Module 1-3 Midterm

Terms in this set (73)

The prediction interval of one member of the population will always be larger than the confidence interval of the mean response for all members of the population when using the same predicting values. (T/F)	True. See 1.7 Regression Line: Estimation & Prediction Examples "Just to wrap up the comparison, the confidence intervals under estimation are narrower than the prediction intervals because the prediction intervals have additional variance from the variation of a new measurement."
In ANOVA, the linearity assumption is assessed using a plot of the response against the predicting variable. (T/F)	False. See 2.2 - Estimation Method Linearity is not an assumption of Anova.

Module 1-3 Midterm

If the model assumptions hold, then the estimator for the variance, σ ^ 2, is a random variable. (T/F)	True See 1.8 Statistical Inference We assume that the error terms are independent random variables. Therefore, the residuals are independent random variables. Since σ ^ 2 is a combination of the residuals, it is also a random variable.
The mean sum of squared errors in ANOVA measures variability within groups. (T/F)	True. See 2.4 Test for Equal Means. MSE = within-group variability
The simple linear regression coefficient, β ^ 0, is used to measure the linear relationship between the predicting and response variables. (T/F)	False See 1.2 Estimation Method β ^ 0 is the intercept and does not tell us about the relationship between the predicting and response variables.
The sampling distribution for the variance estimator in simple linear regression is χ 2 (chi-squared) regardless of the assumptions of the data. (T/F)	False See 1.2 Estimation Method "The sampling distribution of the estimator of the variance is chi-squared, with n - 2 degrees of freedom (more on this in a moment). This is under the assumption of normality of the error terms."

Module 1-3 Midterm

$\hat{\beta}_1$ is an unbiased estimator for β_0. (T/F)	False See 1.4 Statistical Inference "What that means is that $\hat{\beta}_1$ is an unbiased estimator for β_1." It is not an unbiased estimator for β_0.
If the pairwise comparison interval between groups in an ANOVA model includes zero, we conclude that the two means are plausibly equal. (T/F)	True See 2.8 Data Example If the comparison interval includes zero, then the two means are not statistically significantly different, and are thus, plausibly equal.
Under the normality assumption, the estimator for β_1 is a linear combination of normally distributed random variables. (T/F)	True See 1.4 Statistical Inference "Under the normality assumption, β_1 is thus a linear combination of normally distributed random variables... $\hat{\beta}_0$ is also linear combination of random variables"
An ANOVA model with a single qualitative predicting variable containing k groups will have k + 1 parameters to estimate. (T/F)	True See 2.2 Estimation Method We have to estimate the means of the k groups and the pooled variance estimator, s_{pooled}^2.

Module 1-3 Midterm

In simple linear regression models, we lose three degrees of freedom when estimating the variance because of the estimation of the three model parameters β_0, β_1, σ^2. (T/F)	False. See 1.2 Estimation Method "The estimator for σ^2 is $\hat{\sigma}^2$, and is the sum of the squared residuals, divided by n - 2." We lose two degrees of freedom because the variance estimator, $\hat{\sigma}^2$, uses only the estimates for β_0, and β_1 in its calculation.
The pooled variance estimator, s_{pooled}^2, in ANOVA is synonymous with the variance estimator, $\hat{\sigma}^2$, in simple linear regression because they both use mean squared error (MSE) for their calculations. (T/F)	True. See 1.2 Estimation Method for simple linear regression See 2.2 Estimation Method for ANOVA The pooled variance estimator is, in fact, the variance estimator.
The normality assumption states that the response variable is normally distributed. (T/F)	True. See 1.8 Diagnostics "Normality assumption: the error terms are normally distributed." The response may or may not be normally distributed, but the error terms are assumed to be normally distributed.

Module 1-3 Midterm

If the constant variance assumption in ANOVA does not hold, the inference on the equality of the means will not be reliable. (T/F)	True See 2.8 Data Example "This is important since without a good fit, we cannot rely on the statistical inference." Only when the model is a good fit, i.e. all model assumptions hold, can we rely on the statistical inference.
A negative value of $\beta 1$ is consistent with an inverse relationship between the predictor variable and the response variable. (T/F)	True See 1.2 Estimation Method "A negative value of $\beta 1$ is consistent with an inverse relationship"
The p-value is a measure of the probability of rejecting the null hypothesis. (T/F)	False See 1.5 Statistical Inference Data Example "p-value is a measure of how rejectable the null hypothesis is... It's not the probability of rejecting the null hypothesis, nor is it the probability that the null hypothesis is true."

Module 1-3 Midterm

We assess the constant variance assumption by plotting the error terms, ϵ_i, against fitted values. (T/F)	False See 1.2 Estimation Method "We use $\hat{\epsilon}_i$ as proxies for the deviances or the error terms. We don't have the deviances because we don't have β_0 and β_1.
With the Box-Cox transformation, when $\lambda = 0$ we do not transform the response. (T/F)	False. See 1.8 Diagnostics When $\lambda = 0$, we transform using the normal log.
The sampling distribution of $\hat{\beta}_0$ is a: A. t-distribution B. chi-squared distribution C. normal distribution D. None of the above	A. t-distribution See 1.4 Statistical Inference The distribution of β_0 is normal. Since we are using a sample and not the full population, the sampling distribution of $\hat{\beta}_0$ is the t-distribution.

Module 1-3 Midterm

A data point far from the mean of the x's and y's is always:

A. an influential point and an outlier

B. a leverage point but not an outlier

C. an outlier and a leverage point

D. an outlier but not a leverage point

E. None of the above

E. None of the Above.

See 1.9 Outliers and Model Evaluation

We only know that the data point is far from the mean of x's and y's. It only fits the definition of a leverage point because the only information we know is that it is far from the mean of the x's. So you can eliminate the answers that do not include a leverage point. That leaves us with remaining possibilities, "a leverage point but not an outlier" and "an outlier and a leverage point" , both of which we can eliminate. We do not have enough information to know if it is or is not an outlier . None of the answers above fit the criteria of it being always being a leverage point.

The alternative hypothesis of ANOVA can be stated as,

A. the means of all pairs of groups are different

B. the means of all groups are equal

C. the means of at least one pair of groups is different

D. None of the above

C. the means of at least one pair of groups is different

See 2.4 Test for Equal Means

"Using the hypothesis testing procedure for equal means, we test: The null hypothesis, which that the means are all equal (mu 1 = mu 2...=mu k) versus the alternative hypothesis, that some means are different. Not all means have to be different for the alternative hypothesis to be true -- at least one pair of the means needs to be different."

Module 1-3 Midterm

The F-test is a _____ tailed test with ____ and ____ degrees of freedom.

A. one, k, N-1

B. one, k-1, N-k

C. two, k-1, N-k

D. two, k, N-1

E. None of the above.

B. one, k-1, N-k

See 2.4 Test for Equal Means

The F-test is a one tailed test that has two degrees of freedom, namely k - 1 and N - k.

To test if a coefficient is less than a critical value, C, we conduct a one-sided test on the _____ tail of a _____ distribution.

A. left, normal

B. left, t

C. right, normal

D. right, t

E. None of the above

B. left, t

See 1.4 Statistical Inference

"For β_1 greater than zero we're interested on the right tail of the distribution of the $\hat{\beta}_1$."

A multiple linear regression model contains 6 quantitative predicting variables and an intercept. The number of parameters to estimate in this model is 7. (T/F)

False.

See Lesson 3.2: Basic Concepts

The number of parameters to estimate in a multiple linear regression model containing 6 quantitative predicting variables and an intercept is 8: 7 regression coefficients ($\beta_0,\beta_1,...,\beta_6$) and the variance of the error terms (σ^2).

Module 1-3 Midterm

In multiple linear regression, the estimated regression coefficient corresponding to a quantitative predicting variable is interpreted as the estimated expected change in the response variable when there is a change of one unit in the corresponding predicting variable holding all other predictors fixed. (T/F)	True. See Lesson 3.4: Model Interpretation "The estimated value for one of the regression coefficient β_i represents the estimated expected change in y associated with one unit of change in the corresponding predicting variable, X_i, holding all else in the model fixed."
A partial F-Test can be used to test whether the regression coefficients associated with a subset of the predicting variables in a multiple linear regression model are all equal to zero. (T/F)	True. See Lesson 3.7: Testing for Subsets of Regression Parameters We use the Partial F-test to test the null hypothesis that the regression coefficients associated to a subset of the predicting variables are all equal to zero. The alternative hypothesis is that at least one of these regression coefficients is not zero.
The estimated variance of the error terms of a multiple linear regression model with intercept can be obtained by summing up the squared residuals and dividing the sum by n - p , where n is the sample size and p is the number of predictors. (T/F)	False. See Lesson 3.3: Regression Parameter Estimation The estimated variance of the error terms of a multiple linear regression model with intercept should be obtained by summing up the squared residuals and dividing that by n-p-1, where n is the sample size and p is the number of predictors as we lose p+1 degrees of freedom when we estimate the p coefficients and 1 intercept.

Module 1-3 Midterm

For a given predicting variable, the corresponding estimated regression coefficient will likely be different in a conditional model versus a marginal model. (T/F)	True. See Lesson 3.4: Model Interpretation "Importantly, the estimated regression coefficients for the conditional and marginal relationships can be different, not only in magnitude but also in sign or direction of the relationship."
In the case of multiple linear regression, controlling variables are used to control for sample bias. (T/F)	True. See Lesson 3.4: Model Interpretation "Controlling variables can be used to control for bias selection in a sample."
Conducting t-tests on each β parameter in a multiple linear regression model is the preferable to an F-test when testing the overall significance of the model. (T/F)	False. See Lesson 3.7: Testing for Subsets of Coefficients "We cannot and should not select the combination of predicting variables that most explains the variability in the response based on the t-tests for statistical significance because the statistical significance depends on what other variables are in the model."

Module 1-3 Midterm

An example of a multiple linear regression model is Analysis of Variance (ANOVA). (T/F)	True. See Lesson 3.2 Basic Concepts "Earlier, we contrasted the simple linear regression model with the ANOVA model... Multiple linear regression is a generalization of both models."
Given a a quantitative predicting variable and a qualitative predicting variable with 7 categories in a linear regression model with intercept, 7 dummy variables need to be included in the model. (T/F)	False. See Lesson 3.2: Basic Concepts We only need 7 dummy variables. "When we have qualitative variables with k levels, we only include k-1 dummy variables if the regression model has an intercept."
It is good practice to create a multiple linear regression model using a linearly dependent set of predictor variables. (T/F)	True. See Lesson 3.13: Model Evaluation and Multicollinearity It is good practice to create a multiple linear regression model using a linearly independent set of predicting variables. "XTX is not invertible if the columns of X are linearly dependent, i.e. one predicting variable, corresponding to one column, is a linear combination of the others."

Module 1-3 Midterm

The causation of a predicting variable to the response variable can be captured using multiple linear regression on observational data, conditional of other predicting variables in the model. (T/F)	False. See Lesson 3.4 Model Interpretation "This is particularly prevalent in a context of making causal statements when the setup of the regression does not allow so. Causality statements can only be made in a controlled environment such as randomized trials or experiments. "
For a multiple linear regression model to be a good fit, we need the linearity assumption to hold for only one of the predicting variables. (T/F)	False See Lesson 3.11: Assumptions and diagnostics In multiple linear regression, we need the linearity assumption to hold for all of the predicting variables, for the model to be a good fit. "For example, if the linearity does not hold with one or more predicting variables, then we could transform the predicting variables to improve the linearity assumption."
Multicollinearity among the predicting variables will not impact the standard errors of the estimated regression coefficients. (T/F)	False. See Lesson 3.13: Multicollinearity Multicollinearity in the predicting variables can impact the standard errors of the estimated coefficients. "However, the bigger problem is that the standard errors will be artificially large."

Module 1-3 Midterm

The presence of certain types of outliers, such as influential points, can impact the statistical significance of some of the regression coefficients. (T/F)	True. See Lesson 3.11: Assumptions and diagnostics Outliers that are influential can impact the statistical significance of the beta parameters.
Multicollinearity in multiple linear regression means that the rows in the design matrix are (nearly) linearly dependent. (T/F)	False. See Lesson 3.13: Model Evaluation and Multicollinearity Multicollinearity in multiple linear regression means that the columns in the design matrix are (nearly) linearly dependent.
A linear regression model has high explanatory power if the coefficient of determination is close to 1. (T/F)	True. See Lesson 3.3.13: Model Evaluation and Multicollinearity If R2 is close to 1, almost all of the variability in Y can be explained by the linear regression model; hence, the model has high explanatory power.
In multiple linear regression, the prediction of the response variable and the estimation of the mean response have the same interpretation. (T/F)	False. See Lesson 3.2.9: Regression Line and Predicting a New Response. In multiple linear regression, the prediction of the response variable and the estimation of the mean response do not have the same interpretation.

Module 1-3 Midterm

Cook's distance (Di) measures how much the fitted values in a multiple linear regression model change when the ith observation is removed. (T/F)	True. See Lesson 3.11: Assumptions and Diagnostics "This is the distance between the fitted values of the model with all the observations versus the fitted values of the model discarding the i-th observation from the data used to fit the model. "
If the residuals are not normally distributed, we can model the transformed response variable instead, where a common transformation for normality is the Box-Cox transformation. (T/F)	True. See Lesson 3.3.11: Assumptions and Diagnostics If the normality assumption does not hold, we can use a transformation that normalizes the response variable such as Box-Cox transformation.
In multiple linear regression, a VIF value of 6 for a predictor means that 90% of the variation in that predictor can be modeled by the other predictors. (T/F)	False. See Lesson 3.13: Model Evaluation and Multicollinearity A VIF value of 6 for a predictor means that 83.3% of the variation in that predictor can be modeled by the other predictors in the model.

Module 1-3 Midterm

Assuming that the data are normally distributed, the estimated variance has the following sampling distribution under the simple linear model:

A. Chi-square with n-2 degrees of freedom

B. T-distribution with n-2 degrees of freedom

C. Chi-square with n degrees of freedom

D. T-distribution with n degrees of freedom

A. Chi-square with n-2 degrees of freedom

1.1 - Knowledge Check 1

The fitted values are defined as:

A. The difference between observed and expected responses.

B. The regression line with parameters replaced with the estimated regression coefficients.

C. The regression line.

D. The response values.

B. The regression line with parameters replaced with the estimated regression coefficients.

1.1 - Knowledge Check 1

Module 1-3 Midterm

The estimators of the linear regression model are derived by:

A. Minimizing the sum of squared differences between observed and expected values of the response variable.

B. Maximizing the sum of squared differences between observed and expected values of the response variable.

C. Minimizing the sum of absolute differences between observed and expected values of the response variable.

D. Maximizing the sum of absolute differences between observed and expected values of the response variable.

A. Minimizing the sum of squared differences between observed and expected values of the response variable.

1.1 - Knowledge Check 1

The estimators for the regression coefficients are:

A. Biased but with small variance

B. Biased with large variance

C. Unbiased under normality assumptions but biased otherwise.

D. Unbiased regardless of the distribution of the data.

D. Unbiased regardless of the distribution of the data.

1.2 - Knowledge Check 2

Module 1-3 Midterm

The assumption of normality:

A. It is needed for deriving the estimators of the regression coefficients.

B. It is not needed for linear regression modeling and inference.

C. It is needed for the sampling distribution of the estimators of the regression coefficients and hence for inference.

D. It is needed for deriving the expectation and variance of the estimators of the regression coefficients.

C. It is needed for the sampling distribution of the estimators of the regression coefficients and hence for inference.

1.2 - Knowledge Check 2

The estimated versus predicted regression line for a given x*:

A. Have the same variance

B. Have the same expectation

C. Have the same variance and expectation

D. None of the above

B. Have the same expectation

1.2 - Knowledge Check 3

The variability in the prediction comes from:

A. The variability due to a new measurement.

B. The variability due to estimation

C. The variability due to a new measurement and due to estimation.

D. None of the above.

C. The variability due to a new measurement and due to estimation.

1.2 - Knowledge Check 3

Module 1-3 Midterm

Which one is correct?

A. Independence assumption can be assessed using the residuals vs fitted values.

B. Independence assumption can be assessed using the normal probability plot.

C. Residual analysis can be used to assess uncorrelated errors.

D. None of the above

C. Residual analysis can be used to assess uncorrelated errors.

1.3 - Knowledge Check 4

We detect departure from the assumption of constant variance

A. When the residuals increase as the fitted values increase also.

B. When the residuals vs fitted are scattered randomly around the zero line.

C. When the histogram does not have a symmetric shape.

D. All of the above.

A. When the residuals increase as the fitted values increase also.

1.3 - Knowledge Check 4

Module 1-3 Midterm

Which one is correct?

A. If a departure from normality is detected, we transform the predicting variable to improve upon the normality assumption.

B. If a departure from the independence assumption is detected, we transform the response variable to improve upon this assumption.

C. The Box-Cox transformation is commonly used to improve upon the linearity assumption.

D. None of the above

D. None of the above.

1.3 - Knowledge Check 4

In evaluating a simple linear model

A. There is a direct relationship between the coefficient of determination and the correlation between the predicting and response variables.

B. The coefficient of determination is interpreted as the percentage of variability in the response variable explained by the model.

C. Residual analysis is used for goodness of fit assessment.

D. All of the above.

D. All of the Above

1.3 - Knowledge Check 4

Module 1-3 Midterm

Which are all the model parameters in ANOVA?

A. The means of the k populations.

B. The sample means of the k populations.

C. The sample means of the k samples.

D. None of the above.

D. None of the above.

2.1 - Knowledge Check 1

The pooled variance estimator is:

A. The variance estimator assuming equal variances.

B. The variance estimator assuming equal means and equal variances.

C. The sample variance estimator assuming equal means.

D. None of the above.

A. The variance estimator assuming equal variances.

2.1 - Knowledge Check 1

The total sum of squares divided by N-1 is

A. The mean sum of squared errors

B. The sample variance estimator assuming equal means and equal variances

C. The sample variance estimator assuming equal variances.

D. None of the above.

B. The sample variance estimator assuming equal means and equal variances

2.1 - Knowledge Check 2

Module 1-3 Midterm

The mean squared errors (MSE) measures:
A. The within-treatment variability.
B. The between-treatment variability.
C. The sum of the within-treatment and between-treatment variability.
D. None of the above.

A. The within-treatment variability.

2.1 - Knowledge Check 2

Which is correct?
A. If we reject the test of equal means, we conclude that all treatment means are not equal.
B. If we do not reject the test of equal means, we conclude that means are definitely all equal
C. If we reject the test of equal means, we conclude that some treatment means are not equal.
D. None of the above.

C. If we reject the test of equal means, we conclude that some treatment means are not equal.

2.1 - Knowledge Check 2

The objective of the residual analysis is
A. To evaluate goodness of fit
B. To evaluate whether the means are equal.
C. To evaluate whether only the normality assumptions holds.
D. None of the above.

A. To evaluate goodness of fit

2.2 - Knowledge Check 3

Module 1-3 Midterm

The objective of the pairwise comparison is
A. To find which means are equal.
B. To identify the statistically significantly
different means.
C. To find the estimated means which are
greater or lower than other.
D. None of the above.

B. To identify the statistically significantly different means.

2.2 - Knowledge Check 3

The objective of multiple linear regression is
A. To predict future new responses
B. To model the association of explanatory
variables to a response variable accounting
for controlling factors.
C. To test hypothesis using statistical
inference on the model.
D. All of the above.

D. All of the above.

3.1 - Knowledge Check 1

Module 1-3 Midterm

Which one is correct?

A. A multiple linear regression model with p predicting variables but no intercept has p model parameters.

B. The interpretation of the regression coefficients is the same whether or not interaction terms are included in the model.

C. Multiple linear regression is a general model encompassing both ANOVA and simple linear regression.

D. None of the above.

C. Multiple linear regression is a general model encompassing both ANOVA and simple linear regression.

3.1 - Knowledge Check 1

Which one is correct?

A. The regression coefficients can be estimated only if the predicting variables are not linearly dependent.

B. The estimated regression coefficient $\hat{\beta}_i$ is interpreted as the change in the response variable associated with one unit of change in the i-th predicting variable .

C. The estimated regression coefficients will be the same under marginal and conditional model, only their interpretation is not.

D. Causality is the same as association in interpreting the relationship between the

A. The regression coefficients can be estimated only if the predicting variables are not linearly dependent.

3.1 - Knowledge Check 1

Module 1-3 Midterm

Which one correctly characterizes the sampling distribution of the estimated variance?

A. The estimated variance of the error term has a χ^2 distribution regardless of the distribution assumption of the error terms.

B. The number of degrees of freedom for the χ^2 distribution of the estimated variance is n-p-1 for a model without intercept.

C. The sampling distribution of the mean squared error is different of that of the estimated variance.

D. None of the above.

D. None of the above.

3.1 - Knowledge Check 1

The sampling distribution of the estimated regression coefficients is

A. Centered at the true regression parameters.

B. The t-distribution assuming that the variance of the error term is unknown an replaced by its estimate.

C. Dependent on the design matrix.

D. All of the above.

D. All of the above.

3.2 - Knowledge Check 2

Module 1-3 Midterm

The estimators for the regression coefficients are:

A. Biased but with small variance

B. Unbiased under normality assumptions but biased otherwise.

C. Biased regardless of the distribution of the data.

D. Unbiased regardless of the distribution of the data.

D. Unbiased regardless of the distribution of the data.

3.2 - Knowledge Check 2

We can test for a subset of regression coefficients

A. Using the F statistic test of the overall regression.

B. Only if we are interested whether additional explanatory variables should be considered in addition to the controlling variables.

C. To evaluate whether all regression coefficients corresponding to the predicting variables excluded from the reduced model are statistically significant.

D. None of the above.

D. None of the above.

3.2 - Knowledge Check 2

Module 1-3 Midterm

Which one is correct?

A. The prediction intervals need to be corrected for simultaneous inference when multiple predictions are made jointly.

B. The prediction intervals are centered at the predicted value.

C. The sampling distribution of the prediction of a new response is a t-distribution.

D. All of the above.

3.2 - Knowledge Check 3

In evaluating a multiple linear model,

A. The F test is used to evaluate the overall regression.

B. The coefficient of determination is interpreted as the percentage of variability in the response variable explained by the model.

C. Residual analysis is used for goodness of fit assessment.

D. All of the above.

D. All of the Above

3.3 - Knowledge Check 4

Module 1-3 Midterm

In the presence of near multicollinearity,

A. The coefficient of determination decreases.

B. The regression coefficients will tend to be identified as statistically significant even if they are not.

C. The prediction will not be impacted.

D. None of the above.

D. None of the Above

3.3 - Knowledge Check 4

When do we use transformations?

A. If the linearity assumption with respect to one or more predictors does not hold, then we use transformations of the corresponding predictors to improve on this assumption.

B. If the normality assumption does not hold, we transform the response variable, commonly using the Box-Cox transformation.

C. If the constant variance assumption does not hold, we transform the response variable.

D. All of the above.

D. All of the Above

3.3 - Knowledge Check 4

Module 1-3 Midterm

Which one is correct?

A. The residuals have constant variance for the multiple linear regression model.

B. The residuals vs fitted can be used to assess the assumption of independence.

C. The residuals have a t-distribution distribution if the error term is assumed to have a normal distribution.

D. None of the above.

D. None of the Above

3.3 - Knowledge Check 4