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 technically Y_i is the rv and y_i is the outcome but here, we treat y_i as a rv sometimes in notation. 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. We call β_0, β_1 the regression coefficients.

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 higher (alternatively, with a 1 unit change in the covariate).

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).

2.1 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,

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

Re-arranging, we get,

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

With respect to β_1,

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

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 (asserted without proof),

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

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.

2.2 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 $t_{n-2,1-\alpha/2}$ 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 $\hat{\beta}_1 \pm t_{n-2,1-\alpha/2}SE(\hat{\beta}_1)$. 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 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()`.

2.3 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,

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

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: Suppose instead of the mean response, we want the predict the response itself (ie one specific value) 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,

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

Here, $\hat{y}_{new}$ and $-y_{new}$ are independent rvs ($\hat{y}_{new}$ 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 is,

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

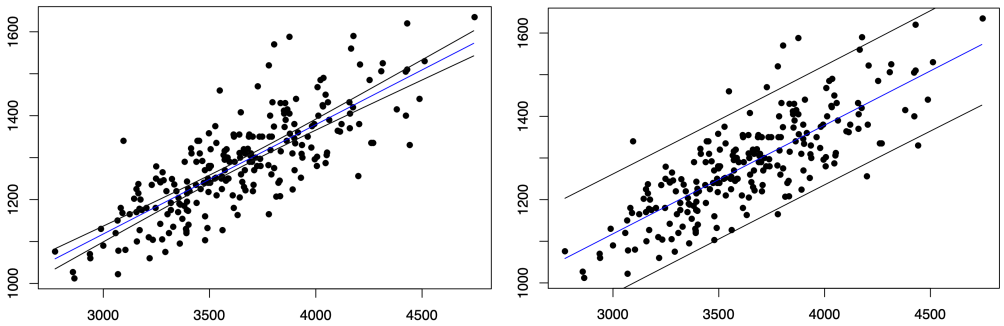
Thus, since the prediction error is normally distributed,

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

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

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

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 comes from variability of any one observation around the mean line (uncertainty in estimating the line itself is accounted for in the rest of the SE after the 1+).

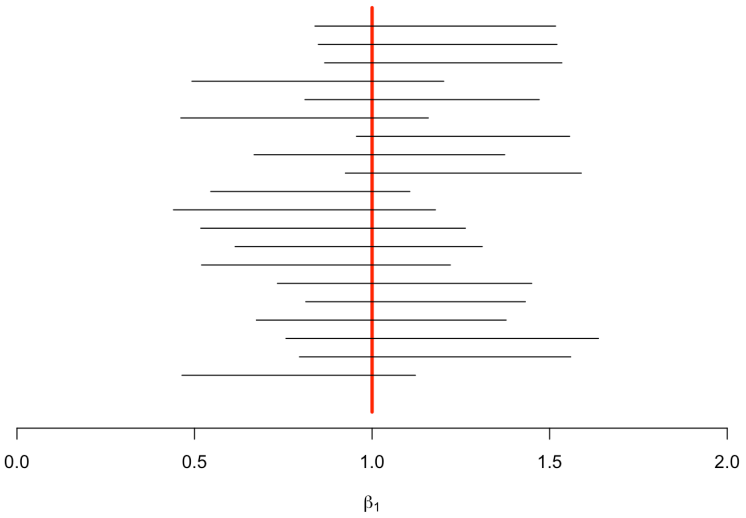


Note that CI is usually for parameters like μ, β_0, β_1 and PI is for a future individual observation, which is a data point that has not yet been observed. For parameters like μ, β_0, β_1, they already exist so it is not a prediction. We use CI for mean response to estimate the mean response for a population of observations at a certain covariate level and PI to predict one specific new response.

2.4 Summary

The important concepts from this section on simple linear regression,

1. The fitted values are $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ where $\hat{\beta}_0, \hat{\beta}_1$ are the LS (or ML) estimates of the parameters based on the observed data (x_i, y_i) for $i = 1, \ldots, n$
2. The LS estimates for β_0, β_1 are $\hat{\beta}_1 = S_{xy}/S_{xx}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and the ML estimates are the same. These estimators are unbiased. The derivation for unbiasedness only relies on the linearity of expectation and did not need the normality, homoskedasticity, or independence assumptions
3. The ML estimate for σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n e_i^2$ where $e_i = y_i - \hat{y}_i$ are the residuals
4. CI, PI, and HT for known ($\sim N(0,1)$) and unknown σ ($\sim t_{n-2}$).
5. $\hat{\beta}_1$ can be expressed as a linear combination of the outcomes, $\sum_{i=1}^n w_i y_i$ for some w_i, thus is also Normal
6. If we collect a new sample, we would get a different CI. And for $100(1-\alpha)\%$ CI, we have that $100(1-\alpha)\%$ of such intervals will cover the true underlying parameter
7. The p value is the probability of a test statistic as extreme or more than what we observe, *under the null* (ie assuming the null is true)
8. Estimating the mean response: $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} = \hat{\beta}_1(x_0 - \bar{x})$
9. Predicting a new response: $\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$
10. A visual interpretation of CI:



Here, the true β_1 is 1 (the red line) and the black lines are possible sampled $100(1-\alpha)$ CIs. We expect $1-\alpha$ proportion of such CIs to cover the true β_1 of 1 as more CIs are sampled.

11. Mean response and predicted response are not two types of response. A response is an outcome (ie y_i). "Mean response" actually refers to the mean of responses $\mu_0 = E[y_i | x_i = x_0] = \beta_0 + \beta_1 x_0$. This is estimated with the estimated mean response $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

    We can also consider a brand new response, y_{new}, which we are trying to predict. The predicted response is our prediction of y_{new}, called $\hat{y}_{new}$.

    μ_0 lies on the line $\beta_0 + \beta_1 x$, which is the true underlying regression line which we do not know. y_{new} varies around this line because of the error term, $\beta_0 + \beta_1 x_{new} + \epsilon_{new}$.

$\hat{\mu}_0$ lies on the fitted regression line $\hat{\beta}_0 + \hat{\beta}_1 x$, which is our best guess for the unknown regression line. As such, there is uncertainty in estimating this line. Even if we had no uncertainty in estimating the regression line, (ie we knew exactly $\beta_0 + \beta_1 x$, there would still be uncertainty in predicting y_{new} since it does not lie on the true regression line because of the error term. That is why there is the 1+ term in the standard error.

Interestingly, the predicted response $\hat{y}_{new}$ lies also on the fitted regression line $\hat{\beta}_0 + \hat{\beta}_1 x$. In other words, if we estimate the mean response at covariate level x^* and predict a new response at covariate level x^*, both will have the same value: $\hat{\mu}_\star = \hat{\beta}_0 + \hat{\beta}_1 x^*$ and $\hat{y}_\star = \hat{\beta}_0 + \hat{\beta}_x^*$. See Quiz 2 Q1f for how this fact can be used to represent a PI in terms of a given CI.

Why does this happen? If we know a new response is not going to fall on the regression line, why is the predicted response the same as the estimate of the mean response? This is because we have no special knowledge of the new observation other than its covariate. Some new observations will be above the line, some below, we do not know. But, their average falls on the regression line. So, the predicted value will still fall on the regression line, as it is an unbiased estimator. Nevertheless, the prediction interval will be much wider since there is variability around the line to account for, in addition to the variability in our estimate of the line.

12. y_{new} is a random variable, which is why it is called a prediction rather than estimation (which is for a fixed parameter). However, like our outcomes, we could observe a realization of it. For example, suppose we drew a random sample of 30 patients in a hospital, and fit our model. We could consider sampling one more patient, with outcome y_{new}, which we would assume is normally distributed from the same model as the one from which our data were drawn. We can make a prediction, construct a prediction interval. And then we could hypothetically measure the outcome on this 31st patient (a realization of this random variable), and see how close our prediction was, whether our prediction interval covered this value, etc.

    A 95% PI can be interpreted as: if we repeatedly drew samples and calculated the PI for a fixed covariate value of x_new, then 95% of the time, we would end up with the true value of y_{new} in our interval.

3 Multiple Linear Regression

3.1 Math Review

Denote vectors and matrices with bold, like in CS 480,

3.1.1 Matrix review

Recall from Math 136,

1. Transpose: $[\boldsymbol{C}^T]_{ij} = [\boldsymbol{C}]_{ji}$
2. $\boldsymbol{C}$ is symmetric if $\boldsymbol{C}^T = \boldsymbol{C}$
3. $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$
4. If square matrix $\boldsymbol{B}$ is nonsingular (aka invertible), then $\boldsymbol{B}\boldsymbol{B}^{-1} = \boldsymbol{B}^{-1}\boldsymbol{B} = \boldsymbol{I}$
5. $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$ if both $\boldsymbol{A}, \boldsymbol{B}$ are square and nonsingular
6. $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$
7. Trace (of a square matrix):
   (a) $\operatorname{tr}(\boldsymbol{A}) = \sum_{j=1}^{n} a_{jj}$, sum of the diagonals
   (b) $\operatorname{tr}(\boldsymbol{A} + \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B})$
   (c) $\operatorname{tr}(c\boldsymbol{A}) = c\operatorname{tr}(\boldsymbol{A})$
   (d) $\operatorname{tr}\big(\boldsymbol{A}^T\big) = \operatorname{tr}(\boldsymbol{A})$
   (e) $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$

3.1.2 Matrix calculus

1. Let $z = f(y_1, \ldots, y_k)$ and $\boldsymbol{y} = (y_1, \ldots, y_k)^T$, then,

$$\frac{\partial z}{\partial \boldsymbol{y}} = \begin{bmatrix} \frac{\partial z}{\partial y_1} \\ \frac{\partial z}{\partial y_2} \\ \vdots \\ \frac{\partial z}{\partial y_k} \end{bmatrix}$$

2. A special case of (1): If $z = \boldsymbol{a}^T\boldsymbol{y}$, where $\boldsymbol{a} = (a_1, \ldots, a_k)^T$ is a vector, then,

$$\frac{\partial z}{\partial y} = \boldsymbol{a}$$

3. Another special case of (1): If $z = \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}$ (quadratic form) where $\boldsymbol{A}$ is a $k \times k$ matrix, then,

$$\frac{\partial z}{\partial \boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{A}^T\boldsymbol{y}$$

   and if $\boldsymbol{A}$ is symmetric, then $\frac{\partial z}{\partial \boldsymbol{y}} = 2\boldsymbol{A}\boldsymbol{y}$

3.2 Random Vectors

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be a random vector (ie, each of $y_1, \ldots, y_n$ are random variables). The mean of $\boldsymbol{y}$ is element-wise,

$$E[\boldsymbol{y}] = \begin{bmatrix} E[y_1] \\ \vdots \\ E[y_n] \end{bmatrix}$$

and a random matrix can be similarly defined element-wise.

3.2.1 Covariance matrix

The variance of a vector $\boldsymbol{y}$ is,

$$\text{Var}(\boldsymbol{y}) = E[(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T]$$

where $\boldsymbol{\mu} = E[\boldsymbol{y}]$. In matrix form,

$$\boldsymbol{V} = \text{Var}(\boldsymbol{y}) = \begin{bmatrix} \text{Var}(y_1) & \text{Cov}(y_1, y_2) & \cdots & \text{Cov}(y_1, y_n) \\ \text{Cov}(y_2, y_1) & \text{Var}(y_2) & \cdots & \text{Cov}(y_1, y_n) \\ \vdots & \vdots & \vdots & \vdots \\ \text{Cov}(y_n, y_1) & \text{Cov}(y_n, y_2) & \cdots & \text{Var}(y_n) \end{bmatrix}$$

Note that $\boldsymbol{V}$ is symmetric and positive semi-definite, meaning that for all $n \times 1$ vectors $\boldsymbol{a}$ in $\mathbb{R}^n$, $\boldsymbol{a}^T\boldsymbol{V}\boldsymbol{a} \geq 0$. As well, the linear regression model does not assume that covariates are independent of each other, so $\boldsymbol{V}$ is not just a diagonal matrix.

The covariance matrix is also often called the variance matrix or the variance-covariance matrix.

3.2.2 Linear combinations

Recall that if $z = \sum_{i=1}^n a_i y_i + c$ and $u = \sum_{i=1}^n b_i y_i + d$ (note that z, y_i, u are rvs),

$$E[z] = \sum_{i=1}^n a_i E[y_i] + c \qquad \text{Cov}(z, u) = \sum_{i=1}^n \sum_{j=1}^n a_i b_j \text{Cov}(y_i, y_j)$$

In matrix form, $z = \boldsymbol{a}^T\boldsymbol{y} + c$ and $u = \boldsymbol{b}^T\boldsymbol{y} + d$ and,

$$E[z] = \boldsymbol{a}^T\boldsymbol{\mu} + c \qquad \text{Cov}(z, u) = \boldsymbol{a}^T\boldsymbol{V}\boldsymbol{b}$$

where $\boldsymbol{\mu} = E[\boldsymbol{y}]$ and $\boldsymbol{V} = \text{Var}(\boldsymbol{y})$.

For the $k \times 1$ random vector $\boldsymbol{z} = (z_1, \ldots, z_k)^T$ of k linear combinations of $n \times 1$ random vector $\boldsymbol{y} = (y_1, \ldots, y_n)^T$,

$$z_i = a_{i1}y_1 + a_{i2}y_2 + \cdots + a_{in}y_n$$

for $i = 1, \ldots, k$. We can equivalently write,

$$\boldsymbol{z} = \boldsymbol{A}\boldsymbol{y}$$

where $\boldsymbol{A}$ is a $k \times n$ matrix with elements a_{ij} and,

$$E[\boldsymbol{z}] = \boldsymbol{A}\boldsymbol{\mu} \qquad \text{Var}(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T$$

Note that if $\boldsymbol{z} = \boldsymbol{y}^T\boldsymbol{B}$, then $E[\boldsymbol{z}] = E[\boldsymbol{y}^T]\boldsymbol{B}$; a matrix of constants can be pulled out of both the left and right sides. Using this fact combined with $E[\boldsymbol{A}\boldsymbol{y}] = \boldsymbol{A}\boldsymbol{\mu}$, the $\text{Var}(\boldsymbol{z})$ expression comes from,

$$\begin{aligned}
\text{Var}(\boldsymbol{z}) &= \text{Var}(\boldsymbol{A}\boldsymbol{y}) \\
&= E[(\boldsymbol{A}\boldsymbol{y} - E[\boldsymbol{A}\boldsymbol{y}])(\boldsymbol{A}\boldsymbol{y} - E[\boldsymbol{A}\boldsymbol{y}])^T] \\
&= E[\boldsymbol{A}(\boldsymbol{y} - E[\boldsymbol{y}])(\boldsymbol{A}(\boldsymbol{y} - E[\boldsymbol{y}]))^T] \\
&= E[\boldsymbol{A}(\boldsymbol{y} - E[\boldsymbol{y}])(\boldsymbol{y} - E[\boldsymbol{y}])^T\boldsymbol{A}^T] \\
&= \boldsymbol{A}E[(\boldsymbol{y} - E[\boldsymbol{y}])(\boldsymbol{y} - E[\boldsymbol{y}])^T]\boldsymbol{A}^T \qquad \text{; pulling out constant matrices} \\
&= \boldsymbol{A}\text{Var}(\boldsymbol{y})\boldsymbol{A}^T \\
&= \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T
\end{aligned}$$

3.2.3 Summary of Useful Properties of Random Vectors

For fixed (constant) vectors $\boldsymbol{a}, \boldsymbol{b}$, fixed matrix $\boldsymbol{A}$, and random vector $\boldsymbol{y}$,

1. $E[\boldsymbol{a}] = \boldsymbol{a}$
2. $E[\boldsymbol{a}^T\boldsymbol{y} + \boldsymbol{b}] = \boldsymbol{a}^T E[\boldsymbol{y}] + \boldsymbol{b}$
3. $E[\boldsymbol{A}\boldsymbol{y}] = \boldsymbol{A}E[\boldsymbol{y}]$
4. $E[\boldsymbol{y}^T\boldsymbol{A}] = E[\boldsymbol{y}^T]\boldsymbol{A}$
5. $\text{Var}(\boldsymbol{y})$ is the covariance matrix
6. $\text{Var}(\boldsymbol{y}) = E[(\boldsymbol{y} - E[\boldsymbol{y}])(\boldsymbol{y} - E[\boldsymbol{y}])^T]$
7. $\text{Var}(\boldsymbol{a}^T\boldsymbol{y}) = \boldsymbol{a}^T\text{Var}(\boldsymbol{y})\boldsymbol{a}$
8. $\text{Var}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}\text{Var}(\boldsymbol{y})\boldsymbol{A}^T$

3.3 Multivariate Normal (MVN) Distribution

Let $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ be a random vector of iid standard normal variables, ie z_i is iid $N(0, 1)$. Then $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu}$ (a linear transformation of $\boldsymbol{z}$, a linear combination of $z_1, \ldots, z_n$) has multivariate normal distribution, ie,

$$\boldsymbol{y} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $E[\boldsymbol{y}] = \boldsymbol{\mu}$ and $\text{Var}(\boldsymbol{y}) = \boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$. Note that $\boldsymbol{y}, \boldsymbol{\mu}$ are $n \times 1$ and $\boldsymbol{\Sigma}$ is $n \times n$. The density function of $\boldsymbol{y}$ is,

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right\}$$

where $|\boldsymbol{\Sigma}|$ is notation for $\det(\boldsymbol{\Sigma})$. Note that if $n = 1$, this is just the Normal distribution.

Some properties of $\boldsymbol{y} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (sometimes MVN is just denoted N if it is clear from context that this is a multivariate case),

1. Linearity: if $\boldsymbol{u} = \boldsymbol{C}\boldsymbol{y} + \boldsymbol{d}$, then,
$$\boldsymbol{u} \sim MVN(\boldsymbol{C}\boldsymbol{\mu} + \boldsymbol{d}, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^T)$$
2. Marginal distribution: if $\tilde{\boldsymbol{y}} = (y_1, \ldots, y_m)^T \subset \boldsymbol{y}$ is a vector subset of $\boldsymbol{y}$ (ie $m \leq n$ and $y_1, \ldots, y_m$ are also elements of $\boldsymbol{y}$, although not necessarily in that order), then $\tilde{\boldsymbol{y}}$ is MVN-distributed with mean $(\mu_1, \ldots, \mu_m)^T$ where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m, \mu_{m+1}, \ldots, \mu_n)^T$ and variance is the sub-matrix of $\boldsymbol{\Sigma}$ containing the $1, \ldots, m$ related elements. Thus, it is easy to identify the MVN distribution of a subvector of $\boldsymbol{y}$, given the $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ of the full vector $\boldsymbol{y}$.

3. Conditional distribution: if $\boldsymbol{y} = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T)^T$ (ie vector $\boldsymbol{y}$ is the concatenation of two vectors $\boldsymbol{y}_1, \boldsymbol{y}_2$), then $\boldsymbol{y}_1^T | \boldsymbol{y}_2^T$ is MVN-distributed.
4. Independence: If $\boldsymbol{\Sigma}_{ij} = 0$, then $\boldsymbol{y}_i, \boldsymbol{y}_j$ are independent. Note that this is not generally true; $\text{Cov}(X, Y) = 0$ does not generally imply that rvs X, Y are independent (although the inverse does hold).

Exercise: Let rvs $X_1 \sim N(0, 1)$ and $B \sim \text{BERN}(0.5)$, independent of X_1. Now consider the rv $X_1 = X_1$ if $B = 0$ or $X_2 = -X_1$ if $B = 1$. X_2 is clearly Normally distributed, since we've just swapped half the signs to the the opposite and since it is symmetric, the distribution is unchanged. However, $Y = X_1 + X_2$ is not Normal. We know this because $P(Y = 0) = P(X_1 = -X_2) = P(B = 1) = 1/2$. Y is not Normal because while X_1, X_2 are Normal, $\boldsymbol{x} = (X_1, X_2)^T$ is not multivariate Normal; ie we cannot find a $2 \times k$ matrix $\boldsymbol{A}$ such that $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z}$ for $\boldsymbol{z} = (Z_1, \ldots, Z_k)^T$ where Z_i are iid $N(0, 1)$.

3.4 Multiple Linear Regression

Mutiple linear regression is useful over simple linear regression in situations where there is more than covariate. Suppose we have a dataset of multivariate data collected from patients at a hospital with satisfaction as the outcome and age, severity of condition, and stress as the covariates. Some questions we might ask, which could be answered by multiple linear regression,

1. Is mean satisfaction associated with stress, conditional on age and severity?
2. How does mean satisfaction differ for older patients vs younger patients with the same severity and stress scores?
3. Given a patient's stress, age, and severity, can we predict their satisfaction?
4. Is the (conditional) association between stress and satisfaction different for older patients than for younger patients?

In multiple linear regression with p covariates (assuming $p < n$, where n is the number of observations),

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where ϵ_i is iid $N(0, \sigma^2)$ for $i = 1, \ldots, n$, and,

$$y_i \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2)$$

where the y_i's are independent. As well, since x_{ij} is fixed, like in simple linear regression, we can consider observations y_i as conditioning on x_i (ie $y_i | x_{i1}, \ldots, x_{ip}$).

We can also represent this formulation using matrices,

$$\begin{bmatrix} y_i \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_P \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

which is denoted using the variables,

$$\boxed{\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}}$$

In this case,

$$\boxed{\boldsymbol{\epsilon} \sim MVN(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \iff \boldsymbol{y} \sim MVN(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})}$$

where $\boldsymbol{0}$ is a $N \times 1$ zero vector. Like in simple linear regression, $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$ are generally unknown and must be estimated. As well, $\boldsymbol{X}$ is often called the design matrix. Specifying the contents of $\boldsymbol{X}$ is all that is needed to define the model, since the other vectors are always present.

Interpreting regression coefficients: Note that,

$$E[y_i|x_{i1},\ldots,x_{ip}] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_i$$

although often, we simply write $E[y_i]$. As well,

$$E[y_i|x_{i1} = \cdots = x_{ip} = 0] = \beta_0$$

which says that β_0 is the mean outcome when all covariates are set to 0. However, this is sometimes not interpretable (eg, how can age be 0 in a sample of adults). And,

$$\begin{aligned} E[y_i|x_{i1} = x_1, x_{i2} = x_2, \ldots, x_{ip} = x_p] &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_P x_p \\ E[y_i|x_{i1} = (x_1+1), x_{i2} = x_2, \ldots, x_{ip} = x_p] &= \beta_0 + \beta_1 (x_1+1) + \beta_2 x_2 + \cdots + \beta_p x_p \end{aligned}$$

Subtracting the first expectation from the second results in just β_1. More generally, β_j is the difference in mean outcome for a one unit change in the jth covariate, holding all other covariates fixed.

3.5 Estimation

Like in the simple linear regression case, we consider Least Squares and Maximum Likelihood Estimation to estimate $\beta_0, \ldots, \beta_P, \sigma^2$,

1. Least Squares: We want to minimize the sum of squares,

$$\begin{aligned} S(\boldsymbol{\beta}) &= . \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

   Setting the partial derivative wrt $\boldsymbol{\beta}$ (using the matrix calculus properties above) to 0 and solving,

$$0 = \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \implies (\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}$$

   And if the columns of $\boldsymbol{X}$ are linearly independent (which is usually the case, if the number of covariates is small compared to the number of data points), then,

$$\boxed{\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}}$$

   This is a very important result in Statistics. See also the Linear Regression section of my CS 480 notes (although this formula is represented using sums rather than matrix multiplication)

2. Maximum Likelihood: Recall $\boldsymbol{y} \sim MVN(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. The likelihood function is,

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma^2\boldsymbol{I}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right\}$$

   and since $\det(\sigma^2\boldsymbol{I}) = (\sigma^2)^n$, the log likelihood is,

$$\ell(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

   And maximizing the log likelihood wrt $\boldsymbol{\beta}$ is equivalent to minimizing $\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ wrt $\boldsymbol{\beta}$, which is exactly the same objective as Least Squares. Thus, $\hat{\boldsymbol{\beta}}_{MLE}$ is the same as the $\hat{\boldsymbol{\beta}}$ derived in LS.

The mean of $\hat{\boldsymbol{\beta}}$, using the properties of random vectors derived earlier, is,

$$E[\hat{\boldsymbol{\beta}}] = E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E[\boldsymbol{y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{\beta}$$

Thus, $\hat{\boldsymbol{\beta}}$ is unbiased. For the variance, again using the properties of random vectors from earlier,

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}) &= \text{Var}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{Var}(\boldsymbol{y})((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)^T \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{Var}(\boldsymbol{y})\boldsymbol{X}((\boldsymbol{X}^T\boldsymbol{X})^{-1})^T \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{Var}(\boldsymbol{y})\boldsymbol{X}((\boldsymbol{X}^T\boldsymbol{X})^T)^{-1} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{Var}(\boldsymbol{y})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma^2\boldsymbol{I})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}
\end{aligned}$$

Moreover, since $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, then $\hat{\boldsymbol{\beta}}$ is MVN distributed, since it is just a linear transformation of $\boldsymbol{y}$, which is MVN distributed. Thus,

$$\boxed{\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})}$$

Moreover, by property 2 of MVN distributions (see above),

$$\boxed{\hat{\beta}_j \sim N(\beta_j, \sigma^2 V_{jj})}$$

where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Next, let the fitted values be $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. This is the multivariate version of $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Expanding,

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] = [\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is called the Hat matrix (or the projection matrix). It maps the outcomes to the vector of fitted values.

Some interesting facts about $\boldsymbol{H}$: it is symmetric ($\boldsymbol{H}^T = \boldsymbol{H}$) and idempotent ($\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$; see Lecture 7 slide 16 for proof). As well, $\boldsymbol{I} - \boldsymbol{H}$ is symmetric and idempotent (see Lecture 8 slide 16 for proof).

The mean of $\hat{\boldsymbol{y}}$ is,

$$E[\hat{\boldsymbol{y}}] = E[\boldsymbol{H}\boldsymbol{y}] = \boldsymbol{H}E[\boldsymbol{y}] = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$$

and the variance is,

$$\text{Var}(\hat{\boldsymbol{y}}) = \text{Var}(\boldsymbol{H}\boldsymbol{y}) = \boldsymbol{H}\text{Var}(\boldsymbol{y})\boldsymbol{H}^T = \boldsymbol{H}\sigma^2\boldsymbol{I}\boldsymbol{H} = \sigma^2\boldsymbol{H}$$

so the distribution of $\hat{\boldsymbol{y}}$ is,

$$\boxed{\hat{\boldsymbol{y}} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{H})}$$

Now consider a vector of residuals $\boldsymbol{e}$ (the multivariate version of $e_i = y_i - \hat{y}_i$),

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

this shows that the residuals can also be written as a linear combination of the outcomes. Note that $\boldsymbol{X}^T\boldsymbol{e} = 0$, since,

$$\boldsymbol{X}^T\boldsymbol{e} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}) = \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = 0$$

Deriving an estimate for σ using Maximum Likelihood, recall,

$$\ell(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$$

and setting the derivative with respect to σ^2 to 0,

$$\begin{aligned}
0 = \frac{\partial \ell(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\
\implies n\sigma^2 &= (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\
\implies \hat{\sigma}^2_{ML} &= \frac{1}{n}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2 = \frac{1}{n}\boldsymbol{e}^T\boldsymbol{e}
\end{aligned}$$

however, like in simple linear regression, we usually do not use the ML estimate and instead use the unbiased estimator for σ^2,

$$\boxed{\hat{\sigma}^2 = \frac{1}{n-(p+1)}\boldsymbol{e}^T\boldsymbol{e}}$$

We now take a closer look at this unbiased estimator. Recall $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and $\boldsymbol{e} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$. Consider the vector $\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \boldsymbol{e}\end{bmatrix}$, which is a linear combination of $\boldsymbol{y}$. Since $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, then this vector is MVN distributed. We want to find the expectation and variance of this vector. Recall $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. And,

$$\begin{aligned}
E[\boldsymbol{e}] &= E[(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}] \\
&= (\boldsymbol{I}-\boldsymbol{H})E[\boldsymbol{y}] \\
&= (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\
&= 0
\end{aligned}$$

Thus, $E\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\beta} \\ 0\end{bmatrix}$. And it can be shown (see Lecture 7 slides 15, 16) that the variance is $\sigma^2\begin{bmatrix}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{I}-\boldsymbol{H})\end{bmatrix}$. The distribution is then,

$$\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \boldsymbol{e}\end{bmatrix} \sim N\left(\begin{bmatrix}\boldsymbol{\beta} \\ 0\end{bmatrix}, \sigma^2\begin{bmatrix}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{I}-\boldsymbol{H})\end{bmatrix}\right)$$

and can also conclude using the theory about marginal MVN distributions that,

$$\boxed{\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2(\boldsymbol{I}-\boldsymbol{H}))}$$

and that $\hat{\boldsymbol{\beta}}$ (whose distribution is already described earlier) and $\boldsymbol{e}$ are independent, since the off-diagonal covariance terms in the block matrix variance above are zero matrices and in the case of MVN, this implies independence.

Note that an alternate way to determine $\text{Var}(\boldsymbol{e})$ using the fact that $\boldsymbol{I}-\boldsymbol{H}$ is idempotent is,

$$\text{Var}(\boldsymbol{e}) = \text{Var}((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}) = (\boldsymbol{I}-\boldsymbol{H})\text{Var}(\boldsymbol{y})(\boldsymbol{I}-\boldsymbol{H})^T = \sigma^2(\boldsymbol{I}-\boldsymbol{H})$$

3.6 Inference

Recall that in simple linear regression, we showed that $\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2}$ using the fact that for independent $Z \sim N(0,1)$ and $V \sim \chi^2_\nu$, then $\frac{Z}{\sqrt{V/\nu}} \sim t_\nu$. We want something similar for the multi variate case.

Since $\frac{1}{\sigma^2}\boldsymbol{e}^T\boldsymbol{e} \sim \chi^2_{n-(p+1)}$ (see Lecture 7 slides $21-24$ for a proof using eigen decomposition; outside of the scope of the course), $\frac{1}{\sigma^2}\boldsymbol{e}^T\boldsymbol{e}$ is independent of $\hat{\boldsymbol{\beta}}$ (follows from the fact that $\boldsymbol{e}$ is independent of $\hat{\boldsymbol{\beta}}$ which was shown earlier), and $\hat{\beta}_j \sim N(\beta_j, \sigma^2 V_{jj})$, then,

$$\frac{\frac{\hat{\beta}_j-\beta_j}{\sqrt{\sigma^2 \boldsymbol{V}_{jj}}}}{\sqrt{\left(\frac{1}{\sigma^2}\boldsymbol{e}^T\boldsymbol{e}\right)/(n-(p+1))}} \sim t_{n-(p+1)} \implies \boxed{\frac{\hat{\beta}_j-\beta_j}{\hat{\sigma}\sqrt{\boldsymbol{V}_{jj}}} \sim t_{n-(p+1)}}$$

where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Intuitively, we have $n-(p+1)$ degrees of freedom in the t distribution because we are estimating $p+1$ different β_i values for $i = 0, \ldots, p$.

Hypothesis Testing: Similarly to the simple linear regression case, we want to test a null hypothesis $H_0 : \beta_j = \theta_0$ (often θ_0 is 0) against an alternative $H_1 : \beta_j \neq \theta_0$. We compute the observed test statistic $t^{(obs)}$ under H_0 (assuming it is true),

$$t^{(obs)} = \frac{\hat{\beta}_j-\beta_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j-\theta_0}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j-\theta_0}{\hat{\sigma}\sqrt{V_{jj}}} \sim t_{n-p-1}$$

and we want to compute $P(|T| \geq |t^{(obs)}|)$, which in R is, `2*pt(|t_obs|, df=n-p-1, lower.tail=FALSE)` and reject if this p value is less than α (eg 0.05). Alternatively, we can find a critical value $t_{n-p-1,1-\alpha/2}$ such that the probability that less than this value is $1-\alpha/2$, implying that the probability to the right is $\alpha/2$. The critical value is a test statistic such that the p value would be exactly α, so we reject if $|t^{(obs)}| \geq t_{n-p-1,1-\alpha/2}$.

As well, recall one-sided tests from Stat 332. If we have $H_0 : \beta_j = \theta_0$ and $H_1 = \beta_j > \theta_0$, we use the same test statistic $t^{(obs)}$, since the null has not changed, but the p value is $P(T > t^{(obs)})$. In R, the command would be `pt(t_obs, df=n-p-1, lower.tail=FALSE)`.

In R, to do a two-sided HT that $H_0 : \beta_j = 0$ on all the coefficients (ie n separate HT; for all $j = 0, 1, \ldots, p$),

```
Model <- lm(data = mydataset, formula = Y ~ X1 + X2 + X3) # implicitly includes intercept
beta.hat <- coef(Model) # numerators of the test statistics
V <- solve(crossprod X) # X is the design matrix
beta.se <- sqrt(sigma(Model)^2 * diag(V)) # denominators of the test statistics (std errs)
beta.se <- sqrt(diag(vcov(Model))) # easier way to do the above using vcov
Tobs <- beta.hat/beta.se # T statistics

pval <- 2 * pt(-abs(Tobs), df = n-p-1) # computes P(|T| > |Tobs|) where T is t(n-p-1)
pval <- 2 * pt(abs(Tobs), df = n-p-1,lower.tail=FALSE) # equivalent to the previous line

# or can just see the pvalues using summary
summary(Model)
```

Confidence Intervals: To construct a $(100(1-\alpha)\%$ CI for a single coefficient β_j, we want,

$$1-\alpha = P\left(-t_{n-p-1,1-\alpha/2} \leq \frac{\hat{\beta}_j-\beta_j}{\hat{\sigma}V_{jj}} \leq t_{n-p-1,1-\alpha/2}\right)$$
$$\implies 1-\alpha = P(\hat{\beta}_j - t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{V_{jj}} \leq \beta_j \leq \hat{\beta}_j + t - n - p - 1, 1-\alpha/2\hat{\sigma}\sqrt{V_{jj}})$$

which implies that a $100(1-\alpha)\%$ CI for β_j is,

$$\hat{\beta}_j \pm t_{n-p-1,1-\alpha/2}SE(\hat{\beta}_j) \implies \boxed{\hat{\beta}_j \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{V_{jj}}}$$

3.7 Prediction

Estimating mean response: For an arbitrary row vector of covariates $\boldsymbol{x}_0 = [1, x_{01}, x_{02}, \ldots, x_{0p}]$, the mean (a scalar) is $\mu_0 = E[y_0] = \boldsymbol{x}_0\boldsymbol{\beta}$. We can estimate this as $\hat{\mu}_0 = \boldsymbol{x}_0\hat{\boldsymbol{\beta}}$. The mean is (since $\hat{\boldsymbol{\beta}}$ is unbiased),

$$E[\hat{\mu}_0] = E[\boldsymbol{x}_0\hat{\boldsymbol{\beta}}] = \boldsymbol{x}_0 E[\hat{\boldsymbol{\beta}}] = \boldsymbol{x}_0\boldsymbol{\beta} = \mu_0$$

which shows that $\hat{\mu}_0$ is an unbiased estimator for μ_0. The variance is,

$$\text{Var}(\hat{\mu}_0) = \text{Var}(\boldsymbol{x}_0\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0\text{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_0^T = \boldsymbol{x}_0\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T = \sigma^2\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T$$

Since $\hat{\mu}_0 = \boldsymbol{x}_0\boldsymbol{\beta} = \boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is a linear combination of $\boldsymbol{y}$ which is Normal, then $\hat{\mu}_0$ is also Normal. Thus,

$$\frac{\hat{\mu}_0 - \mu_0}{\sigma\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim N(0,1)$$

and by the same logic as before,

$$\frac{\hat{\mu}_0 - \mu_0}{\hat{\sigma}\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}} \sim t_{n-p-1}$$

and a $100(1-\alpha)\%$ CI is,

$$\boxed{\hat{\mu}_0 \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}}$$

Prediction: For a new response $y_{new} = \boldsymbol{x}_{new}\boldsymbol{\beta} + \epsilon_{new}$ (where $\boldsymbol{x}_{new}$ is a row vector containing the new covariates) which is independent of all the other outcomes, the prediction is $\hat{y}_{new} = \boldsymbol{x}_{new}\hat{\boldsymbol{\beta}}$ which has mean $E[\hat{y}_{new}] = \boldsymbol{x}_{new}\boldsymbol{\beta}$ and variance $\text{Var}(\hat{y}_{new}) = \sigma^2\boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T$. See Lecture 8 slide 21 for the proof, however it is the exact same as the mean and variance of $\hat{\mu}_0$ above except with $\boldsymbol{x}_{new}$ instead of $\boldsymbol{x}_0$.

Furthermore, we know that y_{new} and $\hat{y}_{new}$ are independent (since $\hat{y}_{new}$ is a function of only the observed y not including y_{new}) and normally distributed. So, since $y_{new} \sim N(\boldsymbol{x}_{new}\boldsymbol{\beta}, \sigma^2)$ by assumption,

$$y_{new} - \hat{y}_{new} \sim N(0, \sigma^2 + \sigma^2\boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T)$$

and as before,

$$\frac{y_{new} - \hat{y}_{new}}{\sigma\sqrt{1 + \boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T}} \sim N(0,1) \qquad \frac{y_{new} - \hat{y}_{new}}{\hat{\sigma}\sqrt{1 + \boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T}} \sim t_{n-p-1}$$

and a $100(1-\alpha)\%$ prediction interval for y_{new} is,

$$\boxed{\hat{y}_{new} \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T}}$$

Like in the simple linear regression case, the extra 1 under the square root captures the variability around the estimated line and the rest captures the variability of estimating the line (mean) itself.

3.8 Categorical Covariates

We now consider categorical covariates, but the outcome is still continuous.

Recall in the multiple linear regression model discussed so far, β_j for $j > 0$ is the mean difference in the outcome for every one unit increase in x_{ij}, holding all other covariates constant. However, this only applies if the covariates are continuous. Even though categorical covariates can be encoded as integers, we do not want to use these integers directly since the numeric values and how widely spaced the integers are imposes additional assumptions on the covariates.

Consider the `fishermen_mercury.csv` dataset (see files) with continuous response `MeHg` and one categorical covariate `fishpart` which can take values N, M, MW, W. Rather than encoding these categories as integers, we instead consider four separate means,

$$\texttt{MeHg}|\{\texttt{fishpart} = N\} \sim N(\gamma_N, \sigma^2) \quad \texttt{MeHg}|\{\texttt{fishpart} = M\} \sim N(\gamma_M, \sigma^2) \quad \cdots$$

In words, the response for data points with N `fishpart` is γ_N, etc. This does not require assumptions on the relative differences between categorical values. The model is then,

$$\texttt{MeHg}_i = \gamma_N I[\texttt{fishpart}_i = N] + \gamma_M I[\texttt{fishpart}_i = M] + \gamma_{MW} I[\texttt{fishpart}_i = MW] + \gamma_W I[\texttt{fishpart}_i = W] + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and are iid and $I[A]$ is 1 if A is true and 0 if A is false (an indicator function). This model essentially has four different intercepts (means) and no slope (ie four horizontal lines when plotting any other non-`fishpart` covariate against the response). An equivalent and more familiar parameterization of this model is,

$$\texttt{MeHg}_i = \beta_0 + \beta_M I[\texttt{fishpart}_i = M] + \beta_{MW} I[\texttt{fishpart}_i = MW] + \beta_W I[\texttt{fishpart}_i = W] + \epsilon_i$$

again with $\epsilon_i \sim N(0, \sigma^2)$ iid. This is the same as the multiple linear regression discussed earlier, where the "covariates" are now indicator functions of the actual covariates and the design matrix would be filled with 0 and 1. The data points with `fishpart` of N is called the referent group, since the indicator is not explicitly included in the model.

The design matrix for this formulation has a column of 1s for the intercept and three columns of 0s and 1s (corresponding to whether the data point has `fishpart` of M, MW, W), a column of 1s. There is no column for the referent group, because from the columns of the three other categories, it can be determined whether the data point is in the referent group (ie if all three columns are 0, then it must be in the referent group; the four indicators for all four categories sum to one since each data point is in exactly one category). As well, the intercept column already is filled with 1s, so there would be linearly dependent columns (and for the $\gamma_N, \gamma_M, \gamma_{MW}, \gamma_W$ model without the intercept, there is no intercept column so this problem does not appear). If the design matrix has linearly dependent columns, then $(\boldsymbol{X}^T\boldsymbol{X})$ would be non-invertible, which would not work with our theory so far. In summary, adding a column for the referent group while there is already an intercept column would be over-parameterizing. See 10:00 in the Lec 10 video for more details.

Furthermore, note that $E[y_i|\texttt{fishpart} = N] = \beta_0, E[y_i|\texttt{fishpart} = M] = \beta_0 + \beta_M, E[y_i|\texttt{fishpart} = MW] = \beta_0 + \beta_{MW}$, etc. This is four unique means described by four unknown parameters $\beta_0, \beta_M, \beta_{MW}, \beta_W$. Relating to the first model,

$$\gamma_N = \beta_0 \qquad \gamma_M = \beta_0 + \beta_M \qquad \gamma_{MW} = \beta_0 + \beta_{MW} \qquad \gamma_W = \beta_0 + \beta_W$$

or alternatively,

$$\beta_0 = \gamma_N \qquad \beta_M = \gamma_M - \gamma_N \qquad \beta_{MW} = \gamma_{MW} - \gamma_N \qquad \beta_W = \gamma_W - \gamma_N$$

so β_0 is the mean of data points with `fishpart` of N, β_M is the mean difference between data points with `fishpart` of M and those of N (the referent group), etc. For example, in an experiment with a placebo, the placebo is the referent group so that the β can represent comparisons to that placebo.

Adding a continuous covariate: Another covariate in the dataset is `weight`, measured in kg, and is continuous. How do we encode a regression model where the expected `MeHg` is linear in `weight` for each level of `fishpart`, with common slope but different intercepts? One way is,

$$\begin{aligned}\texttt{MeHg}_i = \gamma_1 \texttt{weight}_i + \gamma_N I[\texttt{fishpart}_i = N] + \gamma_M I[\texttt{fishpart}_i = M] \\ + \gamma_{MW} I[\texttt{fishpart}_i = MW] + \gamma_W I[\texttt{fishpart}_i = W] + \epsilon_i\end{aligned}$$

where γ_1 is the mean difference in the outcome for a one unit change in `weight`, for any constant value of `fishpart`. And γ_N is the mean outcome when `fishpart` is N and `weight` is 0. Another formulation is,

$$\texttt{MeHg}_i = \beta_0 + \beta_1 \texttt{weight}_i + \beta_M I[\texttt{fishpart}_i = M] + \beta_{MW} I[\texttt{fishpart}_i = MW] + \beta_W I[\texttt{fishpart}_i = W] + \epsilon_i$$

and as usual, ϵ_i is iid $N(0, \sigma^2)$ in both models. In this second model, β_0 is the mean outcome when `weight` is 0 and `fishpart` is N. And, β_1 is the mean difference in outcome with a one unit change in `weight` for any constant value of `fishpart`. Lastly, $\beta_M = E[\texttt{MeHg}|\texttt{weight} = w, \texttt{fishpart} = M] - E[\texttt{MeHg}|\texttt{weight} = w, \texttt{fishpart} = N]$; the mean difference in outcome holding `weight` constant and changing `fishpart` from M to N.

The design matrix is the same as before but with an additional column for the `weight`. The first is the design matrix for the formulation without the intercept (note the column for the referent group `fishpart` of N) and the second is the design matrix for the formulation with the intercept (note the lack of the column for the referent group, the intercept takes care of it),

	weight	fishpartnone	fishpartmuscle	fishpartmuscle_whole	fishpartwhole
1	70	0	0	1	0
2	73	0	1	0	0
3	66	0	0	1	0
4	80	0	1	0	0
5	78	0	1	0	0
6	75	0	1	0	0

	(Intercept)	weight	fishpartmuscle	fishpartmuscle_whole	fishpartwhole
1	1	70	0	1	0
2	1	73	1	0	0
3	1	66	0	1	0
4	1	80	1	0	0
5	1	78	1	0	0
6	1	75	1	0	0

Graphically, if plotting `weight` on the x-axis and `MeHg` on the y-axis, if `fishpart` is N, the regression line is $\gamma_N + \gamma_1 \texttt{weight}$, and similarly when `fishpart` is M, MW, W. These four lines are all parallel with slope γ_1 and intercepts γ_N, γ_M, etc. Using the variables in the second model, the common slope would be β_1 and intercepts $\beta_0, \beta_0 + \beta_M, \beta_0 + \beta_{MW}, \beta_0 + \beta_W$.

See R files for how to represent continuous covariates as factors then build a linear model.

3.8.1 Testing

Using this model where the expected `MeHg` is linear in `weight` for each level of `fishpart`, with common slope but different intercepts, how do we test how different groups compare to one another? The β model is easier to work with for testing.

Comparing a group to the referent: Recall N is the referent group. To test whether

$$E[\texttt{MeHg}|\texttt{weight} = w, \texttt{fishpart} = N] = E[\texttt{MeHg}|\texttt{weight} = w, \texttt{fishpart} = M]$$

the null hypothesis is $H_0 : \beta_M = 0$. This can be done using the test statistic,

$$\frac{\hat{\beta}_M - 0}{SE(\hat{\beta}_M)} \sim t_{n-(p+1)}$$

and constructing CI for β_M can be done similarly as before.

Comparing a group to another (non-referent): To test whether

$$E[\texttt{MeHg}|\texttt{weight} = w, \texttt{fishpart} = M] = E[\texttt{MeHg}|\texttt{weight} = w, \texttt{fishpart} = MW]$$

the null hypothesis is $H_0 : \beta_M = \beta_{MW}$ (ie $\beta_M - \beta_{MW} = 0$). For known σ, the test statistic is,

$$\frac{\hat{\beta}_M - \hat{\beta}_{MW}}{SE(\hat{\beta}_M - \hat{\beta}_{MW})} \sim N(0,1)$$

where,

$$\text{Var}(\hat{\beta}_M - \hat{\beta}_{MW}) = \text{Var}(\hat{\beta}_M) + \text{Var}(\hat{\beta}_{MW}) - 2\text{Cov}(\hat{\beta}_M, \hat{\beta}_{MW}) = \sigma^2(\boldsymbol{V}_{3,3} + \boldsymbol{V}_{4,4} - 2\boldsymbol{V}_{3,4})$$

where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ (the numbers come from "labelling" N, M, MW, W with $1, 2, 3, 4$ respectively. The square root of this variance is the SE. For unknown σ (ie replace σ with $\hat{\sigma}$, the test statistic is $t_{n-(p+1)}$).

Comparing more than two groups at the same time: Suppose we want to know whether the mean `MeHg` varies by `fishpart`, adjusted for weight. The null hypothesis is $H_0 : \boldsymbol{\beta}_\star = (\beta_M, \beta_{MW}, \beta_W)^T = 0$ (ie $\gamma_N = \gamma_M = \gamma_{MW} = \gamma_W$). Recall,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \implies \hat{\boldsymbol{\beta}}_\star \sim N(\boldsymbol{\beta}_\star, \sigma^2\boldsymbol{V}_\star)$$

where $\boldsymbol{V}_\star$ is the corresponding 3×3 sub-matrix of $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$, since $\boldsymbol{\beta}_\star$ is a "subset" of $\boldsymbol{\beta}$ (excluding the β_0, β_1).

We will use the fact that any variance matrix $\boldsymbol{V}$ can be uniquely written as $\boldsymbol{V} = \boldsymbol{L}\boldsymbol{L}^T$ (Cholesky decomposition), where $\boldsymbol{L}$ is a lower triangular matrix with non-negative entries $\boldsymbol{L}_{ii} \geq 0$ on the diagonal. When $\boldsymbol{V}$ is positive-definite (which can be shown is the case), then $\boldsymbol{L}_{ii} > 0$ (meaning $\boldsymbol{L}$ is non-singular since $\det \boldsymbol{L} = \text{tr}(\boldsymbol{L})$.

So, let $\boldsymbol{L}$ such that $\sigma^2\boldsymbol{V}_\star = \boldsymbol{L}\boldsymbol{L}^T$ and define $\boldsymbol{Z} = \boldsymbol{L}^{-1}(\hat{\boldsymbol{\beta}}_\star - \boldsymbol{\beta}_\star)$. It can be shown that (see 55:00 in the Lec 9 video, or slide 20) $Z \sim MVN(0, I)$ and,

$$\sum_{j=1}^{q} \boldsymbol{Z}_j^2 = \boldsymbol{Z}^T\boldsymbol{Z} = (\hat{\boldsymbol{\beta}}_\star - \boldsymbol{\beta}_\star)^T(\boldsymbol{L}^{-1})^T\boldsymbol{L}^{-1}(\hat{\boldsymbol{\beta}}_\star - \boldsymbol{\beta}_\star) = \frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}}_\star - \boldsymbol{\beta}_\star)^T(\boldsymbol{V}_\star)^{-1}(\hat{\boldsymbol{\beta}}_\star - \boldsymbol{\beta}_\star)$$

where q is the dimension of the vector $\hat{\boldsymbol{\beta}}_\star$ (in this case, 3) and $\boldsymbol{Z}_j$ is the jth entry of $\boldsymbol{Z}$.

So, under $H_0 : \boldsymbol{\beta}_\star = 0$,

$$\frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}}_\star)^T(\boldsymbol{V}_\star)^{-1}(\hat{\boldsymbol{\beta}}_\star) = \sum_{j=1}^{q} \boldsymbol{Z}_j^2 \sim \chi^2_q$$

since each $\mathbb{Z}_j \sim N(0,1)$. As well, this is independent of,

$$\frac{n-(p+1)}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-(p+1)}$$

since the $(\hat{\boldsymbol{\beta}}_\star)^T(\boldsymbol{V}_\star)^{-1}(\hat{\boldsymbol{\beta}}_\star)$ is a function of the $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ is a function of the residuals $\boldsymbol{e}$ and we showed earlier that these are independent. Now, define an F statistic,

$$\boxed{F = \frac{\frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}}_\star)^T(\boldsymbol{V}_\star)^{-1}(\hat{\boldsymbol{\beta}}_\star)/q}{\frac{n-(p+1)}{\sigma^2}\hat{\sigma}^2/(n-(p+1))}} = \frac{(\hat{\boldsymbol{\beta}}_\star)^T(\boldsymbol{V}_\star)^{-1}(\hat{\boldsymbol{\beta}}_\star)}{q\hat{\sigma}^2}$$

Recall (from Stat 332), if $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$ are independent, then $W = \frac{X_1/\nu_1}{X_2/\nu_2}$ has an F distribution,

$$W \sim F(\nu_1, \nu_2) \qquad f(w) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\left(\nu_1^{\nu_1}\nu_2^{\nu_2}\frac{w^{\nu_1-2}}{(\nu_2+\nu_1 w)^{(\nu_1+\nu_2)}}\right)^{1/2}$$

Thus, under the null,

$$\boxed{F = \frac{(\hat{\boldsymbol{\beta}}_\star)^T(\boldsymbol{V}_\star)^{-1}(\hat{\boldsymbol{\beta}}_\star)}{q\hat{\sigma}^2} \sim F(q, n-(p+1))}$$

where again, q is the length of the $\hat{\boldsymbol{\beta}}_\star$ vector. In R, the code for testing the null hypothesis is `pf(F_obs, df1=q, df2=n-p-1, lower.tail=FALSE)` (note that the F distribution is non-negative). A one-sided F test is not possible, since the numerator of the test statistic can be considered the square of $\hat{\boldsymbol{\beta}}_\star$ in matrix form, so large positive values cannot be distinguished from large negative values.

Comparison to t-test: In the above formulation, when $q = 1$, the F test is equivalent to the two-sided t test. That is, the t test for $H_0 : \beta_j \neq 0, H_1 : \beta_j \neq 0$ is equivalent to an F test for the same hypothesis. This is because the F statistic simplifies to the square of the t statistic, and it can also be shown that if $T \sim t_\nu$, then $T^2 \sim F(1, \nu)$. However, this is only for the two-sided alternative, since the F test cannot perform one-sided tests, as explained earlier. See Quiz 3 Q1 for more details.

3.8.2 Summary

The only thing which changes with categorical covariates compared to the previous theory when we exclusively considered continuous covariates is the design matrix $\boldsymbol{X}$. This matrix obviously can only contain numbers rather than categories like "low", "medium", etc. The way we incorporate data into $\boldsymbol{X}$ (eg, whether we include a column of $0, 1, 2, 3$ corresponding to `fishpart` of N, M, MW, W, or we include 3 columns of indicators for the 3 non-referent categories, etc) has implications on how we interpret our model parameters.

However, the LS estimators $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ are still the same type of object whether or not we are dealing with 0s and 1s in $\boldsymbol{X}$, or just continuous covariates. $\hat{\boldsymbol{\beta}}$ is still just a linear transformation of the continuous outcomes $\boldsymbol{y}$. So, all the previous theory carries over. As well, the variance matrix for $\hat{\boldsymbol{\beta}}$ is computed exactly the same as before, with the only difference being that $\boldsymbol{X}$ could include 0s and 1s instead of just continuous numbers, etc.

3.9 Interaction

Recall from Stat 332, interaction is when the effects of covariates alone differ from when covariates are present together. That is, the outcome-covariate association varies based on the level of another covariate.

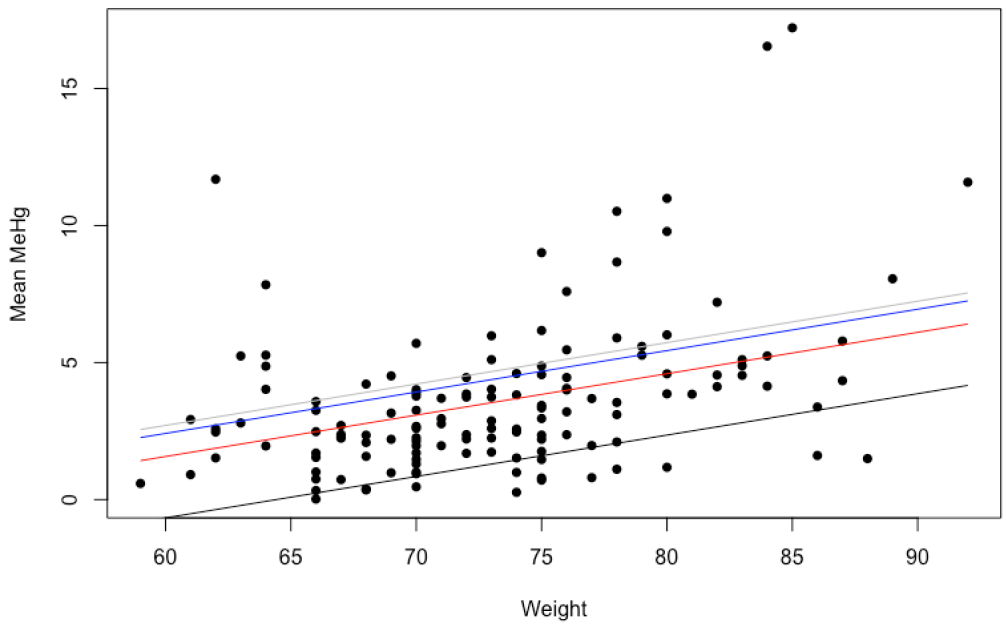
Recall the model,

$$\texttt{MeHg}_i = \beta_0 + \beta_1\texttt{weight}_i + \beta_M I[\texttt{fishpart}_i = M] + \beta_{MW} I[\texttt{fishpart}_i = MW] + \beta_W I[\texttt{fishpart}_i = W] + \epsilon_i$$

where ϵ is iid $N(0, \sigma^2)$ and the referent group is `fishpart` of N. For simplicity of notation, we will re-write this model as,

$$\texttt{MeHg}_i = \beta_0 + \beta_1\texttt{weight}_i + \beta_M M_i + \beta_{MW} MW_i + \beta_W W_i + \epsilon_i$$

which implies a common slope for `weight` for any value of `fishpart` but different intercepts for different `fishpart`. Plotted in R (see `week5(lec9+10).R` for code), where each line is for a different category,



The black data points are not color-coded by category, but depending on which dots belong to which category, it can be seen that forcing every fitted line to have the same slope can be quite restrictive.

To have both different intercepts and different slopes, we consider the model,

$$\begin{aligned}\texttt{MeHg}_i = \beta_0 + \beta_1 \texttt{weight}_i + \beta_M M_i + \beta_{MW} MW_i + \beta_W W_i \\ + \beta_{1M} \texttt{weight}_i M_i + \beta_{1MW} \texttt{weight}_i MW_i + \beta_{1W} \texttt{weight}_i W_i + \epsilon_i\end{aligned}$$

where ϵ_i is iid $N(0, \sigma^2)$. The product of two different covariates (ie $\texttt{weight}_i M_i$) is called an interaction term. The design matrix for this now has a column of 1s, a column for `weight`, three columns of 0 and 1 for M_i, MW_i, W_i, and three additional columns of $\texttt{weight}_i M_i, \texttt{weight}_i MW_i, \texttt{weight}_i W_i$,

	(Intercept)	weight	fishpartmuscle	fishpartmuscle_whole	fishpartwhole	weight:fishpartmuscle
1	1	70	0	1	0	0
2	1	73	1	0	0	73
3	1	66	0	1	0	0
4	1	80	1	0	0	80
5	1	78	1	0	0	78
6	1	75	1	0	0	75

	weight:fishpartmuscle_whole	weight:fishpartwhole
1	70	0
2	0	0
3	66	0
4	0	0
5	0	0
6	0	0

Note that if $N_i = 1$ (ie `fishpart` is N), then,

$$E[\texttt{MeHg}_i | \texttt{weight}_i = w, N_i = 1] = \beta_0 + \beta_1 \texttt{weight}_i$$

so the mean outcome when `fishpart` is N is linear in the `weight`. If $MW_i = 1$, then,

$$E[\texttt{MeHg}_i | \texttt{weight}_i, MW_i = 1] = \beta_0 + \beta_1 \texttt{weight}_i + \beta_{1MW} \texttt{weight}_i + \beta_{MW} = (\beta_0 + \beta_{MW}) + (\beta_1 + \beta_{1MW}) \texttt{weight}_i$$

The mean outcome is again linear in `weight` but with a different intercept and slope as in the case when $N_i = 1$. Using this model, rather than parallel lines as before with slope β_1, there are four lines with possibly different slope and intercept, allowing more flexibility in how the data is modelled.

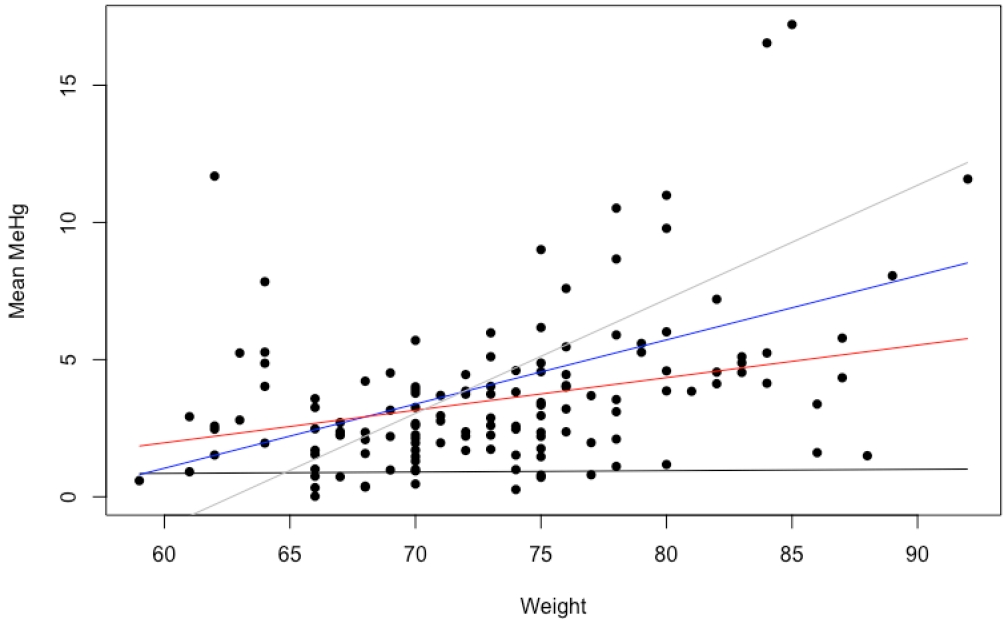
Next, β_1 is the mean difference in response with a one unit increase in `weight`, among those in the group where `fishpart` is N, since,

$$\begin{aligned}\beta_1 &= (\beta_0 + \beta_1(x^* + 1)) - (\beta_1 + \beta_1 x^*) \\ &= E[\texttt{MeHg}_i|\texttt{weight}_i = x^* + 1, N_i = 1] - E[\texttt{MeHg}_i|\texttt{weight}_i = x^*, N_i = 1]\end{aligned}$$

Similarly, $\beta_{1MW} + \beta_1$ is the mean difference in response with a one unit increase in `weight`, in the group where `fishpart` is MW,

$$\begin{aligned}\beta_{1MW} + \beta_1 &= (\beta_0 + \beta_{MW} + \beta_1(x^* + 1) + \beta_{1MW}(x^* + 1)) - (\beta_0 + \beta_{MW} + \beta_1 x^* + \beta_{1MW} x^*) \\ &= E[\texttt{MeHg}_i|\texttt{weight}_i = x^* + 1, MW_i = 1] - E[\texttt{MeHg}_i|\texttt{weight}_i = x^*, MW_i = 1]\end{aligned}$$

Thus, β_{1MW} (a coefficient of an interaction term) is a difference of differences; the increase in the slope of `weight` in the MW group from what it was in the referent group (ie from the referent group slope to the MW group slope). This is a change in the association between `weight` and outcome. This interaction model (note the different slopes and intercepts for the fitted lines for each of the four categories) plotted in R,



3.9.1 Interactions of continuous covariates

Consider,

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

where x_{i1}, x_{i2} may be continuous or categorical. The $x_{i1}x_{i2}$ is called an interaction term (or effect modifier) and x_{i1} and x_{i2} are the main effects. Note that,

$$\beta_1 = E[y_i|x_{i2} = 0, x_{i1} = x^* + 1] - E[y_i|x_{i2} = 0, x_{i1} = x^*]$$

So, β_1 is the difference in mean outcomes for a one unit change in x_{i1} when x_{i2} is 0, which may or may not be interpretable.

A better way to interpret this model is: suppose x_{i2} is fixed at x^*. Then,

$$E[y_i|x_{i1}, x_{i2} = x^*] = \beta_0 + \beta_1 x_{i1} + \beta_2 x^* + \beta_3 x_{i1} x^* = (\beta_0 + \beta_2 x^*) + (\beta_1 + \beta_3 x^*) x_{i1}$$

This shows that at every level of x_2, the conditional mean outcome is linear in x_1 and the intercept and slope of x_1 are different. In other words, a change in mean outcome due to a one unit change in x_1 varies with x_2. In practise, to interpret models like this, first choose some reasonable values to fix x_2 to, then report the change in mean outcome for a one unit change in x_1 at these values.

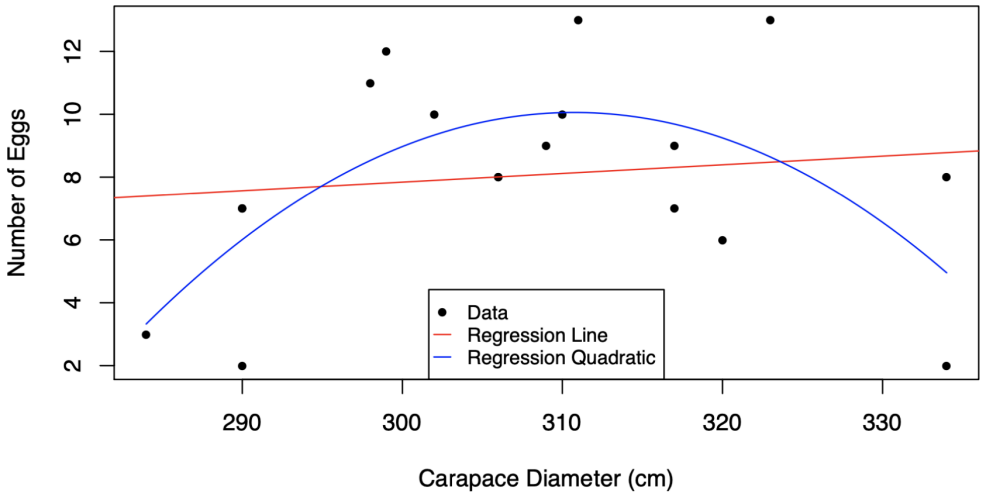
3.10 Non-Linearities

3.10.1 Quadratic Model

Sometimes $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ does not fit the data well. One way to make this model more flexible is to include a quadratic term for x,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

This is still considered a linear regression model, since it is linear to the β coefficients. Linear regression models do not need to be linear in the covariates.



Like in the interaction model, it is hard to explicitly interpret β_1, β_2 in this quadratic model. The change in mean outcome for a one unit change in x_i varies with x_i.

Testing: To test whether the quadratic model is more appropriate than the simple linear model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, we can perform HT on $H_0 : \beta_2 = 0$.

3.10.2 Non-Linear Terms

Beyond polynomial terms, linear regression can be specified quite flexibly,

$$y_i = \sum_{j=1}^{p} \beta_j f_j(x_i) + \epsilon_i$$

where $f_j(\cdot)$ are arbitrary known functions of x_i. Recall generalized linear regression from CS 480 and how kernels can convert data points to a different space to get a non-linear separator in the original space. However, using complex functions can cause overfitting. As well, there is a tradeoff between fit and interpretability.

3.10.3 Hierarchical Principle

Generally, we want to fit hierarchically well-formulated models.

If there is a higher order interaction term, the main effects and lower order interaction terms should also be included. For example, if including x_1x_2, also include x_1 and x_2. If including $x_1x_2x_3$, also include $x_1x_2, x_1x_3, x_2x_3, x_1, x_2$.

If there is a higher order polynomial term, also include the main effects and lower order terms. For example, if including x^3, also include x_2 and x.

Otherwise, there can be unexpected interpretations and implications. For example, suppose we fit the model $y_i = \beta_0 + \beta_2 x_i^2 + \epsilon_i$ and we want to center a covariate to have mean 0 (ie "shift the exposure" by some fixed amount b, where $b = \bar{x}$),

$$y_i = \beta_0 + \beta_2(x_i - b)^2 + \epsilon_i = (\beta_0 + b^2\beta_2) + (-2b\beta_2)x_i + \beta_2 x_i^2 + \epsilon_i$$

And there is now a linear term which was not there before, fundamentally changing the model structure, from just a shift in the data.

3.11 ANOVA

ANOVA (analysis of variance) attempts to characterize how much of the variability in the outcome is explained by our regression model. Recall four important equations, where $\hat{\boldsymbol{\beta}}$ is the LS estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{y}}$ are the fitted values, $\boldsymbol{H}$ is the hat matrix $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, and $\boldsymbol{e}$ are the residuals,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \qquad \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{y} \qquad \boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

We want to decompose variability around the sample mean into: variability associated with the fitted regression line $\hat{\boldsymbol{y}}$ and the remaining (variability associated with the residuals). The ANOVA decomposition to do this is, (where $\mathbf{1}$ is a $n \times 1$ vector of 1s)

$$\begin{aligned}
SSTotal &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = (\boldsymbol{y} - \bar{y}\mathbf{1})^T(\boldsymbol{y} - \bar{y}\mathbf{1}) = \boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2 \\
SSReg &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = (\boldsymbol{H}\boldsymbol{y} - \bar{y}\mathbf{1})^T(\boldsymbol{H}\boldsymbol{y} - \bar{y}\mathbf{1}) = \boldsymbol{y}^T\boldsymbol{H}\boldsymbol{y} - n\bar{y}^2 \\
SSRes &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y})^T(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}) = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}
\end{aligned}$$

where SS stands for sum of squares, $SSTotal$ is the variability of the outcomes around the sample mean (sum of squared differences; like a sample variance but not divided by $n-1$), $SSReg$ is the variability captured/explained by the distance between the fitted regression line and the sample mean, and $SSRes$ is the excess variability between the fitted values and observed mean. $SSReg$ is explained by the regression model, $SSRes$ is not. Note that,

$$\boxed{SSTotal = SSReg + SSRes}$$

The proof for this:

$$\begin{aligned}
SSTotal &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}([y_i - \hat{y}_i] + [\hat{y}_i - \bar{y}])^2 \\
&= \sum_{i=1}^{n}[y_i - \hat{y}_i]^2 + \sum_{i=1}^{n}[\hat{y}_i - \bar{y}]^2 + 2\sum_{i=1}^{n}[y_i - \hat{y}_i][\hat{y}_i - \bar{y}] \\
&= SSRes + SSReg + 0
\end{aligned}$$

using the fact that,

$$\sum_{i=1}^{n}[y_i - \hat{y}_i][\hat{y}_i - \bar{y}] = \sum_{i=1}^{n} e_i[\hat{y}_i - \bar{y}] = \boldsymbol{e}^T[\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}] = \boldsymbol{e}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} - \bar{y}\boldsymbol{e}^T\mathbf{1} = 0 - 0 = 0$$

since it was shown earlier that $\boldsymbol{e}^T\boldsymbol{X} = \mathbf{0}$ and the first column of $\boldsymbol{X}$ is $\mathbf{1}$, thus $\boldsymbol{e}^T\mathbf{1} = 0$ as well.

The **coefficient of determination**, also denoted R^2 (aka R-squared), is the proportion of variability explained by the regression model,

$$\boxed{R^2 = \frac{SSReg}{SSTotal} = 1 - \frac{SSRes}{SSTotal}}$$

and note that $SSReg, SSTotal, SSRes \geq 0$. Thus, $0 \leq R^2 \leq 1$. If R^2 is 1, then $SSRes = 0$ (all $\hat{y}_i = y_i$; all observed data points are on the fitted regression line), and if R^2 is 0, then $SSReg = 0$ (all $\hat{y}_i = \bar{y}$; the fitted regression line is exactly the sample mean line; the model only contains an intercept β_0). A higher value of R^2 implies that more variability in the outcome is explained by the regression model. In R (the language), `Multiple R-squared` in the `summary` output is R^2.

For a model with p covariates (plus a column of ones for the intercept), the ANOVA decomposition is,

Source	SS	df	MS	F
Regression	$\text{SSReg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	p	$\frac{SSR}{p}$	$\frac{SSReg}{p} / \frac{SSRes}{n-(p+1)}$
Residuals	$\text{SSRes} = \sum_{i=1}^{n}(y_i - \hat{y})^2$	$n-(p+1)$	$\frac{SSRes}{n-(p+1)}$	
Total	$\text{SSTotal} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

where MS is the mean squares (SS divided by df). Intuitively, the total variability has df $n-1$ because there are n data points and the -1 comes from computing the sample mean $\bar{y}$ (computing $\bar{y}$ corresponds to an estimate of an intercept β_0 in a model with only the intercept, since recall from simple linear regression, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, so in a model with only the intercept, $\hat{\beta}_0 = \bar{y}$). And the regression variability has df p because we have to estimate p regression coefficients $\beta_1, \ldots, \beta_p$. As well, note that MS for residuals is equal to $\hat{\sigma}^2$.

3.11.1 Testing

***F*-test for the entire model**: Suppose we want to test $H_0 : \beta_1 = \cdots = \beta_p = 0$ (are any of the covariates associated with the outcome; is any variability explained by the covariates; a large $SSReg$ goes against this hypothesis). We can conduct this test using the SS decomposition. Under the null, recall we earlier showed that,

$$\boxed{SSReg/\sigma^2 \sim \chi^2_p} \qquad \boxed{SSRes/\sigma^2 \sim \chi^2_{n-(p+1)}}$$

and since these are independent,

$$\boxed{F = \frac{SSReg/p}{SSRes/(n-(p+1))} \sim F_{p,n-(p+1)}}$$

and we reject the null if the p value is $< \alpha$. Since a large $SSReg$ goes against the null hypothesis that no covariates are associated with the outcome, a larger F test statistic gives more evidence to reject the null. This test is equivalent to the F-test formulation from earlier, if $\hat{\boldsymbol{\beta}}_\star$ were the vector containing $\beta_1, \ldots, \beta_p$. However, this F-test can be used to test a broader class of hypothesis.

***F*-test for a group of covariates**: Suppose we want to test $\beta_j = 0$ for q of the p covariates against the alternative hypothesis that at least one of the β_j are not 0. To do this, we fit the full model as before and additionally, fit a reduced model under the null (the q covariates corresponding to the β_j being tested are left out of the model). Then we consider the additional variation explained by the q covariates,

$$SSReg(FullModel) - SSReg(ReducedModel)$$

which has q degrees of freedom. It can be shown that under the null,

$$\boxed{\frac{(SSReg(FullModel) - SSReg(ReducedModel))/q}{SSRes(FullModel)/(n-(p+1))} \sim F_{q,n-(p+1)}}$$

Again, larger values of the test statistic represents evidence away from the null, since it means that there was more additional variation explained by the q covariates that are missing in the reduced model. This test is a more general version of the F-test for the entire model described above, sicne if $q = p$ (testing coefficients of all covariates), then the reduced model would be just the intercept (ie, just the sample mean line) and $SSReg(ReducedModel) = 0$.

Also for convenience, sometimes it is easier to work with $SSRes$, and use this equivalent test statistic,

$$\frac{(SSReg(FullModel) - SSReg(ReducedModel))/q}{SSRes(FullModel)/(n-(p+1))} = \frac{(SSRes(ReducedModel) - SSRes(FullModel))/q}{SSRes(FullModel)/(n-(p+1))}$$

These are equal since $SSTotal(FullModel) = SSTotal(ReducedModel)$. To compute in R, we can use `anova(ReducedModel, FullModel)`,

```
Analysis of Variance Table

Model 1: MeHg ~ weight
Model 2: MeHg ~ factor(fishpart) + weight
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    133 904.61
2    130 828.03  3    76.576 4.0074 0.009128 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this output, 904.61 is $SSRes(ReducedModel)$, 828.03 is $SSRes(FullModel)$, 76.576 is the difference, and 4.0074 is the F statistic.

***F*-test for general linear hypothesis**: We can use the same structure as described above to test a broader class of null hypotheses, called general linear hypotheses, all of the form,

$$H_0 : \boldsymbol{C\beta} = \mathbf{0}$$

where $\boldsymbol{C}$ is a matrix of rank r. This test can be used to test how "plausible" a "reduced" model (ie, certain relationships between certain coefficients) is. For example, consider $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_i$. Then, $H_0 : \beta_1 = \beta_2 = 0$ is equivalent to taking $\boldsymbol{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$. As another example, we can test $H_0 : \beta_1 = \beta_2$ using $\boldsymbol{C} = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \end{bmatrix}$, since $\boldsymbol{C\beta} = \beta_1 - \beta_2$. This is 0 iff $\beta_1 = \beta_2$.

To perform this test, we again fit the full model as well as the reduced model (ie the model assuming $\boldsymbol{C\beta} = \mathbf{0}$) and then construct the F statistic,

$$\boxed{\frac{(SSReg(FullModel) - SSReg(ReducedModel))/r}{SSRes(FullModel)/(n-(p+1))} \sim F_{r,n-(p+1)}}$$

where r is the rank of the matrix $\boldsymbol{C}$. This statistic is larger (thus the p value is smaller) when the full model is able to capture more variability than the reduced model, implying reason to reject the null.

3.12 Multicollinearity

Perfect multicollinearity: Suppose we have a model $y = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$ where $\boldsymbol{X} = \{1, x_1, x_2, x_3\}$ has a column of ones, a column of x_{i1} (denoted x_i), etc. Then suppose $x_3 = a_0 + a_1 x_1 + a_2 x_2$; the vector x_3 is a linear combination of the other columns of $\boldsymbol{X}$. Thus, the columns of $\boldsymbol{X}$ are linearly dependent; called perfect multicollinearity. In this case, we cannot compute the ordinary least squares (OLS) estimators $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, since $\boldsymbol{X}^T\boldsymbol{X}$ would not be invertible. OLS requires that there is no perfect multicollinearity. The intution for why this is a problem is that x_3 does not explain anything not already explained by x_1, x_2.

Multicollinearity more generally: What if instead of one covariate being exactly a linear combination of other covariates, it is highly correlated with other covariates?

Consider two covariates x_1, x_2 which are highly correlated. If we first regress y (aka fit a regression model) on x_1 and then regress on y on both x_1, x_2, we will see that $SE(\hat{\beta}_1)$ is much lower in the first model and thus, $\hat{\beta}_1$ is more stable (ie does not vary as much with randomly generated data). This is because in the second model, it is difficult to pull apart the variability in the outcome explained by x_1 vs that explained by x_2, since x_1, x_2 are so highly correlated. This causes more uncertainty in estimating β_1 (aka $SE(\hat{\beta}_1)$ is higher). See Lecture 12 code (in `week6.R`) for a demo.

Detecting multicollinearity: To investigate multicollinearity of x_1, x_2, we can use a scatterplot or look at the correlation $\text{Cor}(x_1, x_2)$ between the two. In a large set of covariates, we can use a grid of scatterplots or look at the correlation matrix $\text{Cor}(\boldsymbol{X})$. However, what if one covariate x_j is not strongly correlated with another one covariate but rather, moderately correlated with each of a lot of covariates (eg x_j is the sum of all the other covariates)?

To determine this, we can consider multiple linear regression where x_j is the outcome,

$$x_j = \boldsymbol{X}_{\_j}\boldsymbol{\alpha} + \epsilon^*$$

where $\boldsymbol{X}_{\_j}$ is $\boldsymbol{X}$ but excluding the column corresponding to x_j (ie regressing x_j on all the other covariates). We could consider $\text{Cor}(x_j, \hat{x}_j)$ using the fitted values $\hat{x}_j = \boldsymbol{X}_{\_j}\hat{\alpha}$ from this regression model. This lets us explain the variability in x_j in terms of the variability of all the other covariates.

Recall in simple linear regression (proven on A1) that $r^2_{yx} = R^2$ where r_{yx} is the sample correlation between y_i and x_i and $R^2 = SSReg/SSTotal$ is the coefficient of determination. For multiple linear regression, $r^2_{\boldsymbol{y},\hat{\boldsymbol{y}}} = R^2$ where $r_{\boldsymbol{y},\hat{\boldsymbol{y}}}$ is the sample correlation between the observed outcomes $\boldsymbol{y}$ and fitted values $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. See Lecture 12 slide 22 for proof. This proof uses the fact that $\boldsymbol{e}^T\boldsymbol{1} = 0$, which comes from $\boldsymbol{e}^T\boldsymbol{X} = \boldsymbol{0}$, since the first column of $\boldsymbol{X}$ is $\boldsymbol{1}$. This implies that $\bar{y} = \bar{\hat{y}}$.

Thus, the equivalent is true in the regression of x_j on $\boldsymbol{X}_{\_j}$ and examine the sample correlation $r^2_{x_j,\hat{x}_j} = R^2_j$, where R^2_j is the coefficient of determination for the regression of x_j on $\boldsymbol{X}_{\_j}$. This sample correlation gives a sense of the multicollinearity (how correlated x_j is with all the other covariates).

Recall in simple linear regression of y on x_j (ie $y = \beta_0 + \beta_j x_j + \epsilon$),

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_i (x_{ij} - \bar{x}_j)^2}$$

and in multiple linear regression of y on $\boldsymbol{X}$ (ie $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$),

$$\text{Var}(\hat{\beta}_j) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}$$

It can also be shown that,

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum (x_{ij} - \bar{x}_j)^2} \times \frac{1}{1 - R^2_j}$$

The $\frac{1}{1-R_j^2}$ is called the **variance inflation factor** (denoted VIF_j). From this equation, we see that when R_j^2 is larger (ie more multicollinearity, since the sample correlation between x_j and fitted values based on the other covariates $r^2_{x_j,\hat{x}_j}$ is larger), so is VIF_j, leading to larger standard errors of $\hat{\beta}_j$. As well, $VIF_j \geq 1$, since $0 \leq R_j^2 \leq 1$. Ideally, to reduce $\text{Var}(\hat{\beta}_j)$ overall, we want to balance increases in R_j^2 due to multicollinearity with explaining more variability in y (ie adding a covariate which is maybe correlated with the others, but explains more variability in y), which reduces our estimate of σ^2.

However, note that the σ^2 is different in simple linear regression compared to multiple linear regression. VIF_j is not the ratio of variances for $\hat{\beta}_j$ in multiple linear regression and the univariate slope estimate in simple linear regression $\hat{\beta}_j$. Rather, $VIF_j = \frac{\text{Var}(\hat{\beta}_j)}{\text{Var}(\hat{\beta}_j^*)}$, where $\hat{\beta}_j$ is a coefficient in multiple linear regression and $\boldsymbol{\beta}^*$ is the coefficient vector for a model with an idealized design matrix $\boldsymbol{X}^*$ such that,

1. $\boldsymbol{X}_j^* = \boldsymbol{X}_j$ (column corresponding to covariate j is the same in both matrices)
2. The column space of $\boldsymbol{X}^*$ is the same as that of $\boldsymbol{X}$, hence we get the same fitted values and same $\hat{\sigma}^2$.
3. $\boldsymbol{X}_j^*$ (column corresponding to covariate j) is uncorrelated with all other elements of $\boldsymbol{X}^*$

To create such an idealized design matrix $\boldsymbol{X}^*$, we start by putting the column of $\boldsymbol{X}$ corresponding to covariate j into $\boldsymbol{X}^*$ (thus satisfying the first criteria), then for all $l \neq j$, let the column in $\boldsymbol{X}^*$ corresponding to covariate l be $e_{l|j}$, where $e_{l|j} = x_l - \hat{x}_l$ are the residuals from regressing x_l on x_j (ie consider $x_l = \alpha_0 + \alpha_1 x_j + \epsilon$ then compute fitted values $\hat{x}_l = \hat{\alpha}_0 + \hat{\alpha}_1 x_j$ then take $e_{l|j} = x_l - \hat{x}_l$). See Lecture 12 Practise Q3 for an examplle of what the idealized design matrix looks like and how to compute VIF.

3.13 Summary

Estimation:

1. LS estimators: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ where $\boldsymbol{X}$ is the design matrix
2. Unbiased estimator of σ^2: $\hat{\sigma}^2 = \frac{1}{n-(p+1)}\boldsymbol{e}^T\boldsymbol{e}$ where $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ are the residuals
3. Standard error $SE[\hat{\beta}_j] = \sqrt{\hat{\sigma}^2 \boldsymbol{V}_{jj}}$ where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$

Hypothesis testing:

1. t-test for $H_0 : \beta_j = \theta_0$ has test statistic $\frac{\hat{\beta}_j - \theta_0}{\sqrt{\hat{\sigma}^2 \boldsymbol{V}_{jj}}} \sim t_{n-(p+1)}$
2. F-test for $H_0 : \boldsymbol{\beta}_\star = [0, \ldots, 0]$ against the alternative that at least one is non-zero has test statistic $F = \frac{(\hat{\boldsymbol{\beta}}_\star)^T(\boldsymbol{V}_\star)^{-1}(\hat{\boldsymbol{\beta}}_\star)/q}{\hat{\sigma}^2} \sim F(q, n-(p+1))$

CI and PI:

1. $100(1-\alpha)\%$ CI for β_j is $\hat{\beta}_j \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{V}_{jj}}$
2. $100(1-\alpha)\%$ CI for mean response with covariate vector $\boldsymbol{x}_0$ is $\hat{\mu}_0 \pm t_{n-(p+1),1-\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{x}_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0^T}$
3. $100(1-\alpha)\%$ PI for y_{new} with covariate vector $\boldsymbol{x}_{new}$ is $\hat{y}_{new} \pm t_{n-(p+1),1-\alpha/2}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_{new}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_{new}^T}$

Other interesting points:

1. Beyond the basic multiple linear regression model with continuous covariates, to accommodate for categorical covariates, interactions, and/or non-linear associations, all that needs to be changed is the design matrix (how to encode the covariates into this matrix). For example, to explore non-linear associations, add a column in the design matrix for the covariate squared (or cubed, etc). To explore interactions, add a column for interaction terms (product of one or more covariates). The rest of the theory applies.

2. When there is multicollinearity (covariates highly correlated with others or covariates which are moderately correlated with many others), the variances of estimators of coefficients will increase and this can be quantified with VIF_j for a covariate j.

3. We compute VIF to investigate the effects of multicollinearity. The more correlated your covariates, the higher the VIF. When we are investigating VIFs, it is because we are worried about high multicollinearity (highly correlated covariates), not perfect multicollinearity (one covariate is exactly a linear transformation of other covariates). In perfect multicollinearity, we cannot even compute the OLS estimators $\hat{\boldsymbol{\beta}}$.

4. R^2 is not a perfect measure of which model is "best". For example, in A2Q3a, the model which fits on `type` covariate is found to be not statistically significant according to the overall F-test, yet has the highest R^2, higher than the model which fits on `residue` which is found to be statistically significant. See in later sections about adjusted R^2 how R^2 does not decrease with more covariates, thus this metric never favors models with fewer covariates. And indeed, since `type` was a categorical covariate (which was transformed into a series of indicator covariates), this model regressing on `type` had more covariates, thus a larger R^2, than the model regressing on `residue` which is a single covariate.

5. Note that for any reduced model (ie some covariates are left out from the full model), $SSReg(Full) \geq SSReg(Reduced)$ and equivalently, $SSRes(Full) \leq SSRes(Reduced)$. Thus, the various F-statistics are non-negative, which makes sense since the F-distribution is non-negative. These inequalities hold because: least squares estimation is minimizing the $SSRes$. In a reduced model, we are restricting the space over which we are minimizing the sum of squares (fewer coefficients, since they are taken to be zero), so we cannot do any better than in the full model.

6. Consider going from one covariate to two. If x_2 is strongly correlated with x_1, R_1^2 will be large. Now if x_2 explains no more variability in the outcome, this will have the effect of inflating the variance of $\hat{\beta}_1$ as we have seen. But if it turns out that x_2 explains a whole bunch more variability of the outcome not captured by x_1, it could actually lower our estimate of σ^2, and hence the variance of $\hat{\beta}_1$ could actually be smaller in some cases. Hence there is a tradeoff here. You do not want to add highly correlated covariates that do not explain anything more than what is already explained, but you would consider adding highly correlated covariates if they do explain more variability in the outcome.

7. $SSReg, SSRes$ are independent. The intuition behind this is that $SSReg$ is a function of $\hat{\boldsymbol{\beta}}$ (the fitted model) whereas $SSRes$ is a function of $\boldsymbol{e}$ (the residuals).

8. Interaction vs multicollinearity: these two concepts are fundamentally different. Multicollinearity just means that the covariates are correlated with each other. Interactions allow the outcome-covariate associations to differ based on other covariates. This can happen with or without correlated covariates. For example, consider mentos and coke: mentos and coke consumption may be totally uncorrelated, but the effect of coke is definitely different in the presence of a lot of mentos. Symmetrically, just because two covariates are highly correlated does not imply there should be any interaction.

   That being said, an interaction term x_1x_2 may be strongly correlated with x_1 or x_2 (for example, consider a binary covariate and a continuous one and consider their correlations with their product). In this way, when you have higher order terms it is a good idea to check for multicollinearity.

9. If there are two categorical covariates (eg A with categories a_1, a_2 and B with categories b_1, b_2), a possible model might be: $y_i = \beta_0 + \beta_1 I(A = a_2) + \beta_2 I(B = b_2) + \epsilon_i$, where I are the usual indicator functions and β_0 now represents referent groups for both covariates (ie, β_0 is the mean outcome for those with $A = a_1$ or $B = b_1$).

10. We can deal with multicollinearity by repeatedly removing covariates with high VIF. See A3.

11. When multicollinearity occurs, least squares estimates are still unbiased, but their variances are large so they may be far from the true value.

4 Model Building

Given data, how do we know which model to fit and how do we compare fitted models to find the "best" one? According to statistician George Box, "all models are wrong but some are useful"; in practise, it is impossible to create a regression model which exactly matches the underlying, but we can still choose a reasonably useful model.

The broad goal of regression is to characterize how the outcome varies with the covariates. This might be determining the association between certain covariates and the outcome (inference), or predicting an outcome given covariates (prediction), etc. Different models are better suited to answering these different questions.

4.1 Principles

Some model building principles to decide which model to fit are: interpretability, parsimony, goodness of fit, predictive accuracy.

4.1.1 Interpretability

When the goal of regression analysis is to make inferences about the relationship between the outcome and some covariates. The model is useful to the extent that it can be interpreted. For example, $y_i = \beta_0 + \beta_1 \Gamma(x_i) + \epsilon_i$ might fit the data better but is less interpretable than $y_i = \beta_0' + \beta_1' x_i + \epsilon_i'$. In the second, we can easily interpret β_1' whereas in the first, it is harder to interpret β_1 (how to explain the Γ function).

However, this does not mean we never pick the more complex option. There is often a tradeoff between the complexity and interpretability in models. As well, we may have to do more work to make a complex model interpretable (plot fitted values with CIs, reporting mean differences for specific contrasts, etc).

4.1.2 Parsimony

We tend to prefer models with fewer parameters. Firstly, adjusting for more covariates makes interpreting coefficients more difficult (less interpretable). As well, more parameters increases $\hat{\sigma}^2 = \frac{SSRes}{n-(p+1)} = \frac{\sum_i (y_i - \hat{y}_i)}{n-(p+1)}$ by increasing p, which increases the degrees of freedom (less precise). Lastly, more parameters increases the uncertainty of predictions, since recall $SE(\hat{y}_{new}) = \hat{\sigma}\sqrt{1 + \boldsymbol{x}_{new}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_{new}}$.

4.1.3 Goodness of Fit

Goodness of fit measures how well the model fits the observed data. Four such criteria to determine model fit are: R^2, adjusted R^2, mean squared error, and AIC and related.

1. R^2: Recall $R^2 = 1 - \frac{SSRes}{SSTot}$ and is the proportion of variability explained by the model. However, R^2 never decreases when adding more variables. The intuition behind why is because the OLS estimators $\hat{\boldsymbol{\beta}}$ minimizes $SSRes$ where $\hat{y}_i$ is in the column space of $\boldsymbol{X}$ (since $\hat{y}_i = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is a linear combination of the columns of $\boldsymbol{X}$). Increasing the column space of $\boldsymbol{X}$ (by having more covariates) increases the space over which we are minimizing. Hence in a larger space, we could never do worse than in the reduced space.

   This makes comparing models of different size using R^2 difficult, as it would always favor the larger model.

2. Adjusted R^2: To fix this issue with R^2, we can use adjusted R^2,

$$\boxed{R^2_{adj} = 1 - \frac{SSRes/(n-(p+1))}{SSTot/(n-1)}}$$

where there are p covariates and 1 intercept. The intuition is: since $SSRes$ is non-increasing (ie adding more explanatory variables leads to a better estimate) but p increases with the number of variables, if $SSRes$ decreases only slightly or not at all, it could be outweighed by increasing the degrees of freedom and thus decrease R^2_{adj}. As well, if $SSRes$ decreases a lot, it can outweigh the increase in the degrees of freedom and increase R^2_{adj}.

Overall, R^2_{adj} is more useful for comparing models of different size. We generally prefer the model with higher R^2_{adj}. However, this value no longer represents the proportion of variance explained by the model.

It can also be shown that (see Lecture 13 slide 15),

$$\boxed{R^2_{adj} = R^2 - \left(\frac{p}{n-p-1}\right)(1-R^2)}$$

Since the second term is positive, R^2_{adj} can be considered a penalized version of R^2. A larger p causes a larger penalty. As $n \to \infty$, R^2_{adj} converges to R^2.

3. Mean Squared Error: The mean squared error is $\hat{\sigma}^2 = SSRes/(n-p-1)$.

   For adjusted R^2, it can be shown that,

   $$R^2_{adj} = 1 - \frac{\hat{\sigma}^2}{MSTot}$$

   so since $MSTot$ is a fixed value for a given dataset and does not depend on the covariates, minimizing $\hat{\sigma}$ is equivalent to maximizing R^2_{adj}. Rather than use R^2_{adj}, we could equivalently choose the model with the lowest $\hat{\sigma}$.

4. AIC, BIC, and Related: In the same sense as the penalty term in adjusted R^2, we could use other penalized criteria, such as the Akaike Information Criterion (AIC). Define,

   $$\boxed{AIC = -2\log\mathcal{L}(\hat{\theta}) + 2k}$$

   where $\theta = [\beta_0, \ldots, \beta_p, \sigma^2]$ is a vector of parameters of interest (all the unknown regression coefficients), $\log\mathcal{L}(\hat{\theta})$ is the log likelihood function at $\hat{\theta}$ (log is actually ln; base e), and k is the number of all parameters estimated, including the intercept and σ^2 (ie $p+2$ in the multiple regression case with an intercept; more generally, the length of θ).

   We prefer a model with a lower AIC. We want to maximize the log likelihood (hence minimize the first term) and this is subject to a penalty term for the number of parameters.

   Recall for multiple linear regression, $\hat{\boldsymbol{\beta}}$ is the MLE under the assumption of normality. Using the likelihood function derived earlier, it can be shown that (see Lecture 13),

   $$AIC = n(1 + \log(2\pi) + \log(SSRes/n)) + 2(p+2)$$

   Other information critera: This concept of a penalty term comes from a broader class of measures of model fit. Another example is the Bayesian information criterion,

   $$\boxed{BIC = -2\log\mathcal{L}(\hat{\theta}) + k\log(n)}$$

   which produces a larger penalty term compared to AIC because the $\log(n)$ factor of k replaces the 2. The BIC also depends on the sample size (whereas AIC does not). Note that AIC, BIC can both be computed using the `summary` output in R (using the fact that the `Residual standard error` is $\sqrt{\hat{\sigma}^2} = \sqrt{SSRes/(n-p-1)}$.

Also, DIC, WAIC, and others which apply more broadly (eg non-parametric cases where it is not easy to count the number of variables k). These all try to balance the model fit with a penalty for more variables.

Overfitting: Some of these goodness of fit approaches, such as AIC and adjusted R^2, attempts to penalize overly large models to avoid overfitting (models with many parameters which appear to fit our data very well but do not generalize well; recall from CS 480). Also, we've seen that adding more covariates reduces $SSRes$, causing R^2 to be a poor metric for comparing two models.

4.1.4 Predictive Accuracy

Out-of-sample prediction error: Out-of-sample error measures how well a model predicts new outcomes for new samples. Criteria like AIC, BIC attempt to approximate this error based on in-sample error (ie $SSRes$) and also explicitly penalize the number of parameters.

However, perhaps we can examine out-of-sample error directly.

Holdout approach: We split n observations into a training set S_{train} (of n_{train} observations) and a test set S_{test} (of size $n_{test} = n - n_{train}$), also called the holdout set. Then, we fit our model to the training set and get estimates $\hat{\boldsymbol{\beta}}_{train}$ and use these to predict the outcomes for samples in the test set (ie, $\hat{y}^{new} = \boldsymbol{x}_{test}\hat{\boldsymbol{\beta}}_{train}$ for $\boldsymbol{x}_{test} \in S_{test}$). Then, we measure the prediction accuracy with the mean squared prediction error (MSPE) against the actual observations $y^{new} \in S_{test}$,

$$MSPE = \frac{1}{n_{test}} \sum_{i \in S_{test}} (y_i^{new} - \hat{y}_i^{new})^2$$

or its square root (RMPSE). In summary, we fit the model on only a part of the given data, predict the outcomes for the remaining data, and compare against the actual observed outcomes in the remaining data. Test outcomes do not appear in the fitting process.

Cross validation: The downsides of holdout is that we are not using all the data to fit the model. Alternatively, in cross validation,

1. Randomly divide the data into training set $S_{train,j}$ and test set $S_{test,j}$
2. Fit the model on $S_{train,j}$ only
3. Predict outcomes on the $S_{test,j}$ and compute,

$$MSPE_j = \frac{1}{n_j} \sum_{i \in S_{test,j}} (y_i^{new} - \hat{y}_i^{new})^2$$

4. Repeat steps $1 - 3$ for $j = 1, \ldots, J$, where J is the number of iterations
5. Compute $MSPE_{cv} = \frac{1}{J} \sum_{j=1}^{J} MSPE_j$ (or look at the distribution of all of them, etc).

Depending on the model and how big J is and the complexity of the model, this fitting (training) process can be slow.

k-fold cross validation: The procedure is,

1. Randomly divide the data into K parts (folds)
2. Fit the model on $(K - 1)$ of the K folds (leaving the kth fold out; the $(K - 1)$ folds make up the training data)

3. Predict outcomes on the kth part (as a test set $S_{test,k}$) and compute,

$$MSPE_k = \frac{1}{n_k} \sum_{i \in S_{test,k}} (y_i^{new} - \hat{y}_i^{new})^2$$

4. Repeat steps $2-3$, leaving out each of the k folds once

5. Compute $MSPE_{cv} = \frac{1}{K} \sum_{k=1}^{K} MSPE_k$ (or the best $MSPE$, etc)

See also CS 480 and CS 486.

Leave-one-out cross validation: Consider k-fold cross validation but with $K = n$. This means we fit the model to the entire data set except for one observation and compute $\hat{y}_{i,(-i)}$ (where the ith observation is left out). Then the MPSE is $(y_i - \hat{y}_{i,(-i)})^2$. In linear regression, it turns out this is equal to the square of the PRESS statistic,

$$\frac{e_i}{1 - h_i}$$

where h_i is the ith diagonal of the hat matrix $\boldsymbol{H}$ and e_i is the residual $y_i - \hat{y}_i$, where $\hat{y}_i$ is the prediction from the model fitted on all the data. Hence, we can compute the leave-one-out cross validation MSPEs by just fitting one model on all the data and do not need to refit the model $K = n$ times (although this is not true for more complicated methods).

Training, validation, test data: When reporting prediction error for the final model after selection, we typically consider 3 types of data: training, validation, and test data. The training set was used to fit the model. The validation set is used to compare the predictions to validation outcomes (lets us choose the model with the lowest RMSPE) and also, the best model is eventually refitted on the training and validation set. Everything described earlier regarding the "test set" would actually be the validation set. The test set is used to estimate prediction error based on the final model fit (which was fit on the training and validation set).

Summary: If doing cross validation, we can consider splitting the data into a training and test set then repeatedly splitting the training set further into a training and validation set. The validation set is used to select the best model. The best model is fit on the training and validation set then tested on the test set. Since the test set was never used for training and is only used at the very end, it can be used to report out-of-sample error. See CS 480 for more details.

4.2 Automatic Selection

4.2.1 Manual model selection

We can compare a set of M models manually using the criteria above (such as the four goodness of fit criteria), but this is hard to do if M is large. Instead, we want to automate the selection procedure.

4.2.2 Automatic selection

For example, suppose the number of possible covariates p is large and we want to find the subset of covariates with the "best" properties to build a regression model with. Automatic selection gives us rules for selecting an appropriate subset of these covariates for our model. It is one approach to model building but is not necessarily the go-to and has its own flaws. Some techniques for automatic selection are best subset, forward selection, backward elimination, stepwise selection, LASSO, etc.

4.2.3 Best subset

We could try fitting 2^p models, each corresponding to a subset of the covariates (and there are 2^p subsets of p covariates by Binomial theorem, including the empty subset which corresponds to a model with just an intercept), then choose the best one. This guarantees we choose the best subset but is often impossible in practise (side note: this approach is supposedly doable up to $p = 35$, which is not that large of a p).

4.2.4 Forward selection

The procedure is (where $\mathbf{1}$ is a vector of ones),

1. Start with a model M_0 containing no covariates,

$$M_0 : \boldsymbol{y} = \beta_0 \mathbf{1} + \boldsymbol{\epsilon}$$

2. Fit all p models including exactly one covariate,

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_j + \boldsymbol{\epsilon} \qquad ; j = 1, \ldots, p$$

   and pick $\boldsymbol{x}_*$ which performs the best according to some criteria (can be the smallest p-value for the hypothesis $H_0 : \beta_1 = 0$, AIC, or any other reasonable metric). The current model is then,

$$M_1 : \boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_* + \boldsymbol{\epsilon}$$

3. Fit all $p - 1$ models including exactly two covariates, one of which is $\boldsymbol{x}_*$,

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_* + \beta_2 \boldsymbol{x}_j + \boldsymbol{\epsilon} \qquad ; \text{for all } \boldsymbol{x}_j \neq \boldsymbol{x}_*$$

   and pick $\boldsymbol{x}_{**}$ which performs best according to the same criteria. The current model is then,

$$M_2 : \boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_* + \beta_2 \boldsymbol{x}_{**} + \boldsymbol{\epsilon}$$

4. Repeat in this way (fit all models including an extra covariate and all the ones chosen so far) until we run out of covariates or the next covariate stops improving according to the metric (eg adding any of the remaining covariates causes the AIC to go up, p-value goes above 0.05, etc).

In this procedure, once a variable enters the regression, it is never removed. We start by considering p models then $p - 1$ models, etc. This is at most $(p + 1)p/2$, which is much less than 2^p. This procedure can be considered "greedy" (choose the best covariate at each step). It does not guarantee the "best" model (eg, the first covariate chosen $\boldsymbol{x}_*$ can be the best one out of all models with only one covariate, however does not belong in the overall best model).

Also, note we can do this procedure even when $p > n$, because we would definitely hit a stopping condition (metric stops improving) before this point.

4.2.5 Backward elimination

The procedure is essentially the backwards of forward selection,

1. Start with the full model,

$$M_p : \boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_p \boldsymbol{x}_p + \boldsymbol{\epsilon}$$

   which contains all potential covariates and the intercept.

2. Drop the least important covariate according to our criteria (ie which covariate, when dropped, has the greatest impact on reducing AIC, BIC, or causes the highest p-value greater than 0.05, etc). Dropping means to refit the current model on all covariates excluding this dropped covariate. Note that for the p-value metric, if covariates are categorical, we would need an F-test, since you cannot drop single indicator covariates and must drop an entire categorical covariate at a time. Otherwise, the p-value corresponds to the hypothesis that the corresponding coefficient is zero.

3. Continue in this way until dropping a variable does not improve the metric (eg AIC gets worse by dropping any covariate or if all covariates have p-values less than 0.05).

Unlike forward selection, we cannot have $p > n$ since the first step involves fitting a model with all p covariates and the design matrix would have linearly dependent columns (since there are more columns than rows), and this would cause $\boldsymbol{X}^T\boldsymbol{X}$ to be non-invertible, preventing us from fitting the model.

As well, once a variable is removed, it cannot be added back into the model. Backward elimination can sometimes perform better than forward selection (for example, if the best model is $\boldsymbol{y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 + \boldsymbol{\epsilon}$ and $\boldsymbol{x}_3$ has the strongest association on its own but is weak when $\boldsymbol{x}_1, \boldsymbol{x}_2$ are present).

See A3Q2 for an implementation.

4.2.6 Stepwise selection

The procedure is,

1. Start with a model containing no covariates,

$$M_0 : \boldsymbol{y} = \beta_0\mathbf{1} + \boldsymbol{\epsilon}$$

2. Add one covariate according to the same criteria as forward selection (fit p models each containing one covariate then choose the best one).
3. Assess whether any of the covariates should be removed as in backward selection.
4. Repeat $2 - 3$ and stop when the most recently added covariate is removed.

Note: With forward selection, backward elimination, and stepwise selection, the inference is technically not valid in the final model (post-selection inference); the SEs after these procedures are too small, which affects CI and causes p-values to be also too small. The intuition for why this is the case: we typically start with a model and use the data to conduct some hypothesis test or compute a CI. But with this approach, we use the data to select the model, causing the uncertainty to be underestimated. As well, the covariates chosen by this selection approach influences which hypothesis are even possible to be tested (eg cannot test a hypothesis that a coefficient is 0 if that coefficient is not even in the model). In contrast, classical inference requires the model/hypothesis to be fixed.

4.2.7 LASSO and other shrinkage methods

The three selection approaches discussed so far are fairly primitive and no longer used in practise. LASSO (least absolute shrinkage and selection operator) is a more modern approach for automatic selection.

Recall AIC, BIC which penalize models for having a large number of covariates. The intuition behind LASSO and other shrinkage methods incorporate a penalty term directly into the estimation procedure. Recall the objective of OLS was,

$$\min \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$$

Instead, our penalized objective will be,

$$\min \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \text{penalty}$$

Specifically, the LASSO estimator is defined as the $\boldsymbol{\beta}$ which minimizes,

$$\boxed{\min \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|}$$

Note that $\sum_{j=1}^{p}|\beta_j|$ is the L_1 norm of $\boldsymbol{\beta}$ excluding β_0 (and AIC, BIC penalties instead use an L_0 norm; equal to the number of nonzero elements).

This penalty term has the effect of shrinking parameter estimates towards zero. The choice of L_1 norm is going to shrink certain parameter estimates all the way to zero, which is equivalent to performing automatic variable selection.

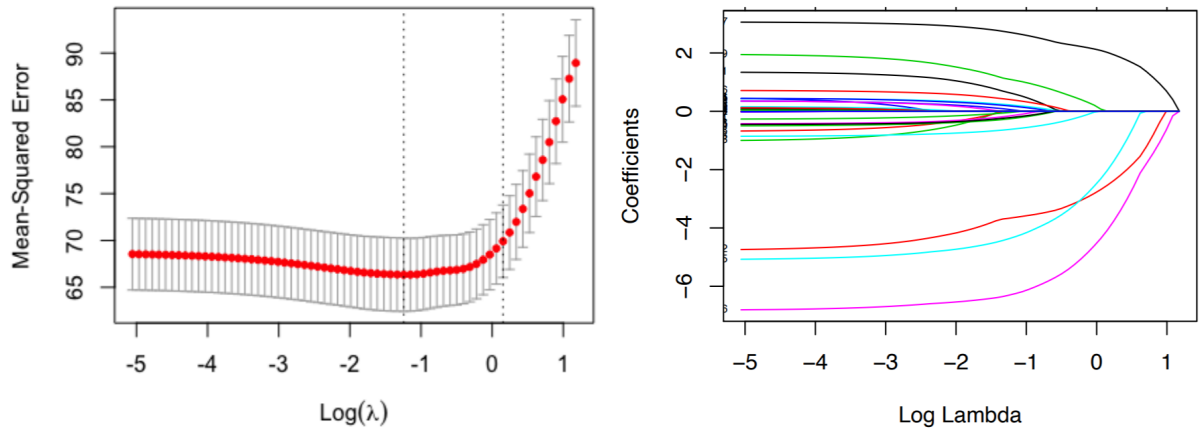
Moreover, this expression with the λ is a Lagrangian and minimizing it to obtain $\hat{\boldsymbol{\beta}}$ is equivalent to,

$$\min \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 \text{ such that } \sum_{j=1}^{p}|\beta_j| \leq t$$

for some t. So, the penalty term really corresponds to a constraint.

Different values of λ will lead to different model fits but λ is unknown. To solve this, we fix λ to some value, fit the model using cross validation, and repeat this for a grid of different λ values. Then, we choose the final λ (which corresponds to a certain model fit) which performs best. This procedure is done in R automatically.

For example, the following left graph visualizes how prediction error changes with λ. For each λ, we might do K-fold cross validation. Then, return the λ with the lowest mean-squared prediction error. The left dotted line is the λ at which the mean-squared prediction error is smallest, the right dotted line is the largest λ which has a mean-squared error within 1 standard deviation of the mean-squared error at the minimum. The right graph shows how the absolute value of coefficients shrink to zero as λ increases (penalty term becomes heavier).

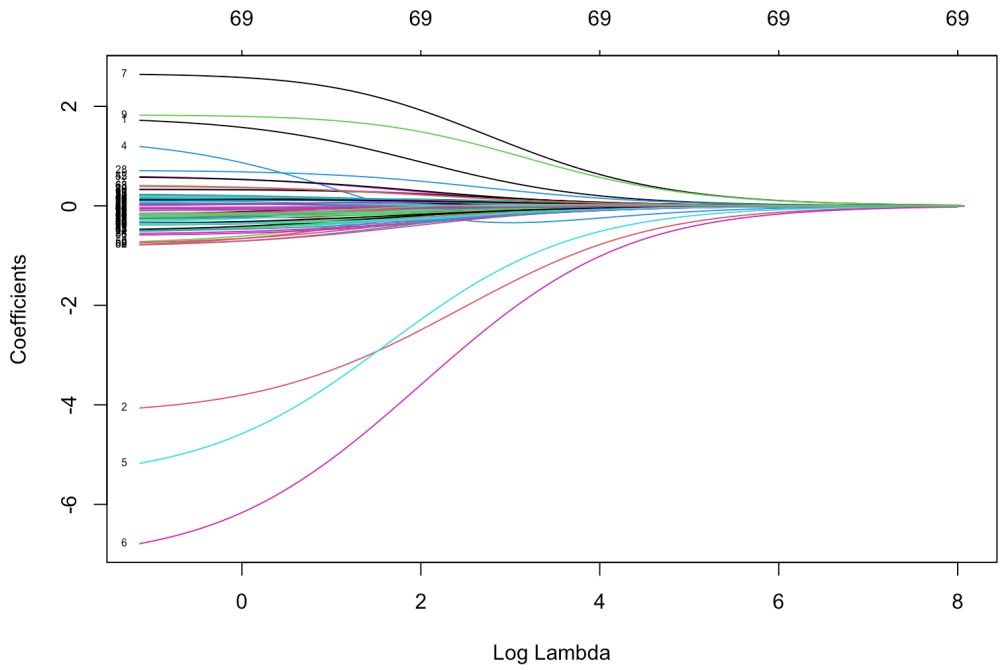


Overall, LASSO can prevent overfitting, especially when you have more parameters than observations. When predictors are correlated, LASSO may give a better selection than stepwise methods.

Ridge regression: Instead of a L_1 norm for the penalty, ridge regression uses a L_2 norm,

$$\min \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda\sqrt{\sum_{j=1}^{p}\beta_j^2}$$

Ridge regression also shrinks estimates (aka regularization, see CS 480). However, unlike LASSO, ridge regression does not shrink them all the way to zero, so does no perform automatic selection like LASSO does. In the following graph, even though the coefficients all approach zero, they do not instantly shrink to exactly zero like in LASSO.



Geometric interpretation of L_1, L_2 penalties: Recall the penalties for LASSO and ridge correspond to constraints,

$$\sum_{j=1}^{p}|\beta_j| \leq t \qquad \sqrt{\sum_{j=1}^{p}\beta_j^2} \leq t$$

Geometrically, in the 2D case (ie $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$), OLS minimization is equivalent to exploring some contour in 2D space and finding the minimum point as $\hat{\boldsymbol{\beta}}$. With the two constraints above, we perform the same minimization but only in a sub-region of 2D space which satisfies the constraints. The LASSO constraint forms a diamond shape in 2D space and the ridge constraint forms a circle. For LASSO, the minimal point tends to be at a "corner" of the diamond (meaning one coefficient is zero) whereas for ridge, the minimal point is likely anywhere on the edge of the circle.

Relaxed LASSO: Both LASSO and ridge have their uses but only LASSO grants variable selection. However, LASSO combines variable selection with shrinkage; these two objectives can be separated with relaxed LASSO. The steps are,

1. Fit LASSO and obtain an optimal λ using cross validation. This λ corresponds to a certain model fit (ie a certain $\hat{\boldsymbol{\beta}}$).
2. Fit LASSO to the subset of covariates whose coefficients were not set to zero in step 1 (ie those selected into the model),

$$\min \sum_{i=1}^{n}(y_i - (\boldsymbol{x}_i^*)^T\boldsymbol{\beta}^*)^2 + \phi\lambda\sum_{l=1}^{p}|\beta_l^*|$$

   where $\boldsymbol{x}_i^*, \boldsymbol{\beta}^*$ only contain covariates whose coefficients were not set to zero in step 1. The ϕ term lets us tune the ultimate level of shrinkage. $\phi = 1$ gives the LASSO estimator, $\phi = 0$ gives the OLS estimator on the subset of selected variables (ie no penalty term but with the subset of selected variables), and $0 < \phi < 1$ allows for different levels of shrinkage, independent of the selection.

Other shrinkage estimators: elastic net (combines L_1, L_2 penalties), fused LASSO, group LASSO, etc.

5 Model Diagnostics

How do we diagnose problems in our regression model? Recall the assumptions we made in our regression model: linearity (ie $E[\boldsymbol{y}|\boldsymbol{x} = \boldsymbol{x}^*] = \boldsymbol{X}\boldsymbol{\beta}$), independence of error terms, error terms are Normally distributed, and error terms have equal variance (homoskedasticity). How can we determine whether these assumptions hold?

What changes when these assumptions do not hold?

Estimation: Minimizing $\sum_i (y_i - x_i^T\boldsymbol{\beta})^2$ is still a reasonable thing to do, so we can get our usual OLS estimator $\hat{\boldsymbol{\beta}}$. However, the estimator $\hat{\boldsymbol{\beta}}$ may no longer be unbiased, since,

$$E[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E[\boldsymbol{y}]$$

and we need linearity to say that $E[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}$, which would lead to $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. Without linearity, $E[\hat{\boldsymbol{\beta}}] \neq \boldsymbol{\beta}$. Note the other three assumptions were not necessary for unbiased estimates.

Inference: Regarding the standard errors, we need independence and homoskedasticity to say that $\text{Var}(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$, which would lead to $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. If either of these two assumptions is not met, our variance estimates will be incorrect. Hence, our SEs, CIs, etc will be invalid.

Normality: Without Normality, $\hat{\boldsymbol{\beta}}$ is no longer a linear transformation of a MVN vector, hence it is no longer Normally distributed, so our CIs, tests are not necessarily valid. However, in large samples, $\hat{\boldsymbol{\beta}}$ is approximately Normally distributed due to the Central Limit Theorem,

$$\frac{\sqrt{n}(\bar{z} - E[z])}{\sqrt{var(z)}} \to N(0,1)$$

So we can get away with valid inference despite non-normal errors in large enough samples. This involves replacing critical $t_{n-p-1,\alpha/2}$ values with $z_{\alpha/2}$, where $z \sim N(0,1)$.

Prediction: Prediction intervals explicitly require Normality: $y_{new} \sim N(\boldsymbol{x}_{new}^T\boldsymbol{\beta}, \sigma^2)$. Without Normality, our prediction intervals are invalid and we cannot apply CLT since we only have one observation y_{new}. However, the predictions are still unbiased, since the $\hat{\boldsymbol{\beta}}$ are still unbiased even without Normality. Prediction intervals are sensitive to all 4 assumptions.

5.1 Residuals

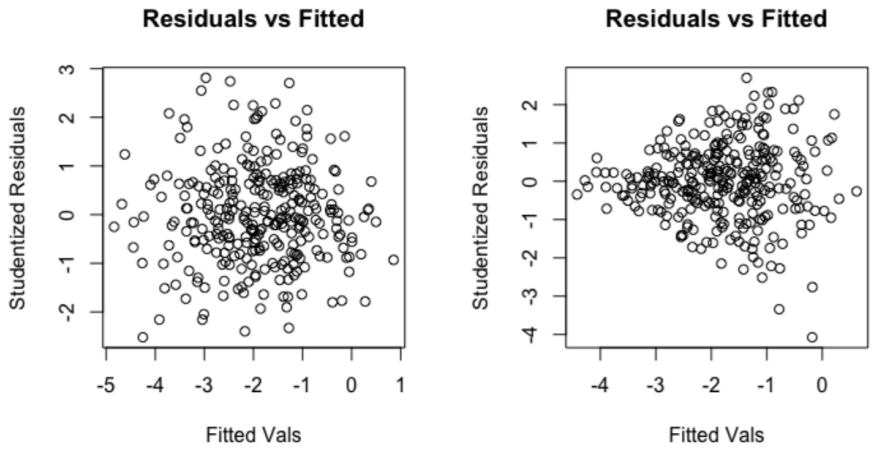
5.1.1 Identifying Issues

Since residuals can be considered estimates of the error terms, one of the best tools for diagnostics is to visualize residuals. We can use ordinary residuals $e_i = y_i - \hat{y}_i$ or studentized residuals $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$ where h_i is the ith diagonal of the hat matrix $\boldsymbol{H}$.

The intuition for studentized residuals is that $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ is $N(0, \sigma^2(\boldsymbol{I} - \boldsymbol{H}))$. Hence, $e_i \sim N(0, \sigma^2(1-h_i))$, meaning e_i have different variances, so it is difficult to learn anything about their distribution. By contrast, $e_i/\sqrt{(1-h_i)}$ has constant variance σ^2 so they should look normally distributed when plotted. Note that in practise, we estimate $\hat{\sigma}$ so the studentized residuals are really t-distributed.

Asseessing Normality: The histogram of studentized residuals should look like the density for $N(0,1)$. We can also use Normal-QQ plots (recall from Stat 231) to compare quantiles of studentized residuals to theoretical quantiles of $N(0,1)$ to assess Normality. Points should fall on a 45 degree line.

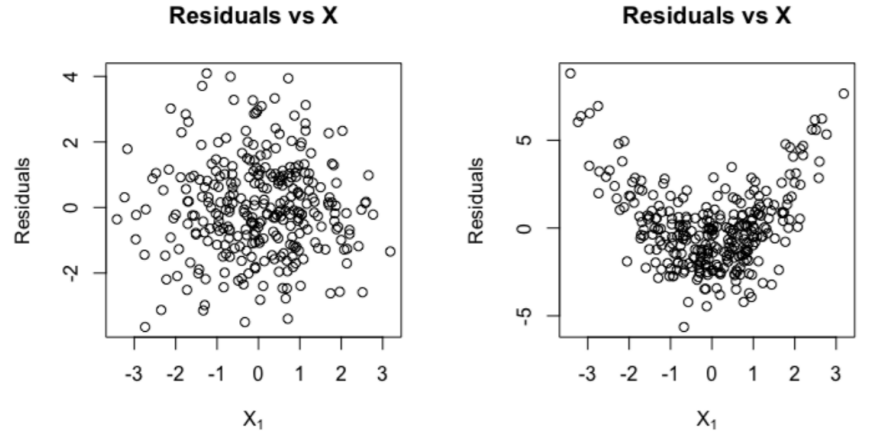
Assessing Heteroskedasticity: We can plot the studentized residuals against fitted values, which lets us detect mean-variance relationships (ie a funnel shape implying that there is higher variance for larger fitted values).



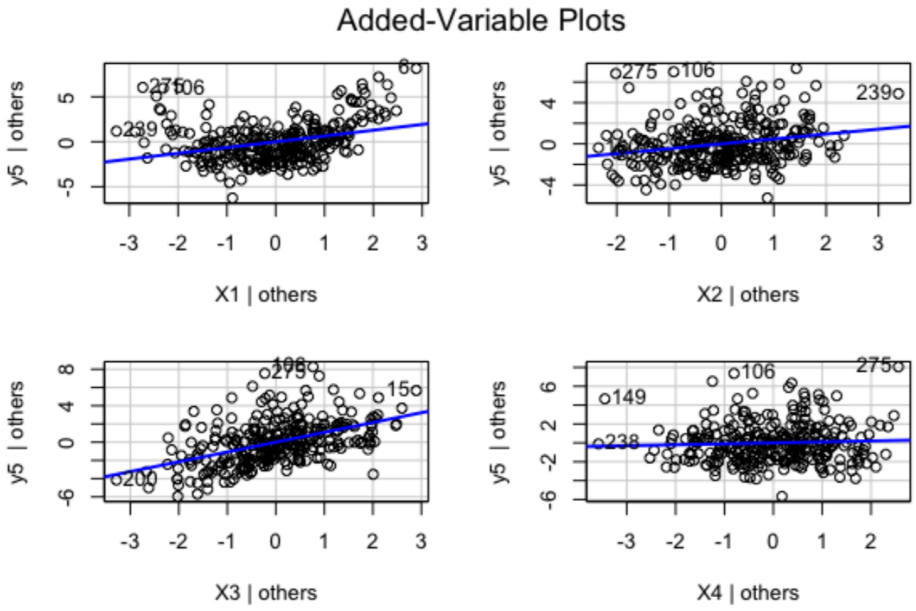
There does not appear to be a mean-variance relationship on the left but does on the right. Do not be confused by a lack of data causing "less" variability. For example, on the left, the variability seems to be less for fitted values of 0 than fitted values of -2 but that is because there are fewer points there.

Assessing Independence: This is difficult to visualize unless you have something like time-series data. Also, note that residuals are not independent even when the errors are, since $\sum e_i = 0$. Instead, we often consider how data was collected. For example: if sampling data from patients, clustered within hospitals, we might expect patients within a hospital or are being treated by the same doctor to have outcomes more similar to each other compared to patients at different hospitals.

Assessing Linearity: For the simple linear regression case, the linear assumption is that $E[y_i] = \beta_0 + \beta_1 x_i$. We can assess linearity by plotting y_i against x_i (should look linear). Or, sometimes it is easier to ideentify non-linearity by plotting residuals $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ against x_i (should look "random", a pattern may indicate non-linearity).



For the multiple linear regression case, plotting y_i against x_i ignores the effect of all the other covariates. Instead, we can visualize partial regression plots (aka added variable plots). To assess linearity in a covariate x^*, we first regress y on all other covariates, get fitted values from this model fit, and compute the residuals e_y. Then, we regress x^* on all the other covariates and get fitted values from this model fit, and compute the residuals e_{x^*}. Then plot e_y against e_{x^*}. Intuitively, we are isolating the $y \sim x^*$ relationship. Using residuals after regressing on all the other covariates is like adjusting for the other covariates.



In the above graphs, there is an outcome $y5$ and four covariates X_1, X_2, X_3, X_4. A linear plot (as is the case for X_2, X_3, X_4) indicates linearity between the outcome and the covariate.

5.1.2 Fixing Issues

Assuming we find that there are violations of any of the four assumptions, how do we fix them?

Linearity: If linearity is not met, we might consider transforming x_j. For example, include $\log(x_j)$ as a covariate or use a quadratic model with x_j and x_j^2, etc. However, this changes the interpretation of coefficients.

Independence: Violations of independence require more advanced regression methods. Without independence, the estimates are still unbiased but the SEs are broken, so we can replace SEs with robust alternatives such as sandwich form SEs, GEE. Or, we can explicitly model the dependence structure with a mixed effects model.

Normality: Violations of normality may not be a big problem if the sample size is large. However, normallity is needed for valid prediction intervals. We could consider transforming the outcome (eg take the log), which may change interpretations, although this is not a problem if we are only concerned with prediction. We could also consider other regression approaches such as generalized linear models.

Homoskedasticity: If errors are heteroskedastic (ie variance of different error terms are different), we can transform the outcome as before (called variance stabilizing transform) or use weighted least squares or bootstrap.

With heteroskedasticity, the model is,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a $n \times n$ matrix with all zeros except σ_i^2 on the ith diagonal. Maximizing the likelihood function (which is the same as before but with σ_i instead of σ) is equivalent to,

$$\min w_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$$

where $w_i = \frac{1}{\sigma_i^2}$. This criterion is called weighted least squares, as opposed to ordinary least squares, since the w_i acts as a weight.

In matrix notation, we want,

$$\min(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

where $\boldsymbol{W} = diag(w_1, \ldots, w_n) = \boldsymbol{\Sigma}^{-1}$. Treating $\boldsymbol{W}$ as known for the moment,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})] \\ &= \frac{\partial}{\partial \boldsymbol{\beta}}[\boldsymbol{y}^T\boldsymbol{W}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta}] \\ &= -2\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y} + 2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})\boldsymbol{\beta}\end{aligned}$$

Setting this to 0, it can be shown that the WLS estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ is,

$$\boxed{\hat{\boldsymbol{\beta}}_{\boldsymbol{W}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}}$$

Furthermore (see Lecture 18 slide 14 for proof),

$$E[\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}] = \boldsymbol{\beta} \qquad \text{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$$

The weighted least squares estimator is unbiased.

An alternative view of WLS: Let $\boldsymbol{W}^{1/2} = diag(w_1^{1/2}, \ldots, w_n^{1/2})$ where $w_i = \frac{1}{\sigma_i^2}$ as before. We could pre-multiply our model by $\boldsymbol{W}^{1/2}$,

$$\boldsymbol{W}^{1/2}\boldsymbol{y} = \boldsymbol{W}^{1/2}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}^{1/2}\boldsymbol{\epsilon} \implies \boldsymbol{y}_w = \boldsymbol{X}_w\boldsymbol{\beta} + \boldsymbol{\epsilon}_w$$

Then,

$$E[\boldsymbol{\epsilon}_w] = 0 \qquad \text{Var}(\boldsymbol{\epsilon}_w) = \text{Var}(\boldsymbol{W}^{1/2}\boldsymbol{\epsilon}) = \boldsymbol{W}^{1/2}\text{Var}(\boldsymbol{\epsilon})\boldsymbol{W}^{1/2} = \boldsymbol{W}^{1/2}\boldsymbol{\Sigma}\boldsymbol{W}^{1/2} = \boldsymbol{I}$$

So, we could achieve $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ just by OLS of $\boldsymbol{y}_w$ on $\boldsymbol{X}_w$.

Fitting WLS: The structure of $\boldsymbol{W}$ may be known by design in some special cases but in practise, it is usually unknown and we need to plug in values for $w_i = 1/\sigma_i^2$. We could estimate σ_i^2 via e_i^2 by methods such as:

1. Directly: set $\sigma_i^2 = e_i^2$, but this is pretty unstable since we are estimating with just one residual
2. Binning: estimate a single σ_i^2 for a group of observations
3. Model σ_i^2. For example, model the absolute values of the residuals by regressing them on the fitted values: $|e_i| = \alpha_0 + \alpha_1\hat{y}_i + \epsilon'$ and then $\hat{\sigma}_i^2 = |\hat{e}_i|^2$. Or, we could regress the squares of the residuals on the fitted values: $e_i^2 = \alpha_0 + \alpha_1\hat{y}_i + \epsilon'$ and then $\hat{\sigma}_i^2 = \hat{e}_i^2$. Alternatively, we could regress against covariates instead of fitted values.

However, how do we obtain $e_i = y_i - \hat{y}_i$ without first obtaining coefficient estimates to estimate $\hat{y}_i$? We use iteratively reweighted least squares algorithm. The steps for this are,

1. Fit OLS: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ to get fitted values $\hat{y}_i$ and residuals e_i
2. Using the fitted values and residuals, estimate σ_i^2 as described above and set $w_i = 1/\sigma_i^2$
3. Fit WLS: $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$ to get updated fitted values and residuals
4. Repeat steps 2, 3 until $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ converges (stops changing), except using the updated fitted values and residuals obtained in step 3 in step 2

Ultimately, you end up with $\hat{w}_i = 1/e_i^2$.

5.2 Outliers

Outliers are unusual or extreme observations. These can be either of outcomes (aka y-outliers) or covariate values (aka x-outliers). Outliers can come from either data entry error or an unusual (but real) observation, and may or may not have a big impact on results. You do not need to already have a model to analyze outliers. Only the hat matrix is needed.

5.2.1 Detecting x-outliers

Leverage: A data point has high leverage if it has "extreme" covariate values (aka is an x-outlier). Consider a single covariate. Intuitively, outliers are far from the mean (ie large $|x_i - \bar{x}|$). For the multiple dimension case, recall the hat matrix $\boldsymbol{H}$ and $\hat{y} = \boldsymbol{H}\boldsymbol{y}$. Leverage for the ith observation, denoted h_i, is defined as the ith diagonal of $\boldsymbol{H}$.

The intuition behind this is ($\boldsymbol{H}_i$ is the ith row of $\boldsymbol{H}$):

$$\hat{y}_i = \boldsymbol{H}_i\boldsymbol{y} = \begin{bmatrix} h_{i1} & h_{i2} & \cdots & h_{in} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j$$

$\hat{y}_i$ is a weighted average of our outcomes and leverage h_{ii} determines how much y_i contributes to the ith fitted value. The ith diagonal of $\boldsymbol{H}$ is $\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$.

Next, recall that $\text{Var}(e_i) = \sigma^2(1 - h_i)$. If h_i is large (close to 1), then $\text{Var}(e_i)$ is small, $|e_i|$ is small, and $\hat{y}_i$ is close to y_i, since $e_i = y_i - \hat{y}_i$.

From simple linear regression,

$$\begin{aligned}
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\
&= \bar{y} + \hat{\beta}_1(x_i - \bar{x}) \\
&= \sum_{j=1}^{n} \frac{1}{n} y_j + (x_i - \bar{x}) \frac{\sum_{j=1}^{n} y_j(x_i - \bar{x})}{S_{xx}} \\
&= \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right] y_i + \sum_{j \neq i} \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} y_j\right]
\end{aligned}$$

Thus, $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$. This is large when $(x_i - \bar{x})^2$ (aka when x_i is far from $\bar{x}$). That is why a data point has high leverage if it has extreme covariate values. Similarly in multiple linear regression.

A statistical rule of thumb is to consider a point "high leverage" (ie is an x-outlier) if,

$$h_i > 2\bar{h}$$

where,

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_i = \frac{1}{n}\,\text{tr}(\boldsymbol{H}) = \frac{p+1}{n}$$

since $\boldsymbol{H}$ is an idempotent matrix and the trace of an idempotent matrix is equal to its rank. See Lecture 19 slide 11 for a proof. Alternatively, this fact follows from $\text{tr}(AB) = \text{tr}(BA)$ for any matrices A, B and,

$$\text{tr}(\boldsymbol{H}) = \text{tr}\big(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big) = \text{tr}\big(\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\big) = \text{tr}(\boldsymbol{I})$$

In summary, a point has high leverage if,

$$\boxed{h_i > \frac{2(p+1)}{n}}$$

Leverage tells us whether $\hat{y}_i$ is close to y_i. But what if y_i is also an outlier?

5.2.2 Detecting y-outliers

Recall the ordinary residuals $e_i = y_i - \hat{y}_i$ and $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \sim N(0, \sigma^2(\boldsymbol{I} - \boldsymbol{H}))$. Hence, $e_i \sim N(0, \sigma^2(1 - h_i))$; these e_i have different variances, so it is difficult to work with.

Instead, we again use studentized residuals $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$. Recall these r_i have constant variance so should look normally distributed (with variance approximately 1 and mean 0) when plotted (although we estimate $\hat{\sigma}$ so these r_i are really t-distributed). As well, recall for a $N(0, 1)$, approximately 68% are within 1 SD, 95% are within 2 SD and 99.7% are within 3 SD. So, we can look closely at observations with large $|r_i|$ (ie, 3 SD away or more).

Sneaky studentized residuals: However, studentized residuals can be sneaky. Based on its formula, the larger the h_i, the smaller the studentized residuals. As well, the larger the $\hat{\sigma}$, the smaller the studentized residuals. But, y-outliers with large residuals would themselves contribute to a large $\hat{\sigma}$, since,

$$\hat{\sigma}^2 = \frac{1}{n - (p+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n - (p+1)} \sum_{i=1}^{n} e_i^2$$

Thus, a large e_i increases the numerator of r_i but also the denominator; outliers can "hide" from us.

Jackknife/LOO residuals: To combat this problem, we can use leave-one-out (LOO) (aka Jackknife residuals), which are defined to be,

$$e_{i(-i)} = y_i - \hat{y}_{i(-i)}$$

where $\hat{y}_{i(-i)}$ is the fitted value for the ith observation based on fitting the model without y_i (ie remove the ith row of data, fit the model on the remaining data to get $\hat{\boldsymbol{\beta}}_{(-i)}$, use those to compute $\hat{y}_{i(-i)}$). We can then use the studentized jackknife residuals,

$$r_{i(-i)} = \frac{e_{i(-i)}}{s_i}$$

where s_i is the appropriate standard deviation so that $r_{i(-i)}$ has constant variance. Specifically,

$$s_i = \sigma\sqrt{1 + \boldsymbol{x}_i(\boldsymbol{X}_{-i}^T\boldsymbol{X}_{-i})^{-1}\boldsymbol{x}_i^T}$$

where $\boldsymbol{x}_i$ is the ith covariate data and $\boldsymbol{X}_{-i}$ is the design matrix without the ith row. This comes from the standard deviation of $y_{new} - \hat{y}_{new}$ (see prediction interval part of the notes); we are essentially doing prediction here, since the ith outcome was not used to fit the model.

By leaving the ith observation out when fitting the model, we avoid the problem of that residual contributing to $\hat{\sigma}$ and concealing itself. However, a problem with this approach is that we need to fit n different models.

We can further simplify this computation. It can be shown that (where r_i is the usual studentized residual),

$$r_{i(-i)} = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(-i)}(1 - h_i)}} = r_i \left[\frac{n - p - 2}{n - p - 1 - r_i^2}\right]^{1/2}$$

which we can extract from a single model fit. See Lecture 20 slide 13 for more details.

Overall, these LOO studentized residuals are all on the same scale (constant variance) and will not conceal themselves by inflating the $\hat{\sigma}$. A y-outlier is one with a high absolute value of $r_{i(-i)}$.

5.2.3 What to do with outliers

If we believe the outlier is incorrect (ie a data entry error or observing a subject from a different population), we should remove them. However, this should not be the default and it is more useful to examine its impact on our results.

5.3 Influential observations

Recall high leverage observations (x-outliers) are those with a potential to impact our regression. Influential observations are those which strongly impact our regression.

Influence is the impact of an observation on our regression. How can we determine the influence of an observation? Using the same intuition as behind jackknife residuals, we could compare the model fit to the full data, to a model fit to the whole data except for that observation. If the model fit changes dramatically, this observation might be influential.

Some classical ways to quantify influence are: DFFITS, Cook's distance, DFBetas. These each measure slightly different things and it is a good idea to investigate all three.

DFFITS: Define for the ith observation,

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{\sqrt{\hat{\sigma}^2_{(-i)} h_i}}$$

where,

$$\hat{y}_{i(-i)} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(-i)} = \boldsymbol{x}_i^T (\boldsymbol{X}_{(-i)}^T \boldsymbol{X}_{(-i)})^{-1} \boldsymbol{X}_{(-i)}^T \boldsymbol{y}_{(-i)}$$

where $\hat{\boldsymbol{\beta}}_{(-i)}$ are the coefficient estimates based on a fit excluding the ith observation and $\hat{\sigma}^2_{(-i)}$ is the MSE for the model fit to all observations except the ith.

Thus, $DFFITS_i$ is a scaled difference between the fitted value for y_i based on the full data fit and the fitted value for y_i based on all data except the ith observation (ie, as if we had not observed y_i). A large value suggests the fitted value changes substantially.

Furthermore, it can be shown that,

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{\sqrt{\hat{\sigma}^2_{(-i)} h_i}} = r_{i(-i)} \sqrt{\frac{h_i}{1 - h_i}}$$

which is a function of $r_{i(-i)}$ (the studentized jackknife residuals) and h_i (ith diagonal of the hat matrix). Note that $r_{i(-i)}$ incorporates information about the y-outliers and h_i incorporates information about the x-outliers. Another benefit of $DFFITS$ that this formulation shows we do not need to refit the model to compute each $DFFITS$, since the $r_{i(-i)}$ do not need separate model fits.

$DFFITS_i$ can quantify influence because a high h_i indicates leverage (the potential to influence the model fit) but depending on the outcome value encoded in $r_{i(-i)}$, a high leverage point may or may not actually influence the model fit. For example, consider in the simple linear regression case, an x-value which is much larger than the rest but the associated outcome is also much larger such that it falls on a line with the other outcomes. Even though there was potential to influence the model fit, this observation did not actually influence it.

A statistical rule of thumb for a high value of $DFFITS_i$ (influence) is,

$$|DIFFITS_i| > 2\sqrt{\frac{p+1}{n}}$$

Cook's distance: Define for the ith observation,

$$D_i = \frac{(\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(-i)})^T (\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(-i)})}{\hat{\sigma}^2 (p+1)} = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(-i)})^2}{\hat{\sigma}^2 (p+1)}$$

where $\hat{y}_{j(-i)}$ is the fitted value for the jth observation on a model fit with all observations except the ith and $\hat{\boldsymbol{y}}_{(-i)} = \boldsymbol{X} \hat{\boldsymbol{\beta}}_{(-i)}$ is the $n \times 1$ vector containing all these $\hat{y}_{j(-i)}$ for each $j \in [1, n]$.

Intuitively, D_i is the scaled measure of average squared distance between the fitted values with and without y_i (ie, how does the ith observation affect the fitted values?). In contrast, $DFFITS_i$ was the impact of excluding the ith observation on just the ith fitted value.

As well, it can be shown that,

$$D_i = \frac{r_i^2}{p+1}\left[\frac{h_i}{1+h_i}\right]$$

which is a function of r_i (the ith studentized residual) and h_i (leverage), thus incorporating info about both x and y outliers in a similar way as $DFFITS_i$. Again, computing each D_i does not require refitting the model.

A statistical rule of thumb for influential points is to compare them to $F_{p+1,n-p-1}$. A large percentile (ie 50th percentile; points such that `(pf(point, p+1, n-p-1, lower.tail=TRUE) > 0.5)` indicates a large effect on the fit.

DFBetas: $DFFITS_i$ measures the ith observation's impact on its fitted value, Cook's distance measures the ith observation's impact on all the fitted values. Sometimes, what we really care about is estimating $\boldsymbol{\beta}$. $DFBETAS$ measures the ith observation's impact on the coefficient estimates. This is especially useful for measuring associations or inference.

For the ith observation's influence on the kth coefficient β_k,

$$DFBETAS_{k,i} = \frac{\hat{\beta}_k - \hat{\beta}_{k(-i)}}{\sqrt{\hat{\sigma}^2_{(-i)}\boldsymbol{V}_{kk}}}$$

where $\hat{\beta}_{k(-i)}$ is the kth element of $\hat{\boldsymbol{\beta}}_{(-i)}$ (the coefficient estimates when fitting on all the observations except the ith). There are a total of $n \times (p+1)$ different $DFBETAS$.

We divide by $\sqrt{\hat{\sigma}^2_{(-i)}\boldsymbol{V}_{kk}}$ because $\text{Var}(\hat{\beta}_k) = \sigma^2\boldsymbol{V}_{kk}$ (recall $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$). Note that $\boldsymbol{V}$ is computed with all observations.

Large values of $DFBETAS$ indicate a large impact on the estimation of β_k. A statistical rule of thumb for influence is,

$$|DFBETAS_{k,i}| > 2/\sqrt{n}$$

5.3.1 What to do with influential points

Exclude them if you have some reason to suspect they are incorrect (eg data entry mistake). More broadly, it is good practise to report them: how do the results look with and without these points? Are our conclusions substantively different?

5.3.2 Influential vs outliers

There is a very good chance that an influential point is also either an x-outlier or y-outlier or both. In general, we care more about influential points because they actually have meaningful impacts on the results. If an observation is an x or y outlier but is not influential, then it is not making a big impact on the results, so does not matter as much.

6 Other

6.1 Assignments

Facts proven on A1,

1. $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$
2. $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where the first sum is $SS(Total)$, second is $SS(Reg)$, and third is $SS(Res)$
3. $SS(Reg)/SS(Total) = r^2$, where r is the sample correlation between x_i and y_i
4. If $\hat{y}_i = y_i$ for all $i = 1, \ldots, n$, then $r = \pm 1$ (the sample correlation between x_i's and y_i's) and all points lie exactly on the fitted regression line

Facts proven on A2,

1. For multiple linear regression, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ minimizes the variances $\text{Var}(\hat{\beta}_0), \text{Var}(\hat{\beta}_1), \ldots, \text{Var}(\hat{\beta}_p)$ (ie the diagonal elements of $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$) among all linear unbiased estimators of $\boldsymbol{\beta}$ (ie among all $\hat{\boldsymbol{\beta}}^* = \boldsymbol{A}\boldsymbol{y}$ for constant matrix $\boldsymbol{A}$), thus giving us the most precise CIs using these β_js.

6.2 Miscellaneous Things

1. CI and HT relationship: recall from Stat 231 that if the same pivotal quantity is used to construct a CI for θ and a test of the hypothesis $H_0 : \theta = \theta_0$, then the parameter value $\theta = \theta_0$ is inside a $100q\%$ CI for θ iff the p value for testing $H_0 : \theta = \theta_0$ is greater than $1 - q$. See the Stat 231 book page 191 for proof.
2. Suppose we construct a single 95% CI. By definition of CI, we expect 95% of CIs constructed using the same procedure to contain the true parameter value across repeated samples, but there is no such assumption for a single CI. This is inherent to the classical/frequentist paradigm of statistics (frequentist meaning that probabilities are interpreted as relative frequencies across repeated trials).

   Alternatively, in the "Bayesian" paradigm (comes from Bayes' theorem), one could construct what are known as "credible intervals", which have an interpretation that is closer to what you want. That is, a credible interval is one within which an unobserved parameter value falls with a particular probability. These Bayesian intervals treat their bounds as fixed and the estimated parameter as a random variable, whereas frequentist CIs treat their bounds as random variables and the parameter as a fixed value.
3. See Lecture 7 slide 15 for how block matrix multiplication works.
4. How hypothesis testing works: consider a null hypothesis and an alternative hypothesis. Compute the probability that, assuming the null hypothesis is true, we observe evidence at least as extreme as the evidence we did observe. To do this, we choose some test statistic which comes from a known distribution (eg t, F distributions) and which produces smaller probabilities the more the data tends away from the null. If this probability is less than some threshold α, then reject the null. Otherwise, say that we do not have evidence to reject the null (but importantly, cannot conclude that the null is true).
5. About MSPE: MSPE is in some ways arbitrary, because it is unknown what is good/bad relative to a certain scientific problem/certain set of data which may or may not explain the data. One strategy would be to compare it to the natural variability in the outcomes. Ie if you knew nothing but the sample mean, how bad would your predictions be. Equivalently, this is like thinking about MSPE relative to the variance of the outcomes. Magnitude of the outcomes could also be considered (MSPE of 0.05 is very bad if outcomes are all around 0.01 but could be good if outcomes are around 100), but the variability of the outcomes is probably more important than the average. Because all your outcomes

could be around 1000, but if they only vary between 999 to 1001, a MSPE of 100 would be bad. But if the outcomes centre around 1 but range from -1000 to 1000, MSPE of 100 might not be so bad.

6.3 R

6.3.1 Summary

The `summary` function in R is commonly used. For a fitted linear regression model `M <- lm(y~x1+x2+x3+x4)`, `summary(M)` might look like,

```
Residuals:
Min      1Q  Median      3Q     Max
-2.2469 -0.7770 -0.3600  0.5899  5.3159

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.2944     4.0169   1.567    0.129
x1              -0.1440     0.7742  -0.186    0.854
x2               1.1802     0.8749   1.349    0.189
x3              -1.0239     0.9281  -1.103    0.280
x4              -0.4191     0.3540  -1.184    0.247

Residual standard error: 1.642 on 26 degrees of freedom
Multiple R-squared:  0.2726,Adjusted R-squared:  0.1606
F-statistic: 2.435 on 4 and 26 DF,  p-value: 0.07256
```

The numbers under the `Estimate` column are the $\hat{\boldsymbol{\beta}}$ (eg, $\hat{\beta}_0 = 6.2944$), the `Std.Error` column is the standard errors of the corresponding estimators (eg $SE(\hat{\beta}_0) = 4.0169$), the `t value` and `Pr(>|t|)` columns are the test statistics and p-values for testing the hypothesis $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$ (two-sided test).

Following this table, the residual standard error is $\hat{\sigma}$ (ie $\hat{\sigma} = 1.642$), where $\hat{\sigma}$ is the MLE estimate of σ and also, $\hat{\sigma}^2 = SSRes/(n-p-1)$. The df beside that (26 in this case) is $n-p-1$. Since there are four covariates ($p = 4$), from this `summary` output, we can determine that $n = 26 + 4 + 1 = 31$.

Then, $R^2 = 0.2726$ is the coefficient of determination and $R^2_{adj} = 0.1606$ is the adjusted R-squared.

Lastly, the F-statistic of 2.435 is for testing $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ against H_1 that at least one of these coefficients it not zero (this tests whether any variability in the outcome is explained by the covariates; are any covariates associated with the outcome). In this F-test, the numerator and denominator dfs are 4 and 26 respectively (ie the test statistic is from a $F_{4,26}$ distribution. The p-value of 0.07256 is also for this test.

6.3.2 Code

```
data <- read.csv("multivariatedata.csv")
pairs(data) # gives a cool visual comparison

Model <- lm(Y ~ X + Y + Z)
X <- model.matrix(M) # get the design matrix X

# See the realestate.csv plot in week5(lec9+10).R for plotting linear models with
# categorical covariates, where data is color-coded and of different shape.
# Also, how to fit a model for: different intercept among categories but same slope,
# different intercept and different slope, different slope and same intercepts.
```

```
# Fit a quadratic model y_i = beta0 + beta1x_i + beta2(x_i)^2 + error_i
Model <- lm(Y ~ X + I(X^2), data = mydata)

# For a categorical covariate x, constructs indicator covariates corresponding to all
# the categories of x except one (which is the referent group)
factor(x)

# See the categories for a categorical covariate x
levels(x)

# Fits a linear regression model, regressing on a categorical covariate x, with indicators
Model <- lm(Y ~ factor(x))

# Generates a table detailing the number of observations in each category
# for cov1 cross cov2. See Lecture 14C_Practice for more R code.
table(mydata[c("cov1", "cov2")])

# Create a design matrix of 30 observations and 4 covariates, where each data point comes
# from a standard normal.
n <- 30
p <- 4
X <- data.frame(matrix(rnorm(n*p),ncol=p))
# Generate 30 outcomes which are also standard normal
y <- rnorm(n)
# Create the data frame which we will give to lm, remember that even though the outcomes are
# not in the design matrix, the data frame given to lm needs to contain it
dat <- X
dat$y <- y
# Doing y ~ . fits outcome "y" on all other covariates in the dataframe
Model <- lm(y ~ .,data=dat)
```

6.4 Bayesian vs Frequentist

This topic was mentioned in passing but is not strictly related to Stat 331.

Suppose a fair coin is flipped and person A observes the outcome but person B does not. Then consider, what is the probability that the coins has landed heads? The Bayeseian perspective is that the probability is either 0 or 1 to person A who has observed the outcome but is 0.5 to person B who has not. The Frequentist perspective is that since the coin has already landed, there is no more randomness associated with it, so even though person B has not observed the outcome, the probability is still either 0 or 1, depending on whether the coin has landed heads. Frequentists should then have a "default" action, such as saying the probability is 1. Then if the coin is revealed to be heads, they would consider their action wrong.

Bayesians understand that the coin has landed on either heads or tails, but are more interested in their own perspective (their own opinion), in which the coin is heads with probability 0.5. However, Frequentists care more about the true answer, which is already fixed, and if this truth is heads, it is 1, and if it is tails, it is 0. Frequentists orient their analysis on the truth, not on their evolving opinion/perspective.

When the outcome is revealed to person B (and say it is tails), both Bayesian and Frequentist perspectives will agree that the probability is 0. The difference is when the coin has landed but the outcome has not been observed.

A Bayesian can never be wrong because they always evaluate from their own perspective. They take a starting

opinion and evolve their opinion as they see more data. There is no notion that this updating method will give the "wrong" answer, since there is no wrong answer to Bayesians.

On the other hand, Frequentists would say: if this procedure were repeated many times, they would guess correctly some of the time and incorrectly in the rest. If guessing heads, the Frequenist would be right 50% of the time, assuming the coin is flipped enough times. Frequentists have a notion of being "correct" and whether they guessed the truth properly.

Overall, **statistics is the discipline of changing your mind under uncertainty**. However, would you like to change your mind about your opinion/prior based on observed data (Bayesian) or change your mind about your methods of guessing and measurement (Frequentist)?

Summary: A Frequenist would say that the parameter is not a random variable but a Bayesian would say that it is. It only makes sense to discuss your method's ability to deliver the right answer under the former, whereas in the latter, there is no notion of right or wrong.

Frequentists are more concerned with confidence intervals, p-values, power, and significance, whereas Bayesians will consider credible intervals, priors, and posteriors. To Bayesians, there is no such thing as statistically significant or rejecting the null, only "more likely" or "less likely" from their perspective.

However, typically people do not "pick a side". Whether to take the Bayesian or Frequentist side depends on whether the situation calls for choosing between actions/methods or forming an evidence-based opinion.