STAT 504: Applied Regression
Problem Set 2

Winter 2022

Due date: Tuesday, January 25th, 2022.

Instructions: Submit your answers in a *single pdf file*. Your submission should be readable and well formatted. **Handwritten answers will not be accepted**. You can discuss the homework with your peers, but *you should write your own answers and code.* ***No late submissions will be accepted.***

1 CEF and OLS as "best" predictors

1.1 CEF is the best predictor of Y using X

Show that, of all possible functions of X, the CEF $\mathbb{E}[Y \mid X]$ is the best predictor of Y using X, in the sense that it minimizes the mean squared error. That is, show that,

$$\mathbb{E}[Y \mid X] = \arg\min_{f} \mathbb{E}[(Y - f(X))^2]$$

1.2 Linear regression is the best *linear* predictor of Y using X

Show that the traditional linear regression of Y on X, namely, $\text{LR}[Y \mid X] := \alpha + \beta X$, with $\alpha = \mathbb{E}[Y] - \beta\,\mathbb{E}[X]$ and $\beta = \text{Cov}(Y, X)/\,\text{Var}(X)$, is the best linear predictor (BLP) of Y using X, in the sense that it minimizes the mean squared error. In other words, show that,

$$\text{LR}[Y \mid X] = \underset{f \,\in\, \text{Linear}}{\arg\min}\ \mathbb{E}[(Y - f(X))^2]$$

1.3 Linear regression is also the best linear approximation of $E[Y \mid X]$

Further show that linear regression is also the best linear approximation of $\mathbb{E}[Y \mid X]$. An immediate consequence of your proof, is that we can estimate the regression of Y on X using aggregate data instead of individual level data. Explain what this means.

1.4 Linear CEF

Finally, show that, if the CEF is linear, than it equals the linear regression of Y on X.

2 Comparing the CEF and Linear Regression

In this question you are going to contrast the CEF and the linear regression in several examples. All these questions should be solved analytically (ie., using mathematics, not simulation).

2.1 Curved-roof distribution

Let the joint distribution of X and Y be:

$$p(x, y) = 3(x^2 + y)/11 \qquad \text{for} \quad 0 \leq x \leq 2,\ 0 \leq y \leq 1$$

and 0 elsewhere.

(a) Find the CEF of Y given X (for $0 \leq x \leq 2$);

(a) Find the BLP of Y given X (the linear regression);

(a) Draw a plot with both the CEF and the linear regression. Are the CEF and the linear regression the same? Where does the BLP seem to approximate the CEF better?

2.2 Binary random variables

Let both X and Y be binary random variables.

(a) Show that the conditional expectation function $\mathbb{E}[Y \mid X]$ is linear on X.

(b) Show that, in general, the conditional variance function $\text{Var}[Y \mid X]$ will *not* be constant.

(c) Some people say it is not appropriate to fit a linear regression when the outcome Y is binary. Is this true in this case? Why or why not? If you fit a linear regression of Y on X, what is the meaning of the regression coefficients?

2.3 Bivariate normal

Let X and Y be bivariate normal random variables:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \right).$$

(a) What are $\mathbb{E}[Y \mid X]$ and $\mathbb{E}[X \mid Y]$?

(a) What are the linear regressions of Y on X and of X on Y? Are the CEF and the linear regressions the same?

(a) What is $\text{Var}(Y \mid X)$? Does it vary as a function of X?

(a) What is the meaning of the regression coefficients?

2.4 Quadratic CEF

Let $X \sim N(0,1)$, $\epsilon \sim N(0,1)$, and $Y = X^2 + \epsilon$.

- Compute $\mathbb{E}[Y \mid X]$.
- Compute $\text{BLP}[Y \mid X]$ (the linear regression of Y on X).
- Plot the CEF and BLP. Are they the same? In general, does the BLP provide a good approximation to the CEF?
- Suppose the researcher is interested in estimating $Q := \mathbb{E}[Y|X = a] - \mathbb{E}[Y \mid X = -a]$. Instead of using the CEF, the researcher approximates Q using $Q' := \text{BLP}[Y \mid X = a] - \text{BLP}[Y \mid X = -a]$. Is Q' a good approximation of Q?
- In this example, if we change the marginal distribution of X, what happens to the CEF—does it also change? And what happens to the BLP?
- Compute the linear regression of Y on X^2, and compare it to the CEF.

3 Properties of error terms

There is much confusion about the meaning of error terms in regression analysis; for instance, many textbooks often conflate properties of error terms with assumptions regarding error terms. Here you will reproduce and interpret the proofs we have seen in class.

3.1 CEF decomposition

For random variables X and Y, show that we can always decompose Y in the following form:

$$Y = f(X) + \epsilon$$

Where:

(a) $\mathbb{E}[\epsilon \mid X] = 0$ and $\mathbb{E}[\epsilon] = 0$;

(b) $\text{Var}[\epsilon \mid X] = \text{Var}(Y \mid X)$ and $\text{Var}[\epsilon] = \mathbb{E}[\text{Var}(Y \mid X)]$;

(c) $\mathbb{E}[h(X)\epsilon] = 0$, $\forall\ h(x)$.

What is the meaning of $f(X)$ and ϵ in this decomposition? Are these *assumptions* about the error term ϵ?

3.2 OLS (linear) decomposition

For random variables X and Y, show that we can always decompose Y in the following form:

$$Y = \alpha + \beta X + e$$

Where:

(a) $\mathbb{E}[e] = 0$;

(b) $\mathbb{E}[Xe] = 0$;

(c) $\text{Var}(e) = \text{Var}(Y) - \beta^2 \text{Var}(X)$.

What is the meaning of $\alpha + \beta X$ and e in this decomposition? Are these *assumptions* about the error term e? What is the interpretation of β?

3.3 Comparing the two decompositions

We just showed above that we decompose Y both as:

$$Y = f(X) + \epsilon$$

and

$$Y = \alpha + \beta X + e$$

If $f(X) \neq \alpha + \beta X$, is this a contradiction? Why or why not?

3.4 Further linear decomposition

Continuing our study of best linear approximations, show we can further decompose Y as:

$$Y = \alpha + \beta X + u + \epsilon$$

Where ϵ is the CEF disturbance as in Question 3.1, and $u := \mathbb{E}[Y \mid X] - (\alpha + \beta X)$. Explain the meaning of each component of this decomposition, namely:

(a) $\alpha + \beta X$;

(b) u;

(c) ϵ.

3.5 Be careful with textbooks

Read the passage of Figure 1, extracted from Weisberg [2005, p.21]. Considering the results you just proved above, regarding properties of the regression error term, what is problematic (or, at least, confusing) with this passage?

Because the variance $\sigma^2 > 0$, the observed value of the ith response y_i will typically not equal its expected value $\mathrm{E}(Y|X = x_i)$. To account for this difference between the observed data and the expected value, statisticians have invented a quantity called a statistical error, or e_i, for case i defined implicitly by the equation $y_i = \mathrm{E}(Y|X = x_i) + e_i$ or explicitly by $e_i = y_i - \mathrm{E}(Y|X = x_i)$. The errors e_i depend on unknown parameters in the mean function and so are not observable quantities. They are random variables and correspond to the *vertical distance between the point* y_i *and the mean function* $\mathrm{E}(Y|X = x_i)$. In the heights data, Section 1.1, the errors are the differences between the heights of particular daughters and the average height of all daughters with mothers of a given fixed height.

We make two important assumptions concerning the errors. First, we assume that $\mathrm{E}(e_i|x_i) = 0$, so if we could draw a scatterplot of the e_i versus the x_i, we would have a null scatterplot, with no patterns. The second assumption is that

Figure 1: Passage of Weisberg [2005].

Extra-credit: Can you find similar problematic passages in other textbooks?

4 ANOVA (Analysis of Variance)

4.1 Nonparametric ANOVA (Law of Total Variance)

Show that we can decompose the variance of Y as:

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}[Y \mid X]) + \mathbb{E}[\mathrm{Var}(Y \mid X)]$$

4.2 Nonparametric R^2 (Pearson's correlation ratio)

This question is for extra-credit. Define the *nonparametric* R^2 (also known as the *Pearson's correlation ratio*) as:

$$\eta^2_{Y \sim X} := \frac{\mathrm{Var}(\mathbb{E}[Y \mid X])}{\mathrm{Var}(Y)}$$

The *nonparametric* R^2 measures the fraction of the variation of Y that is explained by X.

(a) Show that:

$$\eta^2_{Y \sim X} = 1 - \frac{\mathbb{E}(\mathrm{Var}[Y \mid X])}{\mathrm{Var}(Y)} = \mathrm{Cor}^2(Y, \mathbb{E}[Y \mid X])$$

(b) Is the *nonparametric* R^2 symmetric? That is, in general, is $\eta^2_{Y \sim X} = \eta^2_{X \sim Y}$? If not, provide a counter-example.

4.3 Linear ANOVA

Let $Y = \alpha + \beta X + e$ be the linear regression of Y on X. Show that we can decompose the variance of Y as the variance of our linear predictions and the residuals:

$$\mathrm{Var}(Y) = \mathrm{Var}(\alpha + \beta X) + \mathrm{Var}(e)$$

4.4 Linear R^2 (coefficient of determination)

Define the *linear* R^2 (also known as the *coefficient of determination*) as:

$$R^2_{Y\sim X} := \frac{\mathrm{Var}(\alpha + \beta X)}{\mathrm{Var}(Y)}$$

The *linear* R^2 measures the fraction of the variation of Y that is linearly explained by X.

- Show that:

$$R^2_{Y\sim X} = 1 - \frac{\mathrm{Var}(e)}{\mathrm{Var}(Y)} = \mathrm{Cor}^2(Y, \alpha + \beta X) = \mathrm{Cor}^2(Y, X)$$

- Show that the linear R^2 is symmetric, namely:

$$R^2_{X\sim Y} = R^2_{Y\sim X}$$

4.5 Comparing η^2 and R^2

This question is for extra-credit.

(a) Argue that $R^2_{Y\sim X} \leq \eta^2_{Y\sim X} \leq 1$;

(b) Show that:

$$\eta^2_{Y\sim X} = R^2_{Y\sim X} + (1 - R^2_{Y\sim X})\,\mathrm{Cor}^2(u, e)$$

Where u and e are defined as in Questions 3.4 and 3.2.

5 Transformation of variables

5.1 Linear transformation of X

Let $Y = \alpha + \beta X + e$ be the linear regression of Y on X. Now define the new random variable $X' := a + bX$ and let $Y = \alpha' + \beta' X' + e'$ denote the linear regression of Y on X'. Express the regression coefficients α' and β' in terms of the original regression coefficients α and β. Does this transformation change the R^2 of the regression?

5.2 Linear transformation of Y

Let $Y = \alpha + \beta X + e$ be the linear regression of Y on X. Now define the new random variable $Y' = a + bY$ and let $Y' = \alpha' + \beta' X + e'$ denote the linear regression of Y' on X. Express the

regression coefficients α' and β' in terms of the original regression coefficients α and β. Does this transformation change the R^2 of the regression?

5.3 Standardization

Let $Y = \alpha + \beta X + e$ be the linear regression of Y on X. Now define the new random variables $X' := (X - \mathbb{E}[X])/\operatorname{SD}(X)$ and $Y' := (Y - \mathbb{E}[Y])/\operatorname{SD}(Y)$.

(a) What are $\mathbb{E}[X']$ and $\mathbb{E}[Y']$?

(b) What are $\operatorname{Var}(X')$ and $\operatorname{Var}(Y')$?

(c) Let $Y' = \alpha' + \beta' X' + e'$ denote the linear regression of Y' on X'. Express the regression coefficients α' and β' in terms of the original regression coefficients α and β. Does this transformation change the R^2 of the regression?

6 The effect of the regressor distribution on linear regression

As we have discussed in class, if the CEF is not linear, the linear regression of Y on X depends on the distribution of X. Here you are going to verify this both analytically and numerically using simulations. For each item below, let $Y = X^2 + \epsilon$, with $\epsilon \sim N(0, 1)$, but consider different marginal distributions for X.

(a) Let the marginal distribution of X be $X \sim N(1, .2)$. What is the $\operatorname{BLP}(Y \mid X)$ in this case? Plot both the CEF and BLP for $0.1 \leq x \leq 1.9$. Where does the BLP provide a good approximation to the CEF, and where does it fail?

(b) Now let the marginal distribution of X be $X \sim N(-1, .2)$ What is the $\operatorname{BLP}(Y \mid X)$ in this case? Is it similar to the BLP of (a)? Plot both the CEF and BLP for $-0.1 \leq x \leq -1.9$. Where does the BLP provide a good approximation to the CEF, and where does it fail?

(c) Suppose a data scientist fits a linear regression with the data distributed as in (a). Now suppose this model is going to be deployed in a new population, where the data is distributed as in (b). How good are these predictions going to be? Explain what is happening.

(d) Repeat questions (a), (b) and (c) now using the linear regression of Y on X^2. Which conclusions change, and why?

(e) Verify numerically what we showed above in (a), (b), (c) and (d) using simulated data with a sample size of $n = 10,000$ (either in `R` or in `Python`).

References

Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.