STA310 HW1

Olivia Fu

2025-01-20

```
library(tidyverse)
library(tidymodels)
library(knitr)
```

Exercise 1

(a)

The response variable is the number of cricket chirps per minute.

The predictor variable is temperature at the recorded time.

(b)

$$y_i = \mu_i + \epsilon_i = x_i^T \beta + \epsilon_i$$

(c)

Linearity: The relationship between mean of the number of cricket chirps per minute (response Y) and temperature at the recorded time (predictor X) is linear.

Independence: Each observation of the cricket chirps and temperature pair is independent of the others. There is connection between how far any two data points lie from the regression line.

Normality: The number of cricket chirps per minute (response Y) follows a normal distribution at each level/value of temperature (predictor X).

Equal variance: Variance of the number of cricket chirps per minute (response Y) is constant across all values of temperature.

Exercise 2

(a)

The response variable is postnatal depression, specifically patients' depression scores.

The predictor variable is whether or not an estrogen patch is used.

(b) !! ASK

Violation of independence: depression scores were recorded on 6 different visits

Violation of normality: depression score may be skewed as it's less likely to have severely depressed patients?

Exercise 3

(a)

In this new model, we include year as an additional predictor variable alongside track conditions. As observed in the exploratory data analysis before, winning speed varies over time, indicating that year has an impact on winning speed. This model allows us to estimate the difference in winning speeds between fast and non-fast track conditions for a given year. By doing so, it separates the effect of track conditions from the trends over time. When interpreting β_2, it represents the effect of track conditions on winning speed while controlling the time factor. Therefore, it's important to state "holding year constant".

(b)

The equation provides the estimated values of Y_i based on the fitted regression model. This is a regression equation that estimates the function based on the sample data. The error term accounts for random variation not explained by the deterministic component of the model. However, since the regression equation focuses on predicted values, it does not include the potential deviations of actual observations from the predicted values.

Exercise 4

(a)

```r
house <- read.csv("~/Desktop/STA310/sta310-spring25/HW1/kingCountyHouses.csv")
```

```r
lm(price ~ sqft, data = house) |>
  tidy() |>
  kable(digits = 4)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-43580.7431	4402.6897	-9.8987	0
sqft	280.6236	1.9364	144.9204	0

The slope coefficient of model 1 is 280.6236. In this context, when the interior square footage (sqft) increases by 100, we expect the selling price of the house to increase by 28062.36 dollars, on average.

(b)

```r
house <- house |>
  mutate(logprice = log(price))

lm(logprice ~ sqft, data = house) |>
  tidy() |>
  kable(digits = 4)
```

term	estimate	std.error	statistic	p.value
(Intercept)	12.2185	0.0064	1916.8830	0
sqft	0.0004	0.0000	142.2326	0

The slope coefficient of model 2 is 0.0004. In this context, when the interior square footage (sqft) increases by 100, the log of the house price is expected to increase by 0.04, on average.

(c)

Based on model 2, when the interior square footage (sqft) increases by 100, the house price is expected to multiply by a factor of 1.0408 (exp(0.04)), on average.

(d)

```
house <- house |>
  mutate(logsqft = log(sqft))

lm(price ~ logsqft, data = house) |>
  tidy() |>
  kable(digits = 4)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-3451377.1	35169.35	-98.1359	0
logsqft	528647.5	4650.63	113.6722	0

The slope coefficient of model 3 is 528647.5. In this context, a 10% increase in interior square footage (sqft) corresponds to the logsqft to increase by 0.09531 (log(1.1)). Therefore, when sqft increases by 10%, the house price is expected to increase by $528647.5 \times 0.09531 = 50385.39$ dollars, on average.

Exercise 5

(a)

```
college <- read.csv("~/Desktop/STA310/sta310-spring25/HW1/college-data.csv")
```

(b)

(c)

(d)

(e)