HW 23 - Solutions

Problem 1

For $n = 1040$ male college soccer players, the correlation between height and weight is about $r = 0.75$. The sample means for heights and weights are about $\bar{x} = 71$ in and $\bar{y} = 166$ lbs, and the sample standard deviations are about $s_x = 2.5$ in and $s_y = 16$ lbs.

(a) Find the least squares regression line for predicting weight from height. What proportion of the variability in weights is explained by a linear fit on height?

In a SLR model, the estimate for the slope is $\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = r\frac{s_y}{s_x}$ and the estimate for the intercept is $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. Recall that $s_x^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ and $r = \frac{1}{n-1}\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y}$. Thus $\hat{\beta}_1 = (0.75)(16)/2.5 = 4.8$ inches/lb and $\hat{\beta}_0 = 166 - (4.8)(71) = -174.8$ inches.

(b) Find the fitted weight for a 66 inch player and for a 76 inch player. Explain how these fitted values illustrate the regression towards the mean effect in an answer that involves standard deviations relative to the respective means. Hint: You textbook mentions "regression towards mediocrity" but if you google this phrase, you'll find lots of examples and wiki pages on this phenomena!

```
-174.89 + 4.8*66
```

```
## [1] 141.91
```

```
-174.89 + 4.8*76
```

```
## [1] 189.91
```

(c) Use the sample correlation and standard deviation of the weights to find the root mean squared error for the simple regression model. Explain what this number represents in this context.

Please see the class notes from 12/5/22 for a solution to this problem.

Problem 2

Consider the no-intercept linear regression model

$$Y_i \mid X_i = x_i \sim N(\beta x_i, \sigma^2), \quad i = 1, \ldots, n.$$

We should include an intercept in the model even if we believe the mean response when $x = 0$ should be 0, however working with the no-intercept model can help understand the more complicated model since here β is a scalar rather than a vector.

(a) Show that the least squares estimate for β is $\hat{\beta} = \frac{\sum_i x_i Y_i}{\sum_i x_i^2} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, where $\mathbf{X}$ is the $n \times 1$ matrix (vector) of x_i values and $\mathbf{Y}$ is the $n \times 1$ vector of Y_i values.

The least squares estimate for β solves $\min \sum (y_i - \hat{y}_i)^2$ with respect to β. To find this minimizer, we consider

$$\frac{\partial}{\partial \beta} \sum (y_i - \beta x_i)^2 = \sum 2(y_i - \beta x_i)(-x_i) \overset{set}{=} 0$$

which solving for β produces the least squares estimate

$$\hat{\beta}_{LSE} = \sum y_i x_i \sum x_i^2$$

since we can verify this is a minimum by checking

$$\frac{\partial}{\partial \beta} 2 \sum (-x_i)(y_i - \beta x_i) = 2 \sum (-x_i)^2 > 0.$$

(b) Write the joint log-likelihood of (β, σ^2) and explain why the MLE for β is the same as the least squares estimate for β.

$$Lik(\beta, \sigma) = \prod_{i=1}^{n} f(y_i; \beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \beta x_i)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma}\right)^n \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i)^2\}$$

Now we find thee MLE for β by setting the first derivative of the (log) likelihood equal to zero and solving for $\hat{\beta}$:

$$\ln Lik(\beta, \sigma) = const + n(0 - \ln(\sigma)) - \frac{\sum (y_i - \beta x_i)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial \beta} \ln Lik(\beta, \sigma) = \frac{\sum x_i (y_i - \beta x_i)^2}{\sigma^2} \overset{set}{=} 0 \text{ and thus } \hat{\beta}_{MLE} = \frac{\sum y_i x_i}{\sum x_i^2} = \hat{\beta}_{LSE}$$

(c) Find the mean and variance of $\hat{\beta}$.

$$E\left(\hat{\beta}\right) = E\left(\frac{\sum x_i Y_i}{\sum x_i^2}\right) = \frac{\sum x_i E(Y_i)}{\sum x_i^2} = \frac{\sum x_i (\beta x_i)}{\sum x_i^2} = \frac{\beta \sum x_i^2}{\sum x_i^2} = \beta$$

and

$$
\begin{aligned}
Var(\hat{\beta}) &= Var\left(\frac{\sum x_i Y_i}{\sum x_i^2}\right) \\
&= \left(\frac{1}{\sum x_i^2}\right)^2 Var\left(\sum x_i Y_i\right) \\
&= \left(\frac{1}{\sum x_i^2}\right)^2 \sum_i \sum_j x_i x_j Cov(Y_i, Y_j) \\
&= \left(\frac{1}{\sum x_i^2}\right)^2 \sum_i \sum_j x_i x_j Cov(\epsilon_i, \epsilon_j) \\
&= \left(\frac{1}{\sum x_i^2}\right)^2 \sum_i \sum_{j=i} x_i x_j Cov(\epsilon_i, \epsilon_j) \\
&= \left(\frac{1}{\sum x_i^2}\right)^2 \sum_{i=1}^{n} x_i x_i Var(\epsilon_j) \\
&= \left(\frac{\sum x_i^2}{(\sum x_i^2)^2}\right) \sigma^2 \\
&= \frac{\sigma}{\sum x_i^2}
\end{aligned}
$$

Also, recall from our class notes that we expect $Var(Y) = \sigma^2 (X^T X)^{-1}$ and here, $X^T = \begin{pmatrix} x_1 x_2 \dots x_n \end{pmatrix}$ so $X^T X = \sum x_i^2$.

Problem 3

A simple exponential decay models says that the concentration, $C_{(t)}$ of a pesticide remaining after time t is $C_{(t)} = C_0 e^{-\gamma t}$ for $t > 0$ where C_0 is the initial concentration and γ is a constant that determines the rate of decay.

(a) Show how taking the natural log of both sides of the equation above results in a linear model for $Y = \log(C_{(t)})$ on t. What are the slope and intercept?

$$\ln\left(C_{(t)}\right) = \ln\left(C_0 e^{-\gamma t}\right) = ln(C_0) - \gamma t$$

is the equation for a line where the intercept is $ln(C_0)$ and the slope is $-\gamma$.

(b) If you have data on concentrations at n different times, t_i, you could estimate γ by fitting a SLR of Y_i on t_i. This implicitly assumes an additive error term ϵ_i that is approximately normally distributed. Write out the implied model for $C_{(t)}$ and describe how error enters this model.

If we observe t_i for $i = 1, \ldots, n$, and regress these observations on $Y = \ln C_{(t)}$ then we are implying the model for Y is:

$$Y_i = \ln\left(C_0 e^{-\gamma t}\right) = ln(C_0) - \gamma t_i + \epsilon_i, \quad \text{where } \epsilon_i \overset{IID}{\sim} Normal.$$

That is, $C_{(t_i)} = C_0 e^{-\gamma t_i} e^{\epsilon_i}$ where $\epsilon_i \overset{IID}{\sim} Normal$. Hence the error enters this model as a multiplicative factor, rather than an additive one.