Regression Problems – Final Exam Review – Solutions

Stat061-F23

Prof Amanda Luby

Note: these may be harder/longer than what I would expect to give on the final, but wanted to give problems that would help with conceptual understanding and review. If you want more practice with mechanics, try some of the problems from Larsen & Marx or DeGroot & Schervish.

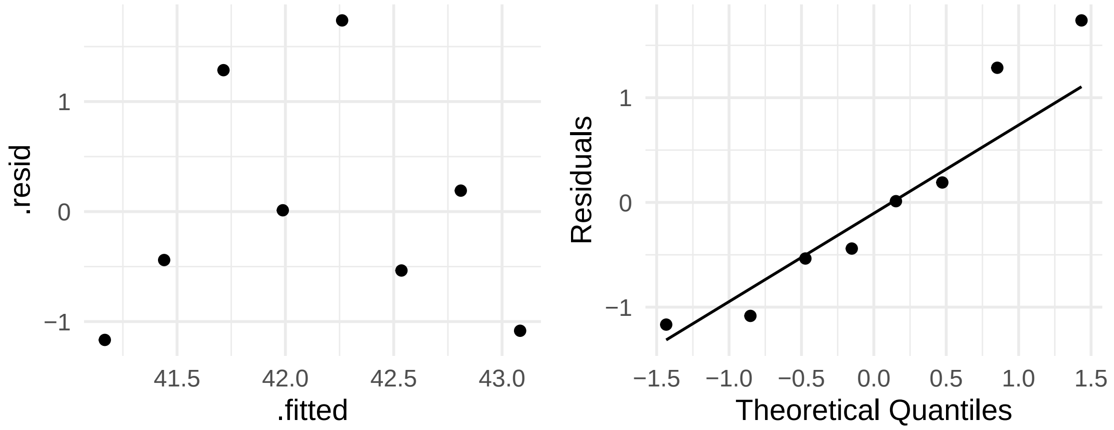
1. Consider a problem of simple linear regression in which the durability Y of a certain type of alloy is to be related to the temperature X at which it was produced. Eight specimens were produced at different temperatures and their durability was recorded. A linear regression model was fit to this data. The coefficient table and residual plots are shown below.

```
Call:
lm(formula = durability ~ temperature, data = durability_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1667 -0.6726 -0.2143  0.4643  1.7381 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  40.8929     0.8842  46.248 6.85e-09 ***
temperature   0.5476     0.3502   1.564    0.169    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.135 on 6 degrees of freedom
Multiple R-squared:  0.2895,	Adjusted R-squared:  0.1711 
F-statistic: 2.445 on 1 and 6 DF,  p-value: 0.1689
```



Using the R output above, answer the following questions:

(a) What are the values of the MLEs $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\sigma}^2$?

(b) Do you have any concerns about the assumptions of the linear model? State how you can tell.

(c) Construct a 95% confidence interval for β_1. Does this interval contain 0?

(d) Test at the $\alpha = .05$ level the hypothesis that the regression line passes through the origin in the x-y plane.

(e) First, note that $(X^T X)^{-1} = \begin{bmatrix} 0.61 & -.214 \\ -.214 & .095 \end{bmatrix}$. Find the covariance of $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

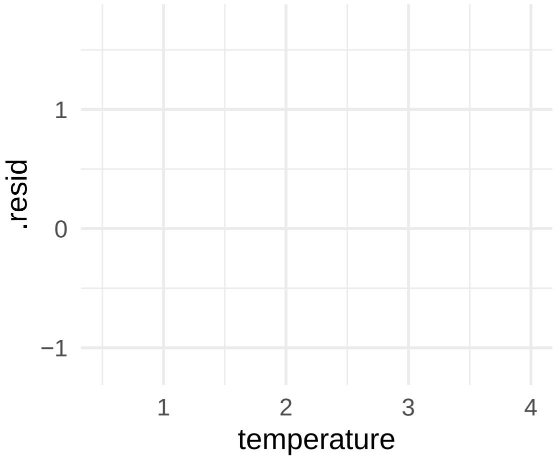
(f) Using your result from the previous part, find an expression for the correlation between $\widehat{\beta}_0$ and $\widehat{\beta}_1$. You should be able to plug in all numbers but you do not need to simplify.

(g) What is the correlation coefficient between `durability` and `temperature`?

(h) Let $\theta = c_1\beta_1 - \beta_0$, where c_1 is a constant. Determine an unbiased estimator $\widehat{\theta}$ for θ.

(i) Show how you would find the MSE $E[(\theta - \widehat{\theta})^2]$

(j) In simple linear regression, it is common to instead make a residual plot of residuals against the x-variable. Fill in the scatterplot below and explain how you were able to do so.



Solution

(a) Read from the table: $\widehat{\beta}_0 = 40.89$, $\widehat{\beta}_1 = .547$ and $\widehat{\sigma^2} = 1.135$.

(b) There may be a curve in the residual scatterplot, suggesting possible nonlinearity.

(c) $.547 \pm 2 \cdot .35$. This would contain zero (either by mental math or by recognizing that the test for $H_0 : \beta_1 = 0$ fails to reject at the $\alpha = .05$ level)

(d) This corresponds to the p-value in the first line: $p = 6.85 \times 10^{-9}$ and so we reject $H_0 : \beta_0 = 0$.

(e) We know that $\Sigma_{\hat{\beta}} = \sigma^2(X^T X)^{-1}$ (formula sheet). So the (estimate) of the covariance would be the estimate of $\sigma^2 \times -.214 = 1.135 \times -.214 = -.243$

(f) $\rho(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{Cov(\widehat{\beta}_0, \widehat{\beta}_1)}{\sigma_{\hat{\beta}_1}\sigma_{\hat{\beta}_0}} = \frac{-.243}{.35 \times .88}$

(g) The multiple $R^2 = .2895$ and so $r = \sqrt{.2895}$

(h) Let $\widehat{\theta} = c_1\widehat{\beta}_1 - \widehat{\beta}_0$. Note that $E(\widehat{\theta}) = c_1 E(\widehat{\beta}_1) - E(\widehat{\beta}_0) = c_1\beta_1 - \beta_0$ so $\widehat{\theta}$ is an unbiased estimator.

(i) Recall that MSE is equal to bias^2 + $V(\widehat{\theta})$. We've shown the bias is 0, so we need to find $V(\widehat{\theta}) = c_1^2 V(\widehat{\beta}_1) + V(\widehat{\beta}_0) - 2Cov(\widehat{\beta}_1, \widehat{\beta}_0) = c_1^2 .35^2 + .88^2 - 2\frac{-.243}{.35 \times .88}$

(j) Scatterplot should look exactly the same as residual vs fitted since there's only 1 x-variable.

2. You are given a dataset with one response variable Y and two predictor variables X_1, X_2, with $n = 5$ observations. You are going to fit the following multiple linear regression model *without an intercept*:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

(a) Write out the matrix form of the multiple linear regression, including the error assumptions. Indicate the dimensions for all matrices.

(b) Express $\frac{1}{n} X^T X$ in terms of sums related to X_{i1}, X_{i2}, Y_i, etc.

(c) To estimate the coefficients, we minimize the sum of squares $\sum (Y_i - \beta_1 X_{i1} - \beta_2 X_{i2})^2$ and derive the following estimating equations:

$$\sum X_{i1} Y_i = \beta_1 \sum X_i^2 + \beta_2 \sum X_{i1} X_{i2}$$

$$\sum X_{i2} Y_i = \beta_1 \sum X_{i1} X_{i2} \beta_2 \sum X_{i2}^2$$

Rewrite the minimization problem using matrices, derive the matrix form of the estimating equations, and show that it has solution $\hat{\beta} = (X^T X)^{-1} X^T Y$

(d) What is the dimension of the hat matrix H for this model?

(e) Express the formula of residual vector $\hat{\epsilon}$ in terms of the hat matrix. Derive the expectation and covariance of $\hat{\epsilon}$ in matrix notation.

Solution

(a) $Y = X\beta + \epsilon$

$$\underset{5\times1}{\begin{bmatrix} y_1 \\ \vdots \\ y_5 \end{bmatrix}} = \underset{5\times2}{\begin{bmatrix} x_{11} & x_{12} \\ & \vdots \\ x_{51} & x_{52} \end{bmatrix}} \underset{2\times1}{\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}} + \underset{5\times1}{\begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_5 \end{bmatrix}}$$

$\epsilon_i \sim N(0, \sigma^2)$ or $\epsilon \sim MVN(\vec{0}, \sigma^2 I)$

(b) $X^T = \begin{bmatrix} x_{11} & \dots & x_{51} \\ x_{12} & \dots & x_{52} \end{bmatrix}$

$$\frac{1}{n} X^T X = \begin{bmatrix} \frac{1}{n} \Sigma x_{i1}^2 & \frac{1}{n} \Sigma x_{i1} x_{i2} \\ \frac{1}{n} \Sigma x_{i1} x_{i2} & \frac{1}{n} \Sigma x_{i2}^2 \end{bmatrix}$$

(c)
$$\begin{aligned} L &= \|Y - X\beta\|^2 \\ &= (Y - X\beta)^T (Y - X\beta) \\ &= (Y - X\beta)^T Y - (Y - X\beta)^T X\beta \\ &= Y^T Y - (X\beta)^T Y - Y^T X\beta + \beta^T X^T X \beta \\ &= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta \end{aligned}$$

To find solutions, take derivative WRT β and set $=0$

$$\frac{\partial L}{\partial \beta} = -2X^T y + 2X^T X \beta = 0$$

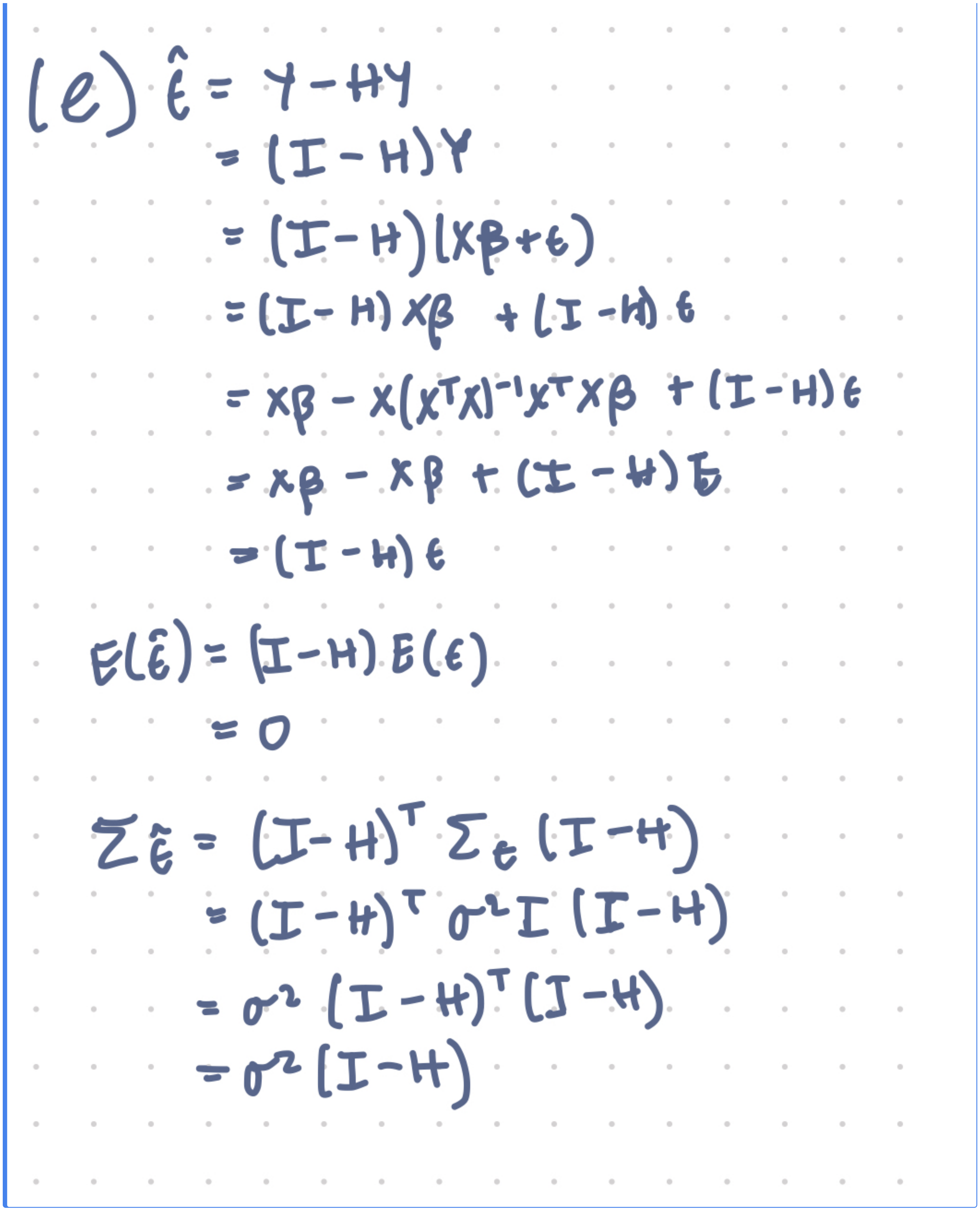
$X^T X \beta = X^T y \leftarrow$ matrix form of estimating equations

To solve for $\hat{\beta}_1$

$$(X^T X)^{-1} (X^T X) \beta = (X^T X)^{-1} X^T y$$
$$\hat{\beta} = (X^T X)^{-1} X^T y$$

(d) $H = \underset{5\times2}{X} \underset{(2\times2)}{(X^T X)^{-1}} \underset{(2\times5)}{X^T}$

H is 5×5



(e) $\hat{\epsilon} = Y - HY$

$$\begin{aligned}
&= (I-H)Y \\
&= (I-H)(X\beta+\epsilon) \\
&= (I-H)X\beta + (I-H)\epsilon \\
&= X\beta - X(X^TX)^{-1}X^TX\beta + (I-H)\epsilon \\
&= X\beta - X\beta + (I-H)\epsilon \\
&= (I-H)\epsilon
\end{aligned}$$

$$\begin{aligned}
E(\hat{\epsilon}) &= (I-H)E(\epsilon) \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\Sigma_{\hat{\epsilon}} &= (I-H)^T \Sigma_\epsilon (I-H) \\
&= (I-H)^T \sigma^2 I (I-H) \\
&= \sigma^2 (I-H)^T (I-H) \\
&= \sigma^2 (I-H)
\end{aligned}$$

3. Recall that in logistic regression, we use the model:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_i$$

That is, 1-unit increase in x_i is associated with a β_1 increase in the logit. It is hard to interpret things on a logit scale. It is often more useful to interpret on the *odds* scale, where the odds = $\frac{p}{1-p}$. Show that an equivalent interpretation is: a 1-unit increase in x_i means that the odds increase by a factor of e^{β_1}.

Solution

Note that $odds_{x_i} = \frac{p_i}{1-p_i} = e^{\beta_0} e^{\beta_1 x_i}$ and $odds_{x_i+1} = \frac{p_i}{1-p_i} = e^{\beta_0} e^{\beta_1 (x_i+1)}$. Therefore when x increases by 1,

$$\frac{odds_{x_i}}{odds_{x_i+1}} = \frac{e^{\beta_0} e^{\beta_1 x_i}}{e^{\beta_0} e^{\beta_1 (x_i+1)}} = e^{\beta_1}$$

So the odds increase by a factor of e^{β_1}.