Exam 01 Review

Solutions

02.24.20

The Data

For this exam review, we will analyze a sample of 300 diamonds that weigh less than 1.1 carats. The data is from the `diamonds` data set in the ggplot2 R package. Type `?ggplot2::diamonds` in the console to read more about the variables in this data set.

We load and prepare the data in the code below.

Part I: Analysis of Variance

How much variation in `price` can be explained by the `clarity` of the diamond?

term	df	sumsq	meansq	statistic	p.value
clarity	7	79659890	11379984	3.398	0.002
Residuals	292	977793811	3348609	NA	NA

1. What is the estimated variance of the distribution of price within each level of clarity?

   The mean square within each group, i.e. mean square of the residuals, 3348609.

2. State the null and alternative hypotheses for the test conducted in the ANOVA table.

   Let $\mu_1, \mu_2, \ldots, \mu_8$ represent the mean price in each level of clarity. Then the hypotheses are

   $$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8$$
   $$H_a : \text{at least one } \mu_i \text{ has a mean price that is not equal to the others}$$

3. Briefly explain how the F test statistic is calculated.

   The F test statistic is MSB / MSW = 11379984 / 3348609.

4. State your conclusion for the test in (2) in the context of the data.

   The p-value 0.002 is small, so we reject H_0 and conclude that the mean price is different for at least one level of clarity.

5. Briefly describe some of the next steps you would take in the analysis.

   We could examine the coefficients and the corresponding p-values and confidence intervals for the model with price as the response and clarity as the predictor. This would give us an idea about how the mean price at each level compares to the baseline level of clarity in the model.

   Another option is to calculate confidence intervals for the mean price for each level of clarity and examine if there is overlap in the intervals.

Part II: Original Model

To better understand how different characteristics of a diamond affect its price, we fit a model with `price` as the response and `caratCent`, `depthCent`, `color`, and the interaction between `caratCent` and `color` as the predictor variables. `caratCent` is the mean-centered version of `carat` and `depthCent` is the mean-centered version of `depth`.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2592.722	122.824	21.109	0.000	2350.965	2834.479
caratCent	7100.854	476.744	14.894	0.000	6162.468	8039.241
depthCent	21.187	33.783	0.627	0.531	-45.309	87.682
colorE	-141.932	170.623	-0.832	0.406	-477.773	193.909
colorF	-340.051	157.736	-2.156	0.032	-650.526	-29.576
colorG	-130.668	158.902	-0.822	0.412	-443.438	182.103
colorH	-466.661	173.258	-2.693	0.007	-807.690	-125.633
colorI	-710.263	197.693	-3.593	0.000	-1099.388	-321.139
colorJ	-980.402	261.296	-3.752	0.000	-1494.717	-466.086
caratCent:colorE	-139.878	664.856	-0.210	0.834	-1448.530	1168.774
caratCent:colorF	-1253.846	596.371	-2.102	0.036	-2427.696	-79.996
caratCent:colorG	-88.720	603.041	-0.147	0.883	-1275.699	1098.260
caratCent:colorH	-1777.497	629.878	-2.822	0.005	-3017.299	-537.695
caratCent:colorI	-1832.264	739.703	-2.477	0.014	-3288.239	-376.289
caratCent:colorJ	-2267.281	1009.529	-2.246	0.025	-4254.361	-280.202

1. Interpret `caratCent` and its 95% confidence interval in the context of the data.

   For each additional carat, we expect the price of a diamond to increase by about 7100.85, on average, holding the color and depth constant.

2. Suppose you fit a new model that includes `carat` instead of `caratCent` as a predictor. All other predictors remain the same as in the model output above. Briefly describe how the estimate of the coefficient of `carat`, the corresponding test statistic, p-value, and confidence interval would compare to those shown in the model output above.

   Nothing would change for `carat`.

3. Interpret `colorF` and its 95% confidence interval in the context of the data.

   Holding carat and depth constant, we expect the mean price of diamonds of color F to be about $340.05 lower than the mean price of diamonds with color D

   We are 95% confident that the mean price for diamonds with color F is between $29.78 to $650.53 lower than the mean price for diamonds with color D, holding carat and depth constant.

4. Interpret `caratCent:colorF` and its 95% confidence interval in the context of the data.

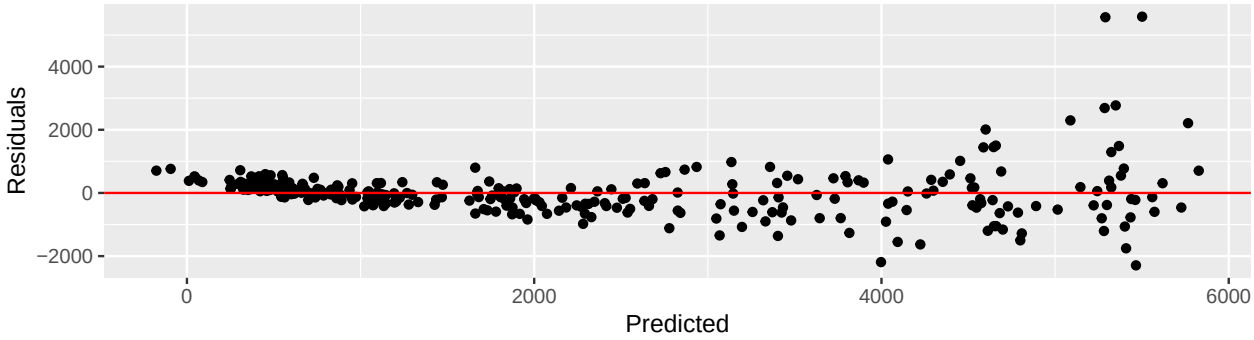
   We expect the change in the mean price for each additional carat to be about $1253.846 less for diamonds with color F than those with color D, holding depth constant.

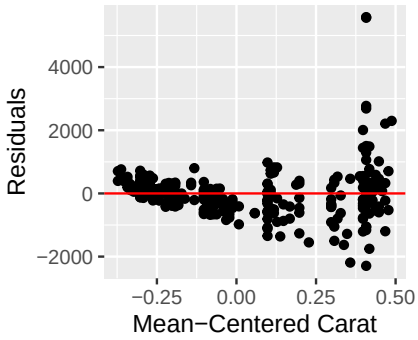
   We are 95% confident that the change in the mean price for each additional carat to be about $80 to $2427.790 less for diamonds with color F than those with color D, holding depth constant.

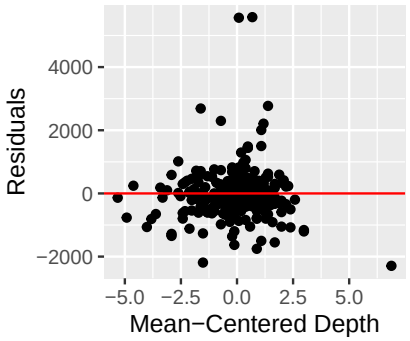
5. Compare the interpretations in questions 3 and 4. Discuss the differences between `colorF` and `caratCent:colorF` and how they affect the model.

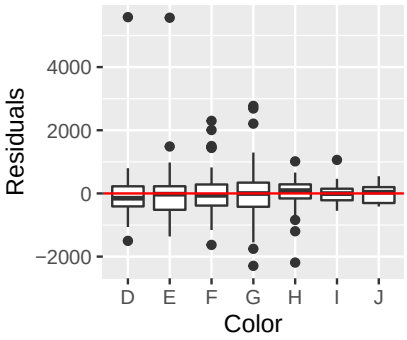
   The main effect `colorF` affects the intercept of the model and the interaction term `caratCent:colorF` affects the coefficient (slope) of `caratCent`.

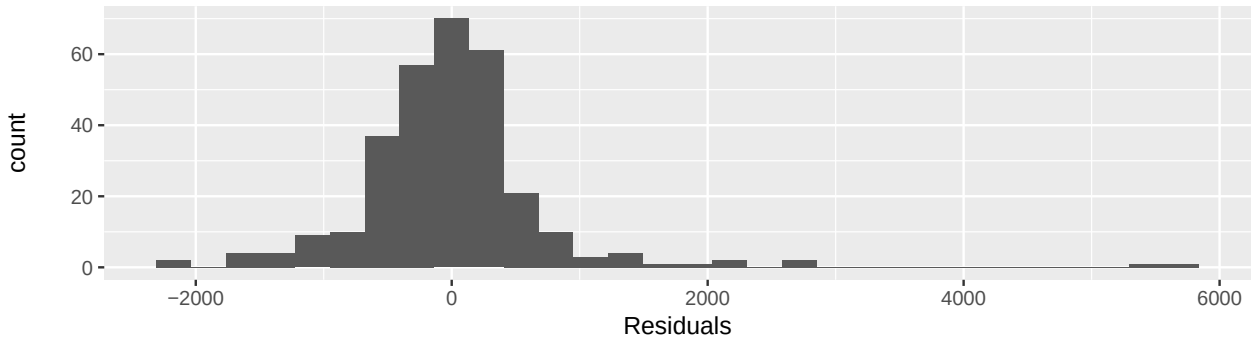
6. What are the assumptions of multiple linear regression? Some of the residuals plots used to check the model assumptions are shown below. For each plot, State the assumption(s) that can be assessed using that plot. Use the plot to determine if the assumption is satisfied.

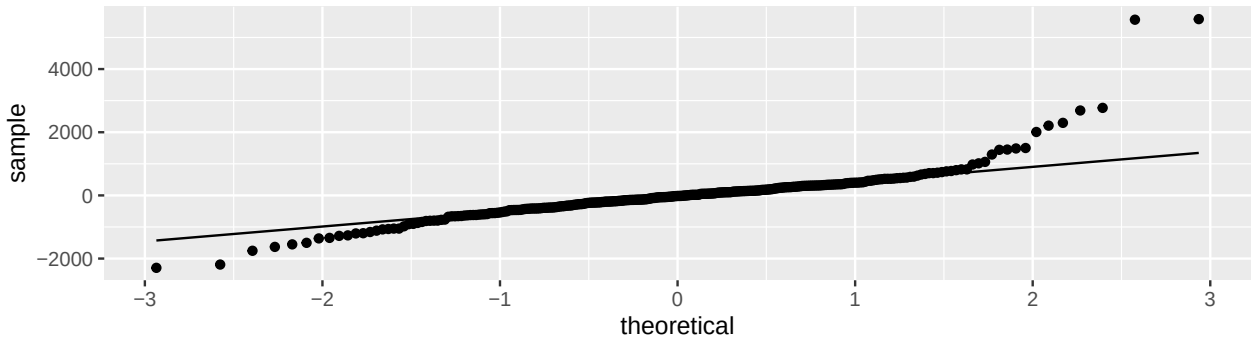












- **Linearity**:
  - *Use plots of residuals vs. predicted and residuals vs. each quantitative predictor. There is no obvious relationship between the values on the x axis and those on the y axis (i.e. a parabola), so we conclude that linearity is satisfied.*
- **Normality**:
  - *Use histogram and normal QQ plot of residuals. There seems to be some right-skewness and high*

outliers, so this assumption is not satisfied.

- **Constant variance**:
  - *Use the plot of the residuals vs. predicted. There is a fan shape in this plot, so this assumption is not satisfied. We should try a log-transformation on the response variable.*
- **Independence**:
  - *Most of the time, you can use a description of the data to determine if the independence assumption is reasonably met. You can can also plot the residuals in time or location order (if relevant). Can also examine boxplots of the residuals vs. a categorical predictor to see if the model systematically over or under predicts for a group. There is nothing to indicate a violation of independence for this dataset, so we conclude this assumption is satisfied.*

Part III: Model with Log-Transformed Response

Next, we fit a model with the `log_price`, the log-transformed version of `price`, as the response variable. We use the same predictor variables as before.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	7.533	0.037	201.531	0.000	7.459	7.606
caratCent	2.948	0.145	20.317	0.000	2.662	3.233
depthCent	-0.004	0.010	-0.355	0.723	-0.024	0.017
colorE	-0.045	0.052	-0.868	0.386	-0.147	0.057
colorF	-0.100	0.048	-2.087	0.038	-0.195	-0.006
colorG	-0.048	0.048	-0.988	0.324	-0.143	0.047
colorH	-0.226	0.053	-4.288	0.000	-0.330	-0.122
colorI	-0.295	0.060	-4.912	0.000	-0.414	-0.177
colorJ	-0.448	0.080	-5.632	0.000	-0.604	-0.291
caratCent:colorE	-0.042	0.202	-0.210	0.834	-0.441	0.356
caratCent:colorF	-0.302	0.181	-1.665	0.097	-0.659	0.055
caratCent:colorG	-0.130	0.184	-0.710	0.478	-0.492	0.231
caratCent:colorH	-0.123	0.192	-0.643	0.520	-0.501	0.254
caratCent:colorI	-0.044	0.225	-0.195	0.846	-0.487	0.399
caratCent:colorJ	0.053	0.307	0.173	0.863	-0.552	0.658

1. Describe the subset of diamonds that are expected to have a median price of exp(7.533) = \$1868.7.

   We are describing the intercept. These are diamonds of color D with average carat weight, and average depth.

2. Interpret the estimated coefficient of `depthCent` in terms of the price.

   For each percentage increase of depth, we expect the median price to multiply by a factor of $\exp\{-0.004\}$*, holding color and carat constant.*

3. Write the model equation to predict `log(price)` for a diamond that's color `D`.

$$\widehat{\log(price)} = 7.533 + 2.948 \times caratCent - 0.004 \times depthCent$$

4. Write the model equation to predict `log(price)` for a diamond that's color `F`.

$$\widehat{\log(price)} = (7.533 - 0.100) + (2.948 - 0.302) \times caratCent - 0.004 \times depthCent$$

5. What is the slope of `caratCent` for a diamond with color `D`? Color `F`?

   The slope is 2.948 for color D and (2.948 - 0.302) for color F.

6. Describe how the price changes when going from a 1 carat diamond that's color `D` with depth of 60 to a 0.5 carat diamond with color `F` and depth of 60.

The mean of carat is 0.602, and the mean of depth is 61.62

Diamond1

$$\widehat{\log(price)} = 7.533 + 2.948 \times (1 - 0.602) - 0.004 \times (60 - 61.62)$$

Diamond2

$$\widehat{\log(price)} = (7.533 - 0.100) + (2.948 - 0.302) \times (0.5 - 0.602) - 0.004 \times (60 - 61.62)$$

Difference in the log(price) between Diamond2 and Diamond1

$$\big[(7.533-0.100)+(2.948-0.302)\times(0.5-0.602)-0.004\times(60-61.62)\big]-\big[7.533+2.948\times(1-0.602)-0.004\times(60-61.62)\big]$$

This equals

$$-0.100 + 2.948 * (0.5 - 1) + 0.302 * (0.602 - 0.5)$$

Therefore, we expect the median price to be multiplied by a factor of $\exp\{-0.100 + 2.948 \times (0.5 - 1) + 0.302 \times (0.602 - 0.5)\}$ *when going from diamonds with the characteristics of Diamond 1 to those with the characteristics of Diamond 2.*