The data for this assignment comes from Wave I of the National Longitudinal Study of Adolescent to Adult Health and is stored in a csv file called **depress2.csv**. We are interested in the association between parental receipt of public assistance and depressive symptom index (from the Center for Epidemiologic Studies Depression Scale or CES-D). The variables are described below. Perform a multiple linear regression predicting depressive symptom index from public assistance, adjusting for age.

Variable Name	Description	Coding/Unit	Type
AID	Subject ID		Numeric
age	Subject's age	Years	Numeric
cesd	Depressive symptom index (CES-D)	Symptom count	Numeric
cesd_16	Depressive symptom index category (≥16 vs. <16)	< 16 symptoms (low) ≥ 16 symptoms (high)	Numeric
publicassist	Parental receipt of public assistance	0=No 1=Yes	Numeric

Preparation: Summary data and visually examination on the relationship.

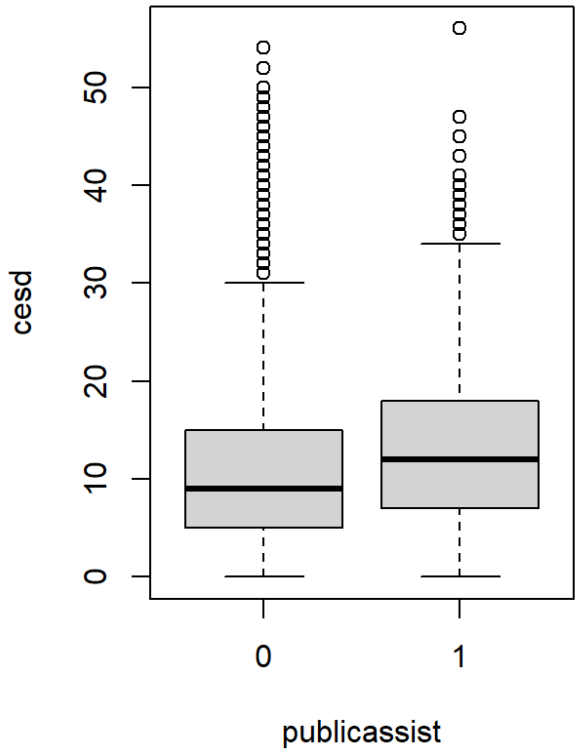
1. Create appropriate summary statistics/sample characteristics to summarize variables (age and cesd only) in this dataset stratified by public assistance status. Then comment on your observations.
2. Generate appropriate plot to show the relationship between depressive symptom index and public assistance and then comment on your observations based on the plot.

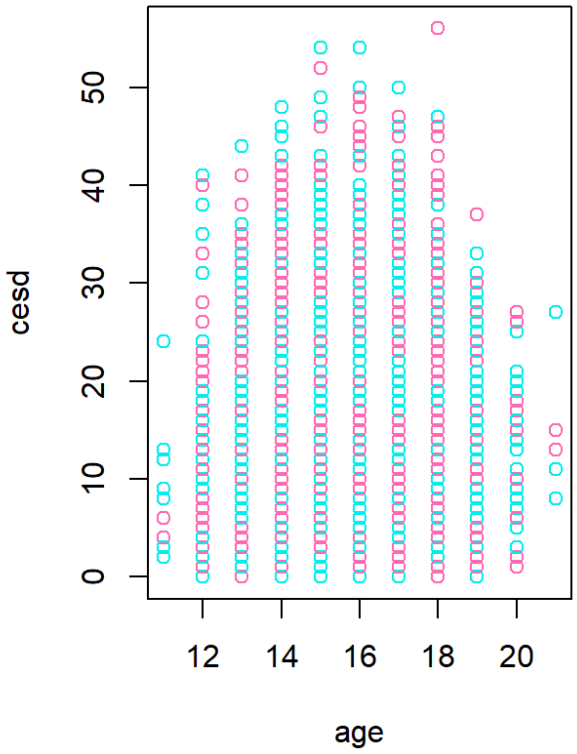
Summary Statistics

Subjects whose parents did not receive public assistance had a mean age of 15.6 (range of 11-21) and mean depressive symptom index score of 10.86 (range of 0-54). Subjects whose parents received public assistance had a mean age of 15.45 (range of 11-20) and mean depressive symptom index score of 13.34 (range of 0-56). From the summary statistics alone, the mean ages appear to be the same between both groups and the mean depressive symptom index score seem to be higher for the subjects whose parents received public assistance. Further hypothesis testing would need to be done in order to come to a definite conclusion.

Boxplot and Scatterplot

The boxplots comparing subjects whose parents received public assistance and subjects whose parents did not receive public assistance show that the depressive symptom index scores between the two groups have approximately the same distribution. The mean depressive symptom index score is slightly higher for those whose parents received public assistance, but the boxplots overlap so a two-sample t-test most likely will not reject the null hypothesis of there being no difference in mean depressive symptom index score between the two groups.





Hypothesis Testing: raw comparison and model building

3. Before going forward with multiple linear regression, perform the two-sample t-test first to determine whether depressive symptom index are the same based on public assistance and then comment on your results.

Two-Sample Two-Tailed T-Test

A two-sample two-tailed t-test was used to test whether there was a difference in mean depressive symptom index score between subjects whose parents received public assistance and subjects whose parents did not receive public assistance. The t-score was =11.615 with 1868.7 degrees of freedom and resulting p-value was <0.0001. With a p-value less than the α=0.05 significance level, the null hypothesis is rejected. There is evidence suggesting that subjects whose parents received public assistance had a higher mean depressive symptom index score (13.3448) than subjects whose parents did not receive public assistance (10.8602).

4. Conduct a multiple linear regression predicting depressive symptom index from public assistance, adjusting for age.
    a. Write a complete report for the **global hypothesis test** for the built model including R2.
    b. Write a complete report for the effect of public assistance, adjusting for age.
    c. Write a complete report for the effect of age, adjusting for public assistance.
    d. What is the predicted value of CESD for a 14 year old subject whose parent receives public assistance?

Multiple Linear Regression

a) A multiple linear regression analysis was used to test whether a subject's depressive symptom index score is linearly associated with the two predictor variables: age and parents' public assistance status. The F-statistic was 168.6 with 2 and 16169 degrees of freedom, and the resulting p-value was < 0.001. With a p-value less than the α=0.05 significance level, the null hypothesis of there being no linear association between depressive symptom index score and the two regressors was rejected. The coefficient of determination was 0.0204, meaning that the predictors age and public assistance status accounted for 2.04% of the variability in the depression scores in this dataset. There is evidence suggesting that the linear association is

$$cesd = 0.4553age + 2.5543publicassist + 3.7553$$

b) An ANCOVA test was used to test whether a subject's depressive symptom index score is linearly associated with their parents' public assistance status, adjusting for age. The F-statistic was 166.5778 with 1 and 16169 degrees of freedom, and the resulting p-value was <0.0001. With a p-value less than the α=0.05 significance level, the null hypothesis of there being no linear association between depressive symptom index score and public assistance status was rejected. There is evidence suggesting that the linear association between a subject's depressive symptom index score and their parents' public assistance status is

$$cesd = 0.4553\overline{age} + 2.5543publicassist + 3.7553$$

c) An ANCOVA test was used to test whether a subject's depressive symptom index score is linearly associated with their age, adjusting for their parents' public assistance status. The F-statistic was 179.4393 with 1 and 16169 degrees of freedom, and the resulting p-value was <0.0001. With a p-value less than the α=0.05 significance level, the null hypothesis of there being no linear association between depressive symptom index score and age was rejected. There is evidence suggesting that the linear association between a subject's depressive symptom index score and their age is

$$cesd = 0.4553age + 2.5543\overline{publicassist} + 3.7553$$

d) The predicted CESD score of a 14-year old subject whose parents receive public assistance is 12.6844.

$$cesd = 0.4553age + 2.5543\overline{publicassist} + 3.7553$$
$$cesd = 0.4553(14) + 2.5543(1) + 3.7553 = 12.6844$$

5. Calculate the age-adjusted least square means for CESD in both ways: manually and with R

Least Squares Mean

The age-adjusted least square means of depressive symptom index score for subjects whose parents do not receive public assistance is 10.8544. The age-adjusted least square means of depressive symptom index score for subjects whose parents receive public assistance is 13.40772.

6. Conduct a multiple linear regression predicting depressive symptom index from public assistance, age and interaction between public assistance and age.
    a. Based on the model, write down the fitted regression line for the people with public assistance and without public assistance, separately.
    b. Perform the hypothesis test to examine whether the interaction should be included in the model.

Interaction Between Public Assistance Status and Age

General Model with Interaction

$$cesd = 3.70696 + 0.4584age + 3.0671publicassist - 0.0332age \times publicassist$$

No Public Assistance

publicassist = 0

$$cesd = 3.70696 + 0.4584age$$

Public Assistance

publicassist = 1

$$cesd = (3.70696 + 3.0671) + (0.4584 - 0.0332)age$$
$$cesd = 6.7767 + 0.42528age$$

Hypothesis Test

A multiple linear regression analysis was used to test whether a subject's depressive symptom index score is linearly associated with the interaction between age and parents' public assistance status. The resulting p-value was 0.7765. With a p-value less than the α=0.05 significance level, the null hypothesis of there being no linear association between depressive symptom index score and the interaction term was not rejected. There is insufficient evidence to conclude that the interaction term should be kept in the multiple linear regression model.

Final Summary

A multiple linear regression analysis was used to test whether a subject's depressive symptom index score is linearly associated with the two predictor variables: age and parents' public assistance status. The F-statistic was 168.6 with 2 and 16169 degrees of freedom, and the resulting p-value was < 0.001. With a p-value less than the α=0.05 significance level, the null hypothesis of there being no linear association between depressive symptom index score and the two regressors was rejected.

ANCOVA tests were used to test whether a subject's depressive symptom index score is linearly associated with each regressor individually while adjusting for the other covariable. The F-statistic for the parents' public assistance status, adjusting for age, was 166.5778 with 1 and 16169 degrees of freedom, and the resulting p-value was <0.0001. The F-statistic for age, adjusting for their parents' public assistance status, was 179.4393 with 1 and 16169 degrees of freedom, and the resulting p-value was <0.0001. With p-values less than the α=0.05 significance level, the null hypothesis of there being no linear association between depressive symptom index score and each regressor was rejected.

Lastly, a multiple linear regression analysis was used to test whether a subject's depressive symptom index score is linearly associated with the interaction between age and parents' public assistance status. The resulting p-value was 0.7765. With a p-value greater than the α=0.05 significance level, the null hypothesis of there being no linear association between depressive symptom index score and the interaction term was not rejected. There is insufficient evidence to conclude that the interaction term should be kept in the multiple linear regression model. There is evidence suggesting that the linear association is

$$cesd = 0.4553age + 2.5543publicassist + 3.7553$$