Problem 1:

i. False. The interpretation of the coefficients depends on assumptions , for example, the exogeneity assumption. The regressors are not correlated with the error term. In addition, there are other factors that are not included in the regression model.

ii. True, both t-tests and Wald tests rely on certain assumptions about the distribution of the error term u_i in a linear regression model. Specifically, these tests assume that the error term u_i is independently and identically distributed with a normal distribution, such that $u_i \sim N(0, \sigma^2)$.

iii. False. For such small sample size which may not meet the assumption that the samples are normally distributed and have equal variances.

iv. False. If the new variables improve the fit of the model, then the adjusted-R2 will increase, but if the new variables do not improve the fit of the model, then the adjusted-R2 will decrease.

v. Flase. The impact of outliers on the $\hat{\beta}$ estimator depends on the degree of influence that the outliers have on the regression line and the sample size. If the sample size is small or the outliers have a significant influence on the regression line, the beta hat estimator may not be consistent.

vi. True. When there is perfect collinearity, the OLS estimator beta hat is not well-defined and the matrix X'X becomes singular, meaning that its inverse does not exist. As a result, the OLS estimator is not consistent, and it does not converge to the true population parameter as the sample size increases.

vii. True. The matrix X'X is nearly singular, which means that the inverse of X'X exists but is very large. The OLS estimator beta hat can be very sensitive to small changes in the data, and a small perturbation in the data can lead to large changes in the estimates of the regression coefficients. $\hat{\beta}$ is biased because it may not converge to the true population.

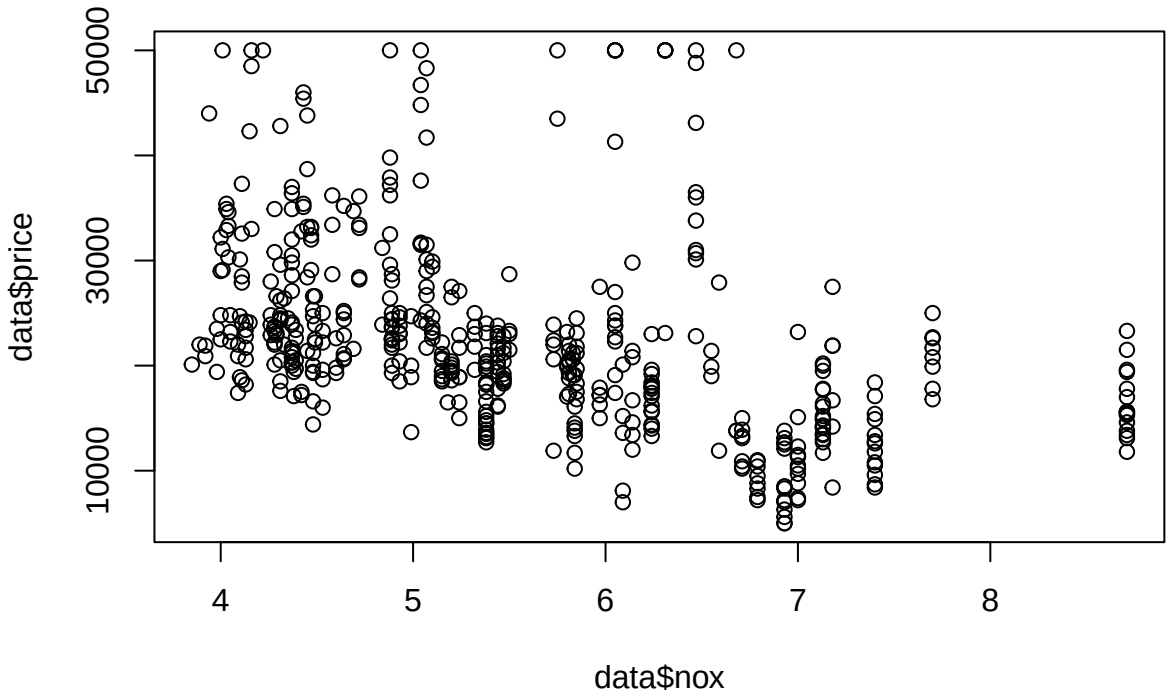
viii. False. If the outliers have infinite fourth moments, the OLS estimator can be biased towards the direction of the outliers. This is because the OLS estimator is based on minimizing the sum of squared errors, and the presence of outliers can cause the estimator to overemphasize the importance of these extreme observations.

Problem 2

i.

```
# package to open data set
library ( foreign )

# open data set
data = read.dta('hprice2.dta')
plot(x = data$nox,
     y = data$price)
```



The scatter plot shows a linear relation ship between *nox* and *price*, thus 1st LSA holds for model 1. There are no large outliers so the 3rd LSA holds for model 1.

ii.

```
model = lm(price ~ nox, data)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ nox, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13682  -5104  -2160   2969  31317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41307.8     1816.2   22.74   <2e-16 ***
## nox          -3386.9      320.4  -10.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8340 on 504 degrees of freedom
## Multiple R-squared:  0.1815, Adjusted R-squared:  0.1799
## F-statistic: 111.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

We can see that the regression model is:

$$price = 41307.8 - 3386.9 * nox$$

```
# Load libraries
library("lmtest")
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library("sandwich")

# Robust t test
coeftest(model, vcov = vcovHC(model, type = "HC0"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 41307.81    1585.86  26.048 < 2.2e-16 ***
## nox         -3386.85     284.36 -11.911 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The heteroskedasticity-robust (HC) standard errors of the intercept is 41307.81 and -3386.85 of the nox. The effect of nox is negative. As we can see from the scatter plot, the effect of nox is negative. And when nox is less 7000, the uncertainty between the two variables is high, thus the model is heteroskedasticity.

iii. As we can see from the results of t-test, the effect of nox on price statistically significant at the 5% and 1% significance level.

```
confint(model,'nox',level=0.95)
```

```
##         2.5 %    97.5 %
## nox -4016.262 -2757.443
```

At 95% confidence interval for β_1, the range is -4016.262 to -2757.443, which is -3386.852 ± 629.410.

iv. Yes, potential omitted variables in OLS can be a concern because they can lead to biased and inconsistent estimates of the regression coefficients. The omitted variable is negatively correlated with the included variable, then the coefficient of the included variable may be underestimated.

v. Fit model 2:

```
model_multi = lm(price ~ nox + rooms + dist + crime + proptax, data)
summary(model_multi)
```

```
##
## Call:
## lm(formula = price ~ nox + rooms + dist + crime + proptax, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16218  -3183   -749   2406  39499
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9060.30    3978.87  -2.277 0.023202 *
## nox         -1737.66     410.78  -4.230 2.78e-05 ***
## rooms        7707.33     399.08  19.313  < 2e-16 ***
## dist         -791.26     197.94  -3.997 7.37e-05 ***
## crime        -150.07      38.12  -3.937 9.41e-05 ***
## proptax       -89.96      23.62  -3.809 0.000157 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5938 on 500 degrees of freedom
## Multiple R-squared:  0.5883, Adjusted R-squared:  0.5842
## F-statistic: 142.9 on 5 and 500 DF,  p-value: < 2.2e-16
```

```
# Robust t test
coeftest(model_multi, vcov = vcovHC(model_multi, type = "HC0"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -9060.303   5366.858 -1.6882 0.0919974 .
## nox         -1737.660    387.347 -4.4861 9.010e-06 ***
## rooms        7707.327    666.642 11.5614 < 2.2e-16 ***
## dist         -791.259    174.699 -4.5293 7.409e-06 ***
## crime        -150.070     30.271 -4.9575 9.795e-07 ***
## proptax       -89.957     26.688 -3.3707 0.0008079 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The heteroskedasticity robust standard error for nox is 387.347. The heteroskedasticity robust standard error for rooms is 666.642. The heteroskedasticity robust standard error for dist is 174.699. The heteroskedasticity robust standard error for crime is 30.271. The heteroskedasticity robust standard error for rooms is 26.688.

```
confint(model_multi,'nox',level=0.95)
```

```
##        2.5 %    97.5 %
## nox -2544.72 -930.5992
```

95% interval for β_1 is (-2544.72, -930.5992). The estimated effect of nox is smaller than the results in ii-iii.

vi. R-square and adjusted R-square

```
# rquare for model 1
print(summary(model)$r.squared)
```

```
## [1] 0.1815076
```

```
# rquare for model 2
print(summary(model_multi)$r.squared)
```

```
## [1] 0.5883483
```

```
# adjusted rquare for model 1
print(summary(model)$adj.r.squared)
```

```
## [1] 0.1798836
```

```r
# rquare for model 2
print(summary(model_multi)$adj.r.squared)
```

```
## [1] 0.5842318
```

The R^2 for model 1 is 0.182, and the R^2 for model 2 is 0.588. The $\bar{R}^2$ of model 1 is 0.180, and the $\bar{R}^2$ of model 2 is 0.584. Model 2 is better because both R^2 and $\bar{R}^2$ is closer to 1.

vii.

```r
-5.5 * confint(model_multi,'nox',level=0.95)
```

```
##        2.5 %   97.5 %
## nox 13995.96 5118.296
```

At 95% confidence interval, decreasing nox by 5.5 will cause the price to increase by 5118.296 to 13995.96.

viii.

```r
-3.5 * confint(model_multi,'crime',level=0.95)
```

```
##           2.5 %   97.5 %
## crime 787.3495 263.1427
```

At 95% confidence interval, decreasing nox by 5.5 will cause the price to increase by 263.1427 to 787.3495. I would prefer reducing pollution than crime, because it will cause more increment of the price.

ix.

```r
(3.5 - 5.5) * confint(model_multi, "nox", level = 0.95) + (70-40) * confint(model_multi, "proptax", lev
```

```
##        2.5 %   97.5 %
## nox 998.7865 554.4215
```

At 95% confidence level, the housing price will increase by 554.4215 to 998.7865.

x.The null hypothesis in multiple linear regression can be written in the usual way as:

$$H0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

Where β_2 is the slope for rooms, $\beta_3$is the slope for dist, β_4 is the slope for crime, β_5 is the slope for proptax.

In matrix notation, the null hypothesis is:

$$\begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix} * \begin{vmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{vmatrix} = \begin{vmatrix} 0 \\ 0 \\ 0 \\ 0 \end{vmatrix}$$

Here q equals 4 because the number of coefficients that are being tested is 4.

xi.

```r
library(aod)
wald.test(Sigma = vcov(model_multi), b = coef(model_multi), Terms = 2:5)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 449.6, df = 4, P(> X2) = 0.0
```

According to Wald test, the P value is 0, which means we can reject H0 at the 95% confidence level.

xii. At the 4 degrees of freedom (df=4), the critical value for rejecting the null hypothesis at a significance level of 0.01 is 13.3. Thus, there is no level in the table at which we cannot reject H_0 for this particular chi-square value and degrees of freedom. This suggest p-value is 0, indicating strong evidence against the null hypothesis.