HW 1

September 10, 2021

1. Consider $m \neq v$ where v is the median. If $m < v$ then:

$$
\begin{aligned}
R(m) - R(v) &= \int_{-\infty}^{\infty} (|y-m| - |y-v|)p(y)dy \\
&= \int_{-\infty}^{m} (|y-m| - |y-v|)p(y)dy + \int_{m}^{v} (|y-m| - |y-v|)p(y)dy + \int_{v}^{\infty} (|y-m| - |y-v|)p(y)dy \\
&= \int_{-\infty}^{m} ((m-y) - (v-y))p(y)dy + \int_{m}^{v} ((y-m) - (v-y))p(y)dy + \int_{v}^{\infty} ((y-m) - (y-v))p(y)dy \\
&= \int_{-\infty}^{m} (m-v)p(y)dy + \int_{m}^{v} (2y-m-v)p(y)dy + \int_{v}^{\infty} (v-m)p(y)dy \\
&\geq (m-v)P(Y \leq m) + (2m-m-v)(0.5 - P(Y \leq m)) + (v-m) * 0.5 \text{ because v is median} \\
&\geq 0.5 * (m-v) + (v-m) * 0.5 \\
&\geq 0
\end{aligned}
$$

Hence, $R(m) \geq R(v)$ in this case. Alternatively, if $m > v$ then:

$$
\begin{aligned}
R(m) - R(v) &= \int_{-\infty}^{\infty} (|y-m| - |y-v|)p(y)dy \\
&= \int_{-\infty}^{v} (|y-m| - |y-v|)p(y)dy + \int_{v}^{m} (|y-m| - |y-v|)p(y)dy + \int_{m}^{\infty} (|y-m| - |y-v|)p(y)dy \\
&= \int_{-\infty}^{v} ((m-y) - (v-y))p(y)dy + \int_{v}^{m} ((m-y) - (y-v))p(y)dy + \int_{m}^{\infty} ((y-m) - (y-v))p(y)dy \\
&= \int_{-\infty}^{v} (m-v)p(y)dy + \int_{v}^{m} (m+v-2y)p(y)dy + \int_{m}^{\infty} (v-m)p(y)dy \\
&\geq (m-v) * 0.5 + (m+v-2m)(0.5 - P(Y \geq m)) + (v-m) * P(Y \geq m) \text{ because v is median} \\
&\geq 0.5 * (m-v) + (v-m) * 0.5 \\
&\geq 0
\end{aligned}
$$

Hence, $R(m) \geq R(v)$ in this case. Since $R(m) \geq R(v)$ in all cases where $m \neq v$, $m = v$ is where $R(m)$ is minimized.

2.

$$
\begin{aligned}
bias^2 + Var[\hat{\theta}] &= (E[\hat{\theta}] - \theta)^2 + E[(\hat{\theta} - E[\hat{\theta}])^2] \\
&= (E[\hat{\theta}])^2 - 2\theta E[\hat{\theta}] + \theta^2 + E[\hat{\theta}^2] - 2(E[\hat{\theta}])^2 + (E[\hat{\theta}])^2 \\
&= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2] \\
&= E[(\hat{\theta} - \theta)^2] \\
&= MSE
\end{aligned}
$$

3a. We know that $\overline{Y}_n$ has distribution $N(\mu, \sigma^2/n)$.
Since the Y_i's are independent, we then get that the difference, $Y_{n+1} - \overline{Y}_n$ has distribution $N(\mu - \mu, \sigma^2 + \sigma^2/n) = N(0, (n+1/n)\sigma^2)$ because Y_{n+1} has distribution $N(\mu, \sigma)$.
Let $Z = Y_{n+1} - \overline{Y}_n$. Then:

$$\begin{aligned} Var[Z] &= E[Z^2] - (E[Z])^2 \\ &= E[Z^2] \text{ because of the previously demonstrated distribution for Z} \\ E[Z^2] &= \frac{n+1}{n}\sigma^2 \end{aligned}$$

3b. $P(Y_{n+1} \in C) = P(-c \leq Y_{n+1} - \overline{Y} \leq c)$. As shown in 3a, $Y_{n+1} - \overline{Y}$ has distribution $N(0, (n+1/n)\sigma^2)$.
If we let $c = \sqrt{(n+1/n)}\sigma z_{\alpha/2}$, then the probability will equal $P(-z_{\alpha/2} \leq \frac{Y_{n+1} - \overline{Y}}{\sqrt{(n+1/n)}\sigma} \leq z_{\alpha/2})$.
Since dividing a normally distributed random variable by a constant results in a new normally distributed random variable with the same mean and with variance divided by the square of that constant, the new random variable has distribution $N(0,1)$, so this probability is equal to $1 - \alpha/2 - \alpha/2 = 1 - \alpha$, so the chosen value for c is the desired one.

4a. $2X \sim N(0,4)$ so $Y = 2X + \epsilon \sim N(0,5)$ because X and ϵ are independent. Hence, the mean of Y is 0 and the variance is 5.

4b. $Var[Y] = E[Y^2] - (E[Y])^2 = E[Y^2]$ because $Y \sim N(0,5)$ from 4a. Hence, $E[Y^2] = Var[Y] = 5$.

4c.

$$\begin{aligned}
E[Y|X=x] &= E[2X + \epsilon|X=x] \\
&= E[2x + \epsilon] \\
&= 2x + E[\epsilon] \\
&= 2x \text{ because } \epsilon \sim N(0,1)
\end{aligned}$$

4d. Since $Y \sim N(0,5)$, the probability distribution of Y is symmetric about 0, so $p_Y(y) = p_Y(-y)$. Then,

$$\begin{aligned}
E[Y^3] &= \int_{-\infty}^{\infty} y^3 p_Y(y)dy \\
&= \int_{-\infty}^{0} y^3 p_Y(y)dy + \int_{0}^{\infty} y^3 p_Y(y)dy \\
&= \int_{\infty}^{0} -u^3 p_Y(-u)(-du) + \int_{0}^{\infty} y^3 p_Y(y)dy \text{ substituting } u = -y \\
&= -\int_{0}^{\infty} u^3 p_Y(u)du + \int_{0}^{\infty} y^3 p_Y(y)dy \text{ by symmetry of Y about 0} \\
&= 0
\end{aligned}$$

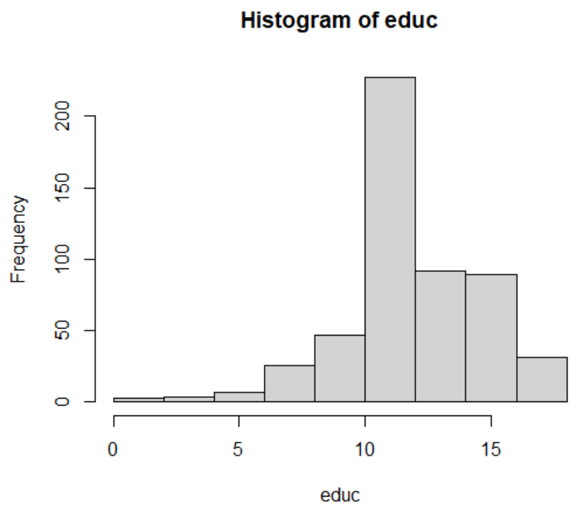
4e. Note that $E[\epsilon] = 0$ from its distribution, $E[\epsilon^2] = 1$ from the same reasoning as 4b, and $E[\epsilon^3] = 0$ from the same reasoning as 4d. Then,

$$\begin{aligned}
Cov(\epsilon, \epsilon^2) &= E[\epsilon * \epsilon^2] - E[\epsilon]E[\epsilon^2] \\
&= E[\epsilon^3] - E[\epsilon]E[\epsilon^2] \\
&= 0 - 0 * 1 \\
&= 0
\end{aligned}$$

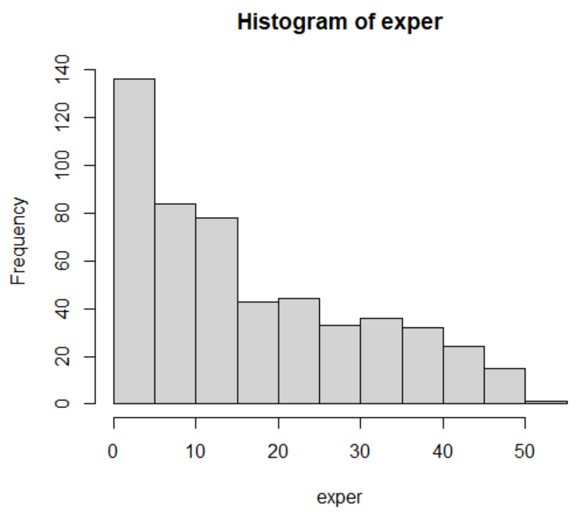
ϵ and ϵ^2 are not independent because determining a value for ϵ also determines a value for ϵ^2 instead of leaving its probability distribution as it was initially.

5a.
> hist(wage)

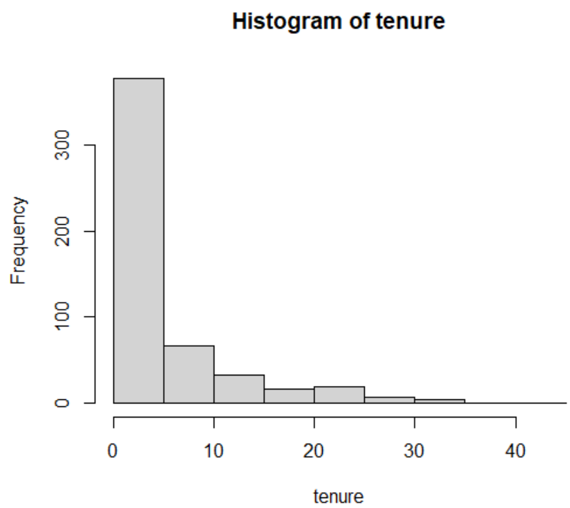
> hist(educ)



> hist(exper)



> hist(tenure)



5b. Note: ith variable from left to right has corresponding mean/standard deviation on the line below in ith position from left to right.

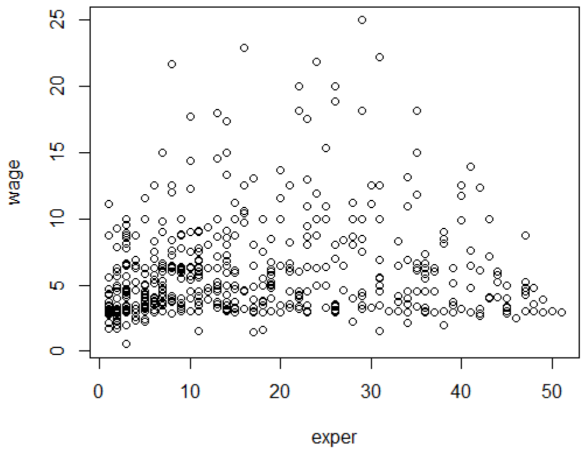
For the mean:
> apply(wage1,2,mean)

wage educ exper tenure nonwhite
5.89610267 12.56273764 17.01711027 5.10456274 0.10266160
female married numdep smsa northcen
0.47908745 0.60836502 1.04372624 0.72243346 0.25095057
south west construc ndurman trcommpu
0.35551331 0.16920152 0.04562738 0.11406844 0.04372624
trade services profserv profocc clerocc
0.28707224 0.10076046 0.25855513 0.36692015 0.16730038
servocc lwage expersq tenursq
0.14068441 1.62326844 473.43536122 78.15019011

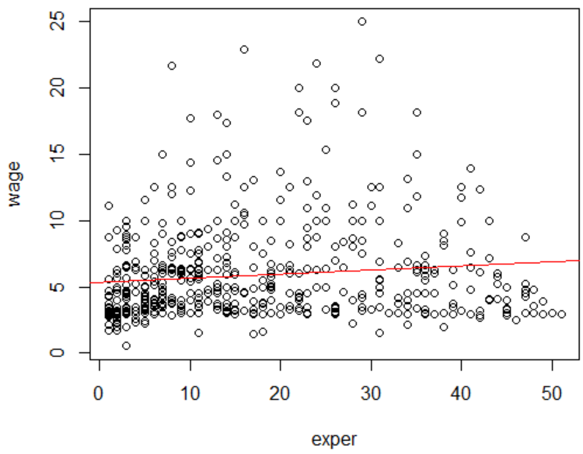
For the standard deviation:
> apply(wage1,2,sd)

wage educ exper tenure nonwhite female
3.6930860 2.7690224 13.5721596 7.2244623 0.3038053 0.5000380
married numdep smsa northcen south west
0.4885804 1.2618915 0.4482246 0.4339728 0.4791242 0.3752867
construc ndurman trcommpu trade services profserv
0.2088743 0.3181970 0.2046800 0.4528262 0.3012978 0.4382574
profocc clerocc servocc lwage expersq tenursq
0.4824233 0.3735991 0.3480267 0.5315382 616.0447716 199.4346635

5c.
> out = lm(wage exper)
> plot(exper,wage)



> abline(out,col='red')



I don't think the line is a good summary of the relationship between exper and wage because a large portion of the points are far away from the line.

5d.

> summary(out)

Call:
lm(formula = wage   exper)

Residuals:
Min 1Q Median 3Q Max
-4.936 -2.458 -1.112 1.077 18.716

Coefficients:
Estimate Std. Error t value Pr(> $|t|$)
(Intercept) 5.37331 0.25699 20.908 ¡ 2e-16 ***
exper 0.03072 0.01181 2.601 0.00955 **
—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.673 on 524 degrees of freedom
Multiple R-squared: 0.01275, Adjusted R-squared: 0.01086

F-statistic: 6.766 on 1 and 524 DF, p-value: 0.009555

This tells us that the intercept is 5.37331 and the slope is 0.03072
> mean(exper)
[1] 17.01711
> mean(wage)
[1] 5.896103
Hence, the mean is (17.01711, 5.896103).

> confint(out,level=0.9)
5 % 95 %
(Intercept) 4.94984261 5.79676788
exper 0.01125999 0.05018375

This indicates that the 90% confidence interval for the slope is [0.01125999, 0.05018375]