Quant I: Final Exam Solutions

December 2013

1 Written Part

Question 1. Simple OLS example.

a. Show that OLS estimates $\hat{\beta}_0 \approx -2.89$ and $\hat{\beta}_1 \approx 1.57$.
The formula for the beta matrix is $(X'X)^{-1}(X'Y)$.
We are already given $(X'X)^{-1}$.

$$X'Y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 3 & -1 & 4 \end{bmatrix} \begin{bmatrix} 4 \\ 2 \\ -6 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 24 \end{bmatrix}$$

$$(X'X)^{-1}(X'Y) = \begin{bmatrix} .536 & -.143 \\ -.143 & .071 \end{bmatrix} \begin{bmatrix} 1 \\ 24 \end{bmatrix} = \begin{bmatrix} -2.89 \\ 1.57 \end{bmatrix} = \hat{\beta}$$

b. Show that $\hat{\sigma} \equiv SEE \approx 3.33$

$$\hat{\sigma} = \sqrt{\frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{N-K-1}} = \sqrt{\frac{1}{N-K-1}(\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}})}$$

First we can calculate $\hat{\mathbf{y}}$:

$$\begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & -1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} -2.89 \\ 1.57 \end{bmatrix} = \begin{bmatrix} .25 \\ 1.82 \\ -4.46 \\ 3.39 \end{bmatrix}$$

$$\hat{\sigma} = \frac{1}{\sqrt{2}}[(\begin{bmatrix} 4 \\ 2 \\ -6 \\ 1 \end{bmatrix} - \begin{bmatrix} .25 \\ 1.82 \\ -4.46 \\ 3.39 \end{bmatrix})'(\begin{bmatrix} 4 \\ 2 \\ -6 \\ 1 \end{bmatrix} - \begin{bmatrix} .25 \\ 1.82 \\ -4.46 \\ 3.39 \end{bmatrix})]^{\frac{1}{2}}$$

$$= (\begin{bmatrix} 3.75 \\ 0.18 \\ -1.54 \\ -2.39 \end{bmatrix}' \begin{bmatrix} 3.75 \\ 0.18 \\ -1.54 \\ -2.39 \end{bmatrix})^{\frac{1}{2}}$$

$$\approx \frac{\sqrt{22.2}}{\sqrt{2}} = 3.33$$

c. Show that $R^2 \approx .61$

$$R^2 = 1 - \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})}$$

From the solution for $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ above:

$$R^2 = 1 - \frac{22.2}{(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})}$$

$$\bar{y} = \frac{\sum_i y_i}{n} = \frac{1}{4}$$

$$(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}}) = (\begin{bmatrix} 4 \\ 2 \\ -6 \\ 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix})'(\begin{bmatrix} 4 \\ 2 \\ -6 \\ 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix}) = 56.75$$

$$R^2 = 1 - \frac{22.2}{56.75} \approx 0.61$$

Question 2. Consider four random variables W, X, Y and Z, where:

$$cov(W,Y) > 0 \quad cov(W,X) = 0;$$
$$cov(Z,Y) = 0 \quad cov(Z,X) < 0;$$
and $cov(X,Y)$ is unknown.

Say whether the following statements are TRUE or FALSE, and explain why. Assume we have a large number of observations of the joint distribution of all four variables from an i.i.d. random sample.

a. If we estimate the equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $\hat{\beta}_1$ is a *biased* estimate of the parameter β_1 due to the omission of w and z.

False. $\hat{\beta}_1$ is biased due to the omission of w iff $cov(W,X) \neq 0$ & $cov(W,Y) \neq 0$. But, from the above, $cov(W,X) = 0$, implying that $\hat{\beta}_1$ is not biased due to the omission of w. Analogously, $\hat{\beta}_1$ is biased by the omission of z iff $cov(Z,X) \neq 0$ & $cov(Z,Y) \neq 0$. But, from the above $cov(Z,Y) = 0$, therefore $\hat{\beta}_1$ is not biased due to the omission of z.

b. The estimate of the parameter β_1 we obtain from the estimated equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ will be *more efficient* than the estimate obtained from the equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i$.

True. $\hat{\beta}_1$ will be distributed normally with mean β_1 and variance $\frac{\sigma^2}{\sum_i (x_i - \bar{x})^2 (1 - R_x^2)}$ where R_x^2 is the R^2 from a regression of x on all other regressors. Since there are no other regressors in the first model $R_x^2 = 0$. And since $cov(X,Z) < 0$ the R_x^2 from the second model is strictly greater than zero. Thus, the variance of $\hat{\beta}_1$ is greater for the second model than for the first.

c. The estimate of the parameter β_1 we obtain from the estimated equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ will be *more efficient* than the estimate obtained from the equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 w_i$.

True. Note that while the inclusion of w in the regression will not affect the true variance of $\hat{\beta}_1$ given by $\frac{\sigma^2}{\sum_i (x_i - \bar{x})^2 (1 - R_x^2)}$ since the R^2 from a regression of x on z is zero; our *estimated variance of* $\hat{\beta}_1$ *will be affected*. This is because we do not observe the true population parameter σ^2. Rather, we must estimate this value by substituting $\hat{\sigma}^2 = \frac{\sum_i \hat{u_i}^2}{N-K-1}$, to obtain $\widehat{Var(\hat{\beta}_1)}$. Since we are increasing the value K (reducing our degrees of freedom), we are reducing the precision of our estimates in any finite sample. Though the magnitude of this effect will be declining as N increases.

Question 3. Consider the three variables X, Y, and Z, where in the population

- X takes on the value zero 50 percent of the time and the value one 50 percent of the time, while
- Z takes on the value zero 3 percent of the time and the value one 97 percent of the time.

You are interested in estimating the *ceteris paribus* association of X with Y as well as the *ceteris paribus* association of Z with Y. To do so, you use the model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i$. Assume that this model is properly specified and the Gauss-Markov assumptions hold.

a. One or more of the following (omitted) statements is true. In a few sentences, identify the correct statement(s) and explain:

The expression $var(\hat{\beta}_1) = \frac{\sigma^2}{nvar(x)(1-R_x^2)}$ is correct. As is the expression $\widehat{var(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{nvar(x)(1-R_x^2)}$. First, note that the expression $Var(\beta_1)$ makes no sense in a frequentist framework. There exists a unique true value β_1 and thus it has a variance of zero. The expression for $Var(\hat{\beta}_1)$ is given by $\frac{\sigma^2}{\sum_i (x_i - \bar{x})^2 (1-R_x^2)}$. Note that $Var(x) = \frac{1}{n}\sum_i (x_i - \bar{x})^2 \Rightarrow Var(\hat{\beta}_1) = \frac{\sigma^2}{nVar(x)(1-R_x^2)}$. Now, since we do not observe the true value of σ^2 we have to estimate this using $\hat{\sigma}^2$. Thus, our estimated variance for the parameter $\hat{\beta}_1$ is $\widehat{Var(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{nvar(x)(1-R_x^2)}$. Since this expression is correct, the final expression given cannot be correct.

b. Why can we say for sure that $var(\hat{\beta}_1) < var(\hat{\beta}_2)$?

Note that the expression $var(\hat{\beta}_j) = \frac{\sigma}{\sum_i (x_{j,i} - \bar{x_j})^2 (1-R_{x_j})}$ for any variable x_j with associated coefficient $\hat{\beta}_j$. Note further that we can think of X as being drawn from a Bernoulli distribution with $p = 0.5$. Z is also drawn from a Bernoulli distribution with $p = 0.03$. For any variable x_j drawn from a Bernoulli distribution $\sum_i (x_{j,i} - \bar{x_j})^2 = np(1-p)$ where $p \in [0, 1]$. Note that this value reaches a maximum when $p = 0.5$. Therefore, the denominator of the expression for $var(\hat{\beta}_1)$ is greater than the denominator for $var(\hat{\beta}_2)$, implying that $var(\hat{\beta}_1) < var(\hat{\beta}_2)$.

c. All things being equal, with which of the two findings should you be more comfortable?

- A failure to reject the null that the ceteris paribus association between Z and Y is zero.
- A failure to reject the null that the ceteris paribus association between X and Y is zero.

Given that the $Var(\hat{\beta}_1) < Var(\hat{\beta}_2)$, I would be more comfortable failing to reject the null that the association between Z and Y is zero. The larger variance on $\hat{\beta}_2$ implies that my confidence interval for this estimate will be wider than that for $\hat{\beta}_1$. As a result, the confidence interval on $\hat{\beta}_2$ is more likely to contain zero - all else equal - than that on $\hat{\beta}_1$.

Question 4. 25 Points Consider the Stata output on the following page. It is an OLS analysis of ratings given to Barak Obama (on a zero to 100 scale) in the 2012 American National Election Studies by a nationally representative sample of American adults. Be sure to explain your answers and show your work.

a. What proportion of the respondents in the sample own guns?

   See the summary statistics following the sum command, from which can see from the minimum and maximum that it is a dummy variable, with mean 32.57.

b. What is the rating predicted to be given to Obama by a (non-Hispanic) African American man born in the US whose education and age are equal to the American average, whose household income is $60,000 and who is a military veteran and a union member but not a gun owner?

$$\hat{y} = 71.74 + 24.16 \cdot black + (-.1059) \cdot age + (.2807) * black \cdot age + (-8.29) \cdot educ +$$

$$1.44 \cdot educ^2 + (-.555) \cdot ln(income) - 3.14 \cdot vet + 4.9 \cdot union = 88.34$$

$$\hat{y} = 71.74 + 24.16 * 1 + (-.1059) * 49.8 + (.2807) * 49.8 + (-8.29) * 3 +$$

$$1.44 * 9 + (-.555) * ln(60,000) - 3.14 * 1 + 4.9 * 1 = 94.4$$

c. How many standard deviations away from y is the typical prediction $\hat{y}$?

   According to the "Root MSE", on average, our predictions miss the true value of y by 29.78 feeling thermometer units. We know that the standard deviation of y in the sample is 28.5 from the summary statistics table, so we have:

$$\frac{29.78}{34.499} = 0.863$$

d. The constant term is about 72. Describe the hypothetical American whose predicted rating of Obama is described by this term.

   Male, age = 0, non-black, non-hispanic, US-born, not a Veteran, not a gun owner, not a Union member, $1 income, education category 1.

e. What is the approximate predicted difference in ratings given to Obama between someone with a household income of $30,000 and someone with an income of $ 45,000, holding all other covariates constant?

$(log(45,000) - log(30,000) * (-.5553) = -.23$

f. What is $\frac{\partial ObamaFT}{\partial AGE}$?

The relevant part of our model, ignoring other covariates that don't vary with `AGE`, is $y = \beta_{AGE} \cdot age + \beta_{BLACK*AGE} \cdot age$. So the change is given by:

$$\frac{\partial ObamaFT}{\partial AGE} = \beta_{AGE} + \beta_{BLACK*AGE} \cdot \mathbf{1}_{black}$$

where $\mathbf{1}_{black}$ is the dummy variable for whether the respondent is black.

g. What is $\frac{\partial ObamaFT}{\partial EDUC}$?

Here the relevant part of our model, ignoring other covariates that don't vary with `EDUC`, is $y = \beta_{EDUC} \cdot EDUC + \beta_{EDUC^2} \cdot EDUC^2$. This gives:

$$\frac{\partial ObamaFT}{\partial EDUC} = \beta_{EDUC} + 2\beta_{EDUC^2} \cdot EDUC$$

meaning that y attains an extremum w.r.t. `EDUC` at `EDUC` $= \frac{\beta_{EDUC}}{2\beta_{EDUC^2}}$, which is a maximum if $\beta_{EDUC} > 0$ and $\beta_{EDUC^2} < 0$ and a minimum if $\beta_{EDUC} < 0$ and $\beta_{EDUC^2} > 0$.

2 Statistical Computing Part

Question 1. Run the command that lists the names of variables, their formats and labels. The command is `describe`

```
  obs:         5,870
 vars:             8
 size:       187,840
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
age             float   %9.0g                 age (in years)
bush_favorabi~y float   %25.0g     caa01      0=very unfavorable, 10=very favorable
economic_situ~n float   %10.0g     ccb04      personal economic situation today
female          float   %9.0g      female     sex (0 = male, 1 = female)
ideology        float   %17.0g     cma06      conservative or liberal
income          float   %24.0g     income     total hh income before taxes
kerry_favorab~y float   %25.0g     cab01      0=very unfavorable, 10=very favorable
pid             float   %14.0g     cma01      party identification
-------------------------------------------------------------------------------
```

Question 2. (18 points) `bush_favorability` is a measure of respondents? favorability ratings of George W. Bush on a scale of zero (least favorable) to ten (most favorable). (NOTE: For reasons that should be obvious, you will need to create a recoded version of this variable.)

a. What is the mode of `bush_favorability`? From tab `bush_favorability`: 0

b. What is the mean? 5.28

c. What is the 95% confidence interval about this mean? (5.19, 5.37)

d. What is the 90% confidence interval about this mean? (5.21, 5.36)

e. Who rates Bush more favorably, men or women? Men rate Bush more favorably, and difference is statistically significant. From the t-test, the difference between men and women is 0.37, with a t-value of 3.91. We get the same results from the simple regression, where the coefficient on `female` is -0.37 with an associated t-value of 3.91.

f. Are higher income Americans more likely to assess Bush favorably?

After recoding, if we simply run `regress bush_favorability income` the coefficient is significant and although it is very small ($4.6x10^6$), this reflects the fact that income ranges

from $2,500 to $150,000. If we run margins after this, however, it is clear that income does have a non-trivial effect:

A better approach is to include the log, and the square of the log, which suggests there is a positive and significant effect of age on `bush_favorability` which is substantively important, with a difference of more than ten percentage points moving from the lowest to the highest category of income.

	(1)	(2)
income	.000***	
	(.000)	
income (log)		3.209***
		(.720)
income sq (log)		-.143***
		(.036)
Constant	4.958***	-12.592***
	(.088)	(3.580)
N	5204	5204

Question 3. (20 points) `kerry_favorability` is a measure of respondents? ratings of John Kerry on the same scale as `bush_favorability`. Create a new variable (as done in class) called `rating_diff` that is equal to `bush_favorability` minus `kerry_favorability`. (You'll again need to do some recoding.)

a. Model I: `reg rating_diff age female ln_inc`

	I	II	III
age (in years)	.007	-.008+	-.008+
	(.005)	(.005)	(.005)
sex (0 = male, 1 = female)	-.772***	-.534***	-.500***
	(.165)	(.148)	(.147)
ln_inc	.717***	.713***	.719***
	(.083)	(.076)	(.075)
conservative or liberal		-2.648***	
		(.073)	
2.conservative or liberal			-.833**
			(.292)
3.conservative or liberal			-4.702***
			(.284)
4.conservative or liberal			-7.142***
			(.308)
5.conservative or liberal			-8.646***
			(.397)
Constant	-7.603***	.674	-3.072***
	(.943)	(.882)	(.878)
N	4827	4706	4706
r2	0.02	0.23	0.24
rmse	5.68	5.02	4.98

b. How well does Model I explain variation in the dependent variable?

Not very well. The R^2 is only 0.02 so we are only "explaining" about 2% of the variation in the variation in the dependent variable. The root mean square error is 5.68, so the standard deviation of the predicted values from the observed y is about 5.68. This is barely smaller than the standard deviation of y itself, which is 5.78.

c. Now run a regression of `rating_diff` that also includes `ideology` – a variable in which respondents rate themselves very conservative (1) to very liberal (5) on a five–point scale. (For the moment, treat ideology as an interval-level variable. Note that some recoding may be necessary.) We'll call this Model II.

i. In a few sentences, explain what happens to the coefficient on female between Model I and Model II and your substantive interpretation of this change.

The coefficient becomes smaller, which reflects that women are more likely to identify as liberal (although the correlation between `female` and `ideology` is only $\rho = .0483$.

ii. How well does Model II explain variation in the dependent variable compared to Model I?

Definitely better; the r^2 is now .23 and the the root mean square prediction error has reduced to 5.02.

iii. What does Model II predict is the difference in `rating_diff` between those who are very conservative and those who are very liberal, holding the other factors constant? Approximately what percentage of the range of the dependent variable is this difference equal to?

The difference is 10.59, which is about 50.4% of the variation in `rating_diff`.

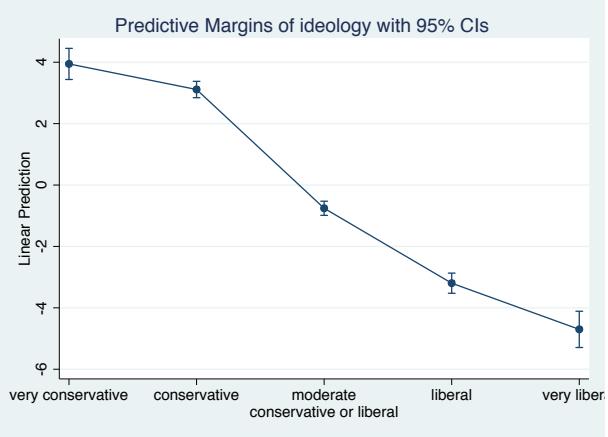
iv. What do your answers to (ii) and (iii) suggest about the importance of ideology in explaining the dependent variable compared to the other variables in Model II?

Not surprisingly, ideology is a very significant predictor of feeling thermometer ratings for Bush vs. Kerry. Changing from the lowest to the highest category is associated with more than 8 percentage points change in `rating_diff`, while switching from male to female is only associated with half a percentage point difference, while income is only associated with less than four percentage points difference, and age even less (although the coefficient on age may understate the case because the effect could be non-linear.

d. Finally, run the estimation in a way that does not require that we assume ideology is an interval–level variable. Call this equation Model III.

i. Now think carefully: what would we need to see in Model III that would give us confidence in treating `ideology` as an interval–level variable in the present context? Do we see this here?

We would like to see that the OLS assumption that the dependent variable is a linear function of the regressors. One way to check this is to dummy out the ideology categories, run a regression, and look at the `marginsiplot`. Using this approach, we see that, while not perfect, the assumption of linearity is not far off:



ii. In a few sentences, say what Model III tells us about the ceteris paribus relationship between ideology and `rating_diff`.

Moving from category 1 to category 5 of ideology is associated with an 8.6 point increase in one's relative liking for Bush compared to Kerry, controlling for sex, age and income. This is clearly a substantial and clear relationship. At the same time, the difference in feeling for Bush vs Kerry does not reduce to differences captured by the ideology question, since this captures less than one-quarter of the observed variation.

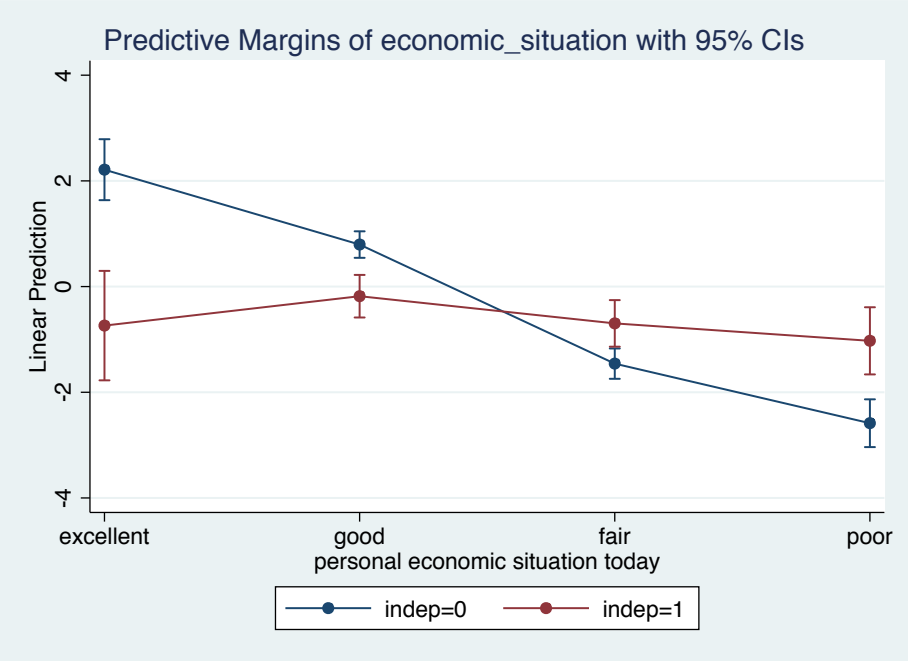
Question 4. Explore the following question using OLS. (25 points)

a. We expect that voters who rated their personal economic situation as poor in 2004 would be likely to blame the incumbent (Bush), and thus have lower values of `rating_diff` than those in better economic circumstances. However, we might theorize that the ceteris paribus association between a voter's personal economic situation and `rating_diff` is even stronger for Independent voters (who are not affiliated with either the Democratic or Republican parties), as these Independents do not have the cue of party identification to rely upon when rating the two candidates.

b. To explore this question, use the variables `pid` (a categorical variable that is a measure of voters' party identification) and `economic_situation` (a variable measuring how voters rate their personal economic situation on a scale of 1 (excellent) to 4 (poor)). For purposes of this question, you may treat this variable as interval–level.

c. From here, you're on your own. You will need to create new variables, estimate the appropriate model, present results both in tabular and graphical form, and interpret the results. Do your estimates confirm the theory? Make sure to describe your results in plain English.

Rather than assuming linearity of age or income, I use dummies for deciles of each of these variables. I interact the `independent` dummy with the `economic_situation` dummies. We see that the difference in economic situation is associated with different levels of rating Bush vs. Kerry for non-independents, and essentially unrelated for independents. This may be as the question suggests that independents do not rely on party cues, or are less informed.



	(1)
2.personal economic situation today	-1.335***
	(.312)
3.personal economic situation today	-3.299***
	(.333)
4.personal economic situation today	-4.685***
	(.394)
indep	-2.084***
	(.623)
2.economic_situation *indep	1.790**
	(.668)
3.economic_situation * indep	2.975***
	(.683)
4.economic_situation * indep	4.531***
	(.745)
2.10 quantiles of age	-.155
	(.319)
3.10 quantiles of age	-.536
	(.336)
4.10 quantiles of age	1.142***
	(.325)
5.10 quantiles of age	-.306
	(.333)
6.10 quantiles of age	-.207
	(.324)
7.10 quantiles of age	-.419
	(.319)
8.10 quantiles of age	.706*
	(.336)
9.10 quantiles of age	-.507
	(.329)
10.10 quantiles of age	-.526
	(.335)
sex (0 = male, 1 = female)	-.350*
	(.147)
10 quantiles of ln_inc	.003
	(.031)
2.conservative or liberal	-.837**
	(.290)
3.conservative or liberal	-4.361***
	(.284)
4.conservative or liberal	-6.871***
	(.309)
5.conservative or liberal	-8.274***
	(.396)
N	4602
r^2	0.28
rmse	4.87