36-401, Chapter 4: Prediction in Simple Linear Regression

Zach Branson, Fall 2025

Example 0.1. Data were gathered from 97 men with advanced prostate cancer, including two variables of interest:

- **PSA Level (`psa`)**: Serum prostate-specific antigen level (ng/mL)
- **Cancer volume (`cavol`)**: Estimate of prostate cancer volume (cc)

It would be useful to predict cancer volume (Y) from PSA levels (X), because PSA level can be measured with a simple blood test, but cancer volume is harder to measure. Knowing (or estimating) cancer volume can help determine the appropriate treatment.

Let's consider some initial EDA. The left side of Figure 0.1 shows histograms of X and Y, and a scatterplot of X and Y. The right side does the same for $\log(X)$ and $\log(Y)$.

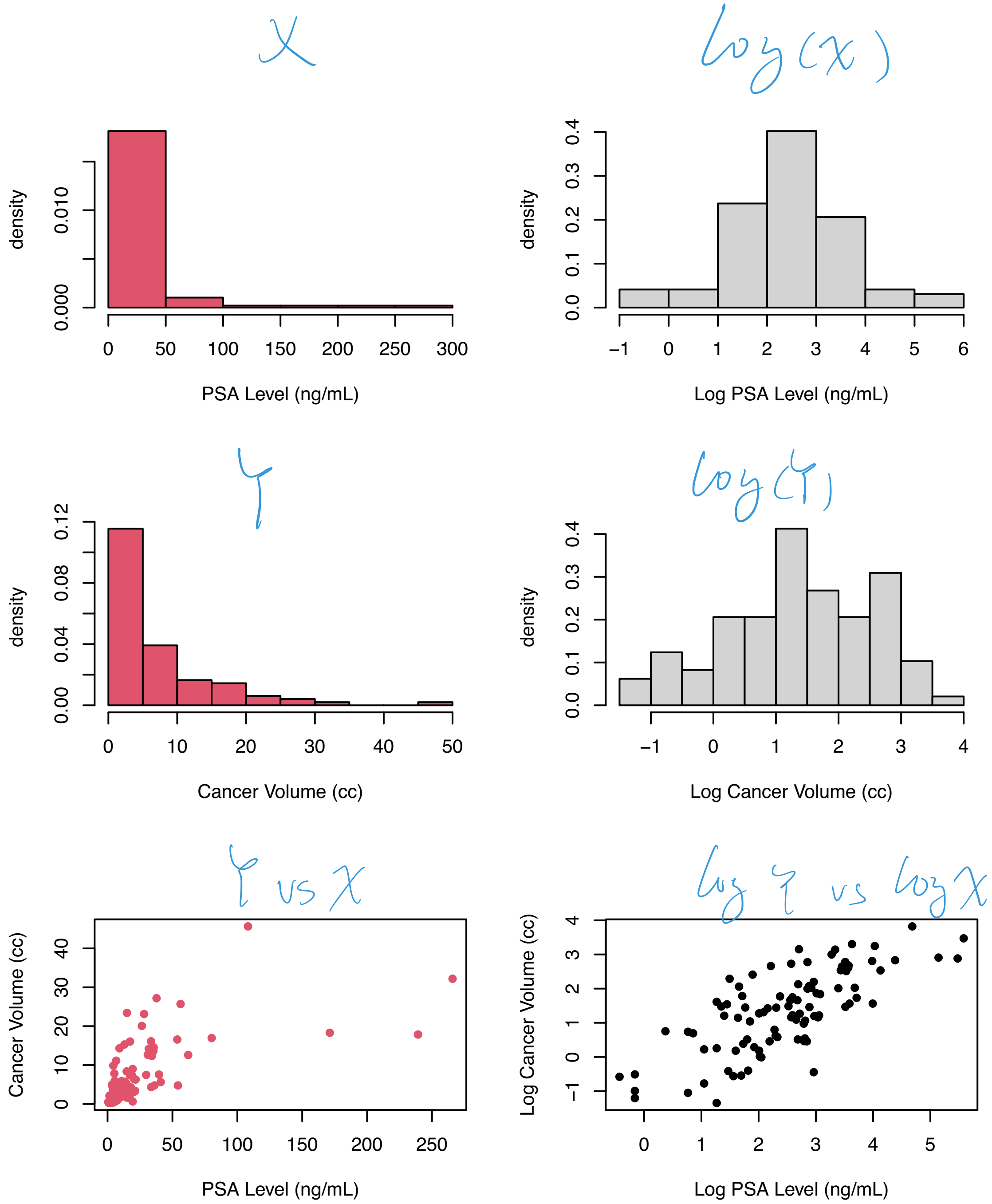


Figure 0.1: Descriptive plots for the prostate bivariate data set.

Exercise 0.1. In this case, do log-transformations help for the purposes of linear regression? If so, why? Furthermore, how does a linear regression model for $\log(Y) \sim \log(X)$ relate to a model for $Y \sim X$?

Note: linear regression assumes $Y|X \sim N(\beta_0 + \beta_1 x, \sigma^2)$

But, we aren't assuming anything about the marginal distribution X and Y.

However, log-transformation makes equal variance assumption more realistic, as well as linearity assump.

The model for $\log(Y) \sim \log(X)$: $\log(Y_i) = \beta_0 + \beta_1 \log(X) + \varepsilon_i$

$$= \beta_0 + \log(X_i^{\beta_1}) + \varepsilon_i$$

$$\Rightarrow Y_i = \exp(\beta_0 + \log(X_i^{\beta_1}) + \varepsilon_i)$$

$$= \exp(\beta_0) \cdot \exp(\log(X_i^{\beta_1})) \cdot \exp(\varepsilon_i)$$

$$= X_i^{\beta_1} \cdot \underbrace{\exp(\beta_0) \cdot \exp(\varepsilon_i)}_{\text{multiplicative error}}$$

Prediction

Under the **normal simple linear regression model**, we assume $Y = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon | X \sim N(0, \sigma^2)$. Thus:

$$Y|X \sim N(\mu(X), \sigma^2), \text{ where } \mu(X) = \mathbb{E}[Y|X] = \beta_0 + \beta_1 X$$

Now we want to use this model to predict Y based on X. How do we make those predictions, how good are they, and how certain are we about them?

Say $X = x$ and we want to predict Y. In Chapter 2, we showed that the MSE-optimal prediction is

$$\mu(x) = \mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x$$

However, we do not know β_0 and β_1. Thus, in practice, we can consider the following prediction at $X = x$:

$$\widehat{\mu}(x) = \widehat{\mathbb{E}}[Y|X = x] = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Here, $\widehat{\mu}(x)$ is called the **fitted value** at x.

Notice that $\widehat{\mu}(x)$ is a random variable, because $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are random variables. We've considered the bias and variance of $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Thus, we can also consider the bias and variance of $\widehat{\mu}(x)$.

We'll again focus on **least-squares estimators**, which are:

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X} \\
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n \left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}
\end{aligned}$$

They are **unbiased**. Their **conditional variances** are:

$$\begin{aligned}
\text{Var}(\widehat{\beta}_0|X) &= \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{j=1}^n (X_j - \overline{X})^2}\right] \\
\text{Var}(\widehat{\beta}_1|X) &= \frac{\sigma^2}{\sum_{j=1}^n (X_j - \overline{X})^2}
\end{aligned}$$

Bias and Variance

Useful Fact: $\widehat{\beta}_0$ and $\widehat{\beta}_1$ can be written in a "signal + noise" form:

$$\widehat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n} \epsilon_i (X_i - \overline{X})}{\widehat{\mathrm{Var}}_n(X)}$$

$$\widehat{\beta}_0 = \beta_0 + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \left\{ 1 - \frac{\overline{X}(X_i - \overline{X})}{\widehat{\mathrm{Var}}_n(X)} \right\}$$

where $\widehat{\mathrm{Var}}_n(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Exercise 0.2. Show $\mathbb{E}(\widehat{\beta}_0 \mid X) = \beta_0$ and $\mathbb{E}(\widehat{\beta}_1 \mid X) = \beta_1$ i.e., the estimators are conditionally unbiased.

Similarly, $\widehat{\mu}(x)$ can be written as “signal + noise”:

$$\begin{aligned}\widehat{\mu}(x) &= \widehat{\beta}_0 + \widehat{\beta}_1 x \\ &= (\beta_0 + \beta_1 x) + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \left\{ 1 - \frac{\overline{X}(X_i - \overline{X})}{\widehat{\text{Var}}(X)} + x\frac{(X_i - \overline{X})}{\widehat{\text{Var}}(X)} \right\} \\ &= (\beta_0 + \beta_1 x) + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \left\{ 1 + \frac{(x - \overline{X})(X_i - \overline{X})}{\widehat{\text{Var}}(X)} \right\}\end{aligned}$$

Following the same process as the previous exercise, we can show $\mathbb{E}[\widehat{\mu}(x)|X] = \mu(x)$, i.e., $\widehat{\mu}(x)$ is also unbiased.

This decomposition also lets us derive the conditional variance of $\widehat{\mu}(x)$ (although it’s algebraically more complicated).

$$\begin{aligned}\text{Var}\{\widehat{\mu}(x) \mid X\} &= \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} \epsilon_i \left\{ 1 + \frac{(x - \overline{X})(X_i - \overline{X})}{\widehat{\text{Var}}_n(X)} \right\} \;\middle|\; X \right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \left\{ 1 + \frac{(x - \overline{X})(X_i - \overline{X})}{\widehat{\text{Var}}_n(X)} \right\}^2 \text{Var}(\epsilon_i \mid X) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \left\{ 1 + \frac{2(x - \overline{X})(X_i - \overline{X})}{\widehat{\text{Var}}_n(X)} + \frac{(x - \overline{X})^2(X_i - \overline{X})^2}{\widehat{\text{Var}}_n(X)^2} \right\} \text{Var}(\epsilon_i \mid X) \\ &= \left(\frac{\sigma^2}{n}\right) \frac{1}{n}\sum_{i=1}^{n} \left\{ 1 + \frac{2(x - \overline{X})(X_i - \overline{X})}{\widehat{\text{Var}}_n(X)} + \frac{(x - \overline{X})^2(X_i - \overline{X})^2}{\widehat{\text{Var}}_n(X)^2} \right\} \\ &= \left(\frac{\sigma^2}{n}\right) \left\{ 1 + \frac{(x - \overline{X})^2}{\widehat{\text{Var}}_n(X)} \right\}\end{aligned}$$

where the second equality follows by the iid assumption, the third by expanding the square, the fourth by constant variance, and the fifth by some simplification. (It is a useful exercise to understand why the fourth line is equal to the fifth line.)

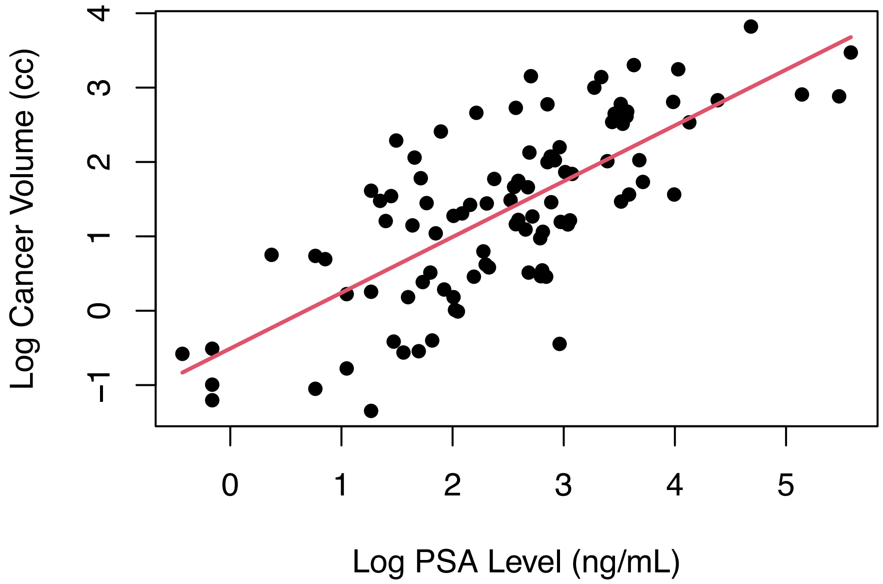
Exercise 0.3. What can we conclude from the form of the conditional variance, $\mathrm{Var}(\widehat{\mu}(x)) = \left(\frac{\sigma^2}{n}\right)\left\{1 + \frac{(x-\overline{X})^2}{\widehat{\mathrm{Var}}_n(X)}\right\}$?

Exercise 0.4. Sketch out how we can derive the variance of $\widehat{\mu}(x)$ using the variances and covariance of $\hat{\beta}$s.

Now we'll consider a linear regression model using the PSA data from Example 0.1. The code below fits the $\log(Y) \sim \log(X)$ linear regression model, and plots predictions along the regression line.

```
prosdat <- read.csv("data/prostate.csv")
proslm <- lm(log(cavol) ~ log(psa), data = prosdat)

plot(log(cavol) ~ log(psa), data = prosdat,
     xlab = "Log PSA Level (ng/mL)",
     ylab = "Log Cancer Volume (cc)", pch = 16)
lines(log(prosdat$psa), predict(proslm), lwd = 2, col = 2)
```



Consider two questions we could answer with this model:

1. **Question about Population**: What is the estimated average cancer volume for the population of men with a PSA level of 10 ng/mL?
2. **Question about Individual**: If an individual has a PSA level of 10 ng/mL, what is our prediction of this man's cancer volume?

Exercise 0.5. Assume we use the simple linear regression model to answer the two questions. Would the answers to these two questions be different?

Nonetheless, the two questions are subtly different. Typically, we're not just interested in point estimation (for a population or individual), but also we want to **quantify our uncertainty** about our estimates. Intuitively, we'll have more certainty about predictions for a population than an individual.

We'll first consider uncertainty quantification when we make predictions about a **population**. This will involve uncertainty for the average prediction, $\widehat{\mu}(x) = \widehat{\mathbb{E}}[Y|X = x]$.

Then, we'll consider uncertainty quantification when we make predictions about an **individual**. Even though this will involve the same prediction $\widehat{\mu}(x)$, we'll have to consider uncertainty about a *specific prediction*, rather than predictions on average.

Interval Estimation for the Mean Response

First we'll consider estimating the mean outcome for a particular covariate value (which we'll call x^*). Our best estimate is:

$$\widehat{\mu}(x^*) = \widehat{\beta}_0 + \widehat{\beta}_1 x^*.$$

We've already seen $\widehat{\mu}(x^*)$ is unbiased. Meanwhile, we showed that

$$\mathrm{Var}(\hat{\mu}(x^*)|X) = \frac{\sigma^2}{n}\left(1 + \frac{(x^* - \overline{X})^2}{\widehat{\mathrm{Var}}_n(X)}\right)$$

Now we'll conduct **uncertainty quantification** for $\hat{\mu}(x^*)$ by considering its distribution and corresponding confidence intervals.

Exercise 0.6. What is the *distribution* of $\hat{\mu}(x^*)|X$?

Exercise 0.7. Construct a $100(1-\alpha)\%$ confidence interval for $\mu(x^*)$.

Exercise 0.8. Return to our "population question": What is the estimated average cancer volume for the population of men with a PSA level of 10 ng/mL? To answer this, construct a 95% confidence interval using the following values:

```
n = nrow(prosdat); n

## [1] 97

coef(proslm)

## (Intercept)    log(psa)
##  -0.5085796   0.7499189

xstar = log(10); xstar

## [1] 2.302585

xbar = mean(log(prosdat$psa))
xbar

## [1] 2.478387

var_n_x = var(log(prosdat$psa))*((n-1)/n)
var_n_x

## [1] 1.318739

summary(proslm)$sigma

## [1] 0.8040745

qt(1 - 0.05/2, df = nrow(prosdat) - 2)

## [1] 1.985251
```

Luckily, R has a built-in command for this confidence interval:

```
interval <- predict(proslm, newdata = data.frame(psa = 10),
                    interval = "confidence", level = 0.95)
interval

##         fit      lwr      upr
## 1 1.218172 1.054206 1.382139

exp(interval)

##         fit      lwr      upr
## 1 3.381003 2.869694 3.983415
```

Note the syntax for how the new value of the predictor is specified via the `newdata` argument using the `data.frame()` function. The variable name must be given **exactly** as it is used in the model (in this case `psa`).

We exponentiate the interval because the model predicts *log* cancer volume. The exponentiated values are confidence intervals for cancer volume.

In text, we would report this as follows:

> Our model estimates that the average cancer volume for patients with a PSA level of 10 ng/mL is 3.38 cc (95% CI [2.87, 3.98]).

Notice we have given units, estimate, and confidence interval together.

Prediction Interval for an Observation

Now we'll consider uncertainty quantification when predicting the outcome for **one observation**, rather than predicting the mean outcome for a population. We will still consider a covariate value $X = x^*$, but now we will form a **prediction interval** for the predicted outcome Y^* (rather than the mean outcome $\mathbb{E}[Y|X = x^*]$). This is relevant when we observe a new observation $X = x^*$, but don't know the corresponding Y^*.

Based on the linear regression model, our best prediction is again:

$$\widehat{\mu}(x^*) = \widehat{\beta}_0 + \widehat{\beta}_1 x^*$$

Key Insight: To quantify the uncertainty for *this particular prediction*, we need to derive $\mathrm{Var}(\widehat{\mu}(x^*) - Y^*|X)$, rather than just $\mathrm{Var}(\widehat{\mu}(x^*)|X)$.

Exercise 0.9. Explain why the above variance is the relevant quantity to measure the precision of the prediction for Y^*.

Exercise 0.10. Derive $\mathrm{Var}(\hat{\mu}(x^*) - Y^*|X)$. What happens as n grows?

Exercise 0.11. Derive the appropriate $100(1-\alpha)\%$ **prediction interval**.

Exercise 0.12. Return to our "individual question": If an individual has a PSA level of 10 ng/mL, what is our prediction of this man's cancer volume? To answer this, construct an appropriate 95% prediction interval.

```
pred <- predict(proslm, newdata = data.frame(psa = 10),
                interval = "prediction", level = 0.95)
pred


##        fit        lwr      upr
## 1 1.218172 -0.3865163 2.822861


exp(pred)


##        fit       lwr      upr
## 1 3.381003 0.6794196 16.82492
```

Again, we exponentiate to get cancer volume on the original scale. We report this in text as:

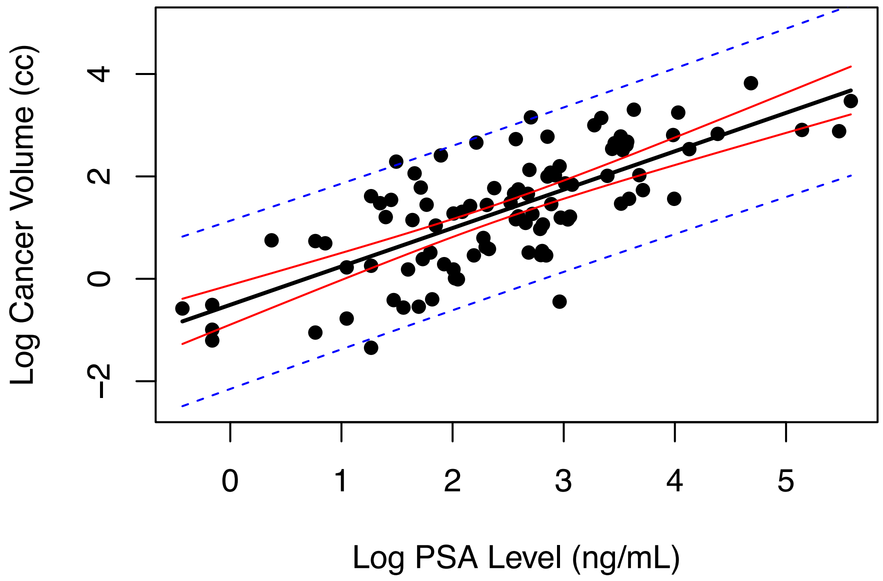
> Based on this model, we predict that the cancer volume of a patient with a PSA level of 10 ng/mL is 3.38 cc (95% prediction interval [0.68, 16.82]).

Again, we have given units, estimate, and prediction interval together.

Exercise 0.13. How does this prediction interval compare to the confidence interval we computed earlier?

We can visually compare the confidence interval and prediction interval in R as follows:

```
#make scatterplot
plot(log(cavol) ~ log(psa), data = prosdat,
     xlab = "Log PSA Level (ng/mL)",
     ylab = "Log Cancer Volume (cc)", pch = 16,
     ylim = c(-2.5, 5))
#compute confidence interval and prediction intervals
confInt = predict(proslm, interval = "confidence", level = 0.95)
predInt = predict(proslm, interval = "prediction", level = 0.95)
#draw regression line
lines(log(prosdat$psa), predict(proslm), col = "black", lwd = 2)
#draw confidence interval
lines(log(prosdat$psa), confInt[,"lwr"], col = "red")
lines(log(prosdat$psa), confInt[,"upr"], col = "red")
#draw prediction interval
lines(log(prosdat$psa), predInt[,"lwr"], col = "blue", lty = 2)
lines(log(prosdat$psa), predInt[,"upr"], col = "blue", lty = 2)
```


Log Cancer Volume (cc)
Log PSA Level (ng/mL)