B705 - Homework 1

Austin Allen

2024-02-01

1. Given the Motor Trend Car Road Tests (mtcars) dataset (mtcars.csv):

Let

$$X_i = \begin{cases} 1, \text{ if *wt} > \text{ median} \\ 0, \text{ if wt} \leq \text{ median} \end{cases}$$

where X is called an indicator or dummy variable.

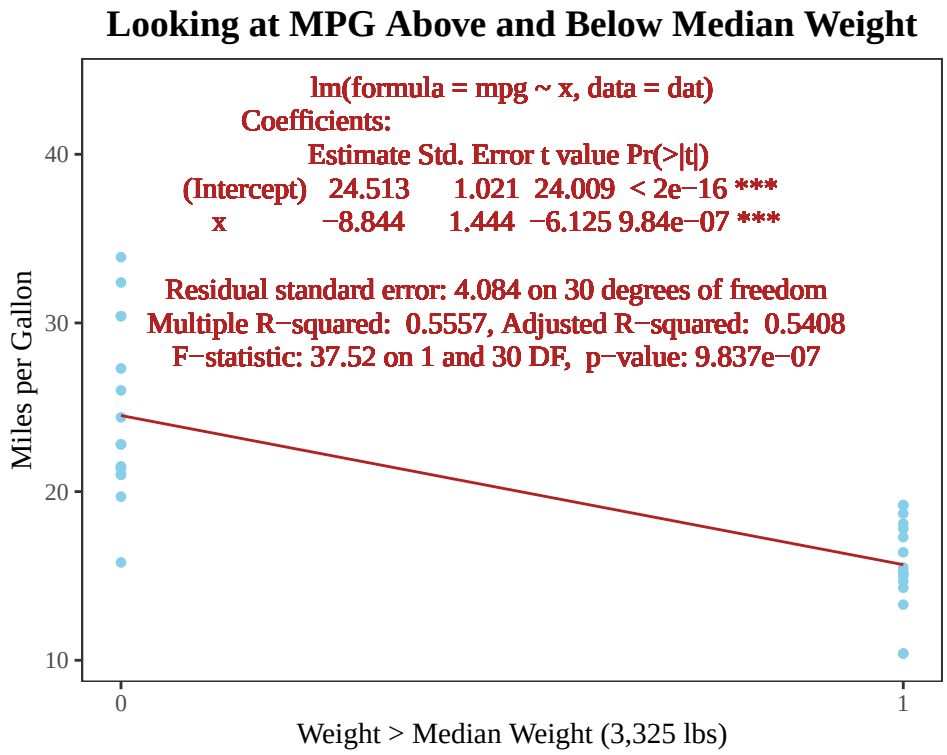
In R type help(mtcars) will provide more details on the dataset and variables name.

*wt = weight (1000 lbs)

a) Fit a simple linear regression, using $\text{mpg}_i = \beta_0 + \beta_1 X_i + \epsilon$. How would you interpret $\hat{\beta}_1$ in the context of the problem?

```
# Read in data
dat <- read.csv("mtcars.csv")
# Create indicator variable X
dat$x <- as.numeric(dat$wt > median(dat$wt))
# Create model
model_a <- lm(mpg ~ x, data = dat)
# Create a plot
ggplot(dat) +
  geom_point(aes(x = x, y = mpg), size = 1.2, col = "skyblue") +
  geom_smooth(aes(x = x, y = mpg), method = "lm", se = FALSE, col = "firebrick", lwd = .5) +
```

```
  labs(x = "Weight > Median Weight (3,325 lbs)", y = "Miles per Gallon", title = "Looking at
  scale_x_continuous(breaks = c(0, 1)) +
  theme_bw() +
  labels_a +
  my_theme
```



Based on the model output, as x increases by 1, the estimated value of mpg decreases by 8.844. Another way to say this is that we expect cars that weigh over 3,325 pounds to get 8.844 less miles for every gallon than cars below the median weight of 3,325 pounds.

b) Write the above model in a matrix form (ie, $Y = X\beta + \epsilon$). Identify the design matrix X.

$$\text{mpg} = X\beta + \epsilon$$

$> X$ is the design matrix containing two columns ($p + 1$ or total number of parameters in the model). The first column contains only 1's to be multiplied by β_0. The second column contains 1's and 0's corresponding to the vehicle with the weight above the median weight. Here's what this matrix looks like:

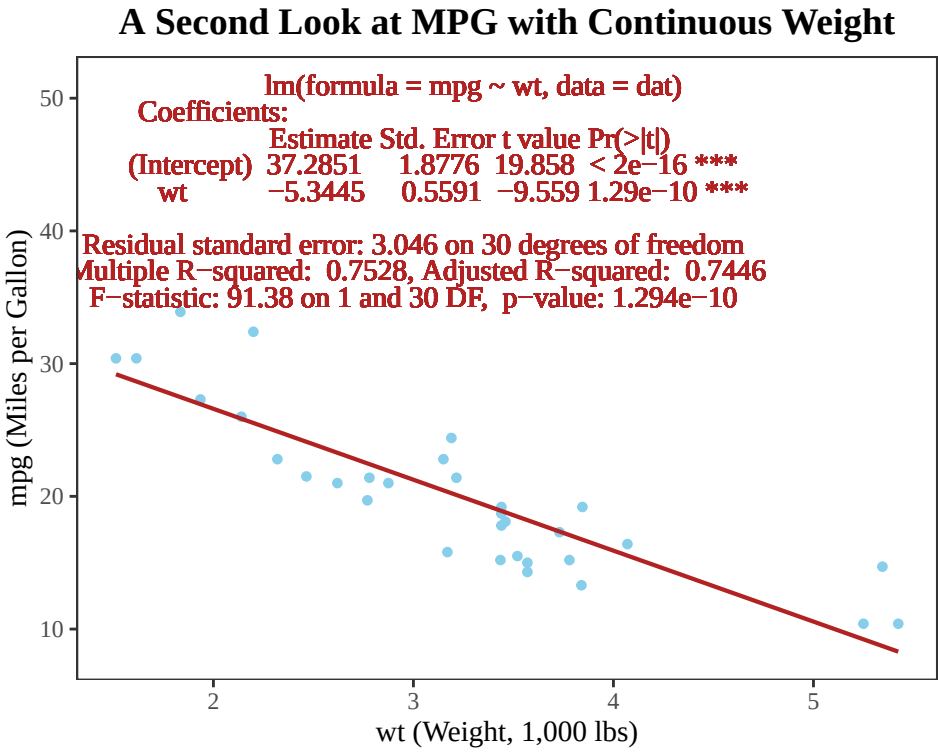
	Intercept	X
1	1	0
2	1	0
...	...	...
32	1	1

c) What is the estimated mean square errors (MSE)?

The estimated `MSE` can be calculated from the summary from `R` shown on the plot above. The `residual standard error` is the square root of the estimated `MSE`, so by squaring this term we get an estimated MSE $= 4.084^2 = 16.68$.

d) Re-do part a) using $\text{mpg}_i = \beta_0 + \beta_1 \text{wt}_i + \epsilon_i$. How does this results differ from part a)? Which model would you recommend? why?

```
# Create model
model_d <- lm(mpg ~ wt, data = dat)
# Create a plot
ggplot(dat) +
  geom_point(aes(x = wt, y = mpg), size = 1.2, col = "skyblue") +
  geom_smooth(aes(x = wt, y = mpg), method = "lm", se = FALSE, col = "firebrick", lwd = .75)
  labs(x = "wt (Weight, 1,000 lbs)", y = "mpg (Miles per Gallon)", title = "A Second Look at
  theme_bw() +
  labels_d +
  my_theme
```



While both slopes are significant, treating `wt` as a continuous variable increases R^2 by over 20%. That's an incredible amount of information that is lost when we dichotomize `wt`. I recommend treating `wt` as a continuous variable, as is show in the model produced in part d).

e) Suppose we fitted a model in part a) by adding hp (horse power) and an interaction, i.e. mpgi$_i = \beta_0 + \beta_1 X_i + \beta_2hp_i + \beta_3 X_ihp_i + \epsilon_i$. How would you interpret each term? Is this model additive?

Let's look at a summary of the model created in R:

```
# Create model
model_e <- lm(mpg ~ x + hp + x*hp, data = dat)

# Print summary of model
summary(model_e)
```

```
Call:
```

```
lm(formula = mpg ~ x + hp + x * hp, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8599 -2.5937  0.0661  1.9823  6.6097

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.81846    1.81824  17.500  < 2e-16 ***
x           -11.47936    3.30937  -3.469 0.001710 **
hp           -0.06966    0.01564  -4.453 0.000124 ***
x:hp          0.04489    0.02104   2.133 0.041786 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.134 on 28 degrees of freedom
Multiple R-squared:  0.7557,    Adjusted R-squared:  0.7295
F-statistic: 28.87 on 3 and 28 DF,  p-value: 1.026e-08
```

Because the interaction term is significant, this model is not additive. That means that we cannot interpret each term independently because X and hp depend on each other.

To interpret this model, we can rearrange the terms as follows:

$$\begin{aligned}\hat{\text{mpg}} &= 31.81 - 11.5X - 0.07\text{hp} + 0.05X\text{hp}_i \\ &= 31.81 - 11.5X - hp(0.07 - 0.05X) \\ &= 31.81 - 0.07\text{hp} - X(11.5 - 0.05hp)\end{aligned}$$

Here's a summary of how to interpret each term:

- To interpret hp, we hold X constant (equation 2) and say that for constant x, the effect of hp on mpg is -(0.07 - 0.05) = -0.02.
- To interpret X, we hold hp constant (equation 3) and say that for constant hp, the effect of X on mpg is -(11.5 - 0.05) = -11.45.

f) Similarly, we added hp and an interaction to the model in part e), i.e. $\textbf{mpg}_i = \beta_0 + \beta_1\textbf{wt}_i + \beta_2\textbf{hp}_i + \beta_3\textbf{wt}_i\textbf{hp}_i + \epsilon_i$. How should the interaction term be interpreted in this model? Can we interpret the main effects? why or why not?

The real question here is how the interpretation of the interaction changes when we moved from a dichotomized variable (x = wt > median(wt)) to a continuous variable (wt).

Let's look at a summary of the `lm()` statement in `R`:

```
# Create model
model_f <- lm(mpg ~ wt + hp + wt*hp, data = dat)

# Print summary of model
summary(model_f)
```

```
Call:
lm(formula = mpg ~ wt + hp + wt * hp, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0632 -1.6491 -0.7362  1.4211  4.5513

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.80842    3.60516  13.816 5.01e-14 ***
wt          -8.21662    1.26971  -6.471 5.20e-07 ***
hp          -0.12010    0.02470  -4.863 4.04e-05 ***
wt:hp        0.02785    0.00742   3.753 0.000811 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.153 on 28 degrees of freedom
Multiple R-squared:  0.8848,    Adjusted R-squared:  0.8724
F-statistic: 71.66 on 3 and 28 DF,  p-value: 2.981e-13
```

You'll notice the interaction term has a p-value that is much lower than the p-value of the interaction of the dichotomized model.

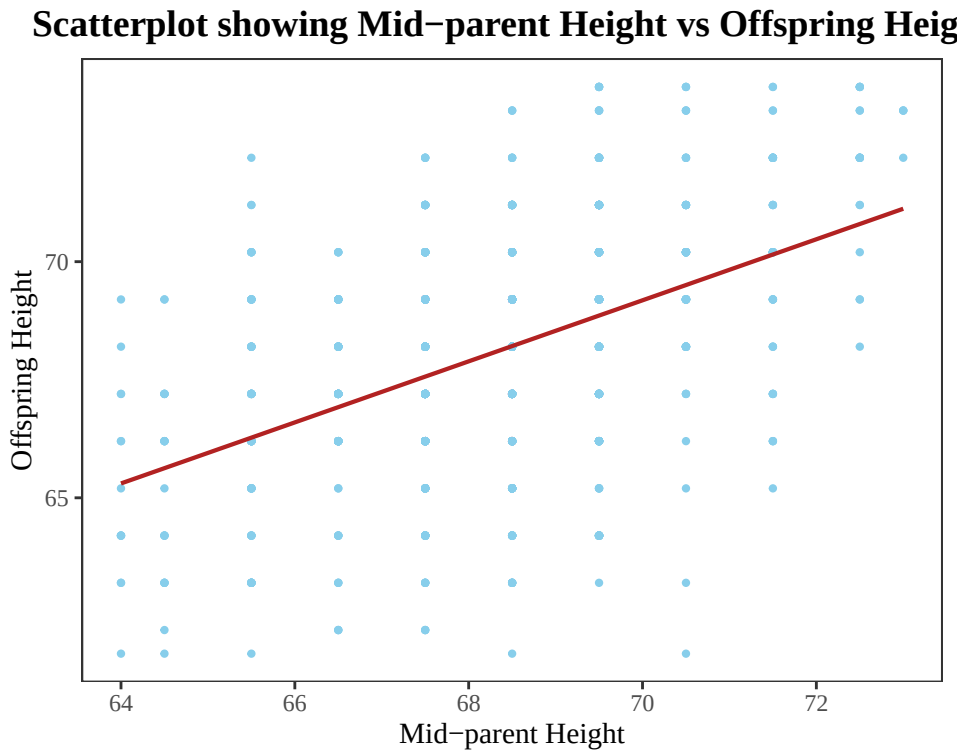
We interpret the model exactly how we did in part e as follows:

$$\begin{aligned}\widehat{\text{mpg}} &= 49.8 - 8.2wt - 0.12\text{hp} + 0.02wt\text{hp}_i \\ &= 49.8 - 8.2wt - hp(0.12 - 0.02wt) \\ &= 49.8 - 0.12\text{hp} - wt(8.2 - 0.02hp)\end{aligned}$$

2. Stigler, History of Statistics pg.285 gives Galton's famous data on heights of sons (Y in inches) and average parents height (X in inches) scaled to represent a male height (essentially sons' heights versus fathers' heights). Data are given in `parents_offsprings.csv`. Consider a statistical model for these data, randomly sampled from some population of interest. In particular, choose a model which accounts for the apparent linear dependence of the mean height of sons on midparent height X.

a) Construct a scatterplot of these data.

```
# Load the data
dat <- read.csv("parents_offsprings.csv")
# Create a scatterplot
ggplot(dat) +
  geom_point(aes(x = midparent_height, y = offspring_height), col = "skyblue", size = .8) +
  geom_smooth(method = "lm", aes(x = midparent_height, y = offspring_height), col = "firebri
  labs(x = "Mid-parent Height", y = "Offspring Height", title = "Scatterplot showing Mid-par
  theme_bw() +
  my_theme
```



b) What is β_1? Is this statistic or parameter?

β_1 is the parameter that defines the *true* relationship (i.e. slope) between a mid-parent's height and the offspring's height for individuals in our source population. It is unknown, unless we can measure every indiviual in the population, which we cannot do. It is a parameter, not a statistic.

c) What is the estimated slope ($\hat{\beta}_1$)? Is this statistic or parameter?

The estimated slope, $\hat{\beta}_1$, is a point estimate for the parameter β_1 defined above. We can actually calculate this estimate by hand, or we can have `R` do it for us. Let's take the easier road and let our computer do the heavy lifting.

```
# Create model
model_c <- lm(offspring_height ~ midparent_height, data = dat)
# Pring model summary
summary(model_c)
```

```
Call:
lm(formula = offspring_height ~ midparent_height, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      23.94153    2.81088   8.517   <2e-16 ***
midparent_height  0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,    Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

In the summary output shown above, $\hat{\beta}_1$ is shown in the `Coefficients` table in the `Estimates` column and `midparent_height` column. You can see that our point estimate is 0.646. This is a statistic calculated from our data which estimates the true parameter, β_1.

d) How much does $\hat{\beta}_1$ vary about β_1 from sample to sample? (Provide an estimate of the standard error, as well as an expression indicating how it was computed)

Again, from the output in R, the point estimate for the standard error of $\hat{\beta}_1$ is 0.041. The expression for calculating $\sigma_{\hat{\beta}_1}$ can be given as follows: $\sigma_{\hat{\beta}_1} = \sqrt{\frac{\text{SSE}}{(n-2)\cdot S^2 X}}$

Just for fun, let's attempt to calculate this with some R code:

```r
# Set variables
n <- nrow(dat)
sse <- sum( (dat$offspring_height - model_c$fitted.values)^2 )
s_squared_x <- sum( (dat$midparent_height - mean(dat$midparent_height))^2)
# Calculate point estimate for the standard error
se_bh1 <- sqrt( sse/( (n - 2) *(s_squared_x) ) )
```

The output is 0.041, which is the same as the ouput in the R summary.

e) What is a region of plausible values for β_1 suggested by the data?

Let's look at a region of plausible values for this parameter by calculating a 95% confidence interval:

```
# Set variables
bh1 <- 0.64629 # from summary output
t <-  qt(c(.05/2, 1-.05/2), (n - 2)) # desired range
se <- 0.04114 # from summary output

confint <- bh1 + t * se
```

From my calculation, the 95% confidence interval is [0.566, 0.727].

f) What is the line that best fits these data, using the criterion that the smallest sum of squared residuals is "best?"

The line that best fits these data according to the criterion given can actually be taken from the R output as well:

$$\widehat{\text{offspring\_height}} = 23.942 + 0.646 \cdot \text{midparent\_height}$$

g) How much of the observed variation in the heights of sons (the y-axis) is explained by this "best" line?

The R^2 value for this model is 0.211. That means that roughly 21% of the variation in the height of sons is explained by the height of the parents, which is surprisingly low, in my opinion.

h) What is the estimated average height of sons whose midparent height is $x = 68$?

Using the "best" line in the previous answer, we can calculate it easily.

```
# Calculate predicted height
predicted_height <- 23.942 + 0.646 * 68
```

The estimated average heigh of sons whose midparent height is $x = 68$ is 67.87.

i) Is this the true average height in the whole population of sons whose midparent height is $x = 68$?

No, it's not! We would need to know the values of β_0 and β_1 to know the true average height. Our line is the "hat" line, so to speak. It's just an estimate.

j) Under the model, what is the true average height of sons with midparent height is $x = 68$?

$$\mu_{68} = \beta_0 + \beta_1 \cdot 68$$

k) What is the estimated standard deviation among the population of sons whose midparent height is $x = 68$? Would you call this standard deviation a "standard error"?

This is the same thing as $S_{Y.X}$, which is 2.239. I would not call this a standard error because it estimates general variability and isn't used to estimate the variability of a sampling distribution.

l) What is the estimated standard deviation among the population of sons whose midparent height is $x = 72$? Bigger, smaller, or the same that for $x = 68$? Is your answer obviously supported or refuted by inspection of the scatterplot?

One assumption we make in linear regression is that the variance is constant for all error terms. This implies that the standard deviation among the population of sons whose midparent height is $x = 72$ will be the same as that of those whose midparent height is $x = 68$, which is 2.239. This isn't obvious when looking at the scatterplot, because it looks to me like the variability is lower around 68 at 72. Variance may not be constant after all, but in my opinion it's not strong enough to say either way.

m) What is the estimated standard error of the estimated average for sons with midparent height $x = 68$, $\hat{\mu}(68) = \hat{\beta}_0 + 68\hat{\beta}_1$? Provide an expression for this standard error.

This can be calculated as follows:

$$\begin{aligned} Var(\hat{\mu}(\hat{68})) &= S^2_{Y.X}[1/n + (68 - \bar{X})^2/(\sum_{i=1}^{n}(X_i - \bar{X})^2)] \\ &= 5.01[1/928 + (68 - 68.31)^2/2961.36] \\ &= 0.00556 \end{aligned}$$

n) Is the estimated standard error of $\hat{\mu}(72)$ bigger, smaller, or the same as that for $\hat{\mu}(68)$?

$$\begin{aligned} Var(\hat{\mu}(\hat{72})) &= S^2_{Y.X}[1/n + (72 - \bar{X})^2/(\sum_{i=1}^{n}(X_i - \bar{X})^2)] \\ &= 5.01[1/928 + (72 - 68.31)^2/2961.36] \\ &= 0.0285 \end{aligned}$$

This is a little bigger because we're a little further away from $\bar{X}$.

o) Is the observed linear association between son's height and midparent height strong? Report a test statistic.

We can think about this association by testing the following hypothesis:

$$H_0 : \beta_1 = 0 H_A : \beta_1 \neq 0$$

In the summary output, we see that the test statistic for the slope estimate is 15.711, yielding a p-value of 2e-16. This p-value is extremely small, indicating that we would reject the null hypothesis. There appears to be a linear relationship.

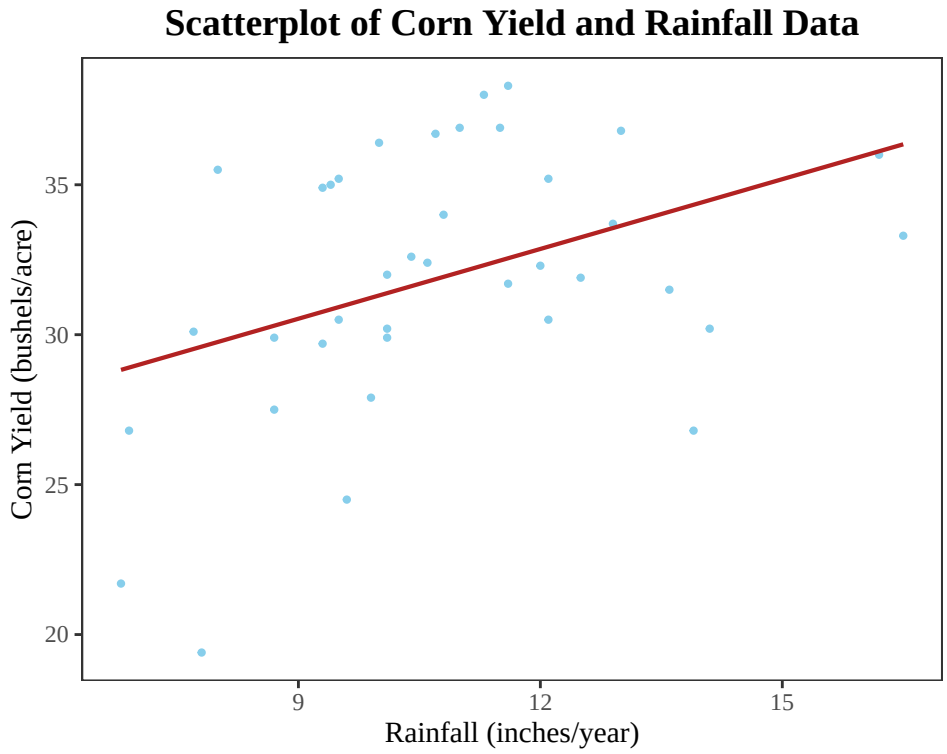
p) What quantity can you use to describe or characterize the linear association between son's height and midparent height in the whole population? Is this a parameter or a statistic?

β_1 is the slope of the true regression line, and it describes the association between son's height and midparent height. It is a parameter.

3. Consider data on corn yield Y (bushels/acre) and rainfall X (inches/yr) in six Midwestern states recorded from 1890 to 1927. Data are given in `corn_yield_and_rainfall.csv`.

a) Construct a scatterplot of corn yield and rainfall measurements.

```
# Load in the data
dat <- read.csv("corn_yield_and_rainfall.csv")
# Create a scatterplot
ggplot(dat) +
  geom_point(aes(x = Rainfall, y = Yield), col = "skyblue", size = .8) +
  geom_smooth(method = "lm", aes(x = Rainfall, y = Yield), col = "firebrick", lwd = .75, se =
  labs(x = "Rainfall (inches/year)", y = "Corn Yield (bushels/acre)", title = "Scatterplot o
  theme_bw() +
  my_theme
```



b) How can we describe the association between yield and rainfall? Does it appear linear?

Looking at the scatter plot, I'm concerned that there's some non-linearity in this plot. The data suggest to me to have some sort of negative quadratic relationship. It seems like optimal rainfall is around 11 inches per year, and anything above or below that is associated with lower corn yield.

While this is something to be concerned about, I'm going to print the summary of the linear model between yield and rainfall below to assist me with subsequent questions:

```
# Create model
model_b <- lm(Yield~ Rainfall, data = dat)
# Print summary
summary(model_b)
```

```
Call:
lm(formula = Yield ~ Rainfall, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2014  -2.3530  -0.2577   3.8929   5.7515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.5521     3.2365   7.277 1.43e-08 ***
Rainfall      0.7755     0.2939   2.639   0.0122 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.049 on 36 degrees of freedom
Multiple R-squared:  0.1621,    Adjusted R-squared:  0.1388
F-statistic: 6.965 on 1 and 36 DF,  p-value: 0.01221
```

c) How can we measure the strength of the linear association?

One way to measure the strength of the linear association is by looking at the R^2 value. The R^2 value in this model is 0.162, which is quite low.

d) To what degree is the variability in yield is described or explained by its association with rainfall? In what units is this variability measured?

We can use adjusted R^2 to answer this question.

Based on the adjusted R-squared value of 0.139, approximately 13.9% of the variability in yield can be explained by its association with rainfall. R-squared is unitless, indicating the proportion of explained variability. However, if the question refers to the units of the actual yield values used in the analysis, the units of the variance of the yield would be $\frac{\text{inches}^2}{\text{years}^2}$

e) How can we use this association to estimate average yield, given a certain level of rainfall, say 14 inches/yr?

We can use the `lm` function in `R` to estimate average yield given a certain level of rainfall. I will show how to calculate a confidence interval in `R`:

```
# Calculate confidence interval
prediction <- predict(model_b, newdata = data.frame(Rainfall = 14), interval = "confidence")
print(prediction)
```

```
       fit      lwr      upr
1 34.40979 32.07566 36.74392
```

This shows that the estimated mean yield given 14 inches of rain is 34.4, with a lower bound of a 95% confidence interval being 32.1 and the upper bound of the interval being 36.7.

f) How can we use this association to predict future yield, if we have an idea about what the rainfall will be? use $x_0 = 14$ (inches/yr).

Again, we can use the `lm` and `predict` functions in `R` to make this prediction (I will also include a 95% confidence interval):

```
# Calculate confidence interval
prediction <- predict(model_b, newdata = data.frame(Rainfall = 14), interval = "prediction")
print(prediction)
```

```
       fit      lwr      upr
1 34.40979 25.87183 42.94775
```

The estimate is exactly the same, but the confidence interval is from 25.9 to 42.9, which is a much larger interval.

g) What is the difference between e) and f) ?

The difference is that part e) is estimating an average and part f) is estimating a single crop yield. The way I think of it is that our confidence interval for the mean is based on the sampling distribution for $\bar{Y}$ which has variance $Var(\bar{Y}) = \sigma^2/n$, whereas the confidence interval for a single crop yield is based on the population distribution, which has variance $Var(Y) = \sigma^2$. Thus, the variance is greater for a single crop yield "yielding" a wider confidence interval.

h) What proportion of variance in yield is explained by the linear regression model?

The model's R^2 value is 0.162. Thus, approximately 16.2% of variance in yield is explained by the linear regression model.

4. An investigator wants to examine the relationship between body temperature Y and heart rate X. Further, he would like to use heart rate to predict the body temperature. (Data in `BodyTemperature.csv`).

a) Construct a simple linear regression model for body temperature using heart rate as the predictor.

```
# Load in the data
dat <- read.csv("BodyTemperature.csv")
# Create model
model_a <- lm(Temperature ~ HeartRate, data = dat)
```

$$\text{BodyTemperature}_i = \beta_0 + \beta_1 \text{HeartRate}_i + \epsilon_i$$

b) Interpret the estimated slope ($\hat{\beta}_1$) and examine its statistical significance.

> The estimated slope in our model is 0.08. This means that for every unit increase in HeartRate, we expect an increase of 0.08 degrees in body temperature. The p-value is at 3.01e-06, indicating statistical significance.
>
> .08, is significant

c) Construct the 95% confidence interval for the population slope β_1. How would you interpret this CI and do es it agree with your conclusion in part (b)?

```
# Define variables
bh1 <- 0.08063 # from summary output
t <- qt(c(.05/2, 1 - .05/2), df = nrow(dat) - 1)
se <- 0.016 # from summary output

confint <- bh1 + t * se
```

> The 95% confidence interval for the population slope is [0.049, 0.112]. Thus, we are 95% confident that the true slope is between these values. This supports my conclusion in the previous response because 0 is not contained in the interval.

d) Calculate adjusted R^2, what does this tell us about the contribution of heart rate in the model?

Adjusted R^2 can be calculated with the following formula:

$$\begin{aligned} R^2_{\text{Adjusted}} &= 1 - \frac{\text{SSE}(n-1)}{\text{SST}(n-2)} \\ &= 1 - \frac{72.48(99)}{90.65(99)} \\ &= 0.192 \end{aligned}$$

Thus, HeartRate explains roughly 19.2% of the variability in Temperature.

e) If someone's heart rate is 75, what would be your estimate of this persons body temperature?

$$\begin{aligned} \text{Estimated Temperature} &= 92.39 + 0.0806\text{HeartRate} \\ &= 92.39 + 0.080675 \\ &= 98.44 \end{aligned}$$

Thus, the estimated termperature given a heart rate of 75 is 98.44.

f) Suppose the investigator believes that adding sex to the model in part (a) will improve model prediction. Is he correct? How would you quantify the contribution of sex to the model?

```
# Add sex to model
model_f <- lm(Temperature ~ HeartRate + Sex, data = dat)
# Print summary of new model
summary(model_f)
```

```
Call:
lm(formula = Temperature ~ HeartRate + Sex, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-2.37056 -0.48862 -0.00963  0.53575  2.68538

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 92.43764    1.18902  77.743  < 2e-16 ***
HeartRate    0.08199    0.01612   5.088 1.77e-06 ***
SexM        -0.30044    0.17041  -1.763    0.081 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8509 on 97 degrees of freedom
Multiple R-squared:  0.2252,    Adjusted R-squared:  0.2093
F-statistic:  14.1 on 2 and 97 DF,  p-value: 4.212e-06
```

He was correct! Our adjusted R^2 increased by 0.017. This constitutes an additional 1.7% percent of explanation in the model.

g) Suppose the investigator added an interaction between heart rate and sex to the model in part (f), how would you Interpret this interaction term?

```r
# Add sex to model
model_g <- lm(Temperature ~ HeartRate + Sex + HeartRate*Sex, data = dat)
# Print summary of new model
summary(model_g)
```

```
Call:
lm(formula = Temperature ~ HeartRate + Sex + HeartRate * Sex,
    data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-2.33872 -0.47940 -0.00335  0.53645  2.69775

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    92.183303   1.854876  49.698  < 2e-16 ***
HeartRate       0.085457   0.025214   3.389  0.00102 **
SexM            0.133777   2.428067   0.055  0.95618
HeartRate:SexM -0.005898   0.032899  -0.179  0.85810
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8552 on 96 degrees of freedom
Multiple R-squared:  0.2255,    Adjusted R-squared:  0.2013
F-statistic: 9.317 on 3 and 96 DF,  p-value: 1.822e-05
```

Unfortunately, the interaction term is not significant, so it does not offer any interpretable value.