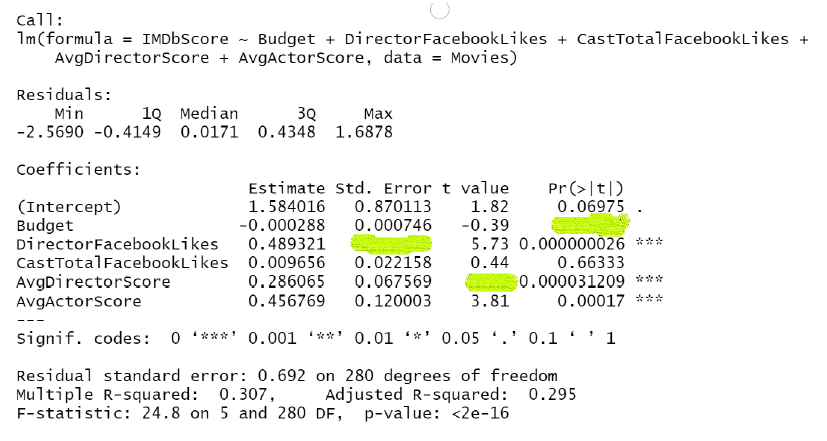
**Regression output:**

This question uses the IMDB movies ratings data from class 13 (with some observations

removed at random). Below is R output from a regression of IMDbScore on Budget,

DirectorFacebookLikes, CastFacebookLikes, AvgDirectorScore, and AvgActorScore.

Use this output to answer the questions below.



(a) What is the SE on DirectorFacebookLikes?

(b) What is the t-value on AvgDirectorScore?

(c) What is the p-value on Budget?

**A/B Testing**

The file AdSmartAB.csv contains data from an experiment done by an advertising agency

in which users were randomized into either being shown a particular ad, or not, and then

their behavior was tracked and recorded. The data has 8077 rows and 9 columns. The only two columns you need to know about are:

. **experiment:** Factor with two values: ”control” & ”exposed”. ”exposed” means user

was shown the ad. ”control” means that user was not shown the ad.

. **yes:** 0-1 variable that equals 1 if the user pressed the ”Yes” button after seeing the

ad, and 0 otherwise

Is there strong evidence in this data that seeing the ad led more users to click

”Yes”? What is the p-value from your hypothesis test? In words, what does the p-value

mean in this test? What do you conclude?

**Earnings regression:**

The nls.csv dataset contains 929 rows where each row is a worker in the US. The data

has the following columns:

. luwe = log weekly wages

. educ = years of eduction

. exper = job-market experience in years

Using this dataset, answer the following questions:

(a) What are the average years of education and average years of job-market

experience in this dataset?

(b) Run a linear regression of luwe on exper. What is the interpretation

of the coefficient on exper in terms of the relationship of job-market experience to

weekly wages?

(c) Now run a linear regression of luwe on educ and exper. Provide a 95%

confidence interval for the effect of exper on luwe computed from this regression.

(d) What is the correlation of exper and educ in the data? Suggest an

explanation for the direction of this correlation.

(e) Explain the difference between the estimated coefficient on exper in the

regression in (b) versus that in the regression in (c).

(f) Create a new variable called expersq that equals the square of exper. Run

a linear regression of luwe on educ, exper, and expersq. Test the null hypothesis

that the true coefficient on expersq is zero. Report the p-value. Do you reject the

null hypothesis at the 1% significance level?

**AirBnB**

The dataset airbnb.csv contains data on airbnb listings in several major cities in the

US. We will use this dataset to create a prediction model for the variable price. (NOTE:

be sure to load in the data with the “stringsAsFactors=TRUE” option!)

There are 57129 rows and 19 columns in the data. The columns are:

. price: price per night in dollars

. property type: Factor noting property type (e.g., “Apartment”)

. room type: Factor noting room type (e.g., “Entire home/apt”)

. accomodates: Number of people that the property accomodates

. bathrooms: Number of bathrooms

. bed type: Factor noting bed type (e.g. “Airbed”)

. cancellation policy: Factor noting cancellation policy (e.g., “flexible”)

. cleaning fee: True/False

. city: Factor noting the city (e.g. “Boston”)

. host has profile pic: True/False/Unknown

. host identity verified: True/False/Unknown

. host response rate: Factor (e.g., “10%”)

. instant bookable: True/False

. neighbourhood: Factor noting neighborhood (e.g., “16th Street Heights”)

. number of reviews: Number of reviews

. review scores rating: Avg review rating (0-100)

. zipcode: Factor with 716 levels (e.g., “02108”)

. bedrooms: Number of bedrooms

. beds: Number of beds

(a) Plot a histogram of price with 200 bins. Does price look normally distributed?

Why or why not?

(b) The variable host response rate is stored incorrectly as a factor instead

of a number (eg., “10%” or “100%” vs 10 or 100). Clean this variable so that

host response rate is stored as a number from 0-100. What is the mean of the

cleaned variable? For how many rows is host response rate missing?

(c) Set your random seed to 2022 and then randomly partition the data into

a 80% training data set and a 20% test data set. Compute the mean and standard

deviation of the variable price in both the training set and the test set.

(d) Run a regression (“Model 1”) on the training data set to predict price using

bedrooms, beds, bathrooms, number of reviews, review scores rating, and

city and report the output. What is the test set RMSE for Model 1?

(e) What is the interpretation of the coefficient on bedrooms in Model 1?

(f) What is the interpretation of the coefficient on “cityChicago” in Model 1?

(g) What is the interpretation of the coefficient on number of reviews in

Model 1? Provide a possible explanation for why this coefficient has the sign that it

does.

(h) Plot the residuals from Model 1 and comment on any patterns you see.

(i) Try running the same regression as Model 1 using the log of price instead

of price. Plot the residuals. Are the residuals from this model better, worse, or

equally good/bad to those in part (h)?

(j) Find a better prediction model for price (not log of price!) that has a

lower test set RMSE than Model 1. Your grade will depend in part on how low the

test set RMSE is for your proposed model.