MATH 60604A Statistical Modelling

Chapter 2 Exercises

Juliana Schulz

Question 1

You're interested in buying a used sports car and you're debating between a Porsche and a Jaguar. You want to make sure you find the best deal, so you collect data on the selling prices of used Porsches and Jaguars, along with information on the car's age and mileage. First, you're interested in understanding the relationship between the car mileage and price. To do this, you fit a linear regression model, the results of which are provided in the `R` output below. The variable `Mileage` gives the car mileage (in 1,000's miles) and the response variable `Price` gives the price (in $1,000's).

```
Call:
lm(formula = Price ~ Mileage, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-33.088 -10.372   2.868   9.233  20.868

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 62.92843    3.16885  19.858  < 2e-16 ***
Mileage     -0.61269    0.07621  -8.039 5.26e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.89 on 58 degrees of freedom
Multiple R-squared:  0.527, Adjusted R-squared:  0.5189
F-statistic: 64.63 on 1 and 58 DF,  p-value: 5.263e-11
```

a) Interpret the coefficient for `Mileage`.

b) Report and interpret the value for the coefficient of determination.

Question 2

Now you're interested in comparing the prices of Porsches and Jaguars. Consider the following `R` output from a linear regression model on the response variable `Price`. Note that the variable `Porsche` is an indicator

that the car is a Porsche, that is `Porsche=1` if the car is a Porsche and `Porsche=0` if the car is a Jaguar.

```
Call:
lm(formula = Price ~ Porsche, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-34.537 -12.057  -1.347  13.823  38.043

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   31.957      2.954  10.816 1.56e-15 ***
Porsche       18.580      4.178   4.447 4.00e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.18 on 58 degrees of freedom
Multiple R-squared:  0.2543,    Adjusted R-squared:  0.2414
F-statistic: 19.77 on 1 and 58 DF,  p-value: 3.997e-05
```

a) Write an expression for the underlying model being fit.

b) What is the predicted price for a Porsche?

c) What is the estimated mean price for a Jaguar?

d) Based on the fitted model, is there a significant difference in the price of a Porsche compared to a Jaguar? Explain.

Question 3

The dataset `salaries.csv` contains information on the salaries of professors. Specifically, the data includes the following variables:

`rank`:	the professor's rank, categorical taking on three levels `AsstProf` (assistant professor), `AssocProf` (associate professor) and `Prof` (full professor)
`discipline`:	the professor's discipline, binary taking on levels `A` and `B`
`yrs_since_phd`:	the number of years since the professor completed their PhD
`yrs_service`:	the number of years the professor has been working
`sex`:	the professor's sex, considered as dichotomous with levels `Male` and `Female`
`salary`:	the professor's salary

a) Produce scatterplots to visualize the pairwise relationships between the variables `salary`, `yrs_since_phd` and `yrs_service`. Comment.

b) Compute the correlations between the variables `salary`, `yrs_since_phd` and `yrs_service`. Are these correlations significant?

c) Fit a linear regression model for `salary` including the following explanatory variables (i) only `yrs_since_phd`, and (ii) only `yrs_service`. For both, provide the fitted model and comment on the significance of the variable effects (use $\alpha = 0.01$).

d) Carry out a residual analysis for model (i). Comment.

e) Using linear regression, test whether there is a significant difference between the salaries of female and male professors.

Question 4

Continue with the `salaries` data.

a) Fit a linear regression model for `salary` using both `yrs_since_phd` and `yrs_service` as predictors. Comment on the significance of the variable effects. Are the results from this model surprising based on what was observed in question Question 3? Explain.

b) Fit a linear regression model for `salary` including `sex` and `rank` as predictors. (Use `Female` as the reference level for sex and `AsstProf` as the reference level for rank).

(i) Provide the fitted model and interpret all of the model parameters.

(ii) In this model, is there a significant difference between the salaries of female and male professors? How do these results compare with that in **Question 3**? Explain.

c) Fit a linear regression model for `salary` including `rank`, `discipline` and `yrs_service`. (Use `AsstProf` as the reference level for rank and `A` as the reference level for discipline).

(i) Provide the fitted model. Interpret the coefficient associated with the variable `yrs_service`.

(ii) Comment on the difference between the test results from the individual parameter effects in comparison to the global effect for the variable `rank`. What can you conclude?

(iii) Carry out a global test for the overall fit of the model.

(iv) Based on this model, estimate the difference between the mean salaries of associate and full professors. Provide a 95% C.I. for this estimated difference.

Question 5

The `credit` data contains several variables regarding credit information for clients. In this question, we will focus on the following variables:

`Rating`:	Credit rating
`Income`:	Income (in \$1,000's)
`Limit`:	Credit limit
`Cards`:	Number of credit cards
`Age`:	Age
`Student`:	A categorical variable with levels No and Yes indicating whether the individual is a student
`Married`:	A categorical variable with levels No and Yes indicating whether the individual is married

a) Fit a linear regression model to assess the simultaneous effects of the above mentioned variables on a client's credit rating. (Use `No` as the reference level for both `Married` and `Student`). Provide the fitted model.

b) Comment on the significance of the variable effects (use $\alpha = 0.05$).

c) Comment on the value of R^2 for the model.

d) Predict the credit rating for a person with an income of 15, a limit of $5,000$, with 2 credit cards, who is a single, 27 years old student. Provide a 95% prediction interval for the predicted value.

e) Estimate the mean credit rating for individuals with an income of 100, a limit of $10,000$, 5 credit cards, is 55 years old, married and not a student. Provide a 95% C.I. for this estimate.

Question 6

Following your analysis of used car prices from questions **1** and **2**, you're now interested in comparing the prices of Porsches, Jaguars *and* BMWs. You're also interested in assessing how the age of the car affects the price. You fit a linear regression model, the results of which are provided below.

Note: The response variable is the car's `Price`, which is measured in $1,000$'s of dollars. The variable `Age` represents the car's age (in years) and the variable `CarType` is a categorical variable indicating the type of car: `BMW`, `Jaguar` and `Porsche`.

```
Call:
lm(formula = Price ~ Age * CarType, data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-17.8887  -6.8037  -0.9248   6.0900  23.0915

Coefficients:
```

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         58.2268     4.7056  12.374  < 2e-16 ***
Age                 -4.8265     0.7522  -6.416 7.85e-09 ***
CarTypeJaguar       -1.2384     6.2742  -0.197  0.84401
CarTypePorsche       5.1483     5.3702   0.959  0.34048
Age:CarTypeJaguar   -0.2135     1.0668  -0.200  0.84186
Age:CarTypePorsche   2.7558     0.8119   3.394  0.00105 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.656 on 84 degrees of freedom
Multiple R-squared:  0.7172,    Adjusted R-squared:  0.7004
F-statistic: 42.61 on 5 and 84 DF,  p-value: < 2.2e-16


Loading required package: carData


Anova Table (Type III tests)

Response: Price
             Sum Sq Df F value    Pr(>F)
(Intercept) 14277.6  1 153.115 < 2.2e-16 ***
Age          3839.0  1  41.170 7.852e-09 ***
CarType       198.1  2   1.062    0.3504
Age:CarType  2025.0  2  10.858 6.394e-05 ***
Residuals    7832.8 84
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) Interpret the intercept in the model.

b) Interpret the coefficient for `Age` in the model.

c) According to this model, what is the estimated price of a new Jaguar?

d) Write an expression for the mean price of a Porsche which is x years old.

e) Carry out a formal statistical test to assess whether the effect of age on the car's price depends on the type of car. Clearly write out the underlying hypotheses, the value of the test-statistic and the corresponding p-value. What can you conclude?

Question 7

Consider the `NCbirths` data (from the `Stat2Data` library in R, see https://cran.r-project.org/web/packages/Stat2Data/Stat2Data.pdf for more details). The data includes information from 1450 births in North Carolina in 2001. In particular, the data includes the following variables

`ID`	Patient ID code
`Plural`	1=single birth, 2=twins, 3=triplets
`Sex`	Sex of the baby, 1=male 2=female
`MomAge`	Mother's age (in years)
`Weeks`	Completed weeks of gestation
`Marital`	Marital status: 1=married or 2=not married
`RaceMom`	Mother's race: 1=white, 2=black, 3=American Indian, 4=Chinese, 5=Japanese, 6=Hawaiian, 7=Filipino, or 8=Other Asian or Pacific Islander
`HispMom`	Hispanic origin of mother: C=Cuban, M=Mexican, N=not Hispanic, O=Other Hispanic, P=Puerto Rico, S=Central/South America
`Gained`	Weight gained during pregnancy (in pounds)
`Smoke`	Smoker mom? 1=yes or 0=no
`BirthWeightOz`	Birth weight in ounces
`BirthWeightGm`	Birth weight in grams
`Low`	Indicator for low birth weight, 1=2500 grams or less
`Premie`	Indicator for premature birth, 1=36 weeks or sooner
`MomRace`	Mother's race: black, hispanic, other, or white

We are interested in investigating the factors influencing a baby's birth weight.

IMPORTANT NOTE: For all questions, consider complete cases only. That is, exclude all patients (observations) with any missing values. (Hint: the `complete.cases` function in `R` could be helpful for this.)

a) Consider a simple linear regression model, treating `BirthWeightGm` as the response variable, and `Sex` as the explanatory variable. Begin by fitting the model by directly including the sex variable exactly as is. Provide the estimated regression coefficients.

b) Based on the model in (a), what is the estimated birth weight (in grams) for male babies? for female babies?

c) Fit the same model again, this time, however, treating the sex variable as categorical (use Sex=1 as the reference level). Provide the fitted model.

d) Based on the model in (c), what is the estimated birth weight (in grams) for male babies? for female babies?

e) Based on the model in (c), is there a significant difference in the birth weights of female babies in comparison to male babies? Formally carry out a statistical test, indicating the underlying hypotheses, the value of the test statistic, the p-value and your conclusion.

f) How do the estimated regression coefficients (i.e., $\hat{\beta}_0$ and $\hat{\beta}_1$) compare in models (a) and (c)? Explain any differences and/or similarities in a few sentences.

Question 8

Continuing with the `NCbirths` data, now focus on the impacts of the mother's race on a baby's birth weight, using the `RaceMom` variable.

a) In one to two sentences, explain why we cannot simply include the variable `RaceMom` in the model directly as is.

b) Fit a linear regression model, treating `BirthWeightGm` as the response variable and `RaceMom` as the explanatory variable (using level 1 as the reference level). Provide the fitted model.

c) Based on the model from (b), comment on the significance of the regression parameters in the context of the problem.

d) Now consider a modified version of the `RaceMom` variable which takes the following levels:

1: if RaceMom=1
2: if RaceMom=2
3: otherwise (i.e., RaceMom=3,4,5,6,7, or 8)

Fit a linear regression model, again using `BirthWeightGm` as the response variable, and this time including this modified version of the race variable (use level 2 as the reference level). Provide the fitted model and interpret the regression parameters.

e) Based on the model in part (d), consider all possible pairwise tests comparing the mean birth weights of babies for mothers of the different levels of the modified race variable. What can you conclude?

Question 9

Again, continuing with the `NCbirths` data, now consider a linear regression model for `BithWeightGm` which includes the covariates `MomAge`, `Weeks`, `Smoke` and `Marital`.

a) Provide the fitted model.

b) Comment on the value of R^2.

c) Interpret the regression coefficients (i.e. the β_j) associated with the `MomAge` and `Smoke` variables.

d) Formally carry out a statistical test to verify whether `Marital` is significant in the model. Be sure to provide a conclusion in the context of the problem.

e) Carry out a residual analysis. Comment on the results.

Question 10

Consider the `GrinnellHouses_mod.csv` data (based on the `GrinnellHouses` data from the `Stat2Data` library in R, see https://cran.r-project.org/web/packages/Stat2Data/Stat2Data.pdf for more details). The data includes information on houses sold between 2005 and 2015 in Grinnell, Iowa. In particular, the data includes the following variables

`Date`	Coded value for date of sale (Jan 1, 2005=16436)
`Address`	Street address of the house
`Bedrooms`	Number of bedrooms
`Baths`	Number of bathrooms
`SquareFeet`	The square footage of the home's living space
`LotSize`	Lot size (in acres)
`YearBuilt`	Year the house was built (note that many pre-1900 homes are listed as 1900)
`YearSold`	The year the house was sold, for this case
`MonthSold`	The month the house was sold (1=Jan, 2=Feb, to 12=Dec)
`DaySold`	Day of the month the house was sold (1 to 31)
`CostPerSqFt`	SalePrice / SquareFeet (round to nearest penny)
`OrigPrice`	List price of the house when originally put on the market (dollars)
`ListPrice`	List price at the time of sale (dollars)
`SalePrice`	Sale price of the house (dollars)
`SPLPPct`	(SalePrice / ListPrice) * 100

In addition to these variables, two new variables were created: one binary variable (winter) indicating whether the house was sold during a winter month and another categorical variable for the age of the home (old/mid/new). Specifically, this was done by creating the following indicator variables:

`winter`	winter month indicator variable (=1 if the month of sale is in the month November, December, January, or Februrary, i.e., MonthSold $\in \{11, 12, 1, 2\}$)
`old`	variable indicating whether the house was built before 1950 (i.e., YearBuilt<1950)
`mid`	variable indicating whether the house was built between 1950 and 1980 (i.e., $1950 \leq$ YearBuilt < 1980)
`new`	variable indicating whether the house was built after 1980 (i.e., YearBuilt ≥ 1980)

We are interested in investigating the factors influencing the sale price of a house.

a) Model the sale price of a house in terms of the square footage, number of bathrooms, number of bedrooms, winter indicator, and age of the home (categorized as old/mid/new), including an interaction between the number of bedrooms and the winter indicator variable, as well as an interaction between the number of bedrooms and the age of the home (categorized as old/mid/new). Use level new as the reference level for the age of the home. Provide the model summary results directly obtained in R.

b) Based on the model in part (a), write an expression for the fitted models for each category of home age, that is, for old homes, mid-aged homes, and new homes, respectively.

c) Based on the model in part (a), what is the estimated effect of the number of bedrooms on the sale price of an old home when the sale occurs in a winter month?

d) Based on the model in part (a), interpret the regression coefficient associated with `SquareFeet` and the main effect of `Bedrooms`.

e) Based on the model in part (a), does the effect of the number of bedrooms on the sale price of the home depend on whether it was sold in a winter month? Justify your answer.

f) Based on the model in part (a), is there a significant difference in the effect of the number of bedrooms on the sale price for mid-aged homes in comparison to new homes? Justify your answer.

g) Based on the model in part (a), does the effect of the number of bedrooms on the sale price of the home depend on the age of the home (categorized as old/mid/new)? Justify your answer.

h) Test whether the sale occurring in a winter month (i.e., the variable `winter`) is globally significant in the model, using an F-test. Justify your answer. Use $\alpha = 1\%$.

Question 11

Continuing with the Grinnell house data, we will now explore non-linear relationships between the square footage of a house and its sale price.

a) Fit a linear regression model to the data allowing to model the sale price in terms of a quadratic relationship with square footage, in interaction with the age of the home (categorized as old/mid/new). (That is, the model should include $\texttt{SquareFeet} + \texttt{SquareFeet}^2$, in interaction with the categorized home age, with no other variables in the model). Use the new category as the reference level. Provide the model summary results directly obtained in R.

b) From the model in part a), provide an expression for the fitted model for each category of home age, that is, for old homes, mid-aged homes and new homes, respectively.

c) What is the estimated difference in the mean sale price of a home with 2000 square feet in comparison to a home with 1000 square feet, for a house built after 1980? What about for a home built between 1950 and 1980? And a home built before 1950? Be sure to show your work!

d) Now fit a model using the log-transformed sale price (i.e. $\ln(\texttt{SalePrice})$) as the response variable, and as covariates the square footage (no quadratic term this time), categorized home age, and their interaction. Again, use new as the reference level for the home age. Provide an expression for the fitted model.

e) Based on the model in part (d), interpret the coefficient corresponding to the main effect of `SquareFeet`.