
 Max.   :70.8   Max.   :73.1
```

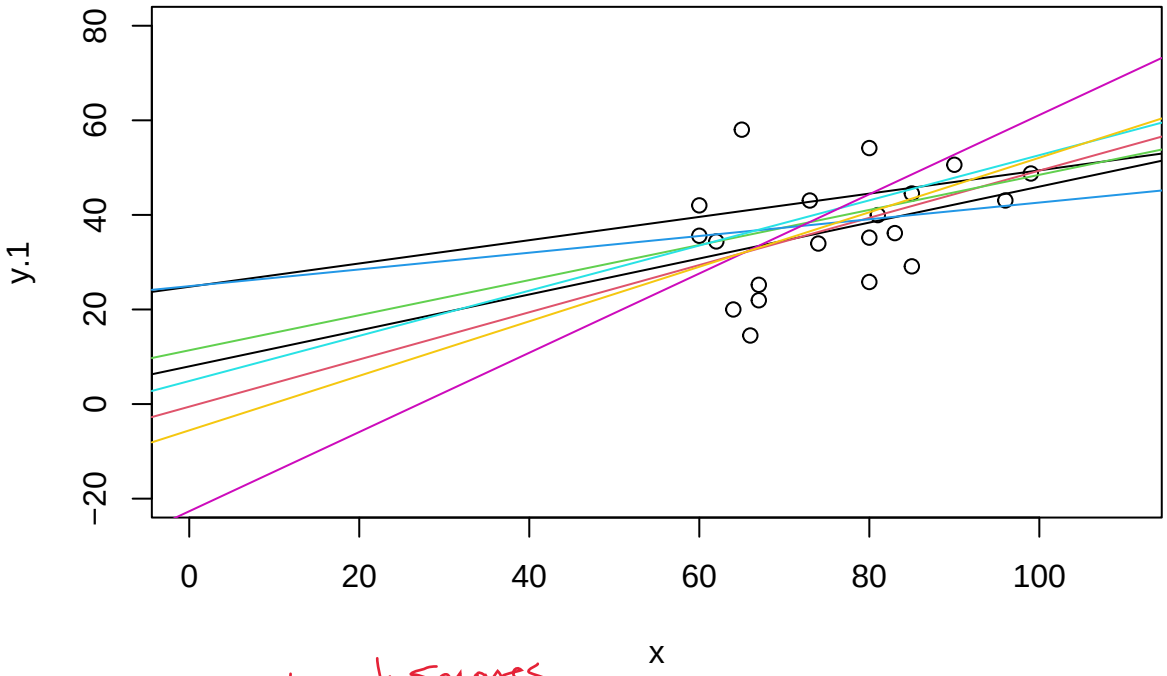
- Observe that the intercept in the regression line won't have a meaningful interpretation because `mheight = 0` is not part of the range of values since all mothers are taller than 0 inches.
- In terms of the slope, here are two equivalent interpretations:
  - If the mother is one inch taller than some given height, we expect the corresponding daughter's height to be $\hat{\beta}_1 = 0.54$ inches taller,
  - Alternatively, daughters of mothers who are one inch taller than some given height would be 0.54 inches taller, on average.

3.3 Properties of Coefficient Estimators

1. With a small amount of algebraic manipulation for the simple linear regression[7], it can be shown that $\hat{\beta}_0$ and $\hat{\beta}_1$ can be expressed as linear combinations of $y_1, y_2, \ldots, y_n$.
2. Because $y_i = (Y|X = x_i)$ are (conditional) random variables before data have been collected, $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of independent random variables and in turn, they are also random variables. Using our previous simulation, let's find new samples (with X values fixed) for different realizations of $y_1, y_2, \ldots, y_n$ to see how our coefficient estimators change:

```
beta0 = 1 #The true intercept
beta1 = .5  # The true slope
sigma = 10 # The true variance

n = 20 # The sample size
set.seed(321) # same random seed for replicability
x = sample(50:100, n, replace = TRUE) # Some sample of predictor values

# let's now plot our first sample (y_i, x_i)
e.1 = rnorm(n, mean = 0, sd = sigma)
y.1 = beta0+beta1*x+e.1
model1 = lm(y.1 ~ x) # The function lm produces the OLS estimators
plot(y.1 ~ x, xlim = c(0,110), ylim = c(-20,80))
abline(model1, col =1) # We plot the line using the OLS est.

# Let's add 7 more OLS lines from 7 new samples
# Note the predictors values (x's) are the same
# But our responses, y, are different because the error terms
# are different

for (j in 1:7){
  e.j = rnorm(n, mean = 0, sd = sigma)
  y.j = beta0+beta1*x+e.j
  model.j = lm(y.j ~ x)
  abline(model.j, col = j)
}
```

[7] or matrix manipulation for multiple linear regression



Ordinary Least Squares

3. OLS estimators are unbiased, i.e., $E(\hat{\beta}_0|X = x) = \beta_0$, $E(\hat{\beta}_1|X = x) = \beta_1$, and $E(\hat{\sigma}^2|X = x) = \sigma^2$. Let's use our simulation to see if we can approximate this result. Using 10^5 replications, we obtain the OLS regression estimates, find the average of values for each coefficient and compare them with the true parameters.

```
set.seed(123)
repli <- 10000
betahat0.vec <-  rep(NA, repli)
betahat1.vec <-  rep(NA, repli)
sigmahat.vec <- rep(NA, repli)
for (j in 1:repli){
  e.j = rnorm(n, mean = 0, sd = sigma)
  y.j = beta0+beta1*x+e.j
  lm.j = lm(y.j ~ x)
  betahat0.vec[j] = coef(lm.j)[1]
  betahat1.vec[j] = coef(lm.j)[2]
  sigmahat.vec[j] <- sigma(lm.j)
}

data.frame("Parameters" = c(beta0,
                            beta1,
                            sqrt(sigma_2)),
           "Estimates"= round(c(mean(betahat0.vec),
                                mean(betahat1.vec),
                                mean(sigmahat.vec)),2),
           row.names = c("beta0", "beta1", "sigma"))
```

```
      Parameters Estimates
beta0        1.0      1.08
beta1        0.5      0.50
```

```
sigma       30.0       9.89
```

The average of OLS estimates, based on 10^5 replicates, are certainly close enough to the true parameters.

4 Inferences about Coefficients

So far, we have not made any assumptions about the distribution of the random variables in our linear model. While it would be possible to make some inferences only based on, for example, large samples and the Central Limit Theorem, many useful results can be applied if we introduce a few assumptions about the distribution of these random variables.

So, let's now assume that the error terms are independent and identically distributed following a normal distribution with mean zero and variance σ^2,

$$e_i \sim \mathcal{N}(0, \sigma^2)$$

for $i = 1, \ldots, n$. This has direct implications in many of our results:

1. Observe that the response, y_i for $i = 1, \ldots, n$, is also normally distributed,
$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$
2. Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables constructed as a linear combination of $y_1, \ldots, y_n$, they also follow normal distributions.
3. The residual mean square, $\hat{\sigma}^2$, follows a multiple of a chi-squared distribution with $n-2$ degrees of freedom,
$$\hat{\sigma}^2 \sim \left[\frac{\sigma^2}{n-2}\right] \chi^2_{n-2}$$

Based on these results, it is possible to use inferential methods similar to those used in ISIR Ch9, 10, and 11.

4.1 Test of significance for OLS Coefficients

It is possible to test for claims about each coefficient in the linear model, β_0 and β_1. For example, for β_1 a useful test is:

$$\begin{aligned} H_0 &: \beta_1 = k \\ H_1 &: \beta_1 \neq k \end{aligned}$$

for some number k. If the null hypothesis is true, the test statistic

$$t = \frac{\hat{\beta}_1 - k}{se(\hat{\beta}_1 | X)} \sim T_{n-2} \tag{3}$$

follows a T distribution with $n-2$ degrees of freedom. So, finding the p-value, comparing to a predefined significance level α, and making decisions about the claim under the null hypothesis is analogous to the work done for the mean, μ, in ISIR ch9 and 10.

When considering the linear regression model, perhaps the most relevant claim is of the form

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

because recall that the relationship between Y and X is given by

$$y_i = \beta_0 + \beta_1 x_i + e_i.$$

Failing to reject the null hypothesis is equivalent to not finding statistical evidence that $\beta_1 = 0$, i.e., not finding evidence that changes in X produce changes in Y.

4.1.1 Example: Height of Mothers and Daughters (continued)

The summary of object `lm()` in R contains all the information needed for these tests. For example, let's use the dataframe `Heights` to test if there is statistical evidence that mother's height, `mheight`, influence daughter's, `dheight`, or

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

```
mod1 <- lm(dheight ~ mheight, data = Heights)
summary(mod1)
```

```
Call:
lm(formula = dheight ~ mheight, data = Heights)

Residuals:
   Min     1Q Median     3Q    Max
-7.397 -1.529  0.036  1.492  9.053

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.917      1.623    18.4   <2e-16
mheight        0.542      0.026    20.9   <2e-16

Residual standard error: 2.27 on 1373 degrees of freedom
Multiple R-squared:  0.241, Adjusted R-squared:  0.24
F-statistic:  435 on 1 and 1373 DF,  p-value: <2e-16
```

The section of "Coefficients" in the middle of the output contains all the information needed. The line starting with `mheight` contains in order:

- The estimate $\hat{\beta}_1 = 0.542$
- The standard error, $se(\hat{\beta}_1|mheight) = 0.026$
- The test statistic, under the null hypothesis, $t = 20.9$.
- The p-value, for a two-tailed test, p-value ≈ 0.

Based on this information, the p-value is nearly zero; therefore, for any significance level α, we reject the null hypothesis that β_1 is equal to 0 and conclude that mother's height influence daughter's height.

While the output is readily available using the R function `summary()` on your `lm()` object, observe that you could also obtain the required results manually in R as we did in ISIR Ch9 and 10. We first construct the test statistic as in (3)

```
beta1hat = coef(mod1)[2] # coef(mod1) provides a vector with all betahats
k = 0
se.beta1hat = summary(mod1)$coef[2,2] # the SE can be obtained from the summary function
t.stat = (beta1hat - k)/se.beta1hat
t.stat
```

```
mheight
 20.868
```

Since my hypothesis is a two-sided problem, we obtain the p-value on both tails:

```
n = dim(Heights)[1] #number of rows in the data set
2*(1 - pt(abs(t.stat), n - 2))
```

```
mheight
      0
```

While doing this manually is not needed for the specific problem at hand, as the output is readily available, changing the hypotheses formulation may require to use the manual process. For example the summary output cannot provide the answer directly for

$$H_0 : \beta_1 \leq 0.5$$
$$H_1 : \beta_1 > 0.5$$

but we can construct this test easily by hand

```
t.stat = (beta1hat - 0.5)/se.beta1hat
t.stat
```

```
mheight
 1.6081
```

```
1 - pt(t.stat, n - 2)
```

```
 mheight
0.054023
```

4.2 Set Estimation: Confidence Intervals for OLS Coefficients

We can construct $(1 - \alpha) \times 100$ confidence intervals for each OLS regression coefficient,

$$\hat{\beta}_j \pm q \times se(\hat{\beta}_j|X)$$

where q is the $(1 - \alpha/2)$-quantile of a T-distribution with $(n - 2)$ degrees of freedom.

4.2.1 Example: Height of Mothers and Daughters (continued)

Let's obtain a 98% confidence intervals. To obtain this directly we use the R function `confint()` on our `lm()` object

```
confint(mod1, level = 0.98)
```

```
                1 %     99 %
(Intercept) 26.13860 33.69628
mheight      0.48128  0.60221
```

We are 98% confident that β_1, the rate of change in daughter's height due to one inch increase in mother's height, is between 0.48 and 0.60 inches.

Observe the function `confint()` provides information for all OLS coefficients. You can also do this manually if you wish

```
beta1hat = coef(mod1)[2]
se.beta1hat = summary(mod1)$coef[2,2]
beta1hat - qt(1 - 0.02/2, n - 2)*se.beta1hat
```

```
mheight
0.48128
```

```
beta1hat + qt(1 - 0.02/2, n - 2)*se.beta1hat
```

```
mheight
0.60221
```

4.3 Prediction

Let's assume a new observation for predictor X is $x*$. If our model is correct, the new response will be given by

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

where e^* is the random error. Based on our linear regression, we can predict the response to be

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

Note that the OLS estimators are random variables (for each different sample drawn different OLS estimates are obtained) and the error term is a random variable; therefore $\hat{y}^*$ is a random variable that has two sources of variation

- Due to the OLS estimates
- Due to the the error term

When using inferential methods then, the standard error of $\hat{y}^*$, $se_y(\hat{y}^*|X = x^*)$, accounts for both sources of variation. The most common inferential method is to obtain a prediction interval, i.e., a confidence interval for y^*, with $(1-\alpha)\times 100$ level of confidence,

$$\hat{y}^* \pm q \times se_y(\hat{y}^*|X = x^*)$$

A less common goal is to estimate what the expected value of the response will be given the new observation x^*,

$$E(Y|X = x^*) = \beta_0 + \beta_1 x^*,$$

and only one source of variation, the OLS estimators, needs to be consider here. The confidence interval for $E(Y|X = x^*)$ is given by

$$\hat{y}^* \pm q \times se_E(\hat{y}^*|X = x^*)$$

Where se_E, a standard error that accounts for only one source of variation, is used instead of se_y.

4.3.1 Example: Height of Mothers and Daughters (continued)

Let's find a 97% prediction interval for the daughter's height given that a new observation shows the height of a mother is 63 inches. To obtain this result in R we require to store the new observation in a dataframe with the same variable names than the original dataframe. Let's observe the names used in the original data frame.

```
colnames(Heights)
```

```
[1] "mheight" "dheight"
```

Now we can build a dataframe with our new observation

```
new.data.height <- data.frame(mheight = 63)
```

We now construct a 97% prediction interval for the new observation. The object needed is our `lm()` object `mod1`, the new data is the data frame constructed above, a `"prediction"` interval argument is needed for the appropriate standard error, $se_y(\hat{y}^*|X = 63)$, and the confidence level needs to be specified if different than 0.95:

```
predict(mod1, newdata = new.data.height, interval = "prediction", level = 0.97)
```

```
     fit    lwr    upr
1 64.047 59.122 68.973
```

So, for a new mother who is 63" tall, we are 97% confident that her daughter's height will be between 59 and 69 inches, not a narrow interval by any means, as two sources of variation need to be taken into account.

If on the other hand, we would be only interested in finding a 97% confidence interval for the average daughter's height for all mothers who are 63 inches tall, then the R code is similar, only changing to a `"confidence"` interval to account for the appropriate standard error, $se_E(\hat{y}^*|X = 63)$:

```
predict(object = mod1, newdata = new.data.height, interval = "confidence", level = 0.97)
```

```
     fit    lwr    upr
1 64.047 63.911 64.184
```

so, for all mothers with height 63", we are 97% confidence that the average daughter's height will be between 63.9 and 64.2 inches, a much narrower interval because we only take into account the variation due to the OLS estimators.

4.4 Revisiting Residuals

When OLS estimators are used to find the values of the response, for the ith row of your sample,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The estimated response, $\hat{y}_i$, is called the fitted value (for the ith observation). In this context, we can redefine the residual for the ith observation, $\hat{e}_i$, as the difference between the fitted value and the observed value,

$$\hat{e}_i = y_i - \hat{y}_i.$$

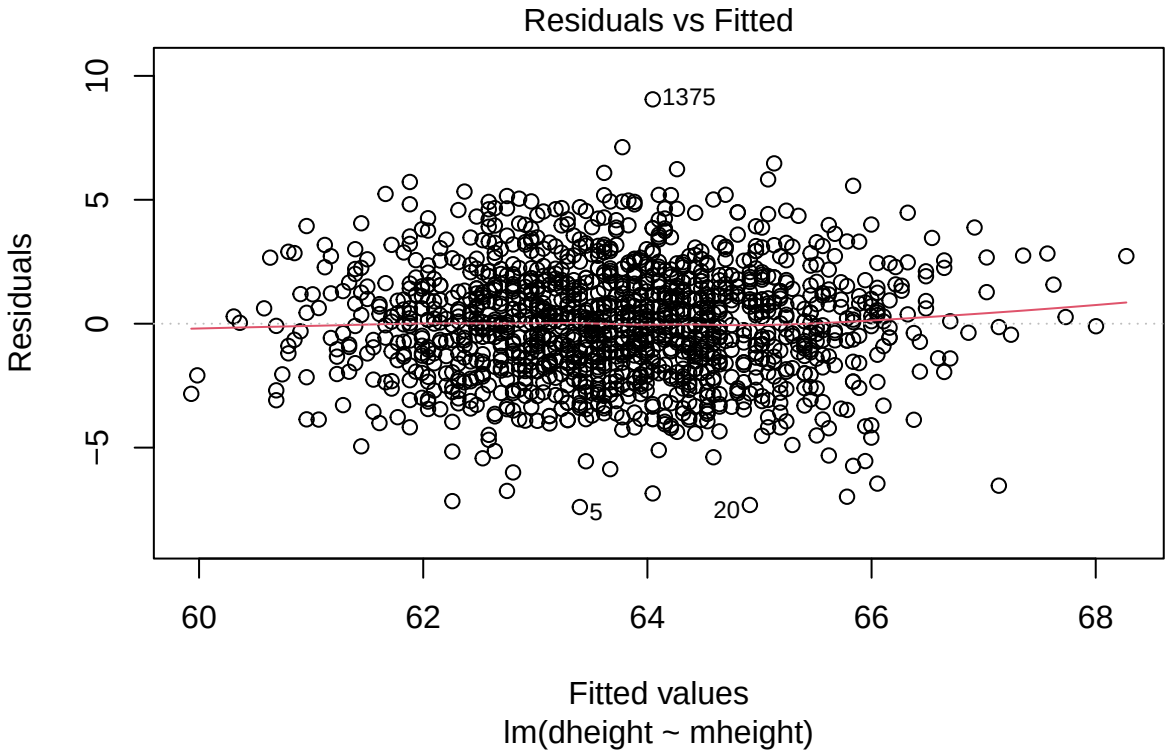
The residuals play an important role in linear regression, because they can help us determine if the data fit some of the assumptions in the model. Our focus will be in graphical considerations, namely in residual plots.[8]

4.4.1 Example: Height of Mothers and Daughters (continued)

Let's obtain the residual plot for the Height problem. For that, we use the R function `plot()` on an `lm()` object. This plot function has 6 different plots and prints 4 of them by default. We are interested only on the first plot, the residuals plot, so we include the argument `which =1`.

```
plot(mod1, which = 1)
```

[8] Many tests of significance have been developed to check for model assumptions using the residual, but they are beyond the scope of this course.



This is a scatterplot where each point corresponds to the ith pair $(\hat{y}_i, \hat{e}_i)$ for $i = 1, \ldots, n$. The vertical distance from each point to the zero-line is the value of each residual. If the model was correctly specified you will expect to see:

- A null plot, that is, a scatterplot where the only pattern is a horizontal line (a slope equal to zero) and no curvature is present. The red line provides the `loess` smoother that could be a useful reference about curvature, but it's sensitive to isolated and extreme observations, so only use it as reference and not as the conclusive tool to determine if the model is well specified. In our example, the plot suggest that the data fit well the linearity assumption of the mean function.
- A (vertical) dispersion that is more or less constant for any levels of the fitted values. This is the assumption of homoskedasticity or constancy of variance. When evaluating homoskedasticity, always take into account the number of points for any given interval, as it is expected to have more variation if you have more points. If the variance is constant and the fitted values are more or less symmetric around the mean, the residual plot has the shape of an ellipsoid. In our example, again, we do not see any evidence against homoskedasticity.

4.5 Coefficient of Determination

The total sum of squares for the response,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

can be decomposed in the regression sum of squares plus the residual sum of squares.[9]

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

A useful statistic using in linear regression that makes use of this decomposition is the coefficient of determination, R^2, defined as

[9] To show this you just need basic algebra

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

The coefficient of determination explains how much variation of the response is explained by the regressors. Due to this decomposition observe that R^2 will always be positive and between 0 and 1, where the closer to 1, the higher the variation explained by the regressors.

Observe for example, using the `Heights` data,

```
summary(mod1)
```

```
Call:
lm(formula = dheight ~ mheight, data = Heights)

Residuals:
   Min     1Q Median     3Q    Max
-7.397 -1.529  0.036  1.492  9.053

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.917      1.623    18.4   <2e-16
mheight        0.542      0.026    20.9   <2e-16

Residual standard error: 2.27 on 1373 degrees of freedom
Multiple R-squared:  0.241, Adjusted R-squared:  0.24
F-statistic:  435 on 1 and 1373 DF,  p-value: <2e-16
```

The coefficient of determination (called `Multiple R-squared` in the output) is 0.24, meaning about 24% of the variation of daughter's height, `dheight`, is explained by mother's height; quite a low value. You can also obtain this result directly from the summary of the `lm()` object:

```
summary(mod1)$r.sq
```

```
[1] 0.2408
```

4.6 Example of Multiple Linear Regression: Fuel Consumption Data

Let's recap what we've learned so far by using the dataframe `fuel2001` from package `alr4` (Note: you need to install this package if you want to run the code in R). These data contain information from US States (and District of Columbia) on motor fuel consumption and related variables. We are interested in studying how fuel consumption changes due to other variables. Here is the description of relevant variables:

- Drivers: Number of Licensed drivers in the state
- FuelC: Gasoline sold for road use in thousands of gallons
- Income: Per capita personal income (year 2000)
- Miles: Miles of Federal-aid highway miles in the state
- Pop: Population age 16 and over
- Tax: Gasoline state tax rate in cents per gallon

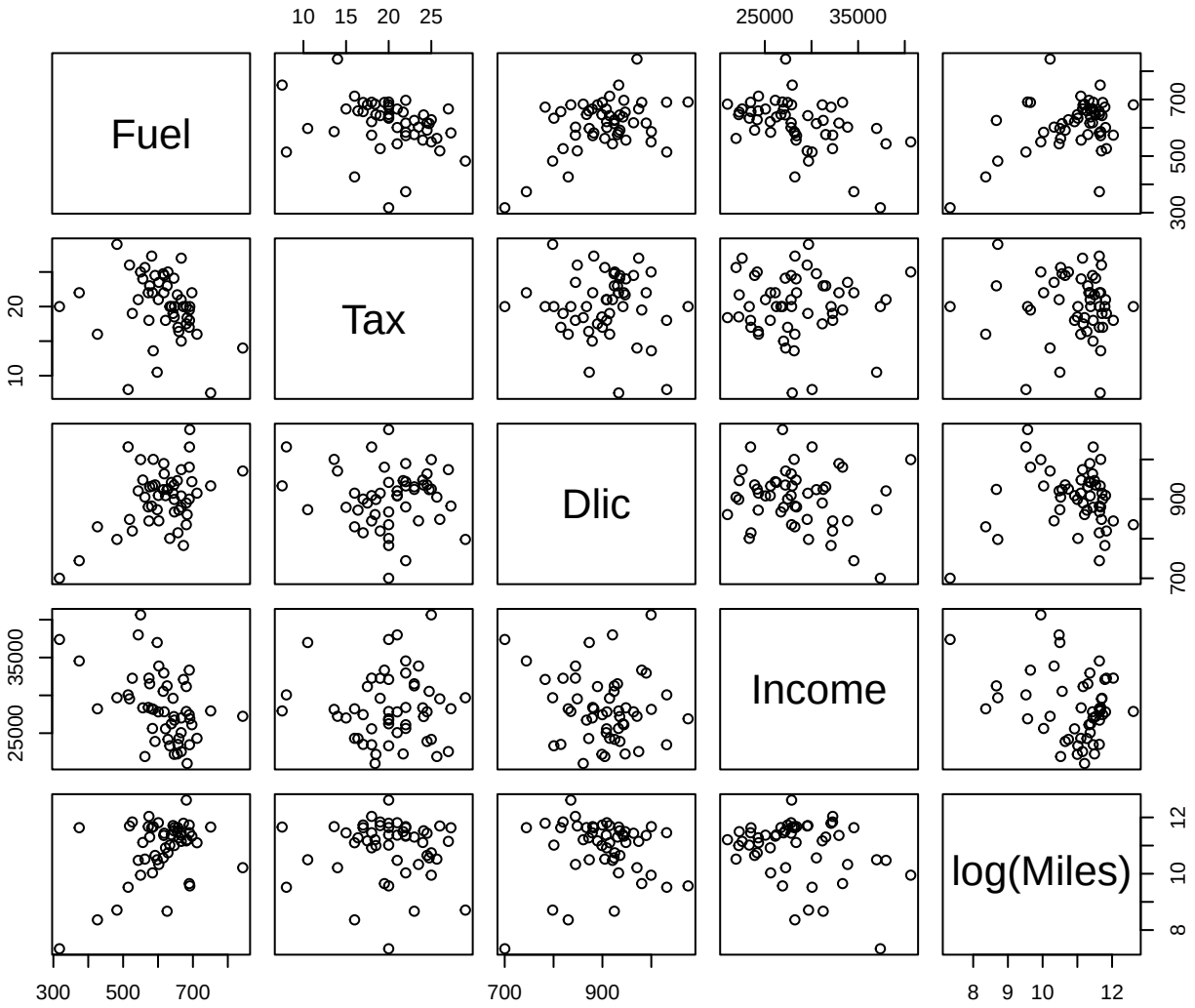
And before starting, let's transform the data in relevant ways

```
data(fuel2001, package = "alr4")
fuel2001 <- transform(fuel2001,
                      Dlic=1000 * Drivers/Pop,
                      Fuel=1000 * FuelC/Pop)
head(fuel2001)
```

```
    Drivers    FuelC Income  Miles     MPC      Pop  Tax    Dlic   Fuel
AL  3559897  2382507  23471  94440 12737.0  3451586 18.0 1031.38 690.26
AK   472211   235400  30064  13628  7639.2   457728  8.0 1031.64 514.28
AZ  3550367  2428430  25578  55245  9411.5  3907526 18.0  908.60 621.48
AR  1961883  1358174  22257  98132 11268.4  2072622 21.7  946.57 655.29
CA 21623793 14691753  32275 168771  8923.9 25599275 18.0  844.70 573.91
CO  3287922  2048664  32949  85854  9722.7  3322455 22.0  989.61 616.61
```

The variable `Dlic` is the proportion of drivers per state times 1000 (to preserve more information), the variable `Fuel` is consumption per capita of gasoline in gallons. To visualize the relationship of the response, `Fuel`, with each regressor, we can produce a scatterplot matrix:

```
pairs(Fuel ~ Tax + Dlic + Income + log(Miles), data=fuel2001)
```



The scatterplots in the first row are the most relevant ones, as they show all relevant regressors against the response, `Fuel`. Observe that, while the relationships does not look particularly strong, there are not clear non-linear relationships that may lead us to try to use remedial measures. Now, let's obtain the results from this model

```
m.fuel = lm(Fuel ~ Dlic + Income + Tax + log(Miles) , data = fuel2001)
summary (m.fuel)
```

```
Call:
lm(formula = Fuel ~ Dlic + Income + Tax + log(Miles), data = fuel2001)

Residuals:
    Min      1Q  Median      3Q     Max
-163.14  -33.04    5.89   31.99  183.50

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.19284  194.90616    0.79  0.43294
Dlic          0.47187    0.12851    3.67  0.00063
Income       -0.00614    0.00219   -2.80  0.00751
Tax          -4.22798    2.03012   -2.08  0.04287
log(Miles)   26.75518    9.33737    2.87  0.00626

Residual standard error: 64.9 on 46 degrees of freedom
Multiple R-squared:  0.51,  Adjusted R-squared:  0.468
F-statistic:   12 on 4 and 46 DF,  p-value: 0.000000933
```

Let `Fuel` be represented by Y and `DLic`, `Income`, `Tax`, and `log(Miles)` be represented by X_1, X_2, X_3 and X_4, respectively. Based on this output, the estimated linear regression model is

$$\widehat{E}(Y|X_1, X_2, X_3, X_4) = 154.19 + (0.47)X_1 + -0.01X_2 + (-4.23)X_3 + 26.76X_4.$$

Let's interpret the coefficient estimate for `Tax`, $\hat{\beta}_3 = -4,23$; if the tax rate increases in one cent per gallon, the consumption of gasoline will decrease, on average, by 4.23 gallons per capita, keeping `Dlic`, `Income`, and `log(Miles)` constant. Assume that we would like to test if changes in taxes levied to gasoline consumption could lead to a change in gasoline consumption per capita, then the test we want to test is

$$\begin{aligned} H_0 &: \beta_3 = 0 \qquad \text{for arbitraty } \beta_1, \beta_2, \text{ and } \beta_4 \\ H_1 &: \beta_3 \neq 0 \qquad \text{for arbitraty } \beta_1, \beta_2, \text{ and } \beta_4 \end{aligned}$$

Based on the output, the p-value for this test is 0.04, marginally significant and perhaps could allow us to reject the null hypothesis and conclude that increase in taxes could lead to change in gasoline consumption.

On the other hand, perhaps a more relevant question would have been to try to find evidence that increases in taxes typically **reduce** gasoline consumption. If this would be the case, then the hypotheses are:

$$\begin{aligned} H_0 &: \beta_3 \geq 0 \qquad \text{for arbitraty } \beta_1, \beta_2, \text{ and } \beta_4 \\ H_1 &: \beta_3 < 0 \qquad \text{for arbitraty } \beta_1, \beta_2, \text{ and } \beta_4 \end{aligned}$$

The OLS estimate for β_3 and the standard error of $\hat{\beta}_3$ do not depend on the test (they are obtained from the sample). Moreover, the test statistic is affected by the value of β_3 under the null hypothesis, which for the purpose of the test, it is the same value given in the output. On the other hand, this is now a left-tailed test, so the p-value is given by

```
pvalue = pt(-2.083, 51 - 5)
pvalue
```

```
[1] 0.021419
```

Observe now that the p-value is 0.021, exactly half of what it was for the two-tailed test (as expected), and the result is statistically significant. Based on this test we have found evidence that an increase in taxes for gasoline will typically reduce gasoline consumption.

4.6.1 Confidence Intervals

Let's find 97% confidence intervals for the coefficients of the Fuel linear model:

```
confint(m.fuel, level = 0.97)
```

```
                  1.5 %      98.5 %
(Intercept) -282.298408 590.6840968
Dlic           0.184066   0.7596763
Income        -0.011048  -0.0012227
Tax           -8.774428   0.3184616
log(Miles)     5.844179  47.6661722
```

We are 97% confident that β_3, the rate of change in gasoline consumption due to an increase in tax for gas, is a number between -8.77 and 0.32. Observe that, because the interval includes zero, it is plausible that $\beta_3 = 0$. This result seem in contradiction with our previous result, when we rejected the hypothesis $H_0 : \beta_3 = 0$, but actually the results are equivalent as long as you use the corresponding significance level, α. Redo the previous test of significance using $\alpha = 1 - 0.97 = 0.03$ and show that in fact, both conclusions are equivalent.

4.6.2 Prediction

Let's observe model `m.fuel` to remember the variable names used:

```
m.fuel
```

```

Call:
lm(formula = Fuel ~ Dlic + Income + Tax + log(Miles), data = fuel2001)

Coefficients:
(Intercept)         Dlic       Income          Tax   log(Miles)
  154.19284      0.47187     -0.00614     -4.22798     26.75518
```

Let's find the 99% prediction interval for a "new" state that has `Tax`= 25 cents, `Dlic` = 950 (.95 of people who can drive has a driver license), `Income` = 30000 dollars of income per capita, and `Miles` = 160000 Federal-aid highway miles in this state.

```
new.data.fuel <- data.frame(Tax = 25, Dlic = 950, Income = 30000, Miles = 160000)
predict(object = m.fuel, newdata = new.data.fuel, interval = "prediction", level = 0.99)
```

```
     fit    lwr    upr
1 633.32 451.22 815.41
```

We are 99% confident that the amount of fuel consumption per capita for this "new" state will be between 451 and 815 gallons per person.