# STAT331: Practice Assignment 1

Ashley Chui (21003047)

January 26, 2026

**1.** In some situations it is appropriate to fit a simple linear regression through the origin. This often arises when theory suggests that one variable is proportional to another, or when including an intercept would imply an unrealistic outcome at zero.

For example, Hooke's law states that when a spring is stretched, the force applied to the spring (X) is proportional to the increase in length (Y). The constant of proportionality can be estimated by fitting the linear model

$$Y_i = \beta_1 x_i + \epsilon_i \text{ for } x = 1, 2, ..., n \tag{1}$$

where the intercept is fixed to be 0 (i.e no applied force and hence no extension) and the random error $\epsilon_i$ are assumed to be independent with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.

(a) Derive the least squares estimate $\hat{\beta}_1$ of $\beta_1$. Be sure to prove that this estimate achieves a minimum.
The LSE is the value of $\beta_1$ that minimizes the sum of squared residuals.

$$S(\beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2.$$

To obtain $\hat{\beta}_1$, we differentiate $S(\beta_1)$ and set it equal to 0:

$$\frac{d}{d\beta_1} S(\beta_1) = 2 \sum_{i=1}^{n} (y_i - \beta_1 x_i)(-x_i).$$

Setting the derivative to zero gives

$$0 = \sum_{i=1}^{n} y_i x_i - \beta_1 \sum_{i=1}^{n} x_i^2,$$

so

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}.$$

To prove this critical point is a minimum, compute the second derivative:

$$\frac{d^2}{d\beta_1^2} S(\beta_1) = 2 \sum_{i=1}^{n} x_i^2.$$

If $\sum_{i=1}^{n} x_i^2 > 0$, then $\frac{d^2}{d\beta_1^2} S(\beta_1) > 0$, so $S(\beta_1)$ is strictly convex and the critical point is the minimizer. Hence,

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}}$$

is the LSE of $\beta_1$ and it achieves a minimum.

(b) Under the no-intercept model, the fitted value is $\hat{y}_i = \hat{\beta}_1 x_i$ and the residual is

$$r_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 x_i.$$

Determine whether the following statements are true:

(i) $\sum_{i=1}^n r_i x_i = 0$
(ii) $\sum_{i=1}^n r_i = 0$

**(i)** From part (a), it was shown that $\hat{\beta}_1$ is found by differentiating $S(\beta_1)$ and setting it to 0 from which we obtained the following:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

Then

$$\sum_{i=1}^n r_i x_i = \sum_{i=1}^n x_i(y_i - \hat{\beta}_1 x_i) = \sum_{i=1}^n x_i y_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0.$$

Therefore, the statement$\sum_{i=1}^n r_i x_i = 0$ is true.

**(ii)** We have

$$\sum_{i=1}^n r_i = \sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i,$$

which is not necessarily equal to zero since it has no relation to the LSE equation. Therefore,

$$\sum_{i=1}^n r_i \neq 0$$

(c) Can you conclude that the sample correlation between the residuals and the explanatory variable is zero? That is, is

$$\frac{\sum_{i=1}^n (r_i - \bar{r})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (r_i - \bar{r})^2}\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = 0?$$

Expanding the numerator, we get

$$\begin{aligned}
\sum_{i=1}^n (r_i - \bar{r})(x_i - \bar{x}) &= \sum_{i=1}^n r_i x_i - \bar{x}\sum_{i=1}^n r_i - \bar{r}\sum_{i=1}^n x_i + n\bar{r}\bar{x} \\
&= \sum_{i=1}^n r_i x_i - \bar{x} n \bar{r} - \bar{r} n \bar{x} + n\bar{r}\bar{x} \\
&= \sum_{i=1}^n r_i x_i - n\bar{r}\bar{x}.
\end{aligned}$$

This only equals 0 if $\bar{r} = 0$ or $\bar{x} = 0$ so we cannot conclude the above statement as true.

(d) Now denote the least squares estimator of $\beta_1$ in the no-intercept model by $\tilde{\beta}_1$. Show that $\tilde{\beta}_1$ is an unbiased estimator and determine its variance.

To show that $\tilde{\beta}_1$ is an unbiased estimator, we can show that $E(\tilde{\beta}_1) = \beta_1$.

From part (a)

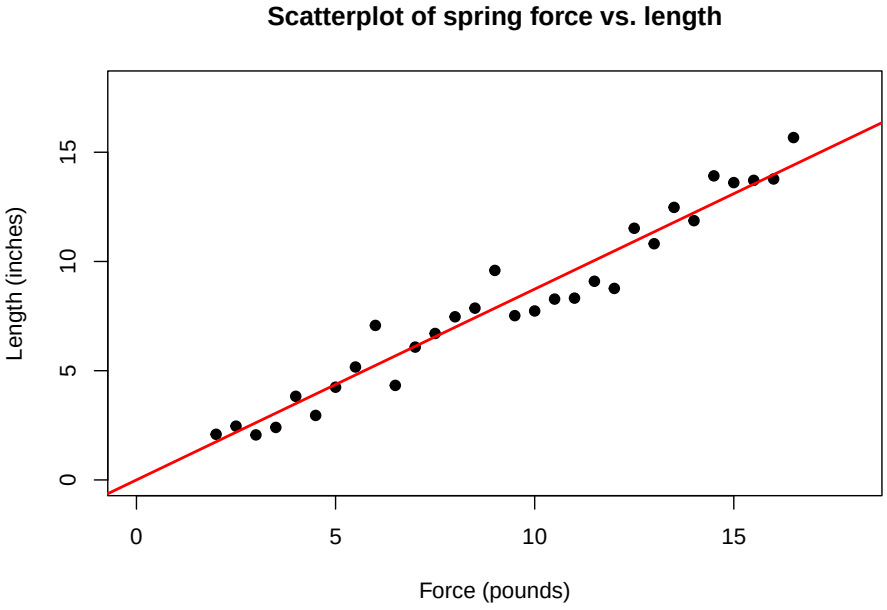
$$\begin{aligned}
\tilde{\beta}_1 &= \frac{\sum_{i=1}^n x_i(\beta_1 x_i + \epsilon_i)}{\sum_{i=1}^n x_i^2} \\
&= \frac{\beta_1 \sum_{i=1}^n x_i^2 + \sum_{i=1}^n x_i \epsilon_i}{\sum_{i=1}^n x_i^2} \\
&= \beta_1 + \frac{\sum_{i=1}^n x_i \epsilon_i}{\sum_{i=1}^n x_i^2}
\end{aligned}$$

Then

$$\begin{aligned} E(\tilde{\beta}_1) &= \beta_1 + \frac{1}{\sum_{i=1}^n x_i} E(\sum_{i=1}^n x_i \epsilon_i) \\ &= \beta_1 + \frac{1}{\sum_{i=1}^n x_i} \sum_{i=1}^n x_i E(\epsilon_i) \\ &= \beta_1 \quad \text{since } Var(\epsilon_i) = \sigma^2 \text{ and by the independence of } \epsilon_i \end{aligned}$$

(e) Plot the data with appropriate labels for x- and y-axis and a title. Determine if the relationship looks approximately linear.

Data: Collected from 30 springs that has been tested, the variables including the force ($x_i$) in pounds and the increase in length ($y_i$) in inches. Available in "spring.txt".




From the scatterplot, the relationship between force and extension appears approximately linear. The data points cluster closely around a straight line with no evident curvature, supporting the use of a linear model through the origin as suggested by Hooke's law.

(f) Fit two simple linear models: one with intercept and one without intercept, using the **lm()** function in R. For example, **lm(y ~ x, data = ...)** for the model with an intercept and **lm(y ~ -1 + x, data = ...)** for the model without. Compare the least squares estimates of $\beta_1$. Does removing the intercept make much difference to the estimate of $\beta_1$?

Fitting models with and without an intercept yields slope estimates of 0.8902 and 0.8735, respectively. Since these values are very close, removing the intercept has little effect on the estimate of $\beta_1$. This supports the use of a no-intercept model for these data.

**2.** Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $E(\epsilon_i) = 0, Var(\epsilon) = \sigma^2$ and $\epsilon_i$'s are independent, $i = 1, 2, ..., n$. Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ denote the least squares estimators for $\beta_0$ and $\beta_1$.

(a) Show that $Cov(\tilde{\beta}_0, \tilde{\beta}_1) = -\bar{x}\sigma^2/S_{xx}$, where $\bar{x} = \sum_{i=1}^n x_i/n$ and $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$.
Using $\hat{\beta}_0 = \bar{Y} - \bar{x}\,\hat{\beta}_1$,

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{Cov}(\bar{Y}, \hat{\beta}_1) - \bar{x}\ \text{Var}(\hat{\beta}_1).$$

First, since

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} Y_i,$$

we have

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{S_{xx}} \right)^2 = \frac{\sigma^2}{S_{xx}}.$$

Next,

$$\mathrm{Cov}(\bar{Y}, \hat{\beta}_1) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{n} \frac{x_j - \bar{x}}{S_{xx}} \mathrm{Cov}(Y_i, Y_j).$$

Since $\mathrm{Cov}(Y_i, Y_j) = \sigma^2$ for $i = j$ and 0 otherwise,

$$\mathrm{Cov}(\bar{Y}, \hat{\beta}_1) = \frac{\sigma^2}{nS_{xx}} \sum_{j=1}^{n} (x_j - \bar{x}) = 0,$$

because $\sum_{j=1}^{n} (x_j - \bar{x}) = 0$.
Therefore,

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x} \frac{\sigma^2}{S_{xx}} = -\frac{\sigma^2 \bar{x}}{S_{xx}}.$$

(b) Show that $Cov(\bar{Y}, \tilde{\beta}_1) = 0$.

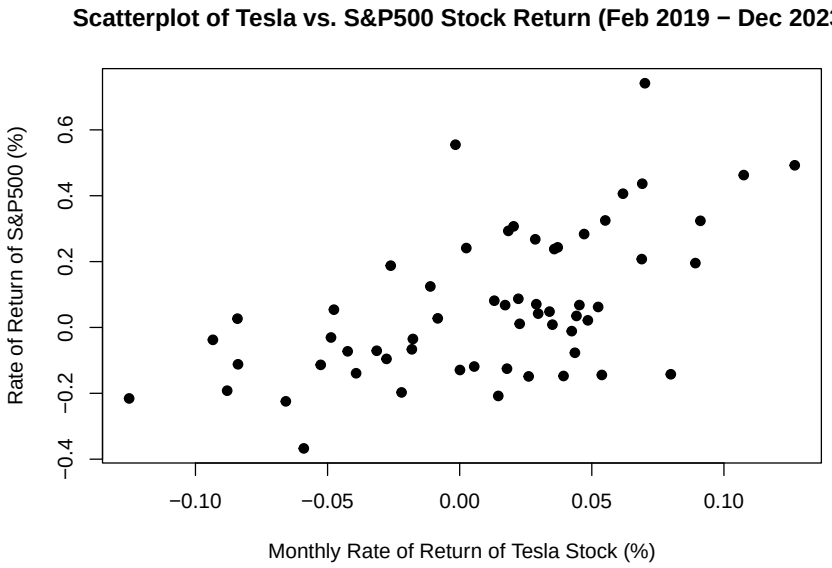
From the above calculation,

$$\mathrm{Cov}(\bar{Y}, \hat{\beta}_1) = \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

**3.** The Capital Asset Pricing Model (CAPM) is an important financial modeling tool where simple linear regression can be applied. This model is used to assess the risk of a security (e.g. stock) in terms of its contribution to the risk of a portfolio of stocks in the market. Here, Y is the monthly rate of return of Tesla stock, and X is the rate of return of S&P 500 (the market portfolio proxy) during the same period. We would like to investigate how Tesla stock returns move relative to overall movements of stock market returns, which reflects the stock's volatility relative to that of an average stock in the market.

The file "tesla.xlsx" (available on LEARN) contains the monthly rate of return of Tesla stock ($y_i$) and the corresponding rate of return of S&P500 ($x_i$) from February 2019 to December 2023.

(a) Before we fit a simple linear model, create a scatterplot of the data with appropriate axis labels and title. Does a linear model seem appropriate?



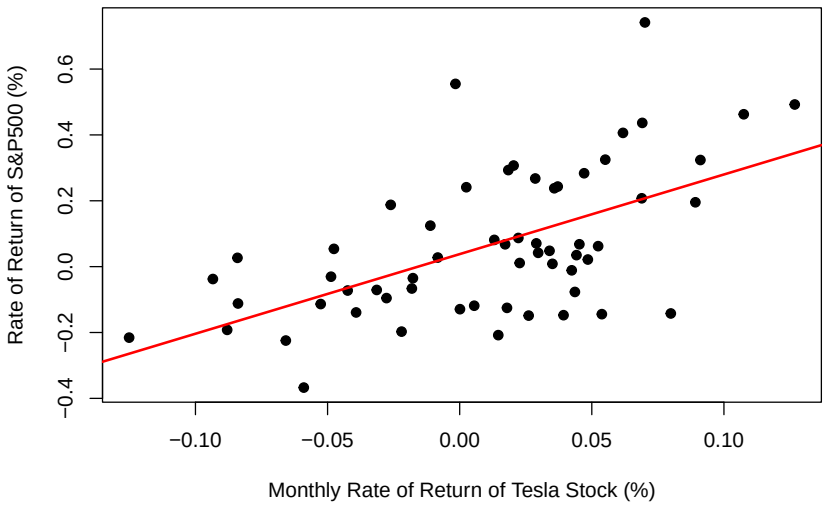

This scatterplot indicates a clear positive association between Tesla's monthly returns vs. S&P500's monthly return. As Tesla's returns increase, S&P 500 returns also tend to increase, therefore a linear model seems appropriate for this data.

(b) Fit the SLR to the data using R. Write down the equation of the fitted line, and add this line to the scatterplot produced in a). Since we assume a simple linear regression model for this data, we must calculate the LSE of $\beta_0$ and $\beta_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

This calculation was done in R and we thus obtained the following equation of the fitted line:

$$Y_i = 0.038 + 2.42 x_i$$



(c) Based on your estimate of the slope $\beta_1$, what does it imply about the volatility of Tesla stock relative ot the market?

The estimated slope coefficient suggests that Tesla stock is substantially more volatile than the overall market. Its returns tend to amplify market movements, indicating high risk relative to an average stock in the S&P 500.

(d) Produce a 95% confidence interval for $\beta_1$, and use it to assess the hypothesis $H_0 : \beta_1 = 0$. What is your conclusion?

A 95% confidence interval for $\beta_1$ was obtained using the fitted simple linear regression model. The resulting interval is

$$(1.509,\ 3.327).$$

Since this interval does not contain 0, we reject the null hypothesis

$$H_0 :\ \beta_1 = 0$$

at the 5% significance level. This conclusion is consistent with the hypothesis test based on the $t$-statistic, which yields a p-value of $1.76 \times 10^{-6}$. Therefore, there is strong statistical evidence of a positive linear relationship between Tesla's monthly returns and the S&P 500's monthly returns, indicating that Tesla stock exhibits significantly higher risk than the overall market.

(e) Suppose an average-risk stock is defined as one whose returns tend to move up or down at the same rate as the overall market. Based on this definition, do you think Tesla stock qualifies as an average-risk stock? Use a hypothesis test to answer this question. Be sure to clearly state

- the null and alternative hypotheses,
- the test statistic
- your conclusion based on this test

An average-risk stock corresponds to the slope coefficient $\beta_1 = 1$. Hence, to test if Tesla is an average-risk stock we choose the following null hypothesis:

$$H_0 :\ \beta_1 = 1 \quad \text{vs.} \quad H_A :\ \beta_1 \neq 1.$$

The test statistic is given by

$$t = \frac{\hat{\beta}_1 - 1}{\text{SE}(\hat{\beta}_1)}.$$

Using the fitted model, $\hat{\beta}_1 = 2.418$ and $\text{SE}(\hat{\beta}_1) = 0.454$, which yields

$$t = \frac{2.418 - 1}{0.454} \approx 3.12.$$

Under the null hypothesis, this statistic follows a $t$-distribution with $n - 2 = 57$ degrees of freedom. The resulting p-value is less than 0.01, leading us to reject $H_0$ at the 5% significance level.

Therefore, there is strong statistical evidence that Tesla's beta is significantly different from 1. We conclude that Tesla stock does not qualify as an average-risk stock.

(f) Estimate the expected (mean) return of Tesla stock when the return rate of the market portfolio is 5%. Construct a 95% confidence interval for this estimate.

Using our fitted regression line, if the market return is $x = 0.05$, the estimated expected return of Tesla is

$$\hat{Y} = 0.038 + 2.42(0.05) = 0.159$$

So when the market return is 5%, the expected monthly return of Tesla stock is 15.9%. The 95% CI for this estimate is $[0.0994, 0.2184]$.

(g) Now, predict the return rate of Tesla stock for the next quarter if the market portfolio is expected to return 5%, as in the previous question. Construct a 95% prediction interval for this prediction.

The 95% prediction interval for the return rate of Tesla stock for the next quarter is $[-0.2132, 0.5311]$.

(h) Compare the 95% prediction interval you calculated in (g) with the 95% confidence interval for the mean return in (f). Explain why the two intervals are different.

The prediction interval is substantially wider than the confidence interval. This is because the confidence interval in part (f) reflects uncertainty only in estimating the average (mean) return of Tesla when the market return is 5%. In contrast, the prediction interval in part (g) accounts for both the uncertainty in estimating the regression line and the additional variability associated with predicting a single future observation.

**4.** An investigative study collected 40 observations from the Wabash river at random locations near Lafayette. Each observation consisted of a measure of water pH (x) and fish count (y). The researchers are interested in how the acidity of the water affects the number of fish.

(a) Complete the following ANOVA table for the regression analysis

| Source | Sum of Squares | Degrees of Freedom | Mean Square |
|---|---|---|---|
| Regression | 50.30 | 1 | 50.30 |
| Error | 9.70 | 38 | 0.255 |
| Total | 60.00 | 39 | |

(b) What is the estimate of the variance of random error?

The estimate of the variance of the random error is the Mean Square Error (MSE). From the table above,

$$\hat{\sigma}^2 = MSE = \frac{SSE}{df_{Error}} = \frac{9.70}{38} = 0.255$$

(c) Use R-square to explain what percentage of variation in fish count can be explained by water pH.

$$R^2 = \frac{SSR}{SST} = \frac{50.30}{60.00} = 0.838$$

About 83.8% of the variation in fish count can be explained by water pH usin this linear regression model.