Assignment 1 Solutions

Question 1

Since $y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, the probability density function for y_i is

$$f(y_i; \beta_0, \beta_1, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$

and the corresponding likelihood function is

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} f(y_i; \beta_0, \beta_1, \sigma) = \frac{1}{(2\pi\sigma^2)^{-n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}.$$

The log-likelihood function is then given by

$$l(\beta_0, \beta_1, \sigma) = \log(L(\beta_0, \beta_1, \sigma)) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

The partial derivatives of this function with respect to β_0 and β_1 are given by:

$$\frac{\partial l}{\partial \beta_0} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$$

and

$$\frac{\partial l}{\partial \beta_1} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i).$$

Solving $\frac{\partial l}{\partial \beta_0} = 0$ and $\frac{\partial l}{\partial \beta_1} = 0$ yield the exact same set of equations as the least squares derivation from class and so I do not reproduce the steps here (but the students should have). As such the resulting ML estimates are exactly the same as the LS estimates:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ and } \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})}.$$

Question 2

The partial derivative of $l(\beta_0, \beta_1, \sigma)$ is the following:

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Solving $\frac{\partial l}{\partial \sigma} = 0$ and substituting the estimates for β_0 and β_1 yields the maximum likelihood estimate of σ:

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}.$$

We can see that this differs slightly from the least squares estimate of σ which has a denominator of $n-2$ rather than n. For large sample sizes the difference is immaterial, but for small sample sizes $\hat{\sigma}_{LSE} \neq \hat{\sigma}_{MLE}$. It turns out that the LSE is to be preferred over the MLE because $\tilde{\sigma}^2_{LSE}$ is an unbiased estimator and $\tilde{\sigma}^2_{MLE}$ is not.

Question 3

a. $\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_{xx}} = \frac{1}{s_{xx}} \left(\sum_{i=1}^{n} x_i - n\bar{x}\right) = \frac{1}{s_{xx}} (n\bar{x} - n\bar{x}) = 0$

b. $\sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})x_i}{s_{xx}} = \frac{1}{s_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{s_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})^2 = 1$

c. $\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{s_{xx}^2} = \frac{1}{s_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{s_{xx}^2} s_{xx} = \frac{1}{s_{xx}}$

Question 4

$$\mathrm{E}\left[\hat{\beta}_1\right] = \mathrm{E}\left[\sum_{i=1}^{n} c_i y_i\right] = \sum_{i=1}^{n} c_i \mathrm{E}\left[y_i\right] = \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i = \beta_0(0) + \beta_1(1) = \beta_1$$

$$\mathrm{Var}\left[\hat{\beta}_1\right] = \mathrm{Var}\left[\sum_{i=1}^{n} c_i y_i\right] = \sum_{i=1}^{n} c_i^2 \mathrm{Var}\left[y_i\right] = \sum_{i=1}^{n} c_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{s_{xx}}$$

Question 5

$$\mathrm{E}\left[\hat{\beta}_0\right] = \mathrm{E}\left[\bar{y} - \hat{\beta}_1 \bar{x}\right] = \mathrm{E}\left[\bar{y}\right] - \mathrm{E}\left[\hat{\beta}_1\right] \bar{x} = (\beta_0 + \beta_1 \bar{x}) - \beta_1 \bar{x} = \beta_0$$

$$\mathrm{Var}\left[\hat{\beta}_0\right] = \mathrm{Var}\left[\bar{y} - \hat{\beta}_1 \bar{x}\right] = \mathrm{Var}\left[\bar{y}\right] + \bar{x}^2 \mathrm{Var}\left[\hat{\beta}_1\right] = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{s_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)$$

Question 6

$$\mathrm{E}\left[\hat{\mu}_0\right] = \mathrm{E}\left[\hat{\beta}_0 + \hat{\beta}_1 x_0\right] = \mathrm{E}\left[\hat{\beta}_0\right] + \mathrm{E}\left[\hat{\beta}_1\right] x_0 = \beta_0 + \beta_1 x_0 = \mu_0$$

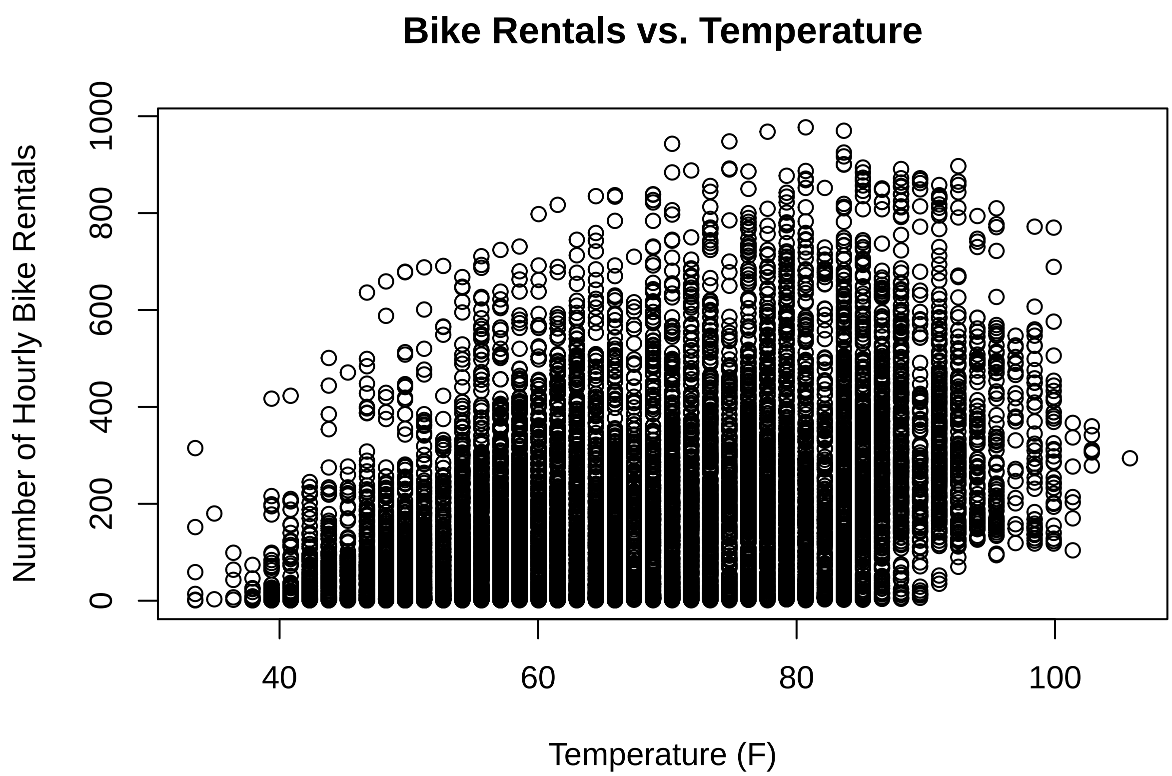
$$\mathrm{Var}\left[\hat{\mu}_0\right] = \mathrm{Var}\left[\hat{\beta}_0 + \hat{\beta}_1 x_0\right] = \mathrm{Var}\left[\bar{y} - \hat{\beta}_1 + \hat{\beta}_1 x_0\right] = \mathrm{Var}\left[\bar{y} + \hat{\beta}_1(x_0 - \bar{x})\right] = \mathrm{Var}\left[\bar{y}\right] + \mathrm{Var}\left[\hat{\beta}_1\right](x_0 - \bar{x})^2$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{s_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)$$

Question 7

(a)

```r
y <- bike$count
x <- bike$temp
plot(x, y, xlab = "Temperature (F)", ylab = "Number of Hourly Bike Rentals",
    main = "Bike Rentals vs. Temperature")
```




```
cor(x, y)
```

```
## [1] 0.3944536
```

The plot above indicates a relatively weak positive linear relationship between the number of bike rentals and the outside temperature. This conclusion is corroborated by the correlation coefficient of 0.3945. Thus as the outside temperature increases, so also does the number of bike rentals, but this relationship is relatively weak.

(b)

```
beta1_hat <- (cor(x, y) * sd(y))/sd(x)
print(beta1_hat)
```

```
## [1] 5.094745
```

```
beta0_hat <- mean(y) - beta1_hat * mean(x)
print(beta0_hat)
```

```
## [1] -156.9856
```

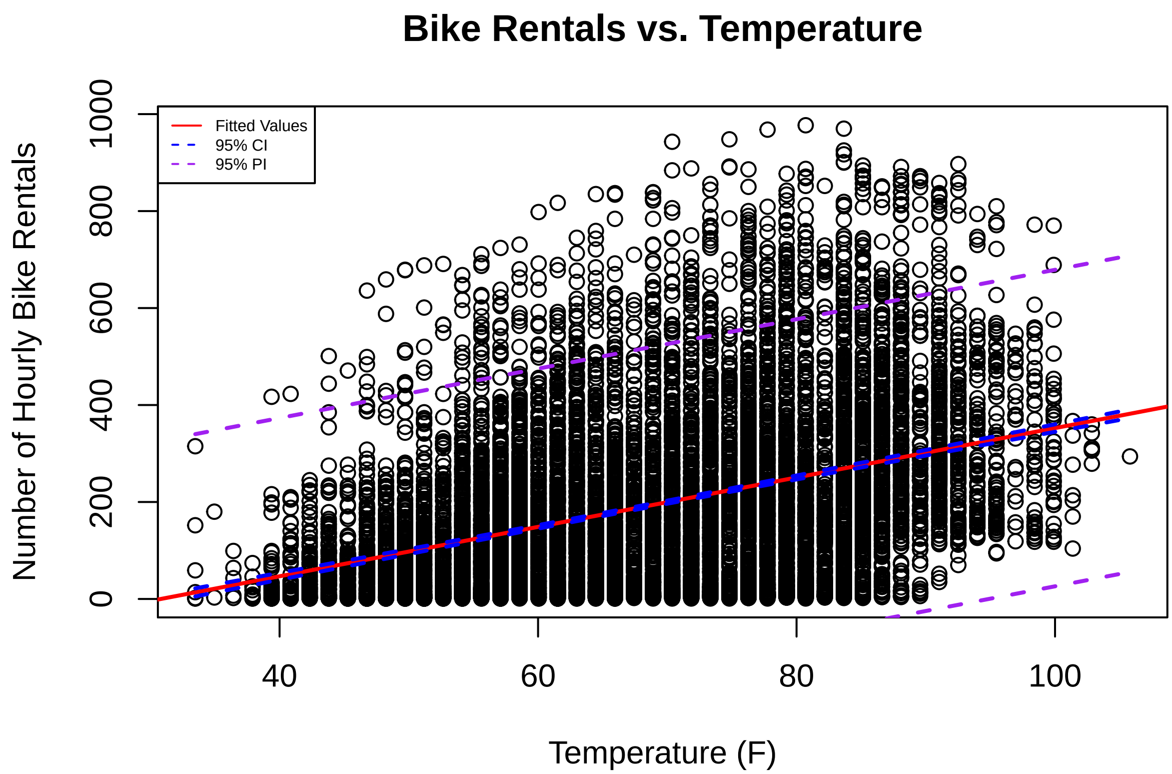
Thus the equation of the line-of-best-fit is $\hat{\mu} = -156.9856 + 5.0947x$.

(c)

- The estimate $\hat{\beta}_0 = -156.9856$ suggests that in hours when the temperature is 0 degrees Fahrenheit we would expect -156.9856 bikes to be rented. This obviously doesn't make any practical sense – but that's because 0 degrees Fahrenheit was not a value observed in the data and so the model doesn't know how to adequately deal with it.
- The estimate $\hat{\beta}_1 = 5.0947$ suggests that for every one-degree increase in outside temperature, we expect the average hourly number of bike rentals to increase by 5.0947.

(d)

```r
plot(x, y, xlab = "Temperature (F)", ylab = "Number of Hourly Bike Rentals",
    main = "Bike Rentals vs. Temperature")
m <- lm(y ~ x)
abline(reg = m, lwd = 2, col = "red")
conf <- data.frame(x, predict(m, interval = "confidence", level = 0.95))
conf <- conf[order(conf[, 1]), ]
lines(conf[, 1], conf[, 3], col = "blue", lwd = 2, lty = 2)
lines(conf[, 1], conf[, 4], col = "blue", lwd = 2, lty = 2)
pred <- data.frame(x, predict(m, newdata = data.frame(x), interval = "prediction",
    level = 0.95))
pred <- pred[order(pred[, 1]), ]
lines(pred[, 1], pred[, 3], col = "purple", lwd = 2, lty = 2)
lines(pred[, 1], pred[, 4], col = "purple", lwd = 2, lty = 2)
legend("topleft", legend = c("Fitted Values", "95% CI", "95% PI"), col = c("red",
    "blue", "purple"), lty = c(1, 2, 2), cex = 0.5)
```



(e)

```r
predict(m, newdata = data.frame(x = 70), interval = "prediction", level = 0.95)
```

```
##        fit       lwr      upr
## 1 199.6465 -126.6687 525.9617
```

As we can see from the output above, in hours when it is 70 degrees Fahrenheit outside, we predict that 199.6465 $\approx$ 200 bikes will be rented. The 95% prediction interval suggests that although we don't know the true number of rentals exactly, we are 95% confident that it is somewhere between -127 and 526.

(f)

```r
m <- lm(y ~ x)
summary(m)
```

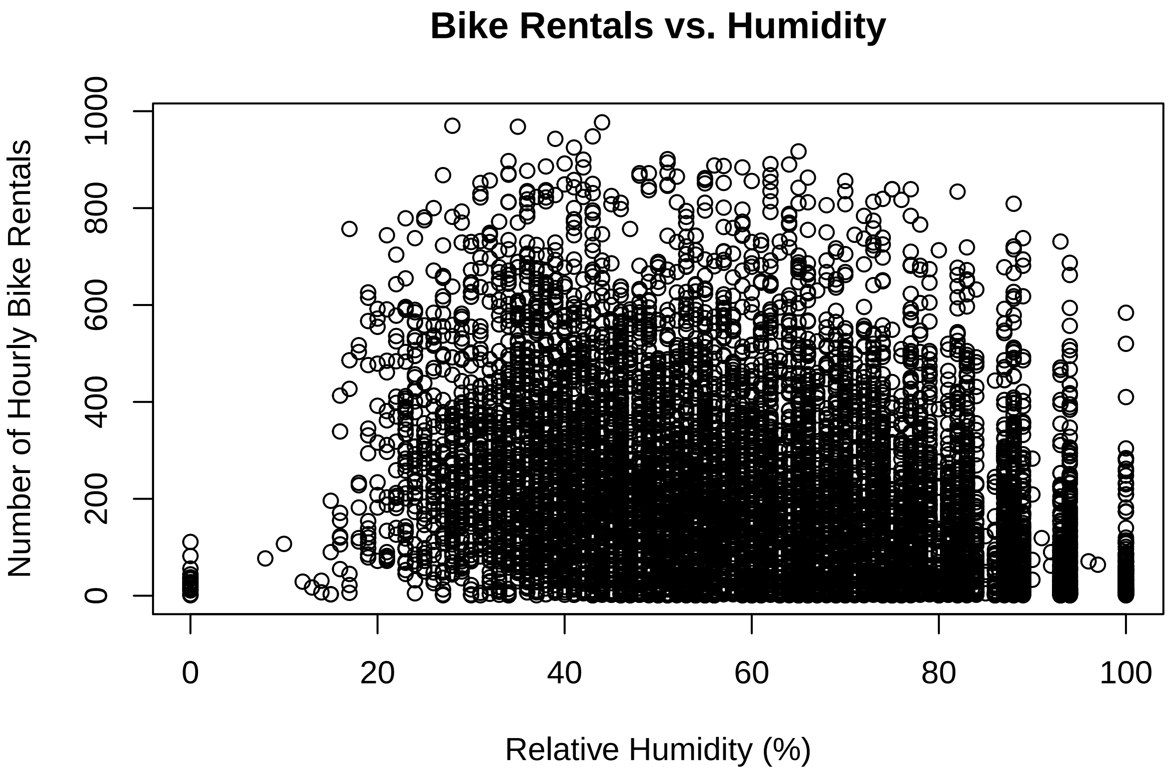
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -293.32 -112.36  -33.36   78.98  741.44
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -156.9856     7.9451  -19.76   <2e-16 ***
## x              5.0947     0.1138   44.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 166.5 on 10884 degrees of freedom
## Multiple R-squared:  0.1556, Adjusted R-squared:  0.1555
## F-statistic:  2006 on 1 and 10884 DF,  p-value: < 2.2e-16
```

We can see from the output that the test statistic associated with $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_A : \beta_1 \neq 0$ is $t = 44.78$ and the associated p-value is $2\mathrm{P}(t_{10884} \geq |t|) < 2.2 \times 10^{-16}$ which is approximately equal to 0. Although the output does not tell us exactly what the p-value is, we do know that it is less than $\alpha = 0.05$ and basically any other typical significance level. As such, we reject $\mathrm{H}_0 : \beta_1 = 0$ and conlcude that bike rentals are significantly influenced by the outside temperature.

Question 8

(a)

```
y <- bike$count
x <- bike$humidity
plot(x, y, xlab = "Relative Humidity (%)", ylab = "Number of Hourly Bike Rentals",
    main = "Bike Rentals vs. Humidity")
```




```
cor(x, y)
```

```
## [1] -0.3173715
```

The plot above indicates a relatively weak negative linear relationship between the number of bike rentals and the relative humidity. This conclusion is corroborated by the correlation coefficient of -0.3174. Thus as humidity increases, the number of bike rentals tends to decrease, but this relationship is relatively week.

(b)

```
beta1_hat <- (cor(x, y) * sd(y))/sd(x)
print(beta1_hat)
```

```
## [1] -2.987269
```

```
beta0_hat <- mean(y) - beta1_hat * mean(x)
print(beta0_hat)
```

```
## [1] 376.4456
```

Thus the equation of the line-of-best-fit is $\hat{\mu} = 376.4456 - 2.9873x$.

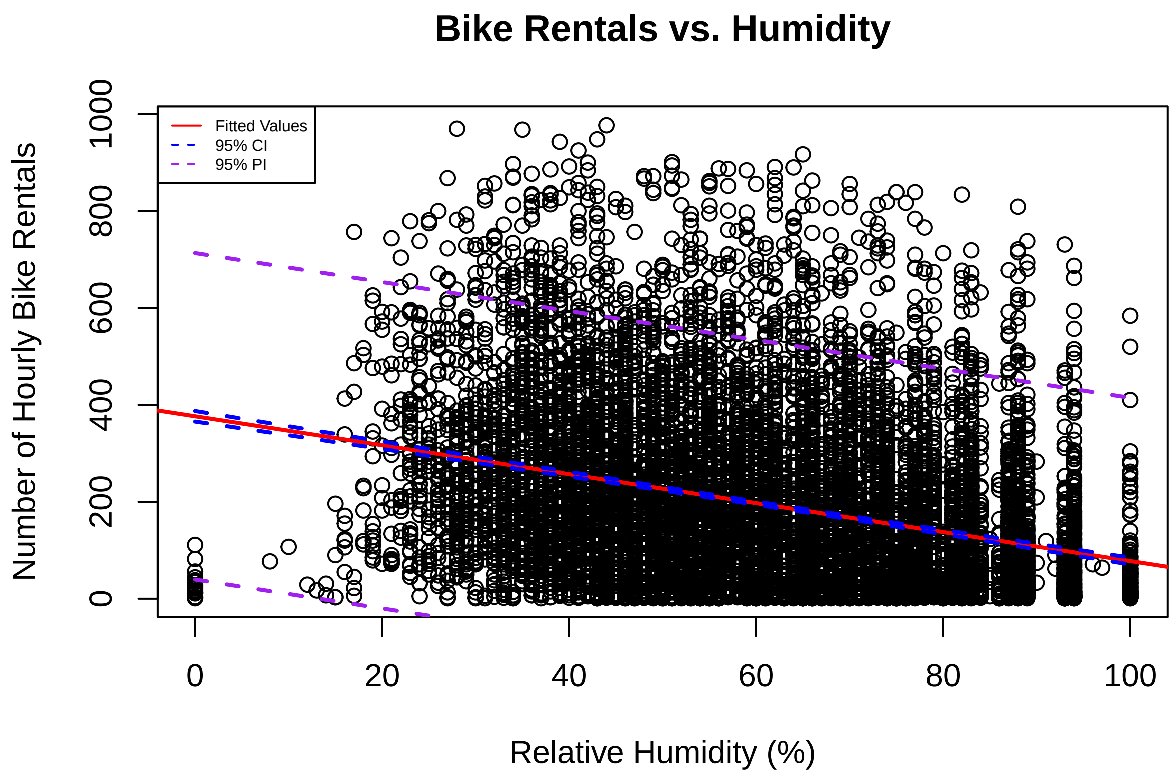
(c)

- The estimate $\hat{\beta}_0 = 376.4456$ suggests that in hours when the humidity is 0% we would expect 376.4456 bikes to be rented.
- The estimate $\hat{\beta}_1 = -2.9873$ suggests that for every 1% increase in humidity, we expect the average hourly number of bike rentals to decrease by 2.9873.

(d)

```
plot(x, y, xlab = "Relative Humidity (%)", ylab = "Number of Hourly Bike Rentals",
    main = "Bike Rentals vs. Humidity")
m <- lm(y ~ x)
```

```r
abline(reg = m, lwd = 2, col = "red")
conf <- data.frame(x, predict(m, interval = "confidence", level = 0.95))
conf <- conf[order(conf[, 1]), ]
lines(conf[, 1], conf[, 3], col = "blue", lwd = 2, lty = 2)
lines(conf[, 1], conf[, 4], col = "blue", lwd = 2, lty = 2)
pred <- data.frame(x, predict(m, newdata = data.frame(x), interval = "prediction",
    level = 0.95))
pred <- pred[order(pred[, 1]), ]
lines(pred[, 1], pred[, 3], col = "purple", lwd = 2, lty = 2)
lines(pred[, 1], pred[, 4], col = "purple", lwd = 2, lty = 2)
legend("topleft", legend = c("Fitted Values", "95% CI", "95% PI"), col = c("red",
    "blue", "purple"), lty = c(1, 2, 2), cex = 0.5)
```



(e)

```r
predict(m, newdata = data.frame(x = 40), interval = "prediction", level = 0.95)
```

```
##        fit       lwr      upr
## 1 256.9549 -79.81519 593.7249
```

As we can see from the output above, in hours when the humidity is 40%, we predict that $256.9549 \approx 257$ bikes will be rented. The 95% prediction interval suggests that although we don't know the true number of rentals exactly, we are 95% confident that it is somewhere between -80 and 594.

(f)

```r
m <- lm(y ~ x)
summary(m)
```

```
##
## Call:
## lm(formula = y ~ x)
```

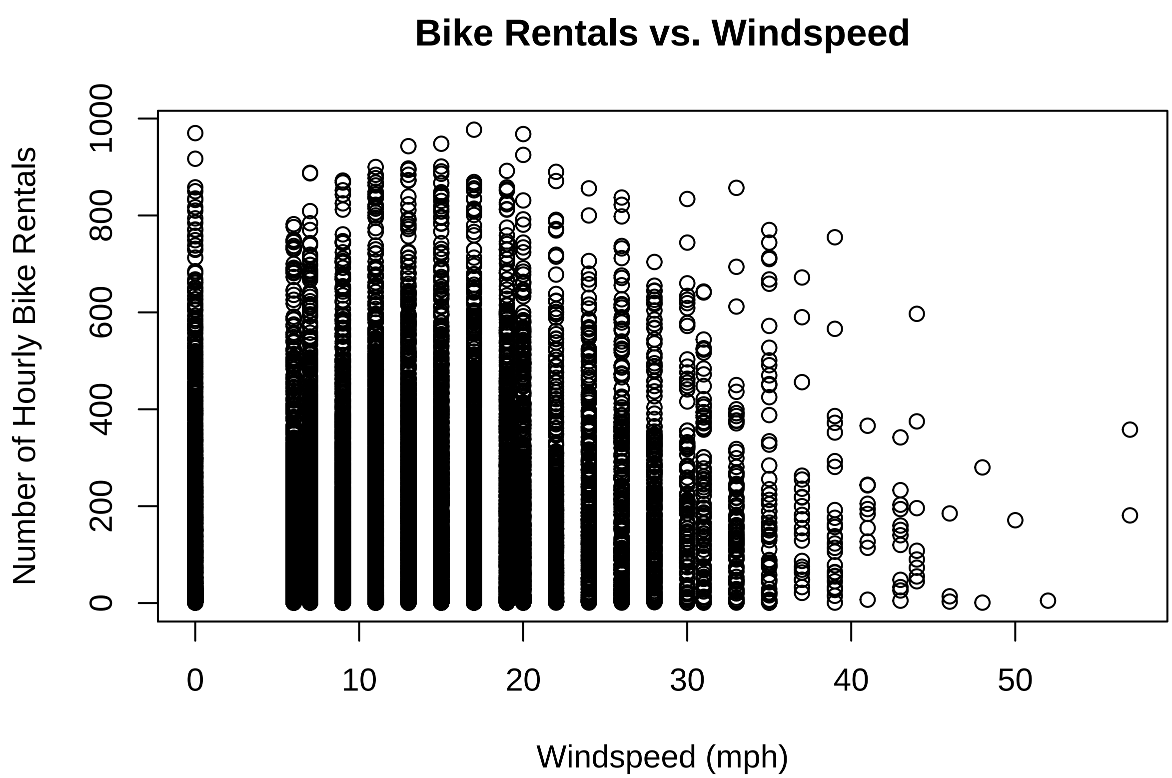
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -375.45 -120.49  -41.86   82.15  734.73
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 376.44561    5.54494   67.89   <2e-16 ***
## x            -2.98727    0.08556  -34.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 171.8 on 10884 degrees of freedom
## Multiple R-squared:  0.1007, Adjusted R-squared:  0.1006
## F-statistic:  1219 on 1 and 10884 DF,  p-value: < 2.2e-16
```

We can see from the output that the test statistic associated with $\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_A : \beta_1 \neq 0$ is $t = -34.91$ and the associated p-value is $2\mathrm{P}(t_{10884} \geq |t|) < 2.2 \times 10^{-16}$ which is approximately equal to 0. Although the output does not tell us exactly what the p-value is, we do know that it is less than $\alpha = 0.05$ and basically any other typical significance level. As such, we reject $\mathrm{H}_0 : \beta_1 = 0$ and conlcude that bike rentals are significantly influenced by the relative humidity.

Question 9

(a)

```r
y <- bike$count
x <- bike$windspeed
plot(x, y, xlab = "Windspeed (mph)", ylab = "Number of Hourly Bike Rentals",
    main = "Bike Rentals vs. Windspeed")
```



```r
cor(x, y)
```

```
## [1] 0.1013695
```

The plots above indicate a very weak positive linear relationship between the number of bike rentals and windspeed. This conclusion is corroborated by the correlation coefficient of 0.1014. Thus as the windspeed increases, the number of bike rentals also tends to increase, but this relationship is very weak.

(b)

```r
beta1_hat <- (cor(x, y) * sd(y))/sd(x)
print(beta1_hat)
```

```
## [1] 2.249058
```

```r
beta0_hat <- mean(y) - beta1_hat * mean(x)
print(beta0_hat)
```

```
## [1] 162.7876
```

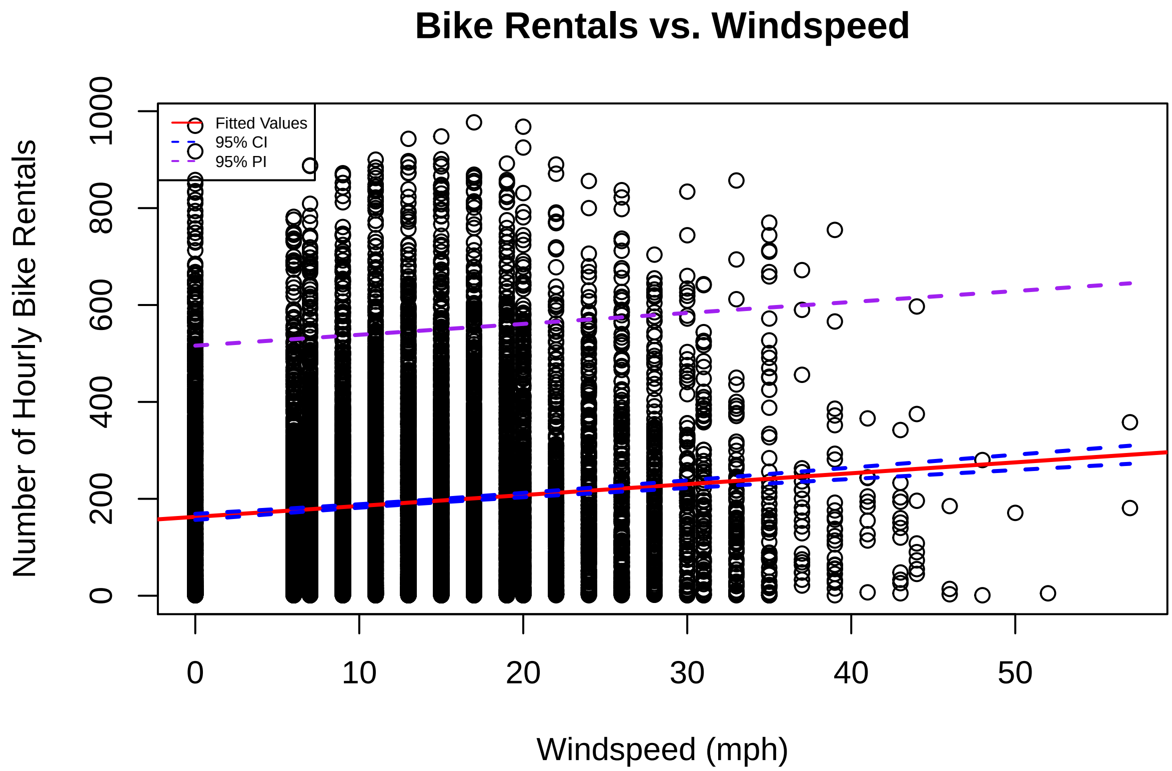
Thus the equation of the line-of-best-fit is $\hat{\mu} = 162.7876 + 2.2491x$.

(c)

- The estimate $\hat{\beta}_0 = 162.7876$ suggests that in hours when the windspeed is 0 miles per hour we would expect 162.7876 bikes to be rented.
- The estimate $\hat{\beta}_1 = 2.2491$ suggests that for every mile per hour increase in windspeed, we expect the average hourly number of bike rentals to increase by 2.249058.

(d)

```r
plot(x, y, xlab = "Windspeed (mph)", ylab = "Number of Hourly Bike Rentals",
    main = "Bike Rentals vs. Windspeed")
m <- lm(y ~ x)
abline(reg = m, lwd = 2, col = "red")
conf <- data.frame(x, predict(m, interval = "confidence", level = 0.95))
conf <- conf[order(conf[, 1]), ]
lines(conf[, 1], conf[, 3], col = "blue", lwd = 2, lty = 2)
lines(conf[, 1], conf[, 4], col = "blue", lwd = 2, lty = 2)
pred <- data.frame(x, predict(m, newdata = data.frame(x), interval = "prediction",
    level = 0.95))
pred <- pred[order(pred[, 1]), ]
lines(pred[, 1], pred[, 3], col = "purple", lwd = 2, lty = 2)
lines(pred[, 1], pred[, 4], col = "purple", lwd = 2, lty = 2)
legend("topleft", legend = c("Fitted Values", "95% CI", "95% PI"), col = c("red",
    "blue", "purple"), lty = c(1, 2, 2), cex = 0.5)
```



(e)

```r
predict(m, newdata = data.frame(x = 10), interval = "prediction", level = 0.95)
```

```
##        fit       lwr      upr
## 1 185.2781 -168.0033 538.5595
```

As we can see from the output above, in hours when the windspeed is 10mph, we predict that $185.2781 \approx 185$ bikes will be rented. The 95% prediction interval suggests that although we don't know the true number of rentals exactly, we are 95% confident that it is somewhere between -168 and 539.

(f)

```r
m <- lm(y ~ x)
summary(m)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -274.74 -145.29  -48.53   92.48  807.21
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 162.7876     3.2120   50.68   <2e-16 ***
## x             2.2491     0.2116   10.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 180.2 on 10884 degrees of freedom
```

```
## Multiple R-squared:  0.01028,    Adjusted R-squared:  0.01018
## F-statistic:   113 on 1 and 10884 DF,  p-value: < 2.2e-16
```

We can see from the output that the test statistic associated with $\text{H}_0 : \beta_1 = 0$ vs. $\text{H}_A : \beta_1 \neq 0$ is $t = 10.63$ and the associated p-value is $2\text{P}(t_{10884} \geq |t|) < 2.2 \times 10^{-16}$ which is approximately equal to 0. Although the output does not tell us exactly what the p-value is, we do know that it is less than $\alpha = 0.05$ and basically any other typical significance level. As such, we reject $\text{H}_0 : \beta_1 = 0$ and conlcude that bike rentals are significantly influenced by the windspeed.

Question 10

The strength of a linear relationship is signified by the the magnitude of the corresponding correlation coefficient. Ranking the three correlation coefficient magnitudes in increasing order indicates the the weakest linear relationship is between bike rentals and windspeed, the strongest linear relationship is between bike rentals and temperature, and the linear relationship between bike rentals and humidity is in the middle.