Assignment 2 Solutions

Question 1

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \text{ and } \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Thus

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots 1 \\ x_1 & x_2 & \cdots x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

and so

$$(X^T X)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - n^2\bar{x}^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} = \frac{n}{n\sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

Next we evaluate $X^T\boldsymbol{y}$:

$$X^T\boldsymbol{y} = \begin{bmatrix} 1 & 1 & \cdots 1 \\ x_1 & x_2 & \cdots x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

Putting all of this together yields:

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{bmatrix} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{bmatrix} \bar{y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i y_i \\ -\bar{x}\bar{y} + \sum_{i=1}^n x_i y_i \end{bmatrix}$$

Therefore

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

and

$$\hat{\beta}_0 = \frac{\bar{y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\bar{y}\sum_{i=1}^n x_i^2 - n\bar{x}^2\bar{y} + n\bar{x}^2\bar{y} - \bar{x}\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\bar{y}(\sum_{i=1}^n x_i^2 - n\bar{x}^2) - \bar{x}(\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \frac{\bar{y}\sum_{i=1}^n (x_i - \bar{x})^2 - \bar{x}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \bar{y} - \hat{\beta}_1\bar{x}$$

Question 2

(a) $\mathrm{E}[y_2] = 6$

(b) $\mathrm{Var}[y_1] = 1$

(c) $\mathrm{Cov}[y_1, y_2] = 0$

(d) $\mathrm{Corr}[y_1, y_3] = \frac{\mathrm{Cov}[y_1,y_3]}{\mathrm{SD}[y_1]\mathrm{SD}[y_3]} = \frac{1}{\sqrt{1}\sqrt{3}} = 0.5773$

(e) $\mathrm{E}[y_2 - y_3] = \mathrm{E}[y_2] - \mathrm{E}[y_3] = 6 - 4 = 2$

(f) $\mathrm{Var}[y_2 - y_3] = \mathrm{Var}[y_2] + \mathrm{Var}[y_3] - 2\mathrm{Cov}[y_2, y_3] = 2 + 3 - 2(-1) = 7$

(g) $\mathrm{E}[\boldsymbol{ay}] = \boldsymbol{a}\mathrm{E}[\boldsymbol{y}] = \begin{bmatrix}1 & 2 & -1\end{bmatrix}\begin{bmatrix}2\\6\\4\end{bmatrix} = (1)(2) + (2)(6) + (-1)(4) = 10$

(h) $\mathrm{Var}[\boldsymbol{ay}] = \boldsymbol{a}\mathrm{Var}[\boldsymbol{y}]\boldsymbol{a}^T = \begin{bmatrix}1 & 2 & -1\end{bmatrix}\begin{bmatrix}1 & 0 & 1\\0 & 2 & -1\\1 & -1 & 3\end{bmatrix}\begin{bmatrix}1\\2\\-1\end{bmatrix} = \begin{bmatrix}0 & 5 & -4\end{bmatrix}\begin{bmatrix}1\\2\\-1\end{bmatrix} = (0)(1) + (5)(2) +$ $(-4)(-1) = 14$

(i) $\mathrm{E}[A\boldsymbol{y}] = A\mathrm{E}[\boldsymbol{y}] = \begin{bmatrix}1 & 0 & 2\\1 & 1 & -1\\2 & 1 & 1\end{bmatrix}\begin{bmatrix}2\\6\\4\end{bmatrix} = \begin{bmatrix}10\\4\\14\end{bmatrix}$

(j) $\mathrm{Var}[A\boldsymbol{y}] = A\mathrm{Var}[\boldsymbol{y}]A^T = \begin{bmatrix}1 & 0 & 2\\1 & 1 & -1\\2 & 1 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 1\\0 & 2 & -1\\1 & -1 & 3\end{bmatrix}\begin{bmatrix}1 & 1 & 2\\0 & 1 & 1\\2 & -1 & 1\end{bmatrix} = \begin{bmatrix}17 & -6 & 11\\-6 & 6 & 0\\11 & 0 & 11\end{bmatrix}$

Question 3

(a)

$$\begin{bmatrix}\hat{\boldsymbol{\mu}}\\ \hline \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}X\hat{\boldsymbol{\beta}}\\ \hline \boldsymbol{y} - X\hat{\boldsymbol{\beta}}\end{bmatrix} = \begin{bmatrix}X(X^TX)^{-1}X^T\boldsymbol{y}\\ \hline \boldsymbol{y} - X(X^TX)^{-1}X^T\boldsymbol{y}\end{bmatrix} = \begin{bmatrix}H\\ \hline I - H\end{bmatrix}\boldsymbol{y} = P\boldsymbol{y}$$

where $H = X(X^TX)^{-1}X^T$ is the "hat" matrix and

$$P = \begin{bmatrix}H\\ \hline I - H\end{bmatrix}$$

is the $2n \times n$ matrix formed by stacking H and $I - H$ on top of eachother.

Since $\boldsymbol{y} \sim \mathrm{MVN}(\boldsymbol{\mu}, \sigma^2 I)$, $P\boldsymbol{y}$ also follows a multivariate normal distribution with mean vector and variance-covariance matrix given by:

- $\mathrm{E}[P\boldsymbol{y}] = P\mathrm{E}[\boldsymbol{y}] = P\boldsymbol{\mu} = \begin{bmatrix}H\\ \hline I - H\end{bmatrix}\boldsymbol{\mu} = \begin{bmatrix}H\boldsymbol{\mu}\\ \hline (I - H)\boldsymbol{\mu}\end{bmatrix} = \begin{bmatrix}X(X^TX)^{-1}X^TX\boldsymbol{\beta}\\ \hline X\boldsymbol{\beta} - X(X^TX)^{-1}X^TX\boldsymbol{\beta}\end{bmatrix} = \begin{bmatrix}X\boldsymbol{\beta}\\ \hline \boldsymbol{0}_{n\times 1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\mu}\\ \hline \boldsymbol{0}_{n\times 1}\end{bmatrix}$
- $\mathrm{Var}[P\boldsymbol{y}] = P\mathrm{Var}[\boldsymbol{y}]P^T = P\sigma^2 IP^T = \sigma^2 PP^T = \sigma^2\begin{bmatrix}H\\ \hline I - H\end{bmatrix}\begin{bmatrix}H^T & \vert & (I - H)^T\end{bmatrix} = \sigma^2\left[\begin{array}{c|c} HH^T & H(I - H)^T\\ \hline (I - H)H^T & (I - H)(I - H)^T\end{array}\right]$
  But because H and $I - H$ are both symmetric ($A = A^T$) and idempotent ($AA = A$) matrices the variance-covariance matrix simplifies to:

$$\mathrm{Var}[P\boldsymbol{y}] = \sigma^2\left[\begin{array}{c|c} H & 0_{n\times n}\\ \hline 0_{n\times n} & (I - H)\end{array}\right]$$

$$\therefore \begin{bmatrix}\hat{\boldsymbol{\mu}}\\ \hline \boldsymbol{e}\end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix}\boldsymbol{\mu}\\ \hline \boldsymbol{0}_{n\times 1}\end{bmatrix}, \sigma^2\left[\begin{array}{c|c} H & 0_{n\times n}\\ \hline 0_{n\times n} & (I - H)\end{array}\right]\right)$$

(b)

Since the off-diagonal blocks of the variance-covariance matrix found in part (a) are filled entirely with zeros, this means that the random vectors $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{e}$ are uncorrelated. And because we are in the context of the multivariate normal distribution (where uncorrelatedness implies independence) we can conclude that $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{e}$.

Question 4

Since

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-p-1)},$$

by the properties of the χ^2 distribution

$$\mathrm{E}\left[\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2}\right] = (n-p-1).$$

Since $(n-p-1)$ and σ^2 are constants they can be factored outside of the expectation, yielding

$$\frac{(n-p-1)\mathrm{E}[\hat{\sigma}^2]}{\sigma^2} = (n-p-1).$$

Rearraning yields the desired result:

$$\mathrm{E}[\hat{\sigma}^2] = \sigma^2.$$

Question 5

(a)

```r
setwd("/Users/nstevens/Dropbox/Teaching/STAT_331/Assignments/Assignment 2/")
worldcup <- read.csv("worldcup.csv", header = TRUE)
summary(worldcup$Position)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.511   4.000   4.000
```

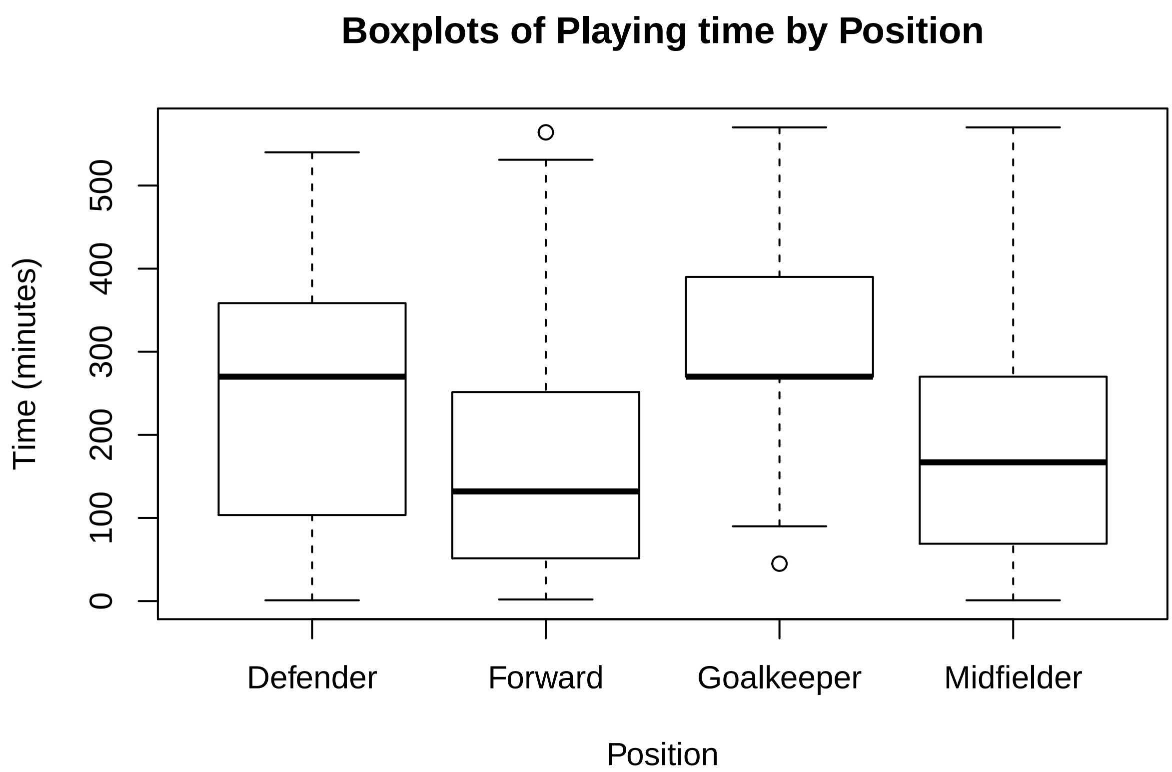
```r
worldcup$Position <- factor(worldcup$Position, levels = 1:4,
                            labels = c("Defender", "Forward", "Goalkeeper", "Midfielder"))
summary(worldcup$Position)
```

```
##   Defender    Forward Goalkeeper Midfielder
##        188        143         36        228
```

As we can from the summaries above, the `Position` variable has been successfully changed from a numeric variable to a categorical factor variable. Note that the summaries are not required for full points, I include them for illustration purposes only.

(b)

```r
boxplot(worldcup$Time ~ worldcup$Position, xlab = "Position", ylab = "Time (minutes)",
          main = "Boxplots of Playing time by Position")
```



As we can see from the plot above, relative to the other positions, goalkeepers tend to play for the longest amount of time, whereas forwards and midfielders tend to play for least amount of time.

(c) The model being fit in this question is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where $x_1 = 1$ if the player is a forward (and 0 otherwise), $x_2 = 1$ if the player is a goalkeeper (and 0 otherwise), and $x_3 = 1$ if the player is a midfielder (and 0 otherwise). The response variable y is `Time`.

```
m <- lm(Time ~ Position, data = worldcup)
summary(m)
```

```
##
## Call:
## lm(formula = Time ~ Position, data = worldcup)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -269.92 -117.13  -11.56   78.44  397.30
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          241.61      10.23  23.611  < 2e-16 ***
## PositionForward      -74.91      15.57  -4.812  1.9e-06 ***
## PositionGoalkeeper    73.30      25.53   2.872 0.004228 **
## PositionMidfielder   -50.05      13.82  -3.621 0.000319 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 140.3 on 591 degrees of freedom
## Multiple R-squared:  0.07394,    Adjusted R-squared:  0.06924
```

```
## F-statistic: 15.73 on 3 and 591 DF,  p-value: 7.464e-10
```

- $\hat{\beta}_0 = 241.61$ indicates that we expect defenders to play for 241.61 minutes.
- $\hat{\beta}_1 = -74.91$ indicates that relative to defenders, we expect forwards to play for 74.91 fewer minutes.
- $\hat{\beta}_2 = 73.30$ indicates that relative to defenders, we expect goalkeepers to play for 73.30 more minutes.
- $\hat{\beta}_2 = -50.05$ indicates that relative to defenders, we expect midfielders to play for 50.05 fewer minutes.

(d) To determine confidence intervals it will be useful to extract the variance-covariance matrix from the model.

```r
vcov(m)
```

```
##                     (Intercept) PositionForward PositionGoalkeeper
## (Intercept)            104.7147       -104.7147          -104.7147
## PositionForward       -104.7147        242.3816           104.7147
## PositionGoalkeeper    -104.7147        104.7147           651.5582
## PositionMidfielder    -104.7147        104.7147           104.7147
##                     PositionMidfielder
## (Intercept)                  -104.7147
## PositionForward               104.7147
## PositionGoalkeeper            104.7147
## PositionMidfielder            191.0584
```

Defender: $\mathrm{E}[y|x_1 = 0, x_2 = 0, x_3 = 0] = \beta_0$

- Estimate: $\hat{\beta}_0 = 241.61$
- Standard Error: $\mathrm{SE}[\hat{\beta}_0] = \sqrt{\mathrm{Var}[\hat{\beta}_0]} = \sqrt{104.7147}$
- CI: $\hat{\beta}_0 \pm t_{(n-p-1)}(1 - \frac{\alpha}{2}) \times \mathrm{SE}[\hat{\beta}_0] = 241.61 \pm t_{(591)}(0.975) \times \sqrt{104.7147} = 241.61 \pm 1.964 \times 10.233 = (221.5125, 261.7075)$

Forward: $\mathrm{E}[y|x_1 = 1, x_2 = 0, x_3 = 0] = \beta_0 + \beta_1$

- Estimate: $\hat{\beta}_0 + \hat{\beta}_1 = 241.61 - 74.91 = 166.70$
- Standard Error: $\mathrm{SE}[\hat{\beta}_0 + \hat{\beta}_1] = \sqrt{\mathrm{Var}[\hat{\beta}_0] + \mathrm{Var}[\hat{\beta}_1] + 2\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1]} = \sqrt{104.7147 + 242.3816 + 2(-104.7147)} = \sqrt{137.6669}$
- CI: $\hat{\beta}_0 + \hat{\beta}_1 \pm t_{(n-p-1)}(1 - \frac{\alpha}{2}) \times \mathrm{SE}[\hat{\beta}_0 + \hat{\beta}_1] = 166.70 \pm t_{(591)}(0.975) \times \sqrt{137.6669} = 166.70 \pm 1.964 \times 11.7332 = (143.6560, 189.744)$

Goalkeeper: $\mathrm{E}[y|x_1 = 0, x_2 = 1, x_3 = 0] = \beta_0 + \beta_2$

- Estimate: $\hat{\beta}_0 + \hat{\beta}_2 = 241.61 + 73.30 = 314.91$
- Standard Error: $\mathrm{SE}[\hat{\beta}_0 + \hat{\beta}_2] = \sqrt{\mathrm{Var}[\hat{\beta}_0] + \mathrm{Var}[\hat{\beta}_2] + 2\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_2]} = \sqrt{104.7147 + 651.5582 + 2(-104.7147)} = \sqrt{546.8435}$
- CI: $\hat{\beta}_0 + \hat{\beta}_2 \pm t_{(n-p-1)}(1 - \frac{\alpha}{2}) \times \mathrm{SE}[\hat{\beta}_0 + \hat{\beta}_2] = 314.91 \pm t_{(591)}(0.975) \times \sqrt{546.8435} = 314.91 \pm 1.964 \times 23.3847 = (268.9824, 360.8376)$

Midfielder: $\mathrm{E}[y|x_1 = 0, x_2 = 0, x_3 = 1] = \beta_0 + \beta_3$

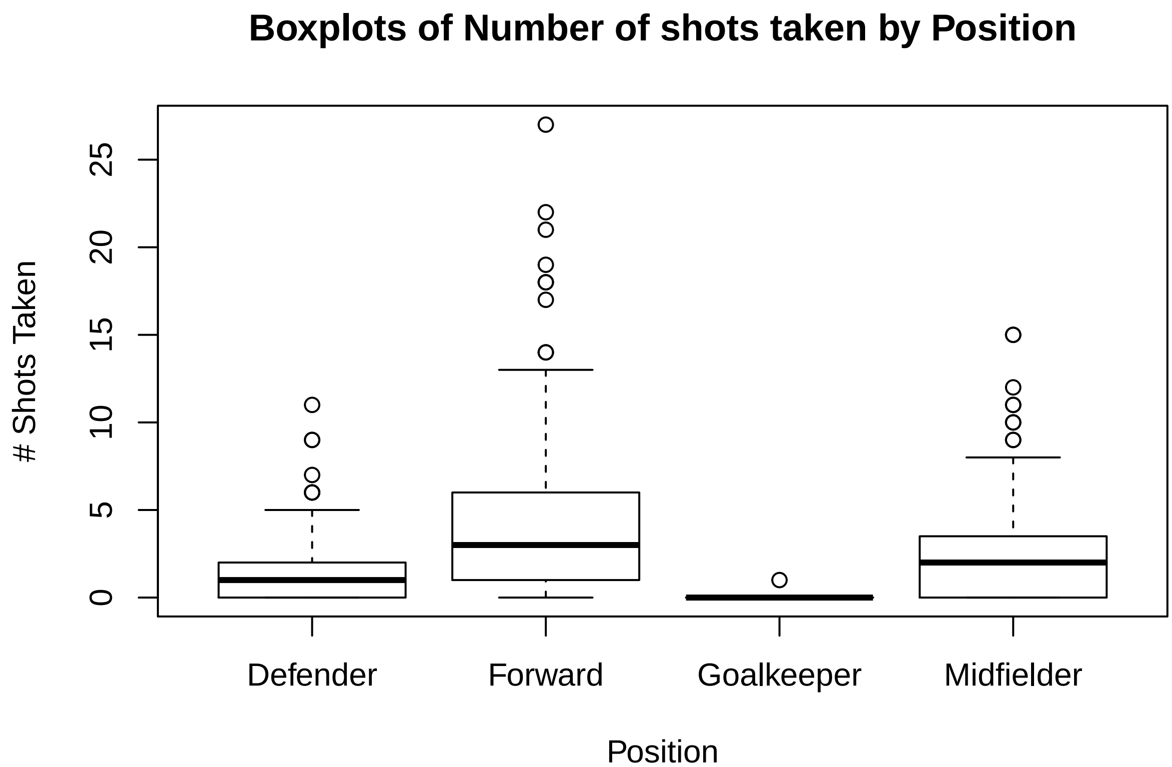
- Estimate: $\hat{\beta}_0 + \hat{\beta}_3 = 241.61 - 50.05 = 191.56$
- Standard Error: $\mathrm{SE}[\hat{\beta}_0 + \hat{\beta}_3] = \sqrt{\mathrm{Var}[\hat{\beta}_0] + \mathrm{Var}[\hat{\beta}_3] + 2\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_3]} = \sqrt{104.7147 + 191.0584 + 2(-104.7147)} = \sqrt{86.3437}$

- CI: $\hat{\beta}_0+\hat{\beta}_3 \pm t_{(n-p-1)}(1-\frac{\alpha}{2}) \times \text{SE}[\hat{\beta}_0+\hat{\beta}_3] = 191.56 \pm t_{(591)}(0.975) \times \sqrt{86.3437} = 191.56 \pm 1.964 \times 9.2921 = (172.3103, 209.8097)$

Note that the intermediate step of separately finding the standard errors is not required.

(e)

```
boxplot(worldcup$Shots ~ worldcup$Position, xlab = "Position", ylab = "# Shots Taken",
          main = "Boxplots of Number of shots taken by Position")
```



Unsurprisingly, the plot above indicates that different positions differ in the number of shots they take; forwards take more shots than midfielders who take more shots than defenders who take more shots than goalkeepers.

(f) The model being fit in this question is the same as in part (b) except that the response variable y is now `Shots`.

```
m <- lm(Shots ~ Position, data = worldcup)
summary(m)
```

```
##
## Call:
## lm(formula = Shots ~ Position, data = worldcup)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2308 -1.3947 -0.3947  0.7692 22.7692
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.1649     0.2264   5.146 3.62e-07 ***
## PositionForward      3.0659     0.3444   8.903  < 2e-16 ***
## PositionGoalkeeper  -1.1371     0.5646  -2.014   0.0445 *
```

```
## PositionMidfielder   1.2298     0.3057   4.022 6.51e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.104 on 591 degrees of freedom
## Multiple R-squared:  0.1447, Adjusted R-squared:  0.1404
## F-statistic: 33.33 on 3 and 591 DF,  p-value: < 2.2e-16
```

- $\hat{\beta}_0 = 1.1649$ indicates that we expect defenders to take 1.1649 shots throughout the World Cup.
- $\hat{\beta}_1 = 3.0659$ indicates that relative to defenders, we expect forwards to take 3.0659 more shots throughout the World Cup.
- $\hat{\beta}_2 = -1.1371$ indicates that relative to defenders, we expect goalkeepers to take 1.1371 fewer shots throughout the World Cup.
- $\hat{\beta}_2 = 1.2298$ indicates that relative to defenders, we expect midfielders to take 1.2298 more shots throughout the World Cup.

(g) To determine confidence intervals it will be useful to extract the variance-covariance matrix from the model.

```r
vcov(m)
```

```
##                      (Intercept) PositionForward PositionGoalkeeper
## (Intercept)            0.0512359      -0.0512359         -0.0512359
## PositionForward       -0.0512359       0.1185950          0.0512359
## PositionGoalkeeper    -0.0512359       0.0512359          0.3188012
## PositionMidfielder    -0.0512359       0.0512359          0.0512359
##                      PositionMidfielder
## (Intercept)                 -0.05123590
## PositionForward              0.05123590
## PositionGoalkeeper           0.05123590
## PositionMidfielder           0.09348305
```

Defender: $\mathrm{E}[y|x_1 = 0, x_2 = 0, x_3 = 0] = \beta_0$

- Estimate: $\hat{\beta}_0 = 1.1649$
- Standard Error: $\mathrm{SE}[\hat{\beta}_0] = \sqrt{\mathrm{Var}[\hat{\beta}_0]} = \sqrt{0.0512}$
- CI: $\hat{\beta}_0 \pm t_{(n-p-1)}(1 - \frac{\alpha}{2}) \times \mathrm{SE}[\hat{\beta}_0] = 1.1649 \pm t_{(591)}(0.975) \times \sqrt{0.0512} = 1.1649 \pm 1.964 \times 0.2263 = (0.7204, 1.6094)$

Forward: $\mathrm{E}[y|x_1 = 1, x_2 = 0, x_3 = 0] = \beta_0 + \beta_1$

- Estimate: $\hat{\beta}_0 + \hat{\beta}_1 = 1.1649 + 3.0659 = 4.2308$
- Standard Error: $\mathrm{SE}[\hat{\beta}_0 + \hat{\beta}_1] = \sqrt{\mathrm{Var}[\hat{\beta}_0] + \mathrm{Var}[\hat{\beta}_1] + 2\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1]} = \sqrt{0.0512 + 0.1186 + 2(-0.0512)} = \sqrt{0.0674}$
- CI: $\hat{\beta}_0 + \hat{\beta}_1 \pm t_{(n-p-1)}(1 - \frac{\alpha}{2}) \times \mathrm{SE}[\hat{\beta}_0 + \hat{\beta}_1] = 4.2308 \pm t_{(591)}(0.975) \times \sqrt{0.0674} = 4.2308 \pm 1.964 \times 0.2596 = (3.7209, 4.7407)$

Goalkeeper: $\mathrm{E}[y|x_1 = 0, x_2 = 1, x_3 = 0] = \beta_0 + \beta_2$

- Estimate: $\hat{\beta}_0 + \hat{\beta}_2 = 1.1649 - 1.1371 = 0.0278$
- Standard Error: $\mathrm{SE}[\hat{\beta}_0 + \hat{\beta}_2] = \sqrt{\mathrm{Var}[\hat{\beta}_0] + \mathrm{Var}[\hat{\beta}_2] + 2\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_2]} = \sqrt{0.0512 + 0.3188 + 2(-0.0512)} = \sqrt{0.2676}$

- CI: $\hat{\beta}_0+\hat{\beta}_2 \pm t_{(n-p-1)}(1-\frac{\alpha}{2}) \times \mathrm{SE}[\hat{\beta}_0+\hat{\beta}_2] = 0.0278 \pm t_{(591)}(0.975) \times \sqrt{0.2676} = 0.0278 \pm 1.964 \times 0.5173 = (-0.9882, 1.0438)$

Midfielder: $\mathrm{E}[y|x_1=0, x_2=0, x_3=1] = \beta_0 + \beta_3$

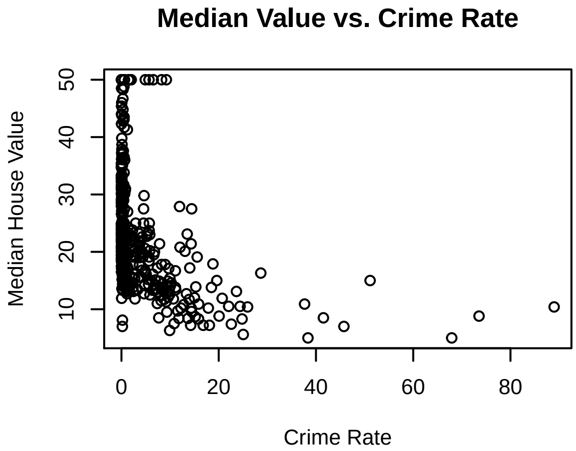
- Estimate: $\hat{\beta}_0 + \hat{\beta}_3 = 1.1649 + 1.2298 = 2.3947$
- Standard Error: $\mathrm{SE}[\hat{\beta}_0+\hat{\beta}_3] = \sqrt{\mathrm{Var}[\hat{\beta}_0] + \mathrm{Var}[\hat{\beta}_3] + 2\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_3]} = \sqrt{0.0512 + 0.0935 + 2(-0.0512)} = \sqrt{0.0423}$
- CI: $\hat{\beta}_0+\hat{\beta}_3 \pm t_{(n-p-1)}(1-\frac{\alpha}{2}) \times \mathrm{SE}[\hat{\beta}_0+\hat{\beta}_3] = 2.3947 \pm t_{(591)}(0.975) \times \sqrt{0.0423} = 2.3947 \pm 1.964 \times 0.2057 = (1.9907, 2.7987)$

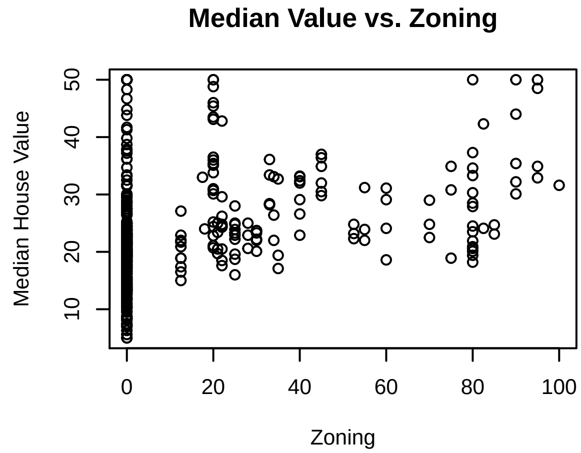
Note that the intermediate step of separately finding the standard errors is not required.

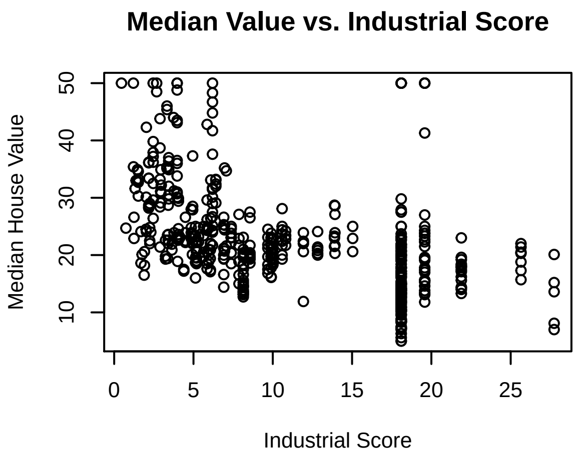
Question 6

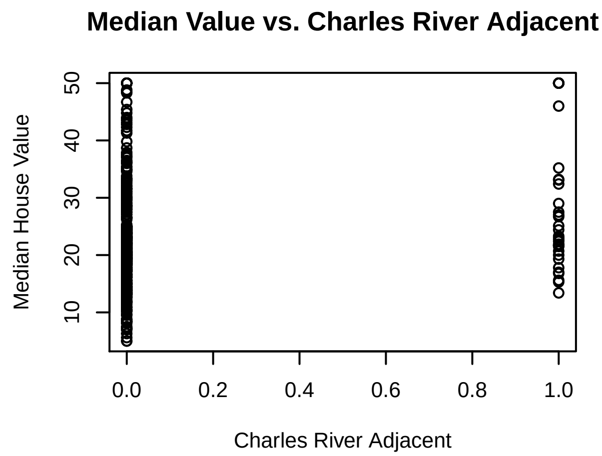
(a)

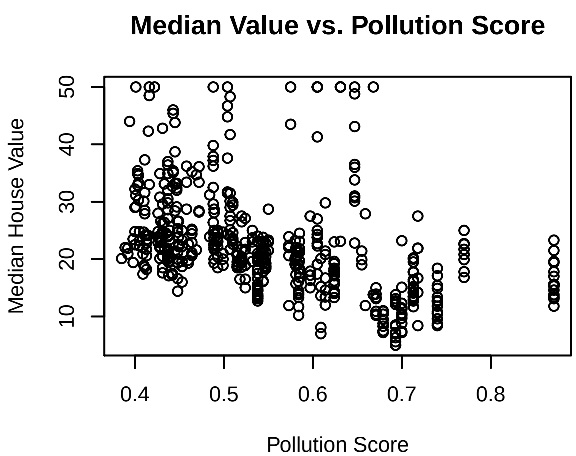
```r
boston <- read.csv("boston.csv", header = TRUE)
var.names <- c("Crime Rate", "Zoning", "Industrial Score", "Charles River Adjacent",
                 "Pollution Score", "Number of Rooms", "Neighbourhood Age", "Distance to Employment",
                 "Highway Access", "Property Taxes", "Student-Teacher Ratio", "% Low SES")
par(mfrow=c(3,2))
for(i in 1:12){
  plot(x = boston[,i], y = boston$medv, xlab = var.names[i], ylab = "Median House Value",
       main = paste("Median Value vs. ", var.names[i], sep = ""))
}
```

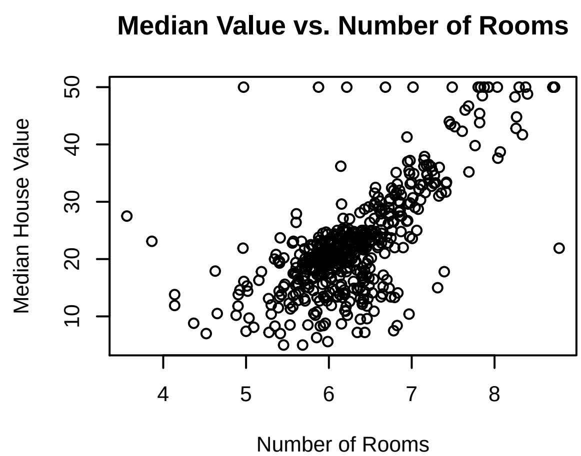

Median Value vs. Crime Rate
Median House Value
10 20 30 40 50
0 20 40 60 80
Crime Rate

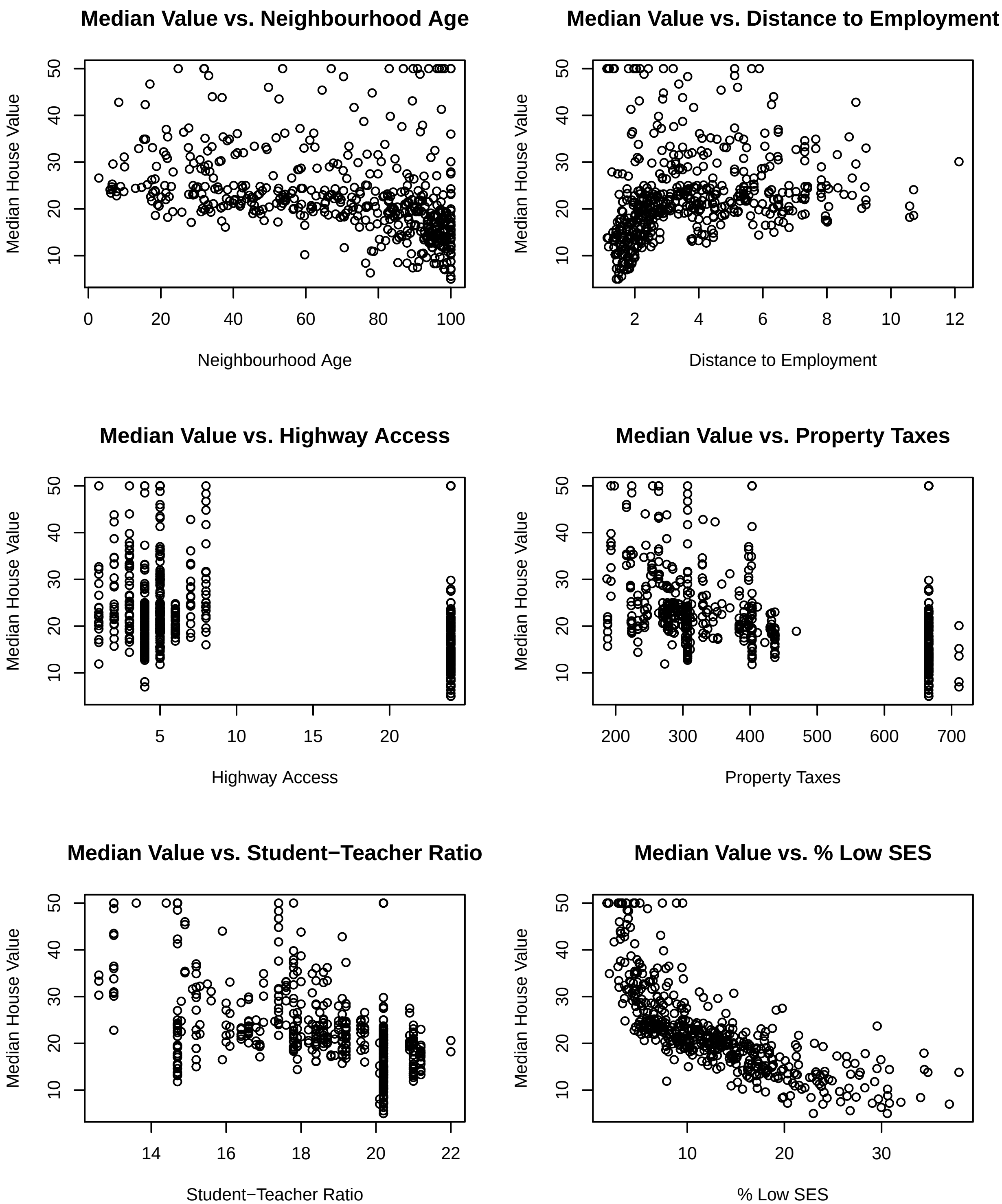

Median Value vs. Zoning
Median House Value
10 20 30 40 50
0 20 40 60 80 100
Zoning


Median Value vs. Industrial Score
Median House Value
10 20 30 40 50
0 5 10 15 20 25
Industrial Score


Median Value vs. Charles River Adjacent
Median House Value
10 20 30 40 50
0.0 0.2 0.4 0.6 0.8 1.0
Charles River Adjacent


Median Value vs. Pollution Score
Median House Value
10 20 30 40 50
0.4 0.5 0.6 0.7 0.8
Pollution Score


Median Value vs. Number of Rooms
Median House Value
10 20 30 40 50
4 5 6 7 8
Number of Rooms



(b)

```
cor(boston)[13,1:12]
```

```
##       crim         zn      indus       chas        nox         rm
## -0.3883046  0.3604453 -0.4837252  0.1752602 -0.4273208  0.6953599
##        age        dis        rad        tax    ptratio      lstat
## -0.3769546  0.2499287 -0.3816262 -0.4685359 -0.5077867 -0.7376627
```

Above are all 12 correlations. Below are the two strongest. We see that they corresopond to `rm` and `lstat`. This implies that median house values are most strongly correlation with the number of rooms in the house and the proportion of the neighbourhood with low socioeconomic status.

```
sort(abs(cor(boston)[13,1:12]))[11:12]
```

```
##        rm     lstat
## 0.6953599 0.7376627
```

(c)

```
m.full <- lm(medv ~ ., data = boston)
summary(m.full)
```

```
##
## Call:
## lm(formula = medv ~ ., data = boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1304  -2.7673  -0.5814   1.9414  26.2526
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.617270   4.936039   8.431 3.79e-16 ***
## crim         -0.121389   0.033000  -3.678 0.000261 ***
## zn            0.046963   0.013879   3.384 0.000772 ***
## indus         0.013468   0.062145   0.217 0.828520
## chas          2.839993   0.870007   3.264 0.001173 **
## nox         -18.758022   3.851355  -4.870 1.50e-06 ***
## rm            3.658119   0.420246   8.705  < 2e-16 ***
## age           0.003611   0.013329   0.271 0.786595
## dis          -1.490754   0.201623  -7.394 6.17e-13 ***
## rad           0.289405   0.066908   4.325 1.84e-05 ***
## tax          -0.012682   0.003801  -3.337 0.000912 ***
## ptratio      -0.937533   0.132206  -7.091 4.63e-12 ***
## lstat        -0.552019   0.050659 -10.897  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.798 on 493 degrees of freedom
## Multiple R-squared:  0.7343, Adjusted R-squared:  0.7278
## F-statistic: 113.5 on 12 and 493 DF,  p-value: < 2.2e-16
```

The test statistic associated wth $\mathrm{H}_0 : \beta_{indus} = 0$ vs. $\mathrm{H}_A : \beta_{indus} \neq 0$ is given by

$$t = \frac{\hat{\beta}_{indus}}{\mathrm{SE}[\hat{\beta}_{indus}]} = \frac{0.013468}{0.062145} = 0.217.$$

The corresponding p-value is:

$$p - \text{value} = 2\mathrm{P}(T \geq |t|) = 2\mathrm{P}(T \geq 0.217) = 0.828520$$

where the null distribution is $T \sim t_{(493)}$. Since this p-value is larger than $\alpha = 0.05$ we do not reject the null hypothesis and we conclude that median house values do not depend significantly on the proportion of non-retail business acreage.

(d)

```r
m.red1 <- update(object = m.full, .~. - indus, data = boston)
summary(m.red1)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + chas + nox + rm + age + dis +
##     rad + tax + ptratio + lstat, data = boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1267  -2.7487  -0.5902   1.9056  26.2609
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.543721   4.919606   8.445 3.42e-16 ***
## crim         -0.121628   0.032950  -3.691 0.000248 ***
## zn            0.046642   0.013786   3.383 0.000773 ***
## chas          2.859128   0.864680   3.307 0.001013 **
## nox         -18.534872   3.707573  -4.999 8.01e-07 ***
## rm            3.650015   0.418175   8.728  < 2e-16 ***
## age           0.003608   0.013317   0.271 0.786563
## dis          -1.499953   0.196913  -7.617 1.33e-13 ***
## rad           0.285390   0.064231   4.443 1.09e-05 ***
## tax          -0.012320   0.003411  -3.611 0.000336 ***
## ptratio      -0.933839   0.130976  -7.130 3.59e-12 ***
## lstat        -0.551115   0.050438 -10.927  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.793 on 494 degrees of freedom
## Multiple R-squared:  0.7343, Adjusted R-squared:  0.7284
## F-statistic: 124.1 on 11 and 494 DF,  p-value: < 2.2e-16
```

The test statistic associated wth $\mathrm{H}_0 : \beta_{age} = 0$ vs. $\mathrm{H}_A : \beta_{age} \neq 0$ is given by

$$t = \frac{\hat{\beta}_{age}}{\mathrm{SE}[\hat{\beta}_{age}]} = \frac{0.003608}{0.418175} = 0.271.$$

The corresponding p-value is:

$$p - \text{value} = 2\mathrm{P}(T \geq |t|) = 2\mathrm{P}(T \geq 0.271) = 0.786563$$

where the null distribution is $T \sim t_{(494)}$. Since this p-value is larger than $\alpha = 0.05$ we do not reject the null hypothesis and we conclude that median house values do not depend significantly on the proportion of owner-occupied houses built prior to 1940.

(e)

```r
m.red2 <- update(object = m.red1, .~. - age, data = boston)
s <-summary(m.red2)
s
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
##     tax + ptratio + lstat, data = boston)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1814  -2.7625  -0.6243   1.8448  26.3920
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.451747   4.903283   8.454 3.18e-16 ***
## crim         -0.121665   0.032919  -3.696 0.000244 ***
## zn            0.046191   0.013673   3.378 0.000787 ***
## chas          2.871873   0.862591   3.329 0.000935 ***
## nox         -18.262427   3.565247  -5.122 4.33e-07 ***
## rm            3.672957   0.409127   8.978  < 2e-16 ***
## dis          -1.515951   0.187675  -8.078 5.08e-15 ***
## rad           0.283932   0.063945   4.440 1.11e-05 ***
## tax          -0.012292   0.003407  -3.608 0.000340 ***
## ptratio      -0.930961   0.130423  -7.138 3.39e-12 ***
## lstat        -0.546509   0.047442 -11.519  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.789 on 495 degrees of freedom
## Multiple R-squared:  0.7342, Adjusted R-squared:  0.7289
## F-statistic: 136.8 on 10 and 495 DF,  p-value: < 2.2e-16
```

- $\hat{\beta}_0 = 41.451747$ suggests that the expected median house price in a neighborhood in which all of the explanatory variables is equal to 0 is \$41,451.75
- $\hat{\beta}_{crim}$ = -0.121665 indicates that for a unit increase in crime rate, we expect the median house price to decrease by \$121.67
- $\hat{\beta}_{zn}$ = 0.046191 indicates that for a unit increase in the proportion of residential land zoned for lots over 25,000 sq. ft., we expect the median house price to increase by \$46.19
- $\hat{\beta}_{chas}$ = 2.871873 indicates that we expect the median price of houses on the Charles river to be \$2,871.87 more than house not on the Charles River.
- $\hat{\beta}_{nox}$ = -18.262427 indicates that for a unit increase in nitrogen oxide levels, we expect the median house price to decrease by \$18,262.43
- $\hat{\beta}_{rm}$ = 3.672957 indicates that for a every additional room in the house, we expect the median house price to increase by \$3,672.96
- $\hat{\beta}_{dis}$ = -1.515951 indicates that for a unit increase in the weighted distance measure (from Boston employment centers), we expect the median house price to decrease by \$1,515.95
- $\hat{\beta}_{rad}$ = 0.283932 indicates that for a unit increase in the index of accessibility to Boston's radial highways, we expect the median house price to increase by \$283.93
- $\hat{\beta}_{tax}$ = -0.012292 indicates that for unit increase in property taxes per \$10,000, we expect the median house price to decrease by \$12.29
- $\hat{\beta}_{ptratio}$ = -0.930961 indicates that for a unit increase in the pupil-to-teacher ratio, we expect the median house price to decrease by \$930.96
- $\hat{\beta}_{lstat}$ = -0.546509 indicates that for a unit increase in the percent of the population that is classified as "low socioeconomic status", we expect the median house price to decrease by \$546.51

(f)

```r
sort(s$coefficients[2:11,4], decreasing = TRUE)
```

```
##         chas           zn          tax         crim          rad
## 9.353905e-04 7.867310e-04 3.396973e-04 2.435786e-04 1.108892e-05
##          nox      ptratio          dis           rm        lstat
## 4.330925e-07 3.391933e-12 5.079900e-15 5.776533e-18 2.292538e-27
```

The output above consists of the p-values from the mode in part (e) ordered from largest to smallest, which corresponds to an ordering of the explanatory variables from least associated to most associated. The ordering is as follows: `chas-zn-tax-crim-rad-nox-ptratio-dis-rm-lstat`.

(g)

```
predict(m.red2, newdata = data.frame(crim=5, zn=25, chas=0, nox=0.6, rm=3, dis=4, rad=10, tax=500,
                                     ptratio=20, lstat=5), interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 11.33761 1.418652 21.25657
```

Thus we predict the median house value in such a neighborhood to be \$11,337.61, with a lower 95% prediction limit of \$1,418.65 and an upper 95% prediction limit of \$21,256.57.