Lecture 6: Gelman Hill Ch 3

Linear Regression

"Linear regression summarizes how the average value of a numerical *outcome* variable vary over subpopulations defined by linear functions of *predictors*... By focusing on regression as a comparison of averages, we (GH) are being explicit about its limitations for defining these relationships casually... Regression can be used to predict an outcome given a linear function of these predictors, and regression coefficients can be thought of as comparisons across predicted values or as comparisons among averages in the data."

Consider a dataset consisting of the volume of beer consumed in Sao Paulo, Brazil. For more information about the data, see https://www.kaggle.com/dongeorge/beer-consumption-sao-paulo. We will work on a cleaned dataset that contains:

- consumed: daily volume of beer consumed in liters
- precip: daily precipitation in (mm)
- max_tmp: daily maximum temperature in C
- weekend: indicator variable for if the day is a weekend.

It is not obvious how the data was collected, but here is a note from the data provider: "The data (sample) were collected in São Paulo, Brazil, in a university area, where there are some parties with groups of students from 18 to 28 years of age (average)."

```
library(readr)
library(dplyr)
beer <- read_csv('http://math.montana.edu/ahoegh/Data/Brazil_cerveja.csv')
```

```
## Parsed with column specification:
## cols(
##   consumed = col_double(),
##   precip = col_double(),
##   max_tmp = col_double(),
##   weekend = col_double()
## )
```

```
beer
```

```
## # A tibble: 365 x 4
##     consumed precip max_tmp weekend
##        <dbl>  <dbl>   <dbl>   <dbl>
## 1      25.5     0     32.5       0
## 2      29.0     0     33.5       0
## 3      30.8     0     29.9       1
## 4      29.8    1.2    28.6       1
## 5      28.9     0     28.3       0
## 6      28.2   12.2    30.5       0
## 7      29.7     0     33.7       0
## 8      28.4   48.6    32.8       0
## 9      24.9    4.4    34         0
## 10     37.9     0     34.2       1
## # ... with 355 more rows
```

One predictor

With a <u>binary predictor</u>, the regression coefficient is the difference between the averages of the two groups.

$$Y_i = B_0 + B_1 I_{day(i) = weekend} + \varepsilon_i$$

B_0: ave beer consumed on a week day.

B_1: difference in beer consumed between a weekend and week day.

```
beer <- beer %>% mutate(weekend = as.factor(weekend))
lm_weekend <- lm(consumed ~ weekend, data = beer)
display(lm_weekend)
          library(arm)
## lm(formula = consumed ~ weekend, data = beer)
##              coef.est coef.se
## (Intercept) 24.00     0.24
## weekend1     4.92     0.44
## ---
## n = 365, k = 2
## residual sd = 3.80, R-Squared = 0.26
```

```
library(ggplot2)
library(gridExtra)
fig1 <- beer %>% ggplot(aes(y = consumed, x = weekend)) +
  geom_jitter(width = .25) + ylab('liters of cerveja consumed') +
  xlab('weekend (1 indicates weekend)') + ggtitle('Cerveja consumed vs. weekend')

fig2 <- beer %>% ggplot(aes(y = consumed, x = weekend)) +
  geom_jitter(width = .25) + ylab('liters of cerveja consumed') +
  xlab('weekend (1 indicates weekend)') + ggtitle('Cerveja consumed vs. weekend') +
  ylim(0, max(beer$consumed)+1)
grid.arrange(fig1, fig2)
```
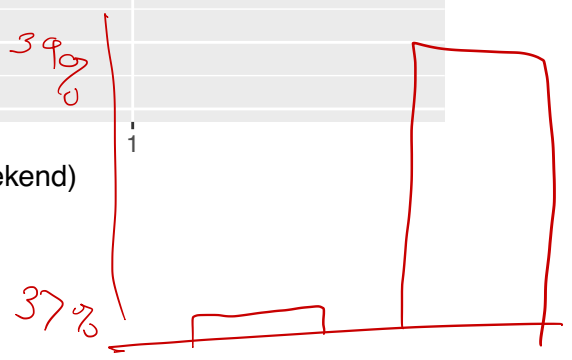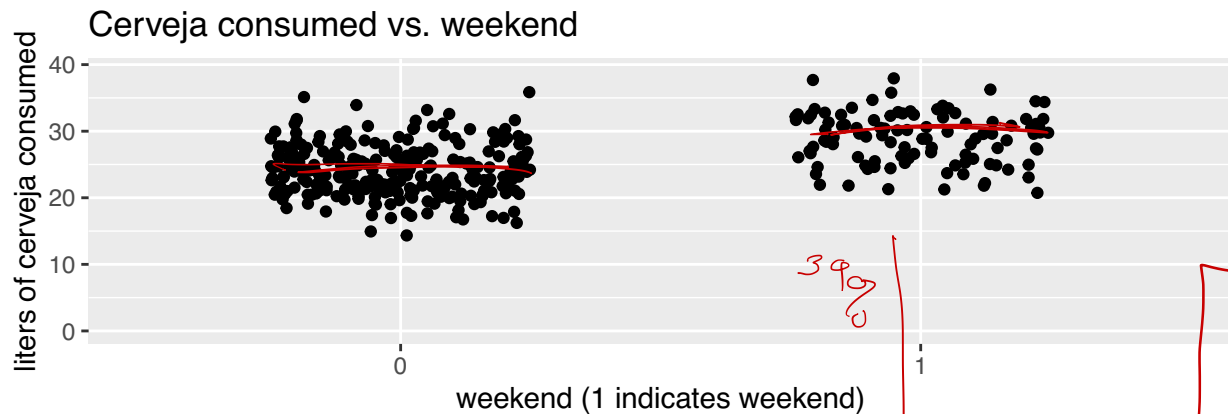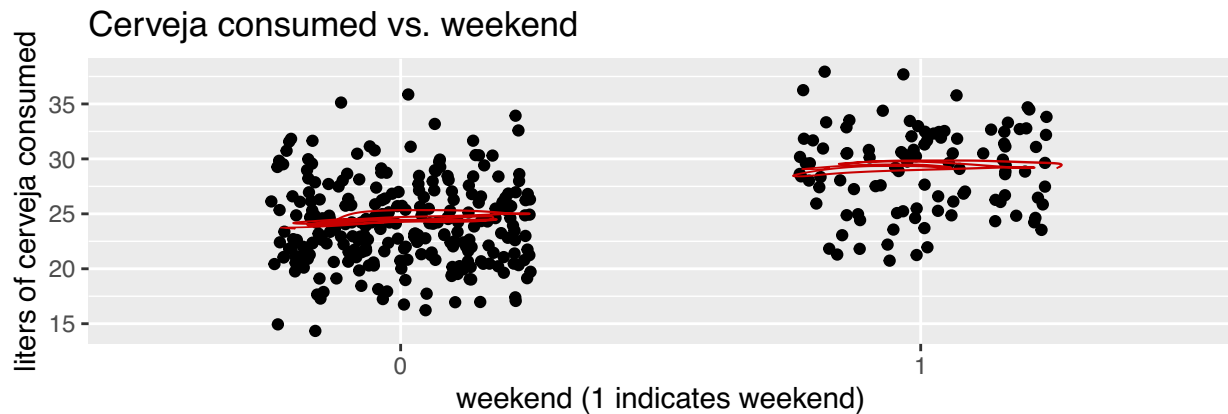

Which graph is preferable? Why?

Regression with a categorical predictor → ANOVA.

With a continuous predictor, say maximum temperature,

$$Y_i = B_0 + B_1 X_{temp,i} + \varepsilon_i$$

Y_i = liters of beer consumed on the ith day

$X_{temp,i}$ is the maximum temperature on the ith day

3

```
lm_temp <- lm(consumed ~ max_tmp, data = beer)
display(lm_temp)

## lm(formula = consumed ~ max_tmp, data = beer)
##             coef.est coef.se
## (Intercept) 7.97     1.10
## max_tmp     0.65     0.04
## ---
## n = 365, k = 2
## residual sd = 3.38, R-Squared = 0.41
```

Interpret the coefficients in this model.

- β_0 : <u>Estimated average</u> amount of beer consumed on a day where the maximum temperature is 0°C. The estimated average is roughly $(6, 10)$ liters

- β_1 : The ^{estimated} average increase of beer consumed for a 1° change in maximum temperature. The predicted increase in the amount of beer consumed for a 1° increase in temp. $(.57, .73)$

```
beer %>% ggplot(aes(y = consumed, x = max_tmp)) +
  geom_jitter(width = .25, alpha = .5) + ylab('liters of cerveja consumed') +
  xlab('Maximum Temperature (Celsius)') + ggtitle('Cerveja consumed vs. Maximum Temperature') +
  ylim(0, max(beer$consumed)+1) + geom_smooth(method ='lm')
```


Cerveja consumed vs. Maximum Temperature
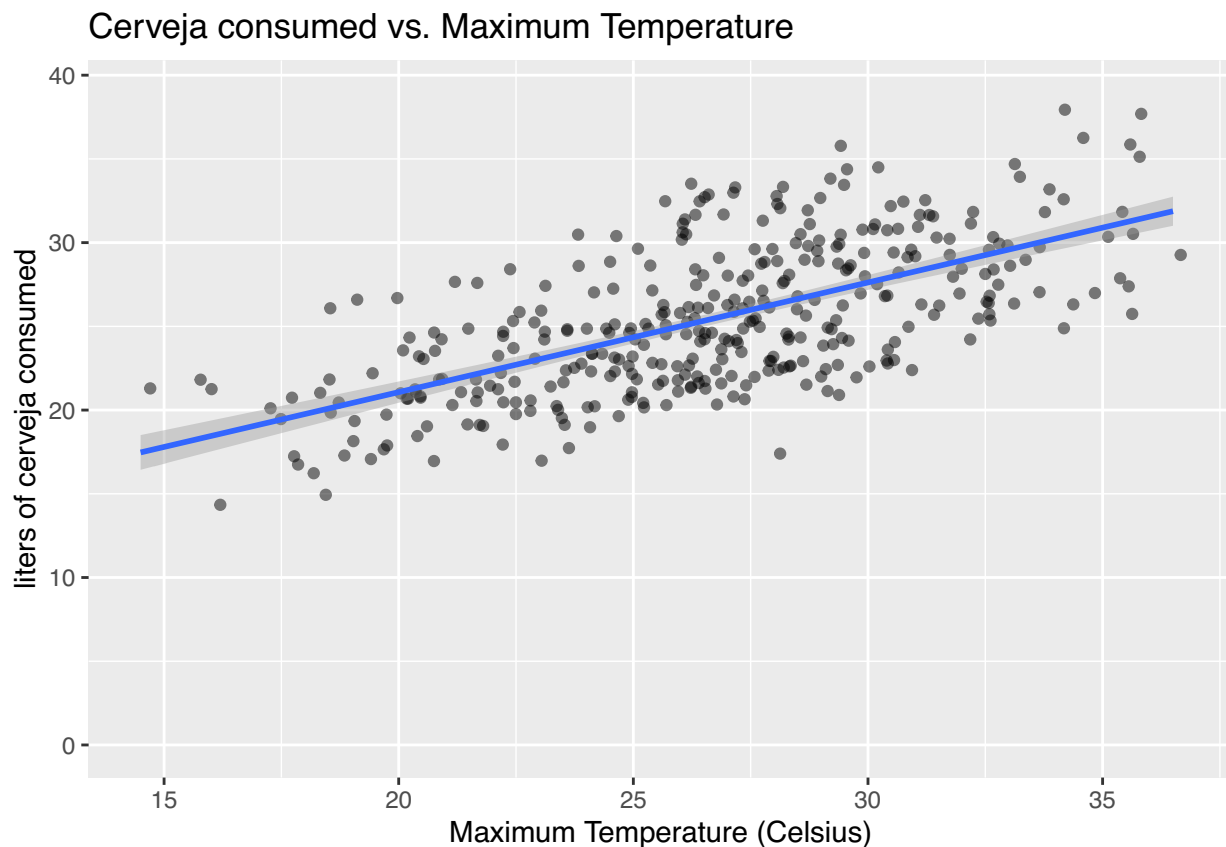
Multiple Regression

As more variables are introduced in a linear model, the intepretation becomes more complicated. Specifically, for any coefficient, the intepretation is

"WITH ALL THE OTHER PREDICTORS HELD CONSTANT"

For instance, write out a model for the mean beer consumed as a function of maximum temperature and whether the day is a weekend. Then interpret each coefficient.

$$Y_i = B_0 + B_1 I_{day(i) = Weekend} + B_2 X_{temp,i} + \varepsilon_i$$

B_0: the average amount of beer consumed on a week day, on a day where the maximum temperature is zero.

Feb-12

B_1: the average increase on beer consumed for a weekend versus a weekday, holding temperature constant.

B_2: the average increase of beer consumed for a one degree increase in temperature, holding day of week constant.

5

```
lm_multi <- lm(consumed ~ max_tmp + weekend, data = beer)
display(lm_multi)
```

```
## lm(formula = consumed ~ max_tmp + weekend, data = beer)
##             coef.est coef.se
## (Intercept) 5.92     0.80
## max_tmp     0.68     0.03
## weekend1    5.18     0.28
## ---
## n = 365, k = 3
## residual sd = 2.43, R-Squared = 0.70
```

(handwritten annotations in red: "(.62 , .74)", "coot", "coef88", ".001", ".0001")

```
beer %>% ggplot(aes(y = consumed, x = max_tmp, color = weekend)) +
  geom_jitter(width = .25, alpha = .5) + ylab('liters of cerveja consumed') +
  xlab('Maximum Temperature (Celsius)') + ggtitle('Cerveja consumed vs. Maximum Temperature') +
  ylim(0, max(beer$consumed)+1) + geom_smooth(method ='lm')
```


Cerveja consumed vs. Maximum Temperature
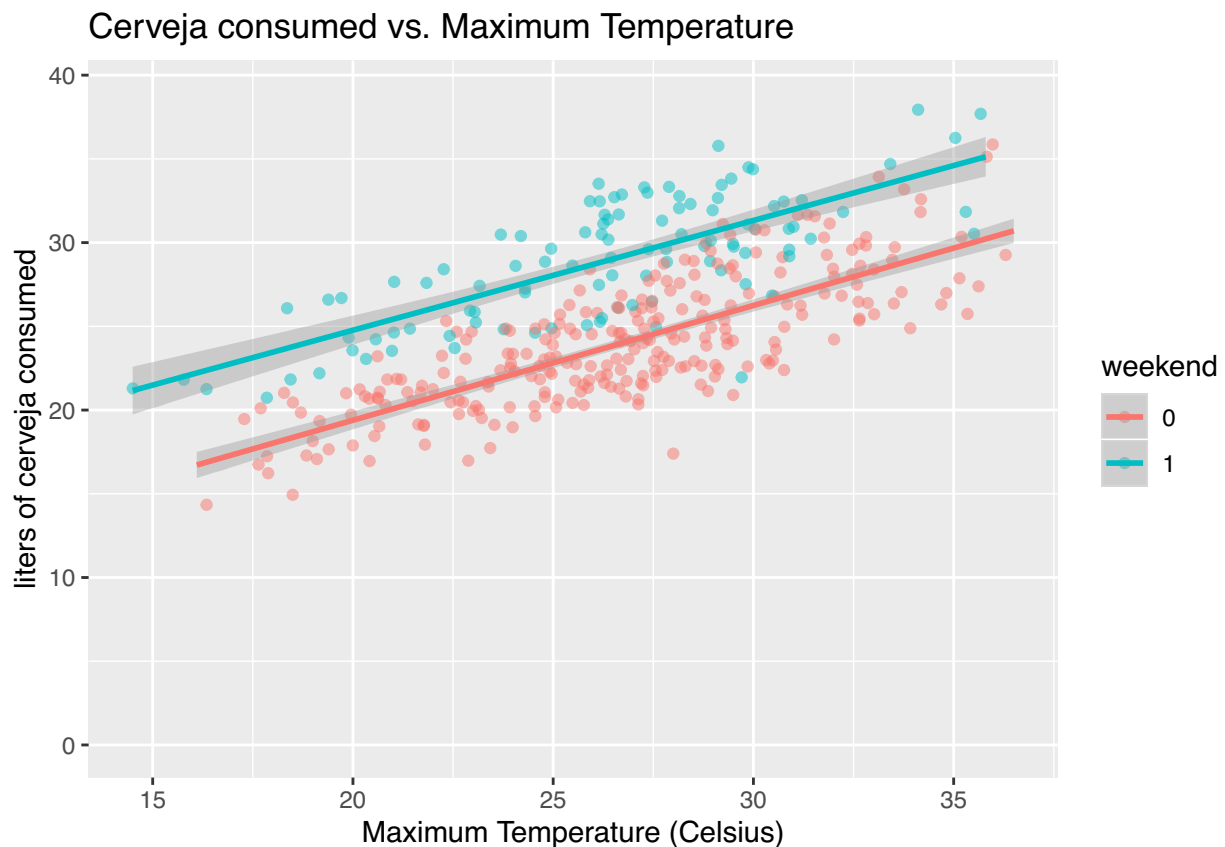
Counterfactual and Predictive Interpretations

There are two ways to interpret regression coefficients.

1. The *predictive interpretation* considers how the outcome variable differs, on average, when comparing two groups of units that differ by 1 in the relevant predictor while being identical in all the other groups.

2. The *counterfactual (causal) interpretation* is expressed in terms of change within individuals, rather than comparisons between individuals. Here, the coefficient is the expected change in the outcome *caused by* adding 1 to the relevant predictor, while leaving all the other predictors in the model unchanged.

Without randomized data (randomly assigned treatment),
          BE CAREFUL WITH ANY CAUSAL LANGUAGE!

Interactions

Now consider an interaction model and interpret the coefficients:

$$consumed_i = \beta_0 + \beta_1 I_{day(i)=weekend} + \beta_2 X_{temp,i} + \beta_3 X_{temp,i} I_{day(i)=weekend} + \varepsilon_i$$

- β_0 : The average amount of beer consumed on a week day where the maximum temperature is $0^\circ C$.

- β_1 : The average amount (increase) of beer consumed on a week day with a maximum temp of $0^\circ C$ vs. a weekend day with a maximum temp of $0^\circ C$

- β_2 : The (estimated/predicted) average increase of beer consumed for a one degree change in temp for a week day.

- β_3 : The difference in slope for temperature, comparing weekday and weekends.
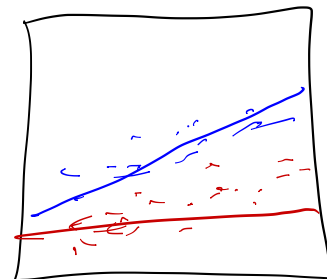
These models (particularly when one of the variables is binary), can be written as separate model:

$$\text{weekday: } consumed_i = \beta_0 + \beta_2 X_{temp,i} + \varepsilon_i \tag{1}$$
$$\text{weekend: } consumed_i = \beta_0 + \beta_1 + (\beta_2 + \beta_3) X_{temp,i} + \varepsilon_i \tag{2}$$

```
lm_interact <- lm(consumed ~ max_tmp * weekend, data = beer)
display(lm_interact)
```

```
## lm(formula = consumed ~ max_tmp * weekend, data = beer)
##                   coef.est coef.se
## (Intercept)          5.68     0.95
## max_tmp              0.69     0.04
## weekend1             5.97     1.74
## max_tmp:weekend1    -0.03     0.06
## ---
## n = 365, k = 4
## residual sd = 2.43, R-Squared = 0.70
```

Interactions

- Interactions are important and they influence how the other main effects can be intrepreted.

- Interactions can occur between categorical and/or continuous variables.

- The textbook mentions smoking as it relates to cancer as a variable that would have large interactions.

- Centering the predictor variables makes for simpler interpretations when models include an interaction.

- Interactions is a way to allow the model to be fit differently for subsets of data. This can also be acheieved with hierarchical models (for categorical variables).

8

Statistical Inference

Notation:

- *units:* individual data points are referred to as units. These are rows in a dataset.

  houses, okCupid profiles

- *outcome:* the outcome of interest

- *predictors* these are the variables used to predict the outcome.

- A linear model can, in general, be written as:

$$\underset{\sim}{\tilde{Y}} = X\underline{\beta} + \underline{\varepsilon}$$

matrix

$$\begin{pmatrix} I & X_{temp} \\ | & day=weekend \\ | & \\ \vdots & \\ | & \\ | & \end{pmatrix}$$

$$
\begin{aligned}
y_i &= X_i\underline{\beta} + \varepsilon && (3)\\
&= \beta_1 X_{i1} + ... + \beta_p X_{1p} + \varepsilon_i && (4)
\end{aligned}
$$

where ε_i are i.i.d and $\sim N(0, \sigma^2)$ and X_i is a matrix of predictors.

$$\underline{\beta} = \begin{pmatrix} \beta_0 & \beta_1 & \cdots & \beta_p \end{pmatrix}$$
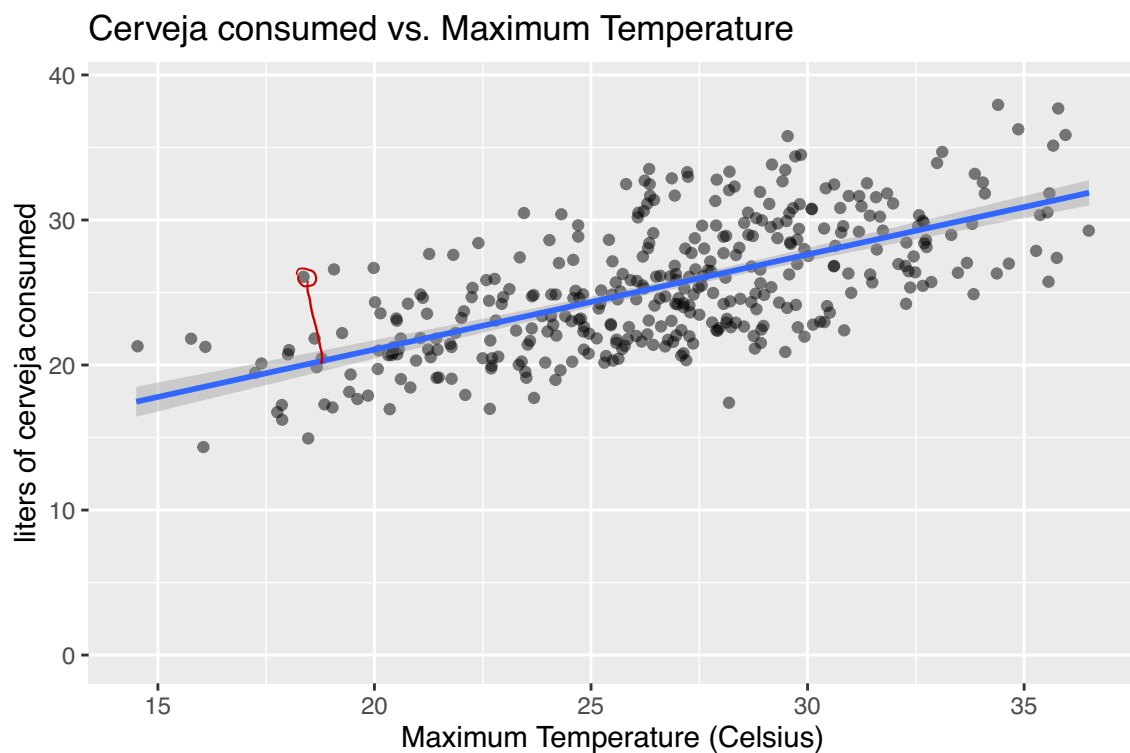
- An equivalent way to write this model is:

$$Y_i \sim N(X_i \underline{\beta}_{-1}, \sigma^2 I)$$

Estimation

With the model $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$, the least squares estimate minimizes the sum of squared errors (or residuals)

$$r_i = (y_i - X_i \hat{\beta})^2$$

Cerveja consumed vs. Maximum Temperature

x^2

$|x|$

The least squares fit is equivalent to the maximum likelihood when the errors are iid $N(0, \sigma^2)$.

The estimates of $\hat{\beta}$ come with standard errors. Using GH's `display()` function only shows the point estimates and standard errors for each coefficient.

Using the CLT (Central Limit Theorem), we can say that the coefficient estimates within 2 standard

errors of $\hat{\beta}$ are consistent with the data.

Variance will tend to be correlated, across predictors, and contained in the covariance matrix

$$\sigma^2 \left(X^T X \right)^{-1}$$

Usually the covariance matrix is not used, except in the case of prediction.

The residual standard deviation $\hat{\sigma}^2$ is defined as:

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{n} \frac{r_i^2}{n - k}},$$

where the

n = # of data points (units)

k = is the number of predictors in the model

The residual standard deviation gives a sense of the model accuracy for an individual unit. For instance, the residual standard deviation for the beer model (with maximum temperature and week day is) 2.43.

Another way to summarize the model fit is using R^2, the fraction of variance explained by the model. The "unexplained variance" is $\hat{\sigma}^2$ and the (adjusted)

$$R^2 = 1 - \frac{\hat{\sigma}^2}{S_y^2}$$

$$0 \leq \hat{\sigma}^2 \leq S_y^2$$

$\hat{\sigma}^2 = 0 \quad R^2 = 1 - \frac{0}{S_y^2} = 1$

$\hat{\sigma}^2 = S_y^2 \quad R^2 = 1 - \frac{S_y^2}{S_y^2} = 0$

Linear Regression Assumptions

GH list the assumption in descreasing order of importance.

1. *Validity:* "Most importantly, the data you are analyzing should map to the research question you are trying to answer... Optimally, this means that the outcome measure should accurately reflect the phenomenon of interest, the model should include all relevant predictors, and the model should generalize to the cases to which it will be applied to. Data used in empirical research rarely meet all (if any) of these criteria precisely. However, keeping these goals in mind can help you be precise about the types of questions you can *and cannot* answer.

2. **Additivity and linearity:** "the most important mathematical assumption of the regression model is that its deterministic component is a linear function of the separate predictors: $y = \beta_1 x_1 + \beta_2 x_2 + ...$ If additivity is violated, it might make sense to transform the data or to add interactions. If linearity is violated, perhaps a predictor should be transformed or included in as a basis function."

3. **Independence of errors:** The simple regression model assumes the errors from the prediction line are independent... more later with multilevel models.

4. **Equal variance of errors:** If the variance of the regression errors are unequal, estimation is more efficiently performed using weighted least squares. However, unequal variance does not affect the predictor $X\underline{\beta}$.

5. **Normality of errors:** "The regression assumption that is generally *least* important is that the errors are normally distributed. GH do *not* recommend diagnostics of the normality of the regression residuals.