Biostat 212a Homework 1

Due Jan 28, 2025 @ 11:59PM

Wenqiang Ge UID:106371961

2025-01-28

Table of contents

Filling gaps in lecture notes (10% pts) 1
ISL Exercise 2.4.3 (10% pts) 3
ISL Exercise 2.4.4 (10% pts) 6
ISL Exercise 2.4.10 (30% pts) 7
ISL Exercise 3.7.3 (20% pts) 15
3.7.15 (20% pts) 17
Bonus question (20% pts) 34

Filling gaps in lecture notes (10% pts)

Consider the regression model

$$Y = f(X) + \epsilon,$$

where $\mathrm{E}(\epsilon) = 0$.

Optimal regression function

Show that the choice

$$f_{\text{opt}}(X) = \mathrm{E}(Y|X)$$

minimizes the mean squared prediction error

$$\mathrm{E}\{[Y - f(X)]^2\},$$

where the expectations averages over variations in both X and Y. (Hint: condition on X.)

Bias-variance trade-off

Given an estimate $\hat{f}$ of f, show that the test error at a x_0 can be decomposed as

$$\mathrm{E}\{[y_0 - \hat{f}(x_0)]^2\} = \underbrace{\mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2}_{\text{MSE of } \hat{f}(x_0) \text{ for estimating } f(x_0)} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{irreducible}},$$

where the expectation averages over the variability in y_0 and $\hat{f}$.

Optimal regression function:

Optimal regression function.

We have $f_{opt}(X) = E(Y|X)$, $(Y - f(X))^2 = (Y - E(Y|X) + E(Y|X) - f(X))^2$

then $(Y - f(X))^2 = (Y - E[Y|X))^2 + 2(Y - E(Y|X)) \cdot (E(Y|X) - f(X))$
$+ (E(Y|X) - f(X))^2$

since $E[Y - E(Y|X)|X] = 0$, then $2(Y - E(Y|X)) \cdot (E(Y|X) - f(X)) = 0$

So we have $E[(Y - f(X))^2|X] = E[(Y - E(Y|X)^2|X] + (E[Y|X] - f(X))^2$

Then $E[(Y - f(X))^2] = E[(Y - E(Y|X)^2|X] + E[(E(Y|X) - f(X))^2]$

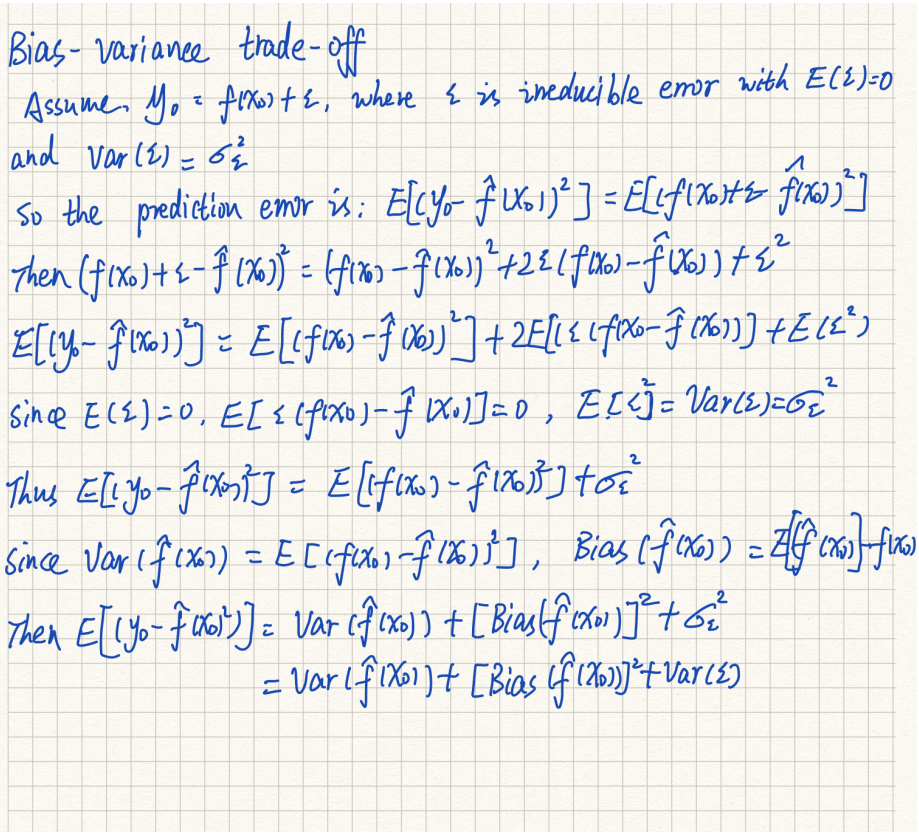
For $E[(Y - E(Y|X)^2|X]$, it's irreducible.

For second one, when $E(Y|X) = f(X)$, it is the minimum (0).

Therefore when $f(X) = E(Y|X)$, the mean squared prediction error: $E[(Y - f(X))^2]$ will be minimized.

$f_{opt}(X) = E(Y|X)$.

Bias-variance trade-off:



ISL Exercise 2.4.3 (10% pts)

```
library(tidyverse)
fit <- lm(sales ~ TV, data = )
```

3. We now revisit the bias-variance decomposition.

   (a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

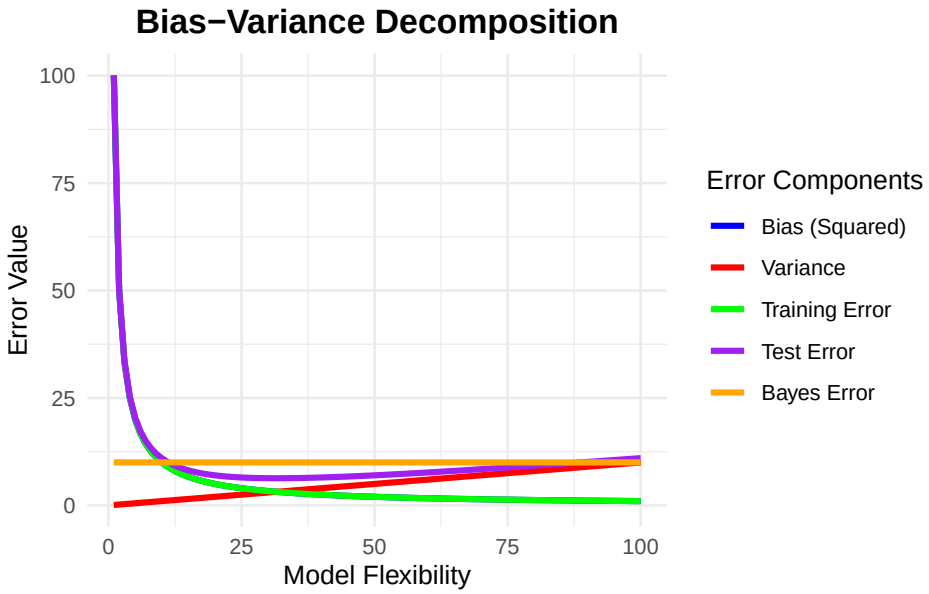
   (b) Explain why each of the five curves has the shape displayed in part (a).

(a)

```
library(ggplot2)

Flexibility <- 1:100
Bias <- 100 / Flexibility
Variance <- Flexibility / 10
TrainingError <- 100 / Flexibility
TestError <- Bias + Variance
BayesError <- rep(10, 100)

# Combine into a data frame
data <- data.frame(
  Flexibility = Flexibility,
  Bias = Bias,
  Variance = Variance,
  TrainingError = TrainingError,
  TestError = TestError,
  BayesError = BayesError
)
data_long <- reshape2::melt(data, id.vars = "Flexibility", variable.name = "ErrorType", v

ggplot(data_long, aes(x = Flexibility, y = ErrorValue, color = ErrorType)) +
  geom_line(size = 1.2) +
  labs(
    title = "Bias-Variance Decomposition",
    x = "Model Flexibility",
    y = "Error Value",
    color = "Error Components"
  ) +
```

```
  scale_color_manual(
    values = c("blue", "red", "green", "purple", "orange"),
    labels = c("Bias (Squared)", "Variance", "Training Error", "Test Error", "Bayes Error
  ) +
  theme_minimal() +
  theme(
    legend.position = "right",
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5)
  )
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



(b) Bias Curve: Decreases because as flexibility increases, the model can better fit the training data, reducing systematic error.Variance Curve: Increases because more flexible models are more sensitive to small fluctuations in the data, leading to overfitting.Training Error Curve: Always decreases because more flexible models can perfectly fit (or nearly fit) the training data.Test Error Curve: U-shaped because it is influenced by both bias and variance. Initially, test error decreases as bias dominates. Later, it increases as variance dominates.Bayes Error: Stays constant as it represents noise or irreducible error in the data.

ISL Exercise 2.4.4 (10% pts)

4. You will now think of some real-life applications for statistical learning.

   (a) Describe three real-life applications in which *classification* might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

   (b) Describe three real-life applications in which *regression* might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

   (c) Describe three real-life applications in which *cluster analysis* might be useful.

(a) Classification application:

Spam email detection: Response: Email is spam (1) or not (0). Predictors: Frequency of certain keywords, length of the email, etc. Goal: Prediction. The model is used to predict whether new emails are spam or not.

Medical Diagnosis: Response: Whether a patient has a disease Yes (1) or not (0).

Predictors: Age, gender, symptoms, test results, and medical history. Goal: Inference and prediction. Inference is used to understand which predictors are most associated with the disease. Prediction is used to diagnose new patients.

Disease classification: Response variable: the disease classification of the patient, such as diabetes (1) , heart disease (2) , health (0) . Predictive variables: age, blood glucose level, cholesterol level, medical history, lifestyle, etc. Goal: Predict. Assist doctors in making quick diagnostic decisions.

(b) Regression application:

Drug dosage and efficacy: Response: Drug efficacy (such as decreased blood glucose levels). Predictive : drug dosage, patient age, weight, etc. Goal: Inference. Assist in drug research, analyze the relationship between dosage and efficacy.

Real estate rent forecast:Response : Rent price. Predictive : geographical location, area, decoration level, surrounding facilities, etc. Goal: Predict. Provide reference prices for the rental market.

Stock market analysis:Response: Future stock price or return. Predictive : historical prices, trading volume, economic indicators, and news sentiment. Goal: Predict. This model helps predict stock prices for investment decisions.

(c) Cluster application:

Grouping students based on their academic performance and learning behavior. Features: classroom performance, exam scores, participation, completion of assignments, etc. Goal: To assist teachers in developing teaching plans for different groups. Genotyping analysis

Grouping genes based on DNA sequence data. Features: gene expression level, sequence similarity, etc. Goal: To discover different types of genomic populations for disease research. Retail store location selection

Grouping urban areas based on population density and consumption behavior. Features: Population characteristics (age, income), traffic flow, consumption level, etc. Goal: Help retailers choose the best location.

ISL Exercise 2.4.10 (30% pts)

Your can read in the `boston` data set directly from url https://raw.githubusercontent.com/ucla-biostat-212a/2024winter/master/slides/data/Boston.csv. A documentation of the `boston` data set is here.

R

```
library(tidyverse)
Boston <- read_csv("https://raw.githubusercontent.com/ucla-biostat-212a/2024winter/master
  print(width = Inf)
```

```
# A tibble: 506 x 13
      crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio lstat
     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl>
 1 0.00632  18    2.31     0 0.538  6.58  65.2  4.09     1   296    15.3  4.98
 2 0.0273    0    7.07     0 0.469  6.42  78.9  4.97     2   242    17.8  9.14
 3 0.0273    0    7.07     0 0.469  7.18  61.1  4.97     2   242    17.8  4.03
 4 0.0324    0    2.18     0 0.458  7.00  45.8  6.06     3   222    18.7  2.94
 5 0.0690    0    2.18     0 0.458  7.15  54.2  6.06     3   222    18.7  5.33
 6 0.0298    0    2.18     0 0.458  6.43  58.7  6.06     3   222    18.7  5.21
 7 0.0883   12.5  7.87     0 0.524  6.01  66.6  5.56     5   311    15.2 12.4
 8 0.145    12.5  7.87     0 0.524  6.17  96.1  5.95     5   311    15.2 19.2
 9 0.211    12.5  7.87     0 0.524  5.63 100    6.08     5   311    15.2 29.9
10 0.170    12.5  7.87     0 0.524  6.00  85.9  6.59     5   311    15.2 17.1
    medv
   <dbl>
 1  24
 2  21.6
 3  34.7
 4  33.4
 5  36.2
```

```
 6  28.7
 7  22.9
 8  27.1
 9  16.5
10  18.9
# i 496 more rows
```

Python

```
# import pandas as pd
# import io
# import requests
#
# url = "https://raw.githubusercontent.com/ucla-econ-425t/2023winter/master/slides/data/B
# s = requests.get(url).content
# Boston = pd.read_csv(io.StringIO(s.decode('utf-8')), index_col = 0)
# Boston
```

10. This exercise involves the `Boston` housing data set.

    (a) To begin, load in the `Boston` data set, which is part of the `ISLP` library.

    (b) How many rows are in this data set? How many columns? What do the rows and columns represent?

    (c) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

    (d) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

    (e) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

    (f) How many of the suburbs in this data set bound the Charles river?

    (g) What is the median pupil-teacher ratio among the towns in this data set?

    (h) Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

    (i) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

(a)

```
library(tidyverse)
Boston <- read_csv("https://raw.githubusercontent.com/ucla-biostat-212a/2024winter/master
  print(width = Inf)
```

```
# A tibble: 506 x 13
      crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio lstat
     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl>
 1 0.00632  18    2.31     0 0.538  6.58  65.2  4.09     1   296    15.3  4.98
 2 0.0273    0    7.07     0 0.469  6.42  78.9  4.97     2   242    17.8  9.14
 3 0.0273    0    7.07     0 0.469  7.18  61.1  4.97     2   242    17.8  4.03
 4 0.0324    0    2.18     0 0.458  7.00  45.8  6.06     3   222    18.7  2.94
 5 0.0690    0    2.18     0 0.458  7.15  54.2  6.06     3   222    18.7  5.33
 6 0.0298    0    2.18     0 0.458  6.43  58.7  6.06     3   222    18.7  5.21
 7 0.0883   12.5  7.87     0 0.524  6.01  66.6  5.56     5   311    15.2 12.4
 8 0.145    12.5  7.87     0 0.524  6.17  96.1  5.95     5   311    15.2 19.2
 9 0.211    12.5  7.87     0 0.524  5.63 100    6.08     5   311    15.2 29.9
10 0.170    12.5  7.87     0 0.524  6.00  85.9  6.59     5   311    15.2 17.1
    medv
   <dbl>
 1  24
 2  21.6
 3  34.7
 4  33.4
 5  36.2
 6  28.7
 7  22.9
 8  27.1
 9  16.5
10  18.9
# i 496 more rows
```

Done!
(b)

```
nrow(Boston)
```

```
[1] 506
```

```
ncol(Boston)
```

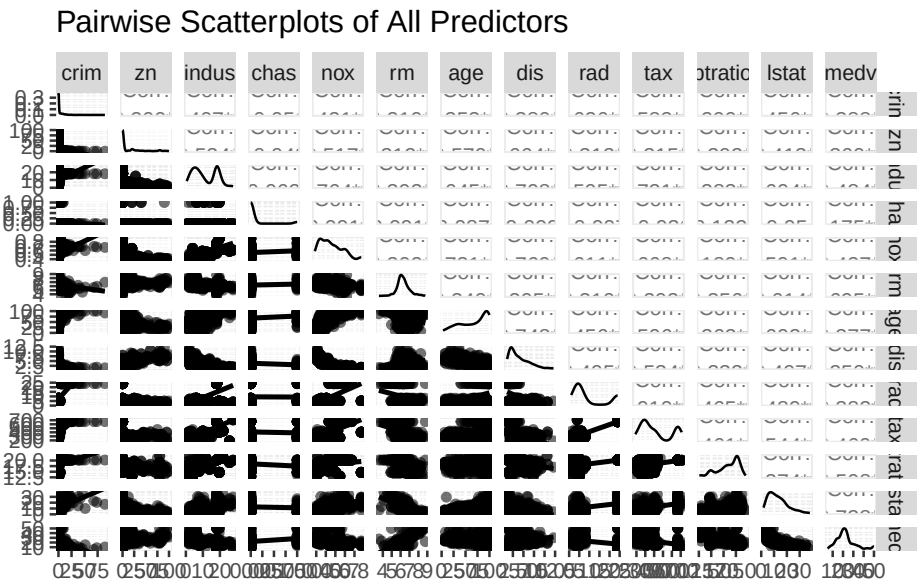
```
[1] 13
```

There are 506 rows and 13 columns. Rows represent suburbs or towns in the Boston area. Each column represents a feature or response variable (e.g., crime

rate, tax rate, median value of homes).

(c)

```
library(ISLR2)
library(GGally)
library(ggplot2)
ggpairs(
  data = Boston,
  mapping = aes(alpha = 0.5),
  upper = list(continuous = wrap("cor", size = 3)),
  lower = list(continuous = wrap("smooth", method = "lm", se = FALSE))
) +
  labs(title = "Pairwise Scatterplots of All Predictors")
```



The correlation coefficient between Zn and Crimea is -0.200, showing a weak negative correlation. When Zn increases, there is a slight downward trend in Crimea.Indus is moderately positively correlated with Crimea. A higher proportion of industrial land is associated with a higher crime rate.chas: The correlation coefficient with Crimea is -0.056, indicating no significant linear relationship.The correlation coefficient between NOx and Crimea is 0.421, showing a moderate positive correlation. Higher concentrations of nitric oxide may be associated with higher crime rates.rm: The correlation coefficient with Crimea is -0.219, showing a weak negative correlation. The crime rate is higher when the average number of rooms is small.

The correlation coefficient between Indus and NOx is 0.764, showing a strong positive correlation, indicating that the higher the proportion of industrial land, the more severe the air pollution (NOx).RM is negatively correlated with both Indus and NOX, indicating that houses with more rooms are usually located in areas with lower industrial land ratios and less pollution.Distribution shape and pattern:Some variables, such as zn and crim, exhibit nonlinear patterns and may require further exploration of their nonlinear relationships.

(d)

```
library(MASS)
data(Boston)
results <- data.frame(Predictor = character(), Coefficient = numeric(), P_Value = numeric
for (predictor in colnames(Boston)[-1]) {  # Exclude `crim` as it's the response variable
  model <- lm(crim ~ Boston[[predictor]], data = Boston)
  summary_model <- summary(model)
  results <- rbind(results, data.frame(
    Predictor = predictor,
    Coefficient = coef(summary_model)[2, 1],  # Slope (relationship strength and directio
    P_Value = coef(summary_model)[2, 4]      # P-value (significance)
  ))
}

significant_results <- subset(results, P_Value < 0.05)
print("Significant Predictors Associated with Crime Rate:")
```

```
[1] "Significant Predictors Associated with Crime Rate:"
```

```
print(significant_results)
```

```
   Predictor Coefficient      P_Value
1         zn -0.07393498 5.506472e-06
2      indus  0.50977633 1.450349e-21
4        nox 31.24853120 3.751739e-23
5         rm -2.68405122 6.346703e-07
6        age  0.10778623 2.854869e-16
7        dis -1.55090168 8.519949e-19
8        rad  0.61791093 2.693844e-56
9        tax  0.02974225 2.357127e-47
10   ptratio  1.15198279 2.942922e-11
11     black -0.03627964 2.487274e-19
12     lstat  0.54880478 2.654277e-27
13      medv -0.36315992 1.173987e-19
```

The crime rate is significantly positively correlated with factors such as the industrial land ratio (indus), nitric oxide concentration (NOx), age of old houses,

highway accessibility (rad), property tax rate (tax), and student teacher ratio (ptratio). This indicates that areas with high levels of economic industrialization, severe air pollution, convenient transportation, but limited educational resources have higher crime rates.

The crime rate is significantly negatively correlated with the proportion of large-scale residential land (Zn), the average number of rooms in a house (RM), the weighted distance from the employment center (DIS), and the proportion of black people (Black). These results indicate that low-density residential communities, higher housing quality, areas further away from urban employment centers, and areas with a higher proportion of black people have lower crime rates.

(e)

```
summary(Boston$crim)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```

```
summary(Boston$tax)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  187.0   279.0   330.0   408.2   666.0   711.0
```

```
summary(Boston$ptratio)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.60   17.40   19.05   18.46   20.20   22.00
```

```
high_crime <- Boston[Boston$crim > quantile(Boston$crim, 0.8), ]
high_tax <- Boston[Boston$tax > quantile(Boston$tax, 0.8), ]
high_ptratio <- Boston[Boston$ptratio > quantile(Boston$ptratio, 0.8), ]

cat("High Crime:", nrow(high_crime),
    "\nHigh Tax:", nrow(high_tax),
    "\nHigh Pupil-Teacher Ratio:", nrow(high_ptratio), "\n")
```

```
High Crime: 101
High Tax: 5
High Pupil-Teacher Ratio: 56
```

The distribution of crime rates is extremely uneven, with some suburban areas (such as crime rates>3.67708) being high crime areas, with the highest values far above the average, and there are obvious extreme values (such as the highest value of 88.97620).

The tax rate distribution is relatively concentrated, with most suburbs having tax rates ranging from 187 to 666, and only a few suburbs (such as>666) at high tax rates.

The distribution of student teacher ratio is relatively even, with only some suburban areas experiencing significant shortage of educational resources (e.g.>20)

(f)

```
sum(Boston$chas == 1)
```

```
[1] 35
```

There are 35 suburbs bound the Charles river.

(g)

```
median(Boston$ptratio, na.rm = TRUE)
```

```
[1] 19.05
```

The median pupil-teacher ratio among the towns is 19.05.

(h)

```
lowest_medv <- Boston[which.min(Boston$medv), ]
lowest_medv
```

```
      crim zn indus chas   nox    rm age    dis rad tax ptratio black lstat
399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.9 30.59
    medv
399    5
```

The high crime rate is an important factor in the decline of housing prices.A high proportion of old buildings may reduce their attractiveness.Severe air pollution and high level of industrialization: have a negative impact on the quality of living environment.High tax rates and limited educational resources will increase the cost of living and reduce the attractiveness of housing.

(i)

```
over_7_rooms <- sum(Boston$rm > 7)
over_8_rooms <- sum(Boston$rm > 8)

over_7_rooms
```

```
[1] 64
```

```
  over_8_rooms
```

```
[1] 13
```

```
  Boston[Boston$rm > 8, ]
```

```
       crim zn indus chas    nox    rm  age    dis rad tax ptratio  black lstat
98  0.12083  0  2.89    0 0.4450 8.069 76.0 3.4952   2 276    18.0 396.90  4.21
164 1.51902  0 19.58    1 0.6050 8.375 93.9 2.1620   5 403    14.7 388.45  3.32
205 0.02009 95  2.68    0 0.4161 8.034 31.9 5.1180   4 224    14.7 390.55  2.88
225 0.31533  0  6.20    0 0.5040 8.266 78.3 2.8944   8 307    17.4 385.05  4.14
226 0.52693  0  6.20    0 0.5040 8.725 83.0 2.8944   8 307    17.4 382.00  4.63
227 0.38214  0  6.20    0 0.5040 8.040 86.5 3.2157   8 307    17.4 387.38  3.13
233 0.57529  0  6.20    0 0.5070 8.337 73.3 3.8384   8 307    17.4 385.91  2.47
234 0.33147  0  6.20    0 0.5070 8.247 70.4 3.6519   8 307    17.4 378.95  3.95
254 0.36894 22  5.86    0 0.4310 8.259  8.4 8.9067   7 330    19.1 396.90  3.54
258 0.61154 20  3.97    0 0.6470 8.704 86.9 1.8010   5 264    13.0 389.70  5.12
263 0.52014 20  3.97    0 0.6470 8.398 91.5 2.2885   5 264    13.0 386.86  5.91
268 0.57834 20  3.97    0 0.5750 8.297 67.0 2.4216   5 264    13.0 384.54  7.44
365 3.47428  0 18.10    1 0.7180 8.780 82.9 1.9047  24 666    20.2 354.55  5.29
    medv
98  38.7
164 50.0
205 50.0
225 44.8
226 50.0
227 37.6
233 41.7
234 48.3
254 42.8
258 50.0
263 48.8
268 50.0
365 21.9
```

There are 64 and 13 suburbs average more than seven and eight rooms per dwelling.

The crime rate in most suburbs is very low, with a minimum of 0.01208. Some suburbs have a high proportion of large residential areas (such as the suburbs with a Zn of 95). The lower concentration of nitric oxide indicates lower air pollution in these areas (such as NOx mostly ranging from 0.4 to 0.6). The proportion of house updates is relatively low in some areas, such as some suburban areas with an age of less than 20. These suburbs are often far from employment centers, indicating that their locations are more remote.

ISL Exercise 3.7.3 (20% pts)

3. Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$.

   (a) Which answer is correct, and why?

      i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

      ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

      iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

      iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

   (b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

   (c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

(a)-(b)

3.7.3

(a) suppose the model: Y (starting salary) $= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$

where $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.

so $Y = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_4 - 10X_5$

High schul Graduates: $X_3 = 0$: $(X_5 = X_1 \cdot X_3 = 0)$

$Y_1 = 50 + 20X_1 + 0.07X_2 + 0.01X_4$

College Graduates: $X_3 = 1$: $(X_5 = X_1 \cdot X_3 = X_1 \cdot 1 = X_1)$

$Y_2 = 50 + 20X_1 + 0.07X_2 + 35 + 0.01X_4 - 10X_5$

For fixed X_1 and X_2 : $Y_{diff} = Y_2 - Y_1 = 35 - 10X_5 = 35 - 10X_1$

so $Y_2 > Y_1$: $35 - 10X_1 > 0 \Rightarrow X_1 < 3.5 \Rightarrow$ College Graduates earn more.

$Y_2 < Y_1$: $35 - 10X_1 < 0 \Rightarrow X_1 > 3.5 \Rightarrow$ High schul Graduates earn more

(i) and (ii): without GPA limits.

(iv): College Graduates should have low GPA.

(b) Since we have $X_1 = 4.0$, $X_2 = 110$, $X_3 = 1$, $X_4 = X_1 \cdot X_2 = 440$, $X_5 = X_1 \cdot X_3 = 4.0$

so $Y = 50 + 20 \times 4.0 + 0.07 \times 110 + 35 \times 1 + 0.01 \times 440 - 10 \times 4.0$

$= 50 + 80 + 7.7 + 35 + 4.4 - 40$

$= 137.1$ (thousands of dollars) $= \$137100$

(c) True. The coefficient for the GPA/IQ interaction term is very small, suggesting that the interaction effect is minimal. The contribution of $\hat{\beta}X$ 4 =0.01 640=6.4, which is relatively small compared to other terms.

3.7.15 (20% pts)

15. This problem involves the `Boston` data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

    (a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

    (b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

    (c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

    (d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$

(a)

```
library(MASS)
data(Boston)

simple_reg_results <- data.frame(Predictor = character(), Coefficient = numeric(), P_Valu

for (predictor in colnames(Boston)[-1]) {
  model <- lm(crim ~ Boston[[predictor]], data = Boston)
  summary_model <- summary(model)
  simple_reg_results <- rbind(simple_reg_results, data.frame(
    Predictor = predictor,
    Coefficient = coef(summary_model)[2, 1],
    P_Value = coef(summary_model)[2, 4]
  ))
```

```
}

significant_predictors <- subset(simple_reg_results, P_Value < 0.05)
print(significant_predictors)
```

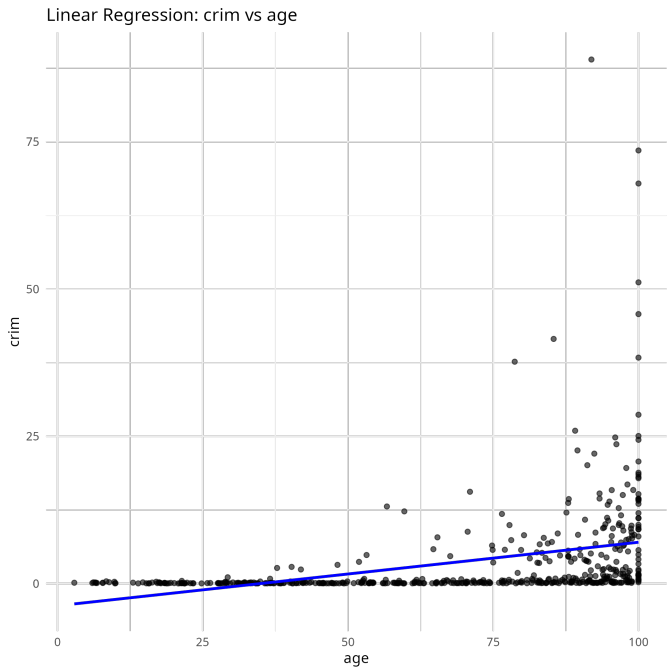
```
   Predictor Coefficient      P_Value
1         zn -0.07393498 5.506472e-06
2      indus  0.50977633 1.450349e-21
4        nox 31.24853120 3.751739e-23
5         rm -2.68405122 6.346703e-07
6        age  0.10778623 2.854869e-16
7        dis -1.55090168 8.519949e-19
8        rad  0.61791093 2.693844e-56
9        tax  0.02974225 2.357127e-47
10   ptratio  1.15198279 2.942922e-11
11     black -0.03627964 2.487274e-19
12     lstat  0.54880478 2.654277e-27
13      medv -0.36315992 1.173987e-19
```

```
library(ggplot2)
output_folder <- "crim_plots"
if (!dir.exists(output_folder)) {
  dir.create(output_folder)
}

for (predictor in significant_predictors$Predictor) {
  p <- ggplot(Boston, aes(x = .data[[predictor]], y = crim)) +
    geom_point(alpha = 0.6) +
    geom_smooth(method = "lm", color = "blue", se = FALSE) +
    ggtitle(paste("Linear Regression: crim vs", predictor)) +
    theme_minimal()
  plot_path <- file.path(output_folder, paste0("crim_vs_", predictor, ".png"))
  ggsave(filename = plot_path, plot = p, width = 7, height = 7)
}
```

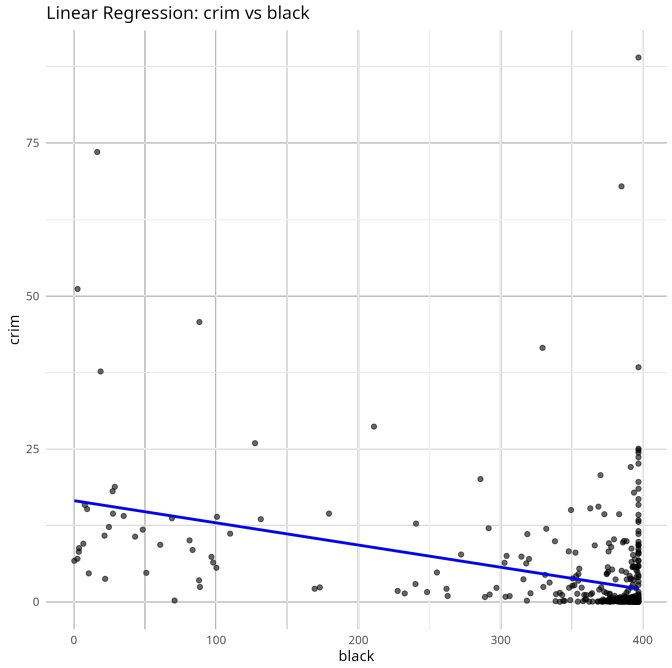
```
library(png)
library(grid)
saved_files <- list.files(output_folder, pattern = "\\.png$", full.names = TRUE)

for (file in saved_files) {
  cat("Displaying:", file, "\n")
  img <- png::readPNG(file)
  grid::grid.newpage()
```

```
  grid::grid.raster(img)
}
```

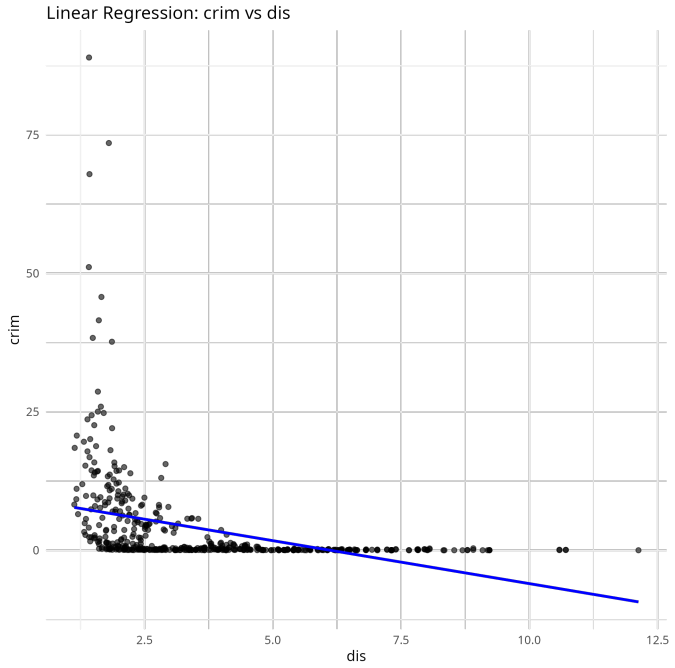
```
Displaying: crim_plots/crim_vs_age.png
```



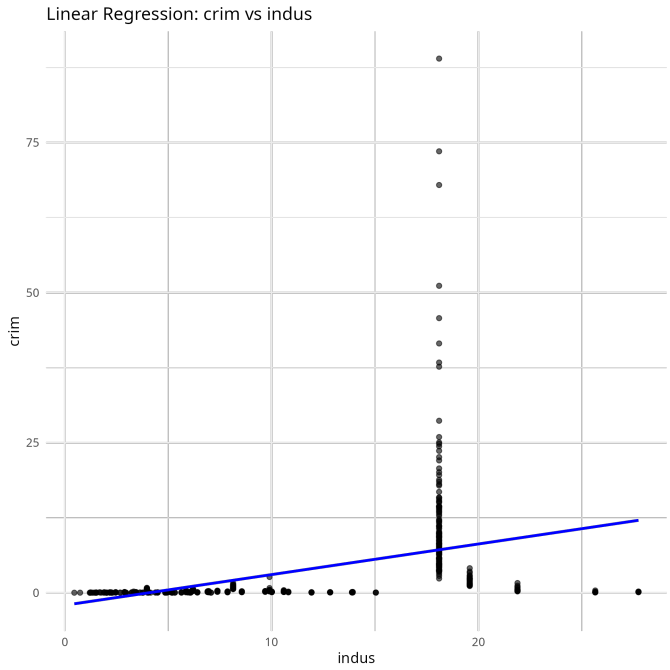

```
Displaying: crim_plots/crim_vs_black.png
```



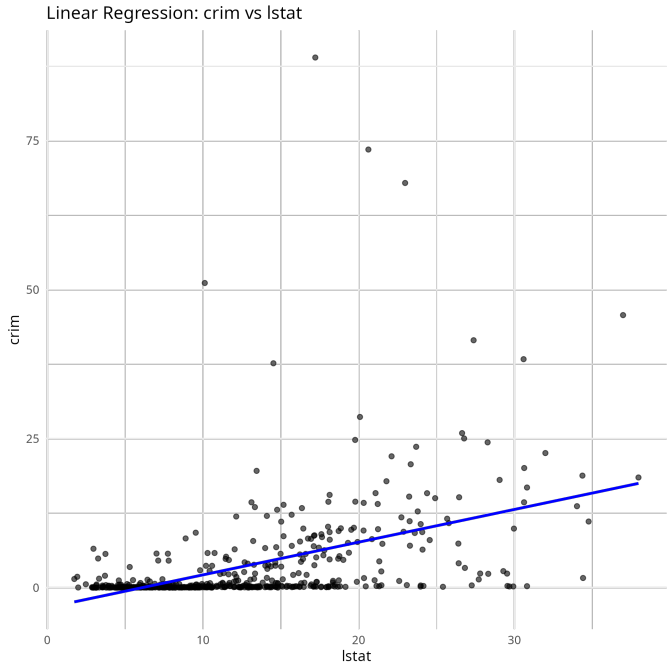
`Displaying: crim_plots/crim_vs_dis.png`



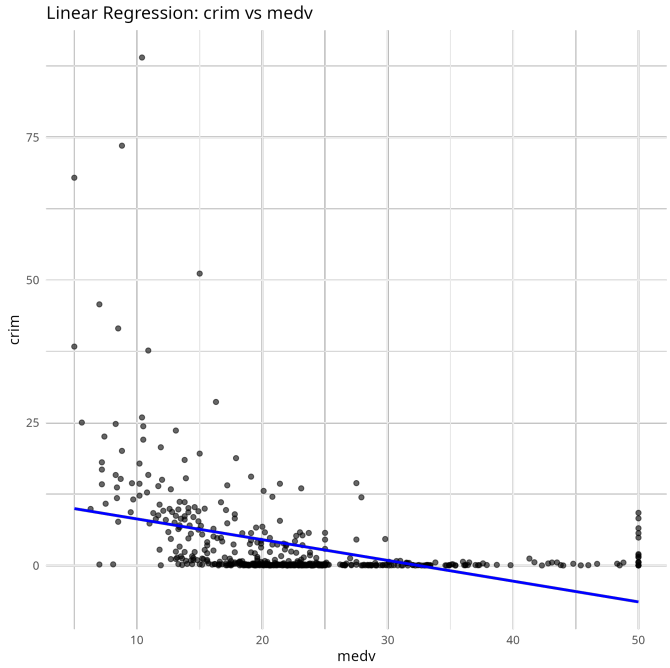
```
Displaying: crim_plots/crim_vs_indus.png
```



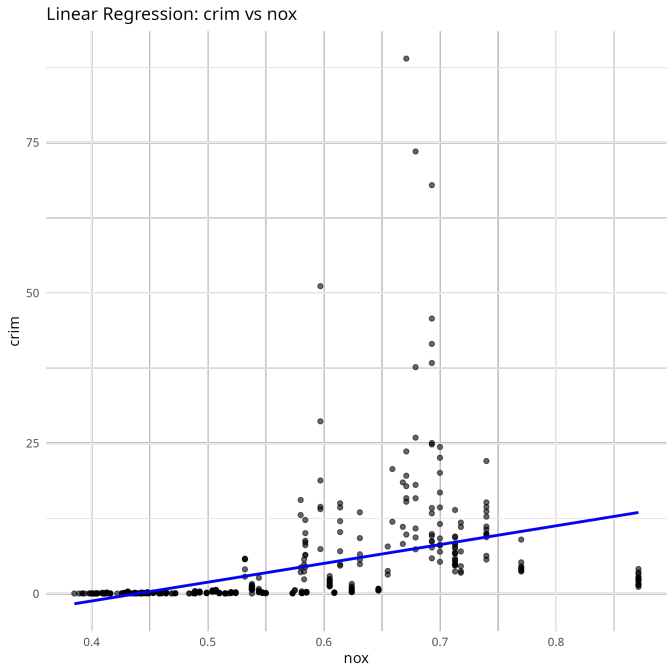

```
Displaying: crim_plots/crim_vs_lstat.png
```



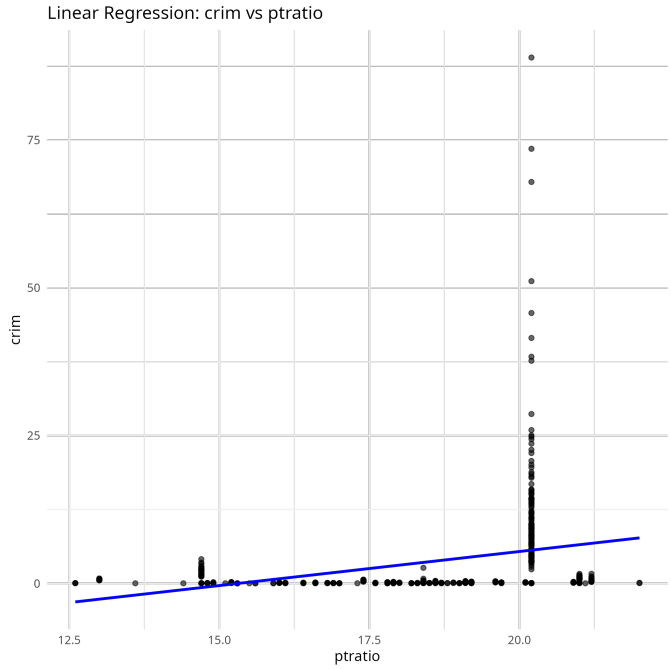
`Displaying: crim_plots/crim_vs_medv.png`



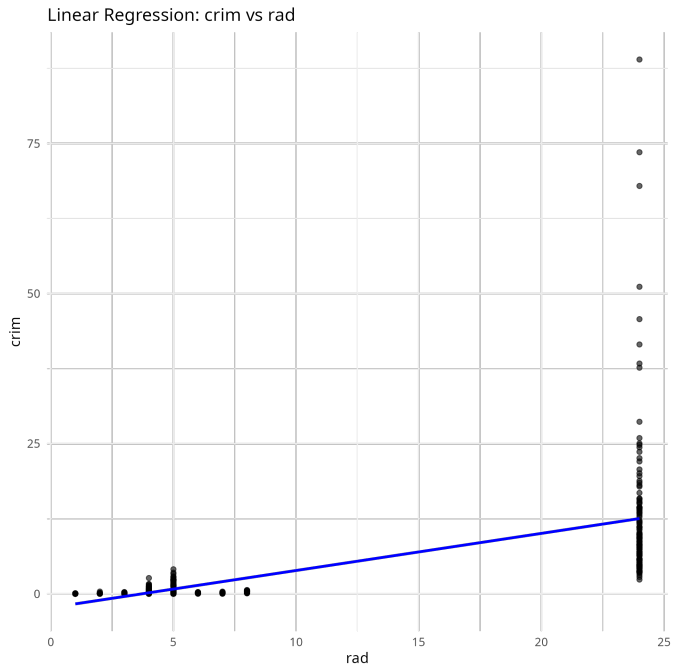
```
Displaying: crim_plots/crim_vs_nox.png
```



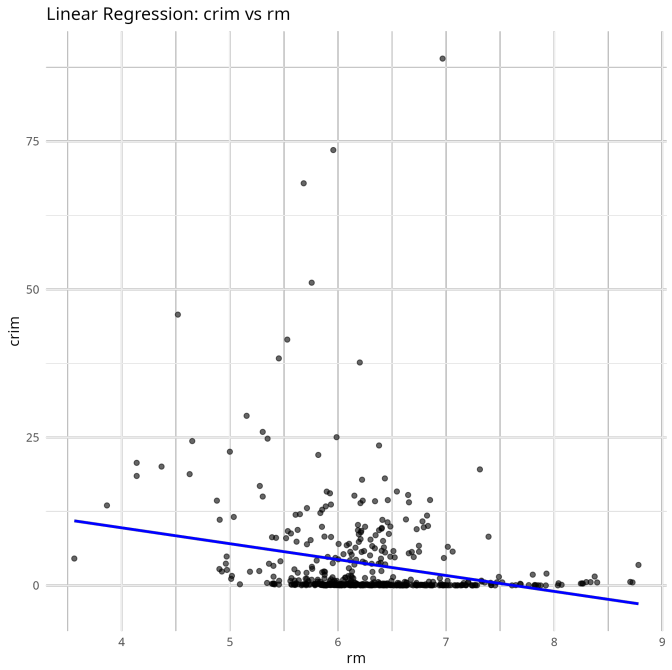

```
Displaying: crim_plots/crim_vs_ptratio.png
```



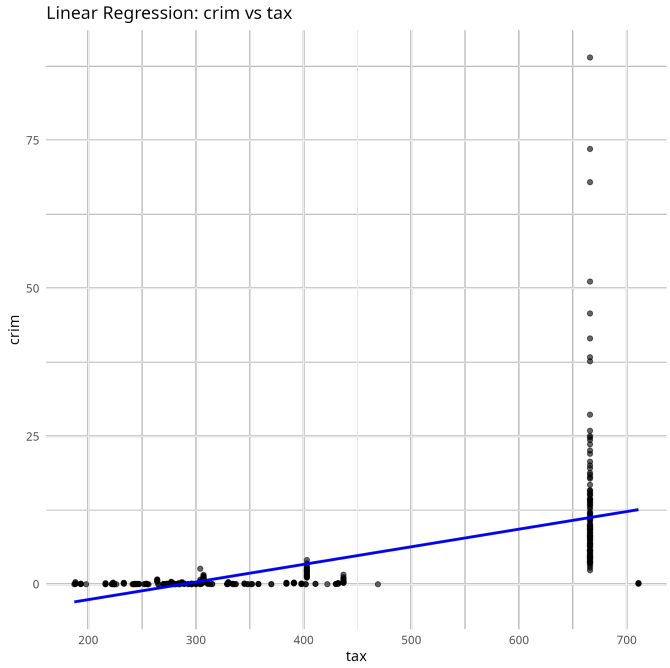
`Displaying: crim_plots/crim_vs_rad.png`



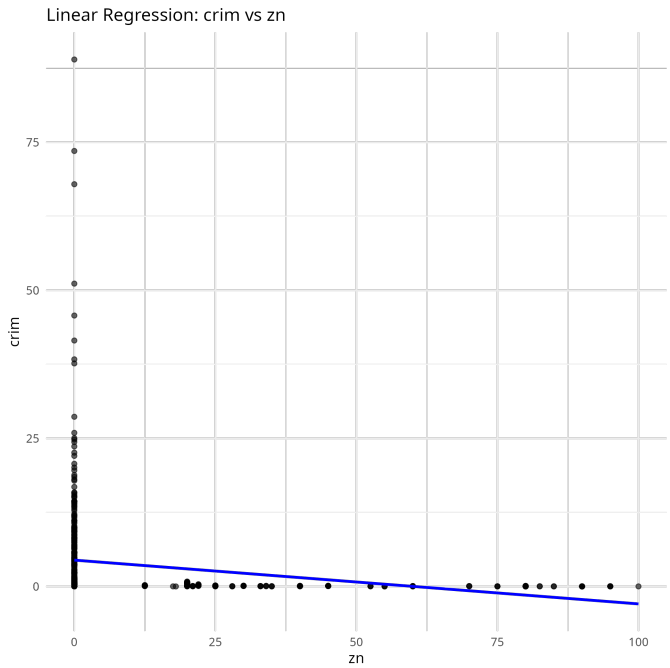
```
Displaying: crim_plots/crim_vs_rm.png
```




```
Displaying: crim_plots/crim_vs_tax.png
```



`Displaying: crim_plots/crim_vs_zn.png`



zn: Negative correlation indicates that a higher proportion of large-scale residential land is associated with lower crime rates. indus: Positive correlation, the higher the proportion of industrial land, the higher the crime rate. nox: Strong positive correlation indicates a significant correlation between air pollution level (nitric oxide concentration) and crime rate. rm: Negative correlation, the more average rooms there are, the lower the crime rate. age: Positive correlation, the higher the proportion of old houses, the slightly higher the crime rate. dis: Negative correlation, the farther away from the employment center, the lower the crime rate. rad: Strong positive correlation, the higher the radiation radius index (indicating high accessibility), the higher the crime rate. tax: Positive correlation, the higher the property tax rate, the slightly higher the crime rate. ptratio: Positive correlation, the higher the student teacher ratio, the higher the crime rate. black: Negative correlation indicates that areas with a higher proportion of black people have lower crime rates.

In all of models, there is a statistically significant association between the preditctor and response.

(b)

```r
library(MASS)
data(Boston)

multi_model <- lm(crim ~ ., data = Boston)

summary_multi_model <- summary(multi_model)
print(summary_multi_model)
```

```
Call:
lm(formula = crim ~ ., data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-9.924 -2.120 -0.353  1.019 75.051

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.033228   7.234903   2.354 0.018949 *
zn            0.044855   0.018734   2.394 0.017025 *
indus        -0.063855   0.083407  -0.766 0.444294
chas         -0.749134   1.180147  -0.635 0.525867
nox         -10.313535   5.275536  -1.955 0.051152 .
rm            0.430131   0.612830   0.702 0.483089
age           0.001452   0.017925   0.081 0.935488
dis          -0.987176   0.281817  -3.503 0.000502 ***
```

```
rad           0.588209   0.088049   6.680 6.46e-11 ***
tax          -0.003780   0.005156  -0.733 0.463793
ptratio      -0.271081   0.186450  -1.454 0.146611
black        -0.007538   0.003673  -2.052 0.040702 *
lstat         0.126211   0.075725   1.667 0.096208 .
medv         -0.198887   0.060516  -3.287 0.001087 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454, Adjusted R-squared:  0.4396
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

```r
multi_reg_results <- data.frame(
  Predictor = rownames(summary_multi_model$coefficients),
  Coefficient = summary_multi_model$coefficients[, 1],
  p_Value = summary_multi_model$coefficients[, 4]
)

significant_vars <- multi_reg_results$Predictor[multi_reg_results$p_Value < 0.05]

print("Significant Predictors:")
```

```
[1] "Significant Predictors:"
```

```r
print(significant_vars)
```

```
[1] "(Intercept)" "zn"          "dis"         "rad"         "black"
[6] "medv"
```

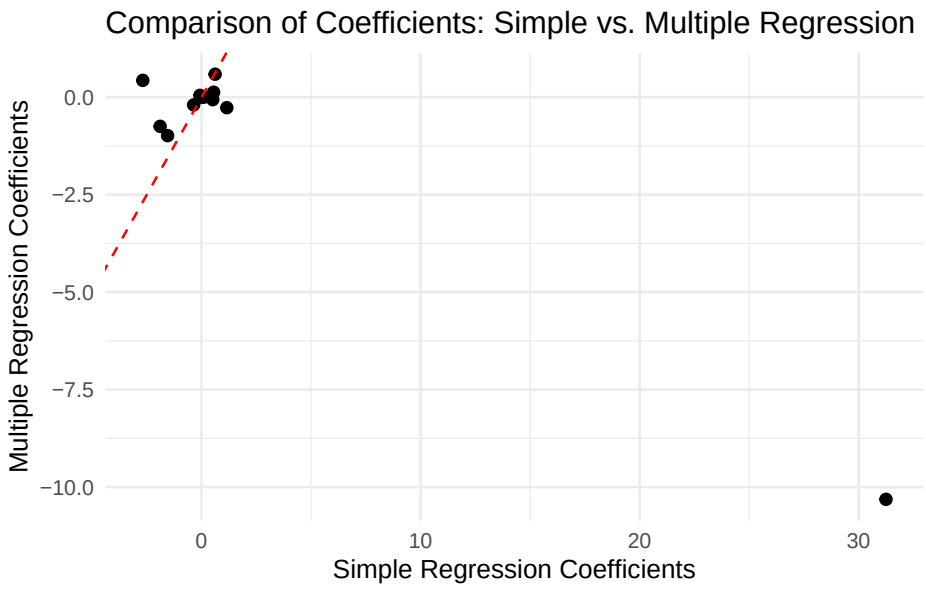
zn:The regression coefficient is 0.0448, indicating a positive correlation between the proportion of large-scale residential land and the crime rate.dis:The regression coefficient is -0.9872, indicating that the further away from the employment center, the lower the crime rate (negative correlation).rad:The regression coefficient is 0.5882, and the higher the radiation index, the higher the crime rate (positive correlation).black:The regression coefficient is -0.0075, indicating a negative correlation between the proportion of black people and the crime rate.medv:The regression coefficient is -0.1988, indicating that the higher the median housing price, the lower the crime rate (negative correlation).

The p-values of Indus, Chas, NOx, RM, Age, Tax, and PTRatio are greater than 0.05, therefore they cannot be considered significantly correlated with crime rates.

The p-values of zn, dis, rad, black, medv are less than 0.05, so they can reject the null hypothesis.

(c)

```r
# Merge results from (a) and (b)
merged_results <- merge(
  simple_reg_results,
  multi_reg_results,
  by.x = "Predictor",
  by.y = "Predictor",
  suffixes = c("_Simple", "_Multiple")
)

ggplot(merged_results, aes(x = Coefficient_Simple, y = Coefficient_Multiple)) +
  geom_point(size = 2) +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  ggtitle("Comparison of Coefficients: Simple vs. Multiple Regression") +
  xlab("Simple Regression Coefficients") +
  ylab("Multiple Regression Coefficients") +
  theme_minimal()
```



(d)

```r
nonlinear_results <- data.frame(Predictor = character(), Linear_P = numeric(), Quadratic_

for (predictor in colnames(Boston)[-1]) {
```

```
  model <- lm(crim ~ Boston[[predictor]] + I(Boston[[predictor]]^2) + I(Boston[[predictor
  summary_model <- summary(model)

  coef_matrix <- coef(summary_model)
  if (nrow(coef_matrix) >= 4) {
    nonlinear_results <- rbind(nonlinear_results, data.frame(
      Predictor = predictor,
      Linear_P = coef_matrix[2, 4],
      Quadratic_P = coef_matrix[3, 4],
      Cubic_P = coef_matrix[4, 4]
    ))
  } else {

    nonlinear_results <- rbind(nonlinear_results, data.frame(
      Predictor = predictor,
      Linear_P = ifelse(nrow(coef_matrix) >= 2, coef_matrix[2, 4], NA),
      Quadratic_P = ifelse(nrow(coef_matrix) >= 3, coef_matrix[3, 4], NA),
      Cubic_P = NA
    ))
  }
}

significant_nonlinear <- subset(nonlinear_results, Quadratic_P < 0.05 | Cubic_P < 0.05, s
print("Significant Non-Linear Predictors:")
```

```
[1] "Significant Non-Linear Predictors:"
```

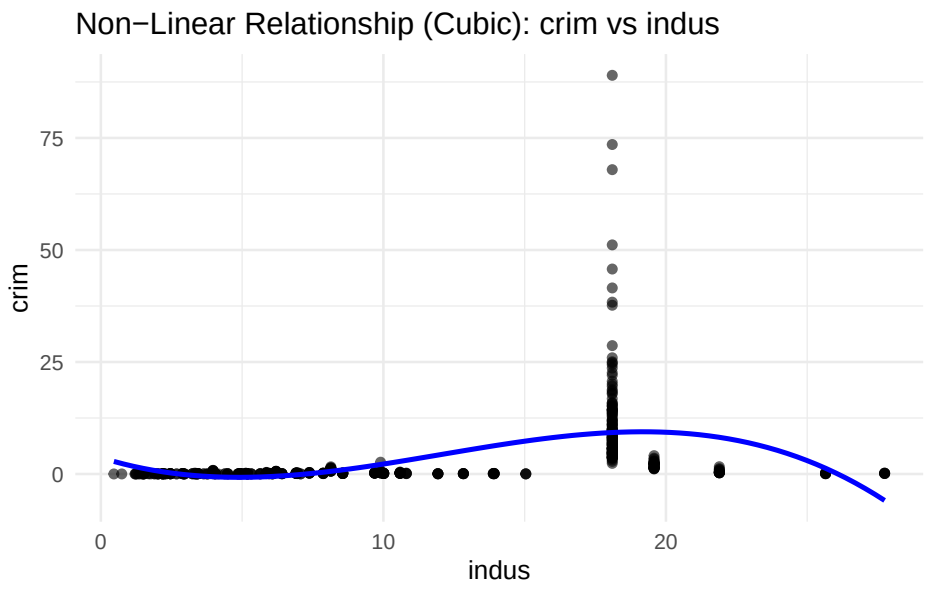
```
print(significant_nonlinear)
```

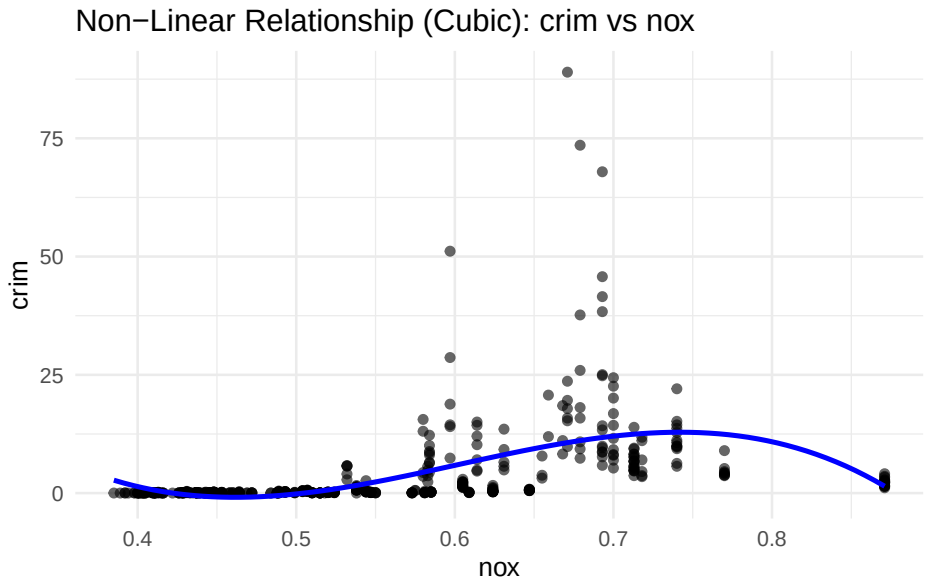
```
   Predictor  Quadratic_P      Cubic_P
2      indus 3.420187e-10 1.196405e-12
4        nox 6.811300e-15 6.961110e-16
6        age 4.737733e-02 6.679915e-03
7        dis 4.941214e-12 1.088832e-08
10   ptratio 4.119552e-03 6.300514e-03
13      medv 3.260523e-18 1.046510e-12
```

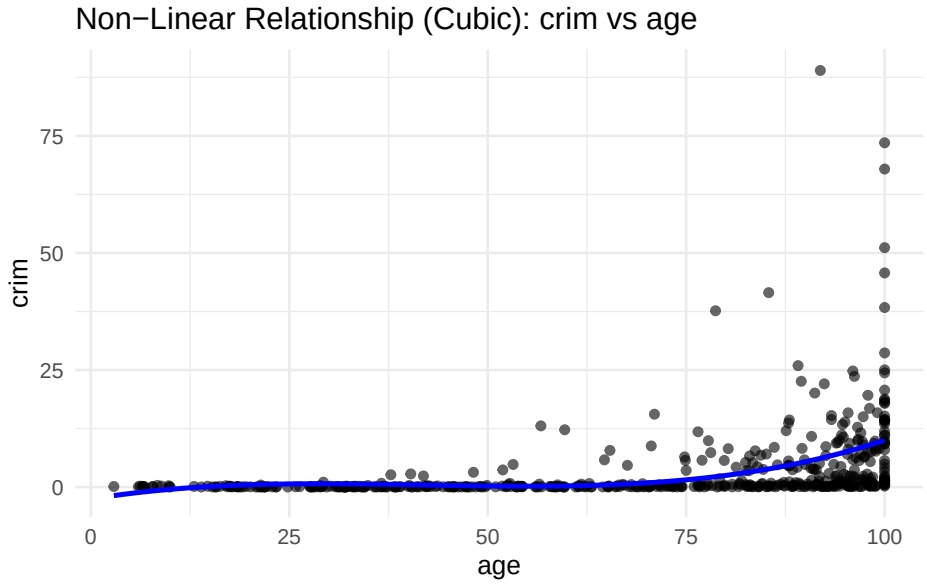
```
library(ggplot2)
for (predictor in significant_nonlinear$Predictor) {
  p <- ggplot(Boston, aes_string(x = predictor, y = "crim")) +
    geom_point(alpha = 0.6) +
    geom_smooth(method = "lm", formula = y ~ poly(x, 3), color = "blue", se = FALSE) +
    ggtitle(paste("Non-Linear Relationship (Cubic): crim vs", predictor)) +
    xlab(predictor) +
```

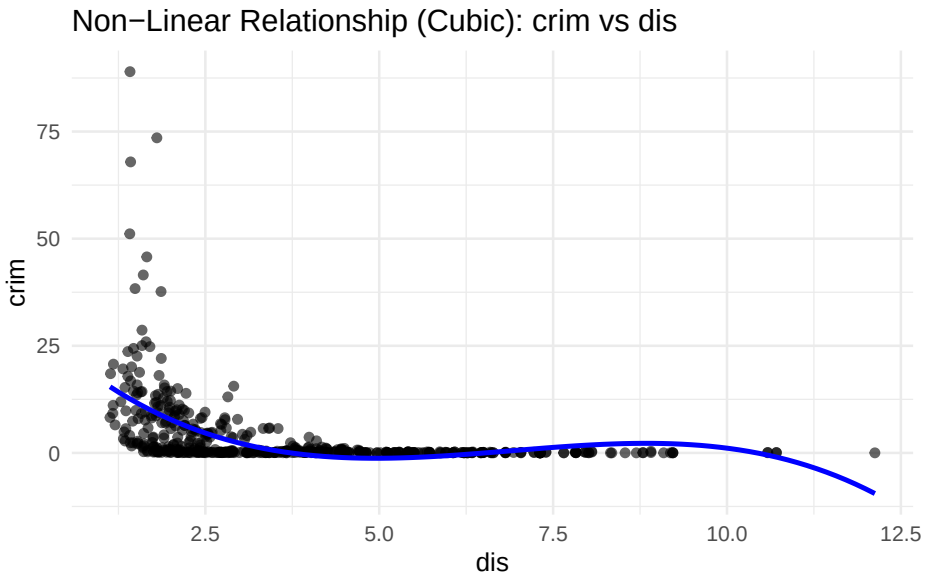
```
    ylab("crim") +
    theme_minimal()

  print(p)
}
```

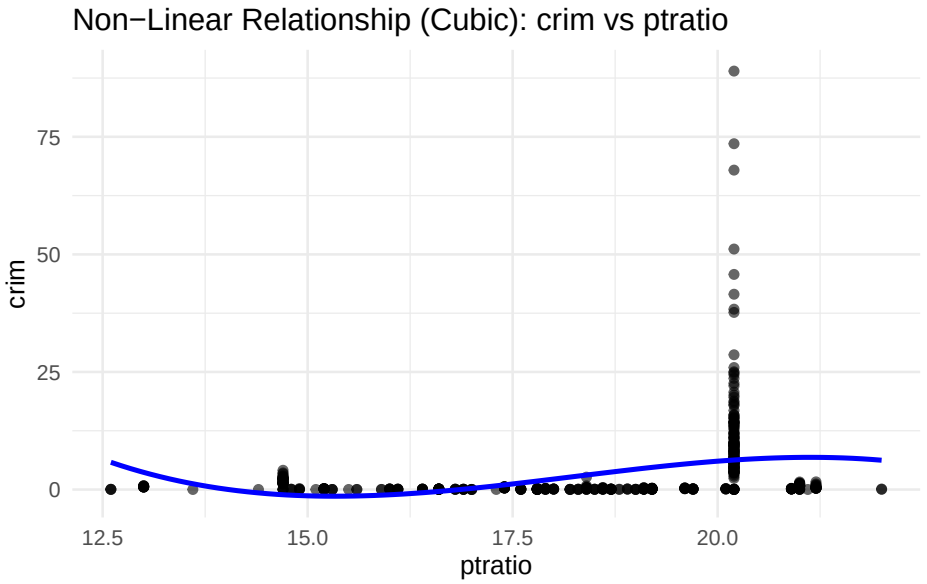
```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
i Please use tidy evaluation idioms with `aes()`.
i See also `vignette("ggplot2-in-packages")` for more information.
```

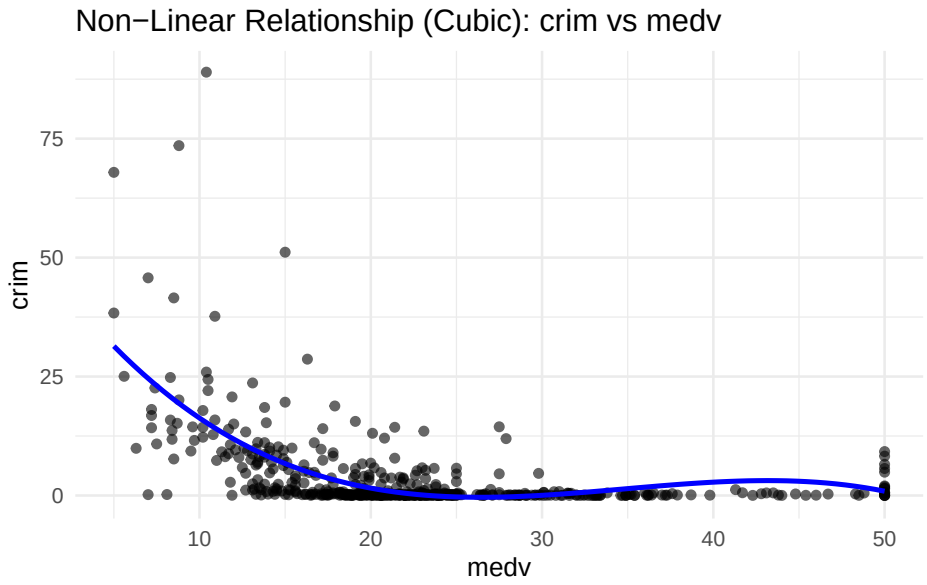



Non−Linear Relationship (Cubic): crim vs nox
75
50
25
0
crim
0.4
0.5
0.6
0.7
0.8
nox


Non−Linear Relationship (Cubic): crim vs age
75
50
25
0
crim
0
25
50
75
100
age


Non−Linear Relationship (Cubic): crim vs dis
75
50
25
0
crim
2.5
5.0
7.5
10.0
12.5
dis


Non−Linear Relationship (Cubic): crim vs ptratio
75
50
25
0
crim
12.5
15.0
17.5
20.0
ptratio



The relationship between these variables (Indus, NOx, age,dis, ptratio and medv.) and Crimea is not a simple linear relationship, but a more complex curve relationship (including U-shaped trends or other nonlinear patterns).

Bonus question (20% pts)

For multiple linear regression, show that R^2 is equal to the correlation between the response vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ and the fitted values $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^T$. That is

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = [\text{Cor}(\mathbf{y}, \hat{\mathbf{y}})]^2.$$

Bonus:

In multiple linear regression, $R^2 = 1 - \frac{RSS}{TSS}$, where $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$, and then $R^2 = \frac{ESS}{TSS}$, where $ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

since $Cor(y, \hat{y}) = \frac{cov(y, \hat{y})}{\sqrt{var(y) \cdot var(\hat{y})}}$, where $cov(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{y})$

$var(y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$, $var(\hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

so $TSS = ESS + RSS$, $ESS = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2 = n \cdot var(\hat{y})$

thus $R^2 = \frac{ESS}{TSS} = \frac{n \cdot var(\hat{y})}{n \cdot var(y)} = \frac{var(\hat{y})}{var(y)}$

$cor^2(y, \hat{y}) = \frac{cov^2(y, \hat{y})}{var(y) \cdot var(\hat{y})} = \frac{(var(\hat{y}))^2}{var(y) \cdot var(\hat{y})} = \frac{var(\hat{y})}{var(y)} = R^2$

therefore $R^2 = 1 - \frac{RSS}{TSS} = [cor(y, \hat{y})]^2$