Minilab 6b Worksheet

Linear Regression: Models with Quantitative and Categorical Predictor Variables

We have seen how to fit a linear model (using the lm() function in R) and how to check the validity of a fitted linear model (regression) using the four diagnostic plots. So far we have fitted linear models with one or two quantitative predictor variables only.

In this minilab, we consider several quantitative and categorical predictor variables and how to compare models in order to select which model best fits the dataset.

1. Regression Models with Two Quantitative Predictor Variables

Remember that *simple linear regression* involves one quantitative predictor variable X and one quantitative response variable Y. However, in a dataset we often have many predictor variables, some of which are quantitative and some of which are categorical.

(1) We will look at a dataset which gives credit card balance ($), income (in thousands), credit card limit, credit rating, age, and years of education for 400 people.

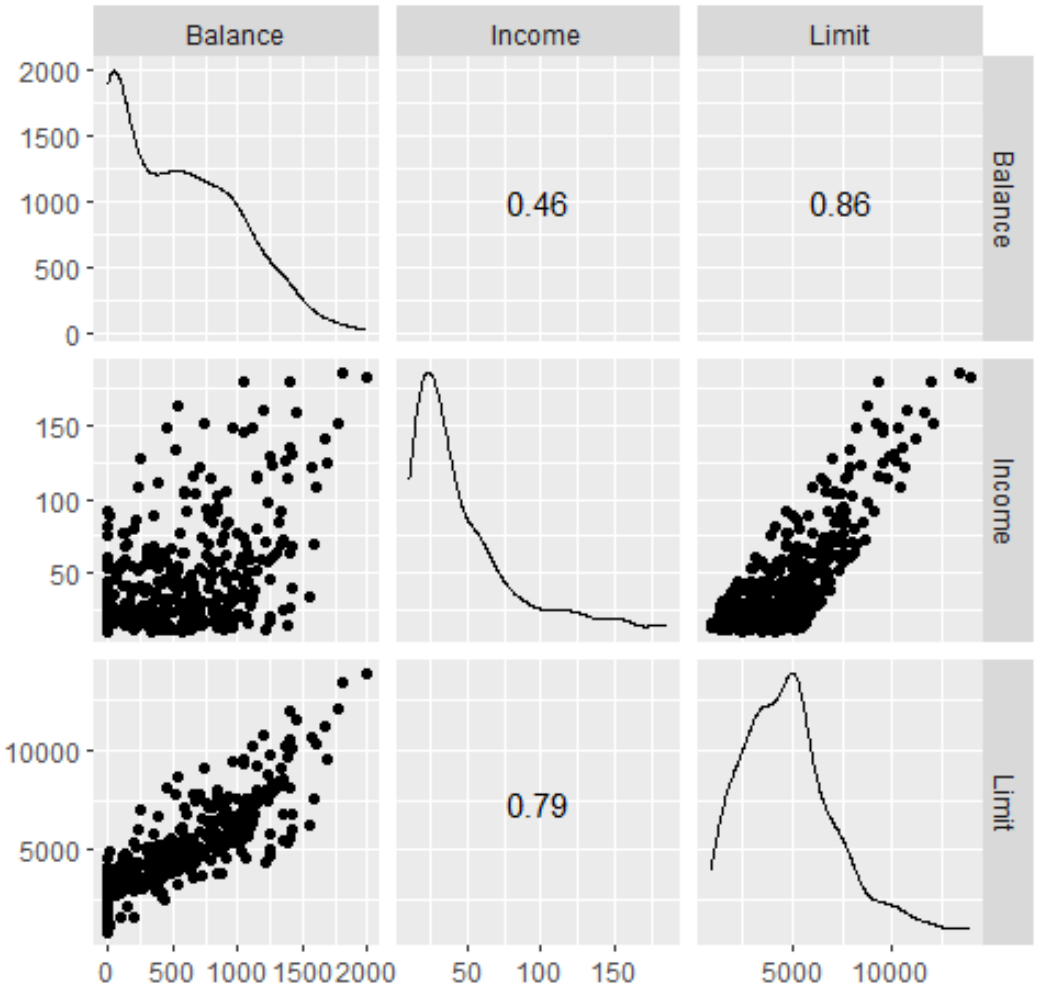
```
#install.packages("ISLR")    # run only once
library(tidyverse)
library(ISLR)
credit = as_tibble(Credit)
credit = select(credit,Balance,Limit,Income)
summary(credit)
```

The summary() function provides a useful summary of each variable separately.

(2) Let us take a look at the correlation between the variables (take them in pairs).

```
# install.packages("GGally")    # run only once
library(GGally)
ggscatmat(select(credit,Balance,Limit,Income))
```

For each pair of variables, this lovely plot gives a scatterplot and a correlation coefficient, together with a smoothed histogram for each variable (along the diagonal). We can see that the variables are clearly all positively correlated.



(3) Suppose we are interested in predicting credit card balance (Balance) using credit card limit (Limit), i.e., the response variable is Balance and the predictor variable is Limit.

```
ggplot(credit, aes(x=Limit,y=Balance)) +
  geom_point() +
  geom_smooth(method="lm",se=FALSE)
model = lm(Balance~Limit, data=credit)
summary(model)
model$coefficients
```

So this model gives the *line of best fit* as:

$$\text{Balance} = -292.7904955 + 0.1716373 \times \text{Limit}$$

Given a value of *Limit* we use this formula to predict a corresponding value for *Balance* (plug the value of *Limit* directly into the formula and evaluate).

(4) *How can we assess how good this model is, i.e., what basis can we use to compare different models?*
Recall (from a previous minilab) that the *Coefficient of Determination* is defined as the square of the correlation between the *observed* (original) and *fitted* (predicted by the model) values of the response variable, i.e.,

$$R^2 = \left(cor(y, \hat{y})\right)^2$$

This is known as the "R squared" value (or "Multiple R-squared" value). It can be interpreted as the proportion of the variance in the response variable that is predictable from the predictor variables. It provides a measure of how well observed responses are replicated by the model, based on the proportion of total variation of responses explained by the model. Higher values R-squared are better.

Only in the case of simple linear regression (one quantitative predictor variable) we have $R^2 = r^2$ when r is the Pearson correlation coefficient, i.e., $r = cor(x, y)$.

So the model "Balance~Limit" has $R^2 = 0.7425$ (the second to last line of the output from summary(model)) so we say that "74.25% of the variance in Balance is explained by Limit."

Exercise. Plot a scatterplot of "Balance~Income" and fit a linear model using lm(). What is the corresponding R^2 value? Is this better or worse than the R^2 value for "Balance~Limit"?

(5) Suppose we use both Limit and Income as predictors of Balance. The webpage https://rpubs.com/moderndive/credit_card_balance_3D_scatterplot has a nice 3D scatterplot of these three variables which you can explore by rotating the plot (hold down the left mouse button and move the mouse). If you rotate the plot "just right" the points line up approximately on a plane in 3D.

We can fit the linear model "Balance~Limit+Income" and find that $R^2 = 0.8711$ which has improved on both "Balance~Limit" and "Balance~Income".

```
model = lm(Balance~Limit+Income, data=credit)
summary(model)
```

This gives

Balance = -385.1793 + 0.24643 × Limit – 7.6633 × Income

with a **negative** multiplier for Income.

Remember that "Balance~Limit" and "Balance~Income" both gave regression lines with **positive** multipliers (slope).

Balance = -292.7905 + 0.1716 × Limit

Balance = 246.515 + 6.048 × Income

This is an example of "Simpson's Paradox"
(see https://en.wikipedia.org/wiki/Simpson%27s_paradox).

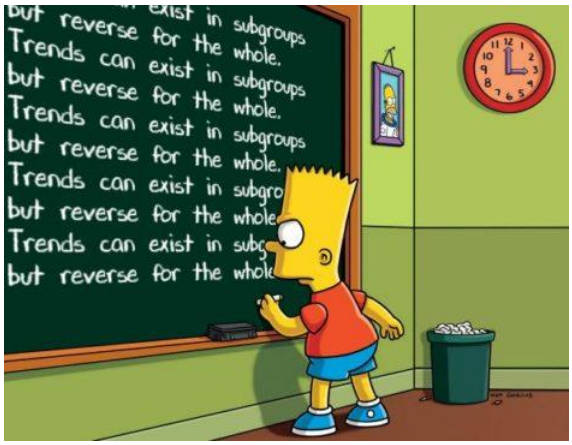


Image from: https://agitator.thedonorvoice.com/what-the-simpsons-can-teach-us-about-retention-rate/

(6) Finally, we should visually check the four diagnostic plots to see if this regression model is valid for this dataset.

```
library(ggfortify)
autoplot(model)
```

Exercise. The credit card data set also has variables for credit rating and age. Compare the models "Balance~Rating", "Balance~Age" and "Balance~Rating+Age" in terms of R^2 value.

```
credit = as_tibble(Credit)
credit = select(credit,Balance,Limit,Income,Rating,Age)
```

2. Regression Models with One Categorical Predictor Variable

Often we have categorical variables in a dataset, so how can they be used for predictions?

(1) We will look again at the "gapminder" dataset, just focussing on data from the year 2007.

```
#install.packages("gapminder")    # run only once
library(tidyverse)
library(gapminder)
gapminder2007 = gapminder %>%
  filter(year==2007) %>%
  select(country, continent, lifeExp, gdpPercap)
gapminder2007
```

(2) Take a look at life expectancy across all countries in 2007. We see that life expectancy ranges from 39.61 years to 82.60 years but with a mean (across countries) of 67.01 years. *We are ignoring the size of countries in this mean.*

```
summary(gapminder2007$lifeExp)
ggplot(gapminder2007, aes(x=lifeExp)) +
  geom_histogram(binwidth=5)
```

(3) The mean of 67.01 years gives our best estimate of life expectancy if we do not have any further information, i.e., we have NO predictor variable. Notice that fitting "lifeExp~1" also gives this same information. This is effectively fitting the line y=constant.

```
model = lm(lifeExp~1, data=gapminder2007)
model
```

(4) Now break this down by continent and we see a histogram for each continent and notice that the mean for Africa is 54.8 years but for Europe is 77.6 years.

```
gapminder2007 %>%
  group_by(continent) %>%
  summarise(count=n(),mean=mean(lifeExp))
ggplot(gapminder2007, aes(x=lifeExp)) +
  geom_histogram(binwidth=5) +
  facet_wrap(~continent)
```

(5) Now consider mean life expectancy above that of Africa.

```
gapminder2007 %>%
  group_by(continent) %>%
  summarise(count=n(),mean=mean(lifeExp-54.8))
```

We see that Americas has mean life expectancy 18.8 years higher than Africa, Asia 15.9, Europe 22.8 and Oceania 25.9.

(6) Now we are in a good place to understand what the linear model "lifeExp~continent" produces.

```
model = lm(lifeExp~continent, data=gapminder2007)
model
```

We see all these same numbers coming out. Africa is chosen as the base for comparison by R only because of alphabetical order. Because there are 5 continents in the dataset, R create 4 "dummy variables" which each have value 0 or 1, e.g., "continentAmericas" is 1 if the country belongs to the Americas continent and 1 if it does not. So a country from Africa has 0 for all 4 of these dummy variables.

R then fits the linear model

"lifeExp~continentAmericas+continentAsia+continentEurope+continentOceania"

and the multipliers (slopes) then just correspond to the continent means.

So now we should have a good understanding of what a linear model with a *single categorical predictor variable* actually means and how the regression coefficients are calculated.

3. Regression Models with One Categorical and One Quantitative Predictor Variable

Consider a dataset where we have two predictor variables but one is categorical and one is quantitative.

(1) We will look at some data which predicts a person's score (on some task) using their age (quantitative) and gender (categorical). Exploratory Data Analysis (EDA) with the scatter matrix plot gives some insights, i.e., score and age are weakly negatively correlated for both genders.

```
library(tidyverse)
load(url("http://www.openintro.org/stat/data/evals.RData"))
evals = as_tibble(evals) %>%
  select(score, age, gender)
summary(evals)
ggscatmat(evals, color="gender")
```

(2) Looking closer we can investigate fitting a linear model to each gender separately. We use geom_jitter rather than geom_point because there are multiple points all plotted over the top of each other.

```
ggplot(evals, aes(x=age,y=score,colour=gender)) +
  geom_jitter() +
  geom_smooth(method="lm", se=FALSE)
female_evals = filter(evals, gender=="female")
female_model = lm(score~age, data=female_evals)
female_model
```

```
male_evals = filter(evals, gender=="male")
male_model = lm(score~age, data=male_evals)
male_model
```

The fitted **female** linear model is: score = 4.88299 – 0.01752 × age

The fitted **male** linear model is: score = 4.436946 – 0.003993 × age

(3) Now we can look at the linear model "score~age+gender".

```
model = lm(score~age+gender, data=evals)
model
```

Since gender is categorical (and by alphabetical order female comes before male), R creates a dummy variable gendermale which is 1 if gender is male and 0 otherwise. It then fits the regression model: "score~age+gendermale" and the best fit is:

score = 4.484116 – 0.008678 × age + 0.190571 × gendermale

For a female (gendermale=0) this is:

score = 4.484116 – 0.008678 × age

For a male (gendermale=1) this is :

score = (4.484116+0.190571) – 0.008678 × age

i.e. score = 4.674687 – 0.008678 × age

These two best fit lines for female and male have the *same slope* but *different intercepts*. Neither of these is what we obtained for female and male data fitted separately in the previous step.

(4) We have ignored what is called the **"interaction effect"** between age and gender, i.e., that exactly how score is predicted from age depends on gender (and how score is predicted from gender depends on age).

In the R code below, notice carefully that we now have a "*" instead of a "+". This indicates an interaction between age and gender.

```
model = lm(score~age*gender, data=evals)
model
```

This time R is fitting the regression model:

"score~age+gendermale+age*gendermale"

and the best fit is:

$$\text{score} = 4.488299 - 0.01752 \times \text{age} - 0.44604 \times \text{gendermale} + 0.01353 \times \text{age} \times \text{gendermale}$$

How do we interpret what is going on here?

For a female (gendermale=0) this is:

$$\text{score} = 4.488299 - 0.01752 \times \text{age}$$

For a male (gendermale=1) this is:

$$\text{score} = (4.488299-0.44604) + (-0.01752+0.01353) \times \text{age}$$

i.e. $\text{score} = 4.43695 - 0.00399 \times \text{age}$

These are EXACTLY (just a tiny bit of rounding error) the best fit models we obtained for female and male separately in step (2). So now we see what including the interaction effect does.

Exercise. Consider using *smoking* status ("yes" or "no") of the mother and length of pregnancy (*gestation*) to predict birth *weight* of babies (grams).

(a) Load the dataset from the TSV file birthsmokers.tsv (on Aula).

(b) Consider the graphical plot produced by the R code below. Carefully explain what you notice in the plot.

```
ggplot(babies, aes(x=gestation,y=weight)) +
  geom_point(aes(colour=smoking)) +
  geom_smooth(method="lm",se=FALSE)
```

(c) Fit, interpret and compare the two linear models: weight ~ gestation and weight ~ gestation + smoking. What can you *conclude* about how smoking influences the relationship between weight and gestation?

(d) What **model** would you recommend to be investigated next and why?

Summary

In this minilab, we have looked at linear models with two quantitative predictor variables (with model comparisons using R^2), one categorical predictor variable, and two predictor variables where one is categorical and one is quantitative. When a predictor variable is categorical, R introduces dummy (0/1) variables.