Assignment 1-2

JARED YU

1. In a simple linear regression analysis, n *independent* paired data $(y_1, x_1), \cdots, (y_n, x_n)$ are fitted to the model
$$y_i = \beta_1(x_i - \mu) + \varepsilon_i, \qquad i = 1, \cdots, n,$$
where x is the only non-random independent variable (or so-called regressor), μ is a known real number, and ε is the random error that has mean zero and unknown constant variance σ^2. Before the data for (y, x) are available, we need to construct estimators for the parameters.
   a. Construct the ordinary least squares (OLS) estimator of β_1.

Ans:

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

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_1(x_i - \mu)\right)(x_i - \mu) \overset{\text{set to}}{\triangleq} 0$$

(*Note: Let* $\sum_i(\cdot) = \sum_{i=1}^{n}(\cdot)$)

$$\rightarrow \sum_i y_i(x_i - \mu) - \hat{\beta}_1\sum_i (x_i - \mu)^2 = 0$$

$$\rightarrow \sum_i y_i(x_i - \mu) = \hat{\beta}_1\sum_i (x_i - \mu)^2$$

$$\rightarrow \boxed{\hat{\beta}_1 = \frac{\sum_i y_i(x_i - \mu)}{\sum_j (x_j - \mu)^2} = \sum_i c_i y_i}$$

where $\boxed{c_i = \frac{(x_i - \mu)}{\sum_j (x_j - \mu)^2}}$

   b. Construct the variance of the OLS estimator of β_1 in a) and construct an unbiased estimator of this variance.

Ans:

The following shows the calculation of the variance of $\hat{\beta}_1$:

$$Var(\hat{\beta}_1) = Var\left(\sum_i c_i y_i\right) = \sum_i c_i^2 Var(y_i) = \sum_i c_i^2\sigma^2 = \sigma^2\sum_i c_i^2$$

$$= \sigma^2\frac{\sum_i (x_i - \mu)^2}{\left[\sum_j (x_j - \mu)^2\right]^2} \boxed{= \frac{\sigma^2}{\sum_j (x_j - \mu)^2}}$$

This is possible since the c_i term has only the constant term μ and fixed $x's$, it can factor out of the variance and the covariance will zero out. The sum of the variance is the variance of the sum (where $Var(y_i) = \sigma^2$), based on the assumption that the y_i terms are uncorrelated.

Next, an unbiased estimator of the variance will be constructed. It requires that an estimator for σ^2, let the estimator of σ^2 be denoted as $\hat{\sigma}^2$.

First, SS_{Res} is as follows:

$$SS_{Res} = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2$$

The variance of y_i is as follows:

$$\sigma^2 = Var(y_i) = E\big(y_i - E(y_i)\big)^2 = E(y_i - \beta_1(x_i - \mu))^2$$

Next, the expectation of SS_{Res} is as follows:
(*Note: Let* $\sum_k(\cdot) = \sum_{k=1}^n(\cdot)$)

$$E(SS_{Res}) = \sum_i E\left[\left(y_i - \hat{\beta}_1(x_i-\mu)\right)^2\right] = \sum_i E\left[\left(y_i - \sum_k c_k y_k\,(x_i-\mu)\right)^2\right]$$

$$= \sum_i E\left[\left(y_i - \beta_1(x_i-\mu) + \beta_1(x_i-\mu) - \sum_k c_k y_k (x_i-\mu)\right)^2\right]$$

$$= \sum_i E\Bigg[\left(y_i - \beta_1(x_i-\mu)\right)^2 + 2\left(y_i - \beta_1(x_i-\mu)\right)\left(\beta_1(x_i-\mu) - (x_i-\mu)\sum_k c_k y_k\right) + \left(\beta_1(x_i-\mu) - (x_i-\mu)\sum_k c_k y_k\right)^2\Bigg]$$

$$= \sum_i \Bigg\{E\left[\left(y_i - \beta_1(x_i-\mu)\right)^2\right] + 2E\left[\left(y_i - \beta_1(x_i-\mu)\right)\left(\beta_1(x_i-\mu) - (x_i-\mu)\sum_k c_k y_k\right)\right] + E\left[\left(\beta_1(x_i-\mu) - (x_i-\mu)\sum_k c_k y_k\right)^2\right]\Bigg\}$$

$$= \sum_i \Bigg\{\sigma^2 + 2E\left[\beta_1 y_i(x_i-\mu) - y_i(x_i-\mu)\sum_k c_k y_k - \beta_1^2(x_i-\mu)^2 + \beta_1(x_i-\mu)^2\sum_k c_k y_k\right] + \beta_1^2(x_i-\mu)^2 - 2\beta_1(x_i-\mu)^2 E\left[\sum_k c_k y_k\right] + E\left[(x_i-\mu)\sum_k c_k y_k\right]^2\Bigg\}$$

$$= \sum_i \Bigg\{\sigma^2 + 2\left[\beta_1 E(y_i)(x_i-\mu) - (x_i-\mu)\sum_{k=1}^n c_k E(y_i y_k) - \beta_1^2(x_i-\mu)^2 + \beta_1(x_i-\mu)^2\sum_{k=1}^n c_k E(y_k)\right] + \beta_1^2(x_i-\mu)^2 - 2\beta_1(x_i-\mu)^2 E\left[\sum_k c_k y_k\right] + E\left[(x_i-\mu)\sum_k c_k y_k\right]^2\Bigg\}$$

To simplify the process, first let's focus on the green section:

$$\beta_1 E(y_i)(x_i-\mu) - (x_i-\mu)\sum_k c_k E(y_i y_k) - \beta_1^2(x_i-\mu)^2 + \beta_1(x_i-\mu)^2\sum_k c_k E(y_k)$$
$$= \beta_1^2(x_i-\mu)^2 - (x_i-\mu)c_i\sigma^2$$
$$-\beta_1^2(x_i-\mu)^2\sum_k c_k(x_k-\mu) - \beta_1^2(x_i-\mu)^2 + \beta_1^2(x_i-\mu)^2\sum_k c_k(x_k-\mu)$$
$$= -(x_i-\mu)c_i\sigma^2$$
$$= -(x_i-\mu)\frac{x_i-\mu}{\sum_i(x_i-\mu)^2}\sigma^2$$
$$= -\frac{(x_i-\mu)^2}{\sum_j(x_j-\mu)^2}\sigma^2$$

Next, let's focus on the blue section:

$$\beta_1^2(x_i-\mu)^2 - 2\beta_1(x_i-\mu)^2 E\left[\sum_k c_k y_k\right] + E\left[(x_i-\mu)\sum_k c_k y_k\right]^2$$
$$= \beta_1^2(x_i-\mu)^2 - 2\beta_1(x_i-\mu)^2\sum_k c_k E(y_k) + (x_i-\mu)^2 E\left[\sum_k c_k y_k\right]^2$$

(*Note: Let* $\sum_l(\cdot) = \sum_{l=1}^n(\cdot)$)

$$= \beta_1^2(x_i-\mu)^2 - 2\beta_1^2(x_i-\mu)^2\sum_k c_k(x_k-\mu) + (x_i-\mu)^2 E\left[\sum_k\sum_l c_k c_l y_k y_l\right]$$

Notice that: $\sum_k c_k(x_k-\mu) = 1$

$$= \beta_1^2(x_i-\mu)^2 - 2\beta_1^2(x_i-\mu)^2 + (x_i-\mu)^2\sum_k\sum_l c_k c_l E(y_k y_l)$$
$$= -\beta_1^2(x_i-\mu)^2 + \beta_1^2(x_i-\mu)^2\sum_k\sum_l c_k c_l(x_k-\mu)(x_l-\mu) + (x_i-\mu)^2\sigma^2\sum_k c_k^2$$
$$= -\beta_1^2(x_i-\mu)^2 + \beta_1^2(x_i-\mu)^2\sum_k\sum_l\left[\frac{x_k-\mu}{\sum_m(x_m-\mu)^2}\right]\left[\frac{x_l-\mu}{\sum_m(x_m-\mu)^2}\right](x_k-\mu)(x_l-\mu)$$
$$+\frac{(x_i-\mu)^2\sigma^2}{\sum_m(x_m-\mu)^2}$$

(*Note: Let* $\sum_m(\cdot) = \sum_{m=1}^n(\cdot)$)

$$= -\beta_1^2(x_i-\mu)^2 + \beta_1^2(x_i-\mu)^2\left[\frac{1}{\sum_m(x_m-\mu)^2}\right]^2\sum_k\sum_l(x_k-\mu)^2(x_l-\mu)^2 + \frac{(x_i-\mu)^2\sigma^2}{\sum_m(x_m-\mu)^2}$$
$$= -\beta_1^2(x_i-\mu)^2 + \beta_1^2(x_i-\mu)^2\left[\frac{1}{\sum_m(x_m-\mu)^2}\right]^2\sum_k(x_k-\mu)^2\sum_l(x_l-\mu)^2 + \frac{(x_i-\mu)^2\sigma^2}{\sum_m(x_m-\mu)^2}$$
$$= \frac{(x_i-\mu)^2\sigma^2}{\sum_m(x_m-\mu)^2}$$

From this is follows that:

$$E(SS_{Res})$$

$$= \sum_i \left\{ \sigma^2 + 2\left[\beta_1 E(y_i)(x_i-\mu) - (x_i-\mu)\sum_k c_k E(y_i y_k) - \beta_1^2 (x_i-\mu)^2 + \beta_1 (x_i-\mu)^2 \sum_k c_k E(y_k) \right] + \beta_1^2 (x_i-\mu)^2 - 2\beta_1 (x_i-\mu)^2 E\left[\sum_k c_k y_k\right] + E\left[(x_i-\mu)\sum_k c_k y_k\right]^2 \right\}$$

$$= \sum_i \left\{ \sigma^2 + 2\left[-\frac{(x_i-\mu)^2}{\sum_j (x_j-\mu)^2}\sigma^2 \right] + \frac{(x_i-\mu)^2\sigma^2}{\sum_m (x_m-\mu)^2} \right\}$$

$$= \sum_i \sigma^2 \left(1 - \frac{(x_i-\mu)^2}{\sum_j (x_j-\mu)^2}\right) = (n-1)\sigma^2$$

It follows that the unbiased estimator of σ^2, $\hat{\sigma}^2$ can be rewritten as:

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-1} = \frac{1}{n-1}\sum_i (y_i - \hat{y}_i)^2$$

Therefore, the unbiased estimator of $Var(\hat{\beta}_1)$ is as follows:

$$\boxed{\widehat{Var(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum_i (x_i-\mu)^2}}$$

Since the following holds true:

$$E\left(\widehat{Var(\hat{\beta}_1)}\right) = \frac{E(\hat{\sigma}^2)}{\sum_i (x_i-\mu)^2} = \frac{\sigma^2}{\sum_i (x_i-\mu)^2} = Var(\hat{\beta}_1)$$

2. In Problem 1, add the intercept term β_0 to the model. Then do a) and b).

Ans:

The model with the intercept added is as follows,

$$y = \beta_0 + \beta_1 (x-\mu) + \varepsilon.$$

The least squares criterion can be written as follows,

$$S(\beta_0, \beta_1) = \sum_i [y_i - \beta_0 - \beta_1(x_i-\mu)]^2.$$

Then the least-squares estimator of β_0, $\hat{\beta}_0$ is as follows:

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_i \left[y_i - \hat{\beta}_0 - \hat{\beta}_1(x_i-\mu)\right] \overset{set\ to}{\triangleq} 0$$

$$\rightarrow \sum_i y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_i (x_i-\mu) = 0$$

$$\rightarrow \hat{\beta}_0 = \frac{1}{n}\left[\sum_i y_i - \hat{\beta}_1 \sum_i (x_i-\mu)\right]$$

$$\boxed{= \bar{y} - \hat{\beta}_1(\bar{x}-\mu)}$$

The least-squares estimator of β_1, $\hat{\beta}_1$ is as follows:

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_i \left[y_i - \hat{\beta}_0 - \hat{\beta}_1(x_i-\mu)\right](x_i-\mu) \overset{set\ to}{\triangleq} 0$$

$$\rightarrow \sum_i y_i(x_i-\mu) - \hat{\beta}_0(x_i-\mu) - \hat{\beta}_1(x_i-\mu)^2 = 0$$

$$\rightarrow \hat{\beta}_1 \sum_i (x_i-\mu)^2 = \sum_i \left[y_i(x_i-\mu) - \hat{\beta}_0(x_i-\mu)\right]$$

$$\rightarrow \hat{\beta}_1 \sum_i (x_i-\mu)^2 = \sum_i \left[y_i(x_i-\mu) - \{\bar{y} - \hat{\beta}_1(\bar{x}-\mu)\}(x_i-\mu)\right]$$

$$\rightarrow \hat{\beta}_1 \sum_i (x_i-\mu)^2 = \sum_i \left[y_i(x_i-\mu)\right] - \left[\bar{y} - \hat{\beta}_1(\bar{x}-\mu)\right]\sum_i (x_i-\mu)$$

$$\rightarrow \hat{\beta}_1 \sum_i (x_i-\mu)^2 = \sum_i \left[y_i(x_i-\mu)\right] - \bar{y}\sum_i (x_i-\mu) + \hat{\beta}_1(\bar{x}-\mu)\sum_i (x_i-\mu)$$

Here, $\sum_i (x_i-\mu) = \sum_i x_i - n\mu = n\,\bar{x} - n\mu = n(\bar{x}-\mu)$

$$\rightarrow \hat{\beta}_1 \sum_i (x_i-\mu)^2 = \sum_i \left[y_i(x_i-\mu)\right] - \bar{y}\sum_i (x_i-\mu) + n\hat{\beta}_1(\bar{x}-\mu)^2$$

$$\rightarrow \hat{\beta}_1 \sum_i (x_i-\mu)^2 - n\hat{\beta}_1(\bar{x}-\mu)^2 = \sum_i (y_i-\bar{y})(x_i-\mu)$$

$$\rightarrow \hat{\beta}_1 \left[\sum_i (x_i-\mu)^2 - n(\bar{x}-\mu)^2\right] = \sum_i (y_i-\bar{y})(x_i-\mu)$$

$$\rightarrow \hat{\beta}_1 = \frac{\sum_i (y_i-\bar{y})(x_i-\mu)}{\sum_i (x_i-\mu)^2 - n(\bar{x}-\mu)^2} \boxed{= \frac{\sum_i (y_i-\bar{y})x_i}{\sum_i (x_i-\bar{x})^2}}$$

Since

$$\sum_i (y_i-\bar{y})(x_i-\mu) = \sum_i (y_i-\bar{y})x_i - \mu\sum_i (y_i-\bar{y}) = \sum_i (y_i-\bar{y})x_i - 0 = \sum_i (y_i-\bar{y})x_i$$

and

$$\sum_i (x_i-\bar{x})^2 = \sum_i \left[(x_i-\mu) - (\bar{x}-\mu)\right]^2$$

$$= \sum_i \left[(x_i-\mu)^2 - 2(x_i-\mu)(\bar{x}-\mu) + (\bar{x}-\mu)^2\right]$$

$$= \sum_i (x_i-\mu)^2 - 2(\bar{x}-\mu)\sum_i (x_i-\mu) + n(\bar{x}-\mu)^2$$

$$= \sum_i (x_i-\mu)^2 - 2(\bar{x}-\mu)[n(\bar{x}-\mu)] + n(\bar{x}-\mu)^2$$

$$= \sum_i (x_i-\mu)^2 - 2n(\bar{x}-\mu)^2 + n(\bar{x}-\mu)^2$$

$$= \sum_i (x_i-\mu)^2 - n(\bar{x}-\mu)^2$$

In trying to derive the variance and an unbiased estimator of the variance, it could be useful to try and re-express the formula in an alternate format:

$$y_i = \beta_0 + \beta_1(x_i-\mu) + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 \tilde{x}_i + \varepsilon_i$$

where $\tilde{x}_i = x_i - \mu$

Then it follows that the estimator of the slope is as follows:

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(\tilde{x}_i - \bar{\tilde{x}})}{\sum_i (\tilde{x}_i - \bar{\tilde{x}})^2}$$

Notice: $\bar{\tilde{x}} = \frac{1}{n}\sum_i (x_i - \mu) = \bar{x} - \mu$

and that:

$$\tilde{x}_i - \bar{\tilde{x}} = (x_i - \mu) - (\bar{x} - \mu) = x_i - \bar{x}$$

Furthermore, $\hat{\beta}_1$ can be rewritten to:

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (\tilde{x}_i - \bar{\tilde{x}})^2} = \frac{\sum_i (y_i - \bar{y}) x_i}{\sum_i (\tilde{x}_i - \bar{\tilde{x}})^2} = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (\tilde{x}_i - \bar{\tilde{x}})^2} = \sum_i d_i y_i$$

where

$$d_i = \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2}$$

and

$$\sum_i (y_i - \bar{y})(x_i - \bar{x}) = \sum_i (y_i - \bar{y}) x_i - \bar{x} \sum_i (y_i - \bar{y}) = \sum_i (y_i - \bar{y}) x_i - 0 = \sum_i (y_i - \bar{y}) x_i$$

Also, for $\hat{\beta}_0$, let the following be shown:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{\tilde{x}} = \bar{y} - \hat{\beta}_1 (\bar{x} - \mu)$$

Next, let's analyze the variance of the coefficients:

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}\right) = Var\left(\sum_i d_i y_i\right) = \sum_i d_i^2 Var(y_i)$$

since d_i consists of x terms that are viewed as constants

$$= \sigma^2 \sum_i d_i^2 = \sigma^2 \sum_i \left[\frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2}\right]^2 \boxed{= \frac{\sigma^2}{\sum_j (x_j - \bar{x})^2}}$$

The unbiased estimator of σ^2 is (based on the textbook results and expanding to the $\tilde{x}$ version):

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 \tilde{x}_i)^2}{n-2} = \frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 (x_i - \mu))^2}{n-2} = \frac{SS_{Res}}{n-2}$$

where

$$SS_{Res} = \sum_i y_i^2 - n\,\bar{y}^2 - \hat{\beta}_1 S_{\tilde{x}y} = \sum_i (y_i - \bar{y})^2 - \hat{\beta}_1 S_{\tilde{x}y}$$

and

$$S_{\tilde{x}y} = \sum_i (\tilde{x}_i - \bar{\tilde{x}})(y_i - \bar{y}) = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = S_{xy}$$

Then it follows that:

$$SS_{Res} = \sum_i (y_i - \bar{y})^2 - \hat{\beta}_1 S_{\tilde{x}y} = \sum_i (y_i - \bar{y})^2 - \hat{\beta}_1 S_{xy}$$

$$= \sum_i (y_i - \bar{y})^2 - \frac{S_{xy}}{S_{xx}} S_{xy} = \sum_i (y_i - \bar{y})^2 - \frac{S_{xy}^2}{S_{xx}} = \sum_i (y_i - \bar{y})^2 - \frac{[\sum_i (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_i (x_i - \bar{x})^2}$$

So, the unbiased estimator of $Var(\hat{\beta}_1)$ is:

$$\boxed{\widehat{Var(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum_j (x_j - \bar{x})^2}}$$

where

$$\hat{\sigma}^2 = \frac{\sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1(x_i - \mu)\right)^2}{n-2}$$

since

$$E\left[\widehat{Var(\hat{\beta}_1)}\right] = \frac{E(\hat{\sigma}^2)}{\sum_j (x_j - \bar{x})^2} = \frac{\sigma^2}{\sum_j (x_j - \bar{x})^2}$$

Next, let us look at the other coefficient, $\hat{\beta}_0$, consider:

$$\begin{aligned} Var(\hat{\beta}_0) &= Var[\bar{y} - \hat{\beta}_1(\bar{x} - \mu)] \\ &= Var(\bar{y}) + \bar{\bar{x}}^2 Var(\hat{\beta}_1) - 2\bar{\bar{x}} Cov(\bar{y}, \hat{\beta}_1) \\ &= Var(\bar{y}) + (\bar{x} - \mu)^2 Var(\hat{\beta}_1) - 2(\bar{x} - \mu) Cov(\bar{y}, \hat{\beta}_1) \end{aligned}$$

Let's examine the above in pieces:

$$Var(\bar{y}) = \frac{1}{n^2} \sum_i Var(y_i) = \frac{\sigma^2}{n}$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

$$\begin{aligned} Cov(\bar{y}, \hat{\beta}_1) &= \frac{1}{n} \sum_i Cov\left(y_i, \sum_j d_j y_j\right) \\ &= \frac{1}{n} \sum_i \sum_j d_j Cov(y_i, y_j) \\ &= \frac{1}{n} \left\{ \sum_{i,j,i=j} d_j Cov(y_i, y_j) + \sum_{i,j,i \neq j} d_j Cov(y_i, y_j) \right\} \\ &= \frac{1}{n} \left\{ \sum_{i,j,i=j} d_j Cov(y_i, y_j) + 0 \right\} \\ &= \frac{1}{n} \sum_i d_i Cov(y_i, y_i) \\ &= \frac{\sigma^2}{n} \sum_i d_i \\ &= \frac{\sigma^2}{n} \sum_i \left[\frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2} \right] \\ &= 0 \end{aligned}$$

Thus, we have:

$$Var(\hat{\beta}_0) = Var(\bar{y}) + (\bar{x} - \mu) Var(\hat{\beta}_1) - 2(\bar{x} - \mu) Cov(\bar{y}, \hat{\beta}_1)$$

$$= \frac{\sigma^2}{n} + (\bar{x} - \mu)\frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} - 0$$

$$\boxed{Var(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{(\bar{x} - \mu)^2}{\sum_i (x_i - \bar{x})^2}\right]}$$

Then, the unbiased estimator of $Var(\hat{\beta}_0)$ is

$$\boxed{\widehat{Var}(\hat{\beta}_0) = \hat{\sigma}^2\left[\frac{1}{n} + \frac{(\bar{x} - \mu)^2}{\sum_i (x_i - \bar{x})^2}\right]}$$

where

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2}$$

and

$$SS_{Res} = \sum_i (y_i - \bar{y})^2 - \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

3. In Problem 1, the μ is a real number but the value is <u>unknown</u>. Please do a) and b).
   a. Construct the ordinary least squares (OLS) estimator of β_1.

<u>Ans:</u>

$$S(\beta_1, \mu) = \sum_i (y_i - \beta_1 x_i - \beta_1 \mu)^2$$

$$\left.\frac{\partial S}{\partial \mu}\right|_{\hat{\mu}} = 2\beta_1 \sum_i (y_i - \beta_1 x_i + \beta_1 \mu) \overset{\text{set to}}{\triangleq} 0$$

$$\rightarrow \beta_1[n\bar{y} - \beta_1 n\bar{x} + n\beta_1 \mu] = 0$$

$$-\beta_1 \mu = \bar{y} - \beta_1 \bar{x}$$

Note: It is assumed that $\beta_1 \neq 0$.

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_1} = -2\sum_i (x_i - \mu)(y_i - \beta_1 x_i - \beta_1 \mu) \overset{\text{set to}}{\triangleq} 0$$

$$\rightarrow -2\left[\sum_i x_i y_i - \beta_1 \sum_i x_i^2 + \beta_1 \mu \sum_i x_i - \mu\left(\sum_i y_i\right) + \mu\beta_1 \sum_i x_i - \beta_1 \mu^2 n\right] = 0$$

$$\rightarrow \sum_i x_i y_i - \beta_1 \sum_i x_i^2 - 2(\bar{y} - \beta_1 \bar{x})\sum_i x_i - \mu\left(\sum_i y_i\right) - \beta_1 \mu^2 n = 0$$

$$\rightarrow \beta_1 \sum_i x_i y_i - \beta_1^2 \sum_i x_i^2 - 2\beta_1(\bar{y} - \beta_1 \bar{x})\sum_i x_i - \beta_1 \mu n\bar{y} - \beta_1^2 \mu^2 n = 0$$

$$\rightarrow \beta_1 \sum_i x_i y_i - \beta_1^2 \sum_i x_i^2 - 2n\bar{x}\beta_1(\bar{y} - \beta_1 \bar{x}) + (\bar{y} - \beta_1 \bar{x})n\bar{y} - (\bar{y} - \beta_1 \bar{x})^2 n = 0$$

$$\rightarrow \beta_1 \sum_i x_i y_i - \beta_1^2 \sum_i x_i^2 - 2n\bar{x}\beta_1\bar{y} + 2n\bar{x}^2\beta_1^2 + +n\bar{y}^2 - \beta_1 n\bar{x}\bar{y} - n\bar{y}^2 + 2n\beta_1\bar{x}\bar{y} - n\beta_1^2\bar{x}^2 = 0$$

$$\rightarrow \beta_1 \sum_i x_i y_i - \beta_1^2 \sum_i x_i^2 - \beta_1 n\bar{x}\bar{y} + n\beta_1^2 \bar{x}^2 = 0$$

$$\rightarrow \beta_1 \left[\sum_i x_i y_i - n\bar{x}\bar{y} - \beta_1^2 \left(\sum_i x_i^2 - n\bar{x}^2\right)\right] = 0$$

Note: It is ignored when $\beta_1 = 0$.

$$\boxed{\hat{\beta}_1 = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2} = \sum_i c_i y_i}$$

where $\boxed{c_i = \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2}}$

Returning back to μ:

$$-\beta_1 \mu = \bar{y} - \beta_1 \bar{x}$$

$$\boxed{\hat{\mu} = \frac{(\bar{y} - \hat{\beta}_1 \bar{x})}{\hat{\beta}_1}}$$

where $\boxed{\hat{\beta}_1 = \sum_i c_i y_i}$

b. Construct the variance of the OLS estimator of β_1 in a) and construct an <u>unbiased</u> estimator of this variance.

<u>Ans:</u>

The following has been shown in part a):

$$\hat{\beta}_1 = \sum_i c_i y_i,$$

where $c_i = \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2}$.

$$Var(\hat{\beta}_1) = \sum_i c_i^2 Var(y_i)$$

Since the c_i term has only the fixed $x's$, it can factor out of the variance.

$$= \sum_i c_i^2 \sigma^2 = \sum_i \frac{(x_i - \bar{x})^2}{\left[\sum_j (x_j - \bar{x})^2\right]^2} \sigma^2 = \frac{\sigma^2}{\sum_j (x_j - \bar{x})^2}$$

Thus, is has been shown that $\boxed{Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_j (x_j - \bar{x})^2}}$.

Next the unbiased estimator will be constructed.

$$SS_{Res} = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \left(y_i - \hat{\beta}_1 (x_i - \hat{\mu})\right)^2$$

$$= \sum_i \left(y_i - \hat{\beta}_1 x_i + \hat{\beta}_1 \hat{\mu}\right)^2 = \sum_i \left(y_i - \hat{\beta}_1 x_i - (\bar{y} - \bar{x}\hat{\beta}_1)\right)^2$$

$$= \sum_i \left(y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})\right)^2$$

$$= \sum_i (y_i - \bar{y})^2 + \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2 - 2\hat{\beta}_1 \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum_i (y_i - \bar{y})^2 + \frac{[\sum_i (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_i (x_i - \bar{x})^2} - 2\frac{[\sum_i (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_i (x_i - \bar{x})^2}$$

$$= \sum_i (y_i - \bar{y})^2 - \frac{[\sum_i (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_i (x_i - \bar{x})^2}$$

$$= \sum_i [\beta_1(x_i - \mu) + \varepsilon_i - \beta_1(\bar{x} - \mu) - \bar{\varepsilon}]^2 - \frac{[\sum_i (x_i - \bar{x})(\beta_1(x_i - \bar{x}) + \varepsilon_i + \bar{\varepsilon})]^2}{\sum_i (x_i - \bar{x})^2}$$

$$= \sum_i [\beta_1(x_i - \bar{x}) + \varepsilon_i - \bar{\varepsilon}]^2 - \frac{[\beta_1 \sum_i (x_i - \bar{x})^2 + \sum_i (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})]^2}{\sum_i (x_i - \bar{x})^2}$$

$$= \beta_1^2 \sum_i (x_i - \bar{x})^2 + 2\beta_1 \sum_i (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) + \sum_i (\varepsilon_i - \bar{\varepsilon})^2 - \beta_1^2 \sum_i (x_i - \bar{x})^2$$
$$- 2\beta_1 \sum_i (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) - \frac{[\sum_i (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})]^2}{\sum_i (x_i - \bar{x})^2}$$

$$= \sum_i (\varepsilon_i - \bar{\varepsilon})^2 - \frac{[\sum_i (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})]^2}{\sum_i (x_i - \bar{x})^2}$$

From here it can be seen that none of the highlighted parts in the section above contain any reference to the variable μ or $\hat{\mu}$. It is the same as the normal SS_{Res} for simple linear regression. Therefore,

$$E(SS_{Res}) = \sigma^2(n-2)$$

or,

$$E(\hat{\sigma}^2) = E\left(\frac{SS_{Res}}{n-2}\right) = \sigma^2$$

Therefore, the unbiased estimator of $Var(\hat{\beta}_1)$ is as follows:

$$\boxed{\widehat{Var(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum_j (x_j - \bar{x})^2}}$$

Since the following holds true:

$$E\left(\widehat{Var(\hat{\beta}_1)}\right) = \frac{E(\hat{\sigma}^2)}{\sum_j (x_j - \bar{x})^2} = \frac{\sigma^2}{\sum_j (x_j - \bar{x})^2} = Var(\hat{\beta}_1)$$

4. Consider a regression model $y = \beta_0 + \beta_1 x + \varepsilon$, where x is a non-random regressor. Discuss whether the ordinary least-squares estimator of the slope β_1 is always unbiased and whether it always has the smallest variance than **any** estimator of β_1, irrespectively of what the value of β_0 is. State assumptions in your discussion. Be careful about the word "**any**".

Ans:

For $\hat{\beta}_1$ to be unbiased, it requires that $E(\hat{\beta}_1) = \beta_1$. For this to be the case, there are two important requirements to that need to be addressed, $E(\varepsilon_i) = 0$ and $\sum_{i=1}^{n} c_i = 0$. The first requirement needs the assumption that the error terms ε_i have a mean of 0. The second requirement needs that the $x's$ are distributed in such a way that the term can exist. More specifically,

$$\sum_{i=1}^{n} c_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2},$$

must be such that the denominator, $\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2$ is nonzero. If the data has the characteristic where $S_{xx} = 0$, then c_i would fail to exist since it's not possible to divide by 0.

To show the steps more carefully, it will be shown how $\hat{\beta}_1$ can be seen as unbiased:
To show that the OLS estimate of β_1 is unbiased, the following two equations must hold true:

$$E\left(\hat{\beta}_1\right) = \beta_1. \tag{3}$$

The formula for $\hat{\beta}_1$ are is follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}. \tag{4}$$

Furthermore, the formula for $\hat{\beta}_1$ can be further rewritten as follows:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i\,, \tag{5}$$

where $c_i = \frac{(x_i - \bar{x})}{S_{xx}}$. To show the unbiasedness, we will first start with $\hat{\beta}_1$ (pp. 18-19):

$$E\left(\hat{\beta}_1\right) = E\left(\sum_{i=1}^{n} c_i y_i\right) \tag{6}$$

In equation (6), we have simply applied the expectation to $\hat{\beta}_1$.

$$= \sum_{i=1}^{n} c_i E(y_i) \tag{7}$$

In equation (7), we move the expectation into the summation, based on the *linearity of expectation*. Furthermore, we are treating c_i as a constant, fixed term, since it consists entirely of x. It is important to note that this step is not possible if x is also a random variable.

$$= \sum_{i}^{n} c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \tag{8}$$

In equation (8), the $E(y_i)$ is rewritten as $\beta_0 + \beta_1 x_i$, since $E(\varepsilon_i) = 0$. Then the summation is distributed to each part of this new term.

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\left[\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2\right]} = \frac{1}{\left[\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2\right]}\left(\sum_{i=1}^{n} x_i - n\bar{x}\right) = 0 \tag{9}$$

In equation (9), it is shown that the $\sum_{i=1}^{n} c_i$ term equals 0, since $\sum_{i=1}^{n} x_i = n\bar{x}$.

$$\sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})x_i}{\left[\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2\right]} = \frac{1}{\left[\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2\right]}\sum_{i=1}^{n}(x_i - \bar{x})x_i \tag{10}$$

In equation (10), the formula $\sum_{i=1}^{n} c_i x_i$ has been expanded. To show that it is equivalent to 1, it must be shown that the numerator $\sum_{i=1}^{n}(x_i - \bar{x})x_i$ is equivalent to the denominator $\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2$. This will be shown as follows:

$$\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})=\sum_{i=1}^{n}(x_i-\bar{x})x_i-\bar{x}\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}(x_i-\bar{x})x_i-0. \qquad (11)$$

So, it follows that $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i x_i = 1$, therefore equation (8) evaluates to β_1. Thus far it has been shown then that $\hat{\beta}_1$ is an unbiased estimator. From here it has been shown that $\hat{\beta}_1$ derived using OLS is always unbiased given the conditions of $E(\varepsilon_i) = 0$ and $\sum_{i=1}^{n} c_i = 0$.

Another important assumption in the model is that the variance of each ε_i terms are all σ^2, in other words they're homoscedastic. It can also be said that they must be uncorrelated. Without this assumption, then the Gauss-Markov theorem for the OLS estimators does not hold. The assumptions of the Gauss-Markov theorem are that $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$, and uncorrelated errors.

The Gauss-Markov theorem makes the case that the OLS coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, given that the assumptions are met, are unbiased and have minimum variance in comparison to other unbiased estimators that are linear combinations of the observations y_i. This definition makes it clear that it only has the minimum variance in comparison to other estimators that are linear combinations of y_i. This excludes certain other models which may obtain a smaller variance with a more complicated structure.

It is worth noting, that based on my entire discussion response to Module 2, that I had stated it is possible to construct OLS estimators without the constant variance assumption holding true within the data. Therefore, it is possible to construct OLS estimates without this assumption. The issue is that the OLS estimates will no longer have the property of best linear unbiased estimators (BLUE) as outlined in the Gauss-Markov theorem. They will still have the unbiased property, but they will not have the minimum variance amongst other unbiased estimators that are linear combinations of the y_i.

5. Use any math/stat software (e.g., www.numbergenerator.org/randomnumbergenerator) of your choice to find a random number generator to randomly select 15 rows of Table for Problem 2.18 (page 63-64) of Textbook and then do (a), (b), (c), (d).
   State assumptions for all steps in your analyses.

Chosen rows: 1, 2, 3, 4, 5, 7, 9, 11, 13, 14, 16, 17, 18, 19, 21.

a. Fit the simple linear regression model to these data.

In the following, the amount a company spends on advertising are represented by $x_i's$ and the retained impressions are represented by the $y_i's$.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \approx 14.4766$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \approx 0.5054$$

Some important assumptions are that the error term ε has mean zero, constant variance and so are uncorrelated. Also, the error terms need to be independent. This also follows from the thinking that x is nonrandom, but such an assumption is not certain and it's possibly a random variable also. In that case, further assumptions would be required for the model to hold. For example, the conditional expectation of the error term w.r.t. x also needs to have mean zero.

b. Is there a significant relationship between the amount a company spends on advertising and retained impressions? Justify your answer statistically.

To test whether there's a significant relationship between the amount a company spends on advertising and retained impressions, a hypothesis test will be done on β_1 to see if it's equal to 0. If it can be shown that statistically $\beta_1 = 0$, then it would be concluded that there is no linear relationship between the two variables. To reject the null hypothesis, it must be shown that $|t_0| > t_{1-\frac{\alpha}{2},n-2}$, where t_0 is the test statistic (shown below) and $t_{1-\frac{\alpha}{2},n-2}$ is the critical value of the Student's *t*-distribution at the $1-\frac{\alpha}{2}$ percentage point and $n-2$ is the degrees of freedom.
Note: The textbook uses $t_{\frac{\alpha}{2},n-2}$ to refer to the same critical value. The difference is $\frac{\alpha}{2}$ versus $1-\frac{\alpha}{2}$, but this is simply differing notation for the same t-table.

$$H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$$

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

$$se(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}} = \sqrt{\frac{\frac{SS_{res}}{n-2}}{S_{xx}}} = \sqrt{\frac{\frac{SS_T - \hat{\beta}_1 S_{xy}}{n-2}}{S_{xx}}} = \sqrt{\frac{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2 - \hat{\beta}_1 S_{xy}}{n-2}}{S_{xx}}} = \sqrt{\frac{259.815}{43912.42}}$$
$$\approx 0.0769$$

$$\rightarrow t_0 \approx \frac{0.5054}{0.0769} \approx 6.5711$$

The critical value of $t_{0.975,2}$ is approximately 4.3027. Since $|t_0| > t_{1-\frac{\alpha}{2},n-2}$, the decision is to reject the null hypothesis at the $\alpha = 0.05$ confidence level. The conclusion then is that there's a 95% probability that $\beta_1 \neq 0$ and that there's a significant relationship between the amount a company spends on advertising and retained impressions.
Note: Changing α to 0.01 leads to the critical value being approximately 9.9248. Therefore, the conclusion would be to fail to reject the null hypothesis that $\beta_1 = 0$ at the 0.01 confidence level.

c. Construct the 95% confidence and prediction bands for these data.

The assumption is that the errors, ε_i, are normally and independently distributed. Then the sampling distribution of $\frac{(\hat{\beta}_1-\beta_1)}{se(\hat{\beta}_1)}$ is a Student's *t*-distribution with $n-2$ degrees of freedom. The $100(1-\alpha)$ percent confidence interval (CI) of
First the 95% confidence bands for the data will be created. To do this, we are looking to estimate the mean response $E(y)$ for a value from x. Then let x_0 be a level from x that we wish to estimate the mean response with, $E(y|x_0)$. An assumption it that x_0 is within the range of the original data of x. The following is an unbiased point estimator of $E(y|x_0)$,

$$E\widehat{(y|x_0)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

The point estimator $\hat{\mu}_{y|x_0}$ is a normally distributed random variable since it's a linear combination of the observations y_i, where it's assumed that the $y_i \sim Normal(\beta_0 + \beta_1 x_i, \sigma^2)$. The variance of $\hat{\mu}_{y|x_0}$ is as follows (proof on p. 31 of the textbook),

$$Var(\hat{\mu}_{y|x_0}) = \sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right].$$

To derive this result, it must be that $Cov(\bar{y}, \hat{\beta}_1) = 0$.Then the sampling distribution of the standardized point estimator,

$$\frac{\hat{\mu}_{y|x_0} - E(y|x_0)}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}},$$

is a Student's t-distribution with $n - 2$ degrees of freedom. Therefore, the $100(1 - \alpha)$ percent CI of the mean response at point $x = x_0$ is as follows,

$$\hat{\mu}_{y|x_0} - t_{1-\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq E(y|x_0)$$
$$\leq \hat{\mu}_{y|x_0} + t_{1-\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$

Each of the variables in the confidence interval (i.e., $t_{1-\frac{\alpha}{2},n-2}$, MS_{Res}, S_{xx}, $\bar{x}$, $\hat{\beta}_0$, and $\hat{\beta}_1$) have been calculated already in the above problems. The result of this calculation can be seen below in Table 1. Furthermore, a graph of the confidence band can be seen below in Figure 1.
Next, the process will be shown on constructing a 95% prediction interval. In the case of a new point x_0 that isn't part of the original data, the point estimate is as follows,

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

This is the same value for $\widehat{E(y|x_0)}$ as seen above with the confidence interval. Then the random variable $\psi = y_0 - \hat{y}_0$ follows a normal distribution with mean zero and the following variance,

$$Var(\psi) = Var(y_0 - \hat{y}_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right].$$

An assumption is that the covariance zeroes out because the point y_0 is independent of $\hat{y}_0$. By predicting y_0 with $\hat{y}_0$, the standard error of ψ can then be used for a prediction interval. The $100(1 - \alpha)$ prediction interval for x_0 then is as follows,

$$\hat{y}_0 - t_{1-\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq y_0$$
$$\leq \hat{y}_0 + t_{1-\frac{\alpha}{2},n-2}\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$

Like before, all the variables have been calculated already in the previous work shown. The result of this calculation can be seen below in Table 1. Furthermore, a graph of the prediction band can be seen below in Figure 1.

PI Lower Bound	CI Lower Bound	x_0	CI Upper Bound	PI Upper Bound	y
-55.92908	-5.563938	5.0	39.57167	89.93681	12.0
-55.53199	-5.069881	5.7	39.78524	90.24735	10.0
-55.30540	-4.788246	6.1	39.90796	90.42512	4.4
-54.45777	-3.736643	7.6	40.37271	91.09384	12.3
-53.55724	-2.623061	9.2	40.87656	91.81075	23.4
-47.95967	4.189374	19.3	44.27420	96.42325	11.7
-47.35919	4.906223	20.4	44.66934	96.93475	21.4
-46.00118	6.515111	22.9	45.58769	98.10399	21.9
-43.84792	9.027485	26.9	47.11891	99.99432	38.0
-43.79440	9.089277	27.0	47.15821	100.04189	40.8
-31.83828	21.853025	50.1	57.74621	111.43752	32.1
-20.27769	31.828172	74.1	72.03261	124.13848	99.6
-16.48111	34.635407	82.4	77.61583	128.73235	60.8
12.66040	52.673942	154.9	132.86741	172.88095	88.9
23.23938	58.950533	185.9	157.92866	193.63981	92.4

Table 1 The table shows the upper and lower bounds for the confidence interval (CI) and prediction interval (PI) of the data. The x_0 indicates the point from which the calculations are based upon. This value is identical to the x variable from the dataset. The y variable is also included as a reference. (Note: the x variable is the amount a company spent on advertising in millions of dollars and the y variable is the retained impressions per week in millions.)

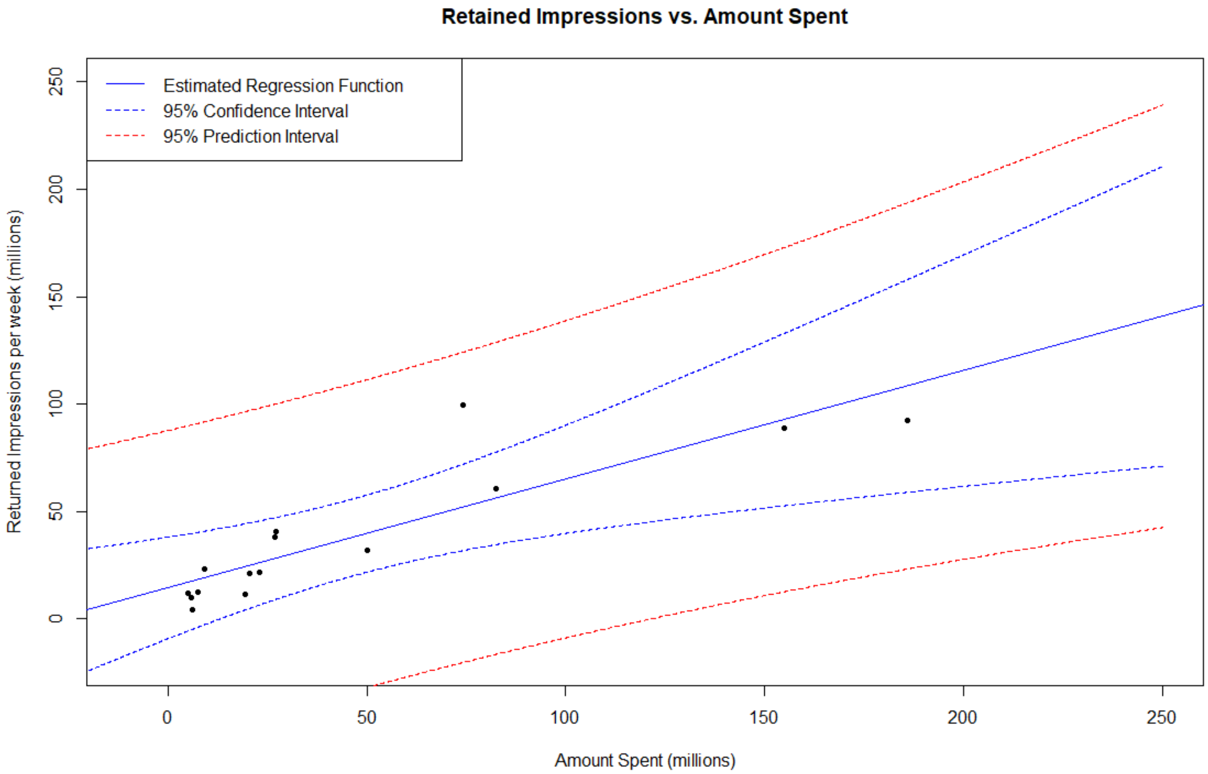

Figure 1 The above figure shows a scatter plot of the retained impressions against the amount spent from the dataset. The points represent each of the companies. The solid blue line is the estimated regression function $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, the dotted blue line is the confidence interval for $E\widehat{(y|x_0)}$ and the dotted red line is the prediction interval for $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

d. Give the 95% confidence and prediction intervals for the number of retained impressions for MCI.

The company MCI has $x = 26.9$ and $y = 50.7$ for the two x and y variables in the dataset. It's worth noting that this company wasn't included in the original dataset since it was not sampled by the random generator. Another important note is that these values aren't outside the range of the original data.
The 95% confidence interval for MCI is as follows:

$$\hat{\mu}_{y|x_0=26.9} \approx 14.4766 + 0.5054 \cdot 26.9 = 28.0732$$

$$28.0732 - 4.3026\sqrt{259.815\left(\frac{1}{15} + \frac{(26.9 - 46.5)^2}{43{,}912.42}\right)} \leq E(y|x_0 = 26.9)$$

$$\leq 28.0732 + 4.3026\sqrt{259.815\left(\frac{1}{15} + \frac{(26.9 - 46.5)^2}{43{,}912.42}\right)}$$

$$\boxed{\approx [9.0275, 47.1189]}$$

The 95% prediction interval for MCI is as follows:

$$\hat{y}_0 \approx 14.4766 + 0.5054 \cdot 26.9 = 28.0732$$

$$28.07324.3026\sqrt{259.815\left(1+\frac{1}{15}+\frac{(26.9-46.5)^2}{43{,}912.42}\right)} \le y_0$$
$$\le 28.0732+4.3026\sqrt{259.815\left(1+\frac{1}{15}+\frac{(26.9-46.5)^2}{43{,}912.42}\right)}.$$
$$\boxed{\approx [-43.8479, 99.9943]}$$

Supplemental Exercises

1. Create a hypothetical data set and then perform simple linear regression analysis and the corresponding inverse regression analysis.

A hypothetical dataset was generated in the following manner. A true slope was set to 1.5 and a true intercept was set to 5. Next, 15 points were generated from the normal distribution with a mean of 25 and a standard deviation of 1.5. Another 15 points were generated from the normal distribution with a mean of 0 and a standard deviation of 1.2. The first set of points represent the hypothetical x values and the second set represent the corresponding ε terms. The set of y values were generated by multiplying the x values with the slope and adding the intercept and error terms.

Using simple linear regression, the following are the resulting regression coefficients:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \approx 6.2517$$
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \approx 1.4466$$

Testing the significance of the regression, the following hypothesis test was conducted:

$$H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$$
$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$
$$se(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}} = \sqrt{\frac{\frac{SS_{res}}{n-2}}{S_{xx}}} = \sqrt{\frac{\frac{SS_T - \hat{\beta}_1 S_{xy}}{n-2}}{S_{xx}}} = \sqrt{\frac{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2 - \hat{\beta}_1 S_{xy}}{n-2}}{S_{xx}}} = \sqrt{\frac{0.7763}{37.9249}}$$
$$\approx 0.1431$$
$$\rightarrow t_0 \approx \frac{1.4466}{0.1431} \approx 10.1113$$

The critical value of $t_{0.995,2}$ is approximately 9.9248. Since $|t_0| > t_{1-\frac{\alpha}{2}, n-2}$, the decision is to reject the null hypothesis at the $\alpha = 0.01$ confidence level. The conclusion then is that there's a 99% probability that $\beta_1 \neq 0$ and that there's a significant relationship between the y and x variables.

The confidence and prediction bands were also generated for each of the 15 points in the hypothetical dataset. They can be seen below in Table 2. This was done in the same style as it was done before with the textbook data. The model is also plotted in Figure 2 (further below).

PI Lower Bound	CI Lower Bound	x_0	CI Upper Bound	PI Upper Bound	y
29.37597	34.87839	22.69007	43.27246	48.77489	38.76599
30.49983	36.40517	23.27851	43.44816	49.35351	39.08863
31.09826	37.22457	23.60715	43.57958	49.70589	40.86360
31.44182	37.69343	23.80149	43.67298	49.92459	42.00515
32.62778	39.25844	24.51065	44.15974	50.79039	42.06864
32.69227	39.33886	24.55118	44.19656	50.84316	41.88284
32.70296	39.35213	24.55792	44.20281	50.85197	41.99688
32.71546	39.36761	24.56581	44.21015	50.86230	40.30719
33.36926	40.13055	24.99135	44.67839	51.43969	43.45205
34.26395	40.97400	25.62196	45.65944	52.36950	41.94790
34.93968	41.43387	26.14539	46.71397	53.20816	44.82241
35.80462	41.83910	26.89443	48.47588	54.51036	44.84783
35.81991	41.84484	26.90864	48.51126	54.53618	45.88579
35.91164	41.87844	26.99470	48.72664	54.69343	44.42174
37.37737	42.27277	28.60698	52.99699	57.89239	47.84194

Table 2 The table shows the upper and lower bounds for the confidence interval (CI) and prediction interval (PI) of the data. The x_0 indicates the point from which the calculations are based upon. This value is identical to the x variable from the dataset. The y variable is also included as a reference.

The process will now be somewhat repeated using inverse regression analysis. The model for inverse regression is as follows,

$$x = -\frac{\alpha}{\beta} + \frac{1}{\beta} y - \varepsilon.$$

The term $\frac{\alpha}{\beta}$ is estimated using $\frac{\hat{\beta}_0}{\hat{\beta}_1}$. The term $\frac{1}{\beta}$ is estimated using $\frac{1}{\hat{\beta}_1}$. The following shows these estimated coefficients from inverse regression:

$$\frac{\hat{\alpha}}{\hat{\beta}} = \frac{\hat{\beta}_0}{\hat{\beta}_1} \approx \frac{6.2517}{1.4466} \approx 4.3216$$

$$\frac{1}{\beta} = \frac{1}{\hat{\beta}_1} \approx \frac{1}{1.4466} \approx 0.6913$$

The critical value of $t_{0.995,2}$ is approximately 9.9248. Since $|t_0| > t_{1-\frac{\alpha}{2},n-2}$, the decision is to reject the null hypothesis at the $\alpha = 0.01$ confidence level. The conclusion then is that there's a 99% probability that $\beta_1 \neq 0$ and that there's a significant relationship between the y and x variables.

The model can be seen below in Figure 2.

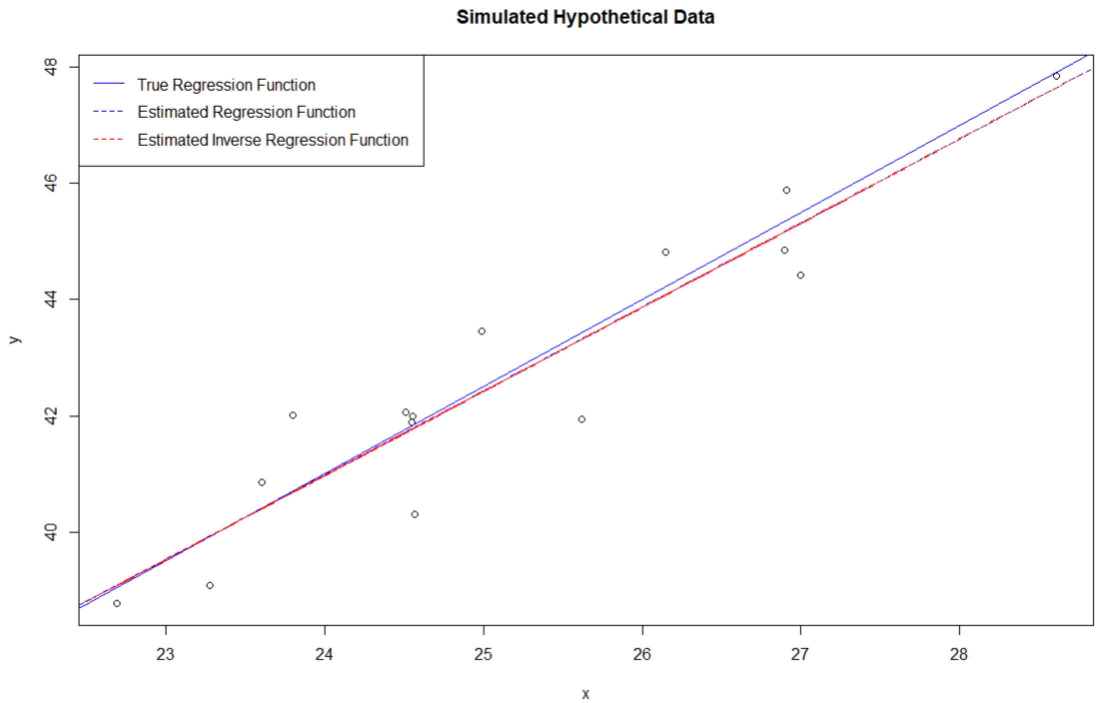

Figure 2 The figure shows the 15 simulated data points. The true regression function that the simulated points are based off is seen as a solid blue line. The estimated regression and estimated inverse regression functions are plotted also as dotted blue and red lines, respectively. They're directly overlapping and so it's difficult to discern them from each other.

2. Construct unbiased estimators for the new intercept and new slope in the regression of x on y.

The formula for the model is given as follows:

$$x_i = -\frac{\alpha}{\beta} + \frac{1}{\beta} y_i - \epsilon_i$$

It will be re-expressed to follow more similarly the traditional simple linear regression formula:

$$-x_i = \frac{\alpha}{\beta} + \frac{1}{\beta}(-y_i) + \epsilon_i$$
$$\tilde{x}_i = \tilde{\alpha} + \tilde{\beta}\tilde{y}_i + \epsilon_i$$

where $\tilde{x}_i = -x_i$ and $\tilde{y}_i = -y_i$

Some important assumptions are that ϵ has mean 0 and constant variance, σ^2. It also follows that the error terms are uncorrelated.

In the case where $\tilde{y}_i$ is nonrandom:

$$\tilde{\beta} = \frac{\sum_i (\tilde{x}_i - \bar{\tilde{x}})(\tilde{y}_i - \bar{\tilde{y}})}{\sum_i (\tilde{y}_i - \bar{\tilde{y}})^2}$$

where

$$\tilde{x}_i - \bar{\tilde{x}} = -x_i - (-\bar{x}) = \bar{x} - x_i$$
$$\tilde{y}_i - \bar{\tilde{y}} = -y_i - (-\bar{y}) = \bar{y} - y_i$$

So,

$$\tilde{\beta} = \frac{\sum_i (\tilde{x}_i - \bar{\tilde{x}})(\tilde{y}_i - \bar{\tilde{y}})}{\sum_i (\tilde{y}_i - \bar{\tilde{y}})^2} = \frac{\sum_i (\bar{x} - x_i)(\bar{y} - y_i)}{\sum_i (\bar{y} - y_i)^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (\bar{y} - y_i)^2}$$

Then it follows that for $\frac{1}{\beta}$ that the unbiased estimator is:

$$\widehat{\left(\frac{1}{\beta}\right)} = \hat{\hat{\beta}} \boxed{= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2}}$$

Lastly, for $\frac{\alpha}{\beta}$, the unbiased estimator is:

$$\frac{\hat{\alpha}}{\hat{\beta}} = \tilde{\alpha} = \bar{\bar{y}} - \bar{\bar{x}}\hat{\hat{\beta}}$$

$$= (-\bar{y}) - (-\bar{x})\hat{\hat{\beta}} = \bar{x}\hat{\hat{\beta}} - \bar{y} \boxed{= \bar{x} \times \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2} - \bar{y}}$$

Code:

```
set.seed(0) # Set seed
random_rows <- sample(1:21, 15) # Select rows
sort(random_rows)
advertisements <- data.frame( # input data
  x = c(50.1, 74.1, 19.3, 22.9, 82.4,
                   185.9, 20.4, 27, 154.9, 5,
                   26.9, 5.7, 7.6, 9.2, 6.1),
  y = c(32.1, 99.6, 11.7, 21.9, 60.8,
                           92.4, 21.4, 40.8, 88.9, 12,
                           38, 10, 12.3, 23.4, 4.4)
)

# part a
# Coefficients
x_bar <- mean(advertisements$x)
y_bar <- mean(advertisements$y)
S_xy <- sum((advertisements$x - x_bar) * (advertisements$y - y_bar))
S_xx <- sum((advertisements$x - x_bar)^2)
beta_hat_1 <- S_xy / S_xx
beta_hat_0 <- y_bar - beta_hat_1 * x_bar

# part b
# Testing significance
SS_T <- sum((advertisements$y - y_bar)^2) # Sum of Squares Total
SS_Res <- SS_T - beta_hat_1 * S_xy # Sum of Squares Residual
n <- nrow(advertisements) # Number of observations
MS_Res <- SS_Res / (n - 2) # Mean Squared Residual
se_beta_hat_1 <- sqrt(MS_Res / S_xx) # Standard error of beta-hat-1
t_0 <- beta_hat_1 / se_beta_hat_1 # Test statistic

alpha <- 0.05
qt(p = c(alpha / 2, 1 - alpha / 2), df = 2)
alpha <- 0.01
qt(p = c(alpha / 2, 1 - alpha / 2), df = 2)

# part c
# confidence band, prediction band
### The band_calculator function calculates the confidence interval
### and prediction interval of the mean of y for a given point x_0
### at a given alpha level.
band_calculator <- function(x_0, alpha = 0.05) {
  mu_hat <- beta_hat_0 + beta_hat_1 * x_0
  y_hat_0 <- mu_hat

  denominator_ci <- sqrt(MS_Res * ((1 / n) + ((x_0 - x_bar)^2) / S_xx))
  denominator_pi <- sqrt(MS_Res * ((1 + 1 / n) + ((x_0 - x_bar)^2) / S_xx))

  critical_value <- qt(p = c(alpha / 2, 1 - alpha / 2), df = 2)

  ci <- c(mu_hat + critical_value[1] * denominator_ci,
    mu_hat + critical_value[2] * denominator_ci)

  pi <- c(y_hat_0 + critical_value[1] * denominator_pi,
```

```
          y_hat_0 + critical_value[2] * denominator_pi)
  return(c(ci, pi))
}
band_vec <- Vectorize(band_calculator, vectorize.args = c("x_0"))
ci_pi <- t(band_vec(x_0 = advertisements$x, alpha = 0.05))
ci_table <- data.frame(lower_bound_pi = ci_pi[,3],
                       lower_bound_ci = ci_pi[,1],
                       x_0 = advertisements$x,
                       upper_bound_ci = ci_pi[,2],
                       upper_bound_pi = ci_pi[,4],
                       y = advertisements$y)
ci_table <- ci_table[order(ci_table$x_0),]
rownames(ci_table) <- NULL
colnames(ci_table) <- c("PI Lower Bound", "CI Lower Bound", "x0",
                        "CI Upper Bound", "PI Upper Bound", "y")
knitr::kable(x = ci_table, "markdown")

# plot confidence and prediction bands
plot(advertisements$x, advertisements$y, pch = 20,
     xlim = c(-10, 250), ylim = c(-20, 250),
     main = "Retained Impressions vs. Amount Spent",
     xlab = "Amount Spent (millions)",
     ylab = "Returned Impressions per week (millions)")
abline(a = beta_hat_0, b = beta_hat_1, col = 'blue')
x_seq <- seq(-20, 250, length.out = 1e4)

# Calculate bands
ci_pi_lines <- band_vec(x_0 = x_seq, alpha = 0.05)

# Plot lines
lines(x_seq, ci_pi_lines[1,], lty = 2, col = 'blue')
lines(x_seq, ci_pi_lines[2,], lty = 2, col = 'blue')
lines(x_seq, ci_pi_lines[3,], lty = 2, col = 'red')
lines(x_seq, ci_pi_lines[4,], lty = 2, col = 'red')
legend("topleft",
       legend = c("Estimated Regression Function",
                  "95% Confidence Interval", "95% Prediction Interval"),
       lty = c(1, 2, 2), col = c("blue", "blue", "red"))

# part d
MCI_x <- 26.9
band_calculator(x_0 = MCI_x, alpha = 0.05)

### Supplemental exercises
# create a hypothetical dataset
hypothetical_intercept <- 5
hypothetical_slope <- 1.5
set.seed(0)
hypothetical_data <- data.frame(
  x = rnorm(n = 15, mean = 25, sd = 1.5),
  epsilon = rnorm(n = 15, mean = 0, sd = 1.2)
)
hypothetical_data$y <- hypothetical_intercept +
  hypothetical_slope * hypothetical_data$x +
  hypothetical_data$epsilon
```

```r
# Coefficients
x_bar <- mean(hypothetical_data$x)
y_bar <- mean(hypothetical_data$y)
S_xy <- sum((hypothetical_data$x - x_bar) * (hypothetical_data$y - y_bar))
S_xx <- sum((hypothetical_data$x - x_bar)^2)
beta_hat_1 <- S_xy / S_xx
beta_hat_0 <- y_bar - beta_hat_1 * x_bar

# Testing significance
SS_T <- sum((hypothetical_data$y - y_bar)^2) # Sum of Squares Total
SS_Res <- SS_T - beta_hat_1 * S_xy # Sum of Squares Residual
n <- nrow(hypothetical_data) # Number of observations
MS_Res <- SS_Res / (n - 2) # Mean Squared Residual
se_beta_hat_1 <- sqrt(MS_Res / S_xx) # Standard error of beta-hat-1
t_0 <- beta_hat_1 / se_beta_hat_1 # Test statistic

alpha <- 0.1
qt(p = c(alpha / 2, 1 - alpha / 2), df = 2)
alpha <- 0.05
qt(p = c(alpha / 2, 1 - alpha / 2), df = 2)
alpha <- 0.01
qt(p = c(alpha / 2, 1 - alpha / 2), df = 2)

# confidence and prediction bands
ci_pi <- t(band_vec(x_0 = hypothetical_data$x, alpha = 0.01))
ci_table <- data.frame(lower_bound_pi = ci_pi[,3],
                       lower_bound_ci = ci_pi[,1],
                       x_0 = hypothetical_data$x,
                       upper_bound_ci = ci_pi[,2],
                       upper_bound_pi = ci_pi[,4],
                       y = hypothetical_data$y)
ci_table <- ci_table[order(ci_table$x_0),]
rownames(ci_table) <- NULL
colnames(ci_table) <- c("PI Lower Bound", "CI Lower Bound", "x0",
                        "CI Upper Bound", "PI Upper Bound", "y")
knitr::kable(x = ci_table, "markdown")

inverse_coeff_1 <- beta_hat_0 / beta_hat_1
inverse_coeff_2 <- 1 / beta_hat_1
est_x <- - inverse_coeff_1 + inverse_coeff_2 * hypothetical_data$y

# plot data and lines
plot(hypothetical_data$x, hypothetical_data$y,
     main = 'Simulated Hypothetical Data',
     xlab = 'x', ylab = 'y')

abline(a = hypothetical_intercept, b = hypothetical_slope, col = 'blue', lty = 1)
abline(a = beta_hat_0, b = beta_hat_1, col = 'blue', lty = 2)
legend("topleft", legend = c("True Regression Function",
                             "Estimated Regression Function",
                             "Estimated Inverse Regression Function"),
       col = c('blue', 'blue', 'red'), lty = c(1, 2, 2))
```

```
lines(x = est_x, y = hypothetical_data$y, col = 'red', lty = 2)
```