Chapter 2: simple regression

We start again by loading the data.

```
# Load the datasets from the RData file
load("../dataCreated/listings_clean.RData")

# Display the first few rows of the cleaned dataset
head(listings_clean)
```

```
##   neighbourhood_cleansed           id
## 1          's-Gravenland 1.012115e+18
## 2          's-Gravenland 1.032655e+18
## 3           Agniesebuurt 8.217094e+17
## 4           Agniesebuurt 9.846371e+17
## 5           Agniesebuurt 7.665157e+17
## 6           Agniesebuurt 6.455105e+17
##                                      listing_url    scrape_id last_scraped
## 1 https://www.airbnb.com/rooms/1012114818168132512 2.024063e+13   2024-06-25
## 2 https://www.airbnb.com/rooms/1032655139418809797 2.024063e+13   2024-06-25
## 3  https://www.airbnb.com/rooms/821709375116932602 2.024063e+13   2024-06-26
## 4  https://www.airbnb.com/rooms/984637143310796841 2.024063e+13   2024-06-25
## 5  https://www.airbnb.com/rooms/766515651027141909 2.024063e+13   2024-06-25
## 6  https://www.airbnb.com/rooms/645510482091510278 2.024063e+13   2024-06-25
##        source                                                name
## 1 city scrape    High-end Serviced Apartment with two bedrooms
## 2 city scrape     High-end Serviced Apartment with one bedroom
## 3 city scrape                                        King Room
## 4 city scrape            Charming Urban Retreat - with parking
## 5 city scrape                       Bed in Mixed Dormitory Room
## 6 city scrape Studio floor with own kitchen & toilet - central
##
## 1
## 2
## 3
## 4              Indulge in the allure of our charming centrally located urban retreat. Enjoy the sunl
## 5                                                                 Sleeping in a dorm room is not or
## 6 10MIN WALK FROM CENTRAL.<br />We have 2 floor apartment and we are renting out our upstairs which i
##
## 1
## 2
## 3
## 4
## 5 Within walking distance to Rotterdam Centraal Station and just a short drive from Rotterdam The Hag
## 6
##
## 1 https://a0.muscache.com/pictures/miso/Hosting-1012114818168132512/original/12df4128-3a61-42a2-a47a-
## 2 https://a0.muscache.com/pictures/miso/Hosting-1032655139418809797/original/69761ba0-3d06-4760-a1e3-
## 3  https://a0.muscache.com/pictures/miso/Hosting-821709375116932602/original/3e333f5e-b739-4f17-873f-
```

```
## 4  https://a0.muscache.com/pictures/miso/Hosting-984637143310796841/original/b16e8b3a-e987-4119-b146-
## 5  https://a0.muscache.com/pictures/miso/Hosting-766515651027141909/original/7b624fb0-bb29-4055-a801-
## 6                                                https://a0.muscache.com/pictures/b3738de7-f406-438a-ab2
##     host_id                                    host_url           host_name
## 1 543381697 https://www.airbnb.com/users/show/543381697 BirdsEye Short Stay
## 2 543381697 https://www.airbnb.com/users/show/543381697 BirdsEye Short Stay
## 3 365217950 https://www.airbnb.com/users/show/365217950              Bianca
## 4 116041691 https://www.airbnb.com/users/show/116041691            Bastiaan
## 5 365217950 https://www.airbnb.com/users/show/365217950              Bianca
## 6  16080777  https://www.airbnb.com/users/show/16080777               Sevim
##   host_since          host_location
## 1 2023-10-25                   <NA>
## 2 2023-10-25                   <NA>
## 3 2020-08-31                   <NA>
## 4 2017-02-12 Rotterdam, Netherlands
## 5 2020-08-31                   <NA>
## 6 2014-05-28 Rotterdam, Netherlands
##
## 1
## 2
## 3
## 4
## 5
## 6 We are a young couple living in Rotterdam both working in corporate. I am Turkish and my boyfriend
##   host_response_time host_response_rate host_acceptance_rate host_is_superhost
## 1     within an hour                96%                  82%              TRUE
## 2     within an hour                96%                  82%              TRUE
## 3     within an hour                85%                  99%             FALSE
## 4     within an hour               100%                  94%              TRUE
## 5     within an hour                85%                  99%             FALSE
## 6 within a few hours               100%                  21%             FALSE
##
## 1 https://a0.muscache.com/im/pictures/user/User-543381697/original/48eae732-3db4-4508-9f9c-ca2a133bd4
## 2 https://a0.muscache.com/im/pictures/user/User-543381697/original/48eae732-3db4-4508-9f9c-ca2a133bd4
## 3                          https://a0.muscache.com/im/pictures/user/cc6977b4-d482-4111-9460-af2fb8e597
## 4                          https://a0.muscache.com/im/pictures/user/db5760d4-54b8-467b-949a-45f95150a0
## 5                          https://a0.muscache.com/im/pictures/user/cc6977b4-d482-4111-9460-af2fb8e597
## 6                          https://a0.muscache.com/im/pictures/user/0d800f6b-1061-4a87-b4d3-cd0d70671a
##
## 1 https://a0.muscache.com/im/pictures/user/User-543381697/original/48eae732-3db4-4508-9f9c-ca2a133bd4
## 2 https://a0.muscache.com/im/pictures/user/User-543381697/original/48eae732-3db4-4508-9f9c-ca2a133bd4
## 3                          https://a0.muscache.com/im/pictures/user/cc6977b4-d482-4111-9460-af2fb8e597
## 4                          https://a0.muscache.com/im/pictures/user/db5760d4-54b8-467b-949a-45f95150a0
## 5                          https://a0.muscache.com/im/pictures/user/cc6977b4-d482-4111-9460-af2fb8e597
## 6                          https://a0.muscache.com/im/pictures/user/0d800f6b-1061-4a87-b4d3-cd0d70671a
##   host_neighbourhood host_listings_count host_total_listings_count
## 1               <NA>                   2                         2
## 2               <NA>                   2                         2
## 3               <NA>                  13                        13
## 4               <NA>                   1                         1
## 5               <NA>                  13                        13
## 6               <NA>                   2                         2
##                host_verifications host_has_profile_pic host_identity_verified
## 1               ['email', 'phone']                 TRUE                   TRUE
```

```
## 2               ['email', 'phone']                 TRUE                   TRUE
## 3                        ['phone']                 TRUE                   TRUE
## 4 ['email', 'phone', 'work_email']                 TRUE                   TRUE
## 5                        ['phone']                 TRUE                   TRUE
## 6 ['email', 'phone', 'work_email']                 TRUE                   TRUE
##                          neighbourhood neighbourhood_group_cleansed latitude
## 1 Rotterdam, Zuid-Holland, Netherlands              Prins Alexander 51.92200
## 2 Rotterdam, Zuid-Holland, Netherlands              Prins Alexander 51.92044
## 3                                 <NA>                        Noord 51.92851
## 4                                 <NA>                        Noord 51.92876
## 5 Rotterdam, Zuid-Holland, Netherlands                        Noord 51.92811
## 6 Rotterdam, Zuid-Holland, Netherlands                        Noord 51.93126
##   longitude        property_type       room_type accommodates bathrooms
## 1  4.553830   Entire rental unit Entire home/apt            4       1.0
## 2  4.554210   Entire rental unit Entire home/apt            2       1.0
## 3  4.474124        Room in hotel    Private room            2       1.0
## 4  4.474194   Entire rental unit Entire home/apt            4       1.0
## 5  4.475240 Shared room in hotel     Shared room            1       1.0
## 6  4.474320         Entire condo Entire home/apt            2       1.5
##   bathrooms_text bedrooms beds
## 1         1 bath        2    2
## 2         1 bath        1    1
## 3 1 private bath        1    1
## 4         1 bath        2    3
## 5  1 shared bath        1    2
## 6      1.5 baths        1    1
##
## 1 ["Private patio or balcony", "Hot water kettle", "Central heating", "Paid street parking off premis
## 2 ["Private patio or balcony", "Hot water kettle", "Central heating", "Paid street parking off premis
## 3
## 4
## 5
## 6
##   price minimum_nights maximum_nights minimum_minimum_nights
## 1   259             3            20                      2
## 2   204             3            20                      2
## 3   198             1           365                      1
## 4   253             2           365                      2
## 5    75             1           365                      1
## 6   116             7            60                      2
##   maximum_minimum_nights minimum_maximum_nights maximum_maximum_nights
## 1                      3                    20                    20
## 2                      3                    20                    20
## 3                      1                  1000                  1000
## 4                     27                   365                   500
## 5                      1                  1000                  1000
## 6                      7                    60                    60
##   minimum_nights_avg_ntm maximum_nights_avg_ntm calendar_updated
## 1                    2.2                  20.0               NA
## 2                    2.2                  20.0               NA
## 3                    1.0                1000.0               NA
## 4                    5.9                 385.8               NA
## 5                    1.0                1000.0               NA
## 6                    6.8                  60.0               NA
```

```
##   has_availability availability_30 availability_60 availability_90
## 1             TRUE              17              32              59
## 2             TRUE              26              56              86
## 3             TRUE              28              58              88
## 4             TRUE              15              20              20
## 5             TRUE              23              51              81
## 6             TRUE              11              24              54
##   availability_365 calendar_last_scraped number_of_reviews
## 1              334           2024-06-25                89
## 2              361           2024-06-25                87
## 3              363           2024-06-26                12
## 4               20           2024-06-25                32
## 5              356           2024-06-25                55
## 6              144           2024-06-25                16
##   number_of_reviews_ltm number_of_reviews_l30d first_review last_review
## 1                    89                    19   2023-12-29  2024-06-20
## 2                    87                     8   2023-12-31  2024-06-16
## 3                     9                     1   2023-04-23  2024-05-30
## 4                    32                     2   2023-10-14  2024-06-02
## 5                    34                     0   2022-12-12  2024-05-18
## 6                     3                     0   2022-06-18  2024-04-03
##   review_scores_rating review_scores_accuracy review_scores_cleanliness
## 1                 4.84                  4.81                       4.89
## 2                 4.86                  4.86                       4.90
## 3                 4.42                  4.67                       4.33
## 4                 4.81                  4.81                       4.72
## 5                 4.33                  4.47                       4.47
## 6                 4.94                  4.88                       4.88
##   review_scores_checkin review_scores_communication review_scores_location
## 1                  4.76                       4.85                    4.60
## 2                  4.69                       4.83                    4.69
## 3                  4.67                       4.00                    4.08
## 4                  4.91                       4.84                    4.84
## 5                  4.76                       4.44                    4.64
## 6                  4.94                       5.00                    4.81
##   review_scores_value                 license instant_bookable
## 1                4.71 0599 F23E 3F4E 6EF1 B897           FALSE
## 2                4.72 0599 9BD6 36FC CB31 F331            TRUE
## 3                3.83                  Exempt            TRUE
## 4                4.69 0599 8A16 99F9 742E C8F8           FALSE
## 5                4.36                  Exempt            TRUE
## 6                4.63 0599 B71E 378D 56FF 1829           FALSE
##   calculated_host_listings_count calculated_host_listings_count_entire_homes
## 1                              2                                           2
## 2                              2                                           2
## 3                             13                                           0
## 4                              1                                           1
## 5                             13                                           0
## 6                              2                                           2
##   calculated_host_listings_count_private_rooms
## 1                                            0
## 2                                            0
## 3                                           12
## 4                                            0
```

```
## 5                                            12
## 6                                             0
##   calculated_host_listings_count_shared_rooms reviews_per_month Coolness
## 1                                           0             14.83        6
## 2                                           0             14.66        6
## 3                                           1              0.84        6
## 4                                           0              3.75        6
## 5                                           1              2.94        6
## 6                                           0              0.65        6
##   Centrality Quietness Fanciness
## 1          5         6         5
## 2          5         6         5
## 3          6         7         5
## 4          6         7         5
## 5          6         7         5
## 6          6         7         5
```

Filter the data so that we use only listings for at most 6 people. From now on use this

```
listings_clean_filtered <- listings_clean %>%
  filter(accommodates <= 6)
```

Regress price on `review_scores_rating`.

```
# Run a linear regression of price on review_scores_rating
model <- lm(price ~ review_scores_rating, data = listings_clean_filtered)

# Display the summary of the regression model
summary(model)
```

```
##
## Call:
## lm(formula = price ~ review_scores_rating, data = listings_clean_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.965  -43.824   -7.964   32.321  163.448
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)             63.92      63.57   1.005    0.315
## review_scores_rating    14.27      13.37   1.067    0.287
##
## Residual standard error: 58.09 on 419 degrees of freedom
## Multiple R-squared:  0.00271,    Adjusted R-squared:  0.0003294
## F-statistic: 1.138 on 1 and 419 DF,  p-value: 0.2866
```

Run a regression of price on how many people can be accommodated.

```
# Run a linear regression of price on accommodates
model_accommodates <- lm(price ~ accommodates, data = listings_clean_filtered)

# Display the summary of the regression model
summary(model_accommodates)
```

```
##
## Call:
```

```
## lm(formula = price ~ accommodates, data = listings_clean_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -144.65  -33.46  -10.91   22.35  159.90
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    60.731      5.735   10.59   <2e-16 ***
## accommodates   25.183      1.855   13.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.48 on 419 degrees of freedom
## Multiple R-squared:  0.3055, Adjusted R-squared:  0.3038
## F-statistic: 184.3 on 1 and 419 DF,  p-value: < 2.2e-16
```

On slide 21, we have said that the sample moment conditions immediately imply that: 1. the average residual is zero 2. the average covariance between the residual and the regressor is zero 3. the regression line goes through the average of the dependent variable and the regressor.

We now verify this.

```
# 1. Verify that the average residual is zero
residuals_accommodates <- residuals(model_accommodates)
mean_residuals <- round(mean(residuals_accommodates),5)
cat("1. Average of residuals:", mean_residuals, "\n")
```

```
## 1. Average of residuals: 0
```

```
# 2. Verify that the average covariance between the residual and the regressor (accommodates) is zero
cov_residuals_regressor <- round(cov(residuals_accommodates, listings_clean_filtered$accommodates),5)
cat("2. Covariance between residuals and accommodates:", cov_residuals_regressor, "\n")
```

```
## 2. Covariance between residuals and accommodates: 0
```

```
# 3. Verify that the regression line goes through the averages of the dependent variable and the regres
mean_price <- mean(listings_clean_filtered$price)
mean_accommodates <- mean(listings_clean_filtered$accommodates)

# Calculate the predicted price at the average of accommodates
predicted_price_at_mean <- coef(model_accommodates)[1] + coef(model_accommodates)[2] * mean_accommodates

cat("3. Average price:", mean_price, "\n")
```

```
## 3. Average price: 131.6746
```

```
cat("   Predicted price at average accommodates:", predicted_price_at_mean, "\n")
```

```
##    Predicted price at average accommodates: 131.6746
```

We have also said that the total sum of squares for the dependent variable is equal to the explained sum of squares plus the residual sum of squares. We can also verify this.

```
# 1. Calculate the Total Sum of Squares (TSS)
mean_price <- mean(listings_clean_filtered$price)
TSS <- sum((listings_clean_filtered$price - mean_price)^2)
cat("Total Sum of Squares (TSS):", TSS, "\n")
```

```
## Total Sum of Squares (TSS): 1417766
```

```
# 2. Calculate the Residual Sum of Squares (RSS)
RSS <- sum(residuals_accommodates^2)
cat("Residual Sum of Squares (RSS):", RSS, "\n")
```

```
## Residual Sum of Squares (RSS): 984669.4
```

```
# 3. Calculate the Explained Sum of Squares (ESS)
# ESS is the difference between TSS and RSS
ESS <- TSS - RSS
cat("Explained Sum of Squares (ESS):", ESS, "\n")
```

```
## Explained Sum of Squares (ESS): 433097
```

```
# 4. Verify that TSS = ESS + RSS
cat("TSS == ESS + RSS:", TSS == (ESS + RSS), "\n")
```

```
## TSS == ESS + RSS: TRUE
```

From this we can also compute the R^2 measure by hand.

```
# 1. Calculate Total Sum of Squares (TSS)
mean_price <- mean(listings_clean_filtered$price)
TSS <- sum((listings_clean_filtered$price - mean_price)^2)

# 2. Calculate Residual Sum of Squares (RSS)
RSS <- sum(residuals_accommodates^2)

# 3. Calculate Explained Sum of Squares (ESS)
ESS <- TSS - RSS

# 4. Calculate R^2 using the formula R^2 = ESS / TSS or 1 - RSS / TSS
R_squared <- 1 - (RSS / TSS)

# Display the result
cat("R^2 (by hand):", round(R_squared,4), "\n")
```

```
## R^2 (by hand): 0.3055
```

It is the same as above in the regressian output (which of course has to be the case)!

Next we use the Stargazer package that we loaded in the beginning to produce a nice table with regression results for three different specifications. See slide 24 for details.

```
# 1. Run a level-level regression of price on accommodates
level_level <- lm(price ~ accommodates, data = listings_clean_filtered)

# 2. Run a log-level regression of log(price) on accommodates
log_level <- lm(log(price) ~ accommodates, data = listings_clean_filtered)

# 3. Run a log-log regression of log(price) on log(accommodates)
log_log <- lm(log(price) ~ log(accommodates), data = listings_clean_filtered)

# 4. Create a table with stargazer displaying all three results
stargazer(level_level, log_level, log_log,
          type = "text",
          title = "Regression Results",
          column.labels = c("level-level", "log-level", "log-log"),
```

```
          dep.var.labels = "price",
          covariate.labels = c("accommodates", "log(accommodates)", "constant"),
          omit.stat = c("f", "ser"),
          digits = 3)
```

```
##
## Regression Results
## ================================================
##                          Dependent variable:
##                    ------------------------------
##                       price        log(price)
##                    level-level log-level log-log
##                        (1)        (2)      (3)
## ------------------------------------------------
## accommodates       25.183***  0.210***
##                     (1.855)    (0.015)
##
## log(accommodates)                         0.639***
##                                           (0.040)
##
## constant           60.731***  4.186***  4.180***
##                     (5.735)    (0.046)  (0.041)
##
## ------------------------------------------------
## Observations          421        421      421
## R2                   0.305      0.322    0.381
## Adjusted R2          0.304      0.320    0.379
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Finally, we produce a figure with the data points and the three fitted regression lines.[1]

```
# 1. Create a new data frame with predictions
listings_clean_filtered$pred_level_level <- predict(level_level, newdata = listings_clean_filtered)
listings_clean_filtered$pred_log_level <- exp(predict(log_level, newdata = listings_clean_filtered))
listings_clean_filtered$pred_log_log <- exp(predict(log_log, newdata = listings_clean_filtered))

# 2. Calculate the average price for each value of accommodates
avg_prices <- listings_clean_filtered %>%
  group_by(accommodates) %>%
  summarise(avg_price = mean(price))

# 3. Scatter plot of the actual data (price vs accommodates)
p <- ggplot(listings_clean_filtered, aes(x = accommodates, y = price)) +
  geom_point(alpha = 0.5) +
  labs(title = "Estimated Empirical Relationships",
       x = "Accommodates",
       y = "Price") +
  theme_minimal()

# 4. Add the fitted values from the level-level model
p <- p + geom_line(aes(y = pred_level_level, color = "level-level"), linewidth = 1)
```

[1] Here, we ignore an issue that the 7th edition of Wooldrige's book discusses on p. 206ff. In brief, the issue is that the predicted value of the dependent variable is not the expenential function evaluated at the fitted value when the dependent variable in the regression was in log form. We ignore this issue here for simplicity.

```r
# 5. Add the fitted values from the log-level model
p <- p + geom_line(aes(y = pred_log_level, color = "log-level"), linewidth = 1)

# 6. Add the fitted values from the log-log model
p <- p + geom_line(aes(y = pred_log_log, color = "log-log"), linewidth = 1)

# 7. Add the average price for each accommodates value as red squares
p <- p + geom_point(data = avg_prices, aes(x = accommodates, y = avg_price),
                    color = "red", shape = 15, size = 3)

# 8. Add legend and show the plot
p <- p + scale_color_manual(name = "model",
                            values = c("level-level" = "blue", "log-level" = "green", "log-log" = "red"))

# Display the plot
print(p)
```