2.6-a

Observation	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	1	16	0	0	1.8	14.2	3.24
2	0	9	-1	1	-5.2	10.2	1.44
3	2	17	1	1	2.8	18.2	1.44
4	0	12	-1	1	-2.2	10.2	3.24
5	3	22	2	4	7.8	22.2	0.04
6	1	13	0	0	-1.2	14.2	1.44
7	0	8	-1	1	-6.2	10.2	4.84
8	1	15	0	0	0.8	14.2	0.64
9	2	19	1	1	4.8	18.2	0.64
10	0	11	-1	1	-3.2	10.2	0.64

$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{10} * (1+0+2+0+3+1+0+1+2+0) = 1$

$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{10} * (16+9+17+12+22+13+8+15+19+11) = 14.2$

$\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{0+5.2+2.8+2.2+15.6+0+6.2+0+4.8+3.2}{0+1+1+1+4+0+1+0+1+1} = 4$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 * \bar{x} = 14.2 - 4 * 1 = 10.2$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i$ (See table for computations)

$MSE = \frac{1}{n-2} * \sum(y_i - \hat{y}_i)^2$

$MSE = \frac{1}{10-2} * (3.24+1.44+1.44+3.24+0.04+1.44+4.84+0.64+0.64+0.64)$

$MSE = 2.2$

$se(\hat{\beta}_1) = \sqrt{\frac{MSE}{\sum(x_i-\bar{x})^2}} = \sqrt{\frac{2.2}{0+1+1+1+4+0+1+0+1+1}} = 0.4690416$

$t^* = 2.306$

$CI \in \hat{\beta}_1 \pm t^* * se(\hat{\beta}_1) = 4 \pm 2.306 * 0.4690416$

$CI \in [2.9183901, 5.0816099]$

The 95% confidence interval for β_1 is $[2.9183901, 5.0816099]$. This means that we are 95% certain that $\beta_1 \in [2.9183901, 5.0816099]$.

2.6-b

```
## t = 2.306
```

```
## t_s = 8.528029
```

The null hypothesis is that there is no linear relationship ($\beta_1 = 0$). The alternate hypothesis is that there is a linear relationship ($\beta_1 \neq 0$). Since the test statistic ($t_s = 8.528029$) is greater than the critical t-value ($t = 2.306$), we reject the null hypothesis. Therefore, there is a linear relationship. The p-value is 0.

2.6-c

```
## [8.670373, 11.72963]
```

The 95% confidence interval for β_0 is [8.670373, 11.72963]. This means that we are 95% certain that $\beta_0 \in [8.670373, 11.72963]$.

2.6-d

```
## t = 2.306
```

```
## t_s = 1.809068
```

The null hypothesis is that the mean number of ampules does not exceed 9.0 ($\beta_0 \leq 9.0$). The alternate hypothesis is that the mean number of ampules does exceed 9.0 ($\beta_0 > 9.0$). Since the test statistic ($t_s = 1.809068$) is greater than the critical t-value (2.306), we accept the null hypotehsis. Therefore, the mean number of ampules does exceed 9.0. Our degrees of freedom is 9. Using the t-table, the p-value is approximately 0.05.

2.6-e

$H_0 : \beta_1 = 0 : \delta = \frac{|2-0|}{.5} = 4, power = .93$ $H0 : \beta_0 \leq 9 : \delta = \frac{|11-9|}{.75} = 2.67, power = .78$

2.9

The variance of $\hat{Y}_h$ is not given because the table is analyzing the variance of the regression parameters (b_0 and b_1), not analyzing the variance of predictors $\hat{Y}_h$.

2.10

a. A prediction interval is appropriate because we are trying to predict what the humidity level will be the next day which is a new observation.
b. Confidence interval is appropriate because we are trying to estimate the average of families based on previous data which is on an existing observation.
c. A confidence interval is appropriate because we are trying to figure out a reasonable electricity usage for next month given nothing changes and this month's electricity usage. Even though we are predicting a future observation, but it is based on existing observations.

2.12

Yes, the variance of pred can be brought closer to 0 as we have more samples because with more samples our overall prediction is usually more precise. This is not the case for the variance of $\hat{Y}_h$ because it is only one point in our data and one of the terms in this equation has $\frac{1}{n}$.

2.27-a

```
## [1] 4.123987e-19
```

$\alpha = 0.05$ The null hypothesis is there is no negative correlation ($\beta_1 \geq 0$). The alternative hypothesis is there is a negative correlation ($\beta_1 < 0$). Using a left tail test where $\alpha = 0.05$, we reject the null hypotehsis when the p-value is less than alpha ($p < 0.05$), while we fail to reject the null when the p-value is greater than alpha ($\alpha > 0.05$). Our p-value is around 0, so we reject the null hypotehsis. Therefore, we conclude there is a negative correlation.

2.27-b

No, because the linear model is too simple to catch all ages, it can only capture the age range in our dataset (it does not capture any women age less than 40 or women of age 0). Therefore, β_0 is not a good estimator for early ages.

2.27-c

```
## Confidence Interval is [-1.370545, -1.009446]
```

It is not necessary to know the ages to make the estimate because we are making an overall estimate, not an estimate at a specific age.

2.29-b

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x          1 11627.5 11627.5  174.06 < 2.2e-16 ***
## Residuals 58  3874.4    66.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.29-d

Proportion is 0.2499332, which is relatively small.

2.29-e

$R^2 = 0.7500668$

$r = -0.866064$

Appendix

```r
knitr::opts_chunk$set(echo = FALSE)

# 2.6-a
n <- 10
x_i <- c(1, 0, 2, 0, 3, 1, 0, 1, 2, 0)
y_i <- c(16, 9, 17, 12, 22, 13, 8, 15, 19, 11)

x_bar <- mean(x_i)
y_bar <- mean(y_i)

beta_1_hat <- sum((x_i - x_bar) * (y_i - y_bar)) / sum((x_i - x_bar) ** 2)
beta_0_hat <- y_bar - beta_1_hat * x_bar
y_i_hat <- beta_0_hat + beta_1_hat * x_i

mse <- 1 / (n - 2) * sum((y_i - y_i_hat) ** 2)

se_beta_1_hat <- sqrt(mse / sum((x_i - x_bar) ** 2))

# df = n - 2 because two restrictions
critical_t_value <- 2.306

confidence_lower_bound <- beta_1_hat - critical_t_value * se_beta_1_hat
confidence_upper_bound <- beta_1_hat + critical_t_value * se_beta_1_hat

nicely_formatted_table <- data.frame(Observation = 1:n)
nicely_formatted_table[["$x_i$"]] <- x_i
nicely_formatted_table[["$y_i$"]] <- y_i
nicely_formatted_table[["$x_i - \\bar x$"]] <- x_i - x_bar
nicely_formatted_table[["$(x_i - \\bar x)^2$"]] <- (x_i - x_bar) ** 2
nicely_formatted_table[["$y_i - \\bar y$"]] <- y_i - y_bar
nicely_formatted_table[["$\\hat y_i$"]] <- y_i_hat
nicely_formatted_table[["$(y_i - \\hat y_i)^2$"]] <- (y_i - y_i_hat) ** 2

# 2.6-b
t_s <- (beta_1_hat - 0) / se_beta_1_hat
cat("t =", critical_t_value)
cat("t_s =", t_s)

# 2.6-c
se_beta_0_hat <- sqrt(mse * (1 / n + x_bar ** 2 / sum((x_i - x_bar) ** 2)))

confidence_lower_bound <- beta_0_hat - critical_t_value * se_beta_0_hat
confidence_upper_bound <- beta_0_hat + critical_t_value * se_beta_0_hat
cat("[", confidence_lower_bound, ", ", confidence_upper_bound, "]", sep = "")
t <- 2.2621
t_s <- (beta_0_hat - 9) / se_beta_0_hat
cat("t =", critical_t_value)
cat("t_s =", t_s)

# 2.27-a
file_path <- "CH01PR27.txt"
```

```r
data <- read.table(file_path)
y <- data$V1
x <- data$V2
n <- nrow(data)

model <- lm(y ~ x, data = data)
summary_model <- summary(model)
summary_model$coefficients["x", "Pr(>|t|)"]

# 2.27-c
coef <- model$coefficients
b0hat <- coef[1]
b1hat <- coef[2]
mse <- summary_model$sigma ** 2
alpha <- 0.05
p <- 1 - alpha / 2
se_b1hat <- summary_model$coefficients["x", "Std. Error"]
lb_b1hat <- b1hat - qt(p, df = n - 2) * se_b1hat
ub_b1hat <- b1hat + qt(p, df = n - 2) * se_b1hat
cat("Confidence Interval is [", lb_b1hat, ", ", ub_b1hat, "]", sep = "")

# 2.29-b
data <- read.table("CH01PR27.txt")
y <- data$V1
x <- data$V2
model <- lm(y ~ x, data = data)
anova_result <- anova(model)
anova_result
sse <- anova_result$"Sum Sq"[length(anova_result$"Sum Sq")]
ssto <- sum(anova_result$"Sum Sq")
proportion <- sse / ssto
big_r_squared <- (ssto - sse) / ssto
little_r <- cor(x, y)
```