Giang Vu - 4445745
STAT 2131
Sep 3, 2020

parts with ⊛ sign will have R scripts & results shown in page 3-5

Homework 2

1) a) From running a linear regression in R, we can obtain the estimated function as follows ⊛

$$\hat{Y} = 168.6 + 2.03438\,X$$

From the plot made in R, we can see that a linear regression seems to be a good fit.

b) When $X = 40$ hours, using the estimated function in a

$\Rightarrow \hat{Y}_h = 168.6 + 2.03438(40) = 249.975$

c) From the estimated function in a, when X increases by 1 hour, mean hardness can be estimated to increase by an amount $\hat{\beta}_1 = 2.03438$

2) From R, we calculate $MSE = \hat{\sigma}^2 = \frac{SSR}{n-2} = 10.4589286$

a) With $X_h = 30 \Rightarrow \hat{Y}_h = 168.6 + 2.03438(30) = 229.6314$

$s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] = 0.6864$ ⊛

$\Rightarrow s\{\hat{Y}_h\} \simeq 0.8285$. Also, we have $t(1-\frac{\alpha}{2}; n-2) = t(0.99, 14) = 2.624$

$\Rightarrow$ 98% CI for $E(Y_h)$ is $\hat{Y}_h \pm t(1-\frac{\alpha}{2}, n-2)\, s\{\hat{Y}_h\}$

which is $229.6314 \pm 2.624(0.8285)$

or $227.457 \leqslant E(Y_h) \leqslant 231.805$

b) With a new $X_h = 30$, $s^2\{pred\} = MSE + s^2\{\hat{Y}_h\} = 10.4589 + 0.6864 = 11.1453$

$\Rightarrow s\{pred\} = 3.338$

$\Rightarrow$ 98% prediction interval for $Y_{h(new)}$ is $\hat{Y}_h \pm t(1-\frac{\alpha}{2}, n-2)\, s\{pred\}$

which is $229.6314 \pm 2.624(3.338)$ or $220.872 \leqslant Y_{h(new)} \leqslant 238.390$

c) With 10 new $X_h = 30$, $s^2\{predmean\} = \frac{MSE}{m} + s^2(\hat{Y}_h) = \frac{10.4589}{10} + 0.6864 = 1.73229$

$\Rightarrow s\{predmean\} = 1.316$

$\Rightarrow$ 98% prediction interval for $\bar{Y}_{h(new)}$ is $\hat{Y}_h \pm t(1-\frac{\alpha}{2}, n-2)\, s\{predmean\}$

which is $229.6314 \pm 2.624(1.316)$ or $226.178 \leqslant \bar{Y}_{h(new)} \leqslant 233.084$

d) The prediction interval in c) is narrower than in b), as it should be, because it involves a prediction of the mean hardness of 10 newly molded test items, not just a single newly molded test item like in b)

e) $W^2 = 2F(1-\alpha, 2, n-2) = 2F(0.98, 2, 14)$
$= 2(5.241) = 10.482$

$\Rightarrow W = 3.238$

$\Rightarrow$ 98% confidence band for the regression line when $X_h = 30$ is $\hat{Y}_h \pm W s\{\hat{Y}_h\}$ which is $229.6314 \pm 3.238(0.8285)$

or $226.948 \leq \beta_0 + \beta_1 X_h \leq 232.314$

This is wider at this point ($X_h = 30$) than the CI in part a), as it should be because it is supposed to ~~contain~~ describe all mean responses for the entire regression line, and not just the mean response at a given X level like in a)

1 & 2 (continued)

R scripts & results for previous questions

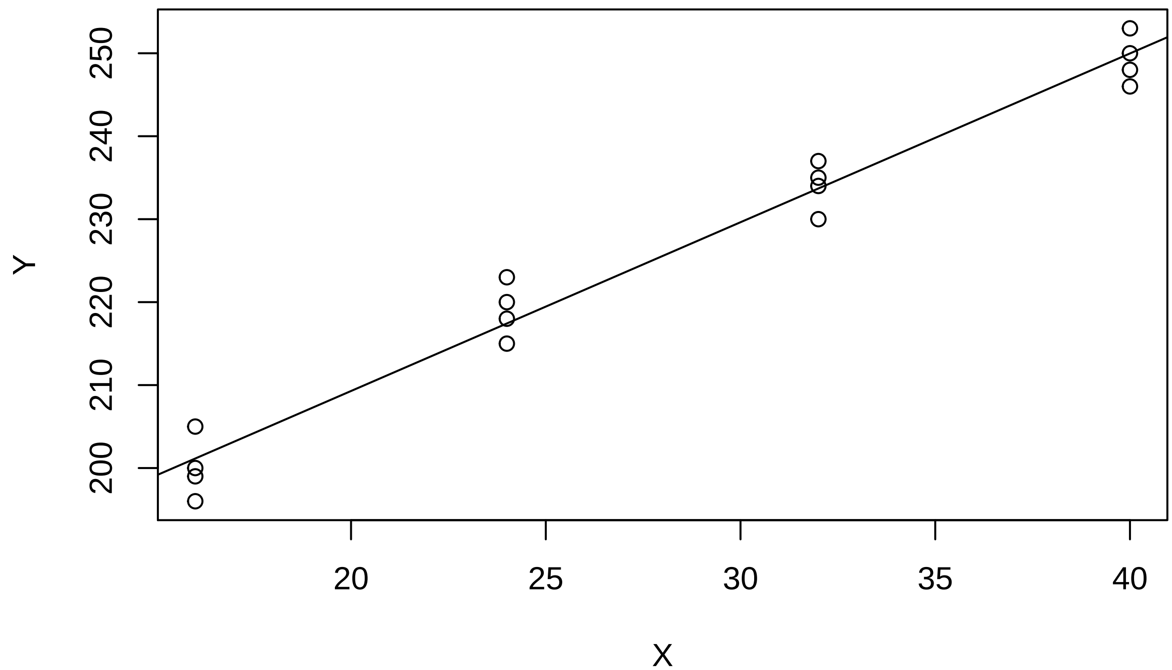
1) a)

```r
#read data
hw2_data <- read.csv("/Users/giangvu/Desktop/STAT 2131 - Applied Stat Methods 1/HW/hw2/hw2_data.csv",
                     header = T,sep = ",")

#fit regression
hw2_model <- lm(Y ~ X, data = hw2_data)
sm <- summary(hw2_model)
sm
```

```
##
## Call:
## lm(formula = Y ~ X, data = hw2_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1500 -2.2188  0.1625  2.6875  5.5750
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 168.60000    2.65702   63.45  < 2e-16 ***
## X             2.03438    0.09039   22.51 2.16e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.234 on 14 degrees of freedom
## Multiple R-squared:  0.9731, Adjusted R-squared:  0.9712
## F-statistic: 506.5 on 1 and 14 DF,  p-value: 2.159e-12
```

```r
#plot
plot(Y ~ X, data = hw2_data)
abline(hw2_model)
```



2) calculating MSE and s^2{Yh}

```
#MSE
mse <- function(sm){
  sum(sm$residuals^2)/(nrow(hw2_data)-2)
}
mse_hw2 <- mse(sm)
mse_hw2
```

```
## [1] 10.45893
```

```
#s^2{Yh}
xbar <- mean(hw2_data$X) #mean of Xi's - X bar
xbar
```

```
## [1] 28
```

```
hw2_data$sq_diff_fr_mean <- (hw2_data$X - xbar)**2
xsum_sqr <- sum(hw2_data$sq_diff_fr_mean) #sum of (Xi - Xbar)^2 for i = 1,...,n
xsum_sqr
```

```
## [1] 1280
```

```r
#with a new point Xh = 30, we calculate the s^2{Y} as follows
s2Yh <- mse_hw2*(1/nrow(hw2_data) + (30 - xbar)**2/(xsum_sqr))
s2Yh
```

```
## [1] 0.6863672
```

3) Because $\tilde{b}_0, \tilde{b}_1$ seek to minimize the $PSS_\lambda(\beta_0, \beta_1)$ function, we can take the derivative of the $PSS_\lambda(\beta_0, \beta_1)$ function and set that equal to 0 to solve for our $\tilde{b}_0$ and $\tilde{b}_1$

We have $$\frac{\partial PSS}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial PSS}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) + 2\lambda\beta_1 = 0$$

$$\Rightarrow \tilde{b}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\tilde{b}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2 + \lambda}.$$

a) With $\lambda = 0 \Rightarrow \tilde{b}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \hat{\beta}_0^{(OLS)}$

$$\tilde{b}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2 + 0} = \hat{\beta}_1^{(OLS)}$$

They are equal to the OLS estimators $(\hat{\beta}_0^{(OLS)}, \hat{\beta}_1^{(OLS)})$ when $\lambda = 0$

As $\lambda = \infty \Rightarrow \tilde{b}_0 = \hat{\beta}_0^{(OLS)}$

$\tilde{b}_1 = 0$ (there's ∞ in the denominator)

So $\tilde{b}_0 = \hat{\beta}_0^{(OLS)}$ but $\tilde{b}_1 = 0$ when $\lambda = \infty$

b) $$Var(\tilde{b}_1) = \frac{\sum(x_i - \bar{x})^2}{(\sum(x_i - \bar{x})^2 + \lambda)^2}\sigma^2 < \frac{\sigma^2}{\sum(x_i - \bar{x})^2} = Var(\hat{\beta}_1^{OLS})$$

$$\Rightarrow Var(\tilde{b}_1) < Var(\hat{\beta}_1^{(OLS)})$$

c) $$E(\tilde{b}_1) = E\left(\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2 + \lambda}\right) \neq E\left(\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right)$$ when $\lambda > 0$

$\Rightarrow E(\tilde{b}_1) \neq E(\hat{\beta}_1^{OLS}) = \beta_1 \Rightarrow \tilde{b}_1$ is biased when $\lambda > 0$

4) a) The OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the following error term

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

=) The first partial derivatives with respect to β_0 and β_1, evaluated at $\hat{\beta}_0$ and $\hat{\beta}_1$ must equal 0

For β_0 =) $$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \Big|_{(\hat{\beta}_0, \hat{\beta}_1)} = 0$$

$$\Rightarrow \quad -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

by definition

$$\Rightarrow \quad -2 \sum_{i=1}^{n} \hat{\epsilon}_i = 0 \quad (\text{as } \hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$\Rightarrow \quad \sum_{i=1}^{n} \hat{\epsilon}_i = 0$$

For β_1 =) $$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \Big|_{(\hat{\beta}_0, \hat{\beta}_1)} = 0$$

$$\Rightarrow \quad -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0$$

$$\Rightarrow \quad -2 \sum_{i=1}^{n} x_i \hat{\epsilon}_i = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} \hat{\epsilon}_i x_i = 0$$

We have $\sum_{i=1}^{n} \hat{\epsilon}_i = \sum_{i=1}^{n} \hat{\epsilon}_i x_i = 0$

b) $$\sum_{i=1}^{n} (\hat{\epsilon}_i - \bar{\hat{\epsilon}})(y_i - \bar{\hat{y}}) = \sum_{i=1}^{n} (\hat{\epsilon}_i \hat{y}_i - \hat{\epsilon}_i \bar{\hat{y}} - \bar{\hat{\epsilon}} \hat{y}_i + \bar{\hat{\epsilon}} \bar{\hat{y}})$$

$$= \sum_{i=1}^{n} \hat{\epsilon}_i \hat{y}_i - \sum_{i=1}^{n} \hat{\epsilon}_i \bar{\hat{y}} - \sum_{i=1}^{n} \bar{\hat{\epsilon}} \hat{y}_i + \sum_{i=1}^{n} \bar{\hat{\epsilon}} \bar{\hat{y}}$$

$$= (1) - (2) - (3) + (4) \quad (\text{for short})$$

For (1) $= \sum_{i=1}^{n} \hat{\epsilon}_i \hat{y}_i = \sum_{i=1}^{n} \hat{\epsilon}_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \sum_{i=1}^{n} (\hat{\epsilon}_i \hat{\beta}_0 + \hat{\epsilon}_i \hat{\beta}_1 x_i)$

$= \sum_{i=1}^{n} \hat{\epsilon}_i \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\epsilon}_i \hat{\beta}_1 x_i = \hat{\beta}_0 \underbrace{\sum_{i=1}^{n} \hat{\epsilon}_i}_{=0} + \hat{\beta}_1 \underbrace{\sum_{i=1}^{n} \hat{\epsilon}_i x_i}_{=0}$

(results from **a**)

$\rightarrow$ (1) $= \hat{\beta}_0 (0) + \hat{\beta}_1 (0) = 0 + 0 = 0$

For (2) $= \sum_{i=1}^{n} \hat{\epsilon}_i \bar{\hat{y}} = \sum_{i=1}^{n} \hat{\epsilon}_i n^{-1} \sum_{i=1}^{n} y_i = n^{-1} \sum_{i=1}^{n} \sum_{i=1}^{n} \hat{\epsilon}_i y_i$

$= n^{-1} \sum_{i=1}^{n} \underbrace{\left(\sum_{i=1}^{n} \hat{\epsilon}_i y_i \right)}$ $\rightarrow$ this is term (1), which is proven to be 0 above

$\Rightarrow$ (2) $= n^{-1} \cdot 0 = 0$

For (3) $= \sum_{i=1}^{n} \bar{\hat{\epsilon}} \hat{y}_i = \sum_{i=1}^{n} n^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i \hat{y}_i = n^{-1} \sum_{i=1}^{n} \sum_{i=1}^{n} \hat{\epsilon}_i \hat{y}_i$

Same as (2), (3) $= n^{-1} \cdot 0 = 0$

For (4) $= \sum_{i=1}^{n} \bar{\hat{\epsilon}} \bar{\hat{y}} = \sum_{i=1}^{n} n^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i \, n^{-1} \sum_{i=1}^{n} y_i$

$= n^{-2} \sum_{i=1}^{n} \sum_{i=1}^{n} \left(\sum_{i=1}^{n} \hat{\epsilon}_i y_i \right) \longrightarrow$ term (1) as well

$\Rightarrow$ (4) $= n^{-2} \cdot 0 = 0$

So we have $\sum_{i=1}^{n} (\hat{\epsilon}_i - \bar{\hat{\epsilon}})(\hat{y}_i - \bar{\hat{y}}) = (1) - (2) - (3) + (4)$

$= 0 - 0 - 0 + 0 = 0$

c) $\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ (by definition)

Also $E(y_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i)$

$\Rightarrow \bar{y} = E(\beta_0) + E(\beta_1 x_i) + E(\epsilon_i)$

$\Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} + 0$ (by definition & OLS properties)

$\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, plug this in the equation above

$\Rightarrow \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} \left((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right)^2$

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

But by definition of OLS estimaters we also have

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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

Our equation becomes

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

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

$$\Rightarrow E\left(\sum_{i=1}^{n} \hat{\epsilon}_i^2\right) = E\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right) - E\left(\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right) \hat{\beta}_1^2\right)$$

$$= (1) - \sum_{i=1}^{n} (x_i - \bar{x})^2 (var(\hat{\beta}_1) + E(\hat{\beta}_1)^2)$$

$$= (1) - \sum_{n} (x_i - \bar{x})^2 (\sigma^2 / \sum (x_i - \bar{x})^2 + \beta_1^2)$$

Examine (1) $= E\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right) = E\left(\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i + \epsilon_i - \beta_0 - \beta_1 \bar{x} - \bar{\epsilon})^2\right)$

$$= E\left(\sum_{i=1}^{n} (\beta_1 (x_i - \bar{x}) + (\epsilon_i - \bar{\epsilon}))^2\right)$$

$$= E\left(\sum_{i=1}^{n} \beta_1^2 (x_i - \bar{x})^2 + \sum_{i=1}^{n} (\epsilon_i - \bar{\epsilon})^2 + 2\beta_1 \sum_{i=1}^{n} (x_i - \bar{x})(\epsilon_i - \bar{\epsilon})\right)$$

$$= \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} E(\epsilon_i - \bar{\epsilon})^2 + 2\beta_1 \sum_{i=1}^{n} (x_i - \bar{x}) E(\epsilon_i - \bar{\epsilon})$$

We have $E(\epsilon_i - \bar{\epsilon}) = E(\epsilon_i) - E(\bar{\epsilon}) = 0 - 0 = 0$

and $\sum_{i=1}^{n} E(\epsilon_i - \bar{\epsilon})^2 = E\left(\sum_{i=1}^{n} (\epsilon_i - \bar{\epsilon})^2\right)$

$$= \sigma^2 (n-1)$$

Then (1) will become $(1) = \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sigma^2 (n-1) + 0$

Plug in our initial equation

4c - continued

$$\Rightarrow E\left(\sum_{i=1}^{n}\hat{\epsilon}_i^2\right) = \beta_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2 + \sigma^2(n-1) - \left(\sigma^2 + \beta_1^2\sum(x_i-\bar{x})^2\right)$$

$$= \sigma^2(n-1) - \sigma^2$$

$$\Rightarrow E\left(\sum_{i=1}^{n}\hat{\epsilon}_i^2\right) = \sigma^2(n-2)$$

Now, with $\hat{\sigma}^2$ defined in the question

$$\Rightarrow E(\hat{\sigma}^2) = E\left((n-2)^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2\right) = \frac{1}{n-2}E\left(\sum_{i=1}^{n}\hat{\epsilon}_i^2\right)$$

$$= \frac{1}{n-2}\cdot\sigma^2(n-2) = \sigma^2$$

$$\Rightarrow E(\hat{\sigma}^2) = \sigma^2$$

5) a) $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Y_i : the amount of sales in \$100
X_i : the amount of radio advertising in hours
} for the ith month ($i = 1, \ldots, 24$)

$E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, $Cov(\epsilon_i, \epsilon_j) = 0$ ($i \neq j$)
(constant variance) (uncorrelated with other ϵ's)

ϵ_i are also independent of X and $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$

We also have $E(Y_i) = \beta_0 + \beta_1 X_i$, $Var(Y_i) = \sigma^2$ and $Cov(Y_i, Y_j) = 0$ ($i \neq j$)

b) $\hat{\beta}_1 = 1.15806$
$se\{\hat{\beta}_1\} = 0.04338$
} from the regression results

$\Rightarrow$ These are the esimated change in Y_i (monthly sales) and the change's standard error when X_i (radio ad) increases by 1 hour. So if X_i increases by 10 hours

$$\Rightarrow \begin{cases} \hat{\beta}_1^{(10hrs)} = 11.5806 \ (= 10 \cdot \hat{\beta}_1) \\ se\{\hat{\beta}_1^{(10hrs)}\} = 0.4338 \ (= \sqrt{10^2 \cdot se^2\{\hat{\beta}_1\}} = 10\, se\{\hat{\beta}_1\}) \end{cases}$$

Also, with a 95% CI we have $\alpha = 0.05$, and we have $n = 24$

$$\Rightarrow t(1-\alpha/2, n-2) = t(1 - \frac{0.05}{2}, 24-2) = t(0.975, 22) = 2.074$$

$\Rightarrow$ 95% CI for $\hat{\beta}_1^{(10hrs)}$ is $11.5806 \pm 2.074\,(0.4338)$
or $10.6809 \leq \hat{\beta}_1^{(10hrs)} \leq 12.4803$

c) (i) With $X_h = 0$ $\Rightarrow \hat{Y}_h = 101.5757 + 0\,(1.15806)$
$\Rightarrow \hat{Y}_h = 101.5757$

From R output, $MSE = 2.74054$

$$s^2(\hat{Y}_h) = MSE\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right)$$

We also have $\bar{Y} = 114.88255 \Rightarrow \bar{X} = \frac{\bar{Y} - \hat{\beta}_0}{\hat{\beta}_1} = \frac{114.88255 - 101.57570}{1.15806}$

$\Rightarrow \bar{X} = 11.4906$

Also $\sum (X_i - \bar{X})^2 = 1952.65488$ (R output)

$$\Rightarrow s^2(\hat{Y}_h) = 2.74054\left(\frac{1}{24} + \frac{(0 - 11.4906)^2}{1952.65488}\right) = 0.2995$$

5c) **continued**

$\Rightarrow s^2\{pred\} = MSE + s^2(\hat{Y}_h)$

$= 2.74054 + 0.2995 = 3.0400$

$\Rightarrow s\{pred\} \simeq 1.74357$

$\Rightarrow$ 95% prediction interval for $\hat{Y}_h$ is $\hat{Y}_h \pm t(1-\frac{\alpha}{2}, n-2)$ **s{pred}**

which is $101.5757 \pm 2.074(1.74357)$

or $97.9595 \leq \hat{Y}_h \leq 105.1919$

(ii) The normality is important for the prediction interval because as opposed to CI, we are focusing on only 1 new value to make the prediction interval. Unlike CI, where we don't really need the normality assumption as much as prediction interval because we ~~have~~ build CI using observed data.

5d) $H_0: \beta_1 = 0$

$H_A: \beta_1 > 0$ (radio ad increases the amount of sales)

Our test statistic is $t^* = 26.69$ (from R output)

the p-value to obtain this test statistic is $< .0001$

Using a significance level $\alpha = 0.05$, we can see that the probability of getting an estimate for β_1 **as extreme as we got** is less than our $\alpha = 0.05$

$\Rightarrow$ We reject the null $H_0: \beta_1 = 0$

$\Rightarrow$ Radio advertisement tends to increase the amount of sales