1 Introduction

Notation: The outcome y is the main variable of interest whose mean we want to explain in terms of other variables; aka the response, output, dependent variable. The covariate x represents the other variables of interest which potentially explain/predict the outcome in some sense; aka the predictor, input, independent variable, feature.

Quantitative description: To describe data, we can use univariate summaries (eg mean or variance of outcome or covariate) or bivariate summaries (eg covariance or correlation).

1. Mean/expectation: We focus on continuous rv in this course,

$$E[Y] = \int yf(y)\, dy$$

Recall the linearity of expectation and for observations $y_1, \ldots, y_n$, the sample mean is $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and is an estimate of the population mean which cannot be observed in practise.

2. Variance:

$$\text{Var}(Y) = E[(Y - E[Y])^2] = E[Y^2] - E[Y]^2$$

Recall $\text{Var}(aY + b) = a^2\text{Var}(Y)$ and for independent rvs X, Y, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. For observations $y_1, \ldots, y_n$, the (unbiased) sample variance is $s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$. Intuitively, we divide by $n - 1$ rather than n since $\bar{y}$ is an estimate rather than an actual observation, we lost one degree of freedom.

3. Covariance:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Recall $\text{Cov}(X, X) = \text{Var}(X)$ and $\text{Cov}(aY + c, bX + d) = ab\text{Cov}(X, Y)$ and $\text{Cov}(U + V, X + Y) = \text{Cov}(U, X) + \text{Cov}(U, Y) + \text{Cov}(V, X) + \text{Cov}(V, Y)$. Using the last property,

$$\text{Var}(X + Y) = \text{Cov}(Y + X, Y + X) = \text{Var}(Y) + \text{Var}(X) + 2\text{Cov}(X, Y)$$

and recall that $\text{Cov}(X, Y) = 0$ for independent rvs X, Y. For observations $(y_1, x_1), \ldots, (y_n, x_n)$, the sample covariance is $\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})$

4. Correlation: the correlation coefficient is,

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

and sample correlation is,

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

where $S_{xy} = \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})$ (note the exclusion of the $\frac{1}{n-1}$) and analogously for S_{xx}, S_{xy}. Correlation gives the strength of a linear relationship but does not characterize it. For example, the correlation of rvs X, Y is equal to the correlation of rvs X, Y' where $Y' = 2Y$, even though the relationship between X, Y is not the same as between X, Y'.

Normal distribution: Recall $Z \sim N(\mu, \sigma^2)$ with parameters μ, σ^2 has pdf

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(z - \mu)^2}{2\sigma^2}\right]$$

and moments $E[Z] = \mu, \text{Var}(Z) = \sigma^2$. As well, for independent $Z_i \sim N(\mu_i, \sigma^2)$, $U = \sum_{i=1}^n (a_i Z_i + b_i)$ is normally distributed,

$$U \sim N\left(\sum_{i=1}^n (a_i\mu_i + b_i), \sum_{i=1}^n a_i^2\sigma_i^2\right)$$

Chi-square distribution: Recall $X \sim \chi^2_\nu$ with ν degrees of freedom has pdf with support $X > 0$,

$$f(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}$$

and moments $E[X] = \nu, \text{Var}(X) = 2\nu$ and for $Z_i \sim N(0,1)$ iid,

$$X = \sum_{i=1}^n Z_i^2 \sim \chi^2_n$$

t-distribution: Recall $Y \sim t_\nu$ with ν degrees of freedom has pdf

$$f(y) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\frac{\nu}{2}} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

and moments $E[Y] = 0$ if $\nu > 1$ and N/A otherwise, and $\text{Var}(Y) = \frac{\nu}{\nu-2}$ if $\nu > 2$ and ∞ otherwise. As well, for independent $Z \sim N(0,1)$ and $X \sim \chi^2_\nu$,

$$\frac{Z}{\sqrt{X/\nu}} \sim t_\nu$$

2 Simple Linear Regression

We want to: characterize the relationship between x, y, predict y given x, evaluate how the mean of y changes when x increases by a. We can do this (in the case of 1 covariate) using a simple linear regression model,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i is the error term for the ith observation. And for cases of multiple covariates, we use a multiple linear regression,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

This course focuses on extending simple linear regression to multiple linear regression: mathematically (derive estimators of unknown parameters $\beta_0, \beta_1, \ldots$), practically (how to fit these models in R), how to choose and compare a model (which x_{ij} to include), and how to evaluate the appropriateness of the model/assumptions (model diagnostics).

For the ith observation, y_i is the outcome, x_i is a covariate, and i indexes the observations in the sample.

Simple Linear Regression Model: Suppose there is 1 covariate. Consider,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i is iid $N(0, \sigma^2)$. Alternatively, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and are independent of each other. Note that often Y_i indicates a rv and y_i indicates an observation but in this course, y_i is the rv. In this model, $\beta_0, \beta_1, \sigma^2$ are fixed, unknown parameters, ϵ_i is an unobserved random variable error term (denoted R_i in Stat 332), and y_i, x_i are the observed data. We treat x_i as fixed but y_i is not, since it is the sum of a fixed piece and a random error component.

Since x_i is fixed, we can consider observations y_i as conditioning on x_i. So,

$$E[y_i|x_i] = \beta_0 + \beta_1 x_i$$

by linearity since the mean of the error term is 0. β_0 can be interpreted as the intercept; $E[y_i|x_i = 0] = \beta_0$. As well, β_1 can be interpreted as a slope; $E[y_i|x_i = x^*] = \beta_0 + \beta_1 x^*$ and $E[y_i|x_i = x^* + 1] = \beta_0 + \beta_1 x^* + \beta_1$, so $\beta_1 = (E[y_i|x_i = x^*] - E[y_i|x_i = x^* + 1])/1$. In other words, the mean difference comparing a population with x to a population with x a unit lower.

Graphically, simple linear regression is a line in 2D space where ϵ_i is the difference between the line at x_i and the actual point y_i; there is variability around the line.

Four assumptions about this model: linearity ($E[y_i|x_i] = \beta_0 + \beta_1 x_i$, or $E[\epsilon_i] = 0$), independence (error terms are iid but note that covariates may or may not be independent of each other), normality (error terms are normally distributed, thus y_i are normal), equal variance (aka homoskedasticity, all error terms have the same variance σ^2).

Estimation: Since β_0, β_1 cannot be observed, we are interested in estimating them (aka finding the line of best fit). We want to find the estimators $\hat{\beta}_0, \hat{\beta}_1$. Note that in Stat 332, estimators are denoted by a tilde, but in this course, it is a hat. Estimators are random variables (rvs) since it is a function of the outcomes, which are themselves rvs. We consider two possible objective functions,

1. Least Squares: We minimize the sum of squares between the y_i and the line at x_i,

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

To do this, we compute the partial derivatives and set to 0,

$$0 = \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) = \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i$$

Re-arranging, we get,

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

With respect to β_1,

$$0 = \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = \sum_{i=1}^{n} y_i x_i - \beta_0 n\bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2$$

Substituting in $\beta_0 = \bar{y} - \beta_1 \bar{x}$ and re-arranging,

$$\beta_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum y_i(x_i - \bar{x})}{\sum x_i(x_i - \bar{x})} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

since $\sum \bar{y}(x_i - \bar{x}) = \bar{y} \sum (x_i - \bar{x}) = \bar{y}(\sum x_i - n\bar{x}) = 0$ and $\sum (\bar{x}^2 - x_i \bar{x}) = 0$. Thus, the LS estimators are,

$$\boxed{\hat{\beta}_1^{LS} = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0^{LS} = \bar{y} - \hat{\beta}_1^{LS} \bar{x}}$$

2. Maximum Likelihood Estimation: Recall y_i is independent $N(\beta_0 + \beta_1 x_i, \sigma^2)$. The likelihood (aka probability of unknown parameters given the observed y) is,

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2 | y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - [\beta_0 + \beta_1 x_i])^2}{2\sigma^2} \right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^{n} \frac{(y_i - [\beta_0 + \beta_1 x_i])^2}{2\sigma^2} \right)$$

The log likelihood is easier to work with,

$$\ell(\beta_0, \beta_1, \sigma^2 | y) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i])^2$$

To solve for $\max[\ell(\beta_0, \beta_1, \sigma^2 | y)] = \max[\mathcal{L}(\beta_0, \beta_1, \sigma^2 | y)]$, we solve a system of three equations,

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta_0} &= \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i]) \right) = 0 \\
\frac{\partial \ell}{\partial \beta_1} &= \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i]) \right) = 0 \\
\frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i])^2 = 0
\end{aligned}$$

Solving the first two equations is equivalent to minimizing the sum of squares. That is, under the assumption of normality,

$$\hat{\beta}_0^{LS} = \hat{\beta}_0^{ML} \qquad \hat{\beta}_1^{LS} = \hat{\beta}_1^{ML}$$

From here on out, we call the LS/ML estimators $\hat{\beta}_0, \hat{\beta}_1$.

We define the fitted values,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

and the residuals,

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Note that the residuals e_i (which are rvs since y_i are rvs) are different from the random errors $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$, which are based on the true, unobservable β_0, β_1.

Back to Maximum Likelihood Estimation, we can rearrange the third equation to solve for $\hat{\sigma}^2_{ML}$,

$$\hat{\sigma}^2_{ML} = \frac{\sum_{i=1}^n (y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2}{n} = \frac{\sum_{i=1}^n e_i^2}{n}$$

However, in practise, we typically use a different estimator,

$$\boxed{\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}}$$

which is unbiased,

$$E[\hat{\sigma}^2] = \sigma^2$$

Intuitively, this is because,

$$\frac{1}{\sigma^2} \sum_{i=1}^n e_i^2 \sim \chi^2_{(n-2)} \qquad E[\chi^2_\nu] = \nu$$

but often the distinction between $\hat{\sigma}^2$ and $\hat{\sigma}^2_{ML}$ does not matter when $n \geq 50$. In this course, we will typically use $\hat{\sigma}^2$. Informally, the $n-2$ comes from 2 fewer degrees of freedom since we are using $\hat{\beta}_0, \hat{\beta}_1$ which are functions of the y_i. For that reason, e_i are not independent of each other. See R code for how to fit a linear model.

2.1 Inference

In two different samples, we will get different estimates of β_0, β_1. We can characterize this uncertainty and variability in our estimates using standard errors (SE), confidence intervals (CI), and hypothesis tests (HT).

Properties of $\hat{\beta}_1$: Recall y_i are independent $N(\beta_0 + \beta_1 x_i, \sigma^2)$ and $\hat{\beta}_1 = \frac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \sum_{i=1}^n w_i y_i$ where $w_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$ are fixed wrt y. Hence, $\hat{\beta}_1$ is a linear combination of independent Normals,

$$\hat{\beta}_1 \sim N\left(\sum_{i=1}^n w_i (\beta_0 + \beta_1 x_i), \sigma^2 \sum_{i=1}^n w_i^2 \right)$$

So, the mean is,

$$\begin{aligned}
E[\hat{\beta}_1] &= \sum_{i=1}^n w_i (\beta_0 + \beta_1 x_i) \\
&= \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} (\beta_0 + \beta_1 x_i) \\
&= \beta_0 \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} + \beta_1 \frac{\sum_{i=1}^n x_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= 0 + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \beta_1
\end{aligned}$$

Hence, this estimator is unbiased. The variance is,

$$\begin{aligned}
\text{Var}(\hat{\beta}_1) &= \sigma^2 \sum_{i=1}^{n} w_i^2 \\
&= \sigma^2 \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \right]^2 \\
&= \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{[\sum_{j=1}^{n}(x_j - \bar{x})^2]^2} \\
&= \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{[\sum_{j=1}^{n}(x_j - \bar{x})^2]^2} \\
&= \sigma^2 \frac{1}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

Thus,

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \implies \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$$

and it can be shown using similar steps that,

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right)$$

Confidence Intervals: Using the known distribution values for $Z \sim N(0,1)$,

$$0.95 = P(-1.96 \leq Z \leq 1.96) = P\left(-1.96\frac{\sigma}{\sqrt{S_{xx}}} \leq \hat{\beta}_1 - \beta_1 \leq 1.96\frac{\sigma}{\sqrt{S_{xx}}}\right)$$

Re-arranging,

$$0.95 = P\left(\hat{\beta}_1 - 1.96\frac{\sigma}{\sqrt{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + 1.96\frac{\sigma}{\sqrt{S_{xx}}}\right)$$

So, a 95% CI for β_1 (assuming σ is known) is,

$$\boxed{\hat{\beta}_1 \pm 1.96\frac{\sigma}{\sqrt{S_{xx}}}}$$

In other words, this random interval (random because it is centered around the rv $\hat{\beta}_1$) will cover the true β_1 95% of the time. However, in practise, σ is unknown and must be estimated. We have,

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1) \qquad V = \frac{1}{\sigma^2}\sum_{i=1}^{n} e_i^2 \sim \chi^2_{(n-2)}$$

and it can be shown that Z, V are independent. Recall that for independent $Z \sim N(0,1)$ and $V \sim \chi^2_\nu$, we have that $\frac{Z}{\sqrt{V/\nu}} \sim t_\nu$. Thus,

$$\frac{Z}{\sqrt{V/(n-2)}} \sim t_{(n-2)} \implies \frac{\frac{\hat{\beta}_1 - \beta}{1/\sqrt{S_{xx}}}}{\sqrt{\sum_{i=1}^{n} e_i^2/(n-2)}} \sim t_{(n-2)} \implies \frac{\hat{\beta}_1 - \beta}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{(n-2)}$$

So, using similar steps as before, a 95% CI for β_1 when σ is unknown (and we use $\hat{\sigma}$) is,

$$\boxed{\hat{\beta}_1 \pm q\frac{\hat{\sigma}}{\sqrt{S_{xx}}} \qquad ; q \sim t_{(n-2)}}$$

Using R code (and see Stat 332 work), we compute q using `qt(p=0.05/2, df=n-2, lower.tail=FALSE)`. q is often denoted $t_{1-\alpha/2,n-2}$ for a $1-\alpha$ CI,

$$\hat{\beta}_1 \pm t_{1-\alpha/2,n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

where $1-\alpha/2$ is used to calculate the p value for HT and $n-2$ is the df for the t.

We can compute CI in R using `confint(Model)`, for some `Model <- lm(y ~ x)`. Note that CI does not represent a probability such as $P(0.235 \leq \beta_1 \leq 0.287) = 0.95$, since β_1 is an (unobservable) constant thus either does or does not fit in the interval. Rather, a CI says that if we draw samples repeatedly and construct intervals in the same way, on average, 95% of such intervals will contain the true β_1.

Standard Error: The standard error of $\hat{\beta}_1$ is the estimated standard deviation of $\hat{\beta}_1$,

$$SD(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{xx}}} \qquad SE(\hat{\beta}_1) = \sqrt{\hat{\text{Var}}(\hat{\beta}_1)} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

so we can write a CI for β_1 unknown σ as $\hat{\beta}_1 \pm t_{1-\alpha/2,n-2}SE(\hat{\beta}_1)$. This is the same as in Stat 332.

We can easily find CI for β_0 by using $\hat{\text{Var}}(\hat{\beta}_0)$ shown earlier instead of $\hat{\text{Var}}(\hat{\beta}_1)$ to get $SE(\hat{\beta}_0)$.

Hypothesis Testing: We want to test a null hypothesis $H_0 : \beta_1 = \theta_0$ against an alternative $H_1 : \beta_1 \neq \theta_0$. Often, in linear regression, $\theta_0 = 0$. The goal is to characterize how much evidence we have against H_0 (how extreme is our data relative to H_0). Under H_0 (assuming it is true),

$$\frac{\hat{\beta}_1 - \theta_0}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{(n-2)}$$

So the probability *under the null* of a test statistic (coming from a discrepancy, as described in Stat 332) as extreme (or more) than what we observe is in a two-sided test is,

$$p = P(|T| \geq |t|) = 2P(T \geq |t|) = 2[1 - P(T \leq -|t|)]$$

The observed test statistic t is,

$$t = \frac{\text{estimate} - H_0\text{value}}{\text{standard error}} = \frac{\hat{\beta}_1 - \theta_0}{\sqrt{\text{Var}(\hat{\beta}_1)}}$$

We reject the null hypothesis at the 5% level (ie $p < 0.05$) and if $p > 0.05$), we cannot accept the null, rather we say we do not have enough evidence to reject. We cannot accept the null because all these calculations were under the assumption that H_0 is true.

As well, it is not true that $P(H_0) = p$. Rather, the p value means: under the null hypothesis, the probability of a test statistic as extreme as the one observed is p. That's why a small p is evidence against the null, since it would be very rare to observe this data under the null.

If we reject the null whenever $p < 0.05$, then if the null is true and we repeat the experiment over and over, we only reject the null incorrectly 5% of the time. This is called the Type-I error rate (incorrectly rejecting a true null). Fun fact, the word "null" means that it is a commonly accepted fact that researchers work to nullify.

Note that the t value and p values in `summary()` in R have $\theta_0 = 0$ in the null hypothesis and the alternative is two sided. So for any other θ_0, you cannot use `summary()`.

2.2 Prediction

Estimating mean response: Recall the mean response is $E[Y_i] = \beta_0 + \beta_1 X_i$, since $E[\epsilon_i] = 0$. For an arbitrary x_0 (may or may not have been observed), the mean is $\mu_0 = E[Y_0] = \beta_0 + \beta_1 x_0$. We can estimate this as,

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_0 = \bar{y} + \hat{\beta}_1 (x_0 - \bar{x})$$

So, by linearity since x_0 is fixed, and since $\hat{\beta}_0, \hat{\beta}_1$ are unbiased estimators,

$$E[\hat{\mu}_0] = E[\hat{\beta}_0 + \hat{\beta}_1 x_0] = E[\hat{\beta}_0] + E[\hat{\beta}_1] x_0 = \beta_0 + \beta_1 x_0$$

Thus, the estimator of the mean response is unbiased. For variance,

$$\begin{aligned}
\text{Var}(\hat{\mu}_0) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\
&= \text{Var}(\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})) \qquad \text{; derived above} \\
&= \text{Var}\left(\left(\sum_{i=1}^{n} \frac{1}{n} y_i \right) + \left(\sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} y_i \right) (x_0 - \bar{x}) \right) \\
&= \text{Var}\left(\sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}} \right) y_i \right) \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}} \right)^2 \sigma^2 \qquad \text{; by independence of } y_i \text{ and } \text{Var}(y_i) = \sigma^2 \\
&= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} + \frac{(x_i - \bar{x})^2 (x_0 - \bar{x})^2}{S_{xx}^2} + 2 \frac{1}{n} \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}} \right) \\
&= \sigma^2 \left(\sum_{i=1}^{n} \frac{1}{n^2} + \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2 (x_0 - \bar{x})^2}{S_{xx}^2} + 2 \sum_{i=1}^{n} \frac{1}{n} \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}} \right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{2(x_0 - \bar{x})}{n S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}^2} S_{xx} + 0 \right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)
\end{aligned}$$

Since $\hat{\mu}_0$ is a linear combination of normal variables, it is itself normal. Using the mean and variance above,

$$\hat{\mu}_0 \sim N\left(\mu_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \right)$$

This implies (for known σ or in the unknown case, $\hat{\sigma}$),

$$\frac{\hat{\mu}_0 - \mu_0}{\sigma \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}} \sim N(0,1) \qquad \frac{\hat{\mu}_0 - \mu_0}{\hat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}} \sim t_{(n-2)}$$

In other words, a $100(1-\alpha)\%$ CI for unknown σ is,

$$\boxed{\hat{\mu}_0 \pm t_{1-\alpha/2, n-2} \hat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}}$$

Note that CIs are wider on the edges (as x_0 gets further away from $\bar{x}$), which makes sense since there is less data at the edges. As well, many points usually fall outside of the mean response CI; what id we do not care about the mean but want to do predictions?

Prediction: Instead of the mean response, we want the predict the response itself for a new covariate value,

$$y_{new} = \beta_0 + \beta_1 x_{new} + \epsilon_{new}$$

Define the predicted value $\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$ and the prediction error $\hat{y}_{new} - y_{new}$. Note that,

$$E[\hat{y}_{new} - y_{new}] = E[(\hat{\beta}_0 + \hat{\beta}_1 x_{new}) - (\beta_0 + \beta_1 x_{new} + \epsilon_{new})] = \beta_0 + \beta_1 x_{new} - (\beta_0 + \beta_1 x_{new}) = 0$$

Note that $\hat{y}_{new}$ and $-y_{new}$ are independent rvs (since $\hat{y}_{new}$ can be expressed as a linear combination of y_i from $i = 1$ to n but y_{new} is a completely new observation that is not included in these/did not form the estimates) and a linear combination of normals. The variance of the prediction error,

$$\begin{aligned}
\text{Var}(\hat{y}_{new} - y_{new}) &= \text{Var}((\hat{\beta}_0 + \hat{\beta}_1 x_{new}) - y_{new}) \\
&= \text{Var}((\hat{\beta}_0 + \hat{\beta}_1 x_{new})) + \text{Var}(y_{new}) \qquad \text{; independence} \\
&= \left[\sigma^2\left(\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right] + \text{Var}(y_{new}) \qquad \text{; same derivation as in mean response} \\
&= \sigma^2\left(\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) + \sigma^2 \qquad \text{; by assumption, since } \text{Var}(\epsilon_{new}) = \sigma^2 \\
&= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)
\end{aligned}$$

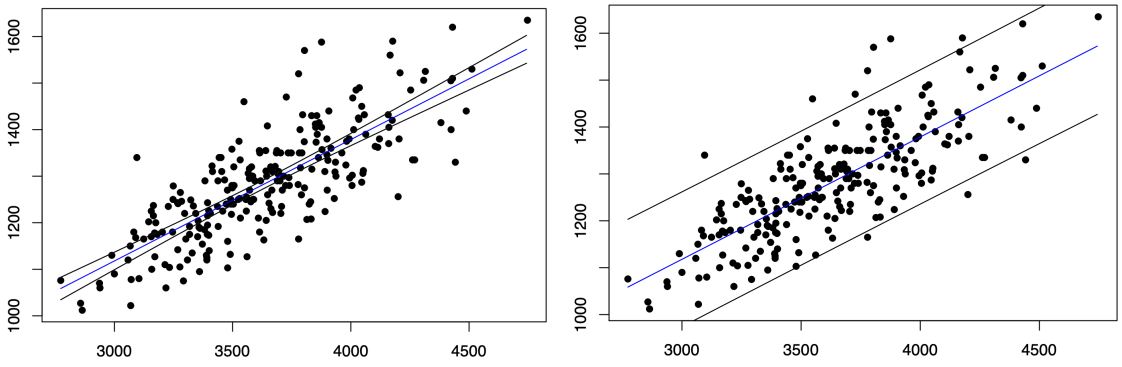
Thus, since the prediction error is normally distributed,

$$\frac{\hat{y}_{new} - y_{new}}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim N(0,1) \qquad \frac{\hat{y}_{new} - y_{new}}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}$$

Thus, a $100(1-\alpha)\%$ prediction interval (for unknown σ) is,

$$\boxed{\hat{y}_{new} \pm t_{1-\alpha/2, n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

CI vs PI: The prediction interval (right) is wider than the CI for the mean response (left) because of the extra 1 in the square root, which accounts for the additional variability around the mean.



Note that CI is usually for parameters like μ, β_0, β_1 and PI is for a future individual observation, which is a data point that has not yet been observed. For parameters like μ, β_0, β_1, they already exist so it is not a prediction.

3 Multiple Linear Regression

4 Model Building

5 Model Diagnostics

6 Extensions