Stat 500 Exam 1 Solutions

Fall Semester - 2015

This study was undertaken in the context of proposals for a guaranteed annual wage (negative income tax). The data are from a national sample of 6000 households with a male head of household and earnings of less than $15,000 annually in 1966. Some demographic subgroups were formed for analysis of the relationship between average hours worked during the year and average hourly wages and other variables. Each data point represents one demographic subgroup.

The variables used were

- **HRS**: Average hours worked during the year (in hours)
- **WAGE**: Average hourly wage (in $)
- **ASSET**: Average family asset holdings (bank accounts, etc.) (in $)
- **AGE**: Average age of respondent (in years)

Here is a summary of the variables:

```
      HRS            WAGE           ASSET            AGE       
 Min.   :1985   Min.   :1.423   Min.   : 1370   Min.   :22.40  
 1st Qu.:2096   1st Qu.:2.505   1st Qu.: 4420   1st Qu.:38.55  
 Median :2134   Median :2.905   Median : 7250   Median :39.20  
 Mean   :2137   Mean   :2.772   Mean   : 6265   Mean   :39.35  
 3rd Qu.:2186   3rd Qu.:3.010   3rd Qu.: 7832   3rd Qu.:39.95  
 Max.   :2267   Max.   :3.636   Max.   :12710   Max.   :57.70  
```

Here is a regression summary output:

```
Call:
lm(formula = HRS ~ WAGE + ASSET + AGE)

Residuals:
    Min      1Q  Median      3Q     Max 
-89.612 -15.789   5.388  20.963  74.645 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2444.78795   93.62216  26.113  < 2e-16 
WAGE         -47.61368   23.00636  -2.070   0.0459 
ASSET          0.02641    0.00393   6.720 8.80e-08 
AGE           -8.66277    1.70598  -5.078 1.27e-05 

Residual standard error: 35.61  on 35 degrees of freedom
Multiple R-Squared: 0.715     Adjusted R-squared: 0.6906 
F-statistic: 29.28 on ??? and ??? DF   p-value: 1.184e-09 

Correlation of Coefficients:
      (Intercept) WAGE  ASSET
```

```
WAGE  -0.81
ASSET  0.73       -0.83
AGE   -0.89        0.48 -0.60
```

1. Find the values of n and RSS.

   $n = 39$, RSS $= \hat{\sigma}^2 * 35 = 44382.52$.

2. Fill in the missing degrees of freedom for the F-statistics, and state the hypothesis the F-statistic in the summary refers to.

   3 and 35. The null hypothesis is $\beta_{WAGE} = \beta_{ASSET} = \beta_{AGE} = 0$.

3. How do you expect the R^2 to change if `WAGE` is dropped from the model? Compare to the effect of dropping `AGE`.

   R^2 will decrease. It is likely to decrease more when AGE is dropped than when WAGE is dropped, because AGE is much more significant than WAGE.

4. From the correlations of coefficients given above, what can you conclude about the correlation between the variables `AGE` and `ASSET`?

   Most likely positive (the opposite of correlation between coefficients).

5. A 95% confidence region for the regression coefficients for `WAGE` and `AGE` is shown in the left panel of Figure 1 (refer to the last page of the test). Give the exact coordinates of the center of the ellipse.

   (-47.61368, -8.66277).

6. The origin lies way outside the ellipse. The knowledge that the origin is outside the ellipse gives the result of a hypothesis test comparing two linear models. State the two **models** (not hypotheses) and the result of the test.

   Model 1 (H_0): $HRS = \beta_0 + \beta_{ASSET} ASSET + \epsilon$
   Model 2 (H_A): $HRS = \beta_0 + \beta_{ASSET} ASSET + \beta_{WAGE} WAGE + \beta_{AGE} AGE + \epsilon$
   Reject H_0 in favor of H_A.

7. A Q-Q plot of the residuals for the model is shown in the right panel of Figure 1. How would you describe the distribution of the residuals: close to normal, long-tailed, short-tailed, skewed (if so, which side)?

   The distribution is skewed, with a long left tail.

8. It is difficult to measure the `ASSET` accurately and we may assume that there is substantial error (but no bias) in the measured values. Suppose some new and more accurate method of retroactively measuring *ASSET* is developed and we replace the old *ASSET* with this new *ASSET*. Would the coefficient of *ASSET* in this new model tend to be smaller, the same as or larger than 0.02641 or there is no way of knowing?

   Errors in predictors tend to shrink coefficients towards zero, so removing the error will likely give a larger coefficient.

9. A diagnostic plot is shown in the left panel of Figure 2. Does it indicate any violations of the model assumptions? If so, which ones?

   Yes, it indicates a violation of the constant error variance assumption (heteroscedasticity).

10. The 4th subgroup in the data set had the following values:

```
   HRS  WAGE  ASSET  AGE
4 2111 2.511  1632  22.4
```

What can you say about its leverage? Explain.

This subgroup is likely to have high leverage because AGE is at the minimum value.

11. The 19th subgroup had the largest (in absolute value) externally studentized residual of -3.45. Can you formally conclude this subgroup is an outlier? You may use the fact that

```
> pt(-3.45,df=35)
[1] 0.0007399137
```

The two-sided test p-value $= 2 \times 0.00074 = 0.00148$, and $\alpha/n = 0.05/39 = 0.00128$. Since p-value $> \alpha/n$, with Bonferroni's correction we cannot reject the hypothesis this subgroup is not an outlier at level $\alpha = 0.05$.

12. Suppose one left out the 19th subgroup and computed estimates of β_{WAGE} and β_{ASSET} being $\hat{\beta}_{(19),WAGE} = -41.52$ and $\hat{\beta}_{(19),AGE} = .0251$, respectively. For this model, less the 19th subgroup, is it likely that either predictor WAGE and/or predictor ASSET would not be statistically significant? Clearly explain your answer, and explain why you cannot be certain.

Assuming that the standard error does not change for each of the coefficients, we conjecture that the new t-statistics are

$$t_{WAGE} = -41.52/23.006 = -1.80$$

and

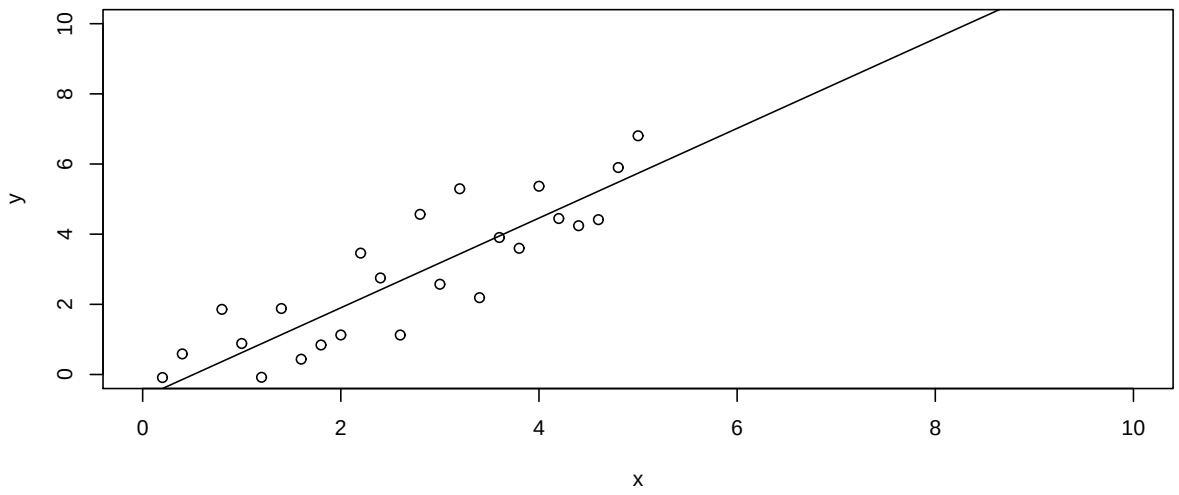
$$t_{ASSET} = .0251/.00393 = 6.39$$

and so since the first is less than 2, it seems that WAGE is likely to insignificant, and since the second is much larger, it seems that ASSET is likely to stay significant. We are uncertain because the standard errors for each coefficient will change as well.

13. The partial regression plot for `WAGE` is shown in the right panel of Figure 2. The least squares line for the points on this plot is also shown. What is the numerical value of the slope of this line? Which model assumption is this plot primarily designed to test? Are there any indications that this assumption is violated?

Slope=-47.61368. This plot is primarily designed to check the linear structure of the model. There are no obvious violations.

14. A dataset with its fitted regression line is shown below. Add two points to the plot: point A that is an outlier but does not have high leverage, and point B that has high leverage but is not an outlier.



There are many ways to answer this correctly. An example: A has coordinates (2,10) and B has coordinates (8,10).

15. Verify that $I - H$, where I is the identity and $H = X(X^\top X)^{-1}X^\top$ is the hat matrix, is a projection matrix. (You may use the fact that H itself is a projection matrix).

    Need to show that $I-H$ is symmetric, and that $(I-H)(I-H) = I-H$. But H is projection matrix, so it is symmetric and $HH = H$. So

    $$(I - H)^\top = I^\top - H^\top = I - H$$

    since I is clearly symmetric. Also,

    $$(I - H)(I - H) = I - H - H + HH = I - 2H + H = I - H$$

16. Suppose the error vector in a linear regression model has covariance matrix of $var(\epsilon) = \Sigma$. Derive the formula for the variance of the generalized least-squares (GLS) estimator $\hat{\beta}$, which is defined as $\hat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$.

$$\begin{aligned}
Var(\hat{\beta}) &= Var((X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y) \\
&= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Var(y)\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} \\
&= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Var(\epsilon)\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} \\
&= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\Sigma\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} \\
&= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} \\
&= (X^T\Sigma^{-1}X)^{-1} \ .
\end{aligned}$$

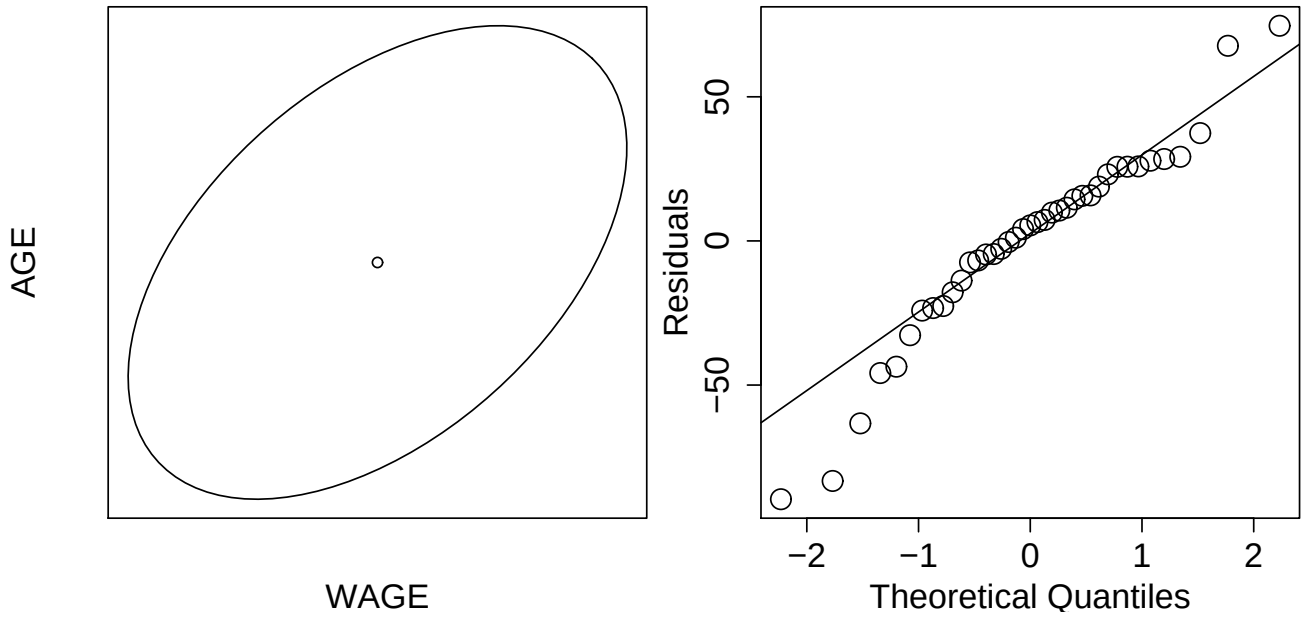


Figure 1: The left panel is used in questions 5 and 6, and the right panel is used in question 7.

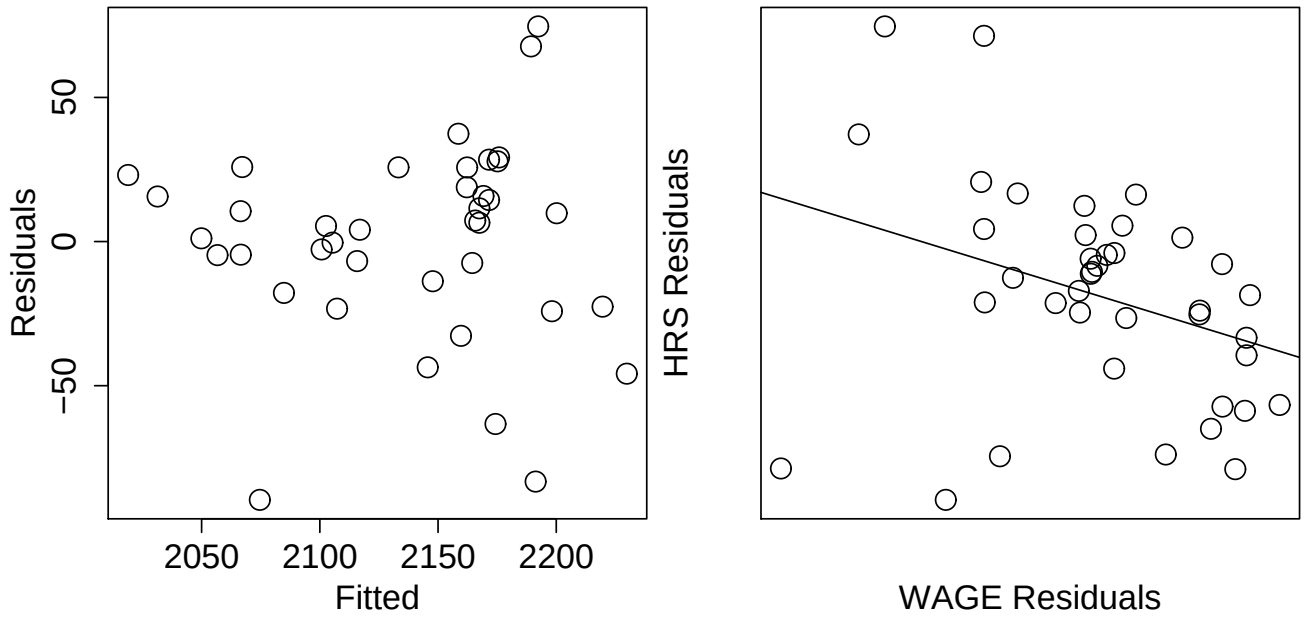


Figure 2: The left panel is used in question 9, and the right panel is used in question 13.