STAT 34300 - Review

Solutions

Exercise 1: The moment assumptions of linear regression are?

Solution 1:

1. $\mathbb{E}(Y|X) = X\beta$ or $Y_i = X_i\beta + \epsilon_i$ where $\mathbb{E}(\epsilon_i) = 0$.
2. $\text{Var}(\epsilon_i) = \sigma^2$ for all i, so variance does not depend on X.
3. All the ϵ_i's are all independent from one another, or at least uncorrelated.

Exercise 2: The normality assumption(s) of linear regression is/are?

Solution 2: $Y \sim N(X\beta, \sigma^2 I)$, which implies all the moment assumptions.

Exercise 3: What is the RSS? TSS?

Solution 3:

1. $\text{RSS} = ||Y - X\beta||_2^2 = \sum_{i=1}^n (Y_i - X_i\beta)^2$.
2. $\text{TSS} = \sum_{i=1}^n (Y_i - \overline{Y})^2$.

Exercise 4: In 1-dimensional regression, what is $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$?

Solution 4:

1. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$,
2. $\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \text{Corr}(X,Y) \cdot \frac{\text{STD}(Y)}{\text{STD}(X)}$.
3. $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$.

Exercise 5: What do the moment / normality assumptions tell you about the estimators in 1-dimensional regression?

Solution 5:
Under the moment assumptions:

1. $\mathbb{E}\hat{\beta}_0 = \beta_0, \mathbb{E}\hat{\beta}_1 = \beta_1, \mathbb{E}\hat{\sigma}^2 = \sigma^2$.
2. $\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_i (X_i - \bar{X})^2} \right)$.
3. $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}$
4. $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \cdot \frac{\bar{X}}{\sum_i (X_i - \bar{X})^2}$.
5. $\hat{\beta} \perp \hat{\sigma}^2$.

Under the normality assumptions:

1. $\hat{\beta} \sim N\left(\beta, \sigma^2 \begin{bmatrix} \frac{1}{n} + \frac{\bar{X}^2}{\sum_i (X_i - \bar{X})^2} & -\frac{\bar{X}}{\sum_i (X_i - \bar{X})^2} \\ -\frac{\bar{X}}{\sum_i (X_i - \bar{X})^2} & \frac{1}{\sum_i (X_i - \bar{X})^2} \end{bmatrix} \right)$.
2. $(n-2) \cdot \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$.
3. $\hat{\beta} \perp \hat{\sigma}^2$.

Exercise 6: (1-dimensional regression) Conduct a z-test and t-test for $\hat{\beta}_1$.

Solution 6:
If the variance is known, then

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}}} \sim N(0,1).$$

Under the null hypothesis ($\beta_1 = 0$), the $1 - \alpha$ confidence interval is

$$\hat{\beta}_1 \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}}.$$

If the variance is unknown, then

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}}} \sim N(0,1) \text{ and } (n-2) \cdot \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.$$

This means that

$$\frac{\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\sigma^2}{\sum_i(X_i-\bar{X})^2}}}}{\sqrt{\frac{(n-2)\cdot\frac{\hat{\sigma}^2}{\sigma^2}}{n-2}}} = \frac{\hat{\beta}_1-\beta_1}{\frac{\hat{\sigma}}{\sqrt{\sum_i(X_i-\bar{X})^2}}} = \frac{\hat{\beta}_1-\beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}.$$

Under the null hypothesis ($\beta_1 = 0$), the $1-\alpha$ confidence interval is

$$\hat{\beta}_1 \pm t_{1-\alpha/2} \cdot SE(\hat{\beta}_1).$$

Exercise 7: (1-dimensional regression) Construct a confidence interval for the mean Y value for individuals with $X = x$.

Solution 7:
The mean Y value with $X = x$ is estimated to be $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$. Under the normality assumptions, we have

$$\hat{\mu} \sim N\left(\beta_0 + \beta_1 x, \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{X})^2}{\sum_i (X_i-\bar{X})^2}\right)\right).$$

Exercise 8: (1-dimensional regression) Construct a prediction interval for a new y at a particular x value.

Solution 8:
A new y value at x has the following form under the model: $y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. This means

$$\begin{aligned} y - \hat{y} &= (\beta_0 + \beta_1 x + \epsilon) - (\hat{\beta}_0 + \hat{\beta}_1 x) \\ &\sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{\sum_i (X_i-\bar{X})^2}\right)\right). \end{aligned}$$

Exercise 9: In p-dimensional regression, what is $\hat{\beta}, \hat{\sigma}^2$?

Solution 9:

1. $\hat{\beta} = (X^T X)^{-1} X^T Y$.
2. $\hat{\sigma}^2 = \frac{||Y - X\hat{\beta}||_2^2}{n-p}$.

Exercise 10: What do the normality assumptions tell you about the estimators in p-dimensional regression? Justify your findings.

Solution 10:

1. $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$
2. $(n-p) \cdot \hat{\sigma}^2 / \sigma^2 \sim \chi_{n-p}$
3. $\hat{\beta} \perp \hat{\sigma}^2$

Exercise 11: In p-dimensional regression, what is the interpretation of β_j?

Solution 11:
All things held constant, increasing the jth covariate by one unit will lead to a β_j increase in the Y value.

Exercise 12: Explain what might happen if covariates are highly correlated. Construct an example of

- If you fit a linear model of Y on covariate X_1 only, then the fitted slope is generally negative, but
- if you fit a linear model of Y on both covariates X_1 and X_2, then the coefficient $\hat{\beta}_1$ on X_1 is generally positive.

Solution 12:
If covariates are highly correlated, then one covariate could stand in as a proxy for another.

Exercise 13: Show that $H = X(X^T X)^{-1} X^T$ is a projection matrix, and interpret the projection. Interpret $I - H$.

Solution 13:

$H^2 = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = H$ and $H^T = H$.

The projection H projects the Y vector onto the subspace spanned by X. On the other hand, $I - H$ projects onto the orthogonal space of X; the residual vector is the projection onto this space.

Exercise 14: Give the definition of χ_p^2 distribution.

Solution 14:
If $X_1, \cdots, X_p \sim_{\text{i.i.d.}} N(0,1)$, then $X_1^2 + \cdots + X_p^2 \sim \chi_p^2$.

Exercise 15: Give the definition of t_k distribution.

Solution 15:
If $Z \sim N(0,1)$ and $V \sim \chi^2_k$, then $\frac{Z}{\sqrt{V/k}} \sim t_k$.

Exercise 16: True or False:

1. $Z \sim N(0, I_m)$ if and only if $Z_1, \cdots, Z_m \sim_{\text{i.i.d.}} N(0,1)$
2. $Z \sim N(\mu, \Sigma)$ if and only if $v^T Z \sim N(v^T\mu, v^T\Sigma v)$ for any $v \in \mathbb{R}^n$.
3. If $Z \sim N(\mu, \Sigma)$ and A, b are fixed, then $AZ + b \sim N(A\mu + b, A\Sigma Z^T)$.

Solution 16:
All are true.

Exercise 17: Prove that if $Z \sim N(0, I_m)$, $A \in \mathbb{R}^{k\times m}, B \in \mathbb{R}^{l\times m}$ with $AB^T = 0$, then $AZ \perp BZ$.

Solution 17:
Write the block matrix $[A^T B^T]^T$. We have

$$\begin{pmatrix} A \\ B \end{pmatrix} Z \sim N\left(0, \begin{bmatrix} AA^T & AB^T \\ BA^T & BB^T \end{bmatrix}\right) = N\left(0, \begin{bmatrix} AA^T & 0 \\ 0 & BB^T \end{bmatrix}\right)$$

Exercise 18: (p-dimensional regression) Conduct a z-test and t-test for $v^T\hat{\beta}$ for some fixed vector v.

Solution 18:
Because $\hat{\beta} \sim N(v^T\beta, \sigma^2(X^TX)^{-1})$, we have $v^T\beta \sim N(v^T\hat{\beta}, \sigma^2 v^T(X^TX)^{-1}v)$. This gives the confidence interval

$$v^T\hat{\beta} \pm z_{1-\alpha/2} \cdot \sigma\sqrt{v^T(X^TX)^{-1}v}.$$

For the t-test, since $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$, we have

$$\frac{v^T\hat{\beta} - v^T\beta}{\hat{\sigma}\sqrt{v^T(X^TX)^{-1}v}} = \frac{v^T\hat{\beta} - v^T\beta}{SE(v^T\hat{\beta})} \sim t_{n-p}.$$

This gives the confidence interval

$$v^T\hat{\beta} \pm t_{1-\alpha/2} \cdot \hat{\sigma}\sqrt{v^T(X^TX)^{-1}v}.$$

Exercise 19: (p-dimensional regression) Construct a confidence interval for the mean Y value for individuals with $X = x$.

Solution 19:
Use the above, but use $v = x$.

Exercise 20: (p-dimensional regression) Construct a prediction interval for a new y at a particular x value.

Solution 20:
From the model, $y = x^T\beta + \epsilon$, where $\epsilon \sim N(0,1)$. This means

$$\begin{aligned} y - \hat{y} &= x^T\beta + \epsilon - x^t\hat{\beta} \\ &\sim N(0, \sigma^2(1 + x^T(X^TX)^{-1}x)) \perp \hat{\sigma}^2. \end{aligned}$$

The prediction interval is given by $\hat{y} \pm t_{1-\alpha/2} \cdot \hat{\sigma}\sqrt{x^T(X^TX)^{-1}x}$.

Exercise 21: Explain R^2 and adjusted R^2. What are the expressions for both, and why might adjusted R^2 be better?

Solution 21:
R-squared is how much the total variance of Y is explained by the linear model. Intuitively, if the model is a perfect linear fit to the data, then the R-squared should be close to one because the model perfectly explains the variance of Y.

Adjusted R-squared is how much the total variance of Y is explained by the linear model, adjusted for the number of parameters in the model. Intuitively, as the number of parameters goes to ∞, the vanilla R-squared should approach one because more parameters are always better from a pure fitting perspective.

$$\begin{aligned} R^2 &= \frac{||\hat{Y} - \bar{Y}||^2}{||Y - \bar{Y}||^2} = 1 - \frac{RSS}{TSS} \\ R^2_{\text{adj}} &= 1 - \frac{\frac{1}{n-p}RSS}{\frac{1}{n-1}TSS} = 1 - \frac{\hat{\sigma}^2}{\hat{\text{Var}}(Y)}. \end{aligned}$$

Exercise 22: What is the F distribution? What is the F-test?

Solution 22:
If $X_1 \sim \chi^2_a$ and $X_2 \sim \chi^2_b$ and $X_1 \perp X_2$, then $\frac{X_1/a}{X_2/b} \sim F_{a,b}$.

The F-test tests the significances of a subset of covariates. For example, we might want to test with the null hypothesis of $\beta_{j_1} = \cdots = \beta_{j_k} = 0$. Under the null hypothesis,

$$\frac{(RSS_{\text{reduced}} - RSS_{\text{full}})/k}{RSS_{\text{full}}/n-p} \sim F_{k,n-p}.$$

We should reject the null if the test statistic is large, since this means that the RSS of the reduced model is substantially larger.

Exercise 23: Give conceptual definitions for (1) outlier, (2) high leverage point, (3) influential point. How do you detect outliers? How do you calculate leverage? How can make decisions about whether or not a point is influential?

Solution 23:
An outlier is a point that strays away from the trend in the Y space - it is a comparative definition that depends on the model that we use. A high leverage point is a point that lies far away in the X space - it is a point that is far from the mean of X. An influential point is a point that makes a large difference in the model.

Outliers can be detected by using studentized residuals and the Bonferroni correction. Let $\hat{\beta}_{(i)}$ be the LSE fitted without point i. If $X_{(i)}$ is the design matrix with the ith row removed, then

$$\hat{\beta}_{(i)} \sim N(\beta, \sigma^2(X_{(i)}^T X_{(i)})^{-1}).$$

Then

$$Y_i - X_i^T \hat{\beta}_{(i)} \sim N(0, \sigma^2(1 + X_i^T (X_{(i)}^T X_{(i)})^{-1})).$$

The test statistic is

$$\frac{Y_i - X_i^T}{\hat{\sigma}\sqrt{1 + X_i^T (X_{(i)}^T X_{(i)})^{-1}}} \sim t_{n-p-1}.$$

Do this for all i, and test against the Bonferroni correction α/n.

Leverage is calculated with the following. Let H be the projection matrix. Then $H_i i$ is the leverage score for point (X_i, Y_i). This is because

$$\begin{aligned}
\text{Var}(\hat{\epsilon}) &= \text{Var}((I-H)Y) \\
&= \text{Var}((I-H)(X\beta + \epsilon)) \\
&= \text{Var}((I-H)\epsilon) \\
&= \sigma^2(I-H),
\end{aligned}$$

which means that if $H_i i$ is large, then $\hat{\epsilon}_i$ is small, which means that the model is overfitting towards the ith point.

To detect influential points, try fitting the model with and without an outlier/high leverage point. If the line fit changes substantially, then we can have reasonable confidence that we have an influential point. Problems may arise if there are multiple suspect influential points that are near each other. We can report our results with and without influential points, use bootstrap, or use robust regression methods.

Exercise 24: What are (1) bootstrapping the sample and (2) boostrapping the residual? How do (1) and (2) behave with relation to

1. Heavy tailed / skewed noise distribution?
2. Nonconstant variance?
3. Nonlinear trends?
4. Leverage points / influential points?

Solution 24:
Bootstrapping the sample is sampling n points from the dataset *with* replacement.

Bootstrapping the residual is

1. Fitting a model to the data to get $Y = X\hat{\beta}$ and residuals $E = \{\hat{\epsilon}_1, \cdots, \hat{\epsilon}_n\}$.
2. Creating n data points of the form $Y_k = X_k^T \hat{\beta} + \hat{\epsilon}_k$, where the error term is sampled without replacement from E.

With heavy tailed / skewed noise distribution, both bootstrapping methods should suffice.

With nonconstant variance, bootstrapping the sample is better than bootstrapping the residuals. For example, if the variance gets larger as the X value grows, bootstrapping the residuals may wrongly assign a small residual term for a large X point.

With nonlinear trends, bootstrapping the sample is better because bootstrapping the residual assumes a linear model a priori.

For leverage points / outlier points, bootstrapping the sample is better because we can get a better estimate of the standard errors of the coefficient estimates. Bootstrapping the residuals may get rid of the exact outlier-ness / high leverage-ness of the point altogether, whereas bootstrapping the sample will create samples that have or do not

have the problem point.

Exercise 25: Explain how bootstrap might be used for standard error estimation of $\hat{\beta}_j$.

Solution 25:
Bootstrap the sample k times, collect the coefficient estimates from each generated sample, and collect the standard errors of these coefficient estimates.

Exercise 26: Why do we want our model to not be too large? Why do we want our model to not be too small?

Solution 26:
The bias-variance tradeoff. Smaller models tend to have higher bias, whereas larger models tend have higher variance.

(Smaller means higher bias.) Suppose X_{-j} is the design matrix with the covariate X_j removed. Suppose β_{-j} is the coefficients without the jth coefficient. Then

$$\begin{aligned}\mathbb{E}(\hat{\beta}^{-j}) &= (X_{-j}^T X_{-j})^{-1} X_{-j}^T \mathbb{E}(Y) \\ &= (X_{-j}^T X_{-j})^{-1} X_{-j}^T X\beta \\ &= (X_{-j}^T X_{-j})^{-1} X_{-j}^T X_{-j}\beta_{-j} \\ &+ (X_{-j}^T X_{-j})^{-1} X_{-j}^T X_j \beta_j \\ &= \beta_j + (X_{-j}^T X_{-j})^{-1} X_{-j}^T X_j \beta_j \neq \beta_j.\end{aligned}$$

(Larger means higher variance.) Suppose true $\beta_j = 0$. Suppose $\hat{\beta}^{-j}$ are the coefficient estimates when regressing on the design matrix with the jth covariate removed. Then it is generally true that

$$\begin{aligned}(X^T X)^{-1} &= \begin{bmatrix} X_{-j}^T X_{-j} & X_{-j}^T X_j \\ X_j^T X_{-j} & X_j^T X_j \end{bmatrix} \\ &\succeq \begin{bmatrix} (X_{-j}^T X_{-j})^{-1} & 0 \\ 0 & 0 \end{bmatrix}.\end{aligned}$$

This means for some $k \neq j$, we have

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_k) &= e_k^T (X^T X)^{-1} e_k \\ &\geq e_k^T (X_{-j}^T X_{-j})^{-1} e_k \\ &= \mathrm{Var}(\hat{\beta}_k^{-j}).\end{aligned}$$

Exercise 27: What is collinearity? Give a few problems that arise from collinearity.

Solution 27:
Collinearity is when a collection of covariates are highly correlated. Equivalently, some covariate (or covariates) is expressable as a linear combination of some subset of other covariates.

Under high collinearity, coefficient estimates may be highly unstable. Suppose $X_1 \approx X_2$. If you were to graph X_1 vs. X_2, the points would be scattered around the identity line. This means that any estimate of $\beta_1 + \beta_2$ is good, but any estimate of $\beta_1 - \beta_2$ is extremely noisy. Hence the eigenvalues of $(X^T X)$ in directions of $(1, 1)$ and $(1, -1)$ will be large and small, respectively. Accordingly, then eigenvalues of $(X^T X)^{-1}$ in the directions of $(1, 1)$ and $(1, -1)$ will be small and large, respectively, which means that standard errors for $\hat{\beta}_0 + \hat{\beta}_1$ and $\hat{\beta}_0 - \hat{\beta}_1$ will be small and large, respectively.

Exercise 28: How does collinearity affect prediction at a new x?

Solution 28:
If $X_j \approx X_k$ in the training data and $x_j \approx x_k$ at the new test point, the prediction will have low variance (look at the eigenvalue discussion above).
If $X_j \approx X_k$ in the training data and $x_j \not\approx x_k$ at the new test point, then the prediction from the full model with both covariates will have high variance.
If $X_j \approx X_k$ in the training data and $x_j \not\approx x_k$ at the new test point, then the prediction from the reduced model with only one of X_j, X_k will have low variance (but this is unreliable).

Exercise 29: Give three methods to measure collinearity.

Solution 29:

1. **Condition number.** $\kappa_2(X) = \sqrt{\lambda_1(X^T X)/\lambda_i(X^T X)} = \sigma_1(X)/\sigma_i(X)$. The condition number expresses the relative proportional difference between the maximum principal component with the minimum principal component. The best possible value is if X is orthogonal i.e. X is perfectly conditioned.
2. **R-squared.** Regress X_j onto X_{-j}. Cal-

culate the R-squared to get an estimate of how correlated X_j is with some linear combination of X_{-j}.

3. **Variance inflation factor (VIF).** $VIF = \frac{((X^TX)^{-1})_{jj}}{((X_r^TX_r)^{-1})_{22}} = \frac{1}{1-R^2}$, where R^2 is the R-squared value from regressing X_j onto X_{-j}. The VIF is the difference in variance of $\hat{\beta}_j$ in the full model vs in the model regressing Y onto X_j only.

Exercise 30: Explain forward selection, backward selection.

Solution 30:

Forward selection. Start with an intercept only model, and push coefficients into the model one-by-one based on whichever one reduces the RSS the most.

Backward selection. Start with the full model, and pop coefficients out of the model one-by-one based on whichever one has the highest p-value (or increases the RSS the least).

Exercise 31: What is the Bayesian Information Criterion (BIC)? Explain how the BIC behaves.

Solution 31:
For a subset of covariate indices $S \subset \{1, \cdots, p\}$,

$$BIC(S) = n \log(RSS(\text{model S})/n) + |S| \log(n).$$

The BIC works like a see-saw. As $|S| \to \infty$, the RSS will undoubtedly decrease, but the second term will increase to infinity. The second term therefore acts as a penalty on the size of the model.

Exercise 32: Give an example of selective inference + multiple testing.

Solution 32:

Multiple testing. "If you give something enough chances to be true, it will be true." Suppose we have one thousand parameters. Then using forward selection and BIC for model selection suffers greatly from multiple testing issues, because some of the parameters will be significant just by random chance.

Selective inference. Suppose we choose to cull some "insignificant points" before running our regression - this is selective inference because our design matrix is dependent on us having looked at the data.

Exercise 33: Why might we do robust regression? What are LAD, least trimmed squares, and Huber regression?

Solution 33:
We might want to do robust regression for robustness against outliers / influential points.

Least absolute deviation (LAD). LAD is solving the optimization problem

$$\hat{\beta}_{LAD} = \arg\min_{\beta} \sum_{i=1}^{n} |Y_i - X_i^T\beta|.$$

Normal LS is optimizing for the mean, whereas LAD is optimizing for the median (which is more robust to outliers).

Huber regression. The Huber loss is defined as

$$l_c(t) = \begin{cases} t^2/2, & |t| \le c \\ c|t| - c^2/2, & |t| \ge c \end{cases}$$

Huber regression is solving the optimization problem

$$\hat{\beta}_H = \arg\min_{\beta} \sum_{i=1}^{n} l_c(Y_i - X_i^T\beta).$$

Least trimmed squares (LTS). LTS is solving the optimization problem

$$\hat{\beta}_{LTS} = \arg\min_{\beta} \min_{S \subset 1, \cdots, n, |S|=q} \sum_{i=S} (Y_i - X_i^T\beta)^2.$$

An iterative method for LTS is running a regression, stripping away the points with highest residuals, refitting again, stripping away the high residual points, refitting, etc. until stabilization.

Exercise 34: What is the optimization problem for weighted least squares? Write it in both sum and vector form.

Solution 34:
Let $W = \text{diag}\{w_1, \cdots w_n\}$. The optimization problem is then

$$\begin{aligned}\hat{\beta}_{WLS} &= \arg\min_{\beta} \sum_{i=1}^{n} w_i (Y_i - X_i^T \beta)^2 \\ &= \arg\min_{\beta} (Y - X\beta)^T W (Y - X\beta).\end{aligned}$$

Exercise 35: Show that if σ_i^2 are known and $W = \text{diag}\{\sigma_i^{-2}\}_{i=1,\ldots,n}$, then weighted least squares is equivalent to doing normal regression on the dataset $(X_i/\sigma_i, Y_i/\sigma_i)$.

Solution 35:
Call $(\tilde{X}_i, \tilde{Y}_i) = (X_i/\sigma_i, Y_i/\sigma_i)$. Then $\tilde{Y} - \tilde{X}\beta = W^{1/2}(Y - X\beta)$. Hence,

$$\begin{aligned}\hat{\beta}_{WLS} &= \arg\min_{\beta} (Y - X\beta)^T W (Y - X\beta) \\ &= \arg\min_{\beta} (W^{1/2}(Y - X\beta))^T (W^{1/2}(Y - X\beta)) \\ &= \arg\min_{\beta} (\tilde{Y} - \tilde{X}\beta)^T (\tilde{Y} - \tilde{X}\beta).\end{aligned}$$

Exercise 36: What is the estimated regression coefficient for weighted least squares? Derive some inference properties (under normality assumptions).

Solution 36:
Call $(\tilde{X}_i, \tilde{Y}_i) = (X_i/\sigma_i, Y_i/\sigma_i)$. Then $\tilde{Y} - \tilde{X}\beta = W^{1/2}(Y - X\beta)$. Using the normal equation, we have

$$\begin{aligned}\hat{\beta}_{WLS} &= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y} \\ &= (X^T W X)^{-1} X^T W Y.\end{aligned}$$

Under the normality assumptions, we assume that $Y \sim N(X\beta, \text{diag}\{\sigma_1^2, \cdots, w_n^2\})$. If $W = \text{diag}\{\sigma_1^{-2}, \cdots, \sigma_n^{-2}\}$, then we have

$$\hat{\beta}_{WLS} \sim N\left(\beta, (X^T W X)^{-1}\right).$$

Exercise 37: True or False: $\hat{\beta}_{WLS}$ is unchanged if we rescale w_i (or σ_i) by a constant for all i. Justify your answer.

Solution 37:
True. Suppose $Y \sim N(\beta, \text{diag}\{\sigma_1^2, \cdots, \sigma_n^2\})$. Take $W = \text{diag}_{i=1,\cdots,n}\{\sigma_i^{-2}\}$ and $W_c = \text{diag}_{i=1,\cdots,n}\{c\sigma_i^{-2}\}$ for some constant c. We have

$$\begin{aligned}\hat{\beta}_{WLS} &= (X^T W X)^{-1} X^T W Y \\ &= c^{-1} c (X^T W X)^{-1} X^T W Y \\ &= (X^T W_c X)^{-1} X^T W_c Y.\end{aligned}$$

Exercise 38: Show that if $Y \sim N(X\beta, \Sigma)$, where $\Sigma = \text{diag}\{\sigma_1^2, \cdots, \sigma_n^2\}$, then weighted least squares with weights $w_i \propto 1/\sigma_i$ is the BLUE of β.

Solution 38:
We prove a weaker version. Suppose we want to estimate $\alpha = a^T\beta$ for some fixed vector a.

- (LE) We want an estimator of the form $\hat{\alpha} = c_0 + c^T Y$ for some c_0, c.
- (U) We want our estimator to have the property that

  $$\mathbb{E}(\hat{\alpha}) = c_0 + c^T \mathbb{E}(Y) = c_0 + c^T X\beta = a^T\beta.$$

  This implies that $c_0 = 0$ and $X^T c = a$.
- (B) We want to minimize $\text{Var}(\hat{\alpha}) = \text{Var}(c^T Y) = c^T \Sigma c$. Our optimization problem is then

  $$\arg\min_{c} \{c^T \Sigma c : X^T c = a\}.$$

  It can be shown through taking projections onto the subspace spanned by X that the best optimizer is $\Sigma^{1/2} c = \tilde{X}(\tilde{X}^T\tilde{X})^{-1} a$, where $\tilde{X} = \Sigma^{-1/2} X$. It can then be shown that

  $$\hat{\alpha} = a^T \hat{\beta}_{WLS}.$$

Exercise 39: What is an M-estimator? Show how you would use iteratively reweighted least squares to find an M-estimator. Describe IRLS for LAD.

Solution 39:
The M-estimator is an estimator that minimizes the average error:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \text{Loss}((X_i, Y_i), \beta)$$

.

The objective of IRLS is to find the M-estimator.

Let $f(\beta) = \sum_{i=1}^n l(Y_i - X_i^T\beta)$. Then

$$\nabla_\beta f = -\sum_{i=1}^n X_i l'(Y_i - X_i^T\beta).$$

Compared to weighted least squares loss function $f_w(\beta) = \sum_{i=1}^n w_i \cdot (Y_i - X_i^T\beta)$, we have

$$\nabla_\beta f_w = \sum_{i=1}^n -w_i \cdot X_i(Y_i - X_i^T\beta).$$

If we knew the optimal β, then we would have weights $w_i = \frac{l'(Y_i - X_i^T\beta)}{Y_i - X_i^T\beta}$.

1. Set weights $w_i = 1$.
2. Fit weighted least squares to get $\hat{\beta}_w$.
3. Re-set weights $w_i = \frac{l'(Y_i - X_i^T\beta)}{Y_i - X_i^T\beta}$.
4. Iterate until convergence.

Exercise 40: What is the Box-Cox transformation? Describe what happens when the parameter(s) go to zero.

Solution 40:
The Box-Cox transformation is defined for positive responses - for positive or negative λ, the transformation is defined as

$$Y_i^\lambda = \beta_0 + \beta_1 X_i + \epsilon_i$$

As $\lambda \to 0$, the transformation goes to $\log(Y_i)$.

Exercise 41: What is the shifted log transformation?

Solution 41:
The shifted log transformation for a positive or lower bounded response is $\log(\alpha + Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i$.

Exercise 42: What is the ridge regression optimization problem? Derive the ridge estimator and its inference properties under the normality assumptions.

Solution 42:

For a regularization parameter λ,

$$\begin{aligned}\hat{\beta} &= \arg\min_\beta\{||Y - X\beta||_2^2 + \lambda||\beta||_2^2\} \\ &= \arg\min\{||Y - X\beta||_2^2 \mid ||\beta||_2^2 \le c\} \text{ for some } c \\ &= (X^TX + \lambda I)^{-1}X^TY.\end{aligned}$$

Ridge regression estimator is **not** unbiased:

$$\mathbb{E}(\hat{\beta}) = (X^TX + \lambda I)^{-1}X^TX\beta.$$

Ridge regression estimator shrinks variance:

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1} \searrow \text{ as } \lambda \nearrow.$$

Exercise 43: How does the ridge estimator behave under high collinearity?

Solution 43:
The ridge estimator spreads the weight across the coefficient estimates.

Exercise 44: What is the lasso regression optimization problem? How does the lasso regression behave?

Solution 44:
For a regularization parameter λ,

$$\begin{aligned}\hat{\beta} &= \arg\min_\beta\{||Y - X\beta||_2^2 + \lambda||\beta||_1\} \\ &= \arg\min\{||Y - X\beta||_2^2 : ||\beta||_1 \le c\}\end{aligned}$$

Encourages sparsity by topology of L^1 ball.

Exercise 45: What is the sparsity optimization problem? Why is this optimization problem hard? Which norm could we use to get closer to a sparsity constraint?

Solution 45:
The sparsity optimization problem is for some k,

$$\hat{\beta} = \arg\min_\beta\{||Y - X\beta||_2^2 : \sum_j 1_{\beta_j \neq 0} \le k\}.$$

This is a non-convex optimization problem. Try lasso regression instead, or use a different "norm" as a surrogate ($||\beta||_p = \left(\sum_j |\beta_j|^p\right)^{1/p}$):

$$\hat{\beta} = \arg\min_\beta\{||Y - X\beta||_2^2 + \lambda||\beta||_p\}.$$

As $p \to 0$, we get closer and closer to the sparsity optimization.

Exercise 46: Define missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Define corrupted data.

Solution 46:

MCAR. X_{ij} is missing with probability p_j i.e. each data point of a covariate is missing uniformly at random.

MAR. X_{ij} has probability p_{ij} of being missing i.e. every point of every covariate has some probability of being missing, and these probabilities may be different. Different individuals have different probabilities of a missing value, but these probabilities are a function of the *observed data* (probabilities can be modeled.)

MNAR. Probabilities of missing values can't be modeled based on observed values.

Corrupted. When data is wrong / tampered.

Exercise 47: How might we deal with missing data? Give three methods and observe how standard errors and coefficient estimates change depending on the method.

Solution 47:

Ignoring missing values. Just ignoring the missing values. This has the effect of increasing the standard errors because we have less data points to work with.

Imputing with sample mean. Impute sample mean $\overline{X_j}$ for missing values. This has the effect of weakening the relationship between the response and the covariates, as well as weakening the correlations between the covariates.

Imputing with regression. Regress X_j onto the other covariates. This has the effect of weakening the relationship between the response and the covariates, but it will strengthen the correlations between the covariates (overcorrection, which could give overconfident standard errors).

Exercise 48: Write down a treatment coding model for a covariate with three treatment levels (0,1,2) and one continuous covariate. Test if we should have three different slopes or all the same slope for the treatment levels. What would you run in R?

Solution 48:
Our model is

$$Y_i = \alpha_{t(i)} + \beta_{t(i)} X_i, \text{ where } t(i) = \text{treatment group}$$

Using treatment coding, the model becomes

$$\begin{aligned} Y_i =& \beta_0 + \beta_{L1}\mathbb{1}\{t(i)=1\} + \beta_{L2}\mathbb{1}\{t(i)=2\} + \beta_X X + \\ & \beta_{X:L1} X \cdot \mathbb{1}\{t(i)=1\} + \beta_{X:L2} X \cdot \mathbb{1}\{t(i)=2\}. \end{aligned}$$

The slopes for all the treatment levels are equivalent if and only if $\beta_{X:L1} = \beta_{X:L2} = 0$. We can use an F-test, where

$$\begin{aligned} \text{Reduced Model} : Y_i =& \beta_0 + \beta_{L1}\mathbb{1}\{t(i)=1\} + \\ & \beta_{L2}\mathbb{1}\{t(i)=2\} + \beta_X X \end{aligned}$$

In R, I would run `anova(lm(response~T*X))`, and look at the last line for significance.

Exercise 49: True or False: the `anova` command in R conducts a different F-test.

Solution 49:
True. The `anova` command compares the reduced model against the completely full model, whereas the F-test that we learned compares the reduced model against the "full" model.

Exercise 50: Consider a factor model with levels $1, \cdots, K$. Construct a test to ask if X has any association with Y.

Derive at test statistic for testing if $\alpha_k = \alpha_l$ for a pair $k \neq l$. Derive Tukey's Honestly Significant Difference test.

Solution 50:
With treatment coding, our model is

$$Y_i = \beta_0 + \beta_1 \mathbb{1}\{t(i)=2\} + \cdots + \beta_{k-1}\mathbb{1}\{t(i)=k\} + \epsilon.$$

To test if there is any association, we can use an F-test with the null hypothesis $\beta_1 = \cdots = \beta_{k-1} = 0$.

To derive a test statistic for pairwise comparisons, let's return to the vanilla model without treat-

ment coding:

$$Y_i = \alpha_{t(i)} + \epsilon.$$

We want to ask if $\alpha_j = \alpha_k$ for $j \neq k$. Assuming that all the treatment groups have the same variance, consider the sample mean $\bar{\alpha}_k \sim N(\alpha_k, \sigma^2/n_k)$. Since $\bar{\alpha}_j$ is determined by completed different points from $\bar{\alpha}_k$, the two sample means are independent, which means that

$$\bar{\alpha}_j - \bar{\alpha}_k \sim N\left(\alpha_j - \alpha_k, \sigma^2\left(\frac{1}{n_j} + \frac{1}{n_k}\right)\right)$$
$$\iff \frac{(\bar{\alpha}_j - \bar{\alpha}_k) - (\alpha_j - \alpha_k)}{\sigma\sqrt{\frac{1}{n_j} + \frac{1}{n_k}}} \sim N(0,1).$$

Because there are K total levels (i.e. total parameters),

$$\frac{(\bar{\alpha}_j - \bar{\alpha}_k) - (\alpha_j - \alpha_k)}{\hat{\sigma}\sqrt{\frac{1}{n_j} + \frac{1}{n_k}}} = \frac{(\bar{\alpha}_j - \bar{\alpha}_k) - (\alpha_j - \alpha_k)}{SE(\bar{\alpha}_j - \bar{\alpha}_k)} \tag{1}$$

$$\sim t_{n-K}. \tag{2}$$

While this can be used to construct a confidence interval, there are $\binom{K}{2}$ pairs to test, which grows quadratically; we need to apply the Bonferroni correction in this case since this is a multiple testing problem, and this may be too conservative.

Let $T_k = \frac{\bar{\alpha}_k - \alpha_k}{\hat{\sigma}/\sqrt{n_k}} \sim t_{n-K}$. Then the studentized range distribution is

$$\max_k T_k - \min_k T_k \sim q_{\#\text{ of levels,d.f. in denom}} = q_{K,n-K}.$$

Let $q_{1-\alpha}$ be the $1-\alpha$ quantile value. If all the group sizes are the same, then

$$(1) = \frac{T_j - T_k}{\sqrt{2}} \leq \frac{|T_j - T_k|}{\sqrt{2}} \leq \frac{\max_k T_k - \min_k T_k}{\sqrt{2}}.$$

Hence we have the following confidence interval for **any pair** $j \neq k$:

$$(\alpha_j - \alpha_k) \pm \frac{q_{1-\alpha}}{\sqrt{2}} \cdot SE(\bar{\alpha}_j - \bar{\alpha}_k).$$

Exercise 51: Describe the scenarios for block randomization and complete randomization.

Solution 51:
Block randomization is when you randomly assign treatments within each group, and complete randomization is when you randomly assign treatments across the entire set.

Exercise 52: Analyze the following model matrix:

$$\begin{bmatrix}\text{BR w/ block effect} & \text{BR w/o block effect}\\ \text{CR w/ block effect} & \text{CR w/o block effect}\end{bmatrix}$$

Solution 52:
Enumerating all four models:

1. BR with block effect: $Y_i = \mu + \alpha_{b(i)} + \beta_{t(i)} + \epsilon$
2. BR without block effect: $Y_i = \mu + \beta_{t(i)} + \epsilon$
3. CR with block effect: $Y_i = \mu + \alpha_{b(i)} + \beta_{t(i)} + \epsilon$, but with the treatments randomly assigned with no regard to block.
4. CR without block effect: $Y_i = \mu + \beta_{t(i)} + \epsilon$, but with the treatments randomly assigned with no regard to block.

Block randomization can reduce variance. For block randomization, there exists a permutation matrix Π such that

$$\Pi X = \begin{bmatrix}\mathbf{1}_{n/4} & \mathbf{0}_{n/4} & \mathbf{0}_{n/4}\\ \mathbf{1}_{n/4} & \mathbf{0}_{n/4} & \mathbf{1}_{n/4}\\ \mathbf{1}_{n/4} & \mathbf{1}_{n/4} & \mathbf{0}_{n/4}\\ \mathbf{1}_{n/4} & \mathbf{1}_{n/4} & \mathbf{1}_{n/4}\end{bmatrix}, \text{ so}$$
$$X^TX = \begin{bmatrix}n & n/2 & n/2\\ n/2 & n/2 & n/4\\ n/2 & n/4 & n/2\end{bmatrix}.$$

For complete randomization, we get that

$$X^TX = \begin{bmatrix}n & n/2 & n/2\\ n/2 & n/2 & n/4+m\\ n/2 & n/4+m & n/2\end{bmatrix}.$$

We get that $(X^TX)^{-1}_{33}$ under complete randomization is greater than or equal to $(X^TX)^{-1}_{33}$ under block randomization (equality if and only if $m = 0$).

To include or not to include blocks. Complete randomization with and without blocks are both valid models. The decision to include blocks (or not include) can effect the variance of our coefficient estimate of the treatment effect.

Take the complete randomization model with blocks, where A_i is a random variable that denotes which block point i is a part of. We have

$$Y_i = \mu + A_i + \beta_{t(i)} + \epsilon, \text{ where } \epsilon \sim N(0,1).$$

Define

$$\bar{\alpha} = \sum_{b=1}^{B} \pi_b \alpha_b \text{ and } \nu^2 = \sum_{b=1}^{B} \pi_b (\alpha_b - \bar{\alpha})^2.$$

Then

$$\begin{aligned} Y_i &= \mu + A_i + \beta_{t(i)} + \epsilon \\ &= \underbrace{(\mu + \bar{\alpha})}_{=\tilde{\mu}} + \beta_{t(i)} + \underbrace{(\epsilon + A_i - \bar{\alpha})}_{=\tilde{\epsilon}} \\ &= \tilde{\mu} + \beta_{t(i)} + \tilde{\epsilon}, \text{ where } \tilde{\epsilon} \sim_{\text{i.i.d.}} N(0, \sigma^2 + \nu^2). \end{aligned}$$

The reason the $\tilde{\epsilon}$'s are iid is because we did complete randomization (i.e. we didn't care about blocks when assigning treatments.) In sum, we see two sources of variance:

1. σ^2 (with block) vs. $\sigma^2 + \nu^2$ (without block).
2. $(X_{\text{big}}^T X_{\text{big}})^{-1}$ (with block) vs. $(X_{\text{small}}^T X_{\text{small}})^{-1}$ (without block). In general, $(X_{\text{big}}^T X_{\text{big}})^{-1} \succeq (X_{\text{small}}^T X_{\text{small}})^{-1}$.

Whether or not the variance is bigger depends on these two factors, which work against each other. If the block effects are very different, then it might be better to include blocks. If instead block effects are all relatively similar, then it might be better to exclude blocks.