Statistics 2 2024/25 Resit

Introduction

The dataset `toyota-corolla.csv` contains data from a random sample of advertisements for used Toyota Corolla cars on the website Craigslist, which is the American equivalent of Marktplaats. Here is an example advertisement from the site:

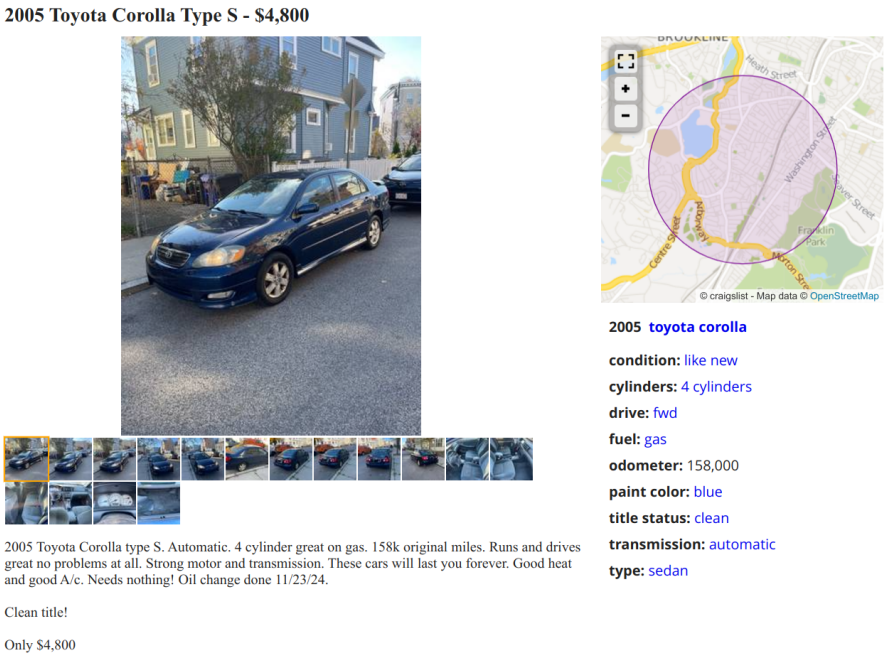

Figure 1: Example Advertisement

There are 857 observations and 7 variables in the dataset:

- `price`: The asking price in dollars for the car in the advertisement.
- `age`: The age of the vehicle in years since its manufacture.
- `odometer`: The total number of miles the car has driven in its lifetime.

This is usually shown on the car's dashboard. *Note:* 1 mile is 1.609344 kilometers.
- `state_ca`: A dummy variable that equals 1 if the car advertisement was placed in the state of California and 0 otherwise.
- `state_fl`: A dummy variable that equals 1 if the car advertisement was placed in the state of Florida and 0 otherwise.
- `state_ny`: A dummy variable that equals 1 if the car advertisement was placed in the state of New York and 0 otherwise.
- `state_tx`: A dummy variable that equals 1 if the car advertisement was placed in the state of Texas and 0 otherwise.

All the advertisements are from one of California, Florida, New York or Texas.

Question 1

What is the sample correlation between `odometer` and `price`?

Question 2

The sample correlation between `age` and `price` is -0.7405305. Choose the answer below which best interprets this correlation.

- A car that is older by one year will on average sell for $0.74 less.
- A car that is newer by one year will on average sell for $0.74 less.
- Older cars tend to have smaller asking prices and newer cars tend to have larger asking prices.
- Older cars tend to have larger asking prices and newer cars tend to have smaller asking prices.

Model 1

Estimate a simple linear regression model with `price` as the dependent variable and `odometer` as the independent variable.

Your estimated sample regression intercept should be 17970.519227.

Use this model to answer the following questions.

Question 3

Choose the answer below which best interprets the sample regression intercept.

- The predicted asking price of a Toyota Corolla is $17970.52.
- All Toyota Corollas sell for $17970.52.
- The predicted asking price of a Toyota Corolla that has driven zero miles is $17970.52.
- All Toyota Corollas that have not driven any miles sell for $17970.52.

Question 4

Report the sample regression slope.

Question 5

Consider two Toyota Corollas, Car A and Car B. Car A has driven 100,000 miles in its lifetime while Car B has driven only 90,000.

According to the model, by how much more is Car B expected to sell for in the market compared to Car A?

In other words, calculate "Predicted asking price car B" - "Predicted asking price car A".

Question 6

Run a regression of the squared residuals against `odometer` to formally test the homoskedasticity assumption in the model. Use a 5% level for the test.

What is the null hypothesis? Choose one of the following options:

- There is no heteroskedasticity.
- There is heteroskedasticity.

What is the p-value of the test?

What is the conclusion of the test? Choose one of the following options:

- There is not sufficient evidence to conclude that there is heteroskedasticity.
- There is sufficient evidence to conclude that there is heteroskedasticity.
- There is not sufficient evidence to conclude that there is homoskedasticity.
- There is sufficient evidence to conclude that there is homoskedasticity.

Model 2

Estimate a simple linear regression model with `price` as the dependent variable and the following two independent variables:

- `odometer`
- `age`

Your estimated sample regression intercept should be 19089.793448.

Use this model to answer the following questions.

Question 7

Holding the odometer reading fixed, according to the model by how much more does a one-year younger Toyota Corolla on average sell for on the market?

Question 8

Provide a 99% confidence interval for the estimated coefficient on `odometer`:

- Lower bound: _____
- Upper bound: _____

Question 9

Perform an appropriate hypothesis test to test the usefulness of the model. Use a 5% significance level.

- The null hypothesis is that *at least one/all/none* (choose one) of $\beta_j <$ / $\leq$ / $>$ / $\geq$ / $=$ / $\neq$ _____ for $j =$ _____ to _____ (choose one comparison operator and fill in values in the blank spaces).
- The alternative hypothesis is that *at least one/all/none* (choose one) of $\beta_j <$ / $\leq$ / $>$ / $\geq$ / $=$ / $\neq$ _____ for the same j (choose one comparison operator and fill in a value in in the blank space).
- The formula for the test statistic is of the form:

$$\frac{\frac{SST-SSE}{a}}{\frac{SSE}{n-k-1}}$$

  What is the value of a in the estimated model? _____
- What is the value of the test statistic? _____
- What is the critical value? _____
- What is your conclusion? (choose one option below):
  - *Reject* H_0. The model is *useful*.
  - *Reject* H_0. The model is *useless*.
  - *Don't reject* H_0. The model is *useful*.
  - *Don't reject* H_0. The model is *useless*.

Question 10

Use the model to test the following claim at the 5% level using a p-value approach:

"Holding the age of a Toyota Corolla fixed, each additional mile driven on average decreases the market value of a Toyota Corolla by more than 4 cents ($0.04)."

Perform this test by answering the questions below.

- What is the null hypothesis? $\beta_1 <$ / $\leq$ / $>$ / $\geq$ / $=$ / $\neq$ _____ (choose one comparison operator and fill in a value in the blank).
- What is the alternative hypothesis? $\beta_1 <$ / $\leq$ / $>$ / $\geq$ / $=$ / $\neq$ _____ (choose one comparison operator and fill in a value in the blank).

- Under the null hypothesis, the test statistic $T = (B_1 - b)/S_{B_1}$, where b is the hinge, follows a t distribution with how many degrees of freedom? _____
- What is the value of the test statistic? _____
- What is the associated p-value? _____
- What is your conclusion? Choose an option below:
  - Reject H0: There is sufficient evidence for the claim.
  - Reject H0: There is not sufficient evidence for the claim.
  - Don't reject H0: There is sufficient evidence for the claim.
  - Don't reject H0: There is not sufficient evidence for the claim.

Model 3

Estimate a simple linear regression model with `price` as the dependent variable and the following three independent variables:

- `odometer`
- `age`
- The interaction between `odometer` and `age`.

Your estimated sample regression intercept should be 20541.174993.

Use this model to answer the following questions.

Question 11

Report the estimated coefficient on the interaction term.

Question 12

According to the model, what is the average impact of driving one additional mile on the resale price for new cars (cars with age zero)?

That is, when `age` equals zero, what is the expected change in the selling price when `odometer` increases by one unit?

Question 13

According to the model, what is the average impact of driving one additional mile on the resale price for cars that are 10 years old (cars with age 10).

That is, when `age` equals ten, what is the expected change in the selling price when `odometer` increases by one unit?

Question 14

Choose the answer below which best interprets the model estimates:

- For newer cars, driving one additional mile on average has a bigger effect on the car's depreciation compared to older cars.
- For older cars, driving one additional mile on average has a bigger effect on the car's depreciation compared to newer cars.
- Driving one additional mile has the same effect on the car's depreciation, regardless of how old it is.
- Driving more miles in a car does not impact its value. All that matters is how old the car is.

Model 4

Estimate a simple linear regression model with `price` as the dependent variable and the following five independent variables:

- `odometer`
- `age`
- `state_fl`
- `state_ny`
- `state_tx`

Your estimated sample regression intercept should be 20106.250435.

Use this model to answer the following questions.

Question 15

Which variables are individually statistically significant at the 1% level?

Question 16

For a given mileage and age, Toyota Corollas on average sell for how much more in New York compared to Texas?

Question 17

Provide an interval that contains with 95% probability the asking price of a 10-year-old Toyota Corolla in Texas with 100,000 miles on its odometer.

- Lower bound: _____
- Upper bound: _____

Question 18

Test the joint usefulness of the state dummy variabes, which are variables 3-5 in your the model. Use a 5% significance level.

Choose one of the options in *italics* and fill in the blanks.

- The null hypothesis is that *all/at least one/none* of β_j _____ _____ for j from _____ to _____.
- The alternative hypothesis is that *all/at least one/none* of β_j _____ _____ for the same j.

The test statistic is of the form:

$$\frac{\frac{SSE_r - SSE_c}{a}}{\frac{SSE_c}{n-k-1}}$$

What is the value of a in the test? _____

What is the value of the test statistic? _____

What is the critical value? _____

Which of the 4 options below is the correct conclusion from the test?

- Reject H0. The variables are useful additions to the model.
- Reject H0. The variables are not useful additions to the model.
- Don't reject H0. The variables are useful additions to the model.
- Don't reject H0. The variables are not useful additions to the model.