Home Assignment 3 Solutions

STAT 151A Linear Modelling: Theory and Applications

Ajay Sharma — Spring 2025

Problem 1: Understand the summary of OLS - Part 1
(No real data is required for this problem.) Consider a hypothetical dataset, `bodyfat.csv`, containing the following columns: `BODYFAT`, `AGE`, `WEIGHT`, `HEIGHT`, `WRIST`, and `THIGH`. I want to run some OLS models for fitting and predicting `BODYFAT`.

(a) A colleague of mine suggests the following model:

$$\begin{aligned}\text{BODYFAT} = \beta_0 + \beta_1\,\text{AGE} + \beta_2\,\text{WEIGHT} + \beta_3\,\text{HEIGHT}\\ + \beta_4\,(\text{WEIGHT} + 3\times\text{HEIGHT}) + \beta_5\,\text{WRIST} + e.\end{aligned}$$

Is this a good model to run? Why or why not?
Solution. It is not a good idea, because `WEIGHT`, `HEIGHT`, and `WEIGHT+3*HEIGHT` have perfect multicollinearity.

(b) I decide against including the variable `WEIGHT+3*HEIGHT` in the model, and instead fit

$$\text{BODYFAT} = \beta_0 + \beta_1\,\text{AGE} + \beta_2\,\text{WEIGHT} + \beta_3\,\text{HEIGHT} + \beta_4\,\text{WRIST} + e.$$

The following Python code fits this model:

```
X = bodyfat[["AGE", "WEIGHT", "HEIGHT", "WRIST"]]
X = sm.add_constant(X)
y = bodyfat["BODYFAT"]
model_M = sm.OLS(y, X).fit()
print(model_M.summary2())
```

The printed summary is:

```
Results: Ordinary least squares
=================================================================
Model: OLS   Adj. R-squared: 0.460
Dependent Variable: BODYFAT   AIC: 1595.0001
Date: ----------   BIC: 1605.5883
No. Observations: 252   Log-Likelihood: -794.50
Df Model: 2   F-statistic: 107.9
Df Residuals: 249   Prob (F-statistic): 1.83e-34
R-squared: 0.464   Scale: 32.449
```

```
-----------------------------------------------------------------
Coef.    Std.Err.    t    P>|t|    [0.025    0.975]
-----------------------------------------------------------------
const   -18.3739    2.5754   -7.1343   0.0000   -23.4464   -13.3015
AGE       0.1827    0.0285    6.4027   0.0000     0.1265    0.2389
WEIGHT    0.1627    0.0122   13.2984   0.0000     0.1386    0.1868
-----------------------------------------------------------------
Omnibus: 0.941   Durbin-Watson: 1.834
Prob(Omnibus): 0.625   Jarque-Bera (JB): 0.876
Skew: 0.144   Prob(JB): 0.645   Kurtosis: 2.984   Condition No.: 1340
=================================================================
```

(i) Using the reported R-squared value in the summary, calculate the RSS of the model m. (Recall that we already know the TSS.)
Solution. Using the R^2 formula, we can calculate the RSS as

$$\mathrm{RSS} = \mathrm{TSS} \times (1 - R^2) = 15079.0166 \times (1 - 0.464) \approx 8082.$$

(ii) Calculate the F-statistic for testing the model m against the model M. Determine the distribution of this test statistic, and calculate the corresponding p-value.
Solution. We've learned from the class that the F-statistic is

$$\frac{\mathrm{RSS}(m) - \mathrm{RSS}(M)}{p-q} \Big/ \frac{\mathrm{RSS}(M)}{n-p-1} \sim F_{p-q,\, n-p-1},$$

where p and q are the number of variables in model M and model m. Thus,

$$\frac{8082 - 6530.9933}{4-2} \Big/ \frac{6530.9933}{252-4-1} \approx 29.32927.$$

As the F-statistic follows $F_{2,247}$, the p-value is $P_{X \sim F_{2,247}}(X > 29.3297) \approx 0.00$.

Problem 2: Understand the summary of OLS - Part 2
(No real data is required for this problem.) Using the same dataset (`bodyfat.csv`) from Problem 1, consider a different linear model:

$$\text{BODYFAT} = \beta_0 + \beta_1\,\text{AGE} + \beta_2\,\text{WEIGHT} + \beta_3\,\text{HEIGHT} + \beta_4\,\text{THIGH} + e.$$

If X represents the design matrix (including an intercept) with columns ordered as intercept, AGE, WEIGHT, HEIGHT, THIGH, then Python provides that:

$$(X^\top X)^{-1} = \begin{pmatrix} 3.740212 & -5.908838\times10^{-3} & 6.662131\times10^{-3} & -3.218478\times10^{-2} & -4.048953\times10^{-2} \\ -5.908838\times10^{-3} & 3.238651\times10^{-5} & -1.222843\times10^{-5} & 3.416435\times10^{-5} & 7.148357\times10^{-5} \\ 6.662131\times10^{-3} & -1.222843\times10^{-5} & 2.632523\times10^{-5} & -4.483899\times10^{-5} & -1.292477\times10^{-4} \\ -3.218478\times10^{-2} & 3.416435\times10^{-5} & -4.483899\times10^{-5} & 3.866748\times10^{-4} & 1.944136\times10^{-4} \\ -4.048953\times10^{-2} & 7.148357\times10^{-5} & -1.292477\times10^{-4} & 1.944136\times10^{-4} & 7.873440\times10^{-4} \end{pmatrix}$$

Additionally, the regression summary is given as follows (with missing entries filled):

```
Results: Ordinary least squares
=================================================================
Model: OLS   Adj. R-squared: ---------
Dependent Variable: BODYFAT   AIC: ---------
Date: ----------   BIC: ---------
No. Observations: 252   Log-Likelihood: ---------
Df Model: 4   F-statistic: 71.01
Df Residuals: 247   Prob (F-statistic): ---------
R-squared: 0.535   Scale: ---------
--------------------------------------------------------------------
Coef.    Std.Err.    t    P>|t|    [0.025    0.975]
--------------------------------------------------------------------
const   -1.0742    10.3055    -0.1042    ------    ------    ------
AGE      0.1890     0.0303     6.2328    ------    ------    ------
WEIGHT   0.1237    0.0273     4.5256    ------    ------   ------
HEIGHT  -0.4607     0.1048    -4.3970    ------    ------    ------
THIGH    0.3655     0.1495     2.4443    ------    ------    ------
--------------------------------------------------------------------
```

(a) In the OLS model with intercept and p variables, the unbiased estimator of variance is given by $\hat{\sigma}^2 = \text{RSS}/(n-p-1)$. Using the provided information, compute $\hat{\sigma}^2$.
Solution. Calculating via the relation between standard errors and $(X^\top X)^{-1}$:

$$\frac{10.3055^2}{3.74021202} \approx 28.3950, \qquad \frac{0.0303^2}{3.23865148\times10^{-5}} \approx 28.3479.$$

The exact value of $\hat{\sigma}^2$ is 28.3952.

(b) Using the computed $\hat{\sigma}^2$, compute RSS and TSS of the model.
Solution. $\text{RSS} = 247 \times \hat{\sigma}^2 \approx 7013.61$. From R^2 and RSS, $\text{TSS} = 15079.0166$.

(c) Fill the five missing values (one in $(X^\top X)^{-1}$ and four in the summary).
Solution.

- Bottom-right entry of $(X^\top X)^{-1}$: 7.873440×10^{-4}.
- F-statistic: 71.01.
- Standard error of WEIGHT: 0.0273.
- t-value of WEIGHT: 4.5256.
- Coefficient of THIGH: 0.3655.

Problem 3: Leverage

Partition the design matrix $\widetilde{X} = [\mathbf{1}_n, X] \in \mathbb{R}^{n\times(p+1)}$, where X is the $n \times p$ matrix of non-intercept covariates, and $H_1 := \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$. The sample covariance of X is

$$S := \frac{1}{n-1}\sum_{i=1}^{n}(x_i^\top - \bar{X})(x_i^\top - \bar{X})^\top.$$

(a) Show that H_1 is symmetric and idempotent.
Solution. Symmetry is immediate. Then $H_1^2 = \frac{1}{n^2}\mathbf{1}_n\mathbf{1}_n^\top\mathbf{1}_n\mathbf{1}_n^\top = H_1$.

(b) Define $X_c = (I_n - H_1)X$. Show that $S = \frac{1}{n-1}X_c^\top X_c$.
Solution. The ith row of $(I_n - H_1)X$ is $x_i - \bar{X}^\top$. Summation yields $S = \frac{1}{n-1}X_c^\top X_c$.

(c) Show that $S = \frac{1}{n-1}X^\top(I_n - H_1)X$.
Solution. Using idempotence and symmetry, $X_c^\top X_c = X^\top(I_n - H_1)X$.

(d) Define $H_c = X_c(X_c^\top X_c)^{-1}X_c^\top$. Show that the ith diagonal of H_c is $h_{ii} - 1/n$.
Solution. Since $\widetilde{X}(\widetilde{X}^\top\widetilde{X})^{-1}\widetilde{X}^\top = H_1 + H_c$ and diagonals of H_1 are $1/n$, it follows.

Problem 4: Leave-one-out

Consider the linear model $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$, with ε_i iid $N(0, \sigma^2)$.

(a) Let $\hat{y}_i$ be the fitted value and $\hat{y}_{i(i)}$ the prediction omitting the ith observation. Write $\hat{y}_i - \hat{y}_{i(i)}$ in terms of ε_i and h_{ii}.

Solution.

$$\hat{y}_{i(i)} = \hat{y}_i - \frac{h_{ii}}{1 - h_{ii}}(y_i - \hat{y}_i) = \hat{y}_i - \frac{h_{ii}}{1 - h_{ii}}\varepsilon_i,$$

thus $\hat{y}_i - \hat{y}_{i(i)} = \frac{h_{ii}}{1-h_{ii}}\varepsilon_i$.

(b) Find the distribution of $\hat{y}_i - \hat{y}_{i(i)}$.

Solution. Since $\varepsilon_i \sim N(0, (1 - h_{ii})\sigma^2)$,

$$\hat{y}_i - \hat{y}_{i(i)} \sim N\Big(0, \sigma^2 \frac{h_{ii}^2}{1 - h_{ii}}\Big).$$

(c) Suggest an unbiased estimator of σ^2.

Solution. From (b), set $\frac{1-h_{ii}}{h_{ii}^2}(\hat{y}_i - \hat{y}_{i(i)})^2$.

Problem 5: Bootstrap confidence interval for bias
The bootstrap bias estimate is defined as

$$\hat{\text{bias}} = \bar{\theta}^* - \hat{\theta}.$$

Use $B = 1000$ replicates to estimate the bias of $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum(Y_i - \bar{Y})^2$ for $Y_i \sim N(0, 50)$, $n = 20$.
Solution. Refers to the `.ipynb` file; bias ≈ -2.39.