Problem 4.4

(a) σ is variable represents the noise level, therefore $\sigma\epsilon$ is the normal distributed noise on the data since ϵ is an independent and identically distributed random variable. We normalize the f function so that noise level σ^2 is calibrated to the function f level.

The orthogonality of Legendre polynomials:

$$\int_{-1}^{1} dx\, L_m(x)L_n(x) = \begin{cases} 0, & m \neq n \\ \frac{2}{2n+1}, & m = n \end{cases}$$

$$f = \sum_{q=0}^{Q_f} a_q L_q(x)$$

To normalize f function and obtain $E[f^2] = 1$:

$$f = \sum_{q=0}^{Q_f} a_q L_q(x)$$

$$E_{a,x}[f^2] = E_{a,x}\left[\sum_{q=0}^{Q_f} a_q L_q(x)\, f\right] = E_a\left[E_x\left[\sum_{q=0}^{Q_f} a_q L_q(x)\right]\right]$$

Applying the orthogonality of the Legendre polynomials:

$$E_x\left[a_i a_j L_i(x) L_j(x)\right] = \begin{cases} 0, & i \neq j \\ \frac{2a_i^2}{2i+1}, & i = j \end{cases}$$

$$E_{a,x}[f^2] = E_a\left[\sum_{q=0}^{Q_f} \frac{a_q^2}{2q+1}\right] = \sum_{q=0}^{Q_f} \frac{E_a\left[a_q^2\right]}{2q+1}$$

Because a_q is a standard normal with $\mu = 0$ and $\sigma^2 = 1$

$$\sigma^2 = E\left[\left(a_q - \mu\right)^2\right] = E\left[a_q^2\right] = 1$$

$$E_{a,x}[f^2] = \sum_{q=0}^{Q_f} \frac{E_a\left[a_q^2\right]}{2q+1} = 1$$

$$\sum_{q=0}^{Q_f} \frac{2E_a\left[a_q^2\right]}{2q+1} = \sum_{q=0}^{Q_f} \frac{2}{2q+1} = 1 = E[f^2]$$

Therefore the normalization factor is $\sqrt{\sum_{q=0}^{Q_f} \frac{1}{2q+1}}$.

(b)

First we have the formula for g_2 and g_{10} in the orthogonal Legendre polynomials form

$$g_2 = a_1 L_1 + a_2 L_2$$

$$g_{10} = a_1 L_1 + a_2 L_2 + \cdots + a_{10} L_{10}$$

With training data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

We calculated the value of Legendre polynomial from order 0 to 10 at each x_i. The training data hence becomes

$$\left\{\left(\left(\left(L_0(x_1), L_1(x_1), L_2(x_1)\right)\right), y_1\right), \left(\left(\left(L_0(x_2), L_1(x_2), L_2(x_2)\right)\right), y_2\right), \ldots, \left(\left(\left(L_0(x_n), L_1(x_n), L_2(x_n)\right)\right), y_n\right)\right\}$$

for g_2

Then g_2 is a three dimensional linear regression.

Similarly g_{10} is an eleven dimensional linear regression.

(c)

The actual target function is $f(x) = \sum_{q=0}^{Q_f} a_q L_q(x)$ and the hypothesis is $g_{10}(x) = \sum_{i=0}^{10} w_i L_i\,(x)$

Therefore we have the E_{out} for g_{10} as below:

$$E_{out10} = E[(g_{10}(x) - y)^2] = E[(g_{10}(x) - f(x) - \sigma\epsilon)^2]$$

$$= E[\left(g_{10}(x) - f(x)\right)^2 - 2\left(g_{10}(x) - f(x)\right)\sigma\epsilon + (\sigma\epsilon)^2]$$

$$= E\left[\left(\sum_{i=0}^{10} w_i L_i\,(x) + \sum_{q=0}^{Q_f} a_q L_q(x)\right)^2 - 2\left(\sum_{i=0}^{10} w_i L_i\,(x) - \sum_{q=0}^{Q_f} a_q L_q(x)\right)\sigma\epsilon + (\sigma\epsilon)^2\right]$$

Because we have ϵ normal distributed with 0 mean and 1 variance

We have

$$E_{out10} = \left[\left(\sum_{i=0}^{10} w_i L_i\,(x)\right)^2 - 2\left(\sum_{q=0}^{\min(10,Q_f)} (w_q a_q L_q(x))^2\right) + \sum_{q=0}^{Q_f}\left(a_q L_q(x)\right)^2\right]$$

$$= \sum_{i=0}^{10} \frac{w_i^2}{2i+1} - 2\sum_{q=0}^{\min(10,Q_f)} \frac{w_q a_q}{2q+1} + \sum_{q=0}^{Q_f} \frac{a_q^2}{2q+1}$$

(d) (The graphs are attached at the end of the answer.)

The over fit measure is significantly positive when the noise level σ^2 is high, while the number of data points and target function complexity are fixed. Because of the noise, the data deviates from the actual target function. Yet, g_{10} has more power to fit the data, which results in more significant out of sample error because g_{10} also tries to fit the noise.

The over fit measure is also significantly positive when the target function complexity is high, while the number of data points is low ($N = 40$). The deterministic noise is dependent on target function complexity. Similarly g_{10} has more power to fit the data including the noise, resulting a significant out of sample error. However, when the number of data points become bigger, the higher model complexity leads to lower error measure.

Additionally, when the number of data points is small, serious overfitting also occurs. In other words, as the number of data points grows larger, the error measure grows smaller. The reason is that with more data points, the trained model is more closed to the actual target.

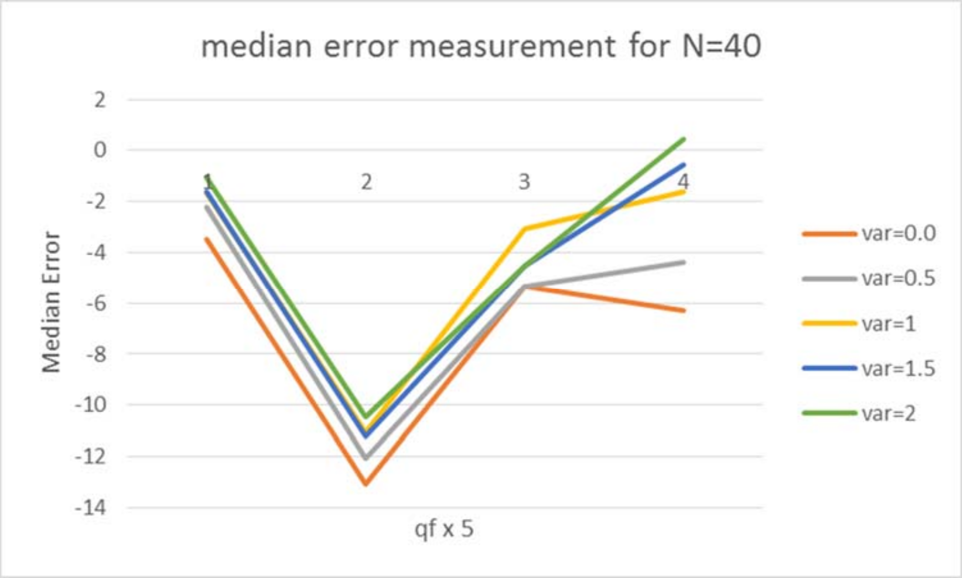
All in all, the over fit measure is negative when the noise is low, the target function complexity is close to g_{10} and the number of data points is large. In that case, g_{10} has a more power to fit the actual correct data (with minimal noise). Therefore the out of sample error of g_{10} is small and hereby the over fit measure is small.

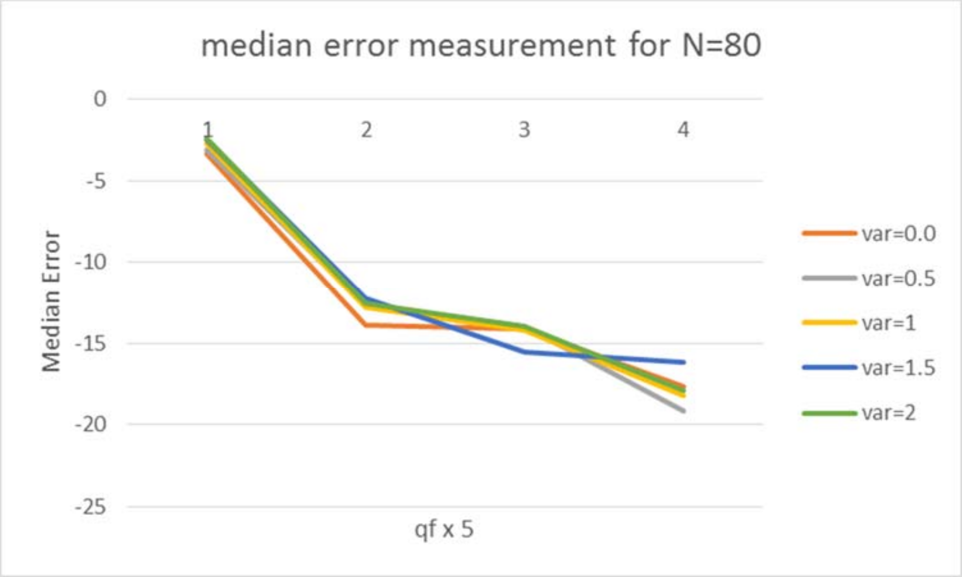
The mean and median are in general very close to each other. However, when the mean overfit measure is high, the difference between the mean and median tends to become larger. This may be that when mean is high, most values are high and it's likely that certain estimate produce a terrible extreme value and hence influence the mean, resulting in a larger difference between the mean and median. On the contrary, median error measurements remains relatively consistent since it is not heavily influenced by extreme values.

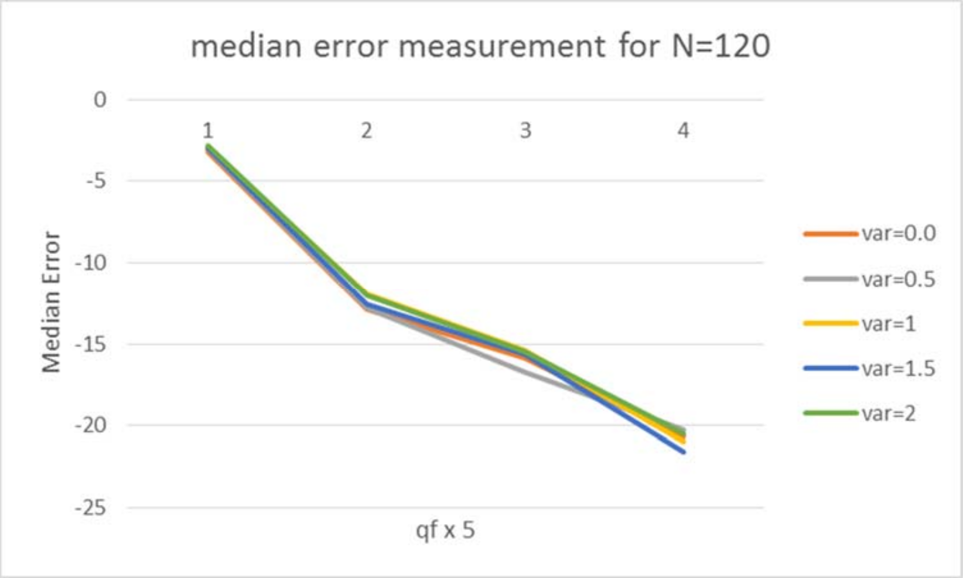
Below is the table for all error measurements for all variances, QFs and sample sizes.

The entries on the left side should be multiplied by the number after var=x. For example, the first entry should be $-0.0045 \times (1.00E + 03) = -4.5$. The right side entries need not to be multiplied.

	Mean Error Measurement					Median Error Measurement		
	N=40	***N=80***	***N=120***			***N=40***	***N=80***	***N=120***
var=0.0	1.00E+03	*			**var=0.0**			
qf=5	-0.0045	-0.0041	-0.004		*qf=5*	-3.462	-3.3257	-3.1846
qf=10	-0.0147	-0.0148	-0.0143		*qf=10*	-13.0608	-13.8716	-12.8414
qf=15	5.1677	0.0109	-0.0058		*qf=15*	-5.3592	-14.1053	-15.8524
qf=20	6.4225	0.0398	-0.0158		*qf=20*	-6.2757	-17.6528	-20.6172
var=0.5	1.00E+03	*			**var=0.5**			
qf=5	0.0247	-0.004	-0.004		*qf=5*	-2.228	-3.1166	-3.0516
qf=10	0.8366	-0.014	-0.0136		*qf=10*	-12.068	-12.607	-12.7495
qf=15	1.8405	-0.0003	-0.0173		*qf=15*	-5.3599	-13.9461	-16.7381
qf=20	3.6846	0.0124	-0.0169		*qf=20*	-4.4027	-19.1197	-20.2496
var=1	1.00E+03	*			**var=1**			
qf=5	0.126	-0.0033	-0.0037		*qf=5*	-1.6752	-2.6889	-2.9523
qf=10	0.0274	-0.0137	-0.014		*qf=10*	-11.0831	-12.8035	-11.94
qf=15	2.69	0.0063	-0.0168		*qf=15*	-3.0472	-14.2093	-15.3873
qf=20	1.0457	0.0274	-0.0187		*qf=20*	-1.5961	-18.2219	-20.9718
var=1.5	1.00E+04	*			**var=1.5**			
qf=5	0.0195	-0.0003	-0.0004		*qf=5*	-1.5977	-2.5139	-2.9355
qf=10	0.405	-0.0013	-0.0014		*qf=10*	-11.2304	-12.2561	-12.5536
qf=15	4.0024	-0.001	-0.0017		*qf=15*	-4.5342	-15.5241	-15.6151
qf=20	2.8696	0.0068	-0.0022		*qf=20*	-0.5351	-16.1906	-21.6218
var=2	1.00E+03	*			**var=2**			
qf=5	0.0778	-0.0031	-0.0037		*qf=5*	-1.0751	-2.4417	-2.79
qf=10	0.0823	-0.0138	-0.0139		*qf=10*	-10.4728	-12.5583	-12.0367
qf=15	1.5139	-0.0041	-0.0018		*qf=15*	-4.5198	-13.9199	-15.4251
qf=20	2.5636	0.0028	-0.0199		*qf=20*	0.4649	-17.8965	-20.4659


median error measurement for N=40
2
0
-2
-4
-6
-8
-10
-12
-14
Median Error
1
2
3
4
qf x 5
var=0.0
var=0.5
var=1
var=1.5
var=2


median error measurement for N=80
0
-5
-10
-15
-20
-25
Median Error
1
2
3
4
qf x 5
var=0.0
var=0.5
var=1
var=1.5
var=2



1. In each graph, we can see that model complexity (qf) have different effects based on the number of data points. When the number of data points are small, high complexity leads to severe overfitting and when the number of data points are big, high complexity leads to smaller median error.
2. And we can see that as N goes from 40 to 120, the line shift down in the median error measure across the graphs.
3. The variance have small effects on the overall error measurements. Yet overall the less the variance, the smaller error measurements.

Problem 4.25

(a) No, we should not select the learner with minimum validation error. First, the VC bound is:

$$E_{out}(g^{-}) \leq E_{val}(g^{-}) + O\left(\sqrt{\frac{\ln M}{2K}}\right)$$

Because each learner is trained on different size of validation set, $O\left(\sqrt{\frac{\ln M}{2K}}\right)$ is different for each model. Therefore selecting the minimum E_{val} does not guarantee the minimum E_{out} due to the fact that $O\left(\sqrt{\frac{\ln M}{2K}}\right)$ is different.

(b)

First, the validation error is an unbiased estimate of E_{out} because the final hypothesis g^{-} was obtained independently of the data points in the validation set.

Second, the validation process is equivalent to learning a hypothesis from g_{val} using the data in the validation set. Therefore, we can apply the VC bound.

(c)

Using Hoeffding inequality $P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$ to obtain a bound on the in-sample and out-of-sample errors:

$$P[|E_{in} - E_{out}| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

The validation errors $E_{val}(m^*)$ are 'in-sample' errors for this learning process. That is

$$P[E_{out}(m) - E_{val}(m) > \epsilon] \leq e^{-\epsilon^2 K_m}$$

Since we have M evaluatations

$$P[E_{out}(m^*) - E_{val}(m^*) > \epsilon] \leq P[E_{out}(1) - E_{val}(1) > \epsilon] + \cdots + P[E_{out}(M) - E_{val}(M) > \epsilon]$$

Because

$$P[E_{out}(1) - E_{val}(1) > \epsilon] + \cdots + P[E_{out}(M) - E_{val}(M) > \epsilon] \leq \sum_{m=1}^{M} e^{-2\epsilon^2 K_m}$$

Let $k(\epsilon) = -\frac{1}{2\epsilon^2}\ln\left(\frac{1}{M}\sum_{m=1}^{M} e^{-2\epsilon^2 K_m}\right)$, we then have

$$P[|E_{val}(m^*) - E_{out}(m^*)| > \epsilon] \leq 2Me^{-2\epsilon^2 k(\epsilon)}$$

$$P[E_{val}(m^*) - E_{out}(m^*) > \epsilon] = P[E_{out}(m^*) - E_{val}(m^*) > \epsilon] \leq Me^{-2\epsilon^2 k(\epsilon)}$$

$$P[E_{out}(m^*) > E_{val}(m^*) + \epsilon] \leq Me^{-2\epsilon^2 k(\epsilon)}$$

Problem 5.4

(a)

(i) The problem is that there is data snooping involved when we manually selects the S&P 500 companies. The hypothesis, which is the companies in this case, is already narrowed down before looking at the data. Therefore the Hoeffding inequality does not apply.

(ii) A better estimate should evaluate all hypothesis (all companies ever traded/is trading), which we have $M = 50000$. We then have the estimate:

$$P[|E_{in} - E_{out}| > 0.02] \leq 2 \times 50000 \times e^{-2\times12500\times0.02^2} \approx 4.54$$

(b)

(i) There is sampling bias involved. Therefore the results are biased because only currently trading companies are evaluated (S&P 500). That is, all the companies that bankrupted or do not make it to the top five hundred is not considered in the evaluation and yet we do not know if our current trading companies will survive or make to five hundred in the future.

Hence, our training data is not representative of the test data and the corresponding conclusion does not work out for general stock trading.

(ii) Based on the information given, buy and hold is a bad strategy because the calculated company survive rate is approximately $10000 \div 50000 = 0.2$, which is low.