# 1 Recitation Feb 16,2016 Note

## 1.1 Colinearity

Perfect Multicollinearity is when one covariate (independent variable) can be expressed as a linear combination of the other covariates. Mathematically, for the example of 3 covariates:
Say our model is of the form:

$$Y_i = \beta_0(1) + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + u_i$$

This model exhibits perfect multicollinearity if

$$\exists(\alpha_0, \alpha_2, \alpha_3) \in R^3 s.t. X_{1,i} = \alpha_0(1) + \alpha_2 X_{2,i} + \alpha_3 X_{3,i} \forall i$$

In simpler math terminology, the model exhibits perfect multicollinearity there exists constant values $\alpha_0, \alpha_2$, and $\alpha_3$, where for every observation in our dataset, if we are given all covariates but one, we can construct the last covariate by a set linear function, say $X_{1,i} = \alpha_0(1) + \alpha_2 X_{2,i} + \alpha_3 X_{3,i}$.

An (important) aside: Note that I am thinking about the intercept as a covariate here. This is why removing the intercept can fix perfect multicollinearity – we no longer can include $\alpha_0$ in our linear combination formula when the intercept is left out.

Examples using the above formula:
1) Repeating a covariate: $X_{1,i} = X_{2,i}$
Perfect multicollinearity, as $\alpha_0 = 0, \alpha_2 = 1, \alpha_3 = 0$ satisfies the linear combination equation.
1) You have 100 dollars you must spend on three goods, and the Xs are the amount out of that sum you spend on each of the goods: $X_{1,i} + X_{2,i} + X_{3,i} = 100$
Perfect multicollinearity, as $\alpha_0 = 100, \alpha_2 = -1, \alpha_3 = -1$ satisfies the linear combination equation.

How do we solve it?
Two methods:
1) Remove one of the covariates that is causing the problem: $Y_i = \beta_0(1) + \beta_1 X_{1,i} + \beta_2 X_{2,i} + u_i$.
2) Remove the intercept (if the intercept is needed for the linear combination): $Y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + u_i$

Be careful about interpretation in these regressions! You have to think first: what is the implied value of the "left out variable"? Then, when we interpret $\beta$ as the expected effect of increasing the covariate by one unit, it is actually the expected effect of increasing the associated covariate by one unit and changing the "left-out" covariate by as many units as is implied by the linear combination between the covariates.

## 1.2 Omitted Variable Bias

First, think about the set up:
The True model is:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \gamma X2, i + \epsilon_i$$

with $\epsilon_i$ iid. Or in other words,

$$Y_i = \beta_0 + \beta_1 X_{1,i} + u_i$$

with

$$u_i = \gamma X_{2,i} + \epsilon_i$$

We run this model with only $X_1$. What happens? $\hat{\beta_1}$ is biased! Specifically, in this case with just one covariate and one omitted variable,

$$\hat{\beta_1} \to \beta_1 + corr(X_{1,i}, u_i)\sqrt{\frac{var(u_i)}{var(X_{1,i}}}$$

Now, $\sqrt{\frac{var(u_i)}{var(X_i}} > 0$ since variances are always positive. So the sign of the bias only depends the sign of $Corr(X_{1,i}, u_i)$. To work this out:

$$Corr(X_{1,i}, u_i) = Corr(X_{1,i}, \gamma X_{2,i} + \epsilon_i)$$

$$= Corr(X_{1,i}, \gamma X_{2,i})$$

by $\epsilon$ iid.

$$= \gamma \frac{Cov(X_{1,i}, \gamma X_{2,i})}{\sqrt{Var(X_{1,i})Var(\gamma X_{2,i})}}$$

$$= \frac{\gamma Cov(X_{1,i}, X_{2,i})}{\sqrt{\gamma^2 Var(X_{1,i})Var(X_{2,i})}}$$

$$= \frac{\gamma}{|\gamma|} \frac{Cov(X_{1,i}, X_{2,i})}{\sqrt{Var(X_{1,i})Var(X_{2,i})}}$$

$$= sign(\gamma)Corr(X_{1,i}, X_{2,i})$$

So the sign of the bias is equal to the "sign of the effect of the omitted variable on the dependent variable" times the "sign of the correlation between the relevant dependent variable and the omitted variable". This generalizes to cases with many variables.

### 1.3 R-squared and Adjusted R-squared

$R^2$ is the fraction of the variation in $Y_i$ explained by the model.

$$R^2 = 1 - \frac{SSR}{TSS}$$

$$\bar{R^2} = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

Also note, it can be derived from these equations that

$$\bar{R^2} = R^2 - \frac{k}{n-k-1}(1-R^2)$$

[1] So the adjusted R squared ($\bar{R^2}$) is the R squared minus a penalty term for having too many variables relative to the number of observations. Note that as $n \to \infty$, the penalty term goes to zero.

Why do we use adjusted R squared then? One way to think about it is that $R^2$ will increase if we add a totally useless variable. Say true model is $Y_i = \beta_0 + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_1)$, i.e. Y is just a normal distributed around a mean. We however, run $Y_i = \beta_0 + \beta_1 X_{1,i} + \epsilon_i$ where $X_{1,i} \sim N(0, \sigma_2)$. $\beta_1 = 0$, but by the randomness in the data, $\hat{\beta_1} \neq 0$. This will generate an $R^2 > 0$, where in fact, our model is explaining none of the true variation.

### 1.4 F-tests

This is section is written on pen and paper and attached.

[1] Ask me if you want the derivation– I have it on paper