7.1 Interpreting multiple regression models

Interpreting regression parameters

The parameter estimates for the sample multiple regression function, $\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$, have the following interpretations:

- b_0 represents the Y-intercept of a sample multiple regression function which is the fitted value of Y when $X_1 = X_2 = \ldots = X_n = 0$.
- $b_1, b_2, \ldots, b_n$ each represent the slope with respect to a particular predictor variable of a sample multiple regression function.

The Y-intercept, b_0, has a practical meaning for datasets in which $X_1 = X_2 = \ldots = X_n = 0$ are meaningful values and sample data exist at or near $X_1 = X_2 = \ldots = X_n = 0$. The measurement unit of b_0 is the same as the measurement unit of the response Y variable. The Y-intercept has no practical meaning for datasets in which $X_1 = X_2 = \ldots = X_n = 0$ are not meaningful values or no sample data are close to $X_1 = X_2 = \ldots = X_n = 0$.

The other parameter estimates, $b_1, b_2, \ldots, b_n$, always have a practical meaning. The measurement unit of b_1 is the measurement unit of Y per the measurement unit of X_1. In other words, the estimated slope parameter for a variable represents the change in the fitted value of Y per unit change in the variable when the other predictors are fixed.

A consideration when constructing multiple regression models is *multicollinearity*, which occurs when two or more predictors in the model are so highly correlated that regression parameter estimates become unreliable with inflated standard errors. Interpreting a regression parameter in the presence of multicollinearity can be challenging and is beyond the scope of this material.

Example 7.1.1: Percent body fat.

Consider the body fat dataset and a model where the response variable Y is percent body fat and the predictor variables $X_1 =$ triceps skinfold thickness (mm) and $X_2 =$ midarm circumference (cm). The sample multiple regression model is given by $\hat{Y} = 6.792 + 1.001X_1 - 0.431X_2$.

For each question below, state the units and practical interpretation of the quantity.

a. What is the estimate for the intercept parameter?
b. What is the estimate for the slope parameter with respect to X_1?

Solution

a. The estimate for the intercept parameter is $b_0 = 6.792$ percent body fat. However, this value does not have any practical meaning, because body measurements cannot be equal to zero.
b. The estimate for the slope parameter with respect to X_1 is $b_1 = 1.001$ percent per millimeter. 1.001 represents the expected change in body fat when triceps skinfold circumference increases by one unit and midarm circumference is fixed.

Analysis

If triceps skinfold circumference and midarm circumference were too highly correlated, then increasing triceps skinfold circumference by one unit while keeping midarm circumference fixed might be impractical. This difficulty in interpreting regression parameters in the presence of multicollinearity means that multicollinearity is generally avoided when parameter interpretation is important for the application at hand.

PARTICIPATION ACTIVITY 7.1.1: Interpreting multiple regression parameters.

Refer to the model in the example above.

1) What is the estimated slope with respect to X_2?
   - 6.792
   - 1.001
   - -0.431

2) What are the units of the estimated slope parameter with respect to X_2?
   - % per mm
   - % per cm
   - cm per %

3) What is the interpretation of -0.431 in the sample regression function?
   - The change in the fitted value of Y per unit change in X_2.

- The change in the fitted value of Y per unit change in X_2, when X_1 is fixed.

Python-Practice 7.1.1: Multiple regression models.

Consider the body fat dataset and a model where the response variable Y is percent body fat and the predictor variables $X1 =$ triceps skinfold thickness (mm) and $X_2 =$ midarm circumference (cm). The model is constructed using the code below.

```
import pandas as pd
from statsmodels.formula.api import ols

fat = pd.read_csv('https://static-resources.zybooks.com/static/fat.csv')

Y = fat['body_fat_percent']

m12 = ols('Y ~ triceps_skinfold_thickness_mm + midarm_circumference_cm', data = fat).fit()
print(m12.summary())
```

`R-squared` measures the proportion of total variation in Y that is accounted for by the multiple regression model, which is 0.786. `Adj. R-squared` is an adjustment to `R-squared` that allows alternative models for the same response variable to be compared. `F-statistic` and `Prob (F-statistic)` tests whether no linear regression relationship exists between Y and the the set $\{X_1, X_2\}$.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.786
Model:                            OLS   Adj. R-squared:                  0.761
Method:                 Least Squares   F-statistic:                     31.25
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           2.02e-06
Time:                        23:19:26   Log-Likelihood:                -45.050
No. Observations:                  20   AIC:                             96.10
Df Residuals:                      17   BIC:                             99.09
Df Model:                           2
Covariance Type:            nonrobust
```

The `coef` column in the table below are the estimates for the parameters, which are $b_0 = 6.7916$, $b_1 = 1.0006$, and $b_2 = -0.4314$. Thus, the equation for the model is $\hat{Y} = 6.7916 + 1.0006X_1 - 0.4314X_2$. The `std err` column contains standard errors of the regression parameter estimators, which measure the precision of the estimators. The `t` column contains individual t-statistics for the regression parameter estimators, equal to each estimate divided by its standard error. The next column contains individual p-values for the regression parameter estimators, equal to the sum of the tail areas beyond the t-statistic. The last two columns give the lower and upper bounds of the 95% confidence interval.

```
==============================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept                      6.7916      4.488      1.513      0.149      -2.678      16.261
triceps_skinfold_thickness_mm  1.0006      0.128      7.803      0.000       0.730       1.271
midarm_circumference_cm       -0.4314      0.177     -2.443      0.026      -0.804      -0.059
```

Run example

Interpreting residual standard error

As with simple linear models, several quantities in multiple regression models use the residuals $\epsilon_i = Y_i - \hat{Y}_i$. These quantities are restated below for review. Recall that Y_i is the actual value of the response variable, $\hat{Y}_i$ is the fitted value of Y_i, and $\bar{Y}$ is the sample mean of the response values.

- The regression sum of squares is $SSR = \sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2$. The SSR represents the explained variation in the dataset.
- The residual sum of squares is the sum of squared residuals for the sample, $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$. The SSE represents the unexplained variation in the dataset.
- The total sum of squares is $SSTO = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$. Note that $SSTO = SSR + SSE$. The $SSTO$ represents the total variation in the dataset.

Intuitively, the magnitude of the residuals indicate the strength of the model. The closer the residuals are to zero, the more effective the model is in predicting the values of the dependent variable. Since residuals can be positive or negative and the mean of the residuals is zero, computing the square root of the mean squared residual is sometimes more useful.

The ***residual mean square error*** (***residual standard error***), denoted by $RMSE$, is the square root of the mean squared residual, which estimates the standard deviation of the residuals. The residual standard error is obtained using the formula

$$RMSE = \sqrt{MSE} = \sqrt{\frac{SSE}{N - p}}$$

where N is the sample size and $p = n + 1$ is the number of regression parameters.

The measurement unit of the residual standard error is the same as the measurement unit of the response Y variable. The residual standard error is a measure of the precision of a model prediction.

The sample multiple regression function can be used to predict a future value of Y for fixed values of $X_1, X_2, \ldots, X_n$. A relatively small residual standard error indicates that the actual future value of Y is likely to be relatively close to the predicted value. Therefore, less residual standard error is better.

zyDE 7.1.1: Analysis of variance table.

The mean standard error and standard residual error can be obtained using an analysi variance table. Consider the body fat dataset and a model where the response variable percent body fat and the two predictor variables are $X1 =$ is triceps skinfold thickness and $X_2 =$ midarm circumference (cm).

In the output, the `sum_sq` column gives the SSR and SSE, which are 389.455334 and 105.934166 respectively. 389.455334 is the variation in percent body fat explained by the variation in triceps skinfold thickness and midarm circumference. 105.934166 is the var in percent body fat unexplained by the variation in triceps skinfold thickness and mida circumference.

The mean square residual can be obtained by dividing $SSE = 105.934166$ by $N - p = 20 - 3 = 17$, which is also given by the residual degrees of freedom df_E in the `df` column. Using these values yields the mean square residual below.

$$MSE = \frac{SSE}{N - p} = \frac{105.934166}{20 - 3} = 6.231$$

Thus, the residual standard error is $RMSE = \sqrt{MSE} = \sqrt{6.231421} = 2.496$.

Current file: main.py ▾ efault template...

Run

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api impor

fat = pd.read_csv('fat.csv')

Regression = fat[['triceps_skinfol

Y = fat['body_fat_percent']

m12 = ols('Y ~ Regression', fat).f

print(sm.stats.anova_lm(m12, typ=2
```

PARTICIPATION ACTIVITY 7.1.2: Interpreting residual standard error.

Consider the body fat model with 3 predictor variables $X_1 =$ triceps skinfold thickness, $X_2 =$ midarm circumference, and $X_3 =$ thigh circumference. The anova table is given below.

	sum_sq	df	F	PR(>F)
Regression	396.984612	3.0	21.515712	0.000007
Residual	98.404888	16.0	NaN	NaN

Run example

1) What is the residual standard error $RMSE$?

- 98.4
- 6.150
- 2.480

2) What is the interpretation of the value of $RMSE$?

- Variation in percent body fat explained by triceps skinfold thickness, midarm circumference, and thigh circumference
- Variation in percent body fat unexplained by triceps skinfold thickness, midarm circumference, and thigh circumference
- Standard deviation of the residuals

3) Would the model with residual standard error of 2.480 tend to give more or less precise predictions than another model with residual standard error of 2.496?

- More precise
- Less precise

7.2 Testing multiple regression parameters

Hypothesis tests for multiple regression

In a multiple regression model, if a parameter is zero, then the corresponding predictor is redundant. To keep the model as simple, interpretable, and generalizable as possible, predictors with a regression parameter of zero should be removed.

Consider the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$. Two types of hypothesis tests are available to determine whether the estimator is sufficiently close enough to zero to warrant the removal of the corresponding predictor variable.

Table 7.2.1: Hypothesis tests for multiple regression.

Test	Research question	Null and alternative hypotheses
Overall F-test	Determine whether a linear relationship exists with at least one predictor variable	$H_0: \beta_1 = \beta_2 = \ldots = \beta_n = 0$ H_a: At least one $\beta_i \neq 0$ for $i = 1, 2, \ldots, n$.
Individual t-test	Determine whether a single variable has an effect	$H_0: \beta_i = 0$ for some $i = 1, \ldots, n$ $H_a: \beta_i \neq 0$

Example 7.2.1: A closer look at the percent body fat model.

Consider the body fat model with response variable $Y =$ percent body fat and three predictor variables $X_1 =$ triceps trifold thickness, $X_2 =$ midarm circumference, and $X_3 =$ thigh circumference.

For each of the questions below, determine the appropriate hypothesis test, and state the null and alternative hypotheses.

a. Is a model with at least one predictor variable useful in predicting a person's percent body fat?
b. Is triceps trifold thickness significantly related to percent body fat?

Solution

a. Overall F-test. The null hypothesis is that the model is not useful and thus all slope parameters $\beta_1, \beta_2, \beta_3$ are equal to zero. That is, $H_0: \beta_1 = \beta_2 = \beta_3 = 0$. If at least one predictor is useful in predicting a person's percent body fat, then at least one of slope parameters should be non-zero. That is, H_a: At least one $\beta_i \neq 0$ for $i = 1, 2, 3$.

b. t-test. The null hypothesis is that triceps trifold thickness is not significantly related to percent body fat. Thus, $H_0: \beta_1 = 0$ and $H_a: \beta_1 \neq 0$.

PARTICIPATION ACTIVITY 7.2.1: Determining which test to use for multiple regression models.

For each research question, determine the appropriate test.

1) A car manufacturer is studying whether a significant linear relationship exists between quality and all predictor variables in the dataset, angle and speed.

- Overall F-test
- Individual t-test

2) A car manufacturer is studying whether a significant linear relationship exists between quality and the speed of a tool that manufactures car parts.

- Overall F-test
- Individual t-test

Overall F-test

The overall F-test considers *all* the regression parameters other than the intercept parameter, β_0. The ***multiple regression overall F-test*** is a method for testing $H_0: \beta_1 = \beta_2 = \ldots = \beta_n = 0$ versus H_a: At least one $\beta_i \neq 0$ for $i = 1, 2, \ldots, n$.

Overall F-test

The steps for the overall F-test are as follows:

- Set the null hypothesis, $H_0: \beta_1 = \beta_2 = \ldots = \beta_n = 0$, versus the alternative hypothesis, H_a: At least one $\beta_i \neq 0$ for $i = 1, 2, \ldots, n$.
- Use statistical software to find the F-statistic, which is the regression mean square divided by the residual mean square.
- Use statistical software to find the p-value that corresponds to the F-statistic. The p-value is the probability of observing an F-statistic at least as far from 0 as the one observed, if the null hypothesis were true.
- Make a decision based on a selected significance level, typically 0.05:
  - If the p-value is less than the significance level, reject the null hypothesis, $H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$, in favor of the alternative hypothesis, H_a: At least one $\beta_i \neq 0$ for $i = 1, 2, \ldots, n$. Conclude that at least one β_k is significantly different from 0, which means that a significant linear relationship exists between Y and the set $\{X_1, X_2, \ldots, X_n\}$.
  - If the p-value is greater than or equal to the significance level, fail to reject the null hypothesis, $H_0: \beta_1 = \beta_2 = \ldots = \beta_n = 0$, in favor of the alternative hypothesis, H_a: At least one $\beta_i \neq 0$ for $i = 1, 2, \ldots, n$. Conclude that no β_k is significantly different from 0, which means that a significant linear relationship does not exist between Y and the set $\{X_1, X_2, \ldots, X_n\}$.

Example 7.2.2: Performing the overall F-test on the percent body fat model.

Consider the body fat model with response variable $Y =$ percent body fat and three predictor variables $X_1 =$ triceps trifold thickness, $X_2 =$ midarm circumference, and $X_3 =$ thigh circumference. Does a linear relationship exist with at least one variable?

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.801
Model:                            OLS   Adj. R-squared:                  0.764
Method:                 Least Squares   F-statistic:                     21.52
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           7.34e-06
Time:                        18:46:57   Log-Likelihood:                -44.312
No. Observations:                  20   AIC:                             96.62
Df Residuals:                      16   BIC:                             100.6
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept                    117.0847     99.782      1.173      0.258     -94.445     328.614
triceps_skinfold_thickness_mm     4.3341      3.016      1.437      0.170      -2.059      10.727
midarm_circumference_cm         -2.1861      1.595     -1.370      0.190      -5.568       1.196
thigh_circumference_cm          -2.8568      2.582     -1.106      0.285      -8.330       2.617
==============================================================================
Omnibus:                        1.200   Durbin-Watson:                   2.243
Prob(Omnibus):                  0.549   Jarque-Bera (JB):                0.830
Skew:                          -0.085   Prob(JB):                        0.660
Kurtosis:                       2.016   Cond. No.                     1.15e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.15e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```


Run example

Solution

The null hypothesis is that no relationship exists between any of the predictor variables X_1, X_2, and X_3 and the response variable Y. The alternative hypothesis is that a relationship exists with at least one variable.

Mathematically, the null and alternative hypotheses are

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$
H_a: At least one $\beta_i \neq 0$ for $i = 1, 2, 3$

The overall F-statistic is $F = 21.52$ with a corresponding p-value of $7.34 \cdot 10^{-6}$. Since this p-value is close to 0, sufficient evidence exists to reject the null hypothesis, which suggests that at least one of X_1, X_2, and X_3 is linearly related to Y.

PARTICIPATION ACTIVITY 7.2.2: Performing the overall F-test on the cars dataset.

Given below are the ANOVA table and the multiple regression model for the Cars dataset where $Y = \text{Quality}$ is the response variable, and $X_1 = \text{Speed}$ and $X_2 = \text{Angle}$ are predictor variables. Does a linear relationship exist with at least one predictor variable?

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.978
Model:                            OLS   Adj. R-squared:                  0.975
Method:                 Least Squares   F-statistic:                     332.2
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           3.80e-13
Time:                        20:48:21   Log-Likelihood:                -21.142
No. Observations:                  18   AIC:                             48.28
Df Residuals:                      15   BIC:                             50.95
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.5382      0.473      1.137      0.273      -0.471       1.547
Speed         -1.9046      0.176    -10.834      0.000      -2.279      -1.530
Angle          4.0280      0.178     22.574      0.000       3.648       4.408
==============================================================================
Omnibus:                        4.358   Durbin-Watson:                   2.121
Prob(Omnibus):                  0.113   Jarque-Bera (JB):                1.414
Skew:                           0.082   Prob(JB):                        0.493
Kurtosis:                       1.637   Cond. No.                         14.4
==============================================================================
```

Run example

1) What is the null hypothesis?

- $H_0: \beta_1 = 0$
- $H_0: \beta_2 = 0$
- $H_0: \beta_1 = \beta_2 = 0$

2) What is the alternative hypothesis?

- $H_a: \beta_1 \neq \beta_2 \neq 0$
- H_a: At least one $\beta_i \neq 0$ for $i = 1, \ldots n$
- H_a: At least one $b_i \neq 0$ for $i = 1, \ldots n$

3) What is the decision based on a significance level of 0.05?

- Reject H_0
- Fail to reject H_0

4) What is the conclusion about whether a linear relationship exists between quality and at least one of the predictors, speed and angle?

- A statistically significant linear relationship exists

- A statistically significant linear relationship does not exist

Individual t-test

An individual t-test considers *one* of the regression parameters. A ***multiple regression individual t-test*** is a method for testing $H_0: \beta_i = 0$ versus $H_a: \beta_i \neq 0$. Ex: The body fat example has three regression parameters other than the intercept parameter. One individual t-test could test $H_0: \beta_1 = 0$. Other t-tests could be $H_0: \beta_2 = 0$ and $H_0: \beta_3 = 0$.

Individual t-test

The steps for an individual t-test are as follows:

- Set the null hypothesis, $H_0: \beta_i = 0$, versus the alternative hypothesis, $H_a: \beta_i \neq 0$.
- Use statistical software to find the t-statistic, which is b_i divided by the standard error of b_i.
- Use statistical software to find the p-value that corresponds to the t-statistic. The p-value is the probability of observing a t-statistic at least as far from 0 as the one observed, if the null hypothesis were true. The reference t-distribution has $n - p$ degrees of freedom.
- Make a decision based on a selected significance level, typically 0.05:
  - If the p-value is less than the significance level, reject the null hypothesis, $H_0: \beta_i = 0$, in favor of the alternative hypothesis, $H_a: \beta_i \neq 0$. Conclude that β_i is significantly different from 0, which means that a significant linear relationship exists between Y and X_i when the remaining predictor variables in the model are fixed. Thus, X_i should not be dropped from the model.
  - If the p-value is greater than or equal to the significance level, fail to reject the null hypothesis, $H_0: \beta_i = 0$, in favor of the alternative hypothesis, $H_a: \beta_i \neq 0$. Conclude that β_i is not significantly different from 0, which means that a significant linear relationship does not exist between Y and X_i when the remaining predictor variables in the model are fixed. Thus, a new model should be investigated in which X_i has been dropped from the model but the remaining predictor variables are retained.

Example 7.2.3: Performing the t-test on the percent body fat model.

Consider the body fat model with response variable $Y =$ percent body fat and two predictor variables $X_1 =$ triceps trifold thickness and $X_2 =$ midarm circumference. At the $\alpha = 0.01$ significance level, does midarm circumference have a statistically significant effect on percent body fat according to the output below?


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.786
Model:                            OLS   Adj. R-squared:                  0.761
Method:                 Least Squares   F-statistic:                     31.25
Date:                Tue, 16 Jul 2019   Prob (F-statistic):           2.02e-06
Time:                        00:12:50   Log-Likelihood:                -45.050
No. Observations:                  20   AIC:                             96.10
Df Residuals:                      17   BIC:                             99.09
Df Model:                           2
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       6.7916      4.488      1.513      0.149      -2.678      16.261
triceps_skinfold_thickness_mm   1.0006      0.128      7.803      0.000       0.730       1.271
midarm_circumference_cm        -0.4314      0.177     -2.443      0.026      -0.804      -0.059
==============================================================================
Omnibus:                        1.363   Durbin-Watson:                   2.371
Prob(Omnibus):                  0.506   Jarque-Bera (JB):                0.873
Skew:                           0.068   Prob(JB):                        0.646
Kurtosis:                       1.985   Cond. No.                         304.
==============================================================================
```

Run example

Solution

The null hypothesis is that no relationship exists between the response variable Y and the predictor variable X_2. The alternative hypothesis is that a relationship exists.

Mathematically, the null and alternative hypotheses are

$H_0: \beta_2 = 0$
$H_a: \beta_2 \neq 0$

The t-statistic for the midarm circumference estimate is -2.443 with a corresponding p-value of 0.026. Since this p-value is greater than the significance level $\alpha = 0.01$, insufficient evidence exists to reject the null hypothesis, which suggests that percent body fat and midarm circumference are not linearly related.

PARTICIPATION ACTIVITY 7.2.3: Performing the individual t-test on the cars dataset.

Given below are the ANOVA table and the multiple regression model for the Cars dataset where $Y = \text{Quality}$ is the response variable, and $X_1 = \text{Speed}$ and $X_2 = \text{Angle}$ are predictor variables. Does X_1 have a statistically significant effect on Y?

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.978
Model:                            OLS   Adj. R-squared:                  0.975
Method:                 Least Squares   F-statistic:                     332.2
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           3.80e-13
Time:                        20:48:21   Log-Likelihood:                -21.142
No. Observations:                  18   AIC:                             48.28
Df Residuals:                      15   BIC:                             50.95
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.5382      0.473      1.137      0.273      -0.471       1.547
Speed         -1.9046      0.176    -10.834      0.000      -2.279      -1.530
Angle          4.0280      0.178     22.574      0.000       3.648       4.408
==============================================================================
Omnibus:                        4.358   Durbin-Watson:                   2.121
Prob(Omnibus):                  0.113   Jarque-Bera (JB):                1.414
Skew:                           0.082   Prob(JB):                        0.493
Kurtosis:                       1.637   Cond. No.                         14.4
==============================================================================
```

Run example

1) What is the null hypothesis?

- $H_0: \beta_1 = 0$
- $H_0: \beta_2 = 0$
- $H_0: \beta_1 \neq 0$

2) What is the alternative hypothesis?

- $H_a: \beta_2 \neq 0$
- $H_a: b_1 \neq 0$
- $H_a: \beta_1 \neq 0$

3) What is the decision based on a significance level of 0.05?

- Reject H_0
- Fail to reject H_0

4) What is the conclusion about whether a linear relationship exists between quality and speed, when angle is fixed?

- A statistically significant linear relationship exists

- A statistically significant linear relationship does not exist

7.3 Multiple regression examples

This section has been set as optional by your instructor.

mtcars is a historical dataset from a 1974 issue of Motor Trend comparing the performance of 32 cars. This dataset has 11 quantitative variables listed in the table below. Each row of the dataset corresponds to a particular car and each column corresponds to a quantitative variable.

Table 7.3.1: Column names with corresponding descriptions of the mtcars dataset.

Column	Description
`mpg`	Miles/(US) gallon
`cyl`	Number of cylinders
`disp`	Displacement (cu. in.)
`hp`	Gross horsepower
`drat`	Rear axle ratio
`wt`	Weight (1000 lbs)
`qsec`	1/4 mile time
`vs`	Engine (0 = V-shaped, 1 = straight)
`am`	Transmission (0 = automatic, 1 = manual)
`gear`	Number of forward gears
`carb`	Number of carburetors

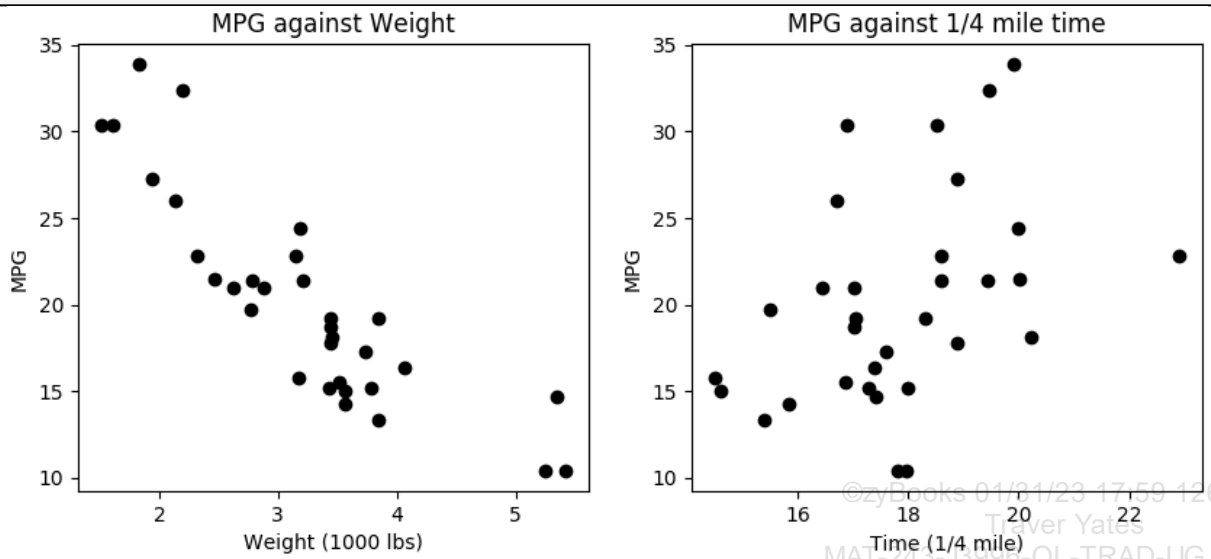
Example 7.3.1: Multiple regression using two predictor variables.

©zyBooks 01/31/23 17:59 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

```
import matplotlib.pyplot as plt
import pandas as pd
from statsmodels.formula.api import ols
import statsmodels.graphics.gofplots as smg

df = pd.read_csv("http://data-analytics.zybooks.com/mtcars.csv")
df_cars = df[["mpg","wt","qsec"]]

print("Correlation Matrix\n")
print (df_cars.corr())
print("\n\n")

fig = plt.figure(figsize = (10, 4))
ax = fig.add_subplot(1, 2, 1)
plt.plot(df_cars["wt"], df_cars["mpg"], 'o', color='black')
plt.title('MPG against Weight')
plt.xlabel('Weight (1000 lbs)')
plt.ylabel('MPG')

ax = fig.add_subplot(1, 2, 2)
plt.plot(df_cars["qsec"], df_cars["mpg"], 'o', color='black')
plt.title('MPG against 1/4 mile time')
plt.xlabel('Time (1/4 mile)')
plt.ylabel('MPG')

Y = df_cars["mpg"]
X = df_cars[["wt","qsec"]]
model = ols('mpg ~ wt+qsec', data=df_cars).fit()

print("Model\n")
print(model.summary())

plt.show()
```


©zyBooks 01/31/23 17:59 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

```
Correlation Matrix

           mpg        wt      qsec
mpg   1.000000 -0.867659  0.418684
wt   -0.867659  1.000000 -0.174716
qsec  0.418684 -0.174716  1.000000



Model

                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.826
Model:                            OLS   Adj. R-squared:                  0.814
Method:                 Least Squares   F-statistic:                     69.03
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           9.39e-12
Time:                        16:10:55   Log-Likelihood:                -74.360
No. Observations:                  32   AIC:                             154.7
Df Residuals:                      29   BIC:                             159.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     19.7462      5.252      3.760      0.001       9.005      30.488
wt            -5.0480      0.484    -10.430      0.000      -6.038      -4.058
qsec           0.9292      0.265      3.506      0.001       0.387       1.471
==============================================================================
Omnibus:                        3.357   Durbin-Watson:                   1.496
Prob(Omnibus):                  0.187   Jarque-Bera (JB):                2.542
Skew:                           0.690   Prob(JB):                        0.281
Kurtosis:                       3.032   Cond. No.                         209.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Run example

PARTICIPATION ACTIVITY 7.3.1: Multiple regression using two predictor variables.

Use the example above to answer the following.

1) Is weight correlated with MPG?

- Yes
- No

2) Does a positively correlated relationship, a negatively correlated relationship, or no relationship at all exist between MPG and weight?

- Positively correlated relationship
- Negatively correlated relationship
- No relationship

3) Let Y represent MPG, X_1 represent weight, and X_2 represent 1/4 mile time. What is the multiple regression model?

- $\hat{Y} = 5.0480X_1 + 0.9292X_2$
- $\hat{Y} = 19.7462X_1 - 5.0480X_2 + 0.9292$
- $\hat{Y} = 19.7462 - 5.0480X_1 + 0.9292X_2$

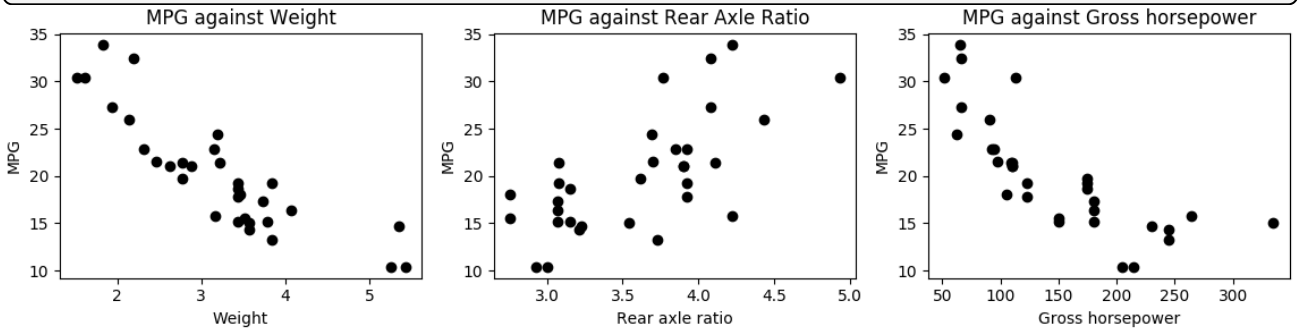
4) What is the coefficient of multiple determination?

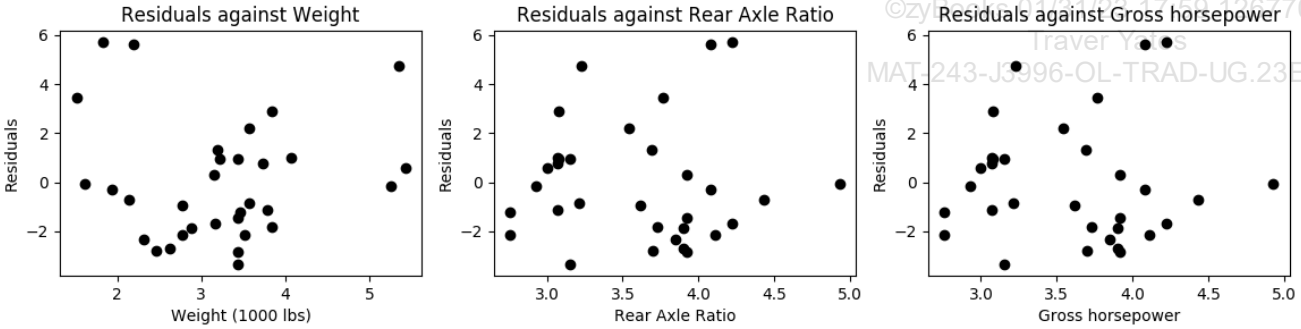
- 0.814
- 0.826
- 0.9292

Example 7.3.2: Multiple regression using three predictor variables.

```
import matplotlib.pyplot as plt
import pandas as pd
from statsmodels.formula.api import ols
import statsmodels.graphics.gofplots as smg

df = pd.read_csv("http://data-analytics.zybooks.com/mtcars.csv")
df_cars2 = df[["mpg","wt","drat","hp"]]

print("Correlation Matrix\n")
print (df_cars2.corr())
print("\n\n")

fig = plt.figure(figsize = (14, 7))
plt.subplots_adjust(hspace = 0.5)
plt.subplot(2, 3, 1)
plt.plot(df_cars2["wt"], df_cars2["mpg"], 'o', color='black')
plt.title('MPG against Weight')
plt.xlabel('Weight')
plt.ylabel('MPG')

plt.subplot(2, 3, 2)
plt.plot(df_cars2["drat"], df_cars2["mpg"], 'o', color='black')
plt.title('MPG against Rear Axle Ratio')
plt.xlabel('Rear axle ratio')
plt.ylabel('MPG')

plt.subplot(2, 3, 3)
plt.plot(df_cars2["hp"], df_cars2["mpg"], 'o', color='black')
plt.title('MPG against Gross horsepower')
plt.xlabel('Gross horsepower')
plt.ylabel('MPG')

model2 = ols('mpg ~ wt+drat+hp', data=df_cars2).fit()
print("Model\n")
print(model2.summary())
print("\n\n")

ax = fig.add_subplot(2, 3, 4)
plt.plot(df["wt"], model2.resid, 'o', color='black')
plt.title('Residuals against Weight')
plt.xlabel('Weight (1000 lbs)')
plt.ylabel('Residuals')

ax = fig.add_subplot(2, 3, 5)
plt.plot(df["drat"], model2.resid, 'o', color='black')
plt.title('Residuals against Rear Axle Ratio')
plt.xlabel('Rear Axle Ratio')
plt.ylabel('Residuals')
plt.show()

ax = fig.add_subplot(2, 3, 6)
plt.plot(df["drat"], model2.resid, 'o', color='black')
plt.title('Residuals against Gross horsepower')
plt.xlabel('Gross horsepower')
plt.ylabel('Residuals')

plt.show()
```

```
Correlation Matrix

           mpg        wt      drat        hp
mpg   1.000000 -0.867659  0.681172 -0.776168
wt   -0.867659  1.000000 -0.712441  0.658748
drat  0.681172 -0.712441  1.000000 -0.448759
hp   -0.776168  0.658748 -0.448759  1.000000



Model

                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.837
Model:                            OLS   Adj. R-squared:                  0.819
Method:                 Least Squares   F-statistic:                     47.88
Date:                Tue, 16 Jul 2019   Prob (F-statistic):           3.77e-11
Time:                        01:31:32   Log-Likelihood:                -73.366
No. Observations:                  32   AIC:                             154.7
Df Residuals:                      28   BIC:                             160.6
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     29.3949      6.156      4.775      0.000      16.784      42.006
wt            -3.2280      0.796     -4.053      0.000      -4.859      -1.597
drat           1.6150      1.227      1.316      0.199      -0.898       4.128
hp            -0.0322      0.009     -3.611      0.001      -0.051      -0.014
==============================================================================
Omnibus:                        5.200   Durbin-Watson:                   1.706
Prob(Omnibus):                  0.074   Jarque-Bera (JB):                4.289
Skew:                           0.896   Prob(JB):                        0.117
Kurtosis:                       3.080   Cond. No.                     2.25e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 2.25e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```



Run example

PARTICIPATION ACTIVITY 7.3.2: Multiple regression using three predictor variables.

Use the example above to answer the following.

1) Is rear axle ratio correlated with MPG?

- Yes
- No

2) Does a positively correlated relationship, a negatively correlated relationship, or no relationship at all exist between rear axle ratio and MPG?

- Positively correlated relationship
- Negatively correlated relationship
- No relationship

3) Does the data appear to satisfy the assumption of linearity?

- Yes
- No

4) Does the data appear to satisfy the assumption of homoscedasticity?

- Yes
- No

5) Let Y represent MPG, X_1 represent weight, X_2 represent rear axle ratio, and X_3 represent gross horsepower. What is the multiple regression model?

- $\hat{Y} = 29.3949 - 3.2280X_1 + 1.6150X_2 - 0.0322X_3$
- $\hat{Y} = 29.3949X_1 - 3.2280X_2 + 1.6150X_3 - 0.0322$
- $\hat{Y} = -3.2280X_1 + 1.6150X_2 - 0.0322X_3$

6) What is the coefficient of multiple determination?

- 0.819
- 0.837
- 0.896

Example 7.3.3: Multiple regression analysis of variance.


```
import pandas as pd
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

df = pd.read_csv("http://data-analytics.zybooks.com/mtcars.csv")
df_cars = df[["mpg","wt","qsec","disp","drat"]]

Y = df_cars['mpg']
X = df_cars[ ['wt'] + ['qsec'] ]
model = ols('Y ~ X', data=df_cars).fit()

anovaResults = anova_lm(model, typ = 2)
print(anovaResults)
```

```
              sum_sq    df          F        PR(>F)
X         930.583556   2.0  69.033106  9.394765e-12
Residual  195.463632  29.0        NaN           NaN
```

Run example

PARTICIPATION ACTIVITY 7.3.3: Coefficient of determination.

Consider an MPG model with two predictor variables, $X_1 =$ weight and $X_2 =$ 1/4 mile time. Use the analysis of variance table above to answer the following.

1) What is the variance explained by X_1 and X_2? Type as: ###.######

Check Show answer

2) What is the total variance? Type as: ####.######

Check Show answer

3) What percent of the variance in MPG can be explained from the variance in X_1 and X_2? Type as: ##.#%

Check Show answer

©zyBooks 01/31/23 17:59 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3