STA238 Tutorial 7

Luis Ledesma

2023-03-15

1 Announcements

- You can upload your work on Crowdmark from the end of the tutorial session to 5pm Friday of that week.
- All questions must be solved using RStudio.

2 Recall: Last tutorial

Last tutorial: We introduced and derived the method of moments estimator for a two-parameter distribution function, Gamma(α, β).

Main takeaways:

1. The method of moments estimator is a technique that will propose an estimator $\hat{\theta}$ of θ, where θ is the parameter for a probability density function $f(x; \theta)$.
2. These estimators are usually consistent (meaning that $\hat{\theta}$ will converge to θ in probability for a large enough sample), but **they can be biased.**
3. To conduct method of moments:

a. Usually, compute k moments, $E(X), \dots, E(X^k)$ where k is the dimension of θ (for the Gamma distribution, $k = 2$). Write out their expression in terms of θ, one should obtain a system of k equations.

b. Substitute $E(X), \dots, E(X^k)$ by $\overline{X}, \dots, \overline{X^k}$ in the system of equations.

c. Substitute θ by $\hat{\theta}$ in the system of equations.

d. Solve for $\hat{\theta}$ in terms of $\overline{X^k}$. This is the method of moments estimator.

3 Tutorial activity

We want to:

1. Visualize two-dimensional data.
2. Conduct linear regression for a sample.
3. Interpret the regression coefficient.
4. Compute the coefficient of determination R^2 and the correlation coefficient r.

3.1 Visualizing the dataset

Given the problem statement, we will code the database:

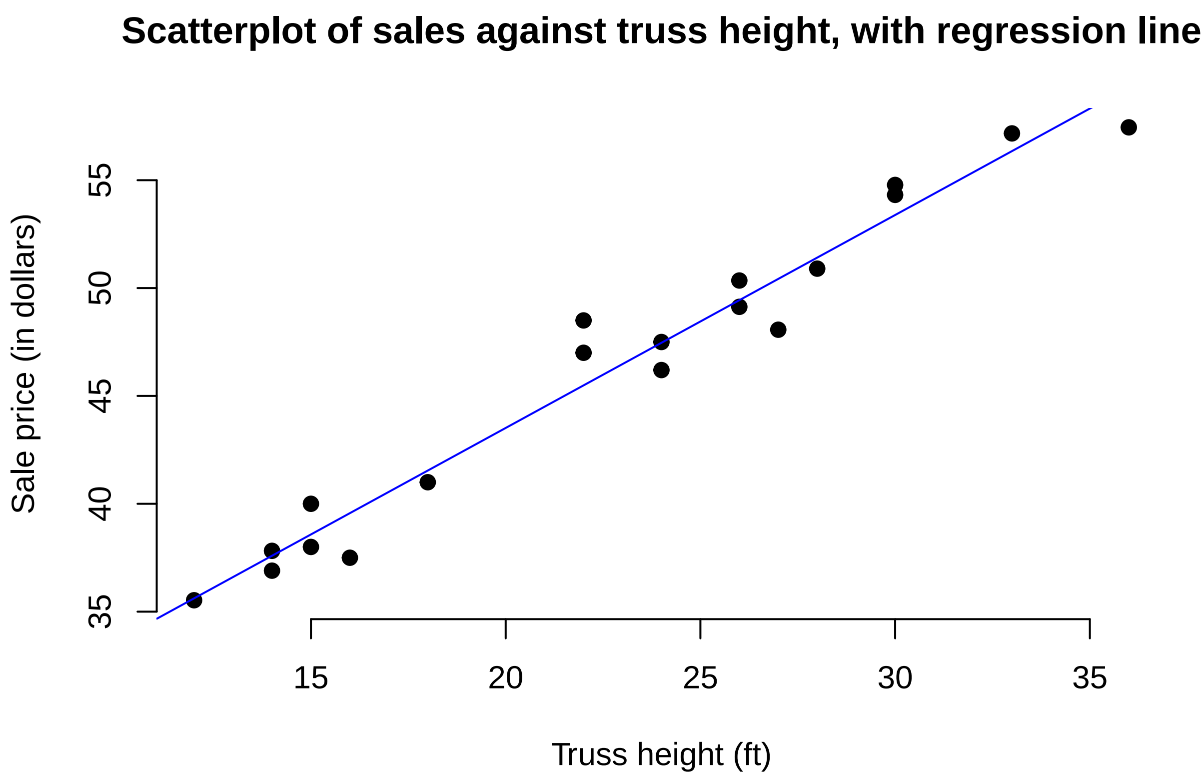
```
x <- c(12, 14, 14, 15, 15, 16, 18, 22, 24, 26, 26, 27, 28, 30, 30, 33, 22, 24, 36)

y <- c(35.53, 37.82, 36.90, 40.00, 38.00, 37.50, 41.00, 48.50, 46.20, 50.35,
       49.13, 48.07, 50.90, 54.78, 54.32, 57.17, 47.00, 47.50, 57.45)
```

We can plot y against x in a scatterplot:

```
plot(x, y, main = "Scatterplot of sales against truss height",
     xlab = "truss height (ft)", ylab = "Sale price",
     pch = 19, frame = FALSE)
```

It does seem that a linear regression may be appropriate for this model.

```
plot(x, y, main = "Scatterplot of sales against truss height, with regression line",
     xlab = "Truss height (ft)", ylab = "Sale price (in dollars)",
     pch = 19, frame = FALSE)
abline(lm(y ~ x), col = "blue")
```



3.2 Obtaining the regression coefficients

To obtain the regression coefficients:

```
linear_model <- lm(y ~ x)
summary(linear_model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.35522 -0.63584 -0.08796  0.92263  3.01053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.77215    1.11347   21.35 1.03e-13 ***
## x            0.98715    0.04684   21.07 1.27e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.416 on 17 degrees of freedom
## Multiple R-squared:  0.9631, Adjusted R-squared:  0.961
## F-statistic: 444.1 on 1 and 17 DF,  p-value: 1.271e-13
```

The `summary` function will give us the coefficients of the linear model, along with test statistics and p-values. Truss height seems to be a significant predictor of sales price, with a one feet increase in height resulting in close to an increase of 1 dollar in sales price.

3.3 Obtaining point estimates from the model

Notice that the model we have fit will be (where x is height):

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

Remark: $\hat{y}$ will be the predicted sales/average value in dollars from the model (not the true value!)

In our formula, if we plug in $x = 25$:

```
predict(linear_model, newdata = data.frame(x=c(25)))
```

```
##        1
## 48.45092
```

Thus, 48.45 will be the average predicted sales value in dollars for a truss height of 25 fit.

3.4 Obtaining the coefficients of determination R^2 and correlation r

To obtain the coefficient of determination R^2, one can extract it from the model output:

```
summary(linear_model)$r.squared
```

```
## [1] 0.9631328
```

In contrast, the correlation coefficient r will be a sample statistic obtained from the data:

```
cor(x, y)
```

```
## [1] 0.9813933
```