Multiple Regression and Logistic Regression I

Dajiang Liu

@PHS 525

Apr-14-2016

Multiple Regression

- Extends simple linear regression to the scenario where
  - Multiple predictors are available

- Multiple regression often results in better models of the outcome, as
  - Very few outcomes are determined by one predictor
  - Typically, the outcome is jointly determined multiple predictors.

- Example:
  - Video game auction:
    - What factors may predict the auction price for the video games:
  - Mario_cart dataset

Mario_Kart Dataset

	price	cond_new	stock_photo	duration	wheels
1	51.55	1	1	3	1
2	37.04	0	1	7	1
⋮	⋮	⋮	⋮	⋮	⋮
140	38.76	0	0	7	0
141	54.51	1	1	1	2

Table 8.1: Four observations from the `mario_kart` data set.

Simple Linear Regression Revisited

- Examine relationships between price and cond_new

$$price = \alpha_0 + \beta \times cond_{new} + \epsilon$$

- What is the estimated values for β
- Is it significantly different from 0?
- Can you make a plot of the data?

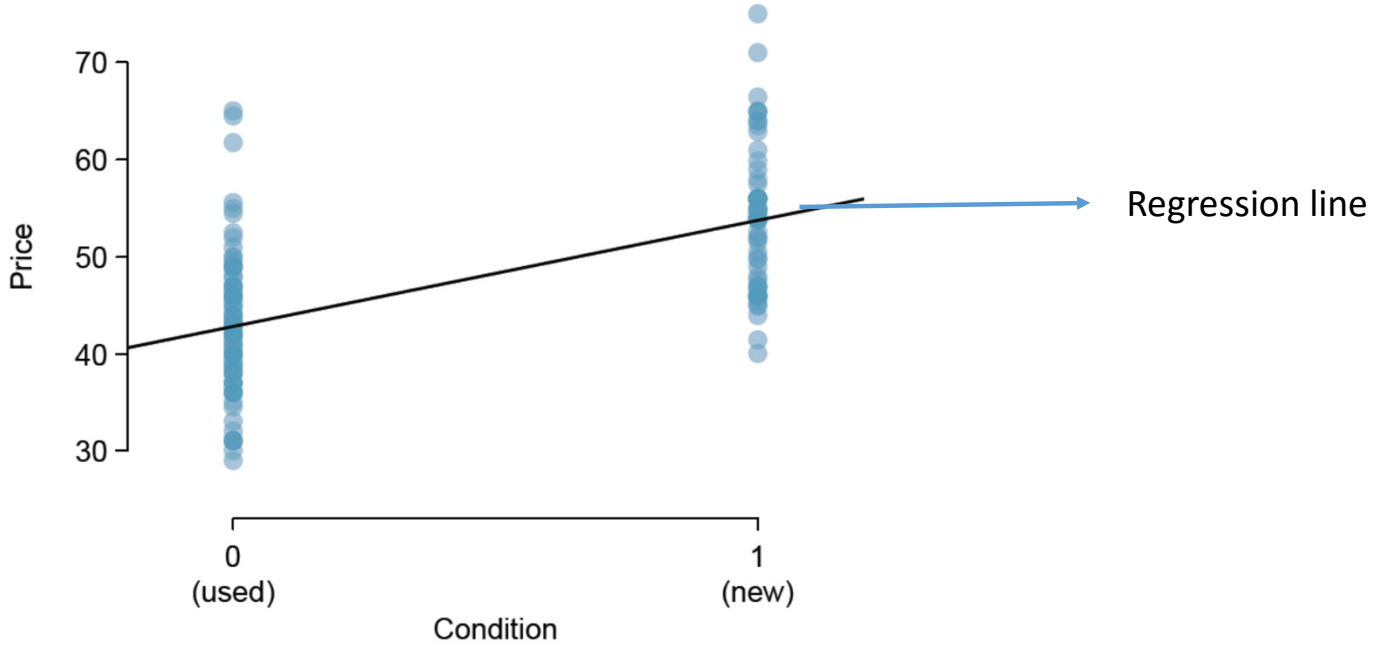

Figure 8.4: Scatterplot of the total auction price against the game's condition. The least squares line is also shown.

Multiple Regression

- In many cases, the price can be determined by multiple predictors
- In order to achieve a better model for the price, we may want to include multiple predictors in the same model
- In the Mario_Kart data, we may consider a model like

$$price = \alpha + \beta_1 \times cond_{new} + \beta_2 STOCK.PHOTO + \beta_3 \times duration + \beta_4 \times wheels + \epsilon$$

Estimating Parameters

- The parameters are estimated so that the sum of squared residuals are minimized, i.e.

$SSE = \sum_i \left(Y_i - \widehat{Y_i}\right)^2 = \sum_i \left(Y_i - \hat{\alpha} - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} \ldots\right)^2,$

where $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{i1} + \hat{\beta}_1 X_{i2} + \cdots$ is the predicted outcome based upon the predictors.

- The model parameters are estimated such that the observed outcome and predicted outcome “agree” the best.

- Can you please estimate the parameters for the Mario_Kart Example?

Why is the Estimate Different from Simple Linear Regression?

Example 8.8 We estimated a coefficient for `cond_new` in Section 8.1.1 of $b_1 = 10.90$ with a standard error of $SE_{b_1} = 1.26$ when using simple linear regression. Why might there be a difference between that estimate and the one in the multiple regression setting?

- How to interpret the estimates from multiple linear regression?

Why is the Estimate Different from Simple Linear Regression?

Example 8.8 We estimated a coefficient for `cond_new` in Section 8.1.1 of $b_1 = 10.90$ with a standard error of $SE_{b_1} = 1.26$ when using simple linear regression. Why might there be a difference between that estimate and the one in the multiple regression setting?

- How to interpret the estimates from multiple linear regression?

Answer: Holding everything else constant, a new game cost 10.90 USD more than an old game.

How to Measure How well the Model Fit: Adjusted R^2

- Estimate the amount of variability that can be explained by the model

The bigger the better ← $R^2 = 1 - \frac{var(\epsilon_i)}{var(Y_i)}$ → Residual variance; the smaller the better

- R^2 is biased

- Adjusted R^2:

$$R^2_{adj} = 1 - \frac{\frac{var(\epsilon_i)}{N-K-1}}{\frac{var(Y_i)}{N-1}}$$

- K is the number of predictors
- N is the number of sample individuals
- R^2_{adj} is always smaller than the R^2 (why??)

How to calculate R^2 from R?

summary(lm(formula = totalPr ~ as.numeric(cond), data = data))

```
Residuals:
   Min     1Q  Median     3Q    Max
-18.168  -7.771  -3.148   1.857 279.362

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        60.393      7.219   8.366 5.24e-14 ***
as.numeric(cond)   -6.623      4.343  -1.525     0.13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.57 on 141 degrees of freedom
Multiple R-squared:  0.01622,   Adjusted R-squared:  0.009244
F-statistic: 2.325 on 1 and 141 DF,  p-value: 0.1296
```

8.3 Baby weights, Part III. We considered the variables `smoke` and `parity`, one at a time, in modeling birth weights of babies in Exercises 8.1 and 8.2. A more realistic approach to modeling infant weights is to consider all possibly related variables at once. Other variables of interest include length of pregnancy in days (`gestation`), mother's age in years (`age`), mother's height in inches (`height`), and mother's pregnancy weight in pounds (`weight`). Below are three observations from this data set.

	bwt	gestation	parity	age	height	weight	smoke
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1236	117	297	0	38	65	129	0

The summary table below shows the results of a regression model for predicting the average birth weight of babies based on all of the variables included in the data set.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-80.41	14.35	-5.60	0.0000
gestation	0.44	0.03	15.26	0.0000
parity	-3.33	1.13	-2.95	0.0033
age	-0.01	0.09	-0.10	0.9170
height	1.15	0.21	5.63	0.0000
weight	0.05	0.03	1.99	0.0471
smoke	-8.40	0.95	-8.81	0.0000

(a) Write the equation of the regression line that includes all of the variables.

(b) Interpret the slopes of `gestation` and `age` in this context.

(c) The coefficient for `parity` is different than in the linear model shown in Exercise 8.2. Why might there be a difference?

(d) Calculate the residual for the first observation in the data set.

(e) The variance of the residuals is 249.28, and the variance of the birth weights of all babies in the data set is 332.57. Calculate the R^2 and the adjusted R^2. Note that there are 1,236 observations in the data set.

```
y.new=data$totalPr[data$cond=='new']
y.used=data$totalPr[data$cond=='used']
y.new
y.used
t.test(y.new,y.used)
history()
names(data)
data$duration
data$wheels
data$stockPhoto
lm(totalPr~duration+stockPhoto+wheels+cond)
lm(totalPr~duration+stockPhoto+wheels+cond,data=data)
summary(lm(totalPr~duration+stockPhoto+wheels+cond,data=data))
history()
dir()
res=read.table('babies.csv',header=T,sep=',');
baby=read.table('babies.csv',header=T,sep=',');
names(baby)
history()
baby$case
names(baby)
lm(btw ~ gestation + parity + age + height + weight + smoke, data=baby)
lm(bwt ~ gestation + parity + age + height + weight + smoke, data=baby)
summary(lm(bwt ~ gestation + parity + age + height + weight + smoke, data=baby))
history
```

Two P-values

- P-values for model fitting
  - $H_0: \beta_1 = \cdots = \beta_J = 0$
  - $H_A: \beta_1 \neq 0$ or $\beta_2 \neq 0$ or … $\beta_J \neq 0$

- P-values for testing the statistical significance for each predictor
  - $H_0: \beta_j = 0$
  - $H_A: \beta_j \neq 0$

An Warmup Exercise

8.5 GPA. A survey of 55 Duke University students asked about their GPA, number of hours they study at night, number of nights they go out, and their gender. Summary output of the regression model is shown below. Note that male is coded as 1.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.45	0.35	9.85	0.00
studyweek	0.00	0.00	0.27	0.79
sleepnight	0.01	0.05	0.11	0.91
outnight	0.05	0.05	1.01	0.32
gender	-0.08	0.12	-0.68	0.50

(a) Calculate a 95% confidence interval for the coefficient of gender in the model, and interpret it in the context of the data.

(b) Would you expect a 95% confidence interval for the slope of the remaining variables to include 0? Explain

Questions of Interest

- Not all predictors are useful
- Including “not useful” predictors in the model will reduce the accuracy of predictors
- **Full model** is the model that contains all predictors
- Question: Determine useful predictors from the full model

Approach I

- Fit the full model that contains the full set of predictors

- Determine which predictors are important by looking at
  - P-values for testing $H_0: \beta_j = 0$
  - Predictor j is important if p-values are significant for testing H_0